Overview of Data-driven Power Flow Linearization

Conference Paper

Author(s):
Jia, Mengshuo; Hug, Gabriela

Publication date:
2023

Permanent link:
https://doi.org/10.3929/ethz-b-000601298

Originally published in:
https://doi.org/10.1109/PowerTech55446.2023.10202779

Overview of Data-driven Power Flow Linearization

Mengshuo Jia
Power Systems Lab
ETH Zürich
Zürich, Switzerland
0000-0002-2027-5314

Gabriela Hug
Power Systems Lab
ETH Zürich
Zürich, Switzerland
0000-0002-4312-616X

***Abstract*—The accuracy limitation of physics-driven power flow linearization approaches and the widespread deployment of advanced metering infrastructure render data-driven power flow linearization (DPFL) methods a valuable alternative. While DPFL is still an emerging research topic, substantial studies have already been carried out in this area. However, a comprehensive overview and comparison of the available DPFL approaches are missing in the existing literature. This paper intends to close this gap and, therefore, provides a narrative overview of the current DPFL research. Both the challenges (including data-related and power-system-related issues) and methodologies (namely regression-based and tailored approaches) in DPFL studies are surveyed in this paper; numerous future research directions of DPFL analysis are discussed and summarized as well.**

***Index Terms*—Power flow linearization, data-driven, machine learning, regression, programming**

I. Introduction

LINEARIZED power flow models are broadly utilized in today's power system operation and analysis [1]–[3]. Traditional power flow linearization methods are mainly physics-driven (a.k.a., model-driven). After decades of development, physics-driven methods have seemingly reached the limit in terms of their accuracy. Due to the widespread deployment of advanced metering infrastructure [4], data-driven power flow linearization (DPFL) approaches have emerged as a valid strategy and drawn increasing attention [5].

DPFL methods use historical steady-state measurements of the system to establish a linear model through data training [6], [7]. The benefits include but are not limited to having (i) higher approximation precision due to being assumption-free [5], [7] and customizable [1], [8], (ii) better applicability to cases where accurate physical parameters are unavailable [9], [10], (iii) implicit integration of power losses [1], [6], [11], (iv) integration of up-to-date system attributes [6], [12], and (v) inclusion of realistic impacts, e.g., control actions [6] or human behaviors [13].

Despite the above advantages, there are a variety of challenges standing in the way of widespread application. These obstacles can be divided into data-related and system-related issues. The former issue includes the problems caused by (i) data multicollinearity [14]–[18], (ii) the presence of outliers [16], [19]–[21], (iii) measurement noises [7], [18], [22], [23], (iv) the temporal correlation among observations [24], and (v) asynchronous data [5], [25]–[27]. The system-related issue refers to the challenges particularly related to power systems, namely (i) the inherent nonlinearity of the physical characteristics in power flows [1], (ii) the lack of consensus on the importance and usage of physical knowledge [1], [7], [28]–[30], (iii) frequent variations in grid topologies [6], [27], [31], (iv) inevitable variations in bus types [6], [32], and (iii) the limited observability of the system [15], [21], [27].

To address the aforementioned issues, numerous DPFL approaches have been designed. They can be classified into two major categories: the classic regression-based methods and more tailored methods. The former type adopts the classic formulations of various regression programming models, while the latter reorganizes the classic regression model by customizing the objective functions and/or constraints. Specifically, the regression DPFL approaches include (i) least squares regression and its variants [1], [3], [5], [7], [11], [14], [21], [24], [25], [27], [33]–[37], (ii) partial least squares and its variants [6], [11], [26], [38], (iii) ridge regression and its variants [15], [18], [31], and (iv) support vector regression and its variants [11], [21], [36], [39], [40]. The tailored DPFL methods consist of (i) linearly constrained programming [7], [30], (ii) chance-constrained programming [41], and (iii) distributionally robust chance-constrained programming [42].

Although the research on DPFL has been extensive and profound to date, the challenges faced by DPFL and the corresponding solutions have not been thoroughly deliberated and agreed upon in the existing literature. Future research directions are also not comprehensively summarized. In a nutshell, DPFL, as an emerging but blooming area, has not been reviewed yet, nor discussed in any previous survey study. Due to this fact, this paper contributes in the following aspects:

- Delivering a comprehensive review of DPFL, including the obstacles and the corresponding methodologies, providing an important reference point for future research on this topic.
- Discussing open research questions for future directions of DPFL studies.

In the following, Section II summarizes the challenges in DPFL studies, while Section III reviews existing DPFL approaches. Future research directions of DPFL are elaborated on in Section IV. Section V concludes this paper.

II. Challenge

DPFL faces a variety of challenges. These issues are summarized in this section.

A. Data-related Challenges

1) Data Multicollinearity: Data multicollinearity (a.k.a., data coherence [17] or data isotropic dispensation [18]) means that the measurements of variables are highly linearly correlated [14]–[16], [26], [34]. The corresponding training datasets are therefore ill-conditioned or even singular [16]. In power systems, the voltages of directly connected nodes [15], [16], [39] tend to rise and fall simultaneously [15], [16], [39]; a similar observation can be made for the active power injections. Hence, the multicollinearity issue is unavoidable in measurements of the power grid state. Ignoring this issue will result in severe negative impacts on the training result, such as significant error, chaos effect, sample-dependent performance, and overfitting [14], [34].

2) Data Outliers: Data outliers (a.k.a., "bad data" [19]) refer to the observations that deviate significantly from others. This data quality issue, caused by measurement variability or error, is a common issue in power systems [21]. If not handled properly, data outliers can severely degrade the performance of the resulting DPFL models [43].

3) Data Noise: Measurements from PMU [22] and SCADA [23] systems are inevitably contaminated with noise [7], [18]. Noise will undoubtedly harm the training process, leading to problematic DPFL models [18]. Note that repeated training cannot always eliminate the impact of noise, particularly when the noise does not have a non-zero mean.

4) Temporal Correlation in Data: The measurements of power flows are, in fact, time series data [44]. For time series data, realizations closer in time have higher correlations. In other words, the observations of a power system within a time window are not independent and not identically distributed (i.i.d) — this violates the prerequisites of many standard data-training algorithms. Dealing with the temporal correlation significantly complicates the data-driven linearization process.

5) Asynchronous Data: Power flow measurements arrive asynchronously and dynamically vary over time. Consequently, it is challenging to find the most suitable training dataset that ideally reflects the status of interest, especially when the operating point changes frequently. In addition, the obtained DPFL model should be continuously updated based on the latest data available [6]. Achieving an efficient but also accurate update for DPFL models is therefore challenging.

B. System-related Challenges

1) Inherent Nonlinearity: While the AC power flow models for many systems indeed show certain linearity according to existing studies, high nonlinearity has been observed even in small-scale systems [1]. It seems impossible to use a linearized hyperplane to match a highly curved hypersurface accurately. Hence, a challenging question is how to ensure the accuracy of DPFL when the target system is highly nonlinear.

2) Incorporation of Physical Knowledge: In order to improve the performance of data-driven models, it is possible to incorporate accessible physical knowledge of the system into the data-driven training process. Although according to [7], [28], [29], the integration of partial physical knowledge seems to promote the accuracy of the DPFL model, the level of the added benefit is debatable because some other studies claim that physical knowledge is not of added value for the model accuracy as long as the training data are sufficient [1], [30]. Furthermore, the lack of a systemic approach to integrating physical knowledge additionally hinders leveraging this additional information.

3) Grid Topology Variation: Lines in a power grid may switch quite often [6], especially in the distribution grid [31]. Accordingly, the measurements within a time window might correspond to different topologies rather than the topology of interest. Differentiating between data belonging to different topologies is generally a difficult task, especially when topology information is unavailable.

4) Bus Type Variation: Bus-type changes are not uncommon in power flow calculations, e.g., some PV buses must transform into PQ buses due to their reactive power output limits [32]. Such variations, however, might render the obtained DPFL model invalid. This is because the DPFL model is usually trained according to a fixed assignment of known/unknown variables. Yet, bus-type changes will alter the assignment of known/unknown variables. As a result, the DPFL model obtained previously is not applicable any more.

5) Limited Observability: Not all parts of the power system are well-observed [15], [21], [27]. E.g., only the measurements at substations and end users are obtainable in some distribution grids [21]. When only part of a system is being monitored, building a DPFL model of the whole system without the complete set of measurements of all buses is challenging.

III. Methodology

Existing DPFL methods, including regression algorithms and tailored algorithms, address part of the aforementioned challenges. In the following, we use the term "independent variable" to denote the known quantities in power flow calculations while using "dependent variable" to represent the unknown quantities (i.e., the power flow calculation results).

A. Regression Methods

To date, regression methods used in DPFL studies include least squares regression, partial least squares regression, ridge regression, support vector regression, and their variants. Bayesian regression has been tried only in [6] but concluded to be unsatisfying. Hence, this regression method is excluded from this paper.

1) Least Squares and Its Variants: Among the least squares regression approaches, the *ordinary least squares* method is most frequently used and has been adopted in a variety of DPFL studies [3], [11], [21], [27], [33], [34]. The ordinary least squares model is an unconstrained programming model that aims to minimize the sum of squared residuals. Notably, the solution, i.e., the estimation of the coefficients in the target DPFL model, can be explicitly expressed using the inverse of the Gramian matrix of the dataset [21]. However, the ordinary least squares method suffers from multiple drawbacks, which we list below. In this discussion, we also provide solutions that have been proposed in the literature.

- **Data Multicollinearity**: In the ordinary least squares method, the training dataset of the independent variables must be full column rank [27]. In other words, the dataset must not exhibit multicollinearity. This condition, however, is generally not satisfied for electrical measurements, as explained in Section II-A1. Correspondingly, the Gramian matrix of the dataset is usually non-invertible, and therefore, the solution of the ordinary least squares is unattainable either. To address this issue, two ideas have been leveraged in DPFL studies. The first one is to compute the inverse of the singular Gramian matrix not directly but using the *Moore–Penrose inverse* [14], *complete orthogonal decomposition* [1], [35], or *singular value decomposition* [27], thus yielding three variant approaches of the least squares. The second idea is based on removing the correlation within the dataset by using *Principal component analysis* [36]. Note that all the variant approaches aforementioned have closed-form solutions.
- **Data Outliers**: The ordinary least squares method is sensitive to data outliers. The main reason is that the squared residual used in the objective function highly emphasizes the observations with large residuals, i.e., outliers. The training process is thus dominated by these outliers, degrading the performance of the ordinary least squares accordingly. To handle this problem, the *Huber loss function* has been introduced to the ordinary least squares model in the DPFL analysis [5], [25]. By replacing the squared residual of the outlier with its absolute residual, the Huber loss function weakens the impact of bad data. The resulting programming model can be

transformed into an equivalent convex problem and therefore solved efficiently [25], [45].

- **Data Noise**: The ordinary least squares method implicitly assumes that only the observations of the dependent variables are noisy. However, the observations of the independent variables are also measured data, which contain noise as well. To avoid this strong assumption, the *total least squares* algorithm has been adopted in [7], such that the noises from different measurement devices are weighted unequally. The resulting problem can then be solved by converting the original model into several approximated linearly constrained quadratic optimization problems [7].
- **Temporal Correlation**: The ordinary least squares method can generate the best linear unbiased estimation only when the measurements have homoskedasticity and no autocorrelation. That is, the observations for training must be i.i.d. This condition might not hold in practice, as discussed in Section II-A4. To overcome this issue, the *generalized least squares* method has been used in [24]. The potential correlation between data points is taken into account via the conditional variance-covariance matrix of the residuals; this matrix is obtained empirically [24]. Note that the generalized least squares method also has a closed-form solution.
- **Inherent Nonlinearity**: The ordinary least squares algorithm can only generate a single, global linear power flow model. This model might not be valid for a wide range of different operating modes. The power flow model under various operating points is a hypersurface, which often cannot be approximated by a unified linear hyperplane. To solve this issue, clustering algorithms such as K-means and Gaussian mixture model are integrated into the ordinary least squares model [37], i.e., carrying out the ordinary least squares fitting for each clustered sub-dataset. The resulting *clustering-based least squares* method therefore generates a piecewise linear model, which can better capture the nonlinearity of power flows [37].

2) Partial Least Squares and Its Variants: Owing to the ability to deal with multicollinearity, partial least squares approaches are widespread in DPFL studies. Similar to the principal-component-analysis-based least squares method, the *ordinary partial least squares* method also attempts to eliminate the correlation in the ill-conditioned dataset [6], [11], [38]. To this end, it projects the datasets of both dependent and independent variables onto lower-dimensional spaces constituted by their orthogonal score vectors. Since the resulting vectors are uncorrelated, the collinear situation has been removed. The most classic method is the nonlinear iterative partial least squares (NIPALS) method proposed in [46], which has been used in several DPFL studies such as [6], [26]. Various improvements toward this approach have been made subsequently and leveraged in DPFL studies, aiming to handle the following issues:

- **Computational Burden**: The NIPALS method suffers from a high computational burden. To speed up the computation of NIPALS and meanwhile lower its memory requirements, [47] presents the *SIMPLS* algorithm, which has been applied in [38] in the context of DPFL. Note that the SIMPLS approach, as one of the most commonly used algorithms in partial least squares analysis, is already available in commercial software such as MATLAB (i.e., the built-in function $\mathrm{plsregress}[\cdot]$).
- **Data Outliers**: While the SIMPLS approach enjoys high computational efficiency, it is sensitive to bad data. To deal with data outliers, a robust version of the SIMPLS method has been developed in [48]; the resulting *robust SIMPLS* (RSIMPLS) is subsequently used in [33] to fit linear power flow models.
- **Asynchronous Data**: Neither SIMPLS nor RSIMPLS can update the DPFL model incrementally — they have to retrain the model to account for new data, which is burdensome. To continually and efficiently update the DPFL model as new observations arrive, a *recursive partial least squares* approach has been proposed in [26]. In this approach, the matrix consisting of previous observations is equivalently replaced with a decomposed matrix with much fewer rows, thereby reducing the corresponding calculation time. The above substitution process is recursive such that new data can be integrated as they emerge.

3) Ridge Regression and Its Variants: The ridge regression approaches can also address the multicollinearity issue, thereby being broadly used in DPFL studies as well. Different from the partial least squares methods, the *ordinary ridge regression* algorithm directly tunes the singular matrix to enable its invertibility [31]. The tuning is achieved by adding the Tikhonov–Phillips regularization factor [18] to the diagonal elements of the singular matrix [31]. It is noteworthy that the regularization factor inevitably injects bias into the fitting process. The tuning of this factor is therefore crucial and should be carried out via systematic tuning approaches [15], [18], [36], [49], e.g., using the ridge trace method [18]. The ordinary ridge regression method has also been improved in multiple aspects in DPFL studies, attempting to address the following:

- **Asynchronous Data**: The ordinary ridge regression approach treats all historical observations equally. However, as mentioned in Section II-A5, historical observations are measured at different time steps. They possibly correspond to different operating modes of the system rather than the operating point of interest. To distinguish the measurements, a *locally weighted ridge regression* approach is proposed in [18]. This approach only focuses on the power flow model around a specific operating point. To this end, this method places a higher emphasis on the data point closer to the operating point of concern. The corresponding weight matrix is further integrated into the model of the ridge regression method. The solution to the resulting model can be explicitly expressed as demonstrated in [18].
- **Inherent Nonlinearity**: Similar to the ordinary least squares approach, the ordinary ridge regression approach can only build a single, global DPFL model. To better capture the nonlinearity of the original power flow model, a *clustering-based ridge regression* method has been developed in [15], which can establish a piecewise DPFL model. The clustering algorithm used in [15] is the K-plane clustering method. It should be emphasized that different from the clustering-based least squares method mentioned in Section III-A1, in the clustering-based ridge regression method, the regression is embedded into each iteration of the clustering procedure, i.e., the solution is found iteratively [15].

4) Support Vector Regression and Its Variants: It is well-known that support vector regression approaches can overcome bad data and multicollinearity issues. They are therefore utilized in DPFL studies as well. Compared to the regression approaches reviewed above, the *ordinary support vector regression* algorithm includes two main differences [11], [39]. First, this algorithm replaces all the squared residual terms in the objective function with the absolute residuals. Accordingly, the objective is less sensitive to data outliers, as explained in Section III-A1. Second, by leveraging the error-insensitive function, it only takes the absolute residuals greater than a preset threshold into account. The other residuals, as well as the corresponding data points, are removed from the training process. The above two changes significantly relax the emphasis on the large residuals, thereby rendering the ordinary support vector regression approach more robust to data outliers [21], [39], [40]. The programming model derived from this approach is typically solved by the sequential minimal optimization algorithm [50]. The support vector regression approach can be further generalized to handle the following issue:

- **Inherent Nonlinearity**: To better manage the inherent nonlinearity of the original power flow model, projecting variables into other spaces can resolve or reduce this issue, as in a new space, the projected variables might enjoy better linearity among them. It is possible, however, that the dimension of the new space is unacceptably large due to the permutation of numerous variables in power systems. In that case, the calculation over the new space could be unacceptably burdensome as well. To solve this issue, the *kernel-based support vector regression* method has been developed and subsequently used in DPFL analysis [21], [40]. In this approach, the projection of variables is carried out using kernel functions selected from the reproducing Hilbert kernel space. These particular functions guarantee that the high-dimensional calculation, e.g., the inner product computation, can be equivalently achieved via low-dimensional, low-cost calculation [21], [40]. As a result, the model of the kernel-based support vector regression can still be solved using the sequential minimal optimization algorithm [51]. Note that the DPFL model derived from this approach is no longer a linear model of the original dependent and independent variables, but rather a linear model of the variables after projection.

B. Tailored Methods

The tailored approaches are derived from adding or modifying the objectives/constraints used in the regression algorithms. These adaptations can be generally applied to various regression approaches. According to the type of the programming model used in the method, existing tailored DPFL algorithms can be divided into the following three categories:

1) Linearly Constrained Programming: The goal of the linearly constrained programming approaches is to *integrate accessible physical knowledge* into the data-driven training process. The integration is achieved by introducing additional constraints into the regression models [7], [30]. These constraints, i.e., the (i) bound constraint [30], (ii) structure constraint [7], and (iii) coupling constraint [30], attempt to enforce physical restrictions on the estimation of the coefficients in the target DPFL model.

- **Bound Constraint**: In the bound constraint, the upper/lower limits of the coefficients are derived from the first-order Taylor series expansion of the AC power flow model. The expansion points of the Taylor approximations are used as the boundary operational conditions of the target system. Accordingly, the coefficients obtained from the Taylor linear approximations are considered the physical limits [28], [30].
- **Structure Constraint**: Since coefficients obtained from DPFL and Taylor approximations share a certain similarity, the coefficient structure derived from the Taylor expansion can thus guide the estimation of the DPFL coefficients. The coefficients derived from the Taylor approximation constitute the well-known Jacobian matrix. Under some mild assumptions, the Jacobian matrix is symmetric; furthermore, parts of the Jacobian matrix are diagonal. These properties can be interpreted as structure constraints (a.k.a., Jacobian-matrix-guided constraints) to restrict the optimization of the coefficients in the DPFL model [7].
- **Coupling Constraint**: According to the AC power flow model, some coefficients in the DPFL model should be highly related. E.g., in the model of a line flow, the voltage angles only appear in the form of an angle difference, i.e., the angle of the “from” bus minus that of the “to” bus. It can therefore be inferred that the coefficients of these two angles in the DPFL model should be opposites. This can be captured by the *coupling constraints* of specific DPFL coefficients [30].

Note that the above three types of constraints are all linear constraints. Adding them to regression models will not deteriorate their solvability.

2) Chance-constrained Programming: Tuning the hyperparameters in support vector regression approaches is often challenging [41]. Hence, the chance-constrained programming method proposed in [41] aims to *remove the hyperparameter* aforementioned. To this end, this method replaces the tolerance of residuals (where the hyperparameter derives from) in the support vector regression model with multiple single chance constraints. The resulting programming model is then converted into a mixed-integer linear programming problem via the classic big-M method. The converted model can be solved efficiently with the help of commercial solvers such as GUROBI or CPLEX.

3) Distributionally Robust Chance-constrained Programming: The DPFL model is generally obtained by minimizing the sum of residuals. Hence, the obtained model is optimal only in an average sense. That is, this model might still yield notable worst-case errors in some other cases. In response to this issue, [42] introduces distributionally robust chance constraints (DRCCs) to explicitly *restrict the worst-case errors* of the target DPFL model. Depending on the type of the distribution ambiguity set, the DRCC-based model can either be equivalently converted into a semi-definite programming model [52], [53] and solved subsequently by MOSEK [42], or be equivalently converted into a traditional chance-constrained model [54] and then solved by GUROBI [42].

IV. Open Question

The DPFL algorithms reviewed above address many challenges discussed in the previous section. However, numerous open problems still exist, suggesting various future research directions, as elaborated in this section.

A. Inherent Nonlinearity

- **Linearity Pre-evaluation**: Directly establishing a DPFL model of a system with low linearity may lead to a noticeable approximation error. It is thus crucial to be able to evaluate the degree of linearity of the target system beforehand. The pre-evaluation result will guide the subsequent choice of the DPFL method. Note that the degree of nonlinearity does not necessarily increase with the scale of the system [1]; see the numerical results in [14] for example. That is, the system scale is not a valid indicator of nonlinearity. Designing a computational-friendly evaluation method or an easily-accessible indicator for the degree of linearity, therefore, deserves further research.
- **Coordinate Transformation**: To deal with nonlinearity, many DPFL approaches employ coordinate transformations for

variables, i.e., projecting original variables to new spaces. The transformation means include the kernel function [21], [40], as reviewed previously, and some other classic manners, such as voltage squaring [15], [27], [30], [31], [39], [41], voltage-angle coupling [21], [31], [40], and dimension lifting [14]. Yet, it remains unclear which method performs the best. More importantly, there is currently no systematic approach to identify or design a better coordinate transformation method.

- **Applicability**: Generally, the DPFL model resulting from a coordinate transformation is not well suited for optimization or control [5]. This is because the obtained model is no longer a linear function of original variables but of transformed variables with limited or no physical meaning. Similarly, the piecewise models derived from clustering-based DPFL approaches are not well suited for practical applications as well, since these models introduce integer decision variables into optimization or control. How to systematically find a better compromise between model accuracy and applicability is an open but fundamental question.

B. Physical Knowledge

- **Access to Information**: DPFL approaches have adopted different types of physical knowledge, including the boundary operating point [7], topology [30], [41], line admittance [3], [7], [38], structure of the original power flow model [3], [27], [38], etc. The question, however, of how easily such information can be accessed in practice arises. Given that using physical knowledge is, to a certain degree, against the motivation of data-driven approaches, the above question should be clarified in future studies.
- **Clarification on Necessity**: As mentioned earlier, it remains unclear whether integrating physical knowledge can reliably improve model accuracy, especially when there are sufficient training data available. Although [1], [30] observe that the extra benefit brought by physical knowledge decreases with the increase of the training data size, each study has only verified one type of knowledge using one specific approach for integration. Sophisticated evaluations regarding different types of physical knowledge, diverse integration approaches, and various data sizes are still needed to gain deeper insights into the added value of integrating physical knowledge. Certainly, developing a general integration approach for various types of physical knowledge will also benefit the above exploration.
- **Effectiveness Evaluation**: While different types of physical information are utilized in DPFL approaches already, which kind of knowledge provides the most considerable improvement in model accuracy is still unclear. It is thus necessary to develop an evaluation method to measure the usefulness of various categories of physical knowledge.

C. Grid Topology Variation

- **Theoretical Guarantees**: So far, topological variations are handled in various but rather heuristic ways. E.g, (i) topology enumeration [27], i.e., establishing DPFL model for each possible topology, (ii) topology identification [31], i.e., assuming the highly correlated measurements are from the same topology [55], and (iii) forgetting factor [5], [25]–[27], i.e., weighting recent measurements heavier than the previous ones. Though these practical, heuristic strategies might apply to some cases, their validity is compromised either by the increase in computational burden due to the combinatorial nature of enumerating topologies, or by the lack of a theoretical guarantees. It is thus important to design a better solution backed by a theoretical basis in light of the frequent topological variations in reality.

D. Bus Type Variation

- **Reliable Solution**: Reference [6] proposes a scheme to handle the bus-type changes by bundling known and unknown variables and assigning them elaborately to both sides of the target linear model. While requiring no retraining after bus-type changes, this scheme might result in very large linearization errors. The reason lies in that an additional inverse calculation is required when employing this strategy, whereas the corresponding matrix, comprised of the estimated coefficients, sometimes is near-singular. Note that the invertibility of this matrix has not been theoretically guaranteed by [6] either. It is hence imperative to find a reliable solution to the issue of bus-type variation in the future.

E. Limited Observability

- **Finding a Solution**: So far, existing DPFL approaches can only build a truncated linear power flow model given that not all variables in the system are measured [15], [21], [40]. Note that the situation with zero-power-injected buses is also a special case of the limited observability issue — while these buses are well observed, their measurements, i.e., the zero values, must be removed from the training dataset, and the corresponding variables have to be excluded from the DPFL model as well. How to build a DPFL model for the whole system with only limited observations is still an open question. Finding a solution should be the first step in future studies.

F. Temporal Correlation

- **Correlation Calculation**: As discussed in Section III-A1, the generalized least squares method has been adopted in DPFL studies to take into account the correlation among time series measurements. The key to using this method lies in knowing the correlation among observations beforehand. Note that this correlation refers to the conditional variance-covariance matrix of the residuals [24]. The ground-truth value of this correlation matrix, however, is generally unknown and challenging to compute. This is because residuals are training results instead of measurements. The correlation matrix computed from the training results can hardly be considered true if the actual correlation has not been incorporated into the training process. As a result, a method to reliably compute the correlation matrix of the residuals is required.

G. Asynchronous Data

- **Sample Selection**: As mentioned before, the training dataset should change dynamically with time as new operating modes continually emerge [37]. Simply weighting forgetting factors to measurements could be imprecise. How to dynamically, automatically, and continually identify which samples to keep for the current training, particularly when the operating point changes frequently, should be but has not been answered sufficiently up to now.

V. Conclusion

This paper provides an overview of existing DPFL studies. The current obstacles faced by DPFL methods and the latest developments in the corresponding methodologies are both outlined and discussed. Meanwhile, this paper proposes a variety of research directions for future DPFL studies. Overall, DPFL is an emerging and promising research topic. It provides an enormous opportunity to resolve the accuracy bottleneck of classic power flow linearization methods.

References

[1] Z. Shao, Q. Zhai, J. Wu, and X. Guan, "Data based linear power flow model: Investigation of a least-squares based approximation," *IEEE Trans. Power Syst.*, vol. 36, no. 5, pp. 4246–4258, 2021.

[2] R. Hu, Q. Li, and S. Lei, "Ensemble learning based linear power flow," in *2020 IEEE PESGM*. IEEE, 2020, pp. 1–5.

[3] X. Li and K. Hedman, "Data driven linearized ac power flow model with regression analysis," *arXiv preprint arXiv:1811.09727*, 2018.

[4] V. Terzija, G. Valverde, D. Cai, P. Regulski, V. Madani, J. Fitch, S. Skok, M. M. Begovic, and A. Phadke, "Wide-area monitoring, protection, and control of future electric power networks," *Proc. IEEE*, vol. 99, no. 1, pp. 80–93, 2010.

[5] Y. Liu, Z. Li, and Y. Zhou, "A physics-based and data-driven linear three-phase power flow model for distribution power systems," *arXiv preprint arXiv:2103.10147*, 2021.

[6] Y. Liu, N. Zhang, Y. Wang, J. Yang, and C. Kang, "Data-driven power flow linearization: A regression approach," *IEEE Trans. Smart Grid*, vol. 10, no. 3, pp. 2569–2580, 2018.

[7] Y. Liu, Y. Wang, N. Zhang, D. Lu, and C. Kang, "A data-driven approach to linearize power flow equations considering measurement noise," *IEEE Trans. Smart Grid*, vol. 11, no. 3, pp. 2576–2587, 2019.

[8] J. Yang, N. Zhang, C. Kang, and Q. Xia, "A state-independent linear power flow model with accurate estimation of voltage magnitude," *IEEE Trans. Power Syst.*, vol. 32, no. 5, pp. 3607–3617, 2016.

[9] K. E. Van Horn, A. D. Dominguez-Garcia, and P. W. Sauer, "Measurement-based real-time security-constrained economic dispatch," *IEEE Trans. Power Syst.*, vol. 31, no. 5, pp. 3548–3560, 2015.

[10] S. Frank, J. Sexauer, and S. Mohagheghi, "Temperature-dependent power flow," *IEEE Trans. Power Syst.*, vol. 28, no. 4, pp. 4007–4018, 2013.

[11] C. Qin, L. Wang, Z. Han, J. Zhao, and W. Wang, "A modified data-driven regression model for power flow analysis," in *2019 IEEE 8th DDCLS*. IEEE, 2019, pp. 794–799.

[12] M. Z. Kamh and R. Iravani, "A sequence frame-based distributed slack bus model for energy management of active distribution networks," *IEEE Trans. Smart Grid*, vol. 3, no. 2, pp. 828–836, 2012.

[13] T. Carriere and G. Kariniotakis, "An integrated approach for value-oriented energy forecasting and data-driven decision-making application to renewable energy trading," *IEEE Trans. Smart Grid*, vol. 10, no. 6, pp. 6933–6944, 2019.

[14] L. Guo, Y. Zhang, X. Li, Z. Wang, Y. Liu, L. Bai, and C. Wang, "Data-driven power flow calculation method: A lifting dimension linear regression approach," *IEEE Trans. Power Syst.*, 2021.

[15] J. Chen, W. Wu, and L. A. Roald, "Data-driven piecewise linearization for distribution three-phase stochastic power flow," *IEEE Trans. Smart Grid*, 2021.

[16] J. Zhang, L. Guan, and C. Chung, "Instantaneous sensitivity identification in power systems-challenges and technique roadmap," in *2016 IEEE PESGM*. IEEE, 2016, pp. 1–5.

[17] P. Li, H. Su, C. Wang, Z. Liu, and J. Wu, "Pmu-based estimation of voltage-to-power sensitivity for distribution networks considering the sparsity of jacobian matrix," *IEEE Access*, vol. 6, pp. 31 307–31 316, 2018.

[18] J. Zhang, Z. Wang, X. Zheng, L. Guan, and C. Chung, "Locally weighted ridge regression for power system online sensitivity identification considering data collinearity," *IEEE Trans. Power Syst.*, vol. 33, no. 2, pp. 1624–1634, 2017.

[19] M. Beza and M. Bongiorno, "Application of recursive least squares algorithm with variable forgetting factor for frequency component estimation in a generic input signal," *IEEE Trans. Ind. Appl.*, vol. 50, no. 2, pp. 1168–1176, 2013.

[20] T. Ahmad and N. Senroy, "Statistical characterization of pmu error for robust wams based analytics," *IEEE Trans. Power Syst.*, vol. 35, no. 2, pp. 920–928, 2019.

[21] J. Yu, Y. Weng, and R. Rajagopal, "Robust mapping rule estimation for power flow analysis in distribution grids," in *2017 North American Power Symposium (NAPS)*. IEEE, 2017, pp. 1–6.

[22] M. Brown, M. Biswal, S. Brahma, S. J. Ranade, and H. Cao, "Characterizing and quantifying noise in pmu data," in *2016 IEEE PESGM*. IEEE, 2016, pp. 1–5.

[23] D. J. Gaushell and W. R. Block, "Scada communication techniques and standards," *IEEE computer applications in power*, vol. 6, no. 3, pp. 45–50, 1993.

[24] C. Mugnier, K. Christakou, J. Jaton, M. De Vivo, M. Carpita, and M. Paolone, "Model-less/measurement-based computation of voltage sensitivities in unbalanced electrical distribution networks," in *2016 PSCC*. IEEE, 2016, pp. 1–7.

[25] Y. Liu, Z. Li, and Y. Zhou, "Data-driven-aided linear three-phase power flow model for distribution power systems," *IEEE Trans. Power Syst.*, 2021.

[26] S. Nowak, Y. C. Chen, and L. Wang, "Measurement-based optimal der dispatch with a recursively estimated sensitivity model," *IEEE Trans. Power Syst.*, vol. 35, no. 6, pp. 4792–4802, 2020.

[27] H. Xu, A. D. Domínguez-García, V. V. Veeravalli, and P. W. Sauer, "Data-driven voltage regulation in radial power distribution systems," *IEEE Trans. Power Syst.*, vol. 35, no. 3, pp. 2133–2143, 2019.

[28] J. Yu, W. Dai, W. Li, X. Liu, and J. Liu, "Optimal reactive power flow of interconnected power system based on static equivalent method using border pmu measurements," *IEEE Trans. Power Syst.*, vol. 33, no. 1, pp. 421–429, 2017.

[29] J. L. Cremer, I. Konstantelos, and G. Strbac, "From optimization-based mach. learn. to interpretable security rules for operation," *IEEE Trans. Power Syst.*, vol. 34, no. 5, pp. 3826–3836, 2019.

[30] Y. Liu, B. Xu, A. Botterud, N. Zhang, and C. Kang, "Bounding regression errors in data-driven power grid steady-state models," *IEEE Trans. Power Syst.*, vol. 36, no. 2, pp. 1023–1033, 2020.

[31] Y. Chen, C. Wu, and J. Qi, "Data-driven power flow method based on exact linear regression equations," *J. Mod. Power Syst. Clean Energy*, 2021.

[32] J. Zhao, H.-D. Chiang, P. Ju, and H. Li, "On pv-pq bus type switching logic in power flow computation," in *2008 PSCC, Glasgow, Scotland*, 2008.

[33] G. Jain, S. Sidar, and D. Kiran, "Alternative regression approach for data-driven power flow linearization methods," in *2021 IEEE 9th ICPS*. IEEE, 2021, pp. 1–6.

[34] X. Li, "Fast heuristic ac power flow analysis with data-driven enhanced linearized model," *Energies*, vol. 13, no. 13, p. 3308, 2020.

[35] S. hentong, Z. Qiaozhu, W. Jiang, and G. Xiaohong, "Data based linearization: Least-squares based approximation," *arXiv preprint arXiv:2007.02494*, 2020.

[36] P. Li, W. Wu, X. Wan, and B. Xu, "A data-driven linear optimal power flow model for distribution networks," *IEEE Trans. Power Syst.*, 2022.

[37] S. Powell, A. Ivanova, and D. Chassin, "Fast solutions in power system simulation through coupling with data-driven power flow models for voltage estimation," *arXiv preprint arXiv:2001.01714*, 2020.

[38] Y. Tan, Y. Chen, Y. Li, and Y. Cao, "Linearizing power flow model: A hybrid physical model-driven and data-driven approach," *IEEE Trans. Power Syst.*, vol. 35, no. 3, pp. 2475–2478, 2020.

[39] J. Chen, W. Li, W. Wu, T. Zhu, Z. Wang, and C. Zhao, "Robust data-driven linearization for distribution three-phase power flow," in *2020 IEEE 4th EI2*. IEEE, 2020, pp. 1527–1532.

[40] J. Yu, Y. Weng, and R. Rajagopal, "Mapping rule estimation for power flow analysis in distribution grids," *arXiv preprint arXiv:1702.07948*, 2017.

[41] Z. Shao, Q. Zhai, Z. Han, and X. Guan, "A linear ac unit commitment formulation: An application of data-driven linear power flow model," *Int. J. Electr. Power Energy Syst.*, vol. 145, p. 108673, 2023.

[42] Y. Liu, Z. Li, and J. Zhao, "Robust data-driven linear power flow model with probability constrained worst-case errors," *arXiv preprint arXiv:2112.10320*, 2021.

[43] F. J, H. T, and T. R, *The elements of statistical learning*. Springer, 2001.

[44] A. Von Meier, E. Stewart, A. McEachern, M. Andersen, and L. Mehrmanesh, "Precision micro-synchrophasors for distribution systems: A summary of applications," *IEEE Trans. Smart Grid*, vol. 8, no. 6, pp. 2926–2936, 2017.

[45] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[46] H. Wold, "Path models with latent variables: The nipals approach," in *Quantitative sociology*. Elsevier, 1975, pp. 307–357.

[47] S. De Jong, "Simpls: an alternative approach to partial least squares regression," *Chemom. Intell. Lab. Syst.*, vol. 18, no. 3, pp. 251–263, 1993.

[48] M. Hubert and K. V. Branden, "Robust methods for partial least squares regression," *J. Chemom.: A Journal of the Chemometrics Society*, vol. 17, no. 10, pp. 537–549, 2003.

[49] A. E. Hoerl, R. W. Kannard, and K. F. Baldwin, "Ridge regression: some simulations," *Commun. Stat. - Theory Methods*, vol. 4, no. 2, pp. 105–123, 1975.

[50] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Stat. Comput .*, vol. 14, no. 3, pp. 199–222, 2004.

[51] G. W. Flake and S. Lawrence, "Efficient svm regression training with smo," *Mach. Learn.*, vol. 46, no. 1, pp. 271–290, 2002.

[52] Y. Zhang, S. Shen, and J. L. Mathieu, "Distributionally robust chance-constrained optimal power flow with uncertain renewables and uncertain reserves provided by loads," *IEEE Trans. Power Syst.*, vol. 32, no. 2, pp. 1378–1388, 2016.

[53] R. Jiang and Y. Guan, "Data-driven chance constrained stochastic program," *Math. Program.*, vol. 158, no. 1, pp. 291–327, 2016.

[54] W. Wei, "Tutorials on advanced optimization methods," *arXiv preprint arXiv:2007.13545*, 2020.

[55] Y. Weng, R. Negi, C. Faloutsos, and M. D. Ilić, "Robust data-driven state estimation for smart grid," *IEEE Trans. Smart Grid*, vol. 8, no. 4, pp. 1956–1967, 2016.