Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

PAPER: Biological modelling and information

Optimal regularizations for data generation with probabilistic graphical models

A Fanthomme, F Rizzato, S Cocco and R Monasson*

Laboratory of Physics of the Ecole Normale Supérieure, PSL Research and CNRS UMR8023, 24 rue Lhomond, 75005 Paris, France
E-mail: remi.monasson@phys.ens.fr

Received 7 December 2021
Accepted for publication 28 March 2022
Published 9 May 2022

Online at stacks.iop.org/JSTAT/2022/053502
https://doi.org/10.1088/1742-5468/ac650c

Abstract. Understanding the role of regularization is a central question in statistical inference. Empirically, well-chosen regularization schemes often dramatically improve the quality of the inferred models by avoiding overfitting of the training data. We consider here the particular case of L_2 regularization in the maximum *a posteriori* (MAP) inference of generative pairwise graphical models. Based on analytical calculations on Gaussian multivariate distributions and numerical experiments on Gaussian and Potts models we study the likelihoods of the training, test, and 'generated data' (with the inferred models) sets as functions of the regularization strengths. We show in particular that, at its maximum, the test likelihood and the 'generated' likelihood, which quantifies the quality of the generated samples, have remarkably close values. The optimal value for the regularization strength is found to be approximately equal to the inverse sum of the squared couplings incoming on sites on the underlying network of interactions. Our results seem to be robust against changes in the structure of the ground-truth underlying interactions that generated the data, when small fluctuations of the posterior distribution around the MAP estimator are taken into account, and when L_1 regularization is considered (instead of L_2). Connections with empirical works on protein models learned from homologous sequences are discussed.

*Author to whom any correspondence should be addressed.

© 2022 IOP Publishing Ltd and SISSA Medialab srl

1742-5468/22/053502+29$33.00

Keywords: inference of graphical models, statistical inference in biological systems, protein function and design

Contents

1. Introduction

Data-driven modeling is now routinely used to address hard challenges in an increasing number of fields of science and engineering for which first-principle approaches have limited success. Applications include the characterization and design of complex materials (Schmidt *et al* 2019), shaped by the pattern of strong and heterogeneous

J. Stat. Mech. (2022) 053502

interactions between their microscopic components. Performance of data-driven models strongly depends on the choice of their hyperparameters, such as the architecture, and the strengths of the regularization penalties. These parameters are generally set through empirical procedures, such as cross-validation with respect to a goodness-of-fit estimator. Unfortunately, this common approach often offers no insight about why these values of the parameters are optimal, and may not guarantee that the obtained models are well-behaved with respect to other estimators. This paper reports some efforts to address these issues for the specific case of L_2-norm regularization and probabilistic graphical models.

Probabilistic graphical models rely on the inference of the set of conditional dependencies between the variables under study, which, in turn, may be used to generate new configurations of these variables (MacKay 2003). Regularization allows the graph of pairwise conditional dependence to satisfy some properties of interests, such as to be sparse or to have dependence factors bounded from above. While the effects of finite sampling and of regularization on the estimation of the covariance matrix have been extensively studied in the statistics community (Ledoit and Wolf 2004, Huang *et al* 2006, Karoui 2008, Ravikumar *et al* 2011) fewer efforts have been devoted to the characterization of the generative performance of the inferred models, which are however crucial in many applications. One of these applications is the modeling of proteins based on homologous, i.e. evolutionary related sequence data. Unveiling the relations between the functional or structural properties of a protein and the sequence of its amino acids is a difficult task. Graphical model-based modeling consists of inferring a graph of effective interactions between the amino acids, which reproduce the low-order (one- and two-point) statistics in the sequence data; for reviews, see Cocco *et al* (2018) for protein modeling and Chau Nguyen *et al* (2017) for general inference of graphical models with discrete variables. In practice, for proteins with few hundreds of amino acids, tens of millions of interaction parameters have to be inferred. To avoid overfitting, regularization of those interactions, often based on pseudocounts, or L_1- and L_2-norms are generally introduced, with intensities varying with the optimality criteria chosen by the authors Barton et *al* (2014), Haldane and Levy (2019). For instance, Ekeberg *et al* chose regularization strength scaling linearly with the number of data (sequences) (Ekeberg *et al* 2013, 2014) to maximize the quality of structural predictions. Hopf *et al* chose linear scaling with the dimension of the data (sequence length) and with the number of possible amino-acid types (generally, $q = 20$) for predicting the fitness effects resulting from mutations along the sequence (Hopf *et al* 2017). The rationale for these scalings and what they tell us about the underlying properties of the protein system remains unclear. In addition, whether these scalings are appropriate for generating new data points, i.e. for the design of new protein sequences having putative properties is not known, and other regularization schemes have been proposed (Barrat-Charlaix *et al* 2021)

In the following, we propose to study the role of regularization in the inference process, replacing Potts models by multivariate Gaussian models in order to make the problem analytically tractable in some limiting cases. We show that two natural definitions for the optimal values of the regularization strength are in practice very close to one another, and that their common value can be related to the amplitude of the ground-truth interactions, in agreement with experimental observations. Our paper is

J. Stat. Mech. (2022) 053502

organized as follows. In section 2, we introduce the Gaussian model and the regularizations of interest. Numerical results are reported in section 3. Section 4 is devoted to the analytical studies of the poor and excellent sampling limits. Last of all, some conclusions and perspectives are drawn in section 5.

2. Gaussian vectors model and regularization

2.1. Expression of likelihood in the large-size limit

In order to be able to model distributions over n-dimensional vectors, we consider first the multidimensional Gaussian distribution, often referred to as Gaussian vectors or spherical model. In the following, we will only consider the case of centered Gaussian vectors, for which the mean value of each component vanishes and the probability density is given by:

$$p(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{C}^{\rm tr})}} e^{-\frac{1}{2}\boldsymbol{x}^{\rm T}(\boldsymbol{C}^{\rm tr})^{-1}\boldsymbol{x}}, \tag{1}$$

where $\boldsymbol{C}^{\rm tr}$ is the $n \times n$-dimensional covariance matrix. Alternatively we may define the underlying data distribution through an interaction matrix $\boldsymbol{J}^{\rm tr}$, which represents the interaction strength between the variables (vector components). This interaction matrix $\boldsymbol{J}^{\rm tr}$ is related to the true covariance matrix $\boldsymbol{C}^{\rm tr}$ of the data through

$$\boldsymbol{C}^{\rm tr} = (\mu^{\rm tr}\boldsymbol{I} - \boldsymbol{J}^{\rm tr})^{-1}, \tag{2}$$

where $\mu^{\rm tr}$ was introduced to impose the spherical normalization constraint $\mathrm{Tr}(\boldsymbol{C}^{\rm tr}) = n$. Denoting as $(j_1^{\rm tr}, \ldots, j_n^{\rm tr})$ the eigenvalues of $\boldsymbol{J}^{\rm tr}$, the normalization condition can be written, in the large n limit, as

$$1 - \frac{1}{n}\sum_{k=1}^{n} \frac{1}{\mu^{\rm tr} - j_k^{\rm tr}} = 0. \tag{3}$$

As the covariance matrix is non-negative we are looking for the unique value of $\mu^{\rm tr}$ in $[\max_k\{j_k^{\rm tr}\}, +\infty]$ that satisfies this equation.

In the following, we will be interested in inferring the interaction matrix $\boldsymbol{J}^{\rm tr}$ from an empirical approximation $\boldsymbol{C}^{\rm emp}$ of the correlation matrix obtained using $p = \alpha n$ samples[1] $(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^p)$ as:

$$\forall (i,j) \in [1,n]^2,\ \boldsymbol{C}^{\rm emp}_{i,j} = \frac{1}{p}\sum_{k=1}^{p} \boldsymbol{x}_i^k \boldsymbol{x}_j^k. \tag{4}$$

We define the posterior density of probability of any interaction matrix $\boldsymbol{J}$ given the empirical covariance matrix $\boldsymbol{C}^{\rm emp}$,

$$p(\boldsymbol{J}|\,\boldsymbol{C}^{\rm emp}) = e^{-n\,E(\boldsymbol{J})}, \tag{5}$$

[1] We here insist on the fact that our notation is standard for physics, and opposite to the one in statistics, where n usually denotes the number of samples and p the number of features.

J. Stat. Mech. (2022) 053502

where the energy function $E(\boldsymbol{J})$ reads

$$E(\boldsymbol{J}) = -\frac{\alpha}{2}\,\mathrm{Tr}(\boldsymbol{J}\boldsymbol{C}^{\mathrm{emp}}) + \alpha\,\log\,Z(\boldsymbol{J}) + \frac{\gamma}{4}\,\mathrm{Tr}(\boldsymbol{J}^2). \tag{6}$$

In the expression above the first two terms correspond to the standard likelihood of a given Gaussian model given the empirical covariance, while the last term expresses a penalty on the L_2 norm of the inferred interaction matrix. The strength of this regularization is controlled by the parameter γ.

The partition function $Z(\boldsymbol{J})$ of the so-called spherical spin model reads

$$\begin{aligned}\log\,Z(\boldsymbol{J}) &= \int_{\boldsymbol{x}\in\mathbb{R}^n} \delta(\boldsymbol{x}^2 = n)\, e^{\frac{1}{2}\sum_{i\neq j} x_i J_{ij} x_j} \\ &\overset{n\to\infty}{\sim} n \min_{\mu}\left(\frac{\mu}{2} - \frac{1}{2n}\,\log(\det(\mu\boldsymbol{I} - \boldsymbol{J}))\right),\end{aligned} \tag{7}$$

to the dominant order in n. The parameter μ can be interpreted as a Lagrange multiplier, introduced to impose the spherical constraint $\mathrm{Tr}(\boldsymbol{C}) = n$, which corresponds exactly to the normalization condition (3) but with the eigenvalues of the true interaction matrix j^{tr} replaced by the ones of $\boldsymbol{J}$.

Our goal will be to minimize the energy (6) with respect to the interaction matrix $\boldsymbol{J}$; the matrix $\boldsymbol{J}^*$ minimizing the energy will be called inferred matrix and will be our primary object of study. We also define μ^* the Lagrange multiplier imposing the spherical constraint on this inferred model, and $\boldsymbol{C}^*$ the covariance matrix of the inferred model. For reference, we define in table 1 all the different quantities that we will be considering and their associated notations.

2.2. MAP estimator of the interaction matrix

When γ is equal to 0, the regularization disappears and the maximum likelihood estimation of $\boldsymbol{J}^*$ is exactly equal to the one computed from the empirical covariance $\boldsymbol{C}^{\mathrm{emp}}$; when γ goes to infinity, the regularization becomes so strong that the inferred interaction matrix is exactly equal to $\boldsymbol{0}$; in the general case of finite γ, we find $\boldsymbol{J}^*$ by computing $\frac{\partial E}{\partial \boldsymbol{J}}(\boldsymbol{J}^*)$, which yields the maximum *a posteriori* (MAP) equation:

$$\gamma\boldsymbol{J}^* - \alpha\,\boldsymbol{C}^{\mathrm{emp}} + \alpha(\mu^*\boldsymbol{I} - \boldsymbol{J}^*)^{-1} = 0. \tag{8}$$

According to equation (8) the inferred interaction matrix $\boldsymbol{J}^*$ is diagonal in the same vector basis as the empirical covariance matrix $\boldsymbol{C}^{\mathrm{emp}}$. It is therefore possible to rewrite this equation in terms of the eigenvalues (respectively, j^*, c^{emp}) of those matrices[2]:

$$\gamma\, j^{*2} - (\gamma\mu^* + \alpha c^{\mathrm{emp}})\, j^* + \alpha(\mu^* c^{\mathrm{emp}} - 1) = 0. \tag{9}$$

[2] Because of equation (8), we know that to each eigenvalue of the empirical covariance matrix corresponds exactly one eigenvalue of the inferred interaction matrix.

J. Stat. Mech. (2022) 053502

Table 1. All quantities used in the inference procedure. Please note that the empirical covariance matrix $\boldsymbol{C}^{\text{emp}}$ and its eigenvalues are stochastic quantities for a given underlying interaction model $\boldsymbol{J}^{\text{tr}}$ (since they depend on the exact samples drawn). Additionally, we will assume the eigenvalues c to be ordered from largest to smallest, and denote with a lower-index k both c_k^{emp} (the kth largest eigenvalue of $\boldsymbol{C}^{\text{emp}}$) and j_k^* the corresponding eigenvalue of $\boldsymbol{J}^*$ (see equation (10)).

Symbol	Quantity
$\boldsymbol{I}$	The identity matrix
n	Dimension of the Gaussian vectors
p	Number of samples
α	Sampling ratio p/n
γ	The strength of the L_2 penalty
$\boldsymbol{J}$	Dummy variable standing for an interaction matrix
$\boldsymbol{C}$	Dummy variable standing for a covariance matrix
$\boldsymbol{J}^{\text{tr}}$	True interaction matrix of the underlying model
$\boldsymbol{C}^{\text{tr}}$	True covariance matrix of the underlying model
$\boldsymbol{C}^{\text{tr,rot}}$	True covariance matrix, in the diagonalizing basis of $\boldsymbol{C}^{\text{emp}}$
c^{tr}	An eigenvalue of the true covariance matrix
μ^{tr}	Lagrange multiplier imposing the spherical constraint on $\boldsymbol{J}^{\text{tr}}$
$\boldsymbol{C}^{\text{emp}}$	Empirical covariance matrix obtained from $p = \alpha n$ samples
c^{emp}	Eigenvalue of the empirical covariance matrix
$\boldsymbol{J}^*$	Interaction matrix obtained from MAP inference
j^*	Eigenvalue of the MAP inferred interaction matrix
μ^*	Lagrange multiplier imposing the spherical constraint on $\boldsymbol{J}^*$

Since the discriminant $\Delta = (\alpha c^{\text{emp}} - \gamma\mu^*)^2 + 4\alpha\gamma \geqslant 0$, the eigenvalue $j^*(c^{\text{emp}})$ always exists in $\mathbb{R}$ and is found to be equal to:

$$j^*(c^{\text{emp}}) = \frac{1}{2\gamma}\left(\alpha c^{\text{emp}} + \gamma\mu^* - \sqrt{(\alpha c^{\text{emp}} - \gamma\mu^*)^2 + 4\alpha\gamma}\right). \tag{10}$$

It should be noted here that this is in fact a self-consistent equation: μ^* is used to compute the eigenvalues j^*, which in turn are used to compute μ^*. In order to solve it, we consider μ^* to be a free parameter and make the expression of the inferred eigenvalues depend on two variables $j^*(c^{\text{emp}},\ \mu^*)$. Introducing the corresponding expression into the normalization condition (3), we find that μ^* is the only root[3] of the residual function:

$$\text{Res}^{\text{norm}}\ (\mu) = 1 - \frac{1}{n}\sum_k \frac{1}{\mu - \frac{1}{2\gamma}\left(\alpha c_k^{\text{emp}} + \gamma\mu - \sqrt{(\alpha c_k^{\text{emp}} - \gamma\mu)^2 + 4\alpha\gamma}\right)}. \tag{11}$$

In practice, the optimization of this residual is performed numerically in Python using the Van Wijngaarden–Dekker–Brent method (Brent 2013), implemented within the SciPy package (Virtanen *et al* 2020). After obtaining the value of μ^*, the inferred

[3] It can easily be shown that $\partial j^*(c^{\text{emp}})/\partial\mu^*$ is always positive; since $j^*(c^{\text{emp}},\ \mu) < \mu$, we have that $\text{Res}(\mu)$ is well-defined for all values of μ; $\partial\text{Res}(\mu)/\partial\mu$ is always positive and therefore $\text{Res}(\mu)$ is monotonically increasing from $-\infty$ when $\mu \to -\infty$ to 1 when $\mu \to +\infty$, ensuring the unicity of the root.

J. Stat. Mech. (2022) 053502

interaction matrix $\boldsymbol{J}^*$ is obtained by computing its spectrum through equation (10) and changing the basis back from the inference basis (which diagonalizes the empirical covariance $\boldsymbol{C}^{\text{emp}}$) to the original basis (in which the true interaction $\boldsymbol{J}^*$ was defined).

2.3. Likelihoods of the training, test, and generated sets

In order to be able to compare the quality of the inferred interaction matrix $\boldsymbol{J}^*$ as a function of the different parameters of the system (namely, α, γ and the true interaction matrix $\boldsymbol{J}^{\text{tr}}$) the first interesting quantity to define is the training likelihood:

$$L_{\text{train}} = \frac{1}{p}\sum_{k=1}^{p}\left[\frac{1}{2}\sum_{i,j} J^*_{ij}\, x_i^k x_j^k - \log\, Z(\boldsymbol{J}^*)\right], \tag{12}$$

which directly quantifies how well the examples of the training set are fit by the MAP estimator $\boldsymbol{J}^*$. By performing the summation over the sample index k, the likelihood can be rewritten as a function of the empirical covariance matrix $\boldsymbol{C}^{\text{emp}}$:

$$L_{\text{train}} = \frac{1}{2}\sum_{i,j} J^*_{ij} C^{\text{emp}}_{ij} - \log\, Z(\boldsymbol{J}^*). \tag{13}$$

A similar reasoning can be performed, this time considering the case where an infinite number of samples are drawn from the true underlying distribution (meaning that C^{emp} is replaced by $\boldsymbol{C}^{\text{tr}}$), corresponding to the average test error on samples independent of the training ones. This leads to the definition of the test likelihood:

$$L_{\text{test}} = \frac{1}{2}\sum_{i,j} J^*_{ij} C^{\text{tr}}_{ij} - \log\, Z(\boldsymbol{J}^*). \tag{14}$$

Finally, one can also consider the likelihoods of a 'generated set' of examples drawn using the inferred interaction matrix, with respect to this same inferred interaction matrix $\boldsymbol{J}^*$:

$$L_{\text{gen}} = \frac{1}{2}\sum_{i,j} J^*_{ij} C^*_{ij} - \log\, Z(\boldsymbol{J}^*). \tag{15}$$

It is possible to rewrite the 'generated' likelihood using the MAP equation:

$$\begin{aligned} L_{\text{gen}} &= \frac{1}{2}\sum_{i,j} J^*_{ij} C^*_{ij} - \log\, Z(\boldsymbol{J}^*) = \frac{1}{2}\sum_{i,j} J^*_{ij} \left.\frac{1}{\mu \boldsymbol{I} - \boldsymbol{J}^*}\right|_{ij} - \log\, Z(\boldsymbol{J}^*) \\ &\overset{(8)}{=} \frac{1}{2}\sum_{i,j} J^*_{ij}\left(\boldsymbol{C}^{\text{emp}}_{ij} - \frac{\gamma}{\alpha}\boldsymbol{J}^*_{ij}\right) - \log\, Z(\boldsymbol{J}^*) = L_{\text{train}} - \frac{\gamma}{2\alpha}\sum_{i,j} {J^*_{ij}}^2. \end{aligned} \tag{16}$$

This form of the generated likelihood can be interpreted as a form of bias-variance trade-off: if an increase in the magnitude of the couplings is necessary to better fit the training set, it will increase the variance of the generated data and consequently decrease the generated set likelihood.

J. Stat. Mech. (2022) 053502

2.4. Generic dependence of the likelihoods upon regularization strength

Figure 1 is a sketch of the typical behaviors expected for the three log-likelihoods defined above as the regularization strength γ is varied:

- For weak regularization i.e. γ close to zero MAP inference is unconstrained, and the inferred covariance coincides with the empirical one. The value of the training likelihood is large, as the details of the training set are fitted. Consequently, the inferred model has poor generalization capability, and the test log-likelihood has a low value. This is a situation of *overfitting*. Generated data look like training data, so the generated likelihood is large.
- For strong regularization, i.e. large γ the regularization term in the energy becomes more important than the likelihood term, so that the MAP estimator $\boldsymbol{J}^*$ tends to zero; this is a case of *under fitting*, as the training, test, and generated likelihoods will be low. When γ goes to infinity, the three likelihoods converge to a common value,

$$L(\gamma \to \infty) = -\frac{n}{2}. \tag{17}$$

- In-between those two regimes, i.e. for intermediate values of γ the training likelihood is monotonically decreasing with γ, reflecting the increasing bias toward small couplings, and so is the generated likelihood. The test likelihood displays a non-monotonic evolution, and reaches a maximum for some regularization penalty γ^{opt}. The presence of γ biases the inference, but also reduces its variance, and hence allows for better generalization of the model to unseen examples. While the test likelihood always remains smaller than the training likelihood (the model cannot generalize better than it fits the available data), the test and generated likelihoods do cross at a certain value γ^{cross}, see figure 1. This can be understood by noting that the inferred couplings $\boldsymbol{J}^*$ will scale as γ^{-1} for large γ, see equation (8), and so will the generated correlations $\boldsymbol{C}^*$, leading to a γ^{-2} scaling for $\sum_{i,j} J^*_{ij} C^*_{ij}$. Conversely, the empirical correlations do not depend on γ, therefore $\sum_{i,j} J^*_{ij} C^{\text{emp}}_{ij}$ scales as γ^{-1}, which implies that L_{test} decreases much slower than L_{gen} (figure 1). Since those two functions share a common limit when γ goes to infinity, there exists a large range of values of γ in which the generated likelihood is lower than the test likelihood.

  In the following we will study, through numerical experiments and analytical calculations the behavior of these two regularization strengths of interest, and their dependence on the model defining parameters (number p of samples compared to the size n, structure of the coupling matrix, ...).

3. Numerical experiments

3.1. Gaussian vectors model

In order to study the dependence of $\gamma^{\text{opt}}, \gamma^{\text{cross}}$ with the different parameters, we implemented the MAP inference procedure in Python (the code is available on GitHub).

J. Stat. Mech. (2022) 053502

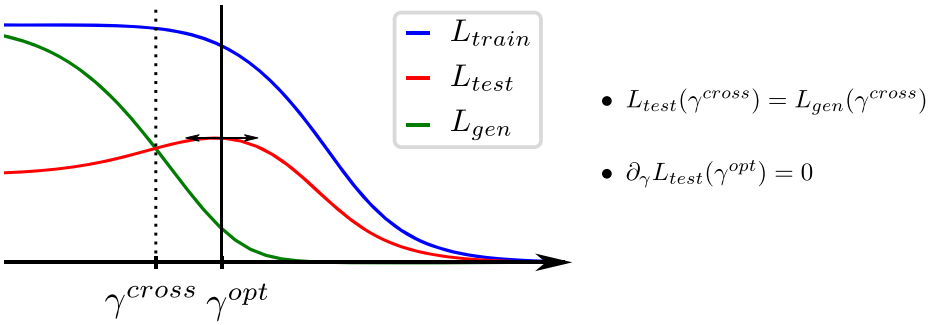

Figure 1. Sketch of the expected behaviors of the likelihoods vs regularization γ, and definitions of the two values of interest: γ^{cross}, for which the test and generated likelihoods are equal; γ^{opt}, for which the test likelihood is maximal. The difference between optimal and crossing likelihoods is strongly magnified for illustration purposes, as in practice they are found to be extremely close to one another in most circumstances.

The general procedure is as follows: first, an interaction matrix $\boldsymbol{J}^{\text{tr}}$ is randomly generated, according to an underlying distribution (see next subsections for details on the distributions we considered); then, a certain number $p = \alpha n$ samples are drawn from the Gaussian vectors model with interactions $\boldsymbol{J}^{\text{tr}}$, and from those samples an empirical covariance matrix $\boldsymbol{C}^{\text{emp}}$ is derived; this matrix is then diagonalized, and the spectrum of the MAP interaction estimator $\boldsymbol{J}^*$ is computed through equation (10); the training and generated set likelihoods are computed directly using those eigenvalues, while the test likelihood requires the inversion of the diagonalization basis change in order to obtain the expression of $\boldsymbol{J}^*$ in the same basis as $\boldsymbol{C}^{\text{tr}}$.[4]

3.1.1. Case of random quenched couplings. **The condensation phase transition.** We assume that the underlying interaction matrix is drawn from the Gaussian orthogonal ensemble, i.e. all its components are drawn at random and independently from a centered Gaussian distribution:

$$\forall\, i, j, \quad J_{ij}^{\text{tr}} \sim \mathcal{G}\left(0, \frac{\sigma}{\sqrt{n}}\right). \tag{18}$$

The presence of this $1/\sqrt{n}$ normalization ensures that the energy is extensive with n. The model is 'infinite range' because all spins are interacting with all other spins with similar strengths, controlled by the parameter σ. As shown in Kosterlitz *et* al (1976) the model exhibits a condensation phase transition when σ crosses the critical value $\sigma_{\text{c}} = 1$. For $\sigma > \sigma_{\text{c}}$ one eigenvalue of the covariance matrix scales linearly with n, while all others remain finite. This transition can be intuitively understood as follows. Since the interaction matrix $\boldsymbol{J}^{\text{tr}}$ has Gaussian entries, its eigenvalue distribution follows Wigner's semi-circle law, and ranges from -2σ and 2σ. As σ increases from small values, the value of the Lagrange multiplier μ imposing the spherical constraint becomes closer and closer to its lower-bound 2σ, and the gaps closes (in the infinite n limit) when $\sigma = \sigma_{\text{c}}$.

[4] Those two basis *a priori* coincide if and only if $\alpha \to \infty$.

J. Stat. Mech. (2022) 053502

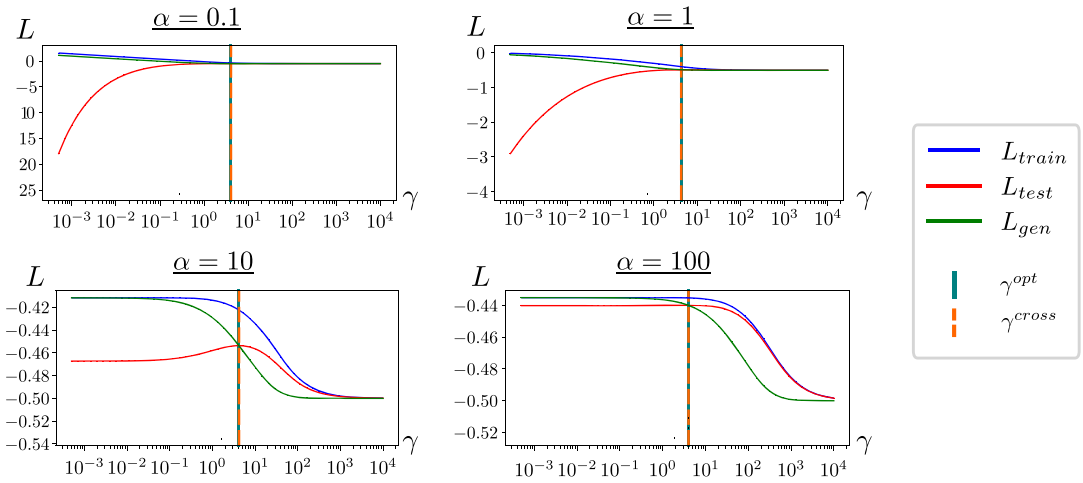

Figure 2. Evolution of the four likelihoods (normalized by n) as functions of the regularization strength γ for four different values of the sampling ratio α. In all cases, both training and generated likelihoods are monotonically decreasing, while the test likelihood is first increasing then decreasing; the training and test likelihoods never cross, while the generated and test likelihoods cross for a value of the regularization extremely close to the optimum of L_{test}.

For $\sigma > \sigma_c \mu$ remains equal to 2σ, and the corresponding top eigenvector of $\boldsymbol{J}^{\text{tr}}$ gives rise to an extensively large eigenvalue in $\boldsymbol{C}^{\text{tr}}$. More precisely, when σ is larger than σ_c, the maximum eigenvalue of $\boldsymbol{C}^{\text{tr}}$ is equal to

$$c^{\text{tr}}_{\max} = n \times \left(1 - \frac{1}{2\pi\sigma^2}\int_{-2\sigma}^{2\sigma} \frac{\sqrt{4\sigma^2 - j^2}}{2\sigma - j} dj\right) = n\left(1 - \frac{1}{\sigma}\right). \tag{19}$$

In this situation, the model generates configurations that are effectively constrained close to a subspace of dimension 1.

Evolution of the log-likelihoods with γ. Figure 2 shows the behaviors of the log-likelihoods with varying γ, for different regimes of low and high sampling fractions α. Vertical lines locate the three values of γ of interest. The overall shape of the curves agree with the expected behaviors sketched in figure 1.

For small γ (overfitting regime), the value of the training likelihood is very large, irrespective of the value of α as the weak regularization allows the inference procedure to fit the training set without bias. The test loss, however, strongly varies with α. For low sampling (small α) $\boldsymbol{C}^{\text{emp}}$ is essentially uncorrelated with $\boldsymbol{C}^{\text{tr}}$, and the test likelihood will be very low. If α is large, $\boldsymbol{C}^{\text{emp}}$ is almost equal to $\boldsymbol{C}^{\text{tr}}$, and the test likelihood will be very close to its training counterpart, both being very high. In all cases the generated and the training log-likelihoods coincides.

When γ is very large, the regularization term in the energy pushes the MAP estimator $\boldsymbol{J}^*$ toward 0. In this *underfitting* regime, all log-likelihoods tend to the same limit value, see equation (17).

J. Stat. Mech. (2022) 053502

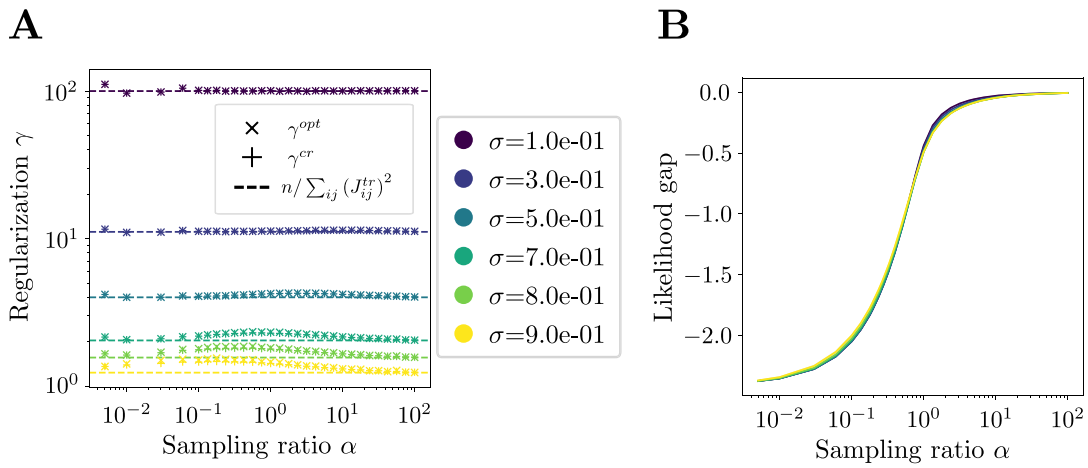

Figure 3. Gaussian vectors model with L_2 regularization. (A) Evolution of the regularizations γ^{opt} and γ^{cross} as functions of the sampling ratio α for different values of the interaction dispersion σ, see equation (18). The theoretical prediction for γ^{cross}, represented here as a dashed line for each value of σ, is given in equation (20) and derived in section 4. (B) Evolution of the likelihood gap $\Delta L = L_{\text{test}}(\alpha, \gamma^{\text{opt}}(\alpha)) - L_{\text{train}}(\alpha = \infty, \gamma = 0)$ as a function of α for the same values of σ as (A). As expected, this gap vanishes as α goes to infinity, meaning that the optimal inferred model (obtained with non-zero regularization) fits the data perfectly in the limit of infinite samples. While the gaps are identical between different values of the interaction strength, we were not able to determine the expression for this evolution.

For intermediate γ, we observe that the location of the maximum of the test likelihood, γ^{opt}, is very close to the value of the regularization strength γ^{cross} for which it crosses the generated log-likelihood. This unexpected results holds in most circumstances as reported in figure 3. This is true both on average and for individual realizations of the underlying interaction matrix $\boldsymbol{J}^{\text{tr}}$ and correlation matrix $\boldsymbol{C}^{\text{emp}}$, although small discrepancies can be observed at low sampling ratio α. We expect those discrepancies to be related to the finite size of the system, and the equality to be recovered in the $n \to \infty$ limit, as detailed analytical calculations for the Gaussian vectors model in section 4 will confirm. These calculations will also allow us to approximate their common value for large sampling ratio α as a function of the 'true' interaction matrix $\boldsymbol{J}^{\text{tr}}$ only:

$$\gamma^{\text{opt}} \simeq \gamma^{\text{cross}} \simeq \frac{n}{\sum_{i,j}(J_{ij}^{\text{tr}})^2}. \tag{20}$$

In order to estimate γ^{opt} and γ^{cross} as precisely as possible, we define in appendix A functions whose roots correspond to those regularizations, and optimize them numerically with care.

3.1.2. Other types of underlying interactions. The empirical coincidence between γ^{opt} and γ^{cross} reported above extends to other choices of the coupling matrices. As an illustration we consider the case where the underlying interaction matrix $\boldsymbol{J}^{\text{tr}}$ is structured,

J. Stat. Mech. (2022) 053502

instead of being randomly drawn. In particular, we present in figure 4 two examples, and show that the presence of structure does not significantly alter our previous observations:

- In panel (A), the interaction matrix is band-diagonal, meaning that the coefficients are given by

$$\forall\,(i,j)\ \text{s.t.}\ |(i-j)\bmod n| < \frac{w}{2},\quad J_{ij}^{\mathrm{tr}} \sim \mathcal{G}\left(0, \frac{\sigma}{\sqrt{w}}\right), \tag{21}$$

  where w is the width of the non-zero band, $\mathcal{G}$ is the Gaussian distribution, and $[n]$ represents the 'modulo n' operation. This means that sites are arranged on a ring, with interactions only between w nearest neighbors, and the value of those non-zero interactions are drawn randomly from a Gaussian distribution.

  This model can be related to the random Schrödinger operator in dimension 1, an object extensively studied in the context of Anderson localization, see Anderson (1958). As observed numerically by Casati *et al* (1990) and later rigorously proved (see Bourgade (2018) for an overview), a phase transition can be observed when $w \sim \sqrt{n}$ between a regime (small w) where the eigenvectors of $\boldsymbol{J}^{\mathrm{tr}}$ are localized i.e. decay exponentially with distance, and another where they are extended (large w).

  Our particular choice of scaling of the individual entries of those band matrices is such that $\frac{1}{n}\sum_{i,j}(J_{ij}^{\mathrm{tr}})^2$ remains constant, and so do the expected values of the regularizations of interest.

- In panel (B), $\boldsymbol{J}^{\mathrm{tr}}$ is a deterministic matrix corresponding to a unidimensional chain:

$$\forall\,(i,j),\quad J_{ij}^{\mathrm{tr}} = \begin{cases} 0 \text{ if } i=j \text{ or } |(i-j)\bmod n| > 1 \\ \sigma \text{ if } |(i-j)\bmod n| = 1 \end{cases}, \tag{22}$$

  meaning that sites are again arranged on a ring, this time with fixed positive interactions between direct neighbors only. This particularly simple model does not exhibit any phase transition.

We find that changing the underlying model of interaction does not significantly impact the phenomenology that we previously observed for infinite-range Gaussian interactions: an optimal regularization still exists for all values of the sampling ratio α.

3.1.3. L_1 regularization. While the L_2 penalty is often used in practice, and encourages smoothness of the energy landscape, it is not the only possible choice. In many cases, it can be interesting to infer sparse interactions models, which is usually done by using an L_1 regularization: in a protein, amino acids which are very distant in the sequence can end up close in the folded structure, and therefore interact strongly so that one has to *a priori* allow interactions between all sites along the sequence; however, in three-dimensional space, each site is close only to a very small fractions, so that the inferred interaction matrix should be sparse. The inference procedure in this case is less straightforward than for the L_2 case, and analytical solutions cannot be obtained in the general case. Instead, one relies on the so-called 'graphical Lasso' method (Friedman *et al* 2008),

J. Stat. Mech. (2022) 053502

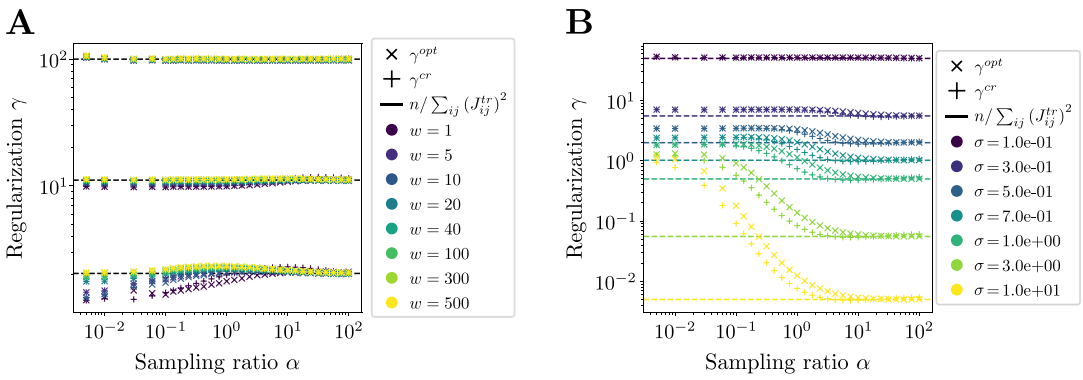

Figure 4. Evolution of the regularizations of interest for two different cases of structured interaction matrices $\boldsymbol{J}^{\mathrm{tr}}$. (A) Case of a random band matrix, for different values of the interaction strength σ (top: $\sigma = 0.1$; middle: $\sigma = 0.3$; bottom: $\sigma = 0.7$). (B) Case of a deterministic, uniform one-dimensional chain. In both cases, the crossing and optimal regularizations are of the same order of magnitude, and remain close to the predicted values in the $\alpha \to \infty$ regime, represented by the dotted lines. For small ratios α the behaviors of the optimal regularizations depend on w, as expected from section 4 by noting that w has an influence on the $\langle\theta\rangle$ quantity defined in equation (44).

which iteratively solves Lasso problems for each column of the interaction matrix using coordinate descent (Wright 2015) until convergence, implemented in Scikit-learn (Pedregosa *et al* 2011).

We show in figure 5 that the behavior of the likelihoods remains qualitatively similar to what we observed in the case of L_2 regularization, despite the difference between the two noticeable regularizations being much higher than previously. A detailed analysis of this inference procedure could both shed light on the difference between the two, and give us a theoretical prediction for the optimal regularization in this regime, but this remains to be done in future work.

3.2. Potts model

In this section, we investigate numerically the effect of L_2 regularization on the generative properties of Potts models. Our motivation is two-fold. First we want to study to what extent the results obtained for Gaussian distributions extrapolate to non Gaussian ones. Second Potts models are especially relevant for the case of protein modeling, see introduction and discussion sections.

3.2.1. Generation of synthetic data and energy model. We now consider a discrete-valued graphical model, in which each (categorical) variables may take one out of q values. The energy of a configuration $\boldsymbol{x}$ is given by

$$E(\boldsymbol{x}; \boldsymbol{h}, \boldsymbol{J}) = -\sum_{i<j} J_{ij}(x_i, x_j) - \sum_i h_i(x_i), \tag{23}$$

J. Stat. Mech. (2022) 053502

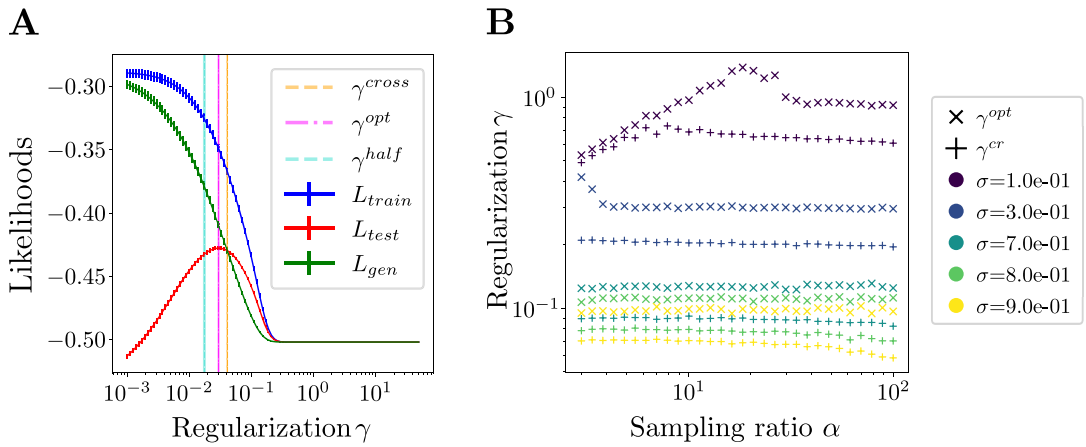

Figure 5. (A) Typical evolution of the likelihoods as a function of the strength of the L_1 regularization. The existence of a finite optimal regularization, as well as the crossing between test and generated likelihoods, remains true as in the L_2 case. (B) Evolution of the crossing and optimal regularizations as a function of the sampling ratio α. While the two noticeable regularizations are no longer equal, they remain of a similar order of magnitude. Results obtained for fully connected, random interactions (18).

The local fields $\boldsymbol{h}$ and the couplings $\boldsymbol{J}$ are, respectively, q-dimensional vectors and $(q \times q)$-dimensional matrices. The corresponding partition function is

$$Z(\boldsymbol{h}, \boldsymbol{J}) = \sum_{\{x_i=1,2,\ldots,q\}} e^{-E(\boldsymbol{x};\boldsymbol{h},\boldsymbol{J})}. \tag{24}$$

We start by drawing the components of $\boldsymbol{h}^{\rm tr}$ and $\boldsymbol{J}^{\rm tr}$ that from Gaussian distributions of zero mean and standard deviations σ_h^2 and σ_J^2. All components of the h vectors and J matrices are chosen at random and independently from each other.

Next, each element of the Gaussian matrix $J_{ij}^{\rm tr}$ is multiplied by a connectivity indicator equal to 0 or 1, which identifies, respectively, the absence or the presence of an edge between the variables i and j in the coupling network. In practice, we choose this connectivity at random, following the prescription of the so-called Erdös–Rényi (ER) random graph ensemble. For each pair i, j of variables we chose to insert an edge in the interaction graph with probability d/n, and to have no connection with probability $1 - d/n$; $d/n \times (n-1)$ is therefore the average degree of each variable in the connectivity graph.

In our simulations, we vary

- The size (number of variables), n; here $n = 25, 50, 100, 150$;
- The number of Potts states, q (here $q = 10, 20$);
- The probability d/n to include edges in the ER graph. Different values of d were tested only for $n = 25$, for which the computation were faster: $d = 1.25, 2.5, 7, 10$.

J. Stat. Mech. (2022) 053502

For each system, a number p of data point, ranging from 10^2 to 10^5 were generated by Markov chain Monte Carlo sampling. Intuition about the sampling level can be obtained by comparing p with the number of parameters to infer from the data, $n \times q + \frac{1}{2}n(n-1) \times q^2$. The parameters defining the Gaussian distribution to generate fields and coupling are here kept constant as n varies: $\sigma_h^2 = 5$, $\sigma_J^2 = 1$.

3.2.2. Behaviors of the train, test, and generated log-likelihoods. Once the data are generated through Monte Carlo sampling of the Gibbs distribution associated to the energy (23) we infer the model parameters $h_i(x), J_{ij}(x, x')$ using two methods. The first one is the pseudo-likelihood method (PLM), a non-Bayesian inference method that bypasses the (intractable) computation of the partition function Z (Ravikumar *et al* 2010, Ekeberg *et al* 2013); Z can then be estimated using the annealed importance aampling (AIS) method. The second one is the so-called adaptive cluster expansion (ACE) algorithm, which recursively computes better and better approximations for the cross-entropy of the data (Cocco and Monasson 2011, Barton *et al* 2016), combined with color compression (Rizzato *et* a*l* 2020); ACE then provides an approximate value for $\log Z$, which we could compare to the estimate found through AIS. In practice, we checked that both methods give quantitatively similar results, both for model parameters and for $\log Z$.

The inference is done with a L_2-norm regularization on the couplings (intensity γ) and on the fields (intensity γ_h). We expect regularization to be much less needed for the fields, because single-site frequencies are much better sampled than pairwise frequencies. We therefore fix the ratio between the regularization of fields and couplings, setting $\gamma_h = \gamma/(10n)$, and vary γ.

In figure 6, we show the average log-likelihoods (normalized by n) of the data in the training set, in the test set (same size as the training set) and the generated data set. For small regularization γ we observe a strong overfitting effect as expected, with similar values for L_{train} and L_{gen}, much above L_{test}. For intermediate regularization values, the test and generated log-likelihoods are similar as the number p of samples available for the inference increases, while the size n is kept fixed. This result is compatible with a weak dependence of γ^{cross} upon α, as found for the Gaussian vectors model. For large γ, L_{gen} may get smaller than L_{test}, a signature of very strong underfitting.

3.2.3. Dependence of optimal regularizations on system and data set sizes. The use of AIS and of ACE allows us to approximation the partition function of the model, and therefore to compute the Kullback–Leibler (KL) divergence of the inferred probability distribution from the ground-truth probability distribution,

$$D_{\text{KL}} = \sum_{\boldsymbol{x}} \frac{e^{-E(\boldsymbol{x};\boldsymbol{h}^*,\boldsymbol{J}^*)}}{Z(\boldsymbol{h}^*,\boldsymbol{J}^*)} \log \left[\frac{e^{-E(\boldsymbol{x};\boldsymbol{h}^*,\boldsymbol{J}^*)}}{Z(\boldsymbol{h}^*,\boldsymbol{J}^*)} \Big/ \frac{e^{-E(\boldsymbol{x};\boldsymbol{h}^{\text{tr}},\boldsymbol{J}^{\text{tr}})}}{Z(\boldsymbol{h}^{\text{tr}},\boldsymbol{J}^{\text{tr}})} \right], \tag{25}$$

We then determine the value of γ for which D_{KL} is minimal, as a function of the various parameters defining the model and the data. We show below that this alternative definition of the optimal regularization is quantitatively consistent with the definition of γ^{opt}.

J. Stat. Mech. (2022) 053502

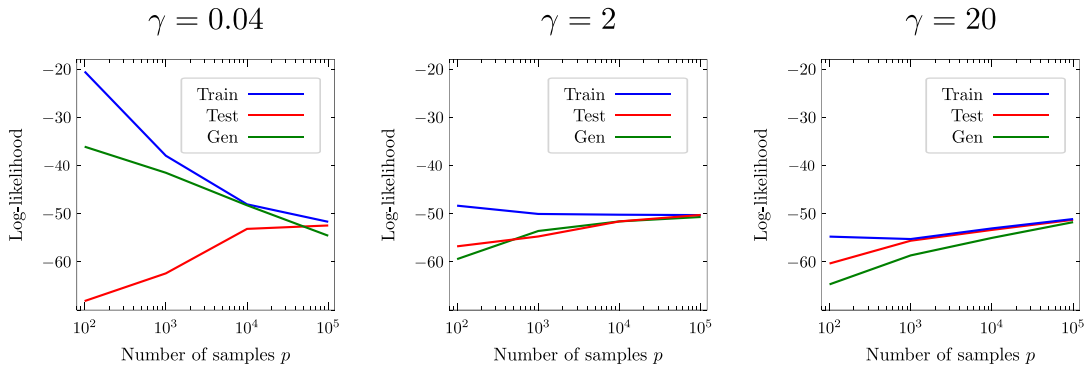

Figure 6. Average log-likelihoods of test, train and generated data (same colors as in previous figures) vs number p of samples for different regularization strengths γ (one panel for each value). Results were averaged over 20 000 sequences for each reported value of γ and p. Parameters: $n = 50$, $q = 10$. Results obtained with PLM.

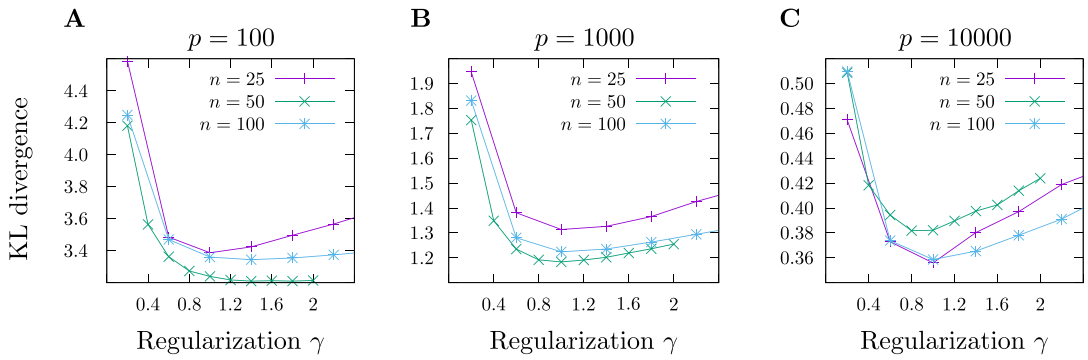

Figure 7. KL divergence between the inferred models and the ground truth for different graph (n) and sampling (p) sizes as a function of the regularization on the couplings (γ). The y-axis was arbitrarily rescaled between the different curves to allow for easier comparison. Parameters: $d = 2.5$, $q = 10$.

Dependence on the size n. We first study if and how the optimal regularization parameter γ changes when we the system size n is increased, while the average connectivity in the graph is fixed by choosing $p = 2.5/n$; we also fix the number of Potts states to $q = 10$. In figure 7 we show the KL divergence for models inferred at different γ for various n and p. The optimal regularization seems to be roughly equal to 0.5 in all the considered cases, independently of p (with some inaccuracy for very poor sampling, i.e. $p = 100$). We have also checked that this optimal value of γ does not seem to depend on q, by repeating the same numerical experiments for $q = 20$ Potts states with similar results, see figure 8.

These two results are in very good agreement with the theoretical prediction reported in equation (20), that is, $\gamma^{\rm opt} \simeq \gamma^{\rm cross} \simeq \frac{1}{d} = 0.4$ for the parameters chosen in figures 7 and 8. Indeed, in ER graphs, the average number of interacting neighbors is equal to d

J. Stat. Mech. (2022) 053502

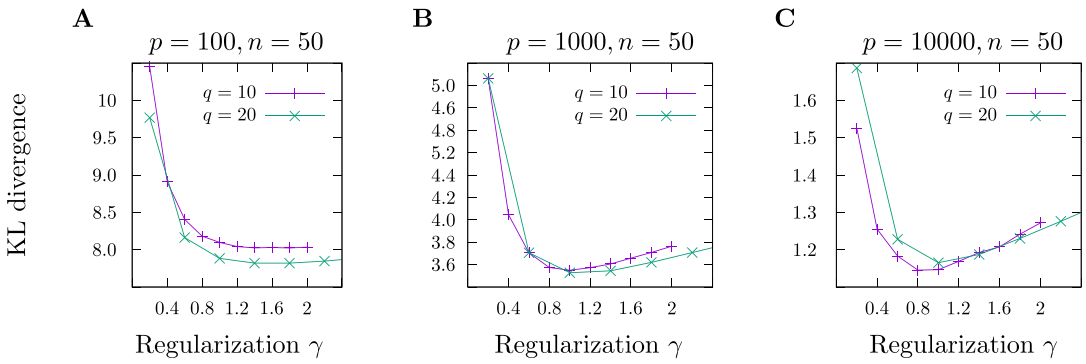

Figure 8. KL divergence between the inferred models and the ground truth for different numbers q of Potts states and p of data points, as a function of the regularization on the couplings (γ) and for diff. The y-axis was arbitrarily rescaled between the different curves to allow for easier comparison. Parameters: $d = 2.5$, $n = 50$.

(on average), independently of n (and p). In addition, since each variable can take one out of q symbol values, the number of variables j interacting with i in the sum at the denominator in equation (20) is independent of q.

Dependence on the structural connectivity of the interaction graph. We then study how the optimal regularization depends on the connectivity of the graph. For this reason we keep the graph size fixed ($n = 25$), and build different ER models with different densities varying d, see section 3.2.1. Once data are generated we infer the model parameters $\boldsymbol{h}^*, \boldsymbol{J}^*$ for different γ and sample sizes p. Results are reported in figure 9, and show a clear dependence on the structural parameter d. We observe that the scaling factor is approximately inversely proportional to the number of neighbors on the interacting graph. This result is in excellent agreement with the outcome of the expected theoretical scaling reported in equation (20).

4. Analytical calculations at low and high sampling ratios

While finding the exact value for the regularization strengths of interest as functions of the model parameters is out of reach we show in this section how this calculation can be done in the case of the Gaussian vectors model for very low and high values of the sampling ratios.

4.1. Asymptotic behavior of γ^{cross}

The crossing regularization γ^{cross} is defined through

$$L_{\text{test}}(\gamma^{\text{cross}}) = L_{\text{gen}}(\gamma^{\text{cross}}). \tag{26}$$

J. Stat. Mech. (2022) 053502

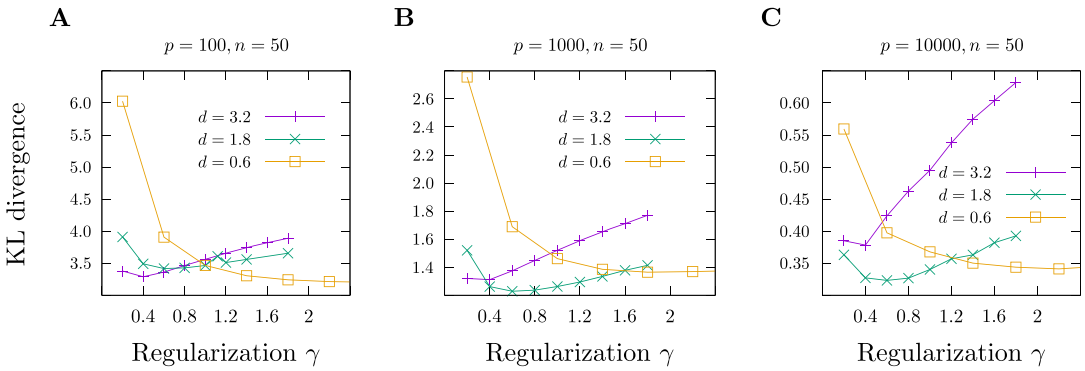

Figure 9. KL divergence between the inferred models and the ground truth for different average number of edge per site (numbers reported in the panels, obtained by varying d), as a function of the regularization (γ) used during inference. The y-axis was arbitrarily rescaled between the different curves to allow for easier comparison. Parameters: $n = 25$, $q = 10$.

Replacing L_{gen} in the equation above with its expression in equation (16) and using the definitions (13) and (14) of the train and test log-likelihoods we obtain

$$\gamma^{\text{cross}} = \alpha \, \frac{L_{\text{train}}(\gamma^{\text{cross}}) - L_{\text{test}}(\gamma^{\text{cross}})}{\sum_{i,j} {J^*_{ij}}^2} = \alpha \, \frac{\sum_{i,j} J^*_{ij} \left(C^{\text{emp}}_{ij} - C^{\text{tr}}_{ij}\right)}{\sum_{i,j} {J^*_{ij}}^2}. \tag{27}$$

4.1.1. $\alpha \to \infty$ regime. We derive below an asymptotic prediction for γ^{cross} in the large sampling regime $\alpha \to \infty$. We begin by considering the $\alpha \gg 1$ limit of the matrix-form MAP equation (8):

$$\boldsymbol{J}^* = \mu \boldsymbol{I} - (\boldsymbol{C}^{\text{emp}})^{-1}. \tag{28}$$

We consider the distribution of the empirical covariance matrix $\boldsymbol{C}^{\text{emp}}$ conditioned to the 'true' correlation matrix $\boldsymbol{C}^{\text{tr}} = (\mu^{\text{tr}} - \boldsymbol{J}^{\text{tr}})^{-1}$, known as the Wishart distribution (Wishart 1928), and defined for $p > n$ as

$$p_{\boldsymbol{J}^{\text{tr}}}(\boldsymbol{C}) \propto e^{n\frac{\alpha}{2}\mathcal{F}(\boldsymbol{C})}, \quad \mathcal{F}(\boldsymbol{C}) = \frac{\alpha - 1}{2} \log\det(\boldsymbol{C}) - \frac{\alpha}{2} \operatorname{Tr}\left((\mu^{\text{tr}} - \boldsymbol{J}^{\text{tr}})\boldsymbol{C}\right), \tag{29}$$

where we omit $\boldsymbol{C}$-independent normalization factor. For large α, we can perform a saddle-point approximation of this density around its maximum $\boldsymbol{C}^{\text{tr}}$:

$$p_{\boldsymbol{J}^{\text{tr}}}(\boldsymbol{C} = \boldsymbol{C}^{\text{tr}} + \Delta\boldsymbol{C}) \propto e^{n\frac{\alpha}{2}\Delta\boldsymbol{C}^\dagger \frac{\partial^2\mathcal{F}}{\partial\boldsymbol{C}\partial\boldsymbol{C}}(\boldsymbol{C}^{\text{tr}})\Delta\boldsymbol{C}}. \tag{30}$$

A straightforward calculation leads to

$$\frac{\partial^2\mathcal{F}}{\partial C_{i,j}\partial C_{a,b}}(C^{\text{tr}}) = \frac{\partial^2 \log\det \boldsymbol{C}}{\partial C_{i,j}\partial C_{a,b}}(\boldsymbol{C}^{\text{tr}}) = -\left(C^{\text{tr}}\right)^{-1}_{a,i}\left(C^{\text{tr}}\right)^{-1}_{b,j}. \tag{31}$$

We deduce from equation (30) that $\left(\boldsymbol{C}^{\text{tr}}\right)^{-1} \times \Delta\boldsymbol{C} = \boldsymbol{U}/\sqrt{n\alpha}$, where $\boldsymbol{U}$ is distributed as an uncorrelated Gaussian matrix, whose entries have zero means and unit standard

J. Stat. Mech. (2022) 053502

deviation. Therefore, using equation (28), we have

$$\boldsymbol{J}^* \simeq \mu \boldsymbol{I} - \left(\boldsymbol{C}^{\mathrm{tr}} + \Delta \boldsymbol{C}\right)^{-1} = \mu \boldsymbol{I} - \left(\boldsymbol{I} - \frac{\boldsymbol{U}}{\sqrt{\alpha n}}\right) \boldsymbol{C}^{\mathrm{tr}\,-1}. \tag{32}$$

This expression for the inferred coupling matrix can be inserted in equation (27) for γ^{cross}. Carrying out the averages over $\boldsymbol{U}$ appearing in $\boldsymbol{J}^*$ and $\boldsymbol{C}^{\mathrm{emp}}$ we obtain

$$\gamma^{\mathrm{cross}} = \frac{n}{\sum_{i,j}(J^*_{ij})^2} \overset{\alpha\to\infty}{\simeq} \frac{n}{\sum_{i,j}(J^{\mathrm{tr}}_{ij})^2}. \tag{33}$$

The stronger the interactions in our underlying model, the weaker the regularization that needs to be applied during inference. One way of intuitively understanding this statement is that stronger interactions will *a priori* generate samples (and therefore MAP estimates) with less undesirable variance, and therefore require less smoothing from the regularization.

4.1.2. $\alpha \to 0$ regime. We now consider the case of very poor sampling. The lowest value of the sampling ratio, $\alpha = \frac{1}{n}$, is reached with a single sample $\boldsymbol{s}$ ($p = 1$). The empirical covariance matrix is then easily written as

$$\boldsymbol{C}^{\mathrm{emp}} = \boldsymbol{s}\boldsymbol{s}^\dagger := n\, \boldsymbol{u}\boldsymbol{u}^\dagger. \tag{34}$$

One eigenvalue of $\boldsymbol{C}^{\mathrm{emp}}$ is non-zero, and is fixed to n to enforce the spherical constraint[5]. In other words, the normalized vector $\boldsymbol{u} = \boldsymbol{s}/\sqrt{n}$ is the unique non-zero eigenvector of $\boldsymbol{C}^{\mathrm{emp}}$.

Eigenvalues of $\boldsymbol{J}^*$. The inferred coupling matrix reads, according to equations (8) and (34),

$$\boldsymbol{J}^* = [j^*(n) - j^*(0)]\, \boldsymbol{u}\boldsymbol{u}^\dagger + j^*(0)\, \boldsymbol{I}, \tag{35}$$

where the eigenvalues $j^*(c^{\mathrm{emp}})$ are given by equation (10). Using $\alpha = \frac{1}{n}$ and expanding in powers of $\frac{1}{n}$, we find

$$j^*(0) = -\frac{1}{n\gamma\mu^*} + \frac{2}{n^2\gamma^2\mu^{*3}} + O(n^{-3}), \tag{36}$$

$$j^*(n) = \frac{1}{2\gamma}\left[1 + \gamma\mu^* - \sqrt{(\gamma\mu^* - 1)^2 + \frac{4\gamma}{n}}\right] + O(n^{-3}). \tag{37}$$

The latter expression can be divided into two cases, depending on whether $\gamma\mu$ is larger or smaller than 1:

$$j^*(n) = \begin{cases} \mu^* - \dfrac{1}{n(1-\gamma\mu^*)} + \dfrac{\gamma}{n^2(1-\gamma\mu^*)^3} + O(n^{-3}) & \text{if } \gamma\mu < 1 \\ \dfrac{1}{\gamma} - \dfrac{1}{n(\gamma\mu^* - 1)} + \dfrac{\gamma}{n^2(\gamma\mu^* - 1)^3} + O(n^{-3}) & \text{if } \gamma\mu > 1, \end{cases} \tag{38}$$

[5] In numerical experiments on finite size n, this constraint is enforced by hand, by rescaling the empirical covariance $\boldsymbol{C}^{\mathrm{emp}}$ to have a trace exactly equal to n. Note that, in the $n \to \infty$ limit and for $\sigma < 1$, this rescaling is not necessary. For σ larger than 1, however, the norm of $\boldsymbol{s}$ fluctuates strongly, as $|\boldsymbol{s}|^2$ follows a chi-square distribution.

J. Stat. Mech. (2022) 053502

which, together with the normalization condition

$$\frac{n-1}{\mu^* - j^*(0)} + \frac{1}{\mu^* - j^*(n)} = n, \tag{39}$$

yields that:

$$\mu^*(\gamma) = \begin{cases} \gamma^{-1/2} & \text{if } \gamma < 1 \\ 1 & \text{if } \gamma\mu^* > 1. \end{cases} \tag{40}$$

Expression for γ^{cross}. We then express the terms appearing in the expression of γ^{cross}, see equation (27), in terms of the eigenvalues $j^*(0), j^*(n)$:

$$\sum_{i,j} (J^*_{ij})^2 = j^*(n)^2 + (n-1)\, j^*(0)^2, \tag{41}$$

$$\sum_{i,j} J^*_{ij}\, C^{\text{emp}}_{ij} = n\, [j^*(n) - j^*(0)] + n\, j^*(0), \tag{42}$$

$$\sum_{i,j} J^*_{ij}\, C^{\text{tr}}_{ij} = n\, [j^*(n) - j^*(0)]\, \theta + n\, j^*(0), \tag{43}$$

where we introduced the matrix element

$$\theta = \frac{1}{n} \sum_{i,j} \boldsymbol{u}_i\, \boldsymbol{C}^{\text{tr}}_{ij}\, \boldsymbol{u}_j. \tag{44}$$

Let us consider this quantity in more details. On average over the sample $\boldsymbol{s}(= \sqrt{n}\boldsymbol{u})$, we have:

$$\langle \boldsymbol{u}_i \boldsymbol{u}_j \rangle = \frac{1}{n} \boldsymbol{C}^{\text{tr}}_{ij}, \tag{45}$$

and thus

$$\langle \theta \rangle = \frac{1}{n^2} \sum_{i,j} \boldsymbol{C}^{\text{tr}^2}_{ij} = \frac{1}{n^2} \sum_{k} c^{\text{tr}^2}_k. \tag{46}$$

Generally, due to the constraint that $\sum_k c^{\text{tr}}_k = n$, we find that $\langle\theta\rangle$ is bounded from below by $1/n$ (when all eigenvalues of $\boldsymbol{C}^{\text{tr}}$ are equal to 1), and from above by 1 (when a single eigenvalue of $\boldsymbol{C}^{\text{tr}}$ is equal to n, and all the other eigenvalues are equal to 0). It should be noted that, while the result $\langle\theta\rangle > 1/n$ is true on average, the value of θ for an individual sample can be arbitrarily close to 0. This possibility will be discussed below.

Analytical expressions for the average value of θ can be obtained in the case of random quenched interactions considered in section 3.1.1 by explicitly integrating over the semi-circle eigenvalue distribution, with the results

J. Stat. Mech. (2022) 053502

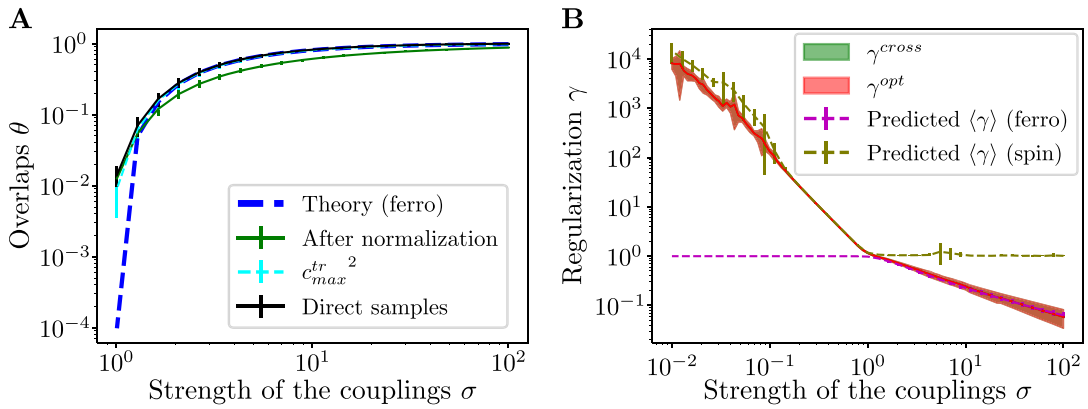

Figure 10. Properties of the inference in the low sampling regime $\alpha = 1/n$ and for the random quenched coupling model. (A) Value of the overlap θ as a function of the scale σ of the interactions in the ferromagnetic regime $\sigma > 1$, see theoretical prediction for $\langle\theta\rangle$ for non rescaled samples in equation (47). For normalized samples the overlap converges toward 1 more slowly. (B) Comparison between the values of $\gamma^{\rm opt}$ and $\gamma^{\rm cross}$ found numerically and the predictions in equations (48) and (49), applied to the empirical distribution of 'rescaled samples' overlaps. In both panels error bars represent the variations across ten choices of the true underlying interaction matrix of the θ, and γ is averaged over 100 random draws from the Gaussian model distribution.

$$\langle\theta\rangle = \begin{cases} \dfrac{1}{n(1-\sigma^2)} & \text{if } \sigma < \sigma_{\rm c} = 1, \\ \left(1 - \dfrac{1}{\sigma}\right)^2 & \text{if } \sigma > \sigma_{\rm c}, \end{cases} \tag{47}$$

see figure 10(A). The last equation comes from the fact that, when σ is larger than 1, the sum in equation (46) is dominated by the single macroscopic eigenvalue of $\boldsymbol{C}^{\rm tr}$, see equation (19). For the rescaled samples we used in practice, the average value of θ still goes to 1 as σ increases, but with a gap closer to $\sim\sigma^{-1/2}$.

Let us now summarize the different cases that can be met, see figure 10(B):

- If σ is below 1, the system is in a disordered phase, and strong regularization is needed. We find that
  - If $\theta > 1/n$, the crossing regularization is

$$\gamma^{\rm cross} = \frac{n\theta}{n\theta - 1}, \tag{48}$$

    which is larger than 1. This corresponds to a situation where the sample is slightly informative, and strong regularization is necessary to avoid overfitting.
  - If $\theta < 1/n$, the two likelihoods never cross, and the optimal regularization appears to be infinite. This corresponds to a situation in which the randomly drawn sample is counter-informative, so that the null answer is better than taking it into account.

J. Stat. Mech. (2022) 053502

- If σ is above 1, the system is in the ferromagnetic phase, so that a single sample conveys significant information about the entire distribution. In that case, we find that for a given (rescaled) sample the crossing regularization is given by

$$\gamma^{\text{cross}} = (1-\theta)^2, \tag{49}$$

which vanishes when $\sigma \to \infty$.

It should be noted that the achieved likelihoods are significantly higher in the ferromagnetic case: since the distribution lives in a single dimension, a single sample is enough to get meaningful information about the entire distribution, a phenomenon related to benign overfitting (Bartlett *et* al 2020). In all cases where σ is either very small or very large, the optimal regularization varies strongly from sample to sample.

4.2. Asymptotic behavior of γ^{opt} for $\alpha \to 0$

While the $\alpha \to \infty$ limit of the optimal regularization is hard to obtain (in particular, because the test likelihood's derivative with respect to γ vanishes uniformly), the computation of γ^{opt} can be carried out in the low ratio regime, $\alpha = 1/n$.

We start from the definition of γ^{opt}:

$$\frac{\partial L^{\text{test}}}{\partial \gamma}(\gamma^{\text{opt}}) = 0 = \frac{\partial}{\partial \gamma}\left[\frac{1}{2}\sum_{i,j} J^*_{ij} C^{\text{tr}}_{ij} - \log\, Z(\boldsymbol{J}^*)\right]. \tag{50}$$

From equations (7) and (35), we have

$$\log\, Z(\boldsymbol{J}^*) = \frac{n}{2}\mu^* - \frac{1}{2}\left[(n-1)\log(\mu^* - j^*(0)) + \log(\mu^* - j^*(n))\right], \tag{51}$$

so that

$$\frac{\partial \log\, Z(\boldsymbol{J}^*)}{\partial \gamma} = \frac{n-1}{2}\frac{\partial_\gamma j^*(0)}{\mu^* - j^*(0)} + \frac{1}{2}\frac{\partial_\gamma j^*(n)}{\mu^* - j^*(n)}. \tag{52}$$

In addition, differentiating equation (43) we get

$$\frac{\partial}{\partial \gamma}\sum_{i,j} J^*_{ij} C^{\text{tr}}_{ij} = n\left[\frac{\partial j^*(n)}{\partial \gamma} - \frac{\partial j^*(0)}{\partial \gamma}\right]\theta + n\frac{\partial j^*(0)}{\partial \gamma.} \tag{53}$$

We now need to evaluate the derivatives $\partial_\gamma j^*(0)$ and $\partial_\gamma j^*(n)$. From equations (36) and (40), at the first order in n, we have

$$\frac{\partial j^*(0)}{\partial \gamma} = \frac{1}{n\gamma^2\mu^*} + \frac{\partial_\gamma \mu^*}{n\gamma\mu^{*2}} = \begin{cases} \dfrac{1}{n\gamma^2} & \text{if } \gamma > 1, \\ \dfrac{1}{2n\gamma^{3/2}} & \text{if } \gamma < 1. \end{cases} \tag{54}$$

J. Stat. Mech. (2022) 053502

Similarly, equations (37) and (40) yield

$$\frac{\partial j^*(n)}{\partial \gamma} = \begin{cases} -\dfrac{1}{\gamma^2} & \text{if } \gamma > 1\ , \qquad (55a) \\ -\dfrac{1}{2\gamma^{3/2}} & \text{if } \gamma < 1\ . \qquad (55b) \end{cases}$$

We may now conclude our calculation of γ^{opt}:

- If $\gamma > 1$,

$$\frac{\partial L^{\text{test}}}{\partial \gamma} = \frac{1}{2\gamma^2}(1 - n\theta) + \frac{1}{2\gamma^2(\gamma - 1)}, \tag{56}$$

and therefore this derivative vanishes for

$$\gamma^{\text{opt}} = \frac{n\theta}{n\theta - 1}, \tag{57}$$

which is the same result as found from the γ^{cross} computation in equation (48).

- If $\gamma < 1$,

$$\frac{\partial L^{\text{test}}}{\partial \gamma} = \frac{n}{4\gamma^{3/2}}\left[(1 - \theta) - \sqrt{\gamma}\right], \tag{58}$$

whose root is given by

$$\gamma^{\text{opt}} = (1 - \theta)^2, \tag{59}$$

in full agreement with the result shown in equation (49).

Therefore, the analytical expressions of γ^{opt} and γ^{cross} coincide in the undersampled regime (single sample), which provides further support to our conjecture that the values of those two regularizations are equal or very close, as suggested by numerical experiments. Unfortunately, the computation of γ^{opt} in the oversampled regime ($\alpha \to \infty$) is more complicated, and we were not able to prove that its value converges to the limit found for γ^{cross} in equation (33).

5. Conclusion

In this work we provided both analytical and numerical evidence for the optimal value of a L_2 penalty term in the likelihood used for MAP inference of graphical models. In addition to showing that a non-zero optimal regularization always exists, we find a remarkable empirical coincidence between two optimality criteria: the maximization of the test log-likelihood, and the condition that test and generated likelihoods are equal, a natural requirement for a generative model, see figure 1. This equality suggests that, while weaker regularizations might give the impression of higher quality generated

J. Stat. Mech. (2022) 053502

data (through higher generated likelihoods), stronger regularizations should actually be employed to achieve the best possible model, and the perceived increase in generated likelihood is actually a form of overfitting.

Analytical expressions for the crossing and optimal regularizations could be obtained in the limiting regimes of poor or good sampling. In the latter case, we obtain an explicit expression for the optimal regularization strengths in terms of the average inverse squared couplings between the variables, see equation (20). This prediction remains remarkably accurate over a wide range of parameter value, and even for case of categorical variables (Potts model), while it was established analytically in the case of the Gaussian multivariate model. This result suggest that our study could also be applied to other interesting classes of models, such as restricted Boltzmann machines, an extension of Ising/Potts models in which multi-body interactions can be introduced. More generally, it has been known for a long time that neural networks benefit from regularization, with extensive research being led on the exact regularization scheme to apply for different tasks, see for example (Wan *et* al 2013, Zaremba *et al* 2015, Louizos *et al* 2018, Haarnoja *et al* 2018, Bartlett *et al* 2021).

Most approaches exhibit some form of 'bias-variance trade off', i.e. a phenomenon in which increasing the strength of the regularization reduces the variance of the estimator (e.g. by increasing the smoothness of the solutions) but biases the inference toward a particular subset of solutions. As a result an optimal value of the regularization exists that balances those two effects, similarly to what we observed in our model. It should however be noted that this simplistic picture might not hold in all circumstances, as suggested by a number of findings recently reviewed by Dar *et* a*l* (2021): in some settings, similarly to our low-sampling limit case with a 'counter-informative' sample, the optimal regularization is infinite (Mignacco *et al* 2020, Loureiro *et al* 2021); in other cases, the optimal regularization can be found to be zero (Hastie *et al* 2020) in the infinitely overparametrized regime. This is to be related to the fact that, contrary to the usual intuition that an increase in number of parameters leads to an increase in variance of the inferred model, some models show an opposite trend (Gerace *et* a*l* 2021).

In terms of modeling protein from sequence data our results suggest that the optimal γ should neither be proportional to $\frac{p}{n}$ nor to q, as proposed in previous works (Ekeberg *et al* 2014, Hopf *et al* 2017), but is related to the inverse sum of the squared couplings incoming onto residues, see equation (20). In particular, our prediction is that the optimal value for γ scales inversely proportional to the number of interacting neighbors on the dependency graph. However, some caution must be brought to this conclusion. The sample size p is not clearly defined for real proteins. The presence of phylogenetic correlations between sequences make the assumption of independent data points only approximate at best. In practice the choices $\gamma = 0.01\frac{p}{n}$ (Ekeberg *et al* 2014) and $\gamma = 0.01q$ (Hopf *et al* 2017) are qualitatively similar when the number of sequences exceed the protein length by a factor 20, which is not unreasonable for a substantial number of protein families.

Our work could be extended in several directions. First, we here focused on MAP inference only, but it would be interesting to analyze what happens within Bayesian inference, i.e. to investigate the effects of fluctuations around the MAP estimator. As a

J. Stat. Mech. (2022) 053502

first step in this direction we show in appendix B that introducing a finite, but small, temperature in the inference procedure yields similar results to what was observed here. Second, the question of what would happen in more complex cases in which the energy landscape is non-convex (e.g. a mixture of Gaussian, or a multi-modal Potts model) remains open, and further investigations would be necessary to understand how regularization might influence the known trade-off between better fit of individual modes of the data (low regularizations) and easier transitions between different modes, which could happen by flattening the energy landscape in-between the modes.

Acknowledgments

We are grateful to J Tubiana for providing us with his code for Annealed Importance Sampling.

Appendix A. Numerical estimation of the regularization strengths

In order to compute the values of γ^{opt} and γ^{cross} as precisely as possible, we derived two *residuals*, i.e. functions of γ which are equal to 0 respectively when the test likelihood is optimal, or when the test and generated likelihoods are equal. Similarly to how μ^* was determined when solving the MAP equation, the roots of those residuals will be minimized using standard convex optimization routines to obtain high precision estimates of the optimal and crossing regularizations.

This approach is easily illustrated in the case of the crossing regularization γ^{cross}. According to equation (27) the following function $\text{Res}^{\text{cross}}(\gamma)$ has its root equal to γ^{cross}:

$$\text{Res}^{\text{cross}}(\gamma) := \alpha \frac{\langle \boldsymbol{J}^*(\gamma)\,(\boldsymbol{C}^{\text{emp}} - \boldsymbol{C}^{\text{tr}})\rangle}{\langle \boldsymbol{J}^*(\gamma)^2\rangle} - \gamma. \tag{60}$$

For the estimation of the optimal regularization, the computation is more involved and relies on finding the derivative of L_{test} with respect to γ. Indeed, γ is equal to γ^{opt} when

$$\text{Res}^{\text{opt}}(\gamma) := \frac{\partial L_{\text{test}}}{\partial \gamma}, \tag{61}$$

is equal to 0.

This derivative can be computed as:

$$\begin{aligned}\frac{\partial L_{\text{test}}}{\partial \gamma} &= \frac{1}{2}\sum_{i,j} \frac{\partial J^*_{ij}}{\partial \gamma} C^{\text{tr}}_{ij} - \frac{\partial \log Z(\boldsymbol{J}^*)}{\partial \gamma} \\ &= \frac{1}{2}\sum_{k} \frac{\partial j^*_{k}}{\partial \gamma} C^{\text{tr,rot}}_{k,k} - \frac{\partial \log Z(\boldsymbol{J}^*)}{\partial \gamma},\end{aligned} \tag{62}$$

where $\boldsymbol{C}^{\text{tr,rot}}$ is the true correlation matrix after changing the basis to the inference basis in which $\boldsymbol{C}^{\text{emp}}$ is diagonal.

J. Stat. Mech. (2022) 053502

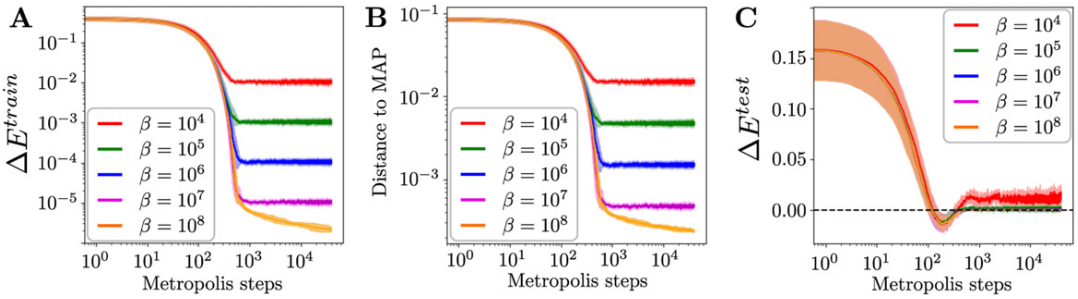

Figure 11. Evolution of the train energy, distance to MAP estimator and test energy as a function of the number of Metropolis steps for different values of the temperature. The energies are given relative to the ones of the MAP. For low temperatures and long enough times, the sampled solutions have very close energies to the MAP estimator. At intermediate times, the test energy of the sampled solutions can get lower than the one of the MAP. Higher temperatures allow the system to stay in states of higher energy, which are further from the MAP. Figure obtained with $n = 20$, $\alpha = 5$, $\sigma = 0.5$, $\gamma = 5$ (larger than the optimal regularization $\gamma^{\text{opt}} = 1/\sigma^2 = 4$).

We begin by computing

$$\begin{aligned} \partial_\gamma j_k^* &= \partial_\gamma \left[\frac{1}{2\gamma} \alpha c_k + \gamma \mu^* - D_k \right] \\ &= A_k \partial_\gamma \mu^* + B_k - \frac{j_k^*}{\gamma}, \end{aligned} \tag{63}$$

where we introduced

$$D_k = \sqrt{(\alpha c_k^{\text{emp}} - \gamma \mu^*)^2 + 4\alpha\gamma}, \tag{64}$$

$$A_k = \frac{1}{2}\left(1 - \frac{\gamma\mu^* - \alpha c_k^{\text{emp}}}{D_k}\right), \tag{65}$$

$$B_k = \frac{1}{\gamma}\left(\mu^* A_k - \frac{\alpha}{D_k}\right). \tag{66}$$

Then, we have that

$$\partial_\gamma \log Z = n\partial_\gamma \mu^* - \frac{1}{2}\sum_k \frac{\partial_\gamma \mu^* - \partial_\gamma j_k^*}{\mu^* - j_k^*}. \tag{67}$$

Finally, we can compute $\partial_\gamma \mu$ by first noting that:

$$\frac{1}{2}\sum_k \frac{1}{\mu^* - j_k^*} = 1, \tag{68}$$

J. Stat. Mech. (2022) 053502

hence

$$\sum_k \frac{\partial_\gamma \mu^* - \partial_\gamma j_k^*}{(\mu^* - j_k^*)^2} = 0, \tag{69}$$

and therefore

$$\partial_\gamma \mu = \left[\sum_k \frac{\partial_\gamma j_k^*}{(\mu^* - j_k^*)^2}\right] / \left[\sum_k \frac{1}{(\mu^* - j_k^*)^2}\right], \tag{70}$$

$$\partial_\gamma \mu^* = \left[\sum_k \frac{A_k \partial_\gamma \mu^* + B_k - j_k^*/\gamma}{(\mu^* - j_k^*)^2}\right] / \left[\sum_k \frac{1}{(\mu^* - j_k^*)^2}\right], \tag{71}$$

$$\partial_\gamma \mu^* \left[\sum_k \frac{1 - A_k}{(\mu^* - j_k^*)^2}\right] = \left[\sum_k \frac{B_k - j_k^* \gamma}{(\mu^* - j_k^*)^2}\right], \tag{72}$$

which finally yields:

$$\partial_\gamma \mu^* = \left[\sum_k \frac{B_k - j_k^*/\gamma}{(\mu^* - j_k^*)^2}\right] / \left[\sum_k \frac{1 - A_k}{(\mu^* - j_k^*)^2}\right]. \tag{73}$$

Putting together equations (62) to (73) yields an explicit expression for the derivative of L_{test} with respect to γ, which is exactly the residual $\text{Res}^{\text{opt}}(\gamma)$ whose root gives the value of γ^{opt}.

Appendix B. Non-zero temperature inference

It is natural to wonder whether our result hold for when sampling the posterior probability at inverse temperature β:

$$p_\beta(\boldsymbol{J}) \propto e^{-\beta\left[\frac{\gamma}{4}\ \text{Tr}(\boldsymbol{J}^2) - \frac{\alpha}{2}\ \text{Tr}(\boldsymbol{J}\boldsymbol{C}^{\text{emp}}) + \alpha\ \log\ Z(\boldsymbol{J})\right]}. \tag{74}$$

While an in-depth study of the different sampling strategies is out of the scope of this work (see Rubinstein and Kroese (2016) for a general overview), we report below numerical and analytical preliminary steps aiming at characterizing this posterior distributions.

We performed some preliminary experiments using a simple Metropolis–Hastings algorithm (Metropolis and Ulam 1949) which consists in starting from a random point in the distribution, proposing a small modification and accepting it with probability $p = \min(1, \exp(-\beta \Delta E))$ depending on the associated change in energy, ΔE. In our case, we start from a symmetric Gaussian matrix in which all the entries above the diagonal are independent and have the same mean and variance as the MAP estimator[6], and the modifications we propose are the addition of small amplitude, sparse, Gaussian matrices.

[6] This initial choice only affects convergence time, as the Metropolis sampling procedure loses information on the initial conditions after a transient regime.

J. Stat. Mech. (2022) 053502

Since increasing the temperature (hence decreasing β) can be seen as a way of letting the system explore areas of higher energy, the matrices sampled at higher temperatures will be further away from the MAP solution, which we illustrate in figures 11(A) and (B) respectively. While the energy used for sampling is computed using the empirical covariance matrix $\boldsymbol{C}^{\text{emp}}$, it is also interesting to consider the evolution of a 'test' energy, computed using the true covariance matrix $\boldsymbol{C}^{\text{tr}}$, which will help quantify the generalization property of these solutions. While at long time scales the test energy converges to a value very close to the one of the MAP estimator, there exists an intermediate regime in which the sampled matrices achieve better test energy than the MAP estimator, as seen in figure 11(C). Notice, however, that the values of the inverse temperature β considered in the simulations are large compared to the canonical inverse temperature, n, defined in the posterior probability over $\boldsymbol{J}$, see equation (5). The results reported above therefore confirm that weak fluctuations of the posterior do not modify the properties of the MAP estimator.

References

Anderson P W 1958 Absence of diffusion in certain random lattices *Phys. Rev.* **109** 1492–505

Barrat-Charlaix P, Muntoni A P, Shimagaki K, Weigt M and Zamponi F 2021 Sparse generative modeling via parameter-reduction of Boltzmann machines: application to protein-sequence families *Phys. Rev.* E **104** 024407

Bartlett P L, Long P M, Lugosi G and Tsigler A 2020 Benign overfitting in linear regression (arXiv:1906.11300)

Bartlett P L, Montanari A and Rakhlin A 2021 Deep learning: a statistical viewpoint (arXiv:2103.09177)

Barton J P, Cocco S, De Leonardis E and Monasson R 2014 Large pseudo-counts and L_2-norm penalties are necessary for the mean-field inference of Ising and Potts models *Phys. Rev.* E **90** 012132

Barton J P, De Leonardis E, Coucke A and Cocco S 2016 ACE: adaptive cluster expansion for maximum entropy graphical model inference *Bioinformatics* **32** 3089–97

Bourgade P 2018 Random band matrices (arXiv:1807.03031)

Brent R P 2013 *Algorithms for Minimization without Derivatives* (North Chelmsford, MA: Courier Corporation)

Casati G, Molinari L and Izrailev F 1990 Scaling properties of band random matrices *Phys. Rev. Lett.* **64** 1851–4

Chau Nguyen H, Zecchina R and Berg J 2017 Inverse statistical problems: from the inverse Ising problem to data science *Adv. Phys.* **66** 197–261

Cocco S and Monasson R 2011 Adaptive cluster expansion for inferring Boltzmann machines with noisy data *Phys. Rev. Lett.* **106** 090601

Cocco S, Feinauer C, Figliuzzi M, Monasson R and Weigt M 2018 Inverse statistical physics of protein sequences: a key issues review *Rep. Prog. Phys.* **81** 032601

Dar Y, Muthukumar V and Baraniuk R G 2021 A farewell to the bias-variance tradeoff? An overview of the theory of overparameterized machine learning (arXiv:2109.02355)

Ekeberg M, Lövkvist C, Lan Y, Martin W and Aurell E 2013 Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models *Phys. Rev.* E **87** 012707

Ekeberg M, Hartonen T and Aurell E 2014 Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences *J. Comput. Phys.* **276** 341–56

Friedman J, Hastie T and Tibshirani R 2008 Sparse inverse covariance estimation with the graphical lasso *Biostatistics* **9** 432–41

Gerace F, Loureiro B, Krzakala F, Mézard M and Zdeborová L 2021 Generalisation error in learning with random features and the hidden manifold model *J. Stat. Mech.* 124013

Haarnoja T, Zhou A, Abbeel P and Levine S 2018 Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor *Int. Conf. Machine Learning PMLR* pp 1861–70

Haldane A and Levy R M 2019 Influence of multiple-sequence-alignment depth on Potts statistical models of protein covariation *Phys. Rev.* E **99** 032405

Hastie T, Montanari A, Rosset S and Tibshirani R J 2020 Surprises in high-dimensional ridgeless least squares interpolation (arXiv:1903.08560)

Hopf T A, Ingraham J B, Poelwijk F J, Schärfe C P I, Springer M, Sander C and Marks D S 2017 Mutation effects predicted from sequence co-variation *Nat. Biotechnol.* **35** 128–35

Huang J Z, Liu N, Pourahmadi M and Liu L 2006 Covariance matrix selection and estimation via penalised normal likelihood *Biometrika* **93** 85–98

J. Stat. Mech. (2022) 053502

Karoui N E 2008 Spectrum estimation for large dimensional covariance matrices using random matrix theory *Ann. Stat.* **36** 2757–90
Kosterlitz J M, Thouless D J and Jones R C 1976 Spherical model of a spin-glass *Phys. Rev. Lett.* **36** 1217–20
Ledoit O and Wolf M 2004 A well-conditioned estimator for large-dimensional covariance matrices *J. Multivariate Anal.* **88** 365–411
Louizos C, Welling M and Kingma D P 2018 Learning sparse neural networks through L_0 regularization (arXiv:1712.01312)
Loureiro B, Sicuro G, Gerbelot C, Pacco A, Krzakala F and Zdeborová L 2021 Learning Gaussian mixtures with generalised linear models: precise asymptotics in high-dimensions (arXiv:2106.03791)
MacKay D J C 2003 *Information Theory, Inference and Learning Algorithms* (Cambridge: Cambridge University Press)
Metropolis N and Ulam S 1949 The Monte Carlo method *J. Am. Stat. Assoc.* **44** 335–41
Mignacco F, Krzakala F, Lu Y M and Zdeborová L 2020 The role of regularization in classification of high-dimensional noisy Gaussian mixture (arXiv:2002.11544)
Pedregosa F *et al* 2011 Scikit-learn: machine learning in Python *J. Mach. Learn. Res.* **12** 2825–30
Ravikumar P, Wainwright M J and Lafferty J D 2010 High-dimensional Ising model selection using ℓ_1-regularized logistic regression *Ann. Stat.* **38** 1287–319
Ravikumar P, Wainwright M J, Raskutti G and Yu B 2011 High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence *Electron. J. Stat.* **5** 935–80
Rizzato F, Coucke A, de Leonardis E, Barton J P, Tubiana J, Monasson R and Cocco S 2020 Inference of compressed Potts graphical models *Phys. Rev.* E **101** 012309
Rubinstein R Y and Kroese D P 2016 *Simulation and the Monte Carlo Method* vol 10 (New York: Wiley)
Schmidt J, Marques M R G, Botti S and Marques M A L 2019 Recent advances and applications of machine learning in solid-state materials science *npj Comput. Mater.* **5** 1–36
Virtanen P *et al* 2020 SciPy 1.0: fundamental algorithms for scientific computing in Python *Nat. Methods* **17** 261–72
Wan L, Zeiler M, Zhang S, Le Cun Y and Fergus R 2013 Regularization of neural networks using DropConnect *Int. Conf. Machine Learning PMLR* pp 1058–66
Wishart J 1928 The generalised product moment distribution in samples from a normal multivariate population *Biometrika* **20A** 32–52
Wright S J 2015 Coordinate descent algorithms *Math. Program.* **151** 3–34
Zaremba W, Sutskever I and Vinyals O 2015 Recurrent neural network regularization. (arXiv:1409.2329)

J. Stat. Mech. (2022) 053502