Learning maximum entropy models from finite-size data sets: A fast data-driven algorithm allows sampling from the posterior distribution

Ulisse Ferrari

Learning Maximum Entropy Models from finite size datasets: a fast Data-Driven algorithm allows sampling from the posterior distribution

Ulisse Ferrari

Sorbonne Universités, UPMC Univ Paris 06, INSERM, CNRS,
Institut de la Vision, 17 rue Moreau, 75012 Paris, France.

Maximum entropy models provide the least constrained probability distributions that reproduce statistical properties of experimental datasets. In this work we characterize the learning dynamics that maximizes the log-likelihood in the case of large but finite datasets. We first show how the steepest descent dynamics is not optimal as it is slowed down by the inhomogeneous curvature of the model parameters space. We then provide a way for rectifying this space which relies only on dataset properties and does not require large computational efforts. We conclude by solving the long-time limit of the parameters dynamics including the randomness generated by the systematic use of Gibbs sampling. In this stochastic framework, rather than converging to a fixed point, the dynamics reaches a stationary distribution, which for the rectified dynamics reproduces the posterior distribution of the parameters.

We sum up all these insights in a "rectified" *Data-Driven* algorithm that is fast and by sampling from the parameters posterior avoids both under- and over-fitting along all the directions of the parameters space. Through the learning of pairwise Ising models from the recording of a large population of retina neurons, we show how our algorithm outperforms the steepest descent method.

Nowadays scientists from many different disciplines face the problem of understanding and characterizing the behavior of large multi-units complex systems with strong correlations [1–8]. Statistical Inference tackles these problems by inferring parameters of a chosen, context inspired, probability distributions to obtain models reproducing the system behavior. The basic strategy consists in choosing a model family described by a set of parameters and tune them to reproduce the dataset properties.

However, in many cases, due to a substantial unawareness of the system properties or to avoid biasing the results, hypotheses on the distribution functional form cannot be suggested nor trusted. To overcome this issue, the Maximum Entropy (MaxEnt) Principle[9] suggests to search for the probability distribution with the largest entropy between those satisfying a set of constraints, which force the model distribution to reproduce the experimental averages of a list of observables. For systems of binary variables a common and fruitful choice is to constraint the model distribution to reproduce the experimental single and pairwise correlations. This choice has been successfully applied in system neuroscience [3, 10–13], gene regulation [14], fitness estimation [15, 16] and many others fields of science. Moreover other possibilities with different observable lists have been also investigated [17, 18], suggesting that depending on the context a careful choice of the observables can improve the inference accuracy and predicting power.

However, once the MaxEnt problem is posed as an inference task, finding its solution can be very hard. For large system size, in fact, the inference problem cannot be solved analytically and specifically devoted algorithms are required. The most known and widely used algorithm [19] was introduced in the eighties and later developed [20, 21]. Other approaches include Selective Cluster Expansions [22, 23], Minimum Probability Flow [24] and several approximation schemes [25–29].

In this paper we develop further the approach of [19]. Our results are based on an analysis of the geometrical structure of the model parameters space. For the MaxEnt models inference this space is shown to benefit of peculiar properties, which allow us to introduce a novel *quasi-Newton* method, the *Data-Driven* algorithm, and to completely characterize its long-time learning dynamics. As it is affected by the randomness of Monta Carlo estimates, the parameters dynamics is stochastic and eventually converges to a stationary distribution around the log-likelihood maximum. The presented method takes advantage of this randomness and shapes the stationary distribution to reproduce the Bayesian posterior distribution and thus to sample from it. This last feature allows to avoid overfitting and endorses our algorithm to be well suited for dataset with highly inhomogeneous noise. We conclude with a test on biological data.

I. MAXIMUM ENTROPY MODELS

Systems of interest for MaxEnt approach are composed by N units that show a stochastic and coupled behavior. In order to be concrete, in this work we focus on binary, 0/1, units, but most of the results can be directly generalized to multiple-state variables, as Potts or Poisson models, or even continuous ones. In this framework, datasets are composed by B independent measurements of the synchronous state of the N system units: $\{\{\sigma_i(b)\}_{i=1}^{N}\}_{b=1}^{B}$. As an example, for a binned spike trains recording of several neurons, $\sigma_i(b) = 1$ may represent the activity (spike) of the i-th neuron in the b-th time-bin and $\sigma_i(b) = 0$ its silence.

The first, crucial, step in the MaxEnt analysis is the choice of a set of observables. Observables are generic

functions of the system units that should be chosen in order to catch the way system components interact. Strictly speaking, if some statistical feature is relevant for the system behavior, the corresponding observable should be included in order to force the model to reproduce that feature. On the other way round, if some feature is not essential, by excluding the corresponding observable, the model will adapt its behavior consistently with the other imposed features. Technically speaking the observables will be the sufficient statistics of the model probability distribution. As an example, the most common choice in the literature considers all the single variable terms (σ_i) and pairwise products ($\sigma_i\sigma_j$) and it leads to the construction of the well know Disorder Pairwise Ising model [3]. This particular choice allows to take into account the pairwise correlations between the units and allows the model to adapt higher order statistics consistently.

Moreover, a carefully choice of the observable should take into account the quality of the dataset. Noisy observables for which the average is not significantly estimated will induce data overfitting with the risk of strongly reducing the model prediction power. From this point of view and overloading the jargon of Bayesian Inference, the observables choice represents a sort of prior term (see sect. I B for details).

For the sake of generality, we like to present the results for an arbitrary observables choice and with this aim in mind we introduce a generic vector of observables $\mathbf{\Sigma}(\sigma) \equiv \{\Sigma_a(\sigma)\}_{a=1}^D$, functions of the system units. The choice of the observables vector completely determines the functional form of the MaxEnt model family [9, 22]. In fact, by searching for the probability distribution which has the largest entropy among those that reproduce the observables averages, we obtain (see, for example [23], for the whole functional calculation):

$$P_{\mathbf{X}}(\sigma) = \exp\{\mathbf{\Sigma}(\sigma) \cdot \mathbf{X}\}/Z[\mathbf{X}] \qquad (1)$$

where $Z[\mathbf{X}]$ is a normalization constant, $\mathbf{X}$ is a D dimensional fields vector, namely the model parameters, conjugated to the observable vector $\mathbf{\Sigma}$ and $\mathbf{\Sigma}(\sigma) \cdot \mathbf{X} = \sum_a \Sigma_a(\sigma)X_a$ is the scalar product in Euclidean space[30].

As the distribution (1) measures the probability of each possible system configurations, we can compute its dataset (log-)likelihood:

$$l\big(\text{ data } | \mathbf{X} \big) \equiv \sum_{b=1}^B \ln P_{\mathbf{X}}\big(\sigma(b)\big). \qquad (2)$$

Even if later we will consider posterior sampling, see sect. I B, for the moment we restrict the inference task to the search for the set of fields $\mathbf{X}^*$ that maximizes the log-likelihood:

$$\mathbf{X}^* \equiv \arg\max_{\mathbf{X}} \Big[\, l[\mathbf{X}] \,\Big] \qquad (3)$$

where

$$l[\mathbf{X}] = B\Big(\mathbf{X} \cdot \overline{\mathbf{P}} - \ln Z[\mathbf{X}] \Big), \qquad (4)$$

and

$$\overline{\mathbf{P}} \equiv \overline{\mathbf{\Sigma}} \equiv \frac{1}{B}\sum_{b=1}^B \mathbf{\Sigma}(\sigma(b)\,) \qquad (5)$$

are the experimental averages of the observables. In fact, finding the fields values $\mathbf{X}^*$ solving

$$\nabla_a l[\mathbf{X}^*] = B\Big(\overline{P_a} - Q_a[\mathbf{X}^*]\Big) = 0 \;, \qquad (6)$$

where $\nabla_a \equiv d/dX_a$, is equivalent to enforce the constraints:

$$\mathbf{Q}[\mathbf{X}^*] = \overline{\mathbf{P}} \qquad (7)$$

where

$$\mathbf{Q}[\mathbf{X}] \equiv \langle\mathbf{\Sigma}\rangle_{\mathbf{X}} \equiv \mathrm{Tr}_\sigma\big[\mathbf{\Sigma}(\sigma)P_{\mathbf{X}}(\sigma)\big] \;. \qquad (8)$$

are the model averages of the observables. Here and in the following $\langle\ldots\rangle_{\mathbf{X}}$ means average over the model distribution (1) with fields $\mathbf{X}$ and $\overline{(\ldots)}$ always refers to dataset averages. Tr_σ means summation over all possible system configurations.

For most of the reasonable observables choice and in case of good data quality, the solution of the maximization problem (3) exists and is unique. However for sake of completeness, in app. A we discuss when and how multiple solutions could arise.

A. The geometry of the X- space

In the following sections we will deal with the non trivial geometry embedding the fields space, the $\mathbf{X}$-space. This geometry is described by three matrices: the (negative) log-likelihood Hessian $\mathcal{H}[\mathbf{X}]$, the Fisher matrix $I[\mathbf{X}]$ and the model susceptibility matrix $\chi[\mathbf{X}]$

$$\mathcal{H}_{ab}[\mathbf{X}] \equiv -\nabla_a\nabla_b l[\mathbf{X}]/B \;, \qquad (9)$$

$$I_{ab}[\mathbf{X}] \equiv \langle\nabla_a \ln P_{\mathbf{X}}(\sigma)\ \nabla_b \ln P_{\mathbf{X}}(\sigma)\rangle_{\mathbf{X}} \;, \qquad (10)$$

$$\chi_{ab}[\mathbf{X}] \equiv \langle\Sigma_a\Sigma_b\rangle_{\mathbf{X}} - \langle\Sigma_a\rangle_{\mathbf{X}}\langle\Sigma_b\rangle_{\mathbf{X}} \;. \qquad (11)$$

$\mathcal{H}[\mathbf{X}]$ describes the concavity of the log-likelihood function, $I[\mathbf{X}]$ describes the covariance of the log-likelihood gradient, whereas $\chi[\mathbf{X}]$ that of the observables. If in a general optimization problem these three matrices differ, in the the MaxEnt model inference they coincide. In fact:

$$I_{ab}[\mathbf{X}] \equiv \ \langle\nabla_a \ln P_{\mathbf{X}}(\sigma)\ \nabla_b \ln P_{\mathbf{X}}(\sigma)\rangle_{\mathbf{X}} \qquad (12)$$

$$= -\langle\nabla_a\nabla_b \ln P_{\mathbf{X}}(\sigma)\rangle_{\mathbf{X}} \qquad (13)$$

$$= -\nabla_a\nabla_b l[\mathbf{X}]/B \ \equiv\ \mathcal{H}_{ab}[\mathbf{X}] \qquad (14)$$

where in the first equality we use a well know properties of the Fisher matrix and in the second the fact that the log-likelihood is linear in $\overline{\mathbf{P}}$, the only data-dependent quantities. Moreover:

$$I_{ab}[\mathbf{X}] = -\langle\nabla_a\nabla_b \ln P_{\mathbf{X}}(\sigma)\rangle_{\mathbf{X}} \qquad (15)$$

$$= \nabla_a\nabla_b \ln Z[\mathbf{X}] \qquad (16)$$

$$= \langle\Sigma_a\Sigma_b\rangle_{\mathbf{X}} - \langle\Sigma_a\rangle_{\mathbf{X}}\langle\Sigma_b\rangle_{\mathbf{X}} \ \equiv\ \chi_{ab}[\mathbf{X}] \;. \qquad (17)$$

Indeed:

$$\chi_{ab}[\mathbf{X}] \ = \ I_{ab}[\mathbf{X}] \ = \ \mathcal{H}_{ab}[\mathbf{X}] \ . \qquad (18)$$

The equations (18) are the keystones of this study. They affect the geometry of the $\mathbf{X}$-space and thus the inference in a peculiar way. In particular, the first equality allows us to introduce an efficient inference method, whereas the second allows us to completely characterize its long-time dynamics.

B. A Bayesian Framework for the Maximum Entropy Models

Until now, we posed the MaxEnt inference as a log-likelihood maximization problem, without considering that the finite size of the dataset could affect the estimate of the observables mean. The error in these estimates will inevitably result in some uncertainty on the fields inference that has to be taken into account. In fact, even if a carefully choice of the $\mathbf{\Sigma}$ avoids to include noisy observables, the system heterogeneity will results in a different precision in the fields estimation.

A Bayesian framework including prior and posterior distributions will exactly account for this uncertainty. In fact, through Bayesian inversion, we can compute the posterior distribution of the fields, $P^{\mathrm{Post}}(\mathbf{X}|\mathrm{data})$, from the prior distribution on the fields and the likelihood function, $P(\mathrm{data}|\mathbf{X}) = \prod_b P_{\mathbf{X}}(\sigma(b))$:

$$P^{\mathrm{Post}}(\mathbf{X} \mid \mathrm{data}) = \frac{P(\mathrm{data} \mid \mathbf{X})\, P^{\mathrm{Prior}}(\mathbf{X})}{\mathrm{Norm}}. \quad (19)$$

where Norm is a $\mathbf{X}$ independent normalization constant. The width of the posterior distribution around its maximum quantifies the intrinsic uncertainty on the fields $\mathbf{X}$ and can be used to test the robustness of the inference. Explicitly, any scientific results based on a particular outcome of the fields inference should remain valid for all the fields sets with large posterior probability. Indeed the possibility to sample from the posterior is a powerful tool to test the robustness of the system analysis.

For the prior distribution we have two possible approaches: either include a flat distribution that does not depend on the fields value, either include a probability distribution that reflects some *a priori* knowledge on them. The first approach lets the dataset account for the whole uncertainly, and in case of very good dataset should be preferred. However, in applications dealing with strongly undersampled data and/or when some *a priori* knowledge is available, the second possibility has shown to be powerful.

In this work, as an example to show how to include a prior term, we focus on the $L2$-regularization of the form:

$$P_{L2}^{\mathrm{Prior}}(\mathbf{X}) \equiv \exp\left\{-\frac{B}{2}\mathbf{X}\cdot\eta\cdot\mathbf{X}\right\} \Big/ \left(\frac{2\pi}{|\eta|B}\right)^{\frac{D}{2}}, \quad (20)$$

where η is an arbitrary D-dimensional positive definite square matrix and $|\eta|$ is its determinant.

II. VANILLA GRADIENT, NEWTON METHOD AND DATA-DRIVEN ALGORITHMS

Many of the algorithms suited for solving the MaxEnt inference task performs a dynamics in fields space that, starting from an initial condition, flows toward the maximum of the log-likelihood. In many applications, in fact, the log-likelihood gradient is fast to compute or estimate and can be used to drive the dynamics to the maximum. In this section we first review two of these approaches and then we propose the *Data-Driven* algorithm, the focus of this work.

A. Review: Vanilla Gradient and Newton Method algorithms

Before introducing the method proposed in this work, we like to review two well known inference method.

Ackley, Hinton and Sejnowski [19] posed the inference problem as a dynamical process ascending the log-likelihood function along the gradient direction:

$$\mathbf{X}_{t+1} - \mathbf{X}_t \equiv \delta\mathbf{X}_t^{\mathrm{VG}} = \alpha(\overline{\mathbf{P}} - \mathbf{Q}[\mathbf{X_t}]), \qquad (21)$$

where α is a learning rate. The algorithm works iteratively: at each time-step the computation of the model averages $\mathbf{Q}[\mathbf{X_t}]$ allows to perform the fields update and to proceed towards the convergence, which is guaranteed for sufficiently small α. We call this approach *Vanilla (Standard) Gradient (VG)* algorithm.

Although it follows the gradient, VG will not go trough the shortest path even for arbitrary small α[31]. The reason lies in the geometrical structure induced by the curvature of the log-likelihood function, namely its Hessian, see eq. (9). To take into account this geometrical effect we can multiply the log-likelihood gradient by the inverse of the Hessian [32] obtaining the well known Newton-Raphson method. However as suggested by Amari[33, 34], in the curved manifold of the log-likelihood the natural local metric is the model Fisher matrix $I[\mathbf{X}]$, see eq. (10). Indeed, in the geometry induced by $I[\mathbf{X}]$, the steepest ascendant direction is $I^{-1}\nabla l$, the contravariant form of the gradient (6) [31]. However for the MaxEnt models inference, Hessian and Fisher matrix coincides and so do Newton-Raphson and Natural gradient methods. Consequently, we do not need to distinguish and simply replace the VG update (21) with

$$\delta\mathbf{X}_t^{\mathrm{NM}} = \alpha\chi^{-1}[\mathbf{X}_t] \cdot (\overline{P} - Q[\mathbf{X}_t]) \qquad (22)$$

to obtain both the *Newton Method (NM)* and the Natural gradient. The positiveness of χ ensures the convergence of this method at least for infinitesimally small α (see later). Despite it is optimal, the Newton Method is slowed down by the time-consuming estimation and inversion of $\chi[\mathbf{X}]$ at each update step. In the following we suggest a way to *bypass* this problem.

B. The Data-Driven Algorithm.

As $\chi[\mathbf{X}]$ depends only on model averages of observables products, we can approximate its value at the solution $\mathbf{X} = \mathbf{X}^*$ with the dataset configurations list:

$$\chi_{ab}[\mathbf{X}^*] \approx \overline{\chi_{ab}} \equiv \overline{\Sigma_a \Sigma_b} - \overline{\Sigma_a} \; \overline{\Sigma_b} \; . \qquad (23)$$

This matrix can be computed before running the inference dynamics and then used to evaluate the fields update. The resulting *Data-Driven* (DD) quasi-Newton Method update rule reads:

$$\delta\mathbf{X}_t^{\mathrm{DD}} = \alpha \; \overline{\chi}^{-1} \cdot (\overline{\mathbf{P}} - \mathbf{Q}[\mathbf{X}_t]) \; . \qquad (24)$$

The quality of the approximation $\chi[\mathbf{X}^*] \approx \overline{\chi}$ depends on two main hypotheses: i) the ability of the MaxEnt model to reproduce dataset statistical properties beyond the mean of $\mathbf{\Sigma}$ and in particular the experimental $\mathbf{\Sigma}$-covariance and ii) the good sampling quality of the experimental dataset. The first hypothesis, however, was already partially assumed when it has been chosen the observables to reproduce and so the MaxEnt model to fit. Otherwise stated, if the approximation is poor, it means that the chosen MaxEnt model was not a good choice. The second hypothesis, instead, reflects the quality of the whole inference task. In case of strong undersampling the inference problem, despite being mathematically well posed, is meaningless as the information encoded in the dataset does not support the fine tuning of the fields. In sect. VI we will give a practical condition to test this second hypothesis.

The quality of the approximation (23) controls the speed up factors of the DD algorithm: the better the approximation is, the better and faster the DD will work. On the contrary, when the approximation is poor, the DD will not be advantageous with respect to the VG. In fact, as $\overline{\chi}$ is positive definite, no matter how bad the approximation is, for sufficiently small α the DD will still converge toward the right solution $\mathbf{X}^*$.

In conclusion, for cases where the inference problem is meaningful and the consequently the approximation (23) is valid, the DD algorithm will speed up the inference, whereas in the other cases DD will be useless but not counterproductive.

III. THE LEARNING DYNAMICS

All the approaches explained before, (VG, NM and DD) solve the learning task through a discrete-time dynamics in the fields space. At each time step the estimation of the model averages of the observables ($\mathbf{Q}[\mathbf{X}]$) allows to compute the fields update and continue the dynamics. For large system size N, however, the exponential complexity of the problem prevents the exact computation of the observables averages and some approximations are required. A standard, but fruitful, choice consists in using Markov-Chain Monte-Carlo (MC) with

Metropolis algorithm to sample M system configurations and use them to estimate:

$$\mathbf{Q}_{\mathbf{X}}^{\mathrm{MC}} \equiv \frac{1}{M} \sum_{b=1}^{M} \mathbf{\Sigma}(\sigma^{\mathrm{MC}}(b) \;) \; . \qquad (25)$$

$\mathbf{Q}_{\mathbf{X}}^{\mathrm{MC}}$ is now a random variable approximating $\mathbf{Q}[\mathbf{X}]$ up to $O(\sqrt{1/M})$ fluctuations.

Even for large M, the randomness of $\mathbf{Q}_{\mathbf{X}}^{\mathrm{MC}}$ will eventually affect the convergence of the dynamics, as for $\mathbf{X}_t$ sufficiently close to $\mathbf{X}^*$, the size of the gradient will become comparable with its fluctuations. However, at the beginning, we expect the dynamics to be almost deterministic and then to become stochastic only at the end. We indeed separate the dynamics in two regimes:

1. *Approaching the convergence*, when $\overline{\mathbf{P}} - \mathbf{Q}_{\mathbf{X}}^{\mathrm{MC}}$ is small but still much larger than its fluctuations.

2. *The long-time stochastic dynamics*, when $\overline{\mathbf{P}} - \mathbf{Q}_{\mathbf{X}}^{\mathrm{MC}} \simeq \sqrt{1/M}$ and thus comparable with its fluctuations.

In appendices B and C we will provide several details of the two regimes and here we only present the mayor results.

A. Approaching the convergence

After an initial transient where the dynamics strongly depends on the chosen algorithm (VG,NM or DD) and on the initial conditions, we expect the $\mathbf{X}_t$ to approach the log-likelihood maximum $\mathbf{X}^*$, so that we can approximate the log-likelihood function up to the quadratic order. Given $\delta l[\mathbf{X}] \equiv l[\mathbf{X}] - l[\mathbf{X}^*]$, we have

$$\delta l[\mathbf{X}] \simeq -\frac{B}{2} \; (\mathbf{X} - \mathbf{X}^*) \; \cdot \; \chi[\mathbf{X}^*] \; \cdot \; (\mathbf{X} - \mathbf{X}^*)$$

$$\approx -\frac{B}{2} \; (\mathbf{X} - \mathbf{X}^*) \; \cdot \; \overline{\chi} \; \cdot \; (\mathbf{X} - \mathbf{X}^*) \; . \qquad (26)$$

In this approximation the dynamics is exactly solvable upon projecting the fields on the $\chi[\mathbf{X}^*]$ Eigenvectors $\{\mathbf{V}^\mu\}_{\mu=1}^{D}$: $\delta X_t^\mu = \sum_a V_a^\mu \delta X_{a,t}$.

Along a μ-Eigenspace the convergence of the VG algorithm scales with the corresponding Eigenvalue λ_μ as: $\delta X_t^\mu \sim (1 - \alpha\lambda_\mu)^t$. Consequently, to ensure the algorithm convergence, we need $\alpha < 2/\lambda_\mu$ along all directions. Moreover, by optimizing the convergence speed along all directions simultaneously we obtain $\alpha_{BEST}^{(\mathrm{VG})} = 2/(\lambda_> + \lambda_<)$, where $\lambda_{>/<}$ are the largest/smallest Eigenvalue. In App. B we present some details and here we simply notice how $\alpha_{BEST}^{(\mathrm{VG})}$ can be squeezed to very small values by a large $\lambda_>$ preventing the learning along all the direction with $\lambda_\mu \ll \lambda_>$. Consequently, for dataset where the Eigenvalues of $\chi[\mathbf{X}^*]$, or of its approximation $\overline{\chi}$, spread over several order of magnitude the convergence of the VG algorithm will be very slow.

Because of the quadratic approximation, the NM and the DD algorithms coincide and we do not distinguish between them. Within their dynamics, $\delta X_t^\mu \sim (1-\alpha)^t$ independently of the Eigenvalues λ_μ. consequently $\alpha < 2$ is the only convergence condition and $\alpha_{BEST}^{(DD)} = 1$.

B. The long-time stochastic dynamics

Before proceeding we like to introduce the shortcut notation:

$$\mathcal{N}[\mathbf{m}; \mathbf{s}](t) \equiv \frac{e^{-\frac{1}{2}\sum_{ab}(t_a - m_a)s_{ab}^{-1}(t_b - m_b)}}{\sqrt{(2\pi)^D |\mathbf{s}|}} \;, \qquad (27)$$

a normal distribution with average $\mathbf{m}$ and covariance $\mathbf{s}$ evaluated at t.

When $\mathbf{X} \to \mathbf{X}^*$, we expect $\nabla l_{\mathbf{X}}^{MC} \equiv B(\overline{\mathbf{P}} - \mathbf{Q}_{\mathbf{X}}^{MC}) \to 0$ only on average, with not negligible fluctuations of $O(\sqrt{1/M})$. In this regime the fields dynamics is not anymore deterministic and the convergence becomes a stochastic process. Consequently, rather than converge to a fixed point, the fields will approach an equilibrium stationary regime around $\mathbf{X}^*$.

The stochastic process is ruled by the discrete-time master equation:

$$P_{t+1}(\mathbf{X}') = \int D\mathbf{X} \, P_t(\mathbf{X}) \, W_{\mathbf{X} \to \mathbf{X}'}, \qquad (28)$$

where the transition rates depend on the distribution of $\nabla l_{\mathbf{X}}^{MC}$.

In the large M limit, the gradient distribution can be approximated as a Gaussian to obtain an analytic expression of $W_{\mathbf{X} \to \mathbf{X}'}$, see app. C. By asking $P_t(\mathbf{X})$ to be invariant under the evolution (28) we can obtain the stationary distribution $P_\infty(\mathbf{X})$:

$$P_\infty^{VG}(\mathbf{X}) = \mathcal{N}\left[\mathbf{X}^*; \frac{\alpha}{M}\left(2\delta_{\mathbf{D}} - \alpha\overline{\chi}\right)^{-1}\right](\mathbf{X}), \quad (29)$$

$$P_\infty^{DD}(\mathbf{X}) = \mathcal{N}\left[\mathbf{X}^*; \frac{\alpha}{M(2-\alpha)}\overline{\chi}^{-1}\right](\mathbf{X}), \qquad (30)$$

where δ_D is the identity matrix in dimension D. Here the typical fluctuations of $\mathbf{X}$ around $\mathbf{X}^*$ must consistently verify the approximation $\mathbf{X} \approx \mathbf{X}^*$: $\langle(\mathbf{X} - \mathbf{X}^*)^2\rangle_{P_\infty}$ should be small enough to allow the expansion (26).

In the stationary regime, on top of the fluctuations induced by the MC, the $\mathbf{X}$ distribution will induce a second source of noise in the actual distribution of $\mathbf{Q}^{MC}$. On average we expect (see app. C):

$$P_\infty^{VG}(\mathbf{Q}^{MC}) = \mathcal{N}\left[\overline{\mathbf{P}}; \frac{2\overline{\chi}\left(2\delta_D - \alpha\overline{\chi}\right)^{-1}}{M}\right](\mathbf{Q}^{MC}) (31)$$

$$P_\infty^{DD}(\mathbf{Q}^{MC}) = \mathcal{N}\left[\overline{\mathbf{P}}; \frac{2\overline{\chi}}{M(2-\alpha)}\right](\mathbf{Q}^{MC}) \;. \qquad (32)$$

Interestingly, the two algorithms provide different $\mathbf{Q}^{MC}$ distributions. In particular, the VG fluctuations along

directions with large (small) λ_μ are larger (smaller) than the DD ones. These differences will have consequences on the ability of both algorithms to reproduce the dataset statistics and thus to avoid both under- and over-fitting.

C. The dynamics under an external stochastic force

For forthcoming purposes, here we characterize the learning dynamics when a linear stochastic force term is added to the gradient term in the learning rules. This analysis will be useful when a $L2$ prior term is included in the inference procedure. We modify eq. (24) as:

$$\delta\mathbf{X}_t^{DD\eta} = \alpha \, \overline{\chi_\eta}^{-1} \cdot \left(\nabla l_{\mathbf{X}}^{MC} + F_{\mathbf{X}}^\eta \right) \;. \qquad (33)$$

where

$$\overline{\chi_\eta} \equiv \overline{\chi} + \eta \qquad (34)$$

$$P(F_{\mathbf{X}}^\eta) = \mathcal{N}\left[-\eta \cdot X; \frac{\eta}{M}\right](F_{\mathbf{X}}^\eta). \qquad (35)$$

The calculation for the stationary fields distribution follows as in the previous section and it results in:

$$P_\infty^{DD\eta}(\mathbf{X}) = \mathcal{N}\left[\mathbf{X}_\eta^*; \frac{\alpha}{M(2-\alpha)}\overline{\chi_\eta}^{-1}\right](\mathbf{X}), \qquad (36)$$

where

$$\mathbf{X}_\eta^* = \overline{\chi_\eta}^{-1} \cdot \overline{\chi} \cdot \mathbf{X}^*. \qquad (37)$$

Analogously to (32), from (36) it follows:

$$P_\infty^{DD\eta}(\mathbf{Q}^{MC} - F^\eta) = \mathcal{N}\left[\overline{\mathbf{P}}; \frac{2\overline{\chi_\eta}}{M(2-\alpha)}\right](\mathbf{Q}^{MC} - F^\eta) \;. \qquad (38)$$

Expressions for the VG algorithm can be easily obtained from (36) and (38) with substitution $\alpha \to \alpha\overline{\chi_\eta}^{-1}$.

IV. AVOIDING UNDER- AND OVER-FITTING BY SAMPLING FROM THE POSTERIOR DISTRIBUTION.

In the previous section we shown how any actual implementation of the algorithm is stochastic. In particular, depending on the chosen algorithm and its parameters, α and M, we expect a whole probability distribution on the output fields $\mathbf{X}$. This stochasticity raises questions on which implementation should be preferred in practical applications. The simple strategy of trying to obtain the best possible approximation of $\mathbf{X}^*$ by increasing M or reducing α will unambiguously lead to data over-fitting thus limiting the model prediction power. The presence of noise in the estimation of $\overline{\mathbf{P}}$ induces some uncertainty on the model fields $\mathbf{X}$ that should be taken into account to avoid over-fitting.

This fields uncertainty is quantified by the posterior distribution of the fields given the data, see eq. (19).

Within the approximation (26), the posterior simplifies to

$$P^{\text{Post}}\big(\ \mathbf{X}\ \big|\ \text{data}\ \big) \propto e^{-\frac{B}{2}(\mathbf{X}-\mathbf{X}^*)\cdot\overline{\chi_{ab}}\cdot(\mathbf{X}-\mathbf{X}^*)} \qquad (39)$$

in the case of flat prior and to

$$P^{\text{Post}}\big(\ \mathbf{X}\ \big|\ \text{data}\ \big) \propto e^{-\frac{B}{2}[\ (\mathbf{X}-\mathbf{X}^*)\cdot\overline{\chi}\cdot(\mathbf{X}-\mathbf{X}^*)+\mathbf{X}\cdot\eta\cdot\mathbf{X}\ ]}$$
$$\propto e^{-\frac{B}{2}(\mathbf{X}-\mathbf{X}^*_\eta)\cdot\overline{\chi_\eta}\cdot(\mathbf{X}-\mathbf{X}^*_\eta)} \qquad (40)$$

when an L2 prior, see eq (20), is considered ($\mathbf{X}_\eta$ and $\overline{\chi_\eta}$ are those defined in eqs. (37) and (34)).

If by tuning the algorithm implementation and settings we can match the a stationary fields distribution and the posterior, we will be able to sample from it, thus avoiding any over- and/or under-fitting. In case of flat prior, we have to compare (39) with (29) and (30), whereas in case of L2 prior (40) with (36) for the DD algorithm and the analogous for the VG one: by setting

$$\alpha = \frac{2M}{B+M} \qquad (41)$$

the DD algorithm will sample from the correct posterior distribution and over- and under-fitting will be avoided along all the D dimensions. On the contrary, for the VG algorithm such a setting does not exist and the algorithm will not sample from the posterior. In particular directions with $\alpha\lambda_\mu \ll 2M/(B+M)$ are strongly over-fitted and those with $\alpha\lambda_\mu \gg 2M/(B+M)$ are strongly under-fitted. Otherwise stated: by choosing the VG algorithm with α small enough to well reproduce directions with large λ_μ will result in overfitting of all directions with small λ_μ.

Among the consequences of overfitting, for finite B we expect an overestimation of the observed model log-likelihood. In fact, as the exact inference fields $\mathbf{X}^*$ reproduce also the noise in the experimental averages $\overline{\mathbf{P}}$, we expect $l^* \equiv l[\mathbf{X}^*] > \widehat{l}$, where $\widehat{l}$ is the log-likelihood of the model that generates the data, the true log-likelihood. In order to quantify this effect, in appendix D we compute the average log-likelihood estimation in the case of the exact inference when the data are synthetically generated by a MaxEnt model with true fields $\widehat{\mathbf{X}}$. By averaging over the distribution of $\overline{\mathbf{P}}$, we find that

$$\langle l^* \rangle - \widehat{l} = \frac{D}{2B}\ , \qquad (42)$$

which shows how the log-likelihood maximization, see eq. (3), induces a finite bias leading to a log-likelihood over-estimation. In the case of the DD algorithm, instead, the average over the posterior distribution exactly cancels the bias, and the true log-likelihood value is recovered.

V. ALGORITHM IMPLEMENTATION

The core of DD algorithm is to iteratively update the fields $\mathbf{X}_t$ with the rule (24), where $\mathbf{Q}[\mathbf{X_t}]$ is approximated by $\mathbf{Q}_\mathbf{X}^{\text{MC}}$ through a MC sampling of M configurations. Ideally $\alpha = \alpha_{\text{BEST}}^{\text{DD}} = 1$ and $M = B$ is the fastest setting that satisfies the condition (41). However any practical implementation will face two main difficulties. The first lies in the fact that the quadratic approximation (26) may not be valid. This will happens at the beginning of the dynamics, when $\mathbf{X}_t$ if far from $\mathbf{X}^*$, but also when B is not large enough to have $(\mathbf{X}_t - \mathbf{X}^*)^2$ small enough to discard third order terms. If it is the case, the distribution (30) will be non-stationary. For this reason we allow the algorithm to adapt the value of α at each time-step. The second difficulty reflects the fact that the algorithm is not suppose to converge to $\mathbf{X}^*$ but rather to a probability distribution. Consequently we need a condition that signals the onset of the thermalization and allows us to start storing $\mathbf{X}_t$s as samples from the posterior distribution. For this reason we introduce the following quantity:

$$\epsilon_t \equiv \sqrt{\frac{B}{2D}\Big(\overline{\mathbf{P}}-\mathbf{Q}_\mathbf{X_t}^{\text{MC}}\Big)\cdot\overline{\chi}^{-1}\cdot\Big(\overline{\mathbf{P}}-\mathbf{Q}_\mathbf{X_t}^{\text{MC}}\Big)}\ . \quad (43)$$

Under the distribution (32) with $\alpha = 1$ and $M = B$, ϵ_∞ will have distribution:

$$P(\varepsilon_\infty) = \frac{2\ D^{\frac{D}{2}}}{\Gamma(\frac{D}{2})\ 2^{\frac{D}{2}}}\varepsilon_\infty^{D-1}\ \exp\Big\{-\frac{D}{2}\varepsilon_\infty^2\Big\} \qquad (44)$$

where $\Gamma(x)$ is the Gamma function. ϵ_∞ has:

$$\langle\varepsilon_\infty\rangle = \sqrt{\frac{2}{D}}\ \frac{\Gamma(\frac{D+1}{2})}{\Gamma(\frac{D}{2})}\ \xrightarrow[D\to\infty]{}\ 1 \qquad (45)$$

$$\sqrt{\langle\varepsilon_\infty^2\rangle - \langle\varepsilon_\infty\rangle^2} = \sqrt{1 - \langle\varepsilon_\infty\rangle^2}\ \xrightarrow[D\to\infty]{}\ 0\ . \quad (46)$$

In the large D limit, the ϵ_∞-distribution shrinks to a Dirac-delta function at $\epsilon_\infty = 1$: if before thermalization we expect with high probability $\varepsilon_t \gg 1$, once the algorithm starts sampling from the stationary distribution (30) we expect $\varepsilon_t \approx 1$. Consequently, once the condition $\varepsilon_t \leq 1$ is full filled, the subsequent $\mathbf{X}_t$ are good estimations of the fields. In order to effectively sample from (30) it will be still necessary to keep running the algorithm in order to decorrelate from the initial condition $\mathbf{X}_0$.

The DD algorithm that we implemented can be sketched as follows:

1. Chose an initial configuration for $\mathbf{X_0}$ compute/evaluate $\mathbf{Q}[\mathbf{X_0}]$, see eq. (8) and compute ε_0 from eq. (43). Then set $\alpha_0 = 1$ and $M_0 = \min(\frac{B}{\varepsilon_0^2}, B)$.

2. Iterate the following step:

   (a) update the $\mathbf{X_t}$ through eq. (24),

   (b) estimate $\mathbf{Q}[\mathbf{X_t}]$ though $M_t = \min(\frac{B}{\varepsilon_{t-1}^2}, B)$ Gibbs samplings, see eq. (25),

   (c) compute ε_t from eq. (43) and

i. if $\varepsilon_t < \varepsilon_{t-1}$ accept the update and set $\alpha_t = \alpha_{t-1}\delta^+$

ii. otherwise discard the update, set $\alpha_t = \alpha_{t-1}/\delta^-$ and estimate again $\mathbf{Q}[\mathbf{X_t}]$.

3. As soon as the condition $\epsilon_t < 1$ is full filled,

(a) either fix α and let the distribution $P_t(\mathbf{X})$ decorrelate from $\mathbf{X}_0$ and thermalize to $P_\infty(\mathbf{X})$, see (30).

(b) either stop the algorithm and retain $\mathbf{X}_t$ as a fields list solving the inference problem.

A variable and adapting α is required because the system can be be far outside the validity range of (26). In order to avoid cycles, we heuristically set $\delta^+ = 1.05$ and $\delta^- = \sqrt{2}$.

The choice of an adapting $M_t = \min(\frac{B}{\varepsilon_t^2}, B) \leq B$, instead, allows to save time during the deterministic dynamics regime, see sect. III A, when the algorithm does not require a high precision in the gradient estimate. ε_t, in fact measures also the norm of the log-likelihood gradient in the metric defined by its fluctuations, $B^2 I[\mathbf{X}]/M = B^2\chi[\mathbf{X}]/M \approx B^2\overline{\chi}/M$, see eq. (C4):

$$\varepsilon_t = \sqrt{\frac{M}{B}} \, \|\nabla l[\mathbf{X_t}]\|_{B^2\overline{\chi}/M} \ . \tag{47}$$

For $M_t = B/\varepsilon_t^2$ we expect the estimate of the gradient to be statistically confident within one standard deviation and consequently enough well estimated. Moreover the randomness induced by small M_t values decorrelates the dynamics from the initial condition. When the algorithm approaches convergence, $\varepsilon_t \to 1$ and consequently $M_t \to B$ thus full filling condition (41).

The complexity of the algorithm depends linearly on NB through the MC sampling. From the tests we performed the number of required steps depends mostly on the dataset properties and in particular on the exactness of $\chi[\mathbf{X}^*] \approx \overline{\chi}$. The hardest limit of the DD lies in the memory allocation for storing $\overline{\chi}$. Working in double precision, for an hardware with $32Gb$ of RAM the maximum number of manageable units is $N \sim 350$[35].

A. The $L2$ prior term

In order to include a $L2$ prior term we add to the dynamics the stochastic force term introduced in sect. III C. At each iteration, we should modify the update rule by adding to the gradient term $\nabla l_{\mathbf{X}}^{\mathrm{MC}}$ an independent realization of the random variable $\mathbf{F}_{\mathbf{X}}^\eta$ according to the distribution (35). As the prior term will biases the inference we need to modify the function (43) in order to account for the difference between (32) and (38):

$$\epsilon_t^\eta = \sqrt{\frac{B}{2D}\left(\mathbf{P} - \mathbf{Q} + \mathbf{F}_{\mathbf{X_t}}^\eta\right) \cdot \overline{\chi_\eta}^{-1} \cdot \left(\mathbf{P} - \mathbf{Q} + \mathbf{F}_{\mathbf{X_t}}^\eta\right)} \tag{48}$$
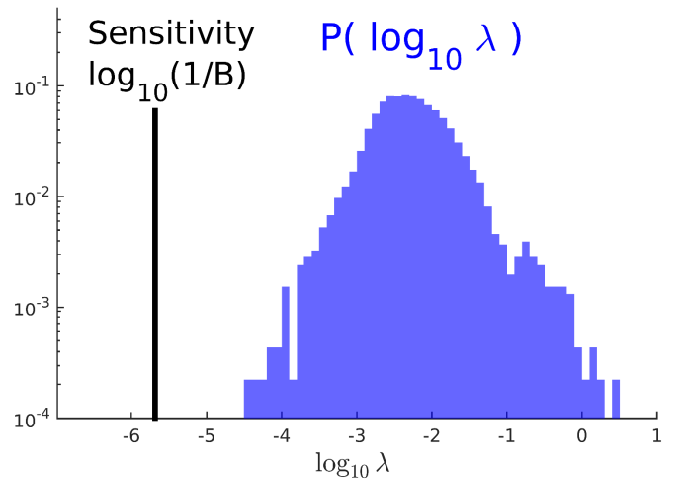

FIG. 1. (Colors online) Eigenvalues distribution of the empirical susceptibility matrix, see eq. (23), for a pairwise Ising model inference problem on biological data. As often happens in biological problems [36], the Eigenvalues span several order of magnitude enhancing the heterogeneity of the parameter space. As will be show in the inset of Fig. 3, the susceptibility matrix of the inferred Ising model has a very similar spectrum. Consequently, even with the optimal learning rate, $\alpha_{BEST}^{(\mathrm{VG})}$ of eq. (B3), VG will be very slow in learning along the directions with the smallest Eigenvalues. As the sensitivity $(1/B)$ lies well below the smallest Eigenvalue all the susceptibility spectrum is informative and has to be reproduced by the inference. See text for more details. Data from a $\sim 2.1h$ (binned at 16ms) recording of 95 rat retinal ganglion cells subject to visual stimulation [12].

which, remarkably, does not require the explicit inversion of the matrix $\overline{\chi}$. Moreover, as $\overline{\chi}$ is, by construction, non-negative, $\overline{\chi_\eta}$ is a positive matrix and can be inverted. As a consequence, the DD algorithm regularized with $L2$ prior can be applied when an undersampling induces zero modes in the empirical estimation of the susceptibility matrix.

VI. TEST

As explained before, the DD algorithm performance depends mostly on how good is the chosen MaxEnt model in modeling the dataset statistical properties. As expected, we succeed in inferring back the fields $\mathbf{X}$ from synthetic dataset obtained through simulation of MaxEnt models. Indeed we find more interesting to present an application to biological data where the underlying statistical distribution does not belong to class of models considered in the Inference task.

We tested the DD algorithm on an *ex-vivo* multielectrode array recording [37] of 95 rat retinal ganglion cells [12]. The retina was stimulated though a video showing two randomly moving bars [12] displayed with a frame rate of $60Hz$. The $\sim 2.1h$ spike trains recording was binned at 16ms to obtain $B \sim 4.8 \cdot 10^5$ system
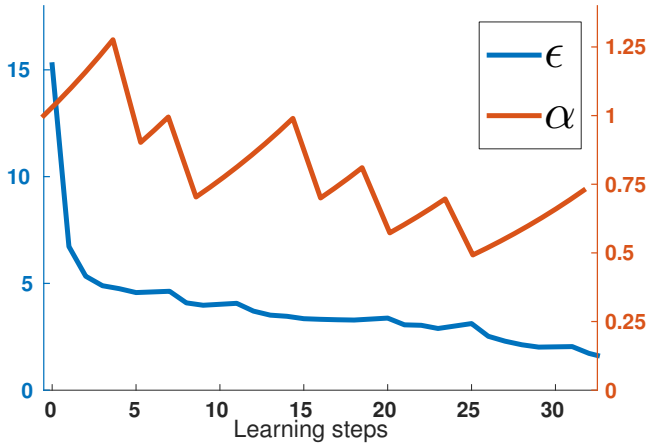
FIG. 2. (Colors online) Behavior of the pairwise Ising model inference of biological data through the Data-Driven algorithm. ε_t (lower blue curve, left axis) and of α_t (upper red curve, right axis) during the algorithm running plotted against the iterative step number. Abscissa is not proportional to the running time as each step performs a different number M of MC sampling, see text. α_t is raised (lowered) at every step where ε_t decreases (increases). The algorithm requires $100 \pm 5s$ to produce a set of fields $\mathbf{X}$ satisfying the early-stop condition $\varepsilon_t < 1$. Dataset as in fig. 1.

FIG. 4. (Colors online) Scatterplots of the inferred couplings (main panel) and biases (inset) with the DD algorithm against those of Naive (blue triangles) and Resummed (red circles) Mean-Field. The black lines show equality. All the approximations fails in the reconstruction with a tendency to overestimate both couplings and fields (notice the different axis scale). Dataset as in fig. 1.
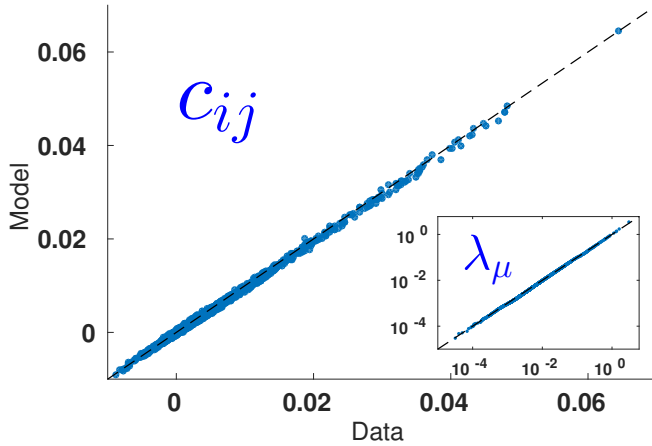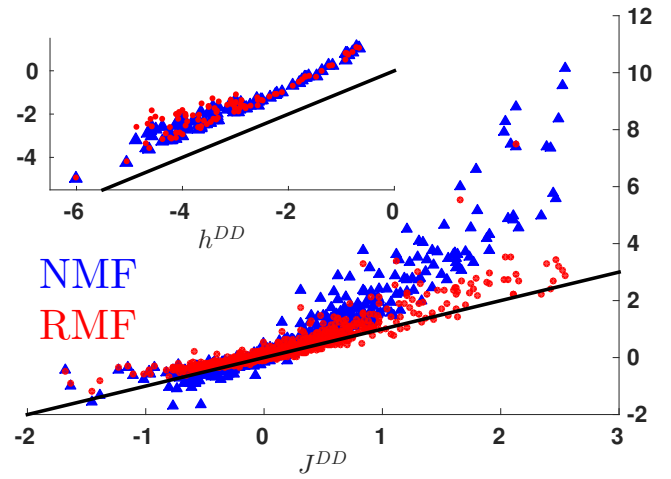

FIG. 3. (Colors online) Scatter-plot of the experimental connected correlations, $c_{ij} = \langle \sigma_i \sigma_j \rangle_{\text{DATA}} - \langle \sigma_i \rangle_{\text{DATA}} \langle \sigma_j \rangle_{\text{DATA}}$, against those estimated with MC sampling of $M = B = 4.8 \ 10^5$ equilibrium configurations of the inferred model distribution. Inset: scatter-plot of the ordered Eigenvalues of $\overline{\chi}$ (x-axis) against those of $\chi[\mathbf{X}^*]$. Dataset as in fig. 1.
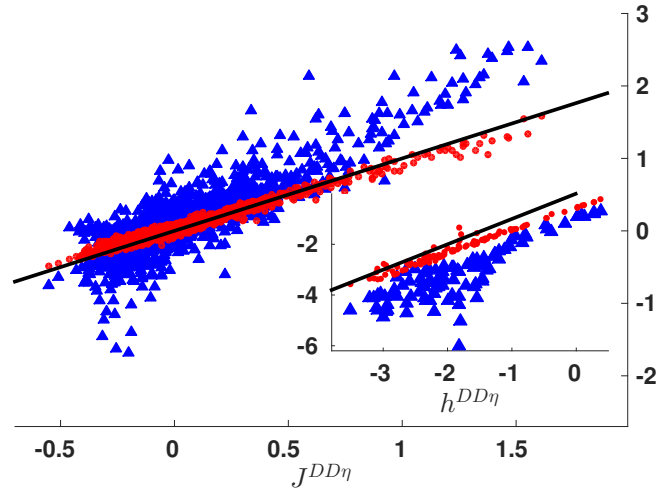

FIG. 5. (Colors online) Scatterplot of the inferred couplings (main panel) and biases (inset) for a pairwise Ising model with and without $L2$ prior term, see eq. (20), with $\eta = 5 \cdot 10^{-3} \ \delta_{\mathbf{D}}$. Blue triangles compare the inferred fields with (x-axis) and without (y-axis) prior term. Red circles compare the inferred fields with regularization (x-axis) with the expected ones (y-axis): $\mathbf{X}_\eta^{\mathbf{DD}} \equiv \overline{\chi_\eta}^{-1} \cdot \overline{\chi} \cdot \mathbf{X}^{\mathbf{DD}}$, see eq. (37). The black line shows equality. As a consequence of the $L2$ regularization the inferred fields are smaller in absolute value. Dataset as in fig. 1.

configurations, where we assign $\sigma_i(b) = 1$ if cell i emitted at least one action potential in the time-bin b and $\sigma_i(b) = 0$ otherwise. We infer a pairwise Ising model thus restricting the observables list to single and pairwise correlations ($\{\Sigma_a\}_{a=1}^D = \{\{\sigma_i\}_{i=1}^N, \{\sigma_i\sigma_j\}_{i<j=1}^N\}$) and consistently the fields to biases and pairwise interactions ($\{X_a\}_{a=1}^D = \{\{h_i\}_{i=1}^N, \{J_{ij}\}_{i<j=1}^N\}$). With this observables choice the model takes the form of the well known

Disordered Ising Model:

$$P_{\mathbf{h},\mathbf{J}}(\sigma) = \exp\left\{ \sum_i h_i \sigma_i + \sum_{i<j} J_{ij} \sigma_i \sigma_j \right\} / Z[\mathbf{h}, \mathbf{J}] \quad (49)$$

where $Z[\mathbf{h}, \mathbf{J}]$ is the normalization constant.

Before performing the inference it is important to check

whether the the number of measurements is large enough to empirically estimate the observables averages with good precision. In the large B limit, we expect an empirical error $\delta\mathbf{P}$ in the estimate of $\overline{\mathbf{P}}$ to have zero mean and covariance $\hat{\chi}/B$, see eq. (D2). By linear regression, en error $\delta\mathbf{P}$ in the data will induce en error:

$$\delta\mathbf{X} = \left(\left.\frac{\delta\mathbf{Q}[\mathbf{X}]}{\delta\mathbf{X}}\right|_{\mathbf{X}=\mathbf{X}^*}\right)^{-1} \cdot \delta\mathbf{P} \approx \hat{\chi}^{-1} \cdot \delta\mathbf{P} \qquad (50)$$

on the inferred fields. Consequently, by approximating $\hat{\chi} \approx \overline{\chi}$:

$$P(\delta\mathbf{X}) = \mathcal{N}\left[\,0\,;\,(B\overline{\chi})^{-1}\,\right](\delta\mathbf{X})\,. \qquad (51)$$

To ask for this error to be small, we should indeed require:

$$\sqrt{B\,\lambda_\mu} \gg 1\,, \quad \text{for all } \mu \Rightarrow \frac{1}{B} \ll \lambda_{\text{MIN}}\,. \qquad (52)$$

As shown in Fig. 1, in the considered dataset, condition (52) is satisfied: the experimental sensitivity, $1/B$ is much smaller then the smaller Eigenvalue of $\overline{\chi}$.

On an Intel® Core™ i7-4770 with eight cores at 3.4GHz the written in Matlab® DD algorithm takes $100 \pm 5s$ to reach convergence. In Fig. 2 we show the behavior of both ϵ_t and α_t against the number of inference steps. As the number M of MC samplings is not constant during the inference, steps are of different time duration.

In order to show the quality of the inferred fields, in the main panel of Fig. 3 we scatter-plot the experimental connected correlations against those estimated through B sampling of the inferred Ising model distribution. In order to give an insight on the validity of (23), in the inset we scatter-plot the ordered Eigenvalues of $\overline{\chi}$ (x-axis) against those of $\chi[\mathbf{X}^*]$.

To convince the reader that the tested inference problem was difficult, we apply two example Mean-Field approaches to the dataset. In Fig. 4 we show the scatter-plot of the inferred fields against those of Naive [25] and Resummed [29] mean-field approximations. Despite Resummed works much better than Naive, for this dataset both approximations largely overestimate couplings and biases.

For comparison, we tested the VG algorithm on the same dataset with $M = B$ and various $\alpha \leq \alpha_{BEST}^{(\text{VG})}$. With $\alpha = 0.2\alpha_{BEST}^{(\text{VG})}$ the VG takes $\sim 4.2 \cdot 10^4 s$ to converge, whereas for larger α values it was not able to satisfies the early-stopping condition $\epsilon < 1$.

In Fig. 5 we scatter plot in blue the inferred fields against those inferred with a $L2$ regularization term, see eq. (20), with $\eta = 5 \cdot 10^{-3}\,\delta_\mathbf{D}$, whereas in red we scatter the inferred with prior fields against the expected ones, see eq. (37).

A part from the retina example, we succeeded in inferring the Ising model fields from synthetic data, human temporal cortex recording [38], rat pre-frontal cortex recording [4] and others. In all the tested cases convergence times are of the order of tens to few hundreds of seconds.

VII. CONCLUSIONS AND DISCUSSION

In this study we introduced a Markov-Chain Monte-Carlo based algorithm for solving the MaxEnt inference problem and sampling from the posterior probability distribution of the MaxEnt fields $\mathbf{X}$, namely couplings and biases of an Ising model or their generalization in the presence of non-pairwise interactions. We carefully analyze the learning dynamics and separate two different regimes: i) a deterministic dynamics that approaches the solution and ii) a stochastic dynamics of the probability distribution of the fields, $P_t(\mathbf{X})$, that thermalizes to a stationary distribution. By tuning the algorithm settings, namely α and M, this distribution can reproduce the posterior distribution of the inference fields thus allowing to sample. We concluded presenting an implementation of the algorithm and a test on a biological dataset showing how the presented algorithm outperforms the standard learning approach. The core of the algorithm is the approximation (23) which requires at first to have enough data to properly estimate the empirical susceptibility then that the probability model chosen for the inference reproduces quite well the data statistics. As the applications to biological data have shown, both conditions have not to be intended as rigid constraints, but rather as requirements for obtaining the largest advantage of the DD approach.

The key properties of the MaxEnt inference that driven this study are the relationships (18). The equality among the log-likelihood Hessian and the model Fisher allows to tune the MC fluctuations of the log-likelihood gradient (ruled by the Fisher) in order to reproduce the fluctuations of Posterior distribution around the log-likelihood maximum (ruled by the Hessian). The equality with the susceptibility matrix allows to obtain an expression of both the Fisher and Hessian that depends only on $\mathbf{X}$-dependent averages of $\mathbf{X}$-independent functions that at the solution $\mathbf{X}^*$ can be approximated through the data.

In sect. III, we have carefully analyzed the learning dynamics and in III B we have characterized it as a stochastic process that converges to a stationary distribution. Later, in sect. IV, we have shown how the DD algorithm stationary distribution can be tuned to the posterior distribution. This characterization has several advantages:

- the computation of $P_\infty^{\text{DD}}(\mathbf{Q}^{\text{MC}})$, together with the running evaluation of ϵ_t, see eq. (43) provides an useful early-stopping condition for the algorithm that allows to check when the inference has been accomplished. Consequently it is useful to save much computational time.

- any inference algorithm has to account for the randomness of the MCMC fluctuation and typically large computational efforts are required to get rid of this noise source (large number of MCMC samplings M). The DD algorithm, instead, takes advantage of these fluctuations to avoid overfitting and to decorrelate from the initial condition.

- as we have shown in appendix D, in comparison with the exact inference, the posterior sampling allows to avoid overfitting the model.

- as in the case of the VG algorithm, an overfitting along several directions happens with all the algorithms that do not *rectify* the fields space.

- as explained in sect. I B, the posterior sampling can provide several consistent lists of inferred fields, thus allowing to test the robustness against the dataset noise of any analysis based on the inferred fields.

Our approach and the use of the MC randomness to induce noise in the algorithm outcome may remind the application of stochastic gradient method for sampling from the fields posterior distribution [39]. Despite the results may look similar, in order the induce fluctuations, here we take advantage from the intrinsic algorithm randomness instead of introducing it by sub-sampling the dataset. The mayor advantage results from the possibility to select and optimize properly the algorithm setting in order to get rid of the MC induced noise that otherwise will affect the dynamics in a spurious way. One important drawback arises from the fact that the algorithm fluctuations may not reproduce the experimental ones when the Gaussian approximation (26) is not valid and second order terms are not enough to approximate the log-likelihood function around the solution. In this case, however, one can notice how the equalities (18) can be extended up to higher order cumulants. These relationships, together with an appropriate update rule that takes into account higher order corrections may extend the equivalence between $P_\infty(\mathbf{X})$ and the posterior. We let this generalization for forthcoming investigations.

The DD approach takes its place in the list of algorithms for exactly solving the MaxEnt model inference problem. Its greatest strengths are the velocity and the possibility to sample from the posterior, whereas its stronger limitation is the memory requirement for the storage of $\overline{\chi}$. As an example, the inference of a $N = 350$ pairwise Ising model requires $32Gb$ of RAM.

Depending on the dataset properties the DD should or should not be preferred to others algorithms as [21], [22] or [24]. In general, if the condition (52) is largely satisfied, we expect DD to be the best choice, because the approximation (23) is expected to be valid. In the opposite case, a largely unsatisfied (52) suggests that the inference of the chosen MaxEnt model is not meaningful and the observables list has to be modified. However for cases in between some tests with different algorithms have to be performed. In particular cases, some available *a priori* knowledge of the system could help in the algorithm choice. For example, if the underlying interaction graph is naturally clusterized in almost non-interacting subcomponents, Selective Cluster Expansion (SCE) [22] is probably the best choice. SCE, in fact, splits the system in many subunits that are recursively joined together to form larger and larger building blocks of the reconstructed interaction network. If the interaction graph is clusterizable, SCE will recognizes these units and accomplish the inference task quickly. However, up to our knowledge, a simple and generic argument for choosing the best suited algorithm for the actual inference problem is still missing and it would be of large interest.

Appendix A: The existence and uniqueness of the inference solution

Because the matrix $\mathcal{H}[\mathbf{X}]$, see eq. (9), measures the concavity of maximization problem (3), its positiveness guarantees the existence and uniqueness of the inference solution $\mathbf{X}^*$. As $\mathcal{H}[\mathbf{X}]$ can be expressed as the covariance matrix $\chi[\mathbf{X}]$, see eq. (18), it is by construction non-negative, but it could have zero modes. A zero mode in a covariance matrix identifies a linear combination of the observables, the corresponding Eigenvector, that does not fluctuate within the model probability distribution. This could happens either for all the values of the parameter $\mathbf{X}$ when some of the functions $\mathbf{\Sigma}(\sigma)$ are linearly dependent either at the solution $\mathbf{X}^*$ when some of the fields diverge quenching (part of) the system[19, 23]. This last case usually happens when the dataset suffers of unsersampling.

As an example, consider a dataset of two non constant spins $\sigma_1(b)$ and $\sigma_2(b)$ that within the dataset are never active together, so that $\sigma_1(b)\sigma_2(b) = 0$ for $b = 1, \ldots, B$. If we take only $\Sigma_1(\sigma) = \sigma_1$ and $\Sigma_2(\sigma) = \sigma_2$ as observables, the inference problem is well posed. However if we include, for example, $\Sigma_3(\sigma) = \sigma_1 + \sigma_2$ or $\Sigma_4(\sigma) = \sigma_1\sigma_2$ the matrix $\chi[\mathbf{X}^*]$ will develop zero modes. In the first case, because $\Sigma_3(\sigma) = \Sigma_1(\sigma) + \Sigma_2(\sigma)$ in the latter case because $X_4 \to -\infty$ in order to fix $\langle\Sigma_4(\sigma)\rangle_{\mathbf{X}^*} = 0$.

Appendix B: On the deterministic convergence of the Vanilla Gradient algorithm.

As introduced in sect. III A, for $\mathbf{X} \approx \mathbf{X}^*$ the deterministic dynamics of the VG algorithm is exactly solvable upon projecting the fields on the $\chi[\mathbf{X}^*]$ Eigenvectors. Along a μ-Eigenspace the convergence of the VG algorithm is not uniform and scales with the corresponding Eigenvalue λ_μ as $(1 - \alpha\lambda_\mu)^t$. Indeed by tuning α to speed up some particular direction, the others can suffers

of very low convergence. The learning rate that optimize the convergence speed along all direction simultaneously is

$$\alpha_{BEST}^{(\text{VG})} \equiv \arg \min_{\alpha} \left[\max_{\mu} \left| 1 - \alpha \lambda_{\mu} \right| \right] \qquad (B1)$$

$$= \arg \min_{\alpha} \max \left[\left| 1 - \alpha \lambda_{<} \right|, \left| 1 - \alpha \lambda_{>} \right| \right]$$

where $\lambda_{>/<}$ are the largest/smallest Eigenvalue. This equation can be solved by equating the two expression in the max:

$$\left| 1 - \alpha_{BEST}^{(\text{VG})} \lambda_{<} \right| = \left| 1 - \alpha_{BEST}^{(\text{VG})} \lambda_{>} \right| \qquad (B2)$$

and the solution reads:

$$\alpha_{BEST}^{(\text{VG})} = \frac{2}{\lambda_{>} + \lambda_{<}}, \qquad (B3)$$

In particular, $\alpha_{BEST}^{(\text{VG})}$ can be squeezed to small value by large $\lambda_{>}$ preventing the learning along the direction with $\lambda_{\mu} \ll \lambda_{>}$. As an example, for the biological data we will consider, see Fig. 1, $\lambda_{>}/\lambda_{<} \simeq 10^5$. Moreover the ratio $\lambda_{>}/\lambda_{<}$ has been shown to diverge in synthetic data of model at criticality [36].

Appendix C: The stationary distribution of the stochastic dynamics

As introduced in sect. III B, the stochastic dynamics of the fields $\mathbf{X}$ is ruled by the discrete-time master equation:

$$P_{t+1}(\mathbf{X}') = \int D\mathbf{X} \, P_t(\mathbf{X}) \, W_{\mathbf{X} \to \mathbf{X}'},$$

where the transition rates depend on the distribution of $\nabla l_{\mathbf{X}}^{\text{MC}}$.

As $\mathbf{Q}_{\mathbf{X}}^{\text{MC}}$ are the average of the observables $\mathbf{\Sigma}$, for large M we expect them to be almost Gaussian distributed with a covariance equal to the $\mathbf{\Sigma}$ covariance, namely $\chi[\mathbf{X}]$, divided by the number of MC measurements:

$$P\left(\mathbf{Q}_{\mathbf{X}}^{\text{MC}}\right) \approx \mathcal{N}\left[\mathbf{Q}[\mathbf{X}] \, ; \, \frac{\chi[\mathbf{X}]}{\mathbf{M}} \right]\left(\mathbf{Q}_{\mathbf{X}}^{\text{MC}}\right). \qquad (C1)$$

Consequently, as $\nabla l_{\mathbf{X}}^{\text{MC}} \equiv B\left(\overline{\mathbf{P}} - \mathbf{Q}_{\mathbf{X}}^{\text{MC}}\right)$ we have:

$$P\left(\nabla l_{\mathbf{X}}^{\text{MC}}\right) \approx \mathcal{N}\left[\nabla l[\mathbf{X}] \, ; \, \frac{\chi[\mathbf{X}]}{M} \right]\left(\nabla l_{\mathbf{X}}^{\text{MC}}\right)$$

If the truncation (26) is valid and $\chi[\mathbf{X} \approx \mathbf{X}^*] \approx \overline{\chi}$, we can replace the gradient mean by the derivative of the approximated log-likelihood,

$$\nabla l[\mathbf{X}] = \nabla\left[-\frac{B}{2}(\mathbf{X} - \mathbf{X}^*)\overline{\chi}(\mathbf{X} - \mathbf{X}^*) \right] \qquad (C2)$$

$$= B \, \overline{\chi}(\mathbf{X}^* - \mathbf{X}), \qquad (C3)$$

to finally obtain:

$$P\left(\nabla l_{\mathbf{X}}^{\text{MC}}\right) \approx \mathcal{N}\left[B \, \overline{\chi}(\mathbf{X}^* - \mathbf{X}); \frac{B^2}{M}\overline{\chi} \right]\left(\nabla l_{\mathbf{X}}^{\text{MC}}\right),$$

The transition rates $W_{\mathbf{X} \to \mathbf{X}'}$ are the probability to measure a value of $l_{\mathbf{X}}^{\text{MC}}$ such that the next fields value in the dynamics is $\mathbf{X}'$. From eqs. (21) and (24) it follows:

$$W_{\mathbf{X} \to \mathbf{X}'}^{\text{VG}} = \mathcal{N}\left[\overline{\chi}(\mathbf{X}^* - \mathbf{X}); \frac{\overline{\chi}}{M} \right]\left(\frac{\mathbf{X}' - \mathbf{X}}{\alpha} \right) \qquad (C4)$$

$$W_{\mathbf{X} \to \mathbf{X}'}^{\text{DD}} = \mathcal{N}\left[\overline{\chi}(\mathbf{X}^* - \mathbf{X}); \frac{\overline{\chi}}{M} \right]\left(\frac{\overline{\chi}(\mathbf{X}' - \mathbf{X})}{\alpha} \right) (C5)$$

By asking $P_t(\mathbf{X})$ to be invariant under the evolution (28) we can obtain the stationary distribution $P_{\infty}(\mathbf{X})$:

$$P_{\infty}^{\text{VG}}(\mathbf{X}) = \mathcal{N}\left[\mathbf{X}^*; \frac{\alpha}{M}\left(2\delta_{\mathbf{D}} - \alpha\,\overline{\chi}\right)^{-1} \right](\mathbf{X}),$$

$$P_{\infty}^{\text{DD}}(\mathbf{X}) = \mathcal{N}\left[\mathbf{X}^*; \frac{\alpha}{M(2 - \alpha)}\,\overline{\chi}^{-1} \right](\mathbf{X}),$$

where δ_D is the identity matrix in dimension D. Here the typical fluctuations of $\mathbf{X}$ around $\mathbf{X}^*$ must consistently verify the approximation $\mathbf{X} \approx \mathbf{X}^*$: $\left\langle (\mathbf{X} - \mathbf{X}^*)^2 \right\rangle_{P_{\infty}}$ should be small enough to allow the expansion (26).

We can obtain conditions on the algorithm convergence by requiring $P_{\infty}(X)$ to be a properly defined probability distribution, namely to be integrable. Remarkably by asking $P_{\infty}(X)$ to have a positive covariance, we re-obtain the upper-bounds for the learning rate α: $\alpha\lambda_{\mu} < 2$, for all μ, for the VG and $\alpha < 2$ for the DD.

In the stationary regime, the fluctuating fields $\mathbf{X}$ will induce a second source of noise in the actual distribution of $\mathbf{Q}^{\text{MC}}$. By inserting the approximation (C2) in the distribution (C1) and then by averaging over $P_{\infty}(\mathbf{X})$, we obtain:

$$P_{\infty}^{\text{VG}}(\mathbf{Q}^{\text{MC}}) = \mathcal{N}\left[\overline{\mathbf{P}}; \frac{2\,\overline{\chi}\,\left(2\delta_D - \alpha\,\overline{\chi}\right)^{-1}}{M} \right](\mathbf{Q}^{\text{MC}}),$$

$$P_{\infty}^{\text{DD}}(\mathbf{Q}^{\text{MC}}) = \mathcal{N}\left[\overline{\mathbf{P}}; \frac{2\,\overline{\chi}}{M(2 - \alpha)} \right](\mathbf{Q}^{\text{MC}}).$$

$P_{\infty}(\mathbf{X})$, see eq.s (29) and (30), allows us to compute the expected deviation of the log-likelihood from $l[\mathbf{X}^*]$. By averaging $\delta l[\mathbf{X}]$, see eq. (26), we obtain:

$$\langle \delta l \rangle_{P_{\infty}^{\text{DD}}} = -\frac{D}{2}\frac{B\,\alpha}{M(2 - \alpha)}$$

$$\sqrt{\langle \delta l^2 \rangle_{P_{\infty}^{\text{DD}}} - \langle \delta l \rangle_{P_{\infty}^{\text{DD}}}^2} = \sqrt{\frac{D}{2}}\frac{B\,\alpha}{M(2 - \alpha)}$$

for the DD algorithm and

$$\langle \delta l \rangle_{P_{\infty}^{\text{VG}}} = -\frac{\alpha B}{2M}\sum_{\mu}\frac{\lambda_{\mu}}{2 - \alpha\lambda_{\mu}}$$

for the VG (we do not report the slightly involved expression of the variance, which also scales as $1/M$). Both estimations are indeed biased to lower values with large fluctuations (of the order of the bias itself). For the DD algorithm the bias depends just on D, α and M and consequently it is data independent. For the VG, instead, it depends strongly on the spectrum of $\overline{\chi}$ and for large

λ_{MAX} or $\alpha \simeq 2/\lambda_{\mathrm{MAX}}$ it can reach very large values, thus nullifying the inference effort.

Note that the three matrix appearing in the covariance of eq. (C1) and in the mean and covariance of eq. (C4) are *a priori* different: the first is the model susceptibility, the second is the log-likelihood Hessian where the third is the model Fisher matrix. However, as explained in section I A for the MaxEnt inference problem these three matrices coincide providing the results (29) and (30).

Appendix D: The posterior sampling avoids to over-estimate the log-likelihood

To better understand the consequences of over-fitting we consider now the case where the system that generates the data is of MaxEnt form with some unknown true fields $\hat{\mathbf{X}}$. By ideally sampling the distribution infinitely many times we can access to the true means of the conjugated observable $\hat{\mathbf{P}}$ and from these compute the true log-likelihood:

$$\widehat{l} = \hat{\mathbf{P}} \cdot \hat{\mathbf{X}} - \ln Z[\hat{\mathbf{X}}] . \qquad (D1)$$

We like to compare the true log-likelihood with that obtained by the exact inference, the one leading to $\mathbf{X}^*$ and with that obtained by the DD algorithm which samples from the posterior. By sampling B times from the true model distribution we can generate synthetic dataset and obtain *empirical* estimates $\overline{\mathbf{P}}$ of $\hat{\mathbf{P}}$. Through the central limit theorem we can approximate the distribution of the expected deviation of the observable means:

$$P\left(\delta\mathbf{P} \right) \equiv P\left(\overline{\mathbf{P}} - \hat{\mathbf{P}} \right) = \mathcal{N}\left[0 ; \frac{\hat{\chi}}{B} \right](\overline{\mathbf{P}}) \qquad (D2)$$

where $\hat{\chi} = \chi[\hat{\mathbf{X}}]$ is the susceptibility matrix of true model.

The exact inference of the fields perfectly reproducing $\overline{\mathbf{P}}$ overestimates the log-likelihood, in fact:

$$\begin{aligned} l^*_{\delta\mathbf{P}} &= \max_{\mathbf{X}} \left[(\hat{\mathbf{P}} + \delta\mathbf{P}) \cdot \mathbf{X} - \ln Z[\mathbf{X}] \right] \\ &= \max_{\delta\mathbf{X}} \left[(\hat{\mathbf{P}} + \delta\mathbf{P}) \cdot (\hat{\mathbf{X}} + \delta\mathbf{X}) - \ln Z[\hat{\mathbf{X}} + \delta\mathbf{X}] \right] \\ &\approx \widehat{l} + \delta\mathbf{P} \cdot \hat{\mathbf{X}} + \max_{\delta\mathbf{X}} \left[\delta\mathbf{P} \cdot \delta\mathbf{X} - \frac{1}{2}\delta\mathbf{X} \cdot \hat{\chi} \cdot \delta\mathbf{X} \right] \\ &= \widehat{l} + \delta\mathbf{P} \cdot \hat{\mathbf{X}} + \frac{1}{2}\delta\mathbf{P} \cdot \hat{\chi}^{-1} \cdot \delta\mathbf{P} \qquad (D3) \end{aligned}$$

where we approximate $\ln Z[\hat{\mathbf{X}} + \delta\mathbf{X}]$ up to the second order. Through the expression (D2) we can approximate the distribution of $l^*_{\delta\mathbf{P}}$ over many realization of the synthetic experiment:

$$P\left(l^* \right) = \mathcal{N}\left[\widehat{l} + \frac{D}{2B} ; \frac{\hat{\mathbf{X}} \cdot \hat{\chi} \cdot \hat{\mathbf{X}}}{B} \right](l^*) \qquad (D4)$$

where as before, D is the dimension of the fields vector and we discard terms of order B^{-2}. l^* is on average positively biased by a factor $\frac{D}{2B}$.
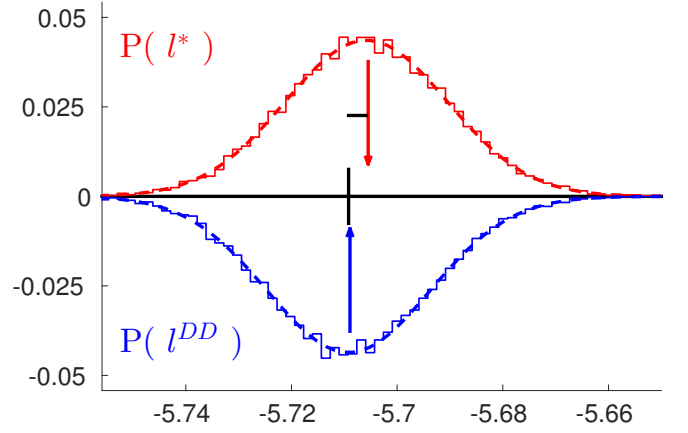

FIG. 6. (Colors online). Histogram of $l^*_{\delta\mathbf{P}}$ (straight) and $l^{\mathrm{DD}}_{\delta\mathbf{P}}$ (reflected) obtained from the inference of $S = 2 \cdot 10^4$ *empirical* estimation of $\overline{\delta\mathbf{P}}$. See text for the details of the synthetic pairwise Ising model used for the simulations. Arrows indicate the corresponding histograms empirical means. The black vertical bar indicates the numerically exact value of $\widehat{l}$, the exact model log-likelihood, whereas the black horizontal segment measures $D/(2B)$. Dotted lines represents the theoretical distributions, see eqs. (D4) and (D7). As predicted, see text, the average of $l^{\mathrm{DD}}_{\delta\mathbf{P}}$ coincides with $\widehat{l}$, whereas the average of $l^*_{\delta\mathbf{P}}$ suffers by a bias of $D/(2B)$

In the case of the DD algorithm ($\alpha = 1$ and $M = B$) we have to substitute the maximization over $\mathbf{X}$ with an integration over the stationary fields distribution (30), which in this case will read:

$$\begin{aligned} P^{\mathrm{DD}}_\infty\left(\delta\mathbf{X} \mid \delta\mathbf{P} \right) &\equiv P^{\mathrm{DD}}_\infty\left(\mathbf{X} - \hat{\mathbf{X}} \mid \delta\mathbf{P} \right) \\ &= \mathcal{N}\left[\hat{\chi}^{-1} \cdot \delta\mathbf{P} ; \frac{\hat{\chi}^{-1}}{B} \right](\delta\mathbf{X}) \end{aligned} \qquad (D5)$$

where the mean equals value of $\delta\mathbf{X}$ after the maximization in the calculation of $l^*_{\delta\mathbf{P}}$, see eq. (D3). For the DD algorithm we obtain:

$$\begin{aligned} l^{\mathrm{DD}}_{\delta\mathbf{P}} &\approx \widehat{l} + \delta\mathbf{P} \cdot \hat{\mathbf{X}} + \left\langle \delta\mathbf{P} \cdot \delta\mathbf{X} - \frac{1}{2}\delta\mathbf{X} \cdot \hat{\chi} \cdot \delta\mathbf{X} \right\rangle_{P^{\mathrm{DD}}_\infty} \\ &= \widehat{l} + \delta\mathbf{P} \cdot \hat{\mathbf{X}} + \frac{1}{2}\delta\mathbf{P} \cdot \hat{\chi}^{-1} \cdot \delta\mathbf{P} - \frac{D}{2B} \\ &= l^*_{\delta\mathbf{P}} - \frac{D}{2B} \qquad (D6) \end{aligned}$$

and again through (D2) we obtain:

$$P\left(l^{\mathrm{DD}} \right) = \mathcal{N}\left[\widehat{l} ; \frac{\hat{\mathbf{X}} \cdot \hat{\chi} \cdot \hat{\mathbf{X}}}{B} \right](l^{\mathrm{DD}}) , \qquad (D7)$$

where again we discard terms of order B^{-2}. Coherently, the integration over the posterior distribution, that prevents to exactly maximize the log-likelihood, cancels the bias in the average of (D7).

To test these results we perform an analysis on a pairwise Ising model of $N = 10$ units, thus restricting the observables list to single and pairwise correlations ($\{\Sigma_a\}_{a=1}^D = \{\{\sigma_i\}_{i=1}^N, \{\sigma_i\sigma_j\}_{i<j=1}^N\}$) and consistently the fields to biases and pairwise interactions

$(\{X_a\}_{a=1}^{D} = \{\{h_i\}_{i=1}^{N}, \{J_{ij}\}_{i<j=1}^{N}\})$. As synthetic model we chose a diluted disorder Ising model on a Erdős-Rényi random graph with average connectivity $c = 4$:

$$
\begin{aligned}
J_{ij} &= 1.33 &, & \quad \text{with prob.}\frac{c}{2(N-1)} \\
J_{ij} &= -1.33 &, & \quad \text{with prob.}\frac{c}{2(N-1)} \\
J_{ij} &= 0, &, & \quad \text{otherwise} \\
h_i &= -0.46 + \sum_{j\neq i} J_{ij} &&
\end{aligned}
\quad \text{(D8)}
$$

We first computed $\widehat{l} = 5.7092$ and then we simulated the model $S = 2 \cdot 10^4$ times to collect S *empirical* estimates of $\overline{\delta\mathbf{P}}$. The number of MCMC sampled was fixed at $B = 2^{13}$. For each of these S realization we estimate $l_{\delta\mathbf{P}}^*$ and $l_{\delta\mathbf{P}}^{\mathrm{DD}}$ and in Fig. 6 we compare their histograms. Dotted lines correspond to the theoretical distributions

(D4) and (D7), whereas the arrows indicate the empirical means. The black vertical lines represents $\widehat{l}$. As can be appreciated by the small difference between the black vertical bar and the blue arrow, the average over the posterior distribution distribution removes the bias produced by the exact inference. Moreover the difference between the red and blue arrows is approximately $D/(2B)$, equals to the length of the horizontal segment. More precisely:

$$
\widehat{l} - \left\langle l_{\delta\mathbf{P}}^* \right\rangle_{P(\ \delta\mathbf{P}\ )} = (-35 \pm 3.0) \cdot 10^{-4}
$$

$$
\frac{D}{2B} + \widehat{l} - \left\langle l_{\delta\mathbf{P}}^* \right\rangle_{P(\ \delta\mathbf{P}\ )} = (-1.2 \pm 3.0) \cdot 10^{-4}
$$

$$
\widehat{l} - \left\langle l_{\delta\mathbf{P}}^{\mathrm{DD}} \right\rangle_{P(\ \delta\mathbf{P}\ )} = (-0.3 \pm 0.6) \cdot 10^{-4}
$$

[1] G. Buzáki. *Large-scale recording of neuronal ensembles* . Nat Neurosci. , **7(5)**:446–51, 2004.

[2] S. J. Phillips, M. Dudk, and R.E. Schapire. *A maximum entropy approach to species distribution modeling.* In Proceedings of the Twenty-First International Conference on Machine Learning , 2004.

[3] E. Schneidman, M. Berry, R. Segev, and W. Bialek. *Weak pairwise correlations imply strongly correlated network states in a population* . Nature , **440**:1007, 2006.

[4] A. Peyrache, M. Khamassi, K. Benchenane, S.I. Wiener, and F.P. Battaglia. *Replay of rule-learning related neural patterns in the prefrontal cortex during sleep* . Nat. Neurosci. , **12** :919–26, 2009.

[5] M. Weigt, R.A. White, H. Szurmant, J.A. Hoch, and T. Hwa. *Identification of direct residue contacts in proteinprotein interaction by message passing* . PNAS, **106(1)**:67–72, 2009.

[6] S. Cocco, R. Monasson, and M. Weigt. *From principal component to direct coupling analysis of coevolution in proteins: Low-eigenvalue modes are needed for structure prediction.* PLoS Comput Biol , **9**:E1003176, 2013.

[7] W. Bialek, A. Cavagna, I. Giardina, T. Mora, O. Pohl, E. Silvestri, M. Viale, and A. Walczak. *Social interactions dominate speed control in driving natural flocks toward criticality..* PNAS, **111(20)**:7212–7217, 2014.

[8] M. Santolini, T. Mora, and V. Hakim. *A General Pairwise Interaction Model Provides an Accurate Description of In Vivo Transcription Factor Binding Sites.* . PLoS Comput Biol , **9(6)**:E99015, 2014.

[9] E. T. Jaynes. *On The Rationale of Maximum-Entropy Method.* Proc. IEEE , **70**:939, 1982.

[10] S. Cocco, S. Leibler, and R. Monasson. *Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods* . Proc. Natl. Acad. Sci. USA , **106**:14058, 2009.

[11] L. S. Hamilton, J. Sohl-Dickstein, A. G. Huth, V. M. Carels, K. Deisseroth, and S. Bao. *Optogenetic Activation of an Inhibitory Network Enhances Feedforward Functional Connectivity in Auditory Cortex.* Neuron , **80**:1066–76, 2013.

[12] T. Mora, S Deny, and O Marre. *Dynamical criticality in the collective activity of a population of retinal neurons* . Phys. Rev. Lett. , **114**:078105, 2015.

[13] T. Tavoni, U. Ferrari, S. Cocco, F.P. Battaglia, and R. Monasson. *Inferred network of the prefrontal cortex activity unveils task-related coupling potentiations and cell assemblies* . submitted , , 2014.

[14] T. R. Lezon, J. R. Banavar, M. Cieplak, A. Maritan, and N. V. Fedoroff. *Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns.* PNAS, **103**:19033–19038, 2006.

[15] A.L. Ferguson, J.K. Mann, S. Omarjee, T. Ndung'u, B.D. Walker, and A. Chakraborty. *Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design..* Immunity , **38**:606–617, 2013.

[16] J.K. Mann, J.P. Barton, A.L. Ferguson, S. Omarjee, B.D. Walker, Chakraborty A., and T. Ndung'u. *The Fitness Landscape of HIV-1 Gag: Advanced Modeling Approaches and Validation of Model Predictions by In Vitro Testing.* PLoS Comput Biol , **10(8)**:e1003776, 2014.

[17] E. Ganmor, R. Segev, and E. Schneidman. *Sparse low-order interaction network underlies a highly correlated and learnable neural population code* . PNAS, **108**:9679–9684, 2011.

[18] G. Tkacik, O. Marre, D. Amodei, E. Schneidman, W Bialek, and Berry M.J. *Searching for collective behaviour in a network of real neurons* . PloS Comput. Biol., **10(1)**:e1003408, 2014.

[19] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski. A learning algorithm for boltzmann machines. *Cognitive Science*, 9:147–169, 1985.

[20] G. Hinton. *Training Products of Experts by Minimizing Contrastive Divergence.* Neural Comput. , **14(8)**:1771–1800, 2002.

[21] T. Broderick, M. Dudik, G. Tkacik, R.E. Schapire, and W. Bialek. *Faster solutions to the inverse pairwise Ising problem* . Arxiv:0712.2437, 2007.

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

[23] J. Barton and S. Cocco. *Ising models for neural activity inferred via selective cluster expansion: structural and coding properties* . J. Stat. Mech. , :P03002, 2013.

[24] J. Sohl-Dickstein, P. B. Battaglino, and M. R. DeWeese. *New Method for Parameter Estimation in Probabilistic*

Models: Minimum Probability Flow. Phys. Rev. Lett. , **107**:220601, 2011.

[25] H.J. Kappen and F.B. Rodriguez. *Efficient learning in boltzmann machines using linear response theory.*. Neural Comput. , **10**:1137–1156, 1997.

[26] T. Tanaka. *Mean-field theory of Boltzmann machine learning.* Phys. Rev. E, **58**:2302, 1998.

[27] E Aurell and M. Ekeberg. *Inverse Ising Inference Using All the Data* . Phys. Rev. Lett. , **108**:090201, 2012.

[28] F Ricci-Tersenghi. *The Bethe approximation for solving the inverse Ising problem: a comparison with other inference methods* . J. Stat Mech , page P08015, 2012.

[29] H. Jacquin and A. Rancon. *Efficient, fast and principled mean-field inference for strongly coupled data, 2015.*

[30] *In order to lighten the notation we do not distinguish between column or row vector and we avoid any transpose symbol.*

[31] *S. Amari.* Natural Gradient Works Efficiently in Learning, Neural Computation. *Neural Comput.* , **10**:251–276, 1998.

[32] *W.H. Press, S. A. Teukolsky, V. T. Vetterling, and B. P. Flannery.* Numerical Recepies. *Cambridge University Press (Cambridge, U.K.), 2007.*

[33] *S. Amari and S.C. Douglas.* Why natural gradient?. *Proc. IEEE* , **2**:1213–16, 1998.

[34] *S. Amari and H. Nagaoka.* Methods of information geometry. *Oxford University Press, Oxford, 2007.*

[35] *As the matrix $\overline{\chi}$ has several symmetries a compressed encoding could decrease the required storage memory.*

[36] *B.B. B. Machta, R. Chachra, M.K. Transtrum, and J.P. Sethna.* Parameter Space Compression Underlies Emergent Theories and Predictive Models. *Science* , **342**:604–607, 2013.

[37] *O. Marre, D. Amodei, N. Deshmukh, K. Sadeghi, F. Soo, T. Holy, and M.J. Berry.* Recording of a large and complete population in the retina. *Journal of Neuroscience* , **32(43)**:1485973, 2012.

[38] *A. Peyrache, N. Dehghani, Eskandar E. N., J. R. Madsen, W. S. Anderson, L.R. Donoghue, J.A. Hochberg, E. Halgren, S.S. Cash, and A. Destexhe.* Spatiotemporal dynamics of neocortical excitation and inhibition during human sleep . *Pnas* , **109** :1731–36, 2012.

[39] *M. Welling and Y.W. Teh.* Bayesian learning via stochastic gradient langevin dynamics. Proceedings of the 28th International Conference on Machine Learning (ICML)*, page 681688, 2011.*