ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095768

PHYSICS-INFORMED TRANSFER LEARNING FOR VOLTAGE STABILITY MARGIN PREDICTION

Manish K. Singh,⋆ Konstantinos D. Polyzos,⋆ Panagiotis A. Traganitis,†
Sairaj V. Dhople,⋆ and Georgios B. Giannakis⋆

⋆ Dept. of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN, USA
†Dept. of Electrical and Computer Engineering, Michigan State University, East Lansing, MI, USA

ABSTRACT

Assessing set-membership and evaluating distances to the related set boundary are problems of widespread interest, and can often be computationally challenging. Seeking efficient learning models for such tasks, this paper deals with voltage stability margin prediction for power systems. Supervised training of such models is conventionally *hard* due to high-dimensional feature space, and a cumbersome label-generation process. Nevertheless, one may find related *easy* auxiliary tasks, such as voltage stability verification, that can aid in training for the hard task. This paper develops a novel approach for such settings by leveraging transfer learning. A Gaussian process-based learning model is efficiently trained using learning- and physics-based auxiliary tasks. Numerical tests demonstrate markedly improved performance that is harnessed alongside the benefit of uncertainty quantification to suit the needs of the considered application.

***Index Terms*—** Gaussian processes, set-membership, transfer learning, voltage stability.

1. INTRODUCTION

Computing the voltage stability margin is critical in power-system operations [1, 2], and forms the primary application considered in this work. The underlying setup, however, is representative for several signal processing, machine learning, and control applications. Addressing a wider audience, the abstract problem setup will be provided next, while the details of voltage stability margin are deferred to Sec. 2.

Problem statement. Consider a compact set $\mathcal{X} \subset \mathbb{R}^M$, and let us define a function $b(\cdot) : \mathbb{R}^M \to \{0, 1\}$ as

$$b(\mathbf{x}) = \begin{cases} 1 & , \ \mathbf{x} \in \mathcal{X} \\ 0 & , \ \text{otherwise.} \end{cases}$$

Function $b(\cdot)$ characterizes the set $\mathcal{X}$. The distance from a given $\mathbf{x}_0 \in \mathcal{X}$ to the boundary of $\mathcal{X}$ is defined as

$$d(\mathbf{x}_0) := \min_{\mathbf{x} \notin \mathcal{X}} \|\mathbf{x} - \mathbf{x}_0\|_2. \tag{1}$$

Part of this work was supported by NSF grants 2128593, 2126052, 2212318, and 2220292. The work of Konstantinos D. Polyzos was also supported by the Onassis Foundation Scholarship.

While metric $d(\mathbf{x}_0)$ could be of critical importance, evaluating it may be challenging. Thus, despite having access to computing $d(\cdot)$, a reliable learning-based surrogate $\widehat{d}(\cdot)$ is valuable for real-time applications.

A typical supervised learning setup to obtain a surrogate $\widehat{d}(\cdot)$ would require a training dataset $\mathcal{D}_{\text{train}} := \{(\mathbf{x}_\tau, d_\tau)\}_{\tau=1}^{T}$, where $d_\tau = d(\mathbf{x}_\tau)$. In several applications, the feature-dimension $\dim(\mathbf{x}) = M$ is large, and building an adequately-sized dataset becomes prohibitive due to the complexity in computing $d(\cdot)$. In such cases, the affordable number of training samples T may not be adequate to satisfactorily learn the surrogate $\widehat{d}(\cdot)$. The membership verification function $b(\cdot)$ may however be relatively easier to compute. This allows to readily obtain an *auxiliary* dataset $\mathcal{D}_{\text{train}}^{\text{aux}} := \{(\mathbf{x}_a, b_a)\}_{a=1}^{T_{\text{aux}}}$, where $b_a = b(\mathbf{x}_a)$. When $b(\cdot)$ is significantly easier to compute than $d(\cdot)$, one can afford $T_{\text{aux}} \gg T$. For the discussed setup, the goal of this paper can be made explicit as follows.

Goal: *Given a training dataset $\mathcal{D}_{\text{train}}$ and an auxiliary dataset $\mathcal{D}_{\text{train}}^{\text{aux}}$, obtain a learning-based surrogate $\widehat{d}(\cdot)$.*

Pertinent applications. The setup above appears in applications as diverse as adversarial machine learning, intrusion-detection for cybersecurity, (in)stability analysis for dynamical systems, and maximal loadability/throughput analysis in physical networks. We will pick a couple to elaborate on.

• *Adversarial machine learning*. Learning-based models are known to misclassify images corrupted by potentially unnoticeable adversarial perturbations [3, 4]. Thus, alongside the inferred class-label, one may be interested in quantifying the change, that is the perturbation, that could alter the label. In such cases, the classifier (assuming binary) serves as $b(\cdot)$, and $\mathcal{X}$ is the set in feature-space that the model assigns to a given class. A small $d(\cdot)$ indicates an adversarial instance or an instance susceptible to adversarial perturbation.

• *Stability-margin prediction*. The stability of nonlinear dynamical systems is contingent on the system state. Assessing stability for a given state constitutes a binary investigation such as $b(\cdot)$, where $\mathcal{X}$ is the stability region. The metric $d(\cdot)$ then quantifies the disturbance that could cause instability.

Prior works on Transfer Learning. Approaches to improve an inference task by utilizing information from a different,

but related, task constitute the paradigm of transfer learning (TL) [5]. Typically, TL is used in supervised learning when the target task has to rely on only a few labeled data, whereas the auxiliary tasks have sufficient data to train a well performing model. TL has gained popularity over the last decade in applications such as ultrasound imaging [6], automated audio captioning [7], and speech recognition [8], to list a few. Several approaches learn the auxiliary task using deep learning models; see e.g., [9], and [10] for Bayesian alternatives involving random models drawn from a joint prior probability density function (pdf). Similar in spirit to TL, self supervised learning (SSL) seeks to transfer information to a target task from so-called pretext tasks [11]. In contrast to the auxiliary input-output instances involved in TL, these pretext tasks rely on synthetically generated auxiliary instances [11].

Prior works on voltage stability. Long-term voltage stability amounts to the solvability of power-flow equations [1]. A pioneering development in this context is known as continuation power flow (CPF) [2]. CPF uses a prediction-correction equation solver along a given linear trajectory of operating conditions, thus yielding a *directional* stability margin. A neural network-based approach for fast prediction of directional margins was reported in [12]; see also [13] for an upper bound on this margin. Identifying the relation between power-flow solvability and the conditioning of the related Jacobian matrix, tractably-computable indices are often used as surrogates to stability margin [14]. Towards directly predicting the stability margins, a data-based approach employing an ensemble of linear LASSO regression models was proposed in [15] placing emphasis on dimensionality reduction of domain-based features. Faced with similar challenges, [16] and [17] argued that building a classifier using binary operator $b(\cdot)$ is much easier, and provides an informative scalar feature-embedding.

Contributions. This work provides a novel *physics-informed transfer learning* (PITL) approach for stability-margin prediction with the following exciting attributes: *i)* Informative low-dimensional features for the hard margin-prediction task are extracted by exploiting an auxiliary classification dataset that is easy to obtain; *ii)* Physics-based readily-computable indices are identified that improve margin-prediction when appended to the aforementioned low-dimensional features; and *iii)* a high-accuracy regression model based on Gaussian process (GP) ensembles (offering uncertainty quantification) is obtained using an extremely small dataset with low-dimensional input features. Finally, the proposed approach is numerically benchmarked against the closely related work of [16], while demonstrating the benefit of using a Gaussian ensemble over a single Gaussian process.

2. VOLTAGE STABILITY MARGIN

Consider a single-phase network with N nodes representing the per-phase equivalent of a three-phase balanced AC power system [18]. Let (p_n, q_n) denote the active- and reactive-power injections at node $n \in \{1, \ldots, N\}$; and $(v_n,\ \theta_n)$ the voltage magnitude and phase. It is typical to capture the network topology and line parameters via the admittance matrix, $\mathbf{Y} \in \mathbb{C}^{N\times N}$, expressed as $\mathbf{Y} = \mathbf{G} + \jmath\mathbf{B}$, where matrices $\mathbf{G}$ and $\mathbf{B}$ are real-valued. The dependence of (p_n, q_n) on (v_n, θ_n) is dictated by the *power flow* equations [18]

$$p_n = v_n \sum_{m=1}^{N} v_m(G_{nm}\cos\theta_{nm} + B_{nm}\sin\theta_{nm}) \tag{2a}$$

$$q_n = v_n \sum_{m=1}^{N} v_m(G_{nm}\sin\theta_{nm} - B_{nm}\cos\theta_{nm}) \tag{2b}$$

where $\theta_{nm} := \theta_n - \theta_m$. In classical power-system analysis, the nodes are classified into three categories based on the known quantities from $\{p_n, q_n, v_n, \theta_n\}$. There is a reference generator node r with a fixed voltage magnitude v_r and phase θ_r (typically $\theta_r = 0$). The remaining generators are treated as PV nodes with fixed (p_n, v_n), and loads are PQ nodes with known (p_n, q_n). Without loss of generality, let us index the first N_G nodes as generators, with the reference node as $r = 1$; and the remaining nodes $n \in \{N_G + 1, \ldots, N\}$ as loads. For brevity, define the vector of *known* quantities

$$\boldsymbol{\alpha} := [\theta_1, v_1, ..v_{N_G}, p_2, .., p_N, q_{N_G+1}, .., q_N]^\top.$$

Given $\boldsymbol{\alpha}$, the power flow problem aims at solving the $2N$ equations in (2) to obtain the remaining $2N$ *unknowns* in

$$\boldsymbol{\gamma} := [p_1, q_1, ..q_{N_G}, \theta_2, .., \theta_N, v_{N_G+1}, .., v_N]^\top.$$

The nonlinear power-flow equations (2) may not admit a solution for arbitrary values of $\boldsymbol{\alpha}$. An operating point $\boldsymbol{\alpha}$ is said to be *long-term voltage stable* if there exists a $\boldsymbol{\gamma}$ satisfying (2) [13, 2]. Oftentimes, power systems are operated such that $v_n \approx 1$ for $n = 1, \ldots, N_G$; and θ_1 is set to zero, limiting the variability in operation to $\mathbf{x} = [p_2, .., p_N, q_{N_G+1}, .., q_N]^\top$. Thus, the set of stable operating conditions can be defined as $\mathcal{X} := \{\mathbf{x} \mid (2) \text{ has a solution}\}$. The binary function $b(\cdot)$ then corresponds to the indicator function for stability, that is, $b(\mathbf{x}) = 1$ implies $\mathbf{x}$ is stable.

Since $\mathcal{X}$ is characterized by the solvability of nonlinear equations (2), computing $d(\cdot)$ as per (1) constitutes a non-convex optimization problem, which is hard to directly solve. A prevalent sub-problem in power systems is to compute the distance of $\mathbf{x}_0$ to the boundary in a particular direction. Specifically, given an $\mathbf{x}' \in \mathbb{R}^M$, the directional distance along $\mathbf{x}_0 \to \mathbf{x}'$ is $\widetilde{d}(\mathbf{x}_0, \mathbf{x}') = \|\lambda_{\min}(\mathbf{x}' - \mathbf{x}_0)\|_2$, where

$$\lambda_{\min} := \min\ \{\lambda \mid \mathbf{x}_0 + \lambda(\mathbf{x}' - \mathbf{x}_0) \notin \mathcal{X}\}. \tag{3}$$

From the definitions of distance, $d(\cdot)$, and directional distance, $\widetilde{d}(\cdot)$, it follows that $d(\mathbf{x}_0) = \min_{\mathbf{x}'} \widetilde{d}(\mathbf{x}_0, \mathbf{x}')$. A robust approach for computing $\widetilde{d}(\mathbf{x}_0, \mathbf{x}')$ is by using the CPF, which is incorporated in the (optimal) power flow toolbox MATPOWER [2, 19]. The value of $d(\mathbf{x}_0)$ can then be approximated by $\min_{\mathbf{x}' \in \{\mathbf{x}'_k\}_{k=1}^K} \widetilde{d}(\mathbf{x}_0, \mathbf{x}')$, where the directional

search is restricted to K random directions [16, 17]. Systems with high dimension, M, may require a large K for reliably approximating $d(\cdot)$; thus introducing large computation cost. To put this in perspective, for the considered setup in our numerical tests, the average time to assess stability (i.e., compute $b(\cdot)$), evaluate $\widetilde{d}(\cdot,\cdot)$, and approximately find $d(\cdot)$ requires 0.01, 0.17, and 50.9 seconds, respectively.

3. PROPOSED APPROACH

Given a dataset $\mathcal{D}_{\text{train}} := \{(\mathbf{x}_\tau, d_\tau)\}_{\tau=1}^T$, our pursuit to learn a model $\widehat{d}(\cdot)$ is challenged by a seemingly small T when considering the large dimension (M) of $\mathbf{x}$. To simplify the task, we advocate a two-step approach: *S1)* Obtain a low-dimensional, yet informative embedding $\mathbf{z} = \mathbf{f}(\mathbf{x})$, where $\mathbf{z} \in \mathbb{R}^m$ and $m \ll M$; and, *S2)* learning the model as $\widehat{d}(\mathbf{x}) \equiv g(\mathbf{z}) = g(\mathbf{f}(\mathbf{x}))$ using a transformed dataset $\widetilde{\mathcal{D}}_{\text{train}} := \{(\mathbf{z}_\tau, d_\tau)\}_{\tau=1}^T$. We next elucidate *S1)* and *S2)*.

In *S1)*, for each $\mathbf{x}_\tau$ in $\mathcal{D}_{\text{train}}$, we seek an embedding $\mathbf{z}_\tau = \left[\mathbf{z}_\tau^{\text{xfer}\top}\ \mathbf{z}_\tau^{\text{phy}\top}\right]^\top$. Embedding $\mathbf{z}_\tau^{\text{xfer}}$ is obtained using the auxiliary dataset $\mathcal{D}_{\text{train}}^{\text{aux}} := \{(\mathbf{x}_a, b_a)\}_{a=1}^{T_{\text{aux}}}$. We can afford $T_{\text{aux}} \gg T$ due to the relatively low complexity of computing $b(\mathbf{x})$. Intuitively, an accurate classifier $\widehat{b}(\cdot)$ assessing membership of $\mathbf{x}$ in $\mathcal{X}$ is anticipated to be cognizant of the set boundary [16]. Thus, we use $\mathcal{D}_{\text{train}}^{\text{aux}}$ to train a neural network-based classifier $\widehat{b}(\cdot)$, where the two outputs of the penultimate layer represent the likelihood of $\mathbf{x}_\tau$ belonging inside and outside $\mathcal{X}$, respectively. Once trained, we use these outputs as $\mathbf{z}_\tau^{\text{xfer}}$. In addition to these *transferred* embeddings, we include additional features $\mathbf{z}_\tau^{\text{phy}}$ motivated by power-system domain knowledge (hence motivating the name physics-informed transfer learning). Specifically, solvability of the power-flow equations (2) is related to the conditioning of the related Jacobian. Two readily-computable pertinent indices include the minimum of real part of the Jacobian's eigenvalues, and the log-determinant of the Jacobian matrix. These indices were chosen as they are often used as surrogates for quantifying voltage stability when computing the margin directly is not viable [14]. Concatenating the transferred and the physics-based embeddings, we obtain $\mathbf{z}_\tau \in \mathbb{R}^4$, allowing us to obtain the transformed dataset $\widetilde{\mathcal{D}}_{\text{train}} = \{(\mathbf{z}_\tau, d_\tau)\}_{\tau=1}^T$.

Step *S2)* involves training a regression model using $\widetilde{\mathcal{D}}_{\text{train}}$. Acknowledging that the motivation for predicting stability margin is safety oriented, uncertainty quantification is of paramount importance. To that end, we build regression models using (ensemble) Gaussian Processes ((E)GP).

Gaussian Processes. GPs have well-documented merits for learning a nonparametric random function along with its posterior pdf that fully quantifies the associated uncertainty in a sample-efficient manner [20]. When learning with GPs, the sought function g is assumed to be drawn from a GP prior; that is $g \sim \mathcal{GP}(0, \kappa(\mathbf{z}, \mathbf{z}'))$ where $\kappa(\cdot)$ denotes a kernel function that captures the pairwise similarity between $\mathbf{z}$ and $\mathbf{z}'$. This implies that the random vector $\mathbf{g}_T := [g(\mathbf{z}_1) \ldots g(\mathbf{z}_T)]^\top \sim \mathcal{N}(\mathbf{g}_T; \mathbf{0}_T, \mathbf{K}_T)$ with $\mathbf{K}_T$ being the $T \times T$ covariance matrix whose (m, m') entry is $[\mathbf{K}_T]_{m,m'} = \text{cov}(g(\mathbf{z}_m), g(\mathbf{z}_{m'})) := \kappa(\mathbf{z}_m, \mathbf{z}_{m'})$ [20]. Focusing on the regression task, where the per-datum likelihood can be written as $p(d_\tau | g(\mathbf{z}_\tau)) = \mathcal{N}(d_\tau; g(\mathbf{z}_t), \sigma_n^2)$, it can be shown that for any instance $\mathbf{z}_*$, the predictive pdf of the corresponding output d_* is given by [20]

$$p(d_* | \mathbf{d}_T; \mathbf{Z}_T) = \mathcal{N}(d_*; \hat{d}_*(\mathbf{z}_*), \sigma_*^2(\mathbf{z}_*)) \tag{4}$$

where

$$\hat{d}_*(\mathbf{z}_*) = \mathbf{k}_T^\top(\mathbf{z}_*)(\mathbf{K}_T + \sigma_n^2 \mathbf{I}_T)^{-1} \mathbf{y}_T \tag{5a}$$

$$\sigma_*^2(\mathbf{z}_*) = \kappa(\mathbf{z}_*, \mathbf{z}_*) - \mathbf{k}_T^\top(\mathbf{z}_*)(\mathbf{K}_T + \sigma_n^2 \mathbf{I}_T)^{-1} \mathbf{k}_T(\mathbf{z}_*) \tag{5b}$$

with $\mathbf{Z}_T := [\mathbf{z}_1, \ldots, \mathbf{z}_T]^\top$, $\mathbf{d}_T := [d_1, \cdots, d_T]^\top$ and $\mathbf{k}_T(\mathbf{z}_*) := [\kappa(\mathbf{z}_1, \mathbf{z}_*), \ldots, \kappa(\mathbf{z}_T, \mathbf{z}_*)]^\top$. Note that (5a) provides a prediction for $\mathbf{z}_*$ with the predictive variance in (5b) quantifying the associated uncertainty. The performance of a single GP predictor hinges on a pre-selected kernel κ which may exhibit limited expressiveness of the learning function space, thus motivating the EGP framework delineated next.

Ensemble GPs. Aiming at a richer function space, an ensemble of M GP models is considered, each relying on a distinct kernel chosen from a given kernel dictionary $\mathcal{K} := \{\kappa^1, \ldots, \kappa^M\}$ that comprises kernels of different types and/or different hyperparameters [21]. This means that for each GP model $m \in \{1, \ldots, M\}$ a unique GP prior is postulated as $f|m \sim \mathcal{GP}(0, \kappa^m(\mathbf{z}, \mathbf{z}'))$. Combining all M GP priors with the corresponding weights $\{w_0^m\}_{m=1}^M$ yields the EGP prior

$$f(\mathbf{z}) \sim \sum_{m=1}^M w_0^m \mathcal{GP}(0, \kappa^m(\mathbf{z}, \mathbf{z}')), \quad \sum_{m=1}^M w_0^m = 1 \tag{6}$$

where $w_0^m := \Pr(i = m)$ is deemed as the prior probability that measures the significance of the corresponding GP model m. Capitalizing on (6), the EGP-based function posterior pdf can be written using the sum-product rule as

$$p(f(\mathbf{z}) | \tilde{\mathcal{D}}_{\text{train}}) = \sum_{m=1}^M \Pr(i = m | \tilde{\mathcal{D}}_{\text{train}}) p(f(\mathbf{z}) | i = m, \tilde{\mathcal{D}}_{\text{train}})$$

which is a mixture of posterior GPs with weights $w_T^m := \Pr(i = m | \tilde{\mathcal{D}}_{\text{train}})$ obtained by

$$w_T^m \propto \Pr(i = m) p(\tilde{\mathcal{D}}_{\text{train}} | i = m) = w_0^m p(\tilde{\mathcal{D}}_{\text{train}} | i = m)\ . \tag{7}$$

Then, for any instance $\mathbf{z}_*$ each GP model forms its (Gaussian in the regression task) predictive pdf $p(d_* | \tilde{\mathcal{D}}_{\text{train}}, m; \mathbf{z}_*) = \mathcal{N}(d_*; \hat{d}_*^m(\mathbf{z}_*), (\sigma_*^m(\mathbf{z}_*))^2)$. Upon combining all GP model predictive pdfs with the adjusted weights $\{w_T^m\}_{m=1}^M$, the EGP predictive pdf is given by [21]

$$\begin{aligned} p(d_* | \tilde{\mathcal{D}}_{\text{train}}; \mathbf{z}_*) &= \sum_{m=1}^M w_T^m p(d_* | \tilde{\mathcal{D}}_{\text{train}}, m; \mathbf{z}_*) \\ &= \sum_{m=1}^M w_T^m \mathcal{N}(d_*; \hat{d}_*^m(\mathbf{z}_*), (\sigma_*^m(\mathbf{z}_*))^2). \end{aligned} \tag{8}$$

Table 1. NMSE and NPLL performance

Method	NMSE	NPLL
GP	1.1628 ± 0.4958	26.1319 ± 10.0676
EGP	1.0391 ± 0.2198	27.9818 ± 2.0642
PITL-GP	0.4892 ± 0.0189	1.5048 ± 0.1601
PITL-EGP	$\mathbf{0.4577 \pm 0.003}$	$\mathbf{0.4927 \pm 0.0592}$
Baseline [16]	0.8847	-

4. NUMERICAL TESTS

In this section, the performance of the proposed approach is evaluated on the IEEE 14-bus power system that comprises 5 generator and 9 load buses. Omitting the buses with zero generation/load, the vector of nominal (re)active powers $\mathbf{x}_{\text{nom}} \in \mathbb{R}^{23}$ was obtained from the MATPOWER `casefile` [19]. To obtain the training and testing datasets for the stability-margin prediction task, we sampled 100 random $\mathbf{x}$'s by scaling $\mathbf{x}_{\text{nom}}$ entry-wise by a scalar drawn independently and uniformly within $[0, 3.5]$. The stability-margin label for each sample was obtained as the minimum of 300 directional margins computed via CPF tool in MATPOWER. These 300 directions were generated per sample by drawing the entries of $\mathbf{x}'$ in (3) from a uniform distribution on $[-25, 25]$. The aforementioned process yielded the set $\{(\mathbf{x}_\tau, d_\tau)\}_{\tau=1}^{100}$, which was equi-partitioned into a training set $\mathcal{D}_{\text{train}}$ and a test set $\mathcal{D}_{\text{test}}$. For the $\mathbf{x}$'s in $\mathcal{D}_{\text{train}}$, the physics-based embeddings $\mathbf{z}_2$ were also obtained using the Jacobian matrices provided by MATPOWER. Next, for the auxiliary task, $T_{\text{aux}} = 10,000$ samples of $\mathbf{x}$'s were generated with the scaling drawn from $[0, 7.5]$. The larger variation was to ensure balanced classes of (un)stable points. The stability label $b(\mathbf{x})$ was obtained via MATPOWER yielding $\{\mathbf{x}_a, b_a\}_{a=1}^{T_{\text{aux}}}$. The training set $\mathcal{D}_{\text{train}}^{\text{aux}}$ was formed with $7,000$ instances and the remaining $3,000$ were used for validation.

We compare the proposed approach with three algorithms. The first two involve a single GP model, and an EGP model, respectively, without exploiting any auxiliary task(s); see e.g., [20, 21]. The third algorithm utilizes a binary classification auxiliary task, followed by a uni-variate linear regression [16]. Being closest to our approach, we refer to this algorithm as the 'baseline,' and implemented it with architecture and parameters provided in [16].

For the proposed PITL approach with GP and EGP models (referred hereon as PITL-GP and PITL-EGP), the first step of obtaining embedding $\mathbf{z}_1$ was carried out by training a single-layer neural-network with 1,000 neurons as a classifier with cross-entropy loss. Other relevant hyperparameters were determined using the maximum validation accuracy criterion. Having trained the classifier, the regression training set $\mathcal{D}_{\text{train}}$ was transformed to include $\mathbf{z}$ yielding $\widetilde{D}_{\text{train}} = \{\mathbf{z}_\tau, d_\tau\}_{\tau=1}^{50}$. For the regression step, two models were tried: a single GP with RBF kernel; and an EGP model with kernel dictionary $\mathcal{K}$ consisting of four kernels: RBF kernels with and without auto-relevance determination, and two Matérn kernels with $\nu = 1.5, 2.5$, respectively. The kernel hyperparameters per GP were obtained by maximizing the marginal log-likelihood.

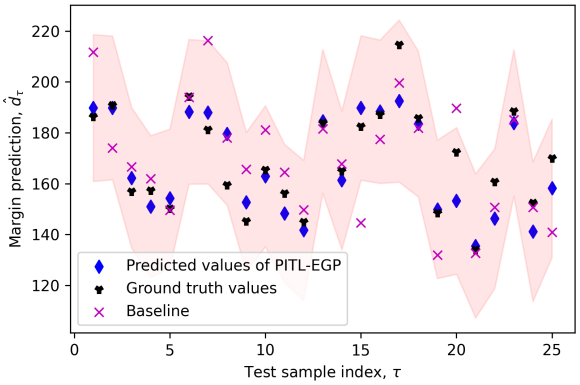

Fig. 1. Performance visualization with 25 test instances.

All competing approaches are assessed on a test set $\mathcal{D}_{\text{test}}$ using the normalized mean-square error (NMSE) to quantify the accuracy of point-predictions; and the negative predictive log-likelihood (NPLL) metrics to further account for the associated uncertainty as in [22]. The test results summarized in Table 1 demonstrate that the advocated PITL-GP and PITL-EGP approaches significantly outperform their single GP and EGP counterparts. These results corroborate the hypothesis that the reduced-dimension, physics-based, and transferred embeddings, serve as excellent predictors. Compared to the baseline in [16], the proposed PITL-GP and PITL-EGP methods not only enjoy lower NMSE, as shown in Table 1, but additionally offer uncertainty quantification through the predictive variance. Figure 1 depicts the predicted values of PITL-EGP along with standard deviation σ-confidence intervals, that the ground truth values were found to lie inside.

5. CONCLUSION

This work sets up the power-system voltage stability margin prediction task as a set-membership problem. In the face of small training datasets for the target regression task, a novel physics-informed transfer learning approach is developed. Knowledge (in form of a low-dimensional embedding) is transferred from an auxiliary classification task that enjoys larger training-data availability. The obtained transferred embeddings are augmented using physics-based readily-computable features that ultimately enable a satisfactory regression performance on the target task. Addressing the need of uncertainty quantification in safety-critical applications, (ensemble) Gaussian processes are employed for predicting the stability margins. Encouraging empirical performance of the developed approach and the generalized problem setup motivate future research with broad application base.

6. REFERENCES

[1] P. Kundur, J. Paserba, V. Ajjarapu, G. Andersson, A. Bose, C. Canizares, N. Hatziargyriou, D. Hill, A. Stankovic, C. Taylor, T. Van Cutsem, and V. Vittal, "Definition and classification of power system stability IEEE/CIGRE joint task force on stability terms and definitions," *IEEE Trans. Power Syst.*, vol. 19, no. 3, pp. 1387–1401, Aug. 2004.

[2] V. Ajjarapu and C. Christy, "The continuation power flow: a tool for steady state voltage stability analysis," *IEEE Trans. Power Syst.*, vol. 7, no. 1, pp. 416–423, Feb. 1992.

[3] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. learn. Representations*, May 2015.

[4] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, July 2017.

[5] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[6] T. Liu, S. Xie, J. Yu, L. Niu, and W. Sun, "Classification of thyroid nodules in ultrasound images using deep model based transfer learning and hybrid features," in *Proc. IEEE Int. Conf. Acoust., Speech, Sig. Process.*, Mar. 2017, pp. 919–923.

[7] X. Xu, H. Dinkel, M. Wu, Z. Xie, and K. Yu, "Investigating local and global information for automated audio captioning with transfer learning," in *Proc. IEEE Int. Conf. Acoust., Speech, Sig. Process.*, June 2021, pp. 905–909.

[8] M. Matassoni, R. Gretter, D. Falavigna, and D. Giuliani, "Non-native children speech recognition through transfer learning," in *Proc. IEEE Int. Conf. Acoust., Speech, Sig. Process.*, Apr. 2018, pp. 6229–6233.

[9] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *Int. Conf. Art. Neural Networks*, Oct. 2018, pp. 270–279.

[10] A. Karbalayghareh, X. Qian, and E. R. Dougherty, "Optimal bayesian transfer learning," *IEEE Trans. Sig. Process.*, vol. 66, no. 14, pp. 3724–3739, July 2018.

[11] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE Trans. Knowl. Data Eng.*, pp. 1–1, June 2021, (early access).

[12] D. Q. Zhou, U. D. Annakkage, and A. D. Rajapakse, "Online monitoring of voltage stability margin using an artificial neural network," *IEEE Trans. Power Syst.*, vol. 25, no. 3, pp. 1566–1574, Aug. 2010.

[13] D. K. Molzahn, I. A. Hiskens, and B. C. Lesieutre, "Calculation of voltage stability margins and certification of power flow insolvability using second-order cone programming," in *Proc. Hawaii Intl. Conf. on Syst. Sciences*, Jan. 2016, pp. 2307–2316.

[14] L. Aolaritei, S. Bolognani, and F. Dörfler, "Hierarchical and distributed monitoring of voltage stability in distribution networks," *IEEE Trans. Power Syst.*, vol. 33, no. 6, pp. 6705–6714, Nov. 2018.

[15] S. Li, V. Ajjarapu, and M. Djukanovic, "Adaptive online monitoring of voltage stability margin via local regression," *IEEE Trans. Power Syst.*, vol. 33, no. 1, pp. 701–713, Jan. 2018.

[16] Y.-h. Lee, Y. Zhao, S.-J. Kim, and J. Li, "Predicting voltage stability margin via learning stability region boundary," in *Proc. IEEE Workshop on Comp. Adv. in Multi-Sensor Adaptive Proc.*, Mar. 2017, pp. 1–5.

[17] J. Li, Y. Zhao, Y.-h. Lee, and S.-J. Kim, "Learning to infer voltage stability margin using transfer learning," in *Proc. IEEE Data Science Workshop*, June 2019, pp. 270–274.

[18] A. J. Wood and B. F. Wollenberg, *Power generation operation and control*, John Wiley & Sons, New York, NY, 2010.

[19] R. D. Zimmerman, C. E. Murillo-Sánchez, and R. J. Thomas, "MATPOWER: Steady-state operations, planning, and analysis tools for power systems research and education," *IEEE Trans. Power Syst.*, vol. 26, no. 1, pp. 12–19, Feb. 2011.

[20] C. E. Rasmussen and C. K. Williams, *Gaussian processes for machine learning*, MIT press Cambridge, MA, 2006.

[21] Q. Lu, G. Karanikolas, Y. Shen, and G. B. Giannakis, "Ensemble Gaussian processes with spectral features for online interactive learning with scalability," in *Proc. Int. Conf. Artif. Intel. and Stats.*, Aug. 2020, pp. 1910–1920.

[22] K. D. Polyzos, Q. Lu, and G. B. Giannakis, "Weighted ensembles for active learning with adaptivity," in *arXiv.2206.05009*, 2022.