Dirichlet Mechanism for Differentially Private KL Divergence Minimization

Anonymous authors
Paper under double-blind review

Abstract

Given an empirical distribution $f(x)$ of sensitive data x, we consider the task of minimizing $F(y) = D_{\mathrm{KL}}(f(x)\|y)$ over a probability simplex, while protecting the privacy of x. We observe that, if we take the exponential mechanism and use the KL divergence as the loss function, then the resulting algorithm is the *Dirichlet mechanism* that outputs a single draw from a Dirichlet distribution. Motivated by this, we propose a Rényi differentially private (RDP) algorithm that employs the Dirichlet mechanism to solve the KL divergence minimization task. In addition, given $f(x)$ as above and $\hat{y}$ an output of the Dirichlet mechanism, we prove a probability tail bound on $D_{\mathrm{KL}}(f(x)\|\hat{y})$, which is then used to derive a lower bound for the sample complexity of our RDP algorithm. Experiments on real-world datasets demonstrate advantages of our algorithm over Gaussian and Laplace mechanisms in supervised classification and maximum likelihood estimation.

1 Introduction

KL divergence is the most commonly used divergence measure in probabilistic and Bayesian modeling. In a probabilistic model, for example, we estimate the model's parameters by maximizing the likelihood function of the parameters, which in turn is equivalent to minimizing the KL divergence between the empirical distribution and the model's distribution. In supervised classification, a standard way to fit a classifier is by minimizing the cross-entropy of the model's predictive probabilities, which is equivalent to minimizing the KL divergence between the class-conditional empirical distribution and the model's predictive distribution.

Such models are widely used in medical fields, social sciences and businesses, where they are used to analyze sensitive personal information. Without privacy considerations, releasing a model to public might put the personal data at risk of being exposed to privacy attacks, such as membership inference attacks (Shokri et al., 2017; Ye et al., 2022). To address the model's privacy issue, we should focus on its building blocks: the KL divergences. How can we minimize the KL divergence over the model's parameters, while keeping the data private?

Differential Privacy (Dwork et al., 2006a;b) provides a framework for quantitative privacy analysis of algorithms that run on sensitive personal data. Under this framework, one aims to design a task-specific algorithm that preserves the privacy of the inputs, while keeping the "distance" between the privatized output and the true output sufficiently small. A simple and well-studied technique is to add a small random noise sampled from a zero-centered probability distribution, such as the Laplace and Gaussian distributions. Another technique is to sample an output from a distribution, with greater probabilities of obtaining points that are closer to the true output, such as the exponential mechanism (McSherry & Talwar, 2007). These techniques have been deployed in many privacy-preserving tasks, from simple tasks such as private counting and histogram queries (Dwork et al., 2006a;b) to complex tasks such as deep learning (Abadi et al., 2016).

In this work, we are interested in a setting where our algorithm outputs an empirical distribution $f(x)$ of some sensitive data x. To protect the privacy of individuals in x, we keep $f(x)$ hidden, and instead release another discrete distribution that approximates $f(x)$ in KL divergence. This setting may not arise, for example, in the task of releasing a normalized histogram, as the distance between histograms is often measured in ℓ^1 or ℓ^2. Nonetheless, there are many tasks where KL divergence arises naturally. Prominent

examples are those in probabilistic modeling, where the outputs—the model's estimated parameters—are obtained from likelihood maximization. Another examples are those in Bayesian modeling, where models are evaluated with adjusted negative log-likelihood scores, such as the Akaike information criterion (AIC) and Bayesian information criterion (BIC). It is also increasingly common in Bayesian practice to evaluate the model with out-of-sample log-likelihood (Vehtari et al., 2016). Again, minimizing these criteria can be formulated as minimizing the KL divergence.

A simple approach to privatize a discrete probability distribution is by adding some random noises from a probability distribution. However, the KL divergence does not behave smoothly with the additive noises, as the following example illustrates: consider count data of 10 people, interpreted as a normalized histogram: $p = (0.1, 0.9)$. Suppose that we draw two sets of noises $z_1 = (-0.090, 0.045)$ and $z_2 = (-0.099, -0.038)$ from Laplace(1/10). Here, the ℓ^1-distances between p and its noisy versions are 0.135 and 0.137, a very small difference. On the other hand, the KL divergences between between p and its noisy versions are 0.186 and 0.499, a 2.68 times increase. This example illustrates that adding noises to a discrete probability vector, even at a small scale, could result in a noisy vector that is too far away from the original vector in terms of KL divergence.

We instead consider the exponential mechanism, a differentially private algorithm that approximately minimizes user-defined *loss functions*. It turns out that, by taking the loss function to be the KL divergence, the exponential mechanism turns into one-time sampling from a Dirichlet distribution; we shall call this the *Dirichlet mechanism*.

The Dirichlet mechanism, however, does not inherit the differential privacy guarantee of the exponential mechanism: the guarantee in (McSherry & Talwar, 2007) requires the loss function to be bounded above, while the KL divergence can be arbitrarily large. In fact, using the original definition of differential privacy (Dwork et al., 2006b), the Dirichlet mechanism is not differentially private (see Appendix A). We thus turn to a relaxation of differential privacy. Specifically, using the notion of the Rényi differential privacy (Mironov, 2017), we study the Dirichlet mechanism and its utility in terms of KL divergence minimization.

1.1 Overview of Our results

Below are summaries of our results.

§3 Privacy. We propose a version of the Dirichlet mechanism (Algorithm 1) that satisfies the Rényi differential privacy (RDP). In this algorithm, we need to evaluate a polygamma function and find the root of a strictly increasing function. Our algorithm is easy to implement, as polygamma functions, root-finding methods and Dirichlet distributions are readily available in many scientific programming languages.

§4 Utility. We derive a probability tail bound for $D_{\mathrm{KL}}(p\|q)$ when q is drawn from a Dirichlet distribution (Theorem 2). From this, we derive a lower bound for the sample complexity of Algorithm 1 that attains a target privacy guarantee, both in general case and on categorical data.

§5 Experiments. We compare the Dirichlet mechanism against the Gaussian and Laplace mechanisms for two learning tasks: naïve Bayes classification and maximum likelihood estimation of Bayesian networks—both tasks can be done with KL divergence minimization. Experiments on real-world datasets show that the Dirichlet mechanism provides smaller cross-entropy loss in classification, and larger log-likelihood in parameter estimation, than the other mechanisms at the same level of privacy guarantee.

1.2 Notations

In this paper, all vectors are d-dimensional, where $d \geq 2$. The number of observations is always N. Let $[d] \coloneqq [1, \ldots, d]$. For any $u \in \mathbb{R}^d$, we let u_i be the i-th coordinate of u, and for any vector-valued function $f : \mathcal{X} \to \mathbb{R}^d$, we let f_i be that i-th component of f. Let $\mathbb{R}^d_{\geq 0}$ be the set of d-tuples of non-negative real numbers, and $\mathbb{R}^d_{>0}$ be the set of d-tuples of positive real numbers. Denote the probability simplex by

$$S^{d-1} \coloneqq \Big\{ p \in \mathbb{R}^d_{\geq 0} : \sum_i p_i = 1 \Big\}.$$

For any $u, u' \in \mathbb{R}^d$ and scalar $r > 0$, we write $u + u' \coloneqq (u_1 + u'_1, \ldots, u_d + u'_d)$ and $ru \coloneqq (ru_1, \ldots, ru_d)$. For any positive-valued functions f, f', the notation $f(x) \propto f'(x)$ means $f(x) = Cf'(x)$ for some constant $C > 0$ and $f(x) \approx f'(x)$ means $cf'(x) \leq f(x) \leq Cf'(x)$ for some $c, C > 0$. Lastly, $\|u\|_2 \coloneqq \sqrt{u_1^2 + \ldots + u_d^2}$ is the ℓ^2 norm of u and $\|u\|_\infty \coloneqq \max_i |u_i|$ is the ℓ^∞ norm of u.

2 Background and related work

2.1 Privacy models

We say that two datasets are *neighboring* if they differ on a single entry. Here, an *entry* can be a row of the datasets, or a single attribute of a row.

Definition 2.1 (Pure and Approximate differential privacy (Dwork et al., 2006a;b))**.** A randomized mechanism $M : \mathcal{X}^n \to \mathcal{Y}$ is (ε, δ)-differentially private ((ε, δ)-DP) if for any two neighboring datasets x and x' and all events $E \subset \mathcal{Y}$,

$$\Pr[M(x) \in E] \leq e^{\varepsilon} \Pr[M(x') \in E] + \delta. \tag{1}$$

If M is $(\varepsilon, 0)$-DP, then we say that it is ε-differential private (ε-DP).

The term *pure differential privacy* (pure DP) refers to ε-differential privacy, while *approximate differential privacy* (approximate DP) refers to (ε, δ)-DP when $\delta > 0$.

In this paper, we shall concern ourselves with Rényi differential privacy, a relaxed notion of differential privacy defined in terms of the Rényi divergence between $M(x)$ and $M(x')$:

Definition 2.2 (Rényi Divergence (Rényi, 1961))**.** Let P and Q be probability distributions. For $\lambda \in (1, \infty)$ the Rényi divergence of order λ between P and Q is defined as

$$\mathrm{D}_\lambda(P \| Q) = \frac{1}{\lambda - 1} \log \left(\mathop{\mathbb{E}}_{y \sim P} \left[\frac{P(y)^{\lambda-1}}{Q(y)^{\lambda-1}} \right] \right).$$

and for $\lambda = 1$, we define $D_1(P \| Q) = D_{\mathrm{KL}}(P \| Q)$,

Definition 2.3 (Rényi differential privacy (Mironov, 2017))**.** A randomized mechanism $M : \mathcal{X}^n \to \mathcal{Y}$ is (λ, ε)-Rényi differentially private ((λ, ε)-RDP) if for any two neighboring datasets x and x',

$$\mathrm{D}_\lambda(M(x) \| M(x')) \leq \varepsilon.$$

Intuitively, ε controls the moments of the privacy loss random variable: $Z \coloneqq \log \frac{P[M(x)=Y]}{P[M(x')=Y]}$, where Y is distributed as $M(x)$, up to order λ. A smaller ε and larger λ correspond to a stronger privacy guarantee.

The following composition property of RDP mechanisms allow us to track the privacy guarantees of using multiple Dirichlet mechanisms. This can be helpful when Dirichlet mechanisms is employed in a more complex algorithms, such as fitting a discrete probabilistic model.

Lemma 1 (Composition of RDP mechanisms (Mironov, 2017))**.** *Let $M_1 : \mathcal{X}^n \to \mathcal{Y}$ be a $(\lambda_1, \varepsilon_1)$-RDP mechanism and $M_2 : \mathcal{X}^n \to \mathcal{Z}$ be a $(\lambda_2, \varepsilon_2)$-RDP mechanism. Then a mechanism $M : \mathcal{X}^n \to \mathcal{Y} \times \mathcal{Z}$ defined by $M(x) = (M_1(x), M_2(x))$ is $(\min(\lambda_1, \lambda_2), \varepsilon_1 + \varepsilon_2)$-RDP.*

2.2 Exponential mechanism with the KL divergence

The exponential mechanism (McSherry & Talwar, 2007) is a privacy mechanism that releases an element from a range $\mathcal{Y}$ that approximately minimizes a given *loss function* $\ell : \mathcal{X}^N \times \mathcal{Y} \to \mathbb{R}$. Given a base measure μ over $\mathcal{Y}$ and a dataset $x \in \mathcal{X}^N$, the mechanism outputs $y \in \mathcal{Y}$ with probability density proportional to:

$$e^{-\beta \ell(x,y)} \mu(y), \tag{2}$$

where β is a function of ε, the privacy parameter.

For the first time, we point out the connection between the exponential mechanism and a well-known family of probability distributions under a specific choice of $\ell(x, y)$. Let $f : \mathcal{X}^N \to \mathbb{R}^d_{\geq 0}$ be an arbitrary vector-valued function on datasets. Let $\mathcal{Y} = S^{d-1}$. Assuming that $N_f \coloneqq \sum_i f_i(x)$ is known and nonzero, we denote the normalized vector $\widetilde{f(x)} = N_f^{-1} f(x) \in S^{d-1}$. In equation 2, let $\ell(x, y) = D_{\mathrm{KL}}(\widetilde{f(x)} \| y)$, $\beta = rN_f$, and $\mu(y)$ be the density of Dirichlet($\boldsymbol{\alpha}$), that is, $\mu(y) \propto \prod_{i=1}^d y_i^{\alpha-1}$. Then, the probability density of the output y of the corresponding exponential mechanism is proportional to:

$$\begin{aligned}
\exp\Big(-rN_f D_{\mathrm{KL}}(\widetilde{f(x)} \| y)\Big) \prod_i y_i^{\alpha-1} &= \exp\left(r \sum_{i, x_i \neq 0} f_i(x) \log(y_i / \widetilde{f_i(x)})\right) \prod_i y_i^{\alpha-1} \\
&\propto \prod_{i, x_i \neq 0} y_i^{r f_i(x)} \prod_i y_i^{\alpha-1} \\
&= \prod_i y_i^{r f_i(x) + \alpha - 1},
\end{aligned}$$

which is exactly the non-normalized density function of Dirichlet($rf(x) + \alpha$). This specific distribution will play a major role in the main privacy mechanism introduced in the next section.

From this derivation, we can see that this particular instance of the exponential mechanism can be used to output y that approximately minimizes the KL divergence $D_{\mathrm{KL}}\Big(\widetilde{f(x)} \| y\Big)$ while keeping x private.

A related problem is learning of a discrete distribution in KL divergence: given a sample from an *unknown* discrete distribution p, one aims to find y with small $D_{\mathrm{KL}}(p \| y)$. See Kamath et al. (2015); Diakonikolas et al. (2015); Canonne et al. (2023) for extensive discussion on this problem. We emphasize that our work solely focuses on private estimation of an empirical distribution, not an unknown discrete distribution, and we leave the latter open for future work.

To see how the choices of r and α affect the "distance" between y_i and $\widetilde{f_i(x)}$, we treat y_i as an estimator of $\widetilde{f_i(x)}$ and look at the bias of y_i:

$$\left|\mathbb{E}[y_i] - \widetilde{f_i(x)}\right| = \left|\frac{r f_i(x) + \alpha}{rN_f + d\alpha} - \frac{f_i(x)}{N_f}\right| = \frac{\alpha |N_f - d f_i(x)|}{N_f (rN_f + d\alpha)}. \tag{3}$$

The bias is reduced when r increases and α decreases. We can also look at the variance of y_i:

$$\mathrm{Var}[y_i] = \frac{(r f_i(x) + \alpha)(r(N_f - f_i(x)) + (d-1)\alpha)}{(rN_f + d\alpha)^2 (rN_f + d\alpha + 1)},$$

which is $O(1/r)$ as $r \to \infty$ and $O(1/\alpha)$ as $\alpha \to \infty$. This implies that draws from Dirichlet($rf(x) + \alpha$) are more concentrated when r and α are large.

Applications. The derivation of the Dirichlet mechanism above suggests that the best use of the Dirichlet mechanism is for privately minimizing KL divergence, which arises in the following scenarios:

1. **Maximum likelihood estimation.** Consider a problem of parameter estimation in a multinomial model with d possible outcomes. Let $x \in [d]^N$ be N observations, $f_1(x), \ldots, f_d(x)$ be the frequencies and $y_1, \ldots, y_d$ be the model's parameters. Then the log-likelihood of x is $\sum_i f_i(x) \log y_i$. Maximizing the log-likelihood with respect to y is equivalent to minimizing the KL divergence:

$$\arg\max_y \sum_i f_i(x) \log y_i = \arg\min_y D_{\mathrm{KL}}\left(\frac{f(x)}{N} \middle\| y\right).$$

   Thus, we can use the Dirichlet mechanism to release the parameters of the model while keeping x private.

2. **Cross-entropy minimization.** Consider the same multinomial model as above. One might instead aim to minimize the cross-entropy loss: $-\frac{1}{N} \sum_i f_i(x) \log y_i$ over y. This is also equivalent to minimizing the KL divergence, so we can use the Dirichlet mechanism to privately solve for y.

2.3 Polygamma functions

In most of this study, we take advantage of several nice properties of the log-gamma function and its derivatives. The *polygamma function of order* m is the $(m+1)$-th derivative of the logarithm of the gamma function. Specifically, when $m = 0$, we have the *digamma function* $\psi(x) := \frac{d}{dx}\log\Gamma(x)$, which is a concave and increasing function.

Our function of interest is the polygamma function of order 1: $\psi'(x)$, which is a positive, convex, and decreasing function (see Figure 1). It has the series representation:

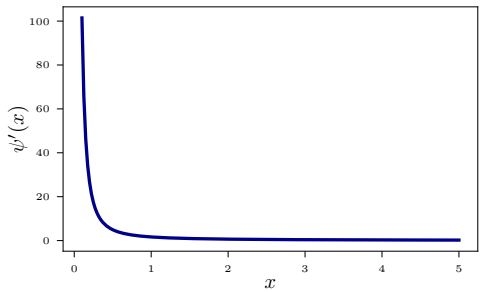

Figure 1: A plot of $\psi'(x)$.

$$\psi'(x) = \sum_{k=0}^{\infty} \frac{1}{(x+k)^2}, \tag{4}$$

which allows for fast approximations of $\psi'(x)$ at any precision. ψ' can also be approximated by the reciprocals:

$$\frac{1}{x} + \frac{1}{2x^2} < \psi'(x) < \frac{1}{x} + \frac{1}{x^2}, \tag{5}$$

which implies that $\psi'(x) \approx \frac{1}{x^2}$ as $x \to 0$ and $\psi'(x) \approx \frac{1}{x}$ as $x \to \infty$.

2.4 Related work

There are several studies on the differential privacy of obtaining a single draw from a probability distribution whose probability density function is of the form $y \mapsto \frac{1}{Z}p(x|y)\mu(y)$. Here, x is sensitive data, $x \mapsto p(x|y)$ is a probability density function for all y in the domain, μ is any positive-valued function, and Z is the normalizing constant. Wang et al. (2015) showed that, when $|\log p(x \mid y)| \leq B$ for some constant B, then a single draw is $4B$-differentially private. However, the densities that we study are not bounded away from zero; they have the form $\prod_i y_i^{rf_i(x)+\alpha}$ which becomes small when one of the y_i's is close to zero. Dimitrakakis et al. (2017) showed that, when p is the density of the binomial distribution and μ is the density of the beta distribution, then a single draw is $(0, \delta)$-DP, and the result cannot be improved unless the parameters are assumed to be above a positive threshold. As a continuation of their work, we prove in the appendix that, when the parameters are bounded below by $\alpha > 0$, sampling from the Dirichlet distribution (which is a generalization of the beta distribution) is (ε, δ)-DP with $\varepsilon > 0$.

Let x be a sufficient statistic of an exponential family with finite ℓ^1-sensitivity. Foulds et al. (2016) showed that sampling $Y \sim p(y \mid \hat{x})$, where $\hat{x} = x +$ Laplace noise, is differentially private and as asymptotically efficient as sampling from $p(y \mid x)$. However, for a small sample size, the posterior over the noisy statistics might be too far away from the actual posterior. Bernstein & Sheldon (2018) thus proposed to approximate the joint distribution $p(y, x, \hat{x})$ using Gibbs sampling, which is then integrated over x to obtain a more accurate posterior over $\hat{x}$.

Geumlek et al. (2017) were the first to study sampling from exponential families with Rényi differential privacy (RDP; Mironov (2017)). Even though they provided a general framework to find (λ, ε)-RDP guarantees for exponential families, explicit forms of λ and the upper bound of λ were not given.

The privacy of data synthesis via sampling from Multinomial(Y), where Y is a discrete distribution drawn from the Dirichlet posterior, was first studied by Machanavajjhala et al. (2008). They showed that the data synthesis is (ε,δ)-DP, where ε grows by the number of draws from Multinomial(Y). In contrast, we show that a single draw from the Dirichlet posterior is approximate DP, which by the post-processing property allows us to sample from Multinomial(Y) as many times as we want while retaining the same privacy guarantee.

Gohari et al. (2021) have recently showed that the Dirichlet mechanism is $(\hat{\varepsilon}(r, \gamma, \eta, \eta'), \delta(r, \gamma, \eta, \eta'))$-DP, where $\gamma, \eta, \eta' \in (0, 1)$ are additional parameters. Not only the guarantee has many parameters to optimize, it is also computational intensive. Specifically, for any $W \subset [d]$, define $\Omega_W^{\eta,\eta'} =$

$\left\{p \in S^{d-1} : p_i > \gamma, \forall i \in W; \sum_{i \in W} p_i \leq 1 - \eta'\right\}$. Gohari et al. proposed $\hat{\varepsilon} = \Theta(r \log(1/\gamma))$ and

$$\delta = 1 - \min_{p \in \Omega_W^{\eta,\eta'}, W \subset [d]} \left\{\Pr[Y_i > \gamma; \forall i \in W] : Y \sim \text{Dirichlet}(rp)\right\}. \tag{6}$$

To compute δ, we have to approximate $\Pr[Y > \gamma]$ with a numerical integration scheme with high precision, otherwise the integral may be greater than one. Even then, the integral is highly dependent on the scheme, and for some choices of the parameters r, η, η', the value of δ cannot go below a certain threshold. We illustrate this in Figure 2. With $r = 171.87, \eta = 0.028$ and $\eta' = 0.114$, the value of δ cannot go below 2.1×10^{-4}. In contrast, our guarantee is much simpler to compute, as the function ψ' can be easily approximated via its series representation (equation 4). Moreover, we are the first to provide the utility of the Dirichlet mechanism in terms of KL divergence minimization.

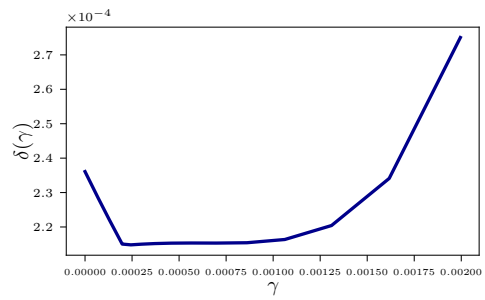

Figure 2: A numerical simulation of δ (defined in equation 6) as a function of γ.

3 Main privacy mechanism

3.1 The Dirichlet mechanism

Let $f : \mathcal{X}^N \to \mathbb{R}^d_{\geq 0}$ be an arbitrary vector-valued function with finite ℓ^2- and ℓ^∞-sensitivities: there exist two constants $\Delta_2, \Delta_\infty > 0$ such that

$$\sup_{x,x' \text{ neighboring}} \|f(x) - f(x')\|_2^2 \leq \Delta_2^2 \quad \text{and} \quad \sup_{x,x' \text{ neighboring}} \|f(x) - f(x')\|_\infty \leq \Delta_\infty.$$

Algorithm 1 below details the Dirichlet mechanism used to privatize $x \in \mathcal{X}^N$.

Algorithm 1 (λ, ε)**-RDP Dirichlet mechanism**

Input: A dataset $x \in \mathcal{X}^N$, A vector-valued function $f : \mathcal{X}^N \to \mathbb{R}^d_{\geq 0}$ with ℓ^2-sensitivity Δ_2 and ℓ^∞-sensitivity Δ_∞

Parameters: $\lambda \geq 1$, $\varepsilon > 0$

1. Use a root-finding algorithm to find $r > 0$ such that $\varepsilon = \frac{1}{2}\lambda r^2 \Delta_2^2 \psi'(1 + 3(\lambda - 1) r \Delta_\infty)$.
2. Let $\alpha = 1 + 4(\lambda - 1) r \Delta_\infty$.
3. Output $y \sim \text{Dirichlet}(rf(x) + \alpha)$.

The following lemma ensures that we can obtain an $r > 0$ in Line 1 for any $\varepsilon > 0$:

Lemma 2. *With $\varepsilon, \Delta_2 > 0, \Delta_\infty > 0$ and $\lambda \geq 1$ held constant, the function $r \mapsto \frac{1}{2}\lambda r^2 \Delta_2^2 \psi'(1 + 3(\lambda - 1) r \Delta_\infty)$ defined on $(0, \infty)$ is strictly increasing from 0 to ∞. Consequently, the equation*

$$\varepsilon = \frac{1}{2}\lambda r^2 \Delta_2^2 \psi'(1 + 3(\lambda - 1) r \Delta_\infty)$$

has a unique solution in r for any $\varepsilon, \Delta_2, \Delta_\infty > 0$ and $\lambda \geq 1$.

The proof of Lemma 2 can be found in Appendix D.

3.2 Privacy guarantee

Theorem 1. *Algorithm 1 is (λ, ε)-RDP.*

The proof of Theorem 1 can be found in Appendix E. A few remarks are in order.

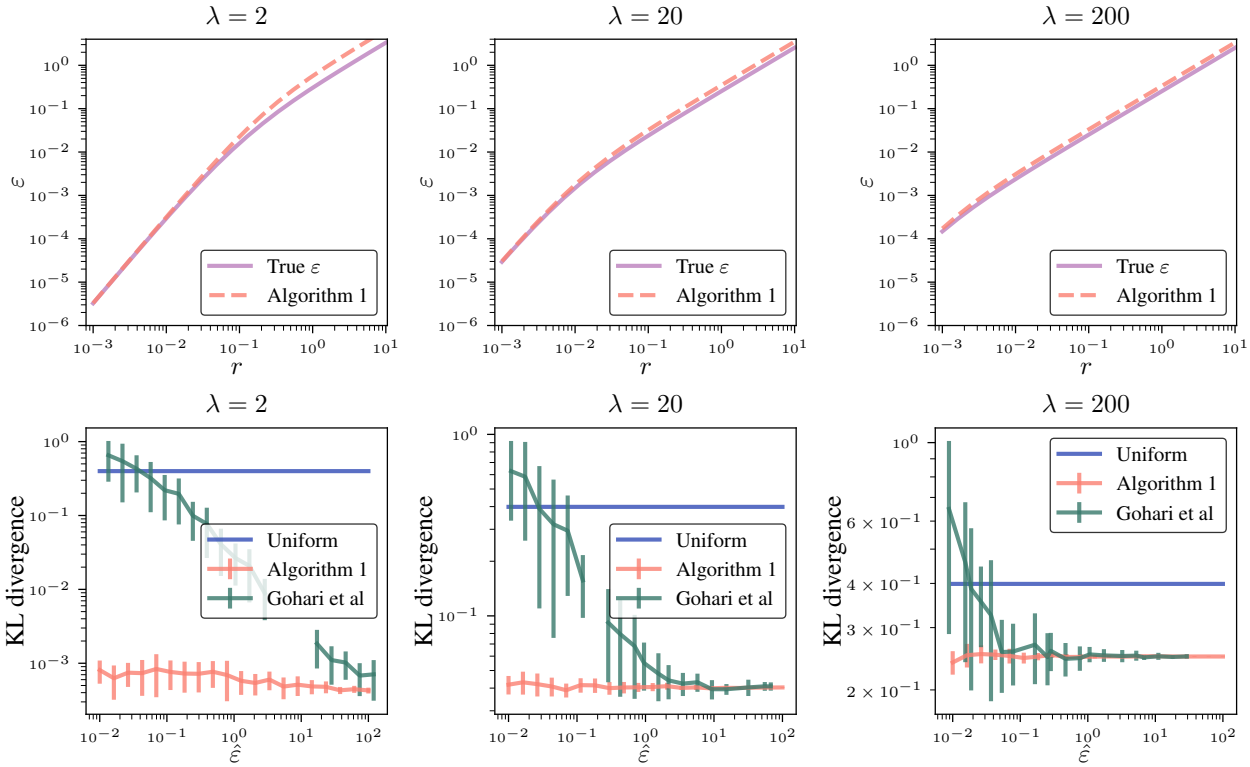

Figure 3: Top: Plots of the Rényi divergence (ε) between Dirichlet($rf(x)+\alpha$) and Dirichlet($rf(x')+\alpha$) using the direct calculations and Algorithm 1 as a function of r for $\lambda \in \{2, 20, 200\}$. Here, $f(x) = (11, 8, 65, 25, 38, 1)$, $f(x') = (11, 7, 65, 25, 38, 0)$ and $\alpha = 1 + 4(\lambda - 1)r$. Bottom: Plots of $D_{\mathrm{KL}}(y \| \widetilde{f(x)})$ for multiple instances of y drawn from Dirichlet($rf(x)+\alpha$), where $f(x) = (119, 74, 618, 272, 13, 187)$, $\alpha = 1 + 4(\lambda - 1)r$, and $\lambda \in \{2, 20, 200\}$. For each $\hat{\varepsilon}$, the privacy parameter r is chosen to satisfy $(\hat{\varepsilon}, 10^{-5})$-DP according to (1) the results of Gohari et al. (2021), and (2) the conversion from our RDP guarantee to approximate DP.

Remark 1. In general, we can replace $\psi'(1+3(\lambda-1)r\Delta_\infty)$ in Line 1 by $\psi'(1+g(r))$, and $\alpha = 1+4(\lambda-1)r\Delta_\infty$ in Line 2 by $\alpha = 1+g(r)+(\lambda-1)r\Delta_\infty$ for any function $g : \mathbb{R}_{>0} \to \mathbb{R}_{\geq 0}$. In particular, choosing $g \equiv 0$ yields $r = \sqrt{2\varepsilon/(\lambda\Delta_2^2\psi'(1))}$ which can be computed without a root-finding algorithm. However, this choice of r makes ε grows as r^2, which becomes too large when $r > 1$. Instead, we choose $g(r)$ to be a constant factor of an existing term $(\lambda-1)r\Delta_\infty$ in α, which allows us to offset the λr^2 factor in ε with $\psi'(1+g(r)) = \Theta\left(\frac{1}{1+(\lambda-1)r}\right)$.

Remark 2. If one has prior knowledge that $f_i(x) > b$ for some $b > 0$ for all $x \in \mathcal{X}^N$ and all $i \in [d]$, then the proof of Theorem 1 can be modified so that (λ, ε)-RDP can be obtained by setting r to be the solution to the equation $\varepsilon = \frac{1}{2}\lambda r^2 \Delta_2^2 \psi'(1 + rb + 3(\lambda - 1)r\Delta_\infty)$. Since ψ' is strictly decreasing, this leads to a larger value of r compared to Algorithm 1.

To demonstrate the tightness of the privacy guarantee of Algorithm 1, we simulate two neighboring histograms: $f(x) = (11, 8, 65, 25, 38, 1)$ and $f(x') = (11, 7, 65, 25, 38, 0)$. As functions of r, we compare ε in Line 1 with the analytic values of the Rényi divergence between Dirichlet($rf(x)+\alpha$) and Dirichlet($rf(x')+\alpha$), where α is given in Line 2. The plots of ε as functions of r in Figure 3 show that our proposed RDP-guarantees are close to the actual Rényi divergences across different values of λ.

We also perform another simulation in order to compare our privacy guarantees with the ones from Gohari et al. (2021) in terms of their effects on the KL divergence. In this simulation, we apply the Dirichlet mechanism with these privacy guarantees to the following count data: $f(x) = (119, 74, 618, 272, 13, 187)$.

For each $\lambda \in \{2, 20, 200\}$, we define $\alpha = 1 + 4(\lambda - 1)r$ as in Algorithm 1. Since the results of Gohari et al. are stated in terms of approximate DP, we have to convert our result from RDP to approximate DP (see Appendix B for more details on the conversion). For each $\hat{\varepsilon}$ ranging from 0.001 to 100, we use Theorem 1 (with the conversion) and Gohari et al.'s results to choose $r > 0$ so that a single draw from Dirichlet$(r\widetilde{f(x)}+\alpha)$ is $(\hat{\varepsilon}, 10^{-5})$-DP. We then draw multiple instances, say y, from the distribution and compute $D_{\mathrm{KL}}(\widetilde{f(x)}\|y)$. Finally, we plot the KL divergence as a function of $\hat{\varepsilon}$, as shown in Figure 3. As a baseline, we also plot the KL divergence between $\widetilde{f(x)}$ and the discrete uniform distribution. We can see that our privacy guarantee generally provides smaller KL divergences than that of Gohari et al.'s. However, as λ becomes very large, the algorithms output discrete probability distributions that are close to being uniform. The missing points in the $\lambda = 2$ and $\lambda = 20$ plots are related to a precision issue with the Gohari et al.'s method that we pointed out in Section 2.4: because of insufficient precision in numerical integration, we could not bring the value of δ down to 10^{-5}.

4 Utility

Let us recap the setting with which we apply the Dirichlet mechanism: we have a sensitive dataset $x \in \mathcal{X}^N$ and an arbitrary vector-valued function $f : \mathcal{X}^N \to \mathbb{R}^d_{\geq 0}$. Let $N_f \coloneqq \sum_i f_i(x)$ and $\widetilde{f(x)} \coloneqq N_f^{-1} f(x) \in S^{d-1}$. We propose the Dirichlet mechanism (Algorithm 1) which aims to output y that minimizes $D_{\mathrm{KL}}\Big(\widetilde{f(x)}\|y\Big)$ while keeping x private. This motivates us to measure the utility of the Dirichlet mechanism in terms of the KL divergence between $\widetilde{f(x)}$ and y. To this end, we can make use of the following bound:

Theorem 2. *For any $\alpha > 0$, $p = (p_1, \ldots, p_d) \in S^{d-1}$ and $q \sim$ Dirichlet$(\beta p + \alpha)$, the following inequality holds for any $\eta > 0$ and any $\beta \geq d\alpha/(e^{\eta/2} - 1)$:*

$$\Pr[D_{\mathrm{KL}}(p\|q) > \eta] \leq e^{-\beta\eta^2/(2(2+\eta)(4+3\eta))}.$$

The proof can be found in Appendix F. Since the Dirichlet mechanism outputs $y \sim$ Dirichlet$(rf(x) + \alpha) =$ Dirichlet$(rN_f\widetilde{f(x)} + \alpha)$, we can apply Theorem 2 with $p = f(x)$, $q = y$ and $\beta = rN_f$. As long as $N_f \geq d\alpha/\big(r(e^{\eta/2} - 1)\big)$, we have the bound

$$\Pr\Big[D_{\mathrm{KL}}\Big(\widetilde{f(x)}\|y\Big) > \eta\Big] \leq e^{-rN_f\eta^2/(2(2+\eta)(4+3\eta))}.$$

We shall assume that $\eta \ll 1$ and $\lambda \geq 2$. To obtain $D_{\mathrm{KL}}\Big(\widetilde{f(x)}\|y\Big) > \eta$ with high probability, one needs $N_f = \Omega\Big(\frac{1}{r\eta^2} + \frac{d\alpha}{r(e^{\eta/2}-1)}\Big)$. Now, we would like to write r and α in terms of ε and λ using the following identities from Algorithm 1.

$$\varepsilon = \frac{1}{2}\lambda r^2 \Delta_2^2 \psi'(1 + 3(\lambda - 1)r\Delta_\infty) \tag{7}$$

$$\alpha = 1 + 4(\lambda - 1)r\Delta_\infty. \tag{8}$$

We recall from Lemma 2 that the right-hand side of equation 7 is a strictly increasing function of r from 0 to ∞. This implies that, as $\varepsilon \to \infty$, we have $r \to \infty$. Under this limit, it follows from equation 5 that $\psi'(1+3(\lambda-1)r\Delta_\infty) = \Theta\Big(\frac{1}{(\lambda-1)r}\Big)$. Thus, equation 7 and 8 give $r = \Theta(\varepsilon)$ and $\alpha = \Theta((\lambda - 1)\varepsilon)$. On the other hand, as if $\varepsilon \to 0$, we have $r \to 0$ which implies $\psi'(1 + 3(\lambda - 1)r\Delta_\infty) = \Theta(1)$. Consequently, $r = \Theta(\sqrt{\varepsilon/\lambda})$ and $\alpha = \Theta(1)$. Therefore, to attain the (λ, ε)-RDP guarantee, one needs

$$N_f = \begin{cases} \Omega\Big(\frac{1}{\varepsilon\eta^2} + \frac{d(\lambda-1)}{e^{\eta/2}-1}\Big) & \text{if } \varepsilon \geq 1 \\ \Omega\Big(\sqrt{\frac{\lambda}{\varepsilon}}\Big[\frac{1}{\eta^2} + \frac{d}{e^{\eta/2}-1}\Big]\Big) & \text{if } \varepsilon < 1. \end{cases}$$

The most common example is when the data is categorical, that is, $x \in [d]^N$ and $f_i(x)$ is the number of i's in x. Then $N_f = \sum_i f_i(x) = N$, and the analysis above implies that the sample complexity for (λ, ε)-RDP and sub-η KL divergence, with λ and η fixed, is $N = \Omega\big(\frac{1}{\varepsilon} + 1\big)$ if $\varepsilon \geq 1$ and $N = \Omega\Big(\frac{1}{\sqrt{\varepsilon}}\Big)$ if $\varepsilon < 1$.

Table 1: UCI datasets used in the experiment

Dataset	#Instances	#Attributes	#Classes	%Positive	Source
CreditCard	30000	23	2	22%	Yeh & hui Lien (2009)
Thyroid	7200	21	3	–	Quinlan et al. (1986)
Shopper	12330	17	2	15%	Sakar et al. (2018)
Digit	5620	64	10	–	Garris et al. (1997)
GermanCredit	1000	20	2	30%	Grömping (2019)
Bank	41188	20	2	11%	Moro et al. (2014)
Spam	4601	57	2	39%	Cranor & LaMacchia (1998)
Adult	48842	13	2	24%	Kohavi (1996)

5 Experiments and discussions

5.1 Naïve Bayes classification

We consider the Dirichlet mechanism for differentially private multinomial naïve Bayes classification. Given a dataset $D = \{(x^{(i)}, y^{(i)})\}_{i=1}^N$, we construct a model to classify labels $y^{(i)} \in [d]$ from discrete features $x^{(i)} = (x_1^{(i)}, \ldots, x_K^{(i)}) \in \prod_{k=1}^K \mathcal{X}_k$, where $\mathcal{X}_1, \ldots, \mathcal{X}_K$ are finite sets. For $j \in [d]$, $k \in [K]$ and $c \in \mathcal{X}_k$, we denote the class count by $N_j \coloneqq \sum_{i=1}^N \mathbb{I}(y^{(i)} = j)$. For the k-th feature, we denote the feature-class count by $N_{jc}^k \coloneqq \sum_{i=1}^N \mathbb{I}(y^{(i)} = j, x_k^{(i)} = c)$. We can use the count data to estimate the class probabilities and the class-conditional feature probabilities:

$$\Pr[y = j] \coloneqq \hat{\pi}_j = N_j/N \quad \text{and} \quad \Pr[x_k = c | y = j] \coloneqq \hat{\theta}_{jc}^k = N_{jc}^k / N_j. \tag{9}$$

The naïve Bayes model assumes that, conditioning on the label, the features are independent. As a result, the probability of $y = j$ conditioned on $(x_1, \ldots, x_K)$ can be computed as follows:

$$\begin{aligned}
\Pr[y = j | x_1, \ldots, x_K] &\propto \Pr[y = j] \prod_{k=1}^K \Pr[x_k = c | y = j] \\
&= \frac{N_j}{N} \prod_{k=1}^K \frac{N_{jx_k}^k}{N_j} \\
&= \hat{\pi}_j \prod_{k=1}^K \hat{\theta}_{jx_k}^k.
\end{aligned}$$

To modify the model with the Dirichlet mechanism, we sample $(\tilde{\pi}_1, \ldots, \tilde{\pi}_d) \sim \text{Dirichlet}(r(N_1, \ldots, N_d) + \alpha)$, where r and α are chosen according to Algorithm 1 (with $\Delta_2^2 = 2$ and $\Delta_\infty = 1$) to attain $(\lambda, \varepsilon/K + 1)$-RDP. Similarly, for each $k \in K$ and $c \in \mathcal{X}_k$, we sample $(\tilde{\theta}_{1c}^k, \ldots, \tilde{\theta}_{dc}^k) \sim \text{Dirichlet}\big(r_c^k(N_{1c}^k, \ldots, N_{dc}^k) + \alpha_c^k\big)$, where r_c^k and α_c^k are chosen to attain $(\lambda, \varepsilon/(K+1))$-RDP as well. We then release $\tilde{\pi}_j$ instead of $\hat{\pi}_j$ and $\tilde{\theta}_{jc}^k$ instead of $\hat{\theta}_{jc}^k$ for all j, k and c, which leads to (λ, ε)-RDP by the basic composition (Lemma 1) and the parallel composition of RDP mechanisms

To benchmark the Dirichlet mechanism, we apply the Gaussian mechanism and the Laplace mechanism to the naïve Bayes model. Specifically, we replace N_j and N_{jc}^k in equation 9 by their noisy versions, namely $\tilde{N}_j \coloneqq N_j + z_j$ and $\tilde{N}_{jc}^k \coloneqq N_{jc}^k + z_{jc}^k$ where $z_j, z_{jc}^k \sim \mathcal{N}(0, \lambda(K+1)/\varepsilon)$ for the Gaussian mechanism and $z_j, z_{jc}^k \sim \text{Laplace}(0, b)$, where b is calculated using Mironov (2017, Corollary 2) to attain $(\lambda, \varepsilon/K)$-RDP for the Laplace mechanism.

In this experiment, the naïve Bayes models with differentially private mechanisms are used to classify 8 UCI datasets (Dua & Graff, 2017) with diverse number of instances/attributes/classes. The details of the datasets are shown in Table 1. For each dataset, we use a 70-30 train-test split. Before fitting the models,

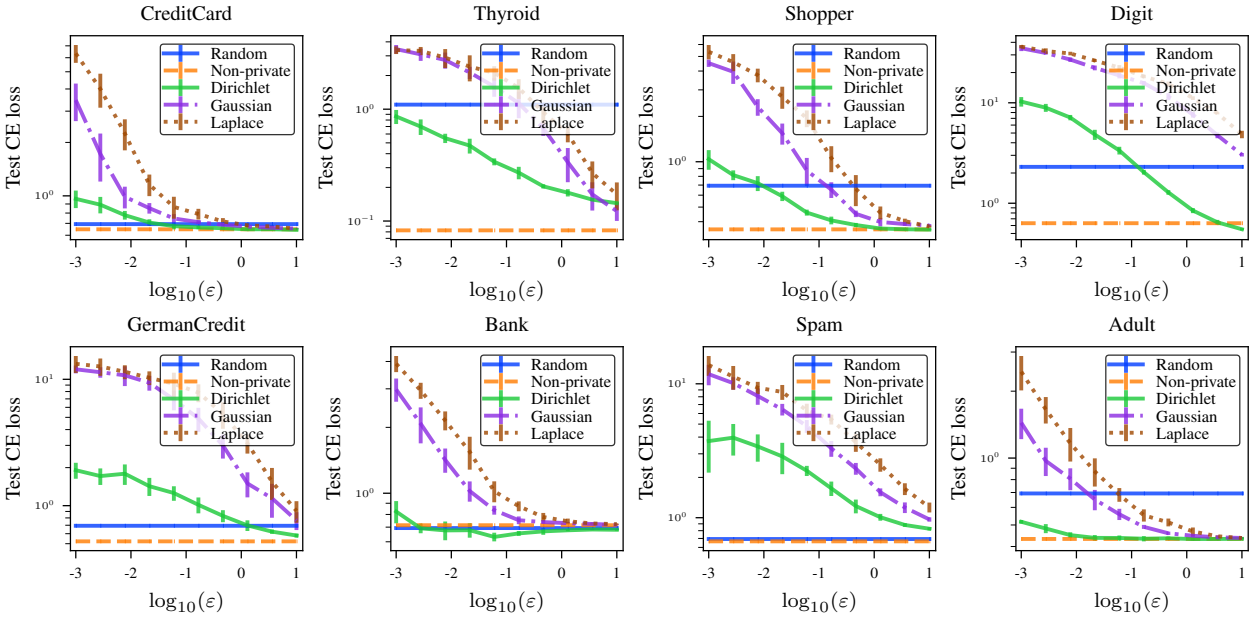

Figure 4: Test CE losses of the original and four $(5, \varepsilon)$-RDP naïve Bayes models on 8 UCI datasets.

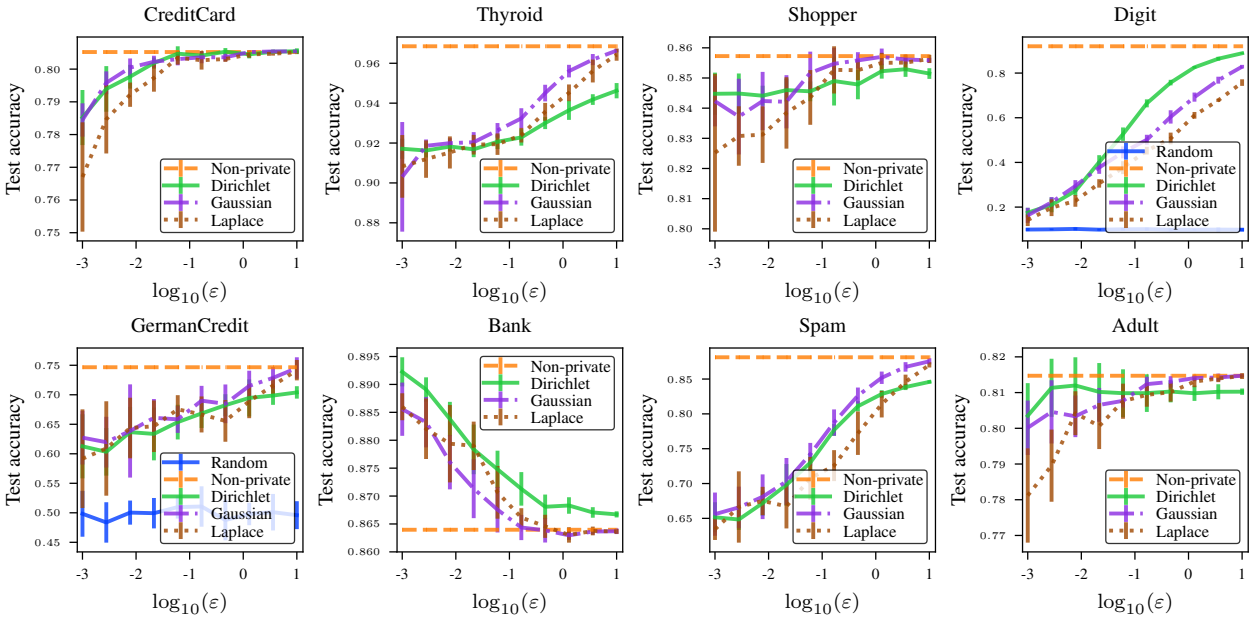

Figure 5: Test accuracies of the original and four $(5, \varepsilon)$-RDP naïve Bayes models on 8 UCI datasets. Plots of the random guessing on some datasets are not shown as its accuracies are well below the other models'.

numerical attributes are transformed into categorical ones using quantile binning, where the number of bins is fixed at 10.

For all privacy mechanisms, we fix $\lambda = 5$ and study their performances as ε increases from 10^{-3} to 10. We also add the random guessing model, which is a $(\lambda, 0)$-RDP model, as the baseline. The classification performances, measured in cross-entropy (CE) loss and accuracy on the test sets, are shown in Figure 4 and 5. We can see that, on all datasets, the test CE losses of the Dirichlet mechanism are substantially less than those of the Gaussian mechanism and Laplace mechanism; they are remarkably close to those

of the non-private model on the CreditCard, GermanCredit, Bank and Adult datasets. This result should not be surprising, as the Dirichlet mechanism is the exponential mechanism that aims to minimize the KL divergence, and thus the cross-entropy between the normalized counts and the parameters.

In terms of accuracy, there is no clear winner among the three mechanisms; the Dirichlet mechanism performs as well as the other mechanisms in most cases. Specifically, it has higher accuracies than the Gaussian mechanism on the Digit dataset for $\varepsilon > 0.1$, on the Adult dataset for $\varepsilon < 0.1$, and on the Bank dataset for all values of ε.

The difference between the two metrics stem from the fact that the cross entropy loss is a continuous function of the predicted probability, while the accuracy is a result of applying a hard threshold on the probability. Thus the accuracy does not distinguish between, for example, two instances, x, x' with $\Pr[y = 1|x] = 0.1$ and $\Pr[y = 1|x'] = 0.4$, but the CE loss will suffer almost three times as much when the true label of x is 1 compared to when the true label of x' is 1. Thus a model with high accuracy can have relatively low CE loss when they are too confident in their incorrect predictions.

All in all, neither metric is an end-all for measuring classification performance, and we should look at more than one metrics when fitting a model. If one wants to publish a naïve Bayes model under privacy constraint that performs well in both CE loss and accuracy, then the Dirichlet mechanism is an attractive option.

5.2 Parameter estimations of Bayesian networks

We use the Dirichlet mechanism for differentially private parameter estimations of discrete Bayesian networks. Consider a dataset $D = \{x^{(i)}\}_{i=1}^N$, where $x^{(i)} = (x_1^{(i)}, \ldots, x_K^{(i)}) \in \prod_{k=1}^K \mathcal{X}_k$ and $\mathcal{X}_1, \ldots, \mathcal{X}_K$ are finite sets. We name the K variables by their index: $1, \ldots, K$. Given a Bayesian network and $k \in [K]$, we denote the set of parents of k, that is, the set of direct causes of k by $Pa(k)$. Let $x_{Pa(k)}^{(i)} \coloneqq (x_\ell^{(i)})_{\ell \in Pa(k)}$ be observed values of $Pa(k)$ and $\mathcal{X}_{Pa(j)} \coloneqq \prod_{\ell \in Pa(j)} \mathcal{X}_\ell$ be the product space of $Pa(k)$. Given $j \in \mathcal{X}_k$ and $c \in \mathcal{X}_{Pa(k)}$, we denote $N_c^k \coloneqq \sum_{i=1}^N \mathbb{I}(x_{Pa(k)}^{(i)} = c)$ and $N_{jc}^k \coloneqq \sum_{i=1}^N \mathbb{I}(x_k^{(i)} = j, x_{Pa(k)}^{(i)} = c)$. The log-likelihood of the parameters $\theta_{jc}^k \coloneqq \Pr[x_k = j \mid x_{Pa(k)} = c]$ is given by:

$$LL(\theta) \coloneqq \sum_{k \in [K]} \sum_{\substack{j \in \mathcal{X}_k \\ c \in \mathcal{X}_{Pa(k)}}} N_{jc}^k \log \theta_{jc}^k. \tag{10}$$

Using the first-derivative test, the maximum-likelihood estimators of the Bayesian network are as follow:

$$\hat{\theta}_{jc}^k \coloneqq \frac{N_{jc}^k}{N_c^k}. \tag{11}$$

We can modify the model using the Dirichlet mechanism: assuming that $\mathcal{X}_k = [d]$, we replace $(\hat{\theta}_{1c}^k, \ldots, \hat{\theta}_{dc}^k)$ by $(\tilde{\theta}_{1c}^k, \ldots, \tilde{\theta}_{dc}^k) \sim \text{Dirichlet}\big(r(N_{1c}^k, \ldots, N_{dc}^k) + \alpha\big)$. Here, r and α are chosen according to Algorithm 1 to attain $(\lambda, \varepsilon/K)$-RDP. By the basic composition (Lemma 1) and the parallel composition, releasing $\tilde{\theta}_{jc}^k$ for all $k \in [K]$, $j \in \mathcal{X}_k$ and $c \in \mathcal{X}_{Pa(k)}$ is (λ, ε)-RDP.

We will compare the Dirichlet mechanism with the Gaussian and Laplace mechanisms. In equation 11, we replace N_{jc}^k by its noisy version: $\tilde{N}_{jc}^k \coloneqq N_{jc}^k + z_{jc}^k$, where $z_{jc}^k \sim \mathcal{N}(0, \lambda K/\varepsilon)$ for the Gaussian mechanism and $z_{jc}^k \sim \text{Laplace}(0, b)$, where b is calculated using Mironov (2017, Corollary 2) to attain $(\lambda, \varepsilon/K)$-RDP for the Laplace mechanism. In addition, we replace N_c^k by $\tilde{N}_c^k \coloneqq \sum_j \tilde{N}_{jc}^k$.

In this experiment, we have prepared Bayesian networks on the Adult, Bank and GermanCredit datasets, which are parts of full networks provided by Le Quy et al. (2022). The Bayesian networks are shown in Figure 6. As in the previous experiment, we use a 70-30 train-test split on each dataset, and continuous attributes are transformed into categorical attributes via quantile binning, with the number of bins fixed at 10.

For all privacy mechanisms, we fix $\lambda = 5$ and study their performances, in terms of the log-likelihoods of the privatized parameters on the test sets, as ε increases from 10^{-3} to 10. The plot of the log-likelihoods

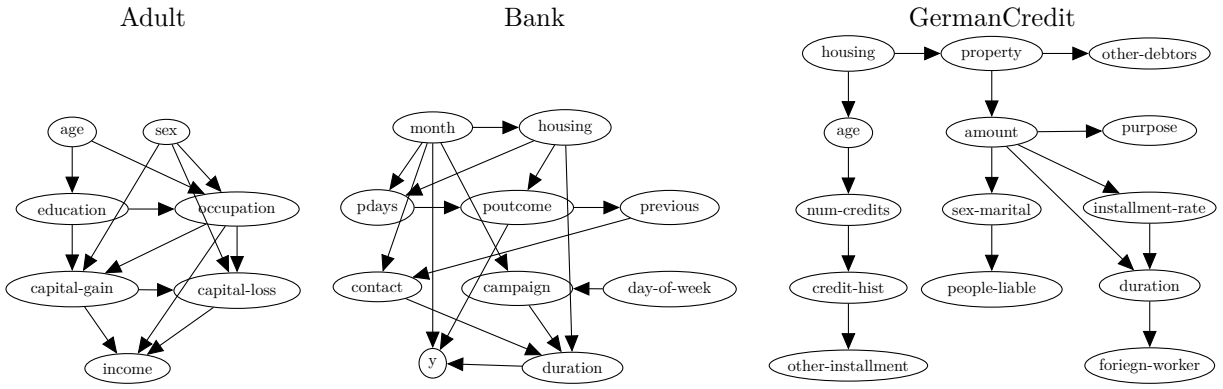

Figure 6: Our Bayesian networks on three datasets.

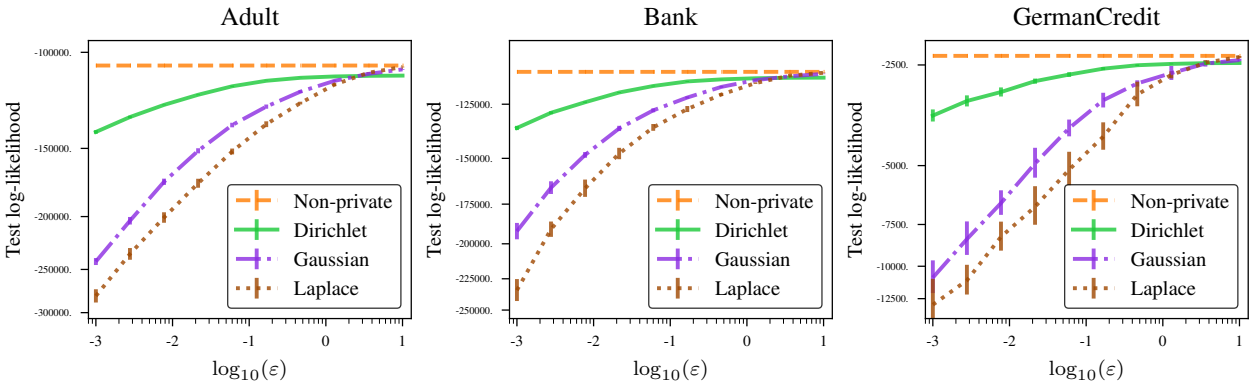

Figure 7: Test log-likelihoods of the parameters obtained from the maximum-likelihood estimation (non-private) and three $(5, \varepsilon)$-RDP mechanisms.

as functions of ε are shown in Figure 7. We can see that, on all datasets, the test log-likelihoods of the Dirichlet mechanism are substantially less than those of the Gaussian mechanism and Laplace mechanism for $\varepsilon < 1$. The results agree with our suggestion to use the Dirichlet mechanism for privacy-aware KL divergence minimization for discrete parameters, as it is equivalent to likelihood maximization.

6 Conclusion

The Dirichlet mechanism is an instance of the exponential mechanism whose loss function is the discrete KL divergence—this motivates us to use the Dirichlet mechanism for private estimation of an empirical distribution in KL divergence. As a consequence, the Dirichlet mechanism can be used for private likelihood maximization and cross–entropy minimization. This work provides a choice for the multiplicative factor r and the prior α that achieves a desired (λ, ε)-RDP guarantee. To demonstrate its efficiency, we compare our mechanism with the Gaussian and Laplace mechanisms for differentially private naïve Bayes classification, and as expected, the Dirichlet mechanism provides significantly lower cross-entropy losses on various datasets compared to the other two mechanisms. We also make a comparison between the mechanisms for maximum likelihood estimations for Bayesian networks. Our experiment on three datasets shows that the Dirichlet mechanism provides significantly higher log-likelihoods than the Gaussian and Laplace mechanisms.

As the KL divergence is a fundamental measure in information theory, we envision that the Dirichlet mechanism would become essential for many privacy-focused information-theoretic models with discrete parameters.

Broader Impact Statement

The Dirichlet mechanism does not provide privacy protection for free, but with a cost of some accuracy loss: the higher the privacy guarantee, the lower the accuracy of the privatized model compared to the original model. Any losses incurred from the inaccuracy must be taken into consideration before deploying the privatized model.

References

Martín Abadi, Andy Chu, Ian J. Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In Edgar R. Weippl, Stefan Katzenbeisser, Christopher Kruegel, Andrew C. Myers, and Shai Halevi (eds.), *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016*, pp. 308–318. ACM, 2016. doi: 10.1145/2976749.2978318. URL https://doi.org/10.1145/2976749.2978318.

Necdet Batir. Some new inequalities for gamma and polygamma functions. *Research report collection*, 7(3), 2004. URL https://vuir.vu.edu.au/17580/.

Garrett Bernstein and Daniel R. Sheldon. Differentially private bayesian inference for exponential families. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 2924–2934, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/08040837089cdf46631a10aca5258e16-Abstract.html.

Clément L. Canonne, Gautam Kamath, and Thomas Steinke. The discrete gaussian for differential privacy. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/b53b3a3d6ab90ce0268229151c9bde11-Abstract.html.

Clément L. Canonne, Ziteng Sun, and Ananda Theertha Suresh. Concentration Bounds for Discrete Distribution Estimation in KL Divergence. February 2023. doi: 10.48550/arXiv.2302.06869. URL https://arxiv.org/abs/2302.06869v1.

Lorrie Faith Cranor and Brian A. LaMacchia. Spam! *Commun. ACM*, 41(8):74–83, aug 1998. ISSN 0001-0782. doi: 10.1145/280324.280336. URL https://doi.org/10.1145/280324.280336.

Ilias Diakonikolas, Moritz Hardt, and Ludwig Schmidt. Differentially Private Learning of Structured Discrete Distributions. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://papers.nips.cc/paper/2015/hash/2b3bf3eee2475e03885a110e9acaab61-Abstract.html.

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.*, 18:11:1–11:39, 2017. URL http://jmlr.org/papers/v18/15-257.html.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014. doi: 10.1561/0400000042. URL https://doi.org/10.1561/0400000042.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In Serge Vaudenay (ed.), *Advances in Cryptology -*

EUROCRYPT 2006, 25th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, Russia, May 28 - June 1, 2006, Proceedings, volume 4004 of *Lecture Notes in Computer Science*, pp. 486–503. Springer, 2006a. doi: 10.1007/11761679_29. URL https://doi.org/10.1007/11761679_29.

Cynthia Dwork, Frank Mcsherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, 2006b.

James R. Foulds, Joseph Geumlek, Max Welling, and Kamalika Chaudhuri. On the theory and practice of privacy-preserving bayesian data analysis. In Alexander T. Ihler and Dominik Janzing (eds.), *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence, UAI 2016, June 25-29, 2016, New York City, NY, USA*. AUAI Press, 2016. URL http://auai.org/uai2016/proceedings/papers/45.pdf.

Michael Garris, J Blue, Gerald Candela, Patrick Grother, Stanley Janet, and Charles Wilson. Nist form-based handprint recognition system, 1997-01-01 1997.

Joseph Geumlek, Shuang Song, and Kamalika Chaudhuri. Renyi differential privacy mechanisms for posterior sampling. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5289–5298, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/56584778d5a8ab88d6393cc4cd11e090-Abstract.html.

Parham Gohari, Bo Wu, Calvin Hawkins, Matthew T. Hale, and Ufuk Topcu. Differential privacy on the unit simplex via the dirichlet mechanism. *IEEE Trans. Inf. Forensics Secur.*, 16:2326–2340, 2021. doi: 10.1109/TIFS.2021.3052356. URL https://doi.org/10.1109/TIFS.2021.3052356.

Ulrike Grömping. South german credit data: Correcting a widely used data set. Reports in mathematics, physics and chemistry, Department II, Beuth University of Applied Sciences Berlin, 4 2019.

Abdolhossein Hoorfar and Mehdi Hassani. Inequalities on the lambert w function and hyperpower function. *J. Inequal. Pure and Appl. Math*, 9(2):5–9, 2008.

Sudeep Kamath, Alon Orlitsky, Dheeraj Pichapati, and Ananda Theertha Suresh. On Learning Distributions from their Samples. In *Proceedings of The 28th Conference on Learning Theory*, pp. 1066–1100. PMLR, June 2015. URL https://proceedings.mlr.press/v40/Kamath15.html.

Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pp. 202–207. AAAI Press, 1996.

Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. A survey on datasets for fairness-aware machine learning. *WIREs Data Mining and Knowledge Discovery*, 12(3):e1452, 2022. doi: https://doi.org/10.1002/widm.1452. URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1452.

Ashwin Machanavajjhala, Daniel Kifer, John M. Abowd, Johannes Gehrke, and Lars Vilhuber. Privacy: Theory meets practice on the map. In Gustavo Alonso, José A. Blakeley, and Arbee L. P. Chen (eds.), *Proceedings of the 24th International Conference on Data Engineering, ICDE 2008, April 7-12, 2008, Cancún, Mexico*, pp. 277–286. IEEE Computer Society, 2008. doi: 10.1109/ICDE.2008.4497436. URL https://doi.org/10.1109/ICDE.2008.4497436.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2007), October 20-23, 2007, Providence, RI, USA, Proceedings*, pp. 94–103. IEEE Computer Society, 2007. doi: 10.1109/FOCS.2007.41. URL https://doi.org/10.1109/FOCS.2007.41.

Ilya Mironov. Rényi differential privacy. In *30th IEEE Computer Security Foundations Symposium, CSF 2017, Santa Barbara, CA, USA, August 21-25, 2017*, pp. 263–275. IEEE Computer Society, 2017. doi: 10.1109/CSF.2017.11. URL https://doi.org/10.1109/CSF.2017.11.

Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, June 2014. doi: 10.1016/j.dss.2014.03.001. URL https://doi.org/10.1016/j.dss.2014.03.001.

John Ross Quinlan, Paul J Compton, KA Horn, and Leslie Lazarus. Inductive knowledge acquisition: a case study. In *Proceedings of the second Australian Conference on the Applications of Expert Systems*, pp. 183–204, 1986.

Alfréd Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.

C. Okan Sakar, S. Olcay Polat, Mete Katircioglu, and Yomi Kastro. Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. *Neural Computing and Applications*, 31(10):6893–6908, May 2018. doi: 10.1007/s00521-018-3523-0. URL https://doi.org/10.1007/s00521-018-3523-0.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pp. 3–18. IEEE Computer Society, 2017. doi: 10.1109/SP.2017.41. URL https://doi.org/10.1109/SP.2017.41.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432, August 2016. doi: 10.1007/s11222-016-9696-4. URL https://doi.org/10.1007/s11222-016-9696-4.

Yu-Xiang Wang, Stephen E. Fienberg, and Alexander J. Smola. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In Francis R. Bach and David M. Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 2493–2502. JMLR.org, 2015. URL http://proceedings.mlr.press/v37/wangg15.html.

Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models. In Heng Yin, Angelos Stavrou, Cas Cremers, and Elaine Shi (eds.), *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, CCS 2022, Los Angeles, CA, USA, November 7-11, 2022*, pp. 3093–3106. ACM, 2022. doi: 10.1145/3548606.3560675. URL https://doi.org/10.1145/3548606.3560675.

I-Cheng Yeh and Che hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2):2473–2480, March 2009. doi: 10.1016/j.eswa.2007.12.020. URL https://doi.org/10.1016/j.eswa.2007.12.020.

A Dirichlet posterior sampling is not ε-differentially private

We show that the Dirichlet posterior sampling does not satisfy the original notion of differential privacy—the pure differential privacy.

Proposition 3. *For any $\varepsilon > 0$, the mechanism that outputs $y \sim \mathrm{Dirichlet}(rf(x) + \alpha)$ is not ε-differentially private.*

Proof. Without loss of generality, let $x = (0, 0, \ldots, 0)$ and $x' = (1, 0, \ldots, 0)$. Let $\alpha > 0$ be any positive number. Let $y \sim \mathrm{Dirichlet}(rf(x) + \alpha)$ and $y' \sim \mathrm{Dirichlet}(rf(x') + \alpha)$. For any $y_0 = (y_1, y_2, \ldots, y_d)$ with

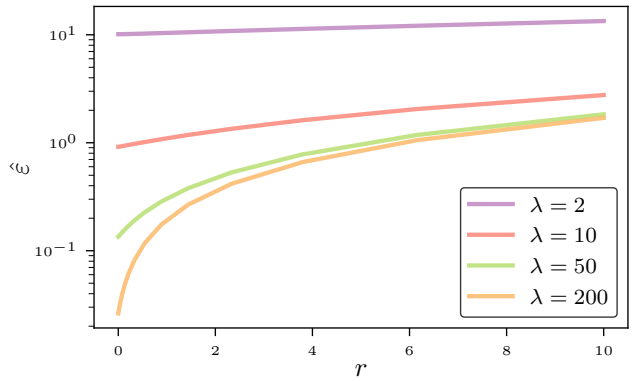

Figure 8: (ε, δ)-DP guarantees of the Dirichlet mechanism following equation 13 with $\lambda \in \{2, 10, 50, 200\}$ and $\delta = 10^{-5}$.

$\sum_i y_i = 1$, we have

$$\begin{aligned}\frac{\Pr[y = y_0]}{\Pr[y' = y_0]} &= \frac{B(rf(x') + \alpha)}{B(rf(x) + \alpha)} \cdot \frac{\prod_i y_i^{rf_i(x)+\alpha}}{\prod_i y_i^{rf_i(x')+\alpha}} \\ &= \frac{B(rf(x') + \alpha)}{B(rf(x) + \alpha)} \cdot \frac{1}{y_1}.\end{aligned}$$

For any $\varepsilon > 0$, we can choose a sufficiently small $y_1 > 0$ so that the right-hand side is larger than e^{ε}. □

B Approximate differential privacy

We can convert from RDP to approximate DP with the following conversion formula:

Lemma 3 (From RDP to Approximate DP (Canonne et al., 2020))**.** *Let $\varepsilon > 0$. If M is a (λ, ε)-RDP mechanism, then it also satisfies $(\hat{\varepsilon}, \delta)$-DP with*

$$\delta = \frac{\exp((\lambda - 1)(\varepsilon - \hat{\varepsilon}))}{\lambda - 1}\left(1 - \frac{1}{\lambda}\right)^{\lambda}. \tag{12}$$

Taking the logarithm of equation 12,

$$\log \delta = (\lambda - 1)(\varepsilon - \hat{\varepsilon}) + (\lambda - 1)\log(\lambda - 1) - \lambda \log(\lambda),$$

which is equivalent to

$$\hat{\varepsilon} = \varepsilon + \log(\lambda - 1) - \frac{\log \delta + \lambda \log(\lambda)}{\lambda - 1}.$$

Plugging in the RDP guarantee in Algorithm 1, we obtain

$$\hat{\varepsilon} = \frac{1}{2}\lambda r^2 \Delta_2^2 \psi'(1 + 3(\lambda - 1) r \Delta_\infty) + \log(\lambda - 1) - \frac{\log \delta + \lambda \log(\lambda)}{\lambda - 1}, \tag{13}$$

which gives a formula for $\hat{\varepsilon}$ in terms of r, λ and δ. Figure 8 shows $\hat{\varepsilon}$ as a function of r at four different values of λ. We can see that, at a fixed δ, $\hat{\varepsilon}$ is increased when we increase r and decrease λ.

C Experiments with approximate DP

We perform the same experiments as those in Section 5. But this time, we focus on approximate DP instead of RDP, and we also include the Dirichlet mechanism with Gohari et al. (2021)'s privacy guarantee in the

experiments. Our (λ, ε)-RDP guarantee of the Dirichlet mechanism is converted to $(\hat{\varepsilon}, \delta)$-DP guarantee, with $\delta = 10^{-5}$, using the material in Section sec:adp. The results of the naïve Bayes and Bayesian network experiments are shown in Figure 9 and Figure 10, and those of the Bayesian networks are shown in Figure 11.

Aside from similar results as those in Section 5, We highlight that our Dirichlet mechanism performs better than Gohari et al.'s in all experiments, and Gohari et al.'s mechanism performs significantly worse for smaller values of $\hat{\varepsilon}$. We also notice that, in contrast to the results in Section 5 the Laplace mechanism performs better than the Gaussian mechanism; this is because the composition property for multiple uses of an $\hat{\varepsilon}$ DP mechanism is better than that of an $(\hat{\varepsilon}, \delta)$-DP for any $\delta > 0$ (see Dwork & Roth (2014, Theorem 3.20)).

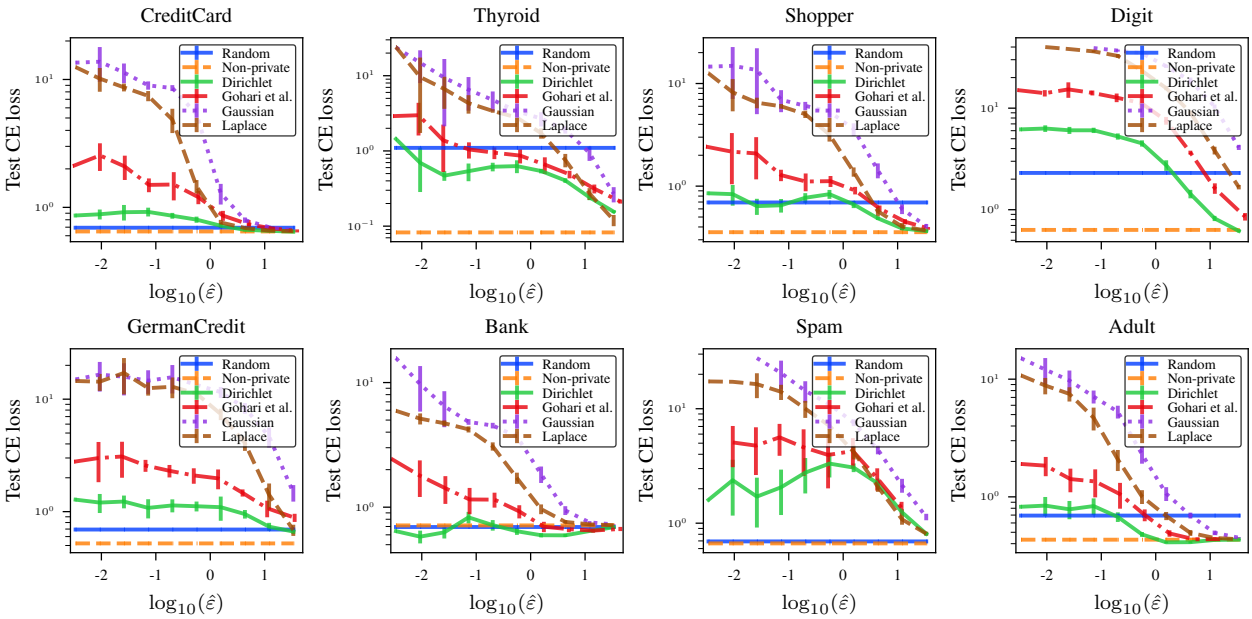

Figure 9: Test CE losses of the original and five $(\hat{\varepsilon}, 10^{-5})$-RDP naïve Bayes models on 8 UCI datasets.

D Proof of Lemma 2

Denote $x = 3(\lambda - 1)r\Delta_\infty$. With $\varepsilon, \lambda, \Delta_2$ and Δ_∞ fixed as constants, we can write the equation as $\varepsilon = Cx^2\psi'(1+x)$ for some constant $C > 0$. From equation 5, we have $\psi'(1+x) = \Theta\left(\frac{1}{(1+x)^2}\right)$ as $x \to 0$ and $\psi'(x) = \Theta\left(\frac{1}{1+x}\right)$ as $x \to \infty$. Consequently,

$$\lim_{x\to 0} x^2\psi'(1+x) = 0 \quad \text{and} \quad \lim_{x\to\infty} x^2\psi'(1+x) = \infty. \tag{14}$$

The conclusion will follow if we can show that the function $\phi(x) := x^2\psi'(1+x)$ is strictly increasing. For this, first we use $\psi'(1+x) < \frac{1}{1+x} + \frac{1}{(1+x)^2}$ to obtain

$$[\psi'(1+x)]^2 < \frac{\psi'(1+x)}{1+x} + \frac{\psi'(1+x)}{(1+x)^2} \leq \frac{2\psi'(1+x)}{1+x} < \frac{2\psi'(1+x)}{x}.$$

In other words, $2\psi'(1+x) > x[\psi'(1+x)]^2$. Combining this with $[\psi'(x)]^2 + \psi''(x) > 0$ (see e.g. Batir (2004, Lemma1.1)), we have

$$\phi'(x) = 2x\psi'(1+x) + x^2\psi''(1+x) > x^2[\psi'(1+x)]^2 + x^2\psi''(1+x) = x^2\big([\psi'(x)]^2 + \psi''(x)\big) > 0.$$

Therefore, $\phi(x)$ is strictly increasing, which, combined with equation 14, implies that the equation $\phi(x) = \varepsilon$ has a unique solution x_ε for any $\varepsilon > 0$. We then obtain a solution in r by letting $r = x_\varepsilon/(3(\lambda-1)\Delta_\infty)$.

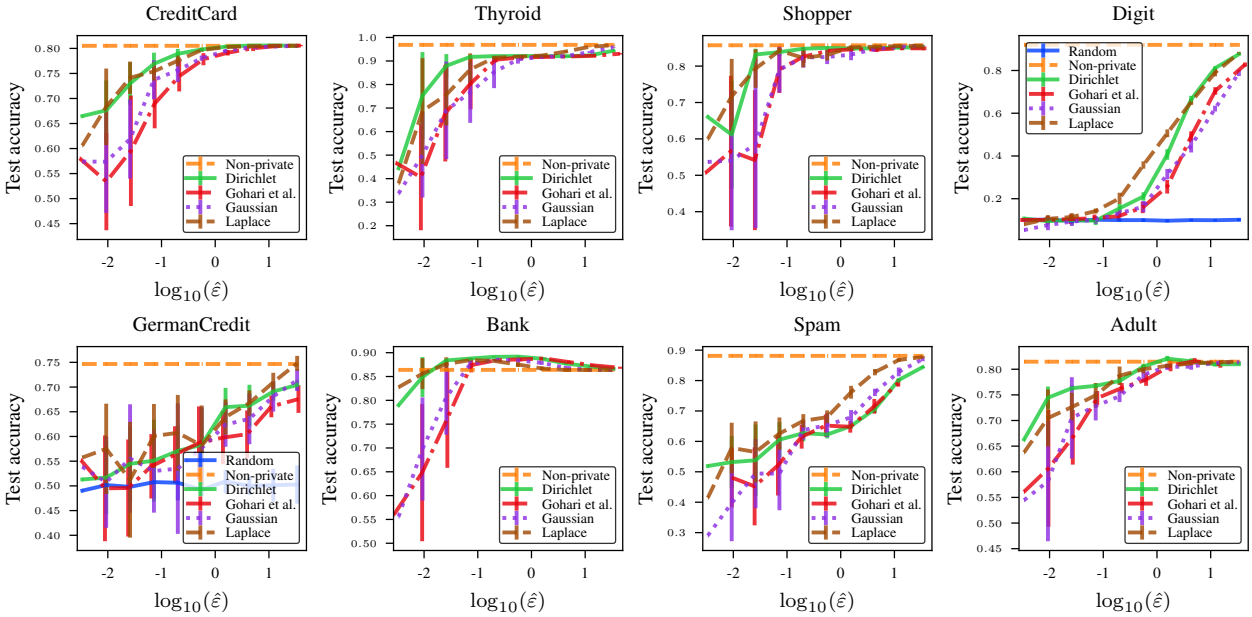

Figure 10: Test accuracies of the original and five $(\hat{\varepsilon}, 10^{-5})$-DP naïve Bayes models on 8 UCI datasets. Plots of the random guessing on some datasets are not shown as its accuracies are well below the other models'.

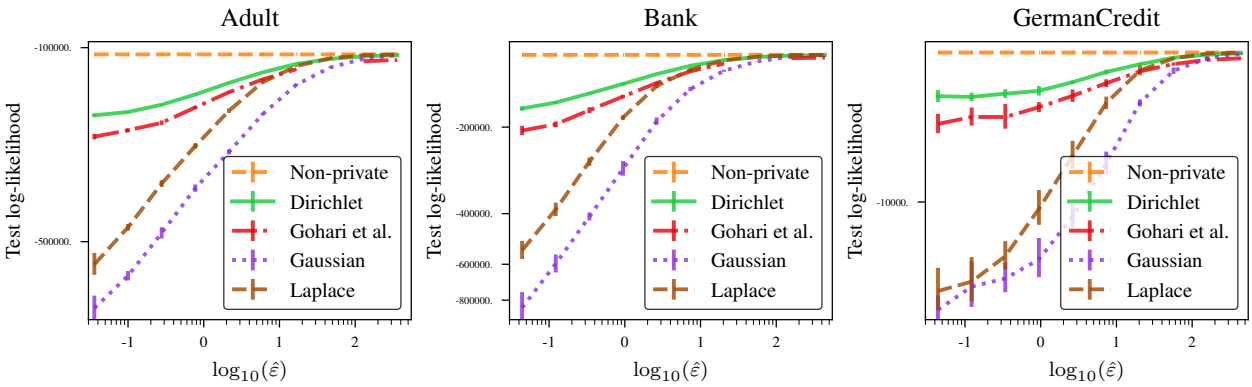

Figure 11: Test log-likelihoods of the parameters obtained from the maximum-likelihood estimation (non-private) and four $(\hat{\varepsilon}, 10^{-5})$-DP mechanisms.

E Proof of Theorem 1

Case 1: $\lambda > 1$.

Let x and x' be neighboring datasets. For notational convenience, let $u \coloneqq rf(x) + \alpha$ and $u' \coloneqq rf(x') + \alpha$. As usual, we write $u = (u_1, \ldots, u_d)$, $u' = (u'_1, \ldots, u'_d)$, $u_0 \coloneqq \sum_i u_i$ and $u'_0 \coloneqq \sum_i u'_i$. Let $P(y)$ be the density of Dirichlet(u) and $P'(y)$ be the density of Dirichlet(u'). To compute the Rényi divergence between $P(y)$ and $P'(y)$, we start with:

$$\begin{aligned}\mathbb{E}_{y\sim P(y)}\left[\frac{P(y)^{\lambda-1}}{P'(y)^{\lambda-1}}\right] &= \frac{B(u')^{\lambda-1}}{B(u)^{\lambda-1}}\mathbb{E}_{y\sim P(y)}\left[y^{(\lambda-1)(u-u')}\right] \\ &= \frac{B(u')^{\lambda-1}}{B(u)^{\lambda-1}} \cdot \frac{B(u+(\lambda-1)(u-u'))}{B(u)}, \end{aligned} \tag{15}$$

where $B(u) = \Gamma(u_0)^{-1}\prod_i \Gamma(u_i)$ is the multivariate beta function. Thus the ratio can be expressed in terms of gamma functions:

$$\frac{B(u')}{B(u)} = \frac{\prod_i \Gamma(u'_i)/\Gamma(\sum_i u'_i)}{\prod_i \Gamma(u_i)/\Gamma(\sum_i u_i)} = \frac{\Gamma(u_0)}{\Gamma(u'_0)}\prod_i \frac{\Gamma(u'_i)}{\Gamma(u_i)},$$

where $u_0 \coloneqq \sum_i u_i$ and $u'_0 \coloneqq \sum_i u'_i$. Similarly,

$$\frac{B(u+(\lambda-1)(u-u'))}{B(u)} = \frac{\Gamma(\sum_i u_i)}{\Gamma(\sum_i u_i + (\lambda-1)\sum_i(u_i-u'_i))}\prod_i \frac{\Gamma(u_i+(\lambda-1)(u_i-u'_i))}{\Gamma(u_i)}.$$

Taking the logarithm on both side of equation 15, we need to find an upper bound of:

$$\log \mathbb{E}_{y\sim P(y)}\left[\frac{P(y)^{\lambda-1}}{P'(y)^{\lambda-1}}\right] = \sum_i (G(u_i,u'_i) + H(u_i,u'_i)) - G(u_0,u'_0) - H(u_0,u'_0), \tag{16}$$

where

$$\begin{aligned} G(u_i,u'_i) &\coloneqq (\lambda-1)(\log\Gamma(u'_i) - \log\Gamma(u_i)) \\ H(u_i,u'_i) &\coloneqq \log\Gamma(u_i+(\lambda-1)(u_i-u'_i)) - \log\Gamma(u_i), \end{aligned}$$

and similarly for $G(u_0,u'_0)$ and $H(u_0,u'_0)$. Using the second-order Taylor expansion, there exists ξ between $u_i+(\lambda-1)(u_i-u'_i)$ and u_i, and ξ' between u_i and u'_i such that

$$\begin{aligned} G(u_i,u'_i) &= -(\lambda-1)(u_i-u'_i)\psi(u_i) + \frac{1}{2}(\lambda-1)(u_i-u'_i)^2\psi'(\xi') \\ &= -(\lambda-1)(f_i(x)-f_i(x'))r\psi(u_i) + \frac{1}{2}(\lambda-1)(f_i(x)-f_i(x'))^2r^2\psi'(\xi') \\ H(u_i,u'_i) &= (\lambda-1)(u_i-u'_i)\psi(u_i) + \frac{1}{2}(\lambda-1)^2(u_i-u'_i)^2\psi'(\xi) \\ &= (\lambda-1)(f_i(x)-f_i(x'))r\psi(u_i) + \frac{1}{2}(\lambda-1)^2(f_i(x)-f_i(x'))^2r^2\psi'(\xi). \end{aligned}$$

We try to find an upper bound of both $\psi'(\xi)$ and $\psi'(\xi')$. If $f_i(x) > f_i(x')$, then $u'_i < u_i < u_i+(\lambda-1)(u_i-u'_i)$. Thus both ξ and ξ' are bounded below by $u'_i \geq \alpha$. On the other hand, if $f_i(x) \leq f_i(x')$, then $u_i + (\lambda-1)(u_i-u'_i) \leq u_i \leq u'_i$. In this case, ξ and ξ' are bounded below by:

$$\begin{aligned} u_i+(\lambda-1)(u_i-u'_i) &= f_i(x)+\alpha-(\lambda-1)(rf_i(x')-rf_i(x)) \\ &\geq \alpha-(\lambda-1)r\Delta_\infty. \end{aligned}$$

Since ψ' is decreasing, both $\psi'(\xi)$ and $\psi'(\xi')$ are bounded above by $\psi'(\alpha-(\lambda-1)r\Delta_\infty)$. Consequently,

$$\begin{aligned} G(u_i,u'_i) + H(u_i,u'_i) &\leq \frac{1}{2}\big((\lambda-1)+(\lambda-1)^2\big)(f_i(x)-f_i(x'))^2r^2\psi'(\alpha-(\lambda-1)r\Delta_\infty) \\ &= \frac{1}{2}\lambda(\lambda-1)(f_i(x)-f_i(x'))^2r^2\psi'(\alpha-(\lambda-1)r\Delta_\infty). \end{aligned}$$

The same argument can be used to show that, there exist ξ_0 and ξ_0' such that:

$$G(u_0, u_0') + H(u_0, u_0') = \frac{1}{2}(\lambda - 1)(u_0 - u_0')^2\psi'(\xi_0') + \frac{1}{2}(\lambda - 1)^2(u_0 - u_0')^2\psi'(\xi_0) > 0.$$

Therefore, continuing from equation 16,

$$\begin{aligned} D_\lambda(P(y)\|P'(y)) &= \frac{1}{\lambda - 1}\left(\sum_i (G(u_i, u_i') + H(u_i, u_i')) - G(u_0, u_0') - H(u_0, u_0')\right) \\ &< \frac{1}{\lambda - 1}\sum_i (G(u_i, u_i') + H(u_i, u_i')) \\ &\leq \frac{1}{2}\lambda \sum_i (f_i(x_i) - f_i(x_i'))^2 r^2 \psi'(\alpha - (\lambda - 1)r\Delta_\infty) \\ &\leq \frac{1}{2}\lambda \Delta_2^2 r^2 \psi'(\alpha - (\lambda - 1)r\Delta_\infty). \end{aligned} \tag{17}$$

Case 2: $\lambda = 1$**.**

We use the following formula for the KL divergence between two Dirichlet distributions:

$$\begin{aligned} D_{\mathrm{KL}}(P(y)\|P'(y)) = &\log\Gamma(u_0) - \sum_i \log\Gamma(u_i) - \log\Gamma(u_0') \\ &+ \sum_i \log\Gamma(u_i') + \sum_i (u_i - u_i')(\psi(u_i) - \psi(u_0)), \end{aligned}$$

From this, we split the right-hand side into two parts and apply the Taylor approximation as before:

$$\begin{aligned} -\sum_i \log\Gamma(u_i) + \sum_i \log\Gamma(u_i') + \sum_i (u_i - u_i')\psi(u_i) &\leq \frac{1}{2}\sum_i (u_i - u_i')^2\psi'(\min\{u_i, u_i'\}) \\ &\leq \frac{1}{2}\sum_i (u_i - u_i')^2\psi'(1) \\ &= \frac{1}{2}\sum_i (f_i(x_i) - f_i(x_i'))^2 r^2\psi'(1) \\ &\leq \frac{1}{2}\Delta_2^2 r^2\psi'(1), \end{aligned}$$

and

$$\begin{aligned} \log\Gamma(u_0) - \log\Gamma(u_0') - \sum_i (u_0 - u_0')\psi(u_0) &\leq -\frac{1}{2}\sum_i (u_i - u_i')^2\psi'(\max\{u_0, u_0'\}) \\ &\leq 0. \end{aligned}$$

Adding these two inequalities yields the same inequality as equation 17 with $\lambda = 1$.

Thus, given any $\lambda \geq 1$, $\varepsilon > 0$ and any $g : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$, if we let r be the solution of $\frac{1}{2}\lambda r^2 \Delta_2^2 \psi'(1 + g(r)) = \varepsilon$ and $\alpha = 1 + g(r) + (\lambda - 1)r\Delta_\infty$, then the inequality above implies $D_\lambda(P(y)\|P'(y)) < \varepsilon$. We conclude that Algorithm 1 by setting $g(r) = 3(\lambda - 1)r\Delta_\infty$.

F Proof of the Utility bound

We first note a pair of inequalities for the digamma function, which hold for all $x > \frac{1}{2}$:

$$\log\left(x - \frac{1}{2}\right) < \psi(x) < \log x. \tag{18}$$

We start with the Chernoff bound: for any $t \leq \beta$,

$$\begin{aligned}
\Pr[D_{\mathrm{KL}}(p\|q) > \eta] &\leq e^{-t\eta}\mathbb{E}\left[e^{tD_{\mathrm{KL}}(p\|q)}\right] \\
&= e^{-t\eta}\mathbb{E}\left[\prod_i (p_i/q_i)^{tp_i}\right] \\
&= e^{-t\eta}\prod_i p_i^{tp_i}\mathbb{E}\left[\prod_i q_i^{-tp_i}\right] \\
&= e^{-t\eta}\prod_i p_i^{tp_i}\frac{1}{B(\beta p+\alpha)}\int\prod_i q_i^{\beta p_i - tp_i + \alpha - 1}\,dq \\
&= e^{-t\eta}\prod_i p_i^{tp_i}\frac{B(\beta p - tp_i + \alpha)}{B(\beta p+\alpha)} \\
&= e^{-t\eta}\frac{\Gamma(\beta+d\alpha)}{\Gamma(\beta-t+d\alpha)}\prod_i p_i^{tp_i}\frac{\Gamma(\beta p_i - tp_i + \alpha)}{\Gamma(\beta p_i+\alpha)}. \qquad (19)
\end{aligned}$$

Using the first-order Taylor approximation, we have the following estimates for log-gamma functions:

$$\begin{aligned}
\log\Gamma(\beta+d\alpha) &\leq \log\Gamma(\beta-t+d\alpha) + t\psi(\beta+d\alpha) \\
\log\Gamma(\beta p_i - tp_i + \alpha) &\leq \log\Gamma(\beta p_i + d\alpha) - tp_i\psi(\beta p_i - tp_i + \alpha).
\end{aligned}$$

Inserting these inequalities and equation 18 into equation 19, we obtain

$$\begin{aligned}
\Pr[D_{\mathrm{KL}}(p\|q) > \eta] &\leq e^{-t\eta}e^{t\psi(\beta+d\alpha)}\prod_i p_i^{tp_i}e^{-tp_i\psi(\beta p_i - tp_i+\alpha)} \\
&< e^{-t\eta}e^{t\log(\beta+d\alpha)}\prod_i p_i^{tp_i}e^{-tp_i\log(\beta p_i - tp_i+\alpha-1/2)} \\
&= e^{-t\eta}(\beta+d\alpha)^t\prod_i p_i^{tp_i}(\beta p_i - tp_i+\alpha-1/2)^{-tp_i} \\
&= e^{-t\eta}(\beta+d\alpha)^t\prod_i\left(\beta - t + p_i^{-1}(\alpha-1/2)\right)^{-tp_i} \\
&= e^{-t\eta}\prod_i\left(\frac{\beta+d\alpha}{\beta-t+p_i^{-1}(\alpha-1/2)}\right)^{tp_i} \\
&< e^{-t\eta}\prod_i\left(\frac{\beta+d\alpha}{\beta-t}\right)^{tp_i} \\
&= e^{-t\eta}\left(\frac{\beta+d\alpha}{\beta-t}\right)^{t} \\
&= \exp\left(-t\eta + t\log\frac{\beta+d\alpha}{\beta-t}\right) \\
&:= \exp(f(t)). \qquad (20)
\end{aligned}$$

The function $f(t)$ is minimized at $t^* \coloneqq \beta\left(1 - W\left(\frac{\beta e^{1+\eta}}{\beta+d\alpha}\right)^{-1}\right)$, where W is the Lambert W function. Note that W satisfies the identity $\log(W(x)/x) = -W(x)$ for all $x \geq -e^{-1}$. Therefore,

$$\begin{aligned}
f(t^*) &= -t^*\eta + t^* \log \frac{\beta + d\alpha}{\beta - t^*} \\
&= -t^*\eta + t^* \log\left\{\frac{\beta + d\alpha}{\beta})W\left(\frac{\beta e^{1+\eta}}{\beta + d\alpha}\right)\right\} \\
&= -t^*\eta + t^* \log\left\{\frac{\beta + d\alpha}{\beta e^{1+\eta}} W\left(\frac{\beta e^{1+\eta}}{\beta + d\alpha}\right)\right\} + t^* \log e^{1+\eta} \\
&= -t^*\eta - t^* W\left(\frac{\beta e^{1+\eta}}{\beta + d\alpha}\right) + t^*(1+\eta) \\
&= t^*\left(1 - W\left(\frac{\beta e^{1+\eta}}{\beta + d\alpha}\right)\right) \\
&= -\beta\left(1 - W\left(\frac{\beta e^{1+\eta}}{\beta + d\alpha}\right)^{-1}\right)\left(W\left(\frac{\beta e^{1+\eta}}{\beta + d\alpha}\right) - 1\right). \qquad (21)
\end{aligned}$$

The assumption $\beta \geq d\alpha/(e^{\eta/2} - 1)$ implies $\beta/(\beta + d\alpha) \geq e^{-\eta/2}$. We use the inequality $W(x) \geq \log x - \log\log x + \log\log x/(2\log x)$ for $x \geq e$ (Hoorfar & Hassani, 2008, Theorem 2.7) to obtain

$$\begin{aligned}
W\left(\frac{\beta e^{1+\eta}}{\beta + d\alpha}\right) &\geq W\left(e^{1+\eta/2}\right) \\
&\geq 1 + \frac{\eta}{2} - \log\left(1 + \frac{\eta}{2}\right) + \frac{\log(1+\eta/2)}{2(1+\eta/2)} \\
&= 1 + \frac{\eta}{2} - \left(\frac{1+\eta}{2+\eta}\right)\log\left(1 + \frac{\eta}{2}\right) \\
&\geq 1 + \frac{\eta}{2} - \frac{\eta}{2} \cdot \frac{1+\eta}{2+\eta} \\
&= 1 + \frac{\eta}{2(2+\eta)}.
\end{aligned}$$

Continuing from equation 21, we have

$$f(t^*) \leq -\beta\left(1 - \left(1 + \frac{\eta}{2(2+\eta)}\right)^{-1}\right)\left(1 + \frac{\eta}{2(2+\eta)} - 1\right) = -\beta\left(\frac{\eta^2}{2(2+\eta)(4+3\eta)}\right).$$

Inserting this inequality back into equation 20, we obtain

$$\Pr[D_{\mathrm{KL}}(p\|q) > \eta] \leq \exp(f(t)) \leq \exp(f(t^*)) \leq e^{-\beta\eta^2/(2(2+\eta)(4+3\eta))},$$

as desired.