Cutoff for non-negatively curved Markov chains

An addition to Justin Salez' lecture notes

J.M. Swart

November 24, 2022

Contents

1 Continuous-time Markov chains

If A is any (finite, square, real or complex) matrix, then we define

$$e^{A} := \sum_{n=0}^{\infty} \frac{1}{n!} A^n, \tag{1.1}$$

where $A^0 := I$, the identity matrix. If A and B commute (i.e., $AB = BA$), then one can check that $e^A e^B = e^{AB}$, but this formula does not hold in general if A and B do not commute. Let $s, t \geq 0$ be real numbers. Then certainly sA commutes with tA, so

$$e^{sA} e^{tA} = e^{(s+t)A} \qquad (s, t \geq 0). \tag{1.2}$$

It is also clear that $e^{0A} = I$, so the operators $(e^{tA})_{t \geq 0}$ form a semigroup. It is easy to check that

$$\tfrac{\partial}{\partial t} e^{tA} = A e^{tA} \qquad (t \geq 0). \tag{1.3}$$

Another useful formula, that is easy to prove, is

$$e^{tA} = \lim_{n \to \infty} (I + \tfrac{t}{n} A)^n. \tag{1.4}$$

Let $\mathcal{X}$ be a finite set and let P be a probability kernel on $\mathcal{X}$. For any integer $n \geq 0$, we let P^n denote the n-th matrix power of P, which corresponds to the n-step transition kernel of the discrete-time Markov chain with transition kernel P. For any real $t \geq 0$, we define

$$P_t := e^{t(P - I)} \qquad (t \geq 0). \tag{1.5}$$

Then clearly, $P_0 = I$ and $P_s P_t = P_{s+t}$. A simple calculation yields

$$\begin{aligned} P_t &= \sum_{n=0}^{\infty} \frac{1}{n!} t^n (P - I)^n = \sum_{n=0}^{\infty} \frac{t^n}{n!} \sum_{k=0}^{n} \binom{n}{k} P^k (-1)^{n-k} \\ &= \sum_{n=0}^{\infty} \sum_{k=0}^{n} \frac{t^n}{(n-k)!\,k!} P^k (-1)^{n-k} = \sum_{k=0}^{\infty} \sum_{n=k}^{\infty} \frac{(-t)^{n-k}}{(n-k)!} \frac{t^k}{k!} P^k \\ &= \sum_{k=0}^{\infty} e^{-t} \frac{t^k}{k!} P^k = \sum_{k=0}^{\infty} p_t(k) P^k, \end{aligned} \tag{1.6}$$

where p_t is the Poisson distribution with parameter t. This means that P_t has the following interpretation. Let $(\sigma_k)_{k \geq 1}$ be i.i.d. exponentially distributed random variables with mean one. Set

$$\tau_n := \sum_{k=1}^{n} \sigma_k \qquad (n \geq 0), \tag{1.7}$$

where $\tau_0 := 0$. Let $(X_n)_{n\geq 0}$ be a (discrete time) Markov chain with transition kernel P and an arbitrary initial law ν, and define $(\hat{X}_t)_{t\geq 0}$ by

$$\hat{X}_t := X_n \quad \text{for} \quad \tau_n \leq t < \tau_{n+1}, \quad n \geq 0. \tag{1.8}$$

Then, for any $0 = t_0 < \cdots < t_m$ and $x_0, \ldots, x_m \in \mathcal{X}$, one has

$$\mathbb{P}\big[\hat{X}_{t_0} = x_0, \ldots, \hat{X}_{t_m} = x_m\big] = \nu(x_0) P_{t_1 - t_0}(x_0, x_1) \cdots P_{t_m - t_{m-1}}(x_{m-1}, x_m). \tag{1.9}$$

Indeed, the set $\{\tau_n : n \geq 1\}$ is a Poisson point set with intensity 1, which means that the number of jumps made by the process $(\hat{X}_t)_{t\geq 0}$ in the time interval $[t_{k-1}, t_k)$ has a Poisson distribution with parameter $t_k - t_{k-1}$, and disjoint time intervals are independent.

We recall that if P is a probability kernel, then we define its associated *lazy kernel* as $\frac{1}{2}(P + I)$. Applying (1.4) to $A = P - I$, we see that

$$P_t = \lim_{n\to\infty} \big(\tfrac{t}{n} P + (1 - \tfrac{t}{n}) I\big)^n. \tag{1.10}$$

This means that we may view continuous-time Markov chains as "extremely lazy" chains. Another way of interpreting (1.10) is as follows: we divide the interval $[0, t]$ into n pieces, and then independently in each time interval apply P with probability t/n and do nothing with the remaining probability. Letting $n \to \infty$, this means, of course, that we apply P at the times of a rate one Poisson point process.

An invariant law is a probability law π such that $\pi P = \pi$. If P is irreducible and aperiodic, then it has a unique invariant law π and $P^n(x, \cdot\,)$ converges to π as $n \to \infty$. Even without aperiodicity, it is true that $P_t(x, \cdot\,)$ converges to π as $t \to \infty$. This should not surprise us, since lazy chains are always aperiodic and continuous-time Markov chains are extremely lazy.

2 The relaxation and mixing times

If μ, ν are probability measures on $\mathcal{X}$, then we let $\Pi(\mu, \nu)$ denote the space of all probability measures γ on $\mathcal{X} \times \mathcal{X}$ whose first and second marginals are μ and ν, respectively:

$$\mu(x) = \sum_{y\in\mathcal{X}} \gamma(x, y) \quad \text{and} \quad \nu(x) = \sum_{x\in\mathcal{X}} \gamma(x, y). \tag{2.1}$$

We call γ a *coupling measure* for μ and ν. The total variation distance between μ and ν is given by

$$d_{\mathrm{TV}}(\mu, \nu) := \inf_{\gamma\in\Pi(\mu,\nu)} \sum_{(x,y)\in\mathcal{X}^2} \gamma(x, y) 1_{\{x \neq y\}}. \tag{2.2}$$

More probabilistically, we can formulate this by saying that

$$d_{\mathrm{TV}}(\mu,\nu) := \inf_{\substack{X\sim\mu\\ Y\sim\nu}} \mathbb{P}[X \neq Y], \tag{2.3}$$

where the infimum is over all possible *couplings* of random variables X and Y with laws μ and ν, respectively. One can check that the infimum is attained for a suitable coupling. There are other, simpler formulas for d_{TV}, for example

$$d_{\mathrm{TV}}(\mu,\nu) = \tfrac{1}{2}\sum_{x\in\mathcal{X}} \big|\mu(x) - \nu(x)\big|. \tag{2.4}$$

If P is an irreducible probability kernel with invariant law π, then we set

$$\begin{aligned} \mathcal{D}_P(n) &:= \sup_{x\in\mathcal{X}} d_{\mathrm{TV}}\big(\pi, P^n(x,\,\cdot)\big),\\ \hat{\mathcal{D}}_P(t) &:= \sup_{x\in\mathcal{X}} d_{\mathrm{TV}}\big(\pi, P_t(x,\,\cdot)\big). \end{aligned} \tag{2.5}$$

It is shown in [Sal22, Lemma 11] that

$$\tfrac{1}{2}\tilde{\mathcal{D}}_P(n) \le \mathcal{D}_P(n) \le \tilde{\mathcal{D}}_P(n) \quad \text{with} \quad \tilde{\mathcal{D}}_P(n) := \sup_{x,y\in\mathcal{X}} d_{\mathrm{TV}}\big(P^n(x,\,\cdot), P^n(y,\,\cdot)\big). \tag{2.6}$$

A similar claim holds in the continuous-time setting. We define the *mixing time* in the discrete and continuous-time settings as:

$$\begin{aligned} t^{(\varepsilon)}_{\mathrm{MIX}}(P) &:= \inf\big\{n\in\mathbb{N} : \mathcal{D}_P(n) \le \varepsilon\big\},\\ \hat{t}^{(\varepsilon)}_{\mathrm{MIX}}(P) &:= \inf\big\{t\ge 0 : \hat{\mathcal{D}}_P(t) \le \varepsilon\big\}. \end{aligned} \tag{2.7}$$

Let Spec(P) denote the spectrum of P (i.e., the set of its complex eigenvalues) and let

$$\lambda_\star(P) := \max\big\{|\lambda| : \lambda\in\mathrm{Spec}(P),\ \lambda\neq 1\big\}. \tag{2.8}$$

Then one can prove that

$$\big(\mathcal{D}_P(n)\big)^{1/n} \underset{n\to\infty}{\longrightarrow} \lambda_\star(P). \tag{2.9}$$

In other words,

$$\mathcal{D}_P(n) = \big(\lambda_\star(P) + o(1)\big)^n, \tag{2.10}$$

where $o(1)$ is a term that tends to zero as $n\to\infty$. Defining the *relaxation time* by

$$\frac{-1}{t_{\mathrm{REL}}(P)} := \log\big(\lambda_\star(P)\big), \tag{2.11}$$

we can rewrite (2.10) as

$$\mathcal{D}_P(n) = e^{-n/t_{\mathrm{REL}}(P) + o(1)}. \tag{2.12}$$

In other words, for large n, the quantity $\mathcal{D}_P(n)$ decays exponentially fast, and the relaxation time $t_{\mathrm{REL}}(P)$ is the time needed for this quantity to get a factor e^{-1} smaller (comparable to the half-life in nuclear decay).

In the continuous-time setting, we can define something similar. We start by observing that

$$\mathrm{Spec}(P - I) = \big\{\lambda - 1 : \lambda \in \mathrm{Spec}(P)\big\}. \tag{2.13}$$

We can order the (complex) eigenvalues of P according to their real parts, such that

$$1 = \lambda_1 \geq \Re(\lambda_2) \geq \cdots \geq \Re(\lambda_n). \tag{2.14}$$

Letting $\lambda_2(P)$ denote the second eigenvalue in this order, we set

$$\hat{t}_{\mathrm{REL}}(P) := \frac{1}{1 - \Re(\lambda_2(P))}. \tag{2.15}$$

(Note that this definition is unambiguous even though there may be several complex eigenvalues with the same real part as $\lambda_2(P)$.) One can prove that

$$-\frac{1}{t}\log \hat{\mathcal{D}}_P(t) \underset{t\to\infty}{\longrightarrow} \frac{1}{\hat{t}_{\mathrm{REL}}(P)}, \tag{2.16}$$

which can be rewritten as

$$\hat{\mathcal{D}}_P(t) = e^{-t/\hat{t}_{\mathrm{REL}}(P) + o(1)}. \tag{2.17}$$

For probability kernels that are irreducible but periodic, the discrete time relaxation time $t_{\mathrm{REL}}(P)$ is infinite while $\hat{t}_{\mathrm{REL}}(P)$ is finite. This is because in this case P has eigenvalues that are different from one, but whose absolute value is equal to one.

3 Reversibility

If π is an invariant law of P, and $(X_0, \ldots, X_n)$ is a Markov chain with transition kernel P and initial law π, then the *reversed chain* $(X_n, \ldots, X_0)$ is a Markov chain with transition kernel

$$P^*(x,y) := \pi(y)P(y,x)\pi(x)^{-1} \qquad (x, y \in \mathcal{X}), \tag{3.1}$$

as follows by observing that

$$\pi(x_0)\prod_{k=1}^{n} P(x_{k-1},x_k) = \pi(x_n)\prod_{k=1}^{n} P^*(x_k,x_{k-1}) \tag{3.2}$$

for all $x_0,\ldots,x_n$. A Markov chain is *reversible* if $P = P^*$, i.e., if the *detailed balance equation*

$$\pi(x)P(x,y) = \pi(y)P(y,x) \qquad (x,y \in \mathcal{X}) \tag{3.3}$$

holds. We can define an inner product on $\mathbb{R}^{\mathcal{X}}$ by

$$\langle f,g\rangle := \sum_{x\in\mathcal{X}} \pi(x)f(x)g(x). \tag{3.4}$$

Then it is easy to see that

$$\langle f,Pg\rangle = \langle P^*f,g\rangle, \tag{3.5}$$

so P^* is the adjoint of P with respect to the inner product $\langle\,\cdot\,,\,\cdot\,\rangle$ and reversibility is equivalent to P being self-adjoint.

If P is reversible, then there exist $\phi_1,\ldots,\phi_n \in \mathbb{R}^{\mathcal{X}}$ that are eigenvectors of P, i.e.,

$$P\phi_i = \lambda_i\phi_i \tag{3.6}$$

for some $\lambda_1,\ldots,\lambda_n \in \mathbb{R}$, and that are moreover *orthonormal* in the sense that

$$\langle \phi_i,\phi_j\rangle = \begin{cases} 1 & \text{if } i=j, \\ 0 & \text{if } i\neq j. \end{cases} \tag{3.7}$$

Without loss of generality we can assume that $1 = \lambda_1 \geq \cdots \geq \lambda_n$. If P is irreducible, then $\lambda_1 > \lambda_2$. If P is moreover aperiodic, then also $\lambda_n > -1$ and hence $|\lambda_i| < 1$ for all $i = 2,\ldots,n$. One has

$$\begin{aligned} P^n(x,y) &= \pi(y) + \pi(y)\sum_{i=2}^{n} \lambda_i^n \phi_i(x)\phi_i(y), \\ P_t(x,y) &= \pi(y) + \pi(y)\sum_{i=2}^{n} e^{(\lambda_i - 1)t}\phi_i(x)\phi_i(y). \end{aligned} \tag{3.8}$$

We observe that

$$\mathrm{Spec}\big(\tfrac{1}{2}(P+I)\big) = \big\{\tfrac{1}{2}(\lambda+1) : \lambda \in \mathrm{Spec}(P)\big\}, \tag{3.9}$$

which implies that the spectrum of a lazy, reversible kernel is contained in $[0,1]$ and as a result

$$\lambda_2\big(\tfrac{1}{2}(P+1)\big) = \lambda_\star\big(\tfrac{1}{2}(P+1)\big). \tag{3.10}$$

4 Covariance formulas

For any probability law μ on $\mathcal{X}$ and functions $f, g \in \mathbb{R}^{\mathcal{X}}$, we let

$$\mathrm{Cov}_\mu(f,g) := \mu(fg) - (\mu f)(\mu g) \tag{4.1}$$

denote the covariance of f and g under μ. There is a nice way to calculate the covariance of two functions of a Markov chain, that is not as well-known as it should be. If P is a probability kernel on $\mathcal{X}$ and $f, g \in \mathbb{R}^{\mathcal{X}}$, then we set

$$\Gamma_P(f,g) := \tfrac{1}{2}\big(P(fg) - (Pf)(Pg)\big). \tag{4.2}$$

The factor $\frac{1}{2}$ is there for historical reasons.

Lemma 4.1 (Covariance formula)**.** *One has*

$$\mathrm{Cov}_{\mu P^n}(f,g) = \mathrm{Cov}_\mu(P^n f, P^n g) + 2\sum_{k=1}^{n} \mu P^{n-k}\Gamma_P(P^{k-1}f, P^{k-1}g). \tag{4.3}$$

Proof The statement is trivial for $n = 0$. Fix $n \geq 1$ and for each $0 \leq k \leq n$ define a function $H_k : \mathcal{X} \to \mathbb{R}$ by

$$H_k := P^k\big((P^{n-k}f)(P^{n-k}g)\big) \qquad (0 \leq k \leq n).$$

Then

$$\begin{aligned}\mu\big(H_n - H_0\big) &= \mu P^n(fg) - \mu\big((P^n f)(P^n g)\big)\\ &= \big[\mu P^n(fg) - (\mu P^h f)(\mu P^n g)\big] - \big[\mu\big((P^n f)(P^n g)\big) - (\mu P^h f)(\mu P^n g)\big]\\ &= \mathrm{Cov}_{\mu P^n}(f,g) - \mathrm{Cov}_\mu(P^n f, P^n g).\end{aligned}$$

It follows that

$$\begin{aligned}\mathrm{Cov}_{\mu P^n}(f,g) - \mathrm{Cov}_\mu(P^n f, P^n g) &= \sum_{k=1}^{n}\mu\big[H_k - H_{k-1}\big]\\ = \sum_{k=1}^{n}\mu\big[P^k\big((P^{n-k}f)(P^{n-k}g)\big) &- P^{k-1}\big((P^{n-k+1}f)(P^{n-k+1}g)\big)\big]\\ = 2\sum_{k=1}^{n}\mu P^{k-1}\Gamma_P(P^{n-k}f, P^{n-k}g).\end{aligned}$$

Changing the summation order (setting $k' := n - k + 1$), we arrive at (4.3). ∎

In the continuous-time setting, we have a similar formula. The *carré du champ* of f and g is the function $\hat\Gamma_P(f,g)$ defined as

$$\hat\Gamma_P(f,g)(x) := \tfrac{1}{2}\sum_{y\in\mathcal{X}} P(x,y)\big(f(y) - f(x)\big)\big(g(y) - g(x)\big). \tag{4.4}$$

It seems there is no English name for this object. We claim that

$$2\hat{\Gamma}_P(f,g) = G(fg) - (Gf)g - f(Gg) \quad \text{with} \quad G := P - I. \tag{4.5}$$

To see this, we calculate

$$\begin{aligned} G(fg)(x) &= \sum_{y\in\mathcal{X}} P(x,y)\big(f(y)g(y) - f(x)g(x)\big) \\ &= \sum_{y\in\mathcal{X}} P(x,y)\Big\{\big(f(y)-f(x)\big)\big(g(y)-g(x)\big) \\ &\qquad\qquad + f(x)\big(g(y)-g(x)\big) + \big(f(y)-f(x)\big)g(x)\Big\} \\ &= 2\hat{\Gamma}_P(f,g)(x) + (Gf)(x)g(x) + f(x)(Gg)(x). \end{aligned} \tag{4.6}$$

Note that $I + \varepsilon G = (1-\varepsilon)I + \varepsilon P$ is a probability kernel for all $\varepsilon \in [0,1]$, and that

$$\begin{aligned} 2\Gamma_{I+\varepsilon G}(f,g) &= (I+\varepsilon G)(fg) - \big((I+\varepsilon G)f\big)\big((I+\varepsilon G)g\big) \\ &= \varepsilon\big[G(fg) + (Gf)g + f(Gg)\big] + O(\varepsilon^2) = 2\varepsilon\hat{\Gamma}_P(f,g) + O(\varepsilon^2) \end{aligned} \tag{4.7}$$

as $\varepsilon \to 0$. This explains why $\hat{\Gamma}_P$ is the right continuous-time analogue of the object Γ_P. We state the following lemma without proof.

Lemma 4.2 (Covariance formula in continuous time)**.** *One has*

$$\mathrm{Cov}_{\mu P_t}(f,g) = \mathrm{Cov}_\mu(P_t f, P_t g) + 2\int_0^t \mu P_{t-s}\hat{\Gamma}_P(P_s f, P_s g)\mathrm{d}s. \tag{4.8}$$

The *Dirichlet form* associated with an irreducible kernel P is the function $\hat{\mathcal{E}}_P : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}$ defined as

$$\hat{\mathcal{E}}_P(f) := \sum_{x\in\mathcal{X}} \pi(x)\hat{\Gamma}_P(f,f), \tag{4.9}$$

where π is the invariant law of P. Contrary to the carré du champ, which is not very well-known outside the French literature, the Dirichlet form is a well-known and much studied object. We claim that

$$\hat{\mathcal{E}}_P(f) = -\langle f, Gf\rangle \quad \text{with} \quad G = P - I. \tag{4.10}$$

Indeed, this follows from (4.5) and the observation that for any function $h \in \mathbb{R}^{\mathcal{X}}$ (and hence in particular for $h = fg$)

$$\sum_{x\in\mathcal{X}} \pi(x)Gh(x) = \pi(P-I)h = \pi Ph - \pi h = 0. \tag{4.11}$$

Formula (4.10) is historically the oldest definition of the Dirichlet form, which explains the factor $\frac{1}{2}$ in the definition of the carré du champ.

Lemma 4.3 (Equilibrium variance)**.** *Let P be an irredicible probability kernel with invariant law π. Then*

$$\mathrm{Var}_\pi(f) = 2\int_0^\infty \hat{\mathcal{E}}_P(P_s f)\,\mathrm{d}s. \tag{4.12}$$

Proof Since π is an invariant law, formula (4.8) simplifies to

$$\mathrm{Cov}_\pi(f,g) = \mathrm{Cov}_\mu(P_t f, P_t g) + 2\int_0^t \pi\,\hat{\Gamma}_P(P_s f, P_s g)\,\mathrm{d}s. \tag{4.13}$$

Setting $f = g$ and letting $t \to \infty$, using (4.9), we arrive at (4.12), where we use that since P is irreducible, $P_t f(x) \to \pi f$ for all $x \in \mathcal{X}$ and the variance of a constant function is zero. ∎

Inserting (4.4) into (4.9) we see that

$$\hat{\mathcal{E}}_P(f) = \tfrac{1}{2}\sum_{x,y\in\mathcal{X}} \pi(x)P(x,y)\big(f(y) - f(x)\big)^2. \tag{4.14}$$

We see from this formula that $\hat{\mathcal{E}}_P(f) \geq 0$, with equality if and only if f is constant. We recall that the *time-reversed* kernel P^* satisfies $\pi(x)P(x,y) = \pi(y)P^*(y,x)$. We see from (4.14) that

$$\hat{\mathcal{E}}_P(f) = \hat{\mathcal{E}}_{P^*}(f) = \hat{\mathcal{E}}_{(P+P^*)/2}(f). \tag{4.15}$$

By definition, the *Poincaré constant* is defined as

$$\gamma(P) := \inf_f \frac{\hat{\mathcal{E}}_P(f)}{\mathrm{Var}_\pi(f)}, \tag{4.16}$$

where we take the infimum over all f that are not constant. Since $\mathcal{E}_P(f)$ and $\mathrm{Var}_\pi(f)$ do not change if we add a constant to f, it suffices to take the infimum over all non-constant functions f with $\pi f = 0$. For such functions $\mathrm{Var}_\pi(f) = \langle f, f\rangle =: \|f\|_2^2$. Since the fraction does not change if we multiply f by a constant, we conclude that

$$\gamma(P) := \inf\big\{\hat{\mathcal{E}}_P(f) : f \in \mathbb{R}^{\mathcal{X}},\ \pi f = 0,\ \|f\|_2 = 1\big\}. \tag{4.17}$$

In [Sal22, Def 19], a similar claim is made but it seems $\|f\|_2$ is replaced by the supremumnorm. I do not see why this should hold. In view of (4.15),

$$\gamma(P) = \gamma(P^*) = \gamma\big((P + P^*)/2\big). \tag{4.18}$$

In [Sal22, Lemma 21], it is proved that

$$\gamma(P) = 1 - \lambda_2\big((P + P^*)/2\big). \tag{4.19}$$

In particular, by (2.15), this implies that

$$\hat{t}_{\mathrm{REL}}(P) = \frac{1}{\gamma(P)} \qquad \text{if } P = P^*. \tag{4.20}$$

5 The Wasserstein distance

Assume that P is irreducible reversible. Then we can equip $\mathcal{X}$ with the structure of a connected graph with set of edges E such that

$$\{x,y\}\in E \quad\Leftrightarrow\quad P(x,y)>0 \quad\Leftrightarrow\quad P(y,x)>0. \tag{5.1}$$

We define the *graph distance* by

$$\mathrm{dist}(x,y):=\inf\big\{n\geq 0: P^n(x,y)>0\big\}. \tag{5.2}$$

Equivalently, $\mathrm{dist}(x,y)$ is the length of the shortest path between x and y in the graph we have just defined. For probability measures μ,ν on $\mathcal{X}$, we define (compare (2.2))

$$\mathcal{W}(\mu,\nu):=\inf_{\gamma\in\Pi(\mu,\nu)}\sum_{(x,y)\in\mathcal{X}^2}\gamma(x,y)\mathrm{dist}(x,y). \tag{5.3}$$

Note that since $\mathrm{dist}(x,y)\leq 1_{\{x\neq y\}}$, an immediate consequence of (2.2) and (5.3) is that

$$d_{\mathrm{TV}}(\mu,\nu)\leq\mathcal{W}(\mu,\nu). \tag{5.4}$$

The infimum in (5.3) is obtained, since it is the infimum of the continuous function

$$\gamma\mapsto\sum_{(x,y)\in\mathcal{X}^2}\gamma(x,y)\mathrm{dist}(x,y) \tag{5.5}$$

over the compact and convex set $\Pi(\mu,\nu)$. Any γ for which the infimum is obtained is called an *optimal coupling.* We claim that the function

$$(\mu,\nu)\mapsto\mathcal{W}(\mu,\nu) \tag{5.6}$$

is convex. To see this, fix $0<p<1$ and (μ_i,ν_i) $(i=1,2)$. Let γ_i be an optimal coupling for μ_i,ν_i. Then $p\gamma_1+(1-p)\gamma_2$ is a coupling measure for $\mu:=p\mu_1+(1-p)\mu_2$ and $\nu:=p\nu_1+(1-p)\nu_2$. Therefore,

$$\begin{aligned}\mathcal{W}(\mu,\nu)&\leq p\sum_{(x,y)\in\mathcal{X}^2}\gamma_1(x,y)\mathrm{dist}(x,y)+(1-p)\sum_{(x,y)\in\mathcal{X}^2}\gamma_2(x,y)\mathrm{dist}(x,y)\\&=p\mathcal{W}(\mu_1,\nu_1)+(1-p)\mathcal{W}(\mu_2,\nu_2).\end{aligned} \tag{5.7}$$

Lemma 5.1 (Wasserstein metric)**.** *The Wasserstein distance is a metric on the space of probability laws on $\mathcal{X}$.*

Proof If $\mathcal{W}(\mu,\nu)=0$, then we can (optimally) couple random variables X and Y with laws μ and ν such that $\mathbb{E}[\mathrm{dist}(X,Y)]=0$. This implies $X=Y$ a.s. and hence $\mu=\nu$. It is also clear that $\mathcal{W}(\mu,\nu)=\mathcal{W}(\nu,\mu)$, so it remains to prove the triangle inequality. Let μ,ν,ρ be probability measures on $\mathcal{X}$. By Lemma 5.2 below, we can construct random variable X,Y,Z so that the law of (X,Y) is an optimal coupling for μ,ν, while (Y,Z) is an optimal coupling for ν,ρ. Now

$$\begin{aligned}\mathcal{W}(\mu,\rho) &= \mathbb{E}\big[\mathrm{dist}(X,Z)\big]\\ &\leq \mathbb{E}\big[\mathrm{dist}(X,Y)+\mathrm{dist}(Y,Z)\big]=\mathcal{W}(\mu,\nu)+\mathcal{W}(\nu,\rho),\end{aligned}\tag{5.8}$$

where we have used the triangle inequality for the graph distance. ∎

Lemma 5.2 (Combined coupling)**.** *Let (X_1,Y_1) and (Y_2,Z_2) be random variables with values in $\mathcal{X}^2$, so that Y_1 and Y_2 have the same law. Then it is possible to construct a random variable (X,Y,Z) with values in $\mathcal{X}^2$ such that (X,Y) has the same law as (X_1,Y_1) and (Y,Z) has the same law as (Y_2,Z_2).*

Proof Let $\mu(y):=\mathbb{P}[Y_1=y]=\mathbb{P}[Y_2=y]$ $(y\in\mathcal{X})$. For each $y\in\mathcal{X}$ such that $\mu(y)>0$, define $P_1(y,x):=\mathbb{P}[X_1=x\,|\,Y_1=y]$ and $P_2(y,z):=\mathbb{P}[Z_2=z\,|\,Y_2=y]$. If $\mu(y)=0$, then define $P_1(y,x)$ and $P_2(y,z)$ in an arbitrary way. Then

$$\gamma(x,y,z):=\mu(y)P_1(y,x)P_2(y,z)\qquad(x,y,z\in\mathcal{X})\tag{5.9}$$

defines a probability measure on $\mathcal{X}^3$, where it does not matter how $P_1(y,x)$ and $P_2(y,z)$ are defined when $\mu(y)=0$. Let (X,Y,Z) be a random variable with law γ. Then (X,Y) has the same law as (X_1,Y_1) and (Y,Z) has the same law as (Y_2,Z_2). Note that moreover, due to our construction, Z and X are conditionally independent given Y. ∎

For any $f\in\mathbb{R}^{\mathcal{X}}$, we define the Lipschitz "norm" as

$$\|f\|_{\mathrm{LIP}}:=\sup_{\{x,y\}\in E}\big|f(x)-f(y)\big|,\tag{5.10}$$

where E is the set of edges defined in (5.1). For any $x,y\in\mathcal{X}$, we can find $x=x_0,\ldots,x_d=y$ with $d=d(x,y)$ and $\{x_{k-1},x_k\}\in E$ for all $1\leq k\leq d$. Using the triangle inequality for $\mathcal{W}$, we then get

$$\big|f(x)-f(y)\big|\leq\|f\|_{\mathrm{LIP}}\,d(x,y)\qquad(x,y\in\mathcal{X}).\tag{5.11}$$

The following lemma says that for any $f\in\mathbb{R}^{\mathcal{X}}$ and probability measures μ,ν on $\mathcal{X}$, one has

$$\big|\mu f-\nu f\big|\leq\mathcal{W}(\mu,\nu)\,\|f\|_{\mathrm{LIP}},\tag{5.12}$$

and $\mathcal{W}(\mu,\nu)$ is the optimal constant for which this inequality holds.

Theorem 5.3 (Wasserstein distance and Lipshitz functions)**.** *For probability measures μ, ν on $\mathcal{X}$, one has*

$$\mathcal{W}(\mu,\nu) = \sup_{\|f\|_{\mathrm{LIP}} \leq 1} \big|\mu f - \nu f\big|. \tag{5.13}$$

Proof Let γ be an optimal coupling for μ and ν and let (X,Y) have law γ. Then for any $f \in \mathbb{R}^{\mathcal{X}}$,

$$\begin{aligned}\Big|\mathbb{E}\big[f(X)\big] - \mathbb{E}\big[f(Y)\big]\Big| &= \Big|\mathbb{E}\big[f(X) - f(Y)\big]\Big| \leq \mathbb{E}\big[|f(X) - f(Y)|\big] \\ &\leq \|f\|_{\mathrm{LIP}}\,\mathbb{E}\big[\mathrm{dist}(X,Y)\big] = \mathcal{W}(\mu,\nu)\,\|f\|_{\mathrm{LIP}}.\end{aligned} \tag{5.14}$$

This proves (5.12) and the inequality $\geq$ in (5.13). The opposite inequality is a bit deeper. Recall that $\mathcal{W}(\mu,\nu)$ is the minimum of the linear function in (5.5) over the compact and convex set $\Pi(\mu,\nu)$. Let

$$\Delta := \sup_{x,y \in \mathcal{X}} \mathrm{dist}(x,y) \tag{5.15}$$

denote the diameter of the graph $(\mathcal{X}, E)$. We claim that $\Delta - \mathcal{W}(\mu,\nu)$ is the maximum of the function

$$\gamma \mapsto \sum_{x,y} \gamma(x,y)\big[\Delta - \mathrm{dist}(x,y)\big] \tag{5.16}$$

subject to the constraints

$$\gamma \geq 0,\ \sum_y \gamma(x,y) \leq \mu(x)\ \forall x,\ \sum_x \gamma(x,y) \leq \nu(y)\ \forall y. \tag{5.17}$$

Indeed, if we find some γ that satisfies the constraints (5.17) that satisfies strict inequality $\sum_{x,y}\gamma(x,y) < 1$, then we can make the function in (5.16) larger by making γ larger, so the optimal γ must be a probability measure which implies that $\sum_y \gamma(x,y) = \mu(x)$ and $\sum_x \gamma(x,y) = \nu(y)$, i.e., its first and second marginals are μ and ν. Comparing with the definition of the Wasserstein distance, we see that the maximum of the function in (5.16) subject to the constraints (5.17) is indeed $\Delta - \mathcal{W}(\mu,\nu)$.

It will be useful to cast (5.16) and (5.17) in a more abstract form. For each $x, y, z \in \mathcal{X}$ and $i = 1, 2$, we define

$$\rho(z,i) := \begin{cases} \mu(z) & \text{if } i = 1, \\ \nu(z) & \text{if } i = 2. \end{cases} \quad \text{and} \quad A(z,i;x,y) := \begin{cases} 1_{\{z=x\}} & \text{if } i = 1, \\ 1_{\{z=y\}} & \text{if } i = 2. \end{cases} \tag{5.18}$$

Then we can rewrite the constraints (5.17) as

$$\gamma \geq 0, \quad \sum_{x,y} A(z,i;x,y)\gamma(x,y) \leq \rho(z,i). \tag{5.19}$$

We can now apply the strong duality theorem of linear programming (Theorem 5.4 below) to conclude that $\Delta - \mathcal{W}(\mu,\nu)$ is the minimum of the function

$$g \mapsto \sum_{z,i} g(z,i)\rho(z,i) \tag{5.20}$$

subject to the constraints

$$g \geq 0, \ \sum_{z,i} A(z,i;x,y)g(z,i) \geq \Delta - \mathrm{dist}(x,y) \ \forall x,y. \tag{5.21}$$

Let us write $g_i(z) := g(z,i)$ and define $f(z) := \Delta - g_2(z)$. Then $\Delta - \mathcal{W}(\mu,\nu)$ is the minimum of the expression

$$\sum_x \mu(x)g_1(x) + \Delta - \sum_x \nu(x)f(x) \tag{5.22}$$

subject to the constraints

$$g_1 \geq 0, \ f \leq \Delta, \ f(y) - g_1(x) \leq \mathrm{dist}(x,y) \ \forall x,y. \tag{5.23}$$

Therefore, $\mathcal{W}(\mu,\nu)$ is the maximum of the expression

$$\sum_x \nu(x)f(x) - \sum_x \mu(x)g(x) \tag{5.24}$$

subject to the constraints

$$g \geq 0, \ f \leq \Delta, \ f(y) - g(x) \leq \mathrm{dist}(x,y) \ \forall x,y. \tag{5.25}$$

Forgetting a constraint will only make the maximum larger, so $\mathcal{W}(\mu,\nu)$ is less or equal than the maximum of the expression

$$\sum_x \nu(x)f(x) - \sum_x \mu(x)g(x) \tag{5.26}$$

subject to the constraints

$$g \geq 0, \ f(y) - g(x) \leq \mathrm{dist}(x,y) \ \forall x,y. \tag{5.27}$$

Since $\mathrm{dist}(x,x) = 0$, these constraints force $f \leq g$. Making f larger will only increase the expression in (5.26), so it suffices to take the maximum over all pairs (f, g) for which $f = g$. Thus, $\mathcal{W}(\mu, \nu)$ is less or equal than the maximum of the expression

$$\sum_x \nu(x) f(x) - \sum_x \mu(x) f(x) \tag{5.28}$$

subject to the constraints

$$f \geq 0, \;\; f(y) - f(x) \leq \operatorname{dist}(x, y) \;\forall x, y. \tag{5.29}$$

Adding a constant to f does not change the expression in (5.28), so we can forget about the constraint $f \geq 0$. Reversing the roles of x and y, our only remaining constraint of course also implies $-\big(f(y) - f(x)\big) = f(x) - f(y) \leq \operatorname{dist}(y, x) = \operatorname{dist}(x, y)$, so

$$\mathcal{W}(\mu, \nu) \leq \sup_{\|f\|_{\mathrm{LIP}} \leq 1} \nu f - \mu f = \sup_{\|f\|_{\mathrm{LIP}} \leq 1} \big|\nu f - \mu f\big|, \tag{5.30}$$

where the equality follows from the fact that we can always replace f by $-f$. The opposite inequality for $\mathcal{W}(\mu, \nu)$ had already been proved, so we conclude that (5.13) holds. ∎

Below is the strong duality theorem of linear programming.

Theorem 5.4 (Strong duality). *Let $A(i, j)_{1 \leq i \leq n,\ 1 \leq j \leq m}$ be a real matrix, and let $\big(b(1), \ldots, b(m)\big)$ and $\big(c(1), \ldots, c(m)\big)$ be real vectors. Assume that the function $x \mapsto \sum_{j=1}^{m} c(j) x(j)$ assumes its maximum M_+ over the set*

$$\Big\{x \in \mathbb{R}^m : x(j) \geq 0 \;\forall 1 \leq j \leq m,\; \sum_{j=1}^{m} A(i, j) x(j) \leq b(i) \;\forall 1 \leq i \leq n\Big\}. \tag{5.31}$$

Then the function $y \mapsto \sum_{i=1}^{n} b(i) y(i)$ assumes its minimum M_- over the set

$$\Big\{y \in \mathbb{R}^n : y(i) \geq 0 \;\forall 1 \leq i \leq n,\; \sum_{i=1}^{n} A(i, j) y(i) \geq c(j) \;\forall 1 \leq j \leq m\Big\}, \tag{5.32}$$

and one has $M_- = M_+$.

6 Curvature

We continue to assume that P is irreducible and reversible. The (Ollivier-Ricci) *curvature* of P is the quantity $\kappa(P) \leq 1$ defined as

$$\kappa(P) := 1 - \sup_{\{x, y\} \in E} \mathcal{W}\big(P(x, \cdot\,), P(y, \cdot\,)\big), \tag{6.1}$$

where E is the set of edges defined in (5.1). A Markov chain is said to have *positive curvature* if $\kappa(P) > 0$. In view of Theorem 5.3 and (5.10),

$$\begin{aligned} 1-\kappa(P) &= \sup_{\{x,y\}\in E}\ \sup_{\|f\|_{\mathrm{LIP}}\le 1} \big|Pf(x)-Pf(y)\big| \\ &= \sup_{\|f\|_{\mathrm{LIP}}\le 1} \|Pf\|_{\mathrm{LIP}}. \end{aligned} \tag{6.2}$$

For positively curved Markov chains, in the light of (2.6) and (5.4), the following lemma gives a bound on the total variation distance to equilibrium.

Lemma 6.1 (Curvature bound)**.** *For any probability measures μ,ν on $\mathcal{X}$, one has*

$$\begin{aligned} &\text{(i)}\ \ \mathcal{W}(\mu P^n,\nu P^n) \le \big(1-\kappa(P)\big)^n\,\mathcal{W}(\mu,\nu) \qquad (n\ge 0),\\ &\text{(ii)}\ \ \mathcal{W}(\mu P_t,\nu P_t) \le e^{-\kappa(P)t}\,\mathcal{W}(\mu,\nu) \qquad (n\ge 0) \end{aligned} \tag{6.3}$$

Proof In the discrete time setting this is proved in [Sal22, Lemma 20]. Note, however, that the notation there is a bit different. What we call $1-\kappa(P)$ is called $e^{-\kappa(P)}$ there. Our definition is more suitable for the continuous-time setting and coincides with the definition in [Sal21].

To prove (6.3) (i), we use (6.2) and induction to obtain

$$\|P^n f\|_{\mathrm{LIP}} \le \big(1-\kappa(P)\big)^n \|f\|_{\mathrm{LIP}} \qquad (n\ge 0). \tag{6.4}$$

By Theorem 5.3 and (5.12), it follows that

$$\begin{aligned} \mathcal{W}(\mu P^n,\nu P^n) &= \sup_{\|f\|_{\mathrm{LIP}}\le 1} \big|\mu P^n f-\nu P^n f\big| \\ &\le \sup_{\|f\|_{\mathrm{LIP}}\le 1} \mathcal{W}(\mu,\nu)\|P^n f\|_{\mathrm{LIP}} \le \big(1-\kappa(P)\big)^n\mathcal{W}(\mu,\nu), \end{aligned} \tag{6.5}$$

proving (6.3) (i). To prove also (6.3) (ii), we use (1.6), which says that

$$P_t = \sum_{n=0}^{\infty} p_t(n)P^n \quad \text{with} \quad p_t(n) := e^{-t}\frac{t^n}{n!} \quad (n\ge 0). \tag{6.6}$$

The Lipschitz "norm" is not really a norm but only a pseudonorm. Using the triangle inequality for this pseudonorm, as well as the continuity of the map $f\mapsto \|f\|_{\mathrm{LIP}}$, we obtain

$$\begin{aligned} \|P_t f\|_{\mathrm{LIP}} &= \Big\|\sum_{n=0}^{\infty} p_t(n)P^n\Big\|_{\mathrm{LIP}} \le \sum_{n=0}^{\infty} p_t(n)\big\|P^n\big\|_{\mathrm{LIP}} \\ &\le \sum_{n=0}^{\infty} p_t(n)(1-\kappa(P))^n\|f\|_{\mathrm{LIP}} = e^{-\kappa(P)t}\|f\|_{\mathrm{LIP}}. \end{aligned} \tag{6.7}$$

The final equality here is not completely obvious. Recall that $\kappa(P) \leq 1$ by definition. If $\kappa(P) \geq 0$, then we can interpret $\kappa = \kappa(P)$ as a probability. Now $\sum_{n=0}^{\infty} p_t(n)(1-\kappa)^n$ is the probability that if we have a random number of particles with a Poisson distribution with mean t, and we perform a random thinning of these particles, where each particle has an independent probability κ that we keep it, then no particles survive the thinning. Now it is well-known that after thinning, the number of particles that is left is Poisson distributed with mean κt. Thus $p_{\kappa t}(0) = e^{-\kappa t}$ is the probability that no particles survive the thinning. For general κ, one can verify the final equality in (6.7) by direct computation. We leave this as an exercise to the reader. Formula (6.3) (ii) follows from (6.7) in the same way we derived (6.3) (i) from (6.4). ∎

7 Cut-off

For any probability kernel P, we define the *sparsity parameter* $\Delta(P)$ by

$$\Delta(P) := \max_{\{x,y\} \in E} \frac{1}{P(x,y)} \tag{7.1}$$

For positive functions f_n, g_n, we write $f_n \ll g_n$ as $n \to \infty$ if $f_n / g_n \to 0$. Recall the definitions of the mixing and relaxation times (in the continuous-time setting) in (2.7) and (2.15). We set $\hat{t}_{\rm MIX}(P_n) := \hat{t}^{(1/4)}_{\rm MIX}(P_n)$. Below is a simplified version of the main result of [Sal21].

Theorem 7.1 (Conditions for cut-off)**.** *Let P_n be a sequence of irreducible reversible probability kernels on finite sets $\mathcal{X}_n$. Assume that $|\mathcal{X}_n| \geq 3$ and $\kappa(P_n) \geq 0$ for all n. Assume that for each $\varepsilon \in (0,1)$,*

$$\big(\hat{t}_{\rm REL}(P_n) \log \Delta(P_n)\big)^2 \ll \hat{t}^{(\varepsilon)}_{\rm MIX}(P_n) \qquad \textit{as } n \to \infty. \tag{7.2}$$

Then for each $0 < \varepsilon < \frac{1}{2}$, there exists a constant C_ε such that

$$\hat{t}^{(\varepsilon)}_{\rm MIX}(P_n) - \hat{t}^{(1-\varepsilon)}_{\rm MIX}(P_n) \leq C_\varepsilon \sqrt{\hat{t}_{\rm MIX}(P_n)}\, \hat{t}_{\rm REL}(P_n) \log \Delta(P_n) \tag{7.3}$$

for all n large enough.

Remark 1 Assmption (7.2) implies that $\hat{t}_{\rm REL}(P_n) \log \Delta(P_n) \ll \sqrt{\hat{t}_{\rm MIX}(P_n)}$ so that the right-hand side of (7.3) satisfies

$$\sqrt{\hat{t}_{\rm MIX}(P_n)}\, \hat{t}_{\rm REL}(P_n) \log \Delta(P_n) \ll \hat{t}_{\rm MIX}(P_n). \tag{7.4}$$

This shows that the sequence of Markov chains with transition kernels P_n exhibits *cut-off*: the total variation distance $\hat{\mathcal{D}}_{P_n}(t)$ changes from being close to one to close to zero in a time interval centered around $\hat{t}_{\mathrm{MIX}}(P_n)$ whose duration is in the limit much shorter than $\hat{t}_{\mathrm{MIX}}(P_n)$.

Remark 2 In these notes, we use a hat to indicate that a certain quantity belongs to the continuous-time setting. This is why our notation is a bit different from the notation in [Sal21].

Remark 3 The main result of [Sal21] is considerably more general than Theorem 7.1. The condition $\kappa(P_n) \geq 0$ says that P_n has non-negative Ollivier-Ricci curvature. In [Sal21], it is shown that the result remains true if Ollivier-Ricci curvature is replaced by Bakry-Émery curvature, a concept that we have not treated here. The main result of [Sal21] moreover applies to a large class of non-reversible chains as well. In that case, however, the relaxation time from Theorem 7.1 has to be replaced by the relaxation time of the symmetrised kernel $(P + P^*)/2$ and one needs the additional assumption that $P(x, y) > 0$ if and only if $P(y, x) > 0$.

Remark 4 The assmption (7.2) is inspired by the so-called *product condition* $\hat{t}_{\mathrm{REL}}(P_n) \ll \hat{t}^{(\varepsilon)}_{\mathrm{MIX}}(P_n)$ $(\varepsilon \in (0, 1))$, that has in the past been conjectured to imply cut-off. This conjecture is known to be wrong, however. Typical counterexamples, however, are extremely non-sparse, in the sense that the Markov chain can jump from any point in the state space to any other point with positive probability. For these counterexamples, $\Delta(P_n)$ would grow very fast with n and (7.2) would not be satisfied. This justifies the occurrence of $\Delta(P_n)$ in (7.2).

Remark 5 The occurrence of the square in (7.2) is a less pleasant feature of the theorem. It seems natural that in order to prove a quantitative statement like (7.3) on the size of the "critical window" where $\hat{\mathcal{D}}_{P_n}(t)$ changes from being close to one to being close to zero, one needs a quantitive version of the product condition, i.e., one needs to say *how much larger* $\hat{t}^{(\varepsilon)}_{\mathrm{MIX}}(P_n)$ is in comparison with $\hat{t}_{\mathrm{REL}}(P_n)$. However, (7.2) is quite a strong assumption in the sense that even when $\Delta(P_n)$ is of order one, it requires $\hat{t}^{(\varepsilon)}_{\mathrm{MIX}}(P_n)$ to be much larger than the *square* of $\hat{t}_{\mathrm{REL}}(P_n)$ (for example, $\hat{t}^{(\varepsilon)}_{\mathrm{MIX}}(P_n)$ could be of order $(\hat{t}_{\mathrm{REL}}(P_n))^{2+\varepsilon}$ for any $\varepsilon > 0$). In [Sal22, Section 5] we have seen that the random walk on the hypercube has cut-off with $t_{\mathrm{REL}}(P_n) \sim n$ and $t_{\mathrm{MIX}}(P_n) \sim \frac{1}{2} n \log n$. In this example, the mixing time grows only a little bit faster than the relaxation time (the difference is the logaritmic term) so this example cannot be covered by Theorem 7.1. In future, one would hope to relax condition (7.2) in this respect..

Remark 6 In order to check assumption (7.2) of Theorem 7.1, one only needs upper bounds on $\hat{t}_{\mathrm{REL}}(P_n)$ and lower bounds on $\hat{t}^{(\varepsilon)}_{\mathrm{MIX}}(P_n)$. In particular, one does *not* need to determine the precise asymptotics of $\hat{t}^{(\varepsilon)}_{\mathrm{MIX}}(P_n)$, including the preconstant, which is often very difficult. Theorem 7.1 is one of very few results of this sort and the first one that applies to a very general class of probability kernels, instead of only to very special classes such as birth-and-death chains etc. In this sense, Theorem 7.1 is a breakthrough result.

8 Strategy of the proof

The *relative entropy* (or *Kullback-Leibler divergence*) or a probability measure μ on a finite set $\mathcal{X}$ with respect to another probability measure π is defined as

$$d_{\mathrm{KL}}(\mu\|\pi) := \sum_{x\in\mathcal{X}} \mu(x)\log\frac{\mu(x)}{\pi(x)}. \tag{8.1}$$

Here, for simplicity, we assume that $\pi(x) > 0$ everywhere. A related quantity, relatively unknown, the *varentropy*, is defined as

$$\mathcal{V}_{\mathrm{KL}}(\mu\|\pi) := \sum_{x\in\mathcal{X}} \mu(x)\Big(\log\frac{\mu(x)}{\pi(x)} - d_{\mathrm{KL}}(\mu\|\pi)\Big)^2. \tag{8.2}$$

Note that

$$\mathcal{V}_{\mathrm{KL}}(\mu\|\pi) = \mathrm{Var}_\mu(f) \quad \text{with} \quad f(x) := \log\frac{\mu(x)}{\pi(x)}, \tag{8.3}$$

since clearly $\mu f = d_{\mathrm{KL}}(\mu\|\pi)$. Following [Sal21], we define

$$\mathcal{V}^\star_{\mathrm{KL}}(t) := \sup_{x\in\mathcal{X}} \mathcal{V}_{\mathrm{KL}}(P_t(x,\,\cdot\,)\|\pi). \tag{8.4}$$

The following theorem, which is [Sal21, Thm 5], is key to the proof of Theorem 7.1.

Theorem 8.1 (Entropic concentration). *Let P be an irreducible probability kernel on a finite set $\mathcal{X}$ and let $0 < \varepsilon < \frac{1}{2}$. Then*

$$\hat{t}^{(\varepsilon)}_{\mathrm{MIX}}(P) - \hat{t}^{(1-\varepsilon)}_{\mathrm{MIX}}(P) \leq 2\varepsilon^{-2}\hat{t}_{\mathrm{REL}}(P)\Big[1 + \sqrt{\mathcal{V}^\star_{\mathrm{KL}}(\hat{t}^{(1-\varepsilon)}_{\mathrm{MIX}}(P))}\Big]. \tag{8.5}$$

The second incredient of the proof of Theorem 7.1 is [Sal21, Thm 5], which we cite here in simplified form.

Theorem 8.2 (Varentropy estimate)**.** *Let P_n be a sequence of irreducible reversible probability kernels with $\kappa(P_n) \geq 0$. Fix $0 < \varepsilon < 1$ and assume that*

$$\sqrt{\hat{t}_{\mathrm{REL}}(P_n)} \ll \hat{t}^{(\varepsilon)}_{\mathrm{MIX}}(P_n) \qquad \text{as } n \to \infty. \tag{8.6}$$

Then there exists a constant $C < \infty$ such that

$$\mathcal{V}^{\star}_{\mathrm{KL}}\big(\hat{t}^{(\varepsilon)}_{\mathrm{MIX}}(P_n)\big) \leq C\,\hat{t}^{(\varepsilon)}_{\mathrm{MIX}}(P_n)\,\big(\log\Delta(P_n)\big)^2 \tag{8.7}$$

for all n large enough.

We now show how Theorems 8.1 and 8.2 imply Theorem 7.1.

Proof of Theorem 7.1 We first make some trivial observations. Since $|\mathcal{X}_n| \geq 3$ and P_n is irreducible and reversible, there must be at least one $x \in \mathcal{X}$ that has degree two in the graph $(\mathcal{X}_n, E_n)$, where $E_n = \{\{x,y\} : P_n(x,y) > 0\}$. It follows that $P(x,y) \leq \frac{1}{2}$ for at least one $\{x,y\} \in E_n$ and hence $\Delta(P_n) \geq 2$ for all n. Since P_n is a probability measure, $|\lambda| \leq 1$ for all eigenvalues λ of P_n. Since P_n is irreducible, the multiplicity of the eigenvalue one is one, so $-1 \leq \Re(\lambda_2(P_n)) < 1$, which by (2.15) implies $\frac{1}{2} \leq \hat{t}_{\mathrm{REL}}(P_n) < \infty$. Thus

$$\Delta(P_n) \geq 2 \quad \text{and} \quad \hat{t}_{\mathrm{REL}}(P_n) \geq \tfrac{1}{2} \quad \forall n. \tag{8.8}$$

In view of this, (7.2) implies that

$$\hat{t}^{(\varepsilon)}_{\mathrm{MIX}}(P_n) \to \infty \quad \text{as } n \to \infty\ \forall 0 < \varepsilon < 1. \tag{8.9}$$

Fix $0 < \varepsilon < \frac{1}{2}$. Since $\Delta(P_n) \geq 2$, (7.2) implies that

$$\begin{aligned} &\big(\hat{t}_{\mathrm{REL}}(P_n)\big)^2 \ll \hat{t}^{(\varepsilon)}_{\mathrm{MIX}}(P_n) \\ \Rightarrow\quad &\sqrt{\hat{t}_{\mathrm{REL}}(P_n)} \ll \hat{t}^{(\varepsilon)}_{\mathrm{MIX}}(P_n)^{1/4} \ll \hat{t}^{(\varepsilon)}_{\mathrm{MIX}}(P_n), \end{aligned} \tag{8.10}$$

where in the last step we have used (8.9). This shows that Theorem 8.2 is applicable. Using moreover Theorem 8.1, inserting (8.7) into (8.5), we obtain

$$\begin{aligned} \hat{t}^{(\varepsilon)}_{\mathrm{MIX}}(P_n) - \hat{t}^{(1-\varepsilon)}_{\mathrm{MIX}}(P_n) &\leq 2\varepsilon^{-2}\hat{t}_{\mathrm{REL}}(P_n)\Big[1 + \sqrt{C\,\hat{t}^{(1-\varepsilon)}_{\mathrm{MIX}}(P_n)\,\big(\log\Delta(P_n)\big)^2}\Big] \\ &\leq 2\varepsilon^{-2}\hat{t}_{\mathrm{REL}}(P_n)\Big[1 + \sqrt{C}\,\sqrt{\hat{t}^{(1-\varepsilon)}_{\mathrm{MIX}}(P_n)}\,\log\Delta(P_n)\Big]. \end{aligned} \tag{8.11}$$

In view of (8.8) and (8.9), the expression under the square root tends to infinity, so for sufficiently large n we can forget about the term that is 1, at the cost of making the constant $\sqrt{C}$ a bit larger. Thus, we see that for some $C_\varepsilon < \infty$

$$\hat{t}^{(\varepsilon)}_{\mathrm{MIX}}(P_n) - \hat{t}^{(1-\varepsilon)}_{\mathrm{MIX}}(P_n) \le C_\varepsilon \sqrt{\hat{t}^{(1-\varepsilon)}_{\mathrm{MIX}}(P_n)}\, \hat{t}_{\mathrm{REL}}(P_n)\, \log \Delta(P_n). \tag{8.12}$$

This is almost the same as (7.3), except that in the right-hand side we need to replace $\hat{t}^{(1-\varepsilon)}_{\mathrm{MIX}}(P_n)$ by $\hat{t}_{\mathrm{MIX}}(P_n)$, which is defined as $\hat{t}^{(1/4)}_{\mathrm{MIX}}(P_n)$. Assmption (7.2) implies that $\hat{t}_{\mathrm{REL}}(P_n) \log \Delta(P_n) \ll \sqrt{\hat{t}^{(1-\varepsilon)}_{\mathrm{MIX}}(P_n)}$. Inserting this into (8.12) yields

$$\hat{t}^{(\varepsilon)}_{\mathrm{MIX}}(P_n) - \hat{t}^{(1-\varepsilon)}_{\mathrm{MIX}}(P_n) \ll \hat{t}^{(1-\varepsilon)}_{\mathrm{MIX}}(P_n). \tag{8.13}$$

This shows that the chain exhibits cut-off, and hence, for each $0 < \varepsilon < \frac{1}{2}$,

$$\frac{\hat{t}^{(\varepsilon)}_{\mathrm{MIX}}(P_n)}{\hat{t}^{(1-\varepsilon)}_{\mathrm{MIX}}(P_n)} \underset{n\to\infty}{\longrightarrow} 1 \tag{8.14}$$

As a result, in the right-hand side of (8.12), it asymptotically does not matter if we write $\hat{t}^{(1-\varepsilon)}_{\mathrm{MIX}}(P_n)$ or $\hat{t}^{(1/4)}_{\mathrm{MIX}}(P_n)$. We can replace one by the other and the inequality will remain true for large n, at the cost of perhaps having to change the constant C_ε a bit. ∎

References

[Sal21] Justin Salez. Cutoff for non-negatively curved Markov chains. Preprint, 2021, arXiv:2102.05597v2.

[Sal22] Justin Salez. *Mixing Times of Markov Chains* Lecture notes, 2022.