Auxiliary-variable Exact Hamiltonian Monte Carlo Samplers for Binary Distributions

Ari Pakman and Liam Paninski
Department of Statistics, Center for Theoretical Neuroscience, Grossman Center for the Statistics of Mind, Columbia University

Introduction

- We map the problem of sampling from binary distributions into continuous variables.
- The Hamiltonian Monte Carlo method is a Markov Chain Monte Carlo algorithm for continuous variables, with usually better performance than Metropolis or Gibbs samplers. It proposes transitions in the Markov chain lying far apart in the sampling space, while maintaining good acceptance rates.
- Implementations of HMC generally involve tuning numerical integration parameters for good acceptance rates. In our algorithm the Hamiltonian equations of motion can be integrated exactly, so there is no need for tuning and the Markov chain always accepts the proposed moves.
- The same ideas can be used to sample from mixtures of binary and Gaussian or Exponential variables.

Mapping Binary to Piecewise Continuous Distributions

We want to sample from a distribution over binary vectors $\mathbf{s} \in \{-1,+1\}^d$

$$p(\mathbf{s}) = \frac{f(\mathbf{s})}{Z}$$

where Z does not need to be known.

- Augment $p(\mathbf{s})$ with continuous variables $\mathbf{y} \in \mathbb{R}^d$ as

$$p(\mathbf{s},\mathbf{y}) = p(\mathbf{s})p(\mathbf{y}|\mathbf{s})$$

  where $p(\mathbf{y}|\mathbf{s})$ is non-zero only in the orthant

$$s_i = sign(y_i) \qquad i = 1,\ldots,d.$$

- Sample $\mathbf{y}$ from the piecewise continuous distribution

$$p(\mathbf{y}) = \sum_{\mathbf{s}'} p(\mathbf{s}')p(\mathbf{y}|\mathbf{s}') = p(\mathbf{s})p(\mathbf{y}|\mathbf{s})$$

  and read out the values of $\mathbf{s}$ from $s_i = sign(y_i)$.
- Choose for $\log p(\mathbf{y}|\mathbf{s})$ a quadratic function of $\mathbf{y}$ on its support and sample from $p(\mathbf{y})$ using Hamiltonian Monte Carlo with exactly-solvable equations of motion.*

*(Pakman and Paninski, Exact Hamiltonian Monte Carlo for Truncated Multivariate Gaussians, JCGS, 2013)

The Exact HMC Sampling Algorithm

- Introduce momentum variables q_i and define the piecewise smooth Hamiltonian

$$H(\mathbf{y},\mathbf{q}) = U(\mathbf{y}) + \frac{\mathbf{q}\cdot\mathbf{q}}{2}$$
$$U(\mathbf{y}) = -\log p(\mathbf{y}|\mathbf{s}) - \log f(\mathbf{s})$$

- In each iteration, sample $\mathbf{q}(0)$ from a standard Gaussian and evolve a d-dimensional particle during a time T according to

$$\dot{\mathbf{y}}(t) = \frac{\partial H}{\partial \mathbf{q}(t)}, \qquad \dot{\mathbf{q}}(t) = -\frac{\partial H}{\partial \mathbf{y}(t)}.$$

- The particle moves smoothly inside an orthant. When a coordinate reaches a boundary $y_i = 0$, the potential energy $U(\mathbf{y})$ is discontinuous and the momentum q_i either is reflected or jumps in order to conserve the total energy.
- The final coordinates $\mathbf{y}(T)$ belong to a Markov chain with invariant distribution $p(\mathbf{y})$, and are the initial coordinates of the next iteration.

Gaussian Augmentation

$$p(\mathbf{y}|\mathbf{s}) \propto \begin{cases} e^{-\frac{\mathbf{y}\cdot\mathbf{y}}{2}} & \text{for } \mathrm{sign}(y_i) = s_i, \qquad i = 1,\ldots,d \\ 0 & \text{otherwise}, \end{cases}$$

- The equations of motion are $\ddot{\mathbf{y}}(t) = -\mathbf{y}(t)$, with solution

$$y_i(t) = y_i(0)\cos(t) + q_i(0)\sin(t),$$

- At each iteration, the boundaries $y_i = 0$ are reached in a fixed order that repeats itself with period $T = \pi$.
- The rate at which each coordinate y_i crosses the $y_i = 0$ boundary coincides with the flip acceptance probability of a Metropolis sampler, but the algorithms have a different global structure.

Exponential Augmentation

$$p(\mathbf{y}|\mathbf{s}) = \begin{cases} e^{-\sum_{i=1}^d |y_i|} & \text{for } \mathrm{sign}(y_i) = s_i, \qquad i = 1,\ldots,d \\ 0 & \text{otherwise}, \end{cases}$$

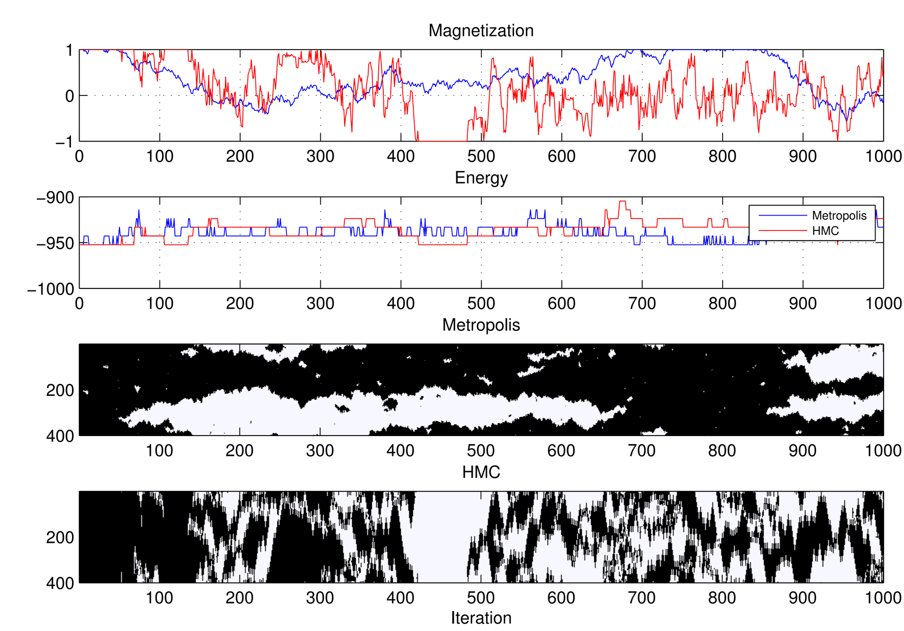
- The equations of motion are $\ddot{y}_i(t) = -s_i$, with solution

$$y_i(t) = y_i(0) + q_i(0)t - \frac{s_i t^2}{2}.$$

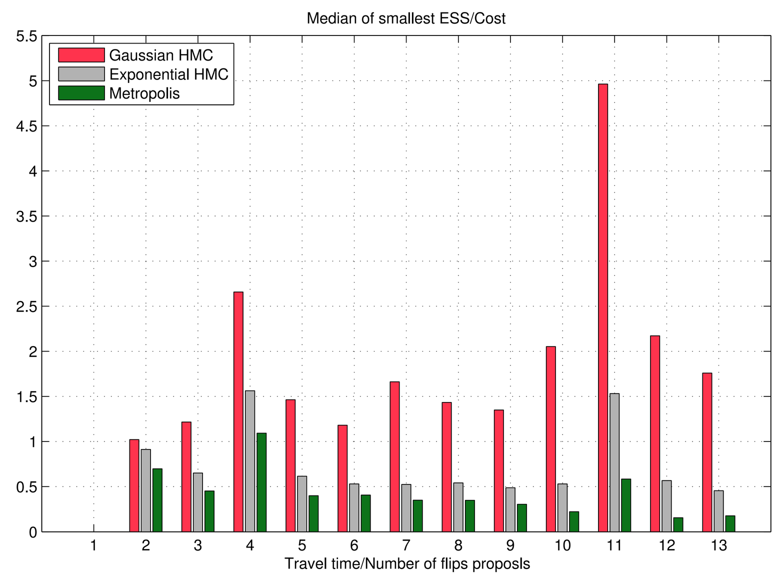
- The order in which the boundaries $y_i = 0$ are hit is not periodic.

Example: 1D Ising model, d=400

Samples of Gaussian HMC vs Metropolis



Efficiency comparison



Exact HMC for Spike-and-Slab Regression with Truncations

In a regression model with log-likelihood

$$\log p(D|\mathbf{w}) = -\frac{1}{2}\mathbf{w}'\mathbf{M}\mathbf{w} + \mathbf{r}\cdot\mathbf{w} + const.$$

and spike-and-slab prior

$$p(\mathbf{w},\mathbf{s}|a,\tau^2) = \prod_{i=1}^d p(w_i|s_i,\tau^2)a^{\frac{(1+s_i)}{2}}(1-a)^{\frac{(1-s_i)}{2}}$$

$$p(w_i|s_i,\tau^2) = \begin{cases} \frac{1}{\sqrt{2\pi\tau^2}}e^{-\frac{w_i^2}{2\tau^2}} & \text{for } s_i = +1, \\ \delta(w_i) & \text{for } s_i = -1 \end{cases}$$

we want to sample from the posterior

$$p(\mathbf{w},\mathbf{s}|D,a,\tau^2) \propto \frac{e^{-\frac{1}{2}\mathbf{w}'\mathbf{M}\mathbf{w}+\mathbf{r}\cdot\mathbf{w}}e^{-\frac{1}{2}\mathbf{w}'_+\mathbf{w}_+\tau^{-2}}}{(2\pi\tau^2)^{|\mathbf{s}_+|/2}}\delta(\mathbf{w}_-)a^{|\mathbf{s}_+|}(1-a)^{|\mathbf{s}_-|}$$

- Replace the delta functions by a factor similar to the slab

$$\delta(w_i) \to \frac{1}{\sqrt{2\pi\tau^2}}e^{-\frac{w_i^2}{2\tau^2}} \qquad \text{for } s_i = -1$$

  This non-singular distribution for the excluded w_i's creates a Reversible Jump sampler.
- As for the binary case, augment the distribution with $\mathbf{y}$ and marginalize over $\mathbf{s}$.
- The Gaussian augmentation gives a piecewise Gaussian distribution which can be sampled using exact HMC

$$p(\mathbf{w},\mathbf{y}|D,a,\tau^2) \propto e^{-\frac{1}{2}\mathbf{w}'_+(\mathbf{M}_+ + \tau^{-2})\mathbf{w}_+ + \mathbf{r}_+\cdot\mathbf{w}_+}e^{-\frac{\mathbf{w}_-\cdot\mathbf{w}_-}{2\tau^2}}e^{-\frac{\mathbf{y}\cdot\mathbf{y}}{2}}a^{|\mathbf{s}_+|}(1-a)^{|\mathbf{s}_-|}$$

- To obtain samples from the original distribution: $(w_i, y_i \geq 0)$ becomes $(w_i, s_i = +1)$ and $(w_i, y_i < 0)$ becomes $(w_i = 0, s_i = -1)$.
- Using exact HMC sampling, it is easy to truncate the $\mathbf{w}$ space (e.g. positivity).

Bayesian Probit with Spike-and-Slab Prior

Given N covariates $\mathbf{x}_i$, we observe N binary variables $b_i = \pm 1$ with probability

$$p(b_i|\mathbf{w},\mathbf{x}_i) = \frac{1}{\sqrt{2\pi}}\int_{z_i b_i \geq 0} dz_i e^{-\frac{1}{2}(z_i + \mathbf{x}_i\mathbf{w})^2}$$

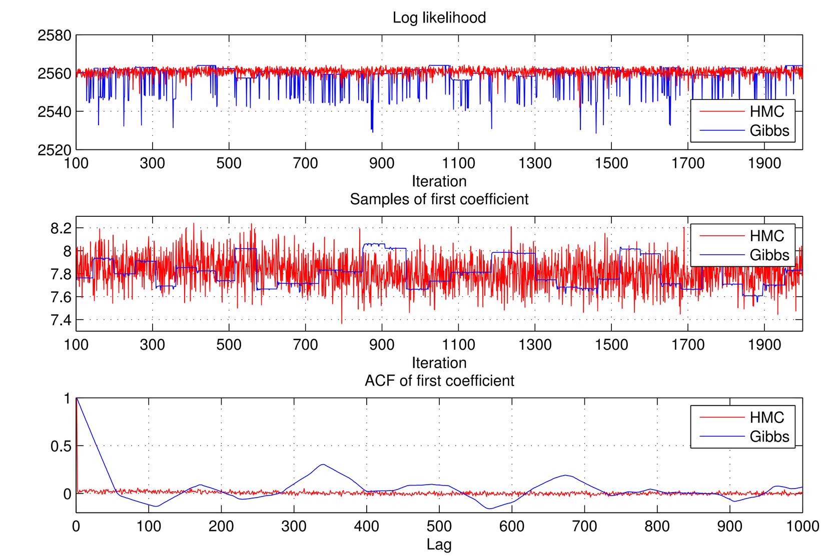
With a spike-and-slab prior for the weights $\mathbf{w}$, the posterior is the marginal over the z_i's of the distribution

$$p(\mathbf{z},\mathbf{w},\mathbf{s}|\mathbf{x},a,\tau^2) \propto \prod_{i=1}^N e^{-\frac{1}{2}(z_i+\mathbf{x}_i\mathbf{w})^2}p(\mathbf{w},\mathbf{s}|a,\tau^2) \qquad z_i b_i \geq 0,$$

As before, this distribution can be transformed into a truncated piecewise Gaussian, defined over the vector $(\mathbf{z},\mathbf{w},\mathbf{y})$, and the $\mathbf{w}$ space can also be truncated. We can sample from this new distribution using exact HMC.

Example: Spike-and-Slab Regression with Positive Coefficients

HMC vs block-Gibbs*



*(Mohamed, Heller and Ghahramani, Bayesian and L1 Approaches to Sparse Unsupervised Learning, ICML 2012).

Email: ari,liam@stat.columbia.edu