

# The Gaussian Process Density Sampler

Ryan Prescott Adams and David J.C. MacKay

Cavendish Laboratory

University of Cambridge

Cambridge CB3 0HE, U.K.

rpa23@cam.ac.uk, mackay@mrao.cam.ac.uk

The Gaussian process is a useful prior on functions for Bayesian regression and classification. Density estimation with a Gaussian process prior has been difficult, however, due to the requirements that densities be nonnegative and integrate to unity. The statistics community has explored the use of a logistic Gaussian process for density estimation, relying on various methods of approximating the normalization constant (e.g. [1, 4]).

We propose the Gaussian Process Density Sampler (GPDS), a nonparametric, practical and consistent method of constructing a Markov chain on the properties of a posterior distribution on an unknown density, without approximation. The GPDS is composed of four parts. The first part is a GP-based prior on density functions. We develop an exchangeable procedure for generating exact samples in data space from a common density drawn from this prior. Second, we show that this prior allows practical inference of specific values of the unnormalized density, using the recently-developed technique of *exchange sampling* [3]. Third, we extend this MCMC algorithm to draw samples from the predictive distribution on data space that arises when the posterior on density functions is integrated out. This is our primary result. Finally, we demonstrate a sampling procedure for inference of the Gaussian process hyperparameters.

**The Prior on Density Functions** We define a prior distribution on densities over a space  $\mathcal{X}$  via a Gaussian process prior over functions  $g(x) : \mathcal{X} \rightarrow \mathbb{R}$  so that each  $g$  corresponds to a density  $f$  via

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

where  $x \in \mathcal{X}$ ,  $\omega(x)$  is an arbitrary probability measure on  $\mathcal{X}$ ,  $\Phi(\cdot)$  is any strictly increasing bounded nonnegative function and  $\mathcal{Z}_\omega[g]$  is the (generally intractable) normalization constant.

We generate exact and exchangeable data samples from a common density drawn from this prior by a rejection sampling procedure that “discovers” the sample of  $g$  as it proceeds. A proposal in data space  $\hat{x}$  is drawn from  $\omega(x)$  and a corresponding value of  $g(\hat{x})$  is drawn from the Gaussian process prior, conditioned on all previous proposals. The proposal is then accepted or rejected based on comparison between  $\Phi(g(\hat{x}))$  and a uniform draw between zero and the upper bound provided by  $\Phi$ .

**Predictive Samples** One of the reasons for the ubiquity of the Metropolis–Hastings algorithm is that it avoids evaluation of the evidence integral in Bayesian posterior distributions: the probability of the data when the parameters are integrated out. Some probabilistic models, however, also involve difficult integrals to evaluate the likelihood function as well. Such distributions are called *doubly-intractable*, and energy-based models such as the Ising and Potts models provide common examples. In these models, determining the acceptance ratio of a Metropolis step involves an intractable ratio of normalization constants.

When exact samples can be generated in data space, however, it is possible to construct a Metropolis algorithm that has both a tractable acceptance ratio and the desired equilibrium distribution [2, 3]. In the GPDS we show that it is possible to use this formulation to construct a Markov chain on a finite set of values of  $g$  such that the values correspond to the posterior distribution on density functions, without approximation of the normalization constant  $Z_\omega[g]$ .

While the values of  $g$  are useful for evaluating the ratio of two density values in data space, the larger utility of this construction is that we may use the generative procedure described above to sample exactly from the predictive distribution. In other words, we may augment the Markov chain to also generate fantasies of the next datum we are likely to see, having already observed  $N$  data. Additionally, if we choose  $\omega$  in such a way that conditional proposals can be easily made, we can fantasize samples from conditional predictive distributions, which are often useful in machine learning.

**Hyperparameter Selection** One of the difficult tasks in traditional kernel density estimation is selection of an appropriate bandwidth parameter. In the GPDS, this corresponds roughly to selection of the Gaussian process hyperparameters. By augmenting the Markov chain to include the hyperparameters we can integrate out these values and include the uncertainty of the “bandwidth” into the predictive estimates. We use a Gibbs-like procedure of making hyperparameter proposals with fixed settings for the finite representation of  $g$ .

**Discussion** Computationally, the GPDS has similar scaling properties to hyperparameter inference in Gaussian process regression:  $O(N^3)$  per step where  $N$  is the number of data. The rejection rate adds an additional constant factor that may be significant in practice.

## References

- [1] P.J. Lenk. Towards a practicable Bayesian nonparametric density estimator. *Biometrika*, 78(3):531–543, 1991.
- [2] J. Møller, A. N. Pettitt, R. Reeves, and K. K. Berthelsen. An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458, 2006.
- [3] Iain Murray, Zoubin Ghahramani, and David J.C. MacKay. MCMC for doubly-intractable distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2006.
- [4] 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.

**Presenter: RPA**

**Category: Learning Algorithms**

**Oral Presentation Preferred**