Interpretable Machine Learning for Power Systems: Establishing Confidence in SHapley Additive exPlanations

T. Ahmad[1], RI Hamilton[2], J. Stiasny[3], S. Chevalier [3], R.Nellikkath[3], I. Murzakhanov[3], S. Chatzivasileiadis[3], PN Papadopoulos[1]
1 - University of Manchester, UK 2 - University of Strathclyde, UK 3- Technical University of Denmark, Lyngby

Paper ID: 81

Motivation

- The Paris Agreement on Climate Change (2021) has set a target to minimise global warming to $1.5°$ C; GHG emissions from electricity and heating account 25% of total emissions.
- Prompted by global calls for decarbonisation, there is an increased trend to integrate more renewable energy (RE) sources (like wind, solar, etc.) for electric power generation.
- Black-box ML models to predict power system behaviour have been developed earlier, however their applicability critical infrastructure like power systems, is limited due to absence of trustworthy explanations.

Introduction & Objectives

- SHAP is an additive feature attribution method for posthoc ML interpretability, which constructs a simple additive *explanation model*, g—which is a linear function of binary variables—to represent the complex *original model*, f [1, 2].
- An effect ϕ_i (where $\phi_i \in R$) is attributed to each feature, the sum of which approximates $f(x)$.
- **Objectives**: To showcase SHAP interpretations as a tool for understanding power system ML models and establish SHAP explanations as a way to capture underlying power system physics.
- This is achieved by proving that *the derivative of SHAP values for an ML model learning DC power flows is equal to the Power Transfer Distribution Factor (PTDF) of the network*.
- PTDF is a state-of-the-art linear sensitivity given by the incremental change in real power that occurs on transmission lines due to real power transfers between two regions and is an important operational index in practical power systems[3].

Derivation of PTDF from SHAP values

- To show the link between SHAP values and the PTDF analytically, we consider $f_{line,i-j}(x)$ to be a linear statistical model $f(x) = w^\top x + b$ whose features, ($x \in P$) are assumed to be independent.
  Using Corollary 1 from [1], the SHAP values $\phi_i(f, x)$ associated with $f(x)$ are given by

$$\phi_0(f, x) = b \tag{1}$$
$$\phi_i(f, x) = w_i(x_i - E[X_i]),\ i \neq 0 \tag{2}$$

where X_i is the training data associated with the i^{th} feature. Perturbing the i^{th} feature (i.e., continuous regressor) yields

$$\frac{\partial \phi_i}{\partial x_i} = w_i. \tag{3}$$

Since w_i relates the sensitivity of line flow to the i^{th} injection, the SHAP derivative is equivalent to a PTDF. To show this experimentally, the sum across all SHAP values yields a model with linear sensitivity to power injections, i.e., an approximate PTDF vector $\hat{D}$, and a constant offset term ϵ:

$$\sum_i \phi_i(f, x) = w^\top (x - E[X]) + b \approx \hat{D}^\top P + \epsilon. \tag{4}$$

By collecting a library of SHAP values Φ associated with a library of sampled injection values $\mathbf{P}$, a regression procedure can yield the PTDF approximation $\hat{D}$ where $(\cdot)^+$ denotes Moore–Penrose pseudoinversion:

$$\begin{bmatrix} \hat{D} \\ \epsilon \end{bmatrix} = \mathbf{P}^+ \Phi, \tag{5}$$

Case Study

- DC power flow analysis is used to evaluate the flow of electricity in the power system network in order to take day-to-day operational decisions.
- Individual XGBoost ML regression models $f_{line,i-j}(x)$ are trained to predict the real power flow ($F_{line,i-j}$) on each line of the test power system.
- The input features $x \in P$ include $PG2$ and $PG3$ drawn independently from a uniform distribution in the range [0, 500] MW for a total of 1001 generation scenarios. Demand is assumed constant at 315 MW
- Real power flows on line (following DC power-flow), $F_{line,i-j}$ are targets for all 1001 operational scenarios.
- A 75-25% train-test split is used for the XGBoost models.
- For the trained models, SHAP values are calculated (using the framework set out in [4], implemented in Python).

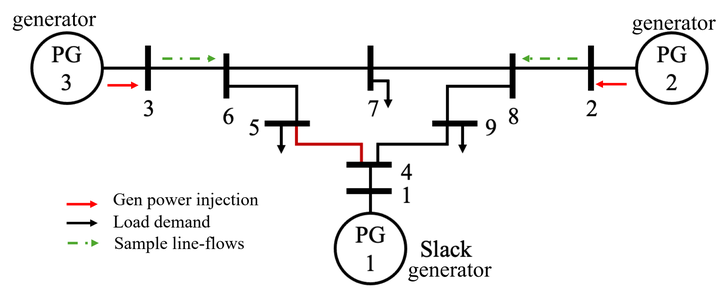

Figure 1: Single line diagram of the IEEE 3 machine 09 bus system

Results

- **Local explanation**: provide the contributions of features (i.e. SHAP values), ϕ_i) in shifting the model prediction from the model expectation ($E[f(x)]$). Figure below shows local SHAP explanation for the model $f_{line,4-5}(x)$. Since $|\phi_{PG2}| > |\phi_{PG3}|$, $PG2$ has a greater effect in this scenario and thus is placed higher on the y-axis.

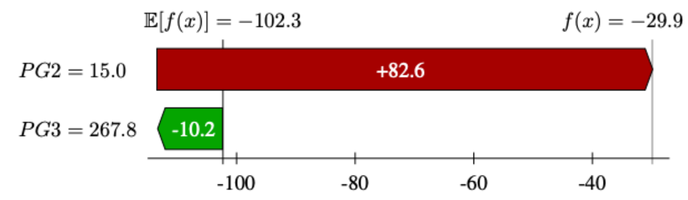

Figure 2: Local explanation for the model $f_{line,4-5}(x)$

- **Global explanation** : comprises of the local explanations for the entire training database. In the Figure below, $PG3$ is found to have a higher importance than $PG2$ – inverse to the local explanation for the scenario above.

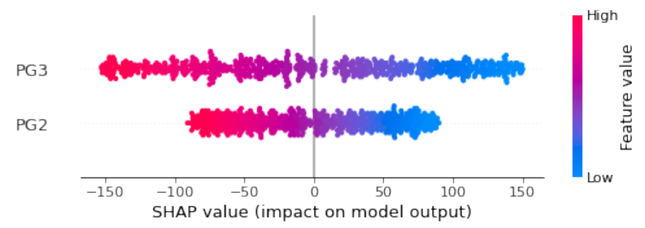

Figure 3: Global explanation for $F_{line,4-5}(x)$

- The analytical equivalence of SHAP and PTDF derived before is also verified below experimentally for the 3-machine 09 bus case.

Table 1: 3 machine 9 bus PTDF & SHAP

	True physical PTDF, D		SHAP-based PTDF, $\hat{D}$	
Line	Bus 2	Bus 3	Bus 2	Bus 3
Line 1-4	-1.0000	-1.0000	-0.9999	-0.9999
Line 4-5	-0.3613	-0.6152	-0.3613	-0.6151
Line 5-6	-0.3613	-0.6152	-0.3613	-0.6151
Line 3-6	0	1.0000	0.0000	0.9999
Line 6-7	-0.3613	0.3848	-0.3613	0.3848
Line 7-8	-0.3613	0.3848	-0.3613	0.3848
Line 8-2	-1.0000	0	-1.0000	0.0000
Line 8-9	0.6387	0.3848	0.6386	0.3848
Line 9-4	0.6387	0.3848	0.6386	0.3848

Summary and Outlook

- Interpretable ML is expected to play an important role in understanding increasingly complex models for decarbonised power systems.
- Physical equivalence is one such way of developing confidence, which we have shown here for SHAP and PTDF in a simple DC power flow case.
- Extending the linear case to more complex nonlinear problems will likely be a fruitful avenue of future research.

KEY REFERENCES

[1] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions.144 Advances in Neural Information Processing Systems, 30, 2017

[2] Robert I. Hamilton and Panagiotis N. Papadopoulos. Using SHAP values and machine learning to understand trends in the transient stability limit. IEEE Trans. on Power Syst., 2023.

[3] Spyros Chatzivasileiadis. Lecture notes on Optimal Power Flow (OPF). CoRR, abs/1811.00943,155 2018.

[4] Scott M Lundberg *et al*. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1):56–67,153 2020.

PARTNERS