Senior Thesis

Markov Chains and Mixing Times on Colorings

Written by
Camille Taltas

Academic Advisor
David Galvin

Contents

Abstract

In this thesis, we address the question of under what circumstances the Glauber dynamics for sampling uniformly from the set of proper q-colorings of a graph mixes rapidly. We start by exploring the conditions under which a chain has a stationary distribution. We then evaluate ways to measure the speed at which a chain converges to its stationary distribution. Lastly, we use those concepts and theorems to apply them to three different proofs of rapidly mixing of uniform samplings on the set of q-colorings of a graph, with a gradual decrease on the lower bound for q in terms of the maximum degree of the graphs inputted.

1 Introduction

A q-coloring of a graph, $G = (V, E)$ is an assignment of colors, which serves as labels, to the set of vertices, V, of the graph. We say that a q-coloring is proper if no two colors on neighboring vertices are the same.

For example, if G is an Amazon network in which V represents the set of products on sale and an edge, $\{x, y\} \in E$, between two vertices represents the fact that two products are frequently bought together, then a color assignment to a vertex would be a set of products recommended to customers. This color assignment partitions the graph in order to avoid redundancy when marketing products.

In this thesis, we attempt to show what a coloring typically looks like. By knowing this information and running a process many times on these typical examples, one can observe the results and understand how efficient the process is.

For example, our process could be a cost function associated with every coloring on our Amazon network. Knowing a typical coloring would allow us to evaluate the cost of marketing.

In order to find a typical example, we want to generate a coloring uniformly at random.

One way of doing that would be to list all the possible q-colorings on G and going to a random point. However, this is not effective as the number of colorings is exponential in the number of vertices.

Thus, we use Markov chain algorithms that move around the space of colorings in a random enough way that after a while, the chain is equally likely to be in any one place as any other.

For example, in this paper, we will study Glauber dynamics. A chain which, at each step, picks a vertex uniformly at random, erases its current color, picks a new color uniformly at random from among all the allowable colors at that vertex, where a color is allowable if it does not violate the proper coloring rule.

We will see, throughout this body of work, that such a process leads to a uniform coloring and does so in polynomial time in the number of vertices.

We will do so by introducing the notions of stationary distribution of Markov chains, ergodicity, total variation distance between probability distributions, mixing time, and coupling of random variables.

Note that the Glauber dynamics chain is particularly useful as it only requires to look at a small portion of the graph at every step. Thus, it can easily run on large and dense data sets. This is also why we would like to evaluate how small we can make q, the number of colors used, for this chain to mix rapidly.

Now, suppose we are working with the class of graphs with maximum degree Δ. For Glauber dynamics, it is easy to see that if q is large compared to Δ then at each step, there is room to move since there are many choices for a new color at randomly chosen vertex. Thus, we expect that this chain will settle rapidly or in polynomial time in the number of vertices of the graph to a uniform distribution. However, if the number of colors is small, there may be places where Glauber dynamics gets stuck, and the mixing time might be slow or exponential in the number of vertices.

It is well known that $\Delta + 2$ is just enough to allow the Glauber chain to be defined in a way that it does indeed always converge rapidly to a uniform distribution. Moreover, there is a long-standing conjecture which states that if $q > \Delta + 1$ then the mixing is rapid.

In this paper we will first prove fairly easily that $q > 3\Delta$ suffices for rapid mixing, we will then, with a harder argument, show the bound of $q > 2\Delta$, and, finally, we will outline the ideas used to show that $q > \frac{11}{6}\Delta$ mixes rapidly.

Note that the bound of $\frac{11}{6}\Delta$ has recently been improved slightly, but that, to this day, the $\Delta + 1$ conjecture remains open.

In order to provide the three proofs mentioned above, we will start by defining Markov chains and some of their basic properties. We will then introduce a proof of the existence and uniqueness of a stationary distribution. This gives us enough knowledge to introduce our problem and Glauber dynamics.

In the second section, we introduce a way to measure the distance from the stationary distribution through total variation distance, mixing time, and coupling. This leads us to proving the convergence theorem, which will then be used to prove our first bound on q of 3Δ.

For our second bound of 2Δ, we introduce the transportation metric and the concept of path coupling. The path coupling theorem will allow us to simplify the analysis of the chain and provide a simple proof to this new and improved bound on q. This proof was first introduced by Russ Bubley and Martin Dyer in 1997.

Lastly, for our lowest bound of $\frac{11}{6}\Delta$ we continue to use the principle of path coupling and apply it to a new chain called Flip Dynamics, introduced by Eric Vigoda in 1999 [2]. This analysis, requiring more algebra and case by case analysis, is summarized in this thesis in order to convey the mechanics and a clearer understanding of how the coupling runs. Also, this proof calls for some use of linear programming as the choice of probabilities is optimized to attain this bound on q.

This paper is written under the assumption that the reader has some fundamental understanding of real analysis, probability, and linear algebra. No prior knowledge of Markov chains or graph theory is needed to read through these proofs.

1.1 Markov Chains

We start this paper by defining what a Markov chain is. An example of a random walk on a graph is provided in order to introduce the intuition behind the Markov chains we will be using for our proofs.

Definition 1.1. *A **Markov Chain** with state space $\mathcal{X}$ and transition matrix P is a sequence of random variables $(X_0, X_1, ...)$ where*

$$\boldsymbol{P}\{X_{t+1} = y \mid X_t = x\} = P(x, y)$$

or the probability of $X_{t+1} = y$ given $X_t = x$ is $P(x, y)$.

Remark 1.2. *The transition matrix P is stochastic; that is, all entries are non-negative and*

$$\sum_{y \in \mathcal{X}} P(x, y) = 1.$$

Remark 1.3. *If μ, a vector whose i^{th} coordinate is the probability that X_0 takes value x_i, is the mass function of X_0, then the mass function μ' of X_1 can be obtained by matrix multiplication: $\mu' = \mu P$, where P is the transition matrix. This is proved by Law of Total Probability.*

Example 1.4 (Random Walk on a Graph)**.** *A graph $G = (V, E)$ consists of a vertex set V and an edge set E, where the elements of E are unordered pairs of vertices. We say that x and y are neighbors and denote it $x \backsim y$ if $\{x, y\} \in E$. We use $deg(x)$ for the degree of a vertex x, the number of neighbos of x.*

Given a graph $G = (V, E)$, we define a simple random walk on G to be the Markov chain with state space V and transition matrix:

$$P(x, y) = \begin{cases} \dfrac{1}{deg(x)} & \textit{if } y \backsim x \\ 0 & \textit{otherwise} \end{cases}$$

In short, when the chain is at vertex x, it examines all neighbors of x, picks one uniformly at random and moves to the chosen vertex.

Remark 1.5. *$P^r(x, \cdot)$ is the distribution of the chain after r steps, given that it started in state x.*

The principles or irreducibility and aperiodicity are introduced as they are integral to proving the existence of, uniqueness of, and convergence to the stationary distribution.

Definition 1.6 (Irreducibility)**.** *A chain P is called irreducible if for any two states $x, y \in \mathcal{X}$, there exists an integer t such that $P^t(x, y) > 0$. In short, this means that it is possible to get to any state from any other states using only transitions of positive probability.*

Definition 1.7 (Aperiodicity)**.** *A chain is aperiodic if all states have period 1, where the period is the greateast common divisor of $\mathcal{T}(x) := \{t \geq 1 : P^t(x,x) > 0\}$.*

Example 1.8. *A chain which has two states where $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, goes back and forth between the two states with certainty and, hence, the greatest common divisor of $\mathcal{T}(x) = \{2, 4, \dots\}$ is 2, which makes it periodic.*

1.2 Stationary Distributions

Here, we introduce the concept of stationary distribution and prove our first two major theorems which guarantee the existence and uniqueness of a stationary distribution.

Definition 1.9 (Stationary Distribution)**.** *Given a distribution π on $\mathcal{X}$ and a transition matrix P, if $\pi = \pi P$ then π is a stationary distribution.*

Remark 1.10. *The importance of being a stationary distribution lies in the fact that the mass function of X_0 is π and, at every time t, the mass function of X_t obtained by matrix multiplication is π.*

Theorem 1.11. *Every finite-state Markov chain has a stationary distribution.*

Proof. Let P be an $n \times n$ transition matrix for a chain with state space $\mathcal{X} = \{x_1, \dots, x_n\}$. We will show that P has a stationary distribution ν. The proof will be carried out in a sequence of claims: Claim 1.12, Claim 1.13, Claim 1.14, and Claim 1.15.

For an arbitrary distribution μ, set $\nu_n = (\mu + \mu P + \cdots + \mu P^{n-1})/n$.

Claim 1.12. *For every state x, $|\nu_n P(x) - \nu_n(x)| \leq 2/n$.*

Proof. (Claim 1.12) By the triangle inequality and since μ is a probability distribution, $|\nu_n P(x) - \nu_n(x)|$

$$\begin{aligned}
&= \frac{|(\mu P(x) + \mu P^2(x) + \cdots + \mu P^n(x) - \mu(x) - \mu P(x) - \cdots - \mu P^{n-1}(x)|}{n} \\
&= \frac{|\mu P^n(x) - \mu(x)|}{n} \\
&\leq \frac{|\mu P^n(x)|}{n} + \frac{|\mu(x)|}{n} \\
&\leq \frac{1}{n} + \frac{1}{n} \\
&= \frac{2}{n}.
\end{aligned}$$

□

Claim 1.13. *There exists a subsequence* $(\nu_{n_1^{(n)}}, \nu_{n_2^{(n)}}, \dots, \nu_{n_m^{(n)}}, \dots)$ *of* $(\nu_1, \nu_2, \dots)$ *for which* $(\nu_{n_m^{(n)}}(x_k))_{m=1}^{\infty}$ *converges to a limit* $\nu(x_k)$ *for each* $k = 1, \dots, n$.

Proof. (Claim 1.13) For state x_1, $(\nu_n(x_1))_{n \geq 1}$ is a sequence of reals bounded between 0 and 1. Hence, by the Bolzano-Weierstrass theorem, we know that there exists a convergent subsequence $(\nu_{n_1^{(1)}}, \nu_{n_2^{(1)}}, \dots, \nu_{n_m^{(1)}}, \dots)$ such that $(\nu_{n_m^{(1)}}(x_1))_{m=1}^{\infty}$ converges to a limit $\nu(x_1)$.

Similarly, for state x_2, $(\nu_{n_1^{(1)}}(x_2), \nu_{n_2^{(1)}}(x_2), \dots, \nu_{n_m^{(1)}}(x_2), \dots)$ is a sequence of reals bounded between 0 and 1, so $(v_{n_1^{(1)}}, \dots)$ has a subsequence $(\nu_{n_1^{(2)}}, \nu_{n_2^{(2)}}, \dots, \nu_{n_m^{(2)}}, \dots)$ such that $(\nu_{n_m^{(2)}}(x_2))_{m=1}^{\infty}$ converges to a limit $\nu(x_2)$ and that along this subsequence, $(\nu_{n_m^{(2)}}(x_1))_{m=1}^{\infty}$ still converges to limit $\nu(x_1)$.

We can continue this process iteratively for every state of the chain in order to get a final subsequence $(\nu_{n_1^{(n)}}, \nu_{n_2^{(n)}}, \dots, \nu_{n_m^{(n)}}, \dots)$ for which $(\nu_{n_m^{(n)}}(x_k))_{m=1}^{\infty}$ converges to $\nu(x_k)$ for each $k = 1, \dots, n$. □

By throwing out some term of the sequence $(\nu_n)_{n \geq 1}$, and by relabelling the rest of the terms, we can assume from here on that in fact $(\nu_m(x_i))_{m=1}^{\infty}$ converges to $\nu(x_i)$ for each i.

Claim 1.14. *The distribution* ν *is a probability distribution.*

Proof. (Claim 1.14) Clearly $\nu(x_n) \geq 0$ for all $n \geq 1$ since it is the limit of a subsequence in which all elements are greater than 0.

For each n_k, v_{n_k} is a probability distribution, so $\sum_{i=1}^{n} \nu_{n_k}(x_i) = 1$. It follows that

$$\sum_{i=1}^{n} \nu(x_i) = \sum_{i=1}^{n} \lim_{k \to \infty} \nu_{n_k}(x_i) = \lim_{k \to \infty} \sum_{i=1}^{n} \nu_{n_k}(x_i) = \lim_{k \to \infty} 1 = 1.$$

□

Claim 1.15. *The distribution* ν *is a stationary distribution.*

Proof. (Claim 1.15) Since,

$$|\nu_n P(x_i) - \nu_n(x_i)| \leq 2/n$$

and

$$\lim_{n \to \infty} \nu_n(x_i) = \nu(x_i)$$

Then:

$$\lim_{n \to \infty} \nu_n P(x_i) = \lim_{n \to \infty} \nu_n(x_i) = \nu(x_i).$$

To complete the proof,we need to show that $\lim_{n \to \infty} \nu_n P(x_i) = \nu P(x_i)$. By the definition of Markov chains we know that,

$$\nu_n P(x_i) = \sum_{j=1}^{n} \nu_n(x_j) P(x_j, x_i)$$

then taking limits we get

$$\lim_{n\to\infty} \nu_n P(x_i) = \sum_{j=1}^{n} \lim_{n\to\infty} \nu_n(x_j)P(x_j, x_i) = \sum_{j=1}^{n} \nu(x_j)P(x_j, x_i) = \nu P(x_i).$$

□

□

Now that we have proved existence of a stationary distribution we will prove uniqueness under additional hypotheses.

Theorem 1.16. *If a finite-state Markov chain is irreducible, then there exists a unique stationary distribution.*

Proof. Since we already proved existence in Therorem 1.11, it suffices to show uniqueness of the stationary distribution. For this, we will first start by introducing harmonic functions:

Definition 1.17. *A function $h : \mathcal{X} \to \mathbb{R}$ is harmonic at x if*

$$h(x) = \sum_{y\in\mathcal{X}} P(x, y)h(y).$$

Remark 1.18. *Note that the definition of h being a harmonic function, when translated into matrix multiplication, says that $Ph = h$, so $Ph - h = 0$, so $(P - I)h = 0$.*

Lemma 1.19. *Suppose P is irreducible, then a function h which is harmonic for all $x \in \mathcal{X}$ is constant.*

Proof. (Lemma 1.19) Since $\mathcal{X}$ is finite, there must be a state x_0 at which the function h reaches its maximum, call it M. Then if for some state z, $P(x_0, z) > 0$ and $h(z) < M$, then

$$h(x_0) = \sum_{y\in\mathcal{X}} P(x_0, y)h(y) < \sum_{y\in\mathcal{X}} P(x_0, y)M = M \sum_{y\in\mathcal{X}} P(x_0, y) = M$$

which contradicts our initial assumption that $h(x_0) = M$.

Since P is irreducible, for all $y \in \mathcal{X}$, there exists a $t \geq 1$ such that $P^t(x_0, y) > 0$. Hence, there exists a sequence $x_0, \ldots, x_n = y$, such that $P(x_i, x_{i+1}) > 0$ for all $x_i \in \mathcal{X}$. Thus, by repeating the argument along this sequence, we know $M = h(x_0) = h(x_1) = \cdots = h(x_n) = h(y)$. □

Since h is constant, the kernel of $P - I$ only includes the space of constant functions and, thus, has a one-dimensional space of solutions. Additionally, if the column rank of $P - I$ has dimension $|\mathcal{X}| - 1$, so does the row rank. Hence, $\pi = \pi P$ has a one-dimensional space of solutions which contains a unique vector whose entries sum to 1. □

Note that stationary distribution exists as long as the chain is finite, but uniqueness requires the chain to be irreducible. We refer to the previous result to show a simple yet important result of a stationary distribution.

Theorem 1.20. *If the Markov chain is irreducible, then in its unique stationary distribution, every state has positive probability i.e values in* $(0, 1]$.

Proof. We will argue by contradiction. Suppose π is stationary for P and that $\pi(x) = 0$ for some $x \in \mathcal{X}$. Let A be the set of all states for which $\pi(x) = 0$ and B be the set of all states for which $\pi(y) > 0$. Then for all $x \in A$,

$$\pi(x) = \sum_{y \in \mathcal{X}} P(y, x)\pi(y) = 0.$$

And so for all $y \in B$ and $x \in A$, $P(y, x) = 0$. Thus, it is impossible to transition from set B to set A which contradicts our assumption that P is irreducible. □

We refer to our previous example on graphs to provide an example of a stationary distribution.

Example 1.21 (Random Walk on a Graph)**.** *Consider the simple random walk on a graph from 1.4. For any vertex* $y \in V$*,*

$$\sum_{x \in V} \deg(x)P(x, y) = \sum_{x \sim y} \frac{\deg(x)}{\deg(x)} = \deg(y).$$

This says that the vector π' *whose* i^{th} *coordinate is the degree of the* i^{th} *vertex of* G *satisfies* $\pi' P = \pi'$. *Then to turn this into a probability vector* π *we normalize by* $\sum_{y \in V} \deg(y) = 2|E|$.

Hence for all $y \in V$*,*

$$\pi(y) = \frac{\deg(y)}{2|E|}.$$

1.3 Reversibility

In order to test for stationarity we may use the Detailed Balance Equations defined below.

Definition 1.22 (Detailed Balance Equations)**.** *A probability distribution* π *on* $\mathcal{X}$ *with transition matrix* P *satisfies the detailed balance equations if*

$$\pi(x)P(x, y) = \pi(y)P(y, x) \ \forall x, y \in \mathcal{X}.$$

Proposition 1.23. *Let* P *be the transition matrix of a Markov chain with state space* $\mathcal{X}$. *Any distribution* π *satisfying the detailed balance equations is stationary for* P.

Proof. Since P is stochastic and π satisfies the detailed balance equations,

$$\sum_{y\in\mathcal{X}} \pi(y)P(y,x) = \sum_{y\in\mathcal{X}} \pi(x)P(x,y) = \pi(x).$$

Hence, π is stationary for P. □

We now define reversibility which, in short, says that given a sample sequence of consecutive states of the chain, there is no way to tell whether the sequence is being presented in order of increasing time or of decreasing time.

Definition 1.24 (Reversible)**.** *A chain which satisfies the detailed balanced equations is called reversible since:*

$$\begin{aligned}
P_\pi(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) &= \pi(x_0)P(x_0,x_1)\ldots P(x_{n-1},x_n) \\
&= \pi(x_1)P(x_1,x_0)P(x_1,x_2)\ldots P(x_{n-1},x_n) \\
&= P(x_1,x_0)\pi(x_1)P(x_1,x_2)\ldots P(x_{n-1},x_n) \\
&= \ldots \\
&= P(x_1,x_0)\ldots P(x_{n-1},x_{n-2})\pi(x_n)P(x_n,x_{n-1}) \\
&= \pi(x_n)P(x_n,x_{n-1})\ldots P(x_1,x_0) \\
&= P_\pi(X_n = x_0, X_{n-1} = x_1, \ldots, X_0 = x_n).
\end{aligned}$$

Once again, we use the example of a random walk on a graph to portray how to test for reversibility.

Example 1.25 (Random Walk on a graph G)**.** *Consider the simple random walk on a graph G. We saw in Example 1.21 that the stationary distribution $\pi(x) = \frac{\deg(x)}{2|E|}$ for all $x \in \mathcal{X}$.*

This chain is reversible since:

$$\pi(x)P(x,y) = \frac{\deg(x)}{2|E|}\frac{\mathbb{1}_{\{x\sim y\}}}{\deg(x)} = \frac{\mathbb{1}_{\{x\sim y\}}}{2|E|} = \frac{\deg(y)}{2|E|}\frac{\mathbb{1}_{\{y\sim x\}}}{\deg(y)} = \pi(y)P(y,x).$$

1.4 Glauber Dynamics

In this last part, we introduce Glauber Dynamics, a Markov chain which we will use for our 2Δ proof. An example, with a visual, is provided to understand a transition.

Definition 1.26 (Proper Coloring)**.** *A proper q-coloring of a graph $G = (V,E)$ is an assignment of colors to the vertices V, subject to the constraint that neighboring vertices do not receive the same color.*

Alternatively, a proper q-coloring of a graph $G = (V,E)$ is an element x of $\{1,\ldots,q\}^V$, the set of functions from V to $\{1,\ldots,q\}$, such that $x(v) \neq x(w)$ for all neighboring vertices w of v.

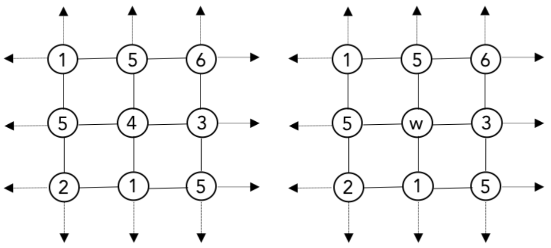
Definition 1.27 (Allowable Color)**.** *For a given coloring x and a vertex v, a color j is allowable at v, if j is different from all colors assigned to the neighbors of v.*

We denote by $A_v(x)$, the set of allowable colors at v in x.

We will now construct a Markov chain on the set of proper q-colorings of G.

Definition 1.28 (Glauber Dynamics on Proper Colorings)**.** *Given a proper q-coloring, generate a new coloring by selecting a vertex $v \in V$ uniformly at random. From the set of allowable colors at v, select a color, j uniformly at random and recolor vertex v with j.*

Example 1.29. *Given a proper 6-coloring on $G(V, E)$, update w from the colors $\{2, 4, 6\}$, each with probability $p = \frac{1}{3}$.*



Here, we provide a proof that the Glauber Dynamics has a uniform stationary distribution.

Claim 1.30. *On any graph, the Glauber Dynamics has a uniform stationary distribution.*

Proof. 1.30 Since we are only updating one vertex at a time in this Markov chain, transitions from one coloring to another are only permitted between colorings differing at a single vertex. Let x and y agree everywhere but at v. Then,

$$P(x, y) = \frac{1}{|V|} * \frac{1}{|A_v(x)|} = \frac{1}{|V|} * \frac{1}{|A_v(y)|} = P(y, x).$$

Thus, the detailed balance equations are satisfied by the uniform distribution. □

Claim 1.31. *The Glauber Dynamics is irreducible for $q > \Delta + 1$.*

Proof. To go from any coloring x to any coloring y, consider an arbitrary ordering of the vertices and attempt to recolor them in that order. Suppose a coloring x' has been reached, that agrees with y up to (but not including) vertex v. When attempting to recolor vertex v with a color c (the color that y has at v), if c does not appear among the neighbors of w, then the recoloring does not encounter any interference. Now suppose that there exists some neighbors w which has the desired color c. Note that w must appear later than v in the ordering since, up to this point, x' and y agree. Thus, for each w, we recolor w with an allowable color not including c, requiring $q > \Delta + 1$. From there, we can recolor v with c as all interferences have been eliminated. □

Claim 1.32. *The Glauber Dynamics is aperiodic.*

Proof. Given a coloring x, since the probability of staying put is

$$P(x,x) = \frac{1}{|V|} * \frac{1}{|A_v(x)|} > 0$$

then the chain is aperiodic. □

2 Markov Chain Mixing and Coupling

Now that the existence of, uniqueness of, and a test for a stationary distribution have been introduced, we can start talking about convergence. In order to do so we must define a measure of distance from the stationary distribution.

2.1 Total Variation Distance

Here, we give a tool to measure the distance between two probability distributions.

Definition 2.1 (Total Variation Distance)**.** *The total variation distance between two probability distributions μ and ν on $\mathcal{X}$ is given by*

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

where A is an event. Note that since X is finite, all subsets are events.

Claim 2.2. *Total variation distance satisfies the triangle inequality.*

Proof.

$$\begin{aligned}
\|\mu - \eta\|_{TV} &= \max_{A \subseteq \mathcal{X}} |\mu(A) - \nu(A) + \nu(A) - \eta(A)| \\
&\leq \max_{A \subseteq \mathcal{X}} \Big(|\mu(A) - \nu(A)| + |\nu(A) - \eta(A)| \Big) \\
&\leq \max_{A \subseteq \mathcal{X}} |\mu(A) - \nu(A)| + \max_{A' \subseteq \mathcal{X}} |\nu(A') - \eta(A')| \\
&= \|\mu - \nu\|_{TV} + \|\nu - \eta\|_{TV}.
\end{aligned}$$

□

2.2 Coupling

We introduce the concept of coupling which will be crucial to proving convergence to the stationary distribution for both Glauber and Flip Dynamics.

Definition 2.3 (Coupling)**.** *A coupling of two probability distributions μ and ν is a pair of random variables (X, Y) each defined on a probability space $\mathcal{X}$ and $\mathcal{Y}$ respectively, so that the possible values of (X, Y) range over $\mathcal{X} \times \mathcal{Y}$. The marginal distribution of X is μ and the marginal distribution of Y is ν i.e $P\{X = x\} = \mu(x)$ and $P\{Y = y\} = \nu(y)$.*

2.3 Coupling and Total Variation Distance

Here, we provide an alternative definition of total variation distance which uses coupling. This will allow us to study convergence via coupling

Proposition 2.4. *Let μ and ν be two probability distributions on $\mathcal{X}$ and let (X, Y) be a coupling of μ and ν. Then,*

$$\|\mu - \nu\|_{TV} = \inf_{(X,Y)} \{P\{X \neq Y\}\}.$$

Remark 2.5. *We will in fact show that there is a coupling (X, Y), referred to as the optimal coupling, which attains the infimum in the equality above.*

Proof. For any coupling (X, Y) of μ and ν and any event $A \subset \mathcal{X}$,

$$\begin{aligned}
\mu(A) - \nu(A) &= P\{X \in A\} - P\{Y \in A\} \\
&= P\{X \in A, Y \in A\} + P\{X \in A, Y \notin A\} \\
&\quad -P\{X \in A, Y \in A\} - P\{X \notin A, Y \in A\} \\
&= P\{X \in A, Y \notin A\} - P\{X \notin A, Y \in A\} \\
&\leq P\{X \in A, Y \notin A\} \\
&\leq P\{X \neq Y\}.
\end{aligned}$$

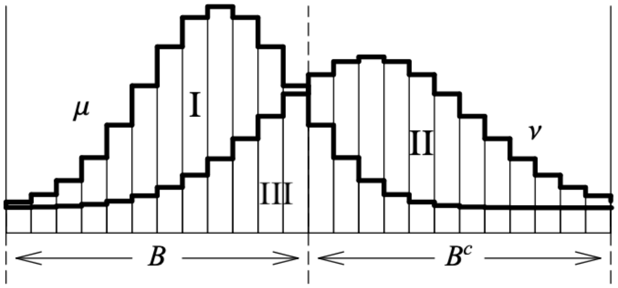
Thus,

$$\|\mu - \nu\|_{TV} = \max_{A \subseteq \mathcal{X}} |\mu(A) - \nu(A)| \leq \inf_{(X,Y)} \{P\{X \neq Y\}\}.$$

In order to prove equality, we will show, by construction, that there exists a coupling (X, Y) which attains this infimum.

Since μ and ν are finite, we can assume, without loss of generality, that both of these probability distributions have the same range by allowing both μ and ν to take on certain values with probability zero.

Now, we order the states x in order to get the following histogram [1]:



Where $B = \{x : \mu(x) \geq \nu(x)\}$, Region I has area $\mu(B) - \nu(B)$, Region II has area $\nu(B^c) - \mu(B^c)$, and region III is bounded by $\mu(x) \wedge \nu(x) = \min\{\mu(x), \nu(x)\}$.

Now, let

$$\begin{aligned} p &= \sum_{x\in\mathcal{X}} \mu(x) \wedge \nu(x) \\ &= \sum_{x\in\mathcal{X}} \min\{\mu(x), \nu(x)\} \\ &= \sum_{\substack{x\in\mathcal{X} \\ \mu(x)\leq\nu(x)}} \mu(x) + \sum_{\substack{x\in\mathcal{X} \\ \mu(x)>\nu(x)}} \nu(x) \\ &= \sum_{\substack{x\in\mathcal{X} \\ \mu(x)\leq\nu(x)}} \mu(x) + \sum_{\substack{x\in\mathcal{X} \\ \mu(x)>\nu(x)}} \nu(x) + \sum_{\substack{x\in\mathcal{X} \\ \mu(x)>\nu(x)}} \mu(x) - \sum_{\substack{x\in\mathcal{X} \\ \mu(x)>\nu(x)}} \mu(x) \\ &= 1 - \sum_{\substack{x\in\mathcal{X} \\ \mu(x)>\nu(x)}} [\mu(x) - \nu(x)]. \end{aligned}$$

To complete the proof of Proposition 2.4, we need the following claim:

Claim 2.6.

$$\|\mu - \nu\|_{TV} = \sum_{\substack{x\in\mathcal{X} \\ \mu(x)>\nu(x)}} [\mu(x) - \nu(x)]$$

Proof. Let $A \subset \mathcal{X}$ be any event.

Then, since any $x \in A \cap B^c$ satisfies $\mu(x) - \nu(x) < 0$ and any $x \in B$ satisfies $\mu(x) - \nu(x) \geq 0$,

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

By exact parallel reasoning,

$$\nu(A) - \mu(A) \leq \nu(A\cap B^c) - \mu(A\cap B^c) \leq \nu(B^c) - \mu(B^c).$$

Additionally, since the total area under μ and ν is 1,

$$\mu(B) + \mu(B^c) = 1 = \nu(B^c) + \nu(B).$$

Thus,

$$\mu(B) - \nu(B) = \nu(B^c) - \mu(B^c).$$

Now, take $A = B \vee B^c$,

$$\begin{aligned}
\|\mu - \nu\|_{TV} &= \max_{A \subseteq \mathcal{X}} |\mu(A) - \nu(A)| \\
&= \frac{1}{2}[\mu(B) - \nu(B) + \nu(B^c) - \mu(B^c)] \\
&= \sum_{x \in B} [\mu(x) - \nu(x)] \\
&= \sum_{\substack{x \in \mathcal{X} \\ \mu(x) > \nu(x)}} [\mu(x) - \nu(x)].
\end{aligned}$$

□

We now return to the proof of Proposition 2.4. By Claim 2.6,

$$p = 1 - \|\mu - \nu\|_{TV}.$$

Here, we are trying to get X and Y to agree with as high a probability a possible. We can get them to agree with probability p by, with probability p, choosing a point uniformly from region III, and letting X equal Y equal the index of whatever column in the histogram the point is in. We use the remaining probability $(1-p)$ to correctly fix the marginal probabilities of X and Y. For X, we choose a point uniformly from region I, and let X equal the index of whatever column in the histogram the point is in; for Y we do the same with region II. In order to do this, we define our coupling as follows:

Step 1: Flip a coin with probability of heads p.
Step 2:

1. If the flip comes up heads:
   Choose a value Z according to the probability distribution

   $$\gamma_{\text{III}}(x) = \frac{\min\{\mu(x), \nu(x)\}}{p}$$

   and set $X = Y = Z$.

2. If the flip comes up tails:
   Choose X according to the probability distribution

   $$\gamma_{\text{I}}(x) = \begin{cases} \dfrac{\mu(x) - \nu(x)}{1-p} & \text{if } \mu(x) > \nu(x) \\ 0 & \text{otherwise} \end{cases}$$

   and independently choose Y according to the probability distribution

   $$\gamma_{\text{II}}(y) = \begin{cases} \dfrac{\nu(y) - \mu(y)}{1-p} & \text{if } \mu(y) > \nu(y) \\ 0 & \text{otherwise} \end{cases}$$

Claim 2.6, ensures that γ_{I} and γ_{II} are probability distributions.

Additionally the probability that $X = x$ under this random process is

$$p\gamma_{\mathrm{III}}(x) + (1-p)\gamma_{\mathrm{I}}(x) = \begin{cases} p\dfrac{\nu(x)}{p} + (1-p)\dfrac{\mu(x)-\nu(x)}{1-p} & \text{if } \mu(x) > \nu(x) \\ p\dfrac{\mu(x)}{p} + 0 & \text{otherwise} \end{cases}$$

and the probability that $Y = y$ is

$$p\gamma_{\mathrm{III}}(x) + (1-p)\gamma_{\mathrm{II}}(y) = \begin{cases} p\dfrac{\mu(y)}{p} + (1-p)\dfrac{\nu(y)-\mu(y)}{1-p} & \text{if } \nu(y) > \mu(y) \\ p\dfrac{\nu(y)}{p} + 0 & \text{otherwise} \end{cases}$$

Thus,

$$p\gamma_{\mathrm{III}} + (1-p)\gamma_{\mathrm{I}} = \mu$$

and

$$p\gamma_{\mathrm{III}} + (1-p)\gamma_{\mathrm{II}} = \nu.$$

So the distribution of X is μ and that of Y is ν and since γ_{I} and γ_{II} are positive on disjoint subsets of $\mathcal{X}$,

$$P\{X \neq Y\} = 1 - p = \|\mu - \nu\|_{TV}.$$

□

2.4 Convergence Theorem

Now that we have proven that every chain has a stationary distribution, and that under the additional hypothesis of irreducibility the stationary distribution is unique, we are going to show that adding one more hypothesis, aperiodicity, suffices to have convergence to stationarity from any starting distribution. More specifically, this means that given any distribution μ on $\mathcal{X}$, the sequence of distributions $\mu, \mu P, \mu P^2, \ldots$ converges to π, the unique stationary distribution.

Theorem 2.7 (Convergence Theorem)**.** *Suppose P is irreducible and aperiodic, with stationary distribution π. Then there exists constants $\alpha \in (0,1)$ and $C > 0$ such that*

$$\max_{x\in\mathcal{X}} \|P^r(x,.) - \pi\|_{TV} \leq C\alpha^r.$$

Proof. The proof will require two claims, Claims 2.8 and 2.9, which we introduce as they are needed.

The first claim establishes a certain uniformity of aperiodicity. By definition, aperiodicity asserts that for each pair of states x, y there is an r such that there is a non-zero probability of going from x to y in r steps; but in fact there is a uniform r that works for all x, y simultaneously.

Claim 2.8. *If P is aperiodic and irreducible, then there is an integer r_0 such that $P^r(x,y) > 0$ for all $x, y \in \mathcal{X}$ and $r \geq r_0$.*

Proof. 2.8 Fix a state x. Since P is aperiodic, the gcd of $\mathcal{T}(x) = \{t \geq 1 : P^t(x,x) > 0\}$ is 1. Additionally, $\mathcal{T}(x)$ is a set of non-negative integers closed under addition as for $s, t \in \mathcal{T}(x)$,

$$P^{s+t}(x,x) \geq P^s(x,x)P^t(x,x) > 0 \Rightarrow x + y \in \mathcal{T}(x).$$

Thus, by Schur's Lemma given on page 19 of Levin, Perez and Wilmer [1], there exists some integer $t(x)$ such that for all $t \geq t(x)$, t can be written as a linear combination of elements of $\mathcal{T}(x)$ with non-negative integer coefficients i.e., $t \in \mathcal{T}(x)$.

Since P is irreducible we also have that for any $y \in \mathcal{X}$, there exists an r such that $P^r(x,y) > 0$.

Therefore, for $t \geq t(x) + r$,

$$P^t(x,y) \geq P^{t-r}(x,x)P^r(x,y) > 0.$$

Thus, taking $t \geq max_{x,y\in\mathcal{X}}\{t(x) + r(x,y)\}$, we have $P^t(x,y) > 0$ for all $x, y \in \mathcal{X}$. □

Let Π be the matrix with $|\mathcal{X}|$ rows, each of which is a copy of the row vector π. Then by Claim 2.8 we know that for sufficiently small $\delta > 0$ and for all $x, y \in \mathcal{X}$,

$$P^t(x,y) \geq \delta\pi(y).$$

Now, let $\theta = 1 - \delta$ and $P^t = (1-\theta)\Pi + \theta Q$. Then Q is stochastic as

$$\begin{aligned}
1 &= \sum_{y\in\mathcal{X}} P^t(x,y) \\
&= \delta \sum_{y\in\mathcal{X}} \Pi(x,y) + (1-\delta) \sum_{y\in\mathcal{X}} Q(x,y) \\
&= \delta \sum_{y\in\mathcal{X}} \pi(y) + (1-\delta) \sum_{y\in\mathcal{X}} Q(x,y) \\
&= \delta + (1-\delta) \sum_{y\in\mathcal{X}} Q(x,y) \\
&= \sum_{y\in\mathcal{X}} Q(x,y).
\end{aligned}$$

Also note that $Q\Pi = \Pi$ since

$$\begin{aligned}
Q\Pi(x,y) &= \sum_{k=1}^{|\mathcal{X}|} Q(x,k)\Pi(k,y) \\
&= \sum_{k=1}^{|\mathcal{X}|} Q(x,k)\pi(y) \\
&= \pi(y)\sum_{k=1}^{|\mathcal{X}|} Q(x,k) \\
&= \pi(y) \\
&= \Pi(x,y)
\end{aligned}$$

and $\Pi M = \Pi$ for any stochastic M such that $\pi M = \pi$ since

$$\Pi M(x,y) = \sum_{k=1}^{|\mathcal{X}|} \Pi(x,k)M(k,y) = \sum_{k=1}^{|\mathcal{X}|} \pi(k)M(k,y) = \pi M(y) = \pi(y) = \Pi(x,y).$$

Claim 2.9. *$P^{tk} = (1-\theta^k)\Pi + \theta^k Q^k$ for $k \geq 1$.*

Proof. We will proceed using a proof by induction.

For $k = 1$, the claim follows by the definition of P^t provided above.

Assume Claim 2.9 holds for $k = n$. Then for $k = n+1$,

$$\begin{aligned}
P^{t(n+1)} &= P^{tn}P^t \\
&= [(1-\theta^n)\Pi + \theta^n Q^n]P^t \\
&= (1-\theta^n)\Pi P^t + \theta^n Q^n[(1-\theta)\Pi + \theta Q] \\
&= (1-\theta^n)\Pi P^t + \theta^n Q^n \Pi - \theta^{n+1}Q^n\Pi + \theta^{n+1}Q^{n+1} \\
&= (1-\theta^n)\Pi + \theta^n\Pi - \theta^{n+1}\Pi + \theta^{n+1}Q^{n+1} \\
&= (1-\theta^{n+1})\Pi + \theta^{n+1}Q^{n+1}.
\end{aligned}$$

□

From Claim 2.9 we get,

$$\begin{aligned}
P^{tk+j} - \Pi &= [(1-\theta^k)\Pi + \theta^k Q^k]P^j - \Pi \\
&= (1-\theta^k)\Pi P^j + \theta^k Q^k P^j - \Pi \\
&= \Pi P^j - \theta^k \Pi P^j + \theta^k Q^k P^j - \Pi \\
&= \Pi - \theta^k\Pi + \theta^k Q^k P^j - \Pi \\
&= \theta^k(Q^k P^j - \Pi).
\end{aligned}$$

Thus by Claim 2.6,

$$
\begin{aligned}
\|P^{tk+j}(x_0,.) - \pi\|_{TV} &= \frac{1}{2}\sum_{i=1}^{|\mathcal{X}|} |P^{tk+j}(x_0,i) - \pi(i)| \\
&= \frac{1}{2}\sum_{i=1}^{|\mathcal{X}|} |P^{tk+j}(x_0,i) - \Pi(x_0,i)| \\
&= \frac{1}{2}\sum_{i=1}^{|\mathcal{X}|} \theta^k [Q^k P^j(x_0,i) - \Pi(x_0,i)] \\
&= \frac{1}{2}\sum_{i=1}^{|\mathcal{X}|} \theta^k [Q^k P^j(x_0,i) - \pi(i)] \\
&= \theta^k \|Q^k P^j(x_0,.) - \pi\|_{TV} \\
&\leq \theta^k \max \|Q^k P^j(x_0,.) - \pi\|_{TV} \\
&\leq \theta^k.
\end{aligned}
$$

For $tk + j = r$, $j \in \{0, \dots, t-1\}$, and since $0 < \theta < 1$,

$$\theta^k = \theta^{\frac{(r-j)}{t}} = \theta^{\frac{r}{t}} \frac{1}{\theta^{\frac{j}{r}}} < \theta^{\frac{r}{t}} \frac{1}{\theta}.$$

Thus, taking $\alpha = \theta^{\frac{1}{t}}$ and $C = \frac{1}{\theta}$ yields

$$\max_{x \in \mathcal{X}} \|P^r(x,.) - \pi\|_{TV} \leq C\alpha^r.$$

□

Note that for a chain to have a unique stationary distribution, it is required to be irreducible; but in order to have convergence to stationarity our chain must also be aperiodic.

2.5 Mixing Time

Now that we know that a finite, aperiodic, and irreducible chain converges to its unique stationary distribution, regardless of starting state, we must now evaluate how fast this happens by introducing the mixing time and some useful related properties.

Definition 2.10. *We denote by $d(t)$ the distance between the stationary distribution, and the stationary distribution of the chain after t steps, given that it starts from state x, taking the worst case over all possible initial states x. Thus,*

$$d(t) := \max_{x \in \mathcal{X}} \|P^t(x,.) - \pi\|_{TV}.$$

We denote by $\overline{d}(t)$ the worst case distance between the distributions reached after t steps, when started at two different states x and y. Thus,

$$\overline{d}(t) := \max_{x,y \in \mathcal{X}} \|P^t(x,.) - P^t(y,.)\|_{TV}.$$

Note that the two parameters ($d(t)$, $\overline{d}(t)$) are closely related and that the next lemma establishes a connection that will allow for toggling back and forth between the two, as needed.

Lemma 2.11.

$$d(t) \leq \overline{d}(t) \leq 2d(t)$$

Proof. We begin by establishing $d(t) \leq \overline{d}(t)$. By the triangle inequality,

$$\begin{aligned} \overline{d}(t) &= \max_{x,y\in\mathcal{X}} \|P^t(x,.) - \pi + \pi - P^t(y,.)\|_{TV} \\ &\leq \max_{x,y\in\mathcal{X}} \|P^t(x,.) - \pi\| + \max_{x,y\in\mathcal{X}} \|\pi - P^t(y,.)\|_{TV} \\ &= 2d(t). \end{aligned}$$

Next, we show that $\overline{d}(t) \leq 2d(t)$. Note from that definition of stationary distribution that for any set A,

$$\pi(A) = \sum_{x\in A} \pi(x) = \sum_{x\in A}\sum_{y\in\mathcal{X}} \pi(y)P^t(y,x) = \sum_{y\in\mathcal{X}} \pi(y)P^t(y,A).$$

Then by the triangle inequality and total variation definition,

$$\begin{aligned} |P^t(x,A) - \pi(A)| &= \left|\sum_{y\in\mathcal{X}} \pi(y)[P^t(y,A) - P^t(y,A)]\right| \\ &\leq \sum_{y\in\mathcal{X}} \pi(y)\left|P^t(y,A) - P^t(y,A)\right| \\ &\leq \max_{A'\subseteq\mathcal{X}} \sum_{y\in\mathcal{X}} \pi(y)\left|P^t(y,A') - P^t(y,A')\right| \\ &= \sum_{y\in\mathcal{X}} \pi(y) \max_{A'\subseteq\mathcal{X}} \left|P^t(y,A') - P^t(y,A')\right| \\ &= \sum_{y\in\mathcal{X}} \pi(y)\|P^t(x,.) - P^t(y,.)\|_{TV}. \end{aligned}$$

Maximizing over A gives us

$$\begin{aligned} \|P^t(x,.) - \pi\|_{TV} &= \max_{A\subseteq\mathcal{X}} |P^t(x,A) - \pi(A)| \\ &= \sum_{y\in\mathcal{X}} \pi(y)\|P^t(x,.) - P^t(y,.)\|_{TV} \\ &\leq \sum_{y\in\mathcal{X}} \pi(y) \max_{y'\in\mathcal{X}} \|P^t(x,.) - P^t(y',.)\|_{TV} \\ &= \max_{y'\in\mathcal{X}} \|P^t(x,.) - P^t(y',.)\|_{TV}. \end{aligned}$$

Now, maximizing over x yields

$$\max_{x\in\mathcal{X}} \|P^t(x,.) - \pi\|_{TV} \leq \max_{x,y\in\mathcal{X}} \|P^t(x,.) - P^t(y,.)\|_{TV}.$$

Thus, we get our desired result

$$\overline{d}(t) \leq 2d(t).$$

□

Lemma 2.12.

$$\overline{d}(s+t) \leq \overline{d}(s)\overline{d}(t)$$

Proof. Fix $x, y \in \mathcal{X}$, and let (X_s, Y_s) be the optimal coupling of $P^s(x,.)$ and $P^s(y,.)$. Then,

$$\|P^s(x,.) - P^s(y,.)\|_{TV} = P\{X_s \neq Y_s\}.$$

Note that,

$$P^{s+t}(x,w) = \sum_z P^s(x,s)P^t(z,w) = \sum_z P\{X_s = z\}P^t(z,w) = E(P^t(X_s,w)).$$

For a set $A \subseteq \mathcal{X}$,

$$\begin{array}{rcl} \sum_{w\in A}[P^{s+t}(x,w) - P^{s+t}(y,w)] & = & P^{s+t}(x,A) - P^{s+t}(y,A) \\ & = & E(P^t(X_s,A)) - E(P^t(Y_s,A)) \\ & = & E(P^t(X_s,A) - P^t(Y_s,A)). \end{array}$$

Maximizing over A gives

$$\begin{array}{rcl} \|P^{t+s}(X_s,.) - P^{t+s}(Y_s,.)\|_{TV} & = & E(P^t(X_s,A') - P^t(Y_s,A')) \\ & \leq & E(\|P^t(X_s,.) - P^t(Y_s,.)\|_{TV}) \\ & \leq & E(\max_{x,y\in\mathcal{X}} \|P^t(X_s,.) - P^t(Y_s,.)\|_{TV}) \\ & \leq & E(\overline{d}(t)) \\ & \leq & E(\overline{d}(t)1_{\{X_s\neq Y_s\}}) \\ & = & P\{X_s \neq Y_s\}\overline{d}(t). \end{array}$$

Thus, maximizing over $x, y \in \mathcal{X}$,

$$\overline{d}(s+t) \leq \overline{d}(s)\overline{d}(t).$$

□

Remark 2.13. *Note that Lemma 2.11 and 2.12 imply that for c and t positive integers,*

$$d(ct) \leq \overline{d}(ct) \leq (\overline{d}(t))^c.$$

Definition 2.14 (Mixing Time). *For each fixed ε, the mixing time is the first time at which the chain is certain to be within distance ε, measured by total variation distance, from stationarity, regardless of the initial state. Thus,*

$$t_{mix}(\varepsilon) := \min\{t : d(t) \leq \varepsilon\}$$

and

$$t_{mix} := t_{mix}(1/4).$$

We say that a chain is rapidly mixing if the mixing time is bounded by a polynomial in n, where $n = |V|$ in our examples on graphs.

Remark 2.15. *Lemma 2.11 and Remark 2.13 show that for l a positive integer,*

$$d(lt_{mix}(\varepsilon)) \leq (\overline{d}(t_{mix}(\varepsilon)))^l \leq (2d(t_{mix}(\varepsilon)))^l \leq (2\varepsilon)^l.$$

And for $\varepsilon = 1/4$,

$$d(lt_{mix}) \leq (2 \times 1/4)^l = 2^{-l}.$$

Taking $l = \lceil \log_2(\varepsilon^{-1}) \rceil$,

$$\begin{aligned} d(\lceil \log_2(\varepsilon^{-1}) \rceil t_{mix}) &\leq 2^{-\lceil \log_2(\varepsilon^{-1}) \rceil} \\ &\leq 2^{-(\log_2(\varepsilon^{-1}))} \\ &= \varepsilon. \end{aligned}$$

Thus,

$$t_{mix}(\varepsilon) \leq \lceil \log_2(\varepsilon^{-1}) \rceil t_{mix}.$$

What the result obtained in this remark shows is that once the chain has been run for enough steps to get within 1/4 of stationarity, in total variation distance, only a small multiple of that number of steps is needed to ensure that the distribution is within ε of stationarity. More specifically, that multiple is $\log(1/\varepsilon)$. So, for example, to get within 1/16 of π requires only 4 times the number of steps needed to get within 1/4 of π; and to get within 1/256 of π only requires 8 times the number of steps needed to get within 1/4 of π. In general, the result of this remark establishes that once the distribution has gotten within 1/4 of π, a linear factor increase in the number of steps the chain is run for leads to an exponential decay in the distance of the distribution from π.

2.6 Upper Bound on the Mixing Time

In this section, we define a Markovian coupling, which will be used throughout the rest of this paper, and provide an upper bound on the mixing time in terms of the coupling time.

Definition 2.16 (Markovian Coupling). *Given a Markov Chain on the state space $\mathcal{X}$ with transition matrix P, a Markovian Coupling of two P-chains is* *a Markov chain $\{(X_t, Y_t)\}_{t\geq 0}$ with state space $\mathcal{X} \times \mathcal{X}$ which satisfies, for all x, y, x', y',*

$$P\{X_{t+1} = x'|X_t = x, Y_t = y\} = P(x, x')$$

and

$$P\{Y_{t+1} = y'|X_t = x, Y_t = y\} = P(y, y').$$

Thus, in a Markovian coupling, if one just looks at the progression of each coordinate of $\{(X_t, Y_t)\}_{t\geq 0}$ on its own, one sees a perfect copy of the P-chain.

Remark 2.17. *From now on, all couplings introduced will be Markovian, with the particular property that the two chains stay together at all times after their first simultaneous visit to a single state. That is, for $t \geq s$,*

$$X_s = Y_s \Rightarrow X_t = Y_t.$$

This will be achieved for each coupling under discussion by running the chains according to that coupling until they meet, and then simply running them together.

Definition 2.18. *τ_{couple} is the time when the two coupled chains X, Y, starting from states x, y respectively, first agree with each other. Thus,*

$$\tau_{couple} := \min\{t : X_s = Y_s, \forall s \geq t\}.$$

Theorem 2.19. *Let $\{(X_t, Y_t\}$ be a coupling for which $X_0 = x$ and $Y_0 = y$. Then,*

$$\|P^t(x, .) - P^t(y', .)\|_{TV} \leq P_{x,y}\{\tau_{couple} > t\}$$

where $P_{x,y}$ is the joint probability mass of the coupled chains X, Y, starting from states x, y respectively.

Proof. Since $P^t(x, z) = P_{x,y}\{X_t = z\}$ and $P^t(y, z) = P_{x,y}\{Y_t = z\}$, (X_t, Y_t) is a coupling of $P^t(x, .)$ with $P^t(y, .)$. Thus, by definition

$$\|P^t(x, .) - P^t(y, .)\|_{TV} \leq P_{x,y}\{X_t \neq Y_t\} = P_{x,y}\{\tau_{couple} > t\}.$$

□

Corollary 2.20. *Suppose that for each $x, y \in \mathcal{X}$, there is a coupling (X_t, Y_t) with $X_0 = x$ and $Y_0 = y$. Then for each (X_t, Y_t),*

$$d(t) \leq \max_{x,y\in\mathcal{X}} P_{x,y}\{\tau_{couple} > t\}$$

and thus

$$t_{mix} \leq 4 \max_{x,y\in\mathcal{X}} E_{x,y}(\tau_{couple}).$$

Proof. The first inequality follows directly from the definition of $\overline{d}(t)$, Theorem 2.19, and Lemma 2.11.

The second follows from Markov's inequality:

$$P\{X > t\} \leq \frac{E(X)}{t},$$

so, setting $t = 4E_{x,y}(\tau_{couple})$ we get

$$\begin{aligned} d(t) &\leq \max_{x,y \in \mathcal{X}} P_{x,y}\{\tau_{couple} > t\} \\ &\leq \max_{x,y \in \mathcal{X}} \frac{E_{x,y}(\tau_{couple})}{t} \\ &= \frac{1}{4}. \end{aligned}$$

□

From these results, the goal will be to try and find couplings of pairs of chains that potentially start from very different initial states, but are designed so that with as high a probability as possible, they move closer to each other at each step. The faster the two coupled chains move together, the smaller the coupling time and, thus, the faster the mixing.

2.7 First Bound on q for Rapid Mixing on Colorings

This section serves as the first proof of rapid mixing for sampling uniform colorings. In order to do so, we must define a Grand coupling and the Metropolis chain.

Definition 2.21 (Grand Coupling)**.** *A* Grand Coupling *is collection of random variables $\{X_t^x : x \in \mathcal{X}, t = 0, 1, 2, \dots\}$ such that for each x, the sequence $(X_t^x)_{t=0}^{\infty}$ is a Markov Chain with transition matrix P started from x.*

Definition 2.22 (Metropolis Chain on Colorings)**.** *Let $G = (V, E)$ be a graph, and x a q-coloring of G (recall: a q-coloring of a graph G is an assignment of colors to the vertices, from a palette of q colors, such that no two adjacent vertices receive the same color). Choose a pair (v, c), where v is a vertex and c is a color, uniformly at random from $V \times \{1, \dots, q\}$ and independently from the past. Re-color the graph by changing the color at v to c. The Metropolis rule accepts the proposed re-coloring if c is allowable and different from $x(v)$ (that is, if the re-coloring is a valid q-coloring different from x) and rejects it otherwise. Thus, the probability of remaining in the current coloring given v is selected is $1 - \frac{a-1}{q}$ where a is the set of allowable colors at v. Note that we apply the Metropolis rule to this chain, where π is the probability measure which is uniform over the space of proper q-colorings.*

We know define a simultaneous coupling of all the Metropolis chains on the space of colorings of a graph, starting from all possible initial colorings.

Definition 2.23 (Grand Coupling on Colorings). *For each $x \in \tilde{\mathcal{X}}$ where $\tilde{\mathcal{X}}$ is the space of all colorings of G, the coloring X_t^x is updated just as for a single Metropolis chain using the same vertex and color.*

Here, with the use of this chain, we prove our first bound on q of 3Δ.

Theorem 2.24. *Let G be a graph with n vertices and maximal degree Δ. For the Metropolis chain on q-colorings of G, if $q > 3\Delta$ then,*

$$t_{mix}(\varepsilon) \leq \left\lceil \frac{1}{1 - \frac{3\Delta}{q}} n \log(n) + \log\left(\frac{1}{\varepsilon}\right) \right\rceil.$$

In particular,

$$t_{mix} \leq Cn \log n$$

for some constant C (depending on Δ and q).

Proof. For two colorings $x, y \in \tilde{\mathcal{X}}$, define

$$\rho(x, y) := \sum_{v \in V} 1_{\{x(v) \neq y(v)\}}.$$

Claim 2.25. *ρ is a metric on $\tilde{\mathcal{X}}$.*

Proof. Clearly $\rho(x, y) \geq 0$ as the minimum number of vertices where x and y disagree is 0.

If $\rho(x, y) = 0$, x and y agree at every vertex so $x = y$.

The number of vertices where x and y disagree is the same for $\rho(x, y)$ and $\rho(y, x)$ so $\rho(x, y) = \rho(y, x)$.

Lastly, to show $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$ we break it down in four cases for $v \in V$:

Case 1 $x(v) = y(v)$ and $y(v) = z(v)$, so $x(v) = z(v)$.

Case 2 $x(v) = y(v)$ and $y(v) \neq z(v)$, so $x(v) \neq z(v)$.

Case 3 $x(v) \neq y(v)$ and $y(v) = z(v)$, so $x(v) \neq z(v)$.

Case 4 $x(v) \neq y(v)$ and $y(v) \neq z(v)$, so $x(v) \neq z(v)$ or $x(v) = z(v)$.

We verify the triangle inequality now for each of these four cases.

Case 1 Using Δ to indicate the contribution of the vertex v to ρ, $\Delta\rho(x, y) = 0$, $\Delta\rho(y, z) = 0$, and $\Delta\rho(x, z) = 0$.

Case 2 $\Delta\rho(x, y) = 0$, $\Delta\rho(y, z) = 1$, and $\Delta\rho(x, z) = 1$.

Case 3 $\Delta\rho(x, y) = 1$, $\Delta\rho(y, z) = 0$, and $\Delta\rho(x, z) = 1$.

Case 4 $\Delta\rho(x, y) = 1$, $\Delta\rho(y, z) = 1$, and $\Delta\rho(x, z) \in \{0, 1\}$.

Now since this is valid for all v, summing over all v proves this inequality. □

Suppose that $\rho(x, y) = 1$, so that x and y agree everywhere except at a vertex v_0.

Let $\mathcal{N}$ be the set colors at the vertices neighboring v_0, i.e the set of non-allowable colors at v_0.

After one step of the grand coupling, we evaluate the possible changes in the value of $\rho(X_1^x, X_1^y)$.

1. $\rho(X_1^x, X_1^y) = 0$ if v_0 is the selected vertex and the color selected is not in $\mathcal{N}$. Thus,

$$P\{\rho(X_1^x, X_1^y) = 0\} = \Big(\frac{1}{n}\Big)\Big(\frac{q - |\mathcal{N}|}{q}\Big) \geq \frac{q - \Delta}{nq},$$

the inequality following since $\max_{v_0 \in V}(|\mathcal{N}|) \leq \Delta$.

2. Since $\rho(X_1^x, X_1^y)$ increases if a neighbor w of v_0 is selected and the proposed color is either $x(v_0)$ or $y(v_0)$ (otherwise the color update is allowable for both x and y) then for d(v) denoting the degree of vertex v,

$$P\{\rho(X_1^x, X_1^y) = 2\} = \Big(\frac{\mathrm{d}(v_0)}{n}\Big)\Big(\frac{2}{q}\Big) \leq \Big(\frac{\Delta}{n}\Big)\Big(\frac{2}{q}\Big).$$

Thus, we can conlude the following about the expected value of $\rho(X_1^x, X_1^y)$,

$$E(\rho(X_1^x, X_1^y) - 1) \leq \frac{q - \Delta}{nq} \times (-1) + \Big(\frac{\Delta}{n}\Big)\Big(\frac{2}{q}\Big) \times 1 = \frac{3\Delta - q}{nq}.$$

Because of the linearity of expectation,

$$E(\rho(X_1^x, X_1^y) - 1) = E(\rho(X_1^x, X^y)) - 1 \leq \frac{3\Delta - q}{nq}$$

and

$$E(\rho(X_1^x, X^y)) \leq 1 + \frac{3\Delta - q}{nq}.$$

Given $q > 3\Delta$,

$$E(\rho(X_1^x, X_1^y)) < 1.$$

So, if two colorings differ at exactly one vertex, then, on average, after one step of the Metropolis chain, the two colorings are closer to one another. We now show how to deal with more general pairs of colorings.

So, suppose $\rho(x, y) = r$. Then there exists a path of colorings
$x_0 = x, \ldots, x_r = y$ such that $\rho(x_k, x_{k-1}) = 1$. Thus by the triangle inequality

in metric spaces,

$$\begin{aligned}
E(\rho(X_1^x, X_1^y)) &\leq E(\sum_{k=1}^{r} \rho(X_1^k, X_1^{k-1})) \\
&= \sum_{k=1}^{r} E(\rho(X_1^k, X_1^{k-1})) \\
&\leq \sum_{k=1}^{r} 1 + \frac{3\Delta - q}{nq} \\
&= r\Big(1 + \frac{3\Delta - q}{nq}\Big) \\
&= \rho(x, y)\Big(1 + \frac{3\Delta - q}{nq}\Big).
\end{aligned}$$

Given that $X_{t-1}^x = x_{t-1}$ and $X_{t-1}^y = y_{t-1}$, (X_t^x, X_t^y) and $(X_1^{x_{t-1}}, X_1^{y_{t-1}})$ have the same distribution. Thus,

$$\begin{aligned}
E(\rho(X_t^x, X_t^y) \mid X_{t-1}^x = x_{t-1}, X_{t-1}^y = y_{t-1}) &= E(\rho(X_1^{x_{t-1}}, X_1^{y_{t-1}})) \\
&\leq \rho(x_{t-1}, y_{t-1})\Big(1 + \frac{3\Delta - q}{nq}\Big).
\end{aligned}$$

Thus, by the law of total expectation, taking an expectation over (X_{t-1}^x, X_{t-1}^y) yields

$$E(\rho(X_t^x, X_t^y)) \leq E(\rho(X_{t-1}^x, X_{t-1}^y)\Big(1 + \frac{3\Delta - q}{nq}\Big).$$

Thus, iterating over t gives

$$\begin{aligned}
E(\rho(X_t^x, X_t^y)) &\leq E(\rho(X_1^x, X_1^y)\Big(1 + \frac{3\Delta - q}{nq}\Big)^{t-1} \\
&\leq \rho(x, y)\Big(1 + \frac{3\Delta - q}{nq}\Big)^{t}.
\end{aligned}$$

By Markov's inequality,

$$\begin{aligned}
P\{\rho(X_t^x, X_t^y) \geq 1\} &\leq \frac{E(\rho(X_t^x, X_t^y))}{1} \\
&\leq \rho(x, y)\Big(1 + \frac{3\Delta - q}{nq}\Big)^{t}.
\end{aligned}$$

We now need a basic inequality: $1 - x \leq e^{-x}$ for $x \geq 0$. To see this, note that by the mean value theorem, there exists a $c \in (0, x)$ such that for $x \geq 0$

$$\frac{e^{-x} - e^0}{x - 0} = -e^{-c}.$$

Thus,

$$1 - e^{-c}x = e^{-x}$$

And

$$1 - x \leq e^{-x}.$$

Hence since $\max_{x,y}(\rho(x,y)) = n$,

$$\begin{aligned}
P\{X_t^x \neq X_t^y\} &= P\{\rho(X_t^x, X_t^y) \geq 1\} \\
&\leq \rho(x,y)\Big(1 + \frac{3\Delta - q}{nq}\Big)^t \\
&\leq n\Big(1 + \frac{3\Delta - q}{nq}\Big)^t \\
&= n\Big(1 - \frac{q - 3\Delta}{nq}\Big)^t \\
&\leq ne^{(-\frac{q-3\Delta}{nq})^t} \\
&= ne^{-t(\frac{q-3\Delta}{nq})}.
\end{aligned}$$

By Corollary 2.20,

$$\begin{aligned}
d(t) &\leq \max_{x,y \in \tilde{\mathcal{X}}} P_{x,y}\{\tau_{couple>t}\} \\
&= \max_{x,y \in \tilde{\mathcal{X}}} P_{x,y}\{X_t^x \neq X_t^y\} \\
&\leq ne^{-t(\frac{q-3\Delta}{nq})}.
\end{aligned}$$

Thus for

$$t \geq \Big(\frac{q - 3\Delta}{nq}\Big)^{-1}\Big(\log(n) + \log\Big(\frac{1}{\varepsilon}\Big)\Big)$$

we have $d(t) \leq \varepsilon$. □

3 The Transportation Metric and Path Coupling

In order to prove our second bound on q we must define a transportation metric and the path coupling theorem, which will also be used for our third proof.

Definition 3.1 (Diameter)**.** *Let P be a transition matrix on a state space $\mathcal{X}$ that has a metric ρ satisfying $\rho(x,y) \geq 1_{\{x \neq y\}}$. The diameter of $\mathcal{X}$ is defined to be*

$$diam(\mathcal{X}) := \max_{x,y \in \mathcal{X}} \rho(x,y).$$

This following fact will be very helpful in simplifying the analysis of our chains as it requires no information about future states to make a conclusion about the mixing time. Actually, all that is required is a coupling that in one step shrinks the expected ρ-distance between two chains.

Remark 3.2. *If for all x and y in $\mathcal{X}$ there exists a coupling (X_1, Y_1) of $P(x, .)$ with $P(y, .)$ which, for some $\alpha > 0$, satisfies*

$$E_{x,y}(\rho(X_1, Y_1)) \leq e^{-\alpha}\rho(x,y) \ for \ all \ x, y \in \mathcal{X},$$

then, by iteration,

$$E_{x,y}(\rho(X_t, Y_t)) \leq e^{-\alpha t} diam(\mathcal{X}).$$

Thus by definition,

$$\begin{aligned} \|P^t(x,.) - P^t(y,.)\|_{TV} &\leq P_{x,y}\{X_t \neq Y_t\} \\ &= P_{x,y}\{\rho(X_t, Y_t) \geq 1\} \\ &\leq E_{x,y}(\rho(X_t, Y_t)) \\ &\leq e^{-\alpha t} diam(\mathcal{X}) \end{aligned}$$

and $d(t) \leq \varepsilon$ if

$$t \geq \left\lceil \frac{1}{\alpha}\Big[\log(diam(\mathcal{X})) + \log\Big(\frac{1}{\varepsilon}\Big)\Big] \right\rceil$$

so that

$$t_{mix}(\varepsilon) \leq \left\lceil \frac{1}{\alpha}\Big[\log(diam(\mathcal{X})) + \log\Big(\frac{1}{\varepsilon}\Big)\Big] \right\rceil.$$

3.1 The Transportation Metric

The transportation metric will serve as an important measure of distance between two distributions for the remaining proofs.

Definition 3.3 (Transportation Metric). *Let ρ be a metric defined on a state space $\mathcal{X}$. Let X, Y be two probability distributions on $\mathcal{X}$, with mass functions μ and ν respectively. The transportation metric is defined by*

$$\rho_K(\mu, \nu) := \inf\{E(\rho(X, Y)) : (X, Y) \textit{ is a coupling of} \mu \textit{ and } \nu\}.$$

Remark 3.4. *The transportation metric is a generalization of total variation distance: if $\rho(x, y) = 1_{\{x \neq y\}}$ then*

$$\begin{aligned} \|\mu - \nu\|_{TV} &= \inf\{P\{X \neq Y\} \\ &= \inf\{E(\rho(X, Y))\} \\ &= \rho_K(\mu, \nu). \end{aligned}$$

Definition 3.5 (Projection onto Coordinates). *When q is a probability distribution on $\mathcal{X} \times \mathcal{X}$, its projection onto the first coordinate is the probability distribution on $\mathcal{X}$ given by*

$$q(\cdot \times \mathcal{X}) = \sum_{y \in \mathcal{X}} q(\cdot, y).$$

Similarly, its projection onto the second coordinate is the probability distribution on $\mathcal{X}$ given by

$$q(\mathcal{X} \times \cdot) = \sum_{x \in \mathcal{X}} q(x, \cdot).$$

Remark 3.6. *Given a coupling (X, Y) of μ and ν, the distribution of (X, Y) on $\mathcal{X} \times \mathcal{X}$ has projection μ on the first coordinate and ν on the second.*

Remark 3.7. *Given a probability distribution q on $\mathcal{X} \times \mathcal{X}$ with projections μ and ν, the identity function on $(\mathcal{X} \times \mathcal{X}, q)$ is a coupling of μ and ν.*

Remark 3.8. *Given a coupling (X, Y) with distribution q, we have $E(\rho(X, Y)) = \sum_{(x,y)\in\mathcal{X}\times\mathcal{X}} \rho(x, y)q(x, y)$. Thus*

$$\rho_K(\mu, \nu) = \inf\left\{ \sum_{(x,y)\in\mathcal{X}\times\mathcal{X}} \rho(x, y)q(x, y) : q(\cdot, \mathcal{X}) = \mu,\ q(\mathcal{X}, \cdot) = \nu \right\}.$$

This following remark proves the existence of an optimal coupling which will be useful in proving the path coupling theorem 3.15.

Remark 3.9 (Existence of the Optimal Coupling)**.** *The set of probability distribution on $\mathcal{X} \times \mathcal{X}$ can be identified with the $(|\mathcal{X}|^2 - 1)$-dimensional simplex, which is a compact (closed and bounded) subset of $\mathbb{R}^{|\mathcal{X}|^2}$. Now, the set of distributions on $\mathcal{X} \times \mathcal{X}$ which project on the first coordinate to μ and on the second coordinate to ν is a closed subset of the $(|\mathcal{X}|^2 - 1)$-dimensional simplex and is, hence, compact since closed subsets of compact sets are compact. Define the continuous function:*

$$q \mapsto \sum_{(x,y)\in\mathcal{X}\times\mathcal{X}} \rho(x, y)q(x, y).$$

Hence, since the set of probability distributions q is compact and this function is continuous, then it is bounded and attains its minimum at $q_\star$,the ρ-optimal coupling of μ and ν.

Equivalently, there is a pair of random variables $(X_\star, Y_\star)$, such that

$$\rho_K(\mu, \nu) = \sum_{(x,y)\in\mathcal{X}\times\mathcal{X}} \rho(x, y)q_\star(x, y) = E(\rho(X_\star, Y_\star)).$$

Lemma 3.10. *The function ρ_K is a metric on the space of probability distributions on $\mathcal{X}$.*

Proof. Since the metric ρ satisfies $\rho(x, y) \geq 1_{\{x\neq y\}}$,

$$\rho_K(\mu, \nu) = \inf\{E(\rho(X, Y)) : (X, Y) \textit{ is a coupling of} \mu \textit{ and } \nu\} \geq 0.$$

Now, $\rho_K(\mu, \nu) = 0$ if the metric $\rho(x, y) = 0$ and, thus, $\mu = \nu$.

Since the metric ρ satisfies $\rho(x, y) = \rho(y, x)$, $\rho_K(\mu, \nu) = \rho_K(\nu, \mu)$.

Lastly, in order to check the triangle inequality, let μ, ν, and η be probability distributions on $\mathcal{X}$. Let p be a probability distribution on $\mathcal{X} \times \mathcal{X}$ which is a coupling of μ and ν, and let q be a probability distribution on $\mathcal{X} \times \mathcal{X}$ which is a coupling of ν and η. Define the probability distribution r on $\mathcal{X} \times \mathcal{X} \times \mathcal{X}$ by

$$r(x, y, z) := \frac{p(x, y)q(y, z)}{\nu(y)}.$$

Note that

$$r(a \times b \times \mathcal{X}) = \sum_{z \in \mathcal{X}} \frac{p(a,b)q(b,z)}{\nu(b)} = \frac{p(a,b)}{\nu(b)} \sum_{z \in \mathcal{X}} q(b,z) = \frac{p(a,b)}{\nu(b)} \nu(b) = p(a,b).$$

Similarly,

$$r(\mathcal{X} \times b \times c) = \sum_{x \in \mathcal{X}} \frac{p(x,b)q(b,c)}{\nu(b)} = \frac{q(b,c)}{\nu(b)} \sum_{x \in \mathcal{X}} p(x,b) = \frac{q(b,c)}{\nu(b)} \nu(b) = q(b,c).$$

And similarly $r(a \times \mathcal{X} \times c)$ is a coupling of μ and η.

Now, assume p is a ρ-optimal coupling of μ and ν and q is a ρ-optimal coupling of ν and η.

Let (X, Y, Z) be a random vector with probability distribution r. Then since ρ is a metric,

$$\rho(X, Z) \leq \rho(X, Y) + \rho(Y, Z).$$

Since (X, Y) is an optimal coupling of μ and ν, (Y, Z) of ν and η, and by linearity of expectation,

$$E(\rho(X, Z)) \leq E(\rho(X, Y)) + E(\rho(Y, Z)) = \rho_K(\mu, \nu) + \rho_K(\nu, \eta).$$

Lastly, since (X, Z) is a coupling of ν and η,

$$\rho_K(\nu, \eta) \leq E(\rho(X, Z)) \leq \rho_K(\mu, \nu) + \rho_K(\nu, \eta).$$

□

Remark 3.11. *The probability distribution r defined above can be thought as three steps of a time-inhomogeneous Markov Chain. The first state X is generated according to μ. Given $X = x$, the second state Y is generated according to $\frac{p(x,\cdot)}{\mu(x)}$. Given $Y = y$, the third state Z is generated according to $\frac{q(y,\cdot)}{\nu(y)}$. Thus,*

$$P\{X = x, Y = y, Z = z\} = \mu(x) \frac{p(x,y)}{\mu(x)} \frac{q(y,z}{\nu(y)} = \frac{p(x,y)q(y,z)}{\nu(y)} = r(x,y,z).$$

3.2 Path Coupling

In this part, we define path length and the path metric in order to prove the path coupling theorem.

Definition 3.12 (Path Length)**.** *Suppose that the state space $\mathcal{X}$ of a Markov Chain (X_t) is the vertex set of a connected graph $G = (\mathcal{X}, E_0)$ and ℓ is a length function defined on E_0 that assigns a length $\ell(x, y) \geq 1$ to each edge $\{x, y\} \in E_0$. If $x_0, x_1, \ldots, x_r$ is a path in G,*

$$\ell(x_0, x_r) = \sum_{i=1}^{r} \ell(x_{i-1}, x_i).$$

Definition 3.13 (Path Metric)**.** *The path metric on $\mathcal{X}$ is defined by*

$$\rho(x,y) = \min\{\ell(x_0,x_r) : x_0 = x, x_1, \ldots, x_r = y \textit{ is a path}\}.$$

Remark 3.14. *Since $\ell(x,y) \geq 1$, it follows that ρ satisfies $\rho(x,y) \geq 1_{\{x \neq y\}}$. Thus for any (X,Y),*

$$\begin{aligned} \|\mu - \nu\|_{TV} &= \inf_{(X,Y)} \{P\{X \neq Y\} \\ &= \inf_{(X,Y)} E(1_{\{X \neq Y\}}) \\ &\leq \inf_{(X,Y)} \{E(\rho(X,Y))\} \\ &= \rho_K(\mu,\nu). \end{aligned}$$

Here, we introduce the path coupling theorem which will be used to prove rapid mixing for the Glauber and Flip Dynamics. This theorem was first proved by Bubley and Dyer in 1997.

Theorem 3.15. *[Path Coupling Theorem] Suppose the state space X of a Markov Chain is the vertex set of a graph with length l. Let ρ be the corresponding path metric. Suppose that for each edge $\{x,y\}$ there exists a coupling (X_1,Y_1) of the distributions $P(x,\cdot)$ and $P(y,.)$ such that*

$$E_{x,y}(E(\rho(X_1,Y_1))) \leq \rho(x,y)e^{-\alpha}.$$

Then for any two probability measures μ and ν on $\mathcal{X}$,

$$\rho_K(\mu P, \nu P) \leq e^{-\alpha}\rho_K(\mu,\nu).$$

Proof. Let $x, y \in \mathcal{X}$ and let $(x_0, x_1, \ldots, x_r)$ be the path used for the path metric. Then, by the triangle inequality for ρ_k,

$$\rho_K(P(x,.),P(y,.)) \leq \sum_{k=1}^{r} \rho_K(P(x_{k-1},.),P(x_k,.)).$$

Since ρ_K is a minimum over all couplings and by assumption, for any edge $\{a,b\}$,

$$\rho_K(P(a,.),P(b,.)) \leq E_{x,y}(\rho(X_1,Y_1)) \leq \rho(a,b)e^{-\alpha} \leq \ell(a,b)e^{-\alpha}.$$

Thus by choice of $(x_0, x_1, \ldots, x_r)$,

$$\rho_K(P(x,.),P(y,.)) \leq \sum_{k=1}^{r} \ell(x_{k-1},x_k)e^{-\alpha} = e^{-\alpha}\sum_{k=1}^{r} \ell(x_{k-1},x_k) = e^{-\alpha}\rho(x,y).$$

Let η be a ρ-optimal coupling of μ and ν so that

$$\rho_K(\mu,\nu) = \sum_{x,y\in\mathcal{X}} \rho(x,y)\eta(x,y).$$

Thus, for all x, y there exists a coupling $\theta_{x,y}$ of $P(x,.)$ and $P(y,.)$ such that

$$\rho_K(P(x,.), P(y,.)) = \sum_{u,w\in\mathcal{X}} \rho(u,w)\theta_{x,y}(u,w) \leq e^{-\alpha}\rho(x,y).$$

Now, let $\theta := \sum_{x,y\in\mathcal{X}} \eta(x,y)\theta_{x,y}$ on $\mathcal{X}\times\mathcal{X}$ be a coupling of μP and νP. Then,

$$\begin{aligned}
\rho_K(\mu P, \nu P) &\leq \sum_{u,w\in\mathcal{X}} \rho(u,w)\theta(u,w) \\
&= \sum_{u,w\in\mathcal{X}} \rho(u,w) \sum_{x,y\in\mathcal{X}} \eta(x,y)\theta_{x,y}(u,w) \\
&= \sum_{x,y\in\mathcal{X}} \eta(x,y) \sum_{u,w\in\mathcal{X}} \rho(u,w)\theta_{x,y}(u,w) \\
&\leq \sum_{x,y\in\mathcal{X}} \eta(x,y)e^{-\alpha}\rho(x,y) \\
&= e^{-\alpha} \sum_{x,y\in\mathcal{X}} \eta(x,y)\rho(x,y) \\
&= e^{-\alpha}\rho_K(\mu,\nu).
\end{aligned}$$

□

Corollary 3.16. *If* $\rho_K(\mu P, \nu P) \leq e^{-\alpha}\rho_K(\mu,\nu)$ *then*

$$d(t) \leq e^{-\alpha t} diam(\mathcal{X}),$$

and thus

$$t_{mix}(\varepsilon) \leq \left\lceil \frac{-\log(\varepsilon) + \log(diam(\mathcal{X}))}{\alpha} \right\rceil.$$

Proof. By iteration,

$$\rho_K(\mu P^t, \nu P^t) \leq e^{-\alpha t}\rho_K(\mu,\nu) \leq e^{-\alpha t} \max_{x,y} \rho(x,y) = e^{-\alpha t} diam(\mathcal{X}).$$

Thus,

$$\|\mu - \nu\|_{TV} \leq e^{-\alpha t}\|\mu - \nu\|_{TV} \leq e^{-\alpha t}\rho_K(\mu,\nu) \leq e^{-\alpha t} diam(\mathcal{X}).$$

Setting $\mu = \delta_x$ and $\nu = \pi$ where δ_x is the probability distribution which puts unit mass on x and π is the unique stationary distribution,

$$\begin{aligned}
d(t) &= \max_{x\in\mathcal{X}} \|P^t(x,.) - \pi\|_{TV} \\
&= \max_{x\in\mathcal{X}} \|\delta_x P^t(.) - \pi\|_{TV} \\
&\leq \max_{x\in\mathcal{X}} \|\delta_x - \pi\|_{TV} \\
&\leq e^{-\alpha t} diam(\mathcal{X}).
\end{aligned}$$

□

3.3 Second Bound on q for Rapid Mixing on Colorings

Now that the path coupling theorem has been proved we can prove that the Glauber dynamics mixes rapidly for $q > 2\Delta$.

Theorem 3.17. *Consider the Glauber dynamics chain for q-colorings of a graph with n vertices and maximum degree Δ. If $q > 2\Delta$, then the mixing time satisfies*

$$t_{mix}(\varepsilon) \leq \left\lceil \left(\frac{q-\Delta}{q-2\Delta}\right) n(\log(n) - \log(\varepsilon)) \right\rceil.$$

Proof. The metric used is $\rho(x,y) = \sum_{v\in V} 1_{\{x(v)\neq y(v)\}}$ i.e the number of vertices at which x and y differ. We say that two colorings are neighbors if and only if they differ at a single vertex i.e $\rho(x,y) = 1$. Let $A_v(x)$ be the set of allowable colors at v in the coloring x (recall that a color is allowable at v if it does not appear on any of the neighbors of v). Let x and y be two neighboring colorings. We describe the simultaneous evolution of the chain started at x and that started at y, such that each chain viewed alone is a Glauber chain.

1. Pick a vertex w uniformly at random from the vertex set V.
2. Update the color of w in the chain at x and at y. There are two cases to study:
    (a) If v is not a neighbor of w, then we can update the two chains with the same color. So w is updated with a color picked uniformly from $A_w(x) = A_w(y)$.
    (b) If v is a neighbor of w, then without loss of generality assume $|A_w(x)| \leq |A_w(y)|$. Generate a color, U, uniformly at random from $A_w(y)$ and set $y(w) = U$. Then, for updating x we have two cases:
        i. If $U \neq x(v)$ then set $x(w) = U$.
        ii. If $U = x(v)$ then we have two cases:
            A. If $|A_w(x)| = |A_w(y)|$ then set $x(w) = y(v)$.
            B. If $|A_w(x)| < |A_w(y)|$ then update w in x with a color chosen uniformly at random from $A_w(x)$.

See the figure labelled Figure 14.3, which has been reproduced from [1].

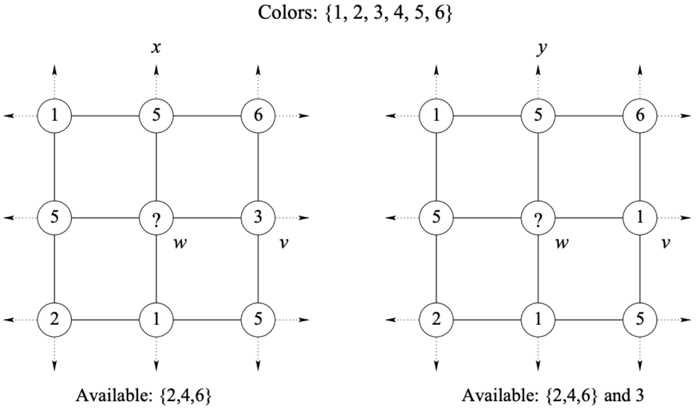


FIGURE 14.3. Jointly updating x and y when they differ only at vertex v and $|A_w(x)| < |A_w(y)|$

Note that the coupling is run is such a way that it exactly runs Glauber dynamics on y.

Claim 3.18. *This process updates x at w to a color chosen uniformly from $A_w(x)$. (So the coloring x transitions to a new coloring in this process, exactly as a step in the Glauber dynamics).*

Proof. 3.18 We provide a case by case analysis:

1. If w is not a neighbor of v then the process picks the color U from $A_w(x)$ with probability $\frac{1}{A_w(x)}$, as desired.
2. If w is a neighbor of v, let S be the set of colors assigned to the neighbors of w, excluding v. Then, given that $|A_w(x)| \leq |A_w(y)|$, we have three cases to evaluate:
   (a) If $y(v) \in S$ and $x(v) \in S$ then the process picks a color U uniformly at random from $A_w(y) = A_w(x)$, as desired.
   (b) If $|A_w(x)| < |A_w(y)|$, then by the last rule of description of the coupling, the update is done with probability $1/|A_w(x)|$, as desired.
   (c) If $|A_w(x)| = |A_w(y)|$ then $A_w(x) = A_w(y) \cap A_w(x) + y(v)$ and $A_w(y) = A_w(y) \cap A_w(x) + x(v)$. Updating $x(w)$ to $y(v)$ only happens when picking $x(v)$ from $A_w(y)$ which happens with probability $1/|A_w(y)| = 1/|A_w(x)|$. For every other possible update at $x(w)$, the probability is that that color was selected by U, which is $1/|A_w(y)| = 1/|A_w(x)|$.

□

Thus, the probability that the two configurations don't update at the same color is $\frac{1}{|A_w(y)|} \leq \frac{1}{q-\Delta}$.

Given two states x and y such that $\rho(x,y) = 1$, we have constructed a coupling (X_1, Y_1) of $P(x,\cdot)$ and $P(y,\cdot)$. There are two cases in which $\rho(X_1, Y_1)$ changes:

1. If $U = x(v)$ and the updates are different in the two configurations then $\rho(X_1, Y_1) = 2$.
2. If $w = v$ then $\rho(X_1, Y_1) = 0$.

Since $2\Delta < q$,

$$\begin{aligned} E_{x,y}(\rho(X_1,Y_1)) &= 1 + E_{x,y}(\rho(X_1,Y_1) - 1) \\ &\leq 1 - 1 \times \frac{1}{n} + 1 \times \frac{d(v)}{n} \times \frac{1}{q-\Delta} \\ &\leq 1 - \frac{1}{n}\left(1 - \frac{\Delta}{q-\Delta}\right) \\ &\leq e^{-\frac{1}{n}\left(1-\frac{\Delta}{q-\Delta}\right)}. \end{aligned}$$

Since $\rho(x,y) = 1$, the hypotheses of the path coupling theorem have been established, with $\alpha = (1/n)(1 - \Delta/(q-\Delta))$, where $\alpha < 1$ since $q > 2\Delta$. Thus from corollary 3.16 we conclude that

$$d(t) \leq ne^{-\frac{1}{n}\left(1-\frac{\Delta}{q-\Delta}\right)},$$

and so

$$t_{mix}(\varepsilon) \leq \left\lceil \left(\frac{q-\Delta}{q-2\Delta}\right) n(\log(n) - \log(\varepsilon)) \right\rceil.$$

□

4 Flip Dynamics

We now give a very brief overview of an alternate chain for sampling from the space of q colorings of a graph, called the flip dynamics. This chain was introduced by Vigoda [2], and allows for a proof of rapid mixing for sampling q-colorings of a maximum degree Δ graph, for $q > \frac{11}{6}\Delta$.

4.1 Markov Chain

Definition 4.1 (Alternating Path)**.** *Let $G = (V, E)$ be a graph. For a coloring x and a color c, a path $v = x_0, x_1, \ldots, x_r = w$ between two vertices v and w is called alternating using colors c and $x(v)$ if for all $i \in \{0, \ldots, r\}$, $\{x_i, x_{i+1}\} \in E$, $x(x_i) \in \{c, x(v)\}$ and $x(x_i) \neq x(x_{i+1})$.*

Definition 4.2 (Flip Dynamics)**.** *Let $\mathcal{X}$ be the state space of the Markov Chain representing the set of all proper q-colorings. Let $S_x(v,c)$ denote the following cluster of vertices:*

$$S_x(v,c) = \left\{ w : \exists v = x_0, x_1, \ldots, x_r = w \textit{ alternating using colors } c \textit{ and } x(v) \right\}.$$

Set $S_x(v, x(v)) = \emptyset$. Note that for each $x_i \in S_x(v,c)$, $S_x(x_i, c) = S_x(v,c)$ if $x(x_i) = x(v)$ or $S_x(x_i, x(v)) = S_x(v,c)$ since $x(x_i) = c$ otherwise. Let $\alpha = |S_x(v,c)|$. For a coloring x, we generate a new coloring y by:

1. *Choosing a vertex v and a color c uniformly at random from V and q respectively.*

2. *Flipping the cluster $S_x(v,c)$ by interchanging colors c and $x(v)$ with probability $\frac{p_\alpha}{\alpha}$.*

Note that we flip $S_x(v,c)$ with probability $\frac{p_\alpha}{\alpha}$ since there are α ways of picking the cluster. Thus, this guarantees that we actually flip the cluster with probability p_α. The properties p_α needs to satisfy in order to make the analysis work turn out to be:

1. $2(i-1)p_i + p_{2i+1} \leq \frac{2}{3}$
2. $(j-1)(p_j - p_{j+1}) \leq \frac{1}{7}$
3. $i(p_i - p_{i+1}) \leq p_1 - p_2 = \frac{29}{42}$
4. $ip_i \leq p_1 = 1$
5. $(i-1)p_i \leq 2p_3 = \frac{1}{3}$
6. *For $c \geq 2$, $(i-c)p_i < \frac{1}{4}$.*

Vigoda showed that as long as the p_α satisfy these conditions, then it is possible to prove rapid mixing for $q > \frac{11}{6}\Delta$. That there is some choice that satisfies the conditions is demonstrated by the following selection:

1. $p_1 = 1$
2. $p_2 = \frac{13}{42}$
3. *For $\alpha > 2$, $p_\alpha = \max\{0, \frac{13}{42} - \frac{1}{7}(1 + \frac{1}{2} + \cdots + \frac{1}{\alpha - 2})\}$.*

Note that in particular the third condition here says that $p_\alpha = 0$ for all $\alpha > 6$. This says that flip dynamics, like Glauber dynamics, is "local" in the sense that each step changes the colors of only a bounded number of vertices (at most 6). Because of this, Vigoda was able to use a "comparison" theorem of Diaconis and Saloff-Coste to show that also Glauber dynamics fir sampling from q-colorings of a maximum degree Δ graph mixes rapidly for $q > \frac{11}{6}\Delta$. Recent work of Chen, Delcourt, Moitra, Perarnau and Postle looks in detail at Vigoda's conditions, and showed that his specific choice of p_α's is optimal (that is, flip dynamics cannot be used to show rapid mixing for $q > C\Delta$ for any $C < \frac{11}{6}$). But they did manage to modify the flip dynamics slightly, to get rapid mixing for $q > \left(\frac{11}{6}\Delta - \varepsilon\right)$ for some very small $\varepsilon > 0$.

Claim 4.3. *The flip dynamics Markov Chain is aperiodic.*

Proof. 4.3 For all $x \in \mathcal{X}$, since flipping $S_x(v, x(v)) = \emptyset$ does not change x, $P(x, x) > 0$. □

Claim 4.4. *The flip dynamics Markov Chain is irreducible.*

Proof. 4.4 In order to prove irreducibly, it is sufficient to assume $p_1 > 0$ and $q > \Delta + 1$ and follow the same proof as for Glauber Dynamics. □

Claim 4.5. *The flip dynamics Markov Chain has uniform stationary distribution π.*

Proof. 4.5 Clearly, if x' is the coloring after flipping $S_x(v, c)$, flipping the cluster $S_{x'}(v, x(v))$ recovers x. So the chain is symmetric and so satisfies the detailed balance equations with the uniform distribution, and so the stationary distribution π is uniform. □

4.2 Third Bound on q for Rapid Mixing on Colorings

Here we provide a very brief summary of the analysis for the last proof of rapid mixing given $q > \frac{11}{6}\Delta$. Our goal is to define a metric and prove that it is expected change is negative to then refer to the path coupling theorem.

Definition 4.6 (Hamming Distance)**.** *Let Φ be a metric on $\mathcal{X} \times \mathcal{X}$, where $\mathcal{X} = C^V$ is the space of colorings, given by $\Phi(x, y)$ being the number of vertices colored differently in x and y.*

For this proof, we focus mainly on explaining how the coupling works and how it affects the Hamming distance. The rest of the algebra in order to use the path coupling theorem and some case by case analysis are omitted.

Theorem 4.7 (Vigoda[2],2000)**.** *The flip dynamics is rapidly mixing, with mixing time $O(nk\log(n))$, provided $q \geq \frac{11}{6}\Delta$.*

Proof. In order to use the path coupling theorem, we will set $\Phi(x, y) = 1$. The path coupling will be defined by flipping a cluster with probability p_α according to the transitions of the Markov chain defined above.

Just as with the previous proof, we begin by choosing a vertex v and a color c uniformly at random. Note that unless $v \in S_x(s, c)$ and/or $S_y(s, c)$, it is possible to flip the same cluster in x and y (the one generated by v and c); this identity coupling leaves $\Phi(x, y)$ unchanged. Thus, for the analysis, we focus on the cases where it is not possible to use the identity coupling.

We partition the set of clusters using colors $x(v), y(v)$, and c and including v in the following way:

$$\Gamma_c = \{w | x(w) = y(w) = c, w \; is \; a \; neighbor \; of \; v\}$$

$$D_c = \{S_x(v,c), S_y(v,c), S_x(w, y(v))_{w \in \Gamma_c}, S_y(w, x(v))_{w \in \Gamma_c}\}.$$

In order to make the path coupling and analysis clearer and more concise, we will ignore the cases in which $c = x(v)$ or $c = y(v)$. So we evaluate the following cases:

1. If $c \neq x(w) = y(w)$, then the flip performed will only flip $\{v\}$ with probability $p_1 = 1$ and color c will be allowable in both x and y. Thus, in this case, the Hamming distance decreases by 1 and $\Phi(x,y) = 0$.

2. If $c = x(w) = y(w)$ then let $\Gamma_c = \{w_1, w_2, \ldots, w_{\delta_c}\}$. When $c \neq x(v)$ we have $S_x(v,c) = \{\cup_i S_y(w_i, x(v))\} \cup \{v\}$. When $c \neq y(v)$ we have $S_y(v,c) = \{\cup_j S_x(w_i, y(v))\} \cup \{v\}$. Notice that this says that a cluster in color x can be partitioned into the singleton v and a collection of clusters from color y. We will refer to the latter as sub-clusters in a moment.

   In this case, let $a_i = a_i(c) = |S_y(w_i, x(v))|, A = A(c) = |S_x(v,c)|$ and similarly $b_j = b_j(c) = |S_x(w_j, y(v))|, B = B(c) = |S_y(v,c)|$.

   We now identify the largest sub-clusters of $S_y(w_i, x(v))$, which we denote by $S_y(w_{i_{\max}}, x(v))$ and similarly for $S_x(w_i, y(v))$, which we denote by $S_x(w_{j_{\max}}, y(v))$.

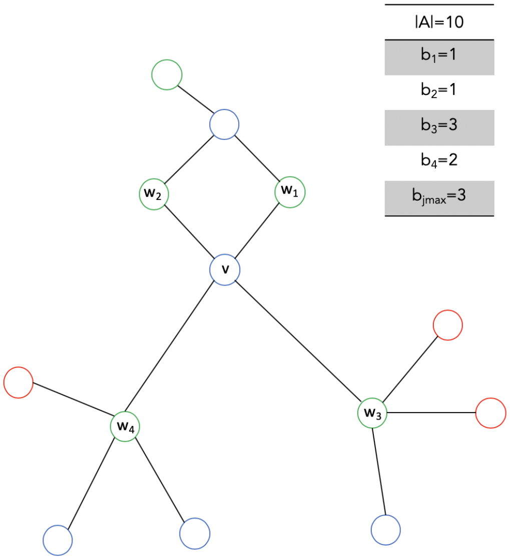
   Note that if sub-clusters occur multiple times we only count them once. This explains in the figure 4.2 why the cluster of w_1 is assigned a size of zero for a_1 as it is the same as the cluster for w_2.

   We can now define the coupling for the moves in D_c for $c \neq x(w) = y(w)$. The main idea is to couple $S_x(v,c)$ and $S_y(v,c)$ with the largest other flips, $S_y(w_{i_{\max}}, x(v))$ and $S_x(w_{j_{\max}}, y(v))$.

   More specifically, a coupling can be thought of as a matrix, with rows indexed by the possible clusters that might be flipped in x, and columns indexed by the possible clusters that might be flipped in y. In both cases, there is a row or column which represents doing nothing to the coloring. Each entry (i,j) of the matrix is the joint probability with which cluster i is flipped in x and cluster j in y. The table provided below is a summary of these probabilities. What needs to be checked to see that it is a valid coupling is that along each row index by a cluster S, the sum of the probabilities is $p_{|S|}$, and the same for the columns.

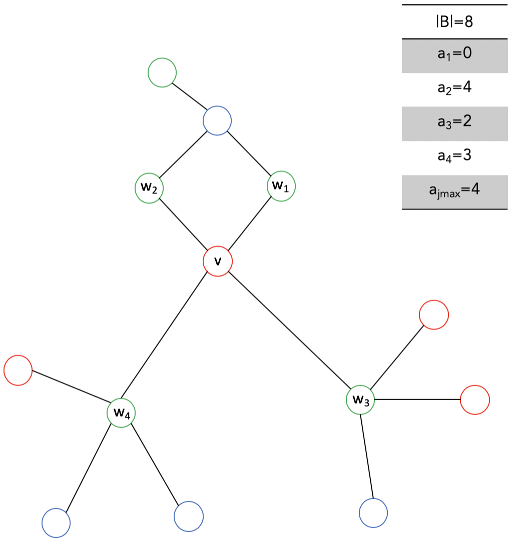
   Thus, the coupling works as follows:

x



|A|=10
b_1=1
b_2=1
b_3=3
b_4=2
b_{jmax}=3

y



|B|=8
a_1=0
a_2=4
a_3=2
a_4=3
a_{jmax}=4

	$S_y(w_{i_{max}}, x(v))$	$S_y(v,c)$	$S_y(w_l, x(v))$		$Total$
$S_x(v,c)$	p_A	0	0	0	p_A
$S_x(w_{j_{\text{max}}}, y(v))$	$p_{a_{i_{\text{max}}}} - p_A$ $p_{b_{j_{\text{max}}}} - p_B$	p_B	p_{a_l} $p_{b_{j_{\text{max}}}} - p_B$	$p_{b_{j_{\text{max}}}} - p_B - p_{a_{i_{\text{max}}}} + p_A$ $p_{b_{j_{\text{max}}}} - p_B - p_{a_l}$	$p_{b_{j_{\text{max}}}}$ $p_{b_{j_{\text{max}}}}$ $p_{b_{j_{\text{max}}}}$ $p_{b_{j_{\text{max}}}}$
$S_x(w_l, y(v))$	$p_{a_{i_{\text{max}}}} - p_A$ p_{b_l}	0	p_{a_l} p_{b_l}	$p_{b_l} - p_{a_{i_{max}}} + p_A$ $p_{b_l} - p_{a_l}$	p_{b_l} p_{b_l} p_{b_l} p_{b_l}
	$p_{a_{i_{\text{max}}}} - p_A - p_{b_{j_{\text{max}}}} + p_B$ $p_{a_{i_{\text{max}}}} - p_A - p_{b_l}$	0	$p_{a_l} - p_{b_{j_{\text{max}}}} + p_B$ $p_{a_l} - p_{b_l}$	0	
$Total$	$p_{a_{i_{\text{max}}}}$ $p_{a_{i_{\text{max}}}}$ $p_{a_{i_{\text{max}}}}$ $p_{a_{i_{\text{max}}}}$	p_B	p_{a_l} p_{a_l} p_{a_l} p_{a_l}	0	

$$p_{a_{i_{\text{max}}}} - p_A = \begin{cases} p_{b_{j_{\text{max}}}} - p_B \\ p_{b_l} \end{cases}$$

$$p_{a_l} = \begin{cases} p_{b_{j_{\text{max}}}} - p_B \\ p_{b_l} \end{cases}$$

$$p_{b_{j_{\text{max}}}} - p_B = \begin{cases} p_{a_{i_{\text{max}}}} - p_A \\ p_{a_l} \end{cases}$$

$$p_{b_l} = \begin{cases} p_{a_{i_{\text{max}}}} - p_A \\ p_{a_l} \end{cases}$$

From here, with some heavy algebra, given p_α defined as above and $q \geq \frac{11}{6}\Delta$, we have that the expected change in Φ is negative. Thus, we can refer to the path coupling theorem in order to prove that it suffices to have $q \geq \frac{11}{6}\Delta$ for the flip dynamics to be rapid mixing. □

References

[1] Levin, D. A., and Peres, Y. *Markov Chains and Mixing Times*, second ed. American Mathematical Soc., Providence, RI, 2017.

[2] Vigoda, E. Improved bounds for sampling colorings. *J. Math. Phys 41* (2000), 1555–1569.