Title
A Mixing Time Bound for the Diaconis Shuffle

Author
Senda, Alto Eugene

Publication Date
2022

Peer reviewed|Thesis/dissertation

A Mixing Time Bound for the Diaconis Shuffle

By

ALTO EUGENE SENDA
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Mathematics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Ben Morris, Chair

Ethan Anderes

Janko Gravner

Committee in Charge

2022

To my parents

Contents

Abstract iv

Acknowledgments v

Chapter 1. Introduction 1

Chapter 2. Mixing Time 2

Chapter 3. Entropy 3

Chapter 4. Reducing Entropy with 3-cycles 15

Chapter 5. The Diaconis Shuffle 28

Chapter 6. The 3-part Process 34

6.1. Stage 1 Analysis 35

6.2. Stage 2 Analysis 53

6.3. Stage 3 Analysis and Total Result 65

Chapter 7. Finding a Mixing Time Bound 67

Appendix A. Pinsker's Inequality 77

Appendix B. Entropy Decompositions 86

Appendix C. Adding Conditioning 97

Appendix D. Removing Conditioning 100

Bibliography 110

A Mixing Time Bound for the Lazy Diaconis Shuffle

Abstract

We prove a theorem that reduces bounding the mixing time of a card shuffle to verifying a condition that involves only triplets of cards, based on earlier work by Morris involving pairs of cards. Then we use it to obtain a bound on the mixing time of the lazy version of a shuffle introduced by Diaconis.

The *Diaconis shuffle* is a variation on the 15-puzzle. There are n^2 tiles arranged in an $n \times n$ torus. Each step, a random tile and direction (North, South, East, West) are selected, and the chosen tile's row or column is cycled in the chosen direction. Diaconis conjectured that the mixing time is $O(n^3 \log n)$. We prove a mixing time bound of $O(n^3 \log^3 n)$.

Acknowledgments

This dissertation would not have been possible without the support of Ben Morris, who agreed to be my advisor and laid out the groundwork for the result.

CHAPTER 1

Introduction

The classic 15-puzzle has 15 tiles placed in a 4×4 grid, leaving one empty space. Tiles can be slid around using the empty space: at each step, a transposition is performed between the empty space and a tile adjacent to it. Starting with a jumbled arrangement, the objective is to move the tiles so that they are in order, usually indicated by labeling the tiles or by having them depict an image.

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	

FIGURE 1.1. A solved 15-puzzle.

In 1988, Diaconis in [**3**] introduced a variant of this puzzle in which the empty space is filled with a tile and the grid is treated as a torus, so that a move is choosing a row or column and then cyclically shifting it by one tile:

$$\begin{array}{c|c|c|c|c|c}
\hline
 & 1 & 4 & 9 & 16 & \\
\hline
\rightarrow & 2 & 3 & 8 & 15 & \rightarrow \\
\hline
 & 5 & 6 & 7 & 14 & \\
\hline
 & 10 & 11 & 12 & 13 & \\
\hline
\end{array}
\quad\Longrightarrow\quad
\begin{array}{|c|c|c|c|}
\hline
1 & 4 & 9 & 16 \\
\hline
15 & 2 & 3 & 8 \\
\hline
5 & 6 & 7 & 14 \\
\hline
10 & 11 & 12 & 13 \\
\hline
\end{array}$$

FIGURE 1.2. A possible move in the Diaconis shuffle in the 4×4 case.

For this variant implemented on an $n \times n$ grid, Diaconis noted that $O(n^3)$ moves are enough to randomize a single square, and conjectured that $O(n^3 \log n)$ moves are enough to randomize the entire puzzle.

We give the first detailed analysis of this shuffle, and show using techniques derived from those featured in [**6**] that it takes $O(n^3 \log^3 n)$ moves to randomize the entire puzzle, which is a bound within a poly log factor of Diaconis's conjecture.

CHAPTER 2

Mixing Time

We begin by reviewing the relevant subject matter, including some parts of [**6**].

Consider a Markov chain on a finite state space $\mathcal{X}$. For probability measures μ and ν on $\mathcal{X}$, the *total variation distance*

$$\|\mu - \nu\|_{\mathrm{TV}} := \max_{\mathcal{A} \subseteq \mathcal{X}} |\mu(\mathcal{A}) - \nu(\mathcal{A})| \tag{2.1}$$

and it also satisfies

$$\|\mu - \nu\|_{\mathrm{TV}} = \frac{1}{2} \sum_{x \in \mathcal{X}} |\mu(x) - \nu(x)| = \frac{1}{2} \|\mu - \nu\|_1 \ ; \tag{2.2}$$

that is, the total variation distance between μ and ν is half the $\mathcal{L}^1$ distance between μ and ν.

Let x and y be elements of $\mathcal{X}$, and let $p^n(x, y)$ denote the n-step transition probabilities. Suppose the stationary distribution of the Markov chain is ω. Then the *mixing time* for this Markov chain is defined as

$$T_{\mathrm{mix}} = \min \left\{ n : \|p^n(x, \cdot) - \omega\|_{\mathrm{TV}} \le \frac{1}{4} \text{ for all } x \in \mathcal{X} \right\}, \tag{2.3}$$

This mixing time is what we will consider to be the number of moves required to randomize the puzzle.

CHAPTER 3

Entropy

Let V be a finite state space. For a probability distribution $a = \{a(x) : x \in V\}$ on V, define the *(relative) entropy* of a as $\mathrm{ENT}(a) = \sum_{x \in V} a(x) \log(|V|\, a(x))$, where we treat $0 \log 0 = 0$.

We clarify that $\mathrm{ENT}(a)$ is the relative entropy of a *with respect to the uniform distribution* $\mathcal{U}$; recall from Section 2.3 of **[2]** that, more generally, given probability distributions a and b on V, the relative entropy of a with respect to b is $D(a||b) = \sum_{x \in V} a(x) \log \frac{a(x)}{b(x)}$, where by convention $0 \log \frac{0}{0} = 0$, $0 \log \frac{0}{t} = 0$ if $t > 0$, and $s \log \frac{s}{0} = \infty$ if $s > 0$.

In the case where b is the uniform distribution on V, we have that $b(x) = \frac{1}{|V|}$ for all $x \in V$, and so we would have that $D(a||b) = \sum_{x \in V} a(x) \log(|V|\, a(x)) = \mathrm{ENT}(a)$, recovering the earlier definition.

As noted in **[2]**, $D(a||b) \geq 0$ for any a and b and $D(a||b) = 0$ if and only if $a = b$. This allows relative entropy to be used as a notion of how "close" two distributions are. However, it is not a metric, as it is not symmetric ($D(a||b) \neq D(b||a)$ generally).

Pinsker's inequality relates entropy to the total variation distance:

$$\sqrt{\frac{1}{2} D(a||b)} \geq \|a - b\|_{\mathrm{TV}} \,. \tag{3.1}$$

[2] proves this as Lemma 11.6.1. We include additional details in Appendix A.

When b is the uniform distribution $\mathcal{U}$, we will have that

$$\sqrt{\frac{1}{2} \mathrm{ENT}(a)} \geq \|a - \mathcal{U}\|_{\mathrm{TV}} \tag{3.2}$$

so an upper bound on the entropy will mean an upper bound on the total variation distance, and hence a bound on the mixing time.

If X is a random variable taking on values in V, we define $\mathrm{ENT}(X)$ as the entropy of the distribution of X.

We will be most interested when $V = S_n$ and π is a random permutation in S_n (so in particular $|V| = n!$) with some distribution, in which case the entropy of π will thus be

$$\mathrm{ENT}(\pi) = \sum_{\nu \in S_n} \mathbb{P}(\pi = \nu) \log(n! \mathbb{P}(\pi = \nu)). \tag{3.3}$$

A key idea in this paper is that of decomposing the relative entropy by using conditional probabilities. We go through the example of conditioning on the last entry of a permutation:

For permutations $\pi \in S_n$, we consider $\pi(j)$ as the position of card j, and $\pi^{-1}(k)$ as the card in position k, which we call the k^{th} *entry* of π. If π is a random permutation in S_n, its last entry $\pi^{-1}(n)$ is also a random variable, taking values in $\{1, \ldots, n\}$, and has relative entropy $\mathrm{ENT}(\pi^{-1}(n))$. We note that if π was uniformly random, then $\pi^{-1}(n)$ would be uniformly random over $\{1, \ldots, n\}$; this gives:

$$\mathrm{ENT}(\pi^{-1}(n)) = \sum_{i=1}^{n} \mathbb{P}(\pi^{-1}(n) = i) \log(n \mathbb{P}(\pi^{-1}(n) = i)). \tag{3.4}$$

Next, suppose $\pi^{-1}(n)$ is known to be a particular value, say k. Then π is a random permutation that we know ends in k; that is to say, we now have the conditional distribution of π given the knowledge that $\pi^{-1}(n) = k$. That conditional distribution is one over all permutations that end in k, and its relative entropy can be calculated as well. This produces the conditional relative entropy of π given $\pi^{-1}(n) = k$. Again, if π were uniform, then knowing that the last entry was k gives a uniform distribution over all permutations ending in k:

$$\mathrm{ENT}(\pi|\pi^{-1}(n) = k) = \sum_{\nu : \nu^{-1}(n) = k} \mathbb{P}(\pi = \nu|\pi^{-1}(n) = k) \log\left((n-1)! \mathbb{P}(\pi = \nu|\pi^{-1}(n) = k)\right). \tag{3.5}$$

This can be done for each possible value of $\pi^{-1}(n)$, and by taking their probabilistic average, we can define the average of the conditional relative entropies:

$$\mathbb{E}(\mathrm{ENT}(\pi|\pi^{-1}(n))) := \sum_{k=1}^{n} \mathbb{P}(\pi^{-1}(n) = k)\, \mathrm{ENT}(\pi|\pi^{-1}(n) = k). \tag{3.6}$$

From this formulation, we can see that if we define $\mathrm{ENT}(\pi|\pi^{-1}(n))$ as the sum of indicators of values of $\pi^{-1}(n)$ multiplied by their corresponding conditional relative entropies, then taking the expectation in the usual sense recovers our definition:

$$\mathrm{ENT}(\pi|\pi^{-1}(n)) := \sum_{k=1}^{n} \mathbf{1}_{\{\pi^{-1}(n)=k\}} \,\mathrm{ENT}(\pi|\pi^{-1}(n) = k). \tag{3.7}$$

Then, as it turns out, the relative entropy of π is in fact the sum of the quantities described above:

$$\mathrm{ENT}(\pi) = \mathbb{E}(\mathrm{ENT}(\pi|\pi^{-1}(n))) + \mathrm{ENT}(\pi^{-1}(n)). \tag{3.8}$$

Let us see how this is true. First, write the probability that π is a particular permutation as the product of the probability that the last entry is correct and the probability that π is that permutation given that the last entry is correct:

$$\mathbb{P}(\pi = \nu) = \mathbb{P}(\pi^{-1}(n) = \nu^{-1}(n)) \cdot \mathbb{P}(\pi = \nu|\pi^{-1}(n) = \nu^{-1}(n)). \tag{3.9}$$

Replacing the instance of the probability term in the logarithm with this product in the expression for $\mathrm{ENT}(\pi)$ from above, we obtain

$$\mathrm{ENT}(\pi) = \sum_{\nu \in S_n} \mathbb{P}(\pi = \nu) \log(n!\mathbb{P}(\pi^{-1}(n) = \nu^{-1}(n)) \cdot \mathbb{P}(\pi = \nu|\pi^{-1}(n) = \nu^{-1}(n))). \tag{3.10}$$

By writing $n! = n \cdot (n-1)!$ and using the fact that a logarithm of a product is equal to the sum of logarithms of the factors, we can rewrite the expression as follows:

$$\begin{aligned}
\mathrm{ENT}(\pi) &= \sum_{\nu \in S_n} \mathbb{P}(\pi = \nu) \log(n\mathbb{P}(\pi^{-1}(n) = \nu^{-1}(n)) \cdot (n-1)!\mathbb{P}(\pi = \nu|\pi^{-1}(n) = \nu^{-1}(n))) \\
&= \sum_{\nu \in S_n} \mathbb{P}(\pi = \nu) \left(\log(n\mathbb{P}(\pi^{-1}(n) = \nu^{-1}(n))) + \log((n-1)!\mathbb{P}(\pi = \nu|\pi^{-1}(n) = \nu^{-1}(n)))\right).
\end{aligned} \tag{3.11}$$

Next, we can reorder the sum over $\nu \in S_n$ by grouping together the ν that have a particular value for the final entry. That is to say, we get

$$\begin{aligned}\mathrm{ENT}(\pi) = \sum_{k=1}^{n} \sum_{\substack{\nu \in S_n: \\ \nu^{-1}(n)=k}} &\mathbb{P}(\pi = \nu)\big(\log(n\mathbb{P}(\pi^{-1}(n) = \nu^{-1}(n))) \\ &+ \log((n-1)!\mathbb{P}(\pi = \nu | \pi^{-1}(n) = \nu^{-1}(n)))\big).\end{aligned} \tag{3.12}$$

In the inner sum, we know that $\nu^{-1}(n) = k$, so we may replace the instances of $\nu^{-1}(n)$ with k:

$$\mathrm{ENT}(\pi) = \sum_{k=1}^{n} \sum_{\substack{\nu \in S_n: \\ \nu^{-1}(n)=k}} \mathbb{P}(\pi = \nu) \left(\log(n\mathbb{P}(\pi^{-1}(n) = k)) + \log((n-1)!\mathbb{P}(\pi = \nu | \pi^{-1}(n) = k))\right). \tag{3.13}$$

Distributing the multiplication and the double sum over the two logarithms gives that

$$\begin{aligned}\mathrm{ENT}(\pi) &= \sum_{k=1}^{n} \sum_{\substack{\nu \in S_n: \\ \nu^{-1}(n)=k}} \big(\mathbb{P}(\pi = \nu) \log(n\mathbb{P}(\pi^{-1}(n) = k)) \\ &\qquad + \mathbb{P}(\pi = \nu) \log((n-1)!\mathbb{P}(\pi = \nu | \pi^{-1}(n) = k))\big) \\ &= \sum_{k=1}^{n} \sum_{\substack{\nu \in S_n: \\ \nu^{-1}(n)=k}} \mathbb{P}(\pi = \nu) \log(n\mathbb{P}(\pi^{-1}(n) = k)) \\ &\quad + \sum_{k=1}^{n} \sum_{\substack{\nu \in S_n: \\ \nu^{-1}(n)=k}} \mathbb{P}(\pi = \nu) \log((n-1)!\mathbb{P}(\pi = \nu | \pi^{-1}(n) = k)).\end{aligned} \tag{3.14}$$

Consider the first double sum: Since $\log(n\mathbb{P}(\pi^{-1}(n) = k))$ is independent of ν, we may pull it outside the inner sum:

$$\sum_{k=1}^{n} \sum_{\substack{\nu \in S_n: \\ \nu^{-1}(n)=k}} \mathbb{P}(\pi = \nu) \log(n\mathbb{P}(\pi^{-1}(n) = k)) = \sum_{k=1}^{n} \log(n\mathbb{P}(\pi^{-1}(n) = k)) \sum_{\substack{\nu \in S_n: \\ \nu^{-1}(n)=k}} \mathbb{P}(\pi = \nu). \tag{3.15}$$

We recognize the inner sum now as the probability that $\pi^{-1}(n) = k$, upon which we recognize the entire sum as $\mathrm{ENT}(\pi^{-1}(n))$:

$$\sum_{k=1}^{n} \sum_{\substack{\nu \in S_n: \\ \nu^{-1}(n)=k}} \mathbb{P}(\pi = \nu) \log(n\mathbb{P}(\pi^{-1}(n) = k)) = \sum_{k=1}^{n} \log(n\mathbb{P}(\pi^{-1}(n) = k))\mathbb{P}(\pi^{-1}(n) = k)$$
$$(3.16) \qquad = \mathrm{ENT}(\pi^{-1}(n)).$$

Now consider the second double sum: We replace $\mathbb{P}(\pi = \nu)$ with the product we saw earlier:

$$(3.17)$$
$$\sum_{k=1}^{n} \sum_{\substack{\nu \in S_n: \\ \nu^{-1}(n)=k}} \mathbb{P}(\pi = \nu) \log((n-1)!\mathbb{P}(\pi = \nu | \pi^{-1}(n) = k))$$
$$= \sum_{k=1}^{n} \sum_{\substack{\nu \in S_n: \\ \nu^{-1}(n)=k}} \mathbb{P}(\pi^{-1}(n) = \nu^{-1}(n))\mathbb{P}(\pi = \nu | \pi^{-1}(n) = \nu^{-1}(n)) \log((n-1)!\mathbb{P}(\pi = \nu | \pi^{-1}(n) = k)).$$

As before, we know in the inner sum that $\nu^{-1}(n) = k$, so we replace the occurrences of $\nu^{-1}(n)$ with k:

$$\sum_{k=1}^{n} \sum_{\substack{\nu \in S_n: \\ \nu^{-1}(n)=k}} \mathbb{P}(\pi = \nu) \log((n-1)!\mathbb{P}(\pi = \nu | \pi^{-1}(n) = k))$$
$$(3.18) \qquad = \sum_{k=1}^{n} \sum_{\substack{\nu \in S_n: \\ \nu^{-1}(n)=k}} \mathbb{P}(\pi^{-1}(n) = k)\mathbb{P}(\pi = \nu | \pi^{-1}(n) = k) \log((n-1)!\mathbb{P}(\pi = \nu | \pi^{-1}(n) = k)).$$

Now since $\mathbb{P}(\pi^{-1}(n) = k)$ is independent of ν, we may pull it out of the inner sum, after which we recognize the inner sum as $\mathrm{ENT}(\pi | \pi^{-1}(n) = k)$, and then the whole sum as $\mathbb{E}(\mathrm{ENT}(\pi | \pi^{-1}(n)))$:

$$
\begin{aligned}
&\sum_{k=1}^{n} \sum_{\substack{\nu\in S_n:\\ \nu^{-1}(n)=k}} \mathbb{P}(\pi=\nu)\log((n-1)!\mathbb{P}(\pi=\nu|\pi^{-1}(n)=k))\\
&\quad=\sum_{k=1}^{n}\mathbb{P}(\pi^{-1}(n)=k)\sum_{\substack{\nu\in S_n:\\ \nu^{-1}(n)=k}} \mathbb{P}(\pi=\nu|\pi^{-1}(n)=k)\log((n-1)!\mathbb{P}(\pi=\nu|\pi^{-1}(n)=k))\\
&\quad=\sum_{k=1}^{n}\mathbb{P}(\pi^{-1}(n)=k)\,\mathrm{ENT}(\pi|\pi^{-1}(n)=k)\\
&\quad=\mathbb{E}(\mathrm{ENT}(\pi|\pi^{-1}(n))).
\end{aligned}
\tag{3.19}
$$

Thus, we have that

$$
\begin{aligned}
\mathrm{ENT}(\pi)&=\sum_{k=1}^{n} \sum_{\substack{\nu\in S_n:\\ \nu^{-1}(n)=k}} \mathbb{P}(\pi=\nu)\log(n\mathbb{P}(\pi^{-1}(n)=k))\\
&\qquad+\sum_{k=1}^{n} \sum_{\substack{\nu\in S_n:\\ \nu^{-1}(n)=k}} \mathbb{P}(\pi=\nu)\log((n-1)!\mathbb{P}(\pi=\nu|\pi^{-1}(n)=k))\\
&=\mathrm{ENT}(\pi^{-1}(n))+\mathbb{E}(\mathrm{ENT}(\pi|\pi^{-1}(n))).
\end{aligned}
\tag{3.20}
$$

as we claimed.

In the above example, we decomposed the relative entropy of a random permutation as the relative entropy of the last entry and the expected value of the relative entropy of the rest of the permutation, given the last entry. There are many other ways to divide the relative entropy: given $1 \le i \le n$, the relative entropy of a random permutation in S_n may be decomposed as the sum of expectations of relative entropies: the relative entropy of the nth entry, the relative entropy of the $(n-1)$th entry given the nth entry, the relative entropy of the $(n-2)$th entry given the $(n-1)$th and nth entry, and so on until the relative entropy of the ith entry given the $(i+1)$ through nth entries, and finally the relative entropy of the rest of the permutation given the ith through nth entries.

For $1 \le i \le n$, let $\mathcal{T}_i^\pi = (\pi^{-1}(i), \ldots, \pi^{-1}(n))$, the tail end of the permutation π starting from entry i, and let $\mathcal{T}_{n+1}$ be the empty string. When the permutation referred to is clear from context, we shall shorten this to $\mathcal{T}_i$. Then we can write the decomposition just described as

$$\text{(3.21)} \qquad \mathrm{ENT}(\pi) = \sum_{k=i}^{n} \mathbb{E}\left(\mathrm{ENT}(\pi^{-1}(k)|\mathcal{T}_{k+1})\right) + \mathbb{E}\left(\mathrm{ENT}(\pi|\mathcal{T}_i)\right),$$

where we note that the $k = n$ term of the sum, $\mathbb{E}\left(\mathrm{ENT}(\pi^{-1}(n)|\mathcal{T}_{n+1})\right)$, is really $\mathrm{ENT}(\pi^{-1}(n)) = \mathbb{E}(\mathrm{ENT}(\pi^{-1}(n)))$.

This is effectively the statement of Proposition 1 in **[6]** but without making reference to sigma-fields, and our initial example is the case where $i = n$.

Proving this decomposition is similar to before:

Let $1 \le i \le n$. Then for any $\nu \in S_n$, we have that

$$\text{(3.22)} \qquad \mathbb{P}(\pi = \nu) = \mathbb{P}(\pi = \nu|\mathcal{T}_i^{\pi} = \mathcal{T}_i^{\nu}) \cdot \prod_{k=i}^{n} \mathbb{P}(\pi^{-1}(k) = \nu^{-1}(k)|\mathcal{T}_{k+1}^{\pi} = \mathcal{T}_{k+1}^{\nu}),$$

by decomposing the probability into a product using conditional probabilities. That is, the probability that π is equal to ν is the probability that they are equal (i.e., all entries agree) given that the entries for i through n agree, multiplied by the probability that the i entries agree given that the entries for $i+1$ through n agree, multiplied by the probability that the $i+1$ entries agree given that the entries for $i+2$ through n agree, and so on until the last factor, which is the probability that the n entries agree.

Then, since $n! = (i-1)! \cdot \prod_{k=i}^{n} k$, we thus have that

(3.23)

$$\begin{aligned}
\mathrm{ENT}(\pi) &= \sum_{\nu \in S_n} \mathbb{P}(\pi = \nu) \log((n!)\mathbb{P}(\pi = \nu)) \\
&= \sum_{\nu \in S_n} \mathbb{P}(\pi = \nu) \log\Big((i-1)!\mathbb{P}(\pi = \nu|\mathcal{T}_i^{\pi} = \mathcal{T}_i^{\nu}) \cdot \prod_{k=i}^{n} (k\mathbb{P}(\pi^{-1}(k) = \nu^{-1}(k)|\mathcal{T}_{k+1}^{\pi} = \mathcal{T}_{k+1}^{\nu}))\Big).
\end{aligned}$$

As before, writing the logarithm of a product as a sum of logarithms yields

(3.24)

$$
\begin{aligned}
&= \sum_{\nu \in S_n} \mathbb{P}(\pi = \nu) \left(\log\left((i-1)!\mathbb{P}(\pi = \nu | \mathcal{T}_i^\pi = \mathcal{T}_i^\nu)\right) + \sum_{k=i}^{n} \log(k\mathbb{P}(\pi^{-1}(k) = \nu^{-1}(k) | \mathcal{T}_{k+1}^\pi = \mathcal{T}_{k+1}^\nu)) \right) \\
&= \sum_{\nu \in S_n} \Bigg(\mathbb{P}(\pi = \nu) \log\left((i-1)!\mathbb{P}(\pi = \nu | \mathcal{T}_i^\pi = \mathcal{T}_i^\nu)\right) \\
&\qquad\qquad + \mathbb{P}(\pi = \nu) \sum_{k=i}^{n} \log(k\mathbb{P}(\pi^{-1}(k) = \nu^{-1}(k) | \mathcal{T}_{k+1}^\pi = \mathcal{T}_{k+1}^\nu)) \Bigg) \\
&= \sum_{\nu \in S_n} \mathbb{P}(\pi = \nu) \log\left((i-1)!\mathbb{P}(\pi = \nu | \mathcal{T}_i^\pi = \mathcal{T}_i^\nu)\right) \\
&\quad + \sum_{\nu \in S_n} \mathbb{P}(\pi = \nu) \sum_{k=i}^{n} \log(k\mathbb{P}(\pi^{-1}(k) = \nu^{-1}(k) | \mathcal{T}_{k+1}^\pi = \mathcal{T}_{k+1}^\nu)).
\end{aligned}
$$

The first sum is similar to one we saw in our previous example. Recall then that we reordered the sum over $\nu \in S_n$ by grouping together the ν that had the same final entry. In the same vein, we reorder the sum over $\nu \in S_n$ here by grouping together the ν that have the same values for entries i through n; that is, the ν that give the same result for $\mathcal{T}_i^\nu$. We do so by letting the variable T range through all of the possible values for $\mathcal{T}_i^\nu$; that is to say, T will be an ordered $(n-i+1)$-tuple whose entries are taken from $\{1, \ldots, n\}$ and are non-repeating. Let $\mathbb{T}_i$ represent the collection of all such possible tails that start from entry i, so then we may write this as $T \in \mathbb{T}_i$; for each T, we consider the $\nu \in S_n$ that have that tail:

$$
\begin{aligned}
&\sum_{\nu \in S_n} \mathbb{P}(\pi = \nu) \log\left((i-1)!\mathbb{P}(\pi = \nu | \mathcal{T}_i^\pi = \mathcal{T}_i^\nu)\right) \\
&= \sum_{T \in \mathbb{T}_i} \sum_{\substack{\nu \in S_n: \\ \mathcal{T}_i^\nu = T}} \mathbb{P}(\pi = \nu) \log\left((i-1)!\mathbb{P}(\pi = \nu | \mathcal{T}_i^\pi = \mathcal{T}_i^\nu)\right).
\end{aligned} \tag{3.25}
$$

Since we know that

$$
\mathbb{P}(\pi = \nu) = \mathbb{P}(\pi = \nu | \mathcal{T}_i^\pi = \mathcal{T}_i^\nu) \cdot \mathbb{P}(\mathcal{T}_i^\pi = \mathcal{T}_i^\nu), \tag{3.26}
$$

substituting this in gets us that

$$\begin{aligned}&\sum_{T\in\mathbb{T}_i}\sum_{\substack{\nu\in S_n:\\ \mathcal{T}_i^\nu=T}}\mathbb{P}(\pi=\nu)\log\left((i-1)!\mathbb{P}(\pi=\nu|\mathcal{T}_i^\pi=\mathcal{T}_i^\nu)\right)\\ &=\sum_{T\in\mathbb{T}_i}\sum_{\substack{\nu\in S_n:\\ \mathcal{T}_i^\nu=T}}\mathbb{P}(\pi=\nu|\mathcal{T}_i^\pi=\mathcal{T}_i^\nu)\mathbb{P}(\mathcal{T}_i^\pi=\mathcal{T}_i^\nu)\log\left((i-1)!\mathbb{P}(\pi=\nu|\mathcal{T}_i^\pi=\mathcal{T}_i^\nu)\right).\end{aligned}\tag{3.27}$$

In the inner sum, we can replace $\mathcal{T}_i^\nu$ with T:

$$=\sum_{T\in\mathbb{T}_i}\sum_{\substack{\nu\in S_n:\\ \mathcal{T}_i^\nu=T}}\mathbb{P}(\pi=\nu|\mathcal{T}_i^\pi=T)\mathbb{P}(\mathcal{T}_i^\pi=T)\log\left((i-1)!\mathbb{P}(\pi=\nu|\mathcal{T}_i^\pi=T)\right)\tag{3.28}$$

and now $P(\mathcal{T}_i^\pi=T)$ is independent of ν, so we may pull it out of the inner sum:

$$=\sum_{T\in\mathbb{T}_i}\mathbb{P}(\mathcal{T}_i^\pi=T)\sum_{\substack{\nu\in S_n:\\ \mathcal{T}_i^\nu=T}}\mathbb{P}(\pi=\nu|\mathcal{T}_i^\pi=T)\log\left((i-1)!\mathbb{P}(\pi=\nu|\mathcal{T}_i^\pi=T)\right),\tag{3.29}$$

but now we recognize what is left of the inner sum as $\mathrm{ENT}(\pi|\mathcal{T}_i^\pi=T)$, and hence the double sum is really the expectation of $\mathrm{ENT}(\pi|\mathcal{T}_i^\pi)$:

$$\begin{aligned}&=\sum_{T\in\mathbb{T}_i}\mathbb{P}(\mathcal{T}_i^\pi=T)\,\mathrm{ENT}(\pi|\mathcal{T}_i^\pi=T)\\ &=\mathbb{E}\left(\mathrm{ENT}(\pi|\mathcal{T}_i^\pi)\right).\end{aligned}\tag{3.30}$$

As for the second sum, the idea is that each entry, conditioned on all the entries that come after it, has its share of relative entropy. We start by switching the order of summation:

$$\begin{aligned}&\sum_{\nu\in S_n}\mathbb{P}(\pi=\nu)\sum_{k=i}^{n}\log(k\mathbb{P}(\pi^{-1}(k)=\nu^{-1}(k)|\mathcal{T}_{k+1}^\pi=\mathcal{T}_{k+1}^\nu))\\ &=\sum_{k=i}^{n}\sum_{\nu\in S_n}\mathbb{P}(\pi=\nu)\log(k\mathbb{P}(\pi^{-1}(k)=\nu^{-1}(k)|\mathcal{T}_{k+1}^\pi=\mathcal{T}_{k+1}^\nu)).\end{aligned}\tag{3.31}$$

Next, similar to before, we shall reorder the sum over $\nu\in S_n$ for each k. We group together the ν that have the same values for entries k through n by first specifying the tail starting from $k+1$ and then enumerating through the possible values of entry k given that tail, which are the numbers from $\{1,\ldots,n\}$ that do not appear in that tail, which we will represent as $j\notin T$. Note that there are k possible values for j. This groups together the ν that all give a particular result for both $\nu^{-1}(k)$ and $\mathcal{T}^{\nu}_{k+1}$:

$$= \sum_{k=i}^{n} \sum_{T \in \mathbb{T}_{k+1}} \sum_{\substack{j=1 \\ j \notin T}}^{n} \sum_{\substack{\nu \in S_n: \\ \nu^{-1}(k)=j, \\ \mathcal{T}^{\nu}_{k+1}=T}} \mathbb{P}(\pi = \nu) \log(k\mathbb{P}(\pi^{-1}(k) = \nu^{-1}(k) | \mathcal{T}^{\pi}_{k+1} = \mathcal{T}^{\nu}_{k+1})). \tag{3.32}$$

We remark that when $k = n$, $\mathcal{T}_{n+1}$ consists of the single unique empty string as discussed earlier. Now we decompose the probability $P(\pi = \nu)$ into three factors:

$$\begin{aligned} &\mathbb{P}(\pi = \nu) \\ &= \mathbb{P}(\pi = \nu | \pi^{-1}(k) = \nu^{-1}(k), \mathcal{T}^{\pi}_{k+1} = \mathcal{T}^{\nu}_{k+1}) \cdot \mathbb{P}(\pi^{-1}(k) = \nu^{-1}(k) | \mathcal{T}^{\pi}_{k+1} = \mathcal{T}^{\nu}_{k+1}) \cdot \mathbb{P}(\mathcal{T}^{\pi}_{k+1} = \mathcal{T}^{\nu}_{k+1}). \end{aligned} \tag{3.33}$$

Substituting this in, we obtain that

$$\begin{aligned} &\sum_{k=i}^{n} \sum_{T \in \mathbb{T}_{k+1}} \sum_{\substack{j=1 \\ j \notin T}}^{n} \sum_{\substack{\nu \in S_n: \\ \nu^{-1}(k)=j, \\ \mathcal{T}^{\nu}_{k+1}=T}} \mathbb{P}(\pi = \nu) \log(k\mathbb{P}(\pi^{-1}(k) = \nu^{-1}(k) | \mathcal{T}^{\pi}_{k+1} = \mathcal{T}^{\nu}_{k+1})) \\ &\quad = \sum_{k=i}^{n} \sum_{T \in \mathbb{T}_{k+1}} \sum_{\substack{j=1 \\ j \notin T}}^{n} \sum_{\substack{\nu \in S_n: \\ \nu^{-1}(k)=j, \\ \mathcal{T}^{\nu}_{k+1}=T}} \big[\mathbb{P}(\pi = \nu | \pi^{-1}(k) = \nu^{-1}(k), \mathcal{T}^{\pi}_{k+1} = \mathcal{T}^{\nu}_{k+1}) \\ &\quad\quad \cdot \mathbb{P}(\pi^{-1}(k) = \nu^{-1}(k) | \mathcal{T}^{\pi}_{k+1} = \mathcal{T}^{\nu}_{k+1}) \mathbb{P}(\mathcal{T}^{\pi}_{k+1} = \mathcal{T}^{\nu}_{k+1}) \log(k\mathbb{P}(\pi^{-1}(k) = \nu^{-1}(k) | \mathcal{T}^{\pi}_{k+1} = \mathcal{T}^{\nu}_{k+1}))\big]. \end{aligned} \tag{3.34}$$

Like before, we can replace $\mathcal{T}^{\nu}_{k+1}$ with T, and $\nu^{-1}(k)$ with j, in the inner sum:

$$\begin{aligned} &= \sum_{k=i}^{n} \sum_{T \in \mathbb{T}_{k+1}} \sum_{\substack{j=1 \\ j \notin T}}^{n} \sum_{\substack{\nu \in S_n: \\ \nu^{-1}(k)=j, \\ \mathcal{T}^{\nu}_{k+1}=T}} \big[\mathbb{P}(\pi = \nu | \pi^{-1}(k) = j, \mathcal{T}^{\pi}_{k+1} = T) \\ &\quad \cdot \mathbb{P}(\pi^{-1}(k) = j | \mathcal{T}^{\pi}_{k+1} = T) \mathbb{P}(\mathcal{T}^{\pi}_{k+1} = T) \log(k\mathbb{P}(\pi^{-1}(k) = j | \mathcal{T}^{\pi}_{k+1} = T))\big]. \end{aligned} \tag{3.35}$$

Now observe that only $\mathbb{P}(\pi = \nu | \pi^{-1}(k) = j, \mathcal{T}^{\pi}_{k+1} = T)$ is dependent on ν, so the rest of the terms may be pulled out of the innermost sum.

$$
\begin{aligned}
&= \sum_{k=i}^{n} \sum_{T\in\mathbb{T}_{k+1}} \sum_{\substack{j=1\\ j\notin T}}^{n} \Bigg[\mathbb{P}(\pi^{-1}(k)=j|\mathcal{T}_{k+1}^{\pi}=T)\mathbb{P}(\mathcal{T}_{k+1}^{\pi}=T)\log(k\mathbb{P}(\pi^{-1}(k)=j|\mathcal{T}_{k+1}^{\pi}=T)) \\
&\qquad \cdot \sum_{\substack{\nu\in S_n:\\ \nu^{-1}(k)=j,\\ \mathcal{T}_{k+1}^{\nu}=T}} \mathbb{P}(\pi=\nu|\pi^{-1}(k)=j,\mathcal{T}_{k+1}^{\pi}=T)\Bigg].
\end{aligned}
\tag{3.36}
$$

The innermost sum is now the sum of the conditional probabilities that $\pi=\nu$ given that π has a specific tail, and the sum is over all ν that have that tail. Consequently, this sum is equal to 1:

$$
= \sum_{k=i}^{n} \sum_{T\in\mathbb{T}_{k+1}} \sum_{\substack{j=1\\ j\notin T}}^{n} \mathbb{P}(\pi^{-1}(k)=j|\mathcal{T}_{k+1}^{\pi}=T)\mathbb{P}(\mathcal{T}_{k+1}^{\pi}=T)\log(k\mathbb{P}(\pi^{-1}(k)=j|\mathcal{T}_{k+1}^{\pi}=T))\cdot 1. \tag{3.37}
$$

Next, since $P(\mathcal{T}_{k+1}^{\pi}=T)$ is independent of j, we may pull it out of the inner sum again, and observe that what is left in the inner sum is the relative entropy of $\pi^{-1}(k)$ given that $\mathcal{T}_{k+1}^{\pi}=T$:

$$
\begin{aligned}
&= \sum_{k=i}^{n} \sum_{T\in\mathbb{T}_{k+1}} \mathbb{P}(\mathcal{T}_{k+1}^{\pi}=T) \sum_{\substack{j=1\\ j\notin T}}^{n} \mathbb{P}(\pi^{-1}(k)=j|\mathcal{T}_{k+1}^{\pi}=T)\log(k\mathbb{P}(\pi^{-1}(k)=j|\mathcal{T}_{k+1}^{\pi}=T)) \\
&= \sum_{k=i}^{n} \sum_{T\in\mathbb{T}_{k+1}} \mathbb{P}(\mathcal{T}_{k+1}^{\pi}=T)\,\mathrm{ENT}(\pi^{-1}(k)|\mathcal{T}_{k+1}^{\pi}=T).
\end{aligned}
\tag{3.38}
$$

We recognize the weighted sum as the expectation of $\mathrm{ENT}(\pi^{-1}(k)|\mathcal{T}_{k+1}^{\pi})$:

$$
= \sum_{k=i}^{n} \mathbb{E}(\mathrm{ENT}(\pi^{-1}(k)|\mathcal{T}_{k+1}^{\pi})). \tag{3.39}
$$

We thus conclude that

$$
\mathrm{ENT}(\pi) = \mathbb{E}\left(\mathrm{ENT}(\pi|\mathcal{T}_i^{\pi})\right) + \sum_{k=i}^{n} \mathbb{E}\left(\mathrm{ENT}(\pi^{-1}(k)|\mathcal{T}_{k+1}^{\pi})\right), \tag{3.40}
$$

which is the decomposition we wanted to show.

So far we have only considered conditioning on the entries of our random permutation, but we may certainly introduce other random variables to condition on to decompose the entropy in additional ways. One particular aspect to condition on that we will consider is the sign of the permutation (when $n \geq 2$): If the sign of the permutation is known to be even, for example, then we know that the permutation will be one out of $\frac{n!}{2}$ possibilities:

$$\mathrm{ENT}(\pi|\,\mathrm{sgn}(\pi)=1) = \sum_{\nu:\mathrm{sgn}\,\nu=1} \mathbb{P}(\pi=\nu|\,\mathrm{sgn}(\pi)=1)\log\left(\frac{n!}{2}\mathbb{P}(\pi=\nu|\,\mathrm{sgn}(\pi)=1)\right). \tag{3.41}$$

A similar expression is obtained for when $\mathrm{sgn}(\pi) = -1$, and in total we would get

$$\mathbb{E}(\mathrm{ENT}(\pi|\,\mathrm{sgn}(\pi))) = \sum_{s=-1,1} \mathrm{ENT}(\pi|\,\mathrm{sgn}(\pi)=s)\mathbb{P}(\mathrm{sgn}(\pi)=s). \tag{3.42}$$

In the same manner as before we can consider the entropy of the sign itself:

$$\begin{aligned}\mathrm{ENT}(\mathrm{sgn}(\pi)) &= \sum_{s=1,-1} \mathbb{P}(\mathrm{sgn}(\pi)=s)\log\left(2\mathbb{P}(\mathrm{sgn}(\pi)=s)\right) \\ &= \mathbb{P}(\mathrm{sgn}(\pi)=1)\log\left(2\mathbb{P}(\mathrm{sgn}(\pi)=1)\right) + \mathbb{P}(\mathrm{sgn}(\pi)=-1)\log\left(2\mathbb{P}(\mathrm{sgn}(\pi)=-1)\right)\end{aligned} \tag{3.43}$$

and decompose the entropy of the permutation like so:

$$\mathrm{ENT}(\pi) = \mathbb{E}\left(\mathrm{ENT}(\pi|\,\mathrm{sgn}(\pi))\right) + \mathrm{ENT}(\mathrm{sgn}(\pi)). \tag{3.44}$$

Furthermore, we can condition on both the sign and the tails of the permutation and obtain the following decomposition (when $n \geq 3$):

$$\mathrm{ENT}(\pi) = \mathbb{E}\left(\mathrm{ENT}(\pi|\,\mathrm{sgn}(\pi), \mathcal{T}_i^\pi)\right) + \sum_{k=i}^{n} \mathbb{E}\left(\mathrm{ENT}(\pi^{-1}(k)|\,\mathrm{sgn}(\pi), \mathcal{T}_{k+1}^\pi)\right) + \mathrm{ENT}(\mathrm{sgn}(\pi)). \tag{3.45}$$

The details of these decompositions may be found in Appendix B.

CHAPTER 4

Reducing Entropy with 3-cycles

In Theorem 9 of [**6**], Morris proves a lower bound on the loss of relative entropy after many steps of a shuffle that involves *collisions*, which are random permutations that are an even mixture of a transposition and the identity. We show here that an analogous result applies when using random permutations that are an even mixture of a 3-cycle and the identity.

First, define a *3-collision* to be a random permutation π in S_n such that there exist distinct $x, y, z \in \{1, \ldots, n\}$ such that π is either the identity with probability $\frac{1}{2}$ or the 3-cycle (x, y, z). with probability $\frac{1}{2}$. If π and μ are permutations in S_n, then we write $\pi\mu$ for the composition $\mu \circ \pi$.

We will consider random permutations π that can be written in the form

$$\pi = \nu c(x_1, y_1, z_1) \cdots c(x_k, y_k, z_k) \tag{4.1}$$

where ν is an arbitrary random permutation, the numbers $x_1, \ldots, x_k, y_1, \ldots, y_k, z_1, \ldots, z_k$ are distinct and $c(x_j, y_j, z_j)$ is a 3-collision of (x_j, y_j, z_j) The values of the x_j, y_j, z_j and the number of 3-collisions (possibly 0) may depend on ν, but conditional on ν the collisions are independent. Such permutations will be called *3-Monte*, in keeping with the terminology used in [**6**]. As noted there, it is true that any random permutation can be considered to be 3-Monte by considering no collisions; we clarify that we want to consider random permutations that can be written usefully with collisions.

Let π be a random permutation that is 3-Monte and let $\pi_1, \pi_2 \ldots$ be independent copies of π. For $t \geq 1$, let $\pi_{(t)} = \pi_1 \cdots \pi_t$, and let $\pi_{(0)}$ be the identity.

We will think of π as representing the order of a deck of cards, with $\pi(i)$ being the location of card i.

For $1 \leq x \leq n$, denote by *card* x the card initially in position x.

For cards x, y, z (with x, y, z distinct), say that x, y, z *collide in that order* at time m if for some i, j, k we have $\pi_{(m)}^{-1}(i) = x$, $\pi_{(m)}^{-1}(j) = y$, $\pi_{(m)}^{-1}(k) = z$ and π_m has a 3-collision of i, j, k.

We recall that we decomposed entropy before by separating out the entropy attributable to the sign:

$$\mathrm{ENT}(\pi) = \mathbb{E}\left(\mathrm{ENT}(\pi|\,\mathrm{sgn}(\pi))\right) + \mathrm{ENT}(\mathrm{sgn}(\pi)). \tag{4.2}$$

This is important because applying 3-collisions to a permutation cannot change its sign, so any entropy change that results must only apply to the entropy not attributable to the sign.

With that in mind, our intent is to generalize Theorem 9 in [**6**] for 3-cycles:

THEOREM 4.0.1. *Let π be a 3-Monte shuffle on n cards. Fix an integer $t > 0$ and suppose that T is a random variable taking values in $\{1, \cdots, t\}$, which is independent of the shuffles $\{\pi_i : i \geq 0\}$. Consider a card x. If x is involved in a 3-collision after time T up to and including time t, then consider the first 3-collision it is involved in after time T; say that x, y, z collide in that order. If that 3-collision is also the first 3-collision after time T that y is involved in and it is the first 3-collision that z is involved in, then we say that* ***that collision matches*** *x (with y and z) and define y to be the* ***front match*** *of x, written as $m_1(x) = y$, and z to be the* ***back match*** *of x, written $m_2(x) = z$. If x is in no such collision, define $m_1(x) = m_2(x) = x$.*

Suppose that for every card i there is a constant $A_i \in [0, 1]$ such that $\mathbb{P}(m_2(i) = j, m_1(i) < i) \geq \frac{A_i}{i}$ for each $j \in \{1, \ldots, i-1\}$. (This also means that with probability at least A_i, $m_1(i) < i$ and $m_2(i) < i$. Note that it cannot be the case that exactly two of i, $m_1(i)$, and $m_2(i)$ are equal; the three are either all the same or all different.) Let μ be an arbitrary random permutation that is independent of $\{\pi_i : i \geq 0\}$. Then

$$\mathbb{E}(\mathrm{ENT}(\mu\pi_{(t)}|\,\mathrm{sgn}(\mu\pi_{(t)}))) - \mathbb{E}(\mathrm{ENT}((\mu|\,\mathrm{sgn}(\mu)))) \leq \frac{-C}{\log n}\sum_{k=1}^{n} A_k E_k, \tag{4.3}$$

where $E_k = \mathbb{E}(\mathrm{ENT}(\mu^{-1}(k)|\mathcal{T}^{\mu}_{k+1}, \mathrm{sgn}(\mu)))$ and C is a universal constant.

PROOF. Let $M = (m_1(i), m_2(i))_{i=1}^{n}$, a list of all matches. For distinct i, j, k with $k < i$, $j < i$, let $c(i, j, k)$ be a 3-collision of i, j, k in that order. Assume that all of the $c(i, j, k)$ are independent of μ, $\pi_{(t)}$, and each other.

Next, define $\pi'_{(t)}$ to be the same as $\pi_{(t)}$ but with all of the 3- collisions that match cards suppressed: If card x has a front match y and a back match z, then there is a specific collision that matches them in $\pi_{(t)}$, since π is 3-Monte. That is, $\pi_{(t)}$ can be written as a product of "matching collisions" and other permutations, and $\pi'_{(t)}$ has these matching collisions removed. For each such matching collision in $\pi_{(t)}$, do nothing instead to obtain $\pi'_{(t)}$.

Let

$$\kappa = \prod_{\substack{i:m_1(i)<i \\ m_2(i)<i}} c(i, m_1(i), m_2(i)). \tag{4.4}$$

Since the matched cards are all distinct from each other, the order of the product does not matter. Furthermore, $\kappa\pi'_{(t)}$ has the same distribution as $\pi_{(t)}$: any particular outcome of $\kappa\pi'_{(t)}$ corresponds to the same outcome of $\pi_{(t)}$ by the choice of matching collision outcomes. That is to say, if card x matches with y and z, and suppose $y < x$, $z < x$, then the outcomes that arise from the matching collision actually performing a 3-cycle in $\pi_{(t)}$ are exactly those outcomes for which the collision $c(x, y, z)$ results in the 3-cycle (x, y, z) in $\kappa\pi'_{(t)}$.

Since they have the same distribution, it follows that

$$\mathbb{E}(\mathrm{ENT}(\mu\pi_{(t)} | \operatorname{sgn}(\mu\pi_{(t)}))) = \mathbb{E}(\mathrm{ENT}(\mu\kappa\pi'_{(t)} | \operatorname{sgn}(\mu\kappa\pi'_{(t)}))). \tag{4.5}$$

By the convexity of $x \log x$, we have that

$$\mathbb{E}(\mathrm{ENT}(\mu\kappa\pi'_{(t)} | \operatorname{sgn}(\mu\kappa\pi'_{(t)}))) \leq \mathbb{E}(\mathrm{ENT}(\mu\kappa\pi'_{(t)} | M, \pi'_{(t)}, \operatorname{sgn}(\mu\kappa\pi'_{(t)}))), \tag{4.6}$$

that is to say, additional conditioning causes an increase in the expected relative entropy. (See Appendix C for a detailed calculation.)

Observe that if $\pi'_{(t)}$ is known, then knowing $\operatorname{sgn}(\mu\kappa\pi'_{(t)})$ is equivalent to knowing $\operatorname{sgn}(\mu\kappa)$; i.e., the sigma-field generated by $\pi'_{(t)}$ and $\operatorname{sgn}(\mu\kappa\pi'_{(t)})$ is the same as the one generated by $\pi'_{(t)}$ and $\operatorname{sgn}(\mu\kappa\pi'_{(t)})$. Consequently, we have that

$$\mathbb{E}(\mathrm{ENT}(\mu\kappa\pi'_{(t)} | M, \pi'_{(t)}, \operatorname{sgn}(\mu\kappa\pi'_{(t)}))) = \mathbb{E}(\mathrm{ENT}(\mu\kappa\pi'_{(t)} | M, \pi'_{(t)}, \operatorname{sgn}(\mu\kappa))). \tag{4.7}$$

Also, if $\pi'_{(t)}$ is known, then the conditional relative entropy of $\mu\kappa\pi'_{(t)}$ is the same as that of $\mu\kappa$, since multiplying by a known permutation will just reindex the sum:

$$\mathbb{E}(\mathrm{ENT}(\mu\kappa\pi'_{(t)}|M, \pi'_{(t)}, \mathrm{sgn}(\mu\kappa))) = \mathbb{E}(\mathrm{ENT}(\mu\kappa|M, \pi'_{(t)}, \mathrm{sgn}(\mu\kappa))). \tag{4.8}$$

Now, since μ, κ, and $\pi'_{(t)}$ are conditionally independent given M, it follows that since we already are conditioning on M, conditioning on $\pi'_{(t)}$ does nothing to the probabilities in the expression for the conditional relative entropy, so we may drop it:

$$\mathbb{E}(\mathrm{ENT}(\mu\kappa|M, \pi'_{(t)}, \mathrm{sgn}(\mu\kappa))) = \mathbb{E}(\mathrm{ENT}(\mu\kappa|M, \mathrm{sgn}(\mu\kappa))). \tag{4.9}$$

Thus, we shall consider $\mathbb{E}(\mathrm{ENT}(\mu\kappa|M, \mathrm{sgn}(\mu\kappa)))$ instead of $\mathbb{E}(\mathrm{ENT}(\mu\kappa|M, \pi'_{(t)}, \mathrm{sgn}(\mu\kappa)))$. (A calculation to show (4.9) is available in Appendix D.)

Next, for $1 \le k \le n$, define

$$\nu_k = \prod_{\substack{i: m_1(i) < i \le k \\ m_2(i) < i}} c(i, m_1(i), m_2(i)). \tag{4.10}$$

Note that $\mathbb{E}(\mathrm{ENT}(\mu\kappa|M, \mathrm{sgn}(\mu\kappa))) = \mathbb{E}(\mathrm{ENT}(\mu\nu_n|M, \mathrm{sgn}(\mu\nu_n)))$ and that $\nu_0 = \mathrm{id}$. Since μ is independent of $\mathcal{M}$, we have $\mathbb{E}(\mathrm{ENT}(\mu|M, \mathrm{sgn}(\mu))) = \mathbb{E}(\mathrm{ENT}(\mu|\,\mathrm{sgn}(\mu)))$, and hence

$$\begin{aligned} &\mathbb{E}(\mathrm{ENT}(\mu\nu_n|M, \mathrm{sgn}(\mu\nu_n))) - \mathbb{E}(\mathrm{ENT}(\mu|\,\mathrm{sgn}(\mu)))) \\ &= \sum_{k=1}^{n} \left[\mathbb{E}(\mathrm{ENT}(\mu\nu_k|M, \mathrm{sgn}(\mu\nu_k))) - \mathbb{E}(\mathrm{ENT}(\mu\nu_{k-1}|M, \mathrm{sgn}(\mu\nu_{k-1})))\right]. \end{aligned} \tag{4.11}$$

Then, it is enough to show that for every k we have

$$\mathbb{E}(\mathrm{ENT}(\mu\nu_k|M, \mathrm{sgn}(\mu\nu_k))) - \mathbb{E}(\mathrm{ENT}(\mu\nu_{k-1}|M, \mathrm{sgn}(\mu\nu_{k-1}))) \le \frac{-CA_kE_k}{\log n}. \tag{4.12}$$

Since $\mathrm{sgn}(\nu_k) = \mathrm{sgn}(\nu_{k-1}) = 1$ as they are products of 3-cycles, $\mathrm{sgn}(\mu\nu_k) = \mathrm{sgn}(\mu\nu_{k-1}) = \mathrm{sgn}(\mu)$, so we can write the above as

$$\mathbb{E}(\mathrm{ENT}(\mu\nu_k|M, \mathrm{sgn}(\mu))) - \mathbb{E}(\mathrm{ENT}(\mu\nu_{k-1}|M, \mathrm{sgn}(\mu))) \le \frac{-CA_kE_k}{\log n}, \tag{4.13}$$

and by linearity of expectation this is

$$\mathbb{E}\Big(\mathrm{ENT}(\mu\nu_k|M,\mathrm{sgn}(\mu) - \mathrm{ENT}(\mu\nu_{k-1}|M,\mathrm{sgn}(\mu)))\Big) \leq \frac{-CA_kE_k}{\log n}. \tag{4.14}$$

More directly, we have that

$$\begin{aligned}
&\mathbb{E}(\mathrm{ENT}(\mu\nu_k|M,\mathrm{sgn}(\mu)) - \mathrm{ENT}(\mu\nu_{k-1}|M,\mathrm{sgn}(\mu))) \\
&= \sum_{m,s} (\mathrm{ENT}(\mu\nu_k|M=m,\mathrm{sgn}(\mu)=s) - \mathrm{ENT}(\mu\nu_{k-1}|M=m,\mathrm{sgn}(\mu)=s))\,\mathbb{P}(M=m,\mathrm{sgn}(\mu)=s)
\end{aligned} \tag{4.15}$$

and note that if m is a matching where $m_1(k) \geq k$ or $m_2(k) \geq k$, then $\nu_k = \nu_{k-1}$, in which case the difference of entropies is 0. Otherwise, if both $m_1(k) < k$ and $m_2(k) < k$, then $\nu_k = \nu_{k-1}c(k, m_1(k), m_2(k))$.

In this case, we can proceed like **[6]** does.

Fix distinct $i, j < k$, let $\lambda = \mu\nu_{k-1}$ and let $\widehat{\lambda} = \lambda(k,i,j)$. $\widehat{\lambda}$ and λ are the same, except that

$$\widehat{\lambda}^{-1}(k) = \lambda^{-1}(j) \qquad \widehat{\lambda}^{-1}(i) = \lambda^{-1}(k) \qquad \widehat{\lambda}^{-1}(j) = \lambda^{-1}(i). \tag{4.16}$$

Also, $\mathrm{sgn}(\widehat{\lambda}) = \mathrm{sgn}(\lambda) = \mathrm{sgn}(\mu)$.

ν_{k-1} has $k+1, \ldots, n$ as fixed points, so $\mathcal{T}^{\widehat{\lambda}}_{k+1} = \mathcal{T}^{\lambda}_{k+1} = \mathcal{T}^{\mu}_{k+1}$.

Consider $\mathrm{ENT}(\lambda c(k,i,j)|M, \mathcal{T}^{\mu}_{k+1}, \mathrm{sgn}(\mu))$. For any particular instance of $M, \mathcal{T}^{\mu}_{k+1}, \mathrm{sgn}(\mu)$, say m, T, and s, we have that

$$\begin{aligned}
&\mathrm{ENT}(\lambda c(k,i,j)|M=m, \mathcal{T}^{\mu}_{k+1}=T, \mathrm{sgn}(\mu)=s) \\
&= \sum_{\eta} \Bigg(\mathbb{P}(\lambda c(k,i,j)=\eta|M=m, \mathcal{T}^{\mu}_{k+1}=T, \mathrm{sgn}(\mu)=s) \\
&\qquad \cdot \log\left(\frac{n!}{2}\mathbb{P}(\lambda c(k,i,j)=\eta|M=m, \mathcal{T}^{\mu}_{k+1}=T, \mathrm{sgn}(\mu)=s)\right)\Bigg).
\end{aligned} \tag{4.17}$$

Observe that the conditional probability of $\lambda c(k,i,j)$ being a particular η can be written as the average of the conditional probabilities of λ and $\widehat{\lambda}$ being that η. Designate those probabilities as p_η and q_η respectively:

$$\mathbb{P}(\lambda c(k,i,j) = \eta | M = m, \mathcal{T}^{\mu}_{k+1} = T, \operatorname{sgn}(\mu) = s)$$

$$(4.18) \quad = \frac{1}{2}\left(\underbrace{\mathbb{P}(\lambda = \eta | M = m, \mathcal{T}^{\mu}_{k+1} = T, \operatorname{sgn}(\mu) = s)}_{p_\eta} + \underbrace{\mathbb{P}(\widehat{\lambda} = \eta | M = m, \mathcal{T}^{\mu}_{k+1} = T, \operatorname{sgn}(\mu) = s)}_{q_\eta} \right)$$

$$= \frac{p_\eta + q_\eta}{2}.$$

Consequently, we have that

$$\begin{aligned} \operatorname{ENT}(\lambda c(k,i,j) | M = m, \mathcal{T}^{\mu}_{k+1} = T, \operatorname{sgn}(\mu) = s) &= \sum_\eta \frac{p_\eta + q_\eta}{2} \log\left(\frac{n!}{2} \cdot \frac{p_\eta + q_\eta}{2}\right) \\ \operatorname{ENT}(\lambda | M = m, \mathcal{T}^{\mu}_{k+1} = T, \operatorname{sgn}(\mu) = s) &= \sum_\eta p_\eta \log\left(\frac{n!}{2} \cdot p_\eta\right) \quad (4.19) \\ \operatorname{ENT}(\widehat{\lambda} | M = m, \mathcal{T}^{\mu}_{k+1} = T, \operatorname{sgn}(\mu) = s) &= \sum_\eta q_\eta \log\left(\frac{n!}{2} \cdot q_\eta\right). \end{aligned}$$

But observe that $\operatorname{ENT}(\lambda | M = m, \mathcal{T}^{\mu}_{k+1} = T, \operatorname{sgn}(\mu) = s) = \operatorname{ENT}(\widehat{\lambda} | M = m, \mathcal{T}^{\mu}_{k+1} = T, \operatorname{sgn}(\mu) = s)$ since the 3-cycle is a fixed permutation (the sums are hence rearrangements of each other, as noted by Proposition 6 in [**6**]), so we have that

$$\sum_\eta p_\eta \log(\frac{n!}{2} \cdot p_\eta) = \sum_\eta q_\eta \log(\frac{n!}{2} \cdot q_\eta) \quad (4.20)$$

and in particular

$$\begin{aligned} \sum_\eta p_\eta \log(\frac{n!}{2} \cdot p_\eta) &= \frac{1}{2}\sum_\eta p_\eta \log(\frac{n!}{2} \cdot p_\eta) + \frac{1}{2}\sum_\eta p_\eta \log(\frac{n!}{2} \cdot p_\eta) \\ &= \frac{1}{2}\sum_\eta p_\eta \log(\frac{n!}{2} \cdot p_\eta) + \frac{1}{2}\sum_\eta q_\eta \log(\frac{n!}{2} \cdot q_\eta). \end{aligned} \quad (4.21)$$

Now recall from [**6**] the following:

For $p, q \geq 0$, define

$$d(p,q) := \frac{1}{2} p \log p + \frac{1}{2} q \log q - \frac{p+q}{2} \log\left(\frac{p+q}{2}\right). \quad (4.22)$$

For a fixed $p \geq 0$, $d(p, \cdot)$ is convex, and by symmetry, for a fixed $q \geq 0$, $d(\cdot, q)$ is also convex. As [**6**] states, this follows from calculating the second derivative. In the case of $d(p, \cdot)$ the second derivative is $\frac{1}{2q} - \frac{1}{2p+2q}$ and in the case of $d(\cdot, q)$ the second derivative is $\frac{1}{2p} - \frac{1}{2p+2q}$, both of which are positive (for $p, q > 0$).

Following [**6**], if $p = \{p_i : i \in V\}$ and $q = \{q_i : i \in V\}$ are both probability distributions on V, we define

$$d(p, q) = \sum_{i \in V} d(p_i, q_i). \tag{4.23}$$

Let $\mathcal{L}(\widehat{\lambda}|M = m, \mathcal{T}^{\mu}_{k+1} = T, \operatorname{sgn}(\mu) = s)$ denote the probability distribution of $\widehat{\lambda}$ given that $M = m$, $\mathcal{T}^{\mu}_{k+1} = T$, and $\operatorname{sgn}(\mu) = s$ (i.e., the *law* of $\widehat{\lambda}$ given those conditions), and similarly for other distributions and conditions. In this case, $\mathcal{L}(\widehat{\lambda}|M = m, \mathcal{T}^{\mu}_{k+1} = T, \operatorname{sgn}(\mu) = s)$ is the collection of q_η and $\mathcal{L}(\lambda|M = m, \mathcal{T}^{\mu}_{k+1} = T, \operatorname{sgn}(\mu) = s)$ is the collection of p_η. Then we have that

$$\begin{aligned}
&\operatorname{ENT}(\lambda c(k,i,j)|M = m, \mathcal{T}^{\mu}_{k+1} = T, \operatorname{sgn}(\mu) = s) - \operatorname{ENT}(\lambda|M = m, \mathcal{T}^{\mu}_{k+1} = T, \operatorname{sgn}(\mu) = s)\\
&= \sum_\eta \frac{p_\eta + q_\eta}{2} \log\left(\frac{n!}{2} \cdot \frac{p_\eta + q_\eta}{2}\right) - \sum_\eta p_\eta \log\left(\frac{n!}{2} \cdot p_\eta\right)\\
&= \sum_\eta \frac{p_\eta + q_\eta}{2} \log\left(\frac{n!}{2} \cdot \frac{p_\eta + q_\eta}{2}\right) - \left(\frac{1}{2}\sum_\eta p_\eta \log\left(\frac{n!}{2} \cdot p_\eta\right) + \frac{1}{2}\sum_\eta q_\eta \log\left(\frac{n!}{2} \cdot q_\eta\right)\right)\\
&= -d(\mathcal{L}(\widehat{\lambda}|M = m, \mathcal{T}^{\mu}_{k+1} = T, \operatorname{sgn}(\mu) = s), \mathcal{L}(\lambda|M = m, \mathcal{T}^{\mu}_{k+1} = T, \operatorname{sgn}(\mu) = s)).
\end{aligned} \tag{4.24}$$

The reason we recalled this function d is to use the projection lemma proved in [**6**]:

LEMMA 4.0.2. *Let X and Y be random variables with distributions p and q, respectively. Fix a function g and let P and Q be the distributions of $g(X)$ and $g(Y)$, respectively. Then $d(p, q) \geq d(P, Q)$.*

Applying the lemma with $g(\lambda) = \lambda^{-1}(k)$, we have that

$$
\begin{aligned}
&d(\mathcal{L}(\widehat{\lambda}|M=m,\mathcal{T}^{\mu}_{k+1}=T,\operatorname{sgn}(\mu)=s),\mathcal{L}(\lambda|M=m,\mathcal{T}^{\mu}_{k+1}=T,\operatorname{sgn}(\mu)=s))\\
&\geq d(\mathcal{L}(\widehat{\lambda}^{-1}(k)|M=m,\mathcal{T}^{\mu}_{k+1}=T,\operatorname{sgn}(\mu)=s),\mathcal{L}(\lambda^{-1}(k)|M=m,\mathcal{T}^{\mu}_{k+1}=T,\operatorname{sgn}(\mu)=s))\\
&= d(\mathcal{L}(\lambda^{-1}(j)|M=m,\mathcal{T}^{\mu}_{k+1}=T,\operatorname{sgn}(\mu)=s),\mathcal{L}(\lambda^{-1}(k)|M=m,\mathcal{T}^{\mu}_{k+1}=T,\operatorname{sgn}(\mu)=s)).
\end{aligned}
\tag{4.25}
$$

We reiterate that if m is a matching where $m_1(k)>k$ or $m_2(k)>k$, or $m_1(k)=m_2(k)=k$, then $\nu_k=\nu_{k-1}$, and if both $m_1(k)<k$ and $m_2(k)<k$, then $\nu_k=\nu_{k-1}c(k,m_1(k),m_2(k))$.

In the case we have such a matching where $m_1(k)<k$ and $m_2(k)<k$, by taking $i=m_1(k)$ and $j=m_2(k)$ we may write the following:

(4.26)

$$
\begin{aligned}
&\operatorname{ENT}(\mu\nu_k|M=m,\mathcal{T}^{\mu}_{k+1}=T,\operatorname{sgn}(\mu)=s)-\operatorname{ENT}(\mu\nu_{k-1}|M=m,\mathcal{T}^{\mu}_{k+1}=T,\operatorname{sgn}(\mu)=s)\\
&=\operatorname{ENT}(\lambda c(k,m_1(k),m_2(k))|M=m,\mathcal{T}^{\mu}_{k+1}=T,\operatorname{sgn}(\mu)=s)\\
&\quad-\operatorname{ENT}(\lambda|M=m,\mathcal{T}^{\mu}_{k+1}=T,\operatorname{sgn}(\mu)=s)\\
&=-d(\mathcal{L}(\widehat{\lambda}|M=m,\mathcal{T}^{\mu}_{k+1}=T,\operatorname{sgn}(\mu)=s),\mathcal{L}(\lambda|M=m,\mathcal{T}^{\mu}_{k+1}=T,\operatorname{sgn}(\mu)=s))\\
&\leq -d(\mathcal{L}(\widehat{\lambda}^{-1}(k)|M=m,\mathcal{T}^{\mu}_{k+1}=T,\operatorname{sgn}(\mu)=s),\mathcal{L}(\lambda^{-1}(k)|M=m,\mathcal{T}^{\mu}_{k+1}=T,\operatorname{sgn}(\mu)=s))\\
&=-d(\mathcal{L}(\lambda^{-1}(m_2(k))|M=m,\mathcal{T}^{\mu}_{k+1}=T,\operatorname{sgn}(\mu)=s),\mathcal{L}(\lambda^{-1}(k)|M=m,\mathcal{T}^{\mu}_{k+1}=T,\operatorname{sgn}(\mu)=s)).
\end{aligned}
$$

Since $m_1(k)<k$ and $m_2(k)<k$, we may write this using indicator functions to specify what $m_2(k)$ is:

$$
\begin{aligned}
&=-\sum_{i=1}^{k-1}(\mathbf{1}_{\{m_2(k)=i,m_1(k)<k\}}(m)\\
&\qquad\cdot d(\mathcal{L}(\lambda^{-1}(i)|M=m,\mathcal{T}^{\mu}_{k+1}=T,\operatorname{sgn}(\mu)=s),\mathcal{L}(\lambda^{-1}(k)|M=m,\mathcal{T}^{\mu}_{k+1}=T,\operatorname{sgn}(\mu)=s))).
\end{aligned}
\tag{4.27}
$$

Recall that $\lambda=\mu\nu_{k-1}$. Since ν_{k-1} does not contain $c(k,m_1(k),m_2(k))$ by definition, it has in particular k and $m_2(k)=i$ as fixed points. This means that $\lambda^{-1}(k)=\mu^{-1}(k)$ and $\lambda^{-1}(i)=\mu^{-1}(i)$ in the previous expression:

$$
\begin{aligned}
&=-\sum_{i=1}^{k-1}(\mathbf{1}_{\{m_2(k)=i,m_1(k)<k\}}(m)\\
&\qquad\cdot d(\mathcal{L}(\mu^{-1}(i)|M=m,\mathcal{T}^{\mu}_{k+1}=T,\operatorname{sgn}(\mu)=s),\mathcal{L}(\mu^{-1}(k)|M=m,\mathcal{T}^{\mu}_{k+1}=T,\operatorname{sgn}(\mu)=s))).
\end{aligned}
\tag{4.28}
$$

Then, since μ is independent of M, the distributions of $\mu^{-1}(k)$ and $\mu^{-1}(i)$ remain the same if we drop the conditioning on M:

(4.29)

$$= -\sum_{i=1}^{k-1} \mathbf{1}_{\{m_2(k)=i, m_1(k)<k\}}(m) d(\mathcal{L}(\mu^{-1}(i)|\mathcal{T}^{\mu}_{k+1} = T, \operatorname{sgn}(\mu) = s), \mathcal{L}(\mu^{-1}(k)|\mathcal{T}^{\mu}_{k+1} = T, \operatorname{sgn}(\mu) = s)).$$

Thus, we have that

(4.30)

$$\begin{aligned}&\mathrm{ENT}(\mu\nu_k|M = m, \mathcal{T}^{\mu}_{k+1} = T, \operatorname{sgn}(\mu) = s) - \mathrm{ENT}(\mu\nu_{k-1}|M = m, \mathcal{T}^{\mu}_{k+1} = T, \operatorname{sgn}(\mu) = s)\\ &\leq -\sum_{i=1}^{k-1} \mathbf{1}_{\{m_2(k)=i, m_1(k)<k\}}(m) d(\mathcal{L}(\mu^{-1}(i)|\mathcal{T}^{\mu}_{k+1} = T, \operatorname{sgn}(\mu) = s), \mathcal{L}(\mu^{-1}(k)|\mathcal{T}^{\mu}_{k+1} = T, \operatorname{sgn}(\mu) = s))\end{aligned}$$

if m is a matching where $m_1(k) < k$ and $m_2(k) < k$.

Also observe that if m is matching where $m_1(k) \geq k$ or $m_2(k) \geq k$, line (4.30) is still true since it says $0 \leq 0$.

By taking expectations we have

$$\begin{aligned}&\mathbb{E}\left(\mathrm{ENT}(\mu\nu_k|M, \mathcal{T}^{\mu}_{k+1}, \operatorname{sgn}(\mu)) - \mathrm{ENT}(\mu\nu_{k-1}|M, \mathcal{T}^{\mu}_{k+1}, \operatorname{sgn}(\mu))\right)\\ &< \mathbb{E}\Big(-\sum_{i=1}^{k-1} \mathbf{1}_{\{m_2(k)=i, m_1(k)<k\}}(M) \cdot d(\mathcal{L}(\mu^{-1}(i)|\mathcal{T}^{\mu}_{k+1}, \operatorname{sgn}(\mu)), \mathcal{L}(\mu^{-1}(k)|\mathcal{T}^{\mu}_{k+1}, \operatorname{sgn}(\mu)))\Big)\end{aligned} \tag{4.31}$$

Since μ and M are independent, it follows that $\mathbf{1}_{\{m_2(k)=i, m_1(k)<k\}}(M)$ and $d(\mathcal{L}(\mu^{-1}(i)|\mathcal{T}^{\mu}_{k+1}, \operatorname{sgn}(\mu)), \mathcal{L}(\mu^{-1}(k)|\mathcal{T}^{\mu}_{k+1}, \operatorname{sgn}(\mu)))$ are independent, so we have

$$\begin{aligned}&\mathbb{E}\left(\mathrm{ENT}(\mu\nu_k|M, \mathcal{T}^{\mu}_{k+1}, \operatorname{sgn}(\mu)) - \mathrm{ENT}(\mu\nu_{k-1}|M, \mathcal{T}^{\mu}_{k+1}, \operatorname{sgn}(\mu))\right)\\ &\leq -\sum_{i=1}^{k-1} \mathbb{E}(\mathbf{1}_{\{m_2(k)=i, m_1(k)<k\}}(M))\mathbb{E}(d(\mathcal{L}(\mu^{-1}(i)|\mathcal{T}^{\mu}_{k+1}, \operatorname{sgn}(\mu)), \mathcal{L}(\mu^{-1}(k)|\mathcal{T}^{\mu}_{k+1}, \operatorname{sgn}(\mu))))\\ &= -\sum_{i=1}^{k-1} \mathbb{P}(m_2(k) = i, m_1(k) < k)\mathbb{E}(d(\mathcal{L}(\mu^{-1}(i)|\mathcal{T}^{\mu}_{k+1}, \operatorname{sgn}(\mu)), \mathcal{L}(\mu^{-1}(k)|\mathcal{T}^{\mu}_{k+1}, \operatorname{sgn}(\mu))))\end{aligned} \tag{4.32}$$

By assumption, we have that $\mathbb{P}(m_2(k)=i, m_1(k)<k) \geq \frac{A_k}{k}$ for each i in the sum. Therefore we have

$$\begin{aligned}
&\leq -\sum_{i=1}^{k-1} \frac{A_k}{k} \mathbb{E}(d(\mathcal{L}(\mu^{-1}(i)|\mathcal{T}^{\mu}_{k+1}, \operatorname{sgn}(\mu)), \mathcal{L}(\mu^{-1}(k)|\mathcal{T}^{\mu}_{k+1}, \operatorname{sgn}(\mu)))) \\
&= -A_k \mathbb{E}\left(\sum_{i=1}^{k-1} \frac{1}{k} d(\mathcal{L}(\mu^{-1}(i)|\mathcal{T}^{\mu}_{k+1}, \operatorname{sgn}(\mu)), \mathcal{L}(\mu^{-1}(k)|\mathcal{T}^{\mu}_{k+1}, \operatorname{sgn}(\mu)))\right).
\end{aligned} \tag{4.33}$$

Since $d(q,q)=0$ for any distribution q, we may add the $i=k$ term to the sum in the expectation, as it adds nothing:

$$= -A_k \mathbb{E}\left(\sum_{i=1}^{k} \frac{1}{k} d(\mathcal{L}(\mu^{-1}(i)|\mathcal{T}^{\mu}_{k+1}, \operatorname{sgn}(\mu)), \mathcal{L}(\mu^{-1}(k)|\mathcal{T}^{\mu}_{k+1}, \operatorname{sgn}(\mu)))\right). \tag{4.34}$$

Recall that for a fixed $p \geq 0$, $d(y,p)$ is convex in y, and that inside the expectation we are taking the average of the values of d. We can thus apply Jensen's inequality to this average to obtain

$$\leq -A_k \mathbb{E}\left(d\left(\textstyle\sum_{i=1}^{k} \frac{1}{k} \mathcal{L}(\mu^{-1}(i)|\mathcal{T}^{\mu}_{k+1}, \operatorname{sgn}(\mu)), \mathcal{L}(\mu^{-1}(k)|\mathcal{T}^{\mu}_{k+1}, \operatorname{sgn}(\mu))\right)\right). \tag{4.35}$$

Now observe that $\displaystyle\sum_{i=1}^{k} \frac{1}{k} L(\mu^{-1}(i)|\mathcal{T}^{\mu}_{k+1}, \operatorname{sgn}(\mu))$ is the uniform distribution over the values in $\{1, \cdots, n\}$ that are not in $\mathcal{T}^{\mu}_{k+1}$.

We recall a lemma from [**6**] that addresses what happens when a uniform distribution is involved in d:

Lemma 4.0.3. *If $\mathcal{U}$ is the uniform distribution on V and q is any distribution on V, then*

$$d(q, \mathcal{U}) \geq \frac{c}{\log |V|} \operatorname{ENT}(q) \tag{4.36}$$

for a universal constant $c > 0$.

Using this lemma, we have that our previous expression can be further bounded by

$$\leq -A_k \frac{c}{\log k} \mathbb{E}(\operatorname{ENT}(\mathcal{L}(\mu^{-1}(k)|\mathcal{T}^{\mu}_{k+1}, \operatorname{sgn}(\mu)))). \tag{4.37}$$

The entropy of the law of $\mu^{-1}(k)$ can be written as just the entropy of $\mu^{-1}(k)$, so this is

$$\leq -A_k \frac{c}{\log k} \mathbb{E}(\mathrm{ENT}(\mu^{-1}(k) | \mathcal{T}^{\mu}_{k+1}, \mathrm{sgn}(\mu))) = -A_k \frac{c}{\log k} E_k. \tag{4.38}$$

Finally, to connect this to our desired result, we argue by decomposition:

We started with

$$\begin{aligned} &\mathbb{E}(\mathrm{ENT}(\mu\nu_k | M, \mathrm{sgn}(\mu)) - \mathrm{ENT}(\mu\nu_{k-1} | M, \mathrm{sgn}(\mu))) \\ &= \sum_{m,s} (\mathrm{ENT}(\mu\nu_k | M = m, \mathrm{sgn}(\mu) = s) - \mathrm{ENT}(\mu\nu_{k-1} | M = m, \mathrm{sgn}(\mu) = s))\, \mathbb{P}(M = m, \mathrm{sgn}(\mu) = s). \end{aligned} \tag{4.39}$$

We may decompose $\mathrm{ENT}(\mu\nu_k | M = m, \mathrm{sgn}(\mu) = s)$ according to the tail entries of μ from the $(k+1)$th entry onward in addition to knowing M and $\mathrm{sgn}(\mu)$ already. That is to say, we get a result similar to equation (3.40):

$$\begin{aligned} &\mathrm{ENT}(\mu\nu_k | M = m, \mathrm{sgn}(\mu) = s) \\ &= \mathbb{E}(\mathrm{ENT}(\mu\nu_k | M = m, \mathrm{sgn}(\mu) = s, \mathcal{T}^{\mu}_{k+1})) + \sum_{i=k+1}^{n} \mathbb{E}(\mathrm{ENT}(\mu\nu_k^{-1}(i) | M = m, \mathrm{sgn}(\mu) = s, \mathcal{T}^{\mu}_{k+1})) \end{aligned} \tag{4.40}$$

Applying the same decomposition to $\mathrm{ENT}(\mu\nu_{k-1} | M = m, \mathrm{sgn}(\mu) = s)$ gets us

$$\begin{aligned} &\mathrm{ENT}(\mu\nu_{k-1} | M = m, \mathrm{sgn}(\mu) = s) \\ &= \mathbb{E}(\mathrm{ENT}(\mu\nu_{k-1} | M = m, \mathrm{sgn}(\mu) = s, \mathcal{T}^{\mu}_{k+1})) + \sum_{i=k+1}^{n} \mathbb{E}(\mathrm{ENT}(\mu\nu_{k-1}^{-1}(i) | M = m, \mathrm{sgn}(\mu) = s, \mathcal{T}^{\mu}_{k+1})). \end{aligned} \tag{4.41}$$

Recall that $\mu\nu_k = \mu\nu_{k-1} c(k, m_1(k), m_2(k))$ if $m_1(k) < k$ and $m_2(k) < k$, and $\mu\nu_k = \mu\nu_{k-1}$ otherwise. Either way, $\mu\nu_k^{-1}(i) = \mu\nu_{k-1}^{-1}(i)$ for $i \geq k+1$, so the expectations of the amount of entropy corresponding to those entries are the same. Consequently, we have that

(4.42)

$$\begin{aligned}
&\mathrm{ENT}(\mu\nu_k|M=m,\mathrm{sgn}(\mu)=s)-\mathrm{ENT}(\mu\nu_{k-1}|M=m,\mathrm{sgn}(\mu)=s)\\
&=\mathbb{E}(\mathrm{ENT}(\mu\nu_k|M=m,\mathrm{sgn}(\mu)=s,\mathcal{T}^{\mu}_{k+1}))-\mathbb{E}(\mathrm{ENT}(\mu\nu_{k-1}|M=m,\mathrm{sgn}(\mu)=s,\mathcal{T}^{\mu}_{k+1}))\\
&=\sum_T\big(\mathrm{ENT}(\mu\nu_k|M=m,\mathcal{T}^{\mu}_{k+1}=T,\mathrm{sgn}(\mu)=s)-\mathrm{ENT}(\mu\nu_{k-1}|M=m,\mathcal{T}^{\mu}_{k+1}=T,\mathrm{sgn}(\mu)=s)\big)\\
&\quad\cdot\mathbb{P}(\mathcal{T}^{\mu}_{k+1}=T|M=m,\mathrm{sgn}(\mu)=s).
\end{aligned}$$

Thus, if we put this into what we had before, we get

(4.43)

$$\begin{aligned}
&\mathbb{E}(\mathrm{ENT}(\mu\nu_k|M,\mathrm{sgn}(\mu))-\mathrm{ENT}(\mu\nu_{k-1}|M,\mathrm{sgn}(\mu)))\\
&=\sum_{m,s}(\mathrm{ENT}(\mu\nu_k|M=m,\mathrm{sgn}(\mu)=s)-\mathrm{ENT}(\mu\nu_{k-1}|M=m,\mathrm{sgn}(\mu)=s))\,\mathbb{P}(M=m,\mathrm{sgn}(\mu)=s)\\
&=\sum_{m,s}\Big(\sum_T\big(\mathrm{ENT}(\mu\nu_k|M=m,\mathcal{T}^{\mu}_{k+1}=T,\mathrm{sgn}(\mu)=s)\\
&\qquad-\mathrm{ENT}(\mu\nu_{k-1}|M=m,\mathcal{T}^{\mu}_{k+1}=T,\mathrm{sgn}(\mu)=s)\big)\cdot\mathbb{P}(\mathcal{T}^{\mu}_{k+1}=T|M=m,\mathrm{sgn}(\mu)=s)\Big)\\
&\quad\cdot\mathbb{P}(M=m,\mathrm{sgn}(\mu)=s)\\
&=\sum_{m,T,s}\big(\mathrm{ENT}(\mu\nu_k|M=m,\mathcal{T}^{\mu}_{k+1}=T,\mathrm{sgn}(\mu)=s)-\mathrm{ENT}(\mu\nu_{k-1}|M=m,\mathcal{T}^{\mu}_{k+1}=T,\mathrm{sgn}(\mu)=s)\big)\\
&\quad\cdot\mathbb{P}(\mathcal{T}^{\mu}_{k+1}=T,M=m,\mathrm{sgn}(\mu)=s)\\
&=\mathbb{E}\left(\mathrm{ENT}(\mu\nu_k|M,\mathcal{T}^{\mu}_{k+1},\mathrm{sgn}(\mu))-\mathrm{ENT}(\mu\nu_{k-1}|M,\mathcal{T}^{\mu}_{k+1},\mathrm{sgn}(\mu))\right).
\end{aligned}$$

We showed that

$$\mathbb{E}\left(\mathrm{ENT}(\mu\nu_k|M,\mathcal{T}^{\mu}_{k+1},\mathrm{sgn}(\mu))-\mathrm{ENT}(\mu\nu_{k-1}|M,\mathcal{T}^{\mu}_{k+1},\mathrm{sgn}(\mu))\right)\le-\frac{c}{\log k}A_kE_k \tag{4.44}$$

since $k\le n$, we have

$$-\frac{c}{\log k}A_kE_k\le-\frac{c}{\log n}A_kE_k \tag{4.45}$$

and so we can conclude that

$$
\begin{aligned}
&\mathbb{E}(\mathrm{ENT}(\mu\nu_k|M,\mathrm{sgn}(\mu)) - \mathrm{ENT}(\mu\nu_{k-1}|M,\mathrm{sgn}(\mu))) \\
&= \mathbb{E}\left(\mathrm{ENT}(\mu\nu_k|M,\mathcal{T}^{\mu}_{k+1},\mathrm{sgn}(\mu)) - \mathrm{ENT}(\mu\nu_{k-1}|M,\mathcal{T}^{\mu}_{k+1},\mathrm{sgn}(\mu))\right) \\
&\leq -\frac{c}{\log n} A_k E_k
\end{aligned} \tag{4.46}
$$

which shows equation (4.12). □

CHAPTER 5

The Diaconis Shuffle

Next, consider the $n \times n$ grid of tiles. We label the tiles $1, \ldots, n^2$ in increasingly larger (backwards) L-shapes, so that each top left $\ell \times \ell$ square contains the labels $1, \ldots, \ell^2$ and each new L-shaped layer adds labels counterclockwise.

1	4	9	16
2	3	8	15
5	6	7	14
10	11	12	13

FIGURE 5.1. Labeling the grid in the case $n = 4$.

Recall that the Diaconis shuffle proceeds as follows: at each step, we select a random row or column (1 out of $2n$ possibilities) and then cyclically shift that row or column left/right or up/down (a fair coin flip). Our key observation is that if a row move is followed by a column move, the result is exactly the same as the column move followed by the row move, followed by a 3-cycle of the tiles at the intersection of the row and column selected. Thus, if we consider pairs of moves and we adopt the convention that we apply the row moves first, we can introduce the possibility that the column move actually happened first as a 3-collision. This shows that the Diaconis shuffle is non-trivally 3-Monte (see Figure 5.2).

Additionally, we observe that this shows that we can generate any L-shaped 3-cycle by the Diaconis shuffle: If r is a row move and c is a column move, and Γ is the L-shaped 3-cycle such that

$$rc = cr\Gamma, \tag{5.1}$$

then it follows that

$$r^{-1}c^{-1}rc = \Gamma \tag{5.2}$$

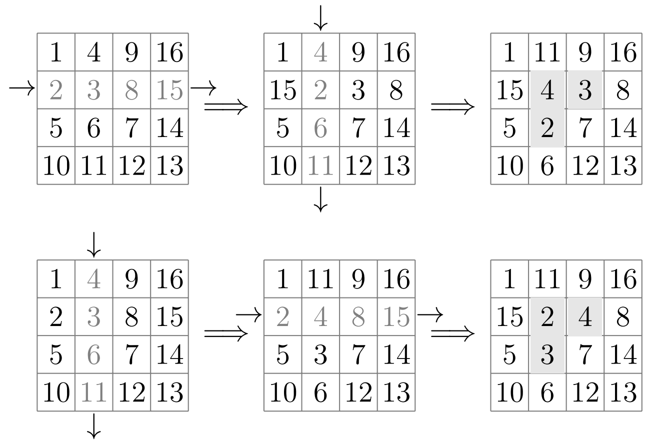

FIGURE 5.2. A possible 3-collision. The first sequence shows the 2nd row shifted right, then the 2nd column shifted down. The second sequence shows the 2nd column shifted down, then the 2nd row shifted right. Note how only the shaded tiles differ in the end result by a 3-cycle.

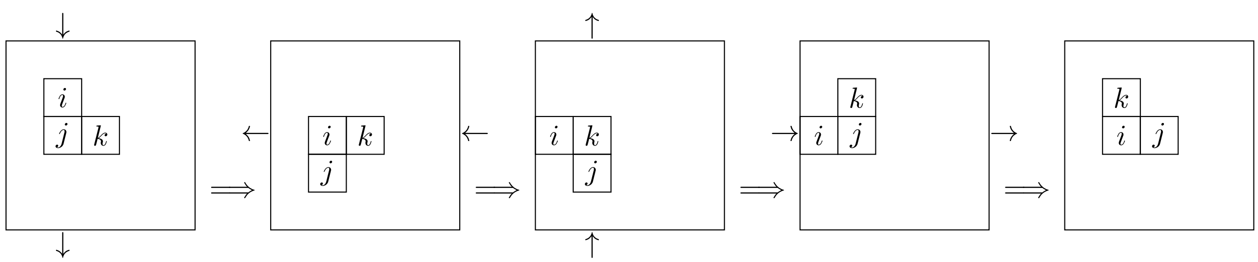

FIGURE 5.3. Generating an L-shaped 3-cycle. Note that the tiles not specifically drawn end up in the same positions.

REMARK 5.0.1. *Since the inverse of a L-shaped 3-cycle is an L-shaped 3-cycle, it follows that both "row-column-row-column" and "column-row-column-row" combinations produce 3-cycles.*

Figure 5.3 gives an example of producing a 3-cycle with a "column-row-column-row" combination. The above remark implies that the corresponding "row-column-row-column" combination will give the inverse 3-cycle. By rotating the grid if necessary, it follows that any arbitrary L-shaped 3-cycle can be generated by the Diaconis shuffle.

This leads to the following:

LEMMA 5.0.2. *Any 3-cycle can be generated from the Diaconis shuffle.*

Proof. First, define the *size* of a 3-cycle to be the smallest integer n such that the 3-cycle can be contained in an $n \times n$ box. For example, the L-shaped 3-cycles generated from the combinations above are size 2.

We induct on the size of the 3-cycle. For the base case, observe that any size 2 3-cycle is necessarily an L-shaped 3-cycle, and therefore can be generated by the Diaconis shuffle as seen above.

Now, let $n \geq 2$, and assume inductively that any 3-cycle that has size from 2 to n can be generated from the Diaconis shuffle. We prove that any size $(n+1)$ 3-cycle can be generated from 3-cycles that we already know can be generated from the Diaconis shuffle. We classify the size $(n+1)$ 3-cycles into types and show sequentially that each type is able to be generated.

Type 1: First, we show that any size $(n+1)$ 3-cycle containable in an $(n+1) \times n$ box or $n \times (n+1)$ box can be generated: Suppose (x, y, z) is such a 3-cycle. Then, without loss of generality, we may assume it is contained in a $(n+1) \times n$ box, and that in that box, x is in the bottom row and y is in the top row.

Type 1a: First suppose z is not in the same row as x or y. Then z must be in a row between x and y.

Let w be the tile directly to the right of z (if this would make w outside the box, take w to be the tile left of z instead). Then observe that

$$(w, z, x)(z, w, y) = (x, y, z), \tag{5.3}$$

while (w, z, x) and (z, w, y) are size n or smaller, so by assumption they can be generated by the Diaconis shuffle, which means that (x, y, z) can be generated by the Diaconis shuffle. See Figure 5.4 for an example of this case.

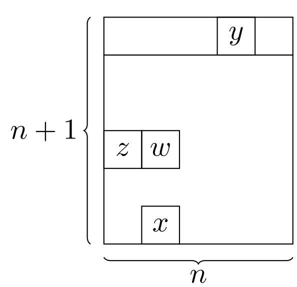

Figure 5.4. If (x, y, z) is a size $(n+1)$ 3-cycle containable in an $(n+1) \times n$ box, with z between x and y, then choose w to be in the same row as z.

Type 1b: Now suppose that z is in the same row as x or y. Without loss of generality, we can assume that z is in the same row as y. (If z is in the same row as x, then have y take the role of x, z take the role of y, and x take the role of z, and flip the box upside down.) Then let w be the tile directly below z. Note that w is not x, since $n \geq 2$. Then observe that

$$(w, y, z)(x, y, w) = (x, y, z). \tag{5.4}$$

(w, y, z) is at most size n, and (x, y, w) is a size $(n + 1)$ cycle of type 1a, since w is not in the same row as x or y. It follows that (x, y, z) can be generated by the Diaconis shuffle. See Figure 5.5 for an example of this case.

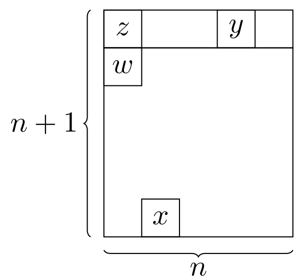

FIGURE 5.5. If (x, y, z) is a size $(n + 1)$ 3-cycle containable in an $(n + 1) \times n$ box, with z in the same row as y, then choose w to be directly below z.

This shows that any size $(n + 1)$ 3-cycle that is containable in an $(n + 1) \times n$ box or $n \times (n + 1)$ box (i.e., a rectangle) can be generated.

Type 2: Now suppose that (x, y, z) is a size $(n + 1)$ 3-cycle that cannot be contained in an $(n + 1) \times n$ box or in an $n \times (n + 1)$ box, and thus requires an $(n + 1) \times (n + 1)$ box to be in. This means that by appropriately choosing x and the box to use (using rotations and reflections if needed), we can set x to be the bottom-left tile. Then (x, y, z) must be of one of the two following forms:

Type 2a: If x and y can be chosen so that x is the bottom left tile and y is the top right tile, so that x and y span the diagonal of the box, then do so. In this case, z could be anywhere else in the box. Since $n \geq 2$, we can choose w to be a tile adjacent to z that is neither x nor y, while still in the box. Then note that

$$(w, z, x)(z, w, y) = (x, y, z), \tag{5.5}$$

and observe that (w, z, x) and (z, w, y) are each either size at most n or a size $(n+1)$ 3-cycle that fits in an $(n+1) \times n$ box (i.e., Type 1), and hence can be generated from the previous arguments. It follows that (x, y, z) can be generated. See Figure 5.6.

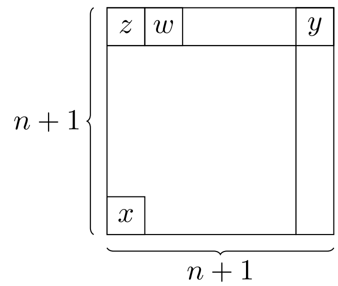

FIGURE 5.6. If x and y span the diagonal and z is elsewhere, then pick w to be adjacent to z.

Type 2b: If the situation described in Type 2a is not possible, then this means that after making x the bottom left tile, neither y nor z is the top right tile. But since (x, y, z) is assumed to require an $(n+1) \times (n+1)$ box, it must be the case that we can assume without loss of generality that y must be in the top row while z is in the rightmost column. We can further assume it is not the case that simultaneously y is the top left tile and z is the bottom right tile, since otherwise by rotating the box this becomes a Type 2a case (since y and z span a diagonal). Therefore, we shall assume that z is in the rightmost column, while neither being the at the top or the bottom. In that case, set w to be the tile left of z. (Note that w is not y since z is not the top right tile.) Then note that

$$(w, y, z)(x, y, w) = (x, y, z) \tag{5.6}$$

and (w, y, z) and (x, y, w) are each either size at most n or a size $(n+1)$ 3-cycle that fits in an $(n+1) \times n$ box (i.e., Type 1), and hence can be generated from the previous arguments. It follows that (x, y, z) can be generated. See Figure 5.7.

It follows that any 3-cycle can be generated by the Diaconis shuffle. □

Since 3-cycles generate the alternating group, it follows that the Diaconis shuffle can generate any even permutation.

Now, consider the following: On an $n \times n$ grid, each move of the Diaconis shuffle is an n-cycle.

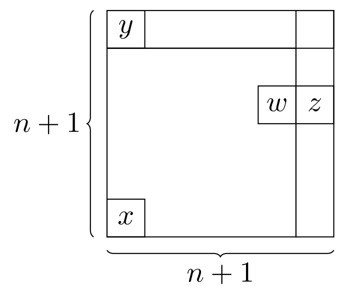

FIGURE 5.7. If neither x and y nor y and z span a diagonal, then pick w to the left of z.

If n is odd, the shuffle leaves the sign unchanged, which means that only even permutations can be attained by the shuffle. Since any even permutation can be generated, it follows that the shuffle mixes in finite time over the set of even permutations. (Note that any permutation can return to itself in both n steps (by iterating an n-cycle n times) and 2 steps (by performing a step and its inverse), so the chain is aperiodic.) For such cases we will keep in mind that when we refer to a uniform distribution we mean uniform over the even permutations.

If n is even, then the shuffle alternates signs each step, which means that the shuffle has no mixing time. However, this also implies that any permutation can be generated by the Diaconis shuffle. To alleviate the problem of the alternating sign, we shall consider a lazy version of the Diaconis shuffle (regardless of the parity of n), where with probability $\frac{1}{2}$ the shuffle does nothing and with probability $\frac{1}{2}$ the shuffle performs a normal step. Then the sign mixes after a single step, after which the fact that any permutation can be generated implies that the shuffle mixes in finite time over the set of all permutations.

Note that in the lazy Diaconis shuffle the chance for it for there to be a 3-collision is only $\frac{1}{4}$ as much as before, since there needs to be two consecutive non-holding moves.

CHAPTER 6

The 3-part Process

Next, we aim to apply Theorem 4.0.1 to the lazy Diaconis shuffle.

To determine the correct t, T, and A_i to use for the theorem, we study the behavior of the tiles under the shuffle.

We claim the following:

LEMMA 6.0.1. *Consider the lazy Diaconis shuffle on an $n \times n$ grid. Then there exist universal constants $\ell_0 > 0$, $C > 0$, $D > 0$ such that if $\ell > \ell_0$ and T' is the least even integer greater than $C\ell^2 n$, then for any triplets of tiles (i, j, k) and (i', j', k') in an $\ell \times \ell$ box within the $n \times n$ grid, the probability that (i, j, k) travels to (i', j', k') (that is, tile i travels to tile i', tile j travels to tile j', and tile k travels to tile k') in T' steps is at least $\frac{D}{\ell^6}$.*

PROOF. Suppose ℓ is sufficiently large in a manner to be described, consider an $\ell \times \ell$ box within the $n \times n$ grid, and let (i, j, k), (i', j', k') be two triplets of tiles within that $\ell \times \ell$ box.

One way that (i, j, k) could move to (i', j', k') is to have i, j, and k move in three stages:

(1) i, j, and k initially travel to three small boxes $\mathcal{B}_i$, $\mathcal{B}_j$, and $\mathcal{B}_k$ that are contained in the $\ell \times \ell$ box, and are spaced away from each other.

(2) i, j, and k move from one position to some specified other position in their respective boxes.

(3) i, j and k then travel from their boxes to their destinations: i', j', and k'.

We claim that Stages 1 and 3 are completed successfully with probability bounded away from 0 and Stage 2 is completed successfully with probability $\Omega(d^{-6})$; note that in order for this construction to make sense, ℓ must be large enough such that the small boxes can contain tiles. Indeed, we prescribe the following: Let $\mathcal{B}_i$, $\mathcal{B}_j$, and $\mathcal{B}_k$ be boxes of side length $c\ell$ positioned as in the left diagram in Figure 6.2, and let $\mathcal{L}_i$, $\mathcal{L}_j$, and $\mathcal{L}_k$ be boxes of side length $c\ell - 2\delta\ell$ centered in $\mathcal{B}_i$, $\mathcal{B}_j$, and $\mathcal{B}_k$, respectively, as demonstrated for $\mathcal{B}_k$ in the right diagram in Figure 6.2, where $c > 0$ and $\delta > 0$ are constants to be determined. More specifically, if the bottom left corner of the $\ell \times \ell$ box is $(0, 0)$, and

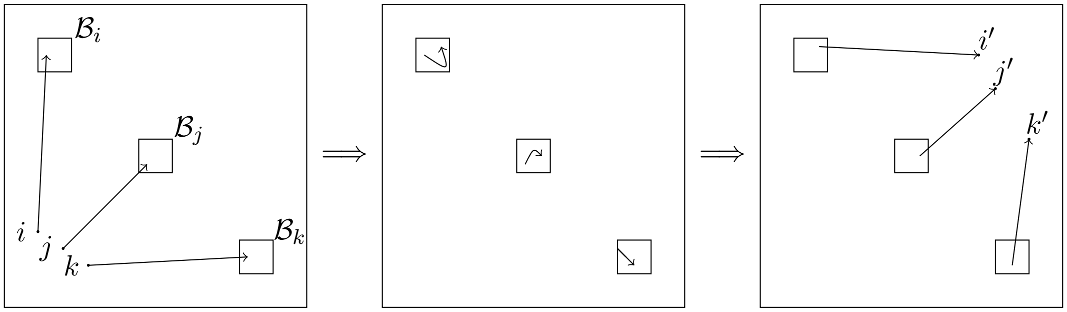

FIGURE 6.1. The three-stage process.

the top right corner is (ℓ, ℓ), then they are centered at $(\frac{\ell}{6}, \frac{5\ell}{6})$, $(\frac{\ell}{2}, \frac{\ell}{2})$, and $(\frac{5\ell}{6}, \frac{\ell}{6})$; these coordinates arise from spacing out $\mathcal{B}_i$, $\mathcal{B}_j$, and $\mathcal{B}_k$ across the $\ell \times \ell$ box. (We will discuss the importance of these even smaller boxes later.)

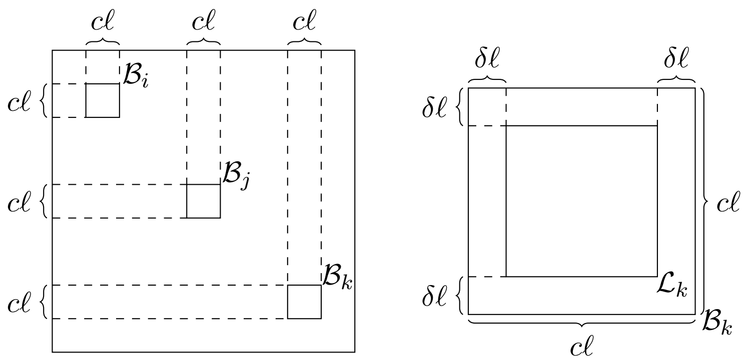

FIGURE 6.2. The boxes involved in the 3-part process. Note that quantities like $c\ell$ may not be integers; we consider a tile to be in a box if it is completely within that box. (Not to scale)

We observe here that we want $c < \frac{1}{3}$ and $\delta < \frac{1}{2}c$, and that if ℓ is sufficiently large, the boxes actually contain tiles.

6.1. Stage 1 Analysis

For the first stage, we wish to show that in $2\ell^2 n$ steps, the three tiles travel to their boxes $\mathcal{B}_i$, $\mathcal{B}_j$, and $\mathcal{B}_k$ with probability bounded away from 0. We start by considering the mechanics of the Diaconis shuffle again:

Consider the movement of a single tile in the lazy Diaconis shuffle. If it moves, it moves one tile in one of the four directions uniformly at random. Thus, when considering a single tile, the Diaconis shuffle induces a random walk on it. This idea can be inverted to produce the shuffle from random walks: For every tile $\mathbb{T}$, assign a random walk sequence $\{m_q^{\mathbb{T}}\}_{q=1}^{\infty}$ to it; the $m_q^{\mathbb{T}}$ are independent and identically distributed uniformly over $(\pm 1, 0), (0, \pm 1)$. Then at each step, move the tiles in the grid in the following way:

(1) Decide whether to move or not with equal probability (laziness).

(2) Arbitrarily, choose n tiles such that no two share the same row or column. For example, one such choice could be a diagonal. Denote these as "master" tiles.

(3) Randomly select one of the n master tiles; call it $\widehat{\mathbb{T}}$.

(4) Suppose $\widehat{\mathbb{T}}$ has already moved $(q-1)$ times. Move $\widehat{\mathbb{T}}$ according to $m_q^{\widehat{\mathbb{T}}}$, and move the other tiles accordingly. Example: If $\widehat{\mathbb{T}}$ has not yet moved (so it has already moved 0 times), then we consult $m_1^{\widehat{\mathbb{T}}}$. Suppose $m_1^{\widehat{\mathbb{T}}} = (0, 1)$. Then we move $\widehat{\mathbb{T}}$ up by one tile, and accordingly move all other tiles in the same column as $\widehat{\mathbb{T}}$ up by one tile.

We note a few things:

- This is equivalent to the Diaconis shuffle: a move in the Diaconis shuffle corresponds to picking the appropriate master tile and moving it in the the correct direction. Note that moving a particular way in the step has probability $\frac{1}{4n}$ in either depiction.
- It does not matter how the n master tiles are selected; any choice produces the same distribution of moves.
- Step 3 considers how many times $\widehat{\mathbb{T}}$ has *moved*, which may be different from how many times $\widehat{\mathbb{T}}$ has been a master tile.

Because it does not matter how we select the master tiles, we shall for the purposes of this lemma choose the master tiles in the following way:

(1) Choose i to be a master tile.

(2) If j is not in the same row or column as i, choose j to be a master tile; i.e., choose j if possible.

(3) If k is not in the same row or column as any of the master tiles chosen so far (i.e., in steps (1) or (2)), choose k; i.e., choose k if possible.

(4) Choose remaining master tiles arbitrarily.

With this choice of master tiles, observe that as a result:

- Tile i moves at a particular time if and only if it is chosen as the master tile to move at that time; moreover, i always moves according to its random walk sequence $\{m_q^i\}$.
- Tiles j and k *mostly* move according to their random walk sequences; any instance in which they do not results from being in the same row or column as i (or j, in the case of k).

We characterize the instances that j and k do not move according to their random walk sequences as *interference*; i interferes with j while i and j interfere with k. However, we will argue that the amount of interference is likely to be small, and hence the random walk sequences of each of i, j, and k serve as a mostly accurate representation of their movement.

In the lazy Diaconis shuffle, let $X_t(\mathbb{T})$ represent the position of tile $\mathbb{T}$ at time t, and call these the P_1 processes, or the *actual process*. That is to say, let $P_{1,t}(\mathbb{T}) = X_t(\mathbb{T})$, for $\mathbb{T} = i, j, k$.

Let ν_i be the number of times i moves after $2\ell^2 n$ steps of the lazy Diaconis shuffle, and let ν_j and ν_k be similarly defined.

Then we know that

$$P_{1,2\ell^2 n}(i) = X_{2\ell^2 n}(i) = X_0(i) + \sum_{s=1}^{\nu_i} m_s^i \tag{6.1}$$

and by our discussion above on interference, we expect that

$$P_{1,2\ell^2 n}(j) = X_{2\ell^2 n}(j) \approx X_0(j) + \sum_{s=1}^{\nu_j} m_s^j \tag{6.2}$$

$$P_{1,2\ell^2 n}(k) = X_{2\ell^2 n}(k) \approx X_0(k) + \sum_{s=1}^{\nu_k} m_s^k. \tag{6.3}$$

To describe these approximations, we introduce what we call the P_2 process, or the *idealized process*:

Both processes start at the same position, so $P_{1,0}(\mathbb{T}) = P_{2,0}(\mathbb{T})$. With the implementation of the Diaconis shuffle using the master tiles as described above, have tile k move in the idealized process only if tile k moves in the Diaconis shuffle; i.e., the actual process, in the following way:

- If tile $\mathbb{T}$ moved for the qth time at time t in P_1 because it was chosen as a master tile, then move tile $\mathbb{T}$ in the same way at time t in P_2; since $\mathbb{T}$ was a master tile, this means it moves in the direction of $m_q^{\mathbb{T}}$.
- If tile $\mathbb{T}$ moved for the qth time at time t in P_1 without being a master tile, then move tile $\mathbb{T}$ randomly according to its random walk at time t in P_2; i.e., move in the direction of $m_q^{\mathbb{T}}$.

Note that either way the tile moves according to $m_q^{\mathbb{T}}$, but in the first case the movement agrees with the Diaconis shuffle, while in the second case the movement is independent of the Diaconis shuffle.

This means that the P_2 processes $P_{2,t}(\mathbb{T})$ for $\mathbb{T} = i, j, k$ may be defined as follows: If at time t, tile $\mathbb{T}$ has moved μ times in the corresponding P_1 process, then

$$P_{2,t}(\mathbb{T}) = X_0(\mathbb{T}) + \sum_{s=1}^{\mu} m_s^{\mathbb{T}}. \tag{6.4}$$

We call these the *idealized processes* since they represent the trajectory a tile would have taken were it not for interference; that is to say, P_1 gives the actual trajectory of tile k including the interference it receives, while P_2 gives the ideal trajectory of tile k had it not received interference.

Thus, using the notation for the P_2 processes, we have from earlier that

$$P_{1,2\ell^2 n}(i) = X_{2\ell^2 n}(i) = X_0(i) + \sum_{s=1}^{\nu_i} m_s^i = P_{2,2\ell^2 n}(i) \tag{6.5}$$

$$P_{1,2\ell^2 n}(j) = X_{2\ell^2 n}(j) \approx X_0(j) + \sum_{s=1}^{\nu_j} m_s^j = P_{2,2\ell^2 n}(j) \tag{6.6}$$

$$P_{1,2\ell^2 n}(k) = X_{2\ell^2 n}(k) \approx X_0(k) + \sum_{s=1}^{\nu_k} m_s^k = P_{2,2\ell^2 n}(k). \tag{6.7}$$

Next, observe that individually, ν_i, ν_j, and ν_k are each binomial random variables with parameters $(2\ell^2 n, \frac{1}{2n})$. However, they are not independent; for example, if tile i moves only a few times (so ν_i is small) then we expect that tiles j and k would move more times (so ν_j, ν_k should be large). Additionally, we would expect to see a difference in, for example, where tile i ends up between the cases where ν_i is small and ν_i is large, as a larger ν_i would give a greater likelihood of being further away from its starting point.

To address these concerns, we will argue that it is likely that ν_i, ν_j, and ν_k are close to their expected value of ℓ^2. Additionally, if they are close to ℓ^2, we can argue that it is likely that the differences between $\sum_{s=1}^{\nu_{\mathbb{T}}} m_s^{\mathbb{T}}$ and $\sum_{s=1}^{\ell^2} m_s^{\mathbb{T}}$ for each tile $\mathbb{T}$ is small. That is to say, we can approximate the movement of $\nu_{\mathbb{T}}$ moves of a random walk by considering what happens in ℓ^2 moves of the same random walk instead, the idea being that a few too many or a few too few moves will not make much of a difference.

We call these further approximations using the expected value of ℓ^2 the P_3 processes, or the *prescribed processes*, which we can write as follows: For tile $\mathbb{T}$, let

$$P_3(\mathbb{T}) = X_0(\mathbb{T}) + \sum_{s=1}^{\ell^2} m_s^{\mathbb{T}}. \tag{6.8}$$

Note that these are not really processes, but rather the prescribed results of running the P_2 processes for a fixed number of moves.

Therefore, we have that

$$P_{1,2\ell^2 n}(i) = X_{2\ell^2 n}(i) = X_0(i) + \sum_{s=1}^{\nu_i} m_s^i = P_{2,2\ell^2 n}(i) \approx X_0(i) + \sum_{s=1}^{\ell^2} m_s^i = P_3(i) \tag{6.9}$$

$$P_{1,2\ell^2 n}(j) = X_{2\ell^2 n}(j) \approx X_0(j) + \sum_{s=1}^{\nu_j} m_s^j = P_{2,2\ell^2 n}(j) \approx X_0(j) + \sum_{s=1}^{\ell^2} m_s^j = P_3(j) \tag{6.10}$$

$$P_{1,2\ell^2 n}(k) = X_{2\ell^2 n}(k) \approx X_0(k) + \sum_{s=1}^{\nu_k} m_s^k = P_{2,2\ell^2 n}(k) \approx X_0(k) + \sum_{s=1}^{\ell^2} m_s^k = P_3(k). \tag{6.11}$$

Recall that we wanted to show that with probability bounded away from zero, $P_{1,2\ell^2 n}(\mathbb{T})$ is in $\mathcal{B}_{\mathbb{T}}$ for $\mathbb{T} = i, j, k$. However, the P_1 processes, as well as the P_2 processes, are difficult to understand. The P_3 processes, however, are the result of a lazy simple two-dimensional random walk, which is well studied. Thus, we will argue that with some probability bounded away from zero, the P_3 processes will be within the smaller box $\mathcal{L}_{\mathbb{T}}$ within each target box, and the P_1 process will not be more than $\delta\ell$ away with respect to the supremum norm from the P_3 process. This will ensures that if the P_3 process lands within $\mathcal{L}_{\mathbb{T}}$, then the corresponding P_1 process must be inside $\mathcal{B}_{\mathbb{T}}$.

Using the triangle inequality, it is enough to specify that the supremum distances between the P_1 and P_2 process and between the P_2 and P_3 processes do not exceed $\frac{1}{2}\delta\ell$.

Thus, we have that our desired probability is at least the probability that each of the P_3 processes lands in their respective $\mathcal{L}_{\mathbb{T}}$, minus the probability that any of the pairs of processes are too far

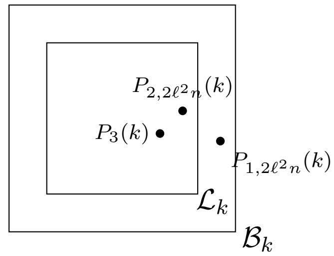

FIGURE 6.3. In the case of k: If P_3 lands in $\mathcal{L}_k$, P_3 and P_2 are close, and P_2 and P_1 are close, then P_1 will be in $\mathcal{B}_k$. (Not to scale)

apart:

$$\begin{aligned}
&\mathbb{P}(P_{1,2\ell^2 n}(\mathbb{T}) \in B_{\mathbb{T}}, \mathbb{T} = i,j,k) \\
&\geq \mathbb{P}\left(d_\infty(P_{1,2\ell^2 n}(\mathbb{T}), \mathcal{L}_{\mathbb{T}}) \leq \delta\ell, \mathbb{T} = i,j,k\right) \\
&\geq \mathbb{P}(P_3(\mathbb{T}) \in \mathcal{L}_{\mathbb{T}}, \mathbb{T} = i,j,k) \\
&\quad - \mathbb{P}\Bigg(\left\|P_{2,2\ell^2 n}(\mathbb{T}) - P_3(\mathbb{T})\right\|_\infty > \frac{1}{2}\delta\ell \text{ or } \left\|P_{1,2\ell^2 n}(\mathbb{T}) - P_{2,2\ell^2 n}(\mathbb{T})\right\|_\infty > \frac{1}{2}\delta\ell, \\
&\qquad \text{for at least one of } \mathbb{T} = i,j,k\Bigg)
\end{aligned} \tag{6.12}$$

where we write $d_\infty(P_{1,2\ell^2 n}(\mathbb{T}), \mathcal{L}_{\mathbb{T}})$ to mean the supremum distance between $P_{1,2\ell^2 n}(\mathbb{T})$ and $\mathcal{L}_{\mathbb{T}}$. By employing a union bound to make the second term more tractable, we get the following:

$$\begin{aligned}
\mathbb{P}(P_{1,2\ell^2 n}(\mathbb{T}) \in B_{\mathbb{T}}, \mathbb{T} = i,j,k) &\geq \mathbb{P}\left(d_\infty(P_{1,2\ell^2 n}(\mathbb{T}), \mathcal{L}_{\mathbb{T}}) \leq \delta\ell, \mathbb{T} = i,j,k\right) \\
&\geq \mathbb{P}(P_3(\mathbb{T}) \in \mathcal{L}_{\mathbb{T}}, \mathbb{T} = i,j,k) \\
&\quad - \sum_{\mathbb{T}=i,j,k} \mathbb{P}\left(\left\|P_{2,2\ell^2 n}(\mathbb{T}) - P_3(\mathbb{T})\right\|_\infty > \frac{1}{2}\delta\ell\right) \\
&\quad - \sum_{\mathbb{T}=i,j,k} \mathbb{P}\left(\left\|P_{1,2\ell^2 n}(\mathbb{T}) - P_{2,2\ell^2 n}(\mathbb{T})\right\|_\infty > \frac{1}{2}\delta\ell\right).
\end{aligned} \tag{6.13}$$

To calculate the first term of the lower bound in (6.13), consider tile i in the prescribed process, where it takes exactly the expected number of ℓ^2 steps of its random walk $\{m_q^i\}_{q=1}^\infty$. We recall that each m_q^i is a random vector in $\mathbb{R}^2$ that is either a unit step in any of the 4 directions with probability

$\frac{1}{4}$ each. The average step is $\langle 0,0\rangle$ and the covariance matrix of a step is $\begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$, and the steps are independent and identically distributed.

The position of tile i relative to its starting point after these ℓ^2 moves is the sum of these steps. (That is to say, tile i performs a 2-dimensional random walk.) Let $\widehat{I^i} = \sum_{q=1}^{\ell^2} m_q^i$ be this sum. Note that $\widehat{I^i}$ has mean $\langle 0,0\rangle$ and covariance matrix $\Sigma = \begin{bmatrix} \frac{\ell^2}{2} & 0 \\ 0 & \frac{\ell^2}{2} \end{bmatrix}$. By the (2-dimensional) central limit theorem, we know that $\widehat{I^i}$ should be approximately a normal random variable. In particular, [**1**] gives a Berry-Esseen bound on how close this approximation is:

THEOREM 6.1.1. *If Z is a 2-dimensional normal random variable with the same (zero) mean and the same covariance matrix as $\widehat{I^i}$, then for any convex set $U \subset \mathbb{R}^2$,*

$$\left|\mathbb{P}(\widehat{I^i} \in U) - \mathbb{P}(Z \in U)\right| \leq C(2)^{\frac{1}{4}}\gamma, \tag{6.14}$$

where C is a universal constant and $\gamma = \sum_{q=1}^{\ell^2} \mathbb{E}\left[\left\|\Sigma^{-1/2} m_q^i\right\|_2^3\right]$.

Since $\Sigma^{-1/2} = \begin{bmatrix} \frac{\sqrt{2}}{\ell} & 0 \\ 0 & \frac{\sqrt{2}}{\ell} \end{bmatrix}$ and m_q^i is a random unit step as described above, it follows that

$$\left\|\Sigma^{-1/2} m_q^i\right\|_2 = \frac{\sqrt{2}}{\ell}, \tag{6.15}$$

so we have that

$$\left\|\Sigma^{-1/2} m_q^i\right\|_2^3 = \frac{2\sqrt{2}}{\ell^3}, \tag{6.16}$$

so $\mathbb{E}\left[\left\|\Sigma^{-1/2} X_t\right\|_2^3\right] = \frac{2\sqrt{2}}{\ell^3}$ for all t. It follows that $\gamma = \frac{2\sqrt{2}}{\ell}$.

Since U can be any convex set, we can apply the theorem to the small box $\mathcal{L}_i$, which we recall is a box of side length $c\ell - 2\delta\ell$. We can thus estimate $\mathbb{P}(P_3(i) \in \mathcal{L}_i)$: by definition, this is exactly the probability that $\widehat{I^i}$ is in $\mathcal{L}_i - X_0(i)$, which is $\mathcal{L}_i$ shifted by the initial position, and hence still a box of side length $c\ell - 2\delta\ell$. By the Berry-Esseen result, this is approximately the probability that the corresponding normal random variable Z is in $\mathcal{L}_i - X_0(i)$. Since Z is bivariate normal and the covariance matrix is diagonal, the coordinates are independent. Then, we can find this probability for Z by considering each coordinate separately. Each coordinate is a normal random variable with mean zero and variance $\frac{\ell^2}{2}$. Hence, if Z_x and Z_y are the coordinates of Z, we have that

$$\begin{aligned}\mathbb{P}(P_3(i) \in \mathcal{L}_i) &= \mathbb{P}(\widehat{I}^i \in \mathcal{L}_i - X_0(i)) \\ &\approx \mathbb{P}(Z \in \mathcal{L}_i - X_0(i)) \\ &= \mathbb{P}(Z_x \in (\mathcal{L}_i)_x - (X_0(i))_x) \cdot \mathbb{P}(Z_y \in (\mathcal{L}_i)_y - (X_0(i))_y),\end{aligned} \tag{6.17}$$

where $(\mathcal{L}_i)_x - (X_0(i))_x$ is the horizontal interval corresponding to the shifted box, and $(\mathcal{L}_i)_y - (X_0(i))_y$ is likewise the vertical interval.

We take a lower estimate on this probability by observing that no matter where the box is relative to i's starting position, the far side of $\mathcal{L}_i$ is no more than a distance ℓ away from i's starting position. Thus, we find that

$$\begin{aligned}\mathbb{P}(Z_x \in (\mathcal{L}_i)_x - (X_0(i))_x) &\geq \mathbb{P}(Z_x \in (\ell - (c\ell - 2\delta\ell), \ell)) \\ &= \mathbb{P}(\bar{Z}_x \in \left(\sqrt{2}(1 - c + 2\delta), \sqrt{2}\right)),\end{aligned} \tag{6.18}$$

where $\bar{Z}_x = \frac{\sqrt{2}}{\ell} Z_x$, the normalized form of Z_x.

Recall that we wanted $0 < c < \frac{1}{3}$ and $0 < \delta < \frac{1}{2}c$; for such c and δ, $(1 - c + 2\delta) < 1$, so $\mathbb{P}(\bar{Z}_x \in \left(\sqrt{2}(1 - c + 2\delta), \sqrt{2}\right)) = \eta$ for some positive constant $\eta > 0$.

So, we find that

$$\mathbb{P}(Z_x \in (\mathcal{L}_i)_x - (X_0(i))_x) \geq \eta > 0. \tag{6.19}$$

The same argument applies for Z_y, so we have that

$$\mathbb{P}(Z_y \in (\mathcal{L}_i)_y - (X_0(i))_y) \geq \eta. \tag{6.20}$$

Therefore, we have that

$$\mathbb{P}(Z \in \mathcal{L}_i - X_0(i)) \geq \eta^2. \tag{6.21}$$

From line (6.14) and the calculation of γ above, it follows that there exists a universal constant $\widehat{C}$ such that

$$\mathbb{P}(P_3(i) \in \mathcal{L}_i) = \mathbb{P}(\widehat{I}^i \in \mathcal{L}_i - X_0(i)) \geq \eta^2 - \frac{\widehat{C}}{\ell}. \tag{6.22}$$

We observe that the same argument can be made for j and k as well, so we obtain the same bound for them.

Since the internal random walks are independent, the P_3 processes are independent. It follows that

$$\begin{aligned}\mathbb{P}(P_3(\mathbb{T}) \in \mathcal{L}_\mathbb{T}, \mathbb{T} = i, j, k) &= \mathbb{P}(P_3(i) \in \mathcal{L}_i) \cdot \mathbb{P}(P_3(j) \in \mathcal{L}_j) \cdot \mathbb{P}(P_3(k) \in \mathcal{L}_k) \\ &\geq \left(\eta^2 - \frac{\widehat{C}}{\ell}\right)^3.\end{aligned} \tag{6.23}$$

Now we consider the second term of the lower bound in (6.13), which we recall is

$$-\sum_{\mathbb{T}=i,j,k} \mathbb{P}\left(\left\|P_{2,2\ell^2 n}(\mathbb{T}) - P_3(\mathbb{T})\right\|_\infty > \frac{1}{2}\delta\ell\right). \tag{6.24}$$

Consider the case for tile i first; we shall see that j and k will be the same.

Recall that $P_3(i)$ denotes the random walk process for ℓ^2 steps and $P_{2,2\ell^2 n}(i)$ denotes the same random walk process for ν_i steps, where ν_i is a binomial random variable with parameters $(2\ell^2 n, \frac{1}{2n})$.

First observe that since the two processes use the same random walk $\{m_q^i\}$, any difference must come from $P_{2,2\ell^2 n}(i)$ using a different number of random walk steps than ℓ^2. We shall argue that this different number is likely to be close to ℓ^2, and when it is, that it is likely that there is not much difference between the two processes using Doob's inequality.

Let $N_{\text{tol}} = \kappa\ell$ represent the *tolerance level* of the number of missing or excess steps, where $\kappa > 0$ is a constant to be determined. Consider what ν_i could be. If ν_i is too far away from ℓ^2, so that $\left|\nu_i - \ell^2\right| \geq N_{\text{tol}}$, then we shall assume (pessimistically) that the two processes will be too far apart.

Otherwise, if $\left|\nu_i - \ell^2\right| < N_{\text{tol}}$, then observe that

$$\begin{aligned} \left\|P_{2,2\ell^2 n}(i) - P_3(i)\right\|_\infty = \left\|\sum_{s=1}^{\nu_i} m_s^i - \sum_{s=1}^{\ell^2} m_s^i\right\|_\infty &\le \max_{\mu:|\mu-\ell^2|<N_{\text{tol}}} \left\|\sum_{s=1}^{\mu} m_s^i - \sum_{s=1}^{\ell^2} m_s^i\right\|_\infty \\ &\le \max_{\mu:|\mu-\ell^2|\le N_{\text{tol}}} \left\|\sum_{s=1}^{\mu} m_s^i - \sum_{s=1}^{\ell^2} m_s^i\right\|_\infty . \end{aligned} \tag{6.25}$$

Note that for sufficiently large ℓ, we have $N_{\text{tol}} = \kappa\ell \le \ell^2$, so the μ in the maximum will be greater than 0.

It follows that by applying a union bound we have that

$$\begin{aligned} &\mathbb{P}\left(\left\|P_{2,2\ell^2 n}(i) - P_3(i)\right\|_\infty > \frac{1}{2}\delta\ell\right) \\ &\le \mathbb{P}(\left|\nu_i - \ell^2\right| \ge N_{\text{tol}}) + \mathbb{P}\left(\max_{\mu:|\mu-\ell^2|\le N_{\text{tol}}} \left\|\sum_{s=1}^{\mu} m_s^i - \sum_{s=1}^{\ell^2} m_s^i\right\|_\infty > \frac{1}{2}\delta\ell\right). \end{aligned} \tag{6.26}$$

By breaking up the maximum into the cases where $\mu < \ell^2$ and when $\mu > \ell^2$ (note that when $\mu = \ell^2$ the quantity is 0), we can use a union bound for the second term:

$$\begin{aligned} &\mathbb{P}\left(\max_{\mu:|\mu-\ell^2|\le N_{\text{tol}}} \left\|\sum_{s=1}^{\mu} m_s^i - \sum_{s=1}^{\ell^2} m_s^i\right\|_\infty > \frac{1}{2}\delta\ell\right) \\ &\le \mathbb{P}\left(\max_{\mu<\ell^2,\ell^2-\mu\le N_{\text{tol}}} \left\|\sum_{s=\mu+1}^{\ell^2} m_s^i\right\|_\infty > \frac{1}{2}\delta\ell\right) + \mathbb{P}\left(\max_{\mu>\ell^2,\mu-\ell^2\le N_{\text{tol}}} \left\|\sum_{s=\ell^2+1}^{\mu} m_s^i\right\|_\infty > \frac{1}{2}\delta\ell\right). \end{aligned} \tag{6.27}$$

Since the m_s^i are independent and identically distributed, these two terms are both bounded above by the probability that the maximum supremum norm attained over the first N_{tol} partial sums of the m_s^i is greater than $\frac{1}{2}\delta\ell$, so we can write that

$$\mathbb{P}\left(\max_{\mu:|\mu-\ell^2|\le N_{\text{tol}}} \left\|\sum_{s=1}^{\mu} m_s^i - \sum_{s=1}^{\ell^2} m_s^i\right\|_\infty > \frac{1}{2}\delta\ell\right) \le 2\mathbb{P}\left(\max_{1\le r\le N_{\text{tol}}} \left\|\sum_{s=1}^{r} m_s^i\right\|_\infty > \frac{1}{2}\delta\ell\right). \tag{6.28}$$

Then the probability that the two processes differ by more than $\frac{1}{2}\delta\ell$ is at most the probability that the number of missing or excess steps exceeds this tolerance level (i.e., we assume that the two processes will be too far apart if ν_i is too large or too small), plus the probability that, given that the number of missing or excess steps is within the tolerance level, the missing or excess steps cause a deviation of more than $\frac{1}{2}\delta\ell$:

(6.29)
$$\mathbb{P}\left(\left\|P_{2,2\ell^2 n}(i) - P_3(i)\right\|_\infty > \frac{1}{2}\delta\ell\right) \le \mathbb{P}(\left|\nu_i - \ell^2\right| > N_{\mathrm{tol}}) + 2\mathbb{P}\left(\max_{1\le r\le N_{\mathrm{tol}}}\left\|\sum_{s=1}^r m_s^i\right\|_\infty > \frac{1}{2}\delta\ell\right).$$

The first probability can be bounded above with Chebyshev's inequality: Since ν_i is a binomial random variable, we know its variance is $\frac{\ell^2(2n-1)}{2n}$, and thus we get that

$$\mathbb{P}(\left|\nu_i - \ell^2\right| \ge N_{\mathrm{tol}}) \le \frac{\operatorname{Var}\nu_i}{N_{\mathrm{tol}}^2} = \frac{\ell^2(2n-1)}{2n\kappa^2\ell^2} \le \frac{1}{\kappa^2}. \tag{6.30}$$

Now consider the second probability. Since having the ℓ^∞ norm exceed $\frac{1}{2}\delta\ell$ means that either the norm of the x-coordinate or the norm of the y-coordinate exceeded $\frac{1}{2}\delta\ell$, a union bound gives that

$$\begin{aligned} &\mathbb{P}\left(\max_{1\le r\le N_{\mathrm{tol}}}\left\|\sum_{s=1}^r m_s^i\right\|_\infty > \frac{1}{2}\delta\ell\right) \\ &\le \mathbb{P}\left(\max_{1\le r\le N_{\mathrm{tol}}}\left|\left(\sum_{s=1}^r m_s^i\right)_x\right| > \frac{1}{2}\delta\ell\right) + \mathbb{P}\left(\max_{1\le r\le N_{\mathrm{tol}}}\left|\left(\sum_{s=1}^r m_s^i\right)_y\right| > \frac{1}{2}\delta\ell\right), \end{aligned} \tag{6.31}$$

where the subscript x and subscript y indicate that we take the x-coordinate and y-coordinate, respectively. Note that

$$\left(\sum_{s=1}^r m_s^i\right)_x = \sum_{s=1}^r \left(m_s^i\right)_x \tag{6.32}$$

and since the m_s^i were random variables uniformly distributed over the four directions, it follows that each $(m_s^i)_x$ is $+1$ or -1 with probability $\frac{1}{4}$ each, and 0 otherwise. Observe that $W_r = \sum_{s=1}^r \left(m_s^i\right)_x$ is a martingale, and hence $|W_r|$ is a submartingale by Jensen's inequality (or observe this directly). By Doob's maximal inequality, we have that

$$\mathbb{P}\left(\max_{1\le r\le N_{\mathrm{tol}}}\left|\left(\sum_{s=1}^r m_s^i\right)_x\right| > \frac{1}{2}\delta\ell\right) \le \frac{\mathbb{E}\left(\left|\left(\sum_{s=1}^{N_{\mathrm{tol}}} m_s^i\right)_x\right|\right)}{\frac{1}{2}\delta\ell}. \tag{6.33}$$

Jensen's inequality tells us that

$$\left(\mathbb{E}\left(\left|\left(\sum_{s=1}^{N_{\mathrm{tol}}} m_s^i\right)_x\right|\right)\right)^2 \leq \mathbb{E}\left(\left|\left(\sum_{s=1}^{N_{\mathrm{tol}}} m_s^i\right)_x\right|^2\right), \tag{6.34}$$

so it follows that

$$\frac{\mathbb{E}\left(\left|\left(\sum_{s=1}^{N_{\mathrm{tol}}} m_s^i\right)_x\right|\right)}{\frac{1}{2}\delta\ell} \leq \frac{\sqrt{\mathbb{E}\left(\left|\left(\sum_{s=1}^{N_{\mathrm{tol}}} m_s^i\right)_x\right|^2\right)}}{\frac{1}{2}\delta\ell}. \tag{6.35}$$

Observe that $\mathbb{E}\left(\left|\left(\sum_{s=1}^{N_{\mathrm{tol}}} m_s^i\right)_x\right|^2\right) = \mathbb{E}\left(\left(\left(\sum_{s=1}^{N_{\mathrm{tol}}} m_s^i\right)_x\right)^2\right)$ is the second moment of the sum; since the first moment of the sum is 0, the second moment is the variance. Since the individual steps $(m_s^i)_x$ are independent, the variance of their sum is the sum of their variances. Each step has variance $\frac{1}{2}$, so it follows that the sum of the variances is $\frac{N_{\mathrm{tol}}}{2}$. Consequently,

$$\mathbb{P}\left(\max_{1\leq r\leq N_{\mathrm{tol}}}\left|\left(\sum_{s=1}^{r} m_s^i\right)_x\right| > \frac{1}{2}\delta\ell\right) \leq \frac{\sqrt{\mathbb{E}\left(\left|\left(\sum_{s=1}^{N_{\mathrm{tol}}} m_s^i\right)_x\right|^2\right)}}{\frac{1}{2}\delta\ell} = \frac{\sqrt{\frac{N_{\mathrm{tol}}}{2}}}{\frac{1}{2}\delta\ell} = \frac{\sqrt{\frac{\kappa}{2}\ell}}{\frac{1}{2}\delta\ell}. \tag{6.36}$$

The same argument applies for the y-coordinate, so we have that

$$\mathbb{P}\left(\max_{1\leq r\leq N_{\mathrm{tol}}}\left\|\sum_{s=1}^{r} m_s^i\right\|_\infty > \frac{1}{2}\delta\ell\right) \leq 2\frac{\sqrt{\frac{\kappa}{2}\ell}}{\frac{1}{2}\delta\ell} = \frac{2\sqrt{2\kappa}}{\delta\sqrt{\ell}}, \tag{6.37}$$

so in total we have that

$$\mathbb{P}\left(\left\|P_{2,2\ell^2 n}(i) - P_3(i)\right\|_\infty > \frac{1}{2}\delta\ell\right) \leq \frac{1}{\kappa^2} + \frac{4\sqrt{2\kappa}}{\delta\sqrt{\ell}}. \tag{6.38}$$

We see that the same argument applies for j and k, so we conclude that

$$-\sum_{\mathbb{T}=i,j,k} \mathbb{P}\left(\left\|P_{2,2\ell^2 n}(\mathbb{T}) - P_3(\mathbb{T})\right\|_\infty > \frac{1}{2}\delta\ell\right) \geq -3\left(\frac{1}{\kappa^2} + \frac{4\sqrt{2\kappa}}{\delta\sqrt{\ell}}\right), \tag{6.39}$$

assuming that we chose the same tolerance level κ for each tile. We shall choose an appropriate κ later.

Finally, we consider the last term of the lower bound in (6.13), which we recall is

$$-\sum_{\mathbb{T}=i,j,k} \mathbb{P}\left(\left\|P_{1,2\ell^2 n}(\mathbb{T}) - P_{2,2\ell^2 n}(\mathbb{T})\right\|_\infty > \frac{1}{2}\delta\ell\right). \tag{6.40}$$

Consider the case for k. Recall that $P_{1,t}(k) = X_t(k)$, the position of tile k after t steps of the lazy Diaconis shuffle.

Let $Z_t^k = X_t(k) - P_{2,t}(k)$, the vector difference between the two processes. Thus, we want to consider the following probability:

$$\mathbb{P}\left(\left\|Z_{2\ell^2 n}^k\right\|_\infty > \frac{1}{2}\delta\ell\right). \tag{6.41}$$

Define N_t^k to be the number of times in the first t steps of the P_1 process that tile k moved without being a master tile.

We claim that $\left\|Z_t^k\right\|_2^2 - 2N_t^k$ is a martingale: Let t increment (i.e., take one step in the lazy Diaconis shuffle, updating $X_t(k)$, and then update $P_{2,t}(k)$ accordingly). Then N_t^k either remains unchanged, or increases by 1. If N_t^k remains unchanged, then either tile k does not move at all, in which case Z_t^k evidently does not change, or tile k moves while being a master tile, in which case tile k moves in the same way in both processes, so that Z_t^k still remains unchanged. Either way, $\left\|Z_t^k\right\|_2^2$ is the same.

Otherwise, if N_t^k increases by 1, then the moves that tile k makes in P_1 and P_2 are independent. In this case, observe that $\left\|Z_t^k\right\|_2^2$ is the sum of the squares of its coordinates. Write $X_t(k) = (x_t, y_t)$ and $P_{2,t}(k) = (p_t, q_t)$. Then we get that

$$\left\|Z_t^k\right\|_2^2 = (x_t - p_t)^2 + (y_t - q_t)^2. \tag{6.42}$$

If we consider just the x-coordinate, then we see that

$$\mathbb{E}((x_{t+1} - p_{t+1})^2) = \mathbb{E}((x_t - p_t)^2 + (\Delta x)^2 + (\Delta p)^2), \tag{6.43}$$

where Δx and Δp represent the step made at $t+1$. They are independent random variables that are 0 with probability $\frac{1}{2}$ and -1 or 1 with probability $\frac{1}{4}$ each. Note that the cross terms have disappeared because Δx and Δp have mean 0, and $\mathbb{E}((\Delta x)^2) = \mathbb{E}((\Delta p)^2) = \frac{1}{2}$. Thus, we get

$$\mathbb{E}(x_{t+1} - p_{t+1})^2 = \mathbb{E}((x_t - p_t)^2 + 1 \tag{6.44}$$

A similar argument holds for the y-coordinate, and thus it follows that

$$\mathbb{E}\left(\left\|Z_{t+1}^k\right\|_2^2\right) = \mathbb{E}\left(\left\|Z_t^k\right\|_2^2\right) + 2 \tag{6.45}$$

It follows that $\left\|Z_t^k\right\|_2^2 - 2N_t^k$ is indeed a martingale.

Additionally, note that this quantity is initially 0, which implies that

$$\mathbb{E}\left(\left\|Z_t^k\right\|_2^2\right) = 2\mathbb{E}(N_t^k). \tag{6.46}$$

Define A_t^k was to be the number of times tile k was not a master tile after t steps. (Recall that we always chose i to be a master tile, then j if possible, then k is possible, then we chose the remaining master tiles arbitrarily.) Since k has a $\frac{1}{2n}$ chance at moving at each step, it follows that

$$\mathbb{E}(N_t^k) = \frac{1}{2n}\mathbb{E}(A_t^k). \tag{6.47}$$

There are two possible reasons that k is not a master tile: either k shares a row or column with i or k shares a row or column with j (or both). So, let $A_t^{k,i}$ be the number of steps taken while k was in the same row or column as i after t steps, and let $A_t^{k,j}$ be the analogous quantity for j. For example, we have that $A_0^{k,i} = 0$; if we suppose that i and k were in the same row at the beginning at time $t = 0$, then it will be the case that $A_1^{k,i} = 1$, since the step that takes us from $t = 0$ to $t = 1$ will be taken while k and i share a row. (Note that it is irrelevant whether k and i share a row or column at time t for the value of $A_t^{k,i}$.)

Then it follows that

$$A_t^k \leq A_t^{k,i} + A_t^{k,j} \tag{6.48}$$

Consider $A_t^{k,i}$. We claim that that $\|X_t(i) - X_t(k)\|_1 - \frac{1}{2n} A_t^{k,i}$ is a martingale: Let t increment. Then either i and k were not in the same row or column or they were.

If they were not in the same row or column, then A_t^i remains unchanged. Additionally, if we write $X_t(i) - X_t(k) = (\Delta x, \Delta y)$, we have that neither coordinate is 0 and that $\|X_t(i) - X_t(k)\|_1 = |\Delta x| + |\Delta y|$. If neither i nor k moved, then $\|X_t(i) - X_t(k)\|_1$ is unchanged. If one of i or k moves, then $(\Delta x, \Delta y)$ becomes one of $(\Delta x + 1, \Delta y)$, $(\Delta x - 1, \Delta y)$, $(\Delta x, \Delta y + 1)$, or $(\Delta x, \Delta y - 1)$ with equal probability, and since neither Δx or Δy was 0, the average change of $\|X_t(i) - X_t(k)\|_1$ is 0. Note that it is not possible for both to move since they are not in the same row or column.

If i and k were in the same row or column, then $A_t^{k,i}$ increases by 1. Suppose without loss of generality that they were in the same row. Then we know that $X_t(i) - X_t(k) = (\Delta x, \Delta y) = (\Delta x, 0)$. With probability $\frac{1}{2n}$, either i or k moves vertically, causing Δy to become either 1 or -1; in either case, $\|X_t(i) - X_t(k)\|_1$ increases by 1. If that is not the case, then either i and k move horizontally in the same direction, in which case $\|X_t(i) - X_t(k)\|_1$ does not change, or neither i nor k move, in which case $\|X_t(i) - X_t(k)\|_1$ does not change either.

It follows that $\|X_t(i) - X_t(k)\|_1 - \frac{1}{2n} A_t^{k,i}$ is a martingale. Consequently,

$$\mathbb{E}\left(\|X_t(i) - X_t(k)\|_1 - \frac{1}{2n} A_t^{k,i} \right) = \|X_0(i) - X_0(k)\|_1, \tag{6.49}$$

from which using the triangle inequality and reverse triangle inequality gives that

$$\begin{aligned} \mathbb{E}\left(\frac{1}{2n} A_t^{k,i} \right) &= \mathbb{E}\left(\|X_t(i) - X_t(k)\|_1 - \|X_0(i) - X_0(k)\|_1 \right) \\ &\leq \mathbb{E}\left(\Big| \|X_t(i) - X_t(k)\|_1 - \|X_0(i) - X_0(k)\|_1 \Big| \right) \\ &\leq \mathbb{E}\left(\|(X_t(i) - X_t(k)) - (X_0(i) - X_0(k))\|_1 \right) \\ &= \mathbb{E}\left(\|(X_t(i) - X_0(i)) - (X_t(k) - X_0(k))\|_1 \right) \\ &\leq \mathbb{E}(\|X_t(i) - X_0(i)\|_1 + \|X_t(k) - X_0(k)\|_1), \end{aligned} \tag{6.50}$$

so we get

$$\frac{1}{2n} \mathbb{E}(A_t^{k,i}) \leq \mathbb{E}(\|X_t(i) - X_0(i)\|_1) + \mathbb{E}(\|X_t(k) - X_0(k)\|_1). \tag{6.51}$$

Recall that

$$\mathbb{E}(\|X_t(i) - X_0(i)\|_1) \leq \sqrt{2}\mathbb{E}(\|X_t(i) - X_0(i)\|_2), \tag{6.52}$$

and observe that by Jensen's inequality, we have

$$(\mathbb{E}(\|X_t(i) - X_0(i)\|_2))^2 \leq \mathbb{E}((\|X_t(i) - X_0(i)\|_2)^2), \tag{6.53}$$

which in turn gives us

$$\sqrt{2}\mathbb{E}(\|X_t(i) - X_0(i)\|_2) \leq \sqrt{2}\sqrt{\mathbb{E}((\|X_t(i) - X_0(i)\|_2)^2)}. \tag{6.54}$$

Write $X_t(i) = (x_t, y_t)$. Then by the definition of the 2-norm we have that

$$\begin{aligned} \mathbb{E}((\|X_t(i) - X_0(i)\|_2)^2) &= \mathbb{E}((x_t - x_0)^2 + (y_t - y_0)^2) \\ &= \mathbb{E}((x_t - x_0)^2) + \mathbb{E}((y_t - y_0)^2). \end{aligned} \tag{6.55}$$

Notice that $x_t - x_0$ is the difference between the initial x-coordinate and the final x-coordinate, and can be written as the sum of differences for each time step:

$$x_t - x_0 = \sum_{s=1}^{t} x_s - x_{s-1}. \tag{6.56}$$

The difference at a time step is really just a step of the lazy Diaconis shuffle from the view point of tile i's x-coordinate, and is -1 or $+1$ with probability $\frac{1}{8n}$ each, and 0 otherwise. Furthermore, each step has mean 0 and the steps are independent, so it follows that $\mathbb{E}((x_t - x_0)^2)$ is in fact the sum of the variances of the steps. Each step has variance $\frac{1}{4n}$ and there are t steps, so

$$\mathbb{E}((x_t - x_0)^2) = \frac{t}{4n}. \tag{6.57}$$

The same argument applies for the y-coordinate, so we have that

$$\mathbb{E}((y_t - y_0)^2) = \frac{t}{4n}, \tag{6.58}$$

upon which we have that

$$\mathbb{E}((\|X_t(i) - X_0(i)\|_2)^2) = \frac{t}{2n}. \tag{6.59}$$

It follows that

$$\sqrt{2}\sqrt{\mathbb{E}((\|X_t(i) - X_0(i)\|_2)^2)} = \sqrt{2}\sqrt{\frac{t}{2n}} = \sqrt{\frac{t}{n}} \tag{6.60}$$

and thus we have that

$$\mathbb{E}(\|X_t(i) - X_0(i)\|_1) \leq \sqrt{\frac{t}{n}}. \tag{6.61}$$

The same argument applies for the movement of tile k, so we get that

$$\frac{1}{2n}\mathbb{E}(A_t^{k,i}) \leq \mathbb{E}(\|X_t(i) - X_0(i)\|_1) + \mathbb{E}(\|X_t(k) - X_0(k)\|_1) \leq 2\sqrt{\frac{t}{n}}. \tag{6.62}$$

A similar argument for $A_t^{k,j}$ gives that

$$\frac{1}{2n}\mathbb{E}(A_t^{k,j}) \leq 2\sqrt{\frac{t}{n}}. \tag{6.63}$$

It thus follows that

$$\mathbb{E}\left(\left\|Z_t^k\right\|_2^2\right) = 2\mathbb{E}(N_t^k) = 2\left(\frac{1}{2n}\mathbb{E}(A_t^k)\right) \leq 2\left(\frac{1}{2n}\mathbb{E}(A_t^{k,i} + A_t^{k,j})\right) \leq 2 \cdot 4\sqrt{\frac{t}{n}} = 8\sqrt{\frac{t}{n}}. \tag{6.64}$$

By Jensen's inequality, we know that

$$\left(\mathbb{E}\left(\left\|Z_t^k\right\|_2\right)\right)^2 \leq \mathbb{E}\left(\left\|Z_t^k\right\|_2^2\right) \tag{6.65}$$

so it follows that

$$\mathbb{E}\left(\left\|Z_t^k\right\|_2\right) = \sqrt{(\mathbb{E}(\left\|Z_t^k\right\|_2))^2} \leq \sqrt{\mathbb{E}(\left\|Z_t^k\right\|_2^2)} = 2\sqrt{2}\left(\frac{t}{n}\right)^{1/4}. \tag{6.66}$$

When $t = 2\ell^2 n$, this becomes

$$\mathbb{E}\left(\left\|Z_{2\ell^2 n}^k\right\|_2\right) \leq 2\sqrt{2}\left(\frac{2\ell^2 n}{n}\right)^{1/4} = 2^{7/4}\sqrt{\ell}. \tag{6.67}$$

Finally, using Markov's inequality, we can get a bound on the probability that Z_t^k is too large. Since $\left\|Z_t^k\right\|_\infty \leq \left\|Z_t^k\right\|_2$, it follows that

$$\mathbb{P}\left(\left\|Z^k_{2\ell^2 n}\right\|_\infty > \frac{1}{2}\delta\ell\right) \leq \mathbb{P}\left(\left\|Z^k_{2\ell^2 n}\right\|_2 > \frac{1}{2}\delta\ell\right) \leq \frac{\mathbb{E}(\left\|Z^k_{2\ell^2 n}\right\|_2)}{\frac{1}{2}\delta\ell} \leq \frac{2^{7/4}\sqrt{\ell}}{\frac{1}{2}\delta\ell} = \frac{2^{11/4}}{\delta\sqrt{\ell}}. \tag{6.68}$$

An analogous argument holds for $\left\|Z_t^j\right\|_\infty$ with the following differences: Since j is not a master tile only if it shares a row or column with i, $A_t^j = A_t^{j,i}$, and thus we get that

$$\mathbb{E}\left(\left\|Z_t^j\right\|_2^2\right) = 2\mathbb{E}(N_t^j) = 2\left(\frac{1}{2n}\mathbb{E}(A_t^j)\right) = 2\left(\frac{1}{2n}\mathbb{E}(A_t^{j,i})\right) \leq 2\cdot 2\sqrt{\frac{t}{n}} = 4\sqrt{\frac{t}{n}}. \tag{6.69}$$

This, in turn, yields that

$$\mathbb{E}\left(\left\|Z_t^j\right\|_2\right) \leq 2\left(\frac{t}{n}\right)^{1/4} \tag{6.70}$$

and that

$$\mathbb{E}\left(\left\|Z^j_{2\ell^2 n}\right\|_2\right) \leq 2^{5/4}\sqrt{\ell}, \tag{6.71}$$

meaning the estimate in the case of j is

$$\mathbb{P}\left(\left\|Z^j_{2\ell^2 n}\right\|_\infty > \frac{1}{2}\delta\ell\right) \leq \frac{2^{9/4}}{\delta\sqrt{\ell}}. \tag{6.72}$$

Finally, note that Z_t^i is always 0 since i is always a master tile and hence will always move according to its internal random walk, so we have that

$$\mathbb{P}\left(\left\|Z^i_{2\ell^2 n}\right\|_\infty > \frac{1}{2}\delta\ell\right) = 0. \tag{6.73}$$

It thus follows that

$$-\sum_{\mathbb{T}=i,j,k}\mathbb{P}\left(\left\|P_{1,2\ell^2 n}(\mathbb{T}) - P_{2,2\ell^2 n}(\mathbb{T})\right\|_\infty > \frac{1}{2}\delta\ell\right) \geq -\frac{2^{11/4}+2^{9/4}}{\delta\sqrt{\ell}}. \tag{6.74}$$

Putting the results from lines (6.23), (6.39), and (6.74) together we find that

$$\begin{aligned}
\mathbb{P}(P_{1,2\ell^2 n}(\mathbb{T}) \in B_{\mathbb{T}}, \mathbb{T}=i,j,k) &\geq \mathbb{P}\left(d(P_{1,2\ell^2 n}(\mathbb{T}), \mathcal{L}_{\mathbb{T}}) \leq \delta\ell, \mathbb{T}=i,j,k\right) \\
&\geq \mathbb{P}(P_3(\mathbb{T}) \in \mathcal{L}_{\mathbb{T}}, \mathbb{T}=i,j,k) \\
&\quad - \sum_{\mathbb{T}=i,j,k} \mathbb{P}\left(\left\|P_{2,2\ell^2 n}(\mathbb{T}) - P_3(\mathbb{T})\right\|_\infty > \frac{1}{2}\delta\ell\right) \\
&\quad - \sum_{\mathbb{T}=i,j,k} \mathbb{P}\left(\left\|P_{1,2\ell^2 n}(\mathbb{T}) - P_{2,2\ell^2 n}(\mathbb{T})\right\|_\infty > \frac{1}{2}\delta\ell\right) \\
&\geq \left(\eta^2 - \frac{\widehat{C}}{\ell}\right)^3 - 3\left(\frac{1}{\kappa^2} + \frac{4\sqrt{2\kappa}}{\delta\sqrt{\ell}}\right) - \frac{2^{11/4}+2^{9/4}}{\delta\sqrt{\ell}}.
\end{aligned} \tag{6.75}$$

Recall that η is a constant. It follows that we may pick κ large enough so that if ℓ is sufficiently large, we have that

$$\mathbb{P}(P_{1,2\ell^2 n}(\mathbb{T}) \in B_{\mathbb{T}}, \mathbb{T}=i,j,k) \geq \left(\eta^2 - \frac{\widehat{C}}{\ell}\right)^3 - 3\left(\frac{1}{\kappa^2} + \frac{4\sqrt{2\kappa}}{\delta\sqrt{\ell}}\right) - \frac{2^{11/4}+2^{9/4}}{\delta\sqrt{\ell}} \geq \frac{\eta^6}{2}. \tag{6.76}$$

6.2. Stage 2 Analysis

In the second stage, we consider the behavior of the tiles i, j, and k, when they are in their respective boxes $\mathcal{B}_i$, $\mathcal{B}_j$, and $\mathcal{B}_k$. Recall that we want i to move from its starting position $i_0 \in \mathcal{B}_i$ to a final position $i_f \in \mathcal{B}_i$, and likewise for j and k. Suppose that we consider R steps of the shuffle for this stage, where R is an integer between $\beta\ell^2 n$ and $\beta\ell^2 n + 2$, where β is a constant to be determined. (This allows for R to be even or odd.) Then if $X_t(i)$ is the position of i in the lazy Diaconis shuffle with $X_0(i) = i_0$, and likewise for j and k, we want a lower bound for

$$\mathbb{P}(X_R(i) = i_f, X_R(j) = j_f, X_R(k) = k_f). \tag{6.77}$$

Let us first try our strategy from the first stage: Assign to every tile $\mathbb{T}$ an internal random walk with a new independent sequence of random walk steps $\{m_q^{\mathbb{T}}\}_{q=1}^{\infty}$. and consider the implementation of the Diaconis shuffle as seen in the first stage. This time, let ν_i be the number of times tile i moves after R steps of the lazy Diaconis shuffle, and define ν_j and ν_k similarly.

For tiles i, j, and k, we can consider the event that their random walks lead them from their starting positions to their target final positions; this would be

$$i_0 + \sum_{s=1}^{\nu_i} m_q^i = i_f, \qquad j_0 + \sum_{s=1}^{\nu_j} m_q^j = j_f, \qquad k_0 + \sum_{s=1}^{\nu_k} m_q^k = k_f. \tag{6.78}$$

We observe that not every such triplet of random walks would result in the three reaching their destination in the lazy Diaconis shuffle, since we saw earlier that there is the potential for interference.

However, if i, j, k are always master tiles, then there will be no interference, and they will move exactly according to their random walks. In such a case, the movement of i, j, and k under the lazy Diaconis shuffle would be the movement according to their random walks, and hence the three would reach their target destinations.

One way to ensure that i, j, and k are always master tiles is to consider only the random walk paths that keep each tile nearby its box. If the trajectories do not go so far out such that there is no possibility that any of the three tiles share a row or a column, then on the event that the random walks do take those paths, then the three tiles will always be master tiles and go to their target tiles as expected, without any potential for interference.

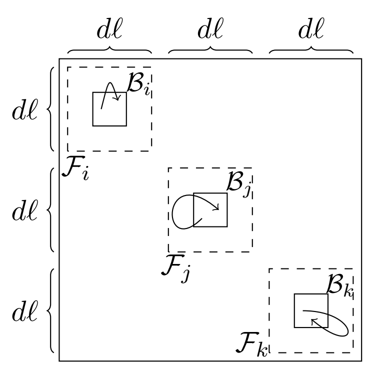

FIGURE 6.4. As long as i, j, and k stay close to $\mathcal{B}_i$, $\mathcal{B}_j$, and $\mathcal{B}_k$ by remaining inside $\mathcal{F}_i$, $\mathcal{F}_j$, and $\mathcal{F}_k$ respectively, there is no potential for interference.

To keep each tile near its box, we will consider paths that stay within buffer zones of each respective boxes. This will mean that we want

$$i_0 + \sum_{s=1}^{\nu_i} m_q^i = i_f, \qquad i_0 + \sum_{s=1}^{r} m_q^i \in \mathcal{F}_i \text{ for } 1 \leq r \leq \nu_i, \tag{6.79}$$

$$j_0 + \sum_{s=1}^{\nu_j} m_q^j = j_f, \qquad j_0 + \sum_{s=1}^{r} m_q^j \in \mathcal{F}_j \text{ for } 1 \leq r \leq \nu_j, \tag{6.80}$$

$$k_0 + \sum_{s=1}^{\nu_k} m_q^k = k_f, \qquad k_0 + \sum_{s=1}^{r} m_q^k \in \mathcal{F}_k \text{ for } 1 \leq r \leq \nu_k \tag{6.81}$$

where, for $\mathbb{T} = i, j, k$, $\mathcal{F}_\mathbb{T}$ is the box of side length $d\ell$ that is centered around $\mathcal{B}_\mathbb{T}$, where d is a constant to be determined and we remark that we should have $\frac{1}{3} > d > c$ (see Figure 6.4, recall that each $\mathcal{B}_\mathbb{T}$ has side length $c\ell$). If ℓ is sufficiently large, then we have that $\mathcal{F}_\mathbb{T}$ strictly contains $\mathcal{B}_\mathbb{T}$ in the sense that there is at least one layer of tiles between $\mathcal{F}_\mathbb{T}$ and $\mathcal{B}_\mathbb{T}$.

To summarize, one way that the three tiles will reach their target final positions in the lazy Diaconis shuffle in R steps is to have that if tiles i, j, and k move ν_i, ν_j, and ν_k times, respectively, then have the tiles have internal random walks that lead them to their targets in the way described earlier, keeping close to their boxes so there is no interference. That is,

$$\begin{aligned}
&\mathbb{P}(X_R(i) = i_f, X_R(j) = j_f, X_R(k) = k_f) \\
&\geq \mathbb{P}(i_0 + \sum_{s=1}^{\nu_i} m_q^i = i_f,\ i_0 + \sum_{s=1}^{r} m_q^i \in \mathcal{F}_i \text{ for } 1 \leq r \leq \nu_i, \\
&\quad j_0 + \sum_{s=1}^{\nu_j} m_q^j = j_f,\ j_0 + \sum_{s=1}^{r} m_q^j \in \mathcal{F}_j \text{ for } 1 \leq r \leq \nu_j, \\
&\quad k_0 + \sum_{s=1}^{\nu_k} m_q^k = k_f,\ k_0 + \sum_{s=1}^{r} m_q^k \in \mathcal{F}_k \text{ for } 1 \leq r \leq \nu_k)
\end{aligned} \tag{6.82}$$

We pause here to point out an issue of parity: if we only consider steps where the tile actually moves as we did before, then establishing a useful lower bound for the probability of moving from one position to another in the small box becomes difficult (since, for example, it is impossible to move to an adjacent space in an even number of moves).

Additionally, observe that checking whether a path stays within $\mathcal{F}_\mathbb{T}$ can be done by considering each coordinate; if a path leaves $\mathcal{F}_\mathbb{T}$, then it must be the case that the path travels too far horizontally or too far vertically. Since checking one dimension at a time is easier and we must also consider the parity issue above, we shall rework this approach to analyze one dimension at a time:

Assign for each tile $\mathbb{T}$ *two* 1-dimensional lazy random walks, a horizontal random walk $\{h_q^{\mathbb{T}}\}_{q=1}^{\infty}$ and a vertical random walk $\{v_q^{\mathbb{T}}\}_{q=1}^{\infty}$. Both are sequences of random variables and the $h_q^{\mathbb{T}}$ and $v_q^{\mathbb{T}}$ are independent and identically distributed, taking values of -1 or 1 with probability $\frac{1}{4}$ each, and 0 with probability $\frac{1}{2}$.

To make use of these lazy random walks, we shall consider a different implementation of the lazy Diaconis shuffle for the second stage in order to reflect the fact that the laziness is now incorporated into the internal random walks, as detailed below.

At each step:

(1) Choose n master tiles as described previously: set i, to be a master tile, then j if possible, then k if possible, then choose the rest arbitrarily.
(2) Randomly select one of the n master tiles; call it $\widehat{\mathbb{T}}$.
(3) Randomly decide whether to try to move $\widehat{\mathbb{T}}$ horizontally or vertically.
   - If $\widehat{\mathbb{T}}$ tries to move horizontally: Suppose $\widehat{\mathbb{T}}$ has already tried to move horizontally $(q-1)$ times. Move $\widehat{\mathbb{T}}$ according to $h_q^{\widehat{\mathbb{T}}}$, and move the other tiles accordingly.
   - If $\widehat{\mathbb{T}}$ tries to move vertically: Suppose $\widehat{\mathbb{T}}$ has already tried to move vertically $(q-1)$ times. Move $\widehat{\mathbb{T}}$ according to $v_q^{\widehat{\mathbb{T}}}$, and move the other tiles accordingly.

Under this framework, we shall consider each time a tile tries to move horizontally (respectively, vertically) walk to be a horizontal (respectively, vertical) "step" for that tile even if that step ends up doing nothing due to laziness. This will allow us to avoid the parity issue described earlier (as now, for example, it is possible to go to an adjacent tile with an even number of steps).

Let H_i, H_j, and H_k be the random amounts of horizontal steps that tiles i, j, and k take, respectively, and let V_i, V_j, and V_k be the random amounts of vertical steps that tiles i, j, and k take, respectively.

Suppose we know how much these quantities are. Then for tiles i, j, and k, having these random walks lead them from their starting positions to their target final positions is the same as having their random walks lead their coordinates to the correct values, so we would want the following, analogous to line (6.78):

$$(i_0)_x + \sum_{s=1}^{H_i} h_s^i = (i_f)_x, \qquad (i_0)_y + \sum_{s=1}^{V_i} v_s^i = (i_f)_y, \tag{6.83}$$

$$(j_0)_x + \sum_{s=1}^{H_j} h_s^j = (j_f)_x, \qquad (j_0)_y + \sum_{s=1}^{V_j} v_s^j = (j_f)_y, \tag{6.84}$$

$$(k_0)_x + \sum_{s=1}^{H_k} h_s^k = (k_f)_x, \qquad (k_0)_y + \sum_{s=1}^{V_k} v_s^k = (k_f)_y. \tag{6.85}$$

As described earlier, we want to consider only those random walks that keep each tile nearby its box, to ensure that all three of i, j, and k are always master tiles. In the context of using these one-dimensional random walks, the stipulation that the path stays inside $\mathcal{F}_{\mathbb{T}}$ becomes the stipulation that each coordinate stays within its corresponding interval derived from $\mathcal{F}_{\mathbb{T}}$. Thus, analogous to line (6.81) we would want

$$(i_0)_x + \sum_{s=1}^{H_i} h_s^i = (i_f)_x, \qquad (i_0)_x + \sum_{s=1}^{r} h_s^i \in [a_i, b_i] \text{ for } 1 \le r \le H_i, \tag{6.86}$$

$$(i_0)_y + \sum_{s=1}^{V_i} v_s^i = (i_f)_y, \qquad (i_0)_y + \sum_{s=1}^{r} v_s^i \in [c_i, d_i] \text{ for } 1 \le r \le V_i, \tag{6.87}$$

$$(j_0)_x + \sum_{s=1}^{H_j} h_s^j = (j_f)_x, \qquad (j_0)_x + \sum_{s=1}^{r} h_s^j \in [a_j, b_j] \text{ for } 1 \le r \le H_j, \tag{6.88}$$

$$(j_0)_y + \sum_{s=1}^{V_j} v_s^j = (j_f)_y, \qquad (j_0)_y + \sum_{s=1}^{r} v_s^j \in [c_j, d_j] \text{ for } 1 \le r \le V_j, \tag{6.89}$$

$$(k_0)_x + \sum_{s=1}^{H_k} h_s^k = (k_f)_x, \qquad (k_0)_x + \sum_{s=1}^{r} h_s^k \in [a_k, b_k] \text{ for } 1 \le r \le H_k, \tag{6.90}$$

$$(k_0)_y + \sum_{s=1}^{V_k} v_s^k = (k_f)_y, \qquad (k_0)_y + \sum_{s=1}^{r} v_s^k \in [c_k, d_k] \text{ for } 1 \le r \le V_k \tag{6.91}$$

where for $\mathbb{T} = i, j, k$, $a_{\mathbb{T}}$ is the x-coordinate of the leftmost column of tiles in $\mathcal{F}_{\mathbb{T}}$, $b_{\mathbb{T}}$ is the x-coordinate of the rightmost column of tiles in $\mathcal{F}_{\mathbb{T}}$, $c_{\mathbb{T}}$ is the y-coordinate of the bottom row of tiles in $\mathcal{F}_{\mathbb{T}}$, and $d_{\mathbb{T}}$ is the y-coordinate of the top row of tiles in $\mathcal{F}_{\mathbb{T}}$.

To have more compact notation, define G_{H_i} to be the "good" event for H_i as described above:

$$G_{H_i} := \left\{ (i_0)_x + \sum_{s=1}^{H_i} h_s^i = (i_f)_x,\ (i_0)_x + \sum_{s=1}^{r} h_s^i \in [a_i, b_i] \text{ for } 1 \leq r \leq H_i \right\} \tag{6.92}$$

and define G_{V_i}, G_{H_j}, G_{V_j}, G_{H_k}, and G_{V_k} analogously.

Then we see that in the implementation of the lazy Diaconis shuffle using two 1-dimensional lazy random walks for each tile as described earlier, one way to have the lazy Diaconis shuffle move i, j and k to their target final positions is to have the three tiles have random walks that satisfy the above events, given the amounts that each tile moves horizontally and vertically.

So then we have, analogous to line (6.82), that

$$\mathbb{P}(X_R(i) = i_f, X_R(j) = j_f, X_R(k) = k_f) \geq \mathbb{P}(G_{H_i}, G_{V_i}, G_{H_j}, G_{V_j}, G_{H_k}, G_{V_k}). \tag{6.93}$$

We know that each of $H_i, V_i, H_j, V_j, H_k, V_k$ is a binomial random variable with parameters $(R, \frac{1}{2n})$, and thus each has average $\frac{R}{2n}$. Since R is roughly $\beta\ell^2 n$, we expect that each will be close to $\frac{\beta\ell^2}{2}$. So, we shall consider the event that each quantity is in the interval $(\frac{\beta\ell^2}{4}, \frac{3\beta\ell^2}{4})$, which we denote as A:

$$A := \left\{ H_i, V_i, H_j, V_j, H_k, V_k \in \left(\frac{\beta\ell^2}{4}, \frac{3\beta\ell^2}{4} \right) \right\}. \tag{6.94}$$

Then, by considering all of the possible choices of amounts $n_{h,i}$, $n_{v,i}$, $n_{h,j}$, $n_{v,j}$, $n_{h,k}$, and $n_{v,k}$ for H_i, V_i, H_j, V_j, H_k, and V_k, respectively, in A, the estimate in line (6.93) can be modified in the following way.

Let $N = (H_i, V_i, H_j, V_j, H_k, V_k)$, the vector of the 6 quantities, and define

$$\mathcal{A} = \left\{ (n_{h,i}, n_{v,i}, n_{h,j}, n_{v,j}, n_{h,k}, n_{v,k}) \middle| n_{h,i}, n_{v,i}, n_{h,j}, n_{v,j}, n_{h,k}, n_{v,k} \in \left(\frac{\beta\ell^2}{4}, \frac{3\beta\ell^2}{4} \right) \right\} \tag{6.95}$$

Then we can write

(6.96)

$$\mathbb{P}(X_R(i) = i_f, X_R(j) = j_f, X_R(k) = k_f) \geq \sum_{n \in \mathcal{A}} \mathbb{P}(G_{H_i}, G_{V_i}, G_{H_j}, G_{V_j}, G_{H_k}, G_{V_k} | N = n) \mathbb{P}(N = n)$$

This lower bound is what we will estimate. We shall find a bound for $\mathbb{P}(G_{H_i}, G_{V_i}, G_{H_j}, G_{V_j}, G_{H_k}, G_{V_k}|N=n)$ that works for all $n \in \mathcal{A}$, allowing us to factor it out, which will leave us with $\sum_{n\in\mathcal{A}} \mathbb{P}(N=n)$, which is $\mathbb{P}(A)$.

First consider $\mathbb{P}(A)$. Since all six quantities are identically distributed, we consider H_i in particular. Recall that H_i is a binomial random variable with parameters $(R, \frac{1}{2n})$, and thus $\mathbb{E}(H_i)$ is $\frac{R}{2n}$. Since R is roughly $\beta\ell^2 n$, we expect that H_I will be approximately $\frac{\beta\ell^2}{2}$; we will want to estimate the probability that $H_i \in (\frac{\beta\ell^2}{4}, \frac{3\beta\ell^2}{4})$.

Since $\operatorname{Var} H_i = \frac{R}{2n} \cdot \frac{2n-1}{2n} \leq \frac{R}{2n}$, by Chebyshev's inequality we have that

$$\mathbb{P}\left(\left|H_i - \frac{R}{2n}\right| \geq \frac{\beta\ell^2}{8}\right) \leq \frac{\operatorname{Var} H_i}{\left(\frac{\beta\ell^2}{8}\right)^2} \leq \frac{\frac{R}{2n}}{\left(\frac{\beta\ell^2}{8}\right)^2} \tag{6.97}$$

But since $\beta\ell^2 \leq R \leq \beta\ell^2 + 2$, we have that $\frac{\beta\ell^2}{2} \leq \frac{R}{2n} \leq \frac{\beta\ell^2}{2} + \frac{1}{n}$, and so

$$\mathbb{P}\left(\left|H_i - \frac{R}{2n}\right| \geq \frac{\beta\ell^2}{8}\right) \leq \frac{\frac{R}{2n}}{\left(\frac{\beta\ell^2}{8}\right)^2} \leq \frac{\frac{\beta\ell^2}{2} + \frac{1}{n}}{\left(\frac{\beta\ell^2}{8}\right)^2} = \frac{32}{\beta\ell^2} + \frac{64}{\beta^2\ell^4 n} \tag{6.98}$$

and for sufficiently large ℓ this will mean that

$$\mathbb{P}\left(\left|H_i - \frac{R}{2n}\right| \geq \frac{\beta\ell^2}{8}\right) \leq \frac{64}{\beta\ell^2}. \tag{6.99}$$

We also see that by the triangle inequality

$$\left|H_i - \frac{\beta\ell^2}{2}\right| \leq \left|H_i - \frac{R}{2n}\right| + \left|\frac{R}{2n} - \frac{\beta\ell^2}{2}\right| \leq \left|H_i - \frac{R}{2n}\right| + \frac{1}{n}. \tag{6.100}$$

If ℓ is sufficiently large, $\frac{1}{n} < \frac{\beta\ell^2}{8}$. Then on the event that $\left|H_i - \frac{R}{2n}\right| < \frac{\beta\ell^2}{8}$ we would have that $\left|H_i - \frac{\beta\ell^2}{2}\right| < \frac{\beta\ell^2}{4}$, so it follows that

$$\mathbb{P}\left(H_i \in \left(\frac{\beta\ell^2}{4}, \frac{3\beta\ell^2}{4}\right)\right) \geq 1 - \mathbb{P}\left(\left|H_i - \frac{R}{2n}\right| \geq \frac{\beta\ell^2}{8}\right) \geq 1 - \frac{64}{\beta\ell^2} \tag{6.101}$$

and the same holds for H_j, H_k, V_i, V_j, and V_k. With the application of a union bound for all six quantities, it follows that

$$\mathbb{P}(A) \geq 1 - \left(6 \cdot \frac{64}{\beta\ell^2}\right) = 1 - \frac{384}{\beta\ell^2} \tag{6.102}$$

Now, for any choice of n for N, the events described in lines (6.86) to (6.91) are independent since the random walks are independent; it follows that the probability of the events in those lines conditioned on the event that $N = n$ is the product of the probabilities of the individual events conditioned on the event that $N = n$:

(6.103)

$$\mathbb{P}(G_{H_i}, G_{V_i}, G_{H_j}, G_{V_j}, G_{H_k}, G_{V_k} | N = n)$$

$$= \mathbb{P}(G_{H_i} | N = n) \cdot \mathbb{P}(G_{V_i} | N = n) \cdot \mathbb{P}(G_{H_j} | N = n) \cdot \mathbb{P}(G_{V_j} | N = n) \cdot \mathbb{P}(G_{H_k} | N = n) \cdot \mathbb{P}(G_{V_k} | N = n)$$

We shall go through

$$\mathbb{P}(G_{H_i} | N = n) = \mathbb{P}\Big((i_0)_x + \sum_{s=1}^{H_i} h_s^i = (i_f)_x, (i_0)_x + \sum_{s=1}^{r} h_s^i \in [a_i, b_i] \text{ for } 1 \leq r \leq H_i | N = n\Big); \tag{6.104}$$

the others will have a similar calculation.

Suppose it is the case that $N = n$ holds, for some $n \in \mathcal{A}$. (We remark that the only relevant part of this for G_{H_i} is that $H_i = n_{h,i} \in \left(\frac{\beta \ell^2}{4}, \frac{3\beta \ell^2}{4}\right)$.)

Let us first consider the probability of just the event that $(i_0)_x + \sum_{s=1}^{H_i} h_s^i = (i_f)_x$; that is, without the consideration of remaining in $[a_i, b_i]$.

Then we note that we have an aperiodic random walk on $\mathbb{Z}$ that takes a step of $+1$ or -1 with probability $\frac{1}{4}$ each, or remains still with a step of 0 with probability $\frac{1}{2}$.

By the local limit theorem considered in **[4]**, we have the following: Let $p_n(k)$ represent the probability that such a random walk (starting at 0) is at position k after n steps. Then we can approximate $p_n(k)$:

$$p_n(k) \approx \overline{p_n}(k) = \frac{1}{\sqrt{\pi n}} e^{-\frac{k^2}{n}} \tag{6.105}$$

and the local central limit theorem tells us that this is approximate in the following sense:

$$|p_n(k) - \overline{p_n}(k)| \leq \frac{\xi}{n^{3/2}} \tag{6.106}$$

for a universal constant ξ, giving an error bound. This will give us an estimate of the probability that $(i_0)_x + \sum_{s=1}^{H_i} h_s^i = (i_f)_x$.

Next, to incorporate the restriction of remaining in $[a_i, b_i]$, we employ the *reflection principle* discussed in [**5**]. Since any path that stays in (a_i, b_i) also stays in $[a_i, b_i]$, we shall consider the former for a cleaner argument.

Starting from $(i_0)_x$, we want to travel to $(i_f)_x$ in H_i steps without becoming less than or equal to a_i or greater than or equal to b_i. Similar to the notation used in [**5**], let X_t be the lazy simple random walk on $\mathbb{Z}$ with $X_0 = (i_0)_x$.

Let τ_{a_i} be the first time the walk hits a_i; i.e., one of the ways to go out of bounds. Then the reflection principle gives that

$$\mathbb{P}_{(i_0)_x}(\tau_{a_i} < H_i, X_{H_i} = (i_f)_x) = \mathbb{P}_{(i_0)_x}(X_{H_i} = 2a_i - (i_f)_x), \tag{6.107}$$

or that the probability that the walk goes to $(i_f)_x$ in H_i steps without hitting a_i is equal to the probability that the walk goes to $2a_i - (i_f)_x$ in H_i steps.

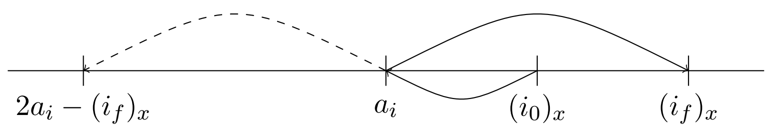

FIGURE 6.5. Any path that starts from $(i_0)_x$ that hits a_i on its way to $(i_f)_x$ is in one-to-one correspondence with a path from $(i_0)_x$ to $2a_i - (i_f)_x$.

An analogous argument for b_i gives that

$$\mathbb{P}_{(i_0)_x}(\tau_{b_i} < H_i, X_{H_i} = (i_f)_x) = \mathbb{P}_{(i_0)_x}(X_{H_i} = 2b_i - (i_f)_x) \tag{6.108}$$

Observe that $\mathbb{P}_{(i_0)_x}(\tau_{b_i} > H_i, \tau_{a_i} > H_i, X_{H_i} = (i_f)_x)$ is the probability that we travel from $(i_0)_x$ to $(i_f)_x$ in H_i steps without leaving (a_i, b_i); i.e., what we want. With a union bound, it follows that we have

$$\begin{aligned} &\mathbb{P}_{(i_0)_x}(\tau_{b_i} > H_i, \tau_{a_i} > H_i, X_{H_i} = (i_f)_x) \\ &\geq \mathbb{P}_{(i_0)_x}(X_{H_i} = (i_f)_x) - \mathbb{P}_{(i_0)_x}(\tau_{a_i} < H_i, X_{H_i} = (i_f)_x) - \mathbb{P}_{(i_0)_x}(\tau_{b_i} < H_i, X_{H_i} = (i_f)_x) \\ &= \mathbb{P}_{(i_0)_x}(X_{H_i} = (i_f)_x) - \mathbb{P}_{(i_0)_x}(X_{H_i} = 2a_i - (i_f)_x) - \mathbb{P}_{(i_0)_x}(X_{H_i} = 2b_i - (i_f)_x) \end{aligned} \tag{6.109}$$

and now notice that all three terms can be estimated by using the local central limit theorem described earlier.

Let now perform some estimates using line (6.105). The number of steps n is H_i.

For $\mathbb{P}_{(i_0)_x}(X_{H_i} = (i_f)_x)$, we will want a lower bound. Thus, we consider the case where $(i_0)_x$ and $(i_f)_x$ are furthest apart, being on opposite sides of $\mathcal{L}_i$. It follows that the furthest the target position could possibly be is no more than $c\ell$, so we have that

$$\mathbb{P}_{(i_0)_x}(X_{H_i} = (i_f)_x) \geq p_{H_i}(c\ell) \tag{6.110}$$

For $\mathbb{P}_{(i_0)_x}(X_{H_i} = 2a_i - (i_f)_x)$, we will want an upper bound, so we consider instead the case where $(i_0)_x$ and $2a_i - (i_f)_x$ are closest to each other. Since $(i_0)_x$ is inside $\mathcal{B}_i$ and $2a_i - (i_f)_x$ is outside $\mathcal{F}_i$ (observe that $2a_i - (i_f)_x = a_i - ((i_f)_x - a_i)$), the closest they could be is no less than $\frac{d-c}{2}\ell$, so we have that

$$\mathbb{P}_{(i_0)_x}(X_{H_i} = 2a_i - (i_f)_x) \leq p_{H_i}\left(\frac{d-c}{2}\ell\right) \tag{6.111}$$

A similar argument holds for $\mathbb{P}_{(i_0)_x}(X_{H_i} = 2b_i - (i_f)_x)$.

Thus, we have that

$$\begin{aligned}
&\mathbb{P}_{(i_0)_x}(\tau_{b_i} > H_i, \tau_{a_i} > H_i, X_{H_i} = (i_f)_x) \\
&\geq \mathbb{P}_{(i_0)_x}(X_{H_i} = (i_f)_x) - \mathbb{P}_{(i_0)_x}(X_{H_i} = 2a_i - (i_f)_x) - \mathbb{P}_{(i_0)_x}(X_{H_i} = 2b_i - (i_f)_x) \\
&\geq p_{H_i}(c\ell) - 2p_{H_i}\left(\frac{d-c}{2}\ell\right) \\
&= p_{A\ell^2}(c\ell) - 2p_{A\ell^2}\left(\frac{d-c}{2}\ell\right).
\end{aligned} \tag{6.112}$$

where, since we have $H_i \in \left[\frac{\beta\ell^2}{4}, \frac{3\beta\ell^2}{4}\right]$, we write $H_i = A\ell^2$, where $A \in \left[\frac{\beta}{4}, \frac{3\beta}{4}\right]$.

Using the approximations from **[4]** discussed earlier at lines (6.105) and (6.106), we have that

$$\overline{p_{A\ell^2}}(c\ell) = \frac{1}{\ell\sqrt{A\pi}} e^{-\frac{c^2}{A}} \tag{6.113}$$

$$|p_{A\ell^2}(c\ell) - \overline{p_{A\ell^2}}(c\ell)| \leq \frac{\xi}{A^{3/2}\ell^3} \tag{6.114}$$

$$\overline{p_{A\ell^2}}\left(\frac{d-c}{2}\ell\right) = \frac{1}{\ell\sqrt{A\pi}}e^{-\frac{(d-c)^2}{4A}} \tag{6.115}$$

$$\left|p_{A\ell^2}\left(\frac{d-c}{2}\ell\right) - \overline{p_{A\ell^2}}\left(\frac{d-c}{2}\ell\right)\right| \le \frac{\xi}{A^{3/2}\ell^3} \tag{6.116}$$

With these approximations and their error bounds, we get that

(6.117)

$$\begin{aligned}
\mathbb{P}_{(i_0)_x}(\tau_{b_i+1} > H_i, \tau_{a_i-1} > H_i, X_{H_i} = (i_f)_x) &\ge p_{A\ell^2}(c\ell) - 2p_{A\ell^2}\left(\frac{d-c}{2}\ell\right) \\
&\ge \frac{1}{\ell\sqrt{A\pi}}e^{-\frac{c^2}{A}} - \frac{\xi}{A^{3/2}\ell^3} - 2\left(\frac{1}{\ell\sqrt{A\pi}}e^{-\frac{(d-c)^2}{4A}} + \frac{\xi}{A^{3/2}\ell^3}\right) \\
&= \frac{1}{\ell\sqrt{A\pi}}e^{-\frac{c^2}{A}} - \frac{2}{\ell\sqrt{A\pi}}e^{-\frac{(d-c)^2}{4A}} - \frac{3\xi}{A^{3/2}\ell^3}
\end{aligned}$$

Since $A \in \left[\frac{\beta}{4}, \frac{3\beta}{4}\right]$, we can lower bound the previous expression:

$$\begin{aligned}
&\frac{1}{\ell\sqrt{A\pi}}e^{-\frac{c^2}{A}} - \frac{2}{\ell\sqrt{A\pi}}e^{-\frac{(d-c)^2}{4A}} - \frac{3\xi}{A^{3/2}\ell^3} \\
&\ge \frac{1}{\ell\sqrt{\left(\frac{3\beta}{4}\right)\pi}}e^{-\frac{c^2}{\left(\frac{\beta}{4}\right)}} - \frac{2}{\ell\sqrt{\left(\frac{\beta}{4}\right)\pi}}e^{-\frac{(d-c)^2}{4\left(\frac{3\beta}{4}\right)}} - \frac{3\xi}{\left(\frac{\beta}{4}\right)^{3/2}\ell^3} \\
&= \frac{1}{\ell\sqrt{\left(\frac{\beta}{4}\right)\pi}}\left(\frac{1}{\sqrt{3}}e^{-\frac{c^2}{\left(\frac{\beta}{4}\right)}} - 2e^{-\frac{(d-c)^2}{4\left(\frac{3\beta}{4}\right)}}\right) - \frac{3\xi}{\left(\frac{\beta}{4}\right)^{3/2}\ell^3} \\
&= \frac{2}{\ell\sqrt{\beta\pi}}\left(\frac{1}{\sqrt{3}}e^{-\frac{4c^2}{\beta}} - 2e^{-\frac{(d-c)^2}{3\beta}}\right) - \frac{24\xi}{\beta^{3/2}\ell^3}
\end{aligned} \tag{6.118}$$

Recall that $\frac{1}{3} > d > c > 0$. In particular, if $d - c \ge \alpha c$ for an $\alpha > 0$ to be determined, to then we find that the first term above can be bounded in the following way:

$$\frac{2}{\ell\sqrt{\beta\pi}}\left(\frac{1}{\sqrt{3}}e^{-\frac{4c^2}{\beta}} - 2e^{-\frac{(d-c)^2}{3\beta}}\right) - \frac{24\xi}{\beta^{3/2}\ell^3} \ge \frac{2}{\ell\sqrt{\beta\pi}}\left(\frac{1}{\sqrt{3}}e^{-\frac{4c^2}{\beta}} - 2e^{-\frac{(\alpha c)^2}{3\beta}}\right) - \frac{24\xi}{\beta^{3/2}\ell^3} \tag{6.119}$$

To take an estimate, we will want a β such that $\frac{1}{2\sqrt{3}}e^{-\frac{4c^2}{\beta}} \ge 2e^{-\frac{(\alpha c)^2}{3\beta}}$. A calculation shows that we will want

$$\beta \le \frac{c^2}{\ln(4\sqrt{3})}\left(\frac{\alpha^2}{3} - 4\right), \tag{6.120}$$

from which we also see that we will need $\alpha > 2\sqrt{3}$. For such α, β, c, and d, we will have that

$$\begin{aligned}\frac{2}{\ell\sqrt{\beta\pi}}\left(\frac{1}{\sqrt{3}}e^{-\frac{4c^2}{\beta}} - 2e^{-\frac{(\alpha c)^2}{3\beta}}\right) - \frac{24\xi}{\beta^{3/2}\ell^3} &\geq \frac{2}{\ell\sqrt{\beta\pi}}\left(\frac{1}{2\sqrt{3}}e^{-\frac{4c^2}{\beta}}\right) - \frac{24\xi}{\beta^{3/2}\ell^3} \\ &= \frac{B}{\ell} - \frac{24\xi}{\beta^{3/2}\ell^3}\end{aligned} \tag{6.121}$$

for a constant $B > 0$.

Finally, if ℓ is sufficiently large then we will have $\frac{24\xi}{\beta^{3/2}\ell^3} \leq \frac{B}{2\ell}$, so we can further estimate this by

$$\frac{B}{\ell} - \frac{24\xi}{\beta^{3/2}\ell^3} \geq \frac{B}{\ell} - \frac{B}{2\ell} = \frac{B}{2\ell} = \frac{C}{\ell} \tag{6.122}$$

for a constant $C > 0$.

We can thus conclude that

$$\mathbb{P}_{(i_0)_x}(\tau_{b_i} > H_i, \tau_{a_i} > H_i, X_{H_i} = (i_f)_x) \geq \frac{C}{\ell} \tag{6.123}$$

and consequently conclude that

$$\mathbb{P}(G_{H_i}|N = n) \geq \frac{C}{\ell} \tag{6.124}$$

and the same holds for G_{H_i}, G_{V_i}, G_{H_j}, G_{V_j}, G_{H_k}, and G_{V_k}. By line (6.103) we have that

$$\mathbb{P}(G_{H_i}, G_{V_i}, G_{H_j}, G_{V_j}, G_{H_k}, G_{V_k}|N = n) \geq \frac{C^6}{\ell^6} \tag{6.125}$$

so with this and lines (6.96) and (6.102) we have that

$$\begin{aligned}\mathbb{P}(X_R(i) = i_f, X_R(j) = j_f, X_R(k) = k_f) &\geq \sum_{n\in\mathcal{A}} \mathbb{P}(G_{H_i}, G_{V_i}, G_{H_j}, G_{V_j}, G_{H_k}, G_{V_k}|N = n)\mathbb{P}(N = n) \\ &\geq \sum_{n\in\mathcal{A}} \frac{C^6}{\ell^6}\mathbb{P}(N = n) \\ &= \frac{C^6}{\ell^6}\sum_{n\in\mathcal{A}} \mathbb{P}(N = n) \\ &= \frac{C^6}{\ell^6}\mathbb{P}(A) \\ &\geq \frac{C^6}{\ell^6}\cdot\left(1 - \frac{384}{\beta\ell^2}\right)\end{aligned} \tag{6.126}$$

and if ℓ is sufficiently large we know that

$$\frac{C^6}{\ell^6} \cdot \left(1 - \frac{384}{\beta\ell^2}\right) \geq \frac{E}{\ell^6} \tag{6.127}$$

for some constant $E > 0$.

6.3. Stage 3 Analysis and Total Result

Stage 3, where the tiles start in their respective boxes and travel to their destinations i', j', and k', is treated in the same way as Stage 1: Notice that for any particular sequence of moves in the Diaconis shuffle, going backwards, i.e., the sequence of moves that undoes what the first sequence did step by step, has the same probability as the first sequence. That is to say, for any sequence of moves that sends i', j', and k' to $\mathcal{B}_i$, $\mathcal{B}_j$, and $\mathcal{B}_k$ respectively, reversing it yields a sequence of moves that sends a tile from somewhere in $\mathcal{B}_i$, $\mathcal{B}_j$, and $\mathcal{B}_k$ to i', j', and k', respectively. It thus follows that the bound obtained for Stage 1 applies for Stage 3.

In total, since the moves of the Diaconis shuffle in different stages are independent, we conclude that the probability of having all three stages be successful in sequence, and hence sending (i, j, k) to (i', j', k') in T' moves, where T is the least even integer greater than $(2 + \beta + 2)\ell^2 n$, is at least

$$\frac{\eta^6}{2} \cdot \frac{E}{\ell^6} \cdot \frac{\eta^6}{2} = \frac{D}{\ell^6} \tag{6.128}$$

for a constant $D > 0$, as we claimed. □

We note that the universal constant C in the statement of the lemma is at least 4 from the proof.

We also remark that the T' can also be the least odd integer greater than $(2 + \beta + 2)\ell^2 n$ since our argument for Stage 2 allowed for R to be odd as well as even, and hence T' can also simply be the least integer greater than $(2 + \beta + 2)\ell^2 n$.

It then follows that lemma is in fact true for all ℓ: For each ℓ that is not sufficiently large, observe that there is always some nonzero probability that a triplet of tiles travels to another triplet of tiles in T' moves, where T' is the smallest even (or odd) integer greater than $(2 + \beta + 2)\ell^2 n$: by moving each tile one at a time, a triplet can be moved to another triplet in no more than 6ℓ moves (each tile needs at most ℓ vertical and ℓ horizontal moves); since we may assume $\ell \geq 2$ as there is nothing to say if $\ell = 1$, we will have that $6\ell < 4\ell^2 n < T'$, upon which the remaining moves can just be lazy.

Hence, the probability is at least $\frac{A}{\ell^6}$ for some A. Since there are only finitely many such insufficiently large ℓ, it follows that taking the minimum of all the A's obtained from these ℓ and the D from the lemma gives a constant $\widehat{D}$ that works for all ℓ.

CHAPTER 7

Finding a Mixing Time Bound

As noted before, the Diaconis shuffle introduces the possibility of a 3-collision when a row move is followed by a column move or vice-versa. As such, we can write the 2-step lazy Diaconis shuffle in 3-Monte form in the following way, by generating it from single steps of the lazy Diaconis shuffle:

Let π_1, π_2 be independent identically distributed copies of the Diaconis shuffle, and let $c_{12} = \pi_2^{-1}\pi_1^{-1}\pi_2\pi_1$. Observe that $\pi_1\pi_2c_{12} = \pi_2\pi_1$. Let $\overline{c_{12}}$ be an even mixture of the identity and c_{12}. Let c_{21} and $\overline{c_{21}}$ be defined similarly. Observe that if π_1 and π_2 consist of one row move and one column move, then c_{12} is a 3-cycle (as is c_{21}), and in that case, $\overline{c_{12}}$ is a 3-collision.

Then define the 2-step lazy shuffle v as follows:

$$v = \begin{cases} \pi_1\pi_2\overline{c_{12}}, & \text{if } \pi_1 \text{ is a row move and } \pi_2 \text{ is a column move} \\ \pi_2\pi_1\overline{c_{21}}, & \text{if } \pi_1 \text{ is a column move and } \pi_2 \text{ is a row move} \\ \pi_1\pi_2, & \text{otherwise} \end{cases} \tag{7.1}$$

Note that in the cases where π_1 and π_2 consist of one row move and one column move, then v assumes the row move came first and then introduces the corresponding collision to allow for the case where the column move came first, and hence has the same distribution as 2 moves of the Diaconis shuffle, as desired.

Let $v_{(t)}$ denote a product of t i.i.d. copies of the 2-step lazy Diaconis shuffle on the $n \times n$ grid. We claim the following:

LEMMA 7.0.1. *There is a universal constant C such that for any random permutation μ there is a value of $t \in \{1, \ldots, Cn^3\}$ such that*

$$\mathbb{E}(\mathrm{ENT}(\mu v_{(t)} |\, \mathrm{sgn}(\mu v_{(t)}))) \leq (1 - f(t))\mathbb{E}(\mathrm{ENT}(\mu |\, \mathrm{sgn}(\mu))) \tag{7.2}$$

where $f(t) = \frac{\gamma}{\log^2 n}\left(\frac{t}{n^2} \wedge 1\right)$, for a universal constant γ.

PROOF. Let $m = \lceil \log_2(n) \rceil$. Let $\ell_0 = 0$, then for $1 \le k \le m+1$, let $\ell_k = 2^{k-1} \wedge n$. Partition the set of locations $\{1, \dots, n^2\}$ into $m+1$ parts as follows: for $1 \le k \le m+1$, let $I_k = \{(\ell_{k-1})^2+1, \dots, \ell_k^2\}$.

Let μ be a random permutation, and consider $\mathbb{E}(\mathrm{ENT}(\mu | \operatorname{sgn}(\mu)))$. Recall the following decomposition:

$$\mathbb{E}(\mathrm{ENT}(\mu | \operatorname{sgn}(\mu))) = \sum_{j=1}^{n^2} \mathbb{E}(\mathrm{ENT}(\mu^{-1}(j) | \operatorname{sgn}(\mu), \mathcal{T}^{\mu}_{j+1})). \tag{7.3}$$

Let $E_j = \mathbb{E}(\mathrm{ENT}(\mu^{-1}(j) | \operatorname{sgn}(\mu), \mathcal{T}^{\mu}_{j+1}))$. We can rewrite the previous decomposition as follows:

$$\mathbb{E}(\mathrm{ENT}(\mu | \operatorname{sgn}(\mu))) = \sum_{k=1}^{m+1} \sum_{j \in I_k} E_j. \tag{7.4}$$

Consequently, if k^* maximizes $\sum_{j \in I_k} E_j$, then

$$\sum_{j \in I_{k^*}} E_j \ge \frac{1}{m+1} \mathbb{E}(\mathrm{ENT}(\mu | \operatorname{sgn}(\mu))). \tag{7.5}$$

We will use Theorem 4.0.1 to get a decay of entropy:

LEMMA 7.0.2. *Fix $1 \le k \le m+1$. There are universal constants $R, C > 0$ and $\psi > 0$ such that if t is the least even integer greater than $Rn\ell_k^2$ and T is such that $2T$ is the least even integer greater than $(Rn\ell_k^2 \wedge Cn^3)$, then*

$$A_x = \begin{cases} \psi(\frac{t}{n^2} \wedge 1) & \textit{if } x \in I_k; \\ 0 & \textit{otherwise}, \end{cases} \tag{7.6}$$

then the assumptions of Theorem 4.0.1 are satisfied by the t, T, and A_x.

PROOF. Let $x \in I_k$. We want that

$$\mathbb{P}(m_2(x) = x, m_1(x) \le x) \ge \frac{A_x}{x} \tag{7.7}$$

for $z < x$.

It will be enough to show the following:

LEMMA 7.0.3. *Let $z < x$, $y < x$ be distinct. Then, x, y, z collide in that order between time T and time t, with that collision being the first after time T that each of x, y, z are involved in, with probability at least $\frac{b_x}{\ell_k^4}$, where $b_x \geq d \cdot A_x$ for some positive constant $d > 0$.*

PROOF. First recall that since we are labeling the grid as in Figure 5.1, $z < x$ and $y < x$ implies that z, y, and x are all in the square box of side length ℓ_k whose upper left tile is the tile 1, which is also the upper left tile in the $n \times n$ grid. Let B_{ℓ_k} denote this box. Note that B_{ℓ_k} consists of the tiles $\{1, \ldots, \ell_k^2\}$ and is in fact the union of all the partitions up to I_{ℓ_k}.

We describe a way that x, y, z can collide in the manner required, which we refer to as "nicely." After the initial T steps, (x, y, z) will have travelled to some (x', y', z'). Then after s additional steps (where $0 \leq s \leq T-1$) with no collisions of x, y, or z, they are in an L-shape Γ where x is the middle tile. Then, they collide at Γ in the next step with probability $\frac{1}{32n^2}$. Considering all possible choices of (x', y', z'), s, and Γ, we have that

$$\mathbb{P}(x, y, z \text{ collide nicely}) \geq \sum_{x', y', z'} \sum_{s} \sum_{\Gamma} P^T((x, y, z), (x', y', z')) \cdot \tilde{P}^s((x', y', z'), \Gamma) \cdot \frac{a}{n^2} \tag{7.8}$$

where $P^T((x, y, z), (x', y', z'))$ is the probability that (x, y, z) moves to (x', y', z') in T steps of the 2-step shuffle, $\tilde{P}^s((x', y', z'), \Gamma)$ is the probability that (x', y', z') moves to the L-shape Γ in s steps of the 2-step shuffle, without involving any collisions for x', y', or z' (note the use of $\tilde{P}$ for the collision-free probability), and $\frac{a}{n^2}$ is the probability that the collision actually happens at Γ (where $a = \frac{1}{32}$).

We then take more lower bounds. By pulling the $\frac{a}{n^2}$ term to the front and considering just the Γ that are contained in B_{ℓ_k}, we get that

$$\mathbb{P}(x, y, z \text{ collide nicely}) \geq \frac{a}{n^2} \sum_{x', y', z'} \sum_{s} \sum_{\Gamma \in B_{\ell_k}} P^T((x, y, z), (x', y', z')) \cdot \tilde{P}^s((x', y', z'), \Gamma). \tag{7.9}$$

We limit ourselves to considering x', y', z' that each have supremum distance at most $\alpha \ell$ from B_{ℓ_k}, where $\alpha \geq 1$ is an integer constant to be determined. Note that if α is sufficiently large, then the set of such tiles is the entire $n \times n$ grid, but otherwise it will be the box with side length $(2\alpha + 1)\ell_k$ that is centered around B_{ℓ_k}. Thus, let $\widehat{\ell_k} = ((2\alpha + 1)\ell_k) \wedge n$ and call this set $B_{\widehat{\ell_k}}$.

Recall that although B_{ℓ_k} is at the top left corner of the $n \times n$ grid, due to how the Diaconis shuffle works we are actually on a torus, so $B_{\widehat{\ell_k}}$ will include tiles from the opposite corner. Since we can simply shift the viewpoint of the torus around, we can visualize $B_{\widehat{\ell_k}}$ as a whole box rather than in pieces.

It follows that we have

$$\mathbb{P}(x, y, z \text{ collide nicely}) \geq \frac{a}{n^2} \sum_{x',y',z' \in B_{\widehat{\ell_k}}} \sum_{s} \sum_{\Gamma \in B_{\ell_k}} P^T((x,y,z),(x',y',z')) \cdot \tilde{P}^s((x',y',z'),\Gamma). \tag{7.10}$$

Observe that T steps of the 2-step lazy Diaconis shuffle is $2T$ steps of the lazy Diaconis shuffle. Thus, as a result of the lemma proved in the previous section, if $2T$ is the least even integer greater than $Cn\widehat{\ell_k}^2$ (so T is the least integer greater than $\frac{Cn\widehat{\ell_k}^2}{2}$) where C is the universal constant from that lemma and thus have t be the least even integer greater than $Cn((2\alpha+1)\ell_k)^2$ to ensure that $2T \leq t$ so that we have enough steps total, then we have that $P^T((x,y,z),(x',y',z')) \geq \frac{D}{\widehat{\ell_k}^6}$ for a universal constant D; in this case we get that

$$\mathbb{P}(x, y, z \text{ collide nicely}) \geq \frac{a}{n^2} \cdot \frac{D}{\widehat{\ell_k}^6} \sum_{x',y',z' \in B_{\widehat{\ell_k}}} \sum_{s} \sum_{\Gamma \in B_{\ell_k}} \tilde{P}^s((x',y',z'),\Gamma). \tag{7.11}$$

Recall that $0 \leq s \leq T-1$. We recall that the universal constant from the lemma is at least 4, and observe that $\ell_k \leq \widehat{\ell_k}$. It follows that $n\ell_k^2 \leq \frac{Cn\widehat{\ell_k}^2}{2} - 1 < T$. Note that $n\ell_k^2$ is an integer, so $n\ell_k^2 \leq T-1$.

Considering only the cases where $1 \leq s \leq n\ell_k^2$, we obtain by rearranging the sums that

$$\mathbb{P}(x, y, z \text{ collide nicely}) \geq \frac{a}{n^2} \cdot \frac{D}{\widehat{\ell_k}^6} \sum_{\Gamma \in B_{\ell_k}} \sum_{s=1}^{n\ell_k^2} \sum_{x',y',z' \in B_{\widehat{\ell_k}}} \tilde{P}^s((x',y',z'),\Gamma). \tag{7.12}$$

We now claim that

$$\sum_{s=1}^{n\ell_k^2} \sum_{x',y',z' \in B_{\widehat{\ell_k}}} \tilde{P}^s((x',y',z'),\Gamma) \geq \frac{n\ell_k^2 \wedge \frac{8n^2}{36}}{2}. \tag{7.13}$$

First, observe that the 2-step chain is time reversible. We can define a time reversal of the 2-step chain that has collisions at the same times as the regular 2-step chain. It follows that the probability of going from (x',y',z') to Γ without any collisions is the same as the probability of going from Γ to

(x', y', z') without any collisions, so we have that

$$\sum_{s=1}^{n\ell_k^2} \sum_{x',y',z' \in B_{\widehat{\ell_k}}} \tilde{P}^s((x',y',z'),\Gamma) = \sum_{s=1}^{n\ell_k^2} \sum_{x',y',z' \in B_{\widehat{\ell_k}}} \tilde{P}^s(\Gamma,(x',y',z')). \tag{7.14}$$

Since the inner sum is over all $x', y', z' \in B_{\widehat{\ell_k}}$, the inner sum is the probability that, starting from Γ, each of the tiles travels to somewhere inside $B_{\widehat{\ell_k}}$ without having a collision for x', y', or z' in s steps, which we write as $\tilde{\mathbb{P}}_\Gamma(X_s \in B_{\widehat{\ell_k}})$, where X_s is the position of the three tiles after s steps:

$$\sum_{s=1}^{n\ell_k^2} \sum_{(x',y',z') \in B_{\widehat{\ell_k}}} \tilde{P}^s(\Gamma,(x',y',z')) = \sum_{s=1}^{n\ell_k^2} \tilde{\mathbb{P}}_\Gamma(X_s \in B_{\widehat{\ell_k}}). \tag{7.15}$$

Let τ represent the amount of time after time T of the first collision involving any of x', y', or z'. Then the event that there is no collision in the first s steps after time T is really the event that $\tau > s$, so we have that

$$\sum_{s=1}^{n\ell_k^2} \tilde{\mathbb{P}}_\Gamma(X_s \in B_{\widehat{\ell_k}}) = \sum_{s=1}^{n\ell_k^2} \mathbb{P}_\Gamma(X_s \in B_{\widehat{\ell_k}}, \tau > s). \tag{7.16}$$

where $\mathbb{P}_\Gamma(X_s \in B_{\widehat{\ell_k}}, \tau > s)$ is the probability that, starting from Γ, each of the tiles travels to somewhere inside $B_{\widehat{\ell_k}}$, and that $\tau > s$.

Now we claim the following:

If $s \leq n\ell_k^2 \wedge \frac{8n^2}{36}$, then $\mathbb{P}_\Gamma(X_s \in B_{\widehat{\ell_k}}, \tau > s) \geq \frac{1}{2}$.

To see this, observe the complements of the individual events. Consider $\mathbb{P}_\Gamma(X_s \notin B_{\widehat{\ell_k}})$. Since X_s denotes the positions of the three tiles, for X_s to not be in $B_{\widehat{\ell_k}}$ means that at least one tile has left $B_{\widehat{\ell_k}}$. We note that if $\widehat{\ell_k} = n$, then evidently such an event is impossible.

Otherwise, $\widehat{\ell_k} = (2\alpha + 1)\ell_k$. Recall that individually, each tile performs a lazy random walk, and hence the probability that a single tile leaves $B_{\widehat{\ell_k}}$ can be estimated using Chebyshev's inequality.

Let ρ be one of the tiles in Γ, say the middle one. Let $X_s^{(\rho,h)}$ represent the horizontal position of the tile, relative to its starting point in Γ, at time s.

One way that ρ might leave $B_{\widehat{\ell_k}}$ is that it travels too far horizonally; since $\Gamma \in B_{\ell_k}$, this would mean that at the very least $X_s^{(\rho,h)}$ travelled more than $\alpha\ell_k$ away from its starting point.

So, we consider $\mathbb{P}\left(\left|X_s^{(\rho,h)}\right| \geq \alpha\ell_k\right)$. Observe that the variance of a single step (of the lazy random walk arising from the lazy Diaconis shuffle) is $\frac{1}{4n}$. Since $s \leq n\ell_k^2$ and s is the number of steps of the 2-step lazy Diaconis shuffle, there are at most $2n\ell_k^2$ steps taken of this lazy random talk in total, so the total variance after s steps is at most $\frac{\ell_k^2}{2}$. It follows that

$$\mathbb{P}\left(\left|X_s^{(x',h)}\right| \geq \alpha\ell_k\right) \leq \frac{\ell_k^2}{2} \cdot \frac{1}{\alpha^2\ell_k^2} = \frac{1}{2\alpha^2}. \tag{7.17}$$

The same calculation applies for the vertical position as well as for the other two tiles; since there are six coordinates total, a union bound gives that the probability that any of them go too far is at most $\frac{6}{2\alpha^2}$. If α is sufficiently large, this is at most $\frac{1}{4}$, and so we find in that case that

$$\mathbb{P}_\Gamma(X_s \notin B_{\widehat{\ell_k}}) \leq \frac{1}{4}. \tag{7.18}$$

Now consider $\mathbb{P}_\Gamma(\tau \leq s)$. Observe that at every step, the probability of a collision involving i' is $\frac{3}{8n^2}$, since the probability that a row-column or column-row combination occurs is $\frac{1}{8}$ and in such a case 3 out of n^2 tiles are in the resulting collision. The same holds true for j' and k', so by taking a union bound the probability that any of them have a collision in a single step is at most $\frac{9}{8n^2}$. The event that $\tau \leq s$ is contained in the union of the events that any of i', j', k' have a collision at time r, for $1 \leq r \leq s$. By a union bound it follows that

$$\mathbb{P}_\Gamma(\tau \leq s) \leq \sum_{r=1}^{s} \frac{9}{8n^2} = s \cdot \frac{9}{8n^2}. \tag{7.19}$$

Since $s \leq \frac{8n^2}{36}$, it follows that $\mathbb{P}_\Gamma(\tau \leq s) \leq \frac{1}{4}$.

Putting the two complement events together and using a union bound, it follows that if $s \leq n\ell_k^2 \wedge \frac{8n^2}{36}$, then $\mathbb{P}_\Gamma(X_s \in B_{\widehat{\ell_k}}, \tau > s) \geq \frac{1}{2}$.

Hence, by limiting the sum to $n\ell_k^2 \wedge \frac{8n^2}{36}$, we get

$$\begin{aligned}\sum_{s=1}^{T} \mathbb{P}_\Gamma(X_s \in B_{\widehat{\ell_k}}, \tau > s) &\geq \sum_{s=1}^{n\ell_k^2 \wedge \frac{8n^2}{36}} \mathbb{P}_\Gamma(X_s \in B_{\widehat{\ell_k}}, \tau > s) \\ &\geq \sum_{s=1}^{n\ell_k^2 \wedge \frac{8n^2}{36}} \frac{1}{2} \\ &\geq \frac{n\ell_k^2 \wedge \frac{8n^2}{36}}{2}.\end{aligned} \tag{7.20}$$

This in turn implies that

$$\mathbb{P}(x, y, z \text{ collide nicely}) \geq \frac{a}{n^2} \cdot \frac{D}{\widehat{\ell_k}^6} \sum_{\Gamma \in B_{\ell_k}} \frac{n\ell_k^2 \wedge \frac{8n^2}{36}}{2}. \tag{7.21}$$

Since there are at least ℓ_k^2 ways to pick an L-shape Γ in B_{ℓ_k}, it follows that

$$\mathbb{P}(x, y, z \text{ collide nicely}) \geq \frac{a}{n^2} \cdot \frac{D}{\widehat{\ell_k}^6} \cdot \ell_k^2 \cdot \left(\frac{n\ell_k^2 \wedge \frac{8n^2}{36}}{2} \right). \tag{7.22}$$

since $\widehat{\ell_k} \leq (2\alpha+1)\ell_k$ we have that

$$\mathbb{P}(x, y, z \text{ collide nicely}) \geq \frac{a}{n^2} \cdot \frac{D}{(2\alpha+1)^6 \ell_k^6} \cdot \ell_k^2 \cdot \left(\frac{n\ell_k^2 \wedge \frac{8n^2}{36}}{2} \right). \tag{7.23}$$

Then, since t is the least even integer greater than $Cn((2\alpha+1)\ell_k)^2$, we observe that since $C > 4$ and $\alpha \geq 1$, we have $2Cn((2\alpha+1)\ell_k)^2 > t$. Therefore,

$$\begin{aligned}\mathbb{P}(x, y, z \text{ collide nicely}) &\geq \frac{a}{n^2} \cdot \frac{D}{(2\alpha+1)^6 \ell_k^6} \cdot \ell_k^2 \cdot \frac{1}{2C(2\alpha+1)^2} \left(\frac{(2C(2\alpha+1)^2) n\ell_k^2 \wedge \frac{(2C(2\alpha+1)^2)8n^2}{36}}{2} \right) \\ &\geq \frac{a}{n^2} \cdot \frac{D}{(2\alpha+1)^6 \ell_k^6} \cdot \ell_k^2 \cdot \frac{1}{2C(2\alpha+1)^2} \left(\frac{t \wedge \frac{(2C(2\alpha+1)^2)8n^2}{36}}{2} \right).\end{aligned} \tag{7.24}$$

A further reduction gets us that

$$\mathbb{P}(x, y, z \text{ collide nicely}) \geq \frac{A}{\ell_k^4} \left(\frac{t}{n^2} \wedge 1 \right). \tag{7.25}$$

where A is a universal constant obtained from the constants in the previous line. This shows the lemma. $\square$

This is enough: Given a $z < x$, there are at least $\frac{\ell_k^2}{8}$ choices for y, while x is at least $\frac{\ell_k^2}{4}$, so we have that

(7.26)
$$\mathbb{P}(m_2(x) = x, m_1(x) \leq x) \geq \sum_y \mathbb{P}(x, y, z \text{ collide nicely}) \geq \frac{\ell_k^2}{8} \frac{A'}{\ell_k^4} \left(\frac{t}{n^2} \wedge 1 \right) \geq \frac{\psi}{x} \left(\frac{t}{n^2} \wedge 1 \right) = \frac{A_x}{x}$$

where ψ is a constant obtained from combining the constants in the previous term.

$\square$

This shows that Theorem 4.0.1 is applicable. As a result, we get that

$$\mathbb{E}(\text{ENT}(\mu\pi_{(t)} | \operatorname{sgn}(\mu\pi_{(t)}))) - \mathbb{E}(\text{ENT}(\mu | \operatorname{sgn}(\mu))) \leq \frac{-C}{\log(n^2)} \sum_{x=1}^{n^2} A_x E_x \tag{7.27}$$

where we recall $E_x = \mathbb{E}(\text{ENT}(\mu^{-1}(x) | \mathcal{T}_{x+1}^{\mu}, \operatorname{sgn}(\mu)))$ and C is a universal constant. By choosing k to be k^*, the previous result gives us that

$$\mathbb{E}(\text{ENT}(\mu\pi_{(t)} | \operatorname{sgn}(\mu\pi_{(t)}))) - \mathbb{E}(\text{ENT}(\mu | \operatorname{sgn}(\mu))) \leq \frac{-C}{\log(n^2)} \sum_{x \in I_{k^*}} \psi \left(\frac{t}{n^2} \wedge 1 \right) E_x \tag{7.28}$$

for the t determined by the previous result. By how we defined k^*, we have that

$$\begin{aligned} &\mathbb{E}(\text{ENT}(\mu\pi_{(t)} | \operatorname{sgn}(\mu\pi_{(t)}))) - \mathbb{E}(\text{ENT}(\mu | \operatorname{sgn}(\mu))) \\ &\leq \frac{-C}{\log(n^2)} \cdot \psi \left(\frac{t}{n^2} \wedge 1 \right) \cdot \frac{1}{m+1} \mathbb{E}(\text{ENT}(\mu | \operatorname{sgn}(\mu))). \end{aligned} \tag{7.29}$$

Recall that $m = \lceil \log_2 n \rceil$; it follows that there is a constant C' such that

$$\mathbb{E}(\text{ENT}(\mu\pi_{(t)} | \operatorname{sgn}(\mu\pi_{(t)}))) - \mathbb{E}(\text{ENT}(\mu | \operatorname{sgn}(\mu))) \leq \frac{-C'}{\log^2 n} \cdot \left(\frac{t}{n^2} \wedge 1 \right) \mathbb{E}(\text{ENT}(\mu | \operatorname{sgn}(\mu))). \tag{7.30}$$

Consequently, we obtain that

$$\mathbb{E}(\text{ENT}(\mu\pi_{(t)} | \operatorname{sgn}(\mu\pi_{(t)}))) \leq (1 - f(t)) \mathbb{E}(\text{ENT}(\mu | \operatorname{sgn}(\mu))) \tag{7.31}$$

where $f(t) = \frac{\gamma}{\log^2 n} \left(\frac{t}{n^2} \wedge 1 \right)$ for some constant γ, which is what we wanted to show. $\square$

Finally, we can obtain a mixing time bound in a similar manner to [**6**]:

LEMMA 7.0.4. *The mixing time for the lazy Diaconis shuffle is $O(n^3 \log^3 n)$.*

PROOF. If t and f are as before, then we have that

$$\frac{t}{f(t)} = \gamma^{-1}(\log^2 n)t\left(\frac{n^2}{t} \vee 1\right) = \gamma^{-1}(\log^2 n)\left(n^2 \vee t\right) \leq \gamma^{-1}(\log^2 n)\left(n^2 \vee cn^3\right) =: R \tag{7.32}$$

The previous lemma implies that there is a $t_1 \in \{1, \ldots, Cn^3\}$ such that

$$\mathbb{E}(\mathrm{ENT}(\pi_1 \cdots \pi_{t_1} | \operatorname{sgn}(\pi_1 \cdots \pi_{t_1}))) \leq (1 - f(t_1))\mathbb{E}(\mathrm{ENT}(\mathrm{id} \,|\, \operatorname{sgn}(\mathrm{id}))) \tag{7.33}$$

as well as a $t_2 \in \{1, \ldots, Cn^3\}$ such that

$$\mathbb{E}(\mathrm{ENT}(\pi_1 \cdots \pi_{t_1+t_2} | \operatorname{sgn}(\pi_1 \cdots \pi_{t_1+t_2}))) \leq (1 - f(t_2))\mathbb{E}(\mathrm{ENT}(\pi_1 \cdots \pi_{t_1} | \operatorname{sgn}(\pi_1 \cdots \pi_{t_1}))) \tag{7.34}$$

and so on, defining $t_3, t_4, \ldots$ recursively.

For $j \geq 1$, let $\tau_j = \sum_{i=1}^j t_i$. Then we get from the above that

$$\begin{aligned}\mathbb{E}(\mathrm{ENT}(\pi_{(\tau_j)} | \operatorname{sgn}(\pi_{(\tau_j)}))) &\leq \left[\prod_{i=1}^{j}(1 - f(t_i))\right] \mathbb{E}(\mathrm{ENT}(\mathrm{id} \,|\, \operatorname{sgn}(\mathrm{id}))) \\ &\leq \exp\left(-\sum_{i=1}^{j} f(t_j)\right) \mathbb{E}(\mathrm{ENT}(\mathrm{id} \,|\, \operatorname{sgn}(\mathrm{id})))\end{aligned} \tag{7.35}$$

Since $t_i \leq Rf(t_i)$ by equation 7.32, we have that

$$\tau_j = \sum_{i=1}^{j} t_i \leq R \sum_{i=1}^{j} f(t_i), \tag{7.36}$$

so it follows that

$$\mathbb{E}(\mathrm{ENT}(\pi_{(\tau_j)} | \operatorname{sgn}(\pi_{(\tau_j)}))) \leq \exp\left(\frac{-\tau_j}{R}\right) \mathbb{E}(\mathrm{ENT}(\mathrm{id} \,|\, \operatorname{sgn}(\mathrm{id}))) \tag{7.37}$$

Since $\mathbb{E}(\mathrm{ENT}(\mathrm{id} \,|\, \operatorname{sgn}(\mathrm{id}))) = \log\left(\frac{n^2!}{2}\right) \leq \log(n^2!) \leq n^2 \log n^2$, it follows that if $\tau_j \geq R \log(8n^2 \log n^2)$ then we find that $\mathbb{E}(\mathrm{ENT}(\pi_{(\tau_j)} | \operatorname{sgn}(\pi_{(\tau_j)}))) \leq \frac{1}{8}$. Since $R = \gamma^{-1} \log^2 n(n^2 \vee cn^3)$, this corresponds to $\tau_j \geq \gamma^{-1} \log^2 n(n^2 \vee cn^3) \log(8n^2 \log n^2)$. For sufficiently large n, the cn^3 dominates in $(n^2 \vee cn^3)$, and hence this corresponds to τ_j being $O(n^3 \log^3 n)$.

Recall that by equation 3.2 it is enough to show that $\mathrm{ENT}(\pi_{(\tau_j)}) \le \frac{1}{8}$ for well-mixing. Recall that we showed the decomposition

$$\mathrm{ENT}(\pi_{(\tau_j)}) = \mathbb{E}(\mathrm{ENT}(\pi_{(\tau_j)} | \operatorname{sgn}(\pi_{(\tau_j)}))) + \mathrm{ENT}(\operatorname{sgn}(\pi_{(\tau_j)})) \tag{7.38}$$

If n is even, the Diaconis shuffle gives an odd permutation each step. Hence, the lazy shuffle will have uniformly randomized the sign after a single step, so $\mathrm{ENT}(\operatorname{sgn}(\pi_{(\tau_j)})) = 0$ so long as $\tau_j \ge 1$. Hence after $O(n^3 \log^3 n)$ steps the lazy Diaconis shuffle well-mixes over all permutations.

On the other hand, if n is odd, then the Diaconis shuffle gives an even permutation each step. This means that the sign cannot change and hence the state space is only the even permutations. In this case, the sign is deterministic and $\mathrm{ENT}(\operatorname{sgn}(\pi_{(\tau_j)})) = 0$. Thus, after $O(n^3 \log^3 n)$ steps, the lazy Diaconis shuffle well-mixes over all even permutations; i.e., over the (smaller) state space.

We conclude that the mixing time is indeed $O(n^3 \log^3 n)$.

□

APPENDIX A

Pinsker's Inequality

This appendix contains a treatment of Pinsker's inequality, following the text in [**2**] with additional details by the author for increased clarity.

Suppose V is a finite set of size n. Let π and μ be two distributions on V. We recall that the total variation distance between π and μ is half the $\mathcal{L}^1$ distance between them:

$$\|\pi - \mu\|_{\mathrm{TV}} = \frac{1}{2}\|\pi - \mu\|_1 = \frac{1}{2}\sum_{i\in V}|\pi(i) - \mu(i)| \tag{A.1}$$

Also recall that the relative entropy between π and μ is

$$D(\pi||\mu) = \sum_{i\in V}\pi(i)\ln\frac{\pi(i)}{\mu(i)} \tag{A.2}$$

with the convention that $a\log\frac{a}{0} = \infty$ if $a > 0$, $0\log\frac{0}{b} = 0$ if $b > 0$, and $0\log\frac{0}{0} = 0$.

Relating them is the result known as *Pinsker's inequality*. In [**6**], the inequality is written as

$$\sqrt{\frac{1}{2}D(\pi||\mu)} \geq \|\pi - \mu\|_{\mathrm{TV}}\ . \tag{A.3}$$

In [**2**], the inequality is written as

$$\widehat{D}(\pi||\mu) \geq \frac{1}{2\ln 2}\|\pi - \mu\|_1^2 \tag{A.4}$$

where

$$\widehat{D}(\pi||\mu) = \sum_{i\in V}\pi(i)\log_2\frac{\pi(i)}{\mu(i)} \tag{A.5}$$

is the relative entropy defined using the base 2 logarithm.

Since [**2**] uses the base 2 logarithm, there is a seeming difference in the constants used. However, the two statements of Pinsker's inequality are in fact equivalent: We observe that

$$\widehat{D}(\pi||\mu) = \sum_{i\in V} \pi(i) \log_2 \frac{\pi(i)}{\mu(i)} = \sum_{i\in V} \pi(i) \frac{\ln \frac{\pi(i)}{\mu(i)}}{\ln 2} = \frac{1}{\ln 2} \sum_{i\in V} \pi(i) \ln \frac{\pi(i)}{\mu(i)} = \frac{1}{\ln 2} D(\pi||\mu) \tag{A.6}$$

and consequently if equation (A.4) holds, then we see that

$$\frac{1}{\ln 2} D(\pi||\mu) = \widehat{D}(\pi||\mu) \geq \frac{1}{2\ln 2} \|\pi - \mu\|_1^2 = \frac{1}{2\ln 2} (2\|\pi - \mu\|_{\mathrm{TV}})^2 = \frac{2}{\ln 2} \|\pi - \mu\|_{\mathrm{TV}}^2 \tag{A.7}$$

and thus

$$D(\pi||\mu) \geq 2\|\pi - \mu\|_{\mathrm{TV}}^2 \tag{A.8}$$

which in turn is equivalent to equation (A.3) since total variation distance and relative entropy are non-negative (as seen in [**2**]).

As for the reverse direction, if equation (A.3), and hence equation (A.8), is true, then rearranging the terms in equation (A.7) gives

$$\widehat{D}(\pi||\mu) = \frac{1}{\ln 2} D(\pi||\mu) \geq \frac{2}{\ln 2} \|\pi - \mu\|_{\mathrm{TV}}^2 = \frac{1}{2\ln 2} (2\|\pi - \mu\|_{\mathrm{TV}})^2 = \frac{1}{2\ln 2} \|\pi - \mu\|_1^2 \; ; \tag{A.9}$$

that is to say, the result in equation (A.4).

We shall now prove Pinsker's inequality.

First note that if π and μ are point distributions, then we see that $D(\pi||\mu) = \|\pi - \mu\|_{\mathrm{TV}} = 0$ if $\pi = \mu$, and otherwise $D(\pi||\mu) = \infty$ and $\|\pi - \mu\|_{\mathrm{TV}} = 1$ if $\pi \neq \mu$, verifying that the inequality holds in those cases.

Next, we prove the other cases of Pinsker's inequality, following the method in [**2**], by recalling the binary case shown there and then providing details for the general case.

Let us first suppose that π and μ are binary distributions with parameters p and q; that is to say, $\pi(1) = p$ and $\pi(0) = 1 - p$, while $\mu(1) = q$ and $\mu(0) = 1 - q$. We have that $0 < p, q < 1$.

First, we see that

$$\begin{aligned}
\frac{1}{2\ln 2}\left\|\pi-\mu\right\|_1^2 &= \frac{1}{2\ln 2}\left(\sum_{i\in V}\left|\pi(i)-\mu(i)\right|\right)^2 = \frac{1}{2\ln 2}(|p-q|+|(1-p)-(1-q)|)^2 \\
&= \frac{1}{2\ln 2}(|p-q|+|-p+q|)^2 = \frac{1}{2\ln 2}(|p-q|+|-1|\,|p-q|)^2 \\
&= \frac{1}{2\ln 2}(2\,|p-q|)^2 = \frac{4}{2\ln 2}(p-q)^2\,,
\end{aligned} \tag{A.10}$$

and that

$$\widehat{D}(\pi||\mu) = \sum_{i\in V}\pi(i)\log_2\frac{\pi(i)}{\mu(i)} = p\log_2\frac{p}{q} + (1-p)\log_2\frac{(1-p)}{(1-q)}\,. \tag{A.11}$$

So, proving equation (A.4) will mean proving that

$$p\log_2\frac{p}{q} + (1-p)\log_2\frac{(1-p)}{(1-q)} \geq \frac{4}{2\ln 2}(p-q)^2 \tag{A.12}$$

or that

$$p\log_2\frac{p}{q} + (1-p)\log_2\frac{(1-p)}{(1-q)} - \frac{4}{2\ln 2}(p-q)^2 \geq 0\,. \tag{A.13}$$

Let $g(p,q)$ be that quantity on the left above:

$$g(p,q) = p\log_2\frac{p}{q} + (1-p)\log_2\frac{(1-p)}{(1-q)} - \frac{4}{2\ln 2}(p-q)^2 \tag{A.14}$$

Then by taking the derivative of g with respect to q, we have that

$$\begin{aligned}
\frac{\partial g(p,q)}{\partial q} &= p\frac{1}{\ln 2}\frac{1}{\frac{p}{q}}\frac{(-p)}{q^2} + (1-p)\frac{1}{\ln 2}\frac{1}{\frac{1-p}{1-q}}\frac{-(1-p)(-1)}{(1-q)^2} - \frac{4}{2\ln 2}2(p-q)(-1) \\
&= \frac{-p}{q\ln 2} + \frac{1-p}{(1-q)\ln 2} - \frac{4}{\ln 2}(q-p) \\
&= \frac{(-p)(1-q)+q(1-p)}{q(1-q)\ln 2} - \frac{4}{\ln 2}(q-p) \\
&= \frac{q-p}{q(1-q)\ln 2} - \frac{4}{\ln 2}(q-p) \\
&= \left(\frac{1}{q(1-q)} - 4\right)\frac{q-p}{\ln 2}
\end{aligned} \tag{A.15}$$

Since $0<q<1$, $q(1-q) = q-q^2 \leq \frac{1}{4}$, so $\frac{1}{q(1-q)} \geq 4$ and hence $\left(\frac{1}{q(1-q)} - 4\right) \geq 0$.

Observe that if $q = p$, then

$$g(p,p) = p\log_2 \frac{p}{p} + (1-p)\log_2 \frac{(1-p)}{(1-p)} - \frac{4}{2\ln 2}(p-p)^2 = 0 + 0 - 0 = 0 \tag{A.16}$$

If $q \geq p$, then $\frac{q-p}{\ln 2} \geq 0$, so $\frac{\partial g(p,q)}{\partial q} = \left(\frac{1}{q(1-q)} - 4\right)\frac{q-p}{\ln 2} \geq 0$. This means that as q increases away from p, $g(p,q)$ increases away from 0; i.e., if $q \geq p$, then $g(p,q) \geq 0$.

Likewise, if $q \leq p$ then $\frac{q-p}{\ln 2} \leq 0$, so $\frac{\partial g(p,q)}{\partial q} = \left(\frac{1}{q(1-q)} - 4\right)\frac{q-p}{\ln 2} \leq 0$. This means that as q increases toward p, $g(p,q)$ must be decreasing toward 0. Thus if $q \leq p$, then $g(p,q) \geq 0$.

Hence, $g(p,q) \geq 0$ generally, which is what we wanted to prove.

Thus Pinsker's inequality holds in the case of binary distributions.

Now for the general case. Let π and μ be any two distributions on V. Let A be the set of $x \in V$ for which π has a higher probability than μ:

$$A = \{x \in V : \pi(x) > \mu(x)\} \tag{A.17}$$

Define $Y = I_A$, the indicator of A, and let $\tilde{\pi}$ and $\tilde{\mu}$ be the resulting distributions of A. Specifically, $\tilde{\pi}$ and $\tilde{\mu}$ are binary distributions with

$$\tilde{\pi}(1) = \sum_{x:\pi(x)>\mu(x)} \pi(x), \qquad \tilde{\mu}(1) = \sum_{x:\pi(x)>\mu(x)} \mu(x), \tag{A.18}$$

$$\tilde{\pi}(0) = \sum_{x:\pi(x)\leq\mu(x)} \pi(x), \qquad \tilde{\mu}(0) = \sum_{x:\pi(x)\leq\mu(x)} \mu(x), \tag{A.19}$$

and in particular $\tilde{\pi}(1) = \pi(A)$ and $\tilde{\mu}(1) = \mu(A)$.

At this point, **[2]** says to use the data-processing inequality for relative entropies to obtain that

$$\begin{aligned} \widehat{D}(\pi||\mu) &\geq \widehat{D}(\tilde{\pi}||\tilde{\mu}) \\ &\geq \frac{4}{2\ln 2}(\pi(A) - \mu(A))^2 \\ &= \frac{1}{2\ln 2}\left\|\pi - \mu\right\|_1^2 , \end{aligned} \tag{A.20}$$

which concludes the proof of Pinsker's inequality for the general case.

However, let us take a closer look at that final argument:

The second inequality is exactly Pinsker's inequality for the binary case, noting that $\pi(A) = \tilde{\pi}(1)$ and $\mu(A) = \tilde{\mu}(1)$ and that this is exactly the expression in equation (A.10).

The final equality we recognize as an equivalent definition of total variation distance.

Thus, the data-processing inequality for relative entropy refers is the first inequality. **[2]** states that it is proved in the same way as the data-processing inequality for mutual information is proved.

However, **[2]** proves the data-processing inequality for mutual information via the chain rule for mutual information, and the way to develop an analogous use of the chain rule for relative entropy to prove the data-processing inequality for relative entropy is not clear.

So instead, we shall just prove the result directly:

Lemma A.0.1 (Data-processing inequality for relative entropy). *Let π and μ be two distributions on X and suppose $Y = f(X)$, with $\tilde{\pi}$ and $\tilde{\mu}$ being the resulting distributions on Y, so that $\tilde{\pi}(y) = \sum_{x \in X : f(x) = y} \pi(x)$ and $\tilde{\mu}(y) = \sum_{x \in X : f(x) = y} \mu(x)$ Then*

$$\widehat{D}(\pi || \mu) \geq \widehat{D}(\tilde{\pi} || \tilde{\mu}). \tag{A.21}$$

Proof. First write out $\widehat{D}(\pi || \mu)$ according to the definition, then rewrite the sum to be over Y:

$$\widehat{D}(\pi || \mu) = \sum_{x \in X} \pi(x) \log_2 \frac{\pi(x)}{\mu(x)} = \sum_{y \in Y} \sum_{x \in X : f(x) = y} \pi(x) \log_2 \frac{\pi(x)}{\mu(x)} \,. \tag{A.22}$$

Then use the log sum inequality, which we shall prove shortly, for each $y \in Y$:

$$\begin{aligned} \sum_{y \in Y} \sum_{x \in X, f(x) = y} \pi(x) \log_2 \frac{\pi(x)}{\mu(x)} &\geq \sum_{y \in Y} \left[\left(\sum_{x \in X : f(x) = y} \pi(x) \right) \log_2 \frac{\sum_{x \in X : f(x) = y} \pi(x)}{\sum_{x \in X : f(x) = y} \mu(x)} \right] \\ &= \sum_{y \in Y} \tilde{\pi}(y) \log_2 \frac{\tilde{\pi}(y)}{\tilde{\mu}(y)} \\ &= \widehat{D}(\tilde{\pi} || \tilde{\mu}) \,. \end{aligned} \tag{A.23}$$

Thus, $\widehat{D}(\pi || \mu) \geq \widehat{D}(\tilde{\pi} || \tilde{\mu})$ as desired.

$\square$

The log sum inequality is proved in [**2**]. We recall the proof here as a lemma, while also adding some extra details:

LEMMA A.0.2 (Log sum inequality). *For nonnegative numbers $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots b_n$,*

$$\sum_{i=1}^{n} a_i \log_2 \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i\right) \log_2 \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}. \tag{A.24}$$

As [**2**] states, we use the convention that $a \log \frac{a}{0} = \infty$ if $a > 0$, $0 \log \frac{0}{b} = 0$ if $b > 0$, and $0 \log \frac{0}{0} = 0$. [**2**] also claims that "these follow easily from continuity," which is the case for the first two but is not clear for the third one.

Indeed, $f(x, y) = x \log_2 \frac{x}{y}$ does not have a limit as $x \to 0$, $y \to 0$:

If $x = y$, then $x \log_2 \frac{x}{y} = x \log_2 1 = 0 \to 0$ as $x \to 0$.

However, if $y = 2^{-\frac{1}{x}}$, then

$$\begin{aligned} x \log_2 \frac{x}{y} &= x \log_2 x - x \log_2(2^{-\frac{1}{x}}) \\ &= x \log_2 x - x \left(-\frac{1}{x}\right) \\ &= x \log_2 x + 1 \to 1 \end{aligned} \tag{A.25}$$

so the limits are not the same.

Thus, we do not understand what is meant by continuity here. Instead, we appeal to a different notion presented by [**2**]: We recall that relative entropy $D(p||q)$, which also uses this expression, is stated to be "a measure of the inefficiency of assuming that the distribution is q when the true distribution is p." If both p and q assign probability 0 to a given outcome x, then intuitively there should be no resulting inefficiency for that outcome.

Additionally, we remark that the analogous statement of the log sum inequality with the natural logarithm is true, since we obtain it by multiplying both sides by $\ln 2$.

PROOF OF LOG SUM INEQUALITY. [**2**] starts by stating that without loss of generality that $a_i > 0$ and $b_i > 0$. We object to this way of presenting it and will cover the additional cases at the end.

So, suppose that $a_i > 0$ and $b_i > 0$. Consider $f(t) = t\log_2 t$. Then $f'(t) = \log_2 t + \frac{1}{\ln 2}$ and $f''(t) = \frac{1}{t\ln 2} > 0$ for positive t, so $f(t)$ is strictly convex. Jensen's inequality then implies that if $\alpha_i \geq 0$ with $\sum_{i=1}^n \alpha_i = 1$ and $t_i > 0$,

$$\sum_{i=1}^n \alpha_i f(t_i) \geq f\left(\sum_{i=1}^n \alpha_i t_i\right) \tag{A.26}$$

(i.e., $(\mathbb{E}(f(t)) \geq f(\mathbb{E}(t)))$. With $f(t) = t\log_2 t$, this is

$$\sum_{i=1}^n \alpha_i t_i \log_2 t_i \geq \left(\sum_{i=1}^n \alpha_i t_i\right) \log_2\left(\sum_{i=1}^n \alpha_i t_i\right) \tag{A.27}$$

Let $\alpha_i = \frac{b_i}{\sum_{j=1}^n b_j}$ and $t_i = \frac{a_i}{b_i}$. Then substituting them into the result from Jensen's inequality gives

$$\sum_{i=1}^n \frac{b_i}{\sum_{j=1}^n b_j} \cdot \frac{a_i}{b_i} \log_2 \frac{a_i}{b_i} \geq \left(\sum_{i=1}^n \frac{b_i}{\sum_{j=1}^n b_j} \cdot \frac{a_i}{b_i}\right) \log_2\left(\sum_{i=1}^n \frac{b_i}{\sum_{j=1}^n b_j} \cdot \frac{a_i}{b_i}\right) \tag{A.28}$$

canceling out the b_i's gives

$$\sum_{i=1}^n \frac{a_i}{\sum_{j=1}^n b_j} \log_2 \frac{a_i}{b_i} \geq \left(\sum_{i=1}^n \frac{a_i}{\sum_{j=1}^n b_j}\right) \log_2\left(\sum_{i=1}^n \frac{a_i}{\sum_{j=1}^n b_j}\right) \tag{A.29}$$

Observe that $\frac{1}{\sum_{j=1}^n b_j}$ can be pulled out of each sum over i, giving

$$\frac{1}{\sum_{j=1}^n b_j} \sum_{i=1}^n a_i \log_2 \frac{a_i}{b_i} \geq \left(\frac{1}{\sum_{j=1}^n b_j} \sum_{i=1}^n a_i\right) \log_2\left(\frac{1}{\sum_{j=1}^n b_j} \sum_{i=1}^n a_i\right) \tag{A.30}$$

Multiply both sides by $\sum_{j=1}^n b_j$ to obtain

$$\sum_{i=1}^n a_i \log_2 \frac{a_i}{b_i} \geq \left(\sum_{i=1}^n a_i\right) \log_2\left(\frac{1}{\sum_{j=1}^n b_j} \sum_{i=1}^n a_i\right) \tag{A.31}$$

and reindex the remaining sum of $b_j's$ to be over the i's to get

$$\sum_{i=1}^n a_i \log_2 \frac{a_i}{b_i} \geq \left(\sum_{i=1}^n a_i\right) \log_2\left(\frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i}\right) \tag{A.32}$$

which is the log sum inequality.

Now, suppose that in our nonnegative numbers $a_1, a_2, \ldots a_n$ and $b_1, b_2, \ldots b_n$, there is an index k for which $a_k = 0$ or $b_k = 0$. We first remark that the inequality is true if we take the subset of the indices where both a and b are positive, as we just showed. We show that the adding back in the indices k where one or both of a_k and b_k are 0 one by one does not change the truth of the inequality.

Concretely, suppose it is already true that

$$\sum_{i=1}^{k-1} a_i \log_2 \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{k-1} a_i\right) \log_2 \frac{\sum_{i=1}^{k-1} a_i}{\sum_{i=1}^{k-1} b_i} \tag{A.33}$$

We see what happens if we "add in" a_k, b_k to both sides:

If $a_k = 0$, $b_k = 0$, then using our convention that $0 \log_2 \frac{0}{0} = 0$, we see that nothing is added to the left side, and since $\sum_{i=1}^{k-1} a_i = \sum_{i=1}^{k} a_i$ and $\sum_{i=1}^{k-1} b_i = \sum_{i=1}^{k} b_i$, the right side remains the same. Thus:

$$\begin{aligned}
\sum_{i=1}^{k} a_i \log_2 \frac{a_i}{b_i} &= \sum_{i=1}^{k-1} a_i \log_2 \frac{a_i}{b_i} \\
&\geq \left(\sum_{i=1}^{k-1} a_i\right) \log_2 \frac{\sum_{i=1}^{k-1} a_i}{\sum_{i=1}^{k-1} b_i} \\
&= \left(\sum_{i=1}^{k} a_i\right) \log_2 \frac{\sum_{i=1}^{k} a_i}{\sum_{i=1}^{k} b_i}
\end{aligned} \tag{A.34}$$

If $a_i = 0$, $b_i > 0$, then nothing is added to the left side ($0 \log \frac{0}{b_k} = 0$), while on the right side, the sum of a_i's remains the same and the quantity inside the logarithm is smaller. Since $\log_2 x$ is an increasing function, the logarithm must be smaller, too. So, we get that

$$\begin{aligned}
\sum_{i=1}^{k} a_i \log_2 \frac{a_i}{b_i} &= \sum_{i=1}^{k-1} a_i \log_2 \frac{a_i}{b_i} \\
&\geq \left(\sum_{i=1}^{k-1} a_i\right) \log_2 \frac{\sum_{i=1}^{k-1} a_i}{\sum_{i=1}^{k-1} b_i} \\
&= \left(\sum_{i=1}^{k} a_i\right) \log_2 \frac{\sum_{i=1}^{k} a_i}{\sum_{i=1}^{k-1} b_i} \\
&\geq \left(\sum_{i=1}^{k} a_i\right) \log_2 \frac{\sum_{i=1}^{k} a_i}{\sum_{i=1}^{k} b_i}
\end{aligned} \tag{A.35}$$

Lastly, if $a_k > 0$ while $b_k = 0$, then since $a_k \log_2 \frac{a_k}{0} = \infty$, the left side becomes ∞ and the statement becomes trivial:

$$\infty = \sum_{i=1}^{k} a_i \log_2 \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{k} a_i \right) \log_2 \frac{\sum_{i=1}^{k} a_i}{\sum_{i=1}^{k} b_i} \tag{A.36}$$

This concludes the proof of the log sum inequality. $\square$

To recap, the log sum inequality was used to prove the data-processing inequality for relative entropies, which in turn was used to prove the general case of Pinsker's inequality, which also used the binary case in its proof.

As a final remark, recall that while relative entropy is not symmetric ($D(\pi||\mu) \neq D(\mu||\pi)$, generally), total variation distance is, so we have that by Pinsker's inequality

$$D(\mu||\pi) \geq 2 \left\| \mu - \pi \right\|_{\mathrm{TV}}^2 = 2 \left\| \pi - \mu \right\|_{\mathrm{TV}}^2 \tag{A.37}$$

i.e., both $D(\pi||\mu)$ and $D(\mu||\pi)$ serve as upper bounds for $2 \left\| \pi - \mu \right\|_{\mathrm{TV}}^2$.

APPENDIX B

Entropy Decompositions

We cover the decomposition of the entropy of a random permutation when given the sign of the permutation, and also when given the sign of the permutation and the tails of the permutation. Much of the calculation is the same as before, just with the addition of the sign condition, but as it is lengthy, we go over it once again.

First, we consider what happens when conditioning on just the sign of the permutation. When $n = 1$, there is no information gained by the sign (since it is always 1), so the decomposition is only meaningful when $n \geq 2$, which we will assume here.

As mentioned previously, we have that

$$\text{(B.1)} \qquad \mathrm{ENT}(\pi|\operatorname{sgn}(\pi) = 1) = \sum_{\nu:\operatorname{sgn}(\nu)=1} \mathbb{P}(\pi = \nu|\operatorname{sgn}(\pi) = 1) \log\left(\frac{n!}{2}\mathbb{P}(\pi = \nu|\operatorname{sgn}(\pi) = 1)\right)$$

as well as

$$\text{(B.2)} \quad \mathrm{ENT}(\pi|\operatorname{sgn}(\pi) = -1) = \sum_{\nu:\operatorname{sgn}(\nu)=-1} \mathbb{P}(\pi = \nu|\operatorname{sgn}(\pi) = -1) \log\left(\frac{n!}{2}\mathbb{P}(\pi = \nu|\operatorname{sgn}(\pi) = -1)\right)$$

and

(B.3)

$$\begin{aligned}\mathrm{ENT}(\operatorname{sgn}(\pi)) &= \sum_{s=-1,1} \mathbb{P}(\operatorname{sgn}(\pi) = s) \log\left(2\mathbb{P}(\operatorname{sgn}(\pi) = s)\right) \\ &= \mathbb{P}(\operatorname{sgn}(\pi) = 1) \log\left(2\mathbb{P}(\operatorname{sgn}(\pi) = 1)\right) + \mathbb{P}(\operatorname{sgn}(\pi) = -1) \log\left(2\mathbb{P}(\operatorname{sgn}(\pi) = -1)\right)\end{aligned}$$

We claimed that

$$\text{(B.4)} \qquad \mathrm{ENT}(\pi) = \mathbb{E}\left(\mathrm{ENT}(\pi|\operatorname{sgn}(\pi))\right) + \mathrm{ENT}(\operatorname{sgn}(\pi))$$

so let us see how this happens.

As usual, we start with the definition:

$$\mathrm{ENT}(\pi) = \sum_{\nu \in S_n} \mathbb{P}(\pi = \nu) \log(n!\mathbb{P}(\pi = \nu)) \tag{B.5}$$

and this time we decompose the probability via conditioning on the sign:

$$\mathbb{P}(\pi = \nu) = \mathbb{P}(\pi = \nu | \operatorname{sgn}(\pi) = \operatorname{sgn}(\nu)) \cdot \mathbb{P}(\operatorname{sgn}(\pi) = \operatorname{sgn}(\nu)) \tag{B.6}$$

By replacing the instance inside the logarithm and splitting the logarithm again, we get

$$\begin{aligned} \mathrm{ENT}(\pi) &= \sum_{\nu \in S_n} \mathbb{P}(\pi = \nu) \log\left(\frac{n!}{2}\mathbb{P}(\pi = \nu | \operatorname{sgn}(\pi) = \operatorname{sgn}(\nu)) \cdot 2\mathbb{P}(\operatorname{sgn}(\pi) = \operatorname{sgn}(\nu))\right) \\ &= \sum_{\nu \in S_n} \mathbb{P}(\pi = \nu) \left(\log\left(\frac{n!}{2}\mathbb{P}(\pi = \nu | \operatorname{sgn}(\pi) = \operatorname{sgn}(\nu))\right) + \log\left(2\mathbb{P}(\operatorname{sgn}(\pi) = \operatorname{sgn}(\nu))\right)\right) \end{aligned} \tag{B.7}$$

Distributing, we get

$$\begin{aligned} \mathrm{ENT}(\pi) = &\sum_{\nu \in S_n} \Bigg(\mathbb{P}(\pi = \nu) \log\left(\frac{n!}{2}\mathbb{P}(\pi = \nu | \operatorname{sgn}(\pi) = \operatorname{sgn}(\nu))\right) \\ &\qquad + \mathbb{P}(\pi = \nu) \log\left(2\mathbb{P}(\operatorname{sgn}(\pi) = \operatorname{sgn}(\nu))\right)\Bigg) \\ = &\sum_{\nu \in S_n} \left(\mathbb{P}(\pi = \nu) \log\left(\frac{n!}{2}\mathbb{P}(\pi = \nu | \operatorname{sgn}(\pi) = \operatorname{sgn}(\nu))\right)\right) \\ &+ \sum_{\nu \in S_n} \left(\mathbb{P}(\pi = \nu) \log\left(2\mathbb{P}(\operatorname{sgn}(\pi) = \operatorname{sgn}(\nu))\right)\right) \end{aligned} \tag{B.8}$$

Reorder the first sum by grouping together the permutations that have the same sign:

$$\begin{aligned} &\sum_{\nu \in S_n} \left(\mathbb{P}(\pi = \nu) \log\left(\frac{n!}{2}\mathbb{P}(\pi = \nu | \operatorname{sgn}(\pi) = \operatorname{sgn}(\nu))\right)\right) \\ &= \sum_{s=-1,1} \sum_{\substack{\nu \in S_n \\ \operatorname{sgn}(\nu) = s}} \left(\mathbb{P}(\pi = \nu) \log\left(\frac{n!}{2}\mathbb{P}(\pi = \nu | \operatorname{sgn}(\pi) = \operatorname{sgn}(\nu))\right)\right) \end{aligned} \tag{B.9}$$

Split the probability as before, then in the inner sum, replace $\operatorname{sgn}(\nu)$ with s:

$$\begin{aligned} &= \sum_{s=-1,1} \sum_{\substack{\nu\in S_n \\ \operatorname{sgn}(\nu)=s}} \Bigg(\mathbb{P}(\pi = \nu | \operatorname{sgn}(\pi) = \operatorname{sgn}(\nu)) \cdot \mathbb{P}(\operatorname{sgn}(\pi) = \operatorname{sgn}(\nu)) \\ &\qquad \cdot \log\left(\frac{n!}{2}\mathbb{P}(\pi = \nu | \operatorname{sgn}(\pi) = \operatorname{sgn}(\nu))\right)\Bigg) \\ &= \sum_{s=-1,1} \sum_{\substack{\nu\in S_n \\ \operatorname{sgn}(\nu)=s}} \left(\mathbb{P}(\pi = \nu | \operatorname{sgn}(\pi) = s)\mathbb{P}(\operatorname{sgn}(\pi) = s) \log\left(\frac{n!}{2}\mathbb{P}(\pi = \nu | \operatorname{sgn}(\pi) = s)\right)\right) \end{aligned} \tag{B.10}$$

Pull $\mathbb{P}(\operatorname{sgn}(\pi) = s)$ out of the inner sum since it is independent of ν, then see that the inner sum is now the entropy given the sign, and thus the whole expression for the first sum is the expected value of the entropy given the sign:

$$\begin{aligned} &= \sum_{s=-1,1} \mathbb{P}(\operatorname{sgn}(\pi) = s) \sum_{\substack{\nu\in S_n \\ \operatorname{sgn}(\nu)=s}} \left(\mathbb{P}(\pi = \nu | \operatorname{sgn}(\pi) = s) \log\left(\frac{n!}{2}\mathbb{P}(\pi = \nu | \operatorname{sgn}(\pi) = s)\right)\right) \\ &= \sum_{s=-1,1} \mathbb{P}(\operatorname{sgn}(\pi) = s) \operatorname{ENT}(\pi | \operatorname{sgn}(\pi) = s) \\ &= \mathbb{E}(\operatorname{ENT}(\pi | \operatorname{sgn}(\pi)) \end{aligned} \tag{B.11}$$

As for the second sum, do the same: reorder the sum by sign:

$$\sum_{\nu\in S_n} \left(\mathbb{P}(\pi = \nu) \log\left(2\mathbb{P}(\operatorname{sgn}(\pi) = \operatorname{sgn}(\nu))\right)\right) = \sum_{s=-1,1} \sum_{\substack{\nu\in S_n \\ \operatorname{sgn}(\nu)=s}} \left(\mathbb{P}(\pi = \nu) \log\left(2\mathbb{P}(\operatorname{sgn}(\pi) = \operatorname{sgn}(\nu))\right)\right) \tag{B.12}$$

Then split the probability:

$$= \sum_{s=-1,1} \sum_{\substack{\nu\in S_n \\ \operatorname{sgn}(\nu)=s}} \left(\mathbb{P}(\pi = \nu | \operatorname{sgn}(\pi) = \operatorname{sgn}(\nu))\mathbb{P}(\operatorname{sgn}(\pi) = \operatorname{sgn}(\nu)) \log\left(2\mathbb{P}(\operatorname{sgn}(\pi) = \operatorname{sgn}(\nu))\right)\right) \tag{B.13}$$

Replace $\operatorname{sgn}(\nu)$ with s:

$$= \sum_{s=-1,1} \sum_{\substack{\nu\in S_n \\ \operatorname{sgn}(\nu)=s}} \left(\mathbb{P}(\pi = \nu | \operatorname{sgn}(\pi) = s)\mathbb{P}(\operatorname{sgn}(\pi) = s) \log\left(2\mathbb{P}(\operatorname{sgn}(\pi) = s)\right)\right) \tag{B.14}$$

Now pull $\mathbb{P}(\operatorname{sgn}(\pi) = s)\log(2\mathbb{P}(\operatorname{sgn}(\pi) = s))$ out of the inner sum because it is independent of ν, then see that the inner sum is then 1, leaving the entropy of the sign:

$$\begin{aligned}
&= \sum_{s=-1,1} \mathbb{P}(\operatorname{sgn}(\pi) = s)\log(2\mathbb{P}(\operatorname{sgn}(\pi) = s)) \sum_{\substack{\nu \in S_n \\ \operatorname{sgn}(\nu)=s}} (\mathbb{P}(\pi = \nu | \operatorname{sgn}(\pi) = s)) \\
&= \sum_{s=-1,1} \mathbb{P}(\operatorname{sgn}(\pi) = s)\log(2\mathbb{P}(\operatorname{sgn}(\pi) = s)) \\
&= \operatorname{ENT}(\operatorname{sgn}(\pi))
\end{aligned} \tag{B.15}$$

So, we get that

$$\operatorname{ENT}(\pi) = \mathbb{E}\left(\operatorname{ENT}(\pi | \operatorname{sgn}(\pi))\right) + \operatorname{ENT}(\operatorname{sgn}(\pi)) \tag{B.16}$$

as we claimed.

Next, recall that we also claimed that for any $1 \leq i \leq n$,

$$\operatorname{ENT}(\pi) = \mathbb{E}\left(\operatorname{ENT}(\pi | \operatorname{sgn}(\pi), \mathcal{T}_i^\pi)\right) + \sum_{k=i}^{n} \left(\mathbb{E}\left(\operatorname{ENT}(\pi^{-1}(k) | \operatorname{sgn}(\pi), \mathcal{T}_{k+1}^\pi)\right)\right) + \operatorname{ENT}(\operatorname{sgn}(\pi)); \tag{B.17}$$

that is to say, after splitting away the portion of entropy attributable to the sign, the remaining entropy can still be split up according to its entries, similar to how it was done before.

We are really claiming that

$$\mathbb{E}\left(\operatorname{ENT}(\pi | \operatorname{sgn}(\pi))\right) = \mathbb{E}\left(\operatorname{ENT}(\pi | \operatorname{sgn}(\pi), \mathcal{T}_i^\pi)\right) + \sum_{k=i}^{n} \left(\mathbb{E}\left(\operatorname{ENT}(\pi^{-1}(k) | \operatorname{sgn}(\pi), \mathcal{T}_{k+1}^\pi)\right)\right), \tag{B.18}$$

and the proof is similar to before. We point out that this decomposition is only really meaningful when $n \geq 3$, since after knowing the sign, knowing all entries after the second is equivalent to knowing all entries–there are only two numbers left to assign and the sign dictates in what order they must be assigned. This means that the constant in the logarithm is different from before. When considering the entropy of the kth entry given the tail from the $k+1$th entry onward and the sign, the number of possibilities will be k if $k \geq 3$; otherwise it will be 1. When considering the entropy of the permutation given the tail end from the ith entry onward and the sign, the number of possible permutations will be $\frac{(i-1)!}{2}$ if $i \geq 3$, otherwise it will be 1 (as the sign will indicate what it should be). Consequently, no entropy is attributable to the second or first entry in this decomposition, so their terms are 0.

Rewinding some of our previous steps, we recall that

$$\begin{aligned}&\mathbb{E}\left(\mathrm{ENT}(\pi|\,\mathrm{sgn}(\pi))\right)\\ &= \sum_{s=-1,1}\mathbb{P}(\mathrm{sgn}(\pi)=s)\sum_{\substack{\nu\in S_n\\ \mathrm{sgn}(\nu)=s}}\left(\mathbb{P}(\pi=\nu|\,\mathrm{sgn}(\pi)=s)\log\left(\frac{n!}{2}\mathbb{P}(\pi=\nu|\,\mathrm{sgn}(\pi)=s)\right)\right).\end{aligned}\tag{B.19}$$

We focus on looking at the inner sum over permutations ν with the correct sign, so consider just

$$\sum_{\substack{\nu\in S_n\\ \mathrm{sgn}(\nu)=s}}\left(\mathbb{P}(\pi=\nu|\,\mathrm{sgn}(\pi)=s)\log\left(\frac{n!}{2}\mathbb{P}(\pi=\nu|\,\mathrm{sgn}(\pi)=s)\right)\right)\tag{B.20}$$

for the time being.

Similar to how we proceeded in the case of $\mathrm{ENT}(\pi)$, we can decompose the conditional probability by further conditioning:

$$\begin{aligned}&\mathbb{P}(\pi=\nu|\,\mathrm{sgn}(\pi)=s)\\ &= \mathbb{P}(\pi=\nu|\mathcal{T}_i^\pi=\mathcal{T}_i^\nu,\mathrm{sgn}(\pi)=s)\cdot\prod_{k=i}^{n}\mathbb{P}(\pi^{-1}(k)=\nu^{-1}(k)|\mathcal{T}_{k+1}^\pi=\mathcal{T}_{k+1}^\nu,\mathrm{sgn}(\pi)=s).\end{aligned}\tag{B.21}$$

Replacing the term in the logarithm gets us

$$\begin{aligned}&\sum_{\substack{\nu\in S_n\\ \mathrm{sgn}(\nu)=s}}\Bigg(\mathbb{P}(\pi=\nu|\,\mathrm{sgn}(\pi)=s)\\ &\quad\cdot\log\left(\frac{n!}{2}\mathbb{P}(\pi=\nu|\mathcal{T}_i^\pi=\mathcal{T}_i^\nu,\mathrm{sgn}(\pi)=s)\cdot\prod_{k=i}^{n}\mathbb{P}(\pi^{-1}(k)=\nu^{-1}(k)|\mathcal{T}_{k+1}^\pi=\mathcal{T}_{k+1}^\nu,\mathrm{sgn}(\pi)=s)\right)\Bigg).\end{aligned}\tag{B.22}$$

We want to factor $\frac{n!}{2}$ according to our earlier discussion regarding the number of possibilities of π given the tail ends and the sign. We write

(B.23)

$$\sum_{\substack{\nu \in S_n \\ \operatorname{sgn}(\nu)=s}} \Bigg(\mathbb{P}(\pi = \nu | \operatorname{sgn}(\pi) = s)$$

$$\cdot \log \left(R_i \mathbb{P}(\pi = \nu | \mathcal{T}_i^\pi = \mathcal{T}_i^\nu, \operatorname{sgn}(\pi) = s) \cdot \prod_{k=i}^{n} S_k \mathbb{P}(\pi^{-1}(k) = \nu^{-1}(k) | \mathcal{T}_{k+1}^\pi = \mathcal{T}_{k+1}^\nu, \operatorname{sgn}(\pi) = s) \right) \Bigg)$$

where R_i is the number of possibilities of π given that $\mathcal{T}_i^\pi = \mathcal{T}_i^\nu$ and $\operatorname{sgn}(\pi) = s$ and S_k is the number of possibilities of $\pi^{-1}(k)$ given that $\mathcal{T}_{k+1}^\pi = \mathcal{T}_{k+1}^\nu$ and $\operatorname{sgn}(\pi) = s$.

As we noted, we have that

$$R_i = \begin{cases} 1 & \text{if } 1 \le i \le 2 \\ \frac{(i-1)!}{2} & \text{if } 3 \le i \le n \end{cases} \tag{B.24}$$

and

$$S_k = \begin{cases} 1 & \text{if } 1 \le k \le 2 \\ k & \text{if } 3 \le k \le n \end{cases} \tag{B.25}$$

and also see that we have

$$R_i \cdot \prod_{k=i}^{n} S_k = \frac{n!}{2} \tag{B.26}$$

so we did indeed factor $\frac{n!}{2}$.

Like before, we split the logarithm to obtain the following expression:

$$\begin{aligned} &\sum_{\substack{\nu \in S_n \\ \operatorname{sgn}(\nu)=s}} \Bigg(\mathbb{P}(\pi = \nu | \operatorname{sgn}(\pi) = s) \\ &\quad \cdot \Bigg(\log \left(R_i \mathbb{P}(\pi = \nu | \mathcal{T}_i^\pi = \mathcal{T}_i^\nu, \operatorname{sgn}(\pi) = s) \right) \\ &\qquad + \sum_{k=i}^{n} \log \left(S_k \mathbb{P}(\pi^{-1}(k) = \nu^{-1}(k) | \mathcal{T}_{k+1}^\pi = \mathcal{T}_{k+1}^\nu, \operatorname{sgn}(\pi) = s) \right) \Bigg) \Bigg). \end{aligned} \tag{B.27}$$

Distributing, we get

(B.28)
$$\sum_{\substack{\nu \in S_n \\ \operatorname{sgn}(\nu)=s}} \mathbb{P}(\pi = \nu | \operatorname{sgn}(\pi) = s) \log \left(R_i \mathbb{P}(\pi = \nu | \mathcal{T}_i^\pi = \mathcal{T}_i^\nu, \operatorname{sgn}(\pi) = s)\right)$$
$$+ \sum_{\substack{\nu \in S_n \\ \operatorname{sgn}(\nu)=s}} \mathbb{P}(\pi = \nu | \operatorname{sgn}(\pi) = s) \sum_{k=i}^{n} \log \left(S_k \mathbb{P}(\pi^{-1}(k) = \nu^{-1}(k) | \mathcal{T}_{k+1}^\pi = \mathcal{T}_{k+1}^\nu, \operatorname{sgn}(\pi) = s)\right).$$

Consider each sum separately in a similar manner to before:

The first sum will be again reordered by tails, by grouping together the permutations ν that share the same tail, while this time also having a particular sign. Let $\mathbb{T}_i$ again represent the collection of all possible tails that start from entry i, then we have that

$$\sum_{\substack{\nu \in S_n \\ \operatorname{sgn}(\nu)=s}} \mathbb{P}(\pi = \nu | \operatorname{sgn}(\pi) = s) \log \left(R_i \mathbb{P}(\pi = \nu | \mathcal{T}_i^\pi = \mathcal{T}_i^\nu, \operatorname{sgn}(\pi) = s)\right)$$
$$= \sum_{T \in \mathbb{T}_i} \sum_{\substack{\nu \in S_n: \\ \mathcal{T}_i^\nu = T, \\ \operatorname{sgn}(\nu)=s}} \mathbb{P}(\pi = \nu | \operatorname{sgn}(\pi) = s) \log \left(R_i \mathbb{P}(\pi = \nu | \mathcal{T}_i^\pi = \mathcal{T}_i^\nu, \operatorname{sgn}(\pi) = s)\right) \tag{B.29}$$

Next, recall that

$$\mathbb{P}(\pi = \nu | \operatorname{sgn}(\pi) = s) = \mathbb{P}(\pi = \nu | \mathcal{T}_i^\pi = \mathcal{T}_i^\nu, \operatorname{sgn}(\pi) = s) \cdot \mathbb{P}(\mathcal{T}_i^\pi = \mathcal{T}_i^\nu | \operatorname{sgn}(\pi) = s), \tag{B.30}$$

and insert this in to get

$$\sum_{T \in \mathbb{T}_i} \sum_{\substack{\nu \in S_n: \\ \mathcal{T}_i^\nu = T, \\ \operatorname{sgn}(\nu)=s}} (\mathbb{P}(\pi = \nu | \mathcal{T}_i^\pi = \mathcal{T}_i^\nu, \operatorname{sgn}(\pi) = s) \mathbb{P}(\mathcal{T}_i^\pi = \mathcal{T}_i^\nu | \operatorname{sgn}(\pi) = s)$$
$$\cdot \log \left(R_i \mathbb{P}(\pi = \nu | \mathcal{T}_i^\pi = \mathcal{T}_i^\nu, \operatorname{sgn}(\pi) = s)\right)) \tag{B.31}$$

Now, inside the inner sum, $\mathcal{T}_i^\nu$ is known to be T, so we replace it accordingly:

$$= \sum_{T \in \mathbb{T}_i} \sum_{\substack{\nu \in S_n: \\ \mathcal{T}_i^\nu = T, \\ \operatorname{sgn}(\nu)=s}} (\mathbb{P}(\pi = \nu | \mathcal{T}_i^\pi = T, \operatorname{sgn}(\pi) = s) \mathbb{P}(\mathcal{T}_i^\pi = T | \operatorname{sgn}(\pi) = s)$$
$$\cdot \log \left(R_i \mathbb{P}(\pi = \nu | \mathcal{T}_i^\pi = T, \operatorname{sgn}(\pi) = s)\right)) \tag{B.32}$$

and now we may pull $\mathbb{P}(\mathcal{T}_i^\pi = T|\operatorname{sgn}(\pi) = s)$ out of the inner sum since it is independent of ν:

$$= \sum_{T\in\mathbb{T}_i} \left(\mathbb{P}(\mathcal{T}_i^\pi = T|\operatorname{sgn}(\pi) = s) \right.$$

$$\left. \cdot \sum_{\substack{\nu\in S_n: \\ \mathcal{T}_i^\nu = T, \\ \operatorname{sgn}(\nu)=s}} \mathbb{P}(\pi = \nu|\mathcal{T}_i^\pi = T, \operatorname{sgn}(\pi) = s) \log\left(R_i \mathbb{P}(\pi = \nu|\mathcal{T}_i^\pi = T, \operatorname{sgn}(\pi) = s)\right) \right) \tag{B.33}$$

and now the inner sum is $\mathrm{ENT}(\pi|T_i^\pi = T, \operatorname{sgn}(\pi) = s)$, so we have

$$= \sum_{T\in\mathbb{T}_i} \mathbb{P}(\mathcal{T}_i^\pi = T|\operatorname{sgn}(\pi) = s)\,\mathrm{ENT}(\pi|T_i^\pi = T, \operatorname{sgn}(\pi) = s). \tag{B.34}$$

We leave this as is for the moment.

Now, for the second sum. First switch the order of summation as before:

$$\begin{aligned} &\sum_{\substack{\nu\in S_n: \\ \operatorname{sgn}(\nu)=s}} \mathbb{P}(\pi = \nu|\operatorname{sgn}(\pi) = s) \sum_{k=i}^{n} \log\left(S_k \mathbb{P}(\pi^{-1}(k) = \nu^{-1}(k)|\mathcal{T}_{k+1}^\pi = \mathcal{T}_{k+1}^\nu, \operatorname{sgn}(\pi) = s)\right) \\ &= \sum_{k=i}^{n} \sum_{\substack{\nu\in S_n: \\ \operatorname{sgn}(\nu)=s}} \mathbb{P}(\pi = \nu|\operatorname{sgn}(\pi) = s) \log\left(S_k \mathbb{P}(\pi^{-1}(k) = \nu^{-1}(k)|\mathcal{T}_{k+1}^\pi = \mathcal{T}_{k+1}^\nu, \operatorname{sgn}(\pi) = s)\right) \end{aligned} \tag{B.35}$$

Then reorder the sum by grouping together the ν that have the same tail from the kth entry onward, by first specifying the tail from the $k+1$th entry onward and then selecting a value for the kth entry, as we previously did:

$$= \sum_{k=i}^{n} \sum_{T\in\mathbb{T}_{k+1}} \sum_{\substack{j=1 \\ j\notin T}}^{n} \sum_{\substack{\nu\in S_n: \\ \nu^{-1}(k)=j, \\ \mathcal{T}_{k+1}^\nu = T, \\ \operatorname{sgn}(\nu)=s}} \mathbb{P}(\pi = \nu|\operatorname{sgn}(\pi) = s) \log\left(S_k \mathbb{P}(\pi^{-1}(k) = \nu^{-1}(k)|\mathcal{T}_{k+1}^\pi = \mathcal{T}_{k+1}^\nu, \operatorname{sgn}(\pi) = s)\right) \tag{B.36}$$

We break down $\mathbb{P}(\pi = \nu | \operatorname{sgn}(\pi) = s)$:

(B.37)
$$\mathbb{P}(\pi = \nu | \operatorname{sgn}(\pi) = s) = \mathbb{P}(\pi = \nu | \pi^{-1}(k) = \nu^{-1}(k), \mathcal{T}^\pi_{k+1} = \mathcal{T}^\nu_{k+1}, \operatorname{sgn}(\pi) = s)$$
$$\cdot \mathbb{P}(\pi^{-1}(k) = \nu^{-1}(k) | \mathcal{T}^\pi_{k+1} = \mathcal{T}^\nu_{k+1}, \operatorname{sgn}(\pi) = s) \cdot \mathbb{P}(\mathcal{T}^\pi_{k+1} = \mathcal{T}^\nu_{k+1}, \operatorname{sgn}(\pi) = s)$$

Putting this in yields

(B.38)
$$= \sum_{k=i}^{n} \sum_{T \in \mathbb{T}_{k+1}} \sum_{\substack{j=1 \\ j \notin T}}^{n} \sum_{\substack{\nu \in S_n: \\ \nu^{-1}(k) = j, \\ \mathcal{T}^\nu_{k+1} = T, \\ \operatorname{sgn}(\nu) = s}} \big[\mathbb{P}(\pi = \nu | \pi^{-1}(k) = \nu^{-1}(k), \mathcal{T}^\pi_{k+1} = \mathcal{T}^\nu_{k+1}, \operatorname{sgn}(\pi) = s)$$
$$\cdot \mathbb{P}(\pi^{-1}(k) = \nu^{-1}(k) | \mathcal{T}^\pi_{k+1} = \mathcal{T}^\nu_{k+1}, \operatorname{sgn}(\pi) = s) \mathbb{P}(\mathcal{T}^\pi_{k+1} = \mathcal{T}^\nu_{k+1} | \operatorname{sgn}(\pi) = s)$$
$$\cdot \log(S_k \mathbb{P}(\pi^{-1}(k) = \nu^{-1}(k) | \mathcal{T}^\pi_{k+1} = \mathcal{T}^\nu_{k+1}, \operatorname{sgn}(\pi) = s)) \big]$$

In the inner sum, $\mathcal{T}^\nu_{k+1}$ is T, and $\nu^{-1}(k)$ is j:

(B.39)
$$= \sum_{k=i}^{n} \sum_{T \in \mathbb{T}_{k+1}} \sum_{\substack{j=1 \\ j \notin T}}^{n} \sum_{\substack{\nu \in S_n: \\ \nu^{-1}(k) = j, \\ \mathcal{T}^\nu_{k+1} = T, \\ \operatorname{sgn}(\nu) = s}} \big[\mathbb{P}(\pi = \nu | \pi^{-1}(k) = j, \mathcal{T}^\pi_{k+1} = T, \operatorname{sgn}(\pi) = s)$$
$$\cdot \mathbb{P}(\pi^{-1}(k) = j | \mathcal{T}^\pi_{k+1} = T, \operatorname{sgn}(\pi) = s) \mathbb{P}(\mathcal{T}^\pi_{k+1} = T | \operatorname{sgn}(\pi) = s)$$
$$\cdot \log(S_k \mathbb{P}(\pi^{-1}(k) = j | \mathcal{T}^\pi_{k+1} = T, \operatorname{sgn}(\pi) = s)) \big]$$

Pulling the terms independent of ν out of the inner sum, we get

(B.40)
$$= \sum_{k=i}^{n} \sum_{T \in \mathbb{T}_{k+1}} \sum_{\substack{j=1 \\ j \notin T}}^{n} \Bigg[\mathbb{P}(\mathcal{T}^\pi_{k+1} = T | \operatorname{sgn}(\pi) = s)$$
$$\cdot \mathbb{P}(\pi^{-1}(k) = j | \mathcal{T}^\pi_{k+1} = T, \operatorname{sgn}(\pi) = s) \log(S_k \mathbb{P}(\pi^{-1}(k) = j | \mathcal{T}^\pi_{k+1} = T, \operatorname{sgn}(\pi) = s))$$
$$\cdot \sum_{\substack{\nu \in S_n: \\ \nu^{-1}(k) = j, \\ \mathcal{T}^\nu_{k+1} = T, \\ \operatorname{sgn}(\nu) = s}} \mathbb{P}(\pi = \nu | \pi^{-1}(k) = j, \mathcal{T}^\pi_{k+1} = T, \operatorname{sgn}(\pi) = s) \Bigg]$$

The inner sum is now just the sum of probabilities that π is a particular permutation given several restrictions, over all such possible permutations, so it ought to be 1. We remark that it *could* be 0 if the sum was empty, but this only happens if $k = 2$ and j is chosen to be an impossible entry due to the sign, in which case the entropy term $\mathbb{P}(\pi^{-1}(k) = j|\mathcal{T}^\pi_{k+1} = T, \operatorname{sgn}(\pi) = s)\log(S_k\mathbb{P}(\pi^{-1}(k) = j|\mathcal{T}^\pi_{k+1} = T, \operatorname{sgn}(\pi) = s))$ is $0\log 0 = 0$ anyway and is of no consequence (i.e., $0 \cdot 1$ is the same as $0 \cdot 0$). So, we shall say that it is 1 without worry. We thus have

$$\begin{aligned}
&= \sum_{k=i}^{n} \sum_{T \in \mathbb{T}_{k+1}} \sum_{\substack{j=1 \\ j \notin T}}^{n} \Big[\mathbb{P}(\mathcal{T}^\pi_{k+1} = T|\operatorname{sgn}(\pi) = s) \\
&\qquad \cdot \mathbb{P}(\pi^{-1}(k) = j|\mathcal{T}^\pi_{k+1} = T, \operatorname{sgn}(\pi) = s)\log(S_k\mathbb{P}(\pi^{-1}(k) = j|\mathcal{T}^\pi_{k+1} = T, \operatorname{sgn}(\pi) = s)) \cdot 1\Big]
\end{aligned} \tag{B.41}$$

and now we may pull $\mathbb{P}(\mathcal{T}^\pi_{k+1} = T|\operatorname{sgn}(\pi) = s)$ since it is independent of j to get

$$\begin{aligned}
&= \sum_{k=i}^{n} \sum_{T \in \mathbb{T}_{k+1}} \Big[\mathbb{P}(\mathcal{T}^\pi_{k+1} = T|\operatorname{sgn}(\pi) = s) \\
&\qquad \cdot \sum_{\substack{j=1 \\ j \notin T}}^{n} \mathbb{P}(\pi^{-1}(k) = j|\mathcal{T}^\pi_{k+1} = T, \operatorname{sgn}(\pi) = s)\log(S_k\mathbb{P}(\pi^{-1}(k) = j|\mathcal{T}^\pi_{k+1} = T, \operatorname{sgn}(\pi) = s))\Big]
\end{aligned} \tag{B.42}$$

and now we recognize that the inner sum is $\operatorname{ENT}(\pi^{-1}(k)|\mathcal{T}^\pi_{k+1} = T, \operatorname{sgn}(\pi) = s)$:

$$= \sum_{k=i}^{n} \sum_{T \in \mathbb{T}_{k+1}} \Big[\mathbb{P}(\mathcal{T}^\pi_{k+1} = T|\operatorname{sgn}(\pi) = s)\operatorname{ENT}(\pi^{-1}(k)|\mathcal{T}^\pi_{k+1} = T, \operatorname{sgn}(\pi) = s)\Big] \tag{B.43}$$

This is our result for the second sum.

Going back to $\mathbb{E}(\operatorname{ENT}(\pi|\operatorname{sgn}(\pi)))$, our work shows that

$$\begin{aligned}
&\mathbb{E}(\operatorname{ENT}(\pi|\operatorname{sgn}(\pi))) \\
&= \sum_{s=-1,1} \mathbb{P}(\operatorname{sgn}(\pi) = s)\Bigg[\sum_{T \in \mathbb{T}_i} \mathbb{P}(\mathcal{T}^\pi_i = T|\operatorname{sgn}(\pi) = s)\operatorname{ENT}(\pi|T^\pi_i = T, \operatorname{sgn}(\pi) = s) \\
&\quad + \sum_{k=i}^{n} \sum_{T \in \mathbb{T}_{k+1}} \mathbb{P}(\mathcal{T}^\pi_{k+1} = T|\operatorname{sgn}(\pi) = s)\operatorname{ENT}(\pi^{-1}(k)|\mathcal{T}^\pi_{k+1} = T, \operatorname{sgn}(\pi) = s)\Bigg]
\end{aligned} \tag{B.44}$$

Distributing the outer sum and $\mathbb{P}(\mathrm{sgn}(\pi) = s)$, we get

(B.45)
$$
\begin{aligned}
&\mathbb{E}\left(\mathrm{ENT}(\pi | \mathrm{sgn}(\pi))\right) \\
&= \sum_{s=-1,1} \sum_{T \in \mathbb{T}_i} \mathbb{P}(\mathrm{sgn}(\pi) = s) \mathbb{P}(\mathcal{T}_i^\pi = T | \mathrm{sgn}(\pi) = s)\, \mathrm{ENT}(\pi | T_i^\pi = T, \mathrm{sgn}(\pi) = s) \\
&\quad + \sum_{s=-1,1} \sum_{k=i}^{n} \sum_{T \in \mathbb{T}_{k+1}} \mathbb{P}(\mathrm{sgn}(\pi) = s) \mathbb{P}(\mathcal{T}_{k+1}^\pi = T | \mathrm{sgn}(\pi) = s)\, \mathrm{ENT}(\pi^{-1}(k) | \mathcal{T}_{k+1}^\pi = T, \mathrm{sgn}(\pi) = s)
\end{aligned}
$$

We observe that we get joint probabilities, and that the double sums yield expectations:

$$
\begin{aligned}
&\mathbb{E}\left(\mathrm{ENT}(\pi | \mathrm{sgn}(\pi))\right) \\
&= \sum_{s=-1,1} \sum_{T \in \mathbb{T}_i} \mathbb{P}(\mathcal{T}_i^\pi = T, \mathrm{sgn}(\pi) = s)\, \mathrm{ENT}(\pi | T_i^\pi = T, \mathrm{sgn}(\pi) = s) \\
&\quad + \sum_{s=-1,1} \sum_{k=i}^{n} \sum_{T \in \mathbb{T}_{k+1}} \mathbb{P}(\mathcal{T}_{k+1}^\pi = T, \mathrm{sgn}(\pi) = s)\, \mathrm{ENT}(\pi^{-1}(k) | \mathcal{T}_{k+1}^\pi = T, \mathrm{sgn}(\pi) = s) \\
&= \mathbb{E}\left(\mathrm{ENT}(\pi | T_i^\pi, \mathrm{sgn}(\pi))\right) \\
&\quad + \sum_{k=i}^{n} \mathbb{E}\left(\mathrm{ENT}(\pi^{-1}(k) | \mathcal{T}_{k+1}^\pi, \mathrm{sgn}(\pi))\right)
\end{aligned}
\tag{B.46}
$$

We conclude that

$$
\mathrm{ENT}(\pi) = \mathbb{E}\left(\mathrm{ENT}(\pi | \mathrm{sgn}(\pi), \mathcal{T}_i^\pi)\right) + \sum_{k=i}^{n} \left(\mathbb{E}\left(\mathrm{ENT}(\pi^{-1}(k) | \mathrm{sgn}(\pi), \mathcal{T}_{k+1}^\pi)\right)\right) + \mathrm{ENT}(\mathrm{sgn}(\pi)) \tag{B.47}
$$

as we wanted.

APPENDIX C

Adding Conditioning

This appendix gives a detailed explanation of equation (4.6), which we recall is

$$\mathbb{E}(\mathrm{ENT}(\mu\kappa\pi'_{(t)}|\operatorname{sgn}(\mu\kappa\pi'_{(t)}))) \leq \mathbb{E}(\mathrm{ENT}(\mu\kappa\pi'_{(t)}|M, \pi'_{(t)}, \operatorname{sgn}(\mu\kappa\pi'_{(t)}))). \tag{C.1}$$

We shall use Jensen's inequality: By conditioning further on M and $\pi'_{(t)}$, we may write, for any permutation ν and sign s,

(C.2)
$$\begin{aligned}
&\mathbb{P}(\mu\kappa\pi'_{(t)} = \nu|\operatorname{sgn}(\mu\kappa\pi'_{(t)}) = s) \\
&\quad = \sum_{m,\rho} \mathbb{P}(\mu\kappa\pi'_{(t)} = \nu|M = m, \pi'_{(t)} = \rho, \operatorname{sgn}(\mu\kappa\pi'_{(t)}) = s) \cdot \mathbb{P}(M = m, \pi'_{(t)} = \rho|\operatorname{sgn}(\mu\kappa\pi'_{(t)}) = s) \\
&\quad = \mathbb{E}(\mathbb{P}(\mu\kappa\pi'_{(t)} = \nu|M, \pi'_{(t)}, \operatorname{sgn}(\mu\kappa\pi'_{(t)}) = s)),
\end{aligned}$$

where the sum is over all possible matchings m for M and all possible permutations ρ for $\pi'_{(t)}$.

We can write $\mathbb{E}(\mathrm{ENT}(\mu\kappa\pi'_{(t)}|\operatorname{sgn}(\mu\kappa\pi'_{(t)})))$ in the following way:

$$\mathbb{E}(\mathrm{ENT}(\mu\kappa\pi'_{(t)}|\operatorname{sgn}(\mu\kappa\pi'_{(t)}))) = \sum_{s=-1,1} \mathrm{ENT}(\mu\kappa\pi'_{(t)}|\operatorname{sgn}(\mu\kappa\pi'_{(t)}) = s)\mathbb{P}(\operatorname{sgn}(\mu\kappa\pi'_{(t)}) = s). \tag{C.3}$$

Consider $\mathrm{ENT}(\mu\kappa\pi'_{(t)}|\operatorname{sgn}(\mu\kappa\pi'_{(t)}) = s)$; recall that

$$\begin{aligned}
&\mathrm{ENT}(\mu\kappa\pi'_{(t)}|\operatorname{sgn}(\mu\kappa\pi'_{(t)}) = s) \\
&\quad = \sum_{\nu:\operatorname{sgn}(\nu)=s} \mathbb{P}(\mu\kappa\pi'_{(t)} = \nu|\operatorname{sgn}(\mu\kappa\pi'_{(t)}) = s) \cdot \log\left(\frac{n!}{2}\mathbb{P}(\mu\kappa\pi'_{(t)} = \nu|\operatorname{sgn}(\mu\kappa\pi'_{(t)}) = s)\right).
\end{aligned} \tag{C.4}$$

Inserting our extra conditioning, we get

$$
\begin{aligned}
&= \sum_{\nu:\mathrm{sgn}(\nu)=s} \mathbb{E}(\mathbb{P}(\mu\kappa\pi'_{(t)} = \nu | M, \pi'_{(t)}, \mathrm{sgn}(\mu\kappa\pi'_{(t)}) = s)) \\
&\qquad \cdot \log\left(\frac{n!}{2}\mathbb{E}(\mathbb{P}(\mu\kappa\pi'_{(t)} = \nu | M, \pi'_{(t)}, \mathrm{sgn}(\mu\kappa\pi'_{(t)}) = s))\right).
\end{aligned} \tag{C.5}
$$

Since $x\log(ax)$ is convex for $x > 0$, for any $a > 0$ (the second derivative is $\frac{1}{x}$), by applying Jensen's inequality to each term in the sum we get

$$
\begin{aligned}
&\leq \sum_{\nu:\mathrm{sgn}(\nu)=s} \mathbb{E}\Bigg(\mathbb{P}(\mu\kappa\pi'_{(t)} = \nu | M, \pi'_{(t)}, \mathrm{sgn}(\mu\kappa\pi'_{(t)}) = s) \\
&\qquad \cdot \log\left(\frac{n!}{2}\mathbb{P}(\mu\kappa\pi'_{(t)} = \nu | M, \pi'_{(t)}, \mathrm{sgn}(\mu\kappa\pi'_{(t)}) = s)\right)\Bigg).
\end{aligned} \tag{C.6}
$$

This is

$$
\begin{aligned}
&\sum_{\nu:\mathrm{sgn}(\nu)=s} \sum_{m,\rho} (\mathbb{P}(M = m, \pi'_{(t)} = \rho | \operatorname{sgn}(\mu\kappa\pi'_{(t)}) = s) \\
&\quad \cdot \Bigg(\mathbb{P}(\mu\kappa\pi'_{(t)} = \nu | M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa\pi'_{(t)}) = s) \\
&\quad \cdot \log\left(\frac{n!}{2}\mathbb{P}(\mu\kappa\pi'_{(t)} = \nu | M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa\pi'_{(t)}) = s)\right)\Bigg)).
\end{aligned} \tag{C.7}
$$

By switching the order of summation and pulling $\mathbb{P}(M = m, \pi'_{(t)} = \rho | \operatorname{sgn}(\mu\kappa\pi'_{(t)}) = s)$ through the now inner sum, we obtain

$$
\begin{aligned}
&\sum_{m,\rho} \mathbb{P}(M = m, \pi'_{(t)} = \rho | \operatorname{sgn}(\mu\kappa\pi'_{(t)}) = s) \\
&\quad \cdot \sum_{\nu:\mathrm{sgn}(\nu)=s} \Bigg(\mathbb{P}(\mu\kappa\pi'_{(t)} = \nu | M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa\pi'_{(t)}) = s) \\
&\qquad \cdot \log\left(\frac{n!}{2}\mathbb{P}(\mu\kappa\pi'_{(t)} = \nu | M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa\pi'_{(t)}) = s)\right)\Bigg)
\end{aligned} \tag{C.8}
$$

and now this inner sum is $\mathrm{ENT}(\mu\kappa\pi'_{(t)} | M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa\pi'_{(t)}) = s)$, so we now have

$$
\begin{aligned}
&\sum_{m,\rho} \mathbb{P}(M = m, \pi'_{(t)} = \rho | \operatorname{sgn}(\mu\kappa\pi'_{(t)}) = s) \,\mathrm{ENT}(\mu\kappa\pi'_{(t)} | M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa\pi'_{(t)}) = s) \\
&= \mathbb{E}\,\mathrm{ENT}(\mu\kappa\pi'_{(t)} | M, \pi'_{(t)}, \mathrm{sgn}(\mu\kappa\pi'_{(t)}) = s).
\end{aligned} \tag{C.9}
$$

Recall that we started with $\mathrm{ENT}(\mu\kappa\pi'_{(t)}|\,\mathrm{sgn}(\mu\kappa\pi'_{(t)}) = s)$, so we have that

$$\mathrm{ENT}(\mu\kappa\pi'_{(t)}|\,\mathrm{sgn}(\mu\kappa\pi'_{(t)}) = s) \leq \mathbb{E}\,\mathrm{ENT}(\mu\kappa\pi'_{(t)}|M, \pi'_{(t)}, \mathrm{sgn}(\mu\kappa\pi'_{(t)}) = s). \tag{C.10}$$

It follows that

$$\begin{aligned}
&\mathbb{E}(\mathrm{ENT}(\mu\kappa\pi'_{(t)}|\,\mathrm{sgn}(\mu\kappa\pi'_{(t)})))\\
&= \sum_{s=-1,1} \mathrm{ENT}(\mu\kappa\pi'_{(t)}|\,\mathrm{sgn}(\mu\kappa\pi'_{(t)}) = s)\mathbb{P}(\mathrm{sgn}(\mu\kappa\pi'_{(t)}) = s)\\
&\leq \sum_{s=-1,1} \mathbb{E}\,\mathrm{ENT}(\mu\kappa\pi'_{(t)}|M, \pi'_{(t)}, \mathrm{sgn}(\mu\kappa\pi'_{(t)}) = s)\mathbb{P}(\mathrm{sgn}(\mu\kappa\pi'_{(t)}) = s).
\end{aligned} \tag{C.11}$$

Unwrapping the notation one more time, this is

$$\begin{aligned}
&= \sum_{s=-1,1}\sum_{m,\rho}(\mathbb{P}(M = m, \pi'_{(t)} = \rho|\,\mathrm{sgn}(\mu\kappa\pi'_{(t)}) = s)\,\mathrm{ENT}(\mu\kappa\pi'_{(t)}|M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa\pi'_{(t)}) = s)\\
&\qquad \cdot \mathbb{P}(\mathrm{sgn}(\mu\kappa\pi'_{(t)}) = s))\\
&= \sum_{s=-1,1}\sum_{m,\rho}\mathbb{P}(M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa\pi'_{(t)}) = s)\,\mathrm{ENT}(\mu\kappa\pi'_{(t)}|M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa\pi'_{(t)}) = s)\\
&= \mathbb{E}(\mathrm{ENT}(\mu\kappa\pi'_{(t)}|M, \pi'_{(t)}, \mathrm{sgn}(\mu\kappa\pi'_{(t)}))).
\end{aligned} \tag{C.12}$$

Therefore,

$$\mathbb{E}(\mathrm{ENT}(\mu\kappa\pi'_{(t)}|\,\mathrm{sgn}(\mu\kappa\pi'_{(t)}))) \leq \mathbb{E}(\mathrm{ENT}(\mu\kappa\pi'_{(t)}|M, \pi'_{(t)}, \mathrm{sgn}(\mu\kappa\pi'_{(t)}))), \tag{C.13}$$

that is to say, additional conditioning causes an increase in the expected relative entropy.

APPENDIX D

Removing Conditioning

We claimed that

$$\mathbb{E}(\mathrm{ENT}(\mu\kappa\pi'_{(t)}|M, \pi'_{(t)}, \mathrm{sgn}(\mu\kappa\pi'_{(t)}))) = \mathbb{E}(\mathrm{ENT}(\mu\kappa|M, \mathrm{sgn}(\mu\kappa))) \tag{D.1}$$

by a series of arguments about independence and conditioning, ending with equation 4.9. We elaborate here by writing out the details.

First, we have that

$$\begin{aligned}\mathbb{E}(\mathrm{ENT}(\mu\kappa\pi'_{(t)}|M, \pi'_{(t)}, \mathrm{sgn}(\mu\kappa\pi'_{(t)}))) = \sum_{s=-1,1}\sum_{\rho}\sum_{m}(&\mathbb{P}(M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa\pi'_{(t)}) = s)\\ &\cdot \mathrm{ENT}(\mu\kappa\pi'_{(t)}|M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa\pi'_{(t)}) = s)).\end{aligned} \tag{D.2}$$

Expand $\mathrm{ENT}(\mu\kappa\pi'_{(t)}|M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa\pi'_{(t)}) = s)$ to get

$$\begin{aligned}&= \sum_{s=-1,1}\sum_{\rho}\sum_{m}\Bigg(\mathbb{P}(M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa\pi'_{(t)}) = s)\\ &\cdot \sum_{\nu:\mathrm{sgn}(\nu)=s}\Big(\mathbb{P}(\mu\kappa\pi'_{(t)} = \nu|M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa\pi'_{(t)}) = s)\\ &\cdot \log\left(\frac{n!}{2}\mathbb{P}(\mu\kappa\pi'_{(t)} = \nu|M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa\pi'_{(t)}) = s)\right)\Big)\Bigg)\end{aligned} \tag{D.3}$$

Next, expand the sum over the signs to get

$$
\begin{aligned}
&= \sum_{\rho} \sum_{m} \Bigg(\mathbb{P}(M = m, \pi'_{(t)} = \rho, \operatorname{sgn}(\mu\kappa\pi'_{(t)}) = 1) \\
&\quad \cdot \sum_{\nu : \operatorname{sgn}(\nu) = 1} \bigg(\mathbb{P}(\mu\kappa\pi'_{(t)} = \nu | M = m, \pi'_{(t)} = \rho, \operatorname{sgn}(\mu\kappa\pi'_{(t)}) = 1) \\
&\qquad \cdot \log\left(\frac{n!}{2} \mathbb{P}(\mu\kappa\pi'_{(t)} = \nu | M = m, \pi'_{(t)} = \rho, \operatorname{sgn}(\mu\kappa\pi'_{(t)}) = 1) \right) \bigg) \Bigg) \\
&+ \sum_{\rho} \sum_{m} \Bigg(\mathbb{P}(M = m, \pi'_{(t)} = \rho, \operatorname{sgn}(\mu\kappa\pi'_{(t)}) = -1) \\
&\quad \cdot \sum_{\nu : \operatorname{sgn}(\nu) = -1} \bigg(\mathbb{P}(\mu\kappa\pi'_{(t)} = \nu | M = m, \pi'_{(t)} = \rho, \operatorname{sgn}(\mu\kappa\pi'_{(t)}) = -1) \\
&\qquad \cdot \log\left(\frac{n!}{2} \mathbb{P}(\mu\kappa\pi'_{(t)} = \nu | M = m, \pi'_{(t)} = \rho, \operatorname{sgn}(\mu\kappa\pi'_{(t)}) = -1) \right) \bigg) \Bigg)
\end{aligned} \tag{D.4}
$$

Now, expand each term over ρ by considering the sign of ρ:

$$
\begin{aligned}
&= \sum_{\substack{\rho: \\ \operatorname{sgn}(\rho) = 1}} \sum_{m} \Bigg(\mathbb{P}(M = m, \pi'_{(t)} = \rho, \operatorname{sgn}(\mu\kappa\pi'_{(t)}) = 1) \\
&\quad \cdot \sum_{\nu : \operatorname{sgn}(\nu) = 1} \bigg(\mathbb{P}(\mu\kappa\pi'_{(t)} = \nu | M = m, \pi'_{(t)} = \rho, \operatorname{sgn}(\mu\kappa\pi'_{(t)}) = 1) \\
&\qquad \cdot \log\left(\frac{n!}{2} \mathbb{P}(\mu\kappa\pi'_{(t)} = \nu | M = m, \pi'_{(t)} = \rho, \operatorname{sgn}(\mu\kappa\pi'_{(t)}) = 1) \right) \bigg) \Bigg)
\end{aligned} \tag{D.5}
$$

$$
\begin{aligned}
&+ \sum_{\substack{\rho: \\ \mathrm{sgn}(\rho)=-1}} \sum_{m} \Bigg(\mathbb{P}(M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa\pi'_{(t)}) = 1) \\
&\quad \cdot \sum_{\nu:\mathrm{sgn}(\nu)=1} \bigg(\mathbb{P}(\mu\kappa\pi'_{(t)} = \nu | M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa\pi'_{(t)}) = 1) \\
&\qquad \cdot \log\left(\frac{n!}{2} \mathbb{P}(\mu\kappa\pi'_{(t)} = \nu | M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa\pi'_{(t)}) = 1) \right) \bigg) \Bigg)
\end{aligned}
$$

$$
\begin{aligned}
&+ \sum_{\substack{\rho: \\ \mathrm{sgn}(\rho)=1}} \sum_{m} \Bigg(\mathbb{P}(M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa\pi'_{(t)}) = -1) \\
&\quad \cdot \sum_{\nu:\mathrm{sgn}(\nu)=-1} \bigg(\mathbb{P}(\mu\kappa\pi'_{(t)} = \nu | M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa\pi'_{(t)}) = -1) \\
&\qquad \cdot \log\left(\frac{n!}{2} \mathbb{P}(\mu\kappa\pi'_{(t)} = \nu | M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa\pi'_{(t)}) = -1) \right) \bigg) \Bigg)
\end{aligned}
$$

$$
\begin{aligned}
&+ \sum_{\substack{\rho: \\ \mathrm{sgn}(\rho)=-1}} \sum_{m} \Bigg(\mathbb{P}(M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa\pi'_{(t)}) = -1) \\
&\quad \cdot \sum_{\nu:\mathrm{sgn}(\nu)=-1} \bigg(\mathbb{P}(\mu\kappa\pi'_{(t)} = \nu | M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa\pi'_{(t)}) = -1) \\
&\qquad \cdot \log\left(\frac{n!}{2} \mathbb{P}(\mu\kappa\pi'_{(t)} = \nu | M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa\pi'_{(t)}) = -1) \right) \bigg) \Bigg)
\end{aligned}
$$

Now observe that in each term, we may rewrite the condition about $\operatorname{sgn}(\mu\kappa\pi'_{(t)})$ to be a condition about $\operatorname{sgn}(\mu\kappa)$:

$$
\begin{aligned}
&= \sum_{\substack{\rho:\\ \operatorname{sgn}(\rho)=1}} \sum_{m} \Bigg(\mathbb{P}(M = m, \pi'_{(t)} = \rho, \operatorname{sgn}(\mu\kappa) = 1) \\
&\quad \cdot \sum_{\nu:\operatorname{sgn}(\nu)=1} \Big(\mathbb{P}(\mu\kappa\pi'_{(t)} = \nu | M = m, \pi'_{(t)} = \rho, \operatorname{sgn}(\mu\kappa) = 1) \\
&\qquad \cdot \log\left(\frac{n!}{2} \mathbb{P}(\mu\kappa\pi'_{(t)} = \nu | M = m, \pi'_{(t)} = \rho, \operatorname{sgn}(\mu\kappa) = 1) \right) \Big) \\
&+ \sum_{\substack{\rho:\\ \operatorname{sgn}(\rho)=-1}} \sum_{m} \Bigg(\mathbb{P}(M = m, \pi'_{(t)} = \rho, \operatorname{sgn}(\mu\kappa) = -1) \\
&\quad \cdot \sum_{\nu:\operatorname{sgn}(\nu)=1} \Big(\mathbb{P}(\mu\kappa\pi'_{(t)} = \nu | M = m, \pi'_{(t)} = \rho, \operatorname{sgn}(\mu\kappa) = -1) \\
&\qquad \cdot \log\left(\frac{n!}{2} \mathbb{P}(\mu\kappa\pi'_{(t)} = \nu | M = m, \pi'_{(t)} = \rho, \operatorname{sgn}(\mu\kappa) = -1) \right) \Big) \Bigg) \\
&+ \sum_{\substack{\rho:\\ \operatorname{sgn}(\rho)=1}} \sum_{m} \Bigg(\mathbb{P}(M = m, \pi'_{(t)} = \rho, \operatorname{sgn}(\mu\kappa) = -1) \\
&\quad \cdot \sum_{\nu:\operatorname{sgn}(\nu)=-1} \Big(\mathbb{P}(\mu\kappa\pi'_{(t)} = \nu | M = m, \pi'_{(t)} = \rho, \operatorname{sgn}(\mu\kappa) = -1) \\
&\qquad \cdot \log\left(\frac{n!}{2} \mathbb{P}(\mu\kappa\pi'_{(t)} = \nu | M = m, \pi'_{(t)} = \rho, \operatorname{sgn}(\mu\kappa) = -1) \right) \Big) \Bigg)
\end{aligned} \tag{D.6}
$$

$$+\sum_{\substack{\rho:\\ \mathrm{sgn}(\rho)=-1}}\sum_{m}\Bigg(\mathbb{P}(M=m,\pi'_{(t)}=\rho,\mathrm{sgn}(\mu\kappa)=1)$$

$$\cdot\sum_{\nu:\mathrm{sgn}(\nu)=-1}\Bigg(\mathbb{P}(\mu\kappa\pi'_{(t)}=\nu|M=m,\pi'_{(t)}=\rho,\mathrm{sgn}(\mu\kappa)=1)$$

$$\cdot\ \log\left(\frac{n!}{2}\mathbb{P}(\mu\kappa\pi'_{(t)}=\nu|M=m,\pi'_{(t)}=\rho,\mathrm{sgn}(\mu\kappa)=1)\right)\Bigg)\Bigg)$$

At this point, consolidating gives us that

$$\mathbb{E}(\mathrm{ENT}(\mu\kappa\pi'_{(t)}|M,\pi'_{(t)},\mathrm{sgn}(\mu\kappa\pi'_{(t)})))=\mathbb{E}(\mathrm{ENT}(\mu\kappa\pi'_{(t)}|M,\pi'_{(t)},\mathrm{sgn}(\mu\kappa))) \tag{D.7}$$

as we claimed. (Switching the second and fourth terms will group the terms with the same condition on $\mathrm{sgn}(\mu\kappa)$ together.)

Continuing on, since $\pi'_{(t)}$ is known, we may rewrite the conditional probability of $\mu\kappa\pi'_{(t)}$ as the conditional probability of $\mu\kappa$:

$$\begin{aligned}
&=\sum_{\substack{\rho:\\ \mathrm{sgn}(\rho)=1}}\sum_{m}\Bigg(\mathbb{P}(M=m,\pi'_{(t)}=\rho,\mathrm{sgn}(\mu\kappa)=1)\\
&\quad\cdot\sum_{\nu:\mathrm{sgn}(\nu)=1}\Bigg(\mathbb{P}(\mu\kappa=\nu\rho^{-1}|M=m,\pi'_{(t)}=\rho,\mathrm{sgn}(\mu\kappa)=1)\\
&\qquad\cdot\ \log\left(\frac{n!}{2}\mathbb{P}(\mu\kappa=\nu\rho^{-1}|M=m,\pi'_{(t)}=\rho,\mathrm{sgn}(\mu\kappa)=1)\right)\Bigg)\Bigg)
\end{aligned} \tag{D.8}$$

$$+\sum_{\substack{\rho:\\ \mathrm{sgn}(\rho)=-1}}\sum_{m}\Bigg(\mathbb{P}(M=m,\pi'_{(t)}=\rho,\mathrm{sgn}(\mu\kappa)=-1)$$

$$\cdot\sum_{\nu:\mathrm{sgn}(\nu)=1}\Bigg(\mathbb{P}(\mu\kappa=\nu\rho^{-1}|M=m,\pi'_{(t)}=\rho,\mathrm{sgn}(\mu\kappa)=-1)$$

$$\cdot\ \log\left(\frac{n!}{2}\mathbb{P}(\mu\kappa=\nu\rho^{-1}|M=m,\pi'_{(t)}=\rho,\mathrm{sgn}(\mu\kappa)=-1)\right)\Bigg)\Bigg)$$

$$
\begin{aligned}
&+ \sum_{\substack{\rho: \\ \mathrm{sgn}(\rho)=1}} \sum_{m} \Bigg(\mathbb{P}(M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa) = -1) \\
&\quad \cdot \sum_{\nu:\mathrm{sgn}(\nu)=-1} \Big(\mathbb{P}(\mu\kappa = \nu\rho^{-1} | M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa) = -1) \\
&\qquad \cdot \log \left(\frac{n!}{2} \mathbb{P}(\mu\kappa = \nu\rho^{-1} | M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa) = -1) \right) \Big) \Bigg)
\end{aligned}
$$

$$
\begin{aligned}
&+ \sum_{\substack{\rho: \\ \mathrm{sgn}(\rho)=-1}} \sum_{m} \Bigg(\mathbb{P}(M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa) = 1) \\
&\quad \cdot \sum_{\nu:\mathrm{sgn}(\nu)=-1} \Big(\mathbb{P}(\mu\kappa = \nu\rho^{-1} | M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa) = 1) \\
&\qquad \cdot \log \left(\frac{n!}{2} \mathbb{P}(\mu\kappa = \nu\rho^{-1} | M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa) = 1) \right) \Big) \Bigg)
\end{aligned}
$$

Now observe that summing the conditional probability that $\mu\kappa = \nu\rho^{-1}$ over the ν that have sign s is the same as summing the conditional probability that $\mu\kappa = \eta$ over the η that have sign $s\,\mathrm{sgn}(\rho^{-1}) = s\,\mathrm{sgn}\,\rho$. Since we have explicitly written out the signs here, the necessary modifications are clearer: We get

$$
\begin{aligned}
&= \sum_{\substack{\rho: \\ \mathrm{sgn}(\rho)=1}} \sum_{m} \Bigg(\mathbb{P}(M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa) = 1) \\
&\quad \cdot \sum_{\eta:\mathrm{sgn}(\eta)=1} \Big(\mathbb{P}(\mu\kappa = \eta | M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa) = 1) \\
&\qquad \cdot \log \left(\frac{n!}{2} \mathbb{P}(\mu\kappa = \eta | M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa) = 1) \right) \Big) \Bigg)
\end{aligned}
\tag{D.9}
$$

$$+ \sum_{\substack{\rho: \\ \mathrm{sgn}(\rho)=-1}} \sum_{m} \Bigg(\mathbb{P}(M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa) = -1)$$

$$\cdot \sum_{\eta:\mathrm{sgn}(\eta)=-1} \Big(\mathbb{P}(\mu\kappa = \eta | M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa) = -1)$$

$$\cdot \log \left(\frac{n!}{2} \mathbb{P}(\mu\kappa = \eta | M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa) = -1) \right) \Big) \Bigg)$$

$$+ \sum_{\substack{\rho: \\ \mathrm{sgn}(\rho)=1}} \sum_{m} \Bigg(\mathbb{P}(M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa) = -1)$$

$$\cdot \sum_{\eta:\mathrm{sgn}(\eta)=-1} \Big(\mathbb{P}(\mu\kappa = \eta | M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa) = -1)$$

$$\cdot \log \left(\frac{n!}{2} \mathbb{P}(\mu\kappa = \eta | M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa) = -1) \right) \Big) \Bigg)$$

$$+ \sum_{\substack{\rho: \\ \mathrm{sgn}(\rho)=-1}} \sum_{m} \Bigg(\mathbb{P}(M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa) = 1)$$

$$\cdot \sum_{\nu:\mathrm{sgn}(\eta)=1} \Big(\mathbb{P}(\mu\kappa = \eta | M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa) = 1)$$

$$\cdot \log \left(\frac{n!}{2} \mathbb{P}(\mu\kappa = \eta | M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa) = 1) \right) \Big) \Bigg)$$

This shows that

$$\mathbb{E}(\mathrm{ENT}(\mu\kappa\pi'_{(t)} | M, \pi'_{(t)}, \mathrm{sgn}(\mu\kappa))) = \mathbb{E}(\mathrm{ENT}(\mu\kappa | M, \pi'_{(t)}, \mathrm{sgn}(\mu\kappa))). \tag{D.10}$$

Finally, since μ, κ, and $\pi'_{(t)}$ are conditionally independent given M, we have that

$$\mathbb{P}(\mu\kappa = \eta | M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa) = s) = \mathbb{P}(\mu\kappa = \eta | M = m, \mathrm{sgn}(\mu\kappa) = s); \tag{D.11}$$

that is to say, we may drop the conditioning on $\pi'_{(t)}$. Doing so gives

(D.12)

$$= \sum_{\substack{\rho: \\ \mathrm{sgn}(\rho)=1}} \sum_{m} \Bigg(\mathbb{P}(M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa) = 1)$$

$$\cdot \sum_{\eta:\mathrm{sgn}(\eta)=1} \left(\mathbb{P}(\mu\kappa = \eta | M = m, \mathrm{sgn}(\mu\kappa) = 1) \log \left(\frac{n!}{2} \mathbb{P}(\mu\kappa = \eta | M = m, \mathrm{sgn}(\mu\kappa) = 1) \right) \right) \Bigg)$$

$$+ \sum_{\substack{\rho: \\ \mathrm{sgn}(\rho)=-1}} \sum_{m} \Bigg(\mathbb{P}(M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa) = -1)$$

$$\cdot \sum_{\eta:\mathrm{sgn}(\eta)=-1} \left(\mathbb{P}(\mu\kappa = \eta | M = m, \mathrm{sgn}(\mu\kappa) = -1) \log \left(\frac{n!}{2} \mathbb{P}(\mu\kappa = \eta | M = m, \mathrm{sgn}(\mu\kappa) = -1) \right) \right) \Bigg)$$

$$+ \sum_{\substack{\rho: \\ \mathrm{sgn}(\rho)=1}} \sum_{m} \Bigg(\mathbb{P}(M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa) = -1)$$

$$\cdot \sum_{\eta:\mathrm{sgn}(\eta)=-1} \left(\mathbb{P}(\mu\kappa = \eta | M = m, \mathrm{sgn}(\mu\kappa) = -1) \log \left(\frac{n!}{2} \mathbb{P}(\mu\kappa = \eta | M = m, \mathrm{sgn}(\mu\kappa) = -1) \right) \right) \Bigg)$$

$$+ \sum_{\substack{\rho: \\ \mathrm{sgn}(\rho)=-1}} \sum_{m} \Bigg(\mathbb{P}(M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa) = 1)$$

$$\cdot \sum_{\nu:\mathrm{sgn}(\eta)=1} \left(\mathbb{P}(\mu\kappa = \eta | M = m, \mathrm{sgn}(\mu\kappa) = 1) \log \left(\frac{n!}{2} \mathbb{P}(\mu\kappa = \eta | M = m, \mathrm{sgn}(\mu\kappa) = 1) \right) \right) \Bigg)$$

Recognizing each inner sum as an entropy term, we get

$$= \sum_{\substack{\rho: \\ \mathrm{sgn}(\rho)=1}} \sum_{m} \mathbb{P}(M = m, \pi'_{(t)} = \rho, \mathrm{sgn}(\mu\kappa) = 1)\, \mathrm{ENT}(\mu\kappa | M = m, \mathrm{sgn}(\mu\kappa) = 1) \tag{D.13}$$

$$+\sum_{\substack{\rho:\\ \operatorname{sgn}(\rho)=-1}}\sum_{m}\mathbb{P}(M=m,\pi'_{(t)}=\rho,\operatorname{sgn}(\mu\kappa)=-1)\,\mathrm{ENT}(\mu\kappa|M=m,\operatorname{sgn}(\mu\kappa)=-1)$$

$$+\sum_{\substack{\rho:\\ \operatorname{sgn}(\rho)=1}}\sum_{m}\mathbb{P}(M=m,\pi'_{(t)}=\rho,\operatorname{sgn}(\mu\kappa)=-1)\,\mathrm{ENT}(\mu\kappa|M=m,\operatorname{sgn}(\mu\kappa)=-1)$$

$$+\sum_{\substack{\rho:\\ \operatorname{sgn}(\rho)=-1}}\sum_{m}\mathbb{P}(M=m,\pi'_{(t)}=\rho,\operatorname{sgn}(\mu\kappa)=1)\,\mathrm{ENT}(\mu\kappa|M=m,\operatorname{sgn}(\mu\kappa)=1)$$

Observe that the first and fourth terms together form the sum over all ρ of their addends, and likewise for the second and third terms. Thus we get

$$\begin{aligned}&=\sum_{\rho}\sum_{m}\mathbb{P}(M=m,\pi'_{(t)}=\rho,\operatorname{sgn}(\mu\kappa)=1)\,\mathrm{ENT}(\mu\kappa|M=m,\operatorname{sgn}(\mu\kappa)=1)\\ &\quad+\sum_{\rho}\sum_{m}\mathbb{P}(M=m,\pi'_{(t)}=\rho,\operatorname{sgn}(\mu\kappa)=-1)\,\mathrm{ENT}(\mu\kappa|M=m,\operatorname{sgn}(\mu\kappa)=-1)\end{aligned}\tag{D.14}$$

The entropy terms are independent of ρ, so if we interchange the order of summation we may pull them out of the sum over ρ:

$$\begin{aligned}&=\sum_{m}\mathrm{ENT}(\mu\kappa|M=m,\operatorname{sgn}(\mu\kappa)=1)\sum_{\rho}\mathbb{P}(M=m,\pi'_{(t)}=\rho,\operatorname{sgn}(\mu\kappa)=1)\\ &\quad+\sum_{m}\mathrm{ENT}(\mu\kappa|M=m,\operatorname{sgn}(\mu\kappa)=-1)\sum_{\rho}\mathbb{P}(M=m,\pi'_{(t)}=\rho,\operatorname{sgn}(\mu\kappa)=-1)\end{aligned}\tag{D.15}$$

The sums of the joint probabilities are the joint probabilities without specifying $\pi'_{(t)}$:

$$\begin{aligned}&=\sum_{m}\mathrm{ENT}(\mu\kappa|M=m,\operatorname{sgn}(\mu\kappa)=1)\mathbb{P}(M=m,\operatorname{sgn}(\mu\kappa)=1)\\ &\quad+\sum_{m}\mathrm{ENT}(\mu\kappa|M=m,\operatorname{sgn}(\mu\kappa)=-1)\mathbb{P}(M=m,\operatorname{sgn}(\mu\kappa)=-1)\end{aligned}\tag{D.16}$$

We can consolidate these two terms into a sum over $\operatorname{sgn}(\mu\kappa)$:

$$=\sum_{s=-1,1}\sum_{m}\mathrm{ENT}(\mu\kappa|M=m,\operatorname{sgn}(\mu\kappa)=s)\mathbb{P}(M=m,\operatorname{sgn}(\mu\kappa)=s)\tag{D.17}$$

But this is $\mathbb{E}(\mathrm{ENT}(\mu\kappa|M, \mathrm{sgn}(\mu\kappa)))$, so we get that

$$\mathbb{E}(\mathrm{ENT}(\mu\kappa|M, \pi'_{(t)}, \mathrm{sgn}(\mu\kappa))) = \mathbb{E}(\mathrm{ENT}(\mu\kappa|M, \mathrm{sgn}(\mu\kappa))) \tag{D.18}$$

as we claimed.

Bibliography

[1] V. Bentkus, *A lyapunov-type bound in rd*, Theory of Probability & Its Applications, 49 (2005), pp. 311–323.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory Second Edition*, John Wiley & Sons, Hoboken, New Jersey, 2006.

[3] P. Diaconis, *Chapter 4: Probabilistic Arguments*, vol. 11 of Lecture Notes–Monograph Series, Institute of Mathematical Statistics, Hayward, CA, 1988, pp. 69–91.

[4] G. F. Lawler and V. Limic, *Random walk: A modern introduction*, 2010.

[5] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov Chains and Mixing Times, second edition*, American Mathematical Society, 2017.

[6] B. Morris, *Improved mixing time bounds for the thorp shuffle and l-reversal chain*, arXiv preprint arXiv:0802.0339v1, (2008).