Statistical learning 4: Markov Chain Monte Carlo

RG sections 4.5, 9.1 and 9.3

Jakob Söhl

Delft University of Technology

Bayesian computation

- Except for a few special cases, the posterior density

$$\pi(\theta) := p(\theta \mid x) = \frac{p(x \mid \theta)p(\theta)}{\int p(x \mid \theta)p(\theta)d\theta}.$$

  is intractable.
- Computing the normalising constant requires (possibly high-dimensional) integration.
- Markov Chain Monte Carlo methods is a collection of techniques for obtaining (dependent) samples from the posterior distribution.
- Main algorithm: Metropolis-Hastings (MH) algorithm.

Metropolis–Hastings algorithm (1953, 1970)

Goal: obtain samples from a target density $\pi(\theta)$ with $\theta \in \Omega \subset \mathbb{R}^d$.

For ease of exposition: first assume Ω is finite.

- Input: an irreducible Markov chain on Ω, say with transition probabilities $q(\cdot, \cdot)$.

  All states communicate: with positive probability state j can be reached from state i.

- Output: a Markov chain $\{\Theta_n\}$ that has π as invariant distribution and

$$\frac{1}{N} \sum_{n=1}^{N} g(\Theta_n) \xrightarrow{\text{a.s.}} \mathbb{E}_\pi g(\Theta),$$

  for functions g for which the right-hand-side is finite
  (if $g(\theta) = \theta$ the RHS is the posterior mean).

There is huge freedom in choosing q.

Metropolis–Hastings algorithm

Define a Markov chain on Ω which evolves $\theta_n = \theta$ to θ_{n+1} as follows

1. propose θ° from a proposal density $q(\theta, \cdot)$;
2. Compute
$$\alpha(\theta, \theta^\circ) = \min\left(1, \frac{\pi(\theta^\circ)}{\pi(\theta)} \frac{q(\theta^\circ, \theta)}{q(\theta, \theta^\circ)}\right).$$
3. Set
$$\theta_{n+1} = \begin{cases} \theta^\circ & \text{with probability } \alpha(\theta, \theta^\circ) \\ \theta & \text{with probability } 1 - \alpha(\theta, \theta^\circ) \end{cases}.$$

It suffices to know π up to a proportionality constant.

Input: proposal q, output: proposal $\bar{q}$, which is q adjusted by the MH-acceptance rule in steps (2) and (3).

Metropolis–Hastings algorithm: discrete case

If the proposed state θ° satisfies $\theta^\circ \neq \theta$, then the probability of the chain proposing and accepting θ° is given by

$$\bar{q}(\theta, \theta^\circ) = q(\theta, \theta^\circ)\alpha(\theta, \theta^\circ).$$

This implies

$$\begin{aligned} \pi(\theta)\bar{q}(\theta, \theta^\circ) &= \pi(\theta)q(\theta, \theta^\circ)\min\left(1, \frac{\pi(\theta^\circ)}{\pi(\theta)}\frac{q(\theta^\circ, \theta)}{q(\theta, \theta^\circ)}\right) \\ &= \min\left(\pi(\theta)q(\theta, \theta^\circ), \pi(\theta^\circ)q(\theta^\circ, \theta)\right) = \pi(\theta^\circ)\bar{q}(\theta^\circ, \theta). \end{aligned}$$

This trivially also holds when $\theta^\circ = \theta$.

Summing over θ gives

$$\sum_\theta \pi(\theta)\bar{q}(\theta, \theta^\circ) = \pi(\theta^\circ).$$

- In case of a "continuous" target distribution, the summation has to be replaced with an integral.

$$\int_\Omega \pi(\theta)\bar{q}(\theta,\theta^\circ)\,\mathrm{d}\theta = \pi(\theta^\circ).$$

- This says that when $\theta \sim \pi$ and we evolve the chain for one step, then $\theta^\circ \sim \pi$.
- Put differently, the MH-chain **preserves** π.
- π is an **invariant distribution** of the MH-chain.
- Under some weak conditions, the law of large numbers then holds

$$\frac{1}{N}\sum_{n=1}^{N} g(\Theta_n) \xrightarrow{\text{a.s.}} \mathbb{E}_\pi g(\Theta).$$

Construction of the proposal kernel: some examples

1. Random walk proposals: choose tuning parameter $\sigma > 0$ and set

$$\theta^\circ = \theta + \sigma Z, \qquad \text{with} \qquad Z \sim N(0,1).$$

   σ should neither be too big nor too small.

2. Independent proposals: Take $q(\theta, \cdot) = h(\cdot)$.

$$\alpha(\theta, \theta^\circ) = \min\left(1, \frac{\pi(\theta^\circ)}{\pi(\theta)} \frac{h(\theta)}{h(\theta^\circ)}\right).$$

   h ideally resembles π.

3. Metropolis Adjusted Langevin Algorithm (MALA): $Z \sim N(0,1)$

$$\theta^\circ = \theta + \frac{1}{2} A\delta \nabla \log \pi(\theta) + \sqrt{\delta A} Z.$$

   Advanced: makes sense since π is invariant for the Langevin diffusion

$$d\theta_t = \frac{1}{2} A \nabla \log \pi(\theta_t) dt + \sqrt{A} dW_t.$$

A simple illustration of the MH-algorithm

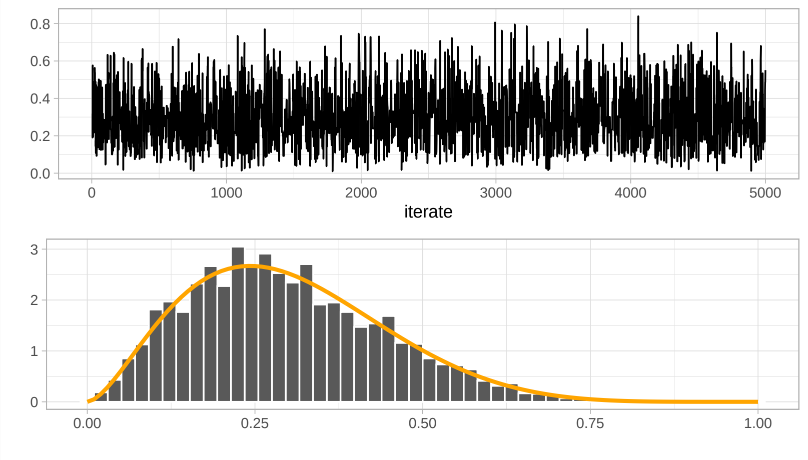
Suppose we wish to simulate from the $Beta(a, b)$-distribution.

- There exist direct ways for simulating *independent* realisations of the beta distribution.
- Use MH-algorithm with
  - independent $U(0, 1)$-proposals, **Independent MH algorithm**;
  - random walk type proposals of the form $\theta^\circ := \theta + U(-\eta, \eta)$, with η a tuning parameter, **Random Walk MH algorithm**.

Results for independent MH algorithm

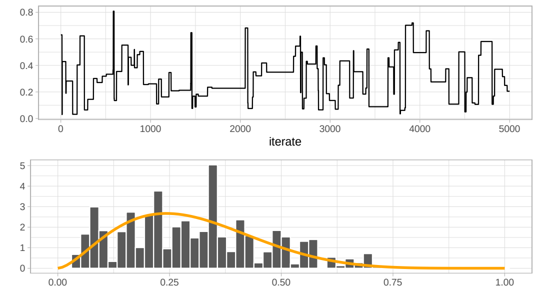
Target density: probability density of $Beta(a = 2.7; b = 6.3)$-distribution.

- Independently propose from $Unif(0, 1)$-distribution.



Results for random-walk MH algorithm

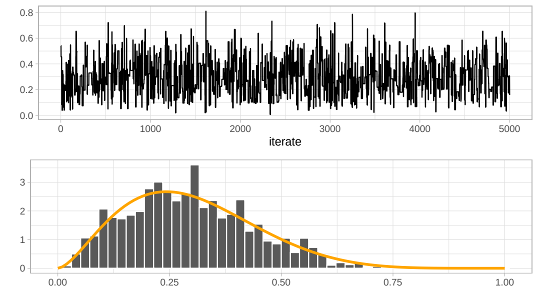
Target density: probability density of $Beta(a = 2.7; b = 6.3)$-distribution.
Random walk: if θ is the current iterate, then propose $\theta + U(-\eta, \eta)$



Results for $\eta = 10$. Steps are too big.
Average acceptance probability equals 0.023.

Results for random-walk MH algorithm

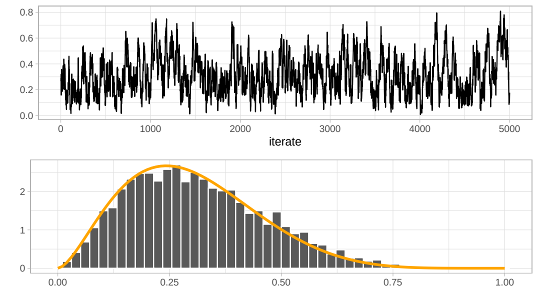
Target density: probability density of $Beta(a = 2.7; b = 6.3)$-distribution.
Random walk: if θ is the current iterate, then propose $\theta + U(-\eta, \eta)$



Results for $\eta = 1$. Steps are still too big.
Average acceptance probability equals 0.224.

Results for random-walk MH algorithm

Target density: probability density of $Beta(a = 2.7; b = 6.3)$-distribution.
Random walk: if θ is the current iterate, then propose $\theta + U(-\eta, \eta)$



Results for $\eta = 0.1$. Steps are a bit too small.
Average acceptance probability equals 0.844.

Second simple illustration of the MH-algorithm

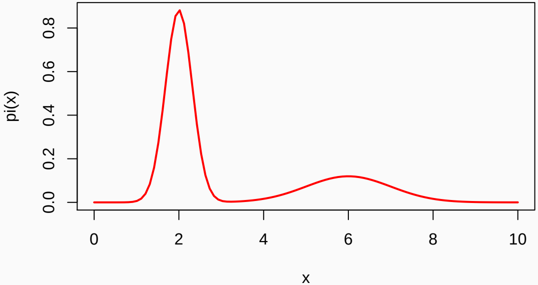
Suppose we wish to simulate from

$$\pi(\theta) = 0.7\varphi(\theta; 2, 0.1) + 0.3\varphi(\theta; 6, 1).$$

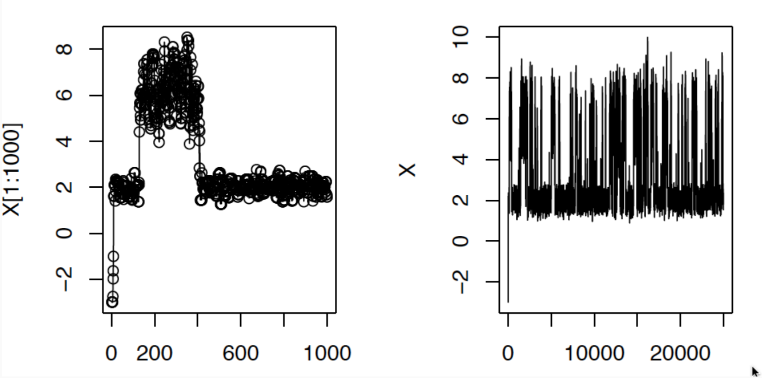
There is a simple direct way to sampling from this density.

Use MH-algorithm with random walk proposals

$$\theta^\circ = \theta + \sigma Z, \qquad \text{with} \qquad Z \sim N(0, 1).$$

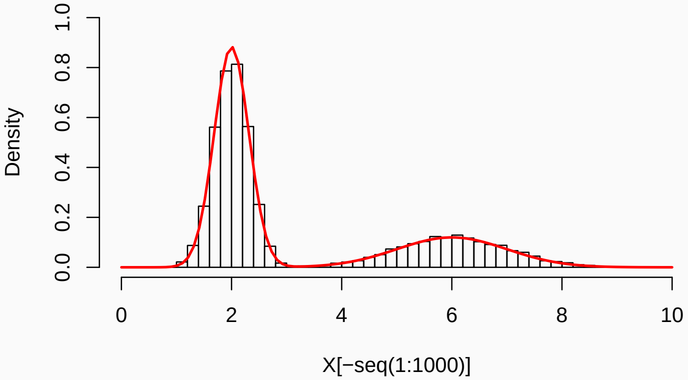


Second simple illustration of the MH-algorithm



Random walk proposals ($\sigma = 1$): average acceptance probability 0.46.

Second simple illustration of the MH-algorithm



Cycles and mixtures of MH-kernels

- **Cycle:** Suppose $\bar{q}_1$ and $\bar{q}_2$ are invariant for the density π, then so is

$$\bar{q}(\theta, \theta^\circ) = \sum_{\psi} \bar{q}_1(\theta, \psi)\bar{q}_2(\psi, \theta^\circ).$$

  Direct extension to cycling with more than two kernels.

- **Mixture:** If each of the kernels $\bar{q}_i, i = 1, \ldots, p$ is invariant for the density π, then so is

$$\bar{q}(\theta, \theta^\circ) = \sum_{i=1}^{p} w_i \bar{q}_i(\theta, \theta^\circ),$$

  where $\sum_{i=1}^{p} w_i = 1$. So we may randomly pick an update mechanism out of p of those that are invariant for π.

Useful if θ is high-dimensional. Then some of the kernels may focus on subsets of $\text{supp}\,(\pi)$.

Gibbs sampler

Goal: sample from $\pi(\theta_1, \ldots, \theta_p)$.

Fixed scan Gibbs sampler. Iterate:

- Sample $\theta_1 \sim \pi(\theta_1 \mid \theta_{-1})$
- Sample $\theta_2 \sim \pi(\theta_2 \mid \theta_{-2})$
- ...
- Sample $\theta_p \sim \pi(\theta_p \mid \theta_{-p})$

Known as iteratively sampling from full conditionals and is a special case of MH (where acceptance probability equals one).

Random scan Gibbs sampler. Iterate:

- Randomly choose an index i from $\{1, \ldots, n\}$
- Sample $\theta_i \sim \pi(\theta_i \mid \theta_{-i})$

Probabilistic programming

A Bayesian hierarchical model consists of

1. observed variables
2. non-observed variables

Form the hierarchical scheme the joint likelihood of all variables is extracted. This is all that is needed for advanced samplers like HMC (Hamiltonian Monte Carlo). Examples:

- BUGS (somewhat old now)
- STAN
- Turing (within Julia language)

Crucially these depend on differentiable programming. Strong influence from computer science!

MCMC for logistic regression

Back to computing the posterior for logistic regression

Define $\psi : \mathbb{R} \to (0,1)$ is defined by

$$\psi(z) = \frac{1}{1+e^{-z}}.$$

Assume

$$y_i \mid \theta \overset{\text{ind}}{\sim} Ber\left(p_i\right), \quad \text{with} \quad p_i = \psi(\theta^T x_i)$$

Posterior density:

$$p(\theta \mid y, X) \propto p(\theta) \prod_{i=1}^{n} p_i^{y_i} (1-p_i)^{1-y_i}.$$

Take $N(0, \sigma^2 I)$ prior on θ.

Random walk proposals

Assume random Walk proposals

$$q(\theta, \theta^{\circ}) = \varphi(\theta^{\circ}; \theta, \sigma^2_{\mathrm{prop}} I)$$

At each iteration accept with probability $\min(1, A)$ where

$$A = \frac{p(\theta^{\circ} \mid y, X)}{p(\theta \mid y, X)} \frac{q(\theta^{\circ}, \theta)}{q(\theta, \theta^{\circ})}.$$

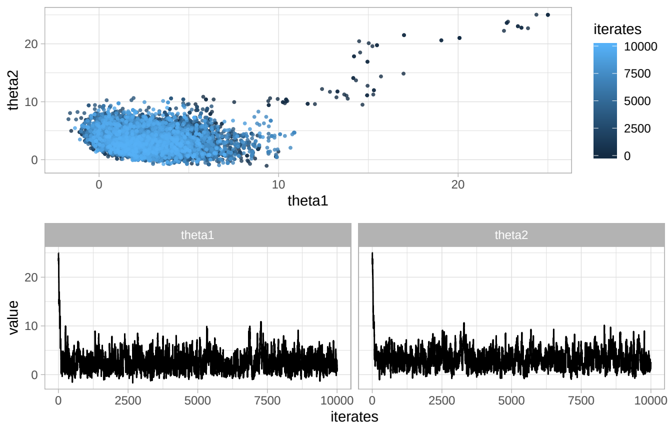
Only requires tuning of σ^2_{prop}.

MCMC

Script `logisticexample.jl`.

1. MCMC with either Random Walk (RW) or MALA.
2. ITER iterations, of which (by default) BURNIN $=$ ITER$/2$ are dropped.
3. Numerically, the only tricky things is to avoid evaluating the log at zero.
4. Choose $\theta \sim N(0, 10 \cdot I_2)$ as prior.

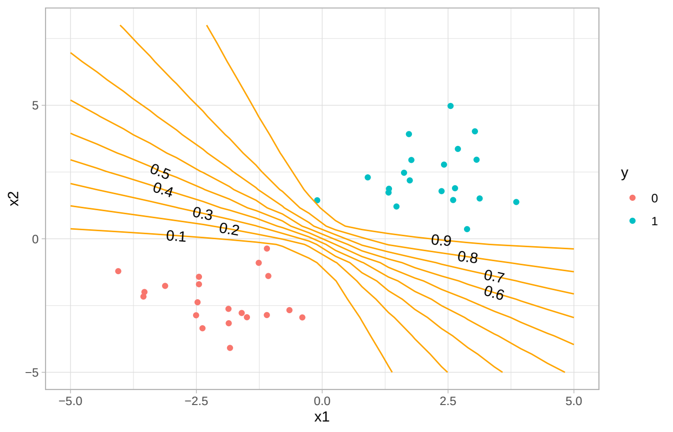
All iterates



Contour map

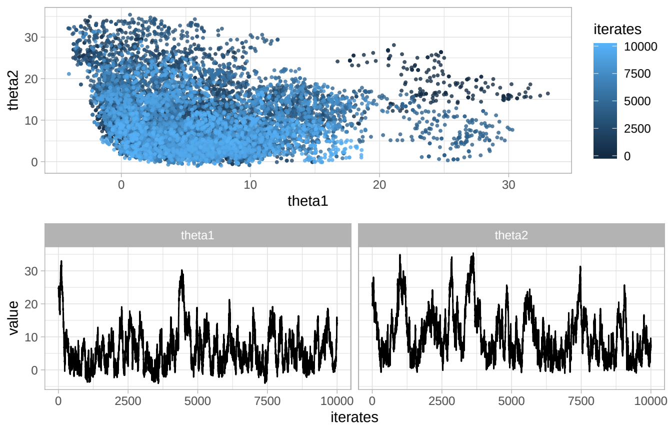
Consider

$$x_{\text{new}} \mapsto \mathbb{P}(Y_{\text{new}} = 1 \mid x_{\text{new}}, X, y).$$



Prior sensitivity

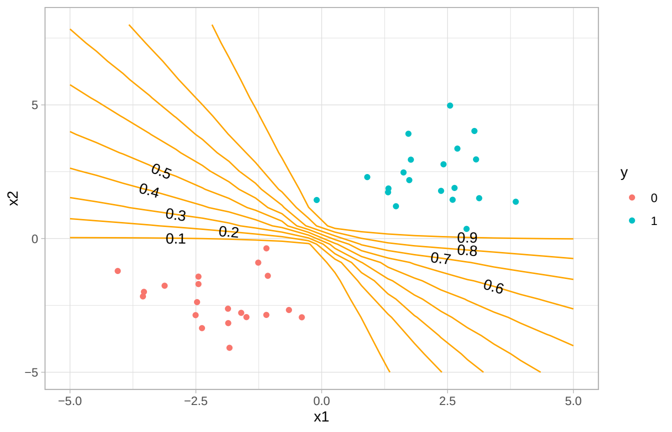
All iterates: prior stdev 3 times larger



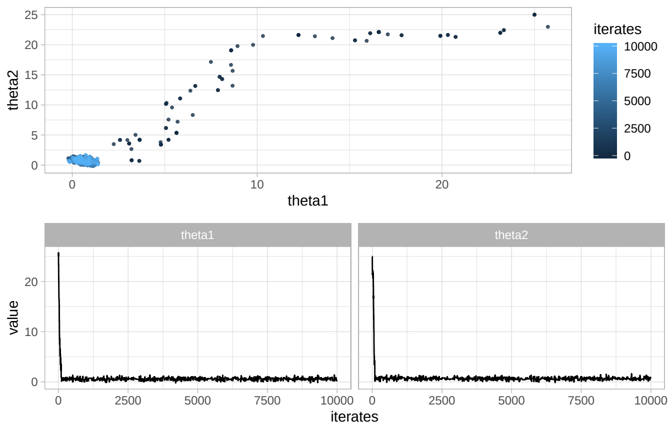
Contour map: prior stdev 3 times larger

Consider

$$x_{\text{new}} \mapsto \mathbb{P}(Y_{\text{new}} = 1 \mid x_{\text{new}}, X, y).$$



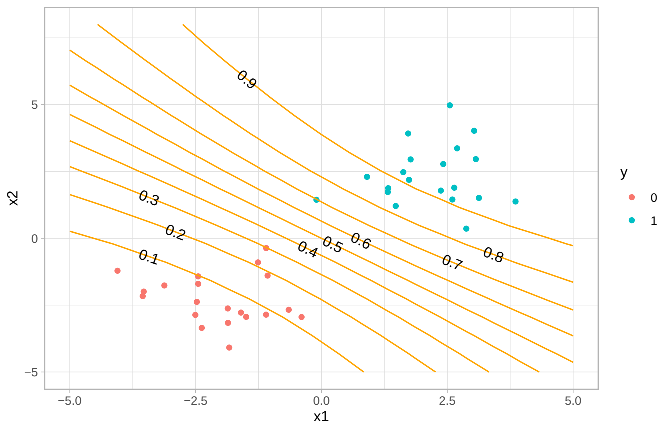
All iterates: prior stdev 10 times smaller



Contour map: prior stdev 10 times smaller

Consider

$$x_{\text{new}} \mapsto \mathbb{P}(Y_{\text{new}} = 1 \mid x_{\text{new}}, X, y).$$



Dealing with prior sensitivity

- Use an empirical Bayes approach, where σ^2 is estimated from the density of $p(y_1, \ldots, y_n)$.
  This is sometimes tractable, but here $p(y_1, \ldots, y_n)$ is not.
- Add an additional layer: hierarchical Bayes approach.

We pursue the second option. Write $\tau = \sigma^2$.

$$\begin{aligned} y_i \mid \theta &\overset{\text{ind}}{\sim} Ber\left(p_i\right), \quad \text{with} \quad p_i = \psi(\theta^T x_i) \\ \theta \mid \tau &\sim N(0, \tau I) \\ \tau &\sim InvGa\left(A, B\right) \end{aligned}$$

We use the inverse gamma distribution, as it has computational advantages in using the Gibbs sampler.

Gibbs sampler: iteratively update (write $\boldsymbol{y} = (y_1, \ldots, y_n)$)

- $\theta \mid \tau, \boldsymbol{y}$ (use MH-step as before)
- $\tau \mid \theta, \boldsymbol{y}$.

Note that

$$p(\tau \mid \theta, \boldsymbol{y}) \propto p(\boldsymbol{y}, \theta, \tau) = p(\boldsymbol{y} \mid \theta, \tau) p(\theta \mid \tau) p(\tau) \propto p(\theta \mid \tau) p(\tau).$$

Therefore (assume $\theta \in \mathbb{R}^k$)

$$\begin{aligned} p(\tau \mid \theta, \boldsymbol{y}) &\propto \tau^{-k/2} \exp\left(-\frac{1}{2\tau}\|\theta\|^2\right) \tau^{-A-1} e^{-B/\tau} \mathbf{1}_{(0,\infty)}(\tau) \\ &\propto \tau^{-(A+k/2)-1} \exp\left(-\frac{B + \|\theta\|^2/2}{\tau}\right) \mathbf{1}_{(0,\infty)}(\tau) \end{aligned}$$

Thus

$$\tau \mid \theta, \boldsymbol{y} \sim InvGa\left(A + k/2, B + \|\theta\|^2/2\right).$$

We say that the prior on τ is *partially conjugate*.