The Gaussian Process Density Sampler

Ryan Prescott Adams*
Cavendish Laboratory
University of Cambridge
Cambridge CB3 0HE, UK
rpa23@cam.ac.uk

Iain Murray
Dept. of Computer Science
University of Toronto
Toronto, Ontario. M5S 3G4
murray@cs.toronto.edu

David J.C. MacKay
Cavendish Laboratory
University of Cambridge
Cambridge CB3 0HE, UK
mackay@mrao.cam.ac.uk

NIPS 2008
as told by Mike Hughes

*some text/figures courtesy of Adams' online presentation
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 $f(x)$?
Motivation

We want a model that:

1) Assigns similar data to similar probabilities for any application-appropriate definition of “similar” kernel functions

2) Complexity of f(x) grows with additional data non-parametric

3) Supports inference tasks
   encode prior beliefs about f(x)
   draw new samples x from f(x)
   evaluate f(x) for any x

Gaussian Process framework = natural answer
Contributions

Naïve Logistic GP Density Model is problematic

\[ g(x) \sim \mathcal{GP}(0, K(x, x')) \]
\[ f(x) = \frac{\exp\{g(x)\}}{\int dx \exp\{g(x)\}} \]

[Adams et al] is exciting because we can …

- draw exact, exchangable samples from \( f(x) \)
- handle high-dimensional \( X \) in practice
- avoid costly + problematic estimation of norm. constant
Model

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

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

\[
K(x, x') = \alpha \exp\left(-\frac{1}{2} \sum_i \ell_i^{-2} (x_i - x_i')^2\right)
\]
Sample Densities

\[ \ell_x = 1, \; \ell_y = 1, \; \alpha = 1 \]

\[ \ell_x = 1, \; \ell_y = 1, \; \alpha = 5 \]

\[ K(x, x') = \alpha \exp\left(-\frac{1}{2} \sum_i \ell_i^{-2} (x_i - x_i')^2\right) \]
Sample Densities

\[ \ell_x = 0.25, \ \ell_y = 0.25, \ \alpha = 2 \]

\[ \ell_x = 0.25, \ \ell_y = 2, \ \alpha = 5 \]

\[ K(x, x') = \alpha \exp(-\frac{1}{2} \sum_i \ell_i^{-2} (x_i - x'_i)^2) \]
How to sample from $f(x)$?

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

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

when $g(x)$ is unknown, we can still use this idea!
Sampling by discovery

- Samples \( \{x_1, x_2, \ldots, x_N\} \) are exact and exchangeable
- Discovered latent function \( g(x) \) in process
  - well, only \( \{g(x_1), g(x_2), \ldots, g(x_N)\} \)
- Never needed that pesky normalization constant

View this sampling scheme as **generative process**
Inference

Goals:

- obtain estimates of $g(x)$
- generate samples from $f(x)$ or predictive distribution
- (optimize hyperparams)

Machinery:

- retrospective MCMC [see Iain Murray's PhD thesis]

(1) Latent History sampling
   - focus of current work
(2) Exchange sampling
   - requires more evaluations of $g(x)$
   - in practice, worse than (1)
Assume observed $X$ were accepted from generative process, we can recover rejected $X$ and $g(x)$ via MCMC

**Sampler state augmented by:**
- rejected data points $X = \{x_1, x_2, \ldots x_M\}$
- corresponding $g(x)$ values $G = \{g(x_1), g(x_2), \ldots g(x_M)\}$

**MCMC result:**
- samples $\{M, G_{\text{accept}}, G_{\text{reject}}, X_{\text{reject}}\}$ from posterior given $X$
  (can also include hyperparams)
Comparing to alternate methods

Parzen Windows

Infinite Mixture of Gaussians ( iMoG )

Dirichlet Diffusion Trees ( DFT )
Toy Data
Macaque Skull “Reconstruction”

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

![Graph showing improvement over Parzen (hats) for Mac T1, Mac T2, and Mac T3]
Concerns

Computational Complexity

- requires matrix decomp: $O((N + M)^3)$
  
  $M$ (# rejections) can be arbitrarily large!

- MCMC sampler efficiency
  no guarantees on convergence time
  poor acceptance rates in high dimensions

Advantages over alternative models?
Discussion Prompts

1) Is the GP Density Sampler worth it?  
   - why not just use iMoG?

2) What are the killer apps?

3) What's involved in a data-space other than R?

4) Possible to use alternative inference?  
   - other MCMC methods  
   - variational Bayes