Theory of Markov Chains Monte Carlo

Some basics

Introduction

- What is Markov chain Monte Carlo (MCMC)?
  - Run an ergodic Markov chain with invariant distribution π,
  - Use sample averages from this Markov chain to compute expectations
- We need a π invariant Markov Probability kernel K

Reading List

- Tierney (1994) Markov Chains for Exploring Posterior Distributions. Ann. Statist.
- Gelman, Gilks, & Roberts (1997) Weak convergence and optimal scaling of random walk Metropolis algorithms, Ann. Appl. Probab.
- Roberts & Rosenthal (2004). General state space Markov chains and MCMC algorithms. Probab. surv., 1, 20-71

Preliminaries

- We want to compute expectations:

$$\pi(\varphi) = \int_{\mathcal{X}} \varphi(x)\pi(dx)$$

- where π is a target distribution on $\mathcal{X}$:

$$\pi(dx) = \frac{\gamma(x)}{Z} dx$$

  with Z **unknown.**

- MCMC sampling procedure:

$$X_0 \sim \nu, X_1 \sim K(X_0, \cdot), X_2 \sim K(X_1, \cdot), \ldots, X_N \sim K(X_{N-1}, \cdot), \ldots$$

  - approximation:

$$\widehat{\pi}(\varphi) = \frac{1}{N} \sum_{i=1}^{N} \varphi(X_i)$$

Introduction to MCMC

- What principles does it make sense to invoke for $\widehat{\pi}(\varphi)$?

1. convergence of K^n to π in some sense (e.g. L^2, total variation norm, Wasserstein distance,...)
2. SLLN $\widehat{\pi}(\varphi) \rightarrow_{N\rightarrow\infty} \pi\left(\varphi\right)$ for $\varphi \in L^1(\pi)$
3. CLT for $\sqrt{N}(\widehat{\pi}(\varphi) - \pi\left(\varphi\right)) \rightarrow \mathcal{N}(0, \sigma^2)$, $\varphi \in L^2(\pi)$,
    3.1 CLT variance useful to characterise asymptotic sampling error in $\widehat{\pi}(\varphi)$
    3.2 can be used to derive measure of Effective Sample Size

Outline

- We will relate with theory of **Markov Chains in general spaces**
  - (...,Revuz 75, Nummelin 84, Kipnis & Varhadhan 86, Meyn & Tweedie 92, ...)
- Given K and x_0, one typically checks
  - π is unique invariant distribution
  - irreducibility, aperiodicity, reversibility
- Want to study convergence of $\widehat{\pi}(\varphi)$
  - Harris recurrence
  - speed of convergence w.r.t x_0 (**geometric ergodicity**)
- Significant MCMC theory relate with tuning K in various contexts
  - e.g. diffusive limits of Roberts et. al.

Stochastic differential equations (SDE) for sampling

- Consider the following (overdampled) **Langevin** Ito-SDE

$$dX_t = \frac{1}{2}\nabla \log \gamma(X_t)dt + dB_t \tag{1}$$

- The stationary distribution for X_t is π
  - rate of convergence to equilibrium depends on the tails of π
- If one could sample exactly $X_0, X_{t_1}, X_{t_2},...$ for $0 < t_1 < t_2 < \dots$ then this is a MCMC procedure
- Of course this is rarely possible so one needs to resort to numerical approximations for solving SDEs.

SDEs for optimisation

- One can use also “annealing”

$$dX_t = \nabla \log \gamma(X_t) dt + \sqrt{2\beta_t^{-1}} dB_t \qquad (2)$$

- If $\beta_t = \beta$ the stationary distribution π^β
- Want to invoke a Laplace or annealing principle (Hwang 81)

$$\pi^{\beta_t} \to \delta_{x^*} \text{ or } \frac{1}{n^*} \sum_{i=1}^{n^*} \delta_{x_i^*}$$

- Simulated annealing uses $\beta_t \propto \log t$ so that $X_t \xrightarrow{\mathbb{P}} x^*$ for large t.
  - starting β_t lower earlier in time balances exploration/exploitation trade-off
  - Many papers: Gidas, Kushner, Geman+Hwang, Hwang+Sheu, Holley+Kusuoka+Stroock ....

Metropolis Hastings

- Resulting Markov transition kernel:

$$K(x, dy) = \alpha(x, y)Q(x, dy) + \delta_x(dy) \int (1 - \alpha(x, y))Q(x, dy)$$

- aking densities w.r.t dx: let $dQ = qdx$

$$\alpha(x, y) = \begin{cases} 1 \wedge \frac{\gamma(y)q(y,x)}{\gamma(x)q(x,y)} & \gamma(x)q(x, y) > 0 \\ 0 & \gamma(x)q(x, y) = 0 \end{cases}$$

- **Reversibility** of K with π holds

$$\pi(dx)K(x, dy) = \pi(dy)K(y, dx)$$

Understanding MCMC

- More formulations for $\alpha(x, y)$ are possible to result to a reversible Markov chain w.r.t π
  - e.g. Barker, Liu book p114
  - MH acceptance ratio is most efficient **(Peskun-Tierney ordering)**
- Reversibility implies

$$\pi K = \pi$$

  - is π a unique invariant distribution?

Understanding MCMC: some questions

$\boxed{\text{Does } \widehat{\pi}(\varphi) \text{ converge to } \pi(\varphi) \text{ and how fast?}}$

1. is π unique?
2. (ergodicity)does $P^n(x_0, \cdot)$ converges to $\pi(\cdot)$?
3. (rate of convergence) how fast?
4. (initialisation) does choice of x_0 matter?

- What additional conditions are needed to establish 1-4?

Basic properties for a Markov kernel K

1. **Irreducibility** (controllability) means every part of state space can be reached
   - or all the support of π here
2. **Aperiodicity** means the state trajectory cannot go through a repeated cycle of subsets $A_1, \ldots, A_T$ w.p.1
3. **Recurrence:** for each B with $\pi(B) > 0$

$$\begin{aligned} \mathbb{P}_{x_0}\left[X_n \in B \text{ i.o.}\right] &> 0, \quad \forall x_0 \in \mathcal{X} \\ \mathbb{P}_{x_0}\left[X_n \in B \text{ i.o.}\right] &= 1, \quad \text{for } \pi \text{ almost all } x_0 \end{aligned}$$

   i.e. all states can (or will) be visited infinitely often

- In general 1-3 are used to establish existence & uniqueness of π

Short answers on ergodicity

- MCMC case: If in addition to $\pi K = \pi$, K is also irreducible and aperiodic
  - π is unique
  - $K^n(x_0, \cdot) \to_{n\to\infty} \pi$, in total variation, for π-almost all x_0
    - and then $\widehat{\pi}(\varphi) \to \pi(\varphi)$ is a bit "weak"
- Convergence holds for π -almost all x_0
  - this is not satisfying as often it is not easy to pick the "right" initial condition
  - need to require more than **irreducibility** and **aperiodicity** for π-invariant K

Short answers on ergodicity

- Typical requirement
  - **Harris recurrence:**
    - $\mathbb{P}_{x_0}[X_n \in B \text{ i.o.}] = 1, \quad \forall x_0 \in \mathcal{X}$
    - there is a small set with a.s. finite hitting times (see below)
  - then $K^n(x_0, \cdot) \to_{n\to\infty} \pi$, in total variation, <u>for all</u> $x_0 \in \mathcal{X}$
    - SLLN $\widehat{\pi}(\varphi) \to \pi(\varphi)$ a.s. for all $x_0 \in \mathcal{X}$ and $\pi(|\varphi|) < \infty$.
- MH case:
  - π-irreducibility implies Harris recurrence

Short answers on rates of convergence

- Basics on convergence properties of Markov chains useful
  - ergodicity requires

$$\|K^n(x_0,\cdot)-\pi\| \le r(x_0,n), \quad r(x_0,n) \to_{n\to\infty} 0$$

- In MCMC $r(x_0,n)$ depends on x_0 directly and also on π, $\mathcal{X}$, Q

Short answers on rates of convergence

- Different types of ergodic behaviour
  - polynomial: there exists a $\kappa(x) > 1$ and $p > 1$ s.t. for all $r < p$

    $$\|K^n(x_0, \cdot) - \pi\| \leq \kappa(x_0) n^{-r},$$

  - **geometric ergodicity:** there exist a $\lambda \in (0, 1)$ and $V(x)$ s.t.

    $$\|K^n(x_0, \cdot) - \pi\| \leq V(x_0) \lambda^n,$$

  - uniform ergodicity $r(x, n) \to 0$ uniformly on x as $n \to \infty$
- General results can be obtained for general classes of MCMC, e.g. MH, Independence sampler,Gibbs, MwG, HMC....

Some definitions for general state spaces

- A chain is ϕ-irreducible if there exists a non-zero measure ϕ on $\mathcal{X}$ s.t for all $A \in \mathcal{X}$ with $\phi(A) > 0$, and for all $x \in \mathcal{X}$, there exists a positive integer n such that $K^n(x, A) > 0$.
  - common example for $\mathbb{R}^d$ is Lebegue measure
  - here we will use $\phi = \pi$
- A set C **is small** if there exists an integer m, a constant ϵ and a probability measure μ s.t.

$$K^m(x, A) \geq \epsilon\mu(A), \quad \forall x \in C \text{ and } A \text{ s.t. } \phi(A) > 0$$

  - small sets are used to extend notion of **atoms** in discrete state spaces
  - every set A with $\phi(A) > 0$ contains a small set
  - here will assume $m = 1$

Detour with discrete states

- Lets consider a singleton state x^* and set $\mu = 1_{y=x^*}$ and $K(x, x^*) \geq \epsilon$ so
$$K \geq \epsilon\mu$$
- Try to solve $\pi K = \pi$ and note null space of $K - I$ is non trivial
- Try instead to invert
$$\pi \left(I - (K - \epsilon\mu) \right) = \epsilon\mu$$
to get
$$\pi = \epsilon\mu G$$
where
$$G = \sum_{n \geq 0} (K - \epsilon\mu)^n$$

Detour with discrete states

- Some calcutions give characterisation of π

$$\pi(x) = \frac{\mathbb{E}_{x^*}\left[\sum_{n=0}^{\tau^*-1} 1_{X_n=x}\right]}{\mathbb{E}_{x^*}\left[\tau^*\right]} = \frac{\rho(x)}{\rho(\mathcal{X})}$$

  with $\rho = \mu G$ and $\tau^* = \min_{n\geq 1}\left\{X_n = x^*\right\}$.
  - need $\rho(\mathcal{X}) < \infty$ i.e. recurrence

- For general state spaces will use small set C instead of x^*

Stability and small sets

- One would like establish **stability (recurrence)** by checking the **return times** to **a small set** C.
  - define stopping times $\tau_C = \min_{n \geq 1} \{X_n \in C\}$
- A weak requirement for existence of an invariant measure π is to check whether

$$\sup_{x \in C} \mathbb{E}_x [\tau_C] < M < \infty$$

- Convergence result is quite weak.
  - $K^n(x_0, \cdot) \to_{n \to \infty} \pi$, holds for $x_0 \in \{x : V(x) < \infty\}$

Stability and drift conditions

- Equivalently can verify Foster's condition:
  - there exists a $V \geq 0$ with $V(x') < \infty$ for some x' s.t.

$$KV(x) \leq -1 + V(x) + b1_{x \in C} \quad x \in \mathcal{X}$$

- This is a Lyapunov type approach

Stability and drift conditions

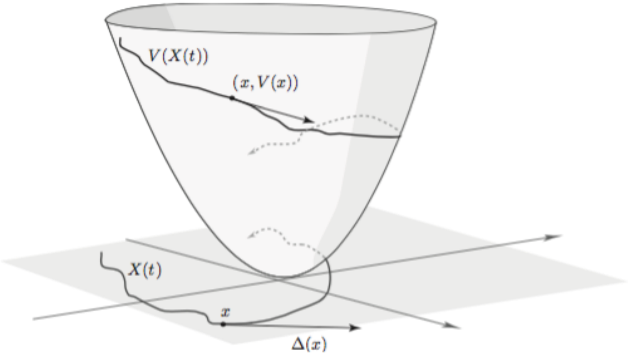

Figure: An illustration of Lyapunov function, here $\Delta = K - I$. Source: S. Meyn (2007) Control Techniques for Complex Networks

Harris recurrence and ergodicity

- Strengthen by requiring **Harris recurrence**: for a small set C

$$\mathbb{P}_{x_0}\left[\tau_C < \infty\right] = 1, \quad \forall x_0 \in \mathcal{X}$$

  - Then $K^n(x_0, \cdot) \to_{n\to\infty} \pi$, holds $\underline{\text{for } x_0 \in \mathcal{X}}$
- **Harris recurrence** is **equivalent** to sample averages converging (SLLN)

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \varphi(X_i) = \int \varphi(y)\pi(dy) \quad a.s. \, \forall x \in \mathcal{X}. \varphi \in L^1(\pi)$$

A primer on Markov chains: geometric ergodicity

- Plain ergodicity is not sufficient here: we want chain to converge fast!
- Recall **geometric ergodicity:** there exist a $\lambda \in (0,1)$ and V s.t.

$$\|K^n(x_0,\cdot) - \pi\| \leq MV(x_0)\lambda^n$$

- This can be shown by requiring either
  - For some small set C, there exist $M < \infty$ and $\kappa > 1$

$$\sup_{x \in C} \mathbb{E}_x\left[\kappa^{\tau_C}\right] < M$$

  - or Foster-Lyapunov drift condition holds: there exists a $V \geq 1$ with $V(x') < \infty$ for some x' s.t.

$$KV(x) \leq (1-\beta)V(x) + b1_{x \in C} \quad x \in \mathcal{X}$$

- (Geometric ergodicity holds for all $x_0 \in \{x : \; V(x) < \infty\}$)

Back to MCMC

- There are also drift conditions like above for polynomial rates
- Showing geometric ergodicity typically requires finding a V
  - typical candidate π^{-p}, $p \in (0,1)$
- Many popular algorithms fail to be geometrically ergodic
  - either due to structure of π or poor design of Q
  - could be expressed via return times to sets of support of π.
  - Example: long excursions in the tails, or certain points to where the transition kernel sticks
- Metropolis Hastings
  - is rarely uniformly ergodic for unbounded state spaces.
  - is geometrically ergodic if and only if the tails π are bounded by $a \exp(-b|x|)$ for positive a and b.

Quiz

- Let $\pi(x) = \exp(-x)$ and $q(x) = k\exp(-kx)$. Consider two cases: $k = 0.01$ and $k = 5$ and implement an independence sampler. Which case is bettter and why? It turns out one case is uniformly ergodic and another not geometrically ergodic. Which is which?

Measuring efficiency: CLT

- $v(\varphi, K)$ is the CLT variance

$$v(\varphi, K) = \mathbb{V}ar_{\pi}[\varphi] + 2\sum_{i\geq 1} \mathbb{C}ov[\varphi(X_0), \varphi(X_i)]$$

- If K is reversible spectral methods are applicable:
  - Kipnis and Varadhan (1986).
  - Let $\varphi \in L^2(\pi)$ and $\pi(\varphi) = 0$.
  - if $v(\varphi, K) = \lim_{n\to\infty} \frac{1}{n}\mathbb{V}ar_K\left[\sum_{i=1}^{n} \varphi(X_i)\right] < \infty$ then CLT holds
- If K is reversible and geometrically ergodic one can show the same CLT for all $\varphi \in L^2$
- There are extensions for non-reversible case: Toth 86

Measuring efficiency: CLT

- CLT variance was used to define
  - the **integrated auto-correlation time** for φ

$$\tau_\varphi = \frac{v(\varphi, P)}{\mathbb{V}ar_\pi[\varphi]} = 1 + 2\sum_{i\geq 1} Cor[\varphi(X_0), \varphi(X_i)]$$

  - or effective sample size

$$ESS = \frac{N}{\tau_\varphi}$$

- Also useful for ordering different MCMC algorithms
  - Low $v(\varphi, P)$ means also higher efficiency asymptotically **(Peskun-Tierney ordering)**

Measuring efficiency: expected square jumping distance

- One diagnostic is expected square jumping distance. Use samples to approximate

$$ESJD = E\left[(X_n - X_{n-1})^2\right]$$

  i.e. just look at first order correlation and linear test functions
- *ESJD* looks like a diffusion quadratic variation
- Is there a link with continuous time MCMC and accept reject schemes such as MH?

Diffusions and rescaling

- Consider
$$dX_t = \frac{1}{2}\Sigma\nabla \log \pi(X_t)dt + \Sigma^{1/2}dB_t$$
- Σ can be viewed as a speed-up function for the time scale
- (Roberts & Rosenthal 12) If we have K_1 and K_2 with Σ_1 and Σ_2 resp. and $\Sigma_1 \leq \Sigma_2$ then
$$v(\varphi, K_1) \geq v(\varphi, K_2)$$

  i.e. the faster the scale better!

Diffusive limits for MH

- Why is all this relevant?
- Let $x = (x^1, \ldots, x^d)$ and allow d to grow.
- Consider the target

$$\pi = \prod_{i=1}^{d} f(x^i)$$

- Let $(X_n;\ n \geq 0)$ be a MH output with $Q(x, \cdot) = \mathcal{N}(x, \frac{\varrho^2}{d} I)$ initialised at $\nu = \pi$
- Then look at the process

$$Z_t = X^1_{[td]}$$

Diffusive limits for MH

- (Roberts, Gelman & Gillks 97) At the limit Z_t with d obeys

$$dZ_t = h(\varrho)\nabla \log f(Z_t)dt + h(\varrho)^{1/2}dB_t$$

with

$$h(\varrho) = \varrho^2 2\Phi(-\frac{\varrho I^{\frac{1}{2}}}{2}) = \varrho^2\alpha(\varrho) = \frac{4}{I}\Phi^{-1}\left(\alpha(\varrho)\right)^2\alpha(\varrho)$$

with $\alpha(\varrho)$ being the limiting acceptance rate and $I = \mathbb{E}_f\left[\nabla \log f(X)^2\right]$.

The scaling problem for Metropolis chains

- Higher speed is better in terms of Peskun ordering so numerical maximisation gives universal constants

$$\alpha(\varrho) = 0.234 \quad \varrho = 0.488$$

- Practioners have realised range of these numbers much quicker!
- This is a very elegant theory and can be applied to many different contexts leading to justification of desired numbers for acceptance ratio
  - see work of Roberts, Rosenthal, Beskos, Breyer, Neal, Sherlock, Bedard, Thiery, Stuart, Pillai
- Similar diffusive limits appeared earlier by Gelfand & Mitter in 91 JOTA paper.

Discussion

- This is just an introduction, many more topics are very useful and important
  - mixing, coupling, splitting, Wasserstein distances, functional inequalities,...
  - minorisation can be restrictive tool
- Not all MCMC algorithms are guaranteed to have good convergence properties
  - this will depend on method used and ingredients
  - for MH: π, Q that construct K
- Understanding from theory often
  - comes later than intuition from observing behaviour in practice
  - and with many conditions...