UNIVERSITÀ DEGLI STUDI DI MILANO

FACOLTÀ DI SCIENZE E TECNOLOGIE

Corso di Laurea Magistrale in Fisica

THE RANDOM MINIMUM SPANNING TREE PROBLEM

Relatore:
Prof. Sergio CARACCIOLO

Correlatore:
Dott. Enrico MALATESTA

Elaborato finale di:
Andrea RIVA
Matricola 897594

PACS: 02.10.Ox
02.50.-r

ANNO ACCADEMICO 2018 - 2019

To my family

and to Elisabetta

Contents

Introduction

The *Minimum Spanning Tree* (MST) *problem* is an archetypal member of the class of *combinatorial optimization problems*, a rather broad and interdisciplinary field, its ideas stretching from theoretical computer science to statistical physics throughout graph theory, the latter providing the ground basis for both exact results and construction of efficient algorithms. Combinatorial optimization may be defined as the set of results and techniques used to find extrema of some function, taking values only on a finite or at most countable set. In this setting, the MST problem on a connected graph consists in finding the spanning acyclic subgraph (tree) with the smallest total edge weight. Despite being simple in its formulation, in practice this problem is quite difficult to handle, as it is clear if one considers that on a graph with N vertices, all connected to each other, the number of different spanning trees is N^{N-2}.

The first generally accepted algorithm for the solution of this problem was published in 1926 by O. Borůvka [11] as a method for constructing an efficient electricity network for the region of Moravia, in the Czech Republic. In that case, the edge weights of the underlying graph were given as a first approximation by the distances between different nodes of the network, so that the problem was actually defined in a *Euclidean* domain. Since then, the MST problem has been one of the most thoroughly studied problems of computational geometry. Moreover, it has found diverse applications in an outstanding number of fields, both theoretical and practical, ranging from computer algorithms design to document clusterings, from the analysis of gene expression data to the modeling of turbulent flows, among many others.

In the MST problem solved by Borůvka, as in most cases of everyday interest, the vertices of the graph are supposed assigned, and therefore the problem is fixed in all its details: no disorder or randomness is present. However, we can consider the problem under a different point of view, supposing for example that the edge weights are *random variables* generated according to some probability distribution density. These can be independent and identically distributed, or even correlated if we consider the vertices of the graph as random points scattered in a Euclidean domain, with the edge weights that become proportional to the distance between them. In such versions of the problem, called *random MST problem*, the specific solution of a given instance of an optimization problem is not of great interest, and one is instead concerned with the average properties of the MST, possibly depending on some parameters and on the way randomness is introduced in the first place.

What we have described is clearly the typical playground of statistical physics, which has developed through the years a plethora of techniques to obtain the average properties of systems with a huge number of degrees of freedom, even in the presence of disorder. In fact, since the seminal work by Mézard and Parisi [25] published in 1985 on the *Euclidean matching problem*, the methods of statistical mechanics, and specifically of spin glasses, has proven extremely powerful to treat random combinatorial optimization problems.

In this thesis we overview the main results concerning with the random MST problem, and we present the results of our investigation on this subject, with the material organized as follows.

In chapter 1 we introduce the basic concepts and definitions of graph theory, together with a brief review of optimization theory. We then concentrate on the MST, both describing its general properties and the main algorithms designed for its search in a given graph. After that we focus on the contextualization of random combinatorial optimization problems, making clear their profound connection to the field of statistical physics, and we provide at the end a description of a relevant technique used to treat both, namely the *replica method*.

In chapter 2 we start dealing with the main subject of the thesis in more detail, considering the random MST problem with i.i.d. edge weights. After a review of the existing literature, we derive the spanning trees generating polynomial for a generic graph as a $q \to 0$ limit of the *q-state Potts model*. We then rewrite this quantity in an interesting way using the matrix-tree theorem and Grassmann variables, and we use it as the starting point to set up a replica calculation for the MST problem on the complete graph.

In chapter 3 we turn our attention to the *random Euclidean MST problem*, where the underlying graph is embedded in a Euclidean domain. After an introduction to the subject, considering edges weighted by a positive power of their Euclidean length, we solve for the first time the one dimensional MST problem on a bipartite graph, in which the vertices are divided in two different sets, providing an exact formula for the average cost. Finally, we perform a numerical investigation in one and two dimensions to analyze the asymptotic behavior of the MST average cost in the thermodynamic limit. This, together with our previous solution, shows that the scaling behaviour does not change by passing from the monopartite to the bipartite case, contrary to what happens in other combinatorial optimization problems such as the *Traveling Salesman Problem* (TSP) and the matching problem. This result represents the main contribution of this thesis to the topic.

Chapter 1

Graphs and optimization

The minimum spanning tree problem is an important combinatorial optimization problem defined on a graph, so the first section is devoted to a short introduction to the basic definitions and results of graph theory, referring mostly to the standard textbook of Diestel [1]. We then turn our attention to optimization problems, with a focus on the main subject of the thesis and the algorithms that solves it. Finally, an interesting connection between random optimization problems and the statistical mechanics of disordered systems is established. For this reason, after a brief description of these two last topics, the chapter concludes with the introduction of a relevant method used to treat both, namely the replica technique.

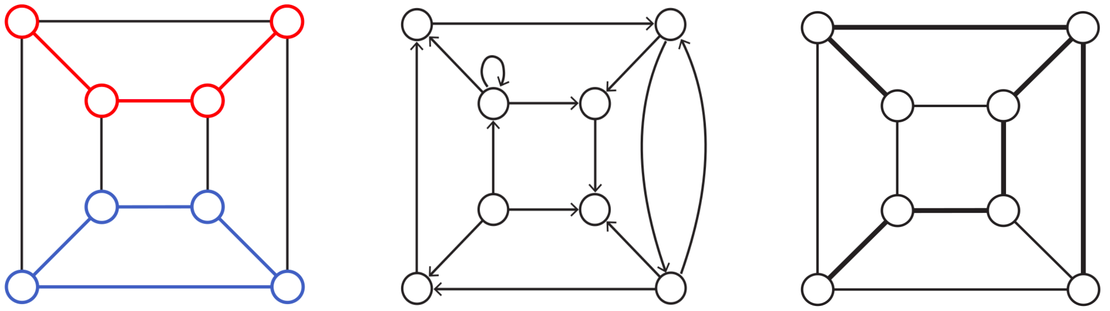
1.1 Graph theory basic definitions

A *graph* $\mathtt{G} = \mathrm{Graph}(\mathcal{V};\mathcal{E})$ is a couple of sets $(\mathcal{V};\mathcal{E})$, such that $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$, whose elements are called *vertices* and *edges* respectively. The cardinality $V = |\mathcal{V}|$ of the vertex set $\mathcal{V}$ is called *order* of $\mathtt{G}$ and we will always suppose that $V \in \mathbb{N}$ is finite, unless otherwise stated.

In the present work an element $e \in \mathcal{E}$ can be uniquely identified by a pair of vertices $u, v \in \mathcal{V}$, i.e. we do not consider graphs containing *multi-edges*. If the ordering does not matter we say that the graph is *undirected*, otherwise every edge is thought to posses an *initial vertex* and a *terminal vertex*, and the graph is called *directed* (or *digraph*). In the last case, an edge in which the initial vertex and the terminal vertex coincides is called a *loop* (Fig. 1.1b).

Usually, one refers to the structure of a graph as a collection of incidence relations, in fact given a vertex v and an edge e we say that v is *incident* with e if $v \in e$, and we write $e \to v$. The number of edges that are incident with a certain vertex v in an undirected graph is called the *degree* (or *coordination number*) of v, and it is denoted by $\deg(v)$. Moreover we say that $u, v \in \mathcal{V}$ are *adjacent* if $(u, v) \in \mathcal{E}$, and we denote by ∂v the set of adjacent vertices to v. Finally, we define the *complete graph* $\mathtt{K}_V$ as the graph in which each of the V vertices is adjacent to all the others (Fig. 1.2a).

Given two graphs $\mathtt{G} = \mathrm{Graph}(\mathcal{V};\mathcal{E})$ and $\mathtt{G}' = \mathrm{Graph}(\mathcal{V}';\mathcal{E}')$, if $\mathcal{V} \subseteq \mathcal{V}'$ and $\mathcal{E} \subseteq \mathcal{E}'$ we say that $\mathtt{G}$ is a *subgraph* of $\mathtt{G}'$ and $\mathtt{G}'$ is a *supergraph* of $\mathtt{G}$, in symbols $\mathtt{G} \subseteq \mathtt{G}'$. Furthermore, if $\mathcal{V} = \mathcal{V}'$ $\mathtt{G}$ is called a *spanning subgraph* of $\mathtt{G}'$ (Fig. 1.1c).



(a) Example of path (red) and cycle (blue)

(b) Directed graph with a loop and a multi-edge

(c) Example of spanning subgraph (bold), in this case a tree

Figure 1.1: Examples of graphs

Let us observe that the set $\mathscr{S}$ of spanning subgraphs is in natural bijection with the *power set* $\mathcal{P}(\mathcal{E})$ of $\mathcal{E}$.

A *path* of length l in a graph $\mathtt{G}$ is a subgraph $\mathtt{P} \subseteq \mathtt{G}$ such that $\mathcal{V}_{\mathtt{P}} = \{v_0, \dots, v_l\}$ is a set of distinct vertices and the edge set is

$$\mathcal{E}_{\mathtt{P}} = \{(v_0, v_1), (v_1, v_2), \dots, (v_{l-1}, v_l)\}. \tag{1.1}$$

Any nontrivial path starting and finishing at the same vertex is called a *cycle* (Fig. 1.1a). A graph $\mathtt{G}$ is said to be *connected* if, for any couple of vertices $u, v \in \mathcal{V}$, there exists a path in $\mathtt{G}$ linking them. Every graph can be expressed as union of maximal connected subgraphs, called *components*. Given a connected graph $\mathtt{G} = \mathrm{Graph}(\mathcal{V}; \mathcal{E})$, the subset $\mathcal{X} \subset \mathcal{V} \cup \mathcal{E}$ is said to be a *separating set* if $\mathtt{G}' = \mathrm{Graph}(\mathcal{V} \setminus \mathcal{X}; \mathcal{E} \setminus \mathcal{X})$ is not connected. Specifically, if $\mathcal{X} \subset \mathcal{E}$ is an edges' subset only we call it a *cut*, and if $\mathcal{X}$ contains only a single vertex this is called a *cutvertex*. Similarly if $\mathcal{X}$ contains only one edge, we say that the selected edge is a *bridge*.

A *Hamiltonian path* in an undirected or directed graph is a path that visits each vertex exactly once. Specifically, if such a path is also a cycle it is called *Hamiltonian cycle*. Analogous definitions hold when we consider paths and cycles traversing all edges exactly once instead of vertices, in which case we refer to *Eulerian paths* and *Eulerian cycles* respectively.

Given two graphs $\mathtt{G} = \mathrm{Graph}(\mathcal{V}; \mathcal{E})$ and $\mathtt{G}' = \mathrm{Graph}(\mathcal{V}; \mathcal{E}')$ with the same vertex set $\mathcal{V}$ we define

$$\mathtt{G} \Delta \mathtt{G}' := \mathrm{Graph}(\mathcal{V}_{\mathtt{G}\Delta\mathtt{G}'}; \mathcal{E} \Delta \mathcal{E}'), \tag{1.2}$$

where

$$\mathcal{E} \Delta \mathcal{E}' := (\mathcal{E} \cup \mathcal{E}') \setminus (\mathcal{E} \cap \mathcal{E}') \tag{1.3}$$

is the *symmetric difference* between the two edge sets and $\mathcal{V}_{\mathtt{G}\Delta\mathtt{G}'}$ is the set of the vertices that are endpoints for the edges in $\mathcal{E} \Delta \mathcal{E}'$.

Let us consider now the set $\mathscr{S}_{\mathtt{G}}$ of the spanning subgraphs of a given graph $\mathtt{G} = \mathrm{Graph}(\mathcal{V}; \mathcal{E})$, which contains the set $\mathscr{S}_{\mathtt{G}}^{\mathrm{E}}$ of Eulerian subgraphs. This space has the peculiar property of being closed under the symmetric difference operation

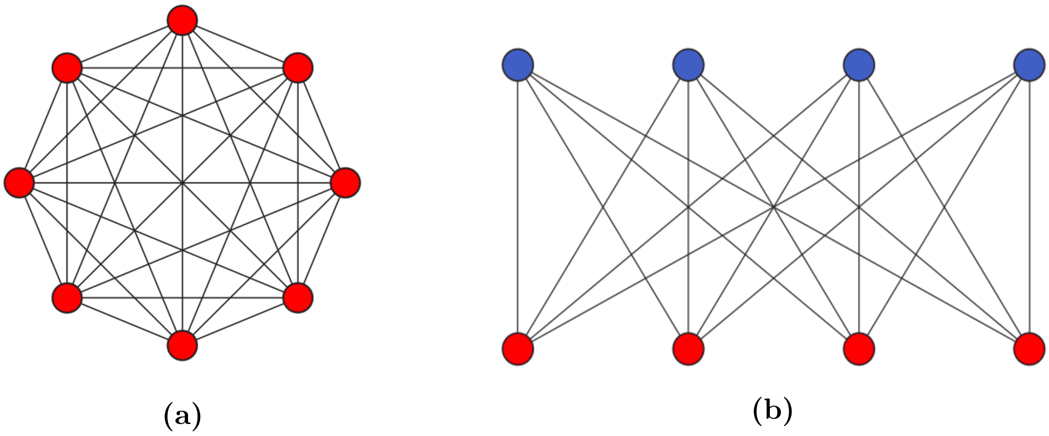


Figure 1.2: A complete graph $\mathtt{K}_V$ with $V = 8$ (a) and a complete bipartite graph $\mathtt{K}_{V,V}$ with $V = 4$ (b).

Δ, that is to say $\mathtt{G}_1 \in \mathscr{S}^{\mathrm{E}}_{\mathtt{G}}$, $\mathtt{G}_2 \in \mathscr{S}^{\mathrm{E}}_{\mathtt{G}} \implies \mathtt{G}_1 \Delta \mathtt{G}_2 \in \mathscr{S}^{\mathrm{E}}_{\mathtt{G}}$. The dimension of $\mathscr{S}^{\mathrm{E}}_{\mathtt{G}}$ with respect to the operation Δ is called *cyclomatic number* L of the graph $\mathtt{G}$, and it corresponds to the number of cycles in $\mathtt{G}$ that cannot be obtained by other subgraphs through symmetric difference. These cycles are called *independent cycles* and play the role of a basis in the space of Eulerian subgraphs.

Here it is worth mentioning the fundamental classic result of Euler (1752), which relates the number of vertices, edges and independent cycles in a general graph, which always sum up to a fixed number depending only on the space in which the graph is embedded, and not on the graph itself [2].

Theorem 1.1.1 (Euler's formula). *Given a graph $\mathtt{G}$ with κ disconnected components, V vertices and E edges, it holds*

$$V + L = E + \kappa. \tag{1.4}$$

This result can be applied easily to the case of *planar graphs*, i.e. graphs that can be drawn on the surface of a sphere in such a way that no edge crossing occurs. In fact, in this case the cyclomatic number is recovered in terms of the number of faces F of the graph as $L = F - 1$.

If the vertex set of a graph $\mathtt{G} = \mathrm{Graph}(\mathcal{V}; \mathcal{E})$ can be partitioned in k subsets, called *classes*,

$$\mathcal{V} = \bigcup_{i=1}^{k} \mathcal{V}_i, \qquad \mathcal{V}_i \cap \mathcal{V}_j = \emptyset \text{ for } i \neq j,$$

in such a way that every edge in $\mathcal{E}$ connects vertices in different classes, we say that $\mathtt{G}$ is *k-partite* (or *multipartite* in general) and we denote it as $\mathtt{G} = \mathrm{Graph}(\mathcal{V}_1, \dots, \mathcal{V}_k; \mathcal{E})$. Furthermore, such a graph is called *complete* and indicated by $\mathtt{K}_{V_1,\dots,V_k}$ if taken any pair of vertices belonging to two different classes there exists an edge connecting them. Note that a multipartite graph with $k = 2$ has no odd cycles and it is referred to as a *bipartite* graph (Fig. 1.2b).

The incidence relations of a graph $\mathtt{G} = \mathrm{Graph}(\mathcal{V};\mathcal{E})$ are usually summarized in the $V \times V$ *adjacency matrix* $\mathbf{A} := (a_{ij})_{ij}$, defined as

$$a_{ij} = \begin{cases} 1 & \text{if } (v_i, v_j) \in \mathcal{E} \\ 0 & \text{otherwise} \end{cases} \tag{1.5}$$

If the graph is undirected the adjacency matrix is symmetric, therefore $\mathbf{A}$ has a real spectrum, called the *spectrum* of $\mathtt{G}$. Another fundamental matrix that can be associated to a graph is the so called *Laplacian matrix* $\mathbf{L} := (L_{ij})_{ij}$, simply given by

$$L_{ij} = \deg(v_i)\, \delta_{ij} - a_{ij}. \tag{1.6}$$

where δ is the Kronecker delta. Finally, a graph $\mathtt{G}$ is said to be *weighted* if there is a function $w : \mathcal{E} \to \mathbb{R}$ that assigns to every edge $e \in \mathcal{E}$ a *weight* $w(e)$. For a given weighted graph we can easily introduce the *weighted adjacency matrix* as $\mathbf{W} := (w(e_{ij})\, a_{ij})_{ij}$.

1.1.1 Trees and forests

In this subsection we introduce a particular class of graphs which will be used extensively throughout the thesis. A connected graph without cycles is called a *tree*, and a disjoint union of trees, i.e. a general acyclic graph, is called a *forest.* The edges of a forest are usually referred to as *branches*, whereas the vertices with degree one are called *leaves* (Fig. 1.3).

Theorem 1.1.2. *The following assertions are equivalent for a graph* $\mathtt{T}$*:*

(i) $\mathtt{T}$ *is a tree;*

(ii) *any two vertices of* $\mathtt{T}$ *are linked by a unique path in* $\mathtt{T}$*;*

(iii) $\mathtt{T}$ *is minimally connected, i.e. the removal of any one edge disconnects the graph;*

(iv) $\mathtt{T}$ *is maximally acyclic, i.e. adding an edge between any two non-adjacent vertices forms a cycle.*

It follows straightforwardly from the theorem above that every connected graph contains a *spanning tree*, in fact by the equivalence of (i) and (iii) any minimally connected spanning subgraph is a tree.

Corollary 1.1.3. *A connected graph with V vertices is a tree iff it has $V - 1$ edges.*

Proof. The vertices of a tree can always be enumerated, say as $v_1, \dots, v_V$, so that every v_i with $i \geq 2$ has a unique neighbour in $\{v_1, \dots, v_{i-1}\}$. Induction on i shows that the subgraph spanned by the first i vertices has $i-1$ edges, and for $i = V$ this proves the forward implication. Conversely, let $\mathtt{G}$ be any connected graph with V vertices and $V - 1$ edges, and let $\mathtt{G}'$ be a spanning tree in $\mathtt{G}$. Since $\mathtt{G}'$ has $V - 1$ edges by the first implication, it follows that $\mathtt{G} = \mathtt{G}'$. □

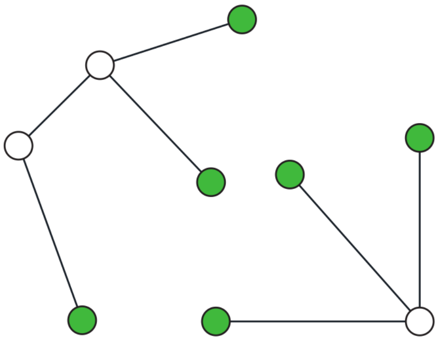


Figure 1.3: A forest composed by two trees, with leaves are colored in green.

Theorem 1.1.2 says that adding just one edge to a spanning tree will create a cycle, which is called *fundamental cycle*. There is a one-to-one correspondence between fundamental cycles and the $E-V+1$ edges not in the spanning tree: in particular for any given spanning tree, the set of all $E-V+1$ fundamental cycles forms a *cycle basis*.

An important graph invariant is the number $\tau(\mathtt{G})$ of spanning trees of a connected graph, also considered as the *complexity* of the graph, because it provides a measure for the global reliability of a network. Its determination is a problem of fundamental interest in mathematics and physics, first addressed by Kirchhoff in his analysis of electric circuits [3]. His theorems provides a universal algorithm for the computation of $\tau(\mathtt{G})$ in terms of the determinant of the Laplacian matrix of the graph (1.6). In the case of a complete (bipartite) graph, $\tau(\mathtt{G})$ is given by the simple following formula.

Theorem 1.1.4 (Cayley's formula). *The number of spanning trees of a complete monopartite graph is*

$$\tau(\mathtt{K}_V) = V^{V-2}, \tag{1.7}$$

while for a complete bipartite graph is

$$\tau(\mathtt{K}_{V,U}) = V^{U-1}U^{V-1}. \tag{1.8}$$

Note that the first formula provides in general the number of unlabelled spanning trees on V vertices. We will explicity derive this result from Kirchoff matrix-tree theorem in Sect. 2.3, but several other interesting proofs exist, one of the simplest and most elegant can be found e.g. in [4].

1.2 Optimization problems

Many problems of both practical and theoretical importance concern themselves with the choice of a "best" configuration or set of parameters to achieve some goal. They are referred to as *optimization problems* and a general definition for them can be stated in the following form. An *instance of an optimization problem* is a pair $(\mathscr{F}, \mathcal{C})$, where $\mathscr{F}$ is any set representing the domain of *feasible solutions*, whereas $\mathcal{C}$ is the *cost function*

$$\mathcal{C} : \mathscr{F} \longrightarrow \mathbb{R}. \tag{1.9}$$

Usually, the target is to find the *globally optimal* solution, i.e. an element $x_0 \in \mathscr{F}$ such that

$$\mathcal{C}[x_0] = \min_{x \in \mathscr{F}} \mathcal{C}[x]. \tag{1.10}$$

whose existence is a priori not guaranteed. For every optimization problem, one can also formulate its so called *decision* versions. For instance, one can wonder if the set

$$S_c = \{x \in \mathscr{F} : \mathcal{C}[x] < c\} \tag{1.11}$$

is empty or not for a given constant c.

In the theory of computational complexity [5], each optimization problem is classified according to the running time (number of computational operations) and memory required to evaluate the decision problem or to find its solution. The knowledge of the algorithmic complexity of a family of problems induces a hyerarchy of classes, depending on the asymptotic resolution time for each problem. For example, an algorithm is said to be *polynomial* if the running time is bounded from above by a certain polynomial in the size of the input, and *superpolynomial* otherwise. We say that an optimization problem belongs to the class $\mathtt{P}$ of *polynomial time problems* if there exists a polynomial algorithm that solves it.

Another relevant class is the one of *non-deterministic polynomial problems* $\mathtt{NP}$, defined as the set of problems whose decision version can be solved in polynomial time. Furthermore, given a complexity class of problems, one defines the subclass of *complete* problems as the set of problems to which all the others can be reduced, up to a polynomial factor in the complexity. It is obvious that $\mathtt{P} \subseteq \mathtt{NP}$, but it is one of the millennium problems proving whether the inclusion is tight or not. If an $\mathtt{NP}$-complete problem is found to be in $\mathtt{P}$, it would follow that $\mathtt{P} = \mathtt{NP}$, with striking consequences in the field of computer science because the $\mathtt{NP}$ class contains a large number of problems, the archetypal of which will be introduced in the next section.

1.2.1 Combinatorial optimization

In this work we concentrate on the so called *combinatorial optimization*, that deals with optimization problems in which the cardinality of $\mathcal{F}$ is *finite* for all instances $|\mathcal{F}| \in \mathbb{N}$. In this case the problem has always at least one feasible solution. However, in many cases the number of feasible solutions is extremely large, so they cannot be faced with a brute-force approach.

To exemplify the aspects described above, let us introduce three classical combinatorial optimization problems: the Traveling Salesman Problem, the Matching problem and the Minimum Spanning Tree problem.

Traveling Salesman Problem. The TSP is considered the archetypal problem in combinatorial optimization [6]. In an instance of the TSP we are given a generic connected graph $\mathtt{G} = \mathrm{Graph}(\mathcal{V}; \mathcal{E})$ with a weight function $w : \mathcal{E} \to \mathbb{R}^+$ that

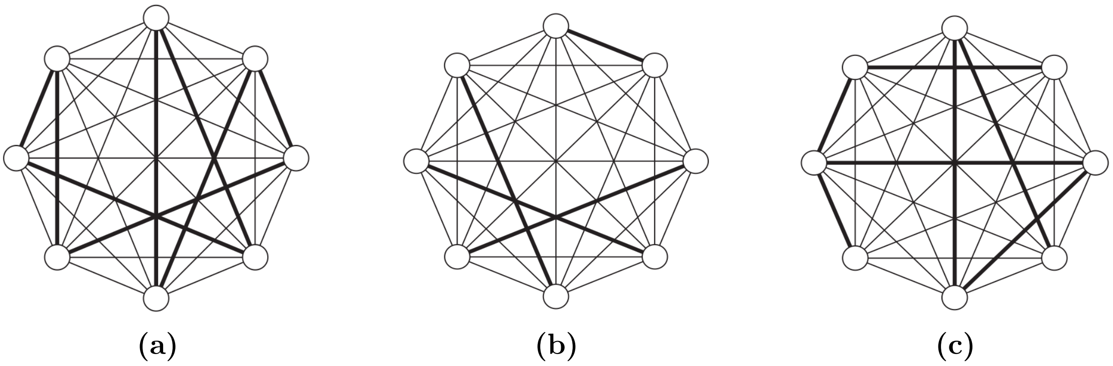


Figure 1.4: Examples of traveling salesman tour (c), perfect matching (a) and spanning tree (b) on a complete graph $\mathtt{K}_V$ with $V = 8$.

associates to each edge the cost $w(e)$ paid to travel along it. Here the set of feasible solutions $\mathscr{F}$ is composed by all possible Hamiltonian cycles $\mathtt{h} = \text{Graph}(\mathcal{V}_{\mathtt{h}}; \mathcal{E}_{\mathtt{h}})$ on the graph $\mathtt{G}$, that is to say the closed paths traversing all vertices exactly once (Fig. 1.4a). The cost function that one aims to minimize to find the cheapest tour can be written as

$$\mathcal{C}^{(\text{TSP})}[\mathtt{h}] := \sum_{e \in \mathcal{E}_{\mathtt{h}}} w(e). \tag{1.12}$$

The TSP belongs to the $\mathtt{NP}$-complete computational complexity class, even when the problem is defined on a bipartite graph, where the tour has to alternate between the two vertices' subsets. Observe that in the complete monopartite case $\mathtt{G} = \mathtt{K}_V$ one has $|\mathscr{F}| = \frac{(V-1)!}{2}$. It is therefore computationally unfeasible, even for small values of V, to try to find the optimal solution with a direct inspection of all possible values of the cost function.

Matching problems. We say that a subgraph $\mathtt{M} = \text{Graph}(\mathcal{V}_{\mathtt{M}}; \mathcal{E}_{\mathtt{M}})$ of a given graph $\mathtt{G} = \text{Graph}(\mathcal{V}; \mathcal{E})$ is a *matching* if no two edges in $\mathcal{E}_{\mathtt{M}}$ have a vertex in common. Moreover when $\mathcal{V}_{\mathtt{M}} = \mathcal{V}$ the matching is said to be *perfect* (Fig. 1.4b). If we consider a weighted graph with weight function $w : \mathcal{E} \to \mathbb{R}^+$, the matching problem consists in finding the perfect matching such that the cost functional

$$\mathcal{C}^{(\text{M})}[\mathtt{M}] = \sum_{e \in \mathcal{E}_{\mathtt{M}}} w(e) \tag{1.13}$$

is minimized. When the matching problem is defined on a complete bipartite graph $\mathtt{K}_{V,V}$ it is called *assignment problem* and the number of feasible solutions is $|\mathscr{F}| = V!$. In fact, in this case one can number the vertices in the two different classes as $\mathcal{V} = \{v_1, \dots, v_V\}$, $\mathcal{U} = \{u_1, \dots, u_V\}$. Assuming that the edge weights are given by $w : (v_i, u_j) \mapsto w_{ij}$, the cost of the optimal matching $\mathtt{M}_0$ becomes

$$\mathcal{C}^{(\text{M})}[\mathtt{M}_0] = \min_{\sigma \in \mathcal{S}_V} \sum_{i=1}^{V} w_{i\sigma(i)}, \tag{1.14}$$

where $\mathcal{S}_V$ is the group of permutations of V elements. Unlike the TSP, from the computational point of view the matching problem can be solved with quite fast algorithm, e.g. the *Hungarian algorithm* in the case of the assignment [7], which belongs to the polynomial complexity class.

Minimum Spanning Tree Problem. Being the main topic of this dissertation, we will devote the next two sections to a detailed description of the minimum spanning tree problem, which we simply define here. We are given a connected weighted graph $\mathtt{G} = \mathrm{Graph}(\mathcal{V};\mathcal{E})$ and the problem consists in finding the spanning tree $\mathtt{T}$ that has the minimal total cost, defined as

$$\mathcal{C}^{(\mathrm{MST})} = \sum_{e\in\mathcal{E}_{\mathtt{T}}} w(e). \tag{1.15}$$

From corollary 1.1.3 we know that the cardinality of the set of feasible solutions is $|\mathscr{F}| = V^{V-2}$ in the case of the complete graph $\mathtt{K}_V$ and $|\mathscr{F}| = V^{U-1}U^{V-1}$ for the bipartite case $\mathtt{K}_{V,U}$. The algorithms that find the MST of a given graph are all polynomial in the running time (see Sect 1.2.3), so the problem is in the $\mathtt{P}$ class. This is true for the related decision problems too, such as determining whether a specific edge is in the MST or if the total weight exceeds a certain threshold c.

1.2.2 The Minimum Spanning Tree

The Minimum Spanning Tree problem is one of the most typical and well-known problems of combinatorial optimization. Methods for its solution, though simple, have generated important ideas of modern combinatorics and have played a central role in the design of computer algorithms [8]. It is standard practice among authors discussing the MST to refer to Kruskal (1956) [9] and Prim (1957) [10] as the sources of the problem and its first efficient solutions, even though one can find references in the literature as early as 1926 [11]. This makes the MST one of the oldest and most thorougly studied problems in *computational geometry*.

In addition to its long-standing theoretical and algorithmic interest, the MST is useful for many practical purposes because its search in a given network stems in several optimization problems. For this reason, the MST is used in document clustering [12], wireless network connectivity [13], analysis of gene expression data [14], percolation analysis [15] and modeling of turbulent flows [16], among other areas. Moreover, the MST is used in several exact and approximation algorithms for other combinatorial optimization problems, such as the TSP and the matching problem [17, 18]. Let us note that all the problems mentioned here are commonly formulated in the Euclidean setting, so the graph $\mathtt{G} = \mathrm{Graph}(\mathcal{V};\mathcal{E})$ on which the MST is originally defined is supposed to be embedded in a Euclidean domain $\Omega \subset \mathbb{R}^d$, i.e. every $v \in \mathcal{V}$ corresponds to a d-dimensional vector.

We now proceed to state two simple properties of the MST, which allow one to add or remove edges from consideration when searching for it in a given weighted graph with weight function $w : \mathcal{E} \to \mathbb{R}^+$. They are intended as crucial for our problem, because all algorithms in the literature have their roots in them in some way. Firstly, let us observe that a sufficient (but not necessary) condition for the MST to be *unique* is that all edge weights in the graph are distinct, i.e. for every $e_i, e_j \in \mathcal{E}$ if $i \neq j$ then $w(e_i) \neq w(e_j)$.

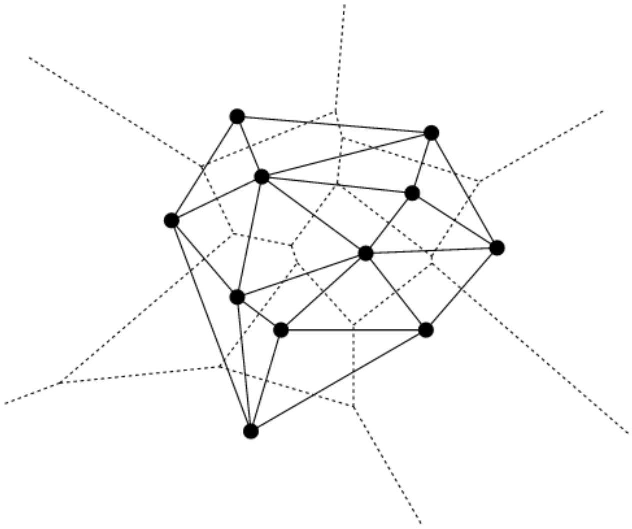


Figure 1.5: Example of Voronoi diagram (dashed) and its dual graph, the Delaunay triangulation (solid), for a given set of points on the plane.

Lemma 1.2.1 (cycle property). *Consider any cycle* $\mathtt{C} \subseteq \mathtt{G}$ *and an edge* $e \in \mathcal{E}_{\mathtt{C}}$ *with maximal cost among all edges of* $\mathtt{C}$. *Then* e *cannot belong to a MST of* $\mathtt{G}$.

Proof. Assume that e belongs to a MST $\mathtt{T}_1 \subset \mathtt{G}$. Then deleting e will break $\mathtt{T}_1$ into two subtrees, which can be reconnected considering the remainder of the cycle $\mathtt{C}$. Hence, there is an edge $e' \in \mathcal{E}_{\mathtt{C}}$ with $w(e') < w(e)$ that forms a new tree $\mathtt{T}_2 \subset \mathtt{G}$ with total cost smaller than that of $\mathtt{T}_1$. □

Lemma 1.2.2 (cut property). *Let* $\mathcal{X}$ *be a cut of the graph* $\mathtt{G}$ *and let* $e \in \mathcal{X}$ *be a minimal-cost edge. Then* e *belongs to all MSTs of the graph* $\mathtt{G}$.

Proof. Suppose that there is a MST $\mathtt{T}$ not containing e. Adding e to $\mathtt{T}$ will produce a cycle which crosses the cut once at e and crosses it back at another edge e'. If we delete e' we obtain a new spanning tree $\mathtt{T}'$ with total edge weight smaller than that of $\mathtt{T}$. □

In the Euclidean setting, the MST has another interesting property which finds its application in algorithms design. To state it, we need to introduce some elements of computational geometry [19] first, focusing for the sake of simplicity on the two dimensional case. Nonetheless, everything can be readily generalized to higher dimensions.

Let $\mathcal{P} \subset \mathbb{R}^2$ be a set of n points, called *seeds.* The *Voronoi diagram* of $\mathcal{P}$ is the subdivision of the plane into n regions, one for each point in $\mathcal{P}$, such that the region of a site $p \in \mathcal{P}$ contains all points in the plane for which p is the closest site. Furthermore, a *triangulation* of $\mathcal{P}$ is the partition of the smallest convex set containing $\mathcal{P}$ (*convex hull*) into non intersecting triangles having vertices in $\mathcal{P}$. A frequently used point set triangulation is the so called *Delaunay triangulation*, composed by the set of triangles that are circumscribed by a circle containing no points in $\mathcal{P}$. Working in two dimensions, it is clear from the definition that a triangulation can be understood as a planar graph $\mathtt{D} = \mathrm{Graph}(\mathcal{V}_{\mathtt{D}}; \mathcal{E}_{\mathtt{D}})$ with set of vertices $\mathcal{V}_{\mathtt{D}} = \mathcal{P}$. This observation allows us to say that, given a set of points $\mathcal{P}$, its Delaunay triangulation is the *dual graph* of its Voronoi diagram (Fig. 1.5). In practice, the former has a vertex for every Voronoi region, and it has an edge between two vertices if the corresponding regions share a boundary.

As anticipated, given the above definitions we can now state the following result

Proposition 1.2.3. *Consider a graph* $\mathtt{G} = \mathit{Graph}(\mathcal{V};\mathcal{E})$ *with* $\mathcal{V} \subset \mathbb{R}^d$ *and edge weights given by* $\|e\|^p$ *for all* $e \in \mathcal{E}$*, where* $\|\cdot\|$ *represents the standard Euclidean norm in* $\mathbb{R}^d$*. Then the MST of* $\mathtt{G}$ *is a subgraph of the Delaunay triangulation of* $\mathcal{V}$*.*

1.2.3 Algorithms for the MST

All classical algorithms for the solution of the MST problem belong to the class of so called *greedy algorithms*, i.e. they follow the problem solving heuristic of making the local optimal choice at each stage, with the purpose of finding a global optimum. They basically rely on lemma 1.2.2 to form cuts in the graph and add the minimum weight edge across each at every stage. Examples of algorithms using this rule are the already mentioned Kruskal's [9] and Prim's [10], which require $\mathcal{O}(E \log V)$ and $\mathcal{O}(E + V \log V)$ time respectively on a graph with V vertices and E edges.

It is worth noting that besides being effective in the direct solution of the problem they are designed for, the algorithms can prove to be very helpful in the theoretical analysis of general aspects of an optimization problem. We will find examples of this in Sects. 2.1 and 3.3 in the case of Kruskal's algorithm, and that is the reason why we describe it here in detail. This also gives us the occasion to show explicitly why the MST problem belongs to the $\mathtt{P}$ computational complexity class. Given a connected weighted graph, Kruskal's algorithm proceeds as described by the following pseudocode (see Fig. 1.6 for a reference).

Algorithm 1 Kruskal's algorithm

```
1: function KRUSKAL(Graph($\mathcal{V};\mathcal{E}$))
2:     $F = \text{Graph}(\mathcal{V};\emptyset) \leftarrow$ starting forest with no edges
3:     $W \leftarrow$ set of the edges sorted in increasing order w.r.t. their weights
4:     while $W \neq \emptyset$ do
5:         Remove the edge $(u,v)$ with minimum weight from $W$
6:         if $(u,v)$ connects two different trees of $F$ then
7:             $F = F \cup \{(u,v)\}$
8:     return $F$
```

Regarding the computational complexity of the above algorithm, the E edges sorting procedure can be carried out in $\mathcal{O}(E \log E)$ time using a simple comparison algorithm: this allows step (3) to operate in $\mathcal{O}(1)$ time. We are then left with $\mathcal{O}(V)$ operations to perform (4), as in each iteration the algorithm finds the components to which u and v belong, possibly joining them. This can be done in $\mathcal{O}(V \log V)$ time by using any *disjoint-set data structures*, which tracks a set of

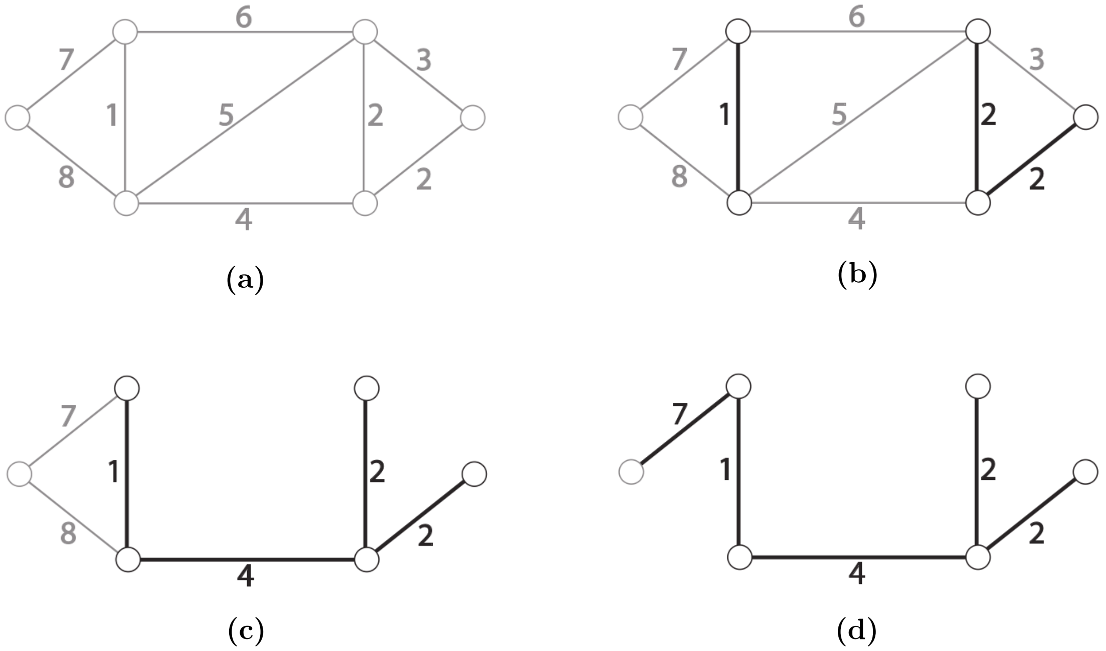


Figure 1.6: Example of MST construction (bold) as performed by Kruskal's algorithm on a given weighted graph (a). Starting from the smallest weight, an edge is added to the MST if it does not form any cycles in the developing spanning forest `F` (b), otherwise it is discarded (c). The algorithm stops when all edges have been considered (d).

elements partitioned into non-overlapping subsets by means of pointers. Considering that E is at most $\mathcal{O}(V^2)$ in a simple graph, and that $V = \mathcal{O}(E)$, we finally obtain the anticipated total running time $\mathcal{O}(E \log V)$.

In recent years new sophisticated algorithms have been developed for the MST problem on general graphs. The fastest non-randomized comparison-based algorithm with known complexity [20] has running time $\mathcal{O}(E\,\alpha(E, V))$, where α is the classical functional inverse of Ackermann function. It grows extremely slowly with its arguments, so that for practical purposes it may be considered a constant smaller than 4, leading to an almost linear time algorithm.

All the cited general algorithms are insufficient for large, metric problems because they depend linearly on the number of edges E. The edge set of a graph consists of all pair of points, therefore linear scaling in E means quadratic scaling in the number of vertices V, which are usually the input data in the Euclidean case. For this reason, most current Euclidean MST algorithms start by computing a set of edges which can be shown to be a superset of the edge set of the MST. For instance, in the two dimensional case a good choice is the Delaunay triangulation, dual of the Voronoi diagram for the V points, which can be constructed in $\mathcal{O}(V \log V)$ time [21]. Unfortunately, this bound worsens to $\mathcal{O}(V^2 \log V)$ in $d \geq 3$, because the edge set of the Delaunay triangulation happens to be the complete graph. This problem has been partially overcome e.g. by the work of Agarwal et al. [22], who related the running time of the minimum spanning tree problem in $\mathbb{R}^d$ to the *bichromatic closest pair problem*, formulated as follows. Given a set of

N red and M blue points in $\mathbb{R}^d$, find a couple such that the distance is minimum among all red-blue pairs.

We end this section noting that all the algorithms described above are designed to work on a monopartite graph. Very little literature exists on algorithms constructed specifically for the MST problem on multipartite graphs. A remarkable example is the recent paper by Biniaz et al. [23], who obtained in two dimensions the time complexity bound $\mathcal{O}(V \log V \log k)$ for the k-partite case with V total vertices.

1.3 Random optimization problems: worst vs typical case

In Sect. 1.2.1 we introduced optimization problems defined on graphs, always supposing that the parameters of the problems, e.g. the weights associated to the edges of the graph, were assigned once and for all. Given an instance of such an optimization problem, specific algorithms allow us to classify the problem according to the computational resources (time and memory) required to solve it.

Note that the theory of computational complexity is based on a pessimistic attitude: a problem's tractability is defined depending on the *worst* possible instance. Quite often this worst case scenario differs considerably from the *typical* case, averaged over a reasonable ensemble of instances. A common observation is that "hard" problems are typically "easy" to solve, and to get real hard instances the parameters must be carefully tuned to certain critical values. Varying the parameters across the critical region often leads to abrupt changes in complexity, related to changes in the structure of the set of feasible solutions [24], very similar to what happens with phase transitions in physical systems.

This arguments justifies the consideration of *random instances* of an optimization problem, in order to study its general properties such as its complexity and its solution *in average*, for large sizes of the input. A first kind of randomization for a combinatorial optimization problem defined on a weighted graph $\mathtt{G} = \mathrm{Graph}(\mathcal{V}; \mathcal{E})$ can be performed on the graph itself. In the simplest case we can choose the weights $\{w(e)\}_{e\in\mathcal{E}}$ as *independently and identically distributed random variables* with distribution density $\rho(w)$. The distribution itself defines therefore an ensemble of random instances and we ask for the typical properties such as the average optimal cost, the optimal cost distribution, and so on, for a given ensemble in the thermodynamic limit $|\mathcal{V}| \to \infty$.

It is easy to see in general that an optimization problem can be stated as a physical problem. Indeed, the set of possible solutions can be interpreted as a configuration space, and the cost can be chosen as the Hamiltonian of the system. Thus, cost minimization turns into finding the ground state of the physical system, when frozen at zero temperature. As we said above, the cost function depends on a large number of parameters, and one is interested in the average case with respect to a measure on the parameters space. Therefore, the corresponding physical system is a *disordered* system, and the probability measure over the disorder corresponds to this measure over the parameters space, i.e. on the space of possible instances of the random optimization problem.

Interestingly, this probabilistic approach to combinatorial optimization problems shed new light on their mathematical properties, and many results have been obtained in recent years. Since the 1985 seminal work by Mézard and Parisi [25], many techniques developed by physicists in the field of statistical physics started to be used effectively on random optimization problems. For example, with a method borrowed from *spin glasses*, namely the *replica method* (see Sect. 1.3.2), the two authors analyzed the random bipartite matching problem (RBMP) with i.i.d. edge weights [26]. In particular, they were able to obtain the average optimal cost of the problem, defined in Eq. (1.14),

$$\mathcal{C}^{(\mathrm{RBMP})} = \lim_{N\to\infty} \overline{\mathcal{C}^{M}[\mathtt{M}_0]} = \zeta(2) = \frac{\pi^2}{6}, \tag{1.16}$$

where $\overline{\bullet}$ denotes the average over the distribution of the weights and $\zeta(z)$ is the Riemann zeta function. Remarkably, only 25 years later Aldous [27] confirmed the obtained result by providing a rigorous mathematical treatmeant of the problem.

1.3.1 Statistical mechanics of disordered systems

As explained in the previous section, in random optimization problems we are not interested on specific instances, but in the average properties of the problem, possibly dependending on some parameters in the cost function and on the way we introduce randomness in the problem itself. Over the years statistical physics has developed a lot of techniques to deal with systems with a huge number of degrees of freedom, even in the presence of *disorder*. For this reason, we will introduce here the fundamental concepts of the statistical mechanichs of disordered systems.

Let us consider a graph $\mathtt{G} = \mathrm{Graph}(\mathcal{V};\mathcal{E})$, $V = |\mathcal{V}|$, and suppose that we assign to each node $v_i \in \mathcal{V}$ an *Ising spin variable* $\sigma_i \in \{-1, 1\}$ and to each edge $e \equiv (v_i, v_j) \in \mathcal{E}$ an interaction energy $-J_{ij}\sigma_i\sigma_j$, with $J_{ij} \in \mathbb{R}\ \forall i, j$. The *Hamiltonian* functional of the system has the following form

$$H_{\mathtt{G}}[\boldsymbol{\sigma}; \mathbf{J}, h] := -\sum_{\langle ij\rangle} J_{ij}\sigma_i\sigma_j - h\sum_{i=1}^{V}\sigma_i, \tag{1.17}$$

with $h \in \mathbb{R}$ a fixed quantity, while the parameters $\mathbf{J} = \{J_{ij}\}_{ij}$ are independently extracted from a certain probability distribution $\rho(J_{ij})$, identical for all $(v_i, v_j) \in \mathcal{E}$. Here we assume that these random quantities are specified once and for all for each instance, and therefore we say that the disorder is *quenched*. In other words, the random parameters are supposed fixed on the time scale in which the degrees of freedom $\boldsymbol{\sigma} := \{\sigma_i\}_{i=1,\dots,V}$ of the system fluctuate. If the system is defined on the hypercubic lattice in d dimensions the model is called *Edwards-Anderson model* (EA-model) and it is an example of *spin glass*. Here we just mention that spin glasses represent the reference frame in which physicists analyze the peculiar effects of disorder on the behaviour of systems with a large number of degrees of freedom, and their importance goes beyond the application to physical systems. An interesting introduction to this argument can be found e.g. in [28, 29].

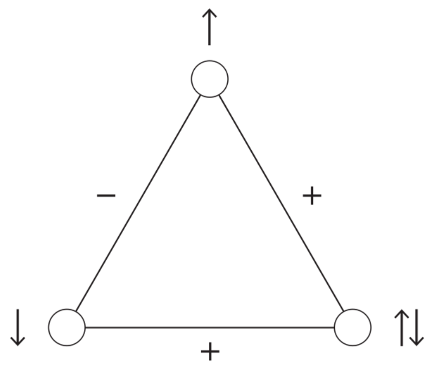


Figure 1.7: The simplest example of frustration

To each configuration $\boldsymbol{\sigma} = \{\sigma_i\}_i$ we can associate a Boltzmann-Gibbs weight

$$\mu_{\mathtt{G}}[\boldsymbol{\sigma}; \mathbf{J}, \beta, h] := \frac{1}{Z_{\mathtt{G}}[\mathbf{J}, \beta, h]} \mathrm{e}^{-\beta H_{\mathtt{G}}[\boldsymbol{\sigma}; \mathbf{J}, h]} \tag{1.18}$$

where β is the inverse *temperature* and with the normalization given by the partition function of the system

$$Z_{\mathtt{G}}[\mathbf{J}, \beta, h] := \sum_{\boldsymbol{\sigma}} \mathrm{e}^{-\beta H_{\mathtt{G}}[\boldsymbol{\sigma}; \mathbf{J}, h]}, \tag{1.19}$$

which plays a central role in the computation of many physical quantities of interest. Finally, given a function $g := g(\boldsymbol{\sigma})$, we denote its expectation as

$$\langle g \rangle_J := \sum_{\boldsymbol{\sigma}} g(\boldsymbol{\sigma}) \mu_{\mathtt{G}}[\boldsymbol{\sigma}; \mathbf{J}, \beta, h]. \tag{1.20}$$

Note that together with randomness, represented here by the fact that the coupling constants are random variables taken with a certain distribution function, a disordered system is characterized by the presence of *frustration*. Differently from a standard ferromagnetic system, in a frustrated one it becomes impossible to satisfy all the couplings at the same time. Formally a system is frustrated if there exists a loop on which the product of the couplings is negative: what happens is that if we fix an initial spin and try to chain-fix the others one at a time to minimize the energy, we are bound to return to the initial spin and flip it (Fig. 1.7). The energy landscape of frustrated systems is thus often non trivial and it is not obvious what is the structure of the minimum energy configuration.

Thanks to the fact that the disorder is quenched, to evaluate a physical quantity using the Hamiltonian of Eq. (1.17) it is appropriate to trace over the spin variables first, with the interactions $\mathbf{J} = \{J_{ij}\}_{ij}$ fixed. For instance, the free energy of the system is calculated as

$$F_{\mathtt{G}}[\mathbf{J}, \beta, h] = -\frac{1}{\beta} \log \sum_{\boldsymbol{\sigma}} \mathrm{e}^{-\beta H_{\mathtt{G}}[\boldsymbol{\sigma}; \mathbf{J}, h]}. \tag{1.21}$$

The next step is to average over the distribution of $\mathbf{J}$, i.e. over the disorder, and this is done through the so called *configurational average*

$$\overline{F_{\mathtt{G}}[\mathbf{J}, \beta, h]} := -\frac{\overline{\log Z_{\mathtt{G}}[\mathbf{J}, \beta, h]}}{\beta} = -\frac{1}{\beta} \left(\prod_{\langle ij \rangle} \int \rho(J_{ij})\, \mathrm{d}J_{ij} \right) \log Z_{\mathtt{G}}[\mathbf{J}, \beta, h]. \tag{1.22}$$

Differentiation of this averaged free energy by the external field h or the inverse temperature β leads to the magnetization or the internal energy, respectively. Note that in general we want that physical observables do not depend on the specific realization of the system, and this is guaranteed by the following crucial property. The free energy density $f_{\mathtt{G}} = F_{\mathtt{G}}/V$ for a disordered system on a generic graph $\mathtt{G}$ has vanishingly small deviations from its mean value $\overline{f_{\mathtt{G}}}$ in the thermodynamic limit $|\mathcal{V}| \to \infty$ [29], i.e. $f_{\mathtt{G}}$ is *self-averaging*

$$\lim_{|\mathcal{V}|\to\infty} \frac{\overline{(f_{\mathtt{G}} - \overline{f_{\mathtt{G}}})^2}}{\overline{f_{\mathtt{G}}}^2} = 0. \tag{1.23}$$

This means that for large systems the raw value $f_{\mathtt{G}}$ for a given set of random parameters $\mathbf{J}$ agrees with its configurational average $\overline{f_{\mathtt{G}}}$ with probability 1. Therefore, we can choose either of these quantities in actual calculations, with the second one much easier to handle having no dependence on $\mathbf{J}$ even for finite-size systems.

1.3.2 The replica method

The dependence of $\log Z_{\mathtt{G}}$ on the random variables $\mathbf{J}$ is very complicated and it is not easy to compute its configurational average. Nonetheless, the manipulations are greatly facilitated by the general relation

$$\overline{\log Z_{\mathtt{G}}} = \lim_{n\to 0} \frac{\overline{Z_{\mathtt{G}}^n} - 1}{n}, \tag{1.24}$$

One prepares n replicas of the original system, evaluates the configurational average of the product of their partition functions $Z_{\mathtt{G}}^n$, and then takes the limit $n \to 0$. This technique, the *replica method*, is very useful because it is easier to evaluate $\overline{Z_{\mathtt{G}}^n}$ than $\overline{\log Z_{\mathtt{G}}}$.

Note that Eq. (1.24) is an identity valid in general, but to carry out the calculations one has to consider n as a positive integer. Only at the end the limit $n \to 0$ is performed as a sort of analitic continuation, and this is quite often the most difficult point of the whole procedure in actual calculations. We will now conclude this chapter showing how does the replica approach work in the simplest possible case, namely the *Sherrington-Kirkpatrick (SK) model.* This task has been fulfilled for the first time by Parisi and collaborators at the beginning of the 80's, and since then the model has become a prototype for mean-field disordered systems. For a more extensive presentation, we refer the reader to [28, 30].

The Sherrington-Kirkpatrick model. The SK model represents the infinite range version of the Edwards-Anderson model, i.e. the graph on which the Hamiltonian (1.17) is defined is the complete graph $\mathtt{G} \equiv \mathtt{K}_N$. The interactions $\mathbf{J} = \{J_{ij}\}_{ij}$ are quenched variables with Gaussian distribution function

$$\rho(J_{ij}) = \frac{1}{J}\sqrt{\frac{N}{2\pi}} \exp\left\{-\frac{N}{2J^2}\left(J_{ij} - \frac{J_0}{N}\right)^2\right\}, \tag{1.25}$$

with the normalization for the mean and for the variance chosen to obtain extensive quantities proportional to N in subsequent calculations.

According to the prescription of the replica method we have to compute the configurational average of the n-th power of the partition function. It follows that

$$\begin{aligned}\overline{Z^n} &= \int \left(\prod_{i<j} \rho(J_{ij})\,\mathrm{d}J_{ij} \right) \mathrm{Tr} \exp \left(\beta \sum_{i<j} J_{ij} \sum_{a=1}^{n} \sigma_i^a \sigma_j^a + \beta h \sum_{i=1}^{N} \sum_{a=1}^{n} \sigma_i^a \right) \\ &= \exp\left(\frac{N\beta^2 J^2 n}{4} \right) \mathrm{Tr} \exp \left[\frac{\beta^2 J^2}{2N} \sum_{a<b} \left(\sum_i \sigma_i^a \sigma_i^b \right)^2 + \right. \\ &\left. + \frac{\beta J_0}{2N} \sum_a \left(\sum_i \sigma_i^a \right)^2 + \beta h \sum_i \sum_a \sigma_i^a \right], \end{aligned} \tag{1.26}$$

where we have carried out the Gaussian integral over J_{ij} independently for each (ij), and we have rewritten the sums over $i < j$ and a, b (namely the replica indices). In order to linearize the exponent with respect to the spin variables, so as to perform the trace operation, we introduce the auxiliary variables q_{ab} and m_a by means of an *Hubbard-Stratonovich transformation*. We thus arrive to

$$\begin{aligned}\overline{Z^n} = \exp\left(\frac{N\beta^2 J^2 n}{4} \right) \int \prod_{a<b} \mathrm{d}q_{ab} \int \prod_a \mathrm{d}m_a \\ \cdot \exp \left(-\frac{N\beta^2 J^2}{2} \sum_{a<b} q_{ab}^2 - \frac{N\beta J_0}{2} \sum_a m_a^2 + N \log \mathrm{Tr}\, \mathrm{e}^L \right), \end{aligned} \tag{1.27}$$

where

$$L = \beta^2 J^2 \sum_{a<b} q_{ab} \sigma^a \sigma^b + \beta \sum_a (J_0 m_a + h) \sigma^a. \tag{1.28}$$

Observe that $\mathrm{Tr}\, \mathrm{e}^L$ appears as a sort of partition function for a set of n coupled spins, each one associated to one of the n replica indices.

The exponent of the above integral is proportional to N, so that it is possible to evaluate the integral by *steepest descent* in the thermodynamic limit $N \to \infty$

$$\begin{aligned}\overline{Z^n} &\approx \exp \left(-\frac{N\beta^2 J^2}{2} \sum_{a<b} q_{ab}^2 - \frac{N\beta J_0}{2} \sum_a m_a^2 + N \log \mathrm{Tr}\, \mathrm{e}^L + \frac{N\beta^2 J^2 n}{4} \right) \\ &\approx 1 + Nn \left(-\frac{\beta^2 J^2}{4n} \sum_{a\neq b} q_{ab}^2 - \frac{\beta J_0}{2n} \sum_a m_a^2 + \frac{1}{n} \log \mathrm{Tr}\, \mathrm{e}^L + \frac{\beta^2 J^2}{4} \right), \end{aligned} \tag{1.29}$$

where in the last line we have taken the limit $n \to 0$ with N kept very large but finite. Note that the correct order for the two limits above should be $N \to \infty$ after $n \to 0$, but we took $N \to \infty$ first so that the saddle point approximation is applicable. This is importante to remark, since at an early stage of research it was suspected that the problems arising in the replica method were due to this apparently inappropriate exchange of limits. Nonetheless, this passage proved to be correct in subsequent years, the source of trouble being elsewhere, as we will point out below.

The values of q_{ab} and m_a in the above expression are those that satisfies the extremality conditions, i.e.

$$q_{ab} = \frac{1}{\beta^2 J^2} \frac{\partial \log \operatorname{Tr} \mathrm{e}^L}{\partial q_{ab}} \equiv \langle \sigma^a \sigma^b \rangle_L \tag{1.30a}$$

$$m_a = \frac{1}{\beta J_0} \frac{\partial \log \operatorname{Tr} \mathrm{e}^L}{\partial m_a} \equiv \langle \sigma^a \rangle_L, \tag{1.30b}$$

with $\langle \cdot \rangle_L$ denoting the average by the weight e^L. Remarkably, q_{ab} are not merely a set of variables introduced for technical convenience, indeed they express a sort of "average superposition" between replicas, being

$$q_{ab} = \overline{\left\langle \sigma_i^a \sigma_i^b \right\rangle_{\mathtt{R}}}, \tag{1.31}$$

where $\langle \bullet \rangle_R$ is the average with respect to the replicated system, whose Hamiltonian is

$$H^{\mathrm{R}}[\{\boldsymbol{\sigma}^a\}_a; \mathbf{J}, h] := \sum_{a=1}^{n} \left(-\sum_{i<j} J_{ij} \sigma_i^a \sigma_j^a - h \sum_{i=1}^{N} \sigma_i^a \right). \tag{1.32}$$

In particular, the variables q_{ab} play the role of *spin glass order parameters*, while m is the ordinary ferromagnetic order parameter according to (1.30b), and is the value of m_a when the latter is independent of a. If we suppose that all replicas are equivalent (*replica symmetric hypothesis*) we have

$$q_{ab} = \overline{\left\langle \sigma_i^a \sigma_i^b \right\rangle_{\mathtt{R}}} = \overline{\left\langle \sigma_i^a \right\rangle_{\mathtt{R}} \left\langle \sigma_i^b \right\rangle_{\mathtt{R}}} = \overline{\langle \sigma_i \rangle^2} = q, \quad a \neq b. \tag{1.33}$$

We expect that for $\beta \to 0$ the spins are randomly oriented, and therefore $q = 0$, whilst in the $\beta \to \infty$ limit $q > 0$, having $\langle \sigma_i \rangle \neq 0$ for each realization.

Proceeding with the computation in the replica symmetric hypothesis, with the notation $\mathrm{D}z := \mathrm{d}z \exp(-z^2/2)/\sqrt{2\pi}$ and $\widetilde{H}(z) = J\sqrt{q}z + J_0 m + h$, one finds [30]

$$-\beta \overline{f} = \frac{\beta^2 J^2}{4}(1 - q^2) - \frac{\beta J_0 m^2}{2} + \int \mathrm{D}z \log[2 \cosh \beta \widetilde{H}(z)], \tag{1.34}$$

from which we obtain the following equations of state through the extremization conditions

$$q = \int \mathrm{D}z \tanh^2 \beta \widetilde{H}(z) \tag{1.35a}$$

$$m = \int \mathrm{D}z \tanh \beta \widetilde{H}(z). \tag{1.35b}$$

Even if the replica approach of the SK-model seems to give a complete solution, an anomalous behaviour emerges at low temperature. For instance, a direct computation gives a negative value of the *entropy* density s

$$\lim_{\beta \to \infty} s = \lim_{\beta \to \infty} \beta^2 \frac{\partial \overline{f}}{\partial \beta} = -\frac{1}{2\pi} \tag{1.36}$$

In particular, Almeida and Thouless [68] found that the replica symmetric solution for the SK-model is not stable in the absence of an external field when $\beta > 1/J$. Then, in 1980, Parisi [69] proposed a solution to the problem by showing that a proper break of the replica symmetry was needed. Remarkably, only years later this solution was proven with mathematical rigour [70, 71]. Despite being extremely interesting, this argument goes beyond the purposes of this introduction, thus we will not go any further, but the reader can find out more in the previosly cited references.

Chapter 2

The MST in the mean field approximation

In this chapter we focus on the random MST problem with independent and identically distributed edge weights, which can be regarded as a sort of *mean field* approximation of the correlated case. After a brief review of the existing literature on the topic, our attention turns to the well known q-state Potts model, from which we derive as the $q \to 0$ limit the *spanning tree generating polynomial* of a general graph. With this quantity, properly rewritten in terms of a Berezin integral defined on two sets of Grassmann variables thanks to the matrix-tree theorem, in the last section we sketch a replica calculation for the purely random MST problem on the complete graph.

2.1 $\zeta(3)$ limit for the random MST

The aim of the present section is to offer a survey of the existing literature concerning the average cost of the random MST defined on a graph with independently and identically distributed edge weights. The starting point is the fundamental result obtained by Frieze in 1985 [31], which can be formulated as follows. Suppose we are given a complete graph $\mathtt{K}_N$ on N vertices in which the edge weights are i.i.d. non-negative random variables. Assume also that their common cumulative distribution function F is differentiable at zero, with $D := F'(0) > 0$. Denoting by W a random variable with this distribution, the following holds for the average optimal cost $\overline{\mathcal{C}_N} := \overline{\mathcal{C}\left[\mathtt{K}_N\right]}$ of the MST.

Theorem 2.1.1. *If W has finite mean, then*

$$\lim_{N\to\infty} \overline{\mathcal{C}_N} = \frac{\zeta(3)}{D}, \qquad \zeta(3) = \sum_{k=1}^{\infty} \frac{1}{k^3} = 1.202\ldots \tag{2.1}$$

Moreover, if W has finite variance, then

$$\lim_{N\to\infty} Pr\left(\left| \mathcal{C}_N - \frac{\zeta(3)}{D} \right| > \epsilon \right) = 0. \tag{2.2}$$

A fact worth noting is that this theorem strongly relies on Kruskal's algorithm procedure. In fact, the author is able to find asymptotics bound for the total average MST cost by writing it in terms of the sequence of successive spanning forests which originates as described in Sect. 1.2.3.

Frieze's result can be generalized in two different ways. First it can be shown [32] that the condition on the distribution function $F'(0) > 0$ is sufficient for convergence in probability, and no other smoothness or moment conditions are required. Furthermore, Theorem 2.1.1 can be stated for graphs different from $\mathtt{K}_N$ [33]. For simplicity, let us consider the case where each edge weight is a uniform random variable on $[0,1]$. Then, e.g. for the complete q-partite graph $(\mathtt{K}_q)_N$ with q classes all of cardinality N, we have

$$\lim_{N\to\infty} \overline{\mathcal{C}\left[(\mathtt{K}_q)_N\right]} \to \frac{q}{q-1}\,\zeta(3) \tag{2.3}$$

We see that for a complete bipartite graph ($q=2$), the difference to the monopartite case is given simply by a factor 2, as it happens in the random matching problem [26].

Note that for uniform distributed edge weights a central limit theorem can be exhibited too for the average cost of the random MST [34, 35], i.e.

$$N^{\frac{1}{2}}\left(\overline{\mathcal{C}_N} - \zeta(3)\right) \xrightarrow{\mathrm{d}} N\left(0,\sigma^2\right). \tag{2.4}$$

with the variance given by $\sigma^2 = 6\zeta(4) - 4\zeta(3)$.

With reference to Eq. (2.1) we have that in the large N limit $\overline{\mathcal{C}_N} = \zeta(3) + o(1)$, but ideally one would like to have an exact expansion for the average cost, as it happens for the random assignment problem [36, 37]. For uniform weights on $[0,1]$, Cooper and Frieze [38] went in this direction by improving the asymptotics for the average cost of the MST to the secondary and tertiary terms

$$\overline{\mathcal{C}_N} = \zeta(3) + \frac{c_1}{n} + \frac{c_2 + o(1)}{n^{4/3}}, \tag{2.5}$$

where

$$c_1 = -1 - \zeta(3) - \frac{1}{2}\int_0^\infty log(1-(1+x)e^{-x})\mathrm{d}x \tag{2.6a}$$

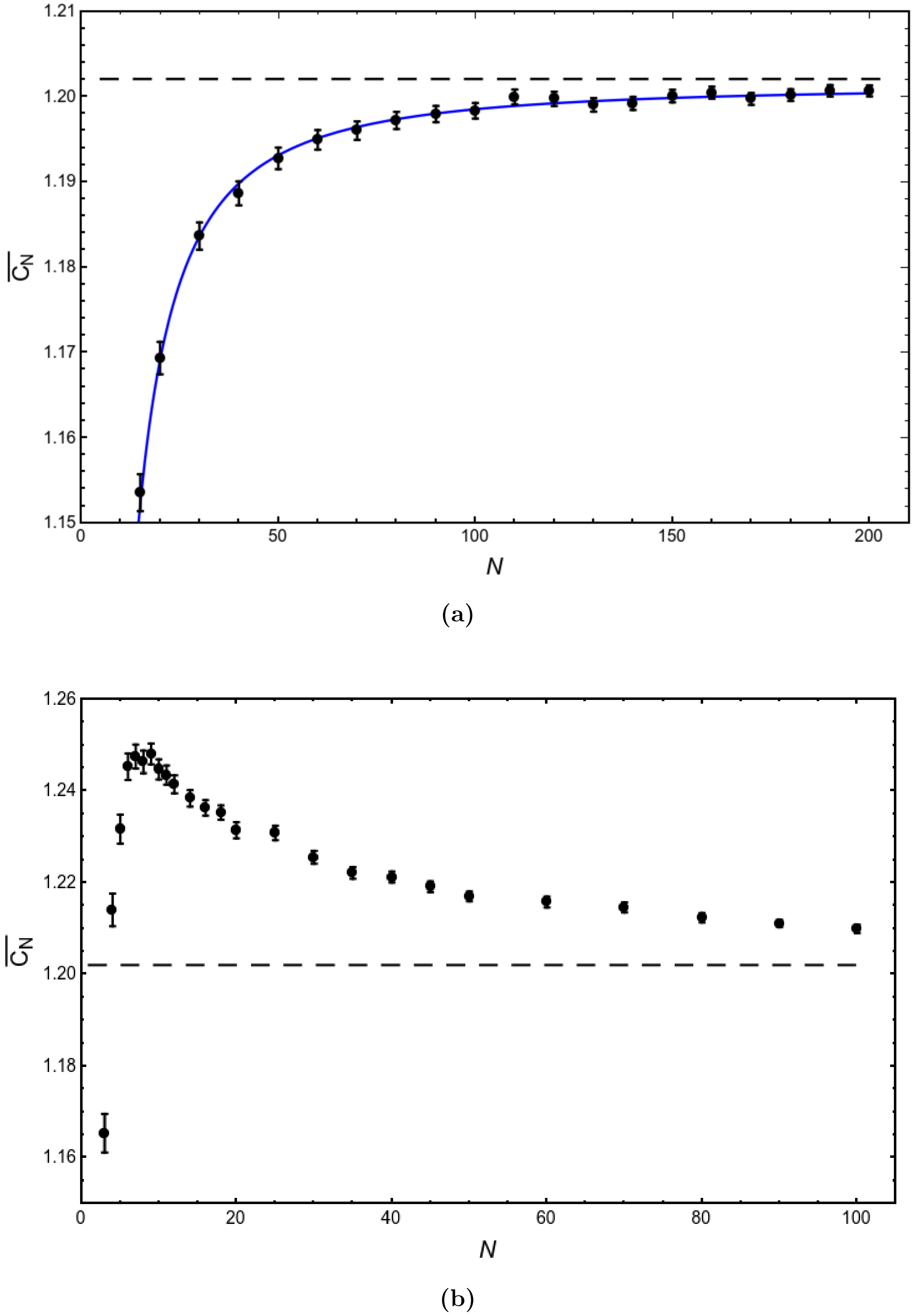
$$c_2 = \frac{2}{3}\int_0^\infty \left(y^{-2}\psi(y)e^{-y^2/24} - y^{-2} - \sqrt{\frac{\pi}{8}}y^{-1} - \frac{1}{2}\right) y^{-1/3}\mathrm{d}y \tag{2.6b}$$

In the last expression, ψ is defined as the moment generating function of the random variable $\mathcal{B}_{\mathrm{ex}}$

$$\psi(t) = \overline{\exp\left(t\mathcal{B}_{\mathrm{ex}}\right)}, \tag{2.7}$$

with $\mathcal{B}_{\mathrm{ex}} = \int_0^1 B_{\mathrm{ex}}(s)\,\mathrm{d}s$ representing the area under a normalized *Brownian excursion* (see [39] for a review on the topic). A numerical investigation performed by the authors yielded $c_1 > 0$, so for (very) large N (see Fig. 2.1a) we expect to find $\overline{\mathcal{C}_N} > \zeta(3)$.

Remarkably, Eq. (2.5) refuted Steele's conjecture [40] of the average cost increasing monotonically with N, based on its exact computation for $N \leq 8$. This



(a)

(b)

Figure 2.1: Numerical simulations for the average optimal cost of the MST with i.i.d. weights generated in the interval $[0, 1]$ with uniform distribution (a) and exponential distribution $\rho(w) = \mathrm{e}^{-w}$ (b). The dashed line represents the $\zeta(3)$ limit, while the solid blue line represents the data fit with Eq. (2.5) ($c_1 = 0.2709$, $c_2 = -2.4751$). For each N, $\overline{\mathcal{C}_N}$ is obtained by averaging over $n = 2 \cdot 10^4$ (a) and $n = 3 \cdot 10^4$ (b) instances.

result was carried out exploiting a very interesting formula involving the so called *Tutte polynomial* $T(\mathtt{G}; x, y)$, which in general contains a lot of information about a graph $\mathtt{G} = \mathrm{Graph}(\mathcal{V}; \mathcal{E})$ (it has been widely studied in the context of the algebraic properties of graphs, see e.g. [41]).

To define this polynomial, one needs to go outside the familiar class of simple graphs, taking into account loops or multi-edges too. Much of the usefulness of the Tutte polynomial comes from its relation to the so called rank function, which measures how much a graph is connected. Given an edge subset $\mathcal{A} \subseteq \mathcal{E}$ the *rank* $r(\mathcal{A})$ of $\mathcal{A}$ is defined by

$$r(\mathcal{A}) = |\mathcal{V}| - k(\mathcal{A}), \tag{2.8}$$

with $k(\mathcal{A})$ number of connected components of the graph $\mathtt{G}' = \mathrm{Graph}(\mathcal{V}; \mathcal{A})$. The Tutte polynomial of the graph $\mathtt{G}$ is then the two-variables polynomial defined as

$$T(\mathtt{G}; x, y) = \sum_{\mathcal{A} \subseteq \mathcal{E}} (x-1)^{r(\mathcal{E})-r(\mathcal{A})} (y-1)^{|\mathcal{A}|-r(\mathcal{A})}. \tag{2.9}$$

By making use of it, one can find the following interesting result valid for any finite number N of vertices of the graph $\mathtt{G}$, contrary to all the previous statements. Here we report a generalization [42] of the original theorem proved by Steele

Theorem 2.1.2. *If* $\mathtt{G} = \mathit{Graph}(\mathcal{V}; \mathcal{E})$ *is a simple, finite and connected graph and its edge weights* W_e *are positive random variables for all* $e \in \mathcal{E}$, *all with the same distribution* $F(x) = \Pr(W_e \leq x)$, *then*

$$\overline{\mathcal{C}[\mathtt{G}]} = \int_0^\infty \frac{1 - F(t)}{F(t)} \frac{T_x(\mathtt{G}; x, y)}{T(\mathtt{G}; x, y)} \, \mathrm{d}t, \tag{2.10}$$

where $x = 1/F(t)$, $y = 1/(1 - F(t))$ *and* $T_x(\mathtt{G}; x, y)$ *is the partial derivative of the Tutte polynomial w.r.t.* x. *In particular, for* $x \in (0, \infty)$ *if* $F_u(x)$ *is the uniform cumulative distribution and* $F_e(x) = 1 - e^{-x}$ *the exponential one, for every connected graph* $\mathtt{G}$ *it holds*

$$\overline{\mathcal{C}_{F_u}[\mathtt{G}]} < \overline{\mathcal{C}_{F_e}[\mathtt{G}]}. \tag{2.11}$$

2.2 The Potts model

The main result of the previous section is that given a graph with i.i.d. edge weights, the average optimal cost of the MST converges in probability to $\zeta(3)$. This limit immediately reminds us of an analogous one in the case of the assignment problem (1.16), which was computed by Mézard and Parisi [26] by means of the replica method. One is then led to wonder whether the same calculation can be performed for the random MST. To try to answer to this question, we need as a starting point an expression for the "partition function" of our system, namely a function that enumerates all the spanning trees in a given graph, weighted with their total edge cost. The aim of this section is thus to derive such an expression, which remarkably emerges as a particular limit of one of the most studied models in statistical physics, the *Potts model.*

The Potts model [43] represents one of several possible generalization of the Ising model: each spin, instead of being allowed to point only up or down, can find itself in q different states (or *colors*). Historically, the model bears its current name after Domb proposed it as a research topic to his student Potts in 1951. Thanks to the outstanding number of problems it is related to, from lattice statistics, to combinatorics, to graph theory, the Potts model has been a subject of increasingly intense research interest in recent years.

Originally, being the Ising model a system of interacting spins that can be either parallel or anti-parallel, a natural extension appeared to be the consideration of spins confined on a plane, each pointing to one of q equally spaced directions specified by the angles

$$\theta_n = 2\pi n/q, \qquad n = 0, \ldots, q-1. \tag{2.12}$$

In the most general form, the nearest-neighbour interaction $\mathbf{J} := \{J_{ij}\}_{ij}$ depends only on the relative angle beetwen the two spins $\theta_{ij} := \theta_{n_i} - \theta_{n_j}$, and the Hamiltonian of the system on a graph $\mathtt{G} = \mathrm{Graph}(\mathcal{V}; \mathcal{E})$ reads

$$H_{\mathtt{G}} = -\sum_{\langle ij \rangle} J(\theta_{ij}). \tag{2.13}$$

This is quite generally known as a system of $\mathbb{Z}(q)$ *symmetry* and it plays an important role in *lattice gauge theories*, (see e.g. [44] for an interesting review).

At first, Potts considered an interaction of the type $J(\theta) = -\epsilon \cos\theta$, but was unable to extend his results on the critical point of the system for $q > 4$. He then focused on another version of the q-state model, known as the *Ashkin-Teller model* in the case $q = 4$ [45], in which there are only two different interaction energies that correspond to nearest-neighbour spins being in the same or different states. This is the q-component model, that has attracted the most attention, and it is the one referred to as the *standard Potts model*, or simply the Potts model.

Given a graph $\mathtt{G} = \mathit{Graph}(\mathcal{V}; \mathcal{E})$ the Potts model Hamiltonian can be written as

$$H_{\mathtt{G}}[\boldsymbol{\sigma}; \mathbf{J}] = -\sum_{\substack{e \in \mathcal{E} \\ e=(i,j)}} J_e \delta(\sigma_i, \sigma_j), \tag{2.14}$$

where the sum runs over all the adjacent vertices, δ is the Kronecker delta and $\boldsymbol{\sigma} = \{\sigma_i\}_i$ denotes the spin configuration of the system. As usual, the interaction is said *ferromagnetic* or *antiferromagnetic* if the coupling J_e is either positive or negative. Finally, the partition function is simply given by the sum over all mappings $\boldsymbol{\sigma} : \mathcal{V} \mapsto \{1, \ldots, q\}$

$$Z_{\mathtt{G}}(q, \mathbf{J}) = \sum_{\boldsymbol{\sigma}} \prod_{\substack{e \in \mathcal{E} \\ e=(i,j)}} \mathrm{e}^{\beta J_e \delta(\sigma_i, \sigma_j)}. \tag{2.15}$$

Let us observe that this model possesses an $\mathcal{S}_q$ symmetry under the group of permutations of q elements, in opposition to the $\mathbb{Z}_2$ symmetry of the Ising model, which is recovered in the special case $q = 2$.

2.2.1 The Fortuin-Kasteleyn representation

The fact that the interaction in the Potts model is expressed through a delta function has strong implications on the combinatorial content of the model. In particular, one is allowed to perform a redefinition in which the parameter q has a natural analytic continuation, as we will show now. This redefinition is called *Fortuin-Kasteleyn representation* of the Potts model [46].

First of all, let us conveniently introduce the following couplings

$$v_e := \mathrm{e}^{\beta J_e} - 1, \tag{2.16}$$

so that the partition function (2.15) reduces to

$$Z_{\mathtt{G}}(q, \mathbf{v}) = \sum_{\boldsymbol{\sigma}} \prod_{\substack{e \in \mathcal{E} \\ e=(i,j)}} \left(1 + v_e \delta(\sigma_i, \sigma_j)\right). \tag{2.17}$$

The ferromagnetic region ($J_e > 0$) is mapped to $v_e > 0$, whereas the antiferromagnetic one ($J_e < 0$) is sent to the interval $-1 \leq v_e \leq 0$. The zero temperature limit is obtained by letting $v_e \to \infty$ for all $e \in \mathcal{E}$ in the former case, and $v_e \to -1$ in the latter. In both cases, the high temperature regime corresponds instead to the specific value $v_e = 0$. Let us note that for v_e values smaller than -1 the system falls into an unphysical region, because the Boltzmann weight is no more a positive quantity, as we expect for a statistical mechanical model.

Theorem 2.2.1 (Fortuin-Kasteleyn representation). *For each positive integer q, we have that*

$$Z_{\mathtt{G}}(q, \mathbf{v}) = \sum_{\mathtt{S} \in \mathscr{S}} q^{k(\mathtt{S})} \prod_{e \in \mathcal{E}_{\mathtt{S}}} v_e, \tag{2.18}$$

where $\mathscr{S}_{\mathtt{G}}$ is the set of all spanning subgraphs of $\mathtt{G}$ and $k(\mathtt{S})$ represents the number of their connected components, including isolated vertices.

Proof. Let us assume for simplicity that $v_e = v \ \ \forall e \in \mathcal{E}$. Looking at Eq. (2.15), each term in the sum is the product of $E = |\mathcal{E}|$ factors, one per edge $e = (i, j)$, that can be either 1 or $v\delta(\sigma_i, \sigma_j)$. The 2^E possible choices of factors are clearly in bijection with the power set $\mathcal{P}(\mathcal{E})$, and thus, as mentioned at the beginning of Sect. 1.1, with the set of spanning subgraphs of $\mathtt{G}$. This tells us that every spanning subgraph $\mathtt{S}$ takes the weight $v^{E_{\mathtt{S}}}$. Furthermore, each connected component is made of spins of the same color q, thanks to the effect of the Kronecker delta, so summing over allowed spin configurations gives a contribution $q^{k(\mathtt{S})}$. Putting all the elements together, the Potts partition function can be written as a sum over spanning subgraphs $\mathtt{S} = \mathrm{Graph}(\mathcal{V}; \mathcal{E}_{\mathtt{S}}) \subseteq \mathtt{G}$ with the following form

$$Z_{\mathtt{G}}(q, v) = \sum_{\mathtt{S} \in \mathscr{S}_{\mathtt{G}}} q^{k(\mathtt{S})} v^{E_{\mathtt{S}}} \tag{2.19}$$

It is quite simple to verify that the same reasoning can be repeated in the case of edge-dependent coupling constants v_e, leading to the desired expression (2.18). □

What we have just proved is that the Fortuin-Kasteleyn representation shows that the partition function $Z_{\mathtt{G}}(q, v)$ of the q-state Potts model on any finite graph $\mathtt{G}$ is in fact a *polynomial* in q and v. This allows us to interpret these two parameters as taking arbitrary real or even complex values.

It should be stressed, however, that the Potts spin model has a probabilistic interpretation, i.e. it has nonnegative weights, only when q is a positive integer and $v \geq -1$. Likewise, the Fortuin-Kasteleyn representation, which extends the Potts model to noninteger q, has a probabilistic interpretation only when $q \geq 0$ and $v \geq 0$. In all other cases, the model belongs to the "unphysical" regime with negative or complex weights, and the ordinary statistical mechanical properties need not to hold. For instance, the free energy needs not to possess the usual convexity properties, or phase transitions can occur even in one-dimensional systems with short-range interactions.

Let us observe that the obtained partition function can be rewritten in an useful way by means of Euler formula (1.4). Denoting by $L(\mathtt{S})$ the cyclomatic number of the spanning subgraph $\mathtt{S}$, i.e. the number of its independent cycles, we have

$$Z_{\mathtt{G}}(q, \mathbf{v}) = q^V \sum_{\mathtt{S} \in \mathscr{S}_{\mathtt{G}}} q^{L(\mathtt{S})} \prod_{e \in \mathcal{E}_{\mathtt{S}}} \frac{v_e}{q}. \tag{2.20}$$

Note that in the mathematical literature, formulas like (2.18) and (2.20) are usually written in terms of the Tutte polynomial $T(\mathtt{G}; x, y)$ defined in Eq. (2.9). In particular, it can be easily shown that the following relation holds

$$T(\mathtt{G}; x, y) = (x-1)^{-k(\mathtt{G})} (y-1)^{-|\mathcal{V}|} Z_{\mathtt{G}}((x-1)(y-1), y-1). \tag{2.21}$$

In other words, the Tutte polynomial $T(\mathtt{G}; x, y)$ and the Potts model partition function $Z_{\mathtt{G}}(q, v)$ are essentially equivalent under the change of variables

$$x = 1 + q/v, \qquad y = 1 + v \tag{2.22}$$

$$q = (x-1)(y-1), \qquad v = y - 1 \tag{2.23}$$

The advantage of the Tutte notation is that it allows a slightly smoother treatment of the $q \to 0$ limit. The disadvantage is that the use of the variables x and y conceals the fact that their particular combinations q and v play very different roles: q is a global variable, while v can be given separate values v_e on each edge.

2.2.2 The $q \to 0$ limit

Now that we have seen how to define the Potts model partition function $Z_{\mathtt{G}}(q, \mathbf{v})$ for arbitrary (even unphysical) values of q and $\mathbf{v}$, we can investigate the limit where these two parameters go to zero, keeping $\mathbf{w} = \mathbf{v}/q$ fixed [47]. As anticipated above, this regime is particularly relevant for us because it has an intriguing combinatorial interpretation, i.e. the partition function reduces to the *generating function* of spanning forest for the graph $\mathtt{G}$. Moreover, the limit we are about to consider acquires further importance in light of the recent discoveries [48] that (a) it can be mapped onto a fermionic theory containing a Gaussian term and a special four-fermion coupling, and (b) this latter theory is equivalent, to all orders in perturbation theory in $1/w$, to the N-vector model, i.e. $O(N)$-invariant σ-model, at $N = -1$ with $\beta = -w$, and in particular it is perturbatively asymptotically free in two dimensions, analogously to two-dimensional σ-models and four-dimensional nonabelian gauge theories.

We now proceed to consider the different ways in which a meaningful limit $q \to 0$ can be taken for the Fortuin-Kasteleyn representation of the Potts model partition function. Let us note that in what follows we will assume that the graph $\mathtt{G} = \text{Graph}(\mathcal{V}; \mathcal{E})$ on which the model is defined is in general not connected, but we will specify every time what happens in the particular case of a single component graph.

The first and simplest approach is to take $q \to 0$ at fixed couplings $\mathbf{v}$. From Eq. (2.18), we see that this selects out the subgraphs $\mathtt{S} \subseteq \mathtt{G}$ having the smallest possible number of connected components. The minimum achievable value is of course $k(\mathtt{G})$ ($= 1$ when the graph is connected), therefore we have

$$\lim_{q\to 0} q^{-k(\mathtt{G})} Z_{\mathtt{G}}(q, \mathbf{v}) = C_{\mathtt{G}}(\mathbf{v}), \tag{2.24}$$

where

$$C_{\mathtt{G}}(\mathbf{v}) = \sum_{\substack{\mathtt{S}\in\mathscr{S}_{\mathtt{G}} \\ k(\mathtt{S})=k(\mathtt{G})}} \prod_{e\in\mathcal{E}_{\mathtt{S}}} v_e \tag{2.25}$$

is the generating function of *maximally connected spanning subgraphs* (connected spanning subgraphs). Observe that all the generating functions we talk about are just polynomials until $\mathtt{G}$ is a finite graph, i.e. with $V = |\mathcal{V}| < \infty$.

A different limit can be obtained by taking $q \to 0$ with fixed values of $\mathbf{w} = \mathbf{v}/q$. Looking at the alternative form (2.20) of the partition function it is clear that only the subgraphs with the smallest possible cyclomatic number survives. The latter is of course 0, so we get

$$\lim_{q\to 0} q^{-V} Z_{\mathtt{G}}(q, q\mathbf{w}) = F_{\mathtt{G}}(\mathbf{w}), \tag{2.26}$$

with

$$F_{\mathtt{G}}(\mathbf{w}) = \sum_{\substack{\mathtt{S}\in\mathscr{S}_{\mathtt{G}} \\ L(\mathtt{S})=0}} \prod_{e\in\mathcal{E}_{\mathtt{S}}} w_e \tag{2.27}$$

generating polynomial of spanning forest (spanning trees).

Suppose now that in $C_{\mathtt{G}}(\mathbf{v})$ we replace each edge weight v_e by λv_e, and then we take the limit $\lambda \to 0$. This chooses, among all the maximally connected spanning subgraphs, those having the fewest edges, i.e. precisely the maximal spanning forests (spanning trees in case $\mathtt{G}$ is connected), with exactly $V - k(\mathtt{G})$ ($V - 1$ for a single component) edges. Hence

$$\lim_{\lambda\to 0} \lambda^{k(\mathtt{G})-V} C_{\mathtt{G}}(\lambda\mathbf{v}) = T_{\mathtt{G}}(\mathbf{v}), \tag{2.28}$$

where

$$T_{\mathtt{G}}(\mathbf{v}) = \sum_{\substack{\mathtt{S}\in\mathscr{S}_{\mathtt{G}} \\ k(\mathtt{S})=k(\mathtt{G}) \\ L(\mathtt{S})=0}} \prod_{e\in\mathcal{E}_{\mathtt{S}}} v_e \tag{2.29}$$

$$\begin{tikzcd}
 & C_G(\mathbf{v}) \arrow[dr, "\mathbf{v}\,\text{infinitesimal}"] & \\
Z_G(q,\mathbf{v}) \arrow[ur, "q \to 0,\ \mathbf{v}\,\text{fixed}"] \arrow[dr, "q \to 0,\ \mathbf{w} = \mathbf{v}/q\,\text{fixed}"'] & & T_G(\mathbf{v}\ \text{or}\ \mathbf{w}) \\
 & F_G(\mathbf{w}) \arrow[ur, "\mathbf{w}\,\text{infinite}"'] &
\end{tikzcd}$$

Figure 2.2: Schematic picture of the $q \to 0$ limit for the Fortuin-Kasteleyn representation of the Potts model.

is the generating function of *unrooted* spanning forests (trees). Note that the same result can be obtained from (2.27) by replacing w_e with λw_e and sending $\lambda \to \infty$. This selects the spanning forests with the greatest number of edges, i.e. once again the maximal spanning forests

$$\lim_{\lambda \to \infty} \lambda^{k(\mathtt{G})-V} F_{\mathtt{G}}(\lambda \mathbf{w}) = T_{\mathtt{G}}(\mathbf{w}). \tag{2.30}$$

Finally, maximal spanning forests (spanning trees) can also be obtained directly from $Z_{\mathtt{G}}(q,\mathbf{v})$ by a one-step process in which the limit $q \to 0$ is taken at fixed $\mathbf{x} = \mathbf{v}/q^\alpha$, where $0 < \alpha < 1$. Indeed, a simple manipulation of Eq. (2.18) with the Euler formula yields

$$Z_{\mathtt{G}}(q, q^\alpha \mathbf{x}) = q^{\alpha V} \sum_{\mathtt{S} \in \mathscr{S}} q^{\alpha L(\mathtt{S}) + (1-\alpha) k(\mathtt{S})} \prod_{e \in \mathcal{E}_{\mathtt{S}}} x_e. \tag{2.31}$$

The quantity $\alpha L(\mathtt{S}) + (1-\alpha)k(\mathtt{S})$ is minimized only on maximal spanning forests, where it takes the value $(1-\alpha)k(\mathtt{G})$, hence

$$\lim_{q \to 0} q^{-\alpha V - (1-\alpha)k(\mathtt{G})} Z_{\mathtt{G}}(q, q^\alpha \mathbf{x}) = T_{\mathtt{G}}(\mathbf{x}). \tag{2.32}$$

In summary, the polynomials that we have obtained are all related one another, as it can be seen pictorially in Fig. 2.2.

We conclude this section by noting that according to the Yang-Lee picture of phase transitions [49], information about their possible loci can be obtained by investigating the zeros of the partition function for finite subsets of the lattice $\mathcal{L}$ on which the system is defined, when one or more physical parameters (e.g. the temperature or the magnetic field) are allowed to take *complex values*. The accumulation points of these zeros in the thermodynamic limit constitute the phase boundaries. For the Potts model, therefore, by studying the zeros of $Z_{\mathtt{G}}(q,v)$ in complex (q,v)-space in the infinite-volume limit, we can learn about its phase diagram even in the real (q,v)-plane. Since those computations are usually quite involved, it is very convenient to study first "slices" of the complex space such the one provided above.

The polynomials that we have introduced are thus of great interest not only from the combinatorial point of view, but also in the study of the physical properties of the Potts model itself. Actually, in the specific case of the spanning trees polynomial (2.29) a number of other interesting physical results exists. For instance, the number of stable configurations that occur with nonzero probability in the steady state of the *abelian sandpile model* [50, 51] on a graph $\mathtt{G}$ equals the number of spanning trees on $\mathtt{G}'$, the graph obtained from $\mathtt{G}$ by connecting one extra site. Furthermore, the spanning tree polynomial is closely related with the analysis of electrical circuits, as we will point out in the following section to introduce the matrix-tree theorem.

2.3 Matrix-tree theorem and Temperley formula

Let $\mathtt{G} = \text{Graph}(\mathcal{V}; \mathcal{E})$ be a connected graph with edge weights $\mathbf{w} = \{w_{ij}\}_{(i,j)\in\mathcal{E}}$, then its Laplacian matrix, defined in Sect. 1.1, is written as

$$(\mathbf{L}_{\mathtt{G}}(\mathbf{w}))_{ij} := \begin{cases} -w_{ij} & \text{if } i \neq j \text{ and } (i,j) \in \mathcal{E} \\ 0 & \text{if } i \neq j \text{ and } (i,j) \notin \mathcal{E} \\ \sum_{k\neq i} w_{ik} & \text{if } i = j \end{cases} \tag{2.33}$$

As we already mentioned, in the case the graph $\mathtt{G}$ is undirected $\mathbf{L}_{\mathtt{G}}$ is symmetric. Moreover it has zero determinant by construction, since each row (and column) elements sum to zero.

If we now consider the graph as an electrical network, the weights w_e represent the *conductance*, while their inverse $1/w_e$ are the *resistance*. Supposing we inject currents $\mathbf{J} = \{J_i\}_{i\in\mathcal{V}}$ into the vertices, one can ask what node voltages $\phi = \{\phi\}_{i\in\mathcal{V}}$ will be produced. By applying Kirchhoff's law of current conservation at each vertex and Ohm's law on each edge, it is quite easy to see that the node voltages and current inflows satisfy the linear system

$$\mathbf{L}_{\mathtt{G}}(\mathbf{w}) \cdot \boldsymbol{\phi} = \mathbf{J}. \tag{2.34}$$

If we further impose that $\sum_{i\in\mathcal{V}} J_i = 0$, namely total current conservation, and fix an arbitrary node $i_0 \in \mathcal{V}$ to zero voltage (the "ground", because only voltage differences are physically observable), the linear system becomes

$$\mathbf{L}_{\mathtt{G}}(\mathbf{w})_{\setminus i_0} \cdot \boldsymbol{\phi}_{\setminus i_0} = \mathbf{J}_{\setminus i_0}, \tag{2.35}$$

where $\mathbf{L}_{\mathtt{G}}(\mathbf{w})_{\setminus i_0}$ denotes the matrix obtained from $\mathbf{L}_{\mathtt{G}}(\mathbf{w})$ by deleting the i_0-th row and column. A natural question is then to ask under which conditions does this system have a (unique) solution, i.e. when the determinant of $\mathbf{L}_{\mathtt{G}}(\mathbf{w})$ is nonzero. In 1847 [3] Kirchhoff answered this question with the following striking result.

Theorem 2.3.1 (matrix-tree theorem). *The determinant of* $\mathbf{L}_{\mathtt{G}}(\mathbf{w})_{\setminus i}$ *is independent of* i *and equals the spanning trees generating polynomial* $T_{\mathtt{G}}(\mathbf{w})$ *(see Eq. (2.29)) for the graph* $\mathtt{G}$

$$\det \mathbf{L}_{\mathtt{G}}(\mathbf{w})_{\setminus i} = T_{\mathtt{G}}(\mathbf{w}). \tag{2.36}$$

Many different proofs of the matrix-tree theorem are now available in the literature, e.g. a simple one based on Cauchy-Binet theorem of matrix theory can be found in [52].

The above result is extremely useful for our purposes, because it provides a way to represent in a very useful form the spanning trees generating function, which will represent the starting point of our replica calculation for the MST problem. What remains to be done is to express Eq. (2.36) in a still simpler fashion, namely eliminating the constraint on the deletion of the i-th row and column. In the meantime, as we have promised at the end of Sect. 1.1.1, we will also provide a proof of Cayley's formula (see Theorem 1.1.4), which counts the number of spanning trees in a complete graph $\mathtt{K}_N$.

Let us start by representing the determinant in Eq. (2.36) in terms of a *Berezin integral* on the couple of *Grassmann variables* $\{\psi_i\}_{i\in\mathcal{V}}$, $\{\bar{\psi}_i\}_{i\in\mathcal{V}}$ that we introduce on each vertex of the graph $\mathtt{G}$ (see Appendix A for an introduction on the Grassmann algebra). Using Eq. (2.29) for the generating polynomial of spanning trees one can write

$$\tau = \sum_{T\in\mathcal{T}} \prod_{e\in T} w_e = \det \mathbf{L}_{\setminus i} = \int \mathcal{D}(\psi,\bar{\psi})\, \bar{\psi}_i \psi_i \, \mathrm{e}^{\left(\bar{\psi},\mathbf{L}\,\psi\right)}, \tag{2.37}$$

where the sum runs over the set of all spanning trees $\mathcal{T}$ in the graph $\mathtt{G}$, and with the shorthands $\mathcal{D}\left(\psi,\bar{\psi}\right) := \prod_i \mathrm{d}\psi_i \mathrm{d}\bar{\psi}_i$ for the integration measure and

$$\left(\bar{\psi},\mathbf{L}\,\psi\right) := \sum_{i,j} \bar{\psi}_i L_{ij} \psi_j . \tag{2.38}$$

for the scalar product in the exponent. As we have already mentioned, the determinant of $\mathbf{L}$ is zero by construction, so by developping it along the k-th row we have

$$\det \mathbf{L} = \sum_l L_{kl} (\operatorname{adj} \mathbf{L})_{lk} = \sum_l L_{kl} \int \mathcal{D}\left(\psi,\bar{\psi}\right) \bar{\psi}_l \psi_k \, \mathrm{e}^{\left(\bar{\psi},\mathbf{L}\,\psi\right)} = 0 \tag{2.39}$$

with the *matrix adjugate* to $\mathbf{L}$ (i.e. the transpose of the cofactor), defined by

$$(\operatorname{adj} \mathbf{L})_{ij} = (-1)^{i+j} \det \mathbf{L}_{\setminus j,i}, \tag{2.40}$$

From Eq. (2.39) it is evident that $\operatorname{adj}\mathbf{L}$ has to be proportional to the projector on the eigenstate with zero eigenvalue of the matrix $\mathbf{L}$, i.e. to the operator $\mathbf{\Pi}$ such that

$$\mathbf{L}\,\mathbf{\Pi} = \mathbf{\Pi}\,\mathbf{L} = 0. \tag{2.41}$$

This corresponds to $\mathbf{\Pi} = \mathbf{J}/N$, with $N = |\mathcal{V}|$ and $J_{ij} = 1 \ \forall i,j$, as can be checked easily given the structure of the Laplacian matrix. The proportionality constant follows again from Eq. (2.39) by decomposing the sum into diagonal and off-diagonal parts, and then using the matrix-tree theorem on the first term

$$\sum_l L_{kl} (\operatorname{adj} \mathbf{L})_{lk} = \tau L_{kk} + \sum_{l\neq k} L_{kl} (\operatorname{adj} \mathbf{L})_{lk}. \tag{2.42}$$

What one obtains is therefore

$$\operatorname{adj}\mathbf{L} = \tau\,\mathbf{J} = \tau N\,\mathbf{\Pi} \tag{2.43}$$

or, in other words, a generalization of the matrix-tree theorem, saying that it does not matter what specific row and column we remove from the Laplacian matrix, due to the fact that

$$\tau = (\operatorname{adj}\mathbf{L})_{ii} = (\operatorname{adj}\mathbf{L})_{jj} = (\operatorname{adj}\mathbf{L})_{ij}. \tag{2.44}$$

Introducing now the orthogonal projector $\mathbf{\Pi}^{\perp} = 1 - \mathbf{J}/N$, thanks to the last relation one can directly check that

$$\int \mathcal{D}\left(\psi,\bar{\psi}\right)\left(\bar{\psi},\mathbf{\Pi}^{\perp}\psi\right)\mathrm{e}^{\left(\bar{\psi},\mathbf{L}\,\psi\right)} = 0, \tag{2.45}$$

thus we can write the following chain of equalities

$$\begin{aligned}
\tau &= \frac{1}{N}\int \mathcal{D}\left(\psi,\bar{\psi}\right)\left(\bar{\psi},\psi\right)\mathrm{e}^{\left(\bar{\psi},\mathbf{L}\,\psi\right)} \\
&= \frac{1}{N}\int \mathcal{D}\left(\psi,\bar{\psi}\right)\left(\bar{\psi},\mathbf{\Pi}\,\psi\right)\mathrm{e}^{\left(\bar{\psi},\mathbf{L}\,\psi\right)} \\
&= \frac{1}{\lambda N}\int \mathcal{D}\left(\psi,\bar{\psi}\right)\mathrm{e}^{\left(\bar{\psi},\mathbf{L}\,\psi\right)+\lambda\left(\bar{\psi},\mathbf{\Pi}\,\psi\right)} \\
&= \frac{1}{\lambda N}\det\left(\mathbf{L}+\lambda\mathbf{\Pi}\right).
\end{aligned} \tag{2.46}$$

In the first passage we simply exploited the independence of τ from the indices i, j, and we used Eq. (2.45) in the second. The third equality follows instead from the fact that the determinant of the Laplacian matrix vanishes, as it happens to higher order in the expansion of λ, thanks to the nilpotency of Grassmann variables.

Note that in the last equation the zero eigenvalue of the Laplacian is substituted by the arbitrary constant λ, which is removed by the denominator. By choosing in particular $\lambda = N$ we finally arrive at the desired result, which is the so called *Temperley's formula* [53]

$$\tau = \frac{1}{N^2}\det\left(\mathbf{L}+\mathbf{J}\right). \tag{2.47}$$

To directly check the validity of the obtained result let us consider a complete graph $\mathtt{K}_N$ with all weights fixed to one. In this case, the generating polynomial $T_{\mathtt{G}}(\mathbf{w})$ coincides with the number of different spanning trees that can be drawn on the graph, and with our formula we find

$$\tau(\mathtt{K}_N) = \frac{1}{N^2}\det\left(\mathbf{L}+\mathbf{J}\right) = \frac{1}{N^2}\det\left(N\,\mathbb{I}\right) = N^{N-2} \tag{2.48}$$

which is exactly Cayley's result (1.7).

2.4 Random MST via replicas

In the previous sections we have provided all the building blocks necessary to set up the final part of our analysis of the random MST problem with i.i.d. edge weights. In particular, we have seen in Sect. 2.2.2 how to obtain the partition function for our system, namely the spanning tree generating polynomial (2.29), as a $q \to 0$ limit of the q-state Potts model. Then we have expressed it as the determinant of the graph Laplacian through the matrix-tree theorem, which allows us to state the problem in terms of a Berezin integral defined on Grassmann variables.

Now we want to proceed with the established path, i.e. with an attempt in the calculation of the average optimal cost of the MST with the replica method, as already performed for other combinatorial optimization problems [25, 26]. Before we start, however, it is important to mention that at least another promising route exists that one can consider to obtain our desired result. In fact, one could think to carry out a replica calculation directly on the q-state Potts model, and only in a second time to consider the limit $q \to 0$ to investigate the case of spanning trees and forests.

In this case, a good starting point would be to consider the work of Elderfield and Sherrington [54], even if a crucial difference has to be taken into account. The two authors applied the replica method to the Potts model by choosing couplings of the form given in Eq. (1.25), i.e. Gaussian random variables, thus finding that only the first two moments of their distribution matter when one averages over the disorder, exactly as it happens in the SK model (see Sect. 1.3.2). Instead, in our case we have to redefine the couplings according to Eq. (2.16), and so we expect that all the moments of the distribution will play a role in the calculations, making them much more involved.

From now on we will restrict ourselves to the case of the complete graph $\mathtt{K}_N$ with N vertices. According to Temperley's formula (2.46) with the arbitrary constant fixed to $\lambda = 1$, the partition function for the weighted spanning trees on the given graph can be written as

$$Z := \sum_{T \in \mathcal{T}} \prod_{e \in T} w_e = \frac{1}{N} \det(\mathbf{L} + \frac{\mathbf{J}}{N}). \tag{2.49}$$

To select the minimum spanning tree $\mathtt{T}_0$ in the low temperature limit $\beta \to \infty$, we assume that the edge weights are of the form

$$w_e := \mathrm{e}^{-\beta N l_e} \tag{2.50}$$

In this case, the cost of the optimal configuration for our system will properly be given by

$$C = -\lim_{\beta \to \infty} \log Z(\beta) = \min_{T \in \mathcal{T}} \sum_{e \in T} l_e = \sum_{e \in T_0} l_e. \tag{2.51}$$

Clearly, as we are considering a random optimization problem, l_e are positive random variables generated according to a probability distribution density $\rho(l)$, which we take for definiteness as the simple exponential one

$$\rho(l) = \theta(l)\mathrm{e}^{-l}. \tag{2.52}$$

We now introduce the Grassmann variables representation (A.11) for the determinant appearing in Eq. (2.49), and according to the prescription of the replica method (see Sect. 1.3.2) we take the n-th power of the partition function, obtaining

$$Z^n = \frac{1}{N^n}\int \mathcal{D}\left(\psi,\bar{\psi}\right)\prod_{a=1}^{n}\exp\left(\sum_{i,j=1}^{N}\left[\frac{1}{2}\left(\bar{\psi}_i^a-\bar{\psi}_j^a\right)w_{ij}\left(\psi_i^a-\psi_j^a\right)+\frac{1}{N}\bar{\psi}_i^a\psi_j^a\right]\right) \tag{2.53}$$

where we have written explicitly the action of the matrices $\mathbf{L}$ and $\mathbf{J}$ on the anticommuting variables ψ, $\bar{\psi}$.

The next step of our recipe consists in performing the average over the disorder, namely the configurational average $\overline{Z^n}$. For this purpose, we exploit the identity

$$\int \mathrm{d}w\rho(w)\mathrm{e}^{xw} = \exp\left(\sum_{k=1}^{\infty}\phi_k\frac{x^k}{k!}\right), \tag{2.54}$$

where ϕ_k are the *cumulants* of the distribution $\rho(w)$. More precisely, being the form of the weights the one given in Eq. (2.50), we have to perform the following integral

$$\int \mathrm{d}l\rho(l)\mathrm{e}^{x\mathrm{e}^{-\beta N l}} = \sum_{k=0}^{\infty}\frac{x^k}{k!}\int \mathrm{d}l\rho(l)\mathrm{e}^{-\beta N k l} \equiv 1+\frac{1}{N}\sum_{k=1}^{\infty}g_k\frac{x^k}{k!}. \tag{2.55}$$

where we have introduced, as in the case of the assignment problem [55], the moments g_k/N of the distribution function ρ. Expanding the last equation in series of N one recovers Eq. (2.54), so the following relations between the cumulants and the moments of the distribution hold

$$\phi_1 = \frac{g_1}{N}, \qquad \phi_2 = \frac{g_2}{N}-\left(\frac{g_1}{N}\right)^2, \qquad \dots \tag{2.56}$$

Remembering that we are interested in the thermodynamic limit $N \to \infty$, it is evident that at the leading order in N the k-th cumulant and the k-th moment coincides. Therefore, the averaged partition function w.r.t. the edge weights distribution is

$$\begin{aligned}\overline{Z^n} =& \frac{1}{N^n}\int \mathcal{D}\left(\psi,\bar{\psi}\right)\exp\left(\frac{1}{N}\sum_{a}\sum_{i,j}\bar{\psi}_i^a\psi_j^a+\right.\\ &\left.+\frac{1}{N}\sum_{k}\frac{g_k}{2^k k!}\sum_{i,j}\left[\sum_{a}\left(\bar{\psi}_i^a-\bar{\psi}_j^a\right)\left(\psi_i^a-\psi_j^a\right)\right]^k\right).\end{aligned} \tag{2.57}$$

Let us remark that thanks to the nilpotency of the Grassmann variables, by expanding the power in the exponent when two indices are equal the product vanishes, so one has

$$\begin{aligned}&\frac{1}{2^k k!}\sum_{i,j=1}^{N}\sum_{a_1,\dots,a_k=1}^{n}\left(\bar{\psi}_i^{a_1}-\bar{\psi}_j^{a_1}\right)\left(\psi_i^{a_1}-\psi_j^{a_1}\right)\cdots\left(\bar{\psi}_i^{a_k}-\bar{\psi}_j^{a_k}\right)\left(\psi_i^{a_k}-\psi_j^{a_k}\right)=\\ &\sum_{1\le i<j\le N}\sum_{1\le a_1<\dots<a_k\le n}\left(\bar{\psi}_i^{a_1}-\bar{\psi}_j^{a_1}\right)\left(\psi_i^{a_1}-\psi_j^{a_1}\right)\cdots\left(\bar{\psi}_i^{a_k}-\bar{\psi}_j^{a_k}\right)\left(\psi_i^{a_k}-\psi_j^{a_k}\right)\end{aligned} \tag{2.58}$$

2.4.1 Sites decoupling

Having in mind what we have done in Sect. 1.3.2 for the Sherrington-Kirkpatrick model, normally the next step in the analysis of the replicated partition function is to transform the part of the integrand of $\overline{Z^n}$ that involves sums on pairs of sites into a form involving only sums over single sites. Having done that, the partition function may be written as an integral where the number of sites N appears just as a parameter.

Unfortunately, in our case this task is not straightforward due to the particular structure of the Laplacian matrix, resulting in the complicated form of the exponent in Eq. (2.57). A similar problem has been already treated in the context of disordered diffusion [56], and faced by the authors with a generalized *functional Hubbard-Stratonovich transformation*, necessary to introduce the auxiliary variables one needs to decouple different sites. This technique has been used in subsequent years in the study of random matrices too, so we refer the reader e.g. to the work of Fyodorov [57], in which one can find an example of the procedure together with a heuristic discussion on its validity. Here we will simply sketch the reasoning, highlighting only the fundamental steps.

The term we want to decouple is of the form

$$\mathrm{e}^{\frac{1}{2N}\sum_{i,j}\mathscr{L}\left(\bar{\psi}_i,\psi_i,\bar{\psi}_j,\psi_j\right)} \tag{2.59}$$

where we have defined the operator

$$\mathscr{L}\left(\bar{\psi}_i,\psi_i,\bar{\psi}_j,\psi_j\right) \equiv 2\sum_a \bar{\psi}_i^a\psi_j^a + \sum_{k=1}^{\infty}\frac{g_k}{2^{k-1}k!}\sum_{a_1,\dots,a_k}\prod_{s=1}^{k}\left(\bar{\psi}_i^{a_s}-\bar{\psi}_j^{a_s}\right)\left(\psi_i^{a_s}-\psi_j^{a_s}\right). \tag{2.60}$$

Note that for simplicity we have dropped the dependence on the replica indices into the arguments of the operator $\mathscr{L}$. Suppose now it is possible to diagonalize the operator, and that the basis over which it is diagonal is spanned by the functions $e_\mu\left(\bar{\psi}_i,\psi_i\right)$ (again, $\bar{\psi}_i$ and ψ_i depend on all replica indices), which are orthonormal with respect to the measure $\mathcal{D}\left(\psi,\bar{\psi}\right)$ in the replica space

$$\int \mathcal{D}\left(\psi,\bar{\psi}\right) e_\mu\left(\bar{\psi}_i,\psi_i\right) e_\nu\left(\bar{\psi}_i,\psi_i\right) = \delta_{\mu\nu}. \tag{2.61}$$

Therefore we have

$$\mathscr{L}\left(\bar{\psi}_i,\psi_i,\bar{\psi}_j,\psi_j\right) = \sum_\mu \lambda_\mu e_\mu\left(\bar{\psi}_i,\psi_i\right) e_\mu\left(\bar{\psi}_j,\psi_j\right) \tag{2.62}$$

so the expression (2.59) can be manipulated as follows

$$\begin{aligned}\mathrm{e}^{\frac{1}{2N}\sum_{i,j}\mathscr{L}\left(\bar{\psi}_i,\psi_i,\bar{\psi}_j,\psi_j\right)} &= \exp\left(\frac{1}{2N}\sum_\mu \lambda_\mu\left[\sum_i e_\mu\left(\bar{\psi}_i,\psi_i\right)\right]^2\right)\\ &= \int\left(\prod_\mu \frac{\mathrm{d}y_\mu}{\sqrt{2\pi}}\right)\exp\left(-\frac{N}{2}\sum_\mu y_\mu^2 + \sum_\mu\sqrt{\lambda_\mu}\sum_i e_\mu\left(\bar{\psi}_i,\psi_i\right)\right).\end{aligned} \tag{2.63}$$

By defining the function

$$G(\bar{\psi},\psi) := \sum_{\mu} y_\mu \sqrt{\lambda_\mu} e_\mu \left(\bar{\psi},\psi\right) \tag{2.64}$$

together with the inverse operator

$$\mathscr{L}^{-1}\left(\bar{\psi}_i,\psi_i,\bar{\psi}_j,\psi_j\right) := \mathscr{F}\left(\bar{\psi}_i,\psi_i,\bar{\psi}_j,\psi_j\right) = \sum_{\mu} \frac{1}{\lambda_\mu} e_\mu\left(\bar{\psi}_i,\psi_i\right) e_\mu\left(\bar{\psi}_j,\psi_j\right), \tag{2.65}$$

one can easily prove by exploiting the orthogonality condition (2.61) that the following equality holds

$$\int \mathcal{D}\left(\psi,\bar{\psi}\right)\mathcal{D}\left(\eta,\bar{\eta}\right) G\left(\bar{\psi},\psi\right) \mathscr{F}\left(\bar{\psi},\psi,\bar{\eta},\eta\right) G\left(\bar{\eta},\eta\right) = \sum_{\mu} y_\mu^2. \tag{2.66}$$

We have thus obtained the generalized Hubbard-Stratonovich transformation we were looking for, which reads

$$\mathrm{e}^{\frac{1}{2N}\sum_{i,j}\mathscr{L}\left(\bar{\psi}_i,\psi_i,\bar{\psi}_j,\psi_j\right)} = \int \mathcal{D}G\, \mathrm{e}^{-\frac{N}{2}\int \mathcal{D}\left(\psi,\bar{\psi}\right)\mathcal{D}(\eta,\bar{\eta}) G\left(\bar{\psi},\psi\right)\mathscr{F}\left(\bar{\psi},\psi,\bar{\eta},\eta\right)G(\bar{\eta},\eta)+\sum_i G\left(\bar{\psi}_i,\psi_i\right)} \tag{2.67}$$

with $\mathcal{D}G$ representing a suitable integration measure. Therefore, we have shown that proceeding this way one finally arrives to the desired form for the configurational average of the replicated partition function

$$\overline{Z^n} = \int \mathcal{D}G\, \mathrm{e}^{-NS[G]}, \tag{2.68}$$

where we have defined the action

$$S[G] \equiv \frac{1}{2}\int \mathcal{D}\left(\psi,\bar{\psi}\right)\mathcal{D}\left(\eta,\bar{\eta}\right) G\left(\bar{\psi},\psi\right) \mathscr{F}\left(\bar{\psi},\psi,\bar{\eta},\eta\right) G\left(\bar{\eta},\eta\right) - \log z[G], \tag{2.69}$$

with

$$z[G] \equiv \int \mathcal{D}\left(\psi,\bar{\psi}\right) \mathrm{e}^{G\left(\bar{\psi},\psi\right)}. \tag{2.70}$$

From these expressions one can perform a saddle point analysis in the thermodynamic limit $N \to \infty$.

Clearly, here we have performed a series of formal manipulations, and the main question is whether it is possible to explicitly diagonalize the operator $\mathscr{L}$ for our specific problem. The integral representation we have found is rather puzzling at first sight, still the method is very useful and led to correct results in various cases [58]. In particular, provided some kind of regularization of the problem is allowed, the functional Hubbard-Stratonovich transformation is correct and identical to the set of Hubbard-Stratonovich identities.

Chapter 3

The Euclidean MST

In the last chapter of this thesis we focus on the so called random Euclidean MST problem, in which correlations between edge weights are present due to the fact that the N vertices of the graph are random points scattered in a Euclidean domain. After an introduction to the subject, we provide for the first time the solution of the problem defined on a bipartite graph in one dimension. Moreover, we perform a numerical investigation in one and two dimensions to study the scaling for large N of the average cost of the random MST. Both of this elements show that the random fluctuations in the positions of the points do not influence the scaling behavior of the average cost in the bipartite setting, which remains identical to the one of the monopartite case, a fact that represents the main contribution of our work to the topic.

3.1 Random Euclidean optimization problems

In the previous chapter, in particular in the first and last sections, we focused on the random minimum spanning tree problem defined on a graph with independent and identically distributed weights. In practice, we excluded the possibility of the presence of *correlations* between different edges, and so we considered what can be thought of as a *mean field* approximation for the case we will examine here. In fact, some optimization problems are actually defined in the geometric space, think e.g. to the TSP to be solved in a certain geographical area, or to the several applications of the MST problem cited in Sect. (1.2.2), so let us introduce the class of so called *Euclidean optimization problems*.

This kind of problem is still defined on a graph $\mathtt{G} = \mathrm{Graph}(\mathcal{V};\mathcal{E})$, but a connected convex domain $\Omega \subset \mathbb{R}^d$ is also given, together with a random point process $\boldsymbol{\Phi}$ which provides the embedding of $\mathtt{G}$ in the Euclidean space,

$$\boldsymbol{\Phi} : \mathcal{V} \to \Omega \text{ such that } v_i \in \mathcal{V} \mapsto \boldsymbol{\Phi}(v_i) \equiv \boldsymbol{\xi}_i \in \Omega. \tag{3.1}$$

In the remaining part of this thesis, we will consider a weight function $w : \mathcal{E} \to \mathbb{R}^+$ defined as

$$w(e_{ij}) \equiv w_{ij}^{(p)} := \|\boldsymbol{\xi}_i - \boldsymbol{\xi}_j\|^p , \quad p \in \mathbb{R}^+ \tag{3.2}$$

so that the weight associated to the edge $e_{ij} \equiv (v_i, v_j)$ is a positive power of the Euclidean distance between the points. Note that apart from this, no other differences emerge with respect to the combinatorial optimization problems on graphs introduced in the first chapter.

Clearly, the Euclidean origin of the weights in the graph is not relevant for the solution of a given instance of the problem, and the algorithms available in the literature work perfectly. However, the study in presence of randomness is more complicated. In the context of random Euclidean optimization problems, randomness is typically introduced in the embedding process of the graph in Ω. In this case, a probability distribution density is given on Ω,

$$\rho : \Omega \to \mathbb{R}^+, \qquad \int_\Omega \rho(\mathbf{x})\, \mathrm{d}^d x = 1. \tag{3.3}$$

and we suppose in practice that $N = |\mathcal{V}|$ points, $\mathscr{X} := \{\boldsymbol{\xi}_i\}_{i=1,\dots,N} \in \Omega$, are independently randomly generated on Ω. Therefore, we associate to each vertex $v_i \in \mathcal{V}$ of the graph a point $\boldsymbol{\xi}_i$ at random.

Doing so, the weights $w(e)$ defined according to Eq. (3.2) are random, but in general *Euclidean correlations* appear, due to the correlations among the distances of the points, such as the one imposed by the triangle inequality. Considering all this, the average procedure becomes more subtle than in the purely uncorrelated case, and great importance is given to the point generation procedure. In this regard, one usually defines the *empirical measure*

$$\rho_{\mathscr{X}}(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} \delta^{(d)}(\mathbf{x} - \boldsymbol{\xi}_i), \tag{3.4}$$

which can be proven to converge to ρ almost surely for $N \to \infty$.

Let us now fix some notations and terminology considering the MST problem defined in a Euclidean domain $\Omega \subset \mathbb{R}^d$, having in mind that everything can be readily adapted to other combinatorial optimization problems, such as the TSP or the matching problem.

In the *random Euclidean monopartite MST problem* (rEm) the problem is defined on the complete graph $\mathtt{K}_N$, with each vertex $v_i \in \mathcal{V}$ associated to a point $\boldsymbol{\xi}_i \in \Omega$ randomly generated according to a given probability density function ρ, as in Eq. (3.3). We are interested in the spanning tree $\mathtt{T}_0 \subseteq \mathtt{K}_N$ that minimizes the functional

$$\mathcal{C}^{(p,\mathrm{Em})}[\mathtt{T}] := \sum_{e \in \mathcal{E}_\mathtt{T}} w^{(p)}(e), \qquad \mathtt{T} = \mathrm{Graph}(\mathcal{V}_\mathtt{T}; \mathcal{E}_\mathtt{T}) \in \mathcal{T}, \tag{3.5}$$

where $w^{(p)}(e)$ is defined in Eq. (3.2) and

$$\mathcal{T} := \{\mathtt{T} \mid \mathtt{T} \subset \mathtt{K}_N \text{ spanning tree}\}. \tag{3.6}$$

In what follows we will denote the MST cost by

$$\mathcal{C}_N^{(p,\mathrm{Em})} := \mathcal{C}^{(p,\mathrm{Em})}[\mathtt{T}_0] = \min_{\mathtt{T} \in \mathcal{T}} \mathcal{C}^{(p,\mathrm{Em})}[\mathtt{T}], \tag{3.7}$$

while its average will be indicated by

$$\mathcal{C}_{N,d}^{(p,\mathrm{rEm})} := \overline{\mathcal{C}_N^{(p,\mathrm{Em})}}. \tag{3.8}$$

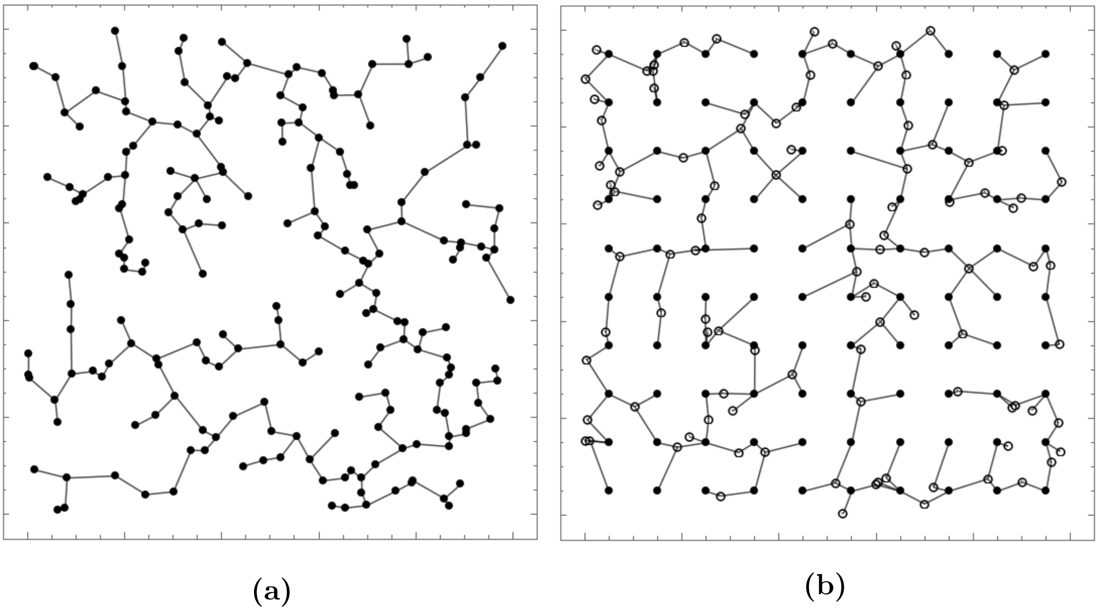


Figure 3.1: Random Euclidean monopartite MST (a) with $N = 200$ and random Euclidean bipartite (grid-Poisson) MST (b) with $N = 100$. In both cases the random points are generated with uniform distribution on $[0, 1]^2$ and open boundary conditions are assumed.

Clearly the average $\overline{\bullet}$ is performed on the positions of the points, and we note that the previous quantity strongly depends on the considered domain Ω, on the number of points N and on their distribution ρ.

If the problem is defined on a complete bipartite graph $\mathtt{K}_{N,N} = \text{Graph}(\mathcal{V},\mathcal{U};\mathcal{E})$ we call it *random Euclidean bipartite MST problem* (rEb), but two different possibilities exist to introduce randomness. Considering the flat distribution $\rho(\mathbf{x}) = \frac{1}{|\Omega|}$, if both vertices in $\mathcal{V}$ and $\mathcal{U}$ are associated to points in Ω that are randomly and independently generated according to ρ, we will call the problem *Poisson-Poisson Euclidean MST problem* (RR-Eb). Whereas if one set of vertices, say $\mathcal{U}$, is mapped to a fixed hypercube lattice on Ω we will refer to the problem as the *grid-Poisson Euclidean MST problem* (GR-rEb). Naturally, all the above definitions for the (rEm) continue to hold (with some obvious modifications), but this time two different averages on the positions of the points can be performed, namely

$$\mathcal{C}_{N,d}^{(p,RR)} := \overline{\mathcal{C}_{N}^{(p,Eb)}} \tag{3.9a}$$

$$\mathcal{C}_{N,d}^{(p,GR)} := \overline{\mathcal{C}_{N}^{(p,Eb)}}\Big|_{\mathcal{U}\text{ on the grid}} . \tag{3.9b}$$

In the first case we average over both the sets of points, in the latter one set is instead supposed fixed.

3.2 Average optimal cost scaling for the rEm

As we did in Sect. 2.1 for the purely uncorrelated case, we now want to review the existing literature on the MST problem when defined in a Euclidean setting.

Before we start to analyze it in detail, let us note that the asymptotic optimal cost of all the classical combinatorial optimization problems has been studied extensively in the hypothesis that the points $\{\boldsymbol{\xi}_i\}_i$ are randomly generated on $\Omega \subset \mathbb{R}^d$ [59]. For instance, using the fact that the Euclidean functionals of the TSP (1.12) and the matching problem (1.13) are *homogeneous* and *translationally invariant*, i.e.

$$\mathcal{C}[\mathtt{S}] \xrightarrow{\boldsymbol{\xi}_i \mapsto \lambda\boldsymbol{\xi}_i + \mathbf{r}} \lambda^p \mathcal{C}[\mathtt{S}], \qquad \lambda \in \mathbb{R}^+, \quad \mathbf{r} \in \mathbb{R}^d, \tag{3.10}$$

with $\mathtt{S}$ proper subgraph depending on the problem considered, Redmond and Yukich [60] proved that on the complete graph $\mathtt{K}_{2N}$, embedded in the unit hypercube in d dimensions, their optimal cost scales as $N^{1-\frac{p}{d}}$ for $0 < p < d$ in the large N limit.

An analogous result for the random Euclidean MST functional of Eq. (3.7) was obtained by Steele in 1988 [61], and it can be formulated in the following form.

Theorem 3.2.1. *Consider a complete graph $\mathtt{G} = Graph(\mathcal{V};\mathcal{E})$ with vertex set given by $\mathcal{V} = \{\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_N\}$, where $\boldsymbol{\xi}_i$ are independent random variables with distribution F having compact support in $\mathbb{R}^d$, $d \geq 2$. If the monotone weighting function $w = w(\|e\|)$, with $\|e\|$ denoting the Euclidean length of the edge $e \in \mathcal{E}$, satisfies $w(x) \sim x^p$ for some $0 < p < d$, then with probability 1*

$$\lim_{N\to\infty} N^{-1+\frac{p}{d}} \, \mathcal{C}_N^{(p,\mathrm{Em})} = c(p,d) \int_{\mathbb{R}^d} f(\mathbf{x})^{1-\frac{p}{d}} \, \mathrm{d}^d x. \tag{3.11}$$

Here f denotes the density of the absolutely continuous part of F and $c(p,d)$ is a stricly positive constant.

First of all, let us note that this theorem lacks the case $d = 1$ simply because it is trivial. In fact, the Euclidean MST for a set of random points scattered on a line is readily constructed by joining them in order from one end to the other. This will not be true anymore in the case of a bipartite graph, in which points in the same class cannot connect, as we will see explicitly in the next section.

Secondly, it is worth observing that the above result is closely related to the general theory of subadditive Euclidean functionals [62], but there are some crucial differences. One issue is that the cost of a monopartite Euclidean MST is not an almost surely increasing sequence of random variables, i.e. it is not true that $\mathcal{C}(\boldsymbol{\xi} \cup A) \geq \mathcal{C}(A)$ for any $\boldsymbol{\xi} \in \mathbb{R}^d$ and finite subsets A of $\mathbb{R}^d$. This forces subtleties on its analysis, which are not present in the study of the TSP or other monotone Euclidean functionals.

At the end of its paper, Steele outlined some open problems concerning with its result, e.g. the possibility to investigate the rate of convergenge for the asymptotic behaviour he provided. This task was carried out by Yukich [63], who also extended the above theorem to the critical range $p \geq d$ in the sense of complete convergence. In particular, for all $d \geq 2$ and $p \geq 1$ he established that for the average cost of the monopartite Euclidean MST it holds

$$\left| N^{-1+\frac{p}{d}} \, \mathcal{C}_{N,d}^{(p,\mathrm{rEm})} - c(p,d) \right| \leq c' N^{-\frac{1}{d}} \tag{3.12}$$

where c' is a simple constant.

Another important question stemming from theorem 3.2.1 was whether the constants $c(p,d)$ could be determined explicitly somehow. Despite the pessimistic attitude of Steele's itself about the topic, a step in this direction was performed some years later [64], exploiting a procedure that remarkably unified the derivation for the MST asymptotic behaviour for the Euclidean model and the purely uncorrelated one. Supposing as usual that the set $\{\boldsymbol{\xi}_i\}_{i=1}^N$ of i.i.d. points is given in $\mathbb{R}^d$, let us define $\mathtt{G}_N(z)$ as the graph composed by all vertices distant at most z, together with the function

$$g_k(y) = \lim_{N\to\infty} P_{k,N}\left[\left(\frac{y}{N v_d}\right)^{\frac{1}{d}}\right], \tag{3.13}$$

where v_d denotes the volume of the unit sphere in d dimensions, while $P_{k,N}(z)$ is the probability that a given point belongs to a component of $\mathtt{G}_N(z)$ having exactly k points. In this setting, the values of the constants appearing in (3.11) for $p \le d$ are given by

$$c(p,d) = \frac{p}{d\, v_d^{p/d}} \sum_{k=1}^{\infty} \frac{1}{k} \int_0^\infty g_k(y)\, y^{\frac{p}{d}-1}\, \mathrm{d}y. \tag{3.14}$$

Truncation of the sum after a finite number of terms yields a sequence of lower bounds for the constants, but unfortunately the functions $g_k(y)$ are increasingly harder to obtain analytically as k increases. Nonetheless, from the expression above one can obtain the following bounds in arbitrary dimension for the case $p = 1$

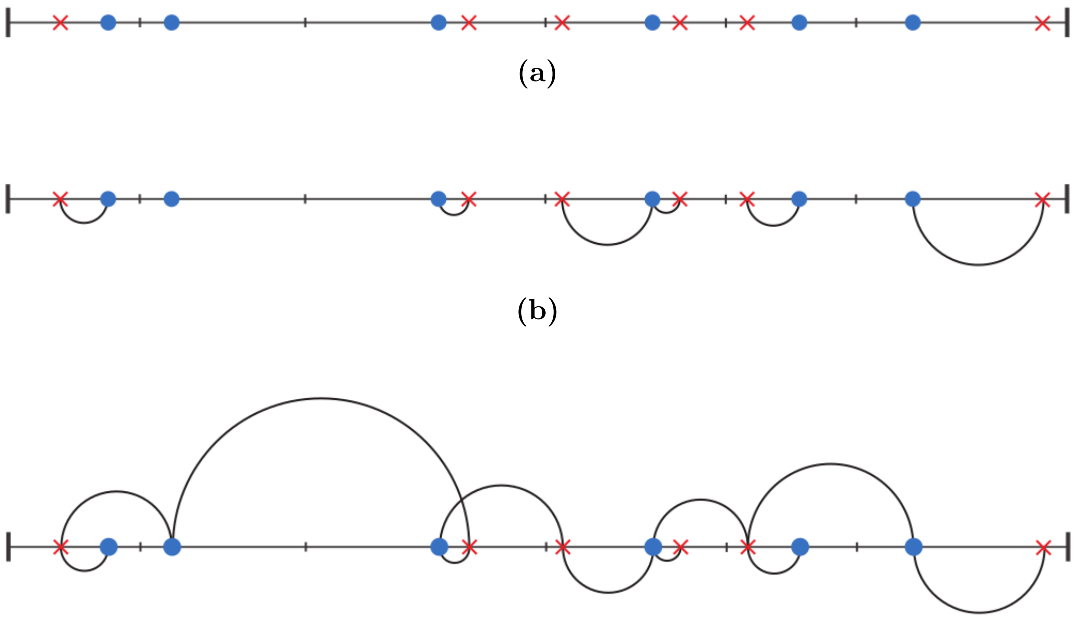
$$\frac{\Gamma(1/d)}{d\, v_d^{1/d}} \le c(1,d) \le \frac{2^{1/d}\,\Gamma(1/d)}{d\, v_d^{1/d}}. \tag{3.15}$$

We conclude this section by noting that, contrary to the monopartite case, no scaling results exist in the literature concerning the cost of the random Euclidean bipartite MST. The fact that, from a mathematical point of view, the treatment of the problem becomes more complicated should not particularly impress the reader. In fact, although the bipartite case appears as a slight modification of the monopartite one, the differences can prove to be crucial with respect not only to the value of the average optimal cost, but also to its scaling behaviour for large N, especially in low dimensions, where local fluctuations of points of different type are much more relevant.

Those discrepancies between the monopartite and bipartite case have been observed in several random Euclidean optimization problems, such as the matching problem [65, 72, 73] and the TSP [66, 74, 75]. Instead, they do not appear in the random Euclidean MST problem, as we will see explicitly in the next sections by analyzing in detail the one and two dimensional cases. We remark that this fact represents the main result of this thesis, and it will be explained in detail at the end of the chapter.

3.3 One dimensional rEb: exact solution for $p > 0$

In this section we want to provide the explicit solution for the random Euclidean bipartite MST problem in one dimension. As in the general setting illustrated



(a)

(b)

(c)

Figure 3.2: (a) $\mathtt{K}_{N,N} = \text{Graph}(\mathcal{R},\mathcal{B};\mathcal{E})$ with $N = 6$ embedded in the segment $[0,1]$, with the vertices chosen uniformly at random. (b)-(c) First and second step of the construction of the MST for the given graph. The lines corresponding to the edges of the MST are drawn as semicircles out of the segment for a matter of visualization.

above, we consider the complete graph $K_{N,N} = Graph(\mathcal{R},\mathcal{B};\mathcal{E})$, $N = |\mathcal{R}| = |\mathcal{B}| \in \mathbb{N}$, and we assume that the points $\{r_i\}_i$ and $\{b_i\}_i$, which correspond to the "red" $\mathcal{R}$ and "blue" $\mathcal{B}$ vertices respectively, are randomly and independently generated with uniform distribution in the interval $[0,1]$ (Fig. 3.2a). Moreover, we consider a weight function $w^{(p)} : \mathscr{E} \to \mathbb{R}^+$ of the form

$$w^{(p)}(e_{ij}) \equiv |r_i - b_j|^p \,, \qquad p \in \mathbb{R}^+ \tag{3.16}$$

i.e. a monotonically increasing function of the distance of the points. The quantity of interest for us is the average cost of the MST, which we have indicated in the previous section as

$$\mathcal{C}_{N,1}^{(p,\mathrm{rEb})} := \overline{\min_{\mathtt{T}\in\mathcal{T}} \sum_{e\in\mathcal{E}_{\mathtt{T}}} w^{(p)}(e)}, \tag{3.17}$$

with $\mathcal{T}$ set of all possible spanning trees of the graph $\mathtt{G}$.

First of all we provide an explicit construction method for the one-dimensional bipartite MST given a specific instance of the $2N$ points on the segment $[0,1]$. This will allow us to write down a general formula for the total cost of the MST, which can be computed exactly for all finite N. Let us consider one of the two sets of points, e.g. the blue one, as the seeds of a Voronoi diagram on the line (see Sect. 1.2.2). In Fig. 3.2a Voronoi cells are represented by the ticks, corresponding to the mean points between successive blue points. In the first step of the construction of the MST every red point connects to the (blue) seed of the Voronoi cell which he belongs to, as shown in Fig. 3.2b. After that, to form the remaining $N-1$ links, we have to connect different cells, and to do so we concentrate on the ticks separating them. For each tick, we select its closest red point, that could end up on its right or left, and connect it to the first blue point on the opposite side of the tick (see Fig. 3.2c). The resulting subgraph is the MST for the starting $2N$ vertices complete graph.

To definitely prove the last statement it is sufficient to consider the functioning of Kruskal's algorithm, explained in Sect. 1.2.3. Let us recall that once the edge weights have been sorted in increasing order, starting from the smallest one the algorithm checks if its corresponding link does form a loop when inserted in the developing spanning forest, discarding it if that happens and adding it otherwise. In our case, when a weight from the list is taken into account two situations can occur.

1. The corresponding link lies completely in the interior of a Voronoi cell (step I). In this case it is necessarily the first edge connecting its red end to the forest, so it will never form any loop.
2. The corresponding link connects two different cells (step II). When this happens, the link is added to the forest if and only if it is the shortest connecting the two cells. In fact, any longer edge will necessary form a loop, since all the red points between the two seeds of the cells have already been connected to them in the first step.

Let us note that by considering the previous procedure, only if more than one consecutive Voronoi cells are empty, i.e. free of red points in our case, it happens for an edge to connect non adjacent cells.

We conclude this proof, together with the section, observing that we never mentioned in our construction the power p which determines the values of the edge weights. This is due to fact that until the weighting function is a monotone increasing function of the distance of the points, the cluster structure formed by the Voronoi cells and the relative distance ordering among the points are unchanged. This means that given a specific instance of the $2N$ random points on $[0,1]$, the MST will be always the same for all $p > 0$.

3.3.1 The Grid-Poisson case

To evaluate explicitly the average cost of the rEb we first focus on its grid-Poisson version, where one set of points (e.g. the blue one) is supposed fixed on a grid. We are thus interested in the cost functional

$$\mathcal{C}_{N,1}^{(p,\mathrm{GR})} := \left.\overline{\mathcal{C}_N^{(p,\mathrm{Eb})}}\right|_{\mathcal{B}\text{ on the grid}}, \tag{3.18}$$

so we set

$$b_j = \frac{2j-1}{2N}, \qquad j = 1, \dots, N. \tag{3.19}$$

In what follows, we will always consider separately the contributions to the cost given by the two different steps of the MST construction given above. So let us start with the first one, taking into account the weights of intra-cells links (Fig. 3.2a).

If the red points are uniformly and independently distributed on $[0, 1]$ all cells are equivalent, so without loss of generality we can concentrate on one of them, say the first $[0, 1/N]$. The probability distribution density for the random variable R representing a red point's position ending up in this interval is simply $\rho_R(r) = N$, so the expected value of the p-th power of its distance from the blue seed at the center of the cell can be written as

$$\mathbb{E}\left[\left|R - \frac{1}{2N}\right|^p\right] = N \int_0^{\frac{1}{N}} \left|r - \frac{1}{2N}\right|^p \mathrm{d}r = \frac{1}{2^p(p+1)N^p} \tag{3.20}$$

This quantity times the number of red points gives the total contribution of the first step to the average cost of our MST. Observe that for $p = 1$ the above result turns into $1/4N$, which can be derived immediately from simmetry considerations too.

Considering now also the second step of the construction, we can write explicitly the total cost of the MST in the grid-Poisson case

$$\mathcal{C}_{N,1}^{(p,\mathrm{GR})} = \frac{1}{2^p(p+1)N^{p-1}} + \overline{\sum_{i=1}^{N-1} \left(\min_{j=1,\dots,N} \left| r_j - \frac{i}{N} \right| + \frac{1}{2N} \right)^p} \Bigg|_{\mathcal{B} \text{ on the grid}} \tag{3.21}$$

Note that in the second term the sum runs over the mean points i/N (ticks in Fig. 3.2c) between successive blue points, and for every link the contribution can be splitted in two parts: a constant $1/2N$ corresponding to the distance between the blue seed and the tick, plus the distance between the tick and its closest red point.

To compute the average over $\mathcal{R}$, i.e. over red points' positions, we first need the distribution function of the random variable

$$M^{(y)} = \min_{j=1,\dots,N} |r_j - y| , \tag{3.22}$$

where $y \in (0, 1)$ is considered fixed. For this purpose, we start from the all order statistics probability distribution of N random samples choosen uniformly in $[0, 1]$ (see Appendix B). Thus, supposing the red points are labeled in such a way that $r_i < r_j$ if $i < j$, their joint distribution can be written as

$$\rho_{R_{(1)},\dots,R_{(N)}}(r_1, \dots, r_N) = N!\theta(r_1) \prod_{i=1}^{N-1} \theta(r_{i+1} - r_i)\theta(1 - r_N). \tag{3.23}$$

The probability distribution for the distance between the N ordered red points and y, let us call it f, follows then by simply imposing N constraints with a delta function on the previous distribution

$$\begin{aligned} f(d_{r_1}^{(y)}, \dots, d_{r_N}^{(y)}) &= \int \rho_{R_{(1)},\dots,R_{(N)}}(r_1, \dots, r_N) \prod_{i=1}^{N} \delta\left[d_{r_i}^{(y)} - (r_i - y)\right] \mathrm{d}r_1 \dots \mathrm{d}r_N = \\ &= N!\, \theta(d_{r_1}^{(y)} + y) \prod_{i=1}^{N-1} \theta(d_{r_{i+1}}^{(y)} - d_{r_i}^{(y)})\theta(1 - d_{r_N}^{(y)} - y) \end{aligned} \tag{3.24}$$

To compute the pdf for the random variable $M^{(y)}$ we start from its cumulative distribution function

$$\Pr\left[\min_{j=1,\dots,N}\left|d^{(y)}_{r_j}\right| < m\right] = 1 - \Pr\left[\left|d^{(y)}_{r_1}\right| \geq m, \dots, \left|d^{(y)}_{r_N}\right| \geq m\right] =$$

$$= 1 - \int_{-y}^{1-y} f(l_1,\dots,l_N)\prod_{i=1}^{N}\theta(|l_i| - m)\,\mathrm{d}l_1\dots\mathrm{d}l_N =$$

$$= 1 - \int_{[-y,-m]\cup[m,1-y]} f(l_1,\dots,l_N)\,\mathrm{d}l_1\dots\mathrm{d}l_N \tag{3.25}$$

where we have used the fact that $d^{(y)}_{ri} \in [-y, 1-y] \quad \forall i$. By observing the interval of integration we note that the result will be different from 0 only if $m \in [0, \max\{y, 1-y\}]$. To simplify the notation from now on we will denote $\int_{[-y,-m]\cup[m,1-y]} \equiv \int$. Taking the derivative of the above cumulative we obtain

$$\rho_{M^{(y)}}(m) = -\frac{\partial}{\partial m}\left[\int_{-y}^{-m}\mathrm{d}l_1\int\mathrm{d}l_2\dots\mathrm{d}l_N f(l_1,\dots,l_N)\right.$$

$$\left. + \int_{m}^{1-y} dl_1 \int dl_2\dots dl_N f(l_1,\dots,l_N)\right]$$

$$= \int \mathrm{d}l_2\dots\mathrm{d}l_N\left[f(-m,l_2,\dots,l_N) + f(m,l_2,\dots,l_N)\right]$$

$$- \int\mathrm{d}l_1\frac{\partial}{\partial m}\int\mathrm{d}l_2\dots\mathrm{d}l_N f(l_1,\dots,l_N) = \dots$$

$$= \sum_{i=1}^{N}\int\mathrm{d}l_1\dots\cancel{\mathrm{d}l_i}\dots\mathrm{d}l_N\left[f(l_1,\dots,-m,\dots,l_N) + f(l_1,\dots,m,\dots,l_N)\right] \tag{3.26}$$

with $-m$ and m both as the i-th argument in the last line. We can now substitute Eq. (3.24) in the previous expression, considering $y \in [0, 1/2]$ and $y \in [1/2, 1]$ separately. Nonetheless, the calculations are exactly the same, so we write explicitly only the first case.

Looking at the form of the pdf in Eq. (3.24) it is clear that one has to pay attention when substituting m and $-m$ if the index of the last sum is $i = 1$ or $i = N$. For this reason we will consider separately these two situations, excluded from the general case which starts below. Proceeding with the substitution, the terms of the sum in Eq. (3.26) become

$$N!\sum_{i=2}^{N-1}\int\mathrm{d}l_1\dots\cancel{\mathrm{d}l_i}\dots\mathrm{d}l_N\,\theta(l_1+y)\dots\left[\theta(-m-l_{i-1})\theta(l_{i+1}+m)\right.$$

$$\left. + \theta(m-l_{i-1})\theta(l_{i+1}-m)\right]\dots\theta(1-l_N-y) \tag{3.27}$$

with the two terms in the square brakets giving the same contribution when one considers the intervals of integration (remember the shorthand), which imply

$$l_{i-1} < -m, \qquad l_{i+1} > -m$$
$$l_{i-1} < m, \qquad l_{i+1} > m.$$

Moreover, from the inequality $l_{i-1} < m$ the integral is nonzero only if $m < y$, so we write

$$2N!\,\theta(y-m)\sum_{i=2}^{N-1}\int_{-y}^{-m}\mathrm{d}l_{i-1}\int_{m}^{1-y}\mathrm{d}l_{i+1}\int\mathrm{d}l_1\ldots\cancel{\mathrm{d}l_{i-1}\mathrm{d}l_i\mathrm{d}l_{i+1}}\ldots\mathrm{d}l_N$$

$$\cdot\,\theta(l_1+y)\ldots\theta(l_{i-1}-l_{i-2})\theta(l_{i+2}-l_{i+1})\ldots\theta(1-l_N-y)=\ldots$$

$$=2N!\theta(y-m)\sum_{i=2}^{N-1}\int_{-y}^{-m}\mathrm{d}l_{i-1}\int_{-y}^{l_{i-1}}\mathrm{d}l_{i-2}$$

$$\cdots\int_{-y}^{l_2}\mathrm{d}l_1\int_{m}^{1-y}\mathrm{d}l_{i+1}\int_{l_{i+1}}^{1-y}\mathrm{d}l_{i+2}\cdots\int_{l_{N-1}}^{1-y}\mathrm{d}l_N \tag{3.28}$$

In the particular case $i=1$ we have instead

$$N!\int\mathrm{d}l_2\ldots\mathrm{d}l_N[\theta(-m+y)\theta(l_2+m)+\theta(m+y)\theta(l_2-m)]\theta(l_3-l_2)\cdots$$

$$=N!\int_{m}^{1-y}\mathrm{d}l_2\int\mathrm{d}l_3\ldots dl_N[\theta(y-m)+1]\theta(l_3-l_2)\cdots\theta(1-l_N-y)=\ldots$$

$$=N!\int_{m}^{1-y}\mathrm{d}l_2\int_{l_2}^{1-y}\mathrm{d}l_3\cdots\int_{l_{N-1}}^{1-y}\mathrm{d}l_N[\theta(y-m)+1] \tag{3.29}$$

Here it is evident that an additional term emerges for $y < m < 1-y$, given by the $+1$ in the braket. The same calculation can be carried out for $i=N$, where no new terms appear with respect to the general case. In the end, putting all together, we have

$$\rho^{(1)}_{M^{(y)}}(m)=2N!\,\theta(y-m)\sum_{i=1}^{N}\int_{-y}^{-m}\mathrm{d}l_{i-1}\int_{-y}^{l_{i-1}}\mathrm{d}l_{i-2}\cdots\int_{-y}^{l_2}\mathrm{d}l_1$$

$$\cdot\int_{m}^{1-y}\mathrm{d}l_{i+1}\int_{l_{i+1}}^{1-y}\mathrm{d}l_{i+2}\cdots\int_{l_{N-1}}^{1-y}\mathrm{d}l_N$$

$$+\,N!\int_{m}^{1-y}\mathrm{d}l_2\int_{l_2}^{1-y}\mathrm{d}l_3\cdots\int_{l_{N-1}}^{1-y}\mathrm{d}l_N\theta(m-y) \tag{3.30}$$

The integrals are easy to evaluate and the expression one obtains is

$$\rho^{(1)}_{M^{(y)}}(m)=2N!\,\theta(y-m)\sum_{i=1}^{N}\frac{1}{(i-1)!(N-i)!}(y-m)^{i-1}(1-y-m)^{N-i}$$

$$+\,N(1-y-m)^{N-1}\theta(m-y)\theta(1-y-m) \tag{3.31}$$

Finally, the finite sum can be computed explicitly using the binomial theorem, so the probability distribution for the minimum distance between the N red points and a fixed one $y \in [0, 1/2]$ is given by

$$\rho^{(1)}_{M^{(y)}}(m) = \begin{cases} 2N(1-2m)^{N-1} & 0 < m < y \\ N(1-y-m)^{N-1} & y < m < 1-y \\ 0 & m > 1-y \end{cases} \qquad y \in (0, \tfrac{1}{2}) \tag{3.32}$$

The result for the specular case $y \in [1/2, 1]$, due to trivial symmetry considerations, can be obtained directly by substituting $y \to 1-y$ in the above expression, so we obtain

$$\rho^{(2)}_{M^{(y)}}(m) = \begin{cases} 2N(1-2m)^{N-1} & 0 < m < 1-y \\ N(y-m)^{N-1} & 1-y < m < y \\ 0 & m > y \end{cases} \qquad y \in (\tfrac{1}{2}, 1) \tag{3.33}$$

At this point we are ready to perform the average over the red points' positions of Eq. (3.21), namely the average over the disorder of our system. For simplicity, let us first consider the case $p = 1$, in which the average cost for the MST can be rewritten as

$$\mathcal{C}^{(1,\mathrm{GR})}_{N,1} = \frac{1}{4} + \frac{N-1}{2N} + \sum_{i=1}^{N-1} \overline{M^{(\frac{i}{N})}}\Bigg|_{\mathcal{B} \text{ on the grid}} \tag{3.34}$$

For $y \in [0, 1/2]$ we have to perform the following integrals

$$\begin{aligned}\overline{M^{(y)}} &= 2N \int_0^y m(1-2m)^{N-1} dm + N \int_y^{1-y} m(1-y-m)^{N-1} dm \\ &= \frac{(1-2y)^{N+1}+1}{2(N+1)}\end{aligned} \tag{3.35}$$

which can be readily asapted as before to the case $y \in [1/2, 1]$. The obtained results can be putted together, so substituting $y = i/N$ in the end we find

$$\mathcal{C}^{(1,\mathrm{GR})}_{N,1} = \frac{1}{4} + \frac{(2N+1)(N-1)}{2N(N+1)} + \frac{1}{N+1} \sum_{i=1}^{\lfloor \frac{N-1}{2} \rfloor} \left(1 - \frac{2i}{N}\right)^{N+1} \tag{3.36}$$

We stress that this result is exact for all number of points N, because no assumptions have been performed during its derivation. Only now, not being able to provide a closed form for the sum appearing above, it is worth performing a large N expansion of the formula, in order to obtain a result comparable with numerical simulations that we will report in Sect. 3.4. By approximating its argument as an exponential, the partial sum turns into a geometric one, which can be easily evaluated

$$\frac{1}{N+1} \sum_{i=1}^{\lfloor \frac{N-1}{2} \rfloor} \left(1 - \frac{2i}{N}\right)^{N+1} \approx \frac{1}{N+1} \sum_{i=1}^{\lfloor \frac{N-1}{2} \rfloor} \mathrm{e}^{-\frac{2(N+1)}{N} i} = \frac{1}{N+1} \frac{1 - \mathrm{e}^{-N+\frac{1}{N}}}{\mathrm{e}^{2+\frac{2}{N}} - 1}. \tag{3.37}$$

Therefore, the expression of the average cost for $N \gg 1$ is given by

$$\mathcal{C}_{N,1}^{(1,\mathrm{GR})} = \frac{5}{4} + \frac{5-3\mathrm{e}^2}{2(\mathrm{e}^2-1)}\frac{1}{N} + \frac{\mathrm{e}^4-5\mathrm{e}^2+2}{(\mathrm{e}^2-1)^2}\frac{1}{N^2} + o\left(\frac{1}{N^2}\right) \tag{3.38}$$

When $p > 0$ the procedure is similar and the computation gets just a little more involved. This time the formula for the average cost reads

$$\mathcal{C}_{N,1}^{(p,\mathrm{GR})} = \frac{1}{2^p(p+1)N^{p-1}} + \sum_{i=1}^{N-1} \overline{\left(M^{\left(\frac{i}{N}\right)} + \frac{1}{2N}\right)^p}\Bigg|_{\mathcal{B}\text{ on the grid}} \tag{3.39}$$

and for $y \in [0, 1/2]$ the argument of the sum corresponds to,

$$I_1+I_2 \equiv 2N\int_0^y (m+c)^p(1-2m)^{N-1}\mathrm{d}m + N\int_y^{1-y}(m+c)^p(1-y-m)^{N-1}\mathrm{d}m \tag{3.40}$$

where we have defined $c := 1/2N$. The two terms can be computed in the same way by iteratively integrating by parts, so let us focus only on the first one (note that $p \in \mathbb{R}^+$ so one can only take the derivative of the term independent of p to get rid of it after $N-1$ steps).

$$\begin{aligned} I_1 &= \frac{2N}{p+1}\left[(y+c)^{p+1}(1-2y)^{N-1} - c^{p+1}\right] \\ &\quad + \frac{4N(N-1)}{p+1}\int_0^y (m+c)^{p+1}(1-2m)^{N-2}\mathrm{d}m = \dots \\ &= N!p!\sum_{j=1}^{N}\frac{2^j}{(N-j)!(p+j)!}\left[(y+c)^{p+j}(1-2y)^{N-j} - c^{p+j}\right] \end{aligned} \tag{3.41}$$

We notice that the above result does not hold for $y = 1/2$, because an indeterminate form 0^0 emerges when $j = N$. By evaluating singularly this particular case, all the terms containing $1-2y$ vanish and we arrive to

$$I_1^{(y=1/2)} = N!\,p!\left[-\sum_{j=1}^{N}\frac{2^j}{(N-j)!\,(p+j)!}c^{p+j} + \frac{2^N}{(N+p)!}\left(\frac{1}{2}+c\right)^{N+p}\right]. \tag{3.42}$$

Note that this problem does not arise in I_2, which is identically zero when $y = 1/2$. Putting together the results of the two integrals we obtain

$$\overline{(M^{(y)}+c)^p} = I_1 + I_2$$

$$= \begin{cases} N!\,p!\left\{\sum_{j=1}^{N}\frac{1}{(N-j)!\,(p+j)!}\left[(2^j-1)(y+c)^{p+j}(1-2y)^{N-j}\right.\right. \\ \qquad \left.\left. -2^jc^{p+j}\right] + \frac{1}{(N+p)!}(1-y+c)^{N+p}\right\} & y \in (0, \frac{1}{2}) \\ N!\,p!\left[-\sum_{j=1}^{N}\frac{2^j}{(N-j)!\,(p+j)!}c^{p+j} + \frac{2^N}{(N+p)!}\left(\frac{1}{2}+c\right)^{N+p}\right] & y = \frac{1}{2} \end{cases} \tag{3.43}$$

which is valid in the case $y \in [1/2, 1]$ too if one substitutes $y \to 1 - y$ as usual.

In conclusion, rearranging the various terms the average optimal cost of the grid-Poisson one-dimensional MST for $p > 0$ and for all N is

$$
\begin{aligned}
\mathcal{C}_{N,1}^{(p,\mathrm{GR})} = {} & \frac{1}{2^p(p+1)N^{p-1}} - \frac{(N-1)N!\,p!}{(2N)^p} \sum_{j=1}^{N} \frac{1}{N^j (N-j)!\,(p+j)!} \\
& + 2N!\,p! \sum_{i=1}^{\left\lfloor \frac{N-1}{2} \right\rfloor} \left[\sum_{j=1}^{N} \frac{2^j - 1}{(N-j)!\,(p+j)!} \left(\frac{i}{N} + \frac{1}{2N} \right)^{p+j} \left(1 - \frac{2i}{N} \right)^{N-j} \right. \\
& \left. + \frac{1}{(N+p)!} \left(1 - \frac{i}{N} + \frac{1}{2N} \right)^{N+p} \right] \\
& + N!\,p! \left[\frac{2^N}{(N+p)!} \left(\frac{1}{2} + \frac{1}{2N} \right)^{N+p} \right] \frac{1 + (-1)^N}{2}
\end{aligned}
\tag{3.44}
$$

Note that the last line survives only when N is even due to the last factor, in fact this contribution arises from the case $y = i/N = 1/2$, appearing in the MST only if one has an even number of Voronoi cells.

3.3.2 The Poisson-Poisson case

Making use of the calculations performed in the previous section we now concentrate on the Poisson-Poisson MST problem, where both red $\mathcal{R} = \{r_i\}_i$ and blue $\mathcal{B} = \{b_i\}_i$ points are chosen uniformly at random in $[0, 1]$. In this case, a generalization of Eq. (3.21) provides the cost functional for our problem in the following form

$$
\begin{aligned}
\mathcal{C}_{N,1}^{(p,\mathrm{RR})} = {} & \sum_{i=1}^{N} \overline{\left(\min_{j=1,\dots,N} |r_i - b_j| \right)^p} \\
& + \sum_{i=1}^{N-1} \overline{\left(\min_{j=1,\dots,N} \left| r_j - \frac{b_{i+1} + b_i}{2} \right| + \frac{b_{i+1} - b_i}{2} \right)^p},
\end{aligned}
\tag{3.45}
$$

where now the average have to be carried out w.r.t. to the positions of both point sets. Again, this formula follows straightforwardly from the construction of the MST illustrated in Sect. 3.3. The first term is the contribution of the links that lie completely in the interior of a Voronoi cell. In fact by choosing the blue points as the seeds of the Voronoi diagram, every red point connects to its closest blue neighbour. Instead, the second term takes into account links between different cells and is composed by two pieces:

1. the distance between the boundary of the cell (indicated with a tick in Fig. 3.2), i.e. the mean point of two consecutive blue seeds, and its closest red point
$$
\min_{j=1,\dots,N} \left| r_j - \frac{b_{i+1} + b_i}{2} \right| \tag{3.46}
$$

2. the distance between the boundary of the cell and the blue seed, e.g. the left one
$$\frac{b_{i+1}+b_i}{2}-b_i=\frac{b_{i+1}-b_i}{2}. \tag{3.47}$$

Although the method adopted is the same, the calculations are quite long when $p>0$, thus we just sketch the derivation of the average cost in the simple case $p=1$, providing the result for the general situation at the end. Let us consider the first term of Eq. (3.45). Being the minimum taken on the positions of the blue points, we can consider the random variables $r_i \in [0,1]$, $i=1,\dots,N$, fixed, and therefore the average over the set $\mathcal{B}$ is given by Eq. (3.35) of the grid-Poisson case, with $y=r_i$,
$$\left.\overline{M^{(r_i)}}\right|_{\mathcal{R}\text{ fixed}}=\frac{|1-2r_i|^{N+1}+1}{2(N+1)} \tag{3.48}$$
At this point, to average over the set $\mathcal{R}$ we need the i-th marginal distribution for the N points order statistics. In fact, we are supposing that the points are ordered, i.e. $r_i<r_j \iff i<j$, so the probability of finding r_i in the interval $\mathrm{d}r\equiv(r,r+\mathrm{d}r)$ is given by Eq. (B.12)
$$\Pr\left[r_i\in\mathrm{d}r\right]=\frac{N!}{(i-1)!\,(N-i)!}r^{i-1}(1-r)^{N-i}\mathrm{d}r, \tag{3.49}$$
and the quantity we have to compute is
$$\overline{M^{(r_i)}}=\frac{1}{2(N+1)}\left(1+\int_0^1|1-2r|^{N+1}\Pr\left[r_i\in\mathrm{d}r\right]\right) \tag{3.50}$$
To evaluate the integral, we note that it is very convenient to carry out first the sum appearing in Eq. (3.45) by exploiting Newton's formula. Doing so we obtain
$$\begin{aligned}\sum_{i=1}^{N}\overline{M^{(r_i)}}&=\frac{N}{2(N+1)}\left(1+\int_0^1\mathrm{d}r\,|1-2r|^{N+1}\right)\\&=\frac{N}{2(N+1)}+\frac{N}{2(N+1)(N+2)}\end{aligned} \tag{3.51}$$

The computation of the second line of Eq. (3.45) proceeds in a similar way, this time with the average over the red points performed first. For this reason, we start again from the result in Eq. (3.35), now with $y=(b_{i+1}+b_i)/2$
$$\left.\overline{M^{\left(\frac{b_{i+1}+b_i}{2}\right)}}\right|_{\mathcal{B}\text{ fixed}}+\frac{b_{i+1}-b_i}{2}=\frac{|1-(b_{i+1}+b_i)|^{N+1}+1}{2(N+1)}+\frac{b_{i+1}-b_i}{2} \tag{3.52}$$
For the second fraction, the average over blue points is easily performed with the i-th and $(i+1)$-th order statistic distributions, exactly as we did above for the red points, by again evaluating the sum appearing in Eq. (3.45) as the first step
$$\sum_{i=1}^{N-1}\overline{\frac{b_{i+1}-b_i}{2}}=\frac{N-1}{2(N+1)} \tag{3.53}$$

For the first franction we need instead the joint probability of finding b_i in $\mathrm{d}x \equiv (x, x + \mathrm{d}x)$ and b_{i+1} in $\mathrm{d}y \equiv (y, y + \mathrm{d}y)$, which is given by Eq. (B.13)

$$\Pr\left[b_i \in \mathrm{d}x, b_{i+1} \in \mathrm{d}y\right] = \frac{N!}{(i-1)!\,(N-i-1)!} x^{i-1} (1-y)^{N-i-1} \theta(y-x) \mathrm{d}x\mathrm{d}y \quad (3.54)$$

We are thus left to evaluate the following expression

$$\begin{aligned}\sum_{i=1}^{N-1} \overline{M^{\left(\frac{b_{i+1}+b_i}{2}\right)}} &= \frac{1}{2(N+1)} \sum_{i=1}^{N-1} \left(1 + \int_0^1 \left|1-(x+y)\right|^{N+1} \Pr\left[b_i \in \mathrm{d}x, b_{i+1} \in \mathrm{d}y\right]\right) \\ &= \frac{N-1}{2(N+1)} \left(1 + N \int_0^1 \left|1-(x+y)\right|^{N+1} (1+x-y)^{N-2} \theta(y-x)\right) \end{aligned} \quad (3.55)$$

Splitting the integral in two parts because of the absolute value, one finds that the two contributions are the same, as can be verified by substituting $x \to 1 - x$, $y \to 1 - y$, and then by renaming $x \leftrightarrow y$. Therefore, considering for a moment only the term containing the integral, the result can be obtained with a series of integrations by parts

$$\begin{aligned}&\frac{N(N-1)}{2(N+1)} \int_0^{\frac{1}{2}} \mathrm{d}x \int_x^{1-x} \mathrm{d}y (1-x-y)^{N+1} (1+x-y)^{N-2} \\ &= \frac{N(N-1)}{2(2N+1)(N+2)(N+1)}. \end{aligned} \quad (3.56)$$

By putting together the results in Eqs. (3.51) and (3.55)-(3.56) we find that the exact value of the average cost for the Poisson-Poisson MST in one dimension for $p = 1$ is

$$\mathcal{C}_{N,1}^{(1,\mathrm{RR})} = \frac{3}{2} - \frac{4}{N+1} + \frac{2}{N+2} + \frac{1}{2(2N+1)} \quad (3.57)$$

which in the large N limit can be written as

$$\mathcal{C}_{N,1}^{(1,\mathrm{RR})} = \frac{3}{2} - \frac{7}{4N} - \frac{1}{8N^2} + o\left(\frac{1}{N^2}\right). \quad (3.58)$$

As we have promised, we now provide the result for the general case $p > 0$ too. Firstly, let us note again that the calculation one has to perform is identical to the one described above. The only difference consists in the fact that the starting point is given by the average over one set of points provided by Eq. (3.43), with $y = r_i$, $c = 0$ and $y = (b_{i+1} - b_i)/2$, $c = (b_{i+1} - b_i)/2$ for the computation of the first and second term of Eq. (3.45) respectively. In particular, the results one obtain for them are the following

$$\sum_{i=1}^{N} \overline{(M^{(r_i)})^p} = \frac{N!\,p!\,N}{2^p (N+p+1)!} \left(2^{p+1} + N - 1\right) \quad (3.59)$$

$$\begin{aligned}\sum_{i=1}^{N-1}\overline{\left(M^{\left(\frac{b_{i+1}+b_i}{2}\right)}+\frac{b_{i+1}-b_i}{2}\right)^p} &= 2N!\,p!\,N(N-1)\left\{\sum_{j=1}^{N}(2^j-1)(N-2)!\right.\\
&\cdot\sum_{k=1}^{N-1}\frac{1}{(N-j+k)!\,(N-k-1)!}\left[\frac{(2N-j-1)!}{(2N+p)!}-\frac{(N-j+k)!}{2^{p+j+1}(N+p+k+1)!}\right]\\
&+\frac{(N-2)!\,\left(1-2^N N+2^N\right)}{2^{N+p+1}}\sum_{j=1}^{N}\frac{1}{(N-j)!\,(N+p+j)!}\\
&\left.+\frac{1}{(N-1)(N+p+1)!}\left(1-\frac{1}{2^{N+p+1}}\right)-\frac{2^{N-1}(N-2)!}{(2N+p)!}\right\}\end{aligned}\tag{3.60}$$

The final result, reached after some manipulations of the expressions above, is given by

$$\begin{aligned}\mathcal{C}_{N,1}^{(p,\mathrm{RR})} &= 2N(N-1)!\,p!\left[\frac{(N-1)(2N+p+1)+2^{p+1}N}{2^p(N+p+1)!}-\frac{4^N N(N-1)!}{(2N+p)!}\right]\\
&+\frac{2(N!)^2p!}{(2N+p)!}\sum_{j=0}^{N}2^j\sum_{k=1}^{N}\binom{2N-j-1}{N-k}+\\
&-\frac{(N!)^2p!\,(N-1)}{2^{p-1}}\sum_{j=0}^{N}\frac{1}{(N-j)!\,(N+p+j)!}\end{aligned}\tag{3.61}$$

3.4 Numerical investigation in one and two dimensions

The aim of the present section is to provide a collection of numerical results concerning the average cost and the cost variance for the Euclidean MST in one and two dimensions. The reasons for this are twofold. First of all we want to check the correctness of the formulas derived in the last two sections for the average cost of the one dimensional bipartite MST problem. Secondly, in light of the considerations we made at the end of Sect. 3.2, we want to analyze the scaling behavior of the average cost of our problem for large N, in order to see how the monopartite result of Theorem 3.2.1 is modified when two different sets of points are present, namely in the bipartite case.

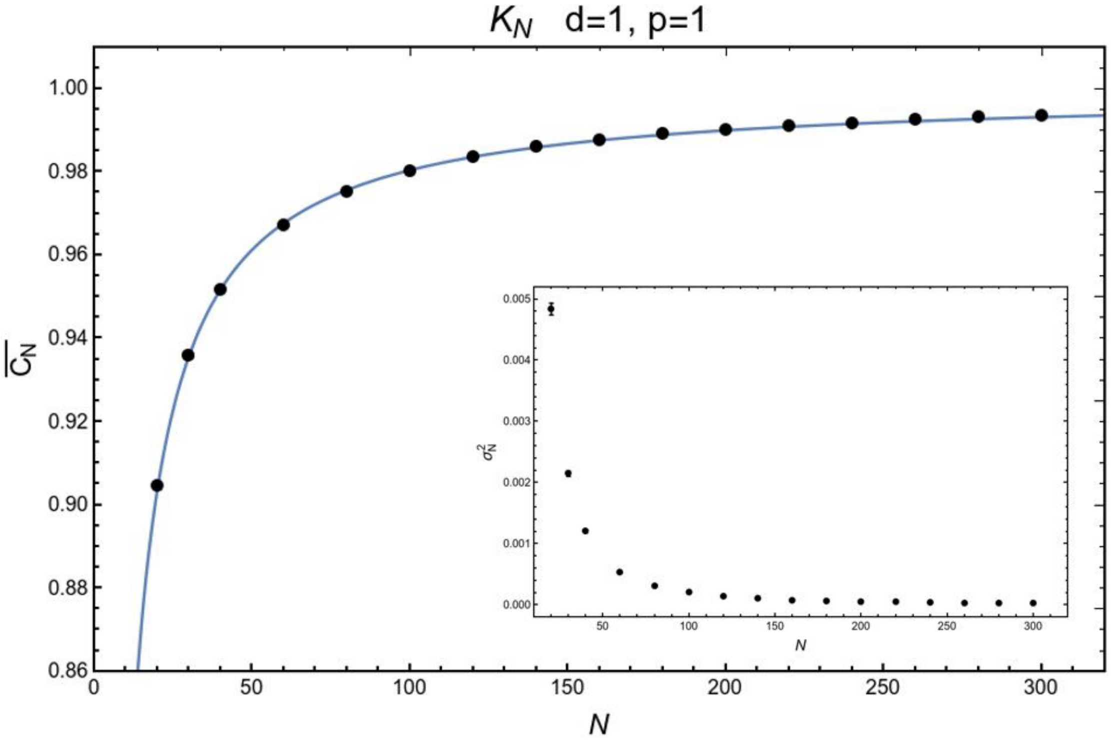
In the plots below the specific problem considered is indicated in the title. We denote by K_N the complete monopartite case, and by $K_{N,N}$ the bipartite one, with the labels GR and RR for its grid-Poisson and Poisson-Poisson versions respectively. Let us note that the positions of the points on the grid are given in one dimension by Eq. (3.19), and more generally are such that all points are equidistant from each other and from the boundary of the Euclidean domain.

Each of the subsequent plots has been obtained using Mathematica and adopting the following procedure. For a given instance of our problem, one (K_N or GR) or two (RR) set of points are uniformly generated on the unit interval ($d = 1$) or square ($d = 2$). The list of their p-th power Euclidean distances, i.e. of the edge weights, together with a complete (bipartite) graph structure, is passed to Kruskal's algorithm to find the MST of the graph, and consequently its cost $\mathcal{C}_N$. This steps are iterated n times to compute the mean $\overline{\mathcal{C}_N}$ and the rescaled variance of the cost (shown in the inset of each plot), which we denote by

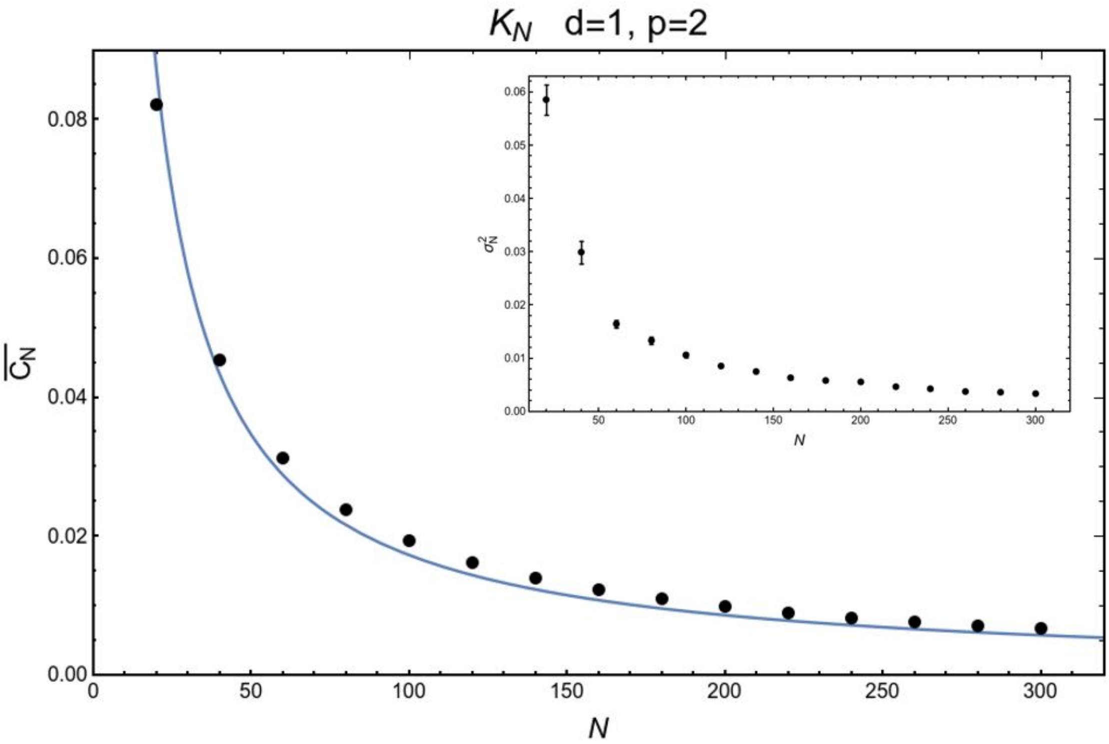
$$\sigma_N^2 := \frac{\overline{\mathcal{C}_N^2} - \overline{\mathcal{C}_N}^2}{\overline{\mathcal{C}_N}^2}, \tag{3.62}$$

together with their standard errors. Note that the number n, that we specify under each plot, is chosen every time in such a way to reach a good compromise between the precision of the numerical results and the running time necessary to obtain them.

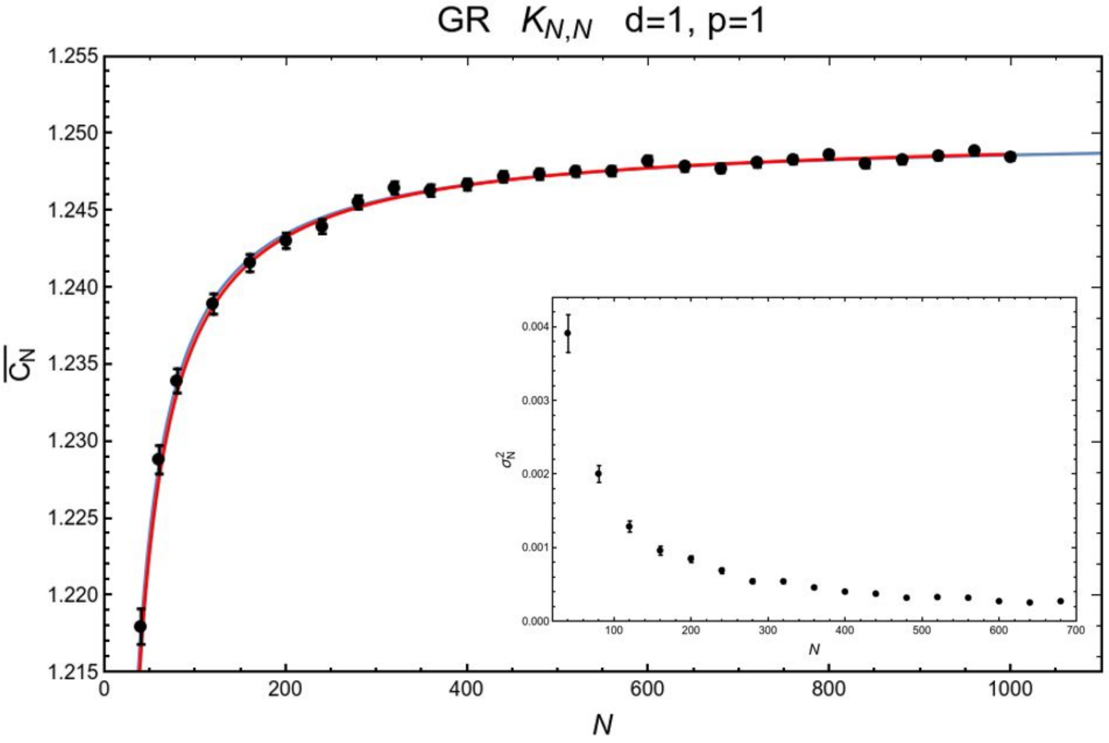
We then perform a fit on the numerical data of the average cost to obtain its scaling behavior for large N: the result is both plotted as a blue line and written below each plot. Moreover, in the one dimensional case, specifically in the bipartite setting, we plot as a red line the numerical values computed with our exact solution given in the previous section. Finally, let us note that error bars are always present, but very often they end up being smaller than the marker sizes.



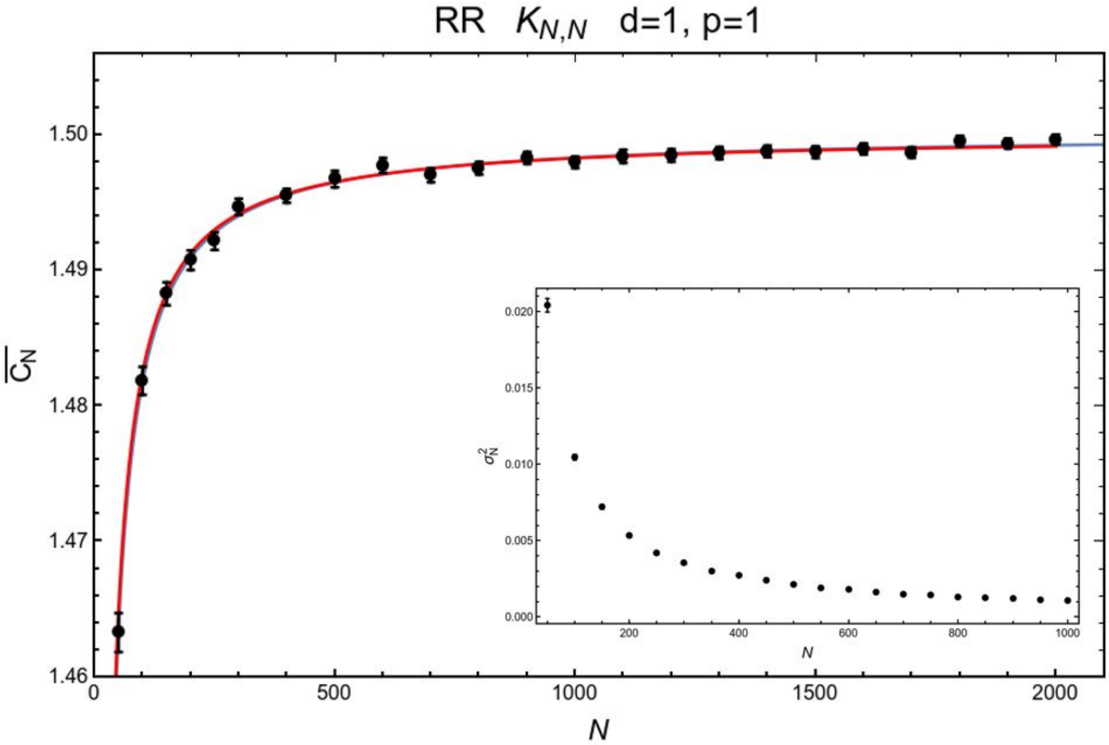
(a) $n = 1000$ $\qquad f(N) = 0.9995 - \frac{1.9071}{N}$



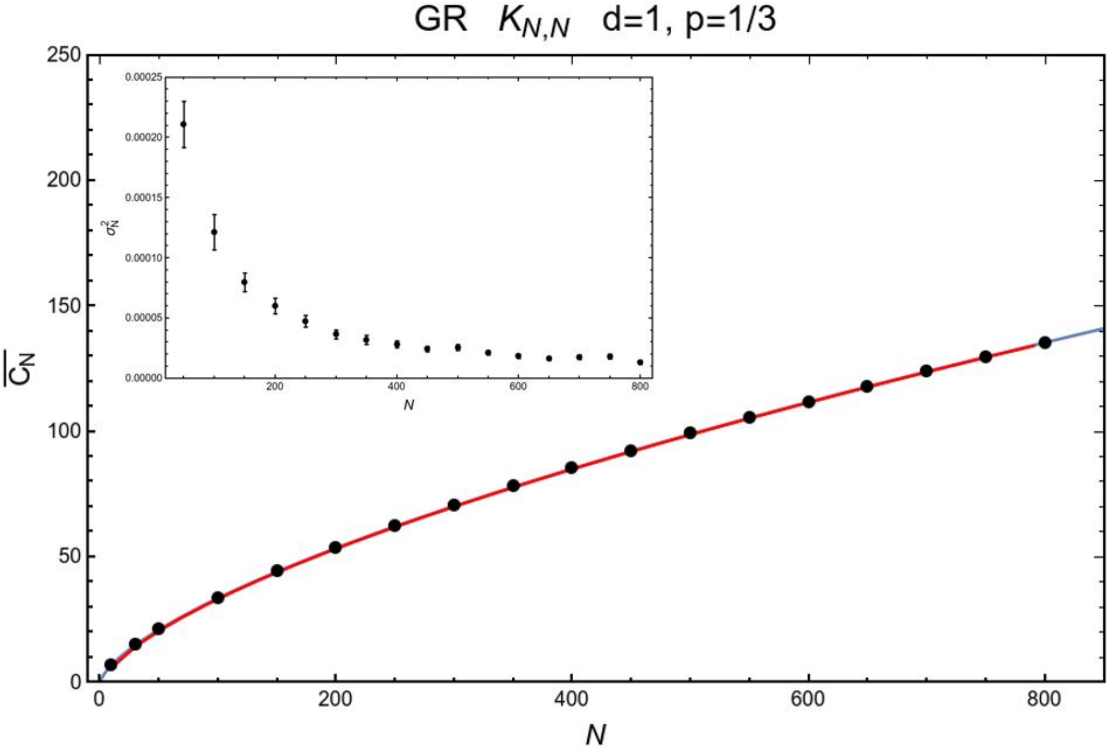
(b) $n = 2000$ $\qquad f(N) = \frac{1.7259}{N}$



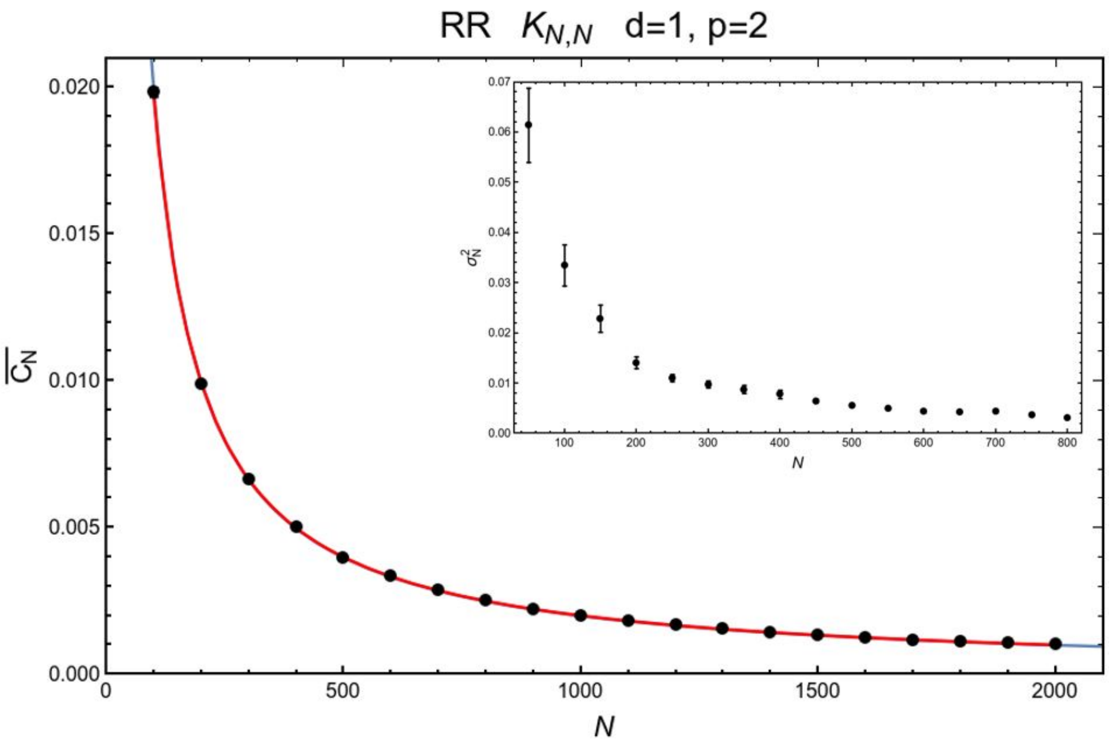
(c) $n = 4000$ $\qquad f(N) = 1.2499 - \frac{1.2783}{N}$



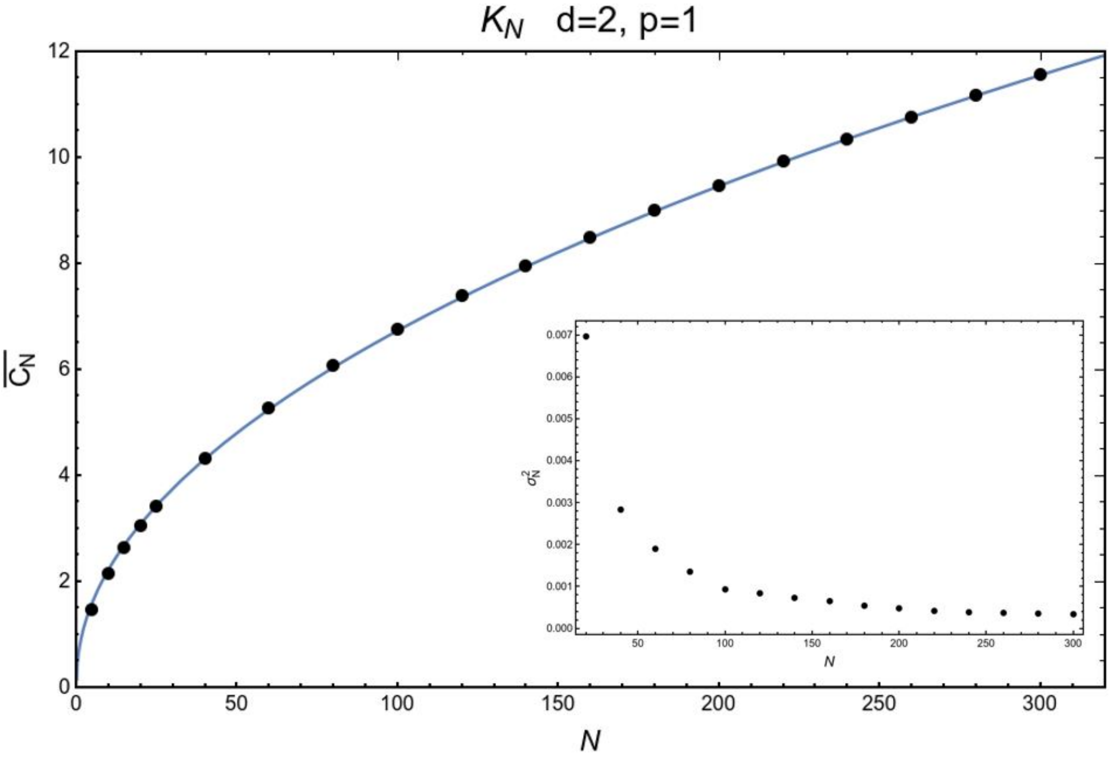
(d) $n = 6000 \qquad f(N) = 1.5001 - \frac{1.8384}{N}$



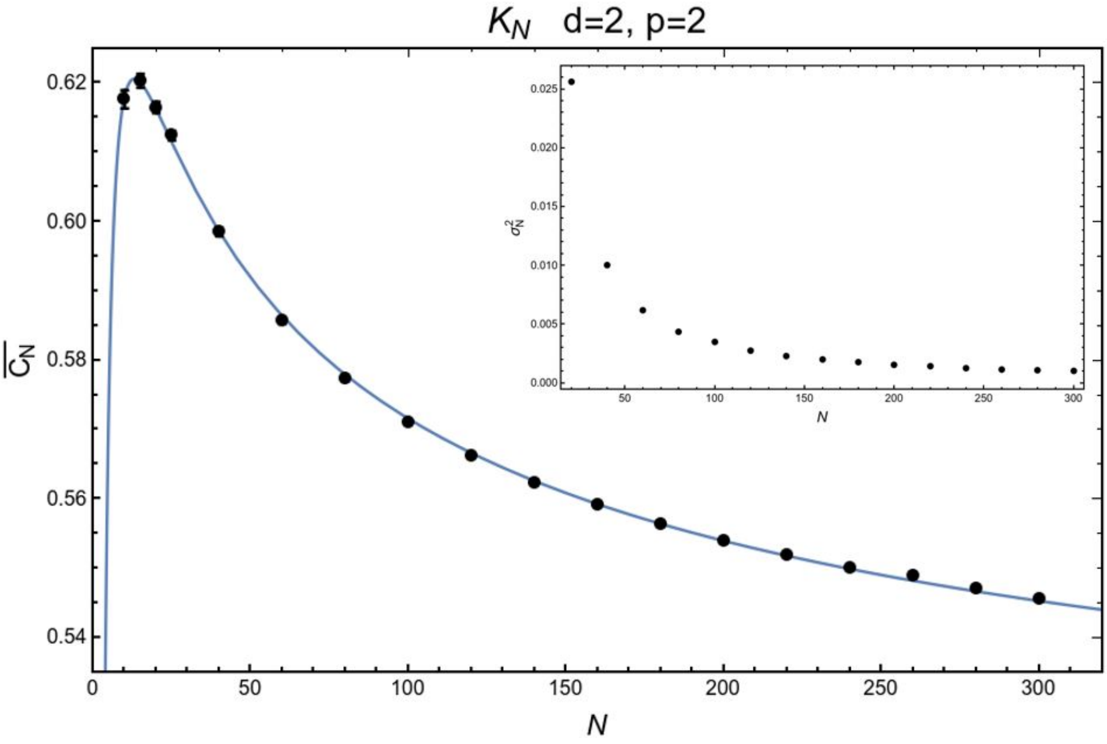
(e) $n = 1000 \qquad f(N) = -0.3259 + \frac{1.5761}{N}$



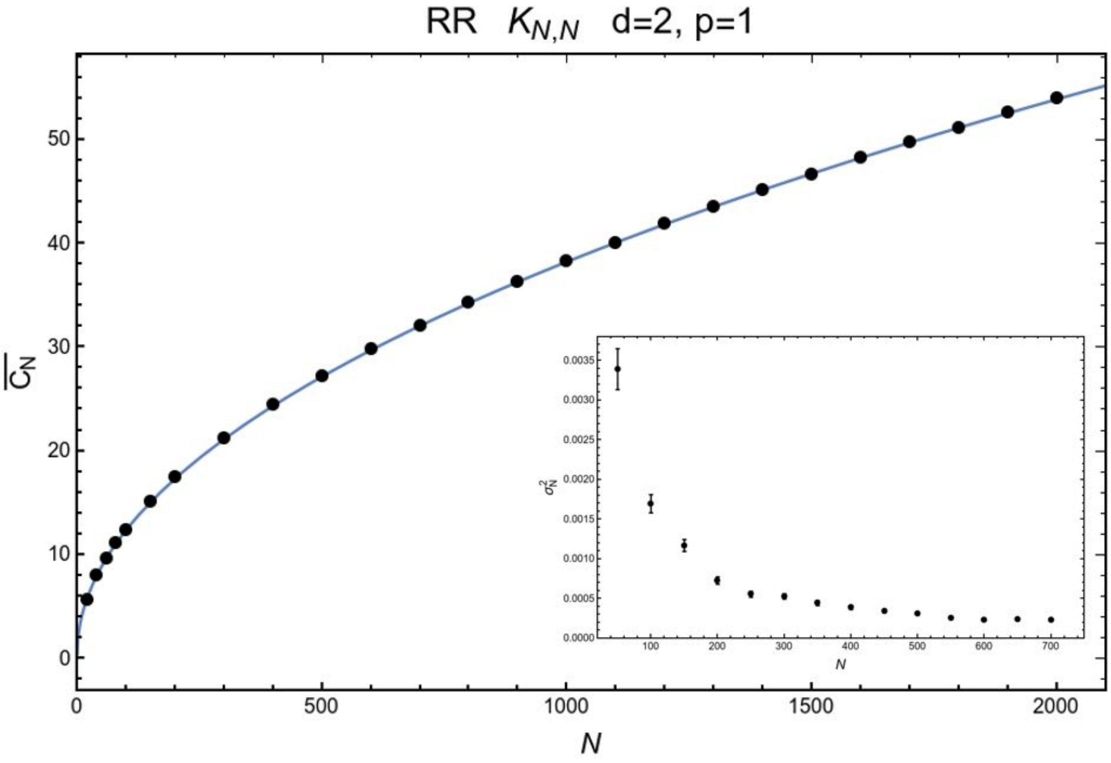
(f) $n = 1000$ $\qquad f(N) = \frac{1.9844}{N}$



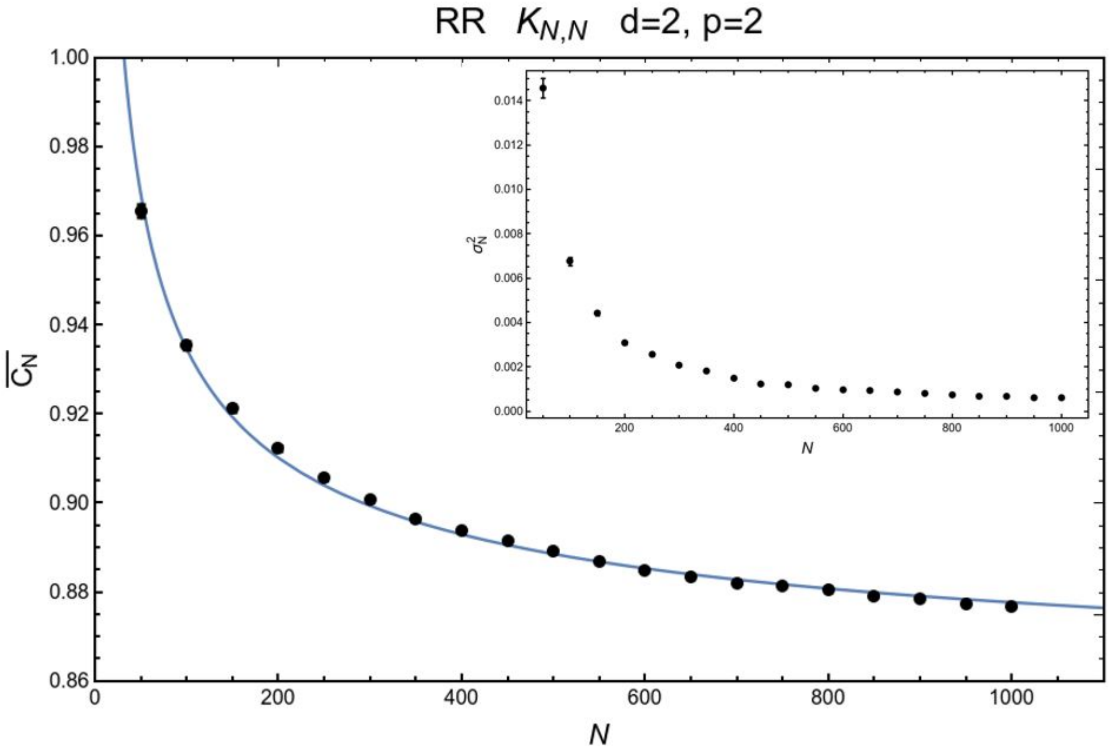
(g) $n = 2000$ $\qquad f(N) = 0.1565 + 0.6578\sqrt{N}$



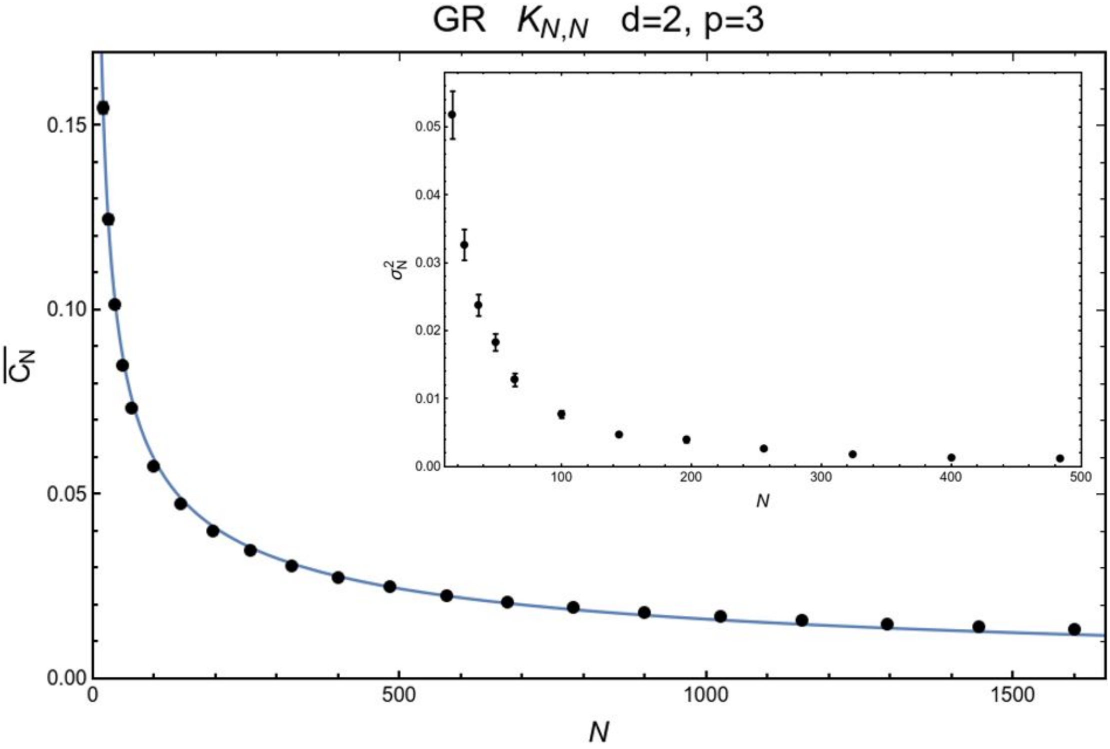
(h) $n = 8000$ $\qquad f(N) = 0.5003 + \frac{0.8702}{\sqrt{N}} - \frac{1.5736}{N}$



(i) $n = 1000$ $\qquad f(N) = 0.4405 + 1.1947\sqrt{N}$



(j) $n = 5000$ $\qquad f(N) = 0.8515 + \frac{0.8286}{\sqrt{N}}$



(k) $n = 1000$ $\qquad f(N) = \frac{0.6301}{\sqrt{N}}$

3.4.1 Comparison with the matching problem and the TSP

The last two sections provided us some analytical and numerical results on which some interesting considerations can be formulated. The final part of this thesis is thus devoted to the analysis of these results, concerning the average cost scaling and the rescaled variance of the cost for the random MST problem in one and two dimensions.

Looking to the plots of the previous section, the thing that stands out from our numerical study is the fact that the average cost of our problem scales for large N in the same way both in the monopartite and bipartite case. This is corroborated for $p = d = 1$ also by our exact solution of the Euclidean MST given in Eqs. (3.38) and (3.58). In practice, the asymptotic behavior provided by Theorem 3.2.1, i.e.

$$\mathcal{C}_{N,d}^{(p,\mathrm{rEm})} \sim N^{1-\frac{p}{d}} \quad \text{as} \quad N \to \infty \tag{3.63}$$

is correct in one and two dimensions also when one considers the MST problem defined on the complete bipartite graph $\mathtt{K}_{N,N}$. As we have already mentioned in Sect. 3.2, this property of the Euclidean MST is quite unexpected considering that in the TSP and in the matching problem the opposite holds. In fact, in the case of the Euclidean assignment (rEa) the scaling of the average optimal cost is known in every dimensions and for every $p > 1$ [65]

$$\overline{\mathcal{C}_{N,d}^{(p,\mathrm{rEa})}} \sim \begin{cases} N^{1-\frac{p}{2}} & d = 1 \\ N^{1-\frac{p}{2}}(\log N)^{\frac{p}{2}} & d = 2 \\ N^{1-\frac{p}{d}} & d > 2. \end{cases} \tag{3.64}$$

Remembering that on the complete monopartite graph $\mathtt{K}_N$ the scaling of the matching problem is $N^{1-\frac{p}{d}}$ for all d, it is clear that an anomalous behavior appears in low dimensions on $\mathtt{K}_{N,N}$. The same happens for the Euclidean TSP, whose average optimal cost in one dimension is twice that of the assignment [66], and thus its scaling is again anomalous w.r.t. the monopartite case.

In general, the fact that in one and two dimensions the scaling does not change for the Euclidean bipartite MST means that the random fluctuations of the positions of the points are irrelevant. This can be understood considering the first part of the construction of the MST as explained in Sect. 3.3 for the one-dimensional case, but valid in general in higher dimensions too. The crucial element is that when two different sets of points are present, a Voronoi diagram arises from one of them, and each point of the other set connects with its closer seed. This allows the presence of vertices in the final subgraph with degree greater than two, differently to what happens in the matching and the TSP, where such configurations are forbidden. Thus, from the perspective of disordered systems, we can say in a sense that the Euclidean bipartite MST is less frustrated than its two classical counterparts, and this reflects on the asymptotic behavior of the average cost for large sizes of the system.

Let us observe that the above discussion is confirmed also by the analysis of the rescaled variance of the cost, depicted in the inset of each of the previous plots. In fact, recalling the definition given in Eq. (1.23), in all the cases examined the cost of the MST proves to be a self-averaging quantity, as one expects when fluctuations do not matter in the thermodynamic limit. Again, this represents a crucial difference w.r.t. the TSP and the matching problem. In fact, in both these cases it has been shown that the average optimal cost is not self-average in $d = 1$ for the bipartite setting [66, 67], while it is self-average in all dimensions when the problem is defined on $\mathtt{K}_N$ [59].

We conclude by saying that if an anomalous behavior in the asymptotic scaling is present, as in the cases of the TSP and the matching problem, it emerges in low dimensions due to the major Euclidean constraints imposed by the structure of the space. For this reason, even if we have not carried out a direct analysis, we expect that what we have found for the Euclidean MST in one and two dimensions continues to hold even in $d \geq 3$, making our arguments valid for all dimensions.

Chapter 4

Conclusions and perspectives

In this thesis we have analyzed the random MST problem both in the purely random case, namely when the edge weights are independent and identically distributed random variables, and in the Euclidean setting, where correlations are present. For both these situations, by considering the results obtained and the approaches used, we can make some interesting final considerations.

In the first case, we have shown in Sect. 2.4 that it is possible to set up a replica calculation for the MST problem on the complete graph, remarkably deriving the partition function for our system as a particular limit of the Potts model. This is certainly an important starting point for an innovative kind of analysis of the random MST problem, considering the effectiveness displayed by the replica method in the study of the matching problem, not only in the purely random case but also in the correlated one. As we have seen, a problem emerges when one performs the configurational average over the disorder, because an immediate decoupling of different sites of the graph appears impossible. However, we illustrated an interesting general technique consisting in a functional Hubbard-Stratonovich transformation that potentially could solve our problem. This method has proven very useful in the study of random matrices in the case of bosonic variables, so it would be very interesting to investigate further whether the same can hold when the problem is formulated with Grassmann variables, as it happens for the random MST.

Regarding the random Euclidean MST problem, one fundamental result of this thesis is surely the explicit solution of the one dimensional bipartite case for $p > 0$ obtained in Sect. 3.3. Unfortunately, in the case $p \neq 1$ we have not been able to provide a closed formula for the average cost, nor to write down an asymptotic expansion for large N for the finite sums which appear in it. This remains an open problem, especially considering the fact that numerical simulations show that for $p = 1$ all the sums diverge, and so they have to cancel each other in the thermodynamic limit. It is possible that a completely novel approach is necessary in order to obtain closed formulas for general values of $p > 0$.

The second fundamental result of our work for the random Euclidean MST has been obtained both with our explicit solution and with a numerical investigation performed in one and two dimensions in Sect. 3.4. This can be formulated simply by saying that random fluctuations do not play any role in the scaling behavior of the average cost in the bipartite setting, being it equal to the one of the monopartite problem. It would be extremely interesting to consider in more detail this fact, for example by computing explicitly the second moment of the cost in the one dimensional case, or by studying directly the two dimensional problem. In particular for the latter case, we expect that something interesting can be obtained for the average properties of the MST by exploiting its relation with the elements of computational geometry introduced in Sect. 1.2.2, namely the Voronoi diagram and the Delaunay triangulation.

Appendix A

Grassmann-Berezin calculus

An N-dimensional Grassmann algebra on $\mathbb{R}$ or $\mathbb{C}$ is the algebra generated by a set of variables $\{\psi_i\}_{i=1}^N$ satisfying the anticommutation relation

$$\{\psi_i, \psi_j\} = 0 \qquad \forall i, j. \tag{A.1}$$

which implies in particular, $\psi_i^2 = 0 \ \forall i$. The most general function that one can write in this algebra has the form

$$\begin{aligned} f(\psi) &= f^{(0)} + \sum_i f^i \psi_i + \sum_{i<j} f^{ij} \psi_i \psi_j + \sum_{i<j<k} f^{ijk} \psi_i \psi_j \psi_k + \dots \\ &= \sum_{0 \le k \le N} \frac{1}{k!} \sum_{\{i\}} f^{i_1, \dots, i_N} \psi_{i_1} \psi_{i_2} \cdots \psi_{i_k}, \end{aligned} \tag{A.2}$$

the coefficients being antisymmetric tensors with k indices, each ranging from 1 to N. Since there are $\binom{N}{k}$ such linearly independent tensors, summing over k from 0 to N yields a 2^N-dimensional algebra. The above expansion can be written also in an alternative form, which highlights the relation with Fermi statistics

$$f(\psi) = \sum_{a_i = 0,1} f_{a_1, a_2, \dots, a_N} \psi_1^{a_1} \psi_2^{a_2} \cdots \psi_N^{a_N}. \tag{A.3}$$

The integers $a_i = 0, 1$ can be tought of as occupation numbers of *states* described by ψ_i. Note that in this setting one refers to the nilpotency property $\psi_i^2 = 0 \ \forall i$ as the Pauli exclusion principle.

Thanks to the anticommuting rule we can define an associative product

$$\begin{aligned} f(\psi) g(\psi) =& f^{(0)} g^{(0)} + \sum_i \left(f^{(0)} g^i + f^i g^{(0)} \right) \psi_i + \\ &+ \frac{1}{2} \sum_{ij} \left(f^{ij} g^{(0)} + f^i g^j - f^j g^i + f^{(0)} g^{ij} \right) \psi_i \psi_j + \dots \end{aligned} \tag{A.4}$$

but in general $fg \neq \pm gf$. Nevertheless, the subalgebra containing terms with an even number of ψ variables commutes with any element f.

In the Grassmann algebra we can define a left and a right derivative $\partial / \partial \psi_i$. When applied to a monomial containing the variable ψ_i, this is moved to the left with the appropriate sign due to the exchanges, and then suppressed. In the case where ψ_i is not present, the result is simply zero. The described operation can be extended by linearity to any element of the Grassmann algebra. From the given definition, one can prove the following rules

$$\left\{\frac{\partial}{\partial\psi_i},\frac{\partial}{\partial\psi_j}\right\}=0 \qquad \left\{\frac{\partial}{\partial\psi_i,\psi_j}\right\}=\delta_{ij}. \tag{A.5}$$

The so called Berezin integral is defined as a linear operation exactly in the same way as the (left) derivative, so we have

$$\int \mathrm{d}\psi f(\psi)=\frac{\partial f}{\partial\psi} \qquad \int \mathrm{d}\psi_2\mathrm{d}\psi_1 f(\psi)=\frac{\partial}{\partial\psi_2}\frac{\partial}{\partial\psi_1}f(\psi) \qquad \dots \tag{A.6}$$

In general, given a permutation $\pi\in\mathcal{S}_N$, it holds

$$\int \mathrm{d}\psi_{\pi(N)}\mathrm{d}\psi_{\pi(N-1)}\cdots\mathrm{d}\psi_{\pi(1)}f(\psi)=\epsilon(\pi)\int \mathrm{d}\psi_N\mathrm{d}\psi_{N-1}\cdots\mathrm{d}\psi_1 f(\psi), \tag{A.7}$$

where $\epsilon(\pi)$ denotes the signature of the permutation.

The integral operation as we have defined it satisfies the constraint of translational invariance, which requires

$$\int \mathrm{d}\psi 1=0 \qquad \int \mathrm{d}\psi\psi=1. \tag{A.8}$$

If a non-singular linear tranformation of the form $\chi_i=\sum_{j=1}^N A_{ij}\psi_j$ is applied, due to the anticommuting structure of the Grassmann algebra one obtains the following relation

$$\int \mathrm{d}\psi_N\psi_{N-1}\cdots\psi_1 f(\psi)=\det A\int \mathrm{d}\chi_N\mathrm{d}\chi_{N-1}\cdots\mathrm{d}\chi_1 F(\chi), \tag{A.9}$$

having set $f(\psi)=F(\chi)$. Let us note that in normal integration such a change of coordinates produces on the right hand side the factor $|\det A|^{-1}$.

Let us now consider a 2^{2N}-dimensional Grassmann algebra comprising two independent sets of generators $\{\psi_i\}_{i=1}^N$ and $\{\bar{\psi}_i\}_{i=1}^N$, with $\bar{\psi}_j$ that can be regarded as the complex conjugate of ψ_i. Together with Eqs. (A.1), (A.5) for both the variable sets, the following relations hold

$$\{\psi_i,\bar{\psi}_j\}=0 \qquad \left\{\frac{\partial}{\partial\psi_i},\bar{\psi}_j\right\}=0 \qquad \left\{\frac{\partial}{\partial\bar{\psi}_i},\psi_j\right\}=0. \tag{A.10}$$

With the ingredients introduced we can prove the fundamental result

$$\int \mathcal{D}_N\left(\psi,\bar{\psi}\right)\exp\left(\sum_{i,j=1}^N\bar{\psi}_i A_{ij}\psi_j\right)=\det A, \tag{A.11}$$

where we have defined the shorthand $\mathcal{D}_N\left(\psi,\bar{\psi}\right)=\prod_{i=1}^N \mathrm{d}\psi_i\mathrm{d}\bar{\psi}_i$. First, we perform the change of variables $\chi_i=\sum_{j=1}^N A_{ij}\psi_j$, obtaining

$$\int \mathcal{D}_N\left(\psi,\bar{\psi}\right)\exp\left(\sum_{i,j=1}^N\bar{\psi}_i A_{ij}\psi_j\right)=\det A\int \mathcal{D}_N\left(\chi,\bar{\psi}\right)\exp\left(\sum_{i=1}^N\bar{\psi}_i\chi_i\right), \tag{A.12}$$

then we observe that due to nilpotency of the Grassmann variables

$$\exp\left(\bar{\psi}_i \chi_i\right) = 1 + \bar{\psi}_i \chi_i, \tag{A.13}$$

so that

$$\int \mathcal{D}_N\left(\chi, \bar{\psi}\right) \exp\left(\sum_{i=1}^{N} \bar{\psi}_i \chi_i\right) = \int \mathcal{D}_N\left(\chi, \bar{\psi}\right) \prod_{i=1}^{N} \left(1 + \bar{\psi}_i \chi_i\right). \tag{A.14}$$

Properties (A.8) guarantees that the integral is non zero only if the integrand contains every variable appearing in the integration measure. Therefore, in the product expansion, only the term $\bar{\psi}_1 \chi_1 \cdots \bar{\psi}_N \chi_N$ contributes to the result. Moreover, the variables order in the integration fixes the factor of this term to $+1$, proving (A.11).

The above result can be extended to expectation values of monomials. If we denote by $A(I|J)$ the submatrix obtained from A deleting the rows $I = (i_1, 1_2, \ldots, i_k)$ and columns $J = (j_1, j_, \ldots, j_k)$ the following general formula can be proven

$$\int \mathcal{D}_N\left(\psi, \bar{\psi}\right) \bar{\psi}_{i_1} \psi_{j_1} \cdots \bar{\psi}_{i_k} \psi_{j_k} \exp\left(\sum_{i,j=1}^{N} \bar{\psi}_i A_{ij} \psi_j\right) = \epsilon(I|J) \, \mathrm{det} A(I|J). \tag{A.15}$$

Indeed, the presence of ψ_k (respectively $\bar{\psi}_k$) annihilates the contribution of terms of the form $\bar{\psi}_i A_{ik} \psi_k$ $(\bar{\psi}_k A_{kj} \psi_j)$, and $\epsilon(I|J) = \pm 1$ accounts for the number of interchanges needed to order the variables before the integration.

Appendix B

Order statistics

Let $X_1, \dots, X_N$ be a sample of i.i.d. random variables generated according to a probability distribution density ρ and with cumulative distribution function given by

$$\Phi(t) = \Pr\left(X_i < t\right) = \int \rho(t)\theta(t-y), \tag{B.1}$$

where $\theta(x)$ is the Heaviside step function. If one considers the ordered sample $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(N)}$ the element $X_{(k)}$ is called *k-th order statistic.*

The joint distribution of all order statistics can be easily derived from the joint distribution of the random variables in exam. Denoting the interval $(x, x + \mathrm{d}x)$ simply by $\mathrm{d}x$, the probability for our random sample to be $X_1 \in \mathrm{d}x_1, \dots, X_N \in \mathrm{d}x_N$, thanks to independence of the X_i's, is given by

$$\Pr\left(X_1 \in \mathrm{d}x_1, \dots, X_N \in \mathrm{d}x_N\right) = \prod_{i=1}^{N} \rho(x_i)\,\mathrm{d}x_i. \tag{B.2}$$

Now since the order does not matter, we have $N!$ possible permutations of the random variables X_i, so the joint distribution for all order statistics reads

$$\Pr\left(X_{(1)} \in \mathrm{d}x_1, \dots, X_{(N)} \in \mathrm{d}x_N\right) = N! \prod_{i=1}^{N} \rho(x_i)\,\mathrm{d}x_i. \tag{B.3}$$

Two special cases of the order statistics are the minimum and the maximum of a given random sample

$$x := X_{(1)} = \min_{i \in [N]} X_i \qquad\qquad X := X_{(N)} = \max_{i \in [N]} X_i \tag{B.4}$$

where $[N] \equiv \{1, \dots, N\}$. Considering e.g. the random variable x, its distribution can be computed by asking which is the probability for an element X_k of the random sample to be smaller than all the others

$$\rho_x(t)\,\mathrm{d}t = \Pr\left(x \in \mathrm{d}t\right) = \Pr\left(\exists!\, k : X_k \in \mathrm{d}t, X_i \geq t \;\; \forall i \neq k\right). \tag{B.5}$$

Due to the fact that the smallest element can be chosen indifferently among N, and exploiting independency of the random variables we have

$$\rho_x(t)\,\mathrm{d}t = N \Pr\left(X_1 \in \mathrm{d}t, X_i \geq t \;\; \forall i \neq k\right) = N \Pr\left(X_1 \in \mathrm{d}t\right) \prod_{i \neq 1} \Pr\left(X_i \geq t\right). \tag{B.6}$$

Proceeding in a similar way for the random variable X in the end we obtain

$$\rho_x(t)\,\mathrm{d}t = N[1-\Phi(t)]^{N-1}\rho(t)\,\mathrm{d}t \tag{B.7a}$$

$$\rho_X(t)\,\mathrm{d}t = N[\Phi(t)]^{N-1}\rho(t)\,\mathrm{d}t. \tag{B.7b}$$

With the same considerations used above, we can compute the probability distribution of the k-th order statistic too. The only difference lies in the fact that we have to choose arbitrarily $k-1$ samples out of $N-1$ that are smaller of $X_{(k)}$. Denoting by $\mathcal{S}_N$ the set of permutation of N elements we have

$$\begin{aligned}
\rho_{X_{(k)}}(t)\,\mathrm{d}t &= \Pr\left(X_{(k)} \in \mathrm{d}t\right) \\
&= \Pr\left(\exists\sigma\in\mathcal{S}_N : X_{\sigma(1)}\in\mathrm{d}t, X_{\sigma(2)}\le t,\dots,X_{\sigma(k)}\le t, X_{\sigma(k+1)}\ge t,\dots\right) \\
&= N\binom{N-1}{k-1}\Pr\left(X_1\in\mathrm{d}t, X_2\le t,\dots,X_k\le t, X_{k+1}\ge t,\dots,X_N\ge t\right) \\
&= N\binom{N-1}{k-1}\Pr\left(X_1\in\mathrm{d}t\right)\left[\prod_{i=2}^{k}\Pr\left(X_i\le t\right)\right]\left[\prod_{j=k+1}^{N}P(X_j\ge t)\right] \\
&= N\binom{N-1}{k-1}[\Phi(t)]^{k-1}[1-\Phi(t)]^{N-k}\rho(t)\mathrm{d}t
\end{aligned} \tag{B.8}$$

An identical procedure allows us to derive also the joint distribution for the two order statistics $X_{(p)}$, $X_{(q)}$ with $p<q$. In this case, for $t<s$ the result is

$$\begin{aligned}
\rho_{X_{(p)},X_{(q)}}(t,s)\,\mathrm{d}t\mathrm{d}s &= \Pr\left(X_{(p)})\in\mathrm{d}t, X_{(q)}\in\mathrm{d}s\right) \\
&= \frac{N!}{(p-1)!\,(q-p-1)!\,(N-q)!}[\Phi(t)]^{p-1}[\Phi(s)-\Phi(t)]^{q-p-1} \\
&\quad\cdot[1-\Phi(s)]^{N-q}\rho(t)\rho(s)\,\mathrm{d}t\mathrm{d}s.
\end{aligned} \tag{B.9}$$

Let us specialize the obtained results in the case of a random sample generated with uniform distribution on the interval $[0,1]$. The probability distribution density ρ and the cumulative distribution function Φ are given by

$$\rho(t)=\theta(t)\theta(1-t), \qquad \Phi(t)=t\,\theta(t)\theta(1-t). \tag{B.10}$$

Substituting into the Eqs. (B.7b), (B.8) and (B.9) we obtain

$$\rho_x(t) = N(1-t)^{N-1}\theta(t)\theta(1-t) \tag{B.11}$$

$$\rho_{X_{(k)}}(t) = N\binom{N-1}{k-1}t^{k-1}(1-t)^{N-k}\theta(t)\theta(1-t) \tag{B.12}$$

$$\begin{aligned}
\rho_{X_{(p)},X_{(q)}}(t,s) =& \frac{N!}{(p-1)!\,(q-p-1)!\,(N-q)!}t^{p-1}(s-t)^{q-p-1}(1-s)^{N-q} \\
&\cdot\theta(t)\theta(s-t)\theta(1-s).
\end{aligned} \tag{B.13}$$

Acknowledgments

First of all, I would like to thank my supervisor, Professor Sergio Caracciolo, for his patient guidance and useful advices. It was a privilege for me to share his exceptional scientific knowledge during these months.

Secondly, a special praise for Enrico Malatesta, definitely the best tutor one could hope for.

I thank my "experimental" fellows, Edoardo and Stefano, for all the fun we have had in the last six competitive years.

My deep and sincere gratitude goes to my family, for their unconditional support and encouragement during all these long years.

Last but not least, a special thank to a special person, Elisabetta, for her endless patience and fundamental help.

References

[1] R. Diestel. *Graph Theory.* Springer Graduate Texts in Mathematics (GTM). 2012.

[2] C. Berge. *Graphs and Hypergraphs.* North-Holland mathematical library. Amsterdam, 1973.

[3] G. Kirchhoff. "Über die Auflösung der Gleichungen, auf welche man bei der Untersuchung der linearen Verteilung galvanischer Ströme gefürht wird". *Annual Review of Physical Chemistry* 72 (1847), pp. 497–508.

[4] N. Hartsfield, G. Ringel. *Pearls in Graph Theory: A Comprehensive Introduction.* Academic Press. 1994.

[5] C. Papadimitriou. *Computational Complexity.* Theoretical computer science. Addison-Wesley, 1994.

[6] E. Lawler, D. Shmoys, A. Kan, J. Lenstra. *The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization.* John Wiley & Sons, Incorporated, 1985.

[7] H. W. Kuhn. "The Hungarian method for the assignment problem". *Naval research logis- tics quarterly* 2 (1955), pp. 83–97.

[8] E. Horowitz, S. Sahni. *Fundamentals of Computer Algorithms.* Computer Science Press, 1984.

[9] J. B. Kruskal. "On the shortest spanning tree of a graph and the traveling salesman problem". *Proceedings of the American Mathematical Society* 7 (1956), pp. 48-50.

[10] R. C. Prim. "Shortest connection networks and some generalizations". *Bell System Technical Journal* 36 (1957), pp. 1389-1401.

[11] J. Nešetřil, E. Milková, H. Nešetřilova. "Otakar Borůvka on minimum spanning tree problem. Translation of both the 1926 papers, comments, history". *Discrete Mathematics* 233 (2001), pp. 3-36.

[12] P. Willet. "Recent trends in hierarchic document clustering: a critical review". *Information Processing & Management* 24(5) (1988), pp. 577-597.

[13] P. J. Wan et al. "Minimum-energy broadcast routig in static ad hoc wireless networks". *IEEE Infocom* (2001).

[14] M. B. Eisen et al. "Cluster analysis and display of genome-wide expression patterns". *PNAS* 95(25) (1998), pp. 14863-14868.

[15] S. P. Bhavsar, R. J. Splinter. "The superiority of the minimal spanning tree in percolation analyses of cosmological data sets". *MNRAS* 282 (1996), pp. 1461-1466.

[16] S. Subramaniam, S. B. Pope. "A mixing model for turbulent reactive flows based on Euclidean minimum spanning trees". *Combustion and Flame* 115(4) (1998), pp. 487-514.

[17] M. Held, R. M. Karp. "The traveling-salesman problem and minimum spanning trees". *Operations Research*, 18(6) (1970), pp. 1138-1162.

[18] K. J. Suppowit, D. A. Plaisted, E. M. Reingold. "Heuristics for weighted perfect matching". *Proceedings of the 12th Annual ACM Symposium on Theory of Computing* (1980), pp. 398-419.

[19] M. de Berg, M. Kreveld, M. Overmars, O. Schwarzkopf. *Computational Geometry: Algorithms and Applications.* Springer, 1997.

[20] B. Chazelle. "A minimum spanning tree algorithm with inverse-ackermann typer complexity". *Journal of the ACM* 47(6) (2000), 1028-1047.

[21] M. Shamos, D. Hoey. "Closest-point problems". *Proceedings of the 16th Annual Symposium on Foundations of Computer Science* (1975), pp. 151-162.

[22] P. Agarwal et al. "Euclidean minimum spanning trees and bichromatic closest pairs". *Discrete Computational Geometry* 6(5) (1991), pp. 407-422.

[23] A. Biniaz et al. "Spanning trees in multipartite geometric graphs". *Algorithmica* 80(11) (2018), pp. 3177-3191.

[24] M. Mézard, A. Montanari. *Information, Physics, and Computation.* Oxford Graduate Texts. OUP Oxford, 2009.

[25] M. Mézard, G. Parisi. "Replicas and optimization". *Journal de Physique Lettres* 46(17) (1985), pp. 771-778.

[26] M. Mézard, G. Parisi. "On the solution of the random link matching problems". *Journal de Physique* 48(9) (1987), pp. 1451-1459.

[27] D. J. Aldous. "The $\zeta(2)$ limit in the random assignment problem". *Random Structures and Algorithms* 2 (2001), pp. 381-418.

[28] M. Mézard, G. Parisi, M. Virasoro. *Spin Glass Theory and Beyond.* Lecture Notes in Physics Series. World Scientific Publishing Company, Incorporated, 1987.

[29] T. Castellani, A. Cavagna. "Spin-glass theory for pedestrians". *Journal of Statistical Mechanics: Theory and Experiment* (2005).

[30] H. Nishimori. *Statistical Physics of Spin Glasses and Information Processing. An Introduction.* Clarendon Press Oxford, 2001.

[31] A. Frieze. "On the value of a random minimum spanning tree problem". *Discrete Applied Mathematics* 10(1) (1985), pp. 47-56.

[32] J. Steele. "On Frieze's $\zeta(3)$ limit for lengths of minimal spanning trees". *Discrete Applied Mathematics* 18(1) (1987), pp. 99-103.

[33] A. Frieze, C. McDiarmid. "On random minimum length spanning trees". *Combinatorica* 9(4) (1989), pp. 363-374.

[34] S. Janson. "The minimal spanning tree in a complete graph and a functional limit theorem in a random graph". *Random Structures and Algorithms* 7(4) (1995), pp. 337-355.

[35] J. Wästlund. "Evaluation of Janson's constant for the variance in the random minimum spanning tree problem". *Linköping studies in mathematics* 7 (2005).

[36] D. Coppersmith, G. B. Sorkin. "Constructive bounds and exact expectations for the random assignment problem". *Random Sttructures & Algorithms* 15(2) (1999), pp. 113-144.

[37] G. Parisi. "A conjecture on random bipartite matching". *arXivorg* cond-mat/9801176 (1998).

[38] C. Cooper, A. Frieze. "On the length of a random minimum spanning tree". *Combinatorics, Probability and Computing* 25(1) (2016), pp. 89-107.

[39] I. Karatzas, S. E. Shreve. *Brownian Motion and Stochastic Calculus.* Springer, 1998.

[40] J. M. Steele. "Minimal spanning trees for graphs with random edge lengths". *Mathematics and Computer Science II: Algorithms, Trees, Combinatorics and Probabilities*, Birkhäuser, pp. 223-245.

[41] I. M. Gessel, B. E. Sagan. "The Tutte polynomial of a graph, depth-first search, and simplicial complex partitions". *Electronic Journal of Combinatorics* 3(2) (1996)

[42] W. V. Li, X. Zhang. "On the difference of expected lengths of minimum spanning trees". *Combinatorics Probability and Computing* 18(3) (2009), pp. 423-434.

[43] F. Y. Wu. "The Potts model". *Review of Modern Physics* 54 (1982), pp. 235-265.

[44] J. B. Kogut. "An introduction to lattice gauge theory and spin systems". *Review of Modern Physics* 51(4) (1979), pp. 659-713.

[45] J. Ashkin, E. Teller. "Statistics of two-dimensional lattices with four components". *Physical Review* 64(5-6) (1943), pp. 178-184.

[46] C. M. Fortuin, P. W. Kasteleyn. "On the random-cluster model. I. Introdution and relation to other models". *Physica* 57(4) (1972). pp. 536-564.

[47] J. L. Jacobsen, J. Salas, D. Sokal. "Spanning forests and the q-state Potts model in the limit $q \to 0$". *Journal of Statistical Physics* 119(5-6) (2005), pp. 1153-1281.

[48] S. Caracciolo, J. L. Jacobsen, H. Saleur, A. D. Sokal, A. Sportiello. "Fermionic field theory for trees and forests". *Physical Review Letters* 93:080601 (2004)

[49] C. N. Yang, T. D. Lee. "Statistical theory of equations of state and phase transitions. I. Theory of condensation". *Physics Review* 87(3) (1952), pp. 404-409.

[50] D. Dhar. "The abelian sandpile and related models". *Physica A. Statistical Mechanics and its Applications* 263 (1999), pp. 4-25.

[51] S. N. Majumdar, D. Dhar. "Equivalence between the abelian sandpile model and the $q \to 0$ limit of the Potts model". *Physica A. Statistical Mechanics and its Applications* 185 (1992), pp. 129-145.

[52] J. W. Moon. "Counting labelled trees". *Canadian Mathematical Congress*, Montreal, 1970.

[53] H. N. V. Temperley. "The transition of the rigid-sphere gas". Proceedings of the Physical Society 84(3) (1964), pp. 339-344.

[54] D. Elderfield, D. Sherrington. "The curious case of the Potts spin glass". *Journal of Physics C: Solid State Physics* 16(15) (1983), pp. 497-503.

[55] S. Caracciolo, M. P. D'Achille, E. M. Malatesta, G. Sicuro. "Finite-size corrections in the random assignment problem". *Physical Review E* 95(5) (2017)

[56] A. J. Bray, G. J. Rodgers. "Diffusion in a sparsely connected space: a model for glassy relaxation". *Physical Review B* 38(16) (1988), pp. 11461-11470.

[57] Y. V. Fyodorov. "Spectral properties of random reactance networks and random matrix pencils". *Journal of Physics A: Mathematical and General* 32 (1999), pp. 7429-7446.

[58] G. M. Cicuta, H. Orland. "Real symmetric random matrices and replicas". *Physical review E* 74:051120 (2006).

[59] J. Yukich. *Probability theory of classical Euclidean optimization problems*. Lecture notes in mathematics. Springer, 1998.

[60] C. Redmond. J. Yukich. "Asymptotics for Euclidean functionals with power-weighted edges". *Stochastic Processes and their Applications* 61(2) (1996), pp. 289-304.

[61] J. M. Steele. "Growth rates of Euclidean minimal spanning trees with power weighted edges". *Annals of Probability* 16(4) (1988), pp. 1767-1787.

[62] J. M. Steele. "Subadditive Euclidean functionals and nonlinear growth in geometric probability". *Annals of Probability* 9 (1981), pp. 365-376.

[63] J. E. Yukich. "Asymptotics for weighted minimal spanning trees on random points". *Stochastic Processes and their Applications* 85 (2000), pp. 123-138.

[64] F. Avram, D. Bertsimas. "The minimum spanning tree constant in geometrical probability and under the independent model: a unified approach". *Annals of Applied Probability* 2(1) (1992), pp. 113-130.

[65] S. Caracciolo, C. Lucibello, G. Parisi, G. Sicuro. "Scaling hypothesis for the Euclidean bipartite matching problem". *Physical Review E* 90(1) (2014), p. 012118.

[66] S. Caracciolo, A. Di Gioacchino, M. Gherardi, E. M. Malatesta. "Solution for a bipartite Euclidean traveling-salesman problem in one dimension". *Physical Review E* 97(5) (2018), p. 052109.

[67] J. Houdayer, J. H. Boutet de Monvel, O. C. Martin. "Comparing mean field and Euclidean matching problems". *The European Physical Journal B - Condensed Matter and Complex Systems* 6(3) (1998), pp. 383-393.

[68] J. R. de Almeida, D. J. Thouless. "Stability of the Sherrington-Kirkpatrick solution of a spin glass model". *Journal of Physics A: Mathematical and General* 11(5) (1978).

[69] G. Parisi. "A sequence of approximated solutions to the S-K model for spin glasses". *Journal of Physics A: Mathematical and General* 13(4) (1980), p. L115.

[70] F. Guerra, F. L. Toninelli. "The thermodynamic limit in mean field spin glass models". *Communications in Mathematical Physics* 230(1) (2002), p. 71.

[71] M. Talagrand. "The Parisi formula". *Annals of Mathematics* (2006), p. 221.

[72] S. Caracciolo, G. Sicuro. "Quadratic stochastic Euclidean bipartite matching problem". *Physical Review Letters* 115 (2015) p. 230601.

[73] E. Boniolo, S. Caracciolo, A. Sportiello. "Correlation function for the grid-Poisson Euclidean matching on a line and on a circle". *Journal of Statistical Mechanics: Theory and Experiment* (2014), p. 11023.

[74] S. Caracciolo, A. Di Gioacchino, E. M. Malatesta, C. Vanoni. "Average optimal cost for the Euclidean TSP in one dimension". *arXivorg* 1811.08265 (2018).

[75] R. Capelli, S. Caracciolo, A. Di Gioacchino, E. M. Malatesta. "Exact value for the average optimal cost of the bipartite traveling salesman and two-factor problems in two dimensions". *Physical Review E* 98 (2018), p. 030101.