Local Landscape of the Sherrington-Kirkpatrick Hamiltonian

Master Thesis

Local Landscape of the Sherrington-Kirkpatrick Hamiltonian

David Opalič

April 2023

Supervised by
Prof. Dr. Afonso Bandeira
Dr. Dmitriy Kunisky

Department of Mathematics, ETH Zürich

Abstract

We look at the behaviour of the Sherrington-Kirkpatrick Hamiltonian as we move around the hypercube. Changes of the Hamiltonian along the edges of a face of the hypercube are encoded by a random variable called *square pattern*. We give a complete description of a distribution of square pattern. We calculate the probability of a vertex being a local optimum and determine the typical value of energy at it. We consider possible generalisations to radius 2 local optima.

Contents

1 Introduction

Throughout the years, physicists proposed many models to describe phenomena in condensed matter physics, with the special focus being on the magnetic disorder. The exploration started with *ferromagnetic models*, perhaps the simplest and most popular being the Ising model (see [13] for a nice introduction). A completely new branch of the field emerged in the 1970s with the introduction of *spin glass models*. An overlaying picture in most of these models is that of a countable weighted graph $G = (V, E)$ with a random variable σ_i at each vertex $i \in V$. A set of edges E together with weights on edges determine the covariance structure of $\sigma = (\sigma_i)_{i \in n}$. Usually, σ takes values in $\Sigma_n = \{\pm 1\}^n$ as in the already mentioned Ising model or as in Sherrington-Kirkpatrick model, where most of our attention will lie. However, there are numerous generalisations to larger spaces Σ, for example, see the Edwards-Anderson model or XY model.

To every realisation of $\sigma = (\sigma_1, \sigma_2, ..., \sigma_n)$, called *configuration*, we associate *Hamiltonian* $H_n(\sigma)$, sometimes also called *energy* of σ due to its usage in physics. Depending on the model it is either a real number or a random variable and it represents how closely aligned coordinates of σ are between themselves with respect to the underlying covariance structure. Next, we define (possibly random) *Gibbs' measure* G_n on Σ_n, given by

$$G_n(\sigma) := \frac{\exp(-\beta H_n(\sigma))}{Z_n},$$

where β is a tuning parameter that in physics represents the inverse temperature, and Z_n is a normalising constant. This definition is not crucial to us and we only mention it here for a general overview of what the field is about and to motivate the calculations in this thesis. Interested reader might want to take a look at [16]. Having a measure on Σ_n, one can start exploring typical configurations at different temperatures. What phenomena can one observe? What happens when we take $n \to \infty$? Are there some critical values of β, where the behaviour of σ changes abruptly?

Spin glass models generalise ferromagnetic spin models in two ways. First, they allow for negative weights on the graph edges. Consequently, sometimes disorder states are preferred. Furthermore, weights on graph edges are random. That makes Hamiltonian random, which means Gibbs' measure is a random measure, further complicating calculations. The name "glass" comes from similarities between the disorder of particles in real glass (e.g. windows) and the preferred chaotic disorder in spin glass. Often, these models are used to describe the behaviour of physical particles and it is desirable that they take into account their spatial positions, e.g. (Euclidean) distances between them. This can be done by choosing a particular type of graph G. However, most such models are, at least at present, intractable, though they are still extensively studied using Monte Carlo methods. To get around this, G is usually assumed to be a clique with edge weights being independent and equally distributed random variables, resulting in a symmetrical framework. Such a family of models is called *mean field spin glass* and it has been extensively studied.

Mean field spin glass models have applications in many fields, among others in physics, chemistry, neural networks and theoretical biology. It is an area where it is feasible to run simulations and get good numerical results, but putting them on rigorous footing requires a lot of hard work. Most of the progress was made by physicists in a heuristic, yet ingenious way. See books [17] and [21] written by Parisi and others for some ideas. Later, mathematicians, led by Michel Talagrand and Francesco Guerra, formalised the field and proved a bunch of important results. Two volume series "Mean Field Models for Spin Glass" by Talagrand ([26] and [27]) gives an exhaustive, rigorous and well-motivated overview of the current state of the field. In particular, the first chapter deals with the Sherrington-Kirkpatrick model.

We stay at the level of Hamiltonian $H_n(\sigma)$, not going into Gibbs' measure G_n. We explore the landscape of $H_n(\sigma)$ and look at optimisation problems connected to it. While motivation comes from physics, our approach is very mathematical. There are many different spin glass models, each having its peculiar characteristics. Some high-level ideas and techniques appear in many of them, though often very model-specific arguments need to be used. As there is no "right" model, the most important is a general picture we get from studying mean field spin glass. Even understanding a single model is a huge task, let alone all of them. We work in the simplest one of them, the Sherrington-Kirpatrick model.

1.1 The Sherrington-Kirkpatrick model

Let $\Sigma_n = \{\pm 1\}^n$ be an n-dimensional hypercube. The Sherrington-Kirkpatrick model has a Hamiltonian $H_n : \Sigma_n \to \mathbb{R}$ defined as

$$H_n(\sigma) \coloneqq \sum_{i<j} \sigma_i \sigma_j g_{ij}, \tag{1}$$

where $(g_{ij})_{i<j}$ are independent standard Gaussians (see Appendix A for a collection of facts about normal random variables we will use). From now on we suppress the dependency on n. Let's look at the example to see how this could model a real-life situation. Assume we have 2 tables available and n people to sit down. Unfortunately, not all of them are friends. We can imagine g_{ij} indicates how fond of each other people i and j are, with $g_{ij} < 0$ meaning they are enemies. Here, we completely unrealistically assumed feelings are always mutual. For a person i, we have $\sigma_i = 1$ if we sit him at table A and $\sigma_i = -1$ if we sit him at table B. The Hamiltonian $H(\sigma)$ describes how well we have done with our sitting σ. Replacing tables with plus and minus electric states and people with particles gives a model of magnetism.

Often, the Hamiltonian in the Sherrington-Kirkpatrick model is defined slightly differently. For example, in [26] we come across

$$H(\sigma) \coloneqq -\frac{1}{\sqrt{n}} \sum_{i<j} \sigma_i \sigma_j g_{ij} + h \sum_i \sigma_i. \tag{2}$$

There are multiple reasons for this. Making the Hamiltonian more complicated results in lighter Gibbs' measure expressions. As these are often the object of interest, many authors choose more convoluted definitions. In addition, equation (2) better captures the physical intuition. The most aligned configurations have the highest Hamiltonian, but in nature, we find the most stable configurations at the lowest energy levels, hence the minus in front. The term $h \sum_i \sigma_i$ represents the external magnetic field acting on spins σ. The square root of n is included for normalisation reasons when considering the normalisation constant Z in G. However, our focus lies on the Hamiltonian H and so we stick to the simpler definition (1).

For a fixed σ, the Hamiltonian $H(\sigma)$ is a simple object. All g_{ij}'s are independent normal random variables and thus $H(\sigma) \sim \mathcal{N}\left(0, \binom{n}{2}\right)$, a centred normal with variance $\binom{n}{2}$. In the words of our little analogy, i being friends with j and j being friends with k, tells us nothing about the relationship between i and k. The situation is more complex if we consider the relations between values of the Hamiltonian at different σ's. Let d be a *Hamming distance* on Σ, that is the proportion of coordinates where σ and σ' differ,

$$d(\sigma, \sigma') = \frac{1}{n} \sum_i \mathbb{1}_{\{\sigma_i \neq \sigma'_i\}}.$$

Since everything is symmetric, the covariance structure is completely determined by distances d.

Lemma 1.1. *Let $\Sigma = \{\pm 1\}^n$. For any $\sigma, \sigma' \in \Sigma$ and d the Hamming distance on Σ, we have*

$$\operatorname{Cov}\left(H(\sigma), H(\sigma')\right) = \frac{n^2}{2}\left(1 - 2d(\sigma, \sigma')\right)^2 - \frac{n}{2}. \tag{3}$$

Proof. For a fixed σ, the hamiltonian $H(\sigma)$ is a centred random variable. Hence,

$$
\begin{aligned}
\mathrm{Cov}\left(H(\sigma), H(\sigma')\right) &= \mathbb{E}\left[H(\sigma)H(\sigma')\right] \\
&= \sum_{i<j} \sigma_i \sigma_j \sigma'_i \sigma'_j \\
&= \frac{n^2}{2}\left(\frac{1}{n}\sum_i \sigma_i \sigma'_i\right)^2 - \frac{n}{2} \\
&= \frac{n^2}{2}\left(1 - 2d(\sigma, \sigma')\right)^2 - \frac{n}{2}.
\end{aligned}
$$

□

As it turns out this is a very rich covariance structure. Consequently, the landscape of $H(\sigma)$, i.e. values the Hamiltonian takes at the edges of the hypercube, is very complicated.

This master thesis is about exploring the local landscape around a vertex $\sigma = (\sigma_1, \sigma_2, ...\sigma_n)$. Let's introduce some notation that describes moving around the hypercube Σ. First, there is a trivial *global spin change* $-\sigma = (-\sigma_1, -\sigma_2, ..., -\sigma_n)$. As σ and $-\sigma$ represent the same partition of $(\sigma_i)_i$ in two sets, it is clear that $H(\sigma) = H(-\sigma)$. We define

$$
\sigma^{(i)} := (\sigma_1, ..., \sigma_{i-1}, -\sigma_i, \sigma_{i+1}, ..., \sigma_n)
$$

to be a neighbour vertex of σ that differs only in coordinate i. To keep the notation lighter we will write $H^{(i)}(\sigma) = H(\sigma^{(i)})$. For a set $I = \{i_1, ..., i_k\}$ we define σ^I and H^I in an analogous way. Note that the order in which we change spins doesn't matter. For example,

$$
(H^{(i)})^{(j)}(\sigma) = (H^{(j)})^{(j)}(\sigma) = H^{(\{i,j\})}(\sigma).
$$

For $0 < |I| < n$, values of $H(\sigma)$ and $H^I(\sigma)$ are almost surely different and from lemma 1.1 we can see that their correlation decreases as the size of I increases.

For a vertex σ and two coordinates i, j, consider the square with vertices $\{\sigma, \sigma^{(i)}, \sigma^{(j)}, \sigma^{(\{i,j\})}\}$. The Hamiltonian takes different values at the vertices of the square and moving along the square edges we can measure energy differences. Signs of these height changes are encoded in a *square pattern*. In the first part of the thesis, we explore different square patterns. They are completely described by a four-dimensional subvector of $H(\sigma)_{\sigma\in\Sigma}$, resulting in easier calculations while still inheriting the original covariance structure. We find that square patterns with matching opposite sides are much more likely to happen. In addition, we are able to give precise probabilities with which different square patterns happen. At the end of Section 2, we look at how these probabilities change if we condition on $H(\sigma)$ having a particular energy value.

Another interesting object is a global minimum of Hamiltonian $H(\sigma)$ over all $\sigma \in \Sigma$. Getting good estimates of its energy levels was perhaps at the heart of mean field spin glass research for many decades. In 1979 Giorgio Parisi, in [20], provided a nonrigorous variational formula describing the free energy that has the following result as a corrollary.

Theorem 1.2. *Let $\Sigma = \{\pm 1\}^n$, $H(\sigma)$ the Sherrington-Kirkpatrick Hamiltonian and $c > 0$ a constant. Then,*

$$
n^{-\frac{3}{2}} \min_{\sigma\in\Sigma} H(\sigma) \to -c \quad \text{in probability.} \tag{4}
$$

It was numerically estimated in [7] that $c \approx 0.7632$. There are also some good bounds known. In a celebrated paper, [25], Talagrand proved Parisi formula.

The proof of the above theorem is well beyond the scope of this thesis. In Section 3 we focus, instead, on local optima. Section 2 and Section 3 can be read independently from each other. A lot of important algorithms for finding high-energy configurations are greedy and stop at local minima/maxima. Therefore, it is important to have a good understanding of them. How many local optima do we expect to see? At what energies do they typically appear? We look at two paths one can take when trying to answer these questions. In the first part of Section 3, we follow [1] and dive into measure concentration theory mixed with some manipulation of normal random variables. In the second part of Section 3, we take a more analytical approach of [28] and look at distribution theory while calculating (too) many integrals. We try to extend both approaches to get similar results about radius two local optima.

2 Square patterns

Uniformly at random choose any vertex $\sigma \in \Sigma$ and two coordinates i, j. We fix all of that for the rest of the chapter. Consider a square $S = \{\sigma, \sigma^{(i)}, \sigma^{(j)}, \sigma^{(\{i,j\})}\}$. The random variables

$$Z_1(\sigma) = \frac{H^{(i)}(\sigma) - H(\sigma)}{2} = -\sum_{k \in [n] \setminus i} \sigma_k \sigma_i g_{ik}, \tag{5}$$

$$Z_2(\sigma) = \frac{H^{(j)}(\sigma) - H(\sigma)}{2} = -\sum_{k \in [n] \setminus j} \sigma_k \sigma_j g_{jk}, \tag{6}$$

$$Z_3(\sigma) = \frac{H^{(\{i,j\})}(\sigma) - H^{(i)}(\sigma)}{2} = -\sum_{k \in [n] \setminus i,j} \sigma_k \sigma_j g_{jk} + \sigma_i \sigma_j g_{ij}, \tag{7}$$

$$Z_4(\sigma) = Z_1(\sigma) + Z_3(\sigma) - Z_2(\sigma) \tag{8}$$

represent changes in the Hamiltonian as we move along edges of the square. Above we have quietly chosen a direction in which we measure height differences and we set σ as a "central point", i.e. we always measure away from it. From now on dependencies on σ will be implicit.

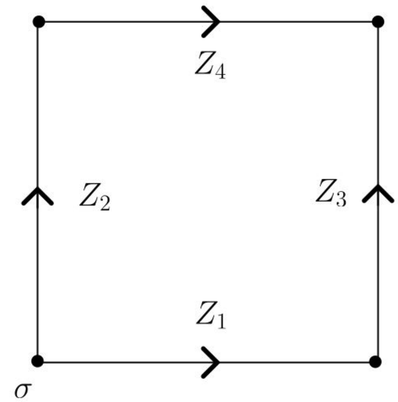

Figure 1: Directions of Z_i's

Square pattern is a random variable $P = (z_1, z_2, z_3, z_4) \in PM = \{+, -\}^4$, where for $1 \leq i \leq 4$,

$$z_i = \begin{cases} + & \text{if } Z_i > 0, \\ - & \text{if } Z_i < 0. \end{cases}$$

Our goal in this Section is to completely describe the probability distribution of P. Even though PM consists of 16 states, there are different symmetries that make our calculations easier. The pairs of edges equipped with (Z_1, Z_3) and (Z_2, Z_4) play the same role and the choice of whether to first consider i or j direction, when numbering Z_i's was arbitrary. A map $(z_1, z_2, z_3, z_4) \mapsto (z_2, z_1, z_4, z_3)$ is called *diagonal symmetry* and it's invariant under probability measure $\mathbb{P}$, i.e.

$$\mathbb{P}[(z_1, z_2, z_3, z_4)] = \mathbb{P}[(z_2, z_1, z_4, z_3)].$$

Intuitively, we reflect differences in the Hamiltonian across the diagonal of S, see Figure 2.

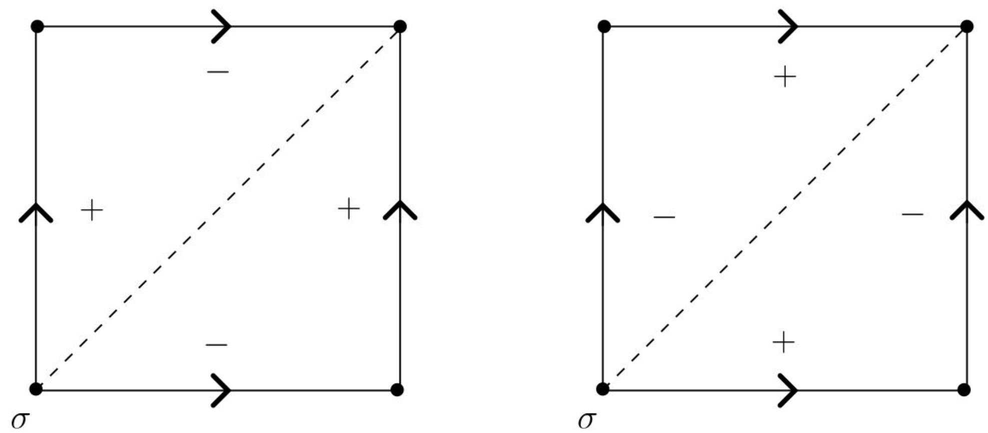

Figure 2: Example of diagonal symmetry

Similarly, all vertices $\sigma \in \Sigma$ play the same role and we were just as likely to start with any other vertex in the square S, but with the same two directions i and j. Then we would be looking at the same square, but from a different central point, resulting in observing different square pattern. Necessarily, these two square patterns are equally likely to be observed, since they depend only on our perspective. Therefore,

$$\mathbb{P}[(z_1, z_2, z_3, z_4)] = \mathbb{P}[(z_3, -z_1, -z_4, z_2)] = \mathbb{P}[(-z_2, z_4, z_1, -z_3)] = \mathbb{P}[(-z_4, -z_3, -z_2, -z_1)].$$

Positional symmetry is a collection of maps $(z_1, z_2, z_3, z_4) \mapsto (z_3, -z_1, -z_4, z_2)$, $(z_1, z_2, z_3, z_4) \mapsto (-z_2, z_4, z_1, -z_3)$ and $(z_1, z_2, z_3, z_4) \mapsto (-z_4, -z_3, -z_2, -z_1)$, all corresponding to a version of described change in perspective. See Figure 3 below.

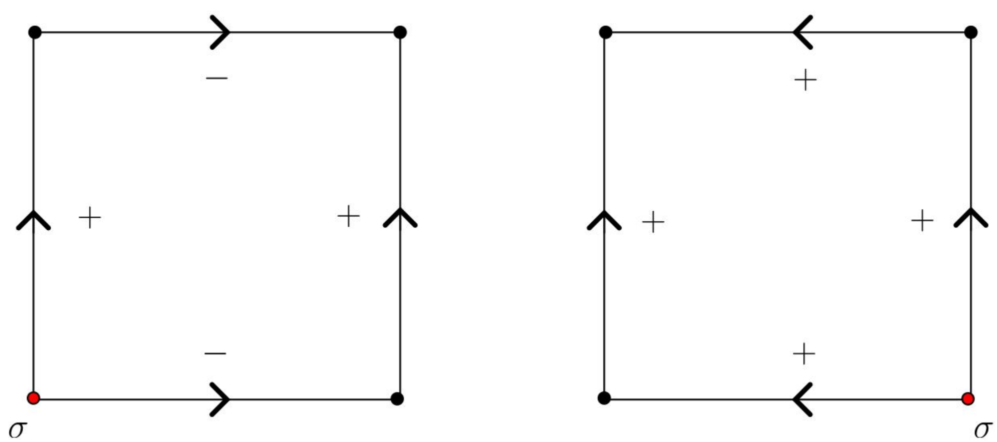

Figure 3: Example of positional symmetry

We say $pm_1, pm_2 \in PM$ are related if we can get from one to another using a sequence of diagonal and positional symmetries. It is easy to see this is an equivalence relation on PM. We call its equivalence classes *square classes*. There are four of them:

- *Normal* square class consists of four patterns: $(+,+,+,+), (-,+,+,-), (+,-,-,+)$ and $(-,-,-,-)$.

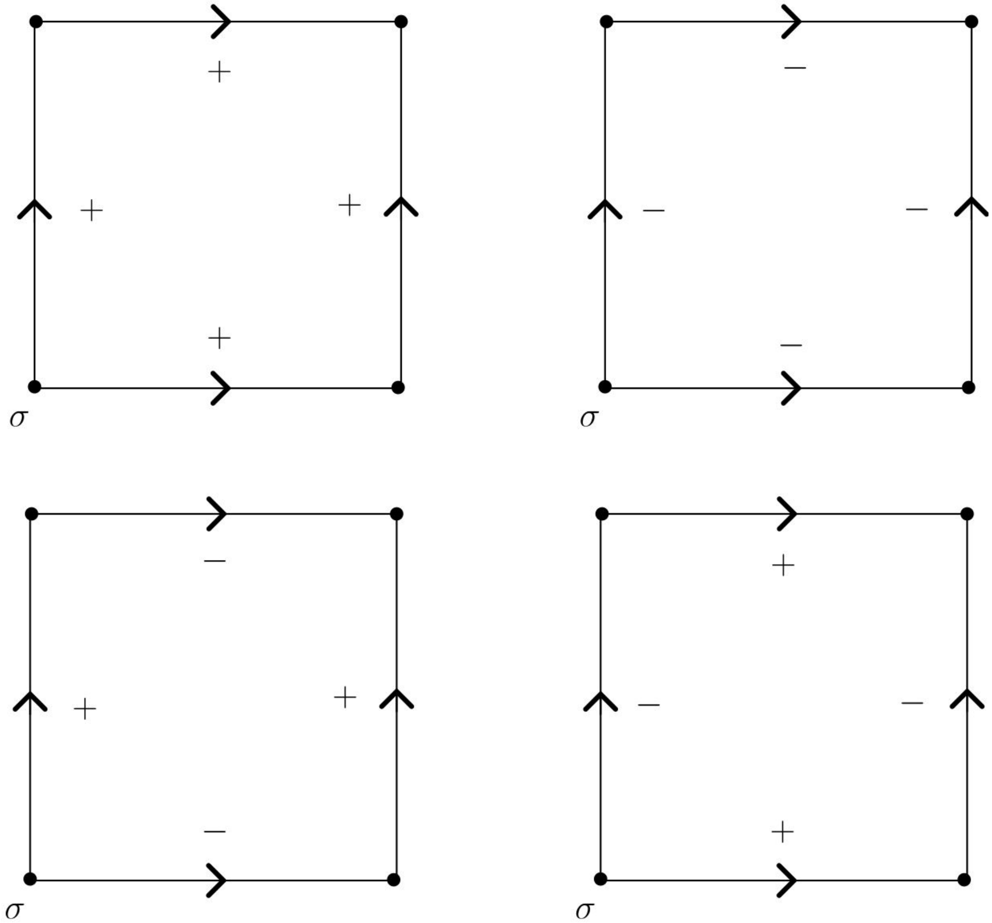

Figure 4: Normal square class

- *Level 1 epistasis* (EP_1) square class consists of eight patterns: $(-,+,+,+), (+,+,+,-), (-,-,-,+)$, $(+,-,-,-), (+,-,+,+), (+,+,-,+), (-,+,-,-)$ and $(-,-,+,-)$.

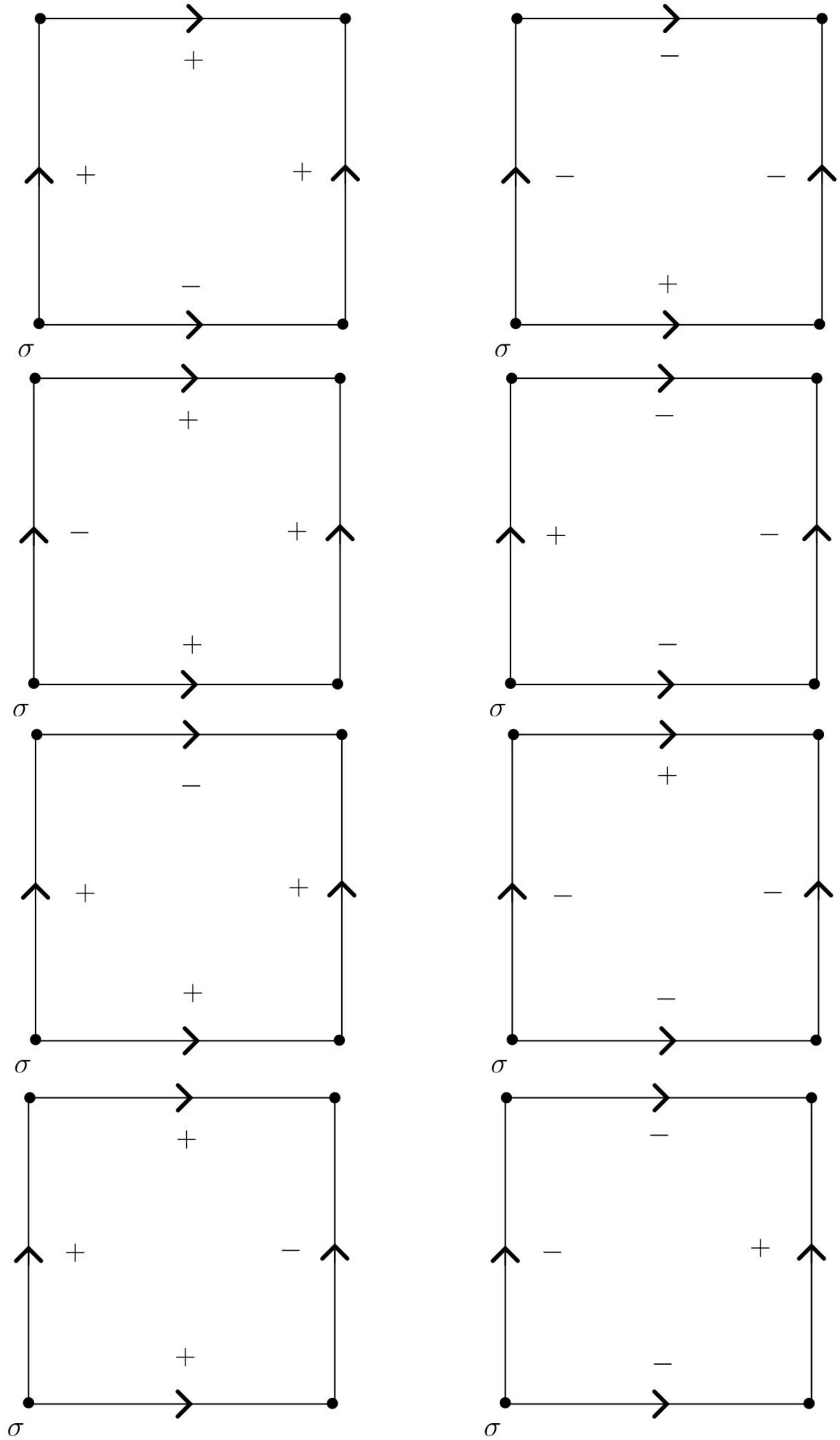

Figure 5: Level 1 epistasis square class

- *Level 2 epistasis* (EP_2) square class consists of two patterns: $(-,-,+,+)$ and $(+,+,-,-)$.

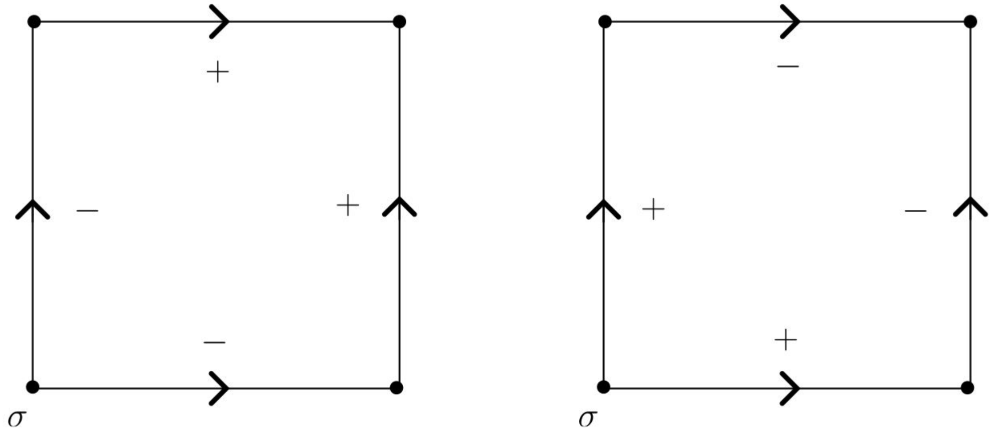

Figure 6: Level 2 epistasis square class

- *Impossible* square class consists of two patterns: $(+,-,+,-)$ and $(-,+,-,+)$.

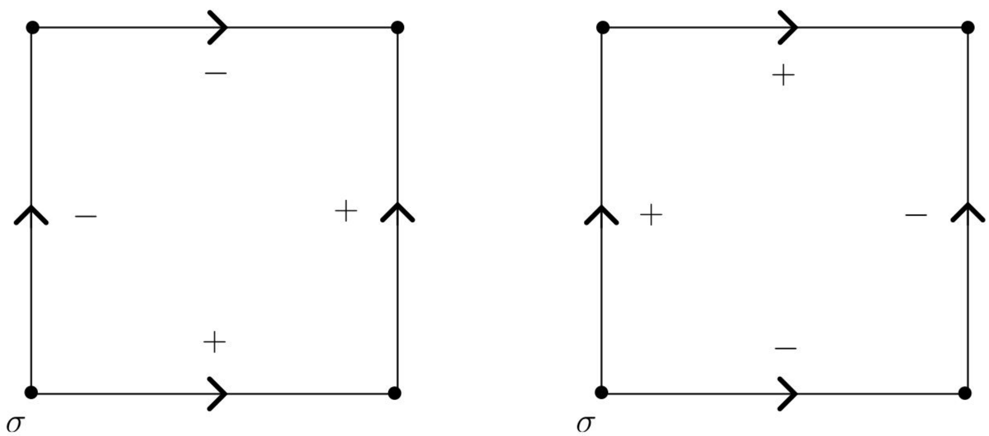

Figure 7: Impossible square class

The event $Z_i = 0$, for some i, is null, hence we are inside of one of the square classes almost surely. We ignore that issue from now on by working on the appropriate subspace of measure 1. Within any square class, all square patterns happen with the same probability, so it's enough to calculate the probability with which each of the selected class representatives appears. Since $Z_1 + Z_3 = Z_2 + Z_4$, impossible class can't happen. Furthermore, the sum of probabilities is 1 and calculating probabilities with which each of the two epistasis classes appear gives a complete description of the distribution of S.

Epistasis classes are characterised by the number of opposite sides of the square that differ, for example, $z_1 = +$ and $z_4 = -$. Level 1 epistasis has one non-matching side, level 2 epistasis has two. In contrast, normal class has both opposite sides matching. We show that opposite sides are highly correlated, so they are not very likely to be different. In theoretical biology, the term epistasis is used to describe, loosely speaking, the possibility of different outcomes in different individuals under the same mutation. There are numerous causes for this phenomenon. Some are environmental (pathogens, temperature), while other are purely stochastic, for example, errors in genes. Different mathematical models were proposed throughout the years to try to capture the effect genes have on phenotype. The most popular is the NK model, see [29], but the Sherrington-Kirkpatrick model is sometimes also used. A biological take on epistasis can be found in [10].

Let $Z = (Z_1, Z_2, Z_3)^T \in \mathbb{R}$ be a multivariate normal random variable. The square S having a certain square pattern is equivalent to Z being in a certain subset of $\mathbb{R}^3$. Thus, we only need a good description of Z and the appropriate subset of $\mathbb{R}^3$. The mean of Z is 0 and the covariance matrix is

$$\Sigma := \operatorname{Cov}((Z_1, Z_2, Z_3)^T) = \begin{pmatrix} n-1 & 1 & -1 \\ 1 & n-1 & n-3 \\ -1 & n-3 & n-1 \end{pmatrix}. \tag{9}$$

We would prefer to work with a standard normal random variable, so we choose a basis in which corresponding quadratic form is given by a diagonal symmetric bilinear form. One can check that if we take a basis

$$\begin{aligned} R_1 &= \frac{Z_1}{\sqrt{n-1}} \\ R_2 &= \frac{Z_2 + Z_3}{\sqrt{2n}} \\ R_3 &= \frac{2Z_1 - (n-1)Z_2 + (n-1)Z_3}{2(n-1)} \end{aligned}$$

then $R = (R_1, R_2, R_3)^T$ is a multivariate normal random variable with mean 0 and identity as a covariance matrix, i.e. R_i's are pairwise independent.

We express Z in terms of R. Sign patterns correspond to signs of coordinates of Z (together with Z_4), which in turn corresponds to different regions in R-coordinate system. We get

$$\begin{aligned} Z_1 &= \sqrt{n-1}R_1, \\ Z_2 &= \frac{R_1}{\sqrt{n-1}} + \sqrt{\frac{n}{2}}R_2 - R_3 \\ Z_3 &= -\frac{R_1}{\sqrt{n-1}} + \sqrt{\frac{n}{2}}R_2 + R_3 \\ Z_4 &= \left(\sqrt{n-1} - \frac{2}{\sqrt{n-1}}\right) R_1 + 2R_3. \end{aligned}$$

Notice that the only class with a non-zero limit probability of appearing is normal square class. This can be explained by the covariance between opposite sides of the square increasing with n. There are $O(2^n)$ squares in Σ. This quick growth might subsidise for the decay of epistasis square patterns and we might still expect to see a lot of them as n grows.

2.1 Level 2 epistasis

Let A be an event that the square S has a pattern $(+,+,-,-)$. This corresponds to the system

$$
\begin{aligned}
R_1 &> 0\\
R_3 &< \frac{R_1}{\sqrt{n-1}} + \sqrt{\frac{n}{2}}R_2\\
R_3 &< \frac{R_1}{\sqrt{n-1}} - \sqrt{\frac{n}{2}}R_2\\
R_3 &< \frac{1}{2}\left(\frac{2}{\sqrt{n-1}} - \sqrt{n-1}\right)R_1.
\end{aligned}
$$

Recall R_i's are independent and $\mathbb{P}(R_1 > 0) = \frac{1}{2}$.

The rest of the calculations (and the thesis in general) rely heavily on manipulation of normal random variables. We once again point to Appendix A, where one can find a collection of terminology and facts about Gaussians. Given two independent random variables A and B we are often interested in the maximum $M = \max(A, B)$. If we know their pdf's f_A, f_B and cdf's F_A, F_B, we have a good description of M. An elementary calculation gives

$$
\begin{aligned}
F_M(x) &= \mathbb{P}[M \le x]\\
&= \mathbb{P}[A \le x]\mathbb{P}[B \le x]\\
&= F_A(x)F_B(x)
\end{aligned}
$$

and upon differentiating

$$f_M(x) = f_A(x)F_B(x) + F_A(x)f_B(x).$$

For this Subsection, we set $A = \left|\sqrt{\frac{n}{2}}R_2\right|$, $B = \left|\frac{\sqrt{n}}{2}R_1\right|$ and M to be the corresponding maximum.

Upper bound

For every $c > 0$, we can bound

$$
\begin{aligned}
\mathbb{P}[A] &= \mathbb{P}\left[A \mid R_1 \le c\sqrt{n-1}\right]\mathbb{P}\left[R_1 \le c\sqrt{n-1}\right] + \mathbb{P}\left[A \mid R_1 \ge c\sqrt{n-1}\right]\mathbb{P}\left[R_1 \ge c\sqrt{n-1}\right]\\
&\le \frac{1}{2}\mathbb{P}\left[A \mid 0 \le R_1 \le c\sqrt{n-1}\right]\cdot 1 + 1 \cdot \mathbb{P}\left[R_1 \ge c\sqrt{n-1}\right]\\
&\le \frac{1}{2}\mathbb{P}\left[R_3 \ge M - c\right] + \mathbb{P}\left[R_1 \ge c\sqrt{n-1}\right].
\end{aligned}
$$

Gaussian tails decay exponentially, so we have a good control over $\mathbb{P}\left[R_1 \ge c\sqrt{n-1}\right]$. The first term can calculated as

$$
\begin{aligned}
\mathbb{P}\left[R_3 \ge M - c\right] &= \int_0^\infty F_M(x) f_{R_3}(x-c)\,dx\\
&= \frac{1}{\sqrt{2\pi}}\int_0^\infty \operatorname{erf}\left(\frac{x}{\sqrt{n}}\right)\operatorname{erf}\left(\frac{x\sqrt{2}}{\sqrt{n}}\right)\exp\left(-\frac{(x-c)^2}{2}\right)dx\\
&= \frac{1}{\sqrt{2\pi}}(I_1 + I_2),
\end{aligned}
$$

where

$$I_1 = \int_0^{\sqrt{\frac{n\pi}{8}}} \operatorname{erf}\left(\frac{x}{\sqrt{n}}\right) \operatorname{erf}\left(\frac{x\sqrt{2}}{\sqrt{n}}\right) \exp\left(-\frac{(x-c)^2}{2}\right) dx \leq \frac{8}{n\sqrt{\pi}} \left(\frac{1}{2} + \epsilon_c\right)(1+c^2),$$

for some constant ϵ_c depending on c that goes to 0 as $c \to 0$, and

$$I_2 = \int_{\sqrt{\frac{n\pi}{8}}}^{\infty} \operatorname{erf}\left(\frac{x}{\sqrt{n}}\right) \operatorname{erf}\left(\frac{x\sqrt{2}}{\sqrt{n}}\right) \exp\left(-\frac{(x-c)^2}{2}\right) dx \leq o\left(\frac{1}{n}\right).$$

The above upper bounds for I_1 and I_2 are derived in Appendix C. Combining this with upper bounding Gaussian tails (Lemma A.1) gives

$$\begin{aligned}
\mathbb{P}(A) &\leq \frac{1}{2}\mathbb{P}\left[R_3 \geq M - c\right] + \mathbb{P}\left[R_1 \geq c\sqrt{n-1}\right] \\
&\leq \frac{1}{2\sqrt{2\pi}}(I_1 + I_2) + \frac{1}{\sqrt{2\pi}} \frac{1}{c\sqrt{n-1}} \exp\left(-\frac{c^2(n-1)}{2}\right) \\
&\leq \frac{2\sqrt{2}}{n\pi}\left(\frac{1}{2} + \epsilon_c\right)(1+c^2) + \frac{1}{\sqrt{2\pi}} \frac{1}{c\sqrt{n}} \exp\left(-\frac{c^2 n}{2}\right) + o\left(\frac{1}{n}\right).
\end{aligned}$$

Now we choose $c = n^{-\frac{1}{3}}$ to see that

$$\mathbb{P}(A) \leq \frac{\sqrt{2}}{\pi n} + o\left(\frac{1}{n}\right).$$

Similar approach can be used (see Appendix C) to lower bound

$$\mathbb{P}[A] \geq \frac{\sqrt{2}}{\pi n} + o\left(\frac{1}{n}\right).$$

Putting it together, we have a precise asymptotic behaviour of $(+,+,-,-)$ pattern. Since all patterns in a square class are equally likely, we have proved the following theorem.

Theorem 2.1. *Probability of observing level 2 epistasis satisfies*

$$\lim_{n\to\infty} n \cdot \mathbb{P}(EP_2) = \frac{4\sqrt{2}}{\pi}.$$

2.2 Level 1 Epistasis

There are 8 different sign patterns that give rise to level 1 epistasis. Since they are all equally likely it is enough to calculate the probability of event $B = (+,+,+,-)$, or equivalently the probability that the following inequalities hold for $R = (R_1, R_2, R_3)$ standard centred normal random variable in $\mathbb{R}^3$,

$$\begin{aligned}
&R_1 > 0 \\
&R_3 < \frac{R_1}{\sqrt{n-1}} + \sqrt{\frac{n}{2}} R_2 \\
&R_3 < \frac{R_1}{\sqrt{n-1}} - \sqrt{\frac{n}{2}} R_2 \\
&R_3 > \frac{1}{2}\left(\frac{2}{\sqrt{n-1}} - \sqrt{n-1}\right) R_1
\end{aligned}$$

Let $X = \sqrt{\frac{n}{2}}|R_2|$ and $Y = \frac{\sqrt{n-1}}{2}|R_1|$. Then X and Y are independent half-normal random variables with variances $\frac{n}{2}$ and $\frac{n-1}{4}$ respectively. Both are also independent of R_3. Conditioning on $R_1 > 0$, our system is equivalent to

$$X + \frac{R_1}{\sqrt{n-1}} < R_3 < Y + \frac{R_1}{\sqrt{n-1}}. \tag{10}$$

Upper bound

Having (10) in mind we can upper bound $\mathbb{P}(B)$, for $c > 0$, as

$$\begin{aligned}\mathbb{P}(B) &= \mathbb{P}\left[B \mid R_1 < c\sqrt{n-1}\right]\mathbb{P}\left[R_1 < c\sqrt{n-1}\right] + \mathbb{P}\left[B \mid R_1 > c\sqrt{n-1}\right]\mathbb{P}\left[R_1 > c\sqrt{n-1}\right] \\ &\le \frac{1}{2}\mathbb{P}[X < R_3 < Y + c] \cdot 1 + 1 \cdot \mathbb{P}\left[R_1 > c\sqrt{n-1}\right].\end{aligned}$$

We have control over Gaussian tails of R_1. As for the other term, we can calculate

$$\begin{aligned}\mathbb{P}[X < R_3 < Y + c] &= \int_0^\infty \int_0^\infty (F_{R_3}(y+c) - F_{R_3}(x)) f_X(x) f_Y(y) \mathbb{1}_{\{x \le y+c\}} dx dy \\ &= \frac{2\sqrt{2}}{\pi\sqrt{n(n-1)}}(J_1 - J_2),\end{aligned}$$

where

$$J_1 = \int_0^\infty \int_{x-c\wedge 0}^\infty \operatorname{erfc}\left(\frac{x}{\sqrt{2}}\right) e^{-\frac{x^2}{n}} e^{-\frac{2y^2}{n-1}} dy dx \le \frac{\sqrt{n}}{2}\left(1 + c\sqrt{\frac{\pi}{2}} + o(1)\right)$$

and

$$J_2 = \int_0^\infty \int_0^{y+c} \operatorname{erfc}\left(\frac{y+c}{\sqrt{2}}\right) e^{-\frac{x^2}{n}} e^{-\frac{2y^2}{n-1}} dx dy \ge (1 + (o(1))\frac{c\sqrt{\pi n}}{2},$$

bounding once again being done in Appendix C. Therefore, we have

$$\begin{aligned}\mathbb{P}[B] &\le \frac{1}{2}\mathbb{P}[X < R_3 < Y + c] \cdot 1 + 1 \cdot \mathbb{P}\left[R_1 > c\sqrt{n-1}\right] \\ &\le \frac{\sqrt{2}}{\pi\sqrt{n(n-1)}}(J_1 - J_2) + \frac{1}{c\sqrt{2\pi(n-1)}} e^{-\frac{c^2(n-1)}{2}} \\ &\le (1 + o(1))\left(\frac{\sqrt{2}}{\pi n}\frac{\sqrt{n}}{2} + \frac{1}{c\sqrt{2\pi n}} e^{-\frac{c^2 n}{2}}\right)\end{aligned}$$

Taking $c = n^{-\frac{1}{3}}$ and letting $n \to \infty$ we get

$$\mathbb{P}[B] \le \frac{1}{\pi\sqrt{2n}} + o\left(\frac{1}{\sqrt{n}}\right).$$

Doing very similar calculations (Appendix C) for the lower bound results in the following theorem.

Theorem 2.2. *Probability of observing level 1 epistasis satisfies*

$$\lim_{n\to\infty} \sqrt{n} \cdot \mathbb{P}(EP_1) = \frac{2\sqrt{2}}{\pi}.$$

Every time one of the pairs of opposite sides of the square differs and the level of epistasis increases, the probability of observing such a pattern is multiplied by $\frac{1}{\sqrt{n}}$. The following simple problem might shine some light on where the $\frac{1}{\sqrt{n}}$ decay is coming from. Let $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(0,1)$ be independent standard normal random variables. What is $\mathbb{P}[Y > \sqrt{n}X \mid Y > 0]$? It's exactly the integral of joint Gaussian density over the area shaded in Figure 8.

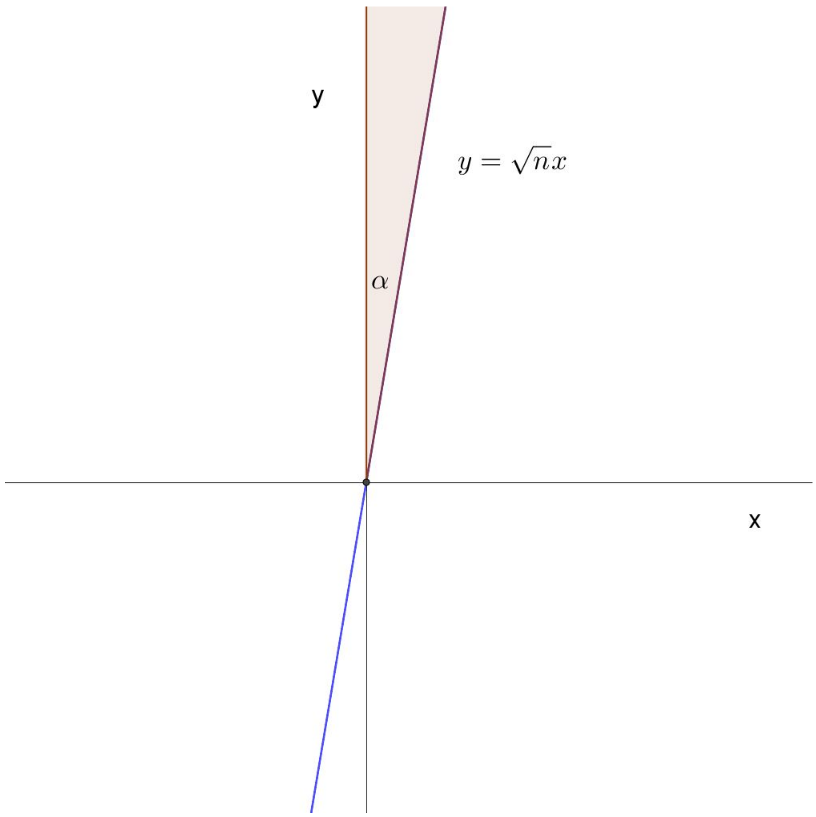

Figure 8: Shaded area represents $\mathbb{P}[Y > \sqrt{n}X \mid Y > 0]$

Gaussian density is circularly symmetric, meaning that the value of pdf at a point depends only on its distance from the origin. Hence, it's enough to calculate the angle α between $y = nx$ line and y-axis. That is,

$$\mathbb{P}[Y > \sqrt{n}X \mid Y > 0] = \frac{\alpha}{2\pi},$$

which can be easily calculated as

$$\frac{1}{2\pi} \arctan\left(\frac{1}{\sqrt{n}}\right) = \frac{1}{2\pi\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right).$$

On a high level, appearance of epistasis for a pair of sides is having $Y > \sqrt{n}X \mid Y > 0$, for X and Y independent standard normals. Opposite sides of the square S are much more correlated than adjacent sides, so each pair acts "independently" enough to force an extra $\frac{1}{\sqrt{n}}$ decay if mismatched.

2.3 Square patterns conditioned on energy

We still work with the same fixed square S, but this time we also have access to the energy value (Hamiltonian) at the vertex σ. If this value is typical for a global maximum, we intuitively expect to see a decrease in energy when we move to a randomly chosen neighbour. In particular, it is not likely that we are at a local minimum. We expect to see more minuses than pluses on the edges of the square S. In this Subsection, we formalize and generalize this intuition. We look at how conditioning on σ having a certain energy changes the expected number of different square patterns.

The Hamiltonian $H = H(\sigma)$ is a centred normal random variable with variance $\binom{n}{2}$, so its typical energy levels are of order n. It follows from the corollary of the Parisi formula (Theorem 1.2), mentioned in the introduction, that typical value of energy at a global maximum/minimum is of order $n^{\frac{3}{2}}$. Paper [1] shows that the Hamiltonian typically has the same order at local maxima/minima. See Section 3 and in particular theorem 3.12 for more details. Note that typical value and expected value are two different concepts. *Expected value* is a weighted mean of possible outcomes, while *typical value* is a type of outcome that happens with a high probability. Imagine drawing of $l \gg 1$ lots. There is only one ticket that pays out l, all other tickets are worthless. The expected (average) prize is then 1, but the typical price is 0.

We still work with Hamiltonian differences Z_1, Z_2, Z_3 and Z_4 from equations (5)-(8). This time the vector of interest is $Z = (H, Z_1, Z_2, Z_3)^T$. It has one dimension more than Z before. However, we condition on H, so effectively we still need to integrate over different subsets of $\mathbb{R}^3$. The covariance matrix of Z is

$$\operatorname{Cov}((H, Z_1, Z_2, Z_3)^T) = \begin{pmatrix} \binom{n}{2} & -(n-1) & -(n-1) & -(n-3) \\ -(n-1) & n-1 & 1 & -1 \\ -(n-1) & 1 & n-1 & n-3 \\ -(n-3) & -1 & n-3 & n-1 \end{pmatrix}.$$

Recall that $(Z_1, Z_2, Z_3) \sim \mathcal{N}(0, \Sigma)$, with Σ is given (9). Conditioned on $H = h$, the distribution of random vector $(Z_1, Z_2, Z_3) \sim \mathcal{N}(\mu(h), \Sigma')$, where

$$\mu(h) = \left(-\frac{2h}{n}, -\frac{2h}{n}, -\frac{2(n-3)h}{n(n-1)}\right)^T = -(1+o(1))\frac{2h}{n}(1,1,1)^T$$

and

$$\begin{aligned}\Sigma' &= \Sigma - \frac{2}{n(n-1)}(n-1, n-1, n-3)^T \cdot (n-1, n-1, n-3) \\ &= (1+o(1)) \begin{pmatrix} n-3 & -1 & -3 \\ -1 & n-3 & n-5 \\ -3 & n-5 & n-3 \end{pmatrix}.\end{aligned}$$

See discussion about Gaussians in Appendix A for more details. We are only interested in a limit behaviour as $n \to \infty$, so it's enough to work up to $o(1)$ term. We proceed as in non-conditioned case and find a basis that up to a $o(1)$ terms diagonalises a quadratic form associated to $\Sigma(h)$. One such basis is

$$\begin{aligned} R_1 &= \frac{Z_1}{\sqrt{n}} \\ R_2 &= \frac{3Z_2 - Z_3}{2\sqrt{n}} \\ R_3 &= \frac{2Z_1 - nZ_2 + nZ_3}{n}. \end{aligned}$$

A vector (R_1, R_2, R_3) is a standard Gaussian with identity as a covariance matrix and mean $m' = \left(\frac{\mu}{\sqrt{n}}, \frac{\mu}{\sqrt{n}}, O\left(\frac{\mu}{n}\right)\right)$, where $\mu = \mu(h)_1 = -\frac{2h}{n}$. Denote by

$$m := \frac{\mu}{\sqrt{n}} = -\frac{2h}{n^{3/2}}.$$

We can express (Z_1, Z_2, Z_3, Z_4) as

$$\begin{aligned}
Z_1 &= (1 + o(1))\sqrt{n}R_1 \\
Z_2 &= (1 + o(1))\left(-\frac{R_1}{\sqrt{n}} + \sqrt{n}R_2 + R_3\right) \\
Z_3 &= (1 + o(1))\left(-\frac{3R_1}{\sqrt{n}} + \sqrt{n}R_2 + 3R_3\right) \\
Z_4 &= (1 + o(1))(\sqrt{n}R_1 + 2R_3).
\end{aligned}$$

We are only interested in a limit behaviour, hence we can ignore $o(1)$ terms.

We denote by P the conditional probability measure $\mathbb{P}[\,\cdot \mid H = h]$. Diagonal symmetry is still invariant under this measure, i.e.

$$P[(z_1, z_2, z_3, z_4)] = P[(z_2, z_1, z_4, z_3)].$$

However, conditioning on energy value at σ introduces a distinguished point in the hypercube and we no longer have invariance of positional symmetry. Consequently, there are many more square classes to consider. There is another symmetry-type equality that can help us reduce the number of cases we need to calculate. Normal random variables are symmetric around zero and so

$$\mathbb{P}[(z_1, z_2, z_3, z_4) \mid H = h] = \mathbb{P}[(-z_1, -z_2, -z_3, -z_4) \mid H = -h].$$

We call this *sign symmetry*. There are two different types of energy we will condition on. For $k \in \mathbb{R}$, there is a typical energy $h = kn$ that we are likely to see at a randomly chosen point and an extreme energy $h = kn^{3/2}$ we usually find at local optima. The behaviour of square patterns depends on the conditioning we choose. We only do sample calculations for the square pattern $(-, +, +, +)$. Other cases follow analogously.

Denote by $D = (-, +, +, +)$ the pattern of interest. It is equivalent to the system

$$\begin{aligned}
&R_1 < 0 \\
&R_3 > \frac{R_1}{\sqrt{n}} - \sqrt{n}R_2 \\
&R_3 > \frac{R_1}{\sqrt{n}} - \frac{\sqrt{n}}{3}R_2 \\
&R_3 > -\frac{\sqrt{n}}{2}R_1.
\end{aligned}$$

To further ease the notation write $m_c = \frac{m}{\sqrt{n}} - c$. For $c \geq 0$ denote by D_c the following system

$$\begin{aligned}
&R_1 < 0 \\
&R_3 > m_c - \sqrt{n}R_2 \\
&R_3 > m_c - \frac{\sqrt{n}}{3}R_2 \\
&R_3 > -\frac{\sqrt{n}}{2}R_1.
\end{aligned}$$

We use the usual bounds on Gaussian tails to get

$$\begin{aligned}
P[D] &= P[D \,|\, R_1 > m - c\sqrt{n}] \cdot P[R_1 > m - c\sqrt{n}] + P[D \,|\, R_1 < m - c\sqrt{n}] \cdot P[R_1 < m - c\sqrt{n}] \\
&\leq P[D_c \,|\, R_1 > m - c\sqrt{n}] \cdot P[R_1 > m - c\sqrt{n}] + 1 \cdot \frac{1}{c\sqrt{2\pi n}} \exp\left(-\frac{c^2 n}{2}\right) \\
&= P[D_c] + \frac{1}{c\sqrt{2\pi n}} \exp\left(-\frac{c^2 n}{2}\right).
\end{aligned}$$

For this bound to be useful, we want $c \gg \frac{1}{\sqrt{n}}$. We will later pick $c = n^{-\frac{1}{3}}$. We can calculate

$$\begin{aligned}
P[D_c] &= P[D_c \mid R_1 < 0] \cdot \Phi(-m) \\
&= \Phi(-m) \cdot P\left[R_3 - m_c > \max\left(-\sqrt{n}R_2, -\frac{\sqrt{n}}{3}R_2, -\frac{\sqrt{n}}{2}R_1 - m_c\right) \mid R_1 < 0\right].
\end{aligned}$$

Denote by $A = -\sqrt{n}R_2, B = -\frac{\sqrt{n}}{3}R_2$ and $C = -\frac{\sqrt{n}}{2}R_1 - m_c \mid R_1 < 0$. Unless R_1 is really small in absolute value compared to c, C is greater than zero and so is the maximum. In that case R_2 can only make a difference when it is negative, but then $A > B$. Hence, $P[D_c]$ is equal to

$$\Phi(-m)\left(\int_0^\infty F_A(x)F_C(x)f_{R_3}\left(x + m_c\right)dx + E\right), \tag{11}$$

where $E = \int_0^{m_c} F_M(x)F_C(x)f_{R_3}\left(x + m_c\right)dx$, for $M = \max(A, B)$, is an error term. Cumulative distribution of A is that of a Gaussian, while we use equation (A33) for F_C, to finally get

$$P[D_c] \propto (1+o(1))\frac{e^{-m^2/2}}{\sqrt{n}}\int_0^\infty \left(1 + \operatorname{erf}\left(\frac{x}{\sqrt{2n}} + \frac{m}{\sqrt{2}}\right)\right)\left(x + m_c\right)\exp\left(-\frac{1}{2}\left(x + m_c\right)^2\right)dx + \Phi(-m)E. \tag{12}$$

Typical energy levels

At typical energy levels $h = kn$, for some $k \in \mathbb{R}$, giving $m = -\frac{2k}{\sqrt{n}}$. We let $c = n^{-1/3}$. For some constant $L > 0$, independent of n, we can bound

$$\begin{aligned}
\Phi(-m)E &\leq \Phi(-m)\int_0^{m_c} F_C(x)f_{R_3}\left(x + m_c\right)dx \\
&\leq \frac{L}{\sqrt{n}}\int_0^{m_c}\left(x + m_c\right)\exp\left(-\frac{1}{2}\left(x + m_c\right)^2\right)dx
\end{aligned}$$

This integral is of a similar form as the other integral in (12), the only real difference is the range over which we are integrating. Now we take limits. All the integrands are bounded by 1 and so limits commute nicely with integrals by the Dominated convergence theorem. We further note that for $a \in (0, \infty]$,

$$\int_0^a xe^{-\frac{x^2}{2}}dx = 1 - e^{-\frac{a^2}{2}} \leq a^2,$$

which converges to 0 with a rate higher than a when a converges to 0. For $a = \infty$, it is equal to 1. That means that $\Phi(-m)E$ is dominated by the first term in (12). Furthermore, the limit values of integrals don't depend on k.

All in all, we see that the $\lim_{n\to\infty}\sqrt{n}\ P[D_c]$ is a constant independent from k. Let D' be the following system

$$\begin{aligned} &R_1 < 0\\ &R_3 > -\sqrt{n}R_2\\ &R_3 > -\frac{\sqrt{n}}{3}R_2\\ &R_3 > -\frac{\sqrt{n}}{2}R_1. \end{aligned}$$

Since $R_1 < 0$, we can trivially add the left inequality in

$$P[D'] \le P[D] \le P[D_c] + \frac{1}{c\sqrt{2\pi n}} \exp\left(-\frac{c^2 n}{2}\right). \tag{13}$$

Calculating $P[D']$ is the same as calculating $P[D_c]$ with $m = 0$ and $c = 0$. Hence,

$$\lim_{n\to\infty}\sqrt{n}\ P[D'] = \lim_{n\to\infty}\sqrt{n}\ P[D_{n^{-1/3}}].$$

Taking limits in (13) we see that $\lim_{n\to\infty}\sqrt{n}\ P[A]$ is a constant that doesn't depend on a value of k in $h = kn$. By the Central limit theorem

$$\frac{H}{\binom{n}{2}^{\frac{1}{2}}} \longrightarrow \mathcal{N}(0,1),$$

implying

$$\mathbb{P}[H = O(n)] = 1.$$

Therefore, the constant must be the one from Theorem 2.2, i.e.

$$\lim_{n\to\infty}\sqrt{n}\ P[A] = \frac{1}{\pi\sqrt{2}}.$$

As long as we are at typical energy levels the precise value of energy at σ doesn't influence the limit probabilities of square patterns of S. The contribution of σ is not too extreme and gets overwritten by other points.

Extreme energy levels

The situation is different if we condition on $h = kn^{\frac{3}{2}}$, i.e. we are at very high energy levels. The conditioning here is much more drastic and the probabilities of observing different sign patterns depend on the value of k. In this case $m = -2k$. Once again we look at equation (12). Using the same reasoning as above we see that the error term E is dominated by the integral term. Taking limits we get, with the help of Dominated convergence theorem,

$$\lim_{n\to\infty}\sqrt{n}\ \mathbb{P}[A_{n^{-1/3}}] \propto e^{-m^2/2}\left(1 + \operatorname{erf}\left(\frac{m}{\sqrt{2}}\right)\right).$$

Going through calculations we see that the m_c term completely disappeared in the limit and so it follows from (13) that

$$\lim_{n\to\infty}\sqrt{n}\ \mathbb{P}[A] \propto e^{-m^2/2}\left(1 + \operatorname{erf}\left(\frac{m}{\sqrt{2}}\right)\right) = e^{-2k^2}\left(1 - \operatorname{erf}\left(\sqrt{2}k\right)\right).$$

Sign pattern	$P[\,\cdot\,] \propto$	**Epistasis level**
$(+,+,+,+)$	$\left(1-\operatorname{erf}\left(\sqrt{2}k\right)\right)^2$	0
$(-,-,-,-)$	$\left(1+\operatorname{erf}\left(\sqrt{2}k\right)\right)^2$	0
$(-,+,+,+)$	$e^{-2k^2}\left(1-\operatorname{erf}\left(\sqrt{2}k\right)\right)$	1
$(+,-,+,+)$	$e^{-2k^2}\left(1-\operatorname{erf}\left(\sqrt{2}k\right)\right)$	1
$(+,+,-,+)$	$e^{-2k^2}\left(1-\operatorname{erf}\left(\sqrt{2}k\right)\right)$	1
$(+,+,+,-)$	$e^{-2k^2}\left(1-\operatorname{erf}\left(\sqrt{2}k\right)\right)$	1
$(+,-,-,-)$	$e^{-2k^2}\left(1+\operatorname{erf}\left(\sqrt{2}k\right)\right)$	1
$(-,+,-,-)$	$e^{-2k^2}\left(1+\operatorname{erf}\left(\sqrt{2}k\right)\right)$	1
$(-,-,+,-)$	$e^{-2k^2}\left(1+\operatorname{erf}\left(\sqrt{2}k\right)\right)$	1
$(-,-,-,+)$	$e^{-2k^2}\left(1+\operatorname{erf}\left(\sqrt{2}k\right)\right)$	1
$(+,-,-,+)$	$1-\operatorname{erf}\left(\sqrt{2}k\right)^2$	0
$(-,+,+,-)$	$1-\operatorname{erf}\left(\sqrt{2}k\right)^2$	0
$(+,+,-,-)$	e^{-4k^2}	2
$(-,-,+,+)$	e^{-4k^2}	2
$(+,-,+,-)$	0	2
$(-,+,-,+)$	0	2

Table 1: Limit probabilities of different square patterns when conditioned on $H(\sigma) = kn^{\frac{3}{2}}$.

We can do similar calculations for other square patterns, using different symmetries to reduce the amount of work required. The results are given in table 1. Note that limit probabilities are appropriately normalised by

$$n^{\frac{\text{Epistasis level}}{2}}.$$

For $A \subset \mathbb{R}$, we would like to use a version of Bayes formula to draw some conclusions about

$$\lim_{n\to\infty} \mathbb{P}[H(\sigma) \in A \mid \text{Square pattern}].$$

However, for A a subset of typical energies, the square pattern doesn't give us any information about the exact level of energies. For A a subset of extreme energies, the probability of observing such energies at a randomly chosen point is a Gaussian tail probability and decays exponentially. We are unlikely to be in such a state and specifying a square pattern doesn't alter the rate of exponential decay.

Yet, we can take a step in this direction and define a family of measures on $\mathbb{R}$, indexed by square patterns P, and given by

$$\mu_P(x) = n^{\frac{\text{Epistasis level of } S}{2}} \cdot \mathbb{P}[\,P \mid H(\sigma) = x], \quad \forall x \in \mathbb{R}.$$

The measures indexed by square patterns with more minuses than pluses have negative mean, confirming our intuition that we are more likely to see more minuses if we are near local maxima. Similarly, square patterns with more pluses than minuses result in measures with a negative mean. If a square pattern has two pluses and two minuses, then its measure is symmetric around zero. Increasing the level of epistasis decreases the variance of the underlying measure. For example, both $S = (+,-,-,+)$ and $S' = (+,+,-,-)$ have means $\mathbb{E}_{\mu_S}[1] = \mathbb{E}_{\mu_{S'}}[1] = 0$, but $\mathrm{Var}_{\mu_S}[1] \approx 0.235$ and $\mathrm{Var}_{\mu_{S'}}[1] \approx 0.111$. We are less likely to see epistasis near high energy points. In particular, epistasis is unlikely to happen near global maxima.

3 Local optima

Previous Section alluded to the richness of the landscape of the Sherrington-Kirkpatrick Hamiltonian. In this Section, we continue our exploration by looking at local optima. For $\sigma \in \Sigma, i \in [n]$, define

$$T_i = T_i(\sigma) \coloneqq \frac{H^{(i)}(\sigma) - H(\sigma)}{2} = -\sum_{j \in [n] \setminus i} \sigma_i \sigma_j g_{ij}, \tag{14}$$

a difference in Hamiltonian after flipping ith coordinate. A vertex σ is a *local minimum* if $T_i > 0$ for all $i \in [n]$. A *global minimum* is a vertex that minimises $H(\sigma)$ over all $\sigma \in \Sigma$. Local maximum and global maximum are defined analogously.

As already mentioned in the introduction, getting rigorous results about global minimum is extremely hard and it took decades of time and numerous advances in our understanding of spin glass models, before Talagrand in [25], building on the work of others, finally proved the celebrated Parisi formula that describes the behaviour of global minimum. To state the precise formula we would need to introduce a lot of extra notation that's not too important for our purposes. Instead, we only restate its corollary.

Theorem 3.1. *Let $\Sigma = \{\pm 1\}^n$, $H(\sigma)$ the Sherrington-Kirkpatrick Hamiltonian and $c > 0$ a constant. Then,*

$$n^{-\frac{3}{2}} \min_{\sigma \in \Sigma} H(\sigma) \to -c \quad \text{in probability.} \tag{15}$$

It was numerically estimated in [7] that $c \approx 0.7632$. It was proven in [15] that $c \leq \sqrt{\frac{2}{\pi}} \approx 0.7979$.

We focus, instead, on local optima. We show that the expected number of local optima grows exponentially as $n \to \infty$ and we find the exact rate of growth. We also touch on techniques Adario-Berry et al used in [1] to derive a version of theorem 3.1 for local optima. They showed, roughly speaking, that the typical value of the Hamiltonian at a randomly chosen local minimum is $dn^{\frac{3}{2}}$ for $d \approx -0.506$. A precise statement of the result can be found as theorem 3.12. We demonstrate two techniques, currently available in the literature, for counting local optima. First, we consider a recent approach [1] by Adarrio-Berry et al. It uses measure concentration ideas, with a special focus on large deviation theory techniques. In the second part of this Section, we look at an older and more analytical approach [28] due to Tanaka and Edwards.

Although less special than global optima, locally optimal configurations are still worth considering. The Hamiltonian takes extreme values at them and it was shown in [3] and [12] (both papers show the result for a local optimum of weighted MaxCut problem, but there is a one-to-one correspondence between the problems; see [1] for more details) that they can be computed in polynomial time by various greedy algorithms, making them quickly accessible. Here, by *greedy algorithm*, we mean any algorithm travelling along vertices of the hypercube, that at every step makes a choice based only on information available from its current position and its neighbours. Surprisingly, many such greedy algorithms tend to terminate at local optima with much more extreme energies than we would expect to see. For example, as qualitatively shown in [12], a basic greedy algorithm that chooses its next move uniformly at random among all neighbour vertices with lower energy usually terminates at a local minimum with energy level around $-0.7n^{\frac{3}{2}}$. That's closer to typical values of global minimum than of local minimum. The phenomenon further hints at the diversity of the Hamiltonian landscape near local extrema.

We take a step further and try to generalise the above. A vertex $\sigma \in \Sigma$ is a *level k local optima* if the Hamiltonian takes more a extreme value at it than at any point within (graph-theoretical) distance k away. Such optima can still be computed easily by greedy algorithms and they behave more like global optima as k increases. We will focus on level 2 local optima. For $i \neq j$, define

$$T_{ij}(\sigma) \coloneqq \frac{H^{(\{i,j\})}(\sigma) - H(\sigma)}{2} = -\sum_{k \in [n] \setminus \{i,j\}} \sigma_k \sigma_i g_{ki} - \sum_{l \in [n] \setminus \{i,j\}} \sigma_l \sigma_j g_{lj}, \tag{16}$$

a difference in the Hamiltonian to a configuration two spins away from σ. We say σ is a *radius 2 local minimum* if $T_{ij} > 0$ for all $i \neq j$. We define *radius 2 local maximum* in the same way. Note that σ being radius 2 local minimum is not enough for σ to be a level 2 local minimum as it can still fail to be a level 1 local minimum. However, to make calculations easier we will focus on radius 2 local optima. They still ensure there is an extreme value at σ or one of it's neighbours and it can be determined in a linear time where exactly it is. The improvement of calculations to proper level 2 local optima should not be too hard.

3.1 Measure concentration approach

3.1.1 Logarithmic Sobolev inequlities

Before we delve deeper into local optima, we take a detour and explore concentration of functions of Gaussian random variables. Measure concentration theory is a branch of probability that explores how random variables concentrate around their mean. It generalises the notion of variance. Most of this section follows [4], a good exhaustive book on concentration of functions of independent random variables.

Perhaps the most basic concencration inequality is Markov inequality.

Theorem 3.2 (Markov inequality)**.** *Let X be a nonnegative random variable. Then, for all $t > 0$,*

$$\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}[X]}{t}.$$

Proof. Let $Y = t\mathbb{1}_{\{X \geq t\}}$. On a coupled probability space, $X \geq Y$ and hence

$$\mathbb{E}[X] \geq \mathbb{E}[Y] = t\mathbb{E}[\mathbb{1}_{\{X \geq t\}}] = t\mathbb{P}[X \geq t].$$

□

This is a rather weak inequality. We can strengthen it by simple algebraic manipulation to get exponential bounds.

Theorem 3.3 (Chernoff inequality)**.** *Let $X_1, \ldots, X_n$ be independent and identically distributed real-valued random variables. Then, for all $t > 0$,*

$$\frac{1}{n} \log \mathbb{P}\left[\sum_i X_i \geq nt\right] \leq \inf_{\lambda \geq 0} \left(\log \mathbb{E}e^{\lambda X_1} - \lambda t\right). \tag{17}$$

Proof. For all $\lambda > 0$ and $i \in [n]$, Markov inequality applied to an increasing function $f(x) = e^{\lambda x}$ gives

$$\mathbb{P}[X_i \geq t] \leq e^{-\lambda t}\mathbb{E}e^{\lambda X_i}.$$

Using the independence of X_i's,

$$\begin{aligned}
\mathbb{P}\left[\sum_i X_i \geq nt\right] &\leq \mathbb{P}[X_i \geq t \ \ \forall\, i] \\
&= \mathbb{P}[X_1 \geq t]^n \\
&\leq \left(e^{-\lambda t}\mathbb{E}e^{\lambda X_1}\right)^n.
\end{aligned}$$

Taking logarithms and optimising over $\lambda \geq 0$ gives the result. □

The bound we obtain from Chernoff inequality is often sharp. It gives the correct rate of asymptotic decay for the average of n independent and identically distributed random variables. However, to get good estimates of more complicated functions of independent random variables we need smarter methods. There are many possible approaches. Mostly, they take into account specific properties of functions and don't generalise too well. One of the most fruitful directions to take, which can be adapted to a lot of different settings, is a class of logarithmic Sobolev inequalities. It's complicated to even define what exactly these inequalities are and we will only be touching on some techniques we need on our way to proving theorem 3.7. For more details on the topic, we refer to [2].

We start by defining an *entropy* of a non-negative integrable random variable Y. Let $\Phi(x) = x \log x$, where we extend it to $[0, \infty)$ by setting $0 \log 0 = 0$. Then,

$$\operatorname{Ent}(Y) = \mathbb{E}\Phi(Y) - \Phi(\mathbb{E}Y).$$

Note that is not the same as more common Shannon entropy, though it is closely related to it. See Chapter 4 of [4] for more details. By Jensen's inequality, $\operatorname{Ent}(Y) \geq 0$. The entropy $\operatorname{Ent}(Y)$ is finite if and only if $\Phi(Y)$ is integrable.

For $X = (X_1, \ldots, X_n)$ and $i \in [n]$, denote by $\mathbb{E}^{(i)}$ the conditional expectation

$$\mathbb{E}^{(i)} := \mathbb{E}[\,\cdot \mid (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)]$$

and define the corresponding conditional entropy by

$$\operatorname{Ent}^{(i)}(Y) = \mathbb{E}^{(i)}\Phi(Y) - \Phi\left(\mathbb{E}^{(i)}Y\right).$$

One of the most important properties of the entropy is its sub-additivity. The proof of the following theorem can be found on page 105 of [4].

Theorem 3.4. *Let $X_1, \ldots, X_n$ be independent random variables and let $Z = f(X_1, \ldots, X_n)$, for f a measurable and nonnegative function such that Z and $\Phi(Z)$ are both integrable. Then*

$$\operatorname{Ent}(Z) \leq 2\sum_{i}^{n} \mathbb{E}\left[\operatorname{Ent}^{(i)}(Z)\right].$$

For any function $f : \mathbb{R}^n \to \mathbb{R}$ we introduce another functional

$$\varepsilon(f) := \frac{1}{2}\mathbb{E}\left[\sum_{i}^{n}\left(f(X) - f\left(\tilde{X}^{(i)}\right)\right)^2\right],$$

where $\tilde{X}^{(i)}$ is obtained by resampling the ith component of X. This functional usually appears in the context of Efron-Stein-type inequalities and has deep implications. Similarly, entropy is widely used all over measure concentration. For us, these functionals are just a stepping stone. Now we state our first Sobolev-type inequality.

Theorem 3.5. *Let $X = (X_1, \ldots, X_n)$ be a uniformly distributed random variable on $\Sigma = \{\pm 1\}^n$ and let $f : \Sigma \to \mathbb{R}$ be any function. Then*

$$\operatorname{Ent}\left(f^2(X)\right) \leq \varepsilon(f(X)).$$

Proof. Write $Z = f(X)$. The discrete version of theorem 3.4 states

$$\operatorname{Ent}\left(Z^2\right) \leq \sum_{i} \mathbb{E}\left[\operatorname{Ent}^{(i)}\left(Z^2\right)\right].$$

Hence, if we can show that for all $i \in [n]$,

$$\mathrm{Ent}^{(i)}\left(Z^2\right) \leq \mathbb{E}^{(i)}\left[\left(f(X) - f\left(\tilde{X}^{(i)}\right)\right)^2\right],$$

the result follows upon taking expectation. Note that each $X_i = \pm 1$, so conditioned on $X^{(i)}$, Z can only take two values, say a and b, with equal probability. It is then enough to prove the following inequality

$$\frac{a^2}{2}\log a^2 + \frac{b^2}{2}\log b^2 - \frac{a^2+b^2}{2}\log\frac{a^2+b^2}{2} \leq \frac{1}{2}(a-b)^2,$$

for any $a, b \in \mathbb{R}$, which follows from simple calculus arguments. □

Now we have all the tools to prove the key theorem of this section

Theorem 3.6. *Let $N = (N_1, ...N_n) \sim \mathcal{N}(0, I_n)$ be a standard multivariate normal random variable and let $f : \mathbb{R}^n \to \mathbb{R}$ be almost everywhere continuously differentiable function. Then*

$$\mathrm{Ent}\left(f(N)^2\right) \leq 2\mathbb{E}\left[\|\nabla f(N)\|^2\right].$$

Proof. We start by proving the result for $n = 1$, that is for N being a one-dimensional standard normal random variable, and then use entropy sub-additivity to extend it to higher dimensions. We also assume that f is a twice differentiable function with compact support. This is a big enough set of functions that we can lift our result to the set of almost everywhere continuously differentiable functions by standard density arguments using the monotone convergence theorem. Finally, we assume f is square integrable since otherwise, the proof is trivial.

For $i \in [n]$, let ϵ_i be independent Rademacher random variables, that is uniform random variables that take values in $\{\pm 1\}$. Denote their normalised sum by

$$S_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i.$$

Logarithmic Sobolev inequality on a hypercube that's described in theorem 3.5 implies

$$\begin{aligned}\mathrm{Ent}\left[f\left(S_n\right)^2\right] &\leq \mathbb{E}\left[\sum_{j=1}^{n}\left(f\left(S_n\right) - f\left(S_n + \frac{\epsilon_j' - \epsilon_j}{\sqrt{n}}\right)\right)^2\right] \\ &= \frac{1}{2}\mathbb{E}\left[\sum_{j=1}^{n}\left(f\left(S_n\right) - f\left(S_n - \frac{2\epsilon_j}{\sqrt{n}}\right)\right)^2\right].\end{aligned}$$

The supremum $K = \sup_x |f''(X)|$ of a twice differentiable function f with compact support is a finite number. Hence, using Taylor approximation, for all $j \in [n]$,

$$\left|f\left(S_n\right) - f\left(S_n - \frac{2\epsilon_j}{\sqrt{n}}\right)\right| \leq \frac{2}{\sqrt{n}}|f'(S_n)| + \frac{K}{n}.$$

and hence

$$\mathrm{Ent}\left[f\left(S_n\right)^2\right] \leq 2\mathbb{E}\left[f'(S_n)^2\right] + o\left(\frac{1}{\sqrt{n}}\right). \tag{18}$$

On the other hand, by the central limit theorem,

$$S_n \longrightarrow N \quad \text{in distribution.}$$

Convergence in distribution can be characterised by test convergence of expectation of all continuous, uniformly bounded functions. Since f^2 and f' are such functions, we have

$$\lim_{n\to\infty} \operatorname{Ent}\left[f\left(S_n\right)^2\right] = \operatorname{Ent}\left[f(N)^2\right]$$

and

$$\lim_{n\to\infty} \mathbb{E}\left[f'\left(S_n\right)\right] = \mathbb{E}\left[f'(N)\right].$$

Taking limits in equation (18) the case $n = 1$ is proven.

For $n \geq 1$, by theorem 3.4,

$$\operatorname{Ent}\left(f(N)^2\right) \leq \sum_{i=1}^{n} \mathbb{E}\left[\operatorname{Ent}^{(i)}\left(f(N)^2\right)\right].$$

Conditional on all but one coordinates, the result in case $n = 1$ implies that, for all $i \in [n]$,

$$\operatorname{Ent}^{(i)}\left(f(N)^2\right) \leq 2\mathbb{E}^{(i)}\left[(\partial_i f(N))^2\right].$$

Taking expectation and noting $\sum_{i=1}^{n} \partial_i f(N))^2 = \|\nabla f(N)\|^2$ finishes the proof.

□

At the beginning of the discussion, we improved Markov inequality by using simple algebraic manipulations. In a similar vein, if f satisfies extra requirements, we can improve theorem 3.6 to get exponential concentration bounds on a function f of a Gaussian. The approach we use is known as Herbst's argument. We have the following inequality.

Theorem 3.7. *Let $N = (N_1, ...N_n) \sim \mathcal{N}(0, I_n)$ be a standard multivariate normal random variable and let $f : \mathbb{R}^n \to \mathbb{R}$ be almost everywhere continuously differentiable function satisfying $\|\nabla f(N)\|^2 \leq af(N)$ for some constant $a \geq 0$. Then, for all $0 \leq \lambda \leq \frac{2}{a}$,*

$$\log \mathbb{E}\left[e^{\lambda f(N)}\right] \leq \lambda \mathbb{E}[f(N)]\left(1 + \frac{a\lambda}{2 - a\lambda}\right).$$

Proof. Write $Z = f(N)$. We use theorem 3.6 with $g(N) = e^{\frac{\lambda Z}{2}}$ to bound

$$\begin{aligned}
\operatorname{Ent}\left(\left(e^{\frac{\lambda Z}{2}}\right)^2\right) &\leq 2\mathbb{E}\left[\|\nabla e^{\frac{\lambda Z}{2}}\|^2\right] \\
&= \frac{\lambda^2}{2}\mathbb{E}\left[e^{\lambda Z}\|\nabla Z\|^2\right] \\
&\leq \frac{a\lambda^2}{2}\mathbb{E}\left[Ze^{\lambda Z}\right].
\end{aligned}$$

Recalling the definition of entropy the above inequality reads

$$\lambda \mathbb{E}\left[Ze^{\lambda Z}\right] - \mathbb{E}\left[e^{\lambda Z}\right]\log \mathbb{E}[e^{\lambda Z}] \leq \frac{a\lambda^2}{2}\mathbb{E}\left[Ze^{\lambda Z}\right].$$

Let $G(\lambda) = \log \mathbb{E}\left[e^{\lambda(Z - \mathbb{E}[Z])}\right]$. After some manipulation, our inequality can be written as a first order differential equation

$$\left(\frac{1}{\lambda} - \frac{a}{2}\right)G'(\lambda) - \frac{G(\lambda)}{\lambda^2} \leq \frac{a\mathbb{E}[Z]}{2}.$$

We notice

$$\left(\frac{1}{\lambda}-\frac{a}{2}\right)G'(\lambda)-\frac{G(\lambda)}{\lambda^2}=\left(\left(\frac{1}{\lambda}-\frac{a}{2}\right)G(\lambda)\right)'.$$

Since $G(0)=G'(0)=0$ and $G'(\lambda)>0$ for all $\lambda>0$, we can simply integrate the differential inequality from 0 to λ (as long as $\frac{1}{\lambda}-\frac{a}{2}>0$) to get

$$G(\lambda)\leq\frac{a\lambda^2\mathbb{E}[Z]}{2-a\lambda}\quad\text{for all}\quad\lambda\in\left[0,\frac{2}{a}\right).$$

The final form follows after some re-arranging.

□

In the next Section, we bring our attention back to local optima. We use theorem 3.7 to bound the probabilities of observing them.

3.1.2 Radius one local optima

We follow [1]. To get the probability of σ being a local minimum we need to integrate $T=(T_1,..,T_n)\in R^n$, for T_i's defined in (14), over the subspace where all its coordinates are positive. The random vector T is a multivariate centred normal random variable with covariance matrix

$$\Sigma=(n-2)I_n+\mathbf{1}_n\mathbf{1}_n^T,$$

where I_n is a n-dimensional identity matrix and $\mathbf{1}_n=(1,\dots,1)\in\mathbb{R}^n$.

We can use the Sherman-Morrison formula (A37) to get the inverse

$$\Sigma^{-1}=\frac{1}{n-2}\left(I_n-\frac{1}{2n-2}\mathbf{1}_n\mathbf{1}_n^T\right).$$

It is easy to see that the eigenvalues of Σ are $2n-2$ with multiplicity 1 and $n-2$ with multiplicity $n-1$. However, one might try to determine the eigenvalues using theorem 1 from [5] to get some practice for the radius two case.

We have

$$\mathbb{P}[\sigma\text{ is local minimum}]=\frac{1}{(2\pi)^{\frac{n}{2}}\det(\Sigma)^{\frac{1}{2}}}\int_{(0,\infty)^n}\exp\left(-\frac{x^T\Sigma^{-1}x}{2}\right)dx.\tag{19}$$

For all coordinates x_i of x in the subspace $(0,\infty)^n$, $|x_i|=x_i$, and so (19) can be further calculated as

$$\begin{aligned}\mathbb{P}[\sigma\text{ is local minimum}]&=\frac{1}{(2\pi)^{\frac{n}{2}}(2n-2)^{\frac{1}{2}}(n-2)^{\frac{n-1}{2}}}\int_{(0,\infty)^n}\exp\left(\frac{\|x\|_1^2}{2(n-2)(2n-2)}-\frac{\|x\|_2^2}{2(n-2)}\right)dx\\&=2^{-n}\frac{1}{(2\pi)^{\frac{n}{2}}(2n-2)^{\frac{1}{2}}(n-2)^{\frac{n-1}{2}}}\int_{\mathbb{R}^n}\exp\left(\frac{\|x\|_1^2}{2(n-2)(2n-2)}-\frac{\|x\|_2^2}{2(n-2)}\right)dx.\end{aligned}$$

Using the substitution $y=\frac{x}{\sqrt{n-2}}$, we get

$$\mathbb{P}[\sigma\text{ is a local minimum}]=2^{-n}\sqrt{\frac{n-2}{2n-2}}\mathbb{E}\exp\left(\frac{\|R\|_1^2}{4(n-1)}\right),$$

where $R\sim\mathcal{N}(0,I_n)$ is a standard multivariate Gaussian.

Precisely evaluating $\mathbb{E}\exp\left(\frac{\|R\|_1^2}{4(n-1)}\right)$ is difficult even in the limit. The following lemma provides us with fairly strong bounds.

Lemma 3.8. *Let R be an n-dimensional multivariate standard Gaussian. Then, for all $0 \le \lambda < \frac{1}{2n}$,*

$$\lambda\mathbb{E}\|R\|_1^2 \le \log\mathbb{E}\exp\left(\lambda\|R\|_1^2\right) \le \lambda\mathbb{E}\|R\|_1^2\left(1+\frac{2n\lambda}{1-2n\lambda}\right).$$

Proof. Since $\exp(x)$ is a convex function, the left-hand side follows directly from Jensen's inequality. For the right-hand side, we use theorem 3.7 with $f(x) = \|x\|_1^2$. The gradient is

$$(\nabla f(R))_i = 2\|R\|_1 \mathrm{sign}(R_i)$$

and so $\|\nabla f(R)\|^2 = 4nf(R)$. Using the theorem 3.7 with $a = 4n$ gives the result. □

The expectation is

$$\mathbb{E}\|R\|_1^2 = nE\left[R_1^2\right] + n(n-1)\mathbb{E}\left[|R_1|\right]^2 = n + n(n-1)\frac{2}{\pi}.$$

Plugging this in lemma 3.8 with $\lambda = \frac{1}{4(n-1)}$, we get

$$2^{-n}\sqrt{\frac{n-2}{2n-2}}\exp\left(\frac{n}{4(n-1)}+\frac{n}{2\pi}\right) \le \mathbb{P}[\sigma \text{ is a local minimum}]$$

and

$$\mathbb{P}[\sigma \text{ is a local minimum}] \le 2^{-n}\sqrt{\frac{n-2}{2n-2}}\exp\left(\left(\frac{n}{4(n-1)}+\frac{n}{2\pi}\right)\frac{2n-2}{n}\right).$$

Overall, we have proved the following theorem.

Theorem 3.9. *For every $\sigma \in \Sigma$, we have*

$$\frac{1}{2\pi} - \log 2 \le \lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}[\sigma \text{ is a local minimum}] \le \frac{1}{\pi} - \log 2.$$

The bounds are approximately $\frac{1}{2\pi} - \log 2 \approx -0.534$ and $\frac{1}{\pi} - \log 2 \approx -0.375$. In [1], Adarrio-Berry et al use large deviation theory to precisely evaluate the above probability in the limit as $n \to \infty$. See the first few chapters of [9] for a thorough introduction to this interesting field. We give a quick overview of what they do.

Large deviations of $\|R\|_1 = \sum_i |R_i|$ can be bounded by theorem 3.3 as

$$\frac{1}{n}\log\mathbb{P}\left[\|R\|_1 \ge nx\right] \le \inf_{\lambda\ge 0}\left(\log\mathbb{E}e^{\lambda|R_1|} - \lambda x\right).$$

For all $x > 0$, define

$$\mu_\star(x) := \sup_{\lambda\ge 0}\left(\lambda x - \log\mathbb{E}e^{\lambda|R_1|}\right),$$

known as the *Fenchel-Légendre transform* of $\log\mathbb{E}e^{\lambda|R_1|}$. The moment generating function can be easily calculated by completing the square to the Gaussian integral.

$$\begin{aligned}\mathbb{E}e^{\lambda|R_1|} &= \int_0^\infty e^{\lambda x}\frac{2}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\,dx\\ &= 2e^{\frac{\lambda^2}{2}}\frac{1}{\sqrt{2\pi}}\int_0^\infty e^{\frac{(x-\lambda)^2}{2}}\,dx\\ &= 2e^{\frac{\lambda^2}{2}}\Phi(\lambda).\end{aligned}$$

In [1], they give a more quantitative description of large deviations of $\|R\|_1$ by taking into account special properties of normal random variables. They show the next key lemma.

Lemma 3.10. *Using the same framework as above, for every $x \geq \sqrt{\frac{2}{\pi}}$ and every $n \in \mathbb{N}$, there exists a real number $r_n(x)$ with*

$$0 \leq r_n \leq \kappa \left(\frac{x - \sqrt{\frac{2}{\pi}}}{\sqrt{n}} + \frac{1}{n} \right),$$

for some $\kappa > 0$ independent of n and x, such that

$$\mathbb{P}\left[\|R\|_1 \geq nx\right] = e^{-(\mu_*(x)+r_n(x))n}.$$

For $F(x)$ nice enough non-decreasing function on $\mathbb{R}$, with $F \geq 0$ and $F(-\infty) = 0$, we can write

$$F(X) = \int_{-\infty}^{X} F'(t)dy = \int_{\{t \leq X\}} F'(t)dt$$

and after taking expectation and using Fubini's theorem

$$\mathbb{E}F(X) = \int_{-\infty}^{\infty} F'(t)\mathbb{P}[X \geq t]dt.$$

They use a multidimensional version of this idea together with lemma 3.10 to calculate $\mathbb{E}\exp\left(\frac{\|R\|_1^2}{4(n-1)}\right)$. After lots of bounding they arrive at the main result of the paper.

Theorem 3.11. *For every $\sigma \in \Sigma$, we have*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}[\sigma \text{ is a local minimum}] = \alpha^* - \log 2,$$

where $\alpha^* = R(v^*)$ and v^* is a unique maximiser of

$$R(x) = \frac{x^2}{4} - \mu_*(x)$$

over $x \geq \sqrt{\frac{2}{\pi}}$.

By lemma 3.8, $\frac{1}{2\pi} \leq \alpha^* \leq \frac{1}{\pi}$. Numerical evaluations show $\alpha^* \approx 0.199$. Similar calculations, also done in [1], describe the energy at a local minimum. With a high probability, the Hamiltonian is concentrated on a small interval around $\frac{v^*}{2}n^{\frac{3}{2}}$. More precisely,

Theorem 3.12. *Let $\sigma \in \Sigma$. Then there exists constants $\epsilon_0 > 0, L > 0$, and n_0 such that for all $0 < \epsilon < \epsilon_0$ and $n > n_0$,*

$$\mathbb{P}\left[-\frac{v^*}{2} - \epsilon \leq n^{-\frac{3}{2}}H(\sigma) \leq -\frac{v^*}{2} + \epsilon \mid \sigma \text{ is local minimum}\right] \geq 1 - \exp\left(L\sqrt{n} - \epsilon^2 n\right).$$

Numerical considerations show $\frac{v^*}{2} \approx 0.506$. The following corollary follows immediately from theorem 3.11.

Corollary 3.13. *Let α^* be defined as above and let L be the number of local minima. Then,*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}[L] = \alpha^*.$$

3.1.3 Radius two local optima

We try to generalise the above to radius 2 local optima. To ease the notation write $N = \binom{n}{2}$. Let

$$T = (T_{12}, T_{13}, \ldots, T_{1n}, T_{23}, \ldots, T_{2n}, \ldots T_{(n-2)n}, T_{(n-1)n}) \in \mathbb{R}^N,$$

for T_{ij} defined in (16). We define a "distance" function d that tells us how far apart are two coordinates of T. More specifically,

$$d(T_{ij}, T_{kl}) = |\{i,k\} \cap \{j,l\}|.$$

Then T has the following covariance matrix

$$C := \mathrm{Cov}(T) = \begin{cases} 2(n-2) & \text{if } d = 2, \\ n-2 & \text{if } d = 1, \\ 4 & \text{if } d = 0. \end{cases}$$

Using ansatz

$$C^{-1} = \begin{cases} a & \text{if } d = 2, \\ b & \text{if } d = 1, \\ c & \text{if } d = 0 \end{cases}$$

and solving the system of equations for different values of d, we get

$$C^{-1} = \begin{cases} a = \frac{n^4 - 14n^3 + 74n^2 - 170n + 144}{4(8-6n+n^2)^2} & \text{if } d = 2, \\ b = -\frac{n^2 - 8n + 14}{4(n-4)(n-2)^2} & \text{if } d = 1, \\ c = \frac{n^2 - 9n + 16}{2(8-6n+n^2)^2} & \text{if } d = 0. \end{cases}$$

Considering the limit behaviour we have, all as $n \to \infty$, $a \to \frac{1}{4}, b \to -\frac{1}{4n}$ and $c \to \frac{1}{2n^2}$. Let M be an $N \times N$ matrix given by $M_{Z_{ij} Z_{kl}} = 1$ if and only if $d(Z_{ij} Z_{kl}) \neq 0$. Then, we have

$$C^{-1} = c\mathbf{1}_N \mathbf{1}_N^T + (b-c)M + (a-b)\mathrm{Id}_N.$$

Square matrices of dimension $\binom{n}{k}$ that are of the form $M_{AB} = f(|A \cap B|)$, for A, B sets of cardinality k are sometimes known as the adjacency algebra of the Johnson association scheme. Their eigenvalues are well-understood and are described in theorem 1.1 of [5]. In particular, it gives the eigenvalues of C as $\lambda_1 = 4(n-2)^2$ with multiplicity 1, $\lambda_2 = (n-4)^2$ with multiplicity $(n-1)$ and $\lambda_3 = 4$ with multiplicity $\binom{n}{2} - n$. Hence, the determinant of C is

$$D := \det(C) = 4(n-2)^2 (n-4)^{(n-1)} 4^{N-n}.$$

Recall that σ is a radius 2 local minimum, denoted $LM2$, if and only if $T_{ij} > 0$ for all $i \neq j$. We calculate

$$\begin{aligned} \mathbb{P}[LM_2] &= \frac{1}{(2\pi)^{\frac{N}{2}} D^{\frac{1}{2}}} \int_{(0,\infty)^N} \exp\left(\frac{-x^T C^{-1} x}{2} \right) dx \\ &= \frac{1}{(2\pi)^{\frac{N}{2}} D^{\frac{1}{2}}} \int_{(0,\infty)^N} \exp\left(-\frac{c\|x\|_1^2}{2} - \frac{(b-c)x^T M x}{2} - \frac{(a-b)\|x\|_2^2}{2} \right) dx \end{aligned}$$

Note that in the second line, in order to get norms, we used that all coordinates are positive, hence $x_{ij} = |x_{ij}|$ for every coordinate ij. We use it once again to rewrite the middle integrand as

$$x^T M x = \sum_{i\in[n]} \left(\sum_{j\in[n]\setminus i} |x_{ij}| \right)^2 = \sum_{i\in[n]} \|x_i\|_1^2,$$

where $x_i \coloneqq \sum_{j\in[n]\setminus i} |x_{ij}|$. Then, we have

$$\begin{aligned}
\mathbb{P}[LM_2] &= \frac{1}{(2\pi)^{\frac{N}{2}} D^{\frac{1}{2}}} \int_{(0,\infty)^N} \exp\left(-\frac{c\|x\|_1^2}{2} - \frac{(b-c)x^T M x}{2} - \frac{(a-b)\|x\|_2^2}{2} \right) dx \\
&= \frac{1}{(2\pi)^{\frac{N}{2}} D^{\frac{1}{2}}} \int_{(0,\infty)^N} \exp\left(-\frac{c\|x\|_1^2}{2} - \frac{(b-c)\sum_i \|x_i\|_1^2}{2} - \frac{(a-b)\|x\|_2^2}{2} \right) dx \\
&= 2^{-N} \frac{1}{(2\pi)^{\frac{N}{2}} D^{\frac{1}{2}}} \int_{\mathbb{R}^N} \exp\left(-\frac{c\|x\|_1^2}{2} - \frac{(b-c)\sum_i \|x_i\|_1^2}{2} - \frac{(a-b)\|x\|_2^2}{2} \right) dx
\end{aligned}$$

and using the change of variables $y = x\sqrt{a-b}$, we finally get

$$\mathbb{P}[LM_2] = 2^{-N} \left(\frac{1}{(a-b)^N D} \right)^{\frac{1}{2}} \mathbb{E}\left[\exp\left(-\frac{c\|R\|_1^2}{2(a-b)} - \frac{(b-c)\sum_i \|R_i\|_1^2}{2(a-b)} \right) \right], \tag{20}$$

where $R \sim \mathcal{N}(0, \mathrm{Id}_N)$ is a standard normal in $\mathbb{R}^N$.

One can hope to evaluate (20) in a similar way as in the radius 1 case. However, the author wasn't able to make much progress. We give some heuristic reasons showing why it is difficult to proceed. Every global minimum (almost surely there is only one) is also a radius 2 local minimum, thus $LM_2 \geq 1$. Due to the richness of the landscape we actually expect to have many more than just one local minimum, but ignore that for now. Hence,

$$\mathbb{P}[LM_2] = 2^{-n}\mathbb{E}[LM_2] \geq 2^{-n}.$$

On the other hand, as $n \to \infty$,

$$2^{-N} \left(\frac{1}{(a-b)^N D} \right)^{\frac{1}{2}} \longrightarrow 2^{-N} \left(\frac{1}{4n} \left(\frac{4e}{n} \right)^n \right)^{\frac{1}{2}}. \tag{21}$$

If we want $-\log \mathbb{P}[LM_2] = O(n)$, we need the big expectation term in (20) to cancel most of (21) and after that we collect the remaining $O(n)$ terms. That requires a very accurate description of expectation term and bounds similar to lemma 3.8 are not enough.

Even the more convoluted approach of [1], which uses large deviations to evaluate the corresponding expectation in radius 1 case, doesn't suggest a clear way to proceed. We might try to define

$$f(R_1, .., R_n) = -\frac{c\|R\|_1^2}{2(a-b)} - \frac{(b-c)\sum_i \|R_i\|_1^2}{2(a-b)}$$

and find a very precise quantitative way to describe large deviations of f. But then the coordinates of f are not even independent. The theory of measure concentration of functions without independence of coordinates is less developed and much more involved than its independence counterpart and the known bounds are less sharp. There are still many techniques available for non-independent case, from Markov random fields to Stein's method of exchanging pairs to name a few, so there are possible directions to take from here.

3.2 Analytical approach

We follow a more analytical approach of [28]. We make a heavy use of distributions (generalised functions), see Appendix E for a quick introduction to them. There are lots of integrals in the upcoming pages and from time to time we will exchange the order of integration. We won't give any comments on why we can do it. In each case, it's simply a matter of showing that the integral of absolute value converges and then using Fubini's theorem. As absolute values of integrands are mostly exponentials of quadratic forms this is not too hard.

3.2.1 Radius one local optima

To ease the notation we introduce the weights $(g_{ii})_i$, but we set them all to 0, i.e. $g_{ii} \equiv 0$ for all i. As already discussed, a configuration $\sigma \in \Sigma$ is a local minimum if and only if

$$T_i(\sigma) = \frac{H^{(i)}(\sigma) - H(\sigma)}{2} = -\sum_j \sigma_i \sigma_j g_{ij} > 0$$

for all $i \in [n]$. That is equivalent to the existence of a vector $\lambda = (\lambda_i)_i \in \mathbb{R}$, with $\lambda_i < 0$, such that

$$\sigma_i \sum_j \sigma_j g_{ij} = \lambda_i \tag{22}$$

for all i. On the other hand, since g_{ij}'s are all symmetric around 0, the condition for σ being a local maximum is the existence of $\lambda = (\lambda_i)_i \in \mathbb{R}$, with $\lambda_i > 0$, satisfying 22. Note that all σ's are equally likely to be global minima and each σ is as likely to be a local minimum as it is to be a local maximum.

Let $G = (g_{ij})_{i \leq j}$ be a vector of Gaussian weights. Let $\delta(x)$ denote the *delta function*. We treat it as a distribution. We start calculating

$$\begin{aligned}
\mathbb{P}[\sigma \text{ is a local minimum}] &= \mathbb{P}\left[(1,\dots,1)^T \text{ is a local maximum}\right] \\
&= \mathbb{E}\left[\mathbb{1}_{\{(1,\dots,1)^T \text{ is a local maximum}\}}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\{(1,\dots,1)^T \text{ is a local maximum}\}} \mid G\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\int_{(0,\infty)^n} \prod_i \delta\left(\sum_j g_{ij} - \lambda_i\right) d\lambda_i \,\middle|\, G\right]\right].
\end{aligned}$$

Replacing the delta function with its integral representation (B.5), we get

$$\mathbb{E}\left[\mathbb{E}\left[\int_{(0,\infty)^n} \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} \exp\left(-\mathrm{i}\sum_i \lambda_i + \mathrm{i}\sum_i \phi_i \sum_j g_{ij}\right) \prod_i d\phi_i \prod_i d\lambda_i \,\middle|\, G\right]\right]$$

and after integrating with respect to normal measure,

$$\frac{1}{(2\pi)^{\frac{D}{2}}} \int_{\mathbb{R}^D} \int_{(0,\infty)^n} \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} \exp\left(-\mathrm{i}\sum_i \lambda_i + \mathrm{i}\sum_i \phi_i \sum_j g_{ij}\right) \prod_i d\phi_i \prod_i d\lambda_i \prod_{i \leq j} \exp\left(-\frac{g_{ij}^2}{2}\right) dg_{ij},$$

where $D = \frac{n(n+1)}{2}$ is the dimension of G. After exchanging the order of integration, integral identity (A38) implies this is equal to

$$\frac{1}{(2\pi)^n} \int_{(0,\infty)^n} \int_{\mathbb{R}^n} \exp\left(-\mathrm{i}\sum_i \lambda_i \phi_i - \frac{1}{2}\sum_{i<j}(\phi_i + \phi_j)^2\right) \prod_i d\phi_i \prod_i d\lambda_i$$

With a change of variables $\phi_i = \sqrt{n-1}\,\phi_i$ and $\lambda_i = \frac{\lambda_i}{\sqrt{n-1}}$ this can be rewritten as

$$\frac{1}{(2\pi)^n}\int_{(0,\infty)^n}\int_{\mathbb{R}^n}\exp\left(-\mathrm{i}\sum_i \lambda_i\phi_i - \frac{1}{2}\sum_i \phi_i^2 - \frac{1}{n-1}\sum_{i<j}\phi_i\phi_j\right)\prod_i d\phi_i \prod_i d\lambda_i. \tag{23}$$

The main difficulty in evaluating the above integral is its restriction to the intersection of half-spaces. To this end, we introduce the well-known *Heaviside function*, treated as a distribution, given by $H(x) = \mathbb{1}_{\{x>0\}}$. Then (23) is equal to

$$\frac{1}{(2\pi)^n}\int_{\mathbb{R}^n}\int_{\mathbb{R}^n}\exp\left(-\mathrm{i}\sum_i \lambda_i\phi_i - \frac{1}{2}\sum_i \phi_i^2 - \frac{1}{n-1}\sum_{i<j}\phi_i\phi_j\right)\prod_i d\phi_i \prod_i H(\lambda_i) d\lambda_i.$$

and replacing H with its integral representation (A39),

$$\frac{1}{(2\pi)^n}\int_{\mathbb{R}^n}\int_{\mathbb{R}^n}\exp\left(-\mathrm{i}\sum_i \lambda_i\phi_i - \frac{1}{2}\sum_i \phi_i^2 - \frac{1}{n-1}\sum_{i<j}\phi_i\phi_j\right)\prod_i d\phi_i \prod_i \lim_{\varepsilon_i\to 0^+}\frac{1}{2\pi\mathrm{i}}\int_{\mathbb{R}}\frac{1}{t_i - \mathrm{i}\varepsilon_i}e^{\mathrm{i}\lambda_i t_i}dt_i d\lambda_i.$$

Once again, we exchange the order of integration. Integrating with respect to λ_i's first gives back delta function, in particular $\delta(t_i - \phi_i)$. In order to make our notation more compact we introduce another distribution

$$D(\phi) \coloneqq \lim_{\varepsilon\to 0^+}\frac{1}{\pi\mathrm{i}}\frac{\exp\left(-\frac{\phi^2}{2}\right)}{\phi - \mathrm{i}\varepsilon} = \delta(\phi) + \frac{1}{\pi\mathrm{i}}PV\left(\frac{\exp\left(-\frac{\phi^2}{2}\right)}{\phi}\right), \tag{24}$$

where PV is the *Cauchy principal value*. See equation (A40) for the definition. Our manipulations so far result in the following expression

$$\mathbb{P}[\sigma \text{ is a local minimum}] = \frac{1}{2^n}\int_{\mathbb{R}^n}\exp\left(-\frac{1}{n-1}\sum_{i<j}\phi_i\phi_j\right)\prod_i D(\phi_i)d\phi_i. \tag{25}$$

Such integrals, where we integrate a function of spin configurations against a (complex) spin weight function, are common in different spin models. There are series expansions in terms of cumulants available that can be informative, for example, see Section 3 of [28]. However, we take a more direct approach. Rewrite (25) as

$$\begin{aligned}\mathbb{P}[\sigma \text{ is a local minimum}] &= \frac{1}{2^n}\int_{\mathbb{R}^n}\exp\left(-\frac{1}{2(n-1)}\sum_{i\neq j}\phi_i\phi_j\right)\prod_i D(\phi_i)d\phi_i \\ &= \frac{1}{2^n}\int_{\mathbb{R}^n}\exp\left(-\frac{1}{2(n-1)}\left(\left(\sum_i \phi_i\right)^2 - \sum_i \phi_i^2\right)\right)\prod_i D(\phi_i)d\phi_i.\end{aligned}$$

Two sums are of different orders of magnitude. We are interested in limit behaviour, so we ignore the smaller one. Our probability can then be approximated by

$$
\begin{aligned}
(1+o(1))\mathbb{P}[\sigma \text{ is a local minimum}] &= \frac{1}{2^n}\int_{\mathbb{R}^n} \exp\left(-\frac{1}{2n}\left(\sum_i \phi_i\right)^2\right)\prod_i D(\phi_i)d\phi_i \\
&= \frac{1}{2^n}\int_{\mathbb{R}^n}\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{n}{2}t^2 + \mathrm{i}t\sum_i \phi_i\right)dt\prod_i D(\phi_i)d\phi_i \\
&=
\end{aligned}
$$

where we used identity (A38). This factorises nicely and can be compactly written as

$$
\frac{1}{2^n\sqrt{2\pi}}\int_{\mathbb{R}} \exp\left(-nf(t)\right)dt \tag{26}
$$

for

$$
f(t) = \frac{1}{2}t^2 - \log\int_{\mathbb{R}} e^{\mathrm{i}t\phi}D(\phi)d\phi.
$$

We can calculate

$$
\begin{aligned}
\int_{\mathbb{R}} e^{\mathrm{i}t\phi}D(\phi)d\phi &= 1 + \lim_{\epsilon\to 0^+}\frac{1}{\pi i}\left(\int_{-\infty}^{-\epsilon} + \int_{\epsilon}^{\infty}\right)\frac{\exp(it\phi)}{\phi}\exp\left(-\frac{\phi^2}{2}\right)d\phi \\
&= 1 + \lim_{\epsilon\to 0^+}\frac{1}{\pi i}\int_{\epsilon}^{\infty}\frac{\exp(it\phi)-\exp(-it\phi)}{\phi}\exp\left(-\frac{\phi^2}{2}\right)d\phi \\
&= 1 + \lim_{\epsilon\to 0+}\frac{2}{\pi}\int_{\epsilon}^{\infty}\frac{\sin(t\phi)}{\phi}\exp\left(-\frac{\phi^2}{2}\right)d\phi \\
&= 1 + \frac{2}{\pi}\int_{0}^{\infty}\frac{\sin(t\phi)}{\phi}\exp\left(-\frac{\phi^2}{2}\right)d\phi.
\end{aligned}
$$

Employing the Feynman integration trick in the last line, we finally get

$$
\int_{\mathbb{R}} e^{\mathrm{i}t\phi}D(\phi)d\phi = 2\Phi(t), \tag{27}
$$

where $\Phi(t)$ is a cumulative distribution of a standard normal.

We use Laplace's method (see Appendix D) to asymptotically evaluate (26). The unique solution t_0 to

$$
\frac{df}{dt} = t - \frac{\Phi'(t)}{\Phi(t)} = 0
$$

was numerically evaluated to be $t_0 \approx 0.5061$ and $f(t_0) \approx -0.19923$. Hence,

$$
\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}[\sigma \text{ is a local minimum}] = -t_0 - \log 2 \approx -0.4939,
$$

This is consistent with Theorem 3.12.

In [28] they go a step further and use a version of the above calculations to get a distribution of energies at local minima, once more confirming the results of [1].

3.2.2 Radius two local optima

We try using the same approach to calculate radius 2 local optima. Recall the definition of

$$T_{ij}(\sigma) = \frac{H^{(\{i,j\})}(\sigma) - H(\sigma)}{2} = -\sum_{k\in[n]\setminus j} \sigma_k\sigma_i g_{ki} - \sum_{l\in[n]\setminus i} \sigma_l\sigma_j g_{lj}.$$

A configuration σ is a radius 2 local minimum if and only is

$$T_{ij}(\sigma) > 0,$$

for all $i<j$, or, equivalently, there exist a vector $\lambda = (\lambda_{ij})_{i<j} \in R^N$ with all coordinates positive, such that

$$\sum_{k\in[n]\setminus j} \sigma_k\sigma_i g_{ki} + \sum_{l\in[n]\setminus i} \sigma_l\sigma_j g_{lj} = -\lambda_{ij},$$

for all $i<j$. Hence, we can calculate

$$\begin{aligned}\mathbb{P}[\sigma \text{ is a radius 2 local minimum}] &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\{(1,\ldots,1)^T \text{ is a radius 2 local maximum}\}} \mid G\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\int_{(0,\infty)^N} \prod_{i<j} \delta\left(\sum_{k\in[n]\setminus j} g_{ki} + \sum_{l\in[n]\setminus i} g_{lj} - \lambda_{ij}\right) d\lambda_{ij} \,\middle|\, G\right]\right].\end{aligned}$$

and replacing δ function with its integral representation (B.5), we get

$$\mathbb{E}\left[\mathbb{E}\left[\int_{(0,\infty)^N} \frac{1}{(2\pi)^N} \int_{\mathbb{R}^N} \exp\left(-\mathrm{i}\sum_{i<j}\lambda_{ij} + \mathrm{i}\sum_{i<j}\phi_{ij}\left(\sum_{k\in[n]\setminus j} g_{ki} + \sum_{l\in[n]\setminus i} g_{lj}\right)\right)\prod_{i<j} d\phi_{ij}\prod_{i} d\lambda_i \,\middle|\, G\right]\right].$$

We integrate with respect to Gaussian measure. Terms g_{ij} and ϕ_{kl} appear together whenever they share exactly one of the indices. Therefore, the above is equal to

$$\frac{1}{(2\pi)^N}\int_{(0,\infty)^N}\int_{\mathbb{R}^N} \exp\left(-\mathrm{i}\sum_{i<j}\phi_{ij}\lambda_{ij} - \frac{1}{2}\sum_{i<j}\left(\sum_{\substack{k,l \\ |\{k,l\}\cap\{i,j\}|}} \phi_{kl}\right)^2\right)\prod_{i<j} d\phi_{ij}\prod_{i<j} d\lambda_{ij}$$

and after substitution $\phi_i = \sqrt{2n-3}\,\phi_i$ and $\lambda_i = \frac{\lambda_i}{\sqrt{2n-3}}$, we have

$$\frac{1}{(2\pi)^N}\int_{(0,\infty)^N}\int_{\mathbb{R}^N} \exp\left(-\mathrm{i}\sum_{i<j}\phi_{ij}\lambda_{ij} - \frac{1}{2}\sum_{i<j}\phi_{ij}^2 - \frac{1}{2n-3}\sum_{i<j} A_{ij}\right)\prod_{i<j} d\phi_{ij}\prod_{i<j} d\lambda_{ij},$$

where

$$A_{ij} = \sum_{\substack{k<l \\ k,l\neq i,j}} \phi_{ik}\phi_{il} + \sum_{\substack{k<l \\ k,l\neq i,j}} \phi_{jk}\phi_{jl} + \sum_{\substack{k\neq l \\ k,l\neq i,j}} \phi_{ik}\phi_{jl}.$$

Following the same steps as for radius 1 local optima, we first introduce the Heaviside function to expand integration limits and then use distribution D, given in equation (24), to compactly write the probability of interest as

$$\mathbb{P}[\sigma \text{ is a local minimum of radius 2}] = \frac{1}{2^N}\int_{\mathbb{R}^N} \exp\left(-\frac{1}{4(n-3)}\sum_{i\neq j} A_{ij}\right)\prod_{i<j} D(\phi_{ij})d\phi_{ij}. \tag{28}$$

Let's take a deeper look at $\sum_{i\neq j} A_{ij}$. For i,j,k,l all distinct, the term $\phi_{ik}\phi_{jl}$ appears once in every A_{xy} for $\{x,y\}\subset\{i,j,k,l\}$. There are 6 such subsets, each appearing twice, resulting in a factor of 12. For i,k,l distinct, the term $\phi_{ik}\phi_{il}$ appears once in all A_{ix} with $x\neq k,l$. That's $2(n-3)$ appearances in total. Hence,

$$\sum_{i\neq j} A_{ij} = 6\left(\left(\sum_{i<j}\phi_{ij}\right)^2 - \sum_{i<j}\phi_{ij}^2\right) + (2(n-3)-6)\sum_i\left(\left(\sum_k \phi_{ik}\right)^2 - \sum_k \phi_{ik}^2\right).$$

We approximate A_{ij} to a leading term to write $(1+o(1))\mathbb{P}[\sigma$ is a local minimum of radius 2] as

$$\frac{1}{2^N}\int_{\mathbb{R}^N}\exp\left(-\frac{1}{2n}\left(3\left(\sum_{i<j}\phi_{ij}\right)^2 + n\sum_i\left(\sum_k\phi_{ik}\right)^2\right)\right)\prod_{i<j}D(\phi_{ij})d\phi_{ij}$$

and after referencing identity (A38),

$$\frac{1}{2^N}\frac{1}{(2\pi)^{\frac{n+1}{2}}}\int_{\mathbb{R}^N}\int_{\mathbb{R}^n}\int_{\mathbb{R}}\exp\left(-\frac{n}{6}t^2-\sum_i s_i^2+\mathrm{i}t\sum_{i<j}\phi_{ij}+\mathrm{i}\sum_i s_i\sum_k\phi_{ik}\right)dt\prod_i ds_i\prod_{i<j}D(\phi_{ij})d\phi_{ij}. \quad (29)$$

Multiple uses of equation (27) give that (29) is equal to

$$\frac{1}{(2\pi)^{\frac{n+1}{2}}}\int_{\mathbb{R}^n}\int_{\mathbb{R}}\exp\left(-\frac{n}{6}t^2-\sum_i s_i^2+\sum_{i<j}\log\Phi(t+s_i+s_j)\right)dt\prod_i ds_i. \quad (30)$$

We try to approximate (30) by once again using Laplace's method. Let

$$f(t,s_1,\dots,s_n) = -\frac{n}{6}t^2-\sum_i s_i^2+\sum_{i<j}\log\Phi(t+s_i+s_j).$$

We have

$$\frac{\partial f}{\partial t} = -\frac{n}{3}t+\sum_{i<j}\frac{\Phi'(t+s_i+s_j)}{\Phi(t+s_i+s_j)}$$

and, for all i,

$$\frac{\partial f}{\partial s_i} = -2s_i+\sum_{j\neq i}\frac{\Phi'(t+s_i+s_j)}{\Phi(t+s_i+s_j)}.$$

Due to a symmetry, a global minimum of f is attained when all s_i's are equal to some $s>0$. Note that s implicitly depends on n. Then

$$-\frac{n}{3}t+\sum_{i<j}\frac{\Phi'(t+2s)}{\Phi(t+2s)} = -2s+\sum_{j\neq i}\frac{\Phi'(t+2s)}{\Phi(t+2s)} = 0.$$

Hence, $t=3s$ and $s=s_n$ is a fixed point of

$$g_n(x) = \frac{n-1}{2}\frac{\Phi'(5x)}{\Phi(5x)}.$$

This equation can't be solved analytically. Numerical considerations show that, although s_n is growing really slowly, we still have $n=o(s_n)$. Laplace's method then suggests that, asymptotically,

$$\mathbb{P}[\sigma \text{ is a local minimum of radius } 2] \propto e^{-\frac{5}{2}s_n^2} \cdot \Phi(5s_n)^{\binom{n}{2}},$$

which decays too quickly as $n \to \infty$. The richness of the Hamiltonian landscape suggests we should have more than one radius 2 local minimum and so it seems that Laplace's method fails.

That's not too surprising. Laplace's method tries to approximate the integrals of the form

$$\int_{\mathbb{R}^p} e^{nf(x)} dx,$$

for $x \in \mathbb{R}^p$. The most basic version is concerned with a framework where the number of dimensions p is fixed. The situation gets much more complicated when p grows with n. As shown in [23], the basic Laplace's method is still valid as long as $p = o(n^{1/3})$, but the situation gets much more delicate for larger p and sometimes non-trivial correction terms are needed. In [23], there is an example for $p = n^{\frac{1}{2}}$. Authors of [23] also express disbelief in the existence of a version of Laplace's method, for regimes when p is big compared to n, that would work for all functions f. They also give a cumulant expansion that can be, under some extra conditions on f, used to show that Laplace's method still works. See also [19].

Another idea could be to first exchange the order of integration in (30) to get

$$\frac{1}{(2\pi)^{\frac{n+1}{2}}} \int_{\mathbb{R}} \exp\left(-\frac{n}{6}t^2\right) \left(\int_{\mathbb{R}^n} \exp\left(-\sum_i s_i^2 + \sum_{i<j} \log \Phi(t + s_i + s_j) \right) \prod_i ds_i \right) dt.$$

The most of the mass comes from the region where $t + s_i + s_j > 0$ for all i, j. Call this region S. Hence, we might try lower bounding (30) by

$$\frac{1}{(2\pi)^{\frac{n+1}{2}}} \int_{\mathbb{R}} \exp\left(-\frac{n}{6}t^2\right) \int_S \exp\left(-\sum_i s_i^2 \right) \prod_i ds_i dt.$$

The inner integral is a Gaussian integral over a certain subspace and can perhaps be very well approximated. After that we can use one-dimensional Laplace's method to evaluate outer integral.

Although the author didn't manage to generalise approaches of [1] and [28] to a radius 2 case, it is clear that these directions might be fruitful. All it's needed is stronger machinery and a bit more ingenuity.

Acknowledgements

I would like to thank my supervisors Prof. Dr. Afonso Bandeira from ETH Zurich and Dr. Dmitriy Kunisky from Yale University for helping me throughout this project. They offered numerous great insights and explanations. I would also like to thank various friends for ideas and stimulating discussions. I am grateful to have my studies sponsored by Public Scholarship, Development, Disability and Maintenance Fund of the Republic of Slovenia through Ad Futura Scholarship.

References

[1] Louigi Addario-Berry, Luc Devroye, Gábor Lugosi, and Roberto I Oliveira, *Local optima of the sherrington-kirkpatrick hamiltonian*, Journal of Mathematical Physics **60** (2019), no. 4, 043301.

[2] Cécile Ané, Sébastien Blachère, Djalil Chafaï, Pierre Fougères, Ivan Gentil, Florent Malrieu, Cyril Roberto, and Grégory Scheffer, *Sur les inégalités de sobolev logarithmiques*, Vol. 10, Société Mathématique de France Paris, 2000.

[3] Omer Angel, Sébastien Bubeck, Yuval Peres, and Fan Wei, *Local max-cut in smoothed polynomial time*, Proceedings of the 49th annual acm sigact symposium on theory of computing, 2017, pp. 429–437.

[4] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart, *Concentration inequalities: A nonasymptotic theory of independence*, Oxford university press, 2013.

[5] Amanda Burcroff, *Johnson schemes and certain matrices with integral eigenvalues*, University of Michigan, Tech. Rep (2017).

[6] Seok-Ho Chang, Pamela C Cosman, and Laurence B Milstein, *Chernoff-type bounds for the gaussian error function*, IEEE Transactions on Communications **59** (2011), no. 11, 2939–2944.

[7] Andrea Crisanti and Tommaso Rizzo, *Analysis of the-replica symmetry breaking solution of the sherrington-kirkpatrick model*, Physical Review E **65** (2002), no. 4, 046137.

[8] Nicolaas Govert De Bruijn, *Asymptotic methods in analysis*, Vol. 4, Courier Corporation, 1981.

[9] Amir Dembo, *Large deviations techniques and applications*, Springer, 2009.

[10] Júlia Domingo, Pablo Baeza-Centurion, and Ben Lehner, *The causes and consequences of genetic interactions (epistasis)*, Annual review of genomics and human genetics **20** (2019), 433–460.

[11] Morris L Eaton, *Multivariate statistics: a vector space approach.*, JOHN WILEY & SONS, INC., 605 THIRD AVE., NEW YORK, NY 10158, USA, 1983, 512 (1983).

[12] Michael Etscheid and Heiko Röglin, *Smoothed analysis of local search for the maximum-cut problem*, ACM Transactions on Algorithms (TALG) **13** (2017), no. 2, 1–12.

[13] Sacha Friedli and Yvan Velenik, *Statistical mechanics of lattice systems: A concrete mathematical introduction*, Cambridge University Press, 2017.

[14] Izrail Solomonovich Gradshteyn and Iosif Moiseevich Ryzhik, *Table of integrals, series, and products*, Academic press, 2014.

[15] Francesco Guerra, *Broken replica symmetry bounds in the mean field spin glass model*, Communications in mathematical physics **233** (2003), 1–12.

[16] Marc Mezard and Andrea Montanari, *Information, physics, and computation*, Oxford University Press, 2009.

[17] Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro, *Spin glass theory and beyond: An introduction to the replica method and its applications*, Vol. 9, World Scientific Publishing Company, 1987.

[18] Edward W Ng and Murray Geller, *A table of integrals of the error functions*, Journal of Research of the National Bureau of Standards B **73** (1969), no. 1, 1–20.

[19] Helen Ogden, *On the error in laplace approximations of high-dimensional integrals*, Stat **10** (2021), no. 1, e380.

[20] Giorgio Parisi, *Infinite number of order parameters for spin-glasses*, Physical Review Letters **43** (1979), no. 23, 1754.

[21] ______, *Field theory, disorder and simulations*, Vol. 49, World Scientific, 1992.

[22] Kaare Brandt Petersen, Michael Syskind Pedersen, et al., *The matrix cookbook*, Technical University of Denmark **7** (2008), no. 15, 510.

[23] Zhenming Shun and Peter McCullagh, *Laplace approximation of high dimensional integrals*, Journal of the Royal Statistical Society: Series B (Methodological) **57** (1995), no. 4, 749–760.

[24] Robert S Strichartz, *A guide to distribution theory and fourier transforms*, World Scientific Publishing Company, 2003.

[25] Michel Talagrand, *The parisi formula*, Annals of mathematics (2006), 221–263.

[26] ______, *Mean field models for spin glasses: Volume i: Basic examples*, Vol. 54, Springer Science & Business Media, 2010.

[27] ______, *Mean field models for spin glasses: Volume ii: Advanced replica-symmetry and low temperature*, Vol. 55, Springer Science & Business Media, 2011.

[28] F Tanaka and SF Edwards, *Analytic theory of the ground state properties of a spin glass. i. ising spin glass*, Journal of Physics F: Metal Physics **10** (1980), no. 12, 2769.

[29] Edward D Weinberger, *Local properties of kauffman's n-k model: A tunably rugged energy landscape*, Physical review A **44** (1991), no. 10, 6399.

[30] Peter Woit, *Notes on distributions* (2020).

A Normal random variables

A one-dimensional random variable X has a *normal distribution* (also called *Gaussian*) with mean μ and variance τ^2, if its probability density function is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{(x-\mu)^2}{2\tau^2}}.$$

We denote it by $X \sim \mathcal{N}(\mu, \tau^2)$. If $\mu = 0$ we say X is *centred* and if, in addition, $\tau^2 = 1$ we say X is a *standard normal (Gaussian) random variable.*

The cumulative distribution of normal random variable appears in a variety of seemingly unrelated mathematical areas. There exists no closed form expression with elementary functions. For $X \sim \mathcal{N}(\mu, \tau^2)$, the cumulative distribution of X is usually expressed in terms of an *error function* as

$$F_X(x) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x-\mu}{\sqrt{2\tau^2}}\right)\right),$$

where for all $x \in \mathbb{R}$,

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt.$$

A Taylor series of an error function is

$$\operatorname{erf}(x) \coloneqq \frac{2}{\sqrt{\pi}} \sum_{i=0}^{\infty} \frac{(-1)^i x^{2i+1}}{i!(2i+1)}. \tag{A31}$$

Sometimes, it is more convenient to use a *complementary error function*

$$\operatorname{erfc}(x) \coloneqq 1 - \operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-t^2} dt, \quad \forall x > 0,$$

which behaves similarly as exponential with quadratic decay. In particular, for all $x > 0$, we have Chernoff-type bounds,

$$\sqrt{\frac{e}{2\pi}} e^{-2x^2} \leq \operatorname{erfc}(x) \leq e^{-x^2}, \tag{A32}$$

as can be seen in [6]. A cumulative distribution of a standard normal random variable is usually denoted by $\Phi(x)$.

One of the reasons normal random variables are so nice is that they are closed under linear combinations as long as they are independent. If $X \sim \mathcal{N}(\mu_X, \tau_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \tau_Y^2)$ are independent, then for any $a, b \in \mathbb{R}$,

$$aX + bY \sim \mathcal{N}(a\mu_X + b\mu_Y, \tau_X^2 + \tau_Y^2).$$

This is no longer true without the independence assumption. In that case, the linear combination is not even guaranteed to be a normal random variable again.

In the thesis, we also come across an absolute value of a centred normal random variable, that is $Y = |X|$ for $X \sim \mathcal{N}(0, \tau^2)$. This is sometimes called *half-normal* distribution. Its density is

$$f_Y(x) = \sqrt{\frac{2}{\pi\tau^2}} e^{-\frac{x^2}{2\tau^2}}$$

and its cumulative distribution is

$$F_Y(x) = \text{erf}\left(\frac{x}{\sqrt{2\tau^2}}\right).$$

Let $a, b \in \mathbb{R}$, with $a > 0$. For a Gaussian $X \sim \mathcal{N}(\mu, 1)$, we want to calculate the cumulative distribution function of $W := -c\sqrt{n}X + b \mid X < 0 \stackrel{d}{=} -c\sqrt{n}X + b \mid -c\sqrt{n}X + b > b$. Define $\xi = x - \mu$ and $\alpha = b - \mu$. Then

$$\begin{aligned}
F_W(x) &= \frac{\Phi(\xi) - \Phi(\alpha)}{1 - \Phi(\alpha)} \\
&= \frac{\text{erf}\left(\frac{\xi}{\sqrt{2}}\right) - \text{erf}\left(\frac{\alpha}{\sqrt{2}}\right)}{1 - \text{erf}\left(\frac{\alpha}{\sqrt{2}}\right)} \\
&= \frac{\text{erf}\left(\frac{\frac{x-b}{c\sqrt{n}} + \mu}{\sqrt{2}}\right) - \text{erf}\left(\frac{\mu}{\sqrt{2}}\right)}{1 - \text{erf}\left(\frac{\mu}{\sqrt{2}}\right)}.
\end{aligned}$$

Taylor expanding *erf* around μ we get

$$F_W(x) \propto \frac{(x-b)e^{-\frac{\mu^2}{2}}}{\sqrt{n}\left(1 - \text{erf}\left(\frac{\mu}{\sqrt{2}}\right)\right)} + o\left(\frac{e^{-\frac{\mu^2}{2}}}{\sqrt{n}}\right). \tag{A33}$$

As described in the following lemma we have good control over the concentration of a normal random variable. The following lemma bounds the probability of a Gaussian attaining high values, sometimes called *Gaussian tails.*

Lemma A.1. *Let $X \sim \mathcal{N}(0, 1)$ and $t \geq 0$. Then,*

$$\mathbb{P}[X \geq t] \leq \frac{1}{\sqrt{2\pi}} e^{-t^2}.$$

Proof. An elementary calculation gives

$$\begin{aligned}
\mathbb{P}[X \geq t] &= \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-x^2/2} dx \\
&= \frac{e^{-t^2/2}}{\sqrt{2\pi}} \int_0^\infty e^{-tx} e^{-x^2} dx \\
&\leq \frac{e^{-t^2/2}}{\sqrt{2\pi}} \int_0^\infty e^{-tx} dx \\
&= \frac{1}{\sqrt{2\pi}} e^{-t^2}.
\end{aligned}$$

□

A higher-dimensional generalisation, called *multivariate normal*, is a random variable $X \in \mathbb{R}^n, n \geq 1$, with a probability density function

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x - \mu)\right),$$

where μ is a mean and variance Σ is positive definite. We write $X \sim \mathcal{N}(\mu, \Sigma)$. Once again X behaves nicely under linear transformations. More precisely, for any n-dimensional matrix A and $b \in \mathbb{R}^n$, we have

$$AX + b \sim \mathcal{N}(A\mu + b, A\Sigma A^T).$$

If we partition a n-dimensional multivariate normal $X = (X_1, X_2)^T$, with $X_1 \in \mathbb{R}$ and $X_2 \in \mathbb{R}^{n-1}$, and accordingly partition it's mean $\mu = (\mu_1, \mu_2)^T$ and covariance

$$\Sigma = \begin{bmatrix} \tau^2 & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

then the conditional distribution $X_2 \mid X_1 = h$ is a multivariate normal $\mathcal{N}(\mu(h), \Sigma')$ with mean

$$\mu(h) := \mu_2 + \frac{1}{\tau^2}\Sigma_{21}(h - \mu_1)$$

and covariance matrix

$$\Sigma' := \Sigma_{22} - \frac{1}{\tau^2}\Sigma_{21}\Sigma_{12}.$$

Somewhat surprisingly the covariance matrix changes, but doesn't depend on h. A proof can be found in [11].

B Integral and other identities

Here is a collection of various identities we use in the thesis.

Lemma B.1. *For $a, b \in \mathbb{R}$, with $a > 0$,*

$$\int_{-\infty}^{\infty} x^2 \exp(-(ax^2 + bx))dx = \frac{\sqrt{\pi}(2a + b^2)}{4a^{\frac{5}{2}}} \exp\left(\frac{b^2}{4a}\right). \tag{A34}$$

Proof. See 3.321 in [14]. □

Lemma B.2. *For all $a, b \in \mathbb{R}$,*

$$\int_0^{\infty} \operatorname{erfc}(bx) \operatorname{erfc}(ax)dx = \frac{1}{ab\sqrt{\pi}}(a + b - \sqrt{a^2 + b^2}). \tag{A35}$$

Proof. Reference is equation 4.7.2 in [18]. □

Lemma B.3. *For all $a, b \in \mathbb{R}$,*

$$\int_0^{\infty} \operatorname{erf}(bx) \operatorname{erfc}(ax)dx = \frac{1}{b\sqrt{\pi}}\left(\frac{\sqrt{a^2 + b^2}}{a} - 1\right). \tag{A36}$$

Proof. Reference is equation 4.7.1 in [18]. □

Lemma B.4 (Sherman-Morrison)**.** *Let $A \in GL_n$ be a n-dimensional invertible matrix and let $b, c \in R^n$. Then,*

$$(A + bc^T)^{-1} = A^{-1} - \frac{A^{-1}bc^T A^{-1}}{1 + c^t A^{-1}b}. \tag{A37}$$

Proof. The identity can be found as 3.2.4 in [22]. □

Lemma B.5 (Integral representation of Dirac delta function)**.** *Viewing both sides as distributions,*

$$\delta(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{\mathrm{i}xt} dt.$$

Proof. See a discussion in [24]. ☐

Lemma B.6. *For any* $a \in \mathbb{R}$*,*

$$\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{\mathrm{i}ax - \frac{x^2}{2}} dx = e^{-\frac{a^2}{2}}. \tag{A38}$$

Proof. Completing the square, we have

$$\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{\mathrm{i}ax - \frac{x^2}{2}} dx = e^{-\frac{a^2}{2}} \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\frac{(\mathrm{i}a - x)^2}{2}} dx.$$

Writing

$$f(a) = \int_{\mathbb{R}} e^{-\frac{(\mathrm{i}a-x)^2}{2}} dx,$$

the integrand is nice enough that we can swap the order of integration and differentiation to get

$$\begin{aligned}\frac{df}{da} &= \int_{\mathbb{R}} \frac{\partial}{\partial a} e^{-\frac{(\mathrm{i}a-x)^2}{2}} dx \\ &= \mathrm{i} \int_{\mathbb{R}} \frac{\partial}{\partial x} e^{-\frac{(\mathrm{i}a-x)^2}{2}} dx \\ &= ie^{-\frac{(\mathrm{i}a-x)^2}{2}} |_{-\infty}^{+\infty} \\ &= 0.\end{aligned}$$

Hence, for all a, $f(a) = f(0)$. However, $f(0)$ is just the usual Gaussian integral equal to 1. ☐

Lemma B.7 (Integral representation of Heaviside function)**.** *Viewing both sides as distributions,*

$$H(x) = \lim_{\varepsilon \to 0^+} \frac{1}{2\pi \mathrm{i}} \int_{\mathbb{R}} \frac{1}{t - \mathrm{i}\varepsilon} e^{\mathrm{i}xt} dt. \tag{A39}$$

Proof. See a discussion in [24]. ☐

C Epistasis calculations

Upper bounding level 2 epistasis

We bound

$$I_1 = \int_0^{\sqrt{\frac{n\pi}{8}}} \mathrm{erf}\left(\frac{x}{\sqrt{n}}\right) \mathrm{erf}\left(\frac{x\sqrt{2}}{\sqrt{n}}\right) \exp\left(-\frac{(x-c)^2}{2}\right) dx$$

and

$$I_2 = \int_{\sqrt{\frac{n\pi}{8}}}^{\infty} \mathrm{erf}\left(\frac{x}{\sqrt{n}}\right) \mathrm{erf}\left(\frac{x\sqrt{2}}{\sqrt{n}}\right) \exp\left(-\frac{(x-c)^2}{2}\right) dx.$$

For I_1, we truncate the Taylor series of an error function (A31) at a linear term, getting an inequality

$$\operatorname{erf}(x) \leq \min\left(\frac{2}{\sqrt{\pi}}x, 1\right).$$

The smaller the x, the better the bound. We use it with n in the denominator, so it strength growths as n increases. We have

$$\begin{aligned} I_1 &\leq \int_0^{\sqrt{\frac{n\pi}{8}}} \frac{2}{\sqrt{\pi}} \frac{x}{\sqrt{n}} \frac{2}{\sqrt{\pi}} \frac{x\sqrt{2}}{\sqrt{n}} \exp\left(-\frac{(x-c)^2}{2}\right) dx \\ &= \frac{4\sqrt{2}}{\pi n} \int_0^{\sqrt{\frac{n\pi}{8}}} x^2 \exp\left(-\frac{(x-c)^2}{2}\right) dx. \end{aligned}$$

Using the identity (A34), we can calculate

$$\int_{-\infty}^{\infty} x^2 \exp\left(-\frac{(x-c)^2}{2}\right) dx = \sqrt{2\pi}(1+c^2).$$

The function $h_c(x) = x^2 \exp\left(-\frac{(x-c)^2}{2}\right)$ is symmetric around c (in the sense $h_c(x) = h_c(c-x)$ for all x). In particular, it is an even function for $c = 0$. Furthermore, if we let $\epsilon_c = \frac{1}{2}\left(\int_0^\infty h_c(x)dx - \int_{-\infty}^0 h_c(x)dx\right)$, then $\epsilon_c \to 0$ as $c \to 0$.

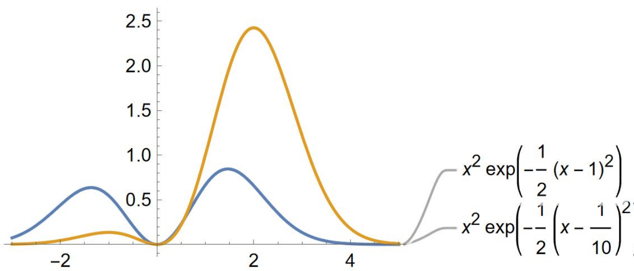

Figure 9: $h_c(x)$ for $c = 1$ and $c = \frac{1}{10}$

Hence, we can bound

$$\begin{aligned} I_1 &\leq \frac{4\sqrt{2}}{\pi n} \int_0^\infty x^2 \exp\left(-\frac{(x-c)^2}{2}\right) dx \\ &= \frac{4\sqrt{2}}{\pi n} \left(\frac{1}{2} + \epsilon_c\right) \sqrt{2\pi}(1+c^2) \\ &= \frac{8}{n\sqrt{\pi}} \left(\frac{1}{2} + \epsilon_c\right)(1+c^2). \end{aligned}$$

For the integral I_2, we have

$$
\begin{aligned}
I_2 &= \int_{\sqrt{\frac{n\pi}{8}}}^{\infty} \operatorname{erf}\left(\frac{x}{\sqrt{n}}\right) \operatorname{erf}\left(\frac{x\sqrt{2}}{\sqrt{n}}\right) \exp\left(-\frac{(x-c)^2}{2}\right) dx \\
&\le \int_{\sqrt{\frac{n\pi}{8}}}^{\infty} \exp\left(-\frac{(x-c)^2}{2}\right) dx \\
&= \sqrt{2} \int_{\sqrt{\frac{n\pi}{8}}}^{\infty} \exp\left(-(x-c)^2\right) dx
\end{aligned}
$$

and using (A32),

$$
I_2 \le \sqrt{\frac{\pi}{2}} \operatorname{erfc}\left(\sqrt{\frac{n\pi}{8}}\right) \le 2\exp\left(-\frac{\pi}{8}n\right) = o\left(\frac{1}{n}\right).
$$

Lower bounding level 2 epistasis

Using similar approach as for the upper bound, we have

$$
\begin{aligned}
\mathbb{P}(A) &= \frac{1}{2}\mathbb{P}[A \mid R_1 \ge 0] \\
&\ge \frac{1}{2}\mathbb{P}[R_3 \ge M] \\
&= \frac{1}{2}\int_0^{\infty} F_M(x) f_{R_3}(x) dx \\
&= \frac{1}{2\sqrt{2\pi}} \int_0^{\infty} \operatorname{erf}\left(\frac{x}{\sqrt{n}}\right) \operatorname{erf}\left(\frac{x\sqrt{2}}{\sqrt{n}}\right) \exp\left(-\frac{x^2}{2}\right) dx \\
&= \frac{1}{2\sqrt{2\pi}} \int_0^{\sqrt{\frac{n}{2}}} \operatorname{erf}\left(\frac{x}{\sqrt{n}}\right) \operatorname{erf}\left(\frac{x\sqrt{2}}{\sqrt{n}}\right) \exp\left(-\frac{x^2}{2}\right) dx.
\end{aligned}
$$

This time we truncate Taylor expansion of erf after a cubic term to get

$$
\operatorname{erf}(x) \ge \begin{cases} \frac{2}{\sqrt{\pi}}\left(x - \frac{x^3}{3}\right) & \text{if } x \le 1 \\ \frac{1}{2} & \text{if } x \ge 1. \end{cases}
$$

Hence, we can further bound

$$
\begin{aligned}
\mathbb{P}[A] &\ge \frac{1}{2\sqrt{2\pi}} \int_0^{\sqrt{\frac{n}{2}}} \operatorname{erf}\left(\frac{x}{\sqrt{n}}\right) \operatorname{erf}\left(\frac{x\sqrt{2}}{\sqrt{n}}\right) \exp\left(-\frac{x^2}{2}\right) dx \\
&\ge \frac{1}{2\sqrt{2\pi}} \int_0^{\sqrt{\frac{n\pi}{8}}} \frac{2}{\sqrt{\pi}}\left(\frac{x}{\sqrt{n}} - \frac{1}{3}\left(\frac{x}{\sqrt{n}}\right)^3\right) \frac{2}{\sqrt{\pi}}\left(\frac{x\sqrt{2}}{\sqrt{n}} - \frac{1}{3}\left(\frac{x\sqrt{2}}{\sqrt{n}}\right)^3\right) \exp\left(-\frac{x^2}{2}\right) dx \\
&= \frac{2}{\pi\sqrt{\pi n}} \int_0^{\sqrt{\frac{n\pi}{8}}} x^2 \exp\left(-\frac{x^2}{2}\right) dx + o(n) \\
&= \frac{2}{\pi\sqrt{\pi n}} \int_0^{\infty} x^2 \exp\left(-\frac{x^2}{2}\right) dx + o(n) \\
&= \frac{\sqrt{2}}{\pi n} + o(n),
\end{aligned}
$$

where we used (A34) in the last line.

Upper bounding level 1 epistasis

We bound

$$J_1 = \int_0^\infty \int_{x-c\wedge 0}^\infty \operatorname{erfc}\left(\frac{x}{\sqrt{2}}\right) e^{-\frac{x^2}{n}} e^{-\frac{2y^2}{n-1}} dydx$$

and

$$J_2 = \int_0^\infty \int_0^{y+c} \operatorname{erfc}\left(\frac{y+c}{\sqrt{2}}\right) e^{-\frac{x^2}{n}} e^{-\frac{2y^2}{n-1}} dxdy.$$

For J_1, we have

$$\begin{aligned} J_1 &= \frac{\sqrt{\pi}}{2}\sqrt{\frac{n-1}{2}} \int_0^\infty \operatorname{erfc}\left(\frac{x}{\sqrt{2}}\right) e^{-\frac{x^2}{n}} \operatorname{erfc}\left(\frac{\sqrt{2}(x-c\wedge 0)}{\sqrt{n-1}}\right) dx \\ &\le \frac{\sqrt{\pi}}{2}\sqrt{\frac{n-1}{2}} \left(\int_0^\infty \operatorname{erfc}\left(\frac{x}{\sqrt{2}}\right) \operatorname{erfc}\left(\frac{\sqrt{2}x}{\sqrt{n-1}}\right) dx + c\cdot \operatorname{erfc}^2(0)\right). \end{aligned}$$

With the help of (A35), combined with Taylor expansion of $\sqrt{1+x} = 1 + \frac{x}{2} - \frac{x^2}{8}\ldots$, we can evaluate this to get

$$\begin{aligned} J_1 &\le \frac{\sqrt{\pi}}{2}\sqrt{\frac{n-1}{2}} \left(\sqrt{\frac{n-1}{\pi}} \left(\sqrt{\frac{1}{2}} + \sqrt{\frac{2}{n-1}} - \sqrt{\frac{1}{2}+\frac{2}{n}}\right) + c\right) \\ &= \frac{\sqrt{n}}{2}\left(1 + c\sqrt{\frac{\pi}{2}} + o(1)\right). \end{aligned}$$

We use similar calculations to bound J_2 as

$$\begin{aligned} J_2 &= \int_0^\infty \int_0^{y+c} \operatorname{erfc}\left(\frac{y+c}{\sqrt{2}}\right) e^{-\frac{x^2}{n}} e^{-\frac{2y^2}{n-1}} dxdy \\ &= \frac{\sqrt{\pi n}}{2} \int_0^\infty \operatorname{erfc}\left(\frac{y+c}{\sqrt{2}}\right) \operatorname{erf}\left(\frac{y+c}{\sqrt{n}}\right) e^{-\frac{2y^2}{n-1}} dy \\ &= (1+o(1))\frac{\sqrt{\pi n}}{2} \int_0^\infty \operatorname{erfc}\left(\frac{y+c}{\sqrt{2}}\right) \operatorname{erf}\left(\frac{y+c}{\sqrt{n}}\right) dy \\ &= (1+o(1))\frac{\sqrt{\pi n}}{2} \left(\int_0^\infty \operatorname{erfc}\left(\frac{y}{\sqrt{2}}\right) \operatorname{erf}\left(\frac{y}{\sqrt{n}}\right) dy - \int_0^c \operatorname{erfc}\left(\frac{y}{\sqrt{2}}\right) \operatorname{erf}\left(\frac{y}{\sqrt{n}}\right) dy\right) \\ &\ge (1+o(1))\frac{\sqrt{\pi n}}{2} \left(\int_0^\infty \operatorname{erfc}\left(\frac{y}{\sqrt{2}}\right) \operatorname{erf}\left(\frac{y}{\sqrt{n}}\right) dy - \int_0^c 1dy\right) \\ &= (1+o(1))\frac{\sqrt{\pi n}}{2} \left(\int_0^\infty \operatorname{erfc}\left(\frac{y}{\sqrt{2}}\right) \operatorname{erf}\left(\frac{y}{\sqrt{n}}\right) dy - c\right). \end{aligned}$$

This time we use Taylor expansion of $\sqrt{1+x}$ with (A36) to get

$$\begin{aligned} J_2 &\ge (1+o(1))\frac{\sqrt{\pi n}}{2} \left(\sqrt{\frac{n}{\pi}} \left(\sqrt{1+\frac{2}{n}} - 1\right) - c\right) \\ &= (1+(o(1))\frac{c\sqrt{\pi n}}{2}. \end{aligned}$$

Lower bounding level 1 epistasis

The calculations are almost identical to finding an upper bound, so we don't give many comments. We have

$$\begin{aligned}
\mathbb{P}[B] &\geq \mathbb{P}[N \mid 0 < R_1 < c\sqrt{n-1}] \cdot \mathbb{P}[0 < R_1 < c\sqrt{n-1}] \\
&\geq \mathbb{P}[X + c < R_3 < Y] \cdot \left(\frac{1}{2} - \mathbb{P}[R_1 > c\sqrt{n-1}]\right).
\end{aligned}$$

We write

$$\begin{aligned}
\mathbb{P}[X + c < R_3 < Y] &= \int_0^\infty \int_0^\infty (F_{R_3}(y) - F_{R_3}(x+c)) f_X(x) f_Y(y) \mathbb{1}_{\{x+c \leq y\}} dx dy \\
&= \frac{2\sqrt{2}}{\pi\sqrt{n(n-1)}}(J_1 - J_2),
\end{aligned}$$

where

$$J_1 = \int_0^\infty \int_{x+c}^\infty \operatorname{erfc}\left(\frac{x+c}{\sqrt{2}}\right) e^{-\frac{x^2}{n}} e^{-\frac{2y^2}{n-1}} dy dx$$

and

$$J_2 = \int_c^\infty \int_0^{y-c} \operatorname{erfc}\left(\frac{y}{\sqrt{2}}\right) e^{-\frac{x^2}{n}} e^{-\frac{2y^2}{n-1}} dx dy.$$

We bound each integral separately. For J_1 we have

$$\begin{aligned}
J_1 &= \int_0^\infty \int_{x+c}^\infty \operatorname{erfc}\left(\frac{x+c}{\sqrt{2}}\right) e^{-\frac{x^2}{n}} e^{-\frac{2y^2}{n-1}} dy dx \\
&= \frac{\sqrt{\pi}}{2}\sqrt{\frac{n-1}{2}} \int_0^\infty \operatorname{erfc}\left(\frac{x+c}{\sqrt{2}}\right) e^{-\frac{x^2}{n}} \operatorname{erfc}\left(\frac{\sqrt{2}(x+c)}{\sqrt{n-1}}\right) dx \\
&\geq (1+o(1))\left(\sqrt{\frac{n\pi}{8}}\left(\int_0^\infty \operatorname{erfc}\left(\frac{x}{\sqrt{2}}\right) \operatorname{erfc}\left(\frac{\sqrt{2}x}{\sqrt{n-1}}\right) dx - c \cdot \operatorname{erfc}^2(0)\right)\right) \\
&= (1+o(1))\left(\sqrt{\frac{n\pi}{8}}\left(\sqrt{\frac{n}{\pi}}\left(\frac{1}{\sqrt{2}} + \sqrt{\frac{2}{n}} - \sqrt{\frac{1}{2} + \frac{2}{n}}\right) - c\right)\right) \\
&= (1+o(1))\frac{\sqrt{n}}{2}\left(1 - c\sqrt{\frac{\pi}{2}}\right).
\end{aligned}$$

Similarly,

$$
\begin{aligned}
J_2 &= \int_c^\infty \int_0^{y-c} \operatorname{erfc}\left(\frac{y}{\sqrt{2}}\right) e^{-\frac{x^2}{n}} e^{-\frac{2y^2}{n-1}} dx dy \\
&= \frac{\sqrt{\pi n}}{2} \int_c^\infty \operatorname{erfc}\left(\frac{y}{\sqrt{2}}\right) \operatorname{erf}\left(\frac{y-c}{\sqrt{n}}\right) e^{-\frac{2y^2}{n-1}} dy \\
&\le \frac{\sqrt{\pi n}}{2} \int_0^\infty \operatorname{erfc}\left(\frac{y}{\sqrt{2}}\right) \operatorname{erf}\left(\frac{y}{\sqrt{n}}\right) dy \\
&= \frac{\sqrt{\pi n}}{2} \sqrt{\frac{n}{\pi}} \left(\sqrt{1+\frac{2}{n}} - 1\right) \\
&= \frac{1}{2}
\end{aligned}
$$

Finally, we once again recall Gaussian tail inequality and we have

$$
\begin{aligned}
\mathbb{P}[B] &\ge \mathbb{P}[X + c < R_3 < Y] \cdot \left(\frac{1}{2} - \mathbb{P}[R_1 > c\sqrt{n-1}]\right) \\
&\ge \frac{2\sqrt{2}}{\pi\sqrt{n(n-1)}} (J_1 - J_2) \left(\frac{1}{2} - \frac{1}{c\sqrt{2\pi n}} e^{-\frac{c^2 n}{2}}\right) \\
&\ge \frac{2\sqrt{2}}{\pi\sqrt{n(n-1)}} \left((1+o(1))\left(\frac{\sqrt{n}}{2} - c\sqrt{\frac{n\pi}{8}}\right) - \frac{1}{2}\right) \left(\frac{1}{2} - \frac{1}{c\sqrt{2\pi n}} e^{-\frac{c^2 n}{2}}\right)
\end{aligned}
$$

Taking c to 0 as $c = n^{-\frac{1}{3}}$, we get

$$
\mathbb{P}[B] \ge \frac{1}{\pi\sqrt{2n}} + o\left(\frac{1}{\sqrt{n}}\right).
$$

D Laplace's method

Consider the *Laplace integral*

$$
F(\lambda) = \int_{-\infty}^\infty f(x) e^{\lambda S(x)} dx
$$

for $f, S \in C^\infty$ and assume S is uniquely maximised at x_0, such that $f(x_0) \neq 0$ and $S''(x_0) \neq 0$. Such integrals are very well understood and can be easily asymptotically approximated. We do some non-rigorous calculations to give some intuition. For precise results with proofs and in a lot more generality see [8].

For $x \neq x_0$,

$$
\exp(\lambda(S(x) - S(x_0))) \to 0 \qquad \text{as } \lambda \to \infty.
$$

Hence, for λ big enough $F(\lambda)$ can be approximated by

$$
F(\lambda) \approx \int_{x_0-\delta}^{x_0+\delta} f(x) e^{\lambda S(x)} dx,
$$

for some $\delta > 0$. The sharper the peak of $S(x)$ at x_0, the better the approximation. In a neighbourhood of x_0, we can Taylor expand $S(x)$ as

$$S(x) \approx S(x_0) + \frac{1}{2}S''(x_0)(x - x_0)^2$$

and we approximate $f(x)$ by $f(x_0)$. Thus,

$$F(\lambda) \approx f(x_0)e^{\lambda S(x_0)} \int_{x_0-\delta}^{x_0+\delta} \exp\left(\lambda S''(x_0)(x - x_0)^2\right) dx.$$

This integral can be approximated, for λ big enough, in a usual way by Gaussian integral, giving

$$F(\lambda) \approx \sqrt{\frac{-2\pi}{\lambda S''(x_0)}} f(x_0)e^{\lambda S(x_0)}.$$

The exact strength of approximation depends on the particular properties of f and S. However, it is already correct to the first order term under very mild assumptions. Laplace's technique can be generalised to a much broader context. Among others to functions with multiple extreme points, complex integrals and higher dimensional integrals.

E Distributions

Distribution is a mathematical object generalising the notion of a function. Let A be a set of functions, called *test functions*. *Distribution* associated to f is a continuous linear map

$$\phi \to \langle f, \phi \rangle \in \mathbb{C}$$

for $\phi \in A$, usually of the form

$$\phi \to \int f(x)\phi(x)dx.$$

The idea is that f doesn't need to be defined pointwise, it only matters how it acts on a set of functions A. The continuity condition depends on the choice of test functions. The two usual choices for A are:

- Set $\mathcal{D}(\mathbb{R})$ of smooth real functions with compact support,
- Set $\mathcal{S}(\mathbb{R})$ of Schwartz functions (see below)

A function g is rapidly decreasing if, for all $n \in \mathbb{N}$, there exist constants M_n and x_n, such that

$$|f(x)| \leq M_n |x|^{-n}$$

for all $x \geq x_n$. A function f is *Schwartz* if f and all its partial derivatives are rapidly decreasing. It is very hard to construct noncontinuous linear functionals using Schwartz functions as test functions, so for all practical purposes, one can ignore continuity condition. We can define Fourier transform of distribution. A set of Schwartz functions is closed under Fourier transforms. When we talk about distributions, we always mean it with respect to $\mathcal{S}(\mathbb{R})$.

One of the simplest examples of a distribution is the *Dirac delta function* δ. It is defined via the integral

$$\int_{\mathbb{R}} f(x)\delta(x)dx = f(0)$$

for all test functions f. Its integral representation is given by equation (B.5).

Given a function f with a unique singularity at some value of x, say at $x = 0$, one can try to make sense of it in the following way. Define a distribution *Principal value* of f by

$$PV(f) : \phi \to \langle PV(f), \phi \rangle := \lim_{\varepsilon \to 0^+} \int_{|x|>\varepsilon} f(x)\phi(x)dx. \tag{A40}$$

For nice enough functions f, this definition makes sense. In particular, it is a sensible definition for $f(x) = \frac{1}{x}$ and one can show

$$\lim_{\varepsilon \to 0^+} \frac{1}{x - \mathrm{i}\varepsilon} = PV\left(\frac{1}{x}\right) - \mathrm{i}\pi\delta(x).$$

See [30] for more details. We recommend [24] to a reader further interested in a distribution theory.