JMLR: Workshop and Conference Proceedings vol 40:1–19, 2015

A Validation Criterion for a Class of Combinatorial Optimization Problems via Helmholtz Free Energy and its Asymptotics

Joachim M. Buhmann[*] JBUHMANN@INF.ETHZ.CH
Julien Dumazert DUJULIEN@STUDENT.ETHZ.CH
Alexey Gronskiy[*] ALEXEY.GRONSKIY@INF.ETHZ.CH
ETHZ Zurich, Switzerland

Wojciech Szpankowski[†] SPA@CS.PURDUE.EDU
Purdue University, USA

Abstract

Most pattern recognition problems are modeled as an optimization of problem-dependent objective functions. The uncertainty in the input, but also the computational constraints arising from "big data" volumes require to regularize such objectives. Conceptually, the model and its regularization have to be rigorously validated from a statistical viewpoint. In this paper we consider an information theoretical concept for validating objective functions of several combinatorial optimization problems. The validation criterion measures the overlap of Gibbs distributions for cost functions that rank solutions for different typical input data sets. The maximum entropy inference method utilizes free energy as an ingredient. We provide both rigorous approaches and empirical insights to determining the asymptotics of the free energy — an important system-defining quantity of a big stochastic systems (problems). More precisely, we obtain asymptotic upper bounds for the free energy for a class of optimization problems. Further, we conjecture an informal empirical correction to it, which allows to reach a more precise asymptotical behavior of the free energy. We verify findings through extensive importance sampling simulations.

Keywords: Helmholz Free energy, Optimization, Gibbs distribution, Random energy model, Partition function asymptotics, Minimum Bisection, Quadratic Assignment Problem

1. Introduction

1.1. Maximum entropy inference for modeling

The search for patterns in data analysis is mostly guided by cost functions that depend on noisy data. Well known examples are graph cut methods to identify clusters in proximity data, Euclidean embeddings of relational data by multidimensional scaling or fitting phylogenetic trees to protein dissimilarities. Robust methods to find good solutions have

[*] This work was supported by SNF Grant # 200021_138117.

[†] This work was supported by NSF Center for Science of Information (CSoI) Grant CCF-0939370, and in addition by NSA Grant 130923, NSF Grant DMS-0800568, and and the MNSW grant DEC-2013/09/B/ST6/02258. W. Szpankowski is also with the Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, Poland.

to average over fluctuations in the data and they should approximate the minimum of the expected cost. This goal is paralleled by the question in computational learning of how well learning algorithms can generalize when we adopt empirical risk minimization as inference technique rather than optimizing the (unknown) expected risk.

Graph cuts and related optimization problems with a combinatorial flavor are characterized by solution spaces that grow exponentially with the number of entities. Therefore, we cannot expect to identify a unique best solution that minimizes the expected risk even in the asymptotic limit. Jaynes (1982) advocated entropy maximizing methods adopted from statistical physics that identify a "stable" set of sufficiently good solutions rather than an "unstable" unique empirical minimizer. Algorithms that sample from the Gibbs distribution maximize the entropy while keeping the average costs of the solution set constant. The Gibbs distribution can be considered as a smoothing measure with a temperature controled width parameter that effectively reduces the resolution of the solution space. Kirkpatrick et al. (1983) prescribed the randomized algorithm *Simulated Annealing* to systematically search for such robust, but still informative sets by exploiting the analogy to annealing solids in material science. We can interpret the Gibbs distribution as posterior probabilities that ranks solutions given the data. A mathematically rigorous treatment of the statistical physics approach to optimization has been pioneered by Talagrand (2003), while Mézard and Montanari (2009) emphasize the connections to computation.

1.2. Boltzmann posteriors for optimal solution validation

One straightforward inspiration which has been imported from statistical physics into learning theory is the usage of Gibbs distributions, also referred to as *Boltzmann posteriors.*

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 solution space $\mathcal{C}$ and X is a random data instance. Then Boltzmann posterior $p_\beta(c|X)$ is a distribution of the form*

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

$Z(\beta, X)$ is known as partition function. Note that the Boltzmann posterior maps a pair (X, β) to a distribution over $\mathcal{C}$. The behavior of this distribution is quite natural: for any β it assigns the highest weights to those solutions, which obtain the smallest costs under the given data instance X. Cost differences are measured relative to the cost scale $1/\beta$. Parameter β of the Boltzmann posterior plays the role of *inverse temperature*, since it controls the level of concentration of $p_\beta(c|X)$ around minimal solutions; thus it rules the variability.

How should we choose β? Too small values of β result in a spread out posterior and we ignore significant cost differences between solutions. Too large β values cause instability of the Gibbs distribution due to fluctuations in the instance. The balance between these two limits of under- and overfitting can be determined by information theoretic considerations (Buhmann, 2013). We will briefly motivate the selection criterion by the following *two-instance* setting.

Let us suppose that we are given *two* instances X' and X''. What is the way in which we could measure how good is a particular choice of β? 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, *i.e.*, Buhmann (2010) introduced the *empirical similarity kernel* for two instances:

Definition 2 *The empirical similarity kernel for two instances X', X'' is given by the function:*

$$\widehat{k}_\beta(X', X'') = \sum_{c \in \mathcal{C}} p_\beta(c|X')p_\beta(c|X'') = \frac{\sum_{c \in \mathcal{C}} \exp(-\beta(R(c, X') + R(c, X'')))}{Z(\beta, X')Z(\beta, X'')} \in [0, 1] \tag{2}$$

The identifiability of solutions given the noisy instances is determined by a cost function specific capacity that plays the role of the mutual information in information theory:

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

$$I := \sup_\beta \mathbb{E}_X \log \Big(\max\{|\mathcal{C}| \, \widehat{k}_\beta(X', X''), 1\} \Big) \,. \tag{3}$$

The expectation is taken w.r.t. the two instances X', X''. The generalization capacity I comes close to the maximum value of $\log |\mathcal{C}|$ when $\widehat{k}_\beta(X', X'') \approx 1$, which means that both posteriors concentrate on the same solution $c^\star$. In this case the posterior is maximally informative of the solution space and completely insensitive to the data fluctuations. I vanishes in the opposite case if the posteriors are concentrated on different sets of solutions without overlap.

To validate models, that is, to select cost functions according to the generalization capacity we have to evaluate expectation values of log partition functions, *i.e.*, $\mathbb{E}_{X'} \log Z_\beta(X')$, $\mathbb{E}_{X''} \log Z_\beta(X'')$ and $\mathbb{E}_{X',X''} \log \sum_{c \in \mathcal{C}} \exp\big(-\beta(R(c, X') + R(c, X''))\big)$. The first two terms are identical for i.i.d. instances. This mathematical challenge is addressed in the remainder of the paper. More precisely, we will be interested in a version of *Helmholtz free energy*, which we define as follows (note a scaling which differs from the conventional definition):

Definition 4 *We will call the quantity*

$$\mathcal{F}(\beta) = \frac{\mathbb{E}_X[\log Z(\beta, X)]}{\log |\mathcal{C}|}. \tag{4}$$

free energy of the set of solutions (configurations) $\mathcal{C}$.

The rest of the paper is devoted to investigating the asymptotics of the free energy in the temperature regime $\beta \to 0$.

2. Notation, Formal Setting and Contribution

2.1. Notation and setting

We consider a class of stochastic optimization problems that can be formulated as follows: let n be an integer (*e.g.*, number of vertices in a graph, size of a matrix, number of keys in a digital tree, etc.), and $\mathcal{S}_n$ a set of objects (*e.g.*, set of edges, elements of a matrix, keys, etc). The data X denote a set of random variables which enter into the definition of an instance (*e.g.*, weights of edges in a weighted graph), which will be clear from the coming passage.

Define $\mathcal{C}_n$ as a set of all feasible solutions (*e.g.* cuts of a graph), and $\mathcal{S}_n(c) \subseteq \mathcal{S}_n$, $c \in \mathcal{C}_n$, as a set of objects belonging to the feasible solution c (*e.g.*, set of edges belonging to a cut), and $w_i(X) = W_i$, $i \in \mathcal{S}_n$, is the weight assigned to the i-th object. For the considered optimization problems, the cost function and optimization task are defined as follows:

$$R(c, X) = \sum_{i \in \mathcal{S}_n(c)} w_i(X) \quad \text{and} \quad c_{\text{opt}}(X) = \arg\min_{c \in \mathcal{C}_n} R(c, X). \tag{5}$$

We also define 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).

Goal. As explained in the introduction, free energy inspired by statistical physics is crucial in assessing robustness and validating solutions to optimization problems. Yet free energy is quite challenging to compute and often intractable in many optimization problems. In this paper, for a class of optimization problems, we aim at better understanding asymptotic behavior of the free energy rate at high temperature (β small). To accomplish it, we need to estimate the expectation of the logarithm of the partition function: $\mathbb{E}_X[\log Z(\beta, X)]$ for $\beta \to 0$. For convenience, we are going to address a scaled version of this quantity, namely free energy (4).

Remark. In the following, we will omit X as an argument of $Z(\beta, X)$ and $R(c, X)$ for the sake of simplicity. (The expectation $\mathbb{E}[.]$, the variance $\text{Var}[.]$ and other probabilistic operations are still meant to be taken with respect to the randomness of X, if otherwise not explicitly stated).

2.2. Contribution

In this paper we focus on two optimization problems in which $\log m = o(N)$. This requires to re-scale β so that it is small and decays as $\widehat{\beta}\sqrt{\log m / N}$ for some constant $\widehat{\beta}$.

We present two types of results: theoretical and experimental. First, in Theorem 5 we establish a fairly tight upper bound on the free energy. Interestingly, we prove that there is a phase transition in the second-order term of the free energy: it grows first quadratically with $\widehat{\beta}$ up to a threshold value, and then linearly. This is in fact confirmed in our experimental results for the bisection problem and the quadratic assignment problem.

Our experiments show a good coincidence with the upper bound for the quadratic part of the free energy, but differ by a small multiplicative constant factor for the linear part. We experimentally conjecture the form of this correcting constant.

To further improve our results, we also propose another derivation of the free energy based on a Taylor expansion. Finally, we conjecture a matching lower bound for the free energy. This is a very challenging problem due to some strong dependencies, but we outline an approach to establish it leaving, however, detailed derivations for a future paper.

3. Two Combinatorial Optimization Problems at a Glance

This section describes two example optimization problems that will be used to describe our findings. They fall into the class specified in Sec. 2.1 and encompass a large range of practical applications.

3.1. The minimum bisection problem

Consider a complete undirected weighted graph $G = (V, E, X)$ of n vertices, where n is an even number. X represents the weights $(W_i)_{i\in E}$ of the edges of the graph. It is the actual "data" contained in the instance of the minimum bisection problem.

A bisection is a balanced partition $c = (U_1, U_2)$ of the vertices in two disjoint sets: $U_1, U_2 \subset V$, $U_1 \sqcup U_2 = V$, $|U_1| = |U_2| = \frac{n}{2}$. Later we also deal with a *sparse* minimum bisection problem in which the disjoint subsets are of the size $|U_1| = |U_2| = \Theta(\log^2 n)$.

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

$$R(c) = \sum_{i\in\mathcal{S}_n(c)} W_i. \tag{6}$$

The minimum bisection problem consists in finding the bisection of the graph with minimum cost.

A simple study shows that $|\mathcal{C}_n| = m = \binom{n}{n/2}$ and $|\mathcal{S}_n(c)| = N = \frac{n^2}{4}$.

$$\log m = \log \binom{n}{n/2} \sim \log\left(2^n\sqrt{\frac{2}{\pi n}}\right) = n\log 2 - \frac{1}{2}\log n + O(1), \tag{7}$$

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

3.2. The quadratic assignment problem

A more complicated example of a problem could be brought as follows. We consider two $n \times n$ 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 objective function is then

$$R(\pi) = \sum_{i,j=1}^{n} V_{ij} \cdot H_{\pi(i),\pi(j)}, \quad \pi \in \mathbf{S}_n. \tag{8}$$

In our terms, the object space is the set of products of entries of V and H constrained by a relation on the indices: $\mathcal{S}_n = \{V_{ij} \cdot H_{\pi(i),\pi(j)} \mid 1 \le i, j \le n; \pi \in \mathbf{S}_n\}$. Thus using the notations of our framework, $N = |\mathcal{S}_n(\pi)| = n^2$ and $m = |\mathcal{C}_n| = n!$ and thus $\log m \sim n \log n = o(N)$ is fulfilled.

4. Upper Bounds for Free Energy

In this section, we present two upper bounds on the second-order term of the free energy rate. The first upper bound applies to the whole class of optimization problems that we consider in this paper. The second one, based on the Sherrington-Kirkpatrick model, is tighter but only applies in the case of the minimum bisection problem. It serves the purpose of showing that the general upper bound is too loose.

4.1. A general upper bound on the free energy rate

The following theorem was stated as a part of (Buhmann et al., 2014, Theorem 1), however, we state it here, because the statement and the proof in (Buhmann et al., 2014) contained certain inprecise points. A complete proof can be found in Appendix A.

Theorem 5 *Consider a class of combinatorial optimization problems in which the cardinality of feasible solutions m and the size N of a feasible solution are related as $\log m = o(N)$. Assume that weights W_i are identically distributed with mean μ and variance σ^2 and that the moment generating function of negative centered weights $(-\overline{W}_i)$ is finite, i.e. $\mathbb{E}[\exp(-t\overline{W}_i)] < \infty$ exists for some $t > 0$. Further assume that within a given solution, the weights are mutually independent, i.e.*

$$\forall c \in \mathcal{C}_n,\ \textit{the set}\ \{W_i \mid i \in \mathcal{S}_n(c)\}\ \ \textit{is a set of mutually independent variables.} \tag{9}$$

Define a scaling $\beta = \widehat{\beta}\sqrt{\log m/N}$, where $\widehat{\beta}$ is a constant. Then the function

$$\widehat{\mathcal{F}}(\beta) := \frac{\mathbb{E}[\log Z(\beta)] + \widehat{\beta}\mu\sqrt{N\log m}}{\log m} = \mathcal{F}(\beta) + \frac{\widehat{\beta}\mu\sqrt{N\log m}}{\log m} \tag{10}$$

satisfies

$$\lim_{n\to\infty} \widehat{\mathcal{F}}(\beta) \le \begin{cases} 1 + \frac{\widehat{\beta}^2\sigma^2}{2}, & \widehat{\beta} < \frac{\sqrt{2}}{\sigma}, \\ \widehat{\beta}\sigma\sqrt{2}, & \widehat{\beta} \ge \frac{\sqrt{2}}{\sigma}. \end{cases} \tag{11}$$

The above theorem shows an interesting phase transition in the second-order term of the free energy rate. For small values of $\widehat{\beta}$, this term grows quadratically up to a threshold value and then linearly. We will verify this surprising phenomenon experimentally on two optimization problems, that is the minimum bisection problem and the quadratic assignment problem.

4.2. A tighter upper bound in the case of the minimum bisection problem

The general upper bound proven above is unfortunately not tight. The following theorem gives a tighter upper bound for the minimum bisection problem on the left side of the phase transition.

Theorem 6 *Consider the minimum bisection problem with n vertices. 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. Assume that edge weights W_i are i.i.d. with mean μ and variance σ^2 and that the moment generating function of centered weights $(-\overline{W}_i)$ is finite, i.e. $\mathbb{E}[\exp(-\beta\overline{W}_i)] < \infty$ exists for some $t > 0$.*

Define a scaling $\beta = \widehat{\beta}\sqrt{\log m/N}$. Then the function $\widehat{\mathcal{F}}(\beta)$ satisfies for $\widehat{\beta} \le \frac{1}{\sqrt{\log 2}\sigma}$

$$\lim_{n\to\infty} \widehat{\mathcal{F}}(\beta) \le 1 + \frac{\widehat{\beta}^2\sigma^2}{4} \tag{12}$$

Sketch of proof. 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 $\mathcal{C}_n^{\text{MBP}} \subset \mathcal{C}_n^{\text{SK}}$. Hence

$$Z^{\text{MBP}}(\beta) = \sum_{c \in \mathcal{C}_n^{\text{MBP}}} \mathrm{e}^{-\beta R(c,X)} \leq \sum_{c \in \mathcal{C}_n^{\text{SK}}} \mathrm{e}^{-\beta R(c,X)} = Z^{\text{SK}}(\beta), \tag{13}$$

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).

A complete proof can be found in Appendix B.

4.3. A Taylor expansion approach to free energy asymptotics

We also propose a new approach to establishing $\mathbb{E} \log Z(\beta)$ using a Taylor expansion of the logarithm function. Namely, by expanding $\log Z(\beta)$ in the Taylor series around $\mathbb{E}[Z(\beta)]$ we have

$$\begin{aligned}\log Z(\beta) = \log \mathbb{E}[Z(\beta)] &+ \frac{Z(\beta) - \mathbb{E}[Z(\beta)]}{\mathbb{E}[Z(\beta)]} - \frac{1}{2}\frac{(Z(\beta) - \mathbb{E}[Z(\beta)])^2}{(\mathbb{E}[Z(\beta)])^2} \\ &+ \sum_{k=3}^{\infty} \frac{(-1)^{k+1}}{k!} \frac{(Z(\beta) - \mathbb{E}[Z(\beta)])^k}{(\mathbb{E}[Z(\beta)])^k}.\end{aligned} \tag{14}$$

Now we take the expectation and obtain

$$\mathbb{E}[\log Z(\beta)] = \log \mathbb{E}[Z(\beta)] - \frac{1}{2}\frac{\operatorname{Var}[Z(\beta)]}{(\mathbb{E}[Z(\beta)])^2} + \sum_{k=3}^{\infty} \frac{(-1)^{k+1}}{k!} \frac{\mathbb{E}\big[(Z(\beta) - \mathbb{E}[Z(\beta)])^k\big]}{(\mathbb{E}[Z(\beta)])^k}. \tag{15}$$

From (Buhmann et al., 2014, Eq. (54)) we know (see also the proof in Appendix C) that

$$\operatorname{Var}[Z(\beta)] = (\mathbb{E}[Z(\beta)])^2 \left(\mathbb{E}_D\left(\frac{G(2\beta)}{(G(\beta))^2}\right)^D - 1\right) = (\mathbb{E}[Z(\beta)])^2(\sigma^2\beta^2\mathbb{E}_D[D] + O(\beta^3)), \tag{16}$$

where $G(\beta) = \mathbb{E}[\exp(-\beta W_i)]$ is a moment generating function of $(-W_i)$, D is a number of objects from $\mathcal{S}_n$ shared by two solutions $c, c' \in \mathcal{C}_n$, chosen *uniformly at random*, *i.e.* $D = |\mathcal{S}_n(c) \cap \mathcal{S}_n(c')|$, and $\mathbb{E}_D$ denotes the expectation w.r.t. this probability space.

Combining (15) and (16) yields the following expansion of $\mathbb{E}[\log Z(\beta)]$:

$$\begin{aligned}\widehat{\mathcal{F}}(\beta) = \frac{\log \mathbb{E}[Z(\beta)] + \widehat{\beta}\mu\sqrt{N \log m}}{\log m} &- \frac{1}{2}\widehat{\beta}^2\sigma^2\frac{\mathbb{E}_D[D]}{N} \\ &+ \frac{1}{\log m}\sum_{k=3}^{\infty} \frac{(-1)^{k+1}}{k!} \frac{\mathbb{E}\big[(Z(\beta) - \mathbb{E}[Z(\beta)])^k\big]}{(\mathbb{E}[Z(\beta)])^k}\end{aligned} \tag{17}$$

5. Simulations and a Conjecture on Free Energy

This section shows some simulations of the second-order term of the free energy rate in the case of the minimum bisection and the quadratic assignment problems. The plots indicate that the quadratic part of the free energy coincides with the upper bound derived in Theorem 5 while the linear part differs by a small multiplicative constant. This allows us to make a conjecture about the behavior of the free energy rate. Eventually, we provide an intuitive explanation.

5.1. Sampling procedure for estimating the partition function

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

$$\widehat{E}_N = \frac{1}{N}\sum_{i=1}^{N}\phi(X_i)\frac{\mathbb{P}(X_i)}{\mathbb{Q}(X_i)}, \quad \text{since} \quad \mathbb{E}_{\mathbb{Q}}\widehat{E}_N = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathbb{Q}}\phi(X_i)\frac{\mathbb{P}(X_i)}{\mathbb{Q}(X_i)} = \mathbb{E}_{\mathbb{P}}\phi(X). \tag{18}$$

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

We adapt it for the computation of Gibbs distribution partition functions. Suppose you have a 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

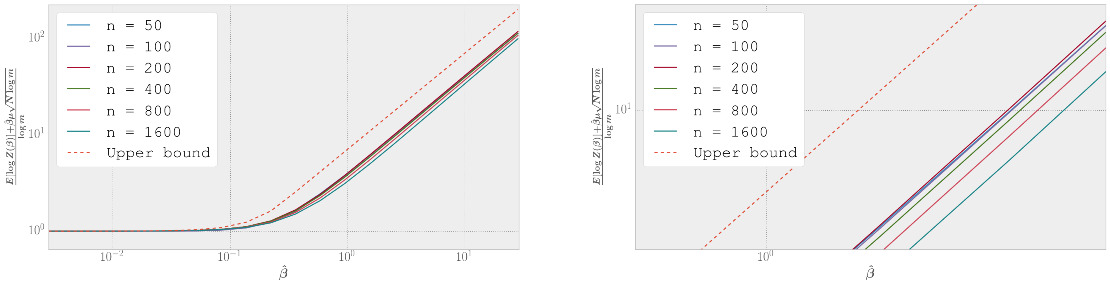
$$Z^*_N(\beta, \beta') = \frac{1}{N}\sum_{i=1}^{N} Z(\beta)\mathrm{e}^{-(\beta'-\beta)R(c_i)} \tag{19}$$

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

$$\mathrm{Var}^{\mathrm{rel}}_{\mathbb{P}(\cdot|\beta)} Z^*_N(\beta, \beta') = \frac{\mathrm{Var}_{\mathbb{P}(\cdot|\beta)} Z^*_N(\beta, \beta')}{\mathbb{E}^2_{\mathbb{P}(\cdot|\beta)} Z^*_N(\beta, \beta')} = \frac{1}{N}\left(\frac{Z(2\beta'-\beta)Z(\beta)}{Z(\beta')^2} - 1\right) \tag{20}$$

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



(a) Full view

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

Figure 1: 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 5.

(19) 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. Simulations of Free Energy Rate

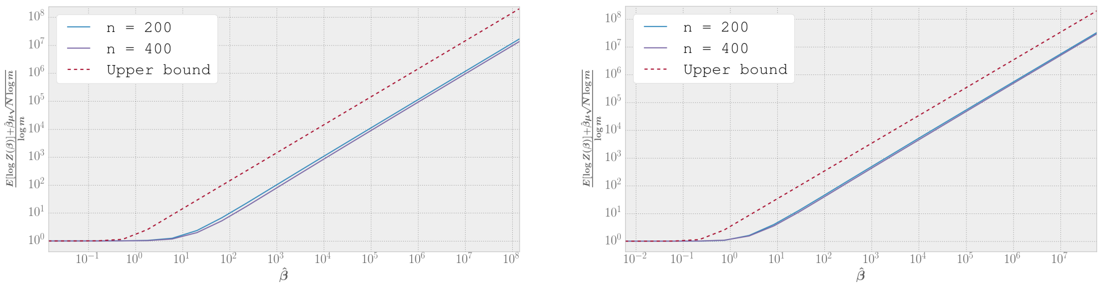
Figure 1 shows the simulation of $\widehat{\mathcal{F}}(\beta)$ in the case of the minimum bisection problem for different graph sizes n. The dashed line corresponds to the upper bound defined in Theorem 5. It appears that the general behavior is quite good for the quadratic part of the free energy while for the linear part there is some discrepancy (a multiplicative factor correction is needed).

Figure 2 shows the simulation of $\widehat{\mathcal{F}}(\beta)$ in the case of the quadratic assignment problem for different graph sizes n. The dashed line corresponds to the upper bound defined in Theorem 5. 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. A conjecture

Based on our empirical results presented in Figures 1 and 2, we are able to conjecture a more precise behavior of the free energy. We shall introduce a correction coefficient α whose value we will determine in the sequel.

Conjecture 7 *Consider a class of combinatorial optimization problems in which the cardinality of feasible solutions m and the size N of a feasible solution are related as $\log m = o(N)$. Assume that weights W_i are identically distributed with mean μ and variance σ^2 and that the moment generating function of negative centered weights $(-\overline{W}_i)$ is finite,*



(a) Standard deviation of $\sigma = 1$ (b) Standard deviation of $\sigma = 2.4$

Figure 2: 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 5.

i.e. $\mathbb{E}[\exp(-t\overline{W}_i)] < \infty$ *exists for some* $t > 0$. *Further assume that within a given solution, the weights are mutually independent, i.e.*

$$\forall c \in \mathcal{C}_n, \text{ the set } \{W_i \mid i \in \mathcal{S}_n(c)\} \text{ is a set of mutually independent variables.} \tag{21}$$

Define a scaling

$$\beta = \widehat{\beta}\sqrt{\log m / N}, \tag{22}$$

where $\widehat{\beta}$ *is a constant. Then the free energy satisfies*

$$\lim_{n\to\infty} \widehat{\mathcal{F}}(\beta) = \begin{cases} 1 + \alpha^2 \frac{\widehat{\beta}^2\sigma^2}{2}, & \widehat{\beta} < \frac{\sqrt{2}}{\alpha\sigma} \\ \alpha\widehat{\beta}\sigma\sqrt{2}, & \widehat{\beta} \geq \frac{\sqrt{2}}{\alpha\sigma} \end{cases} \tag{23}$$

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

$$\alpha = \sqrt{\frac{\mathbb{E}_X \mathrm{Var}_c R(c, X)}{\mathbb{E}_c \mathrm{Var}_X R(c, X)}} = \sqrt{\frac{\mathbb{E}_X \mathrm{Var}_c R(c, X)}{N\sigma^2}} \tag{24}$$

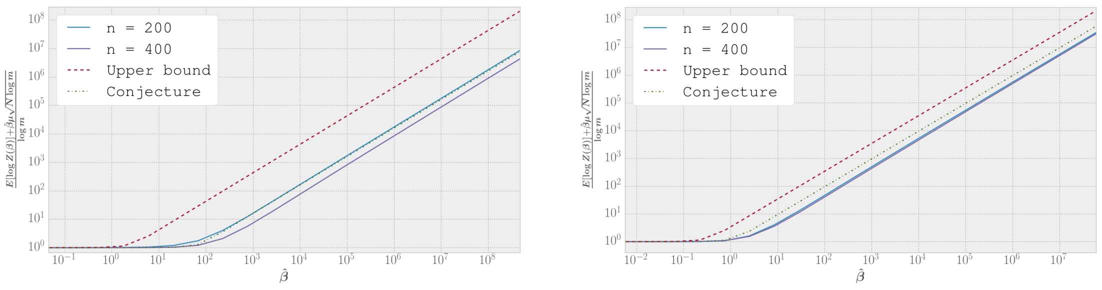
where the expectation $\mathbb{E}_c[\cdot]$ is taken w.r.t. to all feasible solutions selected uniformly.

Let us observe that

$$\mathbb{E}_X \mathrm{Var}_c R(c, X) = \sum_{i,j\in\mathcal{S}_n} \mathbb{E}_X[W_i W_j]\big(\mathbb{P}_c(i, j \in \mathcal{S}_n(c)) - \mathbb{P}_c(i \in \mathcal{S}_n(c)) \cdot \mathbb{P}_c(j \in \mathcal{S}_n(c))\big). \tag{25}$$

When $i \neq j$ implies W_i and W_j are independent (which is true for the minimum bisection problem, wrong for the quadratic assignment problem), then

$$\mathbb{E}_X \mathrm{Var}_c R(c, X) = \sigma^2 \sum_{i\in\mathcal{S}_n} \mathbb{P}_c(i \in \mathcal{S}_n(c))\big(1 - \mathbb{P}_c(i \in \mathcal{S}_n(c))\big). \tag{26}$$



(a) $\sigma = 0.3$

(b) $\sigma = 2.4$

Figure 3: 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 $\sigma_V^2 = \sigma_H^2$.

In particular, for the minimum bisection problem we have

$$\mathbb{E}_X \mathrm{Var}_c R(c, X) = \frac{n^2\sigma^2}{4}\left(1 - \frac{n}{2(n-1)}\right), \tag{27}$$

and then from (24) we find $\alpha = {}^1\!/\!\sqrt{2}$ in the limit. This is quite close to the observed correction ${}^1\!/1.64$. It corresponds *exactly* to the coefficient in the Sherrington-Kirkpatrick model (see equation (12)). For the quadratic assignment problem, denote by V the weight matrix and by H the distance matrix. Assume that elements of V are distributed under a distribution of $\mathcal{N}(\mu_V, \sigma_V^2)$, and those of H under a distribution of $\mathcal{N}(\mu_H, \sigma_H^2)$. We have then $\sigma^2 = \sigma_V^2\sigma_H^2 + \mu_H^2\sigma_V^2 + \mu_V^2\sigma_H^2$. Then

$$\mathbb{E}_X \mathrm{Var}_c R(c, X) = (n^2 - 2)\sigma_V^2\sigma_H^2 \tag{28}$$

and we conclude that $\alpha = \frac{\sigma_V\sigma_H}{\sigma}$. This is illustrated in Figure 3.

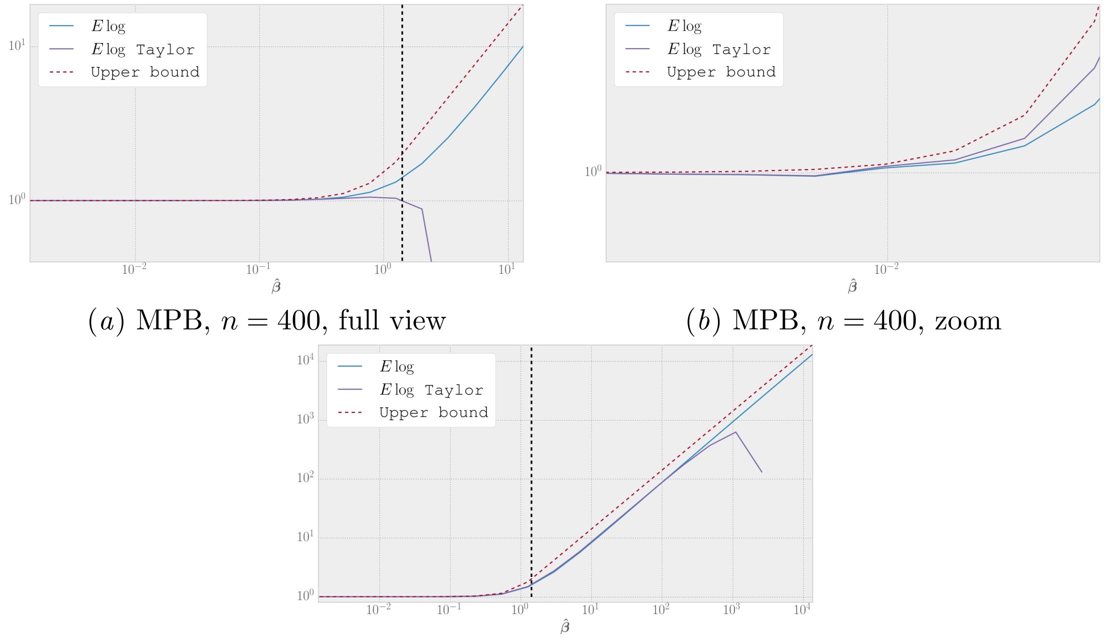
5.4. Simulations of the Taylor expansions

Recall that in Sec. 4.3 we discuss an approach to computing $\widehat{\mathcal{F}}(\beta)$ via the Taylor expansion, namely equation (17). In Figure 4 we present simulation results and compare them to our theoretical results obtained through the Taylor expansion. The curves labelled "E log" represent direct simulations of $\widehat{\mathcal{F}}(\beta)$. Those labelled "$E$ log Taylor" are constructed as the simulation of $\frac{\log \mathbb{E}[Z(\beta)] + \widehat{\beta}\mu\sqrt{N\log m}}{\log m}$ with the correction of $\frac{1}{2}\widehat{\beta}^2\sigma^2\frac{\mathbb{E}_D[D]}{N}$ as proposed in equation (17).

For the minimum bisection problem (Fig. 4(a) and 4(b)), a graph of 400 vertices is used. In Fig. 4(c), we also show a plot for the sparse bisection problem (refer to section 3.1). 3200 vertices are used in this case. For all plots, the edge weights are generated from a normal distribution $\mathcal{N}(0, 1)$.

6. Discussion and Future Work

To obtain precise asymptotics of the free energy, we need to find the matching lower bounds, which turns out to be a difficult task. We will consider the problem setting analogous to



(a) MPB, $n = 400$, full view

(b) MPB, $n = 400$, zoom

(c) Sparse MPB, $n = 3200$

Figure 4: Taylor expansion of $\widehat{\mathcal{F}}(\beta)$ for low values of β around $\frac{\log \mathbb{E}[Z(\beta)]+\widehat{\beta}\mu\sqrt{N\log m}}{\log m}$ for different problems.

that of Sec. 4.1, however with an additional assumption imposed. We present in Appendix D a sketch for the proof of the following result.

Lemma 8 *Consider the setting and requirements of Theorem 5. Additionally we will require that the solution costs are "weakly dependent" in the following sense: for some* $u_n = \Theta(\sqrt{N/\log m})$, *define for* $\bar{R}(c) = -\sum_{i\in\mathcal{S}(c)} \overline{W}_i$ *(as elsewhere) the probability* $a_n := \mathbb{P}(\bar{R}(c) \geq u_n)$ *of exceeding this threshold and assume that*

$$\sum_{c\neq c'\in\mathcal{C}_n} \operatorname{Cov}\big(\mathbb{1}_{\{\bar{R}(c)\geq u_n\}}, \mathbb{1}_{\{\bar{R}(c')\geq u_n\}}\big) = o(m^2 a_n^2). \tag{29}$$

Then the function $\widehat{\mathcal{F}}(\beta)$ *satisfies*

$$\lim_{n\to\infty} \widehat{\mathcal{F}}(\beta) \geq \begin{cases} 1 + \frac{\widehat{\beta}^2\sigma^2}{2}, & \widehat{\beta} < \frac{\sqrt{2}}{\sigma}, \\ \widehat{\beta}\sigma\sqrt{2}, & \widehat{\beta} \geq \frac{\sqrt{2}}{\sigma}. \end{cases} \tag{30}$$

However, the way of proving this lemma indicates that we might need to search for another, stronger technique in order to make the additions assumption less restrictive.

Another direction of research consists in further investigating the Taylor expansion approach and the scope of its applicability to various regimes of β.

References

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

J. M. Buhmann. SIMBAD: emergence of pattern similarity. In *Similarity-Based Pattern Analysis and Recognition*, pages 45–64. Springer Berlin / Heidelberg, 2013.

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

W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 2. Wiley, NY, 2nd edition, 1971.

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

E. T. Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70: 939–952, 1982.

S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

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

D. Sherrington and S Kirkpatrick. Solvable model of a spin glass. *J. Physique Lett.*, 45: 1145–1153, 1975.

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

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

Appendix A. Proof of Theorem 5 (Upper Bound on the Free Energy Rate in General Case)

Proof of Theorem 5. We need to compute $\mathbb{E}[\log Z(\beta)]$. Remember that it can be upper bounded by Jensen's inequality as

$$\mathbb{E}[\log Z(\beta)] \le \log \mathbb{E}[Z(\beta)]. \tag{31}$$

To simplify our analysis, we actually shall investigate the centralized weights $\overline{W} := W - \mu$ and denote by $\widehat{G}(\beta)$, where $\beta > 0$, the moment generating function of $(-\overline{W})$, that is

$$\widehat{G}(\beta) = \mathbb{E}_X[\exp(\beta(-\overline{W}))] < \infty. \tag{32}$$

In order to evaluate $\mathbb{E}[Z(\beta)]$, we proceed as follows

$$\begin{aligned}\mathbb{E}[Z(\beta)] = \mathbb{E}\Big[\sum_{c\in\mathcal{C}}\exp(-\beta R(c))\Big] &= \exp(-\beta N\mu)\mathbb{E}\Big[\sum_{c\in\mathcal{C}}\exp\big(-\beta(R(c)-N\mu)\big)\Big] \\ &= \exp(-\beta N\mu) m \widehat{G}^N(\beta). \qquad (33)\end{aligned}$$

since the random variables W_i are i.i.d. within a given solution c. Thus

$$\log\mathbb{E}[Z(\beta)] = -\beta N\mu + \log m + N\log\widehat{G}(\beta) \qquad (34)$$

From the above relation (34) one can see that in order to get a nontrivial limit of $\frac{\log\mathbb{E}[Z(\beta)]}{\log m}$ we need to choose the limit $\beta \to 0$. Under this assumption, we can expand $\widehat{G}(\beta)$ in the Taylor series to obtain

$$\widehat{G}(\beta) = 1 + \frac{1}{2}\beta^2\sigma^2 + O(\beta^3). \qquad (35)$$

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

$$\begin{aligned}\log\mathbb{E}[Z(\beta)] &= -\beta N\mu + \log m + N\log\widehat{G}(\beta) \\ &= -\beta N\mu + \log m + N\log\Big(1 + \frac{1}{2}\beta^2\sigma^2 + O(\beta^3)\Big) \\ &= -\beta N\mu + \log m + \frac{1}{2}N\beta^2\sigma^2(1 + O(\beta)). \qquad (36)\end{aligned}$$

This suggests that the right choice for β is

$$\beta = \widehat{\beta}\sqrt{\frac{\log m}{N}} \qquad (37)$$

for some constant $\widehat{\beta}$. Thus we arrive at

$$\frac{\log\mathbb{E}[Z(\beta)] + \beta N\mu}{\log m} = 1 + \frac{1}{2}\widehat{\beta}^2\sigma^2(1 + O(\beta)). \qquad (38)$$

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

$$\frac{\mathbb{E}[\log Z(\beta)] + \widehat{\beta}\mu\sqrt{N\log m}}{\log m} \le 1 + \frac{1}{2}\widehat{\beta}^2\sigma^2\left(1 + O\left(\sqrt{\frac{\log m}{N}}\right)\right). \qquad (39)$$

Now make this general bound tighter for ceratin region of $\widehat{\beta}$. Let us denote

$$\phi(\beta) = \mathbb{E}[\log Z(\beta)] + \beta N\mu =: \mathbb{E}[\log\widehat{Z}(\beta)] \qquad (40)$$

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

$$\beta\max_{c\in\mathcal{C}}\overline{R}(c) \le \log\widehat{Z}(\beta). \qquad (41)$$

Using the upper bound obtained in (39) we find

$$\frac{\mathbb{E}[\max_{c\in\mathcal{C}}\overline{R}(c)]}{\log m} \le \sqrt{\frac{N}{\log m}}\left(\widehat{\beta}^{-1} + \frac{1}{2}\widehat{\beta}\sigma^2\right). \qquad (42)$$

Choosing $\widehat{\beta}^* = \sqrt{2}/\sigma$ that minimizes the right-hand side of (42) we arrive at

$$\mathbb{E}[\max_{c\in\mathcal{C}} \overline{R}(c)] \le \sqrt{2\sigma^2 N \log m} \tag{43}$$

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

$$\phi'(\beta) \le \mathbb{E}[\max_{c\in\mathcal{C}} \overline{R}(c)]. \tag{44}$$

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

$$\phi(\beta) \le \phi(\beta^*) + \mathbb{E}[\max_{c\in\mathcal{C}} \overline{R}(c)](\beta - \beta^*), \tag{45}$$

Replacing all the terms eventually yields

$$\mathbb{E}[\log \widehat{Z}(\beta)] \le \widehat{\beta}\sigma\sqrt{2}\log m \tag{46}$$

and the second upper bound in Theorem 5. □

Appendix B. Proof of Theorem 6 (Upper Bound on the Free Energy Rate for the Minimum Bisection Problem)

Proof of Theorem 6. 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.

The problem consists in finding $c \in \{-1, 1\}^n$ so that $\sum_i c_i = 0$ (balance condition) and the sum of the weights of cut edges

$$R(c, X) = \sum_{\substack{c_i = -c_j \\ i<j}} g_{ij} \tag{47}$$

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

Define the partition function as

$$Z(\beta, X) = \sum_{c\in\mathcal{C}_n} \mathrm{e}^{-\beta R(c,X)} \tag{48}$$

where $\mathcal{C}_n = \{c \in \{-1, 1\}^n | \sum_i c_i = 0\}$ is the solution space. Let us now prove Theorem 6. Observe that

$$\begin{aligned}\log Z(\beta, X) + \widehat{\beta}\mu\sqrt{N \log m} &= \log Z(\beta, X) + \beta\mu N \\ &= \log \sum_{c\in\mathcal{C}_n} \mathrm{e}^{-\beta(R(c,X) - N\mu)} \\ &= \log Z(\beta, \overline{X})\end{aligned} \tag{49}$$

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:

$$Z^*(\beta) = \sum_{c \in \mathcal{C}_n^*} \mathrm{e}^{-\beta R(c)} \tag{50}$$

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

$$Z(\beta) \leq Z^*(\beta) \tag{51}$$

Now rewrite the cost function as

$$R(c) = \sum_{\substack{c_i = -c_j \\ i<j}} g_{ij} = \frac{1}{2}\Big(\sum_{i<j} g_{ij} - \sum_{i<j} c_i c_j g_{ij}\Big) = \frac{1}{2}\Big(\sum_{i<j} g_{ij} + \sqrt{n} R^{\mathrm{SK}}(c)\Big) \tag{52}$$

where

$$R^{\mathrm{SK}}(c) = -\frac{1}{\sqrt{n}} \sum_{i<j} c_i c_j g_{ij} \tag{53}$$

is the cost function of the Sherrington-Kirkpatrick model.

This entails

$$Z^*(\beta) = \mathrm{e}^{-\frac{\beta}{2}\sum_{i<j} g_{ij}} \sum_{c \in \mathcal{C}_n^*} \mathrm{e}^{-\frac{\sqrt{n}\beta}{2} R^{\mathrm{SK}}(c)} = \mathrm{e}^{-\frac{\beta}{2}\sum_{i<j} g_{ij}} Z^{\mathrm{SK}}\Big(\frac{\sqrt{n}\beta}{2}\Big) \tag{54}$$

where

$$Z^{\mathrm{SK}}(\beta) = \sum_{c \in \mathcal{C}_n^*} \mathrm{e}^{-\beta R^{\mathrm{SK}}(c)} \tag{55}$$

is the partition function associated with the Sherrington-Kirkpatrick model.

Since the g_{ij} are centered, it follows that

$$\mathbb{E} \log Z^*(\beta) = \mathbb{E} \log Z^{\mathrm{SK}}\Big(\frac{\sqrt{n}\beta}{2}\Big) \tag{56}$$

For the following, we will need an external statement from (Talagrand, 2003), which we cite here.

Theorem 9 (Talagrand, 2003, Theorem 2.2.1) *If $\beta < \frac{1}{\sigma}$, we have*

$$\lim_{n \to \infty} \frac{1}{n} \mathbb{E} \log Z^{SK}(\beta) = \frac{\beta^2 \sigma^2}{4} + \log 2 \tag{57}$$

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

$$\sqrt{n}\beta = \sqrt{n}\cdot\widehat{\beta}\sqrt{\frac{\log m}{N}} = \sqrt{n}\cdot\widehat{\beta}\sqrt{\frac{\log\binom{n}{n/2}}{n^2/4}} \sim \sqrt{n}\cdot\widehat{\beta}\sqrt{\frac{n\log 2}{n^2/4}} = 2\sqrt{\log 2}\widehat{\beta} \tag{58}$$

Thus, we can use Theorem 9 to obtain, for $\widehat{\beta} < \frac{1}{\sqrt{\log 2}\sigma}$

$$\lim_{n\to\infty}\frac{1}{n}\mathbb{E}\log Z^{\mathrm{SK}}\Big(\frac{\sqrt{n}\beta}{2}\Big) = \Big(\frac{\widehat{\beta}^2\sigma^2}{4}+1\Big)\log 2. \tag{59}$$

The equivalence $\frac{\log m}{n} \sim \log 2$ (in $n\to\infty$) and (56) both allow to write

$$\lim_{n\to\infty}\frac{\mathbb{E}\log Z^*(\beta)}{\log m} = \frac{\widehat{\beta}^2\sigma^2}{4}+1 \tag{60}$$

for $\widehat{\beta} < \frac{1}{\sqrt{\log 2}\sigma}$. Now (51) implies that

$$\lim_{n\to\infty}\frac{\mathbb{E}\log Z(\beta)}{\log m} \le \frac{\widehat{\beta}^2\sigma^2}{4}+1 \tag{61}$$

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

Appendix C. Regarding the Variance of Partition Function

In this section we give a clarification of (16) which shows the asymptotics of $\mathrm{Var}[Z(\beta,X)]$:

$$\mathrm{Var}[Z(\beta)] = (\mathbb{E}[Z(\beta)])^2\bigg(\mathbb{E}_D\Big(\frac{G(2\beta)}{(G(\beta))^2}\Big)^D - 1\bigg) = (\mathbb{E}[Z(\beta)])^2(\sigma^2\beta^2\mathbb{E}_D[D] + O(\beta^3)).$$

Recall (from Sec. 4.3) that $G(\beta) = \mathbb{E}[\exp(-\beta W)]$, where W is a random variable with expectation μ and variance σ^2, and D is a number of objects from $\mathcal{S}_n$ shared by two solutions $c, c' \in \mathcal{C}_n$, chosen *uniformly at random, i.e.* $D = |\mathcal{S}_n(c)\cap\mathcal{S}_n(c')|$, and $\mathbb{E}_D$ is the expectation w.r.t. this probability space.

Proof of Eq. (16). The Taylor expansion of $G(\beta)$ around 0 is

$$G(\beta) = 1 - \beta\mu + \frac{\beta^2\mathbb{E}[W^2]}{2} + O(\beta^3). \tag{62}$$

Thus,

$$\begin{aligned}\Big(\frac{G(2\beta)}{(G(\beta))^2}\Big)^D &= \Big(\frac{1-2\beta\mu+2\beta^2\mathbb{E}[W^2]+O(\beta^3)}{[1-\beta\mu+\beta^2\mathbb{E}[W^2]/2+O(\beta^3)]^2}\Big)^D\\ &= \Big(\frac{1-2\beta\mu+2\beta^2\mathbb{E}[W^2]+O(\beta^3)}{1-2\beta\mu+(\mu^2+\mathbb{E}[W^2])\beta^2+O(\beta^3)}\Big)^D\\ &= \big(1+(\mathbb{E}[W^2]-\mu^2)\beta^2+O(\beta^3)\big)^D\\ &= 1+D\sigma^2\beta^2+O(\beta^3),\end{aligned} \tag{63}$$

the last transition taking effect under reasonable assumptions on D and β^2.

Taking the expectation yields

$$\mathbb{E}_D\left(\frac{G(2\beta)}{(G(\beta))^2}\right)^D = 1 + \sigma^2\beta^2\mathbb{E}_D[D] + O(\beta^3). \tag{64}$$

From the above, we obtain the β-asymptotics of $\mathrm{Var}[Z(\beta)]$:

$$\mathrm{Var}[Z(\beta)] = (\mathbb{E}[Z(\beta)])^2\left(\mathbb{E}_D\left(\frac{G(2\beta)}{(G(\beta))^2}\right)^D - 1\right) = (\mathbb{E}[Z(\beta)])^2(\sigma^2\beta^2\mathbb{E}_D[D] + O(\beta^3)), \tag{65}$$

which is the required. □

Appendix D. Sketch of proof of Lemma 8 (Lower Bound on the Free Energy Rate under Assumptions)

Before proceeding to the proof we should state that Lemma 8 is a correction of the erroneous proof of the lower bound in (Buhmann et al., 2014, Theorem 1). Thus the proof of Lemma 8 will be almost the same, with slight modifications.

Sketch of proof of Lemma 8. For the following, note that since the weights are i.i.d. inside each solution, then the centered negative cost function $\overline{R}(c) \xrightarrow{d} \mathcal{N}(0, N\sigma^2)$, where $\mathcal{N}$ represents normal distribution. Let Y be the cardinality of the solution subset for which the $\overline{R}(c)$ is large enough:

$$Y := \mathrm{card}\{c\colon \overline{R}(c) \geq u_n\} \quad \text{for some} \quad u_n = \Theta\left(\sqrt{\frac{N}{\log m}}\right). \tag{66}$$

Denote $a_n := \mathbb{P}(\overline{R}(c) \geq u_n)$.

From the properties of centered Gaussian (see, for example, Talagrand, 2003, (A.37, A.38)), which is the limiting distribution of $\overline{R}(c)$, we get the following bound on a_n (small terms correspond to large deviation bounds):

$$(1 + o(1))\frac{\sigma\sqrt{N}}{L_1 u_n}\exp\left(-\frac{u_n^2}{2\sigma^2 N}\right) \leq a_n \leq (1 + o(1))\exp\left(-\frac{u_n^2}{2\sigma^2 N}\right), \tag{67}$$

where L_1 is a certain constant. Together with the choice of u_n that will be made later, this allows us to write that in the limit $(n \to \infty)$ $m a_n \to \infty$ holds true.

Now let A denote an event $\{Y \leq m a_n/2\}$. By Markov inequality (second transition in the following chain)

$$\mathbb{P}(A) \leq \mathbb{P}\big((Y - \mathbb{E}[Y])^2 \geq m^2 a_n^2/4\big) \leq 4\mathrm{Var}[Y]/(m^2 a_n^2) \to 0, \tag{68}$$

where we used the assumption (29) of the lemma, along with a representation of $\mathrm{Var}[Y]$ as a sum of indicator random variables $\mathbb{1}_{\{\overline{R}(c) \geq u_n\}}$.

Next, we derive lower bounds for $\mathbb{E}[\log \widehat{Z}(\beta)]$ on the events A and $\Omega \setminus A$. For the latter, we have:

$$\widehat{Z}(\beta) = \sum_{c \in \mathcal{C}} \exp(\beta\overline{R}(c)) \geq \sum_{c \in \mathcal{C}} \exp(\beta u_n) \geq \frac{m}{2} a_n \exp(\beta u_n), \tag{69}$$

thus

$$\mathbb{E}[\mathbb{1}_{\Omega\setminus A} \cdot \log \widehat{Z}(\beta)] \geq (1-\mathbb{P}(A))(\log m - \log 2 + \log a_n + \beta u_n). \tag{70}$$

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

$$\mathbb{E}[\mathbb{1}_A \cdot \log \widehat{Z}(\beta)] \geq -\beta\mathbb{E}[-\mathbb{1}_A \overline{R}(c)] \geq -\beta\mathbb{E}[|\overline{R}(c)|] \geq -L\sigma\beta\sqrt{N}, \tag{71}$$

where L is some constant coming from expectation of half-normal distribution, which is the thermodynamic limit distribution for $|\overline{R}(c)|$. Here we use the fact that $|\overline{R}(c)|$ 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 $|\overline{R}(c)|$ (Feller, 1971, Ch. XVI.7), the expectation value of $|\overline{R}(c)|$ also converges to the one of half-normal.

Combining (70) and (71), we obtain

$$\mathbb{E}[\log \widehat{Z}(\beta)] \geq (1-\mathbb{P}(A))(\log m - \log 2 + \log a_n + \beta u_n) - L\sigma\beta\sqrt{N}. \tag{72}$$

thus (72) turns into (we also use here bounds on a_n from (67) normalized by $\log m$)

$$\frac{\mathbb{E}[\log \widehat{Z}(\beta)]}{\log m} \geq (1-\mathbb{P}(A))\Big(1-\frac{\log 2}{\log m}+\frac{\log(\sigma\sqrt{N}/(L_1 u_n))}{\log m}-\frac{u_n^2}{2\log m\sigma^2 N}+\frac{\beta u_n}{\log m}\Big)-\frac{L\sigma\beta\sqrt{N}}{\log m} \tag{73}$$

Now for the regime $\beta \leq \widehat{\beta}^* \sqrt{\log m/N}$ we choose $u_n = \widehat{\beta}\sigma^2\sqrt{N\log m}$, which yields a lower bound

$$\frac{\mathbb{E}[\log \widehat{Z}(\beta)]}{\log m} \geq (1-\mathbb{P}(A))\Big(1+\frac{\widehat{\beta}^2\sigma^2}{2}+O\Big(\frac{\log\log m}{\log m}\Big)\Big)+O\Big(\frac{1}{\sqrt{\log m}}\Big), \tag{74}$$

and for the regime $\beta \geq \widehat{\beta}^* \sqrt{\log m/N}$ we choose $u_n = \sqrt{2\sigma^2 N\log m}$, which yields a lower bound

$$\frac{\mathbb{E}[\log \widehat{Z}(\beta)]}{\log m} \geq (1-\mathbb{P}(A))\Big(\widehat{\beta}\sqrt{2}\sigma+O\Big(\frac{\log\log m}{\log m}\Big)\Big)+O\Big(\frac{1}{\log m}\Big). \tag{75}$$

As $\mathbb{P}(A) \to 0$ due to (68) and the additional terms $O(\cdot)$ are small in the limit, so we obtain the requested asymptotical lower bounds. □