# 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 information-theoretic modeling. In a probabilistic model, for example, we estimate the model's parameters by maximizing the likelihood function of the parameters, which 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; hence they are often applied on 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; this framework allows one to design algorithms that preserve the privacy of their inputs. Many of the designs are results of adding a small amount of random noise to the output of an existing algorithm. Typically, the random noise is usually drawn from a Gaussian or Laplace distribution. These additive noise mechanisms are the backbones of many privacy-preserving algorithms, from simple queries such as counting and histogram queries (Dwork et al., 2006a;b) to complex models such as deep learning (Abadi et al., 2016). The *utility* of such techniques is usually measured in terms of the $\ell^1$- or $\ell^2$-distance between the noisy and true outputs; the counting and histogram queries above are good examples for which small distances are desirable.

If the goal is to minimize the KL divergence, however, additive noise mechanisms might not be appropriate. For example, consider normalized count data of $p = [0.5, 0.5]$ and $q = [0.1, 0.9]$. Suppose that we draw a noise vector of $z = [-0.1, 0.1]$. Then the KL divergences between the true and noisy outputs are $D_{\mathrm{KL}}(p\|p+z) \approx 0.009$ and $D_{\mathrm{KL}}(q\|q+z) = \infty$, which illustrate that adding noises at a fixed scale disproportionately affects the KL divergence of imbalanced normalized counts, implying that additive noise mechanisms do not provide the best utility for private KL divergence minimization.

We instead consider the exponential mechanism (McSherry & Talwar, 2007), 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. Fortunately, 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 reals $x$ and $x'$, the notation $x \propto x'$ means $x = Cx'$ for some constant $C > 0$, $x \approx x'$ means $cx' \leq x \leq Cx'$ for some $c, C > 0$, and $x \lesssim x'$ means $x \leq Cx'$ for some $C > 0$. Lastly, $\|u\|_2 \coloneqq \sqrt{u_1^2 + \ldots + u_d^2}$ is the $\ell^2$ norm of $u$ and $\|u\|_\infty \coloneqq \max_i |u_i|$ is the $\ell^\infty$ 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 $(\varepsilon, \delta)$-differentially private ($(\varepsilon, \delta)$-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 $\varepsilon$-differential private ($\varepsilon$-DP).

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

In contrast to pure and approximate DP, the next definitions of differential privacy are 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 $\lambda$ between $P$ and $Q$ is defined as

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

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

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

Intuitively, $\varepsilon$ 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 $\lambda$. A smaller $\varepsilon$ and larger $\lambda$ correspond to a stronger privacy guarantee.

The composition property allow us to use the Dirichlet posterior sampling as a building block for more complex algorithms.

**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 $\mu$ 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 $\beta$ is a privacy-related 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 $\ell(x, y) = D_{\mathrm{KL}}(\widetilde{f(x)} \| y)$, $\beta = r N_f$, 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:

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

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.

To see how the choices of $r$ and $\alpha$ affect the "distance" between $y_i$ and $\widetilde{f_i(x)}$, we look at the bias of $y_i$, which is

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

We see that the bias can be reduced by increasing the value of $r$ and decreasing the value of $\alpha$.

**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, $f_1(x),\dots,f_d(x)$ be the frequencies and $y_1,\dots,y_d$ 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 an approximate solution 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. Specifically, $\psi(x) \coloneqq \frac{d}{dx}\log\Gamma(x)$ is concave and increasing, while its derivative $\psi'(x)$ is positive, convex, and decreasing (see Figure 1). In addition, $\psi'$ can be approximated by the reciprocals:

$$\frac{1}{x}+\frac{1}{2x^2} < \psi'(x) < \frac{1}{x}+\frac{1}{x^2}, \tag{4}$$

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

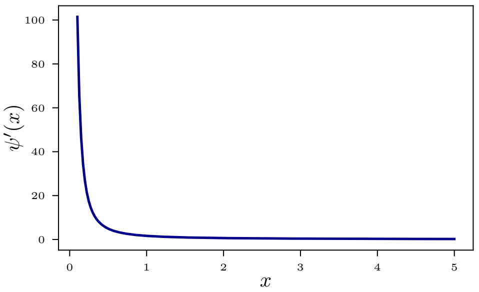

Figure 1: A plot of $\psi'(x)$.

### 2.4 Related work

There are several studies on the differential privacy of probability sampling. Wang et al. (2015) showed that any sampling with the absolute value of the log-density bounded above by a constant $B$ 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 a single draw from the Beta distribution, which is the two-dimensional Dirichlet distribution, 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 is $(\varepsilon,\delta)$-DP with $\varepsilon > 0$.

Let $x$ be a sufficient statistic of an exponential family with finite $\ell^1$-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 $(\lambda, \varepsilon)$-RDP guarantees for exponential families, explicit forms of $\lambda$ and the upper bound of $\lambda$ 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 $(\varepsilon,\delta)$-approximate DP, where $\varepsilon$ 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) has recently provided a privacy guarantee for the Dirichlet mechanism, which is impractical as it requires numerical integrations and optimization over the unit simplex. In contrast, our guarantee is much simpler to compute. We are also the first to provide the utility of the Dirichlet mechanism in terms of KL divergence minimization.

## 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 $\ell^2$- and $\ell^\infty$-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** $(\lambda, \varepsilon)$**-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 $\ell^2$-sensitivity $\Delta_2$ and $\ell^\infty$-sensitivity $\Delta_\infty$
Parameters: $\lambda > 1$, $\varepsilon > 0$
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 $y \sim$ Dirichlet$(rf(x) + \alpha)$.
```

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

**Lemma 2.** *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* $\lambda > 1$.

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

### 3.2 Privacy guarantee

**Theorem 1.** *Algorithm 1 is* $(\lambda, \varepsilon)$*-RDP.*

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

**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

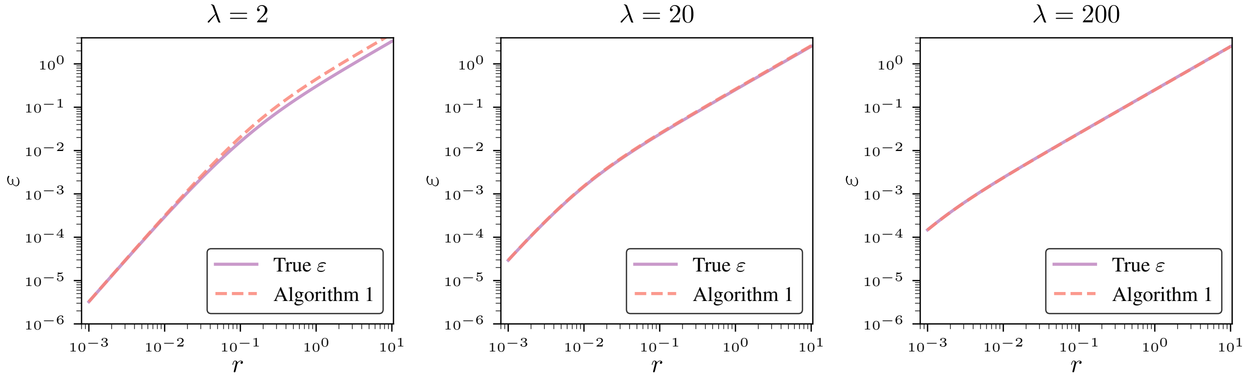

Figure 2: The Rényi divergence ($\varepsilon$) of order $\lambda$ between Dirichlet$(rf(x)+\alpha)$ and Dirichlet$(rf(x')+\alpha)$ as a function of $r$, where $f(x) = (11,8,65,25,38,0)$, $f(x') = (11,7,65,25,38,1)$ and $\alpha$ is given in Algorithm 1 with $\Delta_2^2 = 2$ and $\Delta_\infty = 1$. Here, we plot both the direct calculations of $\varepsilon$ and the suggested values of $\varepsilon$ in Algorithm 1.

$r = \sqrt{2\varepsilon/(\lambda\Delta_2^2\psi'(1))}$ which can be computed without a root-finding algorithm. However, this choice of $r$ makes $\varepsilon$ 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 $\alpha$, which allows us to offset the $\lambda r^2$ factor in $\varepsilon$ 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 $(\lambda,\varepsilon)$-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 $\psi'$ 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: $f(x) = (11,8,65,25,38,0)$ and $f(x') = (11,7,65,25,38,1)$. As functions of $r$, we compare $\varepsilon$ in Line 1 with the analytic values of the Rényi divergence between Dirichlet$(rf(x)+\alpha)$ and Dirichlet$(rf(x')+\alpha)$, where $\alpha$ is given in Line 2. The plots of $\varepsilon$ as functions of $r$ in Figure 2 show that our proposed RDP-guarantees are close to the actual Rényi divergences across different values of $\lambda$.

## 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 $N_f := \sum_i f_i(x)$ and $\widetilde{f(x)} := N_f^{-1} f(x) \in S^{d-1}$. We propose the Dirichlet mechanism (Algorithm 1) which aims to output $y$ that minimizes $D_{\mathrm{KL}}\left(\widetilde{f(x)}\|y\right)$ 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 E. 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

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

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 $\alpha$ in terms of $\varepsilon$ and $\lambda$ using the following identities from Algorithm 1.

$$\varepsilon = \frac{1}{2}\lambda r^2 \Delta_2^2 \psi'(1 + 3(\lambda - 1)r\Delta_\infty) \tag{5}$$

$$\alpha = 1 + 4(\lambda - 1)r\Delta_\infty. \tag{6}$$

We note in Remark 1 above that the right-hand side of equation 5 is an increasing function of $r$. This implies that, if $\varepsilon \geq 1$, then $r > 1$ and so $\psi'(1 + 3(\lambda - 1)r\Delta_\infty) = \Theta\Big(\frac{1}{(\lambda-1)r}\Big)$. Thus, equation 5 and 6 give $r = \Theta(\varepsilon)$ and $\alpha = \Theta((\lambda - 1)\varepsilon)$. On the other hand, if $\varepsilon < 1$, we have $r \leq 1$ 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 $(\lambda, \varepsilon)$-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 $(\lambda, \varepsilon)$-RDP and sub-$\eta$ KL divergence, with $\lambda$ and $\eta$ fixed, is $N = \Omega\big(\frac{1}{\varepsilon} + 1\big)$ if $\varepsilon \geq 1$ and $N = \Omega\Big(\frac{1}{\sqrt{\varepsilon}}\Big)$ if $\varepsilon < 1$.

## 5 Experiments and discussions

### 5.1 Naïve Bayes classification

We consider the Dirichlet mechanism for differentially private multinomial naïve Bayes classification. Let $D = \{(x^{(i)}, y^{(i)})\}_{i=1}^N$ be a dataset. Here, $x^{(i)} = (x_1^{(i)}, \ldots, x_K^{(i)}) \in \prod_{k=1}^K \mathcal{X}_k$ and $y^{(i)} \in [d]$ 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 $N_j := \sum_{i=1}^N \mathbb{I}(y^{(i)} = j)$ and $N_{jc}^k := \sum_{i=1}^N \mathbb{I}(y^{(i)} = j, x_k^{(i)} = c)$. The maximum-likelihood estimators for the naïve Bayes model are:

$$\hat{\pi}_j := N_j/N \quad \text{and} \quad \hat{\theta}_{jc}^k := N_{jc}^k/N_j. \tag{7}$$

With these estimates, the probability of $y = j$ conditioned on $(x_1, \ldots, x_K)$ can be computed as follows:

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

To modify the model with the Dirichlet mechanism, we sample $(\tilde{\pi}_1, \ldots, \tilde{\pi}_d) \sim \mathrm{Dirichlet}(r(N_1, \ldots, N_d) + \alpha)$, where $r$ and $\alpha$ 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 $(\tilde{\theta}_{1c}^k, \ldots, \tilde{\theta}_{dc}^k) \sim \mathrm{Dirichlet}\big(r_c^k(N_{1c}^k, \ldots, N_{dc}^k) + \alpha_c^k\big)$, where $r_c^k$ and $\alpha_c^k$ 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 $(\lambda, \varepsilon)$-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. This can be done by replacing $N_j$ and $N_{jc}^k$ in equation 7 by their noisy versions, namely $\tilde{N}_j := N_j + z_j$ and $\tilde{N}_{jc}^k := N_{jc}^k + z_{jc}^k$ where $z_j, z_{jc}^k \sim \mathcal{N}(0, \lambda(K+1)/\varepsilon)$ for the Gaussian mechanism and $z_j, z_{jc}^k \sim \mathrm{Laplace}(0, \sqrt{2\lambda(K+1)/\varepsilon})$ for the Laplace mechanism.

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) |

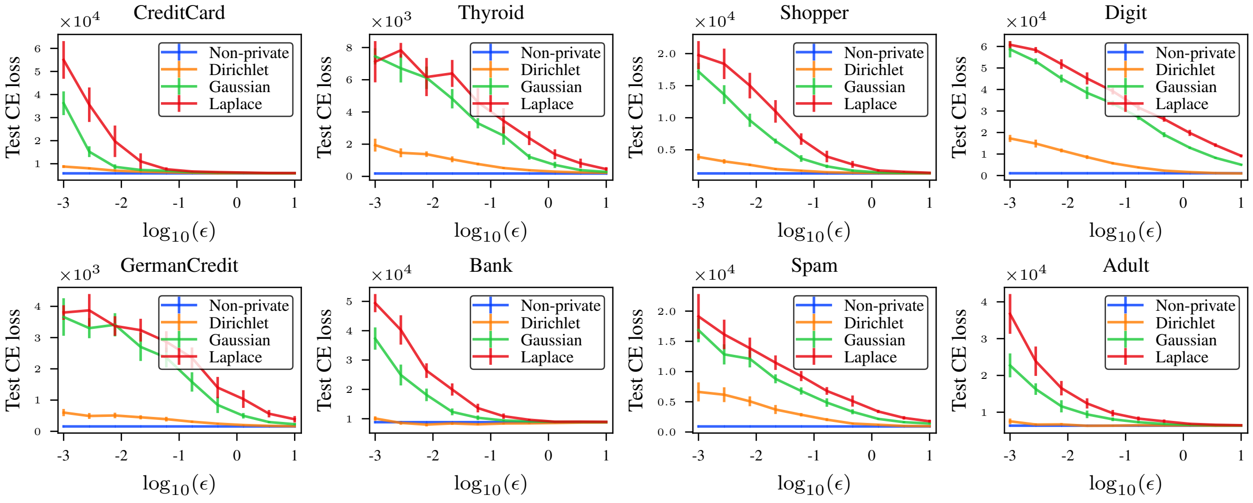

Figure 3: Test CE losses of the original and private naïve Bayes models on 8 UCI datasets

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, 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 $\varepsilon$ increases from $10^{-3}$ to 10. The classification performances, measured in cross-entropy (CE) loss and accuracy on the test sets, are shown in Figure 3 and 4. 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 are 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 $\varepsilon$.

Therefore, if one wants their differentially private naïve Bayes model to perform well in terms of CE loss, or both CE loss and accuracy, then the Dirichlet mechanism is an attractive option.

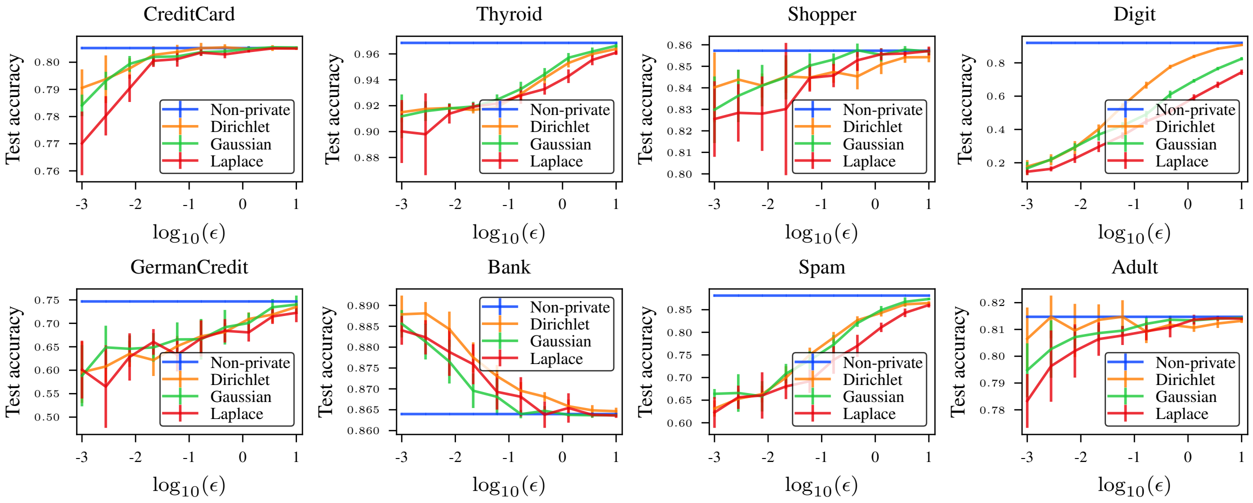

Figure 4: Test accuracies of the original and private naïve Bayes models on 8 UCI datasets

### 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 $k$ by $Pa(k)$. Let $x_{Pa(k)}^{(i)} := (x_\ell^{(i)})_{\ell \in Pa(k)}$ be observed values of $Pa(k)$ and $\mathcal{X}_{Pa(j)} := \prod_{\ell \in Pa(j)} \mathcal{X}_\ell$ be the product space of $Pa(k)$. Given $j \in \mathcal{X}_k$ and $c \in \mathcal{X}_{Pa(k)}$, we denote $N_c^k := \sum_{i=1}^{N} \mathbb{I}(x_{Pa(k)}^{(i)} = c)$ and $N_{jc}^k := \sum_{i=1}^{N} \mathbb{I}(x_k^{(i)} = j, x_{Pa(k)}^{(i)} = c)$. The log-likelihood of the parameters $\theta_{jc}^k := \Pr[x_k = j \mid x_{Pa(k)} = c]$ is given by:

$$LL(\theta) := \sum_{k \in [K]} \sum_{\substack{j \in \mathcal{X}_k \\ c \in \mathcal{X}_{Pa(k)}}} N_{jc}^k \log \theta_{jc}^k. \tag{8}$$

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

$$\hat{\theta}_{jc}^k := \frac{N_{jc}^k}{N_c^k}. \tag{9}$$

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 $(\tilde{\theta}_{1c}^k, \ldots, \tilde{\theta}_{dc}^k) \sim \mathrm{Dirichlet}\big(r(N_{1c}^k, \ldots, N_{dc}^k) + \alpha\big)$. Here, $r$ and $\alpha$ 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 $(\lambda, \varepsilon)$-RDP.

We will compare the Dirichlet mechanism with the Gaussian and Laplace mechanisms. In equation 9, we replace $N_{jc}^k$ by its noisy version: $\tilde{N}_{jc}^k := N_{jc}^k + z_{jc}^k$, where $z_{jc}^k \sim \mathcal{N}(0, \lambda K/\varepsilon)$ for the Gaussian mechanism and $z_{jc}^k \sim \mathrm{Laplace}(0, \sqrt{2\lambda K/\varepsilon})$ for the Laplace mechanism. In addition, we replace $N_c^k$ by $\tilde{N}_c^k := \sum_j \tilde{N}_{jc}^k$.

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 5. 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 $\varepsilon$ increases from $10^{-3}$ to 10. The plot of the log-likelihoods as functions of $\varepsilon$ are shown in Figure 6. We can see that, on all datasets, the test log-likelihoods of the Dirichlet

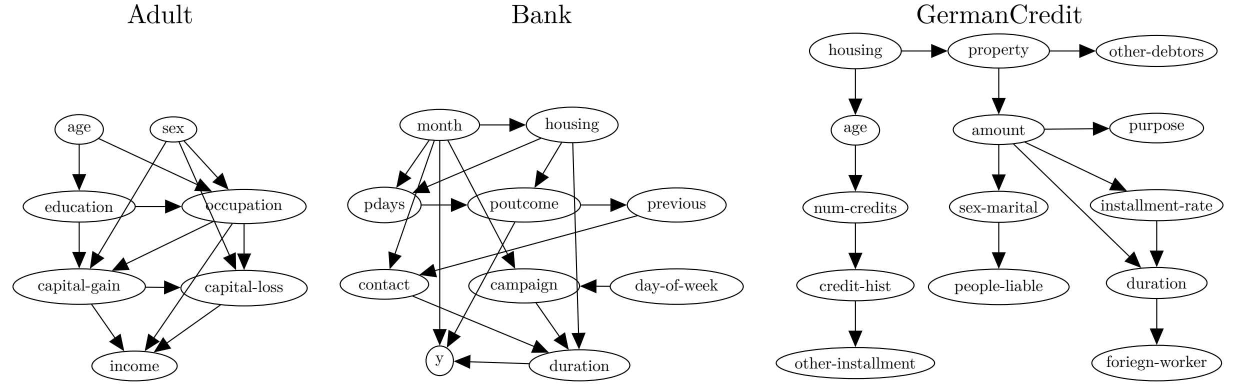

Figure 5: Our Bayesian networks on three datasets.

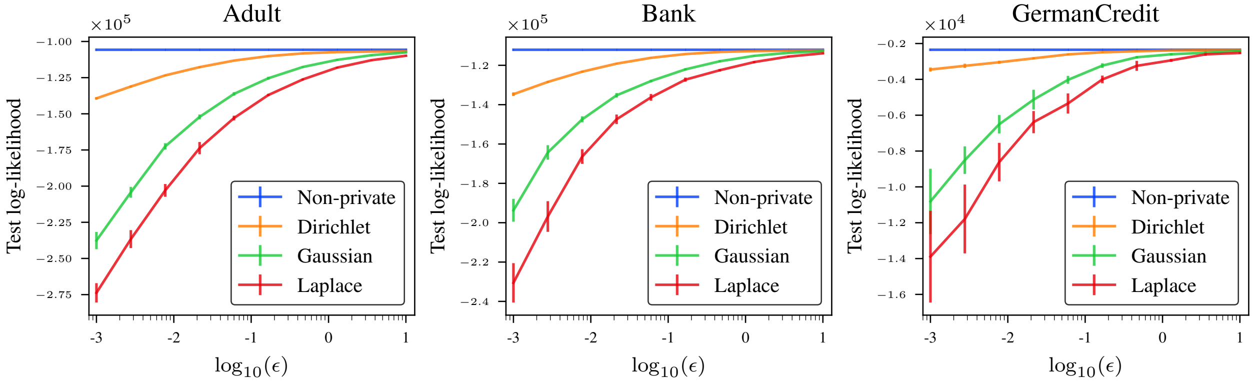

Figure 6: Test log-likelihoods of the parameters obtained from the maximum-likelihood estimation (non-private) and three differential privacy mechanisms.

mechanism are substantially less than those of the Gaussian mechanism and Laplace mechanism. 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

We study derive the Dirichlet mechanism as an instance of the exponential mechanism with the discrete KL divergence as the loss function. Consequently, we suggest using the mechanism for privacy-aware KL divergence minimization, which in turn is equivalent to likelihood maximization and cross–entropy minimization. To this end, we propose a choice of the privacy factor $r$ and the prior $\alpha$ that achieve a desired $(\lambda, \varepsilon)$-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`.

Irmak Aykin, Berk Akgun, Mingjie Feng, and Marwan Krunz. MAMBA: A multi-armed bandit framework for beam tracking in millimeter-wave systems. In *39th IEEE Conference on Computer Communications, INFOCOM 2020, Toronto, ON, Canada, July 6-9, 2020*, pp. 1469–1478. IEEE, 2020. doi: 10.1109/INFOCOM41043.2020.9155408. URL `https://doi.org/10.1109/INFOCOM41043.2020.9155408`.

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`.

Richard J. Boys, Daniel A. Henderson, and Darren J. Wilkinson. Detecting homogeneous segments in dna sequences by using hidden markov models. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 49(2):269–285, 2000. ISSN 00359254, 14679876. URL `http://www.jstor.org/stable/2680853`.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In Martin Hirt and Adam Smith (eds.), *Theory of Cryptography*, pp. 635–658, Berlin, Heidelberg, 2016. Springer Berlin Heidelberg.

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`.

Jukka Corander, Patrik Waldmann, and Mikko J Sillanpää. Bayesian Analysis of Genetic Differentiation Between Populations. *Genetics*, 163(1):367–374, 01 2003. ISSN 1943-2631. doi: 10.1093/genetics/163.1.367.

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`.

Matthieu de Lapparent. Empirical bayesian analysis of accident severity for motorcyclists in large french urban areas. *Accident Analysis & Prevention*, 38(2):260–268, 2006. ISSN 0001-4575. doi: https://doi.org/10.1016/j.aap.2005.09.001. URL `https://www.sciencedirect.com/science/article/pii/S0001457505001466`.

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, 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.

Keegan E. Hines. A primer on bayesian inference for biophysical systems. *Biophysical Journal*, 108(9): 2103–2113, 2015. ISSN 0006-3495. doi: https://doi.org/10.1016/j.bpj.2015.03.042. URL `https://www.sciencedirect.com/science/article/pii/S0006349515003033`.

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.

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.

Michael Lavine and Mike West. A bayesian method for classification and discrimination. *Canadian Journal of Statistics*, 20(4):451–461, 1992. doi: https://doi.org/10.2307/3315614. URL `https://onlinelibrary.wiley.com/doi/abs/10.2307/3315614`.

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`.

Jean-Michel Marin, Kerrie Mengersen, and Christian P. Robert. Bayesian modelling and inference on mixtures of distributions. In D.K. Dey and C.R. Rao (eds.), *Bayesian Thinking*, volume 25 of *Handbook of Statistics*, pp. 459–507. Elsevier, 2005. doi: https://doi.org/10.1016/S0169-7161(05)25016-2. URL `https://www.sciencedirect.com/science/article/pii/S0169716105250162`.

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`.

Imtiaz Nasim, Ahmed S. Ibrahim, and Seungmo Kim. Learning-based beamforming for multi-user vehicular communications: A combinatorial multi-armed bandit approach. *IEEE Access*, 8:219891–219902, 2020. doi: 10.1109/ACCESS.2020.3043301. URL `https://doi.org/10.1109/ACCESS.2020.3043301`.

Viet Cuong Nguyen, Wee Sun Lee, Nan Ye, Kian Ming Adam Chai, and Hai Leong Chieu. Active learning for probabilistic hypotheses using the maximum gibbs error criterion. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pp. 1457–1465, 2013. URL `https://proceedings.neurips.cc/paper/2013/hash/fb89705ae6d743bf1e848c206e16a1d7-Abstract.html`.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pp. 3003–3011, 2013. URL `https://proceedings.neurips.cc/paper/2013/hash/6a5889bb0190d0211a991f47bb19a777-Abstract.html`.

Ted Pedersen and Rebecca F. Bruce. Knowledge lean word-sense disambiguation. In Jack Mostow and Chuck Rich (eds.), *Proceedings of the Fifteenth National Conference on Artificial Intelligence and Tenth Innovative Applications of Artificial Intelligence Conference, AAAI 98, IAAI 98, July 26-30, 1998, Madison, Wisconsin, USA*, pp. 800–805. AAAI Press / The MIT Press, 1998. URL `http://www.aaai.org/Library/AAAI/1998/aaai98-113.php`.

Jerome Pella and Michele Masuda. Bayesian methods for analysis of stock mixtures from genetic characters. *Fishery Bulletin*, 99:151, 01 2001. ISSN 00900656. 1.

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`.

Malcolm J. A. Strens. A bayesian framework for reinforcement learning. In Pat Langley (ed.), *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2, 2000*, pp. 943–950. Morgan Kaufmann, 2000.

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`.

Junge Zhu, Xi Huang, Xin Gao, Ziyu Shao, and Yang Yang. Multi-interface channel allocation in fog computing systems using thompson sampling. In *2020 IEEE International Conference on Communications, ICC 2020, Dublin, Ireland, June 7-11, 2020*, pp. 1–6. IEEE, 2020. doi: 10.1109/ICC40277.2020.9148932. URL `https://doi.org/10.1109/ICC40277.2020.9148932`.

## A Dirichlet posterior sampling is not $\varepsilon$-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 $\varepsilon$-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 $\sum_i y_i = 1$, 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 $e^{\varepsilon}$. □

Since there is no hope for pure differential privacy, we turn our attention to one of relaxed notions of differential privacy. We shall see below that, with Rényi differential privacy (RDP), we can derive the privacy guarantee of the Dirichlet posterior sampling in a simple form.

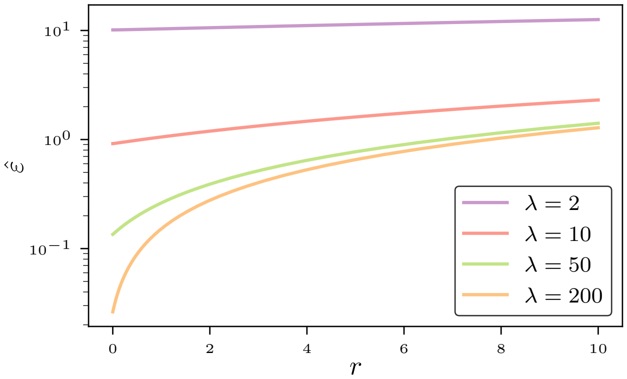

Figure 7: $(\varepsilon, \delta)$-DP guarantees of the Dirichlet mechanism following equation 10 with $\lambda \in \{2, 10, 50, 200\}$ and $\delta = 10^{-5}$.

## 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 $(\lambda, \varepsilon)$-RDP mechanism, then it also satisfies $(\hat{\varepsilon}, \delta)$-DP with*

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

Taking the logarithm of equation 3,

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

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

$$\hat{\varepsilon} = \frac{1}{2}\lambda r^2 \Delta_2^2 \psi'(1 + 3(\lambda - 1)r\Delta_\infty) + \log(\lambda - 1) - \frac{\log \delta + \lambda \log(\lambda)}{\lambda - 1}, \tag{10}$$

which gives a formula for $\hat{\varepsilon}$ in terms of $r$, $\lambda$ and $\delta$. Figure 7 shows $\hat{\varepsilon}$ as a function of $r$ at four different values of $\lambda$. We can see that $\hat{\varepsilon}$ is positively correlated with $r$ and negatively correlated with $\lambda$.

## C Proof of Lemma 2

Denote $x = 3(\lambda - 1)r\Delta_\infty$. With $\varepsilon, \lambda, \Delta_2$ and $\Delta_\infty$ fixed as constants, we can write the equation as $\varepsilon = Cx^2\psi'(1 + x)$ for some constant $C > 0$. From equation 4, 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,

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

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.

## D Proof of Theorem 2

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 $u = (u_1, \dots, u_d)$, $u' = (u'_1, \dots, u'_d)$, $u_0 \coloneqq \sum_i u_i$ 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 $P(\boldsymbol{y})$ and $P'(\boldsymbol{y})$, we start with:

$$\begin{aligned}\mathbb{E}_{y\sim P(y)}\left[\frac{P(y)^{\lambda-1}}{P'(y)^{\lambda-1}}\right] &= \frac{B(u')^{\lambda-1}}{B(u)^{\lambda-1}}\mathbb{E}_{y\sim P(y)}\left[y^{(\lambda-1)(u-u')}\right] \\ &= \frac{B(u')^{\lambda-1}}{B(u)^{\lambda-1}} \cdot \frac{B(u+(\lambda-1)(u-u'))}{B(u)}. \end{aligned} \tag{11}$$

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 11, we need to find an upper bound of:

$$\log \mathbb{E}_{y\sim P(y)}\left[\frac{P(y)^{\lambda-1}}{P'(y)^{\lambda-1}}\right] = \sum_i (G(u_i,u'_i) + H(u_i,u'_i)) - G(u_0,u'_0) - H(u_0,u'_0), \tag{12}$$

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 $\xi$ between $u_i + (\lambda-1)(u_i-u'_i)$ and $u_i$, and $\xi'$ 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 will try to find an upper bound of both $\psi'(\xi)$ and $\psi'(\xi')$. Note that $\psi'$ is increasing. If $f_i(x) > f_i(x')$, then $u'_i < u_i < u_i + (\lambda-1)(u_i-u'_i)$. Thus both $\xi$ and $\xi'$ are bounded below by $u'_i \geq \alpha_m$. 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, $\xi$ and $\xi'$ are bounded below by:

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

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

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

The same argument can be used to show that, there exist $\xi_0$ and $\xi_0'$ 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 12,

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

Thus, given any $\lambda>1$, $\varepsilon>0$ 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 $g(r)=3(\lambda-1)r\Delta_\infty$.

## E Proof of the Utility bound

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

$$\log\left(x-\frac{1}{2}\right)<\psi(x)<\log x. \tag{13}$$

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

$$\begin{aligned}\Pr[D_{\mathrm{KL}}(p\|q)>\eta] &\le e^{-t\eta}\mathbb{E}\Big[e^{tD_{\mathrm{KL}}(p\|q)}\Big]\\ &= e^{-t\eta}\mathbb{E}\left[\prod_i(p_i/q_i)^{tp_i}\right]\\ &= e^{-t\eta}\prod_i p_i^{tp_i}\mathbb{E}\left[\prod_i q_i^{-tp_i}\right]\\ &= e^{-t\eta}\prod_i p_i^{tp_i}\frac{1}{B(\beta p+\alpha)}\int\prod_i q_i^{\beta p_i-tp_i+\alpha-1}\,dq\\ &= e^{-t\eta}\prod_i p_i^{tp_i}\frac{B(\beta p-tp_i+\alpha)}{B(\beta p+\alpha)}\\ &= e^{-t\eta}\frac{\Gamma(\beta+d\alpha)}{\Gamma(\beta-t+d\alpha)}\prod_i p_i^{tp_i}\frac{\Gamma(\beta p_i-tp_i+\alpha)}{\Gamma(\beta p_i+\alpha)}.\end{aligned} \tag{14}$$

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

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

Inserting these inequalities and equation 13 into equation 14, we obtain

$$\begin{aligned}
\Pr[D_{\mathrm{KL}}(p\|q) > \eta] &\leq e^{-t\eta} e^{t\psi(\beta+d\alpha)} \prod_i p_i^{tp_i} e^{-tp_i \psi(\beta p_i - tp_i + \alpha)} \\
&< e^{-t\eta} e^{t\log(\beta+d\alpha)} \prod_i p_i^{tp_i} e^{-tp_i \log(\beta p_i - tp_i + \alpha - 1/2)} \\
&= e^{-t\eta} (\beta+d\alpha)^t \prod_i p_i^{tp_i} (\beta p_i - tp_i + \alpha - 1/2)^{-tp_i} \\
&= e^{-t\eta} (\beta+d\alpha)^t \prod_i \big(\beta - t + p_i^{-1}(\alpha - 1/2)\big)^{-tp_i} \\
&= e^{-t\eta} \prod_i \left(\frac{\beta+d\alpha}{\beta - t + p_i^{-1}(\alpha-1/2)}\right)^{tp_i} \\
&< e^{-t\eta} \prod_i \left(\frac{\beta+d\alpha}{\beta-t}\right)^{tp_i} \\
&= e^{-t\eta} \left(\frac{\beta+d\alpha}{\beta-t}\right)^{t} \\
&= \exp\left(-t\eta + t\log\frac{\beta+d\alpha}{\beta-t}\right) \\
&\coloneqq \exp(f(t)). \qquad (15)
\end{aligned}$$

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,

$$\begin{aligned}
f(t^*) &= -t^*\eta + t^* \log\frac{\beta+d\alpha}{\beta-t^*} \\
&= -t^*\eta + t^* \log\left\{\frac{\beta+d\alpha}{\beta})W\left(\frac{\beta e^{1+\eta}}{\beta+d\alpha}\right)\right\} \\
&= -t^*\eta + t^* \log\left\{\frac{\beta+d\alpha}{\beta e^{1+\eta}}W\left(\frac{\beta e^{1+\eta}}{\beta+d\alpha}\right)\right\} + t^*\log e^{1+\eta} \\
&= -t^*\eta - t^* W\left(\frac{\beta e^{1+\eta}}{\beta+d\alpha}\right) + t^*(1+\eta) \\
&= t^*\left(1 - W\left(\frac{\beta e^{1+\eta}}{\beta+d\alpha}\right)\right) \\
&= -\beta\left(1 - W\left(\frac{\beta e^{1+\eta}}{\beta+d\alpha}\right)^{-1}\right)\left(W\left(\frac{\beta e^{1+\eta}}{\beta+d\alpha}\right) - 1\right). \qquad (16)
\end{aligned}$$

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 16, 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 15, 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.