Statistical Machine Learning
Primer on probability and statistics

Luca Bortolussi

Department of Mathematics and Geosciences
University of Trieste

Office 328, third floor, H2bis
lbortolussi@units.it

DSSC, Spring Semester 2019

Outline

1. Inference by sampling
2. Markov Chain Monte Carlo
3. Hamiltonian Monte Carlo

APPROXIMATE INFERENCE BY SAMPLING

- Generate a sample from a given distribution
- Estimate integrals and expectations

BASIC SAMPLING

- Ancestral Sampling in Bayesian Networks
- Rejection Sampling
- Importance Sampling

Outline

1 Inference by sampling

2 Markov Chain Monte Carlo

3 Hamiltonian Monte Carlo

MARKOV CHAINS

Main idea: to sample from $p(x)$, simulate from an ergodic Markov Chain with stationary probability $p(x)$.

- Definition of Markov Chains
- Stationary distributions and ergodicity
- Reversible MC and detailed balance condition

Basics

Main idea: to sample from $p(x)$, simulate from an ergodic Markov Chain with stationary probability $p(x)$.

- Definition of Markov Chains
- Stationary distributions and ergodicity
- Reversible MC and detailed balance condition
- Metropolis Hastings acceptance criterion

MCMC EXAMPLE - UNIMODAL

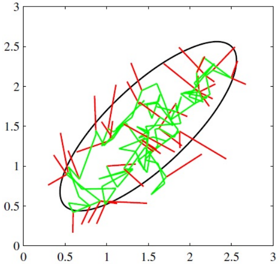

Figure 11.9 from Bishop: A simple illustration using Metropolis algorithm to sample from a Gaussian distribution whose one standard-deviation contour is shown by the ellipse. The proposal distribution is an isotropic Gaussian distribution whose standard deviation is 0.2. Steps that are accepted are shown as green lines, and rejected steps are shown in red. A total of 150 candidate samples are generated, of which 43 are rejected.

MCMC EXAMPLE - MULTIMODAL

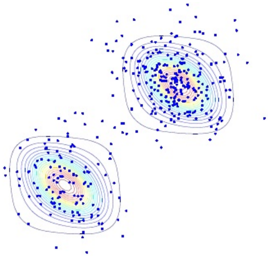

Figure 27.8 from Barber: Metropolis-Hastings samples from a bi-variate distribution $p(x_1, x_2)$ using a proposal $\tilde{q}(\mathbf{x}'|\mathbf{x}) = N(\mathbf{x}'|\mathbf{x}, \mathbf{I})$. We also plot the iso-probability contours of p. Although $p(\mathbf{x})$ is multi-modal, the dimensionality is low enough and the modes sufficiently close such that a simple Gaussian proposal distribution is able to bridge the two modes. In higher dimensions, such multi-modality is more problematic.

Gibbs Sampling

Main idea: to sample from $p(x_1, \ldots, x_n)$, sample iteratively from 1 dimensional conditional distributions.

- Gibbs sampling as MCMC
- Metropolis-within-Gibbs

Gibbs Sampling - ergodicity

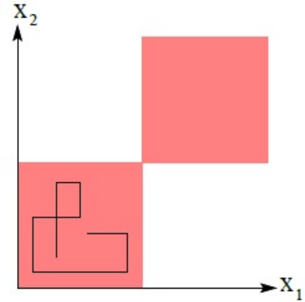

Figure 27.5 from Barber: A two dimensional distribution for which Gibbs sampling fails. The distribution has mass only in the shaded quadrants. Gibbs sampling proceeds from the l^{th} sample state (x_1^l, x_2^l) and then sampling from $p(x_2|x_1^l)$, which we write (x_1^{l+1}, x_2^{l+1}) where $x_1^{l+1} = x_1^l$. One then continues with a sample from $p(x_1|x_2 = x_2^{l+1})$, etc. If we start in the lower left quadrant and proceed this way, the upper right region is never explored.

GIBBS SAMPLING - CORRELATION

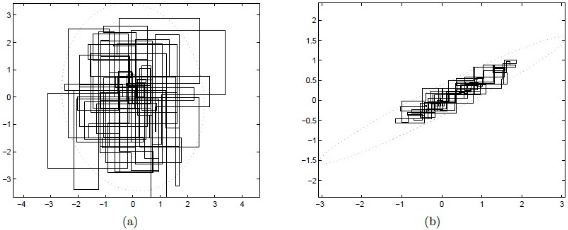

Figure 27.7 from Barber: Two hundred Gibbs samples for a two dimensional Gaussian. At each stage only a single component is updated. (a): For a Gaussian with low correlation, Gibbs sampling can move through the likely regions effectively. (b): For a strongly correlated Gaussian, Gibbs sampling is less effective and does not rapidly explore the likely regions.

Convergence diagnostic

How to check that a MC has reached stationarity

- Compare between variance and within variance between several simulations
- Compute index $\hat{R}$ and effective sample size

Example of MCMC run

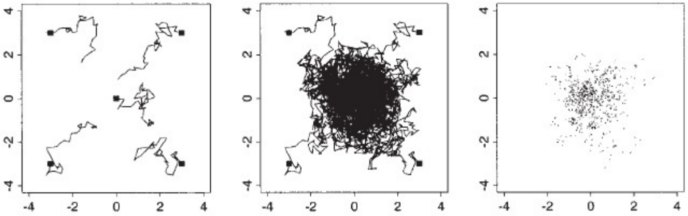

Figure 11.2 from Gelman et al. (2nd Edition): Five independent sequences of a Markov chain simulation for the bivariate unit normal distribution, with over-dispersed starting points indicated by solid squares. (a) After 50 iterations, the sequences are still far from convergence. (b) After 1000 iterations, the sequences are nearer to convergence. Figure (c) shows the iterates from the second halves of the sequences. The points in Figure (c) have been jittered so that steps in which the random walk stood still are not hidden.

CONVERGENCE DIAGNOSTIC - EXAMPLE

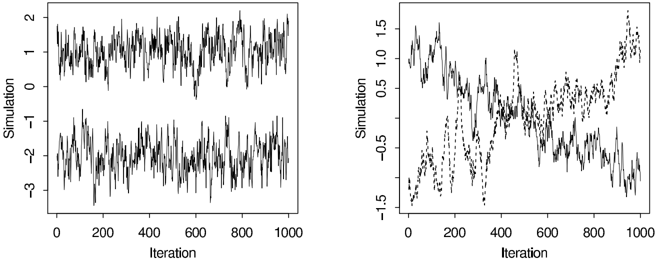

Figure 11.3 *Examples of two challenges in assessing convergence of iterative simulations. (a) In the left plot, either sequence alone looks stable, but the juxtaposition makes it clear that they have not converged to a common distribution. (b) In the right plot, the twosequences happen to cover a common distribution but neither sequence appears stationary. These graphs demonstrate the need to use between-sequence and also within-sequence information when assessing convergence.*

Outline

1. Inference by sampling
2. Markov Chain Monte Carlo
3. Hamiltonian Monte Carlo

HAMILTONIAN MONTE CARLO - ALGORITHM

Algorithm 27.4 Hybrid Monte Carlo sampling

Start from $\mathbf{x}^1$
for $i = 1$ to L **do**
  Draw a new sample $\mathbf{y}$ from $p(\mathbf{y})$.
  Choose a random (forwards or backwards) trajectory direction.
  Starting from $\mathbf{x}^i, \mathbf{y}$, follow Hamiltonian dynamics for a fixed number of steps, giving a candidate $\mathbf{x}', \mathbf{y}'$.
  Accept the candidate $\mathbf{x}^{i+1} = \mathbf{x}'$ if $H(\mathbf{x}', \mathbf{y}') > H(\mathbf{x}, \mathbf{y})$, otherwise accept it with probability $\exp(H(\mathbf{x}', \mathbf{y}') - H(\mathbf{x}, \mathbf{y}))$.
  If rejected, we take the sample as $\mathbf{x}^{i+1} = \mathbf{x}^i$.
end for

– Algorithm 27.4 from Barber

HAMILTONIAN MONTE CARLO - EXAMPLE

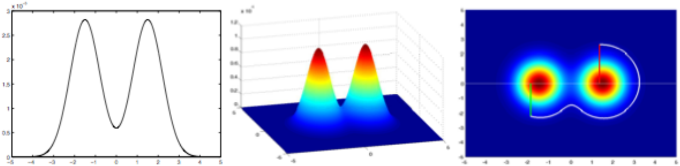

Figure 27.9 from Barber: Hybrid Monte Carlo. (a): Multi-modal distribution $p(x)$ for which we desire samples. (b): HMC forms the joint distribution $p(x)p(y)$ where $p(y)$ is Gaussian. (c): This is a plot of (b) from above. Starting from the point x, we first draw a y from the Gaussian $p(y)$, giving a point (x, y), given by the green line. Then we use Hamiltonian dynamics (white line) to traverse the distribution at roughly constant energy for a fixed number of steps, giving x', y'. We accept this point if $H(x', y') > H(x, y')$ and make the new sample x' (red line). Otherwise this candidate is accepted with probability $\exp(H(x', y') - H(x, y'))$. If rejected the new sample x' is taken as a copy of x.