MA5233 Computational Mathematics

Lecture 25: Markov Chain Monte Carlo

Simon Etter

2019/2020

Markov Chain Monte Carlo

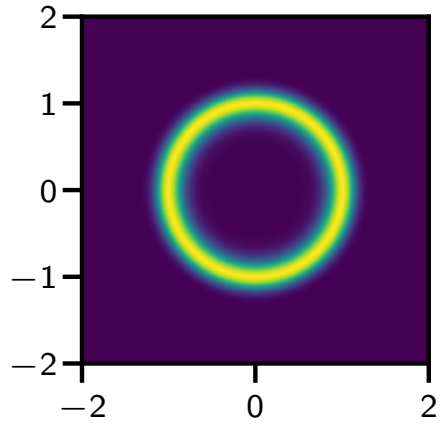
Introductory example

Assume we want to compute

$$I := \frac{\int_{\mathbb{R}^2} h(x)\, g(x)\, dx}{\int_{\mathbb{R}^2} g(x)\, dx}$$

where $h(x)$ is some given smooth function and

$$g(x) := \exp\Big(-10\,(x_1^2 + x_2^2 - 1)^2\Big).$$



Problem: Monte Carlo will be very inefficient unless we use importance sampling, but how do we sample $G \sim g(x)$?

Markov Chain Monte Carlo

Introductory example (continued)

Observation: if G is such that $g(G)$ is large, then $g(\tilde{G})$ is also large for $\tilde{G}$ near G.

Idea: instead of generating samples G_k independently, maybe we can generate samples more efficiently by choosing G_{k+1} as some random perturbation of G_k.

Such a sequence of G_k is a *Markov chain*.

Def: Markov chain

Sequence of random variables X_k such that

$$P(X_{k+1} = x_{k+1} \mid X_k = x_k, \ldots, X_0 = x_0) = P(X_{k+1} = x_{k+1} \mid X_k = x_k).$$

Markov chains are fully specified by

- initial distribution $p_0(x_0) := P(X_0 = x_0)$, and
- transition probability $p(x_{k+1} \mid x_k) := P(X_{k+1} = x_{k+1} \mid X_k = x_k)$.

Markov Chain Monte Carlo

Probability distribution of Markov chains

We have

$$P(X_1 = x_1) = \int P(X_1 = x_1 \mid X_0 = x_0)\, P(X_0 = x_0)\, dx_0$$
$$= \int p(x_1 \mid x_0)\, p_0(x_0)\, dx_0$$

and analogously

$$P(X_k = x_k) = \int \ldots \int p(x_k \mid x_{k-1}) \ldots p(x_1 \mid x_0)\, p_0(x_0)\, dx_{k-1} \ldots dx_0.$$

Introductory example (continued)

We want Markov chain G_k such that $P(G_k = x) = g(x)$.

For $k = 0$, this would require setting $p_0(x) = g(x)$, but doing so is not computationally feasible since we cannot sample from $g(x)$.

Instead, we choose $p_0(x)$ arbitrarily but carefully design $p(x_{k+1} \mid x_k)$ such that $P(X_k = x) \to g(x)$ for $k \to \infty$.

Markov Chain Monte Carlo

Def: Stationary distribution

Probability density $s(x)$ is said to be a stationary distribution of Markov chain with transition probability $p(x_{k+1} \mid x_k)$ if

$$s(x_{k+1}) = \int p(x_{k+1} \mid x_k)\, s(x_k)\, dx_k.$$

Ergodicity theorem

Markov chain $(X_k)_{k=0}^{\infty}$ has a unique stationary distribution and it holds

$$\lim_{k \to \infty} P(X_k = x) \to s(x) \qquad \text{for any } p_0(x_0)$$

if $(X_k)_{k=0}^{\infty}$ is ergodic, i.e. positively recurrent and aperiodic.

Loosely speaking, Markov chain is

- positively recurrent if it is possible to go from any state X to any other state X' using some finite number of steps n, and
- aperiodic if the greatest common divisor of all such n is 1.

Markov Chain Monte Carlo

Remarks

- Stationary distribution is "eigenvector" with eigenvalue 1 of the "matrix" A given by $(Af)(x_{k+1}) := \int p(x_{k+1} \mid x_k)\, f(x_k)\, dx_k$.
- We will assume without rigorous proof that the Markov chain constructed in the following is ergodic.

Thm: Detailed balance

$s(x)$ is a stationary distribution of a Markov chain with transition probability $p(x_{k+1} \mid x_k)$ if for all x_{k+1}, x_k it holds

$$p(x_{k+1} \mid x_k)\, s(x_k) = p(x_k \mid x_{k+1})\, s(x_{k+1}).$$

Remark: detailed balance condition is sufficient but not necessary.

Proof.

$$s(x_{k+1}) = \int p(x_k \mid x_{k+1})\, s(x_{k+1})\, dx_k = \int p(x_{k+1} \mid x_k)\, s(x_k)\, dx_k.$$

where we used $\int p(x_k \mid x_{k+1})\, dx_k = 1$ in first step and detailed balance in second step.

Markov Chain Monte Carlo

Recap

- We want to construct Markov chain G_k with prescribed stationary distribution $g(x)$.
- Ergodicity theorem showed that stationary distribution depends only on transition probability $p(x_{k+1} \mid x_k)$ but not on initial distribution $p_0(x_0)$.
- Detailed balance guarantees that $g(x)$ is a stationary distribution if

$$p(x_{k+1} \mid x_k)\, g(x_k) = p(x_k \mid x_{k+1})\, g(x_{k+1}).$$

- It therefore remains to construct $p(x_{k+1} \mid x_k)$ which satisfies the detailed balance condition.

Markov Chain Monte Carlo

Thm: Metropolis step

Let $s(x)$ and $q(x_{k+1} \mid x_k)$ be probability distributions, where

- $s(x)$ is the desired stationary distribution, and
- $q(x_{k+1} \mid x_k)$ is the proposal distribution.

Consider Markov chain $(X_k)_{k=0}^{\infty}$ generated as follows:

1. Generate proposal $\tilde{X}_{k+1} \sim q(\cdot \mid X_k)$ and $U \sim \text{Uniform}[0, 1]$.
2. Compute acceptance probability $R := \frac{q(X_k \mid \tilde{X}_{k+1})\, s(\tilde{X}_{k+1})}{q(\tilde{X}_{k+1} \mid X_k)\, s(X_k)}$.
3. If $U \leq R$, then set $X_{k+1} := \tilde{X}_{k+1}$. Otherwise, set $X_{k+1} := X_k$.

The transition probability $p(x_{k+1} \mid x_k)$ of $(X_k)_{k=0}^{\infty}$ satisfies the detailed balance condition

$$p(x_{k+1} \mid x_k)\, s(x_k) = p(x_k \mid x_{k+1})\, s(x_{k+1}).$$

Markov Chain Monte Carlo

Proof. We observe that detailed balance is trivially satisfied if $x_{k+1} = x_k$.
We therefore assume $x_{k+1} \neq x_k$ for the remained of the proof.
By construction we have

$$p(x_{k+1} \mid x_k) = q(x_{k+1} \mid x_k) \min\left\{1, \frac{q(x_k|x_{k+1})\, s(x_{k+1})}{q(x_{k+1}|x_k)\, s(x_k)}\right\}.$$

Without loss of generality, let us assume $\frac{q(x_k|x_{k+1})\, s(x_{k+1})}{q(x_{k+1}|x_k)\, s(x_k)} \leq 1$ such that

$$p(x_{k+1} \mid x_k) = q(x_{k+1} \mid x_k) \frac{q(x_k \mid x_{k+1})\, s(x_{k+1})}{q(x_{k+1} \mid x_k)\, s(x_k)},$$
$$p(x_k \mid x_{k+1}) = q(x_k \mid x_{k+1}).$$

Then,

$$\begin{aligned} p(x_{k+1} \mid x_k)\, s(x_k) &= q(x_{k+1} \mid x_k) \frac{q(x_k \mid x_{k+1})\, s(x_{k+1})}{q(x_{k+1} \mid x_k)\, s(x_k)}\, s(x_k) \\ &= q(x_k \mid x_{k+1})\, s(x_{k+1}) \\ &= p(x_k \mid x_{k+1})\, s(x_{k+1}). \end{aligned}$$

Markov Chain Monte Carlo

Remarks

- Metropolis step is similar to rejection sampling, but it differs in that we generate a new sample X_{k+1} even if proposal is rejected.
- If proposal probability is symmetric, $q(x_{k+1} \mid x_k) = q(x_k \mid x_{k+1})$, then acceptance probability simplifies to $R = s(\tilde{X}_{k+1})/s(X_k)$.

  In particular, proposals are always accepted if $s(\tilde{X}_{k+1}) > s(X_k)$.
- See `generate_samples()` and `plot_samples()`.

Error analysis

Markov chain Monte Carlo produces correlated samples X_k; hence central limit theorem does not apply without modification.

Theory is highly technical. Roughly speaking, central limit theorem holds if we introduce an additional factor N_{corr} ("autocorrelation time") in the formula for $\operatorname{Var}[Q]$,

$$\operatorname{Var}[Q] \approx \sqrt{\operatorname{Var}[F]} \sqrt{\frac{N_{\text{corr}}}{N}},$$

i.e. error behaves approximately like $\mathcal{O}\left(\sqrt{\frac{N_{\text{corr}}}{N}}\right)$.

Markov Chain Monte Carlo

Proposal distribution in Metropolis step

Metropolis step satisfies detailed balance for any proposal distribution.

Proposals are typically of the form $\tilde{X}_{k+1} = X_k + \Delta x\, \xi$ for some step length Δx and random noise ξ with $\mathbb{E}[\xi] = 0$, $\|\mathrm{Var}[\xi]\| \approx 1$.

Performance depends heavily on choosing Δx appropriately.

- If Δx is too large, then most proposals get rejected and Markov chain remains constant for a long time.
- If Δx is too small, then most proposals get accepted but it takes a long time to sample the whole space.

In both of these extreme cases, the consequence is that the autocorrelation time N_{corr} becomes large.

See `step_length()`.

Markov Chain Monte Carlo

Review Markov Chain Monte Carlo

- MCMC allows us to do importance sampling in cases where generating samples would be difficult otherwise.
  In particular, MCMC works even when the target distribution is only known up to a normalising constant.
- Error analysis of MCMC methods is much more involved than standard Monte Carlo. In particular, two new parameters appear:
  - Burn-in time N_{start}: number of steps required for the Markov chain to "forget" initial distribution $p_0(x_0)$.
  - Autocorrelation time N_{corr}: number of steps required such that $X_{k+N_{\text{corr}}}$ becomes independent of X_k.
- Burn-in time reduces parallel efficiency of MCMC.