Improving Supervised Phase Identification Through the Theory of Information Losses

Brandon Foggo, *Student Member, IEEE* and Nanpeng Yu, *Senior Member, IEEE*

Abstract—This paper considers the problem of Phase Identification in power distribution systems. In particular, it focuses on improving supervised learning accuracies by focusing on exploiting some of the problem's information theoretic properties. This focus, along with recent advances in Information Theoretic Machine Learning (ITML), helps us to create two new techniques. The first transforms a bound on information losses into a data selection technique. This is important because phase identification data labels are difficult to obtain in practice. The second interprets the properties of distribution systems in the terms of ITML. This allows us to obtain an improvement in the representation learned by any classifier applied to the problem. We tested these two techniques experimentally on real datasets and have found that they yield phenomenal performance in every case. In the most extreme case, they improve phase identification accuracy from 51.7% to 97.3%. Furthermore, since many problems share the physical properties of phase identification exploited in this paper, the techniques can be applied to a wide range of similar problems.

Index Terms—Distribution network, information theoretic machine learning, phase identification, supervised learning.

Fig. 1: Illustration of the Phase Identification Problem. Here, each T represents a transformer, and each combination of the letters A, B, and C refers to a phase connection type.

I. INTRODUCTION

A power distribution circuit encompasses several components. It contains busses, powerlines, substations, regulators, transformers, and more. The physical arrangement of these components constitute the circuit's topology, which dictates much of the system's operation and planning. Over time, a circuit's topology will change. For example, a power outage may initiate a structural change such that the number of the effected customers is minimized. But the topology documentation will often not follow this change - they are only typically updated after major distribution expansion projects. As a result, there are long periods of time in which a circuit's topology is wrong, and this poses a serious problem. Power flow analysis, state estimation, and Volt-VAR control all depend on accurate topological information. When this is not available, the usefulness of those methods diminishes, and the system runs less effectively as a whole. Methods for faster documentation updates are necessary.

Critical to the application of topology identification is the subproblem of phase identification. This subproblem describes the composition of each powerline in the network. Typically, a primary distribution powerline is made up of the four fundamental lines A, B, C and n. The primary feeder, fed directly from the power substation, often consists of all four. However, at some point, a subset of these lines will be branched from the feeder. This change usually happens once along the path from the substation to any customer. As such, we define a customer's 'phase type' as the lines branched along that
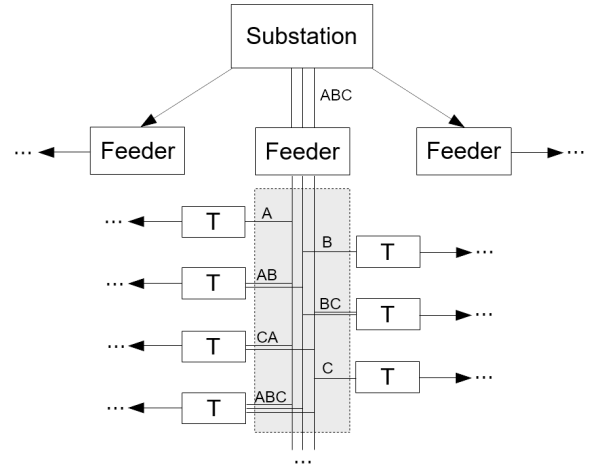
path. This is illustrated in Figure 1. Knowledge of these phase types is necessary for estimation of electrical distances, and for topology reconstruction in general. In fact, since topology estimation is often taken as a precursor to state and parameter estimation in distribution networks, we can think of phase identification as the first step in a pipeline of distribution system modelling techniques.

Literature on phase identification exists, but is limited. Like most power system applications, there are two broad classes of phase identification approaches - one model based, and the other data-driven. Like those other applications, model based methods have high interpretability but low accuracy, and data-driven methods have high accuracy but low interpretability. Our research poses itself as an intermediate between these extremes. This is done through Information Theoretic Machine Learning (ITML). This branch of machine learning connects traditional learning theory to the field of Information Theory. As such, interpretable notions such as entropy and mutual information are studied. These interpretable measures can be connected to the physics of the problems that we wish to solve.

We will focus on a particular regime of ITML which needs further development - the small data regime. Phase identification methods, particularly supervised methods, require a lot of training data to achieve high accuracies. But obtaining this training data requires portions of the lengthy field tests which we were trying to avoid in the first place. Thus we have a trade-off between phase identification accuracy and the amount

of time that a distribution circuit's topology documentation is incorrect. In this paper, we will develop methods to obtain higher phase identification accuracies even when small training datasets are used.

In particular, we will develop two techniques: Inverse Schur Training Data Selection, and Information Loading. The first corresponds to selecting the most informative data-points prior to field testing. The second consists of a modification to the objective function of a standard learning algorithm, with some modifications to the training phase in order to support it. We will heavily emphasize the theoretical reasoning behind these techniques, particularly in why they are applicable to the phase identification problem. Our proposed techniques have the following benefits:

- They require little infrastructure or physical labor.
- They do not require any modeling of the network.
- They are robust with respect to missing data.
- They are easy to implement and tune.
- They can handle any variety of phase connections.
- They provide the best supervised accuracies to date.
- The representations learned are highly meaningful.
- They are generalizable to other networked systems.

II. RELATED WORK

Much phase identification work is based on physical approaches [1]–[4]. References [1] and [2] develop phase identification systems based on high resolution timing measurements communicated between the base station and the feeder transformer secondaries and/or individual electricity consumers. This system is highly accurate and even yields the voltage phasors of the secondaries themselves instead of just the phase names of the wires connected to them. However, the system is quite sophisticated. Deploying such a system for each feeder across the many distribution circuits overseen by a utility is very expensive.

Reference [3] describes a phase identification technique using micro-synchrophasors. The overhead cost of this method is less than the that of reference [1] since micro-synchrophasors are mobile - hence only a few devices are required. However, this reduced overhead cost results in increased labor and time costs, as the micro-synchrophasors must be reinstalled several times throughout the distribution network. Reference [4] patents a method for phase identification through signal injection. A signal generator is placed at the base substation and a unique signal is created for each phase. These signals are detected by a signal discriminator at each customer location. By matching the signals, the phase connectivities can be accurately reproduced. Like reference [1], this method is very intensive in labor and time, but relatively cheap in overhead. These physical methods are the approaches that are typically taken during the field testing projects employed in practice. Furthermore, they act as the base upon which training labels should be acquired in application of this work.

The amount of literature related to solving of phase identification problem with data-driven methods is more limited. Of what does exist, most is unsupervised. This unsupervised branch can be split further into 'model-agnostic' and 'model-based' sub-branches.

In the model-based sub-branch, reference [5] compares simulated power-flow solutions for a given phase configurations to real data. This method is accurate, but requires a correct system model including everything except phase connectivity's - e.g. line parameters, network topology. References [6] and [7] group customers by phase such that the total sum of power injections on each phase matches that of the substation or distribution transformers up to some error. These methods are somewhat non-robust to missing or erroneous data since these will lead to power mismatches between the measured load and the measured supply. Furthermore, phase configuration is often assumed to already exist in methods of network topology estimation and parameter estimation, so we typically like to think of phase identification as the first step in a long sequence of modelling techniques. Assuming realistic models as an input to the phase identification problem is somewhat unrealistic at this time.

In the model-agnostic sub-branch, one of the most popular techniques is to correlate voltage time series data at a household to the voltage time series data captured with other households [8]–[12]. These works showcase the power of using voltage data as a primary predictor of phase type - an idea which we will study in depth in this paper. However, while the statement 'customer A and customer B are on the same phase implies customer A and customer B have correlated voltages' is mostly accurate, these methods struggle in two ways. First, the converse of this statement is not always accurate. That is, customers can be on different phases and still have correlated voltages. Second, knowing that two customers are on the same phase doesn't actually tell us what that phase *is*. This second issue can be resolved by either physical inspection, which are time and labor intensive, or by correlating the relevant voltages to those of the substation [13]. But substation voltages are usually only measured either line to line or line to neutral - not both simultaneously, and so this limits the scope of circuits that we are capable of classifying.

Finally, some unsupervised clustering techniques [14], [15] and frequency domain feature extraction techniques [11] have been tested on this problem. Supervised methods are limited to off-the-shelf algorithms that don't take any consideration into the properties of power distribution networks. To the author's knowledge, the method presented in this paper is the first of its type.

III. BACKGROUND - INFORMATION LOSSES

A. Information Losses

ITML studies the following loss term:

$$I_{Loss} = \left| I(Z,Y) - \mathbb{E}\left[log_2 \frac{\hat{p}(z,y)}{\hat{p}(z)\hat{p}(y)}\right] \right| \qquad (1)$$

This equation defines some notation which we will need to go over. First, $Y \in \mathcal{Y}$ is a random variable denoting the class variable in the machine learning problem. Here, Y refers to the phase type of each customer. Next, we define a random variable $X \in \mathcal{X}$ (not appearing in (1)) used to represent the features of the machine learning problem. Here, X will refer to a voltage time-series. Finally, Z refers to some random variable

generated as a stochastic function of X. This can technically refer to *any* such random variable. However, we think of it as consisting of the internal state of a machine learning algorithm we wish to use. For example, Z may refer to the value of a hidden layer in a neural network.

The term $I(Y; Z)$ refers to the mutual information between Y and Z [16].

$$I(Y; Z) = \int log_2 \frac{dP_{YZ}}{d(P_Y \otimes P_Z)} dP_{YZ} \qquad (2)$$

The $\hat{p}$ notation refers to an estimated probability distribution - for example, a histogram in the discrete case, or a parametric model with a slightly incorrect parameter realization. We will further denote the rightmost term in equation (1) as $\hat{I}(Y; Z)$. In general, all information theoretic terms pertaining to the estimated model will be denoted with a hat.

We are interested, in particular, in these information quantities for two specific types of random variables:

Definition 1. *Let $\epsilon > 0$. Let Z_ϵ^* and $\hat{Z}_\epsilon$ be random variables that are at most ϵ-suboptimal for the following information optimization problems:*

$$\sup_{p(z|x)} I(Y; Z)$$
$$subject\ to\ I(X; Z) = I$$

$$\sup_{p(z|x)} \hat{I}(Y; Z)$$
$$subject\ to\ I(X; Z) = I$$

We think of these representations as follows: Z^* is the optimal estimator when we have perfect knowledge of our dataset, and $\hat{Z}$ is the optimal estimator when using a finite training dataset to estimate it. Both optimizations are subject to a constraint on the complexity of the estimator which comes from the Information Bottleneck method [17]–[19]. Given these definitions, we further define the information losses associated with the sample S by:

$$I_{loss}(S) = |I(Y; Z^*) - I(Y; \hat{Z})| \qquad (3)$$

This quantity is strongly related to 'minimal classification error'. There exist several inequalities indicating that classification error cannot be lower than a monotonic function of $I_{Loss}(S)$ [20]. The most well known inequality of this type is due to Fano [16], which state that, if $I(Y; Z^*) = H(Y)$ (i.e. the best *possible* representation Z^* is actually perfect), then:

$$h_2(P_e) + P_e log_2(|\mathcal{Y}| - 1) \geq I_{loss}(S) \qquad (4)$$

where P_e is the minimum probability of classification error for a classifier using representation $\hat{Z}$.

It has been shown [21] that there is a relationship between $I_{loss}(S)$ and a measure known as the *conditional total variation* which is defined bellow:

Definition 2. *Given a conditional probability distribution $p(y|x)$ and an estimated distribution $\hat{p}(y|x)$, the conditional total variation of $\hat{p}(y|x)$ from $p(y|x)$ is given by*

$$\delta_{\hat{p}} = \mathbb{E}_{\mathbb{P}_X} \left[\frac{1}{2} \sum_y |p(y|x) - \hat{p}(y|x)| \right] \qquad (5)$$

This measure converges under standard machine learning regimes at least as fast as cross-entropy errors. The exact relationship between the term and information losses is given as follows:

Lemma 1. *Let $\hat{p}(y|x)$ be a probability distribution estimate established via the the estimator $\hat{Z}_\epsilon(x)$. Let h_2 denote the binary entropy function $h_2(t) = -t log_2(t) - (1 - t) log_2(1 - t)$. Then*

$$I_{loss}(S) \leq 2 \left(\delta_{\hat{p}} I(X; Z) + h_2(\delta_{\hat{p}}) \right) + \epsilon \qquad (6)$$

In this paper, we will consider techniques of mitigating information losses for the phase identification problem. This will be primarily through careful considerations of (6).

IV. TECHNIQUE - INVERSE SCHUR TRAINING DATA SELECTION

A. Motivation - Field Testing

The first technique addresses the technical limitations of supervised phase identification. Namely, that obtaining labeled data is time consuming. Obtaining phase labels for a given customer requires on-site measurements with phasor measurement unit or phasing meter. Gathering phase connection information for a large number of customers with these equipment to serve as labels would be prohibitively costly.

It is critical that we get as much as we can out of just a few phase labels. This inspires the use of active learning [22], [23]. This is a field of machine learning which focuses on the selection of training data points that best represent the whole data set. It can be loosely divided into unsupervised methods and supervised methods - with the latter division dominating the number of available techniques by a large margin. Supervised methods generally update a machine learning model in conjunction with label selection. But this step is lengthy on its own, and can not be performed in parallel with on-location travelling or with device installation. Thus supervised techniques may be deemed too lengthy when data acquisition requires field testing. Thus we will focus on unsupervised techniques. We will build upon some recent theoretical work on the link between information losses and data selection [24].

B. Training Data Selection via Information Losses

Recent theoretical work has found a bound on the term $\delta_{\hat{p}}$ in equation (6). It requires some notational setup to describe.

We assume that we are given an unlabelled dataset $\mathcal{D}$ of size N. We wish to select a subset $S \subset \mathcal{D}$ of cardinality $M < N$. Let $k(\cdot, \cdot)$ be a continuous, symmetric, positive definite kernel function. Let K denote the matrix

$$\begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \cdots k(x_1, x_N) \\ k(x_2, x_1) & k(x_2, x_2) & \cdots k(x_2, x_N) \\ \cdots & \cdots & \cdots \\ k(x_N, x_1) & k(x_N, x_2) & \cdots k(x_N, x_N) \end{bmatrix} \qquad (7)$$

Let S denote the selected training dataset, $S \subseteq \mathcal{D}$, and let $SS = \{i : x_i \in S\}$. Finally, let K/K_{SS} denote the schur complement of K with respect to K_{SS}. That is, if $\tilde{K}$ is the

permuted form of K such that the training indices correspond to the upper left block

$$\tilde{K} = \begin{bmatrix} \tilde{K}_{SS} & \tilde{K}_{SU} \\ \tilde{K}_{US} & \tilde{K}_{UU} \end{bmatrix} \qquad (8)$$

then

$$K/K_{SS} = \tilde{K}_{UU} - \tilde{K}_{US}\tilde{K}_{SS}^{-1}\tilde{K}_{SU} \qquad (9)$$

Now let T denote the following operator on $L^2_{\mathcal{P}_X}$:

$$T[f](y) = \int k(y,x)f(x)d\mathbb{P}_X(x) \qquad (10)$$

The bound on $\delta_{\hat{p}}$ is expressed in the following lemma

Lemma 2. *Assuming that our learner performs better than kernel optimization of the expected total variation, we have:*

$$\delta_{\hat{p}} \leq \frac{p(Y=1)}{N}Trace(\sqrt{K/K_{SS}}) + \epsilon_{\mathcal{H}} \qquad (11)$$

where $\epsilon_{\mathcal{H}}$ is given by

$$\epsilon_{\mathcal{H}} = \frac{1}{2}\sum_y \|(I-T)p(Y=y|x)\|_{L^1_{\mathcal{P}_X}} \qquad (12)$$

$\sqrt{A}$ *refers to the element-wise square root of the matrix A.*

Thus to reduce information losses, we desire to minimize the term $Trace(\sqrt{K/K_{SS}})$ under a well fitting kernel (i.e. such that the second term is small). For the phase identification problem, and for problems whose datasets are generated by models with well-fitting linearizations, the cosine kernel works well. Thus we wish to solve

$$\min_S Trace\sqrt{K/K_{SS}} \qquad (13)$$

$$s.t.\ k(x,y) = \frac{x^Ty}{\|x\|\|y\|} \qquad (14)$$

However, strict optimization of (13) will take exponential time. Furthermore, since each evaluation of K/K_{SS} for a given S will require a matrix inverse calculation, greedy optimization performs poorly as well. Indeed, the computational complexity grows with $O(NM^4)$. If we choose M such that $M = \rho N$ for some proportion ρ, then this scales with $O(N^5)$, which is quite poor.

However, we can heuristically optimize (13) in $O(N^3)$ time as follows: First, we note that

$$K^{-1} = \begin{bmatrix} \cdot & \cdot \\ \cdot & (K/K_{SS})^{-1} \end{bmatrix} \qquad (15)$$

Thus we can control the value of $Trace((K/K_{SS})^{-1})$ by data point selection in a predictable way. That is, the trace of $(K/K_{SS})^{-1}$ will be large if we pick points corresponding to small diagonal elements in K^{-1}.

We then note a few correspondences between the trace of a matrix and the trace of its inverse. First, by the Cauchy-Schwartz inequality, we have:

$$len(A) \leq Tr(A)Tr(A^{-1}) \qquad (16)$$

where $len(A)$ is the size of either axis of A.

Furthermore, we have the following lemma:

Lemma 3. *Consider the partially ordered set of matrices $M = \{A : A = K/\{\alpha\}$ for some index α of size $M\}$ in the Loewner order. That is, $A \leq B$ iff. $A - B$ is positive semi-definite. Let A^* be the matrix in this poset corresponding to the index α^* which maximizes $(K/\{\alpha^*\})^{-1}$. Then there exists no matrix $B \in M$ such that $A^* > B$.*

This lemma, which we will prove shortly, shows that picking data points to maximize $Tr((K/K_{SS})^{-1})$ will force $Tr(K/K_{SS})$ to be smaller than that of a large range of other datasets. Specifically, if B is comparable to A^* in the Loewner order (i.e. either A^*-B or $B-A^*$ is positive semidefinite) then A^* must be the smaller of the two and so $Tr(A^*) \leq Tr(B)$ by monotonicity of the trace operator. We now proceed to this lemma's proof.

Proof. By the hypothesis of the lemma, A^* has the largest inverse trace of all matrices in M. That is, $Tr(C^{-1}) \leq Tr((A^*)^{-1})$ for all $C \in M$. Now suppose for the sake of contradiction that there exists $B \in M$ such that $A^* > B$. Then $(A^*)^{-1} < B^{-1}$ by monotanacity of the inverse operator. This, in turn, implies that $Trace((A^*)^{-1}) < Trace(B^{-1})$, contradicting the hypothesis. $\square$

V. Technique - Information Loading

A. Voltage Data and Phase Identification

The second technique exploits the properties of our feature space. We will first describe what our feature space is and why we use it. We will then move into deriving the exploitable properties of this space - the main assertion being that a standardized voltage dataset has relatively low entropy. We will then experimentally validate this assertion. Finally, we will describe a technique, called information loading, which will exploit this property.

Our feature space consists of smart-meter voltage magnitude time series data. Voltage data is fairly informative of phase type. Thus it is a good candidate feature space for the phase identification problem. This can be seen through the following example.

Consider a power injection at bus k whose phase type is AB. This induces a current along the lines A and B. Thus, a voltage change will occur along those lines throughout the circuit. Any customer also feeding from either of those lines will notice a change. Due to the capacitive and inductive effects of the primary feeder, both lines will also induce a voltage change along the lines C and n. However, the off-diagonal elements of the phase impedance and shunt admittance matrices are much smaller than the diagonal ones. Hence, the power injection at bus k will have much less effect on phase C than phase A and B. Thus, the customers whose voltage data is most effected by this power injection are those on AB. Second to these customers are customers who share either the phase line A or B (An, Bn, BC, CA). Finally, customers who do not pull from either A or B (Cn) will hardly be effected at all.

Thus, while this current injection somewhat affects all customers, it affects customers of the same phase the most. Hence, voltage data is informative of phase connection type.

B. Properties of Voltage Data

We will now provide some rough quantitative descriptions of distribution systems. We consider a simple model in which each 'customer' corresponds to a distribution transformer branched from the primary feeder. That is, we are ignoring the properties of the secondary distribution. We assume that bus 0 represents the distribution substation and that each bus is a fixed electrical distance Δ_E from the previous bus. We will further assume that the impedance along the primary feeder is a constant z. We assume without loss of generality that the customers are indexed in order by their electrical distance from the distribution transformer. We let $V_p(x)$ denote the voltage on phase p at position x. We assume that we are working in a per unit system and that our system is fairly balanced, such that $V_A(0) \approx 1$, $V_B(0) \approx e^{-\frac{2\pi}{3}}$, $V_C(0) \approx e^{\frac{2\pi}{3}}$, $V_n(0) \approx 0$. We will assume for simplicity that there are no three phase loads in the circuit.

Now, let μ denote the degenerate measure $\sum_{j=1}^{N} \delta(j\Delta_E)$ and let i denote a stochastic process over $[0, N\Delta_E]$ taking complex values. When i is evaluated at any $j\Delta_E, j = 1, 2, \cdots, N$, this is to be interpreted as the current injections into the system by customer j. Finally, we denote as $\vec{\phi}$ the function defining the phase of the current injection i. That is, if the current injection at position x' is on phase An, then $\phi(x') = \begin{bmatrix} 1 & 0 & 0 & -1 \end{bmatrix}^T$ whereas for phase AB we would have $\phi(x') = \begin{bmatrix} 1 & -1 & 0 & 0 \end{bmatrix}^T$.

Now, denoting $\vec{V}(x) \triangleq \begin{bmatrix} V_A(x) & V_B(x) & V_C(x) & V_n(x) \end{bmatrix}^T$, we have:

$$\vec{V}(x) \approx \vec{V}(0) - z \int_0^x x' i(x') \vec{\phi}(x') d\mu(x') \quad (17)$$

Now, the voltage measured at x by a meter, $\tilde{V}(x)$, is given by $\langle \vec{\phi}(x), \vec{V}(x) \rangle$. Thus:

$$\tilde{V}(x = k\Delta_E)$$
$$\approx \langle \vec{\phi}(x), \vec{V}(0) \rangle - z \int_0^x x' i(x') \langle \vec{\phi}(x), \vec{\phi}(x') \rangle d\mu(x') \quad (18)$$
$$= \langle \vec{\phi}(x), \vec{V}(0) \rangle - z\Delta_E \sum_{j \leq k} j \cdot i(j\Delta_E) \langle \vec{\phi}(x), \vec{\phi}(j\Delta_E) \rangle$$

In which we already see our intuition pop out: the customers contributing most to the voltage measurement at x are the customers whose $\vec{\phi}$ vectors have the largest inner product with $\vec{\phi}(x)$, i.e., the customers who share phase lines with the customer at x.

Now, (18) refers to phasor quantities. However, smart meters typically only return time averaged voltage magnitudes. We can take this into account by modifying each $\vec{\phi}$ vector in correspondence with the assumption that the circuit is fairly balanced. As such, we imagine lumping all of the single phase customers into one large wye-connected load and all of the two phase customers into one large delta-connected load. We consider each customer's current injections as contributing to current injections on these loads. Then the effect of all of the A injections, for example, is to drop the voltage magnitude V_A along the primary feeder, but not effect the neutral line at all. Thus, the $\vec{\phi}$ corresponding to phase A should be modified to $\tilde{\phi} = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}$. Similarly, the $\vec{\phi}$ corresponding to phase AB should be modified to $\tilde{\phi} = \begin{bmatrix} \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} & 0 & 0 \end{bmatrix}$ to account

for the transformation from a line-line current magnitude into a line-neutral current magnitude. With all of the above approximations, and denoting as $\bar{V}$ the time-averaged version of $\tilde{V}$ we have

$$\bar{V}(k\Delta_E) \approx g(k\Delta_E) - |z|\Delta_E \sum_{j \leq k} j \cdot |i(j\Delta_E)|c_{k,j} \quad (19)$$

where $g(k\Delta_E)$ is 1 if customer k is single phase and $\sqrt{3}$ if customer k is two phase. $c_{k,j}$ is an inner product with $\frac{1}{\sqrt{3}} \leq |c_{k,j}| \leq \frac{2}{\sqrt{3}}$. We can tighten the lower bound on the diagonal such coefficients as $1 \leq c_{i,i}$.

Collecting all $\bar{V}(k\Delta_E)$ together into one vector $\bar{\mathbf{V}}$, we have:

$$\bar{\mathbf{V}} = \mathbf{g} - |z|\Delta_E \mathbf{CJI} \quad (20)$$

where $\mathbf{g}$ collects the g functions over the customers, $\mathbf{I}$ collects current magnitudes over each customer, $\mathbf{J} = diag(1, 2, \cdots, N)$, and

$$\mathbf{C} = \begin{bmatrix} c_{11} & 0 & 0 & \cdots & 0 \\ c_{21} & c_{22} & 0 & \cdots & 0 \\ c_{31} & c_{32} & c_{33} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ c_{N1} & c_{N2} & c_{N3} & \cdots & c_{NN} \end{bmatrix} \quad (21)$$

Thus the statistics of $\bar{\mathbf{V}}$ are inherited linearly from those of $\mathbf{I}$. Assuming that I is an iid Gaussian distribution with covariance σ^2, this yields a multivariate Gaussian distribution for $\bar{\mathbf{V}}$. The covariance of $\bar{\mathbf{V}}$ is given by $\Sigma = |z|^2 \Delta_E^2 \sigma^2 (\mathbf{CJ})(\mathbf{CJ})^T$. We note that this iid Gaussian assumption is not prohibitive when our goal is in showing that our entropy is low; iid Gaussian random variables *maximize* entropy at a fixed variance level. Thus, if the iid Gaussian assumption of $\mathbf{I}$ does not hold, then the entropy of the dataset will be even lower than what we predict.

C. Entropy Analysis

The above approximate derivation allows us to analyze the entropic properties of smart-meter voltage measurements.

Lemma 4. *For a distribution system following the model of subsection V-B, we have:*

$$ln\left(\frac{e}{N}\right) - \frac{1}{N}ln\left(\frac{\sqrt{\frac{2\pi}{e^2}}}{(N+1)}\right) \leq \frac{H(\bar{\mathbf{V}})}{N}$$

$$\frac{H(\bar{\mathbf{V}})}{N} \leq \frac{1}{2}ln\left(\frac{12e}{N}\right) - \frac{1}{N}ln\left(\frac{\sqrt{\frac{e^2}{4\pi}}}{(N+1)}\right) \quad (22)$$

Proof. We first note the growth rate of the diagonal elements of $\bar{\mathbf{V}}$'s covariance matrix Σ.

$$\Sigma_{11} \propto c_{11}^2 \quad (23)$$
$$\Sigma_{22} \propto c_{21}^2 + 4c_{22}^2 \quad (24)$$
$$\Sigma_{33} \propto c_{31}^2 + 4c_{32}^2 + 9c_{33}^2 \quad (25)$$

And thus we can estimate:

$$\frac{1}{3}\frac{k(k+1)(2k+1)}{6} \leq \frac{\Sigma_{kk}}{\alpha} \leq \frac{4}{3}\frac{k(k+1)(2k+1)}{6} \quad (26)$$

where $\alpha = |z|^2 \Delta_E^2 \sigma^2$.

This estimate is important due to the preprocessing of our dataset. When applying a machine learning algorithm, we typically scale every data-set down such that the diagonal elements of the covariance matrix are equal to 1. This makes the effective covariance $\tilde{\Sigma} = D\Sigma D$ for some matrix D such that the diagonal of $\tilde{\Sigma}$ consists of $1's$.

We can now estimate the entropy of a scaled voltage dataset via the entropy of a Gaussian random variable.

$$H(X) = \frac{1}{2}ln\left((2\pi e)^N det(\tilde{\Sigma})\right) \qquad (27)$$

$$= \frac{1}{2}ln\left((2\pi e)^N det(\alpha DCJ^2 C^T D)\right) \qquad (28)$$

Now, for calculating the determinant portion, we note that

$$det(\alpha DCJ^2 C^T D) = \alpha^N det(D)^2 det(C)^2 det(J)^2 \qquad (29)$$

First, we have $det(J)^2 = N!^2$. Further, we can bound $det(C)$ as:

$$1 \le det(C)^2 \le \left(\frac{4}{3}\right)^N \qquad (30)$$

where the lower bound occurs if every customer is two-phase attached and the upper bound occurs if every customer is single-phase attached. Finally, due to (26) and the scaling property:

$$\frac{9^N}{N!(2N+1)!(N+1)} \le det(\alpha D)^2 \le \frac{36^N}{N!(2N+1)!(N+1)} \qquad (31)$$

Thus

$$\frac{9^N N!}{\alpha^N (3N)!(N+1)} \le det(D)^2 det(J)^2 \le \frac{36^N N!}{\alpha^N (2N)!(N+1)} \qquad (32)$$

Now, by Sterling's approximation, we can further obtain:

$$\frac{\sqrt{\frac{2\pi}{e^2}}\left(\frac{e^2}{\alpha N^2}\right)^N}{(N+1)} \le det(D)^2 det(J)^2 \le \frac{\sqrt{\frac{e^2}{4\pi}}\left(\frac{9e}{\alpha N}\right)^N}{(N+1)} \qquad (33)$$

From which we can finally obtain:

$$Nln\left(\frac{e}{N}\right) - ln\left(\frac{\sqrt{\frac{2\pi}{e^2}}}{(N+1)}\right) \le H(\bar{\mathbf{V}})$$

$$H(\bar{\mathbf{V}}) \le \frac{N}{2}ln\left(\frac{12e}{N}\right) - ln\left(\frac{\sqrt{\frac{e^2}{4\pi}}}{(N+1)}\right) \qquad (34)$$

$\square$

It is easy to see that both the lower bound and the upper bound become quite negative with N. For $N \approx 5000$, these bounds estimate the per-customer entropy to be roughly between -4 and -11 bits. If the smart meters are encoded with 16 bits, then the per-customer entropy that the machine learning algorithm 'sees' is between 5 and 12 bits. Thus we say that this problem is 'inherently low entropy'. We note that the main reason this low entropy occurs is due to the fact that the covariance matrix diagonal values scaled with the distance of the corresponding customer from the substation. Had this

not been the case - i.e., had every customer been of equal uncertainty, then the entropy of the dataset would have been much larger - scaling positively as N grows large.

We obtain a low entropy dataset because the uncertainty of the customer's voltage data is dominated by those customers far from the substation. This will stay true of many problems in networked systems. Since this is the main motivation behind the second technique introduced in this paper (information loading), we conjecture that the technique is highly applicable to problems beyond phase identification.

D. Maximum Mutual Information (MMI) Estimation: Further Evidence of the Low Entropy Feature Space Hypothesis

For the next component of this paper, we will need a way of estimating the amount of mutual information that a given neural network *can* carry about an input variable X. This estimation is the subject of this subsection.

We have from information theory that for any random variable U,

$$H(X) = I(X;U) + H(X|U) \qquad (35)$$

We assume that $H(X) > 0$ and suppose we have a class of conditional distributions $\mathcal{Q}$ from which to search for a variable U meant to approximate X (e.g. the hidden layer in our neural networks). Then if $\mathcal{Q}_{U|X}$ is large enough, it will contain a subset $\mathcal{Q}'$ such that $H(X|U') > 0$ for all U' whose joint distribution follows $p(x)q'(u|x)$ for some $q'(u|x) \in \mathcal{Q}'$. Then for such U', $H(X) \ge I(X;U)$, and we have equality when $H(X|U) = 0$. This motivates the estimator

$$H(X) \ge \sup_{q(u|x)\in\mathcal{Q}} I(X;U) \qquad (36)$$

whose optimum will not occur in $\mathcal{Q}'^c$. Thus we can then adapt the MINE-f mutual information estimator [25] to obtain

$$H(X) \ge \sup_{q(u|x)\in\mathcal{Q}_{U|X}} \sup_{t\in\mathcal{F}} \int t d\mathbb{P}_{XU} - \int e^{t-1} d\left(\mathbb{P}_X \otimes \mathbb{P}_U\right) \qquad (37)$$

where $\mathcal{F}$ is another space of functions. From here, we can simply define $\mathcal{Q}$ and $\mathcal{F}$ as spaces of functions parameterized by deep neural networks with fixed hyper-parameters. Since (37) will only have equality for very large $\mathcal{Q}_{U|X}$, we will give the resulting right hand side a separate name - the *maximum mutual information* (MMI) of $\mathcal{Q}_{U|X}$. This is a number which depends on the input feature space $\mathcal{X}$ and on the space $\mathcal{Q}_{U|X}$, and is the desired amount of mutual information that a given network (with fixed hyper-parameters) can carry about the input variable.

We have experimentally conducted these MMI estimations on five real circuits with $\mathcal{F}$ defined by a feed-forward neural network with a single hidden layer of 1000 units. We have observed a maximum MMI of just 8.67 bits, and an average of just 6.21 bits. For reference, these MMIs are much lower than that of the MNIST dataset, which has an MMI of about 21 bits [26].

E. Information Loading

Let's quickly review what has been seen so far in this paper. In subsection III-A, we have the inequality $I_{loss}(S) \leq 2\delta(\hat{p})I(X;Z) + h_2(\delta(\hat{p}))$. First, we used training data selection to reduce the term $\delta(\hat{p})$. We will now consider the other important term contributing to $I_{loss}(S)$: $I(X;Z)$. At first glance, it would appear that *reducing* $I(X;Z)$ would be a pertinent goal. Indeed this is the premise behind the information bottleneck method [17]. However, there is a hidden trade-off here. This is because reducing $I(X;Z)$ leads to its own form of information loss through the strong data processing inequality:

$$I(Y;Z) \leq \eta I(X;Z), \ \eta \leq 1 \quad (38)$$

Thus we must balance the loss reduction in (1) against these additional losses. Furthermore, since we are already reducing $\delta(\hat{p})$ by using the data selection frameworks above, the marginal loss reduction that can be achieved from reducing $I(X,Z)$ is reduces as well. We thus conjecture that, unless $I(X;Z)$ is large, the losses from the data processing inequality will win out. But voltage data in general has relatively low entropy as we will show. Thus $I(X;Z)$ will be low as well for all random variables Z. Thus, we should attempt to keep $I(X;Z)$ as large as possible.

However, there are losses in $I(X:Z)$ that occur naturally. First, if any stochasticity is introduced to the neural network in use, then we can view the neural network as a lossy channel. Information losses in $I(X;Z)$ will occur as a result. This can be alleviated by just not using a stochastic network, but there is a second form of $I(X;Z)$ losses that are more critical - finite data information losses. These losses come from the fact that, even though X is instantiated for every customer, our neural classifier only sees x instantiations that are accompanied by y labels. This artificially limits the amount of X data that is seen, yielding the losses in $I(X;Z)$.

But this need not be the case. If we can write an estimator of $I(X;Z)$ as a function of neural network parameters, then we can simply add in an information 'anti-regularization' term to whatever loss function we are using. That is, if $\mathcal{L}$ is our current supervised loss function (say cross-entropy), then we can modify this to

$$\mathcal{L} - \beta I(X;Z), \ \beta > 0 \quad (39)$$

effectively performing the *opposite* of the information bottleneck method.

We have such an estimator introduced in this paper already: MINE-f from subsection V-D. Plugging this estimator into (39) will yield the desired result. See figure 2.

VI. Experiments

A. Data

This analysis will be performed over 5 circuits of varying complexity from Southern California Edison, Pacific Gas and Electric Company, and FortisBC. The details of these circuits are contained in the Table I. Where Degree of Balance is measured by the average current coming back on the neutral line in the distribution circuit. The obtained values are partitioned
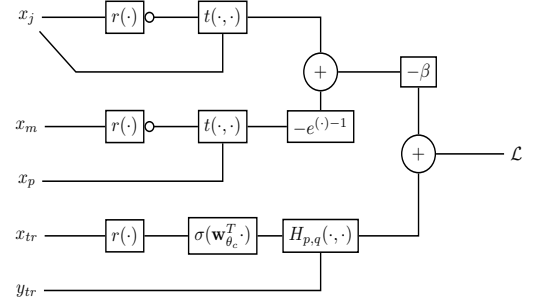

Fig. 2: String diagram representation of the information loading forward pass. $r : \mathcal{X} \rightarrow \mathcal{Z}$ is the representation function. $t : \mathcal{X} \times \mathcal{Z} \rightarrow \mathbb{R}$ is the argument of the information estimator.

TABLE I: Distribution Circuits Characteristics

Name	$N_{consumers}$	Phase Connections	Degree of Balance
I	1892	A, B, C, ABC	Low
II	3166	All	Low
III	4629	A, B, C, ABC	High
IV	3638	B, C, AB, BC, CA	High
V	1773	A, B, C	Low

into two equal-probability bins which we denote as 'Low' and 'High'.

Each circuit contains 31 days of voltage magnitude data, sampled hourly for a feature vector of dimension 744. All experiments are performed with 5% of the total customers used as training data. Every reported number is an average over ten trials.

Empirically, circuits with more potential phase connections (e.g. A, B, C, AB, BC, CA vs. just A, B and C) typically have lower Phase Identification accuracy. This is firstly due to the fact that the difficulty of a classification task is related to the number of classes, but also due to the fact that there are nontrivial dependencies between some of these classes; for example, transformers of the AB class take current from the A line and send it back along the B line, which complicates the dynamics of transformers attached to just A or B. Balanced circuits also have lower Phase Identification accuracy than unbalanced ones, but the effect is less significant. The more phase connections available and the more balanced the circuit is, the more 'difficult' that circuit is to identify.

B. Preprocessing

In many distribution systems, center tapped transformers are abundant. As such, voltage magnitude data will come in two bulk clusters, one near $120V$ and one near $240V$. This distinction has little relevance for the phase connection type of the corresponding customer, and will add instability into any supervised learning algorithm. We take care of this by *self*-normalizing each voltage time series such that its time-average is equal to 1.0. We follow this step with the standard preprocessing technique of *batch*-normalizing the data to have an batch-mean of 0.0 and a batch-standard deviation of 1.0.

C. Results

We first desired to establish some baseline accuracies for the phase identification problem using standard supervised learning

approaches. The results of this analysis are shown in Table II. We see that a two layer neural network with 500 hidden units outperforms the other methods in 4 out of 5 cases. Only on circuit II is the neural network beaten, and barely so. Thus we decided to implement our changes on this classifier.

We tested our proposed techniques in all permutations. These are all shown in Table III. The first column of this table repeats the accuracy of the baseline two layer neural network from figure II. The second column considers information loading in isolation. This yields minor to substantial changes. The effect is highly dependent on the circuit. In general, this technique seems to yield larger improvements to harder circuits. We then tested, training data selection under the facility location method [27]. This yields substantial improvement in every case. We observe in the fourth column that the inverse-matrix heuristic outperformed the facility location method in every case. Finally, we performed both the information loading technique and the inverse-schur heuristic to obtain the rightmost column in the table. These combined techniques yield the highest phase identification accuracies in every case. They are phenomenally improved from the baseline results.

Circuit	Neighbors	Decision Tree	Random Forest	Neural (2-layer)
I	74.6%	71.5%	68.2%	**80.7%**
II	**64.8%**	59.3%	39.3%	64.7%
III	70.6%	59.2%	59.4%	**71.4%**
IV	67.2%	59.3%	51.8%	**75.0%**
V	41.0%	50.00%	37.00%	**51.7%**

TABLE II: Baseline Establishment

Circuit	Baseline	I-loading	Facility	Inverse	Inverse+I-loading
I	80.7%	81.5%	86.7%	89.9%	**91.0%**
II	64.7%	80.6%	90.5%	95.4%	**96.3%**
III	74.1%	75.2%	90.6%	91.5%	**93.1%**
IV	75.0%	78.0%	91.2%	94.6%	**98.8%**
V	51.7%	59.1%	94.2%	96.1%	**97.3%**

TABLE III: Proposed Techniques

Circuit	Correlation	Clustering	Proposed
I	37.8%	75.1%	**91.0%**
II	34.1%	56.4%	**96.3%**
III	46.4 %	65.7%	**93.1%**
IV	40.1%	53.6%	**98.8%**
V	38.4%	38.4%	**97.3%**

TABLE IV: Accuracy comparisons between the literature and the proposed method.

We've also compared the accuracies obtained by our proposed method to two of the methods of phase identification in literature. Both of these methods lie on the physical-intuition side of techniques, in contrast to the more abstract off-the-shelf machine learning techniques. The first, which we've denoted 'correlation', slightly modifies the correlation based methods of references [8]–[12] by computing the empirical voltage correlation matrix over the customers, and using this matrix to link the customers together under complete-linkage.
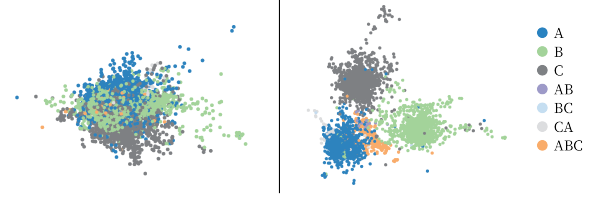

Fig. 3: Learned representations on circuit II. (Left) random selection with no information loading. (Right) Targeted selection with information loading.
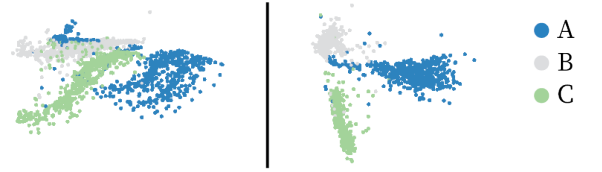

Fig. 4: Learned representations on circuit V. (Left) random selection with no information loading. (Right) Targeted selection with information loading.
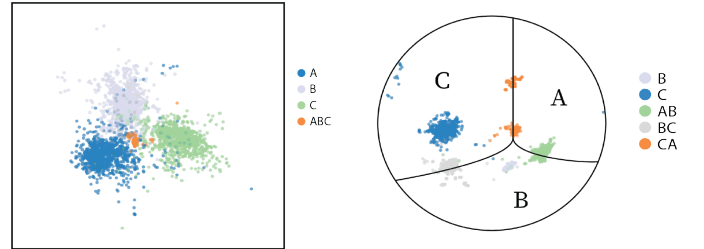

Fig. 5: Learned representations on circuits I (left) and IV (right) with both techniques implemented.

The second, which we've denoted 'clustering' is equivalent to that of reference [14]. We see that the literature is capable of achieving similar accuracies to the more abstract machine learning algorithms in Table II, but without requiring training data. However, our proposed method, which synergistically combines both physical intuition and the more abstract learning algorithms, beats both types of methods individually, yielding the best results on these datasets by a large margin.

Since the central idea of this paper lies in learning better representations, we ought to visualize what the representations have learned. This has been done for two of the circuits, and is visualized in figures 3 and 4. Each point on these plots corresponds to an averaged sample from the distribution $p(z_i|x_i)$ averaged over 50 trials. Each of these averages have dimension 500. We reduced this to 2 dimensions by performing PCA and projecting each point onto the first two principal components.

We observe that the representation has indeed learned much better representations by using these techniques. In the baseline case, there is little separation between the classes - especially in the case of circuit II. By implementing the proposed techniques, we see more class separation in these variables. Furthermore, we observe some learning of the relationships between the classes. For example, in circuit II, the learned representation has placed ABC directly in the middle of the classes A B and C. This is significant since, by subsection V-B, phase ABC truly is a 'combination' of phases A, B, and C when it comes to voltage data. In the case of circuit V, we have learned even more. In the final representation, we see three branches of representation data - one for each class. The distance from the central node of these representations corresponds to distance along the primary feeder of the distribution circuit.

Finally, we have plotted the final representations under both techniques for circuits I, and IV in figure 5. The behavior of circuit III is nearly identical to circuit I, so we are not presenting its plot for the sake of space. In the cases of circuits I and III, we see similar behavior. Independent An Bn and Cn clusters have arisen, and an ABC cluster has appeared in the middle. This is, again, what we expect from circuits consisting of those phase types. Circuit IV is a bit more interesting because it consists of phase types Bn, Cn, AB, BC and CA, with Bn having very little representation. Clusters of each phase type have appeared with decent separation, so classification will at least be easy. More interestingly, however, is the fact that the AB cluster has appeared opposite to the Cn cluster, CA opposite Bn. Assuming the location of the non-existent An cluster to the top right, these positions make much sense.

VII. Conclusion

This paper has used the theory of information losses to propose the application of two novel techniques - inverse schur data selection and information loading - to the phase identification problem. These techniques have synergistically combined the the abstract, problem agnostic, supervised learning techniques found in machine learning research with the physical intuitions that are often employed in more specific power systems projects. As such, we observe substantial improvements in phase identification accuracy over both the purely abstract methods and those methods which only base themselves on the physical intuitions. Furthermore, we have observed that the representations learned upon using these techniques are much more meaningful than the baseline representations, giving us highly interpret-able results which are often not found in techniques which use abstract machine learning algorithms alone. We have argued that these techniques generalize well beyond just phase identification, and have listed properties for which these techniques will be helpful.

References

[1] C.-S. Chen, T.-T. Ku, and C.-H. Lin, "Design of phase identification system to support three-phase loading balance of distribution feeders," *IEEE Transactions on Industry Applications*, vol. 48, no. 1, pp. 191–198, 2012.

[2] V. Arya, V. Chakaravarthy, K. Dontas, S. Hoy, J. Kalagnanam, S. Kalyanaraman, C. Pavlovski, and D. Seetharamakrishnan, "Systems and methods for phase identification," Patent US8 825 416.

[3] 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, 2015, pp. 1–5.

[4] K. J. Caird, "Meter phase identification," Mar. 27 2012, uS Patent 8,143,879.

[5] M. Dilek, R. P. Broadwater, and R. Sequin, "Phase prediction in distribution systems," in *Power Engineering Society Winter Meeting, 2002. IEEE*, vol. 2, 2002, pp. 985–990 vol.2.

[6] and A. Rajeswaran, N. P. Bhatt, and R. Pasumarthy, "A novel approach for phase identification in smart grids using graph theory and principal component analysis," in *2016 American Control Conference (ACC)*, July 2016, pp. 5026–5031.

[7] V. Arya, D. Seetharam, S. Kalyanaraman, K. Dontas, C. Pavlovski, S. Hoy, and J. R. Kalagnanam, "Phase identification in smart grids," in *Smart Grid Communications (SmartGridComm), 2011 IEEE International Conference on*, Oct 2011, pp. 25–30.

[8] 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*, 2018.

[9] H. Pezeshki and P. J. Wolfs, "Consumer phase identification in a three phase unbalanced LV distribution network," in *2012 3rd IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe)*, Oct 2012, pp. 1–7.

[10] B. K. Seal and M. F. McGranaghan, "Automatic identification of service phase for electric utility customers," in *2011 IEEE Power and Energy Society General Meeting*, July 2011, pp. 1–3.

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

[12] 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, Sep. 2018.

[13] 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, June 2013.

[14] 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)*, Dec 2016, pp. 259–265.

[15] F. Ni, J. Q. Liu, F. Wei, C. D. Zhu, and S. X. Xie, "Phase identification in distribution systems by data mining methods," in *2017 IEEE Conference on Energy Internet and Energy System Integration (EI2)*, Nov 2017, pp. 1–6.

[16] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[17] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[18] W. H. Hsu, L. S. Kennedy, and S.-F. Chang, "Video search reranking via information bottleneck principle," in *Proceedings of the 14th ACM international conference on Multimedia*. ACM, 2006, pp. 35–44.

[19] N. Slonim and N. Tishby, "Agglomerative information bottleneck," in *Advances in neural information processing systems*, 2000, pp. 617–623.

[20] O. Shamir, S. Sabato, and N. Tishby, "Learning and generalization with the information bottleneck," in *International Conference on Algorithmic Learning Theory*. Springer, 2008, pp. 92–107.

[21] B. Foggo, N. Yu, J. Shi, and Y. Gao, "Asymptotic Finite Sample Information Losses in Neural Classifiers," *arXiv e-prints*, p. arXiv:1902.05991, Feb 2019.

[22] B. Settles, "Active learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 6, no. 1, pp. 1–114, 2012.

[23] S. Tong, *Active learning: theory and applications*. Stanford University USA, 2001, vol. 1.

[24] B. Foggo and N. Yu, "Interpreting active learning methods through information losses," *arXiv preprint arXiv:1902.09602*, 2019.

[25] I. Belghazi, S. Rajeswar, A. Baratin, R. D. Hjelm, and A. Courville, "Mine: mutual information neural estimation," *arXiv preprint arXiv:1801.04062*, 2018.

[26] O. Rippel and R. P. Adams, "High-dimensional probability estimation with deep density models," *arXiv preprint arXiv:1302.5125*, 2013.

[27] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," *arXiv preprint arXiv:1708.00489*, 2018.