Markov chain Monte Carlo idea

- Situation:
  - Given a target distribution $f(x)$
  - Want to generate samples from $f(x)$

- Idea:
  - construct a Markov chain $\{X_i\}_{i=1}^{\infty}$ so that

$$\lim_{i \to \infty} P(X_i = x) = f(x)$$

  - simulate the Markov chain for many iterations
  - for m large enough $x_m, x_{m+1}, \ldots$ are (essentially) from $f(x)$

How to construct the Markov chain

- How to construct such a Markov chain? ($x \in \Omega$ discrete)
  - Markov chain transition probabilities:
    $P(y|x) = P(X_{i+1} = y | X_i = x)$
  - Need to have

$$f(y) = \sum_{x \in \Omega} f(x) P(y|x) \quad \text{for all } y \in \Omega$$

  - Sufficient condition: Detailed balance condition

$$f(x)P(y|x) = f(y)P(x|y) \quad \text{for all } x, y \in \Omega$$

- Metropolis–Hastings setup for $P(y|x)$:

$$P(y|x) = Q(y|x)\alpha(y|x) \;\; \text{when } y \neq x$$

$$P(x|x) = 1 - \sum_{y \neq x} Q(y|x)\alpha(y|x)$$

where

$$\alpha(y|x) = \min\left\{1, \frac{f(y)}{f(x)} \cdot \frac{Q(x|y)}{Q(y|x)}\right\}$$

Common proposal types

- ⋆ Independent proposals: $Q(y|x) = q(y)$
  - usually not a good alternative (alone)
- ⋆ Random walk proposals: $Q(y|x) = N(y|x, \sigma^2 I)$
  - is used a lot
  - includes a tuning parameter: σ
- ⋆ Langevin proposals: $Q(y|x) = N(y|x + h\nabla \ln f(x), h^2 I)$
  - needs $\nabla \ln f(x)$
  - includes a tuning parameter: h
- ⋆ Gibbs updates: We haven't discussed this yet

Combination of strategies

- Have two (or more) proposal kernels, $Q_1(y|x)$, $Q_2(y|x)$
  - Alternative 1:

$$\begin{aligned} Q(y|x) &= p\,Q_1(y|x) + (1-p)Q_2(y|x) \\ \alpha(y|x) &= \min\left\{1, \frac{f(y)}{f(x)} \cdot \frac{p\,Q_1(x|y) + (1-p)Q_2(x|y)}{p\,Q_1(y|x) + (1-p)Q_2(y|x)}\right\} \end{aligned}$$

  - Alternative 2:

$$\begin{aligned} P_i(y|x) &= \begin{cases} Q_i(y|x)\alpha_i(y|x) & \text{for } y \neq x, \\ 1 - \sum_{z\neq x} Q_i(z|x)\alpha_i(z|x) & \text{for } y = z \end{cases} \\ \alpha_i(y|x) &= \min\left\{1, \frac{f(y)}{f(x)} \cdot \frac{Q_i(x|y)}{Q_i(y|x)}\right\} \\ P(y|x) &= p\,P_1(y|x) + (1-p)P_2(y|x) \end{aligned}$$

  - Alternative 3: We will discuss a third alternative today

Combination of strategies

- Have two (or more) proposal kernels, $Q_1(y|x)$, $Q_2(y|x)$
  - Alternative 1:

$$
\begin{aligned}
Q(y|x) &= p\,Q_1(y|x) + (1-p)Q_2(y|x) \\
\alpha(y|x) &= \min\left\{1, \frac{f(y)}{f(x)} \cdot \frac{p\,Q_1(x|y) + (1-p)Q_2(x|y)}{p\,Q_1(y|x) + (1-p)Q_2(y|x)}\right\}
\end{aligned}
$$

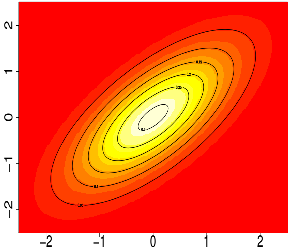
  - Alternative 2:

$$
\begin{aligned}
P_i(y|x) &= \begin{cases} Q_i(y|x)\alpha_i(y|x) & \text{for } y \neq x, \\ 1 - \sum_{z\neq x} Q_i(z|x)\alpha_i(z|x) & \text{for } y = z \end{cases} \\
\alpha_i(y|x) &= \min\left\{1, \frac{f(y)}{f(x)} \cdot \frac{Q_i(x|y)}{Q_i(y|x)}\right\} \\
P(y|x) &= p\,P_1(y|x) + (1-p)P_2(y|x)
\end{aligned}
$$

  - Alternative 3: We will discuss a third alternative today
- Note: Alt. 2 costs less cpu time per iteration than Alt. 1

Toy example: Combination of strategies

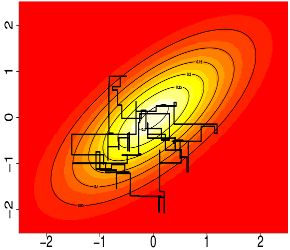
- ⋆ Target distribution $f(x), x = (x^1, x^2) \in \mathbb{R}^2$



- ⋆ Proposal distributions, $p = 1/2$
  - $Q_1(y|x)$:
    - + propose $y^1 \sim N(x^1, \sigma^2)$
    - + keep $y^2 = x^2$ unchanged
  - $Q_2(y|x)$:
    - + propose $y^2 \sim N(x^2, \sigma^2)$
    - + keep $y^1 = x^1$ unchanged
- ⋆ Note: $Q_1(y|x)$ and $Q_2(y|x)$ don't give irreducible Markov chains separately, together they do.

Toy example: Combination of strategies

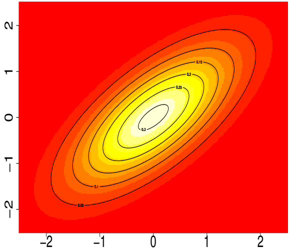
- ⋆ Target distribution $f(x), x = (x^1, x^2) \in \mathbb{R}^2$



- ⋆ Proposal distributions, $p = 1/2$
  - $Q_1(y|x)$:
    - + propose $y^1 \sim N(x^1, 0.3^2)$
    - + keep $y^2 = x^2$ unchanged
  - $Q_2(y|x)$:
    - + propose $y^2 \sim N(x^2, 0.3^2)$
    - + keep $y^1 = x^1$ unchanged
- ⋆ Note: $Q_1(y|x)$ and $Q_2(y|x)$ don't give irreducible Markov chains separately, together they do.

Toy example: Gibbs for a bivariate normal

- ⋆ Target distribution, $x \sim N(0, \Sigma), \Sigma = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}$

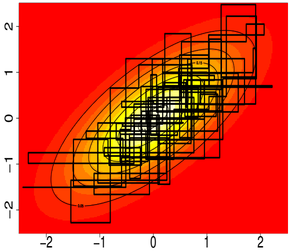


- ⋆ Full conditional distributions
  - $x^1 | x^2 \sim \mathrm{N}(0.7x^2, 0.51)$
  - $x^2 | x^1 \sim \mathrm{N}(0.7x^1, 0.51)$

- ⋆ Note:
  - contains no tuning parameter
  - must be able to find (and sample from) the full conditionals
  - waist of time to update the same coordinate two times in a row

Toy example: Gibbs for a bivariate normal

- ⋆ Target distribution, $x \sim N(0, \Sigma), \Sigma = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}$



- ⋆ Full conditional distributions
  - $x^1|x^2 \sim \mathrm{N}(0.7x^2, 0.51)$
  - $x^2|x^1 \sim \mathrm{N}(0.7x^1, 0.51)$
- ⋆ Note:
  - contains no tuning parameter
  - must be able to find (and sample from) the full conditionals
  - waist of time to update the same coordinate two times in a row

Convergence diagnostics

⋆ When has the Markov chain converged?

⋆ Several theoretical results exist: for a given $\epsilon > 0$

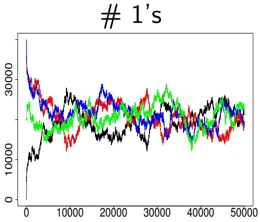
$$||f(\cdot) - \mathrm{P}_n(\cdot)|| \leq \epsilon \text{ for all } n \leq N(\epsilon)$$

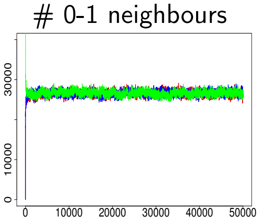
where $N(\epsilon)$ can be computed.

– bounds too weak to be of any practical value

⋆ Standard start to evaluate convergence:

– look at trace plots (ex. Ising model)

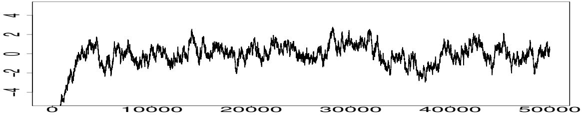




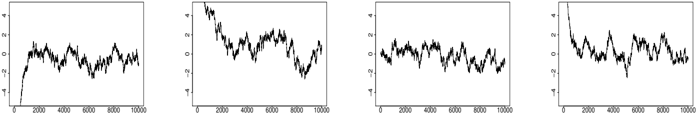
One long chain or many shorter chains?

- With fixed cpu-time available, should we
  - use all time in one long Markov chain run, or
  - run several shorter Markov chain runs?

- One long Markov chain run



- Several shorter Markov chain runs

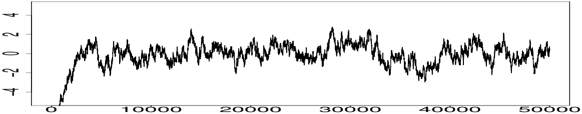


One long chain or many shorter chains?

⋆ With fixed cpu-time available, should we

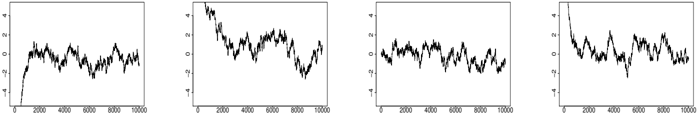
– use all time in one long Markov chain run, or
– run several shorter Markov chain runs?

⋆ One long Markov chain run



– only one burn-in period to discard
– more likely that you really have converged

⋆ Several shorter Markov chain runs



– easier to evaluate the convergence
– easier to estimate estimation variance (the chains are independent)