

# Nonparametric Bayesian Density Modeling with Gaussian Processes

Ryan Prescott Adams  
Cavendish Laboratory  
University of Cambridge  
Cambridge CB3 0HE, UK  
rpa23@cam.ac.uk

Iain Murray  
Department of Computer Science  
University of Toronto  
Toronto, Ontario M5S 3G4, CA  
murray@cs.toronto.edu

David J.C. MacKay  
Cavendish Laboratory  
University of Cambridge  
Cambridge CB3 0HE, UK  
mackay@mrao.cam.ac.uk

## Abstract

We present the Gaussian Process Density Sampler (GPDS), an exchangeable generative model for use in nonparametric Bayesian density estimation. Samples drawn from the GPDS are consistent with exact, independent samples from a fixed density function that is a transformation of a function drawn from a Gaussian process prior. Our formulation allows us to infer an unknown density from data using Markov chain Monte Carlo, which gives samples from the posterior distribution over density functions and from the predictive distribution on data space. We describe two such MCMC methods. Both methods also allow inference of the hyperparameters of the Gaussian process.

## 1 Introduction

We propose a method for incorporating a Gaussian process into a prior on probability density functions. While such constructions have been proposed before (Leonard, 1978; Thorburn, 1986; Lenk, 1988, 1991; Csató, 2002; Tokdar and Ghosh, 2007), ours is the first that allows a procedure for drawing exact and exchangeable data samples from a fixed density drawn from the prior. We call this prior and the associated procedure the *Gaussian process density sampler* (GPDS). Given data, this generative prior allows us to perform inference of the unnormalised density. We present two Markov chain Monte Carlo algorithms for performing this inference, one based on exchange sampling (Murray et al., 2006) and the other based on inferring the latent generative history. In both cases we are also able to perform inference of the Gaussian process hyperparameters and generate samples from the predictive distribution on data space.

Bayesian nonparametric inference is appealing because it promises models of arbitrary complexity while also providing a robust way to control that complexity. The tool of choice for estimation of an unknown distribution is the Dirichlet process (Ferguson, 1973) and related constructions (e.g. Pitman and Yor (1997) and Ishwaran and James (2001)). Samples from the Dirichlet process, however, are discrete distributions with probability one. For many inference problems, we wish to model continuous random variables and in such problems our prior beliefs are often best captured by statements about the associated probability density function.

To fill the gap between nonparametric priors on discrete distributions and nonparametric priors on continuous densities, the Dirichlet process is frequently used to add a countably-infinite number of parameters into a continuous model. The most popular example is the infinite mixture of parametric distributions (Escobar and West, 1995), but other approaches include kernel convolution

(Lo, 1984) and the Dirichlet diffusion tree (Neal, 2001, 2003). Managing complexity in these models requires both an appropriate prior for the Dirichlet concentration parameter and an appropriate prior on the smoothing method.

Prior beliefs about a distribution are often about the probability density function — its support and smoothness properties, for example. There is a rich literature on incorporating prior beliefs about functions into nonparametric Bayesian regression models, using splines, neural networks and stochastic processes (e.g. DiMatteo et al. (2001), MacKay (1992), and O’Hagan (1978)). However, priors on general functions have largely resisted application to density estimation, due to the requirements that probability density functions be nonnegative and integrate to one. This work introduces the first fully-Bayesian nonparametric kernel method for density estimation that does not require a finite-dimensional approximation to perform inference.

## 2 The Gaussian process density sampler prior

We consider densities on a space  $\mathcal{X}$  that we call the *data space*. We assume without loss of generality that  $\mathcal{X}$  is the  $d$ -dimensional real space  $\mathbb{R}^d$ . We first construct a Gaussian process prior with the data space  $\mathcal{X}$  as its input and the one-dimensional real space  $\mathbb{R}$  as its output. The Gaussian process provides a distribution over functions from  $\mathcal{X}$  to  $\mathbb{R}$ . We define a mean function  $m(\cdot) : \mathcal{X} \rightarrow \mathbb{R}$  and a positive definite covariance function  $K(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ . These functions are parameterised by hyperparameters  $\theta$ . Given these two functions and the hyperparameters, for any finite subset of  $\mathcal{X}$  with cardinality  $N$  there is a multivariate normal distribution on  $\mathbb{R}^N$  (Rasmussen and Williams, 2006). We take the mean function to be zero.

We construct a map from a function  $g(x) : \mathcal{X} \rightarrow \mathbb{R}$ ,  $x \in \mathcal{X}$ , to a probability density function  $f(x)$  via

$$f(x) = \frac{1}{\mathcal{Z}_\pi[\mathbf{g}]} \Phi(g(x)) \pi(x) \tag{1}$$

where  $\pi(x)$  is a *base density* that corresponds to an arbitrary base probability measure on  $\mathcal{X}$ . The function  $\Phi(\cdot) : \mathbb{R} \rightarrow (0, 1)$  is a positive function with upper bound 1. We use the bold notation  $\mathbf{g}$  to refer to the function  $g(x)$  compactly as a vector of (infinite) length, versus its value at a particular  $x$ . The normalisation constant is a functional of  $g(x)$ :

$$\mathcal{Z}_\pi[\mathbf{g}] = \int dx' \Phi(g(x')) \pi(x'). \tag{2}$$

Through the map defined by Equation 1, a Gaussian process becomes a prior distribution over normalised probability density functions on  $\mathcal{X}$ . Figure 2 shows several realisations of densities from this prior, along with sample data.

Although we only require that the function  $\Phi(\cdot)$  be positive and bounded, it is convenient for inference if it is a bijective map between  $\mathbb{R}$  and  $(0, 1)$ . If  $\Phi(\cdot)$  is bijective then each realisation  $g(x)$  from the Gaussian process corresponds to a unique function that maps  $\mathcal{X}$  to  $(0, 1)$ . Sigmoids, such as the cumulative normal distribution function and the logistic function, are bijective functions with this domain and range. We take  $\Phi(\cdot)$  to be the logistic function, i.e.  $\Phi(z) = (1 + \exp\{-z\})^{-1}$ .

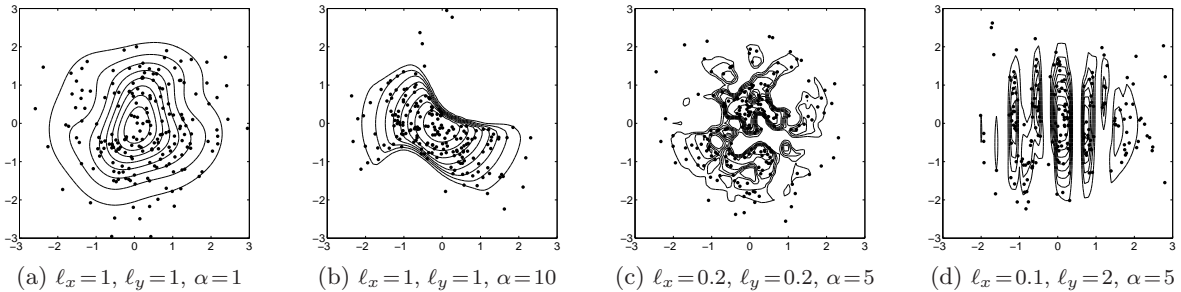


Figure 1: Four samples from the GPDS prior are shown, with 200 data samples. The contour lines show the approximate unnormalized densities. In each case the base density is the zero-mean spherical Gaussian with unit variance. The covariance function was the squared exponential:  $K(x, x') = \alpha^2 \exp(-\frac{1}{2} \sum_i \ell_i^{-2} (x_i - x'_i)^2)$ , with parameters varied as labeled in each subplot.  $\Phi(\cdot)$  is the logistic function in these plots.

### 3 Generating data from the prior

We can use rejection sampling to generate samples from a common density drawn from the the prior described in Section 2. A rejection sampler requires a proposal density that upper bounds the unnormalised density of interest. In this case, the proposal density is  $\pi(x)$  and the unnormalised density of interest is  $\Phi(g(x))\pi(x)$ . We assume that it is possible to draw samples directly from  $\pi(x)$ .

If  $g(x)$  were known, rejection sampling would proceed as follows: First generate proposals  $\{\tilde{x}_q\}$  from the base density  $\pi(x)$ . The proposal  $\tilde{x}_q$  would be accepted if a variate  $r_q$  drawn uniformly from  $(0, 1)$  was less than  $\Phi(g(\tilde{x}_q))$ . These samples would be exact in the sense that they were not biased by the starting state of a finite Markov chain. However, in the GPDS,  $g(x)$  is not known: it is a random function drawn from a Gaussian process prior. We can nevertheless use rejection sampling by “discovering”  $g(x)$  as we proceed at just the places we need to know it, by sampling from the prior distribution of the latent function. As it is necessary only to know  $g(x)$  at the  $\{x_q\}$  to accept or reject these proposals, the samples are still exact. This type of retrospective sampling trick has been used in a variety of MCMC algorithms for infinite-dimensional models Beskos et al. (2006); Papaspiliopoulos and Roberts (2008). Figure 2 shows the the generative procedure graphically.

In practice, we generate the samples sequentially, as in Algorithm 1, so that we may be assured of having as many accepted samples as we require. In each loop, a proposal is drawn from the base density  $\pi(x)$  and the function  $g(x)$  is sampled from the Gaussian process at this proposed coordinate, conditional on all the function values already sampled. We will call these data the *conditioning set* for the function  $g(x)$  and will denote the conditioning inputs as  $\mathbf{X}$  and the conditioning function values as  $\mathbf{G}$ . After the function is sampled, a variate is drawn uniformly from  $(0, 1)$  and compared to the  $\Phi$ -squashed function at the proposal location.

The sequential procedure is exchangeable; the probability of the data is the same under re-ordering. First, the base density draws are i.i.d.. Second, conditioned on the proposals from the base density, the Gaussian process is a simple multivariate Gaussian distribution, which is exchangeable in its components. Finally, conditioned on the draw from the Gaussian process, the acceptance/rejection steps are independent Bernoulli samples, and the overall procedure is exchangeable. This property ensures that the sequential procedure generates data from the same distribution as the simultaneous procedure described above. More broadly, exchangeable priors are

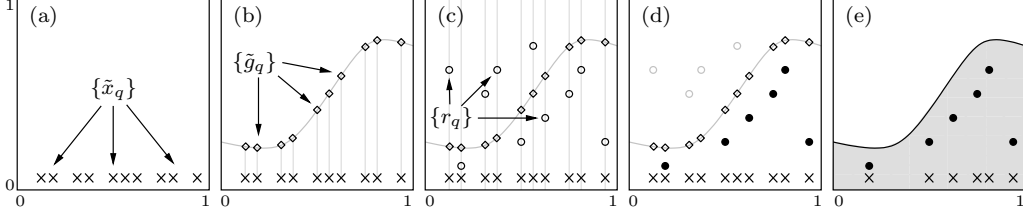


Figure 2: These figures show the procedure for generating samples from a single density drawn from the GPDS prior. (a): Draw  $Q$  samples  $\{\tilde{x}_q\}^Q$  from the base density  $\pi(x)$ , which in this case is uniform on  $[0, 1]$ . (b): Sample the function  $g(x)$  at the randomly chosen locations, generating the set  $\{\tilde{g}_q = g(\tilde{x}_q)\}^Q$ . The squashed function  $\Phi(g(x))$  is shown. (c): Draw a set of variates  $\{r_q\}^Q$  uniformly beneath the bound in the vertical coordinate. (d): Accept only the points whose uniform draws are beneath the squashed function value, i.e.  $r_q < \Phi(\tilde{g}_q)$ . (e): The accepted points  $(\tilde{x}_q, r_q)$  are uniformly drawn from the shaded area beneath the curve and the marginal distribution of the accepted  $\tilde{x}_q$  is proportional to  $\Phi(g(x))\pi(x)$ .

useful in Bayesian modeling because we may consider the data conditionally independent, given the latent density.

---

**Algorithm 1** Generate  $P$  exact samples from the prior

---

**Purpose:** Draw  $P$  exact samples from a common density on  $\mathcal{X}$  drawn from the prior in Equation 1

**Inputs:** GP hyperparameters  $\theta$ , number of samples to generate  $P$

- 1: Initialize empty conditioning sets for the Gaussian process:  $\mathbf{X} = \emptyset$  and  $\mathbf{G} = \emptyset$
  - 2: **repeat**
  - 3:   Draw a proposal from the base density:  $\tilde{x} \sim \pi(x)$
  - 4:   Sample the function from the Gaussian process at  $\tilde{x}$ :  $\tilde{g} \sim \mathcal{GP}(g | \mathbf{X}, \mathbf{G}, \tilde{x}, \theta)$
  - 5:   Draw a uniform variate on  $(0, 1)$ :  $r \sim \mathcal{U}(0, 1)$
  - 6:   **if**  $r < \Phi(\tilde{g})$  (Acceptance rule) **then**
  - 7:     Accept  $\tilde{x}$
  - 8:   **else**
  - 9:     Reject  $\tilde{x}$
  - 10:   **end if**
  - 11:   Add  $\tilde{x}$  and  $\tilde{g}$  to the conditioning sets:  $\mathbf{X} = \mathbf{X} \cup \tilde{x}$  and  $\mathbf{G} = \mathbf{G} \cup \tilde{g}$
  - 12: **until**  $P$  samples have been accepted
- 

## 4 Inference

We now consider the problem of inference with the GPDS. We observe  $N$  data  $\mathcal{D} = \{x_n\}_{n=1}^N$  that we model as having been drawn independently from an unknown density  $f(x)$ . We place the GPDS prior of Section 2 on  $f(x)$ . Even with Markov chain Monte Carlo, inference in this model is difficult due to the normalisation constant  $\mathcal{Z}_\pi[\mathbf{g}]$ . If we write Bayes' theorem for the posterior on  $\mathbf{g}$

$$p(\mathbf{g} | \mathcal{D}, \theta) = \frac{p(\mathbf{g} | \theta) (\mathcal{Z}_\pi[\mathbf{g}])^{-N} \prod_{n=1}^N \Phi(g(x_n)) \pi(x_n)}{\int d\mathbf{g} p(\mathbf{g} | \theta) (\mathcal{Z}_\pi[\mathbf{g}])^{-N} \prod_{n=1}^N \Phi(g(x_n)) \pi(x_n)} \quad (3)$$

we see that the posterior cannot be evaluated without two difficult integrals. It is common for the marginal likelihood in the denominator of the posterior to be intractable; MCMC methods such as

Metropolis–Hastings are well-suited for this situation. Posteriors such as Equation 3 with difficult sums in both the numerator and denominator are called *doubly-intractable*. Doubly-intractable posterior distributions appear most frequently when performing inference in undirected graphical models, where the partition function can be difficult to evaluate (Murray et al., 2006).

To see the difficulty concretely, consider a naïve Metropolis–Hastings Markov chain on  $\mathbf{g}$ , with proposal density  $q(\hat{\mathbf{g}} \leftarrow \mathbf{g})$ :

$$a_{\text{naïve}} = \frac{q(\mathbf{g} \leftarrow \hat{\mathbf{g}})p(\hat{\mathbf{g}} | \theta)}{q(\hat{\mathbf{g}} \leftarrow \mathbf{g})p(\mathbf{g} | \theta)} \left( \frac{\mathcal{Z}_\pi[\hat{\mathbf{g}}]}{\mathcal{Z}_\pi[\mathbf{g}]} \right)^N \prod_{n=1}^N \frac{\Phi(\hat{\mathbf{g}}(x_n))\pi(x_n)}{\Phi(\mathbf{g}(x_n))\pi(x_n)}. \quad (4)$$

We cannot compute this acceptance ratio without evaluating an intractable ratio of normalising constants. We present two Markov chain Monte Carlo algorithms that sidestep this difficulty. The equilibrium distribution in both cases is the posterior in Equation 3, and both algorithms take advantage of the exact data generation procedure described in Section 3.

#### 4.1 Exchange sampling

Exchange sampling (Murray et al., 2006; Murray, 2007) is a variant of the Metropolis–Hastings method that enables sampling from doubly-intractable posterior distributions, subject to the requirement that exact samples can be generated from the model. The procedure is an extension and simplification of the auxiliary variable method of Møller et al. (2006). Exchange sampling introduces additional state into the Markov chain that is chosen so that the intractable constants cancel out of the Metropolis–Hastings acceptance ratio. Murray et al. (2006) use exchange sampling to infer the coupling parameters of Ising and Potts models where exact data could be generated via coupling from the past (Propp and Wilson, 1996). In the GPDS we generate exact samples via the rejection method of Section 3.

Initially, we apply exchange sampling to the posterior on  $\mathbf{g}$  using the Gaussian process prior as the proposal distribution, i.e.  $q(\hat{\mathbf{g}} \leftarrow \mathbf{g}) = p(\hat{\mathbf{g}} | \theta)$ . The joint distribution over the data  $\mathcal{D}$ , the current Markov state  $\mathbf{g}$  and the proposal  $\hat{\mathbf{g}}$  is augmented with  $N$  “fantasy data”  $\mathcal{W} = \{w_n\}_{n=1}^N$ . These fantasy data live on the same space  $\mathcal{X}$  as the true data, but are drawn from the distribution implied by the proposal  $\hat{\mathbf{g}}$ . The augmented joint distribution is

$$p(\mathbf{g}, \{x_n\}_{n=1}^N, \hat{\mathbf{g}}, \{w_n\}_{n=1}^N | \theta) = p(\mathbf{g} | \theta) p(\{x_n\}_{n=1}^N | \mathbf{g}) p(\hat{\mathbf{g}} | \theta) p(\{w_n\}_{n=1}^N | \hat{\mathbf{g}}). \quad (5)$$

Given the current state  $\mathbf{g}$ , we jointly propose  $\hat{\mathbf{g}}$  and  $\mathcal{W}$  by using Algorithm 1 with  $P = N$ . This algorithm simultaneously draws  $\hat{\mathbf{g}}$  from the prior and generates the  $N$  fantasy data  $\mathcal{W}$ . We then propose swapping  $\mathbf{g}$  for  $\hat{\mathbf{g}}$ . The acceptance ratio of the swap proposal is the ratio of the joint density in Equation 5 under each setting:

$$\begin{aligned} a_{\text{exch}} &= \frac{p(\hat{\mathbf{g}} | \theta) p(\{x_n\}_{n=1}^N | \hat{\mathbf{g}}) p(\mathbf{g} | \theta) p(\{w_n\}_{n=1}^N | \mathbf{g})}{p(\mathbf{g} | \theta) p(\{x_n\}_{n=1}^N | \mathbf{g}) p(\hat{\mathbf{g}} | \theta) p(\{w_n\}_{n=1}^N | \hat{\mathbf{g}})} \\ &= \frac{(\mathcal{Z}_\pi[\mathbf{g}])^{-N} (\mathcal{Z}_\pi[\hat{\mathbf{g}}])^{-N} \prod_{n=1}^N \Phi(\hat{\mathbf{g}}(x_n)) \pi(x_n) \prod_{n=1}^N \Phi(\mathbf{g}(w_n)) \pi(w_n)}{(\mathcal{Z}_\pi[\mathbf{g}])^{-N} (\mathcal{Z}_\pi[\hat{\mathbf{g}}])^{-N} \prod_{n=1}^N \Phi(\mathbf{g}(x_n)) \pi(x_n) \prod_{n=1}^N \Phi(\hat{\mathbf{g}}(w_n)) \pi(w_n)} \\ &= \prod_{n=1}^N \frac{\Phi(\hat{\mathbf{g}}(x_n)) \Phi(\mathbf{g}(w_n))}{\Phi(\mathbf{g}(x_n)) \Phi(\hat{\mathbf{g}}(w_n))}. \end{aligned} \quad (6)$$

---

**Algorithm 2** Generate  $T$  samples from the posterior on  $g(x)$  using exchange sampling.

---

**Purpose:** Use exchange sampling to sample from the posterior on  $g(x)$

**Inputs:** GP hyperparameters  $\theta$ , number of samples  $T$ , data  $\mathcal{D} = \{x_n\}_{n=1}^N$

- 1: Sample  $g(x)$  from the GP at the data:  $\{g(x_n)\}_{n=1}^N \sim \mathcal{GP}(g | \mathcal{D}, \theta)$ .
  - 2: Initialise the conditioning sets:  $\mathbf{X} = \mathcal{D}$  and  $\mathbf{G} = \{g(x_n)\}_{n=1}^N$ .
  - 3: **for**  $t = 1$  to  $T$  **do**
  - 4:   Sample  $\hat{g}(x)$  from the GP at the data:  $\{\hat{g}(x_n)\}_{n=1}^N \sim \mathcal{GP}(g | \mathcal{D}, \theta)$ .
  - 5:   Initialise the proposal conditioning sets:  $\hat{\mathbf{X}} = \mathcal{D}$  and  $\hat{\mathbf{G}} = \{\hat{g}(x_n)\}_{n=1}^N$ .
  - 6:   Initialise empty fantasy set  $\mathcal{W} = \emptyset$ .
  - 7:   **repeat**
  - 8:     Draw a proposal from the base density:  $\tilde{x} \sim \pi(x)$ .
  - 9:     Sample  $\hat{g}(\tilde{x})$  from the Gaussian process:  $\tilde{g} \sim \mathcal{GP}(g | \hat{\mathbf{X}}, \hat{\mathbf{G}}, \tilde{x}, \theta)$ .
  - 10:     **if**  $\mathcal{U}(0, 1) < \Phi(\tilde{g})$  (Rejection sampling acceptance rule) **then**
  - 11:       Accept  $\tilde{x}$  and store in the fantasy set  $\mathcal{W}$ .
  - 12:     **else**
  - 13:       Reject  $\tilde{x}$ .
  - 14:     **end if**
  - 15:     Add  $\tilde{x}$  and  $\tilde{g}$  to the proposal conditioning sets:  $\hat{\mathbf{X}} = \hat{\mathbf{X}} \cup \tilde{x}$  and  $\hat{\mathbf{G}} = \hat{\mathbf{G}} \cup \tilde{g}$ .
  - 16:     **until**  $N$  fantasy samples have been accepted
  - 17:     Sample  $g(x)$  from the GP at the fantasy data:  $\{g(w_n)\}_{n=1}^N \sim \mathcal{GP}(g | \mathbf{X}, \mathbf{G}, \mathcal{W}, \theta)$ .
  - 18:     Calculate the acceptance ratio:  $a_{\text{exch}} = \prod_{n=1}^N \frac{\Phi(\hat{g}(x_n)) \Phi(g(w_n))}{\Phi(g(x_n)) \Phi(\hat{g}(w_n))}$ .
  - 19:     **if**  $\mathcal{U}(0, 1) < a_{\text{exch}}$  (Metropolis–Hastings acceptance rule) **then**
  - 20:       Replace the conditioning set:  $\mathbf{X} = \hat{\mathbf{X}}$  and  $\mathbf{G} = \hat{\mathbf{G}}$ .
  - 21:     **else**
  - 22:       Add fantasy data to the conditioning set:  $\mathbf{X} = \mathbf{X} \cup \mathcal{W}$  and  $\mathbf{G} = \mathbf{G} \cup \{g(w_n)\}_{n=1}^N$ .
  - 23:     **end if**
  - 24:   **end for**
- 

The normalisation constants cancel out, and the functions  $g(x)$  and  $\hat{g}(x)$  are only sampled from the Gaussian process at a finite number of locations.

Algorithm 2 shows the exchange sampling inference procedure for the GPDS. Some amount of bookkeeping is required for this procedure to be valid. Specifically, once something is learned about a particular function  $g(x)$ , i.e. sampled from the Gaussian process, it cannot be forgotten until that  $g(x)$  is discarded. For example, when fantasy data is generated from  $\hat{g}(x)$ , as in steps 6 to 16, even if  $\tilde{x}$  is rejected in step 10, the  $(\tilde{x}, \tilde{g})$  pair must be stored in the conditioning set (step 15). If the proposed  $\hat{g}(x)$  is ultimately rejected by step 19, only then can the conditioning set for  $\hat{g}(x)$  be discarded. Similarly, when the current Markov state  $g(x)$  is sampled from the Gaussian process at the fantasy data in step 17, this information must be kept if the proposal is rejected (step 22). Thus step 22 expands the Markov state with every rejection, as information about the current  $g(x)$  accumulates. When the proposal  $\hat{g}(x)$  is accepted, the Markov state reduces in size, as fewer points will typically have been sampled from  $\hat{g}(x)$ .

#### 4.1.1 Improving the acceptance rate

For clarity, we introduced the algorithm with  $q(\hat{\mathbf{g}} \leftarrow \mathbf{g}) = p(\hat{\mathbf{g}} | \theta)$ , but this proposal is a poor choice in practice. To achieve a better acceptance rate, it is better to make conservative, perturbative proposals. We define a set of  $K$  ‘‘control points’’ in  $\mathcal{X}$ , denoted  $\mathcal{C} = \{x_k \in \mathcal{X}\}_{k=1}^K$  and placed at arbitrary locations. We will assume that the control points include the data so that  $\mathcal{D} \subset \mathcal{C}$ .

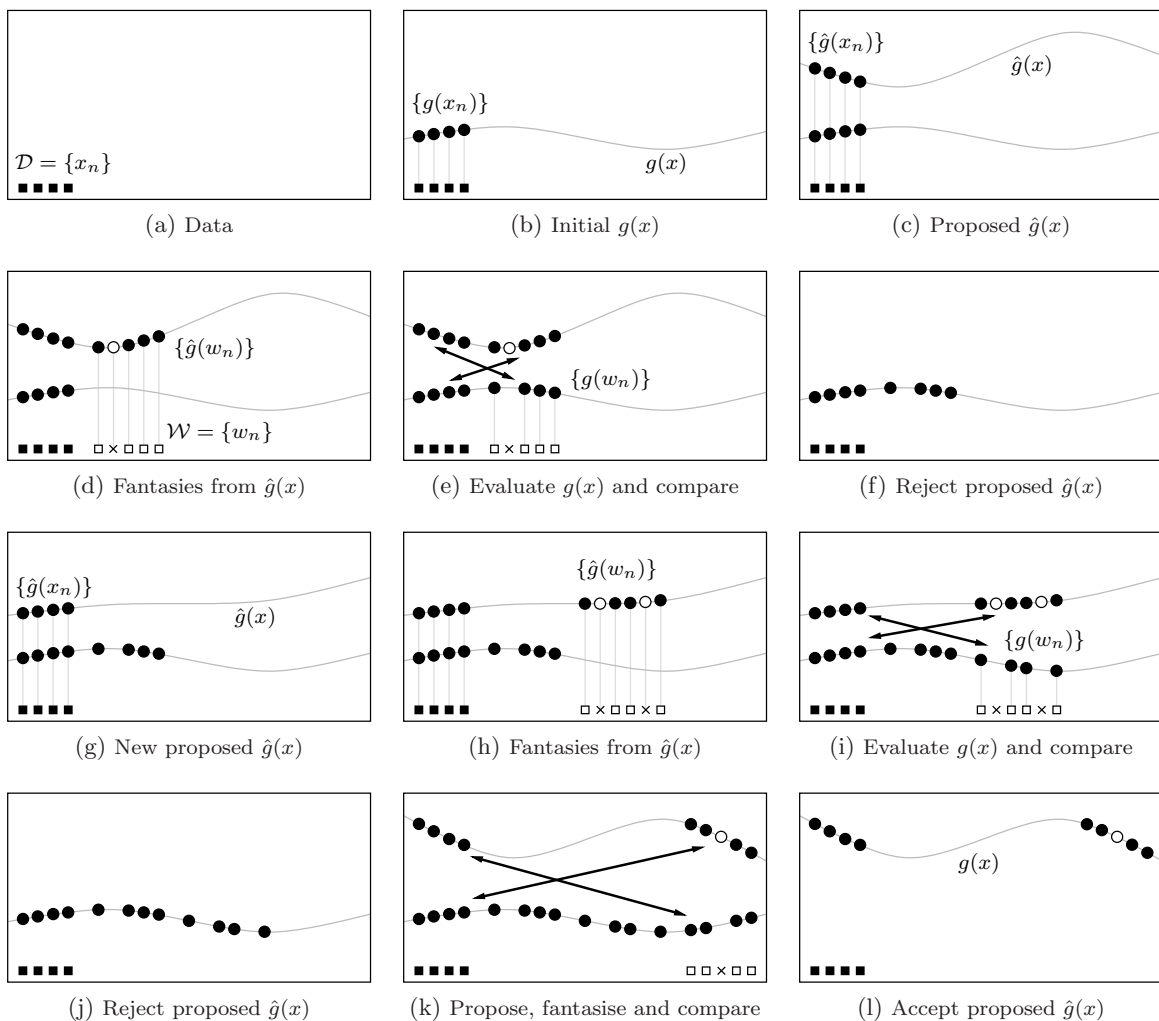


Figure 3: This sequence of figures shows a cartoon of three Metropolis–Hastings steps of the exchange sampler. In the first two M–H steps, the proposals are rejected to demonstrate the accumulation of Markov state. The third proposal is accepted and the previously-accumulated state is discarded. (a): The initial set of data, shown as solid boxes. (b): The initial  $g(x)$  evaluated at the data. (c): The proposed  $\hat{g}(x)$  evaluated at the data. (d): Fantasies are drawn from the density implied by  $\hat{g}(x)$ , shown as empty boxes. The fantasy generation rejected one proposal, shown as an  $x$  and a corresponding empty circle. This rejection contributes to the state of  $\hat{g}(x)$  but not to  $g(x)$ . (e):  $g(x)$  is evaluated at the fantasies and the two explanations are compared. (f): The proposal  $\hat{g}(x)$  is rejected. The conditioning set for  $g(x)$  has expanded to eight data from four. (g): A new proposal  $\hat{g}(x)$  is evaluated at the data. (h): Fantasies are drawn from the density implied by  $\hat{g}(x)$ . In this case, there were two rejections. (i):  $g(x)$  is evaluated at the fantasies and the explanations are compared. (j): The proposal was rejected, but we still must keep previous information about  $g(x)$ , so the conditioning set expands to twelve data. (k): Skipping the intermediate steps, we propose, fantasise, evaluate and compare. (l): We accept the proposal and make it the new  $g(x)$ . The Markov state shrinks from sixteen points in the conditioning set to nine.

This is not required for the algorithm to be valid; the function could be sampled at these locations given the control points. However, this sampling would be required at every Metropolis–Hastings step, so including the data in the control points is preferred. The control points have corresponding function values  $\{g(x_k)\}_{k=1}^K$ . As before, the Markov chain is still defined over the entire function, but we now condition the functions on the values at the control points. By including Gaussian process prior terms for the control points in the acceptance ratio, the equilibrium distribution remains the posterior distribution in Equation 3.

We make incremental proposals on the function values at the control points via a proposal density  $q(\{\hat{g}(x_k)\}_{k=1}^K \leftarrow \{g(x_k)\}_{k=1}^K)$ . When sampling from the Gaussian process to determine the latent functions “retrospectively,” we condition on  $\{g(x_k)\}_{k=1}^K$  or  $\{\hat{g}(x_k)\}_{k=1}^K$  for  $g(x)$  and  $\hat{g}(x)$ , respectively. With these additional points introduced, the exchange sampling acceptance ratio is now

$$a_{\text{exch-control}} = \frac{q(\{g(x_k)\}_{k=1}^K \leftarrow \{\hat{g}(x_k)\}_{k=1}^K) p(\{\hat{g}(x_k)\}_{k=1}^K | \theta)}{q(\{\hat{g}(x_k)\}_{k=1}^K \leftarrow \{g(x_k)\}_{k=1}^K) p(\{g(x_k)\}_{k=1}^K | \theta)} \prod_{n=1}^N \frac{\Phi(\hat{g}(x_n)) \Phi(g(w_n))}{\Phi(g(x_n)) \Phi(\hat{g}(w_n))}. \quad (7)$$

Note that these control points do not replace the conditioning set, but augment it in order to make improved proposals. As before, all information learned about any  $g(x)$  or  $\hat{g}(x)$  must be retained as long as that function is in the current Markov state.

#### 4.1.2 Hyperparameter inference

One of the benefits of the Bayesian approach is the ability to perform hierarchical inference. In this case, it allows us to infer the hyperparameters  $\theta$  of the Gaussian process. We can also introduce hyperparameters  $\phi$  that control the base density. We augment the exchange sampling algorithm slightly to sample from the posterior on hyperparameters: before proposing a new function  $\hat{g}(x)$ , we propose new hyperparameters  $\hat{\theta}$  and  $\hat{\phi}$  from a proposal density  $q(\hat{\theta}, \hat{\phi} \leftarrow \theta, \phi)$ . When samples of the new function are drawn, it is done with these proposed hyperparameters. The new joint distribution is

$$p(\mathbf{g}, \{x_n\}_{n=1}^N, \theta, \phi, \hat{\mathbf{g}}, \{w_n\}_{n=1}^N, \hat{\theta}, \hat{\phi}) = p(\mathbf{g} | \theta) p(\{x_n\}_{n=1}^N | \mathbf{g}, \phi) p(\theta, \phi) q(\hat{\theta}, \hat{\phi} \leftarrow \theta, \phi) p(\hat{\mathbf{g}} | \hat{\theta}) p(\{w_n\}_{n=1}^N | \hat{\mathbf{g}}, \hat{\phi}) \quad (8)$$

where  $p(\theta, \phi)$  is an appropriate hyperprior. The proposal is now to exchange the triplets  $(\mathbf{g}, \theta, \phi)$  and  $(\hat{\mathbf{g}}, \hat{\theta}, \hat{\phi})$ . The proposal of this swap has Metropolis–Hastings acceptance ratio

$$a_{\text{exch-hyper}} = \frac{q(\theta, \phi \leftarrow \hat{\theta}, \hat{\phi}) p(\hat{\theta}, \hat{\phi})}{q(\hat{\theta}, \hat{\phi} \leftarrow \theta, \phi) p(\theta, \phi)} \prod_{n=1}^N \frac{\Phi(\hat{g}(x_n)) \pi(x_n | \hat{\phi}) \Phi(g(w_n)) \pi(w_n | \phi)}{\Phi(g(x_n)) \pi(x_n | \phi) \Phi(\hat{g}(w_n)) \pi(w_n | \hat{\phi})}. \quad (9)$$

This acceptance ratio generalises straightforwardly to the case with control points discussed in Section 4.1.1.

#### 4.1.3 Sampling from the predictive distribution

An important task for in density inference is estimation of the predictive density. The predictive distribution arises on data space when the posterior is integrated out. It can also be thought of as

the distribution on the next datum to arrive, given the  $N$  already seen and taking uncertainty into account. In the GPDS, the predictive density is not available analytically. We nevertheless have all the tools in place to generate samples from the predictive distribution. We do this by using the generative procedure of Section 3 to generate additional data after each Metropolis–Hastings step. We use a very similar method to Algorithm 1, but initialise the conditioning set using the current state of the Markov chain.

## 4.2 Sampling over latent histories

An alternative to inference via exchange sampling is to model the *latent history* of the generative process. By using the GPDS prior to model the data, we are asserting that the data can be explained as the result of the procedure described in Section 3. We do not, however, know what rejections were made en route to accepting the observed data. These rejections are critical to defining the latent function  $g(x)$ . One might think of defining a density as analogous to putting up a tent: pinning the canvas down with pegs (or stakes) is just as important as putting up poles. In density modeling, defining regions with little probability mass is just as important as defining the areas with significant mass. In the exchange sampling algorithm of Section 4.1, the “pegs” are inferred implicitly as rejections along the way to generating fantasy data.

In this alternative approach, we infer the unknown rejections directly. Although the number and locations of the rejections are not known, the GPDS provides a probabilistic model that allows us to traverse the posterior distribution over possible latent histories that resulted in the data. If we define a Markov chain whose equilibrium distribution is the posterior distribution over latent histories, then we may simulate plausible explanations of every step taken to arrive at the data. Such samples capture all the information available about the unknown density, and with them we may ask additional questions about  $g(x)$  or run the generative procedure further to draw predictive samples. Significantly, the generative procedure did not require evaluating the normalisation constant, so we may perform this inference without evaluating it either. This approach is a generalisation of that described by Murray (2007), who performed inference on a coalesced Markov chain sampler that used coupling from the past (Propp and Wilson, 1996).

We model the data  $\mathcal{D} = \{x_n\}_{n=1}^N$  as having been generated exactly as in Algorithm 1, with  $P = N$ , i.e. run until exactly  $N$  proposals were accepted. The state space of the Markov chain on latent histories in the GPDS consists of: 1) the values of the latent function  $g(x)$  at the data, denoted  $\mathcal{G}_N = \{g(x_n)\}_{n=1}^N$ , 2) the number of rejections  $M$ , 3) the locations of the  $M$  rejected proposals, denoted  $\mathcal{M} = \{x_m\}_{m=1}^M$ , and 4) the values of the latent function  $g(x)$  at the  $M$  rejected proposals, denoted  $\mathcal{G}_M = \{g(x_m)\}_{m=1}^M$ .

We perform Gibbs-like sampling of the latent history by alternating between modification of the number of rejections  $M$  and block updating of the rejection locations  $\mathcal{M}$  and latent function values  $\mathcal{G}_M$  and  $\mathcal{G}_N$ . We will maintain an explicit ordering of the latent rejections for reasons of clarity, although this is not necessary due to exchangeability. We could propose at any time a reshuffling of the latent history, subject to it ending in an acceptance, and this proposal would always be accepted, as the two permutations have the same probability under the prior.

### 4.2.1 Modifying the number of latent rejections

We propose a new number of latent rejections  $\hat{M}$  by drawing it from a proposal distribution  $q(\hat{M} \leftarrow M)$ . If  $\hat{M}$  is greater than  $M$ , we must also propose new rejections to add to the latent

state. We take advantage of the exchangeability of the process to generate the new rejections: we imagine these proposals were made *after* the last observed datum was accepted, and our proposal is to call them rejections and move them *before* the last datum. If  $\hat{M}$  is less than  $M$ , we do the opposite by proposing to move some rejections to after the last acceptance.

When proposing additional rejections, we must also propose times for them among the current latent history. There are  $\binom{\hat{M}+N-1}{\hat{M}-M}$  such ways to insert these additional rejections into the existing latent history, such that the sampler terminates after the  $N$ th acceptance. When removing rejections, we must choose which ones to place after the data, and there are  $\binom{M}{M-\hat{M}}$  possible sets. Upon simplification, the proposal ratios for both addition and removal of rejections are identical:

$$\frac{\overbrace{q(M \leftarrow \hat{M}) \binom{\hat{M}+N-1}{\hat{M}-M}}^{\hat{M} > M}}{\overbrace{q(\hat{M} \leftarrow M) \binom{M}{M-\hat{M}}}^{\hat{M} < M}} = \frac{\overbrace{q(M \leftarrow \hat{M}) \binom{M}{M-\hat{M}}}^{\hat{M} < M}}{\overbrace{q(\hat{M} \leftarrow M) \binom{\hat{M}+N-1}{M-\hat{M}}}^{\hat{M} > M}} = \frac{q(M \leftarrow \hat{M}) M! (\hat{M} + N - 1)!}{q(\hat{M} \leftarrow M) \hat{M}! (M + N - 1)!}.$$

When inserting rejections, we propose the locations of the additional proposals, denoted  $\mathcal{M}^+$ , and the corresponding values of the latent function, denoted  $\mathcal{G}_M^+$ . We generate  $\mathcal{M}^+$  by making  $\hat{M} - M$  independent draws from the base density. We draw  $\mathcal{G}_M^+$  jointly from the Gaussian process prior, conditioned on all of the current latent state, i.e.  $(\mathcal{M}, \mathcal{G}_M, \mathcal{D}, \mathcal{G}_N)$ . The joint probability of this state is

$$p(\mathcal{D}, \mathcal{M}, \mathcal{M}^+, \mathcal{G}_N, \mathcal{G}_M, \mathcal{G}_M^+) = \left[ \prod_{n=1}^N \pi(x_n) \Phi(g(x_n)) \right] \left[ \prod_{m=1}^M \pi(x_m) (1 - \Phi(g(x_m))) \right] \left[ \prod_{m=M+1}^{\hat{M}} \pi(x_m) \right] \times \mathcal{GP}(\mathcal{G}_M, \mathcal{G}_N, \mathcal{G}_M^+ | \mathcal{D}, \mathcal{M}, \mathcal{M}^+). \quad (10)$$

The joint distribution in Equation 10 expresses the probability of all the base density draws, the values of the function draws from the Gaussian process, and the acceptance or rejection probabilities of the proposals *excluding* the newly generated points. When we make an insertion proposal, exchangeability allows us to shuffle the ordering without changing the probability; the only change is that now we must account for labeling the new points as rejections. In the acceptance ratio, all terms except for the “labeling probability” cancel. The reverse proposal is similar, however we denote the removed proposal locations as  $\mathcal{M}^-$  and the corresponding function values as  $\mathcal{G}_M^-$ . The overall acceptance ratios for insertions or removals are

$$a_{\text{hist-num}} = \begin{cases} \frac{q(M \leftarrow \hat{M}) M! (\hat{M} + N - 1)!}{q(\hat{M} \leftarrow M) \hat{M}! (M + N - 1)!} \prod_{g \in \mathcal{G}_M^+} (1 - \Phi(g)) & \text{if } \hat{M} > M \\ \frac{q(M \leftarrow \hat{M}) M! (\hat{M} + N - 1)!}{q(\hat{M} \leftarrow M) \hat{M}! (M + N - 1)!} \prod_{g \in \mathcal{G}_M^-} (1 - \Phi(g))^{-1} & \text{if } \hat{M} < M. \end{cases} \quad (11)$$

#### 4.2.2 Modifying rejection locations and function values

Given the number of latent rejections  $M$ , we propose modifying their locations  $\mathcal{M}$ , their latent function values  $\mathcal{G}_M$ , and the values of the latent function at the data  $\mathcal{G}_N$ . We will denote these proposals as  $\hat{\mathcal{M}} = \{\hat{x}_m\}_{m=1}^M$ ,  $\hat{\mathcal{G}}_M = \{\hat{g}(\hat{x}_m)\}_{m=1}^M$  and  $\hat{\mathcal{G}}_N = \{\hat{g}(x_n)\}_{n=1}^N$ , respectively. We make simple perturbative proposals of  $\mathcal{M}$  via a proposal density  $q(\hat{\mathcal{M}} \leftarrow \mathcal{M})$ . However, for the latent

---

**Algorithm 3** Generate  $T$  samples from the posterior on  $g(x)$  and the latent history.

---

**Purpose:** MCMC sample from the posterior on latent histories of the generative process.

**Inputs:** GP hyperparameters  $\theta$ , number of samples  $T$ , data  $\mathcal{D} = \{x_n\}_{n=1}^N$ , step size  $\alpha$

- 1: Sample  $g(x)$  from the GP at the data:  $\mathcal{G}_N \sim \mathcal{GP}(g | \mathcal{D}, \theta)$ .
  - 2: Initialise the number of latent rejections to zero:  $M = 0$ .
  - 3: Initialise the latent rejections to empty sets:  $\mathcal{M} = \emptyset$  and  $\mathcal{G}_M = \emptyset$ .
  - 4: **for**  $t = 1$  to  $T$  **do**
  - 5:   Propose a new number of latent rejections:  $\hat{M} \sim q(\hat{M} \leftarrow M)$ .
  - 6:   **if**  $\hat{M} > M$  (Increasing the number of rejections) **then**
  - 7:     Propose the locations of the  $\hat{M} - M$  new rejections:  $\mathcal{M}^+ \sim \pi(x)$ .
  - 8:     Propose the function values of the new rejections:  $\mathcal{G}_M^+ \sim \mathcal{GP}(g | \mathcal{D}, \mathcal{G}, \mathcal{M}_M, \mathcal{G}_M, \mathcal{M}^+)$ .
  - 9:     Calculate the acceptance ratio:  $a_{\text{hist-num}} = \frac{q(M \leftarrow \hat{M})M!(\hat{M}+N-1)!}{q(\hat{M} \leftarrow M)\hat{M}!(M+N-1)!} \prod_{g \in \mathcal{G}_M^+} (1 - \Phi(g))$
  - 10:    **if**  $\mathcal{U}(0, 1) < a_{\text{hist-num}}$  (Acceptance rule) **then**
  - 11:     Add new rejections to the history in random locations:  $\mathcal{M} = \mathcal{M} \cup \mathcal{M}^+$  and  $\mathcal{G}_M = \mathcal{G}_M \cup \mathcal{G}_M^+$ .
  - 12:    **end if**
  - 13:    **else**
  - 14:     Propose  $M - \hat{M}$  rejections  $\mathcal{M}^-$  to remove from  $\mathcal{M}$ .
  - 15:     Calculate the acceptance ratio:  $a_{\text{hist-num}} = \frac{q(M \leftarrow \hat{M})M!(\hat{M}+N-1)!}{q(\hat{M} \leftarrow M)\hat{M}!(M+N-1)!} \prod_{g \in \mathcal{G}_M^-} (1 - \Phi(g))^{-1}$
  - 16:     **if**  $\mathcal{U}(0, 1) < a_{\text{hist-num}}$  (Acceptance rule) **then**
  - 17:      Remove the rejections in  $\mathcal{M}^-$  from the history:  $\mathcal{M} = \mathcal{M} / \mathcal{M}^-$  and  $\mathcal{G}_M = \mathcal{G}_M / \mathcal{G}_M^-$ .
  - 18:     **end if**
  - 19:    **end if**
  - 20:    Propose new rejection locations:  $\hat{\mathcal{M}} \sim q(\hat{\mathcal{M}} \leftarrow \mathcal{M})$ .
  - 21:    Sample the GP at the new rejections:  $\{g(\hat{x}_m)\}_{m=1}^M \sim \mathcal{GP}(g | \mathcal{D}, \mathcal{G}_N, \mathcal{M}, \mathcal{G}_M, \hat{\mathcal{M}}, \theta)$ .
  - 22:    Draw the GP for the underrelaxed step:  $\mathbf{h} \sim \mathcal{GP}(h | \mathcal{D}, \hat{\mathcal{M}}, \theta)$ .
  - 23:    Take the underrelaxed step:  $(\hat{\mathcal{G}}_N, \hat{\mathcal{G}}_M) = \alpha(\mathcal{G}_N, \{g(\hat{x}_m)\}_{m=1}^M) + \sqrt{1 - \alpha^2}\mathbf{h}$ .
  - 24:    Calculate acceptance ratio:  $a_{\text{hist-locs}} = \frac{q(\mathcal{M} \leftarrow \hat{\mathcal{M}}) \left[ \prod_{m=1}^M \pi(\hat{x}_m)(1 - \Phi(\hat{g}(x_m))) \right] \left[ \prod_{n=1}^N \Phi(\hat{g}(x_n)) \right]}{q(\hat{\mathcal{M}} \leftarrow \mathcal{M}) \left[ \prod_{m=1}^M \pi(x_m)(1 - \Phi(g(x_m))) \right] \left[ \prod_{n=1}^N \Phi(g(x_n)) \right]}$
  - 25:    **if**  $\mathcal{U}(0, 1) < a_{\text{hist-locs}}$  (Acceptance rule) **then**
  - 26:     Accept new history:  $\mathcal{M} = \hat{\mathcal{M}}$ ,  $\mathcal{G}_N = \hat{\mathcal{G}}_N$ , and  $\mathcal{G}_M = \hat{\mathcal{G}}_M$ .
  - 27:    **end if**
  - 28: **end for**
- 

function values perturbative proposals will be poor, as the Gaussian process typically defines a narrow mass. To avoid a low acceptance rate, we propose modifications to the latent function that leave the prior invariant.

We make joint proposals of  $\hat{\mathcal{M}}$ ,  $\hat{\mathcal{G}}_M$  and  $\hat{\mathcal{G}}_N$  in three steps. First, we draw new rejection locations from  $q(\hat{\mathcal{M}} \leftarrow \mathcal{M})$ . Second, we draw a set of  $M$  intermediate function values from the Gaussian process at  $\hat{\mathcal{M}}$ , conditioned on the current rejection locations and their function values, as well as the function values at the data. Third, we propose new function values at  $\hat{\mathcal{M}}$  and the data  $\mathcal{D}$  via an underrelaxation proposal of the form

$$\hat{g}(x) = \alpha g(x) + \sqrt{1 - \alpha^2} h(x)$$

where  $h(x)$  is an independent draw from the Gaussian process prior and  $\alpha \in [0, 1)$ . This is a variant of the overrelaxed MCMC method discussed by Adler (1981) and expanded by Neal (1998), which we call ‘‘underrelaxation.’’ An underrelaxed step leaves the Gaussian process prior invariant, but makes conservative proposals if  $\alpha$  is near one. After making a proposal, we accept or reject via the

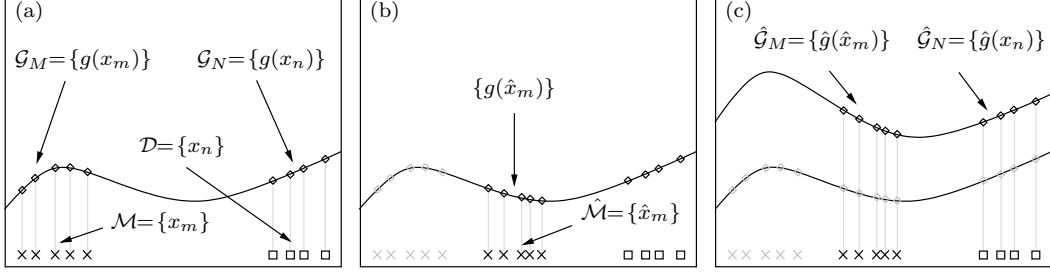


Figure 4: These figures show the sequence of proposing new rejection locations, new function values at those locations, and new function values at the data. (a): The current state, with rejections labeled  $\mathcal{M} = \{x_m\}$  on the left, along with the values of the latent function  $\mathcal{G}_M = \{g_m\}$ . On the right side are the data  $\mathcal{D} = \{x_n\}$  and the corresponding values of the latent function  $\mathcal{G}_N = \{g_n\}$ . (b): New rejections  $\hat{\mathcal{M}} = \{\hat{x}_m\}$  are proposed via  $q(\hat{\mathcal{M}} \leftarrow \mathcal{M})$ , and the latent function is sampled at these points. (c): The latent function is perturbed at the new rejection locations and at the data via an underrelaxed proposal.

ratio of the joint distributions:

$$a_{\text{hist-locs}} = \frac{q(\mathcal{M} \leftarrow \hat{\mathcal{M}}) \left[ \prod_{m=1}^M \pi(\hat{x}_m)(1 - \Phi(\hat{g}(x_m))) \right] \left[ \prod_{n=1}^N \Phi(\hat{g}(x_n)) \right]}{q(\hat{\mathcal{M}} \leftarrow \mathcal{M}) \left[ \prod_{m=1}^M \pi(x_m)(1 - \Phi(g(x_m))) \right] \left[ \prod_{n=1}^N \Phi(g(x_n)) \right]}.$$

The overall latent history algorithm is shown in Algorithm 3.

### 4.2.3 Hyperparameter inference

Given a sample from the posterior on the latent history, we can also perform a Metropolis–Hasting step in the space of hyperparameters. As in Section 4.1.2, we have hyperparameters  $\theta$  for the Gaussian process and  $\phi$  for the base density, with joint prior density  $p(\theta, \phi)$ . We draw proposals  $\hat{\theta}$  and  $\hat{\phi}$  from the proposal density  $q(\hat{\theta}, \hat{\phi} \leftarrow \theta, \phi)$ . The acceptance ratio for a Metropolis–Hastings step in the posterior of the hyperparameters is

$$a_{\text{hist-hp}} = \frac{q(\theta, \phi \leftarrow \hat{\theta}, \hat{\phi}) p(\hat{\theta}, \hat{\phi}) \mathcal{N}(\{\mathcal{G}_M, \mathcal{G}_N\} | \mathcal{M}, \mathcal{D}, \hat{\theta}) \left[ \prod_{m=1}^M \frac{\pi(x_m | \hat{\phi})}{\pi(x_m | \phi)} \right] \left[ \prod_{n=1}^N \frac{\pi(x_n | \hat{\phi})}{\pi(x_n | \phi)} \right]}{q(\hat{\theta}, \hat{\phi} \leftarrow \theta, \phi) p(\theta, \phi) \mathcal{N}(\{\mathcal{G}_M, \mathcal{G}_N\} | \mathcal{M}, \mathcal{D}, \theta) \left[ \prod_{m=1}^M \frac{\pi(x_m | \phi)}{\pi(x_m | \hat{\phi})} \right] \left[ \prod_{n=1}^N \frac{\pi(x_n | \phi)}{\pi(x_n | \hat{\phi})} \right]}.$$

### 4.2.4 Calculating the predictive density

In the latent history case, we can sample from the predictive distribution as we did with the exchange sampling method in Section 4.1.3. We run the generative process forward from the current latent history sample to generate a predictive sample. It may also be desirable to estimate the actual value of the predictive density, and this is available when performing inference via the latent history. We use the method of Chib and Jeliazkov (2001), and consider making a Metropolis–Hastings proposal of moving from  $x$  to  $x'$  in the predictive density. We use the base density  $\pi(x)$  as the proposal density and from the detailed balance requirement of Metropolis–Hastings we get the identity

$$p(x | \mathbf{g}, \theta, \phi) \pi(x') \min \left( 1, \frac{\Phi(g(x'))}{\Phi(g(x))} \right) = p(x' | \mathbf{g}, \theta, \phi) \pi(x) \min \left( 1, \frac{\Phi(g(x))}{\Phi(g(x'))} \right). \quad (12)$$

We integrate this identity over  $x'$

$$\int dx' p(x | \mathbf{g}, \theta, \phi) \pi(x') \min \left( 1, \frac{\Phi(g(x'))}{\Phi(g(x))} \right) = \int dx' p(x' | \mathbf{g}, \theta, \phi) \pi(x) \min \left( 1, \frac{\Phi(g(x))}{\Phi(g(x'))} \right) \quad (13)$$

and then find the expectation under  $p(\mathbf{g}, \theta, \phi | \mathcal{D})$ , the posterior density on the unknowns:

$$\begin{aligned} \int d\theta \int d\phi \int d\mathbf{g} p(\mathbf{g}, \theta, \phi | \mathcal{D}) \int dx' p(x | \mathbf{g}, \theta, \phi) \pi(x') \min \left( 1, \frac{\Phi(g(x'))}{\Phi(g(x))} \right) \\ = \int d\theta \int d\phi \int d\mathbf{g} p(\mathbf{g}, \theta, \phi | \mathcal{D}) \int dx' p(x' | \mathbf{g}, \theta, \phi) \pi(x) \min \left( 1, \frac{\Phi(g(x))}{\Phi(g(x'))} \right). \end{aligned}$$

From this we can find an expression for the predictive density:

$$p(x | \mathcal{D}) = \frac{\int d\theta \int d\phi \int d\mathbf{g} \int dx' p(\theta, \phi, \mathbf{g}, x' | \mathcal{D}) \pi(x) \min \left( 1, \frac{\Phi(g(x))}{\Phi(g(x'))} \right)}{\int d\theta \int d\phi \int d\mathbf{g} \int dx' p(\theta, \phi, \mathbf{g} | x, \mathcal{D}) \pi(x') \min \left( 1, \frac{\Phi(g(x'))}{\Phi(g(x))} \right)} \quad (14)$$

Both the numerator and the denominator in Equation 14 are expectations. The top is an expectation under the posterior and the bottom is an expectation under the posterior where the data has been augmented with  $x$ :

$$p(x | \mathcal{D}) = \frac{\pi(x) \mathbb{E}_{p(\cdot | \mathcal{D})} \left[ \min \left( 1, \frac{\Phi(g(x))}{\Phi(g(x'))} \right) \right]}{\mathbb{E}_{p(\cdot | \mathcal{D}, x)} \left[ \pi(x') \min \left( 1, \frac{\Phi(g(x'))}{\Phi(g(x))} \right) \right]}. \quad (15)$$

The numerator can be estimated directly as a part of the latent history MCMC inference with little additional computation. The denominator requires running a latent history Markov chain that treats  $x$  as an additional observed datum.

### 4.3 Computational issues

Computation with a Gaussian process is expensive. If the GP is realised on  $R$  points, the space complexity of storing the covariance (Gram) matrix is  $O(R^2)$  and the time complexity of decomposing (or inverting) the matrix is  $O(R^3)$ . The time cost of this decomposition will be the asymptotically-dominating factor when performing GPDS inference using either exchange sampling or the latent history method. In the exchange sampling algorithm, the *minimum* number of points from which the Gaussian process is sampled is  $2N$  — at the true data and at the fantasies. It is likely that there will be rejections along the way to generating the fantasies as well. When an exchange sampling proposal is rejected, the size of the conditioning set increases by  $N$  to include the new fantasies.

We expect that the latent history method will typically be more efficient, as it requires fewer evaluations of  $g(x)$ . The Gaussian process in the latent history method maintains  $M + N$  points. When both methods are at equilibrium, we expect that  $M$  in the latent history method will be of roughly the same size as the number of rejections required to generate fantasy data. In this case, however, the exchange sampler must include at least  $N$  more data in the Gaussian process to calculate the Metropolis–Hastings acceptance ratio. Experiments have supported the conclusion that the latent history performs better than exchange sampling.

We are optimistic that more sophisticated Markov chain Monte Carlo techniques may realize constant-factor performance gains over the basic Metropolis–Hasting schemes presented here, without compromising the correctness of the equilibrium distributions. Sparse approaches to Gaussian process regression that improve the asymptotically cubic behavior may also be relevant to the GPDS, but it is unclear that these will be an improvement over other approximate GP-based schemes for density modeling.

#### 4.4 Restricting the function space

With both the exchange sampling and latent history methods, incorporating fewer latent rejections (“tent pegs”) into the Gaussian process results in improved efficiency. For a given  $g(x)$ , the expected number of rejections is  $N(\mathcal{Z}_\pi[g]^{-1} - 1)$ . This expression is derived from the observation that  $\pi(x)$  provides an upper bound on the function  $\Phi(g(x))\pi(x)$  and the ratio of acceptances to rejections is determined by the proportion of the mass of  $\pi(x)$  contained by  $\Phi(g(x))\pi(x)$ . One problem with inference is that there are many functions  $g(x)$  that can explain the data equivalently, as  $\Phi(g(x))\pi(x)$  is unnormalised. Many of these  $g(x)$  will cause  $\Phi(g(x))$  to be close to zero, resulting in many rejections. The Gaussian process prior might only provide weak regularisation to prevent this.

One way to improve this situation is to require that the function  $g(x)$  be pinned to zero for some  $x_0$ . This does not limit the support of the prior, but prevents  $\Phi(g(x))\pi(x)$  from being small everywhere. We use the base density  $\pi(x)$  as a prior on  $x_0$  and treat it as a hyperparameter for the Gaussian process. We can then use the inference methods of Sections 4.1.2 and 4.2.3 to infer an appropriate  $x_0$ .

#### Acknowledgements

The authors wish to thank Radford Neal and Zoubin Ghahramani for valuable comments. Ryan Adams’ research is supported by the Gates Cambridge Trust. Iain Murray’s research is supported by the government of Canada.

#### References

- S. L. Adler. Over-relaxation method for the Monte Carlo evaluation of the partition function for multi-quadratic actions. *Physics Review D*, 23(12):2901–2904, 1981.
- A. Beskos, O. Papaspiliopoulos, G. O. Roberts, and P. Fearnhead. Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion). *Journal of the Royal Statistical Society: Series B*, 68:333–382, 2006.
- S. Chib and I. Jeliazkov. Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association*, 96(453):270–281, 2001.
- L. Csató. *Gaussian processes - iterative sparse approximations*. PhD thesis, Aston University, Birmingham, UK, March 2002.
- I. DiMatteo, C. R. Genovese, and R. E. Kass. Bayesian curve-fitting with free-knot splines. *Biometrika*, 88(4):1055–1071, 2001.
- M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588, June 1995.

- T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.
- H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, March 2001.
- P. J. Lenk. The logistic normal distribution for Bayesian, nonparametric, predictive densities. *Journal of the American Statistical Association*, 83(402):509–516, 1988.
- P. J. Lenk. Towards a practicable Bayesian nonparametric density estimator. *Biometrika*, 78(3):531–543, 1991.
- T. Leonard. Density estimation, stochastic processes and prior information. *Journal of the Royal Statistical Society, Series B*, 40(2):113–146, 1978.
- A. Y. Lo. On a class of Bayesian nonparametric estimates: I. density estimates. *The Annals of Statistics*, 12(1):351–357, March 1984.
- D. J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.
- J. Møller, A. N. Pettit, R. Reeves, and K. K. Bethelsen. An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458, 2006.
- I. Murray. *Advances in Markov chain Monte Carlo methods*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, London, 2007.
- I. Murray, Z. Ghahramani, and D. J. MacKay. MCMC for doubly-intractable distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 359–366, 2006.
- R. M. Neal. Suppressing random walks in Markov chain Monte Carlo using ordered overrelaxation, 1998.
- R. M. Neal. Defining priors for distributions using Dirichlet diffusion trees. Technical Report 0104, Department of Statistics, University of Toronto, 2001.
- R. M. Neal. Density modeling and clustering using Dirichlet diffusion trees. In *Bayesian Statistics 7*, pages 619–629, 2003.
- A. O’Hagan. Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society, Series B*, 40:1–42, 1978.
- O. Papaspiliopoulos and G. O. Roberts. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 95(1):169–186, 2008.
- J. Pitman and M. Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25:855–900, 1997.
- J. G. Propp and D. B. Wilson. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms*, 9(1&2):223–252, 1996.
- C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.
- D. Thorburn. A Bayesian approach to density estimation. *Biometrika*, 73(1):65–75, 1986.
- S. T. Tokdar and J. K. Ghosh. Posterior consistency of logistic Gaussian process priors in density estimation. *Journal of Statistical Planning and Inference*, 137:34–42, 2007.