

# The Gaussian Process Density Sampler

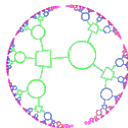
Ryan Prescott Adams

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

14 April 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  $f(x)$
- ▶ 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)

# Short Gaussian Process Introduction

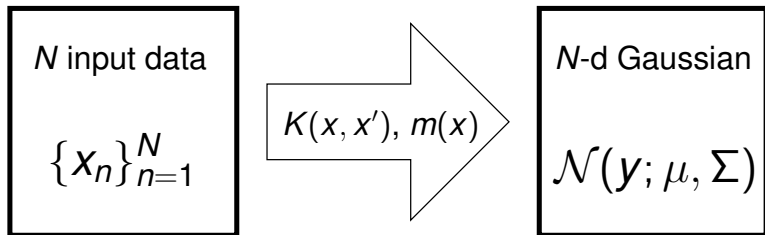
## The Big GP Idea:

Specify a distribution on functions, without choosing an explicit set of basis 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  $m(x; \theta) : \mathcal{X} \rightarrow \mathbb{R}$ .

# Short Gaussian Process Introduction



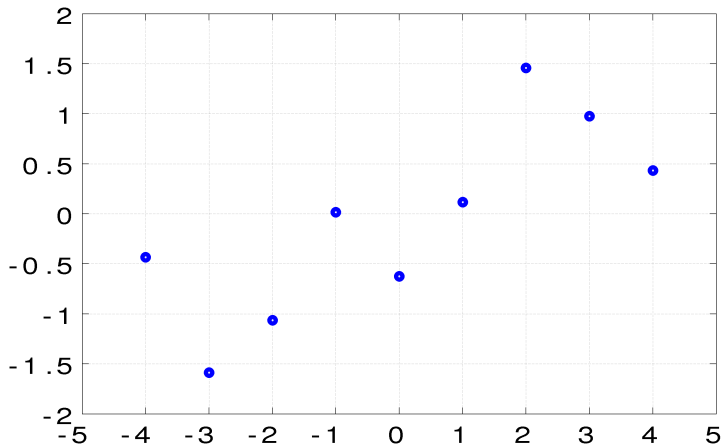
## Typical Covariance Function

“Squared Exponential”:

$$K(x, x') = \theta_1 \exp \left\{ -\frac{1}{2} \sum_i \frac{|x_i - x'_i|^2}{\ell_i^2} \right\} + \theta_2 \delta_{x, x'}$$

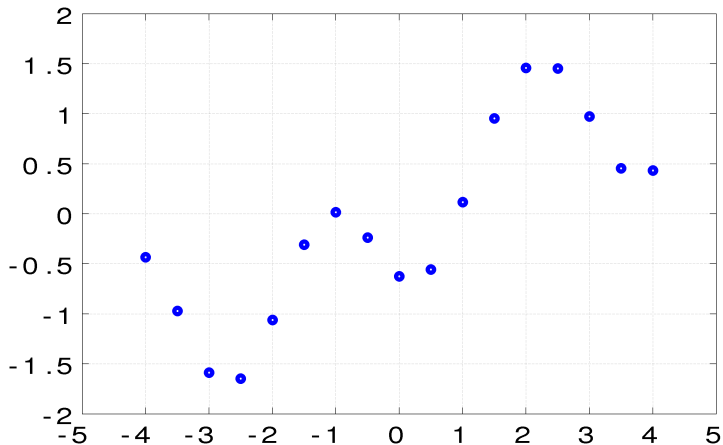
# Short Gaussian Process Introduction

“Nearby inputs have covarying outputs”



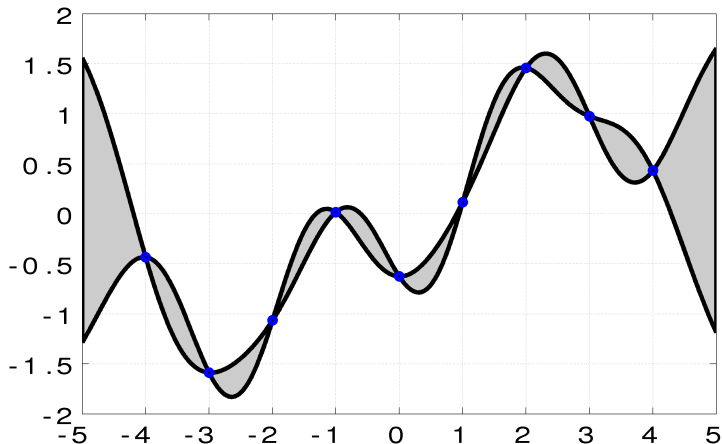
# Short Gaussian Process Introduction

“Nearby inputs have covarying outputs”



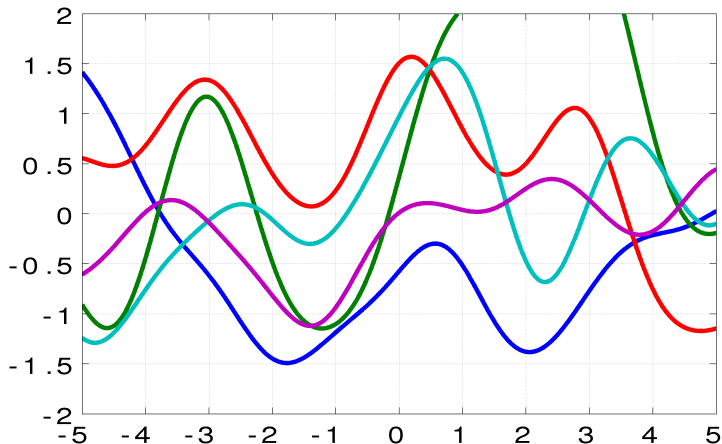
# Short Gaussian Process Introduction

“Nearby inputs have covarying outputs”



# Short Gaussian Process Introduction

“Nearby inputs have covarying outputs”



# What About the Rest of the Space?

How can you ignore the (uncountably infinite) unknown values?

## Gaussian to the Rescue

- ▶ Joint Gaussian  $P(y_1, \dots, y_N) = \mathcal{N}(\mathbf{K})$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} k_{1,1} & k_{1,2} & k_{1,3} & k_{1,4} & k_{1,5} \\ k_{2,1} & k_{2,2} & k_{2,3} & k_{2,4} & k_{2,5} \\ k_{3,1} & k_{3,2} & k_{3,3} & k_{3,4} & k_{3,5} \\ k_{4,1} & k_{4,2} & k_{4,3} & k_{4,4} & k_{4,5} \\ k_{5,1} & k_{5,2} & k_{5,3} & k_{5,4} & k_{5,5} \end{bmatrix} \right)$$

# What About the Rest of the Space?

How can you ignore the (uncountably infinite) unknown values?

## Gaussian to the Rescue

- ▶ Joint Gaussian  $P(y_1, \dots, y_N) = \mathcal{N}(\mathbf{K})$
- ▶ Marginal covariance is the submatrix.

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} k_{1,1} & k_{1,2} & k_{1,3} & k_{1,4} & k_{1,5} \\ k_{2,1} & k_{2,2} & k_{2,3} & k_{2,4} & k_{2,5} \\ k_{3,1} & k_{3,2} & k_{3,3} & k_{3,4} & k_{3,5} \\ k_{4,1} & k_{4,2} & k_{4,3} & k_{4,4} & k_{4,5} \\ k_{5,1} & k_{5,2} & k_{5,3} & k_{5,4} & k_{5,5} \end{bmatrix} \right)$$

# What About the Rest of the Space?

How can you ignore the (uncountably infinite) unknown values?

## Gaussian to the Rescue

- ▶ Joint Gaussian  $P(y_1, \dots, y_N) = \mathcal{N}(\mathbf{K})$
- ▶ Marginal covariance is the submatrix.
- ▶ Conditional is also Gaussian.

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} k_{1,1} & k_{1,2} & k_{1,3} & k_{1,4} & k_{1,5} \\ k_{2,1} & k_{2,2} & k_{2,3} & k_{2,4} & k_{2,5} \\ k_{3,1} & k_{3,2} & k_{3,3} & k_{3,4} & k_{3,5} \\ k_{4,1} & k_{4,2} & k_{4,3} & k_{4,4} & k_{4,5} \\ k_{5,1} & k_{5,2} & k_{5,3} & k_{5,4} & k_{5,5} \end{bmatrix} \right)$$

# Mechanics of GP Regression

The marginal and conditional properties give us the main GP tools:

## Predictive Distribution

Just the conditional  $p(y^* | \mathbf{x}^*, \{\mathbf{x}_n, y_n\}^N)$ :

$$\mu^* = \mathbf{k}^T \mathbf{K}_N^{-1} \mathbf{y}_N$$

$$v^* = \kappa - \mathbf{k}^T \mathbf{K}_N^{-1} \mathbf{k}$$

## Log Evidence

Just the marginal:

$$\ln p(\{\mathbf{x}_n, y_n\}^N) = -\frac{N}{2} \ln 2\pi - \frac{1}{2} \ln |\mathbf{K}_N| - \frac{1}{2} \mathbf{y}_N^T \mathbf{K}_N^{-1} \mathbf{y}_N$$

# Back to Density Modeling

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

## **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 With Known $g(x)$

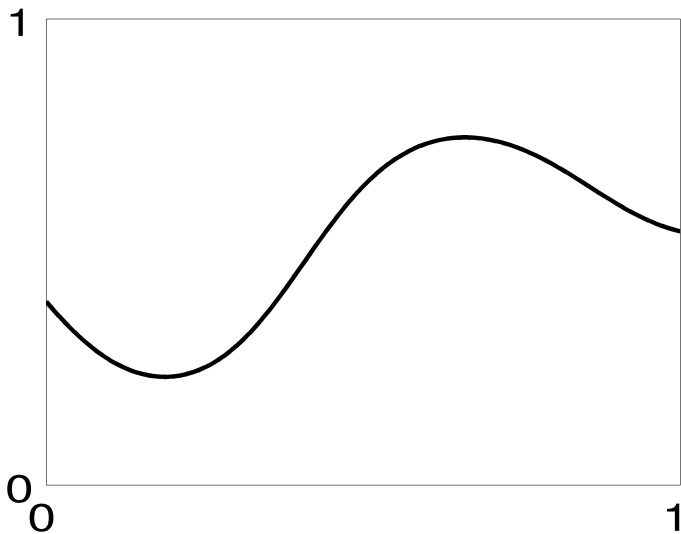
$$f(x) = \frac{1}{Z_{\pi}[g]} \Phi(g(x)) \pi(x)$$

## What if we knew $g(x)$ ?

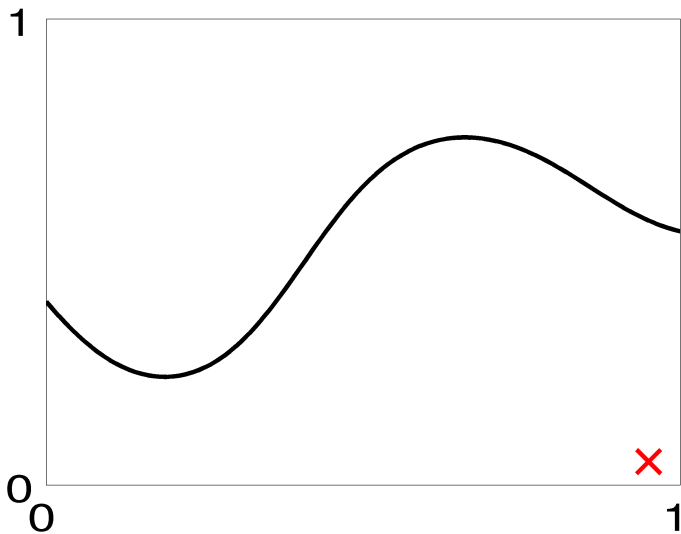
Rejection sampling:

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

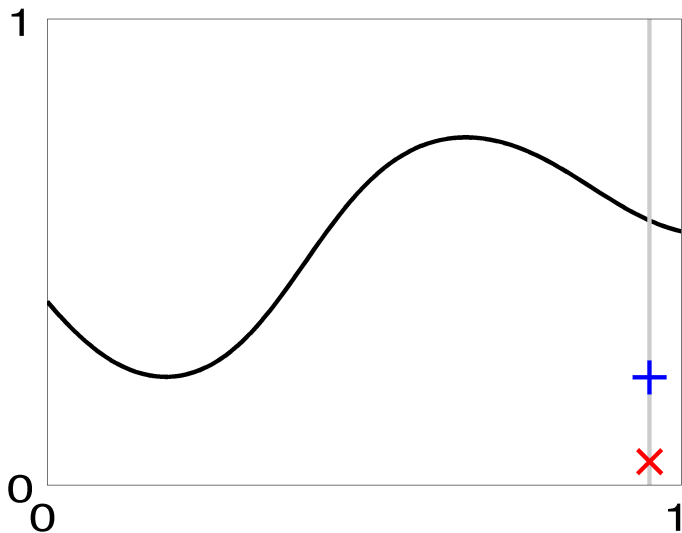
# Sampling With Known $g(x)$



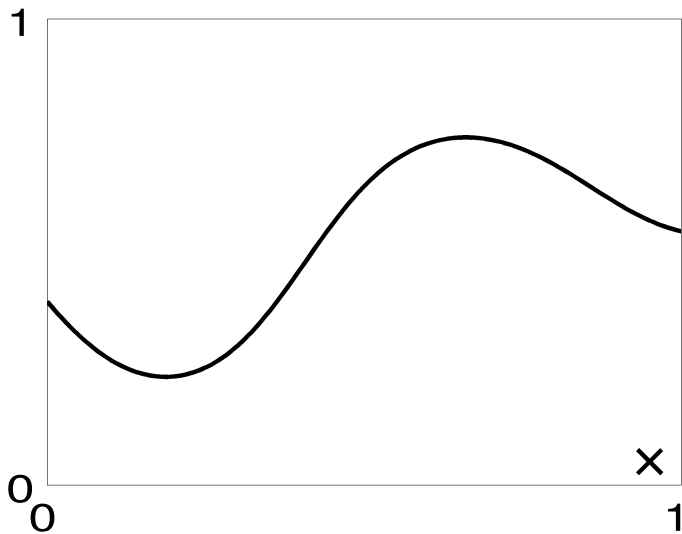
# Sampling With Known $g(x)$



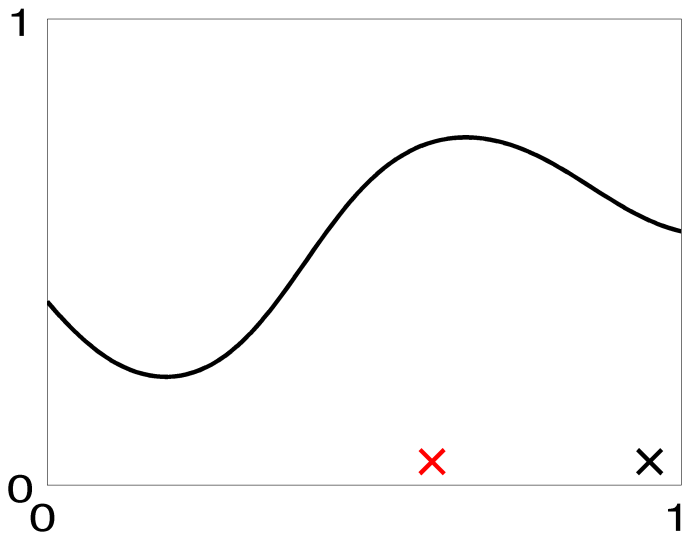
# Sampling With Known $g(x)$



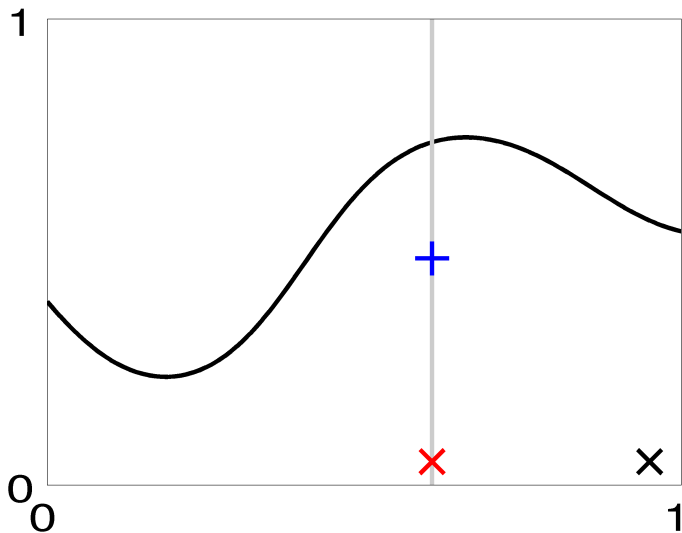
# Sampling With Known $g(x)$



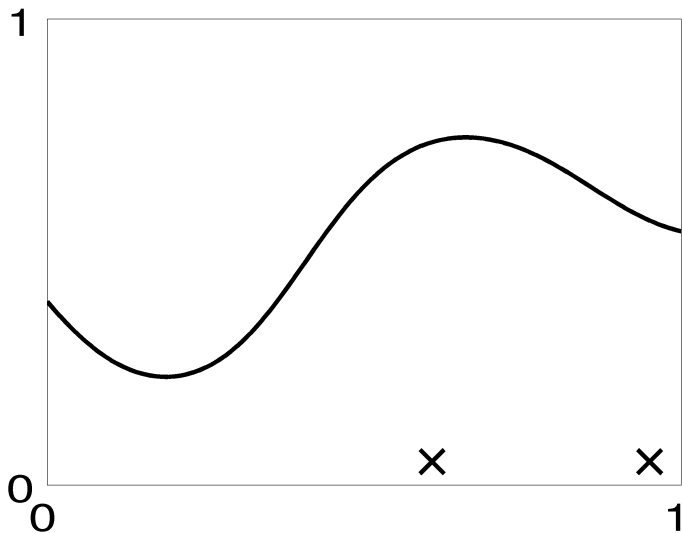
# Sampling With Known $g(x)$



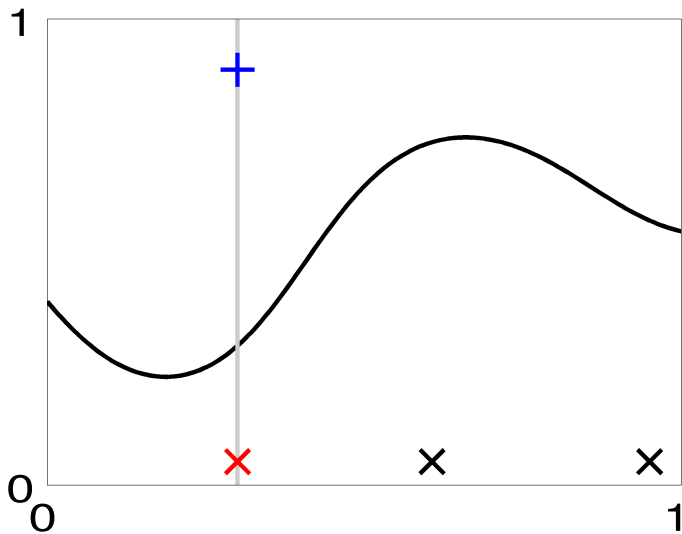
# Sampling With Known $g(x)$



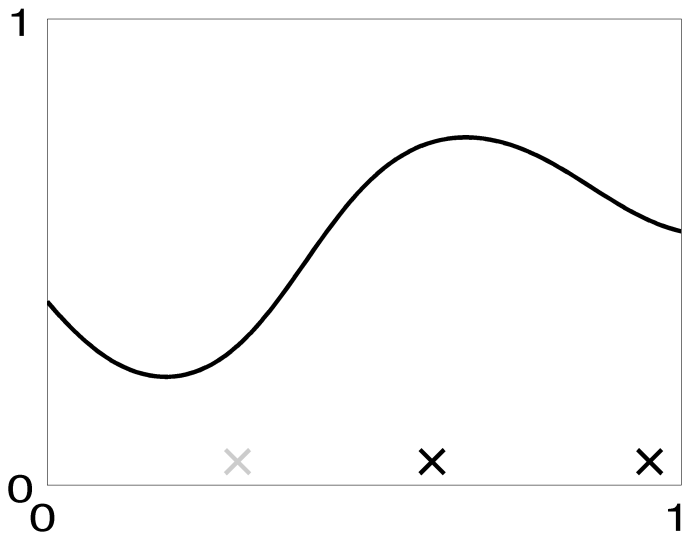
# Sampling With Known $g(x)$



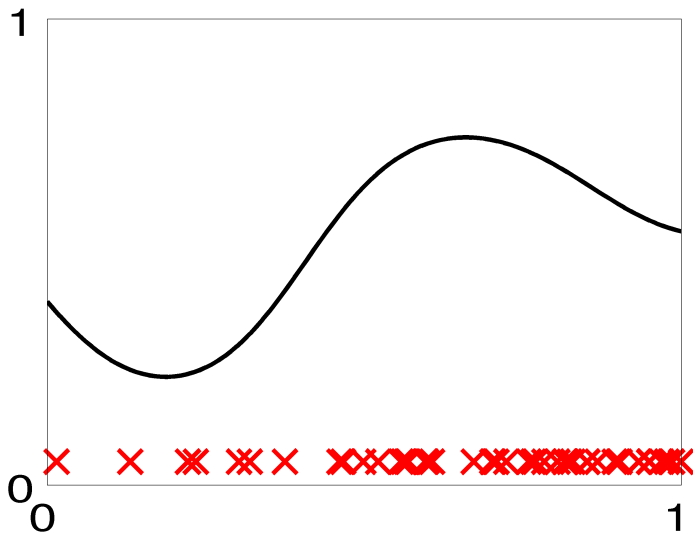
# Sampling With Known $g(x)$



# Sampling With Known $g(x)$



# Sampling With Known $g(x)$



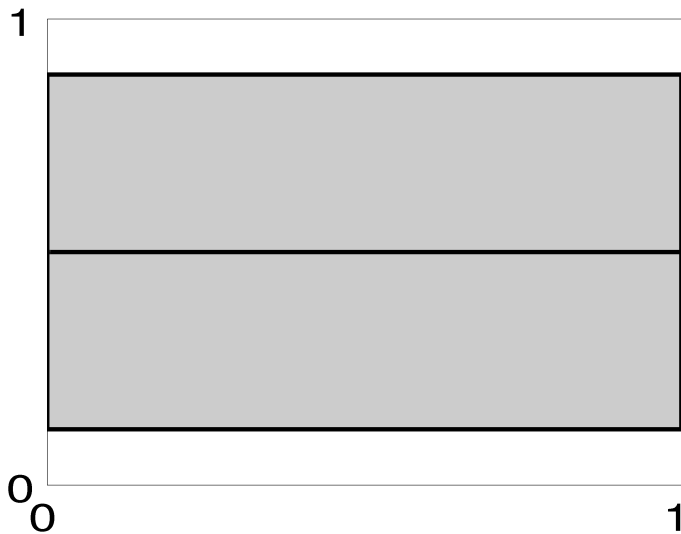
# Sampling While Discovering $g(x)$

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

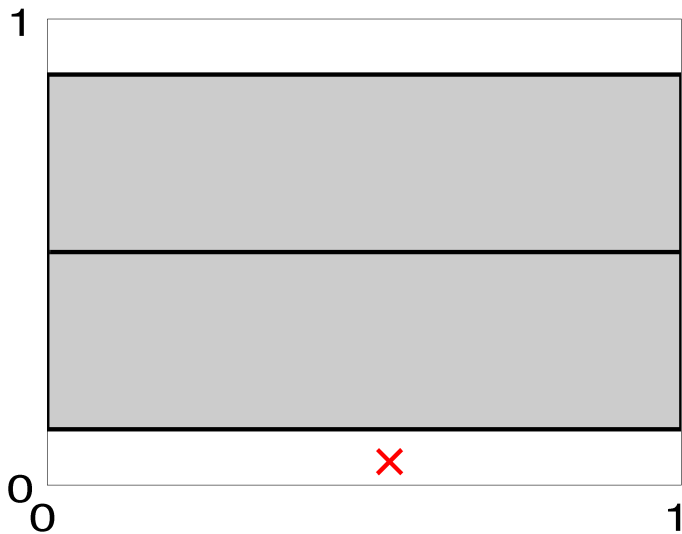
## Rejection sampling:

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

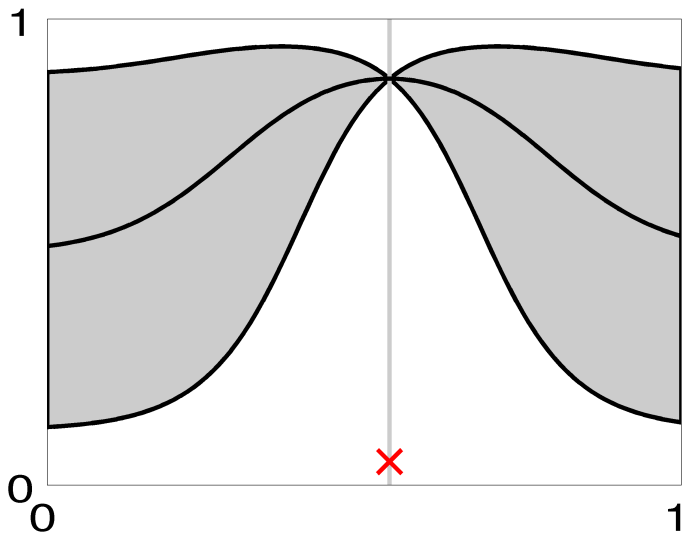
# Sampling While Discovering $g(x)$



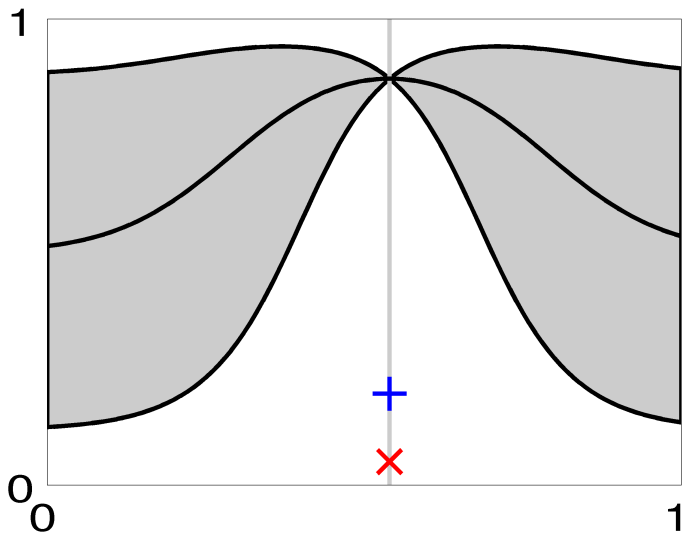
# Sampling While Discovering $g(x)$



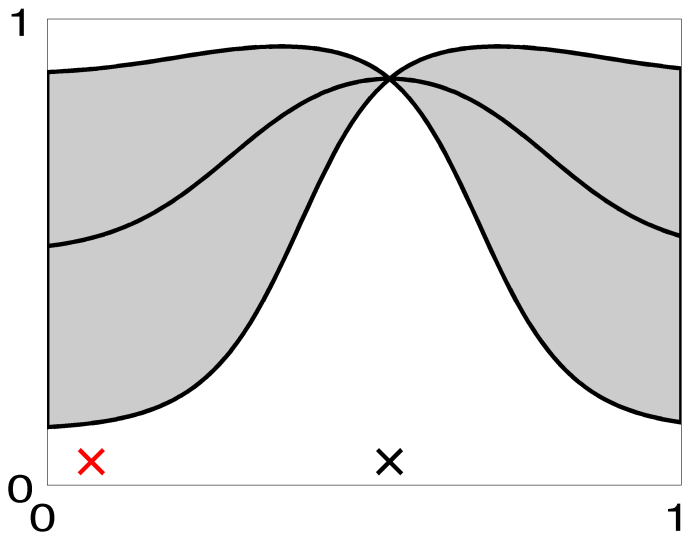
# Sampling While Discovering $g(x)$



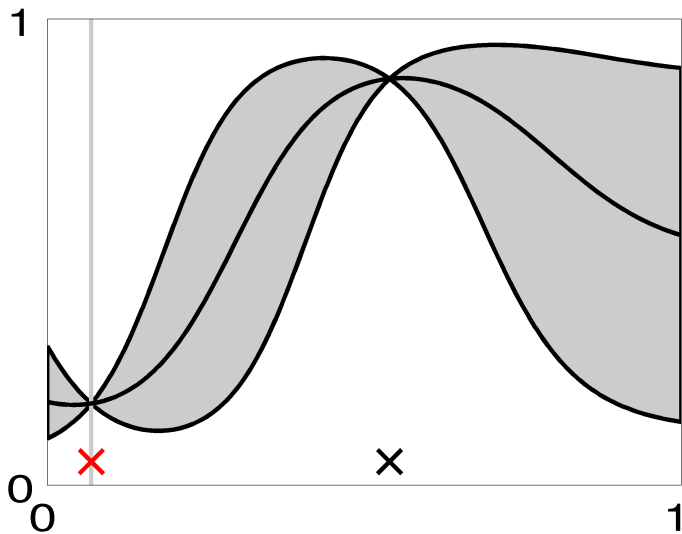
# Sampling While Discovering $g(x)$



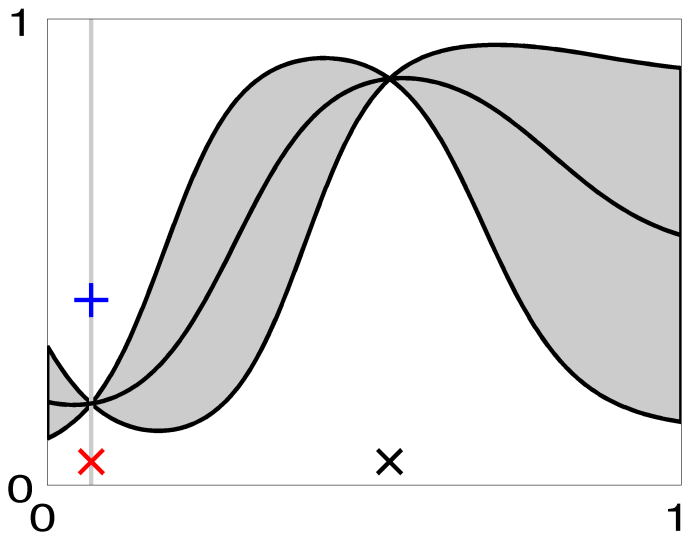
# Sampling While Discovering $g(x)$



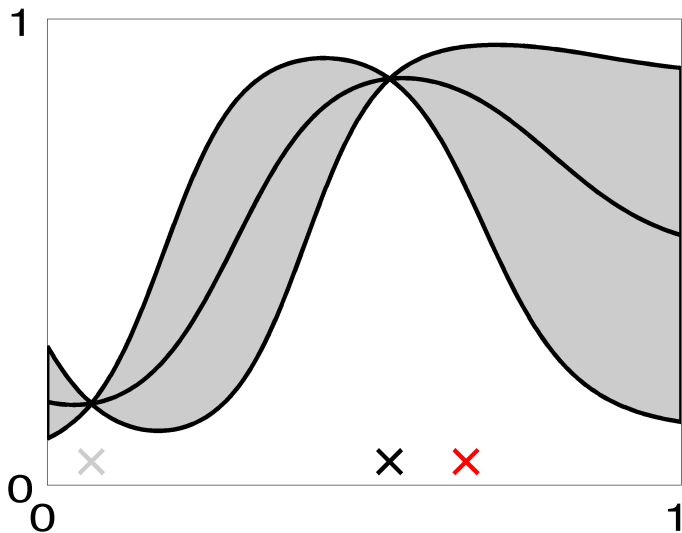
# Sampling While Discovering $g(x)$



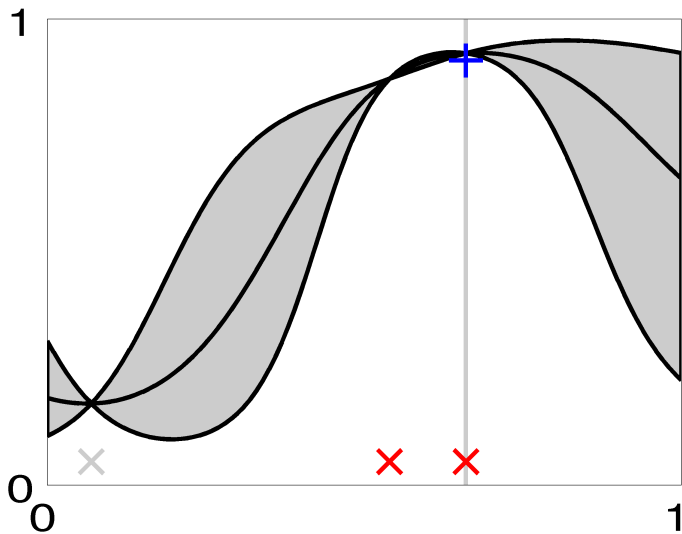
# Sampling While Discovering $g(x)$



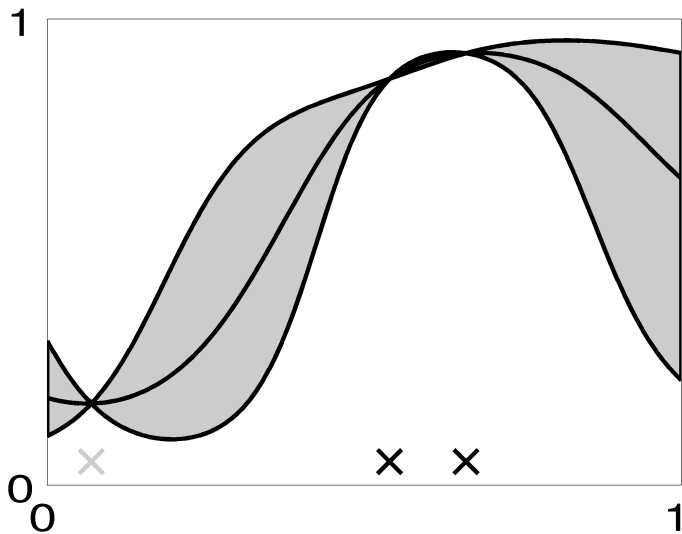
# Sampling While Discovering $g(x)$



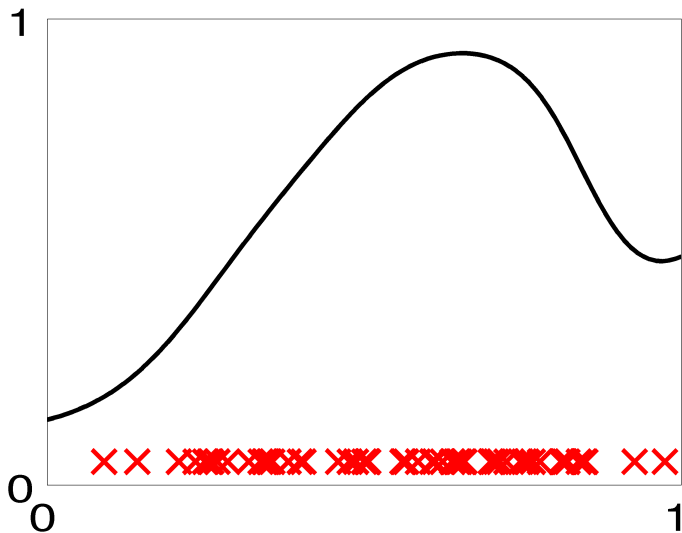
# Sampling While Discovering $g(x)$



# Sampling While Discovering $g(x)$



# Sampling While Discovering $g(x)$



# Properties of the Prior Samples

Rejection sampling is *exact*.

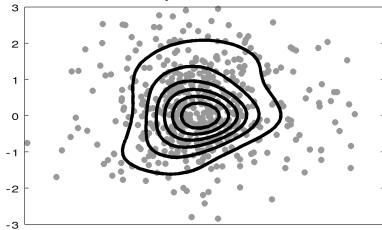
The sampling procedure is *exchangeable*.

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

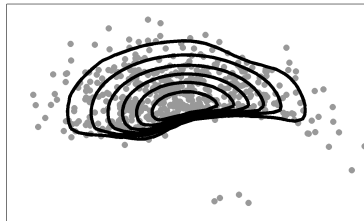
The normalization constant was not evaluated.

# Effect of Hyperparameters

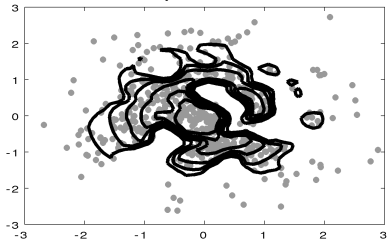
$l_s_x = 1.0, l_s_y = 1.0, \text{amp} = 1.0$



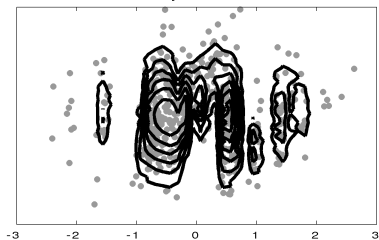
$l_s_x = 1.0, l_s_y = 1.0, \text{amp} = 10.0$



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



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



# One-Slide Intro to Bayesian MCMC

$$p(\theta | \mathcal{D}) = \frac{p(\mathcal{D} | \theta)p(\theta)}{\int d\theta p(\mathcal{D} | \theta)p(\theta)}$$

Rarely analytic in interesting problems!  
We generate samples instead.

## Markov chain Monte Carlo

- ▶ Most common: Metropolis–Hastings
- ▶ Consider jumping to a new state, and make a binary decision on whether or not to do it.
- ▶ The **acceptance ratio** makes sure that this binary decision results in samples from  $p(\theta | \mathcal{D})$ .

# 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:  $q(\hat{\mathbf{g}} \leftarrow \mathbf{g}) = p(\hat{\mathbf{g}})$

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

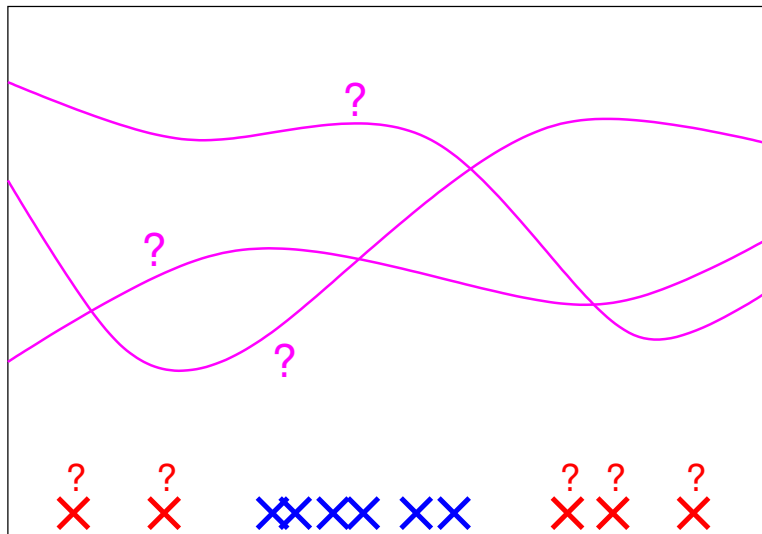
# Modeling the Latent History

Perform MCMC-based inference on the entire *generative process* that resulted in the data.

## Things we don't know:

- ▶ The number of rejections, denoted  $M$ .
- ▶ The locations of the rejections,  $\{x_m\}$ .
- ▶ The values of the latent function  $g(x)$ :  $\{g_m\}$  and  $\{g_n\}$ .

# Modeling the Latent History



# Two Metropolis–Hastings Steps

## 1) Modify the Number of Rejections

- ▶ Draw  $\hat{M} \sim q(\hat{M} \leftarrow M)$ .
- ▶ Propose adding or removing rejections.
- ▶ Condition on all of the current history.

## 2) Modify Everything Else

- ▶ Fix the number of rejections.
- ▶ Propose new rejection locations.
- ▶ Propose new function values.

# Adding Rejections

If  $\hat{M} > M$

- ▶ Draw  $\hat{M} - M$  independently from  $\pi(x)$ .
- ▶ Draw their function values from the GP, conditioned on the latent history.
- ▶ Propose moving these points to random locations before the last data acceptance.

Acceptance Ratio of New Rejections:

$$a = \frac{q(M \leftarrow \hat{M})M!(\hat{M} + N - 1)!}{q(\hat{M} \leftarrow M)\hat{M}!(M + N - 1)!} \prod_{m=M+1}^{\hat{M}} (1 - \Phi(g_m))$$

# Removing Rejections

If  $\hat{M} < M$

- ▶ Choose  $M - \hat{M}$  to move to after the last acceptance.
- ▶ That's it.

Acceptance Ratio of Removal:

$$a = \frac{q(M \leftarrow \hat{M})M!(\hat{M} + N - 1)!}{q(\hat{M} \leftarrow M)\hat{M}!(M + N - 1)!} \prod_{m=\hat{M}+1}^M (1 - \Phi(g_m))^{-1}$$

# Modifying the History, Given $M$

Simple perturbations are very bad proposals.

## Better Proposals:

- ▶ Make a perturbative proposal of  $\{x_m\}$ .
- ▶ Draw from the GP prior at the new location.
- ▶ Make an underrelaxed proposal of the latent function. (Neal 1998).

## Underrelaxation: Prior Invariance

$$\hat{\mathbf{g}} = \alpha \mathbf{g} + \sqrt{1 - \alpha^2} \mathbf{h}$$

$$\mathbf{h} \sim \mathcal{N}(0, K(x, x'))$$

$$\alpha \in (0, 1]$$

# Modifying the History, Given $M$

Acceptance Ratio:

$$a = \frac{q(\{\mathbf{x}_m\} \leftarrow \{\hat{\mathbf{x}}_m\})}{q(\{\hat{\mathbf{x}}_m\} \leftarrow \{\mathbf{x}_m\})} \times \frac{\left[ \prod_{m=1}^M \pi(\hat{\mathbf{x}}_m)(1 - \Phi(\hat{\mathbf{g}}_m)) \right] \left[ \prod_{n=1}^N \Phi(\hat{\mathbf{g}}_n) \right]}{\left[ \prod_{m=1}^M \pi(\mathbf{x}_m)(1 - \Phi(\mathbf{g}_m)) \right] \left[ \prod_{n=1}^N \Phi(\mathbf{g}_n) \right]}$$

# How Many Rejections Do We Expect?

Computation complexity:  $O((N + M)^3)$ .

The Prior, Revisited:

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

- ▶ Green bits are upper-bounded by  $\pi(x)$ .
- ▶  $\mathcal{Z}_\pi[g]$  is the proportion of the mass under  $\pi(x)$  that  $\Phi(g(x))\pi(x)$  takes up.
- ▶ For a particular  $g(x)$ :  $M \simeq N(\mathcal{Z}_\pi[g]^{-1} - 1)$

# 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.

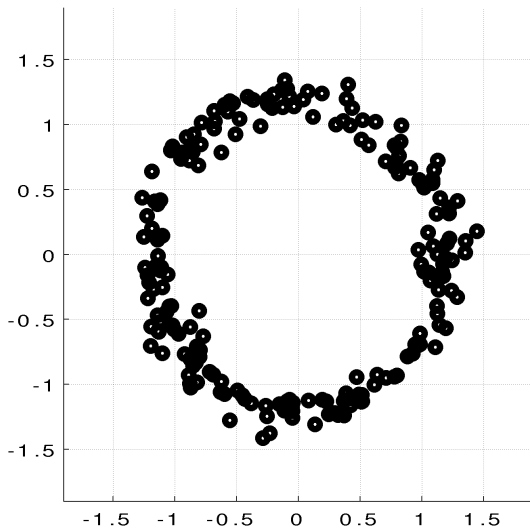
# Hyperparameter Inference

- ▶ Parameters in the covariance function,  $\theta$ .
- ▶ Parameters in the base measure,  $\phi$ .
- ▶ Fix the history, propose new parameters.

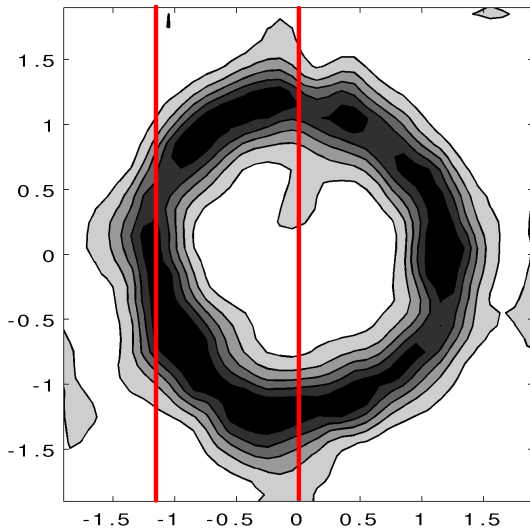
## Acceptance Ratio:

$$a = \frac{q(\theta \leftarrow \hat{\theta})q(\phi \leftarrow \hat{\phi})p(\hat{\theta})p(\hat{\phi}) \mathcal{GP}(\mathbf{g} | \hat{\theta})}{q(\hat{\theta} \leftarrow \theta)q(\hat{\phi} \leftarrow \phi)p(\theta)p(\phi) \mathcal{GP}(\mathbf{g} | \theta)}$$
$$\times \left[ \prod_{m=1}^M \frac{\pi(\mathbf{x}_m | \hat{\phi})}{\pi(\mathbf{x}_m | \phi)} \right] \left[ \prod_{n=1}^N \frac{\pi(\mathbf{x}_n | \hat{\theta})}{\pi(\mathbf{x}_n | \theta)} \right]$$

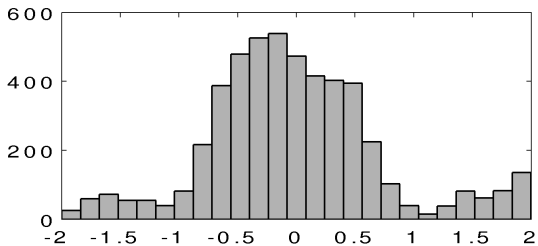
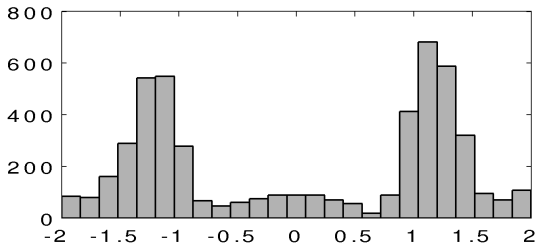
# Ring Data



# Ring Data



# Ring Data



# Summary

“Similar data should have similar probabilities”

## Gaussian Process Density Sampler

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

# Thanks

Thanks to:

- ▶ Iain Murray
- ▶ David MacKay
- ▶ Radford Neal
- ▶ Zoubin Ghahramani
- ▶ Oliver Stegle

Funded by the Gates Cambridge Trust.