

Modeling with Bounded Partition Functions

Ryan Prescott Adams

Cavendish Laboratory
University of Cambridge

<http://www.inference.phy.cam.ac.uk/rpa23/>



16 July 2008



Overall Talk Message

- ▶ Energy functions are nice models for data.
- ▶ Inference in energy models is often hard.
- ▶ If you can draw exact samples, you can do MCMC inference.
- ▶ I have a trick for generating exact data from many energy models.
- ▶ This trick is probably a bad idea.

Outline

Motivation

- Examples of Energy Models

- Inference

- Quick Review of MCMC

- Doubly-Intractable Posterior Distributions

Exchange Sampling

- Concept

- Auxiliary Variables

- Baby and Toy

Exact Sampling from Energy Models

Outline

Motivation

Examples of Energy Models

Inference

Quick Review of MCMC

Doubly-Intractable Posterior Distributions

Exchange Sampling

Concept

Auxiliary Variables

Baby and Toy

Exact Sampling from Energy Models

Energy-based Models of Data

For some space \mathcal{X} , write an energy: $E(\mathbf{x}; \theta)$

Turn this into a probability distribution via:

$$p(\mathbf{x} | \theta) = \frac{1}{\mathcal{Z}(\theta)} \exp \{-E(\mathbf{x}; \theta)\}$$

Big energy implies small probability.

Normalised by $\mathcal{Z}(\theta) = \int_{\mathcal{X}} d\mathbf{x} \exp \{-E(\mathbf{x}; \theta)\}$:

- ▶ Called the **partition function**.
- ▶ Depends on the parameters θ .
- ▶ Intractable in many interesting models.

Examples of Energy-based Models

Exponential Family Distributions

$$E(\mathbf{x}; \boldsymbol{\theta}) = -\boldsymbol{\theta}^\top T(\mathbf{x}) + h(\mathbf{x})$$

- ▶ Gaussian, Gamma, Poisson, etc.
- ▶ Typically easy.

Examples of Energy-based Models

Undirected Graphical Models

$$E(\mathbf{x}; \theta) = -\mathbf{x}^T \mathbf{V} \mathbf{x} - \mathbf{h}^T \mathbf{H} \mathbf{h} - \mathbf{x}^T \mathbf{J} \mathbf{h} - \mathbf{x}^T \boldsymbol{\alpha} - \mathbf{h}^T \boldsymbol{\beta}$$

- ▶ Ising/Potts models, Boltzmann machines
- ▶ Perhaps hidden units \mathbf{h} .
- ▶ Often with finite states.
- ▶ Hard!

Examples of Energy-based Models

Nonparametric Models

$$E(\mathbf{x}; \theta) = g(\mathbf{x})$$

- ▶ $g(\mathbf{x})$ a nonparametric function.
- ▶ Logistic Gaussian process.
- ▶ Hard!

Inference

Given N data $\mathcal{D} = \{\mathbf{x}_n\}_{n=1}^N$, what is θ ?

$$p(\theta | \mathcal{D}) = \frac{p(\theta)}{\int d\theta p(\mathcal{D} | \theta) p(\theta)} \prod_{n=1}^N p(\mathbf{x}_n | \theta)$$

For interesting problems, θ is often complex.

Use Markov chain Monte Carlo?

Quick Review of MCMC

We have $p'(\theta) \propto p(\theta)$ and want to draw samples.

Markov chain (MC): a stochastic rule for wandering around in the space of θ .

MC can have an **equilibrium distribution**.

Simulate a MC for a while and θ is close to being a sample from the equilibrium distribution.

Write down a rule using $p'(\theta)$ so that the MC has $p(\theta)$ as its equilibrium distribution.

Metropolis–Hastings

Metropolis–Hastings is a popular MCMC variant.

MH Markov Transition Rule

Current state is θ .

1. Make a proposal $\hat{\theta} \sim q(\hat{\theta} \leftarrow \theta)$.
2. Evaluate the **acceptance ratio**:

$$a = \frac{q(\theta \leftarrow \hat{\theta}) p'(\hat{\theta})}{q(\hat{\theta} \leftarrow \theta) p'(\theta)}$$

3. Accept $\hat{\theta}$ with probability $\min(a, 1)$, otherwise keep θ .

MH for Inference

Back to the posterior:

$$p(\theta | \mathcal{D}) = \frac{p(\theta)}{\int d\theta p(\mathcal{D} | \theta) p(\theta)} \prod_{n=1}^N p(\mathbf{x}_n | \theta)$$
$$\propto p(\theta) \prod_{n=1}^N p(\mathbf{x}_n | \theta)$$

The “sledgehammer” of Bayesian inference.

Even the MH sledgehammer doesn't work on many energy models!

Doubly-Intractable Posterior

$$p(\theta) \prod_{n=1}^N p(\mathbf{x}_n | \theta) = p(\theta) \prod_{n=1}^N \frac{1}{\mathcal{Z}(\theta)} \exp \{-E(\mathbf{x}_n; \theta)\}$$

Write the Metropolis–Hastings acceptance ratio:

$$a = \frac{q(\theta \leftarrow \hat{\theta}) p(\hat{\theta})}{q(\hat{\theta} \leftarrow \theta) p(\theta)} \frac{\mathcal{Z}(\theta)^N}{\mathcal{Z}(\hat{\theta})^N} \\ \times \prod_{n=1}^N \exp \left\{ -E(\mathbf{x}_n; \hat{\theta}) + E(\mathbf{x}_n; \theta) \right\}$$

Outline

Motivation

Examples of Energy Models

Inference

Quick Review of MCMC

Doubly-Intractable Posterior Distributions

Exchange Sampling

Concept

Auxiliary Variables

Baby and Toy

Exact Sampling from Energy Models

Exchange Sampling

“Murray’s Magical Metropolis Method”

Introduce **auxiliary variables** to the Markov chain so that we can evaluate the Metropolis–Hastings acceptance ratio.

The Catch

We must be able to generate **exact fantasy data** from the model, given a setting of the parameters.

Exchange Sampling

Make our joint distribution bigger:

$$p(\theta, \{\mathbf{x}_n\}_{n=1}^N, \hat{\theta}, \{\mathbf{w}_n\}_{n=1}^N) =$$

$$p(\theta) \quad p(\{\mathbf{x}_n\}_{n=1}^N | \theta) \quad q(\hat{\theta} \leftarrow \theta) \quad p(\{\mathbf{w}_n\}_{n=1}^N | \hat{\theta})$$
The equation is presented as a sequence of four terms, each enclosed in a colored rectangular box. From left to right: a purple box containing $p(\theta)$, a green box containing $p(\{\mathbf{x}_n\}_{n=1}^N | \theta)$, a red box containing $q(\hat{\theta} \leftarrow \theta)$, and a light blue box containing $p(\{\mathbf{w}_n\}_{n=1}^N | \hat{\theta})$. A thin black arrow originates from the right side of the green box and points to the left side of the purple box.

- ▶ Prior on model parameters.

Exchange Sampling

Make our joint distribution bigger:

$$p(\theta, \{\mathbf{x}_n\}_{n=1}^N, \hat{\theta}, \{\mathbf{w}_n\}_{n=1}^N) =$$

$$p(\theta) \quad p(\{\mathbf{x}_n\}_{n=1}^N | \theta) \quad q(\hat{\theta} \leftarrow \theta) \quad p(\{\mathbf{w}_n\}_{n=1}^N | \hat{\theta})$$

- ▶ Prior on model parameters.
- ▶ Likelihood of true data.

Exchange Sampling

Make our joint distribution bigger:

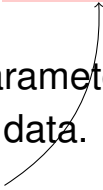
$$p(\theta, \{\mathbf{x}_n\}_{n=1}^N, \hat{\theta}, \{\mathbf{w}_n\}_{n=1}^N) =$$

$$p(\theta)$$

$$p(\{\mathbf{x}_n\}_{n=1}^N | \theta)$$

$$q(\hat{\theta} \leftarrow \theta)$$

$$p(\{\mathbf{w}_n\}_{n=1}^N | \hat{\theta})$$

- ▶ Prior on model parameters.
 - ▶ Likelihood of true data.
 - ▶ Proposal density.
- 

Exchange Sampling

Make our joint distribution bigger:

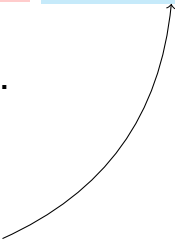
$$p(\theta, \{\mathbf{x}_n\}_{n=1}^N, \hat{\theta}, \{\mathbf{w}_n\}_{n=1}^N) =$$

$p(\theta)$

$p(\{\mathbf{x}_n\}_{n=1}^N | \theta)$

$q(\hat{\theta} \leftarrow \theta)$

$p(\{\mathbf{w}_n\}_{n=1}^N | \hat{\theta})$

- ▶ Prior on model parameters.
 - ▶ Likelihood of true data.
 - ▶ Proposal density.
 - ▶ Likelihood of fantasy data.
- 

Exchange Sampling

Markov Transition Rule

Current state is θ .

1. Make a proposal $\hat{\theta} \sim q(\hat{\theta} \leftarrow \theta)$.
2. Draw N fantasy data $\{\mathbf{w}_n\}_{n=1}^N \sim p(\mathbf{x} | \hat{\theta})$.
3. Propose **exchanging** θ and $\hat{\theta}$.
 - ▶ Acceptance ratio is just the ratio of the joints.

$$a = \frac{q(\theta \leftarrow \hat{\theta}) p(\hat{\theta})}{q(\hat{\theta} \leftarrow \theta) p(\theta)} \prod_{n=1}^N \frac{\cancel{\mathcal{Z}(\hat{\theta})}^{-1} \exp \left\{ -E(\mathbf{x}_n; \hat{\theta}) \right\}}{\cancel{\mathcal{Z}(\theta)}^{-1} \exp \left\{ -E(\mathbf{x}_n; \theta) \right\}}$$
$$\times \prod_{n=1}^N \frac{\cancel{\mathcal{Z}(\theta)}^{-1} \exp \left\{ -E(\mathbf{w}_n; \theta) \right\}}{\cancel{\mathcal{Z}(\hat{\theta})}^{-1} \exp \left\{ -E(\mathbf{w}_n; \hat{\theta}) \right\}}$$

Baby and Toy

The partition function cancel out, leaving:

$$a = \frac{q(\theta \leftarrow \hat{\theta}) p(\hat{\theta})}{q(\hat{\theta} \leftarrow \theta) p(\theta)} \prod_{n=1}^N \frac{\exp \left\{ -E(\mathbf{x}_n; \hat{\theta}) - E(\mathbf{w}_n; \theta) \right\}}{\exp \left\{ -E(\mathbf{x}_n; \theta) - E(\mathbf{w}_n; \hat{\theta}) \right\}}$$

“You can’t take a toy away from a baby without giving it a new one.”

– DJCM (paraphrased)

Outline

Motivation

Examples of Energy Models

Inference

Quick Review of MCMC

Doubly-Intractable Posterior Distributions

Exchange Sampling

Concept

Auxiliary Variables

Baby and Toy

Exact Sampling from Energy Models

How to Generate Exact Samples?

Exchange sampling requires **exact** samples.

Iain used Coupling from the Past (CFTP) to get exact samples with Potts models.

CFTP does not work in models with a lot of structure.

Hence, the “bounded partition function” trick...

Bounding the Partition Function

Rather than this:

$$p(\mathbf{x} | \theta) = \frac{1}{Z(\theta)} \exp \{-E(\mathbf{x}; \theta)\}$$

we do this:

$$p(\mathbf{x} | \theta) = \frac{1}{Z(\theta)} (1 + \exp \{E(\mathbf{x}; \theta)\})^{-1} \pi(\mathbf{x})$$

- ▶ $\pi(\mathbf{x})$ is something easy to sample from.
- ▶ Big energy still has small probability.
- ▶ We can generate exact samples now!

Generating Exact Samples

$$p(\mathbf{x} | \theta) = \frac{1}{\mathcal{Z}(\theta)} (1 + \exp \{E(\mathbf{x}; \theta)\})^{-1} \pi(\mathbf{x})$$

By squashing the energy function, we get

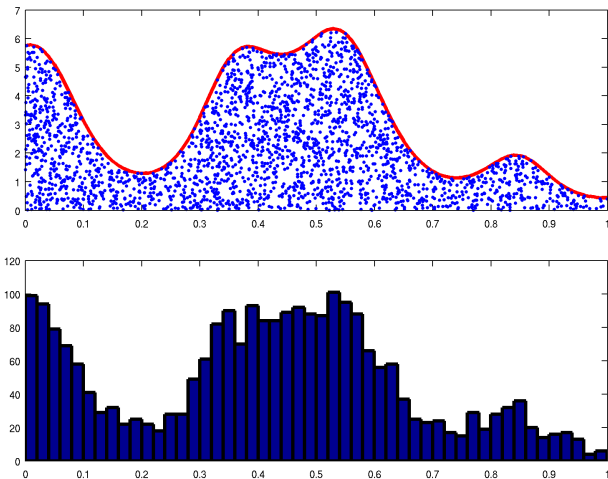
$$0 < (1 + \exp \{E(\mathbf{x}; \theta)\})^{-1} < 1, \forall \mathbf{x}, \theta.$$

$\pi(\mathbf{x})$ is an upper bounding density.

We can use rejection sampling.

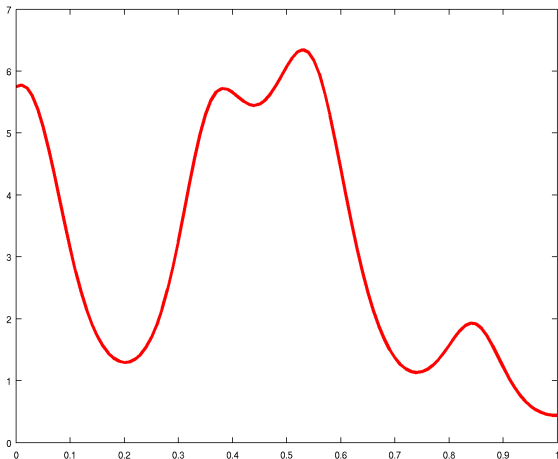
Rejection Sampling

Sample uniformly from the volume, and the marginal of the x coordinates is the density.



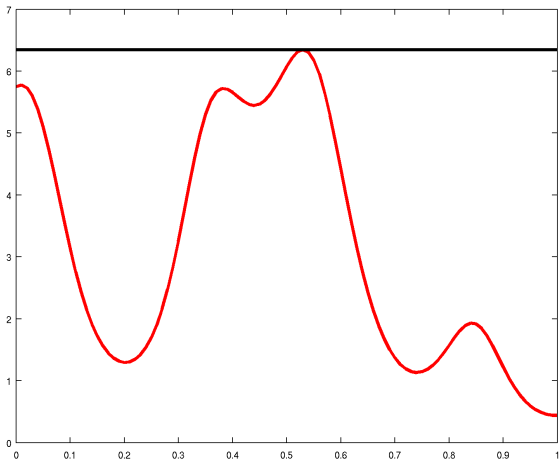
Rejection Sampling

To sample uniformly from a complicated volume, sample from an envelope and then reject the proposals outside.



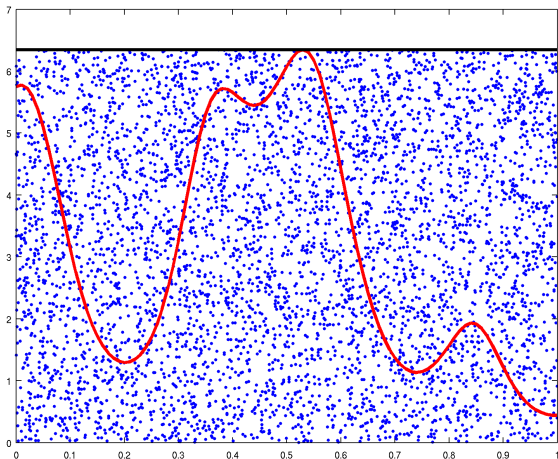
Rejection Sampling

To sample uniformly from a complicated volume, sample from an envelope and then reject the proposals outside.



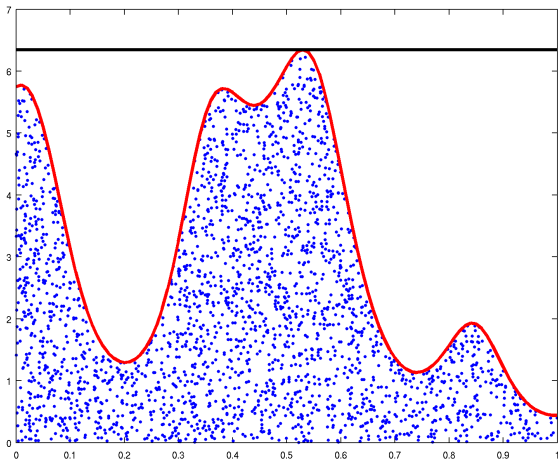
Rejection Sampling

To sample uniformly from a complicated volume, sample from an envelope and then reject the proposals outside.



Rejection Sampling

To sample uniformly from a complicated volume, sample from an envelope and then reject the proposals outside.



Back to Energy Models

$$p(\mathbf{x} | \theta) = \frac{1}{\mathcal{Z}(\theta)} (1 + \exp \{E(\mathbf{x}; \theta)\})^{-1} \pi(\mathbf{x})$$

So:

$$(1 + \exp \{E(\mathbf{x}; \theta)\})^{-1} \pi(\mathbf{x}) \propto p(\mathbf{x} | \theta)$$

and $\pi(\mathbf{x})$ is an envelope.

Exact Sampling from Energy Models

1. Draw $\mathbf{x} \sim \pi(\mathbf{x})$.
2. Draw $r \sim \mathcal{U}(0, 1)$.
3. Accept if $r < (1 + \exp \{E(\mathbf{x}; \theta)\})^{-1}$,
else reject.
4. Loop until you get enough.

- ▶ We can generate exact samples.
- ▶ We can do MCMC inference.

(in principle...)

A Bad Idea

Rejection sampling is **exponentially inefficient** with increasing dimensionality.

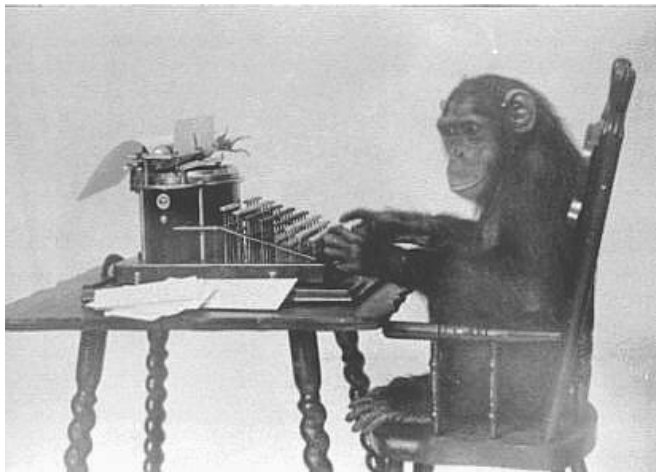
High-dimensional problems are the interesting ones.

Dimensionality is often why the partition function is difficult in the first place.

We've just traded one intractability for another.

“Million Monkey Sampling”

This approach is like modeling Shakespeare by waiting for monkeys to randomly generate it.



How Bad? Really Bad.

Regarding modeling the MNIST digits:

*The best RBMs are of order 100
nats better than mixture of Bernoullis.*

– Iain Murray

So, if $\pi(\mathbf{x})$ is simple you might be waiting until the end of the universe *to take one Metropolis step!*

Any Way Out?

One possibility: use this construction to merge models of local structure.

A very slight bit of good news: rejection sampling is trivially parallelizable.

Developers need to start thinking about software for thousands of cores.

– Intel

Summary

- ▶ Energy functions are a way to model data.
- ▶ A lot of energy-based models are intractable.
- ▶ There exist MCMC algorithms to resolve this when you can generate exact samples.
- ▶ I presented a trick for exact sampling.
- ▶ This trick is a disaster in high dimensions.
- ▶ Maybe there is insight to be gained.