The Dyck bound in the concave 1-dimensional random assignment model

Sergio Caracciolo[1], Matteo P. D'Achille[2], Vittorio Erba[1]
and Andrea Sportiello[3]

[1] Dipartimento di Fisica dell'Università di Milano, and INFN, sez. di Milano,
Via Celoria 16, 20100 Milano, Italy.
[2] Centre CEA de Saclay, Gif-sur-Yvette, France, CIRB Collège de France,
11 Place Marcelin Berthelot, 75231 Paris,
LI-PaRAD Université de Versailles Saint-Quentin-en-Yvelines, Versailles
and Université Paris Saclay, France
[3] LIPN, and CNRS, Université Paris 13, Sorbonne Paris Cité,
99 Av. J.-B. Clément, 93430 Villetaneuse, France.

April 25, 2019

Abstract

We consider models of assignment for random N blue points and N red points on an interval of length $2N$, in which the cost for connecting a blue point in x to a red point in y is the concave function $|x - y|^p$, for $0 < p < 1$. Contrarily to the convex case $p > 1$, where the optimal matching is trivially determined, here the optimization is non-trivial.

The purpose of this paper is to introduce a special configuration, that we call the *Dyck matching*, and to study its statistical properties. We compute exactly the average cost, in the asymptotic limit of large N, together with the first subleading correction. The scaling is remarkable: it is of order N for $p < \frac{1}{2}$, order $N \ln N$ for $p = \frac{1}{2}$, and $N^{\frac{1}{2}+p}$ for $p > \frac{1}{2}$, and it is universal for a wide class of models. We conjecture that the average cost of the Dyck matching has the same scaling in N as the cost of the optimal matching, and we produce numerical data in support of this conjecture. We hope to produce a proof of this claim in future work.

1

1 The problem

1.1 Models of Random Assignment

In this paper we study the statistical properties of the *Euclidean random assignment problem*, in the case in which the points are confined to a one-dimensional interval, and the cost is a *concave* increasing function of their distance.

The *assignment problem* is a combinatorial optimization problem, a special case of the *matching problem* when the underlying graph is bipartite. As for any combinatorial optimization problem, each realisation of the problem is described by an *instance* J, and the goal is to find, within some space of configurations, the particular one that minimises the given cost function. In the assignment problem, the instance J is a real-positive $N \times N$ matrix, encoding the costs of each possible pairing among N blue and N red points (J_{ij} is the cost of pairing the i-th blue point with the j-th red point), the space of configurations is the set of permutations of N objects, $\pi \in \mathfrak{S}_N$ (describing a complete assignment of blue and red points), and the cost function is $H_J(\pi) = \sum_{i=1}^{N} J_{i\pi(i)}$.

A *random assignment problem* is the datum of a probability measure $\mu_N(J)$ on the possible instances of the problem. The interest is in the determination of the statistical properties of this problem w.r.t. the measure under analysis, and in particular the statistical properties of the optimal configuration. The problem can be formulated as the zero-temperature limit of the statistical mechanics properties of a disordered system, where the disorder is the instance J, the dynamical variables are encoded by π, and the Hamiltonian is the cost function $H_J(\pi)$.

The case of random assignment problem in which the entries J_{ij} are random i.i.d. variables, presented already in [1], has been solved at first, in a seminal paper by Parisi and Mézard [2], through the replica trick and afterwards by the Cavity Equations [3] (see also [4] for a recent generalization of those results also at finite system size). The Parisi-Mézard solution also leads to the striking prediction that, calling $\pi_{\mathrm{opt}}(J)$ the optimal matching for the instance J and $H_{\mathrm{opt}}(J) = H_J(\pi_{\mathrm{opt}}(J))$ its optimal cost, the average over all instances of H_{opt}, for N large, tends to $\frac{\pi^2}{6}$.

This problem is simpler than a spin glass, as the determination of the optimal configuration is feasible in polynomial time (for example, through the celebrated Hungarian Algorithm [5]), however it remains non-trivial, and the thermodynamics is replica-symmetric, although the stability of the RS phase becomes marginal in the zero-temperature limit (as evinced, again, via the analysis of the cavity equations). A number of distinct other approaches to the model, and the solution of the Parisi conjecture [6] for the behaviour at finite size in the case of an exponential distribution of the random costs, which is the simplest case of a more general conjecture by Coppersmith and Sorkin [7], have appeared later on [8–10], and it is fair to say that, up to date, this model is one of the best-understood and instructive glassy system.

In analogy with the challenge of understanding spin glasses in finite dimension, there is an interest towards the study of measures $\mu_N(J)$ which are induced by some random process in a d-dimensional domain. For example, the blue and red points could be drawn randomly in some compact domain $\Omega \subset \mathbb{R}^d$ (e.g., uniformly and independently), with the entry J_{ij} given by some cost function $c(x_i, y_j)$ where x_i and y_j are the coordinates in Ω of the i-th blue point, and

the j-th red point, respectively. In the simplest versions of the model, the cost will be invariant by translations and rotations, so it will just be a function of the Euclidean distance between the two points [11], $c(x, y) = c(|x - y|)$. Scaling and universality arguments suggest to consider cost functions with a power-law behaviour, that is, for some parameter p, we shall consider the cost function

$$c(x_i, y_j) = \pm |x_i - y_j|^p \,. \tag{1}$$

Moreover, the properties of the solution of the assignment problem depend strongly on the monotonicity and concavity of the cost function, and power-law costs span a variety of combinations of such behaviours as p varies in $\mathbb{R}$ (and the sign is considered). This observation suggests that models with different values of p, and different signs of the cost function, are potentially in different universality classes, so that it is desirable to perform a study of the model both at arbitrary d and at arbitrary p (and sign).

For a cost function $c(x, y) = s\,|x - y|^p$, with $s = \pm 1$, we say that we are in the *attractive* case if $sp > 0$ (that is, the cost increases as the distance increases), and in the *repulsive* case if $sp < 0$ (that is, the cost decreases as the distance increases). The limit $p \to 0$ also makes sense, as the cost reads (with $s = \pm 1$) $c(x) = s|x|^p = s(1 + p \ln |x| + \mathcal{O}(p^2))$, but, as $\pi_{\mathrm{opt}}(\{J_{ij}\}) = \pi_{\mathrm{opt}}(\{\kappa J_{ij} + \lambda_i + \mu_j\})$, at zero temperature this is equivalent to the cost $c(x) = s' \ln |x| + \mathcal{O}(p)$, where $s' = s$ for the limit $p \searrow 0$ and $s' = -s$ for the limit $p \nearrow 0$.

For the $p > 1$ attractive case (i.e. monotone increasing and convex cost function), if the subset $\Omega \subset \mathbb{R}^d$ is compact, it is known [12] that the average total cost $\mathbb{E}_N(H_{\mathrm{opt}})$ of the optimal assignment scales with the number of points N according to

$$\mathbb{E}_N(H_{\mathrm{opt}}) \simeq \begin{cases} N^{1 - \frac{p}{2}} & \text{for} \quad d = 1 \\ N^{1 - \frac{p}{2}} (\log N)^{\frac{p}{2}} & \text{for} \quad d = 2 \\ N^{1 - \frac{p}{d}} & \text{for} \quad d > 2 \end{cases} \tag{2}$$

where $a(N) \simeq b(N)$ if $\frac{a(N)}{b(N)}$ tends to a non-zero finite constant for $N \to \infty$. In this case a relation with the classical *optimal transport problem* in the continuum has been exploited [13], in particular for the case $p = 2$ where very detailed results have been obtained [14–16].

If, at size N, the points are sampled uniformly in the domain $\Omega_N := N^{\frac{1}{d}}\Omega$ (as to keep the average density of order 1), we just have to scale the result above by the factor $N^{\frac{p}{d}}$.

Recently also the case in which Ω is not compact, and the points are not sampled uniformly, has been considered [17–19].

In the repulsive case $p < 0$ the cost function c is still *convex*. As far as we know, for this version of the problem only the case $d = 1$ has been studied in detail. In this case [20, 21]:

$$\mathbb{E}_N(H_{\mathrm{opt}}) \simeq N \,. \tag{3}$$

More precisely, in $d = 1$ corrections to the leading order in the large-N expansion were studied. Among the results, we could derive the first finite-size corrections

3

for $p < 0$, and an explicit expression in N and p for $p > 1$ [22]:

$$\mathbb{E}_N(H_{\text{opt}}) = \begin{cases} N \dfrac{\Gamma\left(1 + \frac{p}{2}\right)}{p+1} \dfrac{\Gamma\left(N+1\right)}{\Gamma\left(N+1+\frac{p}{2}\right)} \simeq \dfrac{\Gamma\left(1+\frac{p}{2}\right)}{p+1} N^{1-\frac{p}{2}} & \text{for } p > 1 \\[3ex] N \dfrac{1}{2^p}\left[1 + \dfrac{1}{3N}\dfrac{p(p-2)(p-4)}{p-3} + o\left(\dfrac{1}{N}\right)\right] & \text{for } p < 0 \,. \end{cases}$$
(4)

The attractive case $p < 0$ has not been studied so far, but is possibly of interest. However, it is easily seen that in this case the quantity $\mathbb{E}_N(H_{\text{opt}})$ makes sense only when $|p| < d$, otherwise one just gets $\mathbb{E}_N(H_{\text{opt}}) = -\infty$ already from the singular portion of the measure associated to the rare instances in which a red and a blue point almost coincide.

When $0 < p < 1$, in the attractive case, the cost function c is instead *concave*. Also in this case only the case $d = 1$ has been considered, where it has been shown that the optimal solution is always *non-crossing* [23, 24]. A matching is non-crossing if each pair of matched points defines an interval on the line which is either nested to or disjoint from all the others.

Contrary to the case $p > 1$, where in $d = 1$ the optimal matching is completely determined, and to the case $p < 0$, where in $d = 1$ it is known that the optimal matching has certain cyclic properties, the non-crossing property is not sufficient to fully characterize the optimal assignment; the regime $0 < p < 1$ is thus much more challenging to study.

The relevance of non-crossing matching configurations among elementary units aligned on a line has emerged both in physics and in biology. In the latter case, this is due to the fact that they appear in the study of the secondary structure of single stranded DNA and RNA chains in solution [25]. These chains tend to fold in a planar configuration, in which complementary nucleotides are matched, and planar configurations are exactly described by non-crossing matchings between nucleotides. The secondary structure of a RNA strand is therefore a problem of optimal matching on the line, with the restriction on the optimal configuration to be planar [26–28]. The statistical physics of the folding process is highly non-trivial and it has been investigated by many different techniques [29, 30], also in presence of disorder and in search for glassy phases [28, 29, 31–33]. So, as a further motivation for the present work, understanding the statistical properties of the solution to random Euclidean matching problems with a concave cost function could yeld results and techniques to better understand these models of RNA secondary structure.

A summary of the scaling behaviours for the different declinations of the models is provided in the Conclusion section.

1.2   Random Assignment Problems studied in this paper

In this paragraph we describe the problem of $0 < p < 1$ random assignment in detail, and fix our notations.

First, we fix Ω to be a segment (an alternate possible choice, 1-dimensional in spirit, which is not considered here but is considered, for example, in [15, 20, 24], is to take Ω as a circle, or more generally, in dimension d, to take Ω as a d-dimensional torus [13]).

Notice that in the literature Ω is typically taken to be deterministically the unit segment $[0, 1]$, while in our paper we will find easier to work with a segment
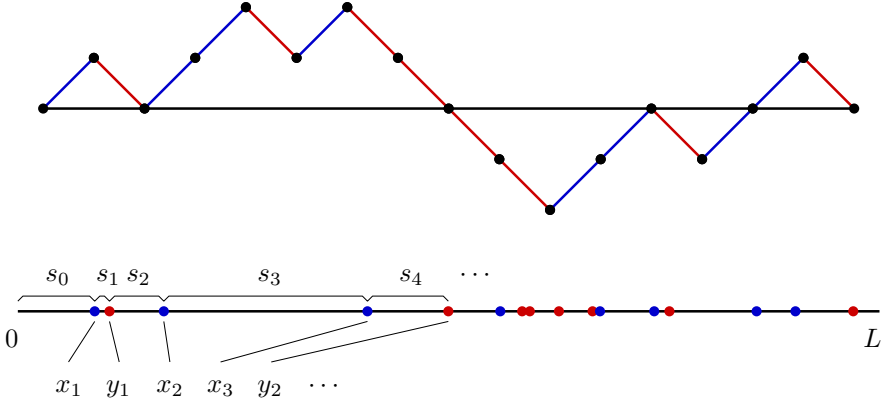
Figure 1: Example of instance $J = (\vec{x}, \vec{y}) = [\vec{s}, \sigma]$, in the PPP model. Bottom: the configuration of points. Top: the Dyck bridge associated to σ. Here $N = 8$, and $\sigma = \{+, -, +, +, -, +, -, -, -, -, +, +, -, +, +, -\}$.

$[0, L]$, where L may be a stochastic variable, whose distribution depends on N. We will work in the framework of constant density, that is $\mathbb{E}(L) \sim N$, so that, for comparison with the existing literature, our results should be corrected by multiplying by a factor of the order N^{-p}.

In dimension 1 there is a natural ordering of the points, so that we can encode an instance $J = (\vec{x}, \vec{y})$ by the ordered lists $\vec{x} = (x_1, x_2, \ldots, x_N)$, $x_i \leq x_{i+1}$, and $\vec{y} = (y_1, y_2, \ldots, y_N)$, $y_i \leq y_{i+1}$. A useful alternate encoding of the instance is $J = [\vec{s}, \sigma]$, where $s = (s_0, s_1, \ldots, s_{2N})$, $s_i \in \mathbb{R}^+$, encodes the distances between consecutive points (and between the first/last point with the respective endpoint of the segment Ω) and the vector $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_{2N}) \in \{-1, +1\}^{2N}$, with $\sum_i \sigma_i = 0$, encodes the sequence of colours of the points (see Figure 1, where the identification blue $= +1$ and red $= -1$ is adopted). In other words, the partial sums of $\vec{s}$, i.e. $(s_0, s_0 + s_1, s_0 + s_1 + s_2, \ldots, s_0 + \cdots + s_{2N-1})$, constitute the ordered list of $\vec{x} \cup \vec{y}$, and σ describes how the elements of $\vec{x}$ and $\vec{y}$ do interlace. In this notation, the domain of the instance $\Omega = [0, L]$ is determined by $L = \sum_{i=0}^{2N} s_i$. Remark that the cardinality of the space of possible vectors σ is just the central binomial,

$$B_N := \binom{2N}{N}. \tag{5}$$

For simplicity, in this paper we consider only the non-degenerate case, in which almost surely all the s_i's are strictly positive, that is, the values in $\vec{x} \cup \vec{y}$ are all distinct.

In this paper, and more crucially in subsequent work, we shall consider two families of measures. In all these measures, we have a factorisation $\mu([\vec{s}, \sigma]) = \mu_1(\vec{s})\mu_2(\sigma)$, and the measure on σ is just the uniform measure.

Independent spacing models (ISM). The measure $\mu(\vec{s})$ is factorised, and the s_i's are i.i.d. with some distribution $f(s)$ with support on $\mathbb{R}^+$ (and, for simplicity, say with all moments finite, $\int ds\, s^k f(s) < \infty$ for all k). Without loss of generality, we will assume that the average of $f(s)$ is 1, i.e. the average of L is $2N + 1$. In particular, we will consider:

Uniform spacings (US): the s_i's are deterministic, identically equal to 1, and thus $L = 2N + 1$ for all instances;

Exponential spacings (ES): the s_i's are i.i.d. with an exponential distribution $f(s) = g_1(s) = \exp(-s)$, and thus L concentrates on $2N+1$, but has a variance of order N.

Exchangeable process model (EPM). This is a generalisation of the ISM above, but now the s_i's are not necessarily i.i.d., they are instead *exchangeable variables*, that is, for all $0 \le i < 2N$,

$$\mu(s_0, \ldots, s_i, s_{i+1}, \ldots, s_{2N}) = \mu(s_0, \ldots, s_{i+1}, s_i, \ldots, s_{2N}). \qquad (6)$$

In particular, within this class of models we could have that μ is supported on the hyper-tetrahedron $\mathbb{T}_{2N}$ described by $s_i \ge 0$, and $L = \sum_i s_i = 2N + 1$. In this paper, we will consider:

Poisson Point Process (PPP): the s_i's are the spacings among the sorted list of 0, $2N+1$, and $2N$ uniform random points in the interval $[0, 2N + 1]$.

Each of these three models has its own motivations. The PPP case is, in a sense, the most natural one for what concerns applications and the comparison with the models in arbitrary dimension d. Implicitly, it is the one described in the introduction. The ES case is useful due to a strong relation with the PPP case (see Remark 1 and Lemma 4 later on in Section 2.2). In a sense, it is the "Poissonisation" of the PPP case (where in this case it is the quantity L that has been "Poissonised", that is, it is taken stochastic with its most natural underlying measure, instead of deterministic). The US case will prove out, in future work, to be the most tractable case for what concerns lower bounds to the optimal cost.

As all of the measures above are factorized in σ and $\vec{s}$, and the measure over σ is uniform, it is useful to define two shortcuts for two recurrent notions of average.

Definition 1 *For any quantity $A(J) = A(\sigma, \vec{s})$, we denote by $\overline{A}$ the average of A over σ*

$$\overline{A} := \mathbb{E}_\sigma(A) = \frac{1}{B_N} \sum_\sigma A(\sigma, \vec{s}); \qquad (7)$$

This average is independent from the choice of model among the classes above. We denote by

$$\langle A \rangle := \mathbb{E}_{\mu(\vec{s})}(A) \qquad (8)$$

the average of A over $\vec{s}$, with its appropriate measure dependence on the choice of model. Finally, we denote by $\mathbb{E}_N(A)$ the result of both averages, in which we stress the dependence from the size parameter N in the measure, that is

$$\mathbb{E}_N(A) := \mathbb{E}_{\sigma, \mu(\vec{s})}(A) = \overline{\langle A \rangle} = \langle \overline{A} \rangle. \qquad (9)$$

For a given instance, parametrised as $J = (\vec{x}, \vec{y})$, or as $J = [\vec{s}, \sigma]$, (and in which the cost function also has an implicit dependence from the exponent p), call π_{opt} one optimal configuration, and $H_{\text{opt}}(J) = H_J(\pi_{\text{opt}})$.

In this paper, we will introduce the notion of *Dyck matching* π_{Dyck} of an instance J and we will compute its average cost $H_{\text{Dyck}} := \mathbb{E}_N(H_J(\pi_{\text{Dyck}}))$ for the measures ES and PPP (with a brief discussion on the US case).

In particular we prove the following theorem:

6

Theorem 1 *For the three measures ES, US and PPP, let $\mathbb{E}_N(H_{\text{Dyck}})$ denote the average cost of the Dyck matchings. Then*

$$\mathbb{E}_N(H_{\text{Dyck}}) \simeq \begin{cases} N & 0 \le p < \frac{1}{2} \\ N \ln N & p = \frac{1}{2} \\ N^{\frac{1}{2}+p} & \frac{1}{2} < p \le 1 \end{cases} \tag{10}$$

where $a(N) \simeq b(N)$ if $\frac{a(N)}{b(N)}$ tends to a finite, non-zero constant when $N \to \infty$.

This theorem follows directly from the combination of suitable lemmas, namely Proposition 1, Proposition 2 and Corollary 1 appearing later on. In fact, our results are more precise than what is stated in the theorem above (we describe the first two orders in a series expansion for large N, including formulas for the associated multiplicative constants), details are given in the forementioned propositions.

The average cost of Dyck matchings provides an upper bound on the average cost of the optimal solution; numerical simulations for the PPP measure, described in Section 4, suggest the following conjecture, that we leave for future investigations:

Conjecture 1 *For the three measures ES, US and PPP, and all $0 < p < 1$,*

$$\lim_{N \to \infty} \frac{\mathbb{E}_N(H_{\text{opt}})}{\mathbb{E}_N(H_{\text{Dyck}})} = k_p \,, \tag{11}$$

with $0 < k_p < 1$.

2 Basic facts

Before starting our main proof, let us introduce some more notations, and state some basic properties of the optimal solution.

2.1 Basic properties of the optimal matching

A *Dyck path of semi-length N* is a lattice path from $(0,0)$ to $(2N,0)$ consisting of N 'up' steps (i.e. steps of the form $(1,1)$) and N 'down' steps (i.e., steps $(1,-1)$), which never goes below the x-axis. There are C_N Dyck path of semi-length N, where

$$C_N = \binom{2N}{N} - \binom{2N}{N+1} = \frac{1}{N+1}\binom{2N}{N} \tag{12}$$

are the *Catalan numbers*. Therefore the generating function for the Dyck paths is

$$C(z) := \sum_{k \ge 0} C_k z^k = \frac{1 - \sqrt{1-4z}}{2z} = \frac{2}{1 + \sqrt{1-4z}} \tag{13}$$

The historical name 'Dyck path' is somewhat misleading, as it leaves us with no natural name for the most obvious notion, that is, the walks of length N with steps in $\{(1,1),(1,-1)\}$. With analogy with the theory of Brownian motion (which relates to lattice walks via the Donsker's theorem) [34], we will define four types of paths, namely walks, meanders, bridges and excursions, according to the following table:

	$y(2N) = 0$	$y(x) \geq 0 \ \forall x$
walk	no	no
meander	no	yes
bridge	yes	no
excursion	yes	yes

(of course, by 'no' we mean 'not necessarily'). Thus, in fact, the 'paths' are the most constrained family of walks, that is the excursions.

In general a Dyck path (i.e., a Dyck excursion) can touch the x-axis several times. We shall call an *irreducible Dyck excursion* a Dyck path which touches the x-axis only at the two endpoints. It is trivially seen that the generating function for the irreducible Dyck excursions is simply $z\,C(z)$. As we said above a *Dyck bridge* is a walk made with the same kind of steps of Dyck paths, but without the restriction of remaining in the upper half-plane, and which returns to the x-axis. The generating function for the Dyck bridges is

$$B(z) := \frac{1}{1 - 2z\,C(z)} = \frac{1}{\sqrt{1 - 4z}} = \sum_{k \geq 0} B_k z^k \tag{14}$$

with B_k the central binomials (5), and k is the semi-length of the bridge (just like excursions, all Dyck bridges must have even length). The factor 2 in the functional form of $B(z)$ in terms of $C(z)$ enters because a bridge is a concatenation of irreducible excursions, each of which can be in the upper- or the lower-half-plane.

Now, it is clear that each configuration σ corresponds uniquely to a Dyck bridge of semi-length N, with $\sigma_i = +1$ or -1 if the i-th step of the walk is an up or down step, respectively.

In a Dyck walk we shall call $(k_{\mathrm{blue}}(i), h_{\mathrm{blue}}(i))$ the two coordinates of the mid-point of the i-th ascending step of the walk (minus $(\frac{1}{2}, \frac{1}{2})$, in order to have integer coordinates and enlighten the notation), and call $(k_{\mathrm{red}}(i), h_{\mathrm{red}}(i))$ the coordinates of the mid-point of the i-th descending step (again, minus $(\frac{1}{2}, \frac{1}{2})$). For $e = (i, j)$ an edge of a matching π, call $\|e\| = k_{\mathrm{blue}}(i) - k_{\mathrm{red}}(j)$ the horizontal distance on the walk, and $|e| = |x_i - y_j|$ the Euclidean distance on the domain segment.

For a given Dyck bridge σ, we say that $\pi \in \mathfrak{S}_N$ is *non-crossing* if, for every pair of distinct edges $e_1 = (i_1, j_1)$ and $e_2 = (i_2, j_2)$ in π, we do not have the pattern $k_{\mathrm{blue}}(i_1) < k_{\mathrm{blue}}(i_2) < k_{\mathrm{red}}(j_1) < k_{\mathrm{red}}(j_2)$, or the analogous patterns with $k_{\mathrm{blue}}(i_1) \leftrightarrow k_{\mathrm{red}}(j_1)$, or $k_{\mathrm{blue}}(i_2) \leftrightarrow k_{\mathrm{red}}(j_2)$, or $(\cdot)_1 \leftrightarrow (\cdot)_2$. The name comes from the fact that, if we represent a matching π as a diagram consisting of the domain segment $[0, L]$, and the set of N semicircles above this segment connecting the x_i's to the $y_{\pi(i)}$'s (as in the bottom part of Figure 2), then these semicircles do not intersect if and only if π is non-crossing. Note that, although these semicircles are drawn on the full $[\vec{s}, \sigma]$ instance, the notion of π being non-crossing only uses the vector σ.

For a given Dyck bridge σ, we say that $\pi \in \mathfrak{S}_N$ is *sliced* if, for every edge $e = (i, j) \in \pi$, we have $h_{\mathrm{blue}}(i) = h_{\mathrm{red}}(j)$.

Two easy lemmas have a crucial role in our analysis.

Lemma 1 *All the optimal matchings are non-crossing.*

Proof. The proof is by absurd. Suppose that π is a crossing optimal matching. If we have a pattern as $k_{\mathrm{blue}}(i_1) < k_{\mathrm{red}}(j_2) < k_{\mathrm{red}}(j_1) < k_{\mathrm{blue}}(i_2)$, then the

matching π' with edges $e'_1 = (i_1, j_2)$ and $e'_2 = (i_2, j_1)$ has $H_J(\pi') < H_J(\pi)$, because $|e'_1| < |e_1|$ and $|e'_2| < |e_2|$.

If we have a pattern as $k_{\text{blue}}(i_1) < k_{\text{blue}}(i_2) < k_{\text{red}}(j_1) < k_{\text{red}}(j_2)$, then again the matching π' with edges $e'_1 = (i_1, j_2)$ and $e'_2 = (i_2, j_1)$ has $H_J(\pi') < H_J(\pi)$, although this holds for a more subtle reason. Calling $a = x_2 - x_1$, $b = y_1 - x_2$ and $c = y_2 - y_1$, we have $|e_1| = a + b$, $|e_2| = b + c$, $|e'_1| = a + b + c$ and $|e'_2| = b$. It is the case that, for $a, b, c > 0$ and $0 < p < 1$,

$$(a + b)^p + (b + c)^p > (a + b + c)^p + b^p. \tag{15}$$

A proof of this inequality goes as follows. Call $F(a, b, c) = (a + b)^p + (b + c)^p - (a + b + c)^p - b^p$. We have $F(a, b, 0) = 0$, and

$$\frac{1}{p} \frac{\partial}{\partial c} F(a, b, c) = \frac{1}{(b + c)^{1-p}} - \frac{1}{(a + b + c)^{1-p}} > 0. \tag{16}$$

All the other possible crossing patterns are in the first or the second of the forms discussed above, up to trivial symmetries. $\square$

Non-crossing properties for assignment and optimal transport problems with concave cost functions were studied in the continuum case in [23] and in the discrete one in [24].

Lemma 2 *All the optimal matchings are sliced.*

Proof. The proof is by absurd. Suppose that π is a non-sliced optimal matching. If we have $(i, j) \in \pi$ with $h_{\text{blue}}(i) \neq h_{\text{red}}(j)$, say $h_{\text{blue}}(i) - h_{\text{red}}(j) = \delta h \neq 0$, we have that the point x_i is matched to y_j, and that, between x_i and y_j, there are n_{blue} and n_{red} blue and red points, respectively, with $n_{\text{blue}} - n_{\text{red}} = -\delta h \neq 0$. So there must be at least $|\delta h|$ points inside the interval (x_i, y_j) which are matched to points outside this interval, and thus, together with (i, j), constitute pairs of crossing edges. So, by Lemma 1, π cannot be optimal. $\square$

The slicing of optimal assignments was studied in [35].

2.2 Reducing the PPP model to the ES model

Theorem 1 (and, hopefully, Conjecture 1), in principle, shall be proven for three different models, Uniform Spacings (US), Exponential Spacings (ES) and the Poisson Point Process (PPP). However, as we anticipated, the PPP case is a minor variant of ES. In this section we give a precise account of this fact.

The starting point is a relation between the two measures:

Remark 1 *We can sample a pair $[\vec{s}', \sigma]$ with the measure μ_N^{ES} by sampling a pair $[\vec{s}, \sigma]$ with the measure μ_N^{PPP}, a value $L \in \mathbb{R}^+$ with the measure $g_{2N+1}(L) = \frac{L^{2N}}{(2N)!} \exp(-L) \, dL$, and then defining $s'_i = s_i \frac{L}{2N+1}$.*

Indeed, the measure on $\vec{s}$ in the PPP can be seen as the measure over independent exponential variables conditioned to the value of the sum; thus, the procedure leads at sight to a measure over $\vec{s}'$ which is unbiased within vectors $\vec{s}'$ with the same value of $L = \sum_i s'_i$. Then, from the independence of the spacings in the ES model we easily conclude that the distribution of L must be exactly $g_1^{*(2N+1)}(L)$, i.e. the convolution of $2N + 1$ exponential distributions.

More precisely, for k an integer, define the Gamma measure

$$g_k(x) := \frac{x^{k-1}}{\Gamma(k)} e^{-x} = g_1^{*k}(x) \,. \tag{17}$$

We use the same notation for its analytic continuation to k real positive.

We introduce (the analytic continuation of) the rising factorial (following the notation due to D. Knuth [36]):

$$n^{\overline{p}} := \frac{\Gamma(n+p)}{\Gamma(n)} = \int \mathrm{d}x \, g_n(x) x^p \,. \tag{18}$$

This choice of notation is motivated by the fact that $n^{\overline{1}} = n^1 = n$, and that, for $n \gg p^2$, $n^{\overline{p}} = n^p(1 + \mathcal{O}(p^2/n))$. More precisely we have:

Lemma 3 *For $0 \leq p \leq 1$ and $n > 1$*

$$(n+p-1)^p \leq n^{\overline{p}} \leq n^p \,. \tag{19}$$

Proof. It is well known that the Gamma function is logarithmically convex [37]. In particular, for any $0 \leq p \leq 1$,

$$\ln \Gamma(n+p) \leq (1-p) \ln \Gamma(n) + p \ln \Gamma(n+1) = \ln \Gamma(n) + p \ln n \,, \tag{20}$$

that is

$$n^{\overline{p}} = \exp(\ln \Gamma(n+p) - \ln \Gamma(n)) \leq \exp(p \ln n) = n^p \,. \tag{21}$$

Analogously, we have

$$\ln \Gamma(n) \leq (1-p) \ln \Gamma(n+p) + p \ln \Gamma(n+p-1) = \ln \Gamma(n+p) - p \ln(n+p-1) \,, \tag{22}$$

that is

$$n^{\overline{p}} = \exp(\ln \Gamma(n+p) - \ln \Gamma(n)) \geq \exp(p \ln(n+p-1)) = (n+p-1)^p \,. \tag{23}$$

$\square$

Summing up, Remark 1 and Lemma 3 allow to prove that

Lemma 4 *The following inequalities hold:*

$$\left(\frac{2N}{2N+1} \right)^{p(1-p)} \mathbb{E}(H_{\mathrm{opt}}^{\mathrm{ES}}) \leq \mathbb{E}(H_{\mathrm{opt}}^{\mathrm{PPP}}) \leq \mathbb{E}(H_{\mathrm{opt}}^{\mathrm{ES}}) \,. \tag{24}$$

Also the corresponding inequalities with H_{opt} replaced by H_{Dyck} do hold, as well as for any other quantity $H(\pi^*(J))$, whenever π^* is some matching determined by the instance, and invariant under scaling of the instance, that is $\pi^*[\vec{s}, \sigma] = \pi^*[\lambda \vec{s}, \sigma]$.

Proof. For compactness of notation, we do the proof only for the H_{opt} case, but the generalisation is straightforward. Of course, $H_{[\lambda \vec{s}, \sigma]}(\pi) = \lambda^p H_{[\vec{s}, \sigma]}(\pi)$ for all instances $[\vec{s}, \sigma]$, all configurations π, and all scaling factors $\lambda > 0$ (so as $\pi^*[\vec{s}, \sigma] = \pi^*[\lambda \vec{s}, \sigma]$, it follows that $H_{[\lambda \vec{s}, \sigma]}(\pi^*[\lambda \vec{s}, \sigma]) = \lambda^p H_{[\vec{s}, \sigma]}(\pi^*[\vec{s}, \sigma])$). In particular, $H_{\mathrm{opt}}([\lambda \vec{s}, \sigma]) = \lambda^p H_{\mathrm{opt}}([\vec{s}, \sigma])$. In light of Remark 1, we can

describe the average over $\mu_{\text{ES}}[\vec{s}, \sigma]$ in terms of an average over $\mu_{\text{PPP}}[\vec{s}, \sigma]$, and over $g_{2N+1}(L)$. More precisely

$$
\begin{aligned}
\mathbb{E}(H_{\text{opt}}^{\text{ES}}) &= \int \mathrm{d}\mu_{\text{ES}}[\vec{s}', \sigma]\, H_{\text{opt}}([\vec{s}', \sigma]) \\
&= \int \mathrm{d}\mu_{\text{PPP}}[\vec{s}, \sigma] \int \mathrm{d}L\, g_{2N+1}(L) \left(\frac{L}{2N+1}\right)^p H_{\text{opt}}([\vec{s}, \sigma]) \\
&= \mathbb{E}(H_{\text{opt}}^{\text{PPP}}) \int \mathrm{d}L\, g_{2N+1}(L) \left(\frac{L}{2N+1}\right)^p \\
&= \mathbb{E}(H_{\text{opt}}^{\text{PPP}}) \frac{(2N+1)^{\overline{p}}}{(2N+1)^p}\,.
\end{aligned}
\tag{25}
$$

The upper bound follows directly from Lemma 3. The lower bound follows as:

$$
\frac{(2N+1)^{\overline{p}}}{(2N+1)^p} \geq \left(1 - \frac{1-p}{2N+1}\right)^p \geq \left(1 - \frac{1}{2N+1}\right)^{p(1-p)}
\tag{26}
$$

where first one uses Lemma 3, then the inequality $(1 - q\epsilon)^p \geq (1 - \epsilon)^{pq}$ (valid for $\epsilon, p, q \in [0,1]$), with $\epsilon = \frac{1}{2N+1}$ and $q = 1 - p$. $\qquad\square$

In light of this lemma, it is sufficient to consider Theorem 1 (and Conjecture 1) only for the US and ES models.

The precise statement of our conclusions is the following:

Corollary 1

$$
\mathbb{E}_N^{PPP}(H_{\text{Dyck}}) = \mathbb{E}_N^{ES}(H_{\text{Dyck}}) \left(1 + \mathcal{O}\left(N^{-1}\right)\right)\,.
\tag{27}
$$

The relevance of this statement lies in the fact that, in the forthcoming equation (28), we provide an expansion for $\mathbb{E}_N^{ES}(H_{\text{Dyck}})$ in which corrections of relative order $1/N$ appear as the third term in the expansion (and we provide explicit results only for the first two terms). As a result, the very same conclusions that we have for the ES model do apply verbatim to the PPP model.

3 The Dyck matching

For every instance $[\vec{s}, \sigma]$, there is a special matching, that we call π_{Dyck}, which is sliced and non-crossing for σ. This is the matching obtained by the canonical pairing of up- and down-steps within every excursion of the Dyck bridge (see Figure 2). In analogy with our notation $H_{\text{opt}}(J) = H_J(\pi_{\text{opt}})$, let us introduce the shortcut $H_{\text{Dyck}}(J) = H_J(\pi_{\text{Dyck}})$.

Remark 2 *The Dyck matching π_{Dyck} is determined by the order of the colours of the points, σ. In particular it does not depend on the actual spacings between them, $\vec{s}$, and it does not depend on the cost exponent p.*

Remark 2 is a crucial fact that makes possible the evaluation of the statistical properties of π_{Dyck}, with a moderate effort. In particular, it will lead to the main result of this paper:
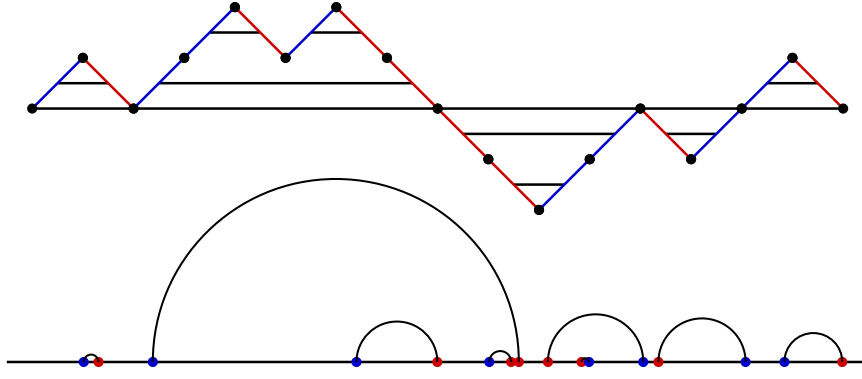
Figure 2: The Dyck matching π_{Dyck} associated to the Dyck bridge σ in the example of Figure 1.

Proposition 1 *For all independent spacing models*

$$
\mathbb{E}_N(H_{\text{Dyck}}) \sim
\begin{cases}
\frac{A_p}{2} N + \frac{2^p \Gamma(p-\frac{1}{2})}{4\Gamma(p+1)} N^{\frac{1}{2}+p} + \mathcal{O}(1) & p < \frac{1}{2} \\
\frac{1}{\sqrt{2\pi}} N \log N + \left(\frac{A_*}{2} + A_*'\right) N + \mathcal{O}(\log N) & p = \frac{1}{2} \\
\frac{2^p \Gamma(p-\frac{1}{2})}{4\Gamma(p+1)} N^{\frac{1}{2}+p} + \frac{A_p}{2} N + \mathcal{O}(N^{-\frac{1}{2}+p}) & p > \frac{1}{2}
\end{cases}
\tag{28}
$$

where A_p and A_* are model-dependent quantities, which are not larger than

$$
A_p^{\max} := \frac{2^{p+1}}{(1-2p)\Gamma(1-p)} , \qquad A_*^{\max} = \sqrt{\frac{2}{\pi}} \left(\log 2 + \gamma_E\right) . \tag{29}
$$

and

$$
A_*' = \frac{\gamma_E + 2\log 2 - 2}{\sqrt{2\pi}} . \tag{30}
$$

In particular, for the ES model

$$
A_p^{\text{ES}} = \frac{\Gamma\left(\frac{1}{2} - p\right)\Gamma(p+1)}{2^{p-1}\sqrt{\pi}\,\Gamma(2-p)} \qquad A_*^{\text{ES}} = \sqrt{\frac{2}{\pi}} \left(5\log 2 + \gamma_E - 4\right) . \tag{31}
$$

The remaining of this section is devoted to the proof of this proposition. First, we factorize the average over the instance $J = [\vec{s}, \sigma]$ in two independent averages, the average over σ and the one over $\vec{s}$ (see Definition 1):

$$
\begin{aligned}
\mathbb{E}_N(H_{\text{Dyck}}) = \mathbb{E}\left(\sum_{i=1}^N J_{i,\pi_{\text{Dyck}}(i)}\right) &= \overline{\sum_{i=1}^N \langle J_{i,\pi_{\text{Dyck}}(i)}\rangle} \\
&= B_N^{-1} \sum_\sigma \sum_{i=1}^N \left\langle |k_{\text{blue}}(i) - k_{\text{red}}(\pi_{\text{Dyck}}^\sigma(i))|^p \right\rangle ,
\end{aligned}
\tag{32}
$$

where in the last line we emphasize that π_{Dyck} depends only on σ, as stated in Remark 2, and we adopted again the shortcut $B_N = \binom{2N}{N}$ for the total number of configurations σ.

Due to the fact that the spacings s_i are independent, the quantity appearing above, $\langle |k_{\text{blue}}(i) - k_{\text{red}}(\pi_{\text{Dyck}}^\sigma(i))|^p \rangle$, only depends on the length $\|e\| = 2k+1$ of

the corresponding link $e = (i, \pi_{\text{Dyck}}(i))$, via the formula

$$S_k^{(p)} := \left\langle |k_{\text{blue}}(i) - k_{\text{red}}(\pi_{\text{Dyck}}^\sigma(i))|^p \right\rangle = \left\langle \left(\sum_{j=0}^{2k} s_j \right)^p \right\rangle, \qquad (33)$$

where the s_j's are i.i.d. variables sampled with the distribution $f(s)$ (that is, in the ES model, i.i.d. exponential random variables).

Then, as a general paradigm for observables of the form $\overline{\sum_{e \in \pi} \langle F(|e|) \rangle}$, we rewrite the sum over all possible σ and over all links e of a given matching $\pi = \pi_{\text{Dyck}}(\sigma)$ as a sum over the forementioned parameter k, with a suitable combinatorial factor $v_{N,k}$:

$$\overline{\sum_{e \in \pi} \langle F(|e|) \rangle} = B_N^{-1} \sum_{k=0}^{N-1} v_{N,k} \left\langle F\left(\sum_{j=0}^{2k} s_j \right) \right\rangle \qquad (34)$$

$$v_{N,k} = \sum_\sigma \sum_{e \in \pi_{\text{Dyck}}(\sigma)} \delta_{\|e\|, 2k+1} . \qquad (35)$$

(Note that $\sum_k v_{N,k} = N B_N = 2\binom{2N-1}{N}$). In particular, in our case,

$$\mathbb{E}_N(H_{\text{Dyck}}) = B_N^{-1} \sum_{k=0}^{N-1} v_{N,k} S_k^{(p)} . \qquad (36)$$

So we face two separate problems: (1) determining the combinatorial coefficients $v_{N,k}$, which are 'universal' (i.e., the same for all independent-spacing models, for all cost exponents p and for all observables F as above); (2) determining the quantities $S_k^{(p)}$, that is, the average over the Euclidean length $|e|$ of the link (which depends from the function $f(s)$ that defines the independent-spacing model, and from the exponent p).

For what concerns $S_k^{(p)}$, this can be computed exactly both in the US and ES cases: in the US case $S_k^{(p)} = (2k + 1)^p$, as in fact $|e| = 2k + 1 = \|e\|$ deterministically, while in the ES case $S_k^{(p)} = \Gamma(2k + 1 + p)/\Gamma(2k + 1)$. More generally, for any model with independent spacings we would have that $S_k^{(p)} = \int \mathrm{d}x\, x^p f^{*2k+1}(x)$ that is, the sum of $2k + 1$ i.i.d. random variables is distributed as the $(2k + 1)$-fold convolution of the single-variable probability distribution. For the ES case this is exactly the Gamma distribution $g_{2k+1}(s)$. Up to this point, we could have also evaluated the analogous quantity for the PPP model, although with a bigger effort (but, from Section 2.2, we know that this is not necessary).

For what concerns $v_{N,k}$, in Appendix A we prove that

$$v_{N,k} = C_k \left[4^{N-k-1} + \frac{N-k}{2} B_{N-k} \right] =: C_k V_{N-k-1} . \qquad (37)$$

In particular, the simple expression for V_{N-k-1} gives in a straightforward way

$$V(z) := \sum_{j=0}^\infty V_j z^j = (1 - 4z)^{-1} + (1 - 4z)^{-\frac{3}{2}} . \qquad (38)$$

13

We pause to study the distribution of the lengths $\|e\|$ of links in π_{Dyck}, that is, the normalised distribution (in k), with parameter N, given by the expression $v_{N,k}/(NB_N)$. It is known that planar secondary structures have a universal behaviour for the tail of such a distribution, with exponent $-\frac{3}{2}$. Indeed, by performing a large N expansion at fixed k, and then studying the large k behaviour, one has

$$
\begin{aligned}
\frac{v_{N,k}}{NB_N} &= \frac{C_k}{NB_N}\Big[4^{N-k-1} + \frac{N-k}{2}B_{N-k}\Big] \\
&\stackrel{N\to\infty}{\approx} 2\,C_k\,4^{-k} \stackrel{k\to\infty}{\approx} \sqrt{\frac{2}{\pi}}\,k^{-\frac{3}{2}}\,,
\end{aligned}
\tag{39}
$$

reproducing the known behaviour.

Equation (36), with the help of (37) and (38), can be used to relate the generating functions

$$
E(z;p) := \sum_{N=0}^{\infty} B_N \mathbb{E}_N(H_{\text{Dyck}})\, z^N
\tag{40}
$$

$$
S(z;p) := \sum_{k=0}^{\infty} C_k\, S_k^{(p)}\, z^k\,.
\tag{41}
$$

Lemma 5

$$
E(z;p) = z\,V(z)\,S(z;p)
\tag{42}
$$

Proof.

$$
\begin{aligned}
E(z;p) &:= \sum_{N=0}^{\infty} B_N \mathbb{E}_N(H_{\text{Dyck}}) z^N \\
&= \sum_{N=0}^{\infty}\sum_{k=0}^{N-1} C_k V_{N-k-1} S_k^{(p)} z^N \\
&= \Big(\sum_{k=0}^{\infty} C_k S_k^{(p)} z^k\Big)\Big(\sum_{N=k+1}^{\infty} V_{N-k-1} z^{N-k}\Big) \\
&= z\,V(z)\,S(z;p)
\end{aligned}
\tag{43}
$$

$\square$

The behaviour at large k of $S_k^{(p)}$ is determined by the theory of large deviations. Said heuristically, the sum of the $2k+1$ i.i.d. variables s_i concentrates on $2k+1$, with tails which are sufficiently tamed that the average of x^p is equal to $(2k+1)^p(1 + \mathcal{O}(k^{-1}))$. That is, $S_k^{(p)} \sim (2k+1)^p \sim 2^p k^p$, and we have

$$
C_k S_k^{(p)} \sim \frac{2^p}{\sqrt{\pi}} 4^k k^{p-\frac{3}{2}}\,.
\tag{44}
$$

We recall a fundamental fact in the theory of generating functions: the singularities of a generating function determine the asymptotic behaviour of its coefficients. In particular, the modulus of the dominant singularity (that nearest to the origin) determines the exponential behaviour, and the nature of the

14

singularity determines the subexponential behaviour (see [38] for a comprehensive treatment of singularity analysis, and Appendix B for a summary of results). This tells us that we just need an expression for $E(z; p)$ around its dominant singularity to extract asymptotic information on the total cost, i.e. we just need to evaluate $S(z; p)$ locally around the dominant singularity of $E(z; p)$.

First of all, one needs to locate the dominant singularity of $S(z; p)$ and compare it with the $z = \frac{1}{4}$ singularity of $V(z)$. From Equation 44, we find an exponential behaviour $\sim 4^n$ of the coefficient of $S(z; p)$, trivially due to the entropy of Dyck walks of length $2n$, thus, the singularity must be in $z = \frac{1}{4}$. Notice that this agrees with the dominant singularity of $V(z)$ (which also is, essentially, a generating function of Dyck walks up to algebraic corrections), so that both generating functions will combine to give the final average-cost asymptotics.

At the dominant singularity, the power-law behaviour of the coefficients is given by a generating function of the kind

$$S(z; p) = A_p + B_p(1 - 4z)^{g_p} + o((1 - 4z)^{g_p}), \tag{45}$$

where A_p encodes the regular terms at the singularity, and $o((1-4z)^{g_p})$ accounts for all other singular terms leading to non trivial subleading corrections (among them one finds power, logarithms... in the variable $1 - 4z$).

In fact, in such a simple situation as in our case, we expect a more precise behaviour, $S(z; p) = A_p(1 + \mathcal{O}(1 - 4z)) + B_p(1 - 4z)^{g_p}(1 + \mathcal{O}(1 - 4z))$, where we have two series of corrections, in integer powers, associated to the regular and singular parts of the expansion around the singularity (up to the special treatment of the degenerate case $g_p \in \mathbb{Z}$).

Notice that B_p and g_p can be found by computing the asymptotic behaviour of the coefficients of Equation 45

$$S_k^{(p)} \sim \frac{B_p}{\Gamma(-g_p)} 4^k k^{-g_p - 1} = \frac{2^p}{\sqrt{\pi}} 4^k k^{p - \frac{3}{2}}, \tag{46}$$

giving, by comparison with Equation 44,

$$g_p = \frac{1}{2} - p \qquad\qquad B_p = \frac{2^p \Gamma\left(p - \frac{1}{2}\right)}{\sqrt{\pi}}. \tag{47}$$

Nothing can be said on the coefficient A_p without performing the exact resummation of the generating function at the singularity (possibly, after having subtracted a suitable diverging part).

This analysis results in an asymptotic expression for $E(z; p)$:

$$E(z; p) \sim \frac{1}{4}\left[A_p(1 - 4z)^{-\frac{3}{2}} + \frac{2^p \Gamma\left(p - \frac{1}{2}\right)}{\sqrt{\pi}}(1 - 4z)^{-(1+p)}\right] \tag{48}$$

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

$$\begin{aligned}
E\left(z; p = \tfrac{1}{2}\right) &= \frac{1}{4}(1 - 4z)^{-\frac{3}{2}}\left[A_{\frac{1}{2}+\epsilon} + \frac{1}{\epsilon}\sqrt{\frac{2}{\pi}} + \sqrt{\frac{2}{\pi}}\log\left(\frac{1}{1 - 4z}\right) + o(\epsilon)\right] \\
&= \frac{1}{4}(1 - 4z)^{-\frac{3}{2}}\left[A^* + \sqrt{\frac{2}{\pi}}\log\left(\frac{1}{1 - 4z}\right)\right]
\end{aligned} \tag{49}$$

for $p = \frac{1}{2}$, where $\epsilon = p - \frac{1}{2}$. The hypothesis of $S(z; p)$ being non-singular in p implies that A_p must cancel the $\frac{1}{\epsilon}$ singularity, leaving a regular part

$$A_* = \lim_{\epsilon \to 0} \left[A_{\frac{1}{2}+\epsilon} + \frac{1}{\epsilon} \sqrt{\frac{2}{\pi}} \right]. \tag{50}$$

This set of results has remarkable consequences, as it unveils a certain universality feature for $E(z; p)$. In fact, for all models in our large classes, the nature of the dominant singularity of $E(z; p)$ is the same, giving a universal asymptotic scaling in N for the average cost of Dyck matchings. Moreover, in the $p \geq \frac{1}{2}$ regime, also the coefficient of the dominant singularity is universal.

We can now expand the generating function using standard techniques (Appendix B, [38]) and the fact that $B_N \sim \frac{4^N}{\sqrt{\pi N}}$, obtaining

$$\mathbb{E}_N \sim \begin{cases} \frac{A_p}{2} N + o(N) & p < \frac{1}{2} \\ \frac{1}{\sqrt{2\pi}} N \log N + o(N \log N) & p = \frac{1}{2} \\ \frac{2^p \Gamma(p-\frac{1}{2})}{4\Gamma(p+1)} N^{\frac{1}{2}+p} + o(N^{\frac{1}{2}+p}) & p > \frac{1}{2} \end{cases}. \tag{51}$$

and in fact, more precisely,

$$\mathbb{E}_N \sim \begin{cases} \frac{A_p}{2} N + \frac{2^p \Gamma(p-\frac{1}{2})}{4\Gamma(p+1)} N^{\frac{1}{2}+p} + \mathcal{O}(1) & p < \frac{1}{2} \\ \frac{1}{\sqrt{2\pi}} N \log N + \left(\frac{A_*}{2} + A_*'\right) N + \mathcal{O}(\log N) & p = \frac{1}{2} \\ \frac{2^p \Gamma(p-\frac{1}{2})}{4\Gamma(p+1)} N^{\frac{1}{2}+p} + \frac{A_p}{2} N + \mathcal{O}(N^{-\frac{1}{2}+p}) & p > \frac{1}{2} \end{cases}. \tag{52}$$

where the terms of the expansion are just the same for the $p < \frac{1}{2}$ and $p > \frac{1}{2}$ cases, but have been arranged differently, in the order of dominant behaviour. The quantity A_* has been defined in (50), for the behaviour of $S(z; \frac{1}{2})$, while the quantity

$$A_*' = \frac{\gamma_E + 2\log 2 - 2}{\sqrt{2\pi}} \tag{53}$$

is a further (universal) correction coming from taking into account how $S(z; \frac{1}{2})$ enters in $E(z; \frac{1}{2})$, via $V(z)$ (and γ_E is the Euler–Mascheroni constant).

The formulas above gives the precise asymptotics, up to relative corrections of the order N^{-1}. As a corollary, we have this very same behaviour in the PPP model, as, from Lemma 4 and Corollary 1, we know that also the relative corrections between ES and PPP models are of the order N^{-1}.

For higher-order corrections, one would need to take into account more subleading terms in Equation 45.

For the ES case the resummation of $E(z; p)$ can be performed analytically by writing the Catalan number in terms of Gamma functions, namely

$$C_k = \frac{4^k \Gamma\left(k + \frac{1}{2}\right)}{\sqrt{\pi} \Gamma(k + 2)}. \tag{54}$$

giving

$$S(z; p) = \Gamma(p+1) F \left(\begin{array}{c} \frac{p+1}{2}, \frac{p+2}{2} \\ 2 \end{array} \middle| 4z \right) \tag{55}$$

where F is the $_2F_1$ hypergeometric function (a reminder is in Appendix B, equation (86)). This allows for an explicit computation of the two non-universal quantities in our expansion:

$$A_p^{\text{ES}} = \frac{\Gamma\left(\frac{1}{2} - p\right)\Gamma\left(p + 1\right)}{2^{p-1}\sqrt{\pi}\,\Gamma\left(2 - p\right)} \qquad A_*^{\text{ES}} = \sqrt{\frac{2}{\pi}}\,\left(5\log 2 + \gamma_E - 4\right). \qquad (56)$$

note how the A_* and A_*' terms involve combinations of quantities of the same algebraic nature. See Appendix B for the details of the derivation.

Similar but more complex resummations seem possible in the independent spacing case, when the function $f(x)$ is a Gamma distribution, $f(x) = ag_a(ax)$ for $a \in \mathbb{N}/2$, and $S(z; p)$ is obtained in terms of generalised hypergeometric functions $_{k+1}F_k$. However no exact resummation seems possible for the US case (which would require a limit $a \to \infty$ in this procedure). Nonetheless, we are able to provide bounds on A_p^{US}.

Proposition 2

$$\frac{\Gamma\left(\frac{1}{2} - p\right)\Gamma\left(p + 1\right)}{2^{p-1}\sqrt{\pi}\,\Gamma\left(2 - p\right)} = A_p^{\text{ES}} \le A_p^{\text{US}} \le A_p^{\text{US,upper}} := \frac{2^{p+1}}{(1 - 2p)\Gamma\left(1 - p\right)}. \qquad (57)$$

Note that, due to the fact $\Gamma(\frac{1}{2} - p)\Gamma(1 - p) = \sqrt{\pi}\,2^{2p}\,\Gamma(1 - 2p)$, the bounds above can be rewritten as

$$\frac{1^{\overline{p}}}{(2 - 2p)^{\overline{p}}}\frac{2^{p+1}}{(1 - 2p)\Gamma\left(1 - p\right)} \le A_p^{\text{US}} \le \frac{2^{p+1}}{(1 - 2p)\Gamma\left(1 - p\right)}. \qquad (58)$$

Proof. We use Lemma 3 and the increasing monotony of $x^{\overline{p}}$ to write

$$(2k + 1)^{\overline{p}} \le (2k + 1)^p \le 2^p(k + \tfrac{3}{2} - p)^{\overline{p}} \le 2^p(k + 2 - p)^{\overline{p}} \qquad (59)$$

so that we can bound A_p^{US} from below by A_p^{ES}, and from above by the regular part of the expansion around $z = \frac{1}{4}$ of the generating function

$$2^p \sum_{k \ge 0} C_k(k + 2 - p)^{\overline{p}} z^k. \qquad (60)$$

This can be easily resummed to a hypergeometric function

$$\frac{2^p}{\Gamma\left(2 - p\right)} F\left(\begin{array}{c} \frac{1}{2}, 1 \\ 2 - p \end{array} \middle| 4z\right) \qquad (61)$$

whose regular part is (see Appendix B)

$$A_p^{\text{US,upper}} = \frac{2^{p+1}}{(1 - 2p)\Gamma\left(1 - p\right)}. \qquad (62)$$

$\square$

The ratio between the upper and lower bound, in the range $p \in [0, \frac{1}{2}]$ (where the corresponding term in the expansion is leading), is rather near to 1: it reaches 1 at the endpoints of the interval, and has a unique maximum, at the value p solving $2\psi(2 - 2p) = \psi(2 - p) + \psi(1 + p)$, or, by using a duplication formula for the digamma function, solving $\psi(p + 1) - \psi(p + \frac{1}{2}) = \ln 4 - \frac{1}{(1-p)(1-2p)} -$

$\pi \tan(\pi p)$. Numerically, the maximum is around $p = 0.2626\ldots$, and is valued $1.1445\ldots$

The analogous treatment in the degenerate case $p \to \frac{1}{2}$ gives bounds also on A_*^{US}

$$\sqrt{\frac{2}{\pi}} \left(\log 2 + \gamma_E + 4\left(\log 2 - 1\right)\right) \le A_*^{\mathrm{US}} \le \sqrt{\frac{2}{\pi}} \left(\log 2 + \gamma_E\right) . \qquad (63)$$

The fact that $A_p^{\mathrm{US}} \ge A_p^{\mathrm{ES}}$, that we used above, is not isolated, but is instead the consequence of the fact that the US distribution $f_{US}(x) = \delta(x-1)$ and the ES distribution $f_{ES}(x) = e^{-x}\Theta(x)$ are 'ordered', $f_{US}(x) \preceq f_{ES}(x)$, w.r.t. the notion of *transportation order* introduced in Appendix C. More generally, as implied by Lemmas 6 and 7 in the appendix, if for two probability distributions $f_1(x) \preceq f_2(x)$ we have determined $A_p^{f_1} \le A_p^{ub}$ and $A_p^{f_2} \ge A_p^{lb}$ (and the analogous statement for A_*), then, for every distribution $f_1(x) \preceq f(x) \preceq f_2(x)$ we have that $A_p^{lb} \le A_p^f \le A_p^{ub}$ (and the analogous statement for A_*), with no need of calculations (notice the inversion of upper and lower bounds due to the concavity of our cost function). This is due to the fact that, for all k, the integral $\int \mathrm{d}x \, x^p f^{*k}(x)$ is bounded as

$$\int \mathrm{d}x \, x^p f_2^{*k}(x) \le \int \mathrm{d}x \, x^p f^{*k}(x) \le \int \mathrm{d}x \, x^p f_1^{*k}(x) \qquad (64)$$

(first, by Lemma 6, $f_1^{*k} \preceq f^{*k} \preceq f_2^{*k}$, then by Lemma 7 and the fact that x^p is concave, we obtain the inequalities above).

In particular, the US distribution is the bottom element of the transportation-order semilattice, thus the upper bound in equation (57) is universal for all independent-spacing models.

4 Numerical results and the average cost of the optimal matching

Our main results concern the leading behaviour of the average cost of the Dyck matching, which, of course, provides an upper bound to the average cost of the optimal matching. The explicit investigation of small-size instances suggests that the optimal matching is often quite similar to the Dyck matching, in the sense that the symmetric difference between π_{opt} and π_{Dyck} typically consists of 'few' cycles, which are 'compact', in some sense. Thus, a natural question arises: could it be that the large-N average properties of optimal matchings and Dyck matchings the same? If not, in which respect do they differ? In order to try to answer this question, we have performed numerical simulations by generating random instances with measure μ_N^{PPP}, and we have computed the average cost associated to the optimal and to the Dyck matching.

Figure 3 gives a comparison between the two average costs by plotting their ratio as a function of N for various values of p. The corresponding fits seem to exclude the possibility that the limit for large N of the ratio of average costs go to zero algebraically in N (and also makes it reasonable that there are no logarithmic factors, although this is less evident), that is, these data suggest the content of Conjecture 1.
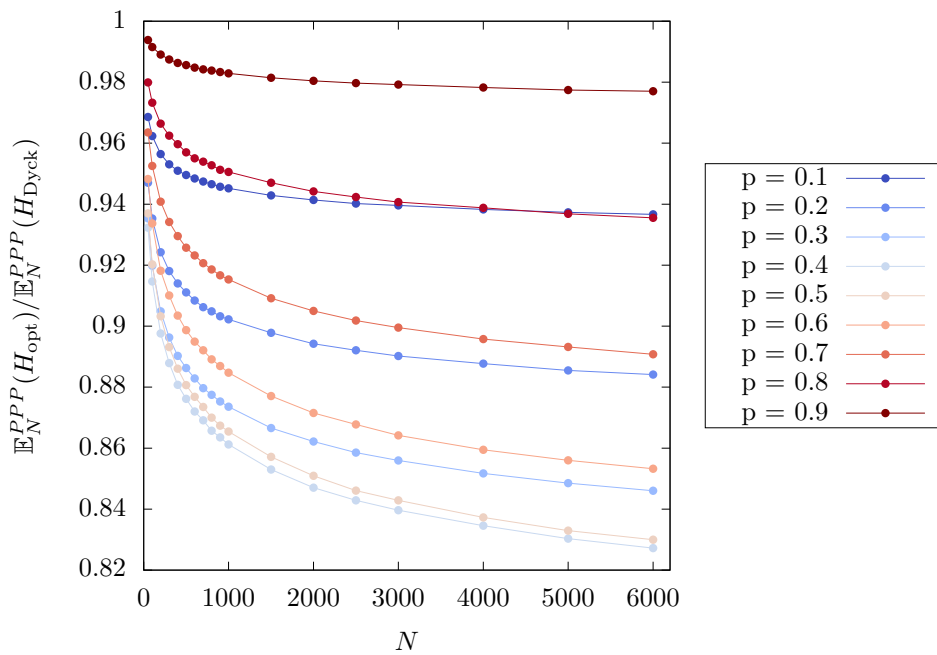
Figure 3: Ratio of the average cost of optimal matchings over that of Dyck matchings as a function of N. The number of simulated instances at each value of (p, n) is 10000, whenever $n \leq 4000$, and 5000, whenever $n = 5000$ or 6000.
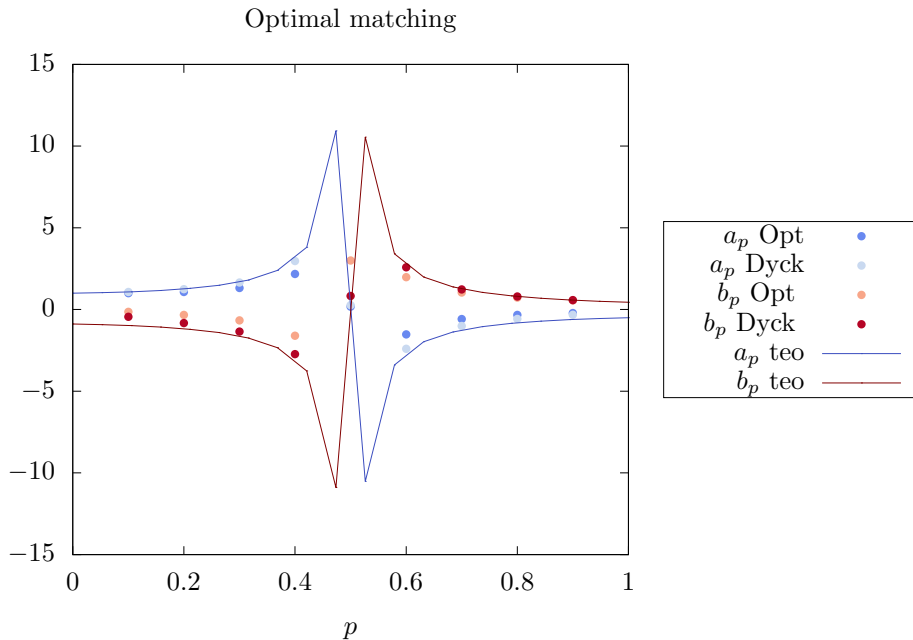
Figure 4: Fitted scaling coefficients as a function of p.

In order to further test this hypothesis, we fitted the optimal average cost to the same scaling behaviour found for the Dyck matching average cost, i.e.

$$\begin{cases} a_p N + b_p N^{\frac{1}{2}+p} & p \neq \frac{1}{2} \\ c\, N\big(\log N + d\big) & p = \frac{1}{2} \end{cases} \tag{65}$$

fixing the scaling exponents and aiming to compute the scaling coefficients. Notice that the term $a_p N$ is leading for $0 < p < \frac{1}{2}$, while $b_p N^{\frac{1}{2}+p}$ is leading for $\frac{1}{2} < p < 1$. Figure 4 summarizes the fitted parameters.

For the Dyck matching, the fitted parameter for the leading scaling coefficient agrees perfectly with the computed coefficient, as expected. The coefficient of the subleading term seems to agree with the computed coefficient in a less precise way, probably due to stronger higher-order corrections. For the optimal matching, the fitted coefficients behave qualitatively as the coefficients that we have computed for the Dyck case, but the agreement is visibly not quantitative. The fit seems to confirm the hypothesis that the two average costs have the same scaling exponents with different coefficients.

To completely confirm such hypothesis, we suggest that lower bounds for the cost could be computed. We expect such lower bounds to share the same scaling as that found in this paper, but with different constants. We leave such matter open for future work.

5 Conclusion

In this paper we have addressed the random Euclidean assignment problem in dimension 1, for points chosen in an interval, with a cost function which is an increasing and concave power law, that is $c(|x|) = |x|^p$ for $0 < p < 1$. We introduced a new special matching configuration associated to an instance of the problem, that we called the *Dyck matching*, as it is produced from the Dyck bridge that describes the interlacing of red and blue points on the domain.

As this is a deterministic configuration, described directly in terms of the instance, instead of involving a complex optimisation problem, this configuration is much more tractable than the optimal matching. On top of this fact, we can exploit a large number of nice facts, from combinatorial enumeration, which provides us also with several results which are exact at finite size, this being, to some extent, surprising. In particular, we computed the average cost of Dyck matchings under a particular choice of probability measure (the one in which the spacings among consecutive points are i.i.d. exponential variables). Finally, we performed numerical simulations that suggest that the average cost of Dyck matchings has the same scaling behaviour of the average cost of optimal matchings. We leave this claim as a conjecture. A promising way to prove this conjecture seems to be that of providing a lower bound on the average cost of optimal matchings with the same scaling as our upper bound, by producing "sufficiently many" or "sufficiently large" sets of edge-lengths which must be taken by the optimal solution. If we assume our conjecture, this result allows to fill in a missing portion in the phase diagram of the model, for what concerns the scaling of the average optimal cost as a function of p (see Figure 5). These new facts highlight a much richer structure that what could have been predicted in light of the previous results alone, with the concatenation of four distinct regions, and a special point with logarithmic corrections.

The case of uniformly spaced points needs further analysis in the region $p < 1/2$. One can define an interpolating family of independent spacing models, which encompasses both the ES and US cases, by taking as function $f(s)$ the Gamma distribution $\alpha g_\alpha(\alpha s)$, for $\alpha > 0$. For example, when α is an integer, each s_i is distributed as a sum of α i.i.d. exponential random variables, each with mean α^{-1}. The ES case is, of course, $\alpha = 1$, while, due to the central limit theorem, the US model is the limit as α tends to infinity. This generalised model appears to be treatable with the same technique that we employed for the pure ES case whenever α is a half-integer: the generating function of the average cost can be computed exactly in this case, and involves more and more complicated hypergeometric functions as α grows (namely, if $\alpha = k/2$, we have a $_kF_{k-1}$ hypergeometric function). Performing singularity analysis over such functions is a challenging task, which builds on classical results on generalised hypergeometric functions (due to Nørlund and Bühring), that we leave for future work.

On a parallel but distinct research direction, Dyck matchings may provide a fruitful framework to study other interesting features of the assignment problem, such as the ground-state degeneracy at $p = 1$ or the properties of the cycle decomposition of optimal matchings, w.r.t. the natural ordering along the domain segment.
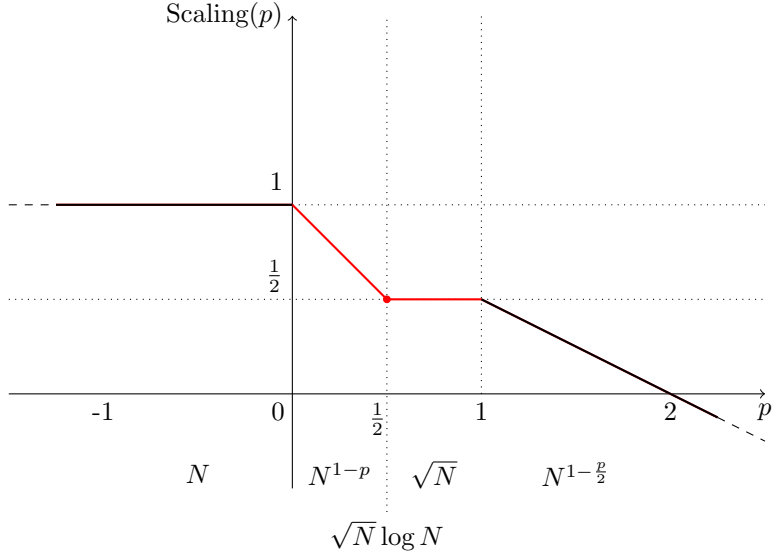
Figure 5: Scaling exponent of the average optimal cost as a function of p, in the case of cost function $c(|x|) = +|x|^p$ (that is, attractive case for $p > 0$ and repulsive case for $p < 0$). In red, our conjectured result. In black, results from [22]. Notice that we rescaled our result by a factor $(2N + 1)^{-p}$, i.e. we plotted the result for $\Omega = [0, 1]$.

A The coefficients $v_{N,k}$

The goal of this section is to compute the coefficients $v_{N,k}$, crucially used in the calculation of the average cost of the Dyck matching, starting with equation (36). These coefficients count, among the N edges of all the possible $\binom{2N}{N}$ Dyck matchings on $2N$ points, the number of edges e of length $\|e\| = 2k + 1$. That is, $v_{N,k} \frac{(N-1)!^2}{2(2N-1)!}$ is the probability that, taking a random Dyck matching π_{Dyck} uniformly, and an edge $e \in \pi_{\text{Dyck}}$ uniformly, we have $\|e\| = 2k + 1$.

Dyck matchings correspond to Dyck 'bridges', w.r.t. the notation introduced in Section 2.1. We proceed with the calculation by first computing the analogous quantity on a restricted ensemble, associated to Dyck 'excursions' (that is, the ordinary Dyck paths), which are Dyck bridges satisfying the extra condition $\sum_{i=1}^{j} \sigma_i \geq 0$ for all $1 \leq j \leq 2N$.

In the whole class of Dyck paths of length $2N$ there are

$$r_{N,k} := \frac{N - k + 1}{2} \, C_k \, C_{N-k} = \frac{1}{2} \, C_k \, B_{N-k} \qquad (66)$$

edges of length $\|e\| = 2k + 1$ (http://oeis.org/A141811), with $0 \leq k \leq n - 1$. These numbers obey the recursion relation, which determines them univocally (together with the initial conditions)

$$r_{N,k} = C_k C_{N-k-1} + \sum_{m=0}^{N-1} \left[r_{m,k} \, C_{N-m-1} + r_{N-m-1,k} \, C_m \right]. \qquad (67)$$

22

The recursion can be understood in terms of a first-return decomposition. If we decompose the path into its *first return*, i.e. the portion between its left endpoint and its first zero (say at position $2m+2$, $0 < m < n-1$), and into its *tail*, i.e. the remaining portion of the path on the right, then:

- the first term counts all the paths in which the link between the first step and the first zero is of the required length. The multiplicity 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 length hidden in the first excursion or in the tail of the path. To count links of the required length 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.

It is easy to prove by induction that

$$r_{N,k} = C_k \, R_{N-k-1} \tag{68}$$

and the recursion reduces to

$$R_s = C_s + 2 \sum_{m=1}^{s-1} C_{s-m} \, R_{m-1} \, . \tag{69}$$

By introducing the generating function

$$R(z) := \sum_{n \geq 0} R_n \, z^n \tag{70}$$

we get the equation

$$R(z) = C(z) + 2 \, z \, C(z) \, R(z) \tag{71}$$

and therefore

$$
\begin{aligned}
R(z) &= \frac{C(z)}{1 - 2 \, z \, C(z)} = C(z) B(z) = \frac{1}{2 \, z} \left[\frac{1}{\sqrt{1 - 4 \, z}} - 1 \right] \\
&= -\frac{1}{2 \, z} + \frac{1}{2 \, z} + \sum_{n \geq 0} \frac{B_{n+1}}{2} \, z^n = \sum_{n \geq 0} \frac{B_{n+1}}{2} \, z^n
\end{aligned}
\tag{72}
$$

indeed

$$\sum_{k=1}^{n} C_k B_{n-k} = \frac{B_{n+1}}{2} \, . \tag{73}$$

It follows that

$$R_{N-k-1} = \frac{B_{N-k}}{2} \tag{74}$$

as announced.

The preliminary computation of the $r_{N,k}$ coefficients suggests to use the same technique for the $v_{N,k}$, and provides an ingredient to write a recursion for

the $v_{N,k}$:

$$v_{N,k} = 2\,C_k\,B_{N-k-1} + 2\sum_{m=0}^{N-1}\left(r_{m,k}\,B_{N-m-1} + v_{N-m-1,k}\,C_m\right)$$

$$= 2\,C_k\,B_{N-k-1} + \sum_{m=k+1}^{N-1} C_k\,B_{m-k}\,B_{N-m-1} + 2\sum_{m=k+1}^{N-1} v_{m,k}\,C_{N-m-1}$$

$$= 2\,C_k\,B_{N-k-1} + \sum_{m=1}^{N-k-1} C_k\,B_m\,B_{N-k-m-1} + 2\sum_{m=1}^{N-k-1} v_{m+k,k}\,C_{N-k-m-1} \tag{75}$$

and if we again set

$$v_{N,k} := C_k V_{N-k-1} \tag{76}$$

we get

$$V_s = 2\,B_s + \sum_{m=1}^{s} B_m\,B_{s-m} + 2\sum_{m=1}^{s} V_{m-1}\,C_{s-m}$$

$$= B_s + \sum_{m=0}^{s} B_m\,B_{s-m} + 2\sum_{m=1}^{s} V_{m-1}\,C_{s-m} \tag{77}$$

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

We introduce now the generating function

$$V(z) := \sum_{k\geq 0} V_k\,z^k \tag{78}$$

to get the relation

$$V(z) = \frac{1}{\sqrt{1-4\,z}} + \frac{1}{1-4\,z} + 2\,z\,\frac{1-\sqrt{1-4\,z}}{2\,z}\,V(z) \tag{79}$$

so that

$$V(z) = \frac{1}{1-4\,z} + \frac{1}{(1-4\,z)^{\frac{3}{2}}} = \sum_{k\geq 0}\left[4^k + \frac{(2k+1)!}{(k!)^2}\right] z^k\,, \tag{80}$$

which is our seeked result. We can finally check that the recursion above is indeed satisfied, as

$$v_{N,k} = C_k\left[4^{N-k-1} + \frac{(2N-2k-1)!}{[(N-k-1)!]^2}\right]$$

$$= C_k\left[4^{N-k-1} + \frac{(N-k)^2}{2\,(N-k)}\,B_{N-k}\right] \tag{81}$$

$$= 4^{N-k-1}\,C_k + (N-k)\,r_{N,k}\,.$$

B Singularity analysis

Singularity analysis is a technique that allows to extract information on the coefficients of a generating function $f(z)$ when an explicit series expansion around $z = 0$ is not available. Roughly speaking, two main principles hold (see e.g. [38, pg. 227]):

1. the moduli of the singularities of f dictate the asymptotic exponential growth of its coefficients. If $z = a$ is a singularity of $f(z) = \sum_{n \geq 0} f_n z^n$, then $f_n \sim |a|^{-n}$;

2. the nature of the singularities of f dictate the asymptotic sub-exponential growth of its coefficients, i.e. they determine the (typically polynomial or logarithmic) function $\theta(n)$ such that $f_n \sim |a|^{-n} \theta(n)$.

We will specialise this analysis to the case of a single singularity, along the real positive axis, which is pertinent to series with positive coefficients, and no oscillatory behaviour. Generalizations of these principles (and of the related theorem below) for the case of multiple singularities at the same radius hold as well, but in our case are not relevant and will not be discussed.

The main result that we are going to need is a theorem (see [38, thm. VI.4]) that states that if $f(z)$ is a "well-behaved" complex function analytic in 0, with a singularity at $z = \zeta + i0$ such that

$$f(z) = \sigma(z/\zeta) + o(\tau(z/\zeta)) \tag{82}$$

for some functions $\sigma = \sum_{n \geq 0} \sigma_n z^n$ and $\tau = \sum_{n \geq 0} \tau_n z^n$ in the span of the reference set

$$\mathcal{S} = \left\{ (1-z)^\alpha \left(\frac{1}{z} \log \frac{1}{1-z} \right)^\beta \ \middle| \ \alpha, \beta \in \mathbb{C} \right\}, \tag{83}$$

then

$$f_n = \zeta^{-n} \sigma_n + o(\zeta^{-n} \tau_n) \,. \tag{84}$$

Here "well behaved" means that there exists an indented disk of radius bigger than ζ, with the indentation that specifically excludes $z = \zeta + i0$, where $f(z)$ can be analytically continued. This means that the theorem is applicable to functions with very general singularities (isolated poles, branch cuts, ...), and in particular to hypergeometric functions.

The reference set $\mathcal{S}$ is composed of functions whose expansion can be computed exactly thanks to generalizations of the binomial theorem. These functions are ubiquitous in series expansions around poles of complex functions, so that the theorem is extremely versatile.

In our specific case of the ES model, the generating function to study is of the form

$$f(z) = s(4z) F \left(\begin{array}{c} a, b \\ c \end{array} \middle| 4z \right) \tag{85}$$

where $s(z) \in \text{span}\,\mathcal{S}$ is singular at $z = 1$, and $F = {}_2F_1$ is the $(2,1)$-hypergeometric function defined by

$$F \left(\begin{array}{c} a, b \\ c \end{array} \middle| z \right) = \sum_{n \geq 0} \frac{\Gamma(n+a)}{\Gamma(a)} \frac{\Gamma(n+b)}{\Gamma(b)} \frac{\Gamma(c)}{\Gamma(n+c)} \frac{z^n}{n!} \,, \tag{86}$$

or, equivalently, by

$$F\left(\begin{array}{c} a,b \\ c \end{array}\middle| z\right) = \sum_{n\geq 0} s_n \frac{z^n}{n!} \,; \tag{87}$$

$$\frac{s_{n+1}}{s_n} = \frac{(a+n)(b+n)}{c+n} \,; \qquad s_1 = 1 \,. \tag{88}$$

To expand and study the hypergeometric function around $z = 1$, a celebrated 'inversion formula' due to Gauss is available

$$F\left(\begin{array}{c} a,b \\ c \end{array}\middle| z\right) = \frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)} F\left(\begin{array}{c} a,b \\ a+b+1-c \end{array}\middle| 1-z\right) + $$
$$\frac{\Gamma(c)\Gamma(a+b-c)}{\Gamma(a)\Gamma(b)} (1-z)^{c-a-b} F\left(\begin{array}{c} c-a,c-b \\ c+1-a-b \end{array}\middle| 1-z\right) \tag{89}$$

This formula restates the seeked expansion around $z = 0$ in terms of an expansion near the singularity at $z = 1$. As the hypergeometric function is analytic in $z = 0$, the singular behaviour at $z = 1$ of the right-hand side combination is described by the power-law prefactors in the inversion formula.

In our specific case, $a = \frac{p+1}{2}$, $b = \frac{p+2}{2}$ and $c = 2$ with $p \in [0,1]$, giving $c - a - b = \frac{1-2p}{2} \in [-\frac{1}{2}, \frac{1}{2}]$. Thus, the leading terms of the expansion of $f(z)$ are:

$$f(z) = s(4z)\left[\frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)} + \frac{\Gamma(c)\Gamma(a+b-c)}{\Gamma(a)\Gamma(b)}(1-4z)^{c-a-b}\right.$$
$$\left. + \Theta((1-4z),(1-4z)^{c-a-b+1})\right]. \tag{90}$$

The above expression is valid only for $p \neq \frac{1}{2}$. A limit procedure combines the diverging Γ's and the $(1-4z)$ term to give

$$f(z) = s(4z)\left[-\frac{1}{\Gamma(\frac{3}{4})\Gamma(\frac{5}{4})}\left(\ln(1-4z) + 2\gamma + \psi_0\left(\tfrac{3}{4}\right) + \psi_0\left(\tfrac{5}{4}\right)\right)\right.$$
$$\left. + \Theta((1-4z)\ln(1-4z))\right]. \tag{91}$$

where γ is the Euler-Mascheroni constant and ψ_0 is the digamma function. The limit is to be performed with care: each term must be written as a function of $\epsilon = p - \frac{1}{2}$ and expanded in powers series. The expansion of the hypergeometric functions must be performed using their definition. When everything is expanded, $o(\epsilon)$ are discarded taking the limit $\epsilon \to 0$, and the leading terms in the $(1-4z)$ are found by taking $n = 0$ in the sum of the hypergeometric function definition.

C Transportation order

Let us introduce a notion of ordering among distributions with the same average, which is a version analogue (but slightly different) from the several ones already existing in literature (such as *second-order stochastic dominance*, or *variability*

order, or *convex order*, see e.g. [39]). We shall call our order the *transportation order*.

If $f(x)$ and $g(x)$ are probability distributions on $\mathbb{R}$ with the same average, we say that $f \preceq g$ if there exists a non-negative function $h(x, y)$ such that, for all x,

$$\int \mathrm{d}y\, h(x, y) = 1 \qquad\qquad \int \mathrm{d}y\, y\, h(x, y) = x \qquad (92)$$

and

$$g(y) = \int \mathrm{d}x\, h(x, y) f(x)\,. \qquad (93)$$

We call h the *transportation map* of the pair (f, g).

We have two crucial facts:

Lemma 6 *Convolution preserves transportation order, that is, if $f_1 \preceq g_1$ and $f_2 \preceq g_2$, then $f_1 * f_2 \preceq g_1 * g_2$.*

Proof. Let us call h_1 and h_2 the associated transportation maps. We have

$$
\begin{aligned}
(f_1 * f_2)(x) &= \int \mathrm{d}x'\, f_1(x') f_2(x - x') \\
&= \int \mathrm{d}x' \mathrm{d}y' \mathrm{d}y\, h_1(x', y') h_2(x - x', y - y') f_1(x') f_2(x - x')
\end{aligned}
\qquad (94)
$$

and

$$
\begin{aligned}
(g_1 * g_2)(y) &= \int \mathrm{d}y'\, g_1(y') g_2(y - y') \\
&= \int \mathrm{d}x' \mathrm{d}y' \mathrm{d}x\, h_1(x', y') h_2(x - x', y - y') f_1(x') f_2(x - x')\,.
\end{aligned}
\qquad (95)
$$

This shows that the function

$$
h(x, y) = \begin{cases} \dfrac{\big((h_1 f_1) * (h_2 f_2)\big)(x, y)}{(f_1 * f_2)(x)} & (f_1 * f_2)(x) > 0 \\ 0 & (f_1 * f_2)(x) = 0 \end{cases}
\qquad (96)
$$

(where convolution of (hf) is both over x and y) is a candidate transportation map for the pair $(f_1 * f_2, g_1 * g_2)$, and is at sight a non-negative function. We shall only verify that the conditions (92) are satisfied. We have

$$
\begin{aligned}
\int \mathrm{d}y\, h(x, y) &= \frac{1}{(f_1 * f_2)(x)} \int \mathrm{d}x' \mathrm{d}y' \mathrm{d}y\, h_1(x', y') h_2(x - x', y - y') f_1(x') f_2(x - x') \\
&= \frac{1}{(f_1 * f_2)(x)} \int \mathrm{d}x' \left(\int \mathrm{d}y_1\, h_1(x', y_1) \right) \left(\int \mathrm{d}y_2\, h_2(x - x', y_2) \right) \\
&\qquad\qquad \times f_1(x') f_2(x - x') \\
&= \frac{1}{(f_1 * f_2)(x)} \int \mathrm{d}x'\, f_1(x') f_2(x - x') = 1
\end{aligned}
\qquad (97)
$$

and

$$\int \mathrm{d}y\,(y-x)h(x,y) = \frac{1}{(f_1 * f_2)(x)} \int \mathrm{d}x' \mathrm{d}y' \mathrm{d}y$$
$$\times \left((y-y') - (x-x') + y' - x'\right) h_1(x',y') h_2(x-x', y-y') f_1(x') f_2(x-x')$$
$$= \frac{1}{(f_1 * f_2)(x)} \int \mathrm{d}x'\, f_1(x') f_2(x-x')$$
$$\times \left[\left(\int \mathrm{d}y_1\,(y_1 - x') h_1(x', y_1)\right) \left(\int \mathrm{d}y_2\, h_2(x-x', y_2)\right) \right.$$
$$\left. + \left(\int \mathrm{d}y_1\, h_1(x', y_1)\right) \left(\int \mathrm{d}y_2\,(y_2 - (x-x')) h_2(x-x', y_2)\right) \right]$$
$$= 0\,.$$
$$(98)$$

$\square$

Lemma 7 *The average of convex/concave observables is monotone w.r.t. transportation order, that is, if $f \preceq g$ and $A(x)$ is convex, then $\int \mathrm{d}x\, A(x) f(x) \leq \int \mathrm{d}x\, A(x) g(x)$, and if $A(x)$ is concave then the opposite inequality holds.*

Proof. Let us call h the transportation maps. We have

$$\Phi(x) := \int \mathrm{d}x\, A(x) g(x) - \int \mathrm{d}x\, A(x) f(x) = \int \mathrm{d}y\,(A(y) - A(x)) h(x,y) f(x) \tag{99}$$

If $A(x)$ is convex, then for all x there exists $m(x)$ such that $A(y) \geq A(x) + m(x)(y-x)$, thus

$$\Phi(x) \geq \int \mathrm{d}y\, m(x)(y-x) h(x,y) f(x) \tag{100}$$

As $\int \mathrm{d}y\,(y-x) h(x,y) = 0$ for all x, and $f(x)$ is non-negative, we get $\Phi(x) \geq 0$, as was to be proven. For the concave case, of course it suffices to consider the convex function $-A(x)$. $\square$

References

[1] Henri Orland. Mean-field theory for optimization problems. *Le Journal de Physique - Lettres*, 46(17):773–770, 1985.

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

[3] Marc Mézard and Giorgio Parisi. Mean-field equations for the matching and the travelling salesman problems. *Europhysics Letters*, 2(12):913–918, 1986.

[4] Sergio Caracciolo, Matteo P. D'Achille, Enrico M. Malatesta, and Gabriele Sicuro. Finite size corrections in the random assignment problem. *Phys. Rev. E*, 95(052129), 2017.

[5] László Lovász and Michael D. Plummer. *Matching Theory*, volume 367 of *AMS Chelsea Publishing Series*. North-Holland; Elsevier Science Publishers B.V., 2009.

[6] Giorgio Parisi. A conjecture on random bipartite matching. `arXiv:cond-mat/9801176`, 1998.

[7] D. Coppersmith and G. Sorkin. Constructive bounds and exact expectations for the random assignment problem. *Random Structures and Algorithms*, 15:113—144, 1999.

[8] David J. Aldous. The $\zeta(2)$ limit in the random assignment problem. *Random Struct. Algorithms*, 2:381–418, 2001.

[9] Chandra Nair, Balaji Prabhakar, and Mayank Sharma. Proofs of the Parisi and Coppersmith-Sorkin random assignment conjectures. *Random Struct. Algorithms*, 27(4):413–444, 2005.

[10] Svante Linusson and Johan Wastlund. Completing a $(k-1)$-assignment. *Combinatorics Probability & Ccomputing*, 16(4):621–629, JUL 2007.

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

[12] Miklós Ajtai, János Komlós, and Gabor Tusnády. On optimal Matchings. *Combinatorica*, 4(4):259–264, 1984.

[13] Sergio Caracciolo, Carlo Lucibello, Giorgio Parisi, and Gabriele Sicuro. Scaling hypothesis for the euclidean bipartite matching problem. *Phys. Rev. E*, 90:012118, 2014.

[14] Sergio Caracciolo and Gabriele Sicuro. Scaling hypothesis for the Euclidean bipartite matching problem. II. Correlation functions. *Phys. Rev. E*, 91(062125), 2015.

[15] Sergio Caracciolo and Gabriele Sicuro. Quadratic stochastic euclidean bipartite matching problem. *Phys. Rev. Lett.*, 115(23):230601, 2015.

[16] Luigi Ambrosio, Federico Stra, and Dario Trevisan. A PDE approach to a 2-dimensional matching problem. *Probab. Theory Relat. Fields*, pages 1–45, 2018.

[17] Sergey Bobkov and Michel Ledoux. One-dimensional empirical measures, order statistics, and Kantorovich transport distances. *Bullettin of the AMS*, 2017.

[18] Sergio Caracciolo, Matteo P. D'Achille, and Gabriele Sicuro. Anomalous scaling of the optimal cost in the one-dimensional random assignment problem. *J. Statist. Phys.*, 174(4):846–864, 2019.

[19] Michel Talagrand. Scaling and non-standard matching theorems. *C. R. Acad. Sci. Paris, Ser. I*, 356(6):692–695, 2018.

[20] Sergio Caracciolo and Gabriele Sicuro. On the one dimensional euclidean matching problem: exact solutions, correlation functions and universality. *Phys. Rev. E*, 90(042112), 2014.

[21] Sergio Caracciolo, Matteo P. D'Achille, and Gabriele Sicuro. Random euclidean matching problems in one dimension. *Phys. Rev. E*, 96:042102, 2017.

[22] Sergio Caracciolo, Andrea Di Gioacchino, Enrico M. Malatesta, and Luca G. Molinari. Selberg integrals in 1d random euclidean optimization problems. *J. Stat. Mech.*, 2019.

[23] McCann Robert. Exact solutions to the transportation problem on the line. *Proc. R. Soc. A: Math., Phys. Eng Sci*, 455:1341–1380, 1999.

[24] Elena Boniolo, Sergio Caracciolo, and Andrea Sportiello. Correlation function for the grid-poisson euclidean matching on a line and on a circle. *J. Stat. Mech.*, 11:P11023, 2014.

[25] Paul G. Higgs. RNA secondary structure: physical and computational aspects. *Q. Rev. Biophys.*, 33(03):199–253, 2000.

[26] Henri Orland and Anthony Zee. RNA folding and large N matrix theory. *Nuclear Physics B*, 620(3):456–476, 2002.

[27] Graziano Vernizzi, Henri Orland, and Anthony Zee. Enumeration of RNA Structures by Matrix Models. *Phys. Rev. Lett.*, 94:168103, 2005.

[28] S. K. Nechaev, A. N. Sobolevski, and O. V. Valba. Planar diagrams from optimization for concave potentials. *Phys. Rev. E*, 87(1):012102, 2013.

[29] R Bundschuh and Terence Hwa. Statistical mechanics of secondary structures formed by random RNA sequences. *Phys. Rev. E*, 65(3):031903, 2002.

[30] M. Mller. Statistical physics of RNA folding. *Phys. Rev. E*, 67:021914, 2003.

[31] Paul G. Higgs. Overlaps between RNA secondary structures. *Phys. Rev. Lett.*, 76(4):704, 1996.

[32] A Pagnani, G Parisi, and F. Ricci-Tersenghi. Glassy Transition in a Disordered Model for the RNA Secondary Structure. *Phys. Rev. Lett.*, 84(9):2026–2029, 2000.

[33] Enzo Marinari, Andrea Pagnani, and Federico Ricci-Tersenghi. Zero-temperature properties of RNA secondary structures. *Phys. Rev. E*, 65:041919, 2002.

[34] Jim Pitman and Marc Yor. A guide to Brownian motion and related stochastic processes. `arXiv:1802.09679`, 2018.

[35] J. Delon, J. Salomon, and A. Sobolevski. Local Matching Indicators for Transport Problems with Concave Costs. *SIAM Journal on Discrete Mathematics*, 26(2):801–827, 2012.

[36] R.L. Graham, D.E. Knuth, and O. Patashnik. *Concrete mathematics.* Addison-Wesley, Reading, MA, 1989.

[37] E. Artin. *The Gamma Function.* Holt, Rinehart, Winston, Inc., 1964.

[38] Philippe Flajolet and Robert Sedgewick. *Analytic combinatorics*. Cambridge University Press, 2009.

[39] H. Colonius. *Mathematical Models of Perception and Cognition Volume I: A Festschrift for James T. Townsend Joseph W. Houpt, Leslie M. Blaha eds.*, chapter Stochastic orders of variability, Ch. 3. Psychology Press, 2016.