Scalable Nonparametric Sampling from Multimodal Posteriors with the Posterior Bootstrap

Edwin Fong 1 2 Simon Lyddon 1 Chris Holmes 1 2

Abstract
Increasingly complex datasets pose a number of challenges for Bayesian inference. Conventional posterior sampling based on Markov chain Monte Carlo can be too computationally intensive, is serial in nature and mixes poorly between posterior modes. Furthermore, all models are misspecified, which brings into question the validity of the conventional Bayesian update. We present a scalable Bayesian nonparametric learning routine that enables posterior sampling through the optimization of suitably randomized objective functions. A Dirichlet process prior on the unknown data distribution accounts for model misspecification, and admits an embarrassingly parallel posterior bootstrap algorithm that generates independent and exact samples from the nonparametric posterior distribution. Our method is particularly adept at sampling from multimodal posterior distributions via a random restart mechanism, and we demonstrate this on Gaussian mixture model and sparse logistic regression examples.

1. Introduction
As datasets grow in complexity and size, Bayesian inference becomes increasingly difficult. The posterior is often intractable, so we resort to simulation methods for inference via Markov chain Monte Carlo (MCMC), which is inherently serial and often too computationally expensive in datasets with a large number of data points (Bardenet et al., 2017). MCMC further struggles with multimodal posteriors which arise in many settings including mixture models (Jasra et al., 2005) or non-convex priors (Seeger et al., 2007), as the MCMC sampler can become trapped in local modes (Rudoy & Wolfe, 2006). Current methods to sample from multimodal posteriors with MCMC include parallel tempering (Neal, 1996) and adaptive MCMC (Pompe et al., 2018), but the associated computational cost is high. Posterior approximation with variational Bayes (VB) (Blei et al., 2017) is a faster alternative, but it is generally difficult to quantify the quality of the approximation, and is thus problematic if accurate uncertainty quantification is desired (Giordano et al., 2015).

A further methodological issue facing Bayesian inference is the fact that all models are false. The increasing scale of datasets exacerbates the effects of model misspecification (Walker, 2013), as the true sampling distribution is meaningfully different from the parametric family of distributions of the model. There is rarely formal acknowledgement of model misspecification which can lead to inconsistencies (Watson et al., 2016; Grünwald & Van Ommen, 2017).

Bayesian nonparametric learning (NPL) introduced by Lyddon et al. (2018) allows for the use of statistical models without assuming the model is true. NPL uses a nonparametric prior centred on a parametric model, and returns a nonparametric posterior over the parameter of interest. The method focuses on accounting for model misspecification and for posterior approximation such as from Variational Bayes (VB) by placing a mixture of Dirichlet processes (Antoniak, 1974) prior on the sampling distribution. In addition to the acknowledgement of model misspecification, the method admits an embarrassingly parallel Monte Carlo sampling scheme consisting of randomized maximizations. However, in most cases this method requires sampling the Bayesian posterior, which is computationally expensive for complex models.

1.1. Our Contribution
In this work, we propose a simplified variant of NPL that utilises a Dirichlet process (DP) prior on \( F_0 \) instead of a mixture of Dirichlet processes (MDP) prior. This allows us to perform inference directly and detaches the nonparametric prior from the prior of the model parameter of interest. Instead of centering on a Bayesian posterior, we center the DP on a sampling distribution which encapsulates our prior beliefs. This simpler choice of prior also has desirable theoretical properties and is highly scalable as we no longer
need to sample from the Bayesian posterior. Our method can handle a variety of statistical models through the choice of the loss functions, and can be applied to a wide range of machine learning settings as we will demonstrate in Section 3. Our method implies a natural noninformative prior, which may be relevant when the number of data points is substantially larger than the number of parameters.

The posterior bootstrap sampling scheme was introduced by Lyddon et al. (2018) under the NPL framework, and we inherit its computational strengths such as parallelism and exact inference under a Bayesian nonparametric model. Independent samples from the nonparametric posterior are obtained through the optimization of randomized objective functions, and we obtain the weighted likelihood bootstrap (Newton & Raftery, 1994) as a special case. Furthermore, sampling from multimodal posteriors now involves a non-convex optimization at each bootstrap sample that we solve through local search and random restart. We demonstrate that our method recovers posterior multimodality on a Gaussian Mixture Model (GMM) problem. We further show that our method is computationally much faster than conventional Bayesian inference with MCMC, and has superior predictive performance on real sparse classification problems. Finally, we utilize the computational speed of NPL to carry out a Bayesian sparsity-path-analysis for variable selection on a genetic dataset.

### 2. Bayesian Nonparametric Learning

Assume that we have observed \( y_{1:n} \overset{iid}{\sim} F_0 \), where \( y_{1:n} \) is a sequence of \( n \) i.i.d. observables and \( F_0 \) is the unknown sampling distribution. We may be interested in a parameter \( \theta \in \Theta \subseteq \mathbb{R}^p \), which indexes a family of probability densities \( F_\theta = \{ f_\theta(y); \theta \in \Theta \} \). Conventional Bayesian updating of the prior to the posterior via Bayes’ theorem formally assumes that \( F_0 \) belongs to the model \( F_\theta \), which is questionable in the presence of complex and large datasets. This assumption is not necessary for NPL. We derive the foundations of NPL by treating parameters as functionals of \( F_0 \), with model fitting as a special case.

#### 2.1. The Parameter of Interest

We define our parameter of interest as

\[
\theta_0(F_0) = \arg \min_\theta \int l(y, \theta) dF_0(y)
\]  

where \( l(y, \theta) \) is a loss function, and its form can be used to target statistics of interest. For example, setting \( l(y, \theta) = |y - \theta| \) returns the median and \( (y - \theta)^2 \) returns the mean.

The loss function of particular interest is \( l(y, \theta) = -\log f_\theta(y) \), where \( f_\theta \) is the density of some parametric model. The value of \( \theta_0 \) minimises the Kullback-Leibler divergence \( \text{KL}(f_0 || f_\theta) \), which is the parameter of interest in conventional Bayesian analysis (Walker, 2013; Bissiri et al., 2016). We have not assumed that \( F_\theta \) contains \( F_0 \), and \( \theta_0 \) in this case does not have any particular generative meaning as it is simply the parameter that satisfies (1).

#### 2.2. The Dirichlet Process Prior

As the sampling distribution is unknown, we place a DP prior on \( F_0 \)

\[
[F|\alpha, F_\pi] \sim \text{DP} (\alpha, F_\pi) 
\]  

where \( F_\pi \) is our prior centering measure, and \( \alpha \) is the strength of our belief.

**The base measure \( F_\pi \)** We encode our prior knowledge about the sampling distribution in the measure \( F_\pi \). If we believe a particular model \( f_\theta \) to be accurate, and have prior beliefs about \( \theta \) encoded in \( \pi(\theta) \), a sensible choice for the density of \( F_\pi \) is \( f_\pi(y) = \int f_\theta(y)d\pi(\theta) \). Alternatively, we could directly specify \( f_\pi \) as a density that accurately represents our beliefs without the burden of defining a joint distribution on \((y, \theta)\). In the presence of historical data \( y_{1:n} \), a suitable choice for \( F_\pi \) is the empirical distribution of the historical data, i.e. \( F_\pi(y) = \frac{1}{n} \sum_{i=1}^{n} \delta_{y_i}(y) \) where \( \delta \) is the Dirac measure. This is in a similar fashion to power priors (Ibrahim et al., 2000). Further intuition is provided in Section A.1 of the Supplementary Material.

It should be noted that we cannot directly include a prior on the parameter of interest \( \theta_0 \), only implicitly through \( (\alpha, F_\pi) \). Our prior is selected independently of the model of interest, and this is appropriate under a misspecified model setting since we do not believe there to be a true \( f_\theta \). As all parameters of interest are defined as a functional of \( F_0 \) as in (1), any informative prior on \( F_0 \) is thus informative of \( \theta_0 \).

**The concentration \( \alpha \)** The size of \( \alpha \) measures the concentration of the DP about \( F_\pi \), and a large value corresponds to a smaller variance in a functional of the DP. We see in (3) that the DP posterior base measure is a weighted sum of the prior \( F_\pi \) and the empirical distribution \( F_\pi = \frac{1}{n} \sum_{i=1}^{n} \delta_{y_i} \), with the weights proportional to \( \alpha \) and \( n \) respectively. We can thus interpret \( \alpha \) as the effective sample size from the prior \( F_\pi \). One method of selecting \( \alpha \) is through simulation of the prior distribution of \( \theta \) via (1) and tuning its variance. Alternatively, we can select \( \alpha \) through the a priori variance of the mean functional (see Section A.2 of the Supplementary Material). The special case of \( \alpha = 0 \) corresponds to the Bayesian bootstrap (Rubin, 1981), which in our case corresponds to a natural way to define an noninformative prior about \( F_0 \) (see Gelman et al. (2013) for a review on noninformative priors). For \( n \gg p \), it may be suitable to set \( \alpha = 0 \) as the prior should have little influence and the Bayesian bootstrap is more computationally efficient.
2.3. The NPL Posterior

From the conjugacy of the DP, the posterior of $F$ is

$$[F|y_{1:n}] \sim \text{DP}(\alpha + n, G_n),$$

$$G_n = \frac{\alpha}{\alpha + n} F + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{y_i}. \quad (3)$$

Our NPL posterior $\tilde{\pi}(\theta|y_{1:n})$ is thus

$$\tilde{\pi}(\theta|y_{1:n}) = \int \pi(\theta|F)d\pi(F|y_{1:n}) \quad (4)$$

where $\pi(\theta|F) = \delta_{\theta_0}(F)$; the delta arises as $\theta$ is a deterministic functional of $F$ as in (1). Properties of the NPL posterior follow from properties of the DP, e.g. draws of $F|y_{1:n}$ are almost surely discrete, so (1) simplifies to

$$\theta(F) = \arg \min_{\theta} \sum_{k=1}^{\infty} w_k l(\tilde{y}_k, \theta) \quad (5)$$

where $w_{1:\infty} \sim \text{GEM}(\alpha + n)$ and $\tilde{y}_{1:\infty}$ are iid $G_n$ from the stick-breaking construction (Sethuraman, 1994). Formally, the GEM distribution is defined

$$v_k \sim \text{Beta}(1, \alpha + n), \quad w_k = v_k \prod_{j=1}^{k-1} (1 - v_j). \quad (6)$$

We preserve the theoretical advantages from Lyddon et al. (2018) due to the symmetries in the limits of the DP and the MDP for $\alpha \to 0$ and $n \to \infty$, where $\alpha$ also denotes the concentration parameter of the MDP.

**Consistency** Under regularity conditions, the NPL posterior is consistent at $\theta_0$ as defined in (1), from the properties of the DP (see van der Vaart (1998); Ghosal (2010); Ghosal & van der Vaart (2017) for details). Interestingly, this is true regardless of the choice of $F_\pi$ and its support. This is not the case in conventional Bayesian inference through Bayes’ rule where the support of the prior must contain $\theta_0$ for posterior consistency. This is particularly reassuring in our misspecified model setting, as inferences about $\theta_0$ are robust to choices of $F_\pi$.

**Asymptotic dominance** The NPL posterior predictive $\tilde{\pi}(|y_{1:n})$ for $\alpha = 0$ asymptotically dominates the conventional Bayesian posterior predictive $\pi(|y_{1:n})$ up to $o(n^{-1})$ under regularity conditions, i.e.

$$E_{y_{1:n} \sim q}[\text{KL}(q(.)||\pi(|y_{1:n})) - \text{KL}(q(.)||\tilde{\pi}(|y_{1:n})]] = K(q(.)) + o(n^{-1}) \quad (7)$$

for all distributions $q$, where $K$ is a non-negative and possibly positive real-valued functional. This states that compared to the Bayesian posterior predictive, the NPL posterior predictive is closer in expected KL divergence to the true $F_0$ up to $o(n^{-1})$. The proof for the MDP case is given in Theorem 1 of Lyddon et al. (2018), and the above follows from the equivalence of the MDP and the DP for $\alpha = 0$.

2.4. Sampling from the NPL Posterior

In almost all cases, $\tilde{\pi}(\theta|y_{1:n})$ is not tractable, but lends itself to a parallelizable Monte Carlo sampling scheme. It may be more intuitive to think of sampling $F$ from the posterior DP, then calculating (1) to generate the sample from $\tilde{\pi}(\theta|y_{1:n})$, as shown in Algorithm 1.

**Algorithm 1** NPL Posterior Sampling

```python
for i = 1 to B do
    Draw $F(i) \sim \text{DP}(\alpha + n, G_n)$
    $\theta(i) = \arg \min_{\theta} \int l(\theta, y) dF(i)(y)$
end for
```

Here $B$ is the number of posterior bootstrap samples. One advantage of this sampling scheme is that it is embarrassingly parallel as each of the $B$ samples can be drawn independently. We can thus take advantage of increasingly available multi-core computing, unlike in conventional Bayesian inference as MCMC is inherently sequential.

2.4.1. The Posterior Bootstrap

Sampling from the DP exactly requires infinite computation time if $F_\pi$ is continuous, but approximate samples can be generated by truncation of the sum in (5). For example, we could truncate the stick-breaking and set the remaining weights to 0. Alternatively, we could approximate $w_{1:T} \sim \text{Dir}(\alpha/T, \ldots, \alpha/T)$ with the finite Dirichlet distribution for large $T$. For further details, see Muller & Secchi (1996); Ishwaran & Zarepour (2002). We opt for the latter suggestion as Dirichlet weights can be generated efficiently, which leads to a simpler variant of the posterior bootstrap algorithm as shown in Algorithm 2.

**Algorithm 2** Posterior Bootstrap Sampling

```python
Define $T$ as truncation limit
Observe samples are $y_{1:n}$
for i = 1 to B do
    Draw prior pseudo-samples $\tilde{y}_{1:T}$ iid $F_\pi$
    Draw $(w_{1:n}^{(i)}, \tilde{w}_{1:T}^{(i)}) \sim \text{Dir}(1, \ldots, 1, \alpha/T, \ldots, \alpha/T)$
    $\theta(i) = \arg \min_{\theta} \left\{ \sum_{j=1}^{n} w_j^{(i)} l(y_j, \theta) + \sum_{k=1}^{T} \tilde{w}_k^{(i)} l(\tilde{y}_k^{(i)}, \theta) \right\}$
end for
```

For $\alpha = 0$, we simply draw $w_{1:n}^{(i)} \sim \text{Dir}(1, \ldots, 1)$, which is no longer an approximation and is equivalent to the Bayesian bootstrap. For $\alpha > 0$, the sampling scheme is asymptotically exact for $T \to \infty$, but this is computationally infeasible. We could fix $T$ to a moderate value, or select it adaptively via adaptive NPL, where we use the stick-breaking construction until the remaining probability is less than $\epsilon$. 
2.5. Tackling Multimodal Posteriors with Initialization

Multimodal posteriors can arise in Bayesian inference if the likelihood function is non-log-concave like in GMMs (Jin et al., 2016; Stephens, 1999), or if the prior is non-log-concave which can arise when selecting sparse priors (Seeger et al., 2007; Park & Casella, 2008; Lee et al., 2010). Unlike the method by Lyddon et al. (2018) with the MDP, our NPL posterior with the DP is now decoupled from the Bayesian posterior. There is thus no reliance on an accurate representation of the Bayesian posterior with potential multimodality, which MCMC and VB can often struggle to capture. If our loss function in (1) is non-convex (e.g. \(-\log f_0(y)\) of a GMM), our NPL posterior may also be multimodal. This now presents an optimization issue: solving (1) requires non-convex optimization. In general, optimizing non-convex objectives is difficult (see Jain & Kar (2017)), but under smoothness assumption of the loss, we can apply convex optimization methods to find local minima.

2.5.1. Random Restart for Multiple Modes

Random restart (see G. E. Boender & H. G. Rinnooy Kan (1987)) can be utilized with convex optimization methods to generate a list of potential global minima then selecting the one with the lowest objective. This involves \(R\) random initializations of \(\theta_{\text{init}} \sim \pi_0\) for each local optimization, and it was shown by Hu et al. (1994) that the uniform measure for \(\pi_0\) has good properties for convergence. If the number of modes is finite, then the global minimum will be achieved asymptotically in the limit of the \(R \to \infty\). The probability of obtaining the correct global minimum for finite \(R\) is related to the size of its basin of attraction. Random restart NPL (RR-NPL) is shown in Algorithm 3.

**Algorithm 3 RR-NPL Posterior Sampling**

\[
\text{for } i = 1 \text{ to } B \text{ do} \\
\text{Draw } F^{(i)} \sim \text{DP}(\alpha + n, G_0) \\
\text{for } r = 1 \text{ to } R \text{ do} \\
\text{Draw } \theta^{(i)}_r \sim \pi_0 \\
\theta^{(i)} = \text{local arg min}_\theta \left( \int l(y, \theta)dF^{(i)}(y), \theta^{(i)}_r \right) \\
\text{end for} \\
\theta^{(i)} = \text{arg min}_\theta \int l(y, \theta^{(i)})dF^{(i)}(y) \\
\text{end for}
\]

This is particularly suited to NPL with non-convex loss functions for the following reasons. Firstly, random restart can utilize efficient convex optimization techniques such as quasi-Newton methods, and the restarts can be easily implemented in parallel which is coherent with our parallelizable sampling scheme. Secondly, we can compromise between accuracy and computational cost by selecting \(R\), as computational cost scales linearly with \(R\) (though we can parallelize). The repercussions of an insufficiently large \(R\) are not severe: our NPL posterior will incorrectly allocate more density to local modes/saddles but all modes will likely still be present for a sufficiently large \(B\). This is demonstrated in Section E.2.2 of the Supplementary Material. Finally, the uniform initialization can sample from nonidentifiable posteriors with symmetric modes as their basins of attraction are selected with equal probability.

Practically, uniform initialization may not be possible if the support of the parameter is infinite, e.g. the variance \(\sigma^2\). In this case, we can pick another \(\pi_0\) (e.g. Gamma for a positive parameter), or sample uniformly from a truncated support. For adaptively setting \(R\), we can utilize stopping rules as discussed in Section B of the Supplementary Material.

2.5.2. Fixed Initialization for Local Modes

We may be interested in targeting local modes of the posterior when we value interpretability of posterior quantities over exact posterior representation. For example in \(K\)-component mixture models, there will be \(K\) symmetrical modes (or sets of modes), and label-switching occurs if the sampler travels between these (Jasra et al., 2005) which impedes useful inference in terms of clustering.

We can target one NPL posterior mode through a fixed initialization scheme by taking advantage of the fact that local optimization methods like expectation-maximization (EM) or gradient ascent are hill-climbers. We initialize each maximization step with the same \(\theta^{\text{init}}\), causing the sampler to stay within the basin of attraction of the local posterior mode with high probability. We can utilize VB’s mode-selection to select \(\theta^{\text{init}}\), assuming the Bayesian and NPL posterior modes are close. Mean-field VB also tends to underestimate posterior variance (Blei et al., 2017), so we are able to obtain accurate local uncertainty quantification of the mode through this scheme. Fixed initialization NPL (FI-NPL) is shown in Algorithm 4.

**Algorithm 4 FI-NPL Posterior Sampling**

\[
\text{Select } \theta^{\text{init}} \text{ from mode of interest} \\
\text{for } i = 1 \text{ to } B \text{ do} \\
\text{Draw } F^{(i)} \sim \text{DP}(\alpha + n, G_0) \\
\theta^{(i)} = \text{local arg min}_\theta \left( \int l(y, \theta)dF^{(i)}(y), \theta^{\text{init}} \right) \\
\text{end for}
\]

2.6. Loss-NPL

As we cannot define priors on \(\theta_0\) directly, we can instead penalize undesirable properties in the loss
\[
l(y, \theta) = -\log f_0(y) + \gamma g(\theta). \tag{8}
\]

For example, \(g(\theta) = |\theta|\) obtains the Bayesian NPL-Lasso, or we can set \(g(\theta) = -\log \pi(\theta)\) if we have some prior preference. We recommend \(\gamma = \frac{1}{n}\) if we desire roughly the same prior regularization as in Bayesian inference, where \(n\) is the size of the training set. The reasoning is outlined in Section D of the Supplementary Material. We could also
tune $\gamma$ through desired predictive performance or properties of $\theta$. Note that we are no longer encoding prior beliefs, and are instead expressing an alternative parameter of interest that minimizes the expectation of (8).

2.7. Related Work

We build on the work of Lyddon et al. (2018) which specifies an MDP prior on $F_0$, and recovers conventional Bayesian inference in the limit of $\alpha \to \infty$. Although the foundations of nonparametric learning are unchanged, our NPL posterior is decoupled from the Bayesian model, offering flexibility in prior measure selection, computational scalability and full multimodal exploration.

NPL unsurprisingly overlaps with other nonparametric approaches to inference. We recover the Bayesian bootstrap (Rubin, 1981) if we set $\alpha = 0$, and further setting $l(y; \theta) = -\log f_0(y)$ gives the weighted likelihood bootstrap (Newton & Raftery, 1994), as discussed in Lyddon et al. (2019). Setting the loss to (8) and $\alpha = 0$ also returns the fixed prior weighted Bayesian bootstrap (Newton et al., 2018). However, these methods were posited as approximations to the true Bayesian posterior, and the Bayesian bootstrap/weighted likelihood bootstrap are unable to incorporate prior information. The NPL posterior on the other hand is exact and distinct to the conventional Bayesian posterior with theoretical advantages, and we are able to incorporate prior information either through $F_{\pi}$ or $l(y; \theta)$.

Treating parameters as functionals of the sampling distribution is akin to empirical likelihood methods (Owen, 1988), in which parameters are defined through estimating equations of the form $\int m(y, \theta) dF_0(y) = 0$. The definition of a parameter of interest through the loss $l(y, \theta)$ is also present in general Bayesian updating introduced by Bissiri et al. (2016), where a coherent posterior over a parameter of interest is obtained without the need to specify a joint generative model. Their target parameter is equivalent to (1), and their methodology is built on a notion of coherency.

3. Examples

We now demonstrate our method on some examples; the code is available online \footnote{https://github.com/edfong/npl}. We compare NPL to conventional Bayesian inference with the No-U-Turn Sampler (NUTS) by Homan & Gelman (2014), and Automatic Differentiation Variational Inference (ADVI) by Kucukelbir et al. (2017) in Stan (Carpenter et al., 2017). We select these as baselines as they are off-the-shelf algorithms that do not require tuning. Similarly, NPL only requires a weighted likelihood optimization procedure. All NPL examples are run on 4 Azure F72s.v2 (72 vCPUs) virtual machines, implemented in Python. The NUTS and ADVI examples cannot be implemented in an embarrassingly parallel manner, so they are run on a single Azure F72s_v2. We avoid running multiple MCMC chains in parallel as the models are multimodal which may impede mixing, and combining unmixed chains is unprincipled. For tabulated results, each run was repeated 30 times with different seeds, and we report the mean with 1 standard error. We emphasize again that our NPL posterior is distinct to the conventional Bayesian posterior, so we are comparing the two inference schemes and their associated sampling methods. We include additional empirical comparisons to importance sampling and NPL with an MDP prior in Sections E.2.3, E.2.4 of the Supplementary Material.

3.1. Gaussian Mixture Model

We demonstrate the ability of RR-NPL to accurately sample from a multimodal posterior in a $K$-component, $d$-dimensional diagonal GMM toy problem, which NUTS and ADVI fail to do. It should be noted that in addition to the $K$! symmetrical modes present from label-switching, further multimodality is present due to the non-log-concavity of the likelihood. We further show how FI-NPL can be used in a clustering example with real data to provide accurate local uncertainty quantification which ADVI is unable to do. Our conventional Bayesian model for $i \in \{1, \ldots, n\}$, $j \in \{1, \ldots, d\}$ and $k \in \{1, \ldots, K\}$ is

\[
\gamma_i | \pi, \mu, \sigma \sim \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \text{diag}(\sigma_k^2)) ,
\]

\[
\pi | a_0 \sim \text{Dir}(a_0, \ldots, a_0), \quad \mu_{kj} \sim \mathcal{N}(0, 1), \quad \sigma_{kj} \sim \text{logNormal}(0, 1).
\]

The posterior is multimodal, and we use ADVI and NUTS for inference. For NPL, we are interested in model fitting, so our loss function is simply the negative log-likelihood

\[
l(y, \pi, \mu, \sigma) = -\log \sum_{k=1}^{K} \pi_k \mathcal{N}(y; \mu_k, \text{diag}(\sigma_k^2)) .
\]

In the case of small $n$, we may want to include a regularization term in the loss to avoid singularities of the likelihood. We select the DP prior separately for each example.

3.1.1. Toy Example: Implementation and Results

We analyze toy data from a GMM with $K = 3$, $d = 1$ and the following parameters:

\[
\pi_0 = \{0.1, 0.3, 0.6\}, \quad \mu_0 = \{0, 2, 4\}, \quad \sigma_0 = \{1, 1, 1\}.
\]

We generate $n_{\text{train}} = 1000$ for model fitting and another $n_{\text{test}} = 250$ held-out for model evaluation with different seeds for each of the 30 runs. For the Bayesian model we set $a_0 = 1$, and for NPL we set $\alpha = 0$ as $n \gg p$. We optimize
We now demonstrate FI-NPL on clustering handwritten digits from MNIST (LeCun & Cortes, 2010), which consists of 28×28 pixel images. In this example $n_{\text{train}} = 10000$, $n_{\text{test}} = 2500$ and $d = 784$. We normalize all pixel values such that they lie in the interval $[0, 1]$, and set $K = 10$. We believe a priori that many pixels are close to 0, so for ease we elicit a tight normal centering measure for the DP

$$f_\pi(y) = \prod_{j=1}^{d} N(y_j; 0, 0.1^2).$$  \hspace{1cm} (11)
we place automatic relevance determination (ARD) priors (MacKay, 1994) on the coefficients with Gamma hyperpriors (Gelman et al., 2008). The conventional Bayesian model for \( i \in \{1, \ldots, n\} \) and \( j \in \{1, \ldots, d\} \) is

\[
y_i|\mathbf{x}_i, \beta, \beta_0 \sim \text{Bernoulli}(\eta_i),
\eta_i = \sigma(\beta^T \mathbf{x}_i + \beta_0),
\beta_j|\lambda_j \sim \mathcal{N}\left(0, \frac{1}{\lambda_j}\right),
\lambda_j|a, b \sim \text{Gamma}(a, b).
\]

Marginally, the prior is the non-standardized t-distribution with (degrees of freedom, location, squared scale)

\[
\beta_j \sim \text{Student-t}\left(2a, 0, \frac{b}{a}\right).
\]

This posterior is intractable and potentially multimodal due to the non-log-concavity of the prior, and we carry out conventional Bayesian inference via NUTS and ADVI. When applying loss-NPL to regression, we assume \( y, x \sim F_0 \), and place a DP prior on the joint distribution \( F_0(y, x) \). We target the parameter which satisfies (1) with loss

\[
l(y, x, \beta, \beta_0) = -(y \log \eta + (1 - y) \log(1 - \eta))
+ \gamma \left(\frac{2a + 1}{2}\right) \sum_{j=1}^{d} \log \left(1 + \frac{\beta_j^2}{2b}\right)
\]

which is the negative sum of the log-likelihood and log-prior, with additional scaling parameter \( \gamma \). Again our NPL posterior may be multimodal due to the non-convexity of the loss, and so we utilize RR-NPL. It should be noted that our target parameter is now different to conventional Bayesian inference, but our method achieves the common goal of posterior sampling from multimodal posteriors with the Posterior Bootstrap (Kohavi, 1996), ‘Polish companies bankruptcy 3rd year’, (Zikeba et al., 2016), and ‘Arcene’ (Guyon et al., 2005) with details in Table 3. We handle categorical covariates with dummy variables, and normalize all covariates to have mean 0 and standard deviation 1. Missing real values were imputed with the mean, and data with missing categorical values were dropped. We carry out a random stratified train-test split for each of the 30 runs, with 80-20 split for ‘Adult’, ‘Polish’ and 50-50 split for ‘Arcene’ due to the smaller dataset. For both NPL and conventional Bayesian inference, the hyperparameters were set to \( a = b = 1 \), which was selected by tuning the sparsity of the Bayesian posterior means to a desired value. For NPL, we set \( \alpha = 0 \) for ‘Adult’ and ‘Polish’ as \( n \) is sufficiently large, and \( \alpha = 1 \) for ‘Arcene’ with \( T = 100 \) as \( n \) is only 100. We set \( \gamma = \frac{1}{n} \) for each dataset as explained in Section 2.6 for a fair comparison to the conventional Bayesian model. We initialize each optimization with \( \beta^0_i \sim \mathcal{N}(0, 1) \), and select the number of restarts to \( R = 1 \) for expediency. Optimization was carried out using the L-BFGS-B algorithm (Zhu et al., 1997) implemented in scipy.optimize (Jones et al., 2001–).

We can see in Table 4 that loss-NPL is predictively similar or better than NUTS and ADVI, and from Table 5 we see that the posterior mean is sparser for loss-NPL. Finally, we see from Table 6 that the loss-NPL run-times for 2000 posterior samples are much faster than for NUTS, and comparable to VB. Further measures of predictive performance are provided in Section E.3.4 of the Supplementary Material.

### 3.2.1. Implementation and Results

We analyze 3 binary classification datasets from the UCI ML repository (Dheeru & Karra Taniskidou, 2017): ‘Adult’ (Kohavi, 1996), ‘Polish companies bankruptcy 3rd year’, ‘Arcene’ (Guyon et al., 2005) with details in Table 3. We handle categorical covariates with dummy variables, and normalize all covariates to have mean 0 and standard deviation 1. Missing real values were imputed with the mean, and data with missing categorical values were dropped. We carry out a random stratified train-test split for each of the 30 runs, with 80-20 split for ‘Adult’, ‘Polish’ and 50-50 split for ‘Arcene’ due to the smaller dataset. For both NPL and conventional Bayesian inference, the hyperparameters were set to \( a = b = 1 \), which was selected by tuning the sparsity of the Bayesian posterior means to a desired value. For NPL, we set \( \alpha = 0 \) for ‘Adult’ and ‘Polish’ as \( n \) is sufficiently large, and \( \alpha = 1 \) for ‘Arcene’ with \( T = 100 \) as \( n \) is only 100. We set \( \gamma = \frac{1}{n} \) for each dataset as explained in Section 2.6 for a fair comparison to the conventional Bayesian model. We initialize each optimization with \( \beta^0_i \sim \mathcal{N}(0, 1) \), and select the number of restarts to \( R = 1 \) for expediency. Optimization was carried out using the L-BFGS-B algorithm (Zhu et al., 1997) implemented in scipy.optimize (Jones et al., 2001–).

### 3.3. Bayesian Sparsity-path-analysis

We now utilize loss-NPL to carry out Bayesian sparsity-path-analysis for logistic regression, which allows us to visualize how the responsibility of each covariate changes with the sparsity penalty as discussed by Lee et al. (2012).
We use the same ARD prior as Section 3.2 with the same initialization scheme, set $\gamma = \frac{1}{7}$, and elicit a noninformative DP prior with $\alpha = 0$. We found empirically that the results for larger values of $R$ are similar and so the approximation with $R = 1$ is sufficient. We fix $\alpha$ and vary the value of $b$ to favour solutions of different sparsity. This varies the squared scale $c = b/a$ of the Student-t prior with fixed degrees of freedom, where a smaller $c$ corresponds to a heavier sparsity penalty and thus more components are set to 0.

### 3.3.1. IMPLEMENTATION AND RESULTS

We analyze the genotype/pseudo-phenotype dataset with $n = 500$ as described by Lee et al. (2012), containing patient covariates $x_i$ which exhibit strong block-like correlations as shown in Figure 3. We normalize the covariates to have mean 0 and standard deviation 1. The pseudo-phenotype data is generated by $y_i \sim Bernoulli(\sigma(\beta^T x_i))$, where $\beta$ has 5 randomly selected non-zero components out of $d = 50$, with the rest set to 0. Each non-zero component is sampled from $N(0, 0.2)$, and the exact values of $\beta$ are provided in Section E.4.1 of the Supplementary Material. We set $\alpha = 1$ and vary $b_t = 0.98^{t-1}$ for $t = \{1, \ldots, 450\}$, and generate 4000 posterior samples for each setting.

The posterior medians of the non-zero components of $\beta$ with 80% central credible interval are shown in Figure 4 for a range of log $c$ values. Both the posterior median and central credible intervals are estimated through the appropriate order statistics of the posterior samples (Gelman et al., 2013). We can see that $\beta_{10}$, $\beta_{14}$ and $\beta_{24}$ have early predictive power as their credible intervals remain large despite a significant sparsity penalty ($\text{small log}(c)$), whilst the other two coefficients $\beta_{31}, \beta_{37}$ are masked. A plot of the absolute medians for all components is included in Section E.4.2 of the Supplementary Material. For $\beta_{10}$ and $\beta_{14}$, the median is close to 0 but the credible interval is large which is due to the multimodality of the marginal posterior. This multimodality is also responsible for the jitter in the median around $\text{log}(c) = -6.5$ for $\beta_{14}$ in Figure 4; the true median likely lies between the two separated modes but the finite posterior sample size causes the sample median to jump between the two. A posterior marginal KDE plot of $\beta_{14}$ changing with log $c$ is shown in Figure 5, allowing us to visualize how the importance of the covariate changes with the sparsity penalty. We observe the bimodality in the marginal posterior for $\text{log}(c) < -4$ as expected from the above discussion.

Loss-NPL required 5 minutes 24 seconds to generate all 450 x 4000 posterior samples. The computational speed of NPL enables fast Bayesian analysis of large datasets with different hyperparameter settings, allowing for Bayesian variable selection analysis.

Figure 4. Lasso-type plot for posterior medians of non-zero $\beta$ with 80% credible intervals against $\text{log}(c)$ from genetic dataset

Figure 5. Posterior marginal KDE of $\beta_{14}$ against $\text{log}(c)$ from genetic dataset

### 4. Discussion

We have introduced a variant of Bayesian nonparametric learning (NPL) with a Dirichlet process (DP) prior on the sampling distribution $F_0$, which leads to highly scalable exact inference under model misspecification, detached from the conventional Bayesian posterior. This method admits a sampling scheme for multimodal posteriors that allows for full mode exploration, which involves a non-convex optimization that we solve through random restart. We demonstrated that NPL can perform predictively better than conventional Bayesian inference, while providing exact uncertainty quantification.

For future work, the small sample performance of NPL could be further explored and compared to conventional Bayesian inference; we currently recommend NPL for moderate to large values of $n$. The scaling of the number of repeats $R$ with increasing dimension for full mode exploration would also be a future avenue of research.
Acknowledgements

We would like to thank our anonymous reviewers for their careful analyses of our paper and their valuable input. We also thank Luke Kelly for useful discussions, Ho Chung Leon Law for helpful feedback and Anthony Lee for providing the genotype dataset. EF is funded by The Alan Turing Institute Doctoral Studentship, under the EPSRC grant EP/N510129/1. SL is funded by the EPSRC OxWaSP CDT, through EP/L016710/1. CH is supported by The Alan Turing Institute, the HDR-UK, the Li Ka Shing Foundation, and the MRC.

References


Scalable Nonparametric Sampling from Multimodal Posteriors with the Posterior Bootstrap


MacKay, D. J. C. Bayesian nonlinear modeling for the prediction competition. ASHRAE Trans, 100, 1994.


