Advanced Simulation - Lecture 7

George Deligiannidis

February 8th, 2016

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 $t = 2, 3, \ldots$

1. Sample $X^\star \sim q\left(\cdot \,|\, X^{(t-1)} \right)$.
2. Compute

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

3. Sample $U \sim \mathcal{U}_{[0,1]}$. If $U \leq \alpha\left(X^\star \,|\, X^{(t-1)} \right)$, set $X^{(t)} = X^\star$, otherwise set $X^{(t)} = X^{(t-1)}$.

Some results

Proposition

If $q(x^\star | x) > 0$ for any $x, x^\star \in supp(\pi)$ 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 $\varphi : \mathbb{X} \to \mathbb{R}$:

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

for every starting value $X^{(1)}$.

Random Walk Metropolis–Hastings

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

$$X^\star = X + \varepsilon, \quad \varepsilon \sim g;$$

  e.g. g is a zero-mean multivariate normal or t-student.
- Acceptance probability becomes

$$\alpha(x^\star \mid x) = \min\left(1, \frac{\pi(x^\star)}{\pi(x)}\right).$$

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

$$\pi(x^\star)/\pi(x) < 1.$$

Independent Metropolis–Hastings

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

$$\alpha(x^\star \mid x) = \min\left(1, \frac{\pi(x^\star)q(x)}{\pi(x)q(x^\star)}\right).$$

  - 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 $\rho\left(X^{(t-1)}, X^{(t)}\right)$ between subsequent values (why?).
- Two sources of correlation:
  - between the current state $X^{(t-1)}$ and proposed value $X \sim q\left(\cdot \mid X^{(t-1)}\right)$,
  - 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

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

- We use a random walk Metropolis—Hastings algorithm with

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

- What is the optimal choice of σ^2?
- We consider three choices: $\sigma^2 = 0.1^2, 1, 10^2$.

Metropolis–Hastings algorithm

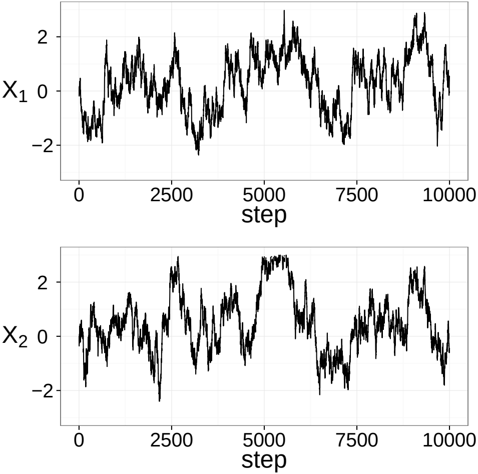

Figure: Metropolis–Hastings on a bivariate Gaussian target. With $\sigma^2 = 0.1^2$, the acceptance rate is $\approx 94\%$.

Metropolis–Hastings algorithm

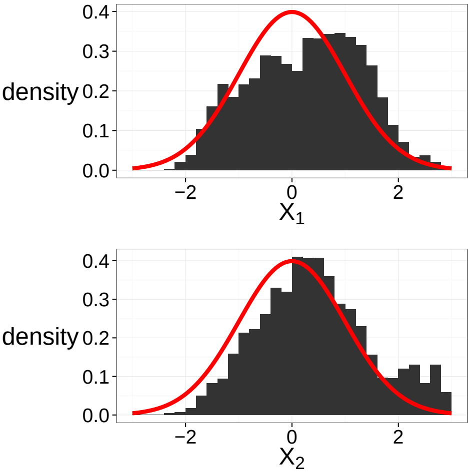

Figure: Metropolis–Hastings on a bivariate Gaussian target. With $\sigma^2 = 0.1^2$, the acceptance rate is $\approx 94\%$.

Metropolis–Hastings algorithm

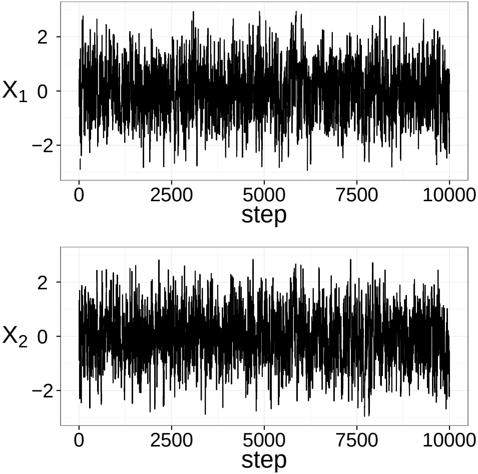

Figure: Metropolis–Hastings on a bivariate Gaussian target. With $\sigma^2 = 1$, the acceptance rate is $\approx 52\%$.

Metropolis–Hastings algorithm

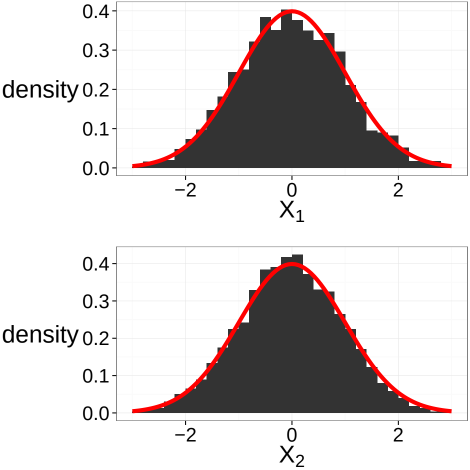

Figure: Metropolis–Hastings on a bivariate Gaussian target. With $\sigma^2 = 1$, the acceptance rate is $\approx 52\%$.

Metropolis–Hastings algorithm

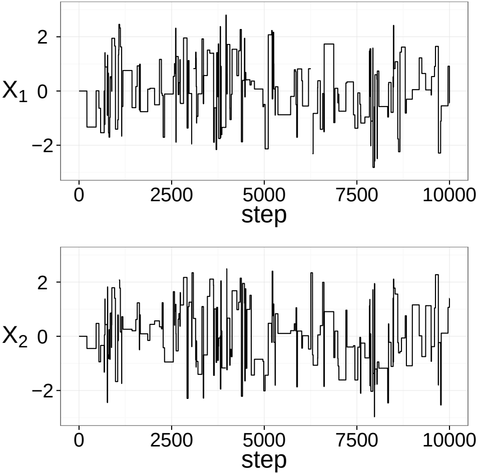

Figure: Metropolis–Hastings on a bivariate Gaussian target. With $\sigma^2 = 10$, the acceptance rate is $\approx 1.5\%$.

Metropolis–Hastings algorithm

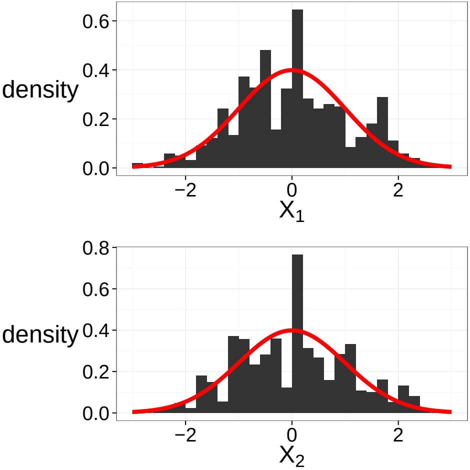

Figure: Metropolis–Hastings on a bivariate Gaussian target. With $\sigma^2 = 10$, the acceptance rate is $\approx 1.5\%$.

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 mimick 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.

Sophisticated Proposals

- "Langevin" proposal relies on

$$X^{\star} = X^{(t-1)} + \frac{\sigma}{2} \nabla \log \pi|_{X^{(t-1)}} + \sigma W$$

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

$$\frac{\pi(X^{\star})q(X^{(t-1)} \mid X^{\star})}{\pi(X^{(t-1)})q(X^{\star} \mid X^{(t-1)})} = \frac{\pi(X^{\star})}{\pi(X^{(t-1)})} \frac{\mathcal{N}(X^{(t-1)}; X^{\star} + \frac{\sigma}{2}.\nabla \log \pi|_{X^{\star}}; \sigma^2)}{\mathcal{N}(X^{\star}; X^{(t-1)} + \frac{\sigma}{2}.\nabla \log \pi|_{X^{(t-1)}}; \sigma^2)}.$$

- Possibility to use higher order derivatives:

$$X^{\star} = X^{(t-1)} + \frac{\sigma}{2} \left[\nabla^2 \log \pi|_{X^{(t-1)}}\right]^{-1} \nabla \log \pi|_{X^{(t-1)}} + \sigma W.$$

Sophisticated Proposals

- We can use

$$q(X^{\star}|X^{(t-1)}) = g(X^{\star};\varphi(X^{(t-1)}))$$

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

$$\frac{\pi(X^{\star})q(X^{(t-1)}|X^{\star})}{\pi(X^{(t-1)})q(X^{\star}|X^{(t-1)})} = \frac{\pi(X^{\star})g(X^{(t-1)};\varphi(X^{\star}))}{\pi(X^{(t-1)})g(X^{\star};\varphi(X^{(t-1)}))}.$$

- 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 π 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 $X^\star = \exp(\log X^{(t-1)} + \varepsilon)$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. What is the acceptance probability then?

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

Why?

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

Random Proposals

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

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

so that

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

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

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

cannot be computed.

Random Proposals

- Instead you decide to accept your proposal with probability

$$\alpha_t = \min\left\{1, \frac{\pi\left(X^\star\right) q_{\sigma^{2,(t-1)}}\left(X^{(t-1)}\middle| X^\star\right)}{\pi\left(X^{(t-1)}\right) q_{\sigma^{2,\star}}\left(X^\star \middle| X^{(t-1)}\right)}\right\}$$

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

- With probability α_t, set $\sigma^{2,(t)} = \sigma^{2,\star}$, $X^{(t)} = X^\star$, 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

$$\widetilde{\pi}\left(x,\sigma^2\right) := \pi\left(x\right) f\left(\sigma^2\right).$$

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

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

- Indeed, we have

$$\frac{\widetilde{\pi}(y,\tau^2)}{\widetilde{\pi}(x,\sigma^2)} \frac{q(x,\sigma^2|y,\tau^2)}{q(y,\tau^2|x,\sigma^2)} = \frac{\pi(y) f(\tau^2)}{\pi(x) f(\sigma^2)} \frac{f(\sigma^2) q_{\sigma^2}(x|y)}{f(\tau^2) q_{\tau^2}(y|x)} = \frac{\pi(y)}{\pi(x)} \frac{q_{\sigma^2}(x|y)}{q_{\tau^2}(y|x)}$$

- **Remark**: we just need to be able to sample from $f(\cdot)$, not to evaluate it.

Using multiple proposals

- Consider a target of density $\pi(x)$ where $x \in \mathbb{X}$.
- To sample from π, you might want to use various proposals for Metropolis-Hastings $q_1(x'|x), q_2(x'|x), ..., q_p(x'|x)$.
- One way to achieve this is to build a proposal

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

and Metropolis-Hastings requires evaluating

$$\alpha\left(X^\star \middle| X^{(t-1)}\right) = \min\left(1, \frac{\pi(X^\star)\, q\left(X^{(t-1)} \middle| X^\star\right)}{\pi\left(X^{(t-1)}\right) q\left(X^\star \middle| X^{(t-1)}\right)}\right),$$

and thus evaluating $q_j\left(X^\star \middle| X^{(t-1)}\right)$ for $j = 1, ..., p$.

Motivating Example

- Let

$$q\left(x' \middle| x\right) = \beta_1 \mathcal{N}\left(x'; x, \Sigma\right) + \left(1 - \beta_1\right) \mathcal{N}\left(x'; \mu\left(x\right), \Sigma\right)$$

where $\mu : \mathbb{X} \to \mathbb{X}$ is a clever but computationally expensive deterministic optimisation algorithm.

- Using $\beta_1 \approx 1$ will make most proposed points come from the cheaper proposal distribution $\mathcal{N}\left(x'; x, \Sigma\right)$…

- …but you won't save time as $\mu\left(X^{(t-1)}\right)$ needs to be evaluated at every step.

Composing kernels

- How to use different proposals to sample from π without evaluating all the densities at each step?

- What about combining different Metropolis-Hastings updates K_j using proposal q_j instead? i.e.

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

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 π-invariant.

Two possibilities of combining 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 $t = 2, 3, ...$

1. Set $Z^{(t,0)} := X^{(t-1)}$.
2. For $j = 1, ..., p$, sample $Z^{(t,j)} \sim K_j \left(Z^{(t,j-1)}, \cdot \right)$.
3. Set $X^{(t)} := Z^{(t,p)}$.

- Full cycle transition kernel is

$$K\left(x^{(t-1)}, x^{(t)}\right) = \int \cdots \int K_1\left(x^{(t-1)}, z^{(t,1)}\right) K_2\left(z^{(t,1)}, z^{(t,2)}\right) \cdots K_p\left(z^{(t,p-1)}, x^{(t)}\right) dz^{(t,1)} \cdots dz^{(t,p-1)}.$$

- K is π-invariant.

Mixture of MCMC updates

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

1. Sample J from $\{1, ..., p\}$ with $\mathbb{P}(J = k) = \beta_k$.
2. Sample $X^{(t)} \sim K_J\left(X^{(t-1)}, \cdot\right)$.

- Corresponding transition kernel is

$$K\left(x^{(t-1)}, x^{(t)}\right) = \sum_{j=1}^{p} \beta_j K_j\left(x^{(t-1)}, x^{(t)}\right).$$

- K is π-invariant.
- The algorithm is *different* from using a mixture proposal

$$q\left(x' \middle| x\right) = \sum_{j=1}^{p} \beta_j q_j\left(x' \middle| x\right).$$

Metropolis-Hastings Design for Multivariate Targets

- If $\dim(\mathbb{X})$ 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 $x = (x_1, ..., x_d)$ and denote $x_{-j} := x \setminus \{x_j\}$.
- We propose "local" proposals where only x_j is updated

$$q_j(x'|x) = \underbrace{q_j\left(x'_j \middle| x\right)}_{\text{propose new component } j} \underbrace{\delta_{x_{-j}}\left(x'_{-j}\right)}_{\text{keep other components fixed}} .$$

Metropolis-Hastings Design for Multivariate Targets

- This yields

$$\alpha_j(x, x') = \min\left(1, \frac{\pi(x'_{-j}, x'_j) q_j(x_j | x_{-j}, x'_j)}{\pi(x_{-j}, x_j) q_j(x'_j | x_{-j}, x_j)} \underbrace{\frac{\delta_{x'_{-j}}(x_{-j})}{\delta_{x_{-j}}(x'_{-j})}}_{=1}\right)$$

$$= \min\left(1, \frac{\pi(x_{-j}, x'_j) q_j(x_j | x_{-j}, x'_j)}{\pi(x_{-j}, x_j) q_j(x'_j | x_{-j}, x_j)}\right)$$

$$= \min\left(1, \frac{\pi_{X_j|X_{-j}}(x'_j | x_{-j}) q_j(x_j | x_{-j}, x'_j)}{\pi_{X_j|X_{-j}}(x_j | x_{-j}) q_j(x'_j | x_{-j}, x_j)}\right).$$

One-at-a-time MH (cycle/systematic scan)

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

- Sample $X^\star \sim q_j(\cdot | X_1^{(t)}, \ldots, X_{j-1}^{(t)}, X_j^{(t-1)}, ..., X_d^{(t-1)})$.
- Compute

$$\alpha_j = \min\left(1, \frac{\pi_{X_j|X_{-j}}\left(X_j^\star \mid X_1^{(t)} \ldots X_{j-1}^{(t)}, X_{j+1}^{(t-1)} \ldots X_d^{(t-1)}\right)}{\pi_{X_j|X_{-j}}\left(X_j^{(t-1)} \mid X_1^{(t)} \ldots X_{j-1}^{(t)}, X_{j+1}^{(t-1)} \ldots X_d^{(t-1)}\right)} \times \frac{q_j\left(X_j^{(t-1)} \middle| X_1^{(t)} ... X_{j-1}^{(t)}, X_j^\star, X_{j+1}^{(t-1)} ... X_d^{(t-1)}\right)}{q_j\left(X_j^\star \middle| X_1^{(t)} ... X_{j-1}^{(t)}, X_j^{(t-1),}, X_{j+1}^{(t-1)} ... X_d^{(t-1)}\right)}\right).$$

- With probability α_j, set $X^{(t)} = X^\star$, 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 $\{1, ..., d\}$ with $\mathbb{P}(J = k) = \beta_k$.
- Sample $X^\star \sim q_J\left(\cdot \mid X_1^{(t)}, ..., X_d^{(t-1)}\right)$.
- Compute

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

- With probability α_J set $X^{(t)} = X^\star$, 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

$$q_j\left(x_j'\middle|\, x\right) = \pi_{X_j|X_{-j}}\left(x_j'\middle|\, x_{-j}\right).$$

Proof.

It follows from

$$\frac{\pi\left(x_{-j}, x_j'\right)}{\pi\left(x_{-j}, x_j\right)} \frac{q_j\left(x_j\middle|\, x_{-j}, x_j'\right)}{q_j\left(x_j'\middle|\, x_{-j}, x_j\right)}$$

$$= \frac{\pi\left(x_{-j}\right) \pi_{X_j|X_{-j}}\left(x_j'\middle|\, x_{-j}\right)}{\pi\left(x_{-j}\right) \pi_{X_j|X_{-j}}\left(x_j\middle|\, x_{-j}\right)} \frac{\pi_{X_j|X_{-j}}\left(x_j\middle|\, x_{-j}\right)}{\pi_{X_j|X_{-j}}\left(x_j'\middle|\, x_{-j}\right)} = 1.$$

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 $X_1^\star \sim \pi(X_1 \mid X_2^{(t-1)})$, then $X_2^\star \sim \pi(X_2 \mid X_1^\star)$. The proposal is then $X^\star = (X_1^\star, X_2^\star)$.
- Compute

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

- Accept $X^\star$ or not based on α_t, where here

$$\alpha_t \neq 1$$

!!