

Differentially Private Distribution Release of Gaussian Mixture Models via KL-Divergence Minimization

Hang Liu, *Member, IEEE*, Anna Scaglione, *Fellow, IEEE*, and Sean Peisert, *Senior Member, IEEE*

Abstract—Gaussian Mixture Models (GMMs) are widely used statistical models for representing multi-modal data distributions, with numerous applications in data mining, pattern recognition, data simulation, and machine learning. However, recent research has shown that releasing GMM parameters poses significant privacy risks, potentially exposing sensitive information about the underlying data. In this paper, we address the challenge of releasing GMM parameters while ensuring differential privacy (DP) guarantees. Specifically, we focus on the privacy protection of mixture weights, component means, and covariance matrices. We propose to use Kullback-Leibler (KL) divergence as a utility metric to assess the accuracy of the released GMM, as it captures the joint impact of noise perturbation on all the model parameters. To achieve privacy, we introduce a DP mechanism that adds carefully calibrated random perturbations to the GMM parameters. Through theoretical analysis, we quantify the effects of privacy budget allocation and perturbation statistics on the DP guarantee, and derive a tractable expression for evaluating KL divergence. We formulate and solve an optimization problem to minimize the KL divergence between the released and original models, subject to a given (ϵ, δ) -DP constraint. Extensive experiments on both synthetic and real-world datasets demonstrate that our approach achieves strong privacy guarantees while maintaining high utility.

Index terms— Gaussian mixture model, density estimation, distribution release, model fitting, differential privacy.

I. INTRODUCTION

In recent years, the remarkable success of data-driven artificial intelligence (AI) has spurred an increasing demand for the sharing and analysis of large-scale, multi-class, and high-dimensional datasets across a variety of domains, such as healthcare records, consumer transactions, and mobility traces. Organizations have recognized the potential of sharing data statistics to enhance data mining, improve public services, optimize recommendations, and facilitate data simulation [1]. However, sharing raw data or even their statistics raise significant privacy concerns, especially when sensitive attributes of individuals might be inferred. This tension underscores the need for privacy-preserving mechanisms that allow the release

of data statistics without compromising personal or proprietary information [2].

Among the various statistical modeling approaches, Gaussian Mixture Models (GMMs) stand out as a versatile tool for representing complex, multivariate, and multi-modal data distributions [3]. GMMs are widely applied in data mining, machine learning (ML), and pattern recognition. By modeling the overall distribution as a mixture of several Gaussian components, GMMs naturally capture latent subgroups or clusters in the data, such as different risk profiles in healthcare or distinct spending behaviors in retail. Furthermore, GMMs provide a compact and interpretable parameterization, including component means, covariances, and mixture weights (also known as categorical frequencies), which can be shared more efficiently than raw data records. As a motivating real-world application, GMMs are accurate in representing the statistics of energy consumption data. In particular, the logarithmic values of load demand profiles in Advanced Metering Infrastructure (AMI) data have been empirically shown to fit well to GMMs [4]. In the context of energy usage, these load demand profiles represent the amount of electricity consumed over time by a given user or household. When fitted to a GMM, the data naturally forms clusters, with each Gaussian component capturing different usage patterns (e.g., high, medium, and low usage times).

Despite their advantages, releasing GMM parameters directly still poses privacy risks, particularly when the underlying dataset contains sensitive information. An adversary could potentially exploit small mixture components or extreme values to infer specific individuals' records or their associated class labels. For instance, an adversary could inspect the component means of the GMM model fitted to the AMI dataset and detect changes that could identify the presence of a specific customer. Small perturbations in the model, such as flipping the label of a single customer, could significantly alter the estimated parameters, making it easier for the adversary to infer private information. To mitigate these risks, differential privacy (DP) [2], [5] offers a robust framework for quantifying and controlling the privacy loss resulting from the inclusion or exclusion of any individual record in a dataset. By introducing carefully calibrated noise, we can estimate mixture model parameters that satisfy strong privacy guarantees while still enabling meaningful statistical analysis.

A central challenge in applying DP mechanisms to GMM fitting is maintaining fidelity between the original non-private data distribution and the released one under a DP constraint.

This research was supported in part by the Director, Cybersecurity, Energy Security, and Emergency Response (CESER) office of the U.S. Department of Energy, via the Privacy-Preserving, Collective Cyberattack Defense of DERs project, under contract DE-AC02-05CH11231. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors of this work.

Hang Liu and Anna Scaglione are with the Department of Electrical and Computer Engineering, Cornell Tech, Cornell University, New York, NY, 10044 USA (e-mails: {hl2382, as337}@cornell.edu). Sean Peisert is with the Computing Sciences Research, Lawrence Berkeley National Laboratory, Berkeley, CA 94720 USA (e-mail: speisert@lbl.gov).

Conventional DP mechanisms typically add artificial noise to perturb the model parameters in order to protect data privacy. However, this perturbation inevitably compromises the accuracy of the released GMM. For instance, the noise added to the Gaussian means could distort the true cluster centers, leading to incorrect clustering and a GMM that fails to represent the underlying data structure. Existing work [6] addressed differentially private GMMs by adding white Laplace noise to the mixture weights, means, and covariances individually. This approach models the accuracy of differentially private GMMs using the mean-square errors (MSEs) of individual parameters.

Rather than the MSEs that measure the accuracy of releasing individual parameters, we argue that distribution divergence, specifically *Kullback–Leibler (KL) divergence*, is another comprehensive metric for quantifying the fidelity of GMMs. KL divergence is a well-established statistical distance metric in information theory that quantifies how much one probability distribution diverges from another [7]. When used to characterize the utility of distribution release, KL divergence measures the “information” lost when substituting the true distribution with an approximate, privacy-preserving model. It is well-known that computing maximum likelihood estimation (MLE) is asymptotically equivalent to minimizing the KL divergence between the empirical and true distributions [8]. Thus, publishing a data distribution with minimal KL divergence ensures that the released model is as accurate as possible in the maximum-likelihood sense. Moreover, when applied to data simulation and augmentation for ML, KL divergence is closely related to the generalization ability of ML models. Extensive research has established PAC-Bayes bounds for quantifying the generalization error of ML models [9], which depends critically on the KL divergence between the training and testing data distributions. In this context, KL divergence serves as a powerful metric for evaluating how well an ML model trained on a given data distribution (e.g., a released GMM under privacy constraints) generalizes to the true data, thus providing an important tool for assessing the performance of data synthesis and augmentation in ML training.

KL divergence provides a unified framework for characterizing the combined impact of noise perturbation on the accuracy of the released GMM, taking into account the effects of noise on all model parameters. However, characterizing the KL divergence between the original and DP-preserved distributions, particularly when applying DP mechanisms to the released GMM parameters, remains an open problem. In this work, we show that the KL divergence for GMM parameter release exhibits a closed-form expression, thereby opening new research avenues for designing DP mechanisms that minimize KL divergence while releasing differentially private GMM models.

A. Our Contributions

Motivated by the above discussions, in this work we propose a differentially private GMM estimation framework that places *KL divergence* at the core of model evaluation. Specifically, we study the fitting of a multi-class labeled dataset to a GMM whose parameters, including mixture weights, Gaussian

means, and covariances, are then released under a given DP requirement. We study the privacy mechanism by adding artificial noise to each parameter following certain controllable statistical distributions. By doing so, we obtain a mechanism design problem that seeks to balance the KL divergence utility and data privacy. We further derive a *closed-form* expression to evaluate the KL divergence between the released private GMM and the original model. We then present a method to bound the achievable DP level by carefully controlling the noise injection. Through this, we optimize the noise distributions to minimize KL divergence subject to a desired DP constraint. The main contributions of this paper are summarized as follows:

- **DP-GMM Model Release Framework:** We propose a two-step approach for the differentially private release of GMMs by fitting and releasing GMM model parameters under any given (ϵ, δ) -DP constraint. Our method first fits the dataset into a GMM and then injects *multivariate* Gaussian and Wishart noise into the Gaussian means and covariances, respectively. For the discrete-valued mixture weights, we introduce a random mapping mechanism that maps the estimated weights to other feasible mixture weights with a controllable mapping probability. We then formulate the DP mechanism design problem as optimizing the privacy budget and the noise distributions to balance the trade-off between DP and data utility, measured by the KL divergence between the released GMM and the original non-private model.
- **Privacy Analysis and Privacy-Utility Trade-Off Characterization:** We conduct an (ϵ, δ) -DP analysis, bounding the combined privacy loss from releasing both the continuous parameters (means and covariances) and the discrete parameters (mixture weights). In addition, we derive a tractable, closed-form expression for the KL divergence between the privatized GMM and the original fitted distribution. This enables us to characterize the trade-off between privacy and the overall fidelity of the released model, providing a clear path to control this balance.
- **DP Mechanism Design via Privacy-Constrained KL-Divergence Minimization:** Based on our analysis of model utility and privacy, we formulate an optimization problem that allocates privacy budgets and optimizes noise distributions by minimizing the expected KL divergence subject to a given (ϵ, δ) -DP constraint. While the problem is non-convex, we propose a low-complexity alternating optimization solution to compute a local optimum efficiently.

Through extensive experiments on both synthetic and real-world datasets, we demonstrate how our method preserves model accuracy while satisfying rigorous privacy criteria. Specifically, we show that our approach provides a differentially private GMM that adheres to stringent DP requirements, while maintaining a much lower KL divergence and preserving model fidelity in comparison to existing DP mechanisms.

B. Organization and Notations

The remainder of this paper is organized as follows. We introduce the system model for differentially private parameter release of GMMs in Section II. In Section III, we analyze the achievable DP and formulate the DP mechanism design problem. In Section IV, we introduce the proposed solution to the DP mechanism design problem. In Section V, we present experimental results to evaluate the proposed method. Finally, this paper concludes in Section VI.

Throughout, we use regular, bold small, and bold capital letters to denote scalars, vectors, and matrices, respectively. We use \mathbf{X}^T to denote the transpose of \mathbf{X} , \mathbf{X}^H to denote the conjugate transpose, $\text{tr}(\mathbf{X})$ to denote the trace, and $|\mathbf{X}|$ to denote the determinant of \mathbf{X} . We use x_i to denote the i -th entry of vector \mathbf{x} , x_{ij} or $X_{i,j}$ interchangeably to denote the (i, j) -th entry of matrix \mathbf{X} , and \mathbf{x}_j to denote the j -th column of \mathbf{X} . The real normal distribution with mean $\boldsymbol{\mu}$ and covariance \mathbf{C} is denoted by $\mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$, and the cardinality of set \mathcal{S} is denoted by $|\mathcal{S}|$. We use $\|\cdot\|_p$ to denote the ℓ_p norm, \mathbf{I}_N to denote the $N \times N$ identity matrix, $\mathbf{1}$ (or $\mathbf{0}$) to denote the all-one (or all-zero) vector with an appropriate size. We use $x \sim p(x)$ to represent that the random variable x is drawn from the distribution $p(x)$.

II. DIFFERENTIALLY PRIVATE DISTRIBUTION RELEASE FOR GMMs

Consider a dataset $\mathcal{D} = \{\mathbf{x}_n, y_n\}_{n=1}^N$, consisting of N labeled samples $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^d$ partitioned into K classes. Each sample \mathbf{x}_n is associated with a class label $y_n \in \{1, \dots, K\}$.¹ Without loss of generality, we assume every class has at least one data point. Our aim is to publicly release a distributional model for these multi-class data in a differentially private manner by fitting them with a *multivariate* GMM.

To achieve this, we estimate the parameters of the following GMM based on \mathcal{D} :

$$p(y = k) = \pi_k(\mathcal{D}), p(\mathbf{x}|y = k) = \mathcal{N}(\boldsymbol{\mu}_k(\mathcal{D}), \boldsymbol{\Sigma}_k(\mathcal{D})), \forall k, \quad (1)$$

where $\pi_k(\mathcal{D}) > 0$ denotes the empirical frequency (mixture weight) of class k satisfying $\sum_k \pi_k(\mathcal{D}) = 1$, and $\boldsymbol{\mu}_k(\mathcal{D})$ and $\boldsymbol{\Sigma}_k(\mathcal{D})$ represent the mean and covariance of the k -th Gaussian component, respectively. For brevity, when the dataset \mathcal{D} is implied by context, we omit the argument in the notation.

To satisfy a prescribed DP constraint, we employ a *two-step* strategy, by first estimating the parameters $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$ from the dataset \mathcal{D} , followed by applying a DP mechanism before releasing these parameters. Concretely, we compute the

parameters via histograms and sample statistics as follows:

$$\pi_k = \frac{1}{N} \sum_{y_n=k} 1, \quad (2a)$$

$$\boldsymbol{\mu}_k = \frac{1}{N\pi_k} \sum_{y_n=k} \mathbf{x}_n, \quad (2b)$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N\pi_k - 1} \sum_{y_n=k} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T. \quad (2c)$$

Define $\boldsymbol{\pi} = [\pi_1, \dots, \pi_K]$. From (2a), it follows that $\boldsymbol{\pi}$ is a discrete histogram belonging to the following set:

$$\mathcal{S} = \left\{ \boldsymbol{\pi} : \sum_{k=1}^K \pi_k = 1, \pi_k \in \left\{ \frac{1}{N}, \frac{2}{N}, \dots, \frac{N-1}{N} \right\}, \forall k \right\}. \quad (3)$$

By the stars-and-bars theorem, the cardinality of \mathcal{S} is given by $|\mathcal{S}| = \binom{N-1}{K-1}$. Unless otherwise specified, when we referred to an element $\boldsymbol{\pi} \in \mathcal{S}$, we assume a fixed ordering of the elements in \mathcal{S} and thus any feasible $\boldsymbol{\pi}$ has a unique index in the order.

After estimating the parameters $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$, we apply a post-processing step to satisfy given DP constraints before making these parameters public. The following definition of DP, introduced in [2], [5], is standard in the literature.

Definition 1 ((ϵ, δ) -DP). Consider a randomized mechanism M that takes a dataset \mathcal{D} as input and outputs a query answer $M(\mathcal{D})$. Let $\epsilon > 0$ and $\delta \in [0, 1]$ be given parameters. M is said to satisfy (ϵ, δ) -DP if, for any two adjacent datasets \mathcal{D} and \mathcal{D}' where \mathcal{D}' differs from \mathcal{D} by altering the class label of exactly one data point², the following inequality holds for all measurable sets \mathcal{H} :

$$\Pr(M(\mathcal{D}) \in \mathcal{H}) \leq e^\epsilon \Pr(M(\mathcal{D}') \in \mathcal{H}) + \delta, \forall \mathcal{H}. \quad (4)$$

In this work, M refers to the mechanism that releases the GMM parameters $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$. Intuitively, an (ϵ, δ) -DP mechanism makes it difficult for any potential attacker to detect the class label of any single data point in \mathcal{D} , especially when ϵ and δ are small. Here, ϵ controls the overall privacy stringency, while δ governs the probability that this privacy guarantee may not hold.

To satisfy a given DP requirement, we leverage well-known DP mechanisms that add artificial noise to $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$ prior to release. Unlike canonical mechanisms designed for unbounded scalars, we must ensure that the added noise respects the dimensionality and retains the feasibility of these parameters. Concretely, we need the perturbed frequency parameters to remain in \mathcal{S} , and the perturbed covariance matrices to stay positive semidefinite (PSD).

Toward this end, we assign each parameter a specific noise distribution. For the continuous-valued mean and covariance of each Gaussian component, we adopt a *multivariate* Gaussian mechanism and a Wishart mechanism [10]. The perturbed

¹When unavailable, the class labels can be calculated *a priori* by using clustering methods such as K -Means clustering.

²For simplicity, we define DP in terms of an adjacent dataset \mathcal{D}' that alters exactly one class label from \mathcal{D} . The same principle extends to the variant where \mathcal{D}' replaces a single data point in \mathcal{D} with a bounded alternative or removes/adds one data point from \mathcal{D} .

means and covariances are given by

$$\tilde{\boldsymbol{\mu}}_k = \boldsymbol{\mu}_k + \mathbf{w}_k, \mathbf{w}_k \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}_k^{-1}), \forall k, \quad (5)$$

$$\tilde{\boldsymbol{\Sigma}}_k = \boldsymbol{\Sigma}_k + \mathbf{W}_k, \mathbf{W}_k \sim \mathcal{W}_d\left(\frac{1}{\gamma_k} \mathbf{I}_d, d+1\right), \forall k, \quad (6)$$

where \mathbf{w}_k and \mathbf{W}_k are the noise terms added to the mean and covariance of the k -th component, respectively, $\boldsymbol{\Gamma}_k$ is the precision matrix of the Gaussian noise, $\mathcal{W}_d(\gamma_k \mathbf{I}_d, d+1)$ is a Wishart distribution over $d \times d$ matrices with $d+1$ degrees of freedom and scale matrix $\gamma_k^{-1} \mathbf{I}_d$, and the parameter $\gamma_k > 0$ controls the variance of the Wishart noise. Because Wishart random matrices are always positive definite, the construction in (6) ensures that the perturbed covariance remains feasible.

Meanwhile, to privatize the discrete-valued frequency vector $\boldsymbol{\pi}$, we release a random vector $\tilde{\boldsymbol{\pi}} \in \mathcal{S}$ according to a tunable conditional probability mass function (PMF) $f(\tilde{\boldsymbol{\pi}}|\boldsymbol{\pi})$. With a fixed ordering of elements in \mathcal{S} , this PMF can be viewed as a transition matrix $\mathbf{F} \in [0, 1]^{|\mathcal{S}| \times |\mathcal{S}|}$ whose entry $F_{i,j}$ specifies the probability of mapping the j -th input element in \mathcal{S} to the i -th output element in \mathcal{S} . In the sequel, we will use $f(\tilde{\boldsymbol{\pi}}|\boldsymbol{\pi})$ and \mathbf{F} interchangeably to denote the mapping PMF.

Remark 1. Our randomization mechanism for the frequency parameter $\tilde{\boldsymbol{\pi}}$ presumes a discrete-valued input $\boldsymbol{\pi}$, matching the histogram-based parameter estimation in (2a). Other GMM estimation methods, such as the expectation-maximization algorithm [11] may yield continuous-valued frequencies, which can be discretized beforehand if needed.

Our goal is to choose $\{\boldsymbol{\Gamma}_k, \gamma_k\}_{k=1}^K$ and \mathbf{F} so that the released parameters $\{\tilde{\boldsymbol{\pi}}_k, \tilde{\boldsymbol{\mu}}_k, \tilde{\boldsymbol{\Sigma}}_k\}_{k=1}^K$ satisfy the (ϵ, δ) -DP. An end user can then reconstruct the private GMM as

$$\tilde{p}(y = k) = \tilde{\pi}_k, \tilde{p}(\mathbf{x}|y = k) = \mathcal{N}(\tilde{\boldsymbol{\mu}}_k, \tilde{\boldsymbol{\Sigma}}_k), \forall k. \quad (7)$$

We use the expected KL divergence, to measure the fidelity of $\tilde{p}(\mathbf{x}, y)$:

$$\mathbb{E}[KL(\tilde{p}(\mathbf{x}, y)||p(\mathbf{x}, y))] = \mathbb{E}\left[\int \tilde{p}(\mathbf{x}, y) \ln\left(\frac{\tilde{p}(\mathbf{x}, y)}{p(\mathbf{x}, y)}\right) d\mathbf{x} dy\right], \quad (8)$$

where the expectations are taken with respect to the randomness in $\{\mathbf{w}_k, \mathbf{W}_k\}_{\forall k}$ and $\tilde{\boldsymbol{\pi}}$. A smaller KL divergence value implies that $\tilde{p}(\mathbf{x})$ is closer to $p(\mathbf{x})$ and therefore a more accurate estimate of the data density.

The following examples demonstrate the effectiveness of the KL divergence for measuring the accuracy of distribution estimation and release.

Example 1 (Classification). Classification is a fundamental problem in data mining and ML. By fitting a GMM to the raw data, one can capture latent structures or subgroups and use the model to classify new data. However, if the private GMM diverges significantly from the true distribution, these classifiers may be distorted or merged incorrectly. For instance, consider a hospital that privately releases classification information about patients based on clinical measurements (high/low blood pressure, high/low cholesterol, etc.). Policymakers rely on accurate classification boundaries to identify high-risk subpopulations and allocate resources effectively. A high KL divergence between the true and released distributions would

imply that the separations are significantly altered, undermining the utility of the published model. Hence, minimizing the KL divergence in (8) ensures that the core data structure is well preserved despite the noise added for privacy.

Example 2 (Density Estimation and Data Simulations). GMMs are widely used to estimate data density and generate synthetic samples that mimic the statistical properties of the real data. KL divergence is a standard metric for accuracy assessment and model selection of density estimation. Consider a public health agency creating a synthetic version of a patient dataset for open research. If the released GMM has a distribution that significantly diverges from the true one in terms of the KL divergence, the synthetic data may incorrectly represent disease prevalence or demographic proportions. This leads to flawed insights, misleading model development, and potential misallocation of medical resources. By monitoring and minimizing the KL divergence, the agency ensures the synthetic dataset remains faithful to genuine population patterns, while still respecting strict privacy requirements.

Example 3 (Data Augmentation for ML). Data augmentation plays a critical role in preventing model overfitting and enhancing model generalization, especially in applications where labeled data are expensive to obtain. A differentially private GMM allows organizations to share model parameters that can be used for sampling additional training data while preserving individual privacy. In machine learning theory, PAC-Bayes methods are commonly used to quantify generalization ability when training and testing data follow different distributions. Typical PAC-Bayes bounds indicate that the model generalization error is bounded by a non-diminishing term related to the KL divergence between the training and testing data distributions [9].

Moreover, in the context of data augmentation or synthesis, [12] reported that the generalization error of a model trained with an augmented or synthetic dataset is upper-bounded by a term determined by the total variation distance (TVD) between the true data distribution and the augmented data distribution. However, this TVD-based bound is often intractable, even for simple distribution models such as GMMs. According to Pinsker's inequality [13], the TVD between two distributions is tightly upper-bounded by the square root of half their KL divergence. In other words, KL divergence serves as a surrogate for TVD and provides a more tractable expression for bounding the model generalization ability. Maintaining low KL divergence ensures that the augmented dataset closely aligns with the true data distribution, resulting in stronger model generalization and more reliable predictive outcomes.

III. PROBLEM FORMULATION

In this section, we present a tractable problem formulation for DP mechanism design by minimizing the KL divergence subject to DP constraints.

A. DP Analysis

We begin by examining the conditions under which releasing $\{\tilde{\boldsymbol{\pi}}_k, \tilde{\boldsymbol{\mu}}_k, \tilde{\boldsymbol{\Sigma}}_k\}_{k=1}^K$ achieves a given (ϵ, δ) -DP requirement,

$$(8) = \sum_{\tilde{\pi} \in \mathcal{S}} f(\tilde{\pi} | \pi(\mathcal{D})) \underbrace{\sum_{k=1}^K \tilde{\pi}_k \left(\ln \frac{\tilde{\pi}_k}{\pi_k} + \frac{1}{2} \left(d \ln \gamma_k + \frac{d+1}{\gamma_k} \text{tr}(\mathbf{\Sigma}_k^{-1}) + \text{tr}(\mathbf{\Sigma}_k^{-1} \mathbf{\Gamma}_k^{-1}) \right) \right)}_{\triangleq g(\tilde{\pi}, \{\gamma_k, \mathbf{\Gamma}_k^{-1}\}_{\forall k})} + \text{Const.} \quad (12)$$

as a function of the controllable parameters $\{\mathbf{\Gamma}_k, \gamma_k\}_{\forall k}$ and $f(\tilde{\pi} | \pi)$. The core idea is to bound the overall privacy loss by composing the individual privacy contributions of each parameter. Following the analysis in [4], [10], we derive a sufficient condition for satisfying (4) when releasing $\{\tilde{\mu}_k, \tilde{\Sigma}_k\}_{k=1}^K$. We then enumerate all possible adjacent datasets to drive the expression of (4) for $\tilde{\pi}$.

Given the dataset \mathcal{D} of size N , let $\mathcal{D}'_{n,k'}$ denote the (n, k') -th adjacent dataset obtained by changing the class label of the n -th data point to k' . There are $N(K-1)$ such adjacent datasets in total. A sufficient condition for satisfying (ϵ, δ) -DP is summarized below.

Theorem 1. Consider the non-private parameters $\{\pi_k, \mu_k, \Sigma_k\}_{k=1}^K$ defined in (2). Releasing the perturbed parameters $\{\tilde{\pi}_k, \tilde{\mu}_k, \tilde{\Sigma}_k\}_{k=1}^K$ in (5) and (6) satisfies (ϵ, δ) -DP if ϵ and δ fulfill the following inequalities:

$$\epsilon \geq \max_{k=1}^K \left\{ \epsilon_k + \frac{3\gamma_k}{2N_k} \right\} + \epsilon_0, \quad (9)$$

$$\frac{\epsilon_k^2}{2 \ln(2/\delta)} \geq \sup_{n, k' \neq k} \left\{ \mu_k(\mathcal{D}) - \mu_k(\mathcal{D}'_{n,k'}) \right\}^T \mathbf{\Gamma}_k (\mu_k(\mathcal{D}) - \mu_k(\mathcal{D}'_{n,k'})), \quad (10)$$

$$\epsilon_0 \geq \left| \ln \frac{f(\tilde{\pi} | \pi(\mathcal{D}))}{f(\tilde{\pi} | \pi(\mathcal{D}'_{n,k'}))} \right|, \forall \tilde{\pi} \in \mathcal{S}, \forall n, k', \quad (11)$$

where N_k is the number of data points in class k , $\epsilon_0 > 0$ and $\epsilon_k > 0, 1 \leq k \leq K$, are auxiliary variables.

Proof: See Appendix A. ■

Intuitively, Theorem 1 indicates that the global privacy guarantee (in terms of ϵ) can be controlled by combining the privacy costs of each parameter release. This result gives us an explicit way to allocate the privacy budget $\epsilon_0, \epsilon_k, \forall k$ across the different noise additions, and then optimize the noise statistics $\{\mathbf{\Gamma}_k, \gamma_k\}_{\forall k}$ and $f(\tilde{\pi} | \pi)$ accordingly.

B. DP-Constrained KL-Divergence Minimization

We now formulate an optimization problem for the DP mechanism design. Our goal is to allocate the privacy budget ϵ_0 and $\epsilon_k, \forall k$, and subsequently choose the noise statistics $\{\mathbf{\Gamma}_k, \gamma_k\}_{\forall k}$ and $f(\tilde{\pi} | \pi)$ so as to minimize the expected KL divergence in (8), subject to a given (ϵ, δ) -DP constraint. The following theorem provides a tractable representation of the objective function in terms of these designing variables.

Proposition 1. The expected KL divergence in (8) can be rewritten as (12), shown at the top of the page. Here, *Const* is a constant that depends solely on the data dimension d .

Proof: See Appendix B. ■

We propose to solve the following KL-divergence minimization problem:

$$(P1) : \min_{\{\gamma_k, \mathbf{\Gamma}_k, \epsilon_k\}_{k=1}^K, \mathbf{F}, \epsilon_0} \quad (12) \quad (13a)$$

$$\text{s.t. } \max_k \left\{ \epsilon_k + \frac{3\gamma_k}{2N_k} \right\} + \epsilon_0 \leq \epsilon, \quad (13b)$$

$$(10), (11) \text{ hold,} \quad (13c)$$

$$\mathbf{\Gamma}_k \succeq \mathbf{0}, \gamma_k > 0, \epsilon_k > 0, \forall k, \quad (13d)$$

$$\epsilon_0 > 0, \mathbf{F}\mathbf{1} = \mathbf{1}, [\mathbf{F}]_{i,j} \geq 0, \forall 1 \leq i, j \leq |\mathcal{S}|. \quad (13e)$$

Constraints (13b) and (13c) ensure the (ϵ, δ) -DP requirement, while (13d) and (13e) enforce feasibility of the precision matrices and the transition PMF. Note that (P1) is a non-convex problem and involves $\mathcal{O}(|\mathcal{S}|^2)$ constraints, posing a challenge for large $|\mathcal{S}|$.

IV. DP MECHANISM OPTIMIZATION

In this section, we present several modifications to (P1) that simplify the DP mechanism optimization. We then propose a low-complexity algorithm to compute a sub-optimal solution based on alternating optimization.

A. Problem Simplification

Recall that $|\mathcal{S}| = \binom{N-1}{K-1}$. The dimension of \mathbf{F} grows exponentially in N , making (P1) computationally prohibitive for large N . However, we note that while \mathcal{S} may be large, the adjacent datasets $\{\mathcal{D}'_{n,k'}\}$ of \mathcal{D} correspond to only a small number of possible frequency vectors $\{\pi(\mathcal{D}'_{n,k'})\}$. Specifically, altering the label of exactly one data point changes exactly two entries in $\pi(\mathcal{D})$. Motivated by this, we define the possible frequency vectors generated by all the adjacent datasets of \mathcal{D} as

$$\begin{aligned} \{\pi(\mathcal{D}'_{n,k'})\} &= \mathcal{S}'(\mathcal{D}) \\ &= \{\pi(\mathcal{D}') \in \mathcal{S} : \|\pi(\mathcal{D}') - \pi(\mathcal{D})\|_0 = 2, \|\pi(\mathcal{D}') - \pi(\mathcal{D})\|_1 = \frac{2}{N}\}. \end{aligned} \quad (14)$$

It follows that $\mathcal{S}'(\mathcal{D}) \subset \mathcal{S}$ and $|\mathcal{S}'(\mathcal{D})| = K(K-1)$. Since $|\mathcal{S}'(\mathcal{D})| \ll |\mathcal{S}|$, to simply the problem in (P1) we impose an additional constraint on our solution of $\{f(\tilde{\pi} | \pi(\mathcal{D}))\}$: there is non-zero mapping probability only when the output frequencies fall within $\{\pi(\mathcal{D})\} \cup \mathcal{S}'(\mathcal{D})$, i.e.,

$$f(\tilde{\pi} | \pi(\mathcal{D})) > 0 \quad \text{only if } \tilde{\pi} \in \mathcal{S}'(\mathcal{D}) \text{ or } \tilde{\pi} = \pi(\mathcal{D}). \quad (15)$$

Hence, the corresponding transition matrix \mathbf{F} has non-zero entries only in a submatrix $\mathbf{F}' \in [0, 1]^{(|\mathcal{S}'(\mathcal{D})|+1) \times (|\mathcal{S}'(\mathcal{D})|+1)}$, whose rows and columns are indexed by $\pi(\mathcal{D})$ and the

elements of $\mathcal{S}'(\mathcal{D})$. This strategy reduces the number of free variables in \mathbf{F} from $\mathcal{O}(e^N)$ to $\mathcal{O}(K^4)$.

Under this restriction, the original constraint (11) becomes:

$$\epsilon_0 \geq \left| \ln \frac{f(\tilde{\pi}|\pi(\mathcal{D}))}{f(\tilde{\pi}|\pi')} \right|, \forall \tilde{\pi} \in \mathcal{S}'(\mathcal{D}) \cup \{\pi(\mathcal{D})\}, \forall \pi' \in \mathcal{S}'(\mathcal{D}). \quad (16)$$

Fixing an indexing order for $\{\pi(\mathcal{D})\} \cup \mathcal{S}'(\mathcal{D})$, let $\pi(\mathcal{D})$ correspond to the j^* -th column of \mathbf{F}' . Then (16) is equivalent to

$$e^{-\epsilon_0} F'_{i,j} \leq F_{i,j^*} \leq e^{\epsilon_0} F'_{i,j}, \forall j \neq j^*, \forall 1 \leq i \leq |\mathcal{S}'(\mathcal{D})| + 1. \quad (17)$$

Next, we simplify the constraints involving $\mathbf{\Gamma}_k$ by switching to $\mathbf{\Gamma}_k^{-1}$. Using the Schur complement on $\mathbf{\Gamma}_k^{-1}$, we have

$$\mathbf{\Gamma}_k \succeq \mathbf{0} \Leftrightarrow \mathbf{\Gamma}_k^{-1} \succeq \mathbf{0}, \quad (18)$$

$$(10) \Leftrightarrow \begin{bmatrix} \epsilon_k^2 / (2 \log(2/\delta)) & \boldsymbol{\mu}_k(\mathcal{D}) - \boldsymbol{\mu}_k(\mathcal{D}'_{n,k'}) \\ \boldsymbol{\mu}_k^T(\mathcal{D}) - \boldsymbol{\mu}_k^T(\mathcal{D}'_{n,k'}) & \mathbf{\Gamma}_k^{-1} \end{bmatrix} \succeq \mathbf{0}. \quad (19)$$

Therefore, we can solve for $\mathbf{\Gamma}_k^{-1}$ directly, ensuring all relevant constraints are convex in $\mathbf{\Gamma}_k^{-1}$.

Gathering all the above modifications, we obtain a simplified optimization problem:

$$(P2) : \min_{\substack{\{\gamma_k, \mathbf{\Gamma}_k^{-1}, \epsilon_k\}_{k=1}^K \\ \mathbf{F}', \epsilon_0}} \sum_{\tilde{\pi}} f(\tilde{\pi}|\pi(\mathcal{D})) g(\tilde{\pi}, \{\gamma_k, \mathbf{\Gamma}_k^{-1}\}_{\forall k}) \quad (20a)$$

$$\text{s.t. } \epsilon_k + \frac{3\gamma_k}{2N_k} + \epsilon_0 \leq \epsilon, \forall k \quad (20b)$$

$$(17), (19), \quad (20c)$$

$$\mathbf{\Gamma}_k^{-1} \succeq \mathbf{0}, \gamma_k > 0, \epsilon_k > 0, \forall k, \quad (20d)$$

$$\epsilon_0 > 0, \mathbf{F}' \mathbf{1} = \mathbf{1}, [\mathbf{F}']_{i,j} \geq 0, \forall i, j, \quad (20e)$$

where $g(\tilde{\pi}, \{\gamma_k, \mathbf{\Gamma}_k^{-1}\}_{\forall k})$ is defined in (12).

Because the feasible region of (P2) is a subset of that for (P1), the solution to (P2) is generally suboptimal but always feasible for (P1). Nonetheless, (P2) reduces the complexity substantially, down to $\mathcal{O}(K^4)$ convex constraints compared to the exponential size of \mathbf{F} in (P1).

B. Proposed Solution to (P2)

Although Problem (P2) is now significantly simplified to a reduced number of convex constraints, the non-convex nature of the objective still makes its optimal solution intractable. In this section, we propose a sub-optimal solution that optimizes the variables in an alternating fashion. The proposed approach first initializes a feasible guess for ϵ_0 and $\epsilon_k, \forall k$, and then alternately solves for $\{\gamma_k\}_{k=1}^K$, $\{\epsilon_k, \mathbf{\Gamma}_k^{-1}\}_{k=1}^K$, and $\{\mathbf{F}', \epsilon_0\}$ until convergence. The details are listed as follows.

- 1) **Update** $\{\gamma_k\}_{k=1}^K$: Fixing the values of the other variables, the optimization of $\{\gamma_k\}_{k=1}^K$ can be decomposed into K independent one-dimensional optimization subproblems

as

$$\begin{aligned} \min_{\gamma_k > 0} \quad & d\gamma_k + \frac{(d+1) \text{tr}(\mathbf{\Sigma}_k^{-1})}{\gamma_k} \\ \text{s.t.} \quad & \gamma_k \leq \frac{2N_k(\epsilon - \epsilon_0 - \epsilon_k)}{3}. \end{aligned} \quad (21)$$

By using the first-order optimality condition, it can be verified that the optimal solution to γ_k is given by

$$\gamma_k = \min \left\{ \frac{d+1}{d} \text{tr}(\mathbf{\Sigma}_k^{-1}), \frac{2N_k(\epsilon - \epsilon_0 - \epsilon_k)}{3} \right\}, \forall k. \quad (22)$$

- 2) **Update** $\{\epsilon_k, \mathbf{\Gamma}_k^{-1}\}_{k=1}^K$: Fixing the other variables, the optimization of $\{\epsilon_k, \mathbf{\Gamma}_k^{-1}\}_{k=1}^K$ can be decomposed into the following K independent convex subproblem:

$$\begin{aligned} \min_{\mathbf{\Gamma}_k^{-1}, \epsilon_k^2} \quad & \text{tr}(\mathbf{\Sigma}_k^{-1} \mathbf{\Gamma}_k^{-1}) \\ \text{s.t.} \quad & \mathbf{\Gamma}_k^{-1} \succeq \mathbf{0}, 0 < \epsilon_k^2 \leq (\epsilon - \epsilon_0 - \frac{3\gamma_k}{2N_k})^2, \\ & \begin{bmatrix} \epsilon_k^2 / (2 \log(2/\delta)) & \boldsymbol{\mu}_k(\mathcal{D}) - \boldsymbol{\mu}_k(\mathcal{D}'_{n,k'}) \\ \boldsymbol{\mu}_k^T(\mathcal{D}) - \boldsymbol{\mu}_k^T(\mathcal{D}'_{n,k'}) & \mathbf{\Gamma}_k^{-1} \end{bmatrix} \succeq \mathbf{0}, \forall n, k'. \end{aligned} \quad (23)$$

This is a semi-definite program (SDP) and can be solved by off-the-shelf solvers such as CVX [14].

- 3) **Update** \mathbf{F}' and ϵ_0 : Fixing $\{\gamma_k, \epsilon_k, \mathbf{\Gamma}_k^{-1}\}_{k=1}^K$, the optimization of \mathbf{F}' and ϵ_0 can be recast as

$$\begin{aligned} \min_{\mathbf{F}', \epsilon_0} \quad & \sum_{i=1}^{|\mathcal{S}'(\mathcal{D})|+1} g_i F'_{i,j^*} \\ \text{s.t.} \quad & \epsilon_0 \leq \epsilon - \max_k \{ \epsilon_k + \frac{3\gamma_k}{2N_k} \} \\ & e^{-\epsilon_0} F'_{i,j} \leq F_{i,j^*} \leq e^{\epsilon_0} F'_{i,j}, \forall j \neq j^*, \forall i. \\ & \epsilon_0 > 0, \mathbf{F}' \mathbf{1} = \mathbf{1}, [\mathbf{F}']_{i,j} \geq 0, \forall i, j. \end{aligned} \quad (24)$$

Here, $g_i = g(\tilde{\pi}, \{\gamma_k, \mathbf{\Gamma}_k\}_{\forall k})$ for $\tilde{\pi}$ that corresponds to the i -th row of \mathbf{F}' and j^* is the index where $\pi(\mathcal{D})$ correspond to the j^* -th column of \mathbf{F}' .

The solution of (24) is given by the following proposition.

Proposition 2. The optimal solution to (24) is given by

$$\epsilon_0 = \epsilon - \max_k \{ \epsilon_k + \frac{3\gamma_k}{2N_k} \}, \quad (25)$$

$$F'_{i,j^*} = \begin{cases} 1/(1 + |\mathcal{S}'(\mathcal{D})|e^{\epsilon_0}), & \text{if } g_i \geq 0 \\ 1/(1 + |\mathcal{S}'(\mathcal{D})|e^{-\epsilon_0}), & \text{otherwise} \end{cases} \quad (26)$$

$$F'_{i,j} = \begin{cases} e^{\epsilon_0} F'_{i,j^*}, & \text{if } g_i \geq 0 \\ e^{-\epsilon_0} F'_{i,j^*}, & \text{otherwise} \end{cases}, \forall j \neq j^*. \quad (27)$$

Proof: See Appendix C. ■

The overall procedure is summarized in Algorithm 1, shown at the top of the next page. In Step 9, we include an early-stopping criterion when the change in the objective value of (20a) between two consecutive iterations falls below a small threshold. Since the objective is bounded below by zero and each iteration produces a non-decreasing sequence of objectives, convergence of Algorithm 1 is guaranteed.

Algorithm 1: The proposed DP mechanism.

- 1: **Input:** $\epsilon, \delta, N, d, \pi(\mathcal{D}), \{\mu_k, \Sigma_k\}, \{\mu_k(\mathcal{D}_{n,k'})\}, \mathcal{S}'(\mathcal{D})$.
 - 2: Initialize $\epsilon_0 = \epsilon_k = \epsilon/3, \forall k$.
 - 3: **for** $\text{iter} = 1, 2, \dots, I$ **do**
 - 4: **for** $k = 1, 2, \dots, K$ **in parallel do**
 - 5: Update γ_k by (22);
 - 6: Update Γ_k^{-1} and ϵ_k by solving (23);
 - 7: **end for**
 - 8: Update \mathbf{F} and ϵ_0 by (25)–(27);
 - 9: **if** $|\text{Obj}(\cdot; \text{iter}) - \text{Obj}(\cdot; \text{iter} - 1)| \leq 10^{-3}$, **early stop**;
 - 10: **end for**
 - 11: Compute Γ_k by the inverse of $\Gamma_k^{-1}, \forall k$;
 - 12: Draw $\{\mathbf{w}_k, \mathbf{W}_K\}$ and $\tilde{\pi}$ by (5)–(6) and $f(\tilde{\pi}|\pi)$.
 - 13: **Output:** $\tilde{\pi}$ and $\{\tilde{\mu}_k, \tilde{\Sigma}_k\}_{\forall k}$.
-

Remark 2 (Complexity Analysis). The primary computational cost of Algorithm 1 arises from solving the SDP problem in (23) and computing \mathbf{F} via (26) and (27). According to [15], the SDP in (23) involves a $d \times d$ matrix variable and $\mathcal{O}(NK^2)$ constraints. Solving the problem via an interior-point method incurs a complexity of $\mathcal{O}(N^{3.5}K^6)$ in the regime of $N \gg d$ [15]. Meanwhile, the computation in (26) and (27) requires $\mathcal{O}(K^4)$ floating-point operations. Consequently, the overall complexity of Algorithm 1 is on the order of $\mathcal{O}(N^{3.5}K^6)$. Notably, this complexity grows in a polynomial order with the dataset size N , due to our restriction that the probability mapping matrix \mathbf{F} only takes non-zero entries for $\pi(\mathcal{D})$ and its adjacent frequency vectors, thus avoiding an exponential increase with N .

V. EXPERIMENTAL RESULTS

In this section, we present the experimental results to evaluate the performance of the proposed approach.

A. Results on Synthetic Data

We begin by assessing the performance of the proposed method on a synthetic dataset drawn from a GMM. Unless otherwise specified, the simulation setup is as follows. First, we generate a ground-truth GMM with the following parameters: the class frequency vector is drawn from a Dirichlet distribution with $K = 5$ categories and concentration parameters equal to one; the mean of each component is drawn from a uniform distribution within the range $[-10, 10]^d$, with data dimension $d = 3$; and the covariance of each component is drawn from a Wishart distribution with $d + 1$ degrees of freedom and a scale matrix \mathbf{I}_d .

Next, we use the GMM with these parameters to generate $N = 1000$ independent and identically distributed (i.i.d.) samples to form the input dataset \mathcal{D} . This dataset is then fitted to an empirical GMM $p(\mathbf{x}, y)$ as defined in (1) using the approach outlined in (2). We then apply the proposed DP mechanism to compute a differentially private GMM $\tilde{p}(\mathbf{x}, y)$ using Algorithm 1, with a predefined (ϵ, δ) -DP requirement. Unless otherwise specified, we set $\delta = 10^{-5}$ and adjust ϵ to

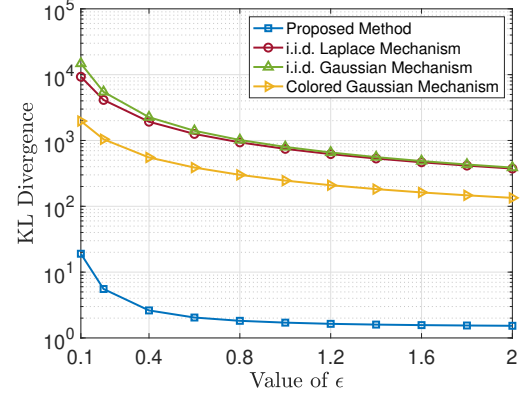


Fig. 1: KL divergence versus the privacy level in terms of the value of ϵ .

control the privacy level. The accuracy of the released model is measured by the KL divergence between the released GMM and the non-private model, i.e., $KL(\tilde{p}(\mathbf{x}, y)|p(\mathbf{x}, y))$.

We compare the performance of the proposed method against the following existing DP mechanisms:

- **i.i.d. Laplace mechanism** [6, Algorithm 2]: This baseline method adds i.i.d. Laplace noise to the estimated parameters $\{\pi_k, \mu_k, \Sigma_k\}_{k=1}^K$ and then projects the perturbed parameters back onto the feasible set. As demonstrated in [6], the noise added is determined by the ℓ_1 sensitivity of the individual parameters. This mechanism guarantees $(\epsilon, 0)$ -DP, making it a stronger privacy condition and sufficient for achieving (ϵ, δ) -DP for any $\delta > 0$.
- **i.i.d. Gaussian mechanism** [2]: In this baseline method, i.i.d. Gaussian noise is added to the estimated parameters $\{\pi_k, \mu_k, \Sigma_k\}_{k=1}^K$, followed by a projection of the perturbed parameters back onto the feasible set to ensure the release satisfies the (ϵ, δ) -DP constraint. The variance of the added noise is determined by the ℓ_2 sensitivity of the parameters. Based on the norm inequality, the ℓ_2 sensitivity is bounded by the corresponding ℓ_1 sensitivity, which is analytically expressed in [6].
- **Colored Gaussian mechanism** [4, Eq. (12)]: Originally designed to enhance DP for K -Means clustering, this method is adapted here to protect the mean vectors of the GMM components. Unlike the standard i.i.d. Gaussian mechanism, colored Gaussian noise is added to the mean components $\{\mathbf{w}_k\}_{k=1}^K$ according to the approach in (5). The covariance matrix $\Gamma_k^{-1}, \forall k$, is optimized by minimizing the MSE between the perturbed and true means, subject to the (ϵ, δ) -DP constraint.

For a fair comparison, we ensure that all the above DP mechanisms all satisfy the (ϵ, δ) -DP with the same ϵ and δ . Note that Laplace mechanisms typically achieve $(\epsilon, 0)$ -DP, which is a stronger condition and, therefore, sufficient to guarantee (ϵ, δ) -DP for any $\delta > 0$. All results are averaged over 1000 Monte Carlo trials.

We begin by examining the utility-privacy trade-off for releasing GMMs. Figure 1 plots the KL divergence of the DP method under varying privacy constraints, as the value of ϵ

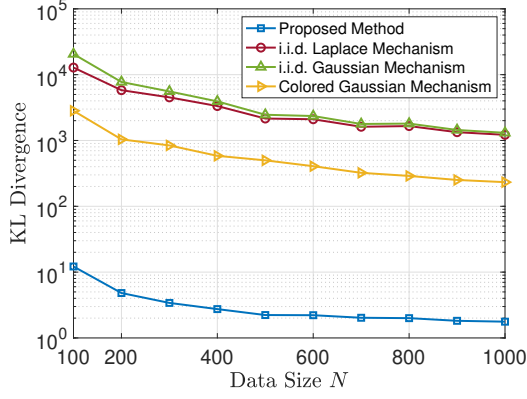


Fig. 2: KL divergence versus the data size N with $\epsilon = 1$.

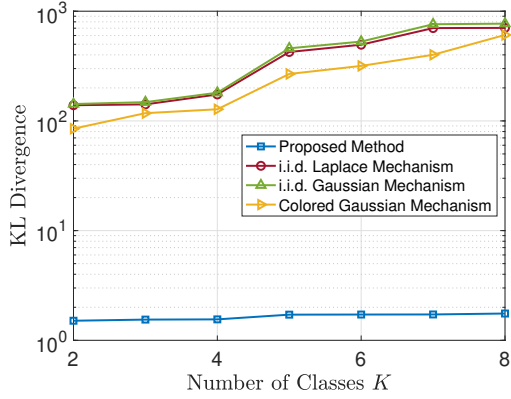


Fig. 3: KL divergence versus the number of classes K , where the total data size is fixed to $N = 1000$.

changes. As expected, a larger ϵ results in more artificial noise, leading to greater distortion in the GMM parameters, which is reflected in a higher KL divergence. The proposed approach consistently achieves a significantly smaller KL divergence than all the other baselines, thanks to its capability to directly minimize this utility measure.

Next, we fix $\epsilon = 1$ and examine the KL divergence performance as a function of the total number of data points N , as shown in Figure 2. We see that the KL divergence decreases as N increases under the same DP constraint. Intuitively, DP assesses the probability of inferring the presence or properties of a particular data point in the entire dataset. Since the data points are i.i.d. and drawn from the same GMM distribution, a larger dataset naturally has a greater capacity to obscure individual data points. As a result, it requires less noise to achieve the same level of privacy. This observation aligns with the DP analysis in Section III.

Another important hyperparameter that affects the performance of differentially private GMMs is the number of data classes K . By fixing $N = 1000$ and $\epsilon = 1$, we plot the KL divergence performance against varying values of K in Figure 3. As shown in Theorem 1, the privacy loss from releasing the component mean \mathbf{w}_k and covariance \mathbf{W}_k for each class k is independently determined by the statistics of the added noise.

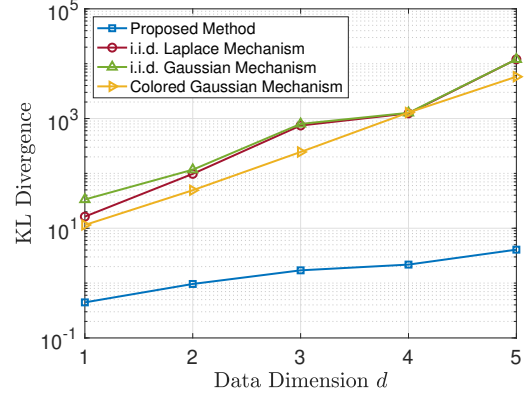


Fig. 4: KL divergence versus the dimension of the data points d with $K = 5$ and $N = 1000$.

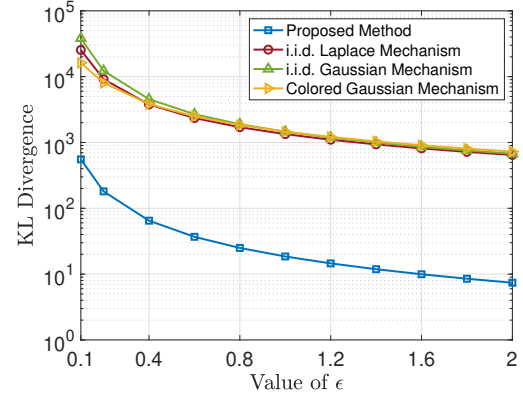


Fig. 5: KL divergence versus the value of ϵ for releasing the GMM fitting to the *Iris* dataset.

The overall privacy loss for releasing $\{\tilde{\mathbf{w}}_k, \tilde{\mathbf{W}}_k\}_{\forall k}$ is governed by the maximum privacy loss for any class k , making it a non-decreasing function of K . Therefore, the KL divergence required to achieve the same privacy constraint increases with K , as confirmed by the results in Figure 3. However, the increase in KL divergence for the proposed approach is much less sensitive to K than in the baseline methods, highlighting the robustness of the proposed approach.

Finally, we evaluate the KL divergence with respect to the data dimension d in Figure 4. From Theorem 1 and Proposition 1, it can be seen that under the same statistics, both the KL divergence and the privacy loss increase with d . Intuitively, due to the curse of dimensionality, data points become sparsely distributed in high-dimensional spaces, increasing their susceptibility to privacy risks as adjacent datasets become easier to distinguish. As a result, more perturbations are required to maintain the same level of DP. As shown in Figure 4, the accuracy of all DP mechanisms significantly deteriorates as d increases, demonstrating the sensitivity of differentially private GMMs to data dimensionality.

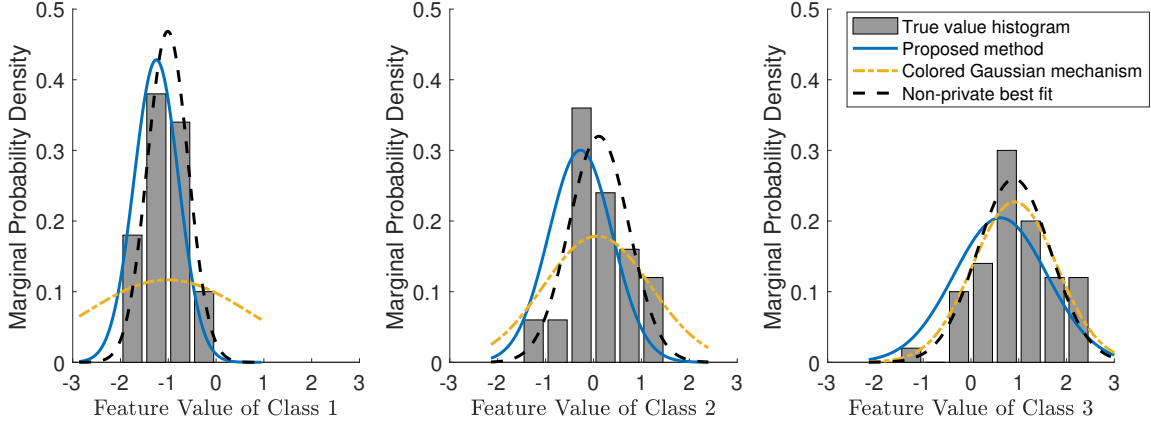


Fig. 6: Marginal probability density of GMMs and empirical histograms of the true values fitted to the *Iris* dataset for the first feature, i.e., the first entry of the data points $\mathbf{x}_{n=1}^N$. Each subplot displays the conditional density for one class.

B. Results on the UCI Machine Learning Dataset

In this section, we evaluate the performance of the proposed DP method using real-world multi-class datasets. We employ the *Iris* dataset from the UCI Machine Learning Repository [16], which is a widely recognized benchmark for classification tasks. The dataset consists of $K = 3$ classes, each representing a type of iris plant, with a total of $N = 150$ instances. Each instance is characterized by a feature vector of $d = 4$ dimensions, measuring the sepal length, sepal width, petal length, and petal width. These features are normalized to have zero mean and unit variance. Notably, the first class (Iris-Setosa) is linearly separable from the other two, while the latter two classes (Iris-Versicolor and Iris-Virginica) exhibit overlapping distributions, making them harder to distinguish.

We fit the dataset to an empirical GMM $p(\mathbf{x}, y)$, as defined in (1), using the method outlined in (2). Subsequently, we apply the proposed DP mechanism to compute a differentially private GMM $\tilde{p}(\mathbf{x}, y)$, with a predefined (ϵ, δ) -DP constraint, where we choose $\epsilon = 2$ and $\delta = 10^{-5}$. Figure 6 visualizes the marginal probability density for the first feature (normalized sepal length) for each class. The gray bars represent the empirical histograms of the true feature values, and the black dashed lines represent the fitted GMM density in (2) without applying DP. The results show that the GMM produced by our method aligns more closely with the true values compared to the baseline method.

Moreover, Figure 5 illustrates the KL divergence between the differentially private GMM and the non-private model, under varying levels of privacy determined by ϵ . Our approach exhibits a superior utility-privacy trade-off, achieving a smaller KL divergence under the same DP constraints. The results in Figures 5 and 6 further justify the use of KL divergence as a utility metric, since a smaller divergence indicates a better fit to the ground truth, demonstrating the effectiveness of our method in preserving data accuracy while maintaining privacy.

Data class	# of data points
Class 1	687
Class 2	428
Class 3	102
Class 4	147
Class 5	14
Class 6	31
Total	1409

TABLE I: The number of data points in each class of the *AMI* data.

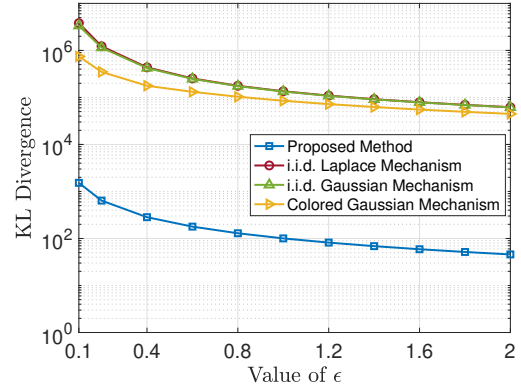


Fig. 7: KL divergence versus the value of ϵ for releasing the GMM fitting to the *AMI* load profile data.

C. Results on Electricity Load Profile Data

We explore the application of differentially private GMMs in real-world scenarios, specifically within the context of power systems. Electricity load demand profiles, which represent the amount of electricity consumption over time, are essential for system analysis, anomaly detection, and fault diagnosis in power systems. These load profile data are typically collected by a utility provider and shared with an analyst for further analysis. It has been shown in [4] that



Fig. 8: Entry-wise marginal probability density of GMMs fitted to the AMI load profile data for each class. The subplots in each column correspond to the marginal density of one class. In these plots, the gray bars represent the empirical histograms of the true values, the blue curves represent the marginal density of the GMM computed using the proposed DP method, and the red curves represent the marginal density of the non-private GMMs computed using (2).

the logarithmic values of load demand profiles fit well to GMMs. However, recent studies have highlighted significant privacy risks associated with releasing load profile data or their statistics, motivating the release of differentially private GMMs for load profile synthesis and analysis.

Here, we examine the performance of fitting and releasing GMMs on real-world load profile data under DP constraints. Following [4], we utilize a dataset from a real-world *AMI* system, which includes data from 1409 houses spread across 12 distribution circuits in California, USA, yielding $N = 1409$ samples. Each sample represents the half-day electricity consumption profile of a single house, with measurements taken hourly over $d = 12$ consecutive hours. These data points are then transformed by taking the logarithm of the power consumption. We then apply K -Means clustering [17] to group the data into $K = 6$ classes, representing different types of electricity consumers (e.g., residential, commercial, agricultural). Table I summarizes the number of data points in each class.

We proceed by fitting the dataset to a GMM using the method in (2) and adding artificial noise to achieve (ϵ, δ) -DP. Figure 7 shows the KL divergence between the released GMM and the non-private model, plotted against different values of ϵ , with $\delta = 10^{-5}$ fixed. The results reveal that our proposed method significantly improves the utility in terms of KL divergence compared to the baseline algorithms, effectively balancing the privacy-utility trade-off.

Next, we fix $\epsilon = 2$ and plot the marginal density of the released GMM for each dimension and class in Figure 8. For Classes 1-4, the density values closely align with the fitted GMM without DP (represented by the red curves) and the ground-truth histograms (represented by the gray bars). However, for Classes 5 and 6, the release of their parameters is more sensitive due to the limited data available, requiring significantly more noise to maintain DP. As a result, the density values from the proposed approach diverge from those of the non-private models in some dimensions for these classes. This highlights the sensitivity of differentially private GMMs to sample size, particularly when fewer data points are available.

D. Results on Image Classification

GMMs are widely adopted models for classification. In this section we explore their use in a differentially private setting for image classification. We consider the well-known *MNIST* dataset [18], which contains 28×28 grayscale images of handwritten digits. We randomly sample $N = 5000$ training images and reduce the dimensionality from 784 to $d = 5$ using principal component analysis [19] to mitigate the curse of dimensionality. Next, we fit these data to a 10-class GMM using the approach in (2) and subsequently apply the DP mechanism to enforce an (ϵ, δ) -DP constraint with $\delta = 10^{-5}$.

Figure 9 plots the KL divergence between the differentially private GMM and the non-private model as a function of the privacy parameter ϵ . The proposed approach achieves a significantly lower divergence, indicating that the released GMM classifier closely approximates the non-private model.

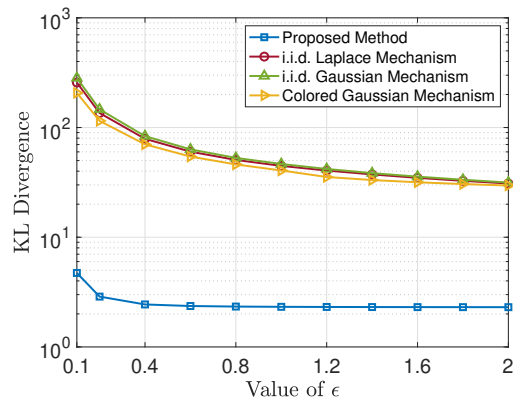


Fig. 9: KL divergence versus the value of ϵ for releasing the GMM fitting to the *MNIST* training data.

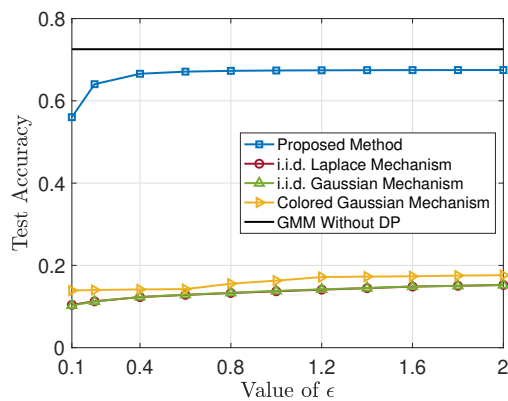


Fig. 10: Test accuracy versus the value of ϵ over the *MNIST* testing data.

Furthermore, we evaluate classification performance by predicting the labels of 10^4 test images, where the class for each test vector is determined via maximum likelihood estimation. Fig. 10 plots the test accuracy of different classifiers, measured by the fraction of correctly predicted labels in the range of $[0, 1]$. We see that our method achieves an accuracy close to the non-private baseline, while the DP baseline approaches suffer from high distortion due to added noise. These results validate the effectiveness of using KL divergence as a utility metric, as a smaller divergence corresponds to a closer match between the training and testing data distributions, thereby enhancing classification performance.

VI. CONCLUSIONS

In this paper, we addressed the privacy protection challenge associated with releasing the parameters of GMMs, including mixture weights, component means, and covariance matrices. We proposed the use of KL divergence as a utility metric to quantify the accuracy of the released GMM, which effectively captures the combined impact of noise perturbation on individual parameters. We then introduced a DP mechanism designed to protect the privacy of GMM parameters. Our analysis reveals the impact of privacy budget allocation and

noise statistics on DP and also offers a tractable expression for evaluating the KL divergence utility. Building on this analysis, we formulated and solved an optimization problem that minimizes the KL divergence between the released and original models, subject to a given (ϵ, δ) -DP constraint. Experimental results on both synthetic and real-world datasets demonstrate the effectiveness of our approach, highlighting its superior performance in achieving a balance between privacy and accuracy.

APPENDIX A PROOF OF THEOREM 1

We first characterize the DP of releasing $\{\tilde{\mu}_k, \tilde{\Sigma}_k\}$ using the following established results.

Lemma 1 (Multivariate Gaussian mechanism [4]). Given \mathcal{D} , the mechanism for publishing $\tilde{\mu}_k$ in (5) satisfies (ϵ_k, δ) -DP given that the following inequality holds for any adjacent dataset $\mathcal{D}'_{n,k'}$:

$$\begin{aligned} & \frac{\epsilon_k^2}{2 \log(2/\delta)} \\ & \geq \sup_{n,k'} \left\{ \mu_k(\mathcal{D}) - \mu_k(\mathcal{D}'_{n,k'}) \right\}^T \Gamma_k \left(\mu_k(\mathcal{D}) - \mu_k(\mathcal{D}'_{n,k'}) \right). \end{aligned}$$

Proof: See [4, Theorem 3]. ■

Lemma 2 (Wishart mechanism [10]). The mechanism for publishing $\tilde{\Sigma}_k$ in (6) satisfies $(\epsilon_0, 0)$ -DP given that

$$3\gamma_k \leq 2N_k \epsilon_0.$$

Proof: See [10, Theorem 4]. ■

Lemma 3 (Parallel composition). Suppose publishing $\tilde{\mu}_k$ and $\tilde{\Sigma}_k$ of an individual class k satisfies (ϵ'_k, δ) -DP. Then, publishing $\{\tilde{\mu}_k, \tilde{\Sigma}_k\}_{k=1}^K$ of all the classes satisfies $(\max_{k=1}^K \epsilon'_k, \delta)$ -DP.

Proof: The result directly follows from the standard parallel mechanism of DP; see, e.g., [5]. ■

Combining the above lemmas, it follows that releasing $\{\tilde{\mu}_k, \tilde{\Sigma}_k\}_{k=1}^K$ adheres to $(\max_{k=1}^K \left\{ \epsilon_k + \frac{3\gamma_k}{2N_k} \right\}, \delta)$ -DP.

Next, we characterize the DP of releasing $\tilde{\pi}$. From Definition 4, it can be verified that $\tilde{\pi}$ satisfies $(\epsilon_0, 0)$ -DP if the following holds for any adjacent dataset $\mathcal{D}'_{n,k'}$:

$$e^{-\epsilon_0} \Pr(\tilde{\pi} | \pi(\mathcal{D}'_{n,k'})) \leq \Pr(\tilde{\pi} | \pi(\mathcal{D})) \leq e^{-\epsilon_0} \Pr(\tilde{\pi} | \pi(\mathcal{D}'_{n,k'})). \quad (28)$$

Combining the definition of $f(\tilde{\pi} | \pi)$, it follows that the condition in (28) is equivalent to (11).

Finally, the overall DP can be obtained by using the following well-know composition theorem:

Lemma 4 (Sequential composition). Let M_1 and M_2 be two randomized mechanisms that apply to an input dataset \mathcal{D} and satisfy (ϵ_1, δ) -DP and $(\epsilon_2, 0)$ -DP, respectively. Then, the sequential execution of them satisfies $(\epsilon_1 + \epsilon_2, \delta)$ -DP.

Proof: See [5]. ■

According to Lemma 4, releasing $\{\tilde{\pi}_k, \tilde{\mu}_k, \tilde{\Sigma}_k\}_{k=1}^K$ satisfies $(\max_{k=1}^K \left\{ \epsilon_k + \frac{3\gamma_k}{2N_k} \right\} + \epsilon_0, \delta)$ -DP.

APPENDIX B PROOF OF PROPOSITION 1

We expend the KL divergence formula in (8) by substituting (1) and (7), yielding the expression in (29) shown on top of the next page. We note that the last term on the right-hand side of (29) is the KL divergence of two Gaussian distributions. Using the formula of the KL divergence between two multivariate Gaussian distributions [20], we have

$$\begin{aligned} & \mathbb{E} \left[\int \mathcal{N}(\tilde{\mu}_k, \tilde{\Sigma}_k) \ln \frac{\mathcal{N}(\tilde{\mu}_k, \tilde{\Sigma}_k)}{\mathcal{N}(\mu_k, \Sigma_k)} d\mathbf{x} \right] \\ &= \frac{1}{2} \mathbb{E} \left[\mathbf{w}_k^T \Sigma_k^{-1} \mathbf{w}_k - d - \ln \frac{|\tilde{\Sigma}_k|}{|\Sigma_k|} + \text{tr}(\Sigma_k^{-1} \tilde{\Sigma}_k) \right] \\ &= \frac{1}{2} \left(\text{tr}(\Sigma_k^{-1} \Gamma_k^{-1}) + \mathbb{E} \left[-\ln \frac{|\tilde{\Sigma}_k|}{|\Sigma_k|} + \text{tr}(\Sigma_k^{-1} \mathbf{W}_k) \right] \right) \\ &= \frac{1}{2} \left(\text{tr}(\Sigma_k^{-1} \Gamma_k^{-1}) + \frac{d+1}{\gamma_k} \text{tr}(\Sigma_k^{-1}) + d \ln \gamma_k \right) \\ & \quad - \frac{d \ln 2 + \psi_d(\frac{d+1}{2})}{2}, \end{aligned} \quad (30)$$

where $\psi_d(\cdot)$ is the multivariate digamma function. Combining (29) and (30) completes the proof.

APPENDIX C PROOF OF PROPOSITION 2

First, we prove by contradiction that the solution satisfies $F'_{i,j^*} > 0$ for $\forall i$. Suppose there exist some i' such that $F'_{i',j^*} = 0$. The corresponding constraint becomes $e^{-\epsilon_0} F'_{i',j} \leq F'_{i',j^*} = 0$ implies $F'_{i',j} = 0$ for any $j \neq j^*$. This contradicts with the row-stochastic constraint of $\sum_j F'_{i,j} = 1 > 0$.

Next, define an auxiliary variable $\theta_{i,j} = \frac{F'_{i,j}}{F'_{i,j^*}}$ for $\forall j \neq j^*$ and $\forall i$. The optimization problem is equivalent to:

$$\begin{aligned} & \min_{\epsilon_0, \{\theta_{i,j}\}, \{F'_{i,j^*}\}_{\forall i}} \sum_i g_i F'_{i,j^*} \\ & \text{s.t.} \quad 0 \leq \epsilon_0 \leq \epsilon - \max_k \left\{ \epsilon_k + \frac{3\gamma_k}{2N_k} \right\}, \\ & \quad e^{-\epsilon_0} \leq \theta_{i,j^*} \leq e^{\epsilon_0}, \forall j \neq j^*, \forall i. \\ & \quad F'_{i,j^*} \left(1 + \sum_{j \neq j^*} \theta_{i,j} \right) = 1, \forall i. \end{aligned}$$

From the last two constraints, it follows that

$$F'_{i,j^*} = \frac{1}{1 + \sum_{j \neq j^*} \theta_{i,j}} \in \left[\frac{1}{1 + |S'(\mathcal{D})|e^{\epsilon_0}}, \frac{1}{1 + |S'(\mathcal{D})|e^{-\epsilon_0}} \right].$$

Consequently, for any given ϵ_0 , the minimization of the linear objective over $\{F'_{i,j^*}\}_{\forall i}$ depends on each sign of g_i . Specifically, for any i we consider the following two subcases:

- If $g_i \geq 0$, the optimal solution is achieved when $F'_{i,j^*} = 1/(1 + |S'(\mathcal{D})|e^{\epsilon_0})$.
- If $g_i < 0$, the optimal solution is achieved when $F'_{i,j^*} = 1/(1 + |S'(\mathcal{D})|e^{-\epsilon_0})$.

The final step is to find the optimal solution for ϵ_0 . Let $A = \sum_{i: g_i > 0} g_i$ and $B = \sum_{i: g_i < 0} g_i$. The problem can be recast as the minimization of a one-dimensional function $J(\epsilon_0) =$

$$\begin{aligned}
(8) &= \mathbb{E} \left[\sum_{k=1}^K \tilde{p}(y=k) \int \tilde{p}(\mathbf{x}|y=k) \ln \frac{\tilde{p}(y=k) \tilde{p}(\mathbf{x}|y=k)}{p(y=k) p(\mathbf{x}|y=k)} d\mathbf{x} \right] \\
&= \mathbb{E}_{\tilde{\pi}|\pi(\mathcal{D})} \left[\sum_{k=1}^K \tilde{p}(y=k) \left(\ln \frac{\tilde{p}(y=k)}{p(y=k)} + \mathbb{E}_{\{\mathbf{r}_k, \gamma_k\}_{\forall k}} \left[\int \tilde{p}(\mathbf{x}|y=k) \ln \frac{\tilde{p}(\mathbf{x}|y=k)}{p(\mathbf{x}|y=k)} d\mathbf{x} \right] \right) \right] \\
&= \sum_{\tilde{\pi} \in \mathcal{S}} f(\tilde{\pi}|\pi(\mathcal{D})) \sum_{k=1}^K \tilde{\pi}_k \left(\ln \frac{\tilde{\pi}_k}{\pi_k} + \mathbb{E}_{\{\mathbf{w}_k, \mathbf{w}_k\}_{\forall k}} \left[\int \mathcal{N}(\tilde{\boldsymbol{\mu}}_k, \tilde{\boldsymbol{\Sigma}}_k) \ln \frac{\mathcal{N}(\tilde{\boldsymbol{\mu}}_k, \tilde{\boldsymbol{\Sigma}}_k)}{\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} d\mathbf{x} \right] \right). \tag{29}
\end{aligned}$$

$\frac{A}{1+|\mathcal{S}'(\mathcal{D})|e^{\epsilon_0}} + \frac{B}{1+|\mathcal{S}'(\mathcal{D})|e^{-\epsilon_0}}$ over the range of $0 < \epsilon_0 \leq \epsilon - \max_k \{\epsilon_k + \frac{3\gamma_k}{2N_k}\}$. The first-order derivative of $J(\epsilon_0)$ is given by

$$J'(\epsilon_0) = \frac{Be^{\epsilon_0}|\mathcal{S}'(\mathcal{D})|}{(1+|\mathcal{S}'(\mathcal{D})|e^{-\epsilon_0})^2} - \frac{Ae^{\epsilon_0}|\mathcal{S}'(\mathcal{D})|}{(1+|\mathcal{S}'(\mathcal{D})|e^{\epsilon_0})^2} \leq 0.$$

Consequently, $J(\epsilon_0)$ is non-increasing in ϵ_0 , and the optimal solution is given by the upper bound of ϵ_0 .

To summarize, the optimal solution is given by

$$\begin{aligned}
\epsilon_0 &= \epsilon - \max_k \left\{ \epsilon_k + \frac{3\gamma_k}{2N_k} \right\}, \\
F'_{i,j^*} &= \begin{cases} 1/(1+|\mathcal{S}'(\mathcal{D})|e^{\epsilon_0}), & \text{if } g_i \geq 0 \\ 1/(1+|\mathcal{S}'(\mathcal{D})|e^{-\epsilon_0}), & \text{otherwise} \end{cases} \\
F'_{i,j} &= \begin{cases} e^{\epsilon_0} F'_{i,j^*}, & \text{if } g_i \geq 0 \\ e^{-\epsilon_0} F'_{i,j^*}, & \text{otherwise} \end{cases}, \forall j \neq j^*.
\end{aligned}$$

REFERENCES

- [1] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*. Springer, 2006, vol. 4, no. 4.
- [2] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. Theory Cryptogr. Conf.*, 2006, pp. 265–284.
- [3] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 381–396, 2002.
- [4] N. Ravi, A. Scaglione, S. Kadam, R. Gentz, S. Peisert, B. Lunghino, E. Levijarvi, and A. Shumavon, "Differentially private k-means clustering applied to meter data analysis and synthesis," *IEEE Trans. Smart Grid*, vol. 13, no. 6, pp. 4801–4814, 2022.
- [5] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3–4, pp. 211–407, 2014.
- [6] Y. Wu, Y. Wu, H. Peng, J. Zeng, H. Chen, and C. Li, "Differentially private density estimation via Gaussian mixtures model," in *IEEE/ACM Int. Symp. Qual. Serv. (IWQoS)*, 2016, pp. 1–6.
- [7] T. M. Cover, *Elements of Information Theory*. John Wiley & Sons, 1999.
- [8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
- [9] F. Hellström, G. Durisi, B. Guedj, and M. Raginsky, "Generalization bounds: Perspectives from information theory and PAC-Bayes," *Found. Trends Mach. Learn.*, vol. 18, no. 1, pp. 1–223, 2025.
- [10] W. Jiang, C. Xie, and Z. Zhang, "Wishart mechanism for differentially private principal components analysis," *Proc. AAAI Conf. Artif. Intell.*, vol. 30, no. 1, Feb. 2016.
- [11] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc. Ser. B (Stat. Methodol.)*, vol. 39, no. 1, pp. 1–22, 1977.
- [12] C. Zheng, G. Wu, and C. Li, "Toward understanding generative data augmentation," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, 2023, pp. 54046–54060.
- [13] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2011.

- [14] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming," <http://cvxr.com/cvx>, Mar. 2014.
- [15] B. Stephen and L. Vandenberghe, *Convex Optimization*. U. K.: Cambridge University Press, 2004.
- [16] R. A. Fisher, "Iris," UCI Machine Learning Repository, 1936. [Online]. Available: <https://doi.org/10.24432/C56C76>
- [17] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. Fifth Berkeley Symp. Math. Stat. Prob.*, vol. 5, 1967, pp. 281–298.
- [18] L. Deng, "The MNIST database of handwritten digit images for machine learning research," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 141–142, 2012.
- [19] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Phil. Trans. R. Soc. A.*, vol. 374, p. 20150202, 2016.
- [20] J. Duchi, "Derivations for linear algebra and optimization," *Berkeley, California*, vol. 3, no. 1, pp. 2325–5870, 2007.