Politecnico di Milano
School of Industrial and Information Engineering
Master of Science in Mathematical Engineering

POLITECNICO
MILANO 1863

Bayesian Nonparametric Privacy-Preserving Synthetic Data Generation

Advisor: Prof. Mario Beraha
Co-advisor: Prof. Stefano Favaro

M.Sc. Thesis of:
Riccardo Lazzarini

Academic Year 2022-2023

POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Bayesian Nonparametric Privacy-Preserving Synthetic Data Generation

Tesi di Laurea Magistrale in
Mathematical Engineering - Ingegneria Matematica

Riccardo Lazzarini, 10692988

Advisor:
Prof. Mario Beraha

Co-advisor:
Prof. Stefano Favaro

Academic year:
2022-2023

Abstract: This work addresses the problem of generating synthetic data from a statistical perspective. In particular, we review some methods known in the literature to generate synthetic data and introduce a new release mechanism based on a Bayesian Nonparametric model. We model the private data as a sample from a Bayesian Nonparametric Prior, namely the Pitman-Yor process, and generate new data by sampling from the posterior predictive distribution. We show the assumptions required by the mechanism to satisfy differential privacy. We study the advantages of our method both in frequentist terms, showing a convergence result for the mean of the 1-Wasserstein distance between the empirical measure based on the released data and the data generating distribution, and in Bayesian terms, showing that it is possible, by releasing data from the predictive distribution, to make posterior inference easily on the private data.

Key-words: Differential Privacy, Synthetic Data, Bayesian Nonparametrics, Pitman-Yor process

1. Introduction

1.1. The problem of releasing data to the public

National agencies and private companies routinely gather sensitive data, e.g. pertaining to individuals' economic, social and health status. Sharing these data with the community or the industry is fundamental to understanding social or scientific phenomena and fostering the development of more advanced machine learning methods. At the same time, data custodians must safeguard the privacy of individuals associated with the data, ensuring their confidential information remains protected.
Different techniques have been introduced to quantify the risk of disclosure when data containing sensitive information are disseminated, such as *k-anonymity* (Sweeney (2002)) or *l-diversity* (Machanavajjhala et al. (2006)). Contemporaneously, various strategies have been developed to avoid disclosing sensitive information. However, traditional techniques like swapping (swapping values between records within a dataset) or suppression (removing certain values from the dataset), which once were effective, are now deemed insufficient to protect the privacy of individuals due to advancements in computing power and the widespread availability of data.
A promising solution to balance the need for data accessibility with privacy concerns is using synthetic data. Synthesising data involves fitting a (statistical) model based on the original dataset and generating simulated data points from this model. Intuitively, if a model can capture the key features of the data on which it is

based, the synthetic data generated from it will allow us to obtain the same answers as we would have had from analysing the original data.
In the statistical community, the idea of releasing synthetic data was initially introduced by Rubin (1993), who proposed a technique known as multiple imputation (MI) to replace sensitive values with "imputed" ones. However, it is only with the framework of differential privacy introduced in the work by Dwork et al. (2006b) that attention began to be focused on the mechanism used to generate the synthetic data so that this would provide mathematically rigorous quantifiable privacy guarantees. In short, differential privacy requires that a change in one record in the database has little impact on the result obtained through the mechanism used. As explained later, knowing that a mechanism satisfies privacy differential leads us to conclude that the synthetic data generated do not allow us to understand information about the individuals to whom the data relate. The differential privacy definition has achieved remarkable success, evident in its widespread adoption as a crucial privacy tool by Google (Erlingsson et al. (2014)), Apple (Apple's Differential Privacy Team (2017)), Microsoft (Ding et al. (2017)) and the US Census Bureau (Abowd et al. (2022)), alongside its substantial presence in academic literature, which continues to expand.
In statistical theory, the work by Rinott et al. (2018) explores methods to protect confidentiality when sharing frequency tables, focusing on the concept of differential privacy. It also discusses techniques to balance data utility and privacy guarantees. Indeed, a main challenge of considerable statistical interest is to ensure that data released by differentially private mechanisms are still useful for making inferences on private data. A landmark paper in this direction is the work by Wasserman and Zhou (2010), where the authors propose various release mechanisms based on nonparametric frequentist statistics satisfying differential privacy and introduce the definition of a "consistent" or "informative" mechanism, i.e. a mathematical request made to preserve the utility of the released data. More recently, Duchi et al. (2018) focused on determining the convergence rate in a minimax framework of estimators based on the synthetic data for various estimation problems. In this regard, we also mention the paper by Butucea et al. (2020), which studies local differential privacy and its impact on optimal density estimation.
As far as the Bayesian approach is concerned, the mechanisms proposed in the literature so far, up to our knowledge, are parametric. In this regard, we recall the work by Dimitrakakis et al. (2017), in which the authors show that under appropriate conditions over the prior, releasing samples from the posterior satisfies differential privacy. Hu et al. (2022) present a novel approach to achieve differential privacy by embedding Bayesian models for synthetic data generation and utilising a censoring mechanism. Bernstein and Sheldon (2019) focus on Bayesian linear regression and likelihoods belonging to the exponential family, showing that, with suitable perturbations, it is possible to generate privacy-preserving data, maintaining the posterior easy to compute. Jewson et al. (2023) introduce a new approach to provide differentially private estimates via posterior sampling in complex classifiers and continuous regression models, including neural networks. Savitsky et al. (2022) propose a mechanism based on a pseudo posterior that satisfies differential privacy. Lastly, the work of Machanavajjhala et al. (2008) generates synthetic data distributed as the posterior predictive distribution of a Dirichlet-Multinomial model.

1.2. Our contribution

In this thesis, we extend the idea of generating synthetic data by sampling from posterior predictive distribution to a Bayesian nonparametric model. Using a discrete Bayesian Nonparametric prior lends itself well to privacy problems where the data are typically discrete, and the number of data categories can be huge. In particular, we consider a Pitman-Yor (PY) process prior, introduced by Pitman and Yor (1997), for modelling the private data, and we disseminate new data distributed as the m-steps posterior predictive distribution. The choice of the PY process as a Bayesian nonparametric prior is a fair compromise between mathematical tractability and model flexibility. Moreover, it includes the well-known Dirichlet Process (DP) prior (Ferguson (1973)) as a special case.
We establish sufficient conditions under which our proposed mechanism satisfies a notion of differential privacy. In a particular case, we find a condition similar to that required by Machanavajjhala et al. (2008) for the Dirichlet-multinomial model. In the general case, on the other hand, the assumption necessary involves a probability that is difficult to calculate analytically but can be straightforwardly approximated via Monte Carlo since sampling from the predictive distributions of a Pitman-Yor process is easy. In addition, we introduce a novel method to perturb the counts of a histogram following a Bayesian Nonparametric approach, and we show the assumptions it requires to have privacy guarantees.
We present the definition of consistency introduced by Wasserman and Zhou (2010), and we show that the BNP mechanism under DP prior satisfies this general (frequentist) consistency property with respect to the 1-Wasserstein distance. This result indicates that our releasing mechanism can also be helpful for statistical purposes in non-Bayesian contexts. Instead, in a Bayesian framework, we show it is possible to obtain samples from the posterior of the model assumed for the private data without using complex MCMC algorithms, contrary

to the case of data subjected to other known privacy-preserving mechanisms, such as noise addition. In this regard, we recall the work by Beraha et al. (2023), which proposes a new MCMC algorithm for making posterior inference in mixture models under differential privacy constraints. In a generic parametric model, we show that the density of the posterior distribution based on data sampled from the predictive distribution has the same analytical expression as the one obtained having the private data but replacing them with the synthetic one. In a Nonparametric setting, we show a similar result considering the predictive distribution in place of the posterior distribution. Finally, we show that, based on data released by the BNP mechanism, it is still possible to derive a closed-form expression for a Bayesian Nonparametric estimator of the missing mass, i.e. the probability of observing a new value with respect to those observed in the private data.

1.3. Outline of the thesis

In Section 2, we introduce the reader to the concept of differential privacy, providing the main definitions: ϵ-differential privacy and (ϵ, δ)-differential privacy. Moreover, we describe the Bayesian modelling framework and some examples of discrete Bayesian nonparametric priors. In Section 3, we review some methods of synthetic data generation commonly used in the statistical literature, both frequentist and Bayesian (parametric). Section 4 describes the BNP mechanism and presents the assumptions required to satisfy differential privacy. Section 5 contains a natural generalization of the definition of consistency for a releasing mechanism and an overview of some basic concepts from the Bayesian Consistency literature. Then, these ideas are used to show the consistency of the BNP mechanism in the case of DP prior and the Dirichlet-Multinomial mechanism. In Section 7, we present the outcomes of applying our mechanism to two real datasets.

2. Background material

2.1. Introduction to Differential privacy

For $n \geq 1$, denote by $X_1, ..., X_n$ the confidential observations, defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, and taking values in a general Polish space $\mathbb{X}$, i.e. a separable completely metrizable topological space, endowed with its Borel σ-algebra $\mathcal{X}$. For $k \geq 1$, let $\mathbb{X}^k = \times_{i=1}^k \mathbb{X}$ and $\mathcal{X}^k$ the associated Borel σ-algebra. Typically, $\mathbb{X}$ is assumed to be bounded, as done by Wasserman and Zhou (2010), who take $\mathbb{X} = [0,1]^r$ for some $r \geq 1$. In statistical privacy research, the goal is to develop a random mechanism which takes in input the data $\mathbf{X} := (X_1, ..., X_n)$ and outputs $\mathbf{Z} := (Z_1, ..., Z_m)$ with $Z_i \in \mathbb{Y}$, possibly different from $\mathbb{X}$, such that if $\mathbf{Z}$ is released then individual privacy is protected. In the following, we will use the term private, sensitive or confidential data to refer to $\mathbf{X}$ and synthetic or released data to refer to $\mathbf{Z}$. We will also assume that $\mathbb{Y} \equiv \mathbb{X}$.
A first important distinction when discussing differential privacy is between *local* and *global* privacy. Indeed, the generation of $\mathbf{Z}$ depends on how the different entities that own the data in $\mathbf{X}$ can interact. In the global privacy setting, the data holders trust a common curator, who has access to the entire dataset $\mathbf{X}$ and uses this information to generate $\mathbf{Z}$. In contrast, this central figure does not exist in the case of local privacy. In the case of local differential privacy, moreover, the mechanism applied to every single datum X_i can also exploit the data generated up to that moment, i.e. $Z_1, ..., Z_{i-1}$, to release the new syntactic datum Z_i. In this case, we speak of *interactive* differential privacy; we refer the interested reader to the work of Duchi et al. (2018), which considers this framework. In this thesis, we will only consider *non-interactive* differential privacy, where the generated data depends only on the starting private data. In a local non-interactive privacy context, the mechanism by which $\mathbf{Z}$ is generated can be modelled as a collection of probability kernels $Q_i(\cdot \mid \cdot) : \mathcal{X} \times \mathbb{X} \to [0,1]$ with $i = 1, ..., n$. Each represents the law of Z_i given $X_i = x_i$. In contrast, in global privacy, the mechanism is regarded as a probability kernel $Q_n(\,\cdot \mid \cdot\,) : \mathcal{X}^m \times \mathbb{X}^n \to [0,1]$, representing the conditional distribution of $\mathbf{Z}$ given $\mathbf{X}$. Note that, by definition, each local privacy mechanism can also be seen as a global with $Q_n(\,\cdot \,|\, \mathbf{X}) = \otimes_{i=1}^n Q_i(\,\cdot \,|\, X_i)$. As can be noted, an initial and essential difference between the two privacy frameworks is the number of data released. In the case of local privacy, the dimensionality of $\mathbf{Z}$ is the same as the dimensionality of $\mathbf{X}$, while in global differential privacy, m is a parameter that can be tuned to satisfy the privacy requirements, which results in greater flexibility. An example of a mechanism shared by the two types of privacy consists of the addition of noise to the private observations, i.e. to release $(Z_1, ..., Z_n)$ where $Z_i = X_i + \epsilon_i$ and $\epsilon_1, ..., \epsilon_n$ are i.i.d. random variables with zero mean. If we assume that the noise distribution has a density g, the (global) random mechanism $Q_n(\,\cdot \mid \mathbf{X} = \mathbf{x})$ has density $q_n(z_1, ..., z_n \mid \mathbf{x}) = \prod_{i=1}^n g(z_i - x_i)$ with $\mathbf{x} := (x_1, ..., x_n)$ and it can be seen as the product measure $\otimes_{i=1}^n Q_i$ of n (local) privacy channels of density $g(\cdot - x_i)$.

2.2. Differential privacy definitions

Having introduced the main notation about data release mechanisms, we can now define differential privacy in terms of constraints imposed on them. In particular, the distinction between local and global privacy is also reflected in the definition of ϵ-differential privacy (Dwork et al. (2006b)), which can be formalized as follows:

Definition 2.1. *Let $\epsilon > 0$. We say that a collection of local privacy mechanisms Q_i for $i = 1, ..., n$ satisfies ϵ-differential privacy if*

$$\sup_{B \in \mathcal{X}^m} \frac{Q_i(B \mid X_i = x)}{Q_i(B \mid X_i = y)} \leq e^{\epsilon} \tag{1}$$

for any $i = 1, ..., n$ and $x, y \in \mathbb{X}$. The ratio in (1) is interpreted to be 1 whenever the numerator and the denominator are both 0.

The parameter ϵ measures the degree of privacy assurance: with ϵ tends to 0, privacy is absolute, whereas increasing ϵ towards infinity reduces the stringency of the privacy requirement. In the case of global privacy, the definition of ϵ-differential privacy is the following

Definition 2.2. *Let $\epsilon > 0$, we say that a global privacy mechanism Q_n satisfies ϵ-differential privacy if*

$$\sup_{\substack{\mathbf{x},\mathbf{y}: \\ h(\mathbf{x},\mathbf{y})=1}} \sup_{B \in \mathcal{X}^m} \frac{Q_n(B \mid \mathbf{X} = \mathbf{y})}{Q_n(B \mid \mathbf{X} = \mathbf{x})} \leq e^{\epsilon}$$

where $h(\mathbf{x}, \mathbf{y}) = |\{i \in \{1, ..., n\} : x_i \neq y_i\}|$ denotes the Hamming distance between the datasets $\mathbf{x}$ and $\mathbf{y}$ and the ratio is interpreted to be 1 whenever the numerator and the denominator are both 0.

As we focus solely on global privacy mechanisms in this thesis, in the following, we will always refer to ϵ-differential privacy as defined in Definition 2.2, unless stated otherwise. To grasp the significance of the ϵ-differential privacy definition, consider the worst case where an attacker seeks to access confidential information about an individual with a unique set of characteristics within a dataset. Suppose the attacker knows the releasing mechanism and has access to all information in the dataset except for the specific entry corresponding to the targeted individual. In this scenario, assuming the attacker possesses knowledge of the releasing mechanism and access to all information in the dataset except for the specific entry corresponding to the targeted individual, she could generate numerous private datasets by systematically exploring all possible attribute combinations for the target. Subsequently, she could compute the likelihood of encountering each possible dataset. If the release mechanism satisfies ϵ-differential privacy with a small ϵ value, these probabilities would be nearly identical, making it exceedingly difficult to ascertain the actual attributes of the target individual.
The idea expressed in the previous example is confirmed from a statistical point of view by the following theorem, shown by Wasserman and Zhou (2010)

Theorem 2.1. *Suppose that $\mathbf{Z}$ is obtained from a data release mechanism that satisfies ϵ-differential privacy. Moreover, assume that $X_1, ..., X_n \overset{i.i.d.}{\sim} \mathfrak{p}_0$. Any level γ test which is a function of $\mathbf{Z}$, $\mathfrak{p}_0$ and Q_n of $H_0 : X_i = s$ versus $H_1 : X_i = t$ has power bounded above by γe^{ϵ}.*

Thus, if parameter ϵ is small, it is impossible to test whether a certain individual in the dataset has a specific combination of attributes, being the power of such a test approximately equal to its level.
Designing mechanisms that meet ϵ-differential privacy is possible. When adding noise, for example, Dwork et al. (2006b) show that choosing ϵ_i distributed as a Laplace distribution of zero mean and variance $8/\epsilon^2$ and assuming $\mathbb{X}$ bounded preserves ϵ-differential privacy. However, the demand made by ϵ differential privacy is very stringent, leading to synthetic data whose distribution differs notably from private data. Therefore, various relaxations of this definition have been proposed; among them, a notable one is the following, introduced by Dwork et al. (2006a):

Definition 2.3. *Let $\epsilon > 0$ and $\delta \in [0, 1)$, we say that Q_n satisfies (ϵ, δ)-differential privacy if for all $B \in \mathcal{X}^m$ and any $\mathbf{x}, \mathbf{y} \in \mathbb{X}^n$ such that $h(\mathbf{x}, \mathbf{y}) = 1$, we have*

$$Q_n(B \mid \mathbf{X} = \mathbf{x}) \leq e^{\epsilon} Q_n(B \mid \mathbf{Y} = \mathbf{y}) + \delta. \tag{2}$$

In the case $\delta = 0$, we recover the definition of ϵ-differential privacy. Informally, the additional parameter δ controls the probability that the loss of privacy is greater than e^{ϵ}. Under this generalisation, the power of the test in Theorem 2.1 has a bound equal to $\gamma e^{\epsilon} + \delta$, as shown by Dong et al. (2022). The advantage of introducing (ϵ, δ)-differential privacy is that the set of mechanisms that meet this definition is much broader. For example, consider the case of Gaussian noise addition. It does not satisfy ϵ-differential privacy for any $\epsilon > 0$, while Dwork et al. (2006a) show that it satisfies (ϵ, δ)-differential privacy.
In practice, verifying if a mechanism satisfies (ϵ, δ)-differential privacy is complex since we must test the condition (2) for every measurable set. The following definition helps in this respect:

Definition 2.4. *A mechanism satisfies (ϵ, δ) probabilistic differential privacy if, given $q_n(\mathbf{z} \mid \mathbf{x})$ the density of $Q_n(\,\cdot \mid \mathbf{X} = \mathbf{x})$, there exists a $\delta > 0$ such that $Q_n(A_{\mathbf{x},\mathbf{y}} \mid \mathbf{X} = \mathbf{x}) > 1 - \delta$ for all $\mathbf{x}, \mathbf{y} :\ h(\mathbf{x}, \mathbf{y}) = 1$ where*

$$A_{\mathbf{x},\mathbf{y}} := \{\mathbf{z} \in \mathbb{X}^m :\ q_n(\mathbf{z} \mid \mathbf{x}) \leq e^{\epsilon} q_n(\mathbf{z} \mid \mathbf{y})\}.$$

Indeed, the work by Rinott, Yosef et al. (2018) shows the following:

Proposition 2.1. *(ϵ, δ) probabilistic differential privacy implies (ϵ, δ)-differential privacy.*

Proof. For all $B \in \mathcal{X}^m$ and any $\mathbf{x}, \mathbf{y}$ with $h(\mathbf{x}, \mathbf{y}) = 1$, we have

$$Q_n(B \mid \mathbf{X} = \mathbf{x}) = \int_{B \cap A_{\mathbf{x},\mathbf{y}}} q_n(\mathbf{z} \mid \mathbf{x}) d\mathbf{z} + \int_{B \cap A^c_{\mathbf{x},\mathbf{y}}} q_n(\mathbf{z} \mid \mathbf{x}) d\mathbf{z} \leq$$
$$e^{\epsilon} \int_{B \cap A_{\mathbf{x},\mathbf{y}}} q_n(\mathbf{z} \mid \mathbf{y}) d\mathbf{z} + \int_{B \cap A^c_{\mathbf{x},\mathbf{y}}} q_n(\mathbf{z} \mid \mathbf{x}) d\mathbf{z} \leq e^{\epsilon} Q_n(B \mid \mathbf{X} = \mathbf{y}) + \delta.$$

□

An essential difference between ϵ-differential privacy and (ϵ, δ)-differential privacy lies in the composition property. Dwork et al. (2006b) show that releasing $(\mathbf{Z}_1, \mathbf{Z}_2)$, where $\mathbf{Z_1}$ and $\mathbf{Z}_2$ are the outputs of an ϵ_1 and ϵ_2-differentially private mechanism respectively, leads to an overall mechanism that is $\epsilon_1 + \epsilon_2$-differentially private. A similar composition theorem applies to (ϵ, δ)-differential privacy, going to sum both parameters in the composition of multiple mechanisms. The problem is that the sum of numerous δ leads to a test power whose upper bound is much larger since δ in the power bound is not multiplied by the test level. Different generalizations of ϵ-differential privacy have been proposed recently to overcome this problem. Here, we briefly recall two of them for completeness. However, in the following, we will consider only ϵ and (ϵ, δ)-differential privacy since they are the most popular in the literature and, unlike the following generalizations, have been introduced for several years now.

The idea of Rényi differential privacy (Mironov (2017)) is to control the statistical closeness of the release mechanism when conditioned to two datasets with a Hamming distance equal to one. Specifically, a mechanism Q_n satisfies (α, ϵ)-Rényi differential privacy if for every $\mathbf{x}, \mathbf{y} \in \mathbb{X}^n : h(\mathbf{x}, \mathbf{y}) = 1$ it holds that

$$D_\alpha(Q_n(\,\cdot \mid \mathbf{X} = \mathbf{x}) \mid\mid Q_n(\,\cdot \mid \mathbf{Y} = \mathbf{y})) \leq \epsilon$$

where $D_\alpha(P \mid\mid Q)$ denotes the Rényi divergence of order $\alpha > 1$ between two probability measures, i.e.

$$D_\alpha(P \mid\mid Q) = \frac{1}{\alpha - 1} \log \left\{ \int_{\mathbb{X}} \left(\frac{dP}{dQ}(x) \right)^{\alpha - 1} P(dx) \right\}$$

where dP/dQ is the Radon–Nikodym derivative. Mironov (2017) shows that this definition is linked with ϵ-differential privacy and satisfies good compositions properties. To conclude, recently the work by Dong et al. (2022) extended the definition of ϵ-differential privacy starting from the result expressed in the theorem 2.1 and introducing a way to compare data releasing mechanism, called Gaussian differential privacy, based on the power of any statistical test of the type $H_0 : X_i = s$ versus $H_1 : X_i = t$.

2.3. The Bayesian approach

Following a Bayesian approach, we consider the private observations $\mathbf{X}$ as a part of an infinite sequence $(X_i)_{i \geq 1}$ of exchangeable $\mathbb{X}$-valued random variables. Let $\mathbb{S}_n$ be the group of permutations of n objects. We point out that $\mathbb{S}_n$ acts naturally to $\mathbb{N}$ by permuting the first n numbers and leaving the remaining unchanged.

Definition 2.5. *A sequence of random variables $(X_i)_{i \geq 1}$ is said to be exchangeable if for any $n \in \mathbb{N}$ and any permutation $\sigma \in \mathbb{S}_n$,*

$$(X_i)_{i \geq 1} \overset{d}{=} (X_{\sigma(i)})_{i \geq 1},$$

where $\overset{d}{=}$ denotes the equality in distribution.

We denote by $\mathcal{P}(\mathbb{X})$ the space of probability measures on $\mathbb{X}$ and by $\mathfrak{B}(\mathcal{P}(\mathbb{X}))$ its Borel-σ field with respect to the topology of weak convergence. De Finetti's representation theorem (de Finetti (1937)) states that a sequence of random variables is exchangeable if and only if it is a mixture of sequences of i.i.d. random variables.

Theorem 2.2. *Let $\mathbb{X}$ be a Polish space and let $(X_i)_{i \geq 1}$ be an exchangeable sequence. Then there exists a probability measure π on $\mathcal{P}(\mathbb{X})$ such that, for every $n \in \mathbb{N}$ and every $A_1, ..., A_n \in \mathcal{X}$,*

$$\mathbb{P}[X_1 \in A_1, ..., X_n \in A_n] = \int_{\mathcal{P}(\mathbb{X})} \left[\prod_{i=1}^{n} \tilde{p}(A_i) \right] \pi(d\tilde{p}) =: \mu_n(A_1 \times ... \times A_n).$$

The theorem above provides a comprehensive way to specify the law of $\mathbf{X}$ through the probability measure π defined on $(\mathcal{P}(\mathbb{X}), \mathfrak{B}(\mathcal{P}(\mathbb{X})))$, the so-called de Finetti measure. It is common to rephrase the result provided by Theorem 2.2 using a random probability $\tilde{p}$, i.e. a measurable function defined on $(\Omega, \mathcal{A}, \mathbb{P})$ and taking values in $(\mathcal{P}(\mathbb{X}), \mathfrak{B}(\mathcal{P}(\mathbb{X})))$, such that

$$\begin{aligned} X_i \mid \tilde{p} &\overset{\text{i.i.d.}}{\sim} \tilde{p} \\ \tilde{p} &\sim \pi \end{aligned} \tag{3}$$

for $i = 1, ..., n$ and any $n \in \mathbb{N}$. Moreover, De Finetti's Theorem justifies the Bayesian approach by guaranteeing the existence of probability of measure π on $\mathcal{P}(\mathbb{X})$. Indeed, the Bayesian statistician pragmatically chooses a possible distribution which takes the role of π, the so-called prior, and then finds the posterior distribution, defined as a probability kernel $\pi_n : \mathfrak{B}(\mathcal{P}(\mathbb{X})) \times \mathbb{X}^n \to [0, 1]$ satisfying the following disintegration

$$\mathbb{P}[X_1 \in A_1, ..., X_n \in A_n, \tilde{p} \in C] = \int_{A_1 \times \cdots \times A_n} \pi_n(C \mid \mathbf{X} = \mathbf{x}) \; \mu_n(d\mathbf{x}) \tag{4}$$

for all $n \geq 1$, $A_1, ..., A_n \in \mathcal{X}$ and $C \in \mathfrak{B}(\mathcal{P}(\mathbb{X}))$.
The exchangeability assumption also provides a natural way to assign the law to an observation X_{n+1} given the first n random variables, i.e. $\mathbb{P}[X_{n+1} \in \cdot \mid \mathbf{X}]$. In particular, this probability measure is called posterior predictive distribution and is rigorously defined by means of a probability kernel $\alpha_n(\cdot \mid \cdot) : \mathcal{X} \times \mathbb{X}^n \to [0, 1]$ for which

$$\mathbb{P}[\mathbf{X} \in (A_1 \times \cdots \times A_n), X_{n+1} \in A_{n+1}] = \int_{A_1 \times \cdots \times A_n} \alpha_n(A_{n+1} \mid \mathbf{X} = \mathbf{x}) \mu_n(d\mathbf{x})$$

is satisfied for all $A_1, ..., A_n, A_{n+1} \in \mathcal{X}$.

2.3.1. Parametric Bayesian models

When π degenerates on a finite-dimensional subspace of $\mathcal{P}(\mathbb{X})$, the inferential problem is usually called *parametric*. Let $\Theta \subset \mathbb{R}^d$ be the parameter space. If one sets $\mathcal{D}(\mathbb{X}) = \{P_\theta : \theta \in \Theta\}$ and assigns a de Finetti measure such that $\pi(\mathcal{D}(\mathbb{X})) = 1$, there exists a bijection $\mathscr{T} : \Theta \to \mathcal{D}(\mathbb{X})$ and π induces a probability measure ϕ (prior) on Θ by means of $\phi(B) = \pi(\mathscr{T}(B))$ for every $B \in \mathfrak{B}(\Theta)$, where $\mathfrak{B}(\Theta)$ denotes the Borel σ-algebra on Θ. Model (10) can be then rewritten as follows

$$\begin{aligned} X_i \mid \theta &\overset{\text{i.i.d.}}{\sim} P_\theta := \mathscr{T}(\theta) \\ \theta &\sim \phi \end{aligned}$$

If we assume that there exists a σ-finite measure μ on $(\mathbb{X}, \mathcal{X})$ such that $P_\theta \ll \mu$ for all $\theta \in \Theta$, we can derive the posterior distribution by Bayes' theorem for dominated models. Let $f_\theta(x)$ (likelihood) be the density of P_θ with respect to μ; the posterior distribution $\psi_n(B \mid \mathbf{X}) = \pi_n(\mathscr{T}(B) \mid \mathbf{X})$ is the probability measure satisfying

$$\psi_n(B \mid \mathbf{X} = \mathbf{x}) = \frac{\int_B [\prod_{i=1}^n f_\theta(x_i)] \; \phi(d\theta)}{\int_\Theta [\prod_{i=1}^n f_\theta(x_i)] \; \phi(d\theta)}$$

for all $B \in \mathfrak{B}(\Theta)$.

2.3.2. Nonparametric Bayesian models

When the support of π is infinite-dimensional, one used to say *nonparametric* inferential problem. Here, we focus on discrete Bayesian Nonparametric priors, i.e. those π which select discrete random probability measures with probability one. These priors have been successfully used in several frameworks, including prediction in species problems (Lijoi et al. (2007b)), genomics and bioinformatics (Lijoi et al. (2007a)) and topic modelling (Teh (2006)). In the following sections, we briefly describe some examples of Bayesian Nonparametric priors and, in particular, the most famous example, the Dirichlet Process (Ferguson (1973)) and one of its generalisations, namely the Pitman-Yor process (Pitman and Yor (1997)).

2.3.3. Dirichlet Process

A way to specify the distribution of a random probability measure is to see it as a stochastic process with index set $\mathcal{X}$. Indeed, as a consequence of Kolmogorov's Extension Theorem, we can characterise the law of $(\tilde{p}(A))_{A \in \mathcal{X}}$ by providing the distribution it attains in any measurable partition of $\mathbb{X}$ (Ferguson (1973)). We denote by $\text{Dir}(\alpha_1, ..., \alpha_k)$ the Dirichlet distribution with parameters $(\alpha_1, ..., \alpha_k)$ with $\alpha_i > 0$ for all i. For $k > 1$,

let $\mathscr{S}^{k-1}$ denotes the set $\{(x_1, ..., x_{k-1}) \in \mathbb{R}^{k-1} : x_i > 0,\ \sum_{i=1}^{k-1} x_i \leq 1\}$. We recall that a Dirichlet distribution of parameters $(\alpha_1, ..., \alpha_k)$ has the following density with respect to the Lebesegue measure on $\mathbb{R}^{k-1}$:

$$f(p_1, ..., p_{k-1}) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^{k} \Gamma(\alpha_i)}\ p_1^{\alpha_1 - 1} \cdots p_{k-1}^{\alpha_{k-1} - 1} \left(1 - \sum_{i=1}^{k-1} p_i\right)^{\alpha_k - 1} \mathbb{1}_{\mathscr{S}^{k-1}}(p_1, ..., p_{k-1})$$

where $\alpha_0 := \sum_{i=1}^{k} \alpha_i$ and $\Gamma(\cdot)$ denotes the Gamma function.

Definition 2.6. *Let $\theta > 0$ and H a probability measure on $(\mathbb{X}, \mathcal{X})$. The Dirichlet Process (DP) with base measure θH is the unique probability law on $\mathcal{P}(\mathbb{X})$ satisfying*

$$(\tilde{p}(A_1), ..., \tilde{p}(A_n)) \overset{d}{=} Dir(\theta H(A_1), ..., \theta H(A_n))$$

for any $n \in \mathbb{N}$ and measurable partition $A_1, ..., A_n$ of $\mathbb{X}$, i.e. $\cup_{i=1}^{n} A_i = \mathbb{X}$, $A_i \cap A_j = \emptyset$ for $i \neq j$ and $A_i \in \mathcal{X}$ for any $i = 1, ..., n$.

We denote a vector of random variables modelled with a DP prior, often called a sample from a Dirichlet process, by

$$\begin{aligned} X_1, ..., X_n \mid \tilde{p} &\overset{\text{i.i.d.}}{\sim} \tilde{p} \\ \tilde{p} &\sim \mathscr{D}(\theta, H) \end{aligned} \tag{5}$$

The parameter θ is called the intensity or concentration parameter, while the probability distribution H is the mean measure, i.e. given $\tilde{p} \sim \mathscr{D}(\theta, H)$ we have that $H(A) = \mathbb{E}[\tilde{p}(A)]$ for any $A \in \mathcal{X}$.
Here, we recall some fundamental properties of the Dirichlet process prior. Using the above definition, Ferguson (1973) shows that the DP prior features posterior conjugacy in the following sense:

Theorem 2.3. *Let $X_1, ..., X_n \mid \tilde{p} \overset{i.i.d.}{\sim} \tilde{p}$ with $\tilde{p} \sim \mathscr{D}(\theta, H)$. Then, the posterior is still a Dirichlet process with concentration parameter $\theta + n$ and mean measure H^*, where*

$$H^* = \frac{\theta}{\theta + n} H + \frac{1}{\theta + n} \sum_{i=1}^{n} \delta_{X_i} \tag{6}$$

and δ_X denotes the Dirac measure concentrated at X.

Moreover, it is also possible to characterise the posterior predictive distributions of a sequence $(X_i)_{i \geq 1}$, which has as de Finetti measure the DP. The work by Pitman (1996) shows that the law of X_{n+1} given a sample of size n from a DP coincides with the probability measure H^* of Equation (6). These predictive distributions are known in the literature as the generalised Pólya or Blackwell-MacQueen urn scheme. Section 2.3.5 presents a result that intuitively shows the discreteness of the Dirichlet process. We note, however, that the expression of the predictive distribution in equation (6) suggests that a sample from the DP may exhibit ties, i.e. $\mathbb{P}[X_i = X_j] > 0$ for all $i \neq j$.

2.3.4. Exchangeable random partition

All Nonparametric Bayesian models considered in this thesis, including the one in the previous section for diffusive H, are examples of species sampling models. A (proper) species sampling model is an exchangeable sequence of random variables $(X_i)_{i \geq 1}$ directed by a de Finetti measure that is a discrete random probability of the following form

$$\tilde{p} = \sum_{j=1}^{\infty} q_j \delta_{Y_j}, \tag{7}$$

where $(Y_j)_{j \geq 1}$ and $(q_j)_{j \geq 1}$ are two independent families of random variables and Y_j are i.i.d. from a non-atomic probability measure H, i.e. $H(\{x\}) = 0$ for all $x \in \mathbb{X}$.
Let us consider a vector of random variables $X_1, ..., X_n$ conditionally distributed as $\tilde{p}$, with $\tilde{p}$ as in equation (7), we can define a random partition $\mathcal{P}_n$ of the set of integers $[n] := \{1, ..., n\}$ according to the rule: two integers i and j belong to the same partition set if and only if $X_i = X_j$. The obtained $\mathcal{P}_n$ is an exchangeable partition (Pitman (2006)). For $\sigma \in \mathbb{S}_n$, let $\sigma(A)$ be the set $\{\sigma(i) : i \in A\}$ with $A \subseteq [n]$. We denote by $\mathcal{C}_n$ the set of compositions of n, i.e. $\{(n_1, ..., n_k) \in \mathbb{N}^k$ for some integer $k \leq n : n_i > 0,\ \sum_{i=1}^{k} n_i = n\}$.

Definition 2.7. *A random partition $\mathcal{P}_n$ of $[n]$ is called exchangeable if for every partition $\{A_1, ..., A_k\}$ of $[n]$ the probability $\mathbb{P}[\mathcal{P}_n = \{\sigma(A_1), ..., \sigma(A_n)\}]$ is the same for every $\sigma \in \mathbb{S}_n$. Equivalently, a random partition $\mathcal{P}_n$ of $[n]$ is exchangeable if there exists a symmetric function $p^{(n)} : \mathcal{C}_n \to [0, 1]$ such that, for every partition $\{A_1, ..., A_k\}$ of $[n]$,*

$$\mathbb{P}[\mathcal{P}_n = \{A_1, ..., A_k\}] = p^{(n)}(|A_1|, ..., |A_k|).$$

The function $p^{(n)}$ is called the exchangeable partition probability function (EPPF) of $\mathcal{P}_n$.

As remarked by Ghosal and van der Vaart (2017), we define the EPPF as a function of compositions $(n_1, ..., n_k)$, but since it is symmetric, we could have defined it on the corresponding unordered sets $\{n_1, ..., n_k\}$, called partition of n. Its value $p^{(n)}(n_1, ..., n_k)$ represents the probability of a particular partition $\{A_1, ..., A_k\}$ with $|A_i| = n_i$ and its number of arguments varies from 1 to n. We denote by $p_k^{(n)}$ the EPPF when the number of its arguments is k.
One way to define a species sampling model is to specify H and a consistent sequence $\{p_k^{(n)},\ n \in \mathbb{N},\ k \leq n\}$ satisfying the following addition rule:

$$p_k^{(n)}(n_1, ..., n_k) = \sum_{j=1}^{k} p_k^{(n+1)}(n_1, ..., n_j + 1, ..., n_k) + p_{k+1}^{(n+1)}(n_1, ..., n_k, 1). \tag{8}$$

for any $(n_1, ..., n_k) \in \mathcal{C}_n$, $k \in [n]$ and $n \in \mathbb{N}$. The n-th element of the EPPF sequence represents the distribution of the random partitions $\mathcal{P}_n$ induced by $(X_i)_{i \geq 1}$ when considered up to n. Condition (5) expresses in a distributional sense the consistency that these partitions must satisfy, i.e. $\mathcal{P}_n$ is equal a.s. to the partition obtained from $\mathcal{P}_{n+1}$ by leaving out the element $n+1$. We refer to Pitman (2006) for a comprehensive overview of exchangeable partitions.
The EPPF allows one to obtain an analytical expression for the predictive distribution. Let $K_n = k \leq n$ be the number of unique values in $\mathbf{X}$; we will denote by $X_1^*, ..., X_{K_n}^*$ their values and by $\mathbf{N}_n = (N_1, ..., N_{K_n}) = (n_1, ..., n_k)$ their frequencies, satisfying $\sum_{i=1}^{k} n_i = n$. The 1-step posterior predictive distribution can be written as follows:

$$\mathbb{P}[X_{n+1} \in A \mid \mathbf{X}] = \frac{p_{k+1}^{(n+1)}(n_1, ..., n_k, 1)}{p_k^{(n)}(n_1, ..., n_k)} H(A) + \sum_{j=1}^{k} \frac{p_k^{(n+1)}(n_1, ..., n_j + 1, ..., n_k)}{p_k^{(n)}(n_1, ..., n_k)} \delta_{X_j^*}(A).$$

for any $A \in \mathcal{X}$.
Furthermore, it is possible to express the marginal law of $\mathbf{X}$ via the EPPF. Indeed, the law of $(X_1, ..., X_n)$ is equal to the law of n observations obtained by first generating a random partition $\mathcal{P}_n$ according to its EPPF and then attaching to the partition sets in order of appearance the values of an independent i.i.d. vector of random variables $X_1^*, ..., X_{K_n}^*$ from H (Pitman (1996)). This result follows from the fact that the EPPF in a species sampling model of the type (7) can be expressed as

$$p_k^{(n)}(n_1, ..., n_k) = \mathbb{E}\left[\sum_{1 \leq i_1 \neq \cdots \neq i_k < \infty} \prod_{j=1}^{k} q_{i_j}^{n_j} \right]$$

and, by the stochastic independence of $(q_j)_{j \geq 1}$ and $(Y_j)_{j \geq 1}$, the marginal law of the observations can be factorized in the probability of observing a particular random partition and the one of observing a specific set of unique values.

2.3.5. Pitman-Yor process

The Pitman-Yor two-parameters family (Pitman and Yor (1997)) is the species sampling model characterized by a non-atomic distribution H and the following sequence of EPPF:

$$p_k^{(n)}(n_1, ..., n_k) = \frac{\prod_{i=1}^{k-1}(\theta + i\sigma)}{(\theta + 1)_{n-1}} \prod_{i=1}^{k} (1 - \sigma)_{n_i - 1} \tag{9}$$

where $(a)_b := \Gamma(a + b)/\Gamma(a)$ for $a, b \geq 0$ denotes the Pochhammer symbol and the couple of parameters (θ, σ) are such that $\sigma \in [0, 1)$ and $\theta > -\sigma$ or $\sigma < 0$ and $\theta = z|\sigma|$ for some integer $z \geq 1$. In the case of $\sigma = 0$, this prior coincides with the Dirichlet Process prior.
Formally, we denote a vector of random variables modelled with a PY process by

$$\begin{aligned} X_1, ..., X_n \mid \tilde{p} &\overset{\text{i.i.d.}}{\sim} \tilde{p} \\ \tilde{p} &\sim \mathscr{PY}(\sigma, \theta, H) \end{aligned} \tag{10}$$

Equation (9) leads to the following expression of the 1-step posterior predictive distribution:

$$\mathbb{P}[X_{n+1} \in \cdot \mid \mathbf{X}] = \frac{\theta + k\sigma}{\theta + n} H(\cdot) + \frac{1}{\theta + n} \sum_{j=1}^{k} (n_j - \sigma) \delta_{X_j^*}(\cdot). \tag{11}$$

Notice that (11) is a linear combination of the probability $(\theta + k\sigma)/(\theta + n)$ that X_{n+1} corresponds to a "new" (unique) value, i.e., hitherto unseen, and the probability $(n_i - \sigma)/(\theta + n)$ that X_{n+1} is equal to X_i^* (Pitman (2006)). Let us assume $\sigma \in [0, 1)$; this parameter controls the rates at which we observe already observed unique values, and new values arise. Under DP prior, i.e. $\sigma = 0$, the probability of generating a new value does not depend on the number of unique values, and the probability of observing X_i^* becomes proportional to its partition set's size.
The stick-breaking construction (Perman et al. (1992)) gives an equivalent definition of PY process.

Theorem 2.4. *Let $\sigma \in [0, 1)$, $\theta > -\sigma$ and $H \in \mathcal{P}(\mathbb{X})$ non-atomic. Let $V_i \overset{ind}{\sim} Beta(1 - \sigma, \theta + i\sigma)$, $W_i = V_i \prod_{l=1}^{i-1}(1 - V_l)$ and $Y_i \overset{i.i.d.}{\sim} H$ independently from $(V_i)_{i\geq 1}$, then*

$$\sum_{i=1}^{\infty} W_i \delta_{Y_i} \sim \mathscr{PY}(\sigma, \theta, H)$$

The parameter $\sigma \in [0, 1)$ governs the tail behaviour of $\tilde{p} \sim \mathscr{PY}(\sigma, \theta, H)$. In particular, let $(W_{(i)})_{i\geq 1}$ be the decreasingly ordered random sequence $(W_i)_{i\geq 1}$ of Theorem 2.4, then as $i \to \infty$ the $W_{(i)}$'s follow a power-law distribution of exponent $c := \sigma^{-1}$, i.e. $\mathbb{E}[W_{(i)}]$ is asymptotically equivalent to $C_1 i^{-c}$ for some constant $C_1 > 0$, as shown by Pitman and Yor (1997). Hence, the larger σ, the heavier the tail of $\tilde{p}$. For $\sigma = 0$, Pitman (2006) shows that the DP features geometric tails, i.e. $\mathbb{E}[W_{(i)}]$ is asymptotically equivalent to $C_2 e^{-i/\theta}$ for $C_2 > 0$.
Another useful property is the updating rule that characterizes the PY process and is given by the following proposition shown in Pitman (1996):

Proposition 2.2. *Let $(X_1, ..., X_n)$ a sample from a random distribution $\tilde{p}$ with $\mathscr{PY}(\sigma, \theta, H)$, then conditionally given $\mathbf{X}$ featuring k distinct values X_i^* with frequencies n_i for $i = 1, ..., k$,*

$$\tilde{p} \mid \mathbf{X} \overset{d}{=} \sum_{j=1}^{k} \tilde{P}_j \delta_{X_j^*} + \tilde{R}_k F_k$$

where $(\tilde{P}_1, ..., \tilde{P}_k, \tilde{R}_k)$ has $Dir(n_1 - \sigma, ..., n_k - \sigma, \theta + k\sigma)$ distribution, independently of the random distribution F_k, which has $\mathscr{PY}(\sigma, \theta + k\sigma, H)$-distribution.

Lastly, we consider the case of $\sigma < 0$ and $\theta = z|\sigma|$. In this case, $p_k^{(n)} = 0$ for $k > \min\{n, z\}$, which entails that the maximum number of unique values in the partition induced by $\mathbf{X}$ is $\min\{n, z\}$ almost surely. In particular, Gnedin and Pitman (2006) show that the random partition for this choice of the parameters can be obtained by sampling from a Multinomial with z categories and probabilities distributed as a symmetric Dirichlet distribution of parameter $|\sigma|$.

3. Differentially private data release

This section briefly describes some releasing mechanisms in the statistical literature that satisfy the definition of ϵ-differential privacy. In particular, we consider the exponential mechanism (Wasserman and Zhou (2010), Rinott et al. (2018), McSherry and Talwar (2007)), the Bayesian (parametric) mechanism proposed by Dimitrakakis et al. (2017) and various methods based on histogram perturbation (Wasserman and Zhou (2010), Machanavajjhala et al. (2008)).

3.1. Expontential mechanism

A general example of mechanism Q_n which satisfies ϵ-differential privacy is the so-called *exponential mechanism* (McSherry and Talwar, 2007). Consider a loss function $\xi : \mathbb{X}^n \times \mathbb{X}^m \to \mathbb{R}$ and define the following quantity:

$$\Delta_{n,m} = \sup_{\substack{\mathbf{x},\mathbf{y}: \\ h(\mathbf{x},\mathbf{y})=1}} \sup_{\mathbf{z}\in\mathbb{X}^m} |\xi(\mathbf{x}, \mathbf{z}) - \xi(\mathbf{y}, \mathbf{z})|,$$

where we assume $\Delta_{n,m} < \infty$. The exponential mechanism consists of generating $(Z_1, ..., Z_m)$ as sample from the density

$$g(\mathbf{z}|\mathbf{x}) = \frac{\exp\left(-\frac{\epsilon\, \xi(\mathbf{x}, \mathbf{z})}{2\Delta_{n,m}}\right)}{\int_{\mathbb{X}^m} \exp\left(-\frac{\epsilon\, \xi(\mathbf{x}, \mathbf{s})}{2\Delta_{n,m}}\right) d\mathbf{s}} \tag{12}$$

where $\epsilon > 0$. McSherry and Talwar (2007) show that this mechanism satisfies ϵ-differential privacy. As stated by Rinott et al. (2018), it is easy to see that Q_n gives a higher probability to datasets $\mathbf{Z}$, which have higher

utility according to the loss ξ. Rinott et al. (2018) also consider different types of loss functions, such as the Hellinger distance, the l_1 and the l_2 norms. This mechanism is simple and intuitive. However, sampling from the density of the Equation 12 is not simple, nor is choosing the best loss function ξ to obtain datasets that are of general utility and do not just preserve specific statistics of the private data, such as the mean.

3.2. Bayesian synthetic data

A remarkable work proposing Bayesian mechanisms to release synthetic data is that of Dimitrakakis et al. (2017). In this section, and only in this one, we assume that the data released by a mechanism belong to a space Θ different to $\mathbb{X}$ in which the observations take values. Accordingly, we model the release mechanism as a probability kernel $Q_n(\,\cdot \mid \cdot\,) : \mathfrak{B}(\Theta^m) \times \mathbb{X}^n \to [0,1]$, where, for $k \geq 1$, $\mathfrak{B}(\Theta^k)$ denotes the Borel-σ field of Θ^k. To present the result of Dimitrakakis et al. (2017), we introduce a generalization of (ϵ, δ)-differential privacy, known as (ϵ, δ)-differential privacy with respect to the metric ϱ (Chatzikokolakis et al. (2013)). This definition relies on a pseudo-metric $\varrho : \mathbb{X}^n \times \mathbb{X}^n \to \mathbb{R}$ which substitutes the Hamming distance to quantify the concept of closeness between two datasets in a more general way than the simple difference in one-data records.

Definition 3.1. *Let $\epsilon > 0$. We say that Q_n satisfies (ϵ, δ)-differential privacy with respect to the pseduo-metric ϱ if, for all $B \in \mathfrak{B}(\Theta^m)$ and for any $\mathbf{x}, \mathbf{y} \in \mathbb{X}^n$,*

$$Q_n(B \mid \mathbf{X} = \mathbf{x}) \leq e^{\epsilon \varrho(\mathbf{x},\mathbf{y})} Q_n(B \mid \mathbf{X} = \mathbf{y}) + \delta \varrho(\mathbf{x}, \mathbf{y})$$

We assume the private data $\mathbf{X}$ modelled using a general Bayesian parametric model, as detailed in Section 2.3.1. The main theorem in Dimitrakakis et al. (2017) states that releasing samples from the posterior is differentially private under suitable assumptions on the prior and the likelihood. In particular let $g_\theta(\mathbf{x}) := \sum_{i=1}^n \log f_\theta(x_i)$ be the log-likelihood, they define

$$k(\theta) := \inf\{u \in \mathbb{R} : |g_\theta(\mathbf{x}) - g_\theta(\mathbf{y})| \leq u \ \varrho(\mathbf{x}, \mathbf{y}),\ \forall \mathbf{x}, \mathbf{y} \in \mathbb{X}^n\}.$$

Next, they consider two assumptions: the first assumption (Lipschitz continuity) requires that there exists $L < \infty$ such that $k(\theta) \leq L$, for all $\theta \in \Theta$, while the second one less restrictive (Stochastic Lipschitz continuity) requires that there are some constants $c, L_0 > 0$ such that, for all $L \geq L_0$:

$$\phi(\{\theta \in \Theta : k(\theta) \leq L\}) \geq 1 - e^{-c(L - L_0)}$$

where ϕ is the prior as defined in Section 2.3.1. Next, they show that under the first assumption, the posterior ψ_n is $(2L, 0)$-differentially private under pseudo metric ϱ, i.e

$$\psi_n(B \mid \mathbf{X} = \mathbf{x}) \leq e^{2L\varrho(\mathbf{x},\mathbf{y})} \psi_n(B \mid \mathbf{X} = \mathbf{y})$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{X}^n$ and $B \in \mathfrak{B}(\Theta)$. Moreover, under the requirement of Stochastic Lipschitz continuity, they show that the posterior ξ is $(0, M)$-differentially private under pseudo metric $\sqrt{\varrho}$ with M suitable constant. They also provide a series of simple Bayesian models in which it is possible to compute the constants of the described results. However, we notice that they only consider releasing a set of samples from the posterior and not using it for generating new data relying on the predictive distributions.

3.3. Sampling from a Histogram

This section describes three concrete mechanisms that satisfy ϵ-differential privacy. They consist of generating synthetic data by sampling from a perturbed histogram. This approach is related to the proposed Bayesian Nonparametric (BNP) mechanism, and it allows us to introduce the work by Machanavajjhala et al. (2008) that our method extends.

We consider $\mathbb{X} = [0,1]^r$ for some integer $r \geq 1$. As stated by Wasserman and Zhou (2010), extensions to more general spaces are possible, but we focus on this space to avoid unnecessary technicalities. We partition $\mathbb{X}$ into k bins $B_1, ..., B_k$ where each bin B_j is a cube with sides of length $h \in (0,1)$ such that $k = 1/h^r$ is an integer. Let $\mathbb{1}_A(x)$ the indicator function of the set A, we denote by $\hat{f}_k(x)$ the histogram estimator for the private data $\mathbf{X}$, defined as follows:

$$\hat{f}_k(x) := \sum_{j=1}^{k} \frac{\hat{p}_j}{h^r} \mathbb{1}_{B_j}(x),$$

where $\hat{p}_j = C_j/n$ and $C_j = \sum_{i=1}^n \mathbb{1}_{B_j}(X_i)$.

The first method we present consists of smoothing the histogram estimator as follows:

$$\hat{f}_{k,\sigma}(x) = (1 - \sigma)\hat{f}_k(x) + \sigma,$$

where $\sigma \in (0,1)$. In particular, we have that sampling from the smoothed histogram satisfies differential privacy as stated by the following result shown in Wasserman and Zhou (2010)

Theorem 3.1. *Let* $\mathbf{Z} = (Z_1, ..., Z_m)$ *where* $Z_1, ..., Z_m \mid \mathbf{X} \overset{i.i.d.}{\sim} \hat{f}_{k,\sigma}(x)$. *If*

$$m \log\left(\frac{(1-\sigma)k}{n\sigma} + 1\right) \leq \epsilon, \tag{13}$$

then ϵ*-differential privacy holds.*

Equation (13) shows an interesting relation between k,m, σ and ϵ. In particular, one can notice that $\sigma = 0$ has no privacy guarantee.

As a second example, we perturb the set of counts by adding noise, and then we build a new histogram estimator based on the perturbed counts and release data sampling from it. In particular, let $D_j = C_j + \nu_j$ where $\nu_1, ..., \nu_m$ are independent, identically distributed draws from a Laplace distribution with 0 mean and variance $8/\epsilon^2$. Since adding noise could lead to negative counts we compute $\tilde{D}_j = \max\{D_j, 0\}$ and $\hat{q}_j = \tilde{D}_j / \sum_{s=1}^{k} \tilde{D}_s$ for $j = 1, ..., k$ and we define

$$\tilde{f}(x) := \sum_{j=1}^{k} \frac{\hat{q}_j}{h^r} \mathbb{1}_{B_j}(x).$$

The work by Wasserman and Zhou (2010) shows that any sample $\mathbf{Z} = (Z_1, .., Z_m)$ from $\tilde{f}(x)$ preserve ϵ-differential privacy for any k.

For the last example, we describe the approach proposed in the work by Machanavajjhala et al. (2008). The authors introduce a new mechanism based on a Dirichlet-Multinomial model, and they apply it to data from a mapping program that shows the commuting patterns of the United States population. In particular, their dataset consists of a table, where each row represents a worker having two attributes: the *origin block*, that is, the census block in which the worker lives, and the *destination block*, i.e., place of employment. Since they consider the *origin block* the sensitive information, for each *destination block* h, they view the data as a histogram, with bin B_j corresponding to the combination "worker lives in *origin block* j and goes to h" for $j = 1, ..., k$. Let $\mathbf{p} := (p_1, ..., p_k)$ be the vector of probabilities $p_j := \mathbb{P}[X_i \in B_j]$ for $j = 1, ..., k$. They model bin counts $(C_1, ..., C_k)$ as a Multinomial$(n, \mathbf{p})$, i.e. the discrete distribution given by

$$\mathbb{P}[(C_1, ..., C_k) = (c_1, ..., c_k) \mid \mathbf{p}\ ] = \frac{n!}{c_1! \cdots c_k!}\ p_1^{c_1} \cdots p_k^{c_k}\ \mathbb{1}_E(c_1, ..., c_k)$$

where $E := \{(c_1, ..., c_k) \in \mathbb{N}^k : c_i \geq 0,\ \sum_{i=1}^{k} c_i = n\}$. They take a Bayesian approach and consider as prior for the vector $\mathbf{p}$ the Dirichlet distribution with parameters $(\alpha_1, ..., \alpha_k)$. Then, they draw a vector of probabilities $\mathbf{q} := (q_1, ..., q_k)$ from the posterior of the model, which is still a Dirichlet with updated parameters $(\alpha_1 + C_1, ..., \alpha_k + C_k)$. Finally, they generate a synthetic vector of bin counts $\mathbf{D} := (D_1, ..., D_k)$ sampling from a Multinomial$(m, \mathbf{q})$. Thus, the distribution of $\mathbf{D}$ given the private data $\mathbf{X}$ is the following

$$\mathbb{P}[\mathbf{D} = (d_1, ..., d_k) \mid \mathbf{X}] = \frac{\Gamma(\alpha_0 + n)\Gamma(m+1)}{\Gamma(m + n + \alpha_0)} \prod_{i=1}^{k} \frac{\Gamma(d_i + \alpha_i + C_i)}{\Gamma(\alpha_i + C_i)\Gamma(d_i + 1)} \tag{14}$$

They show that the described mechanism satisfies ϵ-differential privacy under the condition $\alpha_j + C_j \geq m/(e^{\epsilon} - 1)$ for all $j = 1, ..., k$.

4. Our approach

In this section, we describe the mechanism we have introduced. In addition to an exploratory analysis to understand what levels of privacy are guaranteed by our approach, we also report the main results in which we show that the proposed mechanism satisfies differential privacy.

4.1. BNP mechanism

As a first releasing mechanism, we model the private data $\mathbf{X}$ as a sample from a $\mathscr{PY}(\sigma, \theta, H)$ and we generate synthetic observations $\mathbf{Z}$ distributed as the m-step posterior predictive distribution, i.e. the probability measure that satisfies

$$Q_n(A_1 \times \cdots \times A_m \mid \mathbf{X} = \mathbf{x}) = \int_{\mathcal{P}(\mathbb{X})} \left[\prod_{i=1}^{m} \tilde{p}(A_i)\right]\ \pi_n(d\tilde{p} \mid \mathbf{x})$$

for any $A_1, ..., A_m \in \mathcal{X}$, where π_n is the posterior of the Bayesian model.
The algorithm for generating the dataset $\mathbf{Z}$ is the following and relies on the 1-step posterior predictive distribution of Equation (11).

Algorithm 1 BNP mechanism

Given a discrete dataset $\mathbf{X}$ and the number m of data to be generated;
compute the number k of unique values in $\mathbf{X}$;
for $i = 1, ..., k$ **do**
  compute the frequency n_i of the i-th unique value X_i^*;
end for
for $j = 1, ..., m$ **do**
  sample c from a categorical taking values in $\{1, ..., k+1\}$ with vector of probabilities:

$$\left(\frac{n_1 - \sigma}{\theta + n}, ..., \frac{n_k - \sigma}{\theta + n}, \frac{\theta + \sigma k}{\theta + n} \right)$$

  if $c = k + 1$ **then**
    sample Z_j from H;
    update the vector of frequencies to $(n_1, ..., n_k, 1)$, the number of unique values to $k = k + 1$ and add the new unique value $X_k^* = Z_j$;
  else
    set $Z_j = X_c^*$;
    update $n_c = n_c + 1$;
  end if
end for
return $(Z_1, ..., Z_m)$.

In the proof of the Theorem 4.1, we explain how to derive the density $q_n(\,\cdot \mid \mathbf{x})$ of the BNP mechanism. Using the density expression, we perform an exploratory analysis to understand whether our mechanism satisfies some notion of differential privacy. In particular, we run simulations as specified in the Algorithm 2. The idea of this algorithm is to test the definition of (ϵ, δ)-differential privacy by monitoring for several simulated private data $\mathbf{x}^{(k)}$ with $k = 1, ..., K$, the ratio between the density based on $\mathbf{x}^{(k)}$ and that based on a set of datasets $\mathbf{y}^{(j,k)}$ at Hamming distance one from the private data. We evaluate the ratio of the densities in a set of synthetic datasets generated by the mechanism, $\mathbf{z}^{(i,k)}$ with $i = 1, ..., L$. In this way, we check if the ratio is controlled by e^ϵ for small values of ϵ only on the data that our mechanism is most likely to generate, in the spirit of the definition of (ϵ, δ) probabilistic differential privacy.

Algorithm 2 Check Differential Privacy

Fix n, m, θ, σ and a mean distribution H;
for $k = 1, ..., K$ **do**
  sample $\mathbf{x}^{(k)}$ from a $\mathscr{PY}(\theta, \sigma, H)$;
  for $i = 1, ..., L$ **do**
    generate synthetic data $\mathbf{z}^{(i,k)}$ sampling from $Q_n(\,\cdot \mid \mathbf{X} = \mathbf{x}^{(k)})$;
    for $j = 1, ..., J$ **do**
      consider $\mathbf{y}^{(j,k)} :\ h(\mathbf{x}^{(k)}, \mathbf{y}^{(j,k)}) = 1$;
      add the following to the *density-ratio* list:

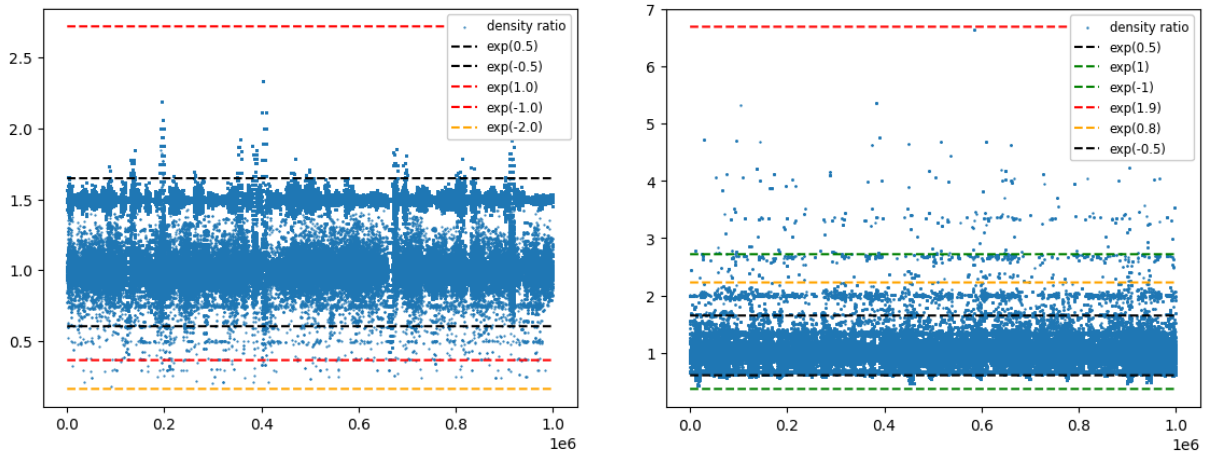
$$\frac{q_n(\mathbf{z}^{(i,k)} \mid \mathbf{x}^{(k)})}{q_n(\mathbf{z}^{(i,k)} \mid \mathbf{y}^{(j,k)})}$$

    end for
  end for
end for
return *density-ratio* list.

In the simulations, we consider the uniform distribution in the interval $[0, 1]$ as the base measure of the PY process. We choose $L = J = K = 100$. Furthermore, to obtain the datasets $\mathbf{y}^{(j,k)}$, we sample t from a discrete uniform distribution with values in $\{0, 1, 2\}$ and we take $\mathbf{y}^{(j,k)}$ equal to $\mathbf{x}^{(k)}$ except for the last observation, which we choose in the following way:

- if $t = 0$, we take randomly another value in $\mathbf{x}^{(k)}$ different to the one we are replacing;
- if $t = 1$, we take randomly a value in $\mathbf{z}^{(i,k)}$ different to the one we are replacing;
- if $t = 2$, we sample a new value from the mean measure of the process.

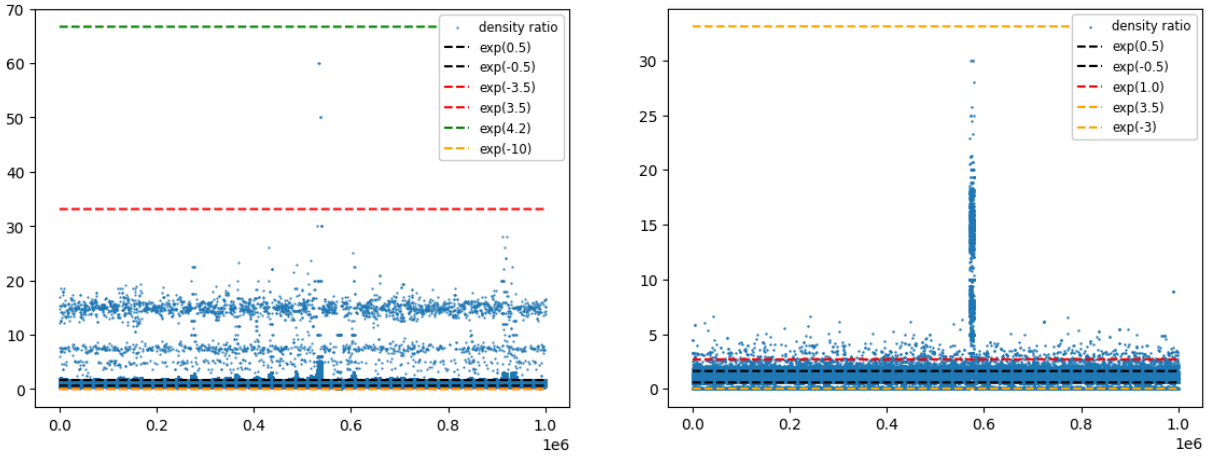
As in the definition of (ϵ, δ)-differential privacy the role of the two datasets $\mathbf{x}, \mathbf{y}$ such that $h(\mathbf{x}, \mathbf{y}) = 1$ is symmetrical, for completeness, we report both the plot when varying the number of iterations of the density ratio as defined in the Algorithm 2, and its inverse.
Figure 1 shows the results obtained with $\theta = 1$ and $\sigma = 0$ (DP prior). We generate 1000 simulated data, and we release $m = 500$ synthetic data. We can see that a choice of $\epsilon = 1.9$ allows us to always control the density ratio and its inverse.



(a) Density ratio (b) Inverse density ratio

Figure 1: Density ratio plot for a DP process with $\theta = 1$: blue dots represent the density ratio or its inverse, while the dashed lines represent the value of e^{ϵ} for different values of ϵ.

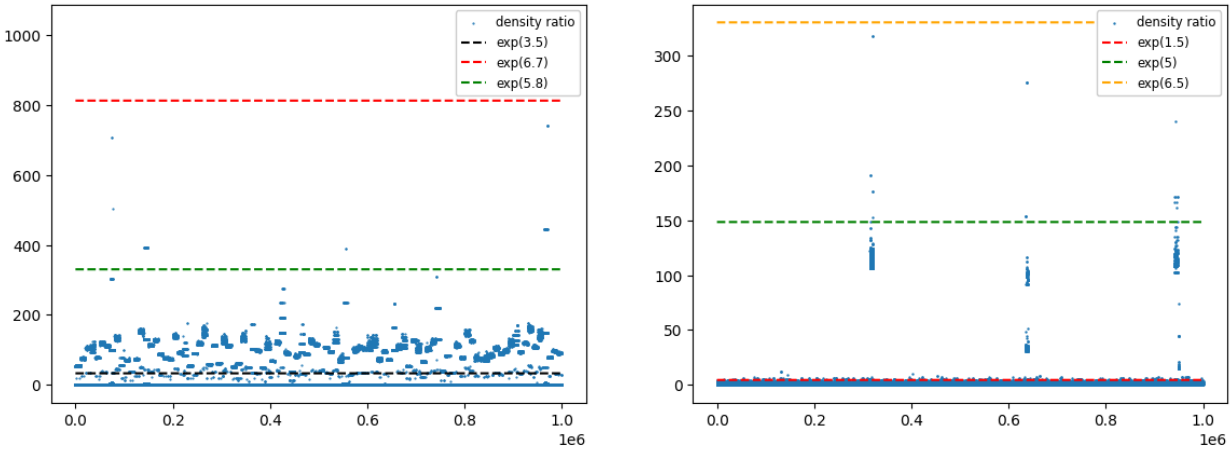
Figure 2 shows the results obtained with $\theta = 10$ and $\sigma = 0$ (DP prior). We generate 1000 simulated data, and we release $m = 500$ synthetic data. As can be seen by increasing the value of θ, the values of ϵ required to bound the density ratio are higher than 4. However, there are few iterations in which $\epsilon > 4$ is required, so the control can occur with a lower value of ϵ with a probability greater than $1 - \delta$ for δ small. We can say that, except for an event of low probability, it is sufficient to take ϵ around 3.



(a) Density ratio (b) Inverse density ratio

Figure 2: Density ratio plot for a DP process with $\theta = 10$: blue dots represent the density ratio or its inverse, while the dashed lines represent the value of e^{ϵ} for different values of ϵ.

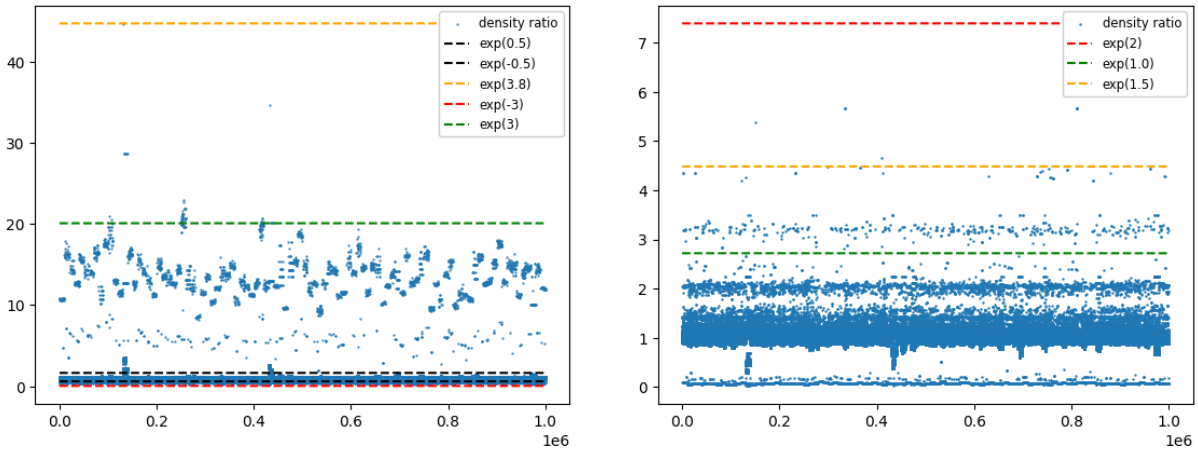
Figure 3 shows the results obtained with $\theta = 2$ and $\sigma = 1/2$. We generate 1000 simulated data, and we release $m = 100$ synthetic data. Although the number of data released is less than in the two previous examples, the values of ϵ required to check densities and its inverse are greater than 6.



(a) Density ratio (b) Inverse density ratio

Figure 3: Density ratio plot for a PY process with $\theta = 2$ and $\sigma = 1/2$: blue dots represent the density ratio or its inverse, while the dashed lines represent the value of e^{ϵ} for different values of ϵ.

Figure 4 shows the results obtained with $\theta = 3$ and $\sigma = 1/5$. We generate 1000 simulated data, and we release $m = 100$ synthetic data. Given the same number of released data in the previous case, we can see that smaller σ values lead to a lower required ϵ.



(a) Density ratio (b) Inverse density ratio

Figure 4: Density ratio plot for a PY process with $\theta = 3$ and $\sigma = 1/5$: blue dots represent the density ratio or its inverse, while the dashed lines represent the value of e^{ϵ} for different values of ϵ.

From this exploratory analysis, we conclude that by moving from $\sigma = 0$ to $\sigma \in (0, 1)$, the guaranteed privacy levels in terms of the parameters ϵ and δ degrade. Furthermore, as θ decreases, smaller values of ϵ can be taken to have δ near 0.

4.2. Privacy preserving theorems

By exchangeability assumption, the data $\mathbf{Z}$ released by the BNP mechanism are the next m observations of the sequence $(X_i)_{i\geq 1}$ to which the private data belongs. Let $X_1^*, ..., X_{K_n}^*$ be the K_n unique values detected in $\mathbf{X}$

and $\mathbf{N}_n := (N_1, ..., N_{K_n})$ the vector of their frequencies. We define

$$L_m^{(n)} := \sum_{i=1}^{m} \prod_{j=1}^{K_n} \mathbb{1}_{\{X_j^*\}^c}(Z_i) = |\{Z_1, ..., Z_m\} \cap \{X_1^*, ..., X_{K_n}^*\}^c|$$

as the number of observations in the synthetic data that do not coincide with any of the K_n unique values in $\mathbf{X}$. We denote by $K_m^{(n)}$ the number of additional unique values observed in $\mathbf{Z}$, i.e. $K_m^{(n)} = K_{n+m} - K_n$ with K_{n+m} the total number of unique values in $(\mathbf{X}, \mathbf{Z})$. At the same time, $X^*_{K_n+1}, ..., X^*_{K_n+K_m^{(n)}}$ denote the $K_m^{(n)}$ additional unique values. Moreover, we define the following random variables

$$S_{K_n+i} := \sum_{j=1}^{m} \mathbb{1}_{\{X^*_{K_n+i}\}}(Z_j), \qquad S_q := \sum_{j=1}^{m} \mathbb{1}_{\{X_q^*\}}(Z_j) \tag{15}$$

for $i = 1, ..., K_m^{(n)}$ and $q = 1, ..., K_n$, where $\sum_{i=1}^{K_m^{(n)}} S_{K_n+i} = L_m^{(n)}$ and $\sum_{q=1}^{K_n} S_q = m - L_m^{(n)}$.
We recall that conditioning on $\mathbf{X}$ is equivalent to conditioning on

$$\{K_n = j\} \cap \{\mathbf{N}_n = \mathbf{n} := (n_1, ..., n_j)\}$$

when considering a discrete nonparametric prior.
Showing that our mechanism satisfies Definition 2.4 is a challenging task. Indeed, note that the definition of (ϵ, δ)-privacy requires a uniform control over all possible datasets $\mathbf{x}$. In our context, this leads to pathological behaviours, e.g., datasets where $K_n = n$ or, if we assume $n_i > 1$ for each i, datasets where n_i is always equal to two. Therefore, we focus on a slightly different question: checking that (ϵ, δ)-differential privacy holds for the specific dataset under analysis. This is a strictly weaker requirement than Definition 2.4, but it still lends itself to an intuitive interpretation in terms of testing. Indeed, suppose that, given $\mathbf{x} \in \mathbb{X}^n$, for any $\mathbf{y} \in \mathbb{X}^n$ such that $h(\mathbf{x}, \mathbf{y}) = 1$

$$Q_n(B \mid \mathbf{X} = \mathbf{x}) \leq e^{\epsilon} Q_n(B \mid \mathbf{Y} = \mathbf{y}) + \delta.$$

Consider an attacker who has explicit knowledge of all the dataset $\mathbf{X} = \mathbf{x}$ except for the value of X_i, then the power of the test $H_0 : X_i = s$ vs $H_1 : X_i = t$ is bounded by $e^{\varepsilon}\gamma + \delta$. The key difference concerning Definition 2.4 is that now the conditions that ϵ and δ must satisfy can depend on the specific instance of $\mathbf{x}$. We observe that the Dirichlet-Multinomial mechanism of Machanavajjhala et al. (2008) as reported by Wasserman and Zhou (2010) satisfies this weaker version of differential privacy since it requires $\alpha_j + C_j \geq m/(e^{\varepsilon} + 1)$ (see below (14)). In the following, we will refer to this weaker version of privacy as "instance-level (ϵ, δ)-differential privacy" or "instance-level ϵ-differential privacy" if $\delta = 0$, and we establish sufficient conditions to ensure that our mechanism satisfies it. We leave it to future research to establish whether the stronger notion of Definition 2.4 is also satisfied by our mechanism.

Proposition 4.1. *Consider data $\mathbf{x} = (x_1, \ldots, x_n)$ displaying $K_n = j$ unique values each appearing n_i, $i = 1, \ldots, j$ times. Suppose $n_i > 1$ for all i. Releasing data sampling from the m-step posterior predictive distribution of a $\mathscr{PY}(\theta, \sigma, H)$, where $\theta = z|\sigma|$ with $z = K_n$ and $\sigma < 0$ satisfies instance-level ϵ-differential privacy if*

$$|\sigma| + n_i - 1 > m/(e^{\epsilon} - 1)$$

holds for all $i = 1, ..., j$.

By choosing the parameters as in the Proposition 4.1, our model coincides in practice with a Dirichlet-Multinomial. Indeed, we find a condition similar to that required by Machanavajjhala et al. (2008) to satisfy differential privacy.
The next theorem considers the case of a PY for $\sigma \in [0, 1)$.

Theorem 4.1. *Consider data $\mathbf{x} = (x_1, \ldots, x_n)$ displaying $K_n = j$ unique values each appearing n_i, $i = 1, \ldots, j$ times. Suppose $n_i > 1$ for all i and that there exists $C_1, C_2 > 0$ such that the mean measure H has a density g with $C_1 \leq g(x) \leq C_2$ for all $x \in \mathbb{X}$. Moreover, let us fix $\epsilon > 0$ and define $\tilde{\delta}(\epsilon)$ as follows*

$$\begin{aligned}
\tilde{\delta}(\epsilon) = 1 - \mathbb{P}\Bigg[& \bigcap_{h=1}^{K_m^{(n)}} \left\{ \frac{S_{j+h} - \sigma}{(\theta + j\sigma)\ C_1} \max_{i=1,\ldots,j} \left(\frac{n_i - 1 - \sigma}{n_i + S_i - 1 - \sigma} \right) \leq e^{\epsilon} \right\}, \\
& \bigcap_{h=1}^{K_m^{(n)}} \left\{ \frac{S_{j+h} - \sigma}{(\theta + j\sigma)\ C_2} \min_{i=1,\ldots,j} \left(\frac{n_i - 1 - \sigma}{n_i + S_i - 1 - \sigma} \right) \geq e^{-\epsilon} \right\}, \\
& \bigcap_{i=1}^{j} \left\{ \frac{\theta + (j + K_m^{(n)})\sigma}{\theta + j\sigma} \frac{n_i - 1 - \sigma}{n_i + S_i - 1 - \sigma} \in [e^{\epsilon}, e^{-\epsilon}] \right\}, \\
& \bigcap_{i=1}^{j} \{S_i \leq (e^{\epsilon} - 1)(n_i - 1 - \sigma)\} \mid K_n = j, \mathbf{N}_n = \mathbf{n} \Bigg].
\end{aligned} \tag{16}$$

Releasing data sampling from the m-step posterior predictive distribution of a $\mathscr{PY}(\theta, \sigma, H)$ with $\sigma \geq 0$ satisfies instance-level (ϵ, δ)-differential privacy for every δ such that $\delta > \tilde{\delta}(\epsilon)$.

Proof. Let $\mathbf{n}$ be a partition of n, $\mathbf{s}_{m-L_m^{(n)}}$ a weak composition of $m - L_m^{(n)}$ composed by $j \leq n$ parts, i.e.

$$\mathbf{s}_{m-L_m^{(n)}} \in \{(s_1, ..., s_j) \in \mathbb{N}^j : \ s_i \geq 0, \ \sum_{i=1}^{j} s_i = m - L_m^{(n)}\}$$

and $\mathbf{s}_{L_m^{(n)}}$ a composition of $L_m^{(n)}$ composed by $k \leq m$ parts, which means

$$\mathbf{s}_{L_m^{(n)}} \in \{(s_1, ..., s_k) \in \mathbb{N}^k : \ s_i > 0, \ \sum_{i=1}^{k} s_i = L_m^{(n)}\}$$

We will denote by $(X_{n+1}, ..., X_{n+m})$ the observations in $\mathbf{Z}$ to be consistent with the usual notation in species sampling problems. Following the approach developed in Lijoi et al. (2007b) and Favaro et al. (2013), we write the density of the m-step posterior predictive distribution as the ratio of the density of the full sample $(\mathbf{Z}, \mathbf{X})$ and the density of $\mathbf{X}$. We remark that the marginal law factorizes in the EPPF and in the distribution of the unique values as detailed in Section 2.3.4. As a consequence, the density of m-step posterior predictive distribution can be written as

$$\begin{aligned}
&\mathbb{P}[X_{n+1} \in dx_{n+1}, ..., X_{n+m} \in dx_{n+m} \mid \mathbf{X} = \mathbf{x}] \\
&\quad = \mathbb{P}[\mathbf{N}_n + \mathbf{S}_{m-L_m^{(n)}} = \mathbf{n} + \mathbf{s}_{m-L_m^{(n)}}, L_m^{(n)} = s, K_m^{(n)} = k, \mathbf{S}_{L_m^{(n)}} = \mathbf{s}_{L_m^{(n)}}, \\
&\qquad\qquad X_{j+1}^* \in dx_{j+1}^*, ..., X_{j+k}^* \in dx_{j+k}^* \mid \mathbf{X}] \\
&\quad = \mathbb{P}[\mathbf{N}_n + \mathbf{S}_{m-L_m^{(n)}} = \mathbf{n} + \mathbf{s}_{m-L_m^{(n)}}, L_m^{(n)} = s, K_m^{(n)} = k, \\
&\qquad\qquad \mathbf{S}_{L_m^{(n)}} = \mathbf{s}_{L_m^{(n)}} \mid K_n = j, \mathbf{N}_n = \mathbf{n}] \prod_{i=j+1}^{j+k} H(dx_i^*)
\end{aligned}$$

where $\mathbf{S}_{L_m^{(n)}} := (S_{K_n+1}, ..., S_{K_n+K_m^{(n)}})$, $\mathbf{S}_{m-L_m^{(n)}} := (S_1, ..., S_{K_n})$ and $X_1^*, ..., X_{j+k}^*$ are the unique values in $X_1, ..., X_{n+m}$. Using the EPPF of the PY process, we obtain

$$\begin{aligned}
&\mathbb{P}[\mathbf{N}_n + \mathbf{S}_{m-L_m^{(n)}} = \mathbf{n} + \mathbf{s}_{m-L_m^{(n)}}, L_m^{(n)} = s, K_m^{(n)} = k, \mathbf{S}_{L_m^{(n)}} = \mathbf{s}_{L_m^{(n)}} \mid K_n = j, \mathbf{N}_n = \mathbf{n}] \prod_{i=j+1}^{j+k} H(dx_i^*) \\
&\quad = \frac{p_{j+k}^{(n+m)}(n_1 + s_1, ..., n_j + s_j, s_{j+1}, ..., s_{j+k})}{p_j^{(n)}(n_1, ..., n_j)} \prod_{i=j+1}^{j+k} H(dx_i^*) \\
&\quad = \frac{1}{(\theta + n)_m} \prod_{i=j}^{j+k-1} (\theta + i\sigma) \prod_{i=1}^{j} (n_i - \sigma)_{s_i} \prod_{r=1}^{k} (1 - \sigma)_{s_{j+r}-1} \prod_{i=j+1}^{j+k} H(dx_i^*)
\end{aligned}$$

Now we consider a dataset $\tilde{\mathbf{x}}$ such that $h(\tilde{\mathbf{x}}, \mathbf{x}) = 1$. By exchangeability, we can assume without loss of generality that the first observation differs in the two samples, i.e. $\tilde{x}_1 \neq x_1$. We assume that $n_i > 1$ for all $i = 1, ..., j$. We will denote by x_l^* the unique value equal to the first observation x_1 while $\tilde{j}, \tilde{\mathbf{n}}, \tilde{k}$ are the information encoded in $(\tilde{\mathbf{x}}, \mathbf{Z})$. We have the following possibilities:

1. $\tilde{x}_1 = x_{j+h}^*$ for some $h = 1, ..., k$. In this case $\tilde{j} = j + 1$, $\tilde{k} = k - 1$, $\tilde{n}_l = n_l - 1$, $\tilde{n}_{j+1} = 1$, $\tilde{s}_{j+1} = s_{j+h}$;
2. $\tilde{x}_1 = x_f^*$ with $x_f^* \neq x_i^*$ for all $i = 1, ..., j + k$. In this case $\tilde{j} = j + 1$, $\tilde{n}_l = n_l - 1$ and $\tilde{n}_f = 1$;
3. $\tilde{x}_1 = x_t^*$ for some $t = 1, ..., j$ and $t \neq l$. In this case $\tilde{n}_t = n_t + 1$, $\tilde{n}_l = n_l - 1$ and $\tilde{j} = j$;

From these considerations, it follows that in the case 1. we want :

$$\begin{aligned}
&\frac{\mathbb{P}[X_{n+1} \in dx_{n+1}, ..., X_{n+m} \in dx_{n+m} \mid \mathbf{X} = \tilde{\mathbf{x}}]}{\mathbb{P}[X_{n+1} \in dx_{n+1}, ..., X_{n+m} \in dx_{n+m} \mid \mathbf{X} = \mathbf{x}]} \\
&\qquad = \frac{1}{(\theta + j\sigma)} \frac{1}{H(dx_{j+h}^*)} \frac{n_l - 1 - \sigma}{n_l + s_l - 1 - \sigma} \ (s_{j+h} - \sigma) \in [e^{-\epsilon}, e^{\epsilon}]
\end{aligned}$$

In the case 2. we want:

$$\begin{aligned}
&\frac{\mathbb{P}[X_{n+1} \in dx_{n+1}, ..., X_{n+m} \in dx_{n+m} \mid \mathbf{X} = \tilde{\mathbf{x}}]}{\mathbb{P}[X_{n+1} \in dx_{n+1}, ..., X_{n+m} \in dx_{n+m} \mid \mathbf{X} = \mathbf{x}]} \\
&\qquad = \frac{\theta + (j + k)\sigma}{\theta + j\sigma} \frac{n_l - 1 - \sigma}{n_l + s_l - 1 - \sigma} \in [e^{-\epsilon}, e^{\epsilon}]
\end{aligned}$$

Finally, in the case 3., we require:

$$\frac{\mathbb{P}[X_{n+1} \in dx_{n+1}, ..., X_{n+m} \in dx_{n+m} \mid \mathbf{X} = \tilde{\mathbf{x}}]}{\mathbb{P}[X_{n+1} \in dx_{n+1}, ..., X_{n+m} \in dx_{n+m} \mid \mathbf{X} = \mathbf{x}]} = \frac{n_t + s_t - \sigma}{n_t - \sigma} \frac{n_l - 1 - \sigma}{n_l + s_l - 1 - \sigma} \in [e^{-\epsilon}, e^{\epsilon}]$$

To satisfy instance-level (ϵ, δ)-differential privacy, we choose δ such that, under $\mathbb{P}[\ \cdot\ |\mathbf{X} = \mathbf{x}]$, the probability assigned to the intersection of the following events is greater than $1 - \delta$:

1. $\frac{1}{(\theta + j\sigma)} \frac{1}{H(dx^*_{j+h})} \frac{n_l - 1 - \sigma}{n_l + S_l - 1 - \sigma} (S_{j+h} - \sigma) \leq e^{\epsilon}\ \forall l, h \iff \frac{S_{j+h} - \sigma}{(\theta + j\sigma)\ C_1} \max_{i=1,...,j} \left(\frac{n_i - 1 - \sigma}{n_i + S_i - 1 - \sigma} \right) \leq e^{\epsilon}$ for all $h = 1, ..., K^{(n)}_m$ under the assumption that H has density $g \geq C_1$;
2. $\frac{1}{(\theta + j\sigma)} \frac{1}{H(dx^*_{j+h})} \frac{n_l - 1 - \sigma}{n_l + S_l - 1 - \sigma} (S_{j+h} - \sigma) \geq e^{-\epsilon}\ \forall l, h \iff \frac{S_{j+h} - \sigma}{(\theta + j\sigma)\ C_2} \min_{i=1,...,j} \left(\frac{n_i - 1 - \sigma}{n_i + S_i - 1 - \sigma} \right) \geq e^{-\epsilon}$ for all $h = 1, ..., K^{(n)}_m$ under the assumption that H has density $g \leq C_2$;
3. $\frac{n_t + S_t - \sigma}{n_t - \sigma} \frac{n_l - 1 - \sigma}{n_l + S_l - 1 - \sigma} \in [e^{-\epsilon}, e^{\epsilon}]$ for all $t \neq l \iff S_i \leq (e^{\epsilon} - 1)(n_i - 1 - \sigma)$ for all $i = 1, ..., j$;
4. $\frac{\theta + (j + k)\sigma}{\theta + j\sigma} \frac{n_l - 1 - \sigma}{n_l + S_l - 1 - \sigma} \in [e^{-\epsilon}, e^{\epsilon}]$ for all $l = 1, .., j$.

□

Theorem 4.1 provides a practical way to calibrate the release mechanism. Chosen ϵ and fixed the parameters of the PY process, it is possible to estimate via Monte Carlo the probability in Equation 4.1 and determine a lower bound for the value of δ. If the required δ is too large for the application we are considering, it is possible to modify the process parameters. If no θ and σ are suitable for obtaining the desired δ, the mechanism cannot guarantee differential privacy for the ϵ considered.
Tables 1 and 2 show the Monte Carlo estimates of $\tilde{\delta}(\epsilon)$ obtained through $N = 10000$ MC iterations, together with the asymptotic error $z_{1-\alpha/2}\ \hat{\sigma}/\sqrt{N}$, where $z_{1-\alpha/2}$ is the quantile of order $1 - \alpha/2$ of a standard normal distribution, with $\alpha = 0.000001$, and $\hat{\sigma}^2$ is the sample variance. In particular, we simulate n private data by sampling from the PY process, assuming the uniform distribution in the interval $[0, 1]$ as the base measure, and then, we estimate $\tilde{\delta}(\epsilon)$ generating for each MC iteration the synthetic dataset consisting of m observations. We can see that by going from Table 1, where we assume $\sigma = 1/2$ to Table 2 where $\sigma = 1/5$ the values of $\tilde{\delta}(\epsilon)$ decrease. In fact, while to have $\tilde{\delta}(\epsilon) = 0$ in the first case it is necessary to choose $\epsilon = 5$, in the second case it is already sufficient to take ϵ values near 4. Moreover, as the number of released data increases, as we might expect, generally $\tilde{\delta}(\epsilon)$ increases.

m	50	125	250
$\epsilon = 1$	0.6973 ± 0.0202	0.9154 ± 0.0122	0.9768 ± 0.0066
$\epsilon = 2$	0.6529 ± 0.0210	0.8714 ± 0.0147	0.9577 ± 0.0088
$\epsilon = 4$	0.0817 ± 0.0120	0.3943 ± 0.0219	0.4459 ± 0.0216
$\epsilon = 5$	0 ± 0	0 ± 0	0 ± 0

Table 1: Estimated $\tilde{\delta}(\epsilon)$ for $n = 500$, $\theta = 3, \sigma = 1/2$ while varying ϵ and m.

m	50	125	250
$\epsilon = 1$	0.5969 ± 0.0216	0.6672 ± 0.0208	0.8807 ± 0.01418
$\epsilon = 2$	0.4009 ± 0.0276	0.6967 ± 0.02032	0.9133 ± 0.0124
$\epsilon = 4$	0.00009 ± 0.00004	0.0004 ± 0.0008	0.0024 ± 0.0022
$\epsilon = 5$	0 ± 0	0 ± 0	0 ± 0

Table 2: Estimated $\tilde{\delta}(\epsilon)$ for $n = 500$, $\theta = 3, \sigma = 1/5$ while varying ϵ and m.

The following corollary considers a sub-case of the Theorem 4.1:

Corollary 4.1. *Consider data $\mathbf{x} = (x_1, \ldots, x_n)$ displaying $K_n = j$ unique values each appearing n_i, $i = 1, \ldots, j$ times. Suppose $n_i > 1$ for all i. We suppose $\mathbb{X} = [0,1]^r$ for some $r \geq 1$ and H uniform distribution on $\mathbb{X}$. Moreover, let us fix $\epsilon > 0$ and define $\tilde{\delta}(\epsilon)$ as follows*

$$\tilde{\delta}(\epsilon) = 1 - \mathbb{P}\left[\bigcap_{h=1}^{K_m^{(n)}} \left\{ \max_{i=1,\ldots,j} \left(\frac{n_i - 1}{n_i + S_i - 1} \right) \frac{S_{j+h}}{\theta} \leq e^{\epsilon} \right\}, \right.$$
$$\bigcap_{h=1}^{K_m^{(n)}} \left\{ S_{j+h} \geq \frac{\theta}{e^{\epsilon}} \max_{i=1,\ldots,j} \left(\frac{n_i + S_i - 1}{n_i - 1} \right) \right\},$$
$$\left. \bigcap_{i=1}^{j} \{ S_i \leq (e^{\epsilon} - 1)(n_i - 1) \} \mid K_n = j, \mathbf{N}_n = \mathbf{n} \right].$$

Releasing data sampling from the m-step posterior predictive distribution of a $\mathscr{D}(\sigma, H)$ satisfies instance-level (ϵ, δ)-differential privacy for every δ such that $\delta > \tilde{\delta}(\epsilon)$.

Proof. H uniform on $[0,1]^r$ implies $C_1 = C_2 = 1$. We take $\sigma = 0$ since we are considering a DP prior.

Using the same procedure as described in the case of the PY process, we obtained Tables 3 and 4. In this case, having chosen uniform distribution in $[0,1]$ as the base measure and $\sigma = 0$, we are under the assumptions of Corollary 4.1. It can be seen that synthetic data generated through the DP allows for a higher privacy guarantee. Indeed, in both tables, $\epsilon = 3$ is sufficient to obtain an estimate of $\tilde{\delta}(\epsilon)$ equal to 0. We note that as θ increases, $\tilde{\delta}(\epsilon)$ also increases. In general, then, our mechanism provides lower privacy levels when the tails of the random probability measures we are assuming to model the private data are heavier.

m	50	125	250
$\epsilon = 1$	0.1561 ± 0.0160	0.23830 ± 0.01882	0.7834 ± 0.0181
$\epsilon = 1.8$	0 ± 0	0.0038 ± 0.00271	0.0558 ± 0.0101
$\epsilon = 2.2$	0 ± 0	0.00009 ± 0.0004	0.0024 ± 0.0021
$\epsilon = 3$	0 ± 0	0 ± 0	0 ± 0

Table 3: Estimated $\tilde{\delta}(\epsilon)$ for $n = 500$, $\theta = 2, \sigma = 0$ while varying ϵ and m.

m	50	125	250
$\epsilon = 1$	0.3367 ± 0.02087	0.6776 ± 0.02064	0.9985 ± 0.001709
$\epsilon = 2$	0.0004 ± 0.00088	0.0759 ± 0.01171	0.0017 ± 0.001819
$\epsilon = 2.8$	0 ± 0	0 ± 0	0.0003 ± 0.00007
$\epsilon = 3$	0 ± 0	0 ± 0	0 ± 0

Table 4: Estimated $\tilde{\delta}(\epsilon)$ for $n = 500$, $\theta = 5, \sigma = 0$ while varying ϵ and m.

Unlike other mechanisms for discrete data previously considered in the literature, our generates synthetic data whose values can be either the same as those observed or completely new (fully synthetic data). A priori, we would expect that generating fully synthetic data in large quantities would increase the "noise" added to the private data and thus lead to better ϵ and δ values. In reality, if the number of unique values K_n observed in the private data is relatively large, moving from one dataset to another at Hamming distance equal to 1, where one of the unique values in $\mathbf{X}$ has been replaced by a fully synthetic one, i.e. only observed in the generated $\mathbf{Z}$, leads to substantial density ratios and therefore not controllable by e^{ϵ} for small values of ϵ. In particular, if the number of unique values observed is large, an attacker who, in the worst case, knows the mechanism, all the data except the one of interest and $\mathbf{Z}$ can, by replacing one of the data in $\mathbf{X}$ with a fully synthetic data and obtaining a new dataset $\tilde{\mathbf{X}}$, understand which unique values $\mathbf{X}$ really contains. In fact, the probability

assigned by the mechanism to the synthetic data $\mathbf{Z}$ conditionally to $\mathbf{X}$ will be much greater than that assigned conditionally to $\tilde{\mathbf{X}}$.
Since the generation of totally synthetic data leads to an increase in differential privacy parameters, we propose the following approach as the last Bayesian Nonparametric mechanism: sample m observations from the predictive of a $\mathscr{PY}(\sigma, \theta)$ with $\sigma \geq 0$ and release only the data corresponding to the unique values already observed in $\mathbf{X}$. The generation of data according to this mechanism is described by Algorithm 3. This mechanism is very similar to the perturbation of counts of a histogram, proposed by Machanavajjhala et al. (2008) but differs from this in that the total sum of the released counts m is not fixed by the user but rather determined ex-post by the mechanism.

Algorithm 3 BNP histogram perturbation

Given a discrete dataset $\mathbf{X}$ and a number m of iterations;
compute the number $\tilde{k}$ of unique values in $\mathbf{X}$;
for $i = 1, ..., \tilde{k}$ **do**
  compute the frequency n_i of the i-th unique value X_i^*;
end for
set $h = 0$, $k = \tilde{k}$;
for $j = 1, ..., m$ **do**
  sample c from a categorical taking values in $\{1, ..., k+1\}$ with vector of probabilities:

$$\left(\frac{n_1 - \sigma}{\theta + n}, ..., \frac{n_k - \sigma}{\theta + n}, \frac{\theta + \sigma k}{\theta + n}\right)$$

  if $c \leq \tilde{k}$ **then**
    set $Z_h = X_c^*$;
    update $n_c = n_c + 1$;
    update $h = h + 1$;
  else if $\tilde{k} < c \leq k$ **then**
    update $n_c = n_c + 1$;
  else
    update the vector of frequencies to $(n_1, ..., n_k, 1)$ and the number of unique values to $k = k+1$;
  end if
end for
return $Z_1, Z_2, ...$

The way this mechanism is designed, it is still possible to write the law of the released data. Indeed, we generate synthetic observations according to the counts of the unique values observed in both the sample from the m-step posterior predictive distribution and the private data. Hence, the law of the observed data is equal to the law of the vector of counts $\mathbf{S}_{m-L_m^{(n)}} = (S_1, ..., S_{K_n})$, where S_q is defined in Equation 15, in the intersection with $m - L_m^{(n)}$, i.e. the number of released data according to the mechanism. Moreover, by Favaro et al. (2013) (Lemma 1), we have that

$$\mathbb{P}[\mathbf{S}_{m-L_m^{(n)}} = \mathbf{s} := (s_1, ..., s_j) \mid K_n = j, \mathbf{N}_n = \mathbf{n}, L_m^{(n)} = s] = (m-s)! \prod_{i=1}^{j} \frac{(n_i - \sigma)_{s_i}}{s_i!} \frac{1}{(n - j\sigma)_{m-s}} \tag{17}$$

representing the probability of observing counts of the old unique values equal to $(s_1, ..., s_j)$ conditioned on $\mathbf{X}$ and $L_m^{(n)} = s$. Favaro et al. (2013) provides also the joint law of $K_m^{(n)}$ and $L_m^{(n)}$. Hence, marginalizing out $K_m^{(n)}$ and multiplying the result to the probability (17), we obtain

$$\mathbb{P}[\mathbf{S}_{m-L_m^{(n)}} = \mathbf{s}, L_m^{(n)} = s \mid K_n = j, \mathbf{N}_n = \mathbf{n}_n] = \prod_{i=1}^{j} \frac{(n_i - \sigma)_{s_i}}{s_i!} \frac{m!}{s!(\theta + n)_m} (\theta + j\sigma)_s$$

This probability allows us to obtain the following result:

Proposition 4.2. *Consider data* $\mathbf{x} = (x_1, \ldots, x_n)$ *displaying* $K_n = j$ *unique values each appearing* n_i, $i =$

$1, \dots, j$ *times. Suppose* $n_i > 1$ *for all* i*. Moreover, let us fix* $\epsilon > 0$ *and define* $\tilde{\delta}(\epsilon)$ *as follows*

$$\tilde{\delta}(\epsilon) = 1 - \mathbb{P}\left[\bigcap_{i=1}^{j} \{S_i \leq (e^{\epsilon} - 1)(n_i - 1 - \sigma)\} \mid K_n = j, \mathbf{N}_n = \mathbf{n}\right]. \tag{18}$$

Releasing the perturbed counts of the unique values observed in $\mathbf{X}$ *through the mechanism described above satisfies instance-level* (ϵ, δ)*-differential privacy for every* δ *such that* $\delta > \tilde{\delta}(\epsilon)$*.*

Let us point out that, in this case, it is also possible to write down the expression of $\tilde{\delta}(\epsilon)$ analytically (see Appendix). Still, the result is a sum over all weak compositions, which can only be calculated in reasonable computational times for small values of n and m. Furthermore, we notice that expression (17) is very close to the predictive of a Dirichlet-Multinomial, but it has additional factors because the sum of the counts is also random.
Tables 5 and 6 show the results of some MC estimations of $\tilde{\delta}(\epsilon)$ according to Equation 18. As in the previous simulations, we consider $N = 10000$ MC iterations, where we generate synthetic data as specified in the Algorithm 3 while varying m. It can be seen that, in this case, we manage to obtain low values, i.e. near 0, of $\tilde{\delta}(\epsilon)$ even taking $\epsilon \leq 1$, unlike the other mechanisms.

m	50	125	250
$\epsilon = 0.5$	0.0005 ± 0.0009	0.0128 ± 0.0049	0.0450 ± 0.0091
$\epsilon = 1.0$	0 ± 0	0 ± 0	0 ± 0

Table 5: Estimated $\tilde{\delta}(\epsilon)$ for $n = 500$, $\theta = 3, \sigma = 0$ while varying ϵ and m.

m	50	125	250
$\epsilon = 0.5$	0.8183 ± 0.0202	0.9915 ± 0.0031	0.999 ± 0.0040
$\epsilon = 0.9$	0.00009 ± 0.00044	0.01470 ± 0.0053	0.1793 ± 0.0169
$\epsilon = 1.5$	0 ± 0	0 ± 0	0 ± 0

Table 6: Estimated $\tilde{\delta}(\epsilon)$ for $n = 500$, $\theta = 5, \sigma = 1/2$ while varying ϵ and m.

Although the number of the data released is a priori unknown, we can still provide its Bayes estimator under the squared loss function; indeed, by Lijoi et al. (2008) (Proposition 2), we know that

$$\mathbb{E}[L_m^{(n)} \mid K_n = j, \mathbf{N}_n = \mathbf{n}] = m\frac{\theta + j\sigma}{\theta + n}$$

Hence, by linearity, the posterior expectation of the number of released data is equal to

$$\mathbb{E}[m - L_m^{(n)} \mid K_n = j, \mathbf{N}_n = \mathbf{n}] = m\frac{n - j\sigma}{\theta + n}$$

This equation provides an interesting trade-off between the process parameters and the number of released data. If the number of unique values in the private data is significant, e.g. $K_n = j$ is close to n, the number of data released will be small unless we take a low value of σ, which in the limit case of $\sigma = 0$ corresponds to the DP prior. Moreover, the parameter θ also controls the number of released data; the higher value of θ leads to a smaller number of data generated by the mechanism.

5. Consistency

5.1. The problem of measuring the statistical utility

Once we have verified that a release mechanism has privacy guarantees, i.e. it satisfies ϵ-differential privacy or one of its generalisations, an essential task of the statistician is to assert the utility of the released data. First, we must emphasise that establishing the "goodness" of the synthetic data depends on their use and the

statistical analysis we want to conduct. In this section we assume that $X_1, ..., X_n \overset{\text{i.i.d}}{\sim} \mathfrak{p}_0$, i.e. there exists a data generating distribution $\mathfrak{p}_0$.
If the statistician's problem is to estimate a certain parameter θ depending on the data distribution, the frequentist approach involves constructing an estimator $\hat{\theta}(X_1, ..., X_n)$. In the case one does not know $\mathbf{X}$ but only $\mathbf{Z}$ is available, it is legitimate to ask which is the release mechanism Q_n such that the estimator based on the synthetic data $\hat{\theta}(Z_1, ..., Z_m)$ is the best possible one, in the sense that that achieves the fastest rate of convergence to the parameter θ. The recent work by Duchi et al. (2018) is concerned with studying this problem in the case of local differential privacy, finding for several estimation problems the mechanism and estimator pair that achieves the ϵ-private minimax risk, defined as follows

$$\inf_{Q \in \mathcal{Q}_\epsilon} \inf_{\hat{\theta}} \sup_{\mathfrak{p}_0 \in \mathcal{P}(\mathbb{X})} \mathbb{E}[\Phi(\varrho(\hat{\theta}(Z_1, ..., Z_n), \theta(\mathfrak{p}_0)))]$$

where Q is a (local) privacy mechanism applied to each of the private data X_i, $\mathcal{Q}_\epsilon$ is the class of all the mechanisms that satisfy ϵ-differential privacy, ϱ a pseudo-metric and Φ a non-decreasing positive function such that $\Phi(0) = 0$. Butucea et al. (2020) follow the same approach considering density estimation problems and develop a novel wavelet-based perturbation mechanism which satisfies ϵ-differential privacy and is minimax optimal up to an additional logarithmic factor.
Although the mechanisms proposed by Duchi et al. (2018) and Butucea et al. (2020) achieve ϵ-private minimax risk, they are designed to perform optimally in the estimation problem under consideration. For example, Butucea et al. (2020) releases Z_i, which is an evaluation of a wavelet basis in X_i suitably perturbed with Laplace noise. This mechanism is ϵ-private minimax optimal for estimating the density because it provides a natural way of defining the density estimator based on the released data but does not guarantee that the released data are statistically useful in general for performing various types of inference.
Here, differently, we follow the approach of Wasserman and Zhou (2010), which allows us to assert whether the data we release are generally useful and not for a specific estimation problem. In particular, we want to establish mathematically if a mechanism is informative, i.e. making inferences on the distribution of the private data $\mathbf{X}$ from the released data $\mathbf{Z}$ is possible. We assume that the user can access the synthetic data $\mathbf{Z}$ but not the mechanism. Since there exists a data-generating distribution, the marginal law of the released data is the probability measure defined as

$$\nu_{n,m}(A) := \int_{\mathbb{X}^n} Q_n(A \mid \mathbf{X} = \mathbf{x}) \; \mathfrak{p}_0(dx_1) \cdots \mathfrak{p}_0(dx_n) \tag{19}$$

for all $A \in \mathcal{X}^m$. We consider $m = m(n)$, i.e. the number of released data is a function of the dimensionality of the private dataset. In this way, any asymptotic statement involving n increasing implies that m also increases. For this reason, with abuse of notation, we will denote the distribution of Equation 19 by ν_n instead of $\nu_{n,m}$. Wasserman and Zhou (2010) study the utility of the privacy mechanisms by finding a rate of convergence at which $d(e_m^{(\mathbf{Z})}, \mathfrak{p}_0)$ goes to 0 in ν_n-probability as n goes to infinity, where

$$e_m^{(\mathbf{Z})} := \frac{1}{m} \sum_{i=1}^{m} \delta_{Z_i}$$

denotes the empirical measure based on the synthetic data, and d is a suitable distance distance between probability distributions. In particular, Wasserman and Zhou (2010) consider as distance d, the L^2 distance between the densities of the two probability distributions and the Kolmogorov-Smirnov (KS) distance, denoted by ρ and defined as follows

$$\rho(\mu_1, \mu_2) := \sup_{t \in \mathbb{R}^d} |F_1(t) - F_2(t)| = \|F_1 - F_2\|_{L^\infty}$$

with F_1, F_2 the comulative distribution functions of the two probability measures μ_1, μ_2 respectively. For instance, let us assume the density f of $\mathfrak{p}_0$ belongs to the class of Lipschitz functions, i.e. it satisfies

$$|f(x) - f(y)| \leq L \; |x - y|$$

for some $L > 0$ and any $x, y \in \mathbb{X}$. Let $a_n \asymp b_n$ hold if and only if there exists $m, M > 0$ such that $ma_n \leq b_n \leq Ma_n$, $\mathbb{E}_\mu[X]$ denotes the expectation of $X \sim \mu$ and $a_n = O(b_n)$ denotes the existence of a constant C such that $a_n \leq Cb_n$. Wasserman and Zhou (2010) show that releasing $\mathbf{Z}$ sampling from a smoothed histogram as detailed in Section 3.3 is consistent. In particular, choosing a number of buckets $k \asymp n^{r/(6+r)}$, a number of released data $m \asymp n^{4/(6+r)}$ and the smoothing parameter $\sigma = mk/n\epsilon$ minimizes $\mathbb{E}_{\nu_n}[\rho(e_m^{(\mathbf{Z})}, \mathfrak{p}_0)]$ under the ϵ-differential privacy constraints and leads to the following rate

$$\mathbb{E}_{\nu_n}\left[\rho(e_m^{(\mathbf{Z})}, \mathfrak{p}_0)\right] = O\left(\frac{\sqrt{\log n}}{n^{2/(6+r)}}\right)$$

where we consider $\mathbb{X} = [0,1]^r$ for some integer $r \geq 1$. This example shows how requiring a certain regularity on the unknown distribution $\mathfrak{p}_0$ is generally necessary to obtain consistency. In the case of the smoothed histogram, we ask the density of the generating distribution to be Liptschiz.

5.2. Integral Probability Metrics

The KS distance is a particular case of a broader family of metrics between probability measures, namely integral probability metrics (Zolotarev (1984), Müller (1997)).

Definition 5.1. *Let $\mathcal{F}$ be a class of $\mathbb{R}$-valued bounded measurable functions on a Polish space $\mathbb{X}$. The integral probability metric between $\mu_1, \mu_2 \in \mathcal{P}(\mathbb{X})$ is*

$$d_{\mathcal{F}}(\mu_1, \mu_2) = \sup_{f \in \mathcal{F}} \left| \int_{\mathbb{X}} f d\mu_1 - \int_{\mathbb{X}} f d\mu_2 \right|$$

To be a distance between distributions, the class of test functions $\mathcal{F}$ have to separate the elements in $\mathcal{P}(\mathbb{X})$, i.e. $\mu_1 = \mu_2$ is implied by the condition: $\mathbb{E}_{\mu_1}[f(X)] = \mathbb{E}_{\mu_2}[f(X)]$ for every $f \in \mathcal{F}$. The choice of $\mathcal{F}$ leads to different well-known metrics. The class of indicator functions of any measurable set $A \in \mathcal{X}$ corresponds to the total variation distance,

$$\|\mu_1 - \mu_2\|_{\mathrm{TV}} := \sup_{A \in \mathcal{X}} |\mu_1(A) - \mu_2(A)|$$

Let $(-\infty, \mathbf{t}] := (-\infty, t_1] \times \cdots \times (-\infty, t_d] \subset \mathbb{R}^d$, if one chooses $\mathcal{F} = \{\mathbb{1}_{(-\infty,\mathbf{t}]}(\cdot) : \mathbf{t} \in \mathbb{R}^d\}$, the integral probability metric coincides with the KS distance. In particular, the class of integral probability metrics also includes the 1-Wasserstein distance. Here, we recall the definition of p-Wasserstein distance (Ambrosio and Gigli (2013),Villani (2008)), with $p \geq 1$, that we will use to present the next sections' results. Let $(\mathbb{X}, d_{\mathbb{X}})$ be a separable complete metric space, we define

$$\mathcal{P}_p(\mathbb{X}) := \left\{ \gamma \in \mathcal{P}(\mathbb{X}) : \int_{\mathbb{X}} [d_{\mathbb{X}}(x, x_0)]^p \gamma(dx) < \infty \ \text{ for some } x_0 \in \mathbb{X} \right\}.$$

The p-Wasserstein distance for $p \geq 1$ is

$$\mathcal{W}_p^{\mathcal{P}_p(\mathbb{X})}(\gamma_1, \gamma_2) := \inf_{\xi \in \Pi(\gamma_1, \gamma_2)} \left(\int_{\mathbb{X}^2} [d_{\mathbb{X}}(x, y)]^p \xi(dxdy) \right)^{1/p} \tag{20}$$

for any $\gamma_1, \gamma_2 \in \mathcal{P}_p(\mathbb{X})$, where $\Pi(\gamma_1, \gamma_2)$ is the Fréchet class of all the couplings between γ_1, γ_2, i.e. the probability measures defined on $(\mathbb{X}^2, \mathcal{X}^2)$ with i-th marginal γ_i, $i = 1, 2$. We assume that $\mathbb{X}$ is totally bounded, so that $\mathcal{P}_p(\mathbb{X})$ coincides with the whole space $\mathcal{P}(\mathbb{X})$. In this setting, by Kantorovich-Rubinstein theorem, taking $\mu_1, \mu_2 \in \mathcal{P}(\mathbb{X})$, we have

$$\mathcal{W}_1^{\mathcal{P}(\mathbb{X})}(\mu_1, \mu_2) = \sup_{f \in \mathrm{Lip}_1(\mathbb{X})} \left| \int_{\mathbb{X}} f d\mu_1 - \int_{\mathbb{X}} f d\mu_2 \right|$$

where $\mathrm{Lip}_1(\mathbb{X})$ denotes the set of all the functions $f : \mathbb{X} \to \mathbb{R}$ that have Lipschitz constant equal to 1. It follows that the 1-Wasserstein distance is an example of integral probability metric choosing $\mathcal{F} = \mathrm{Lip}_1(\mathbb{X})$.
The definition of consistency that we propose follows Wasserman's but broadens the set of possible distances to all integral probability metrics.

Definition 5.2. *Q_n is informative or consistent with respect to the class $\mathcal{F}$ if $d_{\mathcal{F}}(e_m^{(\mathbf{Z})}, \mathfrak{p}_0) \to 0$ in ν_n-probability. Q_n is ϵ_n-informative if $\mathbb{E}_{\nu_n}[d_{\mathcal{F}}(e_m^{(\mathbf{Z})}, \mathfrak{p}_0)] = O(\epsilon_n)$, where $(\epsilon_n)_{n \geq 1}$ is a sequence of positive number such that $\epsilon_n \to 0$ as $n \to \infty$.*

We observe that ϵ_n-informative is a refinement of the definition of consistent mechanism. Indeed, it provides a rate of convergence and implies by Markov's inequality that, for any $a > 0$

$$\nu_n \left(\left\{ d_{\mathcal{F}}(e_m^{(\mathbf{Z})}, \mathfrak{p}_0) \geq a \right\} \right) \leq \frac{\mathbb{E}_{\nu_n} \left[d_{\mathcal{F}}(e_m^{(\mathbf{Z})}, \mathfrak{p}_0) \right]}{a} \leq \frac{\epsilon_n}{a} \to 0$$

as $n \to \infty$, obtaining convergence in ν_n-probability. According to this definition, limiting oneself to the KS distance is no longer necessary, and we include both this distance and the 1-Wasserstein metric in a general framework.

5.3. Basics of Empirical Process theory

The rate in the ϵ_n-informative mechanism definition is strictly related to the Empirical Process theory. This theory plays a crucial role in probability and statistics since it deals with the asymptotic behaviour of an empirical process based on the collected data to the data generating distribution.
As in the previous section, we assume $X_1, ..., X_n \overset{\text{i.i.d.}}{\sim} \mathfrak{p}_0$. Moreover, we consider $\mathbb{X} \subseteq \mathbb{R}$. A main result in Empirical Process theory is the so-called Glivenko-Cantelli's theorem, which states the following

$$\rho(e_n^{(\mathbf{X})}, \mathfrak{p}_0) \to 0$$

as $n \to \infty$, almost surely. Glivenko-Cantelli's theorem is strengthened by the Dvoretzky-Kiefer-Wolfowitz (DKF) inequality (Massart (1990)), that is

$$\mathbb{P}\left[\rho(e_n^{(\mathbf{X})}, \mathfrak{p}_0) > \epsilon\right] \leq 2e^{-2n\epsilon^2}$$

for every $\epsilon > 0$. We remark that this inequality can also be generalized to the case $\mathbb{X} \subseteq \mathbb{R}^d$. In particular, DKF inequality implies the following

$$\mathbb{E}_{\otimes^n \mathfrak{p}_0}\left[\rho(e_n^{(\mathbf{X})}, \mathfrak{p}_0)\right] = O\left(\sqrt{\frac{\log n}{n}}\right) \tag{21}$$

Indeed, let $\epsilon = \epsilon_n$, we have

$$\mathbb{E}_{\otimes^n \mathfrak{p}_0}\left[\rho(e_n^{(\mathbf{X})}, \mathfrak{p}_0)\right] = \int_{A_n} \rho(e_n^{(\mathbf{X})}, \mathfrak{p}_0)\ d\mathbb{P} + \int_{A_n^c} \rho(e_n^{(\mathbf{X})}, \mathfrak{p}_0)\ d\mathbb{P}$$

where $A_n := \{\rho(e_n^{(\mathbf{X})}, \mathfrak{p}_0) \leq \epsilon_n\}$. Thus

$$\int_{A_n} \rho(e_n^{(\mathbf{X})}, \mathfrak{p}_0)\ d\mathbb{P} + \int_{A_n^c} \rho(e_n^{(\mathbf{X})}, \mathfrak{p}_0)\ d\mathbb{P} \leq \epsilon_n + 2\mathbb{P}[\rho(\hat{F}_n, F) > \epsilon_n] \leq \epsilon_n + 4e^{-2n\epsilon_n^2}$$

and considering $\epsilon_n = \sqrt{\log n/n}$, we obtain (21).
Moreover, Glivenko-Cantelli's theorem can be restated by saying that $e_n^{(\mathbf{X})}$ tends weakly to $\mathfrak{p}_0$ as $n \to \infty$, in probability. We refer to van der Vaart and Wellner (1996) for a complete overview and various modes of convergence. Since the p-Wasserstein distance metrizes the weak convergence, we can ask if it is possible to quantify the convergence of the empirical measure by providing a non-asymptotic estimate of the following type

$$\varepsilon_{n,p}(\mathbb{X}, \mathfrak{p}_0) := \mathbb{E}_{\otimes^n \mathfrak{p}_0}[\mathcal{W}_p^{\mathcal{P}(\mathbb{X})}(e_n^{(\mathbf{X})}, \mathfrak{p}_0)] \leq g(n) \tag{22}$$

for all $n \geq 1$, where g is a suitable function. Fournier and Guillin (2015) provide an answer in the case $\mathbb{X} \subseteq \mathbb{R}^d$ with the following theorem

Theorem 5.1. *Let $\mathfrak{p}_0$ a probability measure on $\mathbb{R}^d$ and let $p \geq 1$. Assume that*

$$M_q(\mathfrak{p}_0) := \int_{\mathbb{X}} |x|^q\ \mathfrak{p}_0(dx) < \infty$$

for some $q > p$. There exists a constant C depending only on p, d, q such that for all $n \geq 1$,

$$\varepsilon_{n,p}(\mathbb{X}, \mathfrak{p}_0) \leq C\ M_q^{p/q}(\mathfrak{p}_0)\ \times \\ \times \begin{cases} n^{-1/2} + n^{-(q-p)/q} & \text{if } p > d/2 \text{ and } q \neq 2p \\ n^{-1/2}\log(1+n) + n^{-(q-p)/q} & \text{if } p = d/2 \text{ and } q \neq 2p \\ n^{-p/d} + n^{-(q-p)/q} & \text{if } p \in [1, d/2) \text{ and } q \neq d/(d-p) \end{cases}$$

As Fournier and Guillin (2015) observe, when $\mathfrak{p}_0$ ha sufficiently many moments, the term $n^{-(q-p)/q}$ is small and can be neglected.

5.4. Bayesian consistency

The BNP mechanism's consistency for the DP prior is closely linked to Bayesian consistency. For completeness, we recall the main ideas in the field of Bayesian consistency and then show the recent result of Camerlenghi et al. (2022) on which the proof of our consistency theorem is based. A posterior distribution is consistent if

it concentrates around the true distribution asymptotically. Let us consider a Bayesian nonparametric model, (weakly) consistency requires that, given a true distribution $\mathfrak{p}_0 \in \mathcal{P}(\mathbb{X})$, $\pi_n(U_0^c \mid \mathbf{X}) \to 0$ as $n \to \infty$ in probability for any neighbourhood U_0 of $\mathfrak{p}_0$, where we have assumed $X_1, ..., X_n$ a sample from $\mathfrak{p}_0$. Since our method of releasing synthesized data consists of sampling from the posterior predictive distribution of a Bayesian model, we expect that if the posterior asymptotically describes the data distribution "well", $\mathbf{Z}$ will contain information about the private data.

A way to quantify Bayesian consistency is to provide a posterior contraction rate (PCR) (Ghosal and van der Vaart (2017)). Let $d_{\mathcal{P}(\mathbb{X})}$ be a metric in $\mathcal{P}(\mathbb{X})$ and $\mathcal{Q}(\mathbb{X}) := \mathcal{P}(\mathcal{P}(\mathbb{X}))$ denote the space of probability measure on the space of probability measure on $\mathbb{X}$ (totally bounded Polish space), we have the following definition (Ghosal and van der Vaart (2017), Camerlenghi et al. (2022))

Definition 5.3. *Let π be a prior distribution that belongs to $\mathcal{Q}(\mathbb{X})$. Then, a PCR at $\mathfrak{p}_0$ is defined as any sequence $(\epsilon_n)_{n\geq 1}$ of positive numbers for which, as $n \to \infty$,*

$$\pi_n\left(\left\{\mathfrak{p} \in \mathcal{P}(\mathbb{X}) : d_{\mathcal{P}(\mathbb{X})}(\mathfrak{p}, \mathfrak{p}_0) \geq M_n \epsilon_n\right\} \mid \mathbf{X}\right) \to 0$$

holds in $\otimes^n \mathfrak{p}_0$-probability for every $(M_n)_{n\geq 1}$ of positive numbers such that $M_n \to +\infty$ whenever $n \to \infty$.

Camerlenghi et al. (2022) consider the special case of $d_{\mathcal{P}(\mathbb{X})}$ coincident with the p-Wasserstein metric and assume $\otimes^n \mathfrak{p}_0 \ll \mu_n$ to render the above definition well-defined and not dependent on the choice of the probability kernel $\pi_n(\,\cdot \mid \cdot\,)$ satisfying Equation 4. Indeed, the posterior is not uniquely defined as a point-wise mapping as the solution of the disintegration in Equation 4, and any two solutions π_n and $\pi_n^{'}$ satisfies $\pi_n(\,\cdot \mid \mathbf{x}) = \pi_n^{'}(\,\cdot \mid \mathbf{x})$, as elements of $\mathcal{Q}(\mathbb{X})$, for all $\mathbf{x} \in \mathbb{X}^n \setminus N_n$, where N_n is a μ_n-null set. Thus, the assumption $\otimes^n \mathfrak{p}_0 \ll \mu_n$ entails that $\mathbf{X}$ takes value in N_n with $\otimes^n \mathfrak{p}_0$-probability zero, yielding the desired well-definiteness (Camerlenghi et al. (2022)).

We point out that, if $(\mathbb{X}, d_{\mathbb{X}})$ is a complete separable metric space with $\mathbb{X}$ totally bounded, then $\mathcal{P}(\mathbb{X})$ endowed with the distance $\mathcal{W}_p^{\mathcal{P}(\mathbb{X})}$ is a complete separable metric space too (Villani (2008)). Hence, replacing $(\mathbb{X}, d_{\mathbb{X}})$ with the space $(\mathcal{P}(\mathbb{X}), \mathcal{W}_p^{\mathcal{P}(\mathbb{X})})$ in Equation (20), we can define the p-Wasserstein metric $\mathcal{W}_p^{\mathcal{Q}(\mathbb{X})}$ between elements of $\mathcal{Q}(\mathbb{X})$. Camerlenghi et al. (2022) show that the sequence $(\epsilon_n)_{n\geq 1}$, where

$$\epsilon_n := \mathbb{E}_{\otimes^n \mathfrak{p}_0}\left[\mathcal{W}_p^{\mathcal{Q}(\mathbb{X})}(\pi_n(\,\cdot \mid \mathbf{X}), \delta_{\mathfrak{p}_0})\right]$$

gives a PCR according to Definition 5.3 with respect to the p-Wasserstein distance.

Moreover, they show a general theorem for obtaining a rate ϵ_n of the form $n^{-\gamma}$ for some γ positive constant under different Bayesian Nonparametric models. In particular, assuming the following

$$\mathcal{W}_1^{\mathcal{P}(\mathbb{X})}(\alpha_n(\,\cdot \mid \mathbf{X} = \mathbf{x}), \alpha_n(\,\cdot \mid \mathbf{X} = \mathbf{y})) \leq L\ \mathcal{W}_1^{\mathcal{P}(\mathbb{X})}(e_n^{(\mathbf{x})}, e_n^{(\mathbf{y})}) \tag{23}$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{X}^n$, being α_n the 1-step posterior predictive distribution and $L > 0$ a positive constant, they show is possible to obtain an explicit rate. For the DP prior, Equation (11) with $\sigma = 0$ implies

$$\mathcal{W}_1^{\mathcal{P}(\mathbb{X})}(\alpha_n(\,\cdot \mid \mathbf{X} = \mathbf{x}), \alpha_n(\,\cdot \mid \mathbf{X} = \mathbf{y})) = \frac{n}{\theta + n}\ \mathcal{W}_1^{\mathcal{P}(\mathbb{X})}(e_n^{(\mathbf{x})}, e_n^{(\mathbf{y})}),$$

hence condition (23) holds. As a consequence, we have (Camerlenghi et al. (2022))

Proposition 5.1. *Let $(\mathbb{X}, \mathcal{X})$ be a totally bounded metric space and $\pi \in \mathcal{Q}(\mathbb{X})$ coincide with the DP prior. We assume $\otimes^n \mathfrak{p}_0 \ll \mu_n$ and, for any $C \in \mathcal{X}$, we require that*

$$H(C) + \mathfrak{p}_0(C) \leq K(diam(C))^{\eta d} \tag{24}$$

for some constants $K > 0$ and $\eta \in (0, 1]$. Then

$$\mathbb{E}_{\otimes^n \mathfrak{p}_0}\left[\mathcal{W}_1^{\mathcal{Q}(\mathbb{X})}(\pi_n(\,\cdot \mid \mathbf{X}), \delta_{\mathfrak{p}_0})\right] = \varepsilon_{n,1}(\mathbb{X}, \mathfrak{p}_0) + O(n^{-1/[d(2-\eta)+2]})$$

where d is the dimension of the space $\mathbb{X}$ and $\varepsilon_{n,1}(\mathbb{X}, \mathfrak{p}_0)$ denotes the rate for the mean Glivenko-Cantelli's theorem shown in Section 5.3.

We recall that Definition 5.3 can be reformulated in the case of Bayesian parametric models by replacing the metric space $(\mathbb{X}, d_{\mathbb{X}})$ with the metric space (Θ, d_Θ), where Θ is the parameter space (see Section 2.3.1). We denote $\mathcal{W}_p^{\mathcal{P}(\Theta)}(\cdot, \cdot)$ the p-Wasserstein distance defined in (Θ, d_θ), assuming this a complete and separable metric space. Dolera et al. (2023) show that the sequence $(\epsilon_n)_{n\geq 1}$, where

$$\epsilon_n := \mathbb{E}_{\otimes^n \mathfrak{p}_0}\left[\mathcal{W}_p^{\mathcal{P}(\Theta)}(\psi_n(\,\cdot \mid \mathbf{X}), \delta_{\theta_0})\right]$$

with θ_0 the true parameter, i.e. $\mathscr{T}(\theta_0) = \mathfrak{p}_0$, and ψ_n the posterior of the Bayesian parametric model, is a PCR at θ_0 with respect to the p-Wasserstein distance.
They consider various models, including the Multinomial. Let $X_1, ..., X_n$ be a sequence of categorical random variables taking values in the finite set $\{a_1, ..., a_k\} \subset \mathbb{X}$. It is easy to show that the parameter space Θ, in this case, coincides with the interior of the $k-1$ dimensional simplex, namely

$$\mathscr{S}^{k-1} = \{(\theta_1, ..., \theta_k) \in [0,1]^{k-1} : \sum_{i=1}^{k-1} \theta_i \leq 1\}$$

Dolera et al. (2023) show the following

Proposition 5.2. *Let $k \geq 2$. Let ϕ a prior on $\mathscr{S}^{k-1}$. If ϕ has a density q (with respect to the Lebesgue measure) such that $q \in C^1(\overline{\mathscr{S}^{k-1}})$ and $q(\theta) = 0$ for any $\theta \in \partial\mathscr{S}^{k-1}$, then as $n \to \infty$*

$$\mathbb{E}_{\otimes^n \mathfrak{p}_0}\left[\mathcal{W}_2^{\mathcal{P}(\Theta)}(\psi_n(\,\cdot \mid \mathbf{X}), \delta_{\theta_0})\right] = O(n^{-1/2})$$

which is the optimal rate.

We point out that the above proposition holds, assuming a Dirichlet prior with parameters $(\alpha_1, ..., \alpha_k)$ such that $\alpha_i \geq 2$ for all $i = 1, ..., k$.

5.5. BNP mechanism consistency

As a first result of consistency, we show that the mechanism of releasing samples from the m-step posterior predictive distribution of a Dirichlet-Multinomial model proposed by Machanavajjhala et al. (2008) is consistent with respect to both the KS metric and the 1-Wasserstein distance. We do this because, so far, it has yet to be shown in the literature that this approach satisfies the definition 5.2 but also because, in the case of the proposition 4.1, the BNP mechanism coincides with this mechanism. The preliminary result we show concerns a property of the operator $\mathscr{T} : (\Theta, d_\Theta) \to (\mathcal{P}(\mathbb{X}), \mathcal{W}_1^{\mathcal{P}(\mathbb{X})})$, which for a Dirichlet-Multinomial model is defined as follows

$$\mathscr{T}(\theta) = \sum_{i=1}^{k} \theta_i \delta_{a_i}$$

for all $\theta \in \Theta \equiv \mathscr{S}^{k-1}$, where $\theta_k := 1 - \theta_1 - \cdots - \theta_{k-1}$. We assume d_Θ equal to the metric induced by the l^2 norm in $\mathbb{R}^{k-1}$.

Lemma 5.1. *The map $\mathscr{T} : (\Theta, d_\Theta) \to (\mathcal{P}(\mathbb{X}), \mathcal{W}_1^{\mathcal{P}(\mathbb{X})})$ defined above is Lipschitz continuous, with Lipschitz constant less or equal to $\sqrt{k-1}D$, with $D := diam(\mathbb{X})$.*

Proof. For any $\theta, \phi \in \Theta$, by Villani (2008) (Theorem 6.15), we have

$$\mathcal{W}_1^{\mathcal{P}(\mathbb{X})}(\mathscr{T}(\theta), \mathscr{T}(\phi)) \leq D \, \|\mathscr{T}(\theta) - \mathscr{T}(\phi)\|_{\mathrm{TV}}$$

By Roch (2024) (Lemma 4.1.9), we can say

$$\|\mathscr{T}(\theta) - \mathscr{T}(\phi)\|_{\mathrm{TV}} = \frac{1}{2} \sum_{i=1}^{k} |\theta_i - \phi_i|$$

where

$$\begin{aligned}\sum_{i=1}^{k} |\theta_i - \phi_i| = &\sum_{i=1}^{k-1} |\theta_i - \phi_i| + |\theta_k - \phi_k| = \\ &\sum_{i=1}^{k-1} |\theta_i - \phi_i| + |-\theta_{k-1} + \phi_{k-1} + ... - \theta_1 + \phi_1| \leq \\ &2\sum_{i=1}^{k-1} |\theta_i - \phi_i| \leq 2\sqrt{k-1}\, \|\theta - \phi\|_{l^2}\end{aligned}$$

by triangular inequality and Cauchy-Schwarz inequality. □

We want to remark that the assumption required by the Dirichlet-Multinomial mechanism to satisfy ϵ-differential privacy is the following

$$\alpha_j + C_j \geq m/(e^\epsilon - 1) \tag{25}$$

for all $j = 1, ..., k$, and if n goes to infinity, we can assume also the counts $C_j := |\{i \in \{1, ..., n\} : X_i \in B_j\}|$ goes to infinity. Indeed, the existence of a true distribution $\mathfrak{p}_0 = \mathscr{T}(\theta_0)$ with $\theta_0 = (\theta_{0,j})_{j=1}^k$, implies that $(C_1, ..., C_k) \sim \text{Multinomial}(n, \theta_0)$. Hence, C_j is marginally distributed as a Binomial distribution with parameters $(n, \theta_{0,j})$ for all $j = 1, ..., k$. By Chebyshev's inequality

$$\mathbb{P}\left(|C_j/n - \theta_{0,j}| > \epsilon\right) \leq \frac{\theta_{0,j}(1-\theta_{0,j})}{n\epsilon^2} \to 0,$$

which shows that C_j grows linearly with n, at least in probability. We can, therefore, assume that when $n \to \infty$, m also goes to infinity linearly with n and the condition 25 is still satisfied. To be precise, however, the following result expresses the convergence rate as a function of n and m.

Theorem 5.2. *Let $\alpha_i \geq 2$ for all $i = 1, ..., k$ with $k \geq 2$. We consider $\mathbb{X} \subset \mathbb{R}^d$. If we release $\mathbf{Z}$ from the m-step posterior predictive distribution of a Dirichlet-Multinomial model, then*

$$\mathbb{E}_{\nu_{n,m}}\left[\mathcal{W}_1^{\mathcal{P}(\mathbb{X})}(e_m^{(\mathbf{Z})}, \mathfrak{p}_0)\right] = O(n^{-1/2}) + \begin{cases} O(m^{-1/2}) & , \text{if } d < 2, \\ O(m^{-1/2}\log(1+m)) & , \text{if } d = 2, \\ O(m^{-1/d}) & , \text{if } d > 2 \end{cases}$$

Proof. The mechanism Q_n has the following density with respect to the counting measure

$$q_n(\mathbf{z} \mid \mathbf{X} = \mathbf{x}) = \int_\Theta \left[\prod_{j=1}^m f_\theta(z_j)\right] \psi_n(d\theta \mid \mathbf{X} = \mathbf{x})$$

where $f_\theta(z) = \left(\prod_{i=1}^k \theta_i^{\mathbb{1}_{\{a_i\}}(z)}\right) \mathbb{1}_{\{a_1,...,a_k\}}(z)$. Hence, by Fubini-Tonelli's theorem, we have

$$\begin{aligned}
&\mathbb{E}_{\nu_{n,m}}\left[\mathcal{W}_1^{\mathcal{P}(\mathbb{X})}(e_m^{(\mathbf{Z})}, \mathfrak{p}_0)\right] = \int_{\mathbb{X}^m \times \mathbb{X}^n} \mathcal{W}_1^{\mathcal{P}(\mathbb{X})}(e_m^{(\mathbf{z})}, \mathfrak{p}_0)\ q_n(\mathbf{z} \mid \mathbf{X} = \mathbf{x})\ d\mathbf{z}\ \mathfrak{p}_0^n(d\mathbf{x}) = \\
&\int_{\mathbb{X}^m \times \mathbb{X}^n} \left(\int_\Theta \mathcal{W}_1^{\mathcal{P}(\mathbb{X})}(e_m^{(\mathbf{z})}, \mathfrak{p}_0) \left[\prod_{j=1}^m f_\theta(z_j)\right] \psi_n(d\theta \mid \mathbf{X} = \mathbf{x})\right) d\mathbf{z}\ \mathfrak{p}_0^n(d\mathbf{x}) = \\
&\int_{\Theta \times \mathbb{X}^n} \left(\int_{\mathbb{X}^m} \mathcal{W}_1^{\mathcal{P}(\mathbb{X})}(e_m^{(\mathbf{z})}, \mathfrak{p}_0) \left[\prod_{j=1}^m f_\theta(z_j)\right] d\mathbf{z}\right) \psi_n(d\theta \mid \mathbf{X} = \mathbf{x})\ \mathfrak{p}_0^n(d\mathbf{x}) = \mathbb{E}_{\psi_n \otimes \mathfrak{p}_0^n}\left[\mathbb{E}\left[\mathcal{W}_1^{\mathcal{P}(\mathbb{X})}(e_m^{(\mathbf{Z})}, \mathfrak{p}_0) \mid \theta\right]\right]
\end{aligned}$$

By triangular inequality, we have

$$\mathbb{E}_{\psi_n \otimes \mathfrak{p}_0^n}\left[\mathbb{E}\left[\mathcal{W}_1^{\mathcal{P}(\mathbb{X})}(e_m^{(\mathbf{Z})}, \mathfrak{p}_0) \mid \theta\right]\right] \leq \mathbb{E}_{\psi_n \otimes \mathfrak{p}_0^n}\left[\mathbb{E}\left[\mathcal{W}_1^{\mathcal{P}(\mathbb{X})}(e_m^{(\mathbf{Z})}, \mathscr{T}(\theta)) \mid \theta\right]\right] + \mathbb{E}_{\psi_n \otimes \mathfrak{p}_0^n}\left[\mathbb{E}\left[\mathcal{W}_1^{\mathcal{P}(\mathbb{X})}(\mathscr{T}(\theta), \mathfrak{p}_0) \mid \theta\right]\right].$$

By the property of the conditional expectation

$$\mathbb{E}_{\psi_n \otimes \mathfrak{p}_0^n}\left[\mathbb{E}\left[\mathcal{W}_1^{\mathcal{P}(\mathbb{X})}(\mathscr{T}(\theta), \mathfrak{p}_0) \mid \theta\right]\right] = \mathbb{E}_{\psi_n \otimes \mathfrak{p}_0^n}\left[\mathcal{W}_1^{\mathcal{P}(\mathbb{X})}(\mathscr{T}(\theta), \mathfrak{p}_0)\right].$$

We can write

$$\mathbb{E}_{\psi_n \otimes \mathfrak{p}_0^n}\left[\mathcal{W}_1^{\mathcal{P}(\mathbb{X})}(\mathscr{T}(\theta), \mathfrak{p}_0)\right] = \mathbb{E}_{\mathfrak{p}_0^n}\left[\int_{\Theta^2} \mathcal{W}_1^{\mathcal{P}(\mathbb{X})}(\mathscr{T}(\theta), \mathscr{T}(\phi))\ \psi_n(d\theta|\mathbf{X} = \mathbf{x})\ \delta_{\theta_0}(d\phi)\right].$$

By Lemma 5.1 and Hölder's inequality, we obtain

$$\begin{aligned}
&\mathbb{E}_{\otimes^n \mathfrak{p}_0}\left[\int_{\Theta^2} \mathcal{W}_1^{\mathcal{P}(\mathbb{X})}(\mathscr{T}(\theta), \mathscr{T}(\phi))\ \psi_n(d\theta \mid \mathbf{X} = \mathbf{x})\ \delta_{\theta_0}(d\phi)\right] \leq \\
&D\sqrt{k-1}\ \mathbb{E}_{\otimes^n \mathfrak{p}_0}\left[\int_{\Theta^2} \|\theta - \phi\|_{l^2}\ \psi_n(d\theta|\mathbf{X} = \mathbf{x})\ \delta_{\theta_0}(d\phi)\right] \leq \\
&D\sqrt{k-1}\ \mathbb{E}_{\otimes^n \mathfrak{p}_0}\left[\left(\int_{\Theta^2} \|\theta - \phi\|_{l^2}^2\ \psi_n(d\theta|\mathbf{X} = \mathbf{x})\ \delta_{\theta_0}(d\phi)\right)^{1/2}\right] = \\
&\mathbb{E}_{\otimes^n \mathfrak{p}_0}\left[\mathcal{W}_2^{\mathcal{P}(\Theta)}(\psi_n(\ \cdot \mid \mathbf{X}), \delta_{\theta_0})\right] = O(n^{-1/2})
\end{aligned}$$

where the last equality is due to Proposition 5.2. Moreover, by Theorem 5.1, we cay say the following

$$\mathbb{E}_{\psi_n \otimes \mathfrak{p}_0^n}\left[\mathbb{E}\left[\mathcal{W}_1^{\mathcal{P}(\mathbb{X})}(e_m^{(\mathbf{Z})}, \mathscr{T}(\theta)) \mid \theta\right]\right] = \begin{cases} O(m^{-1/2}) & \text{, if } d < 2, \\ O(m^{-1/2}\log(1+m)) & \text{, if } d = 2, \\ O(m^{-1/d}) & \text{, if } d > 2 \end{cases}$$

since $Z_1, ..., Z_m \mid \theta \overset{\text{i.i.d.}}{\sim} \mathscr{T}(\theta)$ and a discrete measure with finite support has q-moments finite for all $q \geq 1$. □

Suppose we let $m \asymp n$. In that case, we can reformulate the above result by saying that the Dirichlet-Multinomial mechanism is ϵ_n-informative with respect to the class $\text{Lip}_1(\mathbb{X})$, with the rate ϵ_n dependent on the dimensionality of the space $\mathbb{X} = \mathbb{R}^d$.
We also provide a result of consistency with respect to the KS metric.

Proposition 5.3. *Let $\alpha_i \geq 2$ for all $i = 1, ..., k$ with $k \geq 2$. We consider $\mathbb{X} \subset \mathbb{R}$. If we release $\mathbf{Z}$ from the m-step posterior predictive distribution of a Dirichlet-Multinomial model, then*

$$\mathbb{E}_{\nu_{n,m}}\left[\rho(e_m^{(\mathbf{Z})}, \mathfrak{p}_0)\right] = O\left(\sqrt{\frac{\log m}{m}}\right) + O(n^{-1/2})$$

Proof. We denote by $\theta_0 := (\theta_{0,1}, ..., \theta_{0,k})$ the true parameter vector, such that $\mathscr{T}(\theta_0) = \mathfrak{p}_0$. Let $\theta \sim \psi_n$, i.e. a parameter vector distributed as the posterior, by triangular inequality, we have

$$\mathbb{E}_{\nu_{n,m}}\left[\rho(e_m^{(\mathbf{Z})}, \mathfrak{p}_0)\right] \leq \mathbb{E}_{\nu_{n,m}}\left[\rho(e_m^{(\mathbf{Z})}, \mathscr{T}(\theta))\right] + \mathbb{E}_{\nu_{n,m}}\left[\rho(\mathscr{T}(\theta), \mathfrak{p}_0)\right] \tag{26}$$

Following the same procedure used in the proof of Theorem 5.2, we can show

$$\mathbb{E}_{\nu_{n,m}}\left[\rho(e_m^{(\mathbf{Z})}, \mathscr{T}(\theta))\right] = \mathbb{E}_{\psi_n \otimes \mathfrak{p}_0^n}\left[\mathbb{E}\left[\rho(e_m^{(\mathbf{Z})}, \mathscr{T}(\theta)) \mid \theta\right]\right].$$

since $Z_1, ..., Z_m \mid \theta \overset{\text{i.i.d.}}{\sim} \mathscr{T}(\theta)$, by Equation (21) we can say

$$\mathbb{E}_{\psi_n \otimes \mathfrak{p}_0^n}\left[\mathbb{E}\left[\rho(e_m^{(\mathbf{Z})}, \mathscr{T}(\theta)) \mid \theta\right]\right] = O\left(\sqrt{\frac{\log m}{m}}\right)$$

Concerning the second term in the sum of Equation (26), by triangular inequality, Cauchy-Schwarz inequality, and Hölder's inequality, we have

$$\begin{aligned}
&\mathbb{E}_{\nu_{n,m}}\left[\rho(\mathscr{T}(\theta), \mathfrak{p}_0)\right] = \mathbb{E}_{\mathfrak{p}_0^n}[\mathbb{E}_{\psi_n}\left[\rho(\mathscr{T}(\theta), \mathfrak{p}_0)\right]] = \\
&\mathbb{E}_{\mathfrak{p}_0^n}\left[\int_\Theta \sup_{t \in \mathbb{R}} \left| \sum_{i \,:\, a_i \leq t} \theta_i - \sum_{i \,:\, a_i \leq t} \theta_{0,i} \right| \psi(\mathrm{d}\theta \mid \mathbf{X} = \mathbf{x})\right] \leq \\
&\mathbb{E}_{\mathfrak{p}_0^n}\left[\int_\Theta \sum_{i=1}^k |\theta_i - \theta_{0,i}| \; \psi(d\theta \mid \mathbf{X} = \mathbf{x})\right] \leq \\
&2\sqrt{k-1}\; \mathbb{E}_{\mathfrak{p}_0^n}\left[\int_\Theta \|\theta - \theta_0\|_{l^2} \; \psi(d\theta \mid \mathbf{X})\right] \leq \\
&2\sqrt{k-1}\; \mathbb{E}_{\mathfrak{p}_0^n}\left[\mathcal{W}_2^{\mathcal{P}(\Theta)}(\psi_n(\,\cdot \mid \mathbf{X}), \delta_{\theta_0})\right] = O(n^{-1/2})
\end{aligned}$$

where the last equality is due to Proposition 5.2. □

We notice that, in the case of $\mathbb{X} = \mathbb{R}$, up to the logarithmic factor, the rate in 1-Wasserstein distance is equal to the rate in the KS metric. Lastly, we show the result in the case of the BNP mechanism with DP prior. We must emphasise that the privacy theorems do not give us an explicit constraint between m and n, so we report the rate as a function of both dimensions. As we have seen from the plots of the different values ϵ and δ, the number of released data must always be less than n to satisfy differential privacy. However, we are not able to say if the constraint on m is of the following form

$$m \leq D\; n^{1/b}$$

for some constant $D > 0$ and integer $b > 1$.

Theorem 5.3. *Let $(\mathbb{X}, \mathcal{X})$ be a totally bounded metric space of dimension d, and $H \in \mathcal{P}(\mathbb{X})$. If $\otimes^n \mathfrak{p}_0 \ll \mu_n$ and, for any $C \in \mathcal{X}$, we have*

$$H(C) + \mathfrak{p}_0(C) \leq K(diam(C))^{\eta d} \tag{27}$$

for some constants $K > 0$ and $\eta \in (0, 1]$, then, releasing $Z_1, ..., Z_m$ from the m-step posterior predictive distribution of a $\mathscr{D}(\theta, H)$ implies

$$\mathbb{E}_{\nu_{n,m}}\left[\mathcal{W}_1^{\mathcal{P}(\mathbb{X})}(e_m^{(\mathbf{Z})}, \mathfrak{p}_0)\right] = O(n^{-1/[d(2-\eta)+2]}) + \varepsilon_{n,1}(\mathbb{X}, \mathfrak{p}_0) + \varepsilon_{m,1}(\mathbb{X})$$

where $\varepsilon_{n,1}(\mathbb{X}, \mathfrak{p}_0)$ is the rate of convergence defined in Equation 22 and $\varepsilon_{m,1}(\mathbb{X})$ is the rate of convergence for the mean of the 1-Wasserstein distance between $e_m^{(\mathbf{Z})}$ and a random probability measure $\tilde{p} \sim \pi_n$, with π_n the posterior of a sample from a $\mathscr{D}(\theta, H)$.

We recall that, as explained in Section 5.3, in case of $\mathbb{X} \subset \mathbb{R}^d$, we have

$$\varepsilon_{n,1}(\mathbb{X}, \mathfrak{p}_0) \leq C\ M_q^{p/q}(\mathfrak{p}_0) \times \begin{cases} n^{-1/2} + n^{-(q-1)/q} & \text{if } d < 2 \text{ and } q \neq 2 \\ n^{-1/2}\log(1+n) + n^{-(q-1)/q} & \text{if } d = 2 \text{ and } q \neq 2 \\ n^{-1/d} + n^{-(q-1)/q} & \text{if } d \geq 2 \text{ and } q \neq d/(d-1) \end{cases}$$

where $q > 1$ is such that $M_q(\mathfrak{p}_0) = \int_{\mathbb{X}} |x|^q \mathfrak{p}_0(dx) < \infty$. Moreover, If we assume $\mathbb{X} \subset \mathbb{R}^d$ and totally bounded, we have (Fournier and Guillin (2015))

$$\varepsilon_{m,1}(\mathbb{X}) = \begin{cases} O(m^{-1/2}), & \text{if } d < 2 \\ O(m^{-1/2}\log(1+m)) & \text{if } d = 2 \\ O(m^{-1/d}), & \text{if } d > 2 \end{cases}$$

Lastly, we underline that the Assumption (27) is satisfied, for example, with $\eta = 1$ if we consider the space $([0,1]^d, \mathfrak{B}([0,1]^d))$ and both $\mathfrak{p}_0$ and H have continuous density on $[0,1]^d$.

Proof (Theorem 5.3). For every $f : \mathbb{X}^m \to [0, +\infty)$ measurable, by Fubini-Tonelli's theorem, the following holds:

$$\mathbb{E}_{\nu_{n,m}}[f(\mathbf{Z})] = \int_{\mathbb{X}^m \times \mathbb{X}^n} f(z_1, ..., z_m)\ Q_n(dz_1, ..., dz_m \mid \mathbf{x})\ \mathfrak{p}_0^n(d\mathbf{x}) =$$

$$\int_{\mathbb{X}^m \times \mathbb{X}^n} f(z_1, ..., z_m) \left(\int_{\mathcal{P}(\mathbb{X})} \tilde{p}^m(dz_j)\ \pi_n(d\tilde{p} \mid \mathbf{x})\right) \mathfrak{p}_0^n(d\mathbf{x}) =$$

$$\int_{\mathcal{P}(\mathbb{X}) \times \mathbb{X}^n} \mathbb{E}[f(\mathbf{Z}) \mid \tilde{p}\ ]\ \pi_n(d\tilde{p} \mid \mathbf{x})\ \mathfrak{p}_0^n(d\mathbf{x})$$

Taking $f(\mathbf{Z}) = \mathcal{W}_1(e_m^{(\mathbf{Z})}, \mathfrak{p}_0)$, by triangular inequality, we have

$$\mathcal{W}_1(e_m^{(\mathbf{Z})}, \mathfrak{p}_0) \leq \mathcal{W}_1(e_m^{(\mathbf{Z})}, \tilde{q}) + \mathcal{W}_1(\tilde{q}, \mathfrak{p}_0)$$

for any random probability measure $\tilde{q}$. It follows that

$$\mathbb{E}[\mathcal{W}_1(e_m^{(\mathbf{Z})}, \mathfrak{p}_0)] \leq \int_{\mathcal{P}(\mathbb{X}) \times \mathbb{X}^n} \mathbb{E}[\mathcal{W}_1(e_m^{(\mathbf{Z})}, \tilde{p}) \mid \tilde{p}\ ]\ \pi_n(d\tilde{p} \mid \mathbf{x})\ \mathfrak{p}_0^n(d\mathbf{x}) +$$
$$\int_{\mathcal{P}(\mathbb{X}) \times \mathbb{X}^n} \mathbb{E}[\mathcal{W}_1(\tilde{p}, \mathfrak{p}_0) \mid \tilde{p}\ ]\ \pi_n(d\tilde{p} \mid \mathbf{x})\ \mathfrak{p}_0^n(d\mathbf{x})$$

by the property of the conditional expectation $\mathbb{E}[\mathcal{W}_1(\tilde{p}, \mathfrak{p}_0) \mid \tilde{p}] = \mathcal{W}_1(\tilde{p}, \mathfrak{p}_0)$ and, since the only coupling measure between a delta measure and another probability measure is the independent coupling, we can write the following

$$\int_{\mathcal{P}(\mathbb{X}) \times \mathbb{X}^n} \mathcal{W}_1(\tilde{p}, \mathfrak{p}_0)\ \pi_n(d\tilde{p} \mid \mathbf{x})\ \mathfrak{p}_0^n(d\mathbf{x}) = \mathbb{E}_{\mathfrak{p}_0^n}[\mathcal{W}_1^{\mathcal{Q}(\mathbb{X})}(\tilde{p}, \mathfrak{p}_0)]$$

If we show that a random probability measure distributed as the posterior of a Bayesian model with Dirichlet Process prior has q-moments finite, for some integer $q > 1$, at least π_n almost surely, the proof is complete by applying Theorem 5.1 and Proposition 5.1. By Proposition 2.2, $\tilde{p} \sim \pi_n(\ \cdot \mid K_n = j, \mathbf{N}_n = (n_1, ..., n_j))$ implies

$$\tilde{p} \stackrel{d}{=} \sum_{i=1}^{j} W_i \delta_{X_i^*} + R_j F_j$$

where $(W_1, ..., W_j, R_j) \sim \text{Dir}(n_1, ..., n_j, \theta)$, independently of the random distribution F_j, which has $\mathscr{D}(\theta, H)$ distribution. Moreover, by Stick-Breaking construction of the DP (Theorem 2.4 with $\sigma = 0$), we have

$$F_j = \sum_{i=1}^{\infty} Y_i \delta_{\Phi_i}$$

where $\Phi_i \overset{\text{i.i.d.}}{\sim} H$ and

$$Y_1 = V_1, \quad Y_l = V_l \prod_{j=1}^{l-1} (1 - V_j), \quad l \geq 2,$$

with $\{V_i\}_{i \geq 1}$ sequence of independent random variables taking values in $[0, 1]$ and distributed as a Beta$(1, \theta)$. The following holds

$$\int_{\mathbb{X}} |x|^q \, \tilde{p}(dx) = \sum_{i=1}^{j} W_i |X_i^*|^q + R_j \sum_{i=1}^{\infty} Y_i |\Phi_i|^q \leq 2R^q, \qquad \pi_n - a.s.$$

where $R > 0$ is the radius of the ball that contains the space $\mathbb{X}$, which exists because $\mathbb{X}$ is totally bounded. Since we have finite moments for any $q > 1$, we can neglect the terms involving q in the bounds of $\varepsilon_{m,1}(\mathbb{X})$.

□

6. Mechanism's utility

We saw in Section 5 that our release mechanism is consistent, in the sense that assuming a generating distribution of the data $\mathfrak{p}_0$, the distribution of the data we release asymptotically converges to $\mathfrak{p}_0$. Instead, in this section, we show that releasing data by sampling from the posterior predictive distributions allows us to perform posterior inference on private data straightforwardly, unlike other mechanisms in the literature used for generating synthetic data. Hence, this section shows that our mechanism is useful to the Bayesian statistician who wants to perform inference on private data.

6.1. Bayesian utility

Unlike other known privacy-satisfying mechanisms, our approach allows for easy expression of the posterior predictive distribution. Let us consider the case where one wants to make a posterior inference on private data using a Bayesian model but only has access to the released data; we obtain the following model

$$\begin{aligned} Z_1, ..., Z_m \mid X_1, ..., X_n &\sim Q_n(\,\cdot \mid \mathbf{X}) \\ X_1, ..., X_n \mid \tilde{p} &\overset{\text{i.i.d.}}{\sim} \tilde{p} \\ \tilde{p} &\sim \pi \end{aligned} \tag{28}$$

The privacy mechanisms prevalent in the literature assume that the law of $\mathbf{Z}$ is independent of $\tilde{p}$ given the private observations. This means that, marginalizing out $\mathbf{X}$, model (28) is equivalent to

$$\begin{aligned} Z_1, ..., Z_m \mid \tilde{p} &\sim \int_{\mathbb{X}^n} Q_n(\,\cdot \mid \mathbf{X} = \mathbf{x}) \, \tilde{p}^n(d\mathbf{x}) \\ \tilde{p} &\sim \pi \end{aligned} \tag{29}$$

As explained in Section 2.2, typical examples consist of adding noise to the private observations, i.e. $Z_i = X_i + \epsilon_i$ with $(\epsilon_i)_{i=1}^n$ i.i.d. from a distribution with zero mean and density g. In this case, model (29) can be rewritten as

$$\begin{aligned} Z_1, ..., Z_m \mid \tilde{p} &\overset{\text{ind.}}{\sim} \int_{\mathbb{X}} g(\cdot - x_i) \, \tilde{p}(dx_i) \\ \tilde{p} &\sim \pi \end{aligned}$$

which corresponds to a mixture model (Lo (1984)) with a kernel equal to the centred distribution of the noise. Posterior inference in this model type is possible, and many MCMC algorithms exist in the literature; for example, the work of Neal (2000) considers the case where $\tilde{p}$ is a random probability measure almost surely discrete, while Beraha et al. (2023) take in account when $\tilde{p}$ is a mixture model itself. However, obtaining an expression of the posterior predictive distribution in closed form for a mixture model is impossible.
In our case, the posterior predictive distribution of the model (28) has the same form as the one obtained having

access to the private data. The only difference is that we have to replace $\mathbf{X}$ by $\mathbf{Z}$. By the exchangeability assumption, sampling $\mathbf{Z}$ from the m-step posterior predictive distribution is equivalent to assume

$$\begin{aligned} X_1, ..., X_n, Z_1, ..., Z_m \mid \tilde{p} &\overset{\text{i.i.d.}}{\sim} \tilde{p} \\ \tilde{p} &\sim \pi \end{aligned} \tag{30}$$

We denote by π_{n+m} the posterior of the above model based on the full sample $(\mathbf{X}, \mathbf{Z})$. By Fubini-Tonelli's theorem, the posterior predictive distribution obtained marginalizing out the private observations $\mathbf{X}$, i.e. the probability measure defined as

$$\eta(B) = \int_{\mathbb{X}^n} \mathbb{P}[B|\mathbf{X}, \mathbf{Z}] \; \mathbb{P}[\; \mathbf{X} \in d\mathbf{x} \mid \mathbf{Z}]$$

is equal to the following

$$\mu(B) = \int_{\mathcal{P}(\mathbb{X})} \tilde{p}(B) \; \pi_m(d\tilde{p} \mid \mathbf{Z})$$

for all $B \in \mathcal{X}$, where $\pi_m(\,\cdot \mid \mathbf{Z})$ is the posterior of

$$\begin{aligned} Z_1, ..., Z_m \mid \tilde{p} &\overset{\text{i.i.d.}}{\sim} \tilde{p} \\ \tilde{p} &\sim \pi \end{aligned}$$

Indeed, for all $B \in \mathcal{X}$, we have

$$\begin{aligned} \eta(B) = &\int_{\mathbb{X}^n} \left(\int_{\mathcal{P}(\mathbb{X})} \tilde{p}(B) \; \pi_{n+m}(d\tilde{p} \mid \mathbf{X}, \mathbf{Z}) \right) \; \mathbb{P}[\; \mathbf{X} \in d\mathbf{x} \mid \mathbf{Z}] = \\ &\int_{\mathcal{P}(\mathbb{X})} \tilde{p}(B) \left(\int_{\mathbb{X}^n} \pi_{n+m}(d\tilde{p} \mid \mathbf{X}, \mathbf{Z}) \; \mathbb{P}[\; \mathbf{X} \in d\mathbf{x} \mid \mathbf{Z}] \right) = \\ &\int_{\mathcal{P}(\mathbb{X})} \tilde{p}(B) \left(\int_{\mathbb{X}^n} \mathbb{P}[\tilde{p} \in d\tilde{p} \mid \mathbf{X}, \; \mathbf{Z}] \; \frac{\mathbb{P}[\; \mathbf{X} \in d\mathbf{x}, \; \mathbf{Z} \in d\mathbf{z}]}{\mathbb{P}[\mathbf{Z} \in d\mathbf{z}]} \right) = \\ &\int_{\mathcal{P}(\mathbb{X})} \tilde{p}(B) \left(\int_{\mathbb{X}^n} \frac{\mathbb{P}[\tilde{p} \in d\tilde{p}, \; \mathbf{X} \in d\mathbf{x}, \; \mathbf{Z} \in d\mathbf{z}]}{\mathbb{P}[\mathbf{Z} \in d\mathbf{z}]} \right) = \mu(B). \end{aligned}$$

This shows a substantial advantage of our release mechanism as opposed to previously considered ones. Indeed, inference on perturbed data can be cumbersome, owing to the use of MCMC algorithms that typically need to deal with huge parameter spaces as they introduce the private observations as latent variables and update them as part of the algorithm. By contrast, our mechanism allows us to compute the posterior predictive distribution in closed form, from which Bayesian estimates of quantities of interest (e.g., the mean, quantiles, distribution function) are easily obtained as functionals of such distribution either in an analytical form or via straightforward Monte Carlo (and not MCMC) integration.

6.2. Parametric model

The same result shown in the previous section on posterior predictive distribution also applies in the parametric case. Furthermore, for dominated models, it is easy to conclude that the posterior density based on the released data also has an easy expression to compute. In particular, the posterior density will be the same as that based on private data, provided that $\mathbf{X}$ is replaced by $\mathbf{Z}$.
As shown in Section 2.3.1, the posterior $\psi_n(\,\cdot \mid \mathbf{X})$ has density given by

$$g(\theta|\mathbf{x}) = \frac{\prod_{i=1}^n f_\theta(x_i) \; q(\theta)}{m(\mathbf{x})}$$

where $q(\cdot)$ is the density of the prior distribution ϕ and $m(\mathbf{x})$ denotes the marginal density of $\mathbf{X}$ defined as follows

$$m(\mathbf{x}) := \int_\Theta \prod_{i=1}^n f_\theta(x_i) \; q(\theta) \; d\theta.$$

Let us consider the parametric version of the model (28)

$$\begin{aligned} Z_1, ..., Z_m \mid X_1, ..., X_n &\sim Q_n(\,\cdot \mid \mathbf{X}) \\ X_1, ..., X_n \mid \theta &\overset{\text{i.i.d.}}{\sim} P_\theta \\ \theta &\sim \phi \end{aligned} \tag{31}$$

Under this model, if we assume that Q_n corresponds to m-step posterior predictive distribution, we have the following posterior density

$$g(\theta|\mathbf{z}) = \frac{\prod_{i=1}^{m} f_\theta(z_i)\ q(\theta)}{m(\mathbf{z})}$$

where

$$m(\mathbf{z}) := \int_\Theta \prod_{i=1}^{m} f_\theta(z_i)\ q(\theta)\ d\theta$$

Indeed, the exchangeability assumption implies

$$\begin{aligned} Z_1, ..., Z_m, X_1, ..., X_n \mid \theta &\overset{\text{i.i.d.}}{\sim} P_\theta \\ \theta &\sim \phi \end{aligned} \tag{32}$$

We denote by $m(\mathbf{x}|\mathbf{z})$ the conditional density of $\mathbf{X} \mid \mathbf{Z}$ and by $m(\mathbf{z}, \mathbf{x})$ the joint density of the random vector $(\mathbf{X}, \mathbf{Z})$. Moreover, let $g(\theta|\mathbf{x}, \mathbf{z})$ be the posterior of model 32. We have that the posterior based (only) on the released data is the following

$$\begin{aligned} &\int_{\mathbb{X}^n} g(\theta|\mathbf{x}, \mathbf{z})\ m(\mathbf{x}|\mathbf{z})\ d\mathbf{x} \\ &\quad = \int_{\mathbb{X}^n} \frac{\prod_{i=1}^{n} f_\theta(x_i) \prod_{i=1}^{m} f_\theta(z_i)\ q(\theta)}{m(\mathbf{x}, \mathbf{z})} \frac{m(\mathbf{x}, \mathbf{z})}{m(\mathbf{z})}\ \mathrm{d}\mathbf{x} \\ &\quad = \frac{\prod_{i=1}^{m} f_\theta(z_i)\ q(\theta)}{m(\mathbf{z})} \int_{\mathbb{X}^n} \prod_{i=1}^{n} f_\theta(x_i)\ d\mathbf{x} = \frac{\prod_{i=1}^{m} f_\theta(z_i)\ q(\theta)}{m(\mathbf{z})} \end{aligned}$$

In contrast, assume now a general releasing mechanism Q_n such that the distribution of $\mathbf{Z}$ conditioned to $\mathbf{X}$ is independent of the parameter θ. Marginalizing out $\mathbf{X}$, we obtain the following equivalent model

$$\begin{aligned} Z_1, ..., Z_m \mid \theta &\sim \int_{\mathbb{X}^n} Q_n(\ \cdot \mid \mathbf{X} = \mathbf{x})\ P_\theta^n(d\mathbf{x}) \\ \theta &\sim \phi \end{aligned} \tag{33}$$

and, even if we consider a mechanism $Q_n(\ \cdot \mid \mathbf{X} = \mathbf{x}) = \otimes_{i=1}^{n} Q_i(\ \cdot \mid X_i = x_i)$ such that Q_i has density $q_i(\ \cdot \mid x_i)$, which corresponds to $Z_1, ..., Z_m$ conditionally independent with density

$$h(z_i) = \int_{\mathbb{X}} q_i(z_i|x_i)\ f_\theta(x_i)\ dx_i$$

The posterior density of (33) will be

$$g(\theta \mid \mathbf{z}) = \frac{\prod_{i=1}^{n} h(z_i)\ q(\theta)}{\int_\Theta \prod_{i=1}^{n} h(z_i)\ q(\theta)\ d\theta} \tag{34}$$

and obtaining samples from it is a hard task. Indeed, except for rare cases, for example, adding Gaussian noise to the private data and assuming a Gaussian density for the Bayesian model, $h(\cdot)$ cannot be written in closed form. Hence, to sample from (34), we need to replace h with its estimate. This problem is commonly solved using MCMC algorithms with a pseudo-marginal approach, as detailed by Andrieu and Roberts (2009).

6.3. Discovery Probability

A common problem in Statistics and Computer Science consists of estimating the missing mass or discovery probability, i.e. estimating the probability of discovering a new type at the $(n+1)$-th draw, given an n samples from a population of individuals belonging to different types with unknown proportions (Ayed et al. (2018)). Here, we show that releasing synthetic data from the m-step posterior predictive distribution still allows one to obtain a Bayesian Nonparametric estimator of the missing mass.
Let $K_1^{(n+m)}$ the additional number of unique values observed at the $(n+m+1)$-th draw. We can write the discovery probability as

$$\mathbb{P}[K_1^{(n+m)} = 1 \mid K_n = j, K_m^{(n)} = k]$$

where K_n is the number of unique values observed in $(X_1, ..., X_n)$ and $K_m^{(n)}$ denotes the additional number of new unique values observed in the synthetic data $(Z_1, ..., Z_m)$. We assume $(X_1, ..., X_n)$ a sample from a $\mathscr{PY}(\sigma, \theta, H)$ with $\sigma \in (0, 1)$. A direct application of the EPPF leads to the following probability

$$\mathbb{P}[K_1^{(n+m)} = 1 \mid K_n = j, K_m^{(n)} = k] = \frac{\theta + (j+k)\sigma}{\theta + n + m}$$

When we release $\mathbf{Z}$, we do not have K_n because this information is contained in the private data, so we need to estimate the following random variable

$$D = \mathbb{P}[K_1^{(n+m)} = 1 \mid K_n, K_m^{(n)} = k]$$

The Bayes estimate of D, with respect to the squared loss functions, is given by the expected value

$$\begin{aligned}\hat{D} =& \sum_{j=1}^{n} \mathbb{P}[K_1^{(n+m)} = 1 \mid K_n = j, K_m^{(n)} = k]\ \mathbb{P}[K_n = j \mid K_m^{(n)} = k] = \\ & \sum_{j=1}^{n} \mathbb{P}[K_1^{(n+m)} = 1 \mid K_n = j, K_m^{(n)} = k]\ \frac{\mathbb{P}[K_m^{(n)} = k \mid K_n = j]\mathbb{P}[K_n = j]}{\mathbb{P}[K_m^{(n)} = k]} = \\ & \sum_{j=1}^{n} \frac{\theta + (j+k)\sigma}{\theta + n + m} \frac{\mathbb{P}[K_m^{(n)} = k \mid K_n = j]\mathbb{P}[K_n = j]}{\sum_{j=1}^{n} \mathbb{P}[K_m^{(n)} = k \mid K_n = j]\mathbb{P}[K_n = j]}\end{aligned}$$

By Lijoi et al. (2007b), we have

$$\mathbb{P}[K_m^{(n)} = k \mid K_n = j] = \frac{(\theta+1)_{n-1}}{(\theta+1)_{n+m-1}} \frac{1}{\sigma^k} \prod_{i=j}^{j+k-1} (\theta + i\sigma)\ \mathscr{C}(m, k; \sigma, -n + j\sigma)$$

where $\mathscr{C}(a, b; c, \gamma)$ denotes the noncentral generalized factorial coefficient, defined as follows

$$\mathscr{C}(a, b; c, \gamma) := \frac{1}{b!} \sum_{j=0}^{b} (-1)^j \binom{b}{j} (-cj - \gamma)_a \tag{35}$$

Moreover, Lijoi et al. (2007b) provides the law of K_n,

$$\mathbb{P}[K_n = j] = \frac{\prod_{i=1}^{j-1}(\theta + i\sigma)}{\sigma^j (\theta+1)_{n-1}} \mathscr{C}(n, j; \sigma)$$

with $\mathscr{C}(a, b; c)$ the generalized factorial coefficients defined as (35) by letting $\gamma = 0$. Hence we obtain

$$\begin{aligned}& \mathbb{P}[K_m^{(n)} = k \mid K_n = j]\ \mathbb{P}[K_n = j] = \\ & \frac{1}{(\theta+1)_{n+m-1}\sigma} \prod_{i=1}^{j+k-1} \left(\frac{\theta}{\sigma} + i\right)\ \mathscr{C}(n, j; \sigma)\ \mathscr{C}(m, k; \sigma, -n + j\sigma) = \\ & \frac{1}{(\theta+1)_{n+m-1}\sigma}\ (\theta/\sigma + 1)_{j+k-1}\ \mathscr{C}(n, j; \sigma)\ \mathscr{C}(m, k; \sigma, -n + j\sigma)\end{aligned}$$

Obtaining the following estimator

$$\begin{aligned}\hat{D} =& \frac{1}{\theta + n + m} \left(\sum_{j=1}^{n} (\theta + (j+k)\sigma) \frac{1}{(\theta+1)_{n+m-1}\sigma}\ (\theta/\sigma + 1)_{j+k-1}\ \mathscr{C}(n, j; \sigma)\ \mathscr{C}(m, k; \sigma, -n + j\sigma) \right) \times \\ & \frac{1}{\sum_{j=1}^{n} \frac{1}{(\theta+1)_{n+m-1}\sigma}\ (\theta/\sigma + 1)_{j+k-1}\ \mathscr{C}(n, j; \sigma)\ \mathscr{C}(m, k; \sigma, -n + j\sigma)} = \\ & \frac{1}{\theta + n + m} \left(\theta + k\sigma + \sigma \frac{\sum_{j=1}^{n} j\ (\theta/\sigma + 1)_{j+k-1}\ \mathscr{C}(n, j; \sigma)\ \mathscr{C}(m, k; \sigma, -n + j\sigma)}{\sum_{j=1}^{n} (\theta/\sigma + 1)_{j+k-1}\ \mathscr{C}(n, j; \sigma)\ \mathscr{C}(m, k; \sigma, -n + j\sigma)} \right)\end{aligned}$$

If we define a random variable Y with support equal to the set $\{1, ..., n\}$ such that

$$\mathbb{P}[Y = j] \propto (\theta/\sigma + 1)_{j+k-1}\ \mathscr{C}(n, j; \sigma)\ \mathscr{C}(m, k; \sigma, -n + j\sigma)$$

where $\propto$ denotes the symbol of proportionality, then we can write

$$\hat{D} = \frac{\theta + k\sigma + \sigma\mathbb{E}[Y]}{\theta + n + m}.$$

7. Real Application

We present here two real-world applications showcasing the effectiveness of our proposed release mechanism. In particular, we show that the 1-Wasserstein distance between the empirical measure based on synthetic data released using the BNP mechanism and that based on private data decreases as $m \to \infty$. Moreover, at the same level of ϵ and thus privacy assurance, we show that the BNP mechanism beats all other mechanisms described in Section 3.3 in releasing data with statistics close to those of private data: mean, standard deviation, percentiles. We also compare the different mechanisms considered using as metric the L^2 distance between the kernel density estimator based on the histogram of the data released and the kernel density estimator based on the histogram of the private data.

7.1. Invalsi scores

We consider a dataset composed of $n = 39374$ observations, each representing the Invalsi test score in the interval $[0, 10]$ obtained by a student of an Italian school. Grades are expressed using two significant digits so that the number of unique grades in n observations is surely less than 100. We rescaled the data in the $[0, 1]$ range by dividing the scores by 10. We model the private data with a DP prior with a mean measure equal to the uniform distribution in $[0, 1]$. We fix θ using an Empirical Bayes method, i.e. taking the value $\hat{\theta}$ that maximises the density of the observations. In particular, let $k = 95$ the number of possible scores appearing within the dataset, we take $\hat{\theta} = 11.6257$ as follows

$$\hat{\theta} = \arg\max_{\theta} \frac{\theta^{k-1}}{(\theta+1)_{n-1}} \prod_{i=1}^{k} (n_i - 1)!$$

For a number of iterations equal to 100: we fix the seed, we generate synthetic data $Z_1, ..., Z_m$ for $m = 100, 200, 1000, 2000, 5000, 8000, 10000, 15000, 18000, 2000, 30000$ with the BNP mechanism and we compute the 1-Wasserstein distance between the empirical measure based on $\mathbf{Z}$ and the empirical measure based on the real dataset $\mathbf{X}$. Figure 5 represents the mean of the computed 1-Wasserstein distance and a 99% confidence band, varying the number of released data. This allows us to understand what a valid release number could be for the distribution of private data to be preserved in the synthetic data.

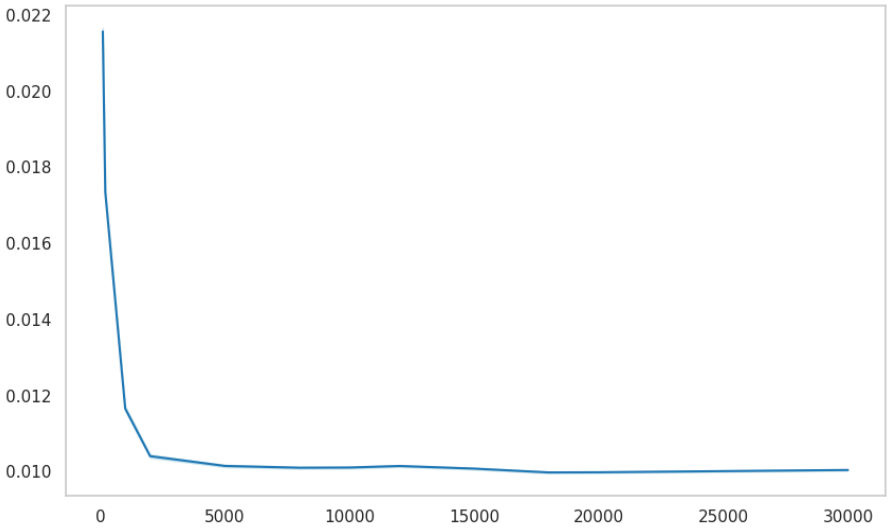
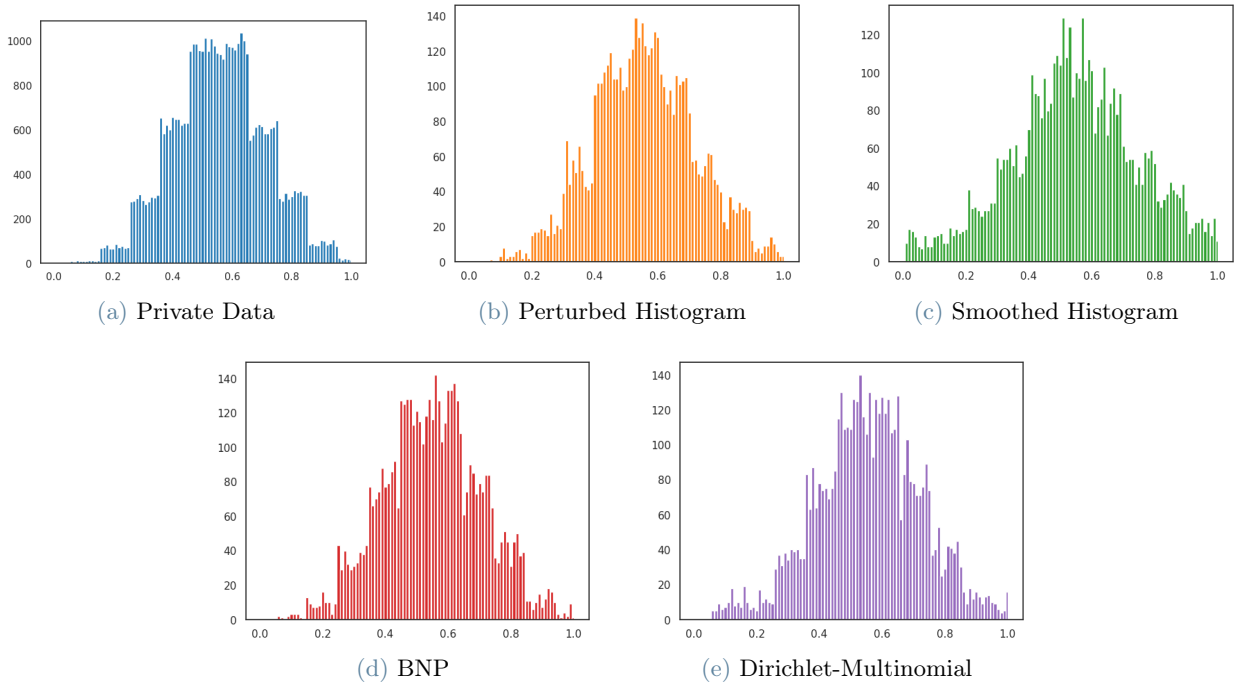


Figure 5: Average 1-Wasserstein distance varying m.

The average Wasserstein distance decreases as the number of released data increases, but after a certain m, it stabilises as the number of private data is fixed. The confidence band is very narrow since the variability of the Wasserstein distance computed by fixing m and varying the seed is very low. To compare the BNP mechanism with those proposed in Section 3.3, we choose $m = 5000$ as the Wasserstein distance stabilises around 0.01 from this number of released data. Since the other methods satisfy ϵ differential privacy, to make the comparison fair, we choose $\epsilon = 4.2$ which allows us to obtain a value of $\tilde{\delta}(\epsilon) = 0$ for the BNP mechanism evaluated via Monte Carlo using 20000 iterations. For the perturbed and smoothed histogram mechanisms, we set a number of buckets equal to 10, one for each grade range $0-1, 1-2, ..., 9-10$. We choose the other parameters equal to the smallest values that allow the mechanisms to satisfy differential privacy with $\epsilon = 4.2$. In particular, we

take a smoothing parameter $\sigma = 0.2321$ in the smoothed histogram mechanism, the variance of the Laplace noise equal to 0.456 for the perturbed histogram mechanism and Dirichlet prior parameters $(\alpha_1, ..., \alpha_{95})$ with $\alpha_j = \max(0.0001, m/(e^\epsilon - 1) - C_j)$ for all $j = 1, ..., 95$. Figure 6 represents the histograms obtained from the real data and the synthetic ones generated through the different mechanisms, taking a number of histogram bins equal to 100, one for every possible score.



(a) Private Data (b) Perturbed Histogram (c) Smoothed Histogram

(d) BNP (e) Dirichlet-Multinomial

Figure 6: Histograms with 100 bins.

Figure 7 represents the boxplots obtained from the real data and the synthetic ones generated through the different mechanisms.

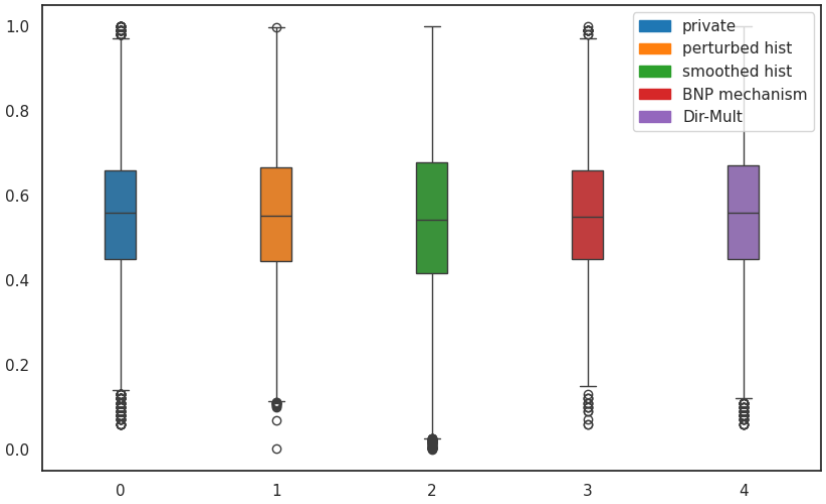


Figure 7: Boxplots: boxplot of the private data (blue) and boxplot of the synthetic data obtained from the perturbed histogram mechanism (orange), the smoothed histogram mechanism (green), the BNP mechanism (red) and the Dirichlet-Multinomial mechanism (purple).

To make the comparison quantitative, we also calculated the error in norm L^2 between the histogram based on the private data and that obtained from the synthetic data, all smoothed using a kernel density estimator. Table 7 contains the value of obtained L^2 distances along with the mean, standard deviation and 25%, 50% and 75% percentiles of the private dataset and all the generated ones.

	Mean	St.dev	25%	50%	75%	L^2-dist
Private Data	0.556774	0.155934	0.450000	0.560000	0.660000	-
Perturbed Hist	0.556711	0.161562	0.444715	0.552957	0.665898	0.037517
Smoothed Hist	0.544161	0.202949	0.417069	0.543319	0.677291	0.241838
BNP	0.550162	0.158369	0.450000	0.550000	0.660000	0.036688
Dir-Multinomial	0.554592	0.169220	0.450000	0.560000	0.670000	0.067266

Table 7: Summary statistics for the private data and the synthetic ones.

As can be seen from the histogram of the private data, the observations are approximately uniformly distributed in each of the 10 sub-intervals $[0, 0.1], [0.1, 0.2], ..., [0.9, 1.0]$ and this benefits Wasserman's methods by having selected a number of bins equal to 10. Nevertheless, the BNP mechanism achieves the smallest error in norm L^2. The Dirichlet-Multinomial model also performs well because the number of data for each category is substantial, so there is no need to force the prior parameters to take large values. In addition, the Dirichlet-Multinomial is the only mechanism that necessarily generates synthetic data equal to one of the 95 scores within the private data. In contrast, the other mechanisms can also generate fully synthetic data, i.e. data different from all the scores already observed in $\mathbf{X}$.

7.2. US Income Data

We consider the 2021 ACS census data publicly available at `https://www.census.gov/programs-surveys/acs/data/experimental-data/2020-1-year-pums.html`. Specifically, we consider a random subsample of 53000 data of the PINCP variable that represents the personal income of the survey responders and is restricted to the citizens of California. The dataset contains missing values and negative income, which we removed for simplicity, resulting in 44816 observations. Since the data have a wide range, we scale the data logarithmically and then apply min-max scaling to obtain values between $[0, 1]$.
The number of possible incomes that recur within the private data is 3263, of which 1436 are present only once. Since Theorem 4.1 requires that private data have no value with frequency one, we double these values within the data. Observe that this procedure creates a bias in the inference, but our numerical experiment shows that this is effectively negligible. The other mechanisms do not require this condition to satisfy differential privacy, so the synthetic data generated by the other mechanisms operate on the original data. When comparing the different mechanisms, we always refer to the original data (without repetitions of the unique values) as the ground truth.
To generate data with the BNP mechanism, we consider a DP prior with $\theta = 2$ and mean measure equal to the uniform distribution in the interval $[0, 1]$. Using an Empirical Bayes approach, we obtain an estimated $\hat{\theta} = 500$. However, such high values of θ require ϵ greater than 10 for the estimated $\tilde{\delta}(\epsilon)$ to be equal to 0, which are too large for common applications of privacy mechanisms, even for a small number of released data, such as $m = 200, 500, 1000$. In contrast, fixing $\theta = 2$, $\epsilon = 2.5$ and $m < 3000$, the Monte Carlo estimate based on 10000 Monte Carlo iterations of $\tilde{\delta}(\epsilon)$ is equal to zero.
Using the same procedure as described in Section 7.1, we compute the average 1-Wasserstein distance between the empirical measure based on $\mathbf{Z}$ and the empirical measure based on $\mathbf{X}$, varying the number of released data. Figure 8 shows the results obtained. As the elbow corresponds to a number of released data equal to $m = 2000$, we choose this value to compare the various release mechanisms.

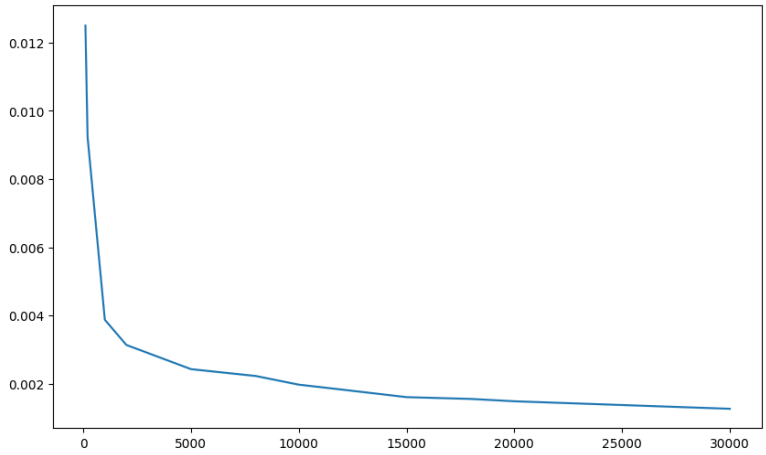
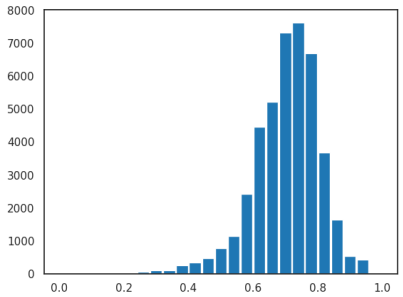
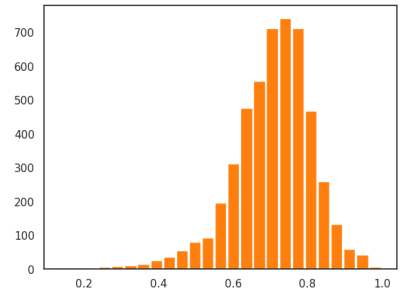


Figure 8: Average 1-Wasserstein distance varying m.

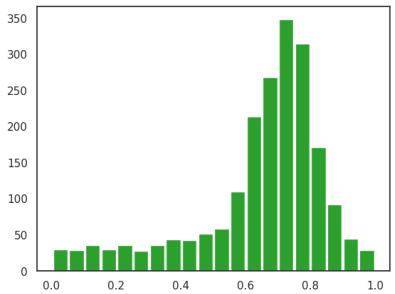
For the mechanism of the smoothed histogram and that of the perturbed histogram, we take a number of bins $k = 25$ and partition the interval $[0, 1]$ into subintervals of length $h = 1/k$. We take the remaining parameters equal to the minimum so that the various conditions of ϵ-differential privacy with $\epsilon = 2.5$ are satisfied.
Figure 9 represents the histograms obtained from the private data and the synthetic ones generated through the different mechanisms, taking a number of histogram bins equal to 25.



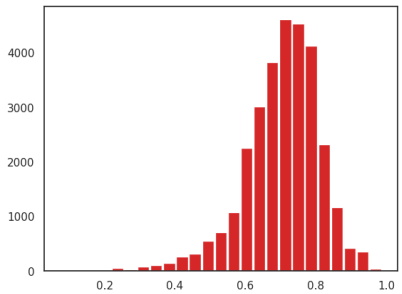
(a) Private Data



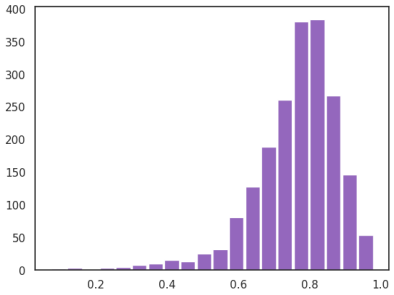
(b) Perturbed Histogram



(c) Smoothed Histogram



(d) BNP



(e) Dirichlet-Multinomial

Figure 9: Histograms with 25 bins.

Figure 10 represents the boxplots obtained from the real data and the synthetic ones generated through the different mechanisms.

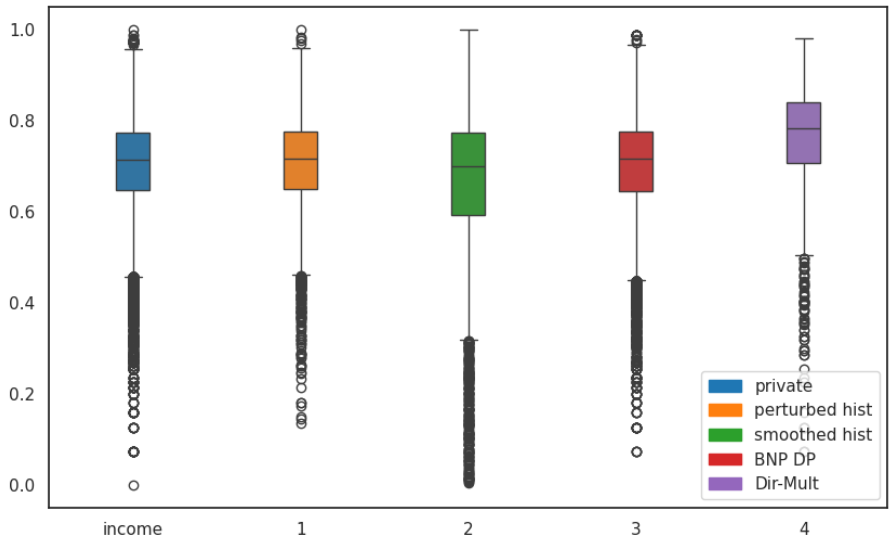


Figure 10: Boxplots: boxplot of the private data (blue) and boxplot of the synthetic data obtained from the perturbed histogram mechanism (orange), the smoothed histogram mechanism (green), the BNP mechanism (red) and the Dirichlet-Multinomial mechanism (purple).

We also calculated the error in norm L^2 between the histogram based on the private data and that obtained from the synthetic data, all smoothed using a kernel density estimator. Table 8 contains the value of obtained L^2 distances along with the mean, standard deviation and 25%, 50% and 75% percentiles of the private dataset and all the generated ones.

	Mean	**St.dev**	**25%**	**50%**	**75%**	**L^2-dist**
Private Data	0.703467	0.103677	0.647091	0.714881	0.773849	-
Perturbed Hist	0.706685	0.105617	0.649485	0.717137	0.775253	0.049465
Smoothed Hist	0.645578	0.201736	0.592295	0.699857	0.774461	0.574438
BNP	0.703734	0.106889	0.644440	0.715106	0.774801	0.032337
Dir-Multinomial	0.762070	0.117283	0.705627	0.783804	0.840608	0.599707

Table 8: Summary statistics for the private data and the synthetic ones.

Contrary to the situation in Section 7.1, note that in this case, the Dirichlet-Multinomial mechanism performs significantly worse than our mechanism. This is likely explained by the fact that there are several categories of income with low counts. The BNP mechanism has the least L^2 distance compared to the private data. The perturbed histogram mechanism also has a very low L^2 distance close to that of the BNP mechanism. The synthetic data generated through these two mechanisms also preserve statistics as the mean and the percentiles. It should be noted, however, that our mechanism starts from data in which the unique values have been doubled and not the original private data.

8. Conclusions

In this work, we proposed a new synthetic data release mechanism based on a Bayesian Nonparametric model. Under the Differential privacy framework, we addressed the problem of showing which assumptions are required by such a mechanism to have privacy guarantees. It is important to note that all mechanisms in the literature have multiple parameters available to adjust the level of privacy required; for example, in the case of release from the predictive of a parametric Bayesian model, it is possible to increase the privacy guarantee by choosing a prior that is more informative. In our case, on the other hand, the model assumed for private data is nonparametric and, thus, tightly data-driven even if the Pitman-Yor hyperparameters are chosen inappropriately. Nevertheless,

we have seen that adjusting the number of data released makes it possible to guarantee privacy at a level specific to the private dataset from which we start.
On the other hand, we have shown that the advantages in terms of the statistical utility of the synthetic data generated by our mechanism are various. In a frequentist context, we can ensure that as the number of data released and that of private data increase, the empirical distribution based on the synthetic data approaches the actual data-generating distribution. Whereas, if the purpose is to analyze the data assuming a Bayesian approach, our mechanism allows us to easily obtain the expression of the one-step posterior predictive based on the synthetic data.
This work opens the door to several interesting research questions:

- What is the best definition of privacy in the literature, even outside the framework of differential privacy, that best fits the mechanism we introduced?
- Is it possible to estimate privacy parameters analytically using an approximation for the law of the vector of frequencies $\mathbf{S}_{m-L_m^{(n)}}$ and $\mathbf{S}_{L_m^{(n)}}$?
- Can we improve the shown convergence rates, and can these rates achieve the minimax rate?
- Is it possible to obtain a convergence rate even in the more general case of the Pitman-Yor process?
- Are there other Bayesian Nonparametric models that allow for greater privacy guarantee?

In particular, regarding the last question, we think it might be interesting to consider a Pólya tree (PT) prior (Mauldin et al. (1992)) and release data sampling from the m-steps posterior predictive. In this case, too, the density of the m-steps predictive has an analytic expression in closed form that allows us to test the ratio of the densities as required by the definition of differential privacy. Furthermore, this density depends on infinite parameters that characterize the PT prior, giving more flexibility for selecting the required level of privacy.

References

J. M. Abowd, R. Ashmead, R. Cumings-Menon, S. Garfinkel, M. Heineck, C. Heiss, R. Johns, D. Kifer, P. Leclerc, A. Machanavajjhala, et al. The 2020 census disclosure avoidance system topdown algorithm. *arXiv*, 2204.08986, 2022.

Luigi Ambrosio and Nicola Gigli. *A User's Guide to Optimal Transport*, pages 1–155. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-32160-3. doi: 10.1007/978-3-642-32160-3_1. URL `https://doi.org/10.1007/978-3-642-32160-3_1`.

Christophe Andrieu and Gareth O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697 – 725, 2009. doi: 10.1214/07-AOS574. URL `https://doi.org/10.1214/07-AOS574`.

Apple's Differential Privacy Team. Learning with privacy at scale. *Apple Machine Learning Journal*, 1, 2017.

Fadhel Ayed, Marco Battiston, Federico Camerlenghi, and Stefano Favaro. On consistent estimation of the missing mass. *arXiv: Statistics Theory*, 2018. URL `https://api.semanticscholar.org/CorpusID:88523838`.

Mario Beraha, Stefano Favaro, and Vinayak Rao. Mcmc for bayesian nonparametric mixture modeling under differential privacy, 2023.

G. Bernstein and D. R. Sheldon. Differentially private bayesian linear regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Cristina Butucea, Amandine Dubois, Martin Kroll, and Adrien Saumard. Local differential privacy: Elbow effect in optimal density estimation and adaptation over Besov ellipsoids. *Bernoulli*, 26(3):1727 – 1764, 2020. doi: 10.3150/19-BEJ1165. URL `https://doi.org/10.3150/19-BEJ1165`.

Federico Camerlenghi, Emanuele Dolera, Stefano Favaro, and Edoardo Mainini. Wasserstein posterior contraction rates in non-dominated bayesian nonparametric models, 2022.

C. A. Charalambides. *Combinatorial Methods in Discrete Distributions*. Wiley Interscience, New York, 2005. ISBN 0-471-68027-3.

Konstantinos Chatzikokolakis, Miguel E. Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. Broadening the scope of differential privacy using metrics. In Emiliano De Cristofaro and Matthew Wright, editors, *Privacy Enhancing Technologies*, pages 82–102, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-39077-7.

Bruno de Finetti. La prévision : ses lois logiques, ses sources subjectives. 1937. URL https://api.semanticscholar.org/CorpusID:194350723.

Christos Dimitrakakis, Blaine Nelson, Zuhe Zhang, Aikaterini Mitrokotsa, and Benjamin I. P. Rubinstein. Differential privacy for bayesian inference through posterior sampling. *J. Mach. Learn. Res.*, 18(1):343–381, jan 2017. ISSN 1532-4435.

B. Ding, J. Kulkarni, and S. Yekhanin. Collecting telemetry data privately. In *Advances in Neural Information Processing Systems*, pages 3571–3580, 2017.

Emanuele Dolera, Stefano Favaro, and Edoardo Mainini. Strong posterior contraction rates via wasserstein dynamics, 2023.

Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(1):3–37, 2022.

John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521):182–201, 2018. doi: 10.1080/01621459.2017.1389735. URL https://doi.org/10.1080/01621459.2017.1389735.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In Serge Vaudenay, editor, *Advances in Cryptology - EUROCRYPT 2006*, pages 486–503, Berlin, Heidelberg, 2006a. Springer Berlin Heidelberg. ISBN 978-3-540-34547-3.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. volume Vol. 3876, pages 265–284, 01 2006b. ISBN 978-3-540-32731-8. doi: 10.1007/11681878_14.

Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, 2014.

Stefano Favaro, Antonio Lijoi, and Igor Prünster. Conditional formulae for Gibbs-type exchangeable random partitions. *The Annals of Applied Probability*, 23(5):1721 – 1754, 2013. doi: 10.1214/12-AAP843. URL https://doi.org/10.1214/12-AAP843.

Thomas S. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2): 209 – 230, 1973. doi: 10.1214/aos/1176342360. URL https://doi.org/10.1214/aos/1176342360.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162:707–738, 2015. doi: 10.1007/s00440-014-0583-7. URL https://doi.org/10.1007/s00440-014-0583-7.

Subhashis Ghosal and Aad van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2017.

A. Gnedin and J. Pitman. Exchangeable gibbs partitions and stirling triangles. *Journal of Mathematical Sciences*, 138:5674–5685, 2006. doi: 10.1007/s10958-006-0335-z.

Jingchen Hu, Matthew R. Williams, and Terrance D. Savitsky. Mechanisms for global differential privacy under bayesian data synthesis. *Statistica Sinica*, 2022. URL https://api.semanticscholar.org/CorpusID:248665552.

Jack E Jewson, Sahra Ghalebikesabi, and Chris C Holmes. Differentially private statistical inference through \beta-divergence one posterior sampling. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 76974–77001. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/f3024ea88cec9f45a411cf4d51ab649c-Paper-Conference.pdf.

Achim Klenke. *Probability Theory: A Comprehensive Course*. Universitext. Springer Cham, 3 edition, 2020. ISBN 978-3-030-56401-8. doi: 10.1007/978-3-030-56402-5. URL https://doi.org/10.1007/978-3-030-56402-5.

Antonio Lijoi, Ramsés H. Mena, and Igor Prünster. A bayesian nonparametric method for prediction in est analysis. *BMC Bioinformatics*, 8(1):339, 2007a. doi: 10.1186/1471-2105-8-339. URL https://doi.org/10.1186/1471-2105-8-339.

Antonio Lijoi, Ramsés H. Mena, and Igor Prünster. Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika*, 94(4):769–786, 2007b. URL https://www.jstor.org/stable/20441417.

Antonio Lijoi, Igor Prünster, and Stephen G. Walker. Bayesian nonparametric estimators derived from conditional gibbs structures. *The Annals of Applied Probability*, 18(4):1519–1547, 2008. ISSN 10505164. URL http://www.jstor.org/stable/25442677.

Albert Y. Lo. On a Class of Bayesian Nonparametric Estimates: I. Density Estimates. *The Annals of Statistics*, 12(1):351 – 357, 1984. doi: 10.1214/aos/1176346412. URL https://doi.org/10.1214/aos/1176346412.

Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam. L-diversity: privacy beyond k-anonymity. *22nd International Conference on Data Engineering (ICDE'06)*, pages 24–24, 2006. URL https://api.semanticscholar.org/CorpusID:679934.

Ashwin Machanavajjhala, Daniel Kifer, John Abowd, Johannes Gehrke, and Lars Vilhuber. Privacy: Theory meets practice on the map. pages 277–286, 04 2008. doi: 10.1109/ICDE.2008.4497436.

P. Massart. The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality. *The Annals of Probability*, 18 (3):1269 – 1283, 1990. doi: 10.1214/aop/1176990746. URL https://doi.org/10.1214/aop/1176990746.

R. Daniel Mauldin, William D. Sudderth, and S. C. Williams. Polya trees and random distributions. *The Annals of Statistics*, 20(3):1203–1221, 1992. ISSN 00905364. URL http://www.jstor.org/stable/2242009.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *Annual IEEE Symposium on Foundations of Computer Science (FOCS)*. IEEE, October 2007. URL https://www.microsoft.com/en-us/research/publication/mechanism-design-via-differential-privacy/.

Ilya Mironov. Rényi differential privacy. In *Proceedings of 30th IEEE Computer Security Foundations Symposium (CSF)*, pages 263–275, 2017. URL https://arxiv.org/abs/1702.07476.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997. ISSN 00018678. URL http://www.jstor.org/stable/1428011.

Radford M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000. doi: 10.1080/10618600.2000.10474879. URL https://www.tandfonline.com/doi/abs/10.1080/10618600.2000.10474879.

Mihael Perman, James Pitman, and Marc Yor. Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields*, 92:21–39, 1992. doi: 10.1007/BF01205234. URL https://doi.org/10.1007/BF01205234.

Jim Pitman. Some developments of the blackwell-macqueen urn scheme. *Lecture Notes-Monograph Series*, 30: 245–267, 1996. ISSN 07492170. URL http://www.jstor.org/stable/4355949.

Jim Pitman. Combinatorial stochastic processes. 2006. URL https://api.semanticscholar.org/CorpusID:118502441.

Jim Pitman and Marc Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855 – 900, 1997. doi: 10.1214/aop/1024404422. URL https://doi.org/10.1214/aop/1024404422.

Yosef Rinott, Christine M. O'Keefe, Natalie Shlomo, and Chris Skinner. Confidentiality and differential privacy in the dissemination of frequency tables. *Statistical Science*, 33(3):358–385, 2018. ISSN 08834237, 21688745. URL https://www.jstor.org/stable/26771006.

Sebastien Roch. *Modern Discrete Probability: An Essential Toolkit*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2024. doi: 10.1017/9781009305129.

Donald B Rubin. Statistical disclosure limitation. *Journal of official Statistics*, 9(2):461–468, 1993.

Terrance D. Savitsky, Matthew R. Williams, and Jingchen Hu. Bayesian pseudo posterior mechanism under asymptotic differential privacy. *J. Mach. Learn. Res.*, 23(1), jan 2022. ISSN 1532-4435.

Latanya Sweeney. k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570, oct 2002. ISSN 0218-4885. doi: 10.1142/S0218488502001648. URL https://doi.org/10.1142/S0218488502001648.

Yee Whye Teh. A hierarchical bayesian language model based on pitman-yor processes. In *Annual Meeting of the Association for Computational Linguistics*, 2006. URL https://api.semanticscholar.org/CorpusID:1541597.

Aad W van der Vaart and Jon A Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer New York, NY, 1 edition, 1996. ISBN 978-1-4757-2547-6. doi: 10.1007/978-1-4757-2545-2. URL https://doi.org/10.1007/978-1-4757-2545-2.

Cédric Villani. *Optimal Transport*. Grundlehren der mathematischen Wissenschaften. Springer Berlin, Heidelberg, 1 edition, 2008. ISBN 978-3-540-71049-3. doi: 10.1007/978-3-540-71050-9. URL https://doi.org/10.1007/978-3-540-71050-9. Hardcover ISBN: 978-3-540-71049-3, Softcover ISBN: 978-3-662-50180-1, eBook ISBN: 978-3-540-71050-9.

Larry Wasserman and Shuheng Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010. ISSN 01621459. URL http://www.jstor.org/stable/29747034.

V. M. Zolotarev. Probability metrics. *Theory of Probability and its Applications*, 28:278–302, 1984.

A. Appendix A - Additional proofs

Proof Proposition 4.2. We consider a dataset $\tilde{\mathbf{x}}$ such that $h(\tilde{\mathbf{x}}, \mathbf{x}) = 1$. By exchangeability, we can assume that the first observation differs in the two samples, i.e. $\tilde{x}_1 \neq x_1$. We will denote by x_l^* the unique value equal to the first observation x_1 for some $l \in \{1, ..., j\}$, while $\tilde{j}$ and $\tilde{\mathbf{n}}$ are the information encoded in $(\tilde{\mathbf{x}}, \mathbf{Z})$. If we assume that $n_l = 1$, then conditionally to $\mathbf{X} = \tilde{\mathbf{x}}$, the probability of having $S_l > 0$ is equal to 0, hence we need to assume $n_i > 1$ for all $i = 1, ..., j$. Under this assumption, conditionally to $\mathbf{X} = \tilde{\mathbf{x}}$, we have $\tilde{n}_l = n_l - 1$, $\tilde{n}_t = n_t + 1$ and $\tilde{j} = j$ for some $t \in \{1, ..., j\}$ with $t \neq l$; We require:

$$\frac{\mathbb{P}[\mathbf{S}_{m-L_m^{(n)}} = \mathbf{s}, L_m^{(n)} = s \mid K_n = \tilde{j}, \mathbf{N}_n = \tilde{\mathbf{n}}]}{\mathbb{P}[\mathbf{S}_{m-L_m^{(n)}} = \mathbf{s}, L_m^{(n)} = s \mid K_n = j, \mathbf{N}_n = \mathbf{n}]} = \frac{n_t + s_t - \sigma}{n_t - \sigma} \frac{n_l - 1 - \sigma}{n_l + s_l - 1 - \sigma} \in [e^{-\epsilon}, e^{\epsilon}]$$

For a fixed $\epsilon > 0$, we satisfy instance-level (ϵ, δ)-differential privacy if we take $\delta > \tilde{\delta}(\epsilon)$, with

$$\begin{aligned}\tilde{\delta}(\epsilon) =& 1 - \sum_{s=0}^{m} \mathbb{P}[(S_1, ..., S_j) \leq (e^{\epsilon} - 1)(n_1 - 1 - \sigma, ..., n_j - 1 - \sigma), L_m^{(n)} = s \mid K_n = j, \mathbf{N}_n = \mathbf{n}] \\ =& 1 - \sum_{s=0}^{m} \sum_{\substack{(s_1,...,s_j):\ s_i \geq 0, \\ \sum_{i=1}^{j} s_i = (m-s),\ s_i \leq \lfloor (e^{\epsilon}-1)(n_i - 1 - \sigma) \rfloor}} \prod_{i=1}^{j} \frac{(n_i - \sigma)_{s_i}}{s_i!} \frac{m!}{s!(\theta + n)_m} (\theta + j\sigma)_s\end{aligned}$$

□

Proof Proposition 4.1. First of all, notice that $z = K_n$ and $\theta = z|\sigma|$, with $\sigma < 0$; hence, the mechanism cannot generate new unique values. It follows that the proof is the same as that of Theorem 4.1, but only case 3. is possible. To satisfy instance-level (ϵ, δ)-differential privacy, we require

$$\frac{n_t + S_t + |\sigma|}{n_t + |\sigma|} \frac{n_l - 1 + |\sigma|}{n_l + S_l - 1 + |\sigma|} \in [e^{-\epsilon}, e^{\epsilon}] \tag{36}$$

for all $t, l = 1, ..., j$ such that $t \neq l$. Since σ can assume any negative value, we require $n_i + |\sigma| - 1 \geq m/(e^{\epsilon} - 1)$ for all $i = 1, ..., j$ and (36) is satisfied with probability 1, i.e. $\delta = 0$. □

Proof Theorem 2.1 (Wasserman and Zhou (2010)). Without loss of generality, take $i = 1$. The marginal distribution of the released data $Z_1, ..., Z_m$ under H_0 is $\nu_0(B) = \int_{\mathbb{X}^n} Q_n(B \mid s, x_2, ..., x_n)\ \mathfrak{p}_0^n(dx_2, ..., dx_n)$, while under H_1 is $\nu_1(B) = \int_{\mathbb{X}^n} Q_n(B \mid t, x_2, ..., x_n)\ \mathfrak{p}_0^n(dx_2, ..., dx_n)$. By the Neyman-Pearson Lemma, the most powerful test has a rejection region given by $U > u$ where $U(z) = (d\nu_1/d\nu_0)(z)$ and u is chosen so that $\int_{\mathbb{X}} \mathbb{1}_{(u,\infty)}(U(z))\ \nu_0(dz) \leq \gamma$. Since $(s, x_2, ..., x_n)$ and $(t, x_2, ..., x_n)$ differ in only one coordinate, $\nu_1(B) \leq e^{\epsilon}\nu_0(B)$ and so we have $\nu_1(U > u) \leq e^{\epsilon}\nu_0(U > u) \leq \gamma e^{\epsilon}$. □

Proof Theorem 3.1 (Wasserman and Zhou (2010)). Take an arbitrary dataset $\tilde{\mathbf{x}}$ such that $h(\tilde{\mathbf{x}}, \mathbf{x}) = 1$. Let $\hat{f}$ denote the smoothed histogram estimator $\hat{f}_{k,\sigma}$ based on $\mathbf{X} = \mathbf{x}$ and let $\hat{g}_{k,\sigma}$ denote the estimator based on $\mathbf{X} = \tilde{\mathbf{x}}$. We also denote by $\hat{p}_j(\mathbf{x})$ and $\hat{p}_j(\tilde{\mathbf{x}})$ the cell proportions. Note that $|\hat{p}_j(\mathbf{x}) - \hat{p}_j(\tilde{\mathbf{x}})| < 1/n$ for all $j = 1, ..., k$ by definition. It is clear that the maximum density ratio for any of the draw z_i occurs in one bin B_j. Now consider $\mathbf{z} = (z_1, ..., z_i)$ such that for all $i = 1, ..., m$, we have $z_i \in B_j$. Then, we have the following two possibilities:

1. Let $\hat{p}_j(\tilde{\mathbf{x}}) = 0$; then in order to maximize $\hat{f}(\mathbf{z})/\hat{g}(\mathbf{z})$, we let $\hat{p}_j(\mathbf{x}) = 1/n$ and obtain

$$\frac{\hat{f}(\mathbf{z})}{\hat{g}(\mathbf{z})} = \prod_{i=1}^{m} \frac{\hat{f}_{k,\sigma}(z_i)}{\hat{g}_{k,\sigma}(z_i)} \leq \left(\frac{(1-\sigma)k1/n + \sigma}{\sigma} \right)^m = \left(\frac{(1-\sigma)k}{n\sigma} + 1 \right)^m$$

2. Otherwise, we let $\hat{p}_j(\tilde{\mathbf{x}}) \geq 1/n$, (as by definition of $\hat{p}_j$, it takes t/n for non-negative integers t) and let $\hat{p}_j(\mathbf{x}) = \hat{p}_j(\tilde{\mathbf{x}}) \pm 1/n$. Now it is clear that in order to maximize the density ratio at $\mathbf{z}$, we may need to reverse the role of $\mathbf{x}$ and $\tilde{\mathbf{x}}$,

$$\max\left(\frac{\hat{g}(\mathbf{z})}{\hat{f}(\mathbf{z})}, \frac{\hat{f}(\mathbf{z})}{\hat{g}(\mathbf{z})} \right) \leq \max\left(\left(\frac{(1-\sigma)k\hat{p}_j + \sigma}{(1-\sigma)k(\hat{p}_j - 1/n) + \sigma} \right)^m, \left(\frac{(1-\sigma)k(\hat{p}_j - 1/n) + \sigma}{(1-\sigma)k\hat{p}_j + \sigma} \right)^m \right)$$
$$\max\left(\frac{(1-\sigma)k1/n}{(1-\sigma)k(\hat{p}_j - 1/n) + \sigma} + 1 \right)^m \leq \left(\frac{(1-\sigma)k}{n\sigma} + 1 \right)^m$$

where the maximum is achieved when $\hat{p}_j(\tilde{\mathbf{x}}) = 1/n$ and $\hat{p}_j(\mathbf{x}) = 0$, given a fixed set of parameters m, n, σ. Thus we have

$$\sup_{\mathbf{z} \in \mathbb{X}^m} \frac{\hat{f}(\mathbf{z})}{\hat{g}(\mathbf{z})} \leq \left(\frac{(1-\sigma)k}{n\sigma} + 1 \right)^m$$

and the theorem holds. □

B. Appendix B - Additional theory

In this section, we present miscellaneous results and definitions used throughout the thesis. In particular, some basics on the theory of probability kernels, taken from the book by Klenke (2020), and some results on random partitions.

B.1. Probability kernel

Definition B.1. *Let $(\Omega_1, \mathcal{A}_1)$, $(\Omega_2, \mathcal{A}_2)$ be measurable spaces. A map $\mathscr{K} : \Omega_1 \times \mathcal{A}_2 \to [0, \infty]$ is called a σ-finite transition kernel (from Ω_1 to Ω_2) if:*

1. *$\omega_1 \mapsto \mathscr{K}(\omega_1, A_2)$ is $\mathcal{A}_1$-measurable for any $A_2 \in \mathcal{A}_2$.*
2. *$A_2 \mapsto \mathscr{K}(\omega_1, A_2)$ is (σ)-finite measure on $(\Omega_2, \mathcal{A}_2)$ for any ω_1 in Ω_1.*

If in (2) the measure is a probability measure for all $\omega_1 \in \Omega_1$, then $\mathscr{K}$ is called a probability or stochastic kernel.

Definition B.2. *Let Y be a random variable with values in a measurable space $(E, \mathcal{E})$ and let $\mathcal{F} \subset \mathcal{A}$ be a sub-σ-algebra. A stochastic kernel $\mathscr{K}_{Y,\mathcal{F}}$ from $(\Omega, \mathcal{F})$ to $(E, \mathcal{E})$ is called regular conditional distribution of Y given $\mathcal{F}$ if*

$$\mathscr{K}_{Y,\mathcal{F}}(\omega, B) = \mathbb{P}[\{Y \in B\} \mid \mathcal{F}](\omega)$$

for $\mathbb{P}$-almost all $\omega \in \Omega$ and for all $B \in \mathcal{E}$; that is, if

$$\int 1_B(Y)\ 1_A\ d\mathbb{P} = \int \mathscr{K}_{Y,\mathcal{F}}(\cdot, B)\ 1_A\ d\mathbb{P}, \quad \text{for all } A \in \mathcal{F}, B \in \mathcal{E}.$$

When $\mathcal{F} = \sigma(X)$, with X random variable taking values in an arbitrary measurable space $(E', \mathcal{E}')$, the stochastic kernel

$$(x, A) \mapsto \mathscr{K}_{Y,X}(x, A) = \mathbb{P}[\{Y \in A\} \mid X = x] = \mathscr{K}_{Y,\sigma(X)}(X^{-1}(x), A)$$

is called regular conditional distribution of Y given X.

Definition B.3. *A measurable space $(E, \mathcal{E})$ is called Borel space if there exists a Borel set $B \in \mathfrak{B}(\mathbb{R})$ such that $(E, \mathcal{E})$ and $(B, \mathfrak{B}(B))$ are isomorphic.*

A separable topological space whose topology is induced by a complete metric is called a Polish space. In particular, $\mathbb{R}^d, \mathbb{Z}^d, \mathbb{R}^d$ and so forth are Polish. Closed subsets of Polish spaces are again Polish. Moreover, we have the following topological result

Theorem B.1. *Let E be a Polish space with Borel-σ-algebra $\mathcal{E}$. Then $(E, \mathcal{E})$ is a Borel space.*

The following theorem states the existence of stochastic kernels for random variables taking values in Borel spaces

Theorem B.2. *Let $\mathcal{F} \subset \mathcal{A}$ be a sub-σ-algebra. Let Y be a random variable with values in a Borel space $(E, \mathcal{E})$. Then there exists a regular conditional distribution $\mathscr{K}_{Y,\mathcal{F}}$ of Y given $\mathcal{F}$.*

Theorem B.3. *Let X be a random variable on $(\Omega, \mathcal{A}, \mathbb{P})$ with values in a Borel space $(E, \mathcal{E})$. Let $\mathcal{F} \subset \mathcal{A}$ be a σ-algebra and let $\mathscr{K}_{X,\mathcal{F}}$ be a regular conditional distribution of X given $\mathcal{F}$. Further, let $f : E \to \mathbb{R}$ be measurable and $\mathbb{E}[|f(X)|] < \infty$. Then*

$$\mathbb{E}[f(X) \mid \mathcal{F}](\omega) = \int f(x)\ \mathscr{K}_{X,\mathcal{F}}(\omega, dx) \quad \text{for } \mathbb{P}-\text{almost all } \omega$$

Theorem B.4. *Let $(\Omega_i, \mathcal{A}_i)$ be measurable spaces, $i = 1, 2$. Let μ be a finite measure on $(\Omega_1, \mathcal{A}_1)$ and let $\mathscr{K}$ be a finite transition kernel from Ω_1 to Ω_2. Assume that $f : \Omega_1 \times \Omega_2 \to \bar{\mathbb{R}}$ is measurable with respect to $\mathcal{A}_1 \otimes \mathcal{A}_2$. If $f \geq 0$ or $f \in \mathcal{L}^1(\mu \otimes \mathscr{K})$, then*

$$\int_{\Omega_1 \times \Omega_2} f\ d(\mu \otimes \mathscr{K}) = \int_{\Omega_1} \left(\int_{\Omega_2} f(\omega_1, \omega_2)\ \mathscr{K}(\omega_1, d\omega_2) \right) \mu(d\omega_1)$$

B.2. More on random partitions

Theorem B.5. *The probability that a random exchangeable partition of $[n]$ consists of j sets, for $j = 1, ..., n$, is equal to the following*

$$\sum_{(n_1,...,n_j)} \frac{1}{j!} \frac{n!}{n_1! \cdots n_j!}\ p_j^{(n)}(n_1, ..., n_j).$$

with the sum over all compositions $(n_1, ..., n_j)$ of n into j parts.

Proof (Ghosal and van der Vaart (2017)). Given a composition $(n_1, ..., n_j)$ of n into j parts, the probability of a particular ordered partition with set sizes $n_1, ..., n_j$ is $p_j^{(n)}(n_1, ..., n_j)/j!$. We can construct all ordered partitions of $[n]$ with composition $(n_1, ..., n_j)$ by first ordering the elements $1, 2, ..., n$ in every possible way and next defining the first set to consist of the first n_1 elements, the second set of the next n_2 elements, etc. There are $n!$ possible orderings of $1, 2, ..., n$, but permuting the first n_1 elements, the next n_2 elements, etc. gives the same ordered partition. Thus there are $n!/\prod_{i=1}^{j} n_i!$ ordered partitions with composition $(n_1, ..., n_j)$. By exchangeability they all have the same probability $p_j^{(n)}(n_1, ..., n_j)/j!$. □

As a consequence, exploiting the following formula (Charalambides (2005)):

$$\sum_{(n_1,...,n_j)} \frac{1}{j!} \frac{n!}{n_1! \cdots n_j!} \prod_{i=1}^{j} (1-\sigma)_{n_i-1} = \frac{1}{\sigma^j} \mathscr{C}(n, j; \sigma) \tag{37}$$

we obtain the following law for the number of unique values of a random partition induced by a PY process of parameter σ and θ:

$$\mathbb{P}[K_n = j] = \frac{\prod_{i=1}^{j-1}(\theta + i\sigma)}{\sigma^j\ (\theta+1)_{n-1}} \mathscr{C}(n, j; \sigma)$$

for all $j = 1, ..., n$.

Lemma B.1 (Lijoi et al. (2007b)). *For each $v \geq 1$ and $j \geq 1$, let $A_{j,v} = \{(v_1, ..., v_j) : v_i \geq 0, \sum_{i=1}^{j} v_i = v\}$. Then*

$$\sum_{(v_1,...,v_j) \in A_{j,v}} \frac{v!}{v_1! \cdots v_j!} \prod_{i=1}^{j} (1-\sigma)_{n_i+v_i-1} = (n - j\sigma)_v \prod_{i=1}^{j} (1-\sigma)_{n_i-1}$$

where $(n_1, ..., n_j)$ is such that $n_i > 0$, for $i = 1, ..., j$, and $\sum_{i=1}^{j} n_i = n$.

Proposition B.1 (Favaro et al. (2013)). *Let $(X_n)_{n \geq 1}$ be a sequence of exchangeable random variables directed by a $\mathscr{PY}(\sigma, \theta, H)$. Then*

$$\mathbb{P}[K_m^{(n)} = k, L_m^{(n)} = s \mid K_n = j, \mathbf{N}_n = \mathbf{n}] = \binom{m}{s} \frac{\prod_{i=j}^{j+k-1}(\theta + i\sigma)}{(\theta+n)_m} (n - j\sigma)_{m-s} \frac{\mathscr{C}(s, k, \sigma)}{\sigma^k}$$

for all $s = 0, ..., m$ *and* $k = 0, ..., s$. *Moreover,*

$$\mathbb{P}[(S_{j+\tau_1}, ..., S_{j+\tau_y}) = (s_{j+\tau_1}, ..., s_{j+\tau_y}) \mid K_n = j, \mathbf{N}_n = \mathbf{n}, L_m^{(n)} = s, K_m^{(n)} = k] =$$
$$\frac{s!}{s_{j+\tau_1}! \cdots s_{j+\tau_y}! \ (s - \sum_{i=1}^{y} s_{j+\tau_i})!} \frac{(k-y)!}{k!} \sigma^y \prod_{i=1}^{y} (1-\sigma)_{s_{j+\tau_i}-1} \frac{\mathscr{C}(s - \sum_{i=1}^{y} s_{j+\tau_i}, k-y, \sigma)}{\mathscr{C}(s, k, \sigma)}$$

for any $1 \leq y \leq k$ *and for any collection of indices* $\tau_1, \tau_2, ..., \tau_y \in \mathbb{N}$ *such that* $1 \leq \tau_1 < \tau_2 < \cdots < \tau_y \leq k$.

Proof of the first statement. For any $n \geq 1$, $m \geq 1$, $j = 1, ..., n$ and $s = 0, ..., m$ let us define the following partition-set

$$\mathcal{D}_{m-s,j}^{(0)} := \left\{ (s_1, ..., s_j) : s_i \geq 0, \sum_{i=1}^{j} s_i = m - s \right\}.$$

and, for any $k = 0, ..., s$, the following

$$\mathcal{D}_{s,k} := \left\{ (s_{j+1}, ..., s_{j+k}) : s_i > 0, \sum_{i=1}^{k} s_{j+i} = s \right\}$$

The proof consists of marginalizing out the random variables $\mathbf{S}_{L_m^{(n)}}$ and $\mathbf{S}_{m-L_m^{(n)}}$ as defined in Section 4.2 from the law of the random partition induced by $X_{n+1}, ..., X_{n+m}$ having observed the random partition induced by $X_1, ..., X_n$, which, using the EPPF, can be written as follows

$$\begin{aligned}
&\mathbb{P}[\mathbf{N}_n + \mathbf{S}_{m-L_m^{(n)}} = \mathbf{n} + \mathbf{s}_{m-L_m^{(n)}}, L_m^{(n)} = s, K_m^{(n)} = k, \mathbf{S}_{L_m^{(n)}} = \mathbf{s}_{L_m^{(n)}} \mid K_n = j, \mathbf{N}_n = \mathbf{n}] \\
&\quad = \frac{p_{j+k}^{(n+m)}(n_1 + s_1, ..., n_j + s_j, s_{j+1}, ..., s_{j+k})}{p_j^{(n)}(n_1, ..., n_j)} \\
&\quad = \frac{1}{(\theta+n)_m} \prod_{i=j}^{j+k-1} (\theta + i\sigma) \prod_{i=1}^{j} (n_i - \sigma)_{s_i} \prod_{r=1}^{k} (1-\sigma)_{s_{j+r}-1}
\end{aligned}$$

Hence, we have

$$\begin{aligned}
&\mathbb{P}[K_m^{(n)} = k, L_m^{(n)} = s \mid K_n = j, \mathbf{N}_n = \mathbf{n}] = \\
&\binom{m}{s} \sum_{\mathcal{D}_{m-s,j}^{(0)}} \frac{(m-s)!}{s_1! \cdots s_j!} \frac{1}{k!} \sum_{\mathcal{D}_{s,k}} \frac{s!}{s_{j+1}! \cdots s_{j+k}!} \times \\
&\frac{1}{(\theta+n)_m} \prod_{i=j}^{j+k-1} (\theta + i\sigma) \prod_{i=1}^{j} \frac{(1-\sigma)_{n_i+s_i-1}}{(1-\sigma)_{n_i-1}} \prod_{r=1}^{k} (1-\sigma)_{s_{j+r}-1}
\end{aligned}$$

and the two sums can be solved using Lemma B.1 and Equation 37.

□

Proposition B.2 (Lijoi et al. (2007b)). *Let* $(X_n)_{n\geq 1}$ *be a sequence of exchangeable random variables directed by a* $\mathscr{PY}(\sigma, \theta, H)$ *and* $L_m^{(n)}$ *the number of observations in* $X_{n+1}, ..., X_{n+m}$ *different from the* $K_n = j$ *unique values observed in* $X_1, ..., X_n$, *one has*

$$\mathbb{E}[L_m^{(n)} \mid K_n = j] = m \frac{\theta + j\sigma}{\theta + n} \tag{38}$$

Proof. The proof works by induction. Let us first note that for any $m \geq 1$ one has

$$L_{m+1}^{(n)} = L_m^{(n)} + H_{n,m}$$

where $H_{n,m} = \mathbb{1}_{\{X_1^*, ..., X_{K_n}^*\}^c}(X_{n+m+1})$. Let us first fix $m = 1$ and determine

$$\mathbb{E}[L_2^{(n)} \mid K_n = j] = \mathbb{E}[L_1^{(n)} \mid K_n = j] + \mathbb{E}[H_{n,1} \mid K_n = j]$$

The first summand is equal to $p_{j+1}^{(n+1)}(n_1, ..., n_j, 1)/p_j^{(n)}(n_1, ..., n_j) = (\theta + j\sigma)/(\theta + n)$. As for the second summand, one can use the assumption of exchangeability which yields

$$\begin{aligned}
\mathbb{E}[H_{n,1} \mid K_n = j] =& \mathbb{E}[\mathbb{1}_{\{X_1^*, ..., X_{K_n}^*\}^c}(X_{n+2}) \mid K_n = j] = \\
&\mathbb{E}[\mathbb{1}_{\{X_1^*, ..., X_{K_n}^*\}^c}(X_{n+1}) \mid K_n = j] = \frac{\theta + j\sigma}{\theta + n}
\end{aligned}$$

Now, suppose Equation 38 is valid for m, and let us show this implies it is still true for $m+1$. This means we shall determine

$$\mathbb{E}[L_{m+1}^{(n)} \mid K_n = j] = \mathbb{E}[L_m^{(n)} \mid K_n = j] + \mathbb{E}[H_{n,m} \mid K_n = j]$$

By assumption, $\mathbb{E}[L_m^{(n)} \mid K_n = j] = m(\theta + j\sigma)/(\theta + n)$. Moreover, exchangeability again entails the second summand above is $(\theta + j\sigma)/(\theta + n)$. □

B.3. Additional MC estimate

In this section, we provide additional tables representing the MC estimate of $\tilde{\delta}(\epsilon)$ in scenarios different from those considered in Section 4.2.

m	500	750	1250
$\epsilon = 1$	0.4399 ± 0.0340	0.6903 ± 0.0216	0.7733 ± 0.0184
$\epsilon = 2$	0.4558 ± 0.0175	0.7591 ± 0.0188	0.5271 ± 0.0220
$\epsilon = 3$	0.0817 ± 0.0120	0.3943 ± 0.0219	0.4459 ± 0.0216
$\epsilon = 4$	0.0002 ± 0.0006	0 ± 0	0.0080 ± 0.00394

Table 9: Estimated $\tilde{\delta}(\epsilon)$ for $n = 5000$, $\theta = 3, \sigma = 1/5$ while varying ϵ and m.

m	50	125	250
$\epsilon = 1$	0.4499 ± 0.0220	0.9133 ± 0.9999	0.9999 ± 0.0004
$\epsilon = 2$	0.00009 ± 0.00044	0.0018 ± 0.0019	0.0939 ± 0.0129
$\epsilon = 3$	0.00009 ± 0.00044	0.00009 ± 0.00044	0.0006 ± 0.0011
$\epsilon = 3.2$	0 ± 0	0 ± 0	0 ± 0

Table 10: Estimated $\tilde{\delta}(\epsilon)$ for $n = 500$, $\theta = 10, \sigma = 0$ while varying ϵ and m.

m	500	750	1250
$\epsilon = 1$	0.5690 ± 0.0210	0.8842 ± 0.01413	0.9531 ± 0.0179
$\epsilon = 2$	0 ± 0	0.0005 ± 0.0009	0.0022 ± 0.0020
$\epsilon = 4$	0 ± 0	0 ± 0	0 ± 0

Table 11: Estimated $\tilde{\delta}(\epsilon)$ for $n = 5000$, $\theta = 3, \sigma = 0$ while varying ϵ and m.

m	50	125	250
$\epsilon = 0.8$	0.0779 ± 0.0118	0.1890 ± 0.0120	0.1007 ± 0.0133
$\epsilon = 1.1$	0 ± 0	0 ± 0	0 ± 0

Table 12: Estimated $\tilde{\delta}(\epsilon)$ for $n = 5000$, $\theta = 3, \sigma = 1/2$ while varying ϵ and m in case we release only the counts of the unique values observed in the private data.

Abstract in lingua italiana

Questo lavoro affronta il problema della generazione di dati sintetici da una prospettiva statistica. In particolare, si passano in rassegna alcuni metodi noti in letteratura per generare dati sintetici e si introduce un nuovo meccanismo di rilascio basato su un modello Bayesiano Nonparametrico. Modelliamo i dati privati come un campione da una prior Bayesiana Nonparametrica, il processo Pitman-Yor, e generiamo nuovi dati campionando dalla distribuzione predittiva a posteriori. Mostriamo le ipotesi richieste dal meccanismo per soddisfare privacy differenziale. Studiamo i vantaggi del nostro metodo sia in termini frequentisti, mostrando un risultato di convergenza per il valore atteso della distanza 1-Wasserstein tra la misura empirica basata sui dati rilasciati e la distribuzione che genera i dati, sia in termini Bayesiani, dimostrando che è possibile, rilasciando i dati dalla distribuzione predittiva, fare facilmente inferenza a posteriori sui dati privati.

Parole chiave: Privacy differenziale, dati sintetici, Bayesiana Nonparametrica, processo Pitman-Yor

Ringraziamenti

Vorrei ringraziare il Professore Mario Beraha per avermi accompagnato con competenza e disponibilità, non solo nel mio percorso di tesi, ma in generale in quest' ultimo anno alla scoperta della Statistica Matematica e alla ricerca di una posizione di dottorato. Ci tengo anche a ringraziare il Professore Stefano Favaro, che ha deciso di supervisionare questa tesi e sostenere le mie candidature. Se ho compreso la mia passione per la Statistica e la ricerca, molto del merito va a voi che mi avete trasmesso tanti stimoli e regalato utili discussioni.
Ringrazio poi tutta la mia famiglia che da sempre crede in me, facendo il possibile per farmi vivere serenamente e sostenendomi in tutti i momenti più difficili. In particolare, grazie Aurora: essere apprezzato da te, che sei un passo avanti a tutti, mi fa sentire più soddisfatto di me stesso ogni giorno.
Ringrazio Alice, che con amore, dolcezza e costanza ha riempito le mie giornate di emozioni e mi ha sopportato anche quando il mio pessimismo cosmico e le mille cose che dovevo fare prendevano il sopravvento.
Ringrazio tutti gli abitanti di Casa Piola per avermi fatto scoprire cosa significa essere parte di una seconda famiglia. Stare a Milano è diventato molto più bello da quando vivo in casa con voi.
Infine ringrazio tutti i miei amici, sia quelli di vecchia data, che quelli di Milano e Bologna. Ognuno di voi è stato un tassello fondamentale della mia formazione come individuo e ha contribuito a suo modo a questa laurea.