Statistics 568 Bayesian Analysis

Markov Chain Monte Carlo

Ruobin Gong

Department of Statistics
Rutgers University

Week 7 - 03/04/21

Recap: approximating expectations

Let $f(\theta)$ be a quantity of interest. Its posterior expectation

$$E\left(f(\theta) \mid y\right) = \int f(\theta) p(\theta|y) d\theta,$$

$$\text{where} \quad p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta}$$

We can (usually) easily evaluate the unnormalized posterior

$$q(\theta|y) = p(y|\theta)p(\theta)$$

for any θ, but not necessarily the integral $\int p(y|\theta)p(\theta)d\theta$.

Use Monte Carlo methods to sample from $p(\theta^{(s)}|y)$ using only $q(\theta^{(s)}|y)$:

$$E\left(f(\theta) \mid y\right) \approx \frac{1}{S}\sum_{s=1}^{S} f(\theta^{(s)})$$

Recap: approximating expectations

A few different approaches of computation in the Bayesian context:

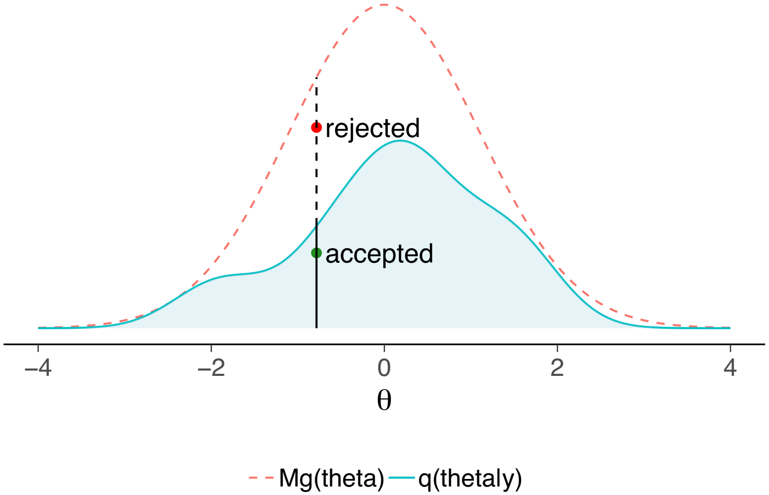
- Conjugate priors and analytic solutions (Ch 1-5)
- Grid integration and other quadrature rules (Ch 3, 10)
- **Independent Monte Carlo, rejection and importance sampling** (Ch 10)
- **Markov Chain Monte Carlo** (Ch 11-12)
- Other distributional approximations (Laplace, variational Bayes, expectation propagation) (Ch 4, 13)

Rejection sampling

Choose a proposal distribution $g(\cdot)$ such that it can form an envelope over the (unormalized) target distribution $q(\cdot) = cp(\cdot)$, i.e. there exists a *covering constant* $M < \infty$ s.t. for all θ,

$$Mg(\theta) \geq q(\theta).$$

Below, the unormalized target $q(\cdot)$ is the blue density. Scaled proposal $Mg(\cdot)$ is the dashed red density.

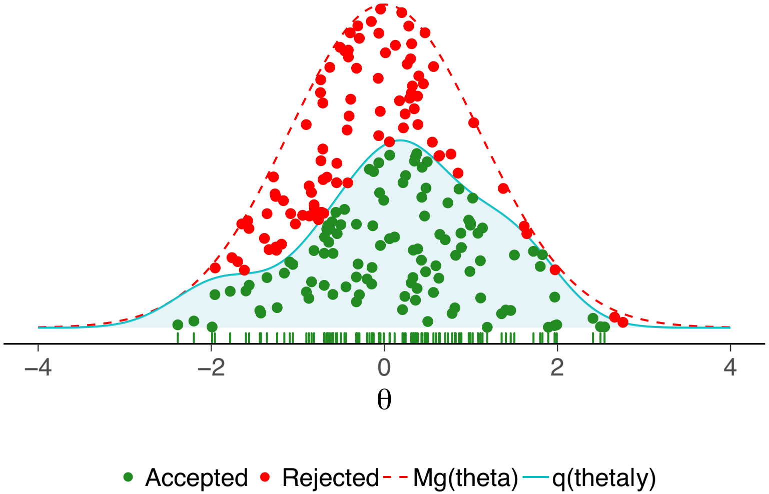


Rejection sampling

(von Neumann, 1951). At step $s = 1, \ldots, S$:

- Draw a sample $\theta^{(s)} \sim g(\cdot)$;
- Accept $\theta^{(s)}$ with probability $r = \frac{q(\theta^{(s)})}{Mg(\theta^{(s)})}$, otherwise go back to the previous step.

The accepted samples $\theta^{(s)}$ follow the target distribution $p(\cdot)$.



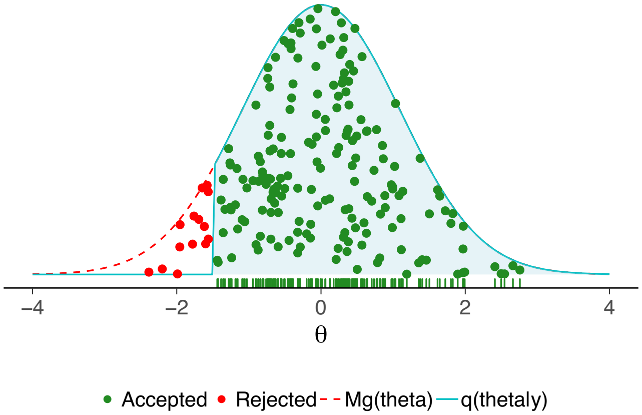
Example: Truncated Gaussian

The (unormalized) target is

$$q(\theta) = \phi(\theta) \mathbf{1}\{\theta > c\}$$

where ϕ is the standard Normal density.

- $c \leq 0$: choose $g(\theta) = \phi(\theta)$. The covering constant $M = 1$. Efficiency is at least 50% (depending on c).



- $c > 0$, especially $c \gg 0$: this choice can be inefficient.

Example: Truncated Gaussian (Homework 6)

Consider proposal distribution in the form of the exponential density (with parameter λ as something we can choose):

$$g(\theta) = \lambda \exp(-\lambda\theta).$$

Need to find smallest M such that

$$p(\theta) = \frac{\phi(\theta + c)}{1 - \Phi(c)} \leq \underbrace{M \cdot \lambda \exp(-\lambda\theta)}_{M \cdot g(\theta)}, \ \forall \theta > 0,$$

$$M \geq \frac{p(\theta)}{g(\theta)} = \odot$$

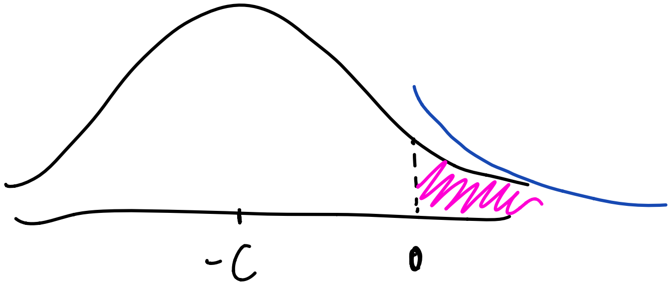
for all x.

which gives the minimal M^*:

$$M^* = \frac{\exp\left\{\left(\lambda^2 - 2\lambda c\right)/2\right\}}{\sqrt{2\pi}\lambda\left(1 - \Phi(c)\right)}.$$

x maximizes $\odot$

at $x^* = \lambda - c$



Example: Truncated Gaussian (Homework 6)

The best acceptance rate for this proposal is $1/M^*$;

$$\Pr(I=1) = \int \Pr(I=1 \mid x)\, g(x)\, dx = \int \frac{p(x)}{M^* \cancel{g(x)}} \cancel{g(x)}\, dx = \frac{1}{M^*}$$

The best λ^* that meets this best acceptance rate is

$$\lambda^* = \left(c + \sqrt{c^2 + 4}\right)/2.$$

$$\frac{\partial}{\partial \lambda}\left[M^*(\lambda)\right]^{-1} = 0$$

$$\Rightarrow \lambda^* = \frac{c + \sqrt{c^2+4}}{2} > 0.$$

Recap: Importance Sampling

Posterior expectation

$$E\left(f(\theta) \mid y\right) = \int f(\theta)p(\theta)d\theta = \frac{\int f(\theta)q(\theta)d\theta}{\int q(\theta)d\theta} = \frac{\int f(\theta)\frac{\overset{cp(\theta)}{q(\theta)}}{g(\theta)}g\left(\theta\right)d\theta}{\int \underset{cp(\theta)}{\frac{q(\theta)}{g(\theta)}}g\left(\theta\right)d\theta}$$

can be estimated using S draws $\theta^{(1)}, \ldots, \theta^{(S)}$ from $g\left(\cdot\right)$ by the expression

$$\hat{f} = \frac{\sum_s w^{(s)} f\left(\theta^{(s)}\right)}{\sum_s w^{(s)}},$$

where

$$w^{(s)} = \frac{q(\theta^{(s)})}{g(\theta^{(s)})} \propto \frac{p(\theta^{(s)})}{g(\theta^{(s)})}$$

are the **importance weights**. See R for an illustration.

Importance Sampling

- In general,

$$\hat{f} = \frac{\sum_s w^{(s)} f(\theta^{(s)})}{\sum_s w^{(s)}}$$

is **biased** for $E\left(f(\theta) \mid y\right)$, due to the estimated weights in the denominator.

- If the ratio $p(\cdot)/g(\cdot)$ is exactly known, then we have an alternative **unbiased** estimate

$$\tilde{f} = \frac{1}{S} \sum_s w^{(s)} f(\theta^{(s)}).$$

However, $\hat{f}$ often has a smaller MSE than $\tilde{f}$.

Choice of IS proposal

The importance weights

$$w^{(s)} = \frac{q(\theta^{(s)})}{g(\theta^{(s)})} \propto \frac{p(\theta^{(s)})}{g(\theta^{(s)})}$$

are used to correct for the “bias” in drawing the θ’s from the proposal $g(\cdot)$ as opposed to the target $p(\cdot)$.

- A good choice for the proposal $g(\cdot)$ is one that’s close in shape to $f(\cdot)p(\cdot)$.
- Unlike rejection sampling, the proposal $g(\cdot)$ need not cover the target $p(\cdot)$. However, it is desirable that it has a longer tail than $p(\cdot)$, to avoid importance weights with large variance.

Effective sample size of IS

Suppose we can compute the normalized target, and the naturally normalized importance weights can be obtained as

$$w_0^{(s)} = \frac{p(\theta^{(s)})}{g(\theta^{(s)})}.$$

Note that here, $E_g\left(w_0\left(\theta\right)\right) = 1$.

The **effective sample size** $\left(ess\right)$ of S independent random θ samples generated from target g is defined as

$$ess\left(S\right) = \frac{S}{1 + var_g\left(w_0\left(\theta\right)\right)}.$$

Effective sample size

If only the unormalized q is known, we cannot compute w_0, only

$$w^{(s)} = \frac{q(\theta^{(s)})}{g(\theta^{(s)})}.$$

The variance of the normalized weights $var_g\left(w_0\left(\theta\right)\right)$, if finite, needs to be estimated with the **coefficient of variation** of the unormalized weights. That is,

$$\hat{ess}\left(S\right) = \frac{S}{1 + cv^2\left(w\right)},$$

where $\bar{w} = \frac{1}{S}\sum_s w^{(s)}$, and

$$cv^2\left(w\right) = \frac{\sum_s \left(w^{(s)} - \bar{w}\right)^2}{\left(S-1\right)\bar{w}^2}.$$

Effective sample size

Exercise. Compare with the *ess* approximation in the textbook, (10.4) on page 266:

$$\tilde{ess}\left(S\right) = \frac{1}{\sum_s \left(\tilde{w}^{(s)}\right)^2},$$

where

$$\tilde{w}^{(s)} = \frac{w^{(s)}}{\sum_s w^{(s)}}$$

Note 1: The textbook has a typo here. There is no S multiplier in the expression of $\tilde{w}$.
Note 2: The approximated effective sample size is small if there are a few extremely large weights. These few but large weights make the estimate itself very noisy, so take it with a grain of salt!

Markov Chain

- Definition. Let $\{x^{(0)}, x^{(1)}, \ldots\}$ be a sequence of random variables defined on a finite state space $\mathcal{X}$. This sequence is called a **Markov chain** if it satisfies the **Markov property**:

$$Pr(x^{(t+1)} \mid x^{(t)}, \ldots, x^{(0)}) = Pr(x^{(t+1)} \mid x^{(t)}),$$

  i.e. the probability of the next event depends only on the state attained in the previous event.

- The **transition function** is

$$T_t(b \mid a) := Pr(x^{(t+1)} = b \mid x^{(t)} = a),$$

  with $\sum_b T_t(b \mid a) = 1$ for all a. T_t can often be time-homogeneous (i.e. same for all t), written as T.

- When $\mathcal{X}$ is continuous, the transition probability function is replaced by the *transition density function*, and summations replaced by integrations.

Markov Chain

The name is due to Andrey Markov, who proved weak law of large numbers and central limit theorem for certain dependent random sequences.

Examples of a Markov chain?

$$X_{t+1} = X_t + \epsilon \quad , \quad \epsilon \sim [0, \sigma^2].$$

$$P(X_{t+1} = b \mid X_t = a, X_{t-1} = \dots \quad)$$

$$= P(X_{t+1} = b \mid X_t = a) = P(\epsilon = b - a)$$

Markov chain Monte Carlo (MCMC)

MCMC for Bayesian computation:

- Produce draws $\theta^{(t)}$ given $\theta^{(t-1)}$ from a Markov chain, constructed in such a way so that its stationary distribution is the target distribution $p(\theta \mid y)$.
- Choice of transition distribution $T_t(\theta^{(t)} \mid \theta^{(t-1)})$ determines the stationary distribution.
- Chain has to be initialized with some starting point $\theta^{(0)}$. Need to run the chain for long enough so that the distribution of current draws is close enough to the stationary distribution (discarding the *burn-in* or *warm up*)

Markov chain Monte Carlo (MCMC)

Pros and cons of MCMC:

- + This is a very generic method to produce samples from a target distribution;
- + If appropriately behaved, central limit theorem holds for expectations;
- - Draws produced by MCMC are dependent. Monitoring convergence is crucial. Also, effective sample size calculation;
- - Construction of efficient Markov chains is not always easy.

The Gibbs sampler

Suppose the parameter vector θ has been divided into d subvectors, $\theta = (\theta_1, \ldots, \theta_d)$. Each iteration of the Gibbs sampler cycles through the subvectors of θ.

Basic algorithm is as follows:

1. Initialize $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_d^{(0)})$;
2. For $t = 1, 2, \ldots$:

   For $j = 1, \ldots, d$:
   Sample θ_j^t from
   $$p(\theta_j \mid \theta_{-j}^{t-1}, y),$$
   where $\theta_{-j}^{t-1} = (\theta_1^t, \ldots, \theta_{j-1}^t, \theta_{j+1}^{t-1}, \ldots, \theta_d^{t-1})$.

$t \backslash j$	1	2	$- - -$	d
0	$\theta_1^{(0)}$	$\theta_2^{(0)}$		$\theta_d^{(0)}$
1	$\theta_1^{(1)}$	$\theta_2^{(1)}$	$\theta_3^{(1)}$	
2				
3				
$\vdots$				

Example: Bivariate Gaussian

(You don't need a Gibbs sampler for this, but we use it to demonstrate the algorithm.)

Consider a single observation (y_1, y_2) from a bivariate normally distributed population with unknown mean $\theta = (\theta_1, \theta_2)$ and known covariance matrix $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.

With a uniform prior distribution, the posterior distribution is

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \sim N\left(\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

Example: Bivariate Gaussian

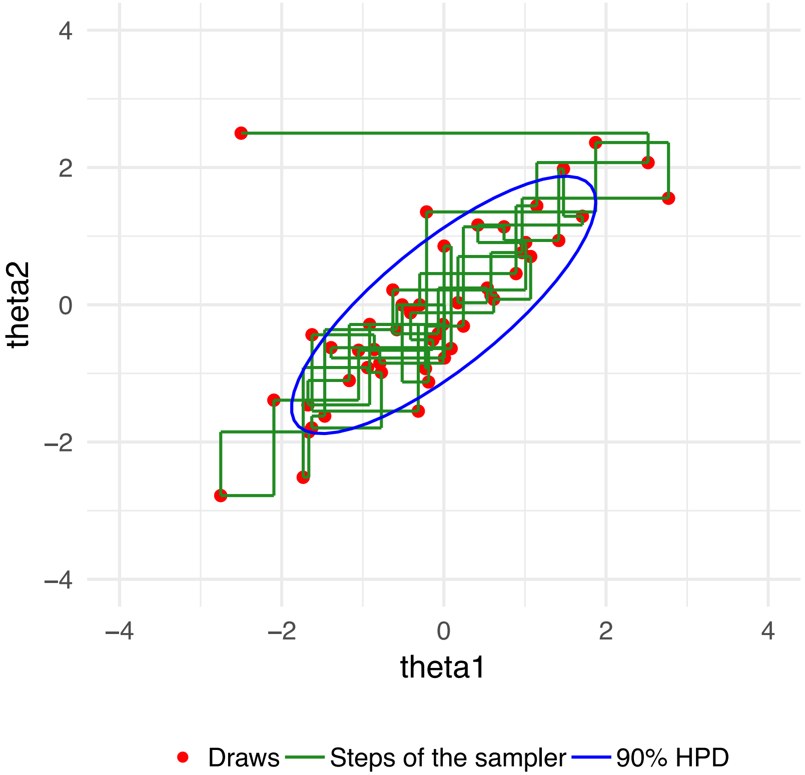
The conditional posteriors

$$\begin{aligned}
\theta_{\cancel{2}1} \mid \theta_{\cancel{1}2}, y &\sim N\left(y_1 + \rho\left(\theta_2 - y_2\right), 1 - \rho^2\right) \\
\theta_{\cancel{1}2} \mid \theta_{\cancel{2}1}, y &\sim N\left(y_2 + \rho\left(\theta_1 - y_1\right), 1 - \rho^2\right).
\end{aligned}$$

The Gibbs sampler proceeds by alternately sampling from these two normal distributions.

Example: Bivariate Gaussian

Here, $(y_1, y_2) = (0, 0)$, $\rho = 0.8$, initialized at $\theta^{(0)} = (-2.5, 2.5)$.

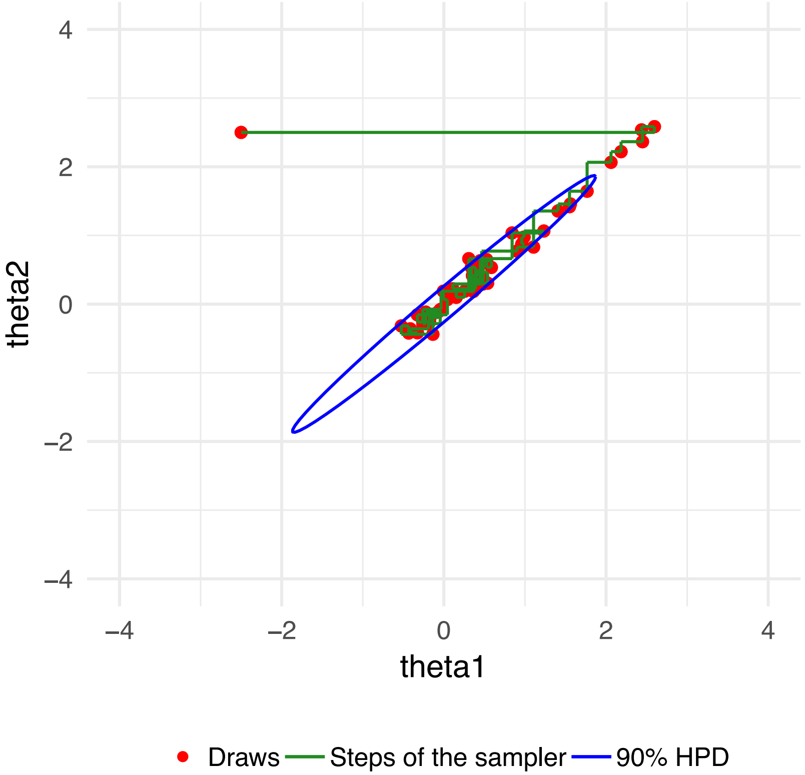


The Gibbs sampler

- With *conditionally* conjugate priors, the sampling from the conditional distributions is easy for wide range of models
- No algorithm parameters to tune (cf. proposal distribution in Metropolis algorithm)
- For not so easy conditionals, use e.g. inverse CDF
- Several parameters can be updated in blocks (*blocking*)
- Slow if parameters are highly dependent in the posterior.

The Gibbs sampler

For example, if $\rho = 0.99$ in the Bivariate Gaussian example:



The Metropolis algorithm

1. Initialize θ^0;
2. For $t = 1, 2, \ldots$,
   (a) pick a proposal θ^* from the proposal distribution $J_t(\theta^* \mid \theta^{t-1})$, also called the **jumping distribution**. In the Metropolis algorithm J is **symmetric**, i.e. $J_t(a \mid b) = J_t(b \mid a)$ for all a, b. This will later be generalized.
   (b) calculate acceptance probability

$$r = \min\left\{1, \frac{p(\theta^* \mid y)}{p(\theta^{t-1} \mid y)}\right\}$$

   (c) set

$$\theta^t = \begin{cases} \theta^* & \text{with probability } r \\ \theta^{t-1} & \text{otherwise} \end{cases}$$

   ie, if $p(\theta^* \mid y) > p(\theta^{t-1})$ accept the proposal always, and otherwise reject the proposal with probability r.

The Metropolis algorithm

$$r = \min\left\{1, \frac{p(\theta^* \mid y)}{p(\theta^{t-1} \mid y)}\right\}$$

Note. In Metropolis,

- Acceptance and rejection of a proposal both increment the time t by one. If rejected, the new state is the same as previous (cf. rejection sampling).
- Step (c) is executed by generating a random number from $U(0,1)$ and compare to r;
- $p(\theta^* \mid y)$ and $p(\theta^{t-1} \mid y)$ have the same normalization terms, and thus instead of $p(\cdot \mid y)$, unnormalized $q(\cdot \mid y)$ can be used since the normalization terms cancel out.

Example. Bivariate Gaussian (again.)

Consider proposal distribution

$$J_t(\theta^* \mid \theta^{t-1}) = \mathbb{N}(\theta^* \mid \theta^{t-1}, \sigma_p^2 I),$$

where σ_p^2 is set by choice. This is also called a **random walk Metropolis** algorithm, because the proposal is of the form

$$\theta^* = \theta^{t-1} + e,$$

where e is a zero-mean spherical random variable.

Demo at `http://elevanth.org/blog/2017/11/28/build-a-better-markov-chain/`.

Metropolis-Hastings Algorithm

Metropolis-Hastings is a generalization of the Metropolis algorithm with **non-symmetric proposal distributions**. That is, it's **not** required that $J_t(a \mid b) = J_t(b \mid a)$.

The acceptance probability now includes the ratio of proposal distributions (Metropolis et al. 1953; Hastings 1970):

$$r = \min\left\{1, \frac{p(\theta^* \mid y) J_t(\theta^{t-1} \mid \theta^*)}{p(\theta^{t-1} \mid y) J_t(\theta^* \mid \theta^{t-1})}\right\}$$

Metropolis-Hastings Algorithm

Other choices of acceptance probability:

- Barker (1965):

  $$r_B = \frac{p(\theta^* \mid y) J_t(\theta^{t-1} \mid \theta^*)}{p(\theta^* \mid y) J_t(\theta^{t-1} \mid \theta^*) + p(\theta^{t-1} \mid y) J_t(\theta^* \mid \theta^{t-1})},$$

  which is always between $[0, 1]$;
- Charles Stein (see Liu, 2001): more generally,

  $$r_C = \frac{\delta(\theta^{t-1} \mid \theta^*)}{p(\theta^{t-1} \mid y) J_t(\theta^* \mid \theta^{t-1})},$$

  where $\delta\,(b \mid a)$ is **any symmetric function** (in a, b) such that $r_C \leq 1$ for all a, b.

Choice of MH proposal.

- Ideal proposal distribution is the distribution itself, i.e. $J(\theta^*|\theta) \equiv p(\theta^*|y)$ for all θ.
  - Acceptance probability is 1, and draws are independent
  - Of course, this is not usually feasible (basically tautology)!
- Good proposal distribution resembles the target distribution
  - if the shape of the target distribution is unknown, usually normal or t distribution is used

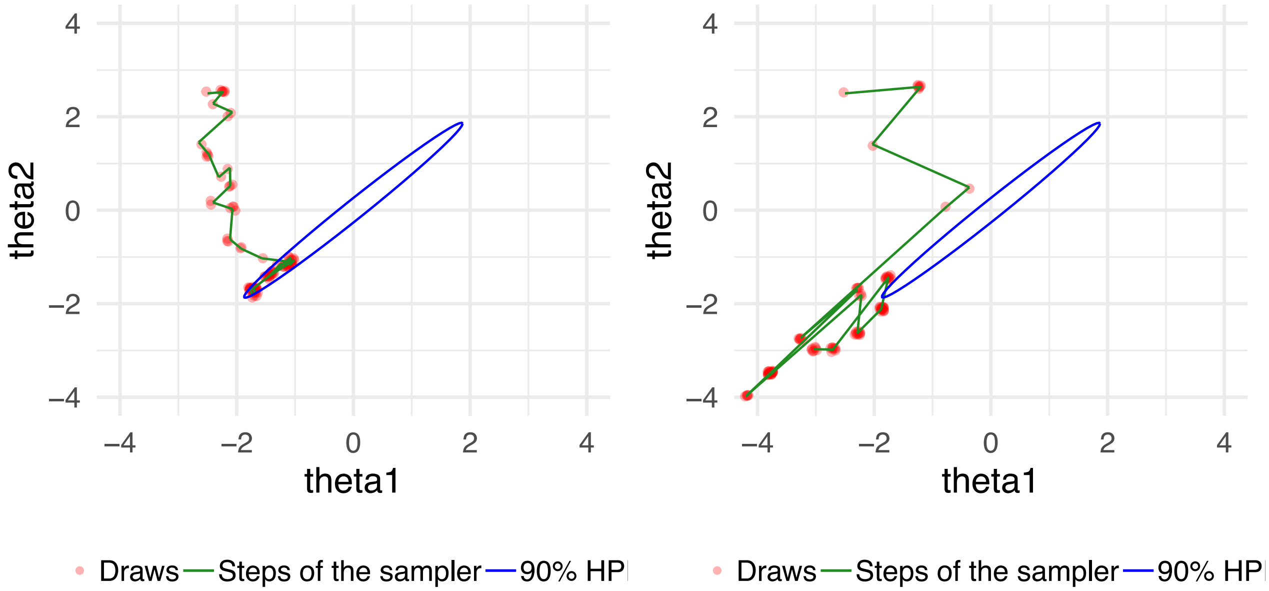
Choice of MH proposal.

- It should be easy to sample from $J(\theta^*|\theta)$ and to compute r for every θ
- After the shape has been selected, it is important to select the scale.
    - scale too small: many steps accepted, but the chain moves slowly due to small steps
    - scale too big: long steps proposed, but many of those rejected and again chain moves slowly

  MH usually doesn't scale well to high dimensions: if the shape of the proposal doesn't match the target, efficiency drops.
- We want to maintain a reasonable rejection rate. Generic rule for rejection rate is 60-90%, but depends on dimensionality and a specific algorithm variation.

Choice of MH proposal scale

Bivariate Gaussian: proposal scale too small (left) compared to the right



Gibbs as a special case of Metropolis-Hastings

Define a MH algorithm to consist of d steps at every iteration t. The jumping distribution at step j of iteration t: $J_{j,t}(\cdot \mid \cdot)$ only jumps along the jth subvector. More precisely, it does so with the conditional posterior density of θ_j given $\theta_{-j}^{(t-1)}$:

$$J_{j,t}^{Gibbs}\left(\theta^* \mid \theta^{(t-1)}\right) = \begin{cases} p\left(\theta_j^* \mid \theta_{-j}^{(t-1)}, y\right) & \text{if } \theta_{-j}^* = \theta_{-j}^{(t-1)} \\ 0 & \text{otherwise.} \end{cases}$$

Gibbs as a special case of Metropolis-Hastings

Viewed as an MH algorithm, the acceptance probability of the Gibbs sampler is

$$\begin{aligned} r &= \min\left\{1, \frac{p\left(\theta^* \mid y\right) / J_{j,t}^{Gibbs}\left(\theta^* \mid \theta^{(t-1)}\right)}{p\left(\theta^{(t-1)} \mid y\right) / J_{j,t}^{Gibbs}\left(\theta^{(t-1)} \mid \theta^*\right)}\right\} \\ &= \min\left\{1, \frac{p\left(\theta^* \mid y\right) / p\left(\theta_j^* \mid \theta_{-j}^{(t-1)}, y\right)}{p\left(\theta^{(t-1)} \mid y\right) / p\left(\theta_j^{(t-1)} \mid \theta_{-j}^{(t-1)}, y\right)}\right\} \\ &= \min\left\{1, \frac{p\left(\theta_{-j}^{(t-1)} \mid y\right)}{p\left(\theta_{-j}^{(t-1)} \mid y\right)}\right\} = 1. \end{aligned}$$

Using Gibbs and Metropolis as building blocks

The Gibbs sampler and the Metropolis algorithm can be used in combination to sample from complicated distributions.

- Gibbs: the simplest of the Markov chain simulation algorithms, first choice for conditionally conjugate models;
- Metropolis: can be used for models that are not conditionally conjugate;
- If some of the conditional posterior distributions in a model can be sampled from directly and some cannot, can use Gibbs where possible, and one-dimensional Metropolis updating otherwise;
- More generally, the parameters can be updated in **blocks**, each block altered using the Gibbs sampler or a Metropolis jump of the parameters within the block.

Why does the Metropolis algorithm work?

Intuitively, we see that we obtain more draws from the higher density areas, as jumps to higher density are always accepted, but only some of the jumps to the lower density are accepted.

To formally establish the legitimacy of the Metropolis algorithm, we need to show that

1 The simulated series is a Markov chain that has a **unique stationary distribution**;

2(!) This stationary distribution is the desired **target distribution**.

Stationary distribution of Markov chains

By the standard Markov chain theory, a chain becomes stationary at its stationary distribution if it is **irreducible**, **aperiodic**, and **recurrent** (not transient).

1. A Markov chain is *irreducible* if it has non-zero probability/density to move from one position in the state space to any other position in a finite number of steps;
2. A Markov chain is *aperiodic* if the maximum common divider of the number of steps it takes to come back to the starting point (any) is one.
3. A Markov chain is *recurrent* if the probability of its eventually returning to a state is one.

MH target distribution

As to why the stationary distribution of the Markov chain is the desired target distribution p? Need to show that

$$\int p(a)\, T_t(b \mid a)\, da = p(b)\,.$$

A *sufficient* (but more restrictive) condition is called **detailed balance**:

$$p(a)\, T_t(b \mid a) = p(b)\, T_t(a \mid b)\,.$$

Markov chains that satisfy detailed balance are called **reversible**.

If detailed balance holds:

$$\int p(a) T_t(b \mid a) da = \int p(b) T_t(a \mid b) da = p(b) \int T_t(a \mid b) da = p(b)$$

MH target distribution

For the Metropolis-Hastings algorithm,

$$T_t(b \mid a) = J_t(b \mid a) \min\left\{1, \frac{p(b)\, J_t(a \mid b)}{p(a)\, J_t(b \mid a)}\right\}.$$

$$p(a)\, T_t(b|a) = p(a)\, J_t(b|a) \min\left\{1, \frac{p(b)\, J_t(a|b)}{p(a)\, J_t(b|a)}\right\}$$

$$= \min\left\{p(a)\, J_t(b|a),\ p(b)\, J_t(a|b)\right\}.$$

$$= p(b)\, T_t(b|a)$$

More generally, as long as $T_t(b \mid a)$ is of the form

$$T_t(b \mid a) = p(b)\, \delta(b \mid a),$$

where δ is symmetric in a and b, detailed balance is satisfied (Stein's r_C).

Warm-up and Convergence Diagnostics

A theoretically valid Markov chain converges to its stationary distribution asymptotically. But in finite time, the initial part of the chain may be non-representative. This is the **warm up** (or **burn-in**). Lower error of the estimate can be obtained by throwing these draws away.

MCMC draws are dependent:

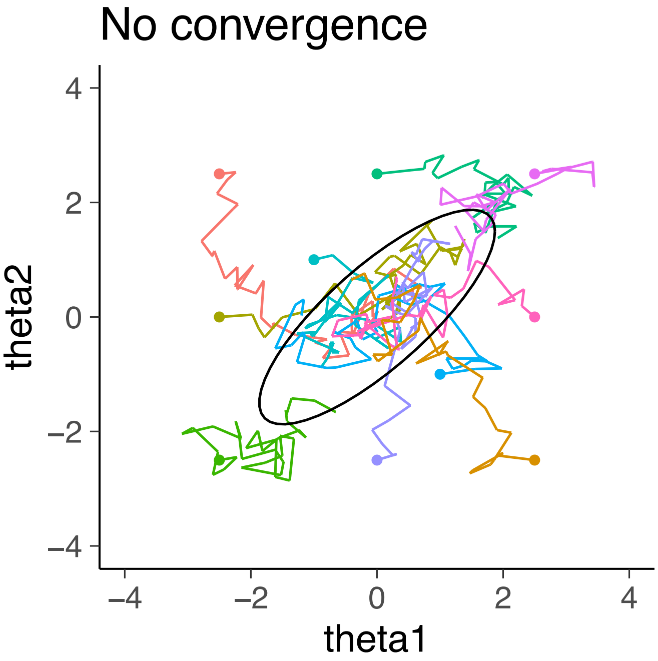
- Monte Carlo estimates still valid if central limit theorem holds. We need to make sure that is the case;
- Estimation of Monte Carlo error is more difficult. This calls for an evaluation of *effective sample size*.

Convergence diagnostics.

Are we getting samples from the target distribution?

- Run multiple chains with **over-dispersed** starting points;
- Separately monitor several scalar estimands;
- Splitting each sequence (post warm-up) into two (or more) parts;
- **Mixing**: only when the distribution of each simulated sequence is close to the distribution of all the sequences mixed together, can they all be (possibly) approximating the target distribution.
- If simulation efficiency is unacceptably low, the algorithm should be altered.

Monitoring mixing



Start chains from different and overdispersed starting points. Remove draws from the beginning of the chains and run chains long enough so that it is not possible to distinguish where each chain started and the chains are well mixed.

Assessing mixing

- M chains, each with N draws
- **Within**-chain variance

$$W = \frac{1}{M}\sum_{m=1}^{M} s_m^2, \qquad \text{where } s_m^2 = \frac{1}{N-1}\sum_{n=1}^{N}(\theta_{nm} - \bar{\theta}_{.m})^2$$

- **Between**-chain variance

$$B = \frac{N}{M-1}\sum_{m=1}^{M}(\bar{\theta}_{.m} - \bar{\theta}_{..})^2,$$

$$\text{where } \bar{\theta}_{.m} = \frac{1}{N}\sum_{n=1}^{N}\theta_{nm}, \quad \bar{\theta}_{..} = \frac{1}{M}\sum_{m=1}^{M}\bar{\theta}_{.m}$$

Thus B/N is the variance of the means of the chains.

Assessing mixing

Estimate marginal posterior variance $var(\theta|y)$ as a weighted mean of W and B:

$$\widehat{var}^{+}(\theta|y) = \frac{N-1}{N}W + \frac{1}{N}B.$$

Note that

- $\widehat{var}^{+}(\theta|y)$ *overestimates* marginal posterior variance if the starting points are *overdispersed*, but is *unbiased* if chain is stationary or when $N \to \infty$;
- On the other hand, for finite N, the within variance W *underestimates* marginal posterior variance, but $\mathbb{E}(W) \to var(\theta|y)$ when $N \to \infty$.

Potential scale reduction factor (PSRF)

Since $\widehat{var}^{+}(\theta|y)$ overestimates and W underestimates $var(\theta|y)$, compute the **PSRF**:

$$\widehat{R} = \sqrt{\frac{\widehat{var}^{+}}{W}},$$

which estimates how much the scale of our current description of θ could reduce if $N \to \infty$.

Potential scale reduction factor (PSRF)

$$\widehat{R} = \sqrt{\frac{\widehat{var}^{+}}{W}},$$

- $\widehat{R} \to 1$, when $N \to \infty$;
- If $\widehat{R}$ is big (e.g., $R > 1.01$), keep sampling;
- If $\widehat{R}$ is close to 1, it is still possible that chains have not converged, for reasons such as
  - if starting points were not overdispersed;
  - if the target distribution far from normal (especially if variance is infinite or nearly infinite);
  - just by chance, when N is finite. Etc..