MCMC, part (III)

From before: Aim is to "sample" (or, e.g., compute integrals) w.r.t. some target measure π on S.

$\Rightarrow$ Create a Markov chain $\{X_n\}_{n\geq 0}$, or process $\{X_t\}_{t\geq 0}$, with π as stationary distribution.

Q: How to analyse convergence?

From before: Let K be the kernel associated w. $\{X_n\}_{n\geq 0}$.

$$P(X_{n+1}\in A \mid X_n = x) = K(x,A)$$

$$K^n(x,A) = \int K^{n-1}(y,A)\,K(x,dy), \quad n\geq 2$$

Set for $k\in\mathbb{N}$, $x\in S$

$$d_k(x) = \sup_{A\in\mathcal{B}(S)} |\mathbb{P}_x(X_k\in A) - \pi(A)| =: \|K^k(x,\cdot) - \pi\|_{TV}$$

Uniform ergodicity:

$$K^m(x,A) \geq \varepsilon\nu(A), \text{ for all } x\in S,\ A\in\mathcal{B}(S),$$

some $\varepsilon\in(0,1)$ and $\nu\in\mathcal{P}(S)$.

Exponential/geometric ergodicity: $\exists\ 0<\gamma<1$ and

$h: S \to \mathbb{R}$ s.t. $d_k(x) \leq h(x) g^k$.

Other common quantities:

Mixing time: (here, w.r.t. TV):

$$\tau(\varepsilon) := \min\{n : \forall x \in S, \|K^n(x,\cdot) - \pi\|_{TV} \leq \varepsilon\}$$

(can replace TV with other "distances", e.g. relative entropy).

Spectral gap: Smallest eigenvalue of

$$I - \frac{1}{2}(K + K^*)$$

If K reversible w.r.t. π, then K self-adjoint

$\Rightarrow$ smallest eigenvalue of $I - K$.

Alt: difference between two largest eigenvalues (in abs. value): $1 - \max_\lambda |\lambda|$.

Large spectral gap $\Rightarrow$ faster convergence.

Now, consider π on the form

$$\pi(dx) \propto e^{-U(x)} dx, \text{ some } U.$$

Optimization: $\min_{x \in \mathcal{X}} U(x)$. Can be "solved" by samples from π.

Common choice: Langevin SDE

$$dX_t = -\nabla U(X_t)\,dt + \sqrt{2}\,dB_t, \quad \{B_t\} \text{ d-dim. BM.}$$

[Exercise: Show that π is invariant for $\{X_t\}$].

One way to analyse convergence properties: via functional analytical methods.

<u>For now</u>: $\{X_t\}$ generic Markov process on $\mathbb{R}^d$, with transition function / kernel: $P(t,\cdot,\cdot)$ $(P_t(\cdot,\cdot))$

Take $f \in C_0^2(\mathbb{R}^d)$ and define

$$(**) \quad (P_t f)(x) := \mathbb{E}[f(X_t) \mid X_0 = x] = \int_{\mathbb{R}^d} f(y)\,P(t,x,dy)$$

<u>Note</u>: Linear operator, $(P_0 f)(x) = f(x)$, $P_0 = Id$.

Moreover:

$$(P_{t+s} f)(x) = \int_{\mathbb{R}^d} f(y)\,P(t+s,x,dy)$$

$$= \int_{\mathbb{R}^d} \overbrace{\int_{\mathbb{R}^d} f(y)\,P(s,z,dy)}^{(P_s f)(z)}\,P(t,x,dz)$$

$$= \int_{\mathbb{R}^d} (P_s f)(z)\,P(t,x,dz) = (P_t \circ P_s)(f)(x)$$

$\therefore$ $P_{t+s} = P_t \circ P_s$, i.e. $(**)$ defines a <u>Markov semigroup</u>.

Can study properties of $\{X_t\}$ by studying $\{P_t\}_{t\geq 0}$.

<u>Invariance</u>: $\mu \in \mathcal{P}(\mathbb{R}^d)$ is an invariant measure for $\{P_t\}_t$.

if for every bounded, meas. $f : \mathbb{R}^d \to \mathbb{R}$,

and every $t \geq 0$,

$$\boxed{\int (P_t f)\, d\mu = \int f\, d\mu.}$$

$\int (P_t f) g\, dx = \int f g\, dx$

The associated (infinitesimal) generator, $\mathcal{L}$, is defined

by
$$\mathcal{L}f = \lim_{t \downarrow 0} \frac{P_t f - f}{t} \qquad (+)$$

for $f \in D(\mathcal{L})$: the set of functions in, e.g., $L^2(\mu)$, for which $(+)$ is well-defined.

[General theory: Hille-Yosida theory, e.g. Ethier & Kurtz]

Clarification: Formula given in class requires additional assumptions. Instead, we can use that if μ is invt, then for $f \in L^1(\mu)$,

$\int \mathcal{L}f\, d\mu = 0$. Conversely, for any class of function $\mathcal{F}$ dense in $L^1(\mu)$

if $\int \mathcal{L}f\, d\mu = 0, \forall f \in \mathcal{F}$, then μ is invariant for $P = \{P_t\}_{t\geq 0}$.

Formally: $P_t = e^{t\mathcal{L}}$

Define $u(x,t) = (P_t f)(x)$ some observable f.

$$\frac{\partial}{\partial t} u(x,t) = \frac{\partial}{\partial t}\left[(P_t f)(x)\right] = \frac{\partial}{\partial t}\left[(e^{t\mathcal{L}} f)(x)\right]$$

$$= \mathcal{L}\,\underbrace{(e^{t\mathcal{L}} f)(x)}_{(P_t f)(x) = u(x,t)} = \mathcal{L} u(x,t)$$

Backward Kolmogorov eq : $\begin{cases} \frac{\partial}{\partial t} u(x,t) = \mathcal{L} u(x,t) \\ u(x,0) = f(x) \end{cases}$

Governs evolution of expectation of an observable f.

<u>Adjoint semigroup</u> : P_t acts on functions.

Can define an adjoint operator, P_t^*, that acts on measures:

$A \in \mathcal{B}(\mathbb{R}^d)$, $\mu \in \mathcal{P}(\mathbb{R}^d)$

$$(P_t^* \mu)(A) = \int_{\mathbb{R}^d} \mathbb{P}(X_t \in A \mid X_0 = x)\, \mu(dx)$$

$$= \int_{\mathbb{R}^d} P(t, x, A)\, d\mu(x)$$

$(P_t^* \mu)$: a probability measure. Formally adjoint of P_t in an L^2-sense :

$$\int_{\mathbb{R}^d} (P_t f(x))\, d\mu(x) = \int_{\mathbb{R}^d} f \, d(P_t^* \mu)(x)$$

Let $\mathcal{L}^*$ be the formal adjoint of $\mathcal{L}$ in an L^2-sense:

$$\int_{\mathbb{R}^d} (\mathcal{L}f)h \, d\mu(x) = \int f(\mathcal{L}^* h) \, d\mu(x)$$

Can write $P_t^* = e^{t\mathcal{L}^*}$

If $\{X_t\}$ is a Markov process w. generator $\mathcal{L}$, $X_0 \sim \mu$,
then $\mu_t = P_t^* \mu$ is the law of X_t.

Arguments similar to derivation of ~~Backward~~ Kolmogorov:
If μ_t, μ have densities ρ_t, ρ w.r.t. Lebesgue,

then $\begin{cases} \frac{\partial}{\partial t} \rho_t(x) = \mathcal{L}^* \rho_t(x) \\ \rho_0(x) = \rho(x) \end{cases}$ Forward Kolmogorov eq. (Fokker-Planck).

Invariance: If μ invariant for $P = \{P_t\}_{t \geq 0}$,
then $P_t^* \mu = \mu$; $\Rightarrow$ stationary Fokker-Planck:

$$\boxed{\mathcal{L}^* \rho(x) = 0}$$

Carré du champ operator: Take $f, g \in D(\mathcal{L})$

Define: $\Gamma(f,g) = \frac{1}{2} [\mathcal{L}(fg) - f\mathcal{L}g - g\mathcal{L}f]$

Notation: $\Gamma(f) = \Gamma(f,f)$

Example: If $\mathcal{L} = \Delta$ [Q: What process?]

then $\Gamma(f,g) = \nabla f \cdot \nabla g$

Dirichlet form: $\mathcal{E}(f,g) = \int \Gamma(f,g)\, d\mu$

Notation: $\mathcal{E}(f) = \mathcal{E}(f,f)$.

Def μ satisfies a Poincaré inequality (PI) w. constant $C_{PI} > 0$ if, for all suitably regular f,

$$\boxed{\underbrace{\mathrm{Var}_\mu(f)}_{\int \left(f - \left[\int f\, d\mu\right]^2\right)^2 d\mu} \leq C_{PI}\, \mathcal{E}(f)}$$

Remark: An alternative def. of the spectral gap λ^* is:

$$\lambda^* = \min_f \left\{ \frac{\mathcal{E}(f)}{\mathrm{Var}_\mu(f)} : \mathrm{Var}_\mu(f) \neq 0 \right\}.$$

From (PI): $\lambda^* \geq \frac{1}{C_{PI}}$. Actually, for any non-zero eigenvalue λ of $-\mathcal{L}$, $\lambda \geq \frac{1}{C_{PI}}$.

Back to Langevin diffusion:

$$(*) \quad dX_t = -\nabla U(X_t)\, dt + \sqrt{2}\, dB_t$$

Here, $\mathcal{L}f = -\nabla U \cdot \nabla f + \Delta f$

Adjoint: $\mathcal{L}^* f = \nabla \cdot (f \nabla U) + \Delta f$ [Exercise: Derive this]

$\Rightarrow$ Forward Kolmogorov:

$$\boxed{\partial_t \rho_t = \nabla_x \cdot (\rho_t \nabla_x U) + \Delta_x \rho_t}$$

<u>Exercise</u>: Using the different characterisations of invariance, show $\pi \propto e^{-U}$ is invariant for $\{X_t\}$.

Suppose π satisfies a PI, with constant C.

Consider χ^2-divergence of $\pi_t = \text{Law}(X_t)$, w.r.t. π,

$$\chi^2(\pi_t | \pi) = \int \left(\frac{d\pi_t}{d\pi} - 1\right)^2 d\pi$$

<u>Aim</u>: Use (PI) to get an upper bound on $\chi^2(\pi_t | \pi)$.

For this $\mathcal{L}$ [<u>Exercise</u>: Check this]

$$\mathcal{E}(f) = \mathbb{E}_\pi \left[\|\nabla f\|^2 \right]$$

$$\frac{\partial}{\partial t} \chi^2(\pi_t | \pi) = \frac{\partial}{\partial t} \int \left(\frac{d\pi_t}{d\pi} - 1\right)^2 d\pi = \frac{\partial}{\partial t} \int \left(\frac{d\pi_t}{d\pi}\right)^2 d\pi$$

$$= \frac{\partial}{\partial t} \int \left(P_t \frac{d\pi_0}{d\pi}\right)^2 d\pi$$

(*) Correction: Use the def. of P_t and conditional expectation

$$= 2\int \left(P_t \frac{d\pi_0}{d\pi}\right) \cdot \frac{\partial}{\partial t}\left(P_t \frac{d\pi_0}{d\pi}\right) d\pi$$

{Exercise:? show by def. of $\mathcal{L}$} $$= 2\int \left(P_t \frac{d\pi_0}{d\pi}\right) \mathcal{L} \left(P_t \frac{d\pi_0}{d\pi}\right) d\pi$$

Now, ~~take~~ $f = \left(P_t \frac{d\pi_0}{d\pi}\right) = \left(\frac{d\pi_t}{d\pi}\right)$

then above is then

(*) Addition: For reversible processes (symmetric semigroups), $\mathcal{E}(f,g) = -\int f \mathcal{L} g \, d\mu$

$$\frac{\partial}{\partial t} \chi^2(\pi_t | \pi) \overset{*}{=} -2\,\mathcal{E}\left(\frac{d\pi_t}{d\pi}\right)$$

Moreover, note:

$$\mathrm{Var}_\pi\left(\frac{d\pi_t}{d\pi}\right) = \int \left[\frac{d\pi_t}{d\pi} - \left(\int \left(\frac{d\pi_t}{d\pi}\right) d\pi\right)^2\right]^2 d\pi$$

$$= \int \left(\frac{d\pi_t}{d\pi} - 1\right)^2 d\pi = \chi^2(\pi_t | \pi)$$

The PI for π then states

$$\chi^2(\pi_t|\pi) \leq C\,\mathcal{E}\left(\frac{d\pi_t}{d\pi}\right)$$

$$\Rightarrow \frac{\partial}{\partial t} \chi^2(\pi_t|\pi) = -2\,\mathcal{E}\left(\frac{d\pi_t}{d\pi}\right) \leq -\frac{2}{C}\chi^2(\pi_t|\pi)$$

Gronwall's: $0 \leq \chi^2(\pi_t|\pi) \leq \chi^2(\pi_0|\pi)\, e^{\int_0^t \left(-\frac{2}{C}\right) ds} = \chi^2(\pi_0|\pi)\, e^{\frac{2t}{C}}$

$$0 \leq \chi^2(\pi_t|\pi) \leq \chi^2(\pi_0|\pi)e^{-\frac{2t}{C}}$$

$\therefore \chi^2(\pi_t|\pi) \to 0$ as $t \to \infty$.

Requires that $\pi \propto e^{-U}$ satisfies a PI.

Requires assumptions on U.

Ex: If $\mathrm{Hess}(U) \geq \lambda I$, some λ, then π satisfies a PI with $C = \frac{1}{\lambda}$.

Implementation-wise: $dX_t = -\nabla U(X_t)dt + \sqrt{2}\,dB_t$.

Vanilla-implementation: EM-scheme

$$X_{k+1} = X_k - \gamma_k \nabla U(X_k) + \sqrt{2\gamma_{k+1}}\, Z_{k+1}$$

$\{Z_k\}$: iid $N(0,1)$, $\{\gamma_k\}$: step-size sequence.

$\gamma_k \equiv \gamma$: $\{X_k\}$ is irreducible, pos. recurrent $\Rightarrow$ unique invariant dist π_γ.

However: $\pi_\gamma \neq \pi$.

$\Rightarrow$ Metropolis-adjusted Langevin (MALA):

Add a accept-reject step to remove bias:

- Propose $Y_{k+1} \sim X_k - \gamma \nabla U(X_k) + \sqrt{2\gamma}\, Z_{k+1}$

- Acceptance ratio: $\alpha_\gamma(x,y) = \min\left\{1, \dfrac{\pi(y)\, r_\gamma(y,x)}{\pi(x)\, r_\gamma(x,y)}\right\}$

  where $r_\gamma(x,y) \propto e^{-\|y-x-\gamma\nabla U(x)\|^2 \cdot \frac{1}{4\gamma}}$

- Accept according to $\alpha_\gamma(X_k, Y_{k+1})$

$\Rightarrow$ Markov chain w. correct invariant dist. π.

- Ethier & Kurtz
- Bakry - Gentil - Ledoux ('14)