

The Gaussian Process Density Sampler

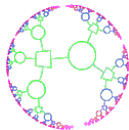
Ryan Prescott Adams

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

7 March 2008



Joint work with Iain Murray
and David MacKay



What is Density Modeling?

N data $\mathcal{D} = \{x_n\}_{n=1}^N$ drawn from a density $f(x)$

We have prior beliefs about f .

What is $p(f | \mathcal{D})$?

What is $p(x_{n+1} | \mathcal{D}) = \int_{f \in \mathcal{F}} p(x | f) p(f | \mathcal{D})$?

Common Density Models

Discrete

- ▶ Dirichlet distribution

Continuous Nonparametric

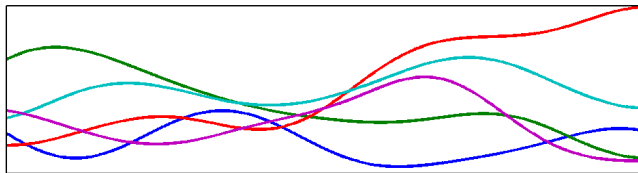
- ▶ Parzen windows
- ▶ Infinite mixtures of parametric distributions
- ▶ Dirichlet diffusion trees (Neal 2001)
- ▶ Gaussian process latent variable models (Lawrence 2005)

Short Intro to Gaussian Processes

A Prior on Functions

- ▶ Input space \mathbb{R}^d , output space \mathbb{R}
- ▶ Covariance function: $K(x, x'; \theta)$
- ▶ Mean function: $m(x; \theta)$

For any finite subset of \mathbb{R}^d of size N there is a multivariate Gaussian distribution on \mathbb{R}^N .



Short Intro to Gaussian Processes

Gaussian Predictive Distribution

$$p(y^* | x^*, \{x_n, y_n\}_{n=1}^N, \theta) = \mathcal{N}(\mu^*, v^*)$$

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

► Previous data

Short Intro to Gaussian Processes

Gaussian Predictive Distribution

$$p(y^* | x^*, \{x_n, y_n\}_{n=1}^N, \theta) = \mathcal{N}(\mu^*, v^*)$$

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

- ▶ Previous data
- ▶ Cross covariance

Short Intro to Gaussian Processes

Gaussian Predictive Distribution

$$p(y^* | x^*, \{x_n, y_n\}_{n=1}^N, \theta) = \mathcal{N}(\mu^*, v^*)$$

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

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

- ▶ Previous data
- ▶ Cross covariance
- ▶ **Marginal covariance of new data**

Short Intro to Gaussian Processes

Gaussian Predictive Distribution

$$p(y^* | x^*, \{x_n, y_n\}_{n=1}^N, \theta) = \mathcal{N}(\mu^*, v^*)$$

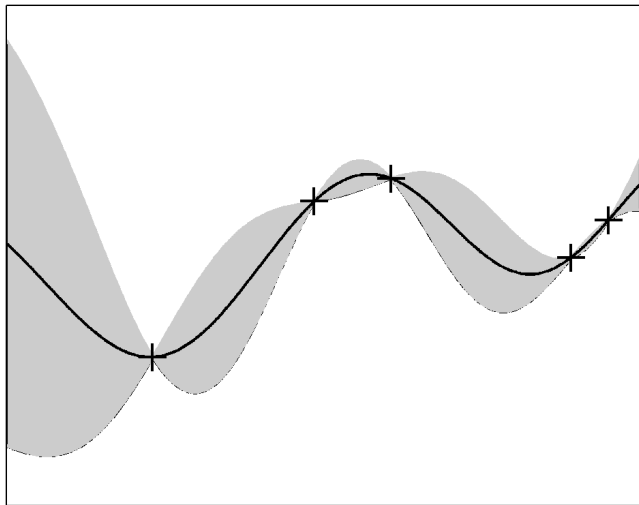
$$\mu^* = \mathbf{k}^T \mathbf{K}_N^{-1} \mathbf{y}$$

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

- ▶ Previous data
- ▶ Cross covariance
- ▶ Marginal covariance of new data
- ▶ Marginal covariance of previous data

Short Intro to Gaussian Processes

The “Sausage Link” Plot



GP Priors for PDFs

Can we use them as priors on probability density functions? Not a new idea...

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

Using GPs for PDFs is hard:

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

$$g(x) \sim \mathcal{GP}$$

GP Priors for PDFs

Can we use them as priors on probability density functions? Not a new idea...

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

Using GPs for PDFs is hard:

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

$$g(x) \sim \mathcal{GP}$$

Gaussian Process Density Sampler

- 1) Specify a GP-based prior on densities
- 2) Construct an MCMC algorithm on $g(x)$
- 3) Samples from the predictive distribution
- 4) Sample from the hyperparameters of the GP

Gaussian Process Density Sampler

- 1) Specify a GP-based prior on densities
- 2) Construct an MCMC algorithm on $g(x)$
- 3) Samples from the predictive distribution
- 4) Sample from the hyperparameters of the GP

The Prior on PDFs

$g(x)$ is drawn from a Gaussian process prior.

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

$\Phi(\cdot)$ is a sigmoid-like function:

- ▶ Nonnegative
- ▶ Bounded
- ▶ Strictly increasing

$\pi(x)$ is a known base measure.

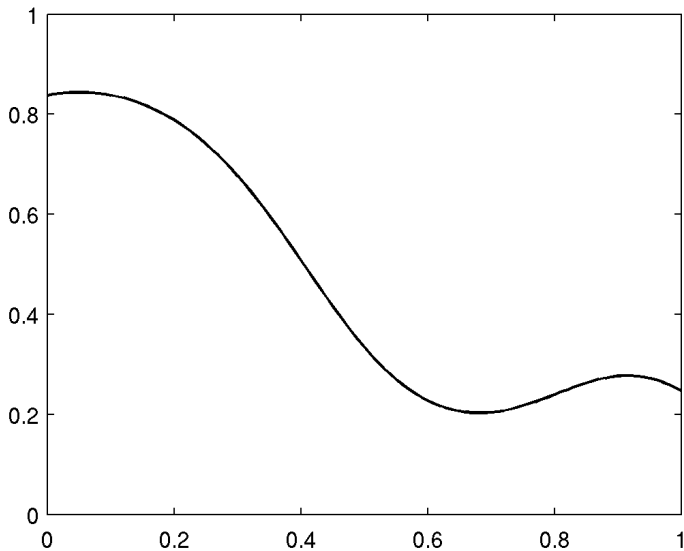
Sampling With Known $g(x)$

What if we knew $g(x)$?

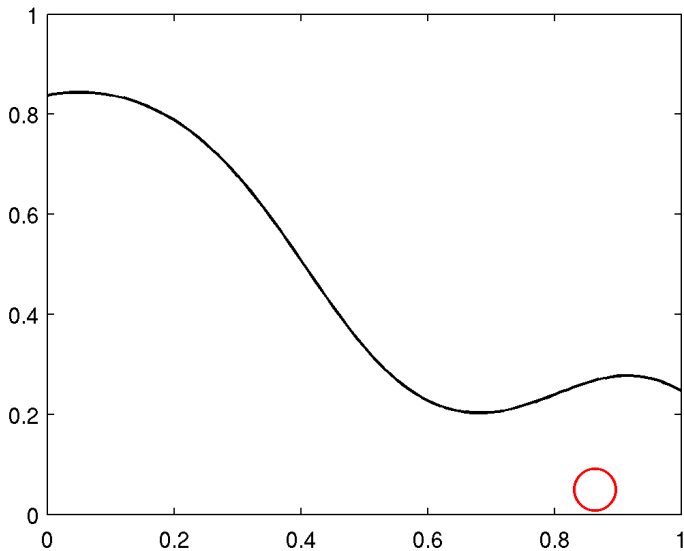
Rejection sampling:

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

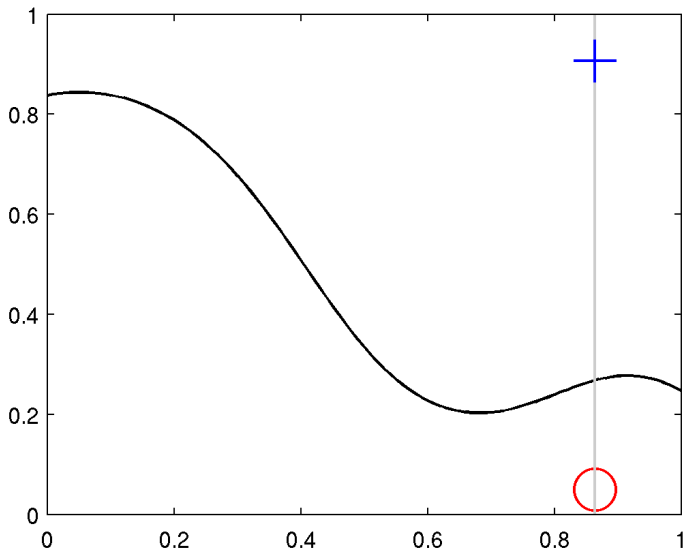
Sampling With Known $g(x)$



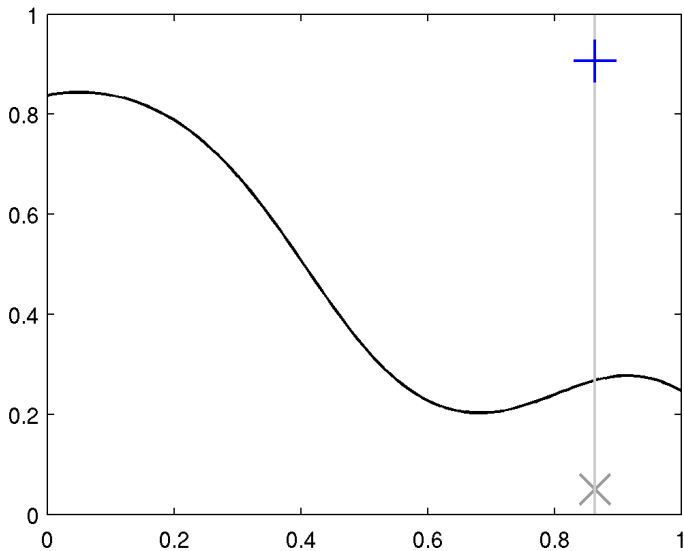
Sampling With Known $g(x)$



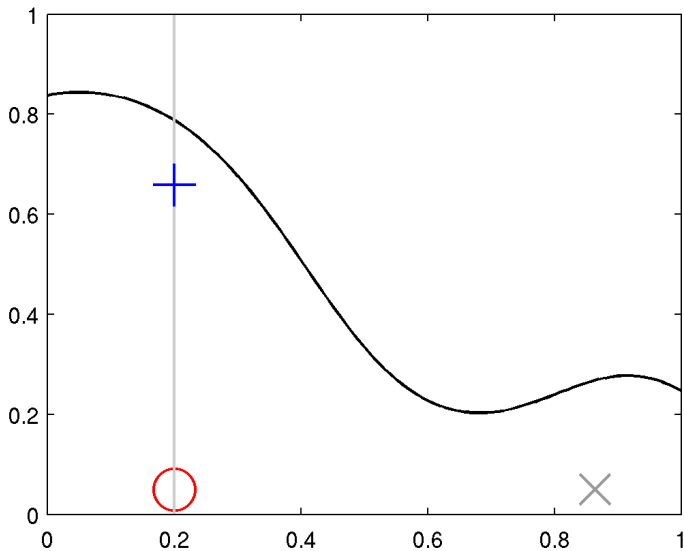
Sampling With Known $g(x)$



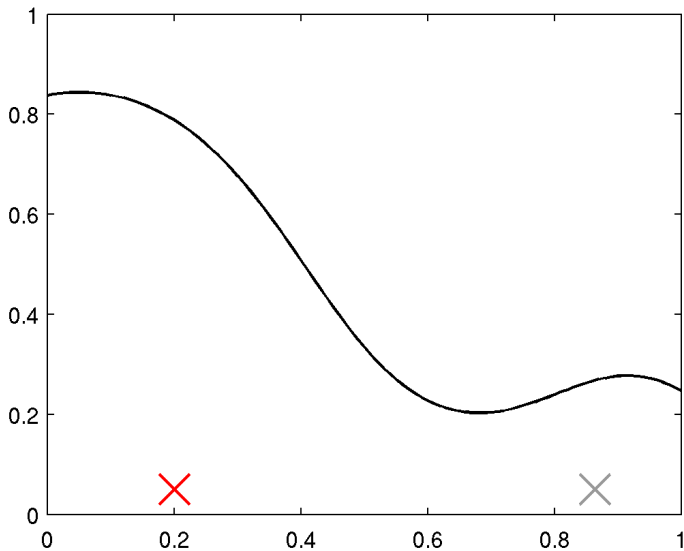
Sampling With Known $g(x)$



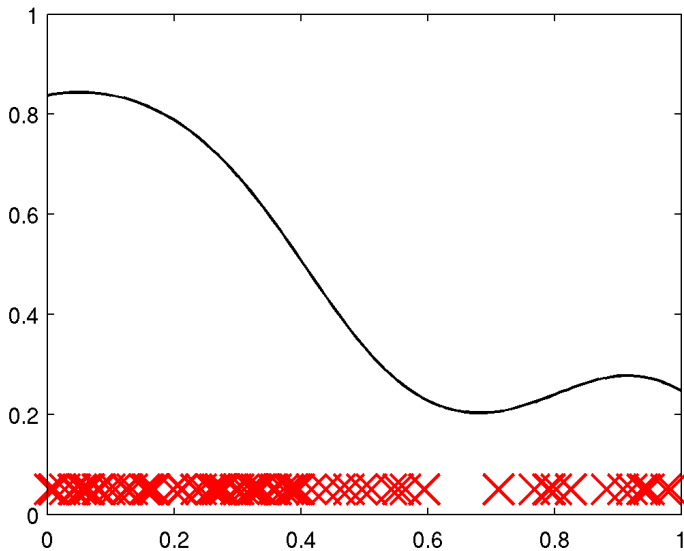
Sampling With Known $g(x)$



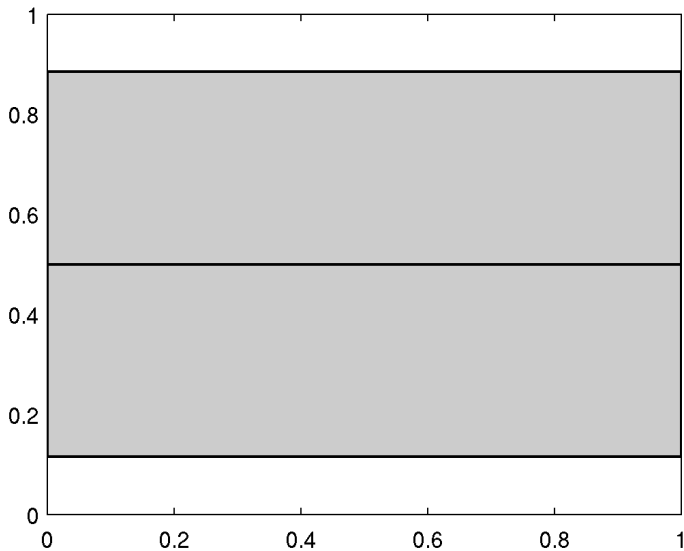
Sampling With Known $g(x)$



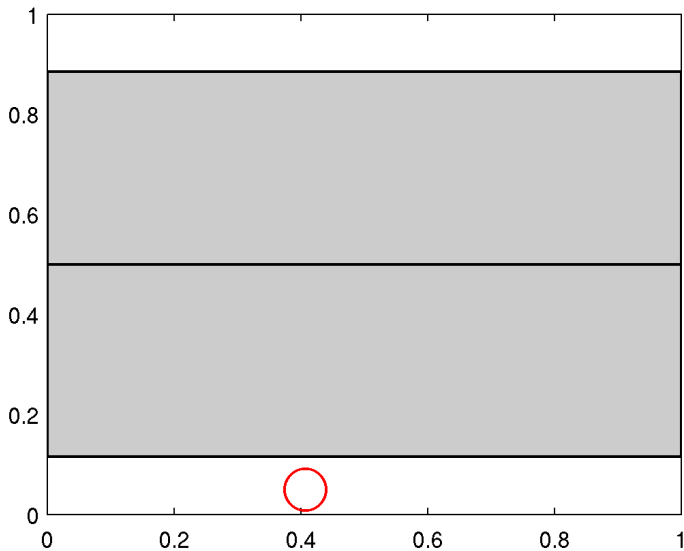
Sampling With Known $g(x)$



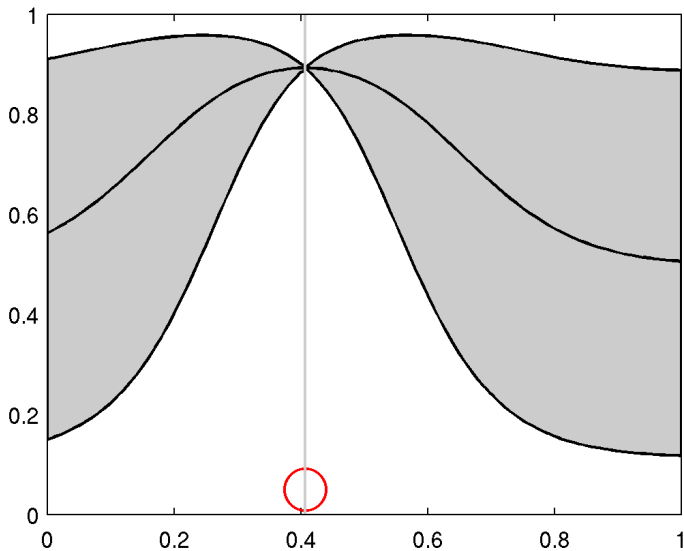
Sampling While Discovering $g(x)$



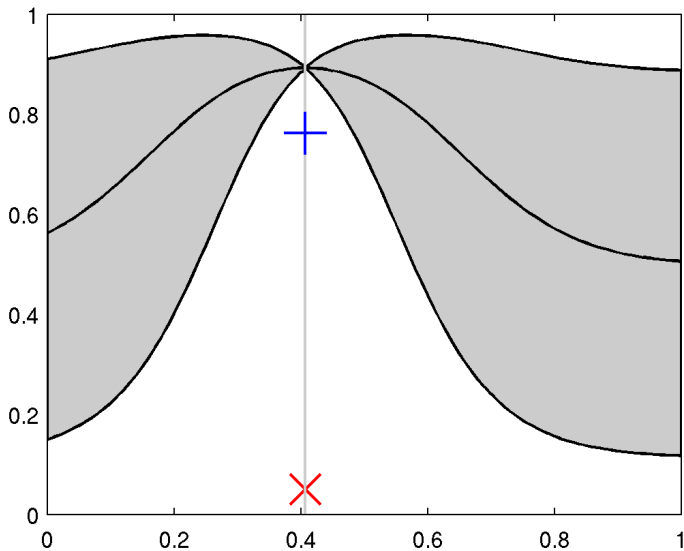
Sampling While Discovering $g(x)$



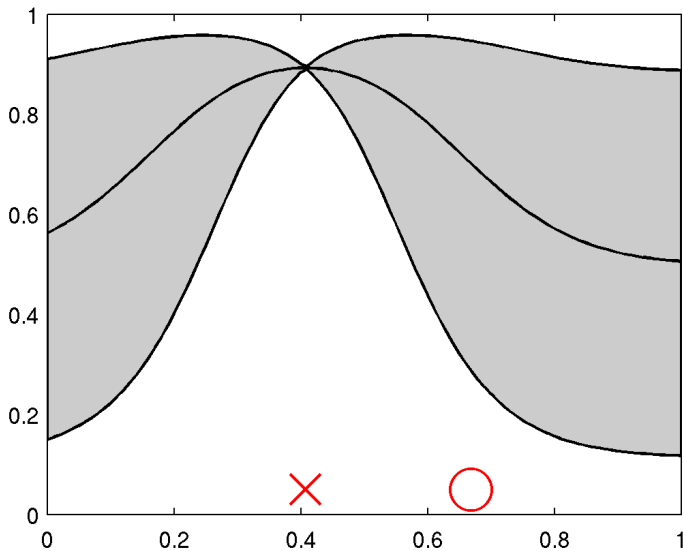
Sampling While Discovering $g(x)$



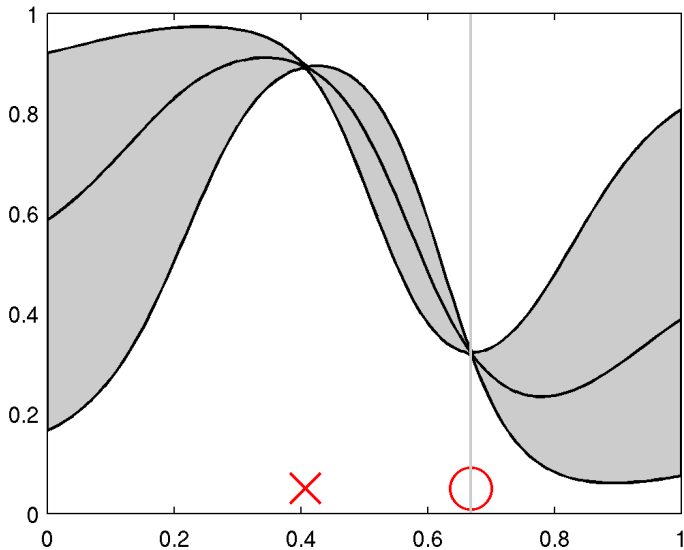
Sampling While Discovering $g(x)$



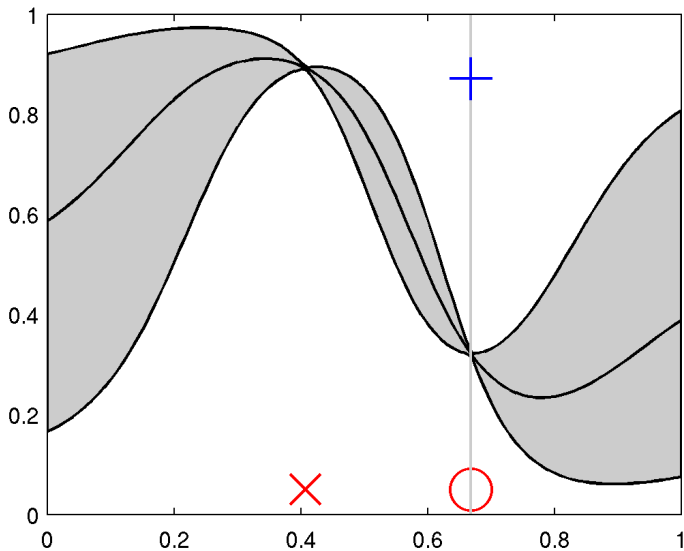
Sampling While Discovering $g(x)$



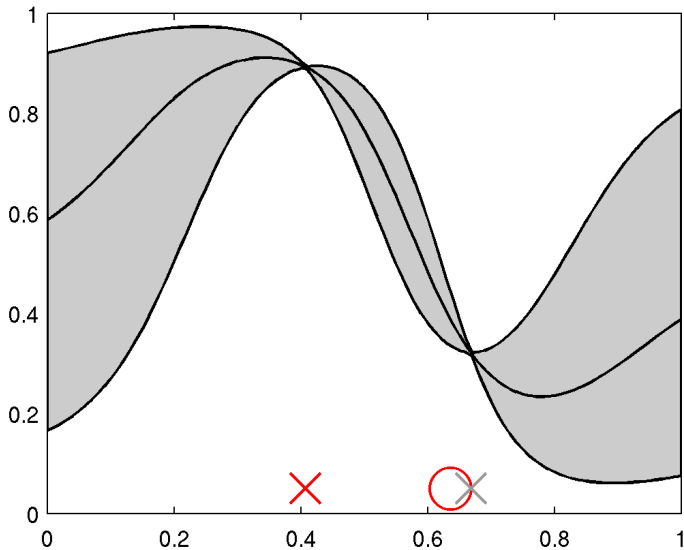
Sampling While Discovering $g(x)$



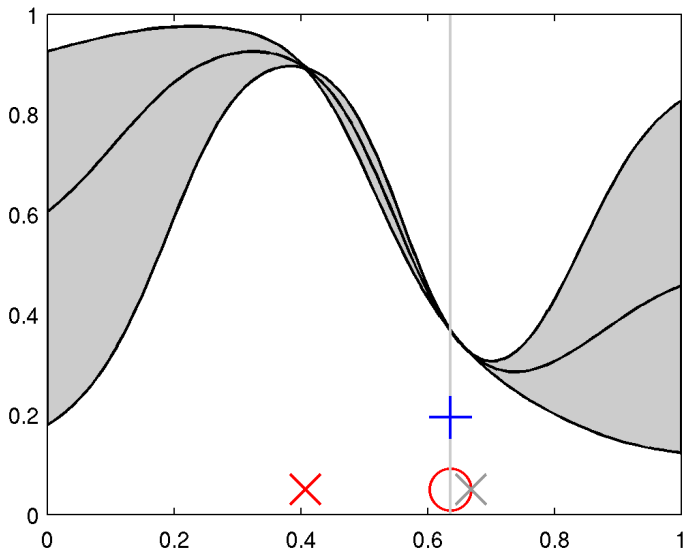
Sampling While Discovering $g(x)$



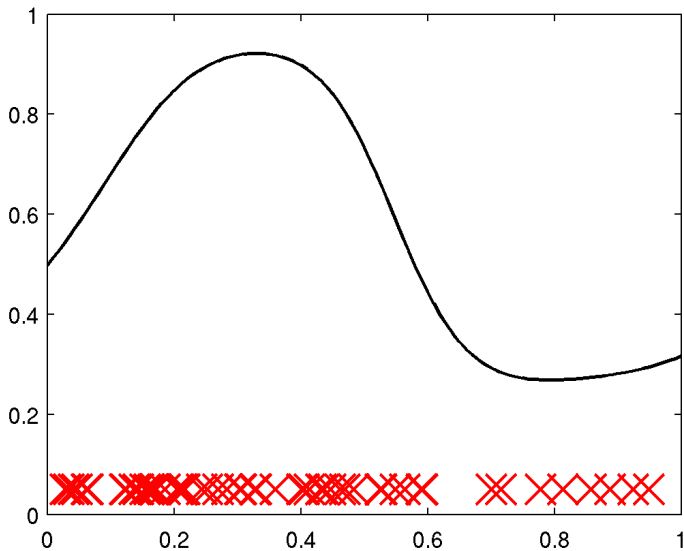
Sampling While Discovering $g(x)$



Sampling While Discovering $g(x)$



Sampling While Discovering $g(x)$



Sampling While Discovering $g(x)$

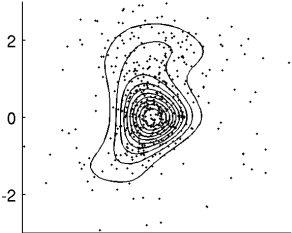
Incrementally sampling $g(x)$

Rejection sampling:

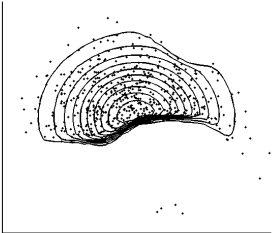
1. Draw \tilde{x} from $\pi(x)$.
2. Sample $g(\tilde{x})$ from GP given all past function samples.
3. Draw r from UNIFORM(0, 1).
4. Accept if $r < \Phi(g(\tilde{x}))$.
5. Goto 1

Effect of Hyperparameters

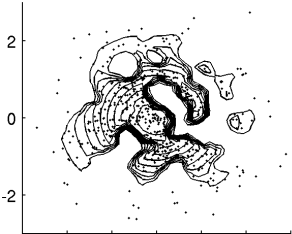
$ls_x = 1.0, ls_y = 1.0, amp = 1.0$



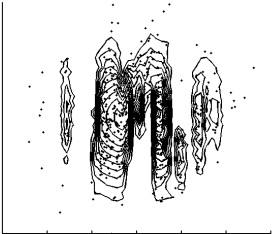
$ls_x = 1.0, ls_y = 1.0, amp = 10.0$



$ls_x = 0.25, ls_y = 0.25, amp = 5.0$



$ls_x = 0.125, ls_y = 2.0, amp = 5.0$



Gaussian Process Density Sampler

- 1) Specify a GP-based prior on densities
- 2) Construct an MCMC algorithm on $g(x)$
- 3) Samples from the predictive distribution
- 4) Sample from the hyperparameters of the GP

A First Stab at Metropolis–Hastings

Given N data $\mathcal{D} = \{x_n\}_{n=1}^N$, the posterior over g :

$$p(g | \{x_n\}_{n=1}^N) = \frac{p(\{x_n\}_{n=1}^N | g)p(g)}{p(\{x_n\}_{n=1}^N)}$$
$$\propto p(g) \prod_{n=1}^N \frac{1}{Z_\pi[g]} \Phi(g(x_n))\pi(x_n)$$

We have a difficult integral, even if we ignore the marginal likelihood!

A First Stab at Metropolis–Hastings

Write the acceptance ratio for
Metropolis–Hastings with proposal $q(\hat{g} \leftarrow g)$:

$$a = \frac{q(g \leftarrow \hat{g})p(\hat{g}) (Z_\pi[\hat{g}])^{-N} \prod_{n=1}^N \Phi(\hat{g}(x_n))}{q(\hat{g} \leftarrow g)p(g) (Z_\pi[g])^{-N} \prod_{n=1}^N \Phi(g(x_n))}$$

Exchange Sampling

Murray, Ghahramani and MacKay, UAI 2006

Add *child variables* to make the acceptance ratio tractable.

The Catch

You must be able to generate exact fantasy data from the model.

The Other Catch

You don't get to find out $Z_{\pi}[g]$, just eliminate it from the acceptance ratio.

Exchange Sampling in the GPDS

Augment $p(g, \{x_n\}_{n=1}^N)$ with the proposal \hat{g} and fantasy data $\{w_n\}_{n=1}^N$ drawn from \hat{g} :

$$\begin{aligned} p(g, \{x_n\}_{n=1}^N, \hat{g}, \{w_n\}_{n=1}^N) \\ = p(g)p(\{x_n\}_{n=1}^N | g)q(\hat{g} \leftarrow g)p(\{w_n\}_{n=1}^N | \hat{g}) \end{aligned}$$

For convenience, take $q(\hat{g} \leftarrow g) = p(\hat{g})$.

$$\begin{aligned} = p(g)p(\hat{g})(Z_\pi[g])^{-N}(Z_\pi[\hat{g}])^{-N} \\ \times \prod_{n=1}^N \Phi(g(x_n))\Phi(\hat{g}(w_n))\pi(x_n)\pi(w_n) \end{aligned}$$

Exchange Sampling in the GPDS

Now, propose the exchange of g and \hat{g} . The acceptance ratio is the ratio of the joint distributions:

$$a = \frac{p(\hat{g})p(g)(Z_\pi[\hat{g}])^{-N}(Z_\pi[g])^{-N}}{p(g)p(\hat{g})(Z_\pi[g])^{-N}(Z_\pi[\hat{g}])^{-N}} \\ \times \frac{\prod_{n=1}^N \Phi(\hat{g}(x_n))\Phi(g(w_n))\pi(x_n)\pi(w_n)}{\prod_{n=1}^N \Phi(g(x_n))\Phi(\hat{g}(w_n))\pi(x_n)\pi(w_n)}$$

Exchange Sampling in the GPDS

Now, propose the exchange of g and \hat{g} . The acceptance ratio is the ratio of the joint distributions:

$$\begin{aligned} a &= \frac{p(\hat{g})p(g)(Z_\pi[\hat{g}])^{-N}(Z_\pi[g])^{-N}}{p(g)p(\hat{g})(Z_\pi[g])^{-N}(Z_\pi[\hat{g}])^{-N}} \\ &\quad \times \frac{\prod_{n=1}^N \Phi(\hat{g}(x_n))\Phi(g(w_n))\pi(x_n)\pi(w_n)}{\prod_{n=1}^N \Phi(g(x_n))\Phi(\hat{g}(w_n))\pi(x_n)\pi(w_n)} \\ &= \frac{\prod_{n=1}^N \Phi(\hat{g}(x_n))\Phi(g(w_n))}{\prod_{n=1}^N \Phi(g(x_n))\Phi(\hat{g}(w_n))} \end{aligned}$$

Exchange Sampling in the GPDS

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

- ▶ Normalisation constant eliminated
- ▶ Only need a finite set of points of $g(x)$

The Catch

We need to track everything we know about any given $g(x)$ and this may get computationally difficult in the GP framework.

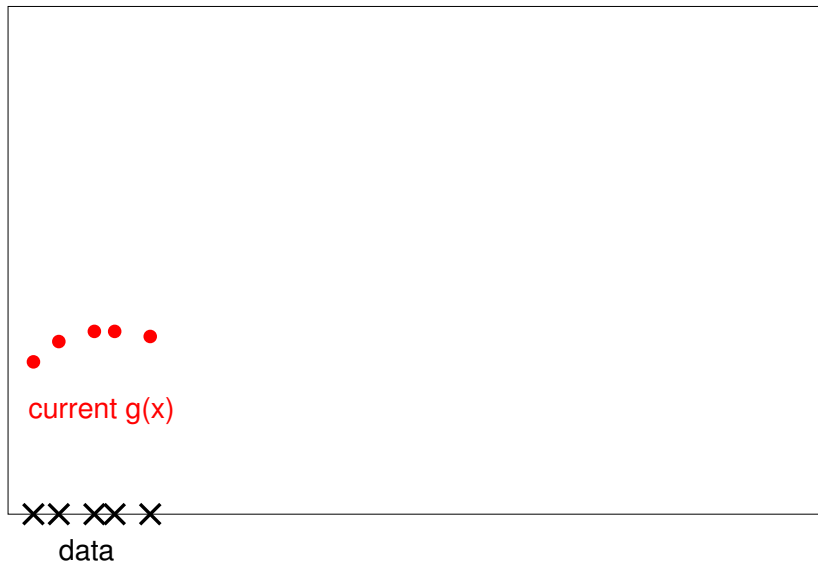
GPDS Metropolis–Hastings Cartoon



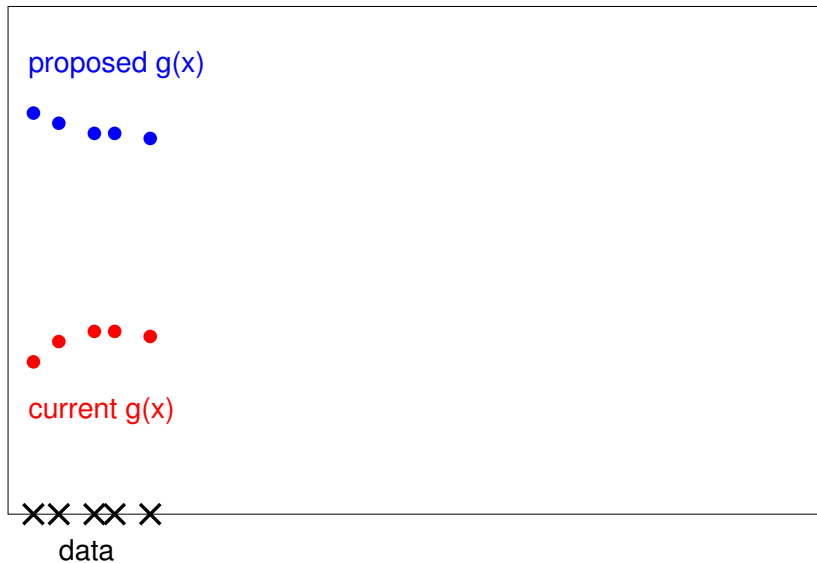
xx xx x

data

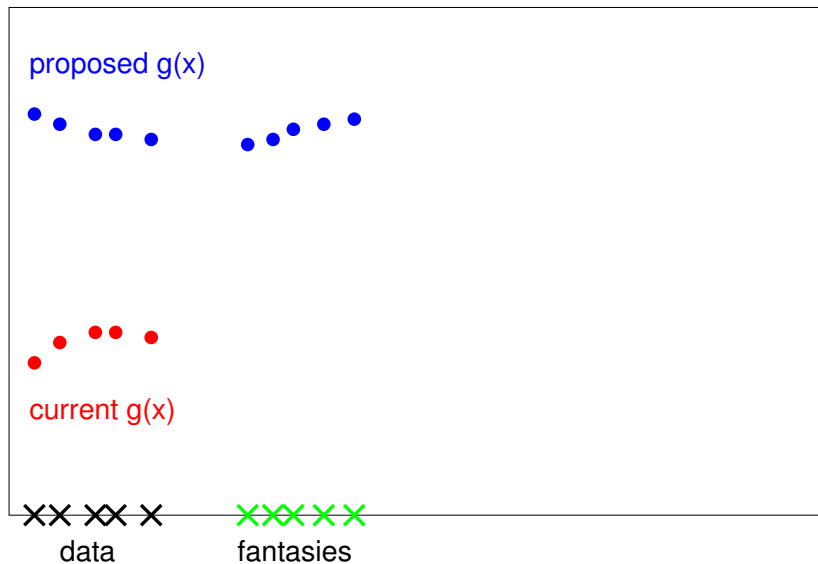
GPDS Metropolis–Hastings Cartoon



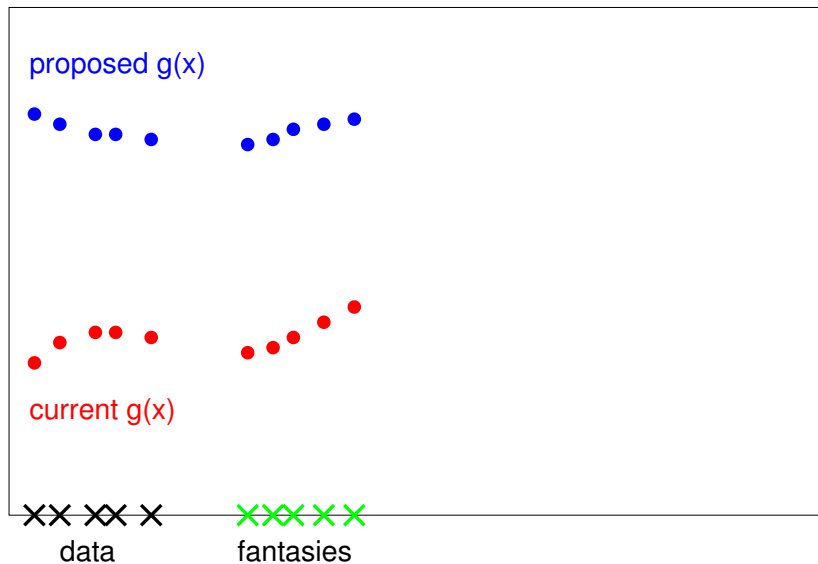
GPDS Metropolis–Hastings Cartoon



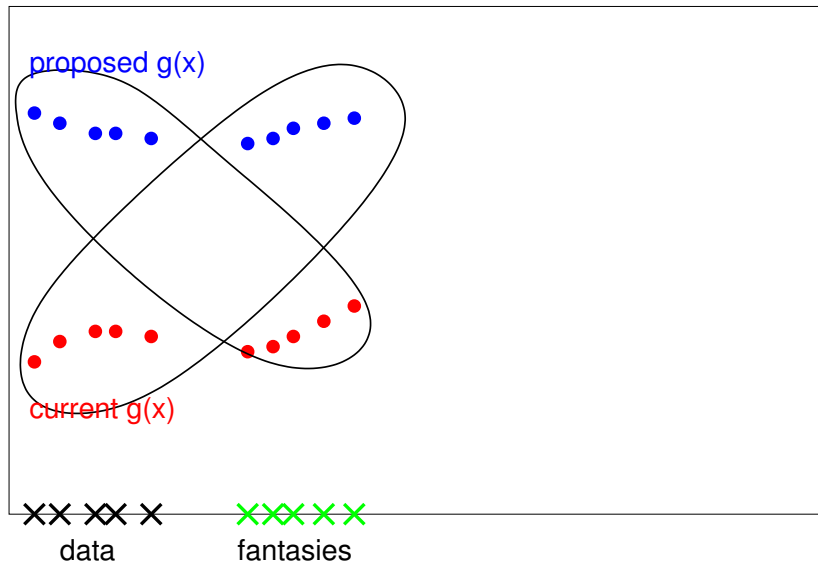
GPDS Metropolis–Hastings Cartoon



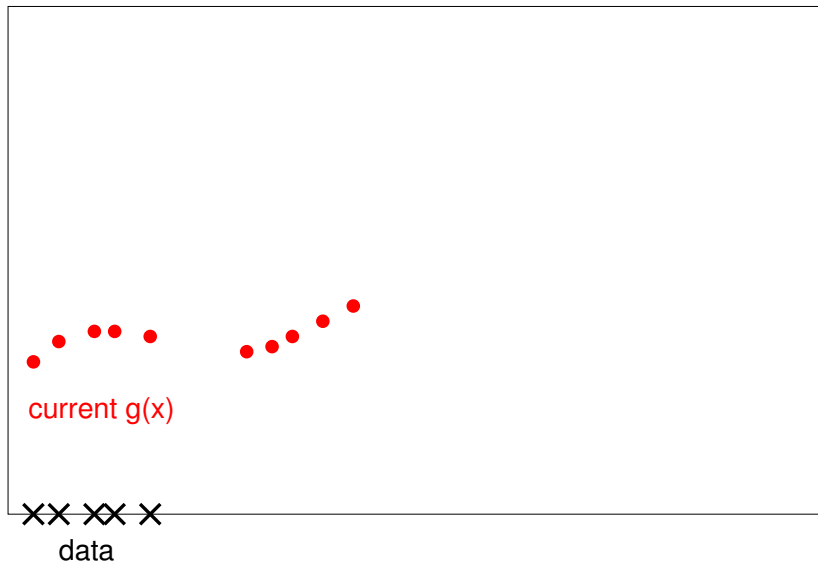
GPDS Metropolis–Hastings Cartoon



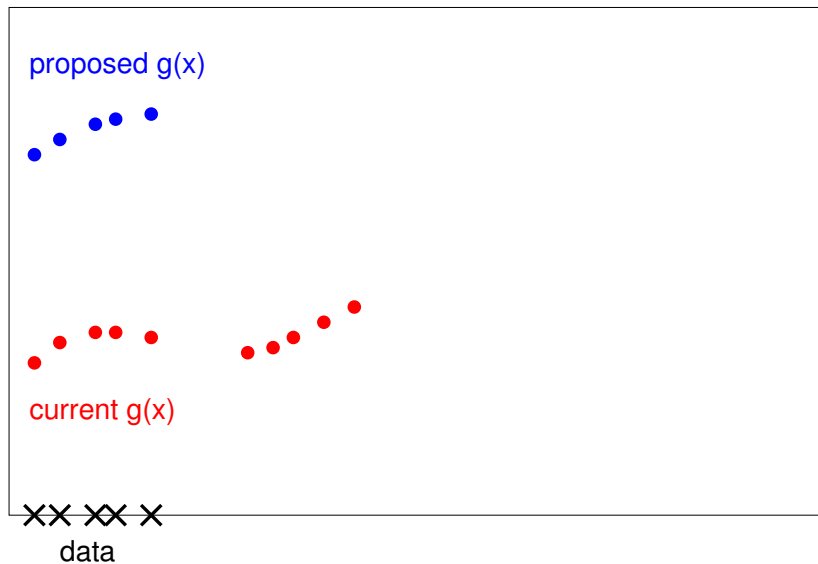
GPDS Metropolis–Hastings Cartoon



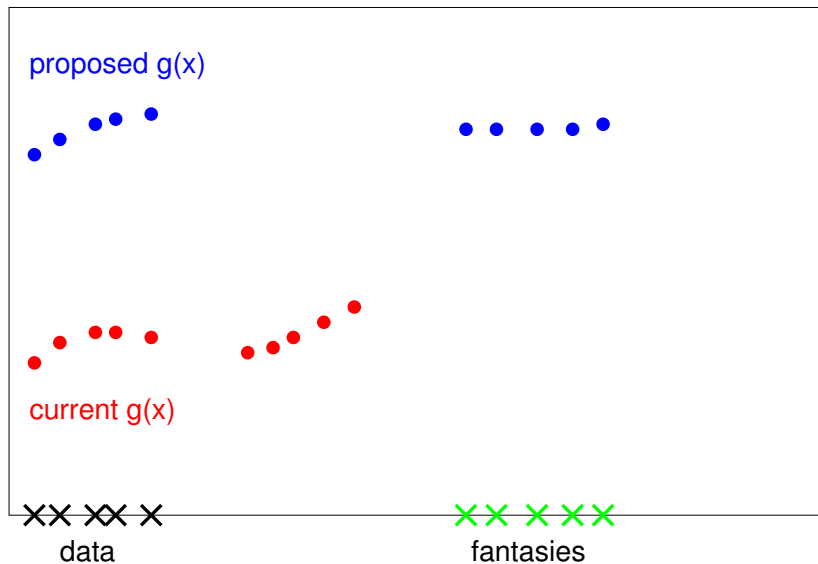
GPDS Metropolis–Hastings Cartoon



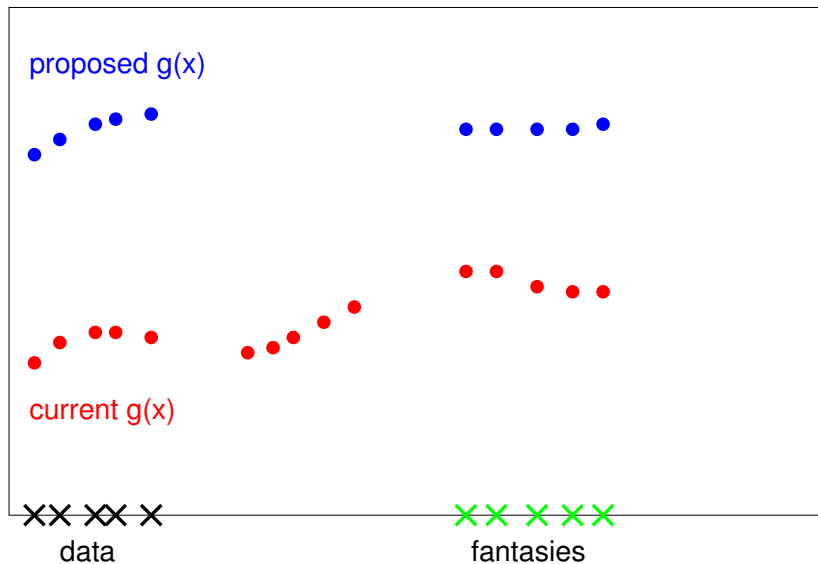
GPDS Metropolis–Hastings Cartoon



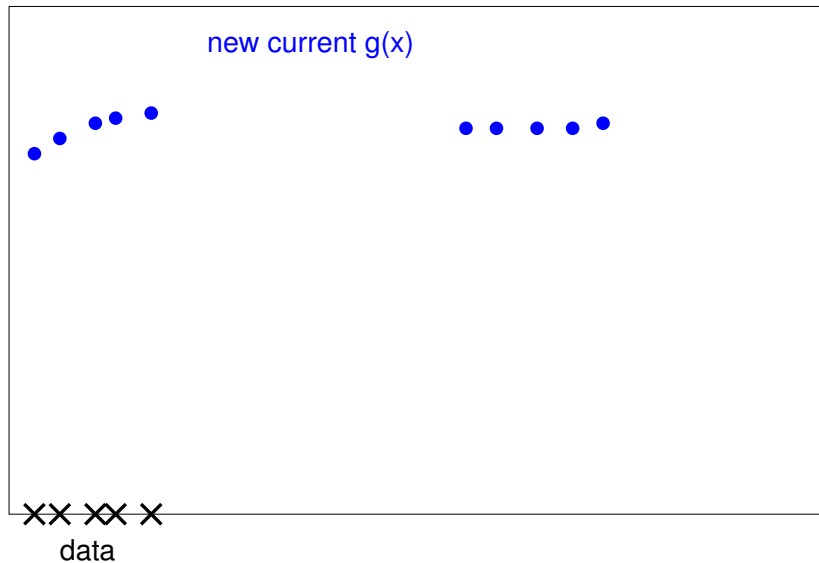
GPDS Metropolis–Hastings Cartoon



GPDS Metropolis–Hastings Cartoon



GPDS Metropolis–Hastings Cartoon



Improving the Acceptance Rate

- ▶ Computational expense goes up with each consecutive rejection.
- ▶ Draws from the prior are bad proposals.
- ▶ Make smaller incremental proposals on a finite set of “control points” $\{x_k, q_k\}_{k=1}^K$.
- ▶ Draw $g(x)$ conditioned on these points.

$$a = \frac{p(\{\hat{g}_k\})q(\{g_k\} \leftarrow \{\hat{g}_k\}) \prod_{n=1}^N \Phi(\hat{g}(x_n))\Phi(g(w_n))}{p(\{g_k\})q(\{\hat{g}_k\} \leftarrow \{g_k\}) \prod_{n=1}^N \Phi(g(x_n))\Phi(\hat{g}(w_n))}$$

Gaussian Process Density Sampler

- 1) Specify a GP-based prior on densities
- 2) Construct an MCMC algorithm on $g(x)$
- 3) Samples from the predictive distribution
- 4) Sample from the hyperparameters of the GP

Generating Predictive Samples

We have samples from the posterior of $g(x)$.

We can generate fantasy data from $g(x)$.

Generate a fantasy after each Metropolis step.

Useful Special Case: $p(x_1 | x_2)$

Generate *conditional* samples by generating from $\pi(x_1 | x_2)$ and then accepting or rejecting as before.

Gaussian Process Density Sampler

- 1) Specify a GP-based prior on densities
- 2) Construct an MCMC algorithm on $g(x)$
- 3) Samples from the predictive distribution
- 4) Sample from the hyperparameters of the GP

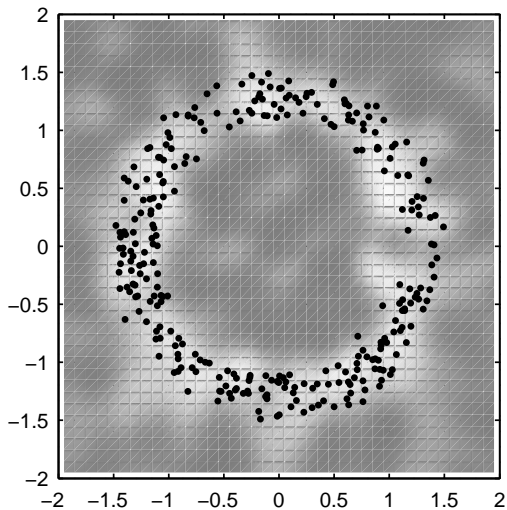
Hyperparameter Inference

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

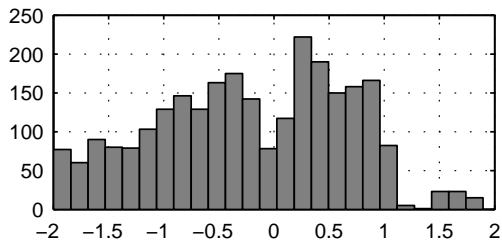
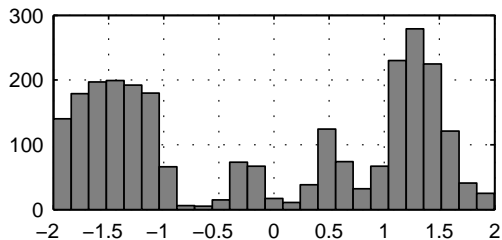
Augment the Markov chain to propose new hyperparameters simultaneously with the new function, then reject or accept together.

$$a = \frac{p(\Theta')p(\{\hat{g}_k\} | \Theta')q(\Theta \leftarrow \Theta')q(\{g_k\} \leftarrow \{\hat{g}_k\})}{p(\Theta)p(\{g_k\} | \Theta)q(\Theta' \leftarrow \Theta)q(\{\hat{g}_k\} \leftarrow \{g_k\})} \times \frac{\prod_{n=1}^N \Phi(\hat{g}(x_n))\Phi(g(w_n))}{\prod_{n=1}^N \Phi(g(x_n))\Phi(\hat{g}(w_n))}$$

Some Examples



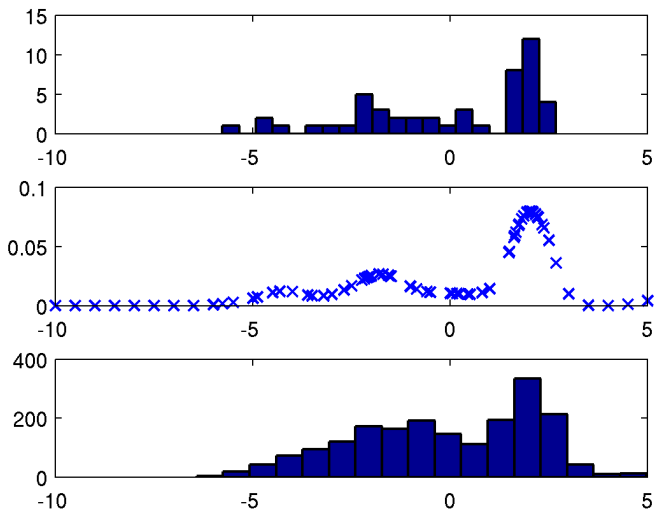
Some Examples



The “Toronto Distribution”



The “Toronto Distribution”



Future Directions

Important Things

- ▶ Prove posterior consistency.
- ▶ Explore performance improvements.
- ▶ Compare to approximate methods.

Speculative Ideas

- ▶ Estimating entropic quantities
- ▶ Clustering
- ▶ Discovering distributed representations

Thanks for coming!

Thanks to Oliver Stegle, David MacKay, Iain Murray, Zoubin Ghahramani, and the Gates Cambridge Trust.

