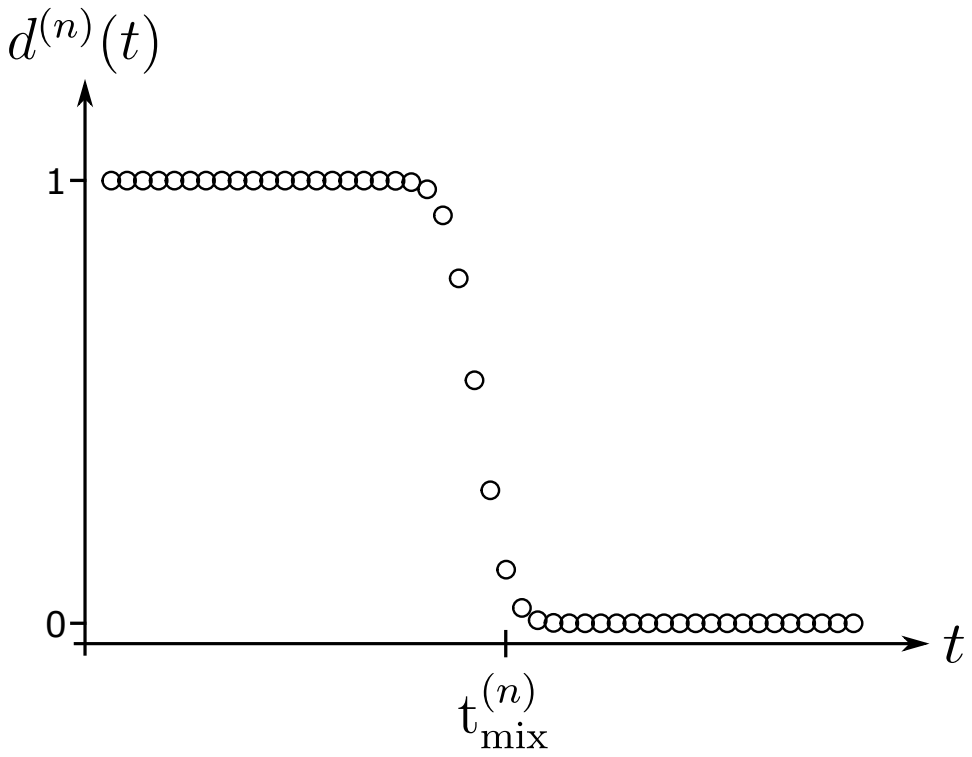
Technische Universität Graz

Bachelor's Thesis:

The Cutoff Phenomenon in Finite Ergodic Markov Chains



Tristan Repolusk

Supervisor: Univ.-Prof. Dipl.-Ing. Dr.rer.nat. Wolfgang Woess

18th June 2018

Contents

1 Abstract

A Markov chain shows *cutoff*, if the approximation of the chain's stationary distribution moves from "bad" to "good" in a short period of time. In this Bachelor thesis, important fundamental results about Markov chains are recalled, notions of distance of the chain's current distribution to the stationary measure are coined, and the basic theory of mixing times and cutoff is presented and developed coherently and rigorously.

2 Introduction

In 1981, Persi Diaconis and Mehrdad Shahshahani discovered the cutoff phenomenon, not yet explicitly naming it, while studying random transpositions on the symmetric group, by investigating the question, "How many transpositions are needed until the permutation is close to random?", in [1, p. 1]. In 1988, P. Diaconis further developed the notion of cutoff, describing it in [2, p. 81: (8)] as "the existence of sharp phase transition [...] cutting down from 1 to zero in a relatively short time. It would be great to understand if this usually happens." In 1996, he published an article about cutoff in finite Markov chains, in which a formal definition is given. From then to now, numerous papers investigating cutoff phenomena have been published, multiple definitions are used, and even the idea of cutoff itself has been generalised to weaker phenomenons like pre-cutoff.

3 Preliminaries

In order to be able to deal with cutoff, we first have to recall some important results concerning Markov chains.

3.1 Basic Definitions

Definition 3.1. *A time homogenous* Markov chain *is a sequence of random variables* $(X_n)_{n\in\mathbb{N}_0}$ *in a state space* $\mathfrak{X}$ *such that the* Markov property *is fulfilled, i.e. that for any* $x, y \in \mathfrak{X}$*, for all* $k \in \mathbb{N}_0$ *and for any tuple* $(x_0, x_1, \ldots, x_{k-1}) \subseteq \mathfrak{X}$ *we have that*

$$\mathbb{P}\left[X_{k+1} = y \,|\, X_k = x,\ X_i = x_i\,(i \in [\![0, k-1]\!])\right] = \mathbb{P}[X_{k+1} = y \,|\, X_k = x] =: p(x, y).$$

Markov chains are a very useful and important means of modelling natural processes. We will denote a time homogeneous Markov chain as a triple $(\mathfrak{X}, P, \mu)$, where $\mathfrak{X}$ symbolises the chain's *finite* state space, $P = [p(x, y)]_{x,y\in\mathfrak{X}}$ its transition matrix, and optionally μ its initial probability distribution. Further, we will write the k-step transition probability for some $x, y \in \mathfrak{X}$ as $p^k(x, y)$. For $k \in \mathbb{N}^*$, we will furthermore write the distribution of X_k contingent on an initial distribution μ as $\mathbb{P}^k_\mu$. In case that $\mu = \delta_x$ is the Dirac measure on some state $x \in \mathfrak{X}$ (this means that the chain almost surely starts at x), we will denote this as $\mathbb{P}^k_x = p^k(x, \cdot)$. Also, we will shorten the notation by defining $\mathbb{P}^k_x(x) := \mathbb{P}^k_x(\{x\})$. Let us recall some other important generic properties of Markov chains and measures:

Definition 3.2. *A Markov chain* $(\mathfrak{X}, P)$ *is called* irreducible *if for any two states* $x, y \in \mathfrak{X}$ *there exists a path from* x *to* y *with non-zero probability, i.e. that there exists some* $k \in \mathbb{N}^*$ *such that* $p^k(x, y) > 0$.

Definition 3.3. *The* period $d(x)$ *of a state* $x \in (\mathfrak{X}, P)$ *is defined as following:* $d(x) := gcd\{n \in \mathbb{N}^* : p^n(x, x) > 0\}$. *We call the state* x aperiodic, *if* $d(x) = 1$. *The chain is said to be* aperiodic *if every state is aperiodic. Given that* $(\mathfrak{X}, P)$ *is irreducible,* $d = d(x)$ *is independent of* x.

Definition 3.4. *We call a chain* $(\mathfrak{X}, P)$ lazy, *if* $p(x, x) \geq \frac{1}{2}$ *for all* $x \in \mathfrak{X}$.

Lemma 3.5. *A lazy Markov chain* $(\mathfrak{X}, P)$ *is aperiodic.*

Proof. Let $x \in \mathfrak{X}$ be an arbitrary state. Due to the chain's laziness, it holds that $p(x, x) > \frac{1}{2} > 0$. Hence, $d(x) = \gcd\{n \in \mathbb{N} : p^n(x, x) > 0\} = 1$. □

Lazy Markov chains are very convenient, since they are always aperiodic. Because of that, it is often very useful to investigate the lazified Markov chain $(\mathfrak{X}, P_L)$ instead of the chain $(\mathfrak{X}, P)$, where $P_L := \frac{1}{2}(P + I)$. As a convex combination of stochastic matrices, P_L is again stochastic and therefore, the lazified chain is well-defined.

3.2 Existence of the Stationary Distribution

Definition 3.6. *A measure* μ *on* $\mathfrak{X}$ *is called* stationary *or* invariant, *if* $\mu P = \mu$, *i.e. for all* $y \in \mathfrak{X}$ *we have* $\sum_{x \in \mathfrak{X}} \mu(x)p(x, y) = \mu(y)$. *A stationary probability distribution, that is a stationary measure fulfilling* $\mu(\mathfrak{X}) = 1$, *is sometimes also called* equilibrium measure.

Under relatively general assumptions, Markov chains show convergence to a unique stationary distribution. For irreducible Markov chains, there is a well-known result for the existence of a stationary distribution, whose proof can be found in most books about stochastic processes, for example in [7, p. 201: Thm. 10.25]:

Theorem 3.7. *An irreducible Markov chain* $(\mathfrak{X}, P)$ *possesses a stationary distribution (this is an invariant probability measure)* π *iff* $(\mathfrak{X}, P)$ *is positive recurrent. In this case,* $\pi(x) = \frac{1}{\mathbb{E}_x(t^x)} > 0$ *for any* $x \in \mathfrak{X}$, *where* $t^x := \inf\{k \in \mathbb{N}^* : X_k = x\}$ *denotes the first return time in the state* x.

Now we want to directly show the uniqueness of the stationary distribution for irreducible chains:

Theorem 3.8. *If a finite, irreducible Markov chain* $(\mathfrak{X}, P)$ *has a stationary distribution* π, *then* π *is unique.*

Proof. Let π_1 and π_2 are two stationary distributions on $(\mathfrak{X}, P)$. Because $\mathfrak{X}$ is finite, we can choose a minimising state $x \in \mathfrak{X}$ such that $x = \text{argmin}_{z \in \mathfrak{X}} \frac{\pi_1(z)}{\pi_2(z)}$ and we define $c := \frac{\pi_1(x)}{\pi_2(x)}$ as the minimum value of said expression. Then it holds that

$$\pi_1(x) = \frac{\pi_1(x)}{\pi_2(x)}\pi_2(x) = c\pi_2(x). \tag{1}$$

Because of the definition of c, $\frac{\pi_1(z)}{\pi_2(z)} \geq c$ is true for any $z \in \mathfrak{X}$. This is equivalent to

$$\forall z \in \mathfrak{X} : \pi_1(z) \geq c\pi_2(z) \tag{2}$$

Due to the fact, that the chain is finite and irreducible, we can find some $n \in \mathbb{N}^*$ such that for any $x, y \in \mathfrak{X}$ there exists a $k \in [\![1, n]\!]$ with $p^k(x, y) > 0$ Since both measures π_1 and π_2 are stationary, we have that $\pi_1 = P^k\pi_1$ and $\pi_2 = P^k\pi_2$. Hence we have for fixed $k \in [\![1, n]\!]$:

$$\begin{aligned} \pi_1(x) = \sum_{z\in\mathfrak{X}} \pi_1(z)p^k(z, x) &\overset{(2)}{\geq} c \sum_{z\in\mathfrak{X}} \pi_2(z)p^k(z, x) = c\pi_2(x) \overset{(1)}{=} \pi_1(x) \\ \Rightarrow \sum_{z\in\mathfrak{X}} \pi_1(z)p^k(z, x) &= c \sum_{z\in\mathfrak{X}} \pi_2(z)p^k(z, x). \end{aligned} \tag{3}$$

All terms in the sum are non-negative, because of this we obtain from (3), that $\forall z \in \mathfrak{X}, p^k(x, z) > 0 : \pi_1(z) = c\pi_2(z)$. By applying this to all $k \in [\![1, n]\!]$, we get that $\forall z \in \mathfrak{X} : \pi_1(z) = c\pi_2(z)$, and subsequently by summation over all $z \in \mathfrak{X}$:

$$1 = \sum_{z\in\mathfrak{X}} \pi_1(z) = c \sum_{z\in\mathfrak{X}} \pi_2(z) = c$$

This directly implies that for all $z \in \mathfrak{X}$ we have $\pi_1(z) = c\pi_2(z) = \pi_2(z)$. This is equivalent to $\pi_1 = \pi_2$, which is exactly what we wanted to show.

□

In the next fundamental theorem we will see that *positive recurrent*, *irreducible* and *aperiodic* Markov chains have very nice properties with respect to the stationary distribution. Therefore, we call chains fulfilling these criteria *ergodic*.

Theorem 3.9. *Let $(\mathfrak{X}, P)$ be an ergodic Markov chain with unique stationary distribution π. Then:*

1. *For arbitrary $x, y \in \mathfrak{X}$: $\lim_{k\to\infty} p^k(x, y) = \pi(y)$.*
2. *For any initial distribution μ and for any $y \in \mathfrak{X}$: $\lim_{k\to\infty} \mathbb{P}^k_\mu(y) = \pi(y)$*

Proof. A proof of *1* can be found in [7, p. 199: Cor. 10.21], so we will only prove statement *2*: Let μ be a distribution on $\mathfrak{X}$ and let $y \in \mathfrak{X}$. We observe that $\mathbb{P}^k_\mu(y) = \mu P^k(y) = \sum_{x\in\mathfrak{X}} \mu(x)p^k(x, y)$ and taking the limit according to *1* yields

$$\lim_{k\to\infty} \mathbb{P}^k_\mu(y) = \sum_{x\in\mathfrak{X}} \mu(x) \lim_{k\to\infty} p^k(x, y) = \pi(y) \sum_{x\in\mathfrak{X}} \mu(x) = \pi(y).$$

□

3.3 Stopping Times and Stationary Times

Stationary times are very important for showing cutoff in Markov chains, as we will see in the following sections of this paper.

Definition 3.10. *A sequence $(\mathscr{F}_n)_{n\in\mathbb{N}^*}$ of σ-algebras is called a* filtration*, if $\mathscr{F}_n \subseteq \mathscr{F}_{n+1}$ for any $n \in \mathbb{N}^*$.*

Definition 3.11. *We call a sequence $(X_n)_{n\in\mathbb{N}^*}$ of random variables* adapted *to a filtration $(\mathscr{F}_n)_{n\in\mathbb{N}^*}$ if X_n is $\mathscr{F}_n$-measurable for all $n \in \mathbb{N}^*$. Let us define $(\mathscr{H}_n)_{n\in\mathbb{N}^*} := (\sigma(X_0, X_1, \ldots, X_n))_{n\in\mathbb{N}^*}$ to be the* natural filtration *with respect to $(X_n)_n$.*

Definition 3.12. *A random variable τ which exclusively assumes values in $\mathbb{N}^*$ is called a* stopping time *for a filtration $(\mathscr{F}_n)_{n\in\mathbb{N}^*}$ if for arbitrary $n \in \mathbb{N}^*$ the event $\{\tau = n\} \in \mathscr{F}_n$.*

Definition 3.13. *Let $(X_n)_{n\in\mathbb{N}^*}$ be an ergodic Markov chain adapted to a filtration $(\mathscr{F}_n)_{n\in\mathbb{N}^*}$, with values in $\mathfrak{X}$ and with stationary distribution π. A* stationary time *τ for $(X_n)_n$ is an $(\mathscr{F}_n)_n$-stopping time with the property, that there exists a $x \in \mathfrak{X}$ such that*

$$\forall y \in \mathfrak{X} : \mathbb{P}_x [X_\tau = y] = \pi(y).$$

If the stronger condition

$$\mathbb{P}_x [\tau = n, X_\tau = y] = \mathbb{P}_x [\tau = n] \pi(y)$$

if fulfilled, i.e. X_τ has distribution π and is independent of τ, we say that τ is a strong stationary time.

Lemma 3.14. *Using the notation of the preceding definition for a strong stationary time τ, the following equality holds for any $n \geq 0, y \in \mathfrak{X}$:*

$$\mathbb{P}_x [\tau \leq t, X_n = y] = \mathbb{P}_x [\tau \leq n] \pi(y).$$

Proof. The rather short proof is based on [10, p. 78: Rem. 6.8]. Since π is stationary, it holds for any natural $k \geq 0$ that $\sum_{z\in\mathfrak{X}} p^k(z, y)\pi(z) = \pi(y)$, whence

$$\begin{aligned}\mathbb{P}_x [\tau \leq t, X_n = y] &= \sum_{s=1}^{n} \sum_{z\in\mathfrak{X}} \mathbb{P}_x [\tau = s, X_s = z, X_n = y] \overset{\text{[str. st.]}}{=} \sum_{s=1}^{n} \mathbb{P}_x [\tau = s] \sum_{z\in\mathfrak{X}} p^{n-s}(z, y)\pi(z) \\ &= \sum_{s=1}^{n} \mathbb{P}_x [\tau = s] \pi(y) = \mathbb{P}_x [\tau \leq s] \pi(y)\end{aligned}$$

□

3.4 Eigenvalues of the Transition Matrix

We will first give the definition of eigenvalues and eigenvectors of a Markov chain. Basically, this is merely the application of the well-known concept of eigenvalues in linear algebra applied to the chain's transition matrix. Since we will exclusively work in finite settings later on, all the common results for eigenvalues and eigenvectors are, of course, applicable. In the following pages, unless otherwise stated, we will assume $(\mathfrak{X}, P)$ to be a general Markov chain.

Definition 3.15. *Let $v : \mathfrak{X} \to \mathfrak{X}$ be a function and $\lambda \in \mathbb{C}$. We call v an* eigenfunction *or* eigenvector *of P with corresponding* eigenvalue *λ, if*

$$Pv = \lambda v,$$

with $Pv(x) := \sum_{y \in \mathfrak{X}} P(x,y)v(y)$. Also, we define $\sigma(P) := \{\lambda \in \mathbb{C} : \lambda$ is an eigenvalue of $P\}$.

Lemma 3.16. *The following statements are true:*

1. *If λ is an eigenvalue of P, then $|\lambda| \leq 1$.*
2. *For irreducible P, the eigenspace associated to eigenvalue 1 is the linear hull $\mathcal{L}(\{\mathbb{1}\})$, where $\mathbb{1}$ is the vector whose entries are solely 1.*
3. *If P is both irreducible and aperiodic, then -1 is not an eigenvalue of the chain.*

Proof. An outline of the proof can be found in [10, p. 176: Ex. 12.1]. □

Definition 3.17. *The* spectral gap *of the chain is defined as*

$$\gamma := 1 - \max\{\lambda : \lambda \in \sigma(P)\}.$$

and the absolute spectral gap *as*

$$\gamma_\star := 1 - \max\{|\lambda| : \lambda \in \sigma(P)\},$$

Also, we define the reciprocal of $\gamma_\star$ to be the relaxation time *of the chain,*

$$t_{\text{rel}} := \frac{1}{\gamma_\star}.$$

Corollary 3.18. *In case that the chain is aperiodic and irreducible, $\gamma_\star > 0$ immediately follows from lemma 3.16.*

3.5 Reversibility of Markov Chains

We will see later that a certain class of Markov chains, the so-called reversible chains, admit very interesting properties with respect to cutoff. Hence, we will now define the notion of reversibility:

Definition 3.19. *A Markov chain $(\mathfrak{X}, P)$ is called* reversible*, if there exists a measure μ such that for any $x, y \in \mathfrak{X}$ we have $\mu(x)p(x,y) = \mu(y)p(y,x)$.*

Lemma 3.20. *A reversible finite chain $(\mathfrak{X}, P)$ with respect to a measure $\mu \not\equiv 0$ possesses a stationary distribution $\pi := \frac{1}{\mu(\mathfrak{X})}\mu$.*

Proof. Let $x \in \mathfrak{X}$ be arbitrary. It is clear that π is a probability distribution. We obtain

$$\pi(x) = \frac{1}{\mu(\mathfrak{X})}\mu(x)\underbrace{\sum_{y \in \mathfrak{X}} p(x,y)}_{=1} \overset{\text{(reversibility)}}{=} \frac{1}{\mu(\mathfrak{X})}\sum_{y \in \mathfrak{X}} \mu(y)p(y,x) = \frac{1}{\mu(\mathfrak{X})}\mu P(x) = \pi P(x).$$

□

4 The Cutoff Phenomenon

From now on, unless otherwise stated, we will always assume that $(\mathfrak{X}, P, \mu)$ is an ergodic finite Markov chain with its stationary distribution called π. We have already seen in section 3, that $\mathbb{P}_\mu^k(y)$ converges to $\pi(y)$ for $n \to \infty$.

4.1 Common distances

We want to quantify the distance between $\mathbb{P}_\mu^k(y)$ and $\pi(y)$. In order to do that, one can theoretically use any meaningful distance which is defined on the space of signed measures, i.e. the differences of two ordinary measures. These distances do not necessarily have to be norms or metrics. In practice, however, there are mostly three types of distances used, which we will introduce in this subsection.

4.1.1 The ℓ^p-Distances

For $1 \leq p \leq \infty$ we can define a distance induced by the ℓ^p-norm:

Definition 4.1. *The ℓ^p-distance between to probability measures μ and ν on $\mathfrak{X}$ with respect to a strictly positive measure π on $\mathfrak{X}$, which in our case will always be a chain's stationary distribution, is for $1 \leq p < \infty$ defined as*

$$\|\mu - \nu\|_{p,\pi} := \left(\sum_{x \in \mathfrak{X}} \pi(x)^{1-p} |\mu(x) - \nu(x)|^p \right)^{\frac{1}{p}}$$

and for $p = \infty$ as

$$\|\mu - \nu\|_{\infty,\pi} := \max_{x \in \mathfrak{X}} \frac{|\mu(x) - \nu(x)|}{\pi(x)}.$$

If clear, we will drop π from the notation and simply write $\|\mu - \nu\|_p$ instead of $\|\mu - \nu\|_{p,\pi}$. Also, if we just look at the ℓ^∞-difference between some probability measure μ and π with respect to π, we can simplify the term to $\|\mu - \pi\|_\infty = \max_{x \in \mathfrak{X}} \left| \frac{\mu(x)}{\pi(x)} - 1 \right|$. Cutoff with respect to the ℓ^p-distance for $1 < p \leq \infty$ (note that the case $p = 1$ is excluded) has some interesting properties for lazy and reversible Markov chains, which are mentioned in [9, p. 3: Thm. A].

4.1.2 The Total Variation Distance

One of the most natural ways to quantify a distance between two probability measures is the *total variation*. Early definitions of cutoff have already been using this norm, see for example [3] from 1996. Nevertheless, this type of cutoff is still often used in modern papers regarding cutoff phenomena.

Definition 4.2. *Let μ, ν be two probability measures on $\mathfrak{X}$. The* total variation distance *is defined by*

$$\|\mu - \nu\|_{TV} := \max_{A \subseteq \mathfrak{X}} |\mu(A) - \nu(A)|.$$

We will now see, that the total variation distance is closely interwoven with the ℓ^1-norm, which is easier to handle.

Theorem 4.3. *Let ν, μ be two distributions on $\mathfrak{X}$. Then*

$$\begin{aligned} &1. : \|\mu - \nu\|_{TV} = \frac{1}{2}\|\mu - \nu\|_1 = \frac{1}{2}\sum_{x\in\mathfrak{X}} |\mu(x) - \nu(x)| \\ &2. : \|\mu - \nu\|_{TV} = \sum_{\substack{x\in\mathfrak{X} \\ \nu(x)\leq\mu(x)}} (\mu(x) - \nu(x)) \end{aligned}$$

Proof. The proof is based on [10, p. 48: Prop. 4.2]. Define $B := \{x \in \mathfrak{X} : \mu(x) - \nu(x) \geq 0\}$ and let $A \subseteq \mathfrak{X}$. Since $1 = \mu(\mathfrak{X}) = \mu(B) + \mu(B^C)$ and $1 = \nu(\mathfrak{X}) = \nu(B) + \nu(B^C)$ we get $\mu(B) + \mu(B^C) = \nu(B) + \nu(B^C)$ which is equivalent to

$$\mu(B) - \nu(B) = \nu(B^C) - \mu(B^C). \tag{4}$$

Also, we have that

$$\begin{aligned} \mu(A) - \nu(A) =& \mu(A\cap B) - \nu(A\cap B) + \underbrace{\mu(A\cap B^C) - \nu(A\cap B^C)}_{\leq 0} \leq \mu(A\cap B) - \nu(A\cap B) \\ \leq& \mu(A\cap B) - \nu(A\cap B) + \underbrace{\mu(A^C\cap B) - \nu(A^C\cap B)}_{\geq 0} = \mu(B) - \nu(B). \end{aligned} \tag{5}$$

Analogously, we obtain

$$\nu(A) - \mu(A) \leq \nu(B^C) - \mu(B^C) \overset{(4)}{=} \mu(B) - \nu(B). \tag{6}$$

Note that $\mu(B) - \nu(B) \geq 0$ by definition of B. By this, we get

$$\|\mu - \nu\|_{\mathrm{TV}} = \max_{A\subseteq\mathfrak{X}} |\mu(A) - \nu(A)| \overset{(5)}{=} |\mu(B) - \nu(B)| = \mu(B) - \nu(B) = \sum_{\substack{x\in\mathfrak{X} \\ \nu(x)\leq\mu(x)}} (\mu(x) - \nu(x)),$$

which concludes the proof of 2. Now we prove 1: By the preceding equation, it holds that

$$\begin{aligned} \|\mu - \nu\|_{\mathrm{TV}} =& \mu(B) - \nu(B) = \sum \overset{(6)}{=} \frac{1}{2}[\mu(B) - \nu(B) + \nu(B^C) - \mu(B^C)] \\ =& \frac{1}{2}\left[\sum_{x\in B}(\mu(x) - \nu(x)) + \sum_{x\in B^C}(\nu(x) - \mu(x))\right] \\ =& \frac{1}{2}\left[\sum_{\mu(x)-\nu(x)\geq 0} |\mu(x) - \nu(x)| + \sum_{\mu(x)-\nu(x)<0} |\mu(x) - \nu(x)|\right] \\ =& \frac{1}{2}\sum_{x\in\mathfrak{X}} |\mu(x) - \nu(x)| = \frac{1}{2}\|\mu - \nu\|_{L^1} \end{aligned}$$

□

4.1.3 The Separation Distance

Definition 4.4. *Let again be μ, ν two probability measures on $\mathfrak{X}$. We define the* separation distance *between μ and ν as*

$$\operatorname{sep}(\mu, \nu) := \max_{x \in \mathfrak{X}} \left(1 - \frac{\mu(x)}{\nu(x)}\right)$$

For a strictly positive measure π, $\operatorname{sep}(\mu, \pi)$ looks very similar to the simplified expression of the ℓ^∞-distance. But $\operatorname{sep}(\mu, \pi)$ is not a metric due to its asymmetry, which also reflects in its notation.

Using *strong stationary times*, which were introduced in section 3, we can obtain the following useful result:

Lemma 4.5. *For a chain with strong stationary time τ and arbitrary $k \in \mathbb{N}^*, x \in \mathfrak{X}$ we have that*

$$\operatorname{sep}(\mathbb{P}_x^k, \pi) \leq \mathbb{P}_x[\tau > k].$$

Proof. The proof of this lemma is based on [6, p. 17: Lem. 4.1]. For fixed $x, y \in \mathfrak{X}$ we observe that

$$\mathbb{P}_x^k(y) = \mathbb{P}_x[X_k = y, \tau > k] + \mathbb{P}_x[X_k = y, \tau \leq k] \geq \mathbb{P}_x[X_k = y, \tau \leq k]$$

By this and using the fact that τ is a strong stationary time, it follows that

$$1 - \frac{\mathbb{P}_x^k(y)}{\pi(y)} \leq 1 - \frac{\mathbb{P}_x[X_k = y, \tau \leq k]}{\pi(y)} = 1 - \frac{\mathbb{P}_x[\tau \leq k]\pi(y)}{\pi(y)} = 1 - \mathbb{P}_x[\tau \leq k] = \mathbb{P}_x[\tau > k].$$

By taking the supremum over all $y \in \mathfrak{X}$ we obtain the inequality. □

4.2 Defining Cutoff

Investigating certain sequences of Markov chains, one can observe a sudden decrease of a distance between $\mathbb{P}_\mu^k$ and the stationary distribution π from approximately 1 to nearly 0 in a short time period. This is called the *cutoff phenomenon*. Note that the notion of cutoff is only meaningful for a sequence of chains rather than a single Markov chain.

These sequences $(\mathfrak{X}_k, P_k)_{k \in \mathbb{N}^*}$ usually have strictly monotonically increasing sizes of the state spaces $\mathfrak{X}_k$, and can, for example, show one of the following properties:

- For any $k \in \mathbb{N}^*$, $(\mathfrak{X}_k, P_k)$ are of the same type with similar state spaces $\mathfrak{X}_k$, which are increasing in size. Notable examples include mixing processes (e.g. riffle shuffle on k cards) with each state space $\mathfrak{X}_k = \mathfrak{S}_k$ being the symmetric group over k elements.
- Proceeding from a sequence $(\mathfrak{X}_k, P_k)_{k \in \mathbb{N}^*}$ of Markov chains, one can construct another sequence $(\mathfrak{Y}_k, Q_k)_{k \in \mathbb{N}^*}$, with $\mathfrak{Y}_k := \left(\mathfrak{X}_k^{(1)}, \ldots, \mathfrak{X}_k^{(k)}\right)$ being the product of k independent copies $(\mathfrak{X}_k^{(i)})_{i=1}^k$ of $\mathfrak{X}_k$ with $Q_k((x_1, \ldots, x_k), (y_1, \ldots, y_k)) := \prod_{i=1}^k p_k(x_i, y_i)$. The sequence $(\mathfrak{Y}_k, Q_k)_k$ is called a sequence of product chains.

The occurrence of cutoff phenomena is heavily dependent on both the investigated chains, see for example [9], as well as the distance used to quantify the distribution's deviation from the stationary distribution. Hence, general results regarding cutoff are very difficult to obtain. Due to this, most papers focus on sequences of concrete classes of Markov chains, where cutoff or non-cutoff can be shown explicitly.

In order to rigorously formalise the notion of cutoff with regard to sequences of Markov chains, we first need to introduce *worst-case distances* and *mixing times*. For the following definitions we will mostly use the *total variation distance* or the *separation distance*. Analogously, all these notions may as well be defined for different distances, yielding different types of cutoff.

4.2.1 Worst-Case Distances

Definition 4.6. *For $k \in \mathbb{N}^*$, the* worst-case total variation distance *in k is defined as*

$$d(k) := \max_{x \in \mathfrak{X}} \left\| \mathbb{P}_x^k - \pi \right\|_{\mathrm{TV}}$$

and for $1 \leq p \leq \infty$ the worst-case ℓ^p-distance *in k as*

$${}_p d(k) := \max_{x \in \mathfrak{X}} \left\| \mathbb{P}_x^k - \pi \right\|_p .$$

Note, that $d = \frac{1}{2}\, {}_1 d$. Analogously, we write the worst-case separation distance *in k as*

$$s(k) := \max_{x \in \mathfrak{X}} \operatorname{sep}(\mathbb{P}_x^k, \pi).$$

Also, we define the worst-case starting total variation distance *to be*

$$\bar{d}(k) := \max_{x,y \in \mathfrak{X}} \left\| \mathbb{P}_x^k - \mathbb{P}_y^k \right\|_{\mathrm{TV}} .$$

The following lemma will show us that *worst-case total variation distance* and *worst-case starting total variation distance* are indeed very meaningful and intuitive names for $d(k)$ respectively $\bar{d}(k)$.

Lemma 4.7. *Let us denote the space of all probability measures on $\mathfrak{X}$ by $\mathscr{P}(\mathfrak{X})$. Then the following equalities holds for any $k \in \mathbb{N}^*$:*

1. $\sup_{\mu \in \mathscr{P}(\mathfrak{X})} \left\| \mathbb{P}_\mu^k - \pi \right\|_{\mathrm{TV}} = d(k)$
2. $\sup_{\mu,\nu \in \mathscr{P}(\mathfrak{X})} \left\| \mathbb{P}_\mu^k - \mathbb{P}_\nu^k \right\|_{\mathrm{TV}} = \bar{d}(k)$

Note that the proof works for any distance on $\mathscr{P}$ fulfilling the triangle inequality.

Proof. We only prove *1*, since *2* can be proved analogously. Let $k \in \mathbb{N}^*$ be arbitrary. We prove the equality by proving the inequalities $\geq$ and $\leq$.

$\geq$: Due to $\mathbb{P}_x^k = \mathbb{P}_{\delta_x}^k$ with δ_x denoting the Dirac measure in x, it holds that

$$d(k) = \max_{x \in \mathfrak{X}} \left\| \mathbb{P}_x^k - \pi \right\|_{\mathrm{TV}} = \sup_{\substack{\mu \in \mathscr{P} \\ \exists x \in \mathfrak{X}\colon \mu = \delta_x}} \left\| \mathbb{P}_\mu^k - \pi \right\|_{\mathrm{TV}} \leq \sup_{\mu \in \mathscr{P}(\mathfrak{X})} \left\| \mathbb{P}_\mu^k - \pi \right\|_{\mathrm{TV}}$$

$\leq$: Let $\mu \in \mathscr{P}$ be arbitrary. We first note that $\mu = \sum_{x\in\mathfrak{X}} \mu(x)\delta_x$ and $\delta_x P^k = \mathbb{P}_x^k$ for any $x \in \mathfrak{X}$, as well as $\sum_{x\in\mathfrak{X}} \mu(x) = 1$. With this and the triangle equality of $\|\cdot\|_{\mathrm{TV}}$ we get that

$$\left\|\mathbb{P}_\mu^k - \pi\right\|_{\mathrm{TV}} = \left\|\mu P^k - \pi\right\|_{\mathrm{TV}} = \left\|\sum_{x\in\mathfrak{X}} \mu(x)\delta_x P^k - \pi\right\|_{\mathrm{TV}} = \left\|\sum_{x\in\mathfrak{X}} \mu(x)\mathbb{P}_x^k - \pi\right\|_{\mathrm{TV}}$$
$$= \left\|\sum_{x\in\mathfrak{X}} \mu(x)\mathbb{P}_x^k - \sum_{x\in\mathfrak{X}} \mu(x)\pi\right\|_{\mathrm{TV}} \leq \sum_{x\in\mathfrak{X}} \mu(x)\underbrace{\left\|\mathbb{P}_x^k - \pi\right\|_{\mathrm{TV}}}_{\leq d(k)} \leq d(k)\sum_{x\in\mathfrak{X}} \mu(x) = d(k).$$

Taking the supremum over all $\mu \in \mathscr{P}$ on both sides of preceding inequation yields the desired result.

□

Lemma 4.8. *$d(k)$ and $\bar{d}(k)$ are monotonically decreasing functions in $k \in \mathbb{N}^*$.*

Proof. We first show, that $\|\mu P - \nu P\|_{\mathrm{TV}} \leq \|\mu - \nu\|_{\mathrm{TV}}$ for fixed $\mu, \nu \in \mathscr{P}$ and any stochastic matrix P over $\mathfrak{X}$ using the triangle inequality and interchanging order of summation:

$$\|\mu P - \nu P\|_{\mathrm{TV}} = \frac{1}{2}\sum_{x\in\mathfrak{X}} |\mu P(x) - \nu P(x)| = \frac{1}{2}\sum_{x\in\mathfrak{X}} \left|\sum_{y\in\mathfrak{X}} \mu(y)p(y,x) - \nu(y)p(y,x)\right|$$
$$\leq \frac{1}{2}\sum_{x\in\mathfrak{X}}\sum_{y\in\mathfrak{X}} |\mu(y)p(y,x) - \nu(y)p(y,x)| = \frac{1}{2}\sum_{y\in\mathfrak{X}} |\mu(y) - \nu(y)| \underbrace{\sum_{x\in\mathfrak{X}} p(y,x)}_{=1} = \|\mu - \nu\|_{\mathrm{TV}}.$$

Taking the supremum over all $\mu, \nu \in \mathscr{P}$ on both sides yields the result for $\bar{d}$. Also, it immediately follows that for a stationary distribution π and for all $k \in \mathbb{N}^*$ we have

$$\left\|\mu P^{k+1} - \pi\right\|_{\mathrm{TV}} = \left\|\mu P^{k+1} - \pi P^{k+1}\right\|_{\mathrm{TV}} \leq \left\|\mu P^k - \pi P^k\right\|_{\mathrm{TV}} = \left\|\mu P^k - \pi\right\|_{\mathrm{TV}}.$$

This time, taking the supremum over $\mu \in \mathscr{P}$ gives us that $d(k+1) \leq d(k)$, from which by induction follows that d is non-increasing. □

We will show later, that a meaningful method for proving cutoff is finding upper and lower bounds for the worst-case distances. Hence, we will now explicitly state important inequalities regarding them:

Lemma 4.9. *The* worst-case separation distance, worst-case ℓ^p-distance *and the* worst-case starting total variation distance *are submultiplicative, i.e. for any $k, l \in \mathbb{N}^*$:*

1. $s(k+l) \leq s(k)s(l)$
2. ${}_p d(k+l) \leq {}_p d(k)_p d(l)$
3. $\bar{d}(k+l) \leq \bar{d}(k)\bar{d}(l)$

Proof. A sketch of the proof of *1* can be found in [10, p. 86: Ex. 6.4], *2* is given as an exercise in [10, p. 59: Lem. 4.18] and *3* is shown in [10, p. 54: Lem. 4.11]. □

Lemma 4.10. *The* separation distance $\mathrm{sep}(k)$ *is non-increasing in* $k \in \mathbb{N}^*$.

Proof. Let us first note, that for any $k \in \mathbb{N}^*$ and $\mu, \nu \in \mathscr{P}$,

$$\mathrm{sep}(\mu, \nu) = 1 - \underbrace{\min_{x \in \mathfrak{X}} \frac{\mu(x)}{\nu(x)}}_{\leq 1} \leq 1.$$

The minimum is bounded by 1 from above, because if we assume that $\min_{x\in\mathfrak{X}} \frac{\mu(x)}{\nu(x)} > 1$, it follows that $\nu < \mu$ on $\mathfrak{X}$, whence $1 = \sum_{x\in\mathfrak{X}} \nu(x) < \sum_{x\in\mathfrak{X}} \mu(x) = 1$, which is a contradiction.

Hence, also $s(k) \leq 1$ by definition for all $k \in \mathbb{N}^*$. From this and the first point from lemma 4.9, we obtain that for arbitrary $k, l \in \mathbb{N}^*$ with $k \leq l$,

$$s(l) \leq s(k)s(l-k) \leq s(k).$$

□

Lemma 4.11. *The* worst-case ℓ^p-distances *are non-decreasing in* p*, i.e. for* $1 \leq p \leq q \leq \infty$ *and* $k \in \mathbb{N}^*$,

$$_p d(k) \leq {}_q d(k).$$

Proof. The statement can be found without explicit proof in [10, p. 56: (4.37)]. □

Lemma 4.12. *The following two statements are true for any* $k \in \mathbb{N}^*$*:*

1. $\left\|\mathbb{P}_x^k - \pi\right\|_{\mathrm{TV}} \leq \mathrm{sep}(\mathbb{P}_x^k, \pi)$.
2. $d(k) \leq s(k)$.

Proof. This short proof is based on [10, p. 80: Lem. 6.16].

1. Let $x \in \mathfrak{X}, k \in \mathbb{N}^*$ be arbitrary. Then by (2.) of theorem 4.3 it holds that

$$\begin{aligned}\left\|\mathbb{P}_x^k - \pi\right\|_{\mathrm{TV}} &\overset{(4.3)}{=} \sum_{\substack{y\in\mathfrak{X} \\ \mathbb{P}_x^k(y)<\pi(y)}} (\pi(y) - \mathbb{P}_x^k(y)) = \sum_{\substack{y\in\mathfrak{X} \\ \mathbb{P}_x^k(y)<\pi(y)}} \pi(y)\left(1 - \frac{\mathbb{P}_x^k(y)}{\pi(y)}\right) \\ &\leq \max_{y\in\mathfrak{X}} \left(1 - \frac{\mathbb{P}_x^k(y)}{\pi(y)}\right) = \mathrm{sep}(\mathbb{P}_x^k, \pi),\end{aligned}$$

   which concludes the proof of *1.*

2. By taking the maximum over all $x \in \mathfrak{X}$ on both sides of 1., we obtain the desired inequation.

□

Lemma 4.13. *The following inequalities between the* worst-case total variation distance *and the* worst-case starting total variation distance *hold for any* $k \in \mathbb{N}^*$*:*

$$d(k) \leq \bar{d}(k) \leq 2d(t)$$

Proof. This proof is based on the proof given in [10, p. 53: Lem. 4.10]. We subsequently show both inequalities:

1. First, fix $k \in \mathbb{N}^*$ and $x \in \mathfrak{X}$ and note that due to $\pi = \pi P$, we have $\pi(A) = \sum_{y\in\mathfrak{X}} \pi(y)p^k(y, A)$ for any $A \subseteq \mathfrak{X}$. Since π is a distribution, $\sum_{y\in\mathfrak{X}} \pi(y) = 1$ holds. With this and the triangle inequality we get

$$\left|\mathbb{P}_x^k(A) - \pi(A)\right| = \left|\sum_{y\in\mathfrak{X}} \left(\mathbb{P}_x^k(A) - \mathbb{P}_y^k(A)\right)\right| \leq \sum_{y\in\mathfrak{X}} \pi(y)\left\|\mathbb{P}_x^k - \mathbb{P}_y^k\right\|_{\mathrm{TV}} \leq \bar{d}(t) \sum_{y\in\mathfrak{X}} \pi(y) = \bar{d}(t).$$

   By taking the supremum over all $x \in \mathfrak{X}, A \subseteq \mathfrak{X}$ on both sides of the inequation we obtain the desired inequality.

2. Using the triangle inequality, we immediately obtain

$$\bar{d}(k) = \max_{x,y\in\mathfrak{X}} \left\|\mathbb{P}_x^k - \mathbb{P}_y^k\right\|_{\mathrm{TV}} \leq \max_{x,y\in\mathfrak{X}} \left(\left\|\mathbb{P}_x^k - \pi\right\|_{\mathrm{TV}} + \left\|\pi - \mathbb{P}_y^k\right\|_{\mathrm{TV}}\right) = 2d(t).$$

□

We also get a similar estimate for the inverse direction if the Markov chain is reversible:

Lemma 4.14. *In case that the underlying Markov chain is reversible, the following inequalities hold:*

$$s(2t) \leq 1 - (1 - \bar{d}(t))^2 \leq 2\bar{d}(t) \leq 4d(t)$$

Proof. A complete proof of the first inequality can be found in [10, p. 80: Lem. 6.17]. The latter two are clear by expanding the square and applying lemma 4.13. □

4.2.2 Mixing Times

In order to quantify the time after which the chain gets arbitrarily close to equilibrium, we introduce the so-called *mixing times*.

Definition 4.15. *Let* $\varepsilon > 0$ *be arbitrary. We define the* ε-mixing time *with respect to the* total variation distance *to be*

$$t_{\mathrm{mix}}(\varepsilon) := \min\left\{k \in \mathbb{N}^* : d(k) \leq \varepsilon\right\}.$$

Analogously, we define ${}_pt_{\mathrm{mix}}(\varepsilon)$ *to be the* ε*-mixing time with respect to the* ℓ^p*-norm, and* ${}_{\mathrm{sep}}t_{\mathrm{mix}}(\varepsilon)$ *to be the separation distance* ε*-mixing time. In order to shorten the notation, we set* $t_{\mathrm{mix}} := t_{\mathrm{mix}}(\frac{1}{4})$ *and* ${}_pt_{\mathrm{mix}} := {}_pt_{\mathrm{mix}}(\frac{1}{2})$*. Since* $t_{\mathrm{mix}} = {}_1t_{\mathrm{mix}}$ *and the* ℓ^p*-distances are submultiplicative by lemma 4.9, these definitions are indeed consistent.*

$t_{\text{mix}}(\varepsilon)$ can be interpreted as the smallest time, such that the total variation distance between $\mathbb{P}_{\mu}^{t_{\text{mix}}(\varepsilon)}$ and the stationary distribution π is at most ε for any starting distribution μ.

Lemma 4.16. *For any mixing time t_{mix} with respect to a non-increasing worst-case distance $w : \mathbb{N}^* \to \mathbb{R}_0^+$, it holds for any $k \in \mathbb{N}^*$ and $\varepsilon > 0$ that*

$$w(k) \leq \varepsilon \Leftrightarrow t_{\text{mix}}(\varepsilon) \leq k.$$

Proof. We show each implication:

$\Rightarrow$: Let $w(k) \leq \varepsilon$. From the definition of the mixing time, $t_{mix}(\varepsilon) = \min\{l \in \mathbb{N}^* : w(l) \leq \varepsilon\}$, it immediately follows that $t_{mix}(\varepsilon) \leq k$.

$\Leftarrow$: Now, let $t_{\text{mix}}(\epsilon) \leq k$ and we assume that $w(k) > \varepsilon$. Since w is non-increasing, we obtain that $w(t_{\text{mix}}(\varepsilon)) \geq w(k) > \varepsilon$, which contradicts $w(t_{\text{mix}}(\varepsilon)) \leq \varepsilon$.

□

Lemma 4.17. *The ε-mixing time is non-increasing in ε. This means that for any $0 < \delta \leq \varepsilon$ it holds that*

$$t_{\text{mix}}(\delta) \geq t_{\text{mix}}(\varepsilon).$$

Note that this result is independent of the used worst-case distance.

Proof. Let $0 < \delta \leq \varepsilon$. For arbitrary $c > 0$ we define $T_c := \{k \in \mathbb{N}^* : d(k) \leq c\}$ and observe, that $T_\delta \subseteq T_\varepsilon$. By definition of the mixing times,

$$t_{\text{mix}}(\delta) = \min T_\delta \geq \min T_\varepsilon = t_{\text{mix}}(\epsilon).$$

□

Lemma 4.18. *For any $\varepsilon > 0$ we have $t_{\text{mix}}(\varepsilon) \leq \left\lceil \log_2 \varepsilon^{-1} \right\rceil t_{\text{mix}}$.*

Proof. The proof is taken from [10, p. 54: (4.34)]. First, note that due to lem. 4.13 and the submultiplicativity of $\bar{d}$ (lem. 4.9) for any $k \in \mathbb{N}^*$ we have that

$$d(kt_{\text{mix}}(\delta)) \overset{4.13}{\leq} \bar{d}(kt_{\text{mix}}(\delta)) \overset{4.9}{\leq} \bar{d}(t_{\text{mix}}(\delta))^k \overset{4.13}{\leq} (2d(kt_{\text{mix}}(\delta)))^k = (2\delta)^k.$$

If we set $\delta = \frac{1}{4}$, we obtain $d(kt_{\text{mix}}) \leq 2^{-k}$. Now let $\varepsilon > 0$ be arbitrary and choose the smallest $k \in \mathbb{N}^*$ such that $2^{-k} \leq \varepsilon$, i.e. $k = \left\lceil \log_2 \varepsilon^{-1} \right\rceil$. Therewith, we have $d\left(\left\lceil \log_2 \varepsilon^{-1} \right\rceil t_{\text{mix}}\right) \leq 2^{-k} \leq \epsilon$, which implies $t_{\text{mix}}(\varepsilon) \leq \left\lceil \log_2 \varepsilon^{-1} \right\rceil t_{\text{mix}}$ by definition of the mixing time.

□

4.2.3 Cutoff and Pre-Cutoff

We have finally acquired the knowledge necessary to define cutoff. From now on, we will always consider $(\mathfrak{X}_n, P_n)_{n\in\mathbb{N}^*}$ to be a sequence of ergodic Markov chains with stationary distributions π_k and ε-mixing times $t_{\text{mix}}^{(n)}(\varepsilon)$, $t_{\text{mix}}^{(n)}$ the $\frac{1}{4}$-mixing time and $d^{(n)}$ the worst case total variation norm of the n-th chain for each $n \in \mathbb{N}^*$.

Definition 4.19. *We say that this sequence of chains exhibits cutoff, if for any $\varepsilon \in (0,1)$*

$$\lim_{n\to\infty} \frac{t_{\text{mix}}^{(n)}(\varepsilon)}{t_{mix}^{(n)}(1-\varepsilon)} = 1.$$

Cutoff has the following equivalent characterisation:

Theorem 4.20. *The sequence $(\mathfrak{X}_n, P_n)_{n\in\mathbb{N}^*}$ has cutoff iff*

$$\lim_{n\to\infty} d^{(n)}(ct_{\text{mix}}^{(n)}) = \begin{cases} 1, & \text{if } c \in (0,1) \\ 0, & \text{if } c > 1 \end{cases}$$

Of course, $d^{(n)}(ct_{\text{mix}}^{(n)}))$ is only defined for $ct_{\text{mix}}^{(n)} \in \mathbb{N}^*$. Also note that $c \neq 1$, because by definition $d^{(n)}(t_{\text{mix}}^{(n)}) \leq \frac{1}{4}$.

Note that this equivalence does hold in any setting, where non-increasing worst-case distances which are bounded from above by 1 are used, for example, the separation distance and, of course, the total variation distance, for which the statement is proved explicitly. For the ℓ^p-distances, this theorem is not true without further modifications.

Proof. We subsequently show both implications:

$\Rightarrow$: Assume that the sequence of chains exhibits cutoff, this means that for any $\varepsilon \in (0,1)$,

$$\lim_{n\to\infty} \frac{t_{\text{mix}}^{(n)}(\varepsilon)}{t_{\text{mix}}^{(n)}(1-\varepsilon)} = 1.$$

(a) First, let $c > 1$ and $\varepsilon \in (0, \frac{1}{4})$ fixed. Then, there exists some $N = N(\varepsilon, c) \in \mathbb{N}^*$ such that for all $n \geq N$ we have

$$\frac{t_{\text{mix}}^{(n)}(\varepsilon)}{t_{\text{mix}}^{(n)}(1-\varepsilon)} < c. \tag{7}$$

Since $1 - \varepsilon \geq \frac{3}{4} > \frac{1}{4}$, we obtain with the preceding inequality and lemma 4.17, which states that the ε-mixing time is non-increasing in ε, that

$$\frac{1}{c} t_{\text{mix}}^{(n)}(\varepsilon) < t_{\text{mix}}^{(n)}(1-\varepsilon) \leq t_{\text{mix}}^{(n)}\left(\frac{3}{4}\right) \leq t_{\text{mix}}^{(n)}(\frac{1}{4}) = t_{\text{mix}}^{(n)}.$$

This, in particular, implies that

$$t_{\text{mix}}^{(n)}(\varepsilon) \leq ct_{\text{mix}}^{(n)}$$

from which, in turn, follows from the definition of the mixing time and the non-decreasingness of the worst-case total variation distance shown in lemma 4.8, that for any $n \geq N$, $0 < \varepsilon < \frac{1}{4}$,

$$0 \leq d^{(n)}\left(ct_{\text{mix}}^{(n)}\right) \leq d^{(n)}\left(t_{\text{mix}}^{(n)}(\varepsilon)\right) \leq \varepsilon,$$

from which, by taking the limit with respect to n, we obtain for arbitrary $c > 1$,

$$\lim_{n\to\infty} d^{(n)}\left(ct_{\text{mix}}^{(n)}\right) = 0.$$

(b) Now let $c \in (0,1)$ and $\varepsilon \in (0, \frac{1}{4})$. Set $\tilde{c} := \frac{1}{c} > 1$. Then by statement (7) from point (a), there is some $N = N(\varepsilon, \tilde{c}) \in \mathbb{N}^*$ such that

$$\frac{t_{\text{mix}}^{(n)}(\varepsilon)}{t_{\text{mix}}^{(n)}(1-\varepsilon)} < \tilde{c} = \frac{1}{c} \Leftrightarrow ct_{\text{mix}}^{(n)}(\varepsilon) < t_{\text{mix}}^{(n)}(1-\varepsilon).$$

Due to $\varepsilon < \frac{1}{4}$ and the monotonic decrease of $t_{\text{mix}}^{(n)}$, we get that $t_{\text{mix}}^{(n)}(\varepsilon) \geq t_{\text{mix}}^{(n)}$, and consequently

$$t_{\text{mix}}^{(n)}(1-\varepsilon) > ct_{\text{mix}}^{(n)}(\varepsilon) \geq ct_{\text{mix}}^{(n)}.$$

Lemma 4.16 implies

$$d^{(n)}(ct_{\text{mix}}^{(n)}) > 1 - \varepsilon,$$

for any $\varepsilon \in (0, \frac{1}{4}), c \in (0,1)$. Therefore, by taking the limit over n and using that our distance is bounded from above by 1, it follows that

$$\lim_{n\to\infty} d^{(n)}(ct_{\text{mix}}^{(n)}) = 1.$$

This concludes the first implication.

$\Leftarrow$: Let $\varepsilon \in (0,1)$ and $c \in (0,1)$, i.e. $\frac{1}{c} > 1$. By precondition, $\lim_{n\to\infty} d^{(n)}(ct_{\text{mix}}^{(n)}) = 1$ as well as $\lim_{n\to\infty} d^{(n)}(\frac{1}{c}t_{\text{mix}}^{(n)}) = 0$. Hence, we can find some $N, N', N'', N''' \in \mathbb{N}^*$ dependent on c, ε such that

(a) For all $n \geq N$:

$$d^{(n)}(ct_{\text{mix}}^{(n)}) > 1 - \varepsilon \overset{4.16}{\Rightarrow} t_{\text{mix}}^{(n)}(1-\varepsilon) > ct_{\text{mix}}^{(n)}$$

(b) For all $n \geq N'$:

$$d^{(n)}\left(\frac{1}{c}t_{\text{mix}}^{(n)}\right) \leq \varepsilon \overset{4.16}{\Rightarrow} t_{\text{mix}}^{(n)}(\varepsilon) \leq \frac{1}{c}t_{\text{mix}}^{(n)}$$

(c) Using (a) with $\tilde{\varepsilon} := 1 - \varepsilon \in (0,1)$ ensures the existence of some $N'' \in \mathbb{N}^*$ such that for any $n \geq N''$:

$$t_{\text{mix}}^{(n)}(\varepsilon) = t_{\text{mix}}^{(n)}(1-\tilde{\varepsilon}) > ct_{\text{mix}}^{(n)}$$

(d) Using (b) with the same trick used in (c) again guarantees the existence of some $N''' \in \mathbb{N}^*$ such that for any $n \geq N'''$:

$$t_{\text{mix}}^{(n)}(1-\varepsilon) = t_{\text{mix}}^{(n)}(\tilde{\varepsilon}) \leq \frac{1}{c} t_{\text{mix}}^{(n)}$$

From (a) to (d) we can conclude that for all $c \in (0,1)$ and $n \geq \max\{N, N', N'', N'''\}$,

$$c^2 \overset{\text{(c),(d)}}{\leq} \frac{t_{\text{mix}}^{(n)}(\varepsilon)}{t_{\text{mix}}^{(n)}(1-\varepsilon)} \overset{\text{(a),(b)}}{\leq} \frac{1}{c^2}.$$

Therefore, by taking the limits $c \to 1$ and $n \to \infty$, we get

$$\lim_{n\to\infty} \frac{t_{\text{mix}}^{(n)}(\varepsilon)}{t_{\text{mix}}^{(n)}(1-\varepsilon)} = 1.$$

□

The preceding lemma provides a very demonstrative interpretation of cutoff, see fig. 1.

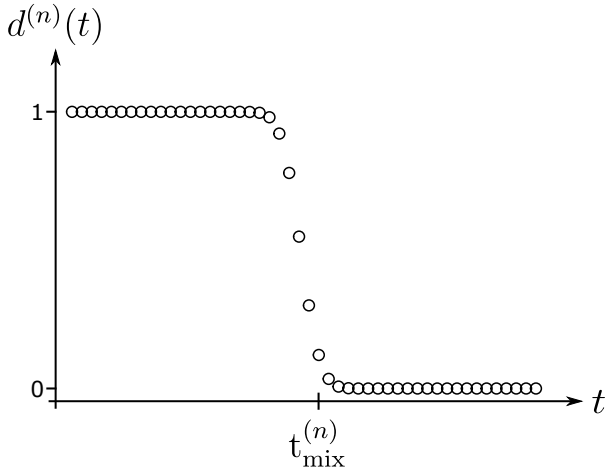


Figure 1: Graphical interpretation of the cutoff phenomenon. For $n \to \infty$ the graph approximates a step function. This figure is based on [10, p. 262: Fig. 18.1] .

In order to better understand the definitions, we will now see a very simple sequence of Markov chains showing cutoff:

Example 4.21. *Consider the sequence of Markov chains with state spaces* $\mathfrak{X}_n := [\![1, n]\!]$ *and transition matrices* $P_n := \frac{1}{n}\mathbb{1}_{n\times n}$ *for any* $n \in \mathbb{N}^*$*, where* $\mathbb{1}_{n\times n}$ *denotes the* $n \times n$*-matrix with every entry being* 1*. This can be interpreted as following: In every step of the chain, some element from the state space is chosen using uniform distribution on* $\mathfrak{X}_n$*, completely independent of the state's current state. It is easy to check that the chain is both irreducible and aperiodic.*

First, fix $n \in \mathbb{N}^*$ *and* $k \in \mathbb{N}^*$*. Since* $P_n^k = P_n$*, also* $\mathbb{P}_x^k = p^k(x, \cdot) = \frac{1}{n}\mathbb{1}_n^t$ *for any* $x \in \mathfrak{X}$*. Because of that, the stationary distribution* $\pi = \frac{1}{n}\mathbb{1}_n^t$ *is uniform.*

This means that for any starting distribution μ*, we attain* $\mu P = \pi$ *after one single step, hence* $t_{\text{mix}}^{(n)}(\varepsilon) = 1$ *and also* $\frac{t_{\text{mix}}^{(n)}(\varepsilon)}{t_{\text{mix}}^{(n)}(1-\varepsilon)} = 1$ *for arbitrary* $\varepsilon > 0$*. By definition 4.19, one can easily see that this rather trivial sequence of chains exhibits cutoff.*

Definition 4.22. *We say that the sequence of Markov chains has a* pre-cutoff *if*

$$\sup_{0<\varepsilon<\frac{1}{2}} \limsup_{n\to\infty} \frac{t_{\text{mix}}^{(n)}(\varepsilon)}{t_{\text{mix}}^{(n)}(1-\varepsilon)} < \infty$$

We can directly see that the presence of cutoff implies pre-cutoff. Nevertheless, the converse does not hold: Hubert Lacoin constructed in [8] a class of sequences of Markov chains, which always exhibits pre-cutoff, but not necessarily cutoff. He is using sequences of product chains, which were briefly mentioned in the introduction to subsection 4.2.

Definition 4.23. *The sequence of Markov chains is said to exhibit cutoff with a* (cutoff)-window *of size* $O(w_n)$ *for positive* $w_n = o(t_{mix}^{(n)})$*, i.e.* $\lim_{n\to\infty} \frac{w_n}{t_{mix}^{(n)}} = 0$*, if*

$$\lim_{\alpha\to-\infty} \liminf_{n\to\infty} d^{(n)}(t_{mix}^{(n)} + \alpha w_n) = 1 \tag{8}$$

$$\lim_{\alpha\to\infty} \limsup_{n\to\infty} d^{(n)}(t_{mix}^{(n)} + \alpha w_n) = 0. \tag{9}$$

There are used many different definitions of cutoff with window, which differ from above definition of cutoff with window in small details only. Notable examples include Diaconis' definition of cutoff, which is given in [3], and the definition used by D'Angeli and Donno in [5].

Lemma 4.24. *Cutoff with a window of size* $O(w_n)$ *implies cutoff.*

Proof. We show cutoff by using the equivalent characterisation 4.20:

1. First, let $c < 1$ and $\varepsilon > 0$ be arbitrary. Because of 8, there exists some $A > 0$ such that for all $\alpha < -A$,
$$\liminf_{n\to\infty} d^{(n)}(t_{\text{mix}}^{(n)} + \alpha w_n) > 1 - \varepsilon.$$
Because of this, we can find some $N \in \mathbb{N}^*$ such that $n \geq N$ implies
$$d^{(n)}(t_{\text{mix}}^{(n)} + \alpha w_n) > 1 - \varepsilon.$$
Since
$$\frac{t_{\text{mix}}^{(n)} + \alpha w_n}{t_{\text{mix}}^{(n)}} = 1 + \alpha\frac{w_n}{t_{\text{mix}}^{(n)}}$$

and $\lim_{n\to\infty} \frac{w_n}{t_{\text{mix}}^{(n)}} = 0$ (by $w_n = o(t_{mix}^{(n)})$) and $c < 1$, we can find $N' \geq N$ such that for $n \geq N'$,

$$\frac{t_{\text{mix}}^{(n)} + \alpha w_n}{t_{\text{mix}}^{(n)}} > c \Leftrightarrow ct_{\text{mix}}^{(n)} < t_{\text{mix}}^{(n)} + \alpha w_n.$$

We can now apply 4.16, which yields

$$1 \geq d^{(n)}(ct_{\text{mix}}^{(n)}) \geq d^{(n)}(t_{\text{mix}}^{(n)} + \alpha w_n) > 1 - \varepsilon.$$

Since $\varepsilon > 0$ was chosen arbitrarily, we obtain

$$\lim_{n\to\infty} d^{(n)}(ct_{\text{mix}}^{(n)}) = 1$$

2. The case $c > 1$ can be shown analogously.

□

Ultimately, we want to mention, that a very general notion of cutoff for a family of non-decreasing non-negative functions can be defined, which is used by Guan-Yu Chen and Laurent Saloff-Coste in [4, p. 5: Def. 2.1]. This type of cutoff is completely independent of probabilistic settings.

5 Conclusion

In the course of this Bachelor's thesis, we have motivated and formulated the ideas behind the concept of *cutoff* in Markov chains in detail. To conclude this paper, we want to mention that notable examples of Markov chains show cutoff, for example the *random walk on the hypercube*, which is presented in [10, p. 266].

In addition, random walks on the symmetric group $\mathfrak{S}_n$ of $n \in \mathbb{N}^*$ elements, which are used in modelling shuffling of cards, frequently show cutoff. A very basic, yet impractical, example is the so-called *top-to-random shuffle*. A more practical shuffle is the riffle shuffle, described by Gilbert, Shannon and Reeds, which is hence also called *GSR-shuffle*. These two shuffles show cutoff, and detailed proofs of their cutoffs can be found in [6, p. 10-15]. This paper, for example, also uses the technique of *strong stationary times*.

References

[1] Diaconis, P., Shahshahani, M.: *Generating a Random Permutation with Random Transpositions.* Z. Wahrscheinlichkeitstheorie verw. Gebiete. **57**: 159-179 (1981)

[2] Diaconis, P.: *Group Representations in Probability and Statistics.* Institute of Mathematical Statistics. Hayward, California. (1988)

[3] Diaconis, P.: *The cutoff phenomenon in finite Markov chains.* Proc. Natl. Acad. Sci. USA. **93**: 1659-1664 (1996)

[4] Chen, G., Saloff-Coste, L.: *The Cutoff Phenomenon for Ergodic Markov Processes.* Electron. J. Probab. **13.3**: 26-78 (2008)

[5] D'Angeli, D., Donno, A.: *No Cut-Off Phenomenon for the "Insect Markov Chain".* Monatsh. Math. **156**: 201-210 (2009)

[6] Nooitgedagt, H.: *Two Convergence Limits of Markov Chains: Cutoff and Metastability.* Mathematisch Instituut, Universiteit Leiden (2010)

[7] Mürmann, M.: *Wahrscheinlichkeitstheorie und Stochastische Prozesse.* Springer-Verlag Berlin Heidelberg (2014)

[8] Lacoin, H.: *A Product Chain without Cutoff.* Electron. Commun. Probab. **20.19**: 1-9 (2015)

[9] Hermon, J., Lacoin, H., Peres, Y.: *Total Variation and Separation Cutoffs are not Equivalent and Neither One Implies the Other.* Electron. J. Probab. **21.44**: 1-36 (2016)

[10] Levin, D. A., Peres, Y., Wilmer, E. L.: *Markov Chains and Mixing Times.* American Mathematical Society (2017)