MCMC with constellations of points

Alex Shestopaloff
U of T Department of Statistics
Supervised by Radford Neal

U of T Graduate Student Research Day, April 28, 2011

What are constellations of points?

- A constellation is simply a collection of points in some parameter space.
- Constellations are constructed by taking an initial point and increasing or decreasing each coordinate by a constant, keeping the others fixed.
- An example of a constellation in one dimension would be $\{\theta, \theta + \epsilon, \theta - \epsilon\}$, for some $\epsilon > 0$.
- Constellations can be defined in many other ways as well. We can make ϵ random, we can add or subtract multiples or powers of epsilon.

Ensemble MCMC methods

- Our general framework is ensemble MCMC, introduced by Neal in 2010.
- Here, we first probabilistically map from the space we want to sample from, say Θ, to its K-fold Cartesian product, Θ^K.
- We then perform updates in the product space, and map back to the original space to obtain a sample from our distribution.
- This must be done in a way that leaves the distribution we want to sample from invariant.

A constellation Metropolis sampler (1)

- Consider a (parameter) space Θ, and some point $\theta = (\theta_1^p, \ldots, \theta_K^p) \in \Theta$.
- Consider a constellation consisting of the $2K + 1$ points $\{(\theta_1^c, \ldots, \theta_K^c), (\theta_1^c + \epsilon, \ldots, \theta_K^c), \ldots, (\theta_1^c, \ldots, \theta_K^c - \epsilon)\} = \{\theta_1^c, \ldots, \theta_{2K+1}^c\}$.
- Map to the constellation by identfiying θ with a point of this constellation. This is done by uniformly selecting an index $i \in \{1, \ldots, 2K + 1\}$.
- Compute the remaining points of the constellation, relative to the selected point.

A constellation Metropolis sampler (2)

- Once we have mapped to a constellation, we want to perform an update in the constellation space.
- Suppose we are sampling from a density π, and our constellation consists of the $2K+1$ points $\{\theta_1, \ldots, \theta_{2K+1}\}$.
- We propose a new constellation by shifting the current one. For example, we can propose a new 'centre' point, and shift all other points relative to it.
- The constellation move is accepted or rejected with probability

$$\min\left(1, \frac{\sum_i \pi(\theta_i^*)}{\sum_i \pi(\theta_i)}\right)$$

- The constellation can then be mapped back to a single point i with probabilities $\pi(\theta_i)/\sum_i \pi(\theta_i)$.
- Note that we can perform multiple updates in the constellation space before mapping back to a point.

When can this help?

- The hope is that this sampler accepts more of the moves we propose, especially larger moves, compared to the usual Metropolis algorithm.
- But... such a method makes sense only if we can evaluate the density for every constellation point faster than evaluating the density at each point separately.
- If this is not possible, we may well just use the extra time to generate more samples.
- Fortunately, efficient computations are possible for a large class of models: neural networks, logistic regression, linear regression, and so on.

A neural network model and example problem

- Consider a neural network for regression with P inputs and H hidden units.

$$t = a + \sum_{j=1}^{H} v_j \tanh(b_j + \sum_{k=1}^{P} x_k w_{jk})$$
$$= a + \sum_{j=1}^{H} v_j h_j$$

- We want to fit this network to 100 observations from

$$t = 0.7x_1^2 + 0.8\sin(0.3 + (4.5 + 0.5x_1)x_2) +$$
$$+ 0.85\cos(0.1 + 5x_3 + 0.1x_2^2) + N(0, 0.4^2)$$

  where, in addition to the 3 relevant inputs, we include 9 additional inputs, 6 that are highly correlated with the relevant ones, and 3 irrelevant ones. This example is from (Neal, 2010).

Fitting the network (1)

- We want to fit this network using a Bayesian approach, by averaging over predictions made using a sequence of parameter samples from the posterior.

- We assume that the regression noise is Gaussian, and put a mean zero Gaussian prior on the parameters.

- We update the noise variance and the variances of different parameter groups using Gibbs sampling from an Inverse Gamma distribution.

- When using constellation MCMC, we will consider a constellation over the 'hidden to output weights' v_j only.

Fitting the network (2)

- For efficient evaluations of the posterior density at $v_j + \epsilon$, we save the values of h_j's and $\sum_{j=1}^{H} v_j h_j$. We then add ϵh_j to the latter sum.

- Note that this requires a total of 2 arithmetic operations, as opposed to $2H$ if we only saved the $h_j's$ and recomputed the sum from scratch. If we didn't save the $h_j's$, we would have to do even more.

- Similar computational savings apply if we want to evaluate the density for constellations over other parameters.

Performance of constellation MCMC

- We use a spherical Gaussian as our proposal distribution, fixing the standard deviation at 0.0425. For constellation MCMC, we set ϵ to 0.1. We draw a total of 100000 samples.

- At each constellation MCMC step, we map to a single point, update the prior variances, and then map to a new constellation.

- To compare constellation and ordinary Metropolis, we look at autocorrelation times of $\sum_{j=1}^{H} w_{jk}^2$, the sums of squared weights for each of the 12 inputs, and the acceptance rates.

- The autocorrelation times are estimated using an AR model. On average, the reduction is about 37%. In particular, the average autocorrelation time over all inputs is reduced from about 832 to about 607.

- The acceptance rate increases by about 20%, from about 18.6% to about 22.4%.

Dynamics of constellation MCMC

- We want to understand whether proposals are accepted due to moving to a new constellation point. This can help in selecting a good value of ϵ
- If ϵ is too large, it is possible that we may get stuck at a small subset of the constellation points, mapping back to the same point for long sequences of steps.
- In an extreme case, we may always map to a single point. The algorithm thus essentially works in a single dimension.
- Asymptotically, the probabilities of selecting a point with a particular index from the constellation become uniform. This is another diagnostic measure.

Some illustrations

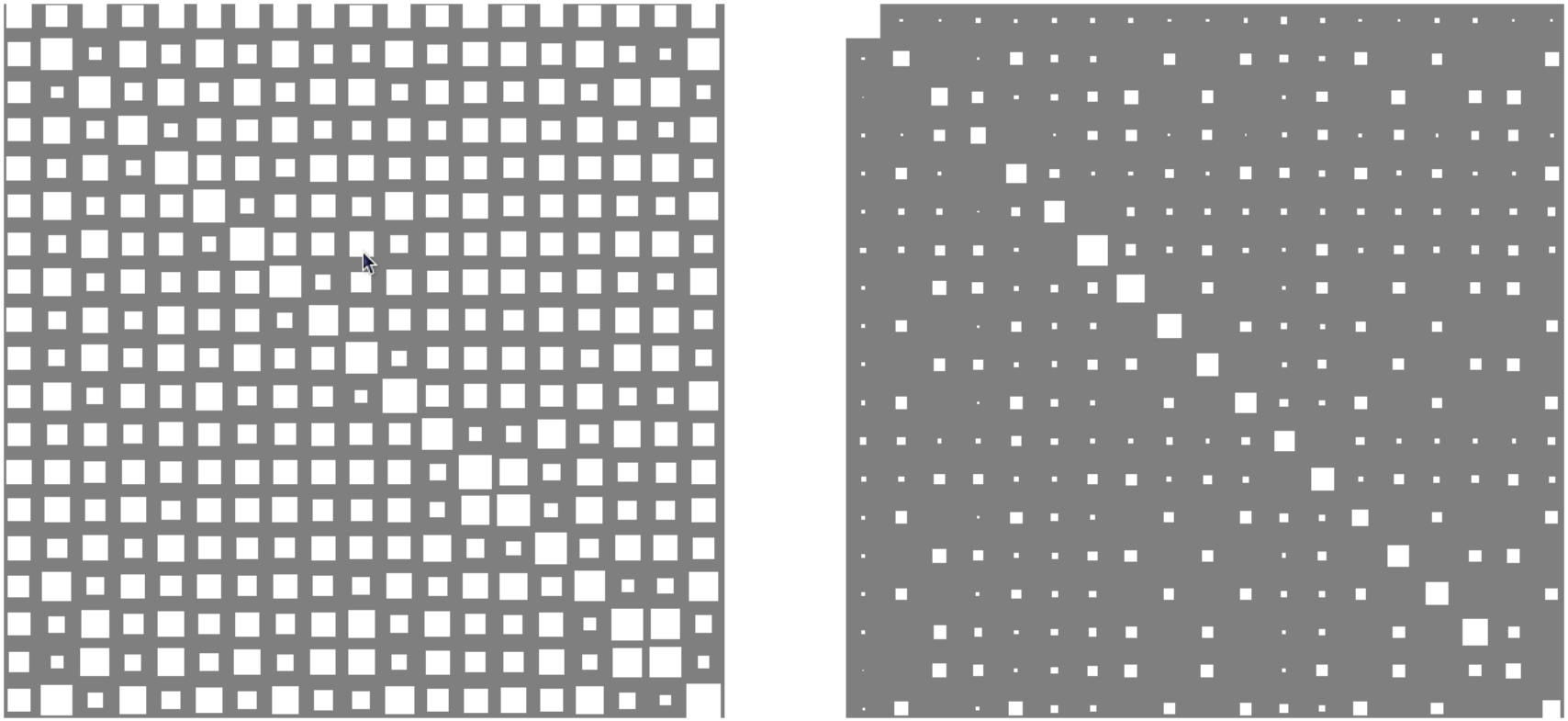

Figure 1: Estimated transition probabilities, epsilon=0.1 and epsilon=0.4

Future plans

- Selection of a good value of epsilon. We can link this to the prior variances.
- Choosing the number of 'shells', when to map back and forth from a constellation to a point.
- Selecting the parameters we construct a constellation for.
- Better understanding of the dynamics of constellation Metropolis.