Mathematics and numerics for data assimilation and state estimation – Lecture 12

Summer semester 2020

Overview

1. Sampling of the posterior probability density function
   - Monte Carlo method
2. Importance sampling applied to posterior densities
3. Discrete-time continuous-space Markov chains
4. Markov chain Monte Carlo sampling/dynamics

Summary of lecture 11

- Bayesian inversion and optimization

$$Y = G(U) + \eta, \qquad \eta \sim N(0, \Gamma)$$

  MAP estimator corresponds to a form of Tikhonov regularization when prior is Gaussian, and to LASSO regression when it is component-wise iid-Laplace distributed.

- Kullback-Leibler divergence and information gain and fitting of Gaussian to posteriors.

Overview

1. Sampling of the posterior probability density function
   - Monte Carlo method
2. Importance sampling applied to posterior densities
3. Discrete-time continuous-space Markov chains
4. Markov chain Monte Carlo sampling/dynamics

We recall that by Bayesian inversion

$$Y = G(U) + \eta,$$

and observation $Y = y$ leads to the posterior density

$$\pi(u|y) = \frac{g(u)\pi_U(u)}{Z}, \quad \text{where} \quad g(u) := \pi_\eta(y - G(u)).$$

Problem: Given a quantity of interest (QoI) $f : \mathbb{R}^d \to \mathbb{R}$, we seek to estimate the mean

$$\int f(u)\pi(u|y)du$$

Plan: Present different Monte Carlo methods for approximating said mean.

Plain Monte Carlo method

Given a pdf $\pi \in \mathcal{M}$ on $\mathbb{R}^d$, we seek to approximate

$$\pi[f] := \mathbb{E}^{\pi}[f].$$

We introduce the empirical (random) probability measure

$$\pi_{MC}^{M} := \frac{1}{M}\sum_{k=1}^{M} \delta_{U_k}, \quad \text{where } U_k \sim \pi \text{ are iid} \tag{1}$$

and the Monte Carlo estimator

$$\pi_{MC}^{M}[f] = \frac{1}{M}\sum_{k=1}^{M} f(U_k)$$

Comment: regardless of whether $\hat{\pi}$ is a pdf or a probability measure we denote by $\hat{\pi}[f]$ the expectation of $f(X)$ where $X \sim \hat{\pi}$.

Theorem 1 (Convergence results SST 5.1)

For any $f : \mathbb{R}^d \to \mathbb{R}$ *such that* $\pi[|f|] < \infty$,

$$\mathbb{E}\left[\pi^M_{MC}[f]\right] = \pi[f] \quad \textit{and} \quad \mathbb{E}\left[\left(\pi^M_{MC}[f] - \pi[f]\right)^2\right] = \frac{Var[f]}{M},$$

where $Var[f] = \pi[f^2] - (\pi[f])^2$.

Proof ideas: The assumption $\pi[|f|] < \infty$ is a sufficient condition for $\pi[f]$ being well-defined, and

$$\mathbb{E}\left[\pi^M_{MC}[f]\right] = \qquad\qquad\qquad\qquad = \pi[f].$$

And using that $\pi[f] = \mathbb{E}[f(U)]$ and $\{f(U_k) - \mathbb{E}[f(U)]\}_k$ are iid,

$$\mathbb{E}\left[\left(\pi^M_{MC}[f] - \pi[f]\right)^2\right] =$$

Remark:

Whenever $\|f\|_{L^\infty(\mathbb{R}^d)} \leq 1$, then

$$\text{Var}[f] = \pi[f^2] - (\pi[f])^2 \leq 1$$

which implies that

$$\sup_{\|f\|_{L^\infty(\mathbb{R}^d)} \leq 1} \mathbb{E}\left[\left(\pi_{MC}^M[f] - \pi[f] \right)^2 \right] \leq \frac{1}{M},$$

Example 2 (Volume ratio unit ball / smallest containing cube)

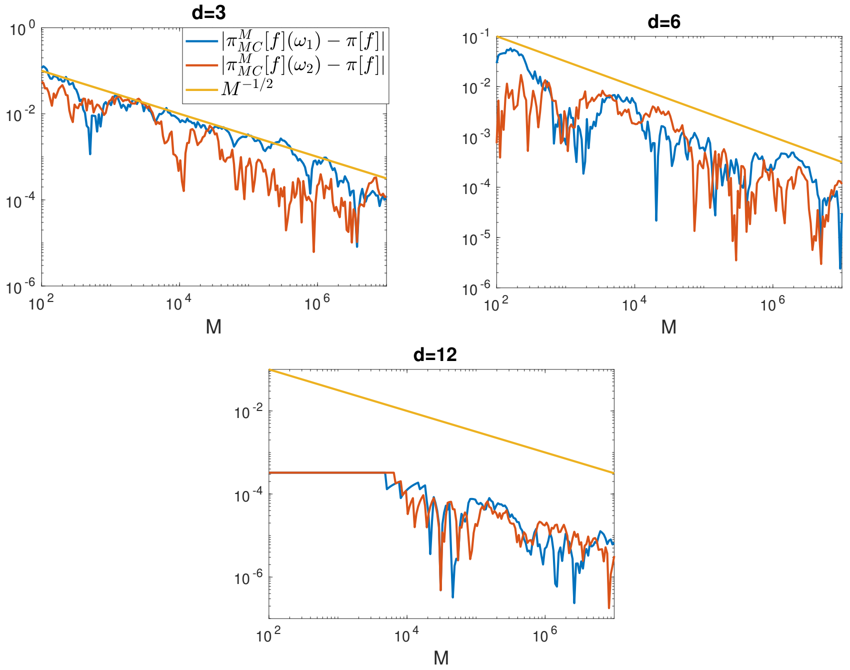
Let $B_d = \{x \in \mathbb{R}^d \mid |x| \leq 1\}$ and $\pi = PDF(U(-1,1)^d)$. Then for $f(x) = \mathbb{1}_{B_d}(x)$

$$\begin{aligned}\pi[f] &= 2^{-d} \int_{[-1,1]^d} \mathbb{1}_{B_d}(x) dx \\ &= 2^{-d} \mathrm{Leb}(B_d) = \frac{\mathrm{Leb}(B_d)}{\mathrm{Leb}([-1,1]^d)} \\ &= \frac{\pi^{d/2}}{2^d \, \Gamma(d/2+1)}\end{aligned}$$

Since $\|f\|_\infty \leq 1$,

$$\sqrt{\mathbb{E}\left[\left(\pi^M_{MC}[f] - \pi[f]\right)^2\right]} \leq \frac{1}{\sqrt{M}}.$$

Let us numerically confirm that root-mean-square convergence rate is independent of d.


d=3
d=6
d=12
$|\pi_{MC}^{M}[f](\omega_1) - \pi[f]|$
$|\pi_{MC}^{M}[f](\omega_2) - \pi[f]|$
$M^{-1/2}$
M

Efficiency of Monte Carlo

Given an accuracy constraint

$$\mathbb{E}\left[(\pi_{MC}^{M}[f] - \pi[f])^2 \right] \leq \epsilon^2$$

it is sufficient to use $M = \lceil \mathrm{Var}[f]/\epsilon^2 \rceil$.

Verification: By theorem

$$\mathbb{E}\left[(\pi_{MC}^{M}[f] - \pi[f])^2 \right] = \mathrm{Var}[\pi_{MC}^{M}[f]] = \frac{\mathrm{Var}[f]}{M}$$

Note: It is often possible to reduce the magnitude of $\mathrm{Var}[f]$ and improve the efficiency of Monte Carlo methods.

Importance sampling

$$\pi[f] = \int_{\mathbb{R}^d} f(x)\pi(x)\,dx$$

can alternatively be computed by the expectation wrt to another pdf $\hat{\pi}$ provided $f\pi$ is **dominated** by $\hat{\pi}$, meaning that

$$\hat{\pi}(x) = 0 \implies f(x)\pi(x) = 0.$$

Then

$$\pi[f] = \int_{\mathbb{R}^d} f(x)\pi(x)\,dx = \int_{\mathbb{R}^d} f(x)\underbrace{\frac{\pi(x)}{\hat{\pi}(x)}}_{W(x)}\hat{\pi}(x)\,dx = \hat{\pi}[Wf]$$

IS algorithm:

1. Select a $\hat{\pi}$ that domintes $f\pi$.
2. Generate $U_1, \ldots, U_M \overset{iid}{\sim} \hat{\pi}$ and compute

$$\hat{\pi}^M_{MC}[Wf] = \frac{1}{M}\sum_{i=1}^{M} W(U_i)f(U_i) = \frac{1}{M}\sum_{i=1}^{M} \frac{\pi(U_i)}{\hat{\pi}(U_i)} f(U_i)$$

The convergence rate of IS

$$\mathbb{E}\left[(\hat{\pi}^M_{MC}[Wf] - \pi[f])^2\right] = \mathrm{Var}[\hat{\pi}^M_{MC}[Wf]] = \frac{\mathrm{Var}_{\hat{\pi}}[Wf]}{M}$$

So performance of IS compared to plain Monte Carlo relates to ratio

$$\frac{\mathrm{Var}[\hat{\pi}^M_{MC}[Wf]]}{\mathrm{Var}[\pi^M_{MC}[f]]} = \frac{\mathrm{Var}_{\hat{\pi}}[Wf]}{\mathrm{Var}_{\pi}[f]}.$$

Optimization: Find $\hat{\pi}$ dominating $f\pi$ that minimizes

$$\mathrm{Var}_{\hat{\pi}}\left[\frac{\pi}{\hat{\pi}}f\right].$$

In real optimization problem, i.e., for efficiency of method rather than convergence rate, the cost of sampling from $\hat{\pi}$ should also be included.

Convergence of random variables – in the probability space

A different viewpoint, leading to analogous results as the above, is to extend the sampling theory in Lecture 4 to mixed and continuous rv. Drawing iid $X_k \sim \mathbb{P}_X$ that may continuous, mixed or discrete, the sample average

$$\bar{X}_M := \frac{1}{M} \sum_{k=1}^{M} X_k \tag{2}$$

satisfies the following:

- it is unbiased $\mathbb{E}\left[\bar{X}_M\right] = \mathbb{E}[X]$,
- if $X \in L^2(\Omega)$, then, as we know, it converges in root-mean-square sense with rate $1/2$

$$\|\bar{X}_M - \mu\|_{L^2(\Omega)} = \frac{\|X - \mathbb{E}[X]\|_{L^2(\Omega)}}{\sqrt{M}}, \tag{3}$$

- if $\mathbb{E}[|X_k|] < \infty$, then the weak law of large numbers applies: for any $\epsilon > 0$,

$$\mathbb{P}(|\bar{X}_M - \mathbb{E}[X]| > \epsilon) \to 0 \quad \text{as} \quad M \to \infty.$$

Overview

Importance sampling of the posterior

Objective: Given a posterior

$$\pi(u|y) = \frac{g(u)\pi_U(u)}{Z} \quad \text{we seek to estimate} \quad \mathbb{E}^{\pi(\cdot|y)}[f].$$

Problem setting: We do not know Z, and we cannot sample from the posterior directly. But we can sample from the prior π_U and we can evaluate $g(u)$.

Approach:

$$\mathbb{E}^{\pi(\cdot|y)}[f] = \int_{\mathbb{R}^d} f(u)\pi(u|y)\,du = \frac{1}{Z}\int_{\mathbb{R}^d} f(u)g(u)\pi_U(u)du = \frac{\pi_U[fg]}{\pi_U[g]}.$$

Using the shorthand $\pi := \pi_U$, we introduce the sampling estimator

$$\frac{\pi_{MC}^M[fg]}{\pi_{MC}^M[g]}.$$

The estimator

Simulate $U_1, \ldots, U_M \overset{iid}{\sim} \pi$ and compute

$$\frac{\pi_{MC}^M[fg]}{\pi_{MC}^M[g]} = \frac{M^{-1}\sum_{i=1}^M f(U_i)g(U_i)}{M^{-1}\sum_{j=1}^M g(U_j)} = \sum_{i=1}^M \frac{g(U_i)}{\sum_{j=1}^M g(U_j)} f(U_i) = \sum_{i=1}^M W_i f(U_i)$$

with

$$W_i := \frac{g(U_i)}{\sum_{j=1}^M g(U_j)}.$$

Introducing the weighted, random empirical measure

$$\pi_{IS}^M := \sum_{i=1}^M W_i \delta_{U_i} \quad \text{we define} \quad \pi_{IS}^M[f] := \sum_{i=1}^M W_i f(U_i).$$

NB! error analysis of $\pi_{IS}^M[f] \to \mathbb{E}^{\pi(\cdot|y)}[f]$ is more complicated than before, since this estimator may be biased, meaning

$$\mathbb{E}\left[\pi_{IS}^M[f]\right] \neq \mathbb{E}^{\pi(\cdot|y)}[f]$$

Convergence rates

For pdfs $\pi, \hat{\pi} \in \mathcal{M}_+$, we define the χ^2-divergence from π to $\hat{\pi}$ as

$$d_{\chi^2}(\pi||\hat{\pi}) = \int_{\mathbb{R}^d} \left(\frac{\pi(u)}{\hat{\pi}(u)} - 1\right)^2 \hat{\pi}(u)\, du$$

Theorem 3 (SST 5.4)

For any $f : \mathbb{R}^d \to \mathbb{R}$ with $\|f\|_{L^\infty(\mathbb{R}^d)} \leq 1$, it holds that (where we expand our shorthand $\pi = \pi_U$ on the RHS for clarity)

$$\left|\mathbb{E}\left[\pi_{IS}^M[f] - \mathbb{E}^{\pi(\cdot|y)}[f]\right]\right| \leq 2\frac{1 + d_{\chi^2}(\pi(\cdot|y)||\pi_U)}{M}$$

and

$$\mathbb{E}\left[\left(\pi_{IS}^M[f] - \mathbb{E}^{\pi(\cdot|y)}[f]\right)^2\right] \leq 4\frac{1 + d_{\chi^2}(\pi(\cdot|y)||\pi_U)}{M}$$

Bias convergence rate: 1, root-mean-square convergence rate: 1/2.

Overview

From discrete-space to continuous space Markov chains

Essential components of discrete-space Markov chains on $\mathbb{S}$

- Initial distribution $\pi^0 : \mathbb{S} \to [0, 1]$.
- Transition function $p : \mathbb{S} \times \mathbb{S} \to [0, 1]$:

$$p(x, y) = \mathbb{P}(X_{n+1} = y | X_n = x) \quad \text{whenever} \mathbb{P}(X_n = x) > 0,$$

  (also time-inhomogeneous transition functions p(x,y,n))
- Dynamics for path:

$$X_{n+1} \sim p(X_n, \cdot)$$

- Dynamics for distribution

$$\pi^{n+1} = \pi^n p.$$

Continuous-space Markov chains

$X_0, X_1, \ldots$ is a time-discrete Markov chain on state-space $\mathbb{R}^d$ provided it

- has initial distribution $X_0 \sim \mathbb{P}^0$
- has transition kernel $K : \mathbb{R}^d \times \mathcal{B}^d \to [0, \infty)$ satisfying that
  1. for every $x \in \mathbb{R}^d$,
     $$K(x, \cdot) \quad \text{is a probability measure,}$$
  2. for each $A \in \mathcal{B}^d$, $K(\cdot, A)$ is a measurable mapping,
  3. Markov property and conditioning on probability 0 events defined through the kernel and limits: For any $A \in \mathcal{B}^d$ and $x_0, \ldots, x_n \in \mathbb{R}^d$,
     $$\mathbb{P}(X_{n+1} \in A \mid X_{0:n} = x_{0:n}) := \mathbb{P}(X_{n+1} \in A \mid X_n \in dx_{0:n})$$
     and
     $$\mathbb{P}(X_{n+1} \in A \mid X_{0:n} \in dx_{0:n}) = \mathbb{P}(X_{n+1} \in A \mid X_n \in dx_n) := K(x_n, A).$$

Dynamics

Dynamics of the Markov chain $X_0 \sim \mathbb{P}^0$ and

$$X_{n+1} \sim K(X_n, \cdot).$$

Example difference equation

$$X_{n+1} = \theta X_n$$

Then

$$X_{n+1} | X_n = x_n \sim \delta_{\theta x_n} = K(x_n, \cdot)$$

A Markov chain may be deterministic (but it is then probably not practical to study it as a random process).

Example

Auto regressive AR(1) process on $\mathbb{R}$:

$$X_{n+1} = \theta X_n + \eta_n,$$

with $\theta \in \mathbb{R}$ iid sequence $\eta_k \sim N(0, \sigma^2)$.

Transition kernel:

$$X_{n+1} | X_n = x_n \sim N(\theta x_n, \sigma^2) = K(x_n, \cdot).$$

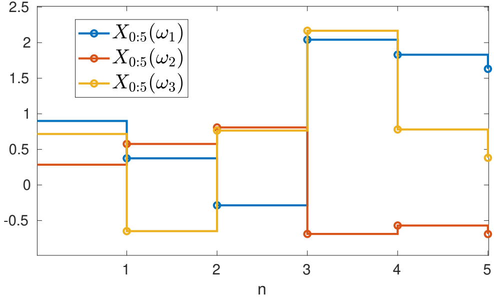


Figure: Simulations of AR(1) when $\theta = 1/2$, $\sigma = 1$ and $X_0 \sim U(0, 1)$.

Chapman-Kolmogorov equation

By the Markov property and law of total probability,

$$\mathbb{P}(X_2 \in A_2, X_1 \in A_1 | X_0 = x_0) = \int_{A_1} \mathbb{P}(X_2 \in A_2, X_1 \in dx_1 | X_0 = x_0)$$

$$= \int_{A_1} \mathbb{P}(X_2 \in A_2 | X_{0:1} = x_{0:1}) \mathbb{P}(X_1 \in dx_1 | X_0 = x_0)$$

$$= \int_{A_1} K(x_1, A_2) K(x_0, dx_1).$$

This leads to the Chapman-Kolmogorov equation: for any $A_1, \ldots, A_n \in \mathcal{B}^d$,

$$\mathbb{P}(X_{1:n} \in A_{1:n} | X_0 = x_0)$$

$$= \int_{A_{n-1}} \cdots \int_{A_2} \int_{A_1} K(x_{n-1}, A_n) K(x_{n-2}, dx_{n-1}) \ldots K(x_1, dx_2) K(x_0, dx_1)$$

Compare to discrete-space Markov chain on A:

$$\mathbb{P}(X_n = x_n | X_0 = x_0) = \sum_{x_{1:n-1} \in A^{n-1}} p(x_0, x_1) p(x_1, x_2) \ldots, p(x_{n-1}, x_n).$$

Remarks

Dynamics of the chain can also be described as dynamics $\mathcal{P}$, the space probability measures on $\mathbb{R}^d$:
Let the transition mapping $T : \mathcal{P} \to \mathcal{P}$ be defined by

$$(T\mathbb{P})(A) := \int_{\mathbb{R}^d} K(x, A)\mathbb{P}(dx)$$

and

$$\mathbb{P}^{n+1} = T\mathbb{P}^n$$

Invariant measure $\mathbb{P}$ is an invariant measure provided

$$\mathbb{P} = T\mathbb{P},$$

(Trivial example: AR(1) with $\theta = 0$ has invariant measure $\mathbb{P} = N(0, \sigma^2)$.)

Markov chains - density point of view

If there exists a function $k : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$ such that

$$K(x, A) = \int_A k(x, y) dy \qquad \forall x \in \mathbb{R}^d \quad A \in \mathcal{B}^d,$$

then k is the density kernel function, i.e., $k(x, \cdot) \in \mathcal{M}$ for every x.

And we can describe the Markov chain dynamics for densities: Let $T : \mathcal{M} \to \mathcal{M}$ be defined by

$$(T\pi)(y) = \int_{\mathbb{R}^d} k(x, y)\pi(x) dx$$

Invariant density π is an invariant density provided

$$\pi = T\pi, \quad \text{i.e. if } d_{TV}(\pi, T\pi) = 0.$$

And for the dynamics of the chain $X_0, X_1, \dots,$

$$X_{n+1} \sim k(X_n, \cdot).$$

Overview

1. Sampling of the posterior probability density function
   - Monte Carlo method
2. Importance sampling applied to posterior densities
3. Discrete-time continuous-space Markov chains
4. Markov chain Monte Carlo sampling/dynamics

Accept reject sampling

Problem setting: We have a **target density** π that we want to sample from.

Accept reject algorithm: Assume that we a **proposal density** $\hat{\pi}$ which we can draw samples from, and that for some $N \geq 1$, it holds that $N\hat{\pi} \geq \pi$.

Sample $X \sim \pi$ as follows:

1. sample $Y \sim \hat{\pi}$ and $U \sim U[0,1]$ with $U \perp Y$.
2. accept $X = Y$ with **acceptance probability** $U \leq \pi(Y)/(N\hat{\pi}(Y))$; otherwise return to step 1.

Verification that $X \sim \pi$:

$$\pi_X(x) = \frac{\mathbb{P}(Y \in dx \mid U \leq \pi(Y)/(N\hat{\pi}(Y)))}{dx} = \ldots \text{ubung 5}$$

Markov Chain Monte Carlo method (MCMC)

Input: target pdf π, a conditional proposal $q(y|x)$ (i.e., $q(\cdot|x) \in \mathcal{M}$ for every $x \in \mathbb{R}^d$).

Output: Markov chain $X_0, X_1, \ldots$ with objective that $\pi^M_{MCMC} = \frac{1}{M}\sum_{k=1}^{M} \delta_{X_k}$ approximates measure associated to π.

Metropolis-Hastings algorithm

Given X_n,

1. generate proposal $Y_n \sim q(\cdot|X_n)$
2. set

$$X_{n+1} = \begin{cases} Y_n & \text{with probability} \quad \rho(X_n, Y_n) \\ X_n & \text{with probability} \quad 1 - \rho(X_n, Y_n) \end{cases}$$

   where the M-H acceptance probability is defined by

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

Assumptions and properties of Metropolis Hastings

Assumptions

- must be able to sample from $q(\cdot|x)$ for relevant x
- π must be known up to a constant (i.e., relevant for posterior densities with Z unknown),
- $q(\cdot|x)$ must be known up to a constant that is independent of x.

Properties:

- When $q(x|y) = q(y|x)$ the test ratio becomes
$$\frac{\pi(y)}{\pi(x)}\frac{q(x|y)}{q(y|x)} = \frac{\pi(y)}{\pi(x)}.$$
- If $q(x|y) > q(y|x)$, then (compared to not having a q ratio in the acceptance probability), the probability accepting transitions $x \mapsto y$ increases. So transitions for which the reverse transition $q(x|y)$ is more often proposed than the transition itself, increases likelihood.
- If $q(x|y) < q(y|x)$, then (compared to not having a q ratio in the acceptance probability), the probability accepting transitions $x \mapsto y$ decreases.

M-H dynamics is associated to the transition kernel (ubung 5)

$$K(x,A) = \underbrace{\int_A \rho(x,y)q(y|x)dy}_{r(x,A)} + \Big(1 - r(x,\mathbb{R}^d)\Big)\delta_x(A)$$

Idea:

$$\begin{aligned} K(x,A) &= \mathbb{P}(X_1 \in A \mid X_0 = x) \\ &= \mathbb{P}(Y_0 \in A, X_1 = Y_0 \mid X_0 = x) + \mathbb{P}(x \in A, X_1 = x \mid X_0 = x) \end{aligned}$$

M-H properties

If $q(\cdot|x)$ dominates π for all x, then the M-H kernel satisfies detailed balance wrt π:

$$\int_A K(x,B)\pi(x)dx = \int_B K(x,A)\pi(x)dx \qquad \forall A,B \in \mathcal{B}^d,$$

and π is an invariant pdf of the M-H Markov chain.

Sketch of proof: Assume that $X_0 \sim \pi$. Then

$$\begin{aligned}
\mathbb{P}_{X_1}(A) &= \int_{\mathbb{R}^d} K(x,A)\mathbb{P}_{X_0}(dx) \\
&= \int_{\mathbb{R}^d} K(x,A)\pi(x)dx \\
&= \int_{\mathbb{R}^d} \left(\rho(x,y)q(y|x) - \Big(1 - r(x,\mathbb{R}^d)\Big)\delta_x(A)\right)\pi(x)\,dx \\
&= \dots \\
&= \int_A \pi(x)dx.
\end{aligned}$$

Remarks

Challenges in real applications: Choosing a proposal such that (1) one achieves convergence $\pi^n \to \pi$, (2) the convergence is fast in n, and (3) that acceptance of the proposal is frequent (for efficiency of MCMC).

See SST 6.4.2 for assumptions on prior and likelihood for $\pi(\cdot|y)$ in combination with Gaussian proposal $q(\cdot|x)$ which ensures convergence of the chain distribution.

If interested, "Monte Carlo Statistical Methods" by Robert and Casella is a good book on Monte Carlo and MCMC methods.

Next time

- Smoothing and filtering for discrete-time continuous state-space Markov chains.
- Discrete-time Kalman filtering and smoothing.