The Number of Optimal Matchings for Euclidean Assignment on the Line

Sergio Caracciolo[1] · Vittorio Erba[1] · Andrea Sportiello[2]

Abstract

We consider the Random Euclidean Assignment Problem in dimension $d = 1$, with linear cost function. In this version of the problem, in general, there is a large degeneracy of the ground state, i.e. there are many different optimal matchings (say, $\sim \exp(S_N)$ at size N). We characterize all possible optimal matchings of a given instance of the problem, and we give a simple product formula for their number. Then, we study the probability distribution of S_N (the zero-temperature entropy of the model), in the uniform random ensemble. We find that, for large N, $S_N \sim \frac{1}{2} N \log N + Ns + \mathcal{O}(\log N)$, where s is a random variable whose distribution $p(s)$ does not depend on N. We give expressions for the moments of $p(s)$, both from a formulation as a Brownian process, and via singularity analysis of the generating functions associated to S_N. The latter approach provides a combinatorial framework that allows to compute an asymptotic expansion to arbitrary order in $1/N$ for the mean and the variance of S_N.

1 Introduction

The Euclidean Assignment Problem (EAP) is a combinatorial optimization problem in which one has to pair N white points to N black points minimizing a total cost function that depends on the Euclidean distances among the points. The Euclidean Random Assignment Problem (ERAP) is the statistical system in which these $2N$ points are drawn from a given probability measure. In the latter version of the problem, one is interested in characterizing the statistical properties of the optimal solution, such as its average cost, average structural properties, etc…

✉ Vittorio Erba
vittorio.erba@unimi.it

[1] Dipartimento di Fisica, University of Milan and INFN, via Celoria 16, 20133 Milan, Italy

[2] LIPN, and CNRS, Université Paris 13, Sorbonne Paris Cité, 99 Av. J.-B. Clément, 93430 Villetaneuse, France

The EAP problem has applications in Computer Science, where it has been used in computer graphics, image processing and machine learning [20], and it is also a discretised version of a problem in functional analysis, called *optimal transport problem*, where the N-tuples of points are replaced by probability measures. The optimal transport problem has recently seen a growing interest in pure mathematics, where it has found applications in measure theory and gradient flows [2].

To be more definite, the EAP problem is the optimization problem defined by

$$\min_{\pi \in \mathfrak{S}_N} H_J(\pi) \tag{1}$$

where

- $J = (\{w_i\}, \{b_j\})$ is the *instance* of the problem, i.e., in the Euclidean version in dimension d, the datum of the positions of N white points $\{w_i\}$ and N black points $\{b_j\}$ in $\mathbb{R}^d$;
- π is a bijection between the two sets of points, linking biunivocally each white point to a unique black point. In other words, is a perfect matching. We denote by $\mathfrak{S}_N$ the set of all possible bijections between sets of size N;
- $H_J(\pi)$ is the cost function

$$H_J(\pi) = \sum_{i=1}^{N} c\left(\text{dist}\left(w_i, b_{\pi(i)}\right)\right) \tag{2}$$

i.e. the sum of the costs of the links of π, where a link is weighted using a *link cost function* $c(x) : \mathbb{R}^+ \to \mathbb{R}$, depending only on the Euclidean distance among the two points.

In the EAP, J is considered as fixed, while in the ERAP, J is a random variable with a fixed probability distribution.

It is a longstanding project to understand the "phase diagram" of the model, in the plane (p, d), say, for definiteness, in the version of this problem where J is given by $2N$ i.i.d. points from the d-dimensional hypercube $[0, 1]^d$, and the link cost function is given by $c(x) = x^p$. Even the definition of phase diagram requires a clarification, as it is not obvious which observable should be adopted as order parameter. A first relevant quantity is the scaling in N of the cost of the optimal matching, a quantity that, as we will see in a moment, undergoes phase transitions that are connected to changes in the structural properties of the optimal matchings.

This phase diagram is, to a certain extent, still mysterious, although some progress has been made recently (see for example [1]), and all the information collected so far is coherent with the assumption that, for all (p, d), there exist exponents γ and γ' such that the optimal cost at size N scales as $E_N \sim N^\gamma (\ln N)^{\gamma'}$, where $\gamma = \gamma(p, d)$ is continuous and piece-wise smooth, and $\gamma' = \gamma'(p, d)$ is zero everywhere except possibly on the lines of discontinuity of $\gamma(p, d)$ (which are the critical lines). It is possible that one critical line is the half-line $d = 2$ and $p \geq 1$, and two other critical lines reach a tricritical point $(p, d) = (1, 2)$, starting from $d = 0$, and passing through $p = \frac{1}{2}$ and $p = 1$ when $d = 1$. In particular, when $d = 1$, the analytical characterization of optimal solutions is quite tractable [4,16], and the emerging landscape is as follows:

- For $p > 1$, the optimal matching is *ordered*[1] independently on J, and this is due to the fact that the link cost function is increasing and convex. In this case, plenty of results have been obtained on the statistics of the average optimal cost [7–9,11].
- For $0 < p < 1$, the optimal matching is *non-crossing*, meaning that pairs of matched points are either nested one inside the other, or disjoint (that is, matched pairs do not interlace) [16]. This property is imposed by the concavity of the link cost function. In this case, the optimal matching is not uniquely determined by the ordering of the points (although the viable candidates are typically reduced, roughly, from $n!$ to $\sqrt{n!}$), and a comparatively smaller number of results has been found so far for the random version of the problem [3,5,10,15]. Note that, remarkably, it is expected that, within this interval, there are two transitions: one at $p = 1$, with $\gamma'(1, 1) = 0$, somewhat corresponding to the ordered/non-crossing structural transition, and one at $p = \frac{1}{2}$, with $\gamma'(\frac{1}{2}, 1) = 1$, corresponding to a transition in which, in the cost of the optimal matching, the leading contribution comes from the few longest edges (of length $\mathcal{O}(1)$, when $p > \frac{1}{2}$) or from the many shortest edges (of length $\mathcal{O}(N^{-1})$, when $p < \frac{1}{2}$), and the logarithmic correction at $p = \frac{1}{2}$ comes from the fact that, in this case, all length scales contribute to the leading part of the optimal cost [5].
- For $p < 0$, due to the fact that an overall positive factor in $c(x)$ is irrelevant in the determination of the optimal matching, it is questionable if one should consider the analytic continuation of the cost function $c(x) = x^p$ (which has the counter-intuitive property that the preferred links are the longest ones), or of the cost function $c(x) = p\, x^p$ (which has the property that the preferred links are the shortest ones, and that the limit $p \to 0$ is well-defined, as it corresponds to $c(x) = \log x$, but has the disadvantage of having average cost $-\infty$ when $p \leq -d$). In the first case, the optimal matching is *cyclical*, meaning that the permutation that describes the optimal matching has a single cycle, and some result on the average optimal cost where obtained in [7]. In the second case, in the pertinent range $-d < p \leq 0$, it seems that the qualitative features of the $0 < p < \frac{1}{2}$ regime are preserved.

We notice that in all the cases above, as well as in all the cases with $d \neq 1$ and any p, the optimal matching is 'effectively unique', meaning that it is almost surely unique, and, even in presence of an instance showing degeneracy of ground states, almost surely an infinitesimal perturbation immediately lifts the degeneracy.

This generic non-degeneracy property does not hold only for the $d = p = 1$ case (note that $p = 1$ is the exponent at which the cost function changes concavity, that is, the critical point for the ordered/non-crossing structural transition). It is known [4] that in this case there are (almost surely) at least two distinct optimal matchings for each instance of the problem, the *ordered matching* and the *Dyck matching* [5] which coincide only in the rather atypical case (with probability $N!/(2N − 1)!!$) in which, for all i, the i-th white and black points are consecutive along the segment.

These observations suggest some fundamental questions for the $(d, p) = (1, 1)$ problem:

1. is it possible to characterize all the optimal matchings of a fixed instance J of the EAP?
2. how many optimal configurations are there for a fixed instance J of the EAP?
3. for random J's, what are the statistical properties of the number of optimal configurations?

The aim of this paper is to answer the three questions above.

---

[1] The ordered matching is the one in which the first white point from the left is matched with the first black point and so on, and is represented by the identity permutation $\pi_{\mathrm{ord}}(i) = i$ if the points are sorted by increasing coordinate.

Our interest is not purely combinatorial. In fact, the ERAP is a well-known and well-studied toy model of Euclidean spin glass [6,17–19]. The characterization of the set $\mathcal{Z}_J \subseteq \mathfrak{S}_N$ of the optimal matchings of J is thus related to the computation of the zero-temperature partition function $Z_J = |\mathcal{Z}_J|$ of the disordered model, and of its zero-temperature entropy $S_N(J) = \log(Z_J)$.

In this manuscript we will focus on the statistical properties of the zero-temperature entropy, the thermodynamic potential that rules the physics of the model when the disorder is *quenched*, i.e. when the timescale of the dynamics of the disorder degrees of freedom of the model is much larger than that of the microscopic degrees of freedom. We leave to future investigations the study of the *annealed* and *replicated* partition functions Z_J and Z_J^k, which are less fundamental from the point of view of statistical physics, but, as we will show elsewhere, have remarkable combinatorial and number-theoretical properties.

1.1 Summary of Results

In Sect. 2.1 we prove that, given a fixed instance J of the $(d, p) = (1, 1)$ EAP problem, a matching π is optimal if and only if

$$k_{\mathrm{LB}}(z) = k_\pi(z) \tag{3}$$

where $k_{\mathrm{LB}}(z)$ is a function of J that counts the difference of the numbers of white and black points on the left of z, and $k_\pi(z)$ counts the number of links of π whose endpoints lie on opposite sides of z.

In Sect. 2.2, we show that $\mathcal{Z}_J$, the set of optimal matchings, depends on J only through the ordering of the points, while it is independent on their positions (provided that the ordering is not changed). Using a straightforward bijection between the ordering of bi-colored point configurations on the line and a class of lattice paths (the *Dyck bridges*), we provide a combinatorial recipe to construct the set $\mathcal{Z}_J$. As a corollary, we obtain a rather simple product formula for the cardinality $Z_J := |\mathcal{Z}_J|$, that is, roughly speaking,

$$Z_J = \prod_{\substack{\text{descending} \\ \text{steps}}} \text{height of the step} \tag{4}$$

where the product runs over the descending steps of the Dyck bridge associated with the ordering of the points of configuration J, and the height of a step is, roughly speaking, the absolute value of its vertical position. This implies an analogous sum formula for the entropy $S_N(J) = \log(Z_J)$ (we stress that the size of J equals $2N$ using a subscript).

Then, we study the statistics of $S_N(J)$ when J is a random instance of the problem of fixed size N. Our techniques apply equally well, and with calculations that can be performed in parallel, to two interesting statistical ensembles of $2N$ points:

Dyck bridges: the case in which white and black points are just i.i.d. on the unit interval.
Dyck excursions: the restriction of the previous ensemble to the case in which, for all $z \in [0, 1]$, there are at least as many white points on the left of z than black points.

The fact that we can study these two ensembles in parallel is present also in our study for the distribution of the energy distribution of the "Dyck matching", that we perform elsewhere [5,10].

In Sect. 3.1, we highlight a connection between S_N (in the two ensembles) and the observable

$$s[\sigma] = \int_0^1 dt \log\left(|\sigma(t)|\right) \tag{5}$$

over Brownian bridges (or excursions) σ. Simple scaling arguments imply that

$$s = \lim_{N\to\infty} \frac{S_N - \frac{1}{2}N\log N}{N} \tag{6}$$

is a random variable with a non-trivial limit distribution for large N.

We provide integral formulas for the integer moments of s, and we use these formulas to compute analytically its first two moments in the two ensembles.

In Sect. 3.2, we complement this analysis with a combinatorial framework at finite size N. We use this second approach to provide an effective strategy for the computation of finite-size corrections, that we illustrate by calculating the first and second moment, for both bridges and excursions.

In particular, we can establish that

$$S_N \stackrel{d}{=} \tfrac{1}{2}N\log N + Ns + \mathcal{O}(\log(N)) \tag{7}$$

where s is a random variable whose distribution depends on the ensemble (among bridges and excursions). For Dyck bridges we have

$$\begin{aligned}
\langle s\rangle_B &= -\frac{\gamma_E + 2}{2} + \mathcal{O}\left(\frac{\log N}{\sqrt{N}}\right) \\
\langle s^2\rangle_B &= \frac{4}{3} + \frac{\gamma_E^2}{4} + \gamma_E - \frac{\pi^2}{72} + \mathcal{O}\left(\frac{(\log N)^2}{\sqrt{N}}\right)
\end{aligned} \tag{8}$$

and for Dyck excursions we have

$$\begin{aligned}
\langle s\rangle_E &= -\frac{\gamma_E}{2} + \mathcal{O}\left(\frac{\log N}{\sqrt{N}}\right) \\
\langle s^2\rangle_E &= \frac{\gamma_E^2}{4} + \frac{5\pi^2}{24} - 2 + \mathcal{O}\left(\frac{(\log N)^2}{\sqrt{N}}\right)
\end{aligned} \tag{9}$$

In Sect. 4 we provide numerical evidence that the distribution of the rescaled entropy s is non-Gaussian (but we leave unsolved the question whether the centered distribution for the excursions is an even function), and we confirm that the predicted values for the first two moments match with the simulated data.

With both our approaches it seems possible to access also higher-order moments, however we cannot prove at present that, for any finite moment, the evaluation can be performed in closed form, and that (as we conjecture) the result is in the form of a rational polynomial that only involves γ_E and (multiple) zeta functions. We will investigate these aspects in future works.

2 Optimal Matchings at $p = 1$

In the following, we focus on the EAP with $p = d = 1$. We have N white points $W = \{w_i\}_{i=1}^N$ and N black points $B = \{b_i\}_{i=1}^N$ on a segment $[0, L]$ (i.e., we have 'closed' boundary

conditions, instead of 'periodic', as we would have if the points were located on a circle). We assume that the instance is generic (i.e., no two coordinates are the same), and we label the points so that the lists above are ordered ($w_i < w_{i+1}$ and $b_i < b_{i+1}$).

2.1 Characterization of Optimal Matchings

We start by giving an alternate, integral representation of the cost function $H_J(\pi)$. For a given matching π, and every $z \in \mathbb{R}$ which is not a point of W or B, define the function

$$k_\pi(z) = \sum_{i=1}^{N} \kappa(z, w_i, b_{\pi(i)}) \qquad \kappa(z, x, y) = \begin{cases} 1 \text{ if } x < z < y \text{ or } y < z < x \\ 0 \text{ otherwise} \end{cases} \tag{10}$$

In words, $\kappa(z, x, y)$ is just the indicator function over the segment with extreme points x and y, and $k_\pi(z)$ counts the number of links of π that have endpoints on opposite sides of z.

Furthermore, define the function

$$k_{\mathrm{LB}}(z) = \big| \#(W \cap [0, z]) - \#(B \cap [0, z]) \big|, \tag{11}$$

where $\#(I)$ denotes the cardinality of the set I. As well as $k_\pi(z)$, also k_{LB} is defined for all $z \in \mathbb{R} \setminus (W \cup B)$, and counts the excess of black or white points on the left of z.

Then, we have two simple observations

Proposition 1 *At $p = 1$*

$$H_J(\pi) = \int \mathrm{d}z \, k_\pi(z). \tag{12}$$

Moreover, at $p = 1$, for all π and all z,

$$k_\pi(z) \geq k_{\mathrm{LB}}(z). \tag{13}$$

This has the immediate corollary that, for all π,

$$H_J(\pi) \geq H_J^{\mathrm{LB}} := \int \mathrm{d}z \, k_{\mathrm{LB}}(z). \tag{14}$$

Now, call π_{id} the identity permutations, that is the so-called *ordered matching*. By simple inspection, we have that $H_J(\pi_{\mathrm{id}}) = H_J^{\mathrm{LB}}$. This implies that π_{id} is optimal, and more generally

Corollary 1 *$\pi \in \mathcal{Z}_J$ iff the functions k_{LB} and k_π coincide.*

See Fig. 1 for the description of all optimal matchings at $N = 2$.

In the following we will provide a simple algorithm to construct the optimal matchings of a given instance. In order to do this, we shall now give another characterization of optimal matchings:

Definition 1 Let π be a matching. Let us call $P = (p_1, \ldots, p_{2N})$ the ordered list of the points in $W \cup B$. For $1 \leq i \leq 2N$, we call $P_i(\pi)$ the *stack of π at i*, that is, the set of points in $\{p_1, \ldots, p_i\}$ that are paired by π to points in $\{p_{i+1}, \ldots, p_{2N}\}$ (see Fig. 2).

Proposition 2 *$\pi \in \mathcal{Z}_J$ iff, for all $1 \leq i \leq 2N$, the stack of π at i is either empty or monochromatic, i.e. if $P_i(\pi) \cap W = \varnothing$ or $P_i(\pi) \cap B = \varnothing$.*
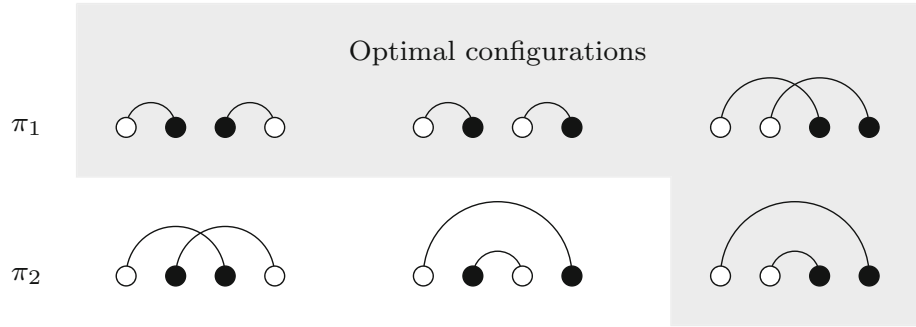
Fig. 1 Optimal matchings at $p = 1$ for $2N = 4$ points. π_1 is the matching in which the first white point is matched with the first black point; π_2 is the only other possible matching. All optimal configurations satisfy $k_\pi(x) = k_{\mathrm{LB}}(x)$
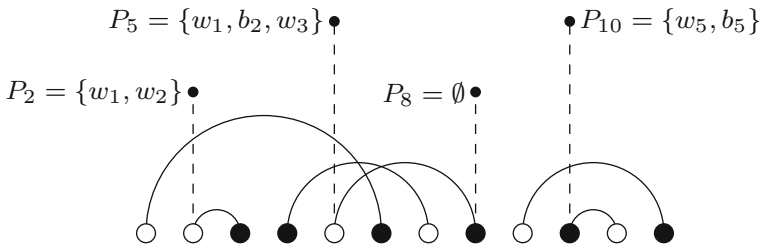
Fig. 2 The stack of a matching at position i is given by all the points on the left side of point i (including i itself) that are matched to a point on the right side of point i. In the picture, w_i is the i-th white point from the right, and analogously b_i is the i-th black point from the right. At the locations specified by the dashed lines, we show the stack of the represented matching

Proof Suppose that for, some $1 \leq i \leq 2N$, the stack of π at i is non-empty and non-monochromatic. Then there are $p_w \in P_i(\pi) \cap W$ and $p_b \in P_i(\pi) \cap B$ which are matched to $q_b \in P_i^c(\pi) \cap B$ and $q_w \in P_i^c \cap W$, respectively. In the matching π' in which we swap these two pairs, we have $k_{\pi'}(z) = k_\pi(z) - 2$ for all $\max(p_w, p_b) < z < \min(q_w, q_b)$, and $k_{\pi'}(z) = k_\pi(z)$ elsewhere, thus, by Corollary 1, π cannot be optimal.

Viceversa, let π have only empty or monochromatic stacks. This means that, in a right neighbourhood of p_i, the cardinality of the stack, which by definition coincides with $k_\pi(x)$, is exactly given by k_{LB}. Furthermore, both these functions are constant on the intervals between the points (where they jump by ± 1), so the two functions coincide everywhere on the domain. This means, by Corollary 1, that π must be optimal. □

2.2 Enumeration of Optimal Matchings

We are now interested in enumerating the optimal matchings at $p = 1$ for a fixed configuration J of size N. First of all, we give an alternative representation of J (already adopted in [5]) that will be useful in the following. A configuration J can be encoded by sorting the $2N$ points in order of increasing coordinate (as in Definition 1 above), and defining

- a vector of spacings $\mathbf{s}(J) \in (\mathbb{R}^+)^{2N}$, given by $\mathbf{s}(J) = (p_1, p_2 - p_1, p_3 - p_2, \ldots, p_{2N} - p_{2N-1})$;

– a vector of signs, $\boldsymbol{\sigma}(J) \in \{-1, +1\}^{2N}$ such that if p_i is white (resp. black), $\sigma_i = +1$ (resp. -1). Note that $\sum_i \sigma_i = 0$.

It is easily seen that the criterium in Proposition 2 is stated only in terms of $\boldsymbol{\sigma}(J)$. This makes clear that the set $\mathcal{Z}_J$ itself is fully determined by $\boldsymbol{\sigma}(J)$. Thus, from this point onward, we understand that $\mathcal{Z}(\boldsymbol{\sigma})$, $Z(\boldsymbol{\sigma})$ and $S(\boldsymbol{\sigma})$ are synonims of the quantities $\mathcal{Z}_J$, Z_J and S_J, for any J with $\boldsymbol{\sigma}(J) = \boldsymbol{\sigma}$.

Binary vectors can be represented as lattice paths, i.e. paths in the plane, starting at the origin and composed by up-steps (or rises) $(+1, +1)$ and down-steps (or falls) $(+1, -1)$. So we have a bijection among zero-sum binary vectors $\boldsymbol{\sigma}$ and lattice bridges, in which the i-th step of the path is (i, σ_i). The bijection between color orderings, binary vectors and lattice paths is so elementary that in the following, with a slight abuse of notation, we will just identify the three objects.

We are interested in two classes of binary vectors / lattice paths:

– **Dyck bridges** $\mathcal{B}_N$ of semi-lenght (size) N. These are lattice paths with an equal number of up- and down-steps. They are precisely in bijection with the color orderings of N white and N black points, i.e. with the color ordering of all the possible configurations J. There are $B_N = |\mathcal{B}_N| = \binom{2N}{N}$ Dyck bridges of size N.
– **Dyck paths** $\mathcal{C}_N$ of semi-length (size) N. These are lattice bridges that never reach negative ordinate (we will also call them *excursions*, in analogy with their continuum counterparts). They are in bijection with configurations J in the forementioned "Dyck excursions" ensemble. There are $C_N = |\mathcal{C}_N| = \frac{1}{N+1} B_N$ Dyck paths of size N.

In the following, the generating function of the series B_N and C_N will turn useful. We have

$$B(z) = \sum_{N \geq 0} B_N z^N = (1 - 4z)^{-\frac{1}{2}},$$

$$C(z) = \sum_{N \geq 0} C_N z^N = \frac{1 - \sqrt{1 - 4z}}{2z}. \tag{15}$$

Our notion of height will be associated to the steps of the path. We call $h_i(\boldsymbol{\sigma})$ the *height of the path at step i*, that is, the height of the midpoint of the i-th step of the path, that in terms of the binary vector reads

$$h_i(\boldsymbol{\sigma}) = \sigma_1 + \sigma_2 + \cdots + \sigma_{i-1} + \frac{\sigma_i}{2}. \tag{16}$$

The choice of the midpoint to compute the height is arbitrary, but has the advantage of being a symmetric definition with respect to reflections w.r.t. the x-axis, while taking values in an equispaced range of integers. We can then define $\bar{h}_i$ as the positive integers

$$\bar{h}_i = |h_i| + \frac{1}{2}. \tag{17}$$

Then we have

Lemma 1

$$Z(\boldsymbol{\sigma}) = \prod_{\substack{i=1,\ldots,2N \\ h_i \sigma_i < 0}} \bar{h}_i(\boldsymbol{\sigma}). \tag{18}$$

Proof The proof goes through the characterisation of the stacks of optimal configurations, given in Proposition 2. First of all, notice that the list of positions given by the condition $h_i \sigma_i < 0$ is the one at which the stack at $i - 1$ decreases its size, because one point in the stack is paired to the i-th point. We shall call *closing steps* the elements of this set (and *closing points* the associated points in P), and *opening steps* those in the complementary set. Indeed, the cardinality of the stack at $i - 1$ is exactly $\bar{h}_i$, while the sign of h_i determines the colour of the points in the stack at $i - 1$ (which is also the colour of the stack at i, unless the latter is empty). So, there are exactly $\bar{h}_i$ choices for the pairing at i, while, if i is not in the list above, the choice is unique. As the cardinalities of the stacks are the same for all optimal configurations, the choice at i does not affect the number of possible choices at $j > i$, and we end up with Eq. (18). □

Notice that Eq. (18) is trivially equivalent to

$$Z(\boldsymbol{\sigma}) = \prod_{i=1}^{2N} \sqrt{\bar{h}_i(\boldsymbol{\sigma})} = \prod_{\substack{i=1,\dots,2N \\ \sigma_i = -1}} \bar{h}_i(\boldsymbol{\sigma}) . \tag{19}$$

The proof above has a stronger implication. Write each optimal matching $\pi \in Z_j$ in the form $\pi = ((i_1, j_1), \dots, (i_N, j_N))$ where i_a and j_a are in the range $\{1, \dots, 2N\}$, and are the indices of the white and black points (altogether), ordered from left to right. Order the pairs canonically, by setting $i_a < j_a$ for all $a = 1, \dots, N$, and the i_a's in increasing order. Observe that, calling $\boldsymbol{i}(\pi)$ the list of the i_a's in the resulting order, we have that all the optimal matchings have the same list $\boldsymbol{i}$ (and no non-optimal matching has this same list). Then, call $\boldsymbol{j}(\pi)$ the list of the j_a's, in the order induced by the i_a's. Finally, order the set Z_J according to the lexicographic order of the strings $\boldsymbol{j}(\pi)$.

Despite the fact that, as is apparent from Lemma 1, the typical values of Z_J are potentially at least exponential in N, we can construct the m-th of the Z_J solutions (in the order above) by a polynomial-time algorithm (which takes on average time $\sim N \log N$ and space $\sim \sqrt{N}$ if the suitable data structure is used). The algorithm goes as follows. First, rewrite m in the form $m - 1 = a_1 + a_2 \bar{h}_1 + a_3 \bar{h}_1 \bar{h}_2 + \dots + a_N \bar{h}_1 \bar{h}_2 \cdots \bar{h}_{N-1}$, with $0 \le a_j < \bar{h}_j$. Then, say that $i(j) = i$ if the j-th closing point is p_i. Now, produce the N pairs of the m-th optimal matching by pairing the closing points, in order of increasing j, by pairing this point to the a_j-th of the stack in $i(j) - 1$, when this is sorted in increasing order.

3 Statistical Properties of S_N

We are now interested in the statistical properties of the entropy

$$S_N(\boldsymbol{\sigma}) = \log Z(\boldsymbol{\sigma}) = \frac{1}{2} \sum_{i=1}^{2N} \log(\bar{h}_i(\boldsymbol{\sigma})) \tag{20}$$

when $\boldsymbol{\sigma}(J)$ is a random variable induced by some probability measure on the space of configurations J of $2N$ points, and N tends to infinity. The subscript N reminds us that we are at size $|W| = |B| = N$.

In matching problems, the typical choices for the configurational probability measure are factorized over a measure on the spacings and a measure on the color orderings, i.e.

$$\mu(J) = \mu_{\text{spacing}}(\mathbf{s}(J)) \, \mu_{\text{color}}(\boldsymbol{\sigma}(J)) \tag{21}$$

(see [5] for more details and examples). As $S_N(J) = S_N(\sigma(J))$, we can again forget about the spacing degrees of freedom, and study the statistics of $S_N(\sigma)$ induced by some measure $\mu_{\text{color}}(\sigma)$. In particular, we will study the cases in which σ is uniformly drawn from the set of Dyck paths, or uniformly drawn from the set of Dyck bridges.

3.1 Integral Formulas for the Integer Moments of S_N via Wiener Processes

It is well known (see Donsker's theorem [12]) that lattice paths such as Dyck paths and bridges converge, as $N \to \infty$ and after a proper rescaling, to Brownian bridges and Brownian excursions. Brownian bridges are Wiener processes constrained to end at null height, while Brownian excursions are Wiener processes constrained to end at null height and to lie in the upper half-plane. The correct rescaling of the steps of the lattice paths that highlights this convergence is given by $(+1, \pm 1) \to \left(+\frac{1}{N}, \pm\frac{1}{\sqrt{N}} \right)$.

These scalings suggest to consider a rescaled version of the entropy

$$s(\sigma) = \frac{S_N(\sigma) - \frac{1}{2} N \log N}{N} = \frac{1}{2N} \sum_{i=1}^{2N} \log \left(\frac{\bar{h}_i(\sigma)}{\sqrt{N}} \right). \tag{22}$$

In the limit $N \to \infty$, the rescaled entropy will converge to an integral operator over Wiener processes

$$s[\sigma] = \int_0^1 dt \, \log\left(|\sigma(t)|\right) \tag{23}$$

where $\sigma(x)$ is a Brownian bridge/excursion.

The integer moments of $s[\sigma]$ can be readily computed as correlation functions of the Brownian process:

$$\langle (s[\sigma])^k \rangle_{\text{B/E}} = \int \mathcal{D}_{\text{B/E}}[\sigma] \int_0^1 dt_1 \dots dt_k \prod_{a=1}^k \log\left(|\sigma(t_a)|\right)$$

$$= k! \int_{\Delta_k} dt_1 \dots dt_k \int_{\mathbb{R}} dx_1 \dots dx_k \prod_{a=1}^k \log\left(|x_a|\right) \int \mathcal{D}_{\text{B/E}}[\sigma] \prod_{a=1}^k \delta(\sigma(t_a) - x_a), \tag{24}$$

where $\Delta_k \subset \mathbb{R}^k$ is the canonical symplex $\{0 = t_0 < t_1 < t_2 < \dots < t_k < t_{k+1} = 1\}$, and $\mathcal{D}_{\text{B/E}}[\sigma]$ is the standard measure on the Brownian process of choice among bridges and excursions. The last integral is the probability that the Brownian process we are interested in starts and ends at the origin and visits the points $(t_1, x_1), \dots, (t_k, x_k)$, while subject to its constraints.

Let us consider Brownian bridges first. In this case, the probability that a Wiener process travels from (t_i, x_i) to (t_f, x_f) is given by $\mathcal{N}(x_f - x_i | 2(t_f - t_i))$ where $\mathcal{N}(x|\sigma^2)$ is the p.d.f. of a centered Gaussian distribution with variance σ^2. The factor 2 comes from Donsker's theorem, and is due to the fact that the variance of the distribution of the steps in the lattice paths is exactly 2. Thus, for Brownian bridges

$$\int \mathcal{D}_{\text{B}}[\sigma] \prod_{a=1}^k \delta(\sigma(t_a) - x_a) = \frac{\sqrt{4\pi}}{\prod_{a=0}^k \sqrt{4\pi(t_{a+1} - t_a)}} \exp\left[-\sum_{a=0}^k \frac{(x_{a+1} - x_a)^2}{4(t_{a+1} - t_a)} \right] \tag{25}$$

where $x_0 = 0$, $x_{k+1} = 0$ and the factor $\sqrt{4\pi}$ is a normalization, so that

$$\langle (s[\sigma])^k \rangle_{\mathrm{B}} = k! \int_{\Delta_k} dt_1 \ldots dt_k \int_{\mathbb{R}} dx_1 \ldots dx_k \frac{\sqrt{4\pi} \prod_{a=1}^k \log(|x_a|)}{\prod_{a=0}^k \sqrt{4\pi(t_{a+1} - t_a)}} \exp\left[-\sum_{a=0}^k \frac{(x_{a+1} - x_a)^2}{4(t_{a+1} - t_a)} \right].$$

(26)

Brownian excursions can be treated analogously using the reflection principle. In this case, the conditional probability that a Wiener process travels from (t_i, x_i) to (t_f, x_f) without ever reaching negative heights, given that it already reached (t_i, x_i), is given by $\mathcal{N}(x_f - x_i | 2(t_f - t_i)) - \mathcal{N}(x_f + x_i | 2(t_f - t_i))$ for $x_{i,f} > 0$, while for $x_i = 0$ and $x_f = x$ (or viceversa) it equals $\frac{|x|}{2(t_f - t_i)} \mathcal{N}(x | 2(t_f - t_i))$. Moreover, now all x_i's are constrained to be positive. Thus, for Brownian excursions

$$\langle (s[\sigma])^k \rangle_{\mathrm{E}} = k! \int_{\Delta_k} dt_1 \ldots dt_k \int_{[0,+\infty)} dx_1 \ldots dx_k \frac{\sqrt{4\pi} x_1 x_k \prod_{a=1}^k \log(x_a)}{t_1(1 - t_k) \prod_{a=0}^k \sqrt{4\pi(t_{a+1} - t_a)}}$$

$$\times \exp\left[-\frac{x_1^2}{4t_1} - \frac{x_k^2}{4(1 - t_k)} \right] \prod_{a=1}^{k-1} \left\{ \exp\left[-\frac{(x_{a+1} - x_a)^2}{4(t_{a+1} - t_a)} \right] - \exp\left[-\frac{(x_{a+1} + x_a)^2}{4(t_{a+1} - t_a)} \right] \right\}.$$

(27)

In both cases, the Gaussian integrations on the heights x_i can be explicitly performed. First of all, we replace

$$\log|x_a| = \frac{1}{2} \partial_{\kappa_a} \left[x_a^{2\kappa_a} \right]_{\kappa_a=0}.$$

(28)

Then, we treat the contact terms. In the case of bridges, the contact terms can be rewritten (under integration) as

$$\exp\left[\frac{x_{a+1} x_a}{2(t_{a+1} - t_a)} \right] \rightarrow \cosh\left(\frac{x_{a+1} x_a}{2(t_{a+1} - t_a)} \right),$$

(29)

where the hyperbolic sine term is discarded due to the parity of the rest of the integrand in the variables x_a. In the case of excursions, the contact term instead reads

$$\exp\left[\frac{x_{a+1} x_a}{2(t_{a+1} - t_a)} \right] - \exp\left[-\frac{x_{a+1} x_a}{2(t_{a+1} - t_a)} \right] = 2 \sinh\left(\frac{x_{a+1} x_a}{2(t_{a+1} - t_a)} \right).$$

(30)

In both cases, we can expand the hyperbolic function in power-series, so that the integrations in the x_a variables are now factorized and of the kind

$$\int_{\mathbb{R}} dx \, x^{2k} \exp\left[-\frac{x^2}{\lambda} \right] = \Gamma\left(k + \frac{1}{2} \right) \lambda^{k + \frac{1}{2}}$$

(31)

(in the case of excursions, a factor $1/2$ must be added to take into account the halved integration domain).

Using these manipulations, the first two moments for both bridges and excursions can be analytically computed. We detail the computations in Appendix A. The results are:

$$\langle s[\sigma] \rangle_{\mathrm{B}} = -\frac{\gamma_E + 2}{2}$$

$$\langle s[\sigma] \rangle_{\mathrm{E}} = -\frac{\gamma_E}{2},$$

(32)

where γ_E is the Euler–Mascheroni constant, and

$$\langle (s[\boldsymbol{\sigma}])^2 \rangle_{\mathrm{B}} = \frac{4}{3} + \gamma_E + \frac{\gamma_E^2}{4} - \frac{\pi^2}{72}$$

$$\langle (s[\boldsymbol{\sigma}])^2 \rangle_{\mathrm{E}} = \frac{\gamma_E^2}{4} + \frac{5\pi^2}{24} - 2 \,. \tag{33}$$

The approach presented in this Section is simple in spirit, and allows to connect our problem to the vast literature on Wiener processes. Moreover, it is suitable for performing Monte Carlo numerical integration to retrieve the moments of $s(\boldsymbol{\sigma})$.

In this Section we worked directly in the continuum limit. In the next Section, we provide a combinatorial approach that allows to recover the values of the first two moments in a discrete setting, and to compute finite-size corrections in the limit $N \to \infty$.

3.2 Combinatorial Properties of the Integer Moments of S_N at Finite N

In this section, we introduce a combinatorial method to compute the moments of $S_N(\boldsymbol{\sigma})$ in the limit $N \to \infty$. This new approach allows to retain informations on the finite-size corrections.

The underlying idea is to reproduce Eq. (24) in the discrete setting for the variable $S_N(\boldsymbol{\sigma})$, and to study its large-N behaviour using methods from analytic combinatorics.

We start again from

$$S_N(\boldsymbol{\sigma}) = \sum_{\substack{i=1\dots 2N \\ \sigma_i = -1}} \log(\bar{h}_i(\boldsymbol{\sigma})) \tag{34}$$

In the following, the superscript/subscript T = E, B will stand for Dyck paths (excursions, E) and Dyck bridges (B) respectively, $T_N = C_N, B_N$ and $\mathcal{T}_N = \mathcal{C}_N, \mathcal{B}_N$; we will mantain the notation unified whenever possible.

The k-th integer moment equals

$$\langle (S_N(\boldsymbol{\sigma}))^k \rangle_{\mathrm{T}} := M_{N,k}^{(\mathrm{T})} = \frac{1}{T_N} \sum_{\boldsymbol{\sigma} \in \mathcal{T}_N} [S_N(\boldsymbol{\sigma})]^k = \frac{k!}{T_N} \sum_{\boldsymbol{\sigma} \in \mathcal{T}_N} \sum_{\substack{1 \le t_1, \dots, t_k \le 2N \\ \sigma_{t_1} = \cdots = \sigma_{t_k} = -1}} \prod_{a=1}^{k} \log(\bar{h}_{t_a}(\boldsymbol{\sigma}))$$

$$= k! \sum_{c=1}^{k} \sum_{\substack{1 \le t_1 < t_2 < \cdots < t_c \le 2N \\ v_1, \dots, v_c \ge 1 \\ v_1 + \cdots + v_c = k \\ \bar{h}_1, \dots, \bar{h}_c > 0}} \prod_{a=1}^{c} \left(\frac{(\log \bar{h}_a)^{v_a}}{v_a!} \right) \frac{\mathcal{M}_N^{(\mathrm{T})}(t_1, \dots, t_c; \bar{h}_1, \dots, \bar{h}_c)}{T_N} \tag{35}$$

where $\mathcal{M}_N^{(\mathrm{T})}(t_1, \dots, t_c; \bar{h}_1, \dots, \bar{h}_c)$ is the number of paths of type T that has closing steps at horizontal positions $t_1, \dots t_c$, and at heights $h_1 = \pm(\bar{h}_1 - 1/2), \dots, h_c = \pm(\bar{h}_c - 1/2)$.

The last equation reproduces, as anticipated earlier, Eq. (24) in the discrete setting. Notice that here we must take into account the multiplicities v_a, while in the continuous setting we could just set $c = k$ and $v_a = 1$ for all a, as the contribution from the other terms is washed out in the continuum limit. This suggests that, in this more precise approach, we will verify explicitly that the leading contributions in the large-N limit comes from the $c = k$ term of Eq. (35).

In order to study Eq. (35), we take the following route. As this equation depends on N only implicitly through the summation range, and explicitly through a normalization, we would

like to introduce a generating function

$$M_k^{(T)}(z) = \sum_{N \geq 1} z^N T_N M_{N,k}^{(T)} \tag{36}$$

that will decouple the summation range over the variables $t_{i+1} - t_i$. By singularity analysis [14], the asymptotic expansion for $N \to \infty$ of $M_{N,k}^{(T)}$ will be then retrieved by the singular expansion of $M_k^{(T)}(z)$ around its dominant singularity.

We start by giving an explicit form for $\mathcal{M}_N^{(T)}$ for Dyck paths and Dyck bridges.

Proposition 3 *In the case of Dyck bridges, we have*

$$\mathcal{M}_N^{(B)}(t_1, \ldots, t_c; \bar{h}_1, \ldots, \bar{h}_c)$$

$$= 2B_{t_1-1,\bar{h}_1} \left(B_{t_2-t_1-1,\bar{h}_2-(\bar{h}_1-1)} + B_{t_2-t_1-1,\bar{h}_2+(\bar{h}_1-1)} \right) \cdots \tag{37}$$

$$\cdots \left(B_{t_c-t_{c-1}-1,\bar{h}_c-(\bar{h}_{c-1}-1)} + B_{t_c-t_{c-1}-1,\bar{h}_c+(\bar{h}_{c-1}-1)} \right) B_{2N-t_c,\bar{h}_c-1},$$

where

$$B_{a,b} = \begin{cases} \binom{a}{\frac{a+b}{2}} & \text{if } a, b \in \mathbb{Z}^+ \text{ and } a + b \text{ is even} \\ 0 & \text{otherwise} \end{cases} \tag{38}$$

is the number of unconstrained paths that start at (x, y) and end at $(x + a, y + b)$.

In the case of Dyck paths, we have

$$\mathcal{M}_N^{(E)}(t_1, \ldots, t_c; \bar{h}_1, \ldots, \bar{h}_c)$$

$$= C_{t_1-1,\bar{h}_1,0} \, C_{t_2-t_1-1,\bar{h}_2-(\bar{h}_1-1),\bar{h}_1-1} \cdots \tag{39}$$

$$\cdots C_{t_c-t_{c-1}-1,\bar{h}_c-(\bar{h}_{c-1}-1),\bar{h}_{c-1}-1} C_{2N-t_c,-(\bar{h}_c-1),\bar{h}_c-1}$$

where

$$C_{a,b,d} = \left(B_{a,b} - B_{a,b+2(d+1)} \right) \theta(b + d) \qquad a, b, d \in \mathbb{Z}^+, \tag{40}$$

is the number of paths that start at (x, y), end at $(x + a, y + b)$ and never fall below height $y - d$, and $\theta(x) = 1$ for $x \geq 0$ and zero otherwise,

Notice that, while in general the θ factors are necessary for the definition of $C_{a,b,d}$ in terms of $B_{a,b}$, in our specific case they are all automatically satisfied, as $\bar{h}_a \geq 1$ for all $1 \leq a \leq c$.

Proof Let us start by considering Dyck bridges. The idea is to decompose a path contributing to the count of $\mathcal{M}_N^{(B)}(t_1, \ldots, t_c; \bar{h}_1, \ldots, \bar{h}_c)$ around its closing steps:

- the first closing step starts at coordinate $(t_1 - 1, \pm\bar{h}_1)$. There are $B_{t_1-1,\bar{h}_1} + B_{t_1-1,-\bar{h}_1} = 2B_{t_1-1,\bar{h}_1}$ different portions of path joining the origin to the starting point of the first closing step.
- the a-th closing step happens $t_a - t_{a-1} - 1$ steps after the $(t - 1)$-th one, and, based on the relative sign of the heights of the two closing steps, their difference in height equals $\bar{h}_a - (\bar{h}_{a-1} - 1)$ or $\bar{h}_a + (\bar{h}_{a-1} - 1)$. Thus, there are $B_{t_a-t_{a-1}-1,\bar{h}_a-(\bar{h}_{a-1}-1)} + B_{t_a-t_{a-1}-1,\bar{h}_a+(\bar{h}_{a-1}-1)}$ different portions of path connecting the two closing steps.
- the last closing step happens $2N - t_c$ steps before the end of the path and at height $h_c = \pm\left(\bar{h}_c - \frac{1}{2}\right)$. Thus, there are $B_{2N-t_c,\bar{h}_c-1}$ portions of path concluding the original path.

The product of the contribution of each subpath recovers Eq. (37).

The case of Dyck paths can be treated analogously, with a few crucial differences. In fact, each of the portions of path between the i-th and $(i + 1)$-th closing steps (which, for excursions, are just down-steps) has now the constraint that it must never fall below the horizontal axis, i.e. must never reach a height $\left(\bar{h}_i - \frac{1}{2}\right)$ lower with respect to its starting step. Let us count these paths. A useful trick to this end is the discrete version of the *reflection method*, that we already used in Sect. 3.1. Call a the total number of steps, b the relative height of the final step with respect to the starting step, and c the maximum fall allowed with respect to the starting step. Moreover, call *bad paths* all paths that do not respect the last constraint. A bad path is characterized by reaching relative height $-c - 1$ at some point (say, the first time after s steps). By reflecting the portion of path composed of the first s steps, we obtain a bijection between bad paths and unconstrained paths that start at relative height $-2(c + 1)$, and reach relative height b after a steps. Thus, the total number of good paths $C_{a,b,d}$ is given by subtraction as

$$C_{a,b,d} = B_{a,b} - B_{a,b+2(d+1)} . \tag{41}$$

This line of thought holds for all values of $a, d > 0$ and $b \geq -d$; if $b < -d$ we just have $C_{a,b,d} = 0$. Moreover, by properties of $B_{a,b}$, $C_{a,b,d} = 0$ if $a + b$ is not an even number.

Equation (39) can be easily established by decomposing a generic (marked) path around its closing steps, and by applying our result above.   □

The fact that we want to exploit now is that, while a given binomial factor $B_{a,b}$ (and its constrained variant $C_{a,b,d}$) are not easy to handle exactly, their generating function in a have simple expressions, induced by analogously simple decompositions, that we collect in the following:

Proposition 4

$$\mathcal{B}_b(z) := \sum_a z^{\frac{a}{2}} B_{a,b} = B(z)(\sqrt{z}C(z))^{|b|} , \tag{42}$$

$$\mathcal{C}_{b,d}(z) := \sum_a z^{\frac{a}{2}} C_{a,b,d} = B(z)\left[(\sqrt{z}C(z))^{|b|} - (\sqrt{z}C(z))^{|b+2(d+1)|}\right]\theta(b+d) , \tag{43}$$

$$\mathcal{C}_b(z) := \sum_a z^{\frac{a}{2}} C_{a,b,0} = B(z)\left(1 - zC(z)^2\right)(\sqrt{z}C(z))^b\theta(b) , \tag{44}$$

where, as in (15),

$$B(z) = \sum_{k \geq 0} z^k B_k = \frac{1}{\sqrt{1 - 4z}} ; \qquad C(z) = \sum_{k \geq 0} z^k C_k = \frac{1 - \sqrt{1 - 4z}}{2z} . \tag{45}$$

Proof To obtain Eq. (42), observe that a path going from $(0, 0)$ to (a, b) with non-negative b can be uniquely decomposed as $w = w_0 u_1 w_1 u_2 w_2 \ldots u_h w_h$ where u_i is the right-most up-step of w at height $i - 1/2$, and w_i is a (possibly empty) Dyck path, for all $i = 1, \ldots, h$, while w_0 is a (possibly empty) Dyck bridge. Thus,

$$B_{a,b} = \sum_{\substack{\ell_0,\ldots,\ell_b \geq 0 \\ 2\sum_{i=0}^{b} \ell_i + b = a}} B_{\ell_0} C_{\ell_1} \cdots C_{\ell_b} . \tag{46}$$

For negative b, the same reasoning holds with u_i's replaced by down-steps, hence the absolute value on b in the result. Equation (42) then follows easily.

Equation (43) follows from $C_{a,b,d} = (B_{a,b} - B_{a,b+2(d+1)}) \theta(b+d)$.

Equation (44) can be derived either as a special case of Eq. (43), or as a variation of (42) where w_0 must be a Dyck path. The equivalence of these two decompositions is granted by the fact that $C(z) = (1 - zC(z)^2) B(z)$. □

Let us introduce the symbol $x = x(z)$ for the recurrent quantity

$$x(z) = zC(z)^2 = C(z) - 1 \tag{47}$$

which, if used to parametrise the other relevant quantities, gives

$$z(x) = \frac{x}{(1+x)^2} ; \qquad\qquad B(z(x)) = \frac{1+x}{1-x} . \tag{48}$$

Then, Eq. (36) reads

$$M_k^{(T)}(z) = k! \sum_{c=1}^{k} \sum_{\substack{v_1,\ldots,v_c \geq 1 \\ \bar{h}_1,\ldots,\bar{h}_c > 0 \\ \sum_a v_a = k}} \prod_{a=1}^{c} \left(\frac{(\log \bar{h}_a)^{v_a}}{v_a!} \right) \mathcal{M}^{(T)}(z; \bar{h}_1, \ldots, \bar{h}_c), \tag{49}$$

where

$$\mathcal{M}^{(T)}(z; \bar{h}_1, \ldots, \bar{h}_c) = \sum_{N \geq 0} z^N \sum_{1 \leq t_1 < t_2 < \ldots < t_c \leq 2N} \mathcal{M}_N^{(T)}(t_1, \ldots, t_c; \bar{h}_1, \ldots, \bar{h}_c). \tag{50}$$

Proposition 5 *Using x to denote $x(z)$, we have that for bridges*

$$\mathcal{M}^{(B)}(z; \bar{h}_1, \ldots, \bar{h}_c) = 2z^{\frac{c}{2}} B(z)^{c+1} \sqrt{x}^{\bar{h}_1} \left(\sqrt{x}^{|\bar{h}_2 - \bar{h}_1 + 1|} + \sqrt{x}^{\bar{h}_2 + \bar{h}_1 - 1} \right)$$
$$\cdots \left(\sqrt{x}^{|\bar{h}_c - \bar{h}_{c-1} + 1|} + \sqrt{x}^{\bar{h}_c + \bar{h}_{c-1} - 1} \right) \sqrt{x}^{\bar{h}_c - 1}, \tag{51}$$

and for excursions

$$\mathcal{M}^{(E)}(z; \bar{h}_1, \ldots, \bar{h}_c) = z^{\frac{c}{2}} B(z)^{c+1} (1-x) \sqrt{x}^{\bar{h}_1} \left(\sqrt{x}^{|\bar{h}_2 - \bar{h}_1 + 1|} - \sqrt{x}^{\bar{h}_2 + \bar{h}_1 + 1} \right)$$
$$\cdots \left(\sqrt{x}^{|\bar{h}_c - \bar{h}_{c-1} + 1|} - \sqrt{x}^{\bar{h}_c + \bar{h}_{c-1} + 1} \right) (1-x) \sqrt{x}^{\bar{h}_c - 1}. \tag{52}$$

Proof First of all, we notice that

$$\mathcal{M}_N^{(T)}(t_1, \ldots, t_c; \bar{h}_1, \ldots, \bar{h}_c)$$
$$= f_1(t_1 - 1; \bar{h}_1) f_2(t_2 - t_1 - 1; \bar{h}_2, \bar{h}_1) \ldots f_c(t_c - t_{c-1} - 1; \bar{h}_c, \bar{h}_{c-1}) f_{c+1}(2N - t_c; \bar{h}_c) \tag{53}$$

for some functions f_i that depend on the type of paths T that we are studying. Thus, by performing the change of summation variables $\{t_1, \ldots, t_c, N\} \to \{\alpha_1, \ldots, \alpha_{c+1}\}$ such that

$$\alpha_1 = t_1 - 1,$$
$$\alpha_i = t_i - t_{i-1} - 1, \qquad 2 \leq i \leq c, \tag{54}$$
$$\alpha_{c+1} = 2N - t_c,$$

we have that

$$
\begin{aligned}
\mathcal{M}^{(\mathrm{T})} & (z; \bar{h}_1, \ldots, \bar{h}_c) \\
&= \sum_{N \geq 0} z^N \sum_{1 \leq t_1 < t_2 < \cdots < t_c \leq 2N} \mathcal{M}_N^{(\mathrm{T})}(t_1, \ldots, t_c; \bar{h}_1, \ldots, \bar{h}_c) \\
&= z^{c/2} \sum_{\alpha_1, \ldots, \alpha_{c+1} \geq 0} f_1(\alpha_1; \bar{h}_1) z^{\alpha_1/2} \ldots f_c(\alpha_c; \bar{h}_c, \bar{h}_{c-1}) z^{\alpha_2/2} f_{c+1}(\alpha_{c+1}; \bar{h}_c) z^{\alpha_{c+1}/2},
\end{aligned}
\tag{55}
$$

so that all summations are now untangled. Equations (51) and (52) can now be recovered by using the explicit form of the functions f_i for Dyck paths and Dyck bridges given in Proposition 3, and the analytical form for the generating functions given in Proposition 4. Again, notice that the θ functions are all automatically satisfied as $\bar{h}_i \geq 1$ for all $1 \leq i \leq c$. □

At this point, we have obtained a quite explicit expression for $M_k^{(\mathrm{T})}(z)$. In the following sections, we will study the behaviour near the leading singularities of the quantities above, for the first two moments, i.e. $k = 1, 2$. Higher-order moments require a more involved computational machinery that will be presented elsewhere.

3.2.1 Singularity Analysis for $k = 1$

We start our analysis from the simplest case, $k = 1$, to illustrate how singularity analysis is applied in this context. We expect to recover Eq. (32). From now on, for simplicity, as the $\bar{h}$ indices are mute summation indices, we will call them simply h. We have

$$
M_1^{(\mathrm{B})}(z) = 2\sqrt{z} B(z)^2 \sum_{h_1 \geq 1} (\log h_1) \sqrt{x}^{\,2h_1 - 1} = \frac{2(1+x)}{(1-x)^2} \sum_{h \geq 1} \log h \; x^h = \frac{2(1+x)}{(1-x)^2} \mathrm{Li}_{0,1}(x),
\tag{56}
$$

and

$$
M_1^{(\mathrm{E})}(z) = \sqrt{z} B(z)^2 (1-x)^2 \sum_{h_1 \geq 1} (\log h_1) \sqrt{x}^{\,2h_1 - 1} = (1+x) \sum_{h \geq 1} \log h \; x^h = (1+x) \mathrm{Li}_{0,1}(x),
\tag{57}
$$

where $\mathrm{Li}_{s,r}(x) = \sum_{h \geq 1} h^{-s} (\log(h))^r x^h$ is the generalized polylogarithm function [13].

In both cases, the dominant singularity is at $x(z) = 1$, i.e. at $z = 1/4$. We have

$$
x(z) = 1 - 2\sqrt{1-4z} + 2(1-4z) + \mathcal{O}\left((1-4z)^{\frac{3}{2}}\right) \quad \text{for } z \to \left(\tfrac{1}{4}\right)^-,
\tag{58}
$$

and

$$
\mathrm{Li}_{0,1}(x) = \frac{\mathrm{L}(x) - \gamma_E}{1-x} + \mathcal{O}\left(\mathrm{L}(x)\right) \quad \text{for } x \to 1^-
\tag{59}
$$

where $\mathrm{L}(x) = \log\left((1-x)^{-1}\right)$. Here and in the following, the rewriting of $\mathrm{Li}_{\alpha,r}(x)$ (for $-\alpha, r \in \mathbb{N}$) in the form $P(\mathrm{L}(x))/(1-x)^{1-\alpha}$, with $P(y)$ a polynomial of degree r, can be done either by matching the asymptotics of the coefficients in the two expressions (and

appealing to the Transfer Theorem), or by using the explicit formulas in [14, Thm. VI.7]. In this paper we mostly adopt the first strategy. Passing to the variable z gives

$$\text{Li}_{0,1}(x(z)) = \frac{\text{L}(4z) - 2\gamma_E - 2\log 2}{4\sqrt{1-4z}} + \mathcal{O}(\text{L}(4z)) \quad \text{for } z \to \left(\tfrac{1}{4}\right)^-. \tag{60}$$

Thus, the singular expansion of $M_1^{(T)}(z)$ is given by

$$M_1^{(B)}(z) = \frac{\text{L}(4z) - 2\gamma_E - 2\log 2}{4(1-4z)^{\frac{3}{2}}} + \mathcal{O}\left(\frac{\text{L}(4z)}{1-4z}\right) \quad \text{for } z \to \left(\tfrac{1}{4}\right)^- \tag{61}$$

and

$$M_1^{(E)}(z) = \frac{\text{L}(4z) - 2\gamma_E - 2\log 2}{2\sqrt{1-4z}} + \mathcal{O}(\text{L}(4z)) \quad \text{for } z \to \left(\tfrac{1}{4}\right)^-. \tag{62}$$

The behaviour of $T_N M_{N,1}^{(T)} = [z^N] M_1^{(T)}(z)$ for large N can be now estimated by using the so-called *transfer theorem*, that allows to jump back and forth between singular expansion of generating functions and the asymptotic expansion at large order of their coefficients (see [14], in particular Chapter VI for general informations, the table in Fig. VI.5 for the explicit formulas). The partinent result is also reported here in Appendix B, namely in Eq. (102). In practice, we can expand the approximate generating functions given in Eqs. (61) and (62) to get an asymptotic approximation for $T_N M_{N,1}^{(T)}$. Finally, using the asymptotic expansion for T_N, i.e.

$$B_N = \frac{4^N}{\sqrt{\pi} N^{\frac{1}{2}}} \left(1 + \mathcal{O}(N^{-1})\right), \qquad C_N = \frac{4^N}{\sqrt{\pi} N^{\frac{3}{2}}} \left(1 + \mathcal{O}(N^{-1})\right) \tag{63}$$

for $N \to \infty$, we obtain an asymptotic expansion for the first moment of S_N (which agrees with what we already found in Eq. (32))

$$\begin{aligned} M_{1,N}^{(B)} &= B_N^{-1}[z^N] M_1^{(B)}(z) = \frac{1}{2} N \log N - \frac{\gamma_E + 2}{2} N + \mathcal{O}\left(\sqrt{N}\log(N)\right) \\ M_{1,N}^{(E)} &= C_N^{-1}[z^N] M_1^{(E)}(z) = \frac{1}{2} N \log N - \frac{\gamma_E}{2} N + \mathcal{O}\left(\sqrt{N}\log(N)\right). \end{aligned} \tag{64}$$

Notice that, although we have truncated our perturbative series at the first significant order, in principle the combinatorial method gives us access to finite-size corrections at arbitrary finite order.

3.2.2 Singularity Analysis for $k = 2$

For $k = 2$, we compute Eq. (49) by studying separately terms at different values of c. Let us start from bridges. For $c = 1$, and thus $\nu_1 = 2$, we have

$$M_2^{(B)}(z)|_{c=1} = 4\sqrt{z} B(z)^2 \sum_{h_1 \geq 1} \frac{(\log h_1)^2}{2} \sqrt{x}^{2h_1 - 1} = \frac{2(1+x)}{(1-x)^2} \text{Li}_{0,2}(x) \tag{65}$$

while for $c = 2$, and thus $v_1 = v_2 = 1$, we have

$$M_2^{(B)}(z)|_{c=2} = 4zB(z)^3 \sum_{h_1,h_2 \geq 1} \log h_1 \log h_2 \sqrt{x}^{h_1} \left(\sqrt{x}^{|h_2-h_1+1|} + \sqrt{x}^{h_2+h_1-1} \right) \sqrt{x}^{h_2-1}.$$

(66)

The presence of the absolute value $|h_2 - h_1 + 1|$ forces us to consider separately the case $h_1 > h_2$ and $h_1 \leq h_2$. In the first case we get

$$4zB(z)^3 \sum_{h_1 > h_2 \geq 1} (\log h_1 \log h_2) \sqrt{x}^{h_1} (\sqrt{x}^{h_1-h_2-1} + \sqrt{x}^{h_1+h_2-1}) \sqrt{x}^{h_2-1}$$

$$= 4zB(z)^3 \left(\sum_{h_1 \geq 1} (\log(h_1+1)\log(h_1!)) x^{h_1} + \sum_{h_1 > h_2 \geq 1} (\log h_1 \log h_2) x^{h_1+h_2-1} \right),$$

(67)

while in the second case we obtain

$$4zB(z)^3 \sum_{1 \leq h_1 \leq h_2} (\log h_1 \log h_2) \sqrt{x}^{h_1} (\sqrt{x}^{h_2-h_1+1} + \sqrt{x}^{h_1+h_2-1}) \sqrt{x}^{h_2-1}$$

$$= 4zB(z)^3 \left(\sum_{h_2 \geq 1} (\log h_2 \log h_2!) x^{h_2} + \sum_{1 \leq h_1 \leq h_2} (\log h_1 \log h_2) x^{h_1+h_2-1} \right).$$

(68)

The combination of these two terms gives

$$M_2^{(B)}(z)|_{c=2} = \frac{4x(1+x)}{(1-x)^3} \left(\sum_{h \geq 1} \left(\log(h^2+h) \log h! \right) x^h + \frac{1}{x} (\text{Li}_{0,1}(x))^2 \right).$$

(69)

In the case of excursions, the computations are completely analogous, and give

$$M_2^{(E)}(z)|_{c=1} = 2\sqrt{z}B(z)^2(1-x)^2 \sum_{h_1 \geq 1} \frac{(\log h_1)^2}{2} \sqrt{x}^{2h_1-1} = (1+x)\text{Li}_{0,2}(x) \quad (70)$$

$$M_2^{(E)}(z)|_{c=2} = \frac{2x(1+x)}{1-x} \left(\sum_{h \geq 1} \left(\log(h^2+h)\log h! \right) x^h - (\text{Li}_{0,1})^2 \right).$$

(71)

In order to compute the singular expansion of $\text{Li}_{0,2}(x)$ and of $\sum_{h \geq 1}(\log(h^2+h)\log(h!)x^h)$, one can again use the transfer theorem, obtaining

$$\text{Li}_{0,2}(x) = \frac{\text{L}(x)^2 - 2\gamma_E \text{L}(x) + \gamma_E^2 + \frac{\pi^2}{6}}{1-x} + \mathcal{O}\left(\text{L}(x)^2\right)$$

(72)

and

$$\sum_{h \geq 1} (\log(h^2+h)\log(h!)z^h) =$$

$$\frac{2\text{L}(x)^2 + 2(1-2\gamma_E)\text{L}(x) + \frac{\pi^2}{3} + 2\gamma_E^2 - 2\gamma_E - 2}{(1-x)^2} + \mathcal{O}\left(\frac{\text{L}(x)^2}{1-x}\right).$$

(73)

We see that, for both Dyck bridges and paths, the $c = 1$ term is subleading with respect to the $c = 2$ term, by a factor $1 - x$ which, after use of the Transfer Theorem, implies a factor $\mathcal{O}(N^{-1})$. It is easy to imagine (and in agreement with the discussion in Sect. 3.1) that this pattern will hold also for all subsequent moments, that is, the leading term for the k-th
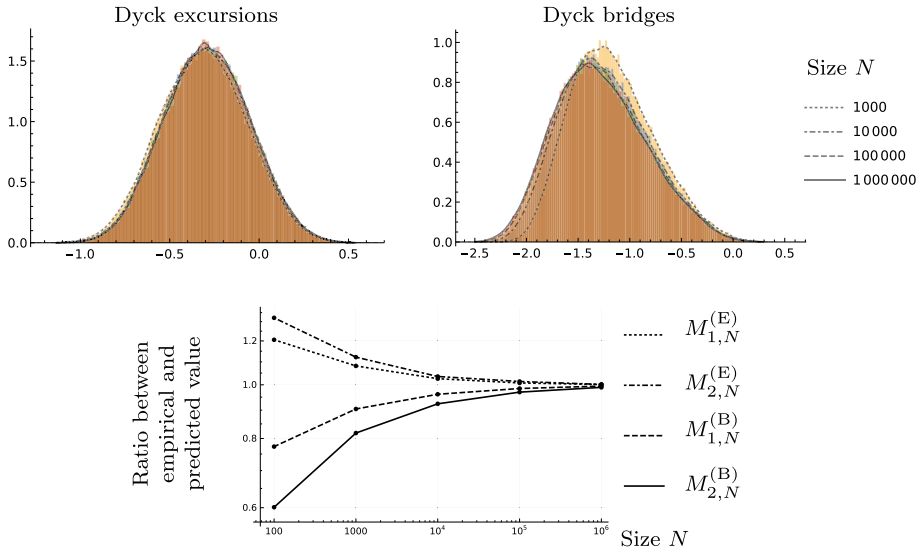
Fig. 3 Top: distribution of s for the Dyck excursions (left) and Dyck bridges (right) ensembles at various sizes N. The shaded fillings are histograms, while the black profiles are kernel density estimates. **Bottom:** ratio between empirical and predicted values of the first two moments as a function of the size N, for both ensembles

moment will be given by the $c = k$ contribution alone, all other terms altogether giving a correction $\mathcal{O}(N^{-1})$.

After substituting $x(z)$ with its expansion around $z = 1/4$, and after having performed a series of tedious but trivial computations, we obtain

$$M_{2,N}^{(B)} = \frac{1}{4}N^2 (\log N)^2 - \frac{\gamma_E + 2}{2} N^2 \log N + \left(\frac{4}{3} + \frac{\gamma_E^2}{4} + \gamma_E - \frac{\pi^2}{72}\right) N^2 + \mathcal{O}\left(N^{\frac{3}{2}} (\log N)^2\right),$$

$$M_{2,N}^{(E)} = \frac{1}{4}N^2 (\log N)^2 - \frac{\gamma_E}{2} N^2 \log N + \left(\frac{\gamma_E^2}{4} + \frac{5\pi^2}{24} - 2\right) N^2 + \mathcal{O}\left(N^{\frac{3}{2}} (\log N)^2\right).$$

$$(74)$$

Finally, we recover the moments of $s(\sigma)$

$$\langle (s(\sigma))^2 \rangle_B = \frac{4}{3} + \frac{\gamma_E^2}{4} + \gamma_E - \frac{\pi^2}{72} + \mathcal{O}\left(\frac{(\log N)^2}{\sqrt{N}}\right),$$

$$\langle (s(\sigma))^2 \rangle_E = \frac{\gamma_E^2}{4} + \frac{5\pi^2}{24} - 2 + \mathcal{O}\left(\frac{(\log N)^2}{\sqrt{N}}\right).$$

$$(75)$$

Details for these computations can be found in Appendix B.

4 Numerical Results

To check our analytical predictions, we performed an exact sampling of our configurations, and collected a statistics on the resulting entropies, from which we estimated the distribution

of the rescaled entropy

$$s(\boldsymbol{\sigma}) = \frac{1}{N} \left(S_N(\boldsymbol{\sigma}) - \frac{1}{2} \log N \right) . \tag{76}$$

For both bridges and excursions, and for values of $N = 10^3, 10^4, 10^5, 10^6$, we sampled uniformly 10^5 random paths.

Figure 3 summarises our results. We clearly see that as N grows larger and larger, the prediction for the first two moments of s matches better and better the empirical value in both ensembles. The distribution of s is clearly non-Gaussian in the case of Dyck bridges. For Dyck excursions, a quick Kolmogorov test rules out the Gaussian hypothesis (in particular, more easily, the 4-th centered moment is only ~ 2.7 times the 2-nd centered moment squared, instead of a factor 3 required for gaussianity). However, at the present statistical precision we cannot rule out the hypothesis that the centered distribution is symmetric, i.e. that all the centered odd moments vanish, although we have no theoretical argument for conjecturing this fact, neither from the probabilistic approach of Sect. 3.1, nor from the combinatorial approach of Sect. 3.2. It would be interesting to understand the reasons of this unexpected (to us) apparent symmetry in the distribution for Dyck excursions.

The code and the raw data used to produce Fig. 3 are available at https://github.com/vittorioerba/EntropyMatching.

Declarations

A The First Two Moments of the Rescaled Entropy in the Integral Representation

In this Appendix we provide the computation for the first two moments of the rescaled entropy in the bridges ensemble, using the integral representation. The case of excursion can be treated analogously.

A.1 Useful Formulas

We start by providing some useful identities.

We start giving an explicit representation for the non-integer Gaussian moments:

$$\int_{-\infty}^{\infty} dy \, y^{2k} e^{-\frac{y^2}{\lambda}} = \Gamma\left(k + \frac{1}{2}\right) \lambda^{k+\frac{1}{2}}. \tag{77}$$

In the following we shall also use the duplication formula for the Gamma function

$$\Gamma(s)\Gamma\left(s + \frac{1}{2}\right) = 2^{1-2s} \sqrt{\pi} \Gamma(2s), \tag{78}$$

and the expansion for the hyperbolic cosine

$$\cosh(2z) = \sum_{s \geq 0} \frac{(2z)^{2s}}{(2s)!} = \sqrt{\pi} \sum_{s \geq 0} \frac{1}{\Gamma\left(s + \frac{1}{2}\right)} \frac{z^{2s}}{s!}. \tag{79}$$

Finally, we recall the definition of the hypergeometric function

$$_2F_1(a, b; c; z) = \frac{\Gamma(c)}{\Gamma(a)\Gamma(b)} \sum_{s \geq 0} \frac{\Gamma(s+a)\Gamma(s+b)}{\Gamma(s+c)} \frac{z^s}{s!}, \tag{80}$$

and the Euler identity

$$_2F_1(a, b; c; z) = (1 - z)^{c-b-a} {}_2F_1(c - a, c - b; c; z) \tag{81}$$

which implies that

$$\sum_{s \geq 0} \frac{\Gamma(s+a)\Gamma(s+b)}{\Gamma(s+c)} \frac{z^s}{s!} = (1 - z)^{c-b-a} \frac{\Gamma(a)\Gamma(b)}{\Gamma(c-a)\Gamma(c-b)} \sum_{s \geq 0} \frac{\Gamma(s+c-a)\Gamma(s+c-b)}{\Gamma(s+c)} \frac{z^s}{s!}. \tag{82}$$

A.2 First Moment

We wish to compute

$$\langle s[\boldsymbol{\sigma}] \rangle_{\mathrm{B}} = \int_0^1 dt \int_{-\infty}^{\infty} dx \frac{\log x^2}{2} \frac{e^{-\frac{x^2}{4t(1-t)}}}{\sqrt{4\pi t(1-t)}} = \frac{1}{2} \int_0^1 dt \, I_1(t). \tag{83}$$

First of all, we substitute $\log x^2 \to x^{2k}$. We will later take the derivative in k and evaluate our expressions for $k = 0$ to obtain back the logarithmic contribution.

Thus, we start from

$$I_1(t, k) = \int_{-\infty}^{\infty} dx \, x^{2k} \frac{e^{-\frac{x^2}{4t(1-t)}}}{\sqrt{4\pi t(1-t)}} = \frac{\lambda^{k+\frac{1}{2}} \Gamma\left(k + \frac{1}{2}\right)}{\sqrt{\pi \lambda}} = \frac{\lambda^k \Gamma\left(k + \frac{1}{2}\right)}{\sqrt{\pi}} \tag{84}$$

where $\lambda = 4t(1 - t)$ so that

$$I_1(t, k) = \frac{[4t(1 - t)]^k}{\sqrt{\pi}} \Gamma\left(k + \frac{1}{2}\right) \tag{85}$$

Then,

$$I_1(t) = \partial_k I_1(t, k)|_{k=0} = \left[\log[4t(1-t)] + \psi_0\left(\frac{1}{2}\right) \right] \tag{86}$$

where $\psi_0(z) = \frac{d}{dz} \log \Gamma(z)$ is the digamma function, and

$$\psi_0\left(\frac{1}{2}\right) = -\gamma_E - \log 4. \tag{87}$$

Finally,

$$\langle s[\sigma] \rangle_B = \frac{1}{2} \int_0^1 dt \{ \log[t(1-t)] - \gamma_E \} = -\frac{2 + \gamma_E}{2}. \tag{88}$$

A.3 Second Moment

We wish to compute

$$
\begin{aligned}
\left\langle (s[\sigma])^2 \right\rangle_B &= 2! \int_{\Delta_2} dt_1 dt_2 \frac{1}{4\pi \sqrt{t_1(t_2 - t_1)(1 - t_2)}} \\
&\quad \times \int_{-\infty}^{\infty} dxdy \frac{\log x^2}{2} \frac{\log y^2}{2} \exp\left[-\frac{x^2}{4t_1} - \frac{(x-y)^2}{4(t_2 - t_1)} - \frac{x^2}{4(1 - t_2)} \right] \\
&= \frac{1}{2} \int_{\Delta_2} dt_1 dt_2 \, I_2(t_1, t_2).
\end{aligned}
\tag{89}
$$

Again, we substitute $\log x^2 \to x^{2k_1}$ and $\log y^2 \to y^{2k_2}$, and we will recover the correct logarithmic factors by taking derivatives in k_1 and k_2 later. We start by dealing with the contact term:

$$
\begin{aligned}
\exp\left[-\frac{(x-y)^2}{4(t_2 - t_1)} \right] &= \exp\left[-\frac{x^2 + y^2}{4(t_2 - t_1)} \right] \cosh\left(\frac{xy}{2(t_2 - t_1)} \right) \\
&= \exp\left[-\frac{x^2 + y^2}{4(t_2 - t_1)} \right] \sum_{s \geq 0} \frac{\sqrt{\pi}}{\Gamma\left(s + \frac{1}{2}\right) s!} \left(\frac{xy}{4(t_2 - t_1)} \right)^{2s}
\end{aligned}
\tag{90}
$$

where the hyperbolic sine was discarded due to parity in x and y. With this substitution, the integrals in x and y are decoupled, and can be evaluated as

$$
\begin{aligned}
I_2&(t_1, t_2, k_1, k_2) = \\
&= \frac{1}{4\pi \sqrt{t_1(t_2 - t_1)(1 - t_2)}} \int_{-\infty}^{\infty} dxdy \, x^{2k_1} y^{2k_2} \exp\left[-\frac{x^2}{4t_1} - \frac{(x-y)^2}{4(t_2 - t_1)} - \frac{x^2}{4(1 - t_2)} \right] \\
&= \frac{2}{\pi \sqrt{\Delta_1 \Delta_2 \Delta_3}} \sum_{s \geq 0} \frac{\sqrt{\pi}}{\Gamma\left(s + \frac{1}{2}\right) s! \Delta_2^{2s}} \\
&\quad \int_{-\infty}^{\infty} dx x^{2(s+k_1)} \exp\left[-\frac{x^2}{\lambda_1} \right] \int_{-\infty}^{\infty} dy y^{2(s+k_2)} \exp\left[-\frac{y^2}{\lambda_2} \right] \\
&= \frac{2\lambda_1^{k_1 + \frac{1}{2}} \lambda_2^{k_2 + \frac{1}{2}}}{\sqrt{\pi} \Delta_1 \Delta_2 \Delta_3} \sum_{s \geq 0} \frac{\Gamma\left(k_1 + s + \frac{1}{2}\right) \Gamma\left(k_2 + s + \frac{1}{2}\right)}{\Gamma\left(s + \frac{1}{2}\right) s!} \left(\frac{\lambda_1 \lambda_2}{\Delta_2^2} \right)^s
\end{aligned}
\tag{91}
$$

where

$$\lambda_1 = \frac{4t_1(t_2 - t_1)}{t_2}, \qquad \lambda_2 = \frac{4(t_2 - t_1)(1 - t_2)}{1 - t_1}, \qquad \Delta_i = 4(t_i - t_{i-1}), \tag{92}$$

(we recall that $t_0 = 0$ and $t_3 = 1$ by convention). We also notice that the following identity holds

$$\frac{2\sqrt{\lambda_1 \lambda_2}}{\sqrt{\Delta_1 \Delta_2 \Delta_3}} \left(1 - \frac{\lambda_1 \lambda_2}{\Delta_2^2}\right)^{-\frac{1}{2}} = 1 \tag{93}$$

as it can be explicitly verified by substituting the definitions.

By using the previous identity and Eq. (82), we can rewrite $I_2(t_1, t_2, k_1, k_2)$ as

$$I_2(t_1, t_2, k_1, k_2) =$$

$$= \frac{\lambda_1^{k_1} \lambda_2^{k_2}}{\sqrt{\pi}} \left(1 - \frac{\lambda_1 \lambda_2}{\Delta_2^2}\right)^{-k_1 - k_2} \frac{\Gamma\left(k_1 + \frac{1}{2}\right) \Gamma\left(k_2 + \frac{1}{2}\right)}{\Gamma(-k_1) \Gamma(-k_2)} \sum_{s \geq 0} \frac{\Gamma(s - k_1) \Gamma(s - k_2)}{\Gamma\left(s + \frac{1}{2}\right) s!} \left(\frac{\lambda_1 \lambda_2}{\Delta_2^2}\right)^s$$

$$= k_1 k_2 \left[\sqrt{\pi} \sum_{s \geq 1} \frac{\Gamma(s) \Gamma(s)}{\Gamma\left(s + \frac{1}{2}\right) s!} \left(\frac{\lambda_1 \lambda_2}{\Delta_2^2}\right)^s + \mathcal{O}(k_1) \mathcal{O}(k_2)\right]$$

$$+ \frac{\lambda_1^{k_1} \lambda_2^{k_2}}{\pi} \left(1 - \frac{\lambda_1 \lambda_2}{\Delta_2^2}\right)^{-k_1 - k_2} \Gamma\left(k_1 + \frac{1}{2}\right) \Gamma\left(k_2 + \frac{1}{2}\right), \tag{94}$$

where in the last line we used the fact that $\Gamma(-k) = k^{-1} + \dots$ as k goes to zero to highlight the linear dependence in k_1 and k_2 in the first term. Notice that

$$\sum_{s \geq 1} \frac{\Gamma(s) \Gamma(s)}{\Gamma\left(s + \frac{1}{2}\right) s!} \left(\frac{\lambda_1 \lambda_2}{\Delta_2^2}\right)^s = \frac{2}{\sqrt{\pi}} \arcsin^2\left(\sqrt{z}\right). \tag{95}$$

We can now take the derivative with respect to k_1 and k_2 and evaluate the expression in $k_1 = k_2 = 0$ to obtain $I_2(t_1, t_2)$

$$I_2(t_1, t_2) = \partial_{k_1, k_2}^2 I_2(t_1, t_2, k_1, k_2)|_{k_1 = k_2 = 0}$$

$$= 2 \arcsin^2\left(\sqrt{\frac{\lambda_1 \lambda_2}{\Delta_2^2}}\right)$$

$$+ \left[\psi_0\left(\frac{1}{2}\right) + \log \lambda_1 - \log\left(1 - \frac{\lambda_1 \lambda_2}{\Delta_2^2}\right)\right]$$

$$\left[\psi_0\left(\frac{1}{2}\right) + \log \lambda_2 - \log\left(1 - \frac{\lambda_1 \lambda_2}{\Delta_2^2}\right)\right] \tag{96}$$

$$= 2 \arcsin^2\left(\sqrt{\frac{t_1(1 - t_2)}{(1 - t_1)t_2}}\right)$$

$$+ \left[\psi_0\left(\frac{1}{2}\right) + \log\left(4t_1(1 - t_1)\right)\right] \left[\psi_0\left(\frac{1}{2}\right) + \log\left(4t_2(1 - t_2)\right)\right].$$

The second line is easily treated by noting that it is symmetric under $t_1 \to 1 - t_1$, giving

$$
\begin{aligned}
\frac{1}{2} \int_0^1 dt_1 &\int_{t_1}^1 dt_2 \left[\psi_0\left(\frac{1}{2}\right) + \log\left(4t_1(1 - t_1)\right) \right] \left[\psi_0\left(\frac{1}{2}\right) + \log\left(4t_2(1 - t_2)\right) \right] \\
&= \frac{1}{4} \int_0^1 dt_1 \int_0^1 dt_2 \left[\psi_0\left(\frac{1}{2}\right) + \log\left(4t_1(1 - t_1)\right) \right] \left[\psi_0\left(\frac{1}{2}\right) + \log\left(4t_2(1 - t_2)\right) \right] \\
&= \left[\frac{1}{2} \int_0^1 dt \left(\log\left(t(1 - t)\right) - \gamma_E\right) \right]^2 \\
&= \left(-\frac{\gamma_E + 2}{2} \right)^2 = (\langle s[\sigma] \rangle_{\mathrm{B}})^2 \; .
\end{aligned}
\tag{97}
$$

Thus, the variance of $s[\sigma]$ is given by

$$
\begin{aligned}
\left\langle (s[\sigma])^2 \right\rangle_{\mathrm{B}} - (\langle s[\sigma] \rangle_{\mathrm{B}})^2 &= \frac{1}{2} \int_0^1 dt_1 \int_{t_1}^1 dt_2 \, 2 \arcsin^2 \left(\sqrt{\frac{t_1(1 - t_2)}{(1 - t_1)t_2}} \right) \\
&= \int_0^1 dt \int_0^1 dz \, \frac{t(1 - t)}{(t + z - tz)^2} \arcsin^2\left(\sqrt{z}\right) \\
&= \int_0^1 dz \left[-\frac{2(1 - z) + (1 + z)\log z}{(1 - z)^3} \right] \arcsin^2\left(\sqrt{z}\right) \\
&= \frac{1}{3} - \frac{\pi^2}{72} \; .
\end{aligned}
\tag{98}
$$

B Computations Needed in Sect. 3.2.2

In this Appendix, we present in greater detail the computations sketched in Sect. 3.2.2. First of all, let us state the transfer theorem of singularity analysis in a slightly unprecise, but operatively correct form. See [14, Chapter VI] for the details.

Theorem 1 *Let $f(z)$, $g(z)$ and $h(z)$ be generating function with unit radius of convergence, and let f_n, g_n and h_n be their coefficients. Then*

$$
f(z) = g(z) + \mathcal{O}(h(z)) \quad z \to 1^- \tag{99}
$$

if and only if

$$
f_n \sim g_n + \mathcal{O}(h_n) \quad n \to \infty \,. \tag{100}
$$

In practice, one builds a *standard scale* of functions whose coefficient expansion is known, expands a generic generating function over the standard scale (assuming that the scale is rich enough to admit the generating function of interest in its linear span), and uses the transfer theorem to guarantee that an asymptotic equivalence at the level of the generating functions implies (and is implied) by an asymptotic equivalence of their coefficients. A standard scale adopted already in [14, Section VI.8], and which is rich enough for our purposes, is given by functions of the form

$$
f(z) = \sum_{n \geq 0} f_n z^n = \left(\frac{1}{1 - x} \right)^\alpha \mathrm{L}(x)^\beta \tag{101}
$$

where $L(x) = \log\left((1-x)^{-1}\right)$. In this case, if α is not a negative integer, the coefficients f_n satisfy

$$f_n = [z^n]\left(\frac{1}{1-x}\right)^\alpha L(x)^\beta \sim \frac{1}{\Gamma(\alpha)} N^{\alpha-1} \log^\beta(N) \tag{102}$$

for $n \to \infty$. See [14, Fig. VI.5] for the behaviour of subleading corrections.

B.1 Singular Behaviour of $\mathrm{Li}_{0,2}(x)$

The coefficient of $\mathrm{Li}_{0,2}(x)$ equals $\log^2(h)$. It is easy to see, using [14, Fig. VI.5], that the singular expansion of $\mathrm{Li}_{0,2}(x)$ onto the standard scale must be made by the terms $(1-x)^{-1}L(x)^2$, $(1-x)^{-1}L(x)$, $(1-x)^{-1}$ and higher-order terms. The coefficient of the expansion can be retrieved by basic linear algebra:

- the coefficient of $(1-x)^{-1}L(x)^2$ must be 1 to reconstruct the $\log(h)^2$ term;
- the coefficient of $(1-x)^{-1}L(x)$ must be $-2\gamma_E$ to cancel the $\log(h)$ term introduced by the first standard scale function;
- the coefficient of $(1-x)^{-1}$ must be $\zeta(2) + \gamma_E^2 = \pi^2/6 + \gamma_E^2$ to cancel the constant terms introduced in the expansion by the previous standard scale functions.

The error term can be obtained by observing that all expansions used above are valid up to order $\mathcal{O}\left(h^{-1}\log h\right)$. Thus

$$\mathrm{Li}_{0,2}(x) = \frac{L(x)^2 - 2\gamma_E\,L(x) + \gamma_E^2 + \frac{\pi^2}{6}}{1-x} + \mathcal{O}\left(L(x)^2\right). \tag{103}$$

B.2 Singular Behaviour of $\sum_{h\geq 1}\left(\log(h^2 + h)\log h!\right)x^h$

The procedure is analogous to the computation for the $\mathrm{Li}_{0,2}(x)$. We just need to use Stirling's approximation for $\log h!$ to obtain the expansion of the coefficients in h. We have, for the coefficient of order k of this function,

$$\begin{aligned}
&\left(\log(h^2 + h)\log h!\right)\\
&= 2h\log^2(h) - 2h\log(h) + \log^2(h)\\
&\quad + (1 + \log(2) + \log(\pi))\log(h) - 1 + \mathcal{O}\left(\frac{\log(h)}{h}\right)
\end{aligned} \tag{104}$$

so that the singular expansion must be made by the terms $(1-x)^{-2}L(x)^2$, $(1-x)^{-2}L(x)$, $(1-x)^{-2}$, and higher-order ones. The coefficients can be found using the same strategy adopted for $\mathrm{Li}_{0,2}(x)$, obtaining

$$\begin{aligned}
&\sum_{h\geq 1}\left(\log(h^2 + h)\log h!\right)x^h\\
&= \frac{2L(x)^2 + 2(1 - 2\gamma_E)L(x) + 2\gamma^2 + \frac{\pi^2}{3} - 2 - 2\gamma_E}{(1-x)^2} + \mathcal{O}\left(\frac{L(x)^2}{1-x}\right).
\end{aligned} \tag{105}$$

B.3 The Change of Variable $x(z)$

First of all, we rewrite the change of variable $x(z) = C(z) - 1$ in terms of the singular variables $X = (1-x)^{-1}$ and $Z = (1-4z)^{-1}$. We have that

$$X(Z) = \frac{1}{2}\left(\sqrt{Z} + 1\right) \tag{106}$$

so that

$$(1-x)^{-\alpha} = X^{\alpha} = \left(\frac{\sqrt{Z}}{2}\right)^{\alpha}\left(1 + \frac{\alpha}{\sqrt{Z}} + \mathcal{O}(Z^{-1})\right),$$

$$L(x)^k = \log^k(X) = \left(\frac{1}{2}\log Z - \log 2 + \mathcal{O}(1/\sqrt{Z})\right)^k \tag{107}$$

$$= \left(\frac{1}{2}\log Z - \log 2\right)^k + \mathcal{O}\left(\frac{\log^{k-1}(Z)}{\sqrt{Z}}\right).$$

These relations are enough to convert singular expansions in X into singular expansions in Z at the leading algebraic order. The formulas of this subsection are useful, in principle, also at higher moments.

B.4 Explicit Expressions for $M_2^{(T)}(z)$

We report here the explicit expression of $M_2^{(T)}(z)$ both as a function of x and as a function of z:

$$M_2^{(E)}(z(x)) = \frac{4L(x)^2 + 8(1-\gamma_E)L(x) - 8 - 8\gamma_E + 4\gamma_E^2 + \frac{4\pi^2}{3}}{(1-x)^3} + \mathcal{O}\left(\frac{L(x)^2}{(1-x)^2}\right)$$

$$M_2^{(E)}(z) = \left(\frac{1}{8}L(4z)^2 - \frac{1-\gamma_E-\log(2)}{2}L(4z)\right)\frac{1}{(1-4z)^{\frac{3}{2}}}$$

$$+ \frac{\pi^2 - 6 + 3\gamma_E(\gamma_E - 2 + \log(4)) + (\log(2) - 2)\log(8)}{6(1-4z)^{\frac{3}{2}}} + \mathcal{O}\left(\frac{L(4z)^2}{1-4z}\right)$$

$$M_2^{(B)}(z(x)) = \frac{24L(x)^2 + 16(1-3\gamma_E)L(x) - 16 - 16\gamma_E + 24\gamma_E^2 + \frac{8\pi^2}{3}}{(1-x)^5} + \mathcal{O}\left(\frac{L(x)^2}{(1-x)^4}\right)$$

$$M_2^{(B)}(z) = \left(\frac{3}{16}L(4z)^2 - \frac{3\gamma_E + 1 - 3\log(2)}{4}L(4z)\right)\frac{1}{(1-4z)^{\frac{5}{2}}}$$

$$+ \frac{9\gamma_E^2 + \pi^2 - 6 + 6\gamma_E(\log(8) - 1) + (\log(8) - 2)\log(8)}{12(1-4z)^{\frac{5}{2}}} + \mathcal{O}\left(\frac{L(4z)^2}{(1-4z)^2}\right). \tag{108}$$

These expressions can be easily recalculated also with smaller error terms, repeating the procedure of the previous paragraphs with more terms in the Taylor expansions, and provide an evaluation of the variance of our quantities of interest, at the desired order in N.

References

1. D'Achille, M.P.: Statistical properties of the Euclidean random assignment problem. Theses, Université Paris-Saclay (2020). https://tel.archives-ouvertes.fr/tel-03098672
2. Ambrosio, L., Gigli, N.: A User's Guide to Optimal Transport. In: Ambrosio, L., Bressan, A., Helbing, D., Klar, A., Zuazua, E. (eds.) Modelling and Optimisation of Flows on Networks: Cetraro, Italy 2009, Editors: Benedetto Piccoli, Michel Rascle, Lecture Notes in Mathematics, pp. 1–155. Springer, Berlin, Heidelberg (2013). https://doi.org/10.1007/978-3-642-32160-3_1
3. Bobkov, S.G., Ledoux, M.: Transport Inequalities on Euclidean spaces with non-Euclidean metrics. J. Fourier Anal. Appl. **26**(4), 60 (2020). https://doi.org/10.1007/s00041-020-09766-2
4. Boniolo, E., Caracciolo, S., Sportiello, A.: Correlation function for the Grid-Poisson Euclidean matching on a line and on a circle. J. Stat. Mech. **2014**(11), P11023 (2014). https://doi.org/10.1088/1742-5468/2014/11/P11023
5. Caracciolo, S., D'Achille, M.P., Erba, V., Sportiello, A.: The Dyck bound in the concave 1-dimensional random assignment model. J. Phys. A **53**(6), 064001 (2020). https://doi.org/10.1088/1751-8121/ab4a34
6. Caracciolo, S., D'Achille, M.P., Malatesta, E.M., Sicuro, G.: Finite size corrections in the random assignment problem. Phys. Rev. E **95**, 052129 (2017). https://doi.org/10.1103/PhysRevE.95.052129
7. Caracciolo, S., D'Achille, M.P., Sicuro, G.: Random euclidean matching problems in one dimension. Phys. Rev. E **96**, 042102 (2017). https://doi.org/10.1103/PhysRevE.96.042102
8. Caracciolo, S., D'Achille, M.P., Sicuro, G.: Anomalous scaling of the optimal cost in the one-dimensional random assignment problem. J. Statist. Phys. **174**(4), 846–864 (2019). https://doi.org/10.1007/s10955-018-2212-9
9. Caracciolo, S., Di Gioacchino, A., Malatesta, E.M., Molinari, L.G.: Selberg integrals in 1D random Euclidean optimization problems. J. Stat. Mech. **2019**, 063401 (2019). https://doi.org/10.1088/1742-5468/ab11d7
10. Caracciolo, S., Erba, V., Sportiello, A.: The p-Airy distribution. arXiv preprint arXiv:2010.14468 (2020)
11. Caracciolo, S., Sicuro, G.: On the one dimensional euclidean matching problem: exact solutions, correlation functions and universality. Phys. Rev. E **90**, 042112 (2014). https://doi.org/10.1103/PhysRevE.90.042112
12. Donsker, M.D.: Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems, pp. 277–281. The Institute of Mathematical Statistics (1952). https://doi.org/10.1214/aoms/1177729445
13. Fill, J.A., Flajolet, P., Kapur, N.: Singularity analysis, Hadamard products, and tree recurrences. J. Comput. Appl. Math. **174**(2), 271–313 (2005). https://doi.org/10.1016/j.cam.2004.04.014
14. Flajolet, P., Sedgewick, R.: Analytic Combinatorics. Cambridge University Press, Cambridge; New York (2009)
15. Juillet, N.: On a solution to the Monge transport problem on the real line arising from the strictly concave case. SIAM J. Math. Anal. **52**(5), 4783–4805 (2020). https://doi.org/10.1137/19M1277242
16. McCann, R.J.: Exact solutions to the transportation problem on the line. Proc. R. Soc. Lond. A **455**(1984), 1341–1380 (1999). https://doi.org/10.1098/rspa.1999.0364
17. Mézard, M., Parisi, G.: Replicas and optimization. Le Journal de Physique - Lettres **46**(17), 771–778 (1985). https://doi.org/10.1051/jphyslet:019850046017077100
18. Mézard, M., Parisi, G.: Mean-field equations for the matching and the travelling salesman problems. Europhys. Lett. **2**(12), 913–918 (1986). https://doi.org/10.1209/0295-5075/2/12/005
19. Orland, H.: Mean-field theory for optimization problems. Le Journal de Physique - Lettres **46**(17), 773–770 (1985). https://doi.org/10.1051/jphyslet:019850046017076300
20. Peyré, G., Cuturi, M.: Computational optimal transport: With applications to data science. Found. Trends Mach. Learn. **11**(5–6), 355–607 (2019). https://doi.org/10.1561/2200000073