Statistical mechanics of sparse generalization and graphical model selection

Alejandro Lage-Castellanos\textsuperscript{1,2}, Andrea Pagnani\textsuperscript{2} and Martin Weigt\textsuperscript{2}

\textsuperscript{1} Physics Faculty, University of Havana, La Habana, CP 10400, Cuba
\textsuperscript{2} Institute for Scientific Interchange, Viale Settimio Severo 65, Villa Gualino, I-10133 Torino, Italy

Received 21 July 2009
Accepted 26 September 2009
Published 13 October 2009

Online at stacks.iop.org/JSTAT/2009/P10009
doi:10.1088/1742-5468/2009/10/P10009

Abstract. One of the crucial tasks in many inference problems is the extraction of an underlying sparse graphical model from a given number of high-dimensional measurements. In machine learning, this is frequently achieved using, as a penalty term, the $L_p$ norm of the model parameters, with $p \leq 1$ for efficient dilution. Here we propose a statistical mechanics analysis of the problem in the setting of perceptron memorization and generalization. Using a replica approach, we are able to evaluate the relative performance of naive dilution (obtained by learning without dilution, following by applying a threshold to the model parameters), $L_1$ dilution (which is frequently used in convex optimization) and $L_0$ dilution (which is optimal but computationally hard to implement). Whereas both $L_p$ diluted approaches clearly outperform the naive approach, we find a small region where $L_0$ works almost perfectly and strongly outperforms the simpler to implement $L_1$ dilution.

Keywords: cavity and replica method, disordered systems (theory), neuronal networks (theory)

ArXiv ePrint: 0907.3241
1. Introduction

The problem of extracting sparse information from high-dimensional data is common to various fields of scientific data analysis: computational biology, computer science, combinatorial chemistry, neuroscience, and text processing are just a few examples (see [1, 2] for a general introduction on the subject). Its importance becomes particularly evident in the analysis of biological high-throughput experiments. To give an example, the number of gene probes analyzed simultaneously ranges from the order of tens of thousands in gene expression experiments (e.g. \( \sim 30,000 \) for human DNA chips) to hundreds of thousands in the case of single-nucleotide polymorphisms (\( \sim 500,000 \) for standard genotyping platforms). The information about certain phenotypical traits is, however, expected to be contained in an \textit{a priori} unknown, but small fraction (e.g. \(< 100\)) of all measured probes. These probes may act in a combinatorial way, making their one-by-one extraction impossible. As a further complication, also the number of independent measurements rarely exceeds the order of a few hundreds. Therefore the problem of extracting information from a \textit{few} high-dimensional data points has become a major challenge in biological research. Both the extraction of features related to the phenotypical traits (i.e. purely topological information, here denoted as \textit{graphical model selection}) and the construction of an explicit functional relation between the measured values of these features and the phenotype are of enormous interest.
The literature about feature selection has so far been concentrated around two main strategies: (i) *wrapper* which utilizes learning to score signatures according to their predictive value, (ii) *filters* that fix the signature as a preprocessing step independent from the classification strategy used in the second step. In this work we will present a replica computation on a *wrapper* strategy which falls into the subclass of *embedded* methods where variable selection is performed in the training process of the classifier. More concretely, we will present an analytical teacher–student computation on the properties of a continuous diluted perceptron (i.e. a perceptron where a finite fraction of the coupling parameters are zero). Dilution will be introduced via an external field forcing the student to set as many variables as possible to zero. This external field will be coupled to the $L_p$ norm $\|\vec{J}\|_p = \sum_i |J_i|^p$ of the coupling vector of the student perceptron. For $p \leq 1$, the cusp-like singularity of this function at zero actually sets a fraction of all model parameters exactly to zero, as required for graphical model selection.

This strategy is not new, but so far, most of the more mathematically minded studies in the context of linear regression and various non-linear models [3]–[8] have been concentrating (a) on the case $p = 1$, which is the only case of a convex $L_p$ norm with a cusp at zero, and therefore dilution can be achieved within the framework of convex optimization (this case is well known under the name LASSO [3]); and (b) on the case of a large amount of available data (our model parameter $\alpha$ would scale like $\ln N$ instead of being constant as in our setting), where mathematically rigorous performance guarantees concerning the exact extraction of the topology of the underlying graphical model can be given.

It is, however, obvious that the most efficient dilution should be obtained for $p = 0$, where non-zero parameters are penalized independently of their non-zero value. The non-convexity of the $L_0$ norm introduces however computational complexity. The problem of sparse generalization has already been studied quite extensively by the neural network community, where the problem is also known as node pruning, and a general introduction to the subject can be found in [9]. Some clever greedy local strategy like Optimal Brain Damage [10], and the related Optimal Brain Surgery [11] have been proposed. Interesting statistical mechanics approaches have been presented using the replica method, analyzing both the memorization and generalization performance of optimal diluted perceptrons [12]–[17]. More recently a comparative analysis of continuous versus discrete classifiers has been presented Malzahn [18]. The problem of the inference of a classifier with discrete weights has been analyzed in [19]–[21], where a theoretical computation for the average case together with a message passing algorithm has been proposed. Another way of attacking the problem has been recently proposed by Uda and Kabashima in [22], where dilution is obtained by masking the couplings of a continuous perceptron by Boolean variables.

Here we aim at bridging these different approaches: using the techniques developed in the statistical physics of disordered systems, in particular the replica trick in the replica symmetric approximation, we provide an analysis of the relative performance of various strategies for selecting diluted graphical models in the case of limited data availability. The article is organized as follows. In section 2 we describe the generalization problem, and the replica approach used for its analytical description. In sections 3 and 4 we apply the general results of the replica trick to non-diluted generalization and $L_p$ diluted generalization respectively. The performance of non-diluted and $L_1$ and $L_0$
diluted generalizations are compared in section 5. In section 6 the memorization problem is treated as a noise-dominated limiting case of generalization, and at the end, the main results are reviewed and put into context in the conclusions 7. Three appendices are added to clarify some technical aspects of the mathematical derivations.

2. Generalization and replicas

Two common problems in machine learning are the so-called memorization and generalization problems. In each of them, a number of patterns \( \{ \vec{x}^\mu, \mu \in (1\ldots M) \} \) are classified with labels \( y^\mu \), and one aims at memorizing or inferring a rule that reproduces the given classification. We will study these problems, for the perceptron with continuous weights.

Let us consider the case of \( N \) binary variables \( x_i = \pm 1 \) defining each pattern \( \vec{x}^\mu \). We assume the existence of a hidden relation among these variables and the labels \( y^\mu = \pm 1 \) of each pattern:

\[ y^\mu = \sigma^0(\vec{x}^\mu). \]

The function \( \sigma^0(\vec{x}) \) could be, e.g., the one relating the activated/repressed expression states of genes \( x_i \) to the presence or absence \( y = \pm 1 \) of a disease, or to the expression of another gene not contained in \( \vec{x} \). Unfortunately, \( \sigma^0(\vec{x}) \) is unknown and all we have in general is a set of \( M \) experiments \( \{(y^\mu, \vec{x}^\mu), \mu \in (1\ldots M)\} \), linking patterns \( \vec{x}^\mu \) to labels \( y^\mu \). In supervised learning these experimental data are used as a ‘training set’ to infer the real relations among the variables. As a first approximation, one could mimic the output function \( \sigma^0(\vec{x}) \) as the sign of a linear combination,

\[ \sigma(\vec{J}, \vec{x}^\mu) = \text{sgn}(\vec{J} \cdot \vec{x}^\mu) \]

where the \( N \) weights \( J_i \), also called couplings, are parameters to be tuned in order to reproduce the experimental (training) data. Such a function is called a perceptron. Here the weights \( J \) are allowed to take continuous real values.

The memorization [23,24] and generalization [25,26] problems concern the question of inferring the optimal values of the \( J_s \) from the training data \( \{(y^\mu, \vec{x}^\mu)\} \). For this scope we define the training energy (cost function)

\[ E(\vec{J}) = \sum_{\mu} \Theta(-y^\mu \vec{J} \cdot \vec{x}^\mu) \]  

counting the number of misclassified patterns when \( \vec{J} \) is used to reproduce the training data. The function \( \Theta(\cdot) \) is the Heaviside step function: \( \Theta(x) = 1 \) if \( x > 0 \), and zero otherwise. Note that the function \( E(\vec{J}) \) depends only on the orientation of the vector and not on its length, i.e. \( E(\vec{J}) = E(c\vec{J}) \) for all \( c \neq 0 \).

In general, the real unknown output function \( \sigma^0(\vec{x}) \) will be a complex one, and attempts at reproducing it using a linear perceptron may fail. This means that the training energy will eventually become non-zero if the number of training patterns is sufficiently large. However, we will focus on the case of realizable rules, this is, when the output function \( \sigma^0(\vec{x}) \) is actually a perceptron, and there is always at least one set of weights with zero energy.
The possibility of non-realizability will be accounted for as a random noise affecting the output. In mathematical terms, the training patterns are generated by

$$y^\mu = s^0(\vec{x}) = \text{sgn}(\vec{J}^0 \cdot \vec{x}^\mu + \eta^\mu)$$

where the noises $\eta^\mu$ are i.i.d. Gaussian variables, with variance $\gamma^2$, and the hidden perceptron parameters $\vec{J}^0$ give the rule that we are interested to ‘discover’. We will refer to $\vec{J}^0$ as the teacher, and to the free parameters of our problem $\vec{J}$ as the student, since the latter attempts to reproduce the patterns generated by the former. Note that the training energy (1) does not change when $\vec{J}$ is multiplied by a global scalar factor. To cope with this invariance, we will look for student vectors subject to the spherical constraint $\vec{J} \cdot \vec{J} = N$.

In the zero-noise limit ($\gamma \to 0$), there will be at least one student capable of correctly classifying any amount of training data, namely $\vec{J} = \vec{J}^0$. Upon increasing the noise level ($\gamma > 0$), the correlation between the patterns and the teacher becomes shadowed by the noise, and the student will need a larger amount of patterns to learn from the teacher. If the noise dominates completely $\gamma \to \infty$, there is no information left in the training data about the teacher’s structure, and the student will memorize all patterns up to a critical threshold above which it starts to fail.

In the case of a feasible rule, the number of perfect solutions for the student ($E(\vec{J}) = 0$) is generally large. The entropy of the space of perfect solutions is a decreasing function of the number $M$ of training patterns, since every new pattern imposes a constraint on the student. We can further restrict this space by looking at diluted solutions inside the space of perfect students. A general dilution term can be added to the training energy to form the following Hamiltonian:

$$\beta H(\vec{J}) = \beta E(\vec{J}) + h\|\vec{J}\|_p$$

where $\|\vec{J}\|_p = \sum_i |J_i|^p$ is the $L_p$ norm of the student. The dilution field $h$ will be used to force dilution, and non-diluted generalization corresponds to $h = 0$. Among the different choices of $p$, the case $p = 1$ corresponds to the $L_1$ norm $\|\vec{J}\|_1 = \sum_i |J_i|$ used in the celebrated Tibshirani paper [3], while $p = 0$ corresponds to the $L_0$ norm $\|\vec{J}\|_0 = \sum_i^N (1 - \delta_{J_i})$, where $\delta_J$ is the Kronecker delta. A particular feature of the $L_p$ norm is that, for $p \leq 1$, it sets a finite fraction of the model parameters exactly to zero, whereas it is convex for $p \geq 1$. The only parameter common to these two ranges is $p = 1$, explaining the popularity of the $L_1$ norm for convex optimization approaches.

In the following we apply the replica trick to compute the volume of the space of solutions [23,27], as well as other relevant quantities (order parameters) for the generalization problem.

2.1. Replica calculation

Let us consider the space of optimal solutions for the supervised learning of a realizable rule. The standard situation would be that where a training set $\{(y^\mu, \vec{x}^\mu)\}$ of $M$ experiments is presented to be classified by a linear perceptron with $N$ continuous weights $J_i$. The number of patterns relative to the amount of variables, $\alpha = M/N$, will play an
essential role as a control parameter. We define the Gibbs measure for the student vector \( \vec{J} \) as

\[
P_{\text{Gibbs}}(\vec{J}) = \frac{1}{Z(\beta, h)} e^{-\beta E(\vec{J}) - h\|\vec{J}\|_p}
\]

It depends on the inverse temperature \( \beta \), and the dilution field \( h \). In the \( \beta \to \infty \) limit, the partition function

\[
Z(\beta, h) = \int \prod_{i=1}^N dJ_i \exp(-\beta E(\vec{J}) - h\|\vec{J}\|_p)
\]

contains only terms of minimal training energy. So, by computing \( Z \) we can obtain the properties of the desired space. Although not explicitly indicated, the integration should be over the sphere \( \vec{J} \cdot \vec{J} = N \) to remove the scale invariance in the energy term in equation (3).

In the partition function above, the degrees of freedom are the \( N \) couplings \( J_i \), while the \( \vec{x}^\mu \) and \( y^\mu \) present in the Hamiltonian give the so-called quenched disorder. As we are concerned with the properties of the solutions in the typical case, we will have to average over these quenched variables. In particular, the \( \vec{x}^\mu \) will be i.i.d. random variables in \( \{\pm 1\}^N \), while the labels \( y^\mu \) are generated from the hidden structure of the couplings via equation (2). The teacher weights \( \vec{J}^0 \) too are i.i.d. random variables distributed as

\[
\rho(J^0) = (1 - n_{\text{eff}}^0)\delta_{J^0} + n_{\text{eff}}^0\rho'(J^0).
\]

The first term introduces the sparsity of the teacher, and the second term contains all non-zero couplings. Later we will use the letter \( t \) to refer to the variance of this distribution. The effective fraction of couplings, \( n_{\text{eff}}^0 = N_{J \neq 0}/N \), is the relative amount of non-zero couplings, and sparse models correspond to small \( n_{\text{eff}}^0 \ll 1 \). The fact that \( \vec{J}^0 \) is involved directly in the computation will allow us to compare the student vector \( \vec{J} \) to it.

The free energy \( f = -(1/\beta) \log Z \) is the relevant thermodynamic quantity, and the one that should be averaged over the quenched disorder. However, direct integration over \( \vec{x}^\mu \) and \( \vec{J}^0 \) in \( \log Z \) is out of reach. To work around this obstacle, we use the replica trick [28], which consists of using the known property

\[
\log Z = \lim_{n \to 0} \frac{Z^n - 1}{n}
\]

to average over \( Z^n \), instead of \( \log Z \), and sending \( n \) to zero afterward. Note that \( Z^n \) is the partition function of an \( n \)-times-replicated system if \( n \) is integer, which is the origin of the method’s name. In our case the averaged and replicated partition function would be

\[
Z^n = 2^{-MN} \sum_{x^\mu_{\kappa} = \pm 1} \int \prod_{\mu=1}^M D\gamma^\mu \int \prod_{i=1}^N dJ_i^0 \prod_{a=1}^N \rho(J^0_i) \int \prod_{a=1}^n \prod_{i=1}^N dJ_i^a \times \exp \left\{ -\beta \sum_{a=1}^n \sum_{\mu=1}^M \Theta \left( -\sum_{i=1}^N J_i^a x_i^\mu + \eta^\mu \right) \left( \sum_{i=1}^N J_i^0 x_i^\mu \right) - h \sum_{a=1}^n \|\vec{J}^a\|_p \right\}
\]
where $D_{\gamma}q_{\mu}$ stands for the Gaussian distributions of the noise variable $\eta_{\mu}$, with variance $\gamma^2$,

$$D_{\gamma}q_{\mu} = \frac{e^{-(\eta_{\mu}^2/2\gamma^2)}}{\sqrt{2\pi}\gamma}d\eta_{\mu}.$$  

This notation will be used throughout the paper, and if the subindex $\gamma$ is omitted, it refers to $\gamma = 1$.

After some standard steps detailed in appendix A, the replica symmetric estimate of the free energy is obtained as

$$-\beta\mathcal{F} = \text{extr}_{q,\hat{q},r,\hat{r},\lambda}\{ -r\hat{r} + \frac{1}{2}q\hat{q} - \lambda + G_J + \alpha G_X \}.$$  

The order parameters $q, \hat{q}, r, \hat{r}$ and $\lambda$ were introduced via Dirac delta functions in the replica calculation. In particular $q = N^{-1}\langle \hat{J}^0 \cdot \hat{J}^0 \rangle$ is the overlap between two (independent) student solutions. The notation $\langle \cdot \rangle$ stands for the expectation value w.r.t. the Gibbs measure. Note that $0 \leq q \leq 1$; it will be 1 when the Gibbs measure is condensed into a single $\hat{J}$, and it will be smaller than 1 when the measure is more spread. The parameter $r = N^{-1}\langle \hat{J} \cdot \hat{J}^0 \rangle$ is the overlap between the student vectors and the teacher one, and will be crucial in our understanding of the performance of generalization. The parameters $\hat{q}, \hat{r}, \lambda$ and $\mu$ are the corresponding associated Fourier variables (to represent the deltas introduced in the replica calculation). The last one, $\lambda$, corresponds to the spherical constraint $\hat{J} \cdot \hat{J} = N$.

The terms $G_J$ and $G_X$ are given by

$$G_J = \int dx \int dJ^0 \rho(J^0) \log \int dJ \exp \left( - \left( \frac{\hat{q}}{2} - \lambda \right) J^2 - h\|J\|_p + (\hat{r}J^0 - \sqrt{q}x)J \right),$$

$$G_X = 2 \int dx H \left( \frac{x^r}{\sqrt{q^2 + qx - r^2}} \right) \log \left( (e^{-\beta} - 1)H \left( -\sqrt{\frac{q}{1-q}}x \right) + 1 \right)$$

with $H(x) = \int_x^\infty (dy/\sqrt{2\pi})e^{-y^2/2}$. From the replica calculation, the term $G_J$ can be interpreted as the effective free energy of a single $J$. The inner term

$$Z_J(J^0, x) = \int dJ \exp \left( - \left( \frac{\hat{q}}{2} - \lambda \right) J^2 - h\|J\|_p + (\hat{r}J^0 - \sqrt{q}x)J \right)$$

plays the role of a single-$J$ partition function, while the term $\log Z_J$ corresponds to its free energy. The dependence of $Z_J(J^0, x)$ on $J^0$ and $x$ is conditioning the free energy of the single $J$ to the different values $J^0$ of the corresponding element in the teacher vector, and to the effective ‘noise’ from the realization of the training patterns $\bar{x}^p$. So the integration over $J^0$ and $x$ gives the average effective free energy of a single $J$.

This interpretation of $G_J$ allows also for formulating the following joint probability distribution of $x, J^0$ and $J$:

$$P(x, J^0, J) = \frac{e^{-(x^2/2)}}{\sqrt{2\pi}} \rho(J^0) \frac{\exp((-\hat{q}/2 - \lambda)J^2 - h\|J\|_p + (\hat{r}J^0 - \sqrt{q}x)J)}{\int dJ \exp((-\hat{q}/2 - \lambda)J^2 - h\|J\|_p + (\hat{r}J^0 - \sqrt{q}x)J)}$$

such that any expectation value of a generic function $g(x, J^0, J)$ can be found as

$$\mathbb{E}[g(x, J^0, J)] = \int dx \int dJ^0 \int dJ g(x, J^0, J) P(x, J^0, J).$$

doi:10.1088/1742-5468/2009/10/P10009
The limit $\beta \to \infty$ is trivial in equation (7). It concentrates the Gibbs measure onto the subspace of students with minimum training energy (error), and in the case of a feasible rule to the perfect solutions $E(\vec{J}) = 0$. The actual values of the variational parameters are determined by the saddle point condition for the free energy $\nabla_{q,r,...}f = 0$. With all the previous definitions, at zero temperature ($\beta \to \infty$) this condition is given by

$$\hat{q} = \frac{r\hat{r}}{q} + \frac{\alpha \sqrt{2}}{\sqrt{\pi}} \frac{1}{(1 - q)q} \int Dx \ H \left( \frac{x\sqrt{1 - q}}{\sqrt{q\gamma^2 + qt - r^2}} \right) \frac{x}{H(\sqrt{qx})}$$

$$\hat{r} = \frac{-2\alpha}{\sqrt{2\pi}} \frac{1}{\sqrt{q\gamma^2 + qt - r^2}} \int Dx \ x \log H \left( \frac{\sqrt{q}}{1 - q} \sqrt{1 - \frac{r^2}{q\gamma^2 + qt}} \right)$$

$$q = 1 + \frac{1}{\sqrt{q}} \mathbb{E}[x,J]$$

$$r = \mathbb{E}[J^0,J]$$

$$1 = \mathbb{E}[J^2].$$

This set of equations has to be solved numerically for each $\alpha = M/N$ and each dilution field $h$. The resulting values of the variational parameters $q, r, \hat{q}, \hat{r}$ and $\lambda$ are used to describe the solution space. For instance the generalization error, i.e. the probability that a new pattern (independently generated from those used for training) is misclassified by the student, depends only on the overlap between teacher and student $r$ (see [26])

$$\epsilon = \frac{1}{\pi} \arccos \left( \frac{r}{\sqrt{t}} \right)$$

The square root of the variance of the teacher $t = \int \rho(J^0, J^0) dJ^0$ is required because the teacher is not necessarily normalized to unity.

However, as mentioned before, the problem of inferring the topology of the data generator, i.e. of extracting its non-zero couplings, is also of crucial interest. In this context, the solution of the fixed point equations can be used to construct the precision versus recall curve, which is a standard check in network inference. In the case of graphical model selection we can use the information given by the student solution $\vec{J}$ to classify the couplings as relevant, $J_i > J_{ih}$, or not relevant, $J_i < J_{ih}$, where $J_{ih}$ is a sensitivity parameter. This means that we will disregard all inferred $J_i$ which are not strong enough. With the joint probability distribution (8) we can compute the probability of having any of the following situations:

- true positive, TP $J_i \neq 0$ $J_i^0 \neq 0$
- false positive, FP $J_i \neq 0$ $J_i^0 = 0$
- true negative, TN $J_i = 0$ $J_i^0 = 0$
- false negative, FN $J_i = 0$ $J_i^0 \neq 0$.

For instance the probability of having a true positive (TP) is $P_{TP} = \mathbb{E}[\Theta(|J| - J_{ih})(1 - \delta_{j^0})]$.

The recall (sensitivity) and the precision (specificity) are defined as follows:

$$\text{RC} = \frac{P_{TP}}{P_{TP} + P_{FN}} = \frac{P_{TP}}{n_{\text{eff}}^0} \quad \text{PR} = \frac{P_{TP}}{P_{TP} + P_{FP}} = \frac{P_{TP}}{n_{\text{eff}}^{th}}$$

where $n_{\text{eff}}^0$ is the real sparsity of the teacher (see (4)) and $n_{\text{eff}}^{th} = \mathbb{E}[|J| > J_{ih}]$ is the dilution of the student when the threshold value for a relevant coupling is $J_{ih}$. Note that both the
recall and the precision depend on $J_{th}$, as well as on the variational parameters $q, r, \hat{q}, \hat{r}$ and $\lambda$ that solve the fixed point equations (9). The PR–RC curve is the parametric curve $\text{RC}(J_{th})$ versus $\text{PR}(J_{th})$: the closer we can get to $\text{RC} = 1$ and $\text{PR} = 1$, the better the student perceptron has understood the topological structure of the teacher.

3. Non-diluted generalization

To avoid confusion we will call \textit{sparse} the case of teachers with many trivial couplings $J_{th}^0 = 0$, while the term \textit{diluted} will be saved for the generalization method (non-diluted/diluted). The replica calculation hitherto developed is general in a set of aspects. First, the teacher distribution (4) can be of any kind, including a non-sparse teacher $n_{th}^0 = 1$, although we will focus on the case of sparse models. Second, the possibility of a non-diluted generalization can be accounted by setting the dilution field $h = 0$, and for $h \neq 0$ different choices of regularization are possible. In this paper we will show the results for $L_0$ and $L_1$. For each of these cases (non-diluted, $L_0$ and $L_1$) the replica calculation has its particularities, which we present hereafter.

The simplest case is the non-diluted generalization ($h = 0$), as some of the equations simplify considerably, being equivalent to those in [26]. The absence of the dilution term in (7) makes the expression integrable, such that

$$G_J = \frac{t^2 + \hat{q}}{2(\hat{q} - 2\lambda)} - \frac{1}{2} \log(\frac{\hat{q}}{2} - \lambda).$$

The first two fixed point equations in (9) do not change, while the last three can be reduced to two algebraic equations without $\lambda$,

$$q = (\hat{r}^2t + \hat{q})(1 - q)^2$$
$$r = \hat{r}t(1 - q).$$

(12)

The value of $\lambda$ can be recovered using $(1 - q)(\hat{q} - 2\lambda) = 1$. The fixed point equations can be solved numerically for evaluating the generalization error (10) as well as the PR–RC curve. The calculation of the expectation values using (8) is also simplified since

$$P(J^0, J) = \frac{\frac{\hat{q} - 2\lambda}{\sqrt{2\pi(2\hat{q} - 2\lambda)}} \exp\left[-\frac{(\hat{q} - 2\lambda)J - \hat{r}J^0)^2}{2(2\hat{q} - 2\lambda)}\right]}$$

(13)

and the terms involved in recall and precision (11) are easier to obtain.

Let us take as a toy example the case of a sparse teacher with only $n_{th}^0 = 5\%$ non-zero couplings. We set the noise to $\gamma = 0$, such that there is always a perfect student solution. In particular, we will use a discrete teacher $J^0 \in \{-1, 0, 1\}$,

$$\rho(J^0) = (1 - 0.05)\delta_{J^0} + 0.05 \frac{1}{2} (\delta_{j^0, -1} + \delta_{j^0, 1}).$$

(14)

With such a simple structure it happens to be the case that the teacher’s dilution and variance are equal, $n_{th}^0 = t = 0.05$.

The solution of the fixed point equations ((9) and (12)) is found numerically for different values of the amount of training data $\alpha$. For each $\alpha$, the different PR–RC curves are shown in figure 1. It is clear from the figure that for sufficiently large $\alpha$, for instance $\alpha \geq 2.0$ in this example, the generalization is capable of a good classification of the

doi:10.1088/1742-5468/2009/10/P10009

9
couplings $J$, achieving both high precision and high recall, i.e. a good performance in model selection. This is seen in the figure as a curve that approaches the PR ≡ 1 line.

It is no surprise that more training data result in better model selection. However, we can gain some information about how the solution approaches perfect graphical model selection by looking at the statistical distributions of the student’s $J$s. Figure 2 shows how $P(J) = \int dJ^0 P(J^0, J)$ (equation (13)) concentrates around the discrete (and rescaled) values of $J^0$ with a set of Gaussians that have negligible overlap for large values of $\alpha$. Above a critical $\alpha \approx 1.6$ (in this example) we can start to discriminate the $J$s from different Gaussians because local minima in $P(J)$ emerge. It is expected that above this point a reasonable value for $J_{th}$ is the one satisfying

$$\frac{\partial P(J)}{\partial J} = 0, \quad \frac{\partial^2 P(J)}{\partial J^2} > 0.$$ 

This choice for $J_{th}$ leads to the recall and precision, which can be seen in figure 1 marked by the square symbols. For $\alpha = 2.5$ the Gaussians in $P(J)$ are almost perfectly distinguishable. The optimal choice for $J_{th}$ has a precision and a recall near 1 (RC ≈ 0.991, PR ≈ 0.996), meaning that generalization is achieving an almost perfect graphical model selection.

4. Diluted generalization

At zero temperature (infinite $\beta$) and $h = 0$, the Gibbs measure gives the same probability to all perfect student solutions, since they have the same energy, $E(\vec{J}) = 0$, while suppressing completely positive-cost students. Working at zero temperature, a dilution field $h > 0$ gives the chance of imposing a different measure over the set of perfect solutions. This measure favors the students with the lowest values of $\|\vec{J}\|_p$, and in the
Figure 2. The statistical distribution of the student couplings $J$ for three different values of $\alpha$. When enough training data are given ($\alpha = 2.5$ in this figure), the distribution $P(J)$ can be recognized as the superposition of Gaussian distributions located around the discrete (and rescaled) values of the teacher’s couplings. In such a case, setting to 0 all of those student couplings $J_i$ that are around 0 results in a nearly perfect model reconstruction. As less information is used for training, the Gaussians overlap, and any threshold for the relevance of a coupling $J_{th}$ will misclassify some couplings, resulting in a worse graphical model selection.

limit of $h \to \infty$, it concentrates to the perfect solution with the highest $L_p$ dilution. We will now study the properties of the subset of perfect students with the smallest $L_p$ norm.

Unlike the trivial $\beta \to \infty$ limit (see equation (7)), the large dilution limit has to be taken carefully as some parameters diverge. For large dilution field $h \to \infty$ we have $q \to 1$, meaning that different students are very close to each other; and in the limit $h = \infty$ there is only one student which is at the same time zero cost ($E(\vec{J}) = 0$) and maximally diluted. As can be seen from the fixed point equations (9), when $q$ tends to 1, the order parameters $\hat{q}$, $\hat{r}$ and $\lambda$ diverge. The scaling behavior of these variables is the following:

$$\begin{align*}
(1 - q) & \simeq \frac{Q}{h} \quad \hat{r} \simeq \hat{R} h \\
\hat{q} & \simeq \hat{Q} h^2 \quad \frac{\hat{q}}{2} - \lambda \simeq \frac{K}{2} h.
\end{align*}$$

In terms of these new variables, the fixed point equations for $\hat{q}$ and $\hat{r}$ (9) become

$$\begin{align*}
\dot{\hat{q}} &= \frac{\alpha}{\pi Q^2} \left[ \arccot \frac{r}{\sqrt{\gamma^2 + t - r^2}} - \frac{r \sqrt{\gamma^2 + t - r^2}}{\gamma^2 + t} \right] \\
\dot{\hat{R}} &= \frac{\alpha \sqrt{\gamma^2 + t - r^2}}{Q \pi (\gamma^2 + t)}
\end{align*}$$

(16)

and do not depend on the dilution $p$. Furthermore, this scaling makes the exponent of the exponential term inside $P[x, J^0, J]$ (equation (8)) proportional to $h$. So, in the remaining
three equations

\[
Q = \frac{1}{\sqrt{\hat{Q}}} \mathbb{E}[xJ] \\
 r = \mathbb{E}[J^0 J] \\
1 = \mathbb{E}[J^2]
\]  

(17)

the expectation values for \( h \to \infty \) are dominated by the largest values of the exponent in (8). The specific details of this saddle point calculation depend on the actual dilution \( \|\tilde{J}\|_p \).

Among all possible values of \( p \), \( p = 1 \) and \( p = 0 \) give cases that are special for both their meaning and their simplicity in the calculations. The \( L_1 \) norm is extremely popular in machine learning because it maintains the convexity of a convex cost function while forcing sparse solutions [3]. On the other hand, \( L_0 \) lacks the convexity preserving property completely (it is not even continuous), and therefore is not a suitable penalty for convex optimization. However, the \( L_0 \) norm is optimal in the sense that it does not deform the Hamiltonian beyond penalizing non-zero couplings. We will compare the dilution achieved by the \( L_1 \) approach with the largest possible dilution (the one obtained using \( L_0 \)), and give a qualitative description of this widely used regularization. Another simple and common choice for the penalty term is \( p = 2 \), but in our model setting it is meaningless since the student is constrained to the sphere and therefore has a fixed \( \|\tilde{J}\|_2 = N \).

4.1. \( L_1 \) dilution

We first discuss the case of \( L_1 \) regularization, \( \|\tilde{J}\|_1 = \sum_{i=1}^{N} |J_i| \). For \( h \to \infty \) the expectation value of an arbitrary function \( g(x, J^0, J) \) is given by

\[
\mathbb{E}[g(x, J^0, J)] = \int \mathcal{D}x \int dJ^0 \rho(J^0) \\
\times \begin{cases} 
  g(x, J^0, 0) & |\hat{R}J^0 - \sqrt{\hat{Q}}x| < 1 \\
  g \left( x, J^0, \frac{\hat{R}J^0 - \sqrt{\hat{Q}}x - \text{sgn}(\hat{R}J^0 - \sqrt{\hat{Q}}x)}{K} \right) & \text{otherwise.}
\end{cases}
\]

The derivation of this expectation value is shown in appendix B. As already mentioned, one of the virtues of the \( L_1 \) regularization is that it forces the solution to be diluted by setting a fraction of the couplings exactly to zero. This fact becomes evident in the previous equation.

Using the scaling behavior (15), writing \( S(J^0, x) = \text{sgn}(\hat{R}J^0 - \sqrt{\hat{Q}}x) \), and defining the functional

\[
L[\cdot] = \int \mathcal{D}x \ [\cdot] \Theta(|\hat{R}J^0 - \sqrt{\hat{Q}}x| - 1),
\]
the resulting fixed point equations in the $h \to \infty$ limit are (16) and

$$Q = \frac{1}{K} \int dJ^0 \rho(J^0)L[1]$$

$$r = \frac{1}{K} \int dJ^0 \rho(J^0) \left[ \hat{R}J^0 L[1] - \sqrt{\hat{Q}} J^0 L[x] - J^0 L[S(J^0, x)] \right]$$

$$K = (\hat{Q} + 1)Q + r\hat{R} + \frac{1}{K} \int dJ^0 \rho(J^0) \left[ \sqrt{\hat{Q}} L[xS(J^0, x)] - \hat{R}J^0 L[S(J^0, x)] \right].$$

(18)

Note that the original parameter $q$ is no longer present, since it is 1, but the overlap between teacher and student, $r$, is still a non-trivial order parameter.

4.2. $L_0$ dilution

For the $L_0$ regularization the dilution term in the Hamiltonian (3) is $\|\vec{J}\|_0 = \sum_i^N (1 - \delta J_i)$, punishing only the fact that a given $J_i$ is non-zero, but otherwise making no distinction between different non-zero $J$ values. In a strict mathematical sense, introducing the $L_0$ norm in the Hamiltonian is meaningless, since a finite single-point discontinuity cannot alter the integration over the continuous range of $J$ values in the partition function. So the $L_0$ norm can only be understood as the limiting case $p \to +0$ of a family of continuous functions (see appendix C).

Using an approach similar to the one presented in appendix B for $L_1$, the expectation value of an arbitrary function $g(x, J^0, J)$ in the $h \to \infty$ limit reads

$$E[g(x, J^0, J)] = \int Dx \int dJ^0 \rho(J^0)$$

$$\times \begin{cases} 
  g(x, J^0, 0) & \frac{|\hat{R}J^0 - \sqrt{\hat{Q}}x|}{\sqrt{2K}} < 1 \\
  g\left(x, J^0, \frac{\hat{R}J^0 - \sqrt{\hat{Q}}x}{K}\right) & \text{otherwise.}
\end{cases}$$

(19)

The fixed point equations for $\hat{Q}$ and $\hat{R}$ are exactly the same as for $L_1$ dilution (equation (16)), while the other three order parameters are now given by

$$Q = \frac{1}{K} \int dJ^0 \rho(J^0) \left[ M[1] - \frac{1}{\sqrt{\hat{Q}}} \hat{M}[xS(J^0, x)] \right]$$

$$r = \frac{1}{K} \int dJ^0 \rho(J^0) \left[ \hat{R}J^0 M[1] - \sqrt{\hat{Q}} J^0 M[x] \right]$$

$$K = r\hat{R} + Q\hat{Q}$$

(20)

where

$$M[\cdot] = \int Dx [\cdot] \theta \left( \frac{|\hat{R}J^0 - \sqrt{\hat{Q}}x|}{\sqrt{2K}} - 1 \right)$$

doi:10.1088/1742-5468/2009/10/P10009
and $\tilde{M}[xS(J^0, x)]$ is defined as
\[
\tilde{M}[xS(J^0, x)] = -\frac{2\sqrt{2K}}{2\pi} \exp - \frac{\hat{R}^2 J^0^2 + 2K}{2Q} \cosh \frac{\hat{R} J^0 \sqrt{2K}}{Q}.
\]

4.3. Dilution, recall and precision

The numerical solution of the fixed point equations for the $L_1$ and $L_0$ dilutions gives us the overlap $r$ between the teacher and the student. The generalization error is obtained using equation (10).

The most striking effect of the norms is the emergence of an extensive number of couplings that are exactly zero (see $P(J)$ in the appendices). The fraction of non-zero couplings is the effective dilution $n_{\text{eff}}$ achieved by the student, and it is obtained as

\[
n_{\text{eff}} = E[|J| > 0] = \begin{cases}
\int dJ^0 \rho(J^0) L[1] = QK & \text{L}_1 \text{ norm} \\
\int dJ^0 \rho(J^0) M[1] & \text{L}_0 \text{ norm}.
\end{cases}
\]

It is expected (and numerically observed) that for large values of $\alpha$ the effective dilution $n_{\text{eff}}$ converges to the real dilution of the teacher $n^0_{\text{eff}}$.

Along the same lines as the development for the non-diluted case, we can further restrict the set of non-zero couplings by setting a threshold for relevant couplings. In other words, we interpret as non-relevant all those couplings that are not strong enough, $J_i < |J_{\text{th}}|$. In this case, the fraction of relevant couplings equals

\[
n_{\text{eff}}^{\text{th}} = E[\Theta(|J| - J_{\text{th}})]
\]

and can be used to calculate the precision according to equation (11). The other terms appearing in the recall and precision are also computed using the expectation value $E[\cdot]$ for each dilution scheme. For instance, the probability of having a false positive is given by

\[
P_{FP} = E[\Theta(|J| - J_{\text{th}}) \delta J^0].
\]

5. How well does dilution work?

The mathematical machinery discussed can shed some light on this question. To see the differences between diluted and non-diluted generalizations, and the performance in sparse graphical model selection, let us use the same toy example as was used for the non-diluted case, with a teacher of dilution $n^0_{\text{eff}} = 5\%$ and discrete values $J^0_i \in \{-1, 0, 1\}$; see equation (14).

The functions $n_{\text{eff}}(\alpha)$ for the $L_1$ and $L_0$ dilutions are presented in figure 3. It can be seen that $L_0$ diluted generalization goes monotonically from below to the correct value $n_{\text{eff}} = 0.05$, in a somehow Occam-optimal way. In other words, $L_0$ dilution adds non-zero couplings just when strictly required by the empirical (training) evidence. The $L_1$ norm is not that effective. It is an interesting result that, for a certain range in $\alpha$, the $L_1$ optimal solution requires more non-zero couplings ($n_{\text{eff}} > n^0_{\text{eff}} = 0.05$) than are actually present in the teacher. This overshooting is the cost that we pay for deforming the Hamiltonian through the $L_1$ penalty of large couplings. Unlike for $L_0$ regularization, $L_1$ approaches the correct dilution from above, not from below.
Figure 3. The dilution $n_{\text{eff}}(\alpha)$ achieved by the $L_1$ and $L_0$ dilutions, as a function of the amount of training patterns $\alpha$. The $L_0$ regularization approaches the dilution of the teacher $n_{\text{eff}}^0 = 0.05$ from below, including non-zero couplings only when strictly required to correctly classify the training data. The $L_1$ dilution is not that efficient, and for $\alpha > 0.45$ it uses more non-zero couplings than are actually needed.

One could be tempted to call the change of slope of $L_0$ near $\alpha = 0.8$ in figure 3 a transition to a perfect student solution, but it is not. The generalization error in figure 4 shows that errors persist also for larger $\alpha$. On the other hand, while the $L_1$ norm goes smoothly to $\epsilon = 0$, the $L_0$ undergoes an abrupt reduction of the generalization error near $\alpha = 0.8$. This might be a sign of a transition to an (almost) perfect graphical model selection, such that for $\alpha > 0.8$ the student has identified the correct $J_i = 0$, and its mistakes are restricted to the actual values of those $J_i$s that are non-zero. Such a discontinuity was already noted in [17].

To compare the $L_1$ and $L_0$ dilutions to non-diluted generalization, we show the PR–RC curves in figure 5, for four typical values of $\alpha$. The curves for the diluted generalization seem to be missing the right part—but they are not. As the $L_1$ and $L_0$ methods set a fraction of the couplings exactly to zero, lowering the threshold $J_{\text{th}}$ will never achieve including them as non-zero couplings, and this is why we cannot arrive at recall equal to 1.

Looking at the PR–RC curves, the first obvious fact is that the non-diluted generalization is much worse than any of the diluted ones. The next interesting fact is that $L_1$ performs slightly better than $L_0$ for low values of $\alpha$, something that could be seen also from the generalization error (figure 4). Finally the sudden change to PR $\approx 1$, for $\alpha = 0.8$, of the PR–RC curve for the $L_0$ dilution is showing that $L_0$ quickly moves to almost perfect graphical model selection, as we guessed from the generalization error curve. However, there is no critical $\alpha$, and the sudden change is not a phase transition. This can be seen more clearly working with less diluted teachers (for instance $n_{\text{eff}}^0 = 0.1$; data not shown). To gain some more understanding of the onset of an almost perfect graphical model selection it is interesting to see the distributions of couplings, $P(J)$, which are shown in appendices B and C.
Figure 4. The generalization errors at different values of $\alpha$. The two curves at the center are the generalization errors achieved using $L_1$ and $L_0$ dilutions. While the $L_1$ error decreases smoothly with the training data, the one corresponding to $L_0$ undergoes a sudden drop near $\alpha = 0.8$. This is a consequence of a sudden move to almost perfect graphical model selection, where the set of non-zero interactions has been identified with very good precision. Both the upper and lower curves correspond to the non-diluted generalization, where the lower one has been plotted with a rescaled $x$-axis, $\epsilon(\alpha/n^{\text{eff}})$. If the student could know from the beginning which are the non-zero couplings, it could use all the training data to tune the values of these couplings, resulting in a huge reduction of the generalization error.

6. The memorization limit

In the calculations presented so far, the noise $\eta^\mu$ affecting the output $y^\mu$ in equation (2) was neglected by setting its variance to $\gamma^2 = 0$. By doing this, we guaranteed that for any $\alpha$, there is always at least one zero-cost solution for the student, namely $\vec{J} = \vec{J}^{0}$. Let us now study the opposite extreme case where the noise is extremely large. In that case, the output function is given by

$$y^\mu = \sigma^0(\vec{x}) = \text{sgn}(\eta^\mu),$$

i.e. the patterns are randomly classified by $y^\mu = \pm 1$. The teacher’s couplings $J^{0}_i$ become irrelevant, and the student will try to learn (generalize) a non-existing hidden relation. This limit is equivalent to the well-studied memorization problem of a random input–output relation. It is a classical result [24] that for $\alpha > 2$ the student will fail to correctly classify all patterns, while for $\alpha < 2$ the student can find a solution of zero energy. In the latter case, the student vector $\vec{J}$ reproduces correctly the relation between the $M$ patterns $\vec{x}_\mu$ and the corresponding labels $y^\mu$ through $y^\mu = \text{sgn}(\vec{J} \cdot \vec{x}_\mu)$. The student was capable of memorizing the labeling of the input patterns.

Therefore memorization can be studied as the noise-dominated limit of generalization. By computing the limit $\gamma \rightarrow \infty$ in the fixed point equations for $\hat{q}$ and $\hat{r}$ (9), we found...
that $\hat{r} = 0$ while
\begin{equation}
\hat{q} = \frac{\alpha}{\sqrt{2\pi q(1-q)}} \int \mathcal{D}x \frac{x}{H(\sqrt{q}x)}.
\end{equation}
The expectation value of a function $g(x, J)$ becomes
\begin{equation}
\mathbf{E}[g(x, J)] = \int \mathcal{D}x \frac{\int dJ g(x, J)e^{-((\hat{q}/2-\lambda)J^2-h\|J\|_p-\sqrt{\hat{q}}x)J}}{\int dJe^{-(\hat{q}/2-\lambda)J^2-h\|J\|_p-\sqrt{\hat{q}}x)J}}.
\end{equation}
In particular we have $r = \mathbf{E}[J^0 J] = 0$, meaning that the overlap between student and teacher is zero, which is an obvious consequence of the large noise limit. So, the set of variational parameters describing our problem reduces to $q, \hat{q}$ and $\lambda$, and the fixed point equations are (22) and
\begin{equation}
q = 1 + \frac{1}{\sqrt{\hat{q}}} \mathbf{E}[xJ] \quad 1 = \mathbf{E}[J^2].
\end{equation}
The generalization error and the precision versus recall curve are meaningless in this context. However, we can still check the efficiency of the $L_1$ and $L_0$ memorizations in...
using as few non-zero couplings as possible to memorize a set of patterns. There is a first trivial conclusion, coming from the already stated fact that a continuous perceptron is capable of memorizing without error until $\alpha = 2$. This means that a perceptron with $N$ couplings $J_i$ can remember the classification of $M = 2N$ patterns. It follows directly from this that if $\alpha < 2$ patterns are given, we can set to zero any fraction $1 - \alpha/2$ of the couplings, and still be capable of memorizing without error with the remaining $\alpha/2$ couplings. We are interested in how much more dilution can be obtained by the introduction of a dilution term in the Hamiltonian. Note that, instead of setting to zero a random group of $(1 - \alpha/2)N$ couplings, we optimize their selection, we can go far below the trivial $n_{\text{eff}} = \alpha/2$ dilution.

We can solve the fixed point equations in the limit of large dilution fields $h \to \infty$. Once again the solution space reduces to only one solution $q \to 1$, so there are some divergences in the equations. The scaling behavior of the variational parameters is the following:

$$
(1 - q) \simeq \frac{Q}{h} \\
\hat{q} \simeq \hat{Q}h^2 \\
\frac{\hat{q}}{2} - \lambda \simeq \frac{K}{2}h.
$$

Using this scaling, the expectation values are given by

$$
E[g(x, J)] = \int Dx \begin{cases} 
  g(x, 0) & x^2 < \frac{1}{Q} \\
  g \left( x, \frac{\sqrt{Q}x - \text{sgn}(x)}{K} \right) & \text{otherwise}
\end{cases}
$$

for the $L_1$ case, and

$$
E[g(x, J)] = \int Dx \begin{cases} 
  g(x, 0) & x^2 < \frac{2K}{Q} \\
  g \left( x, \frac{\sqrt{Q}x}{K} \right) & \text{otherwise}
\end{cases}
$$

for the $L_0$ case.

Solving numerically the corresponding fixed point equations, we can compute the dilution achieved by each method:

$$
n_{\text{eff}} = 1 - D[\delta_{J,0}] = \begin{cases} 
  2H \left( \hat{Q}^{-1/2} \right) & L_1 \text{ norm} \\
  2H \left( \sqrt{2K}\hat{Q}^{-1/2} \right) & L_0 \text{ norm}
\end{cases}
$$

The resulting functions $n_{\text{eff}}(\alpha)$ are shown in figure 6. It shows that both $L_1$ and $L_0$ achieve a much stronger dilution than the trivial random one. As in the generalization case, the $L_1$ regularization works less well than $L_0$, the reason being that it penalizes large $J$ values. The dilution achieved in memorization for the $L_1$ and $L_0$ dilutions is always above the corresponding generalization curves in figure 3. Although this is not shown in
Figure 6. The maximum dilution achieved when using the $L_1$ and $L_0$ regularizations in the memorization problem. The trivial dilution $n_{\text{eff}} = \alpha/2$ is outperformed by both $L_0$ and $L_1$ dilutions, and $L_0$ is the most efficient of all. For $\alpha > 2$ the student fails to memorize all patterns and there is no student solution of zero energy.

either figure, we checked that near $\alpha = 0$ the corresponding curves coincide, as there is no difference between learning and memorizing when too few training data are given.

7. Conclusions

In this paper, we have presented an analytical replica computation on the generalization properties of a sparse continuous perceptron. Dilution has been achieved in different ways: first, it can be imposed naively by using non-diluted inference, followed by deleting all those couplings which are below some threshold value. Second, it can be achieved by introducing a dilution field which is coupled to the $L_p$ norm of the coupling vector, penalizing thereby vectors of high norm. For $p \leq 1$, the cusp-like singularity of the $L_p$ norm at zero forces a finite fraction of all couplings to be exactly zero. We have studied in particular two special cases: (i) $p = 1$ is a popular choice in convex optimization since it is the only value of $p$ which corresponds to both a convex penalty function and dilution; (ii) $p = 0$ achieves optimal dilution since it penalizes equally all non-zero couplings independently of their actual value, but due to the non-convex character of this penalty, it easily leads to computational intractability.

As a first finding, we see that both $L_p$ schemes work fundamentally better than the naive scheme, both as regards the questions of graphical model selection (i.e. for the identification of topological properties of the data-generating perceptron given by its non-zero couplings) and as regards the generalization ability. For a very small or a very large amount of training data, $L_0$ and $L_1$ achieve very comparable results. We find, however, an intermediate regime where $L_0$ suddenly improves its performance toward almost perfectly graphical model selection, whereas $L_1$ dilution shows a more gradual increase in performance. This is very interesting since this regime is found for relatively small data
sets, and in many current inference tasks (e.g. in computational biology) the quantity of data is the major limiting factor for the computational extraction of information. It might be in this parameter region where statistical physics based algorithms like the ones presented in [19]–[22] may outperform methods based on convex optimization proposed in [3].

These analytic results call for efficient algorithms in real case studies. At odds with the linear regression case with $L_1$ norm, in the case of a continuous perceptron, a simple gradient descent strategy does not work due to the presence of a zero mode in the energetic term equation (1) ($E(\vec{J}) = E(c\vec{J})$ for every scalar $c > 0$). The zero mode has been removed in the computation by fixing the modulus of the classification vector ($\vec{J} \cdot \vec{J} = N$). Unfortunately this spherical constraint breaks the convexity of the problem and it is not clear to us whether there are more ingenious ways of removing the zero mode that could work, at least in the $L_1$ norm case.

Another possibility that we are planning to follow up is that of considering variational approximation schemes like belief propagation/message passing for continuous perceptrons [20]–[22], [29] which are able to overcome also the problem of the non-convexity of the $L_0$ regularization.

One point of our analysis deserves further scrutiny: all analytical results were derived under the assumption of replica symmetry. The non-convex character in particular of $L_0$ dilution might lead to a region in parameter space which has to be characterized by full replica symmetry breaking. Whereas we expect even in that case the replica symmetric solution to quite accurately describe the relative behavior of the different dilution schemes, a major problem might emerge for the use of belief propagation type algorithms which normally do not converge when replica symmetry is locally unstable. Detailed studies of the stability of the solution presented here are in progress.

During the preparation of this paper, a related study on the efficiency of $L_p$ dilution in systems of linear equations was posted online [30]. Also there, the relative importance of $L_0$ and $L_1$ dilutions was studied, with conclusions which are highly compatible with ours.

Acknowledgments

AL and MW acknowledge support by the EC-founded STREP GENNETEC (‘Genetics Networks: emergence and complexity’).

Appendix A. Replica calculation details

The calculation of $\bar{Z}^n$ in equation (6) is done by the introduction of an overlap matrix $Q_{a,b}$ using the constraints

$$\delta \left( NQ_{a,b} - \sum_i J_i^a J_i^b \right)$$

for any $0 \leq a \leq b \leq n$. As there is a symmetry in the replica indices $Q_{a,b} = Q_{b,a}$, only half of the matrix is needed. The value $a = 0$ refers to the teacher, while $a = 1, \ldots, n$ refers to the $n$-fold-replicated student. Among these constraints, there are some that are particular. For instance the term $Q_{0,0}$ is the variance of the teacher, and it should be
equal to the variance $t$ of the teacher distribution (4). Similarly, the $n$ terms $Q_{a,a}$ are set to 1, in order to impose the spherical constraint on the student, since the energy (1) is invariant under rescaling of the student vector.

Using Fourier representation of the Dirac deltas, the replicated partition function is

$$Z^n = \int \frac{dQ_{a,b} d\hat{Q}_{a,b} d^n \lambda^a}{(2\pi)^{3/2} N^{n-1}} \exp \left( iN \sum_{a<b} Q_{a,b} \hat{Q}_{a,b} + iN \sum_{a>0} \lambda_a + iN t \lambda_0 \right)$$

$$\times \left( \int d^n J^a \rho(J^0) \exp \left( -h \sum_a \|J^a\|_p - i \sum_{a\leq b} J^a \hat{Q}_{a,b} J^b \right) \right)^N$$

$$\times \left( \int D\gamma \frac{d^n X^a d^n \hat{X}^a}{2\pi} \exp \left( -\beta \sum_a \theta(-X^0 + \eta) X^a \right) + i \sum_a X^a \hat{X}^a - \frac{1}{2} \sum_{a,b} \hat{X}^a Q_{a,b} \hat{X}^b \right)^M \quad (A.1)$$

where $\hat{Q}_{a,b}$ are the conjugated parameters in the Fourier representation of the deltas. In particular, $\lambda^0$ and $\lambda^a$ are the ones corresponding to the teacher variance and the spherical constraint. To save some space, we used the shorthand notation $d^n A^a$ as a substitute for $\prod_{a=0}^n dA^a$, and $dQ_{a,b}$ for the differential of all the terms in the overlap matrix.

The next step in the replica calculation is to assume a structure for the overlap matrix. In the replica symmetric case, the overlap matrix and its Fourier counterpart have the structure (exemplified for $n = 3$)

$$Q_{a,b} = \begin{pmatrix} t & 1 \\ r & q & 1 \\ r & q & q & 1 \end{pmatrix} \quad -i\hat{Q}_{a,b} = \begin{pmatrix} \lambda^0 \\ \hat{r} & \lambda \\ \hat{r} & \hat{q} & \lambda \\ \hat{r} & \hat{q} & \hat{q} & \lambda \end{pmatrix}. \quad (A.2)$$

The Fourier mode corresponding to the variance of the teacher $t$ can be shown to be $\lambda^0 = 0$, while that of the spherical constraint remains a variational parameter $\lambda = -i\lambda^a$. The other parameters are $q$, the self-overlap of two student solutions, $r$, the overlap between a student and the teacher, and their conjugate Fourier modes $\hat{q}$ and $\hat{r}$.

A standard feature of the replica trick is the inversion of the order of the limits, taking $N \to \infty$ first, and then $n \to 0$, profiting thereby from the saddle point method to solve the integral in (A.1). Note that the last two lines in (A.1) can be brought to the exponential form by using $X = \exp \log X$. Thus the value of the free energy $-\beta \bar{f} = \lim_{n \to 0} \lim_{N \to \infty} (1/Nn) \log Z^n$ is given by extremizing the equation

$$-\beta \bar{f} = -r \hat{r} + \frac{1}{2} \beta \hat{q} \hat{q} - \lambda + G_J + \alpha \ G_X$$

with respect to the variational parameters $(q, r, \hat{q}, \hat{r}, \lambda)$, where we have introduced

$$G_J = \int D\gamma \int dJ^0 \rho(J^0) \log \int dJ \exp \left( -\left( \frac{\hat{q}}{2} - \lambda \right) J^2 - h\|J\|_p + (\hat{r} J^0 - \sqrt{\hat{q}} \gamma) J \right)$$

$$G_X = 2 \int D\gamma H \left( \frac{\sqrt{q} r}{\sqrt{q^2 + q t - r^2}} \right) \log \left( (e^{-\beta} - 1) H \left( -\sqrt{\frac{q}{1-q}} \gamma \right) + 1 \right)$$
Statistical mechanics of sparse generalization and graphical model selection

\[ H(x) = \int_x^\infty \frac{dy}{\sqrt{2\pi}} e^{-y^2/2}. \]

**Appendix B. The limit \( h \to \infty \)**

The scaling behaviors of the parameters \( q, r, \hat{q}, \hat{r} \) and \( \lambda \) in the limit \( h \to \infty \)

\[
(1 - q) \simeq \frac{Q}{h} \quad \hat{r} \simeq \hat{R} h \\
\hat{q} \simeq \hat{Q} h^2 \quad \frac{\hat{q}}{2} - \lambda \simeq \frac{K}{2} h
\]

were first obtained by looking at the solutions of the fixed point equations for growing values of \( h \), and their consistency was checked later in the fixed point equations.

Considering this scaling, the expectation value of a generic function \( g(x, J^0, J) \) is given by

\[
\mathbb{E}[g(x, J^0, J)] = \int D x \, d J^0 \rho(J^0) \frac{\int dJ \, g(x, J^0, J) e^{-h((K/2)J^2 + ||J||_p - (\hat{R}J^0 - \sqrt{\hat{Q}}J))}}{\int dJ e^{-h((K/2)J^2 + ||J||_p - (\hat{R}J^0 - \sqrt{\hat{Q}}J))}}. \tag{B.1}
\]

The diverging prefactor \( h \) in the exponentials forces the main contribution to the \( J \) integration to come from the largest value of the exponent (saddle point approximation):

\[
J_p^* = \text{arg min}_J \left( \frac{K}{2} J^2 + ||J||_p - (\hat{R}J^0 - \sqrt{\hat{Q}}J) \right). \tag{B.2}
\]

In the case of the \( L_1 \) norm (\( ||J||_1 = |J| \)) the solution of the previous equation is given by

\[
J = \begin{cases} 
0 & |\hat{R}J^0 - \sqrt{\hat{Q}}x| < 1 \\
\frac{\hat{R}J^0 - \sqrt{\hat{Q}}x - \text{sgn}(\hat{R}J^0 - \sqrt{\hat{Q}}x)}{K} & \text{otherwise}.
\end{cases} \tag{B.3}
\]

The expectation value is thus

\[
\mathbb{E}[g(x, J^0, J)] = \int D x \int d J^0 \rho(J^0) \\
\times \begin{cases} 
g(x, J^0, 0) & |\hat{R}J^0 - \sqrt{\hat{Q}}x| < 1 \\
g \left( x, J^0, \frac{\hat{R}J^0 - \sqrt{\hat{Q}}x - \text{sgn}(\hat{R}J^0 - \sqrt{\hat{Q}}x)}{K} \right) & \text{otherwise}.
\end{cases} \tag{B.4}
\]

The probability distribution of the student couplings \( P(J) \) can be obtained as

\[
P(J) = (1 - n_{\text{eff}}) \delta(J) + \frac{K}{\sqrt{Q} \sqrt{2\pi}} \int d J^0 \rho(J^0) \exp \left( -\frac{(\hat{R}J^0 - \text{sgn}J - KJ)^2}{2\hat{Q}} \right)
\]

doi:10.1088/1742-5468/2009/10/P10009
where \( n_{\text{eff}} = 1 - \int \mathcal{D}x \int \mathcal{D}J^0 \rho(J^0) \Theta[|\hat{R}J^0 - \sqrt{\hat{Q}}x| - 1] \). The continuous part of this distribution is shown in figure B.1 for the same four values of \( \alpha \) as the precision–recall curves were shown for in figure 5. We can see that for growing values of \( \alpha \), the distribution \( P(J) \) is more concentrated around the discrete values of \( J \), and the amount of couplings that are small but not zero reduces continuously. This explains the high performance in graphical model selection of the \( L_1 \) dilution for \( \alpha > 1.0 \).

Note that the calculation of the smallest value of the exponent in (B.1) is particularly simple for \( L_0 \) and \( L_1 \). Other values of \( p \) may require a numerical solution. It is simple to see that in the \( p > 1 \) case no dilution is obtained.

### Appendix C. \( L_0 \) as the \( p \to 0 \) limit

The \( L_0 \) dilution corresponds to a term \( \sum_i (1 - \delta_{J_i}) \) in the Hamiltonian (3). However, the Kronecker delta is zero for all non-zero arguments, with an isolated and finite discontinuity at the origin. This single-point discontinuity is irrelevant in the integration over continuous \( J \)s in the partition function as well as in (B.1). Therefore using the \( L_0 \) dilution from the beginning gives the same results as for the non-diluted case \( h = 0 \). Nevertheless, we can still interpret the \( L_0 \) norm as the \( p \to +0 \) limit of the \( L_p \) norm.

For general \( p > 0 \) there is no explicit solution for equation (B.2). We will argue that the limit \( p \to 0 \) of such solutions is exactly the solution of

\[
J^*_0 = \arg\min J \left( \frac{K}{2} J^2 + (1 - \delta_J) - (\hat{R}J^0 - \sqrt{\hat{Q}}x)J \right)
\]

just as if we had introduced the \( L_0 \) norm from the beginning, and taken naively the saddle point including the isolated singularity. There are two candidate values for \( J^*_0 \); one is 0 and the other one is the zero-derivative point of the quadratic function \( J^*_0 = (\hat{R}J^0 - \sqrt{\hat{Q}})/T \).
Figure C.1. The $p = 0$ and $p > 0$ cases of the function $Y_p(J) = (K/2)J^2 + \|J\|_p - (\hat{R}J^0 - \sqrt{Qx})J$ in the two characteristic situations where $J_p^* = 0$ and $J_p^* > 0$. The closer $p$ is to zero, the closer the function $Y_p(J)$ is to $J_0(J)$.

Figure C.2. The distribution $P(J)$ of the student couplings for the four values of $\alpha$ used for the precision–recall curves of figure 5. Comparing this result for $L_0$ also with the one for $L_1$ in figure B.1, we can understand the difference of their performances.

The latter will be the actual solution if and only if

$$\left(\frac{\hat{R}J^0 - \sqrt{Qx}}{2T}\right)^2 < 1.$$
whenever $(\hat{R}J^0 - \sqrt{Q})^2/(2T) \neq 1$. The point where the equality holds corresponds to the negligible case when the value at $J = 0$ is exactly equal to that at the point of zero derivative. We conclude that except for this single point, $J^*_p \to J^*_0$ as $p \to 0$, and therefore we can replace the $L_0$ norm directly into the steepest descent condition to obtain the $p \to 0$ result.

Repeating the steps shown in appendix B, a similar computation for the $L_0$ dilution gives the expectation value reported in equation (19), and the following probability distribution for the student couplings:

$$P(J) = (1 - n_{eff}) \delta(J) + \frac{K}{\sqrt{Q}\sqrt{2\pi}} \int dJ^0 \rho(J^0) \exp\left(-\frac{(\hat{R}J^0 - KJ)^2}{2Q}\right) \Theta\left(|J| - \sqrt{\frac{T}{2}}\right).$$

This distribution is shown in figure C.2 for the same four values of $\alpha$ as the precision–recall curves were shown for in figure 5. Note that the main difference between this distribution and that corresponding to the $L_1$ dilution, figure B.1, is the presence of the $\Theta(\cdot)$ function in the former. When the Gaussians of the continuous part of the distribution have a standard deviation smaller than the gap in the $\Theta$ function, the presence of false positives corresponding to the Gaussian around $J^0 = 0$ is suppressed by the $\Theta$ function, and this is why we observe such a good performance in model selection for $\alpha > 0.8$ in figure 5.

References

[1] Guyon I and Eliseeef A, 2003 J. Mach. Learning Res. 3 1157
[2] Guyon I, Gunn S, Nikravesh M and Zadeh L, 2006 Feature Extraction: Foundations and Applications (Berlin: Springer)
[3] Tibshirani R, 1994 J. R. Stat. Soc. B 58 267
[4] Ravikumar P, Wainwright M and Lafferty J, 2006 Advances in Neural Information Processing Systems 19: Proc. 20th Annual Conf. (NIPS 2006) (Cambridge, MA: MIT Press) pp 1455–72
[5] Banerjee O, El Ghaou L, d’Aspremont A and Natsoulis G, 2006 ACM International Conference Proceeding Series vol 148, pp 12–8
[6] Lee S-I, Ganapathi V and Koller D, 2007 Advances in Neural Information Processing Systems (NIPS 2006)
[7] Schmidt M, Niculescu-Mizil A and Murphy K, 2007 Proc. 22nd AAAI Conf. on Artificial Intelligence (AAI)
[8] Meinshausen N and Bühlmann P, 2006 Ann. Stat. 34 1436
[9] Bishop C M, 1996 Neural Networks for Pattern Recognition (Oxford: Oxford University Press)
[10] LeCun Y, Denker J S, Solla S, Howard R E and Jackel L D, 1990 Advances in Neural Information Processing Systems 2 (NIPS*89) ed D Touretzky (Denver, CO: Morgan Kaufman)
[11] Hassibi B and Stork D G, 1993 Advances in Neural Information Processing Systems 5 ed D S Lippman, J E Moody and D S Touretzky (Denver, CO: Morgan Kaufman) pp 164–171
[12] Bouten M, Engel A, Komoda A and Serneels R, 1990 J. Phys. A: Math. Gen. 23 4643
[13] Garces R, Kuhlmann P and Eissfeller H, 1992 J. Phys. A: Math. Gen. 25 L1335
[14] Kuhlmann P, Garces R and Eissfeller H, 1992 J. Phys. A: Math. Gen. 25 L593
[15] Wong K Y M and Bouten M, 1991 Europhys. Lett. 16 525
[16] Komoda A, Serneels R, Wong K Y M and Bouten M, 1991 J. Phys. A: Math. Gen. 24 L743
[17] Kuhlmann P and Muller K R, 1994 J. Phys. A: Math. Gen. 27 3759
[18] Malzahn D, 2000 Phys. Rev. E 61 6261
[19] Braunstein A, Pagnani P, Weigt M and Zecchina R, 2008 J. Phys.: Conf. Ser. 95 012016
[20] Braunstein A, Pagnani A, Weigt M and Zecchina R, 2008 J. Stat. Mech. P12001
[21] Pagnani A, Trias F and Weigt M, 2009 J. Stat. Mech. P05001
[22] Ueda S and Kabashima Y, 2005 J. Phys. Soc. Japan 74 2233
[23] Gardner E, 1987 Europhys. Lett. 4 481
[24] Gardner E, 1988 J. Phys. A: Math. Gen. 21 257
[25] Györgyi G, 1990 Phys. Rev. Lett. 64 2057
[26] Seung H S, Sompolinsky H and Tishby N, 1992 Phys. Rev. A 45 6056

doi:10.1088/1742-5468/2009/10/P10009
[27] Engel A and Van Den Broeck C P, 2001 Statistical Mechanics of Learning (New York: Cambridge University Press)
[28] Mézard M, Parisi G and Virasoro M, 1987 Spin Glass Theory and Beyond (Singapore: World Scientific)
[29] Kabashima Y, 2008 J. Phys.: Conf. Ser. 95 012001
[30] Kabashima Y, Wadayama T and Tanaka T, 2009 arXiv:0907.0914