Posterior Agreement for Large Parameter-Rich Optimization Problems

Joachim M. Buhmann[a,1], Julien Dumazert[a], Alexey Gronskiy[a,1,*], Wojciech Szpankowski[b,2]

[a] *ETH Zurich, 8092 Zurich, Switzerland*
[b] *Purdue University, West Lafayette, IN 47907, USA*

Abstract

Most real world combinatorial optimization problems are affected by noise in the input data, thus behaving in the high noise limit like large disordered particle systems, e.g. spin glasses or random networks. Due to uncertainty in the input, optimization of such disordered instances should infer stable posterior distributions of solutions conditioned on the noisy input instance. The maximum entropy principle states that the most stable distribution given the noise influence is defined by the *Gibbs distribution* and it is characterized by the *free energy*. In this paper, we first provide rigorous asymptotics of the difficult problem to compute the free energy for two combinatorial optimization problems, namely the sparse Minimum Bisection Problem (sMBP) and Lawler's Quadratic Assignment Problem (LQAP). We prove that both problems exhibit phase transitions equivalent to the discontinuous behavior of Derrida's Random Energy Model (REM). Furthermore, the derived free energy asymptotics lead to a theoretical justification of a recently introduced concept (Buhmann, 2010) of *Gibbs posterior agreement* that measures stability of the Gibbs distributions when the cost function fluctuates due to randomness in the input. This relatively new stability concept may potentially provide a new method to select robust solutions for a large class of optimization problems.

Keywords: Free energy, Uncertain optimization, Gibbs distribution, Random energy model, Partition function asymptotics, Minimum Bisection, Quadratic Assignment, Information Criterion

*Corresponding author
Email addresses: jbuhmann@inf.ethz.ch (Joachim M. Buhmann), julien.dumazert@gmail.com (Julien Dumazert), alexeygr@inf.ethz.ch (Alexey Gronskiy), szpan@purdue.edu (Wojciech Szpankowski)

[1]This work was supported by SNF Grant # 200021_138117.
[2]This work was supported by NSF Center for Science of Information (CSoI) Grant CCF-0939370, and in addition by NSF Grants CCF-1524312, and NIH Grant 1U01CA198941-01.

1. Introduction

1.1. Overview

Combinatorial optimization arises in many real world settings and these problems are often notoriously difficult to solve due to data dependent noise in the parameters defining such instances. Algorithms that minimize these noisy instances or approximate their global minimum return a solution that is a random variable due to input randomness and that is most often highly unstable. Therefore, we ask the natural questions: What is the distribution of the output returned by the algorithm? Can we stabilize such an output distribution by regularizing the algorithm?

Algorithm design in noise affected real world settings requires both statistical as well as computational considerations: first, we have to ensure that outputs of algorithms are typical in a statistical sense, i.e., they have to occur with high probability. Second, such typical outputs have to be computable in an efficient way with efficient resources. The reader should notice that statistical requirements dominate computational ones in an epistemological sense: A computational result has to be rejected if it is atypical since it lacks predictive power. Computationally, we might require significantly different algorithmic resources (time and space) to calculate typical solutions for typical inputs compared to minimizing the empirical risk .

Due to the statistical nature of inference, we have to efficiently compute posterior distributions of solutions given input data. Open theoretical issues emerge for this strategy, e.g., analytical computation of macroscopic properties like entropy, expected log-partition function or expected costs (Frenk et al., 1985; Talagrand, 2003). The expected log-partition function known also as the *free energy*, appeared in the context of combinatorial optimization since the mid 80's; see e.g., Vannimenus and Mézard (1984) which explored the free energy properties of the traveling salesman problem. An intriguing property of free energy is the emergence of discontinuities of certain order when changing the concentration of the posterior distribution. Such abrupt changes of macroscopic properties, also known as *phase transitions*, are characteristic features of various large systems and have generated a long-lasting interest in theory of discrete structures (see Cohen, 1988; Łuczak, 1994).

The concept "*free energy*" found also applications in theoretical computer science. Recently, in a series of papers on robust learning, Buhmann (2010); Busse et al. (2013) introduced a robustness score function called the *expected log-posterior agreement* (eLPA) for measuring "goodness" of robust solutions. Although the eLPA arose in a different field, it is tightly connected to computing free energies, as we see later in the paper. Furthermore, estimating the free energy for combinatorial optimization problems allow us to justify theoretically some experimental results obtained for these problems.

For the sake of completeness we should mention here that the statistical physics community developed an equally intensive research interest for finding *theoretical* laws that govern the behavior of macroscopic thermodynamic properties as the free energy. Many interesting models of such large systems were

introduced relatively early, e.g. the *Sherrington-Kirkpatrick (SK) spin glass model* (see Sherrington and Kirkpatrick, 1975). It required, however, considerable time and effort to develop rigorous techniques for solving them. For example, Derrida (1981) introduced a very simple, but exactly solvable model called *random energy model (REM)* as the limit of the SK models family. Later, Aizenman et al. (1987) published an exact solution in the high-temperature phase for the SK model. The general question of the exact free energy behavior became increasingly fascinating: it triggered a new wave of latest research (see Bovier et al., 2002; Talagrand, 2003). The reader should also note that many interesting heuristic tools have been developed in the context of statistical physics over the last several decades, such as the replica method (Parisi, 2009), the cavity method (Mézard and Parisi, 2003) and meanfield approximation schemes with belief propagation algorithms.

1.2. Notation and setting

We consider optimization problems that can be formulated as follows (for explanation see Fig. 1): let n be some integer determining the size of the problem (e.g., number of vertices in a graph, size of a matrix, etc.), and $\mathcal{S}_n$ a finite set of objects (e.g., set of edges, elements of a matrix, etc). Let X denote the input to the problem (data).

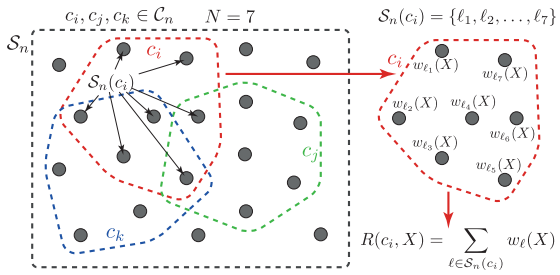


Figure 1: Illustration of the notation: each of the solutions (examples shown in the figure are c_i, c_j, c_k) includes N (in the figure $N = 7$) objects from the underlying set $\mathcal{S}_n$. The cost function of a solution is the sum of weights assigned to the objects, which belong to that solution.

Define $\mathcal{C}_n$ as a finite set of all feasible solutions (e.g. bisections of a graph), and $\mathcal{S}_n(c) \subseteq \mathcal{S}_n$, $c \in \mathcal{C}_n$, as a finite set of objects belonging to the feasible solution c (e.g., set of edges belonging to a bisection). Let $w_i(X) = W_i$, $i \in \mathcal{S}_n$, be the weight assigned to the i-th object. In this paper we consider optimization problems for which the cost function and optimization task are defined as follows:

$$R(c, X) = \sum_{i \in \mathcal{S}_n(c)} w_i(X) \quad \text{and} \quad c_{\text{opt}}(X) = \arg\min_{c \in \mathcal{C}_n} R(c, X). \tag{1}$$

We also denote the cardinality of the feasible set as m (i.e., $m := |\mathcal{C}_n|$) and the cardinality of $\mathcal{S}_n(c)$ as N for all $c \in \mathcal{C}_n$ (i.e., $N := |\mathcal{S}_n(c)|$). In this paper,

we focus on optimization problems in which $\log m = o(N)$ holds true (see Szpankowski, 1995). We call these optimization problems *parameter rich* since the logarithm of the solution space cardinality scales sub-linearly with the number N of objects that belong to a solution c.

1.3. Finding the optimal Gibbs posterior

Most real world combinatorial optimization problems are affected by noise in the input data X. Therefore, they behave like large disordered particle systems, e.g., random networks or spin glasses. Like physical systems, they optimize an application dependent functional (cost function), and their solutions are characterized by the maximum entropy method. In view of this stochastic setting, we suggest to "robustify" the solution by sampling it from some *posterior distribution* $p(c|X)$. In the framework of maximum entropy, it is well justified (see Vannimenus and Mézard (1984)) to use Gibbs distributions, known also as *Gibbs posteriors*, for the posterior distribution $p(c|X)$ leading to a channel presented in Fig. 2.

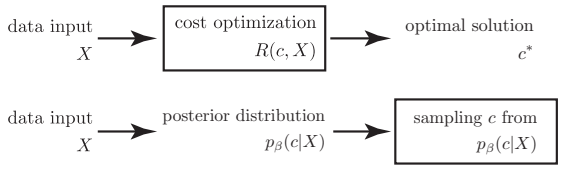


Figure 2: Standart risk minimization (upper) solution and a solution obtained via sampling from approximating posterior distribution (lower).

Definition 1. *Suppose we are given an optimization problem defined by a cost function $R(c, X) \in \mathbb{R}$, where c is a solution from the finite solution space $\mathcal{C}$ and X is a random data instance. Then the* ***Gibbs posterior distribution*** *$p_\beta(c|X)$ is defined as*

$$p_\beta(c|X) = \frac{1}{Z(\beta, X)} \exp(-\beta R(c, X)) \quad \textit{with} \quad Z(\beta, X) = \sum_{c' \in \mathcal{C}} \exp(-\beta R(c', X)) \,. \tag{2}$$

The term $Z(\beta, X)$ is known as the *partition function*. The Gibbs distribution is parameterized by a parameter β which is called the *inverse temperature*. For any β the Gibbs posterior assigns the highest weights to those solutions that have the smallest costs, and β controls the level of concentration of $p_\beta(c|X)$ around minimal solutions (see Fig. 3a).

In passing we remark that we will sometimes omit X as an argument of $Z(\beta, X)$ and $R(c, X)$ for the sake of brevity. Expectation $\mathbb{E}[.]$, variance $\mathrm{Var}[.]$ and other probabilistic operations are meant to be evaluated with respect to the distribution of X, if not explicitly stated otherwise.

Obviously, β somewhat contributes to robustness of $p_\beta(c|X)$. But then the question arises: what is the right way to measure how good a particular choice of β is? To answer that, we investigate what happens to $p_\beta(c|X)$ when the input

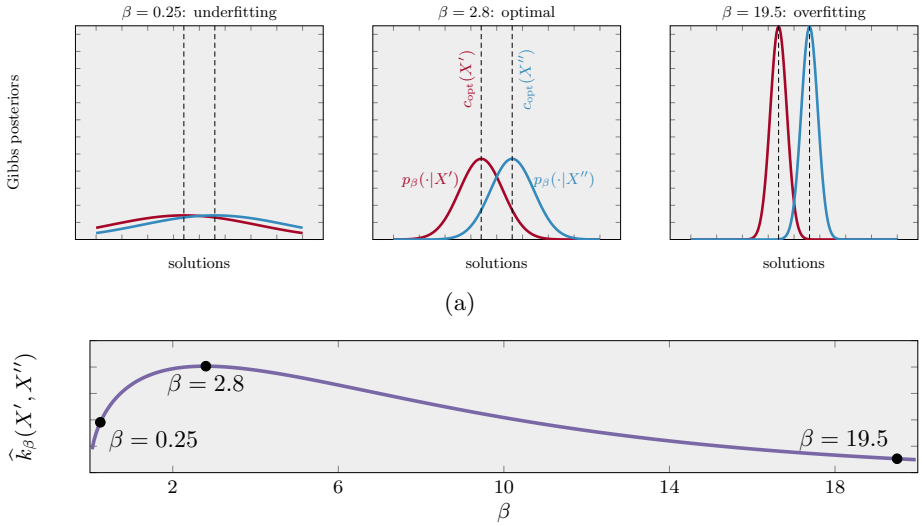


Figure 3: (a): Schematic depiction of two Gibbs posteriors $p_\beta(\cdot|X')$ and $p_\beta(\cdot|X'')$, which may underfit (low β), be optimal (intermediate β) or overfit (high β) depending on the regularizing inverse temperature β. (b): the value of the empirical agreement kernel $\widehat{k}_\beta(X', X'')$ as a function of β, computed for the toy example of the Figure (a). The value $\beta = 2.8$ maximizes this kernel, meaning that the two posteriors are possibly "stable" *and* "informative".

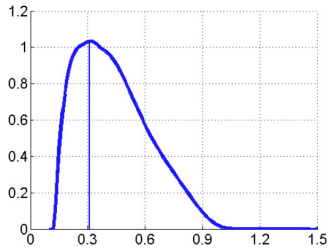


Figure 4: Experimental results for the averaged log-posterior agreement of a clustering problem (Chehreghani et al., 2012).

data fluctuates. Let us assume that *two* noisy instances X' and X'', come from the same source. Intuitively (see Fig. 3a), for values of β that are *very small*, the posteriors $p_\beta(c|X')$ and $p_\beta(c|X'')$ are very similar (we will informally say "stable"), but they do not carry much information due to their large variance (we will informally say "non-informative"). Conversely, using values of β that are *very high* result in very informative posteriors but they are simultaneously very sensible to noise (observe that the best solution to X'' is highly improbable under $p_\beta(\cdot|X')$). One of the ways to balance between these two limits of under- and overfitting is to introduce the posterior agreement kernel for two data instances that show how "close" $p(c|X)$ and $p(c|X')$ are.

A natural measure of agreement between $p_\beta(c|X')$ and $p_\beta(c|X'')$ is defined by the overlap between the two posteriors in the solution space. Buhmann (2010) introduced the *log-posterior agreement kernel* for two instances defined below.

Definition 2. *The* ***posterior agreement kernel*** *for two instances X', X'' is defined as*

$$\begin{aligned}\widehat{k}_\beta(X', X'') &= \sum_{c\in\mathcal{C}} p_\beta(c|X')p_\beta(c|X'') \\ &= \frac{\sum_{c\in\mathcal{C}} \exp(-\beta(R(c, X') + R(c, X')))}{Z(\beta, X')Z(\beta, X'')} \in [0, 1]\,. \end{aligned} \tag{3}$$

Then Buhmann (2010) also introduced a *generalization capacity* that quantitatively measures the maximum of expected log-posterior agreement:

Definition 3. *The* ***generalization capacity*** *I of a cost function $R(c, X)$ is defined as*

$$I := \sup_\beta \mathbb{E}_{X',X''} \log|\mathcal{C}| \sum_{c\in\mathcal{C}} p_\beta(c|X')p_\beta(c|X'') := \mathbb{E}_{X',X''} \log\big(|\mathcal{C}|\ \widehat{k}_\beta(X', X'')\big)\,. \tag{4}$$

The optimal β is thus, according to Buhmann (2010), obtained through maximizing the expected log-posterior agreement. In a toy example presented in Figure 3b the posterior kernel is shown. We see that it has a clear maximum w.r.t. the temperature, and this behavior is usually observed. In fact, this is also confirmed by recent experimental results from (Chehreghani et al., 2012) shown in Figure 4. In this paper, in Theorem 4 we provide theoretical justification for such behavior by considering in details two optimization problems, namely sparse Minimum Bisection Problem (sMBP) and Lawler's Quadratic Assignment Problem (LQAP).

1.4. Computing free energy density

In order to estimate the posterior kernel (3) we need to evaluate $\mathbb{E}\log Z(\beta, X')$, $\mathbb{E}\log Z(\beta, X'')$ as well as $\mathbb{E}\log\sum_{c\in\mathcal{C}}\exp\big(-\beta(R(c, X') + R(c, X''))\big)$, i.e. the expected log-partition functions. For a large n this task represents a computational bottleneck and is known to pose a notoriously difficult mathematical challenge

(see Talagrand (2003)). We address this issue in our paper and provide new solutions and novel lower bounding techniques. More precisely, we compute the *Helmholtz free energy* density defined next.

Definition 4. *The free energy of a set of solutions (configurations) $\mathcal{C}$ is defined as*

$$\mathcal{F}(\beta) = -\mathbb{E}_X[\log Z(\beta, X)]/\log|\mathcal{C}| \,. \tag{5}$$

It is known (Bovier et al., 2002; Talagrand, 2003), that obtaining asymptotic bounds for this quantity is a difficult mathematical problem. In this paper we address it for tghe two special cases sMBP and LQAP.

1.5. Contributions and structure of this paper

The two main contributions of our paper are: (i) mathematically rigorous asymptotic analysis of the free energy for two optimization problems (i.e., sparse MBP and Lawler QAP) in the high-temperature regime. We shall find phase transitions which are equivalent to the discontinuities of REM and high-temperature SK (Derrida, 1981; Aizenman et al., 1987) (ii) rigorous asymptotic computation of the quantity called the expected log-posterior agreement (eLPA) and interpreting the semantics of this quantity (Buhmann, 2010; Chehreghani et al., 2012; Busse et al., 2013).

The paper is organized as follows. Formal definitions and main result theorems are stated in Section 2. In Section 2.1 the problems under consideration are described, while in Sections 2.2 (namely, Theorems 2, 3 and 4) and 2.3 the statements of our contribution are provided and discussed. Proofs of the main results can be found in Sections 3 and 4. In Section 5 we provide simulation details and some new conjectures.

2. Main Results

In this paper we focus on two optimization problems, namely the sparse Minimum Bisection Problem (sMBP) and the Lawler Quadratic Assignment Problem (LQAP). Formal definitions are given below. However, we should add that many of our results hold for a larger class of optimization problems as long as $\log m = o(N)$ (see Szpankowski, 1995). In the rest of the paper we will utilize the temperature rescaling $\beta = \widehat{\beta}\sqrt{\log m/N}$ with $\widehat{\beta} = O(1)$ which together with $\log m = o(N)$ explains $\beta \to 0$ limit. This rescaling was justified in (Buhmann et al., 2014).

For these two problems we shall provide tight asymptotics for the free energy (5), and compute asymptotically the log-posterior agreement as well as $\widehat{\beta}^*$ that maximizes the posterior kernel.

2.1. Minimum bisection and quadratic assignment optimization problems

This section introduces combinatorial optimization problems that will be used to describe our findings. These problems fall into the $\log m = o(N)$ class specified in Sec. 1.2 and cover a wide range of practical applications in signal processing and neural information processing.

Minimum bisection problem (MBP). Consider a complete undirected weighted graph $G = (V, E, X)$ of n vertices, where n is an even number. The input data instance X is represented by (random) weights $(W_i)_{i\in E}$ of the graph edges.

A *bisection* is a balanced partition $c = (U_1, U_2)$ of the vertices in two disjoint sets: $U_1, U_2 \subset V$, $U_1 \sqcup U_2 = V$, $|U_1| = |U_2| = \frac{n}{2}$. Now $\mathcal{S}_n = E$ and $\mathcal{C}_n$ is the set of all bisections of graph G, while $\mathcal{S}_n(c)$ is the set of all edges cut by the bisection c. The cost of a bisection c is the sum of the weights of all cut edges $R(c) = \sum_{i\in\mathcal{S}_n(c)} W_i$.

The minimum bisection problem finds the bisection of the graph with minimum cost. A simple calculation (we omit here $1/2$ constant for the sake of brevity) shows that $|\mathcal{C}_n| = m = \binom{n}{n/2}$ and $|\mathcal{S}_n(c)| = N = \frac{n^2}{4}$, and that

$$\log m = \log \binom{n}{n/2} \sim \log\left(2^n \sqrt{\frac{2}{\pi n}}\right) = n\log 2 - \frac{1}{2}\log n + O(1), \tag{6}$$

which shows that the minimum bisection problem belongs to the class of stochastic optimization problems discussed in this paper (i.e., $\log m = o(N)$).

Sparse minimum bisection problem (Sparse MBP or sMBP). We actually will focus on the *sparse* Minimum Bisection Problem in which the disjoint subsets are of the size $|U_1| = |U_2| \equiv d$ where d grows at least logarithmically and at most linearly which we write as $\log \ll d \ll n$. Thus, $N = d^2$ and the following holds

$$\log m = \log \binom{n}{d}\binom{n-d}{d} = \log \frac{n!}{d!(n-2d)!} \sim 2d\log n. \tag{7}$$

Thus the problem falls into the class $\log m = o(N)$ since we assume $\log n \ll d$.

Quadratic Assignment Problem (QAP). We consider two $n \times n$ real-, positive-valued matrices, namely the weight matrix V and the distance matrix H. The solution space $\mathcal{C}_n$ is the set of the n-element permutations $\mathbf{S}_n$. The cost function is then $R(\pi, V, H) = \sum_{i,j=1}^n V_{ij} \cdot H_{\pi(i),\pi(j)}$ for $\pi \in \mathbf{S}_n$. In our terms, the object space is the set of products of entries of V and H constrained by a relation on the indices: $\mathcal{S}_n = \{V_{ij} \cdot H_{\pi(i),\pi(j)} \mid 1 \le i, j \le n; \pi \in \mathbf{S}_n\}$. In our notation, $N = |\mathcal{S}_n(\pi)| = n^2$ and $m = |\mathcal{C}_n| = n!$ and thus $\log m \sim n\log n = o(N)$ is satisfied.

Lawler Quadratic Assignment Problem (Lawler QAP or LQAP). Lawler (1963) introduced a generalization of the QAP where the distance and weight matrices are replaced by a 4-dimensional matrix Q with i.i.d. values: $R(\pi, Q) = \sum_{i,j=1}^n Q_{i,j,\pi(i),\pi(j)}$ for $\pi \in \mathbf{S}_n$. It is interesting to see that this generalization does not change the combinatorial structure of the problem: a Lawler QAP can be built from a normal QAP and thus falls into our class.

2.2. Free energy, expected log-posterior agreement and its phase transition

In order to give a full picture, before presenting our main results we first derive a tight upper bound on the free energy as discussed in (Buhmann et al.,

2014). Interestingly, it shows that there is a phase transition in the second-order term of the upper bound of the free energy. Such a phase transition is a characteristic feature of various large-scale systems (see Łuczak, 1994; Talagrand, 2003; Mézard and Montanari, 2009).

First, let us state our main assumptions that we use throughout the paper.

Common Theorem Setting. *Consider a class of combinatorial optimization problems in which:*

(A) *the cardinality m of the set of feasible solutions and the size N of every feasible solution are related as $\log m = o(N)$, and we adopt the scaling $\beta = \widehat{\beta}\sqrt{\log m/N}$;*

(B) *weights W_i are identically (not necessarily independently) distributed with mean μ and variance σ^2 and that the moment generating function of the negative centered weights $(-\overline{W}_i)$ is finite, i.e. $\overline{G}(t) \equiv \mathbb{E}[\exp(-t\overline{W}_i)] < \infty$ exists for some $t > 0$;*

(C) *within a given solution c, the weights are mutually independent, i.e. for all $c \in \mathcal{C}_n$, the set $\{W_i \mid i \in \mathcal{S}_n(c)\}$ is a set of mutually independent variables.*

To get a flavour of bounding $\mathbb{E}[\log Z]$ we first observe that $\mathbb{E}[\log Z] \leq \log \mathbb{E}[Z]$ (by Jensen's inequality). We can evaluate $\mathbb{E}[Z]$ as follows:

$$\begin{aligned}\mathbb{E}[Z] &= \mathbb{E}\Big[\sum_{c\in\mathcal{C}} \exp(-\beta R(c))\Big] = \exp(-\beta N\mu)\mathbb{E}\Big[\sum_{c\in\mathcal{C}} \exp\big(-\beta(R(c) - N\mu)\big)\Big] \\ &= \exp(-\beta N\mu) m \overline{G}^N(\beta). \end{aligned} \tag{8}$$

Thus

$$\log \mathbb{E}[Z] = -\beta N\mu + \log m + N \log \overline{G}(\beta) \tag{9}$$

since the set of random variables W_i belonging to the same solution are mutually independent variables. Throughout we write $\overline{R}(c) = R(c) - \mathbb{E}[R] = R(c) - N\mu$ for the centralized cost, and $\overline{G}(\beta)$ for the moment generating function of the centralized weight $\overline{W}_i = W_i - \mu$.

We can expand $\overline{G}(\beta)$ into the Taylor series around zero and obtain

$$\overline{G}(\beta) = 1 + \frac{1}{2}\beta^2\sigma^2 + O(\beta^3). \tag{10}$$

We find as long as $\beta \to 0$

$$\begin{aligned}\log \mathbb{E}[Z] &= -\beta N\mu + \log m + N \log \overline{G}(\beta) \\ &= -\beta N\mu + \log m + N \log\Big(1 + \frac{1}{2}\beta^2\sigma^2 + O(\beta^3)\Big) \\ &= -\beta N\mu + \log m + \frac{1}{2}N\beta^2\sigma^2(1 + O(\beta)). \end{aligned} \tag{11}$$

Now we apply the rescaling from the Common Theorem Setting

$$\beta = \widehat{\beta}\sqrt{\frac{\log m}{N}} \tag{12}$$

for some constant $\widehat{\beta}$ leading to

$$\frac{\log \mathbb{E}[Z] + \beta N\mu}{\log m} = 1 + \frac{1}{2}\widehat{\beta}^2\sigma^2(1 + O(\beta)). \tag{13}$$

In terms of $\mathbb{E}[\log Z]$ we find

$$\frac{\mathbb{E}[\log Z] + \widehat{\beta}\mu\sqrt{N \log m}}{\log m} \leq 1 + \frac{1}{2}\widehat{\beta}^2\sigma^2\left(1 + O\left(\sqrt{\frac{\log m}{N}}\right)\right). \tag{14}$$

But there is a surprise! Let us denote

$$\phi(\beta) = \mathbb{E}[\log Z] + \beta N\mu =: \mathbb{E}[\log \widehat{Z}(\beta)] \tag{15}$$

where $\widehat{Z}(\beta) = \sum_{c\in\mathcal{C}} \exp(\beta \overline{R}(c))$ with $\overline{R}(c) = -\sum_{i\in\mathcal{S}(c)} \overline{W}_i$. It is easy to observe that

$$\beta \max_{c\in\mathcal{C}} \overline{R}(c) \leq \log \widehat{Z}(\beta). \tag{16}$$

Using the upper bound obtained in (14) we find

$$\frac{\mathbb{E}[\max_{c\in\mathcal{C}} \overline{R}(c)]}{\log m} \leq \sqrt{\frac{N}{\log m}}\left(\widehat{\beta}^{-1} + \frac{1}{2}\widehat{\beta}\sigma^2\right). \tag{17}$$

Choosing $\widehat{\beta}^* = \sqrt{2}/\sigma$ that minimizes the right-hand side of (17) we arrive at

$$\mathbb{E}[\max_{c\in\mathcal{C}} \overline{R}(c)] \leq \sqrt{2\sigma^2 N \log m} \tag{18}$$

Now proceeding as in Talagrand (2003, Proposition 1.1.3) we obtain

$$\phi'(\beta) \leq \mathbb{E}[\max_{c\in\mathcal{C}} \overline{R}(c)]. \tag{19}$$

But for $\beta > \beta^* := \widehat{\beta}^*\sqrt{\log m/N}$,

$$\phi(\beta) \leq \phi(\beta^*) + \phi'(\beta^*)(\beta - \beta^*), \tag{20}$$

since $\phi(\beta)$ is known to be convex. Applying the upper bound for $\phi'(\beta)$ yields

$$\mathbb{E}[\log \widehat{Z}] \leq \widehat{\beta}\sigma\sqrt{2}\log m \tag{21}$$

and the upper bound for the second $\widehat{\beta}$ region is obtained. Observe that in this region the growth is linear with respect to $\widehat{\beta}$.

In summary, we have the following upper bounds.

Theorem 1 (Buhmann et al. (2014))**.** *Under the common setting of the current section the following holds:*

$$\lim_{n\to\infty} \frac{\mathbb{E}[\log Z(\beta, X)] + \widehat{\beta}\mu\sqrt{N \log m}}{\log m} \leq \begin{cases} 1 + \frac{\widehat{\beta}^2\sigma^2}{2}, & \widehat{\beta} < \frac{\sqrt{2}}{\sigma}, \\ \widehat{\beta}\sigma\sqrt{2}, & \widehat{\beta} \geq \frac{\sqrt{2}}{\sigma}. \end{cases} \tag{22}$$

Remark. The general upper bound proven above is unfortunately not tight. Consider the (non-sparse) minimum bisection problem with $d = n/2$. Under

the same general assumptions for the weights, it can be shown that a tighter bound holds for $\widehat{\beta} \leq \frac{1}{\sqrt{\log 2}\sigma}$

$$\lim_{n\to\infty} \frac{\mathbb{E}[\log Z(\beta, X)] + \widehat{\beta}\mu\sqrt{N \log m}}{\log m} \leq 1 + \frac{\widehat{\beta}^2\sigma^2}{4}. \tag{23}$$

We prove (23) in Appendix A. □

We proceed now to state our results. For some combinatorial optimization problems, the asymptotical upper bound of Theorem 1 turns out to be tight. Below we present two main results which give the asymptotically matching lower bounds for the Sparse MBP and Lawler QAP. For the Sparse MBP we develop a novel approach of proving it since the techniques proposed by Talagrand (2003, Chapter 1) seem not to work.

Theorem 2. *Consider Sparse MBP complying with Common Theorem Setting whose edge weights have mean μ and variance σ^2. Then the following holds:*

$$\lim_{n\to\infty} \frac{\mathbb{E}[\log Z(\beta, X)] + \widehat{\beta}\mu\sqrt{N \log m}}{\log m} = \begin{cases} 1 + \frac{\widehat{\beta}^2\sigma^2}{2}, & \widehat{\beta} < \frac{\sqrt{2}}{\sigma}, \\ \widehat{\beta}\sigma\sqrt{2}, & \widehat{\beta} \geq \frac{\sqrt{2}}{\sigma} \end{cases} \tag{24}$$

provided $\log \ll d \ll n^{2/7}/\sqrt{\log n}$.

Let us now consider the Lawer QAP. In this case, we apply a slightly modified approach developed in Talagrand. However, we should point out that LQAP has some dependency that were not present in Derrida's model for which Talagrand proposed his method.

Theorem 3. *Consider Lawler QAP complying with Common Theorem Setting, whose matrix entries have mean μ and variance σ^2. Then the following holds:*

$$\lim_{n\to\infty} \frac{\mathbb{E}[\log Z(\beta, X)] + \widehat{\beta}\mu\sqrt{N \log m}}{\log m} = \begin{cases} 1 + \frac{\widehat{\beta}^2\sigma^2}{2}, & \widehat{\beta} < \frac{\sqrt{2}}{\sigma}, \\ \widehat{\beta}\sigma\sqrt{2}, & \widehat{\beta} \geq \frac{\sqrt{2}}{\sigma}. \end{cases} \tag{25}$$

The matching lower bounds for sMBP and LQAP given above in Theorems 2 and 3 allow us to present theoretical justification for the behavior of the posterior agreement kernel as shown in Figure 4. To see that, we observe that

$$\mathbb{E}_{X',X''} \log \sum_{c\in\mathcal{C}} p_\beta(c|X')p_\beta(c|X'') \tag{26}$$

$$= \mathbb{E}_{X',X''} \log \sum_{c\in C} \frac{\exp(-\beta(R(c, X') + R(c, X'')))}{Z(\beta, X')Z(\beta, X'')}$$

$$= \mathbb{E}_{X',X''} \log Z(\beta, X', X'') - \mathbb{E}_{X'} \log Z(\beta, X') - \mathbb{E}_{X''} \log Z(\beta, X''), \tag{27}$$

where we $Z(\beta, X', X'')$ can naturally be defined as a "partition function".

$$Z(\beta, X', X'') := \sum_{c\in C} \exp(-\beta(R(c, X') + R(c, X''))). \tag{28}$$

Eventually this allows us to use the above theorems to compute all the three terms of (26).

To make the final step, we first need to formalize how exactly X' and X'' are obtained: let us assume that the two instances X' and X'' are both represented by two sets of weights $X' = \{W_i'\}$ and $X'' = \{W_i''\}$ through adding two "noise" instances $\delta X' = \{\delta W_i'\}$ and $\delta X'' = \{\delta W_i''\}$ to the same "signal" instance $X = \{W_i\}$ all the mentioned sets being of the same size $|\mathcal{S}|$:

$$W_i' = W_i + \delta W_i', \quad W_i'' = W_i + \delta W_i'' \quad \text{for} \quad i \in \mathcal{S}. \tag{29}$$

We also require that the signal and noise weights have certain means and variances:

$$\mathbb{E}[W_i] = \mu \qquad \mathrm{Var}[W_i] = \sigma^2 \tag{30}$$

$$\mathbb{E}[\delta W_i'] = \mathbb{E}[\delta W_i''] = 0 \qquad \mathrm{Var}[\delta W_i'] = \mathrm{Var}[\delta W_i''] = \tilde{\sigma}^2. \tag{31}$$

We define the noise-to-signal ratio as $\gamma = \tilde{\sigma}/\sigma$.

Applying Theorems 2 and 3 we are led to the following result for the posterior agreement kernel.

Theorem 4. *Consider Sparse MBP or Lawler QAP complying with the Common Theorem Setting. Let set X be "signal" weights with mean μ and variance σ^2 and two sets $\delta X'$, $\delta X''$ be "noise" with mean 0, and variance $\tilde{\sigma}^2$, all the sets of the same size. Let $X' = X + \delta X'$ and $X'' = X + \delta X''$ (elementwise sum) be the two problem instances. Let $\gamma := \tilde{\sigma}/\sigma$ be noise-to-signal ratio. Then the expectation of the log-posterior agreement (3) satisfies*

$$\lim_{n\to\infty} \frac{\mathbb{E}_{X,\delta X',\delta X''} \log\bigl(|\mathcal{C}|\, \widehat{k}_\beta(X', X'')\bigr)}{\log m} = \eta(\widehat{\beta}), \tag{32}$$

where

$$\eta(\widehat{\beta}) = \begin{cases} (\widehat{\beta}\sigma)^2, & \widehat{\beta}\sigma < \frac{\sqrt{2}}{\sqrt{4+2\gamma^2}} \\ \widehat{\beta}\sigma\sqrt{2}\sqrt{4+2\gamma^2} - (\widehat{\beta}\sigma)^2(1+\gamma^2) - 1, & \frac{\sqrt{2}}{\sqrt{4+2\gamma^2}} \le \widehat{\beta}\sigma < \frac{\sqrt{2}}{\sqrt{1+\gamma^2}} \\ \widehat{\beta}\sigma\sqrt{2}\bigl(\sqrt{4+2\gamma^2} - 2\sqrt{1+\gamma^2}\bigr) + 1, & \frac{\sqrt{2}}{\sqrt{1+\gamma^2}} \le \widehat{\beta}\sigma \end{cases} \tag{33}$$

In particular, the expected log-posterior agreement is maximized at the eLPA-optimal inverse temperature:

$$\widehat{\beta}^* \equiv \widehat{\beta}^*_{\mathrm{eLPA}} = \frac{\sqrt{2+\gamma^2}}{\sigma(1+\gamma^2)}. \tag{34}$$

2.3. Discussion of the results

First, the reader should notice that the free energy of Sparse MBP (24) and of Lawler QAP (25) exhibit a phase transition similar to that of Derrida's *Random Energy Model (REM)* (Derrida, 1981, Section V). We like to emphasize that Sparse MBP and Lawler QAP *introduce some correlation* between costs of

pairs of solutions, while REM defines a technically much simpler setting without any correlations between cost values.

Second, regarding the behavior of expected log-posterior agreement (4) shown in Theorem 4, we can notice that it shows two phase transitions with quadratic, mixed and linear phases, which corresponds to three phases of the Generalized REM, well explained in (Derrida and Gardner, 1986, Section 3). Its behavior is visualized in Fig. 5. We make the following brief observations, explaining the combinatorial meaning for an approximate learning process:

— The normalized eLPA in (33) depends on the temperature by the product $\widehat{\beta}\sigma$, which pronounces the fact that the reference scale for the temperature of Gibbs posteriors is adjusted by the amount of signal in data.

— The noise-to-signal ratio γ plays the crucial role. For a fixed signal σ, the optimal temperature $\widehat{\beta}^*$ grows to $\sqrt{2}/\sigma$, as the noise-to-signal vanishes ($\gamma \to 0$, i.e. $\tilde{\sigma} \ll \sigma$). This behavior supports our intuition that a posterior adapted to the signal variance is better to choose in the absence of noise. Optimal $\widehat{\beta}^*$ is located in the so called *retrieval phase*.

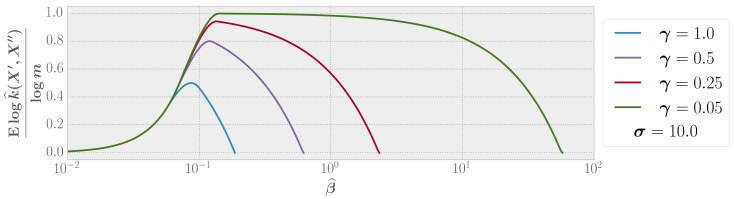


Figure 5: Behavior of the expected log-posterior agreement (4).

— The *high temperature phase* $\widehat{\beta} \to 0$ results in low eLPA meaning that informative solutions cannot be found by sampling from Gibbs posteriors (which are too "broad") in this phase.

— *Freezing phase* $\widehat{\beta} \to \infty$: the decreasing expected low-posterior agreement reflects the instability of local minima under perturbations $\delta X'$, $\delta X''$. Solutions do not generalize and a learning algorithm cannot extract information from dataset X' and test it on X''.

Third, we compare the log-posterior agreement behavior to the previous experimental evidence of Chehreghani et al. (2012, Fig. 2), which shows the same shape of log-posterior agreement. Although the referred paper has an experimental nature and considers another optimization problem, it proves the concept in a nutshell, thereby showing that the theorems presented in this paper support to the approach pioneered in (Buhmann, 2010; Busse et al., 2013).

Forth, Theorems 2 and 3 allow to directly optimize the expected Gibbs risk

$$\mathbb{E}_{p_\beta(c|X),X}[R(c,X)] = -\frac{\partial}{\partial\beta}\mathbb{E}_X \log Z(\beta, X), \tag{35}$$

by means of applying differentiation to the results of these theorems on the right-hand side. As a simple corollary, we thus obtain the following theorem:

Theorem 5 (Minimizing expected Gibbs risk)**.** *The expected Gibbs risk* (35) *is*

minimized at the GR-optimal inverse temperature:

$$\widehat{\beta}^*_{\mathrm{GR}} := \arg\min_{\widehat{\beta}} \mathbb{E}_{p_\beta(c|X),X}[R(c,X)] = \frac{\sqrt{2+2\gamma^2}}{\sigma(1+\gamma^2)}. \tag{36}$$

It is interesting to compare GR-optimal (36) and eLPA-optimal (34) inverse temperatures:

$$\widehat{\beta}^*_{\mathrm{eLPA}} = \frac{\sqrt{2+\gamma^2}}{\sigma(1+\gamma^2)} \quad \text{and} \quad \widehat{\beta}^*_{\mathrm{GR}} = \frac{\sqrt{2+2\gamma^2}}{\sigma(1+\gamma^2)} \tag{37}$$

and not that they have a slight difference: eLPA selects slightly less (by a factor of $\sqrt{1+\frac{\gamma^2}{2+\gamma^2}}$) inverse temperature. This can be interpreted as follows: eLPA approach tends to be a bit more conservative, selecting slightly "broader" Gibbs posterior as opposed to expected Gibbs risk minimization.

3. Proof of Theorem 2: Matching Lower Bound for Sparse MBP

In this section we present a proof of the matching lower bound for Sparse MBP. The proof technique that we propose here is novel to the best of our knowledge and was also used in (see Magner et al., 2015, 2016).

The proof is broken into several lemmas. Let us start with defining D as elementwise overlap between two solutions (i.e. number of shared *edges*) sampled *uniformly at random*. We will refer to this uniform distribution as $\mathcal{D}$.

Lemma 6. *The following holds*

$$\frac{\#\{\textit{vertex-non-overlapping}\}}{m^2} = 1 - \Theta(d^2/n). \tag{38}$$

Proof. Observe that

$$\frac{\#\{\text{vertex-non-overlapping}\}}{m^2} = \frac{\binom{n}{d}\binom{n-d}{d}\binom{n-2d}{d}\binom{n-3d}{d}}{\binom{n}{d}^2\binom{n-d}{d}^2} = \frac{\binom{n-2d}{d}\binom{n-3d}{d}}{\binom{n}{d}\binom{n-d}{d}}. \tag{39}$$

We now use Stirling's approximation, for any integer c to find

$$\binom{n-cd}{d} \le \frac{(n-cd)^d}{d!} = \frac{n^d(1-cd/n)^d}{d!} \sim \frac{n^d(1-cd^2/n)}{d!}. \tag{40}$$

Similarly,

$$\binom{n-cd}{d} \ge \frac{(n-(c+1)d)^d}{d!} = \frac{n^d(1-(c+1)d/n)^d}{d!} \sim \frac{n^d(1-(c+1)d^2/n)}{d!}. \tag{41}$$

Applying these bounds we find

$$\frac{\#\{\text{vertex-non-overlapping}\}}{m^2} \le \frac{(1-2d^2/n)(1-3d^2/n)}{(1-d^2/n)(1-2d^2/n)} \sim 1-2d^2/n \tag{42}$$

and

$$\frac{\#\{\text{vertex-non-overlapping}\}}{m^2} \geq \frac{(1-3d^2/n)(1-4d^2/n)}{(1-d^2/n)} \sim 1-6d^2/n. \tag{43}$$

This completes the proof. □

Lemma 7. *The following holds:*

$$\mathbb{P}_{\mathcal{D}}(D=0) \sim \frac{\#\{\textit{vertex-non-overlapping}\}}{m^2} \tag{44}$$

Proof.

$$\mathbb{P}_{\mathcal{D}}(D=0) = \frac{\#\{\text{vertex-non-overlapping}\}}{m^2} + \frac{\#\{\text{edge-non-overlapping} \mid \text{vertex-overlapping}\}}{m^2}. \tag{45}$$

Since the following inclusion holds:

$$\{\text{edge-non-overlapping}|\text{vertex-overlapping}\} \subseteq \{\text{vertex-overlapping}\}, \tag{46}$$

we can conclude that

$$\frac{\#\{\text{edge-non-overlapping} \mid \text{vertex-overlapping}\}}{m^2} \leq \frac{\{\text{vertex-overlapping}\}}{m^2} \tag{47}$$

$$= \frac{m^2 - \{\text{vertex-non-overlapping}\}}{m^2} = 1 - 1 + \Theta(d^2/n) = o(1), \tag{48}$$

where the last equation comes from Lemma 6 and $d = o(n)$. Hence, the the following holds:

$$\frac{\mathbb{P}_{\mathcal{D}}(D=0)}{\#\{\text{vertex-non-overlapping}\}/m^2} = 1 + \frac{\#\{\text{edge-non-overlapping} \mid \text{vertex-overlapping}\}/m^2}{\#\{\text{vertex-non-overlapping}\}/m^2} \tag{49}$$

$$= 1 + \frac{o(1)}{1+o(1)} = 1 + o(1), \tag{50}$$

which proves the lemma. □

These two lemmas allow us to estimate the expected value of D.

Lemma 8. *The following holds:*

$$\mathbb{E}_{\mathcal{D}} D = \mathcal{O}(d^4/n). \tag{51}$$

Proof. To compute $\mathbb{E}_{\mathcal{D}} D$, observe

$$\mathbb{E}_D D = 0 \cdot \mathbb{P}_{\mathcal{D}}(D=0) + \sum_{k=1}^{N} k \cdot \mathbb{P}_{\mathcal{D}}(D=k) \leq N \sum_{k=1}^{N} \mathbb{P}_{\mathcal{D}}(D=k) \tag{52}$$

$$= d^2 \cdot \mathbb{P}_{\mathcal{D}}(D \neq 0) = d^2\big(1 - \mathbb{P}_{\mathcal{D}}(D=0)\big) \sim \Theta(d^4/n), \tag{53}$$

where the last asymptotic equivalence follows from Lemmas 6 and 7. The less-than-equal sign turns Θ into $\mathcal{O}$. The lemma is proven. $\square$

Now we are in the position to prove Theorem 2. Let us now introduce an event A for some ϵ we choose later:

$$A := \{Z \geq \epsilon \mathbb{E}Z\}. \tag{54}$$

This implies, by Chebychev inequality,

$$1 - \mathbb{P}(A) \leq \mathbb{P}\big(|Z - \mathbb{E}Z| \geq (1-\epsilon)\mathbb{E}Z\big) \leq \frac{\mathrm{Var}Z}{(1-\epsilon)^2(\mathbb{E}Z)^2}. \tag{55}$$

In Lemma 13 of Appendix B we prove the following (see also (Buhmann et al., 2014)):

$$\mathrm{Var}Z = (\mathbb{E}Z)^2\Big(\mathbb{E}_{\mathcal{D}}\Big(\frac{G(2\beta)}{G^2(\beta)}\Big)^D - 1\Big). \tag{56}$$

Expanding $G(2\beta)$ and $G^2(\beta)$ in Taylor's series, we find

$$\mathrm{Var}Z \sim (\mathbb{E}Z)^2\big(\sigma^2\beta^2\mathbb{E}_{\mathcal{D}}D\big). \tag{57}$$

Thus (55) can be further rewritten as

$$1 - \mathbb{P}(A) \leq \frac{\mathrm{Var}Z}{(1-\epsilon)^2(\mathbb{E}Z)^2} \sim \frac{\sigma^2\beta^2\mathbb{E}_{\mathcal{D}}D}{(1-\epsilon)^2} = \mathcal{O}\Big(\frac{\beta^2\mathbb{E}_{\mathcal{D}}D}{(1-\epsilon)^2}\Big) \tag{58}$$

$$= \mathcal{O}\Big(\frac{d^4\log m}{n(1-\epsilon)^2 N}\Big) = \mathcal{O}\Big(\frac{d^3\log n}{n(1-\epsilon)^2}\Big), \tag{59}$$

where we used Lemma 8 for $\mathbb{E}_{\mathcal{D}}D$ asymptotics.

We now proceed to compute $\mathbb{E}\log Z$ along the way of (Magner et al., 2015):

$$\mathbb{E}\log Z = \mathbb{E}[\log Z \mid A]\cdot\mathbb{P}(A) + \mathbb{E}[\log Z\mathbb{1}(\bar{A})] \tag{60}$$

$$\geq (\log\mathbb{E}Z + \log\epsilon)\mathbb{P}(A) + \mathbb{E}[\log Z\mathbb{1}(\bar{A})]. \tag{61}$$

But by (11) we find

$$\log\mathbb{E}Z = -\beta N\mu + \log m + \frac{1}{2}N\beta^2\sigma^2 + o(\beta^2). \tag{62}$$

Let the above expression be denoted as $L(\beta, N, m, \sigma)$ for the sake of brevity. So, using (59), we rewrite (61):

$$\mathbb{E}\log Z \geq \big(L(\beta,N,m,\sigma) + \log\epsilon\big)\cdot\left(1 - \mathcal{O}\Big(\frac{d^3\log n}{n(1-\epsilon)^2}\Big)\right) + \mathbb{E}[\log Z\mathbb{1}(\bar{A})] \tag{63}$$

$$= L(\beta,N,m,\sigma) + \log\epsilon - \big(L(\beta,N,m,\sigma) + \log\epsilon\big)\cdot\mathcal{O}\Big(\frac{d^3\log n}{n(1-\epsilon)^2}\Big) \tag{64}$$

$$+ \mathbb{E}[\log Z\mathbb{1}(\bar{A})].$$

Thus,

$$\frac{\mathbb{E}\log Z+\beta N\mu}{\log m}\geq 1+\frac{\widehat{\beta}^2\sigma^2}{2}+\frac{\log\epsilon}{\log m}-\big(L(\beta,N,m,\sigma)+\log\epsilon\big)\cdot\mathcal{O}\bigg(\frac{d^3\log n}{n(1-\epsilon)^2}\bigg) \tag{65}$$
$$+\,\mathbb{E}[\log Z\mathbb{1}(\bar{A})].$$

Now we introduce below assumption (66) proved later to be true: assume that

$$\frac{d^3\log n}{n(1-\epsilon)^2}\to 0\quad(n\to\infty). \tag{66}$$

Having that we notice

$$\big(L(\beta,N,m,\sigma)+\log\epsilon\big)\cdot\mathcal{O}\bigg(\frac{d^3\log n}{n(1-\epsilon)^2}\bigg)=o(1), \tag{67}$$

i.e. it is small and thus is further neglected. So we can rewrite

$$\frac{\mathbb{E}\log Z+\beta N\mu}{\log m}\gtrsim 1+\frac{\widehat{\beta}^2\sigma^2}{2}+\frac{\log\epsilon}{\log m}+\frac{\mathbb{E}[\log Z\mathbb{1}(\bar{A})]}{\log m}. \tag{68}$$

We now estimate the term $\mathbb{E}[\log Z\mathbb{1}(\bar{A})]$. For some solution c,

$$\mathbb{E}[\log Z\mathbb{1}(\bar{A})]\geq\mathbb{E}[\log e^{-\beta R(c)}\mathbb{1}(\bar{A})]=\mathbb{E}[-\beta R(c)\cdot\mathbb{1}(\bar{A})] \tag{69}$$
$$=\mathbb{E}[-\beta(\bar{R}(c)+\mathbb{E}R)\cdot\mathbb{1}(\bar{A})]=\mathbb{E}[-\beta\bar{R}(c)\mathbb{1}(\bar{A})]-\beta\mathbb{E}R\cdot\mathbb{P}(\bar{A}) \tag{70}$$
$$\geq-\beta\mathbb{E}[|\bar{R}(c)|]-\beta\mathcal{O}(N)(1-\mathbb{P}(A)) \tag{71}$$
$$\geq-\beta\mathcal{O}(\sqrt{N})-\beta\mathcal{O}\bigg(N\frac{d^3\log n}{n(1-\epsilon)^2}\bigg). \tag{72}$$

Thus, essentially,

$$\frac{\mathbb{E}[\log Z\mathbb{1}(\bar{A})]}{\log m}\geq-\frac{\beta}{\log m}\mathcal{O}\bigg(N\frac{d^3\log n}{n(1-\epsilon)^2}\bigg)\sim-\mathcal{O}\bigg(\frac{d^{7/2}\sqrt{\log n}}{n(1-\epsilon)^2}\bigg)=o(1) \tag{73}$$

provided $d=o(n^{2/7}/\sqrt{\log n})$ which we also write as $d\ll n^{2/7}/\sqrt{\log n}$. Consequently, (68) becomes

$$\frac{\mathbb{E}\log Z+\beta N\mu}{\log m}\gtrsim 1+\frac{\widehat{\beta}^2\sigma^2}{2}+\frac{\log\epsilon}{\log m}-\mathcal{O}\bigg(\frac{d^{7/2}\sqrt{\log n}}{n(1-\epsilon)^2}\bigg). \tag{74}$$

We will now choose ϵ in order to produce the lower bounds, and then check that assumption (66) is satisfied. For $\widehat{\beta}>\widehat{\beta}^*:=\sqrt{2}/\sigma$ we choose

$$\epsilon=m^{-(1-\widehat{\beta}\sigma\sqrt{2}+\frac{\widehat{\beta}^2\sigma^2}{2})}.$$

This gives

$$\frac{\mathbb{E}\log Z+\beta N\mu}{\log m}\gtrsim\widehat{\beta}\sigma\sqrt{2}-o(1), \tag{75}$$

since for this choice $\epsilon = o(1)$, yielding

$$\mathcal{O}\left(\frac{d^{7/2}\sqrt{\log n}}{n(1-\epsilon)^2}\right) = o(1) \tag{76}$$

by our choice of d. For this choice of ϵ the assumption (66) holds.

For $\widehat{\beta} \le \widehat{\beta}^* := \sqrt{2}/\sigma$ we choose $\epsilon = 1/2$, yielding

$$\frac{\mathbb{E}\log Z + \beta N\mu}{\log m} \gtrsim 1 + \frac{\widehat{\beta}^2\sigma^2}{2} + o(1), \tag{77}$$

since

$$\frac{\log \epsilon}{\log m} = o(1), \quad \mathcal{O}\left(\frac{d^{7/2}\sqrt{\log n}}{n(1-\epsilon)^2}\right) = o(1) \tag{78}$$

and the assumption (66) holds. This completes the proof of Theorem 2. □

4. Proof of Theorem 3: Matching Lower Bound for Lawler QAP

In this section we present the proof of Theorem 3 for the matching lower bound for the Lawler QAP.

Theorem 1 gives us a general upper bound. To find the matching lower bound we follow Talagrand (2003) that we briefly review. We should point out up front that Talagrand's technique was designed for proving the matching lower bound for the Random Energy Model (REM) without any dependency. In our case, there are clear dependency between solutions, however, not strong enough to destroy the the essence of our argument, thought the details are much more involved as discussed below.

To start, as in Talagrand, we define Y to be the cardinality of the solution subset for which the centered negative cost function $\overline{R}(c)$ is large enough, that is,

$$Y := \text{card}\{c\colon \overline{R}(c) \ge u_n(\widehat{\beta})\}, \tag{79}$$

where we set in our case

$$u_n(\widehat{\beta}) = \begin{cases} \widehat{\beta}\sigma^2\sqrt{N\log m}, & \widehat{\beta} < \widehat{\beta}^* \\ \widehat{\beta}^*\sigma^2\sqrt{N\log m}, & \widehat{\beta} \ge \widehat{\beta}^*. \end{cases} \tag{80}$$

We also define

$$a_n = \mathbb{P}(\overline{R}(c) \ge u_n(\widehat{\beta})), \tag{81}$$

and the event A as

$$A = \{Y \le ma_n/2\}.$$

Observe that

$$\mathbb{E}[Y] = ma_n \tag{82}$$

and by Markov inequality

$$\mathbb{P}(A) \le \mathbb{P}\big((Y - \mathbb{E}[Y])^2 \ge m^2a_n^2/4\big) \le \frac{4\text{Var}[Y]}{m^2a_n^2} \le \frac{4\mathbb{E}[Y^2]}{m^2a_n^2} - 1. \tag{83}$$

To follow Talagrand's approach, w need to show that $\mathbb{E}[Y^2]/(ma_n)^2 \to 1$ so that $\mathbb{P}(A) \to 0$ which we formulate as the following lemma proved at the end of this section.

Lemma 9. *There exits $\varepsilon > 0$ such that*

$$\mathbb{P}(A) = O(n^{-\varepsilon}) \tag{84}$$

for large n.

In order to establish it and present a concise proof of our matching bound, we need one more technical result regarding large deviations of $\mathbb{P}(\overline{R}(c) \geq u_n(\widehat{\beta}))$ formulated next.

Lemma 10. *Define $\widehat{\beta}^* < \sqrt{2}/\sigma$. The following holds*

$$\log \mathbb{P}(\overline{R}(c) \geq u_n(\widehat{\beta})) = -\frac{u_n(\widehat{\beta})^2}{2N\sigma^2} + o(\log m) \tag{85}$$

where $u_n(\widehat{\beta})$ is defined in (80) above.

Proof We will demonstrate the result only for $\widehat{\beta} < \widehat{\beta}^*$. The proof is identical for the right region of $\widehat{\beta}$. The main tool of this demonstration is the Gärtner-Ellis large-deviation theorem, as best described in (Dembo and Zeitouni, 2009). The following developments introduce the quantities at play in this theorem.

First observe that

$$\mathbb{P}(\overline{R}(c) \geq u_n(\widehat{\beta})) = \mathbb{P}(\frac{1}{\sqrt{N \log m}} \overline{R}(c) \geq \widehat{\beta}\sigma^2).$$

Define the logarithmic generating function of $\overline{R}(c)/\sqrt{N \log m}$:

$$\Lambda_n(\lambda) := \log \mathbb{E}\mathrm{e}^{\lambda \frac{1}{\sqrt{N \log m}} \overline{R}(c)} = N \log \widehat{G}\left(\frac{\lambda}{\sqrt{N \log m}}\right), \tag{86}$$

where $\widehat{G}(\lambda)$ is the moment generating function of a negative centered weight $(-\overline{W})$. The last transition is valid because we assume that the weights within a solution are independent as expressed by assumption (C). Define the limiting moment generating function as

$$\Lambda(\lambda) := \lim_{n\to\infty} \frac{1}{\log m} \Lambda_n(\lambda \log m) \tag{87}$$

$$= \lim_{n\to\infty} \frac{N}{\log m} \log \widehat{G}\left(\lambda \sqrt{\frac{\log m}{N}}\right) \tag{88}$$

$$= \frac{\lambda^2 \sigma^2}{2} \tag{89}$$

because $\widehat{G}(\lambda) = 1 + \frac{\lambda^2\sigma^2}{2} + o(\lambda^2)$ for $\lambda \to 0$.

The Gärtner-Ellis theorem (Dembo and Zeitouni, 2009, Theorem 2.3.6) yields

$$\lim_{n\to\infty} \frac{1}{\log m} \log \mathbb{P}(\overline{R}(c) \geq u_n(\widehat{\beta})) = -\Lambda^*(\widehat{\beta}\sigma^2) \tag{90}$$

where $\Lambda^*(x) = \sup_{\lambda>0} \lambda \cdot x - \Lambda(\lambda) = x^2/\left(2\sigma^2\right)$ is the Fenchel-Legendre transform of $\Lambda(\lambda)$. Hence,

$$\log \mathbb{P}(\overline{R}(c) \geq u_n(\widehat{\beta})) = -\frac{\widehat{\beta}^2\sigma^2}{2}\log m + o(\log m) \tag{91}$$

$$= -\frac{u_n(\widehat{\beta})^2}{N\sigma^2} + o(\log m) \tag{92}$$

The proof for $\widehat{\beta} \geq \widehat{\beta}^*$ is similar in all aspects. □

Now, we are in the position to prove Theorem 3 granted Lemma 9 that we prove at the end of this section. To accomplish it, we need a lower bound for $\mathbb{E}[\log \widehat{Z}(\beta)]$. We consider two cases conditioning on event A defined above and its complementary event $\Omega \backslash A$.

We have on event $\Omega \backslash A$

$$\widehat{Z}(\beta) = \sum_{c \in \mathcal{C}_n} \exp(\beta \overline{R}(c)) \geq \sum_{c:\overline{R}(c) \geq u_n(\widehat{\beta})} \exp(\beta u_n(\widehat{\beta})) \tag{93}$$

$$\geq Y \exp(\beta u_n(\widehat{\beta})) \tag{94}$$

$$\geq \frac{ma_n}{2} \exp(u_n(\widehat{\beta})). \tag{95}$$

Therefore,

$$\mathbb{E}[\mathbb{1}_{\Omega \backslash A} \log \widehat{Z}(\beta)] \geq (1 - \mathbb{P}(A)) \left(\log m + \log a_n - \log 2 + \beta u_n(\widehat{\beta})\right). \tag{96}$$

On event A, we derive the lower bound in the following way. Choosing an arbitrary solution c_0, we notice that $\widehat{Z}(\beta) \geq \exp(\beta \overline{R}(c_0))$ and thus

$$\mathbb{E}[\mathbb{1}_A \log \widehat{Z}(\beta)] \geq -\beta \mathbb{E}[-\mathbb{1}_A \overline{R}(c_0)] \geq -\beta \mathbb{E}[|\overline{R}(c_0)|] \geq -L\sigma\beta\sqrt{N} + o(1), \tag{97}$$

where L is some constant coming from expectation of half-normal distribution, which is the limiting distribution for $|\overline{R}(c_0)|$. Here we use the fact that $|\overline{R}(c_0)|$ converges in distribution to a half-normal (due to CLT), and then we determine that, due to the dominated convergence theorem and uniform integrability of $|\overline{R}(c_0)|$ (Feller, 1971, Ch. XVI.7), the expectation value of $|\overline{R}(c_0)|$ also converges to the one of half-normal.

Combining (96) and (97), we obtain

$$\mathbb{E}[\log \widehat{Z}(\beta)] \geq (1 - \mathbb{P}(A)) \left(\log m + \log a_n - \log 2 + \beta u_n(\widehat{\beta})\right) - L\sigma\beta\sqrt{N} + o(1). \tag{98}$$

In summary, by Lemmas 9 and 10 we arrive at

$$\frac{\mathbb{E}[\log \widehat{Z}(\beta)]}{\log m} \geq 1 - \frac{u_n(\widehat{\beta})^2}{2\sigma^2 N \log m} + \frac{\beta u_n(\widehat{\beta})}{\log m} - \frac{L\sigma\beta\sqrt{N}}{\log m} + o(1) \tag{99}$$

$$= 1 - \frac{u_n(\widehat{\beta})^2}{2\sigma^2 N \log m} + \frac{\beta u_n(\widehat{\beta})}{\log m} + o(1). \tag{100}$$

Now for the regime $\widehat{\beta} < \widehat{\beta}^*$, recall that $u_n(\widehat{\beta}) = \widehat{\beta}\sigma^2\sqrt{N \log m}$, which yields

$$\frac{\mathbb{E}[\log \widehat{Z}(\beta)]}{\log m} \geq 1 + \frac{\widehat{\beta}^2\sigma^2}{2} + o(1). \tag{101}$$

For regime $\widehat{\beta} \geq \widehat{\beta}^*$, $u_n(\widehat{\beta}) = \widehat{\beta}^*\sigma^2\sqrt{N \log m}$, hence

$$\frac{\mathbb{E}[\log \widehat{Z}(\beta)]}{\log m} \geq 1 - \frac{\widehat{\beta}^{*\,2}\sigma^2}{2} + \widehat{\beta}\widehat{\beta}^*\sigma^2 + o(1). \tag{102}$$

All in all, we have

$$\lim_{n\to\infty} \frac{\mathbb{E}[\log Z(\beta, X)] + \widehat{\beta}\mu\sqrt{N \log m}}{\log m} \geq \begin{cases} 1 + \dfrac{\widehat{\beta}^2\sigma^2}{2}, & \widehat{\beta} < \widehat{\beta}^*, \\ 1 - \dfrac{\widehat{\beta}^{*\,2}\sigma^2}{2} + \widehat{\beta}\widehat{\beta}^*\sigma^2, & \widehat{\beta} \geq \widehat{\beta}^*. \end{cases} \tag{103}$$

Theorem 3 is proven if we establish Lemma 9, which we do next.

Proof of Lemma 9. First, note that if solutions (permutations) π and π' have k common points, then they share k^2 entries of the 4-dimensional matrix Q (out of the $N = n^2$ entries appearing in $R(\pi, X)$). Besides, since the solution space $\mathcal{C}_n$ (the set of permutations of size n) is a group, there exists a permutation π'' such that $\pi' = \pi \circ \pi''$. Thus, counting the common points between π and π' is equivalent to counting the fixed points of π'', which is a well-studied problem.

The number of permutations with k fixed points is the rencontre number (see e.g., Szpankowski (2001))

$$D_{n,k} = \frac{n!}{k!}\sum_{j=0}^{n-k}\frac{(-1)^j}{j!}. \tag{104}$$

Therefore, the number of ordered pairs of permutations sharing k fixed points is

$$B_{n,k} = n!D_{n,k} = \frac{n!^2}{k!}\sum_{j=0}^{n-k}\frac{(-1)^j}{j!}. \tag{105}$$

Similarly to the case of the sMBP, we can break down $\mathbb{E}[Y^2]$ using those elements

$$\mathbb{E}[Y^2] = \sum_{\pi,\pi'\in C_n} \mathbb{P}\left(\overline{R}(\pi, X) \geq u_n(\widehat{\beta}) \text{ and } \overline{R}(\pi', X) \geq u_n(\widehat{\beta})\right) \tag{106}$$

$$= \sum_{k=0}^{N} B_{n,k}\mathbb{P}\left(O_{n,k} + I_{n,k} \geq u_n(\widehat{\beta}) \text{ and } O'_{n,k} + I_{n,k} \geq u_n(\widehat{\beta})\right) \tag{107}$$

where $O_{n,k}, O'_{n,k} \sim \mathcal{N}(0, N - k^2)$, $I_{n,k} \sim \mathcal{N}(0, k^2)$ are independent of each other, $I_{n,k}$ represents the sum of the entries shared by the two solutions, and $O_{n,k}, O'_{n,k}$ the entries exclusive to one of the two.

Let us now bound the probability

$$p_{n,k}(\widehat{\beta}) = \mathbb{P}\left(O_{n,k} + I_{n,k} \geq u_n(\widehat{\beta}) \text{ and } O'_{n,k} + I_{n,k} \geq u_n(\widehat{\beta})\right) \tag{108}$$

that two solutions with k^2 shared entries exceed the threshold $u_n(\widehat{\beta})$. This is exactly the probability that the two coordinates of a multivariate centered normal vector with covariance matrix $\begin{pmatrix} n^2 & k^2 \\ k^2 & n^2 \end{pmatrix}\sigma^2$ exceed $u_n(\widehat{\beta})$. Applying the results of (Savage, 1962) about multivariate Gaussian bounds, we have

$$p_{n,k} \leq \frac{\sigma^2}{2\pi u_n(\widehat{\beta})^2}\sqrt{\frac{(n^2+k^2)^3}{n^2-k^2}}\exp\left(-\frac{u_n(\widehat{\beta})^2}{(n^2+k^2)\sigma^2}\right) \tag{109}$$

$$= \frac{1}{2\pi\widehat{\beta}^2\sigma^2}\frac{1}{n^2\log n!}\sqrt{\frac{(n^2+k^2)^3}{n^2-k^2}}\exp\left(-\widehat{\beta}^2\sigma^2\frac{n^2\log n!}{(n^2+k^2)\sigma^2}\right) \tag{110}$$

for $k < n$.

For $k = n$, $p_{n,n} = a_n$ and we know that

$$a_n \sim \frac{n\sigma}{\sqrt{2\pi}u_n(\widehat{\beta})}\exp\left(-\frac{u_n(\widehat{\beta})^2}{2n^2\sigma^2}\right) \tag{111}$$

$$= \frac{1}{\sqrt{2\pi\log n!}\widehat{\beta}\sigma}\exp\left(-\frac{\widehat{\beta}^2\sigma^2}{2}\log n!\right). \tag{112}$$

Combining equations (105), (107), (110) and (112) yield

$$\frac{\mathbb{E}[Y^2]}{m^2a_n^2} \lesssim S_n = \sum_{k=0}^{n-1}\frac{1}{k!}\left(\sum_{j=0}^{n-k}\frac{(-1)^j}{j!}\right)\frac{1}{n^2}\sqrt{\frac{(n^2+k^2)^3}{n^2-k^2}}\mathrm{e}^{\left(1-\frac{n^2}{n^2+k^2}\right)\widehat{\beta}^2\sigma^2\log n!}$$
$$+ \frac{\sqrt{2\pi\log n!}\widehat{\beta}\sigma}{n!}\exp(\frac{\widehat{\beta}^2\sigma^2}{2}\log n!). \tag{113}$$

It is obvious that the term outside of the sum will tend to 0 as long as $\widehat{\beta} < \sqrt{2}/\sigma$. Let us now address the asymptotics of

$$S_n = \sum_{k=0}^{n-1}\frac{1}{k!}\left(\sum_{j=0}^{n-k}\frac{(-1)^j}{j!}\right)\frac{1}{n^2}\sqrt{\frac{(n^2+k^2)^3}{n^2-k^2}}\mathrm{e}^{\left(1-\frac{n^2}{n^2+k^2}\right)\widehat{\beta}^2\sigma^2\log n!}. \tag{114}$$

For that, set $k = o\left(\frac{n}{\log n}\right)$ and consider the following approximations of various terms of S_n.

First,

$$\sum_{j=0}^{n-k}\frac{(-1)^j}{j!} = \frac{1}{\mathrm{e}} + \mathcal{O}\left(\frac{1}{(n-k)!}\right). \tag{115}$$

Second,

$$\frac{1}{n^2}\sqrt{\frac{(n^2+k^2)^3}{n^2-k^2}} = 1 + \mathcal{O}\left(\frac{k^2}{n^2}\right). \tag{116}$$

Eventually,

$$\mathrm{e}^{\left(1-\frac{n^2}{n^2+k^2}\right)\widehat{\beta}^2\sigma^2\log n!} \sim \mathrm{e}^{\frac{k^2}{n^2}n\log n} = 1+\mathcal{O}\Big(\frac{k\log n}{n}\Big). \tag{117}$$

Plugging back into S_n, we find

$$S_n = \sum_{k=0}^{\infty}\frac{1}{k!}\cdot\frac{1}{\mathrm{e}}\cdot\left(1+\mathcal{O}\Big(\frac{k^2}{n^2}\Big)\right)\cdot\left(\mathcal{O}\Big(\frac{k\log n}{n}\Big)\right) = 1+\mathcal{O}\Big(\frac{1}{n^{\epsilon}}\Big), \tag{118}$$

provided that $k=\frac{n^{1-\epsilon}}{\log n}$. Thus,

$$\frac{\mathbb{E}[Y^2]}{m^2a_n^2}\to 1+O(n^{-\varepsilon}) \ \text{ and } \ \mathbb{P}(A)\le\frac{4\mathrm{Var}[Y]}{m^2a_n^2}=O(n^{-\varepsilon}). \tag{119}$$

This proves the lemma. □

5. Remarks, Conjectures and Future Work

This paper investigates the asymptotics of the information score called the *expected log-posterior agreement* to validate cost functions and algorithms for "parameter rich" combinatorial optimization problems. As subtasks, first we provided rigorous derivations for free energy of Sparse MBP and Lawler QAP. However, for general MBP and QAP we do not expect the lower bound to match the upper bound found in Theorem 1. In fact, based on extensive simulation we concluded that there is an additional scaling in the part of linear growth. To establish it, we realize that we need some new techniques to prove lower bounds. Second, we showed that two second order phase transitions occur for the expected log-posterior agreement. Our analysis and experimental results show three regions of the expected log-posterior agreement: a high temperature phase with low information, a retrieval phase and a disordered frozen phase. Only the retrieval phase can be used for efficient sampling solutions. While investigating the asymptotics of the log-posterior agreement and free energy we faced a challenging mathematical problem leading to new research on the interplay between statistical physics and computation. We hope that techniques presented here can be successfully used for a large class of different combinatorial structures and problems.

We also have proposed empirically-inspired conjectures for approximating free energy for general problems (i.e. for MBP, QAP, and potentially other problems). These conjectures are well supported by our experiments and by a rigorous analysis for special cases (i.e. conjecture turns into proven asymptotics for Sparse MBP, Lawler QAP), as explained below.

5.1. Sampling procedure for simulating the free energy

To produce simulations of the partition function for any given optimization problem, we use a Metropolis-Hastings procedure to sample solutions at a given temperature $1/\beta$, coupled with an *importance sampling* cooling schedule scheme

to efficiently sample solutions at low temperature levels. Below, we provide a brief review of importance sampling.

Importance sampling. Let us assume that samples from a distribution $\mathbb{Q}$ over a random variable X are given. Then, the expectation $\mathbb{E}_{\mathbb{P}}\phi(X)$ of a function $\phi(X)$ under a distribution $\mathbb{P}$ can be estimated by sampling X under $\mathbb{Q}$ with

$$\widehat{E}_N = \frac{1}{N}\sum_{i=1}^{N}\phi(X_i)\frac{\mathbb{P}(X_i)}{\mathbb{Q}(X_i)}, \quad \text{since} \quad \mathbb{E}_{\mathbb{Q}}\widehat{E}_N = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathbb{Q}}\phi(X_i)\frac{\mathbb{P}(X_i)}{\mathbb{Q}(X_i)} = \mathbb{E}_{\mathbb{P}}\phi(X). \tag{120}$$

This method is called the importance sampling since each sample is "reweighted" using the target distribution.

We adapt it for the computation of Gibbs distribution partition functions. Suppose we have Gibbs distribution at temperature $1/\beta$ over a space $\mathcal{C}$ defined by a cost function $R : \mathcal{C} \to \mathbb{R}$. The probability of $c \in \mathcal{C}$ is $\mathbb{P}(c|\beta) = \mathrm{e}^{-\beta R(c)}/Z(\beta)$ where $Z(\beta) = \sum_{c\in\mathcal{C}}\mathrm{e}^{-\beta R(c)}$ is the partition function. Let us assume the partition function $Z(\beta)$ is given and we can sample from the Gibbs distribution at temperature $1/\beta'$. Then

$$Z_N^*(\beta,\beta') = \frac{1}{N}\sum_{i=1}^{N} Z(\beta)\mathrm{e}^{-(\beta'-\beta)R(c_i)} \tag{121}$$

is an unbiased estimator of $Z(\beta')$ when sampled under $\mathbb{P}(\cdot|\beta)$. Its precision is controlled by the relative variance:

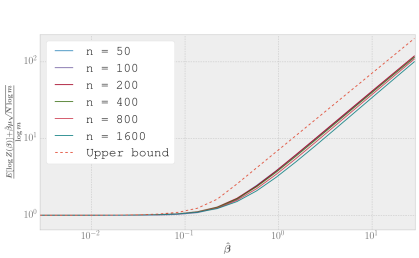
$$\mathrm{Var}^{\mathrm{rel}}_{\mathbb{P}(\cdot|\beta)} Z_N^*(\beta,\beta') = \frac{\mathrm{Var}_{\mathbb{P}(\cdot|\beta)} Z_N^*(\beta,\beta')}{\mathbb{E}^2_{\mathbb{P}(\cdot|\beta)} Z_N^*(\beta,\beta')} = \frac{1}{N}\left(\frac{Z(2\beta'-\beta)Z(\beta)}{Z(\beta')^2} - 1\right). \tag{122}$$

Observe that when β differs significantly from β', the variance may be large, leading to poor simulations results. Furthermore, when β is close to β', the variance is small, thus simulations are more accurate.

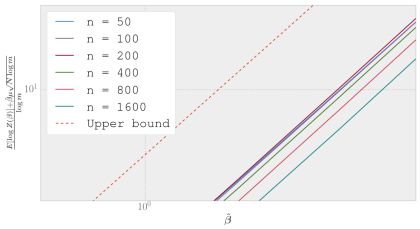
Our goal is to estimate the partition function for a wide range of β. The difficulties arise mostly for large values of β, since the partition function is then very concentrated. To overcome this, we apply our importance sampling philosophy and simulate first the partition function for small values of β (this makes the partition function more uniform and easier to estimate). Once we have computed the partition function for small β, we use equation (121) to evaluate it for the targeted value β'. But that is not the end of the story since we need to proceed in small steps using a cooling schedule $\beta_0 = 0 < \beta_1 < \cdots < \beta_k$ in order to reach the regions of the solution space contributing the most to the partition function. This is called a cooling schedule (Huber, 2012). In practice, we use a Metropolis-Hastings procedure to sample from the Gibbs distribution at a given temperature.

5.2. Results of the simulation

Figure 6 shows the simulation of the free energy in the case of the minimum bisection problem ($d = n/2$) for different graph sizes n. The dashed line corre-

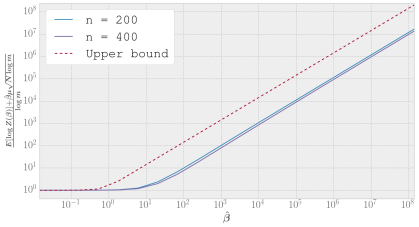


(a) Full view

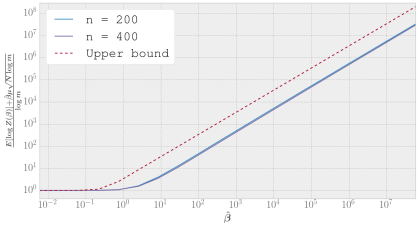


(b) A zoom on the right part of the plot

Figure 6: Second-order terms of the free energy rate in the case of the minimum bisection problem. The edge weights are i.i.d. and generated from a Gaussian distribution $\mathcal{N}(\mu = 20, \sigma = 5)$. Every curve is the average of 10 different problem instances. The curve labeled "upper bound" corresponds to the prediction of Theorem 1.



(a) Standard deviation of $\sigma = 1$



(b) Standard deviation of $\sigma = 2.4$

Figure 7: Second-order terms of the free energy rate in the case of the quadratic assignment problem. The distance and weight matrix entries are i.i.d. and generated from equal Gaussian distributions so that the product of two entries has mean $\mu = 4$ and varying standard deviation. Every curve is the average of 10 different problem instances. The curve labeled "upper bound" corresponds to the prediction of Theorem 1.

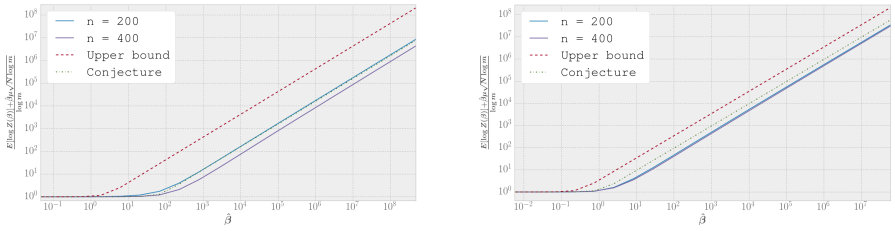
sponds to the upper bound defined in Theorem 1. It appears that the general behavior is quite good for the quadratic part of the free energy while for the linear part there is some discrepancy (a multiplicative factor correction is needed).

Figure 7 shows the simulation of the free energy in the case of the quadratic assignment problem for different graph sizes n. The dashed line corresponds to the upper bound defined in Theorem 1. The two plots correspond to different variances. Interestingly, in this problem the correction coefficient depends on the variance, which was not the case for the minimum bisection problem. Indeed, the correction coefficient is around $1/12$ for $\sigma = 1.0$ and near $1/8$ for $\sigma = 2.4$.

5.3. Conjecture

Based on our empirical results presented in Figures 6 and 7, we are able to conjecture a more precise behavior of the free energy for the two optimization problems discussed in this paper. We shall introduce a correction coefficient α whose value is determines experimentally in the sequel.

Conjecture 1. *Consider a class of combinatorial optimization problems complying with Common Theorem Setting, weights W_i having mean μ and variance*



(a) $\sigma = 0.3$ (b) $\sigma = 2.4$

Figure 8: Influence of the correction in the case of the quadratic assignment problem for different standard deviation. The mean is $\mu = 4$ and $\mu_V = \mu_H$ and $\sigma_V^2 = \sigma_H^2$.

σ^2. *Then the free energy satisfies*

$$\lim_{n\to\infty} \frac{\mathbb{E}[\log Z(\beta, X)] + \widehat{\beta}\mu\sqrt{N \log m}}{\log m} = \begin{cases} 1 + \alpha^2 \frac{\widehat{\beta}^2\sigma^2}{2}, & \widehat{\beta} < \frac{\sqrt{2}}{\alpha\sigma} \\ \alpha\widehat{\beta}\sigma\sqrt{2}, & \widehat{\beta} \geq \frac{\sqrt{2}}{\alpha\sigma} \end{cases} \tag{123}$$

for some $\alpha \geq 1$.

The correction coefficient α is related to the variance of the partition function which involves strong correlations between feasible solutions (that was largely ignored in (Buhmann et al., 2014)). Based on our experimental results, we conclude that α is well approximated by the following formula

$$\alpha = \sqrt{\frac{\mathbb{E}_X \mathrm{Var}_c R(c, X)}{\mathbb{E}_c \mathrm{Var}_X R(c, X)}} = \sqrt{\frac{\mathbb{E}_X \mathrm{Var}_c R(c, X)}{N\sigma^2}} \tag{124}$$

where the expectation $\mathbb{E}_c[\cdot]$ is taken w.r.t. to all feasible solutions selected uniformly.

Appendix A. A Tighter Upper Bound for MBP

The general upper bound proven in Theorem 1 above is unfortunately not tight. To show it we consider the general minimum bisection problem with n vertices, that is, $d = n$. In this case, $N = |\mathcal{S}_n(c)| = n^2/4$ is the number of edges cut in a bisection, and $m = |\mathcal{C}_n| = \binom{n}{n/2}$ is the number of possible bisections. Thus we are still in our framework of $\log m = o(N)$, and therefore we define a scaling $\beta = \widehat{\beta}\sqrt{\log m/N}$.

Theorem 11. *For the general minimum bisection problem the following holds for* $\widehat{\beta} \leq \frac{1}{\sqrt{\log 2}\sigma}$

$$\lim_{n\to\infty} \frac{\mathbb{E}[\log Z(\beta, X)] + \widehat{\beta}\mu\sqrt{N \log m}}{\log m} \leq 1 + \frac{\widehat{\beta}^2\sigma^2}{4}. \tag{A.1}$$

Remark. The idea of the proof is that the minimum bisection problem is a constrained version of the Sherrington-Kirkpatrick model, which is a spin model where all the spins are independent (cf. Sherrington and Kirkpatrick, 1975). In

the minimum bisection problem, it is required that the partition of the graph be balanced, or equivalently rephrased in spin model terms, it is required that there is the same number of up-spins as down-spins.

Therefore, the only difference between the two problems is the solution space. More precisely, we have $\mathcal{C}_n^{\mathrm{MBP}} \subset \mathcal{C}_n^{\mathrm{SK}}$. Hence

$$Z^{\mathrm{MBP}}(\beta) = \sum_{c \in \mathcal{C}_n^{\mathrm{MBP}}} \mathrm{e}^{-\beta R(c,X)} \leq \sum_{c \in \mathcal{C}_n^{\mathrm{SK}}} \mathrm{e}^{-\beta R(c,X)} = Z^{\mathrm{SK}}(\beta), \tag{A.2}$$

which allows us to extend any upper bound on Z^{SK} to Z^{MBP}. In particular, Talagrand provides such an upper bound in (Talagrand, 2003).

Proof of Theorem 11 First, we introduce some alternate notations for the minimum bisection problem in order to ease the transition to the Sherringkton-Kirkpatrick formalism. Denote by G an undirected weighted complete graph with n vertices. The problem consists in finding a bisection of the graph (a partition in two subsets of equal size) of minimum cost. More formally, define by g_{ij} the weight assigned to the edge between vertices i and j ($g_{ij} = g_{ji}$). Denote by $c_i \in \{-1, 1\}$ an indicator of the subset containing vertex i.

We need to find $c \in \{-1, 1\}^n$ such that $\sum_i c_i = 0$ (balance condition) and the sum of the weights of cut edges

$$R(c, X) = \sum_{\substack{c_i = -c_j \\ i<j}} g_{ij} \tag{A.3}$$

is minimal. Here, X denotes a problem instance of size n, i.e. the particular values $(g_{ij})_{ij}$ of the edge weights.

Define the partition function as

$$Z(\beta, X) = \sum_{c \in \mathcal{C}_n} \mathrm{e}^{-\beta R(c,X)} \tag{A.4}$$

where $\mathcal{C}_n = \{c \in \{-1, 1\}^n | \sum_i c_i = 0\}$ is the solution space. Let us now prove Theorem 11. Observe that

$$\begin{aligned} \log Z(\beta, X) + \widehat{\beta}\mu\sqrt{N \log m} &= \log Z(\beta, X) + \beta\mu N \\ &= \log \sum_{c \in \mathcal{C}_n} \mathrm{e}^{-\beta(R(c,X) - N\mu)} = \log Z(\beta, \overline{X}), \end{aligned} \tag{A.5}$$

where the edge weights of $\overline{X}$ are defined by $\bar{g}_{ij} = g_{ij} - \mu$. Hence without loss of generality, we will only consider centered problem instances in the rest of the proof. For clarity, the explicit mention of the dependence to X is dropped in the partition function, i.e. $Z(\beta, X) := Z(\beta)$.

Then, let us relax our problem by allowing the partitions to be unbalanced:

$$Z^*(\beta) = \sum_{c \in \mathcal{C}_n^*} \mathrm{e}^{-\beta R(c)} \tag{A.6}$$

where $\mathcal{C}_n^* = \{-1, 1\}^n$ is the relaxed set of solutions. Since $\mathcal{C}_n \subset \mathcal{C}_n^*$, it follows

that

$$Z(\beta) \leq Z^*(\beta). \tag{A.7}$$

Now rewrite the cost function as

$$R(c) = \sum_{\substack{c_i=-c_j \\ i<j}} g_{ij} = \frac{1}{2}\Big(\sum_{i<j} g_{ij} - \sum_{i<j} c_i c_j g_{ij}\Big) = \frac{1}{2}\Big(\sum_{i<j} g_{ij} + \sqrt{n} R^{\mathrm{SK}}(c)\Big) \tag{A.8}$$

where

$$R^{\mathrm{SK}}(c) = -\frac{1}{\sqrt{n}} \sum_{i<j} c_i c_j g_{ij} \tag{A.9}$$

is the cost function of the Sherrington-Kirkpatrick model. This entails

$$Z^*(\beta) = \mathrm{e}^{-\frac{\beta}{2}\sum_{i<j} g_{ij}} \sum_{c\in\mathcal{C}_n^*} \mathrm{e}^{-\frac{\sqrt{n}\beta}{2} R^{\mathrm{SK}}(c)} = \mathrm{e}^{-\frac{\beta}{2}\sum_{i<j} g_{ij}} Z^{\mathrm{SK}}\Big(\frac{\sqrt{n}\beta}{2}\Big) \tag{A.10}$$

where

$$Z^{\mathrm{SK}}(\beta) = \sum_{c\in\mathcal{C}_n^*} \mathrm{e}^{-\beta R^{\mathrm{SK}}(c)} \tag{A.11}$$

is the partition function associated with the Sherrington-Kirkpatrick model. Since the g_{ij} are centered, it follows that

$$\mathbb{E}\big[\log Z^*(\beta)\big] = \mathbb{E}\Big[\log Z^{\mathrm{SK}}\Big(\frac{\sqrt{n}\beta}{2}\Big)\Big]. \tag{A.12}$$

We need now the following statement from (Talagrand, 2003), that we present next.

Theorem 12 (Talagrand, 2003, Theorem 2.2.1)**.** *If $\beta < \frac{1}{\sigma}$, we have*

$$\lim_{n\to\infty} \frac{1}{n}\mathbb{E}\big[\log Z^{SK}(\beta)\big] = \frac{\beta^2\sigma^2}{4} + \log 2. \tag{A.13}$$

Now let us determine the limit of $\sqrt{n}\beta$:

$$\sqrt{n}\beta = \sqrt{n}\,\widehat{\beta}\sqrt{\frac{\log m}{N}} = \sqrt{n}\,\widehat{\beta}\sqrt{\frac{\log\binom{n}{n/2}}{n^2/4}} \sim \sqrt{n}\,\widehat{\beta}\sqrt{\frac{n\log 2}{n^2/4}} = 2\sqrt{\log 2}\,\widehat{\beta} \tag{A.14}$$

Thus, we can use Theorem 12 to obtain, for $\widehat{\beta} < \frac{1}{\sqrt{\log 2}\sigma}$

$$\lim_{n\to\infty} \frac{1}{n}\mathbb{E}\Big[\log Z^{\mathrm{SK}}\Big(\frac{\sqrt{n}\beta}{2}\Big)\Big] = \Big(\frac{\widehat{\beta}^2\sigma^2}{4} + 1\Big)\log 2. \tag{A.15}$$

The equivalence $\frac{\log m}{n} \sim \log 2$ (in $n\to\infty$) and (A.12) both allow to write

$$\lim_{n\to\infty} \frac{\mathbb{E}[\log Z^*(\beta)]}{\log m} = \frac{\widehat{\beta}^2\sigma^2}{4} + 1 \tag{A.16}$$

for $\widehat{\beta} < \frac{1}{\sqrt{\log 2\sigma}}$. Now (A.7) implies that

$$\lim_{n\to\infty} \frac{\mathbb{E}[\log Z(\beta)]}{\log m} \le \frac{\widehat{\beta}^2\sigma^2}{4} + 1 \tag{A.17}$$

for $\widehat{\beta} < \frac{1}{\sqrt{\log 2\sigma}}$. □

Appendix B. Proof of (56)

We prove the following lemma that establishes (56).

Lemma 13. *For any $\beta > 0$ we have*

$$\mathrm{Var} Z = (\mathbb{E}Z)^2 \Big(\mathbb{E}_{\mathcal{D}} \Big(\frac{G(2\beta)}{G^2(\beta)} \Big)^D - 1 \Big) \tag{B.1}$$

where D is a random variable denoting the size of the elementwise overlap for two solutions $c, c' \in \mathcal{C}$, chosen uniformly at random (this uniformness is addresses in $\mathcal{D}$). Here, $G(\beta)$ is the moment generating function of the negative weights $(-W_i)$.

Proof. Let

$$Z(\beta) = \exp(-\beta N\mu)\widehat{Z}(\beta), \tag{B.2}$$

where, as previously, $\widehat{Z}(\beta) = \sum_{c\in\mathcal{C}} \exp(\beta \overline{R}(c))$ with $\overline{R}(c) = -\sum_{i\in\mathcal{S}(c)} \overline{W}_i$. To compute $\mathrm{Var}\widehat{Z}$, we proceed as follows

$$\begin{aligned} \mathbb{E}\widehat{Z}^2 &= \mathbb{E}\Big[\sum_{c\in\mathcal{C}} \exp(\beta\overline{R}(c)) \cdot \sum_{c'\in C} \exp(\beta\overline{R}(c')) \Big] \\ &= \sum_{c,c'\in\mathcal{C}} \mathbb{E}\exp\Big(-\beta \Big(\sum_{i\in\mathcal{S}(c)} \overline{W}_i + \sum_{j\in\mathcal{S}(c')} \overline{W}_j \Big) \Big). \end{aligned} \tag{B.3}$$

Now define the elementwise overlap between the solutions c and c' as $\mathcal{S}_{\mathrm{ovr}}(c,c') := \mathcal{S}(c)\cap\mathcal{S}(c')$, and its cardinality $d = d(c,c') := |\mathcal{S}(c,c')|$. We also define the symmetric difference $\bar{\mathcal{S}}_{\mathrm{ovr}}(c,c') := \mathcal{S}(c)\triangle\mathcal{S}(c')$ and continue the chain of equalities:

$$\mathbb{E}\widehat{Z}^2 = \sum_{c,c'\in\mathcal{C}} \mathbb{E}\exp\Big(-\beta \Big(2 \sum_{i\in\mathcal{S}_{\mathrm{ovr}}(c,c')} \overline{W}_i + \sum_{j\in\bar{\mathcal{S}}_{\mathrm{ovr}}(c,c')} \overline{W}_j \Big) \Big). \tag{B.4}$$

Here the sets of weights $\mathcal{S}_{\mathrm{ovr}}(c,c')$ and $\bar{\mathcal{S}}_{\mathrm{ovr}}(c,c')$ are independent, allowing us to decompose the expectation into the product:

$$\begin{aligned} \mathbb{E}\widehat{Z}^2 &= \sum_{c,c'\in\mathcal{C}} \mathbb{E}\exp\Big(-\beta \Big(2 \sum_{i\in\mathcal{S}_{\mathrm{ovr}}(c,c')} \overline{W}_i \Big) \Big) \cdot \mathbb{E}\exp\Big(-\beta \Big(\sum_{j\in\bar{\mathcal{S}}_{\mathrm{ovr}}(c,c')} \overline{W}_j \Big) \Big) \\ &= \sum_{c,c'\in\mathcal{C}} \big(\widehat{G}(2\beta)\big)^d \big(\widehat{G}(\beta)\big)^{2(N-d)} = \big(\widehat{G}(\beta)\big)^{2N} \sum_{c,c'\in\mathcal{C}} \Big(\frac{\widehat{G}(2\beta)}{(\widehat{G}(\beta))^2} \Big)^d. \end{aligned} \tag{B.5}$$

Now assume that the probability of the two solutions c and c', chosen uniformly at random, to have a d-element overlap is $\mathbb{P}_{\mathcal{D}}(d)$ and rewrite the above as follows:

$$\mathbb{E}\widehat{Z}^2 = \big(\widehat{G}(\beta)\big)^{2N} \sum_{d=0}^{N} m^2 \mathbb{P}_{\mathcal{D}}(d) \bigg(\frac{\widehat{G}(2\beta)}{(\widehat{G}(\beta))^2} \bigg)^d = m^2 \big(\widehat{G}(\beta)\big)^{2N} \sum_{d=0}^{N} \mathbb{P}_{\mathcal{D}}(d) \bigg(\frac{\widehat{G}(2\beta)}{(\widehat{G}(\beta))^2} \bigg)^d$$
$$= (\mathbb{E}\widehat{Z})^2 \sum_{d=0}^{N} \mathbb{P}_{\mathcal{D}}(d) \bigg(\frac{\widehat{G}(2\beta)}{(\widehat{G}(\beta))^2} \bigg)^d. \tag{B.6}$$

We conclude that

$$\mathrm{Var}\widehat{Z} = \mathbb{E}\widehat{Z}^2 - (\mathbb{E}\widehat{Z})^2 = (\mathbb{E}\widehat{Z})^2 \bigg(\mathbb{E}_{\mathcal{D}} \bigg(\frac{\widehat{G}(2\beta)}{(\widehat{G}(\beta))^2} \bigg)^D - 1 \bigg). \tag{B.7}$$

Recalling that $Z(\beta) = \exp(-\beta N \mu) \widehat{Z}(\beta)$ and, as well, $G(\beta) = \exp(-\beta\mu)\widehat{G}(\beta)$, we obtain the version without hats:

$$\mathrm{Var}Z = \mathbb{E}Z^2 - (\mathbb{E}Z)^2 = (\mathbb{E}Z)^2 \bigg(\mathbb{E}_{\mathcal{D}} \bigg(\frac{G(2\beta)}{(G(\beta))^2} \bigg)^D - 1 \bigg). \tag{B.8}$$

This proves Lemma 13.

References

Aizenman, M., Lebowitz, J. L., Ruelle, D., 1987. Some rigorous results on the Sherrington-Kirkpatrick spin glass model. Communications in Mathematical Physics 112 (1), 3–20.

Bovier, A., Kurkova, I., Löwe, M., 2002. Fluctuations of the free energy in the rem and the p-spin sk models. The Annals of Probability 30 (2), 605–651.

Buhmann, J. M., 2010. Information theoretic model validation for clustering. In: International Symposium on Information Theory (ISIT). Austin, TX, USA, pp. 1398–1402.

Buhmann, J. M., Gronskiy, A., Szpankowski, W., 2014. Free energy rates for a class of very noisy optimization problems. In: Analysis of Algorithms (AofA). France, pp. 67–78.

Busse, L., Chehreghani, M., Buhmann, J. M., 2013. German Conference on Pattern Recognition. Springer Berlin Heidelberg, Berlin, Heidelberg, Ch. Approximate Sorting, pp. 142–152.

Chehreghani, M. H., Busetto, A.-G., Buhmann, J. M., 2012. Information theoretic model validation for spectral clustering. In: International Conference on Artificial Intelligence and Statistics (AISTATS). Vol. 22. pp. 495–503.

Cohen, J. E., 1988. Threshold phenomena in random structures. Discrete Applied Mathematics 19 (1), 113–128.
"alex5.bbl" 121 lines, 5452 characters Mathematics 19 (1), 113–128.

Dembo, A., Zeitouni, O., 2009. Large deviations techniques and applications. Vol. 38. Springer Science & Business Media.

Derrida, B., 1981. Random-energy model: An exactly solvable model of disordered systems. Phys. Rev. B 24, 2613–2626.

Derrida, B., Gardner, E., 1986. Solution of the generalised random energy model. Journal of Physics C: Solid State Physics 19 (13), 2253–2274.

Feller, W., 1971. An Introduction to Probability Theory and Its Applications, 2nd Edition. Vol. 2. Wiley, NY.

Frenk, J. B. G., van Houweninge, M., Kan, A. H. G. R., 1985. Asymptotic properties of the quadratic assignment problem. Mathematics of Operations Research 10 (1), 100–116.

Huber, M. L., 2012. Approximation algorithms for the normalizing constant of gibbs distributions. arXiv preprint arXiv:1206.2689.

Lawler, E. L., 1963. The quadratic assignment problem. Management science 9 (4), 586–599.

Łuczak, T., 1994. Phase transition phenomena in random discrete structures. Discrete Mathematics 136 (1), 225 – 242.

Magner, A., Szpankowski, W., Kihara, D., 2015. On the Origin of Protein Superfamilies and Superfolds. Scientific Reports, 5: 8166.

Magner, A., Kihara, D., Szpankowski, W., 2016. The boltzmann sequence-structure channel. *Proceedings of the IEEE*, 2016; also IEEE International Symposium on Information Theory, ISIT 2016, Barcelona, Spain, July 10-15, 2016. pp. 255–259.

Mézard, M., Montanari, A., 2009. Information, Physics, and Computation. Oxford University Press.

Mézard, M., Parisi, G., 2003. The cavity method at zero temperature. Journal of Statistical Physics 111 (1), 1–34.

Parisi, G., 2009. The mean field theory of spin glasses: The heuristic replica approach and recent rigorous results. Letters in Mathematical Physics 88 (1), 255–269.

Savage, I. R., 1962. Mills' ratio for multivariate normal distributions. J. Res. Nat. Bur. Standards Sect. B 66, 93–96.

Sherrington, D., Kirkpatrick, S., 1975. Solvable Model of a Spin-Glass. Physical Review Letters 35, 1792–1796.

Szpankowski, W., 1995. Combinatorial optimization problems for which almost every algorithm is asymptotically optimal. Optimization 33, 359–368.

W. Szpankowski. *Avarage Case Analysis of Algorithms on Sequences*, Wiley, New York, 2001.

Talagrand, M., 2003. Spin Glasses: A Challenge for Mathematicians: Cavity and Mean Field Models. Springer Verlag.

Vannimenus, J., Mézard, M., 1984. On the statistical mechanics of optimization problems of the travelling salesman type. Phys. Rev. Lett. 35, 1792–1796.