BML lecture #2: MCMC

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

Rémi Bardenet

CNRS & CRIStAL, Univ. Lille, France

CRIStAL
Centre de Recherche en Informatique,
Signal et Automatique de Lille

Outline

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

Expected utility requires computing integrals

Minimizing the posterior expected loss

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

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

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

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

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

- or for estimation with squared loss

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

Let π be a pdf w.r.t. $\mathrm{d}\theta$.

The problem of numerical integration

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

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

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,

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

Riemann-like integration is impractical when $d \gg 1$.

- For modern developments, see quasi-Monte Carlo integration Dick and Pilichshammer, 2010.

The Monte Carlo principle

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

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

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.

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

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

- 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 $\sqrt{T}\mathcal{E}_T(f) \to \mathcal{N}(0, \sigma^2_{\mathrm{NIS}}(f))$.
- Problem is that for reasonable choices of f, q, π, $\log \sigma_{\mathrm{NIS}}(f) \propto d$.

Outline

(Mostly) friendly faces

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

Markov chain Monte Carlo (MCMC; (Robert and Casella, 2004))

- 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).

A law of large numbers for Markov chains

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

- There exists π s.t.

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

- 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 $\int |f|\mathrm{d}\pi < \infty$, for any initial distribution μ_0 of θ_0, almost surely

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

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

The Metropolis-Hastings algorithm

$\mathrm{MH}\big(\pi_u,\ q(\cdot|\cdot),\ \theta_0,\ T\big)$

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

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

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

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

  if $u < \rho,$

    $\theta_t \leftarrow \theta'$ ▷ *Accept*

  else $\theta_t \leftarrow \theta$ ▷ *Reject*

return $(\theta_t)_{t=1,\ldots,N_{\mathrm{iter}}}$

The MH Markov kernel...

... is given by

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

where

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

MH leaves π invariant and satisfies the LLN

- 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

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

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

Outline

The random scan Gibbs sampler

- Consider MH with

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

- 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

Collapsed Gibbs sampling for LDA

Outline

Hamiltonian dynamics is the source of inspiration

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

$$\dot{x}_i = \frac{\partial H}{\partial \xi_i} \qquad \dot{\xi}_i = -\frac{\partial H}{\partial x_i}. \tag{1}$$

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

Hamiltonian Monte Carlo mimics a physical system

- Let $\log \pi(x, \xi) = \log \pi(x) + \frac{1}{2}\xi^T M(x)\xi$.
- For $t > 0$ fixed, consider the Markov kernel $P((x, \xi), (x, \xi'))$ corresponding to
$$\xi \sim \mathcal{N}(0, M(x)^{-1})$$
followed by
$$(x', \xi') = \varphi_T(x, \xi).$$
Then $\pi(x, \xi)$ is invariant for P, and so is its marginal $\pi(x)$.

- Integrating the Hamilton flow can lead to long jumps compared to MH with a Gaussian proposal, especially in high dimensions.
- In practice, φ_T has to be approximated, thus requiring an acceptance step. Parameters like T have to be tuned, as in NUTS (Hoffman and Gelman, 2014), which favors long jumps with no U-turns.

What can go wrong?

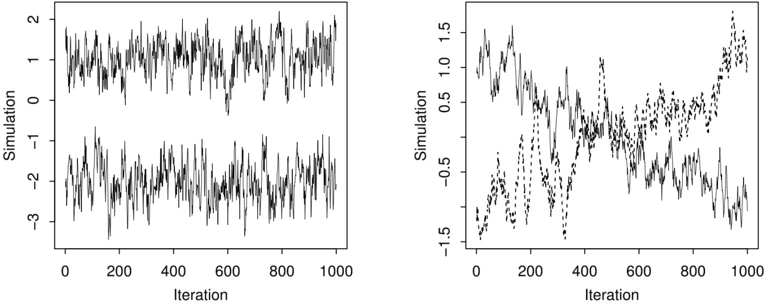

Figure: Taken from (Gelman et al., 2013)

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

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

More convergence diagnostics

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 (**Gew04**)

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).

Conclusion

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 Pilichshammer, 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 Pilichshammer, 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] M. D. Hoffman and A. Gelman. "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo.". In: *Journal of Machine Learning Research* 15.1 (2014), pp. 1593–1623.

[6] C. P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer, 2004.