Università degli Studi di Milano

Facoltà di Scienze e Tecnologie

Corso di Laurea Magistrale in Fisica

Random Euclidean Bipartite Matching with concave cost functions in 1d

Relatore
Sergio Caracciolo

Candidato:
Vittorio Erba
Matricola: 896632

Correlatori
Andrea Di Gioacchino
Enrico Malatesta

Anno Accademico: 2017/18

Abstract

Euclidean matching problems have been extensively studied in their random formulation since early works by Mézard and Parisi [MP85]; the interest arises from the technical similarities that random matching problems share with disordered systems like spin glasses. Due to Euclidean constraints that correlate the random variables of the problem, random Euclidean matching is difficult to treat. In one dimension, a number of results were proven (see for example [CDS17]) for a power law cost function ($f(x) = x^p$) due to the fact that optimal matchings are independent from the specific istance of the problem in the case $p > 1$. This circumstance does not hold in the concave cost function regime ($0 < p < 1$).

In this Thesis, new simulated data for the concave cost function regime are presented and a concept of approximate optimal matching is introduced. Approximate matchings are more tractable than exact optimal matchings and have well known combinatorial properties, allowing for explicit computation of average quantities of interest.

Contents

CHAPTER 1

Motivation

The Euclidean matching problem dates back to 1781, when Gaspard Monge first introduced and studied it in his treatise [Mon81]. The problem is simply stated: given mines and deposits, what's the cheapest way to transport the extracted goods from each mine to the deposits? The natural assumption was that to transport goods there should be a cost, and that the cost should be function of the Euclidean distance of the transport route. In this formulation, the problem is a combinatorial optimization problem, i.e. an optimization problem whose domain consists in a finite set of possible optimal solutions. Such problems can be *a priori* tackled by a brute force approach; interesting problems tipically arise when the set of possible optimal solutions is of large size, typically $N!$ if N is a measure of the size of the istance of the problem (in the matching case, N could be the number of mines).

The Euclidean matching problem was soon abandoned in its combinatorial formulation due to the lack of computational power and practical uses of algorithmic approaches to the problem. Today, combinatorial Euclidean matchings can be solved in polynomial time thanks to the *Hungarian algorithm* ideated by Kuhn in 1955 [Kuh55]. The lack of interest in the combinatorial matching was balanced by a reformulation of the problem in the continuous setting. The new formulation statement is the following: consider a continuous distribution of mass to be excavated and a continuously shaped deposit ready to recieve goods; what's the optimal way to transport all the mass to the deposit? In this contiuum formulation, the problem becomes of interest both to measure theory and geometry, and is referred to as the *optimal transport problem*. The optimal transport problem was and still is a fertile field of research. As proof, the Nobel in Economics of 1975 was awarded to Kantorovič for works related to optimal transport theory, and one of the 2018 Fields medals was awarded to Figalli for results in the same field.

From a physical point of view, both the combinatorial and the continuous versions of the problem are quite uninteresting *per se*; both problems deal with the solution of specific istances of matching, aiming to derive general properties of the solutions given hypotesis on the initial data. Moreover, thanks to the already mentioned *Hun-*

garian algorithm, the combinatorial version of the problem can be practically solved in a couple of hours by a common laptop for a number of mines and deposits in the order of the tens of thousands; this satisfies every practical need for a solution to a specific istance of the problem. Still, the Euclidean matching found a new formulation that appeals researchers from statistical mechanics, in particular from the field of disordered systems. The **random Euclidean matching problem** studies the average properties of the solutions of Euclidean matching problems when the points, or the costs of expedition, are generated randomly according to some probability law. This version of the problem is not trivial at all: when points are randomly generated, their distances, and thus the expedition costs, are correlated by Euclidean constraints like the triangular inequality; average properties are then difficult to compute and study. The reason for this difficulty is mainly computational: averages of disordered systems are usually computed through the *replica trick*, that is intractable when the probability distribution of the possible configuration of the system does not factor over the degrees of freedom. The random Euclidean matching problem is now studied as a toy model to test and create techniques to tackle correlated random variables problems in physics; moreover, the fact that single istances of the problem can be easily solved through computer simulations gives experimental data to confront with.

A first step in the computation of average properties in the random Euclidean matching problem was performed thanks to the study of a simpler version of the problem, the **random-link** matching problem. In this formulation, randomness is introduced by randomly generating the expedition costs; such costs are thus non correlated, and averages are easier to be computed. This uncorrelated matching problem can be seen as an infinite dimensional version of the problem, where Euclidean constraints are not important anymore. Mézard and Parisi studied the random-link problem and proposed a perturbative study for the Euclidean version based on the random-link results in a series of papers published in the eighties ([MP85], [MP86], [MP87] and [MP88]).

The other dimensional limit in which Euclidean matching seems tractable is the one dimensional case. In recent years a number of exact results were found for one dimensional matchings (see for example [CS14], [CDS17] and [CDS18]). The key observation is that the concavity of the cost function, i.e. the function of the distance which computes the expedition costs, determines the structure of the solution of the matching in one dimension. For convex cost functions, the solution to the matching problem can be stated in a way that is independent on the particular istance of the problem. This feature allows to relate average properties of the solution to known properties of Brownian bridges, easing computations. This unfortunately does not remain true for the concave cost function case, which is the main subject of this Thesis. In fact, while some features of the solution are known (see [McC99]), they are highly dependent on the particular istance of the problem; thus no average property is known in the concave case.

The interest in concave cost function random matching problems arises also because their solutions can be interpreted as folding structures of polymeric chains (in a sense that will be made precise in Subsection 2.1.1). RNA folding is a challenging problem: it consists in the study of the equilibrium structure of the molecule given external properties like temperature. Equilibrium structures determine the functionality of the molecule, so that exposed portions of the chain are used by the cell, while hidden portions remain inactive. Recent works on RNA folding share quite a lot of vocabulary and techniques with the matching literature, such as in [NSV13] or [TN07]. Average results for concave cost function random matchings may be able to provide new tools, techniques and computational algorithms to study RNA folding.

This work of Thesis consisted first in extensive simulations of the concave cost function regime for Euclidean matching problem to study and understand important features of optimal solutions, and to collect average data. Then an approximation to the optimal matching was introduced and studied, allowing to compute approximate average properties using techniques from generating functions theory. The Thesis has the following structure:

- Chapter 2 formally introduces matching problems and reviews important results for both combinatorial and random versions of the problem;
- Chapter 3 presents new simulated data for concave cost function problems, focusing on the comparison between optimal matching and approximate matching;
- Chapter 4 develops the theory of approximate matchings in a rather general way, and presents computations for some average properties of approximate matchings; finally, computed properties are compared with simulated data.

CHAPTER 2

Random Euclidean Matching in 1d

Matching problems arise often in real life situations. What's the best way to assign jobs to workers? What's the optimal transport plan to send some goods from factories to resellers?

In this chapter, a particular version of the matching problem will be introduced, namely the monodimensional random Euclidean bipartite matching. For a review of general matching problems see [Sic16] and [D'A15].

2.1 Definition of the problem

Consider N red points $R = \{r_i\}$ and N blue points $B = \{b_i\}$, distributed on the segment $[0, 1] \subset \mathbb{R}$. Usually the coordinates are considered ordered, i.e. $i < j \implies b_i < b_j, r_i < r_j$.

A **matching** over B and R consists in assigning to each blue point one and only one red point. Two points mutually assigned in a matching are said to be a **link** of that matching, and are denoted as (b_i, r_j). In the following, the interval whose endpoints are b_i and r_j will be denoted as I_{b_i,r_j}; notice that this notation does not imply $b_i < r_j$. A matching can be seen equivalently as:

- a bijective map $\pi : B \to R$ such that $\pi(b_i) = r_j \iff r_j$ is assigned to b_i;
- a permutation of N objects π such that $\pi(i) = j \iff r_j$ is assigned to b_i.

As matchings can be seen as permutations, the total number of matchings over two differently colored sets of N points is given by $N!$, i.e. the total number of permutations of N objects.

In the matching problem, we are interested in finding an optimal matching. In order to define what an optimal matching is, to each possible matching π over given B and R a cost $E[\pi]$ is assigned. The optimal matching will be the minimum of $E[\pi]$ over all possible matchings.

One way to build the cost functional E is to assign to each possible link (b_i, r_j) a cost $w_{i,j}$, and define

$$E[\pi] := \sum_{i=1}^{N} w_{i,\pi(i)}, \tag{2.1}$$

i.e. E is the total cost of the links of the matching π. Among the various possibilities for a choice of costs $w_{i,j}$, the Euclidean version of the problem defines

$$w_{i,j} := g(|b_i - r_j|) \tag{2.2}$$

for a generic function $g : \mathbb{R} \to \mathbb{R}$, called **cost function**. In this thesis, the cost function analysed is $g(x) = x^p$, as it has well defined monotonicity and convexity properties and it is smooth for each value of p. Thus, the cost functional considered will be

$$E[\pi, p] = \sum_{i=1}^{N} |b_i - r_{\pi(i)}|^p. \tag{2.3}$$

The optimal matching will be denoted as $\tilde{\pi}(p)$ and its cost will be denoted as $\tilde{E}(p) = E[\tilde{\pi}, p]$. Notice that for $p = 0$ every matching has the same cost. Examples of optimal matchings for various values of p can be seen in Figure 2.2.

The random version of the problem assumes that costs are randomly generated following some specified procedure, and asks questions about average properties of the optimal matching (denoted with an overline, for example $\overline{E}$). The most natural way to introduce randomness in the Euclidean problem is to consider the points as randomly extracted with uniform probability on $[0, 1] \subset \mathbb{R}$. This implementation of randomness is highly non trivial. The costs generated this way are in fact correlated through Euclidean inequalities, leading to a non factorized probability density.

Another difficulty arises by the fact that typical quantitites of interest are functions of the optimal matching, which in turn is, a priori, highly dependent on the particular istance of randomness. This is a key observation: most of the known and new results presented in this thesis follow from the fact that in certain regimes and approximations the optimal matching depends weakly, or does not depend at all on the instance of randomness.

Tipical quantities of interest are:

- the average optimal cost per link $\epsilon(p, N) = \frac{1}{N}\overline{\tilde{E}(p, B, R)}$, where the dependence over the sets of random points is explicitly indicated to stress the fact that of the input of the problem, only the number of points N is significant in the averaged quantity;

- the distribution of optimal costs per link;

- the distribution of links' lenghts in the optimal solution. The lenght of a link (b_i, r_j) can be defined both in an Euclidean fashion, as the distance $|r_j - b_i|$, and

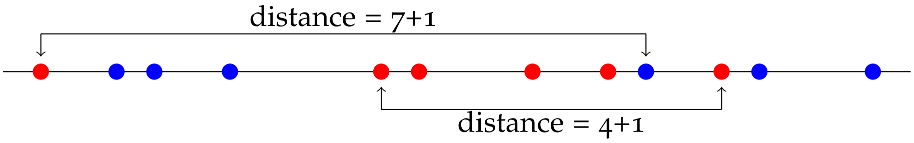


Figure 2.1: Lattice distance between points

Lattice distance between two points is computed by counting the spacings between consecutive points between the two, i.e. counting the points between the two and adding 1.

in a lattice fashion, as the number of points between b_i and r_j plus one. Examples of lattice distances are shown in Figure 2.1. Both distances give distributions worth studying.

2.1.1 Visual representations

One dimensional Euclidean matching allows for a number of different graphical representations. Here we briefly review some of them.

Link representation

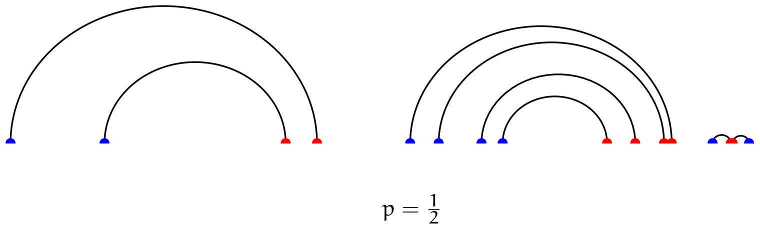
The easiest way to visualize a matching is by explicitly linking matched points using arcs. In the following, this visualization will be called the **link representation**. The convention is that every arc has to lie in the upper halfplane, if the segment over which the matching is done is embedded in $\mathbb{R}^2$ as the segment with endpoints $(0,0)$ and $(1,0)$. Figure 2.2 uses this kind of representation to depict optimal matchings.

Definition 2.1. A matching is called **non crossing** if in its link representation no arc crosses another one; such a graph is usually called a **planar** graph.

Figure 2.2 shows a non crossing matching for $p = \frac{1}{2}$.

Height diagrams

If a matching is non crossing a common way to represent it is by the use of a **height diagram**. The height diagram counts how many arcs pass above a certain point, giving informations about the overall structure of the matching in a syntetic way. To each point, one associates the number of arcs passing above it, adding or subtracting $\frac{1}{2}$ if the point itself is a leftmost (starting) or rightmost (ending) point of a link. Given the height diagram, the matching is recovered by matching contiguous points at the same height; notice that this recovery process assumes that the matching is non crossing.



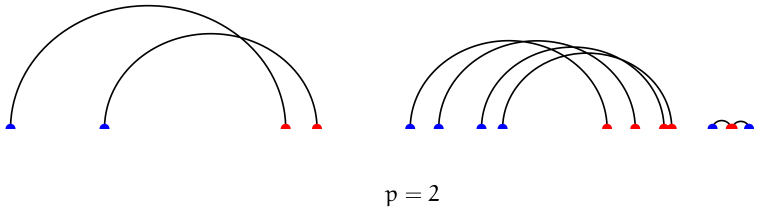


Figure 2.2: Examples of Bipartite Matchings

For $N = 8$ and $p = 0.5, 2$, the optimal matching is shown via link representation. Notice the ordering of the $p = 2$ solution and the non crossing of the $p = 0.5$ one. See Theorem 2.5 for the details of this observation.

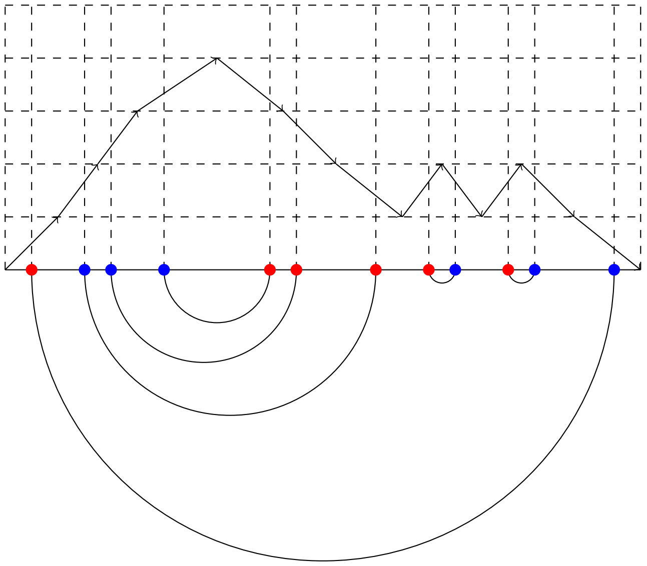


Figure 2.3: Example of the height diagram of a matching

Comparison between the link representation and the height diagram of a matching. Given a matching, the height diagram is built by drawing an "up" step over each leftmost point of a link, and a "down" step otherwise. The matching is recovered from its height diagram by linking contiguous points at the same height.

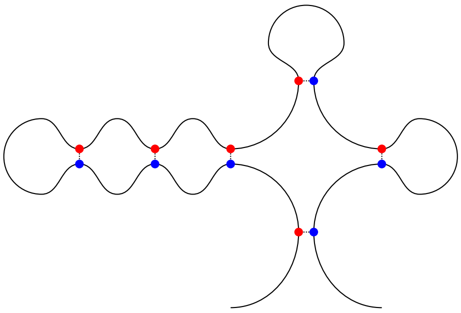


Figure 2.4: Example of chain folding induced by a matching

Example of chain folding induced by the matching represented in Figure 2.3. The segment containing the points is folded such that linked points are near in the ambient space. The folded segment provides the same amount of information as the matching itself. Notice that as we started from a non crossing matching, no pseudoknots are formed; all loops are independent and not intertwined.

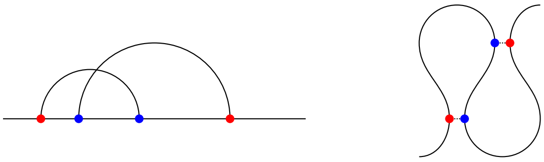


Figure 2.5: Example of pseudoknot

A crossing matching with its link and its chain fold representations. Crossing implies that a pseudoknot is formed, i.e. that two loops are intertwined.

Chain folding

Matchings give a useful instrument to describe patterns of folding of 1d chains of points. The reason is clear if a third graphical representation of matchings is introduced. Start from the link representation of a matching, then deform the $[0, 1]$ segment in a way such that linked points are superimposed. The $[0, 1]$ segment thus folds on itself, creating a structure of loops that encodes the same information of the initial matching. See Figure 2.4 for an example.

This representation motivates the interest in the study of matchings as equilibrium configurations of 1d polymeric chains. The parallel is strenghtened by the observation that in nature, RNA molecules fold on themself without creating pseudoknots, i.e. particular loop configurations; an example is given in Figure 2.5. Pseudoknots prevent the chain from properly stacking parallel subchains, thus giving energetically unfavorable configurations. These configurations are precisely described by crossing link representations. Thus, RNA seems to fold according to a non crossing matching.

2.1.2 Properties

In the following, a number of useful definitions are introduced.

Definition 2.2. A matching is called **ordered** if for any pair of links (b_1, r_1), (b_2, r_2) with $b_1 < b_2$, then $r_1 < r_2$.

Notice that, if all points are labeled in order of ascending coordinate, then the ordered matching is given by the identity permutation $\pi(i) = i$.

Definition 2.3. A matching is called **non crossing** if for any pair of links (b_1, r_1), (b_2, r_2), the intervals I_{b_1,r_1} and I_{b_2,r_2} whose endpoints are the endpoints of the links are either disjoint (i.e. $I_{b_1,r_1} \cap I_{b_2,r_2} = \emptyset$) or nested (i.e. $I_{b_1,r_1} \subset I_{b_2,r_2}$ or $I_{b_2,r_2} \subset I_{b_1,r_1}$). A matching is **crossing** if it is not non crossing. This definition is equivalent to Definition 2.1.

Definition 2.4. (Rule of three) Consider two nested links $(b_1, r_1), (b_2, r_2)$ of a matching, and suppose without loss of generality that $b_1 < r_1$ and $I_{b_2,r_2} \subset I_{b_1,r_1}$. Suppose that the two links are not equally oriented, i.e. $(r_1 - b_1)(r_2 - b_2) < 0$. With our assumptions, this restricts the mutual ordering of points to $b_1 < r_2 < b_2 < r_1$.
If $2b_2 - r_2 < r_1$ and $2r_2 - b_2 > b_1$, then the pair of links satisfies the **rule of three**. In a link representation, this means that the outer link must be large enough to enclose the inner link and two copies of it positioned just on its sides (see Figure 2.6).
If every pair of links of a matching is either disjoint or satifies the rule of three, than the matching is said to satisfy the rule of three.

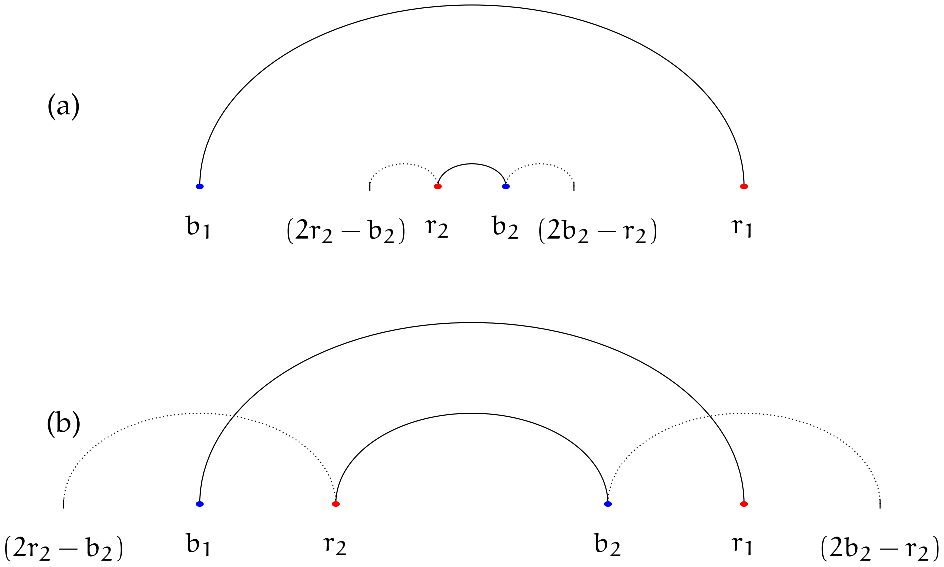


Figure 2.6: Rule of three

Examples of pairs of links (a) satisfing the rule of 3 and (b) not satisfing it.

2.2 Analytical results

A number of analytical results are available for the Euclidean Bipartite Matching in one dimension, both in its random and non random versions. Non random properties are usually the baseground onto which random results are derived.

2.2.1 Non random properties

The main analytical result about the optimal matching, at fixed positions of the points, is that for $p > 1$ the matching is ordered, and for $0 < p < 1$ the matching is non crossing. Figure 2.2 gives a graphical comparison of the two regimes for a fixed distribution of points.

Theorem 2.5. For $p > 1$, the optimal matching is ordered, and if the points are labeled in order of ascending coordinate, the optimal permutation is the identity permutation $\pi(i) = i$.
For $0 < p < 1$, the optimal matching is non crossing.

Proof. The proof has the following structure:

1. the theorem is first proven in the case $N = 2$ by direct inspection of the possible cases;

2. for $p > 1$, the local ordering of the optimal matching deduced in the $N = 2$ case immediately implies the ordering of the global matching for $N > 2$;

3. for $0 < p < 1$, there is still the need to show that each "uncrossing" operation involving two links not only lowers the cost of the matching as shown in the $N = 2$ case, but it also lowers the number of total crossings. This is again proven by explicit inspection of the possible cases;

4. for $p = 1$, point 1) and 2) guarantee that the ordered matching is optimal, but don't rule out the possibility of having other optimal matchings as well.

1) $N = 2$ **case**

Factoring out the exchange and reflection symmetries, [●●●●], [●●●●] and [●●●●] are the only possible orderings of 2 red and 2 blue points. Call T_1 the cost of the matching that links the leftmost red with the leftmost blue, and T_2 the cost of the other possible matching. By extension, call T_1 and T_2 also the relative matchings. We can proceed computing the costs of all possible matchings to find the optimal one.

First case [●●●●]: apart from a scaling factor that does not alter the reasoning, fix the position of the four points at $0, 1, 1 + x_1, 1 + x_2$, with $0 < x_1 < x_2$ (see Figure 2.7). Then:

$$\begin{aligned} T_1 &= (1 + x_1)^p + x_2^p \\ T_2 &= (1 + x_2)^p + x_1^p \end{aligned} \tag{2.4}$$

and $T_1 \leqslant T_2$ if and only if

$$f(x_1) \leqslant f(x_2) \tag{2.5}$$

for $f(x) = (1 + x)^p - x^p$. But $f(x)$ is monotone increasing for $p > 1$, constant for $p = 0, 1$ and monotone decreasing for $0 < p < 1$ by a check of its first derivative. Thus T_1 is optimal for $p > 1$, T_2 is optimal for $0 < p < 1$ and the two matchings are degenerate for $p = 0, 1$. Notice that T_1 is ordered and crossing, T_2 is non ordered and non crossing.

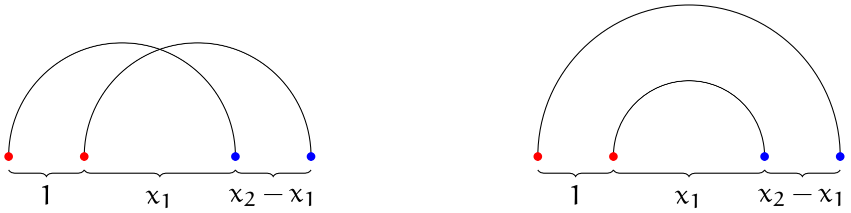


Figure 2.7: Theorem 2.5 - First case

Second case [●●●●]: apart from a scaling factor that does not alter the reasoning, fix the position of the four points at $0, 1 - x_1, 1, 1 + x_2$, with $0 < x_1 < 1$ (see Figure 2.8). Then:

$$\begin{aligned} T_1 &= (1 - x_1)^p + x_2^p \\ T_2 &= (1 + x_2)^p + x_1^p \end{aligned} \tag{2.6}$$

and $T_1 \leqslant T_2$ implies

$$(1-x_1)^p - x_1^p \leqslant (1+x_2)^p - x_2^p. \tag{2.7}$$

For $p \geqslant 1$, $T_1 \leqslant T_2$ as

$$T_1 \leqslant 1 \leqslant T_2. \tag{2.8}$$

For $p < 1$, the optimal matching is not always the same. Numerical checks confirm indeed that for different values of x_1 and x_2 T_1 can be greater, equal or less than T_2. Notice that T_1 is ordered, T_2 is not and both matchings are non crossing.

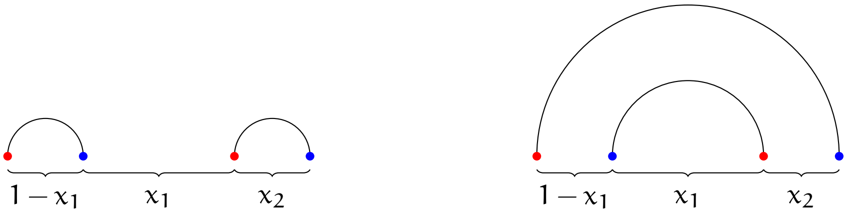


Figure 2.8: Theorem 2.5 - Second case

Third case [●●●●]: apart from a scaling factor that does not alter the reasoning, fix the position of the four points at $0, x_1, x_2, 1$, with $0 < x_1 < x_2 < 1$ (see Figure 2.9). Then:

$$\begin{aligned} T_1 &= x_1^p + (1-x_2)^p \\ T_2 &= x_2^p + (1-x_1)^p \end{aligned} \tag{2.9}$$

and $T_1 \leqslant T_2$ if and only if

$$f(x_1) \leqslant f(x_2) \tag{2.10}$$

for $f(x) = (1+x)^p + x^p$. But a check on the first derivative confirms that $f(x)$ is monotone increasing for $p > 0$ and constant for $p = 0$. Thus T_1 is optimal for $p > 0$. Notice that T_1 is ordered and non crossing, T_2 is non ordered and crossing.

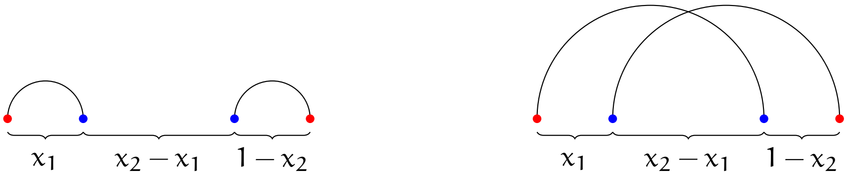


Figure 2.9: Theorem 2.5 - Third case

3) See Theorem 2 in [DSS12b].
Consider a matching with 2 crossing links (a, b) and (c, d); Figure 2.10 shows the possible crossing configurations (modulo symmetries). Uncrossing creates the two new links (a, d) and (c, b) in each of the two configurations. Consider a third link (x, y) that in the uncrossed configuration of the original links crosses (a, d) or (c, b). Its possible positions are shown in Figure 2.10 as dotted lines, both in the crossed and

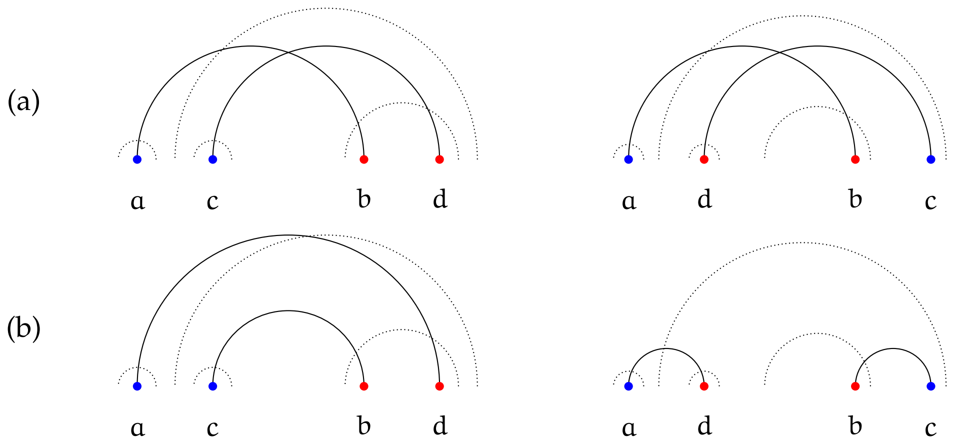


Figure 2.10: Possible crossing configurations

(a) Possible crossing configurations of two links. Dotted lines show the possible positions of a third link that crosses at least one of the original links.
(b) Uncrossed configurations of the two links.

uncrossed situation. Direct inspection confirms that uncrossing operations eliminate exactly one crossing. Thus, a finite number of uncrossings generates a non crossing matching and lowers the total matching cost for $0 < p < 1$. □

Theorem 2.5 is crucial: for $p > 1$, the optimal matching is uniquely determined independently on the positions of the points. This greatly simplifies the treatment of the random properties of the problem, as averages are decoupled from the actual construction of the optimal matching in each random instance. On the other side, for $0 < p < 1$ only non crossing is guaranteed, and unfortunately this does not suffice to uniquely determine the optimal matching. Still, a remarkable restriction of possible optimal matchings is achieved.

Finally, $p = 1$ shows degeneracy properties that place this regime at the boundary of ordered and non crossing optimal solutions.

More results are available for the $0 < p < 1$ cases.

Lemma 2.6. (Rule of three) If $0 < p < 1$, the optimal matching satifies the rule of three.

Proof. See [McC99], Lemma 2.1. Consider two links in the configuration $b_1 < r_2 < b_2 < r_1$ as in Definition 2.4. If the pair of links are part of an optimal matching, then it must be true that

$$w(b_1, r_1) + w(b_2, r_2) \leqslant w(b_1, r_2) + w(b_2, r_1). \tag{2.11}$$

Now, b_2 is the midpoint between $2b_2 - r_2$ and r_2, thus $w(b_2, r_2) = w(b_2, 2b_2 - r_2)$; by ordering, r_2 is nearer to b_1 than to $2b_2 - r_2$ giving, by monotonicity of the cost

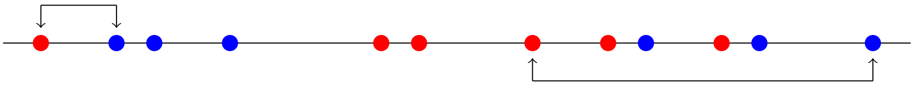


Figure 2.11: Examples of opposite neighbours

The opposite neighbour of a point is the nearest point of the opposite color, considering the lattice distance.

function, $w(b_1, r_2) < w(b_1, 2b_2 - r_2)$ (see Figure 2.6). Thus, Equation 2.11 implies

$$w(b_1, r_1) + w(b_2, 2b_2 - r_2) < w(b_2, r_1) + w(b_1, 2b_2 - r_2). \tag{2.12}$$

Equation 2.12 can be interpreted as an optimality condition for the links (b_1, r_1) and $(b_2, 2b_2 - r_2)$ over their counterparts $(b_1, 2b_2 - r_2)$ and (b_2, r_1). This optimality condition is subject to non crossing, giving that $b_1 < 2b_2 - r_2 < r_1$. This results and its symmetric counterpart for $2r_2 - b_2$ prove that the pair of links satisfies the rule of three, and by extension that the optimal matching satisfies the rule of three.

□

Lemma 2.6 describes another property of the optimal matching in the $0 < p < 1$ regime, unfortunately not sufficient to determine it uniquely. In fact, this Lemma only forbids pairs of link breaking the rule of three; nothing is said if they satisfy the rule. Moreover, the rule is quite difficult to implement in computations and, to best of our knowledge, was never succesfully used.

The last non random result presented is the key concept exploited in the following chapters to deduce analytical results in the $0 < p < 1$ regime.

Definition 2.7. (Opposite neighbour) Define the **opposite neighbour (o.n.)** of a point to be the nearest point of the other color such that between the two there is an equal number of red and blue points (possibly zero). Here nearest is to be intended in a "lattice" way, i.e. the distance between two points is $k + 1$ if between them there are k other points.

Notice that the definition is symmetric (b is o.n. of $r \iff r$ is o.n. of b).

Lemma 2.8. Each distribution of N red/blue points admits a **natural non crossing matching**, i.e. a non crossing matching uniquely determined by the mutual ordering of differently colored points.

Proof. Match every point with its opposite neighbour. This matching is uniquely determined by the color ordering and well defined thanks to the symmetry of the o.n. definition.

This matching is non crossing: given a link (b, r) of opposite neighbours, the opposite neighbours of all the points inside $I_{b,r}$ must lay inside the same interval. In fact, suppose without loss of generality that $b < r$ and consider the first point f and last point l inside $I_{b,r}$ (supposed non empty, else the statement is trivial); f must be blue, otherwise b's o.n. would be f, and l must be red for the same reason. Between f and l there is an equal number of red and blue points, thus either they are o.n., or their o.n.'s lay between them by iteration. □

Another way to introduce natural matchings is the following: consider the base points of the matching and build a height diagram by assigning an "up" step to red points and a "down" step to blue points, and by reflecting all negative portions of the graph to the positive halfplane. The matching induced by this height diagram reproduces the above defined natural matching.

Lemma 2.8 seems quite unaccomplishing: counterexamples are easily found confirming that the natural matching defined is not in general the optimal matching. For example, consider the $N = 2$ configuration [●●●●]: as seen in the proof of Theorem 2.5, both possible matchings are non crossing, and can both be optimal in different ranges of distances between the points. The natural matching selects always the ordered matching in this ambiguous case, which is not the known behaviour of the optimal matching.

Surprisingly, when simulations are run the natural matching seems quite similar to the optimal one. This is fortunate: manipulating natural matchings to extract average properties is much easier, though not trivial, than manipulating optimal matchings. Chapters 3 and 4 will focus on the comparison of simulated quantities for optimal and natural matchings, and will develop a number of results for the average properties of the natural matchings.

2.2.2 Random properties

As sketched in Subsection 2.2.1, average results are known only in the $p > 1$ case. The main point is that independently on the random positions of the points, the matching is ordered. Labeling the points in order of ascending coordinate shows that the matching is composed by all the links (b_i, r_i), $1 \leqslant i \leqslant N$.

Definition 2.9. The **transport field** of a matching π is defined as

$$\phi_i(\pi) = b_i - r_{\pi(i)}, \quad 1 \leqslant i \leqslant N. \tag{2.13}$$

For $p > 1$, $\pi = \mathbb{1}$ and $\phi_i = b_i - r_i$. The optimal cost is

$$\tilde{E}(p) = \sum_{i=1}^{N} |\phi_i|^p. \tag{2.14}$$

In [CDS17], the transport field ϕ_i is related to the difference of two Brownian processes. This connection allows for the computation of the average optimal cost in the large N limit, as well as correlation functions of the kind $\overline{\phi_i \phi_j}$. The average optimal cost is

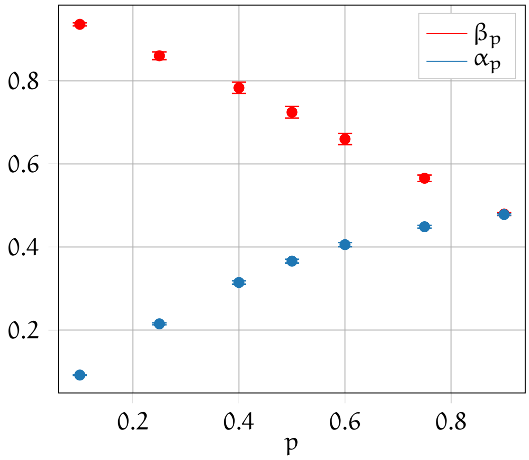
$$\epsilon(p, N) = N^{-\frac{p}{2}} \frac{\Gamma(1+\frac{p}{2})}{p+1} \left[1 + \frac{p(p+2)}{8} \frac{1}{N} + o\left(\frac{1}{N}\right)\right]. \tag{2.15}$$

2.3 Simulations

Simulations where performed by Matteo D'Achille in his thesis work [D'A15] for N up to 1000. The average optimal cost per link scaling behaviour was studied extensively, simulating up to 10000 random istances per each value of the parameters p and N. For $p > 1$, simulations are in agreement with the results of Subsection 2.2.2. For $0 < p < 1$ its scaling behaviour, i.e. the leading coefficient of the expansion in the limit of large N was extracted by fitting data against

$$\epsilon(p, N) = \frac{\beta_p}{N^{\alpha_p}}(1 + o(1)). \tag{2.16}$$

Results are shown in Figure 2.12. The dependence on p of the coefficient was then fitted against simple polynomials, with results shown in Figure 2.13. The scaling exponent is in good agreement with a parabolic behaviour $p(1 - \frac{p}{2})$, and the scaling coefficient seems to have a linear behaviour $1 - \frac{p}{2}$.



p	α_p	β_p
0.1	0.0917(9)	0.936(4)
0.25	0.215(2)	0.860(9)
0.4	0.314(4)	0.78(1)
0.5	0.366(5)	0.72(1)
0.6	0.406(5)	0.66(1)
0.75	0.449(3)	0.565(8)
0.9	0.478(2)	0.480(3)

Figure 2.12: Scaling exponent α_p and coefficient β_p

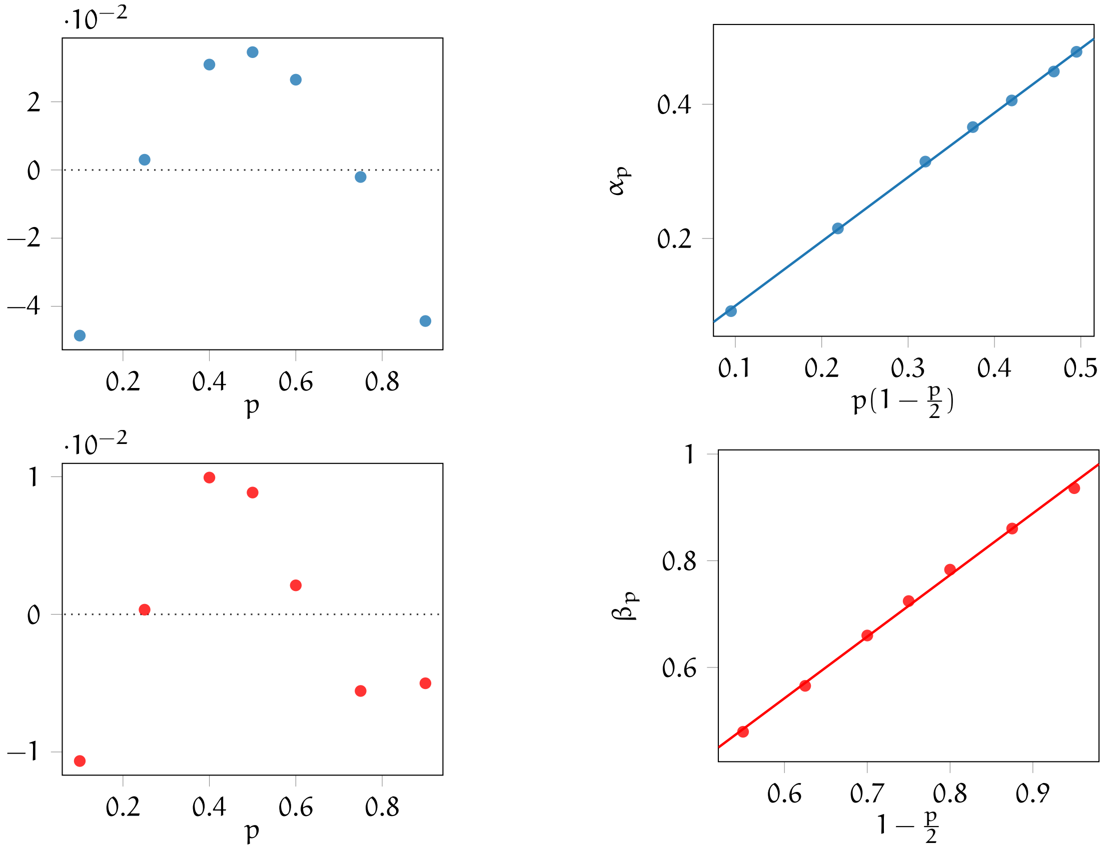


Figure 2.13: Behaviour of α_p and β_p

Plots of residuals of the linear regression for α_p vs p (top-left) and β_p vs p (bottom-left) suggest higher-than-linear dependence of parameters on p. Instead, on the right very good agreement with the ansatz $\alpha_p \sim -p(1-\frac{p}{2})$ (top-right) and $\beta_p \sim 1-\frac{p}{2}$ is found.

2.4 Variants

In this Section, a brief review on the literature regarding variants of the problem will be considered.

Random link

Euclidean costs for links are a difficult feature to treat. In fact, Euclidean correlations do not allow for the usage of mean field techniques. A more tractable version of the matching problem is then created by considering each link cost as a independent random variable distributed with a certain probability law. Such version of the problem loses any information about the underlying geometry of the space of points, resulting in a infinite range version of the matching problem. The random link matching problem was studied in [MP85] by Mezard and Parisi by usage of the so called replica trick, and the techniques developed ware later adapted to the correlated problem in [HDMM98].

Monopartite

The monopartite version of the problem considers 2N points of a single kind, and looks for the optimal matching without any color restriction. Optimal matchings found this way induce a bipartition on the points, or better 2^N distinct bipartitions: for all matched points, let one be red and the other be blue. This observation allows to study the monopartite problem as the minimum over possible bipartition, i.e. over possible colorings, of the optimal bipatite matching. Results for the $p > 1$ regime can be found in [CDS17].

Negative p

The structure of optimal matching in the $p < 0$ regime was studied in [CDS17]. In particular, the permutation describing the matching is found to satisfy certain ciclic properties, that allow to treat the problem in the same way as the $p > 1$ regime. The average optimal cost per link behaves as

$$\epsilon(p,N) = \frac{1}{2^p}\left[1 + \frac{p(p-2)(p-4)}{3(p-3)}\frac{1}{N}\right] + o\left(\frac{1}{N}\right) \tag{2.17}$$

in the large N limit.

Grid-poisson unbalanced

In [BCS14], the Grid-poisson problem was studied, in which red points are taken on an equispaced lattice, while blue points are randomly uniformly distributed. Moreover, the problem addressed was unbalanced, i.e. with different numbers of red and blue points. The focus was on the correlation function study for the problem with respect to a density parameter ρ, defined as the ration between the number of red and blue points; criticality was studied in the $\rho \approx 1$ regime.

Higher dimensions

Higher dimensions Euclidean matching problems have been studied extensively. In the convex regime, the average optimal cost per link leading behaviour in the large N limit was conjectured in [MP88] and later proven, for $d \geqslant 3$ in [Tal92]:

$$\epsilon(p,N,d) \sim N^{1-\frac{p}{d}}. \tag{2.18}$$

Further results were obtained in low dimensions for the behaviour of the first order corrections, see [CDS18] and [CLPS14].

CHAPTER 3

$0 < p < 1$ simulations

This chapter summarizes the results of simulations in the $0 < p < 1$ regime. The main reasons that motivate the simulations are:

- studying the low N behaviour of optimal matchings to gain insights on its properties and dependences;

- studying the large N limit for the average cost, to extract its leading behaviour;

- studying natural matchings (Lemma 2.8) in both N regimes to understand their relevance as approximate optimal matchings.

The first point of the simulations was carried out qualitatively, generating tens of random instance of the problem at fixed conditions, for example at fixed points positions or at fixed color ordering. Color ordering was found to be the variable with the strongest impact on the structure of optimal matchings, while not determinant. This observation motivated the introduction of natural matchings and their study.

The second and third phase of the simulations were performed thanks to the access to the **LCM farm** (`https://lcm.mi.infn.it/farm/`), which allowed for up to 1000 parallel simulations to be run at a single time. In the following sections results obtained in this phase are presented.

Given as input the pair (p, N), a random istance of the Euclidean matching problem was produced and solved. The output gave back:

- the total cost of the optimal matching and the cost of the natural one;

- the structure of computed matchings, i.e. the positions of all red and blue points, along with the permutation describing the optimal and natural matchings.

Simulations were performed in two batches:

1. given the sizable output of the structure of computed matchings, the first batch focused only on the simulation of the costs. For each value of p in $[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]$ and for a range of values of N from 50 to 6000,

between 5000 and 10000 random istances where simulated, and the costs computed. Section 3.2 presents the analysis of this dataset;

2. for the same values of p and for a range of values of N from 500 to 4000, the entire structure of computed matching was simulated and saved. Subsections following Section 3.2 present the analysis of this second dataset.

3.1 Code

The program used for the simulations is a custom made C++ code revolving around the usage of an external library for graph optimization, namely the library **LEMON** (**L**ibrary for **E**fficient **M**odeling and **O**ptimization in **N**etworks), an open source project available at `http://lemon.cs.elte.hu/trac/lemon`.

The code is straightforward: it defines a class BIPARTITE that stores all the relevant information of the problem, i.e. positions of the points, number of points, dimension, p parameter, etc... Moreover, it stores the graph structure using **LEMON** types. After random generation of the distribution of points, the weights are computed and the optimal assignement is found using an Edmond's maximum weighted perfect matching algorithm (see `http://lemon.cs.elte.hu/pub/doc/latest/a00256.html` for the details). The optimal cost is extracted by the algorithm itself. The natural matching is computed directly by building the height diagram and using it to recover the matching itself.

3.2 Average optimal cost

The first dataset was used to compute average total costs per link (from now on, we will omit the phrasings "total" and "per link"). For each fixed (p, N), all the costs produced for each random instance were averaged, and the statistical error was extracted as the standard error. Results are plotted Figure 3.1, and shown in Table 3.1 and Table 3.2.

Data was fitted against a power law leading behaviour of the kind

$$\epsilon(p, N) = \frac{\beta_p}{N^{\alpha_p}} \tag{3.1}$$

and compared with results presented in Section 2.3, i.e. with smaller N simulation data and predicted behaviours $\alpha_p = p(1 - \frac{p}{2})$ and $\beta_p = 1 - \frac{p}{2}$. Results for the fit are plotted in Figure 3.2 and shown in Table 3.3.

The parabolic behaviour of α_p starts to deform. The new data was analysed progressively removing high N data: this procedure showed that the parabolic behaviour deforms more and more as N grows, suggesting that $N = 6000$ is still not large enough to single out the leading behaviour. The behaviour of β_p is evidently non linear.

↓ p	N= 50	100	200	300	400	500	600	700	800
0.1	0.654(2)	0.615(1)	0.5768(5)	0.5554(9)	0.5405(4)	0.5292(4)	0.5200(2)	0.5124(2)	0.5059(4)
0.2	0.445(2)	0.394(9)	0.349(2)	0.324(5)	0.3077(8)	0.2954(2)	0.2855(8)	0.2774(6)	0.2706(1)
0.3	0.313(9)	0.265(3)	0.222(4)	0.199(9)	0.1853(1)	0.1744(8)	0.1663(2)	0.1596(4)	0.1540(0)
0.4	0.229(7)	0.185(8)	0.148(5)	0.130(3)	0.118(4)	0.1099(1)	0.1033(7)	0.0981(9)	0.0936(7)
0.5	0.173(8)	0.135(5)	0.104(6)	0.089(7)	0.080(3)	0.0735(7)	0.0685(4)	0.0644(9)	0.0610(8)
0.6	0.135(6)	0.103(0)	0.077(0)	0.065(0)	0.057(4)	0.0521(8)	0.0481(8)	0.0452(4)	0.0425(2)
0.7	0.108(9)	0.080(7)	0.059(3)	0.049(1)	0.043(3)	0.0392(7)	0.0359(4)	0.0332(7)	0.0313(7)
0.8	0.089(3)	0.065(0)	0.047(0)	0.038(7)	0.033(7)	0.0304(0)	0.0278(3)	0.0257(7)	0.0242(2)
0.9	0.074(3)	0.053(1)	0.037(9)	0.031(2)	0.0270(0)	0.0243(7)	0.0224(0)	0.0206(8)	0.0193(8)
↓ p	N= 900	1000	1500	2000	2500	3000	4000	5000	6000
0.1	0.5002(7)	0.4952(2)	0.4761(1)	0.4629(8)	0.4530(8)	0.4451(3)	0.4326(9)	0.4233(9)	0.4158(7)
0.2	0.2647(1)	0.2595(1)	0.2403(5)	0.2276(2)	0.2181(3)	0.2107(0)	0.1993(9)	0.1910(2)	0.1843(5)
0.3	0.1492(2)	0.1450(3)	0.1298(6)	0.1198(2)	0.1127(1)	0.1072(4)	0.0990(1)	0.0930(2)	0.0884(2)
0.4	0.0899(7)	0.0869(3)	0.0754(3)	0.0683(9)	0.0632(3)	0.0593(3)	0.0536(0)	0.0495(8)	0.0465(1)
0.5	0.0583(6)	0.0560(3)	0.0474(9)	0.0423(4)	0.0386(5)	0.0359(4)	0.0318(6)	0.0291(5)	0.0269(8)
0.6	0.0403(7)	0.0385(0)	0.0323(1)	0.0284(3)	0.0256(5)	0.0237(1)	0.0208(9)	0.0188(0)	0.0173(3)
0.7	0.0296(6)	0.0283(1)	0.0234(1)	0.0203(8)	0.0183(9)	0.0168(6)	0.0147(5)	0.0131(8)	0.0121(4)
0.8	0.0228(7)	0.0217(7)	0.0179(2)	0.0155(6)	0.0139(4)	0.0128(5)	0.0111(4)	0.0099(2)	0.0091(0)
0.9	0.0182(1)	0.0173(0)	0.0141(8)	0.0123(8)	0.0110(0)	0.0100(4)	0.0087(4)	0.0078(4)	0.0071(9)

Table 3.1: Average total cost for the **optimal** matching

↓ p	N= 50	100	200	300	400	500	600	700	800
0.1	0.675(4)	0.639(1)	0.603(1)	0.5828(3)	0.5684(2)	0.5573(4)	0.5483(1)	0.5408(7)	0.5345(0)
0.2	0.470(1)	0.422(2)	0.377(8)	0.353(4)	0.336(7)	0.324(3)	0.3143(6)	0.3061(7)	0.2990(6)
0.3	0.335(6)	0.288(5)	0.245(8)	0.223(0)	0.208(2)	0.196(9)	0.188(4)	0.1814(8)	0.1755(0)
0.4	0.246(4)	0.203(1)	0.165(5)	0.146(8)	0.134(4)	0.125(4)	0.118(5)	0.1129(8)	0.1082(0)
0.5	0.185(5)	0.147(2)	0.115(8)	0.100(4)	0.090(6)	0.083(5)	0.078(2)	0.0738(3)	0.0702(1)
0.6	0.143(0)	0.110(3)	0.083(9)	0.071(4)	0.063(5)	0.058(1)	0.0538(3)	0.0507(1)	0.0478(2)
0.7	0.113(0)	0.084(7)	0.063(0)	0.052(6)	0.046(6)	0.042(4)	0.0389(3)	0.0361(3)	0.0341(5)
0.8	0.091(1)	0.066(7)	0.048(6)	0.040(2)	0.035(1)	0.0317(7)	0.0291(4)	0.0270(1)	0.0254(2)
0.9	0.074(8)	0.053(6)	0.038(3)	0.031(6)	0.027(4)	0.0247(2)	0.0227(5)	0.0210(2)	0.0197(0)
↓ p	N= 900	1000	1500	2000	2500	3000	4000	5000	6000
0.1	0.5289(7)	0.5239(3)	0.5049(5)	0.4918(0)	0.4818(9)	0.4737(5)	0.4611(5)	0.4517(0)	0.4439(9)
0.2	0.2930(6)	0.2876(4)	0.2677(1)	0.2545(4)	0.2445(2)	0.2366(9)	0.2246(0)	0.2157(2)	0.2085(1)
0.3	0.1704(8)	0.1660(1)	0.1498(6)	0.1389(7)	0.1312(9)	0.1252(9)	0.1162(5)	0.1096(3)	0.1045(1)
0.4	0.1041(9)	0.1009(4)	0.0884(3)	0.0807(4)	0.0750(2)	0.0706(7)	0.0642(2)	0.0597(1)	0.0562(2)
0.5	0.0672(9)	0.0647(4)	0.0554(1)	0.0497(5)	0.0456(8)	0.0426(4)	0.0380(5)	0.0349(9)	0.0325(0)
0.6	0.0455(1)	0.0435(1)	0.0368(3)	0.0326(3)	0.0295(6)	0.0274(3)	0.0243(0)	0.0219(6)	0.0203(1)
0.7	0.0323(6)	0.0309(3)	0.0257(5)	0.0225(2)	0.0203(9)	0.0187(4)	0.0164(6)	0.0147(5)	0.0136(2)
0.8	0.0240(5)	0.0229(0)	0.0189(3)	0.0164(8)	0.0147(9)	0.0136(6)	0.0118(6)	0.0105(9)	0.0097(3)
0.9	0.0185(2)	0.0176(0)	0.0144(5)	0.0126(2)	0.0112(2)	0.0102(5)	0.0089(3)	0.0080(2)	0.0073(6)

Table 3.2: Average total cost for the **natural** matching

p=	0.1	0.2	0.3	0.4	0.5
α_p optimal	0.096(4)	0.187(6)	0.27(0)	0.33(9)	0.39(4)
α_p natural	0.090(1)	0.17(4)	0.24(9)	0.31(4)	0.36(8)
β_p optimal	0.9626(9)	0.946(2)	0.930(4)	0.89(8)	0.84(3)
β_p natural	0.9742(6)	0.953(3)	0.923(3)	0.87(8)	0.81(7)
p=	0.6	0.7	0.8	0.9	
α_p optimal	0.43(3)	0.46(1)	0.47(8)	0.489(0)	
α_p natural	0.41(1)	0.44(5)	0.46(8)	0.485(4)	
β_p optimal	0.76(4)	0.68(0)	0.58(9)	0.50(7)	
β_p natural	0.74(0)	0.66(3)	0.57(9)	0.50(3)	

Table 3.3: Fit results

The comparison between optimal and natural matching results looks promising: as $N = 6000$ isn't large enough to single out the leading behaviour of the cost, it's still possible that both matchings share the same leading behaviour.

3.3 Distribution of links' lenghts

The distribution of links' lenghts was computed by averaging over 1000 istances of disorder the histogram of discrete links' lenghts. Discrete lenghts are to be intended as distances in lattice units, i.e. two points are at distance $2l + 1$ id they have $2l$ other points between them.

An example of links' lenghts distribution for optimal matchings can be found in Figure 3.3, and for natural matchings in Figure 3.4. Compared to the optimal matchings distribution, the natural matchings one seems to have no dependence on the value of p. The tail shows a different behaviour too, indicating a significative lack of longer links in natural matchings. This can be explained considering that natural matchings favor ordered non crossing matchings whenever possible, and ordered matching have shorter links. Nevertheless, in the range $[0, 0.5]$ the distributions for optimal and natural matchings seem to agree. A comparison at fixed values of p can be found in Figure 3.5

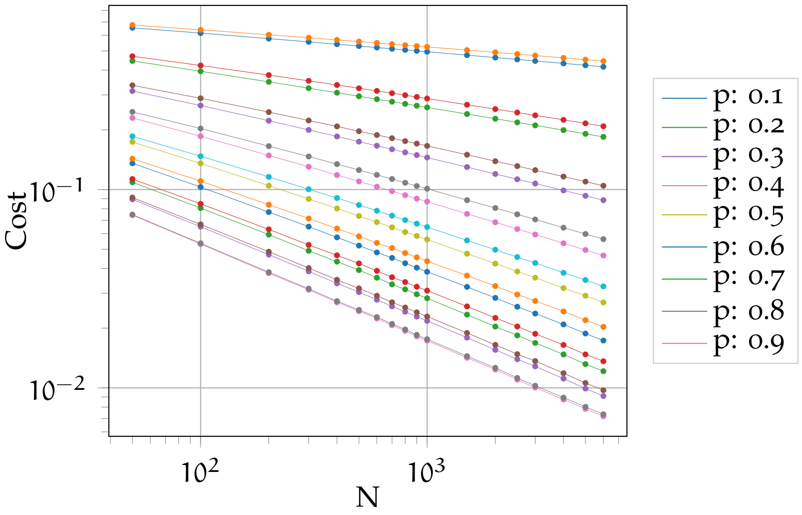


Figure 3.1: Average total cost for optimal and natural matchings

Each pair of lines describes the behaviour at a different value of p. For each pair of lines, the below one represents the average total cost of the optimal matching, the other the average total cost of the natural matching. Each data line shows linearity, suggesting that the power law fit is viable.

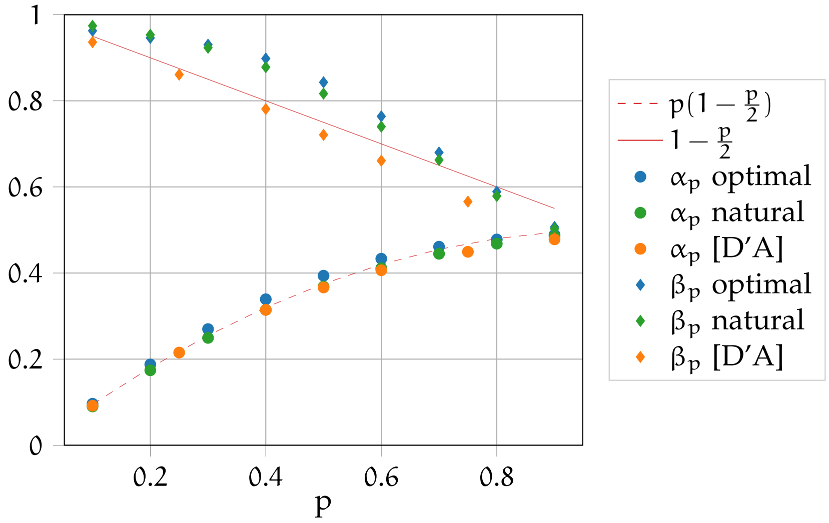


Figure 3.2: Fit results

Power law fit results for the cost of optimal and natural matchings, compared with previous results by D'Achille [D'A15].

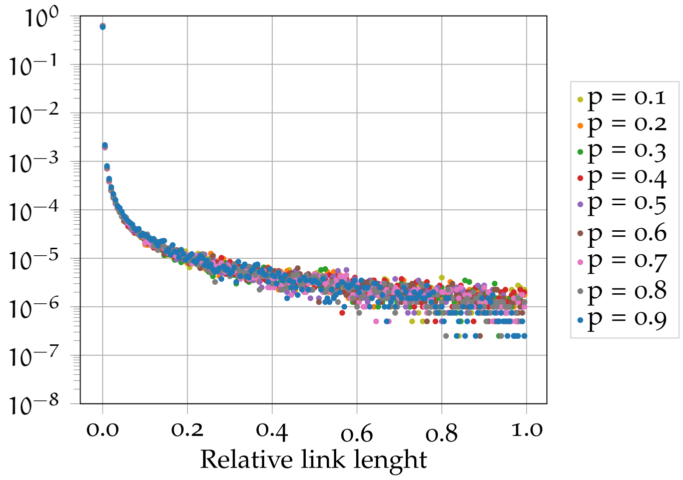


Figure 3.3: Links' lenght distribution at $N = 4000$ for optimal matchings

The distribution of links' lengths was computed by averaging the links' lenghts histogram over the 1000 random istances generated. A further normalization factor of $\frac{1}{N}$ was included such that the area of the histogram equals the total number of links per istance, i.e. N. To be able to plot the distributions at $N = 4000$, a filtering process was performed, discarding all datapoints but 200 equispaced ones.

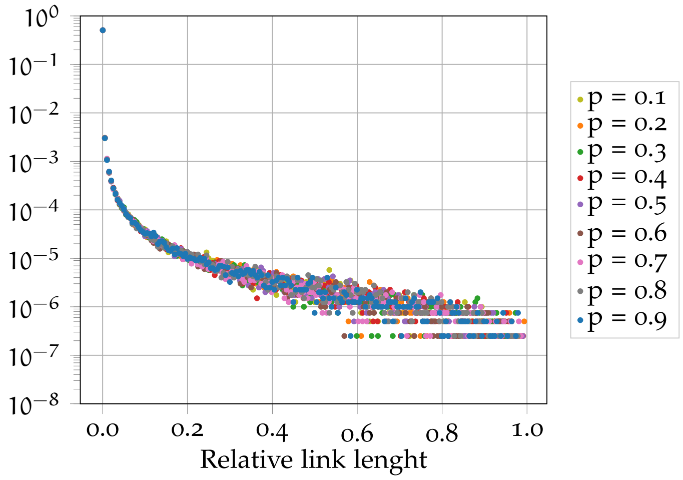


Figure 3.4: Links' lenght distribution at $N = 4000$ for natural matchings

The distribution for natural matchings was extracted and normalized the same way as Figure 3.3. Despite the coarseness of the tail of the plot, a different behaviour can be observed.

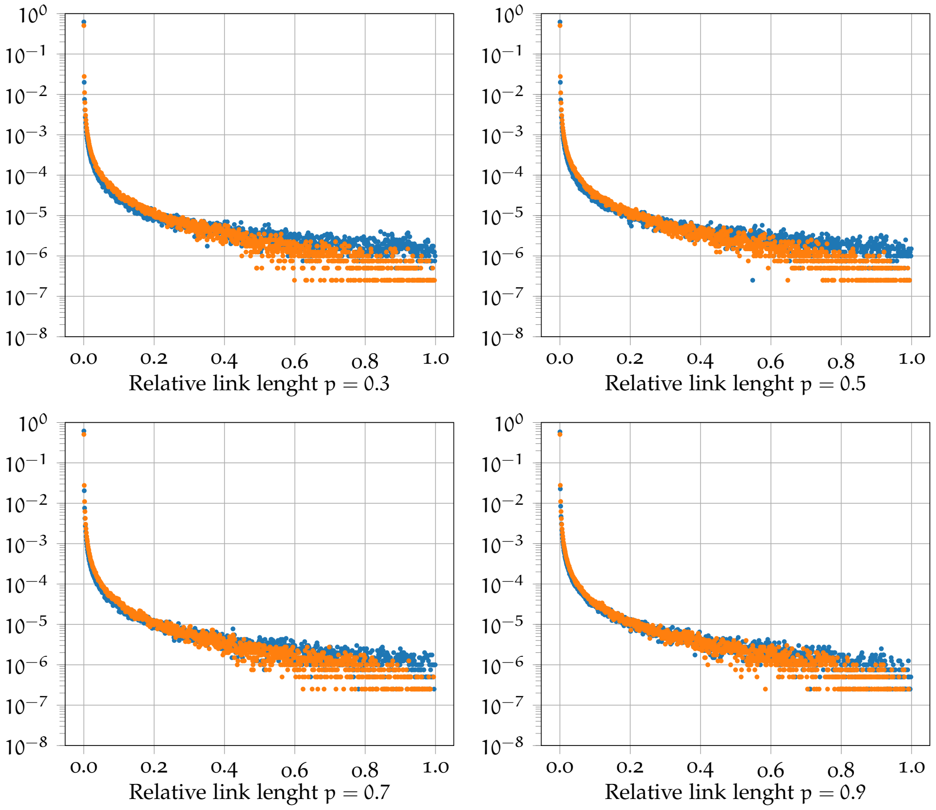


Figure 3.5: Links' lenght distribution at $N = 4000$, $p = [0.3, 0.5, 0.7, 0.9]$ for both optimal and natural matchings

For each value of p, the blue plot is the distribution for optimal matchings, the orange plot fot natural matchings. A different behaviour in tha tail of the distribution is present.

CHAPTER 4

Combinatorics of natural matchings

In this chapter the main analytical results obtained in this work of thesis are presented.

The observation that natural matchings and optimal matchings share similar properties motivated the interest in the analytical study of natural matchings. Contrary to the optimal ones, natural matchings depend only on the color ordering of the points, and not on their actual positions. This allows to study random averages in two independent steps: an average over color ordering, on which the natural matching depends, and an average over points' positions, that determines only the average separation between two points, knowing that they have $2l$ points between them.

First, colored sequences and their natural matchings are studied, counting properties of interest are presented as well as technical details that will be useful in the following. Then the actual expression for the average cost is given, analyzed and manipulated to obtain closed expressions for the scaling behaviour of the cost. Finally the analytical data is confronted with the simulations.

4.1 Paths and bridges

The combinatorics of natural matchings relies heavily on the combinatorics of sequences of colored points.

Definition 4.1. A **bridge** is a sequence composed by an equal number of two kinds of letters, say u and d. In the following, u will be associated with red points, d with blue points.
A **Dyck path** is a bridge such that any of its left subsequences (i.e. contiguous subsequences starting from the leftmost letter) contains at least as many u letters as d letters.

In the literature, a useful graphical representation for Dyck paths is found (see Figure 4.1): consider the two dimensional lattice of integer coordinates, and represent each u letter as a vector moving toward the nearest north-east lattice point, each d letter as a vector moving towards the nearest south-east lattice point (respectively, an *up* and

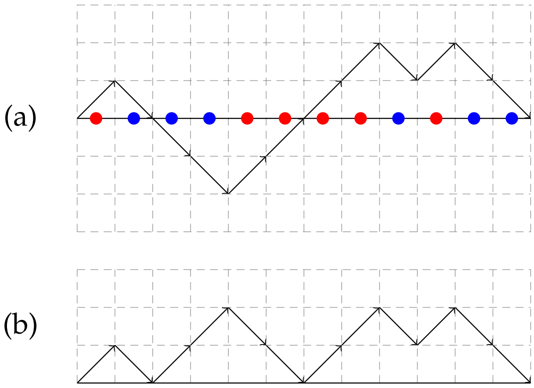


Figure 4.1: Examples of Dyck paths and bridges

(a) A bridge fo lenght $N = 6$. Contrary to Dyck paths, bridges has no height restriction. The sequence of colored points in bijection with the bridge is drawn on the horizontal axis, with the choice that red points are up steps.
(b) A Dyck path of lenght $N = 6$. Notice that the profile always lies above the "horizon". This Dyck path is the Dyck path assigned to the bridge in (a), and can be interpreted as an height diagram for a non crossing matching.

a *down* movement). Then bridges are paths of up and down movements constrained to start and finish at the same height (y-coordinate value). Dyck paths are bridges that never fall below the height of the starting point. In the following:

- Dyck paths will be called just paths for the sake of brevity, where no other indication is given;
- paths and bridges will be assumed to start at the origin of the lattice;
- paths and bridges will be said to be of lenght N if they are composed by $2N$ letters;
- the set of bridges of lenght N will be denoted as $\mathcal{B}_N$, and the set of paths of lenght N will be denoted as $\mathcal{C}_N$.

Dyck paths arise in the study of **first excursions** of bridges, i.e. portion of a bridge between its first up step and its first step to arrive at height zero. First excursions are a powerful way to decompose paths and bridges in order to find recursions and formulas about their properties. Notice that, by reflecting all the portions of a bridge that fall below the horizontal axis above, a Dyck path is recovered. In this sense, each Dyck path corresponds to a family of 2^k bridges, where k is the number of zeroes of the path, including the starting point and excluding the arrival.

Bridges are important because they are in bijection with the color ordering of the points in the matching problem, ordering that determines the natural matching over the sequence. Moreover, consider the Dyck path assigned to a bridge. This path can be interpreted naturally as an height diagram representing some non crossing matching over equispaced points. This matching is precisely the natural matching assigned to the colored point sequence in bijection with the starting bridge. In this sense, to every bridge a natural matching is assigned.

4.2 Technical preliminaries

Definition 4.2. Let $x \in \mathbb{C}$, $N \in \mathbb{N}$.
The **falling factorial** is defined as $x^{\underline{N}} = x(x-1)\dots(x-N+1)$.
The **rising factorial** is defined as $x^{\overline{N}} = x(x+1)\dots(x+N-1)$.
The notation follows [GKPL89].

Lemma 4.3. Properties of rising and falling factorials

1. $x^{\underline{N}} = N!\binom{x}{N}$
2. $x^{\overline{N}} = N!\binom{x+N-1}{N}$
3. $x^{\overline{N}} = (x+N-1)^{\underline{N}}$
4. $x^{\overline{N}} = (-)^N(-x)^{\underline{N}}$
5. $x^{\overline{N}} = \frac{\Gamma(x+N)}{\Gamma(x)}$
6. $x^{\underline{N}} = \frac{\Gamma(x+1)}{\Gamma(x-N+1)}$

Proof. Properties 1), 2), 3), 4) follow from the definition.
Properties 5), 6) follow from the definition and the factorial property of the Euler gamma function. These two properties extend the definition of the rising and falling factorial to complex N. □

Lemma 4.4. Generalized binomial theorem

$$(1-x)^a = \sum_{n=0}^{\infty} \frac{\Gamma(n-a)}{\Gamma(-a)} \frac{x^n}{n!} \qquad \forall a \in \mathbb{R}/\mathbb{N}. \tag{4.1}$$

Proof. Follows from the straightforward generalization of the binomial theorem to real exponents, along with the use of properties 1) and 4) of Lemma 4.3. □

A further generalization of binomial theorem will be needed:

Lemma 4.5.

$$\log(1-x)\,(1-x)^a = \sum_{k=0}^{\infty} \frac{x^k}{k!} \frac{\Gamma(k-a)}{\Gamma(-a)} \left[\psi_0(-a) - \psi_0(k-a)\right] \tag{4.2}$$

where $\psi_0(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ is the digamma function.

Proof. Using Lemma 4.4:

$$\begin{aligned}\log(1-x)\,(1-x)^a &= \lim_{\epsilon\to 0}\frac{(1-x)^\epsilon-1}{\epsilon}(1-x)^a\\ &= \lim_{\epsilon\to 0}\frac{1}{\epsilon}\left[(1-x)^{\epsilon+a}-(1-x)^a\right]\\ &= \sum_{k=0}^{\infty}\frac{x^k}{k!}\lim_{\epsilon\to 0}\frac{1}{\epsilon}\left[\frac{\Gamma(k-a-\epsilon)}{\Gamma(-a-\epsilon)}-\frac{\Gamma(k-a)}{\Gamma(-a)}\right]\\ &= \sum_{k=0}^{\infty}\frac{x^k}{k!}\lim_{\epsilon\to 0}\frac{1}{\epsilon}\left[\epsilon\frac{\Gamma(k-a)}{\Gamma(-a)}\left(\psi_0(-a)-\psi_0(k-a)\right)+o(\epsilon)\right]\\ &= \sum_{k=0}^{\infty}\frac{x^k}{k!}\frac{\Gamma(k-a)}{\Gamma(-a)}\left[\psi_0(-a)-\psi_0(k-a)\right].\end{aligned} \tag{4.3}$$

□

Lemma 4.6. Properties of central binomials
Let $B_N=\binom{2N}{N}$ be the central binomial of order N. Then:

1. $(1-4z)^{-\frac{1}{2}}=\sum_{n=0}^{\infty}B_n z^n$
2. $\sum_{m=0}^{n}B_m B_{n-m}=4^N$
3. $B_N=\frac{4^N}{\sqrt{\pi N}}(1+o(N^{-1}))$

Proof. 1) Follows from Lemma 4.4 using the duplication formula for the Γ function. In the following $B(z)=(1-4z)^{-\frac{1}{2}}$.
2) Follows from property 1). In fact:

$$\begin{aligned}\sum_{n=0}^{\infty}\sum_{m=0}^{n}B_m B_{n-m}z^n &= \sum_{m=0}^{\infty}\sum_{n=m}^{\infty}B_m B_{n-m}z^m z^{n-m}=[B(z)]^2\\ &=\frac{1}{1-4z}=\sum_{n=0}4^n z^n.\end{aligned} \tag{4.4}$$

3) Follows using Stirling's approximation for factorials. □

Definition 4.7. Generating function
Let a_N be a sequence. It's generating function is defined as $a(z)=\sum_{N=0}^{\infty}a_N z^N$.

Property 1) of Lemma 4.6 gives a closed expression for the generating function of central binomials. Generating functions are to be considered formal series, possibly with null radius of convergence. If they converge in a finite disk around the origin, than in that disk they are considered analytical functions. In the following, every sequence not defined for negative index will be supposed to satisfy $a_{-N}=0$ for all $N\geqslant 1$.

4.3 Counting properties

A typical problem in combinatorics is to count how many objects of a certain kind there are, or how many objects of a certain kind satisfy some property. In our case, as bridges represent natural matchings, counting techniques can allow the characterization of properties of the ensemble such as its size, the average distribution of links' lenghts in a natural matching, and more.

4.3.1 Number of bridges

The total number of bridges of size N, B_N, can be found simply by considering the possible ways to select N letters between $2N$ to be u's, letting all the others be d's. Then

$$B_N = \binom{2N}{N} \tag{4.5}$$

the combinations of $2N$ objects of class N, also called $N-th$ **central binomial**.

4.3.2 Number of paths

A more sofisticate technique is required to compute the number of Dyck paths of size N, C_N: the **first return decomposition**. Consider a path of size N such that it has its first zero after $2m+2$ steps (notice that only after an even number of steps there can be a zero). Its first portion, to the first zero, is called **first excursion**; the rest of the path is called the **tail**. Then, we can write that

$$C_N = \sum_{m=0}^{N-1} C_m C_{N-1-m} \qquad N \geqslant 1 \tag{4.6}$$

meaning that a path is always composed by two smaller paths, the first excursion and the tail, possibly empty (when $m = N-1$). The recursion written has starting condition $C_0 = 1$, meaning that the only path with no letters is the empty one. The initial condition can be enforced by adding a Kroneker delta $\delta_{n,0}$ in the recursion, validating it for $N \geqslant 0$. Recursions are often solved introducing a generating function $C(z) = \sum_{N=0}^{\infty} C_N z^N$. The recursion implies, by multiplying by z^N, summing and paying attention to the particular case $N = 0$:

$$\begin{aligned} C(z) = \sum_{N=0}^{\infty} C_N z^N &= 1 + \sum_{N=1}^{\infty} \sum_{m=0}^{N-1} C_m C_{N-m-1} z^N \\ &= 1 + z \sum_{m=0}^{\infty} \sum_{N=m+1}^{\infty} C_m z^m C_{N-m-1} z^{N-m-1} \\ &= 1 + z \sum_{m=0}^{\infty} C_m z^m \sum_{i=0} C_i z^i \\ &= 1 + zC(z)^2. \end{aligned} \tag{4.7}$$

The equation for $C(z)$ leads to

$$C(z) = \frac{1 \pm \sqrt{1-4z}}{2z}. \tag{4.8}$$

The *plus* solution diverges in $z = 0$, leading to $C_0 = \infty$ and thus not respecting the initial condition of the recursion. The other solution gives $C_0 = 1$ after removal of the discontinuity, and allows to compute the coefficients C_N explicitly by series expansion using Lemma 4.4:

$$\begin{aligned} C(z) &= \frac{1}{2z}\left(1 - \sum_{N=0}^{\infty} \frac{\Gamma\left(N-\frac{1}{2}\right)}{\Gamma\left(-\frac{1}{2}\right)N!}(4z)^N\right) \\ &= \frac{1}{2z}\left(1 - 1 - \sum_{N=1}^{\infty} \frac{2\sqrt{\pi}4^{-N}\Gamma(2N)}{\frac{1}{2}(2N-1)\Gamma(N)} \frac{1}{-\frac{\sqrt{\pi}}{2}N!}(4z)^N\right) \\ &= \sum_{N=1}^{\infty} \frac{\sqrt{\pi}(2N-2)!}{\sqrt{\pi}(N-1)!N!} z^{N-1} = \sum_{N=0}^{\infty} \frac{(2N)!}{(N+1)!N!} z^N \\ &= \sum_{N=0}^{\infty} \frac{B_N}{N+1} z^N \end{aligned} \tag{4.9}$$

where the duplication formula for the Γ function was used, as well as its value at $\frac{1}{2}$. Thus:

$$C_N = \frac{1}{N+1} B_N. \tag{4.10}$$

The numbers found are known in the literature and widely studied as the **Catalan numbers** (https://oeis.org/A000108). These numbers count a variety of different combinatorial objects, justifing the interest in results concering them. Among interesting results, there are a number of parameters according to which Dyck path can be counted, as peaks, returns, and others. A complete review on counting methods and results is [Deu99].

Lemma 4.8. Properties of Catalan numbers

1. $C_{N+1} = \frac{2(2N+1)}{N+2} C_N$
2. $C_N = \frac{4^N \Gamma\left(N+\frac{1}{2}\right)}{\sqrt{\pi}\,\Gamma(N+2)}$
3. $C_N = \frac{4^N}{\sqrt{\pi}\,N^{\frac{3}{2}}}(1 + o(N^{-1}))$

Proof. 1) Follows from the definition.
2) Follows from the definition, using the duplication formula for the Γ function. This provides an analytical continuation for Catalan numbers.
3) Follows from the definition, using Stirling's approximation for factorials. □

4.4 Links' lenghts distribution

Among the various properties according to which bridges and paths can be counted and characterized, the following is of relevance for a later exstimate of the average cost of natural matchings. It's the **links' lenghts distribution**, i.e. the average number of links of a certain lenght in natural matchings built using paths or bridges. In this section, lenght will be used to mean "the number of lattice spacings", imagining each point fixed on a equispaced lattice of step 1. A link of lenght $2l+1$ is therefore a link such that between its two endpoints there are $2l$ other points. Non crossing forbids even lenght links.

The average distribution can be computed as follows: consider a bridge b of lenght N, such that its natural matching has $n_l(b)$ links of lenght $2l+1$. Then the average distribution $L_N^{\mathcal{B}}(l)$ is:

$$L_N^{\mathcal{B}}(l) = \frac{1}{B_N} \sum_{b \in \mathcal{B}_N} n_l(b) = \frac{1}{B_N} \nu_{N,l} \tag{4.11}$$

where $\nu_{N,l}$ is the total number of links of lenght $2l+1$ in natural matchings of all bridges. An analogous result holds for paths, where the numbers $\nu_{N,l}$ are called $r_{N,l}$.

This distribution is important because will allow us to write an explicit expression for the average cost of natural matchings. The computation will be performed for paths first, as the technique is instructive and useful for the bridge case.

4.4.1 Paths

The total number of link of lenght $2l+1$ in paths of lenght $2N$ will be called $r_{N,l}$.

As for counting paths, the first return decomposition is useful to count the total number of links of lenght $2l+1$. A recursion can be written:

$$\begin{aligned} r_{N,l} &= C_l C_{N-l-1} + \sum_{m=0}^{N-1} [r_{m,l} C_{N-m-1} + C_m r_{N-m-1,l}] \\ &= C_l C_{N-l-1} + 2 \sum_{m=0}^{N-1} r_{m,l} C_{N-m-1} \\ &= C_l C_{N-l-1} + 2 \sum_{m=l+1}^{N-1} r_{m,l} C_{N-m-1}, \end{aligned} \tag{4.12}$$

where $2m+2$ is to be interpreted as the position of the first zero of the path. The logic is the following:

- the first term counts all the paths in which the link between the first step and the first zero is of the required lenght. The molteplicity of paths in which this situation arise is given by all the possible paths composing the first excursion times all the possible paths composing the tail;

- the sum counts, for all the possible positions of the first zero, the possible links of the required lenght hidden in the first excursion or in the tail of the path. To count links of the required lenght hidden in the first excursion, one can use $r_{m,l}$ itself, times all the possible tails C_{N-m-1}. The tail case is symmetric; the second passage exploits this symmetry.

The starting condition for this recursion is that $r_{N,l} = 0$ for $l \geqslant N$: short paths cannot have long links. The starting condition is used in the third line, where it simplifies the lower bound of the sum.

The recursion is solved by noting first that $r_{N,l} = C_l \tilde{R}_{N,l}$: all $r_{N,l}$ are divisible by C_l by induction. The remaining recursion for the ratio $\tilde{R}_{N,l}$ is better analysed by shifting its dependence from (N, l) to $(N-l-1, l)$, i.e. defining $R_{N-l-1,l} = \tilde{R}_{N,l}$; the recursion becomes:

$$R_{s,l} = C_s + 2\sum_{m=1}^{s} R_{m-1,l}C_{s-m} \qquad s \geqslant 0, \tag{4.13}$$

where we see that as for all l the initial condition of the recursion is $R_{0,l} = 0$, $R_{N-l-1,l}$ is really just function of $s = N - l - 1$.

The recursion in Equation 4.13 can be solved introducing a generating function $R(z) = \sum_{s=0}^{\infty} R_s z^s$. The recursion implies

$$\begin{aligned}
R(z) = \sum_{s=0}^{\infty} R_s z^s &= \sum_{s=0}^{\infty} C_s z^s + 2\sum_{s=0}^{\infty}\sum_{m=1}^{s} R_{m-1}C_{s-m}z^s \\
&= C(z) + 2\sum_{m=1}^{\infty}\sum_{s=m}^{\infty} R_{m-1}C_{s-m}z^s \\
&= C(z) + 2\sum_{n=0}^{\infty}\sum_{k=0}^{\infty} R_n C_k z^{k+n+1} \\
&= C(z) + 2zC(z)R(z).
\end{aligned} \tag{4.14}$$

Thus $R(z) = \frac{1}{2z}((1-4z)^{-\frac{1}{2}} - 1)$. Lemma 4.4 allows to expand $R(z)$:

$$\begin{aligned}
R(z) &= \frac{1}{2z}\left(\sum_{m=0}^{\infty} \frac{\Gamma\left(m+\frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)}\frac{4^m z^m}{m!} - 1\right) = \frac{1}{2z}\sum_{m=1}^{\infty} \frac{\Gamma\left(m+\frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)}\frac{4^m z^m}{m!} \\
&= \sum_{m=1}^{\infty} \frac{2^{1-2m}\sqrt{\pi}\,\Gamma(2m)}{2\sqrt{\pi}\,\Gamma(m)}\frac{4^m z^{m-1}}{m!} = \sum_{m=0}^{\infty} \frac{\Gamma(2m+2)}{\Gamma(m+1)}\frac{z^m}{(m+1)!} \\
&= \sum_{m=0}^{\infty} \frac{(2m+1)!}{m!(m+1)!}z^m = \sum_{m=0}^{\infty} \frac{(2m+2)!}{2(m+1)!(m+1)!}z^m \\
&= \sum_{m=0}^{\infty} \frac{B_{m+1}}{2}z^m,
\end{aligned} \tag{4.15}$$

obtaining $R_{N-l-1} = \frac{1}{2}B_{N-l}$, and $r_{N,l} = \frac{1}{2}C_l B_{N-l}$.

The average distribution of links' lenghts is

$$L_N^e(l) = \frac{C_l B_{N-l}}{2C_N}. \tag{4.16}$$

4.4.2 Bridges

The total number of link of lenght $2l+1$ in paths of lenght $2N$ will be called $\nu_{N,l}$.

The first return decomposition with first zero at step $2m+2$ can be used in this case too, leading to the recursion

$$\begin{aligned}\nu_{N,l} &= 2\left[C_l B_{N-l-1} + \sum_{m=0}^{N-1}\left(r_{N,m}B_{N-m-1} + C_m \nu_{N-m-1,l}\right)\right]\\ &= 2C_l B_{N-l-1} + \sum_{m=l+1}^{N-1} C_l B_{m-l} B_{N-m-1} + 2\sum_{m=0}^{N-1} C_{N-m-1}\nu_{m,l}\\ &= 2C_l B_{N-l-1} + C_l \sum_{m=1}^{N-l-1} B_m B_{N-l-1-m} + 2\sum_{m=l+1}^{N-1} C_{N-m-1}\nu_{m,l}\\ &= 2C_l B_{N-l-1} + C_l \sum_{m=1}^{N-l-1} B_m B_{N-l-1-m} + 2\sum_{m=1}^{N-l-1} C_{N-l-1-m}\nu_{m+l,l}\end{aligned} \tag{4.17}$$

The logic is the following:

- the bridge is supposed to start with an up step. This undercounts the total number of links by a factor one half. This is corrected by the initial factor 2;
- each term is the same as in Equation 4.12, with carefulness due to the fact that the first excursion is a path, while the tail is a bridge.

The initial condition is, as for paths, $\nu_{N,l} = 0$ for $l \geqslant N$.

By induction, $\nu_{N,l} = C_l V_{N-l-1,l}$ as in the paths case. The recursion for the ratio gives:

$$V_{s,l} = 2B_s + \sum_{m=1}^{s} B_m B_{s-m} + 2\sum_{m=1}^{s} C_{s-m} V_{m-1,l} \qquad s \geqslant 0, \tag{4.18}$$

which again shows that $V_{N-l-1,l}$ is only function of $s = N-l-1$. Now, the second term is simplified thanks to propery 2) of Lemma 4.6, giving

$$V_s = 4^s + B_s + 2\sum_{m=1}^{s} C_{s-m} V_{m-1}. \tag{4.19}$$

The recursion is solved introducing a generating function $V(z) = \sum_{s=0}^{\infty} V_s z^s$. Equation 4.19 implies

$$\begin{aligned}
V(z) &= \sum_{s=0}^{\infty} [4^s + B_s] z^s + 2\sum_{s=0}^{\infty}\sum_{m=1}^{s} C_{s-m}V_{m-1}z^s \\
&= (1-4z)^{-1} + (1-4z)^{-\frac{1}{2}} + 2\sum_{m=1}^{\infty}\sum_{s=m}^{\infty} C_{s-m}V_{m-1}z^s \\
&= (1-4z)^{-1} + (1-4z)^{-\frac{1}{2}} + 2\sum_{n=0}^{\infty}\sum_{k=0}^{\infty} C_k V_n z^{k+n+1} \\
&= (1-4z)^{-1} + (1-4z)^{-\frac{1}{2}} + 2zC(z)V(z),
\end{aligned} \tag{4.20}$$

giving $V(z) = (1-4z)^{-1} + (1-4z)^{-\frac{3}{2}}$.

Lemma 4.4 gives

$$\begin{aligned}
V(z) &= \sum_{m=0}^{\infty}\left[1 + \frac{\Gamma\left(m+\frac{3}{2}\right)}{\Gamma\left(\frac{3}{2}\right)m!}\right]4^m z^m \\
&= \sum_{m=0}^{\infty}\left[1 + \frac{\left(m+\frac{1}{2}\right)2^{1-2m}\sqrt{\pi}\,\Gamma(2m)}{\frac{\sqrt{\pi}}{2}\Gamma(m)m!}\right]4^m z^m \\
&= \sum_{m=0}^{\infty}\left[4^m + 2\frac{(2m+1)(2m-1)!}{(m-1)!m!}\right]z^m \\
&= \sum_{m=0}^{\infty}\left[4^m + \frac{(2m+2)(2m+1)2m(2m-1)!}{(2m+2)m(m-1)!m!}\right]z^m \\
&= \sum_{m=0}^{\infty}\left[4^m + \frac{m+1}{2}B_{m+1}\right]z^m
\end{aligned} \tag{4.21}$$

resulting in $V_{N-l-1} = 4^{N-l-1} + \frac{N-l}{2}B_{N-l}$ and $v_{N,l} = C_l 4^{N-l-1} + (N-l)r_{N,l}$.

The average distribution of links' lenghts is

$$L_N^{\mathcal{B}}(l) = \frac{C_l}{B_N}\left[4^{N-l-1} + \frac{N-l}{2}B_{N-l}\right]. \tag{4.22}$$

4.5 Unified notation for paths and bridges

From now on, all models and results can be equally defined for paths or bridges. It's useful to adopt a unified notation in all the following, except where differently specified.

The cardinality of the set of possible color orderings will be denoted as P_N (for paths, $P_N = C_N$ etc...), and the set itself will be denoted as $\mathcal{P}_N$. The average distribution of links' lenghts will be denoted as $L_N(l)$, and as seen has the form

$$L_N(l) = \frac{C_l}{P_N}D_{N-l-1}. \tag{4.23}$$

The generating function of D_s will be denoted as $D(z)$.

4.6 Average cost computation

The average cost per link of natural matchings will be denoted as $\epsilon(p,N)$ with a slight abuse of notation. Notice that every colored sequence is equiprobable, thus every natural matching/bridge is equiprobable. The average cost of natural matchings can be written as:

$$\begin{aligned}\epsilon(p,N) &= \frac{1}{NN^p P_N}\sum_{b\in\mathcal{P}_N}\sum_{l=0}^{N-1} n_l(b)\phi(p,l,N)\\ &= \frac{1}{N^{p+1}P_N}\sum_{l=0}^{N-1}\left(\sum_{b\in\mathcal{P}_N} n_l(b)\right)\phi(p,l,N) \qquad (4.24)\\ &= \frac{1}{N^{p+1}}\sum_{l=0}^{N-1} L_N(l)\phi(p,l,N)\end{aligned}$$

where $N^{-p}\phi(p,l,N)$ is the average cost of a link of lenght $2l+1$, and the factor N^{-p} is highlighted to absorb the N dependence of ϕ in the large N limit, allowing to consider $\phi(p,l,N)=\phi(p,l)$. $n_l(b)$ is the number of links of lenght $2l+1$ in the natural matching of the sequence b.

After the introduction of the average cost function, explicit summation of Equation 4.24 will be performed with two independent methods.

4.6.1 Average cost of links

To study Equation 4.24, the average cost of a link of lenght $2l+1$, $N^{-p}\phi(p,l,N)$ must be computed. Three cases are considered.

Equidistant points

The easiest case for the computation of $\phi(p,l,N)$ is the equidistant points case, in which all the points are considered equidistant, with the first being at coordinate $\frac{1}{2N+1}$ and the last at coordinate $\frac{2N}{2N+1}$.

$$N^{-p}\phi_{eq}(p,l) = \left(\frac{2l+1}{2N+1}\right)^p . \qquad (4.25)$$

Strangely enough, as simple as it seems the equidistant case is not exactly nor approximately solvable. Only its leading behaviour for $\frac{1}{2}<p<1$ is known: in fact, it will be shown that ϕ functions that agree in the large l limit provide the same $\frac{1}{2}<p<1$ leading behaviour.

Random positions at fixed color ordering

In [CDS17] the probability density for the distance between successive random points is computed (Equation 95a).

One has

$$\rho_N^{(1)}(\varphi_i = x_{i+1} - x_i) = 2N e^{2N\varphi_i} \tag{4.26}$$

if x_i is the position of the i-th point in order of ascending coordinate.

Then

$$N^{-p}\phi_{\text{rnd}}(p,l,N) = \overline{(x_{i+2l+1} - x_i)^p} = \overline{\left(\sum_{j=i}^{i+2l} \varphi_j\right)^p} = \overline{z_i^p} \tag{4.27}$$

where $z_i = \sum_{j=i}^{i+2l} \varphi_j$. The last average is intended over the probability distribution of z, which is the convolution of $2l+1$ exponentials:

$$p_Z(z) = (2N)^{2l+1} z^{2l} e^{-2Nz}. \tag{4.28}$$

The average value is then

$$\begin{aligned} N^{-p}\phi_{\text{rnd}}(p,l) = \overline{z^p} &= \frac{\int_0^\infty (2N)^{2l+1} z^p z^{2l} e^{-2Nz}}{\int_0^\infty (2N)^{2l+1} z^{2l} e^{-2Nz}} \\ &\overset{x=2Nz}{=} \frac{\int_0^\infty (2N)^{-2l-p-1} x^{2l+p} e^{-x}}{\int_0^\infty (2N)^{-2l-1} x^{2l} e^{-x}} \\ &= \left(\frac{1}{2N}\right)^p \frac{\Gamma(2l+1+p)}{\Gamma(2l+1)}. \end{aligned} \tag{4.29}$$

The behaviour for large l of $\phi_{\text{rnd}}(p,l,N)$ is given by

$$\phi_{\text{rnd}}(p,l) = l^p \left(1 + \frac{p(1+p)}{4l} + o\left(\frac{1}{l}\right)\right), \tag{4.30}$$

coinciding at first order with the equidistant points case.

Custom cost function

The possibility of using custom made average costs $\phi(p,l,N)$ allows to choose cost functions that simplify computation. Clearly custom cost function will not describe precisely the average distance in matching problems; it's important then to understand when these custom cost function approximate well some matching problem.

We found that

$$\phi_o(p,l) = \frac{\Gamma\left(l+\frac{1}{2}+p\right)}{\Gamma\left(l+\frac{1}{2}\right)} = l^p\left(1 + \frac{p^2}{2l} + o\left(\frac{1}{l}\right)\right). \tag{4.31}$$

allows certain recursions to close in exact form, allowing easier explicit computations of the average cost.

In fact, consider the total cost on the ensamble of paths

$$E_N[\phi] = \frac{1}{C_n} \sum_{l=0}^{N-1} r_{N,l}\, \phi(l) \tag{4.32}$$

as a functional of the average cost function ϕ. Note that the $r_{n,k}$ satisfies the following recursion:

$$r_{N+1,l} = \begin{cases} \frac{1}{2} B_{N+1} & l = 0 \\ \frac{1}{N+2}[(4N-4l+2)r_{N,l} + (4l-2)r_{N,l-1}] & 1 \leqslant l \leqslant N-1 \\ C_N & l = N \end{cases} . \tag{4.33}$$

In turns, this implies

$$(4N+2)E_{N+1}[\phi] = (4N+4)E_N[\phi] + 4E_N[\tilde{\phi}] + (2N+1)\phi(0) + \phi(N) \tag{4.34}$$

where

$$\tilde{\phi}(l) = \left(l + \frac{1}{2}\right)[\phi(l+1) - \phi(l)]. \tag{4.35}$$

The family of functions $\phi_0(p,l,N)$ is such that $\tilde{\phi}_0(p,l) = p\phi_0(p,l)$. This allows to iterate the recursion to the initial condition $E_0[\phi] = 0$, explaining our interest; the actual computation is performed in Section 4.6.6.

4.6.2 Generating functions for the average cost

The first method to compute Equation 4.24 uses generating functions techniques, allowing the summation for both paths and bridges with ϕ_0 cost function. The same technique allows for the computation of the leading behaviour in the case of ϕ_{rnd} average cost function, and for the equidistant case in the $p > \frac{1}{2}$ regime. At the cost of tedious computations, this method allows also for the computation of subleading corrections.

The idea of the generating functions method is to introduce a generating function $\mathcal{E}(z;p,\phi) = \sum_{N=0}^{\infty} \left[N^{p+1} P_N \epsilon(p,N)\right] z^N$; then, using the definition of ϵ and the factoriazion of $L_N(l) = \frac{C_l}{P_N} D_{N-l-1}$:

$$\begin{aligned} \mathcal{E}(z;p,\phi) &= \sum_{N=0}^{\infty} \sum_{l=0}^{N-1} D_{N-l-1} C_l \phi(l,p) z^N \\ &= z \sum_{l=0}^{\infty} C_l \phi(l,p) z^l \sum_{N=l+1}^{\infty} D_{N-l-1} z^{N-l-1} \\ &= zD(z) \sum_{l=0}^{\infty} C_l \phi(l,p) z^l \\ &= zD(z)\Phi(z;p) \end{aligned} \tag{4.36}$$

where $\Phi(z;p)$ generates the sequence $C_l\phi(l,p)$. Computation of $\Phi(z;p)$ allows to compute the coefficients of the expansion of $\mathcal{E}(z;p,\phi)$.

Random positions at fixed color ordering

For $\phi(p, l) = \phi_{\mathrm{rnd}}(l, p)$, one has

$$\begin{aligned}
\Phi_{\mathrm{rnd}}(z;p) &= \sum_{k=0}^{\infty} z^k C_k \frac{\Gamma(2k+1+p)}{\Gamma(2k+1)2^p} \\
&= 2^{-p} \sum_{k=0}^{\infty} z^k \frac{(2k)!}{(k+1)!k!} \frac{\Gamma(2k+1+p)}{(2k)!} \\
&= 2^{-p}\Gamma(p+1) \sum_{k=0}^{\infty} \frac{(4z)^k}{k!} \frac{1}{(k+1)!} 4^{-k} \frac{\Gamma(2k+1+p)}{\Gamma(1+p)} \\
&= 2^{-p}\Gamma(p+1) \sum_{k=0}^{\infty} \frac{(4z)^k}{k!} \frac{\left(\frac{1+p}{2}\right)^{\overline{k}} \left(\frac{2+p}{2}\right)^{\overline{k}}}{2^{\overline{k}}} \\
&= 2^{-p} p\Gamma(p)\, F\left(\begin{matrix} \frac{1+p}{2}, \frac{2+p}{2} \\ 2 \end{matrix} \middle| 4z \right),
\end{aligned} \tag{4.37}$$

where the hypergeometric function F was introduced (the notation follows [GKPL89]).

The hypergeometric function is defined precisely by its series expansion, and it can be shown that it converges absolutely for $|4z| < 1$, diverges for $|4z| > 1$ and, on the circle $|4z| = 1$ converges absolutely if the sum of its lower parameters its greater than the sum of its higher parameters. In this case

$$2 - \frac{p+1}{2} - \frac{p+2}{2} > 0 \quad \Longrightarrow \quad p < \frac{1}{2} \tag{4.38}$$

is the condition for convergence on the circle.

Expanding around the possible pole $4z = 1$, one finds (Mathematica):

$$\begin{aligned}
\Phi_{\mathrm{rnd}}(z;p) = \frac{p\Gamma(p)\sqrt{\pi}\sec(p\pi)}{2^p} \Bigg[&- \frac{2^p}{\Gamma(\frac{3}{2}-p)\Gamma(p+1)} (1-4z)^{\frac{1}{2}-p} \\
&+ \frac{2^{1-p}}{\Gamma(2-p)\Gamma(p+\frac{1}{2})} \Bigg]
\end{aligned} \tag{4.39}$$

for $p \neq \frac{1}{2}$, and

$$\Phi_{\mathrm{rnd}}(z; p = \frac{1}{2}) = -\frac{\Gamma\left(\frac{1}{2}\right)}{\pi} \log(1-4z) = -\frac{1}{\sqrt{\pi}} \log(1-4z). \tag{4.40}$$

Custom cost function

For $\phi(p,l) = \phi_0(l,p)$, one has

$$\begin{aligned}
\Phi_0(z;p) &= \sum_{l=0}^{\infty} C_l \frac{\Gamma(l+p+\frac{1}{2})}{\Gamma(l+\frac{1}{2})} z^l = \sum_{l=0}^{\infty} \frac{4^l \Gamma(l+\frac{1}{2})}{\sqrt{\pi}\,\Gamma(l+2)} \frac{\Gamma(l+p+\frac{1}{2})}{\Gamma(l+\frac{1}{2})} z^l \\
&= \frac{1}{\sqrt{\pi}} \sum_{l=0}^{\infty} (4z)^l \frac{\Gamma(l+p+\frac{1}{2})}{\Gamma(l+2)} = \frac{1}{4z\sqrt{\pi}} \sum_{l=1}^{\infty} (4z)^l \frac{\Gamma(l+p-\frac{1}{2})}{l!} \\
&\overset{p\neq\frac{1}{2}}{=} \frac{1}{4z\sqrt{\pi}} \left[\sum_{l=0}^{\infty} (4z)^l \frac{\Gamma(l+p-\frac{1}{2})}{l!} - \Gamma(p-\frac{1}{2}) \right] \\
&= \frac{\Gamma(p+\frac{1}{2})}{2z(1-2p)\sqrt{\pi}} \left[1 - \sum_{l=0}^{\infty} (4z)^l \frac{\Gamma(l+p-\frac{1}{2})}{l!\,\Gamma(p-\frac{1}{2})} \right] \\
&= \frac{\Gamma(p+\frac{1}{2})}{2z(1-2p)\sqrt{\pi}} \left[1 - (1-4z)^{\frac{1}{2}-p} \right],
\end{aligned} \tag{4.41}$$

where in the second step the analytic continuation of Catalan numbers was used, and in the last step Lemma 4.4 was used.

For $p = \frac{1}{2}$, using standard MacLaurin expansions one finds

$$\Phi_0(z, p = \frac{1}{2}) = -\frac{\log(1-4z)}{4z\sqrt{\pi}} = \frac{1}{4z\sqrt{\pi}} \sum_{l=1}^{\infty} \frac{(4z)^l}{l}, \tag{4.42}$$

whose expansion matches the fifth passage of Equation 4.41.

4.6.3 Explicit summation with generating functions: paths

For paths, $D(z) = \frac{1}{2z}\left((1-4z)^{-\frac{1}{2}} - 1\right)$.

Random positions at fixed color ordering

In this case

$$\mathcal{E}(z;p,\phi_{\text{rnd}}) = \frac{1}{2}\left((1-4z)^{-\frac{1}{2}} - 1\right) \frac{p\Gamma(p)}{2^p} F\left(\begin{matrix} \frac{1+p}{2}, \frac{2+p}{2} \\ 2 \end{matrix} \middle| \, 4z \right) \tag{4.43}$$

and no exact expansion could be found. This is not a problem for the asymptotic analysis of expansion coefficients. In fact, a number of asymptotic properties are related to analytical properties of the generating function. In particular, to compute the leading order behaviour of the coefficients of a generating function, only its expansion at the pole is needed (see Theorem 5.11 of [Wil05]).

For $0 < p < \frac{1}{2}$ one has:

$$\begin{aligned}
\mathcal{E}(z; p, \phi_{\mathrm{rnd}}) &= \frac{p\Gamma(p)\sqrt{\pi}\,\sec(p\pi)}{4^p\Gamma(2-p)\Gamma\left(p+\frac{1}{2}\right)}\left((1-4z)^{-\frac{1}{2}}-1\right) \\
&= \frac{p\Gamma(p)\Gamma\left(\frac{1}{2}-p\right)\Gamma\left(\frac{1}{2}+p\right)}{4^p\sqrt{\pi}\,\Gamma(2-p)\Gamma\left(p+\frac{1}{2}\right)}\left((1-4z)^{-\frac{1}{2}}-1\right) \\
&= \frac{p\Gamma(p)\Gamma\left(\frac{1}{2}-p\right)}{4^p\sqrt{\pi}\,\Gamma(2-p)}\sum_{N=1}^{\infty}\frac{\Gamma\left(N+\frac{1}{2}\right)}{\sqrt{\pi}}\frac{4^N z^N}{N!}.
\end{aligned} \tag{4.44}$$

This implies that

$$\begin{aligned}
\epsilon(p, N) &\sim \frac{N^{-1-p}}{C_N}\frac{p\Gamma(p)\Gamma\left(\frac{1}{2}-p\right)}{4^p\,\pi\,\Gamma(2-p)}4^N\frac{\Gamma\left(N+\frac{1}{2}\right)}{N!} \\
&\sim \frac{N^{-1-p}\sqrt{\pi}N^{\frac{3}{2}}}{4^N}\frac{p\Gamma(p)\Gamma\left(\frac{1}{2}-p\right)}{4^p\,\pi\,\Gamma(2-p)}4^N\frac{1}{\sqrt{N}} \\
&\sim N^{-p}\frac{p\Gamma(p)\Gamma\left(\frac{1}{2}-p\right)}{4^p\sqrt{\pi}\,\Gamma(2-p)}
\end{aligned} \tag{4.45}$$

where $\sim$ is inteded as equality for the leading term in the large N expansion. In the above the following fact was used:

$$\frac{\Gamma(N+a)}{\Gamma(N+b)} \sim N^{a-b} \quad \text{for} \quad N \to \infty. \tag{4.46}$$

For $p = \frac{1}{2}$ one has, using Lemma 4.5, and considering only the dominant singular part of $D(z)$:

$$\begin{aligned}
\mathcal{E}(z; p = \frac{1}{2}, \phi_{\mathrm{rnd}}) &= -\frac{1}{2}(1-4z)^{-\frac{1}{2}}\frac{1}{\sqrt{\pi}}\log(1-4z) \\
&= \frac{1}{2\sqrt{\pi}}\sum_{N=0}^{\infty}\frac{\Gamma\left(N+\frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)}\left[\psi_0\left(N+\frac{1}{2}\right)-\psi_0\left(\frac{1}{2}\right)\right]\frac{4^N z^N}{N!}.
\end{aligned} \tag{4.47}$$

This implies:

$$\begin{aligned}
\epsilon(p = \frac{1}{2}, N) &\sim \frac{N^{-\frac{3}{2}}\sqrt{\pi}N^{\frac{3}{2}}}{4^N}\frac{1}{2\pi}\frac{4^N}{\sqrt{N}}\log(N) \\
&\sim \frac{1}{2\sqrt{\pi}}\frac{1}{\sqrt{N}}\log(N).
\end{aligned} \tag{4.48}$$

For $\frac{1}{2} < p < 1$ one has, considering only the dominant singular part of $D(z)$:

$$\begin{aligned}
\mathcal{E}(z; p, \phi_{\mathrm{rnd}}) &= -\frac{p\Gamma(p)\sqrt{\pi}\sec(p\pi)}{2\Gamma(p+1)\Gamma\left(\frac{3}{2}-p\right)}(1-4z)^{-p} \\
&= -\frac{\sqrt{\pi}\sec(p\pi)}{2\Gamma\left(\frac{3}{2}-p\right)}\sum_{N=0}^{\infty}\frac{\Gamma(N+p)}{\Gamma(p)}\frac{4^N z^N}{N!} \\
&= -\frac{\Gamma\left(\frac{1}{2}-p\right)\Gamma\left(\frac{1}{2}+p\right)}{2\sqrt{\pi}\left(\frac{1}{2}-p\right)\Gamma\left(\frac{1}{2}-p\right)}\sum_{N=0}^{\infty}\frac{\Gamma(N+p)}{\Gamma(p)}\frac{4^N z^N}{N!} \\
&= \frac{\Gamma\left(p-\frac{1}{2}\right)}{2\sqrt{\pi}}\sum_{N=0}^{\infty}\frac{\Gamma(N+p)}{\Gamma(p)}\frac{4^N z^N}{N!}
\end{aligned} \tag{4.49}$$

This implies that

$$\begin{aligned}
\epsilon(p, N) &\sim \frac{N^{-1-p}\sqrt{\pi}N^{\frac{3}{2}}}{4^N}\frac{\Gamma\left(p-\frac{1}{2}\right)}{2\Gamma(p)\sqrt{\pi}}4^N\frac{\Gamma(N+p)}{N!} \\
&\sim \frac{N^{\frac{1}{2}-p}\Gamma\left(p-\frac{1}{2}\right)}{2\Gamma(p)}N^{p-1} \\
&\sim \frac{\Gamma\left(p-\frac{1}{2}\right)}{2\Gamma(p)}\frac{1}{\sqrt{N}}.
\end{aligned} \tag{4.50}$$

To sum it up:

$$\epsilon(p, N) \sim \begin{cases}
\frac{p\Gamma(p)\Gamma\left(\frac{1}{2}-p\right)}{4^p\sqrt{\pi}\Gamma(2-p)}\frac{1}{N^p} & 0 < p < \frac{1}{2} \\
\frac{1}{2\sqrt{\pi}}\frac{\log(N)}{\sqrt{N}} & p = \frac{1}{2} \\
\frac{\Gamma\left(p-\frac{1}{2}\right)}{2\Gamma(p)}\frac{1}{\sqrt{N}} & \frac{1}{2} < p < 1
\end{cases} \tag{4.51}$$

Custom cost function

In this case:

$$\mathcal{E}(z; p, \phi_0) = \frac{1}{2}\left((1-4z)^{-\frac{1}{2}}-1\right)\frac{\Gamma(p+\frac{1}{2})}{2z(1-2p)\sqrt{\pi}}\left[1-(1-4z)^{\frac{1}{2}-p}\right] \tag{4.52}$$

which can be exactly expanded, allowing for an exact expression at finite N for the average cost.

For $p \neq \frac{1}{2}$:

$$\begin{aligned}
\mathcal{E}(z;p,\phi_0) &= zD(z)\Phi_0(z;p) \\
&= \frac{1}{2}\left((1-4z)^{-\frac{1}{2}}-1\right)\frac{-\Gamma(p-\frac{1}{2})}{4z\sqrt{\pi}}\left(1-(1-4z)^{\frac{1}{2}-p}\right) \\
&= \frac{\Gamma(p-\frac{1}{2})}{8z\sqrt{\pi}}\left[1+(1-4z)^{-p}-(1-4z)^{-\frac{1}{2}}-(1-4z)^{\frac{1}{2}-p}\right] \\
&= \frac{\Gamma(p-\frac{1}{2})}{2\sqrt{\pi}}\sum_{N=0}^{\infty}(4z)^N\frac{1}{(N+1)!}\left[\frac{\Gamma(N+1+p)}{\Gamma(p)}-\frac{\Gamma(N+\frac{3}{2})}{\Gamma(\frac{1}{2})}\right. \\
&\qquad\left.-\frac{\Gamma(N+p+\frac{1}{2})}{\Gamma(p-\frac{1}{2})}\right].
\end{aligned} \tag{4.53}$$

This implies:

$$\begin{aligned}
\epsilon(N,p) &= \frac{1}{N^{p+1}C^N}\frac{\Gamma(p-\frac{1}{2})}{2\sqrt{\pi}}\frac{4^N}{(N+1)!}\left[\frac{\Gamma(N+p+1)}{\Gamma(p)}-\frac{\Gamma(N+\frac{3}{2})}{\sqrt{\pi}}-\frac{\Gamma(N+p+\frac{1}{2})}{\Gamma(p-\frac{1}{2})}\right] \\
&= \frac{1}{N^{p+1}}\left[-\frac{(2N+1)\Gamma(p-\frac{1}{2})}{4\sqrt{\pi}}+\frac{\Gamma(p-\frac{1}{2})\Gamma(N+p+1)-\Gamma(p)\Gamma(N+p+\frac{1}{2})}{2\Gamma(p)\Gamma(N+\frac{1}{2})}\right]
\end{aligned} \tag{4.54}$$

and at leading order

$$\begin{aligned}
\epsilon(N,p) &\sim -\frac{N^{-p}\Gamma(p-\frac{1}{2})}{2\sqrt{\pi}}+\frac{\Gamma(p-\frac{1}{2})N^{-\frac{1}{2}}-\Gamma(p)N^{-1}}{2\Gamma(p)} \\
&\sim \frac{\Gamma(p-\frac{1}{2})}{2}\left[\frac{1}{\Gamma(p)\sqrt{N}}-\frac{1}{\sqrt{\pi}N^p}\right].
\end{aligned} \tag{4.55}$$

For $p = \frac{1}{2}$ one has, using Lemma 4.5:

$$\begin{aligned}
\mathcal{E}(z;p=\frac{1}{2},\phi_0) &= -\frac{1}{2}\left[(1-4z)^{-\frac{1}{2}}-1\right]\frac{1}{4z\sqrt{\pi}}\log(1-4z) \\
&= \frac{1}{2\sqrt{\pi}}\sum_{N=0}^{\infty}(4z)^N\left[\frac{\Gamma(N+\frac{3}{2})}{\sqrt{\pi}(N+1)!}\left(\psi_0(N+\frac{3}{2})-\psi_0(\frac{1}{2})\right)-\frac{1}{N}\right]
\end{aligned} \tag{4.56}$$

This implies:

$$\epsilon(p=\frac{1}{2},N) = N^{-\frac{3}{2}}\frac{(N+1)!}{2\Gamma(N+\frac{1}{2})}\left[\frac{(2N+1)\Gamma(N+\frac{1}{2})}{2\sqrt{\pi}(N+1)!}\left(\psi_0(N+\frac{3}{2})-\psi_0(\frac{1}{2})\right)-\frac{1}{N}\right] \tag{4.57}$$

and at leading order

$$\epsilon(p=\frac{1}{2},N) \sim N^{-\frac{3}{2}}\left[\frac{N}{2\sqrt{\pi}}\log(N)-\frac{\sqrt{N}}{2}\right] = \frac{1}{2\sqrt{\pi}}\frac{\log(N)}{\sqrt{N}}. \tag{4.58}$$

To sum it up:

$$\epsilon(p,N) \sim \begin{cases} \frac{-\Gamma\left(p-\frac{1}{2}\right)}{2\sqrt{\pi}}\frac{1}{N^p} & 0<p<\frac{1}{2} \\ \frac{1}{2\sqrt{\pi}}\frac{\log(N)}{\sqrt{N}} & p=\frac{1}{2} \\ \frac{\Gamma\left(p-\frac{1}{2}\right)}{2\Gamma(p)}\frac{1}{\sqrt{N}} & \frac{1}{2}<p<1 \end{cases} \tag{4.59}$$

Comparison

The behaviour of the average cost per link with the two cost functions considered is similar: both predict the same exponent in the whole range of values of p.

The coefficient of the leading order is the same for $\frac{1}{2} \leqslant p < 1$. This is expected: in this regime of values of p, a finite initial portion of the sum defining the average cost can be summed separately, giving a subleading contribution of order N^{-p}. Then, only the large l behaviour of the average cost functions $\phi(p,l)$ matters, and since the ϕ's considered agree in this limit, the same leading coefficient is found. For $0 < p < \frac{1}{2}$ the opposite is true: finite initial portion of the sum contribute exactly to the leading behaviour, and in fact different coefficients are found for different average cost functions ϕ.

4.6.4 Explicit summation with generating functions: bridges

For bridges, $D(z) = (1-4z)^{-\frac{3}{2}} + (1-4z)^{-1}$.

Random positions at fixed color ordering

In this case

$$\mathcal{E}(z;p,\phi_{\text{rnd}}) = z\left[(1-4z)^{-\frac{3}{2}} + (1-4z)^{-1}\right]\frac{p\Gamma(p)}{2^p}\, F\left(\begin{matrix}\frac{1+p}{2}, \frac{2+p}{2} \\ 2\end{matrix}\,\middle|\, 4z\right) \tag{4.60}$$

whose explicit expansion is not known. In the vicinity of the pole $4z=1$, the asymptotic behaviour of the coefficients of this generating function are dominated by the term $(1-4z)^{-\frac{3}{2}}$ and by the leading term of the hypergeometric function expansion in Equation 4.39; the z term equals $\frac{1}{4}$ to this order.

Thus, for $0 < p < \frac{1}{2}$

$$\begin{aligned} \mathcal{E}(z;p,\phi_{\text{rnd}}) &= \frac{2p\Gamma(p)\sqrt{\pi}\,\sec(p\pi)}{4^p\Gamma(2-p)\Gamma\left(p+\frac{1}{2}\right)}\frac{1}{4}(1-4z)^{-\frac{3}{2}} \\ &= \frac{2p\Gamma(p)\Gamma\left(\frac{1}{2}-p\right)}{4^{p+1}\sqrt{\pi}\,\Gamma(2-p)}\sum_{N=1}^{\infty}\frac{2\Gamma\left(N+\frac{3}{2}\right)}{\sqrt{\pi}}\frac{4^N z^N}{N!}. \end{aligned} \tag{4.61}$$

This implies that

$$\begin{aligned} \epsilon(p,N) &\sim \frac{N^{-1-p}}{B_N}\frac{p\Gamma(p)\Gamma\left(\frac{1}{2}-p\right)}{4^p\,\pi\,\Gamma(2-p)}4^N\frac{\Gamma\left(N+\frac{1}{2}\right)}{N!} \\ &\sim \frac{N^{-1-p}\sqrt{\pi}\sqrt{N}}{4^N}\frac{p\Gamma(p)\Gamma\left(\frac{1}{2}-p\right)}{4^p\,\pi\,\Gamma(2-p)}4^N\sqrt{N} \\ &\sim N^{-p}\frac{p\Gamma(p)\Gamma\left(\frac{1}{2}-p\right)}{4^p\,\sqrt{\pi}\,\Gamma(2-p)}. \end{aligned} \tag{4.62}$$

For $p=\frac{1}{2}$ one has, using Lemma 4.5, and considering only the dominant singular part of $D(z)$:

$$\begin{aligned} \mathcal{E}(z;p=\frac{1}{2},\phi_{\text{rnd}}) &= -\frac{1}{4}(1-4z)^{-\frac{3}{2}}\frac{1}{\sqrt{\pi}}\log(1-4z) \\ &= \frac{1}{4\sqrt{\pi}}\sum_{N=0}^{\infty}\frac{\Gamma\left(N+\frac{3}{2}\right)}{\Gamma\left(\frac{3}{2}\right)}\left[\psi_0\left(N+\frac{3}{2}\right)-\psi_0\left(\frac{3}{2}\right)\right]\frac{4^N z^N}{N!}. \end{aligned} \tag{4.63}$$

This implies:

$$\begin{aligned} \epsilon(p=\frac{1}{2},N) &\sim \frac{N^{-\frac{3}{2}}\sqrt{\pi}\sqrt{N}}{4^N}\frac{2}{4\pi}4^N\sqrt{N}\log(N) \\ &\sim \frac{1}{2\sqrt{\pi}}\frac{1}{\sqrt{N}}\log(N). \end{aligned} \tag{4.64}$$

For $\frac{1}{2}<p<1$ one has, considering only the dominant singular part of $D(z)$:

$$\begin{aligned} \mathcal{E}(z;p,\phi_{\text{rnd}}) &= -\frac{1}{4}\frac{p\Gamma(p)\sqrt{\pi}\sec(p\pi)}{\Gamma(p+1)\Gamma\left(\frac{3}{2}-p\right)}(1-4z)^{-1-p} \\ &= -\frac{1}{4}\frac{\sqrt{\pi}\sec(p\pi)}{\Gamma\left(\frac{3}{2}-p\right)}\sum_{N=0}^{\infty}\frac{\Gamma(N+p+1)}{\Gamma(p+1)}\frac{4^N z^N}{N!} \\ &= -\frac{1}{4}\frac{\Gamma\left(\frac{1}{2}-p\right)\Gamma\left(\frac{1}{2}+p\right)}{\sqrt{\pi}\,\left(\frac{1}{2}-p\right)\Gamma\left(\frac{1}{2}-p\right)}\sum_{N=0}^{\infty}\frac{\Gamma(N+p+1)}{\Gamma(p+1)}\frac{4^N z^N}{N!} \\ &= \frac{1}{4}\frac{\Gamma\left(p-\frac{1}{2}\right)}{\sqrt{\pi}}\sum_{N=0}^{\infty}\frac{\Gamma(N+p+1)}{\Gamma(p+1)}\frac{4^N z^N}{N!} \end{aligned} \tag{4.65}$$

This implies that

$$\begin{aligned} \epsilon(p,N) &\sim \frac{N^{-1-p}\sqrt{\pi}\sqrt{N}}{4^N}\frac{\Gamma\left(p-\frac{1}{2}\right)}{4p\Gamma(p)\sqrt{\pi}}4^N\frac{\Gamma(N+p+1)}{N!} \\ &\sim \frac{N^{-\frac{1}{2}-p}\Gamma\left(p-\frac{1}{2}\right)}{4p\Gamma(p)}N^p \\ &\sim \frac{\Gamma\left(p-\frac{1}{2}\right)}{4p\Gamma(p)}\frac{1}{\sqrt{N}}. \end{aligned} \tag{4.66}$$

To sum it up:

$$\epsilon(p,N) \sim \begin{cases} \frac{p\Gamma(p)\Gamma(\frac{1}{2}-p)}{4^p\sqrt{\pi}\Gamma(2-p)}\frac{1}{N^p} & 0<p<\frac{1}{2} \\ \frac{1}{2\sqrt{\pi}}\frac{\log(N)}{\sqrt{N}} & p=\frac{1}{2} \\ \frac{\Gamma(p-\frac{1}{2})}{4p\Gamma(p)}\frac{1}{\sqrt{N}} & \frac{1}{2}<p<1 \end{cases} \tag{4.67}$$

Notice that the first order corrections at $0 < p < \frac{1}{2}$ are given by the leading behaviour at $\frac{1}{2} < p < 1$ and viceversa. Thus the computation is valid at the second order in the asymptotic expansion for $N \to \infty$.

Custom cost function

In this case:

$$\mathcal{E}(z;p,\phi_0) = z\left[(1-4z)^{-\frac{3}{2}} + (1-4z)^{-1}\right]\frac{\Gamma(p+\frac{1}{2})}{2z(1-2p)\sqrt{\pi}}\left[1-(1-4z)^{\frac{1}{2}-p}\right] \tag{4.68}$$

which can be exactly expanded, allowing for an exact expression at finite N for tha average cost.

For $p \neq \frac{1}{2}$:

$$\begin{aligned} \mathcal{E}(z;p,\phi_0) &= zD(z)\Phi_0(z;p) \\ &= z\left[(1-4z)^{-\frac{3}{2}} + (1-4z)^{-1}\right]\frac{-\Gamma(p-\frac{1}{2})}{4\sqrt{\pi}}\left(1-(1-4z)^{\frac{1}{2}-p}\right) \\ &= -\frac{\Gamma(p-\frac{1}{2})}{4\sqrt{\pi}}\left[(1-4z)^{-\frac{3}{2}} + (1-4z)^{-1} - (1-4z)^{-1-p} - (1-4z)^{-\frac{1}{2}-p}\right] \\ &= -\frac{\Gamma(p-\frac{1}{2})}{4\sqrt{\pi}}\sum_{N=0}^{\infty}\frac{(4z)^N}{N!}\left[\frac{2\Gamma\left(N+\frac{3}{2}\right)}{\sqrt{\pi}} + N! - \frac{\Gamma(N+p+1)}{\Gamma(p+1)}\right. \\ &\qquad\qquad\qquad\left. - \frac{\Gamma(N+p+\frac{1}{2})}{\Gamma(p+\frac{1}{2})}\right]. \end{aligned} \tag{4.69}$$

This implies:

$$\begin{aligned} \epsilon(N,p) &= \frac{\sqrt{\pi}\Gamma(N+1)}{N^{1+p}4^N\Gamma(N+\frac{1}{2})}\frac{4^N\Gamma(p-\frac{1}{2})}{4\sqrt{\pi}N!}\left[\frac{\Gamma(N+p+1)}{\Gamma(p+1)} + \frac{\Gamma(N+p+\frac{1}{2})}{\Gamma(p+\frac{1}{2})} - N! - \frac{2\Gamma(N+\frac{3}{2})}{\sqrt{\pi}}\right] \\ &= \frac{\Gamma(p-\frac{1}{2})}{4N^{1+p}\Gamma(N+\frac{1}{2})}\left[\frac{\Gamma(N+p+1)}{\Gamma(p+1)} + \frac{\Gamma(N+p+\frac{1}{2})}{\Gamma(p+\frac{1}{2})} - N! - \frac{2\Gamma(N+\frac{3}{2})}{\sqrt{\pi}}\right] \end{aligned} \tag{4.70}$$

and at leading order

$$\begin{aligned}\epsilon(N,p) &\sim \frac{\Gamma(p-\frac{1}{2})}{4}\left[\frac{1}{\Gamma(p+1)\sqrt{N}}+\frac{1}{N\,\Gamma(p+\frac{1}{2})}-\frac{1}{N^{p+\frac{1}{2}}}-\frac{2}{\sqrt{\pi}N^p}\right]\\ &\sim \frac{\Gamma(p-\frac{1}{2})}{4}\left[\frac{1}{p\Gamma(p)\sqrt{N}}-\frac{2}{\sqrt{\pi}\,N^p}\right].\end{aligned} \tag{4.71}$$

For $p=\frac{1}{2}$ one has, using Lemma 4.5:

$$\begin{aligned}\mathcal{E}(z;p=\frac{1}{2},\phi_0) &= -z\left[(1-4z)^{-\frac{3}{2}}+(1-4z)^{-1}\right]\frac{1}{4z\sqrt{\pi}}\log(1-4z)\\ &= \frac{1}{4\sqrt{\pi}}\sum_{N=0}^{\infty}\frac{(4z)^N}{N!}\left[\frac{\Gamma\left(N+\frac{3}{2}\right)}{\Gamma\left(\frac{3}{2}\right)}\left(\psi_0\left(N+\frac{5}{2}\right)-\psi_0\left(\frac{3}{2}\right)\right)\right.\\ &\quad\left.+N!\left(\psi_0(N+1)-\gamma_E\right)\right]\end{aligned} \tag{4.72}$$

This implies:

$$\begin{aligned}\epsilon(p=\frac{1}{2},N) &= \frac{\sqrt{\pi}N!}{4^N N^{\frac{3}{2}}\Gamma(N+\frac{1}{2})}\frac{4^N}{4\sqrt{\pi}N!}\left[\frac{\Gamma\left(N+\frac{3}{2}\right)}{\Gamma\left(\frac{3}{2}\right)}\left(\psi_0\left(N+\frac{5}{2}\right)-\psi_0\left(\frac{3}{2}\right)\right)\right.\\ &\quad\left.+N!\left(\psi_0(N+1)-\gamma_E\right)\right]\\ &= \frac{1}{4N^{\frac{3}{2}}\Gamma(N+\frac{1}{2})}\left[\frac{\Gamma\left(N+\frac{3}{2}\right)}{\Gamma\left(\frac{3}{2}\right)}\left(\psi_0\left(N+\frac{5}{2}\right)-\psi_0\left(\frac{3}{2}\right)\right)\right.\\ &\quad\left.+N!\left(\psi_0(N+1)-\gamma_E\right)\right]\end{aligned} \tag{4.73}$$

and at leading order

$$\epsilon(p=\frac{1}{2},N)\sim\frac{1}{2\sqrt{\pi}}\frac{\log(N)}{\sqrt{N}}. \tag{4.74}$$

To sum it up:

$$\epsilon(p,N)\sim\begin{cases}\frac{-\Gamma(p-\frac{1}{2})}{2\sqrt{\pi}}\frac{1}{N^p} & 0<p<\frac{1}{2}\\ \frac{1}{2\sqrt{\pi}}\frac{\log(N)}{\sqrt{N}} & p=\frac{1}{2}\\ \frac{\Gamma(p-\frac{1}{2})}{4p\Gamma(p)}\frac{1}{\sqrt{N}} & \frac{1}{2}<p<1\end{cases} \tag{4.75}$$

Comparison

Again, as in the paths case, the large l behaviour determines the leading coefficient in the $\frac{1}{2} < p < 1$ regime.

Moreover, the comparison between paths and bridges at fixed average cost function shows that the leading coefficient is simply related: its the same for $0 < p < \frac{1}{2}$, and its different for a factor $2p$ in the $\frac{1}{2} < p < 1$ regime. The reason for this is explained in Section 4.6.5.

4.6.5 Universal relation between paths and bridges asymptotic coefficents

The reason for the relation between leading coefficients in the expansion of the average cost in the paths and bridges case is justified here. For later convenience, define the reduced quantities

$$\begin{aligned} c_N &:= 2^{-2N-1} C_N \sim \kappa N^{-\frac{3}{2}} \qquad \kappa = \frac{1}{2\sqrt{\pi}} \\ b_N &:= 2^{-2N-1} B_N \sim \kappa N^{-\frac{1}{2}}. \end{aligned} \tag{4.76}$$

Every bridge is either a path, or the concatenation of a non-empty path and a non-empty bridge. This implies the formula

$$b_N = 2c_N + 2\sum_{k=1}^{N-1} c_k b_{N-k}. \tag{4.77}$$

Suppose that we know the normalized total cost scaling

$$E_N^{(\text{paths})} = N^{1+p} \epsilon^{(\text{paths})}(p, N) = \frac{N^p}{C_N} \sum_{T \in \mathcal{C}_N} \sum_{e \in T} \phi(\ell(e)) \sim \alpha N^\gamma (1 + \mathcal{O}(N^{-\delta})) \tag{4.78}$$

where both α and γ are known, and $\gamma \geqslant 1$. The normalization is precisely chosen such that $\gamma \geqslant 1$ considering the behaviours found in the previous sections. Suppose (and indeed that is verified in the previous sections) that the bridge asymptotics has the same exponent

$$E_N^{(\text{bridges})} = N^{1+p} \epsilon^{(\text{bridges})}(p, N) = \frac{N^p}{B_N} \sum_{T \in \mathcal{B}_N} \sum_{e \in T} \phi(\ell(e)) \sim \beta N^\gamma (1 + \mathcal{O}(N^{-\delta})). \tag{4.79}$$

We want to determine β. The recursion above can be used to rewrite the expression for the total cost of bridges of lenght $2N$ as follows:

$$\begin{aligned} b_N \, \beta \, N^\gamma (1 + \mathcal{O}(N^{-\delta})) = {} & 2 \, c_N \, \alpha \, N^\gamma (1 + \mathcal{O}(N^{-\delta})) + \\ & 2 \, \alpha \sum_{k=1}^{N-1} c_k b_{N-k} k^\gamma (1 + \mathcal{O}(k^{-\delta})) + \\ & 2 \, \beta \sum_{k=1}^{N-1} c_k b_{N-k} (N-k)^\gamma (1 + \mathcal{O}((N-k)^{-\delta})). \end{aligned} \tag{4.80}$$

The first sum can be approximated by an integral:

$$\begin{aligned} &2\,\alpha \sum_{k=1}^{N-1} c_k b_{N-k} k^{\gamma}(1+\mathcal{O}(k^{-\delta})) = \\ &= 2\,\alpha\,\kappa^2 N^{\gamma-1} \int_0^1 dx\, x^{-3/2+\gamma}(1-x)^{-1/2}\ (1+\mathcal{O}(N^{-1},N^{-\delta})) \\ &= 2\alpha\,\kappa^2 N^{\gamma-1} \frac{\sqrt{\pi}\,\Gamma\left(\gamma-\frac{1}{2}\right)}{\Gamma(\gamma)}(1+\mathcal{O}(N^{-1},N^{-\delta})) \\ &= \alpha\, b_N \frac{\Gamma\left(\gamma-\frac{1}{2}\right)}{\Gamma(\gamma)} N^{\gamma-\frac{1}{2}}(1+\mathcal{O}(N^{-1},N^{-\delta})) \end{aligned} \tag{4.81}$$

where the fact that $\gamma > \frac{1}{2}$ is fundamental to the convergence of the integral.

The second sum can be approximated by an integral only by removing and treating separately a singularity:

$$\begin{aligned} &\sum_{k=1}^{N-1} c_k b_{N-k}(N-k)^{\gamma}(1+\mathcal{O}((N-k)^{-\delta})) \\ &= \sum_{k=1}^{N-1} c_k b_N N^{\gamma} \left[\sum_{s\geqslant 0}(-1)^s \left(\frac{k}{N}\right)^s \binom{\gamma-1/2}{s}\right] (1+\mathcal{O}(N^{-\delta})) \end{aligned} \tag{4.82}$$

where the $s=0$ term is singular due to the $k^{-\frac{3}{2}}$ behaviour. For $s=0$ the k sum gives

$$\sum_{k=1}^{N-1} c_k b_N N^{\gamma} = b_N N^{\gamma}\left(1+\frac{1}{\sqrt{\pi}N}+\mathcal{O}(N^{-\frac{3}{2}})\right) \tag{4.83}$$

where the sum of reduced Catalans was computed by explicitly finding and reexpanding their generating function. For $s \geqslant 1$ the integral approximation is valid, giving:

$$\begin{aligned} \sum_{k=1}^{N-1} c_k \left[\sum_{s\geqslant 1}(-1)^s\left(\frac{k}{N}\right)^s \binom{\gamma-1/2}{s}\right] &\simeq \kappa N^{-\frac{1}{2}} \int_0^1 dx \left[\sum_{s\geqslant 1}(-1)^s x^{-\frac{3}{2}+s}\binom{\gamma-1/2}{s}\right] \\ &= N^{-\frac{1}{2}}\left(\frac{1}{\sqrt{\pi}} - \frac{\Gamma\left(\gamma+\frac{1}{2}\right)}{\Gamma(\gamma)}\right) \end{aligned} \tag{4.84}$$

for a combined value for the whole sum

$$2\,\beta \sum_{k=1}^{N-1} c_k b_{N-k}(N-k)^{\gamma} \simeq \beta\, b_N N^{\gamma}\left(1-N^{-\frac{1}{2}}\frac{2\,\Gamma\left(\gamma+\frac{1}{2}\right)}{\Gamma(\gamma)}\right). \tag{4.85}$$

Plugging the sums in Equation 4.80 leads to cancellations of the leading order, and at subleading order to the condition

$$\frac{\beta}{\alpha} = \frac{1}{2\gamma-1}. \tag{4.86}$$

Now, from the explicit computation in the paths case, taking into account the different normalization (a factor N^{1+p}):

$$\gamma = \begin{cases} -p+1+p=1 & 0<p\leqslant\frac{1}{2} \\ -\frac{1}{2}+1+p=p+\frac{1}{2} & \frac{1}{2}<p<1 \end{cases}, \tag{4.87}$$

giving

$$\frac{\beta}{\alpha} = \begin{cases} 1 & 0<p\leqslant\frac{1}{2} \\ 1/2p & \frac{1}{2}<p<1 \end{cases}, \tag{4.88}$$

4.6.6 Explicit summation: exact recursion solution method

The second method revolves on exact solution of Equation 4.89, that allows to compute both in the path and bridge case the value of Equation 4.24 for average cost function ϕ_0.

Paths

It was already shown that the following recursion holds in the ϕ_0 case, using the omogeneity of $E[\phi]$ and reorganizing the terms:

$$E_{N+1}[\phi_p] = \left(1+\frac{p+\frac{1}{2}}{N+\frac{1}{2}}\right)E_N[\phi_p]+\frac{1}{2}\phi_p(0)+\frac{1}{4N+2}\phi_p(N) \tag{4.89}$$

Iterating the recursion up to $E_0[\phi_p]=0$ gives

$$\begin{aligned} E_N[\phi_0] &= \sum_{k=0}^{N-1}\frac{\Gamma(N+p+1)}{\Gamma(n+\frac{1}{2})}\frac{\Gamma(k+\frac{3}{2})}{\Gamma(k+p+2)}\left(\frac{1}{2}\phi_0(0)+\frac{1}{4k+2}\phi_0(k)\right) \\ &= \frac{\Gamma(N+p+1)}{\Gamma(N+\frac{1}{2})}\left[\frac{\Gamma(p+\frac{1}{2})}{2\Gamma(\frac{1}{2})}\sum_{k=0}^{N-1}\frac{\Gamma(k+\frac{3}{2})}{\Gamma(k+p+2)}+\frac{1}{4}\sum_{k=0}^{N-1}\frac{\Gamma(k+p+\frac{1}{2})}{\Gamma(k+p+2)}\right] \end{aligned} \tag{4.90}$$

The sums can be performed exactly, giving:

$$E_N[\phi_0] = -(2N+1)\frac{\Gamma(p-\frac{1}{2})}{4\sqrt{\pi}}+\frac{\Gamma\left(p-\frac{1}{2}\right)\Gamma(N+p+1)-\Gamma(p)\Gamma\left(N+p+\frac{1}{2}\right)}{2\Gamma\left(N+\frac{1}{2}\right)\Gamma(p)} \tag{4.91}$$

that reproduces the result obtained with the other method.

Bridges

The sum can be performed exactly for bridges as well, using the result for paths and computing some of the sums exactly. This gives again the already obtained result.

4.6.7 Limiting cases: $p = 0, 1$

Two limiting cases are interesting. At $p = 0$, the problem is trivial due to the fact that every link has weight 1, giving $\epsilon(p, N) = 1$. At $p = 1$, the problem has known scaling behaviour thanks to the results in the $p > 1$ regime (Equation 2.15):

$$\epsilon(p = 1, N) = \frac{\sqrt{\pi}}{4} \frac{1}{\sqrt{N}}. \tag{4.92}$$

The bridges results agree with both behaviours at $p = 0, 1$.

4.7 Comparision with simulations

4.7.1 Average cost

The theorical prediction for the leading behaviour of average cost is given in Equation 4.67 and recalled here:

$$\epsilon(p, N) \sim \begin{cases} \frac{p\Gamma(p)\Gamma\left(\frac{1}{2}-p\right)}{4^p\sqrt{\pi}\Gamma(2-p)} \frac{1}{N^p} & 0 < p < \frac{1}{2} \\ \frac{1}{2\sqrt{\pi}} \frac{\log(N)}{\sqrt{N}} & p = \frac{1}{2} \\ \frac{\Gamma\left(p-\frac{1}{2}\right)}{4p\Gamma(p)} \frac{1}{\sqrt{N}} & \frac{1}{2} < p < 1 \end{cases} \tag{4.93}$$

Recall that the first correction to the leading order in the case $0 < p < \frac{1}{2}$ is the leading behaviour for $\frac{1}{2} < p < 1$ and vice versa; in the following comparisons, the average cost was theoretically predicted using also the first correction.

Direct comparison with data must be performed with carefulness: Equation 4.93 is an asymptotic estimate for the average cost that is expected to be valid only for large N. It was already argued in Chapter 3 that simulated data seems to have a too low value of N to be confronted with asymptotic estimates. Nevertheless is important to at least confirm that simulated data for natural matchings agrees qualitatively with theorical computations; comparison with optimal matching data is needed too to understand if the natural matching approximation is a solid techinque to approach the problem.

A first comparison can be performed by simulating through Equation 4.93 the average cost at the simulated values of p and N for a graphical comparison: results are shown in Figure 4.2. It seems that natural matchings are always described properly by the theorical prediciton as N grows. Optimal matchings seems well described only for high values of p.

With these computed data, a fitting procedure can be performed on the same line as described in Chapter 3. Fitting results are shown alongside to the simulated data in Figure 4.3; notice that fitting for theorical prediction was performed only for $N \geqslant 1000$, as for low N the asymptotic approximation gets worse and worse. Simulated data

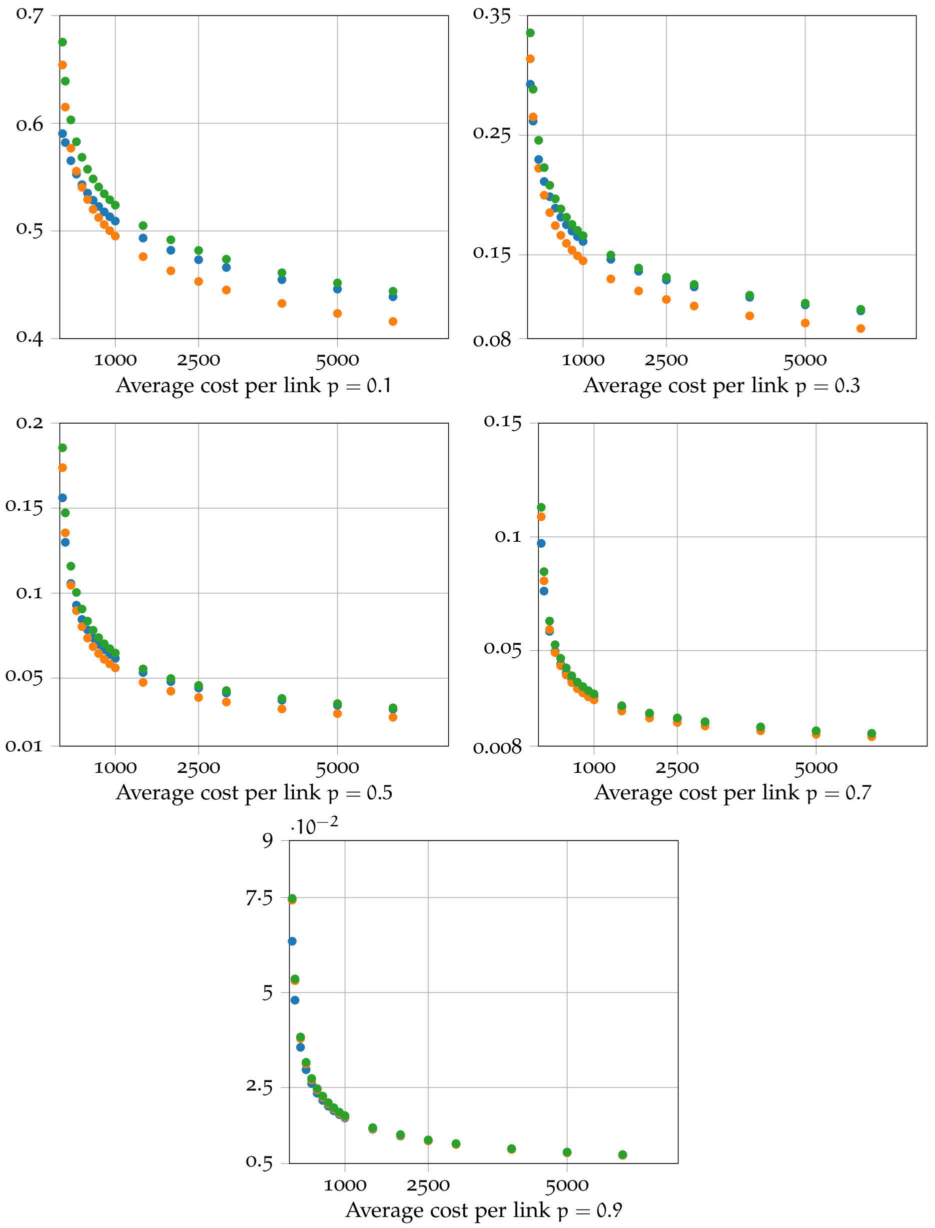


Figure 4.2: Average cost per link comparison at $p = [0.1, 0.3, 0.5, 0.7, 0.9]$

For each value of p, the blue plot is the theorical average cost, the green plot is the simulated average cost for natural matchings and the orange plot is the simulated average cost for optimal matchings. For all values of p, as N grows the theorical prediction and the natural matching simulated data coincide more and more. As p grows, optimal matching gets more and more well described by the theorical prediction.

is not perfectly reproduced by the theorical prediction, probably due to higher order corrections to the leading behaviour that were discarded in this analysis. Future work focused on higher order correction will be needed to confirm the goodness of this theorical computation.

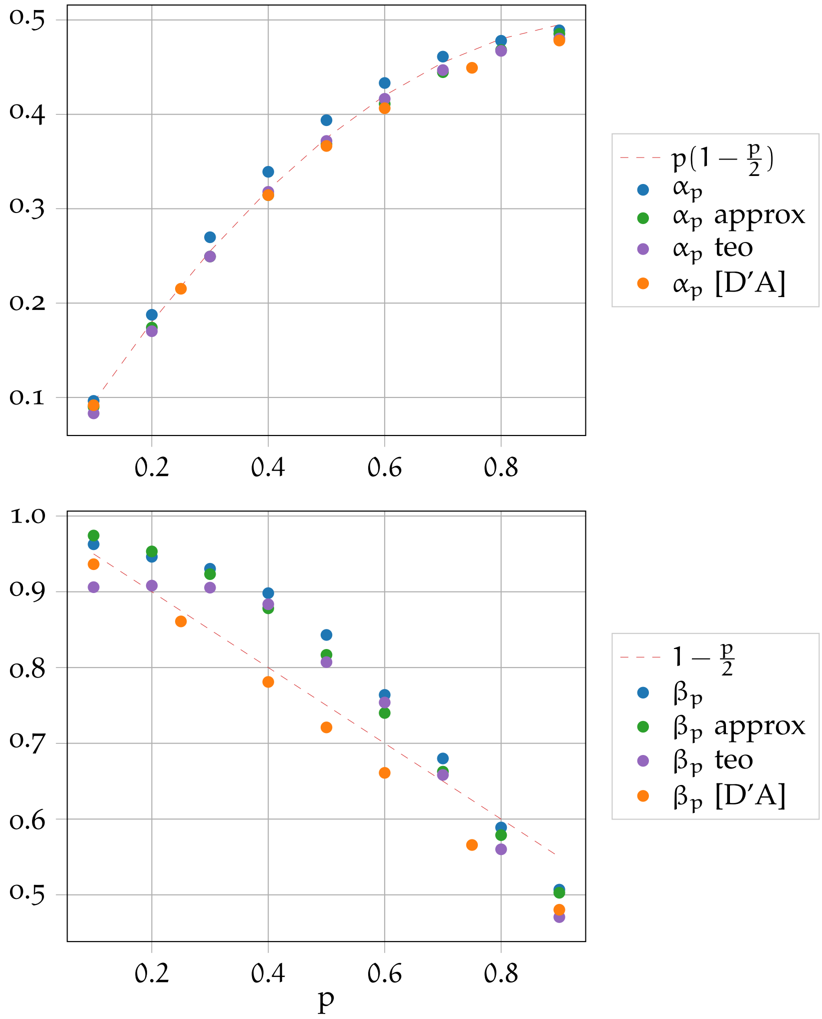


Figure 4.3: Fit results compared with theorical prediction

Theorical predictions are added to Figure 3.2. Qualitatively the predictions reproduce the behaviour of the simulated data.

4.7.2 Links' lenghts distribution

The theorical prediction for the links' lenghts distribution is given in Equation 4.22 and recalled here

$$L_N^{\mathcal{B}}(l) = \frac{C_l}{B_N}\left[4^{N-l-1} + \frac{N-l}{2}B_{N-l}\right]. \tag{4.94}$$

Comparison between theorical predictions and simulated data can be found in Figure 4.4. Theorical predictions seem to well reproduce the natural matching simulated data as expected. Optimal matchings are well reproduced in the range of shorter links, but have a significant discrepancy in the longer link range. No particular p dependence is observed.

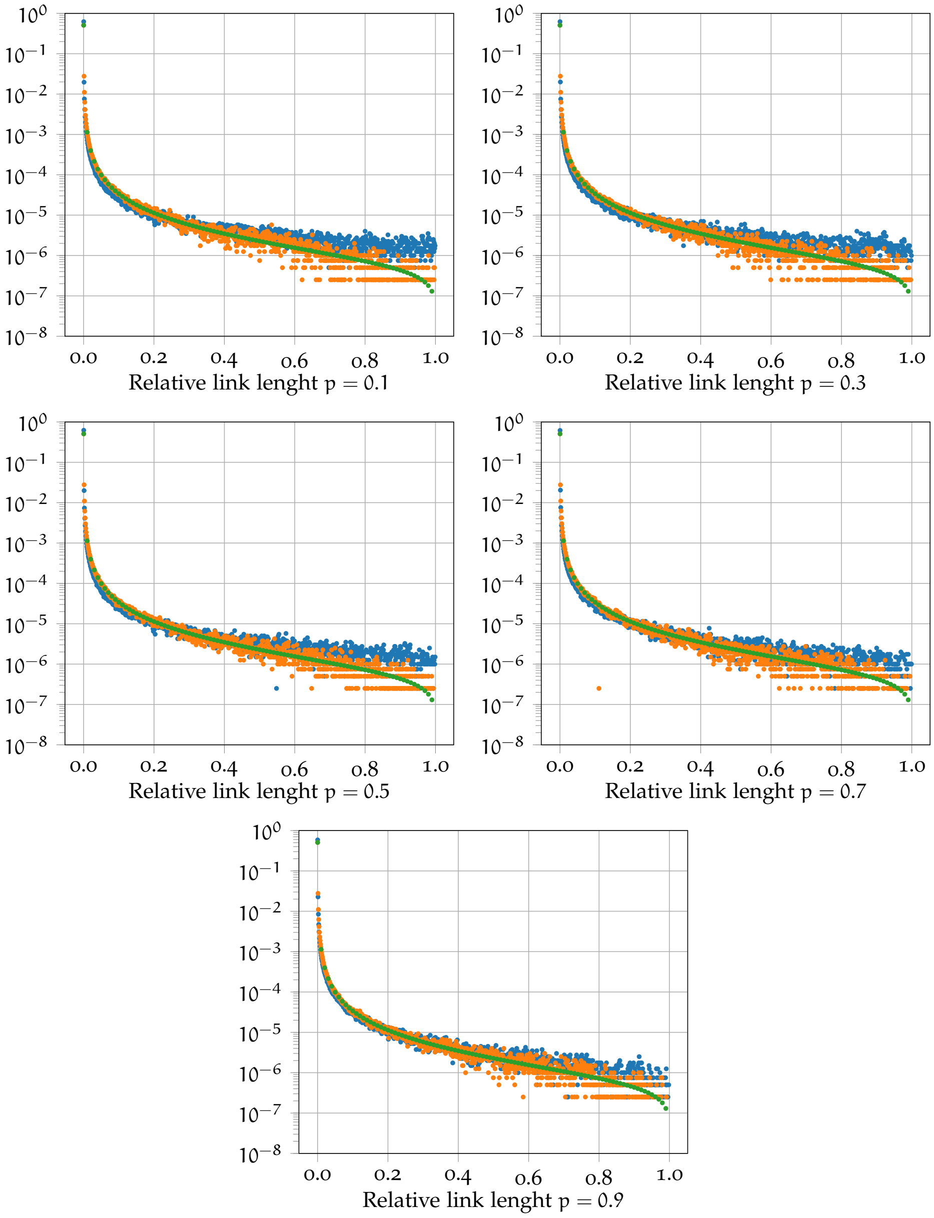


Figure 4.4: Links' lenghts distribution compared with theorical computations at $p = [0.1, 0.3, 0.5, 0.7, 0.9]$

Links' lenghts distributions: in blue data from optimal matchings, in orange data from natural matchings and in green theorical predictions from Equation 4.94 To be able to plot the distributions at $N = 4000$, a filtering process was performed, discarding all datapoints but 200 equispaced ones.

CHAPTER 5

Conclusions and outlook

In this Thesis the one dimensional random Euclidean matching problem with concave cost function was addressed. A new concept of approximate matching was introduced, namely the natural matching, and some of its average properties were exactly computed. Moreover, the same average properties were computed for formally similar combinatorial objects, Dyck paths.

The principal open question still to be completely answered is whether natural matchings share the same average properties of optimal matchings in the large N limit. Data acquired during this work of Thesis show that natural matchings qualitatively behave as optimal matchings; still, it seems that $N = 6000$ is not a large enough N to consider our data as depicting large N limit properties. New data should be simulated to better compare the two kind of matchings, and to understand if natural matchings can be really used as a proper approximation. Moreover, simulated data was not used in its fullness in this work of Thesis; the distribution of average costs could be studied, along with links' properties distributions.

Future work will focus on two main points:

- taken for granted that natural matching can be an effecive way to compute average properties of optimal matchings, we are interested in understanding how wide their application in the study of matching problems can be. Is it possible to compute properties in the monopartite problem? Is it possible to compute properties for more general concave cost functions, other than power laws? Moreover, for what model natural matchings are the real optimal solutions? Addressing these questions could improve the toolbox of techniques and results that are used in the study of matching problems, possibly allowing for new exact results;
- as mentioned in Subsection 2.1.1, non crossing matchings are closely related to favored configurations for the folding of polymeric chains, as RNA molecules. First, a thorough review of existent literature will be needed to understand known results in the light of natural matchings techniques. Then toy models for RNA folding could be formulated with the aim of reproducing experimental observations about phase transitions in the folding structure of these molecules.

Bibliography

[ABNK$^+$92] Alok Aggarwal, Amotz Bar-Noy, Samir Khuller, Dina Kravets, and Baruch Schieber. Efficient minimum cost matching using quadrangle inequality. In *Foundations of Computer Science, 1992. Proceedings., 33rd Annual Symposium on*, pages 583–592. IEEE, 1992.

[BCS14] Elena Boniolo, Sergio Caracciolo, and Andrea Sportiello. Correlation function for the grid-poisson euclidean matching on a line and on a circle. *Journal of Statistical Mechanics: Theory and Experiment*, 2014(11):P11023, 2014.

[CC05] Tommaso Castellani and Andrea Cavagna. Spin-glass theory for pedestrians. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(05):P05012, 2005.

[CDMS17] Sergio Caracciolo, Matteo P D'Achille, Enrico M Malatesta, and Gabriele Sicuro. Finite-size corrections in the random assignment problem. *Physical Review E*, 95(5):052129, 2017.

[CDS17] Sergio Caracciolo, Matteo D'Achille, and Gabriele Sicuro. Random euclidean matching problems in one dimension. *Physical Review E*, 96(4):042102, 2017.

[CDS18] Sergio Caracciolo, Matteo D'Achille, and Gabriele Sicuro. Anomalous scaling of the optimal cost in the one-dimensional random assignment problem. *arXiv preprint arXiv:1803.04723*, 2018.

[CLPS14] Sergio Caracciolo, Carlo Lucibello, Giorgio Parisi, and Gabriele Sicuro. Scaling hypothesis for the euclidean bipartite matching problem. *Physical Review E*, 90(1):012118, 2014.

[CS14] Sergio Caracciolo and Gabriele Sicuro. One-dimensional euclidean matching problem: exact solutions, correlation functions, and universality. *Physical Review E*, 90(4):042112, 2014.

[D'A15] Matteo Pietro D'Achille. On two linear assignment problems: Random assignment and euclidean bipartite matching. 2015.

[Deu99] Emeric Deutsch. Dyck path enumeration. *Discrete Mathematics*, 204(1-3):167–202, 1999.

[DSS12a] Julie Delon, Julien Salomon, and Andrei Sobolevski. Local matching indicators for transport problems with concave costs. *SIAM Journal on Discrete Mathematics*, 26(2):801–827, 2012.

[DSS12b] Julie Delon, Julien Salomon, and Andrei Sobolevski. Minimum—weight perfect matching for nonintrinsic distances on the line. *Journal of Mathematical Sciences*, 181(6):782–791, 2012.

[FS09] Philippe Flajolet and Robert Sedgewick. *Analytic combinatorics*. cambridge University press, 2009.

[GKPL89] Ronald L Graham, Donald E Knuth, Oren Patashnik, and Stanley Liu. Concrete mathematics: a foundation for computer science. *Computers in Physics*, 3(5):106–107, 1989.

[HDMM98] J Houdayer, JH Boutet De Monvel, and OC Martin. Comparing mean field and euclidean matching problems. *The European Physical Journal B-Condensed Matter and Complex Systems*, 6(3):383–393, 1998.

[Kuh55] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[McC99] Robert J McCann. Exact solutions to the transportation problem on the line. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 455, pages 1341–1380. The Royal Society, 1999.

[MM01] Andrea Montanari and Marc Mézard. Hairpin formation and elongation of biomolecules. *Physical review letters*, 86(10):2178, 2001.

[Mon81] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

[MP85] Marc Mézard and Giorgio Parisi. Replicas and optimization. *Journal de Physique Lettres*, 46(17):771–778, 1985.

[MP86] Marc Mézard and Giorgio Parisi. Mean-field equations for the matching and the travelling salesman problems. *EPL (Europhysics Letters)*, 2(12):913, 1986.

[MP87] Marc Mézard and Giorgio Parisi. On the solution of the random link matching problems. *Journal de Physique*, 48(9):1451–1459, 1987.

[MP88] Marc Mézard and Giorgio Parisi. The euclidean matching problem. *Journal de Physique*, 49(12):2019–2025, 1988.

[Nis01] Hidetoshi Nishimori. *Statistical physics of spin glasses and information processing: an introduction*, volume 111. Clarendon Press, 2001.

[NSV13] SK Nechaev, AN Sobolevski, and OV Valba. Planar diagrams from optimization for concave potentials. *Physical Review E*, 87(1):012102, 2013.

[PP16] Youngja Park and SeungKyung Park. Enumeration of generalized lattice paths by string types, peaks, and ascents. *Discrete Mathematics*, 339(11):2652–2659, 2016.

[Sic16] Gabriele Sicuro. *The Euclidean matching problem*. Springer, 2016.

[Tal92] Michel Talagrand. Matching random samples in many dimensions. *The Annals of Applied Probability*, pages 846–856, 1992.

[TN07] MV Tamm and SK Nechaev. Necklace-cloverleaf transition in associating rna-like diblock copolymers. *Physical Review E*, 75(3):031904, 2007.

[VOZ05] Graziano Vernizzi, Henri Orland, and A Zee. Enumeration of rna structures by matrix models. *Physical review letters*, 94(16):168103, 2005.

[Wil05] Herbert S Wilf. *generatingfunctionology*. AK Peters/CRC Press, 2005.

[Zde08] Lenka Zdeborová. Statistical physics of hard optimization problems. *arXiv preprint arXiv:0806.4112*, 2008.