Appendix A Definitions, proofs, and related work

Here, we provide missing definitions of the KLD and the $\beta$-divergence.

**Definition 2 (Kullback-Leibler divergence [53]).** The KLD between probability densities $g(\cdot)$ and $f(\cdot)$ is given by

$$KLD(g||f) = \int g(x) \log \frac{g(x)}{f(x)} dx.$$ 

**Definition 3 ($\beta$-divergence [9, 63]).** The $\beta$-divergence is defined as

$$D_{\beta}^{(g)}(g||f) = \frac{1}{\beta(\beta - 1)} \int g(x)^{\beta} dx + \frac{1}{\beta} \int f(x)^{\beta} dx - \frac{1}{\beta - 1} \int g(x)f(x)^{\beta - 1} dx,$$

where $\beta \in \mathbb{R} \setminus \{0, 1\}$. $D_{\beta}^{(g)}$ is a member of the Bregman-divergence family [16] with $\psi(t) = \frac{1}{\beta(\beta - 1)} t^\beta$.

When $\beta \to 1$, $D_{\beta}^{(g)}(g||f(x)) \to KLD(g(x)||f(x))$.

The $\beta$-divergence has often been referred to as the density-power divergence in the statistics literature [9] where it is often parameterised with $\beta_{DPD} = \beta - 1$.

Intuition for how $\beta$D-Bayes provides DP estimation is provided in Figure 5 which shows the divergence between the posterior before and after adding an observation $y$ that is $|y - \mu|$ standard deviations away from the posterior mean $\mu$ when updating using a Gaussian distribution under KLD-Bayes and $\beta$D-Bayes. The influence of observations under KLD-Bayes is steadily increasing, making the posterior sensitive to extreme observations and therefore leaking their information. Under $\beta$D-Bayes, the influence initially increases before being maximised at a point depending on the value of $\beta$, before decreasing to 0. Therefore, each observation has bounded influence on the posterior, allowing for DP estimation.

![Figure 5: The influence of adding an observation $y$ with $|y - \mu|$ on the posterior conditioned on a sample of 1000 points from a $N(0, 1)$ when fitting a $N(\mu, \sigma^2)$.](image)

### A.1 Bernstein-von Mises theorem for $\beta$D-Bayes

The general Bernstein-von Mises theorem for generalised posteriors [Theorem 4; 64] can be applied to the $\beta$D-Bayes posterior to show that

$$\int \widetilde{\pi}^{(\beta)}(\phi) - N\left(\phi; 0, (H_0^{(\beta)})^{-1}\right) d\phi \underset{n \to \infty}{\longrightarrow} 0 \quad (5)$$

where $\widetilde{\pi}^{(\beta)}$ denotes the density of $\sqrt{n}(\hat{\theta} - \hat{\theta}_n^{(\beta)})$ when $\hat{\theta} \sim \pi^{(\beta)}(\cdot; D)$, $N(x; \mu, \sigma^2)$ denotes the normal distribution with mean $\mu$ and variance $\sigma^2$, and $\hat{\theta}_n^{(\beta)} := \arg \min_{\theta \in \Theta} \sum_{i=1}^{n} \ell^{(\beta)}(D_i, f(\cdot; \theta))$, $\theta_0^{(\beta)} := \arg \min_{\theta \in \Theta} E_g \left[ \ell^{(\beta)}(D, f(\cdot; \theta)) \right]$

$H_0^{(\beta)} := \left( \frac{\partial}{\partial \theta_i \partial \theta_j} E_D \left[ \ell^{(\beta)}(D, f(\cdot; \theta_0^{(\beta)})) \right] \right)_{i,j}$.

15
That is to show that the $\beta D$-Bayes posterior converges to a Gaussian distribution centered around the $\beta D$ minimising parameter $\theta_0^{(\beta)}$ in total variation distance.

### A.2 Proofs

#### A.2.1 Proof of Lemma 1

**Lemma 1** (Bounded sensitivity of the $\beta D$-Bayes loss). Under Condition 1 the sensitivity of the $\beta D$-Bayes-loss for any $\beta > 1$ is $|\ell^{(\beta)}(D, f(\cdot; \theta)) - \ell^{(\beta)}(D', f(\cdot; \theta))| \leq \frac{M^{\beta-1}}{\beta - 1}$.

**Proof.** By (4), for $\beta > 1$

$$
|\ell^{(\beta)}(D, f(\cdot; \theta)) - \ell^{(\beta)}(D', f(\cdot; \theta))| = \frac{1}{\beta - 1} (f(D'; \theta)^{\beta-1} - f(D; \theta)^{\beta-1}) 
\leq \max_{D} \frac{1}{\beta - 1} f(D; \theta)^{\beta-1} 
\leq \frac{M^{\beta-1}}{\beta - 1}
$$

\[\square\]

#### A.2.2 Proof of Theorem 1

**Theorem 1** (Differential privacy of the $\beta D$-Bayes posterior). Under Condition 1, a draw $\hat{\theta}$ from the $\beta D$-Bayes posterior $\pi^{(\beta)}(\theta | D)$ in (3) is $(\frac{2M^{\beta-1}}{\beta - 1}, 0)$-differentially private.

**Proof.** Define $D = \{D_1, \ldots, D_n\}$, $D' = \{D'_1, \ldots, D'_n\}$ and let $j$ be the index such that $D_j \neq D'_j$ with $D_j = D'_j$ for all $i \neq j$. Firstly, the normalising constant of the $\beta D$-Bayes posterior combining (3) with (4) is

$$
P^\ell(D) := \int \pi(\theta) \exp \left( -w \sum_{i=1}^{n} \ell(D_i, f(\cdot; \theta)) \right) d\theta
$$

Then,

$$
\log \frac{\pi^{(\beta)}(\theta | D)}{\pi^{(\beta)}(\theta | D')} = \sum_{i=1}^{n} \ell^{(\beta)}(D_i, f(\cdot; \theta)) - \sum_{i=1}^{n} \ell^{(\beta)}(D_i, f(\cdot; \theta)) + \log \frac{P^{(\beta)}(D')}{P^{(\beta)}(D)}
\leq \ell^{(\beta)}(D'_j; f(\cdot; \theta)) - \ell^{(\beta)}(D_j; f(\cdot; \theta)) + \log \frac{P^{(\beta)}(D')}{P^{(\beta)}(D)}
$$

where $P^{(\beta)}(D')$ is the normaliser of the general Bayesian posterior defined in (3).

Now, by Condition 1 and Lemma 1,

$$
\ell^{(\beta)}(D'_j; f(\cdot; \theta)) - \ell^{(\beta)}(D_j; f(\cdot; \theta)) \leq \frac{M^{\beta-1}}{\beta - 1},
$$

and

$$
P^{(\beta)}(D') = \int \exp \left\{ - \sum_{i=1}^{n} \ell^{(\beta)}(D'_i, f(\cdot; \theta)) \right\} \pi(\theta) d\theta
$$

$$
= \int \exp \left\{ \ell^{(\beta)}(D_j, f(\cdot; \theta)) - \ell^{(\beta)}(D'_j, f(\cdot; \theta)) - \sum_{i=1}^{n} \ell^{(\beta)}(D_i, f(\cdot; \theta)) \right\} \pi(\theta) d\theta
$$

$$
= \exp \left\{ \frac{M^{\beta-1}}{\beta - 1} \right\} \int \exp \left\{ - \sum_{i=1}^{n} \ell^{(\beta)}(D_i, f(\cdot; \theta)) \right\} \pi(\theta) d\theta,
$$

which combined provides that

$$
\log \frac{\pi(\theta | D)}{\pi(\theta | D')} \leq 2 \frac{M^{\beta-1}}{\beta - 1}.
$$

\[\square\]
A.2.3 Proof of Theorem 2

**Theorem 2** (Consistency of βD-Bayes sampling). Under the conditions of Theorem 4 of [64],
1. a posterior sample \( \hat{\theta} \sim p^{(\beta)}(\theta|D) \) is a consistent estimator of \( \theta^{(\beta)}_0 \).
2. if data \( D_1, \ldots, D_n \sim g(\cdot) \) were generated such that there exists \( \theta_0 \) with \( g(D) = f(D; \theta_0) \), then \( \hat{\theta} \sim p^{(\beta)}(\theta|D) \) for all \( 1 \leq \beta \leq \infty \) is consistent for \( \theta_0 \).

**Proof.** For part 1), define \( B_r(x_0) = \{ x \in \mathbb{R}^p : |x - x_0| < r \} \). Theorem 4 of [64] applied to βD-Bayes posterior proves that
\[
\int_{B_r(\theta^{(\beta)}_0)} \pi^{(\beta)}(\theta|D)d\theta \rightarrow 1 \quad \text{as} \quad n \rightarrow \infty
\]
for all \( \varepsilon > 0 \). This is enough to show that for \( \hat{\theta} \sim p^{(\beta)}(\theta|D) \) → \( \theta^{(\beta)}_0 \) in probability.

For part 2), note that if \( g(D) = f(D; \theta_0) \), then for all \( 1 \leq \beta \leq \infty \)
\[
\hat{\theta}^{(\beta)}_0 := \arg \min_{\theta \in \Theta} \mathbb{E}_q \left[ f^{(\beta)}(D; f(\cdot; \theta)) \right]
\]
\[
= \arg \min_{\theta \in \Theta} D^{(\beta)}_\theta (g||f(\cdot; \theta))
\]
\[
= \theta_0.
\]

A.2.4 Proof of Proposition 1

**Proposition 1** (Asymptotic efficiency). Under the conditions of Theorem 4 of [64], \( \hat{\theta} \sim p^{(\beta)}(\theta|x) \) is asymptotically distributed as \( \sqrt{n} (\hat{\theta} - \theta^{(\beta)}_0) \) \( \rightarrow \mathcal{N}(0, (H^{(\beta)}_0)^{-1} K^{(\beta)}_0 (H^{(\beta)}_0)^{-1}) \), where \( K^{(\beta)}_0 \) and \( H^{(\beta)}_0 \) are defined in Appendix A.1.

**Proof.** Let \( \hat{\theta} \sim p^{(\beta)}(\theta|D) \). By the Bernstein-von Mises theorem [64] applied to βD-Bayes in (5),
\[
\sqrt{n} (\hat{\theta} - \theta^{(\beta)}_n) \rightarrow \mathcal{N}(0, (H^{(\beta)}_0)^{-1}).
\]

By the asymptotic normality of \( \theta^{(\beta)}_n \) [10], we have that
\[
\sqrt{n} (\theta^{(\beta)}_n - \theta^{(\beta)}_0) \rightarrow \mathcal{N}(0, (H^{(\beta)}_0)^{-1} K^{(\beta)}_0 (H^{(\beta)}_0)^{-1})
\]
for \( K_0 := \left( \frac{\partial^2}{\partial \theta \partial \theta} E_D \left[ f^{(\beta)}(D; f(\cdot; \theta^{(\beta)}_0)) \right] \right)_{i,j} \). The result then comes from the asymptotic independence of \( \hat{\theta} - \theta^{(\beta)}_n \) and \( \theta^{(\beta)}_n \) [see e.g. 80] □

A.2.5 Proof of Proposition 2

**Proposition 2** (DP-MCMC methods for the βD-Bayes-Posterior). Under Condition 1, the penalty algorithm of [Algorithm 1; 82], DP-HMC of [Algorithm 1; 72] and DP-Fast MH of [Algorithm 2; 84] and under further Condition 2 DP-SGLD of [Algorithm 1; 56] can be used to produce \( (\epsilon, \delta) \)-DP estimation with \( \delta > 0 \) without requiring the clipping of any gradients.

**Condition 2** (Boundedness of the model density/mass function gradient). The model density or mass function \( f(\cdot; \theta) \) is such that there exists \( 0 < C^{(\beta)}_\theta < \infty \) such that \( |\nabla_\theta f(D; \theta) \times f(D; \theta)^{\beta - 2}| \leq C^{(\beta)}_\theta, \forall \theta \in \Theta \).

**Proof.** Algorithm 1 of [82], Algorithm 1 of [72] and Algorithm 2 of [84] requires a posterior whose log-likelihood has bounded sensitivity. For βD-Bayes posterior, this requires βD-Bayes-loss has bounded sensitivity which is provided by Condition 1 and Lemma 1.
Algorithm 1 of [56] requires a posterior whose log-likelihood has bounded gradient. For $\beta$D-Bayes posterior, this requires $\beta$D-Bayes-loss to have bounded gradient:

\[
|\nabla_\theta \ell^{(\beta)}(D; \theta)| = |\nabla_\theta f(D; \theta) \times f(D; \theta)^{\beta - 2} - \int \nabla_\theta f(D; \theta) \times f(D; \theta)^{\beta - 1} dD |
\]

\[
= |\nabla_\theta f(D; \theta) \times f(D; \theta)^{\beta - 2} - \int \nabla_\theta f(D; \theta) \times f(D; \theta)^{\beta - 2} \times f(D; \theta) dD |
\]

\[
\leq \max\{G^{(\beta)}, G^{(\beta)} M\},
\]

assuming we can interchange integration and differentiation and as $|\nabla_\theta f(D; \theta) \times f(D; \theta)^{\beta - 2}| \leq G^{(\beta)}$ by Condition 2 not requiring the clipping of any gradients.

\section{Related work}

Here, we would like to extend our discussion of two important areas within the related work.

### A.3.1 Differentially private logistic regression

Chaudhuri et al. [19] propose a regularised DP logistic regression, solving (1). (1) adds the regulariser to the average loss and as a result, the impact of the regulariser does not diminish as $n \to \infty$. Even though the scale of the Laplace noise decreases as $n$ grows, Chaudhuri et al. [19] consistently estimate a parameter that is not the data generating parameter. Alternatively, one could choose a regulariser $\lambda' := \frac{1}{n}$ whose influence decreases as $n$ grows. This would allow for unbiased inference as $n \to \infty$ (assuming a Bayesian model with corresponding prior distribution), but the $n$ cancels in the scale of the Laplace noise and therefore the perturbation scale does not decrease in $n$, and the estimator is inconsistent. Choosing instead $\lambda':=\frac{1}{n}$ with $0 < r < 1$, would help in constructing unbiased and consistent estimators. In our experiments, we did not find this choice to help.

### A.3.2 Differentially private Monte Carlo methods

Wang et al. [80] propose using Stochastic Gradient Langevin Dynamics [SGLD; 81] with a modified burn-in period and bounded step-size to provide DP sampling when the log-likelihood has bounded gradient. Li et al. [56] improve upon [80], taking advantage of the moments accountant [1] to allow for a larger step-size and faster mixing for non-convex target posteriors. Foulds et al. [28] extend their privatisation of sufficient statistics to a Gibbs sampling setting where the conditional posterior distribution for a Gibbs update is from the exponential family. Yıldırım and Ermiş [82] use the penalty algorithm which adds noise to the log of the Metropolis-Hastings acceptance probability. Heikkilä et al. [38] use Barker’s acceptance test [8, 75] and provide RDP guarantees. Räisä et al. [72] derive DP-HMC also using the penalty algorithm. Zhang and Zhang [84] propose a random batch size implementation of Metropolis-Hasting for a general proposal distribution that takes advantage of the inherent randomness of Metropolis-Hasting and is asymptotically exact. Lastly, Awan and Rao [7] consider DP rejection sampling.

### A.4 Attack optimality

\textbf{Remark 1.} Let $p(\hat{\theta}|D)$ be the density of the privacy mechanism—i.e the Laplace density for [19] or the posterior (i.e. Equations 2,3) for OPS. An attacker estimating $M(\hat{\theta}, D, D') = \frac{p(\hat{\theta}|D')}{(p(\hat{\theta}|D) + p(\hat{\theta}|D'))}$ is Bayes optimal. For OPS, $M(\hat{\theta}, D, D') = \exp\{\ell(D'; f(\cdot; \hat{\theta})) - \ell(D; f(\cdot; \hat{\theta}))\} \exp\{\ell(D'; f(\cdot; \theta)) - \ell(D; f(\cdot; \theta))\} \pi(\theta|D)d\theta$ where $D, D' s.t. D \setminus D' = \{D\}$ and $D' \setminus D = \{D'\}$ (see Appendix A.4).

The privacy attacks outlined in Section 4 require the calculation of $M(\hat{\theta}, D, D') := p(m = 1; \hat{\theta}, D, D') = p(\hat{\theta}|D')/(p(\hat{\theta}|D)) + p(\hat{\theta}|D'))$.

by Bayes Theorem. For [19], it is

\[
p(\hat{\theta}|D) = L\left(\hat{\theta}(D), \frac{2}{n\lambda}\right),
\]

where $\hat{\theta}(D)$ was defined in (1).
For the OPS methods, Minami et al. [65] and βD-Bayes, \( p(\hat{\theta} | D) \) is the posterior

\[
p(\hat{\theta} | D) = \pi(\ell) \langle \hat{\theta} | D \rangle \propto \pi(\theta) \exp\{-\sum_{i=1}^{n} \ell(D_i ; \theta)\}
\]

where for [65] \( \ell(D_i ; f(\cdot ; \theta)) = -w \log f(D_i ; \theta) \), and for βD-Bayes \( \ell(D_i ; f(\cdot ; \theta)) = \ell^{(\beta)}(D_i ; f(\cdot ; \theta)) \) given in (4). Without loss of generality, index observations within \( D \) and \( D' \) such that \( D \setminus D' = \{D_i\} \) and \( D' \setminus D = \{D'_i\} \). Then,

\[
\frac{\hat{\pi}(\hat{\theta} | D)}{\hat{\pi}(\hat{\theta} | D')} = \frac{\pi(\hat{\theta}) \exp\{-\sum_{i=1}^{n} \ell(D_i ; f(\cdot ; \hat{\theta}))\}}{\int \pi(\hat{\theta}) \exp\{-\sum_{i=1}^{n} \ell(D_i ; f(\cdot ; \hat{\theta}))\} d\theta} / \frac{\pi(\hat{\theta}) \exp\{-\sum_{i=1}^{n} \ell(D'_i ; f(\cdot ; \hat{\theta}))\}}{\int \pi(\hat{\theta}) \exp\{-\sum_{i=1}^{n} \ell(D'_i ; f(\cdot ; \hat{\theta}))\} d\theta}
\]

\[
= \exp\{\ell(D'_i ; f(\cdot ; \hat{\theta})) - \ell(D_i ; f(\cdot ; \hat{\theta}))\} \int \frac{\pi(\hat{\theta}) \exp\{-\sum_{i=1}^{n} \ell(D_i ; f(\cdot ; \hat{\theta}))\}}{\int \pi(\hat{\theta}) \exp\{-\sum_{i=1}^{n} \ell(D'_i ; f(\cdot ; \hat{\theta}))\} d\theta} d\theta
\]

\[
= \exp\{\ell(D'_i ; f(\cdot ; \hat{\theta})) - \ell(D_i ; f(\cdot ; \hat{\theta}))\} \int \exp\{\ell(D'_i ; f(\cdot ; \hat{\theta})) - \ell(D'_i ; f(\cdot ; \hat{\theta}))\} \pi(\theta | D) d\theta
\]

\[
\approx \exp\{\ell(D'_i ; f(\cdot ; \hat{\theta})) - \ell(D_i ; f(\cdot ; \hat{\theta}))\} \frac{1}{N} \sum_{j=1}^{N} \exp\{\ell(D_i ; f(\cdot ; \theta_j)) - \ell(D'_i ; f(\cdot ; \theta_j))\},
\]

where \( \{\theta_j\}_{j=1}^{N} \sim \pi(\theta | D) \). The adversary only needs to sample from the posterior based on dataset \( D \) to be able to estimate \( \mathcal{M}(\hat{\theta}, D, D') \) for all \( D' \) differing from \( D \) in only one index \( l \).

Appendix B  Additional experimental details and results

Additional experimental details  Unless otherwise specified, we choose \( d = 2 \) in the simulated experiments. The MCMC methods are run for 1000 warm-up steps, and 100 iterations. DPSGD is run for \( 15 + |\epsilon| \) epochs, with clipping norm 1, batch size 100, and learning rate of \( 10^{-2} \). All other implementation details can be found on https://anonymous.4open.science/r/beta-bayes-ops-6626.

Neural network classification  Similarly to neural network regression, we can use βD-Bayes for neural network classification. As we see in Figure 6, βD-Bayes regularly outperforms DPSGD for \( \epsilon > 0.2 \) on simulated and real data, except on abalone.

Sensitivity in number of features  Please refer to Figure 7 for the sensitivity of the private methods w.r.t. the number of features in the data set. We see that the RMSE of the data generating parameter \( \theta \) (divided by the number of dimensions of \( \theta \)) increases. The reason for this is two-fold: 1) The methods of [19] and [65] provide their privacy guarantees w.r.t. the number of features. While more noise has to be added for [19], the influence of the prior increases for [65] when the number of features increases for a fixed privacy budget. 2) A single sample from a posterior is of higher variance than the higher-dimensional the posterior is, negatively influencing OPS methods such as [65] and βD-Bayes.

Membership inference attacks  For \( \epsilon \in \{0.2, 1, 2, 7, 10, 20\} \), we run 10,000 rounds of the attack presented in Section 4. In Figure 8, we use the approach presented by [44] to estimate a lower bound on \( \epsilon \) given the false positive and negative rates of the attacks. Note that these lower bounds are unrealistic for \( \epsilon < 1 \). We see that, for any RMSE value, βD-Bayes achieves a lower practical bound on \( \epsilon \) than [19], which gives exact privacy guarantees.

Compute  While the final experimental results can be run within approximately two hours on a single Intel(R) Xeon(R) Gold 5118 CPU @ 2.30GHz core, the complete compute needed for the final results, debugging runs, and sweeps amounts to around 11 days.

Licenses  The UCI data sets are licensed under Creative Commons Attribution 4.0 International license (CC BY 4.0).
Figure 6: Test set predictive ROC-AUC of DP estimation for neural network classification as the number of observations $n$ increases on simulated and UCI data.

Figure 7: Parameter log RMSE of DP logistic regression (first row), test set predictive log RMSE of DP neural network regression (second row), and test set ROC-AUC of DP neural network classification (third row) as the number of features $d$ increases on simulated data with $n = 1000$.

Figure 8: Lower bound on $\varepsilon$ against log RMSE. Points correspond to values of $\varepsilon$. 
References


