Efficient Creation of Datasets for Data-Driven Power System Applications

Andreas Venzke, *Student Member, IEEE,* Daniel K. Molzahn, *Senior Member, IEEE,*
and Spyros Chatzivasileiadis, *Senior Member, IEEE*

***Abstract*—Advances in data-driven methods have sparked renewed interest for applications in power systems. Creating datasets for successful application of these methods has proven to be very challenging, especially when considering power system security. This paper proposes a computationally efficient method to create datasets of secure and insecure operating points. We propose an infeasibility certificate based on separating hyperplanes that can a-priori characterize large parts of the input space as insecure, thus significantly reducing both computation time and problem size. Our method can handle an order of magnitude more control variables and creates balanced datasets of secure and insecure operating points, which is essential for data-driven applications. While we focus on N-1 security and uncertainty, our method can extend to dynamic security. For PGLib-OPF networks up to 500 buses and up to 125 control variables, we demonstrate drastic reductions in unclassified input space volumes and computation time, create balanced datasets, and evaluate an illustrative data-driven application.**

***Index Terms*—Convex relaxation, data-driven, machine learning, optimal power flow, power system operation.**

I. Introduction

Recent advances in data-driven methods have shown substantial potential for power system applications including security assessment under uncertainty [1]–[6], e.g., by rapidly estimating line flows [2], training accurate security classifiers [3], or applying them in the context of data-driven security-constrained optimal power flow [4] and deep learning toolboxes [5]. The performance of these methods, however, relies on the quality of the underlying dataset. As historical data is often limited and does not contain many abnormal situations, the datasets have to be enriched through simulation. This, however, is a highly computationally demanding task. The resulting datasets should be balanced between secure and insecure samples to improve classifier performance, take into consideration all degrees of freedom of the system, and be able to accurately represent the security boundary. In this work, we propose an efficient method to create datasets with these properties for data-driven applications in power systems.

The steady-state operational constraints are described by the AC optimal power flow (AC-OPF) problem. The degrees of freedom of the system, i.e. the inputs characterizing each operating point, are defined by the control variables, which in the AC-OPF problem are generator active power and voltage set-points. By defining these, the remaining state variables are determined by solving the AC power flow equations [7]. Even for medium-sized systems, the number of control variables renders the task of creating datasets covering a wide range of operating points very computationally challenging.

To address this challenge, we can directly classify operating points that are infeasible with respect to the AC-OPF problem as insecure and avoid any further stability or static security assessment. Ref. [8] formulated infeasibility certificates to the AC-OPF problem based on hyperspheres that certify a wide range of operating points a-priori as insecure. Inspired by [8], our previous work in [9] used such certificates to generate large datasets, reducing the input space and decreasing computation time, while considering both N-1 security and small-signal stability. Both works [8] and [9] consider systems with up to 11 control variables. In this work, instead of hyperspheres, we propose the use of separating hyperplanes, which, among other important benefits, allows us to consider numbers of control variables that are at least an order of magnitude greater than previous methods (up to 125 in our test cases).

Another popular approach to create such datasets is through importance sampling, e.g., [10], [11]. In power systems, however, the initial sampling space is largely unbalanced, i.e., the volume of insecure space is several orders of magnitude larger than the secure space, and, as we observed in [9], it can be challenging to obtain an adequate number of secure samples. In this work, we show how our proposed method can lead to a balanced dataset, as it enables us to sample from inside the secure space. A related strand of research uses historical data and enriches them through sampling methods such as composite modelling approaches and vine-copulas [12], [13]. However, this can neglect parts of the secure space or might not capture abnormal operating regions.

To create representative datasets for data-driven power system applications, we propose a computationally efficient method which a) can deal with high input dimensionality (our test cases have up to 125 control variables), b) provides a detailed description of the security boundary, and c) creates *balanced* classes. We apply this method to AC-OPF problems including N-1 security and uncertainty in power injections. The main contributions of our work are:

1) We propose an infeasibility certificate based on separating hyperplanes. This certificate is computed using convex relaxations of AC-OPF problems and consid-

A. Venzke and S. Chatzivasileiadis are with the Department of Electrical Engineering, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark e-mail: {andven, spchatz}@elektro.dtu.dk.

D. K. Molzahn is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30313, USA as well as the Energy Systems Division, Argonne National Laboratory, Lemont, IL 60439 USA, e-mail: molzahn@gatech.edu.

This work is supported by the multiDC project, funded by Innovation Fund Denmark, Grant Agreement No. 6154-00020B.

ers both N-1 security and uncertainty. Compared to the hypersphere-based method proposed in [8], our algorithm shows two key improvements: First, separating hyperplanes allow the classification of substantially larger parts of the input space as insecure. Second, as these hyperplanes form a convex polyhedron, efficient methods to sample uniformly from inside the remaining unclassified space are available. Based on these, we propose an efficient algorithm to maximize the volume of the input space classified a-priori as insecure.

2) We evaluate this algorithm on PGLib-OPF networks with up to 500 buses and number of control variables up to 125. Compared to initial normalized input space volumes of 1 (i.e. 10^0) based on specified control variable bounds, the infeasibility certificates reduce the unclassified input space volumes significantly up to 10^{-40}.
3) We propose a computationally efficient method to create datasets for data-driven power system applications which can handle systems where the number of control variables is at least one order of magnitude greater than state-of-the-art methods (e.g., [9]). Computing infeasibility certificates allows us to efficiently characterize the security boundary in detail and sample from inside the secure space. We create balanced datasets for PGLib-OPF networks up to 500 buses and train neural network classifiers as an illustrative data-driven application.

This paper is structured as follows: In Section II, we describe the AC-OPF problem including N-1 security and uncertainty, and its convex relaxation. In Section III, we outline our proposed methodology to create datasets, including the infeasibility certificate, boundary description, and sampling from inside the secure space. Section IV presents simulation results on PGLib-OPF networks up to 500 buses. Section V concludes.

II. Optimal Power Flow Formulation

This section presents the AC-OPF formulations necessary for deriving the dataset creation methodology. In particular, we formulate the N-1 security-constrained preventive AC-OPF problem considering uncertainty in power injections, and its quadratic convex (QC) relaxation. For a detailed survey on AC-OPF and convex relaxations of the AC-OPF, the reader is referred to [7], [14]. Here, for brevity, we build our formulation upon the AC-OPF formulation of [15] to facilitate the derivation of the QC relaxation. We use the QC relaxation as it represents a good trade-off between computational complexity and tightness of the relaxation [15]. Note that the following derivations could be readily extended via the many other convex relaxations of the power flow equations [7].

A. Security-Constrained AC-OPF under Uncertainty

A power system is defined by its set $\mathcal{N}$ of buses. A subset of those buses, which are denoted by $\mathcal{G}$, have a controllable generator connected. A second subset denoted by $\mathcal{U}$, which can be either generation or load buses, are subject to uncertain power injections. It is assumed that all buses of the power system are connected by a set $(i,j) \in \mathcal{L}$ of power lines from bus i to bus j. To ensure the N-1 security criterion during operation, we consider the potential outage of a list of critical candidate lines defined by the set $\mathcal{C} \subset \mathcal{L}$. Note that we define the first entry of $\mathcal{C}$ to correspond to the intact system state $\{0\}$, i.e., no transmission line is outaged.

The optimization variables in the security-constrained AC-OPF are the complex bus voltages V_k^c for each bus $k \in \mathcal{N}$ and contingency $c \in \mathcal{C}$, the complex power dispatch of generator $S_{G_k}^c$ for each bus $k \in \mathcal{G}$ and contingency $c \in \mathcal{C}$, and the uncertain complex power injections S_{U_k} for each bus $k \in \mathcal{U}$. The uncertain power injections do not change upon outage of system components, i.e., $S_U = S_U^c \ \forall c \in \mathcal{C}$. We assume that the uncertain reactive power injection $Q_U = \Im\{S_U\}$ is determined through a fixed power factor $\cos\phi$ in relation to the uncertain active power injection $P_U = \Re\{S_U\}$, i.e., $Q_U = \sqrt{\frac{1-\cos^2\phi}{\cos^2\phi}}\, P_U$. The following constraints must be satisfied for the intact system and for each contingency $c \in \mathcal{C}$:

$$(V_k^{\min})^2 \leq V_k^c (V_k^c)^* \leq (V_k^{\max})^2 \qquad \forall k \in \mathcal{N} \quad (1a)$$
$$S_{G_k}^{\min} \leq S_{G_k}^c \leq S_{G_k}^{\max} \qquad \forall k \in \mathcal{G} \quad (1b)$$
$$|S_{ij}^c| \leq S_{ij}^{\max} \qquad \forall (i,j) \in \mathcal{L} \quad (1c)$$
$$S_{G_k}^c - S_{D_k} + S_{U_k} = \sum_{(k,j)\in\mathcal{L}} S_{kj}^c \qquad \forall k \in \mathcal{N} \quad (1d)$$
$$S_{ij}^c = (Y_{ij}^c)^* V_i^c (V_i^c)^* - (Y_{ij}^c)^* V_i^c (V_j^c)^* \qquad \forall (i,j) \in \mathcal{L} \quad (1e)$$
$$S_{U_k}^{\min} \leq S_{U_k} \leq S_{U_k}^{\max} \qquad \forall k \in \mathcal{U} \quad (1f)$$
$$-\theta_{ij}^{\max} \leq \angle(V_i^c (V_j^c)^*) \leq \theta_{ij}^{\max} \qquad \forall (i,j) \in \mathcal{L} \quad (1g)$$

The bus voltage magnitudes are constrained in (1a) by upper and lower limits $V_k^{\min}$ and $V_k^{\max}$. The superscript $*$ denotes the complex conjugate. Similarly, the generators' complex power outputs are limited in (1b) by upper and lower bounds $S_{G_k}^{\min}$ and $S_{G_k}^{\max}$. The inequality constraints for complex variables are defined as bounds on the real and imaginary parts. The apparent power flow S_{ij} on the line from i to j is upper bounded in (1c) by $S_{ij}^{\max}$. The nodal complex power balance (1d) including the load S_D, generation S_G and uncertain injections S_U has to hold for each bus. The apparent power flow S_{ij} on the line from i to j is defined in (1e). The term Y denotes the admittance matrix of the power grid. Constraint (1f) models minimum and maximum bounds $S_{U_k}^{\min}$, $S_{U_k}^{\max}$ on the uncertain injections. The flow on the line from i to j is limited in (1g) by an upper limit on angle differences $\theta_{ij}^{\max}$.

We consider preventive actions in the security-constrained AC-OPF formulation, i.e., the generator set-points remain fixed during an outage. As a result, we include the following linking constraints between the intact system state and the outaged system states:

$$|V_k^0| = |V_k^c| \qquad \forall k \in \mathcal{G},\ \forall c \in \mathcal{C}\backslash\{0\} \quad (2a)$$
$$P_{G_k}^0 = P_{G_k}^c \qquad \forall k \in \mathcal{G}\backslash\{\text{slack}\},\ \forall c \in \mathcal{C}\backslash\{0\} \quad (2b)$$

The first constraint sets the generator voltage set-points $|V_k|$ of the outaged system states to the values from the intact system state. The second constraint does the same for the active power generation dispatch, excluding the slack bus which compensates the difference in active power losses.

B. Quadratic Convex (QC) Relaxation

The QC relaxation proposed in [15] uses convex envelopes of the polar representation of the AC-OPF problem to relax the dependencies among voltage variables. As proposed in [15], [16], an additional auxiliary matrix variable W^c is introduced for the intact system state and each contingency $c \in \mathcal{C}$, which denotes the product of the complex bus voltages:

$$W_{ij}^c = V_i^c (V_j^c)^* \quad \forall c \in \mathcal{C} \tag{3}$$

This allows reformulation of (1a), (1e), (1g), and (2a) as:

$$(V_k^{\min})^2 \leq W_{kk}^c \leq (V_k^{\max})^2 \qquad \forall k \in \mathcal{N} \quad (4a)$$
$$S_{ij} = (Y_{ij}^c)^* W_{ii}^c - (Y_{ij}^c)^* W_{ij}^c \qquad \forall (i,j) \in \mathcal{L} \quad (4b)$$
$$S_{ij} = (Y_{ij}^c)^* W_{jj}^c - (Y_{ij}^c)^* (W_{ij}^c)^* \qquad \forall (i,j) \in \mathcal{L} \quad (4c)$$
$$\tan(-\theta_{ij}^{\max}) \leq \frac{\Re\{W_{ij}^c\}}{\Im\{W_{ij}^c\}} \leq \tan(\theta_{ij}^{\max}) \qquad \forall (i,j) \in \mathcal{L} \quad (4d)$$
$$W_{kk}^0 = W_{kk}^c \qquad \forall k \in \mathcal{G},\ \forall c \in \mathcal{C}\backslash\{0\} \quad (4e)$$

The non-convexity is encapsulated in the voltage product (3). To obtain a convex relaxation, the non-convex constraint (3) is removed from the optimization problem and variables for voltages, $v_i^c \angle \theta_i^c\ \forall i \in \mathcal{N}\ \forall c \in \mathcal{C}$, and squared current flows, $l_{ij}^c\ \forall (i,j) \in \mathcal{L}\ \forall c \in \mathcal{C}$, are added. The following convex constraints and envelopes are introduced for the intact system state and contingency $c \in \mathcal{C}$ [15]:

$$W_{kk}^c = \left\langle v_k^2 \right\rangle^T \qquad \forall k \in \mathcal{N} \quad (5a)$$
$$\Re\{W_{ij}^c\} = \left\langle \left\langle v_i^c v_j^c \right\rangle^M \left\langle \cos(\theta_i^c - \theta_j^c) \right\rangle^C \right\rangle^M \qquad \forall (i,j) \in \mathcal{L} \quad (5b)$$
$$\Im\{W_{ij}^c\} = \left\langle \left\langle v_i^c v_j^c \right\rangle^M \left\langle \sin(\theta_i^c - \theta_j^c) \right\rangle^S \right\rangle^M \qquad \forall (i,j) \in \mathcal{L} \quad (5c)$$
$$S_{ij}^c + S_{ji}^c = Z_{ij}^c l_{ij}^c \qquad \forall (i,j) \in \mathcal{L} \quad (5d)$$
$$|S_{ij}^c|^2 \leq W_{ii}^c l_{ij}^c \qquad \forall (i,j) \in \mathcal{L} \quad (5e)$$

The superscripts T, M, C, S denote convex envelopes for the square, bilinear product, cosine, and sine functions, respectively. The term Z_{ij} denotes the line impedance. Refer to [15] for the complete QC formulation. The resulting relaxation of the preventive security-constrained AC-OPF under uncertainty is a second-order cone program (SOCP) that minimizes an objective function, e.g., generation cost, subject to (1b)–(1d), (1f), (2b), (4), and (5).

III. Methodology to Create Datasets

The goal of the following methodology is to create a dataset which maps operating points described by the input vector x to a power system security classification, e.g., secure or insecure. The desired properties of the dataset are that it should be balanced between secure and insecure samples, take into consideration the degrees of freedom of the system, and have a detailed description of the security boundary. The power system security classification we consider is feasibility with respect to the N-1 security-constrained AC-OPF problem under uncertainty defined in (1) and (2). The resulting dataset can be complemented with further assessment of dynamic security criteria, e.g. small-signal stability [9]. The input vector x, i.e., the control variables that define the relevant degrees of freedom, is defined as follows:

$$x = \begin{bmatrix} P_{G_i}^0 \\ |V_j^0| \\ P_{U_k} \end{bmatrix} \qquad \forall i \in \mathcal{G}\backslash\{\text{slack}\},\ \forall j \in \mathcal{G},\ \forall k \in \mathcal{U} \tag{6}$$

We denote with P_G the active power dispatch, i.e., $P_G = \Re\{S_G\}$. Using the input x, all other states in the AC-OPF problem can be determined by solving the non-linear AC power flow equations. The minimum and maximum bounds on input vector $x^{\max}$ and $x^{\min}$ are defined in (1a), (1b), and (1f).

The main challenge in creating a representative and balanced dataset is the large number of control variables. The dimensionality of the input vector x grows substantially with increasing system size. For instance, the IEEE 118-bus system has 72 control variables, i.e., the dimensionality $|x|$ is 72. A naïve approach to create a dataset would be to sample with a prespecified discretization interval, e.g., by specifying 10 steps in each dimension of the control variables, $x_1, x_2, x_3, \ldots$. For the 118-bus system, this would require power flow solutions for 10^{72} operating points, which is computationally intractable. Further, as we will empirically show in Section IV-C, large parts of the input space $x \in [x^{\min}, x^{\max}]$ are infeasible. As a result, identifying secure samples by naïvely sampling from the entire input space is not possible for larger test cases.

To address these challenges, we present an efficient method for creating such datasets. First, to a-priori classify large parts of the input space as insecure, we propose an infeasibility certificate based on separating hyperplanes in Section III-A. Focusing on the unclassified regions, we then characterize the security boundary in detail in Section III-B. Finally, we sample inside the secure space in Section III-C.

A. Constructing Infeasibility Certificates

We propose an infeasibility certificate which can a-priori certify regions in which the non-convex security-constrained AC-OPF problem under uncertainty is infeasible. This exploits the following property of a convex relaxation: if a relaxation is infeasible for a given operating point, the original non-convex problem is also guaranteed to be infeasible for that operating point. The proposed infeasibility certificate has three components: First, we employ bound tightening to tighten both the QC relaxation and the input bounds; this better approximates the secure region, while also reducing the sample space. Second, we propose an infeasibility certificate based on separating hyperplanes. Third, we present an efficient algorithm to maximize the input region classified as infeasible.

1) Bound Tightening Algorithms: The tightness of the QC relaxation relies on the tightness of the envelopes used in (5) including the envelopes on cosine and sine terms. These in turn depend on the tightness of the bounds on the voltage magnitudes and angle differences. The goal of bound tightening is to iteratively tighten voltage magnitudes and angle differences, and, as a result, obtain a tighter relaxation. In the context of our work, the benefits of bound tightening are twofold: First, it tightens the QC relaxation, i.e., shrinks its feasible space, making the infeasibility certificate based on

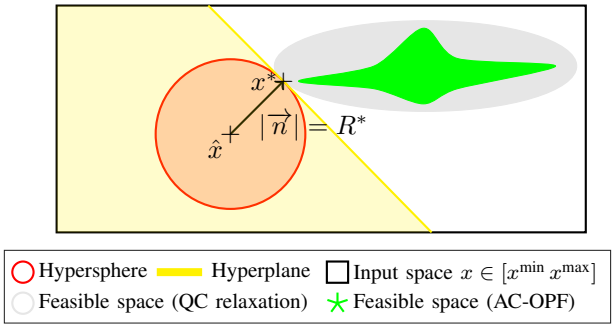

Fig. 1. Illustrative example of the differences between the infeasibility certificates using hyperspheres and hyperplanes. For a given infeasible point $\hat{x}$, the closest point x^* is computed which is feasible to the QC relaxation. The normal vector $\overrightarrow{n}$ is perpendicular to the feasible space of the QC relaxation. The feasible space of the non-convex AC-OPF problem is contained within that of the convex QC relaxation. All points inside the hypersphere or all points that are on the left side of the hyperplane are guaranteed to be infeasible with respect to both the QC relaxation and the non-convex AC-OPF problem, respectively. Note that the sets are not drawn to scale.

separating hyperplanes more effective, and second, it allows us to directly tighten the bounds on the input vector x.

We use two bound tightening algorithms from the literature: First, we rely on a computationally light-weight bound tightening technique for the branch angle differences $\theta_{ij}^{\max}$ in (1g) from [17]. Second, we use an optimization-based bound tightening algorithm from [18] which tightens the voltage magnitude bounds at each bus $V^{\max}, V^{\max}$ in (1a), and further tightens the angle differences for each line $\theta_{ij}^{\max}$ in (1g). For this purpose, we iteratively solve convex optimization problems to calculate the maximum and minimum values that the optimization variable under study, i.e., a voltage magnitude or a voltage angle difference, can obtain in the relaxed problem. Note that by tightening one variable bound, it may be possible to further tighten a previously tightened bound. This procedure can be executed for a defined number of iterations or until a fixed point is reached. As a final step in the bound tightening, we compute the tightened bounds for the input vector x, i.e., the bounds on active power of generators and uncertain injections. All inputs x which are outside the tightened minimum and maximum input bounds $x^{\text{BT,min}}$, $x^{\text{BT,max}}$ are guaranteed to be infeasible with respect to the non-convex AC-OPF problem. We calculate the volume of the remaining unclassified input space volume, normalized by the originally specified bounds on x:

$$V_{BT} = \prod_{k \in \mathcal{X}} \frac{x_k^{\text{BT,max}} - x_k^{\text{BT,min}}}{x_k^{\max} - x_k^{\min}} \tag{7}$$

The input set $\mathcal{X}$ is defined as $\mathcal{X} : \{\mathcal{G} \backslash \{\text{slack}\},\ k \in \mathcal{G},\ k \in \mathcal{U}\}$.

2) Separating Hyperplanes: We next propose an infeasibility certificate based on separating hyperplanes. Consider a particular operating point $\hat{x}$ that is infeasible with respect to the non-convex security-constrained AC-OPF. We solve the following optimization problem to compute the closest dispatch x^* which is feasible to the convex QC relaxation:

$$\min_{x,V,S_U,S_G,R} \quad R \tag{8a}$$

$$\text{s.t. (1b)–(1d), (2b), (4), (5), (6)} \tag{8b}$$

$$\sqrt{\sum_{k \in \mathcal{X}} (x_k - \hat{x}_k)^2} \leq R \tag{8c}$$

If the obtained radius R^* is greater than zero, i.e., the operating point $\hat{x}$ is infeasible with respect to the relaxation, no operating point x exists which is closer to $\hat{x}$ than the obtained point x^*. This property has been used in [8] to construct infeasibility certificates in the form of hyperspheres and ellipses by assigning different weights to the components in (8c). Here, we propose to use hyperplanes as infeasibility certificates in order to significantly enlarge the volume classified as infeasible:

Proposition 1: For a given infeasible point $\hat{x}$, if the solution to (8) yields a non-zero radius R^* and optimal solution x^*, all vectors x which fulfill the following criterion are infeasible with respect to the AC-OPF constraints (1) and (2):

$$\overrightarrow{n}^T (x - x^*) < 0 \tag{9}$$

The normal of the hyperplane is defined as $\overrightarrow{n} := x^* - \hat{x}$ and the operator T denotes the transpose.

Proof of Proposition 1: Proof by contradiction: Assume there exists a feasible point $\tilde{x}$ that is inside the region classified as infeasible by the hyperplane: $\overrightarrow{n}^T(\tilde{x} - x^*) < 0$. As the feasible space of optimization problem (8) is convex, it must hold that any linear combination between $\tilde{x}$ and x^* is also feasible: $\lambda\tilde{x} + (1-\lambda)x^*, \lambda \in [0,1]$. Then, there exists a point $\tilde{x}^* = \lambda^*\tilde{x} + (1-\lambda^*)x^*$ which has a radius $\tilde{R}^*$ to the initial infeasible point $\hat{x}$ that is smaller than R^*. Since the optimization problem (8) is convex, we obtained the globally optimal solution x^* with the smallest radius R^*. As a result, there cannot exist an input $\tilde{x}$ that has a smaller radius than R^*. We have shown by contradiction that there cannot exist a feasible point $\tilde{x}$ that is inside the region classified as infeasible by the hyperplane. The infeasibility certificate with respect to the non-convex AC-OPF problem (1) and (2) follows from the property that infeasibility with respect to the QC relaxation constraints (8b) is sufficient to ensure infeasibility with respect to (1) and (2).

An illustrative comparison of both infeasibility certificates is shown in Fig. 1. By solving the same optimization problem, it is evident that the infeasibility certificate based on hyperplanes is able to classify significantly larger spaces as infeasible. This is quantitatively analysed through simulation studies in Section IV-B.

3) An Efficient Algorithm to Minimize the Unclassified Input Space: Using the infeasibility certificate, we propose an efficient algorithm to maximize the portion of the input space that can be classified a-priori as infeasible. Our algorithm relies on an insight related to the hyperplanes: together with the initial input space restriction, subsequent hyperplanes form a convex polyhedron which can be described as $Ax \leq b$. We can write the row of A and entry in b corresponding to the hyperplane in (9) as $A_k := \overrightarrow{n}^T$ and $b_k := \overrightarrow{n}^T x^*$. Efficient methods to sample uniformly from inside a convex polyhedron are available, e.g., "Hit-and-Run" sampling [19].

Algorithm 1 Computing Infeasibility Certificates

Run bound tightening and obtain $x^{\text{BT,min}}$ and $x^{\text{BT,max}}$
Set iteration count: $k \leftarrow 0$
Initialize unclassified region $A^{(0)}x \leq b^{(0)}$:
  $A^{(0)} := [\mathbf{I}^{|x|\times|x|} \ -\mathbf{I}^{|x|\times|x|}]^T$
  $b^{(0)} := [(x^{\text{BT,max}})^T \ (x^{\text{BT,min}})^T]^T$
while $k \leq N_1$ **do**
  draw random $x^{(k)}$ from inside $A^{(k)}x \leq b^{(k)}$
  solve (8) with $\hat{x} := x^{(k)}$ and obtain x^*
  if $R > 0$ **then**
    reduce unclassified region by adding hyperplane:
      $A^{(k+1)} = [(A^{(k)})^T \ \overrightarrow{n}^T]^T$
      $b^{(k+1)} = [(b^{(k)})^T \ \overrightarrow{n}^T x^*]^T$
  end if
  $k \leftarrow k + 1$
end while.

This allows us to iteratively construct hyperplanes while sampling only inside the currently unclassified region. Thus, the hyperplane certificates facilitate a significant improvement on the "rejection" sampling approach used with hypersphere certificates in [8], [9].

The steps of the algorithm to compute infeasibility certificates are detailed in Algorithm 1. We start with a description of the convex polyhedron restricted to the tightened input bounds. We iteratively sample uniformly from inside the convex polyhedron and add identified hyperplanes until we reach an upper iteration limit of N_1 samples. This ensures that only samples which have not yet been classified as infeasible by previously added hyperplanes are considered in optimization problem (8). In Section IV-C, we will demonstrate the performance of this algorithm on a range of PGLib-OPF networks up to 500 buses by calculating the remaining unclassified volume as the volume of the convex polyhedron $A^{(N_1)}x \leq b^{(N_1)}$. This shows substantial reductions of unclassified input space volumes.

B. Security Boundary Identification

After computing the infeasibility certificates, we perform sampling and directed walks, similar to [9], to obtain a detailed description of the security boundary. For this purpose, we first uniformly draw a large number N_2 of samples from the convex polyhedron describing the remaining unclassified input region: $A^{(N_1)}x \leq b^{(N_1)}$. For each sample, we first run AC power flows for the intact and the outaged system states and check if any of the constraints in (1) are violated. If not, we add the current point to the dataset as a feasible point, otherwise as an infeasible point. If constraints are violated, we run additional AC power flows for which we enforce the reactive power limits of generators, i.e., if any generator violates its reactive power limit it is converted from a PV to a PQ bus in the power flow. This is based on the observation that reactive power limits are often the only constraints violated. If the obtained power flow solutions satisfy all constraints in (1), the point is added as feasible point to the dataset. Note that the voltage set-points of generators in x are updated accordingly. If both stages are not feasible, we solve the following non-convex optimization problem which computes the closest feasible dispatch to the non-convex AC-OPF problem in (1) and (2):

$$\min_{x,V,S_U,S_G,R} \quad R \tag{10a}$$

$$\text{s.t. } (1), (2a), (2b), (6), (8c) \tag{10b}$$

We add the obtained locally optimal point x^* to the dataset as feasible point. We repeat this procedure for all N_2 samples and obtain as a result a detailed security boundary description.

C. Sampling from Inside the Secure Space

To obtain a more detailed description of the entire secure space, we fit a multivariate normal distribution $\mathcal{N}$ to the feasible points obtained. To this end, we estimate both the mean μ and the covariance matrix Σ from the feasible data points. To bias the sampling towards inside the boundary, we reduce the magnitude of all entries of the covariance matrix, i.e., $\Sigma_{\text{red}} = s_{\text{red}} \cdot \Sigma$, by a constant scaling factor $s_{\text{red}} < 1$. We draw a large number, denoted N_3, of samples from $\mathcal{N}(\mu, \Sigma_{\text{red}})$. For each of these samples, we first run AC power flows for the intact and the outaged system states, check feasibility with respect to all AC-OPF constraints, and add the sample with the corresponding classification to the dataset. If the sample is infeasible, we run a second round of AC power flows in which we enforce the generators' reactive limits and again evaluate the feasibility with respect to all AC-OPF constraints. If the sample is feasible, we add it to the dataset with the generator voltage set-points adjusted accordingly. Our simulations indicate that sampling from a multivariate normal distribution $\mathcal{N}(\mu, \Sigma_{\text{red}})$ results in identification of feasible samples inside the secure space. We did not observe improvements by fitting a Gaussian mixture model.

IV. Simulation and Results

We analyse the performance of our proposed methodology for a range of test cases from the PGLib-OPF networks. First, we compare the proposed infeasibility certificate based on separating hyperplanes with the certificate based on hyperspheres from [8]. Second, we compute the volume of the unclassified input space using the infeasibility certificates and show substantial reductions. Third, we create balanced datasets and demonstrate their applicability using an illustrative data-driven application.

A. Simulation Setup

In the following, we first evaluate our proposed methods on 13 PGLib-OPF networks (v19.05) [20] up to 500 buses for which we do not consider N-1 security and uncertainty, i.e., we use the test cases as specified in [20]. Second, we use two test cases for which we include both N-1 security and uncertainty. We use *case39_epri* and *case162_ieee_dtc* with the following line contingencies $\mathcal{C} = \{0, 7, 22, 24, 36, 43\}$ and $\mathcal{C} = \{6, 8, 24, 50, 128\}$, respectively. We assume the same parameters for the outaged system state as for the intact system state. Furthermore, we place 3 wind farms with rated power of 500 MW and consider 3 uncertain loads with $\pm 50\%$ variability, i.e., a total of 6 uncertain power

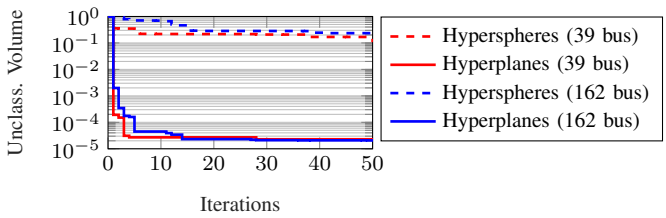

Fig. 2. For *case39_epri* and *case162_ieee_dtc*, we compare the remaining unclassified volume between an infeasibility certificate based on hyperspheres from [8] and the proposed certificate in Section III-A2 based on hyperplanes.

injections, at buses $\mathcal{U} = \{3, 21, 27, 4, 25, 28\}$ for *case39_epri* and $\mathcal{U} = \{60, 90, 145, 3, 8, 52\}$ for *case162_ieee_dtc*. For all uncertain injections, we assume a power factor $\cos\phi = 1$. Note that in the following, all inputs x are normalized with respect to their maximum and minimum limits as specified in [20], i.e., if x has dimension $|x|$, then $x \in [0, 1]^{|x|}$. This normalization step is standard practice for many data-driven applications including neural networks and improves performance [21]. For both AC power flow and AC optimal power flow computations, we rely on MATPOWER [22] with the IPOPT solver for AC-OPF problems [23]. For the bound tightening, we use the implementations in [17], [18]. Note that we only tighten the bounds of the intact system state, and we run the optimization-based bound tightening for up to three iterations. Extension of these toolboxes to the full N-1 case is a direction for future work. To solve the QC relaxation we use MOSEK [24]. To approximate the volumes of the convex polyhedrons describing the remaining unclassified input space, we use a volume approximation toolbox in C++ [25]. Note that an exact volume computation is considered intractable for dimensionality 10 or higher [25]. The relative approximation error threshold is set to be less than one order of magnitude, which is sufficiently accurate for our purposes since we compute volumes of spaces several orders of magnitudes smaller than the initial volume.

B. Comparison of Infeasiblity Certificates

We compare the infeasibility certificate based on hyperspheres proposed in [8] with the infeasibility certificate based on hyperplanes proposed in Section III-A2. The main metric for comparison is the volume of the remaining unclassified space after applying the infeasibility certificates. We consider *case39_epri* and *case162_ieee_dtc*. We assume that both algorithms rely on the QC relaxation, all the bounds have been tightened as described in Section III-A1, we only consider the active power generation in the input variables x, and we do not consider N-1 security or uncertainty. For the certificate based on hyperplanes, we follow the algorithm outlined in Algorithm 1. For the certificate based on hyperspheres, we assume that in each iteration we draw a random sample from the entire input space, and if it is infeasible we compute the closest feasible input x by solving (8). If the distance is non-zero, we have obtained an infeasibility certificate. To estimate the volume of the unclassified space not covered by the hyperspheres we use Monte Carlo sampling with 10^6 samples. Fig. 2 shows the variation of the unclassified input volume with up to 50 iterations. It can be observed, first, that the hyperplane certificates shrink the unclassified region by four orders of magnitude more than the hyperspheres, i.e., the unclassified volumes evaluate to 10^{-5} versus 10^{-1} compared to the initial unit hypercube's normalized volume of 1. Second, the algorithm using hyperplanes requires significantly fewer iterations. After the first iteration, the hyperplanes classify a substantially larger space as infeasible than the hyperspheres after 50 iterations. The reasons for this are twofold: First, as evident in Fig. 1, certificates based on separating hyperplanes cover a larger volume than hyperspheres for the same sample and second, the hyperplanes enable the use of efficient methods for sampling uniformly from inside the associated convex polyhedron [19], [25].

TABLE I
UNCLASSIFIED INPUT VOLUMES FOR PGLIB-OPF NETWORKS

case	$\lvert x\rvert$	V^{BT}	$\lvert HP\rvert$	V^{HP}	$\frac{-\log_{10}(V)}{\lvert x\rvert}$
AC-OPF without N-1 security and without uncertainty					
case3_lmbd	4	6.3e-02	28	3.3e-02	37.0%
case5_pjm	7	1.0e+00	99	6.9e-03	30.9%
case14_ieee	6	2.4e-01	54	6.9e-04	52.7%
case24_ieee_rts	20	9.2e-01	184	2.3e-06	28.2%
case30_ieee	7	6.2e-03	61	8.8e-06	72.2%
case39_epri	19	9.9e-02	203	7.0e-08	37.7%
case57_ieee	10	3.8e-02	231	4.9e-06	53.1%
case73_ieee_rts	62	1.0e+00	608	6.1e-16	24.5%
case118_ieee	72	1.7e-02	1000	1.6e-17	23.3%
case162_ieee_dtc	23	6.1e-04	371	1.5e-11	47.1%
case200_tamu	69	9.3e-01	1000	6.0e-11	14.8%
case300_ieee	125	1.0e-12	1000	3.4e-40	31.6%
case500_tamu	111	8.6e-02	1000	5.4e-26	22.8%
AC-OPF considering N-1 security and uncertainty					
case39_epri	25	2.6e-01	271	2.0e-05	18.8%
case162_ieee_dtc	29	2.2e-04	394	6.0e-10	31.8%
Median all cases	23	8.6e-02	271	7.0e-08	31.6%

C. Estimating Unclassified Volumes for PGLib-OPF Networks

In the following, we compute infeasibility certificates and the volume of the remaining unclassified input space for a range of test cases. For this purpose, we run Algorithm 1 with the number of iterations N_1 set to 1000. We evaluate the remaining estimated volume for the bound tightening V^{BT} according to (7) and for the separating hyperplanes described as a convex polyhedron by running the volume approximation algorithm in [25]. In Table I, the dimensionality $|x|$, the reduced unclassified volume V^{BT} after bound tightening, the number of hyperplanes $|HP|$, and the reduced unclassified volume V^{HP} enclosed by the separating hyperplanes is listed. Note that both volumes are defined with respect to the unit hypercube $x \in [0\,1]^{|x|}$ normalized by the original power system limits with volume 1. We can make several observations. First, the bound tightening results in a moderate reduction in input dimensionality of several orders of magnitude (10^{-1} to 10^{-4})

TABLE II
CREATED DATASETS FOR AC-OPF PROBLEMS

Power system case	Boundary $N_2 = 10^4$	Inside (MVND) $N_3 = 10^5$	Overall Secure	Overall Points
AC-OPF without N-1 security and without uncertainty				
case3_lmbd	69.5%	36.5%	40.6%	114'389
case5_pjm	68.6%	69.4%	69.3%	125'432
case14_ieee	73.3%	59.0%	61.0%	147'047
case24_ieee_rts	66.8%	44.3%	48.7%	131'158
case30_ieee	75.0%	50.2%	54.0%	124'944
case39_epri	57.2%	29.9%	33.9%	154'635
case57_ieee	58.9%	35.2%	38.9%	150'865
case73_ieee_rts	63.9%	51.1%	52.7%	222'730
case118_ieee	53.2%	47.0%	47.6%	209'996
case162_ieee_dtc	50.0%	40.1%	41.7%	129'165
case200_tamu	50.2%	36.6%	38.1%	177'023
case300_ieee	50.0%	32.6%	34.7%	163'087
case500_tamu	50.0%	35.4%	37.1%	174'774
AC-OPF considering N-1 security and uncertainty				
case39_epri	58.2%	78.2%	75.2%	139'756
case162_ieee_dtc	50.0%	17.9%	23.2%	121'358
Average all cases	59.7%	44.2%	46.5%	152'424

for most test cases. Second, the infeasibility certificates based on hyperplanes enable further substantial reductions in the unclassified volume. As a result, the total unclassified volume compared to the unit hypercube is reduced between 2 and 40 orders of magnitude (10^{-2} to 10^{-40}). The median of the unclassified volume is 10^{-8}. This means that in order to identify one sample inside the unclassified volume, we would have to uniformly draw 10^8 samples from the original bounds on the input x. This highlights the necessity of first computing the infeasibility certificates to be able to identify the secure space. The number of hyperplanes is below 1000 for most test cases, indicating that Algorithm 1 has obtained a good estimation of the unclassified volume. For the four test cases for which 1000 hyperplanes are added, the unclassified input volume could be further reduced by increasing N_1.

To allow for comparability between test cases, we propose to use a metric defined as $\frac{-\log_{10}(V)}{|x|}$. The metric is motivated as follows: If one wants to sample 10 steps in each dimension, i.e., $10^{|x|}$, then this metric quantifies by how much the exponent is reduced. Note that the value obtained in percent is not the dimensionality reduction itself but relates to the reduction in the orders of magnitudes of the dimensionality. This value is between 14.8% and 72.2% for all test cases, showcasing the general applicability of the proposed infeasibility certificate for AC-OPF problems.

D. Dataset Creation for PGLib-OPF Networks

We create datasets of operating points classified based on their feasibility with respect to AC-OPF problems including N-1 security and uncertainty. To this end, we first draw a number of samples $N_2 = 10^4$ from the inside of the remaining unclassified volume described in Table I and obtain a detailed security boundary description following the approach in Section III-B. We fit a multivariate normal distribution with $s_{\text{red}} = 0.25$ and classify $N_3 = 10^5$ samples as secure or insecure following the approach in Section III-C. In Table II, we list the characteristics of the obtained datasets. First, note that in the boundary identification stage, if the percentage of secure points is above 50%, then sampling directly from the remaining unclassified volume results in identifying secure operating points. This is the case for the majority of test cases, demonstrating that the infeasibility certificate is able to provide a tight approximation of the secure spaces of non-convex AC-OPF problems. For the test cases where the sampling did not find any secure samples, the number of iterations for the feasibility certificate could be enlarged or other relaxations such as moment-based relaxations described in [7] could be used to further reduce the unclassified space in Algorithm 1. Second, the results show that sampling from a multivariate normal distribution fitted to the boundary samples results in identification of a large number of secure samples. The resulting datasets are well balanced with on average 46.5% secure samples. Note that this is an important metric for the successful application of data-driven methods such as neural networks [1]. The number of overall points is dependent on the number of feasible samples identified by enforcing the generators' reactive power limits in the AC power flows and differs between the test cases. Regarding the computational tractability, all simulations were carried out on a laptop and the dataset creation for the largest test cases took a few hours, with the most computationally intense task being the AC-OPF evaluations in the boundary identification and the optimization-based bound tightening [18]. By using high-performance computing and parallelizing both the boundary identification and the AC power flow computations, we expect that our approach can scale to systems with thousands of buses.

TABLE III
TEST SET ACCURACY OF NEURAL NETWORK CLASSIFIERS

case	full dataset	only boundary
AC-OPF without N-1 security and without uncertainty		
case14_ieee	78.2%	60.5%
case39_epri	74.6%	38.5%
case162_ieee_dtc	84.4%	49.8%
AC-OPF including N-1 security and uncertainty		
case39_epri	81.0%	80.4%
case162_ieee_dtc	93.4%	31.9%

E. Training Neural Network Classifiers

As an illustrative data-driven application, we evaluate the performance of a neural network classifier trained on several of the created datasets. The neural network predicts a binary classification, i.e., whether the input x belongs to the class "secure" or "insecure". We choose neural network structures with five hidden layers with the numbers of neurons of each hidden layer selected to be 10 times the input dimension $|x|$.

We split the dataset into a training set consisting of 85% of all samples and a test set of the remaining 15%. Note that the classifier has no information of the test set during training, and its performance is evaluated on the test set only. This gives a metric for how well the classifier generalizes to unseen data. We train the neural networks using TensorFlow [21] with standard training parameters and 250 epochs. Table III shows the test set accuracy, i.e., the share of correctly predicted labels for the test set. First, we use 85% of the full dataset for training and 15% of the full dataset for testing. Second, we only use the boundary samples from Section III-B as training data and then test on 15% of the full dataset. This gives us an estimation of the benefit of the additional sampling from the fitted multivariate distribution in Section III-C. We observe that the neural network classifier is able to generalize from the training to the test set and achieve high accuracy when using the full dataset. To further increase the classification accuracy, deeper neural networks or a deep autoencoder to identify lower-dimensional features could be used. We also observe that only relying on the boundary samples for prediction is not sufficient for most test cases, higlighting the importance of obtaining a representative dataset.

V. Conclusion

Successful application of data-driven methods in power systems requires datasets of sufficient size, covering a wide range of operating points. Creating a dataset that characterizes the security boundary and sufficiently covers both secure and insecure operating regions is a highly computationally demanding task, even for medium-sized systems, as we showed in [9]. In this paper, we propose an efficient method to create such datasets. We focus on AC-OPF feasibility and N-1 security, as any operating point should first satisfy static security criteria. Future work will extend this to include dynamic security criteria, similar to [9]. We develop an infeasibility certificate based on separating hyperplanes which is able to classify large portions of the input space as insecure. We show that the infeasibility certificates reduce the unclassified input space volumes significantly, by up to 10^{-40} compared to an initial normalized input space volume of 1 (i.e., 10^{0}) based on defined control variable bounds. Although the secure operating region is a very small portion of the original input space, our method is able to produce balanced datasets of secure and insecure operating points, a property desired for successful applications of data-driven methods. As an illustrative application, we used the generated datasets to assess the performance of neural network classifiers. Future work is directed towards (i) utilizing convex restrictions from [26], [27] to characterize secure spaces and (ii) exploiting high-performance computing.

References

[1] L. A. Wehenkel, *Automatic learning techniques in power systems*. Springer Science & Business Media, 2012.

[2] B. Donnot *et al.*, "Fast power system security analysis with guided dropout," *arXiv:1801.09870*, Jan. 2018.

[3] J. L. Cremer *et al.*, "Data-driven power system operation: Exploring the balance between cost and risk," *IEEE Transactions on Power Systems*, vol. 34, no. 1, pp. 791–801, Jan. 2019.

[4] L. Halilbasic *et al.*, "Data-driven security-constrained AC-OPF for operations and markets," in *20th Power Systems Computation Conference (PSCC)*, June 2018.

[5] J.-M. H. Arteaga, F. Hancharou, F. Thams, and S. Chatzivasileiadis, "Deep learning for power system security assessment," in *13th IEEE PowerTech 2019*, June 2019.

[6] L. Duchesne, E. Karangelos, and L. Wehenkel, "Using machine learning to enable probabilistic reliability assessment in operation planning," in *2018 Power Systems Computation Conference*, 2018, pp. 1–8.

[7] D. K. Molzahn and I. A. Hiskens, "A Survey of Relaxations and Approximations of the Power Flow Equations," *Foundations and Trends in Electric Energy Systems*, vol. 4, no. 1-2, pp. 1–221, Feb. 2019.

[8] D. K. Molzahn, "Computing the feasible spaces of optimal power flow problems," *IEEE Transactions on Power Systems*, vol. 32, no. 6, pp. 4752–4763, Nov. 2017.

[9] F. Thams, A. Venzke, R. Eriksson, and S. Chatzivasileiadis, "Efficient database generation for data-driven security assessment of power systems," to appear in *IEEE Transactions on Power Systems*, 2019.

[10] V. Krishnan, J. D. McCalley, S. Henry, and S. Issad, "Efficient database generation for decision tree based power system security assessment," *IEEE Transactions on Power Systems*, vol. 26, no. 4, pp. 2319–2327, Nov. 2011.

[11] C. Hamon, M. Perninge, and L. Söder, "An importance sampling technique for probabilistic security assessment in power systems with large amounts of wind power," *Electric Power Systems Research*, vol. 131, pp. 11–18, 2016.

[12] Mingyang Sun, I. Konstantelos, S. Tindemans, and G. Strbac, "Evaluating composite approaches to modelling high-dimensional stochastic variables in power systems," in *19th Power Systems Computation Conference (PSCC)*, June 2016.

[13] I. Konstantelos, M. Sun, S. H. Tindemans, S. Issad, P. Panciatici, and G. Strbac, "Using vine copulas to generate representative system states for machine learning," *IEEE Transactions on Power Systems*, vol. 34, no. 1, pp. 225–235, Jan 2019.

[14] M. B. Cain, R. P. Oneill, and A. Castillo, "History of optimal power flow and formulations (OPF Paper 1)," *Federal Energy Regulatory Commission*, vol. 1, pp. 1–36, 2012.

[15] C. Coffrin, H. L. Hijazi, and P. Van Hentenryck, "The QC relaxation: A theoretical and computational study on optimal power flow," *IEEE Transactions on Power Systems*, vol. 31, no. 4, pp. 3008–3018, 2016.

[16] J. Lavaei and S. H. Low, "Zero duality gap in optimal power flow problem," *IEEE Transactions on Power Systems*, vol. 27, no. 1, pp. 92–107, Feb 2012.

[17] D. Shchetinin, T. T. De Rubira, and G. Hug, "Efficient bound tightening techniques for convex relaxations of AC optimal power flow," *IEEE Transactions on Power Systems*, vol. 34, no. 5, pp. 3848–3857, 2019.

[18] K. Sundar *et al.*, "Optimization-based bound tightening using a strengthened QC-relaxation of the optimal power flow problem," *arXiv:1809.04565*, 2018.

[19] D. P. Kroese, T. Taimre, and Z. I. Botev, *Handbook of Monte Carlo methods*. John Wiley & Sons, 2013, vol. 706.

[20] IEEE PES Task Force on Benchmarks for Validation of Emerging Power System Algorithms, "The power grid library for benchmarking AC optimal power flow algorithms," *arXiv:1908.02788*, Aug. 2019. [Online]. Available: https://github.com/power-grid-lib/pglib-opf

[21] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[22] 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 Transactions on Power Systems*, vol. 26, no. 1, pp. 12–19, 2010.

[23] A. Wächter and L. T. Biegler, "On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming," *Mathematical Programming*, vol. 106, no. 1, pp. 25–57, 2006.

[24] MOSEK ApS, *MOSEK 8.0.0.37* , 2018.

[25] I. Z. Emiris and V. Fisikopoulos, "Practical polytope volume approximation," *ACM Transactions on Mathematical Software (TOMS)*, vol. 44, no. 4, p. 38, 2018.

[26] D. Lee, H. D. Nguyen, K. Dvijotham, and K. Turitsyn, "Convex restriction of power flow feasibility sets," *IEEE Transactions on Control of Network Systems*, vol. 6, no. 3, pp. 1235–1245, 2019.

[27] B. Cui and X. A. Sun, "Solvability of power flow equations through existence and uniqueness of complex fixed point," *arXiv:1904.08855*, April 2019.