The kernel of the Gibbs sampler (case $d=2$) is

$$K(x^{(t-1)}, x^{(t)}) = \pi_{X_1|X_2}(x_1^{(t)} | x_2^{(t-1)}) \pi_{X_2|X_1}(x_2^{(t)} | x_1^{(t)})$$

Case $d > 2$:

$$K(x^{(t-1)}, x^{(t)}) = \prod_{j=1}^{d} \pi_{X_j|X_{-j}}(x_j^{(t)} | x_{1:j-1}^{(t)}, x_{j+1:d}^{(t-1)})$$

**Proposition**

The systematic scan Gibbs sampler kernel admits $\pi$ as invariant distribution.
Invariance of the Gibbs sampler II

Proof for $d = 2$.

Let $x = (x_1, x_2)$ and $y = (y_1, y_2)$. Then we have

\[
\int K(x, y)\pi(x)\,dx = \int \pi(y_2 \mid y_1)\pi(y_1 \mid x_2)\pi(x_1, x_2)\,dx_1\,dx_2
\]

\[
= \pi(y_2 \mid y_1) \int \pi(y_1 \mid x_2)\pi(x_2)\,dx_2
\]

\[
= \pi(y_2 \mid y_1)\pi(y_1) = \pi(y_1, y_2) = \pi(y).
\]
Irreducibility and Recurrence

Proposition

Assume \( \pi \) satisfies the positivity condition, then the Gibbs sampler yields a \( \pi \)-irreducible and recurrent Markov chain.

Proof.

Recurrence. Will follow from irreducibility and the fact that \( \pi \) is invariant, \(^a\)

(One step)Irreducibility. Let \( \mathbb{X} \subset \mathbb{R}^d \), such that \( \pi(\mathbb{X}) = 1 \).

Write \( K \) for the kernel and let \( A \subset \mathbb{X} \) such that \( \pi(A) > 0 \).

Then for any \( x \in \mathbb{X} \)

\[
K(x, A) = \int_A K(x, y)dy
\]

\[
= \int_A \pi_{X_1 | X_{-1}}(y_1 | x_2, \ldots, x_d) \times \cdots \times \pi_{X_d | X_{-d}}(y_d | y_1, \ldots, y_{d-1})dy.
\]

\(^a\)Meyn and Tweedie, Markov chains and stochastic stability, Prop’n 10.1.1.
Proof.

Thus if for some \( x \in X \) and \( A \) with \( \pi(A) > 0 \) we have \( K(x, A) = 0 \), we must have that

\[
\pi_{X_1|X_{-1}}(y_1 \mid x_2, \ldots, x_d) \times \cdots \times \pi_{X_d|X_{-d}}(y_d \mid y_1, \ldots, y_{d-1}) = 0,
\]

for almost all \( y = (y_1, \ldots, y_d) \in A \).

Therefore, by the Hammersley-Clifford theorem, we must also have that

\[
\pi(y_1, y_2, \ldots, y_d) \propto \prod_{j=1}^{d} \frac{\pi_{X_j|X_{-j}}(y_j \mid y_1:j-1, x_{j+1:d})}{\pi_{X_j|X_{-j}}(x_j \mid y_1:j-1, x_{j+1:d})} = 0,
\]

for almost all \( y = (y_1, \ldots, y_d) \in A \) and thus \( \pi(A) = 0 \) obtaining a contradiction.

Note: Positivity not necessary for irreducibility; e.g. \( f \propto 1_{|x| \leq 1} \).
Theorem

If the positivity condition is satisfied then for any \( \pi \)-integrable function \( \varphi : \mathbb{X} \rightarrow \mathbb{R} \):

\[
\lim_{t \to \infty} \frac{1}{t} \sum_{i=1}^{t} \varphi \left( X^{(i)} \right) = \int_{\mathbb{X}} \varphi (x) \pi (x) \, dx
\]

for \( \pi \)-almost all starting values \( X^{(1)} \).
Example: Bivariate Normal Distribution

- Let \( X := (X_1, X_2) \sim \mathcal{N}(\mu, \Sigma) \) where \( \mu = (\mu_1, \mu_2) \) and

\[
\Sigma = \begin{pmatrix}
\sigma_1^2 & \rho \\
\rho & \sigma_2^2
\end{pmatrix}.
\]

- The Gibbs sampler proceeds as follows in this case

(a) Sample \( X_1^{(t)} \sim \mathcal{N} \left( \mu_1 + \rho/\sigma_2^2 \left( X_2^{(t-1)} - \mu_2 \right), \sigma_1^2 - \rho^2/\sigma_2^2 \right) \)

(b) Sample \( X_2^{(t)} \sim \mathcal{N} \left( \mu_2 + \rho/\sigma_1^2 \left( X_1^{(t)} - \mu_1 \right), \sigma_2^2 - \rho^2/\sigma_1^2 \right) \).

- By proceeding this way, we generate a Markov chain \( X^{(t)} \) whose successive samples are correlated. If successive values of \( X^{(t)} \) are strongly correlated, then we say that the Markov chain mixes slowly.
Bivariate Normal Distribution

Figure: Case where $\rho = 0.1$, first 100 steps.
Bivariate Normal Distribution

Figure: Case where $\rho = 0.99$, first 100 steps.
Bivariate Normal Distribution

Figure: Histogram of the first component of the chain after 1000 iterations. Small $\rho$ on the left, large $\rho$ on the right.
Figure: Histogram of the first component of the chain after 10000 iterations. Small $\rho$ on the left, large $\rho$ on the right.
Figure: Histogram of the first component of the chain after 100000 iterations. Small $\rho$ on the left, large $\rho$ on the right.
Gibbs Sampling and Auxiliary Variables

- Gibbs sampling requires sampling from $\pi_{X_j|X_{-j}}$.
- In many scenarios, we can include a set of auxiliary variables $Z_1,...,Z_p$ and have an “extended” distribution of joint density $\pi(x_1,...,x_d,z_1,...,z_p)$ such that

$$\int \pi(x_1,...,x_d,z_1,...,z_p) dz_1...dz_d = \pi(x_1,...,x_d).$$

which is such that its full conditionals are easy to sample.
- Mixture models, Capture-recapture models, Tobit models, Probit models etc.
Independent data $y_1, \ldots, y_n$

$$Y_i | \theta \sim \sum_{k=1}^{K} p_k \mathcal{N} \left( \mu_k, \sigma_k^2 \right)$$

where $\theta = \left( p_1, \ldots, p_K, \mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2 \right)$. 
Bayesian Model

- Likelihood function

\[
p(y_1, \ldots, y_n | \theta) = \prod_{i=1}^{n} p(y_i | \theta) = \prod_{i=1}^{n} \left( \sum_{k=1}^{K} \frac{p_k}{\sqrt{2\pi \sigma_k^2}} \exp \left( -\frac{(y_i - \mu_k)^2}{2\sigma_k^2} \right) \right).
\]

Let’s fix \( K = 2, \sigma_k^2 = 1 \) and \( p_k = 1/K \) for all \( k \).

- Prior model

\[
p(\theta) = \prod_{k=1}^{K} p(\mu_k)
\]

where

\( \mu_k \sim \mathcal{N}(\alpha_k, \beta_k) \).

Let us fix \( \alpha_k = 0, \beta_k = 1 \) for all \( k \).

- Not obvious how to sample \( p(\mu_1 | \mu_2, y_1, \ldots, y_n) \).
Auxiliary Variables for Mixture Models

- Associate to each $Y_i$ an auxiliary variable $Z_i \in \{1, \ldots, K\}$ such that

$$\mathbb{P}(Z_i = k|\theta) = p_k \quad \text{and} \quad Y_i|Z_i = k, \theta \sim \mathcal{N}\left(\mu_k, \sigma_k^2\right)$$

so that

$$p(y_i|\theta) = \sum_{k=1}^{K} \mathbb{P}(Z_i = k) \mathcal{N}(y_i; \mu_k, \sigma_k^2)$$

- The extended posterior is given by

$$p(\theta, z_1, \ldots, z_n|y_1, \ldots, y_n) \propto p(\theta) \prod_{i=1}^{n} \mathbb{P}(z_i|\theta) p(y_i|z_i, \theta).$$

- Gibbs samples alternately

$$\mathbb{P}(z_{1:n}|y_{1:n}, \mu_{1:K})$$

$$p(\mu_{1:K}|y_{1:n}, z_{1:n}).$$
Gibbs Sampling for Mixture Model

- We have

\[ P(z_{1:n} \mid y_{1:n}, \theta) = \prod_{i=1}^{n} P(z_i \mid y_i, \theta) \]

where

\[ P(z_i \mid y_i, \theta) = \frac{P(z_i \mid \theta) p(y_i \mid z_i, \theta)}{\sum_{k=1}^{K} P(z_i = k \mid \theta) p(y_i \mid z_i = k, \theta)} \]

- Let \( n_k = \sum_{i=1}^{n} 1_{\{k\}}(z_i), n_k \bar{y}_k = \sum_{i=1}^{n} y_i 1_{\{k\}}(z_i) \) then

\[ \mu_k \mid z_{1:n}, y_{1:n} \sim N \left( \frac{n_k \bar{y}_k}{1 + n_k}, \frac{1}{1 + n_k} \right). \]
Mixtures of Normals

Figure: 200 points sampled from $\frac{1}{2} \mathcal{N}(-2, 1) + \frac{1}{2} \mathcal{N}(2, 1)$. 
Mixtures of Normals

Figure: Histogram of the parameters obtained by 10,000 iterations of Gibbs sampling.
Mixtures of Normals

Figure: Traceplot of the parameters obtained by 10,000 iterations of Gibbs sampling.
Gibbs sampling in practice

- Many posterior distributions can be automatically decomposed into conditional distributions by computer programs.

- This is the idea behind **BUGS** (Bayesian inference Using Gibbs Sampling), **JAGS** (Just another Gibbs Sampler).
Gibbs Recap

• Given a target $\pi(x) = \pi(x_1, x_2, ..., x_d)$, Gibbs sampling works by sampling from $\pi_{X_j|X_{-j}}(x_j|x_{-j})$ for $j = 1, ..., d$.

• Sampling exactly from one of these full conditionals might be a hard problem itself.

• Even if it is possible, the Gibbs sampler might converge slowly if components are highly correlated.

• If the components are not highly correlated then Gibbs sampling performs well, even when $d \to \infty$, e.g. with an error increasing "only" polynomially with $d$.

• Metropolis–Hastings algorithm (1953, 1970) is a more general algorithm that can bypass these problems.

• Additionally Gibbs can be recovered as a special case.
Metropolis–Hastings algorithm

- Target distribution on $\mathbb{X} = \mathbb{R}^d$ of density $\pi(x)$.
- Proposal distribution: for any $x, x' \in \mathbb{X}$, we have $q(x'|x) \geq 0$ and $\int_{\mathbb{X}} q(x'|x) \, dx' = 1$.
- Starting with $X^{(1)}$, for $t = 2, 3, \ldots$
  (a) Sample $X^* \sim q(\cdot \mid X^{(t-1)})$.
  (b) Compute

$$\alpha(X^* \mid X^{(t-1)}) = \min \left( 1, \frac{\pi(X^*) \, q(X^{(t-1)} \mid X^*)}{\pi(X^{(t-1)}) \, q(X^* \mid X^{(t-1)})} \right).$$

  (c) Sample $U \sim \mathcal{U}_{[0,1]}$. If $U \leq \alpha(X^* \mid X^{(t-1)})$, set $X^{(t)} = X^*$, otherwise set $X^{(t)} = X^{(t-1)}$. 

Metropolis–Hastings algorithm

Figure: Metropolis–Hastings on a bivariate Gaussian target.
Metropolis–Hastings algorithm

Figure: Metropolis–Hastings on a bivariate Gaussian target.
Metropolis–Hastings algorithm

Figure: Metropolis–Hastings on a bivariate Gaussian target.
Metropolis–Hastings algorithm

Figure: Metropolis–Hastings on a bivariate Gaussian target.
Figure: Metropolis–Hastings on a bivariate Gaussian target.
Metropolis–Hastings algorithm

Figure: Metropolis–Hastings on a bivariate Gaussian target.
Metropolis–Hastings algorithm

Figure: Metropolis–Hastings on a bivariate Gaussian target.
Metropolis–Hastings algorithm

Figure: Metropolis–Hastings on a bivariate Gaussian target.
Metropolis–Hastings algorithm

Figure: Metropolis–Hastings on a bivariate Gaussian target.
Metropolis–Hastings algorithm

Figure: Metropolis–Hastings on a bivariate Gaussian target.
Metropolis–Hastings algorithm

Figure: Metropolis–Hastings on a bivariate Gaussian target.
Metropolis–Hastings algorithm

Figure: Metropolis–Hastings on a bivariate Gaussian target.
Metropolis–Hastings algorithm

Figure: Metropolis–Hastings on a bivariate Gaussian target.
Metropolis–Hastings algorithm

Figure: Metropolis–Hastings on a bivariate Gaussian target.
Figure: Metropolis–Hastings on a bivariate Gaussian target.
Metropolis–Hastings algorithm

Figure: Metropolis–Hastings on a bivariate Gaussian target.
Metropolis–Hastings algorithm

Figure: Metropolis–Hastings on a bivariate Gaussian target.
Metropolis–Hastings algorithm

Figure: Metropolis–Hastings on a bivariate Gaussian target.
Metropolis–Hastings algorithm

Figure: Metropolis–Hastings on a bivariate Gaussian target.
Figure: Metropolis–Hastings on a bivariate Gaussian target.
Metropolis–Hastings algorithm

- Metropolis–Hastings only requires point-wise evaluations of $\pi(x)$ up to a normalizing constant; indeed if $\tilde{\pi}(x) \propto \pi(x)$ then
  
  \[
  \frac{\pi(x^*)}{\pi(x^{(t-1)})} \frac{q(x^{(t-1)} | x^*)}{q(x^* | x^{(t-1)})} = \frac{\tilde{\pi}(x^*)}{\tilde{\pi}(x^{(t-1)})} \frac{q(x^{(t-1)} | x^*)}{q(x^* | x^{(t-1)})}.
  \]

- At each iteration $t$, a candidate is proposed.
- The **average acceptance probability** from the current state is
  
  \[
  a(x^{(t-1)}) := \int_X \alpha(x | x^{(t-1)}) q(x | x^{(t-1)}) \, dx
  \]
  
  in which case $X^{(t)} = X$, otherwise $X^{(t)} = X^{(t-1)}$.
- This algorithm clearly defines a Markov chain $(X^{(t)})_{t \geq 1}$. 
Lemma

The kernel of the Metropolis–Hastings algorithm is given by

$$K(y | x) \equiv K(x, y) = \alpha(y | x)q(y | x) + (1 - a(x))\delta_x(y).$$

Proof.

We have

$$K(x, y)$$

$$= \int q(x^* | x)\{\alpha(x^* | x)\delta_{x^*}(y) + (1 - \alpha(x^* | x))\delta_x(y)\}dx^*$$

$$= q(y | x)\alpha(y | x) + \left\{\int q(x^* | x)(1 - \alpha(x^* | x))dx^*\right\}\delta_x(y)$$

$$= q(y | x)\alpha(y | x) + \left\{1 - \int q(x^* | x)\alpha(x^* | x)dx^*\right\}\delta_x(y)$$

$$= q(y | x)\alpha(y | x) + \{1 - a(x)\}\delta_x(y).$$
Reversibility

Proposition

The Metropolis–Hastings kernel $K$ is $\pi$–reversible and thus admits $\pi$ as invariant distribution.

Proof.

For any $x, y \in \mathbb{X}$, with $x \neq y$

$$
\pi(x)K(x, y) = \pi(x)q(y \mid x)\alpha(y \mid x)
$$

$$
= \pi(x)q(y \mid x) \left(1 \land \frac{\pi(y)q(x \mid y)}{\pi(x)q(y \mid x)}\right)
$$

$$
= \left(\pi(x)q(y \mid x) \land \pi(y)q(x \mid y)\right)
$$

$$
= \pi(y)q(x \mid y) \left(\frac{\pi(x)q(y \mid x)}{\pi(y)q(x \mid y)} \land 1\right) = \pi(y)K(y, x).
$$

If $x = y$, then obviously $\pi(x)K(x, y) = \pi(y)K(y, x)$. \qed
Reducibility and periodicity of Metropolis–Hastings

- Consider the target distribution

\[ \pi(x) = \left( \mathcal{U}_{[0,1]}(x) + \mathcal{U}_{[2,3]}(x) \right) / 2 \]

and the proposal distribution

\[ q(x^* | x) = \mathcal{U}_{(x-\delta, x+\delta)}(x^*) \cdot \]

- The MH chain is reducible if \( \delta \leq 1 \): the chain stays either in \([0,1]\) or \([2,3]\).

- Note that the MH chain is aperiodic if it always has a non-zero chance of staying where it is.
Some results

**Proposition**

If \( q(x^* | x) > 0 \) for any \( x, x^* \in \text{supp}(\pi) \) then the Metropolis-Hastings chain is irreducible, in fact every state can be reached in a single step (strongly irreducible).

Less strict conditions in (Roberts & Rosenthal, 2004).

**Proposition**

If the MH chain is irreducible then it is also Harris recurrent (see Tierney, 1994).
LLN for MH

Theorem

If the Markov chain generated by the Metropolis–Hastings sampler is $\pi$–irreducible, then we have for any integrable function $\varphi : X \to \mathbb{R}$:

$$\lim_{t \to \infty} \frac{1}{t} \sum_{i=1}^{t} \varphi \left( X^{(i)} \right) = \int_{X} \varphi \left( x \right) \pi \left( x \right) dx$$

for every starting value $X^{(1)}$. 
Random Walk Metropolis–Hastings

- In the Metropolis–Hastings, pick \( q(x^* \mid x) = g(x^* - x) \) with \( g \) being a symmetric distribution, thus

\[
X^* = X + \varepsilon, \quad \varepsilon \sim g;
\]
e.g. \( g \) is a zero-mean multivariate normal or t-student.

- Acceptance probability becomes

\[
\alpha(x^* \mid x) = \min\left(1, \frac{\pi(x^*)}{\pi(x)}\right).
\]

- We accept...
  - a move to a more probable state with probability \( 1 \);
  - a move to a less probable state with probability

\[
\pi(x^*) / \pi(x) \leq 1.
\]
Independent Metropolis–Hastings

- **Independent proposal**: a proposal distribution \( q(x^* \mid x) \) which does not depend on \( x \).

- Acceptance probability becomes

  \[
  \alpha(x^* \mid x) = \min\left(1, \frac{\pi(x^*) q(x)}{\pi(x) q(x^*)}\right).
  \]

- For instance, multivariate normal or t-student distribution.

- If \( \pi(x)/q(x) < M \) for all \( x \) and some \( M < \infty \), then the chain is **uniformly ergodic**.

- The acceptance probability at stationarity is at least \( 1/M \) (Lemma 7.9 of Robert & Casella).

- On the other hand, if such an \( M \) does not exist, the chain is not even geometrically ergodic!
Choosing a good proposal distribution

- **Goal:** design a Markov chain with small correlation $\rho(X^{(t-1)}, X^{(t)})$ between subsequent values (why?).

- Two sources of correlation:
  - between the current state $X^{(t-1)}$ and proposed value $X \sim q(\cdot | X^{(t-1)})$,
  - correlation induced if $X^{(t)} = X^{(t-1)}$, if proposal is rejected.

- Trade-off: there is a compromise between
  - proposing large moves,
  - obtaining a decent acceptance probability.

- For multivariate distributions: covariance of proposal should reflect the covariance structure of the target.
Choice of proposal

• Target distribution, we want to sample from

\[ \pi(x) = \mathcal{N}(x; \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}) . \]

• We use a random walk Metropolis—Hastings algorithm with

\[ g(\varepsilon) = \mathcal{N}(\varepsilon; 0, \sigma^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}) . \]

• What is the optimal choice of \( \sigma^2 \)?
• We consider three choices: \( \sigma^2 = 0.1^2, 1, 10^2 \).
Metropolis–Hastings algorithm

![Graph showing Metropolis-Hastings for a bivariate Gaussian target.]

Figure: Metropolis–Hastings on a bivariate Gaussian target. With $\sigma^2 = 0.1^2$, the acceptance rate is $\approx 94\%$. 
Metropolis–Hastings algorithm

Figure: Metropolis–Hastings on a bivariate Gaussian target. With $\sigma^2 = 0.1^2$, the acceptance rate is $\approx 94\%$. 
Figure: Metropolis–Hastings on a bivariate Gaussian target. With $\sigma^2 = 1$, the acceptance rate is $\approx 52\%$. 
Figure: Metropolis–Hastings on a bivariate Gaussian target. With $\sigma^2 = 1$, the acceptance rate is $\approx 52\%$. 
Metropolis–Hastings algorithm

Figure: Metropolis–Hastings on a bivariate Gaussian target. With $\sigma^2 = 10$, the acceptance rate is $\approx 1.5\%$. 
Figure: Metropolis–Hastings on a bivariate Gaussian target. With $\sigma^2 = 10$, the acceptance rate is $\approx 1.5\%$. 
Choice of proposal

- Aim at some intermediate acceptance ratio: 20%? 40%? Some hints come from the literature on “optimal scaling”.
- Literature suggest tuning to get .234...
- Maximize the expected square jumping distance:

\[ \mathbb{E} \left[ \| X_{t+1} - X_t \|^2 \right] \]

- In multivariate cases, try to mimic the covariance structure of the target distribution.

Cooking recipe: run the algorithm for \( T \) iterations, check some criterion, tune the proposal distribution accordingly, run the algorithm for \( T \) iterations again . . .

“Constructing a chain that mixes well is somewhat of an art.”

*All of Statistics*, L. Wasserman.
The adaptive MCMC approach

• One can make the transition kernel $K$ adaptive, i.e. use $K_t$ at iteration $t$ and choose $K_t$ using the past sample $(X_1,\ldots,X_{t-1})$.

• The Markov chain is not homogeneous anymore: the mathematical study of the algorithm is much more complicated.

• Adaptation can be counterproductive in some cases (see Atchadé & Rosenthal, 2005)!

• Adaptive Gibbs samplers also exist.
  △ Extreme care is needed when designing adaptive algorithms: it’s easy to make an algorithm with the wrong invariant distribution.
Sophisticated Proposals

- “Langevin” proposal relies on

\[
X^* = X^{(t-1)} + \frac{\sigma}{2} \nabla \log \pi \left( X^{(t-1)} \right) + \sigma W
\]

where \( W \sim \mathcal{N}(0, I_d) \), so the Metropolis-Hastings acceptance ratio is

\[
\frac{\pi(X^*) q(X^{(t-1)} | X^*)}{\pi(X^{(t-1)}) q(X^* | X^{(t-1)})} = \frac{\pi(X^*)}{\pi(X^{(t-1)})} \frac{\mathcal{N}(X^{(t-1)}; X^* + \frac{\sigma}{2} \nabla \log \pi (X^*); \sigma^2)}{\mathcal{N}(X^*; X^{(t-1)} + \frac{\sigma}{2} \nabla \log \pi (X^{(t-1)}); \sigma^2)}.
\]

- Possibility to use higher order derivatives:

\[
X^* = X^{(t-1)} + \frac{\sigma}{2} \left[ \nabla^2 \log \pi \left( X^{(t-1)} \right) \right]^{-1} \nabla \log \pi \left( X^{(t-1)} \right) + \sigma W.
\]
Sophisticated Proposals

- We can use

\[ q(X^*|X^{(t-1)}) = g(X^*; \varphi(X^{(t-1)})) \]

where \( g \) is a distribution on \( \mathbb{X} \) of parameters \( \varphi(X^{(t-1)}) \) and \( \varphi \) is a deterministic mapping

\[
\frac{\pi(X^*) q(X^{(t-1)}|X^*)}{\pi(X^{(t-1)}) q(X^*|X^{(t-1)})} = \frac{\pi(X^*) g(X^{(t-1)}; \varphi(X^*))}{\pi(X^{(t-1)}) g(X^*; \varphi(X^{(t-1)}))}.
\]

- For instance, use heuristics borrowed from optimization techniques.
Sophisticated Proposals

The following link shows a comparison of

- adaptive Metropolis-Hastings,
- Gibbs sampling,
- No U-Turn Sampler (e.g. Hamiltonian MCMC)

on a simple linear model.

twiecki.github.io/blog/2014/01/02/visualizing-mcmc/
Sophisticated Proposals

- Assume you want to sample from a target \( \pi \) with \( \text{supp}(\pi) \subset \mathbb{R}^+ \), e.g. the posterior distribution of a variance/scale parameter.

- Any proposed move, e.g. using a normal random walk, to \( \mathbb{R}^- \) is a waste of time.

- Given \( X^{(t-1)} \), propose \( X^* = \exp(\log X^{(t-1)} + \epsilon) \) with \( \epsilon \sim \mathcal{N}(0, \sigma^2) \). What is the acceptance probability then?

\[
\alpha(X^* \mid X^{(t-1)}) = \min \left( 1, \frac{\pi(X^*)}{\pi(X^{(t-1)})} \frac{q(X^{(t-1)} \mid X^*)}{q(X^* \mid X^{(t-1)})} \right)
\]

\[
= \min \left( 1, \frac{\pi(X^*)}{\pi(X^{(t-1)})} \frac{X^*}{X^{(t-1)}} \right).
\]

Why?

\[
q(y \mid x) = \frac{1}{y \sigma \sqrt{2\pi}} \exp \left[ - \frac{(\log y - \log x)^2}{2\sigma^2} \right] = x
\]

\[
q(x \mid y) = \frac{1}{x \sigma \sqrt{2\pi}} \exp \left[ - \frac{(\log x - \log y)^2}{2\sigma^2} \right] = \frac{x}{y}.
\]
Random Proposals

- Assume you want to use $q_{\sigma^2}(X^*|X^{(t-1)}) = \mathcal{N}(X; X^{(t-1)}, \sigma^2)$ but you don’t know how to pick $\sigma^2$. You decide to pick a random $\sigma^2, *$ from a distribution $f(\sigma^2)$:

$$\sigma^2, * \sim f(\sigma^2, *), \quad X^*|\sigma^2, * \sim q_{\sigma^2, *}(\cdot|X^{(t-1)})$$

so that

$$q(X^*|X^{(t-1)}) = \int q_{\sigma^2, *}(X^*|X^{(t-1)}) f(\sigma^2, *) d\sigma^2, *.$$

- Perhaps $q(X^*|X^{(t-1)})$ cannot be evaluated, e.g. the above integral is intractable. Hence the acceptance probability

$$\min\{1, \frac{\pi(X^*) q(X^{(t-1)}|X^*)}{\pi(X^{(t-1)}) q(X^*|X^{(t-1)})}\}$$

cannot be computed.
Random Proposals

• Instead you decide to accept your proposal with probability

\[ \alpha_t = \min \left\{ 1, \frac{\pi(X^*) q_{\sigma^2,(t-1)} \left( X^{(t-1)} \mid X^* \right)}{\pi(X^{(t-1)}) q_{\sigma^2,*} \left( X^* \mid X^{(t-1)} \right)} \right\} \]

where \( \sigma^2,(t-1) \) corresponds to parameter of the last accepted proposal.

• With probability \( \alpha_t \), set \( \sigma^2,(t) = \sigma^2,* \), \( X^{(t)} = X^* \), otherwise \( \sigma^2,(t) = \sigma^2,(t-1) \), \( X^{(t)} = X^{(t-1)} \).

• Question: Is it valid? If so, why?
Random Proposals

• Consider the extended target

\[ \tilde{\pi}(x, \sigma^2) := \pi(x) f(\sigma^2). \]

• Previous algorithm is a Metropolis-Hastings of target \( \tilde{\pi}(x, \sigma^2) \) and proposal

\[ q(y, \tau^2 | x, \sigma^2) = f(\tau^2) q_{\tau^2}(y|x) \]

• Indeed, we have

\[
\frac{\tilde{\pi}(y, \tau^2)}{\tilde{\pi}(x, \sigma^2)} \frac{q(x, \sigma^2 | y, \tau^2)}{q(y, \tau^2 | x, \sigma^2)}
= \frac{\pi(y) f(\tau^2) f(\sigma^2) q_{\sigma^2}(x|y)}{\pi(x) f(\sigma^2) f(\tau^2) q_{\tau^2}(y|x)}
= \frac{\pi(y)}{\pi(x)} \frac{q_{\sigma^2}(x|y)}{q_{\tau^2}(y|x)}
\]

• **Remark**: we just need to be able to sample from \( f(\cdot) \), not to evaluate it.
Using multiple proposals

- MH with target $\pi(x)$ where $x \in \mathbb{X}$.
- Can’t choose between proposals $q_1(x'|x)$, $q_2(x'|x)$, ..., $q_p(x'|x)$.
- If you build a **mixture proposal**

\[
q(x'|x) = \sum_{j=1}^{p} \beta_j q_j(x'|x), \quad \beta_j > 0, \quad \sum_{j=1}^{p} \beta_j = 1,
\]

then you have to evaluate $q_j(X^*|X^{(t-1)})$ for $j = 1, \ldots, p$. 
Composing kernels

• How to use different proposals to sample from $\pi$ without evaluating all the densities at each step?

• Instead combine Metropolis-Hastings updates $K_j$ using proposal $q_j$ instead? i.e.

$$K_j(x, x') = \alpha_j(x'|x) q_j(x'|x) + (1 - a_j(x)) \delta_x(x')$$

where

$$\alpha_j(x'|x) = \min \left(1, \frac{\pi(x') q_j(x|x')}{\pi(x) q_j(x'|x)} \right)$$

$$a_j(x) = \int \alpha_j(x'|x) q_j(x'|x) dx'.$$
Composing kernels

Generally speaking, assume

- $p$ possible updates characterised by kernels $K_j(\cdot, \cdot)$,
- each kernel $K_j$ is $\pi$-invariant.

Two ways to combine the $p$ MCMC updates:

- **Cycle**: perform the MCMC updates in a deterministic order.
- **Mixture**: Pick an MCMC update at random.
Cycle of MCMC updates

- Starting with $X^{(1)}$ iterate for $t = 2, 3, ...$
  (a) Set $Z^{(t,0)} := X^{(t-1)}$.
  (b) For $j = 1, ..., p$, sample $Z^{(t,j)} \sim K_j \left( Z^{(t,j-1)}, . \right)$.
  (c) Set $X^{(t)} := Z^{(t,p)}$.

- Full cycle transition kernel is
  \[
  K(x, y) = \int \cdots \int K_1(x, z_1) K_2(z_1, z_2) \\
  \cdots K_p \left( z_{p-1}, y \right) dz_1 \cdots dz_p.
  \]

- $K$ is $\pi$-invariant.
Mixture of MCMC updates

• Starting with $X^{(1)}$ iterate for $t = 2, 3, ...$
  (a) Sample $J$ from $\{1, ..., p\}$ with $\mathbb{P}(J = k) = \beta_k$.
  (b) Sample $X^{(t)} \sim K_J(X^{(t-1)}, \cdot)$.
• Corresponding transition kernel is
  \[
  K(x, y) = \sum_{j=1}^{p} \beta_j K_j(x, y).
  \]
• $K$ is $\pi$-invariant.
• The algorithm is different from using a mixture proposal
  \[
  q(x'|x) = \sum_{j=1}^{p} \beta_j q_j(x'|x).
  \]
If \( \dim(\mathcal{X}) \) is large, it might be very difficult to design a “good” proposal \( q(x'|x) \).

As in Gibbs sampling, we might want to partition \( x \) into \( x = (x_1, \ldots, x_d) \) and denote \( x_{-j} := x \setminus \{x_j\} \).

We propose “local” proposals where only \( x_j \) is updated

\[
q_j(x'|x) = q_j(x'_j|x) \delta_{x_{-j}}(x'_{-j}).
\]

- propose new component \( j \)
- keep other components fixed
Metropolis-Hastings Design for Multivariate Targets

This yields

\[
\alpha_j(x, x') = \min \left( 1, \frac{\pi(x'_j, x_j) q_j(x_j \mid x_j, x'_j) \delta_{x'_j}(x'_j)}{\pi(x_j, x'_j) q_j(x'_j \mid x_j, x'_j) \delta_{x_j}(x'_j)} \right) = 1
\]

\[
= \min \left( 1, \frac{\pi(x'_j, x_j) q_j(x_j \mid x'_j, x_j)}{\pi(x_j, x'_j) q_j(x'_j \mid x_j, x'_j)} \right)
\]

\[
= \min \left( 1, \frac{\pi_{X_j|x_j}(x'_j \mid x_j) q_j(x_j \mid x'_j, x_j)}{\pi_{X_j|x_j}(x_j \mid x_j) q_j(x'_j \mid x_j, x_j)} \right).
\]
One-at-a-time MH (cycle/systematic scan)

Starting with \( X^{(1)} \) iterate for \( t = 2, 3, \ldots \)
For \( j = 1, \ldots, d \),
- Sample \( X^* \sim q_j(\cdot|X_1^{(t)}, \ldots, X_{j-1}^{(t)}, X_j^{(t-1)}, \ldots, X_d^{(t-1)}) \).
- Compute
\[
\alpha_j = \min \left( 1, \frac{\pi_{X_j|X_{-j}} \left( X_j^* | X_1^{(t)} \ldots X_{j-1}^{(t)}, X_{j+1}^{(t)} \ldots X_d^{(t-1)} \right)}{\pi_{X_j|X_{-j}} \left( X_j^{(t-1)} | X_1^{(t)} \ldots X_{j-1}^{(t)}, X_{j+1}^{(t-1)} \ldots X_d^{(t-1)} \right)} \times \frac{q_j \left( X_j^{(t-1)} | X_1^{(t)} \ldots X_{j-1}^{(t)}, X_j^*, X_{j+1}^{(t-1)} \ldots X_d^{(t-1)} \right)}{q_j \left( X_j^* | X_1^{(t)} \ldots X_{j-1}^{(t)}, X_j^{(t-1)}, X_{j+1}^{(t-1)} \ldots X_d^{(t-1)} \right)} \right).
\]
- With probability \( \alpha_j \), set \( X^{(t)} = X^* \), otherwise set \( X^{(t)} = X^{(t-1)} \).
One-at-a-time MH (mixture/random scan)

Starting with $X^{(1)}$ iterate for $t = 2, 3, ...$

- Sample $J$ from $\{1, \ldots, d\}$ with $\mathbb{P}(J = k) = \beta_k$.
- Sample $X^* \sim q_J\left(\cdot | X_1^{(t)}, \ldots, X_d^{(t-1)}\right)$.
- Compute

$$\alpha_J = \min\left(1, \frac{\pi_{X_J|X_{-J}}\left(X_J^* | X_1^{(t-1)} \ldots X_{J-1}^{(t-1)}, X_{J+1}^{(t-1)} \ldots\right)}{\pi_{X_J|X_{-J}}\left(X_J^{(t-1)} | X_1^{(t-1)} \ldots X_{J-1}^{(t-1)}, X_{J+1}^{(t-1)} \ldots\right)} \times \frac{q_J\left(X_J^{(t-1)} | X_1^{(t-1)} \ldots X_{J-1}^{(t-1)}, X_J^*, X_{J+1}^{(t-1)} \ldots X_d^{(t-1)}\right)}{q_J\left(X_J^* | X_1^{(t-1)} \ldots X_{J-1}^{(t-1)}, X_J^{(t-1)}, X_{J+1}^{(t-1)} \ldots X_d^{(t-1)}\right)}\right).$$

- With probability $\alpha_J$ set $X^{(t)} = X^*$, otherwise $X^{(t)} = X^{(t-1)}$. 


Gibbs Sampler as a Metropolis-Hastings algorithm

**Proposition**

The systematic Gibbs sampler is a cycle of one-at-a time MH whereas the random scan Gibbs sampler is a mixture of one-at-a time MH where

\[
q_j \left( x'_j \mid x \right) = \pi_{X_j \mid X_{-j}} \left( x'_j \mid x_{-j} \right).
\]

**Proof.**

It follows from

\[
\frac{\pi \left( x_{-j}, x'_j \right) q_j \left( x_j \mid x_{-j}, x'_j \right)}{\pi \left( x_{-j}, x_j \right) q_j \left( x'_j \mid x_{-j}, x_j \right)} = \frac{\pi \left( x_{-j} \right) \pi_{X_j \mid X_{-j}} \left( x'_j \mid x_{-j} \right) \pi_{X_j \mid X_{-j}} \left( x_j \mid x_{-j} \right)}{\pi \left( x_{-j} \right) \pi_{X_j \mid X_{-j}} \left( x_j \mid x_{-j} \right) \pi_{X_j \mid X_{-j}} \left( x'_j \mid x_{-j} \right)} = 1. \]
This is not a Gibbs sampler

Consider a case where $d = 2$. From $X_1^{(t-1)}, X_2^{(t-1)}$ at time $t-1$:

- Sample $X_1^* \sim \pi(X_1 | X_2^{(t-1)})$, then $X_2^* \sim \pi(X_2 | X_1^*)$. The proposal is then $X^* = (X_1^*, X_2^*)$.

- Compute

$$
\alpha_t = \min \left( 1, \frac{\pi(X_1^*, X_2^*)}{\pi(X_1^{(t-1)}, X_2^{(t-1)})} \right) \frac{q(X^{(t-1)} | X^*)}{q(X^* | X^{(t-1)})}
$$

- Accept $X^*$ or not based on $\alpha_t$, where here

$$
\alpha_t \neq 1
$$

!!
Pseudomarginal

- Suppose the target is itself an intractable integral

\[ \pi(x) = \int f(x, u) \, du, \]

where \( u \) is potentially high-dimensional, and \( f(x, u) \) is a p.d.f.

- Common scenario when there are a lot of latent variables, e.g. random effects models, hidden-Markov models, latent Dirichlet allocation from machine learning.

- Computing \( \pi(\cdot) \) is infeasible or impossible.

- BUT, we can estimate \( \pi(x) \) unbiasedly as

\[ \hat{\pi}(x) = \frac{1}{N} \sum_{i=1}^{N} p(x, U_i), \quad U_i \overset{\text{i.i.d.}}{\sim} f_{U|x}(\cdot|x). \]
Pseudomarginal

- Idea: use $\hat{\pi}(x)$ instead of $\pi$ in the Metropolis acceptance probability,

$$
\alpha(x, y) = \min \left\{ 1, \frac{\hat{\pi}(y) q(y, x)}{\hat{\pi}(x) q(x, y)} \right\}.
$$

- Amazingly this works and gives an MCMC algorithm targeting $\pi$.

- Such algorithms are collectively known as **pseudo-marginal** algorithms.
Pseudomarginal

- Why do they work?
- Useful to think of estimator as

\[ \hat{\pi}(x) = \pi(x) Z_x, \]

where the random variable \( Z_x \)

- ♠️ is non-negative;
- ♣️ has density \( g_x(\cdot) \);
- ♢️ has unit expectation \( \int z g_x(z) dz = 1 \).
Pseudomarginal

- Define a new target with density

\[
\tilde{\pi}(x, z) = \pi(x) z g_x(z), \quad \int \tilde{\pi}(x, z) \, dz = \pi(x);
\]

- and apply Metropolis-Hastings with proposal

\[
\tilde{q}(\langle x, z \rangle, \langle y, w \rangle) := q(x, y) g_y(w).
\]

- The acceptance probability will be

\[
\min \left\{ 1, \frac{\tilde{\pi}(y, w) q(\langle y, w \rangle, \langle x, z \rangle)}{\tilde{\pi}(x, z) q(\langle x, z \rangle, \langle y, w \rangle)} \right\}
\]

\[
= \min \left\{ 1, \frac{\pi(y) w g_y(w) q(y, x) g_z(x)}{\pi(x) z g_x(z) q(x, y) g_y(w)} \right\}
\]

\[
= \min \left\{ 1, \frac{\pi(y) w g_y(w) q(y, x) g_z(x)}{\pi(x) z g_x(z) q(x, y) g_y(w)} \right\}
= \min \left\{ 1, \frac{\pi(y) q(y, x) w}{\pi(x) q(x, y) z} \right\}
\]

where \( z \sim g_x(\cdot) \) and \( w \sim g_w(\cdot) \).
Pseudomarginal

The algorithm is as follows:

- given \((X^{(t-1)}, Z^{(t-1)})\), for an unbiased estimator 
  \(Z^{(t-1)} = \hat{\pi}(X^{(t-1)}) \sim g_{X^{(t-1)}}(\cdot)\);

1. propose \(Y \sim q(x, \cdot)\);
2. sample \(\hat{\pi}(Y) \sim g_Y(\cdot)\);
3. With probability

\[
\min \left\{ 1, \frac{\hat{\pi}(Y) q\left(y, X^{(t-1)}\right)}{\hat{\pi}(X^{(t-1)}) q(X^{(t-1)}, Y)} \right\}
\]

set

\(X^{(t)} = Y, Z^{(t)} = \hat{\pi}(Y),\)

4. Otherwise set \( (X^{(t)}, Z^{(t)}) = (X^{(t-1)}, Z^{(t-1)}). \)
Tuning the pseudo-marginal

- The variance of the estimator

\[ \hat{\pi}(x) = \frac{1}{N} \sum_{i=1}^{N} p(x, U_i), \quad U_i \text{i.i.d. } f_{U|x}(\cdot|x), \]

will affect the performance and the cost.
- Small \( N \) is cheap to compute but gives large variance and poor mixing;
- Large \( N \) is expensive but improves mixing;
- Optimal? Variance needs to be \( \in [1, 2] \);
- In practice estimate mode \( \hat{x} \) and estimate \( \pi(\hat{x}) \) for various \( N \); choose \( N \) that gives variance close to 1-2.
Convergence diagnostics

• Goal: assess whether MCMC chains have converged.

• In general, impossible to know for sure that there is no problem.

• But we can sometimes know for sure that there is a problem.
Visual diagnostics: traceplot

Target: $\pi = \mathcal{N}(-2, 0.2^2)$, proposal $q(y \mid x) = \mathcal{N}(y; x, 0.5^2)$. 
Visual diagnostics: autocorrelogram

Target: $\pi = \mathcal{N}(-2, 0.2^2)$, proposal $q(y \mid x) = \mathcal{N}(y; x, 0.5^2)$. 
Visual diagnostics: convergence of estimators

Target: $\pi = \mathcal{N}(-2,0.2^2)$, proposal $q(y \mid x) = \mathcal{N}(y; x, 0.5^2)$.

Could be also computed on different non-overlapping subsequences, leading to Geweke’s diagnostics.
Visual diagnostics: traceplot

Target: \( \pi = \frac{1}{2} \mathcal{N}(-2, 0.2^2) + \frac{1}{2} \mathcal{N}(+2, 0.2^2) \), same proposal.
Visual diagnostics: autocorrelogram

Target: $\pi = \frac{1}{2} \mathcal{N}(-2,0.2^2) + \frac{1}{2} \mathcal{N}(+2,0.2^2)$, same proposal.
Visual diagnostics: convergence of estimators

Target: \( \pi = \frac{1}{2} \mathcal{N}(-2, 0.2^2) + \frac{1}{2} \mathcal{N}(+2, 0.2^2) \), same proposal.
Multiple starting points

We start $M$ chains from various starting points.

After enough iterations the starting point should not matter and hence we should obtain the same results based on each chain.

We have the classical “sum of squares” decomposition in “intra group” and “inter group” terms:

$$\sum_{m=1}^{M} \sum_{t=1}^{T} (X_{m,t} - \bar{X}_{..})^2 = \sum_{m=1}^{M} \sum_{t=1}^{T} (\bar{X}_{m..} - \bar{X}_{..})^2 \quad \text{inter-group}$$

$$+ \sum_{m=1}^{M} \sum_{t=1}^{T} (X_{m,t} - \bar{X}_{m..})^2 \quad \text{intra-group}$$
Multiple starting points

This leads to considering

\[ W = \frac{1}{M} \sum_{m=1}^{M} \frac{1}{T-1} \sum_{t=1}^{T} (X_{m,t} - \bar{X}_{m,.})^2 \]

\[ B = \frac{1}{M-1} \sum_{m=1}^{M} (\bar{X}_{m,.} - \bar{X}_{.,.})^2 \]

\[ V = \left(1 - \frac{1}{T}\right)W + B \]

In principle \( W \) and \( V \) should both converge to the true variance of the target distribution.

\( V \) would be unbiased if starting points were drawn from the target, whereas \( W \) under-estimates the variance.

We can thus plot \( \sqrt{V/W} \) and compare to 1. This is the idea behind Gelman-Rubin diagnostics.
Visual diagnostics: Gelman-Rubin diagnostics

Target: $\pi = \mathcal{N}(-2, 0.2^2)$, $M = 4$ chains.
Visual diagnostics: Gelman-Rubin diagnostics

Target: \( \pi = \frac{1}{2} \mathcal{N}(-2, 0.2^2) + \frac{1}{2} \mathcal{N}(+2, 0.2^2) \), \( M = 4 \) chains.
Visual diagnostics: traceplot with $M$ chains

Target: $\pi = \frac{1}{2} \mathcal{N}(-2,0.2^2) + \frac{1}{2} \mathcal{N}(+2,0.2^2)$, $M = 4$ chains.
Parallelization

In the past (and in the next?) years, many more parallel cores, but not much more clockspeed.

Among the methods seen so far, which are parallelizable?

MCMC methods are by definition iterative methods. Sometimes the likelihood evaluation itself can be parallelized.

We can run independent MCMC in parallel, as in the Gelman-Rubin diagnostics.

Should we make the chains interact?
Parallelization of the likelihood evaluation

Consider the evaluation of the likelihood in the normal mixture case: the observations $Y_1, \ldots, Y_n$ come from

$$\forall i \in \{1, \ldots, n\} \quad Y_i \sim \sum_{k=1}^{K} p_k \mathcal{N}(\mu_k, \sigma_k^2).$$

The likelihood can be written

$$L(\theta; y_1, \ldots, y_n) = \prod_{i=1}^{n} \left( \sum_{k=1}^{K} p_k \varphi(y_i; \mu_k, \sigma_k^2) \right)$$

which can be done by evaluating the $n$ terms in the product in parallel and then taking the product.

Or $n \times K$ terms in parallel, and then partial sums and a product.
Parallelization of the likelihood evaluation

For i.i.d. data the likelihood evaluation can be parallelized.

In cases where

the likelihood is not so expensive,

or the likelihood evaluation cannot be efficiently parallelized.

then a single-chain Metropolis-Hastings algorithm cannot benefit from multiple processors.

However we can run multiple chains!
Parallel Tempering

The idea of parallel tempering is to run $N$ chains targeting different versions of $\pi$, of “increasing difficulty”.

Introduce “inverse temperatures”:

$$0 < \gamma_1 < \gamma_2 < \ldots < \gamma_N = 1.$$

Introduce “tempered” distributions $\pi^{\gamma_n}$ for $n = 1, \ldots, N$ and $N$ chains one for each $\pi^{\gamma_k}$.

For $\gamma \approx 0$, $\pi^\gamma$ is considered easier to sample because the variations of $\pi$ are smaller.
Parallel Tempering

The "joint chain" is targeting

\[ \pi^\gamma_1 \otimes \pi^\gamma_2 \otimes \ldots \otimes \pi^\gamma_N. \]

We occasionally perform a swap move:
Sample indices \( k_1, k_2 \) uniformly in \( \{1, \ldots, N\} \).
With acceptance probability

\[
\min \left( 1, \frac{\pi^\gamma_{k_1}(x_{k_2}) \pi^\gamma_{k_2}(x_{k_1})}{\pi^\gamma_{k_1}(x_{k_1}) \pi^\gamma_{k_2}(x_{k_2})} \right).
\]

exchange the value of \( x_{k_1} \) and \( x_{k_2} \).

**FACT:** The swap moves preserve detailed balance.
This doesn’t change the joint target distribution

\[ \pi^\gamma_1 \otimes \pi^\gamma_2 \otimes \ldots \otimes \pi^\gamma_N. \]

In particular the \( N \)-th chain still targets \( \pi^\gamma_N = \pi \).
Parallel Tempering

Figure: Target density function.
Parallel Tempering

Figure: Target density function.
Figure: Target density function.
Parallel Tempering

Figure: Target density function.
Parallel Tempering

Let’s use $N = 10$ chains and $\gamma_1 = 0.1, \gamma_2 = 0.2, \ldots, \gamma_{10} = 1$. No swapping.

Figure: Trace plot of the “low temperature chains”.
Parallel Tempering

Let’s use $N = 10$ chains and $\gamma_1 = 0.1, \gamma_2 = 0.2, \ldots, \gamma_{10} = 1$.

Figure: Trace plot of the “high temperature chains”.
Parallel Tempering

If we want to find the modes of $\pi$, we might just use the high temperature chains and forget about sampling directly from $\pi$.

If we want to sample from $\pi$, can we use the “high temperature” chains to improve the mixing of the chain targeting $\pi$?

Parallel tempering works by proposing moves where chains of different temperatures are swapped.
Parallel Tempering

Figure: Trace plot of the “low temperature chains” using swap moves.
Parallel Tempering

Figure: Histogram of the chain targeting $\pi^{\gamma_1}$. 
Parallel Tempering

Figure: Histogram of the chain targeting $\pi^{\gamma_4}$. 
Figure: Histogram of the chain targeting $\pi^{\gamma_7}$.
Parallel Tempering

Figure: Histogram of the chain targeting $\pi^{\gamma_{10}}$.

Swap moves improve the mixing of chains with high values of $\gamma$. 