

BML lecture #2: MCMC

<http://github.com/rbardenet/bml-course>

Rémi Bardenet

CNRS & CRIStAL, Univ. Lille, France



- 1** Introduction
- 2** Monte Carlo methods
- 3** The Metropolis-Hastings algorithm
- 4** Gibbs sampling
- 5** Hamiltonian Monte Carlo
- 6** Convergence diagnostics for MCMC

- 1 Introduction**
- 2 Monte Carlo methods
- 3 The Metropolis-Hastings algorithm
- 4 Gibbs sampling
- 5 Hamiltonian Monte Carlo
- 6 Convergence diagnostics for MCMC

What comes to *your* mind when you hear "Monte Carlo"?

Minimizing the posterior expected loss

If we partition $s = (s_{\text{obs}}, s_u)$, then, given s_{obs} , we choose

$$a^* = \delta(s_{\text{obs}}) = \arg \min_{a \in \mathcal{A}} \mathbb{E}_{s_u | s_{\text{obs}}} L(a, s).$$

The bottleneck is computing integrals w.r.t. the posterior

- ▶ E.g. for binary prediction with 0-1 loss

$$y^* \in \arg \max_{y \in \{0,1\}} \int p(y|x, \theta) p(\theta | x_{1:n}, y_{1:n}) d\theta$$

- ▶ or for estimation with squared loss

$$\theta^* = \int \theta p(\theta | y_{1:n}) d\theta.$$

Let π be a pdf w.r.t. $d\theta$.

The problem of numerical integration

Find T nodes (θ_t) and weights (w_t) so that

$$\int f(\theta)\pi(\theta)d\theta \approx \sum_{t=1}^N w_t f(\theta_t), \quad \forall f \in \mathcal{C},$$

where \mathcal{C} is a large class of functions.

A constraint for Bayesians: π is only known up to a constant

E.g. in estimation,

$$\pi(\theta) = p(\theta|y_{1:n}) \propto p(y_{1:n}|\theta)p(\theta) =: \pi_u(\theta).$$

Or in classification/regression,

$$\pi(\theta) = p(\theta|x_{1:n}, y_{1:n}) \propto p(y_{1:n}|x_{1:n}, \theta)p(\theta) =: \pi_u(\theta).$$

- ▶ For modern developments, see quasi-Monte Carlo integration [Dick and Pilichshammer, 2010](#).

- 1 Introduction
- 2 Monte Carlo methods**
- 3 The Metropolis-Hastings algorithm
- 4 Gibbs sampling
- 5 Hamiltonian Monte Carlo
- 6 Convergence diagnostics for MCMC

The Monte Carlo principle

Find a distribution on $\theta_1, \dots, \theta_T$ and weights w_t such that

$$\mathcal{E}_T(f) = \sum_{t=1}^T w_t f(\theta_t) - \int f(\theta) \pi(\theta) d\theta$$

is small (with large probability, in quadratic mean, converges in law at some rate, etc.)

- If you knew how to sample from π , you could take $\theta_t \sim \pi$ i.i.d., $w_t = 1/T$, and prove e.g.

$$\mathbb{P} \left(\mathcal{E}_T(f) \geq \alpha \frac{\sigma(f)}{\sqrt{T}} \right) \leq \frac{1}{\alpha^2}, \quad \forall \alpha,$$

as soon as $\sigma(f)^2 := \mathbb{V}_\pi[f(\theta) - \int f(\theta) \pi(\theta) d\theta] < +\infty$.

- ▶ Let $\pi_u(\theta) = Z\pi(\theta)$ be the unnormalized target pdf.
- ▶ Sample $\theta_{1:T}$ i.i.d. from q , and take

$$w_t = \frac{\pi_u(\theta_t)}{q(\theta_t)} \times \left(\sum_{t=1}^T \frac{\pi_u(\theta_t)}{q(\theta_t)} \right)^{-1}$$

so that $\sum w_t = 1$.

- ▶ Then
-
-
-
-
-
-
-
-
-
-
- ▶ One can show that $\sqrt{T}\mathcal{E}_T(f) \rightarrow \mathcal{N}(0, \sigma_{\text{NIS}}^2(f))$.
 - ▶ Problem is that for reasonable choices of f, q, π , $\log \sigma_{\text{NIS}}(f) \propto d$.

- 1 Introduction
- 2 Monte Carlo methods
- 3 The Metropolis-Hastings algorithm**
- 4 Gibbs sampling
- 5 Hamiltonian Monte Carlo
- 6 Convergence diagnostics for MCMC

(Mostly) friendly faces



Figure: A few MCMC pioneers: N. Metropolis, S. Ulam, A. Rosenbluth, W. K. Hastings

- ▶ The idea is to take (θ_t) to be an ergodic Markov chain with limiting distribution π , so that for $f \in L^1(\pi)$,
- ▶ In MCMC research, when a new Markov kernel comes out, we typically first prove a **law of large numbers**, and then a **central limit theorem**, i.e., that under weak conditions on π and f ,

and that $\sigma^2(f)$ can be estimated; see (Douc, Moulines, and Stoffer, 2014).

Let $(\theta_t)_{t \in \mathbb{N}}$ be a Markov chain on Θ , with Markov kernel P . If

- There exists π s.t.

$$\int d\pi(x)P(x, B) = \pi(B).$$

- For any A with $\pi(A) > 0$, for any $\theta \in \Theta$,

$$\mathbb{P}_\theta \left(\sum_{t=0}^{\infty} 1_{\theta_t \in A} = +\infty \right) = 1,$$

then for any f such that $\int |f| d\pi < \infty$, for any initial distribution μ_0 of θ_0 , almost surely

$$\frac{1}{T} \sum_{t=1}^T f(\theta_t) \rightarrow \int f d\pi.$$

See e.g. (Douc, Moulines, and Stoffer, 2014).

The Metropolis-Hastings algorithm

MH(π_u , $q(\cdot|\cdot)$, θ_0 , T)

1 **for** $t \leftarrow 1$ **to** T

2 $\theta \leftarrow \theta_{t-1}$

3 $\theta' \sim q(\cdot|\theta)$, $u \sim \mathcal{U}_{(0,1)}$,

4 $\rho = \frac{\pi(\theta')}{\pi(\theta)} \frac{q(\theta|\theta')}{q(\theta'|\theta)}$.

5 **if** $u < \rho$,

6 $\theta_t \leftarrow \theta'$ \triangleright *Accept*

7 **else** $\theta_t \leftarrow \theta$ \triangleright *Reject*

8 **return** $(\theta_t)_{t=1,\dots,N_{\text{iter}}}$

... is given by

$$P_{\text{MH}}(\theta, \theta') = \alpha(\theta, \theta') q(\theta' | \theta) + \delta_{\theta}(\theta') \left[1 - \int \alpha(\theta, \vartheta) q(\vartheta | \theta) \right] d\vartheta,$$

where

$$\alpha(\theta, \theta') = 1 \wedge \frac{\pi(\theta')}{\pi(\theta)} \frac{q(\theta | \theta')}{q(\theta' | \theta)}.$$

- ▶ We first show detailed balance, i.e., $\pi(\theta)P(\theta, \theta') = \pi(\theta')P(\theta', \theta)$.
- ▶ We deduce that P leaves π invariant.

Theorem (Robert and Casella, 2004)

If $\pi(A) > 0 \Rightarrow (\forall x)q(A|x) > 0$, then P_{MH} satisfies the LLN.

Some additional useful properties

- ▶ Note that if P_1 and P_2 leave π invariant, then so does

$$P_1 P_2(\theta, \theta') = \int P_1(\theta, \vartheta) P_2(\vartheta, \theta') d\vartheta.$$

- ▶ The MH error scales polynomially with the dimension; see blog post.

- 1 Introduction
- 2 Monte Carlo methods
- 3 The Metropolis-Hastings algorithm
- 4 Gibbs sampling**
- 5 Hamiltonian Monte Carlo
- 6 Convergence diagnostics for MCMC

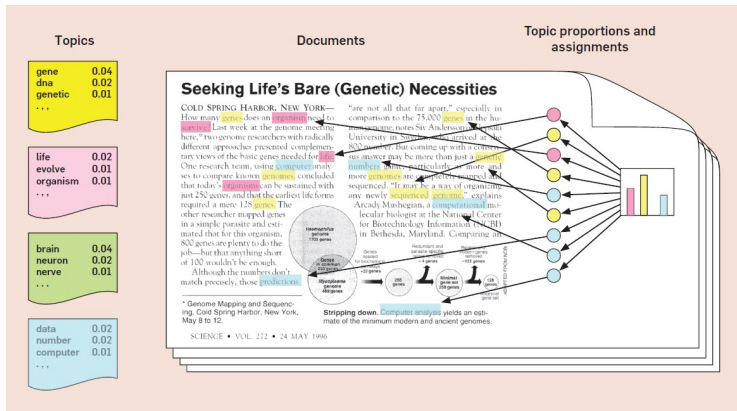
- ▶ Consider MH with

$$q(\theta'|\theta) = \frac{1}{d} \sum_{k=1}^d \pi(\theta'_k|\theta_{\setminus k}) \mathbf{1}_{\theta'_k=\theta_k}, \quad \theta_{\setminus k} := (\theta_1, \dots, \theta_{k-1}, \theta_{k+1}, \dots, \theta_d).$$

- ▶ Then the probability of acceptance $\alpha(\theta, \theta')$ is always 1.

- ▶ In practice, the systematic scan Gibbs sampler is more common, which consists in repeatedly: drawing $\theta_1|\theta_{\setminus 1}$, then $\theta_2|\theta_{\setminus 2}$, etc. always conditioning on the newest values available of each θ_k .
- ▶ You can also partition θ in arbitrary blocks.

An example: Latent Dirichlet allocation



- 1 Introduction
- 2 Monte Carlo methods
- 3 The Metropolis-Hastings algorithm
- 4 Gibbs sampling
- 5 Hamiltonian Monte Carlo**
- 6 Convergence diagnostics for MCMC

- ▶ Let S be a linear involution of $\mathcal{X} \subset \mathbb{R}^{2d}$, such that $\eta \circ S = \eta$ for some (possibly unnormalized) PDF η .

- ▶ Let further $\Phi : \mathbb{R}^{2d} \rightarrow \mathbb{R}^{2d}$ be a C^1 -diffeomorphism such that $S \circ \Phi = \Phi^{-1} \circ S$.

- ▶ Now let

$$\alpha(x) \triangleq 1 \wedge \frac{\eta(\Phi(x))}{\eta(x)} |\Phi'(x)|, \quad (1)$$

and consider the Markov kernel

$$P_{aHMC}(x, A) = \alpha(x) 1_{\Phi(x) \in A} + (1 - \alpha(x)) 1_{S(x) \in A}.$$

Proposition

P_{aHMC} leaves η invariant.

Hamilton's equations of motion

Consider a physical system described by Hamiltonian $H(x, \xi)$ in phase space $(x, \xi) \in \mathbb{R}^{2d}$. Then the trajectories are prescribed by

$$\dot{x}_i = \frac{\partial H}{\partial \xi_i} \quad \dot{\xi}_i = -\frac{\partial H}{\partial x_i}. \quad (2)$$

- ▶ Given an initial point (x, ξ) , solve (2) and denote the corresponding position in \mathbb{R}^{2d} at time $t > 0$ by $\Phi_t(x, \xi)$.
- ▶ (2) implies that $t \mapsto H(\Phi_t(x, \xi))$ is constant.
- ▶ As an example, consider $H(x, \xi) = \frac{1}{2}x^2 + \frac{1}{2}\xi^2$.

- ▶ One idea would be to put some monotone function of the target in the Hamiltonian, such as $H(q, p) = -\log \pi(q) + \frac{1}{2}\xi^T M \xi$.
- ▶ We know approximations of the Hamiltonian flow, such as the leapfrog (aka velocity Verlet) integrator. It is defined as $\psi_h^n = \psi_h \circ \dots \circ \psi_h$, where $(p', q') = \psi_h(p, q)$ is

$$p_{1/2} = p + \frac{h}{2} \nabla \log \pi(q)$$

$$q' = q + hM^{-1}p_{1/2}$$

$$p' = p_{1/2} + \frac{h}{2} \nabla \log \pi(q');$$

Proposition

The leapfrog integrator satisfies $S \circ \psi_h^n = (\psi_h^n)^{-1} \circ S$ for $S(q, p) = (q, -p)$, and $|\det(\psi_h^n)'(q, p)| = 1$.

Hamiltonian Monte Carlo mimics a physical system

- ▶ Let $\log \tilde{\pi}(x, \xi) = \log \pi(x) + \frac{1}{2} \xi^T M(x) \xi$.
- ▶ Consider the Markov kernel $\tilde{P}((x, \xi), (x, \xi'))$ given by the product of

$$\xi \sim \mathcal{N}(0, M(x)^{-1})$$

and

$$P_{aHMC}(x, A) = \alpha(x) 1_{\Phi(x) \in A} + (1 - \alpha(x)) 1_{S(x) \in A}.$$

where

$$\alpha(x) \triangleq 1 \wedge \frac{\tilde{\pi}(\psi_h^n(x))}{\tilde{\pi}(x)} \frac{|\psi_h^n(x)|}{|(\psi_h^n)'(x)|}, \quad (3)$$

Then \tilde{P} leaves π invariant.

- 1 Introduction
- 2 Monte Carlo methods
- 3 The Metropolis-Hastings algorithm
- 4 Gibbs sampling
- 5 Hamiltonian Monte Carlo
- 6 Convergence diagnostics for MCMC**

What can go wrong?

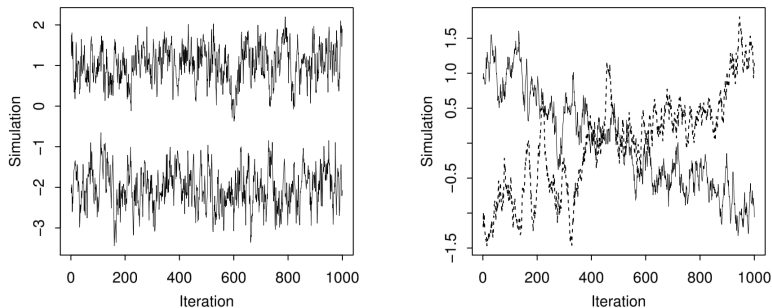


Figure: Taken from (Gelman et al., 2013)

We need to monitor both cross-chain and within-chain behavior.

Comparing P chains with overdispersed starting points

- ▶ The behaviour of the P traces should become similar.
- ▶ Always make visual sanity checks!
- ▶ Scalar estimates should converge to the same value.
- ▶ We can also compare the variance of a scalar estimate within- and across chains

The Gelman-Rubin diagnostic

- ▶ Choose an f of interest, e.g. $f(\theta) = \theta_1$.
- ▶ Compute $B := \frac{T}{P-1} \sum_{p=1}^P (\bar{f}_{\cdot p} - \bar{f}_{\cdot\cdot})^2$.
- ▶ Compute $W := \frac{1}{P} \sum_{p=1}^P \left[\frac{1}{T-1} \sum_{t=1}^T (\bar{f}_{tp} - \bar{f}_{\cdot p})^2 \right]$.
- ▶ Then check whether

$$\hat{R} = \sqrt{\frac{\frac{T-1}{T} W + \frac{1}{T} B}{W}} \in [1, 1.1].$$

- ▶ See (Vats and Knudson, 2021) for an insightful discussion.

Single-chain diagnostics

- ▶ The idea is to compare different chunks of a single chain.
- ▶ At stationarity, large chunks should be statistically hard to distinguish.
- ▶ The **Geweke diagnostic** tests this similarity (Geweke, 1992)

Effective sample size

- ▶ Autocorrelation in each chain is what increases the variance of scalar estimands, compared to i.i.d. draws from π .
- ▶ We can estimate this autocorrelation, and build an estimator for the ratio of the two variances $\widehat{ESS} \in [1, PT]$, called the **effective sample size**; see Section 11.5 of (Gelman et al., 2013).

- ▶ Note that

$$\hat{R} \approx \sqrt{1 + P/\widehat{ESS}},$$

so $\hat{R} = 1.1$ **only** corresponds to $\widehat{ESS} = 5P$ (Vats and Knudson, 2021).

Take-home message

- ▶ MCMC approximates the integrals in the expected utility framework.
 - ▶ Try to **leverage the problem's structure** to design your kernels.
 - ▶ Otherwise, try standard kernels like HMC.
 - ▶ Always monitor convergence.
-
- ▶ HMC with NUTS is the default choice in most probabilistic programming frameworks.
 - ▶ MCMC is a **rich research topic**. Some keywords: Wang-Landau, Langevin, equi-energy, hit-and-run, bouncy particle sampler.
 - ▶ Besides Markov chains, checkout **sequential Monte Carlo samplers** (Del Moral, Doucet, and Jasra, 2006).
 - ▶ Deterministic methods are also investigated: **quasi-Monte Carlo methods** (Dick and Pillichshammer, 2010) have the best convergence rates as soon as the integrand is smooth.

Take-home message

- ▶ MCMC approximates the integrals in the expected utility framework.
 - ▶ Try to **leverage the problem's structure** to design your kernels.
 - ▶ Otherwise, try standard kernels like HMC.
 - ▶ Always monitor convergence.
-
- ▶ HMC with NUTS is the default choice in most probabilistic programming frameworks.
 - ▶ MCMC is a **rich research topic**. Some keywords: Wang-Landau, Langevin, equi-energy, hit-and-run, bouncy particle sampler.
 - ▶ Besides Markov chains, checkout **sequential Monte Carlo samplers** (Del Moral, Doucet, and Jasra, 2006).
 - ▶ Deterministic methods are also investigated: **quasi-Monte Carlo methods** (Dick and Pillichshammer, 2010) have the best convergence rates as soon as the integrand is smooth.

- [1] P. Del Moral, A. Doucet, and A. Jasra. “Sequential Monte Carlo samplers”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.3 (2006), pp. 411–436.
- [2] J. Dick and F. Pilichshammer. *Digital Nets and Sequences. Discrepancy Theory and Quasi-Monte Carlo Integration*. Cambridge University Press, 2010.
- [3] R. Douc, É. Moulines, and D. Stoffer. *Nonlinear time series*. Chapman-Hall, 2014.
- [4] A. Gelman et al. *Bayesian data analysis*. 3rd. CRC press, 2013.
- [5] J. Geweke. “Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments”. In: *Bayesian statistics 4* (1992), pp. 641–649.
- [6] C. P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer, 2004.
- [7] D. Vats and C. Knudson. “Revisiting the Gelman–Rubin Diagnostic”. In: *Statistical Science* 36.4 (2021), pp. 518–529.