Towards Robust and Scalable Power System State Estimation

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

***Abstract*— Power system state estimation is an important instance of data-driven decision making in power systems. Yet due to the nonconvexity of the problem, existing approaches based on local search methods are susceptible to spurious local minima. 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 when the amount of measured data is not too low. Furthermore, we develop a robustness metric called "mutual incoherence," which provides robustness guarantees in the presence of bad data. The proposed method demonstrates superior performance over existing methods in terms of both estimation accuracy and bad data detection for an array of benchmark systems. This technique is shown to be scalable to large systems with more than 13,000 nodes and can achieve an accurate estimation within a minute.**

I. INTRODUCTION

Power system state estimation (PSSE) is conducted on a regular basis to monitor the state of the grid by collecting and filtering a wealth of sensor data from transmission and distribution infrastructures [1], [2]. 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 [3], [4]. As a result, there is a long tradition of studying this problem [2], [5]–[11].

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], [5]. However, the estimator is prone to outliers and sparse noise/errors, which can arise from sensor faults, topological errors [12]–[14], or adversarial attack [15], [16]. To deal with large and sparse noise, one common approach is to perform bad data detection (BDD) on residual errors [17]. This method relies on statistical assumptions on the 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], [6], [18]–[21]. 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 a global minimum [21], [22]. Even though some recent works have shed light on the possibility of the non-existence of local minima in certain scenarios [23], the conditions are difficult to verify for PSSE [21].

Recently, the semidefinite programming (SDP) relaxation technique has been applied to PSSE following its success for the optimal power flow problem [24], which has shown a satisfactory numerical performance even in the presence of topological errors and bad data [8], [9], [11], [13]. Theoretical analysis of the estimator has been conducted in [9], [11]. Furthermore, [16] analyzes the vulnerabilities of AC PSSE against potential cyber attacks. While SDP relaxation is a promising approach with both numerical success and theoretical guarantees, this method requires that the solution be rank-1 to recover the true state. Since most interior point methods for solving SDPs produce the highest-rank solution by default, one may need to add an extra rank penalty to the objective function (e.g., nuclear norm [8] or custom-designed norm [9], [11]), which forces the solution to be near-global optimal. Furthermore, the addition of the positive semidefinite constraint limits the solvability of large-scale problems, since most conic numerical algorithms scale on the order of $O(n^6)$, where n is the number of variables.

In this study, we propose a method to solve large-scale AC PSSE with quadratic programming 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 to fully capture the properties of the power grid topology. Furthermore, we also provide a 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 the 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 presentation.

II. POWER SYSTEM AC-MODEL

A. Notations

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

[†]This work was supported by the ONR grant N00014-17-1-2933, 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 significantly extended journal version of this paper has been submitted to the IEEE Transactions on Smart Grid.

$\mathrm{supp}(\boldsymbol{x})$ 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 $\mathcal{J}$. 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. We use $\mathbb{P}$ to denote probability, and $\mathbb{E}$ to denote expectation. The notations $\|\boldsymbol{x}\|_1$, $\|\boldsymbol{x}\|_2$ and $\|\boldsymbol{x}\|_\infty$ represent the 1-norm, 2-form and ∞-norm of $\boldsymbol{x}$.

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 sets 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 g_ℓ^{sh} 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,

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

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 square 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:

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

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

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

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

  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 $\boldsymbol{x}_\natural = (\{x_k^{\mathrm{mg}}\}_{k \in \mathcal{N}}, \{x_\ell^{\mathrm{im}}, x_\ell^{\mathrm{im}}\}_{\ell \in \mathcal{L}}) \in \mathbb{R}^{n_x}$ 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 $i \in [n_m]$. 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 [8], [9], [11], [13], and it efficiently exploits the sparsity of the network.

D. Measurement model

We consider the measurement model as follows:

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

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

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 [25]. Let $\mathcal{J} \coloneqq \operatorname{supp}(\boldsymbol{b}) \subset [n_m]$ 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 $\boldsymbol{Q}_{\mathcal{J}} \coloneqq \begin{bmatrix}\boldsymbol{A} & \boldsymbol{I}_{\mathcal{J}}^\top\end{bmatrix}$, 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 $\mathcal{J}^c$.

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 [26], [27]. In our case, 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 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.

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.

III. Two-Stage State Estimation

This section describes the proposed two-stage state estimation method.

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.

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

In this case, the dense noise is negligible, i.e., $\boldsymbol{w}_\natural = \mathbf{0}$, and 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.

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 $\hat{\boldsymbol{v}} = \boldsymbol{v}$.

IV. Theoretical Analysis

This section presents several theoretical analyses for the proposed framework, where we 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 1. *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 2. *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 1 and 2 are coincident. In the case of negligible dense noise, a strong global recovery is achieved, meaning that 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 2 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.

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 1. *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 $\mathbf{e} \in \mathbb{R}^{n_l}$ *has 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.*

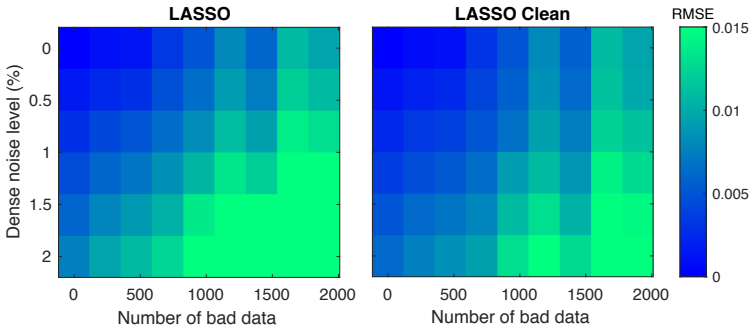
V. Experiments

Numerical evaluations are performed on benchmark systems from MATPOWER [28]. With the exception of the last experiment, 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. Due to space restrictions, although we cannot offer more simulations on cases with a high amount of data but not all possible measurements, we have observed similar behaviors to what to be presented next. All the experiments are performed on a personal laptop with 3.3GHz Intel Core i7 and 16GB memory.

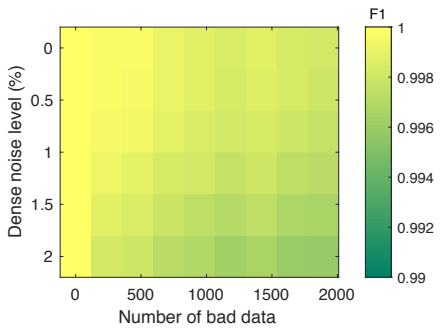
In each case, we randomly generate 50 sets of dense noise $\boldsymbol{w}$ and sparse bad data $\boldsymbol{b}$. The dense noise for each measurement is zero-mean Gaussian variable, with standard deviation of $0.1 \times c_n$ (per unit) for voltage magnitude measurements and c_n (per unit) for all the other measurements, where c_n is the dense noise level. For the sparse bad data, 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 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

TABLE I: 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



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



(b) F1 score for (S1-LASSO)

Fig. 1: 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).

of elements). The F1 score 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 [21]. Throughout the experiment, we choose λ in (S1-LASSO) to be $3 \times 10^{-4}/n_m$, 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 Table I, 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. As is shown in Fig. 1, 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 to large systems with 13659 buses, which is the largest system provided by MATPOWER. We fix the dense noise level to 0.5% and the percentage of bad data to 2%, which amounts to 2457 number of arbitrarily bad measurements. In addition, we experiment with two sets of measurements: case A includes full branch flow measurements and PVQ nodal measurements on PQ buses as well as PV measurements on PV buses; case B has full branch flow measurements and full nodal measurements. In case A and B, the estimator can achieve RMSE of .008 and .006, respectively, and F1 score of .996 and .998, respectively. Moreover, the average time of computation is less than a minute.

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 breaks down the NP-hardness of the PSSE under mild conditions that are 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. A 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 the Newton method with bad data detection scheme and the least mean absolute value regression using ℓ_1 norm. 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 PSSE 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] D. Bienstock and A. Verma, "Strong NP-hardness of AC power flows feasibility," *arXiv preprint arXiv:1512.07315*, 2015.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] R. Baldick, K. Clements, Z. Pinjo-Dzigal, and P. Davis, "Implementing nonquadratic objective functions for state estimation and bad data rejection," *IEEE Transactions on Power Systems*, vol. 12, no. 1, pp. 376–382, 1997.

[21] 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

[22] 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.

[23] 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.

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

[25] M. Jin, J. Lavaei, and K. Johansson, "A semidefinite programming relaxation under false data injection attacks against power grid ac state estimation," in *Proc. of IEEE Annual Allerton Conference on Communication, Control, and Computing*, 2017, pp. 236–243.

[26] 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.

[27] 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.

[28] 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.

APPENDIX

A. Proof of Theorem 1

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

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

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:

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

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

where $\partial\|\boldsymbol{b}_\natural\|_1$ denotes the subgradiet of $\|\boldsymbol{b}_\natural\|_1$. By the definition of $\mathcal{J} \coloneqq \text{supp}(\boldsymbol{b}_\natural)$, we need to find a vector $\boldsymbol{h}_\star$ such that $\boldsymbol{h}_{\star\mathcal{J}} = \text{sign}(\boldsymbol{b}_{\natural\mathcal{J}})$ and $\|\boldsymbol{h}_{\star\mathcal{J}^c}\|_\infty \leq 1$. In fact, we can meet a slightly stronger condition for strict feasibility by choosing $\boldsymbol{h}_{\star\mathcal{J}^c} = -\boldsymbol{A}_{\mathcal{J}^c}^{\top+}\boldsymbol{A}_{\mathcal{J}}^\top \text{sign}(\boldsymbol{b}_{\natural\mathcal{J}})$, 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 $\boldsymbol{Q}_{\mathcal{J}} \coloneqq \begin{bmatrix}\boldsymbol{A} & \boldsymbol{I}_{\mathcal{J}}^\top\end{bmatrix}$ has full column rank. Let $\tilde{\mathcal{J}} = \text{supp}(\tilde{\boldsymbol{b}})$;

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 $(\boldsymbol{x}_\natural, \boldsymbol{b}_\natural)$.

B. Proof of Theorem 2

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 [27]):

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

Lemma 1. *If the PDW procedure succeeds, then $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{b}})$ is the unique optimal solution of (S1-LASSO), where $\hat{\boldsymbol{b}} = (\hat{\boldsymbol{b}}_{\mathcal{J}}, \boldsymbol{0})$.*

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 $F(\boldsymbol{x}, \boldsymbol{b}) = \frac{1}{2n_m} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2$. One can write:

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

and hence,

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

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

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

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

Proof of Theorem 2: 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 reorganized as:

$$\begin{aligned} -\frac{1}{n_m} \begin{bmatrix} \boldsymbol{I}_{\mathcal{J}} \boldsymbol{A} & \boldsymbol{I}_{\mathcal{J}} \boldsymbol{I}_{\mathcal{J}}^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_\natural - \hat{\boldsymbol{x}} \\ \boldsymbol{b}_{\mathcal{J}\natural} - \hat{\boldsymbol{b}}_{\mathcal{J}} \end{bmatrix} \\ - \frac{1}{n_m} \boldsymbol{I}_{\mathcal{J}} \boldsymbol{w}_\natural + \lambda \hat{\boldsymbol{z}}_{\mathcal{J}} = \boldsymbol{0}. \end{aligned} \tag{22}$$

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

$$\hat{\boldsymbol{z}}_{\mathcal{J}^c} = \frac{1}{n_m \lambda} \boldsymbol{I}_{\mathcal{J}^c} \left(\boldsymbol{A}(\boldsymbol{x}_\natural - \hat{\boldsymbol{x}}) + \boldsymbol{w}_\natural \right). \tag{23}$$

Similarly, combining (20) and (22) leads to

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

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

$$\boldsymbol{\Delta} = -(\boldsymbol{Q}_{\mathcal{J}}^\top \boldsymbol{Q}_{\mathcal{J}})^{-1} \boldsymbol{Q}_{\mathcal{J}}^\top \boldsymbol{w}_\natural + n_m \lambda (\boldsymbol{Q}_{\mathcal{J}}^\top \boldsymbol{Q}_{\mathcal{J}})^{-1} \begin{bmatrix} \boldsymbol{0} \\ \hat{\boldsymbol{z}}_{\mathcal{J}} \end{bmatrix}. \tag{24}$$

Recall that $\boldsymbol{I}_x$ and $\boldsymbol{I}_b$ denote the matrices consisting of the first n_x rows and last $|\mathcal{J}|$ rows of the identity matrix of size

$n_x + |\mathcal{J}|$, respectively. Therefore,

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

By the mutual incoherence condition (i.e., $\rho(\mathcal{J}) = 1-\gamma$ for $\gamma > 0$), we have $\|\boldsymbol{\mu}\|_\infty \le 1-\gamma$. 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. It can be verified that

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

due to $\boldsymbol{I}_{\mathcal{J}^c}\boldsymbol{I}_{\mathcal{J}}^\top = \boldsymbol{0}$. Since the elements of $\boldsymbol{w}$ are zero-mean sub-Gaussian with the parameter σ^2 and the projection operator has spectral norm one, it can be concluded that

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

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

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

By the design of λ, we conclude that

$$\mathbb{P}\left(\|\hat{\boldsymbol{z}}_{\mathcal{J}^c}\|_\infty \ge 1-\frac{\gamma}{2}\right) \le 2\exp\left(-c_1n_m^2\lambda^2\right).$$

Part 2): Now, we will bound the estimation error $\boldsymbol{\Delta}$ in (24). First, we bound the infinity norm of $\boldsymbol{b}_{\mathcal{J}\natural} - \hat{\boldsymbol{b}}_{\mathcal{J}} = \boldsymbol{I}_b\boldsymbol{\Delta}$. It follows from the triangle inequality that

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

Since the second term is deterministic, one can 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

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

Thus, by the union bound, we have

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

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

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

with probability greater than $1 - 2\exp(-c_2n_m^2\lambda^2)$. This indicates that 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 $\boldsymbol{x}_\natural - \hat{\boldsymbol{x}} = \boldsymbol{I}_x\boldsymbol{\Delta}$ as

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

For the first term, by the application of standard sub-Gaussian concentration, one can write

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

It can be verified that

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

due to the lower eigenvalue condition (5) and the normalized measurement assumption (6). Similarly,

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

and

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

Together, it can be concluded that

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

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

C. Proof of Proposition 1

The ℓ-th component of the vector $\hat{\boldsymbol{\theta}}_\Delta$ can be written as

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

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

$$|[\hat{\boldsymbol{\theta}}_\Delta]_\ell - [\boldsymbol{\theta}_\Delta]_\ell| \le \left|\frac{\hat{x}_\ell^{\text{im}}x_\ell^{\text{re}} - x_\ell^{\text{im}}\hat{x}_\ell^{\text{re}}}{\hat{x}_\ell^{\text{re}}x_\ell^{\text{re}}}\right| = |e_\ell|$$

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