

Nonparametric Bayesian Density Modeling with Gaussian Processes

Ryan Prescott Adams

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

9 July 2008

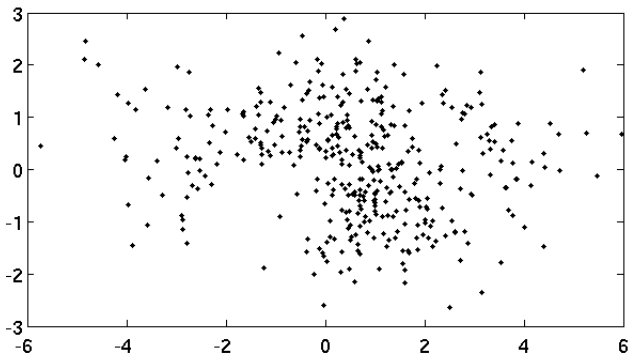
Joint work with Iain Murray
and David MacKay



The Density Modeling Problem

The setup:

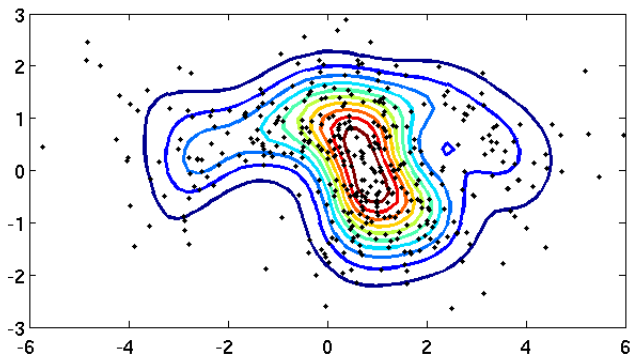
- ▶ Data $\{x_n\}_{n=1}^N$ from an unknown density $f(x)$
- ▶ Prior beliefs about $f(x)$



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

Classic Approach:

- ▶ Kernel density estimation (Parzen windows)

Some Bayesian Approaches:

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

This Talk

Using a Gaussian process to specify the prior on the probability density function.

Why Gaussian Processes?

The GP is a nonparametric prior on functions.

GP Components:

- ▶ Input space \mathcal{X} (e.g. \mathbb{R}^d)
- ▶ Output space $\mathcal{Y} = \mathbb{R}$
- ▶ Covariance function
 $K(x, x'; \theta) : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$.
- ▶ Mean function $m(x; \theta) : \mathcal{X} \rightarrow \mathbb{R}$.

The predictive distribution and the marginal likelihood are easy.

GPs for probability density functions

Not a new idea...

- ▶ Leonard (1978)
- ▶ Lenk (1988, 1991)
- ▶ Tokdar and Ghosh (2007)

The Logistic Gaussian Process

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

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

Gaussian Process Density Sampler

Our Idea

A *generative* GP-based prior on densities.

We can then perform inference **without making finite-dimensional approximations**, as has been necessary in the logistic Gaussian process.

The GPDS Prior

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

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

- ▶ $g(x)$ has a GP prior.
- ▶ $\pi(x)$ is a known “base density.”
- ▶ $\Phi(x)$ is nonnegative and bounded. e.g. logistic $\Phi(z) = (1 + \exp(-z))^{-1}$

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

Sampling With Known $g(x)$

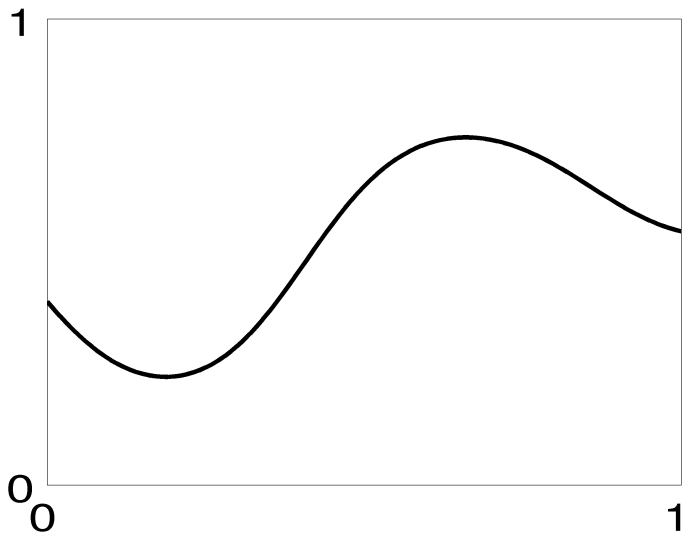
$$f(x) = \frac{1}{Z_{\pi}[g]} \Phi(g(x)) \pi(x)$$

What if we knew $g(x)$?

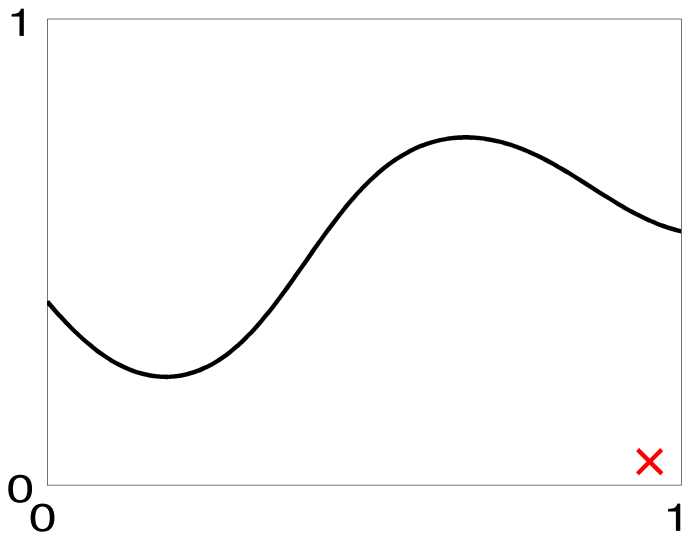
Rejection sampling:

1. Draw \tilde{x} from $\pi(x)$.
2. Draw r from UNIFORM(0, 1)
3. Accept if $r < \Phi(g(\tilde{x}))$
4. Goto 1

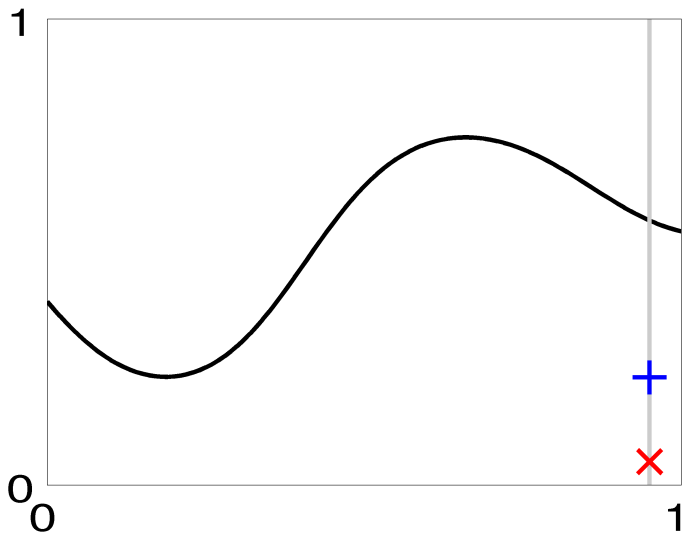
Sampling With Known $g(x)$



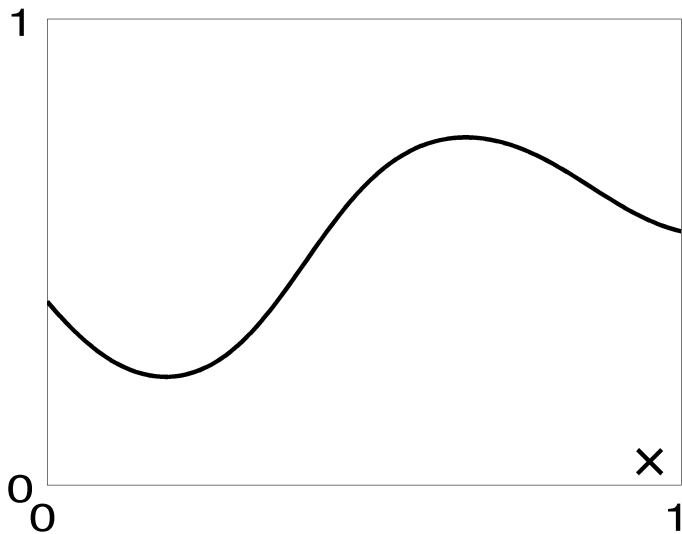
Sampling With Known $g(x)$



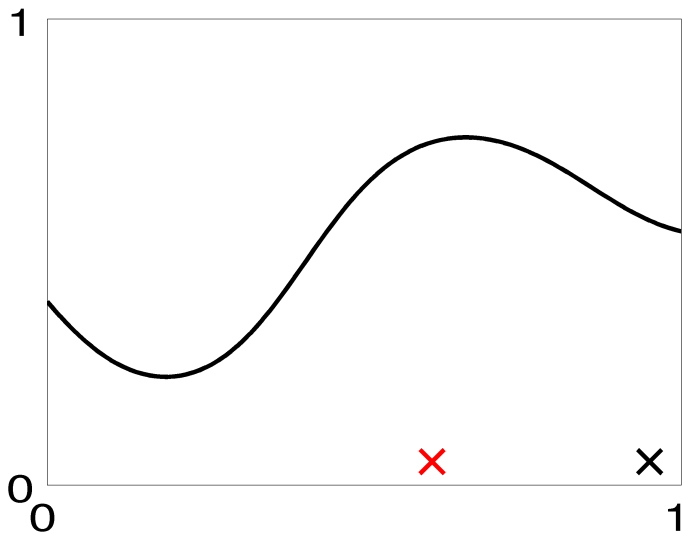
Sampling With Known $g(x)$



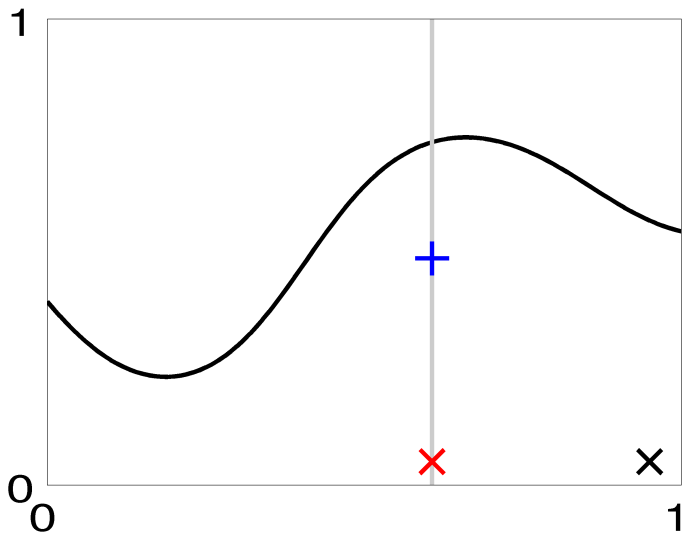
Sampling With Known $g(x)$



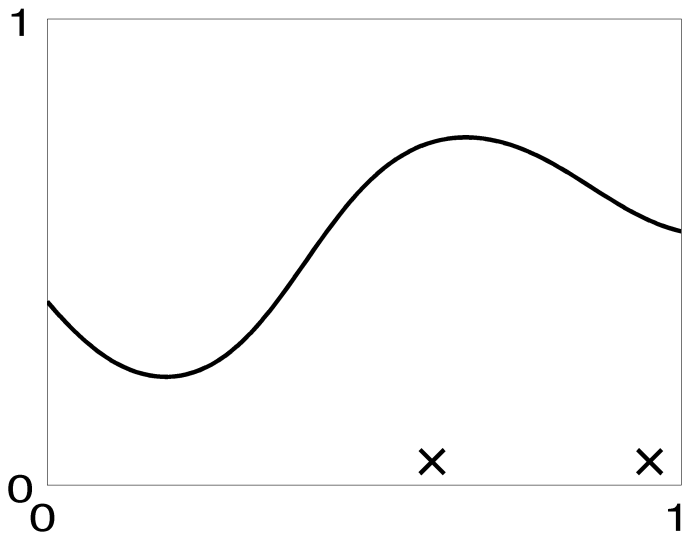
Sampling With Known $g(x)$



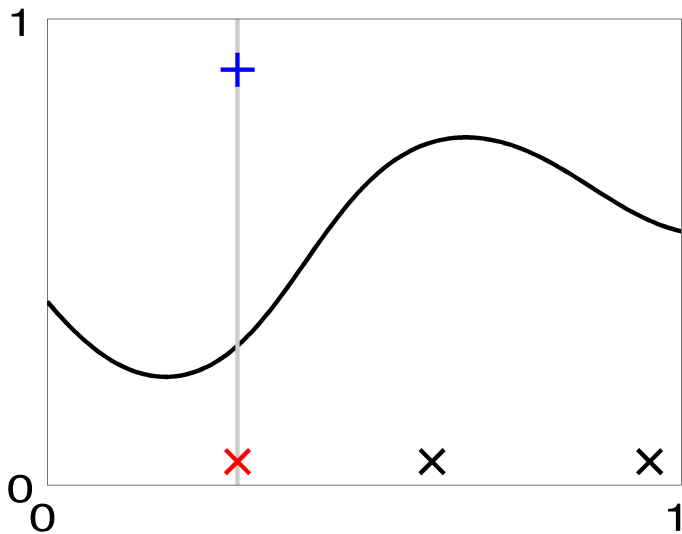
Sampling With Known $g(x)$



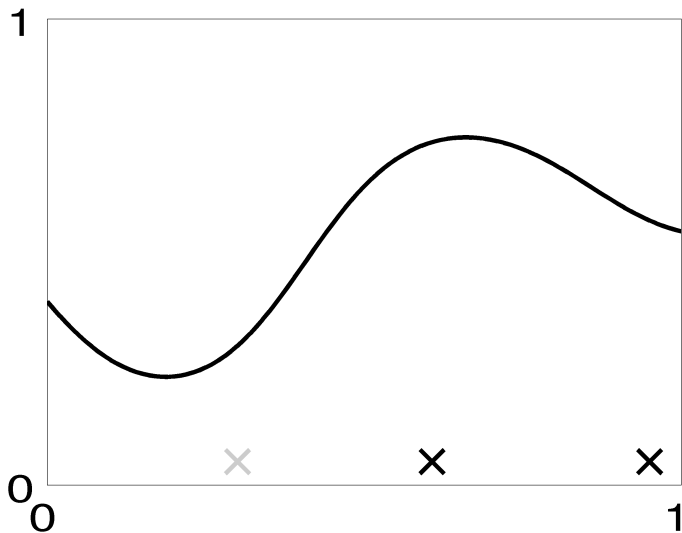
Sampling With Known $g(x)$



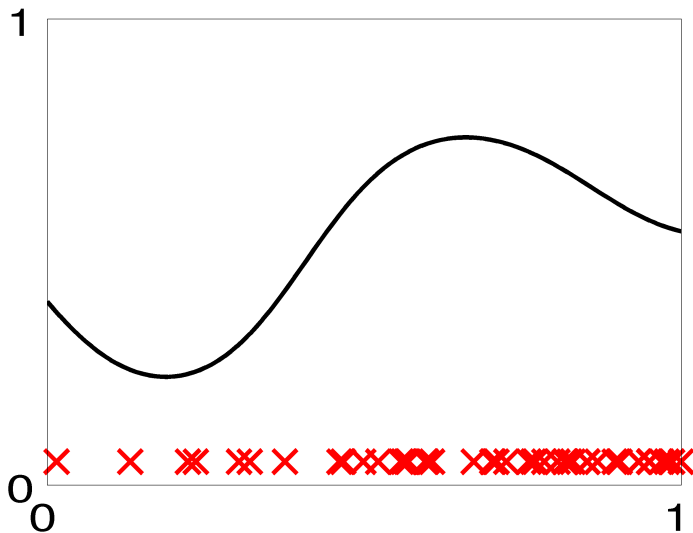
Sampling With Known $g(x)$



Sampling With Known $g(x)$



Sampling With Known $g(x)$



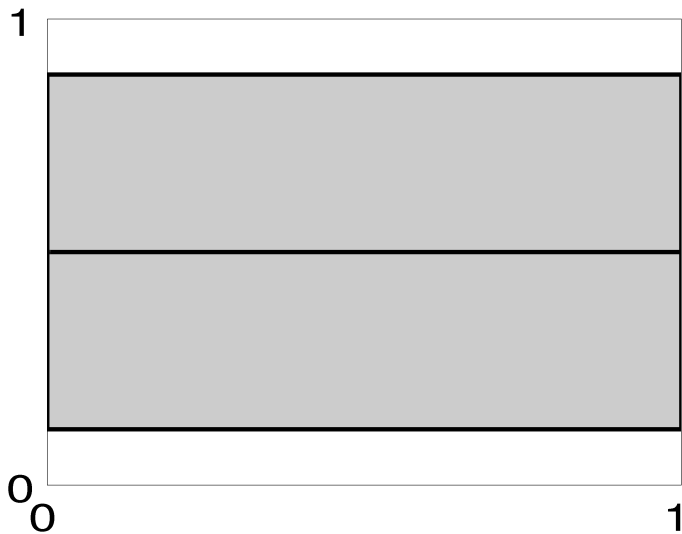
Sampling While Discovering $g(x)$

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

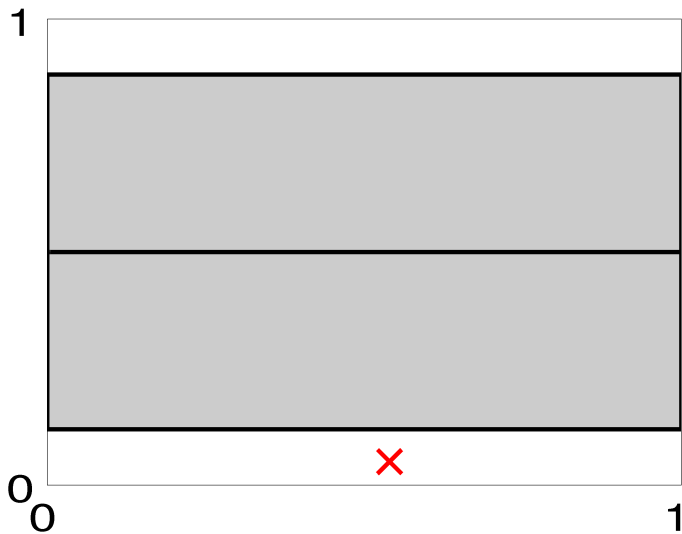
Rejection sampling:

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

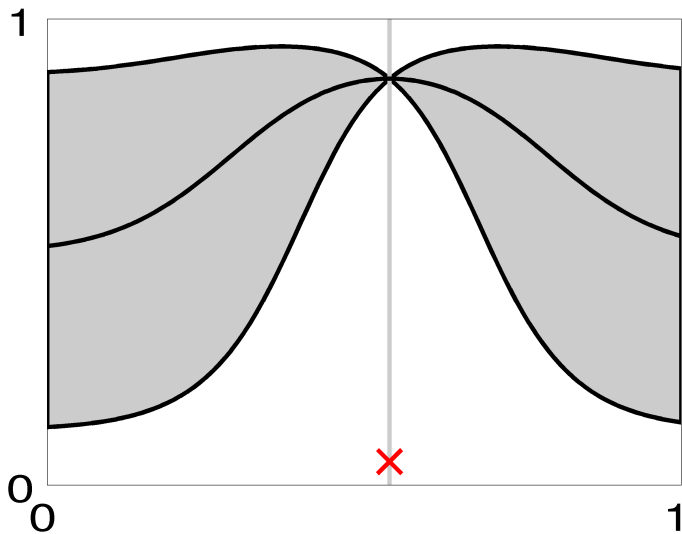
Sampling While Discovering $g(x)$



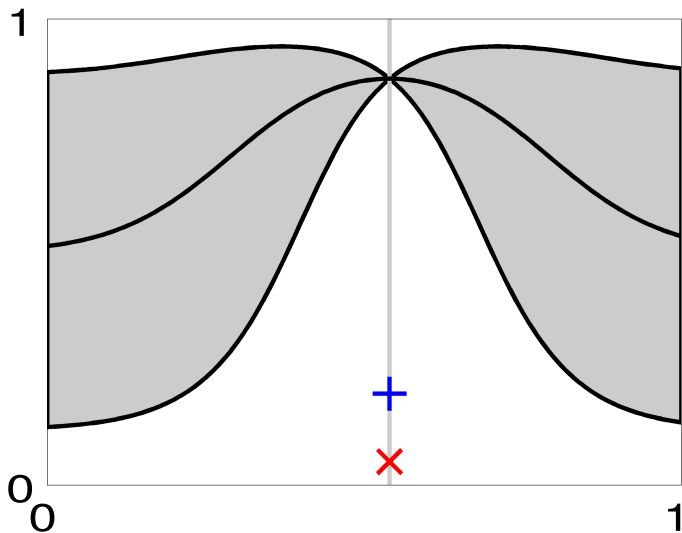
Sampling While Discovering $g(x)$



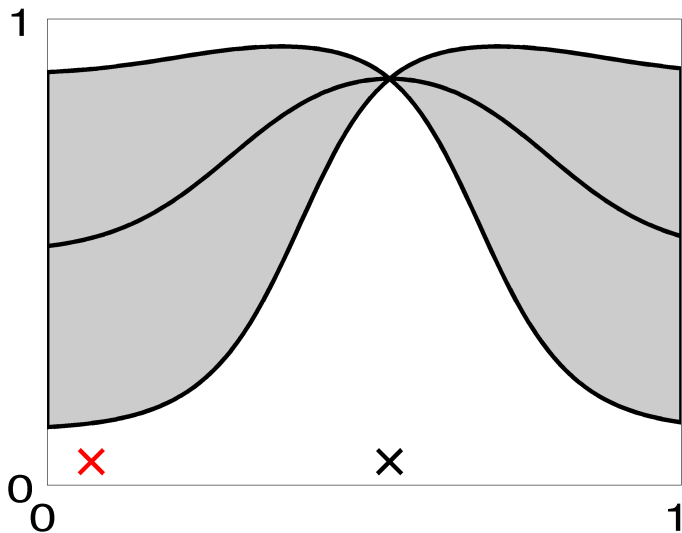
Sampling While Discovering $g(x)$



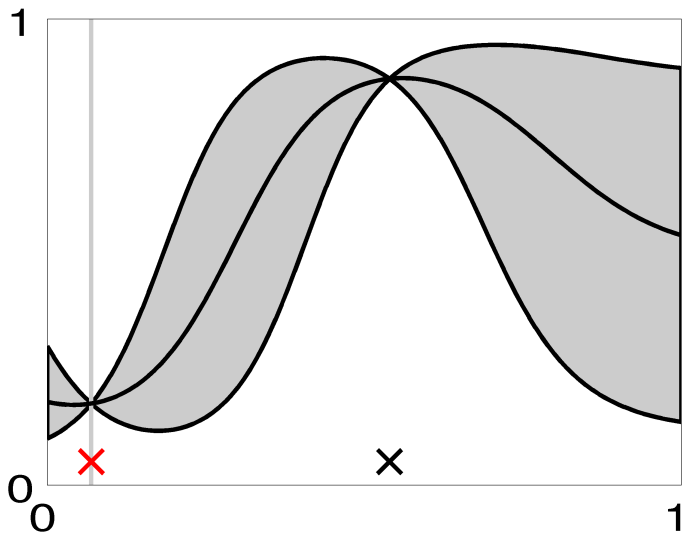
Sampling While Discovering $g(x)$



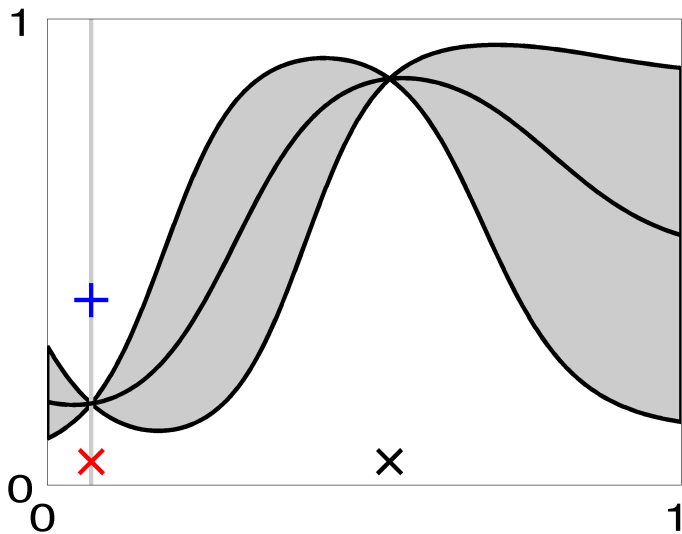
Sampling While Discovering $g(x)$



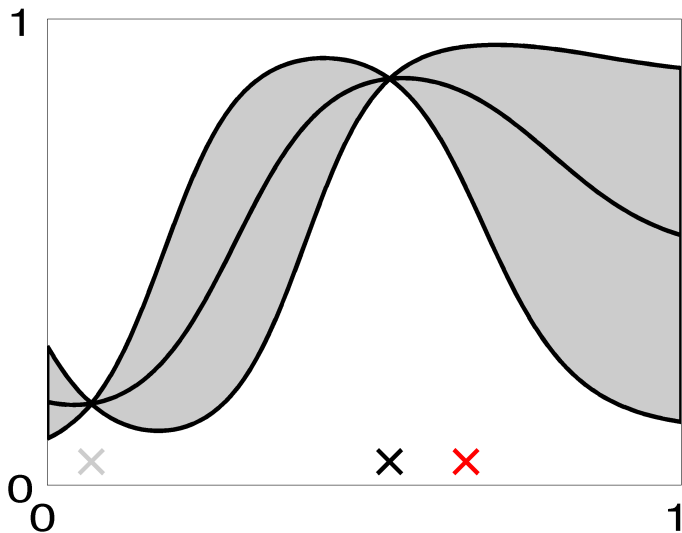
Sampling While Discovering $g(x)$



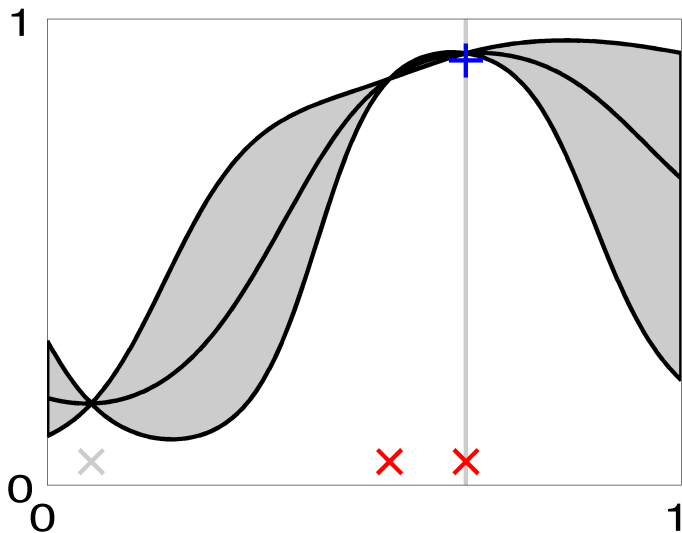
Sampling While Discovering $g(x)$



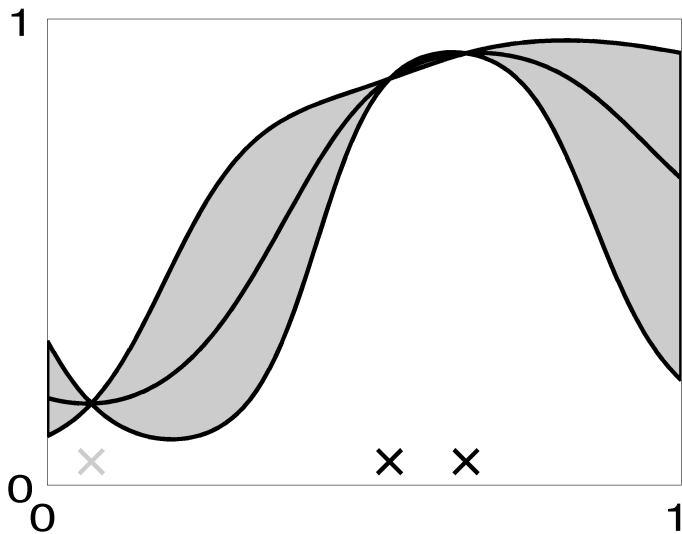
Sampling While Discovering $g(x)$



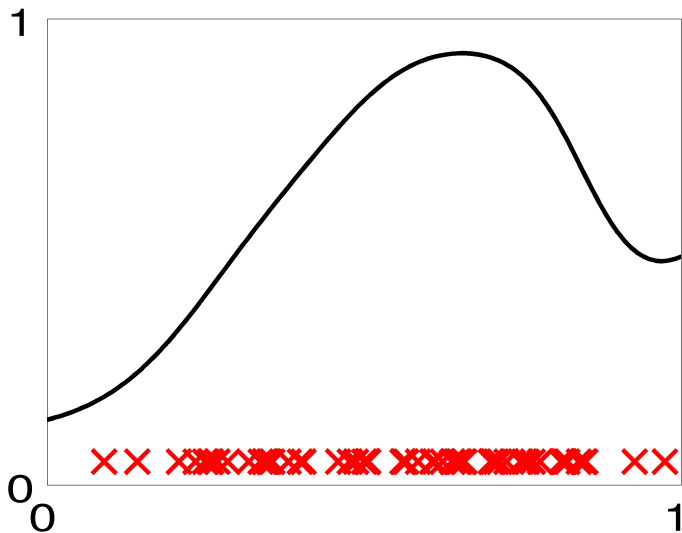
Sampling While Discovering $g(x)$



Sampling While Discovering $g(x)$



Sampling While Discovering $g(x)$



Properties of the GPDS

Rejection sampling is *exact*.

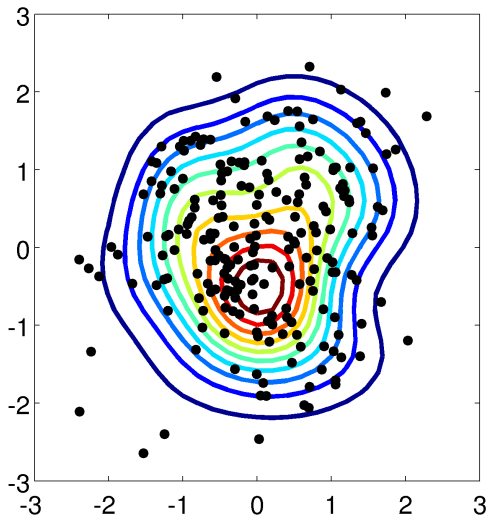
The sampling procedure is *exchangeable*.

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

The normalization constant was not evaluated.

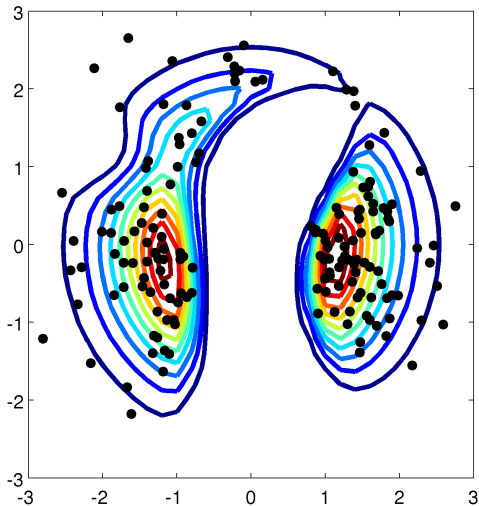
Hyperparameter Effects

$$l_x = 1, l_y = 1, \alpha = 1$$



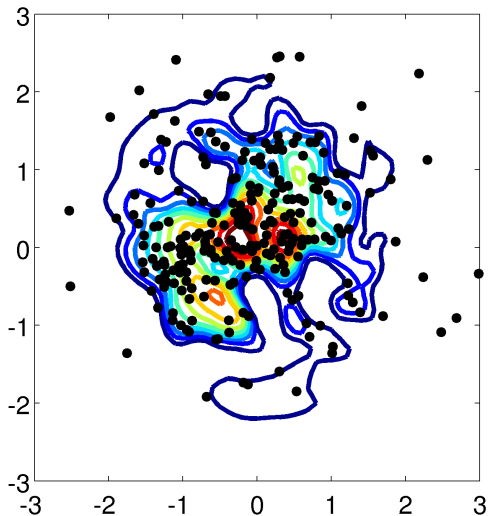
Hyperparameter Effects

$$l_x = 1, l_y = 1, \alpha = 5$$



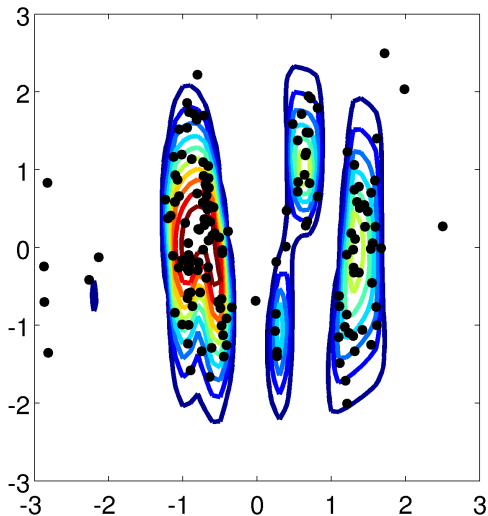
Hyperparameter Effects

$$l_x = 0.25, l_y = 0.25, \alpha = 2$$



Hyperparameter Effects

$$\ell_x = 0.25, \ell_y = 2, \alpha = 5$$



Inference with the GPDS Prior

Up to now we have discussed the **prior**.

We now look at **inference**:

Given data $\mathcal{D} = \{x_n\}_{n=1}^N$, find $p(g(x) | \mathcal{D})$.

We use Markov chain Monte Carlo (MCMC), specifically **Metropolis–Hastings**.

Inference is Difficult

Recall the likelihood:

$$p(x_n | g(x)) = \frac{1}{\mathcal{Z}_\pi[\mathbf{g}]} \Phi(g(x_n)) \pi(x_n)$$

$\mathcal{Z}_\pi[\mathbf{g}]$ is intractable, like the **partition function** in undirected graphical models.

Acceptance ratio of naïve Metropolis–Hastings:

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

Saved by the Generative Procedure

- ▶ We can get **exact samples** for a given $g(x)$.
- ▶ We borrow methods for inference UGMs:
CFTP in Ising and Potts models.

Two MCMC Methods:

- ▶ **Exchange Sampling**
(Murray, Ghahramani & MacKay, UAI 2006)
- ▶ **Latent History Inference**
(Murray 2007)

Exchange Sampling

ES is an auxiliary variable method.

The main idea is:

- ▶ Propose a new function $\hat{g}(x)$ as in M–H.
- ▶ Draw **fantasy data** from the proposal $\hat{g}(x)$.
- ▶ Propose swapping $g(x)$ for $\hat{g}(x)$.

Likelihood terms appear on the top and bottom of the acceptance ratio, so the normalization constants cancel out.

Latent History Inference

Using the GPDS prior to model data is saying the data are the result of the generative procedure.

Infer the history of the generative procedure:

- ▶ Number of rejections
- ▶ Location of rejections
- ▶ Values of the latent function

The generative procedure didn't require the normalisation constant, so inference doesn't either.

Other Inferences

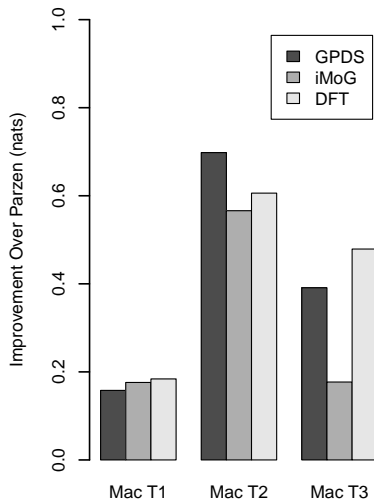
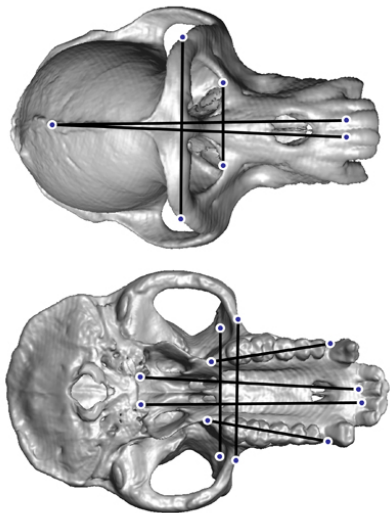
Can infer the GP hyperparameters.

Can also infer the parameters of the base density $\pi(x)$.

Easy to draw samples from the predictive distribution: after every Metropolis–Hastings step, just run the GPDS forward.

Comparison on Macaque Skull Data

10 linear distances, 200 training, 28 test, 3 trials



Computational Concerns

Gaussian processes are expensive: $O(N^3)$

Rejection samplers are inefficient in high dimensions

- ▶ Exchange sampling: expensive to fantasize
- ▶ Latent histories: many latent rejections

Summary

- ▶ GP-based prior on probability densities
- ▶ Can generate exact data from the prior
- ▶ Two MCMC inference methods
- ▶ Hyperparameter inference
- ▶ Predictive samples
- ▶ No approximation of partition function
- ▶ Generally expensive

Thank you!