Large deviations in the perceptron model and consequences for active learning

IOP Publishing Mach. Learn.: Sci. Technol. 2 (2021) 045001 https://doi.org/10.1088/2632-2153/abfbbb

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
5 December 2020

REVISED
3 April 2021

ACCEPTED FOR PUBLICATION
26 April 2021

PUBLISHED
15 July 2021

PAPER

Large deviations in the perceptron model and consequences for active learning

H Cui*, L Saglietti and L Zdeborová

Institute of Physics, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
* Author to whom any correspondence should be addressed.

E-mail: hugo.cui@epfl.ch

Keywords: perceptron model, large deviations, active learning

Abstract

Active learning (AL) is a branch of machine learning that deals with problems where unlabeled data is abundant yet obtaining labels is expensive. The learning algorithm has the possibility of querying a limited number of samples to obtain the corresponding labels, subsequently used for supervised learning. In this work, we consider the task of choosing the subset of samples to be labeled from a fixed finite pool of samples. We assume the pool of samples to be a random matrix and the ground truth labels to be generated by a single-layer teacher random neural network. We employ replica methods to analyze the large deviations for the accuracy achieved after supervised learning on a subset of the original pool. These large deviations then provide optimal achievable performance boundaries for any AL algorithm. We show that the optimal learning performance can be efficiently approached by simple message-passing AL algorithms. We also provide a comparison with the performance of some other popular active learning strategies.

1. Introduction

1.1. Motivation

Supervised learning consists in presenting a parametric function (often a neural network) with a series of samples (samples) and labels, and adjusting (training) the parameters (network weights) so as to match the network output with the labels as closely as possible. Active learning (AL) is concerned with choosing the most informative samples so that the training requires the least number of labeled samples to reach the same test accuracy. AL is relevant in situations where the potential set of samples is large, but obtaining the labels is expensive (computationally or otherwise). There exist many strategies for AL, see e.g. [1] for a review. In membership-based AL [2–4] the algorithm is allowed to query the label of any sample, most often one it generates itself. In stream-based AL [5] an infinite sequence of samples is presented to the learner which can decide whether or not to query its label. In pool-based AL, which is the object of the present work, the learner can only query samples that belong to a pre-existing, fixed pool of samples. It therefore needs to choose according to some strategy which samples to query so as to have the best possible test accuracy.

Pool-based AL is relevant for many machine learning applications, e.g. because not every possible input vector is of relevance. A beautiful recent application of AL is in computational chemistry [6] where a neural network is trained to predict inter-atomic potentials. In this case the pool of data is large and consists in all possible alloys, but not of arbitrary input vectors, and labelling is extremely expensive, as it demands resource-intensive *ab-initio* simulations. Consequently, only a limited number of samples can be labeled, i.e. one only possesses a certain budget for the cardinal of the training set. Another setting where a cheap large pool of input data is readily available but labelling is expensive is drug discovery [7], where given a target molecule one aims to find new compounds among the pool able to bind it. Another example would be on text classification [8–10], where labelling a text requires non-negligible human input, while a large pool of texts is readily available on the internet. Establishing efficient pool-based AL procedures in this case implies to select *a priori* the most informative data samples for labelling.

Main-stream works on AL focus on designing heuristic algorithmic strategies in a variety of settings, and analyzing the performance thereof. It is very rarely known what are the information-theoretic limitations an AL algorithm can face and hence evaluating the distance from optimality is mostly an open question. The main contribution of the present work is to provide a toy model that is at the one hand challenging for AL, and at the same time where the optimal performance of pool-based AL can be computed and heuristic algorithms hence evaluated and bench-marked against the optimal solution. To our knowledge, this is the first work to derive optimal performance results for pool-based AL procedures are computed. More specifically we study the random perceptron model [11]. The available pool of samples is assumed to be i.i.d. vectors following a normal distribution, the teacher generating the labels is taken to be also a perceptron with the vector of teacher-weights having i.i.d. normal components. We compute the large deviation function for how likely it is to find a subset of the samples that leads to a given learning accuracy. Our results are obtained through the non-rigorous yet asymptotically exact (under appropriate probabilistic assumptions on the data) replica method, from theoretical statistical physics [12]. While the presented analysis is based on the so-called replica symmetric (RS) ansatz, we also provide a stability analysis of our results against possible symmetry-breaking effects. Providing a rigorous proof of the obtained results or turning them into rigorous bounds would be a natural, and rather challenging, next step. In the algorithmic part of this work we benchmark several existing algorithms and also propose two new algorithms relying on the approximate-message-passing algorithm for estimation of the label uncertainty for yet unlabeled sample, showing that they closely achieve, in the studied cases, the relevant information-theoretic limitations.

The paper is organized as follows: the problem is defined and related work discussed in section 1.2. In section 2, we propose a measure to quantify the informativeness of given subsets of samples. In section 3, we derive the large deviation function over all possible subset choices and deduce performance boundaries that apply to any pool-based AL strategy. In section 4 we then summarize our results on the large deviations and support them with numerical simulations. In section 5, we then compare these theoretical results with the performance of existing AL algorithms and propose two new ones, based on approximate-message-passing.

1.2. Definition of the problem and related work

A natural modeling framework for analyzing learning processes and generalization properties is the so-called teacher-student (or planted) perceptron model [13], where the input samples are assumed to be random i.i.d. vectors, and the ground truth labels are assumed to be generated by a neural network (denoted as the teacher) belonging to the same hypothesis class as the student-neural-network. In this work we will restrict to single-layer neural networks (without hidden units) for which this setting was defined and studied in [14]. Specifically we collect the input vectors into a matrix $\boldsymbol{F} \in \mathbb{R}^{P\times N}$ where N is the dimension of the input space and P is the number of samples. The teacher generating the labels, called teacher perceptron, is characterized by a teacher-vector of weights $\mathbf{x}^0$ and produces the label vector $\boldsymbol{Y} \in \mathbb{R}^P$ according to $\mathbf{Y} = \mathrm{sign}(\mathbf{F}\cdot\mathbf{x}^0)$. Learning is then done using a student perceptron and consists in finding a vector x so that for the training set $\mathbf{F}$ we have as closely as possible $\mathbf{Y} = \mathrm{sign}(\mathbf{F}\cdot\mathbf{x})$. The relevant notion of error is the test accuracy (generalization error) measuring the agreement between the teacher and the student on a new sample not presented in the training set. Since both teacher and student possess the same architecture, the training process can be rephrased in terms of an inference problem (as discussed for instance in [13]): the student aims to infer the teacher weights, used to generate the labels, from the knowledge of a set of input-output associations. This scenario allows for nice geometrical insights (see for example [15]), as the generalization properties are linked to the distance in weight space between teacher and student functions. Note that, in the case of a noiseless labelling process, the teacher–student scenario guarantees that perfect training is always possible.

Active learning was previously studied in the context of the teacher–student perceptron problem. Best known is the line of work on query by committee (QBC) [4, 16, 17], dealing with the membership based AL setting, i.e. where the samples are picked one by one into the training set and can be absolutely arbitrary N-dimensional vectors. The AL is in that case more a strategy for designing the samples rather than one for selecting them smartly from an predefined set. In the original work [4] the new samples are chosen so that a committee of several student-neural-networks has the maximum possible disagreement on the new sample. The paper shows that in this way one can reach a generalization error that decreases exponentially with the size of the training set, while for a random training set the generalization error can decrease only inversely proportionally to the size of the set [15]. However, in many practical applications the possible set of samples to be picked into the training set is not arbitrarily big, e.g. not every input vector represents an encoding of a molecular structure. We hence argue that the pool-based AL, studied in the present paper, where the samples are selected from a pre-defined set is of larger relevance to many applications.

The theoretical part of this paper is presented for a generalization of the perceptron model, specifically the for the random teacher–student generalized linear models (GLM), see e.g. [18]. The teacher-student GLM setup is an inference problem over a dataset $\{\boldsymbol{F}^\mu, y^\mu\}_{\mu=1}^P$. The labels y^μ are generated from the data $\boldsymbol{F}^\mu$

through a probability distribution $P_{\rm out}(y^\mu|\boldsymbol{F}^\mu\cdot\boldsymbol{x}^0)$, where $\boldsymbol{x}^0$ is a fixed vector, generated once and for all from a prior measure $P_X(\cdot)$. Given the dataset $\{\boldsymbol{F}^\mu, y^\mu\}_{\mu=1}^P$, the goal of the inference problem is to infer the weight vector $\boldsymbol{x}^0$ that was used to generate the labels. A completely equivalent formulation of the teacher student GLM setting is to consider that the labels y^μ have been generated by a single-layer 'teacher' neural network with weight vector $\boldsymbol{x}^0$. A 'student' network sharing the exact same architecture is then trained on the resulting dataset $\{\boldsymbol{F}^\mu, y^\mu\}_{\mu=1}^P$. Since a common architecture is shared between the teacher and student network, training is in this case equivalent to the student trying to match its own weight vector with the teacher weights $\boldsymbol{x}^0$. An instance of a GLM is thus specified by a prior measure $P_X(\cdot)$ on the weights $\boldsymbol{x}$, from which the true generative model is assumed to be sampled, and an output channel measure $P_{\rm out}(\cdot|\cdot)$, defining the generative process for the labels y^μ given the pre-activations $\boldsymbol{F}^\mu\cdot\boldsymbol{x}$. In the part where results of this work are presented we focus on the prototypical case of the noiseless continuous perceptron, where $P_X(x)=e^{-\frac{x^2}{2}}/\sqrt{2\pi}$ and $P_{\rm out}(y|h)=\delta(y-\mathrm{sign}(h))$ where for example μ we have $h^\mu=\mathbf{F}^\mu\cdot\mathbf{x}$. Moreover, we will consider the setting where the learning model is matched to the generative model and thus the student has perfect knowledge of the correct form of the two above defined measures.

The pool-based AL task can now be more formally stated as follows: given a set of N-dimensional samples $\mathcal{S}=\{\boldsymbol{F}^\mu\}$ of cardinality $|\mathcal{S}|=P=\alpha N$, the goal is to select and query the labels of a subset $S\in\mathcal{S}$ of cardinality $|S|=nN$, $0<n\leqslant\alpha$, according to some AL criterion. We will refer to n as the budget of the student. The true labels are then obtained through $y^\mu\sim P_{\rm out}(y^\mu|\boldsymbol{F}^\mu\cdot\boldsymbol{x}^0)$, $\boldsymbol{x}^0\sim P_X(\boldsymbol{x}^0)$. Henceforth measures with vector arguments are understood to be products over the coordinates of the corresponding scalar measures. For technical reasons, we rely on the strong (but customary) assumption that the samples are i.i.d. Gaussian distributed, $F_i^\mu\sim\mathcal{N}(0,1)$, $\forall i\in\{1,\ldots,N\}$, $\forall\mu\in\{1,\ldots,P\}$. Note that, while this assumption implies that the full set $\mathcal{S}$ of input data is generally unstructured and uncorrelated, it does not prevent non-trivial correlations to appear in any smaller labeled subset S, selected through an AL procedure.

In pool-based AL settings, it is assumed that the student has a fixed budget n for building its training set, i.e. that only up to nN labels can be queried for training. The AL goal is to select, among the pool $\mathcal{S}$ of available samples, the nN *most informative* labels, to present to the student so that the latter achieves the best possible generalization performance. While many criteria of informativeness have been considered in the literature, see e.g. [1], in the teacher-student setting there exist a natural measure of informativeness, which we shall define in the next section.

2. The Gardner volume as a measure of optimality

A natural strategy for ranking the possible subset selections is to evaluate the mutual information $\mathcal{I}(\boldsymbol{x}^0;\boldsymbol{Y}|\boldsymbol{F})$ between the teacher vector random variable $\boldsymbol{x}^0$ and the random variable corresponding to the subset of labels $\boldsymbol{Y}$, conditioned on the realization of the corresponding inputs $\boldsymbol{F}$. The mutual information between two random variables $\boldsymbol{x}^0, \boldsymbol{Y}$ quantifies the mutual dependence between these variables, or, in other words, how much information about one random variable is gained if the other is observed [19]. More precisely, $\mathcal{I}(\boldsymbol{x}^0;\boldsymbol{Y}|\boldsymbol{F})$ is mathematically defined as:

$$\mathcal{I}(\boldsymbol{x}^0;\boldsymbol{Y}|\boldsymbol{F})=\mathcal{H}(\boldsymbol{Y}|\boldsymbol{F})-\mathcal{H}(\boldsymbol{Y}|\boldsymbol{F},\boldsymbol{x}^0), \tag{1}$$

where

$$\mathcal{H}(\boldsymbol{Y}|\boldsymbol{F})=-\mathbb{E}_{\boldsymbol{Y},\boldsymbol{x}^0}\ln P(\boldsymbol{Y}|\boldsymbol{F}) \tag{2}$$

$$\mathcal{H}(\boldsymbol{Y}|\boldsymbol{F},\boldsymbol{x}^0)=-\mathbb{E}_{\boldsymbol{Y},\boldsymbol{x}^0}\ln P(\boldsymbol{Y}|\boldsymbol{F},\boldsymbol{x}^0), \tag{3}$$

are so-called conditional entropies. $\mathbb{E}_{\boldsymbol{Y},\boldsymbol{x}^0}$ denotes the average with respect to the random variables $\boldsymbol{Y},\boldsymbol{x}^0$. The entropy of a random variable quantifies the uncertainty about its realization [19]. Therefore, the mutual information admits the very intuitive interpretation of the gain of information about the realization of $\boldsymbol{Y}$ if in addition to the selected data $\boldsymbol{F}$ the realization of $\boldsymbol{x}^0$ is also known. Good selections contain larger amounts of information about the ground truth, encoded in the labels, and make the associated inference problem for the student easier. Conversely, bad selections are characterized by less informative labels. In the case of the teacher-student perceptron, where the output channel $P_{\rm out}(\cdot|\cdot)$ is completely deterministic and binary, the mutual information can be rewritten (following [18]) as:

$$\begin{aligned}\mathcal{I}(\boldsymbol{x}^0;\boldsymbol{Y}|\boldsymbol{F})&=\mathcal{H}(\boldsymbol{Y}|\boldsymbol{F})-\mathcal{H}(\boldsymbol{Y}|\boldsymbol{F},\boldsymbol{x}^0)=\mathcal{H}(\boldsymbol{Y}|\boldsymbol{F})\\&=-\mathbb{E}_{\boldsymbol{x}^0,\boldsymbol{Y}}\ln\int d\boldsymbol{x}P_X(\boldsymbol{x})P_{\rm out}(\boldsymbol{Y}|\boldsymbol{F}\cdot\boldsymbol{x}).\end{aligned} \tag{4}$$

In going from the first to the second line, we used the fact that $P(\boldsymbol{Y}|\boldsymbol{F},\boldsymbol{x}^{\boldsymbol{0}}) = \delta(\boldsymbol{Y} - \text{sign}(\boldsymbol{F}\cdot\boldsymbol{x}^{\boldsymbol{0}}))$ and that the entropy associated to a point mass distribution is vanishing ($\mathcal{H}(\boldsymbol{Y}|\boldsymbol{F},\boldsymbol{x}^{\boldsymbol{0}}) = 0$), see for example [19]. Equation (4) allows a connection with a quantity well-known in statistical physics, the so-called Gardner volume [11, 15, 20], denoted in the following by v:

$$\ln v \equiv \frac{1}{N}\mathbb{E}_{\boldsymbol{x}^{\boldsymbol{0}},\boldsymbol{Y}}\ln\int d\boldsymbol{x}P_X(\boldsymbol{x})P_{\text{out}}(\boldsymbol{Y}|\boldsymbol{F}\cdot\boldsymbol{x}) = -\frac{1}{N}\mathcal{I}(\boldsymbol{x}^{\boldsymbol{0}};\boldsymbol{Y}|\boldsymbol{F}). \tag{5}$$

The Gardner volume differs from the mutual information (4) just by a trivial multiplicative factor $-1/N$. The Gardner volume represents the extent of the version space [21], i.e. the entropy of hypotheses in the model class consistent with the labeled training set. This provides a natural measure of the quality of the student training. A narrower volume implies less uncertainty about the ground truth $\boldsymbol{x}^{\boldsymbol{0}}$ and is thus a desirable objective in an AL framework. We shall focus the rest of our discussion on the large deviation properties of the Gardner volume while inviting the reader to keep in mind its connection with the mutual information given in (4) and (5).

There exist other natural measures of informativeness, e.g. the student generalization error ε_g and the magnetization (or teacher/student overlap) $m = \boldsymbol{x}\cdot\boldsymbol{x}^{\boldsymbol{0}}/N$. In the thermodynamic limit $N\uparrow\infty$, ε_g is a decreasing function of m (see the appendix C for more details). Moreover we will show analytical and numerical evidence that all these measures co-vary, at least in the simple teacher-student setting studied in this work. A numerical check at finite N of the correlation between v and m can also be found in section 4.1. In contrast to the Gardner volume v, the computation of the best achievable values for the test error ε_g or the magnetization m is harder and to the author's knowledge a methodologically open question even in the setting of the present article.

3. Large deviations of the Gardner volume

We consider the problem of sampling labeled subsets of cardinality nN, $0 < n \leqslant \alpha$, from a fixed pool of data of cardinality αN, $\alpha \sim \mathcal{O}(1)$, and study the variations in the associated Gardner volumes. We will hereby assume (as is usual in statistical physics) that, for any fixed pool and subset size, the Gardner volume probability distribution follows a large deviation principle, i.e. that there exist an exponential number $e^{N\Sigma(n,v)}$ of subsets choices that produce Gardner volumes equal to v. Employing a statistical physics terminology, we will refer to the rate function, $\Sigma(n,v)$, as the *complexity* of labeled subsets associated to a budget n and a volume v.

In the large N limit, the overwhelming majority of subsets will thus realize a Gardner volume $v^\star$, such that $v^\star = \text{argmax}_v\Sigma(n,v)$. We will call $v^\star$ the typical Gardner volume because for a randomly drawn subset the resulting Gardner volume is with high probability close to this value. This is because of the large deviation principle that implies the fluctuations around this typical value to be exponentially rare, random sampling will thus almost certainly yield Gardner volumes extremely close to $v^\star$. However, the aim of AL is to find strategies for accessing the atypically informative subsets (i.e. the atypically small volumes $v < v^\star$), whence the necessity of analyzing the large deviations properties of the subset selection process.

It is convenient to introduce a vector of selection variables $\{\sigma_\mu\}_{0\leqslant\mu\leqslant\alpha N} \in \{0,1\}^{\alpha N}$, such that $\sigma_\mu = 1$ when the sample $\boldsymbol{F}_\mu \in \mathcal{S}$ is selected (and added to the labeled training set), while $\sigma_\mu = 0$ otherwise. In this notation the selected subset $S \subset \mathcal{S}$ is easily defined as $S = \{\boldsymbol{F}_\mu \in \mathcal{S}|\sigma_\mu = 1\}$.

Since a direct computation of the complexity is not straightforward, as customary in this type of analyses [22] we derive it by first evaluating its Legendre transform. We introduce the (unnormalized) measure over the selection variables, for any reals β,ϕ:

$$\mathbb{P}_{\beta,\phi}(\{\sigma_\mu\}) = \left[\int d\boldsymbol{x}P_X(\boldsymbol{x})\prod_{\mu=1}^{\alpha N}P_{\text{out}}(y^\mu|\boldsymbol{F}^\mu\cdot\boldsymbol{x})^{\sigma_\mu}\right]^\beta e^{\phi\sum_\mu\sigma_\mu}, \tag{6}$$

and the associated free entropy:

$$\Phi(\beta,\phi) = \mathbb{E}_{\boldsymbol{F},\boldsymbol{x}^{\boldsymbol{0}}}\frac{1}{N}\ln\Xi = \mathbb{E}_{\boldsymbol{F},\boldsymbol{x}^{\boldsymbol{0}}}\frac{1}{N}\ln\sum_{\sigma_\mu}\mathbb{P}_{\beta,\phi}(\{\sigma_\mu\}). \tag{7}$$

From a statistical physics perspective, Ξ can be regarded as a grand-canonical partition function, with β playing the role of an inverse temperature, the Gardner volume being the associated energy function, and where ϕ is an effective chemical potential controlling the cardinality of the selection subset, $|S|$. In the

thermodynamic limit $N\uparrow\infty$, by applying the saddle-point method one can easily see that $\Phi(\beta,\phi)$ will be dominated by a subset of selection vectors $\{\sigma_\mu\}$ whose budget and energy, $n^\star$ and $\nu^\star$, are given by:

$$\Phi(\beta,\phi)=\underset{\nu,n}{\text{extr}}\left\{\Sigma(n,\nu)+\beta\ln\nu+\phi n\right\}. \tag{8}$$

Thus, inverting the Legendre transform yields the sought complexity:

$$\Sigma(n,\nu)=\Phi(\beta,\phi)-\beta\ln\nu-n\phi|_{\partial_\beta\Phi=\ln\nu,\partial_\phi\Phi=n}. \tag{9}$$

At fixed budget n, the range of values of the volume ν associated to positive complexities, i.e. with $\Sigma(n,\nu)>0$, effectively spans all the achievable Gardner volumes for subsets of that given cardinality, agnostic of the actual strategy for selecting them. In particular, $\inf_\nu\{\nu|\Sigma(n,\nu)>0\}$ and $\sup_\nu\{\nu|\Sigma(n,\nu)>0\}$ define the minimal and maximal Gardner volumes and provide theoretical algorithmic boundaries for all realizable AL strategies. Note that this means that our prototypical model, albeit being idealized, constitutes a nice benchmark for comparing known pool-based AL heuristics.

3.1. Replica symmetric free energy for a GLM

This section details the computation of the free entropy (7) for the generic case of GLM, with arbitrary teacher/student posteriors/priors. The specialization to the particular case of the teacher-student perceptron with no mismatch (Bayes-optimal) will be carried out section 3.2. Our computation borrows from [18, 23] which study the simple measure. Because we study large deviations, our formalism has some semblance with one-step replica breaking (1RSB) equations, a discussion of which can be found for example in [24–26].

3.1.1. Notations and assumptions

We consider a student GLM [23] with N-dimensional weights learning from αN samples $\boldsymbol{F}^\mu\in\mathbb{R}^N$ stacked in a matrix $\boldsymbol{F}\in\mathcal{M}_{\alpha N,N}(\mathbb{R})$ and the corresponding labels y^μ stacked into $\boldsymbol{Y}\in\mathbb{R}^{\alpha N}$. We assume the student–teacher (or planted) setting [13], where the labels are generated by the ground truth (teacher weights) $\boldsymbol{x}^0$ with the channel measure $\overline{P_{\text{out}}}(y^\mu|\boldsymbol{F}^\mu\cdot\boldsymbol{x}^0)$. The teacher weight itself is drawn with prior $\overline{P_X}(\cdot)$. Given $\boldsymbol{F}$ and $\boldsymbol{Y}$, the student perceptron is trained so that its own weight vector $\boldsymbol{x}$ tries to match the ground truth $\boldsymbol{x}^0$. The inference is carried out with the student prior $P_X(\cdot)$ and posterior $P_{\text{out}}(\boldsymbol{Y}|\boldsymbol{F}\cdot\boldsymbol{x})$. Note that the cases where $P_X(\cdot)\neq\overline{P_X}(\cdot)$ or $P_{\text{out}}(\cdot)\neq\overline{P_{\text{out}}}(\cdot)$ mean that the student ignores the precise Markov chain wherefrom the labels are generated, as discussed in section 1.2, see also [13]. The likelihood that a vector $\boldsymbol{x}$ is the ground truth vector is then $P_X(\boldsymbol{x})P_{\text{out}}(\boldsymbol{Y}|\boldsymbol{F}\cdot\boldsymbol{x})$. In this case the Gardner volume reads:

$$\nu=\left(\int d\boldsymbol{x}P_X(\boldsymbol{x})P_{\text{out}}(\boldsymbol{Y}|\boldsymbol{F}\cdot\boldsymbol{x})\right)^{\frac{1}{N}}. \tag{10}$$

The smaller ν, the easier the student inference, see section 1.2. The validity of the Gardner volume as a measure of informativeness is justified for the Bayes-optimal perceptron in section 2.

We consider here pool-based AL, where only a subset S of the pool $\mathcal{S}=\{\boldsymbol{F}^\mu\}_{1\leqslant\mu\leqslant\alpha N}$ is used for training. The choice of subset can be conveniently parametrized by the Boolean $\sigma_\mu\in\{0,1\}$, where $\sigma_\mu=1$ means sample $\boldsymbol{F}_\mu$ is used, while $\sigma_\mu=0$ means $\boldsymbol{F}_\mu$ is not selected. For a given budget $0\leqslant n=\frac{1}{N}|S|=\frac{1}{N}\sum_{\mu=1}^{\alpha N}\sigma_\mu\leqslant\alpha$, we intend to find the selection S that minimizes the Gardner volume $\nu(\sigma_\mu)$, viz. that allows the best student guess. To do this we shall compute the complexity $\Sigma(n,\nu)$, with $e^{N\Sigma(n,\nu)}$ the number of ways to select nN samples so that the Gardner volume associated with the training of the student is ν, as in section 3.

To simplify, the samples are taken to be identically and independently distributed according to a normal distribution $\forall(i,\mu),\ F_i^\mu\overset{d}{=}\mathcal{N}(0,\frac{1}{\sqrt{N}})$. Moreover all measures over vectors are assumed to be separable, that is factorizable as a product of identical measures over the components. Notation-wise $P_X(\boldsymbol{x})$ for example is therefore understood to mean $\prod_{i=1}^N P_X(x_i)$.

3.1.2. Replica trick

The goal is to compute the averaged log partition function (free entropy in statistical physics terms):

$$\Phi(\beta,\phi)=\mathbb{E}_{\boldsymbol{F},\boldsymbol{Y},\boldsymbol{x}^0}\frac{1}{N}\ln\Xi=\mathbb{E}_{\boldsymbol{F},\boldsymbol{Y},\boldsymbol{x}^0}\frac{1}{N}\ln\sum_{\sigma_\mu}\left[\int d\boldsymbol{x}P_X(\boldsymbol{x})\prod_{\mu=1}^{\alpha N}P_{\text{out}}(y^\mu|\boldsymbol{F}^\mu\cdot\boldsymbol{x})^{\sigma_\mu}\right]^\beta e^{\phi\sum_\mu\sigma_\mu}, \tag{11}$$

β can be seen as an inverse temperature and ϕ as a chemical potential, see section 3. Note that the selection variables $\{\sigma_\mu\}$ play in the grand-canonical partition function (11) the role of an annealed disorder in disordered systems terminology [26], and shall be sometimes referred to as such in the following.

The standard way of taking care of the logarithm in equation (11) is the replica trick [3.1, 26, 28],

$$\Phi(\beta,\phi)=\lim_{s\to 0}\frac{1}{s}\mathbb{E}_{\boldsymbol{F},\boldsymbol{Y},\boldsymbol{x}^0}\Xi^s. \tag{12}$$

To compute $\mathbb{E}_{\boldsymbol{F},\boldsymbol{Y},\boldsymbol{x}^0}\Xi^s$, one needs to further replicate β times to care for the power β involved in the summand in equation (11):

$$\begin{aligned}\mathbb{E}_{\boldsymbol{F},\boldsymbol{x}^0}\Xi^s&=\mathbb{E}_{\boldsymbol{F}}\int d\boldsymbol{x}^0\overline{P_X}(\boldsymbol{x}^0)\int d\boldsymbol{y}\prod_{\mu=1}^{\alpha N}\overline{P_{\rm out}}(y^\mu|\boldsymbol{F}^\mu\cdot\boldsymbol{x}^0)\\&\quad\times\sum_{S_a^\mu}\int\prod_{a=1}^{s}\prod_{\alpha=1}^{\beta}(d\boldsymbol{x}^{a\alpha}P_X(\boldsymbol{x}^{a\alpha}))\prod_{a\alpha}\prod_{\mu}P_{\rm out}(y^\mu|\boldsymbol{F}^\mu\cdot\boldsymbol{x}^{a\alpha})^{\sigma_\mu^a}e^{\phi\sigma_\mu^a}\\&=\mathbb{E}_{\boldsymbol{F}}\sum_{S_a^\mu}\int\prod_{a\alpha}(d\boldsymbol{x}^{a\alpha}P_X(\boldsymbol{x}^{a\alpha}))e^{-\phi}\prod_{a\alpha}\prod_{\mu}P_{\rm out}(y^\mu|\boldsymbol{F}^\mu\cdot\boldsymbol{x}^{a\alpha})^{\sigma_\mu^a}e^{\phi\sigma_\mu^a}.\end{aligned} \tag{13}$$

In the present problem we thus introduced two replication levels. Each replica is hence characterized by a set of two indices: the first a index runs from 1 to s and specifies the disorder replica, the second α index, running from 1 to β is related to the replication in β. In total there are therefore $s\times\beta$ replicas. The teacher is set as replica 0. Implicitly henceforth $a\alpha$ when summed over will be running over $[1,s]\times[1,\beta]\cup 0$. But

$$\begin{aligned}\mathbb{E}_{\boldsymbol{F}}\prod_{a\alpha}\prod_{\mu}P_{\rm out}(y^\mu|\boldsymbol{F}^\mu\cdot\boldsymbol{x}^{a\alpha})^{\sigma_\mu^a}&=\int\prod_{\mu}\prod_{a\alpha}dh_{a\alpha}^\mu(\det 2\pi\boldsymbol{Q})^{-\frac{\alpha N}{2}}e^{-\frac{1}{2}\sum_\mu\sum a\alpha,c\gamma h_{a\alpha}^\mu(\boldsymbol{Q}^{-1})^{a\alpha\ c\gamma}h_{c\gamma}^\mu}\\&\quad\times\int\prod_{a\alpha\neq c\gamma}dq_{a\alpha,c\gamma}\int\prod_{a\alpha\neq c\gamma}d\hat{q}_{a\alpha,c\gamma}e^{\sum_{a\alpha\neq c\gamma}\hat{q}_{a\alpha,c\gamma}(\boldsymbol{x}^{a\alpha}\cdot\boldsymbol{x}^{c\gamma}-Nq_{a\alpha,c\gamma}}\\&\quad\times\prod_{a\alpha}\prod_{\mu}P_{\rm out}(y^\mu|h_{a\alpha}^\mu)^{\sigma_\mu^a},\end{aligned} \tag{14}$$

where we defined $h_{a\alpha}^\mu\equiv\boldsymbol{F}^\mu\cdot\boldsymbol{x}^{a\alpha}$ Gaussian because of the central limit theorem and enforced the definition of its covariance matrix $\boldsymbol{Q}$ with integral representations of Dirac deltas. The conjugate matrix is $\hat{\boldsymbol{Q}}$. Matrix elements are noted with small q. Then,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{F},\boldsymbol{x}^0}\Xi^s&=\int d\hat{\boldsymbol{Q}}d\boldsymbol{Q}e^{-N{\rm Tr}\hat{\boldsymbol{Q}}\boldsymbol{Q}}\left(\int\prod_{a\alpha}d\boldsymbol{x}^{a\alpha}P_X(\boldsymbol{x}^{a\alpha})e^{\sum_{a\alpha\neq c\gamma}x^{a\alpha}\hat{q}_{a\alpha,c\gamma}x^{c\gamma}}\right)^N\\&\quad\times\left(\sum_{S^a}e^{\phi(\sum_a S^a-1)}\int d\boldsymbol{y}\prod_{a\alpha}dh_{a\alpha}(\det 2\pi\boldsymbol{Q})^{-\frac{1}{2}}e^{-\frac{1}{2}\sum_{a\alpha,c\gamma}h_{a\alpha}(\boldsymbol{Q}^{-1})^{a\alpha\ c\gamma}h_{c\gamma}}\prod_{a\alpha}P_{\rm out}(y^\mu|h_{a\alpha})^{S^a}\right)^{\alpha N},\end{aligned} \tag{15}$$

where we factorized both in i indices (first parenthesis) and in μ indices (second parenthesis). The free entropy defined in (11) then reads:

$$\Phi(\beta,\phi)=\lim_{s\to 0}\frac{1}{s}\underset{\hat{\boldsymbol{Q}},\boldsymbol{Q}}{\rm extr}\left\{-{\rm Tr}\hat{\boldsymbol{Q}}\boldsymbol{Q}+\ln I_X(\hat{\boldsymbol{Q}})+\alpha\ln I_Y(\boldsymbol{Q})\right\}, \tag{16}$$

with,

$$I_X=\int\prod_{a\alpha}d\boldsymbol{x}^{a\alpha}P_X(\boldsymbol{x}^{a\alpha})e^{\sum_{a\alpha\neq c\gamma}x^{a\alpha}\hat{q}_{a\alpha,c\gamma}x^{c\gamma}}, \tag{17}$$

$$I_Y=\sum_{S^a}e^{\phi(\sum_a S^a-1)}\int d\boldsymbol{y}\prod_{a\alpha}dh_{a\alpha}(\det 2\pi\boldsymbol{Q})^{-\frac{1}{2}}e^{-\frac{1}{2}\sum_{a\alpha,c\gamma}h_{a\alpha}(\boldsymbol{Q}^{-1})^{a\alpha\ c\gamma}h_{c\gamma}}\prod_{a\alpha}P_{\rm out}(y^\mu|h_{a\alpha})^{S^a}. \tag{18}$$

3.1.3. Replica symmetric (RS) ansatz

The extremization in equation (16) is hard to carry out. As is now standard in the disordered systems literature we can reduce the number of parameters to be extremized over by enforcing the so-called RS ansatz [26] on both replication levels:

$$q^{0,0} = r^0, \;\; \hat{q}^{0,0} = \hat{r}^0 \tag{19}$$

$$q^{a\alpha,0} = m, \;\; \hat{q}^{a\alpha,0} = \hat{m} \tag{20}$$

$$q^{a\alpha,a\alpha} = r, \;\; \hat{q}^{a\alpha,a\alpha} = -\frac{1}{2}\hat{r} \tag{21}$$

$$q^{a\alpha,a\gamma} = Q, \;\; \hat{q}^{a\alpha,a\gamma} = \hat{Q} \tag{22}$$

$$q^{a\alpha,c\gamma} = q, \;\; \hat{q}^{a\alpha,c\gamma} = \hat{q}, \tag{23}$$

where $q < Q$. Physically, the ansatz (19)–(23) means that two replicas seeing the same realization of disorder (i.e. possessing the *same* first index) have an overlap Q greater than the overlap between students seeing different realizations (an thus possessing *different* a-index). The $-\frac{1}{2}$ in the definition of $\hat{r}$ (21) is just introduced for latter convenience.

Note finally that while the ansatz (19)–(23) is replica-symmetric for both replications, it gives a set of equations that are formally those of a 1RSB problem [25]. This is also a reason why taking 1RSB ansatz [26] in the present large deviation calculation would be rather involved as it would lead to equations in the usual 2RSB form that are numerically involved to be solved.

We plug the RS ansatz (19)–(23) into the three contributions that make up equation (16). The trace term is:

$$-\mathrm{Tr}\hat{\boldsymbol{Q}}\boldsymbol{Q} = -\hat{r}^0 r^0 - \beta s m\hat{m} + \frac{1}{2}\beta s r\hat{r} - s\frac{\beta(\beta-1)}{2}Q\hat{Q} - \beta^2\frac{s(s-1)}{2}q\hat{q}. \tag{24}$$

We can decompose the exponent in (17) according to the ansatz (19)–(23):

$$\sum_{a\alpha\neq c\gamma} x^{a\alpha}\hat{q}_{a\alpha,c\gamma}x^{c\gamma} = \hat{r}^0(x^0)^2 + \hat{m}x^0\sum_{a\alpha\neq 0}x^{a\alpha} - \frac{\hat{r}+\hat{Q}}{2}\sum_{a\alpha\neq 0}(x^{a\alpha})^2 + \frac{\hat{Q}-\hat{q}}{2}\sum_{a\neq 0}\sum_{\alpha,\gamma}x^{a\alpha}x^{a\gamma} + \frac{\hat{q}}{2}\sum_{a\alpha\neq 0, c\gamma\neq 0}x^{a\alpha}x^{c\gamma}. \tag{25}$$

In the last but one term index 0 does not intervene. Introducing Hubbard–Stratonovitch fields $\{\lambda_a\}$ for the last but one term and Hubbard–Stratonovith field ξ for the last, I_X reads:

$$I_X = \int D\xi \int dx^0 \overline{P_X}(x^0) e^{\hat{r}^0(x^0)^2}\left[\int D\lambda\left(\int dx P_X(x) e^{\hat{m}x^0 x - \frac{\hat{r}+\hat{Q}}{2}x^2 + \sqrt{\hat{Q}-\hat{q}}\lambda x + \sqrt{\hat{q}}\xi x}\right)^\beta\right]^s. \tag{26}$$

To carry out the computation for I_Y (equation (18)) we need to explicitly compute the inverse of the Parisi matrix $\boldsymbol{Q}$ involved in equation (18). This is done in the following subsection.

3.1.4. Inverse of the overlap matrix

Name $\tilde{\boldsymbol{Q}} \equiv \boldsymbol{Q}^{-1}$ the inverse of the overlap matrix $\boldsymbol{Q}$. Since $\tilde{\boldsymbol{Q}}$ is clearly of the same form as $\boldsymbol{Q}$, we can parametrize its coefficient in an identical fashion as those of $\boldsymbol{Q}$ $\tilde{r}^0, \tilde{m}, \tilde{r}, \tilde{q}, \tilde{Q}$. $\tilde{\boldsymbol{Q}}\boldsymbol{Q} = \mathbb{1}_{\beta s+1}$ means:

$$r^0\tilde{r}^0 + \beta s m\tilde{m} = 1 \tag{27}$$

$$r^0\tilde{m} + m\tilde{r} + (\beta-1)\tilde{Q}m + \beta(s-1)\tilde{q}m = 0, \tag{28}$$

$$\tilde{r}^0 m + \tilde{m}(r + (\beta-1)Q + \beta(s-1)q) = 0 \tag{29}$$

$$m\tilde{m} + r\tilde{r} + (\beta-1)Q\tilde{Q} + \beta(s-1)q\tilde{q} = 1, \tag{30}$$

$$m\tilde{m} + r\tilde{Q} + \tilde{r}Q + (\beta-2)Q\tilde{Q} + \beta(s-1)q\tilde{q} = 0 \tag{31}$$

$$m\tilde{m} + r\tilde{q} + (\beta-1)Q\tilde{q} + q\tilde{r} + (\beta-1)q\tilde{Q} + \beta(s-2)q\tilde{q} = 0, \tag{32}$$

yielding:

$$\tilde{r}^0 = \frac{r + (\beta-1)Q + \beta(s-1)q}{r^0(r + (\beta-1)Q + \beta(s-1)q) - \beta s m^2}, \tag{33}$$

$$\tilde{m} = \frac{-m}{r^0(r + (\beta-1)Q + \beta(s-1)q) - \beta s m^2} \tag{34}$$

$$\begin{aligned}\tilde{r} = {} & \frac{\beta m^2(q - 2Q + r) + \beta((1-s)q^2 + Q(3Q-2r) + (s-2)q(2Q-r))r^0}{(Q-r)(r + (\beta-1)Q - \beta q)(r^0(r + (\beta-1)Q + \beta(s-1)q) - \beta s m^2)} \\ & + \frac{\beta^2(q-Q)(-m^2 s + ((-1+s)q + Q)r^0) + (Q-r)(m^2 + (-2Q+r)r^0)}{(Q-r)(r + (\beta-1)Q - \beta q)(r^0(r + (\beta-1)Q + \beta(s-1)q) - \beta s m^2)},\end{aligned} \tag{35}$$

$$\tilde{Q} = \frac{(Q-r)(m^2 - Qr^0) + \beta(q-Q)(m^2 s - ((s-1)q + Q)r^0}{(Q-r)(r + (\beta-1)Q - \beta q)(r^0(r + (\beta-1)Q + \beta(s-1)q) - \beta s m^2)}, \tag{36}$$

$$\tilde{q} = \frac{-m^2 + qr^0}{(r + (\beta-1)Q - \beta q)(r^0(r + (\beta-1)Q + \beta(s-1)q) - \beta s m^2)}. \tag{37}$$

To compute the determinant of $\boldsymbol{Q}$, the simplest way is to guess the eigenvectors. For $(x, 1, 1, \ldots, 1)^T$ we get two eigenvalues $\lambda_\pm$ whose product is:

$$\lambda_+ \lambda_- = r^0(r + (\beta-1)Q + (s-1)\beta q) - \beta s m^2. \tag{38}$$

Then come $s(\beta-1)$ eigenvectors $\boldsymbol{e_i} - \boldsymbol{e_{i+1}}$, $i \not\equiv 0[\beta]$ (we are indexing starting from 0), with associated eigenvalues $(r-Q)$. Then for $0 \leqslant s \leqslant s-2$,

$$\sum_{k=s\beta+1}^{(s+1)\beta} \boldsymbol{e_k} - \sum_{k=(s+1)\beta+1}^{(s+2)\beta} \boldsymbol{e_k}, \tag{39}$$

is an eigenvector with eigenvalue $r + (\beta-1)Q - \beta q$. Then

$$\begin{aligned}\ln\det\boldsymbol{Q} = {} & \ln(r^0(r + (\beta-1)Q + (s-1)\beta q) - \beta m^2) + (s-1)\ln(r + (\beta-1)Q - \beta q) \\ & + (\beta-1)s\ln(r-Q).\end{aligned} \tag{40}$$

The same equality holds with tilde quantities in the right-hand side provided the signs are inverted, since $\ln\det\boldsymbol{Q} = -\ln\det\tilde{\boldsymbol{Q}}$. Identifying term by term results straightforwardly in a set of relations between tilde and non-tilde quantities (henceforth referred as determinant relations),

$$r^0(r + (\beta-1)Q + (s-1)\beta q) - \beta m^2 = [\tilde{r}^0(\tilde{r} + (\beta-1)\tilde{Q} + (s-1)\beta\tilde{q}) - \beta\tilde{m}^2]^{-1} \tag{41}$$

$$r + (\beta-1)Q - \beta q = [\tilde{r} + (\beta-1)\tilde{Q} - \beta\tilde{q}]^{-1} \tag{42}$$

$$r - Q = [\tilde{r} - \tilde{Q}]^{-1}. \tag{43}$$

3.1.5. Evaluating the replica symmetric free entropy for GLM

Now decomposing the exponent in I_Y (18)

$$-\frac{1}{2}\sum_{a\alpha,c\gamma} h_{a\alpha}\tilde{q}_{a\alpha,c\gamma}h_{c\gamma} = -\frac{1}{2}\tilde{r}^0(h^0)^2 - \tilde{m}h^0\sum_{a\alpha\neq 0} h_{a\alpha} - \frac{\tilde{r}-\tilde{Q}}{2}\sum_{a\alpha\neq 0}(h_{a\alpha})^2 + (\tilde{q}-\tilde{Q})\sum_a\sum_{\alpha,\gamma} h_{a\alpha}h_{a\gamma} - \tilde{q}\sum_{a\alpha\neq 0,c\gamma\neq 0} h_{a\alpha}h_{c\gamma}. \tag{44}$$

Introducing Hubbard–Stratonovich fields $\{\zeta_a\}$ and η for the last two sums of (44) and factorizing in the index a

$$I_Y = \frac{1}{\sqrt{\det 2\pi\mathbf{Q}}}\int dy\int D\eta\int dh^0\overline{P_{\text{out}}}(y|h^0)e^{-\frac{1}{2}\tilde{r}^0(h^0)^2} \times\left[\sum_{S=0,1}\int D\zeta e^{\phi S}\left(\int dh P_{\text{out}}(y|h)^S e^{-\tilde{m}h^0h-\frac{\tilde{r}-\tilde{Q}}{2}h^2+\sqrt{\tilde{q}-\tilde{q}}\zeta h+\sqrt{-\tilde{q}}\eta h}\right)^\beta\right]^s. \tag{45}$$

Now that all terms are computed the next step is then to divide by s and take the $s\to 0$ limit as prescribed by the replica trick (12). First, we need to enforce that all non-vanishing order 0 contribution cancel out, since the free entropy should not be diverging. Then, one needs to actually compute first order terms that will contribute in Φ (16).

At order 0, $I_Y = 1$ since

$$\lim_{s\to 0}\ln\det(2\pi\mathbf{Q}) = \frac{1}{2}\ln(2\pi\tilde{r}^0) = -\ln\int dy\int D\eta\int dh^0\overline{P_{\text{out}}}(y|h^0)e^{-\frac{1}{2}\tilde{r}^0(h^0)^2}. \tag{46}$$

The cancellation of order 0 terms imposes:

$$0 = \hat{r}^0 r^0 + \ln\int dx^0\overline{P_X}(x^0)e^{\hat{r}^0(x^0)^2}, \tag{47}$$

where the first term comes from the trace term (24). It follows that $\hat{r}^0 = 0$. Moreover, because of the saddle point equality:

$$q_{a\alpha,c\gamma} = \frac{\int\prod_{d\delta}(dx^{d\delta}P_X(x^{d\delta}))x^{a\alpha}x^{c\gamma}e^{\sum_{d\delta\neq e\epsilon}x^{d\delta}\hat{q}_{d\delta,e\epsilon}x^{e\epsilon}}}{\int\prod_{d\delta}(dx^{d\delta}P_X(x^{d\delta}))e^{\sum_{d\delta\neq e\epsilon}x^{d\delta}\hat{q}_{d\delta,e\epsilon}x^{e\epsilon}}}, \tag{48}$$

derived straightforwardly from (16) we also have

$$r^0 = \int dx^0\overline{P_X}(x^0)(x^0)^2. \tag{49}$$

The order 1 contribution of I_X can be rewritten by carrying out a change of variables $\xi\to\xi+\hat{q}^{-\frac{1}{2}}\hat{m}x^0$ in equation (26), I_X assumes the compact form:

$$\lim_{s\to 0}\frac{1}{s}I_X = \int D\xi I_X^0(\xi)\ln I_X^1(\xi), \tag{50}$$

where

$$I_X^0(\xi) = \int dx^0\overline{P_X}(x^0)e^{-\frac{\hat{m}^2}{2\hat{q}}(x^0)^2+\frac{\hat{m}}{\sqrt{\hat{q}}}\xi x^0} \tag{51}$$

$$I_X^1(\xi) = \int D\lambda\left[\int dx P_X(x)e^{-\frac{\hat{r}+\hat{Q}}{2}x^2+(\sqrt{\hat{Q}-\hat{q}}\lambda+\sqrt{\hat{q}}\xi)x}\right]^\beta. \tag{52}$$

Assessing the order 1contribution from I_Y requires more work. Changing $\eta\to\eta-\frac{\tilde{m}}{\sqrt{-\tilde{q}}}h^0$ in (45) yields:

$$I_Y = \frac{1}{\sqrt{\det 2\pi\mathbf{Q}}}\int dy\int D\eta g^0(y,\eta)(g^1(y,\eta))^s, \tag{53}$$

with

$$g^0(y,\eta)=\int dh^0\overline{P_{\text{out}}}(y|h^0)e^{-\frac{1}{2}(\tilde{r}^0-\frac{\tilde{m}^2}{\tilde{q}})(h^0)^2-\frac{\tilde{m}}{\sqrt{-\tilde{q}}}h^0\eta} \tag{54}$$

$$g^1(y,\eta)=\sum_{S=0,1}\int D\zeta e^{\phi S}\left(\int dhP_{\text{out}}(y|h)^S e^{-\frac{\tilde{r}-\tilde{Q}}{2}h^2+\sqrt{\tilde{q}-\tilde{Q}}\zeta h+\sqrt{-\tilde{q}}\eta h}\right)^\beta. \tag{55}$$

Expanding $\ln I_Y$ to $\mathcal{O}(s)$ (subscripts in parentheses standing for order in s) gives

$$\frac{1}{s}\ln I_Y=-\frac{1}{2}\ln(\det 2\pi\mathbf{Q})_{(1)}+\frac{1}{\sqrt{2\pi r^0}}\int D\eta\int dy\left[g^0_{(1)}(y,\eta)+g^0_{(0)}(y,\eta)\ln g^1_{(0)}(y,\eta)\right]. \tag{56}$$

We used the fact that at order 0 terms canceled, and the identity $\int D\eta\int dy g^0(y,\eta)=\sqrt{2\pi r^0}+\mathcal{O}(s)$. But

$$\begin{aligned}\frac{1}{\sqrt{2\pi r^0}}\int D\eta\int dy g^0_{(1)}(y,\eta)&=\lim_{s\to 0}\frac{1}{\sqrt{2\pi r^0}}\partial_s\left(\int D\eta\int dy g^0(y,\eta)\right)\\&=\lim_{s\to 0}\frac{1}{\sqrt{2\pi r^0}}\partial_s\sqrt{\frac{2\pi}{\tilde{r}^0}}\\&=\frac{-\beta m^2}{2r^0(r+(\beta-1)Q-\beta q)},\end{aligned} \tag{57}$$

thus

$$\begin{aligned}\frac{1}{s}\ln I_Y=&-\frac{1}{2}\frac{\beta q}{r+(\beta-1)Q-\beta q}-\frac{1}{2}\ln(r+(\beta-1)Q-\beta q)-\frac{1}{2}(\beta-1)\ln(r-Q)\\&+\frac{1}{\sqrt{2\pi r^0}}\int dy\int D\eta g^0_{(0)}(y,\eta)\ln g^1_{(0)}(y,\eta).\end{aligned} \tag{58}$$

It is actually possible to proceed to Gaussian changes of variables in the last term so as to exactly cancel the first three contributions in (58). To do this:

$$h\to\sqrt{r-Q}h+(r-Q)(\sqrt{\tilde{q}-\tilde{Q}}\zeta+\sqrt{-\tilde{q}}\eta) \tag{59}$$

$$\zeta\to\frac{1}{\sqrt{1-\beta(r-Q)(\tilde{q}-\tilde{Q}}}\zeta-\frac{\beta\sqrt{-\tilde{q}(\tilde{q}-\tilde{Q})}(r-Q)}{1-\beta(r-Q)(\tilde{q}-\tilde{Q}}\eta, \tag{60}$$

(we used the determinant relations (41)–(43)) allowing to rewrite the last term in (58) as:

$$\begin{aligned}&\frac{1}{\sqrt{2\pi r^0}}\int dy\int D\eta g^0_{(0)}(y,\eta)\ln g^1_{(0)}(y,\eta)\\&=\frac{\beta}{2}\ln(r-Q)-\frac{1}{2}\ln(1-\beta(r-Q)(\tilde{q}-\tilde{Q}))\\&\quad+\int\frac{dy}{\sqrt{2\pi r^0}}\int D\eta g^0_{(0)}(y,\eta)\ln\left(1+e^\phi\int D\zeta\left(\int dhP_{\text{out}}(y|*)\right)^\beta\right)\\&\quad+\frac{1}{2}\left(-\beta(r-Q)\tilde{q}+\frac{\beta^2(-\tilde{q}(\tilde{q}-\tilde{Q})(r-Q)^2}{1-\beta(r-Q)(\tilde{q}-\tilde{Q}}\right)\int D\eta\eta^2\int\frac{dy}{\sqrt{2\pi r^0}}g^0_{(0)}(y,\eta),\end{aligned} \tag{61}$$

with

$$*=\sqrt{r-Q}h+(r-Q)\left(\sqrt{\frac{\tilde{q}-\tilde{Q}}{1-\beta(r-Q)(\tilde{q}-\tilde{Q})}}\zeta+\frac{\sqrt{-\tilde{q}}}{1-\beta(r-Q)(\tilde{q}-\tilde{Q})}\eta\right). \tag{62}$$

It is straightforward to see:

$$\int D\eta\eta^2\int\frac{dy}{\sqrt{2\pi r^0}}g^0_{(0)}(y,\eta)=\frac{\tilde{r}^0\tilde{q}-\tilde{m}^2}{\tilde{r}^0\tilde{q}}, \tag{63}$$

so the last term in (61) is:

$$\begin{aligned}&\frac{1}{2}\left(-\beta(r-Q)\tilde{q}+\frac{\beta^2(-\tilde{q}(\tilde{q}-\tilde{Q})(r-Q)^2}{1-\beta(r-Q)(\tilde{q}-\tilde{Q}}\right)\int D\eta\eta^2\int\frac{dy}{\sqrt{2\pi r^0}}g^0_{(0)}(y,\eta)\\&=-\frac{\beta}{2}\frac{\tilde{r}^0\tilde{q}-\tilde{m}^2}{\tilde{r}^0(\tilde{r}+(\beta-1)\tilde{Q}-\beta\tilde{q})}\\&=\frac{1}{2}\frac{\beta q}{r+(\beta-1)Q-\beta q}.\end{aligned}\tag{64}$$

Similarly the first terms in (61) are:

$$\begin{aligned}&\frac{\beta}{2}\ln(r-Q)-\frac{1}{2}\ln(1-\beta(r-Q)(\tilde{q}-\tilde{Q}))\\&=-\frac{\beta-1}{2}\ln(\tilde{r}-\tilde{Q})-\frac{1}{2}\ln(\tilde{r}+(\beta-1)\tilde{Q}-\beta\tilde{q})\\&=\frac{1}{2}\ln(r+(\beta-1)Q-\beta q)+\frac{1}{2}(\beta-1)\ln(r-Q).\end{aligned}\tag{65}$$

We again used the determinant identity (41)–(43) in the last line. Then tilde quantities in the $s=0$ limit can be accordingly be replaced by their expressions (33)–(37):

$$\lim_{s\to0}\tilde{r}^0-\frac{\tilde{m}^2}{\tilde{q}}=\frac{q}{qr^0-m^2}\tag{66}$$

$$\lim_{s\to0}\frac{\tilde{m}}{\sqrt{-\tilde{q}}}=\sqrt{\frac{m^2}{r^0(qr^0-m^2)}}\tag{67}$$

$$\lim_{s\to0}\sqrt{\frac{\tilde{q}-\tilde{Q}}{1-\beta(r-Q)(\tilde{q}-\tilde{Q})}}=\frac{\sqrt{Q-q}}{r-Q}\tag{68}$$

$$\lim_{s\to0}\frac{\sqrt{-\tilde{q}}}{1-\beta(r-Q)(\tilde{q}-\tilde{Q})}=\sqrt{\frac{qr^0-m^2}{r^0(r-Q)^2}}.\tag{69}$$

Ultimately some changes of variables can be used to bring $g^0_{(0)}$ to a more compact form:

$$h\to\sqrt{\frac{qr^0-m^2}{q}}h^0+\sqrt{\frac{m^2(qr^0-m^2)}{qr^0}},\quad\eta\to\sqrt{\frac{qr^0}{qr^0-m^2}}\eta.\tag{70}$$

Finally

$$\lim_{s\to0}\frac{1}{s}I_Y=\int D\eta\int dyI^0_Y(y,\eta)\ln I^1_Y(y,\eta),\tag{71}$$

with

$$I^0_Y(y,\eta)=\int Dh^0\overline{P_{\rm out}}\left(y|\sqrt{\frac{qr^0-m^2}{q}}h^0+\sqrt{\frac{m^2}{q}}\eta\right),\tag{72}$$

$$I^1_Y(y,\eta)=1+e^{\phi}\int D\zeta\left(\int dhP_{\rm out}\left(y|\sqrt{r-Q}h+\sqrt{Q-q}\zeta+\sqrt{q}\eta\right)\right)^{\beta}.\tag{73}$$

3.1.5.1. Replica symmetric free entropy for GLM

Putting everything together the replica free entropy (16) reads:

$$\begin{aligned}\Phi_{\rm RS}=\underset{\hat{m},\hat{r},\hat{q},\hat{Q},r,q,Q}{\rm extr}\Big\{&-\beta m\hat{m}+\frac{\beta}{2}r\hat{r}-\frac{\beta(\beta-1)}{2}Q\hat{Q}+\frac{\beta^2}{2}q\hat{q}+\int D\xi I^0_X(\xi)\ln I^1_X(\xi)\\&+\alpha\int D\eta\int dyI^0_Y(y,\eta)\ln I^1_Y(y,\eta)\Big\}\end{aligned}\tag{74}$$

$$r^0 = \int dx^0 \overline{P_X}(x^0)(x^0)^2 \tag{75}$$

$$I_X^0(\xi) = \int dx^0 \overline{P_X}(x^0) e^{-\frac{\hat{m}^2}{2\hat{q}}(x^0)^2 + \frac{\hat{m}}{\sqrt{\hat{q}}}\xi x^0} \tag{76}$$

$$I_X^1(\xi) = \int D\lambda \left[\int dx P_X(x) e^{-\frac{\hat{r}+\hat{Q}}{2}x^2 + \left(\sqrt{\hat{Q}-\hat{q}}\lambda + \sqrt{\hat{q}}\xi\right)x} \right]^\beta \tag{77}$$

$$I_Y^0(y,\eta) = \int Dh^0 \overline{P_{\text{out}}}\left(y | \sqrt{\frac{qr^0 - m^2}{q}} h^0 + \sqrt{\frac{m^2}{q}}\eta \right) \tag{78}$$

$$I_Y^1(y,\eta) = 1 + e^\phi \int D\zeta \left(\int dh P_{\text{out}}\left(y | \sqrt{r-Q}h + \sqrt{Q-q}\zeta + \sqrt{q}\eta \right) \right)^\beta. \tag{79}$$

3.2. Replica free energy for the perceptron

The Bayes-optimal teacher-student setting for the perceptron is defined by the following measures (see section 1.2):

$$\overline{P_X}(x) = P_X(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}x^2} \tag{80}$$

$$\overline{P_{\text{out}}}(y|h) = P_{\text{out}}(y|h) = \delta(y - \text{sgn}(h)). \tag{81}$$

3.2.1. Replica symmetric free entropy for the perceptron

We shall simply plug into the generic GLM (74) expressions the particular priors and posteriors for the perceptron (80) and (81). First,

$$I_X^0(\xi) = \int \frac{1}{\sqrt{2\pi}} dx^0 e^{-\frac{1}{2}(1+\frac{\hat{m}^2}{\hat{q}})(x^0)^2 + \frac{\hat{m}}{\sqrt{\hat{q}}}\xi x^0} \tag{82}$$

$$= \frac{1}{\sqrt{2\pi(1+\frac{\hat{m}^2}{\hat{q}})}} e^{\frac{1}{2}\frac{\hat{m}^2}{\hat{q}+\hat{m}^2}\xi^2}, \tag{83}$$

while two straightforward Gaussian integrals yield:

$$I_X^1(\xi) = \int D\lambda \left[\int \frac{1}{\sqrt{2\pi}} dx P_X(x) e^{-\frac{1}{2}(\hat{r}+\hat{Q})x^2 + (\sqrt{\hat{Q}-\hat{q}}\lambda + \sqrt{\hat{q}}\xi)x} \right]^\beta \tag{84}$$

$$= \int d\lambda \frac{1}{(1+\hat{r}+\hat{Q})^{\frac{\beta}{2}}} e^{-\frac{1}{2}(1-\beta\frac{\hat{Q}-\hat{q}}{1+\hat{r}+\hat{Q}})\lambda^2 + \beta\frac{\sqrt{\hat{q}(\hat{Q}-\hat{q})}}{1+\hat{r}+\hat{Q}}\xi\lambda + \frac{\beta}{2}\frac{\hat{q}}{1+\hat{r}+\hat{Q}}\xi^2} \tag{85}$$

$$= (1+\hat{r}+\hat{Q})^{\frac{1-\beta}{2}} (1+\hat{r}-(\beta-1)\hat{Q}+\beta\hat{q})^{-\frac{1}{2}} e^{\frac{\beta\hat{q}}{2(1+\hat{r}-(\beta-1)\hat{Q}+\beta\hat{q})}\xi^2}. \tag{86}$$

Thus,

$$\int D\xi I_X^0(\xi) \ln I_X^1(\xi) = -\frac{\beta-1}{2}\ln(1+\hat{r}+\hat{Q}) - \frac{1}{2}\ln(1+\hat{r}-(\beta-1)\hat{Q}+\beta\hat{q}) + \frac{\beta}{2}\frac{\hat{q}+\hat{m}^2}{1+\hat{r}-(\beta-1)\hat{Q}+\beta\hat{q}}. \tag{87}$$

Now defining the special function: $H(x) \equiv \frac{1}{\sqrt{2\pi}} \int_x^\infty Dt$

$$I_Y^0(y,\eta) = \int Dh^0 \delta\left(y - \text{sgn}\left(\sqrt{\frac{q-m^2}{q}} h^0 + \sqrt{\frac{m^2}{q}}\eta \right) \right) = H\left(-y\sqrt{\frac{m^2}{q-m^2}}\eta \right). \tag{88}$$

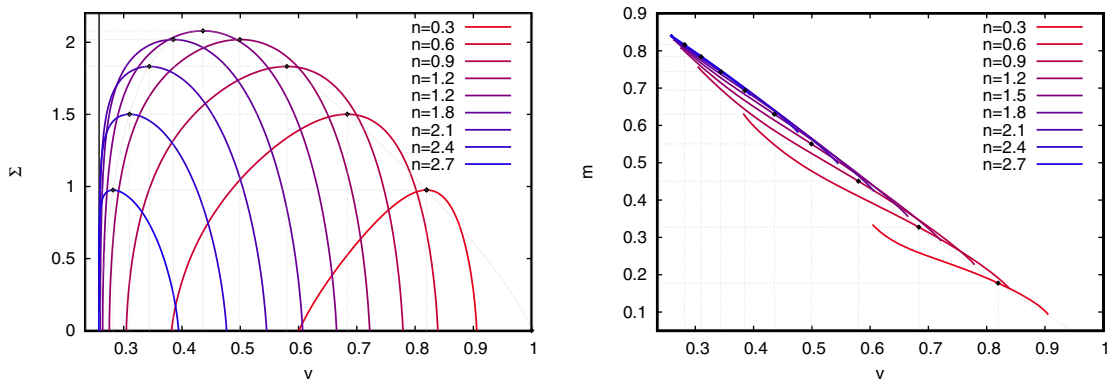

Figure 1. (Left) Complexity–volume curves $\Sigma(n, v)$ for various budgets n, at pool size $\alpha = 3$ extracted from the large deviation computations. These curves reach their maxima at a point with coordinates corresponding to the Gardner volume of randomly chosen nN samples, and log-number of choices of nN elements among αN ones. (Right) The magnetization order parameter m (in other words the teacher–student overlap) as a function of the Gardner volume v for a pool of cardinality $\alpha = 3$, as extracted from the large deviation computations. As is physically intuitive, smaller Gardner volumes imply larger values of the magnetization.

In writing so we took into account the fact that that $y = \pm 1$, which implies also to replace the integral over y in the energetic part by a sum over $\{\pm 1\}$. Furthermore,

$$I_Y^1(y,\eta) = 1 + e^{\phi} \int D\zeta H\left(-\frac{y}{\sqrt{r-Q}}\left(\sqrt{Q-q}\zeta + \sqrt{q}\eta\right)\right)^{\beta}, \tag{89}$$

from which it follows that the energetic term in equation (74) reads:

$$\alpha \int D\eta \sum_{y=\pm 1} I_Y^0(y,\eta) \ln I_Y^1(y,\eta) = 2\alpha \int D\eta H\left(-\sqrt{\frac{m^2}{q-m^2}}\eta\right) \times \ln\left[1 + e^{\phi} \int D\zeta H\left(-\frac{1}{\sqrt{r-Q}}\left(\sqrt{Q-q}\zeta + \sqrt{q}\eta\right)\right)^{\beta}\right]. \tag{90}$$

The $y = 1$ and $y = -1$ being equal modulo a double change of variable $\zeta, \eta \to -\zeta, -\eta$ whence the factor 2. Thus for the perceptron:

$$\Phi_{\mathrm{RS}} = \underset{\hat{m},\hat{r},\hat{q},\hat{Q},r,q,Q}{\mathrm{extr}} \left\{ \frac{\beta}{2} r\hat{r} - \beta m\hat{m} - \frac{\beta(\beta-1)}{2} Q\hat{Q} + \frac{\beta^2}{2} q\hat{q} - \frac{\beta-1}{2} \ln(1 + \hat{r} + \hat{Q}) - \frac{1}{2} \ln(1 + \hat{r} - (\beta-1)\hat{Q} + \beta\hat{q}) + \frac{\beta}{2} \frac{\hat{q} + \hat{m}^2}{1 + \hat{r} - (\beta-1)\hat{Q} + \beta\hat{q}} + 2\alpha \int D\eta H\left(-\sqrt{\frac{m^2}{q-m^2}}\eta\right) \ln\left[1 + e^{\phi} \int D\zeta H\left(-\frac{1}{\sqrt{r-Q}}\left(\sqrt{Q-q}\zeta + \sqrt{q}\eta\right)\right)^{\beta}\right] \right\}. \tag{91}$$

4. Large deviation results

In figure 1, we show the results of the large deviation analysis at $\alpha = 3$. Note that the qualitative picture is unaltered when α is varied (e.g. equivalent results for $\alpha = 10$ are shown in section 3.2). The different curves, obtained at fixed values of the budget n, show the complexity (i.e. the exponential rate of the number) of possible subset choices, Σ, that realize the corresponding Gardner volumes v. As expected, the maximum of each curve is observed at $\beta = 0$, and yields the typical Garner volume of a teacher-student perceptron that has learned to correctly classify nN i.i.d. Gaussian input patterns. The associated complexity is simply given by the binomial distribution.

The cases where the extremum in equation (9) is realized for positive values of β describe choices of the labeled subsets that induce atypically large Gardner volumes: these correspond to AL scenarios where the student query is worse than random sampling. The number of possible realizations of these scenarios

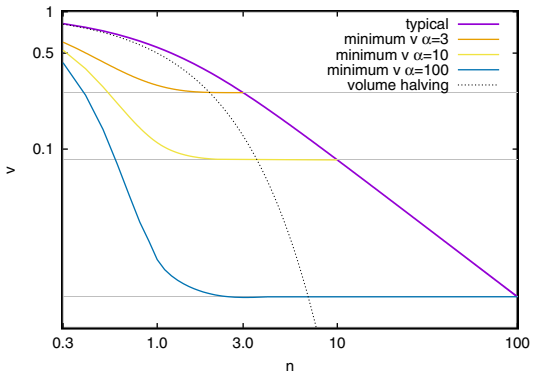

Figure 2. Typical (maximum complexity) Gardner volume (purple, decreasing linearly at large α) and information theoretically smallest achievable one (orange, yellow and blue) extracted from the large deviation computation for $\alpha \in \{3, 10, 100\}$. The horizontal lines depict the value of the Gardner volume corresponding to the whole pool, we see the fast saturation of the lowest volumes at these lines. The information-theoretic volume-halving limit 2^{-n} for label-agnostic AL procedures is plotted in a dotted line. We notice that the qualitative picture is essentially unchanged when α is varied.

decreases exponentially as one approaches the right-hand extremum of where the complexity curve is positive, describing the largest possible volume at that given budget n. An important remark is that as soon as $\beta > 0$, the statistics of the input patterns in the labeled set is no longer i.i.d. but has increasing correlations for larger β.

On the other side, negative β induces atypically small Gardner volumes and labeled subsets with high information content. Again, as one spans smaller and smaller volumes the associated complexity drops, making the problem of finding these desirable subsets harder and harder. The left positive-complexity extremum of the curves in the left plot of figure 1 corresponds to the smallest reachable Gardner volumes. We observe in the figure that for larger values of budgets the complexity curves saturate fast very close to the smallest possible Garner volume corresponding to the Gardner volume for entire pool of samples $v(\alpha)$, suggesting that past a certain budget querying additional examples does not add much information.

In the right plot of figure 1, we also show the prediction for the typical value of the magnetization, i.e. the overlap between teacher and students, as the Gardner volume is varied. As mentioned in section 2, small Gardner volumes induce high magnetizations and thus low generalization errors.

In figure 2 the typical (purple) and corresponding minimum (orange, yellow, cyan) Gardner volumes are depicted as a function of the budget n for various pool sizes $\alpha = 3,\ 10,\ 100$. Note that the qualitative picture is unaltered when α is varied. We further observe that the minimum volume becomes very close to the Gardner volume of the entire pool of samples $v(\alpha)$ already for very small budgets n.

4.1. Additional numerical confirmation

We supply numerical evidence for some assumptions made in this work, in particular the replica trick (12) and the use of the Gardner volume as a measure of informativeness (see section 2 in the main text).

First, we sample numerically at random subsets of cardinalities $n \in \{0.3, 0.6, 0.9, 2.7\}$ out of a pool of cardinal given by $\alpha = 3$, and plot the complexity extracted therefrom in figure 3. The volumes were evaluated using the approximate message passing (AMP) algorithm 2, and simulations were performed at $N = 20$, with 10^7 draws, for a fixed teacher. Such a large number of samples and small system size is needed in order to access exponentially rare events. For larger sizes the probability of rare event is exponentially smaller and proportionally more samples would be needed, which is not computationally accessible. Because of the $\mathcal{O}(10^7 \times N^2)$ complexity N has been kept small, while AMP is known to be valid only in the $N \uparrow \infty$ limit, hence inducing errors due to finite size. Nevertheless, the agreement with the theoretical curves for $\Sigma(n, v)$ is quite good.

We finally present a numerical check of the theoretical prediction for the $m(v)$ curves, see figure 1. At large instance size, $N = 2000$, it is not computationally feasible to obtain sufficient statistics for observing the large deviations of the volume through passive subset sampling, as it was done in the previous experiment. Thus, we resorted to the label-informed AL-AMP AL strategy (see section 5, algorithm 1 and table 1) for biasing the subset selection towards more/less informative subsets. In particular, we constructed each subset by mixing varying ratios of maximally informative samples (selected according to the informed AL-AMP procedure) and minimally informative samples (selected according to the same procedure but with the reversed sorting order). In figure 4, the pool size is $\alpha = 10$ and the budget is fixed to $n = 1.5$. For each subset, the AMP algorithm 2 was run to get the estimator $\hat{\boldsymbol{x}}$ and the magnetization $m = \frac{x^0 \cdot \hat{x}}{N}$ was deduced therefrom.

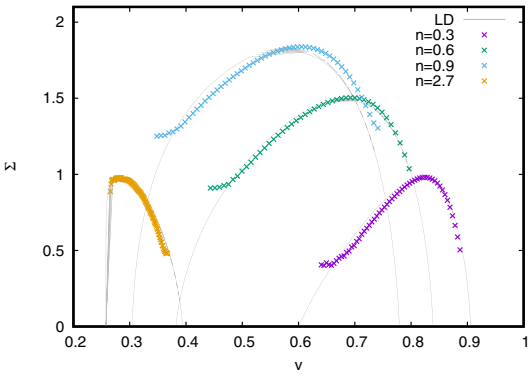

Figure 3. Complexity vs volume curves for $\alpha = 3$, and $n \in \{0.3, 0.6, 0.9, 2.7\}$. The dots are the values extracted from numerical experiments performed at $N = 20$ by repeatedly sampling passively 10^7 times a subset of cardinality n out of a fixed pool of size $\alpha = 3$. Solid lines are the theoretical complexities as predicted by the large deviation computations, see also figure 1. Volumes were evaluated using the AMP algorithm 2. The agreement is rather good knowing the discrepancies that ought to be expected because of running AMP algorithm 2 at finite and small N.

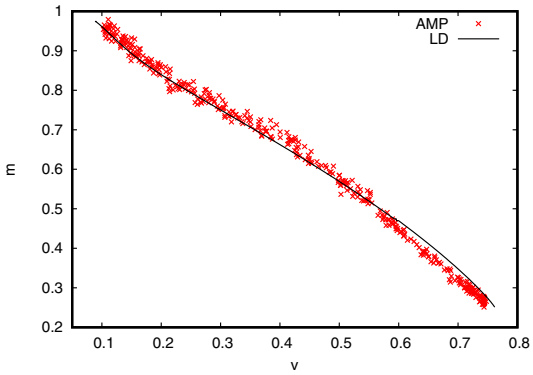

Figure 4. Magnetization m against Gardner volume v for various subsets. The experiments were performed at system size $N = 2 \times 10^3$, pool size $\alpha = 10$ and budget $n = 1.5$. Subsets covering a wide range of volumes were designed by varying the ratio of informative samples (using label-informed AL-AMP, see section 5) and uninformative samples (selected using simple passive learning). Magnetizations and volumes were evaluated for each subset by training the model using the AMP procedure 2. In solid line is the typical $m(v)$ curve predicted by the large deviation computations, which agrees quite well with the numerical simulations.

This incidentally corroborates once more that using the Gardner volume instead of the magnetization to judge for the informativeness of a selection is coherent.

4.2. Stability of the replica symmetric solution

We remark at this point that the presented replica calculation was obtained in the so-called RA ansatz. In general, it is possible for the RS result not to be asymptotically exact, requiring replica symmetry breaking (RSB) in order to evaluate the correct free entropy $\Phi(\beta, \phi)$ [26]. In this model, while RSB is surely not needed close to the maximum of the complexity curves as implied by results in [18], it conversely has to be taken into account away from typical volumes. The volumes for which the RS ansatz (19)–(23) ceases to be valid can be evaluated by considering an infinitesimal perturbation thereof of 1 step RSB form as is detailed in appendix B, yielding the stability condition:

$$\left|\left(2\beta\frac{\hat{q}+\hat{m}^2}{(1+\hat{r}-(\beta-1)\hat{Q}+\beta\hat{q})^3}-\frac{1}{(1+\hat{r}-(\beta-1)\hat{Q}+\beta\hat{q})^2}\right)\right| \times\left|-\frac{\partial}{\partial q}\left(\frac{2\alpha e^{2\phi}}{\beta^2(r-Q)(Q-q)}\int D\rho H\left(-\sqrt{\frac{m^2}{q-m^2}}\rho\right)\left(\frac{\int D\zeta\,\zeta H^\beta(X)}{1+e^\phi\int D\zeta H^\beta(X)}\right)^2\right)\right|<1. \tag{92}$$

Figure 5 shows in solid lines the regions in v space where the RS assumption holds. At the same time, the presence of RSB usually entails corrections that are very small in magnitude. We hence believe the error bounds here reported under the assumption of replica symmetry to achieve a good degree of accuracy.

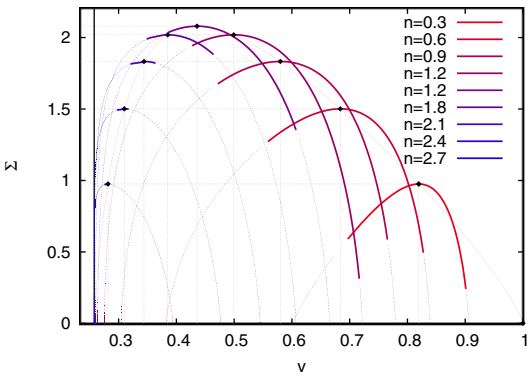

Figure 5. Complexity vs volume curves for $\alpha = 3$, and $n \in \{0.3, 0.6, 0.9, 2.7\}$, see also figure 1. Solid lines corresponds to ranges of values for v where the replica symmetric ansatz is consistent (stable). As n increases the window of validity of the ansatz shrinks, signaling the emergence of non-trivial structures in the geometrical organization of the selection vectors associated with atypical Gardner volumes.

5. Algorithmic implications for active learning

5.1. Generic considerations

The setting investigated in this paper provides a unique chance to benchmark the algorithmic performance of any given pool-based AL algorithm against the optimal achievable performance, and to measure how closely are the large deviations results approached. This offers a controlled setup complementary to the usual one of simply comparing different algorithms on standard datasets, where no point of reference exists to evaluate the performances. The aim of the present section is hence to illustrate how the results on Gardner volumes reported above may serve to evaluate existing AL procedures. Before moving to such algorithmic performance we should make a distinction between two possible classes of AL scenarios:

- Label-agnostic settings, where the student has no prior knowledge on the ground truth labels. In other words, the AL selection must be based solely on the knowledge of the input patterns $\{\boldsymbol{x}\}$, and make no use of the true labels $\{y\}$. In this case, for binary labels there is a simple lower bound on the Gardner volume reachable with nN samples, $v \geqslant 2^{-n}$, which is obtained by the argument that every new sample can at best divide the current volume by a factor two, see [4]. This strategy is explored in the famous QBC AL strategy, and the classical work [4] argues that the volume halving can be actually achieved when an unlimited set of samples is available. Plotting this volume-halving bound in figure 2 we see that even though there exist subsets of the pool that would lead to smaller Gardner volumes, they cannot be found in a label-agnostic way.
- Label-informed settings, where external knowledge on the true labels is available and can be used for extracting more information during the selection process. In many real-world applications the structure in the input data could be exploited (e.g. through clustering, transfer learning, etc) for making unsupervised guesses of the labels and for bootstrapping an AL strategy. A concrete example where external insight is available is drug discovery [7], where additional information can be inferred from the presence of chemical functional groups (or absence thereof) on the molecules in the data pool. In the present work, we study whether it is possible, with full access to the labels, to devise an efficient method for finding a subset of samples that achieves close to the minimal Gardner volume bound (note that this is still an algorithmically non-trivial problem).

In this section we will investigate both the label-agnostic and label-informed strategies. We will benchmark several well known AL algorithms on the model studied in the present paper as well as design and test a new message passing based AL algorithm. Before doing that let us describe the general strategy.

Many of the commonly used AL criteria rely on some form of label-uncertainty measure. Uncertainty sampling [1, 29] is an AL scheme based on the idea of iteratively selecting and labelling data-points where the prediction of the available trained model is the least confident. In general, the computational complexity associated to this type of scheme is of order $\mathcal{O}(N^3)$, requiring an extensive number of runs of a training algorithm (which can scale as $\mathcal{O}(N^2)$ at best). Since even training a single model per pattern addition can become expensive in the large N setting, in all our numerical tests we opted for adding to the labeled set batches of $k = 20$ samples instead of a single sample per iteration. We remark that, despite the k-fold speed-up, the observed performance deterioration is negligible. The structure of this type of algorithm is sketched in algorithm 1.

Algorithm 1. Uncertainty sampling.

Select *heuristic strategy* from table 1
Define batch size k
Initialize $S \subset \mathcal{S} = \{\boldsymbol{F}_\mu\}_{1\leqslant\mu\leqslant\alpha N}$ ($|S| > 0$)
while $|S| < nN$ **do**
  Obtain *required estimates* given S
  Obtain model predictions at data-points in S^c
  Sort predictions according to *sorting criterion*
  Add first k elements in the sorting permutation to S
end while

Algorithm 2. Single-instance AMP for the perceptron.

Initialize $\hat{\boldsymbol{x}} \leftarrow 0$
Initialize $\hat{\boldsymbol{\Delta}} \leftarrow 1$
Initialize $\boldsymbol{g}_{\text{out}} \leftarrow 0$
while Convergence criterion not satisfied **do**
  $\Gamma_\mu^t \leftarrow \sum_i (F_i^\mu)^2 \hat{\Delta}_i^{t-1}$
  $\omega_\mu^t \leftarrow \sum_i F_i^\mu \hat{x}_i^{t-1} - \Gamma_\mu^t g_{\text{out}}^{t-1}$
  $g_{\text{out},\mu}^t \leftarrow \frac{y^\mu}{\sqrt{2\pi\Gamma_\mu^t}} \frac{e^{-\frac{(\omega_\mu^t)^2}{2\Gamma_\mu^t}}}{H\left(-\frac{y^\mu \omega_\mu^t}{\sqrt{\Gamma_\mu^t}}\right)}$
  $(\Sigma_i^t)^{-1} \leftarrow -\sum_\mu (F_i^\mu)^2 (-\frac{\omega^t}{V^t} g_{\text{out}}^t - (g_{\text{out}}^t)^2)$
  $R_i^t \leftarrow \hat{x}_i^{t-1} + \Sigma_i^t \sum_\mu F_i^\mu g_{\text{out}}^t$
  $\hat{x}_i^t \leftarrow \frac{R_i^t}{1+\Sigma_i^t}$
  $\hat{\Delta}_i^t \leftarrow \frac{\Sigma_i^t}{1+\Sigma_i^t}$
end while

5.2. Approximate message passing for AL (AL-AMP)

In general, estimating the Gardner volume on a given training set or the label-uncertainty of a new sample is a computationally hard problem. However, in perceptrons (or more general GLMs) with i.i.d. Gaussian input data $\boldsymbol{F}$, at large system size N one can rely on the estimate provided by a well known algorithm for approximate inference, AMP (historically also referred to as the Thouless–Anderson–Palmer (TAP) equations, see [30]). AMP is a standard iterative procedure used for Bayesian inference on a factor graph associated to a probability measure. We refer the interested reader to the ample literature dedicated to message-passing algorithms ([12, 23, 31] for example) for more detailed discussion thereon. The AMP algorithm [13, 32, 33], yields (at convergence) an estimator of the posterior means, $\hat{\boldsymbol{x}}$, and variances, $\hat{\boldsymbol{\Delta}}$, thus accounting for uncertainty in the inference process including the label of a new sample. The Gardner volume v (corresponding to the so-called Bethe free entropy) can then be expressed as a simple function of the AMP fixed-point messages (see [34] for an example). We provide a pseudo-code of AMP in the case of the perceptron in algorithm 2. An important remark is that when the training set is not sampled randomly from the pool, as in the AL context, correlations can arise and AMP is no longer rigorously guaranteed to converge nor to provide a consistent estimate of the Gardner volume. In the present work, we can only argue that its employment seems to be justified *a posteriori* by observing the agreement between theoretical predictions and numerical experiments for instance for the generalization error.

We use the AMP algorithm to introduce a new uncertainty sampling procedure relying on the information contained in the AMP messages, denoted as AL-AMP in the following. At each iteration, the single-instance AMP equations are run on the current training subset to yield posterior mean estimate $\hat{\boldsymbol{x}}$ and variance $\hat{\boldsymbol{\Delta}}$. These quantities can then be used to evaluate, for all the unlabeled samples, the output magnetization (i.e. the Bayesian prediction) defined as

$$m_{\text{out}}^\mu = \text{erf}(\omega^\mu/\sqrt{2\Gamma^\mu}) \qquad \forall\mu|\sigma_\mu = 0, \tag{93}$$

where we introduced the output overlaps $\boldsymbol{\omega} = \boldsymbol{F}' \cdot \hat{\boldsymbol{x}}$ and variances $\boldsymbol{\Gamma} = (\boldsymbol{F}' \odot \boldsymbol{F}') \cdot \hat{\boldsymbol{\Delta}}$, where $\odot$ is the component-wise product. The output magnetizations correspond to the weighted output average over all the estimators contained in the current version space, and their magnitude represents the overall confidence in

Table 1. table summarizing the specifics of the uncertainty sampling strategies considered in this paper.

Uncertainty sampling strategies		
Heuristic	Required estimates	Sorting criterion
Agnostic AL-AMP	$\hat{\boldsymbol{x}}_{\rm AMP}, \hat{\boldsymbol{\Delta}}_{\rm AMP}$	$\mathrm{argmin}_{\mu} \left\lvert \mathrm{erf}\left(\frac{\boldsymbol{F}'^{\mu}\hat{\boldsymbol{x}}_{\rm AMP}}{\sqrt{2(\boldsymbol{F}'^{\mu})^2\hat{\boldsymbol{\Delta}}_{\rm AMP}}} \right) \right\rvert$
Informed AL-AMP	$\hat{\boldsymbol{x}}_{\rm AMP}, \hat{\boldsymbol{\Delta}}_{\rm AMP}$	$\mathrm{argmax}_{\mu} \left\lvert y^{\mu} - \mathrm{erf}\left(\frac{\boldsymbol{F}'^{\mu}\hat{\boldsymbol{x}}_{\rm AMP}}{\sqrt{2(\boldsymbol{F}'^{\mu})^2\hat{\boldsymbol{\Delta}}_{\rm AMP}}} \right) \right\rvert$
Query by committee	$\{\boldsymbol{x}^k_{\rm SGD}\}_{k=1}^{K}$	$\mathrm{argmin}_{\mu} \lvert \sum_{k=1}^{K} \mathrm{sign}\left(\boldsymbol{F}'^{\mu}\cdot\boldsymbol{x}^k_{\rm SGD}\right) \rvert$
Logistic regression	$\boldsymbol{x}_{\rm log}$	$\mathrm{argmin}_{\mu} \lvert \boldsymbol{F}'^{\mu}\cdot\boldsymbol{x}_{\rm log} \rvert$
Perceptron learning	$\boldsymbol{x}_{\rm perc}$	$\mathrm{argmin}_{\mu} \lvert \boldsymbol{F}'^{\mu}\cdot\boldsymbol{x}_{\rm perc} \rvert$

the classification of the still unlabeled samples. This means that AMP provides an extremely efficient way of obtaining the information on uncertainty. The specifics of the algorithm can be found in table 1.

We also explore numerically the label-informed AL regime introduced in the previous section. We consider its limiting case by introducing the informed AL-AMP strategy, which can fully access the true labels $\boldsymbol{Y}$ in order to query the samples $\boldsymbol{F}^{\mu}$ whose output magnetisation $m^{\mu}_{\rm out}$ (93) is maximally distant from the correct classification y^{μ}. This selection process can iteratively reduce the Gardner volume of factors larger than 2. Again, the relevant specifics of informed AL-AMP algorithm are detailed in table 1.

5.3. Other tested measures of uncertainty

One of the widely used uncertainty sampling procedure is the so-called QBC strategy [4, 16]. In QBC, at each time step, a committee of K students is to be sampled from the version space (e.g. via the Gibbs algorithm). The committee is then employed to choose the labels to be queried, by identifying the samples where maximum disagreement in the committee members outputs is observed. The QBC algorithm was introduced as a proxy for doing bisection, i.e. cutting version space into two equal-volume halves. As already mentioned, this constitutes the optimal information gain in an label-agnostic setting [35]. Note that, however, the QBC procedure can achieve volume-halving only in the infinite-size committee limit, $K\uparrow\infty$, with uniform version space sampling and with availability of infinitely many samples. Obviously, running a large number K of ergodic Gibbs sampling procedures quickly becomes computationally unfeasible. Moreover in the pool-based AL the pool of samples is limited. In order to allow comparison with other strategies at finite sizes, we approximated the uniform sampling with a set of greedy optimization procedures (e.g. stochastic gradient descent) from random initialization conditions, checking numerically that this yields a committee of students reasonably spread out in version space. It is possible to ensure a greater coverage of the version space by performing a short Monte-Carlo random walk for each committee member. The effect has been found to be small for computationally reasonable lengths of walk.

We also implemented an alternative uncertainty sampling strategy, relying on a single training procedure (e.g. training with the perceptron algorithm or logistic regression) per iteration: in this case, the uncertainty information is extracted from the magnitude of the pre-activations measured at the unlabeled samples after each training cycle. This strategy implements the intuitive geometric idea of looking for the samples that are most orthogonal to the available updated estimator, which are more likely to halve the version space independently of the value of the true label.

All the tested procedures, with the exception of QBC, display a complexity of approximately $\mathcal{O}(N^3)$, possibly to be corrected by a factor accounting for the number of iterations required at each training step. The adapted QBC algorithm similarly has complexity $\mathcal{O}(KN^3)$, as it involves a committee of K models. Ideally, the original QBC algorithm [4] would require at each step to sample uniformly the current version space, thus implying an exponentially costly Monte Carlo step.

5.4. Algorithmic results

In figure 6, we compare the minimum Gardner volume obtained from the large deviation calculation with the algorithmic performance obtained on synthetic data at finite size, $N = 2\times10^3$, by the AL-AMP algorithms detailed in algorithm 1 and table 1. The data-pool size is fixed to $\alpha = 3$. The large deviation analysis yields values for the minimum and maximum achievable Gardner volumes at any budget n. We compare the algorithmic results also with the prediction for the typical case and with the volume-halving curve 2^{-n}. Since in the considered pool-based setting the volume-halving performance cannot be achieved for volumes smaller than the Gardner volume corresponding to the entire pool $v(\alpha)$, the relevant volume-halving bound should be more precisely $\max(2^{-n}, v(\alpha))$. Random sampling displays good

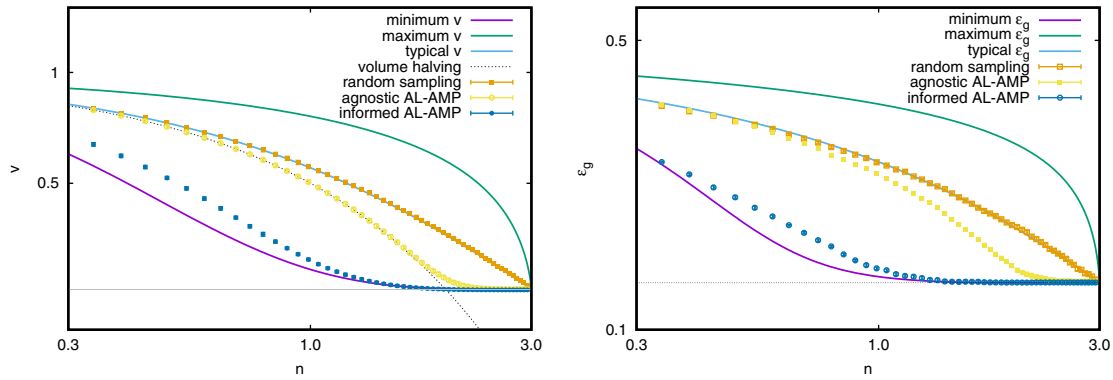

Figure 6. (Left) Performance of the label-agnostic (yellow circles) and label-informed (blue circles) AL-AMP, plotted together with the minimum and maximum values of the Gardner volume extracted from the large deviation computation (purple and green) and the volume-halving curve (dotted black). For comparison we also plot the typical Gardner volume (cyan) and the one obtained by random sampling (orange squares). Numerical experiments were run for system size $N = 2 \times 10^3$ and pool size $\alpha = 3$. For each algorithmic performance curve the average over 10 samples is presented. Fluctuations were found to be negligible around the average and are not shown. (Right) The same plot with the Gardner volume replaced by the Bayesian test accuracy, derived in appendix C. For the AL-AMP algorithm the accuracy is evaluated using a test set of size $P_{\text{test}} = 5 \times 10^4$. The qualitative picture is very similar to the one for the Gardner volume curves (left), once more confirming that Gardner volumes and generalization errors both constitute good measures for informativeness.

agreement with the expected typical volumes. Most notably, the label-agnostic AL-AMP algorithm tightly follows the volume-halving bound $\max(2^{-n}, v(\alpha))$, thus reaching close to optimal possible performance. Since for large α the behavior of $v(\alpha) = \text{const.}/\alpha$ [15] we conclude that the AL-AMP algorithm will reach close to minimum possible Gardner volumes for a budget $n \sim \mathcal{O}[\log(\alpha)]$.

Theoretical justification of this very good performance is however yet to be established: as a matter of fact, due to the purely sequential nature of the employed AL scheme we are not accounting for possible correlations between subsequent queries. Moreover, in machine learning practice, higher order methods for AL (like expected model change [36], with a $\mathcal{O}(N^4)$ complexity) have been shown to perform potentially better than uncertainty sampling. We conjecture the observed optimal performance of AL-AMP can be traced back to the noiseless nature of the considered learning setting and leave the problem of studying AL in the presence of noise for future work. Note that, even in our pool-based setting, we thus obtain an exponential reduction in the number of samples, similar to the original QBC work [4].

The label-informed AL-AMP also approaches the theoretically minimal volume but not as closely. We remark that an important limit of the AL-AMP algorithm comes from the fact that AMP is not guaranteed to provide good estimators (or converge at all) with correlated data. For example, in the numerical experiments for obtaining the informed AL-AMP curve, we had to resort to mild damping schemes in the message-passing to allow fixed-points being reached. This effect was stronger for the label-informed algorithm than for the label-agnostic one.

In figure 7, we provide a numerical comparison of the performance of the agnostic AL-AMP and the other above mentioned label-agnostic AL algorithms. The finite size experiments were run at $N = 2 \times 10^3$, while here we set $\alpha = 10$. Note that, while the mentioned different AL strategies where employed for selecting the labeled subset, in all cases supervised learning and the related performance estimates were obtained by running AMP. In the plot, we can see that, while AL-AMP is able to extract very close to the maximum amount of information from each query (one bit per pattern, until the volume $v(\alpha)$ is saturated), other heuristics with the same computational complexity are sub-optimal. In particular, in the simplified QBC we observe that increasing the size K of the committee does not yield very noticeable change in its performance, most probably because the committee cannot cover a sufficient portion of the version space if the computational cost is to be kept reasonable. On the other hand, using the information of the magnitude pre-activations allows better performance while being also more time-efficient, since only a single perceptron, rather than a committee thereof, has to be trained at each step. The logistic loss allows a rather good performance, close to that of AL-AMP, while the uncertainty sampling with the perceptron algorithm yields a mitigated performance.

We leave a more systematic bench-marking of the many existing strategies for future work, stressing the fact that, while there certainly exist more involved procedures that can yield better performance than the presented heuristics, the absolute performance bounds still apply, agnostic of the implemented AL strategy. Because the reported results in the present work are specific to the GLM setup with Gaussian data, future investigations should also be concerned with extending the AL-AMP procedure to real-word datasets, as

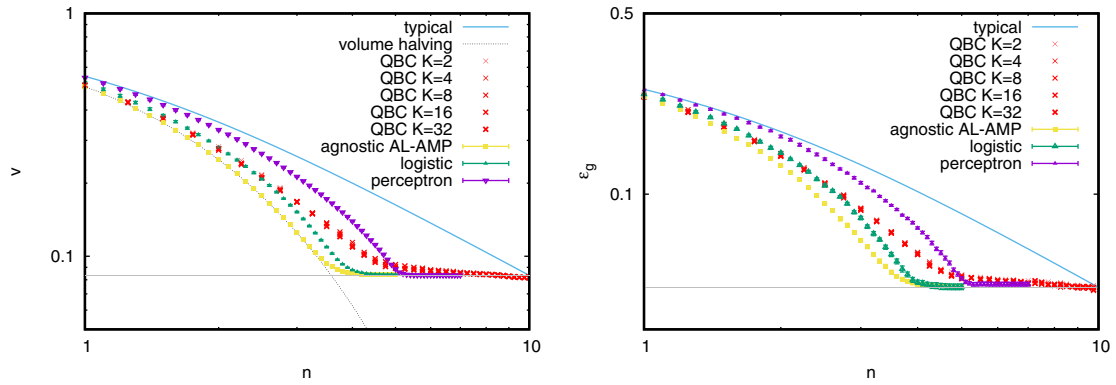

Figure 7. (Left) Performance of the label-agnostic algorithms presented in table 1 is plotted against the budget n and compared to the volume-halving lower bound. Experiments were performed at system size $N = 2 \times 10^3$, and pool size $\alpha = 10$. For each algorithm the average over 10 samples is presented. Note that error bars are smaller than marker size. (Left) (Bayesian) test accuracy of the same heuristics for various budgets n. The test set size was chosen to be $P_{\text{test}} = 5 \times 10^4$. In blue the Bayesian test accuracy for a typical subset, see appendix C. Again, the qualitative picture is unchanged going from the Gardner volume to the test accuracy.

opposed to the synthetic data used in the present work, and observe whether the good performance generalizes. It should be stressed that such endeavor would likely require to use a variant of AMP, namely vectorial AMP [37], more robust to mismatches between data and model assumptions.

6. Conclusions

Using the replica method for large deviation calculation of the Gardner volume, we computed for the teacher-student perceptron model the minimum Gardner volume (equivalently, maximum mutual information) achievable by selecting a subset with fixed cardinality from a pre-existing pool of i.i.d. normal samples. We evaluated the large deviation function based on the RS assumption; checking for RSB and evaluating the eventual corrections to the presented results is left for future work, as well as rigorous establishment of the presented results. Our result for the information-theoretic limit of pool-based AL in this setting complements the already known volume-halving bound for label-agnostic strategies. We hope our result may serve as a guideline to benchmark future heuristic algorithms on the present model, while our modus operandi regarding the derivation of the large deviations may help for future endeavor in theoretical analysis of AL in more realistic settings. We presented the performance of some known heuristics, plus we suggested the AL-AMP algorithms to perform the uncertainty based AL. We show numerically that on the present model the label-agnostic AL-AMP algorithm performs very close to the optimal bound, thus being able to achieve accuracy corresponding to the entire pool of samples with exponentially fewer samples.

Data availability statement

No new data were created or analysed in this study.

Acknowledgments

We want to thank Guilhem Semerjian for clarifying discussions in early stages on this work. This work is supported by the ERC under the European Union's Horizon 2020 Research and Innovation Program 714608-SMiLe.

Appendix A. Saddle point equations for the perceptron

A.1. Saddle-point equations for the perceptron

The canonical way of carrying out the extremization in equation (91) is to take the saddle-point equations (zero-gradient conditions) and to solve them. The saddle point equations associated to $\Phi(\beta, \phi)$ (equation (91)) read:

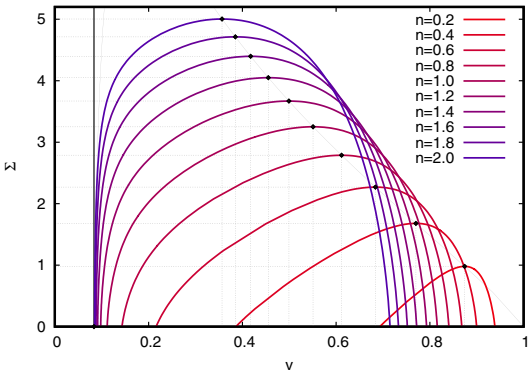

Figure A1. Complexity $\Sigma(n, v)$ as a function of the Gardner volume v for budgets $n \leqslant 2$ at pool size $\alpha = 10$ extracted from the large deviation computations, see also figure 1 for the same curves at different pool size. For any budget n the maximum complexity corresponds to the volume reached by random subset selection (passive learning), see discussion in section 3 in the main text. We invite the reader to notice that the qualitative allure of the complexity curves remains significantly unchanged as the pool size α is varied (see figure 1 for $\alpha = 3$).

$$m^t = \frac{\hat{m}^t}{1+\hat{r}^t-(\beta-1)\hat{Q}^t+\beta\hat{q}^t} \tag{A.1}$$

$$q^t = \frac{\hat{q}^t+(\hat{m}^t)^2}{(1+\hat{r}^t-(\beta-1)\hat{Q}^t+\beta\hat{q}^t)^2} \tag{A.2}$$

$$Q^t = \frac{\hat{q}^t+(\hat{m}^t)^2}{(1+\hat{r}^t-(\beta-1)\hat{Q}^t+\beta\hat{q}^t)^2} + \frac{1}{\beta}\frac{1}{1+\hat{r}^t-(\beta-1)\hat{Q}^t+\beta\hat{q}^t} - \frac{1}{\beta}\frac{1}{1+\hat{r}^t+\hat{Q}^t} \tag{A.3}$$

$$r^t = \frac{\hat{q}^t+(\hat{m}^t)^2}{(1+\hat{r}^t-(\beta-1)\hat{Q}^t+\beta\hat{q}^t)^2} + \frac{1}{\beta}\frac{1}{1+\hat{r}^t-(\beta-1)\hat{Q}^t+\beta\hat{q}^t} + \frac{\beta-1}{\beta}\frac{1}{1+\hat{r}^t+Q^t} \tag{A.4}$$

$$X^t = -\frac{1}{\sqrt{r^t-Q^t}}\left(\sqrt{Q^t-q^t}\zeta+\sqrt{q^t}\eta\right) \tag{A.5}$$

$$\hat{q}^{t+1} = 2\alpha\int D\eta H\left(-\sqrt{\frac{(m^t)^2}{q^t-(m^t)^2}}\eta\right)\frac{e^{2\phi}}{2\pi(r^t-Q^t)}\left[\frac{\int D\zeta H(X^t)^{\beta-1}e^{-\frac{1}{2}(X^t)^2}}{1+e^{\phi}\int D\zeta H(X^t)^{\beta}}\right]^2 \tag{A.6}$$

$$\hat{m}^{t+1} = 2\alpha\int d\eta\frac{1}{(2\pi)^{\frac{3}{2}}\sqrt{r-Q}}\frac{1}{\sqrt{1-\frac{(m^t)^2}{q^t}}}e^{-\frac{1}{2}\frac{\eta^2}{1-\frac{(m^t)^2}{q^t}}}\frac{e^{\phi}\int D\zeta H(X^t)^{\beta-1}e^{-\frac{1}{2}(X^t)^2}}{1+e^{\phi}\int D\zeta H(X^t)^{\beta}} \tag{A.7}$$

$$\hat{Q}^{t+1} = \frac{2\alpha}{r^t-Q^t}\int D\eta H\left(-\sqrt{\frac{(m^t)^2}{q^t-(m^t)^2}}\eta\right)\frac{\int D\zeta H(X^t)^{\beta-2}e^{-(X^t)^2}\frac{1}{2\pi}}{1+e^{\phi}\int D\zeta H(X^t)^{\beta}}, \tag{A.8}$$

$$\hat{r}^{t+1} = -\frac{2\alpha}{r^t-Q^t}\int D\eta H\left(-\sqrt{\frac{(m^t)^2}{q^t-(m^t)^2}}\eta\right)\frac{\int D\zeta H(X^t)^{\beta-1}X^te^{-\frac{1}{2}(X^t)^2}\frac{1}{\sqrt{2\pi}}}{1+e^{\phi}\int D\zeta H(X^t)^{\beta}}. \tag{A.9}$$

In practice, equations (A.1)–(A.9) are iterated until convergence. The time indices are derived from an independent computation using AMP [13, 24], not shown here. They indicate in which order the equations ought to be iterated in order to converge. Remark that the update schedule very simply consists in updating in parallel all order parameters m, q, Q, r then all auxiliary (hatted) order parameters $\hat{m}, \hat{q}, \hat{Q}, r$. After convergence the order parameters can be used in equation (91) to evaluate the free entropy Φ and subsequently evaluate the complexity $\Sigma(n, v)$ by inverting the Legendre transform (9).

We present for illustration in figure A1 the complexity curves for $\alpha = 10$ for some budgets n, and refer the interested reader to figure 1 for the same plot at pool size $\alpha = 3$. As the budget n is increased, smaller values of Gardner volumes become accessible, provided sufficiently informative subset are found. A detailed discussion can further be found in section 3.

Appendix B. Stability of the replica symmetric ansatz

In this appendix we investigate the stability of the RS ansatz (19)–(23) under an infinitesimal perturbation, which we choose to be of one-step RSB (1RSB) form. This is tantamount to ascertaining whether the extremum in the optimization problem (16) is reached outside of the subspace of RS matrices (19)–(23). We first give the expression of the 1RSB free entropy in the general case of a GLM, before specializing it to the perceptron. Eventually, we analyze the stability of the 1RSB saddle-point equations under infinitesimal departure from the RS ansatz, and give a stability condition for the RS assumption to be consistent.

B.1. 1RSB ansatz

We depart from the RS setting and assume the disorder (selection variables σ^a) to be 1RSB, with Parisi parameter noted τ [25, 26]. For convenience we shall replace $1 \leqslant a \leqslant s$ by a double index (a', a), with $1 \leqslant a' \leqslant \frac{s}{\tau}$ the 1RSB cluster index and $1 \leqslant a \leqslant \tau$ indexing the selection variables inside a same block. As a consequence the $\boldsymbol{x}$ variables will carry a triple index (a', a, α). We also need to differentiate the former q index into q_1 (for students seeing different disorders pertaining to the same 1RSB cluster) and q_0 (for students seeing disorders from different clusters).

B.2. Inverse and determinant of a 1RSB planted matrix

Let us note as before $\tilde{\boldsymbol{Q}} = \boldsymbol{Q}^{-1}$ with elements $\tilde{r}^0, \tilde{m}, \tilde{r}, \tilde{Q}, \tilde{q_1}, \tilde{q_0}$. Inverting the matrix is tantamount to solving the equations:

$$r_0\tilde{r_0} + \beta s m \tilde{m} m = 1, \tag{B.1}$$

$$r_0\tilde{m} + m\tilde{r} + (beta - 1)m\tilde{Q} + \beta(\tau - 1)m\tilde{q_1} + \beta(s - \tau)m\tilde{q_0} = 0, \tag{B.2}$$

$$m\tilde{r_0} + r\tilde{m} + (\beta - 1)Q\tilde{m} + \beta(\tau - 1)q_1\tilde{m} + \beta(s - \tau)q_0\tilde{m} = 0, \tag{B.3}$$

$$m\tilde{m} + r\tilde{r} + (\beta - 1)Q\tilde{Q} + \beta(\tau - 1)q_1\tilde{q_1} + \beta(s - \tau)q_0\tilde{q_0} = 1, \tag{B.4}$$

$$m\tilde{m} + r\tilde{Q} + Q\tilde{r} + (\beta - 2)Q\tilde{Q} + \beta(\tau - 1)q_1\tilde{q_1} + \beta(s - \tau)q_0\tilde{q_0} = 0, \tag{B.5}$$

$$m\tilde{m} + r\tilde{q_1} + (\beta - 1)Q\tilde{q_1} + q_1\tilde{r} + (\beta - 1)q_1\tilde{Q} + \beta(\tau - 2)q_1\tilde{q_1} + \beta(s - \tau)q_0\tilde{q_0} = 0, \tag{B.6}$$

$$\begin{aligned} &m\tilde{m} + r\tilde{q_0} + (\beta - 1)Q\tilde{q_0} + \beta(\tau - 1)q_1\tilde{q_0} + q_0\tilde{r} + (\beta - 1)q_0\tilde{Q} + \beta(\tau - 1)q_0\tilde{q_1} \\ &+ \beta(s - 2\tau)q_0\tilde{q_1} + \beta(s - 2\tau)q_0\tilde{q_0} = 0. \end{aligned} \tag{B.7}$$

Solutions are :

$$\tilde{r_0} = \frac{(1-\beta)Q - r + \beta(\tau - s)q_0 + \beta(1-\tau)q_1}{-rr^0 + Qr^0(1-\beta) + \beta m s^2 + \beta r^0(q_1(1-\tau) + q_0(\tau - s))}, \tag{B.8}$$

$$\tilde{m} = \frac{m}{-rr^0 + Qr^0(1-\beta) + \beta m s^2 + \beta r^0(q_1(1-\tau) + q_0(\tau - s))}, \tag{B.9}$$

$$\begin{aligned} \tilde{r} = &\frac{1}{\beta}\left(\frac{\beta - 1}{r - Q} + \frac{1}{(\beta - 1)Q - \beta q_1 + r}\right) \\ &+ \frac{r_0}{s\beta[(\beta - 1)Qr_0 + rr_0 + \beta(-ms^2 + q_0(q_1(\tau - 1) + q_0(s - \tau)))]} \\ &+ \frac{1}{\beta}\frac{Q(1-\beta) - r + \beta(q_1(1-s) + q_0 s)}{s((1-\beta)Q - \beta q_1 + r)((1-\beta)Q + r + \beta q_1(\tau - 1) - \beta q_0\tau)}, \end{aligned} \tag{B.10}$$

$$\tilde{Q} = \frac{r_0}{-\beta s(-rr^0 + Qr^0(1-\beta) + \beta ms^2 + \beta r^0(q_1(1-\tau) + q_0(\tau - s)))} + \frac{1}{(\beta-1)Q - \beta q_1 + r}\left[\frac{Q-q_1}{Q-r} + \frac{Q(1-\beta) - r + \beta(q_1(1-s) + q_0 s)}{\beta s((\beta-1)Q + r + \beta q_1(\tau-1) - \beta q_0 \tau))}\right], \tag{B.11}$$

$$\tilde{q_1} = \frac{r^0}{-\beta s(-rr^0 + Qr^0(1-\beta) + \beta ms^2 + \beta r^0({}_1(1-\tau) + q_0(\tau - s)))} + \frac{Q(1-\beta) - r + \beta(q_1(1-s) + q_0 s)}{\beta s((\beta-1)Q - \beta q_1 + r)((\beta-1)Q + r + \beta q_1(\tau-1) - \beta q_0 \tau))}, \tag{B.12}$$

$$\tilde{q_0} = \frac{-q_0 r^0 + m^2}{(r^0(Q(1-\beta) - r + \beta(q_1(1-\tau) + q_0(\tau - s))) + \beta ms^2)} \times \frac{+}{(\beta-1)Q + r + \beta(q_1(\tau-1) - q_0\tau)}. \tag{B.13}$$

One can also compute det$\boldsymbol{Q}$ by finding the eigenvectors. There is a couple of eigenvectors with product

$$r^0(r + (\beta-1)Q + \beta(\tau-1)q_1 + \beta(s-\tau)q_0) - \beta s m^2, \tag{B.14}$$

and $\frac{s}{\tau}$ eigenvectors with eigenvalue:

$$r + (\beta-1)Q - \beta\tau q_0 + \beta(\tau-1)q_1, \tag{B.15}$$

and $\frac{s}{\tau}(\tau-1)$ eigenvectors with eigenvalue:

$$r + (\beta-1)Q - \beta q_1. \tag{B.16}$$

Thus,

$$\begin{aligned}\det\boldsymbol{Q} = {} & \ln(r^0(r + (\beta-1)Q + \beta(\tau-1)q_1 + \beta(s-\tau)q_0) - \beta s m^2) \\ & + (\frac{s}{\tau} - 1)\ln(r + (\beta-1)Q - \beta\tau q_0 + \beta(\tau-1)q_1) + \frac{s}{\tau}(\beta\tau - 1)\ln(r + (\beta-1)Q - \beta q_1).\end{aligned} \tag{B.17}$$

B.3. Computing the 1RSB free entropy

The trace term Tr$\hat{\boldsymbol{Q}}\boldsymbol{Q}$ in (16) now reads with the 1RSB ansatz

$$r_0\hat{r}_0 + \beta s m\hat{m} - \frac{r\hat{r}}{2}\beta s + \frac{\beta(\beta-1)}{2}sQ\hat{Q} + \frac{\beta s}{2}\beta(\tau-1)q_1\hat{q}_1 + \frac{\beta s}{2}\beta(s-\tau)q_0\hat{q}_0. \tag{B.18}$$

For the I_X term we proceed in very similar manner as in the RS case, see section 3.1. The end result is:

$$I_X = \frac{1}{\tau}\int D\eta I_X^0 \ln I_X^1, \tag{B.19}$$

$$I_X^0 = \int dx^0 \overline{P_X}(x^0) e^{-\frac{\hat{m}^2}{2\hat{q}_0}(x^0)^2 + \frac{\hat{m}}{\hat{q}_0}\sqrt{\hat{q}_0}\eta x^0}, \tag{B.20}$$

$$I_X^1 = \int D\xi\left[\int D\lambda\left(\int dx P_X(x) e^{-\frac{\hat{r}+\hat{Q}}{2}x^2 + (\sqrt{\hat{Q}-\hat{q}_1}\lambda + \sqrt{\hat{q}_1-\hat{q}_0}\xi + \sqrt{\hat{q}_0}\eta)}\right)^\beta\right]^\tau. \tag{B.21}$$

Note that the results

$$\hat{r}^0 = 0 \tag{B.22}$$

$$r^0 = \mathbb{E}_{x^0}(x^0)^2, \tag{B.23}$$

also carry through to the 1RSB case. The treatment of the I_Y term requires, as in the non-RSB case, more care. The computation is very analogous (though more lengthy) and shall not be further detailed presently. We report only the end result:

$$I_Y = \frac{\alpha}{\tau}\int D\rho I_Y^0 \ln I_Y^1, \tag{B.24}$$

$$I_Y^0 = \int Dh^0 \overline{P_{\text{out}}}\left(y|\sqrt{\frac{q_0 r^0 - m^2}{q_0}}h^0 + \sqrt{\frac{m^2}{q_0}}\right), \tag{B.25}$$

$$I_Y^1 = \int D\eta \left[1 + e^{\phi}\int D\zeta \left(P_{\text{out}}\left(y|\sqrt{r-Q}h + \sqrt{Q-q_1}\zeta + \sqrt{q_1-q_0} + \sqrt{q_0}\rho\right)\right)^{\beta}\right]^{\tau}. \tag{B.26}$$

In summary the 1RSB free entropy reads:

$$\begin{aligned}\Phi_{\text{1RSB}} = \underset{r,m,Q,q_1,q_0,\hat{r},\hat{m},\hat{Q},\hat{q_1},\hat{q_0}}{\text{extr}} & -\beta m\hat{m} + \frac{\beta}{2}r\hat{r} - \frac{\beta(\beta-1)}{2}Q\hat{Q} - \frac{\beta\tau}{2}\beta(\tau-1)q_1\hat{q}_1 \\ & + \frac{\beta^2\tau}{2}q_0\hat{q}_0 + \frac{1}{\tau}\int D\eta I_X^0 \ln I_X^1 + \frac{\alpha}{\tau}\int D\rho I_Y^0 \ln I_Y^1,\end{aligned} \tag{B.27}$$

$$I_X^0 = \int dx^0 \overline{P_X}(x^0) e^{-\frac{\hat{m}^2}{2\hat{q}_0}(x^0)^2 + \hat{m}\sqrt{\hat{q}_0}\eta x^0}, \tag{B.28}$$

$$I_X^1 = \int D\xi \left[\int D\lambda \left(\int dx P_X(x) e^{-\frac{\hat{r}+\hat{Q}}{2}x^2 + (\sqrt{\hat{Q}-\hat{q}_1}\lambda + \sqrt{\hat{q}_1-\hat{q}_0}\xi + \sqrt{\hat{q}_0}\eta)}\right)^{\beta}\right]^{\tau}, \tag{B.29}$$

$$I_Y^0 = \int Dh^0 \overline{P_{\text{out}}}\left(y|\sqrt{\frac{q_0 r^0 - m^2}{q_0}}h^0 + \sqrt{\frac{m^2}{q_0}}\right), \tag{B.30}$$

$$I_Y^1 = \int D\eta \left[1 + e^{\phi}\int D\zeta \left(P_{\text{out}}\left(y|\sqrt{r-Q}h + \sqrt{Q-q_1}\zeta + \sqrt{q_1-q_0} + \sqrt{q_0}\rho\right)\right)^{\beta}\right]^{\tau}. \tag{B.31}$$

B.4. Specialization to the perceptron

We now proceed to specialize the 1RSB free entropy (B.27), derived above for a generic GLM, for the special case of a perceptron. The only term that differs non-trivially from the RS specialization reported in section 3.2 is I_X^1, for which one further (Gaussian) integration has to be carried out. We give the final formula for the perceptron free entropy:

$$\begin{aligned}\Phi_{\text{1RSB}} = & -\beta m\hat{m} + \frac{\beta}{2}r\hat{r} - \frac{\beta(\beta-1)}{2}Q\hat{Q} - \frac{\beta}{2}\beta(\tau-1)q_1\hat{q}_1 + \frac{\beta^2\tau}{2}q_0\hat{q}_0 \\ & - \frac{\beta-1}{2}\ln(1+\hat{r}+\hat{Q}) + \frac{1}{2}\left(\frac{1}{\tau}-1\right)\ln(1+\hat{r}-(\beta-1)\hat{Q}+\beta\hat{q}_1) \\ & - \frac{1}{2\tau}\ln(1+\hat{r}-(\beta-1)\hat{Q}+\beta\hat{q}_1-\beta\tau(\hat{q}_1-\hat{q}_0)) \\ & + \frac{\beta}{2}\frac{\hat{q}_0+\hat{m}^2}{1+\hat{r}-(\beta-1)\hat{Q}+\beta\hat{q}_1-\beta\tau(\hat{q}_1-\hat{q}_0)} + \frac{2\alpha}{\tau}\int D\rho H\left(-\sqrt{\frac{m^2}{q_0-m^2}}\rho\right) \\ & \times \ln\left\{\int D\eta\left[1+e^{\phi}\int D\zeta H\left(-\frac{\sqrt{Q-q_1}\zeta + \sqrt{q_1-q_0}\eta + \sqrt{q_0}\rho}{\sqrt{r-Q}}\right)^{\beta}\right]^{\tau}\right\}.\end{aligned} \tag{B.32}$$

B.5. Stability analysis

Having derived the free entropy (B.32) in the 1RSB ansatz, we can now study the stability of the RS solution with respect to an infinitesimal 1RSB perturbation. To that end we consider a 1RSB ansatz departing infinitesimally from the RS form:

$$q_1 - q_0 = \epsilon = o(1), \tag{B.33}$$

$$\hat{q}_1 - \hat{q}_0 = \hat{\epsilon} = o(1). \tag{B.34}$$

We name $q, \hat{q}$ the common order 0 value for these overlaps. The entropic part (trace term and I_X) of the free entropy (B.32) then reads to first order in $\epsilon, \hat{\epsilon}$

$$-\beta m\hat{m}+\frac{\beta}{2}r\hat{r}-\frac{\beta(\beta-1)}{2}Q\hat{Q}-\frac{\beta-1}{2}\ln(1+\hat{r}+\hat{Q})-\frac{\beta^2(\tau-1)}{2}(q\hat{q}+\epsilon\hat{q}+\hat{\epsilon}q)$$
$$+\frac{\beta^2\tau}{2}q\hat{q}-\frac{1}{2}\ln(1+\hat{r}-(\beta-1)\hat{Q}+\beta\hat{q})+\frac{\beta}{2}\frac{\hat{q}+\hat{m}^2}{1+\hat{r}-(\beta-1)\hat{Q}+\beta\hat{q}}$$
$$-\frac{\beta}{2}\frac{(\hat{q}+\hat{m}^2)\beta(1-\tau)\hat{\epsilon}}{(1+\hat{r}-(\beta-1)\hat{Q}+\beta\hat{q})^2}. \tag{B.35}$$

Prior to expanding the energetic part I_Y it is convenient to define the shorthand

$$X=-\frac{1}{\sqrt{r-Q}}(\sqrt{Q-q}\zeta+\sqrt{q}\rho), \tag{B.36}$$

as we did in the RS computations, see section 3.2. Expanding the argument of $*=\log\int D\eta\left[1+e^{\phi}\int D\zeta H(\ldots)^{\beta}\right]^{\tau}$

$$*=\log\left\{\int D\eta[1+e^{\phi}\int D\zeta H(X)^{\beta}\right.$$
$$\left.+e^{\phi}\int D\zeta(H^{\beta})'(X)\left(\frac{\zeta\epsilon}{2\sqrt{r-Q}\sqrt{Q-q}}-\frac{\sqrt{\epsilon}\eta}{\sqrt{r-Q}}\right)+\frac{1}{2}(H^{\beta})''(X)\frac{\epsilon\eta^2}{r-Q}+\ldots]^{\tau}\right\} \tag{B.37}$$

$$=\log\left\{\left(1+e^{\phi}\int D\zeta H(X)^{\beta}\right)^{\tau}\right.$$
$$\times\int D\eta\left[1++\epsilon\tau\frac{e^{\phi}\int D\zeta((H^{\beta})'(X)\frac{\zeta}{2\sqrt{r-Q}\sqrt{Q-q}}+\frac{1}{2}(H^{\beta})''(X)\frac{\eta^2}{r-Q})}{1+e^{\phi}\int D\zeta H^{\beta}(X)}\right.$$
$$\left.\left.+\epsilon\eta^2\frac{\tau(\tau-1)}{2}\left(\frac{\frac{e^{\phi}}{r-Q}\int D\zeta(H^{\beta})'(X)}{1+e^{\phi}\int D\zeta H^{\beta}(X)}\right)^2\right]\right\} \tag{B.38}$$

$$=\tau\log(1+e^{\phi}\int D\zeta H(X)^{\beta})+\log\left\{1+\epsilon\frac{\tau(\tau-1)}{2}\left(\frac{\frac{e^{\phi}}{r-Q}\int D\zeta(H^{\beta})'(X)}{1+e^{\phi}\int D\zeta H^{\beta}(X)}\right)^2\right\}$$
$$=\tau\log(1+e^{\phi}\int D\zeta H(X)^{\beta})+\epsilon\frac{\tau(\tau-1)}{2}\left(\frac{\frac{e^{\phi}}{r-Q}\int D\zeta(H^{\beta})'(X)}{1+e^{\phi}\int D\zeta H^{\beta}(X)}\right)^2. \tag{B.39}$$

In going from the first to the second line we used that the $\mathcal{O}(\epsilon^{\frac{1}{2}})$ term is killed by the integration over η. The first fraction in the second line is found to be vanishing using an integration by parts.

The variables $\epsilon,\hat{\epsilon}$ intervene in the saddle-point Equations. If iterating the SP equations induce $\epsilon,\hat{\epsilon}$ to become large, then the assumption that the RS fixed point is stable ceases to hold. To obtain the dynamical equations for the pair $\epsilon,\hat{\epsilon}$ one has to derive the SP equations associated to zero-gradient conditions in the q and $\hat{q}$ direction. Note that the derivatives of terms not involving ε or $\hat{\epsilon}$ shall eventually sum to zero, since we assume to be at the RS ($\epsilon=\hat{\epsilon}=0$) fixed points. The first $\partial_{\hat{q}}\Phi_{\mathrm{1RSB}}=0$ equation reads

$$\epsilon=-\left(2\beta\frac{\hat{q}+\hat{m}^2}{(1+\hat{r}-(\beta-1)\hat{Q}+\beta\hat{q})^3}-\frac{1}{(1+\hat{r}-(\beta-1)\hat{Q}+\beta\hat{q})^2}\right)\hat{\epsilon}, \tag{B.40}$$

while the $\partial_q\Phi_{\mathrm{1RSB}}=0$ implies:

$$\hat{\epsilon}=\left(-\frac{2\alpha}{\beta^2}\frac{\partial}{\partial q}\int D\rho H\left(-\sqrt{\frac{m^2}{q-m^2}}\rho\right)\left(\frac{\frac{e^{\phi}}{r-Q}\int D\zeta(H^{\beta})'(X)}{1+e^{\phi}\int D\zeta H^{\beta}(X)}\right)^2\right)\epsilon \tag{B.41}$$

$$=-\frac{\partial}{\partial q}\left(\frac{2\alpha e^{2\phi}}{\beta^2(r-Q)(Q-q)}\int D\rho H\left(-\sqrt{\frac{m^2}{q-m^2}}\rho\right)\left(\frac{\int D\zeta\,\zeta H^{\beta}(X)}{1+e^{\phi}\int D\zeta H^{\beta}(X)}\right)^2\right)\epsilon. \tag{B.42}$$

We choose not to explicit the q derivative as the expression is rather large. The stability condition, expressing the fact that the dynamical system (B.40) and (B.41) converges to $\epsilon=\hat{\epsilon}=0$, then reads:

$$\left|\left(2\beta\frac{\hat{q}+\hat{m}^2}{(1+\hat{r}-(\beta-1)\hat{Q}+\beta\hat{q})^3}-\frac{1}{(1+\hat{r}-(\beta-1)\hat{Q}+\beta\hat{q})^2}\right)\right| \times\left|-\frac{\partial}{\partial q}\left(\frac{2\alpha e^{2\phi}}{\beta^2(r-Q)(Q-q)}\int D\rho H\left(-\sqrt{\frac{m^2}{q-m^2}}\rho\right)\left(\frac{\int D\zeta\,\zeta H^\beta(X)}{1+e^\phi\int D\zeta H^\beta(X)}\right)^2\right)\right|<1. \tag{B.43}$$

The stability condition (92) can be evaluated numerically using the values for the order parameters resulting from the convergence of the saddle-point equations (A.1)–(A.9). The results are presented in figure 5. As the budget n increases, the region of validity of the RS assumptions diminishes. This seems to suggest that the selections $\{\sigma_\mu\}_{1\leqslant\mu\leqslant\alpha N}$ leading to atypical enough volumes v are grouped into isolated clusters in configuration space, i.e. the landscape is 1RSB, this effect being more pronounced for larger values of budget. That the RS assumption should break down away from typicality is expected. However, in most known settings, this simple ansatz proves to be a very good approximation nonetheless, and taking into account further steps of RS breaking usually yields only minor improvement [12]. We therefore believe that the error bounds here reported have a good degree of accuracy, although they may certainly be improved if further symmetry breaking are taken into account. The precise evaluation of these corrections is left for future investigation.

Appendix C. Optimal generalization error for the large deviation perceptron

We derive here the expression for the optimal generalization error ε_g (in the Bayesian sense) associated to a subset of volume v and of budget n as a function of the perceptron order parameters m and q, see sections 3.1 and 3.2. The Bayesian ε_g was introduced for example in [38] and, unlike the test error yielded by training the perceptron on some loss, is independent of the training procedure and thus may serve as a nice measure of informativeness for subsets. For the usual perceptron model, it is known that the optimal generalization error is achieved when the student classification is performed by averaging the predicted label over the student measure $P_X(\cdot)P_{\rm out}(\boldsymbol{Y}|\boldsymbol{F}\cdot)$ and taking the sign thereof. Note that the average predicted label is but the output magnetization $\boldsymbol{m}_{\rm out}$ discussed in section 5 of the main text. Transposition to the large deviation setting, which allows to fix the budget n and the volume v, is straightforward provided one averages over the large deviation measure (11). By definition the test error is the probability that a new sample $\boldsymbol{F}_{\rm new}\overset{d}{=}\mathcal{N}(0,1)$ is correctly classified by the student according to the output magnetization $\mathbb{E}_{\boldsymbol{x}}^{\beta,\phi}{\rm sgn}(\boldsymbol{x}\cdot\boldsymbol{F}_{\rm new})$

$$1-\epsilon_g=\mathbb{E}_{\boldsymbol{x}^0,\boldsymbol{Y},\boldsymbol{F},\boldsymbol{F}_{\rm new}}\Theta[{\rm sgn}(\boldsymbol{x}^0\cdot\boldsymbol{F}_{\rm new})\mathbb{E}_{\boldsymbol{x}}^{\beta,\phi}{\rm sgn}(\boldsymbol{x}\cdot\boldsymbol{F}_{\rm new})], \tag{C.1}$$

where $\mathbb{E}_{\boldsymbol{x}}^{\beta,\phi}$ denote the average with respect to the large deviation posterior measure (11) with control parameters β and ϕ. Introducing an integral representation for a Dirac delta and expanding the resulting exponential:

$$1-\epsilon_g=\mathbb{E}_{\boldsymbol{x}^0,\boldsymbol{Y},\boldsymbol{F}_{\rm new}}\int d\nu d\hat{\nu}e^{i\nu\hat{\nu}}\Theta[{\rm sgn}(\boldsymbol{x}^0\cdot\boldsymbol{F}_{\rm new})\nu]\sum_{j=0}^{\infty}\frac{(i\hat{\nu})^j}{j!}\mathbb{E}_{\boldsymbol{F}}(\mathbb{E}_{\boldsymbol{x}}^{\beta,\phi}{\rm sgn}(\boldsymbol{x}\cdot\boldsymbol{F}_{\rm new}))^j \tag{C.2}$$

$$=\mathbb{E}_{\boldsymbol{F}_{\rm new}}\sum_{j=0}^{\infty}\int d\nu d\hat{\nu}e^{i\nu\hat{\nu}}\frac{(i\hat{\nu})^j}{j!}\mathbb{E}_{\boldsymbol{x}^0,\boldsymbol{Y}}\Theta[{\rm sgn}(\boldsymbol{x}^0\cdot\boldsymbol{F}_{\rm new})\nu]\mathbb{E}_{\boldsymbol{F}}(\mathbb{E}_{\boldsymbol{x}}^{\beta,\phi}{\rm sgn}(\boldsymbol{x}\cdot\boldsymbol{F}_{\rm new}))^j. \tag{C.3}$$

For any fixed j, the computation of $\mathbb{E}_{\boldsymbol{x}^0,\boldsymbol{Y}}\Theta[{\rm sgn}(\boldsymbol{x}^0\cdot\boldsymbol{F}_{\rm new})\nu]\mathbb{E}_{\boldsymbol{F}}(\mathbb{E}_{\boldsymbol{x}}^{\beta,\phi})^j$ is formally very similar to the one detailed in section 3.1 and follows the same lines. First notice that using the replica trick $\mathbb{E}_{\boldsymbol{x}}^{\beta,\phi}{\rm sgn}(\boldsymbol{x}\cdot\boldsymbol{F}_{\rm new})$ prescribes to introduce as precedently βs replicas

$$\mathbb{E}_{\boldsymbol{x}}^{\beta,\phi}{\rm sgn}(\boldsymbol{x}\cdot\boldsymbol{F}_{\rm new})=\lim_{s\to0}\Xi^{s-1}\sum_{\{\sigma_\mu\}}e^{\phi\sum_\mu S_\mu}\left[\int d\boldsymbol{x}P_X(\boldsymbol{x})\prod_\mu P_{\rm out}(y^\mu|\boldsymbol{F}^\mu\boldsymbol{x})^{\sigma_\mu}\right]^{\beta-1} \times\int d\boldsymbol{x}P_X(\boldsymbol{x})\prod_\mu P_{\rm out}(y^\mu|\boldsymbol{F}^\mu\boldsymbol{x})^{\sigma_\mu}{\rm sgn}(\boldsymbol{x}\cdot\boldsymbol{F}_{\rm new}) \tag{C.4}$$

$$=\lim_{s\to0}\sum_{\boldsymbol{\sigma}^1,\ldots,\boldsymbol{\sigma}^n}\int\prod_{a=1}^{s}\prod_{\alpha=1}^{\beta}d\boldsymbol{x}^{a\alpha}P_X(\boldsymbol{x})\prod_\mu P_{\rm out}(y^\mu|\boldsymbol{F}^\mu\boldsymbol{x}^{a\alpha})^{\sigma_\mu^a}e^{\phi\sum_{a,\mu}\sigma_\mu^a} \times{\rm sgn}(\boldsymbol{x}^{11}\cdot\boldsymbol{F}_{\rm new}). \tag{C.5}$$

From which it follows that:

$$(\mathbb{E}_{\boldsymbol{x}}^{\beta,\phi}\mathrm{sgn}(\boldsymbol{x}\cdot\boldsymbol{F}_{\mathrm{new}}))^j = \lim_{s\to 0}\sum_{\substack{\boldsymbol{\sigma}^{la}\\ 1\leqslant l\leqslant j, 1\leqslant a\leqslant s}}\int\prod_{a=1}^{s}\prod_{l=1}^{j}\prod_{\alpha=1}^{\beta}d\boldsymbol{x}^{a\alpha}P_X(\boldsymbol{x}^{la\alpha})\prod_{\mu}P_{\mathrm{out}}(y^\mu|\boldsymbol{F}^\mu\boldsymbol{x}^{la\alpha})^{\sigma_\mu^{la}}$$
$$\times e^{\phi\sum_{l,a,\mu}\sigma_\mu^{la}}\prod_{l=1}^{j}\mathrm{sgn}(\boldsymbol{x}^{l11}\cdot\boldsymbol{F}_{\mathrm{new}}). \tag{C.6}$$

The net effect is simply to transform the first index a into a double index (l,a), thereby introducing a third level of replication. Pursuing equation (C.2),

$$1-\epsilon_g = \lim_{s\to 0}\sum_{j=0}^{\infty}\int d\nu d\hat{\nu}e^{i\nu\hat{\nu}}\frac{(i\hat{\nu})^j}{j!}\mathbb{E}_{\boldsymbol{x}^0,\boldsymbol{Y},\boldsymbol{F}}\sum_{\substack{\boldsymbol{\sigma}^{la}\\ 1\leqslant l\leqslant j, 1\leqslant a\leqslant s}}\int\prod_{a=1}^{s}\prod_{l=1}^{j}\prod_{\alpha=1}^{\beta}d\boldsymbol{x}^{a\alpha}P_X(\boldsymbol{x}^{la\alpha})$$
$$\times\prod_{\mu}P_{\mathrm{out}}(y^\mu|\boldsymbol{F}^\mu\boldsymbol{x}^{la\alpha})^{\sigma_\mu^{la}}e^{\phi\sum_{l,a,\mu}\sigma_\mu^{la}}\mathbb{E}_{\boldsymbol{F}_{\mathrm{new}}}\prod_{l=1}^{j}\mathrm{sgn}(\boldsymbol{x}^{l11}\cdot\boldsymbol{F}_{\mathrm{new}})\Theta[\mathrm{sgn}(\boldsymbol{x}^0\cdot\boldsymbol{F}_{\mathrm{new}})\nu] \tag{C.7}$$

$$=\lim_{s\to 0}\sum_{j=0}^{\infty}\int d\nu d\hat{\nu}e^{i\nu\hat{\nu}}\frac{(i\hat{\nu})^j}{j!}e^{sjN\Phi(\beta,\phi)}\int dh^0\prod_{l=1}^{j}dh^l\frac{e^{-\frac{1}{2}\boldsymbol{h}^T\mathcal{Q}^{-1}\boldsymbol{h}}}{\sqrt{(2\pi)^{j+1}\det\mathcal{Q}}}\Theta[\mathrm{sgn}(h^0)\nu]\prod_{l=1}^{j}\mathrm{sgn}(h^l). \tag{C.8}$$

In going from the second to the last line we introduced overlap variables h as in section 3.1. The rest of the large deviation measure in equation (C.6) factorize into $e^{Nsj\Phi(\beta,\phi)}$, with Φ the free entropy computed in sections 3.1 and 3.2, and goes to 1 as the $s\to 0$ limit is taken. We also introduced an overlap matrix $\mathcal{Q}$ of RS form

$$\mathcal{Q}_{00} = r^0 \tag{C.9}$$

$$\mathcal{Q}_{0l} = m\ \ \forall 1\leqslant l\leqslant j \tag{C.10}$$

$$\mathcal{Q}_{ll} = r\ \ \forall 1\leqslant l\leqslant j \tag{C.11}$$

$$\mathcal{Q}_{lk} = q\ \ \forall 1\leqslant l\neq k\leqslant j. \tag{C.12}$$

The order parameters r^0, r, m and q were defined in the replica ansatz equations (19)–(23). Note that the relevant overlap is q, rather than Q, since the j-fold replication in equation (C.6) affects also the selection variables σ and thus the variables $\boldsymbol{x}^{l11}$ see different disorders, see section 3.1. The inverse $\tilde{\mathcal{Q}}=\mathcal{Q}^{-1}$ is characterized by the coefficients:

$$\tilde{r}^0 = \frac{r+(j-1)q}{r^0(r+(j-1)q)-jm^2} \tag{C.13}$$

$$\tilde{r} = \frac{r^0(r+(j-2)q)-(j-1)m^2}{(r-q)(r^0(r+(j-1)q)-jm^2)} \tag{C.14}$$

$$\tilde{m} = \frac{-m}{(r^0(r+(j-1)q)-jm^2} \tag{C.15}$$

$$\tilde{r} = \frac{m^2-r^0q}{(r-q)(r^0(r+(j-1)q)-jm^2)}. \tag{C.16}$$

The last integral in equation (C.8) can be taken care of in the usual manner, by decomposing the exponent and introducing a Hubbard–Stratonovitch field η, see for example section 3.1. Similarly the expression can then be factorized in l indices. The result is:

$$\int dh^0 \prod_{l=1}^{j} dh^l e^{-\frac{1}{2}\boldsymbol{h}^T \mathcal{Q}^{-1}\boldsymbol{h}} \Theta[\mathrm{sgn}(h^0)v] \prod_{l=1}^{j} \mathrm{sgn}(h^l)$$

$$= \int \frac{D\eta dh^0}{\sqrt{(2\pi)^{j+1}\det\mathcal{Q}}} e^{-\frac{1}{2}\tilde{r}^0(h^0)^2} e^{\frac{j}{2}\left(\sqrt{\frac{-\tilde{q}}{\tilde{r}-\tilde{q}}}\eta - \frac{\tilde{m}}{\sqrt{\tilde{r}-\tilde{q}}}h^0\right)^2} \left[1 - 2H\left(\sqrt{\frac{-\tilde{q}}{\tilde{r}-\tilde{q}}}\eta - \frac{\tilde{m}}{\sqrt{\tilde{r}-\tilde{q}}}h^0\right)\right]^j, \tag{C.17}$$

$$= \int D\eta Dh^0 \left[1 - 2H\left(\sqrt{\frac{m^2 - qr^0}{r^0(q-r)}}\eta - \sqrt{\frac{m^2}{r^0(r-q)}}h^0\right)\right]^j. \tag{C.18}$$

This terminates the computation of ϵ_g, since from equation (C.8):

$$1 - \epsilon_g = \int D\eta Dh^0 \Theta\left[h^0\left(1 - 2H\left(\sqrt{\frac{m^2 - qr^0}{r^0(q-r)}}\eta - \sqrt{\frac{m^2}{r^0(r-q)}}h^0\right)\right)\right] \tag{C.19}$$

$$= \int D\eta Dh^0 \Theta[h^0(\sqrt{r^0 q - m^2}\eta - mh^0)] \tag{C.20}$$

$$= 1 - \frac{1}{\pi}\cos^{-1}\frac{m}{\sqrt{q}}. \tag{C.21}$$

We used the fact that $x \to 1 - 2H(x)$ was odd and the fact that $r^0 = 1$ for the perceptron model with Gaussian priors, see section 3.2.

ORCID iD

H Cui https://orcid.org/0000-0003-4648-244X

References

[1] Settles B 2009 Active learning literature survey computer sciences *Technical Report*
[2] Angluin D 1988 *Mach. Learn.* **2** 319–42
[3] Cohn D A, Atlas L E and Ladner R E 1994 *Mach. Learn.* **15** 201–21
[4] Seung H S, Opper M and Sompolinsky H 1992 *Conf. Learning Theory* vol 5 pp 287–94
[5] Atlas L E, Cohn D A and Ladner R E 1990 *Advances in Neural Information Processing Systems* vol 2 pp 566–73
[6] Zhang L, Lin D Y, Wang H, Car R and Weinan E 2019 *Phys. Rev. Mater.* **3** 023804
[7] Warmuth M K, Rätsch G, Mathieson M, Liao J and Lemmen C 2001 *Advances in Neural Information Processing Systems* vol 14 pp 1449–56
[8] McCallum A K and Nigam K 1998 *Int. Conf. Machine Learning* pp 350–8
[9] Tong S and Koller D 2000 *J. Mach. Learn. Res.* **2** 45–66
[10] Hoi S C, Jin R, Zhu J and Lyu M R 2006 *Int. Conf. Machine Learning* vol 6 pp 417–24
[11] Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271–84
[12] Mézard M and Montanari A 2002 *Information, Physics and Computation* (Oxford : Oxford University Press)
[13] Zdeborová L and Krzakala F 2016 *Adv. Phys.* **5** 453–552
[14] Gardner E and Derrida B 1989 *J. Phys. A: Math. Gen.* **22** 1983
[15] Engels A and van den Broeck C P 2001 *Statistical Mechanics of Learning* (Cambridge: Cambridge University Press)
[16] Freund Y, Shamir E, Seung H S and Tishby N 1992 *Advances in Neural Information Processing Systems* vol 5 pp 483–90
[17] Zhou H J 2019 *Commun. Theor. Phys.* **71** 243
[18] Barbier J, Krzakala F, Macris N, Miolane L and Zdeborová L 2019 *Proc. Natl Acad. Sci.* **116** 5451–60
[19] Cover T and Thomas J 1991 *Elements of Information Theory* (New York: Wiley)
[20] Nishimori H 2001 *Statistical Physics of Spin Glasses and Information Processing* (Oxford: Oxford Science Publications)
[21] Mitchell T M 1982 *Artif. Intell.* **18** 203–26
[22] Dotsenko V, Franz S and Mézard M 1994 *J. Phys. A: Math. Gen.* **27** 2351–65
[23] Krzakala F, Mézard M, Sausset F, Yifan S and Zdeborová L 2012 *J. Stat. Mech.* **2012** 08009
[24] Antenucci F, Krzakala F, Urbani P and Zdeborová L 2019 *J. Stat. Mech.* **2019** 023401
[25] Mézard M, Parisi G, Sourlas N, Toulouse G and Virasoro M A 1984 *J. Phys.* **45** 843–54
[26] Mézard M, Virasoro M A and Parisi G 1986 *Spin Glass Theory and Beyond* (*Lecture Notes in Physics*) (Singapore: World Scientific)
[27] Parisi G 1979 *Phys. Lett.* **73** 203–5
[28] Parisi G 1983 *Phys. Rev. Lett* **50** 1946–8
[29] Lewis D D and Gale W A 1994 *Special Interest Group on Information Retrieval (SIGIR)* vol 17 pp 3–12
[30] Thouless D J, Anderson P W and Palmer R G 1977 *Phil. Mag.* **35** 593–601
[31] Bayati M and Montanari A 2011 *IEEE Trans. Inf. Theory* **57** 764–85
[32] Donoho D L, Maleki A and Montanari A 2009 *Proc. Natl Acad. Sci.* **106** 18914–9
[33] Rangan S 2011 Generalized approximate message passing for estimation with random linear mixing *2011 IEEE Int. Symp. Information Theory Proc.* (IEEE) pp 2168–72
[34] Krzakala F, Manoel A, Tramel E W and Zdeborová L 2014 *IEEE Int. Symp. Information Theory* pp 1499–503

[35] Dasgupta S 2005 *Advances in Neural Information Processing Systems* vol 17 pp 337–44
[36] Cai W, Zhang Y and Zhou J 2013 Maximizing expected model change for active learning in regression *2013 IEEE 13th Int. Conf. Data Mining* (IEEE) pp 51–60
[37] Rangan S, Schniter P and Fletcher A K 2019 *IEEE Trans. Inf. Theory* **65** 6664–84
[38] Baldassi C, Ingrosso A, Lucibello C, Saglietti L and Zecchina R 2015 *Phys. Rev. Lett.* **115** 128101