# More on Markov chain Monte Carlo

**First winter school in eScience**
**Geilo, Thursday February 1st 2007**

**Pdf file available from**
**http://www.math.ntnu.no/∼haakont/vinterskole/**

**Håkon Tjelmeland**
**Department of Mathematical Sciences**
**Norwegian University of Science and Technology**
**Trondheim, Norway**

# Some repetition

- **Given target distribution:** $\pi(x), x \in \mathbb{R}^N$
- **Want to understand the properties of** $\pi(x)$

$$\mu_f = \mathbf{E}[f(x)] = \int f(x)\pi(x)\mathbf{d}x$$

  - **or what is the probability distribution of** $f(x)$**?**

- **Generate realisations** $x_1, \ldots, x_n$ **from** $\pi(x)$

$$\widehat{\mu}_f = \frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

  - **or make a histogram of** $f(x_1), \ldots, f(x_n)$

- **Example: Ising model,** $x = (x^1, \ldots, x^N)$**,** $x^i \in \{0, 1\}$

$$\pi(x) = c\exp\left\{-\beta\sum_{i\sim j} I(x^i \neq x^j)\right\}$$

  - **let**

$$f_1(x) = \frac{1}{N}\sum_{i=1}^{N} I(x^i = 1) \quad \textbf{and} \quad f_2(x) = \frac{1}{N}\sum_{i\sim j} I(x^i = x^j)$$

  - **Note: because of symmetry** $\mathbf{E}[f_1(x)] = 1/2$

# Some repetition (cont.)

- **Note: In principle we can compute**

$$\mathbf{E}[f_2(x)] = \sum_x f_2(x)\pi(x)$$

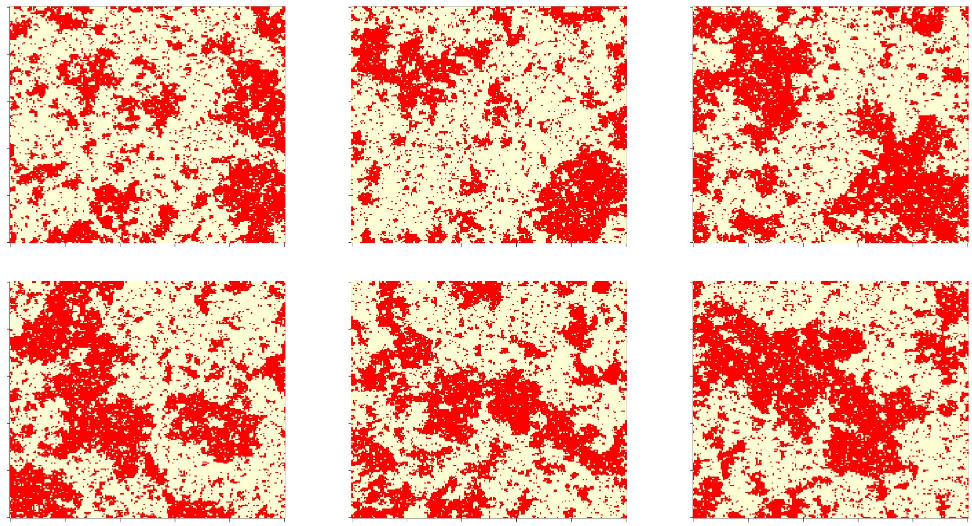
  - **but the sum has** $2^{200\cdot 200} \approx 10^{12041}$ **terms**

  - **and to find the normalising constant of** $\pi(x)$ **we need to compute a sum with the same number of terms**

- **So in practice it is not possible (in my lifetime)**

# Some repetition (cont.)

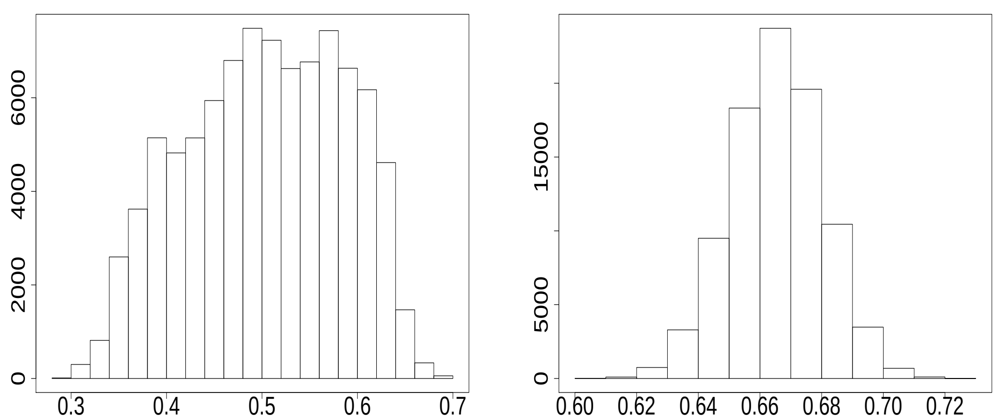
- **Realisations from $\pi(x)$ with $\beta = 0.87$**



- **Emperical mean values**

$$\widehat{\mu}_{f_1} = 0.5034 \quad \textbf{and} \quad \widehat{\mu}_{f_2} = 0.665$$

- **Histograms**



# Some repetition (cont.)

- **Metropolis–Hastings transition kernel**

$$\mathbf{P}(y|x) = \mathbf{Q}(y|x)\alpha(y|x) \text{ , } y \neq x$$

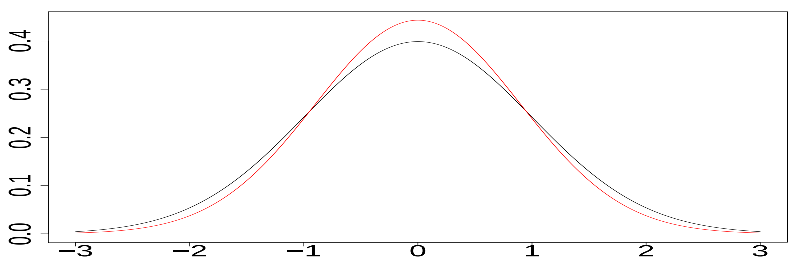
$$\alpha(y|x) = \min\left\{1, \frac{\pi(x)\mathbf{Q}(x|y)}{\pi(y)\mathbf{Q}(y|x)}\right\}$$

- **Metropolis–Hastings algorithm**
  - **generate initial state** $x_0 \sim f(x_0)$
  - **for** $i = 1, 2, \ldots$
    - **propose potential new state** $y_i \sim \mathbf{Q}(y_i|x_{i-1})$
    - **compute acceptance probability** $\alpha(y_i|x_{i-1})$
    - **generate** $u_i \sim$ **Uniform**$(0, 1)$
    - **if** $u_i \leq \alpha(y_i|x_{i-1})$ **accept** $y_i$**, i.e. set** $x_i = y_i$**, otherwise reject** $y_i$ **and set** $x_i = x_{i-1}$

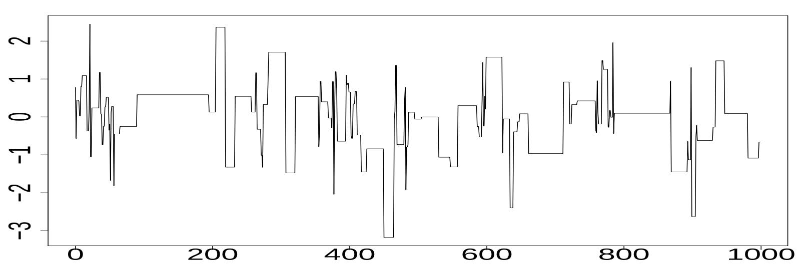
- **Next question: What** $\mathbf{Q}(y|x)$ **to use?**
  - **simple choices is often ok — but not always**

# Independent proposal MH

- **Target density:** $\pi(x), x \in \mathbb{R}^N$
- **Proposal density:** $Q(y|x) = q(y)$
  - **does not depend on current state** $x$
  - $q(y)$ **is an approximation to** $\pi(x)$
- **Toy example**
  - **target distribution:** $x \sim \mathbf{N}_{250}(0, I)$
  - **proposal distribution:** $y|x \sim \mathbf{N}_{250}(0, 0.9^2 \cdot I)$

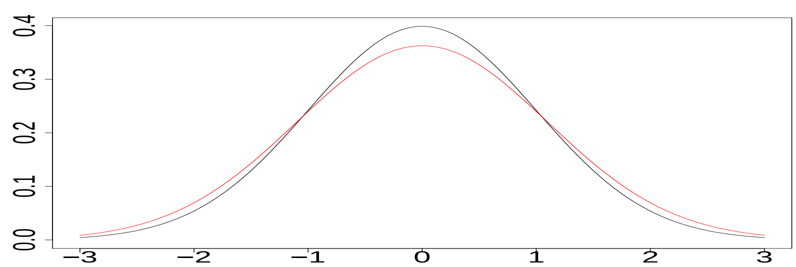



  - **trace plot of** $x^1$

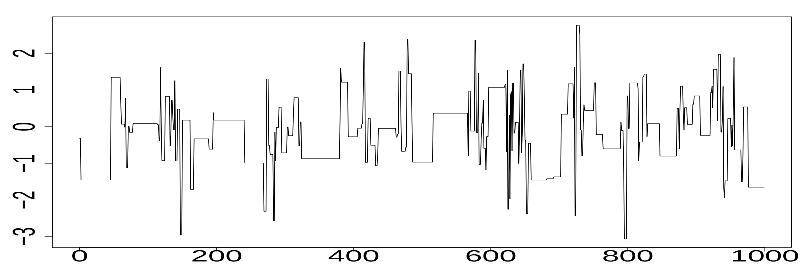


# Independent proposal MH (cont.)

- **Another toy example**
  - **target distribution:** $x \sim \mathbf{N}_{250}(0, I)$
  - **proposal distribution:** $y|x \sim \mathbf{N}_{250}(0, 1.1^2 \cdot I)$




  - **trace plot of** $x^1$




- **Experience:**
  - **Except in low dimensional spaces: Convergence of independent proposal MH is either very good or very bad, usually very bad**
  - **The tails of the proposal distribution must at least be as heavy as the tails of the target distribution**

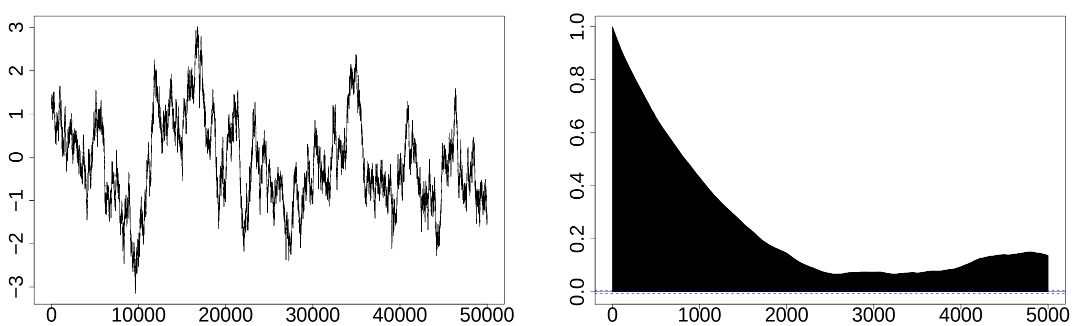
# Random walk proposal MH

- **Target density:** $\pi(x), x \in \mathbb{R}^N$

- **Proposal density:** $Q(y|x) = q\left(\frac{y-x}{\sigma}\right)$
  - **typically: Gaussian proposal**
  - **proposal mean is current state**
  - **tuning parameter:** $\sigma$

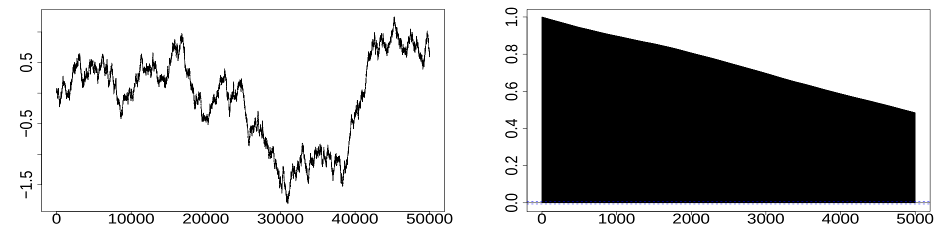
- **Toy example**
  - **target distribution:** $x \sim \mathbf{N}_{250}(0, I)$
  - **proposal distribution:** $y|x \sim \mathbf{N}_{250}(x, \sigma^2 \cdot I)$
  - **trace plot and acf of** $x^1$

$\sigma = 0.05$, **acceptance rate** $= 0.69$

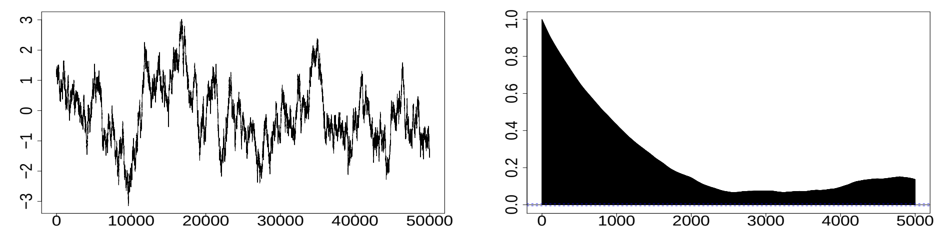


# Random walk proposal MH (cont.)

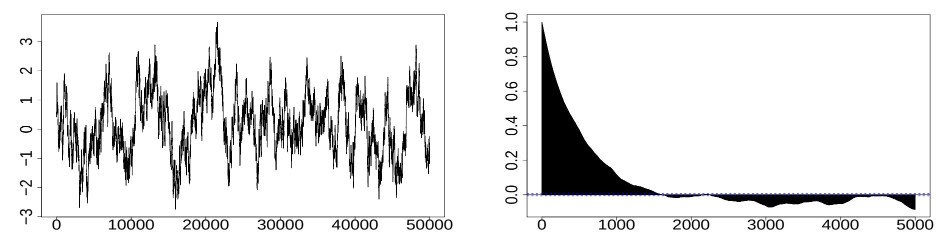
**$\sigma = 0.01$, acceptance rate $= 0.94$**



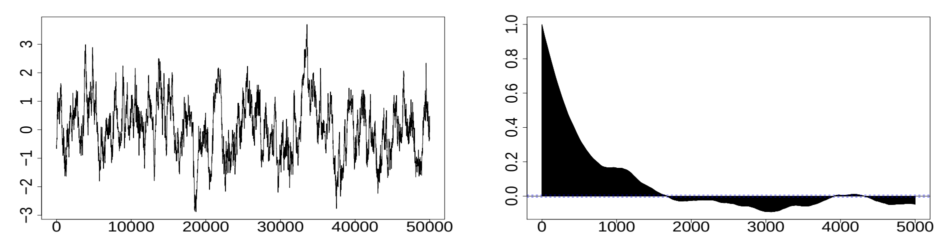

**$\sigma = 0.05$, acceptance rate $= 0.69$**



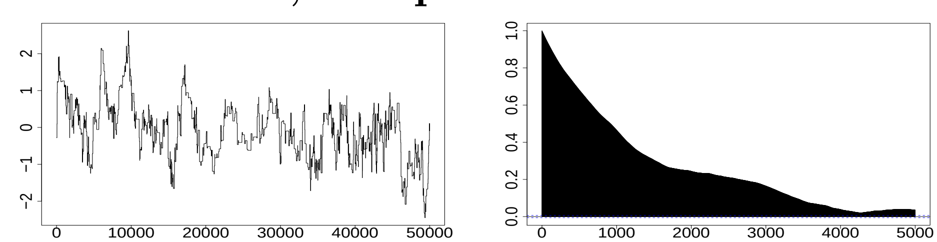

**$\sigma = 0.10$, acceptance rate $= 0.426$**




**$\sigma = 0.20$, acceptance rate $= 0.11$**




**$\sigma = 0.30$, acceptance rate $= 0.018$**



# Random walk MH (cont.)

- **Result (Roberts et al., 1997):**

  - **let**
  $$\pi(x) = \prod_{i=1}^{n} f(x^i)$$
  **where $f(\cdot)$ fulfil some conditions**

  - **use Gaussian random walk MH algorithm to sample $\pi(x)$**

  - **asymptotically, as $n \to \infty$, the optimal tuning parameter $\sigma$ gives acceptance rate $0.234$.**

- **Rule of thumb for random walk MH:**
  - **tune $\sigma$ to get acceptance rate $0.234$**
  - **between $0.15$ and $0.5$ is ok.**

# Langevin proposals

- **Intuition: Should more oftenly propose new values in high probability area**

- **Suboptimal to have $x$ as proposal mean**

- **Proposal mean should be shifted in the gradient direction**

- **Langevin proposal**

$$Q(y|x) = \mathbf{N}(x + h\nabla\pi(x), h^2 I)$$

- **Can also be motivated from stochastic differential equation theory when $h \to 0$**

- **For us $h$ should not be too small**

- **Again one can ask how to choose $h$, or what is the optimal acceptance rate**

- **The answer is acceptance rate about $0.5$**

# Combination of strategies

- **Target distribution:** $\pi(x)$
- **Two proposal distributions:** $Q_1(y|x)$ **and** $Q_2(y|x)$
- **How to combine the proposal distributions?**
  - **first alternative**

$$
\begin{aligned}
Q(y|x) &= pQ_1(y|x) + (1-p)Q_2(y|x) \\
\alpha(y|x) &= \min\left\{1, \frac{\pi(y)(pQ_1(x|y) + (1-p)Q_2(x|y))}{\pi(x)(pQ_1(y|x) + (1-p)Q_2(y|x))}\right\}
\end{aligned}
$$

  - **second alternative (notation for discrete $x$)**

$$
P(y|x) = pP_1(y|x) + (1-p)P_2(y|x)
$$

**where**

$$
\begin{aligned}
P_i(y|x) &= \begin{cases} Q_i(y|x)\alpha_i(y|x) & \textbf{if } y \neq x \\ 1 - \sum_{z \neq x} Q_i(z|x)\alpha_i(z|x) & \textbf{if } y = x \end{cases} \\
\alpha_i(y|x) &= \min\left\{1, \frac{\pi(y)Q_i(x|y)}{\pi(x)Q_i(y|x)}\right\}
\end{aligned}
$$

- **first alternative give higher acceptance rate**
- **second alternative cost less per iteration**
- **is the second alternative correct?**

# Proof of correctness

- **The $Q_1(y|x)$ gives $P_1(y|x)$ for which**

$$\pi(y) = \sum_{x \in \Omega} \pi(x) P_1(y|x)$$

- **The $Q_2(y|x)$ gives $P_2(y|x)$ for which**

$$\pi(y) = \sum_{x \in \Omega} \pi(x) P_2(y|x)$$

- **For $P(y|x) = pP_1(y|x) + (1-p)P_2(y|x)$, need to verify**

$$\pi(y) = \sum_{x \in \Omega} \pi(x) P(y|x)$$

- **Start with the sum on the right**

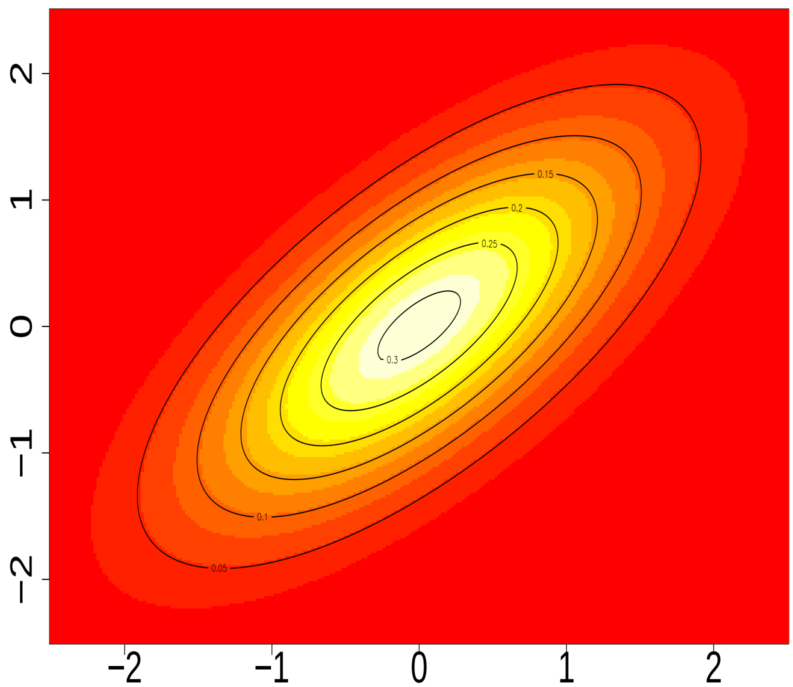
$$\begin{aligned}
\sum_{x \in \Omega} \pi(x) P(y|x) &= \sum_{x \in \Omega} \pi(x)(pP_1(y|x) + (1-p)P_2(y|x)) \\
&= p \sum_{x \in \Omega} \pi(x) P_1(y|x) + (1-p) \sum_{x \in \Omega} \pi(x) P_2(y|x) \\
&= p\pi(y) + (1-p)\pi(y) = \pi(y)
\end{aligned}$$

- **$P(y|x)$ fulfils detailed balance if $P_1(y|x)$ and $P_2(y|x)$ do**

$$\pi(x) P(y|x) = \pi(y) P(x|y)$$

# Combination of strategies - example

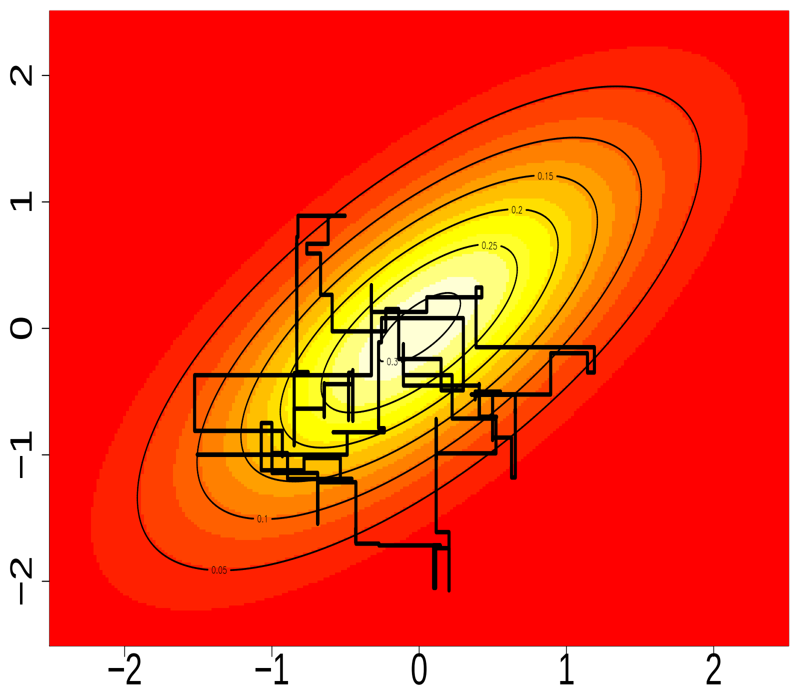
- **Target distribution** $\pi(x), x = (x^1, x^2) \in \mathbb{R}^2$




- **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.**

# Combination of strategies - example

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




- **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.**

# Combination of strategies - Ising

- **Probability distribution**

$$\pi(x) = c \cdot \exp\left\{-\beta \sum_{i \sim j} I(x^i \neq x^j)\right\}$$

- $N$ **proposal distributions,** $Q_i(y|x)$ **is**
  - **propose** $y^i = 1 - x^i$
  - **keep** $y^k = x^k, k \neq i$ **unchanged**
  - **thus**

$$Q_i(y|x) = \begin{cases} 1 & \textbf{if } y^i = 1 - x^i \textbf{ and } y^k = x^k, k \neq i, \\ 0 & \textbf{otherwise} \end{cases}$$

$$\alpha_i(y|x) = \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}$$

- **In each iteration: draw** $i \in \{1, \ldots, n\}$ **at random**

- **Note:**
  - **same algorithm as before**
  - **don't need to be any randomness in** $Q_i(y|x)$

- **Can we "visit" the nodes sequentially in stead?**

# Combination of strategies

- **Target distribution:** $\pi(x)$
- **Two proposal distributions:** $Q_1(y|x)$ **and** $Q_2(y|x)$
- **How to combine the proposal distributions?**
  - **third alternative**

$$P(y|x) = \sum_{z \in \Omega} P_1(z|x) P_2(y|z)$$

    * **update** $x^1$**, update** $x^2$**, update** $x^1$ **and so on**

- **Is this third alternative correct?**

# Proof of correctness

- **The $Q_1(y|x)$ gives $P_1(y|x)$ for which**

$$\pi(y) = \sum_{x \in \Omega} \pi(x) P_1(y|x)$$

- **The $Q_2(y|x)$ gives $P_2(y|x)$ for which**

$$\pi(y) = \sum_{x \in \Omega} \pi(x) P_2(y|x)$$

- **For $P(y|x) = \sum_{z \in \Omega} P_1(z|x) P_2(y|z)$, need to verify**

$$\pi(y) = \sum_{x \in \Omega} \pi(x) P(y|x)$$

- **Start with the sum on the right**

$$\begin{aligned}
\sum_{x \in \Omega} \pi(x) P(y|x) &= \sum_{x \in \Omega} \left[ \pi(x) \sum_{z \in \Omega} P_1(z|x) P_2(y|z) \right] \\
&= \sum_{z \in \Omega} \left[ P_2(y|z) \sum_{x \in \Omega} \pi(x) P_1(z|x) \right] \\
&= \sum_{z \in \Omega} P_2(y|z) \pi(z) = \pi(y)
\end{aligned}$$

- **$P(y|x)$ does not fulfil detailed balance even if $P_1(y|x)$ and $P_2(y|x)$ do**

# Gibbs sampler

- **Let** $x = (x^1, \ldots, x^n)$

- $N$ **proposal distributions,** $Q_i(y|x)$ **is**
  - **propose** $y^i \sim \pi(y^i | x^1, \ldots, x^{i-1}, x^{i+1}, \ldots, x^n)$
  - **keep** $y^k = x^k, k \neq i$ **unchanged**

- **Notation:** $x^{-i} = (x^1, \ldots, x^{i-1}, x^{i+1}, \ldots, x^n)$

- **Acceptance probability**

$$
\begin{aligned}
\alpha_i(y|x) &= \min\left\{1, \frac{\pi(y)Q_i(x|y)}{\pi(x)Q_i(y|x)}\right\} = \min\left\{1, \frac{\pi(y)\pi(x^i|y^{-i})}{\pi(x)\pi(y^i|x^{-i})}\right\} \\
&= \min\left\{1, \frac{\pi(y)\,\frac{\pi(x^i, y^{-i})}{\pi(y^{-i})}}{\pi(x)\,\frac{\pi(y^i, x^{-i})}{\pi(x^{-i})}}\right\} = \min\left\{1, \frac{\pi(y)\,\pi(x^i, y^{-i})}{\pi(x)\,\pi(y^i, x^{-i})}\right\} \\
&= \min\left\{1, \frac{\pi(y)\pi(x^i, x^{-i})}{\pi(x)\pi(y^i, y^{-i})}\right\} = 1
\end{aligned}
$$

- **Thus: always accept**

# Gibbs for Ising

- **Ising probability distribution**

$$\pi(x) = c \cdot \exp\left\{-\beta \sum_{k\sim l} I(x^k \neq x^l)\right\}$$

- **Full conditional distribution**

$$\begin{aligned}
\pi(x^i|x^{-i}) &= \frac{\pi(x^i, x^{-i})}{\pi(x^{-i})} \propto \pi(x^i, x^{-i}) = \pi(x) \\
&= \exp\left\{-\beta \sum_{k\sim l} I(x^k \neq x^l)\right\} \\
&\propto \exp\left\{-\beta \sum_{k\sim i} I(x^k \neq x^i)\right\}
\end{aligned}$$

**Thus**

$$\pi(x^i|x^{-i}) = c \exp\left\{-\beta \sum_{k\sim i} I(x^k \neq x^i)\right\}$$

**where**

$$c = \left[\sum_{x^i=0}^{1} \exp\left\{-\beta \sum_{k\sim i} I(x^k \neq x^i)\right\}\right]^{-1}$$

- **Should we here prefer Gibbs, or always propose to change the value of $x_i$?**

# Ising: Gibbs or propose to change?

- **Probability for a changed value:**
  - **Gibbs**

$$
\begin{aligned}
\pi(1-x^i|x^{-1}) &= \frac{e^{-\beta\sum_{k\sim i} I(x^k\neq 1-x^i)}}{e^{-\beta\sum_{k\sim i} I(x^k\neq x^i)} + e^{-\beta\sum_{k\sim i} I(x^k\neq 1-x^i)}} \\
&= \frac{e^{-\beta\cdot(\#\ \textbf{equal})}}{e^{-\beta\cdot(\#\ \textbf{unequal})} + e^{-\beta\cdot(\#\ \textbf{equal})}} \\
&= \frac{e^{-\beta\cdot(\#\ \textbf{equal}\ -\ \#\ \textbf{unequal})}}{1 + e^{-\beta\cdot(\#\ \textbf{equal}\ -\ \#\ \textbf{unequal})}}
\end{aligned}
$$

  - **always propose to change**

$$
\begin{aligned}
\alpha(y|x) &= \min\left\{1, e^{-\beta\sum_{j\sim i}\left[I(x^j\neq 1-x^i)-I(x^j\neq x^i)\right]}\right\} \\
&= \min\left\{1, e^{-\beta\cdot(\#\ \textbf{equal}\ -\ \#\ \textbf{unequal})}\right\}
\end{aligned}
$$

- **See that**

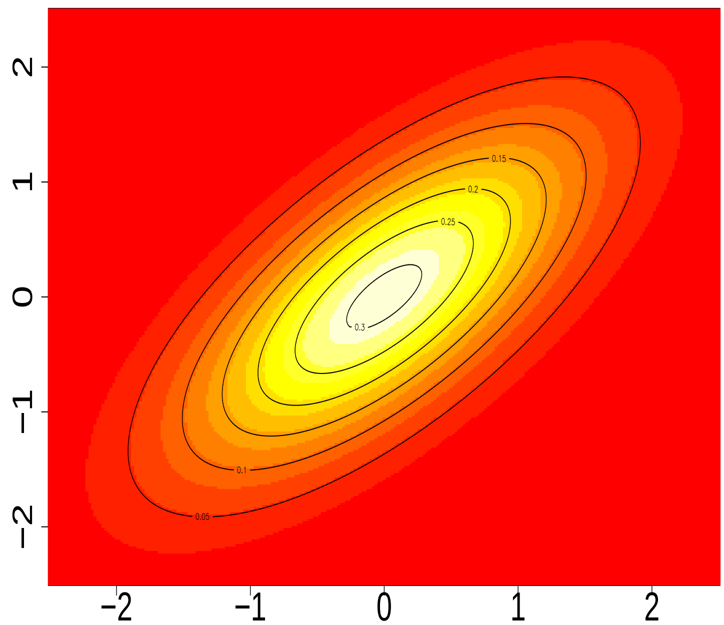
$$\pi(1-x^i|x^{-i}) < \alpha(y|x)$$

- **Better always to propose a change**

# Gibbs for a bivariate normal

- **Toy example, you should never use MCMC here!**
- **Target distribution**

$$\pi(x) = \frac{1}{2\pi}\frac{1}{\sqrt{|\Sigma|}}\exp\left\{-\frac{1}{2}x^T\Sigma^{-1}x\right\}\ ,\ x \in \mathbb{R}^2,\ \Sigma = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}$$



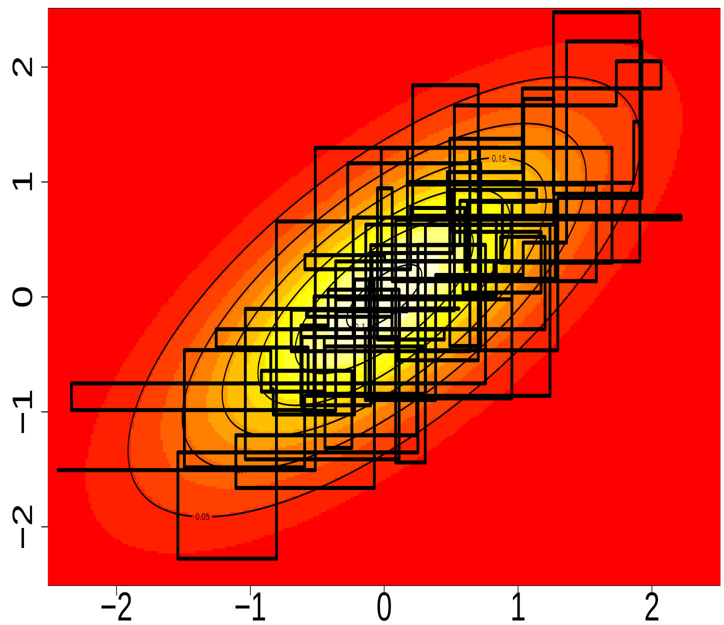

- **Full conditional distributions**
  - $x^1|x^2 \sim \mathbf{N}(0.7x^2, 0.51)$
  - $x^2|x^1 \sim \mathbf{N}(0.7x^1, 0.51)$

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

# Gibbs for a bivariate normal

- **Toy example, you should never use MCMC here!**
- **Target distribution**

$$\pi(x) = \frac{1}{2\pi} \frac{1}{\sqrt{|\Sigma|}} \exp\left\{ -\frac{1}{2} x^T \Sigma^{-1} x \right\} \;, \; x \in \mathbb{R}^2, \; \Sigma = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}$$




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

# Plan

- **The Markov chain Monte Carlo (MCMC) idea**
- **Some Markov chain theory**
- **Implementation of the MCMC idea**
  - **Metropolis–Hastings algorithm**
- **MCMC strategies**
  - **independent proposals**
  - **random walk proposals**
  - **combination of strategies**
  - **Gibbs sampler**
- **Convergence diagnostics**
  - **trace plots**
  - **autocorrelation functions**
  - **one chain or many chains?**
- **Typical MCMC problems — and some remedies**
  - **high correlation between variables**
  - **multimodality**
  - **different scales**

# Convergence diagnostics

- **When has the Markov chain converged?**
- **Several theoretical results exist: for a given $\epsilon > 0$**

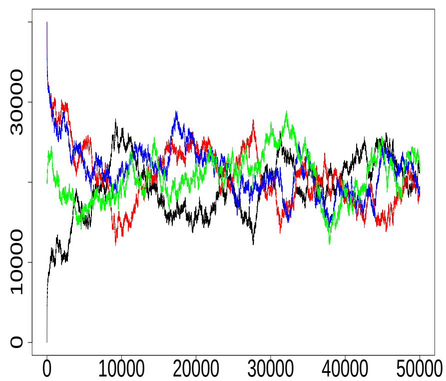
$$||\pi(\cdot) - \mathbf{P}^n(\cdot)|| \leq \epsilon \text{ for all } n \geq M(\epsilon)$$

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

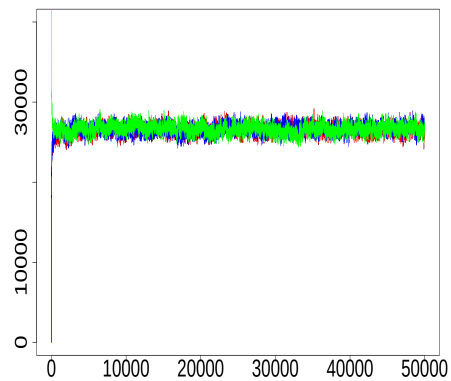
  – **bounds too weak to be of any practical value**

- **Standard start to evaluate convergence:**
  – **look at trace plots**
    * **Ising example:**

**# 1's**




**# 0-1 neighbours**




- **Result:**

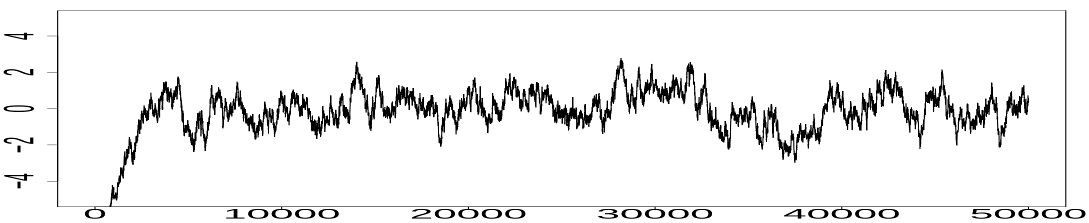
$$\mathbf{P}^n(\cdot) \ \rightarrow^d \ \pi(\cdot)$$
$$\Updownarrow$$
$$\int f(\cdot) d\mathbf{P}_n \ \rightarrow \ \int f(\cdot) d\pi \text{ for all}$$

**bounded real-valued ($\mu$-measurable) functions $f(\cdot)$**

# One chain or many chains?

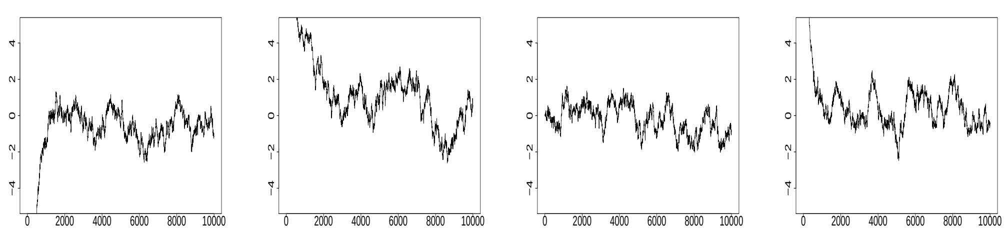
- With fixed cpu-time available, should we
  - use all time in one long Markov chain run, or
  - run several shorter Markov chain runs?

- 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

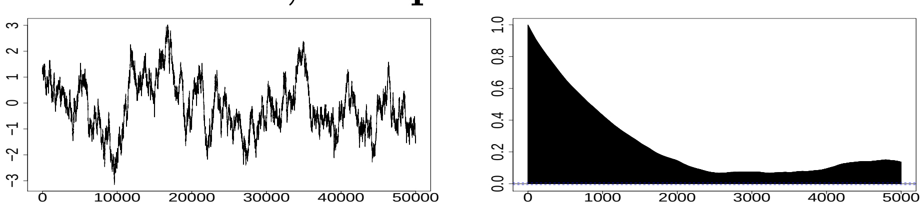
# Convergence diagnostics

- **many more formal convergence diagnostics exists**
  - **some based on a single Markov chain run**
  - **some based on several Markov chain runs**
- **To see when a chain has convergence, we need to simulate much longer than to convergence**
- **If some properties of the target distribution is known: use it to check convergence!**
- **All convergence diagnostics can (and do) fail**
  - **we can construct situations where it fails**

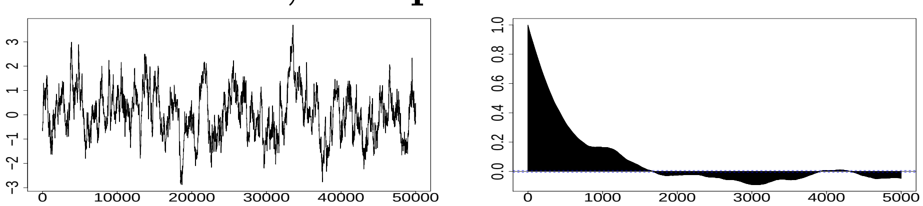
# Compare algorithms

- **Assume: have two (or more) Markov chains with limiting distribution $\pi(x)$**
- **Which one should we prefer?**
- **Estimate and compare autocorrelation functions**
  - **ignore burn-in periods!**
  - **assume stationary time series**
  - **must again consider scalar functions $f(x)$**
  - **random walk proposal example, choice of tuning parameter**

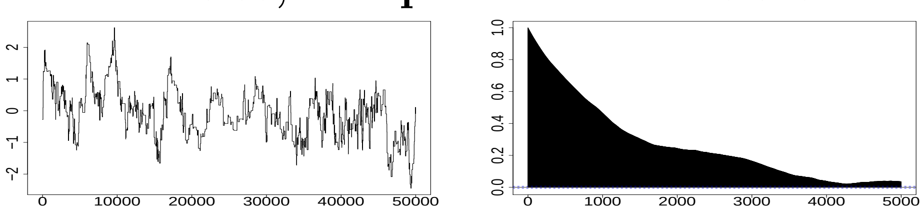
$\sigma = 0.05$, **acceptance rate** $= 0.69$




$\sigma = 0.20$, **acceptance rate** $= 0.11$




$\sigma = 0.30$, **acceptance rate** $= 0.018$



# Variance estimation in MCMC

- **Standard Monte Carlo gives independent samples**

$$x_1, \ldots, x_n \sim \pi(x) \textbf{ and independent}$$

  – **unbiased estimator for** $\mu_f = \int f(x)\pi(x)\mathbf{d}x$

$$\widehat{\mu}_f = \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

  – **variance estimation is easy**

$$\mathbf{Var}[\widehat{\mu}_f] = \frac{1}{n^2}\sum_{i=1}^{n} \mathbf{Var}[f(x_i)] = \frac{\mathbf{Var}[f(x)]}{n}$$

$$\widehat{\mathbf{Var}}[f(x)] = \frac{1}{n-1}\sum_{i=1}^{n} \left(f(x_i) - \widehat{\mu}_f\right)^2$$

- **MCMC gives dependent samples**

$$x_1, \ldots, x_n \sim \pi(x) \textbf{ and dependent}$$

  – **unbiased estimator for** $\mu_f$

$$\widehat{\mu}_f = \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

  – **variance estimation is not so easy**

$$\mathbf{Var}[\widehat{\mu}_f] = \frac{1}{n^2}\left[\sum_{i=1}^{n} \mathbf{Var}[f(x_i)] + \sum_{i=1}^{n}\sum_{j\neq i} \mathbf{Cov}[f(x_i), f(x_j)]\right]$$

# Variance estimation in MCMC (cont.)

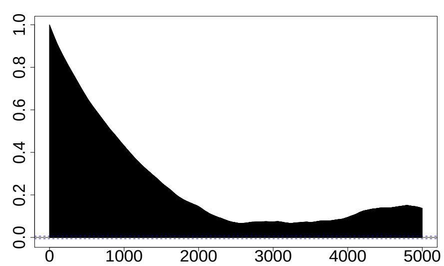
- **Recall**

$$\widehat{\mu}_f = \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

$$\begin{aligned}\mathbf{Var}[\widehat{\mu}_f] &= \frac{1}{n^2}\left[\sum_{i=1}^{n}\mathbf{Var}[f(x_i)] + \sum_{i=1}^{n}\sum_{j\neq i}\mathbf{Cov}[f(x_i), f(x_j)]\right] \\ &= \frac{\mathbf{Var}[f(x)]}{n}\left[1 + 2\sum_{h=1}^{\infty}\rho(h)\right]\end{aligned}$$

  - **note: negative correlations are good!**

- **Two approaches**
  - **estimate the correlation structure**




    * **needs to "cut" the sum somewhere**
    * **different strategies exist**
  - **do several independent runs**
    * **or divide a long run into (almost) independent batches**

# Variance estimation in MCMC (cont.)

- **Do $K$ independent MCMC runs**

$$\widehat{\mu}_f^{(k)} = \frac{1}{n} \sum_{i=1}^{n} f(x_i^{(k)})$$

$$\widehat{\mu}_f = \frac{1}{K} \sum_{k=1}^{K} \widehat{\mu}_f^{(k)}$$

  - **then $\widehat{\mu}_f^{(1)}, \dots, \widehat{\mu}_f^{(K)}$ are independent**

$$\mathbf{Var}[\widehat{\mu}_f] = \frac{\mathbf{Var}[\widehat{\mu}_f^{(\cdot)}]}{K}$$

$$\widehat{\mathbf{Var}}[\widehat{\mu}_f^{(\cdot)}] = \frac{1}{K-1} \sum_{k=1}^{K} \left( \widehat{\mu}_f^{(k)} - \widehat{\mu}_f \right)^2$$

- **Alternatively divide one long run into $K$ batches and treat the batches as independent**

  - **batch lengths must be long compared to correlation length**

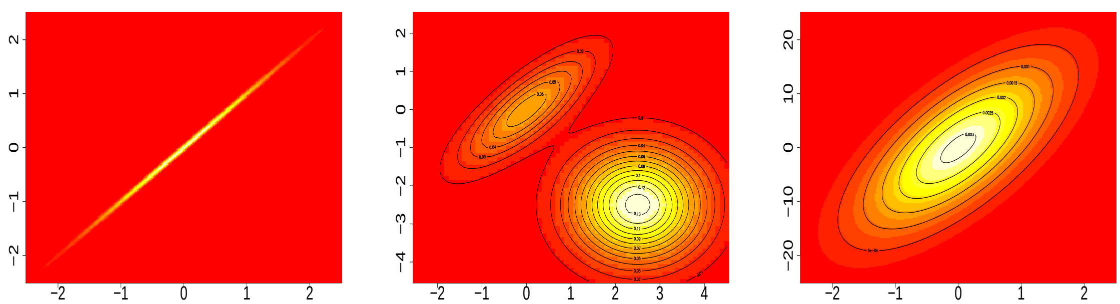
# Plan

- **The Markov chain Monte Carlo (MCMC) idea**
- **Some Markov chain theory**
- **Implementation of the MCMC idea**
  - **Metropolis–Hastings algorithm**
- **MCMC strategies**
  - **independent proposals**
  - **random walk proposals**
  - **combination of strategies**
  - **Gibbs sampler**
- **Convergence diagnostics**
  - **trace plots**
  - **autocorrelation functions**
  - **one chain or many chains?**
- **Typical MCMC problems — and some remedies**
  - **high correlation between variables**
  - **multimodality**
  - **different scales**

# Typical MCMC problems

- **Note: If you knows the solution, it is easy to solve a problem!**

- **Properties of $\pi(x)$ that may make MCMC difficult**
  - **strong dependency between variables**
  - **several modes**
  - **different scales on different variables**



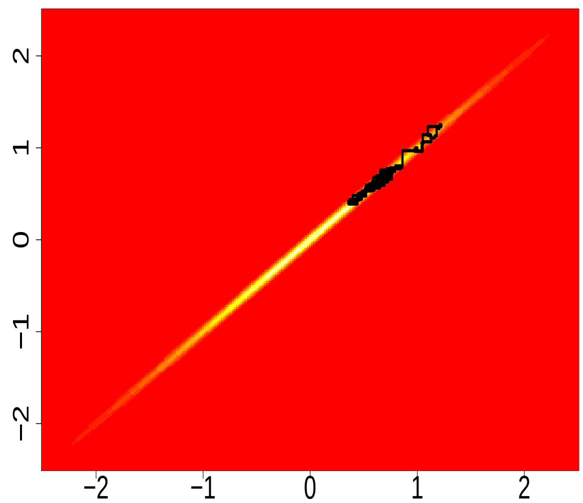
- **In toy examples: this is not a problem**
  - **we know how $\pi(x)$ looks like**

- **In real problems: this may be difficult**
  - **we have a formula for $\pi(x)$**
  - **we don't know how $\pi(x)$ looks like**

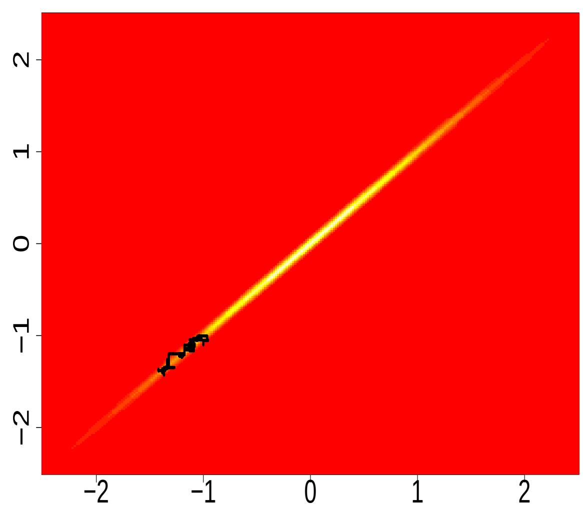
- **Need to iterate**

# Strong dependencies

- Gibbs sampling doesn't work



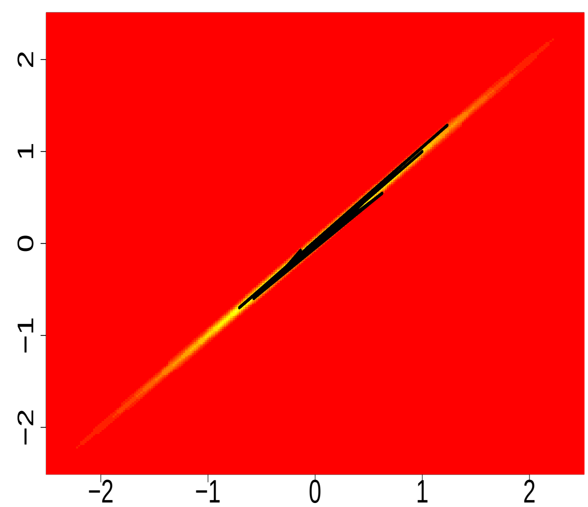

- Changing one variable at a time doesn't work



# Strong dependencies

- Blocking may solve the problem
  - $x = (x^1, x^2, \ldots, x^N)$
  - $x^1$ and $x^2$ are highly correlated
  - propose joint updates for $x^1$ and $x^2$
    - block Gibbs: $(y^1, y^2)|x \sim \pi(y^1, y^2|x^{-\{1,2\}})$
    - random walk Metropolis–Hastings:

$$(y^1, y^2)|x \sim \mathbf{N}_2\left(\left[\begin{array}{c} x^1 \\ x^2 \end{array}\right], R\right)$$



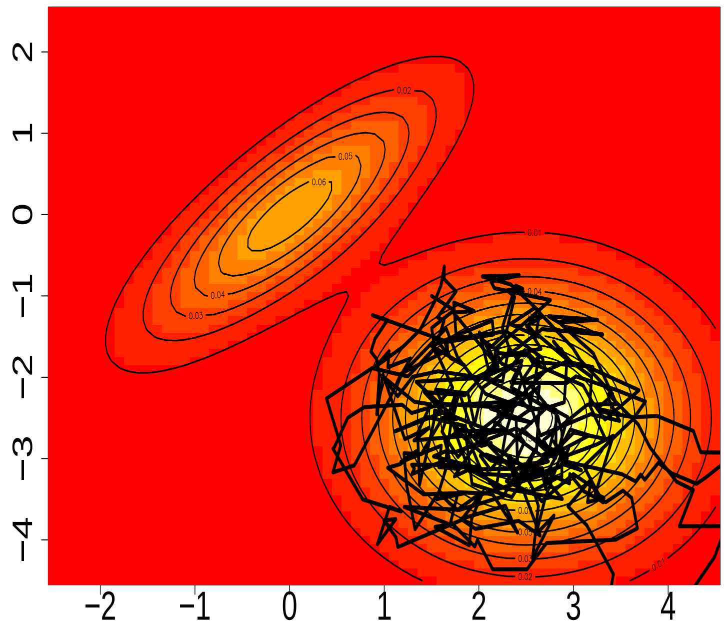

    - in toy example:
      - target: correlation $0.999$
      - in proposal: correlation $0.90$

# Strong dependencies

- **Reparameterisation may solve the problem**
  - $x = (x^1, x^2, \ldots, x^N)$
  - $x^1$ **and** $x^2$ **are highly correlated**
  - **define**
    $$\begin{bmatrix} \tilde{x}^1 \\ \tilde{x}^2 \end{bmatrix} = A \begin{bmatrix} x^1 \\ x^2 \end{bmatrix}$$
    **and**
    $$\tilde{x}^i = x^i \quad \textbf{for } i = 3, \ldots, N$$
  - **with suitable choice of matrix $A$, the correlation between $\tilde{x}^1$ and $\tilde{x}^2$ in $\pi(\tilde{x})$ will be much lower**

# Multimodal target distribution

- Random walk proposals doesn't work




- To come from one mode to another: needs to visit low probability area — happens very seldomly

# Multimodal target distributions

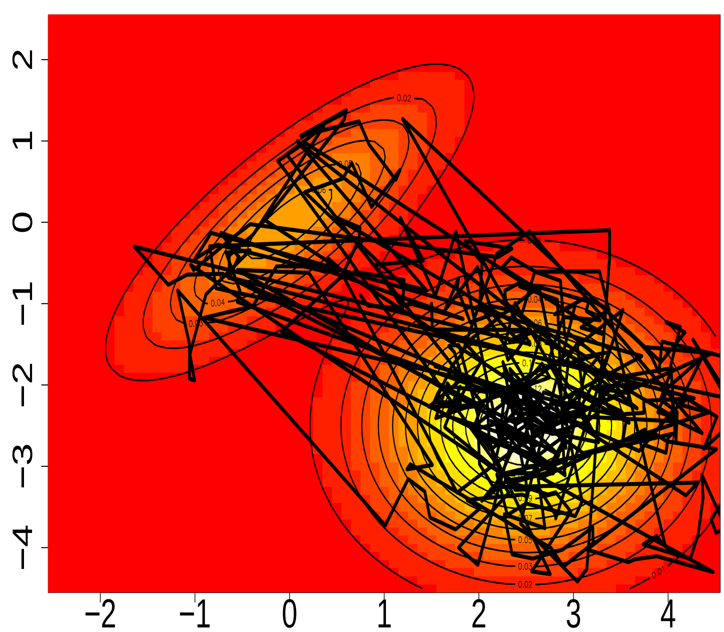
- **If you know (approximately) the modes**
  - **can combine**
    - **independent proposals**

$$y|x \sim \frac{1}{2}g_1(y) + \frac{1}{2}g_2(y)$$

    - **random walk proposals**

$$y|x \sim \mathbf{N}(x, R)$$

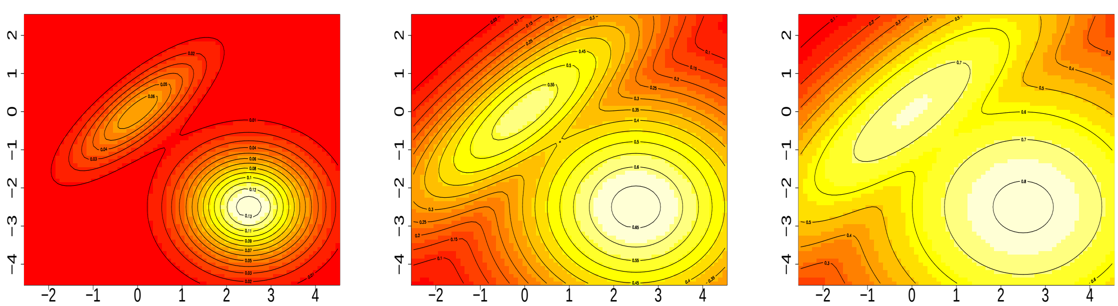
  - **randomly or systematically**



# Multimodal target distributions

- **Simulated tempering**
  - **let**
$$\pi(x) = c \exp\{-U(x)\}$$
  - **introduce an extra variable,** $k \in \{0, 1, 2, \ldots, K\}$
  - **define** $K$ **temperatures:** $1 = T_0 < T_1 < T_2 < \ldots < T_K$
  - **define** $K$ **distributions and constants** $c_0, c_1, \ldots, c_K$
$$\pi_k(x) = c_k \exp\left\{-\frac{1}{T_k}U(x)\right\}$$
    * **note:** $\pi_0(x) = \pi(x)$

$T = 1$ $\qquad$ $T = 5$ $\qquad$ $T = 10$




  - **define joint distribution for** $x$ **and** $k$
$$\pi(x, k) \propto \pi_k(x)$$
  - **simulate from** $\pi(x, k)$ **with Metropolis–Hastings**
  - **keep simulated** $x$**'s that corresponds to** $k = 0$

- **Note: the** $T_k$**'s and** $c_k$**'s must be chosen carefully**

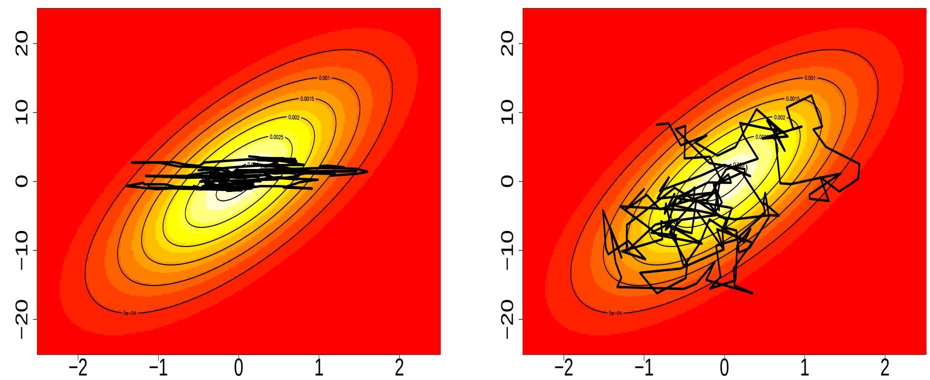
# Multimodal target distributions

- **Other solutions has been proposed**
  - **MCMCMC: Metropolis coupled MCMC**
    - **simulate one $x_k$ for each temperate $T_k$**
    - **simulate each $x_k$ by standard Metropolis-Hastings**
    - **occasionally propose to swap two "neighbour" states $x_k$ and $x_{k+1}$**
    - **accept/reject according to MH acceptance probability**
  - **mode-jumping**
    - **in a Metropolis–Hastings algorithm: use local optimisation to locate a local maximum, then propose a new value from that mode**
    - **more on this in an example later (?)**

# Different scales

- With Gibbs: different scales are not a problem
  - Gibbs finds the appropriate scale
- If Gibbs not possible: have to tune to find appropriate scales

equal scale in proposals | tuned scales in proposals




- Tempting to tune the proposal scales automatically based on the history of the Markov chain
  - careful!! it is no longer Markov
  - more difficult to get the required limiting distribution
  - some *adaptive MCMC* algorithms exist — more later (?)