ÉCOLE NORMALE SUPÉRIEURE DE LYON

ENS de Lyon
Università degli Studi Roma Tre

Internship 2021 (M1)
Department of Mathematics

On the role of Entropy in the cutoff phenomenon

Student :
Alexandre Bristiel

Internship tutor :
Pietro Caputo

December 29, 2021

Contents

1 Convergence of discrete Markov chains

General definition Let $G = (V, E)$ be a finite graph with weights $c_e \geq 0$ for $e \in E$. We can always assume the graph to be complete and put 0 weights on edges that didn't exist. We will at first assume the graph is directed. The *continuous time Markov chain* on G is a family of random variables $(X_t^x)_{t \in \mathbb{R}_+, x \in V}$ such that :

$$\mathbb{P}(X_t^x = y) = \exp(tL)(x, y)$$

Where L is the *discrete Laplacian* on G, also know as the *generator* of X. It is a linear operator defined by :

$$Lf(x) = \sum_{y \in V} c_{xy}(f(y) - f(x))$$

Since $\mathscr{P}_t = e^{tL}$ satisfies the heat equation :

$$\partial_t \mathscr{P}_t = L\mathscr{P}_t$$

then $\mathscr{P}_t$ is often called the heat kernel of the Markov chain X_t and L is know as the generator since it generates the semi-group $(\mathscr{P}_t)_{t \geq 0}$. On the other hand a semi-group of stochastic operators that is differentiable will always be written as $\exp(tL)$ for a certain discrete Laplacian L.

Define $q(x) = \sum_{y \neq x} c_{xy}$, q is the time spend at x. Markov chains are mostly studied only looking at their law and forgetting the underlying stochastic process. But to get an idea of what such a process is, X^x starts at the point x and stay there for a time $t \sim \mathcal{E}(q(x))$, then jumps to a neighboring vertex with law $c_{xy}/q(x)$. Then if it jumps at y it has the same law as X^y and the process is repeated.

Proposition 1.1. *X_t^x is well defined, that is to say for any positive $t \geq 0$ $\mathscr{P}_t$ is a stochastic matrix. On the other hand any semi-group of stochastic matrix differentiable at 0 will be written $\exp(tL)$ with $L = \partial_t(\mathscr{P}_t)_{t=0}$ a generator.*

Proof. Since $L\mathbf{1} = 0$ and L commutes with $\mathscr{P}_t$, we have $\partial_t(\mathscr{P}_t\mathbf{1}) = 0$, so $\mathscr{P}_t\mathbf{1} = \mathbf{1}$. By the semi-group propriety ($\mathscr{P}_{t+s} = \mathscr{P}_t\mathscr{P}_s$) we only need to prove that close to 0 $\mathscr{P}_t$ has positive entries.

Let f be a positive function and x be such that $f(x) = 0$, then :

$$\partial_t \mathscr{P}_t f(x)_{|t=0} = Lf(x) \geq 0$$

so at the entries of that are zero f $\mathscr{P}_t f$ increases, so locally $\mathscr{P}_t f$ is positive, and by the semi-group propriety for any positive time $\mathscr{P}_t f \geq 0$. In particular for $f = \delta_x$, $\mathscr{P}_t$ has positive entries.
On the other hand if $\mathscr{P}_t$ is a semi-group of differentiable operators since $\mathscr{P}_0 = \mathrm{Id}$ then for $x \neq y$ $\mathscr{P}_t(x, y)' \geq 0$. Finally by differentiating the stochastic condition:

$$L(x, x) = -\sum_{y \neq x} L(x, y)$$

□

In a more general setting a Markov chain is a family of linear operators on measurable functions that satisfy the semi-group propriety as well as leaving constant function invariant and being increasing. The definition given here can be transposed as is for discrete spaces and letting L act on at least on finitely supported functions.

Invariant law An invariant law π is a probability measure on V such that for any function $f : V \to \mathbb{R}$:

$$\mathbb{E}_\pi[Lf] = 0$$

By differentiating this is equivalent to :

$$\forall t \geq 0,\ \mathbb{E}_\pi[\mathscr{P}_t f] = \mathbb{E}_\pi[f]$$

and by taking a basis this is also equivalent to $\pi L = 0$ or $\pi \mathscr{P}_t = \pi$. In terms of stochastic process this means that picking a point following π and then following the process for any positive time will have the same law as π.

As soon as X is irreducible, that is to say for any positive time $\mathbb{P}(X_t^x = y) > 0$, then Perron-Frobenius's theorem assure the existence an uniqueness of such a invariant law.

Theorem 1 (Perron-Frobenius)**.** *Let A be a positive irreducible matrix.*
Then there exists a maximal eigenvalue $r > 0$ such that r is a simple eigenvalue and there exits a positive eigenvector for the eigenvalue r. Furthermore any other eigenvalue s is such that $|s| \leq r$.

Proof. Let x be a positive non-zero vector, then $Ax > 0$. Indeed we know that $Ax \geq 0$, but if i is such that $x_i \neq 0$ then for k such that $A^k_{i,i} \neq 0$ we have $A^k x \neq 0$, therefore $Ax \neq 0$.
We can now consider for $x \geq 0$ $f(x) = \max\{t > 0, tx \leq Ax\}$, we know that $f(x) > 0$ for $x \geq 0$ non zero. Let :

$$r = \sup_{\substack{\|x\| \leq 1 \\ x \geq 0, x \neq 0}} f(x) = \sup_{\substack{\|x\| = 1 \\ x \geq 0}} f(x)$$

f is upper semi-continuous therefore r is in fact a max. Indeed if $x \to x_0$ then we have :

$$f(x)x \leq Ax \implies \limsup_{x \to x_0} f(x)x_0 \leq Ax_0$$

therefore $\limsup_{x \to x_0} f(x) \leq f(x_0)$.

If $\|x\| = 1, x \geq 0$ is such that $f(x) = r$ then $Ax - rx \geq 0$, if we had $Ax - rx \neq 0$ then $AAx > rAx$ thus $f(Ax) > f(x)$ which is impossible, therefore x is a positive eigenvector for the value $r > 0$.
Assuming $x_i \neq 0$ let k_j be such that $A^{k_j}_{ij} > 0$ then we have $x_j = 1/r^{k_j}(A^{k_j}x)_j > 0$, thus $x > 0$.

Let y be an other eigenvector for the value r. Let t be the smallest such that $(tx - y) \geq 0$, in particular we have a least one coordinate i such that $tx_i - y_i = 0$. Assuming $tx - y \neq 0$ then $tx - y/\|tx - y\|$ meets the maximum of f that implies $tx - y > 0$ which is impossible, therefore $tx = y$ in other words $E_r(A)$ is of dimension 1.

Let $\lambda \neq r$ be another eigenvalue and y an normalized eigenvector. By the triangular inequality :

$$A|y| \geq |Ay| = |\lambda||y|$$

therefore $|\lambda| < r$. $\square$

In the following we will assume X is irreducible and π is the unique invariant law and strictly positive. A focal point in the study of Markov chains is the convergence of these to their invariant measure.

Total variation distance The privileged method of estimating the convergence of the chain is the total variation distance :

Definition 1.1 (Total variation distance)**.** The *total variation* distance of probability measures on a measurable space $(E, \mathcal{F})$ is defined as:

$$\|\mu - \nu\|_{TV} = \sup_{A \subset \mathcal{F}} |\mu(A) - \nu(A)|$$

In the discrete case we have a useful representation of this norm :

Proposition 1.2. *If $\mathcal{H}$ is at most a countable set then:*

$$\|\mu - \nu\|_{TV} = \frac{1}{2}\|\mu - \nu\|_1 = \frac{1}{2}\sum_{x \in \mathcal{H}} |\mu(x) - \nu(x)|$$

Proof. Let $A = \{x, \mu(x) \geq \nu(x)\}$, then :

$$|\mu(A) - \nu(A)| = \mu(A) - \nu(A) = \sum_{x \in A} |\mu(x) - \nu(x)|$$

but we also have :

$$|\mu(A) - \nu(A)| = \nu(A^c) - \mu(A^c)| = \sum_{x \in A^c} |\mu(x) - \nu(x)|$$

We now only need to prove that A meets the sup. Let $B \subset \mathcal{H}$, then :

$$\begin{aligned} |\mu(B) - \nu(B)| &\leq |\mu(B \cap A) - \nu(B \cap A)| - |\mu(B \cap A^c) - \nu(B \cap A^c)| \\ &= \mu(B \cap A) - \nu(B \cap A) + \mu(B \cap A^c) - \nu(B \cap A^c) \\ &= \mu(A) - \nu(A) = |\mu(A) - \nu(A)| \end{aligned}$$

This proves also that A is optimal. $\square$

This can be generalized to any measurable (σ-finite) space.

Proposition 1.3. *If $\nu \ll \mu$, let $h = \frac{\partial \nu}{\partial \mu}$ then :*

$$\|\mu - \nu\|_{TV} = \frac{1}{2}\|h - 1\|_{L^1(\mu)}$$

An other representation of the total variation distance can be convenient :

Proposition 1.4. *For any two laws μ, ν we have :*

$$\|\mu - \nu\|_{TV} = \inf_{\substack{\pi \in \Gamma \\ X,Y \sim \pi}} \mathbb{P}(X \neq Y)$$

where Γ is the set of product laws that have projection coinciding with μ and ν. And an explicit minimizing coupling exists.

The prove of both these propositions are quite involved and are detailed in the measure theory annex. An important corollary is the Markov contraction of the L^1 norm.

Corollary 1.1 (Markov contraction)**.** *If $\mathscr{P}_t$ is a Markov kernel then :*

$$\|\mu \mathscr{P}_t - \nu \mathscr{P}_t\|_{TV} \leq \|\mu - \nu\|_{TV}$$

for any initial law μ, ν.

Proof. Let π be a coupling of μ, ν and $X, Y \sim \pi$. Let X_t^x be a Markov chain resulting from $x \in V$. We define a coupling of $\mu \mathscr{P}_t$ and $\nu \mathscr{P}_t$ by taking:

$$X_t(\omega) = Z_t^{X(\omega)}(\omega), \quad Y_t(\omega) = Z_t^{Y(\omega)}(\omega)$$

Knowing Y, $X \sim \mu$ therefore $Z_t^X \sim \mu \mathscr{P}_t$, the same can be said for Y, we have thus defined a coupling of $\mu \mathscr{P}_t$ and $\nu \mathscr{P}_t$. But noticing :

$$X(\omega) = Y(\omega) \implies X_t(\omega) = Y_t(\omega)$$

we have :

$$\|\mu \mathscr{P}_t - \nu \mathscr{P}_t\|_{TV} \leq \mathbb{P}(X_t \neq Y_t) \leq \mathbb{P}(X \neq Y)$$

this is true for any coupling, proving the lemma. □

1.1 Poincare constant

L^2 structure of the invariant measure We can associate several objects to such a generator that will be useful to study the convergence of X towards equilibrium. Notably an euclidean structure on $L^2(\pi)$.

Definition 1.2 (Dirichlet form)**.** The *Dirichlet form* of L is :

$$\mathcal{E}(f, g) = -\langle Lf, g \rangle_\pi = -\mathbb{E}_\pi[fLg]$$

If L is auto-adjointed in $L^2(\pi)$ then $\mathcal{E}(f, g)$ defines the privileged euclidean structure on $L^2(\pi)$. In that case we say the Markov chain is *reversible*.

Definition 1.3 (Carre du champ)**.** The *carre du champ* operator of L is :

$$\Gamma(f, g) = \frac{1}{2}[L(fg) - fLg - gLf]$$

By rearranging we can see that :

$$\Gamma(f, g)(x) = \frac{1}{2} \sum_{y \neq x} c_{xy}(f(y) - f(x))(g(y) - g(x))$$

In particular $\Gamma(f, f) \geq 0$.

Spectral gap Finally let $L^\star$ by the adjoint of L in $L^2(\pi)$, that is to say :

$$L^\star(x,y) = \frac{\pi(y)L(y,x)}{\pi(x)}$$

The Poincarre constant of L or spectral gap, is λ the lowest non-zero eigenvalue of $-L^* = -(L^\star + L)/2$. Since this operator is auto-adjoint in $L^2(\pi)$ then this operator as real eigenvalues. Since L^* is auto adjoint:

$$\mathcal{E}(f,f) = -\langle f, Lf\rangle_\pi - \langle f, L^* f\rangle_\pi = \mathbb{E}_\pi[\Gamma^*(f,f)] \geq 0$$

Where Γ^* is the carre du champ of the symmetrized generator. So by definition of λ :

$$\mathcal{E}(f,f) \geq \lambda \mathrm{Var}_\pi[f]$$

By the minmax principle this is the best constant for such an inequality. This constant gives a great number of information on the structure of the underlying graph and also quantifies the convergence of the Markov chain. Indeed we have:

$$\|\mathscr{P}_t(x,.) - \pi\|_{TV} = \frac{1}{2}\|h-1\|_{1,\pi}$$

where $h(y) = \frac{\mathscr{P}_t(x,y)}{\pi(y)}$, by C.S :

$$\|h-1\|_{1,\pi} \leq \|h-1\|_{2,\pi} = \sqrt{\mathrm{Var}_\pi[h]}$$

Proposition 1.5 (Poincare inequality)**.** *We have for any positive time:*

$$\|\mathscr{P}_t(x,.) - \pi\|_{TV} \leq \frac{1}{2}e^{-\lambda t}\frac{1}{\pi(x)}$$

Proof. Since $\partial_t \mathrm{Var}_\pi[h] = -2\mathcal{E}(h,h)$ we have :

$$\partial_t \mathrm{Var}_\pi[h] \leq -2\lambda \mathrm{Var}_\pi[h]$$

so by Gronwall's lemma :

$$\mathrm{Var}_\pi[h] \leq e^{-2\lambda t}\mathrm{Var}_\pi[\delta_x/\pi] = e^{-2\lambda t}\big(1/\pi(x)^2 - 1\big)$$

□

This can be easily generalized to any starting law μ and with the bound $\mathrm{Var}_\pi(\partial\mu/\partial\pi) \leq \|\mu/\pi\|_\infty^2$ we have :

$$\|\mu\mathscr{P}_t - \pi\|_{TV} \leq \frac{1}{2}e^{-\lambda t}\left\|\frac{\mu}{\pi}\right\|_\infty$$

Finally we denote $t_{\mathrm{rell}} = 1/\lambda$ the relaxation time.

1.2 Modified logarithmic Sobolev constant

If λ quantifies the decay of the L^2 norm, ρ the modified logarithmic Sobolev constant quantifies the decay of entropy. More precisely the decay of relative entropy of Kullback Leibler divergence :

Definition 1.4 (Kullback Leibler divergence)**.** Let $\nu \ll \mu$ be two measures then:

$$d_{KL}(\nu\|\mu) = \mathbb{E}_\nu\left[\log\frac{\partial\nu}{\partial\mu}\right] = \mathbb{E}_\mu\left[\frac{\partial\nu}{\partial\mu}\log\frac{\partial\nu}{\partial\mu}\right]$$

The mutual entropy has several interesting proprieties notably it is positive and 0 if an only if $\mu = \nu$. Unfortunately it is not symmetric. The *Pinsker's inequality* relates the entropy to the L^1 norm :

Proposition 1.6 (Pinsker's inequality)**.** *Let $\nu \ll \mu$ be two measures then:*

$$\|\nu - \mu\|_{TV}^2 \leq \frac{1}{2}d_{KL}(\nu\|\mu)$$

So the decay of entropy controls the convergence of the chain. Just as $\partial_t \mathrm{Var}_\pi \mathscr{P}_t = -2\mathcal{E}$ a useful differential relation holds for entropy :

$$\partial_t d_{KL}(\mathscr{P}_t(x,.)\|\pi) = \mathcal{E}(h, \log h)$$

The *modified logarithmic Sobolev constant* is the best constant such that the exponential decay of entropy holds, be take infinitesimally we can define it as the biggest constant such that for any law $\nu = f d\pi$ we have :

$$\mathcal{E}(f, \log f) \geq \rho d_{KL}(\nu\|\pi)$$

Since both d_{KL} and $\mathcal{E}$ are homogeneous (if L is auto-adjoint), this relation can prolonged to any positive function in that case d_{KL} becomes :

$$\mathrm{Ent}_\pi[f] = \mathbb{E}_\pi\left[f\log\frac{f}{\mathbb{E}_\pi[f]}\right] = \mathbb{E}_\pi[f\log f] - \mathbb{E}_\pi[f]\log\mathbb{E}_\pi[f] \tag{1}$$

The convexity of $x\log x$ guaranties that $\mathrm{Ent}_\pi[f] \geq 0$ and $\mathrm{Ent}_\pi[f] = 0$ if and only if $f = \mathbb{E}_\pi[f]$, i.e. f is constant. By construction we have exponential decay of entropy at rate ρ :

Proposition 1.7 (Modified logarithmic Sobolev inequality)**.** *We have for any positive time :*

$$d_{KL}(\mathscr{P}_t(x,.)\|\pi) \leq e^{-t\rho}\log\frac{1}{\pi(x)}$$

Like in the Poincare inequality this can be generalized to any initial law μ :

$$d_{KL}(\mu\mathscr{P}_t\|\pi) \leq e^{-t\rho}d_{KL}(\mu\|\pi)$$

The Poincare constant can be defined as an eigenvalue of an operator so the usual tools of linear algebra can be used however no such representation is know for ρ which, among other reasons, makes it notoriously hard to study or even estimate. Also we can't symmetrize the operator since $\mathcal{E}(f, \log f) \neq \mathcal{E}^*(f, \log f)$ in general.

1.3 Mixing times and the cutoff phenomenon

Mixing times have a self explanatory definition :

Definition 1.5 (Mixing time)**.** For $1 > \varepsilon > 0$ the mixing time t_{mix} of a Markov chain with heat kernel $\mathscr{P}_t$ is :

$$t_{\text{mix}}(\varepsilon) = \inf \left\{ t | \max_{x \in V} \|\mathscr{P}_t(x,.) - \pi\|_{TV} \leq \varepsilon \right\}$$

Cutoff phenomenon An interesting problem is to precisely estimate these mixing times and specifically their dependence on the size underlying graph. The 'cutoff phenomenon' is the observation that $t_{\text{mix}}(\varepsilon)$ does not depend on ε as the size of the graph goes to infinity, this phenomenon has been cataloged in multiple example. It can be precisely defined as :

Definition 1.6. The sequence of Markov chains X_n with state space of size diverging, exhibits a cutoff if for any fixed ε we have :

$$t^n_{\text{mix}}(\varepsilon) \underset{n\to\infty}{\sim} t^n_{\text{mix}}(1-\varepsilon)$$

Remark. Often the dependence in n of the mixing times will be omitted, and we will for example write $t_{\text{mix}}(\varepsilon) \sim t_{\text{mix}}(1-\varepsilon)$.Often the dependence in n of the mixing times will be omitted, and we will for example write $t_{\text{mix}}(\varepsilon) \sim t_{\text{mix}}(1-\varepsilon)$.

Example. A prototypical example are product chain. If G is a graph with a Markov chain X the product chain on G^n is a chain where each component is independently updated one at the time. If t_{rell} is the relation time of the original graph G then the product chain X_n have mixing time :

$$t_{\text{mix}} \sim \frac{t_{\text{rell}}}{2} n \log n$$

for any ε.

Question This cutoff phenomenon has mostly been studied on chain using a great knowledge of the behavior the chain but very general criteria to characterize cutoff are unknown.
Recent works however seem to point out the crucial role of entropy in this phenomenon. A generic cutoff condition has been established under curvature condition in Salez [Sal21].

2 Entropy concentration

2.1 Entropy concentration implies cutoff

We will establish several very general bound on mixing times using the entropies at time t:

Definition 2.1. (Maximal entropy) The *maximal entropy* of a Markov chain is defined as :

$$d^\star(t) = \max_{x \in V} d_{KL}(\mathscr{P}_t(x,.)\|\pi) \tag{2}$$

Lemma 1 (Entropic upper-bound)**.** *For any $t \geq 0$ and $\varepsilon \in]0,1[$ we have :*

$$t_{mix}(\varepsilon) \leq t + \frac{t_{rell}}{\varepsilon}(1 + d^\star(t))$$

Proof. Take any law μ on V and consider the set :

$$A = \{x \in V, \ln \mu(x)/\pi(x) < 1 + 2d_{KL}(\mu\|\pi)/\varepsilon\}$$

The set of point that don't deviate from the mean of $\ln \mu/\pi$, $d_{KL}(\mu\|\pi)$, we expect A to have a large measure, indeed by definition :

$$\begin{aligned}\left(1 + \frac{2d_{KL}(\mu\|\pi)}{\varepsilon}\right)\mu(A^c) &\leq \mathbb{E}_\mu[\ln \mu/\pi \mathbf{1}_{A^c}] \\ &\leq d_{KL}(\mu\|\pi) - \mathbb{E}_\mu[(\mu/\pi - 1)\mathbf{1}_A] \\ &= d_{KL}(\mu\|\pi) + \pi(A) - \mu(A) \\ &= d_{KL}(\mu\|\pi) + \mu(A^c) + \pi(A) - 1 \\ &\leq d_{KL}(\mu\|\pi) + \mu(A^c)\end{aligned}$$

where we used $\ln x \leq x - 1$. After rearranging we have $\mu(A^c) \leq \varepsilon/2$, in particular A is non empty.

Now consider $\hat{\mu} = \mu_{|A} = \mathbf{1}_A\mu/\mu(A)$, then :

$$\left\|\frac{\hat{\mu}}{\pi}\right\|_\infty = \frac{1}{\mu(A)} \max_{x\in A} \frac{\mu(x)}{\pi(x)} \leq \frac{e}{1 - \varepsilon/2} \exp(2d_{KL}(\mu\|\pi)/\varepsilon) \leq e^2 \exp(2d_{KL}(\mu\|\pi)/\varepsilon)$$

So by Proposition 1.5, we have :

$$\|\hat{\mu}\mathscr{P}_t - \pi\| \leq \frac{1}{2}\exp(1 + d_{KL}(\mu\|\pi)/\varepsilon - t/t_{\mathrm{rell}})$$

Using $e^{1-1/x} \leq x$ for positive x, we have for $t_0 = t_{\mathrm{rell}}/\varepsilon(1 + d_{KL}(\mu\|\pi))$, $\|\hat{\mu}\mathscr{P}_{t_0} - \pi\| \leq \varepsilon/2$. Then using the following lemma we have $\|\hat{\mu}\mathscr{P}_{t_0} - \mu\mathscr{P}_{t_0}\|_{TV} \leq \|\hat{\mu} - \mu\|_{TV} = \mu(A^c) \leq \varepsilon/2$. Therefore :

$$\|\mu\mathscr{P}_{t_0} - \pi\|_{TV} \leq \varepsilon$$

We can then conclude using $\mu = \delta_x\mathscr{P}_t$ □

With this lemma and exponential decay of entropy we can already establish a general criteria for cutoff :

Theorem 2. *(Modified logarithmic Sobolev condition for cutoff) Let X_n be a sequence of irreducible reversible Markov chains with modified logarithmic Sobolev constant ρ. We have :*

$$t_{mix}(\varepsilon) - t_{mix}(1-\varepsilon) \leq \frac{t_{rell}}{\varepsilon}(1 + e^{-\rho t_{mix}(1-\varepsilon)} \log 1/\pi^\star)$$

Where $\pi^\star = \min_{x\in V} \pi(x)$. In particular if :

$$t_{mix}(1-\varepsilon) \geq \frac{1}{\rho} \log\log \frac{1}{\pi^\star} + o\left(1/\rho \log\log \frac{1}{\pi^\star}\right)$$

Then we have cutoff.

The issue with this condition is that as previously said an estimation of ρ is very difficult, furthermore as demonstrated in the product case the cutoff happens often precisely at the threshold $t_{\text{mix}} = 1/2\lambda \log\log 1/\pi^*$ (Where π^* in the product case is $\pi^{*,n} = (\pi^*)^n$. So $\log\log \pi^* = \log n + \mathcal{O}(1)$.)
The following proposition shows that this condition is not really useful to prove cutoff in without a precise knowledge of ρ.

Proposition 2.1. $\rho \leq 2\lambda$.

Proof. The Poincare inequality is in reality a linearization of the modified logarithmic Sobolov inequality. Let $f = 1 + \varepsilon g$, with g of mean 0 and ε small enough such that f is positive, then :

$$\text{Ent}_\pi[f] = \mathbb{E}_\pi[f \log f] = \frac{\varepsilon^2}{2}\mathbb{E}_\pi[g^2] + o(\varepsilon^2)$$

and since $\pi L = 0$:

$$\mathcal{E}(f, \log f) = -\varepsilon \mathbb{E}_\pi[Lg \log f] = -\varepsilon^2 \mathbb{E}_\pi[gLg] + o(\varepsilon^2)$$

By taking $\varepsilon \to 0$ and since $\mathcal{E}(f,f)$ is invariant by translation (because Γ is), we have for any g:

$$\mathcal{E}(g,g) \geq \frac{\rho}{2}\text{Var}_\pi[g]$$

$\square$

Varentropy One refinement would be to not look at the entire decay of entropy until time $t_{\text{mix}}(1-\varepsilon)$ but only look at the *concentration of entropy* that will be quantified using *varentropy* introduced in a somewhat different context for log-concave measures (see [FLM20]).

Definition 2.2 (Varentropy)**.** If $\nu \ll \mu$ are two measures then their *varentropy* is:

$$\mathscr{V}_{KL}(\nu\|\mu) = \text{Var}_\nu\left[\log\frac{\partial \nu}{\partial \mu}\right]$$

in the same way the *maximal varentropy* for a Markov chain is defined as :

$$\mathscr{V}^\star(t) = \max_{x\in V} \mathscr{V}_{KL}(\mathscr{P}_t(x,.)\|\pi)$$

Modifying the Chebychev inequality and using a total variation bound we can upper bound the mutual entropy using varentropy.

Lemma 2. *For any two probability measures $\mu \ll \nu$ we have :*

$$d_{KL}(\mu\|\nu) \leq \frac{1 + \sqrt{\mathscr{V}_{KL}(\mu\|\nu)}}{1 - \|\mu-\nu\|_{TV}}$$

Proof. Take $\varepsilon = 1 - \|\mu - \nu\|_{TV} > 0$ since $\nu \ll \mu$. Let $\theta = d_{KL}(\mu\|\nu) - \sqrt{\mathscr{V}_{KL}(\mu\|\nu)}/\varepsilon$ and let $h = \frac{\partial\mu}{\partial\nu}$ and :

$$A = \{x | \ln h(x) \geq \theta\}$$

by Chebychev's inequality we have :

$$\begin{aligned}\mu(A) &= \int_A d\mu(x) \\ &\geq 1 - \int_{A^c} \frac{(h(x) - \mathbb{E}_\mu[h])^2}{(\theta - \mathbb{E}_\mu[h])^2} d\mu(x) \\ &\geq 1 - \varepsilon^2\end{aligned}$$

because $\mathscr{V}_{KL}(\mu\|\nu) = \mathbb{E}_\mu[\ln h]$. Then by definition of A :

$$\mu(A) = \int_A h(x) d\nu(x) \geq e^\theta \int_A d\nu(x) = e^\theta \nu(A)$$

With these two inequalities we have :

$$\begin{aligned}1 - \varepsilon = |\mu - \nu\|_{TV} &\geq \mu(A) - \nu(A) \\ &\geq (1 - e^{-\theta})\mu(A) \\ &\geq (1 - e^{-\theta})(1 - \varepsilon^2)\end{aligned}$$

By rearranging we get :

$$\begin{aligned}&1 - \varepsilon \geq (1 - e^{-\theta})(1 - \varepsilon^2) \\ \iff & e^{-\theta} \geq 1 - \frac{1}{1+\varepsilon} \\ \iff & -\ln \frac{\varepsilon}{1+\varepsilon} = \ln(1 + 1/\varepsilon) \geq \theta\end{aligned}$$

using again $\ln x \leq x - 1$ we have $\theta \leq 1/\varepsilon$ which gives us the desired inequality. □

Using these lemmas we can easily prove the following theorem :

Theorem 3 (Entropic concentration implies cutoff). *For any Markov chain on a finite number of states and any $\varepsilon \in]0, 1[$ we have :*

$$t_{mix}(\varepsilon) - t_{mix}(1 - \varepsilon) \leq \frac{2t_{rell}}{\varepsilon^2}\left(1 + \sqrt{\mathscr{V}^\star(t_{mix}(1 - \varepsilon))}\right)$$

Proof. We apply lemma 1 to $t = t_{\text{mix}}(1 - \varepsilon)$. And by definition of $t_{\text{mix}}(1 - \varepsilon)$ we have :

$$\forall x \in V,\ \|\mathscr{P}_{t_{\text{mix}}(1-\varepsilon)}(x, .) - \pi\|_{TV} \leq 1 - \varepsilon$$

Then by lemma 2 we have for any $x \in V$:

$$d^\star(t_{\text{mix}}(1 - \varepsilon)) \leq \frac{1 + \sqrt{\mathscr{V}^\star(t_{\text{mix}}(1 - \varepsilon))}}{1 - \max_x \|\delta_x \mathscr{P}_{t_{\text{mix}}(1-\varepsilon)} - \pi\|_{TV}} = \frac{1 + \sqrt{\mathscr{V}^\star(t_{\text{mix}}(1 - \varepsilon))}}{\varepsilon}$$

Finally with $1 \leq 1/\varepsilon$ we get the desired inequality. □

In particular if the following *entropic concentration* condition is met then the sequence exhibits a cutoff :

$$1 + \sqrt{\mathscr{V}^\star(t_{\text{mix}}(1 - \varepsilon))} \ll \frac{t_{\text{mix}}(1 - \varepsilon)}{t_{\text{rell}}}$$

This condition can be clarified under a positive curvature condition.

2.2 Discrete Ricci curvature

Transportation distance The transportation distance of two measure is a way to relate the topology of a Polish space (X, d) and the divergence of borelian measures.

Definition 2.3 (L^1 transportation distance)**.** The L^1 *transportation distance* (or 1 Wasserstein distance) between μ and ν is the best average distance between μ and ν :

$$W_1(\mu, \nu) = \inf_{\xi \in \Pi(\mu,\nu)} \mathbb{E}_\xi[d]$$

this can be see as the best way to transport the mass of μ towards ν.

W_1 defines a distance on the set of measure with finite first order. The proof is detailed in the annex but isn't really relevant to the subject. Furthermore we have a dual interpretation of this distance using 1 Lipschitz functions.

Theorem 4 (Kantorovich duality)**.** *When either side is non-infinite :*

$$W_1(\mu, \nu) = \sup_{f \in \text{ 1-Lip}} \int_X f d\mu - \int_X f d\nu$$

This dual interpretation is much easier to manipulate and has very useful proprieties, for instance the contraction of the Markov operator.

Ricci curvature We can now define the Ricci curvature along xy. In a Riemannian manifold the curvature measures the contraction of two parallel geodesics starting from two points, in our case the chain will play the role of geodesics and the optimal coupling will be an analogue of parallelism.

Definition 2.4 (Coarse Ricci curvature)**.** Let (X, d) be a Polish space and $\mathscr{P}_t$ a Markov semi-group on X. Let $x \neq y \in X$. The *coarse Ricci curvature* at time t, $\kappa_t(x, y)$ of $\mathscr{P}$ along xy is defined by :

$$W_1(\mathscr{P}_t(x, .), \mathscr{P}_t(y, .)) = (1 - \kappa_t(x, y))d(x, y)$$

The scalar Ricci curvature of m is $\kappa_t = \inf_{x \neq y} \kappa_t(x, y)$. We denote $W_{1,t}(x, y) = W_1(\mathscr{P}_t(x, .), \mathscr{P}_t(x, .))$ or $W_1(x, y)$ when there is no confusion possible. In our finite case the metric is almost exclusively the graph metric, i.e. the length of the shortest path.
We also have a dual interpretation of curvature.

Proposition 2.2. $\|\mathscr{P}_t\|_{Lip} = 1 - \kappa_t$. *Where:*

$$\|M\|_{Lip} = \sup_{f \text{ 1-Lip}} \|Mf\|_{Lip}$$

Proof. If $\kappa_t = 1$ then $\mathscr{P}_t$ is a multiple of $\mathbf{1}$ and $\mathrm{im}\mathscr{P}_t$ is the subspace of constant functions. So $\|\mathscr{P}_t\|_{Lip} = 1 - \kappa_t = 0$. Assume now $\kappa_t < 1$.
Let f be 1-Lipschitz then :

$$(\mathscr{P}_t f)(x) - (\mathscr{P}_t f)(y) = \int_X f d\mathscr{P}_t(x, .) - \int_X f d\mathscr{P}_t(y, .) \leq W_1(x, y) \leq (1 - \kappa_t)d(x, y)$$

So we have proven that $\|\mathscr{P}_t\|_{Lip} \leq 1 - \kappa_t$.

Let $1-\kappa_t > \varepsilon > 0$. Take x, y such that $\kappa_t(x,y) \leq \kappa_t + \varepsilon/2$ then f 1-Lip such that $\int_X f d\mathscr{P}_t(x,.) - \int_X f \mathscr{P}_t(x,.) \geq W_1(x,y) - \varepsilon d(x,y)/2$. Then we have :

$$(\mathscr{P}_t f)(x) - (\mathscr{P}_t f)(y) \geq (1-\kappa_t-\varepsilon)d(x,y) \geq 0$$

so for any ε $\|\mathscr{P}_t\|_{Lip} \geq 1-\kappa_t-\varepsilon$ concluding the proof. □

Corollary 4.1. *For any positive times t, s we have :*

$$1-\kappa_{t+s} \leq (1-\kappa_s)(1-\kappa_t)$$

Positive curvature condition By density of $\mathbb{Q}$ and by $1-x \leq e^{-x}$ if κ_{t_0} is positive then for all $t \geq t_0$:

$$1-\kappa_t \leq e^{-(\kappa_{t_0}/t_0)t}$$

If κ_t is differentiable at 0 then let $\kappa = \lim_{t\to 0^+} \kappa_t/t$ and we have the bound:

$$1-\kappa_t \leq e^{-\kappa t}$$

The *positive curvature condition* can be either $\forall t \geq 0$ $\kappa_t \geq 0$, or κ is positive, in that case we say that the chain satisfies the κ-curvature condition.

Remark. Under stochastic normalizations $L = P - \mathrm{Id}$, where P is a stochastic operator, so by convexity :

$$\|\mathscr{P}_t\|_{Lip} \leq e^{t(\|P\|_{Lip}-1)}$$

in this case positive curvature can also refer to $\|P\|_{Lip} \leq 1$.

2.3 Entropy concentration under positive curvature condition

As seen with Proposition (2.2) the positive curvature condition means that the chain is a contraction in the sense of Lipschitz norm. Since $c_{xy} \neq 0$ if and only x y are adjacent, i.e. at distance 1 in the graph metric, we can always bound :

$$\Gamma(\mathscr{P}_t f, \mathscr{P}_t f)(x) \leq \frac{q(x)}{2}\|f\|_{Lip}^2\|\mathscr{P}_t\|_{Lip}^2$$

And with κ-curvature we have:

$$\Gamma(\mathscr{P}_t f, \mathscr{P}_t f)(x) \leq \frac{q(x)}{2}\|f\|_{Lip}^2 e^{-2\kappa t}$$

Lemma 3 (Local concentration under positive curvature)**.** *For all x:*

$$\mathscr{P}_t(f^2)(x) - (\mathscr{P}_t f)(x)^2 \leq \frac{1-e^{-2\kappa t}}{2\kappa} q(x)\|f\|_{Lip}^2$$

the left factor is to be understood as t if $\kappa = 0$.

Proof. By differentiating both side it can be checked that :

$$\mathscr{P}_t(f^2)(x) - (\mathscr{P}_t f)(x)^2 = 2\int_0^t \mathscr{P}_{t-s}\Gamma(\mathscr{P}_s f, \mathscr{P}_s f)(x) ds$$

Since $\mathscr{P}$ is a increasing operator then :

$$\mathscr{P}_{t-s}\Gamma(\mathscr{P}_s f, \mathscr{P}_s f)(x) \leq \frac{q(x)}{2}\|f\|_{Lip}^2 e^{-2\kappa s}$$

yielding the desired inequality. □

By convexity applying the Lemma to $\log \mathscr{P}_t(x,.)/\pi$ we have the bound:

$$\mathscr{V}^\star(t) \leq \frac{1-e^{-2\kappa t}}{2\kappa}\mathbb{E}_\pi[q]\max_x \|\log \mathscr{P}_t(x,.)/\pi\|^2_{Lip}$$

This last Lipschitz norm can be controlled with elementary bounds on the generator. A first Lemma will be useful, which proof is immediate by the telescoping.

Lemma 4. *If (X,d) is a α-geodesic Polish space, that is for any x,y there exits a sequence $x_1=x,\cdots,x_n=y$ such that :*

$$\forall i\leq n-1,\ d(x_i,x_{i+1})\leq\alpha,\qquad \sum_{i=1}^{n-1}d(x_i,x_{i+1})=d(x,y)$$

then for all Lipschitz functions :

$$\|f\|_{Lip}=\inf_{0<d(x,y)\leq\alpha}\frac{d'(f(x),f(y))}{d(x,y)}$$

Since graphs are 1-geodesic this means we can restrict finding the minimum on adjacent vertices. For simplicity we will now assume L to be auto-adjoint. Denote :

$$P=L+\operatorname{diag}(q)$$

Then:

$$\mathscr{P}_t(x,y)=e^{-q(x)t}\sum_{k=0}^{\infty}\frac{t^k}{k!}P^k(x,y)$$

Fix a base point o and let $f(x)=\mathscr{P}_t(o,x)/\pi(x)$. Since L is auto-adjoint we have the equation $\pi(x)P(x,y)f(y)=P(y,x)\mathscr{P}_t(o,y)$ summing over y yields :

$$\begin{aligned}\pi(x)Pf(x)&=e^{-q(o)t}\sum_{k=0}^{\infty}\frac{t^k}{k!}\sum_y P(x,y)P^k(o,y)\\&=e^{-q(o)t}\frac{1}{t}\sum_{k=0}^{\infty}k\frac{t^k}{k!}P^k(o,x)\\&\leq e^{-q(o)t}\frac{1}{t}\sum_{k=0}^{\infty}k\frac{t^k}{k!}\left(\max_x q(x)\right)^k\end{aligned}$$

By concavity :

$$\frac{Pf(x)}{f(x)}\leq\frac{1}{t}\log\frac{e^{-q(o)t}\sum_{k=0}^{\infty}e^k\frac{t^k}{k!}(\max_x q(x))^k}{\mathscr{P}_t(o,x)}=e\max_x q(x)-q(o)+\frac{1}{t}\log\frac{1}{\mathscr{P}_t(o,x)}$$

Replacing L by $\operatorname{diag}(q)3/4+L/4$, then for $k\geq\operatorname{diam}(G)$ then $P^k\geq 1/4\Delta$, so for $t\geq\operatorname{diam}(G)/4$:

$$\frac{1}{\mathscr{P}_t(o,x)}\leq\Delta^{\operatorname{diam}(G)/\max q(x)}$$

where $\Delta=\sup_x q(x)/\inf_{c_{xy}\neq 0}c_{xy}$. So for $t\geq\operatorname{diam}(G)/4$ we have :

$$\max_{x\sim y}\frac{f(y)}{f(x)}\leq(e+1)\Delta+\Delta\log\Delta$$

Yielding for $t \geq \mathrm{diam}(G)/4$:

$$\max_x \left\| \log \frac{\mathscr{P}_t(x,.)}{\pi} \right\|_{Lip} \leq K \log \Delta$$

where K is an universal constant, with Lemma 3 up to an universal constant :

$$\mathscr{V}^\star(t) \leq K \mathbb{E}_\pi[q] t \log \Delta \tag{3}$$

Proving that $t_{\mathrm{mix}} \geq \mathrm{diam}(G)/4$ for any graph large enough will prove the following theorem

Theorem 5. *For any sequence of graphs with non-negative curvature:*

$$t_{mix}(\varepsilon) - t_{mix}(1-\varepsilon) = \mathcal{O}\Big(\mathbb{E}_\pi[q] \log \Delta t_{rell} \sqrt{t_{mix}(1-\varepsilon)}\Big)$$

Under a stochastic normalization $q = 1$ so :

$$\Delta = \frac{1}{\inf_{c_{xy}>0} c_{xy}}$$

and in particular we have cutoff if :

$$t_{mix}(1-\varepsilon) \gg (t_{rell} \log \Delta)^2$$

for all $\varepsilon > 1/2$.

For simplicity in the rest of the report we will assume stochastic normalization ($q = 1$).

Lemma 5. *For any graph under stochastic normalization :*

$$diam(G) \leq 2t_{mix}(\varepsilon) + 2\sqrt{\frac{2t_{mix}(\varepsilon)}{1-\varepsilon}} + 2\sqrt{\frac{2t_{rell}}{1-\varepsilon}}$$

Proof. Fix o a starting point and $t = t_{\mathrm{mix}}(\varepsilon)$ and set :

$$A = \left\{ d(o,\cdot) \leq t + \sqrt{\frac{2t}{1-\varepsilon}} \right\}$$

If X_t is the chain proceeding from o, then the distance at the origin is the number of jumps and they occur at rate at most a Poisson variable of mean $t \max q = t$. So :

$$\mathbb{P}(X_t \in A) \geq 1 - \mathbb{P}\left(Z \geq t + \sqrt{\frac{2t}{1-\varepsilon}}; Z \sim \mathrm{Pois}(t) \right) > 1 - \frac{1-\varepsilon}{2} = \frac{1+\varepsilon}{2}$$

by Tchebychev's inequality (since $\mathrm{Var}(Z) = t$). Since $t = t_{\mathrm{mix}}(\varepsilon)$ then $\|X_t - \pi\|_{TV} \leq \varepsilon$ in particular :

$$\pi(A) > \frac{1+\varepsilon}{2} - \varepsilon = \frac{1-\varepsilon}{2} \tag{4}$$

Next, by definition of t_{rell} :

$$\mathrm{Var}_\pi(f) \leq t_{\mathrm{rell}} \mathcal{E}(f,f)$$

where $f = d(o, \cdot)$ and since f is 1-Lipschitz, $\mathcal{E}(f,f) \leq 1$. If $U \sim \pi$ denotes a variable distributed according to π, then this last inequality can be rewritten as:

$$\mathrm{Var}(f(U)) \leq t_{\text{rell}}$$

Again by Tchebychev:

$$\mathbb{P}\left(|f(U) - \mathbb{E}[f(U)]| \geq \sqrt{\frac{2t_{\text{rell}}}{1-\varepsilon}} \right) \leq \frac{1-\varepsilon}{2}$$

but by (4), $f(U) \in A$ and $|f(U) - \mathbb{E}[f(U)]| \leq \sqrt{\frac{2t_{\text{rell}}}{1-\varepsilon}}$ cannot be disjoint proving :

$$\mathbb{E}[f(U)] \leq t_{\text{mix}}(\varepsilon) + \sqrt{\frac{2t_{\text{mix}}(\varepsilon)}{1-\varepsilon}} + \sqrt{\frac{2t_{\text{rell}}}{1-\varepsilon}}$$

This is true for all origin points o, so for any two points x, y, by the triangular inequality :

$$2\left(t_{\text{mix}}(\varepsilon) + \sqrt{\frac{2t_{\text{mix}}(\varepsilon)}{1-\varepsilon}} + \sqrt{\frac{2t_{\text{rell}}}{1-\varepsilon}} \right) \geq \mathbb{E}[d(x,U) + d(y,U)] \geq d(x,y)$$

□

Finally by a classic argument we have $t_{\text{rell}} = \mathcal{O}(t_{\text{mix}}(\varepsilon))$.
Let f be a eigenfunction of L for the value $-\lambda = -1/t_{\text{rell}}$, such that $\|f\|_\infty = 1$. Take x such that $|f(x)|$ is maximal, since the eigenvectors of a symmetric operator are orthogonal $\mathbb{E}_\pi[f] = \langle f, \mathbf{1} \rangle_\pi = 0$:

$$|\mathscr{P}_t f(x)| = e^{-\lambda t}|f(x)| = |(\mathscr{P}_t(x,\cdot) - \pi)f| \leq \|f\|_\infty 2\|\mathscr{P}_t(x,\cdot) - \pi\|_{TV}$$

Proving $t_{\text{mix}}(\varepsilon) \geq -t_{\text{rell}} \log(2\varepsilon)$. So under the weak assumption $t_{\text{mix}}(\varepsilon) \to \infty$:

$$\mathrm{diam}(G) \leq 2t_{\text{mix}}(\varepsilon) + o(t_{\text{mix}}(\varepsilon))$$

proving that for large enough graphs $t_{\text{mix}}(\varepsilon) \geq \mathrm{diam}(G)/4$. Note that if $t_{\text{mix}}(\varepsilon) = \mathcal{O}(1)$ then the theorem is trivially valid.

Refinements The bound $\frac{1-e^{-2\kappa t}}{2\kappa} \leq t$ can be replaced by a bound $1/2\kappa$ when κ is positive, so under slow decrease of κ we have an other condition on t_{mix} to prove cutoff :

$$t_{\text{mix}}(\varepsilon) \gg \frac{t_{\text{rell}} \log \Delta}{\sqrt{\kappa}}$$

Corollary 5.1. *A generic condition for cutoff on a sequence of graphs using only the geometry of the chain is :*

$$\sqrt{\frac{\log n}{(\log \Delta)^3}} \gg t_{rell}$$

where $n = |G|$.

Proof. By the following lemma $\log n / \log \Delta = \mathcal{O}(t_{\text{mix}}(\varepsilon))$ so:

$$(t_{\text{rell}} \log \Delta)^2 \ll \frac{\log n}{\log \Delta}$$

is enough for cutoff. □

Lemma 6. *For all $\varepsilon \in]0,1[$ we have :*

$$t_{mix}(\varepsilon) \geq K(\varepsilon) \frac{\log[n(1-\varepsilon)/2)]}{\log \Delta}$$

as soon as the right side is bigger then 1. Where

$$K(\varepsilon) = \left(1 + \sqrt{\frac{2}{1-\varepsilon}}\right)^{-1}$$

Proof. Let $\deg(x)$ by the degree of x. Under stochastic normalization, since $\sum_{y\sim x} c_{xy} = 1$, we have :

$$\min_{y\sim x} c_{xy} \geq 1/\deg(x)$$

So $\Delta \leq \max \deg(x)$. Let $d = \max \deg(x)$, if $1 \leq t \leq K(\varepsilon) \frac{\log n(1-\varepsilon)/2}{\log d}$ take:

$$A = \left\{ d(o,\cdot) \leq t + \sqrt{\frac{2t}{1-\varepsilon}} \right\}$$

By Lemma 5 we have :

$$\mathscr{P}_t(o, A) \geq \frac{\varepsilon + 1}{2}$$

We have $\sqrt{2t/(1-\varepsilon)} \leq \sqrt{2/(1-\varepsilon)}t$, and at most d^y points are at distance $\leq y$ from o. Therefore:

$$\pi(A) \leq d^{K(\varepsilon)^{-1}t}/n \leq \exp\left(\log(n(1-\varepsilon)/2) - \log n\right) = (1-\varepsilon)/2$$

So :

$$\|\mathscr{P}_t(o,\cdot) - \pi\|_{TV} \geq \mathscr{P}_t(o,A) - \pi(A) \geq \frac{\varepsilon + 1 - (1-\varepsilon)}{2} = \varepsilon$$

Proving :

$$t_{\text{mix}}(\varepsilon) \geq K(\varepsilon) \frac{\log n(1-\varepsilon)/2}{\log d} \geq K(\varepsilon) \frac{\log n(1-\varepsilon)/2}{\log \Delta}$$

□

The condition :

$$t_{\text{rell}} \ll \sqrt{\frac{\log n}{(\log \Delta)^3}}$$

can be understood as a control of the expansion of the underlying graph. The isoperimetric constant for graphs was introduced in [Moh89] as an analogue of the isoperimetric constant for Riemanninan manifolds. It serves to quantify the expansion of a graph that is to say the minimal proportions of edges of any set that are on the periphery of the set. Indeed the classic Cheeger inequality relates the isoperimetric constant $h(G)$ and the spectral gap $1/t_{\text{rell}}$ (with no assumptions on the normalization) :

Proposition 2.3 (Cheeger inequalities). *For any connected chain on a graph G:*

$$\frac{h(G)^2}{2\max_x q(x)} \leq \frac{1}{t_{rell}} \leq 2h(G)$$

where :

$$h(G) = \min_{0<\pi(S)\leq 1/2} \frac{\mathcal{E}^*(\mathbf{1}_S, -\mathbf{1}_{S^c})}{\pi(S)}$$

Remark. If G is a symmetric unweighted graph then π is uniform and for $y \notin S$:

$$L\mathbf{1}_S(y) = \sum_{x\sim y} \mathbf{1}_S(x) = |\{x \sim y; x \in S\}|$$

So :

$$\mathcal{E}^*(\mathbf{1}_S, \mathbf{1}_{S^c}) = -|\{x \sim y; x \in S, y \notin S\}|/n = -|\partial S|/n$$

Yielding the more common definition of h :

$$h(G) = \min_{0<|S|\leq n/2} \frac{|\partial S|}{|S|}$$

furthermore $q(x) = \deg(x)$ yielding the classical Cheeger inequalities

Proof. Assume L is auto-adjoint, if not replace all mentions of L by L^*.

First Inequality Let g by a eigenfunction for the value λ. By developing :

$$\mathcal{E}(g,g) = \frac{\mathcal{E}(g,g)\mathcal{E}(g,-g)}{\frac{1}{2}\sum_{x,y} c_{xy}\pi(x)(g(x)+g(y))^2}$$

The quotient can be bounded with :

$$\begin{aligned}\frac{1}{2}\sum_{x,y} c_{xy}\pi(x)(g(x)+g(y))^2 &= 2\sum_{xy} c_{xy}\pi(x)g(x)^2 - \mathcal{E}(g,g) \\ &\leq 2\max_x q(x)\mathrm{Var}_\pi(g)\end{aligned} \tag{5}$$

Furthermore by C.S :

$$\mathcal{E}(g,g)\mathcal{E}(g,-g) \geq \frac{1}{4}\left(\sum_{xy} \pi(x)c_{xy}|g(x)^2 - g(y)^2|\right) \tag{6}$$

Label the vertices $1, \cdots, n$ such that $g(i)$ is non-increasing. Then :

$$\begin{aligned}\sum_{xy} \pi(x)c_{xy}|g(x)^2 - g(y)^2| &= 2\sum_{i<j} \pi(i)c_{ij}(g(i)^2 - g(j)^2) \\ &= 2\sum_{k=0}^{n}\sum_{i\leq k}\sum_{j>k} \pi(i)c_{ij}(g(k)^2 - g(k+1)^2)\end{aligned}$$

Because $\sum_{j>k\geq i}(g(k)^2 - g(k+1)^2) = g(i)^2 - g(j)^2$. Let $S_k = [\![1,k]\!]$, we have :

$$\mathbb{E}_\pi[\mathbf{1}_{S_k} L \mathbf{1}_{S_k^c}] = \sum_{i\leq k}\sum_{j>k} \pi(i)c_{ij} = \mathcal{E}(\mathbf{1}_{S_k}, -\mathbf{1}_{S_k^c})$$

So by definition of $h(G)$:

$$\begin{aligned} 2\sum_{k=0}^{n}\sum_{i\leq k}\sum_{j>k}\pi(i)c_{ij}(g(k)^2 - g(k+1)^2) &\geq 2h(G)\sum_{k=0}^{n}\pi(S_k)(g(k)^2 - g(k+1)^2) \\ &= 2h(G)\mathrm{Var}_\pi(g) \end{aligned}$$

where we used $\pi(S_{k+1}) = \pi(k+1) + \pi(S_k)$. Combining this with (5) and (6) we get :

$$\frac{1}{t_{\text{rell}}} = \frac{\mathcal{E}(g,g)}{\mathrm{Var}_\pi(g)} \geq \frac{h(G)^2}{2\max_x q(x)}$$

Second Inequality Take S achieving the bound $h(G)$ and let $f = \mathbf{1}_S/\pi(S) - \mathbf{1}_{S^c}/\pi(S^c)$. We have $\mathrm{Var}_\pi f = \frac{1}{\pi(S)} + \frac{1}{\pi(S^c)}$ and:

$$\mathcal{E}(f,f) = -\frac{1}{\pi(S^c)}\mathbb{E}_\pi[Lf\mathbf{1}_{S^c}] + \frac{1}{\pi(S)}\mathbb{E}_\pi[Lf\mathbf{1}_S]$$

Since on $x \in S$:

$$Lf(x) = (1/\pi(S) + 1/\pi(S^c))L\mathbf{1}_{S^c}(x)$$

and on S^c we have the same relation with the inverse sign. Since $1/\pi(S^c) \leq 1/\pi(S)$:

$$\begin{aligned} \mathcal{E}(f,f) &\leq \frac{1}{\pi(S)}(1/\pi(S) + 1/\pi(S^c))\mathbb{E}_\pi[L\mathbf{1}_S\mathbf{1}_{S^c} + L\mathbf{1}_{S^c}\mathbf{1}_S] \\ &= \mathrm{Var} f 2\frac{\mathcal{E}(\mathbf{1}_S, -\mathbf{1}_{S^c})}{\pi(S)} \\ &= \mathrm{Var} f 2h(G) \end{aligned}$$

thus $\lambda \leq \mathcal{E}(f,f)/\mathrm{Var}(f) \leq 2h(G)$. □

In conclusion this condition can also by viewed as a condition on good expansion of the graph. To sum up, if the graphs has non-negative curvature and has good expansion, i.e. :

$$h(G) \gg \sqrt[4]{\frac{(\log[\max_x \deg(x)])^3}{\log n}}$$

Then the sequence exhibits a cutoff.

3 Application to random walks on abelian groups

Random walks on abelian groups are understood as random walks one of their associated Cayley graphs. That is to say a abelian group G and a set of generators S, in the associated graph x is linked to $x+s$ with $s \in S$, here we will assume the graph is undirected i.e.

$S^{-1} = S$. Then under stochastic normalization the generator of the simple random walk on this graph is :

$$Lf(x) = \frac{1}{d}\sum_{s\in S} f(x+s) - f(x) = Pf(x) - f(x)$$

where $d = |S|$.

Proposition 3.1. *Any simple random walk on a abelian Cayley graph has non-negative curvature.*

Proof. Let f be a 1-Lipschitz function and x, y be a distance 1 in the Cayley graph, i.e. there is $s \in S$ such that $y = x + s$. Then :

$$Pf(y) - Pf(x) = \frac{1}{d}\sum_{s'\in S} f(x+s+s') - f(x+s')$$

but since the graph is abelian $x+s+s' = x+s'+s$ so $d(x+s+s', x+s) = 1$ and $Pf(y) - Pf(x) \leq 1$. Proving $\|P\|_{Lip} \leq 1$ so $\|\mathscr{P}_t\|_{Lip} \leq 1$, i.e. $\kappa_t \geq 0$ for all t. □

So what we proved yields that any sequence of abelian Cayley graph such that :

$$t_{\text{rell}} \ll \sqrt{\frac{\log n}{(\log d)^3}}$$

exhibit a cutoff.

References

[FLM20] Matthieu Fradelizi, Jiange Li, and Mokshay Madiman. Concentration of information content for convex measures. *Electronic Journal of Probability*, 25, Jan 2020.

[Moh89] Bojan Mohar. Isoperimetric numbers of graphs. *Journal of Combinatorial Theory, Series B*, 47(3):274–291, 1989.

[Sal21] Justin Salez. Cutoff for non-negatively curved markov chains. 2021.

[Vil09] Cédric Villani. *Optimal Transport*. Springer Berlin Heidelberg, 2009.

Measure theory annex

Let Γ be the set of measurable sets such that $A \in \Gamma$ iif:

$$\forall B \subset A \text{ measurable } , \mu(B) \geq \nu(B)$$

This propriety is stable by reunion so $\Gamma_{\max} = \cup_{A \in \Gamma} A$ is the maximal measurable set with this propriety

Proposition 3.2. *Using these notations we have :*

$$\mu(\Gamma_{\max}) - \nu(\Gamma_{\max}) = \|\mu - \nu\|_{TV}$$

Proof. Let A be such that $\mu(A) - \nu(A) > \|\mu - \nu\|_{TV} - \varepsilon$, if there exists $B \subset A$ such that $\nu(B) - \mu(B) \geq \varepsilon$ then :

$$\mu(A) - \nu(A) \leq -\varepsilon + \mu(A - B) - \nu(A - B) \leq \|\mu - \nu\|_{TV} - \varepsilon$$

so for any measurable subset of A we have $\nu(B) < \mu(B) + \varepsilon$. Denoting Γ^ε the set of sets with this propriety, and $\Gamma^\varepsilon_{\max}$ the maximal set. Then we have :

$$\begin{aligned} \mu(\Gamma^\varepsilon_{\max}) - \nu(\Gamma^\varepsilon_{\max}) &= \mu(A) - \nu(A) + \mu(\Gamma^\varepsilon_{\max} - A) - \nu(\Gamma^\varepsilon_{\max} - A) \\ &> \|\mu - \nu\|_{TV} - 2\varepsilon \end{aligned}$$

because $\Gamma^\varepsilon_{\max} - A \subset \Gamma^\varepsilon_{\max}$. But the $\Gamma^\varepsilon_{\max}$ are decreasing sets therefore :

$$\mu\left(\lim_{\varepsilon \to 0} \Gamma^\varepsilon_{\max}\right) - \nu\left(\lim_{\varepsilon \to 0} \Gamma^\varepsilon_{\max}\right) \geq \|\mu - \nu\|_{TV}$$

but we have exactly $\Gamma_{\max} = \lim_{\varepsilon \to 0} \Gamma^\varepsilon_{\max}$. □

We can now generalize proposition 1.2 :

Corollary 5.2. *If $\nu \ll \mu$ let $h = \frac{\partial \nu}{\partial \mu}$ then :*

$$\|\mu - \nu\|_{TV} = \frac{1}{2}\|h - 1\|_\mu$$

Proof. If this case $\{x \in \mathcal{H} | h(x) \leq 1\}$ also realizes the sup of the total variation distance. Indeed let $A \subset \{x \in \mathcal{H} | h(x) \leq 1\}$ (measurable), then we have :

$$\nu(A) = \int_A h(x) d\mu(x) \leq \int_A d\mu(x) = \mu(A)$$

therefore $\Gamma_{\max} \supset \{x \in \mathcal{H} | h(x) \leq 1\}$ and :

$$\begin{aligned} \mu(\Gamma_{\max}) - \nu(\Gamma_{\max}) &\geq \mu(\{x \in \mathcal{H} | h(x) \leq 1\}) - \nu(\{x \in \mathcal{H} | h(x) \leq 1\}) \\ &= \int_{\Gamma_{\max}} (1 - h(x)) \mathbf{1}_{h(x) \leq 1} d\mu(x) \\ &\geq \int_{\Gamma_{\max}} (1 - h(x)) d\mu(x) \\ &= \mu(\Gamma_{\max}) - \nu(\Gamma_{\max}) \end{aligned}$$

Therefore we have :

$$\|\mu-\nu\|_{TV}=\int_{\mathcal{H}}(1-h(x))\mathbf{1}_{h(x)\leq 1}d\mu(x)$$

but also :

$$\|\mu-\nu\|_{TV}=\nu(\{x\in\mathcal{H}|h(x)>1\})-\mu(\{x\in\mathcal{H}|h(x)>1\})=\int_{\mathcal{H}}(h(x)-1)\mathbf{1}_{h(x)>1}d\mu(x)$$

Therefore : $2\|\mu-\nu\|_{TV}=\int_{\mathcal{H}}|h(x)-1|d\mu(x)$. □

We also have a expression of the distance in terms of coupling.

Proposition 3.3. *For any two laws μ,ν we have :*

$$\|\mu-\nu\|_{TV}=\inf_{\substack{\pi\in\Gamma\\ X,Y\sim\pi}}\mathbb{P}(X\neq Y)$$

where Γ is the set of product laws that have projection coinciding with μ and ν.

Proof. Let A be a measurable set and (X,Y) a coupling, then :

$$\mathbb{P}(X\neq Y)\geq\mathbb{P}(X\in A,Y\notin A)\geq\mathbb{P}(X\in A)-\mathbb{P}(Y\in A)=\mu(A)-\nu(A)$$

therefore we have proved the first inequality.

We know the TV sup is met by $\Gamma_{\max}$ here we prove that the coupling bound is also met. If $\mu(\Gamma_{\max})=\nu(\Gamma_{\max})$ then $\mu=\nu$ and if $\mu(\Gamma_{\max})-\nu(\Gamma_{\max})=1$ then μ and ν are of disjoint support and the coupling bound is met by any coupling. In the general case we assume $x=\mu(\Gamma_{\max})-\nu(\Gamma_{\max})\in]0,1[$. We can now define 3 law, for any measurable set A:

$$\begin{cases}\gamma_0(A) &= \frac{\mu(A\cap\Gamma_{\max}^c)+\nu(A\cap\Gamma_{\max}))}{1-x}=\frac{\mu(A\cap\Gamma_{\min})+\nu(A\cap\Gamma_{\min}^c))}{1-x}\\ \gamma_1(A) &= \frac{\mu(A\cap\Gamma_{\max})-\nu(A\cap\Gamma_{\max})}{x}\\ \gamma_2(A) &= \frac{\nu(A\cap\Gamma_{\min})-\mu(A\cap\Gamma_{\min})}{x}\end{cases}$$

where $\Gamma_{\min}$ is the maximal set such that all subsets verify $\mu(B)\leq\nu(B)$, by symmetry $\Gamma_{\min}$ also meets the sup of the TV, thus all these law are positives.

On $\Gamma_{\min}\cap\Gamma_{\max}$ we have $\nu=\mu$, but both γ_1 and γ_2 are zero, therefore the support of γ_1 is in $\Gamma_{\max}-\Gamma_{\min}$ and γ_2 is in $\Gamma_{\min}-\Gamma_{\max}$, in particular there are of disjoint support. Take $X_i\sim\gamma_i$ and $\xi\sim\mathcal{B}(x)$ (independent) then we have the following coupling:

$$\begin{cases}X=Y=X_1 \text{ if } \xi=0\\ X=X_1,Y=X_2 \text{ else}\end{cases}$$

we have a.s $X_1\neq X_2$ therefore we have exactly :

$$\mathbb{P}(X\neq Y)=\mathbb{P}(\xi\neq 0)=x=\|\mu-\nu\|_{TV}$$

We now only need to prove that we have the correct projections. But since ξ is independent of Y then :

$$\mathbb{E}[X|Y]\sim\gamma_0\mathbb{E}[1-\xi]+\gamma_1\mathbb{E}[\xi]=\mu$$

The same holds for Y proving the proposition. □

Proposition 3.4. *W_1 defines a distance on the set of measure with finite first order.*

Proof. Since d is symmetric clearly W_1 is too. Then is μ, ν have finite first order, that is to say if we fix on point $x \in X$ we have :

$$\mathbb{E}_\mu[d(x, .)] < \infty$$

by the triangular inequality this quantity is finite iif it is finite for any $x \in X$. Then take ξ the product coupling we have :

$$\mathbb{E}_\xi[d] \leq \mathbb{E}_\nu[d(x, .) + \mathbb{E}_\mu[d(x, .)] = \mathbb{E}_\mu[d(x, .)] + \mathbb{E}_\nu[d(x, .)] < \infty$$

If $\mu = \nu$ we can take diagonal measure :

$$\xi(A \times B) = \mu(A \cap B)$$

For $f(x, y) = \mathbf{1}_A(x)\mathbf{1}_B(y)$ we have :

$$\mathbb{E}_\xi[f] = \mu(A \cap B) = \mathbb{E}_\mu[f(., .)]$$

so by linearity and density for any measurable function :

$$\mathbb{E}_\xi[f] = \mathbb{E}_\mu[f(., .)]$$

but $d(x, x) = 0$ so $\mathbb{E}_\xi[d] = 0$ and $W_1(\mu, \mu) = 0$.

If $W_1(\mu, \nu) = 0$. If $X, Y \sim \xi$ we have :

$$\mathbb{P}(d(X, Y) \geq \varepsilon) \leq \frac{\mathbb{E}_\xi[d]}{\varepsilon}$$

Fix $r > 0$ let ξ be a coupling such that $\mathbb{E}_\xi[d] \leq \varepsilon r$. We have :

$$\begin{aligned}
\sup_{x \in X} \mathbb{P}(X \in B(x, r)) - \mathbb{P}(Y \in B(x, r)) &\leq \mathbb{P}(\exists x \in X,\ X \in B(x, r), Y \notin B(x, r)) \\
&\leq \mathbb{P}(d(X, Y) \geq r) \\
&\leq \varepsilon
\end{aligned}$$

As $\mathbb{P}(X \in B(x, r)) - \mathbb{P}(Y \in B(x, r)) = \mu(B(x, r)) - \nu(B(x, r))$ is independent on couplings we take $\varepsilon \to 0$, thus for every open set O we have :

$$\mu(O) = \nu(O)$$

the σ-algebra is the borelian σ-algebra so $\mu = \nu$.

Let ξ_1 be coupling of μ, ν and ξ_2 a coupling of ν, π. Let V be the subspace of bounded measurable function of X^3 such that there exists ϕ_1, ϕ_2 measurable with :

$$f(x, y, z) = \phi_1(x, y) + \phi_2(y, z)$$

Let $G : V \to \mathbb{R}$ be such that :

$$Gf = \mathbb{E}_{\xi_1}[\phi_1] + \mathbb{E}_{\xi_2}[\phi_2]$$

If we take an other choice of ϕ_1', ϕ_2' we have :

$$\phi_1(x,y) - \phi_1'(x,y) = \phi_2(y,z) - \phi_2'(y,z)$$

Then :

$$\begin{aligned}\mathbb{E}_{\xi_1}[\phi_1 - \phi_1'] &= \mathbb{E}_{\xi_1}[\phi_2(.,z) - \phi_2'(.,z)] \\ &= \mathbb{E}_{\nu}[\phi_2(.,z) - \phi_2'(.,z)] \\ &= \mathbb{E}_{\nu}[\phi_1'(x,.) - \phi_1(x,.)] \\ &= \mathbb{E}_{\xi_2}[\phi_2' - \phi_2]\end{aligned}$$

thus G is well defined on V and positive and leaves constant functions invariant.

Let $p(f) = \inf\{Gg; g \in V, f \leq g\}$ defined on the set of bound measurable functions. This is well defined since constant functions are in V. p is sub-additive and positively homogeneous, therefore Hahn-Banach theorem gives a extension of G to the space of measurable function with the condition :

$$Gf \leq p(f)$$

If f is positive then :

$$Gf \geq -p(-f)$$

but we have $0 \in V$ and $0 \geq -f$ then $p(-f) \leq 0$ and $Gf \geq 0$. Thus G is a positive form on bounded measurable function of X^3.

Now for B a measurable set of X^3 let :

$$\xi(B) = G\mathbf{1}_B$$

this defines a measure on X^3. If we take $B = A \times X$ with $A \subset X^2$, then we have :

$$\mathbf{1}_B(x,y,z) = \mathbf{1}_A(x,y)$$

so by definition of G :

$$\xi(A \times X) = G\mathbf{1}_B = \mathbb{E}_{\xi_1}[\mathbf{1}_A] = \xi_1(A)$$

the same can by said for sets of form $X \times A$. ξ is the "gluing" of ξ_1 and ξ_2. In conclusion if $X, Y, Z \sim \xi$ then $X, Y \sim \xi_1$,$Y, Z \sim \xi_2$ and X, Z is a coupling of μ and π.

With the triangular inequality we have :

$$d(X,Z) \leq d(X,Y) + d(Y,Z)$$

so :

$$W_1(\mu,\pi) \leq \mathbb{E}[d(X,Z)] \leq \mathbb{E}_{\xi_1}[d] + \mathbb{E}_{\xi_2}[d]$$

this is true for any coupling ξ_1, ξ_2 so we finally have the triangular inequality :

$$W_1(\mu,\pi) \leq W_1(\mu,\nu) + W_1(\nu,\mu)$$

□

Since X is Polish space $\Pi(\mu,\nu)$ is in fact weak precompact but also closed, therefore an optimal coupling does exist.

Convex analysis annex

Using this optimal coupling in the W_1 distance and convex analysis there is an dual representation for W_1 (see [Vil09] for a rigorous proof)

Theorem 6 (Kantorovich duality). *When either side is non-infinite :*

$$W_1(\mu,\nu) = \sup_{f\in \text{ 1-Lip}} \int_X f d\mu - \int_X f d\nu$$

Proof. I will give the basic idea of the proof. If ξ is any law on X^2 if ξ is a coupling of μ and ν then whenever it is defined we always have :

$$\mathbb{E}_\mu[f] + \mathbb{E}_\nu[g] = \mathbb{E}_\xi[f+g]$$

If ξ is not a coupling WLOG we can for example assume $\xi(A\times X) < \mu(A)$ Take $f = \mathbf{1}_A, g = 0$ then we have :

$$\mathbb{E}_\mu[f] > \mathbb{E}_\xi[f]$$

Taking $f_n = nf$ we have :

$$\sup_{f\in L^1(\mu), g\in L^1(\nu)} \langle f,g\rangle = \mathbb{E}_\mu[f] + \mathbb{E}_\nu[g] - \mathbb{E}_\xi[f+g] = \infty \iff \xi \notin \Pi(\mu,\nu)$$

if not this sup is 0, so we can rewrite W_1 as :

$$\begin{aligned} W_1(\mu,\nu) &= \inf_{\xi\in P(X^2)} \mathbb{E}_\xi[d] + \sup_{f\in L^1(\mu), g\in L^1(\nu)} \langle f,g\rangle \\ &= \inf_{\xi\in P(X^2)} \sup_{f\in L^1(\mu), g\in L^1(\nu)} \mathbb{E}_\xi[d] + \langle f,g\rangle \end{aligned}$$

$P(X^2)$ with $d\in L^1$ and $L^1(\mu)\times L^1(\nu)$ are both convex set and the function $\xi, f, g \to \mathbb{E}_\xi[d] + \langle f,g\rangle$ stratifies conditions of quasi-convexity and semi-continuity of Sion's minimax theorem such that we can exchange both quantities :

$$W_1(\mu,\nu) = \sup_{f\in L^1(\mu), g\in L^1(\nu)} \inf_{\xi\in P(X^2)} \mathbb{E}_\xi[d-(f+g)] + \mathbb{E}_\mu[f] + \mathbb{E}_\nu[g]$$

If $d-(f+g)$ is positive then the $\inf_\xi \mathbb{E}_\xi[d-(f+g)] \geq 0$ take $\xi = \mu\otimes\mu$ we have ξ p.s. $d-(f+g) = 0$ thus is $d-(f+g)$ is positive :

$$\inf_{\xi\in P(X^2)} \mathbb{E}_\xi[d-(f+g)] = 0$$

If not there exists μ,ν-a.s. the exists x, y such that $d(x,y) - (f(x)+g(y)) < 0$. Let $h = (f+g-d)^+ \in L^1(\mu)\otimes L^1(\nu)$ take ξ_n with marginal $d\xi_n(x,y) = h^n(x,y)d\mu(x)d\nu(x)/\mathbb{E}_{\mu\otimes\nu}[h^n]$. Then we have :

$$\mathbb{E}_{\xi_n}[d-(f+g)] = -\frac{\mathbb{E}_{\mu\otimes\nu}[h^{n+1}]}{\mathbb{E}_{\mu\otimes\nu}[h^n]}$$

this divergence to $-\infty$ if $h\geq 1$ with non 0 probability if not replace n by $-n$ and get the same result. In conclusion:

$$\inf_{\xi\in P(X^2)} \mathbb{E}_\xi[d-(f+g)] = -\infty \iff d \leq f+g$$

if not this inf is 0, so we can rewrite :

$$W_1(\mu,\nu) = \sup_{\substack{f\in L^1(\mu), g\in L^1(\nu)\\ f+g\leq d}} \mathbb{E}_\mu[f] + \mathbb{E}_\nu[g]$$

If $f+g \leq d$ then :

$$g(x) \leq \inf_y \{d(x,y) - f(y)\}$$

We can define $g = \inf_y\{d(x,y) - f(y)\} = f^d(x)$ that maximizes the sup. If f is not 1 Lip then :

$$\sup_y f(y) - d(x,y) > f(x)$$

then $f < -g$, thus not maximizing W_1, the sup appends exactly when $f(x) = -g(x)$, that is to say when f is 1 Lip, proving the theorem. $\square$