Maximum Marginal Likelihood Estimation of Phase Connections in Power Distribution Systems

Wenyu Wang, *Student Memeber, IEEE* and Nanpeng Yu, *Senior Member, IEEE*

***Abstract*—Accurate phase connectivity information is essential for advanced monitoring and control applications in power distribution systems. The existing data-driven approaches for phase identification lack precise physical interpretation and theoretical performance guarantee. Their performance generally deteriorates as the complexity of the network, the number of phase connections, and the level of load balance increase. In this paper, by linearizing the three-phase power flow manifold, we develop a physical model, which links the phase connections to the smart meter measurements. The phase identification problem is first formulated as a maximum likelihood estimation problem and then reformulated as a maximum marginal likelihood estimation problem. We prove that the correct phase connection achieves the highest log likelihood values for both problems. An efficient solution method is proposed by decomposing the original problem into subproblems with a binary least-squares formulation. The numerical tests on a comprehensive set of distribution circuits show that our proposed method yields very high accuracy on both radial and meshed distribution circuits with a combination of single-phase, two-phase, and three-phase loads. The proposed algorithm is robust with respect to inaccurate feeder models, incomplete measurements, and bad measurements. It also outperforms the existing methods on complex circuits.**

***Index Terms*—Distribution network, maximum marginal likelihood estimation, phase identification.**

NOMENCLATURE

I_n	$n \times n$ identity matrix.
$Im(\cdot)$	Imaginary part of a complex variable.
M	Number of loads in a circuit.
N	Number of three-phase non-substation nodes in the distribution network.
$Re(\cdot)$	Real part of a complex variable.
Y^{ij}	Bus admittance matrix between phase i and j.
$\text{diag}(\cdot)$	$\text{diag}(\boldsymbol{x})$ of a vector $\boldsymbol{x}$ is a diagonal matrix with $\boldsymbol{x}$ on the main diagonal. $\text{diag}(X_1, ..., X_n)$ is a block diagonal matrix with diagonal matrices of $X_1, ..., X_n$.
$\boldsymbol{v}$, $\boldsymbol{\theta}$, $\boldsymbol{p}$, $\boldsymbol{q}$	Vector of voltage magnitudes, voltage angles, real power injections, and reactive power injections of 3 phases of the nodes.
$\check{\boldsymbol{v}}$, $\check{\boldsymbol{\theta}}$, $\check{\boldsymbol{p}}$, $\check{\boldsymbol{q}}$	Non-substation nodes' voltage magnitude and angle difference with the substation, and their real and reactive power.
$\hat{\boldsymbol{v}}$, $\hat{\boldsymbol{p}}$, $\hat{\boldsymbol{q}}$	Vectors of load measurements of voltage magnitudes, real power, and reactive power injections.
$\tilde{\boldsymbol{v}}$, $\tilde{\boldsymbol{p}}$, $\tilde{\boldsymbol{q}}$	Time differenced load measurements of voltage magnitudes, real, and reactive power injections.
$\overline{\boldsymbol{v}}$, $\overline{\boldsymbol{\theta}}$	Flat voltage solution of three-phase power flow.
$\hat{\boldsymbol{v}}^{\text{ref}}$	Vector of reference voltage for loads.
x_m^i	Decision variable of load m's phase connection.
$\boldsymbol{x}$	Vector of decision variables x_m^i.
$\boldsymbol{x}^*$	True value of the decision variable vector $\boldsymbol{x}$.
α	Rotation operator, $\alpha = e^{-j\frac{2\pi}{3}}$.
$\mathbf{1}_n$	An all-1 vector of size n.
$(\cdot)^i$	A variable in phase i.
$(\cdot)^{ij}$	A variable between phase i and j.
$(\cdot)_n$	A variable at node or load n.
$(\cdot)_{-n}$	A variable excluding node or load n.
$\cdot(t)$	The value of a variable at time t.

Manuscript received August 22, 2019; revised January 2, 2020.

The authors are with the Department of Electrical and Computer Engineering, University of California Riverside, Riverside, CA 92521 USA (e-mail: wwang032@ucr.edu; nyu@ece.ucr.edu).

I. INTRODUCTION

With declining costs, distributed energy resources (DERs) such as energy storage systems, distributed generation, and electric vehicles are rapidly penetrating power distribution systems around the world. To coordinate the operations of a large number of heterogeneous DERs, advanced distribution system control applications such as Volt-VAR control, network reconfiguration, and three-phase optimal power flow need to be implemented. The successful implementation of these applications requires accurate information about the phase connectivity of power distribution systems. However, the phase connectivity information in electric utilities is usually missing or highly unreliable.

Traditionally, electric utilities send field crews to measure phase angles and determine phase connections with special equipment such as phase meters [1]. Although such practices provide very accurate phase connections information, they are very labor-intensive, time-consuming, and expensive. The time synchronized measurements from micro-phasor measurement units (μPMUs) can also provide highly accurate estimations of phase connections [2], [3]. However, a system-wide installation is cost prohibitive. State estimation can also be used to verify phase connection information [4]. However, this method only applies to circuits with mostly accurate phase connections and the area of incorrect phase connections needs to be known. In order to develop more cost effective phase identification algorithms, researchers have turned to data-driven methods, which use measurements from the advanced metering infrastructure (AMI). The existing data-driven approaches can be categorized into three approaches: energy supply and consumption matching, correlation-based analysis, and clustering-based analysis.

The energy supply and consumption matching approach is based on the principle of conservation of energy. With complete coverage of load measurements, the aggregate power consumption of downstream loads in each phase plus losses is equal to the corresponding phase's power flow measured at the upstream point. In this approach, Ref. [5] formulates the problem as integer programming and solves it using tabu search. Ref. [6] uses relaxed integer programming and improves the phase identification accuracy by actively managing the power injections of DERs. In [7], principal component analysis (PCA) and its graph-theoretic interpretation are used to infer phase connections. However, algorithms in this approach cannot identify phase connections in the presence of delta-connected two-phase loads.

In the correlation-based analysis approach, correlation analysis is performed using smart meters' and the substation's measurements or the three-phase primary line's measurements. Each smart meter is assigned to a phase, which has the highest correlation coefficient with it. In this approach, Ref. [8], [9] use voltage magnitude profiles for the correlation analysis. In [10], salient features are extracted from load profiles for the correlation analysis. Although the correlation-based analysis has achieved good performance on radial circuits with only single-phase loads, it does not work well for a meshed circuit, which has all seven possible phase connections of single-phase, two-phase, and three-phase loads.

In the clustering-based approach, smart meters are grouped based on the mutual similarity of their voltage magnitude profiles. It is assumed that each resulting cluster represents a single phase connection. Ref. [11], [12] project the voltage magnitude profiles onto low-dimension spaces and leverage constrained clustering algorithms to identify both single-phase and two-phase connections. Ref. [13] designs an algorithm by combining clustering and the minimum spanning tree method to identify phase connections. However, it has been shown that the performance of the clustering-based approach deteriorates as the feeder becomes more balanced [12].

To further improve the phase identification accuracy and provide a theoretical foundation for the problem, we develop a physically inspired machine learning method for phase identification. By linearizing the three-phase power flow manifold, we first develop a physical model, which links phase connections to the smart meter measurements. We then formulate the phase identification task as a maximum likelihood estimation (MLE) problem and prove that the correct phase connection yields the highest log likelihood value. The nonlinearity and nonconvexity nature of the MLE problem makes it difficult to solve. Thus, we reformulate the MLE problem as a maximum marginal likelihood estimation (MMLE) problem and prove that the correct phase connection also yields the highest marginal log likelihood value. Finally, an efficient solution algorithm is developed for the MMLE problem by dividing it into sub-problems, which can be solved by least squares integer programming.

Compared to the existing data-driven phase identification algorithms, our approach has the following advantages: first, the physically interpretable MMLE formulation brings a solid theoretical foundation to the phase identification problem; second, our proposed algorithm not only works for radial distribution feeders, but also heavily meshed networks; third, our proposed algorithm achieves higher accuracy for complex circuits with both single-phase and two-phase connections and a lower level of unbalance, which create a lot of problems to existing data-driven methods; fourth, our proposed algorithm is robust with respect to inaccurate feeder models, incomplete measurements, and bad measurements.

The rest of the paper is organized as follows. Section II covers the problem setup and the linearized three-phase power flow model. Section III derives the model that links the phase connections to the smart meter measurements. Section IV formulates the phase identification problem as an MLE and MMLE problem and presents an efficient solution algorithm. A comprehensive numerical test is performed in Section V to evaluate the performance of the proposed MMLE-based phase identification method. Section VI states the conclusion.

II. Problem Setup and Linearized Three-Phase Power Flow Model

A. Problem Setup

We intend to identify the type of phase connection for all loads on a distribution feeder. The distribution feeder's three-phase primary line contains $N+1$ nodes, indexed as node 0 to N, in which node 0 is the source/substation. A load can connect to a three-phase node directly, or indirectly through a single-phase or two-phase branch (e.g., the dashed lines and dash-dot lines in Fig. 1). Note that nodes and loads are two different concepts. In the technical derivation, all variables are in per unit or radian angles unless otherwise specified.

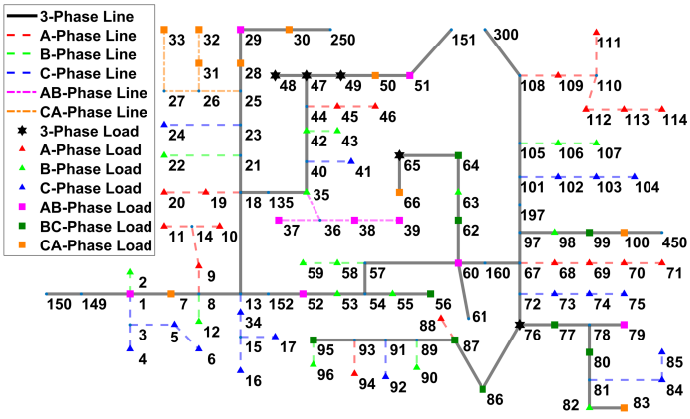


Fig. 1. Schematic of a modified IEEE 123-node test feeder.

B. Assumptions

Note that the assumptions described below are only used to prove that the correct phase connection yields the highest log likelihood value of the MLE and MMLE problem formulated in this paper. Some of these assumptions may not hold in the real world. However, the numerical study will show that our proposed algorithm still works well even when some of these assumptions no longer hold. In these cases, we can no longer guarantee that our proposed algorithm will result in 100% accurate phase identification results.

1) Data and Model Availability: First, the information about whether the load is single-phase, two-phase, or three-phase is assumed to be available. Usually, this information can be deduced by examining the distribution transformer configuration and customer billing information. Second, for a single-phase load on phase i, we know its power injection (both real and reactive power) and voltage magnitude of phase i. Third, for a two-phase delta-connected load between phase i and j, we know its power injection and voltage magnitude across phase i and j. Fourth, for a three-phase load, we know its total power injection and the voltage magnitude of one of the phases, which needs to be identified. Fifth, for the source node, we know the voltage measurement. Sixth, the connectivity model and the parameters of the primary feeder are known. Finally, we assume that the distribution feeder is not severely unbalanced. The task of phase identification is to determine which phase(s) each single-phase or two-phase load connects to and which phase's voltage magnitude the three-phase smart meter measures. Note that our proposed algorithm does not assume a 100% smart meter penetration rate. The numerical study will show that our algorithm is robust with respect to incomplete measurements.

2) Statistical Assumptions: First, it is assumed that the incremental changes in measured real, reactive power, and voltage magnitudes across one time interval are independent over time. Second, it is assumed that the noise terms which represent the model errors and the measurement errors are i.i.d. Gaussian. Note that the noise terms will be derived later in Section IV. Third, it is assumed that theses noise terms are independent of the incremental changes in smart meter measurements. Note that these statistical assumptions will be verified in the numerical study section.

C. The Linearized Power Flow Model for Primary Feeders

The very first step of our phase identification framework is to build a three-phase power flow model for the primary feeder. To do so, we need a procedure that we call *reduction*, and the resulting network is called a *reduced network*. The reduction is simply converting any loaded single-phase or two-phase branch into an equivalent load so that the reduced network contains only three-phase lines. The details of the reduction procedure is explained in Appendix A. In the rest of the paper, we use M to denote the number of loads in the reduced network and *load* refers to the equivalent load in the reduced network.

From the reduced primary feeder, by following [14], we can derive the linearized three-phase power flow model shown in (1), with the variables organized by phase. The linearized model ignores shunt admittance because it is very small. Numerical study results will verify that ignoring shunt admittance does not affect the phase identification accuracy.

$$A\begin{bmatrix}\boldsymbol{v}-\overline{\boldsymbol{v}}\\ \boldsymbol{\theta}-\overline{\boldsymbol{\theta}}\end{bmatrix}=\begin{bmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix}\begin{bmatrix}\boldsymbol{v}-\overline{\boldsymbol{v}}\\ \boldsymbol{\theta}-\overline{\boldsymbol{\theta}}\end{bmatrix}=\begin{bmatrix}\boldsymbol{p}\\ \boldsymbol{q}\end{bmatrix} \tag{1}$$

Here A_{11}, A_{12}, A_{21}, and A_{22} are $3(N+1)\times 3(N+1)$ matrices. $\boldsymbol{v}$, $\boldsymbol{\theta}$, $\boldsymbol{p}$, and $\boldsymbol{q}$ are the nodes' voltage magnitude, voltage angle, and real and reactive power of three phases. $\overline{\boldsymbol{v}}=\mathbf{1}_{3(N+1)}$ and $\overline{\boldsymbol{\theta}}=[0\times\mathbf{1}_{N+1}^T,-\frac{2\pi}{3}\times\mathbf{1}_{N+1}^T,\frac{2\pi}{3}\times\mathbf{1}_{N+1}^T]^T$ are the flat feasible solution for the underlying nonlinear power flow model. Let $\alpha=e^{-j\frac{2\pi}{3}}$, define $\Phi\triangleq\text{diag}(I_{(N+1)},\ \alpha I_{(N+1)},\ \alpha^2 I_{(N+1)})$ and define

$$Y\triangleq\begin{bmatrix}Y^{aa} & Y^{ab} & Y^{ac}\\ Y^{ba} & Y^{bb} & Y^{bc}\\ Y^{ca} & Y^{cb} & Y^{cc}\end{bmatrix} \tag{2}$$

where Y^{ij} is the $(N+1)\times(N+1)$ nodal admittance matrix between phase i and j. Then A_{11}, A_{12}, A_{21}, and A_{22} can be calculated as $A_{11}=-A_{22}=Re(\Phi^{-1}Y\Phi)$ and $A_{12}=A_{21}=-Im(\Phi^{-1}Y\Phi)$.

It has been shown in [15] that for a connected three-phase network, $\text{rank}(Y)=3N$. Thus, $\text{rank}(A)$ is at most $6N$. For subsequent derivations, we need to transform A into a nonsingular form. Following Appendix B, the transformed power flow model becomes

$$\check{A}\begin{bmatrix}\check{\boldsymbol{v}}\\ \check{\boldsymbol{\theta}}\end{bmatrix}=\begin{bmatrix}\check{A}_{11} & \check{A}_{12}\\ \check{A}_{21} & \check{A}_{22}\end{bmatrix}\begin{bmatrix}\check{\boldsymbol{v}}\\ \check{\boldsymbol{\theta}}\end{bmatrix}=\begin{bmatrix}\check{\boldsymbol{p}}\\ \check{\boldsymbol{q}}\end{bmatrix} \tag{3}$$

where $\check{A}_{mn}$ is a $3N\times 3N$ matrix obtained by removing the rows and columns corresponding to the substation node in A_{mn}. We denote the difference of voltage magnitudes and voltage angles between the non-substation nodes and the substation nodes as $\check{\boldsymbol{v}}$, $\check{\boldsymbol{\theta}}$. We denote the non-substation nodes' real and reactive power as $\check{\boldsymbol{p}}$ and $\check{\boldsymbol{q}}$.

In theory, $\check{A}$ is not guaranteed to be invertible. However, for the majority of real-world distribution feeders, $\text{rank}(\check{A})=6N$. It will be shown in the numerical study section that for all IEEE distribution test feeders, $\check{A}$ has a full rank.

Solving for $\check{v}$ with $\check{p}$ and $\check{q}$ from (3), we have

$$\begin{aligned}\check{\boldsymbol{v}}=&(\check{A}_{11}-\check{A}_{12}\check{A}_{22}^{-1}\check{A}_{21})^{-1}\check{\boldsymbol{p}}\\ &-(\check{A}_{11}-\check{A}_{12}\check{A}_{22}^{-1}\check{A}_{21})^{-1}\check{A}_{12}\check{A}_{22}^{-1}\check{\boldsymbol{q}}\end{aligned} \tag{4}$$

or in condensed form as

$$\check{\boldsymbol{v}}=K\check{\boldsymbol{p}}-L\check{\boldsymbol{q}} \tag{5}$$

It can be shown that $(\check{A}_{11}-\check{A}_{12}\check{A}_{22}^{-1}\check{A}_{21})$ is invertible if $\check{A}$ is invertible. Similarly, we can link $\check{\boldsymbol{\theta}}$ with $\check{\boldsymbol{p}}$ and $\check{\boldsymbol{q}}$ as

$$\begin{aligned}\check{\boldsymbol{\theta}}=&(\check{A}_{12}-\check{A}_{11}\check{A}_{21}^{-1}\check{A}_{22})^{-1}\check{\boldsymbol{p}}\\ &-(\check{A}_{12}-\check{A}_{11}\check{A}_{21}^{-1}\check{A}_{22})^{-1}\check{A}_{11}\check{A}_{21}^{-1}\check{\boldsymbol{q}}\end{aligned} \tag{6}$$

or in condensed form as

$$\check{\boldsymbol{\theta}}=\mathcal{K}\check{\boldsymbol{p}}-\mathcal{L}\check{\boldsymbol{q}} \tag{7}$$

III. Model for Phase Identification

In this section, we develop a mathematical model that relates the phase connections of loads to voltage magnitude and power injection measurements. Section III-A explains how to express smart meter measurements in terms of nodal voltages and power injections of the three-phase power flow model. Section III-B derives the phase connection model, which relates phase connections to network measurements.

A. Link Smart Meter Measurements with the Nodal Voltages and Power Injections

The linearized three-phase power flow models (5) and (7) are derived in terms of nodal voltages and power injections $\check{\boldsymbol{v}}$, $\check{\boldsymbol{\theta}}$, $\check{\boldsymbol{p}}$, and $\check{\boldsymbol{q}}$, which are often not directly measured by smart meters. Thus, we need to embed the smart meter measurements into these two equations. This is straightforward for single-phase and three-phase loads. For a single-phase load m on node n, its voltage measurement $\hat{v}_m$ is equal to one of the three phase-to-neutral voltage magnitudes v_n^i $(i = a, b, c)$, which is related to $\check{v}_n^i$ in (3) via $\check{v}_n^i \triangleq v_n^i - v_0^i$, where v_0^i is the source voltage magnitude in phase i. Similarly, a single-phase load's power injection measurement $\hat{p}_m + j\hat{q}_m$ corresponds to the power injection of one of the three phases $\check{p}_n^i + j\check{q}_n^i$ at node n. For a three-phase load m at node n, the single-phase voltage measurement $\hat{v}_m$ is equal to one of the three nodal voltage magnitudes v_n^i $(i = a, b, c)$. We can assume that the three-phase power injections $\hat{p}_m + j\hat{q}_m$ is distributed relatively evenly to three phases at node n. For a delta-connected two-phase load, we need the following derivations to link its measurements to the three-phase power flow model.

1) Link Power Injection Measurements with Power Flow Model: Without loss of generality, we use a phase AB load as an example. Suppose the two-phase power injection measurement is $S_{ab} = P_{ab} + jQ_{ab} = S_a + S_b = (P_a + jQ_a) + (P_b + jQ_b)$. Here, S_a and S_b are the power injections at the phase A and phase B ports. We can estimate S_a and S_b based on S_{ab} as follows: (see the proof in Appendix C)

$$S_a \approx \left(\frac{1}{2}P_{ab} + \frac{\sqrt{3}}{6}Q_{ab}\right) + j\left(\frac{1}{2}Q_{ab} - \frac{\sqrt{3}}{6}P_{ab}\right) \tag{8}$$

$$S_b \approx \left(\frac{1}{2}P_{ab} - \frac{\sqrt{3}}{6}Q_{ab}\right) + j\left(\frac{1}{2}Q_{ab} + \frac{\sqrt{3}}{6}P_{ab}\right) \tag{9}$$

2) Link Voltage Magnitude Measurements with Power Flow Model: Here we need to establish a relationship between the phase-to-phase voltage magnitude measurements and the nodal phase-to-neural voltage magnitudes in (5) and (7). For a load m across phase ij ($ij \in \{ab, bc, ca\}$) at node n, the relationship can be written as: (see the proof in Appendix D)

$$\begin{aligned}\hat{v}_m - v_0^{ij} \approx & \frac{\sqrt{3}}{2}(v_n^i - v_0^i) + \frac{\sqrt{3}}{2}(v_n^j - v_0^j) \\ & + \frac{1}{2}(\theta_n^i - \theta_0^i) - \frac{1}{2}(\theta_n^j - \theta_0^j)\end{aligned} \tag{10}$$

where $\hat{v}_m$ is load m's voltage magnitude measurement. v_0^{ij} is the voltage magnitude across phase ij at the substation. v_n^i and v_0^i are the voltage magnitudes of phase i at node n and the substation. θ_n^i and θ_0^i are the voltage angles of phase i at node n and the substation. Note that in above derivations, voltages are in per unit and angles are in radian.

B. Modeling Phase Connections in Three-phase Power Flow

1) Decision Variables for Phase Connections: We use three decision variables, x_m^1, x_m^2, and x_m^3 to denote the phase connection for each load m. $x_m^i = 0$ or 1, and $\sum_i x_m^i = 1$, $\forall\, m$. If load m is single-phase, then x_m^1, x_m^2, and x_m^3 represent AN, BN, and CN connections. If m is two-phase, then x_m^1, x_m^2, and x_m^3 represent AB, BC, and CA connections. If m is three-phase, and the measured voltage is between one phase and the neutral, then x_m^1, x_m^2, and x_m^3 represent which of the phases AN, BN, and CN is measured. As stated in the assumptions, we know whether a load is single-phase, two-phase, or three-phase from the distribution transformer configuration and customer billing information. The phase connection decision variables form an $M \times 3M$ matrix X defined as $X \triangleq \text{diag}([x_1^1\ x_1^2\ x_1^3], ..., [x_M^1\ x_M^2\ x_M^3])$.

2) Additional Definitions: Several matrices and variables are defined here to build the model for phase connections.

Define matrices W_1 and W_2 as

$$W_1 \triangleq \begin{bmatrix}1 & 1 & 0\\ 0 & 1 & 1\\ 1 & 0 & 1\end{bmatrix},\ W_2 \triangleq \begin{bmatrix}1 & -1 & 0\\ 0 & 1 & -1\\ -1 & 0 & 1\end{bmatrix} \tag{11}$$

Let I_n denote an identity matrix of size n, $\mathbb{0}_{k\times l}$ denote a $k \times l$ all-0 matrix, and $\mathbb{1}_{k\times l}$ denote a $k \times l$ all-1 matrix. Define U^1 and U^2 as $3M \times 3N$ matrices of 3×3 blocks. Define $\hat{U}^1$ and $\hat{U}^2$ as $3N \times 3M$ matrices of 3×3 blocks. Define U_{mn}^1 and U_{mn}^2 as the mn-th block of U^1 and U^2. Define $\hat{U}_{nm}^1$ and $\hat{U}_{nm}^2$ as the nm-th block of $\hat{U}^1$ and $\hat{U}^2$. If load m is not connected to node n, then U_{mn}^1, U_{mn}^2, $\hat{U}_{nm}^1$, and $\hat{U}_{nm}^2$ are equal to $\mathbb{0}_{3\times 3}$. If load m is connected to node n, then U_{mn}^1, U_{mn}^2, $\hat{U}_{nm}^1$, and $\hat{U}_{nm}^2$ are defined based on load m's phase connection type, as shown in Table I.

TABLE I
VALUES OF 3×3 BLOCKS BY PHASE CONNECTION TYPE IF LOAD m IS CONNECTED TO NODE n

Load m's Phase Connection Type	U_{mn}^1	U_{mn}^2	$\hat{U}_{nm}^1$	$\hat{U}_{nm}^2$
single-phase	I_3	$\mathbb{0}_{3\times 3}$	I_3	$\mathbb{0}_{3\times 3}$
two-phase	$\frac{\sqrt{3}}{2}W_1$	$\frac{1}{2}W_2$	$\frac{1}{2}W_1^T$	$\frac{\sqrt{3}}{6}W_2^T$
three-phase	I_3	$\mathbb{0}_{3\times 3}$	$\frac{1}{3}\mathbb{1}_{3\times 3}$	$\mathbb{0}_{3\times 3}$

Define $\hat{\boldsymbol{v}}^{\text{ref}} \triangleq [\hat{\boldsymbol{v}}_1^{\text{ref}}, \ldots, \hat{\boldsymbol{v}}_M^{\text{ref}}]^T$, where $\hat{\boldsymbol{v}}_m^{\text{ref}} = [v_0^a, v_0^b, v_0^c]$ if load m is single-phase or three-phase; $\hat{\boldsymbol{v}}_m^{\text{ref}} = [v_0^{ab}, v_0^{bc}, v_0^{ca}]$ if load m is two-phase. Here, v_0^i denotes the substation's voltage magnitude of phase i, and v_0^{ij} denotes the substation's voltage magnitude across phase ij.

3) Phase Connection Model: Now we can build the model, which links phase connections with the smart meter measurements. Let $\hat{\boldsymbol{v}}$, $\hat{\boldsymbol{p}}$, and $\hat{\boldsymbol{q}}$ be $M \times 1$ vectors of measured voltage magnitudes, real power, and reactive power of the M loads. From (8) - (10), Section III-B1, and III-B2, we have:

$$\check{\boldsymbol{p}} \approx \hat{U}^1 X^T \hat{\boldsymbol{p}} + \hat{U}^2 X^T \hat{\boldsymbol{q}} \tag{12}$$

$$\check{\boldsymbol{q}} \approx -\hat{U}^2 X^T \hat{\boldsymbol{p}} + \hat{U}^1 X^T \hat{\boldsymbol{q}} \tag{13}$$

$$\hat{\boldsymbol{v}} \approx X\hat{\boldsymbol{v}}^{\text{ref}} + XU^1\check{\boldsymbol{v}} + XU^2\check{\boldsymbol{\theta}} \tag{14}$$

With a slight abuse of notations, the entries of $\check{\boldsymbol{p}}$, $\check{\boldsymbol{q}}$, $\check{\boldsymbol{v}}$, and $\check{\boldsymbol{\theta}}$ are organized by node in (12)-(14) (instead of by phase as in (5) and (7)). Equations (12) and (13) map the measured power injection of each load to the corresponding nodal power

injections in the linearized power flow model. Take load m connected to node n as an example and suppose $x_m^1 = 1$. If load m is single-phase, then its power injection is mapped to phase A at node n. If load m is two-phase, then its power injection is distributed to phase A and B at node n according to (8) and (9). If load m is three-phase, then its power injection is evenly distributed to all three phases of node n.

Equation (14) links the voltage measurement $\hat{\boldsymbol{v}}$ with $\check{\boldsymbol{v}}$ and $\check{\boldsymbol{\theta}}$, i.e., the nodal line-to-neutral voltage magnitude and angle difference with the substation in the linearized power flow model. Take load m connected to node n as an example and suppose $x_m^1 = 1$. If load m is single-phase or three-phase, then (14) can be reduced to $\hat{v}_m = v_0^a + (v_n^a - v_0^a)$, where v_n^a is node n's voltage magnitude in phase A. If load m is two-phase, then (14) is equivalent to (10).

Substituting (5), (7), (12) and (13) into (14) yields

$$\hat{\boldsymbol{v}} \approx X\hat{\boldsymbol{v}}^{\text{ref}} + X\hat{K}X^T\hat{\boldsymbol{p}} + X\hat{L}X^T\hat{\boldsymbol{q}} \tag{15}$$

where $\hat{K} \triangleq [(U^1K + U^2\mathcal{K})\hat{U}^1 + (U^1L + U^2\mathcal{L})\hat{U}^2]$ and $\hat{L} \triangleq [(U^1K+U^2\mathcal{K})\hat{U}^2-(U^1L+U^2\mathcal{L})\hat{U}^1]$. Here, with a slight abuse of notations, K, L, $\mathcal{K}$, and $\mathcal{L}$'s entries are organized by node (instead of by phase as in (5) and (7)). Thus, (15) provides the physical model, which relates power injection measurements and phase connections to voltage magnitude measurements.

To remove trends and seasonality in time series data, we define the difference of the voltage measurement and its lagged variable as $\tilde{\boldsymbol{v}}(t)$, with $\tilde{\boldsymbol{v}}(t) \triangleq \hat{\boldsymbol{v}}(t) - \hat{\boldsymbol{v}}(t-1)$. $\tilde{\boldsymbol{v}}^{\text{ref}}(t)$, $\tilde{\boldsymbol{p}}(t)$, and $\tilde{\boldsymbol{q}}(t)$ are defined in a similar way. Thus, we have the time difference version of the physical model:

$$\tilde{\boldsymbol{v}}(t) = X\tilde{\boldsymbol{v}}^{\text{ref}}(t) + X\hat{K}X^T\tilde{\boldsymbol{p}}(t) + X\hat{L}X^T\tilde{\boldsymbol{q}}(t) + \boldsymbol{n}(t) \tag{16}$$

where $\boldsymbol{n}(t)$ is the "noise term" representing the error of the linearized power flow model, the measurement error, and all the other sources of noise not considered. In (16), $\tilde{\boldsymbol{v}}(t)$, $\tilde{\boldsymbol{p}}(t)$, $\tilde{\boldsymbol{q}}(t)$, and $\tilde{\boldsymbol{v}}^{\text{ref}}(t)$ can be calculated from the smart meter and substation measurements. $\hat{K}$ and $\hat{L}$ can be derived from the feeder model. Thus, the task of phase identification is to estimate the phase decision variables in X.

IV. Maximum Marginal Likelihood Estimation of Phase Connections

In this section, we first formulate phase identification as an MLE problem and then as an MMLE problem. Next, we prove that the correct phase connection is a global optimizer of the MMLE problem. Lastly, we develop a computationally efficient algorithm to solve the MMLE problem.

A. MLE Problem Formulation

Let $\boldsymbol{x} \triangleq [x_1^1, x_1^2, x_1^3, ..., x_M^1, x_M^2, x_M^3]^T$ be the phase connection decision variable vector. Define $\tilde{\boldsymbol{v}}(t, \boldsymbol{x})$ as the theoretical differenced voltage measurement $\tilde{\boldsymbol{v}}(t)$ with phase connection $\boldsymbol{x}$:

$$\tilde{\boldsymbol{v}}(t, \boldsymbol{x}) \triangleq X\tilde{\boldsymbol{v}}^{\text{ref}}(t) + X\hat{K}X^T\tilde{\boldsymbol{p}}(t) + X\hat{L}X^T\tilde{\boldsymbol{q}}(t) \tag{17}$$

Then $\tilde{\boldsymbol{v}}(t) = \tilde{\boldsymbol{v}}(t, \boldsymbol{x}) + \boldsymbol{n}(t)$, where $\boldsymbol{x}$ is the phase connection decision variable vector that we need to estimate.

As stated in Section II-B, we assume that the noise $\boldsymbol{n}(t)$ is independent of $\tilde{\boldsymbol{v}}^{\text{ref}}(t)$, $\tilde{\boldsymbol{p}}(t)$, and $\tilde{\boldsymbol{q}}(t)$ and is i.i.d. Gaussian $\boldsymbol{n}(t) \sim \mathcal{N}(\mathbb{0}_{M\times 1}, \Sigma_n)$, where Σ_n is an unknown underlying covariance matrix. Given these conditions, $\boldsymbol{n}(t)$ is also independent of $\tilde{\boldsymbol{v}}(t, \boldsymbol{x})$. Thus, the likelihood of observing $\{\tilde{\boldsymbol{v}}(t)\}_{t=1}^T$ given $\{\tilde{\boldsymbol{v}}^{\text{ref}}(t)\}_{t=1}^T$, $\{\tilde{\boldsymbol{p}}(t)\}_{t=1}^T$, and $\{\tilde{\boldsymbol{q}}(t)\}_{t=1}^T$ is a function of $\boldsymbol{x}$:

$$Prob(\{\tilde{\boldsymbol{v}}(t)\}_{t=1}^T | \{\tilde{\boldsymbol{v}}^{\text{ref}}(t)\}_{t=1}^T, \{\tilde{\boldsymbol{p}}(t)\}_{t=1}^T, \{\tilde{\boldsymbol{q}}(t)\}_{t=1}^T; \boldsymbol{x}) = \frac{|\Sigma_n|^{-\frac{T}{2}}}{(2\pi)^{\frac{MT}{2}}} \times \exp\Big\{-\frac{1}{2}\sum_{t=1}^T [\tilde{\boldsymbol{v}}(t) - \tilde{\boldsymbol{v}}(t, \boldsymbol{x})]^T \Sigma_n^{-1} [\tilde{\boldsymbol{v}}(t) - \tilde{\boldsymbol{v}}(t, \boldsymbol{x})]\Big\} \tag{18}$$

Taking the negative logarithm of (18), removing the constant term, and scaling by $\frac{2}{T}$, we get

$$f(\boldsymbol{x}) \triangleq \frac{1}{T}\sum_{t=1}^T [\tilde{\boldsymbol{v}}(t) - \tilde{\boldsymbol{v}}(t, \boldsymbol{x})]^T \Sigma_n^{-1} [\tilde{\boldsymbol{v}}(t) - \tilde{\boldsymbol{v}}(t, \boldsymbol{x})] \tag{19}$$

It will be shown in Lemma 1 that the correct phase connection $\boldsymbol{x}^*$ maximizes the likelihood function (18) and minimizes $f(x)$ under two mild assumptions.

Lemma 1. *Let $\boldsymbol{x}^*$ be the correct phase connection. If the following two conditions are satisfied, then as $T \to \infty$, $\boldsymbol{x}^*$ is a global optimizer to minimize $f(\boldsymbol{x})$.*

1) *$\boldsymbol{n}(t_k)$ is i.i.d. and independent of $\tilde{\boldsymbol{v}}^{ref}(t_l)$, $\tilde{\boldsymbol{p}}(t_l)$, and $\tilde{\boldsymbol{q}}(t_l)$, for $\forall t_k, t_l \in Z^+$.*
2) *$\tilde{\boldsymbol{v}}^{ref}(t_k)$, $\tilde{\boldsymbol{p}}(t_k)$, and $\tilde{\boldsymbol{q}}(t_k)$ are independent of $\tilde{\boldsymbol{v}}^{ref}(t_l)$, $\tilde{\boldsymbol{p}}(t_l)$, and $\tilde{\boldsymbol{q}}(t_l)$, for $\forall t_k, t_l \in Z^+,\ t_k \neq t_l$*

The proof of Lemma 1 can be found in Appendix E. By substituting (17) into (19), we can see that directly minimizing $f(\boldsymbol{x})$ is very difficult due to its nonlinearity and nonconvexity. Furthermore, the actual value of Σ_n is unknown. To address this technical challenge, in Section IV-B, we will convert the phase identification problem into an MMLE problem and prove that the correct phase connection is also a global optimizer of the MMLE problem.

B. MMLE Problem Formulation

Let $\tilde{v}_m(t)$ be the mth entry of $\tilde{\boldsymbol{v}}(t)$, $\tilde{v}_m(t, \boldsymbol{x})$ be the mth entry of $\tilde{\boldsymbol{v}}(t, \boldsymbol{x})$, and $n_m(t)$ be the mth entry of $\boldsymbol{n}(t)$. The marginal likelihood of observing $\{\tilde{v}_m(t)\}_{t=1}^T$ given $\{\tilde{\boldsymbol{v}}^{\text{ref}}(t)\}_{t=1}^T$, $\{\tilde{\boldsymbol{p}}(t)\}_{t=1}^T$, and $\{\tilde{\boldsymbol{q}}(t)\}_{t=1}^T$ is a function of $\boldsymbol{x}$:

$$Prob(\{\tilde{v}_m(t)\}_{t=1}^T | \{\tilde{\boldsymbol{v}}^{\text{ref}}(t)\}_{t=1}^T, \{\tilde{\boldsymbol{p}}(t)\}_{t=1}^T, \{\tilde{\boldsymbol{q}}(t)\}_{t=1}^T; \boldsymbol{x}) = \frac{\Sigma_n(m,m)^{-\frac{T}{2}}}{(2\pi)^{\frac{T}{2}}} \exp\Big\{-\frac{1}{2}\sum_{t=1}^T \frac{[\tilde{v}_m(t) - \tilde{v}_m(t, \boldsymbol{x})]^2}{\Sigma_n(m,m)}\Big\} \tag{20}$$

where $\Sigma_n(m,m)$ is the mth diagonal entry of Σ_n. Taking the negative logarithm of (20), removing the constant term, and scaling by $\frac{2\Sigma_n(m,m)}{T}$, we have

$$f_m(\boldsymbol{x}) \triangleq \frac{1}{T}\sum_{t=1}^T [\tilde{v}_m(t) - \tilde{v}_m(t, \boldsymbol{x})]^2 \tag{21}$$

Lemma 2. *Let $\boldsymbol{x}^*$ be the correct phase connection. If the two conditions in Lemma 1 hold, then $\boldsymbol{x}^*$ is a global optimizer to*

minimize $f_m(\boldsymbol{x})$ as $T \to \infty$. In addition, any $\boldsymbol{x}$ is a global optimizer of $f_m(\boldsymbol{x})$ if it satisfies all the following conditions:

1) $x_m^i = x^{*i}{}_m, \forall i$;

2) $x_k^i = x^{*i}{}_k, \forall i$, $k \neq m$ and load k is not three-phase.

The proof of Lemma 2 can be found in Appendix F.

C. Solution Method for the MMLE Problem

Directly minimizing $f_m(\boldsymbol{x})$ from (21) is still a difficult task. Thus, we further simplify the optimization problem by first solving three subproblems $min f_{m,i}(\boldsymbol{x}_{-m}), i \in \{1,2,3\}$. $f_{m,i}(\boldsymbol{x}_{-m})$ are defined as

$$\begin{aligned} &f_{m,i}(\boldsymbol{x}_{-m}) \triangleq f_m(\boldsymbol{x}) \\ &\text{subject to } x_m^i = 1 \text{ and } x_m^j = 0 \text{ for } j \neq i \end{aligned} \tag{22}$$

where $\boldsymbol{x}_{-m}$ is a $(3M-3)\times 1$ vector containing every element in $\boldsymbol{x}$ except x_m^1, x_m^2, and x_m^3. Since $x_m^i = 0$ or 1, and $\sum_i x_m^i = 1$, then from (22) we have:

$$\min_{\boldsymbol{x}} f_m(\boldsymbol{x}) = \min_{i=1,2,3} \min_{\boldsymbol{x}_{-m}} f_{m,i}(\boldsymbol{x}_{-m}) \tag{23}$$

To solve the sub-problems, we first define $\tilde{v}_{m,i}(t, \boldsymbol{x}_{-m})$ as

$$\begin{aligned} &\tilde{v}_{m,i}(t, \boldsymbol{x}_{-m}) \triangleq \tilde{v}_m(t, \boldsymbol{x}) \\ &\text{subject to } x_m^i = 1 \text{ and } x_m^j = 0 \text{ for } j \neq i \end{aligned} \tag{24}$$

Substituting (17) into (24), we have

$$\begin{aligned} \tilde{v}_{m,i}(t, \boldsymbol{x}_{-m}) =& \tilde{v}^{\text{ref}}_{m,i}(t) + \hat{K}_{m,i} X^T \tilde{\boldsymbol{p}}(t) + \hat{L}_{m,i} X^T \tilde{\boldsymbol{q}}(t) \\ &\text{subject to } x_m^i = 1 \text{ and } x_m^j = 0 \text{ for } j \neq i \end{aligned} \tag{25}$$

where $\tilde{v}^{\text{ref}}_{m,i}(t)$ is the entry of $\tilde{\boldsymbol{v}}^{\text{ref}}(t)$ corresponding to x_m^i, $\hat{K}_{m,i}$ and $\hat{L}_{m,i}$ are the row vectors of $\hat{K}$ and $\hat{L}$ corresponding to x_m^i.

Define an $M \times 3M$ matrix $\mathfrak{D}$ as:

$$\mathfrak{D} \triangleq \text{diag}(\underbrace{[1\ 1\ 1], \ldots, [1\ 1\ 1]}_{\text{repeat } M \text{times}}) \tag{26}$$

Then matrix X can be expressed by decision vector $\boldsymbol{x}$ as $X = \mathfrak{D}\ \text{diag}(\boldsymbol{x})$. Thus, we can simplify the second term on the right-hand-side (RHS) of (25) as

$$\begin{aligned} &\hat{K}_{m,i} X^T \tilde{\boldsymbol{p}}(t) = \hat{K}_{m,i}\ \text{diag}(\boldsymbol{x})\ \mathfrak{D}^T\ \tilde{\boldsymbol{p}}(t) \\ =&\boldsymbol{x}^T\ \text{diag}(\hat{K}_{m,i})\ \mathfrak{D}^T\ \tilde{\boldsymbol{p}}(t) = \boldsymbol{x}^T\ \boldsymbol{\zeta}_{m,i}(t) = \boldsymbol{\zeta}^T_{m,i}(t)\ \boldsymbol{x} \end{aligned} \tag{27}$$

where $\boldsymbol{\zeta}_{m,i}(t) \triangleq \text{diag}(\hat{K}_{m,i})\ \mathfrak{D}^T\ \tilde{\boldsymbol{p}}(t)$. Similarly, simplify the third term on the RHS of (25) as

$$\hat{L}_{m,i} X^T \tilde{\boldsymbol{q}}(t) = \boldsymbol{\xi}^T_{m,i}(t)\ \boldsymbol{x} \tag{28}$$

where $\boldsymbol{\xi}_{m,i}(t) \triangleq \text{diag}(\hat{L}_{m,i})\ \mathfrak{D}^T\ \tilde{\boldsymbol{q}}(t)$.

Substituting (27) and (28) into equation (25), we have

$$\begin{aligned} &\tilde{v}_m(t) - \tilde{v}_{m,i}(t, \boldsymbol{x}_{-m}) \\ =&\tilde{v}_m(t) - \tilde{v}^{\text{ref}}_{m,i}(t) - \boldsymbol{\zeta}^T_{m,i}(t)\boldsymbol{x} - \boldsymbol{\xi}^T_{m,i}(t)\boldsymbol{x} \\ =&\tilde{v}_m(t) - \tilde{v}^{\text{ref}}_{m,i}(t) - \boldsymbol{\psi}^T_{m,i}(t)\boldsymbol{x} \\ =&\tilde{v}_m(t) - \tilde{v}^{\text{ref}}_{m,i}(t) - [\boldsymbol{\varphi}^T_{m,i}(t)\boldsymbol{x}_{-m} + \eta_{m,i}(t)] \\ =&v^{\text{tot}}_{m,i}(t) - \boldsymbol{\varphi}^T_{m,i}(t)\boldsymbol{x}_{-m} \end{aligned} \tag{29}$$

Where $\boldsymbol{\psi}_{m,i}(t) \triangleq \boldsymbol{\zeta}_{m,i}(t) + \boldsymbol{\xi}_{m,i}(t)$. $\boldsymbol{\varphi}_{m,i}(t)$ is a vector containing all the elements in $\boldsymbol{\psi}_{m,i}(t)$ except the three elements corresponding to x_m^1, x_m^2, and x_m^3. $\eta_{m,i}(t)$ is the element in $\boldsymbol{\psi}_{m,i}(t)$ corresponding to x_m^i. In the last line of (29), $v^{\text{tot}}_{m,i}(t)$ is defined as $v^{\text{tot}}_{m,i}(t) \triangleq \tilde{v}_m(t) - \tilde{v}^{\text{ref}}_{m,i}(t) - \eta_{m,i}(t)$.

Note that our proposed phase identification method still works even if there is a topology change in the primary feeder. If such topology change occurs at time t_c, then we can simply update $v^{\text{tot}}_{m,i}(t)$ and $\boldsymbol{\varphi}_{m,i}(t)$ in (29) according to the new primary feeder topology.

With (29), the function $f_{m,i}(\boldsymbol{x}_{-m})$ can be transformed into

$$f_{m,i}(\boldsymbol{x}_{-m}) = \frac{1}{T}\sum_{t=1}^{T} [v^{\text{tot}}_{m,i}(t) - \boldsymbol{\varphi}^T_{m,i}(t)\boldsymbol{x}_{-m}]^2 \tag{30}$$

Now each MMLE sub-problem in (23) can be formulated as

$$\begin{aligned} \text{find} \quad & \boldsymbol{x}^{\dagger}_{-m,i} = \arg\min_{\boldsymbol{x}_{-m}}\ f_{m,i}(\boldsymbol{x}_{-m}) \\ \text{subject to} \quad & x_k^j = 0 \text{ or } 1 \quad \forall j \text{ and } k \neq m \\ & \sum_j x_k^j = 1 \quad \forall k \neq m. \end{aligned} \tag{31}$$

This is a binary least-square problem. To solve it efficiently, we can further relax the problem by replacing the binary constraint by its convex hull. Now the problem is equivalent to convex quadratic programming, which can be solved in polynomial time [16]. The continuous solution of $\boldsymbol{x}_{-m}$ in the convex hull can then be rounded to binary values as follows: for each load $k \neq m$, round x_k^j to 1 if it is the largest among x_k^1, x_k^2, and x_k^3, and round the other two variables to 0.

D. Phase Identification Algorithm

Our proposed MMLE-based phase identification algorithm is summarized in Algorithm 1 and explained as follows. From step 1 to 6, we solve M MMLE problems, each of which contains three binary least-square sub-problems. Step 3 solves the sub-problems of MMLE based on (31). Based on (23), step 5 solves the mth MMLE problem by finding which of the three $\boldsymbol{x}^{\dagger}_{-m,i}$ $(i = 1,2,3)$ minimizes $f_m(\boldsymbol{x})$. The chosen $\boldsymbol{x}^{\dagger}_{-m,i}$, combined with the corresponding $x_m^i = 1$ and $x_m^j = 0$ $(j \neq i)$, forms the $3M \times 1$ solution $\boldsymbol{x}^{\dagger}{}_m$ of the mth MMLE problem. The M sets of $\boldsymbol{x}^{\dagger}{}_m$ may not be all correct due to the limited number of measurements and measurement noise. Thus, in step 7, we design two approaches to integrate M sets of $\boldsymbol{x}^{\dagger}{}_m$ into two phase identification solutions:

1) *Target-only Approach*. The phase connection of each load m is the corresponding connection shown in the mth solution $\boldsymbol{x}^{\dagger}{}_m$.
2) *Voting Approach*. For a single-phase or two-phase load m, the phase connection is the corresponding phase connection that receives the most votes in the M sets of $\boldsymbol{x}^{\dagger}{}_m$. For a three-phase load m, the phase connection is still determined by the target-only approach.

In step 8, we calculate $\sum_{m=1}^{M} f_m(\boldsymbol{x})$ based on the phase identification solution of both the target-only and the voting approaches. The final phase identification solution is the one that has the lower sum of square error.

Algorithm 1 Phase Identification Algorithm

Input: $\tilde{\boldsymbol{v}}(t)$, $\tilde{\boldsymbol{v}}^{\text{ref}}(t)$, $\tilde{\boldsymbol{p}}(t)$, $\tilde{\boldsymbol{q}}(t)$, $\hat{K}$, and $\hat{L}$, $t = 1, ..., T$.
Output: Estimated phase connections for the M loads.

1: **for** $m = 1$ to M **do**
2: **for** $i = 1$ to 3 **do**
3: Use the input to calculate $v^{\text{tot}}_{m,i}(t)$ and $\boldsymbol{\varphi}^T_{m,i}(t)$ and find the solution $\boldsymbol{x}^{\dagger}_{-m,i}$ to the sub-problem in (31).
4: **end for**
5: Use $\boldsymbol{x}^{\dagger}_{-m,i}$, $i \in \{1, 2, 3\}$ to find the $\boldsymbol{x}$ that minimizes $f_m(\boldsymbol{x})$ in (21). Record the solution as ${\boldsymbol{x}^{\dagger}}_m$.
6: **end for**
7: Generate two phase identification results based on M sets of ${\boldsymbol{x}^{\dagger}}_m$ using two approaches: the target-only approach and the voting approach.
8: Calculate $\sum_{m=1}^{M} f_m(\boldsymbol{x})$ based on both the target-only and the voting approach. Select the solution with the lower sum of square error.

V. Numerical Study

A. Setup for Numerical Tests

The performance of our proposed MMLE-based algorithm is evaluated using the IEEE 37-bus, 123-bus, and 342-bus test circuits. Fig. 1 illustrates the schematic of the 123-bus circuit. The results show that the proposed algorithm works well for distribution networks with either tree structured feeders (37-bus and 123-bus) or heavily meshed primary feeders (342-bus). The 342-bus feeder represents meshed distribution systems that are often used in urban centers, which have a high load density and require very high reliability. In North America, about 80 cities currently operate such distribution systems [17]. To make the task more difficult, we modify the test feeders to include all possible phase connection types (single-phase, two-phase, and three-phase). The number of loads by phase connection type is summarized in Table II.

TABLE II
Number of Loads Per Phase in the IEEE Test Circuits and Level of Unbalance

Feeder	Phase Connection								Level of Unbalance
	A	B	C	AB	BC	CA	ABC	Total	
37-bus	5	5	6	3	2	2	2	25	0.027
123-bus	18	17	17	9	9	10	5	85	0.0164
342-bus	30	38	31	35	31	33	10	208	0.0097

The hourly average real power consumption measurements from smart meters of a distribution feeder managed by FortisBC are used in test feeders. The length of the real power consumption time series is 2160, which represents 90 days of hourly smart meter measurements. The reactive power time series are generated by randomly sampling power factors from a uniform distribution $\mathcal{U}(0.9, 1)$ to represent lagging loads. The peak loads for the three IEEE test circuits are 2.4 MW, 4 MW, and 43 MW. The power flows of the test circuits are simulated using OpenDSS. All smart meter measurements contain noise that follows zero-mean Gaussian distributions with three-sigma deviation matching 0.1% to 0.2% of the nominal values. The 0.1 and 0.2 accuracy class smart meters established in ANSI C12.20-2015 are typical in real-world implementations. To make the phase identification task even more challenging, we assume that older generations of smart meters are adopted. That is to say, after adding measurement noise, the voltage measurements are rounded to the nearest 1 V for primary line loads and 0.1 V for secondary loads. The real and reactive power measurements are rounded to the nearest 0.1 kW or 0.1 kVAr. The relaxed optimization problems in equation (31) are solved using CPLEX on a DELL workstation with 3.3 GHz Intel Xeon CPU and 16 GB of RAM.

In the simulation, the power consumption time series are allocated relatively evenly to each phase so that the test feeders are close to balance. Following [12], the level of unbalance of a feeder at time interval t can be measured as

$$u(t) = \frac{|I_A(t) - I_m(t)| + |I_B(t) - I_m(t)| + |I_C(t) - I_m(t)|}{3I_m(t)} \tag{32}$$

where $I_m(t) = \frac{1}{3}(I_A(t) + I_B(t) + I_C(t))$ is the mean of the distribution substation line current magnitudes of the three phases at time interval t. $u(t)$ represents the power deviation of each phase from the average value at time interval t. We use the 90-day average value of $u(t)$ in this simulation to measure the level of unbalance of a feeder. The level of unbalance of the three feeders are shown in Table II.

Before presenting the main numerical results, we first verify the Gaussianity assumption for the noise term $\boldsymbol{n}(t)$ in equation (16). The Kolmogorov-Smirnov test is used to verify the Gaussianity assumption. With a significance level of 5%, the noise terms for all loads pass the test except 9 loads at 0.1% meter accuracy level and 1 load at 0.2% meter accuracy level in the 342-bus circuit. By checking the normalized auto-correlations of $n(t)$, we found the noise to be uncorrelated over time. For Gaussian random variables, this indicates independence over time.

B. Performance of the Proposed Phase Identification Method

The phase identification accuracy of our proposed MMLE-based algorithm is shown in Table III, which covers three IEEE test feeders, two meter accuracy classes (0.1% and 0.2%), and three time windows (30 days, 60 days, 90 days). With 90 days of hourly meter measurements and both accuracy class meters, the proposed algorithm achieved 100% accuracy for all three IEEE distribution test circuits. The proposed algorithm works well not only for radial feeders (37-bus, 123-bus), but also the meshed circuit (342-bus). As shown in the table, the accuracy of the MMLE-based phase identification algorithm increases as the smart meter measurement error decreases. When additional smart meter data becomes available, the phase identification accuracy of the proposed algorithm also increases as expected. The average computation time of the algorithm with 90 days of data is only around 1.3 seconds, 6.5 seconds, and 256 seconds for the three circuits, respectively.

We also test our proposed method with higher unbalance levels by adjusting the load levels in each phase of the test feeders. The result shows that the method is very accurate even if the feeder is severely unbalanced. Table IV shows the phase identification accuracy using 0.1% meter class in three

TABLE III
ACCURACY OF THE PROPOSED PHASE IDENTIFICATION METHOD WHEN FEEDERS ARE CLOSE TO BALANCE

Feeder	Meter Class	30 Days	60 Days	90 Days
37-bus	0.1%	100%	100%	100%
	0.2%	92%	100%	100%
123-bus	0.1%	96.47%	100%	100%
	0.2%	63.53%	96.47%	100%
342-bus	0.1%	96.63%	100%	100%
	0.2%	72.60%	99.52%	100%

unbalance levels: the original close-to-balance level, 0.05, and 0.1. The feeders are considered to be moderately unbalanced when the unbalance level is 0.05. The feeders are deemed as severely unbalanced, when the unbalance level is 0.1. As shown in Table IV, the phase identification accurate of our proposed algorithm gradually decreases as the unbalance level increases for the 123-bus feeder. This is because our proposed method is derived from an approximate model of distribution feeders that are close to balance. Thus, the approximation error will increase when the feeder becomes more unbalanced. However, at 0.1 unbalance level, our proposed method still attains high accuracy with sufficient smart meter data.

TABLE IV
ACCURACY OF THE PROPOSED PHASE IDENTIFICATION METHOD WITH DIFFERENT UNBALANCE LEVELS (METER ACCURACY CLASS 0.1%)

Feeder	Level of Unbalance	30 Days	60 Days	90 Days
37-bus	0.027	100%	100%	100%
	0.05	100%	100%	100%
	0.1	100%	100%	100%
123-bus	0.0164	96.47%	100%	100%
	0.05	95.29%	100%	100%
	0.1	85.88%	100%	100%
342-bus	0.0097	96.63%	100%	100%
	0.05	96.63%	99.52%	100%
	0.1	96.63%	99.52%	100%

We verify the robustness of our proposed phase identification algorithm against bad data. We assume that a meter with bad data has erroneous voltage and power measurements in 10% of the hours. The erroneous voltage measurements are assumed to have a uniform distribution within $\pm 20\%$ of the true values. The erroneous real and reactive power measurements are assumed to follow uniform distributions within $\pm 100\%$ of the true values. We test our method when 1%, 5%, and 10% of the smart meters have bad data on the 123-bus and 243-bus feeders. The 37-bus feeder only has 25 smart meters. Thus, we increase percentage of meters with bad data to 4%, 8%, and 12%. Table V shows the phase identification accuracy of our proposed method using 0.1% meter class when the feeders are close to balance. The average accuracy of 30 test cases are reported in the table. In each test case, we randomly select the meters with bad data. As shown in Table V, the phase identification accuracy of our proposed algorithm gradually decreases as more meters are compromised with bad data. However, even with 10% bad meters, our algorithm can still achieve over 93% accuracy on the most complex circuit with 90 days of smart meter data.

TABLE V
AVERAGE ACCURACY OF THE PROPOSED PHASE IDENTIFICATION METHOD WITH BAD DATA (METER ACCURACY CLASS 0.1%)

Feeder	% of Meters with Bad Data	30 Days	60 Days	90 Days
37-bus	0%	100%	100%	100%
	4%	98.53%	98.53%	100%
	8%	96.53%	97.20%	99.07%
	12%	91.47%	95.73%	97.87%
123-bus	0%	96.47%	100%	100%
	1%	94.94%	99.80%	99.92%
	5%	90.51%	99.33%	99.76%
	10%	88.51%	98.43%	99.61%
342-bus	0%	96.63%	100%	100%
	1%	95.06%	98.93%	99.17%
	5%	91.54%	96.47%	97.15%
	10%	86.71%	93.35%	93.83%

C. Comparison With Existing Methods

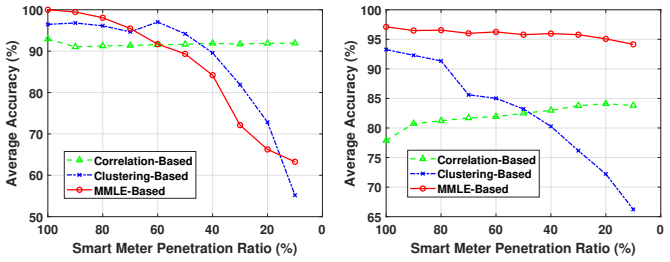
The phase identification accuracy of our proposed MMLE-based method is compared with two state-of-the-art methods: the correlation-based approach [10] and the clustering-based approach [12]. We also evaluate the robustness of the phase identification algorithms with respect to inaccurate feeder models, incomplete measurements, and bad data.

The 123-bus and 342-bus test feeders with 90 days of 0.1% accuracy class smart meter measurements are used for the comparison. To introduce incomplete smart meter measurements, we gradually decrease the penetration ratio of smart meters from 100% to 10% with a 10% step. To create inaccurate feeder models, we introduce noisy network parameters and inaccurate topology information. Specifically, we add zero-mean Gaussian noise with three-sigma deviation matching 30% of the nominal values to the actual line admittance of the 123-bus and 342-bus feeders. Eight secondary branches are assumed to be missing in the topology model of the 342-bus feeder.

Note that the correlation-based method [10] was originally designed to handle single-phase loads only. Thus, we extend it to accommodate two-phase loads. To make it a fair comparison, we assume that the information of whether a particular load is one-phase, two-phase, or three-phase is known to all algorithms. Inaccurate feeder models and incomplete measurements do not affect the correlation-based and clustering-based algorithms directly. This is because these two methods do not rely on the primary feeder model. Similarly, the MMLE-based method simply constructs a formulation with a smaller decision vector $\boldsymbol{x}$ when dealing with incomplete meter measurements.

The average phase identification accuracies of the proposed algorithm and two benchmark algorithms with different smart meter penetration ratios and inaccurate feeder models are shown in Fig. 2. When the smart meter penetration rate is not 100%, we randomly select the location of smart meters around 50 times and calculate the average accuracies.

As shown in Fig.2, our proposed MMLE-based algorithm achieves around 97% accuracy on the 342-bus feeder at the 100% smart meter penetration rate. This is lower than the 100% accuracy reported in Table III due to an inaccurate



(a) 123-bus (b) 342-bus

Fig. 2. Average phase identification accuracy of three methods with inaccurate feeder models (0.1% meter class, 90 days' data).

primary feeder model. Our proposed algorithm yields higher accuracy for the 123-bus radial feeder when the smart meter penetration rate is at 70% or higher. Note that the smart meter penetration level is already higher than 70% in hundreds of thousands of around the world and keeps increasing around the world. In the U.S. more than 40 electric companies have fully deployed smart meters [18] by the end of 2016. The smart meter penetration level in North America is expected to reach 81% in 2024 [19]. In European countries such as Italy, Sweden, Finland, and the Netherlands, smart meter penetration levels have reached 80% and are expected to pass 95% in 2020 [20]. As the penetration level of smart meters continue to increase around the world, the comparative advantage of our proposed algorithm will become more pronounced.

For the more complex 342-bus feeder, which is heavily meshed, our proposed algorithm outperforms both existing algorithms across all smart meter penetration levels. Our proposed algorithm is more robust with respect to incomplete measurements on the heavily meshed 342-bus feeder than on the radial 123-bus feeder. To explain this phenomenon, we examine the sensitivity of $\tilde{v}_m(t, \boldsymbol{x})$, the smart meter voltage measurement for load m, with respect to the phase connection decision vector $\boldsymbol{x}$. It turns out that in the 342-bus feeder, load m's voltage measurement is more sensitive to its own phase connection decision variables and less sensitive to the phase connection decision variables of other loads.

Finally, we compare the performance of all three phase identification methods when there is bad data in the smart meter measurements. Table VI shows the phase identification accuracy under different ratios of meters with bad data, using 90 days of 0.1% accuracy class meter measurements. The result is based on 100% smart meter penetration level and accurate feeder parameters. The average accuracies over 30 test cases are reported. In each test case, we randomly select the meters with bad data. As shown in Table VI, in the presence of bad data, our proposed phase identification algorithm always yields higher accuracy than the two benchmark algorithms when the smart meter penetration rate is 100%.

VI. Conclusion

This paper develops a physically inspired data-driven algorithm for the phase identification in power distribution systems. The phase identification problem is first formulated as an MLE and MMLE problem based on the three-phase power flow manifold. We prove that the correct phase connection is a global optimum for both the MLE and the MMLE problems. A computationally efficient algorithm is developed to solve the MMLE problem, which involves synthesizing the solutions from the sub-problems via the voting and the target-only approaches. The sub-problems are further transformed into an equivalent binary least square form and solved efficiently by relaxing the binary constraints. Comprehensive simulation results with real-world smart meter data and IEEE distribution test circuits show that our proposed phase identification algorithm yields high accuracy and outperforms existing methods. The proposed algorithm is also fairly robust with respect to inaccurate feeder models, incomplete measurements, and bad measurements.

TABLE VI
Average Accuracy of Three Phase Identification Methods With Bad Data (Meter Accuracy Class 0.1%, 90 Days' Data)

Feeder	% of Meters with Bad Data	Correlation-Based Approach [10]	Clustering-Based Approach [12]	MMLE-Based Algorithm
37-bus	0%	92%	100%	100%
	4%	91.87%	98.40%	100%
	8%	92%	97.60%	99.07%
	12%	91.73%	96.40%	97.87%
123-bus	0%	92.94%	96.47%	100%
	1%	88.35%	96.43%	99.92%
	5%	89.14%	94.24%	99.76%
	10%	89.33%	92.51%	99.61%
342-bus	0%	77.88%	93.27%	100%
	1%	80.51%	51.76%	99.17%
	5%	80.54%	49.55%	97.15%
	10%	79.25%	48.51%	93.83%

Appendix A
Simplification of Single-Phase and Two-Phase Branches

To convert loaded single-phase and two-phase branches into a load directly connected to the primary feeder, we need to estimate each branch's equivalent power injection and voltage magnitude. In other words, given the line impedances of single-phase and two-phase branches, the voltage magnitudes and power injections of the loads, we need to calculate the equivalent power injection and voltage magnitude on the primary feeder. The conversion of single-phase and two-phase branches is carried out separately below.

1) Simplification of a Single-Phase Line: Suppose there is a single-phase line with impedance z serving a load with power injection S and voltage magnitude $|V|$. It is assumed that the power injection S and the voltage magnitude $|V|$ are given. Thus, the current injection magnitude $|I|$ and power factor angle ϕ can be calculated. Then, at the upstream port of the primary feeder, the single-phase line's equivalent voltage magnitude is $||V| - z|I|\angle -\phi|$ and the equivalent power injection is $S - z|I|^2$.

2) Simplification of a Two-Phase Line: For a two-phase line serving a load, the voltage drop along the line section can be described by

$$\begin{bmatrix} V_n^1 \\ V_n^2 \end{bmatrix} = \begin{bmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \end{bmatrix} \begin{bmatrix} -I \\ I \end{bmatrix} + \begin{bmatrix} V_m^1 \\ V_m^2 \end{bmatrix} \tag{33}$$

where z_{11}, z_{12}, z_{21}, and z_{22} form the line impedance matrix, which is assumed to be known. V_n^1, V_n^2, V_m^1, and V_m^2 are the nodal voltage phasors of the upstream port and the load, which are assumed to be unknown. I is the current injection phasor of the load. Subtracting row 2 from row 1 in (33), we have

$$V_n^{12} = (z_{12} + z_{21} - z_{11} - z_{22})I + V_m^{12} = z_{sum}I + V_m^{12} \quad (34)$$

where $V_n^{12} = V_n^1 - V_n^2$ and $V_m^{12} = V_m^1 - V_m^2$. For load m, using the measured voltage magnitude $|V_m^{12}|$ and power injection S_m, we calculate the current injection magnitude $|I|$ and the power factor angle ϕ. Then, at the upstream port of the primary feeder, the two-phase line's equivalent voltage magnitude is $||V_m^{12}| + z_{sum}|I|\angle -\phi|$ and the equivalent power injection is $S_m + z_{sum}|I|^2$.

Appendix B
Derivation of the Transformed Linearized Three-phase Power Flow Model

Let A_{mn}^{ij} be the $(N+1)\times(N+1)$ block in matrix A_{mn} corresponding to phase ij. Suppose the first row and column of A_{mn} correspond to the substation node, then A_{mn}^{ij} can be divided into 4 blocks as follows:

$$A_{mn}^{ij} = \begin{bmatrix} d_{mn}^{ij} & (\boldsymbol{b}_{mn}^{ij})^T \\ \boldsymbol{b}_{mn}^{ij} & \check{A}_{mn}^{ij} \end{bmatrix} \quad (35)$$

where $\check{A}_{mn}^{ij}$ is a nonsingular $N\times N$ matrix. Define $\check{A}_{mn}$ as the collection of $\check{A}_{mn}^{ij}$ over all i and j, B_{mn} as the collection of $\boldsymbol{b}_{mn}^{ij}$ over all i and j, C_{mn} as the collection of $(\boldsymbol{b}_{mn}^{ij})^T$ over all i and j, and D_{mn} as the collection of d_{mn}^{ij} over all i and j. By permuting the variables and corresponding matrix rows and columns, (1) can be transformed into

$$\begin{bmatrix} \check{A}_{11} & \check{A}_{12} & B_{11} & B_{12} \\ \check{A}_{21} & \check{A}_{22} & B_{21} & B_{22} \\ C_{11} & C_{12} & D_{11} & D_{12} \\ C_{21} & C_{22} & D_{21} & D_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_{-0} - \overline{\boldsymbol{v}}_{-0} \\ \boldsymbol{\theta}_{-0} - \overline{\boldsymbol{\theta}}_{-0} \\ \boldsymbol{v}_0 - \overline{\boldsymbol{v}}_0 \\ \boldsymbol{\theta}_0 - \overline{\boldsymbol{\theta}}_0 \end{bmatrix} = \begin{bmatrix} \boldsymbol{p}_{-0} \\ \boldsymbol{q}_{-0} \\ \boldsymbol{p}_0 \\ \boldsymbol{q}_0 \end{bmatrix} \quad (36)$$

where $(\cdot)_{-0}$ denotes a vector excluding the substation node, and $(\cdot)_0$ denotes a vector of the substation node.

Define Matrix $\mathcal{D}$ as follows:

$$\mathcal{D} = \text{diag}(\mathbf{1}_N, \mathbf{1}_N, \mathbf{1}_N, \mathbf{1}_N, \mathbf{1}_N, \mathbf{1}_N) \quad (37)$$

From the property of admittance matrix Y^{ij}, we have $A_{mn}^{ij}\mathbf{1}_{N+1} = \mathbb{0}_{(N+1)\times 1}$ and $[\check{A}_{mn}^{ij}, \boldsymbol{b}_{mn}^{ij}]\mathbf{1}_{N+1} = \mathbb{0}_{N\times 1}$.

Thus, we have the following equality relationship:

$$\begin{bmatrix} \check{A}_{11} & \check{A}_{12} & B_{11} & B_{12} \\ \check{A}_{21} & \check{A}_{22} & B_{21} & B_{22} \end{bmatrix} \begin{bmatrix} \mathcal{D} \\ I_{6\times 6} \end{bmatrix} = \mathbb{0}_{6N\times 6} \quad (38)$$

Now, it can be easily shown that

$$\begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = -\begin{bmatrix} \check{A}_{11} & \check{A}_{12} \\ \check{A}_{21} & \check{A}_{22} \end{bmatrix} \mathcal{D} \quad (39)$$

Plugging equation (39) into equation (36), we have

$$\begin{bmatrix} \check{A}_{11} & \check{A}_{12} \\ \check{A}_{21} & \check{A}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_{-0}^a - \mathbf{1}_N v_0^a \\ \boldsymbol{v}_{-0}^b - \mathbf{1}_N v_0^b \\ \boldsymbol{v}_{-0}^c - \mathbf{1}_N v_0^c \\ \boldsymbol{\theta}_{-0}^a - \mathbf{1}_N \theta_0^a \\ \boldsymbol{\theta}_{-0}^b - \mathbf{1}_N \theta_0^b \\ \boldsymbol{\theta}_{-0}^c - \mathbf{1}_N \theta_0^c \end{bmatrix} = \begin{bmatrix} \boldsymbol{p}_{-0} \\ \boldsymbol{q}_{-0} \end{bmatrix} \quad (40)$$

where $\boldsymbol{v}_{-0}^i$ and $\boldsymbol{\theta}_{-0}^i$ denote the phase i variables in $\boldsymbol{v}_{-0}$ and $\boldsymbol{\theta}_{-0}$. v_0^i and θ_0^i denote the substation's voltage magnitude and angle of phase i. (40) is exactly the same as (3).

Appendix C
Estimation of Nodal Power Injection of a Two-phase Load

Define I_{ab} as the current phasor flowing out of the load's phase A port and into the load's phase B port. Let I_a be the injected current phasor from phase A port, and let I_b be the injected current phasor from phase B port. By definition, we know that $I_a = -I_b = I_{ab}$. Let the angle of V_{ab} be the reference angle, i.e., $V_{ab} = |V_{ab}|\angle 0°$, then

$$\begin{aligned} S_{ab} &= P_{ab} + jQ_{ab} \\ &= V_{ab}I_{ab}^* \\ &= |V_{ab}|[Re(I_{ab}) - jIm(I_{ab})] \end{aligned} \quad (41)$$

Thus,

$$\begin{aligned} Re(I_{ab}) &= \frac{P_{ab}}{|V_{ab}|} \\ Im(I_{ab}) &= -\frac{Q_{ab}}{|V_{ab}|} \end{aligned} \quad (42)$$

When the three-phase voltages are close to balance, the nodal phase-to-neutral power injection can be estimated by the two-phase power injection as follows:

$$\begin{aligned} S_a &= V_a I_a^* \\ &\approx \frac{\sqrt{3}}{3}|V_{ab}|\angle -30° \cdot I_{ab}^* \\ &= \frac{\sqrt{3}}{3}|V_{ab}|\angle -30°\left(\frac{P_{ab}}{|V_{ab}|} + j\frac{Q_{ab}}{|V_{ab}|}\right) \\ &= \frac{\sqrt{3}}{3}[\cos(-30°) + j\sin(-30°)](P_{ab} + jQ_{ab}) \\ &= \left(\frac{1}{2}P_{ab} + \frac{\sqrt{3}}{6}Q_{ab}\right) + j\left(\frac{1}{2}Q_{ab} - \frac{\sqrt{3}}{6}P_{ab}\right) \end{aligned} \quad (43)$$

This is exactly the same as (8). Equation (9) can be derived in a similar way.

Appendix D
Link the Voltage Magnitude Measurements of Two-Phase Loads to Nodal Values in the Power Flow Model

In the following derivations, the voltages are in per unit and angles are in radian. For a two-phase load m across phase ij ($ij \in \{ab, bc, ca\}$) at node n, we have

$$\hat{v}_m = v_n^{ij} = \sqrt{(v_n^i)^2 + (v_n^j)^2 - 2v_n^i v_n^j \cos\theta_n^{ij}} \quad (44)$$

where $\hat{v}_m$ is load m's magnitude measurement, v_n^{ij} is the voltage magnitude between phase ij at node n, v_n^i is the voltage of phase i at node n, and θ_n^{ij} is the voltage phase angle between phase ij at node n.

Similarly, at the substation, we also have

$$v_0^{ij} = \sqrt{(v_0^i)^2 + (v_0^j)^2 - 2v_0^i v_0^j \cos\theta_0^{ij}} \quad (45)$$

where v_0^{ij}, v^i, and θ_0^{ij} are the corresponding nodal values at the substation. Under normal operating conditions, $v_n^i \approx v_n^j \approx 1$, $\theta_n^{ij} \approx \frac{2\pi}{3}$. From (44) we have

$$\frac{\partial v_n^{ij}}{\partial v_n^i} \approx \frac{\sqrt{3}}{2}, \quad \frac{\partial v_n^{ij}}{\partial v_n^j} \approx \frac{\sqrt{3}}{2}, \quad \frac{\partial v_n^{ij}}{\partial \theta_n^{ij}} = \frac{\partial v_n^{ij}}{\partial(\theta_n^i - \theta_n^j)} \approx \frac{1}{2} \tag{46}$$

Under normal operating conditions, voltage and angle differences between non-substation nodes and the substation node is very small. Thus, we can easily derive (10) from (46) to approximate $\hat{v}_m - v_0^{ij}$.

Appendix E
Proof of Lemma 1

Proof: By definition, $\tilde{\boldsymbol{v}}(t) = \tilde{\boldsymbol{v}}(t, \boldsymbol{x}^*) + \boldsymbol{n}(t)$. Plugging it into equation (19), we have

$$\begin{aligned}
&\lim_{T\to\infty} f(\boldsymbol{x}) \\
&= \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} [\tilde{\boldsymbol{v}}(t, \boldsymbol{x}^*) - \tilde{\boldsymbol{v}}(t, \boldsymbol{x}) + \boldsymbol{n}(t)]^T \Sigma_n^{-1} \\
&\quad [\tilde{\boldsymbol{v}}(t, \boldsymbol{x}^*) - \tilde{\boldsymbol{v}}(t, \boldsymbol{x}) + \boldsymbol{n}(t)] \\
&= \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} [\tilde{\boldsymbol{v}}(t, \boldsymbol{x}^*) - \tilde{\boldsymbol{v}}(t, \boldsymbol{x})]^T \Sigma_n^{-1} [\tilde{\boldsymbol{v}}(t, \boldsymbol{x}^*) - \tilde{\boldsymbol{v}}(t, \boldsymbol{x})] \\
&\quad + \lim_{T\to\infty} \frac{2}{T} \sum_{t=1}^{T} [\tilde{\boldsymbol{v}}(t, \boldsymbol{x}^*) - \tilde{\boldsymbol{v}}(t, \boldsymbol{x})]^T \Sigma_n^{-1} \boldsymbol{n}(t) \\
&\quad + \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{n}(t)^T \Sigma_n^{-1} \boldsymbol{n}(t) \\
&\geq \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{n}(t)^T \Sigma_n^{-1} \boldsymbol{n}(t)
\end{aligned} \tag{47}$$

It should be noted that $\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} [\tilde{\boldsymbol{v}}(t, \boldsymbol{x}^*) - \tilde{\boldsymbol{v}}(t, \boldsymbol{x})]^T \Sigma_n^{-1} [\tilde{\boldsymbol{v}}(t, \boldsymbol{x}^*) - \tilde{\boldsymbol{v}}(t, \boldsymbol{x})] \geq 0$ because $\Sigma_n^{-1} \succ 0$. As stated in condition 1 of Lemma 1, $\boldsymbol{n}(t)$ is independent of $\tilde{\boldsymbol{v}}(t, \boldsymbol{x})$ and $\tilde{\boldsymbol{v}}(t, \boldsymbol{x}^*)$, so we have $E([\tilde{\boldsymbol{v}}(t, \boldsymbol{x}^*) - \tilde{\boldsymbol{v}}(t, \boldsymbol{x})]^T \Sigma_n^{-1} \boldsymbol{n}(t)) = 0$. Condition 1 and 2 of Lemma 1 also make $[\tilde{\boldsymbol{v}}(t, \boldsymbol{x}^*) - \tilde{\boldsymbol{v}}(t, \boldsymbol{x})]^T \Sigma_n^{-1} \boldsymbol{n}(t)$ a sequence of independent variables. Under normal system operating conditions, $[\tilde{\boldsymbol{v}}(t, \boldsymbol{x}^*) - \tilde{\boldsymbol{v}}(t, \boldsymbol{x})]^T \Sigma_n^{-1} \boldsymbol{n}(t)$ has limited variance. By Kolmogorov's Strong Law of Large Numbers [21], $\lim_{T\to\infty} \frac{2}{T} \sum_{t=1}^{T} [\tilde{\boldsymbol{v}}(t, \boldsymbol{x}^*) - \tilde{\boldsymbol{v}}(t, \boldsymbol{x})]^T \Sigma_n^{-1} \boldsymbol{n}(t) \to 0$. Therefore, inequality (47) holds. In addition, the minimum of $\lim_{T\to\infty} f(\boldsymbol{x})$ is achieved when $\boldsymbol{x} = \boldsymbol{x}^*$. ■

Appendix F
Proof of Lemma 2

Proof: Following a procedure similar to Appendix E, we can prove that $\lim_{T\to\infty} f_m(\boldsymbol{x}) \geq \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} n_m(t)^2$, and the minimum of $\lim_{T\to\infty} f_m(\boldsymbol{x})$ is achieved when $\boldsymbol{x} = \boldsymbol{x}^*$. Condition 1) and 2) in Lemma 2 simply mean that we can assign any three-phase loads except load m to any phase and get the same optimum value. This is true, because changing three-phase loads' decision variables does not change the power injections in the system. As long as condition 1) and 2) of Lemma 2 hold, $\tilde{v}_m(t, \boldsymbol{x}) = \tilde{v}_m(t, \boldsymbol{x}^*)$. This can also be verified by the structure of $\hat{U}^1$ and $\hat{U}^2$ for three-phase loads. ■

References

[1] W. S. Bierer, "Long range phasing voltmeter," Oct. 5 2010, US Patent 7,808,228.

[2] M. H. Wen, R. Arghandeh, A. von Meier, K. Poolla, and V. O. Li, "Phase identification in distribution networks with micro-synchrophasors," in *2015 IEEE Power & Energy Society General Meeting*. IEEE, Jul. 2015, pp. 1–5.

[3] Y. Liao, Y. Weng, G. Liu, Z. Zhang, C. W. Tan, and R. Rajagopal, "Unbalanced three-phase distribution grid topology estimation and bus phase identification," *arXiv preprint arXiv:1809.07192 [cs.SY]*, Sep. 2018.

[4] V. D. Krsman and A. T. Sarić, "Verification and estimation of phase connectivity and power injections in distribution network," *Electric Power Systems Research*, vol. 143, pp. 281–291, Feb. 2017.

[5] M. Dilek, "Integrated design of electrical distribution systems: Phase balancing and phase prediction case studies," Ph.D. dissertation, Virginia Polytechnic Institute and State University, 2001.

[6] P. Kumar, V. Arya, D. A. Bowden, and L. Kohrmann, "Leveraging DERs to improve the inference of distribution network topology," in *2017 IEEE International Conference on Smart Grid Communications (SmartGridComm)*. IEEE, Oct. 2017, pp. 52–57.

[7] S. J. Pappu, N. Bhatt, R. Pasumarthy, and A. Rajeswaran, "Identifying topology of low voltage distribution networks based on smart meter data," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 5113–5122, Mar. 2018.

[8] T. A. Short, "Advanced metering for phase identification, transformer identification, and secondary modeling," *IEEE Transactions on Smart Grid*, vol. 4, no. 2, pp. 651–658, Jun. 2013.

[9] W. Luan, J. Peng, M. Maras, J. Lo, and B. Harapnuk, "Smart meter data analytics for distribution network connectivity verification," *IEEE Transactions on Smart Grid*, vol. 6, no. 4, pp. 1964–1971, Jul. 2015.

[10] M. Xu, R. Li, and F. Li, "Phase identification with incomplete data," *IEEE Transactions on Smart Grid*, vol. 9, no. 4, pp. 2777–2785, Jul. 2018.

[11] W. Wang, N. Yu, B. Foggo, J. Davis, and J. Li, "Phase identification in electric power distribution systems by clustering of smart meter data," in *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, Dec. 2016, pp. 259–265.

[12] W. Wang and N. Yu, "Advanced metering infrastructure data driven phase identification in smart grid," in *The Second International Conference on Green Communications, Computing and Technologies*, Sep. 2017, pp. 16–23.

[13] F. Olivier, A. Sutera, P. Geurts, R. Fonteneau, and D. Ernst, "Phase identification of smart meters by clustering voltage measurements," in *2018 Power Systems Computation Conference (PSCC)*. IEEE, Jun. 2018, pp. 1–8.

[14] S. Bolognani and F. Dörfler, "Fast power system analysis via implicit linearization of the power flow manifold," in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, Sep. 2015, pp. 402–409.

[15] A. M. Kettner and M. Paolone, "On the properties of the compound nodal admittance matrix of polyphase power systems," *IEEE Transactions on Power Systems*, vol. 34, no. 1, pp. 444–453, Aug. 2019.

[16] S. A. Vavasis, *Complexity theory: Quadratic programming*. Boston, MA: Springer US, 2001, pp. 304–307. [Online]. Available: https://doi.org/10.1007/0-306-48332-7_65

[17] K. Schneider, P. Phanivong, and J.-S. Lacroix, "IEEE 342-node low voltage networked test system," in *2014 IEEE PES General Meeting| Conference & Exposition*. IEEE, Jul. 2014, pp. 1–5.

[18] A. Cooper, "Electric company smart meter deployments: Foundation for a smart grid," The Edison Foundation, Tech. Rep., Dec. 2017. [Online]. Available: https://www.edisonfoundation.net/iei/publications/Documents/IEI_Smart%20Meter%20Report%202017_FINAL.pdf

[19] Berg Insight AB, "Smart metering–world 2019," Berg Insight AB, Tech. Rep., Jul. 2019. [Online]. Available: https://www.researchandmarkets.com/reports/4793286/smart-metering-world-2019

[20] IoT Analytics, "Smart meter market report 2019-2024," IoT Analytics, Tech. Rep., Nov. 2019. [Online]. Available: https://iot-analytics.com/smart-meter-market-2019-global-penetration-reached-14-percent/

[21] W. H. Greene, *Econometric Analysis*, 7th ed. Pearson, 2011.