Foundations of Bayesian Learning from Synthetic Data

Harrison Wilde
Department of Statistics
University of Warwick

Jack Jewson
Barcelona GSE
Universitat Pompeu Fabra

Sebastian Vollmer
Department of Statistics,
Mathematics Institute
University of Warwick

Chris Holmes
Department of Statistics
University of Oxford;
The Alan Turing Institute

Abstract

There is significant growth and interest in the use of synthetic data as an enabler for machine learning in environments where the release of real data is restricted due to privacy or availability constraints. Despite a large number of methods for synthetic data generation, there are comparatively few results on the statistical properties of models learnt on synthetic data, and fewer still for situations where a researcher wishes to augment real data with another party's synthesised data. We use a Bayesian paradigm to characterise the updating of model parameters when learning in these settings, demonstrating that caution should be taken when applying conventional learning algorithms without appropriate consideration of the synthetic data generating process and learning task at hand. Recent results from general Bayesian updating support a novel and robust approach to Bayesian synthetic-learning founded on decision theory that outperforms standard approaches across repeated experiments on supervised learning and inference problems.

1 Introduction

Privacy enhancing technologies comprise an area of rapid growth (The Royal Society, 2019). An important aspect of this field concerns publishing privatised versions of datasets for learning; it is known that simply anonymising the data is not sufficient to guarantee individual privacy (e.g. Rocher et al., 2019). We instead adopt the Differential Privacy (DP) framework (Dwork et al., 2006), to define working bounds on the probability that an adversary may identify whether a particular observation is present in a dataset, given that they have access to all other observations in the dataset. DP's formulation is context-dependent across the literature; here we amalgamate definitions regarding adjacent datasets from Dwork et al. (2014); Dwork and Lei (2009):

Definition 1 ((ε, δ)-differential privacy) *A randomised function or algorithm $\mathcal{K}$ is said to be (ε, δ)-differentially private if for all pairs of **adjacent, equally-sized** datasets D and D' that differ in one observation and all $S \subseteq \text{Range}(\mathcal{K})$,*

$$\Pr[\mathcal{K}(D) \in S] \leq e^{\varepsilon} \times \Pr\left[\mathcal{K}\left(D'\right) \in S\right] + \delta \qquad (1)$$

Current state-of-the-art approaches involve the privatisation of generative modelling architectures such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs) or Bayesian Networks. This is achieved through adjustments to their learning processes such that their outputs fulfil a DP guarantee specified at the point of training (e.g. Zhang et al., 2017; Xie et al., 2018; Jordon et al., 2018; Rosenblatt et al., 2020). Despite these contributions, a fundamental question remains regarding how, from a statistical perspective, one should learn from privatised synthetic data. Progress has been made for simple exponential family and regression models (Bernstein and Sheldon, 2018, 2019), but these model classes are of limited use in modern machine learning applications.

We characterise this problem for the first time via an adoption of the *M*-open world viewpoint (Bernardo and Smith, 2001) associated with model misspecification; unifying the privacy and synthetic data generation literature alongside recent results in generalised Bayesian updating (Bissiri et al., 2016) and minimum divergence inference (Jewson et al., 2018) to ask what it means to learn from synthetic data, and how can we improve upon our inferences and predictions given that we acknowledge its privatised synthetic nature?

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

This characterisation results in generative models that are 'misspecificed by design', owing to the constraints imposed upon their design by requiring the fulfilment of a DP guarantee. This inevitably leads to discrepancy between the learner's final model and the one that they would otherwise have formulated if not for this DP restriction. In real-world, finite data contexts where synthesis methods are often 'black-box' in nature, it is difficult for a learner to fully capture and understand the inherent differences in the underlying distributions of the real and synthetic data that they have access to.

There are two key insights that we explore in this paper following the characterisation above: Firstly, when left unchecked, the Bayesian inference machine learns model parameters minimising the Kullback-Leibler divergence (KLD) to the synthetic data generating process (S-DGP) (Berk et al., 1966; Walker, 2013) rather than the true data generating process (DGP); Secondly, robust inference methods offer improved performance by acknowledging this misspecification, where in some cases synthetic data can otherwise significantly hinder learning rather than helping it.

In order to investigate these behaviours, we experiment with models based on a mix of simulated-private and real-world data to offer empirical insights on the learning procedure when a varying amount of real data is available, and explore the optimalities in the amounts of synthetic data with which to augment this real data.

The contributions of our work are summarised below:

1. Learning from synthetic data can lead to unpredictable and negative outcomes, due to varying levels of model misspecification introduced by its generation and associated privacy constraints.
2. Robust Bayesian inference offers improvements over classical Bayes when learning from synthetic data.
3. Real and synthetic data can be used in tandem to achieve practical effectiveness through the discovery of desirable stopping points for learning, and optimal model configurations.
4. Consideration of the preferred properties of the inference procedure are critical; the specific task at hand can determine how best to use synthetic data.

We adopt a Bayesian standpoint throughout this paper, but note that many of the results also hold in the frequentist setting.

2 Problem Formulation

We outline the inference problem as follows,

- Let $x_{1:n}$ denote a training set of n exchangeable observations from Nature's true DGP, $F_0(x)$ with density $f_0(x)$ with respect to the Lebesque measure, such that $x_{1:n} \sim F_0(x)$; we suppose $x_i \in \mathbb{R}^d$. These observations are held privately by a data keeper K.
- K uses data $x_{1:n}$ to produce an (ε, δ)-differentially private synthetic data generating mechanism (S-DGP). With a slight abuse of notation we use $\mathcal{G}_{\varepsilon,\delta}(x_{1:n})$ to denote the S-DGP, noting that $\mathcal{G}_{\varepsilon,\delta}$ could be a fully generative model, or a private release mechanism that acts directly on the finite data $x_{1:n}$ (see discussion on the details of the S-DGP below). We denote the density of this S-DGP as $g_{\varepsilon,\delta}$.
- Let $f_\theta(x)$ denote a learner L's model likelihood for $F_0(x)$, parameterised by θ with prior $\tilde{\pi}(\theta)$, and marginal (predictive) likelihood $p(x) = \int_\theta f_\theta(x)\tilde{\pi}(\theta)d\theta$.
- L's prior may already encompass some other set of real-data drawn from F_0 leading to $\tilde{\pi}(\theta) = \pi(\theta \mid x^L_{1:n_L})$, for $n_L \geq 0$ prior observations.

We adopt a decision theoretic framework (Berger, 2013), in assuming that L wishes to take some optimal action $\hat{a}$ in a prediction or inference task; satisfying:

$$\hat{a} = \arg\max_{a \in \mathcal{A}} \int U(x, a) F_0(x) dx. \tag{2}$$

This is with respect to a user-specified utility-function $U(x, a)$ that evaluates actions in the action space $\mathcal{A}$, and makes precise L's desire to learn about F_0 in order to accurately identify $\hat{a}$.

Details of the synthetic data generation mechanism. In defining $\mathcal{G}_{\varepsilon,\delta}$, we believe it is important to differentiate between its two possible forms:

1. $\mathcal{G}_{\varepsilon,\delta}(x_{1:n}) = G_{\varepsilon,\delta}(z \mid x_{1:n})$: G is a privacy-preserving generative model fit on the real data, such as the PATE-GAN (Jordon et al., 2018), DP-GAN (Xie et al., 2018) or PrivBayes (Zhang et al., 2017). Privatised synthetic data is produced by injecting potentially heavy-tailed noise into gradient-based learning and/or through partitioned training leveraging marginal distributions, aggregations and subsets of the data. The S-DGP provides conditional independence between $z_{1:m}$ and $x_{1:m}$ and therefore no longer queries the real data after training.
2. $\mathcal{G}_{\varepsilon,\delta} = \int K_{\varepsilon,\delta}(x, dz) F_0(dx)$: A special case of this integral comprises the convolution of F_0 with some noise distribution H, such that $\mathcal{G}_{\varepsilon,\delta} = F_0 \star H_{\varepsilon,\delta}$. The sampling distribution is therefore not a function of the private data $x_{1:n}$. In this case, the number of samples that we can draw is limited to $m \leq n$ as drawing one data item requires using one sample of K's data. Examples of this formulation include the Laplace mechanism (Dwork et al., 2014) and transformation-based privatisation (Aggarwal and Yu, 2004).

The fundamental problem of synthetic learning. L wants to learn about F_0 but only has access to their prior $\tilde{\pi}(\theta)$ and to $z_{1:m} \sim \mathcal{G}_{\varepsilon,\delta}$, where $\mathcal{G}_{\varepsilon,\delta} \not\equiv F_0$. That is, the S-DGP $\mathcal{G}_{\varepsilon,\delta}(\cdot)$ is 'misspecified by design'. This claim is supported by a number of observations:

- L specifies a model $p(x)$ using beliefs about the target F_0 to be built using real data $x_{1:n}$, they are then constrained by a subsequently imposed requirement of guaranteeing DP which instead requires consideration of the resulting S-DGP $\mathcal{G}_{\varepsilon,\delta}$; this leads to an inevitable change in their beliefs such that the resulting model would be misspecified relative to the original 'best' model for the true DGP F_0.
- Correctly modelling a privacy preserving mechanism as part of an S-DGP such as a 'black-box' GAN or complex noise convolution is often intractable.
- There is inherent finiteness to real-world sensitive data contexts that makes it difficult for a generative model to capture the true DGP of even the non-private data it seeks to emulate. Considering sufficiently large quantities of data, and sufficiently flexible models, would in theory allow for the true DGP to be learnt, but relying on asymptotic behaviour in this setting is at odds with the definition of DP given that the identifiability of individuals would also naturally diminish as data availability increases. Moreover, for non-trivial high-dimensional models, the amount of data required to properly capture the DGP often becomes infeasible, regardless of the magnitude of DP constraints.

Therefore, the posterior predictive converges to a different distribution under real and synthetic data generating processes such that $p(x \mid z_{1:m\to\infty}) \not\equiv p(x \mid x_{1:n\to\infty})$. Learning from synthetic data is an intricate example of learning under model misspecification, where the misspecification is by K's design. It is important, as shown below, that this is recognised in the learning of models. Fortunately we can adapt recent advances in Bayesian inference under model misspecification to help optimise learning with respect to L's task.

2.1 Bayesian Inference under model misspecification

Bayesian inference under model misspecification has recently been formalised (Walker, 2013; Bissiri et al., 2016) and represents a growing area of research, see Watson and Holmes (2016); Jewson et al. (2018); Miller and Dunson (2018); Lyddon et al. (2018); Grünwald et al. (2017); Knoblauch et al. (2019) to name but a few. Traditional Bayes rule updating in this context can be seen as an approach that learns about the parameters of the model that minimises the logarithmic score, or equivalently, the Kullback-Leibler divergence (KLD) of the model from the DGP of the data (Berk et al., 1966; Walker, 2013; Bissiri et al., 2016), where $\mathrm{KLD}(g_{\varepsilon,\delta} \,\|\, f) = \int \log\left(g_{\varepsilon,\delta}/f\right) d\mathcal{G}_{\varepsilon,\delta}$.

As a result, if L updates their model $f_\theta(x)$ using synthetic data $z_{1:m} \sim \mathcal{G}_{\varepsilon,\delta}(x_{1:n})$, then as $m \to \infty$ they will be learning about the limiting parameter that minimises the KLD to the S-DGP:

$$\theta^{\mathrm{KLD}}_{\mathcal{G}_{\varepsilon,\delta}} = \arg\min_{\theta\in\Theta} \mathrm{KLD}\left(g_{\varepsilon,\delta}(\cdot) \,\|\, f_\theta(\cdot)\right), \tag{3}$$

and under regularity conditions the posterior distribution concentrates around that point, $\pi(\theta \mid z_{1:m}) \to \mathbf{1}_{\theta^{\mathrm{KLD}}_{\mathcal{G}_{\varepsilon,\delta}}}$ as $m \to \infty$.

Furthermore, this posterior will concentrate *away* from the model that is closest to F_0 in KLD, corresponding to the limiting model that would be learnt given an infinite real sample $x_{1:\infty}$ from F_0:

$$\theta^{\mathrm{KLD}}_{\mathcal{G}_{\varepsilon,\delta}} \not\equiv \theta^{\mathrm{KLD}}_{F_0} = \arg\min_{\theta\in\Theta} \mathrm{KLD}\left(f_0(\cdot) \,\|\, f_\theta(\cdot)\right). \tag{4}$$

This problem is exacerbated by the fact that S-DGP's can be prone to generating 'outliers' relative to the private data as a result of the injection of additional noise into their training or generation to ensure (ε,δ)-DP, combined with the fact that $\theta^{\mathrm{KLD}}_{\mathcal{G}_{\varepsilon,\delta}}$ is known to be non-robust to outliers (e.g. Basu et al., 1998). Therefore, given that as we collect more synthetic data our inference is no longer minimising the KLD towards F_0, we must carefully consider and investigate whether our inference is still 'useful' for learning about F_0 at all.

2.2 The approximation to F_0

Before we proceed any further we must consider what it means for data from $\mathcal{G}_{\varepsilon,\delta}$ to be 'useful' for learning about F_0. We can do so using the concepts of *scoring rules* and statistical divergence.

Definition 2 (Proper Scoring Rule) *The function $s : \mathcal{X} \times \mathcal{P}$ is a strictly proper scoring rule provided its difference function D satisfies*

$$D(f_0 \,\|\, f) = \mathbb{E}_{x\sim f_0}\left[s(x, f(\cdot))\right] - \mathbb{E}_{x\sim f_0}\left[s(x, f_0(\cdot))\right]$$

$$D(f_1 \,\|\, f_2) \geq 0,\ D(f \,\|\, f) = 0 \text{ for all } f, f_1, f_2 \in \mathcal{P}(x)$$

$$\mathcal{P}(x) := \left\{ f(x) : f(x) \geq 0\, \forall x \in \mathcal{X}, \int_{\mathcal{X}} f(x)dx = 1 \right\},$$

Where the function D measures a distance between two probability distributions. $s(x, f)$ arises as the divergence and is minimised when $f_0 = f$ (Gneiting and Raftery, 2007; Dawid, 2007). A further advantage of

this representation is that it allows for the minimisation of $D(f_0 \,\|\, \cdot)$ using only samples from F_0,

$$\arg\min_{f\in\mathcal{F}} D(f_0 \,\|\, f) = \arg\min_{f\in\mathcal{F}} \mathbb{E}_{x\sim f_0}[s(x, f(\cdot))]$$
$$\xleftarrow[n\to\infty]{} \frac{1}{n}\sum_{i=1}^{n} s(x_i, f(\cdot)), \quad x_i \sim F_0 \tag{5}$$

Henceforth, we define any concepts of closeness (or 'usefulness') in terms of a chosen divergence D and associated scoring rule s, where the approximating density f is given by the predictive inferences resulting from synthetic $z_{1:m} \sim \mathcal{G}_{\varepsilon,\delta}$. Given that inference using $\mathcal{G}_{\varepsilon,\delta}$ is no longer able to exactly capture f_0, L can use this notion of closeness to define what aspects of F_0 they are most concerned with capturing. The importance of this specification is illustrated in Section 4.

3 Improved learning from the S-DGP

The classical assumptions underlying statistics are that minimising the KLD is the optimal way to learn about the DGP, and that more observations provide more information about this underlying DGP; such logic does not necessarily apply here. L wishes to learn about the private DGP, F_0, but must rely on observations from the S-DGP $\mathcal{G}_{\varepsilon,\delta}$ to do so. In this section we acknowledge this setting to propose a framework for improved learning from synthetic data. In so doing we pose the following question and detail our solutions in turn: Given the scoring criteria D, is $\theta^{\text{KLD}}_{\mathcal{G}_{\varepsilon,\delta}}$ the best the learner can do?

1. Can the robustness of the *learning procedure* be improved to better approximate F_0 by acknowledging the misspecification and outlier prone nature of $z_{1:m}$?
2. Starting from the prior predictive, $p(x)$, for a given learning method, when does learning using $z \sim \mathcal{G}_{\varepsilon,\delta}$ stop improving inference for $F_0(x)$? That is, when

$$\mathbb{E}_z\left[D\left(f_0(\cdot) \,\|\, p(\cdot \mid z_{1:j+1})\right)\right] > \mathbb{E}_z\left[D\left(f_0(\cdot) \,\|\, p(\cdot \mid z_{1:j})\right)\right]$$

3.1 General Bayesian Inference

In order to address these issues we adopt a general Bayesian, minimum divergence paradigm for inference (Bissiri et al., 2016; Jewson et al., 2018) inspired by model misspecification, where L can coherently update beliefs about their model parameter θ from prior $\pi(\theta)$ to posterior $\pi(\theta \,|\, z_{1:m})$ using:

$$\pi^{\ell}(\theta \,|\, z_{1:m}) \propto \frac{\tilde{\pi}(\theta)\exp\left(-\sum_{i=1}^{m}\ell(z_j, f_\theta)\right)}{\int \tilde{\pi}(\theta)\exp\left(-\sum_{i=1}^{m}\ell(z_j, f_\theta)\right)d\theta}, \tag{6}$$

where $\ell(z, f_\theta)$ is the loss function used by L for inference. Note that in this formulation, the logarithmic score $\ell_0(z, f_\theta) = -\log f_\theta(z)$ recovers traditional Bayes rule updating. The predictive distribution associated with such a posterior and the model f_θ is:

$$p^{\ell}(x|z_{1:m}) = \int f_\theta(x)\pi^{\ell}(\theta|z_{1:m})d\theta \tag{7}$$

3.2 Robust Bayes and dealing with outliers

In the absence of the ability to correctly model the S-DGP, robust statistics (see e.g. Berger et al., 1994) provide an alternative option to guard against artefacts of the generated synthetic data. We can gain increased robustness in our learning procedure to data $z_{1:m}$ by changing the loss function $\ell(z, f_\theta)$ used for inference in Eq. (6). We consider two alternative loss functions to the standard logarithmic score underpinning standard Bayesian statistics,

$$\ell_w(z, f_\theta) := -w\log f_\theta(z) \tag{8}$$
$$\ell^{(\beta)}(z, f_\theta) := \frac{1}{\beta+1}\int f_\theta(y)^{\beta+1}dy - \frac{1}{\beta}f_\theta(z)^{\beta}. \tag{9}$$

Here, $\ell_w(z, f_\theta)$ introduces a learning parameter $w > 0$ into the Bayesian update (e.g. Lyddon et al., 2018; Grünwald et al., 2017; Miller and Dunson, 2018; Holmes and Walker, 2017). Down-weighting, $w < 1$ will generally produce a less confident posterior than in the case of traditional Bayes' rule, with a greater dependence on the prior. Conversely, $w > 1$ will have the opposite effect. The value of w can have ramifications for inference and prediction (Rossell and Rubio, 2018; Grünwald et al., 2017). However, we note that as the sample size grows using the weighted likelihood posterior will still learn about θ^*_G if w is fixed. Choosing $w = {}^{e}/_{m}$ instead can be seen to average the log-likelihood, with e providing a notion of effective sample size.

Alternatively, minimising $\ell^{(\beta)}(x, f(\cdot))$ in expectation over the DGP is equivalent to minimising the β-divergence (βD) (Basu et al., 1998). Therefore, analogously to the KLD and the log-score, using $\ell^{(\beta)}(x, f(\cdot))$ (Bissiri et al., 2016; Jewson et al., 2018; Ghosh and Basu, 2016) produces a Bayesian update targeting:

$$\theta^{\beta\text{D}}_{\mathcal{G}_{\varepsilon,\delta}} := \arg\min_{\theta\in\Theta} \beta\text{D}\left(g_{\varepsilon,\delta}(\cdot) \,\|\, f_\theta\right). \tag{10}$$

As $\beta \to 0$, then $\beta\text{D} \to$ KLD, but as β increases it provides increased robustness through skepticism of new observations relative to the prior. We demonstrate the robustness properties of the βD in some simple scenarios in the Supplementary material (see A.1) and refer the reader to e.g. Knoblauch et al. (2019, 2018) for further examples. We note there are many possible divergences providing greater robustness properties than the KLD, e.g. Wasserstein or Stein discrepancy (Barp et al., 2019), but for our exposition we focus on the βD for its convenience and simplicity.

A key difference between the two robust loss functions considered above is that while $\ell_w(z, f_\theta)$ down-weights the log-likelihood of each observation equally, $\ell^{(\beta)}(x, f(\cdot))$ does so adaptively, based on how likely the new observation is under the current inference (Cichocki et al., 2011). It is this adaptive down-weighting that allows the βD to target a different limiting parameter to $\theta^{\text{KLD}}_{\mathcal{G}_{\varepsilon,\delta}}$. This, in particular, allows the βD to be robust to outliers and/or heavy tailed contaminations. As a result, we believe that

$$D\left(F_0 \,\|\, f_{\theta^{\beta\text{D}}_{\mathcal{G}_{\varepsilon,\delta}}}\right) < D\left(F_0 \,\|\, f_{\theta^{\text{KLD}}_{\mathcal{G}_{\varepsilon,\delta}}}\right)$$

across a wide range of S-DGPs, i.e. the βD minimising approximation to $\mathcal{G}_{\varepsilon,\delta}$ is a better approximation of F_0 than the KLD minimising approximation. Proving that this is the case in general proves complicated and is hampered by the intractability both of the βD minimisation and of many popular S-DGP's. However, we further justify this claim in A.1.2 where we show that for the prevalent Laplace DP mechanism, this holds uniformly over the privacy level parameterised by the scaling parameter of the Laplace distribution λ for $D = \text{KLD}$.

A strength of the βD is that, unlike standard robust methods using heavier tailed models or losses (Berger et al., 1994; Huber and Ronchetti, 1981; Beaton and Tukey, 1974), $\ell^{(\beta)}(x, f(\cdot))$ does not change the model used for inference. In the absence of any specific knowledge about the S-DGP, updating using the βD maintains the model L *would* have used to estimate F_0, but updates its parameters robustly. This also has advantages in the data combination scenario where L is combining inferences from their own private data $x^L_{1:n_L}$ with synthetic data $z_{1:m}$. They can maintain the same model for both datasets, with the same model parameters, yet update robustly about $z_{1:m}$ whilst still using the log-score for $x^L_{1:n_L}$ (i.e. to condition their prior $\tilde{\pi}(\theta)$).

3.3 The Learning Trajectory

The concept of closeness provided by D allows us to consider how L's approximation to F_0 changes as more data is collected from the S-DGP. To do so we consider the 'learning trajectory', a function in m of the expected divergence to F_0 of inference using m observations from $\mathcal{G}_{\varepsilon,\delta}$:

$$T_\ell\left(m; D, f_0, g_{\varepsilon,\delta}\right) = E_z\left[D\left(f_0(\cdot) \,\|\, p^\ell(\cdot \mid z_{1:m})\right)\right],$$

where $p^\ell(\cdot \mid z_{1:m})$ is the general Bayesian posterior predictive distribution (using ℓ) based on (synthetic) data $z_{1:m}$. This 'learning trajectory' traces the path taken by Bayesian inference from its prior predictive ($m = 0$) towards the synthetic data's DGP under increasing amounts of data (e.g. Figure 1; Section A.4).

We provide a proposition that says using more data and approaching the limit, $\theta^{\text{KLD}}_{\mathcal{G}_{\varepsilon,\delta}}$ is not necessarily the optimal target to learn about according to criteria D.

Proposition 1 (Suboptimality of the S-DGP) *For S-DGP $\mathcal{G}_{\varepsilon,\delta}$, model $f_\theta(\cdot)$, and divergence D, there exists prior $\tilde{\pi}(\theta)$, private DGP F_0 and at least one value of $0 \leq m < \infty$ such that,*

$$T_{\ell_0}\left(m; D, f_0, g_{\varepsilon,\delta}\right) \leq D\left(F_0 \,\|\, f_{\theta^{\text{KLD}}_{\mathcal{G}_{\varepsilon,\delta}}}\right), \tag{11}$$

where $\theta^{\text{KLD}}_G := \arg\min_{\theta\in\Theta} \text{KLD}(\mathcal{G}_{\varepsilon,\delta} \,\|\, f_\theta)$ and T_{ℓ_0} is the learning trajectory of the Bayesian posterior predictive distribution (using ℓ_0) based on (synthetic) data $z_{1:m}$, see Eq. (7).

The proof of Proposition 1 involves a simple counter example in which the prior is a better approximation to F_0 according to divergence D than $\mathcal{G}_{\varepsilon,\delta}$. While trivial, this could reasonably occur if L has strong, well-calibrated expert belief judgements, or if they have a considerable amount of their own data before beginning to incorporate synthetic data. Furthermore, we argue next that by considering this learning trajectory path between the prior and the S-DGP $\mathcal{G}_{\varepsilon,\delta}$, in terms of the number of synthetic observations m, it is possible to get even 'closer' to F_0 in terms of a chosen D.

Changing the divergence used for inference as suggested in Section 3.2 changes these trajectories by changing their limiting parameter. However, Proposition 1, which considered learning minimising the KLD, can equally be shown for learning minimising the βD (see A.2.1). In the following sections we talk generally about optimising such a trajectory for a given learning method, before focusing on comparisons in these methods and the resulting trajectories in Section 4.

3.4 Optimising the Learning Trajectory

The insights from the previous section raise two questions for a learner L using synthetic data:

1. Is synthetic data able to improve L's inferences about F_0 according to divergence D? If so,
2. What is the optimal quantity to use for getting the learner closest to F_0 according to D?

Both questions can be solved by the estimation of

$$m^* := \underset{0\leq m\leq M}{\arg\min}\, T_\ell\left(m; D, f_0, g_{\varepsilon,\delta}\right), \tag{12}$$

However, clearly the learner never has access to the data generating density. Instead we take advantage of the representation of proper scoring rules and propose

using a 'test set' $x'_{1:N} \sim F_0$ to estimate

$$\hat{m} := \underset{0\leq m\leq M}{\arg\min} \frac{1}{N}\frac{1}{B}\sum_{b=1}^{B}\sum_{j=1}^{N} s(x'_j, p^{\ell}(\cdot|z^{(b)}_{1:m})) \quad (13)$$
$$\text{with } \{z^{(b)}_{1:m}\}_{b_{1:B}} \sim \mathcal{G}_{\varepsilon,\delta}.$$

As such we use a small amount of data from F_0 to guide the synthetic data inference towards F_0. We consider two procedures for doing so: Firstly, by tailoring to L's specific inference problem, and when this is not possible, putting the onus on K to evaluate the general ability of their S-DGP to capture F_0. Before so doing the following remark briefly considers optimising the learning trajectory for a specific stream of data rather than average over the S-DGP.

Remark 1 *We note that although previously the learning trajectory was defined to average across samples from the* S-DGP*, there may be cases where it is preferable to define the learning trajectory based on a concrete stream of data $z_{1:m}$. In such a scenario we may consider data-dependent*

$$m'^{*}(z_{1:M}) := \underset{0\leq m\leq M}{\arg\min}\, D\left(f_0(\cdot)\,\|\,p^{\ell}(\cdot \mid z_{1:m})\right). \quad (14)$$

This can be seen as learning the optimal point to stop learning for a specific stream of data rather than the optimal size for the average synthetic dataset. Such a setting introduces an undesirable dependence between $m'^{*}(z_{1:M})$ and the ordering of the data, but this can be somewhat mitigated by averaging different realisations of the synthetic data (see Prop. 2 and A.2.3/A.5.2 for more details). For the rest of this paper, however, we focus on the more generally defined m^* from Eq. (12)*.*

3.4.1 Optimising for L's inference

In the first instance we consider the learner L has access to an independent test set $x'_{1:N} \sim F_0$ allowing them to calculate the $\hat{m}$ associated with their specific learning trajectory. Two potential sources are for $x'_{1:N}$ are 1) for L to sacrifice some of their own data $x^{L}_{1:n_L}$ when constructing their prior, or 2) require that K hold $x'_{1:N}$ out when it trains the S-DGP, which can then be queried by L in order to estimate $\hat{m}$. Clearly K is not able to share the observations with L as this would violate the DP guarantee. Instead a secure protocol for two-way communication between L's model and K's test set must be established; promising directions include (Cormode et al., 2019; de Montjoye et al., 2018) and a practical use case (UK HDR Alliance, 2020).

3.4.2 A Broader Study

When the previous, problem, and data-specific methods are not available we have to fall back on a broader study. Here we recommend that alongside releasing synthetic data, K optimises the learning trajectory themselves, under some default model, loss and prior setting by repeatedly partitioning $x_{1:n}$ into test and training sets. For example, when releasing classification data, K could release an $\hat{m}$ associated with logistic regression and BART, for the log-score, under some vaguely informative priors, providing learners an idea of what to expect in terms of utility from the synthetic data. Whilst this is less tailored to any specific inference problem, it still allows K to communicate a broad measure of the quality of its released data for learning about F_0, and is advisable given our results.

3.5 Posthoc improvement through averaging

Once $\hat{m}$ has been estimated, the question then remains of how to conduct inference given $\hat{m}$. In particular, if more synthetic data is available (e.g. $\hat{m} << M$ or sampling $z_{1:m}, m \to \infty$ from a GAN), we can average the posterior predictive distribution across different realisations, using $\hat{m}$ each time, ensuring we do not waste any synthetic data. Jensen's inequality allows us to prove that performance of the predictive distribution will not diminish if we consider convex proper scoring rules such as the logarithmic score:

Proposition 2 (Predictive Averaging) *Given divergence D with convex scoring rule, averaging over different realisations of the posterior predictive depending on different synthetic data sets cannot deteriorate inference:*

$$\mathbb{E}_z D\left(F_0 \,\|\, \frac{1}{B}\sum_{b=1}^{B}\tilde{p}\left(x|z^{(b)}_{1:m}\right)\right) \overset{\textit{Jensen's ineq.}}{\leq}$$
$$\mathbb{E}_z \frac{1}{B}\sum_{b=1}^{B} D\left(F_0 \,\|\, \tilde{p}\left(x|z^{(b)}_{1:m}\right)\right) = \mathbb{E}_z D\left(F_0 \,\|\, \tilde{p}\left(x|z^{(b)}_{1:m}\right)\right).$$

A more detailed proof is provided in A.2.2. The significance of this is that more synthetic data cannot hinder the predictive distribution, but only if we do not naïvely use all of it to learn at once.

4 Experimental Setup and Results

In order to investigate the concepts and methodologies outlined above, we consider two experiment types:

1. Learning the location and variance of a univariate Gaussian distribution.
2. Using Bayesian logistic regression for binary classification on a selection of real-world datasets.

In these contexts we investigate the learning trajectory of classical Bayesian updating alongside the robust ad-

justments discussed in Section 3.2. In order to draw comparisons between these methods, we study the trajectories' dependence on values spanning a grid of data quantities n_L and m, robustness parameters w and β, prior values, and the parameters of chosen DP mechanisms (see A.6.3 for full experimental specifications). The varying amounts of non-private data available to L were used to construct increasingly informative priors $\tilde{\pi}(\theta) = \pi(\theta \mid x^L_{1:n_L})$ through the use of standard Bayesian updating, as robustness is not required when learning using data drawn from F_0. Learning trajectories are then estimated utilising an unseen dataset $x'_{1:N}$ (mimicking either K's data or some subset of L's data not used in training).

To this end we use optimised MCMC sampling schemes (e.g. Hoffman and Gelman, 2014) to sample from all of the considered posteriors in each experiment's case; drawing comparisons across the grid laid out above and repeating experiments to mitigate any sources of noise. This results in an extensive computational task, made feasible through a mix of Julia's Turing PPL (Ge et al., 2018), MLJ (Blaom et al., 2020) and Stan (Carpenter et al., 2017).

The majority of the experiments are carried out with $\varepsilon = 6$, which is seen to be a realistic value respective of practical applications (Lee and Clifton, 2011; Erlingsson et al., 2014; Tang et al., 2017; Differential Privacy Team at Apple, 2017) and upon observation of the relationship between privacy and misspecification shown in Figure A.4. We evaluate our experiments according to the divergences and scoring rules discussed in detail in A.6.2, presented across a range of the figures in the main text and supplementary material.

4.1 Simulated Gaussian Models

We first introduce a simple but illustrative simulated example in which we infer the parameters of a Gaussian model $f_\theta = \mathcal{N}(\mu, \sigma^2)$ where $\theta = (\mu, \sigma^2)$. We place conjugate priors on θ with $\sigma^2 \sim \text{InverseGamma}(\alpha_p, \beta_p)$ and $\mu \sim \mathcal{N}(\mu_p, \sigma_p \times \sigma)$ respectively. We consider $x_{1:n}$ drawn from DGP $F_0 = \mathcal{N}(0, 1^2)$ and adopt the Laplace mechanism (Dwork et al., 2014) to define our S-DGP. This mechanism works by perturbing samples drawn from the DGP with noise drawn from the Laplace distribution of scale λ, calibrated via the sensitivity $\mathcal{S}$ of the DGP in order to provide $(\varepsilon, 0)$-DP per the Laplace mechanism's definition with $\varepsilon = \mathcal{S}/\lambda$. To achieve finiteness of $\mathcal{S}$ in this case, we adjust our model to be that of a truncated Gaussian; restricting its range to $\pm 3\sigma$ to allow for meaningful ε's to be calculated under the Laplace mechanism (See A.3 for a proof of this result).

We then compare and evaluate the empirical performance of the competing methods defined below (see A.6.1.1 for explicit formulations):

1. The standard likelihood adjusted with an additional reweighting parameter w as in Eq. (8).
2. The posterior under the βD loss as in Eq. (9).
3. The 'Noise-Aware' likelihood where the S-DGP can be tractably modelled using the Normal-Laplace convolution (Reed, 2006; Amini and Rabbani, 2017).

4.1.1 Results and Discussion

We observe that three different categories of learning trajectory occur across the models; these are illustrated in the 'branching' plots in Figure 1 (explained and analysed further in A.6.4 and A.6.5):

1. The prior $\tilde{\pi}$ is sufficiently inaccurate or uninformative (in this case due to low n_L) such that the synthetic data continues to be useful across the range of m we consider. As a result the learning trajectory is a monotonically decreasing curve in the criteria of interest.
2. A turning point is observed; synthetic data initially brings us closer to F_0 before the introduction of further synthetic observations moves the inference away. We see that in the majority of cases these trajectories lie under the limiting KLD and βD approximations to $\mathcal{G}_{\varepsilon,\delta}$ demonstrating the efficacy of 'optimising the learning trajectory' through the existence of these optimal turning points.
3. The final scenario occurs under a sufficiently informative prior $\tilde{\pi}$ (here due to a large n_L) such that synthetic data is not observably of any use at all; it can be seen to immediately cause model performance to deteriorate.

We can further quantify the turning points which are perhaps the most interesting characteristic of these experiments. To do this we formulate bootstrapped averages of the number of 'effective real samples' that correspond to the estimated optimal quantity of synthetic data. This is done by comparing these minima with the black curves in the 'branching' plots representing the learning trajectory under an increasing n_L and $m = 0$ (see A.6.4 for more details). These calculations are shown in Figure 2 (and discussed further in A.6.5).

In general, we observe a significant increase in performance from the βD (see Figures 1, 3), indicated by its proximity to even the noise-aware model at lower values of n_L, alongside more modest improvements from reweighting methods. The βD achieves more desirable minimum-trajectory log score, KLD and Wasserstein values compared to other model types; exhibiting greater robustness to larger amounts of synthetic data where other approaches lose out significantly.

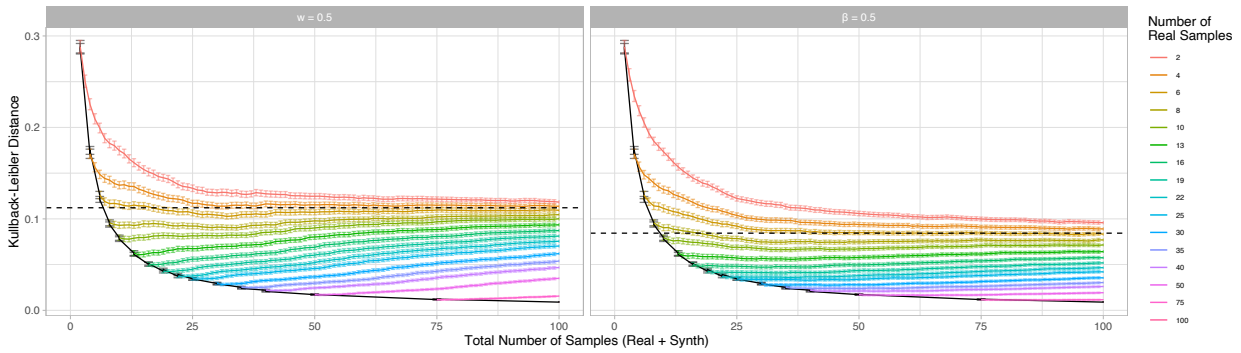


Figure 1: Shows how the KLD to F_0 changes as we add more synthetic data, starting with increasing amount of real data. This demonstrates the effectiveness of the optimal discovered βD configuration compared to the the closest alternative traditional model in terms of performance with down-weighting $w = 0.5$; the black dashed line illustrates $\text{KLD}(F_0 \,\|\, f_{\theta^*})$ for $\theta^* = \theta^{\text{KLD}}_{\mathcal{G}_{\varepsilon,\delta}}$ (left) and $\theta^* = \theta^{\beta\text{D}}_{\mathcal{G}_{\varepsilon,\delta}}$ (right), representing the approximation to F_0 given an infinite sample from $\mathcal{G}_{\varepsilon,\delta}$ under the two learning methods, and exhibiting the superiority of the βD.

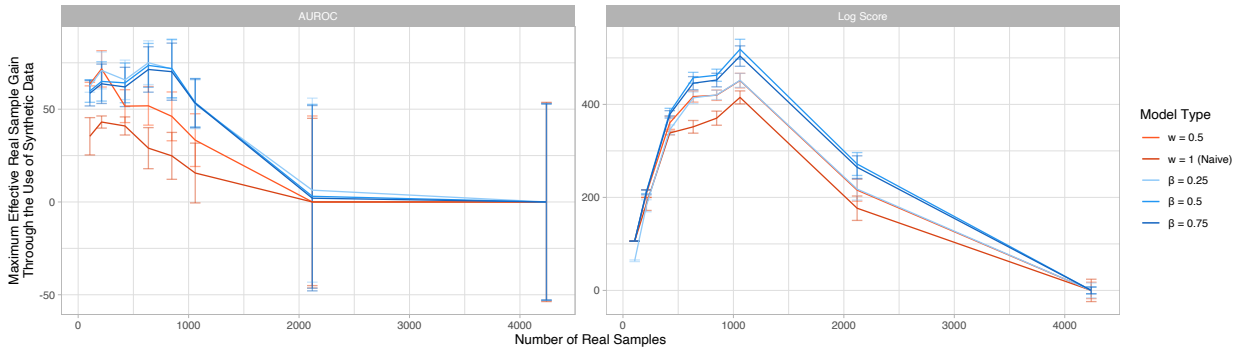
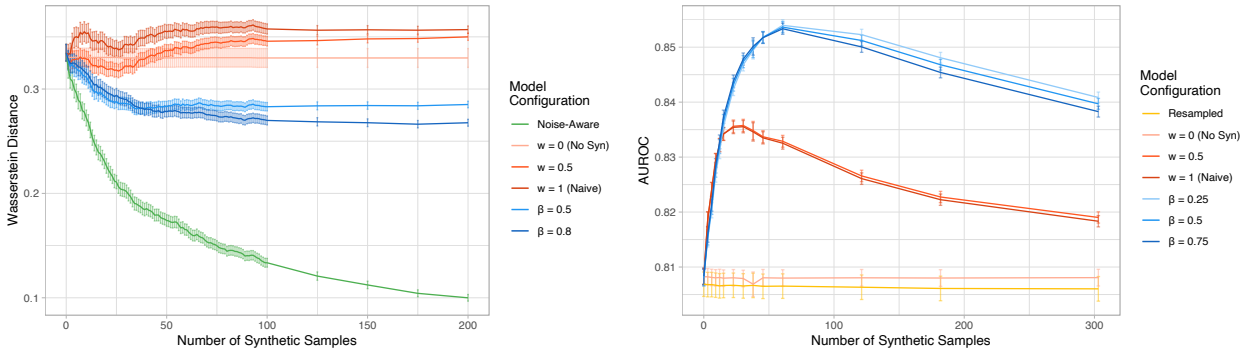


Figure 2: Shows the effective number of real samples gained through optimal $\hat{m}$ synthetic observations alongside varying amounts of real data usage with respect to the AUROC and log-score performance criteria. These are calculated and presented here via bootstrapped averaging under a logistic regression model learnt on the Framingham dataset. The amount of effective real samples is significantly affected by the learning task's criteria.



(a) Simulated Gaussian, real $n = 10$

(b) UCI Heart Dataset, real $n = 30$ (10%)

Figure 3: Given a fixed real amount of data, we can compare model performances directly by focusing on one of the 'branches' in the class of diagrams shown in Figures 1 & 4, to see that the βD's performance falls between that of the noise-aware model and the other models, exhibiting robust and desirable behaviour across a range of β. Naïve and reweighting-based approaches fail to gain significantly over not using synthetic data (shown by $w = 0$'s flat trajectory); the resampled model in the logistic case can also be seen to perform very poorly in comparison to models that leverage the synthetic data.

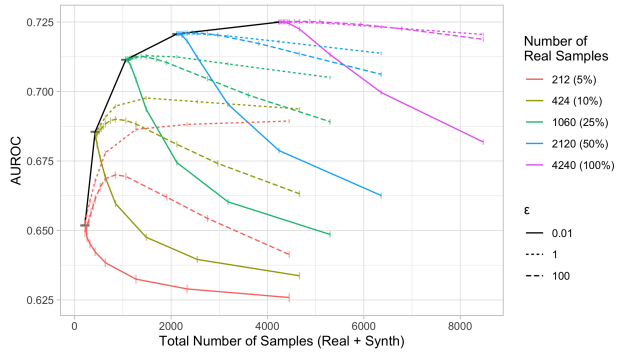


Figure 4: This plot illustrates an interesting and important observation made when varying ε for a GAN based model, we observe that there is a privacy 'sweet-spot' around $\varepsilon = 1$ whereby more private data performs better than less private data (see the curves for $\varepsilon = 100$ which essentially represent non-private data).[1]

4.2 Logistic Regression

We now move on to a more prevalent and practical class of models that also exhibit the potentially dangerous behaviours of synthetic data in real-world contexts, via datasets concerning subjects that have legitimate potential privacy concerns. Namely, we build logistic regression models for the UCI Heart Disease dataset (Dua and Graff, 2017) and the Framingham Cohort dataset (Splansky et al., 2007). Clearly, we are now only able to access the empirical distribution F_{n^*}, where n^* is the total amount of data present in each dataset. We use $x_{1:n^T}$ to train an instance of the aforementioned PATE-GAN $\mathcal{G}_{\varepsilon,\delta}$ and keep back $x_{1:(n^*-n^T)}$ for evaluation; we then draw synthetic data samples $z_{1:m} \sim \mathcal{G}_{\varepsilon,\delta}$. As before, we investigate how the learning trajectories are affected across the experimental parameter grid.

Again, we consider learning using ℓ_w and ℓ_β applied to the logistic regression likelihood, f_θ (see A.6.1.2 for exact formulations). In this case we cannot formulate a 'Noise-Aware' model due to the black-box nature of the GAN, highlighting the reality of the model misspecification setting we find ourselves in aside from simple or simulated examples. We can instead define a 'resampled' model that recycles the real data used in formulating the prior.

4.2.1 Results and Discussion

Here the learning trajectories are defined with respect to the AUROC as well as the log score; whilst not technically a divergence, this gives us a decision theoretic criteria to quantify the closeness of our inference to F_0.

[1]This plot exhibits the effect under the βD model on the Framingham dataset with $\beta = 0.5$, but is observable across all model types in both AUROC and log score.

Referring to Figures 2, 3 and 4 we see that the learning trajectories observed in this more realistic example mirror those observed in our simulated Gaussian experiments. There are however some cases in which the reweighted posterior outperforms the βD, and we see large discrepancies in $\hat{m}$ when comparing log score to AUROC values, reinforcing the importance of carefully defining the learning task to prioritise.

Additionally, experiments using synthetic data from a GAN offer the unique observation that performance can actually improve as ε decreases. We believe this is due to potential mode collapse in the GAN learning process on imbalanced datasets, and concentrations of realisations of $\mathcal{G}_{\varepsilon,\delta}$ as the injected noise increases such that a small number of synthetic samples can actually be *more* representative of F_0 than even the real data. This effect is short-lived as more synthetic observations are used in learning, as presumably these samples then become over-represented through the posterior distribution and performance begins to deteriorate, see Figure 4 for performance comparisons across ε.

5 Conclusions

We consider foundations of Bayesian learning from synthetic data that acknowledge the intrinsic model misspecification and learning task at hand. Contrary to traditional statistical inferences, conditioning on increasing amounts of synthetic data is not guaranteed to help you learn about the true data generating process or make better decisions. Down-weighting the information in synthetic data (either using a weight w or divergence βD) provides a principled approach to robust optimal information processing and warrants further investigation. Further work could consider augmenting these general robust techniques with tailored adjustments to inference based on the specific synthetic data acknowledging the discrepancies between its generating density and the Learner's true target.

Acknowledgements

HW is supported by the Feuer International Scholarship in Artificial Intelligence. JJ was funded by the Ayudas Fundación BBVA a Equipos de Investigación Cientifica 2017 and Government of Spain's Plan Nacional PGC2018-101643-B-I00 grants whilst working on this project. SJV is supported by The Alan Turing Institute (EPSRC grant EP/N510129/) and the University of Warwick IAA funding. CH is supported by The Alan Turing Institute, Health Data Research UK, the Medical Research Council UK, the Engineering and Physical Sciences Research Council (EPSRC) through the Bayes4Health programme Grant EP/R018561/1,

and AI for Science and Government UK Research and Innovation (UKRI).

References

Charu C Aggarwal and Philip S Yu. A condensation approach to privacy preserving data mining. In *Advances in Database Technology - EDBT 2004*, pages 183–199. Springer Berlin Heidelberg, 2004.

Zahra Amini and Hossein Rabbani. Letter to the editor: Correction to "the normal-laplace distribution and its relatives". *Communications in Statistics-Theory and Methods*, 46(4):2076–2078, 2017.

Alessandro Barp, Francois-Xavier Briol, Andrew Duncan, Mark Girolami, and Lester Mackey. Minimum stein discrepancy estimators. In *Advances in Neural Information Processing Systems*, pages 12964–12976, 2019.

Ayanendranath Basu, Ian R Harris, Nils L Hjort, and MC Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85 (3):549–559, 1998.

Albert E Beaton and John W Tukey. The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16(2):147–185, 1974.

James O Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Science & Business Media, March 2013.

James O Berger, Elías Moreno, Luis Raul Pericchi, M Jesús Bayarri, José M Bernardo, Juan A Cano, Julián De la Horra, Jacinto Martín, David Ríos-Insúa, Bruno Betrò, et al. An overview of robust bayesian analysis. *Test*, 3(1):5–124, 1994.

Robert H Berk et al. Limiting behavior of posterior distributions when the model is incorrect. *The Annals of Mathematical Statistics*, 37(1):51–58, 1966.

José M Bernardo and Adrian FM Smith. Bayesian theory, 2001.

Garrett Bernstein and Daniel R Sheldon. Differentially private bayesian inference for exponential families. In *Advances in Neural Information Processing Systems*, pages 2919–2929, 2018.

Garrett Bernstein and Daniel R Sheldon. Differentially private bayesian linear regression. In *Advances in Neural Information Processing Systems*, pages 523–533, 2019.

PG Bissiri, CC Holmes, and Stephen G Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016.

Anthony D Blaom, Franz Kiraly, Thibaut Lienart, Yiannis Simillides, Diego Arenas, and Sebastian J Vollmer. MLJ: A Julia package for composable Machine Learning. July 2020.

Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

Andrzej Cichocki, Sergio Cruces, and Shun-ichi Amari. Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy*, 13(1):134–170, 2011.

Graham Cormode, Tejas Kulkarni, and Divesh Srivastava. Answering range queries under local differential privacy, 2019.

A Philip Dawid. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59 (1):77–93, 2007.

Yves-Alexandre de Montjoye, Sébastien Gambs, Vincent Blondel, Geoffrey Canright, Nicolas de Cordes, Sébastien Deletaille, Kenth Engø-Monsen, Manuel Garcia-Herranz, Jake Kendall, Cameron Kerry, Gautier Krings, Emmanuel Letouzé, Miguel Luengo-Oroz, Nuria Oliver, Luc Rocher, Alex Rutherford, Zbigniew Smoreda, Jessica Steele, Erik Wetter, Alex Sandy Pentland, and Linus Bengtsson. On the privacy-conscientious use of mobile phone data. *Sci Data*, 5: 180286, December 2018.

Differential Privacy Team at Apple. Learning with privacy at scale. 2017.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 371–380, 2009.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067, 2014.

Hong Ge, Kai Xu, and Zoubin Ghahramani. Turing: a language for flexible probabilistic inference. In *In-*

ternational Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain, pages 1682–1690, 2018. URL `http://proceedings.mlr.press/v84/ge18b.html`.

Abhik Ghosh and Ayanendranath Basu. Robust bayes estimation using the density power divergence. *Annals of the Institute of Statistical Mathematics*, 68 (2):413–437, 2016.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102 (477):359–378, 2007.

Peter Grünwald, Thijs Van Ommen, et al. Inconsistency of bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069–1103, 2017.

Matthew D Hoffman and Andrew Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15 (1):1593–1623, 2014.

CC Holmes and SG Walker. Assigning a value to a power likelihood in a general bayesian model. *Biometrika*, 104(2):497–503, 2017.

Peter J Huber and EM Ronchetti. Robust statistics, series in probability and mathematical statistics, 1981.

Jack Jewson, Jim Smith, and Chris Holmes. Principles of Bayesian inference using general divergence criteria. *Entropy*, 20(6):442, 2018.

James Jordon, Jinsung Yoon, and Mihaela van der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations*, 2018.

Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. Doubly robust Bayesian inference for non-stationary streaming data using β-divergences. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 64–75, 2018.

Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. Generalized variational inference. *arXiv preprint arXiv:1904.02063*, 2019.

Jaewoo Lee and Chris Clifton. How much is enough? choosing ε for differential privacy. In *International Conference on Information Security*, pages 325–340. Springer, 2011.

SP Lyddon, CC Holmes, and SG Walker. General bayesian updating and the loss-likelihood bootstrap. *Biometrika*, 2018.

Jeffrey W Miller and David B Dunson. Robust bayesian inference via coarsening. *Journal of the American Statistical Association*, pages 1–13, 2018.

William J Reed. The normal-laplace distribution and its relatives. In *Advances in distribution theory, order statistics, and inference*, pages 61–74. Springer, 2006.

Luc Rocher, Julien M Hendrickx, and Yves-Alexandre De Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature communications*, 10(1):1–9, 2019.

Lucas Rosenblatt, Xiaoyan Liu, Samira Pouyanfar, Eduardo de Leon, Anuj Desai, and Joshua Allen. Differentially private synthetic data: Applied evaluations and enhancements. *arXiv preprint arXiv:2011.05537*, 2020.

David Rossell and Francisco J Rubio. Tractable bayesian variable selection: beyond normality. *Journal of the American Statistical Association*, 113(524): 1742–1758, 2018.

Greta Lee Splansky, Diane Corey, Qiong Yang, Larry D Atwood, L Adrienne Cupples, Emelia J Benjamin, Ralph B D'Agostino Sr, Caroline S Fox, Martin G Larson, Joanne M Murabito, et al. The third generation cohort of the national heart, lung, and blood institute's framingham heart study: design, recruitment, and initial examination. *American journal of epidemiology*, 165(11):1328–1335, 2007.

Jun Tang, Aleksandra Korolova, Xiaolong Bai, Xueqiang Wang, and Xiaofeng Wang. Privacy loss in apple's implementation of differential privacy on macos 10.12. *arXiv preprint arXiv:1709.02753*, 2017.

The Royal Society. Privacy Enhancing Technologies: protecting privacy in practice. 2019.

UK HDR Alliance. Trusted research environments (TRE). `https://ukhealthdata.org/wp-content/uploads/2020/07/200723-Alliance-Board_Paper-E_TRE-Green-Paper.pdf`, 2020. Accessed: 2020-10-15.

Stephen G Walker. Bayesian inference with misspecified models. *Journal of Statistical Planning and Inference*, 143(10):1621–1633, 2013.

James Watson and Chris Holmes. Approximate models and robust decisions. *Statistical Science*, 31(4):465–489, 2016.

Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*, 2018.

Jun Zhang, Graham Cormode, Cecilia M Procopiuc, Divesh Srivastava, and Xiaokui Xiao. PrivBayes: Private data release via bayesian networks. *ACM Trans. Database Syst.*, 42(4):1–41, October 2017.