AN INTRO TO HYPERCONTRACTIVITY

LIAM HODGKINSON

ABSTRACT. The rate of convergence of an ergodic stochastic process to its stationary distribution is a valuable and well-studied problem for MCMC and Stein's method alike. In this talk, I will discuss the concept of hypercontractivity (essentially exponential convergence in relative entropy), and the many benefits it entails.

The rate of convergence of an ergodic stochastic process to its stationary distribution is perhaps one of the most well-studied and valued topics in the theory of stochastic processes. It is a classical result of Kolmogorov that processes of the form

$$\mathrm{d}X_t^M = \nabla \log \pi \left(X_t^M\right) \mathrm{d}t + \mathrm{d}W_t^M, \tag{1}$$

where W_t^M denotes Brownian motion on the Riemannian manifold M, comprise all time-reversible diffusion processes with stationary distribution $\pi \in \mathcal{C}^1$. These processes are significant for many reasons; in particular, I am interested in the use of these diffusion processes in Markov Chain Monte Carlo, as well as in Stein's method. In this talk, I will look at the rate of convergence of this stochastic process in the case that $M = \mathbb{R}^n$ (flat Euclidean space). Most of you would be somewhat familiar (or at least, not surprised by) the following result due to Gareth Roberts and Richard Tweedie:

Theorem 1 (EXPONENTIAL ERGODICITY). *Suppose that π is a light-tailed distribution on $\mathbb{R}^n$. Denoting by $p_t(x,y)$ the transition kernel of X_t, there exist constants $C, \lambda > 0$ such that*

$$\sup_x \int_{\mathbb{R}^n} |p_t(x,y) - \pi(y)| \,\mathrm{d}y \le Ce^{-\lambda t}, \qquad t \ge 0, \tag{2}$$

The measure of discrepancy here is in *total variation*, which stems from initial studies into the concept of geometric ergodicity involving discrete state spaces. In these cases, it is a perfectly adequate and completely natural measure of discrepancy between probability measures. However, for uncountable state spaces, the total variation norm is well-known to be rather unnatural to work with. In fact, while the above result on exponential ergodicity is highly regarded in the MCMC literature, I claim that it is a poor method of measuring rate of convergence. The biggest problems are obvious: C, λ are unknown, and virtually impossible to estimate in black-box applications.

Instead, I suggest shifting the discussion to the theory of *hypercontractivity*. Instead of the total variation metric as a measure of discrepancy, consider the relative entropy (or KL divergence):

$$H(\mu|\nu) = \int \log\left(\frac{\partial\mu}{\partial\nu}\right)\mathrm{d}\mu, \qquad H(p|q) = \int_{\mathbb{R}^n} p(x)\log\left(\frac{p(x)}{q(x)}\right)\mathrm{d}x,$$

and diffusions which satisfy

$$H\left(X_t|\pi\right) \leq H\left(X_0|\pi\right)e^{-\lambda t}, \qquad t \geq 0.$$

Although a surprisingly non-trivial fact, this is stronger than exponential ergodicity, that is, exponential convergence in relative entropy implies exponential ergodicity. However, as a functional concept, this provides us with relatively little information. Instead, I shall introduce an entropy operator corresponding to the Markov process. Recall that the transition semigroup of the Markov process is defined by

$$P_t f\left(x\right) = \mathbb{E}\left[f\left(X_t\right)|X_0 = x\right], \qquad t \geq 0.$$

We now define H_t for $t \geq 0$ by

$$H_t f = \mathbb{E}_\pi\left[P_t f \cdot \log P_t f\right] - \left(\mathbb{E}_\pi f\right)\log\left(\mathbb{E}_\pi f\right), \qquad f \in \mathcal{C}_0\left(\mathbb{R}^n\right),$$

where $\mathbb{E}_\pi f = \int_{\mathbb{R}^n} f\left(x\right)\pi\left(x\right)\mathrm{d}x$. Observe that if p_0 is the density of X_0, then $H_t\left(p_0/\pi\right) = H\left(X_t|\pi\right)$, so if these entropy operators diminish exponentially in time, we obtain our desired exponential convergence in relative entropy. This is precisely what hypercontractivity is.

Definition 2. A stochastic process X_t is said to be hypercontractive if $H_t f \leq e^{-\lambda t}H_0 f$ for some λ, called the hypercontractivity constant.

This definition differs from that seen in most of the literature on the subject, but can be shown to be equivalent, and I generally find this to be much easier to digest. Indeed, the original definition of hypercontractivity imposed that

$$\left\|P_t f\right\|_{L^q(\pi)} \leq \left\|f\right\|_{L^p(\pi)}, \qquad \text{whenever } e^{-\lambda t} \leq \frac{p-1}{q-1}.$$

This definition of hypercontractivity comes from Edward Nelson in 1966 in quantum field theory. Original motivations and work on the concept were intimate with harmonic analysis. However, it was the revolutionary paper of Leonard Gross in 1975 that really brought the concept to the forefront. Gross proved that X_t is hypercontractive if and only if the stationary distribution π satisfies

$$\int_{\mathbb{R}^n} f\left(x\right)^2 \log\left(\frac{f\left(x\right)^2}{\mathbb{E}_\pi f^2}\right)\pi\left(x\right)\mathrm{d}x \leq \frac{2}{\lambda}\int_{\mathbb{R}^n}\left\|\nabla f\left(x\right)\right\|^2 \pi\left(x\right)\mathrm{d}x,$$

called a *logarithmic Sobolev inequality* (here, I have written this in integral form to make the connection to analysis clearer). At this point, the Ornstein-Uhlenbeck process (so π is the standard Gaussian distribution) was known to be hypercontractive, which led to what is commonly referred to as THE logarithmic Sobolev inequality. There are a number of important consequences of this inequality:

- (**Probability**): Aside from the obvious value in developing general probabilistic estimates involving π, the log-Sobolev inequality implies the *Poincare inequality*, which states that

$$\mathrm{Var}_\pi f\left(X\right) \leq \frac{2}{\lambda}\mathbb{E}_\pi\left\|\nabla f\left(X\right)\right\|^2.$$

In many cases, it is significantly easier to compute or estimate expectations of derivatives, than those of the function itself. I have used the estimate of this form for bounded random variables on numerous occasions throughout my own work.

- (**Analysis and PDEs**): The origin of the name 'logarithmic Sobolev inequality' stems from the fact that it is a dimension-independent generalisation of the fundamental Sobolev inequalities in PDE theory. In particular, it implies the embedding
$$W^{1,2} \hookrightarrow L^2 \log L^2 (\pi) .$$
Therefore, the logarithmic Sobolev inequality is valuable in the study of non-linear partial differential equations.
- (**Geometry**): Perhaps the most significant of all of these applications is the work of Perelman. Yes, THAT Perelman. Fundamental to his proof of the Poincare conjecture is the construction of a $\mathcal{W}$ functional which behaves nicely under the Ricci flow. The functional itself is complicated; however, it is worthwhile to mention that the non-negativity (which provides an essential lower bound for the functional for convergence results) of $\mathcal{W}$ is precisely the log-Sobolev inequality. The connections between hypercontractivity and geometry go even deeper, but I shall not discuss them here.

Hopefully, it should be clear that hypercontractivity is a highly desirable property for an ergodic diffusion. But which of these processes are hypercontractive? The following necessary condition, which stemmed from initial correspondence between Gross and Herbst, provides a pretty clear picture.

- **Necessary:** The diffusion X_t is hypercontractive only if π is subgaussian, that is, $\mathbb{E}_\pi e^{\alpha X^2} < \infty$ for some $\alpha > 0$.

This is clearly a much stronger condition than light tails (subexponential). But you get more bang for your buck! More interesting are sufficient conditions, which stick pretty tight to the necessary condition. In fact, one of the benefits of dealing with hypercontractivity is that it is relatively straightforward to obtain the hypercontractivity constant λ. There is a two step process to this. The first is the Bakry-Emery condition:

- **Sufficient (Bakry-Emery):** The diffusion X_t is hypercontractive if $\text{Hess}V \succcurlyeq \lambda I$ ($\text{Hess}V - \lambda I$ is positive-definite) for some $\lambda > 0$ where $V = -\log \pi$ is the potential. In this case, λ is a hypercontractivity constant.

Basically, X_t is hypercontractive if π is a strongly log-concave distribution. We can weaken this substantially using Holley-Stroock perturbation.

- **Sufficient (Holley-Stroock):** Suppose that $\pi(x) = q(x) \exp(-V_0(x))$ where $\text{Hess}V_0 \succcurlyeq \lambda_0 I$ and $\lambda_1^{-1} \leq q(x) \leq \lambda_1$ for all $x \in \mathbb{R}^n$. Then X_t is hypercontractive with hypercontractivity constant λ_0/λ_1.

The Bakry-Emery condition possesses a rather involved proof, but Holley-Stroock perturbation is a simple consequence of the Radon-Nikodym theorem. Together, these conditions imply hypercontractivity for a large number of subgaussian distributions, and with explicit rates of convergence too!

Example 3. Let $\pi(x) \propto e^{-\frac{1}{2}(x+a)^2} + e^{-\frac{1}{2}(x-a)^2}$ be a mixture of two standard normal distributions with modes at $-a$ and a. This distribution is strongly log-concave whenever $a < 1$, and so the Bakry-Emery condition applies with

$$\lambda = 1 - a^2.$$

Remember that for $a = 0$, $\lambda = 1$ which coincides with the Gaussian case. To extend to larger a, we need to apply Holley-Stroock perturbation. The process is less clear, but by choosing $q(x) = \min\left\{\exp\left((1+a)^2 x^2\right), (1+a)^2\right\}$, we can achieve

$$\lambda \leq \frac{2}{1+a}.$$

I suspect this is about the best you can hope for. Therefore, the mixing times increase proportional to the width between modes. Performing a discretisation with Metropolis correction, this increase will get much worse.