Fast computation of exact solutions of generic and degenerate assignment problems

Patrice Koehl [1] and Henri Orland [2]
[1]*Department of Computer Science and Genome Center, University of California, Davis, California 95616, USA*
[2]*Institut de Physique Théorique, Université Paris-Saclay, CNRS, CEA, 91191 Gif/Yvette Cedex, France*

(Received 2 December 2020; accepted 1 March 2021; published 2 April 2021)

The linear assignment problem is a fundamental problem in combinatorial optimization with a wide range of applications, from operational research to data science. It consists of assigning "agents" to "tasks" on a one-to-one basis, while minimizing the total cost associated with the assignment. While many exact algorithms have been developed to identify such an optimal assignment, most of these methods are computationally prohibitive for large size problems. In this paper, we propose an alternative approach to solving the assignment problem using techniques adapted from statistical physics. Our first contribution is to fully describe this formalism, including all the proofs of its main claims. In particular we derive a strongly concave effective free-energy function that captures the constraints of the assignment problem at a finite temperature. We prove that this free energy decreases monotonically as a function of β, the inverse of temperature, to the optimal assignment cost, providing a robust framework for temperature annealing. We prove also that for large enough β values the exact solution to the generic assignment problem can be derived using simple roundoff to the nearest integer of the elements of the computed assignment matrix. Our second contribution is to derive a provably convergent method to handle degenerate assignment problems, with a characterization of those problems. We describe computer implementations of our framework that are optimized for parallel architectures, one based on CPU, the other based on GPU. We show that the latter enables solving large assignment problems (of the orders of a few 10 000s) in computing clock times of the orders of minutes.

DOI: 10.1103/PhysRevE.103.042101

I. INTRODUCTION

Imagine that there are N flour milling plants around Paris, France that serve N bakeries within Paris, and let us assume balance, namely, that there is as much flour produced by one plant as needed by one bakery. A company in charge of the distribution of the flour will take into account the individual cost of transporting flour from one plant to one bakery to find an "optimal distribution plan," namely, an assignment of an exclusive flour milling plant to each bakery that leads to a minimal overall cost for the transport. Finding a solution to this seemingly simple practical task has become a classical problem in combinatorial optimization referred to as the assignment problem or alternatively, using the language of graph theory, as the bipartite weighted matching problem (for a comprehensive analysis of assignment problems, see, for example, Ref. [1]). Interests in solving it have been stimulated by applications in operational research, economics, and data science, among others. With such a wide range of applications, it has been and remains a topic of research of equal importance for mathematicians, statisticians, and computer scientists. As a consequence, many solutions, exact or approximate, have been proposed. In this paper, we are interested in filling the gap in one group of approximate solutions based on mean-field theory and show that they can be modified to yield an exact solution in non degenerate as well as in degenerate situations in a computer efficient manner.

Let P be the set of plants, and B the set of bakeries. We are concerned with the balanced assignment problem, namely, we assume $|P| = |B| = N$. We note that the unbalanced problem (i.e., when there are different numbers of plants and bakeries) can always be reduced to the balanced case by adding pseudo plants or bakeries so that the two corresponding sets have the same cardinality. If we define as $C(i, j)$ the cost of transport between plant i and bakery j, then the assignment problem can be formalized as finding a bijection f between P and B that minimizes

$$U = \sum_{i \in P} C(i, f(i)). \tag{1}$$

Note that f can be seen as a permutation of $\{1, \ldots, N\}$. This is a linear problem. It can be solved naively by testing all possible bijections f, or equivalently all permutations of $\{1, \ldots, N\}$: this is, however, extremely inefficient, as the number of such permutations is $N!$ and unnecessary. There are indeed polynomial time algorithms to solve the assignment problem. The most famous of such algorithms was most likely originally proposed by Ref. [2] and published posthumously in Latin, and rediscovered 60 years later by Ref. [3] and dubbed the Hungarian algorithm. Initially developed as a $O(N^4)$ algorithm, it has since been sped up and its fastest exact, general version is of order $O(N^3 + N^2 \ln N)$ when using Fibonacci heaps [4]. The Hungarian algorithm remains the most efficient exact algorithm when applied to a generic cost

matrix C (there are faster versions for special cost matrices C; see, for example, Refs. [5,6]). It is a global algorithm that iteratively identifies assignments between the two sets of points, or, in the language of graph theory, by augmenting paths between the two graphs to be matched. While it has polynomial worst-case running time, its main limitation is that it is serial, i.e., it cannot be improved with parallelization. There are alternate solutions to the assignment problem such as the auction algorithms [7,8] that are based on finding local updates, rather than full augmenting paths between the two graphs. These methods have worse asymptotic computing time behaviors, but they often work better in practice [9]. These algorithms have an average time complexity of $O(N^2)$ and their structure is such that it is possible to parallelize them (see, for example, Ref. [10]). Note, however, that their gains in computing time compared to the Hungarian algorithm are highly problem-dependent: The parallelization gain may be modest for some cost matrices C, and in some degenerate cases, auction algorithms may even run forever [11].

The Hungarian and the auction algorithms are iterative methods aimed at finding the best bijection f within the discrete set of all possible permutations, of cardinality $N!$. On par with the invisible hand algorithm (IHA) proposed by Ref. [12], we propose instead to use continuous systems motivated by statistical physics. We have adapted an algorithm we have recently proposed to solve the balanced optimal transport problem [13,14] to solve specifically the assignment problem. We focus on the balanced assignment problem (i.e., with the same number of points in the two sets of points considered), with minimal cost, with an understanding that our method could be easily extended to handle unbalanced and/or maximal cost assignment problems. Our goals in this paper are to

(1) Establish and validate a continuous framework for solving the assignment problem using statistical physics,

(2) Establish that, in the generic case in which the assignment problem has a unique solution, the framework proposed above is guaranteed to converge arbitrarily close to that solution, and derive criteria to generate this solution,

(3) Describe a modification of the method that is guaranteed to find at least one solution for degenerate assignment problems with multiple solutions, and

(4) Demonstrate that the implementation of this framework can be efficiently parallelized on multiple cores / CPUs and/or a general purpose GPU.

Note that the first two goals were already achieved by the IHA algorithm [12]. In this paper, we propose a different statistical physics formulation of a relaxed version of the assignment problem, validate that it has similar theoretical properties as the IHA in terms of convergence, and include a comparison of its performance with respect to the IHA algorithm, showing that the latter becomes significantly slower for large systems. An analysis of the differences is provided. In addition, the IHA does not handle degenerate cases; a significant contribution in this paper is that such cases are considered explicitly within our framework (goal 3 listed above).

The following four sections map with the four goals listed above. We conclude with a discussion in which we compare our framework with alternate methods for solving continuous assignment problems, as well as with a presentation of possible extensions to very large assignment problems.

II. A FINITE-TEMPERATURE ASSIGNMENT PROBLEM

We consider two sets of points S_1 and S_2 of the same cardinality N. We encode the cost of transport between S_1 and S_2 as a positive matrix $C(k,l)$ with $(k,l) \in \{1,\ldots,N\}^2$. The assignment problem can then be formulated as finding a binary permutation matrix G of correspondence between points in S_1 and points in S_2 that minimizes the matching cost U defined as

$$U(G) = \sum_{k,l} G(k,l)C(k,l), \tag{2}$$

where the summations extend over all k in S_1 and l in S_2. The minimum of U is to be found for the values of $G(k,l)$ that satisfy the following constraints:

$$\forall k, \quad \sum_{l} G(k,l) = 1, \tag{3a}$$

$$\forall l, \quad \sum_{k} G(k,l) = 1, \tag{3b}$$

$$\forall (k,l), \quad G(k,l) \in \{0,1\}. \tag{3c}$$

The solution to the assignment problem provides an optimal permutation matrix G^* and the corresponding minimum matching cost $U^* = U(G^*)$. Minimizing Eq. (2) under the constraints Eq. (3) is a discrete optimization problem, namely, an integer linear program problem. We solve it using a statistical physics approach by rephrasing it as a temperature-dependent problem with real variables, with the integer optimal solution found at the limit of zero temperature. This relaxed version of the assignment problem is a special case of a discrete optimal transport (OT) problem [15,16] in which the masses associated to the points in S_1 and S_2 are all equal to 1. Many methods have been proposed for solving the OT problem, from directly solving the linear system stem to solving entropy-regularized version of this system [17]. Here we introduce a modified version of our statistical physics approach for solving this problem, adapting it to the specifics of the assignment problem. Note that this algorithm is a generalization of the so-called invisible hand algorithm [12].

A. An effective free energy for the assignment problem

In statistical physics, a system that is in thermal equilibrium at finite temperature will sample many states. The corresponding Gibbs distribution represents the probability of this system to exist in any specific state. The most probable state is then the one with lowest energy. Hence, minimizing an energy function can be reformulated as the problem of finding the most probable state of the system it defines. In the assignment problem between two sets S_1 and S_2, the "system" is identified with the different binary transport plans between S_1 and S_2 that satisfy the marginal constraints Eqs. (3a) and (3b) as well as the constraint (3c). Those plans belong to the permutation polytope which we denote as $\mathcal{G}$.

Each state in this system is identified with a transport plan $G \in \mathcal{G}$, and its corresponding energy $U(G)$ is defined in

Eq. (2). The probability $P(G)$ associated with a transport plan G is defined as

$$P(G) = \frac{1}{Z(\beta)} e^{-\beta U(G)}. \tag{4}$$

In this equation, $\beta = 1/k_B T$ where k_B is the Boltzmann constant and T the temperature, and $Z(\beta)$ is the partition function computed over all states of the system. This partition function is given by

$$Z(\beta) = e^{-\beta \mathcal{F}(\beta)} = \int_{G \in \mathcal{G}} e^{-\beta U(G)} dG, \tag{5}$$

where dG can be seen as the Lebesgue measure for the space of transport plans $\mathcal{G}$ and $\mathcal{F}(\beta)$ is the free energy of the system. This free energy is of limited practical interest as it cannot be computed explicitly. We propose a scheme for approximating it using the saddle point approximation.

Taking into account the constraints on the transport plan G, the partition function can be written as

$$Z(\beta) = \sum_{G(k,l) \in \{0,1\}} e^{-\beta \sum_{kl} C(k,l) G(k,l)} \prod_k \delta\left(\sum_l G_{kl} - 1\right) \prod_l \delta\left(\sum_k G_{kl} - 1\right). \tag{6}$$

The first sum imposes that the $G(k,l)$ take values of 0 or 1 only. The constraints that there is only one 1 per line and only one 1 per column are imposed through the δ functions. We use the Fourier representation of those δ functions, thereby introducing new auxiliary variables $\lambda(k)$ and $\mu(l)$, with $(k,l) \in \{1,\ldots,N\}^2$. After rearrangements, the partition function can be written as (up to a multiplicative constant),

$$Z(\beta) = \int_{-\infty}^{+\infty} \prod_k d\lambda(k) \int_{-\infty}^{+\infty} \prod_l d\mu(l) e^{\beta(\sum_k i\lambda(k) + \sum_l i\mu_l)} \sum_{G(k,l) \in \{0,1\}} e^{-\beta \sum_{k,l} G(k,l)[C(k,l) + i\lambda(k) + i\mu(l)]}. \tag{7}$$

Note that we have scaled the auxiliary variables λ and μ by a factor β for scale consistency with the energy term. Performing the summations over the variables $G(k,l)$, we get

$$Z(\beta) = \int_{-\infty}^{+\infty} \prod_k d\lambda(k) \int_{-\infty}^{+\infty} \prod_l d\mu_l e^{-\beta F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})}, \tag{8}$$

where F_β is a functional, or effective free energy defined by

$$F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu}) = -\left(\sum_k i\lambda(k) + \sum_l i\mu(l)\right) - \frac{1}{\beta} \sum_{kl} \ln\left(1 + e^{-\beta[C(k,l) + i\lambda(k) + i\mu(l)]}\right). \tag{9}$$

Note that compared to the internal energy U defined in Eq. (2) that depends on N^2 constrained binary variables $G(k,l)$, the effective free energy $F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})$ depends on $2N$ unconstrained variables $\lambda(k)$ and $\mu(l)$. In the following we will show how finding the extremum of this function allows us to solve the assignment problem.

B. Optimizing the effective free energy

Let $\bar{G}(k,l)$ be the expectation value of $G(k,l)$ with respect to the Gibbs distribution given in Eq. (4). As mentioned above, it is unfortunately not possible to compute this value directly as the partition function defined in Eq. (8) is not known analytically. Instead, we derive a saddle point approximation (SPA) by looking for extrema of the effective free energy with respect to the variables λ and μ:

$$\frac{\partial F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \lambda_k} = 0 \quad \text{and} \quad \frac{\partial F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \mu_l} = 0. \tag{10}$$

After some rearrangements, those two equations can be written as

$$\forall k, \quad \sum_l X(k,l) = 1, \tag{11a}$$

$$\forall l, \quad \sum_k X(k,l) = 1, \tag{11b}$$

where

$$X(k,l) = h\{\beta[C_{kl} + i\lambda(k) + i\mu(l)]\} \tag{12}$$

and

$$h(x) = \frac{1}{e^x + 1}. \tag{13}$$

We will prove that in the limit $\beta \to \infty$ (or equivalently $T \to 0$), the matrix X converges to the solution of the assignment problem G^* (see above).

As is often the case, the saddle-point may be purely imaginary. In the present case, one can easily see from Eq. (11) that the variables $i\lambda(k)$ and $i\mu(l)$ must be real and in the following, we will replace $\{i\lambda(k), i\mu(l)\}$ by $\{\lambda(k), \mu(l)\}$. We observe that the values of the matrix $X(k,l)$ are invariant under the translation $\{\lambda(k) + K, \mu(l) - K\}$ where K is an arbitrary constant. This translational degree of freedom leaves the free energy F_{eff} unchanged.

To analyze the SPA, we need to check the existence and assess the unicity of the critical points of the free energy. The following theorem shows that $F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})$ is weakly concave and can be made strictly concave on a subspace of the parameter space that is easily defined.

Theorem 1. The Hessian of the effective free energy $F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})$ is negative semi-definite with $(2N-1)$ negative eigenvalues and one zero eigenvalue. Furthermore, the eigenvector corresponding to the zero eigenvalue is (1,...,1, -1, ...-1) (with N 1s, and N -1s), and thus corresponds to the constant translation invariance of this energy. Setting one of the parameters $\lambda(k)$ or $\mu(l)$ as zero, the free-energy function on this restricted parameter space is strictly concave.

Proof. See Appendix A. ■

For a given value of the parameter β, the $X(k,l)$ that are solutions to the system of Eqs. (11) form a transport plan X_β^{MF} between S_1 and S_2 that is optimal with respect to the free energy defined in Eq. (9). We can associate to this transport plan an optimal free energy $F^{\text{MF}}(\beta)$ and an optimum energy $U^{\text{MF}}(\beta) = \sum_{k,l} X_\beta^{\text{MF}}(k,l) C(k,l)$. Note that those two values are the mean-field approximations of the exact free energy and internal energy of the system, respectively. We now list important properties of $U^{\text{MF}}(\beta)$ and $F^{\text{MF}}(\beta)$:

Property 1. $F^{\mathrm{MF}}(\beta)$ and $U^{\mathrm{MF}}(\beta)$) are, respectively, monotonic increasing and monotonic decreasing functions of the parameter β.

Proof. See Appendix B for $F^{\mathrm{MF}}(\beta)$ and Appendix C for $U^{\mathrm{MF}}(\beta)$. ■

Theorem 1 and the property 1 above highlight a number of advantages of the proposed framework that rephrases the assignment problem as a temperature-dependent process. First, at each temperature the assignment problem is turned into a strongly concave problem with a unique solution. This problem has a linear complexity in the number of variables, compared to the quadratic complexity of the original problem. The concavity allows for the use of simple algorithms for finding a minimum of the effective free-energy function [Eq. (9)]. We note also that Eq. (12) provides good numerical stability for computing the transport plan, because of the behavior of the function $h(x)$ (see below). Finally, the convergence as a function of temperature is monotonic.

C. Rewriting the free energy

Equation (9) provides an expression of the free energy as a function of the unconstrained variables $\lambda(k)$ and $\mu(l)$. This free energy is not "standard" as it does not include the corresponding energy U. We derive a new form for this free energy. To simplify notations, let us define

$$\begin{aligned} x_{kl} &= C(k,l) + \lambda(k) + \mu(l), \\ X(k,l) &= h(\beta x_{kl}), \\ U_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu}) &= \sum_{kl} C(k,l) X(k,l), \end{aligned} \tag{14}$$

where $h(x)$ is the function defined above. We have the following property:

Theorem 2. The effective free energy of the assignment problem can be written as

$$\begin{aligned} F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu}) = U_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu}) - T S_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu}) \\ + \sum_k \lambda_k \left(\sum_l X(k,l) - 1 \right) \\ + \sum_l \mu_l \left(\sum_k X(k,l) - 1 \right), \end{aligned} \tag{15}$$

where we have defined the entropy S as

$$S_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \sum_{kl} (\beta x_{kl} h(\beta x_{kl}) + \ln[1 + e^{-\beta x_{kl}}]). \tag{16}$$

In particular, at a maximum of the free energy,

$$F^{\mathrm{MF}}(\beta) = U^{\mathrm{MF}}(\beta) - T S^{\mathrm{MF}}(\beta). \tag{17}$$

This form of the free energy has an intuitive physical interpretation. The first term is the original assignment energy, the second is $-T$ times an entropy term, and the third and fourth terms impose constraints via Lagrange multipliers.

Proof. Using the definition of the free energy [Eq. (9)], and adding and subtracting the internal energy, we get

$$\begin{aligned} F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu}) = U_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu}) - \sum_{kl} C(k,l) X(k,l) \\ - \left(\sum_k \lambda(k) + \sum_l \mu(l) \right) \\ + \frac{1}{\beta} \sum_{kl} \ln[1 + e^{-\beta x_{kl}}]. \end{aligned} \tag{18}$$

As $C(k,l) = x_{kl} - \lambda(k) - \mu(l)$, we get

$$\begin{aligned} F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu}) = U_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu}) + \sum_k \lambda_k \left(\sum_l X(k,l) - 1 \right) \\ + \sum_l \mu_l \left(\sum_k X(k,l) - 1 \right) \\ \frac{1}{\beta} \sum_{kl} (\beta x_{kl} X(k,l) + \ln[1 + e^{-\beta x_{kl}}]), \end{aligned} \tag{19}$$

which concludes the proof. ■

By noticing that the function $H(x) = -\ln(1 + e^{-x})$ is an antiderivative of the function $h(x) = 1/(1 + e^x)$, we have more general definitions for the internal energy U_β and the entropy S_β as a function of $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$,

$$\begin{aligned} U_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu}) &= \sum_{kl} C(k,l) h(\beta x_{kl}), \\ S_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu}) &= \sum_{kl} t(\beta x_{kl}) \\ &= \sum_{kl} [\beta x_{kl} h(\beta x_{kl}) - H(\beta x_{kl})], \end{aligned} \tag{20}$$

or alternatively, as a function of the transport plan X,

$$\begin{aligned} U_\beta(X) &= \sum_{kl} C(k,l) X(k,l) \\ S_\beta(X) &= \sum_{kl} J[X(k,l)] \\ &= \sum_{kl} \left(X(k,l) h^{-1}[X(k,l)] - H\{h^{-1}[X(k,l)]\} \right). \end{aligned} \tag{21}$$

In the specific case considered here in which $h(x) = 1/(1 + e^x)$, the functions $t(x)$ and $J(x)$ are defined as

$$\begin{aligned} &t : \mathbb{R} \to \mathbb{R}, \quad t(x) = \frac{x}{1 + e^x} + \ln(1 + e^{-x}), \\ &J : [0,1] \to \mathbb{R}, \quad J(x) = -x \ln(x) - (1 - x) \ln(1 - x). \end{aligned} \tag{22}$$

Note that $J'(x) = \ln\left(\frac{1-x}{x}\right) = h^{-1}(x)$, and therefore that $J(x)$ is the Legendre transform of $-H(x)$. $J(x)$ is positive on [0,1], null only for $x = 0$ and $x = 1$, and maximum for $x = 0.5$ in which case it is equal to $\ln(2)$. In Fig. 1, we illustrate the different functions $h(x), H(x), t(x)$ and $J(x)$ as their properties are central to the rest of the paper.

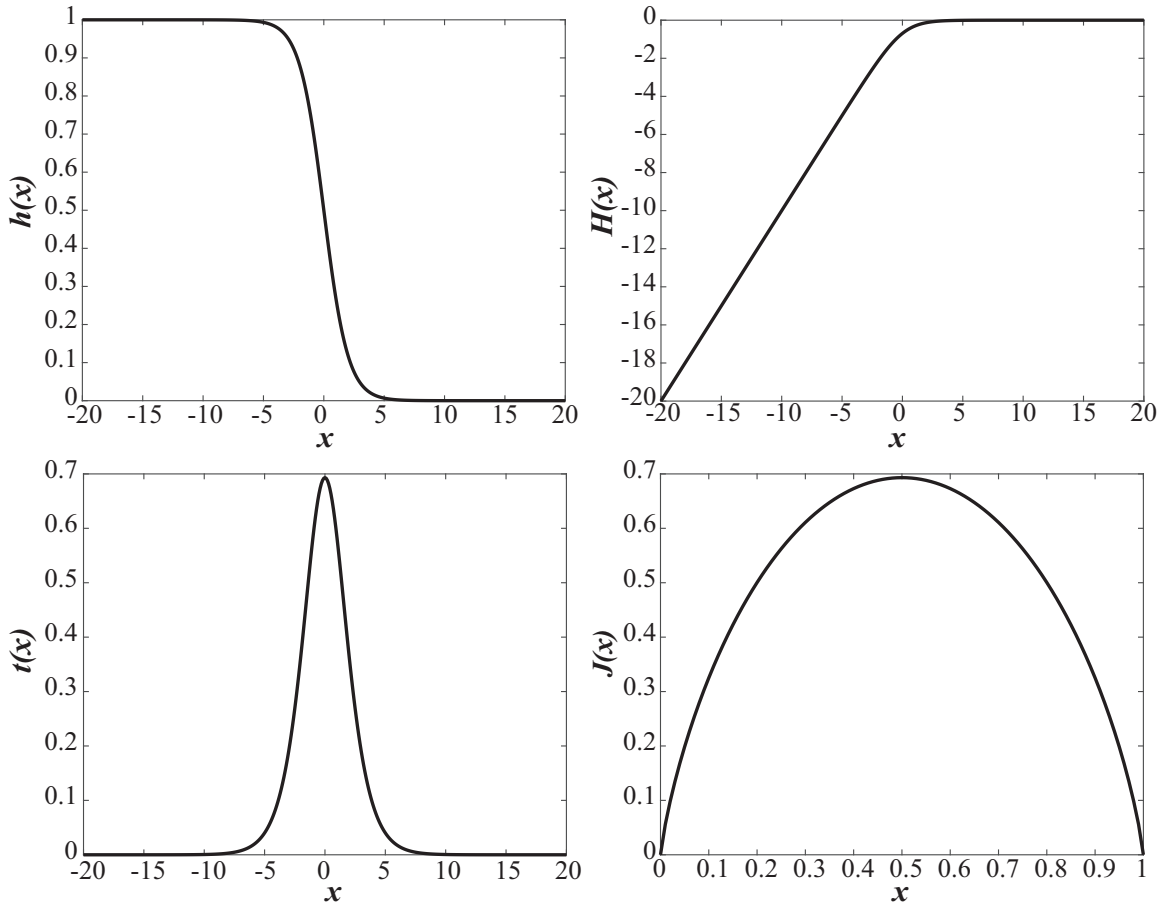

FIG. 1. The different functions $h(x)$, $H(x)$, $t(x)$, and $J(x)$ (see text for details).

III. SOLVING THE GENERIC ASSIGNMENT PROBLEM

In the previous section, we have described a formalism based on statistical physics for solving the assignment problem. We have derived an effective free energy, $F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})$, that depends on $2N$ unconstrained variables $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$. We have shown that this free energy is (weakly) concave and that its maximum is found by solving a system of nonlinear equations, at each inverse temperature β. We have also shown that the trajectory of the maxima $F^{\mathrm{MF}}(\beta)$ as a function of β is monotonic, increasing. We need to establish now that this trajectory allows us to find the actual solution of the assignment problem. Recall that this solution is defined by a permutation matrix G^* and its corresponding energy U^*. In this section, we will assume that the assignment problem is non degenerate and that it has a unique solution. We will fully characterize what it means in the next section.

We first prove that the optimal assignment energy U^*, is equal to the infinite inverse temperature limit of both the mean-field free energy and the internal energy:

Theorem 3.

$$U^* = \lim_{\beta \to +\infty} F^{\mathrm{MF}}(\beta),$$
$$U^* = \lim_{\beta \to +\infty} U^{\mathrm{MF}}(\beta). \tag{23}$$

Proof. See Appendix D. ■

As the trajectories of $F^{\mathrm{MF}}(\beta)$ and $U^{\mathrm{MF}}(\beta)$ as a function of β were already found to be respectively monotonically increasing and monotonically decreasing, this theorem adds the information at the infinite inverse temperature limit (or equivalently at the zero temperature limit), both converge to the optimal assignment energy. These results validate our statistical physics approach and the saddle-point approximation in particular. They are however results at convergence, i.e., at infinite inverse temperature, and we need to assess how well the solution at a finite inverse temperature approximates the exact solution.

The following theorem puts bounds on the entropy, internal energy, and free energy at the SPA. Let us define $A(N) = N^2 \ln(N) - N(N-1)\ln(N-1)$; then

Theorem 4.

$$0 \leqslant S^{\mathrm{MF}}(\beta) \leqslant A(N), \tag{24}$$

$$U^* - \frac{A(N))}{\beta} \leqslant F^{\mathrm{MF}}(\beta) \leqslant U^*, \tag{25}$$

$$U^* \leqslant U^{\mathrm{MF}}(\beta) \leqslant U^* + \frac{A(N)}{\beta}. \tag{26}$$

Proof. See Appendix E. ■

The two previous theorems are valid for all assignment problems. We establish now bounds on the element of the assignment matrix X_β^{MF} *in the specific case that this assignment problem has a unique solution*. The matrix X_β^{MF} denotes the

unique doubly stochastic matrix associated with the minimum of the free energy at the inverse temperature β. The next theorem bounds how close this doubly stochastic matrix is to the *unique* permutation matrix, G^*, representing the optimal solution to the assignment problem.

Theorem 5. Suppose that the assignment problem associated with the $N \times N$ cost matrix C admits a unique optimal assignment matrix, G^*. Let Δ be the difference in total cost between the optimal solution and the second best solution. Then,

$$\max_{k,l} \left| X_\beta^{\mathrm{MF}}(k,l) - G^*(k,l) \right| \leqslant \frac{A(N)}{\beta \Delta}. \tag{27}$$

Proof. See Appendix F. ■

This theorem validates that in the generic case for which the solution to the assignment problem is unique, the converged solution matrix X_∞^{MF} when $\beta \to +\infty$ is this unique solution to the assignment problem, G^*. In addition, it provides bounds to how close X_β^{MF} is from the optimal solution at any inverse temperature β. We can use this result to find bounds on the inverse temperature required to recover the optimal assignment from the free-energy optimum:

Theorem 6. Suppose that the assignment problem associated with the $N \times N$ cost matrix C admits a unique optimal assignment matrix, G^*. Let Δ be the difference in total cost between the optimal solution and the second best solution. Then, rounding off each of the entries of X_β^{MF} to the nearest integer yields the permutation matrix G^* that solves the assignment problem whenever

$$\beta > \frac{2A(N)}{\Delta}. \tag{28}$$

Proof. The proof follows directly from Theorem 5. Rounding off to the nearest integer will yield the optimal assignment matrix whenever

$$\max_{k,l} \left| X_\beta^{\mathrm{MF}}(k,l) - G^*(k,l) \right| < \frac{1}{2}. \tag{29}$$

This condition is met if

$$\frac{A(N)}{\beta \Delta} < \frac{1}{2}, \tag{30}$$

i.e., if

$$\beta > \frac{2A(N)}{\Delta}. \tag{31}$$

■

We can then conclude that in the generic case we can solve the assignment problem exactly at finite, although sufficiently high inverse temperature β. Assuming finite precision, the inverse temperature required for convergence is $O[A(N)]$. Since $A(N) = N^2 \ln(N) - N(N-1)\ln(N-1)$, $A(N)$ is $O[N \ln(N)]$, and therefore the inverse temperature is of order $N \ln(N)$. Theorem 6 is important theoretically as it validates that the mean-field approach converges to the solution of a generic assignment problem. It also provides a recipe for setting a cutoff for the value of the inverse temperature β that guarantees that the optimal solution has been found. It is, however, not easy to implement as it is difficult to estimate Δ. A simpler procedure is based on the following result:

Theorem 7. Suppose that the assignment problem associated with the $N \times N$ cost matrix C admits a unique optimal assignment matrix, G^*. Let us assume that at an inverse temperature β, the current solution matrix X_β^{MF} is strictly row dominant. Then, rounding off each of the entries of X_β^{MF} to the nearest integer yields the permutation matrix G^* that solves the assignment problem.

This theorem defines a criteria that is easily implemented to terminate the annealing process in β when solving the assignment problem with our method. Note that this theorem is not equivalent to Theorem 6. It does not guarantee convergence, i.e., it does not establish that the matrix X_β^{MF} becomes row dominant, but only claims that if it does, then the annealing process can be stopped. From Theorem 6 we know however that if the assignment problem has a unique solution, then our framework will converge to this solution, and, in doing so, the trajectory of the X_β^{MF} matrices is guaranteed to reach a row dominant matrix.

Proof. See Appendix G. ■

In Sec. VI, we will compare these possible cutoff schemes for β.

IV. SOLVING DEGENERATE ASSIGNMENT PROBLEMS

In our statistical physics approach described in Sec. II, the binary assignment problem has been relaxed. Indeed, we build a collection of real matrices X_β^{MF} that minimizes the assignment cost and that are doubly stochastic, but at a finite inverse temperature this matrix cannot be binary as X_β^{MF} is given as $h(\beta x_{kl})$, where all values $h(x)$ are strictly in (0,1) (see Fig. 1). In the previous section, we have shown that if this relaxed assignment problem has a unique solution, then the trajectories of the X_β^{MF} as β increases converge to the permutation matrix G^* that solves the integer assignment problem. The question remains as to whether this is always the case, and if it is not, then how we can still find an integer solution to the assignment problem.

Let S_1 and S_2 be two sets of points of cardinality N and let C be a cost matrix between S_1 and S_2. The relaxed assignment problem can be formulated as finding the assignment matrix G_r that minimizes

$$U(G) = \sum_{k,l} G_r(k,l) C(k,l), \tag{32}$$

where the summations extend over all k in S_1 and l in S_2. The minimum of U is to be found for the values of $G_r(k,l)$ that satisfy the following constraints:

$$\forall k, \quad \sum_l G_r(k,l) = 1, \tag{33a}$$

$$\forall l, \quad \sum_k G_r(k,l) = 1, \tag{33b}$$

$$\forall (k,l), \quad 0 \leqslant G_r(k,l) \leqslant 1. \tag{33c}$$

Note that the assignment problem defined in Eqs. (2) and (3) is a special case of this relaxed problem. The two problems have the same solution under circumstances that will be described below. In general, it is not expected that solving this relaxed problem will lead to an optimal matrix G_r^* that

is binary. However, Ref. [18] have proved that this relaxed assignment problem always has an optimal solution where G_r^* take integer values. We rewrite the corresponding theorem here and sketches its proof, as it contains elements that we will use later.

Theorem 8 ([18]). If the relaxed assignment problem has at least one feasible solution, then it has at least one integral optimal solution. This solution is an optimal solution for the corresponding integer assignment program.

Proof. Let G_r^* be an optimal solution to the relaxed assignment problem described in Eqs. (32) and (33) and let $U(G_r^*)$ be the associated minimum assignment cost. Let us denote as K the number of nonintegral values in G_r^*. If $K = 0$, then G_r^* is a permutation matrix and we are done. If $K > 0$, then let $G_r^*(k_1, l_1)$ be one of its nonintegral values:

$$0 < G_r^*(k_1, l_1) < 1. \tag{34}$$

Since

$$\sum_k G_r^*(k, l_1) = 1, \tag{35}$$

there exists $k_2 \in S_1$ with $k_2 \neq k_1$ such that $G_r^*(k_2, l_1)$ is nonintegral. Similarly, we can find $l_2 \neq l_1$ in S_2 such that $G_r^*(k_2, l_2)$ is nonintegral. We can continue in this manner, leading to a path $[(k_1, l_1), (k_2, l_1), \ldots]$ with nonintegral values in G_r^*. As the number of points in S_1 and S_2 is finite, we will ultimately reach a pair that we have already visited. This means that we have identified a cycle A among all edges between S_1 and S_2; the cardinality of this cycle is even (bipartite graph). We write this cycle as

$$A = \{(a_1, b_1), (a_2, b_2), \ldots, (a_{2M}, b_{2M})\}, \tag{36}$$

where $2M = |A|$. For a small real number ϵ, we define the matrix G_ϵ as

$$G_\epsilon(k, l) = G_r^*(k, l) \quad (k, l) \notin A,$$

$$G_\epsilon(a_{2i}, b_{2i}) = G_r^*(a_{2i}, b_{2i}) + \epsilon \quad i \in \{1, \ldots, M\},$$

$$G_\epsilon(a_{2i+1}, b_{2i+1}) = G_r^*(a_{2i+1}, b_{2i+1}) - \epsilon \quad i \in \{1, \ldots, M\}.$$

As two consecutive pairs in A leads to the addition and subtraction of the same quantity ϵ on one row or one column of G_r^*, it is easy to verify that G_ϵ is doubly stochastic and therefore satisfies the constraints of the assignment problem. In addition, for sufficiently small ϵ, we have

$$0 \leqslant G_\epsilon(a_i, b_i) \leqslant 1 \tag{37}$$

for all $(a_i, b_i) \in A$, as by construction those pairs where chosen such that $0 < G_r^*(a_i, b_i) < 1$. Let us now compute the assignment cost U_ϵ associated with G_ϵ:

$$\begin{aligned} U_\epsilon &= \sum_{k,l} G_\epsilon(k, l) C(k, l) \\ &= \sum_{(k,l)\notin A} G_\epsilon(k, l) C(k, l) + \sum_{i=1}^{2M} G_\epsilon(a_i, b_i) C(a_i, b_i) \\ &= \sum_{(k,l)\notin A} G_r^*(k, l) C(k, l) + \sum_{i=1}^{2M} [G_r^*(a_i, b_i) + (-1)^i \epsilon] C(a_i, b_i) \\ &= U(G_r^*) + \epsilon \Gamma, \end{aligned} \tag{38}$$

where we have defined $\Gamma = \sum_{i=1}^{2M} (-1)^i C(a_i, b_i)$. Since G_r^* is optimal, we have $\Gamma = 0$, for otherwise, we would have $U_\epsilon < U(G_r^*)$ either by choosing $\epsilon > 0$ if $\Gamma < 0$, or by choosing $\epsilon < 0$ for $\Gamma > 0$. This means that G_ϵ is another optimal solution of the relaxed assignment problem. By choosing the largest $\epsilon > 0$ for which the constraints $0 \leqslant G_\epsilon(a_i, b_i) \leqslant 1 \quad \forall i$ are still satisfied, one of the $(a_i, b_i) \in A$ will be such that $G_\epsilon(a_i, b_i) \in \{0, 1\}$. Therefore, G_ϵ has fewer nonintegral elements than G_r^* and the procedure can be repeated until $K = 0$. ∎

The proof described above provides an algorithm for modifying a fractional optimal solution to the relaxed assignment problem into an integer solution with the same optimal assignment cost. This algorithm, however, is numerically unstable and difficult to implement for large N. Indeed, a fractional solution to the assignment problem will not satisfy the constraints exactly (because of numerical imprecision) and therefore cycles are difficult to identify. An alternate solution to following this algorithm would be to add a penalty term to the energy function of the form $\sum_{kl} G(k, l)[1 - G(k, l)]$ that would be minimum when $G(k, l)$ is 0 or 1, therefore pushing the solution towards integer values. We propose a different solution. We first list an interesting side result from the proof above as a property on its own:

Property 2. If the relaxed assignment problem has an optimum solution that contains fractional elements, then there exists (at least) one cycle $A = \{(a_1, b_1), (a_2, b_2), \ldots, (a_{2M}, b_{2M})\}$ in the cost matrix C for which $\Gamma = \sum_{i=1}^{2M} (-1)^i C(a_i, b_i) = 0$. Reversely, if the cost matrix C does not contain any cycle of the form $A = \{(a_1, b_1), (a_2, b_2), \ldots, (a_{2M}, b_{2M})\}$ for which $\sum_{i=1}^{2M} (-1)^i C(a_i, b_i) = 0$, then the corresponding assignment problem has a unique integer solution.

Proof. The proof of the first part of the proposition follows exactly the proof of Theorem 8 that is sketched above. The second part is basically its contrapositive. Briefly, we start from the fact that the cost matrix C does not contain any cycle of the form $A = \{(a_1, b_1), (a_2, b_2), \ldots, (a_{2M}, b_{2M})\}$ for which $\sum_{i=1}^{2M} (-1)^i C(a_i, b_i) = 0$. Let us assume now that the corresponding assignment matrix has an optimal solution matrix that contains fractional elements. Then, based on the proof of Theorem 8, we can identify (at least) one cycle in the cost matrix, which is contradictory to our hypothesis. Therefore, the assignment problem has only integer solutions that are permutation matrices. Let us assume now that it has (at least) two different optimal permutation matrices π_1 and π_2 as solutions, i.e., with the same optimal cost U^*. We can then build a doubly stochastic matrix $G_a = a\pi_1 + (1 - a)\pi_2$ for each $a \in [0, 1]$. The cost associated with G_a is

$$\begin{aligned} U(G) &= a \sum_{kl} C(k, l) \pi_1(k, l) + (1 - a) \sum_{kl} C(k, l) \pi_2(k, l) \\ &= a \sum_k C(k, \pi_1(k)) + (1 - a) C(k, \pi_2(k)) \\ &= aU^* + (1 - a)U^* = U^*, \end{aligned} \tag{39}$$

for all $a \in [0, 1]$. This would mean that G_a is also an optimal solution to the assignment problem. However, for a strictly in (0,1), G_a is fractional, which contradicts the fact that the assignment problem only has integer solutions. Therefore, the assignment problem has a unique solution. ■

This property implies that when solving the assignment problem for generic cost matrices that do not contain specific cycles, we can follow the strategy described in Sec. III. If the cost matrix is degenerate and contains (at least) one cycle, then we propose to randomly perturb that matrix to bring it back to the generic problem. We do need to specify the perturbation level that guarantees that a solution of the perturbed problem is also a solution to the original problem. This is the purpose of the following theorem:

Theorem 9. Suppose that the solution $X_\beta^{\rm MF}$ to the assignment problem associated with the $N \times N$ cost matrix C has a nonzero entropy $S^{\rm MF}(\beta)$ when $\beta \to +\infty$. Let Δ be the difference in total cost between the optimal solution and the second best solution. Then, adding random uniform noise with support $[0, \alpha]$ to each value of C and solving the assignment problem on this perturbed matrix will generate one integer solution that is also solution to the unperturbed assignment problem with probability one, whenever

$$\alpha < \frac{\Delta}{2N}. \tag{40}$$

If all the entries of the cost matrix C are scaled to be integers, then $\Delta \geqslant 1$ and it suffices to have

$$\alpha < \frac{1}{2N}. \tag{41}$$

Proof. See Appendix H. ■

This theorem gives us a general strategy for solving a minimum cost assignment problem for any cost matrix C:

(a) Solve the assignment problem using the approach described in Secs. II A and II B. If the trajectory of the entropy converges to 0 as $\beta \to +\infty$, then the solution is guaranteed to be a permutation. All results from Sec. III apply and in particular the solution matrix G^* obtained by rounding off each element of $X_\beta^{\rm MF}$ when β is large enough is guaranteed to be an optimal solution to the assignment problem. In practice, β is set to be large enough when the entropy $S^{\rm MF}(\beta)$ falls below a cutoff value, usually 10^{-6}.

(b) If the approach described in (a) fails as the entropy does not converge to 0, then scale the cost matrix to be integer, and add random uniform noise in the interval $[0, \frac{1}{2N}]$. Solve the assignment problem with the perturbed matrix: its solution is guaranteed to solve also the unperturbed assignment problem.

V. IMPLEMENTATION

We have implemented the finite-temperature assignment framework described here in a C++ program matching that is succinctly described in Algorithm 1.

Algorithm 1. Matching: A temperature-dependent framework for solving the assignment problem

Input: The size of the assignment problem, N, the cost matrix C, and the initial value β_0 for β
Initialize: Initialize arrays $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ to 0. Set $STEP = \sqrt{10}$. Set $\beta^0 = \beta_0/STEP$
for $k = 1, \ldots$ until convergence **do**
(1) Set $\beta^k = STEP * \beta^{k-1}$.
(2) Solve nonlinear Eqs. (11) for $\boldsymbol{\lambda}^{\rm MF}$ and $\boldsymbol{\lambda}^{\rm MF}$ at saddle point
(3) Compute current optimal assignment matrix $X_\beta^{\rm MF}$ and the corresponding assignment $U^{\rm MF}(\beta)$ and entropy $S^{\rm MF}(\beta)$
(4) Check for convergence: If $|U^{\rm MF}(\beta^k) - U^{\rm MF}(\beta^{k-1})|/U^{\rm MF}(\beta^{k-1}) < TOL$, or if $X_\beta^{\rm MF}$ strictly row dominant, or if the entropy falls below a cutoff value, then stop
end for
Output: The converged transport plans $G_\beta^{opt}(k, l)$ and the corresponding transport costs $U^{\rm MF}(\beta)$.

Matching is based on an iterative procedure in which the parameter β (inverse of the temperature) is gradually increased. At each value of β, the nonlinear system of equations defined by Eq. (11) is solved using an iterative Newton-Raphson method. At each iteration for this Newton method, the Jacobian of the system of equations is computed, and the corresponding linear system of equations is solved using a preconditioned conjugate gradient approach (we use an incomplete LU decomposition of the Jacobian matrix as a preconditioner). Solutions of this system provide updated estimates for the arrays of parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$. These new estimates are then used to assess how well the SPA equations are satisfied. Once the errors on the SPA equations fall below a tolerance TOL (usually set to 10^{-4}), the optimal cost matrix $X_\beta^{\rm MF}$ and the corresponding assignment $U^{\rm MF}(\beta)$ and entropy $S^{\rm MF}(\beta)$ are computed. If the latter falls within the tolerance TOL, or if the matrix the procedure is deemed to have converged, then the program is stopped. The values in the corresponding $X_\beta^{\rm MF}$ are rounded off and its corresponding cost defines the minimal assignment cost. Note that the converged values of $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ at a given β serve as input for solving the SPA nonlinear system of equation at the following β, in spirit of an annealing procedure.

In some cases, matching has converged in energy, but the entropy remains nonzero and the current assignment matrix $X_\beta^{\rm MF}$ is not row dominant. This indicates that the assignment problem does not have a unique solution, and that matching has identified a fractional solution. We then rerun matching by applying the relaxed procedure described in Sec. IV, by introducing random noise to each element of the cost matrix (see above for details).

VI. NUMERICAL SIMULATIONS

In order to confirm our theoretical results, we applied our method to solve random linear assignment problem with random cost matrices of size $N \times N$ whose elements are independent identically distributed (iid) variables with exponential distribution with mean 1. The authors of Ref. [19] have conjectured that the expectation value for the optimal, minimal

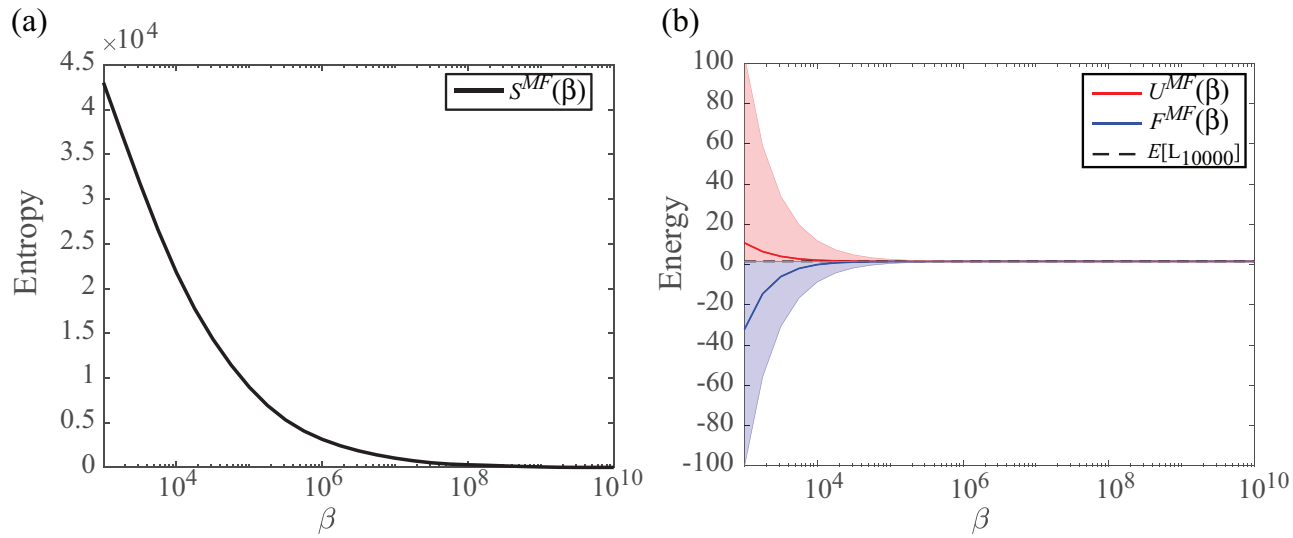

FIG. 2. Convergence of the entropy $S^{\mathrm{MF}}(\beta)$ (a), and of the internal energy $U^{\mathrm{MF}}(\beta)$ and free energy $F^{\mathrm{MF}}(\beta)$ as a function of β when solving a random assignment problem with a cost matrix C of size 10 000 whose elements are independent identically distributed values drawn from exponential distributions with mean 1. In panel (b), we show the theoretical bounds on the internal energy and on the free energy computed from the theoretical bounds given in Theorem 4 as red and blue shaded areas, respectively. The dotted horizontal line show the expected value for the minimal cost of an exponential random assignment problem of the same size.

assignment cost L_N satisfies

$$\lim_{N\to+\infty} E[L_N] = \frac{\pi^2}{6}, \tag{42}$$

and a few years later [20] they conjectured that

$$E[L_N] = \sum_{i=1}^{N} \frac{1}{i^2}. \tag{43}$$

Proofs of the two conjectures were subsequently provided by Refs. [21] and [22], respectively. We note that such random problems are guaranteed to have a unique solution: as the elements of the cost matrix are iid variables, there is a zero probability that they can form cycles (as defined in the previous section) and then, based on proposition 2, the corresponding assignment problem has a unique solution matrix whose entries are 0 or 1.

Matching is based on an iterative procedure in which the parameter β (inverse of the temperature) is gradually increased. At each value of β, the nonlinear system of equations defined by Eq. (11) is solved using an iterative Newton-Raphson method. At each iteration for this Newton method, the Jacobian of the system of equations is computed, and the corresponding linear system of equations is solved using a preconditioned conjugate gradient approach (we use an incomplete LU decomposition of the Jacobian matrix as a preconditioner). Solutions of this system provide updated estimates for the arrays of parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$. These new estimates are then used to assess how well the SPA equations are satisfied. Once the errors on the SPA equations fall below a tolerance TOL (usually set to 10^{-4}), the optimal cost matrix X_β^{MF} and the corresponding assignment $U^{\mathrm{MF}}(\beta)$ and entropy $S^{\mathrm{MF}}(\beta)$ are computed. If the latter falls within the tolerance TOL, then the procedure is deemed to have converged and the program is stopped. The values in the corresponding X_β^{MF} are rounded off and its corresponding cost defines the minimal assignment cost. Note that the converged values of $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ at a given β serve as input for the following β, in spirit of an annealing procedure.

A. Simple example

We ran the procedure described above on a random cost matrix with exponential distributions with mean 1 of size 10 000 × 10 000. In Fig. 2(a), we show the trajectory of the entropy $S^{\mathrm{MF}}(\beta)$ (left panel) as a function of β generated while solving the corresponding assignment problem. As this assignment problem has a unique matrix solution whose elements are either 0 or 1 (i.e., a permutation matrix), it is expected that the entropy converges to 0, as observed. Based on Theorem 4, the entropy is bounded in $[0, A(10\,000)]$ where $A(10\,000) \approx 10^5$. In Fig. 2(b), we show the corresponding trajectories of the internal energy $U^{\mathrm{MF}}(\beta)$ and free energy $F^{\mathrm{MF}}(\beta)$ as well as the theoretical bounds on those values given in Theorem 4. As expected, the internal energy is monotonically decreasing while the free energy is monotonically increasing, and both converge to the same value, 1.6341. Note that from Eq. (43), the expected value of the minimum cost associated with a matrix of this size is $E[L_{10\,000}] = 1.6445$, i.e., very close to the value observed with the specific cost matrix that was generated for this example.

B. Solving large assignment problems

We ran simulations on random cost matrices with exponential distributions with mean 1 of sizes N ranging in size between 50 × 50 and 15 000 × 15 000. We ran five independent simulations for each size. As mentioned above, each of these assignment problems have a unique solution; we verified that we obtained the correct assignment by running in parallel the Hungarian algorithm.

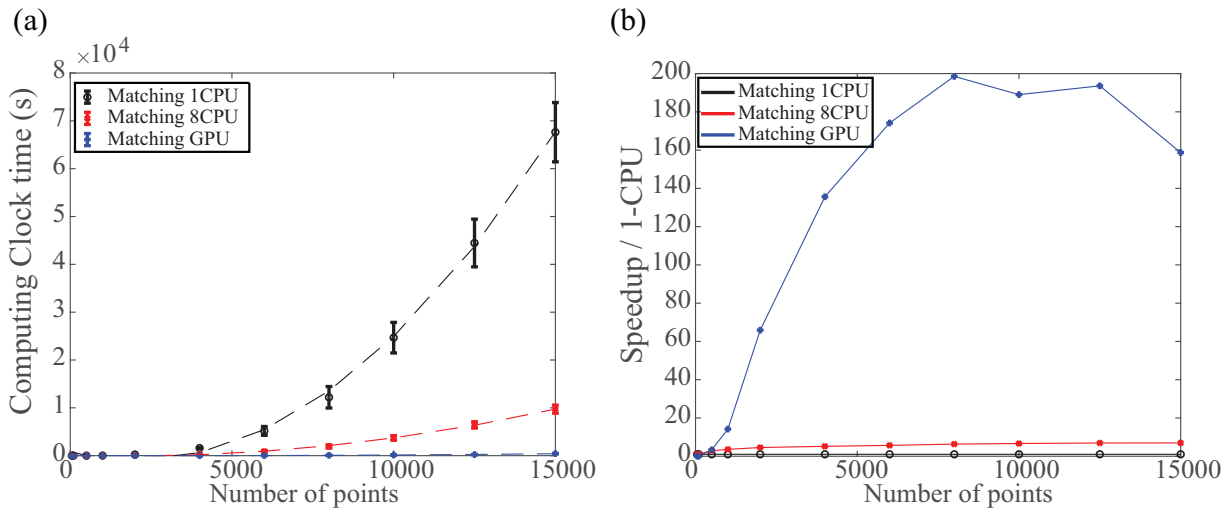

FIG. 3. CPU time and efficiency of matching for solving the assignment problem. Two sets of calculations are performed, on two different computers. The first set is based on the implementation of matching for CPUs. It is run on an Intel Core i7 processor running at 4.00 GHz, and 64 GB of memory. The second set is based on the implementation of matching for GPUs. It is run on a Linux server, with Xeon Platinum 8168 CPU at 2.7 GHz, and a NVIDIA RT2080 Ti GPU card with 11 GB of memory. In each set, we run matching on random cost matrices with exponential distributions with mean 1 of sizes N ranging in size between 50×50 and $15\,000 \times 15\,000$. (a) The mean computing times (clock time) are plotted against the sizes of the cost matrices for computation on one core (black), on 8 cores (red), and on the GPU (blue). The dashed lines represent quadratic polynomial fits to the means. (b) The speedup (computed as the ratio of total computing time over clock time) is plotted against the size of the assignment problem.

We have claimed that the temperature-based method we propose enables a fast and robust solution to the assignment problem. To check that it is indeed the case, we have monitored the running times for our procedure for the different simulations described above. We first note that our implementation relies heavily on linear algebra, as at each inverse temperature we solve a nonlinear system of equation iteratively, with each iteration involving the solution of a linear system of equation. It is therefore expected that the whole algorithm can benefit greatly from parallelization. We have therefore implemented two versions of matching, one that runs on possibly multiple CPUs, and another that runs on GPUs. Both rely heavily on the optimized BLAS and LAPACK libraries for the corresponding processors. The computing times for the two versions of matching, averaged over five simulations, are plotted against the size N of the assignment problem in Fig. 3. As expected, we observe a significant speedup when matching is run on multiple processors: a factor of nearly 7 for large matrices on a 8 CPUs, and nearly a factor of 200 when matching is run on GPU. The gain in time is significant: the mean computing time for solving a random assignment problem with a cost matrix of size $15\,000 \times 15\,000$ is 67 000 s for a serial computation on a single CPU, and 9700 s and 425 s on a modern 8-CPU computer and on a modern GPU card, respectively. While we cannot fully take credit for the effectiveness of the different implementations of matching as they are based on the highly efficient machine-specific BLAS and LAPACK libraries, we note that the method we have presented here provides the framework for such significant improvement in computing time compared to a serial computation.

To estimate the overall time complexity of matching, we need to consider the number M of β values considered, the number P of iterations required to solve the nonlinear system of equations at each β, the number of conjugate gradient iterations required to solve the linear Jacobian system at each of those iterations, and finally the cost of each of those conjugate gradients. For a cost matrix of size $N \times N$, the linear Jacobian system is of size $2N \times 2N$, and the worst case complexity for solving such a system of equation using conjugate gradient is $2N \times 2N \times 2N$, assuming $2N$ iterations to reach convergence, namely, a $O(N^3)$ time complexity. The total worst case complexity of matching is therefore of order $M \times P \times N^3$, where M is a constant (see below), while P depends on the quality of the initial guess for the solution of the system (also discussed below). In practice, we observe a N^2 time complexity [see Fig. 3(a)]. This quadratic time complexity indicates that the conjugate gradient procedure converges in a small number of steps which is nearly independent of N.

Matching includes an annealing procedure with respect to the temperature. In practice, this means that the values of the converged parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ at one value of β are used as input to the next value of β considered. This is found to significantly improve convergence. We did repeat the calculations with matching in which $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ are reset to zero for each β value. The reset was found to lead to significantly less efficient convergence. This is expected, as the efficiency of solving the system of nonlinear equations at each β is strongly dependent on the quality of the initial guess for the solution, with zero being a poor guess, and the value computed at the prior β a more reasonable guess.

One option to reduce the computing cost associated with matching is to limit the number of inverse temperatures β considered. In all the simulations described above, the temperature annealing is performed until the total entropy $S^{\mathrm{MF}}(\beta)$ falls below a small tolerance (10^{-6}). When this happens, the

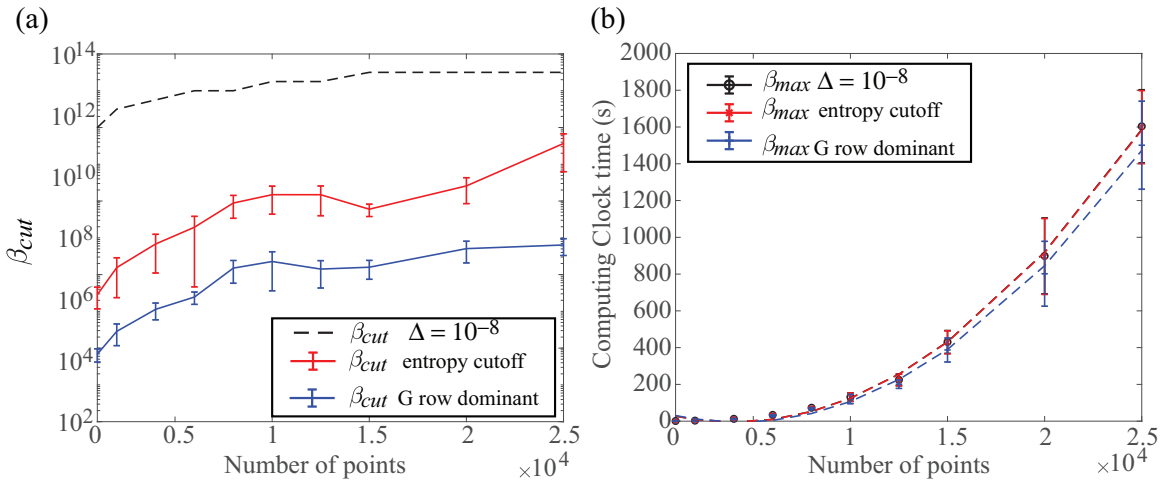

FIG. 4. Comparing different cutoff criteria for stopping the temperature annealing in matching. (a) Cutoff β values (averaged over 5 simulations) of the theoretical (based on Theorem 6, with $\Delta = 10^{-8}$) (black dashed line), entropy-based (red), and detection of strict row dominance of the assignment matrix X_β (blue), as a function of the assignment problem size. (b) The corresponding mean total computing times (on GPU) as a function of the assignment problem size. The dashed lines represent quadratic polynomial fits to the means. Note that the red and black curves overlap exactly.

values X_β^{MF} of the assignment matrix are either very close to 0, or very close to 1, and it is then safe to round them off to the nearest integer. There are two other possible criteria that we can use to determine when to stop the temperature annealing process and still be guaranteed that we can recover the optimal assignment. One is to consider the theoretical cutoff defined in Theorem 6, namely, that β be larger than $\frac{2A(N)}{\Delta}$, where $A(N)$ is a number defined by N, and Δ is the difference in total cost between the optimal solution and the second best solution. While this cutoff guarantees that the solution at such a β is the exact solution, after roundoff to the nearest integer, it is difficult to implement as Δ is not known. In practice, for computations with fixed precision, it can be set to be this precision, i.e., of the order of 10^{-8}. The second possibility is to continue the annealing process until the computed assignment matrix X_β^{MF} becomes strictly row dominant and then stop and roundoff the elements of X_β^{MF} to the nearest integer. The validity of this approach is established in Appendix H. In Fig. 4, we compare the three stopping criteria and their impacts on computing time, for random assignment problems with exponential distributions with mean 1 of sizes N ranging in size between 50×50 and $25\,000 \times 25\,000$. All computations are performed on GPU. Each simulation is run with β up to 10^{14}, with the different stopping criteria being computed during the annealing procedure. Five simulations are run for each matrix size.

There are significant differences in the required maximum β values to reach convergence with guaranteed exact solutions after rounding off to the nearest integer, based on the criteria considered, with the condition of row dominance of the assignment matrix giving the smallest β cutoff. This is expected as the other criteria cannot be satisfied before the matrix is strictly row dominant. For large assignment problems, the difference is significant, with a cutoff of the order of 10^9 for row dominance, and of the order of 10^{11} for entropy cutoff. The cutoff based on Δ is even larger (close to 10^{14}), however this cutoff was set arbitrarily large as it is difficult to actually estimate Δ.

Interestingly, while the stopping criteria based on entropy and based on an estimate of Δ differ significantly, the corresponding computing times do not and in fact overlap exactly. When the entropy cutoff is satisfied, the assignment matrix is basically integer and the system has converged; adding a few steps in β will not change the computing time, as the initial values from the entropy converged step will satisfy the nonlinear SPA systems at larger values of β. Computing times based on the row dominance cutoff are shorter (mean value of 1500 s for matrices of size $25\,000 \times 25\,000$, compared to 1600 s on the same matrices for the entropy cutoff), but by less than 10% while the differences in the maximum β value is of two orders of magnitude. Again, when the procedure has converged, independent of the cutoff scheme, additional steps will come at minimal computing costs. We do note, however, that large values of β (of the order of 10^9) are required for large assignment problems (of the order of 25 000), are therefore the computing framework is expected to be numerically stable, which is the case for our own procedure.

C. Comparisons with other algorithms for solving the assignment problem

Based on the results presented above. Matching is expected to provide a fast and robust solution to the assignment problem. To check that it is indeed the case, we have compared matching, with our own implementations of the invisible hand algorithm and of the entropy regularized approach to the assignment problem.

The IHA [12] is very similar to the method proposed in this paper. Indeed, in both approach the relaxed assignment problem is rewritten as the problem of finding a saddle point approximation for a free-energy functional. The derivations and consequently the free-energy functionals differ; both,

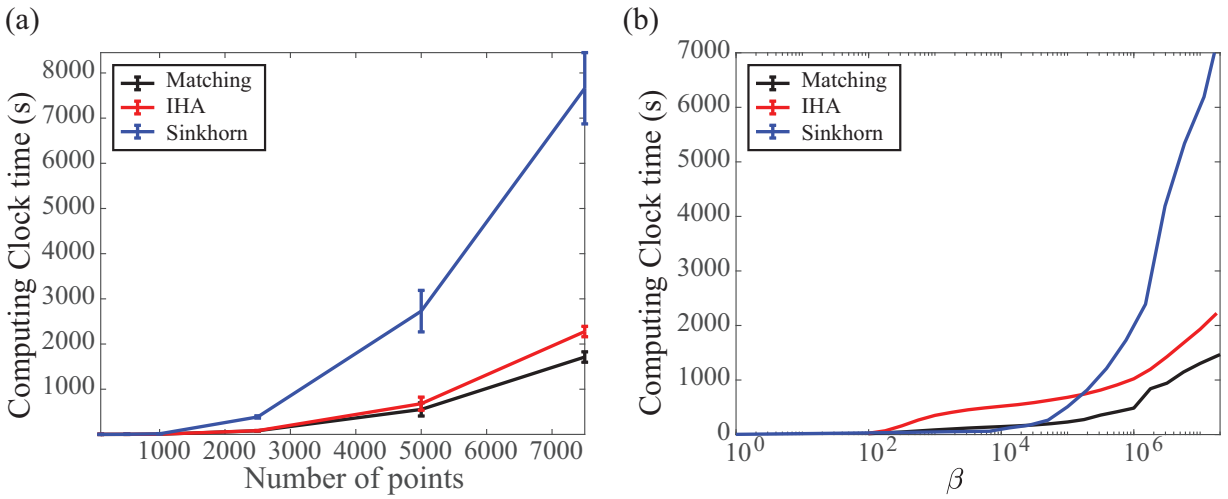

FIG. 5. Time complexity for matching, IHA, and Sinkhorn algorithms. (a) The running times for matching, the discrete relaxed solver of the assignment problem introduced in this paper, for the Invisible Hand Algorithm, with an implementation that mirrors matching (see text for details), and for a stabilized version of the Sinkhorn algorithm are given as a function of the size N of balanced random assignment problems, with randomized cost matrix with exponential distributions. Results are shown as mean values over five experiments at each value with N, with error bars at one standard deviation. (b) The three algorithms are run on the same cost matrix of size $7\,500 \times 7\,500$. Theirs computing times are shown as a function of the annealing parameter β, the inverse of the temperature. For the Sinkhorn algorithm, β is the inverse of the relaxation parameter ϵ, i.e., the factor in front of the entropy regularization term. All timings are computed on a single I7 processor running at 4.0 GHz with 64 GB of RAM.

however, are written as functions of unconstrained variables λ and μ, which satisfy a system of equations of the form

$$\forall k, \quad \sum_l G(k,l) = 1, \tag{44a}$$

$$\forall l, \quad \sum_k G(k,l) = 1, \tag{44b}$$

where

$$G(k,l) = h\{\,\beta[C_{kl} + \lambda(k) + \mu(l)]\}, \tag{45}$$

where C is the cost matrix, and G the coupling matrix to be found. The main difference between matching and IHA lays in the function $h(x)$, with $h(x) = 1/(1+e^x)$ for the matching algorithm, and $h(x) = e^{-x}$ for the IHA. The simplicity of the latter makes it possible to eliminate the variables λ and solve only for the variables μ. This can be done using a Sinkhorn-like fixed point algorithm, a steepest descent algorithm (both approaches were described in the original IHA paper [12]), or using a Newton approach, as proposed for matching. Our implementation of the IHA follows the latter.

Our implementation of the Sinkhorn algorithm for solving the relaxed assignment problem is based on a log-domain stabilization and eta-scaling heuristic [23] and an overrelaxation scheme [24]. These two modifications to the original algorithm of Cuturi [25] are expected to improve convergence of the iterative scaling algorithm, as well as robustness for small values of the relaxation parameter ϵ through the use of logarithmic stabilization.

We have experimented with applications of matching, IHA, and Sinkhorn on random cost matrices based on exponential distributions, as described above. We have solved the corresponding assignment problems using all three methods, for problem size N between 100 and 7,500, with five independent runs for each value of N. The optimization is performed until convergence, using the row dominance criterium of Theorem 7. All computational experiments were performed on an iMac computer with a 4.0 GHz Intel I7 processor, with 64 GB of memory. The computing times are plotted against N in Fig. 5(a). With the exceptions of small problem sizes, matching and IHA are found to be faster than Sinkhorn, with matching becoming faster as the problem size increases. We have assigned this difference to the fact that Sinkhorn is known to slow down significantly for very small ϵ values, despite the log-stabilization and ϵ scaling, as illustrated in Fig. 5(b). Matching and our implementation of IHA use the same strategy of solving the nonlinear system of equations to find the saddle point with an iterative Newton's approach. We have tried a steepest descent approach as well as a rewriting of IHA using s Sinkhorn iterative scheme, as proposed in Ref. [12]), but found that those implementations were slower than the Newton's approach described here. Matching is faster than IHA for large values of N; we believe that this is a consequence of the modified free-energy functional we have introduced. This will be discussed below.

D. Solving pathological assignment problems

All the numerical experiments presented above relate to assignment problems with random cost matrices drawn from exponential distributions. To further analyze the behavior and efficiency of our approach, we repeated our analyses on two other types of cost matrices, namely, real matrices whose elements are drawn from the Cauchy distribution, and integer matrices whose elements are drawn uniformly from a given interval.

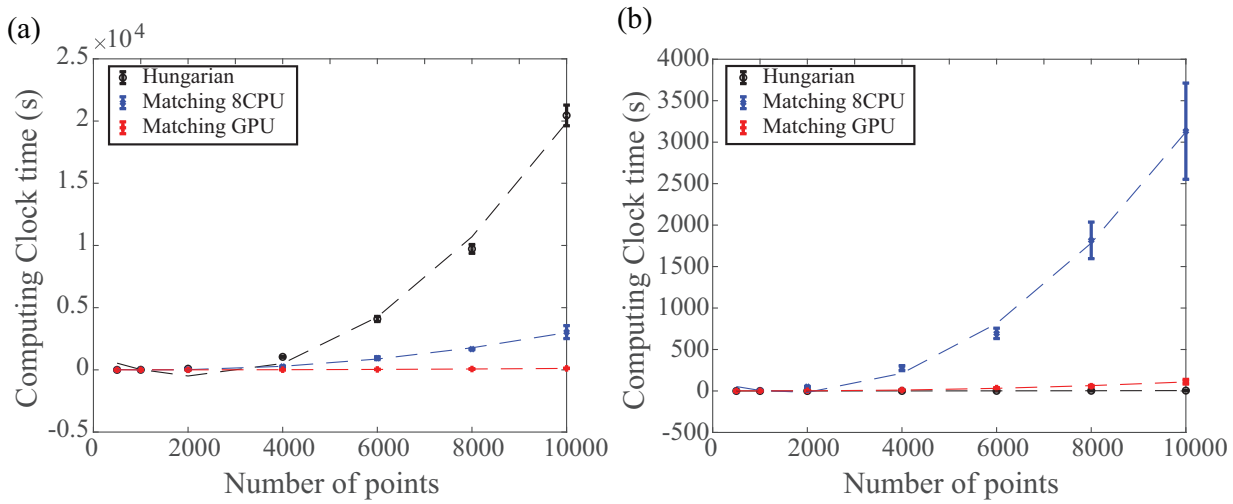

FIG. 6. CPU time for solving pathological assignment problems. We compare the computing times of the Hungarian algorithm (black) with the computing times of two versions of our matching algorithm, one running on an 8-core CPU (blue) and one running on GPU (red), when those algorithms are applied to two types of cost matrices, random real matrices whose elements are drawn from a standard Cauchy distribution (a), and random integer matrices whose elements are drawn uniformly from the interval [0,10] (b). The mean computing times (clock time) over five independent calculations are plotted against the sizes of the cost matrices. The dashed lines represent quadratic polynomial fits to the means. Technical details are provided in the caption of Fig. 3.

The standard Cauchy distribution is defined with the probability distribution

$$f(x) = \frac{1}{\pi(1+x^2)}.$$

It is a canonical example of a "pathological" distribution since both its expected value and its variance are undefined. We ran simulations on random cost matrices whose elements are drawn from this standard Cauchy distribution. Those matrices vary in sizes between 50×50 and $10\,000 \times 10\,000$. We ran five independent simulations for each size. To our knowledge, there are no known theoretical results on the expected values of the optimal cost for the assignment problems associated to those cost matrices. For each experiment, we have then verified that we obtained the correct optimal assignment cost by running in parallel the Hungarian algorithm. We note that for all those experiments, the Hungarian and our algorithm not only found the same optimal cost but also the same assignment, hinting that these assignment problems have a unique solution. The computing times for the Hungarian algorithm and for the two versions of matching (i.e., CPU-based and GPU-based), averaged over five independent simulations, are plotted against the size N of the assignment problem in Fig. 6, left panel. Much akin to the simulations based on random cost matrices derived from exponential distributions, we observe that matching provides a significant speed improvement compared to the Hungarian algorithm. This improvement is a consequence of the fact that matching benefits from parallelization (see above): the difference between applications of the Hungarian algorithm and of the GPU-based matching algorithm is of order 200 in favor of the latter for matrices of size $10\,000 \times 10\,000$.

We observe very different behaviors, however, when we consider random integer matrices. We ran simulations on such random cost matrices whose integer elements are drawn uniformly in the interval $[0, M]$, with $M = 10$, and with sizes N ranging in size between 50×50 and $10\,000 \times 10\,000$. For all those simulations, simple applications of the matching algorithm lead to non integer assignment matrices, indicative of the fact that the corresponding cost matrices are degenerate. We applied the method described in Sec. IV (i.e., addition of small random noise to the cost matrix) to identify an integer assignment with the same optimal cost. We note also that in all cases, the solutions found by the Hungarian algorithm and by matching had the same optimal cost but different assignments. The computing times for the Hungarian algorithm and for the two versions of matching (i.e., CPU-based and GPU-based), averaged over five independent simulations, are plotted against the size N of the assignment problem in Fig. 6, right panel. In opposition to the random assignment problems based on real matrices, the Hungarian algorithm was always found to be faster than matching, for all matrix sizes considered. The Hungarian algorithm is an algorithm that proceeds by iteratively removing ambiguities when attempting assignments between "agents" and "tasks" through modifications of the cost matrix that do not affect the optimal solution. Those modifications proceeds by subtractions between rows or between columns to reach values of zeros, and an unambiguous zero defines an assignment. When the matrix elements are integer values, drawn from a small interval, the chances of getting many zeros when performing those operations are significantly higher than if the matrix elements are real. The Hungarian algorithm greatly benefits from this fact, while matching handles integer values as if they were real values. Figure 6 shows that the Hungarian algorithm is significantly faster than the two implementations of matching, with a speedup of approximatively 700 compared to the 8-CPU version, and of approximately 20 for the GPU version, for matrices of size $10\,000 \times 10\,000$. We also investigated the importance of M, that defines the size of the interval from which the random integer elements of the cost matrices are drawn. Results are shown in Fig. 7. As matching does not

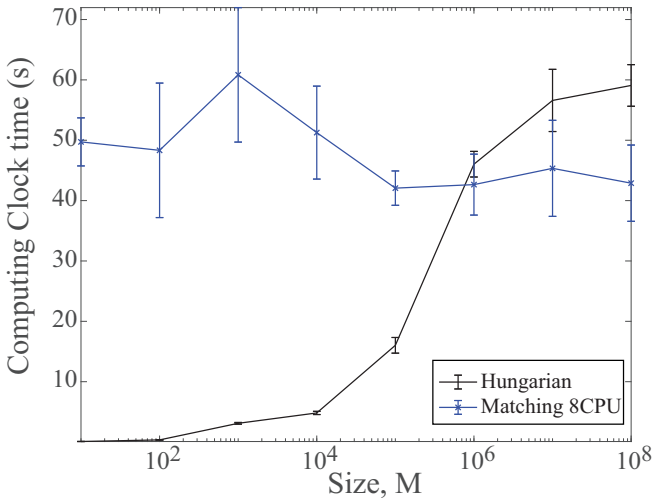

FIG. 7. CPU time for solving integer assignment problems. We compare the computing times of the Hungarian algorithm (black) with the computing times of the multicore CPU version of our matching algorithm,when those algorithms are applied to random integer matrices of sizes 2000×2000 whose elements are drawn uniformly from the interval $[0, M]$. The mean computing times (clock time) over five independent calculations are plotted against the sizes M of the intervals from which the elements of the cost matrices are drawn. note that the computing times for matching remain constant, while those for the Hungarian algorithm increase as M increases. Technical details are provided in the caption of Fig. 3.

differentiate if the cost matrix is real or integer, we observe that its computing cost is independent of M, for a given matrix size. The Hungarian algorithm, however, is found to be strongly dependent on the value of M, with computing times that increase as M increases. This corroborate our assertion that the diversity inside the cost matrix affects the performance of the Hungarian algorithm.

VII. DISCUSSION

In this paper, we have proposed a statistical physics framework to solve the balanced assignment problem. Given two sets of points S_1 and S_2 with the same cardinality N, and a cost matrix between those sets, we have constructed a weakly concave free energy parametrized by temperature that captures the constraints of the assignment problem. Its maximum defines an optimal assignment between the two sets of points. We proved that this free energy decreases monotonically as a function of β (the inverse of temperature) to the optimal assignment cost, providing a robust framework for temperature annealing. We proved also that for large enough β values (i.e., small enough temperature), the exact solution to the generic assignment problem can be derived directly from the maximum of the free energy using simple roundoff to the nearest integer of the elements of the assignment matrix associated with this maximum. We have also derived a provably convergent method to handle degenerate assignment problems, with a characterization of those problems. We have described two computer implementations of our framework that are optimized for parallel architectures, one based on CPU, the other based on GPU, and have shown that the latter enables solving large assignment problems (of the orders of a few 10 000 s) in computing clock times of the orders of minutes.

A. Comparison with other algorithms coming from physics

Statistical physics provides a framework for addressing otherwise difficult optimization problems. For example, statistical physicists have long been interested in the assignment problem (for examples, see Refs. [12,19,26–28]. Of direct relevance to this paper, the "invisible hand algorithm" [12], solves the assignment problem using a statistical physics approach similar to the one we have proposed. Both approaches use temperature schemes that are provably guaranteed to converge to the exact assignment solution at zero temperature for generic problems. For both approaches, schemes are proposed to extract the exact solution in bounded computing time. While we have expanded beyond generic assignment problems with guaranteed unique solution by building a provably convergent scheme for solving degenerate assignment problems (see Sec. IV), the main differences between our method and invisible hand algorithm sit elsewhere and are worth discussing. Both methods rely on the construction of a temperature-dependent free energy, weakly convex for the invisible hand algorithm and weakly concave in our case. While energy functions are derived using different formalisms when constructing the partition function for the system considered, they do take similar forms. If C is the cost matrix between the two sets of points considered, and G is an assignment matrix between those two sets, then the free-energy functionals take the form

$$F(\beta) = \sum_{kl} C(k,l)G(k,l) - \frac{1}{\beta}\sum_{kl} s[G(k,l)] + \sum_k \lambda_k \left[\sum_l G(k,l) - 1\right] + \sum_l \mu_l \left[\sum_k G(k,l) - 1\right], \tag{46}$$

where β is the inverse of the temperature T. Note that we do not write the exact formulation given in Ref. [12], but an equivalent form proposed by Ref. [28]. From a physics point of view, this form for the free energy is intuitive: the first term is the internal energy, i.e., the assignment cost that needs to be minimized, the second term is an entropic term, which can be seen as a regularization term that renders the problem convex (or concave) as well as a barrier function that will prevent the $G(k,l)$ to take some values, and the third and fourth terms impose the row sums and column sums constraints via Lagrange multipliers, respectively. The two energy functions differ in expression of the function $s(x)$ that encodes the entropy.

In the invisible hand algorithm, the function $s(x) = -x\ln(x)$, namely, takes the traditional form of the Gibbs entropy. It serves as a barrier at zero, thereby maintaining the positivity of the $G(k,l)$. Interestingly, with this formulation, the invisible hand algorithm is equivalent to the entropy regularized method that was proposed for solving the optimal transport (OT) problem, i.e., a generalized assignment problem not limited to binary assignments. Just like for the invisible hand algorithm, the entropic penalization for the OT

problem has the advantage that it defines a strongly convex problem with a unique solution [25]. In addition, its solution can be found efficiently through the so-called iterative proportional fitting procedure [29], also known as the Sinkhorn's algorithm [30], or Sinkhorn-Knopp algorithm [31]. Note that the use of this algorithm has led Cuturi [25] to propose a "Lightspeed Computation of Optimal Transport" (in the title of this paper), which we paraphrased for the title of this paper. Many variants of those algorithms have been developed for solving regularized OT problems; we refer to [32–34] for overviews on those methods. Those algorithms find solutions for a given value of the relaxation parameter ϵ, which plays the role of a temperature. For small values of this parameter, numerical issues can arise and a stabilization of the algorithm is necessary [35]. Despite such stabilization, convergence of a stabilized Sinkhorn-Knopp algorithm can nevertheless be very slow when ϵ is small, and sometimes numerically unstable. Such small values are, however, desirable for finding good approximations to the solution of the original problem. The same difficulty can be mentioned for the invisible hand algorithm, as it can also be solved using the Sinkhorn's algorithm (see Ref. [12]).

In contrast, the function $s(x)$ in our formalism takes the form $s(x) = -x\ln(x) - (1-x)\ln(1-x)$. Note that this is a typical mixture entropy, where the first term is the entropy of "particles," and the second term is the entropy of "holes." As such, it introduces barriers both a zero for positivity and at one to ensure that points are only assigned once. It also provides a simple and stable expression for the terms of the assignment matrix as a function of the internal variables of the free energy, given by the function $h(x) = 1/(1+e^x)$. $h(x)$ is continuous, monotonic, bounded between 0 and 1, and bijective. With this function and the double-barrier entropy function s we consider, we have run routinely computations with temperatures of the order of 10^{-13} without numerical instabilities.

B. Computational complexity: How large can we go?

Our implementations of the method presented in this paper were found to be efficient with nearly optimal use of parallelization, both on CPU and on GPU processors. While we cannot fully take credit for the effectiveness of these implementations as they are based on the highly efficient machine-specific BLAS and LAPACK libraries, we note that the method we have presented here provides the framework for such significant improvements in computing time compared to a serial computation. In addition, the apparent time complexity of those implementations were found to be $O(N^2)$, an improvement compared to the $O(N^3)$ time complexity of the Hungarian algorithm (though this needs to be considered with caution as the former is based on a small sample of empirical running time averages, while the latter is defined theoretically). The space complexity of our implementations is also $O(N^2)$, as we need to store both the cost matrix and a work array of similar size that contains either the assignment matrix, or part of the Jacobian matrix needed to solve the nonlinear systems of equations at the saddle point approximations. Both matrices are of size $N \times N$. Such a requirement limits the use of our implementations to problems of size up to $25\,000 \times 25\,000$. Indeed, with $N = 30\,000$, handling two matrices of size N^2 in double precision requires 14.4 GB of memory, which is beyond the capacity of the GPU cards we have used in our numerical simulations. While GPU cards with larger memory are available (currently up to 32 GB), it remains that a $O(N^2)$ algorithm in memory complexity is ultimately limited to assignment problems of up to a few 10 000 points. Handling larger problem sizes for which the cost matrix is sparse will require some redesign of our algorithm. We will pursue this in future studies.

ACKNOWLEDGMENTS

The work discussed here originated from a visit by P.K. at the Institut de Physique Théorique, CEA Saclay, France. He thanks them for their hospitality and financial support.

APPENDIX A: PROOF OF THEOREM 1: CONCAVITY OF THE EFFECTIVE FREE ENERGY

We first prove that the effective free energy $F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})$ is weakly concave, by showing that its Hessian H is negative semidefinite. H is a symmetric matrix of size $2N \times 2N$, such that its rows and columns correspond to all N λ values first, followed by all N μ values. Let h' be the derivative of the function $h(x) = 1/(1+e^x)$, i.e.,

$$h'(x) = -\frac{e^x}{(1+e^x)^2}. \tag{A1}$$

We note first that $h'(x) \in [\frac{-1}{4}, 0) \quad \forall x \in \mathbb{R}$, i.e., that $h'(x)$ is always strictly negative. We define the matrix X' such that

$$X'(k,l) = h'\{\beta[C(k,l) + \lambda(k) + \mu(l)]\} \tag{A2}$$

From Eqs. (11), we obtain

$$H(k,i) = \frac{\partial^2 F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial\lambda(k)\partial\lambda(i)} = \beta\delta_{ki}\sum_l X'(k,l), \tag{A3}$$

$$H(k,l) = \frac{\partial^2 F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial\lambda(k)\partial\mu(l)} = \beta X'(k,l), \tag{A4}$$

$$H(l,m) = \frac{\partial^2 F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial\mu(l)\partial\mu(m)} = \beta\delta_{lm}\sum_k X'(k,l), \tag{A5}$$

where δ are Kronecker functions, the indices k and i belong to $[1, N]$, and the indices l and m belong to $[1, N]$.

Let $\mathbf{x} = (\mathbf{x_1}, \mathbf{x_2})$ be an arbitrary vector of size $2N$. The quadratic form $Q(\mathbf{x}) = \mathbf{x}^T H \mathbf{x}$ is equal to

$$\begin{aligned} Q(\mathbf{x}) = & \sum_{i,k} x_1(k)H(k,i)x_1(i) + 2\sum_{k,l} x_1(k)H(k,l)x_2(l) \\ & + \sum_{l,m} x_2(l)H(l,m)x_2(m) \\ = & \ \beta\sum_{k,l} x_1(k)^2 H'(k,l) + 2\beta\sum_{k,l} x_1(k)X'(k,l)x_2(l) \\ & + \beta\sum_{k,l} x_2(l)^2 X'(k,l) \\ = & \ \beta\sum_{k,l}[x_1(k) + x_2(l)]^2 X'(k,l). \end{aligned} \tag{A6}$$

As $X'(k,l)$ is based on the function h' that is strictly negative, the summands in the equation above are negative for all k and l, and therefore $Q(\mathbf{x})$ is negative for all vector $\mathbf{x}$. The Hessian H is negative, semidefinite. As a consequence $F_\beta(\boldsymbol{\lambda},\boldsymbol{\mu})$ is (weakly) concave.

As $Q(\mathbf{x})$ is a sum of negative terms, it is 0 if and only if all the terms are equal to 0. This means that $\forall(k,l) \quad x_1(k)+x_2(l)=0$. This is realized when all the coordinates to x_1 are equal, and set to a parameter K, and all the coordinates to x_2 are equal, and set to $-K$. Therefore, 0 is an eigenvalue of H, with eigenvector $\mathbf{x}=(1,\ldots,1,-1,\ldots,-1)$. This eigenvector corresponds to the translation invariance for the free energy. It can be removed by setting one of the parameters $\lambda(k)$ or $\mu(l)$ to zero; the free-energy functional $F_\beta(\boldsymbol{\lambda},\boldsymbol{\mu})$ on this restricted parameter space is then strictly concave.

APPENDIX B: MONOTONICITY OF $F^{\mathrm{MF}}(\beta)$

The effective free energy $F_\beta(\boldsymbol{\lambda},\boldsymbol{\mu})$ defined in Eq. (9) is a function of the cost matrix C and of real unconstrained variables $\lambda(k)$ and $\mu(l)$. For the sake of simplicity, for any $(k,l)\in[1,N]^2$, we define

$$x_{kl}=C(k,l)+\lambda(k)+\mu(l). \tag{B1}$$

The effective free energy is then

$$F_\beta(\boldsymbol{\lambda},\boldsymbol{\mu})=-\left[\sum_k\lambda(k)+\sum_l\mu_l\right]-\frac{1}{\beta}\sum_{kl}\ln(1+e^{-\beta x_{kl}}). \tag{B2}$$

As written above, $F_\beta(\boldsymbol{\lambda},\boldsymbol{\mu})$ is a function of the independent variables β, $\lambda(k)$ and $\mu(l)$. However, under the saddle point approximation, the variables $\lambda(k)$ and $\mu(l)$ are constrained by the conditions

$$\begin{aligned}\frac{\partial F_\beta(\boldsymbol{\lambda},\boldsymbol{\mu})}{\partial\lambda(k)}&=0,\\ \frac{\partial F_\beta(\boldsymbol{\lambda},\boldsymbol{\mu})}{\partial\mu(l)}&=0,\end{aligned} \tag{B3}$$

and the free energy under those constraints is written as $F^{\mathrm{MF}}(\beta)$. In the following, we will use the notations $\frac{dF^{\mathrm{MF}}(\beta)}{d\beta}$ and $\frac{\partial F^{\mathrm{MF}}(\beta)}{\partial\beta}$ to differentiate between the total derivative and partial derivative of $F^{\mathrm{MF}}(\beta)$ with respect to β, respectively. Based on the chain rule,

$$\begin{aligned}\frac{dF^{\mathrm{MF}}(\beta)}{d\beta}=&\frac{\partial F_\beta(\boldsymbol{\lambda},\boldsymbol{\mu})}{\partial\beta}+\sum_k\frac{\partial F_\beta(\boldsymbol{\lambda},\boldsymbol{\mu})}{\partial\lambda(k)}\frac{\partial\lambda(k)}{\partial\beta}\\ &+\sum_l\frac{\partial F_\beta(\boldsymbol{\lambda},\boldsymbol{\mu})}{\partial\mu(l)}\frac{\partial\mu(l)}{\partial\beta}.\end{aligned} \tag{B4}$$

Using the constraints Eq. (B3), we find that

$$\frac{dF^{\mathrm{MF}}(\beta)}{d\beta}=\frac{\partial F_\beta(\boldsymbol{\lambda},\boldsymbol{\mu})}{\partial\beta}, \tag{B5}$$

namely, that the total derivative with respect to β is in this specific case equal to the corresponding partial derivative, which is easily computed to be

$$\frac{dF^{\mathrm{MF}}(\beta)}{d\beta}=\frac{1}{\beta^2}\sum_{kl}\left[\ln\left(1+e^{-\beta x_{kl}^{\mathrm{MF}}}\right)+\frac{\beta x_{kl}^{\mathrm{MF}}}{1+e^{\beta x_{kl}^{\mathrm{MF}}}}\right]. \tag{B6}$$

Let $t(x)=\ln(1+e^{-x})+\frac{x}{1+e^x}$. The function $t(x)$ is continuous and defined over all real values x and is bounded below by 0 (see Fig. 1), i.e., $t(x)\geqslant 0 \quad \forall x\in\mathbb{R}$.

As

$$\frac{dF^{\mathrm{MF}}(\beta)}{d\beta}=\frac{1}{\beta^2}\sum_{kl}t(\beta x_{kl}), \tag{B7}$$

we conclude that

$$\frac{dF^{\mathrm{MF}}(\beta)}{d\beta}\geqslant 0, \tag{B8}$$

namely, that $F^{\mathrm{MF}}(\beta)$ is a monotonically increasing function of β.

APPENDIX C: MONOTONICITY OF $U^{\mathrm{MF}}(\beta)$

Let

$$U_\beta(\boldsymbol{\lambda},\boldsymbol{\mu})=\sum_{kl}C(k,l)h(\beta x_{kl}), \tag{C1}$$

where we have used the same definition for $x_{kl}=C(k,l)+\lambda(k)+\mu(l)$ as above, and let the corresponding mean-field approximation of the internal energy at the saddle point,

$$U^{\mathrm{MF}}(\beta)=U_\beta(\boldsymbol{\lambda}^{\mathrm{MF}},\boldsymbol{\mu}^{\mathrm{MF}}). \tag{C2}$$

Before computing $\frac{dU^{\mathrm{MF}}(\beta)}{d\beta}$, we prove the following property.

Property 3.

$$U^{\mathrm{MF}}(\beta)=F^{\mathrm{MF}}(\beta)+\beta\frac{dF^{\mathrm{MF}}(\beta)}{d\beta}, \tag{C3}$$

i.e., it extends the well-known relationship between the free energy and the average energy to their mean-field counterparts.

Proof. Using Eqs. (B2) and (B6), and the definition of $h(x)=1/(1+e^x)$, we find that

$$\begin{aligned}\beta\frac{dF^{\mathrm{MF}}(\beta)}{d\beta}=&-F^{\mathrm{MF}}(\beta)-\sum_k\lambda^{\mathrm{MF}}(k)-\sum_l\mu^{\mathrm{MF}}(l)\\ &+\sum_{kl}x_{kl}^{\mathrm{MF}}h\left(\beta x_{kl}^{\mathrm{MF}}\right).\end{aligned} \tag{C4}$$

Let us recall that

$$x_{kl}^{\mathrm{MF}}=C(k,l)+\lambda_k^{\mathrm{MF}}+\mu_l^{\mathrm{MF}}.$$

In addition, all mean-field values correspond to the maximum of the effective free energy, for which the constraints

are satisfied, namely, $\sum_l h(\beta x_{kl}^{\mathrm{MF}}) = 1$ and $\sum_k h(\beta x_{kl}^{\mathrm{MF}}) = 1$. Replacing in Eq. (C4), we get

$$\begin{aligned}\beta\frac{dF^{\mathrm{MF}}(\beta)}{d\beta} &= -F^{\mathrm{MF}}(\beta) - \sum_{kl}\lambda^{\mathrm{MF}}(k)h\big(\beta x_{kl}^{\mathrm{MF}}\big)\\ &\quad - \sum_{kl}\mu^{\mathrm{MF}}(l)h\big(\beta x_{kl}^{\mathrm{MF}}\big)\\ &\quad + \sum_{kl}(C(k,l)+\lambda^{\mathrm{MF}}(k)+\mu^{\mathrm{MF}}(l))h\big(\beta x_{kl}^{\mathrm{MF}}\big),\end{aligned} \tag{C5}$$

i.e.,

$$\begin{aligned}\beta\frac{dF^{\mathrm{MF}}(\beta)}{d\beta} &= -F^{\mathrm{MF}}(\beta) + \sum_{kl}C(k,l)h\big(\beta x_{kl}^{\mathrm{MF}}\big)\\ &= -F^{\mathrm{MF}}(\beta) + U^{\mathrm{MF}}(\beta),\end{aligned} \tag{C6}$$

which concludes the proof. ■

Based on the chain rule,

$$\begin{aligned}\frac{dU^{\mathrm{MF}}(\beta)}{d\beta} &= \frac{\partial U^{\mathrm{MF}}(\beta)}{\partial\beta} + \sum_k \frac{\partial U^{\mathrm{MF}}(\beta)}{\partial\lambda(k)}\frac{\partial\lambda(k)}{\partial\beta}\\ &\quad + \sum_l \frac{\partial U^{\mathrm{MF}}(\beta)}{\partial\mu(l)}\frac{\partial\mu(l)}{\partial\beta}.\end{aligned} \tag{C7}$$

Let us compute all partial derivatives in this equation using Proposition 4:

$$\begin{aligned}\frac{\partial U^{\mathrm{MF}}(\beta)}{\partial\lambda(k)} &= \frac{\partial F^{\mathrm{MF}}(\beta)}{\partial\lambda(k)} + \beta\frac{\partial}{\partial\lambda_k}\left(\frac{\partial F^{\mathrm{MF}}(\beta)}{\partial\beta}\right)\\ &= \frac{\partial F^{\mathrm{MF}}(\beta)}{\partial\lambda(k)} + \beta\frac{\partial}{\partial\beta}\left(\frac{\partial F^{\mathrm{MF}}(\beta)}{\partial\lambda(k)}\right)\\ &= 0,\end{aligned} \tag{C8}$$

where the zero is a consequence of the SPA constraints. Similarly, we find

$$\frac{\partial U^{\mathrm{MF}}(\beta)}{\partial\mu(l)} = 0. \tag{C9}$$

Finally,

$$\begin{aligned}\frac{\partial U^{\mathrm{MF}}(\beta)}{\partial\beta} &= 2\frac{\partial F^{\mathrm{MF}}(\beta)}{\partial\beta} + \beta\frac{\partial}{\partial\beta}\left(\frac{\partial F^{\mathrm{MF}}(\beta)}{\partial\beta}\right)\\ &= 2\frac{\partial F^{\mathrm{MF}}(\beta)}{\partial\beta}\\ &\quad +\beta\left(-\frac{2}{\beta}\frac{\partial F^{\mathrm{MF}}(\beta)}{\partial\beta} + \frac{1}{\beta^2}\sum_{kl}\beta x_{kl}^{\mathrm{MF}}t'\big(\beta x_{kl}^{\mathrm{MF}}\big)\right),\end{aligned} \tag{C10}$$

i.e.,

$$\frac{\partial U^{\mathrm{MF}}(\beta)}{\partial\beta} = \frac{1}{\beta}\sum_{kl}\beta x_{kl}^{\mathrm{MF}}t'\big(\beta x_{kl}^{\mathrm{MF}}\big), \tag{C11}$$

where f is defined above [see Eq. (B6)]. As $t'(x) = -\frac{x}{(1+e^x)^2}$, we get

$$\frac{\partial U^{\mathrm{MF}}(\beta)}{\partial\beta} = -\frac{1}{\beta}\sum_{kl}\frac{\big(x_{kl}^{\mathrm{MF}}\big)^2}{(1+e^{\beta x^{\mathrm{MF}}(k,l)})^2}. \tag{C12}$$

Therefore,

$$\frac{dU^{\mathrm{MF}}(\beta)}{d\beta} = \frac{\partial U^{\mathrm{MF}}(\beta)}{\partial\beta} \leqslant 0, \tag{C13}$$

and the function $U^{\mathrm{MF}}(\beta)$ is a monotonically decreasing function of β.

APPENDIX D: PROOF OF THEOREM 3: CONVERGENCE OF THE MEAN-FIELD FREE ENERGY AND THE INTERNAL ENERGY TO THE OPTIMAL ASSIGNMENT COST

We prove first that the optimal assignment energy U^* is equal to the limit of the mean-field free energy when the inverse temperature $\beta \to +\infty$. For simplicity in notation, we define $F^{\mathrm{MF}}(\infty) = \lim_{\beta\to+\infty} F^{\mathrm{MF}}(\beta)$.

We first prove that $U^* \leqslant F^{\mathrm{MF}}(\infty)$.

Let $U^{\mathrm{MF}}(\beta)$ be the mean-field internal energy at the inverse temperature β:

$$U^{\mathrm{MF}}(\beta) = \sum_{k,l} C(k,l)X_\beta^{\mathrm{MF}}(k,l), \tag{D1}$$

where X_β^{MF} is the solution to the SPA system of equations. At a finite inverse temperature β, X_β^{MF} is strictly nonintegral, as each of its terms is of the form $h\beta(x_{kl})$, where $h(x) = 1/(1+e^x)$, and therefore strictly in (0,1). However, X_β^{MF} satisfies the constraints on row sums and column sums, it is a doubly stochastic matrix. The set $\mathcal{S}_N$ of doubly stochastic matrices of size $N \times N$ forms a convex polytope that is the convex hull of the set of permutation matrices. In addition, the vertices of $\mathcal{S}_N$ are exactly the permutation matrices (Birkhoff–von Neumann theorem, see Ref. [36]). Therefore, X_β^{MF} can be written as a linear combination of the permutation matrices $\pi_k \in \Pi_N$,

$$X_\beta^{\mathrm{MF}} = \sum_{\pi\in\Pi_N} a_\pi \pi, \tag{D2}$$

with all $a_\pi \in [0,1]$ and $\sum_{\pi\in\Pi_N} a_\pi = 1$. The summation extends over all $N!$ permutations in Π_N. Therefore,

$$\begin{aligned}U^{\mathrm{MF}}(\beta) &= \sum_{k,l} C(k,l)X_\beta^{\mathrm{MF}}(k,l)\\ &= \sum_{\pi\in\Pi_N} a_\pi \sum_k C[k,\pi(k)].\end{aligned} \tag{D3}$$

As U^* is the minimum matching cost over all possible permutations of $\{1,N\}$, for all $\pi \in \Pi_N$, we have

$$\sum_k C[k,\pi(k)] \geqslant U^*. \tag{D4}$$

Combining Eqs. (D3) and (D4), we get

$$U^{\mathrm{MF}}(\beta) \geqslant \sum_{\pi\in\Pi_N} a_\pi U^*, \tag{D5}$$

from which we conclude that at each β,

$$U^* \leqslant U^{\mathrm{MF}}(\beta). \tag{D6}$$

The mean-field free energy and internal energy are related by Eq. (17). In this equation, the entropy can be written as

$$S^{\mathrm{MF}}(\beta) = -\sum_{kl} X_\beta^{\mathrm{MF}}(k,l)\ln\left[X_\beta^{\mathrm{MF}}(k,l)\right] - \sum_{kl}\left[1-X_\beta^{\mathrm{MF}}(k,l)\right]\ln\left[1-X_\beta^{\mathrm{MF}}(k,l)\right]. \tag{D7}$$

This entropy is positive (see Fig. 1) and satisfies the following constraints:

$$0 \leqslant S^{\mathrm{MF}}(\beta) \leqslant N^2\ln(2). \tag{D8}$$

Using Eq. (17), after rearrangement we obtain

$$U^{\mathrm{MF}}(\beta) - \frac{1}{\beta}N^2\ln(2) \leqslant F^{\mathrm{MF}}(\beta) \leqslant U^{\mathrm{MF}}(\beta). \tag{D9}$$

Taking the limits when $\beta \to +\infty$, we get

$$F^{\mathrm{MF}}(\infty) = U^{\mathrm{MF}}(\infty), \tag{D10}$$

and since $U^* \leqslant U^{\mathrm{MF}}(\beta)$ for all β, $U^* \leqslant F^{\mathrm{MF}}(\infty)$.

We now prove the converse inequality, $F^{\mathrm{MF}}(\infty) \leqslant U^*$. Let us first recall the definition of the free energy,

$$F_\beta(\boldsymbol{\lambda},\boldsymbol{\mu}) = -\sum_k \lambda(k) - \sum_l \mu(l) - \frac{1}{\beta}\sum_{kl}\ln\left(1+e^{-\beta[C(k,l)+\lambda(k)+\mu(l)]}\right). \tag{D11}$$

For sake of clarity, let us write again $x(k,l) = C(k,l) + \lambda(k) + \mu(l)$. Note first this property of limits:

$$\lim_{\beta\to+\infty}\frac{\ln(1+e^{-a\beta})}{\beta} = \begin{cases} 0 & \text{if } a \geqslant 0, \\ -a & \text{if } a \leqslant 0. \end{cases} \tag{D12}$$

Therefore,

$$\lim_{\beta\to+\infty} F_\beta(\boldsymbol{\lambda},\boldsymbol{\mu}) = -\sum_k \lambda(k) - \sum_l \mu(l) + \sum_{kl|x(k,l)\leqslant 0} x(k,l). \tag{D13}$$

In this limit, the third term on the right only includes the terms $C(k,l)+\lambda(k)+\mu(l)$ that are negative.

Let us consider a permutation π of $\{1,N\}$. We can write

$$\sum_{k=1}^N C[k,\pi(k)] = \sum_{k=1}^N \{C(k,l)+\lambda(k)+\mu[\pi(k)]\} - \sum_k \lambda(k) - \sum_l \mu(l), \tag{D14}$$

i.e.,

$$\sum_{k=1}^N C[k,\pi(k)] = \sum_{k=1}^N x[k,\pi(k)] - \sum_k \lambda(k) - \sum_l \mu(l). \tag{D15}$$

For each index k, the summand included in the first term on the right is always larger or equal to the sum of all the corresponding terms that are negative:

$$x(k,\pi(k)) \geqslant \sum_{l|x(k,l)\leqslant 0} x(k,l). \tag{D16}$$

Therefore,

$$\sum_{k=1}^N C[k,\pi(k)] \geqslant \sum_{kl|x(k,l)\leqslant 0} x(k,l) - \sum_k \lambda(k) - \sum_l \mu(l), \tag{D17}$$

i.e.,

$$\sum_{k=1}^N C[k,\pi(k)] \geqslant \lim_{\beta\to+\infty} F_\beta(\boldsymbol{\lambda},\boldsymbol{\mu}), \tag{D18}$$

where this inequality follows from Eq. (D13).

Equation (D18) is valid for all permutations π: It is therefore valid for the optimal permutation π^* that solves the assignment problem. Since $U^* = \sum_k C(k,\pi^*(k)$, we have

$$U^* \geqslant \lim_{\beta\to+\infty} F_\beta(\boldsymbol{\lambda},\boldsymbol{\mu}). \tag{D19}$$

As this equation is true for all $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$, it is true in particular for $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{\mathrm{MF}}$ and $\boldsymbol{\mu} = \boldsymbol{\mu}^{\mathrm{MF}}$, leading to

$$U^* \geqslant \lim_{\beta\to+\infty} F^{\mathrm{MF}}(\beta) = F^{\mathrm{MF}}(\infty). \tag{D20}$$

We have shown that $U^* \leqslant F^{\mathrm{MF}}(\infty)$ and $F^{\mathrm{MF}}(\infty) \leqslant U^*$; therefore, $U^* = F^{\mathrm{MF}}(\infty)$. The corresponding result for the internal energy, $U^* = U^{\mathrm{MF}}(\infty)$ follows directly from Eq. (D10).

APPENDIX E: PROOF OF THEOREM 4: BOUNDS ON THE ENTROPY, INTERNAL ENERGY, AND FREE ENERGY

1. Bounds on the entropy

In the previous Appendix, we have already derived bounds on the entropy, see Eq. (D13). These bounds were found from the behavior of the function $J(x)$ that defines the entropy, which is bound in the interval $[0,\ln(2)]$. However, a tighter upper bound can be found by noticing that the values of the variable x, i.e., the different $X_\beta(k,l)$ are constrained. Using Lagrange multipliers to optimize the entropy $S = \sum_{kl} J[X(k,l)]$ under the constraints $\sum_l X(k,l) = 1$ and $\sum_k X(k,l) = 1$, we find that the maximum is found when $X(k,l) = 1/N$, in which case,

$$S^{\mathrm{MF}}(\beta) \leqslant N^2 J\left(\frac{1}{N}\right) \leqslant N^2\left[-\frac{1}{N}\ln\left(\frac{1}{N}\right) - \left(1-\frac{1}{N}\right)\ln\left(1-\frac{1}{N}\right)\right] \leqslant A(N), \tag{E1}$$

where we have defined $A(N) = N^2\ln(N) - N(N-1)\ln(N-1)$. As the entropy is positive, we conclude

$$0 \leqslant S^{\mathrm{MF}}(\beta) \leqslant A(N). \tag{E2}$$

2. Bounds on the free energy

In Appendix C, we have shown that [see Eq. (C6)]

$$\beta \frac{dF^{\mathrm{MF}}(\beta)}{d\beta} = -F^{\mathrm{MF}}(\beta) + U^{\mathrm{MF}}(\beta). \tag{E3}$$

Using this equation and the relationship between free energy, energy, and entropy at SPA [see Eq. (17)], we obtain

$$\frac{dF^{\mathrm{MF}}(\beta)}{d\beta} = \frac{1}{\beta^2} S^{\mathrm{MF}}(\beta). \tag{E4}$$

From the bounds on the entropy,

$$0 \leqslant \frac{dF^{\mathrm{MF}}(\beta)}{d\beta} \leqslant \frac{A(N)}{\beta^2}. \tag{E5}$$

By integrating over β between β and $+\infty$,

$$0 \leqslant F^{\mathrm{MF}}(\infty) - F^{\mathrm{MF}}(\beta) \leqslant \frac{A(N)}{\beta}. \tag{E6}$$

Finally, as $F^{\mathrm{MF}}(\infty) = U^*$,

$$U^* - \frac{A(N)}{\beta} \leqslant F^{\mathrm{MF}}(\beta) \leqslant U^*. \tag{E7}$$

3. Bounds on the energy

As $U^{\mathrm{MF}}(\beta) = F^{\mathrm{MF}}(\beta) + \frac{1}{\beta} S^{\mathrm{MF}}(\beta)$, using the inequalities in Eqs. (E2) and (E7), we get

$$U^{\mathrm{MF}}(\beta) \leqslant U^* + \frac{A(N)}{\beta}. \tag{E8}$$

In addition, as $U^{\mathrm{MF}}(\beta)$ is monotonic, decreasing, with limit U^* as $\beta \to +\infty$, $U^* \leqslant U^{\mathrm{MF}}(\beta)$. Therefore,

$$U^* \leqslant U^{\mathrm{MF}}(\beta) \leqslant U^* + \frac{A(N)}{\beta}. \tag{E9}$$

APPENDIX F: PROOF OF THEOREM 5: BOUNDS ON ASSIGNMENT MATRIX X_β^{MF}

This proof is inspired by the proof of Theorem 6 in Appendix 2 of Ref. [12].

We first recall that X_β^{MF} is a doubly stochastic matrix, it can be written as a linear combination of the permutation matrices $\pi_k \in \Pi_N$,

$$X_\beta^{\mathrm{MF}} = \sum_{\pi \in \Pi_N} a_\pi \pi_k, \tag{F1}$$

with all $a_\pi \in [0, 1]$ and $\sum_{\pi \in \Pi_N} a_\pi = 1$ (see Appendix D for details).

To prove that $\max_{k,l} |X_\beta^{\mathrm{MF}}(k, l) - G^*(k, l)| \leqslant \frac{A(N)}{\beta\Delta}$, where G^* is the optimal solution of the assignment problem, $\Delta = U^{2*} - U^*$ the difference in total cost between the second best solution and the optimal solution ($\Delta > 0$ as we have assumed that the assignment problem has a unique solution), and $A(N) = N^2 \ln(N) - N(N-1)\ln(N-1)$, we use a proof by contradiction. We assume that there exists a pair (i, j) such that

$$\frac{A(N)}{\beta\Delta} < \left| X_\beta^{\mathrm{MF}}(i, j) - G^*(i, j) \right|. \tag{F2}$$

Let us denote $B(i, j) = |X_\beta^{\mathrm{MF}}(i, j) - G^*(i, j)|$. As G^* is a permutation matrix, $G^*(i, j) = 0$ or $G^*(i, j) = 1$.

In the first case,

$$\begin{aligned} B(i, j) &= X_\beta^{\mathrm{MF}}(i, j) \\ &= \sum_{\pi \in \Pi_N} a_\pi \pi(i, j). \end{aligned} \tag{F3}$$

Since G^* is a permutation matrix, it is included in the decomposition of X_β^{MF}, and therefore,

$$\begin{aligned} B(i, j) &= a_{G^*} G^*(i, j) + \sum_{\pi \in \Pi_N - \{G^*\}} a_\pi \pi(i, j) \\ &= \sum_{\pi \in \Pi_N - \{G^*\}} a_\pi \pi(i, j) \\ &< \sum_{\pi \in \Pi_N - \{G^*\}} a_\pi = 1 - a_{G^*}, \end{aligned} \tag{F4}$$

where the final equality follows from the fact that the sum of all coefficients a is equal to 1.

In the second case, $G^*(i, j) = 1$,

$$\begin{aligned} B(i, j) &= 1 - X_\beta^{\mathrm{MF}}(i, j) \\ &= 1 - \sum_{\pi \in \Pi_N} a_\pi \pi(i, j). \end{aligned} \tag{F5}$$

Again, as G^* is included in the decomposition of X_β^{MF},

$$\begin{aligned} B(i, j) &= 1 - a_{G^*} G^*(i, j) - \sum_{\pi \in \Pi_N - \{G^*\}} a_\pi \pi(i, j) \\ &= 1 - a_{G^*} - \sum_{\pi \in \Pi_N - \{G^*\}} a_\pi \pi(i, j) \\ &< 1 - a_{G^*}, \end{aligned} \tag{F6}$$

where the final inequality follows from the fact that $\sum_{\pi \in \Pi_N - \{G^*\}} a_\pi \pi(i, j)$ is positive.

In both cases, we have

$$\frac{A(N)}{\beta\Delta} < 1 - a_{G^*}. \tag{F7}$$

Now, let us look at the energy associated with X_β^{MF}:

$$\begin{aligned} U^{\mathrm{MF}}(\beta) &= \sum_{kl} C(k, l) X_\beta^{\mathrm{MF}}(k, l) \\ &= \sum_{\pi \in \Pi_N} a_\pi \sum_k C[k, \pi(k)] \\ &= a_{G^*} U^* + \sum_{\pi \in \Pi_N - \{G^*\}} a_\pi \sum_k C[k, \pi(k)] \\ &\geqslant a_{G^*} U^* + \left(\sum_{\pi \in \Pi_N - \{G^*\}} a_\pi \right) U^{2*} \\ &\geqslant a_{G^*} U^* + (1 - a_{G^*}) U^{2*} \\ &\geqslant U^* + (1 - a_{G^*})\Delta. \end{aligned} \tag{F8}$$

In Theorem 4, we have shown that

$$U^* \leqslant U^{\mathrm{MF}}(\beta) \leqslant U^* + \frac{A(N)}{\beta}. \tag{F9}$$

Therefore,

$$U^* + (1 - a_{G^*})\Delta \leqslant U^* + \frac{A(N)}{\beta}, \tag{F10}$$

i.e.,

$$(1 - a_{G^*}) \leqslant \frac{A(N)}{\beta\Delta}, \tag{F11}$$

as Δ is strictly positive.

We have shown that $\frac{A(N)}{\beta\Delta} < 1 - a_{G^*}$ [Eq. (G3)] and $(1 - a_{G^*}) \leqslant \frac{A(N)}{\beta\Delta}$ [Eq. (G6)], i.e., we have reached a contradiction. Our hypothesis is wrong, and therefore $\max_{k,l} |X_\beta^{\rm MF}(k,l) - G^*(k,l)| \leqslant \frac{A(N)}{\beta\Delta}$.

APPENDIX G: PROOF OF THEOREM 7: SIMPLE TERMINATION CRITERIA FOR THE GENERIC ASSIGNMENT PROBLEM

Let us start by proving the following lemma (note that this lemma is at the core of the Hungarian algorithm for solving the assignment problem):

Lemma 1. Let S_1 and S_2 be two sets of points with the same cardinality N and let C be a real-valued cost matrix between S_1 and S_2. Let G be an assignment matrix between S_1 and S_2 that satisfies the constraints on row sum and column sum, namely, G is a doubly stochastic matrix, and let $U(G,C)$ be the total cost associated with G, namely, $U(G,C) = \sum_{k,l} C(k,l)G(k,l)$. Let $\mathbf{a}$ and $\mathbf{b}$ be any two real-valued vectors of size N, and let $D_{\mathbf{a},\mathbf{b}}$ be the matrix defined as $D_{\mathbf{a},\mathbf{b}}(k,l) = C(k,l) + a(k) + b(l)$. Then,

$$U(D_{\mathbf{a},\mathbf{b}}, G) = U(C,G) + m, \tag{G1}$$

where $m = \sum_k a(k) + \sum_l b(l)$ is a constant, independent of G.

Proof. From the definition of D

$$\begin{aligned} U(D_{\mathbf{a},\mathbf{b}}, G) &= \sum_{kl} (C(k,l) + a(k) + b(l))G(k,l) \\ &= U(C,G) + \sum_{kl} a(k)G(k,l) \\ &\quad + \sum_{l} b(l)G(k,l) \\ &= U(C,G) + \sum_{k} a(k) \sum_{l} G(k,l) \\ &\quad + \sum_{l} b(l) \sum_{k} G(k,l) \\ &= U(C,G) + \sum_{k} a(k) + \sum_{l} b(l), \end{aligned} \tag{G2}$$

where the last equality comes from the fact that G is doubly stochastic. ■

It is clear from Lemma 1 that solving the assignment problem between S_1 and S_2 with the cost matrix C is equivalent to solving the assignment problem with the cost matrix $D_{\mathbf{a},\mathbf{b}}$. In general this is of little help within our approach to solving the assignment problem, as the latter has no reason to be simpler than the former. There is one significant exception, however, corresponding to the setting of Theorem 7. Indeed, let us consider an inverse temperature β and let $\boldsymbol{\lambda}^{\rm MF}$ and $\boldsymbol{\mu}^{\rm MF}$ be the mean-field solutions at that temperature. Let us suppose that the matrix $X_\beta^{\rm MF}$ is strictly row dominant. We write first what it means to be strictly row dominant. On each row k of $X_\beta^{\rm MF}$, there is one element, which we will write as $\pi(k)$, such that

$$\left|X_\beta^{\rm MF}[k,\pi(k)]\right| > \sum_{l \neq \pi(k)} \left|X_\beta^{\rm MF}(k,l)\right|. \tag{G3}$$

As $X_\beta^{\rm MF}$ satisfies the row sum and row column constraints, the vector $\{\pi(1),\ldots,\pi(N)\}$ forms a permutation of $\{1,\ldots,N\}$.

As all $X_\beta^{\rm MF}(k,l)$ are positive, Eq. (G3) is equivalent to

$$2X_\beta^{\rm MF}[k,\pi(k)] > \sum_l X_\beta^{\rm MF}(k,l). \tag{G4}$$

As the matrix $X_\beta^{\rm MF}$ is a solution to the assignment problem at the inverse temperature β, it satisfies the row constraints, and therefore the sum on the right side is 1, and we have

$$X_\beta^{\rm MF}[k,\pi(k)] > \tfrac{1}{2}. \tag{G5}$$

It is equally easy to show that $X_\beta(k,l) < \frac{1}{2}$ for all $l \neq \pi(k)$. Since $X_\beta(k,l) = h(x_{kl}^{\rm MF})$, where $x_{kl}^{\rm MF} = C(k,l) + \lambda^{\rm MF}(k) + \mu^{\rm MF}(l)$, we get

$$\begin{aligned} &x_{k\pi(k)}^{\rm MF} < 0, \\ &x_{kl}^{\rm MF} > 0 \quad \forall l \neq \pi(k). \end{aligned} \tag{G6}$$

By setting the vectors $\mathbf{a}$ and $\mathbf{b}$ in Lemma 1 to be $\boldsymbol{\lambda}^{\rm MF}$ and $\boldsymbol{\mu}^{\rm MF}$, respectively, we have $D_{\boldsymbol{\lambda}^{\rm MF},\boldsymbol{\mu}^{\rm MF}}(k,l) = x_{kl}^{\rm MF}$, and, therefore,

$$\begin{aligned} &D_{\boldsymbol{\lambda}^{\rm MF},\boldsymbol{\mu}^{\rm MF}}[k,\pi(k)] < 0, \\ &D_{\boldsymbol{\lambda}^{\rm MF},\boldsymbol{\mu}^{\rm MF}}(k,l) \quad \forall l \neq \pi(k). \end{aligned} \tag{G7}$$

As the assignment problem associated with this matrix $D_{\boldsymbol{\lambda}^{\rm MF},\boldsymbol{\mu}^{\rm MF}}$ corresponds to finding the assignment with minimal cost, element k in S_1 is trivially associated with element $\pi(k)$ in S_2, as the corresponding cost is negative and therefore minimal compared to all the other costs $D_{\boldsymbol{\lambda}^{\rm MF},\boldsymbol{\mu}^{\rm MF}}(k,l), l \neq \pi(k)$ that are positive. Therefore, the assignment problem associated with the cost matrix $D_{\boldsymbol{\lambda}^{\rm MF},\boldsymbol{\mu}^{\rm MF}}$ has for solution the permutation matrix Π corresponding to π, and based on Lemma 1, it is also the solution to the original assignment problem.

Finally, we note that since $X_\beta^{\rm MF}[k,\pi(k)] > \frac{1}{2}$ and $X_\beta^{\rm MF}(k,l) < \frac{1}{2}, \quad \forall l \neq \pi(k)$, the permutation matrix Π is constructed from $X_\beta^{\rm MF}$ by simply rounding off its elements to the nearest integer.

APPENDIX H: PROOF OF THEOREM 9: SOLVING THE ASSIGNMENT PROBLEM FOR DEGENERATE COST MATRICES

Let us introduce first some notations. Let S_1 and S_2 be two sets of points with cardinality N, and let C be the cost matrix between S_1 and S_2. The assignment problem between S_1 and S_2 amounts to minimizing $U(G) = \sum_{kl} C(k,l)G(k,l)$, where G is a permutation matrix. We note G^* one solution to this

problem, and U^* the minimum cost associated with G^*. We note U^{2*} the second best cost, such that $\Delta = U^{2*} - U^* > 0$.

Suppose now that we perturb each element of C by a uniform random number:

$$C_\alpha(k,l) = C(k,l) + \alpha\eta(k,l), \tag{H1}$$

where $\eta(k,l)$ is uniform in [0,1] and the different η are independent of each other. The perturbed assignment problem between S_1 and S_2 amounts to minimizing $U_\alpha(G) = \sum_{kl} C_\alpha(k,l)G(k,l)$ where G is a permutation matrix. We note G^*_α one solution to this problem, and U^*_α the minimum cost associated with G^*_α. We first prove the following property:

Property 4. Let S_1 and S_2 be the two sets of points with cardinality N, and let C be the cost matrix between S_1 and S_2. Adding random uniform noise with support $[0,\alpha]$ to each value of C and solving the assignment problem on this perturbed matrix will generate a unique integer solution.

Proof. Let us assume first that the perturbed assignment problem has (at least) two different integer solutions, namely, two permutations π_1 and π_2 such that $U^*_\alpha = U_\alpha(\pi_1) = U_\alpha(\pi_2)$. It is easy to show that the matrix $G_a = a\pi_1 + (1-a)\pi_2$ where a is a real number in (0,1) is also an optimal solution to the perturbed problem. Indeed, G_a is a doubly stochastic matrix as it is a combination of two permutation matrices (see Birkhoff–von Neumann theorem [36]). In addition,

$$\begin{aligned} U_\alpha(G_a) &= \sum_{kl} C_\alpha(k,l)G_a(k,l) \\ &= a\sum_k C_\alpha[k,\pi_1(k)] + (1-a)\sum_k C_\alpha[k,\pi_2(k)] \\ &= aU_\alpha(\pi_1) + (1-a)U_\alpha(\pi_1) = U^*_\alpha. \end{aligned} \tag{H2}$$

Therefore, if the perturbed assignment problem has more than one solution, then it has a solution with fractional components. Based on Proposition 2, this means that there exists (at least) one cycle $A = \{(a_1,b_1),(a_2,b_2),\ldots,(a_{2M},b_{2M})\}$ in the cost matrix C_α for which $\Gamma = \sum_{i=1}^{2M}(-1)^i C_\alpha(a_i,b_i) = 0$, in which case we would have

$$\alpha\sum_{i=1}^{2M}(-1)^i\eta(a_i,b_i) = -\sum_{i=1}^{2M}(-1)^i C(a_i,b_i). \tag{H3}$$

As the variables η are independent random uniform variables and the term on the right side of the equation is constant, the probability to have this linear relationship on the η is 0. Therefore, there are no cycles within the matrix C_α, the perturbed assignment problem does not have solution with fractional value and consequently it has a unique integer solution.

We now provide an upper bound on α such that the solution G^*_α is also an optimal solution to the unperturbed assignment problem. First, we note that G^* and G^*_α are both permutation matrices, and as G^* is one optimal solution to the unperturbed assignment problem,

$$U^* \leqslant U(G^*_\alpha). \tag{H4}$$

Reversely, as G^*_α is the optimal solution to the perturbed assignment problem,

$$U_\alpha(G^*_\alpha) \leqslant U_\alpha(G^*). \tag{H5}$$

From this equation, we have

$$\sum_{kl} C_\alpha(k,l)G^*_\alpha(k,l) \leqslant \sum_{kl} C_\alpha(k,l)G^*(k,l), \tag{H6}$$

which can be rewritten as

$$U(G^*_\alpha) + \alpha\sum_{kl}\eta(k,l)G^*_\alpha(k,l) \leqslant U^* + \alpha\sum_{kl}\eta(k,l)G^*(k,l). \tag{H7}$$

Moving all the terms containing η on the right side,

$$U(G^*_\alpha) \leqslant U^* + \alpha\sum_{kl}\eta(k,l)[G^*(k,l) - G^*_\alpha(k,l)]. \tag{H8}$$

The matrices G^* and G^*_α contains exactly N ones and $N^2 - N$ zeros, therefore there are at most $2N$ nonzero values of the form $G^*(k,l) - G^*_\alpha(k,l)$. As $\eta(k,l) \leqslant 1$, we have

$$U(G^*_\alpha) \leqslant U^* + 2N\alpha. \tag{H9}$$

If we impose that $\alpha < \frac{\Delta}{2N}$, then

$$U(G^*_\alpha) < U^* + \Delta, \tag{H10}$$

i.e.,

$$U(G^*_\alpha) < U^{2*}. \tag{H11}$$

Combining Eqs. (H4) and (H11),

$$U^* \leqslant U(G^*_\alpha) < U^{2*}. \tag{H12}$$

As U^{2*} is by definition the second best cost for the assignment problem, there are no solutions to the assignment problem whose cost is strictly between the optimal cost U^* and the second best cost, U^{2*}. Therefore, $U^* = U(G^*_\alpha)$ and G^*_α is an optimal solution of the unperturbed assignment problem whenever $\alpha < \frac{\Delta}{2N}$. ■

[1] R. Burkard, M. Dell'Amico, and S. Martello, *Assignment Problems* (Society for Industrial and Applied Mathematics, Philadelphia, PA, 2009).

[2] F. Ollivier, Appl. Algebra Eng. Commun. Comput. **20**, 7 (2009).

[3] H. Kuhn, Nav. Res. Logist. **2**, 83 (1955).

[4] M. Fredman and R. Tarjan, J. ACM **34**, 596 (1987).

[5] H. Gabow and R. Tarjan, SIAM J. Comput. **18**, 1013 (1989).

[6] M. Thorup, J. Comput. Syst. Sci. **69**, 330 (2004).

[7] D. Bertsekas, SIAM J. Optim. **1**, 425 (1991).

[8] D. Bertsekas, Comput. Optim. Appl. **1**, 7 (1992).

[9] D. Bertsekas, S. Pallotino, and M. Scutellà, Comput. Optim. Appl. **4**, 99 (1995).

[10] E. Riedy and J. Demmel, *Parallel Weighted Bipartite Matching and Applications*, SIAM Parallel Processing for Scientific Computing (SIAM, Philadelphia, PA, 2004).

[11] L. Ramshaw and R. E. Tarjan, On minimum-cost assignments in unbalanced bipartite graphs, Tech. Rep. HPL-2012-40R1 (HP Labs, Palo Alto, CA, 2012).

[12] J. Kosowsky and A. Yuille, Neural Netw. **7**, 477 (1994).
[13] P. Koehl, M. Delarue, and H. Orland, Phys. Rev. E **100**, 013310 (2019).
[14] P. Koehl, M. Delarue, and H. Orland, Phys. Rev. Lett. **123**, 040603 (2019).
[15] C. Villani, *Topics in Optimal Transportation*, Vol. 58, Graduate Studies in Mathematics (American Mathematical Society, Providence, RI, 2003).
[16] C. Villani, *Optimal Transport: Old and New*, Grundlehren der mathematischen Wissenschaften (Springer, Berlin, 2008).
[17] G. Peyré and M. Cuturi, Found. Trends Mach. Learn. **11**, 355 (2019).
[18] B. Gärtner and J. Matoušek, *Understanding and using Linear Programming* (Springer, Berlin, 2006).
[19] M. Mézard and G. Parisi, J. Phys. Lett. **46**, 771 (1985).
[20] G. Parisi, arXiv:cond-mat/9801176 [cond-mat] (1998).
[21] D. Aldous, Random Struct. Algorithms **18**, 381 (2001).
[22] S. Linusson and J. Wästlund, Probab. Theory Relat. Fields **128**, 419 (2004).
[23] B. Schmitzer, SIAM J. Sci. Comput. **41**, A1443 (2019).
[24] A. Thibault, L. Chizat, C. Dossal, and N. Papadakis, in *NIPS'17 Workshop on Optimal Transport & Machine Learning* (Curran Associates, Inc., San Jose, CA, 2017).
[25] M. Cuturi, in *Advances in Neural Information Processing Systems 26*, edited by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Curran Associates, Red Hook, NY, 2013), pp. 2292–2300.
[26] H. Orland, J. Phys. Lett. **46**, 763 (1985).
[27] M. Mézard and G. Parisi, Europhys. Lett. **2**, 913 (1986).
[28] A. Yuille and J. Kosowsky, Neural Comput. **6**, 341 (1994).
[29] W. E. Deming and F. F. Stephan, Ann. Math. Stat. **11**, 427 (1940).
[30] R. Sinkhorn, Ann. Math. Stat. **35**, 876 (1964).
[31] R. Sinkhorn and P. Knopp, Pac. J. Math. **21**, 343 (1967).
[32] J. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré, SIAM J. Sci. Comput. **37**, A1111 (2015).
[33] A. Genevay, M. Cuturi, G. Peyré, and F. Bach, in *Advances in Neural Information Processing Systems 29* (Curran Associates, Red Hook, NY, 2016), pp. 3440–3448.
[34] P. Dvurechensky, A. Gasnikov, and A. Kroshnin, in *Proceedings of the 35th International Conference on Machine Learning* (PMLR, 2018), pp. 1367–1376.
[35] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard, Math. Comput. **87**, 2563 (2018).
[36] A. Schrijver, *Combinatorial Optimization: Polyhedra and Efficiency* (Springer, Berlin, 2002).