

The Gaussian Process Density Sampler

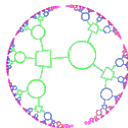
Ryan Prescott Adams

Cavendish Laboratory
University of Cambridge
<http://www.inference.phy.cam.ac.uk/rpa23/>

1 April 2008
Snowbird Learning Workshop



Joint work with Iain Murray
and David MacKay



The Density Modeling Problem

The setup:

- ▶ Data $\{x_n\}_{n=1}^N$ from an unknown density $f(x)$
- ▶ Prior beliefs about $f(x)$
- ▶ What is the posterior on $f(x)$?

Nonparametric density models:

- ▶ Parzen windows
- ▶ Infinite mixtures of parametric distributions
- ▶ Dirichlet diffusion trees (Neal 2001)
- ▶ Gaussian process latent variable models (Lawrence 2005)

Density Modeling with GPs

Gaussian processes:

- ▶ Useful priors on functions
- ▶ Specify behavior via covariance kernel

GPs for density functions - not a new idea:

- ▶ Leonard (1978)
- ▶ Lenk (1988, 1991)

Using GPs for PDFs is hard:

$$f(x) = \frac{\exp\{g(x)\}}{\int dx \exp\{g(x)\}}$$

$$g(x) \sim \mathcal{GP}(0, K(x, x'))$$

Gaussian Process Density Sampler

Four parts to the GPDS:

- 1) Specify a GP-based prior on densities.
- 2) Construct an MCMC algorithm on the density.
- 3) Draw samples from the predictive distribution.
- 4) Sample from the hyperparameters of the GP.

The Prior on Densities

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

$$\mathcal{Z}_\pi[g] = \int dx \Phi(g(x)) \pi(x)$$

- ▶ $g(x) \sim \mathcal{GP}(0, K(x, x'))$
- ▶ $\Phi(\cdot)$ is nonnegative and bounded.
 - ▶ We'll use the logistic: $\Phi(z) = (1 + \exp(-z))^{-1}$.
- ▶ $\pi(x)$ is a known base measure.

We can generate exact, exchangeable samples from a common density drawn from this prior.

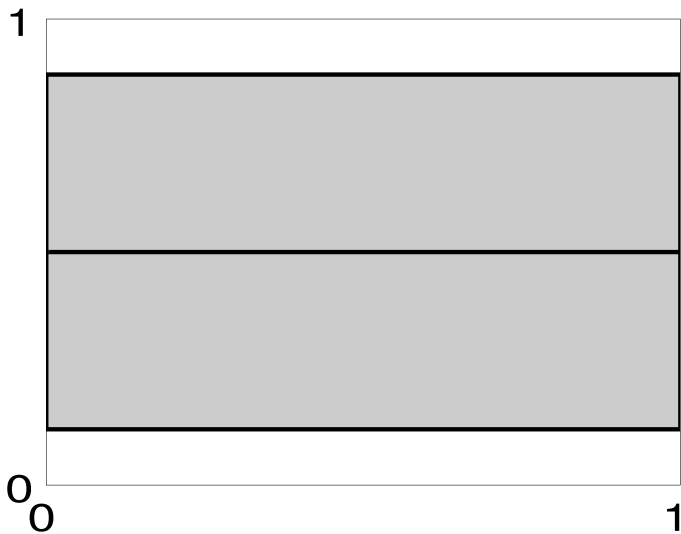
Sampling While Discovering $g(x)$

$$f(x) = \frac{1}{Z_{\pi}[g]} \Phi(g(x)) \pi(x)$$

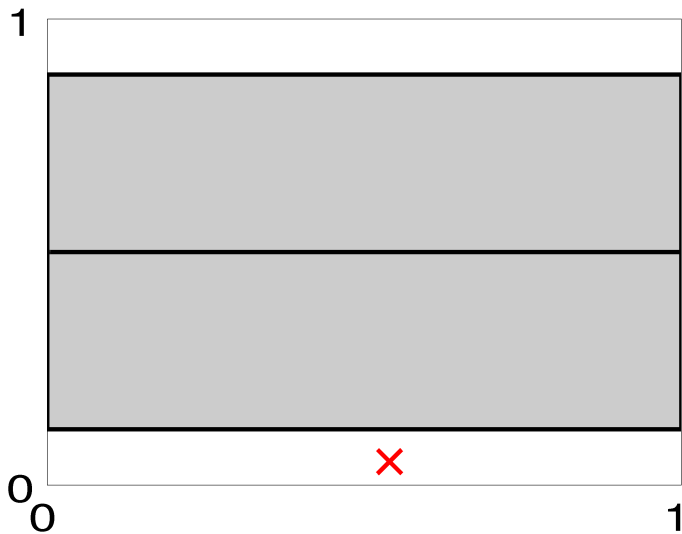
Rejection sampling:

1. Draw \tilde{x} from $\pi(x)$.
2. Sample $g(\tilde{x})$ from GP given all past function samples.
3. Draw r from UNIFORM(0, 1).
4. Accept if $r < \Phi(g(\tilde{x}))$.
5. Goto 1

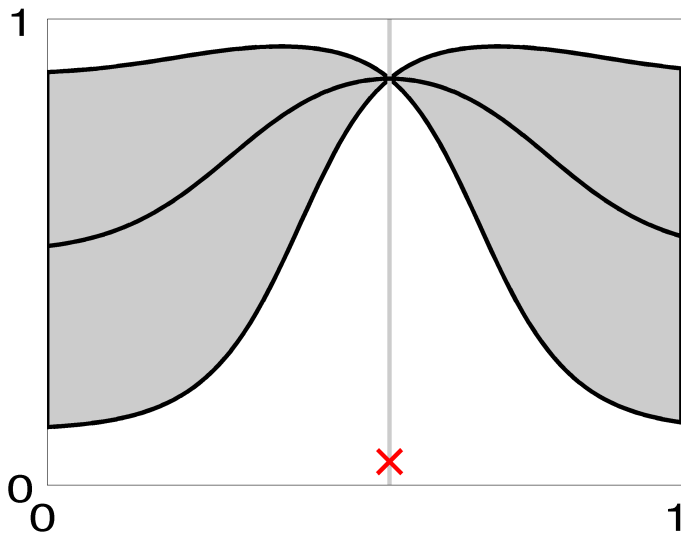
Sampling While Discovering $g(x)$



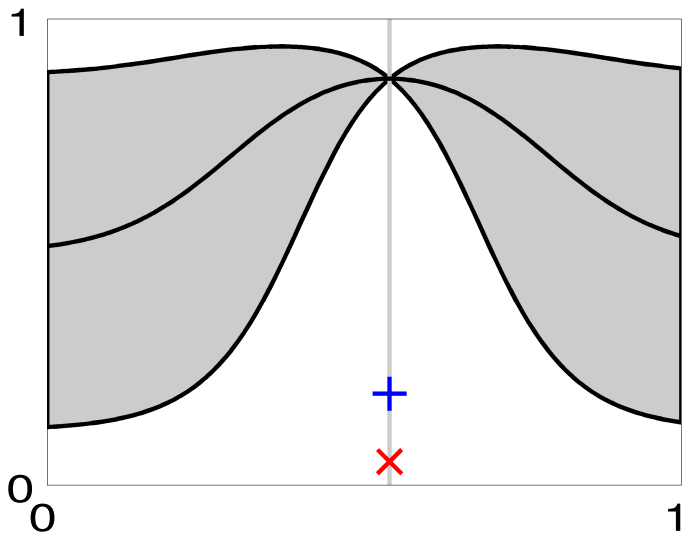
Sampling While Discovering $g(x)$



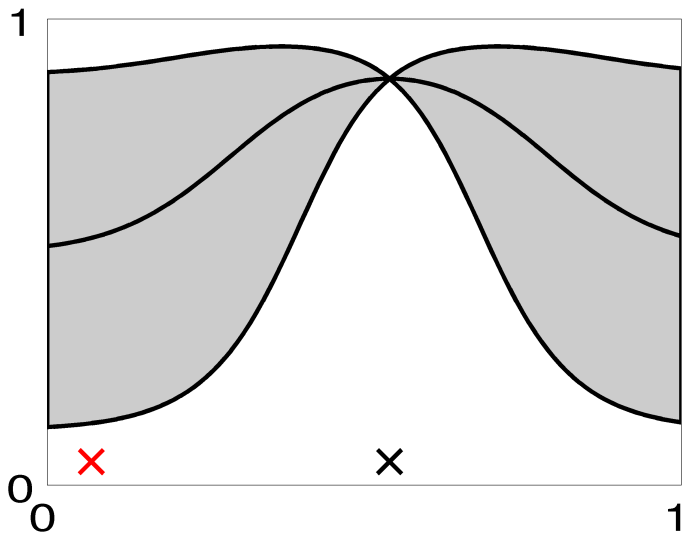
Sampling While Discovering $g(x)$



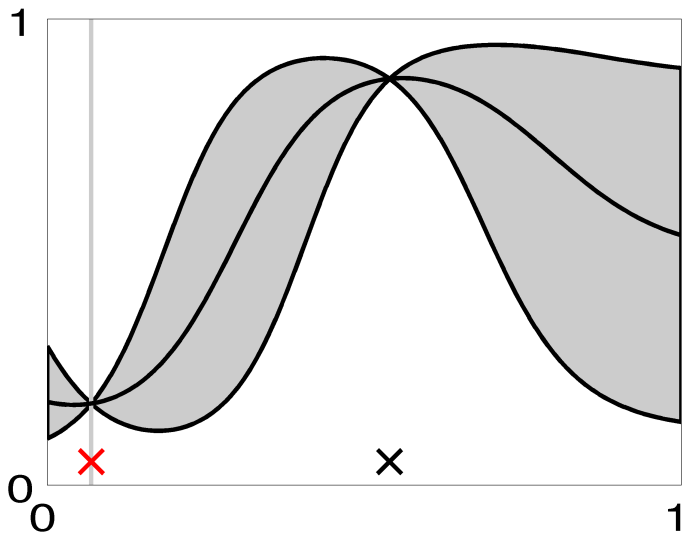
Sampling While Discovering $g(x)$



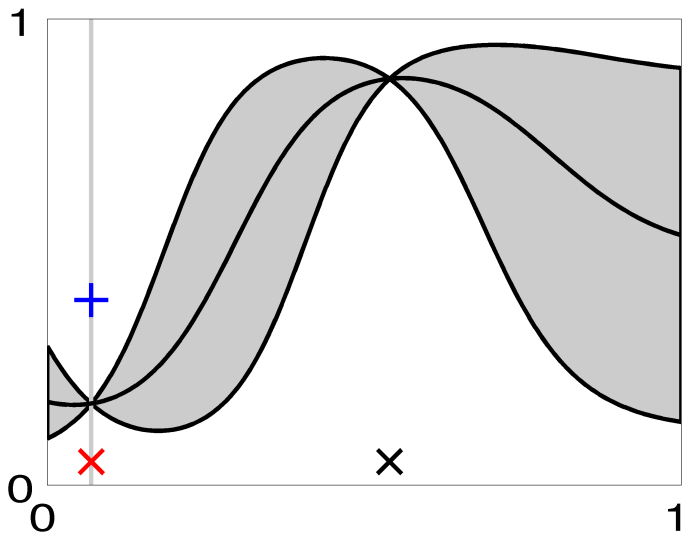
Sampling While Discovering $g(x)$



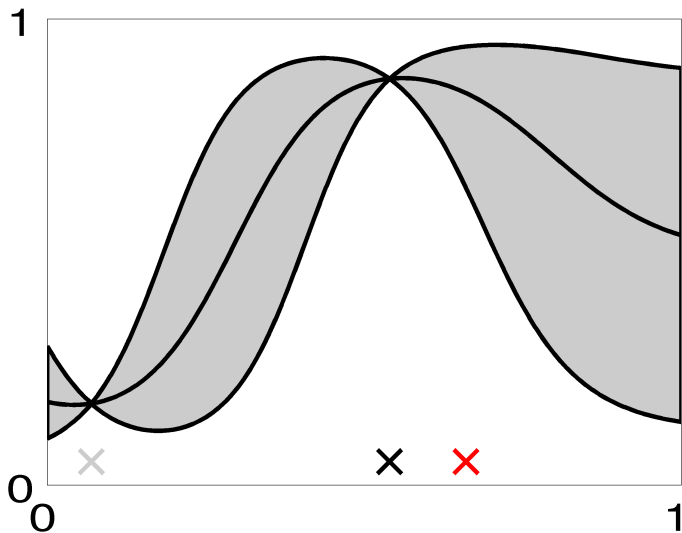
Sampling While Discovering $g(x)$



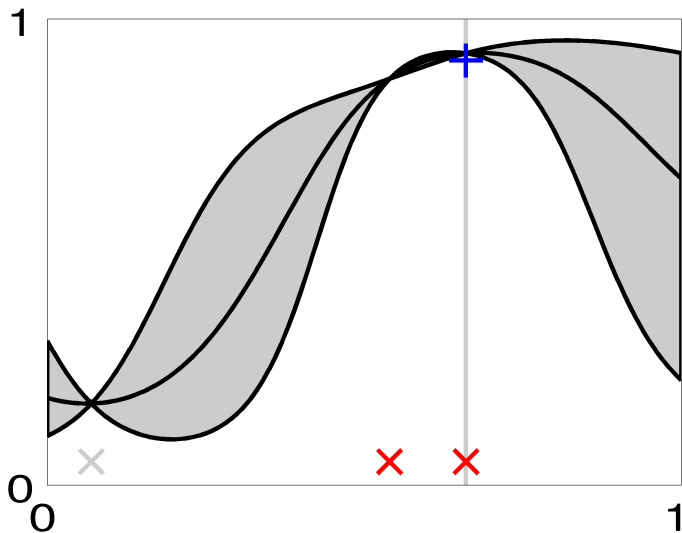
Sampling While Discovering $g(x)$



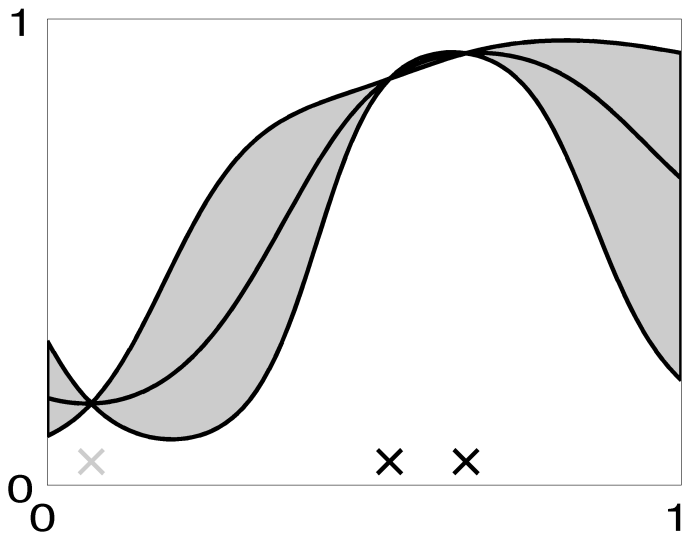
Sampling While Discovering $g(x)$



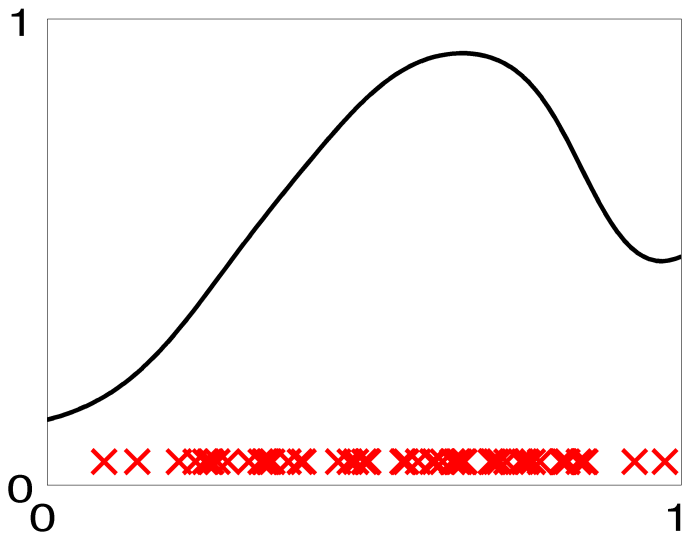
Sampling While Discovering $g(x)$



Sampling While Discovering $g(x)$



Sampling While Discovering $g(x)$



Properties of the Prior Samples

Rejection sampling is *exact*.

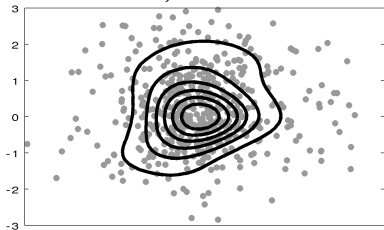
The sampling procedure is *exchangeable*.

The latent function was sampled at a finite number of locations.

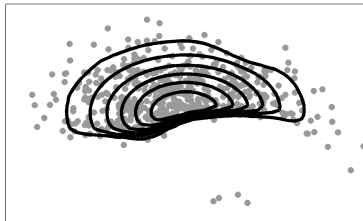
The normalization constant was not evaluated.

Effect of Hyperparameters

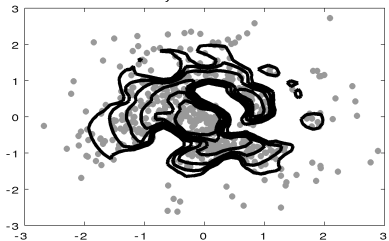
$l_s_x = 1.0, l_s_y = 1.0, \text{amp} = 1.0$



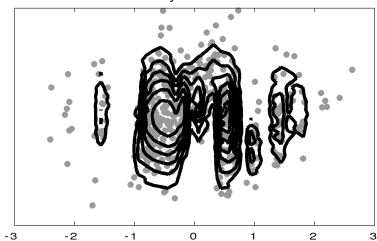
$l_s_x = 1.0, l_s_y = 1.0, \text{amp} = 10.0$



$l_s_x = 0.25, l_s_y = 0.25, \text{amp} = 5.0$



$l_s_x = 0.125, l_s_y = 2.0, \text{amp} = 5.0$



Performing Inference

We have a generative model from a Gaussian process to data. How to invert it?

$$p(\mathbf{g} \mid \{x_n\}_{n=1}^N) = \frac{p(\mathbf{g}) \mathcal{Z}_\pi[\mathbf{g}]^{-N} \prod_{n=1}^N \Phi(\mathbf{g}(x_n)) \pi(x_n)}{p(\{x_n\}_{n=1}^N)}$$

Naïve Metropolis–Hastings:

- ▶ Markov chain on the *entire function* $g(x)$
- ▶ Independent proposals: $q(\hat{\mathbf{g}} \leftarrow \mathbf{g}) = p(\hat{\mathbf{g}})$

$$a = \left(\frac{\mathcal{Z}_\pi[\mathbf{g}]}{\mathcal{Z}_\pi[\hat{\mathbf{g}}]} \right)^N \left(\prod_{n=1}^N \frac{\Phi(\hat{\mathbf{g}}(x_n))}{\Phi(\mathbf{g}(x_n))} \right)$$

Exchange Sampling

Add *child variables*:

- ▶ The proposed function $\hat{g}(x)$
- ▶ Fantasy data $\{w_n\}_{n=1}^N$ drawn from $\hat{g}(x)$.

New joint distribution:

$$\begin{aligned} p(\mathbf{g}, \{x_n\}, \hat{\mathbf{g}}, \{w_n\}) \\ = p(\mathbf{g})p(\{x_n\} | \mathbf{g})q(\hat{\mathbf{g}} \leftarrow \mathbf{g})p(\{w_n\} | \hat{\mathbf{g}}) \end{aligned}$$

Given the current \mathbf{g} and the data $\{x_n\}$:

1. Draw a proposal $\hat{\mathbf{g}}$.
2. Fantasize data from $\hat{\mathbf{g}}$.
3. **Propose swapping \mathbf{g} and $\hat{\mathbf{g}}$.**

Exchange Sampling

$$p(\mathbf{g}, \{x_n\}, \hat{\mathbf{g}}, \{w_n\}) = p(\mathbf{g})p(\hat{\mathbf{g}})(\mathcal{Z}_\pi[\mathbf{g}] \mathcal{Z}_\pi[\hat{\mathbf{g}}])^{-N} \\ \times \prod_{n=1}^N \Phi(g(x_n))\Phi(\hat{g}(w_n))\pi(x_n)\pi(w_n)$$

Acceptance ratio of the swap:

$$a = \frac{\prod_{n=1}^N \Phi(\hat{g}(x_n))\Phi(g(w_n))}{\prod_{n=1}^N \Phi(g(x_n))\Phi(\hat{g}(w_n))}$$

We can accept or reject:

- ▶ Keeping only a finite Markov chain state.
- ▶ Without the ratio of normalization constants.

Exchange Sampling

A Big Catch

For consistency, you must remember everything you find out about a $g(x)$. The Markov state expands as a result.

but...

You can throw away old data after an accept.

GPs scale badly, so acceptance rate is critical.

The Predictive Distribution

The distribution *on the data space* when the latent function is integrated out:

$$p(x | \{x_n\}) = \int d\mathbf{g} p(x | \mathbf{g})p(\mathbf{g} | \{x_n\})$$

- ▶ We sample from the posterior on $g(x)$.
- ▶ We can generate fantasies from $g(x)$.
- ▶ Generate a fantasy after each M–H step.

To generate *conditional samples*, generate fantasies from the conditional base measure.

Hyperparameter Inference

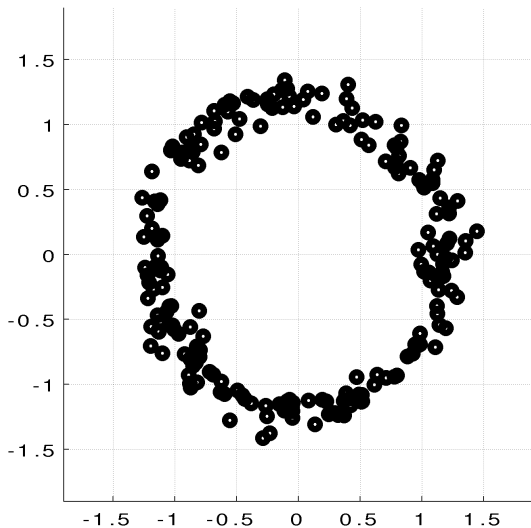
Hierarchical Inference:

- ▶ Gaussian process hyperparameters
- ▶ Base measure parameters

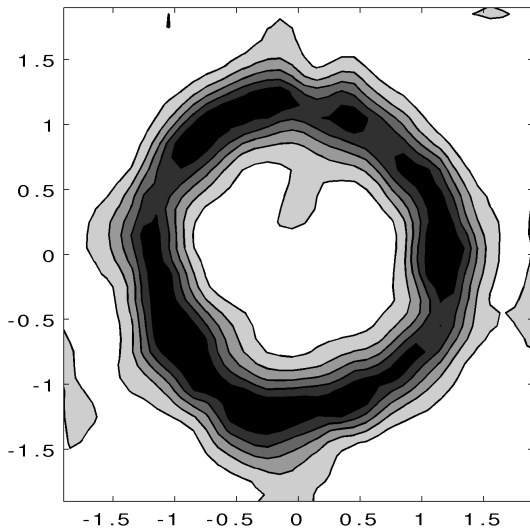
Roughly: “To what degree should similar data have similar probabilities?”

- ▶ Add the parameters to the MCMC state.
- ▶ Joint proposals of parameters and functions.
- ▶ Accept or reject both together.

Example Data



Example Data



Summary

“Similar data should have similar probabilities”

- ▶ GP-based prior on density functions.
- ▶ MCMC scheme for density inference.
- ▶ Samples from the predictive distribution.
- ▶ Infer GP hyperparameters.

This work funded by the Gates Cambridge Trust.