Differentially Private Algorithms for Synthetic Power System Datasets

Vladimir Dvorkin, Jr., *Member, IEEE*, and Audun Botterud, *Member, IEEE*

***Abstract*— While power systems research relies on the availability of real-world network datasets, data owners (e.g., system operators) are hesitant to share data due to privacy risks. To control these risks, we develop privacy-preserving algorithms for synthetic generation of optimization and machine learning datasets. Taking a real-world dataset as input, the algorithms output its noisy, synthetic version, which preserves the accuracy of the real data on a specific downstream model or even a large population of those. We control the privacy loss using Laplace and Exponential mechanisms of differential privacy and preserve data accuracy using a post-processing convex (or mixed-integer) optimization. We apply the algorithms to generate synthetic network parameters and wind power data.**

***Index Terms*— Differential privacy, machine learning, power systems optimization, synthetic datasets**

I. Introduction

POWER system datasets are instrumental for enhancing solutions to many problems, including optimal power flow (OPF) and wind power forecasting. Releasing real data, however, is challenging due to security and privacy concerns. Indeed, detailed network datasets inform cyberattacks on SCADA systems and can be used by strategic market players to maximize profits at the expense of deteriorating social welfare. These concerns motivate producing synthetic datasets – a sanitized version of private datasets that approximately preserve accuracy of data for power system applications.

Differential privacy (DP) is an algorithmic notion of privacy preservation that enables trade-offs between data privacy and accuracy in optimization [1] and machine learning [2]. It has also found applications in the context of privacy-preserving OPF computations, e.g., in distributed algorithms [3] and centralized solvers for power grids [4], [5], as well as in machine learning problems specific to power systems [6]. Models in [3]– [6], however, only control data leakages in computations and do not provide synthetic datasets per se.

Producing synthetic datasets in a DP way is achieved by corrupting data with privacy-preserving noise [7], [8]. However, applications of the standard noise-additive DP mechanisms in power systems, such as the Laplace mechanism, may no longer admit a meaningful result. Indeed, adding noise to data may fundamentally alter important statistics and trends in machine learning datasets [9]. In the OPF context, [10] and [11] showed that the Laplacian perturbation of network data almost surely violates OPF feasibility. As a remedy, they proposed an optimization-based post-processing which restores the accuracy of synthetic OPF datasets without altering the privacy guarantee. The proposed restoration, however, renders the synthetic dataset feasible only for a particular OPF model. Repeated applications of the Laplace mechanism to restore accuracy on many OPF models (e.g., for different instances of variable renewable production) may not be possible, as noise must be scaled drastically as per composition of DP [12].

In this paper, we introduce private synthetic dataset generation algorithms for power systems, which ensure the accuracy of synthetic datasets for downstream models. They enjoy known DP mechanisms and convex (or mixed-integer) post-processing optimization of data. Specifically, we develop:

1) Wind power obfuscation (WPO) algorithm which privately generates wind power measurements, while guaranteeing DP of the real data and ensuring accuracy in terms of the outcomes of a regression analysis.
2) Transmission capacity obfuscation (TCO) algorithm, which generates synthetic line parameters, while ensuring their feasibility and cost-consistency on a population of OPF models. Here, we use both Laplace and Exponential mechanisms of DP to substantially reduce the noise compared to using the Laplace mechanism alone.

Next section reviews the basic DP results. In Sections III and IV we present the two algorithms and their theoretical properties. Section V provides numerical experiments, and Section VI concludes. Proofs are relegated to the Appendix.

Notation: I is an identity matrix, e_i is the basis vector with element 1 at position i. Schur product is denoted by $\circ$. By $\|\cdot\|_1$ and $\|\cdot\|$ we denote L_1 and L_2 norms, respectively.

II. Preliminaries on Differential Privacy

This section reviews basic DP mechanisms used as building blocks for our synthetic dataset generation algorithms.

Consider a vector $y \in \mathcal{Y} \subseteq \mathbb{R}^n$, with n private records from universe $\mathcal{Y}$, and a query $Q : \mathcal{Y} \mapsto \mathcal{R}$ as a mapping from universe $\mathcal{Y}$ to range $\mathcal{R}$. Queries of interest include simple numerical queries, i.e., identity $Q(y) = y$, and optimization and ML queries, such as OPF or regression models. The goal is to make *adjacent* vectors $y, y' \in \mathcal{Y}$ of private records, statistically indistinguishable in query answers.

Definition 1 (Adjacency [13]): Vectors $y, y' \in \mathcal{Y}$ are said to be $\alpha-$adjacent, denoted as $y \sim_\alpha y'$, if $\exists i \in 1, \dots, n$, s.t. $y_j = y'_j, \forall j \in \{1, \dots, n\} \backslash i$, and $|y_i - y'_i| \leqslant \alpha$ for $\alpha > 0$.

Vladimir Dvorkin and Audun Botterud are with the Laboratory for Information & Decision Systems, Massachusetts Institute of Technology (MIT), Cambridge, MA 02139, USA. Vladimir Dvorkin is also with the MIT Energy Initiative. {dvorkin,audunb}@mit.edu

This work is supported by the Marie Sklodowska-Curie Actions and Iberdrola Group, Grant №101034297 – project Learning ORDER.

A statistical similarity of $\alpha-$adjacent datasets in query answers is captured by the notion of differential privacy.

Definition 2 ($\varepsilon-$differential privacy [12]): Random query $\tilde{Q} : \mathcal{Y} \mapsto \mathcal{R}$ is $\varepsilon-$DP if, for any output $r \subseteq \mathcal{R}$ and any $\alpha-$adjacent vectors $y, y' \in \mathcal{Y}$, the following ratio holds

$$\frac{\Pr[\tilde{Q}(y\ ) = r]}{\Pr[\tilde{Q}(y') = r]} \leqslant \exp(\varepsilon). \tag{1}$$

where probability is with respect to the randomness of $\tilde{Q}$.

Privacy parameter $\varepsilon > 0$ is termed *privacy loss*: smaller ε provides stronger privacy protection, i.e., for small ε we have $\exp(\varepsilon) \approx 1 + \varepsilon$, thereby making two adjacent datasets y and y' statistically similar in the random query answer.

Theorem 1 (Composition [12]): A series $\tilde{Q}_1(y), ..., \tilde{Q}_k(y)$ of ε_i-DP queries on dataset y satisfies $\sum_{i=1}^{k} \varepsilon_i-$DP.

Theorem 2 (Post-processing immunity [12]): If query $\tilde{Q}$ satisfies ε-DP, then $g \circ \tilde{Q}(y)$, where g is any data-independent post-processing of the query answer, also satisfies ε-DP.

The former bounds the privacy loss over multiple queries, and the latter states that any data-independent transformation of a DP query answer preserves the privacy guarantee.

A numerical query is made DP by adding random noise to its output. The noise magnitude depends on the worst-case sensitivity δ_Q of query Q to adjacent datasets, i.e.,

$$\delta_Q = \max_{y \sim_\alpha y'} \|Q(y) - Q(y')\|_1 .$$

Let $\text{Lap}(\lambda)^k$ be a sample from the $k-$dimensional Laplace distribution with zero mean and scale parameter λ. DP of a numerical query is then achieved with the following result.

Theorem 3 (Laplace mechanism [14]): Let Q be a query that maps datasets to $\mathbb{R}^k$. Then, the Laplace mechanism which outputs $Q(y) + \text{Lap}(\delta_Q/\varepsilon)^k$ achieves $\varepsilon-$DP.

We also like to limit privacy losses when answering non-numerical queries. For example, given a population $\mathcal{Q}$ of queries, we would like to answer the question: *which query $Q \in \mathcal{Q}$ gives the maximum value on a private dataset y?* The Exponential mechanism answers this question privately.

Theorem 4 (Exponential mechanism [15]): For a query population $\mathcal{Q}$ and a score function $u : \mathcal{Y} \times \mathcal{Q} \mapsto \mathbb{R}$ with sensitivity δ_u, the Exponential mechanism, which outputs query $Q \in \mathcal{Q}$ proportionally to $\exp\left(\frac{\varepsilon u(y,Q)}{2\delta_u}\right)$, attains $\varepsilon-$DP.

For discrete populations, i.e., $\mathcal{Q} = Q_1, \ldots, Q_m$, we can adopt the report-noisy-max algorithm [12, §3.3] – an efficient implementation of the exponential mechanism for finite $\mathcal{Q}$.

Next, we leverage these results to design DP algorithms for synthetic dataset generation as applicable to power systems.

III. Privacy-Preserving Wind Power Data Release

Consider the problem of a wind turbine operator (data owner) who releases synthetic wind power records in a DP way. The real dataset $\mathcal{D} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ consists of m records, where each record i includes some public weather data $x_i \in \mathbb{R}^n$ and a private power measurement $y_i \in \mathbb{R}$ subject to obfuscation. The release of the synthetic dataset takes the form $\tilde{\mathcal{D}} = \{(x_1, \tilde{y}_1), \ldots, (x_m, \tilde{y}_m)\}$, where $\tilde{y}_i$ is a synthetic measurement. To provide formal privacy guarantees in this release, we could perturb each real record y_i with the Laplace mechanism of Theorem 3. However, the application of the Laplace mechanism alone is ignorant of the accuracy of the resulting dataset in the downstream analysis, and such a release may not be useful. We discuss the dataset accuracy in terms of the outcomes of a regression (downstream) problem

$$\underset{\beta}{\text{minimize}} \quad \|X\beta - y\| + \lambda \|\beta\| , \tag{2}$$

which minimizes the loss function by optimally choosing regression weights $\beta \in \mathbb{R}^p$, given some small regularization parameter λ to prevent overfitting. Here, matrix $X^{m \times p}$ collects weather features; we do not require $p = n$, as model (2) may not include all meteorological data from $\mathcal{D}$ and may also enjoy certain feature transformations (e.g., squared wind speeds). The goal is thus to release a synthetic dataset $\tilde{\mathcal{D}}$ whose regression loss and weights are consistent with those on the real dataset. On a particular vector of measurements $\overline{y}$, we denote the regression loss and weights by $\ell(\overline{y})$ and $\beta(\overline{y})$, respectively. To estimate them on the real dataset privately, we need to bound their sensitivities to adjacent datasets.

Lemma 1 (Regression sensitivity bounds): For any two α-adjacent vectors of power measurements $y, y' \in \mathbb{R}^m$, the worst-case sensitivity of regression weights is bounded as

$$\delta_\beta = \max_{y \sim_\alpha y'} \|\beta(y) - \beta(y')\|_1 \leqslant \left\|(X^\top X + \lambda I)^{-1} X^\top\right\|_1 \alpha,$$

and the worst-case sensitivity of the regression loss

$$\delta_\ell = \max_{y \sim_\alpha y'} \|\ell(y) - \ell(y')\|_1$$

is bounded by the solution of the following problem:

$$\delta_\ell \leqslant \underset{i=1,\ldots,m}{\text{maximize}} \left\|(X(X^\top X + \lambda I)^{-1} X^\top - I)(e_i \circ \alpha)\right\| .$$

Importantly, the two bounds only depend on public data, i.e., features, regularization and adjacency parameters, and completely independent from private measurements y.

Algorithm 1: Differentially private WPO

Input : WP records $\mathcal{D} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$;
DP param. $\varepsilon_1, \varepsilon_2, \alpha$; regularization param. γ_β, γ_y

Output: Synthetic WP records $\tilde{\mathcal{D}} = \{(x_1, \tilde{y}_1), \ldots, (x_m, \tilde{y}_m)\}$

1 Initialize synthetic measurements $\tilde{y}^0 = y + \text{Lap}\,(\alpha/\varepsilon_1)^m$

2 Laplace mechanism to privately compute regression results

$$\overline{\ell} = \ell(y) + \text{Lap}\,(\delta_\ell/\varepsilon_2) \quad \overline{\beta} = \beta(y) + \text{Lap}\,(\delta_\beta/\varepsilon_2)$$

3 Post-processing optimization of $\tilde{y}^0$:

$$\tilde{y} \in \underset{\tilde{y}}{\text{argmin}} \ \left\|\overline{\ell} - \ell\right\| + \gamma_\beta \left\|\overline{\beta} - \beta\right\| + \gamma_y \left\|\tilde{y}^0 - \tilde{y}\right\| \tag{3a}$$

$$\text{subject to} \quad \mathbb{0} \leqslant \tilde{y} \leqslant \mathbb{1}, \tag{3b}$$

$$\beta \in \underset{\beta}{\text{argmin}} \ \|X\beta - \tilde{y}\| + \lambda \|\beta\| \tag{3c}$$

return: synthetic wind power measurements $\tilde{y}$

A. Differentially Private WPO Algorithm

We now introduce the privacy-preserving wind power obfuscation (WPO) Algorithm 1. The algorithm takes the real dataset, privacy and regularization parameters as inputs, and

produces a consistent synthetic dataset of wind power records. It relies on Lemma 1 to privately reveal regression results on a real dataset, and then leverages them to restore the consistency of the synthetic dataset using a post-processing optimization. Specifically, at Step 1, the algorithm initializes the synthetic datasets using the Laplace mechanism. Then Step 2 computes the approximate regression loss and weights using the Laplace mechanism twice. At the last Step 3, the synthetic dataset is post-processed using optimization to ensure the consistency of regression results on the synthetic and real datasets.

The post-processing is based on the hierarchical optimization (3), where the upper-level problem (3a)–(3b) optimizes the synthetic dataset $\tilde{y}$ in response to the outcomes of the embedded lower-level regression problem (3c). In the upper-level objective, the first term improves the consistency in terms of regression loss, while the second and third terms are used for regularizing the synthetic dataset. Indeed, the losses l and $\bar{l}$ can be matched with infinitely many assignments of β and $\tilde{y}$. Thus, by setting a small parameter $\gamma_\beta > 0$, the matching is achieved with a close approximation of the regression weights on the real data. Similarly, by setting a small parameter $\gamma_y > 0$, we regularize the new data points according to the perturbation of real data points at Step 1. Finally, constraint (3b) respects the nominal power limits.

While the hierarchical optimization (3) is originally intractable, we arrive at its tractable convex reformulation by substituting the lower-level problem (3c) with

$$\beta = (X^\top X + \lambda I)^{-1} X^\top \tilde{y}, \tag{4a}$$

$$\|X\beta - \tilde{y}\| \leqslant \ell, \tag{4b}$$

where the equation (4a) is the closed-form solution to regression weights, and the conic constraint (4b) computes the regression loss. This reformulation results in a convex quadratic program with polynomial-time complexity [16].

Theorem 5 (DP of the WPO Algorithm): Setting $\varepsilon_1 = \varepsilon/2$ and $\varepsilon_2 = \varepsilon/4$ renders Algorithm 1 $\varepsilon-$DP for $\alpha-$adjacent wind power datasets.

IV. Privacy-Preserving DC-OPF Data Release

We now consider a problem of releasing a synthetic network dataset in a DP way. The goal is to guarantee both privacy and accuracy with respect to possible downstream computations on the synthetic dataset. We consider the DC-OPF problem as the main computational task. Although we specifically focus on the release of transmission capacity data, other network parameters, such as electrical loads, can be released similarly.

In a network with n buses and e transmission lines, the OPF problem seeks the least-cost generation dispatch $p \in \mathbb{R}^n$ which satisfies loads $d \in \mathbb{R}^n_+$, generation limits in set $\mathcal{P} = \{p \mid \underline{p} \leqslant p \leqslant \overline{p}\}$, and line capacities $\overline{f} \in \mathbb{R}^e_+$. Generators produce at linear costs $c \in \mathbb{R}^n_+$. The DC power flows are modeled using the power transfer distribution matrix $F \in \mathbb{R}^{e\times n}$, such that the resulting power flows are $\varphi = F(p-d) \in \mathbb{R}^e$.

Suppose that there is a set $1, \ldots, m$ of OPF models, where each model i includes specific costs c_i, generation limits $\mathcal{P}_i$, and loads d_i. The transmission data, i.e., topology encoded in F and capacity $\overline{f}$, remain the same. Each OPF model i is then described by a tuple $\langle c_i, d_i, \mathcal{P}_i, F, \overline{f}\rangle$. Given the real OPF dataset $\langle c_i, d_i, \mathcal{P}_i, F, \overline{f}\rangle_{i=1}^m$, the goal is to produce its synthetic version $\langle c_i, d_i, \mathcal{P}_i, F, \overline{\varphi}\rangle_{i=1}^m$ with an obfuscated transmission capacity vector $\overline{\varphi}$, which permits feasible and cost-consistent – with respect to real data – OPF outcomes across m models.

Towards the goal, we formulate a DC-OPF problem parameterized by the synthetic transmission capacity $\overline{\varphi}$:

$$\mathcal{C}_i(\overline{\varphi}) = \underset{p\in\mathcal{P}_i}{\text{minimize}} \quad c_i^\top p \tag{5a}$$

$$\text{subject to} \quad \mathbb{1}^\top(p - d_i) = 0, \tag{5b}$$

$$\|F(p-d_i)\|_1 \leqslant \overline{\varphi}, \tag{5c}$$

where function (5a) is to minimize the OPF costs, denoted by $\mathcal{C}_i(\overline{\varphi})$, subject to power balance (5b), flow and generation limits in (5c) and $\mathcal{P}_i$, respectively; all specific to a particular model i. We make two assumptions on problem (5).

Assumption 1 (Feasibility): $\mathcal{C}_i(\overline{f})$ exists $\forall i = 1, \ldots, m$.

Assumption 2 (Sensitivity): For any two $\overline{\varphi}_1 \sim_\alpha \overline{\varphi}_2$ capacity vectors, $\|\mathcal{C}_i(\overline{\varphi}_1) - \mathcal{C}_i(\overline{\varphi}_2)\|_1 \leqslant \overline{c}\alpha, \forall i = 1, \ldots, m$, where $\overline{c}$ is the maximum cost coefficient.

The former requires OPF feasibility of the real data across historical records, and the latter bounds the change in OPF costs to the cost of the most expensive unit.

As a perturbed capacity vector may not be OPF feasible, we additionally introduce the relaxed OPF problem to give a numerical value to infeasibility of a particular vector $\overline{\varphi}$:

$$\mathcal{C}_i^R(\overline{\varphi}) = \underset{p\in\mathcal{P}_i, v\geqslant\mathbb{0}}{\text{minimize}} \quad c_i^\top p + \psi\mathbb{1}^\top v \tag{6a}$$

$$\text{subject to} \quad \mathbb{1}^\top(p - d_i) = 0, \tag{6b}$$

$$\|F(p-d_i)\|_1 \leqslant \overline{\varphi} + v, \tag{6c}$$

where the slack variable $v \in \mathbb{R}^e$ renders the OPF solution feasible for any assignment $\overline{\varphi}$ using penalty scalar $\psi \gg \overline{c}$.

A. Differentially Private TCO Algorithm

We now introduce the privacy-preserving transmission capacity obfuscation (TCO) Algorithm 2 for DC-OPF datasets, where Step 1 initializes synthetic dataset $\overline{\varphi}^0$ by perturbing real data using the Laplace mechanism, and the remaining steps post-process the synthetic dataset. Step 2 runs the report-noisy-max algorithm, a discrete version of the Exponential mechanism [12], to privately identify the worst-case OPF model. Here, the score function $\Delta\mathcal{C}$ is the L_1 norm measuring the distance between OPF costs on real and synthetic data. Step 3 uses the Laplace mechanism to estimate the cost of the worst-case OPF model on the real data. Step 4 post-processes the synthetic dataset using optimization in (7), where $\mathcal{C}_{k^\tau}(\overline{\varphi})$ is the OPF costs obtained from the embedded DC-OPF problem (5) for some fixed vector $\overline{\varphi}$. By embedding the OPF problem as a constraint, we require feasibility and cost-consistency of $\overline{\varphi}$ with respect to the worst-case OPF models identified at previous steps. In addition, the last term in (7a) regularizes solution $\overline{\varphi}$ to ensure that the changes in synthetic capacities are only guided by feasibility and cost-consistency requirements.

The major difference between WPO and TCO algorithms is in repeating Steps 2 to 4 T times. The OPF feasibility for one model does not guarantee feasibility across the whole

Algorithm 2: Differentially private TCO for DC-OPF

Input : OPF dataset $\langle c_i, d_i, \mathcal{P}_i, F, \overline{f}\rangle_{i=1}^m$; DP parameters $\varepsilon_1, \varepsilon_2, \alpha$; iteration limit T

Output: Synthetic OPF data $\langle c_i, d_i, \mathcal{P}_i, F, \overline{\varphi}\rangle_{i=1}^m$

Step 1: Initialize synthetic dataset $\overline{\varphi}^0 = \overline{f} + \text{Lap}\,(\alpha/\varepsilon_1)^e$

for $t \in 1, \ldots, T$ **do**

  Step 2: Exponential mech. to find the worst-case OPF:

  for $i \in 1, \ldots, m$ **do**

    $\Delta\mathcal{C}_i = \left\|\mathcal{C}_i(\overline{f}) - \mathcal{C}_i^R(\overline{\varphi}^{t-1})\right\|_1 + \text{Lap}\,(\bar{c}\alpha/\varepsilon_2)$

  end

  return: index $k^t = \text{argmax}_i \Delta\mathcal{C}_i$ of the worst-case OPF

  Step 3: Laplace mech. to compute the worst-case cost:

$$\overline{\mathcal{C}}_t = \mathcal{C}_{k^t}(\overline{f}) + \text{Lap}\,(\bar{c}\alpha/\varepsilon_2)$$

  Step 4: Post-processing optimization of synthetic data:

$$\overline{\varphi}^t \in \underset{\overline{\varphi}}{\text{argmin}} \textstyle\sum_{\tau=1}^t \left\|\overline{\mathcal{C}}_\tau - \mathcal{C}_{k^\tau}(\overline{\varphi})\right\| + \left\|\overline{\varphi} - \overline{\varphi}^{t-1}\right\| \quad (7a)$$

$$\text{subject to DC-OPF problem (5) on } \overline{\varphi}, \forall\tau \quad (7b)$$

end

return: synthetic line capacity $\overline{\varphi} \leftarrow \overline{\varphi}^T$

population of models. By increasing T, the TCO algorithm finds more worst-case OPF models with the largest cost $\mathcal{C}_i^R$ of violations, thereby improving the accuracy (feasibility) of the synthetic dataset across the population.

To solve the bilevel problem in (7), we obtain a single-level equivalent by substituting the embedded OPF problems with their Karush–Kuhn–Tucker conditions (KKTs) [17, §6]. The complementarity slackness, i.e., a non-convex subset of KKTs, is addressed using Special Ordered Set of Type 1 variables, allowing for a global solution to this problem. We refer to the online repository (see the link below) for details. Although the resulting mixed-integer problem renders the entire algorithm NP-hard, we use practical heuristics, such as a mixed-integer solver *Gurobi*, making the TCO algorithm more practical than its worst-case complexity would imply. Finally, as problem (7) only relies on obfuscated data, it thus does not induce any privacy loss. We now state the DP guarantee of the algorithm.

Theorem 6 (DP of the TCO Algorithm): Setting $\varepsilon_1 = \varepsilon/2$ and $\varepsilon_2 = \varepsilon/(4T)$ renders Algorithm 2 $\varepsilon-$DP for $\alpha-$adjacent DC-OPF datasets.

Remark 1 (Relation to prior work): When $m = T = 1$, Step 2 of TCO algorithm is redundant, and the algorithm replicates the Laplace-based PLO mechanism in [10], when applied to the capacity obfuscation in the DC-OPF setting. The difference between the two algorithms reveals when the synthetic data must be feasible and cost consistent on a population of OPF models, i.e., $m \gg 1$. Indeed, the worst-case OPF model and cost can also be estimated using Laplace perturbations, but the induced privacy loss would reach $mT\varepsilon_2$. The combination of the Exponential and Laplace mechanisms in Algorithm 2, however, reduces the privacy loss to $2T\varepsilon_2$.

V. Numerical Experiments

In our experiments, we fix the privacy loss $\varepsilon = 1$ and vary adjacency α, hence increasing the range of adjacent datasets, which are required to be statistically indistinguishable. All data, codes, and additional experiments are available online:

`https://github.com/wdvorkin/SyntheticData`

A. Synthetic Wind Power Records Generation

We first demonstrate the WPO Algorithm 1 using a wind power curve of the General Electric GE-2.75.103 turbine from [18], considering a medium range of wind speeds between 2.5 and 12.5 $^m/_s$, where generation is most sensitive to speed. We then perturb each power record with a Gaussian noise $\mathcal{N}(0, 0.1)$ to introduce some variation among the records; the data is thus not completely real, but resembles real-life datasets which we hope to eventually release with our algorithm. In the dataset, we have $m = 1,000$ normalized power measurements $y \in [0, 1]^m$ and corresponding wind speeds x.

To specify regression (2), we transform the wind speed records using $p = 5$ Gaussian radial basis functions:

$$\varphi_j(x) = e^{-\left(\frac{1}{2}\|x-\mu_j\|\right)^2}, \forall j = 1, \ldots, p,$$

positioned at $\mu = \{2.5, 5, 7.5, 10, 12.5\}$ $^m/_s$. Each feature in X is obtained as $X_{ij} = \varphi_j(x_i), \forall i = 1, \ldots, m, \forall j = 1, \ldots, p$, and the regularization parameter is set to $\lambda = 10^{-3}$.

We use the standard Laplace mechanism for the reference, which perturbs power records as $\tilde{y} = y + \text{Lap}(\alpha/\varepsilon)^m$, and projects them onto feasible range $[0, 1]^m$. The resulting synthetic records satisfy $\varepsilon-$DP for $\alpha-$adjacent datasets, as per Theorems 2 and 3. In the WPO algorithm, we set ε_1 and ε_2 according to Theorem 5. We also set regularization parameters $\gamma_y, \gamma_\beta = 10^{-5}$ for post-processing in (3).

Figure 1 demonstrates some examples of synthetic wind power dataset releases. Here, we measure adjacency α in % of the nominal capacity of the wind turbine. Observe, that with increasing α, the regression-agnostic Laplace mechanism yields a larger deviation of the synthetic records from the real data. While the WPO algorithm introduces even more noise, i.e., $\times 3$ more noise at Step 1 and more noise at Step 2 due to sensitivities δ_ℓ and δ_β growing in α, the post-processing of the noisy records at Step 3 results in a better accuracy of the synthetic dataset. Figure 2 demonstrates the statistical significance of this observation by plotting the loss on synthetic datasets under the two methods. With increasing α, the Laplace mechanism demonstrates a notable deviation from the loss on real data. The WPO algorithm, on the other hand, converges to the real loss on average and does not deviate significantly. Moreover, with the WPO algorithm, the regression model on synthetic data remains substantially closer to the model on real data, as further shown in Fig. 1.

B. Synthetic Transmission Data Generation

We apply the TCO algorithm to a network data release from the IEEE 73-Bus Reliability Test System. To make the case more challenging, we reduce the transmission capacity to 60% of the nominal level to increase network congestion. We generate $m = 10^3$ feasible DC-OPF datasets by sampling demand and generation limits from uniform distributions with bounds $\pm 12.5\%$ of their nominal values. The cost data is

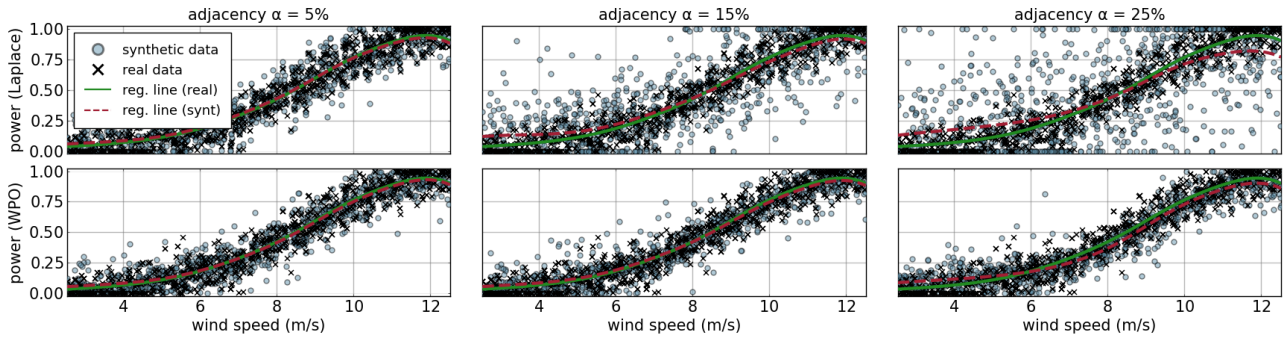

Fig. 1. Wind power records obfuscation with Laplace (top row) and WPO (bottom row) algorithms for varying adjacency. The green lines depict the regression model on the real data (black marks), while the dashed red lines depict the regression model on synthetic data (round marks).

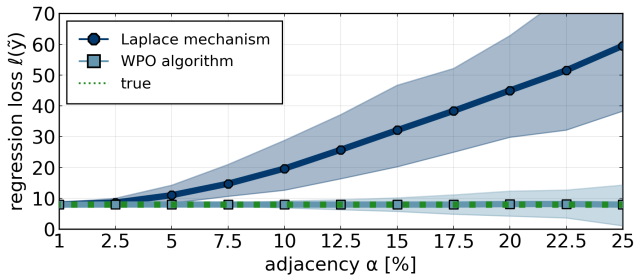

Fig. 2. The mean and 90% confidence band of regression losses on synthetic data for 300 runs of Laplace and WPO algorithms.

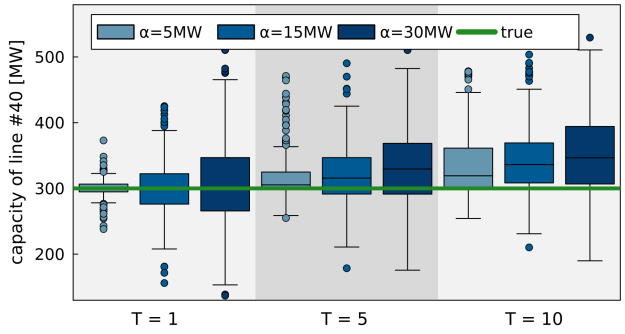

Fig. 3. Distributions of obfuscation outcomes for line #40 across 300 runs of the TCO algorithm for varying adjacency $\boldsymbol{\alpha}$ and iteration limit $\boldsymbol{T}$.

sampled from a uniform distribution $\mathcal{U}(80, 100)$ \$/MWh, and we set penalty $\psi = 3 \cdot 10^3$ in (6a) for flow limit violations. The privacy loss ε is split according to Theorem 6. Finally, we vary adjacency parameter α from 5 to 30 MW and iteration limit T from 1 to 10. Each iteration of the TCO algorithm does not take more than 30 seconds of CPU time.

By increasing α, we increase the noise magnitude at Step 1 of the TCO algorithm, resulting in a broader distribution of synthetic dataset outcomes, as depicted by box plots in Fig. 3 for one selected transmission line. However, as noise increases, the probability of obtaining an infeasible synthetic dataset also increases. We thus increase the iteration limit T to improve the accuracy of the synthetic dataset. By setting T, we require feasibility and cost-consistency with respect to the set of T worst-case OPF models and outcomes, provided at Steps 2 and 3, respectively. Such deeper post-processing results in distributional shifts, as further shown in Fig. 3 for increasing T. The virtue of these shifts is revealed in Fig.

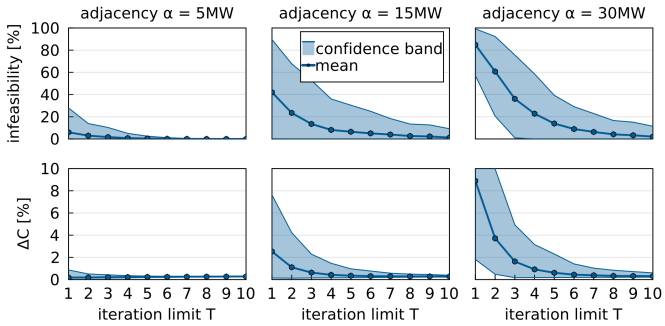

Fig. 4. Infeasibility and sub-optimality of synthetic DC-OPF datasets. Top row: percentage of infeasible OPF solutions across a population of OPF models. Bottom row: the mean sub-optimality of OPF costs on synthetic data. The shaded areas are 90% confidence bands.

4, where the top row demonstrates how the probability of infeasible OPF outcomes on synthetic datasets reduces as the iteration limit increases. For smaller adjacency, it takes fewer iterations to restore the feasibility of the synthetic dataset. For example, for $\alpha = 5$MW, it is enough to leverage 6 worst-case OPF models in the post-processing optimization at Step 4 to restore feasibility across the entire population of $1,000$ OPF models. For larger adjacency, it takes as much as 10 iterations on average. The bottom row in Fig. 4 depicts the mean sub-optimality of OPF solutions on synthetic data, computed as:

$$\Delta\mathcal{C} = \frac{1}{m}\sum_{i=1}^{m} \frac{\left\|\mathcal{C}_i(\overline{f}) - \mathcal{C}_i^R(\overline{\varphi}^T)\right\|}{\mathcal{C}_i(\overline{f})} \times 100\%. \tag{8}$$

The sub-optimality increases in adjacency parameter α, as more noise corrupts the real data. However, as we increase T, the OPF cost on synthetic data gets closer to that on the real data, eventually reaching near zero optimality gap.

VI. Conclusions

We developed two algorithms for privacy-preserving releases of synthetic power system datasets that enjoy both Laplace and Exponential DP mechanisms to guarantee privacy and leverage optimization-based post-processing to ensure data accuracy for downstream optimization and machine learning problems. Although the proposed algorithms focus on wind power and transmission capacity data releases, other network parameters can be released similarly, e.g., see the electric load obfuscation (ELO) algorithm in the online repository accompanying this paper. We also note that related privacy notions,

such as $(\varepsilon, \delta)-$DP with similar composition properties, are also relevant for synthetic power system datasets.

APPENDIX

A. Proof of Lemma 1

The worst-case sensitivity of weights is bounded as:

$$\delta_\beta = \underset{y \sim_\alpha y'}{\text{maximize}} \ \|\beta(y) - \beta(y')\|_1 \tag{9a}$$
$$= \underset{y \sim_\alpha y'}{\text{maximize}} \ \left\|(X^\top X + \lambda I)^{-1} X^\top (y - y')\right\|_1 \tag{9b}$$
$$\leqslant \left\|(X^\top X + \lambda I)^{-1} X^\top\right\|_1 \cdot \underset{y \sim_\alpha y'}{\text{maximize}} \ \|y - y'\|_1 \tag{9c}$$
$$\leqslant \left\|(X^\top X + \lambda I)^{-1} X^\top\right\|_1 \cdot \alpha \tag{9d}$$

where equality (9b) is from the closed-form solution to weights, inequality (9c) is due to the Hölders inequality, and inequality (9d) is due to Definition 1 of adjacent datasets.

The sensitivity of regression loss ℓ is bounded as:

$$\delta_\ell = \underset{y \sim_\alpha y'}{\text{maximize}} \ \|\ell(y) - \ell(y')\|_1 \tag{10a}$$
$$= \underset{y \sim_\alpha y'}{\text{maximize}} \ \big\|\|X\beta - y\| - \|X\beta - y'\|\big\|_1 \tag{10b}$$
$$\leqslant \underset{y \sim_\alpha y'}{\text{maximize}} \ \|X(\beta(y) - \beta(y')) - (y - y')\| \tag{10c}$$
$$= \underset{y \sim_\alpha y'}{\text{maximize}} \ \left\|(X(X^\top X + \lambda I)^{-1} X^\top - I)(y - y')\right\| \tag{10d}$$
$$= \underset{i=1,\ldots,m}{\text{maximize}} \ \left\|(X(X^\top X + \lambda I)^{-1} X^\top - I)(e_i \circ \alpha)\right\| \tag{10e}$$

where (10c) is due to the reverse triangle inequality, (10d) is from the closed-form solution to regression weights. Equality (10e) originates from Definition 1 of adjacent datasets, i.e., different in one element by at most α. It is thus enough to find index i of that element which maximizes the norm.

B. Proof of Theorem 5

The WPO algorithm queries real data in the interest of the following computations:

1) Dataset initialization at Step 1 using the Laplace mechanism with parameters α/ε_1. Since the worst-case sensitivity of identity queries is α [13], this computation is ε_1-DP by Theorem 3.
2) Estimation of the regression loss on the real data at Step 2 using the Laplace mechanism with parameters $\delta_\ell/\varepsilon_2$. By Lemma 1 and Theorem 3, this estimation is ε_2-DP.
3) Estimation of regression weights on the real data at Step 2 using the Laplace mechanism with parameters $\delta_\beta/\varepsilon_2$. By Lemma 1 and Theorem 3, this estimation is ε_2-DP.

As the post-processing optimization at Step 3 only uses obfuscated data, it does not induce any privacy loss per Theorem 2. Per Theorem 1, the total privacy loss becomes $\varepsilon_1 + 2\varepsilon_2$, yielding ε when setting parameters $\varepsilon_1 = \varepsilon/2$ and $\varepsilon_2 = \varepsilon/4$.

C. Proof of Theorem 6

Algorithm 2 queries private transmission capacity vector $\overline{f}$ for the following computations:

1) Initial dataset $\overline{\varphi}^0$: the algorithm uses a private identity query with privacy budget α/ε_1. Since the sensitivity of identity queries on $\alpha-$adjacent datasets is α [13], by Theorem 3 this computation is ε_1-DP.
2) Worst-case OPF index: found by the discrete variant of the Exponential mechanism with privacy budget $\overline{c}\alpha/\varepsilon_2$. Since the sensitivity of the score function $\Delta\mathcal{C}_i$ is the same as that of $\mathcal{C}_i$, by Theorems 2 and 4 and Assumption 2, this is ε_2-DP.
3) Worst-case OPF cost: Step 3 uses a private identity query of the worst-case OPF cost using privacy budget $\overline{c}\alpha/\varepsilon_2$. Per Assumption 2 and Theorem 3, this computation is ε_2-DP.

Let $\overline{\varepsilon}$ be the total privacy loss of the algorithm. Step 1 accumulates privacy loss of ε_1. Since Steps 2 and 3 repeat T times, per Theorem 1, they accumulate the privacy loss of $2T\varepsilon_2$. The total loss is then $\overline{\varepsilon} = \varepsilon_1 + 2T\varepsilon_2$, which amounts to ε when setting DP parameters $\varepsilon_1 = \varepsilon/2$ and $\varepsilon_2 = \varepsilon/(4T)$.

REFERENCES

[1] R. Iyengar *et al.*, "Towards practical differentially private convex optimization," in *2019 IEEE Symposium on Security and Privacy*. IEEE, 2019, pp. 299–316.

[2] M. Gong *et al.*, "A survey on differentially private machine learning," *IEEE Comput. Intell. Mag.*, vol. 15, no. 2, pp. 49–64, 2020.

[3] V. Dvorkin *et al.*, "Differentially private optimal power flow for distribution grids," *IEEE Trans. Power Syst.*, vol. 36, no. 3, pp. 2186–2196, 2021.

[4] ——, "Differentially private distributed optimal power flow," in *2020 59th IEEE Conference on Decision and Control (CDC)*, 2020, pp. 2092–2097.

[5] F. Zhou, J. Anderson, and S. H. Low, "Differential privacy of aggregated DC optimal power flow data," in *2019 American Control Conference (ACC)*. IEEE, 2019, pp. 1307–1314.

[6] V. Dvorkin *et al.*, "Privacy-preserving convex optimization: When differential privacy meets stochastic programming," *arXiv:2209.14152*, 2022.

[7] G. Vietri *et al.*, "New oracle-efficient algorithms for private synthetic data release," in *International Conference on Machine Learning*. PMLR, 2020, pp. 9765–9774.

[8] M. Hardt, K. Ligett, and F. McSherry, "A simple and practical algorithm for differentially private data release," *Advances in neural information processing systems*, vol. 25, 2012.

[9] H. Wang and C. Wu, "Privacy preservation for time series data in the electricity sector," *IEEE Trans. Smart. Grid.*, 2022.

[10] F. Fioretto, T. W. Mak, and P. Van Hentenryck, "Differential privacy for power grid obfuscation," *IEEE Trans Smart Grid*, vol. 11, no. 2, pp. 1356–1366, 2019.

[11] T. W. Mak *et al.*, "Privacy-preserving power system obfuscation: A bilevel optimization approach," *IEEE Trans. Power Syst.*, vol. 35, no. 2, pp. 1627–1637, 2019.

[12] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[13] K. Chatzikokolakis *et al.*, "Broadening the scope of differential privacy using metrics," in *Privacy Enhancing Technologies: 13th International Symposium, Bloomington, IN, USA*. Springer, 2013, pp. 82–102.

[14] C. Dwork *et al.*, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*. Springer, 2006, pp. 265–284.

[15] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*. IEEE, 2007, pp. 94–103.

[16] Y. Nesterov and A. Nemirovskii, *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.

[17] D. Pozo, E. Sauma, and J. Contreras, "Basic theoretical foundations and insights on bilevel models and their applications to power systems," *Ann. Oper. Res.*, vol. 254, pp. 303–334, 2017.

[18] I. Staffell and S. Pfenninger, "Using bias-corrected reanalysis to simulate current and future wind power output," *Energy*, vol. 114, pp. 1224–1239, 2016.