Contractivity of Markov Processes

Sam Power

University of Bristol

4 April, 2023

Contractivity of Markov Processes

Lecture 2

Sam Power

University of Bristol

4 April, 2023

Basic Examples

- Many idealised Markov chains satisfy curvature conditions with explicit constants.
- Here, I will catalog some examples which are relatively transparent.
- The choice of examples is guided by our shared interests in Monte Carlo methods with applications to statistical computation.

Advice for ‘Curvature Hunters’

1. On your first try, make generous assumptions.
2. For calculations, it can help to work with continuous-time processes.
3. Try to reduce to studying a lower-dimensional problem.
4. For harder problems, be flexible about the metric.
   - This will be the focus of Section 3.

Example 1: Curvature in the Discrete Metric

- Recall the discrete metric on E, given by $\mathsf{d}(x, y) = \mathrm{I}[x \neq y]$.
- In this context, one can identify

$$\mathcal{T}_{1,\mathsf{d}}(P(x, \cdot), P(y, \cdot)) = \mathrm{TV}(P(x, \cdot), P(y, \cdot)),$$

  and the contractivity condition reads as

$$x \neq y \implies \mathrm{TV}(P(x, \cdot), P(y, \cdot)) \leqslant 1 - \kappa$$

- Consequently, there exists a probability measure $\nu \in \mathcal{P}(E)$ such that

$$x \in E \implies \mathrm{TV}(P(x, \cdot), \nu) \leqslant 1 - \kappa.$$

- c.f. ‘Doeblin condition’, ‘(uniform) (one-step) minorisation’

Example 1: Application to IMH (1)

- Let π, $q \in \mathcal{P}(E)$ such that $M = \|w\|_\infty < \infty$, where $w := \frac{\mathrm{d}\pi}{\mathrm{d}q}$.
- Define a π-invariant Markov kernel P as the Independent Metropolis-Hastings kernel with proposal q, i.e.

$$P(x, \mathrm{d}z) = q(\mathrm{d}z) \cdot \alpha(x, z) + (1 - \alpha(x)) \cdot \delta(x, \mathrm{d}z),$$

with

$$\alpha(x, z) = \min\left\{1, \frac{w(z)}{w(x)}\right\}$$

$$\alpha(x) = \int q(\mathrm{d}z) \cdot \alpha(x, z).$$

Example 1: Application to IMH (2)

- Note that since $w(x) \leqslant M$ almost surely, we can always bound $w(x)^{-1} \geqslant M^{-1}$.
- We can then write

$$\alpha(x, z) = \min\left\{1, \frac{w(z)}{w(x)}\right\} \geqslant \min\left\{1, M^{-1} \cdot w(z)\right\} = M^{-1} \cdot w(z),$$

  and hence

$$P(x, \mathrm{d}z) \geqslant q(\mathrm{d}z) \cdot \alpha(x, z) \geqslant q(\mathrm{d}z) \cdot M^{-1} \cdot w(z) \geqslant M^{-1} \cdot \pi(\mathrm{d}z).$$

- Thus, the IMH has curvature $\geqslant M^{-1}$ with respect to the discrete metric.
- Approach generalises to i-SIR, i-cSMC, and related algorithms.

Example 2: Curvature in the Euclidean Metric

- Recall the Euclidean metric on $E = \mathrm{R}^d$, given by $\mathrm{d}\left(x, y\right) = \|x - y\|_2$.
- I will give some examples in both continuous and discrete time.
- In the Euclidean case, sometimes you can even have *almost-sure* contractivity, rather than just 'on average'.
  - Not always realistic, but can be insightful to study.
  - Often easier to prove these results.
  - Consequences are often much stronger than 'simple' curvature.

Example 2: Almost-Sure Contraction of Diffusions

- Consider the overdamped Langevin diffusion, as applied to sampling from an m-strongly log-concave target $\pi(\mathrm{d}x) = \exp(-V(x))\,\mathrm{d}x$, whose trajectories follow the stochastic differential equation

$$\mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t.$$

- Couple two copies of the diffusion by simply feeding the same driving Brownian motion W to each path:

$$\begin{aligned}
\mathrm{d}(X_t - Y_t) &= -(\nabla V(X_t) - \nabla V(Y_t))\,\mathrm{d}t \\
\mathrm{d}\|X_t - Y_t\|_2^2 &= -2 \cdot \langle X_t - Y_t, \nabla V(X_t) - \nabla V(Y_t)\rangle\,\mathrm{d}t \\
&\leqslant -2 \cdot m \cdot \|X_t - Y_t\|_2^2\,\mathrm{d}t.
\end{aligned}$$

Example 2: Almost-Sure Contraction of Diffusions

- By Grönwall's inequality, holds almost-surely that

$$\| X_t - Y_t \|_2^2 \leqslant \exp \left(-2 \cdot m \cdot t \right) \cdot \| X_0 - Y_0 \|_2^2 \,,$$

  i.e. curvature $\geqslant m$.
- Can also study $\mathrm{d} X_t = b \left(X_t \right) \, \mathrm{d} t + \sqrt{2} \, \mathrm{d} W_t$ with right conditions on b.
- Can also study hypoelliptic diffusions; typically need some 'twist' of the metric.

Example 2: Contraction of IRFS

- ‘Iterated Random Function Systems’
- At each step, draw a random function from some distribution over mappings $E \to E$, and then apply it to your current state, i.e.

$$f_n \sim F, \quad X_n = f_n(X_{n-1}).$$

  - c.f. random dynamical systems, stochastic flows, ...
- Assume that F is supported on functions which are almost-surely Lipschitz with respect to d.
  - If $\mathrm{Lip}_{E,E}(f) \leqslant 1 - \kappa$ almost surely, quite simple.
  - If $\mathrm{E}_F\left[\mathrm{Lip}_{E,E}(f)\right] < 1$, still reasonably simple.
  - Actually, $\mathrm{E}_F\left[\log \mathrm{Lip}_{E,E}(f)\right] < 0$ is sufficient for exponential ergodicity.

Example 2: Contraction of IRFS (ULA)

- Unadjusted Langevin Algorithm: draw $\xi \sim \mathcal{N}(0, I_d)$, set

$$f = f_\xi : x \mapsto x - h\nabla V(x) + \sqrt{2h}\xi.$$

- Write

$$\begin{aligned} f_\xi(x) - f_\xi(y) &= \left(x - h\nabla V(x) + \sqrt{2h}\xi\right) - \left(y - h\nabla V(y) + \sqrt{2h}\xi\right) \\ &= \int_0^1 \left(I - h\nabla^2 V(t \cdot x + (1-t) \cdot y)\right)(x - y) \end{aligned}$$

- Assume now that V is m-strongly convex and L-smooth, so that $\mathrm{eigs}(\nabla^2 V) \in [m, L]$; write $q = m \cdot L^{-1} \in [0, 1]$.

Example 2: Contraction of IRFS (ULA)

- It then holds that

$$\begin{aligned}\|f_\xi(x) - f_\xi(y)\|_2 &\leqslant \int_0^1 \left\|I - h\nabla^2 V(t\cdot x + (1-t)\cdot y)\right\|_{\mathrm{op}} \cdot \|x-y\|_2 \\ &\leqslant \max\{|1 - h\cdot\lambda| : m \leqslant \lambda \leqslant L\} \cdot \|x-y\|_2 .\end{aligned}$$

- Taking $h = \frac{2}{L+m}$ yields that $\mathrm{Lip}_{E,E}(f_\xi) \leqslant \frac{1-q}{1+q}$ a.s., so that $\kappa \geqslant \frac{2\cdot q}{1+q} \sim 2\cdot q$.
- Exercise: Generalise this analysis to MYULA.
- Exercise: Generalise this analysis to SGLD.

Example 2: Some Other Euclidean Contractions

- Ideal Hamiltonian Monte Carlo (on e.g. well-conditioned potentials)
- Generalised Doubling Maps (deterministic dynamics, backwards in time)
- Simple Slice Sampling (reduction to spherical symmetry; convexity)
- Mean-Field Diffusions (convex confinement, convex interaction)

Example 3: Curvature in the Hamming Metric

- Recall the Hamming metric on $E = \{\pm 1\}^d$, given by $\mathrm{d}(x, y) = \sum_{i \in [d]} \mathrm{I}[x_i \neq y_i]$.
- Relevant to study of graphical models, spin systems.
- Algorithm: (Poissonised) Random-Scan Gibbs Sampler
  1. For each $i \in [d]$, have a rate-one Poisson clock.
  2. When clock i rings, replace x_i by a draw from its full conditional distribution.
- Intuition: under weak dependence, this should mix well.

Example 3: Dobrushin's Criterion

- Let A be a finite alphabet, let V be a finite index set, and let π be a probability measure on A^V. Write

$$\pi_i(\cdot \mid x) := \mathrm{Law}_\pi\left(X_i \mid X_j = x_j \text{ for } j \in V \setminus \{i\}\right)$$

$$c_{i,j} := \sup\{\mathrm{TV}(\pi_i(\cdot \mid x), \pi_i(\cdot \mid y)) : x = y \text{ off } j\}.$$

- If the spectral radius of C satisfies $\rho(C) < 1$, then $\mathrm{pRSGS}(\pi)$ is exponentially ergodic with rate $\lambda \leqslant \rho(C)$.
- (usually stated in terms of a particular estimate on $\rho(C)$)

Example 3: Dobrushin-Wu Criterion

- Is the discreteness of the spins relevant? **No!**
- Let (E, d) be a metric space, let V be a finite index set, and let π be a probability measure on E^V. Write

$$c_{i,j} := \sup \left\{ \frac{\mathcal{T}_{1,\mathsf{d}}\left(\pi_i\left(\cdot \mid x\right), \pi_i\left(\cdot \mid y\right)\right)}{\mathsf{d}\left(x_j, y_j\right)} : x = y \text{ off j} \right\}$$

- If the spectral radius of C satisfies $\rho\left(C\right) < 1$, then $\mathrm{pRSGS}(\pi)$ is exponentially ergodic with rate $\lambda \leqslant \rho\left(C\right)$.

Example 3: Hybrid Metric Perspective

- These results can be viewed as ‘stitching together several near-contractions’.
- The conditions of the Theorem imply that if we define

$$\mathrm{Lip}_i\left(f\right) := \sup\left\{\frac{\left|f\left(x\right)-f\left(y\right)\right|}{d_i\left(x,y\right)} : x = y \text{ off } i\right\},$$

then it holds for $j \neq i$ that

$$\mathrm{Lip}_j\left(P^i f\right) \leqslant \mathrm{Lip}_j\left(f\right) + \mathrm{Lip}_i\left(f\right)\cdot c_{ij}.$$

- This implies (with some care) that

$$\mathrm{Lip}\left(P_t f\right) \preceq \mathrm{Lip}\left(f\right)\cdot \exp\left(-t\left(I-C\right)\right),$$

which is **stronger** than a single Lipschitz contraction.
- There exists a more general result involving this principle; see Section 3.

Example 3: Curie-Weiss Model

- Consider the Curie-Weiss Model on $\{\pm 1\}^N$ with law

$$\pi\left(x_1, \cdots, x_N\right) \propto \exp\left(h \cdot \sum_{i \in [N]} x_i + \frac{c}{N} \cdot \sum_{i,j \in [N]} x_i x_j\right).$$

- The conditionals are then given by

$$\pi\left(x_i \mid x_{-i}\right) \propto \exp\left(x_i \cdot \left\{h + \frac{c}{N} \cdot \sum_{j \in [N] \setminus \{i\}} x_j\right\}\right),$$

and in particular,

$$\mathbb{E}\left[x_i \mid x_{-i}\right] = \tanh\left(h + \frac{c}{N} \cdot \sum_{j \in [N] \setminus \{i\}} x_j\right).$$

Example 3: Curie-Weiss Model

- As the conditional laws are binary, we bound TV by the difference in means:

$$\begin{aligned}\mathrm{TV}\left(\pi_i\left(\cdot \mid x\right), \pi_i\left(\cdot \mid y\right)\right) &\leqslant \frac{1}{2} \cdot \left|\tanh\left(h + c \cdot m_i(x)\right) - \tanh\left(h + c \cdot m_i(y)\right)\right| \\ &\leqslant \frac{c}{2} \cdot \left|m_i(x) - m_i(y)\right| \\ &\leqslant c \cdot N^{-1}\end{aligned}$$

  if $x = y$ off j.
- So, if $c < 1$, then $pRSGS(\pi)$ will have a spectral gap of at least $1 - c$.
- n.b. estimate holds *uniformly* in N.

Example 3: Some Other Spin Systems

- Can apply to other mean-field spin systems (e.g. SK); not necessarily well-suited
- Can apply to Ising models on sparse graphs (e.g. d-regular); often well-suited
- Can apply to Gaussian graphical models with ‘diagonally dominant’ precision
- Can apply to strongly log-concave, non-Gaussian graphical models

Basic Examples: Recap

- Some recurrent elements:
  1. Kernels have large probability overlap.
  2. Kernels have large spatial overlap.
- These examples haven't fought back too much: the natural metric was enough.