Optimal errors and phase transitions in high-dimensional generalized linear models

Jean Barbier[a,b,1,2], Florent Krzakala[b,1], Nicolas Macris[c,1], Léo Miolane[d,1,2], and Lenka Zdeborová[e,1]

[a]Quantitative Life Sciences, International Center for Theoretical Physics, 34151 Trieste, Italy; [b]Laboratoire de Physique de l'Ecole Normale Supérieure, Université Paris-Sciences-et-Lettres, Centre National de la Recherche Scientifique, Sorbonne Université, Université Paris-Diderot, Sorbonne Paris Cité, 75005 Paris, France; [c]Communication Theory Laboratory, School of Computer and Communication Sciences, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland; [d]Département d'Informatique de l'Ecole Normale Supérieure, Université Paris-Sciences-et-Lettres, Centre National de la Recherche Scientifique, Inria, 75005 Paris, France; and [e]Institut de Physique Théorique, Centre National de la Recherche Scientifique et Commissariat à l'Energie Atomique, Université Paris-Saclay, 91191 Gif-sur-Yvette, France

Edited by David L. Donoho, Stanford University, Stanford, CA, and approved January 18, 2019 (received for review February 14, 2018)

Generalized linear models (GLMs) are used in high-dimensional machine learning, statistics, communications, and signal processing. In this paper we analyze GLMs when the data matrix is random, as relevant in problems such as compressed sensing, error-correcting codes, or benchmark models in neural networks. We evaluate the mutual information (or "free entropy") from which we deduce the Bayes-optimal estimation and generalization errors. Our analysis applies to the high-dimensional limit where both the number of samples and the dimension are large and their ratio is fixed. Nonrigorous predictions for the optimal errors existed for special cases of GLMs, e.g., for the perceptron, in the field of statistical physics based on the so-called replica method. Our present paper rigorously establishes those decades-old conjectures and brings forward their algorithmic interpretation in terms of performance of the generalized approximate message-passing algorithm. Furthermore, we tightly characterize, for many learning problems, regions of parameters for which this algorithm achieves the optimal performance and locate the associated sharp phase transitions separating learnable and nonlearnable regions. We believe that this random version of GLMs can serve as a challenging benchmark for multipurpose algorithms.

high-dimensional inference | generalized linear model | Bayesian inference | perceptron | approximate message-passing algorithm

As datasets grow larger and more complex, modern data analysis requires solving high-dimensional estimation problems with very many parameters. Developing algorithms for this task and understanding their limitations have become a major challenge in computer science, machine learning, statistics, signal processing, communications, and related fields.

In the present contribution, we address this challenge in the case of generalized linear estimation models (GLMs) (1, 2) where data are generated as follows: Given an n-dimensional vector $\mathbf{X}^*$, hidden to statisticians, they observe instead an m-dimensional vector $\mathbf{Y}$ where each component reads

$$Y_\mu = \varphi\left(\frac{1}{\sqrt{n}}[\mathbf{\Phi X}^*]_\mu, A_\mu\right), \qquad 1 \le \mu \le m, \qquad \textbf{[1]}$$

where $\mathbf{\Phi}$ is an $m \times n$ "measurement" or "data" matrix, and the random variables $(A_\mu) \overset{\text{iid}}{\sim} P_A$ account for noise/randomness of the model. The model is "linear" because the output Y_μ depends on a linear combination of the data $z_\mu = \frac{1}{\sqrt{n}}[\mathbf{\Phi X}^*]_\mu = \frac{1}{\sqrt{n}}\sum_{i=1}^n \Phi_{\mu i} X_i^*$. The GLM generalizes the ordinary linear regression by allowing the output function $\varphi(z, A)$ to be nonlinear and/or stochastic; in the case of a deterministic model we simply write $\varphi(z)$. Explicit examples are given below.

GLMs belong to the realm of supervised learning and arise in a wide variety of scientific fields. In signal processing one usually observes Y_μ given as a linear combination of the signal elements $\mathbf{X}^*$. In a range of applications these observations are obtained via a nonlinear function φ. In optics or X-ray crystallography one often measures only the amplitude of $[\mathbf{\Phi X}^*]_\mu$, leading to the phase retrieval problem (3). A real-valued analog is the problem of sign retrieval when we observe only $|[\mathbf{\Phi X}^*]_\mu|$ (4, 5). Observations are sometimes quantized to reduce the storage, leading for instance to the problem of 1-bit compressed sensing (6). In statistics and machine learning, classification is often described via a GLM where the output function φ is discrete and corresponds to the labels that classify the data points $\mathbf{\Phi}_\mu$ (1, 2, 7). GLMs with nonlinear output functions are also the basic building blocks of each layer of neural networks (8): φ corresponds to the activation, the rows of the matrix $\mathbf{\Phi}$ are different data samples, and $\mathbf{X}^*$ is the set of synaptic weights to be learned.

There are two main learning problems in GLMs: (i) The estimation task requires, knowing the measured vector $\mathbf{Y}$ and the matrix Φ, inference of the unknown vector $\mathbf{X}^*$; (ii) the prediction or generalization task instead requires, again knowing $\mathbf{Y}$ and Φ, accurate prediction of new values Y_{new} when new rows (i.e., data points) are added to the matrix $\mathbf{\Phi}$.

Significance

High-dimensional generalized linear models are basic building blocks of current data analysis tools including multilayers neural networks. They arise in signal processing, statistical inference, machine learning, communication theory, and other fields. We establish rigorously the intrinsic information-theoretic limitations of inference and learning for a class of randomly generated instances of generalized linear models, thus closing several decades-old conjectures. Moreover, we delimit regions of parameters for which the optimal error rates are efficiently achievable with currently known algorithms. Our proof technique is able to deal with the output nonlinearity and is hence of independent interest, opening ways to establish similar results for models of neural networks where nonlinearities are essential but in general difficult to account for.

Author contributions: J.B., F.K., N.M., L.M., and L.Z. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: Code related to this paper has been deposited in Github (https://github.com/sphinxteam/GeneralizedLinearModel2017).

[1] J.B., F.K., N.M., L.M., and L.Z. contributed equally to this work.

[2] To whom correspondence may be addressed. Email: jbarbier@ictp.it or leo.miolane@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1802705116/-/DCSupplemental.

Published online March 1, 2019.

In the present paper we build a rigorous theory for both of these tasks for random instances of the GLM. In this setting each element $\Phi_{\mu i}$ of the matrix is sampled independently from a probability distribution of zero mean and unit variance, and the unknown vector $\mathbf{X}^*$ has been also created randomly from a probability distribution P_0, with each of its components $X_1^*, \ldots, X_n^* \overset{\text{iid}}{\sim} P_0$. Since our main aim is to study the intrinsic information-theoretic and algorithmic limitations caused by the lack of samples and/or the amplitude of the noise, we assume throughout this paper that P_0 and φ are known to the statistician (if they are not, the task can only be harder). Our results are derived in the challenging and interesting high-dimensional limit where $m, n \to \infty$ and $m/n \to \alpha$ a constant. Random instances of GLMs are both practically and theoretically relevant in many different contexts:

i) In signal processing, GLM estimation with a random matrix $\mathbf{\Phi}$ has been studied with considerable attention in the context of compressed sensing (9–11), where an n-dimensional sparse signal is recovered from $m < n$ noisy measurements. While standard compressed sensing focused on the linear case—where $\varphi(z, A) = z + A$ with a Gaussian noise A—the generalized case was also widely studied (12, 13), especially for quantized output (14) and 1-bit compressed sensing (6, 15) where $\varphi(z, A) = \text{sign}(z + A)$, as well as for compressive phase retrieval when $\varphi(z, A) = |z + A|$ (16).

ii) In statistical learning, a substantial amount of activity is dedicated to understanding the limitation of learning with data generated by GLMs, both in the linear case, e.g., in the context of ridge regression or least absolute shrinkage and selection operator (LASSO) (17), or with nonlinear probabilistic output, e.g., logistic regression. Random instances were studied in particular in the context of so-called M estimators (18–21).

iii) In studies of artificial neural networks there has been a large amount of work using random instances of GLMs, with φ playing the role of a nonlinear activation function. In this context the random GLM was introduced as the teacher–student setting for the perceptron in the pioneering work of Gardner and Derrida (22). A large volume of work followed and is reviewed, e.g., in refs. 23–25. While initial works concentrated on a simple activation function $\varphi(z) = \text{sign}(z - K)$ (K is the threshold constant), many other functions were considered, e.g., in refs. 26–28. Recently, the study of random instances of neural networks has emerged as a key ingredient in understanding the performance of deep-learning algorithms (29, 30). Computing mutual information in GLMs is also a critical issue in confirming the information bottleneck scenario of refs. 31 and 32.

iv) In communications, error-correcting codes that use random constructions are particularly efficient, as discussed by Shannon in his seminal paper (33). Random instances of GLMs describe both the setting of code-division multiple access—a multiuser access method used in communication technologies (34, 35)—and an error correction scheme called sparse superposition codes, which have been shown to achieve the Shannon capacity for any type of noisy channel (36–40).

Interestingly there is an important gap in the above volume of work. On the one hand there are studies that rely on the algorithmic performance of the so-called generalized approximate message-passing (GAMP) algorithm (11, 12, 41). GAMP is remarkable in that its asymptotic ($n, m \to \infty$, $m/n \to \alpha$) performance can be analyzed rigorously using the so-called state evolution (42–45). However, GAMP is not expected to be always information-theoretically optimal. On the other hand, other results are concerned with the linear case of the GLM with additive Gaussian noise for which the information-theoretically optimal performance was established in refs. 46–48 (the methodology of these works unfortunately does not generalize straightforwardly to the important nonlinear case or to other types of additive noise). All of the other works, which provide information-theoretic results for the nonlinear case, are based on powerful and sophisticated but nonrigorous techniques originating in statistical physics of disordered systems, such as the cavity and replica methods (49). Historically, the first of these nonrigorous, yet correct, results on information-theoretic limitations of learning was for the perceptron with binary weights and was established using the replica method in refs. 22, 50, and 51, including a discontinuous phase transition to perfect learning that appears as the ratio between the number of samples and the dimension exceeds $\alpha \approx 1.249$.

In the present paper we close the above gap between mathematically rigorous work and conjectures (some of them several decades old) from statistical mechanics. In particular, we prove that the results for GLMs stemming from the replica method are indeed correct and imply the optimal value of both the estimation and generalization error. These results are summarized in *Main Results*. The proof is based on the adaptive interpolation method recently developed in ref. 52 and is of independent interest as it is applicable to a range of other models. We present it in *Methods and Proofs* and in *SI Appendix*. We compare our information-theoretic results to the performance of the GAMP algorithm and its state evolution (reviewed briefly in *Main Results*). We determine regions of parameters where this algorithm is or is not information-theoretically optimal. Up to technical assumptions (specified below), our results apply to all activation functions φ and priors P_0, thus unifying a large volume of previous work where many particular functions have been analyzed on a case-by-case basis. This generality allows us to provide a unifying understanding of the types of phase transitions and phase diagrams that we can encounter in GLMs, which is as well of independent interest and we devote *Application to Learning and Inference* to its presentation.

Main Results

This section summarizes our main results. Their formal statement and all technical assumptions and full proofs are provided in *Methods and Proofs* and in *SI Appendix*.

For the random GLM problem as defined in the Introduction, the optimal way to estimate the ground-truth signal/weights $\mathbf{X}^*$ relies on its posterior probability distribution

$$P(\mathbf{x}|\mathbf{Y}, \mathbf{\Phi}) = \frac{1}{\mathcal{Z}(\mathbf{Y}, \mathbf{\Phi})} \prod_{i=1}^{n} P_0(x_i) \prod_{\mu=1}^{m} P_{\text{out}}\left(Y_\mu \middle| \frac{[\mathbf{\Phi x}]_\mu}{\sqrt{n}}\right), \qquad \textbf{[2]}$$

where we used the prior P_0 of $\mathbf{X}^*$ and introduced the likelihood P_{out} that an output Y_μ is observed given $\frac{1}{\sqrt{n}}[\mathbf{\Phi x}]_\mu$. $P_{\text{out}}(\cdot\,|\,z)$ is the probability density function of $\varphi(z, A)$ [where again the random variable (r.v.) $A \sim P_A$ accounts for noise]. We are concerned with the so-called Bayes-optimal setting where the prior P_0 and the likelihood P_{out} that appear in the posterior **2** were also used to generate the ground-truth signal $\mathbf{X}^*$ and the labels $\mathbf{Y}$, with a known random matrix $\mathbf{\Phi}$.

A first quantity of interest is the free entropy (which is the free energy up to a sign) defined as $f_n(\mathbf{Y}, \mathbf{\Phi}) \equiv \frac{1}{n} \ln \mathcal{Z}(\mathbf{Y}, \mathbf{\Phi})$. The expectation of the free entropy is equal to minus the conditional entropy density of the observation $-\frac{1}{n} H(\mathbf{Y}|\mathbf{\Phi})$, as well as (up to an additive constant) to the mutual information density between the signal and the observations $\frac{1}{n} I(\mathbf{X}^*; \mathbf{Y}|\mathbf{\Phi})$.

The Free Entropy. Our first result is the rigorous determination of the free entropy, in the high-dimensional asymptotic regime $n, m \to \infty$, $m/n \to \alpha$. For a random matrix $\mathbf{\Phi}$ with independent entries of zero mean and unit variance, for output $\mathbf{Y}$ that was generated using **[1]**, and under appropriate technical assumptions stated precisely in *Methods and Proofs*, the free entropy converges in probability to

$$f_n(\mathbf{Y},\mathbf{\Phi}) \equiv \frac{1}{n}\ln \mathcal{Z}(\mathbf{Y},\mathbf{\Phi}) \xrightarrow[n\to\infty]{\mathbb{P}} \sup_{q\in[0,\rho]}\inf_{r\geq 0} f_{\mathrm{RS}}(q,r;\rho), \qquad \textbf{[3]}$$

where $\rho \equiv \mathbb{E}_{P_0}[(X^*)^2]$ and where the potential $f_{\mathrm{RS}}(q,r;\rho)$ is

$$f_{\mathrm{RS}}(q,r;\rho) \equiv \psi_{P_0}(r) + \alpha \Psi_{P_{\mathrm{out}}}(q;\rho) - rq/2\,, \qquad \textbf{[4]}$$

$$\text{with}\quad \psi_{P_0}(r) \equiv \mathop{\mathbb{E}}_{[Z_0,X_0]} \ln \int dP_0(x)\, e^{rxX_0+\sqrt{r}xZ_0-rx^2/2}\,, \qquad \textbf{[5]}$$

$$\Psi_{P_{\mathrm{out}}}(q;\rho) \equiv \mathop{\mathbb{E}}_{[V,W,\tilde{Y}_0]} \ln \int \mathcal{D}w P_{\mathrm{out}}(\tilde{Y}_0|\sqrt{q}\,V+\sqrt{\rho-q}\,w)\,, \qquad \textbf{[6]}$$

where $\mathcal{D}w = dw\exp(-w^2/2)/\sqrt{2\pi}$ is a standard Gaussian measure and the scalar r.v. are independently sampled from $X_0 \sim P_0$, then $V, W, Z_0 \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and $\tilde{Y}_0 \sim P_{\mathrm{out}}(\cdot|\sqrt{q}\,V + \sqrt{\rho-q}\,W)$. Only the special linear case with Gaussian P_{out} is known rigorously so far (46–48). Convergence of the averaged free entropy is precisely stated in *Theorem 1*; the one in probability follows from concentration results in *SI Appendix*.

One can check by explicit comparison that for specific choices of P_0 and P_{out} the expression **4** is the replica-symmetric free entropy derived in numerous statistical physics papers (thus the RS in f_{RS}) and in particular in refs. 22, 41, 50, and 51 for $\varphi(z)=\mathrm{sign}(z)$. The formula for general P_0 and P_{out} was conjectured based on the statistical physics derivation in ref. 13. Establishing **[3]** closes these old conjectures and yields an important step toward vindication of the cavity and replica methods for inference, along with, e.g., refs. 43 and 53. We now discuss the main consequences of this formula.

Overlap and Optimal Estimation Error. Our second result concerns the overlap between a sample $\mathbf{x}$ from the posterior **2** and the ground truth. We obtain that as $n, m\to\infty$, $n/m\to\alpha$,

$$\frac{1}{n}\left|\mathbf{x}\cdot\mathbf{X}^*\right| \xrightarrow[n\to\infty]{\mathbb{P}} q^* \qquad \textbf{[7]}$$

whenever $q^*=q^*(\alpha)$ the maximizer in formula **3** is unique. This is the case for almost every α (*SI Appendix*).

It is a simple fact of Bayesian inference that, given the measurements $\mathbf{Y}$ and the measurement matrix $\mathbf{\Phi}$, the estimator $\hat{\mathbf{X}}$ that minimizes the mean-square error with the ground-truth $\mathbf{X}^*$ is the mean of the posterior distribution **2**; i.e., $\hat{\mathbf{X}} = \mathbb{E}_{P(\mathbf{x}|\mathbf{Y},\mathbf{\Phi})}[\mathbf{x}]$. The minimum mean-square error (MMSE) that is achieved by such a "Bayes-optimal" estimator is deduced, again in the limit $n\to\infty$, $m/n\to\alpha$, as follows:

$$\mathrm{MMSE} = \frac{1}{n}\mathbb{E}[\|\mathbf{X}^*-\hat{\mathbf{X}}\|^2] \to \rho - q^*. \qquad \textbf{[8]}$$

We refer to *Theorem 2* in *Methods and Proofs* for rigorous statements. Again the value of the MMSE is known rigorously so far only for the linear case with Gaussian noise (46–48) (and conjectured for the nonlinear case, e.g., in ref. 13).

Optimal Generalization Error. Our third result concerns the prediction error, also called generalization error. Consider again the statistical model **1**. To define the Bayes-optimal generalization error, one is given a new row of the matrix/data point, denoted $\mathbf{\Phi}_{\mathrm{new}}\in\mathbb{R}^n$ (in addition to the data $\mathbf{\Phi}$ and associated outputs $\mathbf{Y}$ used for the learning), and is asked to estimate the corresponding output value Y_{new}. We seek for an estimator $\hat{Y}_{\mathrm{new}} = \hat{Y}_{\mathrm{new}}(\mathbf{Y},\mathbf{\Phi},\mathbf{\Phi}_{\mathrm{new}})$ that achieves $\mathcal{E}_{\mathrm{gen}} \equiv \min_{\hat{Y}_{\mathrm{new}}}\mathbb{E}[(Y_{\mathrm{new}}-\hat{Y}_{\mathrm{new}})^2]$, i.e., that minimizes the MSE with the true Y_{new} obtained using the ground-truth weights $\mathbf{X}^*$. Such an estimator is again obtained from the posterior: $\hat{Y}_{\mathrm{new}} = \mathbb{E}_{P_A(a)}\mathbb{E}_{P(\mathbf{x}|\mathbf{Y},\mathbf{\Phi})}\varphi(\frac{1}{\sqrt{n}}\mathbf{\Phi}_{\mathrm{new}}\cdot\mathbf{x}, a)$. Note that this is different from the plug-in estimator $\tilde{Y}_{\mathrm{new}} = \varphi(\frac{1}{\sqrt{n}}\mathbf{\Phi}_{\mathrm{new}}\cdot\hat{\mathbf{X}})$, which leads to a worse MSE than $\hat{Y}_{\mathrm{new}}$. Yet it is often used in practice for deterministic models since most algorithms for generalized linear regression do not provide the full posterior distribution.

Our result states that the optimal generalization error follows from the I-MMSE theorem (54) applied to the free entropy **3** (see *SI Appendix* for details). The optimal generalization error reads as $n\to\infty$, $m/n\to\alpha$ (q^* is the maximizer in **[3]**),

$$\mathcal{E}_{\mathrm{gen}} \to \mathop{\mathbb{E}}_{V,a}\left[\varphi(\sqrt{\rho}\,V,a)^2\right] - \mathop{\mathbb{E}}_{V}\left[\mathop{\mathbb{E}}_{w,a}\left[\varphi(\sqrt{q^*}\,V+\sqrt{\rho-q^*}w,a)\right]^2\right], \qquad \textbf{[9]}$$

where $V, w \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and $a\sim P_A$. See again *Theorem 2* in *Methods and Proofs* for the precise statement (and *SI Appendix, Theorems 3* and *4*).

Note that for labels $\mathbf{Y}$ belonging to a discrete set the MSE might not be a suitable loss and we are more often interested in maximizing the so-called overlap, i.e., the probability of obtaining the correct label. In this case the Bayes-optimal estimator is computed as the argmax of the posterior marginals, rather than as their mean; i.e., for discrete labels $\bar{Y}_{\mathrm{new}} = \mathrm{argmax}_y \mathbb{P}(y = \varphi(\frac{1}{\sqrt{n}}\mathbf{\Phi}_{\mathrm{new}}\cdot\mathbf{x}, a))$ where again $\mathbf{x}$ is distributed according to **[2]**, $a\sim P_A$. The replica method has been used to compute the optimal generalization error for the perceptron where $\varphi(x) = \mathrm{sign}(z)$ in the pioneering works of refs. 23, 50, and 55. We note that in this special case the plug-in estimator $\tilde{Y}_{\mathrm{new}}$ is actually equal to the optimal one $\bar{Y}_{\mathrm{new}}$.

A final note concerns the issue of overfitting. In optimization-based approaches to learning overfitting may lead to a generalization error which is too large compared with the training error. In the Bayes-optimal setting the estimators are constructed to not overfit. This is related to general properties of Bayes-optimal inference and learning that are called "Nishimori conditions" in the physics literature (13) and that turn out to be crucial in our proofs.

Optimality of Approximate Message Passing. Although the three results stated above are of an information-theoretic nature, our fourth one concerns the performance of an algorithm for solving random instances of GLMs called GAMP (11–13), which is closely related to the Thouless–Anderson–Palmer (TAP) equations developed in statistical physics (41, 56, 57).

The GAMP algorithm can be summarized as follows (11–13): Given initial estimates $\widehat{\mathbf{x}}^0, \mathbf{v}^0$ for the marginal posterior means and variances of the unknown signal vector $\mathbf{X}^*$ entries, GAMP iterates the following equations, with $g_\mu^0 = 0$:

$$\begin{cases} V^t = \overline{\mathbf{v}^{t-1}} & \\ \boldsymbol{\omega}^t = \mathbf{\Phi}\widehat{\mathbf{x}}^{t-1}/\sqrt{n} - V^t\mathbf{g}^{t-1} & \\ g_\mu^t = g_{P_{\mathrm{out}}}(Y_\mu, \omega_\mu^t, V^t) & \forall\,\mu = 1,\dots m \\ \lambda^t = \alpha\,\overline{g_{P_{\mathrm{out}}}^2(\mathbf{Y},\boldsymbol{\omega}^t, V^t)} & \\ \mathbf{R}^t = \widehat{\mathbf{x}}^{t-1} + (\lambda^t)^{-1}\mathbf{\Phi}^\intercal\mathbf{g}^t/\sqrt{n} & \\ \widehat{x}_i^t = g_{P_0}(R_i^t, \lambda^t) & \forall\, i = 1,\dots n \\ \mathrm{v}_i^t = (\lambda^t)^{-1}\,\partial_R g_{P_0}(R,\lambda^t)|_{R=R_i^t} & \forall\, i=1,\dots n \end{cases}$$

(here we denote by $\bar{\mathbf{u}}$ the average over all of the components of a vector $\mathbf{u}$). The so-called thresholding function $g_{P_0}(R,\lambda)$ is defined as the mean of the normalized distribution $\propto P_0(x)\exp(-\lambda(R-x)^2/2)$ and the output function

$g_{P_{\rm out}}(Y,\omega,V)$ is similarly the mean of the normalized distribution (of x) $\propto P_{\rm out}(Y|\omega+\sqrt{V}x)\exp(-x^2/2)$.

The heuristic derivation of GAMP in statistical physics (13) suggests via the definition of the function $g_{P_{\rm out}}$ that $\boldsymbol{\omega}$ and V are the estimates of the means and average variance of the components of the variable $\mathbf{z}=\mathbf{\Phi}\mathbf{x}$. This, in turn, suggests a GAMP prediction of labels of new data points,

$$\hat{Y}^{GAMP,t}_{\rm new} = \int y\, P_{\rm out}(y|\omega^t_{\rm new}+z\sqrt{V^t})\, dy\mathcal{D}z,$$

where $\omega^t_{\rm new}\equiv\frac{1}{\sqrt{n}}\mathbf{\Phi}_{\rm new}\cdot\hat{\mathbf{x}}^{t-1}$. Comparing it with the test-set labels, this serves to compute GAMP's generalization error.

One of the strongest assets of GAMP is that its performance can be tracked via a closed-form procedure known as state evolution (SE), again in the asymptotic limit when $n,m\to\infty$, $m/n\to\alpha$. For proofs of SE see refs. 43 and 44 for the linear case and ref. 45 for the generalized one. In our notations, SE tracks the correlation (or "overlap") between the true weights $\mathbf{X}^*$ and their estimate $\hat{\mathbf{x}}^t$ defined as $q^t\equiv\lim_{n\to\infty}\frac{1}{n}\mathbf{X}^*\cdot\hat{\mathbf{x}}^t$ via

$$q^t = 2\psi'_{P_0}(r^t)\,, \qquad r^t = 2\alpha\Psi'_{P_{\rm out}}(q^{t-1};\rho)\,. \qquad \textbf{[10]}$$

The derivatives are with respect to (w.r.t.) the first argument. Similarly for the evolution of GAMP's generalization error $\mathcal{E}^{\rm GAMP,t}_{\rm gen}$ (*SI Appendix*) we obtain that it is asymptotically, and with high probability, given by the right-hand side (r.h.s.) of formula **9** but with q^* replaced by q^t.

It is a simple algebraic fact that the fixed points of the SE Eqs. **10** correspond to the critical points of the potential **4**. The question of GAMP achieving asymptotically optimal MMSE or generalization error therefore reduces to the study of the extrema of the two-scalar-variables potential **4**. If the SE **10** converges to the same couple (q,r) as the extremizer (q^*,r^*) of **[3]**, then GAMP is optimal, and if it does not, then GAMP is suboptimal. In the next section we illustrate this result on several examples, delimiting regions where GAMP reaches optimality. We note that optimality of AMP-based algorithms in terms of the MMSE on the ground-truth vector $\mathbf{X}^*$ was proved for several cases where the extremizer q^* in **[3]** is unique, e.g., ref. 58, or in the linear case of GLM in ref. 47. Our results allow us to complete the characterization of regions of parameters where the algorithm reaches optimal performance in terms of the estimation and generalization errors. While the asymptotic value of the Bayes-optimal generalization error was predicted for some cases of $P_{\rm out}$ and P_0 (55), and TAP-based algorithms were argued to reach this performance in refs. 59 and 60, it was not known whether this error can be achieved provably or for what exact regions of parameters the algorithm is suboptimal. Our present work settles this question due to the state evolution of the GAMP algorithm. Interestingly, heuristic arguments based on the glassy nature of the corresponding probability measure were used to argue that direct sampling or optimization-based approaches will not be able to match this performance (51). Whether this statement is correct goes beyond the scope of the present paper.

Application to Learning and Inference

In this section, we report what our results imply for the information-theoretically optimal errors and those reached by the GAMP algorithm for several interesting cases of output functions φ and prior distributions P_0. We do not seek to be exhaustive in any way; we simply aim to illustrate the kind of insights about the GLM that can be obtained from our results. We focus on determination of phase transitions in performance as we vary parameters of the model, e.g., the number of samples or the sparsity of the signal. We use careful numerical procedures to compute the expectations required in formula **4** and check that the reported results are stable toward the choice of various precision parameters. In this section we, however, do not seek rigor in bounding formally the corresponding numerical errors. Many of the codes used in this section are given online in a github repository (62).

General Observations About Fixed Points and Terminology.

Noninformative fixed point and its stability. It is instrumental to analyze under what conditions $q^*=0$ is the optimizer in **[3]**. Our result **8** about the MMSE implies that if $q^*=0$, then the MMSE is as large as if we had no samples/measurements at our disposition. A necessary condition for $q^*=0$ is that it is a fixed point of the state evolution. In turn, a sufficient condition for the state evolution **10** to have such a fixed point is that (i) the output density $P_{\rm out}(y|z)$ is even in the argument z and that (ii) the prior P_0 has zero mean. A proof of this is given in *SI Appendix*. For $q^*=0$ to be a fixed point to which the state evolution **10** converges, it needs to be stable. We detail in *SI Appendix* that under properties i and ii this fixed point is stable when

$$\alpha\int dy\,\frac{\left(\int\mathcal{D}z(z^2-1)P_{\rm out}(y|\sqrt{\rho}z)\right)^2}{\int\mathcal{D}zP_{\rm out}(y|\sqrt{\rho}z)}<1. \qquad \textbf{[11]}$$

In what follows we denote α_c the largest value of α for which the above condition holds. Consequently the error reachable by the GAMP algorithm is as bad as random guessing for both the estimation and generalization errors as long as $\alpha<\alpha_c$. For $\alpha>\alpha_c$, starting with infinitesimal positive q the state evolution will move toward larger q as in ref. 63. Note that condition **11** also appears in a recent work (61) as a barrier for performance of spectral algorithms.

Concerning the information-theoretically optimal error, we call the phase where $\mathrm{MMSE}=\rho$, i.e., $q^*=0$ is the extremizer of **[4]**, the noninformative phase. Existing literature sometimes refers to such behavior as the retarded learning phase (64), in the sense that in this case a critical number of samples is required for the generalization error to be better than random guessing. Below we evaluate condition **11** explicitly for several examples.

Almost exact recovery fixed point. Another fixed point of **[10]** that is worth our particular attention is the one corresponding to almost exact recovery, meaning with average error per coordinate going to 0 as $n\to\infty$, where $q^*=\rho$. A sufficient and necessary condition for this to be a fixed point is that $\lim_{q\to\rho}\Psi'_{P_{\rm out}}(q;\rho)=+\infty$. This means that the integral of the Fisher information of the output channel diverges,

$$\int dyd\omega\,\frac{e^{-\frac{\omega^2}{2\rho}}}{\sqrt{2\pi\rho}}\,\frac{P'_{\rm out}(y|\omega)^2}{P_{\rm out}(y|\omega)}=+\infty\,,$$

where $P'_{\rm out}(y|\omega)$ denotes the partial derivative w.r.t. ω. This typically means that the output channel should be noiseless. For example, for the Gaussian channel with noise variance Δ, the above expression equals $1/\Delta$. For the probit channel where $P_{\rm out}(y|z)=\mathrm{erfc}(-yz/\sqrt{2\Delta})/2$ the above expression at small Δ is proportional to $1/\sqrt{\Delta}$.

Stability of the almost exact recovery fixed point depends nontrivially on the properties of both the output channel and the prior. Below we give several examples where almost exact recovery either is or is not possible. In what follows we call the region of parameters for which $\mathrm{MMSE}=0$, i.e., $q^*=\rho$ is the extremizer in **[3]**, the almost exact recovery phase.

Hard phase. As can be anticipated from the statement of our main algorithmic result, there are regions of parameters for which the error reached by GAMP is asymptotically equal to the optimal error and regions where it is not. We call the hard phase the region of parameters where $\mathrm{MMSE}<\mathrm{MSE}_{\rm AMP}$ with a strict inequality. Focusing on the ratio α between the number of samples and the dimensionality, we denote $\alpha_{\rm IT}$ the ratio for which the hard phase appears and $\alpha_{\rm AMP} > \alpha_{\rm IT}$ the ratio for which it disappears. In other words, the hard phase is an interval $(\alpha_{\rm IT}, \alpha_{\rm AMP})$ and is associated to a first-order phase transition in the Bayes-optimal posterior probability distribution.

It remains a formidable open question of average computational complexity whether in the setting of this paper (and for problems that are NP complete in the worst case) there exists an efficient algorithm that achieves better performance than GAMP in the hard phase. We are not aware of any and tend to conjecture that there is not.

Sensing Compressively with Nonlinear Outputs. Existing literature covers in detail the case of noiseless compressed sensing, i.e., when the output function $\varphi(z) = z$. The representative sparse prior distribution is the Gauss–Bernoulli (GB) distribution $P_0 = \rho\mathcal{N}(0,1) + (1-\rho)\delta_0$, where ρ is the average fraction of nonzeros, which are in this case standard Gaussians. The phase diagram of this case is well known (67, 68). In noiseless compressed sensing with random i.i.d. matrices and GB prior, almost exact recovery of the signal is possible for $\alpha > \alpha_{\rm IT} = \rho$ and GAMP recovers the signal for $\alpha > \alpha_{\rm AMP,CS}$ where $\alpha_{\rm AMP,CS}$ is plotted in Fig. 1 (*Left*) with a dotted red line, thus delimiting the hard phase of compressed sensing. We note that the Donoho–Tanner phase transition (9) known as the performance limit of the LASSO ℓ_1 regularization is slightly higher than $\alpha_{\rm AMP,CS}$.

Signless output channel. The phase diagram of noiseless compressed sensing changes intriguingly when only the absolute value of the output is measured, i.e., when $\varphi(z) = |z|$ instead of $\varphi(z) = z$. Such an output channel is reminiscent of the widely studied phase retrieval problem (3) where the signal is complex valued and only the amplitude is observed. The generalization of our results for the complex case would require extensions, as done for the algorithmic aspects in ref. 69. The real-valued case was studied under the name "sparse recovery from quadratic measurements" in the literature, e.g., ref. 70 and references therein, when the number of nonzero variables grows slower than linearly with the dimension n. Our results give access to the phase diagram of sparse recovery from quadratic (or equivalently signless) measurements that is presented in Fig. 1 (*Left*) for the GB prior.

We observe that the information-theoretical phase transition $\alpha_{\rm IT}$ is the same in the signless sparse recovery as in the canonical linear case; i.e., almost exact recovery is possible whenever $\alpha > \rho$. However, the algorithmic phase transition $\alpha_{\rm AMP}$ above which GAMP is able to find the sparse signal is strikingly larger for the signless case (solid red line in Fig. 1, *Left*). (We note that to break the symmetry that prevents GAMP from finding the signal in a constant number of iteration steps, we mismatch infinitesimally the output function φ used in the algorithm from the symmetric one used to generate the data. Another way to deal with this issue is related to a spectral initialization as discussed recently in ref. 61.) We note that even for a dense signal $\rho = 1$ almost exact recovery is algorithmically possible only for $\alpha > \alpha_{\rm AMP}(\rho = 1) \approx 1.128$. For very sparse signals, small ρ, the situation is even more striking because the measurement rate of at least $\alpha > \alpha_c = 1/2$ is needed for algorithmically tractable almost exact recovery for every ρ. This is in sharp contrast with the canonical compressed sensing where $\alpha_{\rm AMP,CS} \to 0$ as $\rho \to 0$. The nature of this algorithmic difficulty of GAMP is related to the symmetry of the output channel due to which the noninformative fixed point is stable for $\alpha < \alpha_c = 1/2$. Summarizing this result in one sentence, tractable compressive sensing is impossible (for $\alpha < 1/2$) if we have lost the signs. We reiterate that this result holds in the setting of the present paper, i.e., in particular when the sparsity ρ is of constant order. For signals where $\rho = o(1)$ the situation is expected to be different (70).

Rectified linear unit output channel. Another case of output channel that attracted our interest is the rectified linear unit (ReLU), $\varphi(z) = \max(0, z)$, as widely used in multilayer feedforward neural networks. In the present single-layer case reconstruction with the ReLU output is interesting mathematically. With GB signals, roughly half of the measurements are given without noise, but the only information we have about the other half is its sign. A straightforward upper bound for both information-theoretic and tractable almost exact recovery is simply twice as many measurements than needed in the canonical

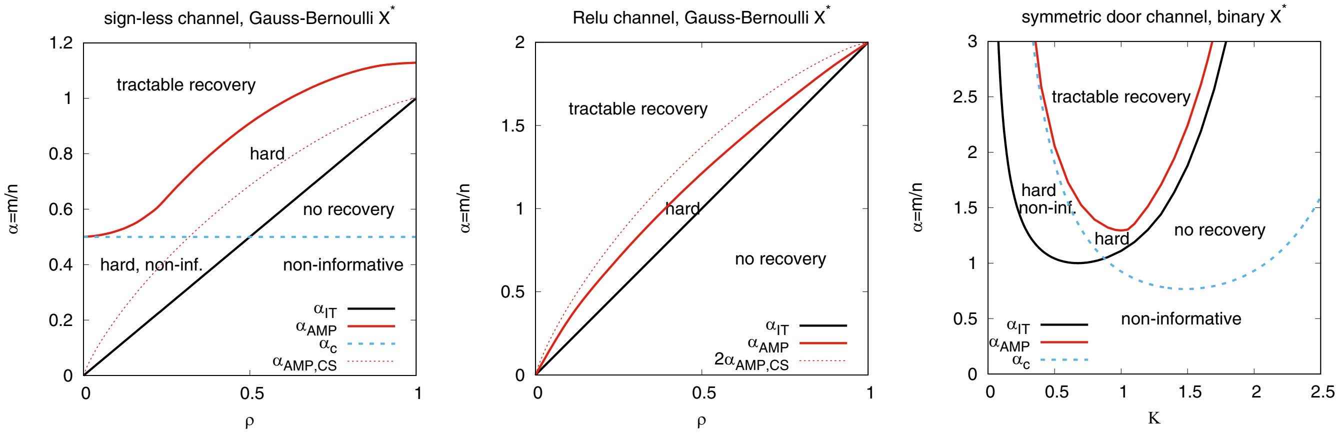

Fig. 1. Phase diagrams showing boundaries of the region where almost exact recovery is possible (in absence of noise). (*Left*) The case of signless sparse recovery, $\varphi(x) = |x|$ with a Gauss–Bernoulli signal, as a function of the ratio between number of samples/measurements and the dimension $\alpha = m/n$, and the fraction of nonzero components ρ. Evaluating **[4]** for this case, we find that a recovery of the signal is information-theoretically impossible for $\alpha < \alpha_{\rm IT} = \rho$. Recovery becomes possible starting from $\alpha > \rho$, just as in the canonical compressed sensing. Algorithmically the signless case is much harder. Evaluating **[11]**, we conclude that GAMP is not able to perform better than a random guess as long as $\alpha < \alpha_c = 1/2$, and the same is true for spectral algorithms (61). For larger values of α, the inference using GAMP leads to better results than a purely random guess. GAMP can recover the signal and generalize perfectly only for values of α larger than $\alpha_{\rm AMP}$ (solid red line). The dotted red line shows for comparison the algorithmic phase transition of the canonical compressed sensing. (*Center*) Analogous to *Left*, for the ReLU output function, $\varphi(x) = \max(0, x)$. Here it is always possible to perform better than random guessing using GAMP. The dotted red line shows the algorithmic phase transition when using information only about the nonzero observations. (*Right*) Phase diagram for the symmetric door output function $\varphi(z) = \text{sign}(|z| - K)$ for a Rademacher signal, as a function of α and K. The stability line α_c is depicted as a dashed blue line, the information-theoretic phase transition to almost exact recovery $\alpha_{\rm IT}$ is a solid black line, and the algorithmic one $\alpha_{\rm AMP}$ is a solid red line.

noiseless compressed sensing. It is interesting to ask whether this bound is tight. Results in the present paper imply that for the information-theoretic performance this bound indeed is tight. However, the phase transition α_{AMP} above which almost exact recovery is possible with the GAMP algorithm is strictly lower than twice the phase transition of compressed sensing; both are depicted in Fig. 1, *Center*. This implies that while the negative outputs are not useful information theoretically, they do help to achieve better performance algorithmically.

Perceptron and Similar Problems.

Binary and Gauss–Bernoulli perceptron. One of the most studied problems that fits in the setting of the present paper is the problem of the perceptron (71), where $\varphi(z) = \mathrm{sign}(z)$, that has been analyzed for random patterns $\mathbf{\Phi}$ in the statistical physics literature; see refs. 23–25 for reviews. We plot in Fig. 2 the optimal generalization error **9** as follows from our results for the binary perceptron, i.e., weights taken from the Rademacher distribution $P_0 = \frac{1}{2}\delta_{+1} + \frac{1}{2}\delta_{-1}$ (Fig. 2, *Left*) and for the GB perceptron where $P_0 = \rho\mathcal{N}(0,1) + (1-\rho)\delta_0$ (Fig. 2, *Center*). The information-theoretically optimal value of the generalization error that we report and prove agrees with existing predictions obtained by the nonrigorous replica method from refs. 50, 51, and 55. Notably, we see that for the GB case the optimal generalization error decreases smoothly as α increases, while for the binary case the generalization error has a first-order (i.e., discontinuous) phase transition toward perfect generalization at $\alpha_{\mathrm{IT}} \approx 1.249$ as predicted already in ref. 50. Our results provide rigorous validation for these old conjectures.

Furthermore, our results together with recent literature on GAMP provide a refreshing clarification of the algorithmic questions. It is natural to ask for what region of parameters the optimal generalization error can be provably achieved with efficient algorithms. This question remained unanswered until now. Indeed, for the spherical perceptron the optimal generalization error was computed in ref. 55 and argued empirically in small instances to be achievable with a TAP-like algorithm (59). The state evolution of GAMP together with our formulas for the generalization error (**[9]** for the average optimal one and with q^t replacing q^* in this formula for GAMP) imply that the optimal generalization error is indeed achievable asymptotically for all α in the GB perceptron.

For the binary perceptron the optimal generalization error was computed in refs. 50 and 51. By comparing with the state evolution of GAMP we obtain that it can also be asymptotically achieved by GAMP, but this time only outside of the hard phase $(\alpha_{\mathrm{IT}}, \alpha_{\mathrm{AMP}})$ with $\alpha_{\mathrm{AMP}} \approx 1.493$. The literature was unclear on the algorithmic question; ref. 50 identified the spinodal of the replica-symmetric solution to be at $\alpha \approx 1.493$, but did not attribute it to any algorithmic or physical meaning. Ref. 51 argues that metastable states exist at least up to $\alpha_{\mathrm{RSB}} \approx 1.628$ and speculates that Gibbs sampling-based algorithms will not be able to reach perfect generalization before that point (23). Taking our results into account, the main algorithmic question that remains open is whether efficient algorithms can reach perfect generalization for $\alpha_{\mathrm{IT}} < \alpha < \alpha_{\mathrm{AMP}}$.

Symmetric door. Out of interest we explored an example of a binary output channel for which $P_{\mathrm{out}}(y|z)$ is even in the argument z, so that the noninformative fixed point $q^* = 0$ exists. Specifically we analyzed the symmetric door channel with $\varphi(z) = \mathrm{sign}(|z| - K)$ and Rademacher prior P_0. In literature such a perceptron is studied with the replica method in the context of lossy data compression (28). In Fig. 1, *Right* we report the phase diagram in terms of the stability line of the noninformative fixed point α_c (below which GAMP is not better than random guesses), the information-theoretic phase transition toward perfect generalization α_{IT}, and the phase transition of GAMP to perfect generalization α_{AMP}.

A simple-counting lower bound states that for binary outputs and weights X_i^* perfect generalization is not possible for $\alpha < 1$. Thus it is interesting to note that the symmetric door channel is able to saturate this lower bound for $K \approx 0.6745$ for which the probability of $y_\mu = 1$ is $1/2$. This saturation was already stated in ref. 28. Our results also, however, imply that in that case the perfect generalization will not be achievable with the GAMP (and we conjecture no other efficient) algorithm unless $\alpha > \alpha_{\mathrm{AMP}} \approx 1.566$. The generalization error that GAMP provides for this case is depicted in Fig. 2, *Right*.

Empirical comparison with general-purpose algorithms. In this section we argue that many cases that fit into the setting of the present paper could serve as useful benchmarks for existing machine-learning algorithms. We believe that the situation is perhaps similar to Shannon coding theorems that have driven algorithmic developments in error-correcting codes, achieving the Shannon bound being the primary goal in many works in communications. In machine learning, classification is a natural task and algorithms are usually benchmarked using open access databases. In current state-of-the-art applications of machine learning we usually have very little insight about what is the

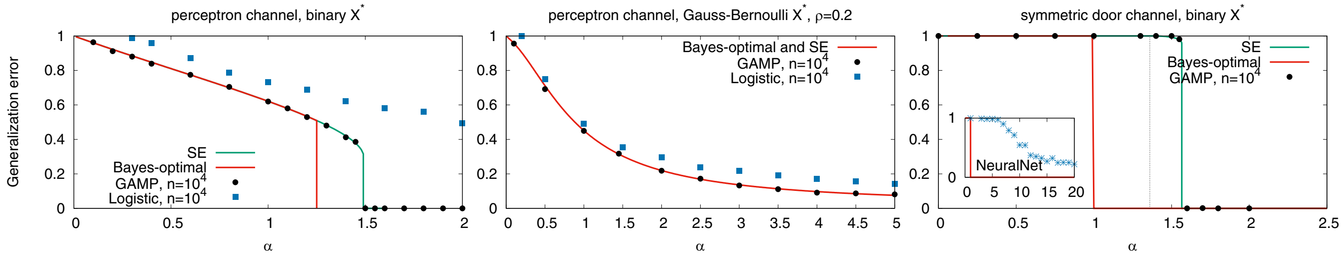

Fig. 2. Optimal generalization error in three classification problems vs. the sample complexity α, the size of the training set being αn. The solid red line is the Bayes-optimal generalization error **9** while the solid green line shows the (asymptotic) performances of GAMP as predicted by the state evolution **10**. For comparison, we also show the results of GAMP (black circles) and the performance of a standard out-of-the-box solver (blue squares). (*Left*) Perceptron, with $\varphi(x) = \mathrm{sign}(x)$ and a binary Rademacher signal. While a perfect generalization is information-theoretically possible starting from $\alpha_{\mathrm{IT}} \approx 1.249$, the state evolution predicts that GAMP will achieve such perfect prediction only above $\alpha_{\mathrm{AMP}} \approx 1.493$. The results of a logistic regression with fine-tuned regularizations with the software scikit-learn (65) are shown for comparison. (*Center*) Perceptron with Gauss–Bernoulli distribution of the weights. No phase transition is observed in this case, but a smooth decrease of the error with α. The results of a logistic regression are very close to optimal. (*Right*) The symmetric door activation rule with parameter K chosen to observe the same number of occurrences of the two classes. In this case there is a sharp phase transition from as bad as random to perfect generalization at $\alpha_{\mathrm{IT}} = 1$. GAMP identifies the rule perfectly, starting only from $\alpha_{\mathrm{AMP}} \approx 1.566$. The noninformative fixed point is stable up to $\alpha_c = 1.36$ (dashed gray line). Interestingly, this nonlinear rule seems very hard to learn for standardly used solvers. Using Keras (66), a neural network with two hidden layers was able to learn only approximately the rule, only for considerably larger training set sizes and a much larger number of iterations than GAMP.

sample complexity, i.e., how many samples are truly needed so that a given generalization error can be achieved. In our setting the situation is different: We can present samples $(y_\mu, \mathbf{\Phi}_\mu)$ to generic out-of-the-box classification algorithms and see how their performances compare with the information-theoretic optimal performance and to the one of the GAMP algorithm that is fine-tuned to the problem.

In Fig. 2 we present examples of state-of-the-art classification algorithms that are compared with our results. In Fig. 2, *Left* and *Center* we compare the optimal and GAMP performances to a simple logistic regression, fine-tuned by manually optimizing the ridge penalty (for ℓ_2 regularization) and LASSO penalty (for a sparsity-enhancing ℓ_1 regularization) with the software scikit-learn (65). We observe that for the GB case the logistic regression is comparable to the performance of GAMP, whereas for binary weights perfect generalization is not achieved close to the GAMP phase transition.

In Fig. 2, *Right* we study classification for labels generated by the symmetric door channel. A general-purpose algorithm would not know about the form of the channel. A neural network with only two hidden units is in principle able to represent the corresponding function (each of the hidden neurons can learn one of the two planes that separate data in the symmetric door function). A more intriguing question is whether a more generic multilayer neural network is indeed able to learn this rule and how many samples it may need. In the example used in Fig. 2, using the software Keras (66) with a tensorflow backend, we show the performance of a network with two hidden layers, ReLU activation and dropout [the details for this particular run can be found in the github repository (62)]. The symmetric door function thus provides a challenging benchmark that could be used to study how to improve performance of the general-purpose multilayer neural network classifiers. In *SI Appendix* we provide additional examples comparing the optimal performance to general-purpose algorithms for regression.

Methods and Proofs

In this section we give the main theorem for the free entropy and main ideas of the proof. An essential tool is the adaptive interpolation method recently introduced in ref. 52 which is a powerful evolution of the Guerra and Toninelli (72) interpolation method developed for spin glasses. Ref. 52 analyzed simpler inference problems. In particular, the proof for the upper bound in ref. 52 does not apply to GLMs and requires nontrivial additional ingredients. One such additional ingredient is to work with a potential $f_{\mathrm{RS}}(q, r; \rho)$ depending on two parameters (q, r) instead of a single one as in ref. 52. This allows us to use convexity arguments that are crucial to finish the proof, discussed in *Matching Bounds and End of Proof*. We stress that the present analysis heavily relies on properties of Bayes-optimal inference that translate into remarkable identities between correlation functions (called Nishimori identities by physicists; see *SI Appendix* for their formulation) valid for all values of parameters. These identities are used in the derivation of **[17]** and **[18]** below, which are two essential steps of our proof. The formula from *Theorem 1* relies on the Nishimori identities and does not hold out of the Bayes-optimal setting.

Main Theorems. For the proof it is necessary to work with a slightly different model with an additive regularizing Gaussian noise with variance $\Delta \geq 0$,

$$Y_\mu = \varphi\left(\frac{1}{\sqrt{n}}[\mathbf{\Phi X^*}]_\mu, A_\mu\right) + \sqrt{\Delta} Z_\mu, \qquad 1 \leq \mu \leq m, \qquad \textbf{[12]}$$

where $(Z_\mu) \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, and $(A_\mu) \overset{\text{iid}}{\sim} P_A$ are r.v. that represent the stochastic part of φ. It is also instrumental to think of the measurements as the outputs of a "channel" $Y_\mu \sim P_{\mathrm{out}}(\cdot | \frac{1}{\sqrt{n}}[\mathbf{\Phi X^*}]_\mu)$ with transition density $P_{\mathrm{out}}(y|z) = (2\pi\Delta)^{-1/2} \int dP_A(a) \exp\{-\frac{1}{2\Delta}(y - \varphi(z,a))^2\}$ if $\Delta > 0$, or $P_{\mathrm{out}}(y|z) = \int dP_A(a) \mathbf{1}(y = \varphi(z,a))$ else, where $\mathbf{1}(\cdot)$ is the indicator function. Our main theorem holds under the following rather general hypotheses:

h1) The prior distribution P_0 admits a finite third moment and has at least two points in its support.
h2) The sequence $(\mathbb{E}[|\varphi(\frac{1}{\sqrt{n}}[\mathbf{\Phi X^*}]_1, \mathbf{A}_1)|^{2+\gamma}])_{n \geq 1}$ is bounded for some $\gamma > 0$.
h3) The r.v. $(\Phi_{\mu i})$ are independent with zero mean, unit variance, and finite third moment bounded with n.
h4) For almost all values of a (w.r.t. the distribution P_A), the function $x \mapsto \varphi(x, a)$ is continuous almost everywhere.
h5) $(\Delta > 0)$ or $(\Delta = 0$ and φ takes values in $\mathbb{N})$.

In general, when φ is continuous, the condition $\Delta > 0$ (but arbitrarily small) is necessary for the existence of a finite limit of the free entropy [for particular choices of (φ, P_A) this might not be needed, e.g., $\varphi(z, A) = z + A$ with $A \sim \mathcal{N}(0, \sigma^2)$]. We also assume that the kernel P_{out} is informative; i.e., there exists y such that $P_{\mathrm{out}}(y\,|\,\cdot)$ is not equal almost everywhere to a constant. If P_{out} is not informative, it is not difficult to show that estimation is then impossible.

We define the set of the critical points of f_{RS}, **[4]**, also called "state evolution fixed points" (as is clear from **[10]**):

$$\Gamma \equiv \left\{ (q,r) \in [0,\rho] \times (\mathbb{R}_+ \cup \{+\infty\}) \;\middle|\; \begin{array}{l} q = 2\psi'_{P_0}(r) \\ r = 2\alpha\Psi'_{P_{\mathrm{out}}}(q;\rho) \end{array} \right\}.$$

Define $f_n \equiv \mathbb{E} f_n(\mathbf{Y}, \mathbf{\Phi}) = \frac{1}{n}\mathbb{E} \ln \mathcal{Z}(\mathbf{Y}, \mathbf{\Phi})$. Then the main theorem of this paper is stated as follows:

Theorem 1 (Replica-Symmetric Free Entropy). *Suppose that* (*h*1)–(*h*2)–(*h*3)–(*h*4)–(*h*5) *hold. Then, for the GLM* **12**,

$$\lim_{n\to\infty} f_n = \sup_{q \in [0,\rho]} \inf_{r \geq 0} f_{\mathrm{RS}}(q,r) = \sup_{(q,r)\in\Gamma} f_{\mathrm{RS}}(q,r)\,.$$

Moreover, as one can see in *SI Appendix*, the "sup inf" and the supremum over Γ above are achieved over the same couples. Under stronger assumptions on P_0 and P_{out}, one can show (*Theorem 6* in *SI Appendix*) that $f_n(\mathbf{Y}, \mathbf{\Phi})$ concentrates around its mean f_n and thus obtains convergence in probability **3**.

An immediate corollary of *Theorem 1* is the limiting expression of the mutual information $I(\mathbf{X^*}; \mathbf{Y}|\mathbf{\Phi}) \equiv \mathbb{E} \ln P(\mathbf{Y}, \mathbf{X^*}|\mathbf{\Phi}) - \mathbb{E} \ln(P(\mathbf{Y}|\mathbf{\Phi})P(\mathbf{X^*}))$ between the observations and the unknown vector:

Corollary 1 (Mutual Information). *Under the same hypotheses as in Theorem 1, the mutual information for the GLM* **12** *verifies*

$$\lim_{n\to\infty} \frac{1}{n} I(\mathbf{X^*}; \mathbf{Y}\,|\mathbf{\Phi}) = \inf_{q\in[0,\rho]} \sup_{r\geq 0} i_{\mathrm{RS}}(q,r) = \inf_{(q,r)\in\Gamma} i_{\mathrm{RS}}(q,r)\,,$$

$$i_{\mathrm{RS}}(q,r) \equiv \alpha\Psi_{P_{\mathrm{out}}}(\rho;\rho) - \alpha\Psi_{P_{\mathrm{out}}}(q;\rho) - \psi_{P_0}(r) + rq/2\,.$$

Finally, we gather our main results related to the optimal errors in a single theorem (see *SI Appendix* for more details), including results on the optimality of the GAMP algorithm:

Theorem 2 (Optimal Errors). *Assume the same hypotheses as in Theorem 1. Then formula* **9** *for the generalization error is true as* $n, m \to \infty$, $m/n \to \alpha$ *whenever the maximizer* $q^*(\alpha)$ *of* **[3]** *is unique, which is the case for almost every* α. *If moreover all of the moments of* P_0 *are finite, then formula* **7** *for the overlap and the matrix-MMSE formula*

$$\frac{1}{n^2}\mathbb{E}[\|\mathbf{X^*X^{*\intercal}} - \mathbb{E}_{P(\mathbf{x}|\mathbf{Y},\mathbf{\Phi})}[\mathbf{xx^\intercal}]\|_{\mathrm{F}}^2] \to \rho^2 - q^*(\alpha)^2 \qquad \textbf{[13]}$$

are true, where $\| - \|_{\mathrm{F}}$ *is the Frobenius norm.*

There are cases of GLMs (e.g., the signless output channel $\mathbf{Y} = |\mathbf{\Phi X^*}|/\sqrt{n} + \mathbf{Z}$) where the sign of $\mathbf{X^*}$ simply cannot be estimated (thus the absolute value in **[7]**). This is why our general theorem is related to an error metric **13** insensitive to this ± symmetry. Nevertheless formula **8** for the signal MSE is formally valid when there is no such sign symmetry.

Sketch of Proof by the Adaptive Interpolation Method. We now give the main ideas behind the proof of *Theorem 1*. We defer to *SI Appendix* the details, as well as those of *Corollary 1* and *Theorem 2*.

We note a clarification about notation. The r.v. $\mathbf{Y}$ (and also $\mathbf{\Phi}$, $\mathbf{X^*}$, $\mathbf{A}$, and $\mathbf{Z}$) are called quenched because once the measurements are acquired, they are fixed. The expectation w.r.t. all quenched r.v. is denoted by $\mathbb{E}$ without a subscript. In contrast, expectation of annealed variables w.r.t. a posterior distribution at fixed quenched variables is denoted by Gibbs brackets $\langle - \rangle$.

Two scalar inference channels. An important role in the proof is played by two simple scalar inference channels. The free entropy is expressed in terms of the free entropies of these channels. This "decoupling property" stands at the root of the replica approach in statistical physics.

The first scalar channel is an additive Gaussian channel. Suppose that we observe $Y_0 = \sqrt{r}\, X_0 + Z_0$ where $X_0 \sim P_0$ and $Z_0 \sim \mathcal{N}(0, 1)$ are independent. Consider the inference problem consisting of retrieving X_0 from the observation Y_0. The free entropy associated with this channel is the expectation of the logarithm of the normalization factor of the associated posterior $dP(x|Y_0)$ that is given by **[5]** (up to a constant).

The second scalar channel that appears naturally in the problem is linked to the channel $P_{\rm out}$ through the following inference model. Suppose that $V, W^* \overset{\rm iid}{\sim} \mathcal{N}(0, 1)$ where V is known while the inference problem is to recover the unknown W^* from the observation $\widetilde{Y}_0 \sim P_{\rm out}(\cdot\,|\sqrt{q}\, V + \sqrt{\rho - q}\, W^*)$ where $\rho > 0$ and $q \in [0, \rho]$. The free entropy for this model, again given by a normalization factor of the posterior of w given $\widetilde{Y}_0$ and V, is exactly **[6]**.

Interpolating the estimation problem. To carry out the proof, we introduce an "interpolating estimation problem" that interpolates between the original problem $Y_\mu \sim P_{\rm out}(\cdot|\frac{1}{\sqrt{n}}[\mathbf{\Phi X^*}]_\mu)$ at $t = 0$, with $t \in [0, 1]$ being the interpolation parameter, and the two scalar problems described above at $t = 1$. For $t \in (0, 1)$ the interpolating estimation problem is a mixture of the original and the scalar problems. This interpolation scheme is inspired by the interpolation paths used by Talagrand (73) to study the perceptron. Due to a novel ingredient specific to the adaptive interpolation method (52), it allows us to obtain in a unified manner a complete proof of the replica formula for the free entropy and in the whole phase diagram.

Let $q(t)$ and $r(t)$ be two interpolation functions. Moreover define $S_{t,\mu} = S_{t,\mu}(\mathbf{X}^*, W_\mu^*, V_\mu, \mathbf{\Phi})$ as

$$S_{t,\mu} \equiv \sqrt{\tfrac{1-t}{n}}\, [\mathbf{\Phi X^*}]_\mu + \sqrt{\textstyle\int_0^t q(v)dv}\, V_\mu + \sqrt{\textstyle\int_0^t (\rho - q(v))dv}\, W_\mu^*,$$

where $V_\mu, W_\mu^* \overset{\rm iid}{\sim} \mathcal{N}(0, 1)$. Consider the following observation channels, with two types of observations obtained through

$$\begin{cases} Y_{t,\mu} & \sim \quad P_{\rm out}(\,\cdot\,|\,S_{t,\mu})\,, \quad \text{for } 1 \le \mu \le m, \\ Y'_{t,i} & = \quad \sqrt{\int_0^t r(v)dv}\, X_i^* + Z'_i\,, \quad \text{for } 1 \le i \le n, \end{cases} \qquad \textbf{[14]}$$

where $(Z'_i) \overset{\rm iid}{\sim} \mathcal{N}(0, 1)$. We assume that $\mathbf{V} = (V_\mu)_{\mu=1}^m$ is known and that the inference problem is to recover both $\mathbf{W}^* = (W_\mu^*)_{\mu=1}^m$ and $\mathbf{X}^* = (X_i^*)_{i=1}^n$ from the "t-dependent" observations $\mathbf{Y}_t = (Y_{t,\mu})_{\mu=1}^m$ and $\mathbf{Y}'_t = (Y'_{t,i})_{i=1}^n$.

We now understand that the integral of $r(t)$ appearing in the second set of measurements in **[14]** and $1 - t$ as well as the two integrals appearing in the first set all play the role of signal-to-noise ratios (SNRs) in the interpolating problem, with t giving more and more "power" (or weight) to the scalar inference channels when increasing. Here is the first crucial ingredient of our interpolation scheme. In classical interpolations, these SNRs would all take a trivial form, i.e., be linear in t, but here, the nontrivial integral dependency in t of the two latter SNRs allows for much more flexibility when choosing the interpolation path. This will allow us to actually choose the "optimal interpolation path" (this will become clear below).

Define $u_y(x) \equiv \ln P_{\rm out}(y|x)$ and, with a slight abuse of notations, $s_{t,\mu} = s_{t,\mu}(\mathbf{x}, w_\mu, V_\mu, \mathbf{\Phi}) \equiv S_{t,\mu}(\mathbf{x}, w_\mu, V_\mu, \mathbf{\Phi})$, the expression above with $\mathbf{X}^*, W_\mu^*$ replaced by $\mathbf{x}, w_\mu$. We introduce the interpolating Hamiltonian $\mathcal{H}_t = \mathcal{H}_t(\mathbf{x}, \mathbf{w}; \mathbf{Y}_t, \mathbf{Y}'_t, \mathbf{\Phi}, \mathbf{V})$

$$\mathcal{H}_t \equiv -\sum_{\mu=1}^m u_{Y_{t,\mu}}(s_{t,\mu}) + \frac{1}{2}\sum_{i=1}^n \Big(Y'_{t,i} - \sqrt{\textstyle\int_0^t r(v)dv}\, x_i\Big)^2$$

and the corresponding (t-dependent) Gibbs bracket $\langle - \rangle_t$ which is the expectation w.r.t. the joint posterior distribution of $(\mathbf{x}, \mathbf{w})$ given the observations $\mathbf{Y}_t, \mathbf{Y}'_t$ (and $\mathbf{\Phi}, \mathbf{V}$), defined as

$$\langle L(\mathbf{x}, \mathbf{w})\rangle_t \equiv \mathcal{Z}_t(\mathbf{Y}_t, \mathbf{Y}'_t, \mathbf{\Phi}, \mathbf{V})^{-1} \int dP_0(\mathbf{x})\mathcal{D}\mathbf{w} L(\mathbf{x}, \mathbf{w}) e^{-\mathcal{H}_t}\,,$$

for every continuous bounded test function L. Here $\mathcal{Z}_t \equiv \int dP_0(\mathbf{x})\mathcal{D}\mathbf{w} \exp\{-\mathcal{H}_t(\mathbf{x}, \mathbf{w}; \mathbf{Y}_t, \mathbf{Y}'_t, \mathbf{\Phi}, \mathbf{V})\}$ is the appropriate normalization, and $\mathcal{D}\mathbf{w}$ is the standard Gaussian measure. Finally we introduce

$$f_n(t) \equiv \frac{1}{n}\mathbb{E} \ln \mathcal{Z}_t(\mathbf{Y}, \mathbf{Y}', \mathbf{\Phi}, \mathbf{V})$$

which is the interpolating free entropy. One verifies easily that

$$\begin{cases} f_n(0) & = \quad f_n - \frac{1}{2}\,, \\ f_n(1) & = \quad \psi_{P_0}(\int_0^1 r(t)dt) - \frac{1 + \int_0^1 r(t)dt\,\rho}{2} + \frac{m}{n}\Psi_{P_{\rm out}}(\int_0^1 q(t)dt; \rho)\,. \end{cases} \qquad \textbf{[15]}$$

Now comes another crucial property of the interpolating model: It is such that at $t = 0$ we recover the original problem and thus $f_n(0) = f_n - 1/2$ (the constant 1/2 comes from the purely noisy measurements of the second channel in **[14]**), while at $t = 1$ we have the two scalar inference channels and thus the associated terms ψ_{P_0} and $\Psi_{P_{\rm out}}$ appear in $f_n(1)$. These are precisely the terms appearing in the free entropy potential **4**.

Entropy variation along the interpolation. From the understanding of the previous section, it is natural to evaluate the variation of entropy along the interpolation, which allows us to "compare" the original and purely scalar models due to the identity

$$f_n = f_n(0) + \tfrac{1}{2} = f_n(1) - \textstyle\int_0^1 f'_n(t) + \tfrac{1}{2}\,, \qquad \textbf{[16]}$$

where the first equality follows from **[15]** (the prime means the derivative). Then by choosing the optimal interpolation path due to the nontrivial SNR dependencies in t, we will be able to show the equality between the replica formula and the true free entropy f_n.

We thus compute the t derivative of the free entropy (see *SI Appendix* for the details of this calculation). It is given by

$$f'_n(t) = \frac{r(t)q(t)}{2} - \frac{r(t)\rho}{2} + \mathcal{O}_n(1)$$

$$-\frac{1}{2}\mathbb{E}\Big\langle \Big(\frac{1}{n}\sum_{\mu=1}^m u'_{Y_{t,\mu}}(S_{t,\mu}) u'_{Y_{t,\mu}}(s_{t,\mu}) - r(t)\Big)(Q - q(t))\Big\rangle_t, \qquad \textbf{[17]}$$

where $\mathcal{O}_n(1)$ is a quantity that goes to 0 in the $n, m \to \infty$ limit, uniformly in t, and the overlap is $Q = Q_n \equiv n^{-1}\sum_{i=1}^n X_i^* x_i$.

We now state a crucial result in an informal way and refer to *SI Appendix* for precise statements. Formally, the overlap concentrates around its mean (for all $t \in [0, 1]$), a behavior called "replica-symmetric" in statistical physics. To make this statement mathematically rigorous, one has to slightly modify the interpolating model by adding a "side channel" that brings vanishingly small additional information about $\mathbf{X}^*$ without affecting the asymptotic free entropy density. This perturbation forces the overlap to concentrate. Effectively, one can use the following formal formula (see *SI Appendix*, section 4.3, *Lemma 2* for a precise statement):

$${\rm Var}_t(Q) = \mathbb{E}\langle\, (Q - \mathbb{E}\langle Q\rangle_t)^2\rangle_t = \mathcal{O}_n(1)\,. \qquad \textbf{[18]}$$

Canceling the remainder. Note from **[15]** and **[4]** that the first two terms appearing in **[17]** are precisely the missing ones to obtain the expression of the potential on the r.h.s. of **[16]**. Thus, we want to "cancel" the Gibbs bracket in **[17]**. This term is called the remainder. To prove the replica formula, we have to show that this remainder vanishes, which was until now a difficult task. But due to the freedom of choice of the interpolation path allowed by the interpolating function q, we are able to do so by "adapting" the interpolation (thus the name of the method). Thus, we want to choose $q(t) = \mathbb{E}\,\langle Q\rangle_t \approx Q$ because of **[18]**. However, $\mathbb{E}\,\langle Q\rangle_t$ is a function of $\int_0^t q(v)dv$. The equation $q(t) = \mathbb{E}\,\langle Q\rangle_t \in [0, \rho]$ is therefore an order 1 differential equation over $t \mapsto \int_0^t q(v)dv$. Assume for the moment that this equation has a solution over [0,1]. Once the solution $q_n^{(r)}$ is selected, the Cauchy–Schwarz inequality applied to the remainder allows us to show that its absolute value is upper bounded by $C\sqrt{{\rm Var}_t(Q)}$ for some constant $C > 0$ independent of n and t. Therefore from **[17]** and **[18]**, for $0 \le t \le 1$ we get

$$f'_n(t) = \tfrac{r(t)}{2} q_n^{(r)}(t) - \tfrac{r(t)\rho}{2} + \mathcal{O}_n(1)\,.$$

Finally combining this with **[15]** and **[16]** leads to

$$f_n = \psi_{P_0}(\textstyle\int_0^1 r(t)dt) + \frac{m}{n}\Psi_{P_{\rm out}}(\int_0^1 q_n^{(r)}(t)dt; \rho) - \frac{1}{2}\int_0^1 r(t) q_n^{(r)}(t)dt + \mathcal{O}_n(1)\,. \qquad \textbf{[19]}$$

This important equality is obtained due to the choice of the optimal interpolation path $q_n^{(r)}(t)$ permitted by the method.

Matching bounds and end of proof. We now possess all of the necessary tools to finish the sketch of the proof of *Theorem 1*. We first prove that $\lim_{n\to\infty} f_n = \sup_{r\ge 0}\inf_{q\in[0,\rho]} f_{\rm RS}(q, r)$. Then in *SI Appendix*, we show that (*i*) this is also equal to $\sup_{q\in[0,\rho]}\inf_{r\ge 0} f_{\rm RS}(q, r)$, which gives the first equality of the theorem, and (*ii*) that this sup inf is attained at the supremum of the state evolution fixed points, which gives the second equality.

Lower bound. Choose *r(t)* = *r* the constant function. Identity **19** implies $\liminf_{n\to\infty} f_n \geq \inf_{q\in[0,\rho]} f_{\mathrm{RS}}(q,r)$. This is true for all $r \geq 0$ and thus

$$\liminf_{n\to\infty} f_n \geq \sup_{r\geq 0} \inf_{q\in[0,\rho]} f_{\mathrm{RS}}(q,r)\,. \qquad \mathbf{[20]}$$

Upper bound. We show in *SI Appendix* that $\Psi'_{P_{\mathrm{out}}}$ is nonnegative, continuous, and bounded and thus we can define $K \equiv 2\alpha \max_{q\in[0,\rho]} \Psi'_{P_{\mathrm{out}}}(q;\rho) \in \mathbb{R}_+$. Consequently we can complete the general differential equation satisfied by *q(t)* by choosing *r(t)* as the solution of (see *SI Appendix* for more details)

$$r(t) = 2\alpha\Psi'_{P_{\mathrm{out}}}(\int_0^1 q_n^{(r)}(t)dt;\rho) \in [0,K]$$

In *SI Appendix* we show that a solution exists. Moreover from **[19]** and convexity of **[5]** and **[6]** we can assert

$$f_n \leq \int_0^1 f_{\mathrm{RS}}(q_n^{(r)}(t), r(t))dt + \mathcal{O}_n(1)\,.$$

Finally note that if we denote r_n^* the solution of the ODE

$$f_{\mathrm{RS}}(\int_0^1 q_n^{(r_n^*)}(t)dt, r_n^*) = \inf_{q\in[0,\rho]} f_{\mathrm{RS}}(q, r_n^*)\,.$$

Indeed, the function $g_{r_n^*}: q \in [0,\rho] \mapsto f_{\mathrm{RS}}(q, r_n^*)$ is convex (*SI Appendix*) and its derivative is $g'_{r_n^*}(q) = \alpha\Psi'_{P_{\mathrm{out}}}(q) - r_n^*/2$. Since $g'_{r_n^*}(\int_0^1 q_n^{(r_n^*)}(t)dt) = 0$ by definition of r_n^*, the minimum of $g_{r_n^*}$ is necessarily achieved at $\int_0^1 q_n^{(r_n^*)}(t)dt$. We thus have

$$\limsup_{n\to\infty} f_n \leq \sup_{r\geq 0} \inf_{q\in[0,\rho]} f_{\mathrm{RS}}(q,r)$$

which, when combined with **[20]**, allows us to deduce the result

$$\lim_{n\to\infty} f_n = \sup_{r\geq 0} \inf_{q\in[0,\rho]} f_{\mathrm{RS}}(q,r)\,.$$

ACKNOWLEDGMENTS. F.K. and L.Z. acknowledge funding from the European Research Council under the European Union's 7th Framework Program Grant Agreement 307087-SPARCS and under Horizon 2020 Research and Innovation Programme Grant Agreement 714608-SMile. F.K. and N.M. acknowledge support from the PAIL bilateral grant of Agence Nationale de la Recherche and Swiss National Foundation for Science (200021E-175541).

1. Nelder J, Wedderburn R (1972) Generalized linear models. *J R Stat Soc Ser A* 135:370–384.
2. McCullagh P (1984) Generalized linear models. *Eur J Oper Res* 16:285–292.
3. Fienup JR (1982) Phase retrieval algorithms: A comparison. *Appl Opt* 21:2758–2769.
4. Demanet L, Hand P (2014) Stable optimizationless recovery from phaseless linear measurements. *J Fourier Anal Appl* 20:199–221.
5. Candes EJ, Strohmer T, Voroninski V (2013) Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Commun Pure Appl Math* 66:1241–1274.
6. Boufounos PT, Baraniuk RG (2008) 1-bit compressive sensing. *42nd Annual Conference on Information Sciences and Systems (CISS)* (IEEE, Piscataway, NJ), pp 16–21.
7. Bühlmann P, Van De Geer S (2011) *Statistics for High-Dimensional Data: Methods, Theory and Applications* (Springer, Berlin).
8. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521:436–444.
9. Donoho DL, Tanner J (2005) Sparse nonnegative solution of underdetermined linear equations by linear programming. *Proc Natl Acad Sci USA* 102:9446–9451.
10. Candes EJ, Tao T (2006) Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans Inf Theory* 52:5406–5425.
11. Donoho DL, Maleki A, Montanari A (2009) Message-passing algorithms for compressed sensing. *Proc Natl Acad Sci USA* 106:18914–18919.
12. Rangan S (2011) Generalized approximate message passing for estimation with random linear mixing. *IEEE International Symposium on Information Theory Proceedings (ISIT)*, eds Kuleshov A, Blinovsky VM, Ephremides A (IEEE, Piscataway, NJ), pp 2168–2172.
13. Zdeborová L, Krzakala F (2016) Statistical physics of inference: Thresholds and algorithms. *Adv Phys* 65:453–552.
14. Kamilov U, Goyal VK, Rangan S (2011) Optimal quantization for compressive sensing under message passing reconstruction. *IEEE International Symposium on Information Theory Proceedings (ISIT)* (IEEE, Piscataway, NJ), pp 459–463.
15. Xu Y, Kabashima Y, Zdeborová L (2014) Bayesian signal reconstruction for 1-bit compressed sensing. *J Stat Mech Theory Exp* 2014:P11015.
16. Schniter P, Rangan S (2015) Compressive phase retrieval via generalized approximate message passing. *IEEE Trans Signal Process* 63:1043–1055.
17. Bayati M, Montanari A (2012) The lasso risk for Gaussian matrices. *IEEE Trans Inf Theory* 58:1997–2017.
18. El Karoui N, Bean D, Bickel PJ, Lim C, Yu B (2013) On robust regression with high-dimensional predictors. *Proc Natl Acad Sci USA* 110:14557–14562.
19. Donoho D, Montanari A (2016) High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probab Theory Relat Fields* 166:935–969.
20. Gribonval R, Machart P (2013) Reconciling "priors" & "priors" without prejudice? *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems*, eds Burges CJC, Bottou L, Ghahramani Z, Weinberger KQ (Neural Information Processing Systems Foundation, La Jolla, CA). Available at https://papers.nips.cc/book/advances-in-neural-information-processing-systems-26-2013. Accessed February 20, 2019.
21. Advani M, Ganguli S (2016) An equivalence between high dimensional Bayes optimal inference and m-estimation. *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems*, eds Lee DD, Sugiyama M, von Luxburg U, Guyon I, Garnett R (Neural Information Processing Systems Foundation, La Jolla, CA). Available at https://papers.nips.cc/book/advances-in-neural-information-processing-systems-29-2016. Accessed February 20, 2019.
22. Gardner E, Derrida B (1989) Three unfinished works on the optimal storage capacity of networks. *J Phys A Math Gen* 22:1983–1994.
23. Seung HS, Sompolinsky H, Tishby N (1992) Statistical mechanics of learning from examples. *Phys Rev A* 45:6056–6091.
24. Watkin TLH, Rau A, Biehl M (1993) The statistical mechanics of learning a rule. *Rev Mod Phys* 65:499–556.
25. Engel A, Van den Broeck C (2001) *Statistical Mechanics of Learning* (Cambridge Univ Press, New York).
26. Engel A, Reimers L (1994) Reliability of replica symmetry for the generalization problem in a toy multilayer neural network. *Europhys Lett* 28:531–536.
27. Bex GJ, Serneels R, den Broeck CV (1995) Storage capacity and generalization error for the reversed-wedge Ising perceptron. *Phys Rev E* 51:6309–6312.
28. Hosaka T, Kabashima Y, Nishimori H (2002) Statistical mechanics of lossy data compression using a nonmonotonic perceptron. *Phys Rev E* 66:066126.
29. Baldassi C, et al. (2016) Unreasonable effectiveness of learning neural networks: From accessible states and robust ensembles to basic algorithmic schemes. *Proc Natl Acad Sci USA* 113:E7655–E7662.
30. Martin CH, Mahoney MW (2017) Rethinking generalization requires revisiting old ideas: Statistical mechanics approaches and complex learning behavior. arXiv:1710.09553. Preprint, posted October 26, 2017.
31. Tishby N, Pereira FC, Bialek W (1999) The information bottleneck method. *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing* (Univ of Illinois, Champaign, IL), pp 368–377.
32. Shwartz-Ziv R, Tishby N (2017) Opening the black box of deep neural networks via information. arXiv:1703.00810. Preprint, posted March 2, 2017.
33. Shannon CE (1948) A mathematical theory of communication, part i, part ii. *Bell Syst Tech J* 27:623–656.
34. Tanaka T (2002) A statistical-mechanics approach to large-system analysis of CDMA multiuser detectors. *IEEE Trans Inf Theory* 48:2888–2910.
35. Guo D, Verdú S (2005) Randomly spread CDMA: Asymptotics via statistical physics. *IEEE Trans Inf Theory* 51:1983–2010.
36. Barron AR, Joseph A (2010) Toward fast reliable communication at rates near capacity with Gaussian noise. *IEEE International Symposium on Information Theory (ISIT)* (IEEE, Piscataway, NJ), pp 315–319.
37. Barbier J, Krzakala F (2017) Approximate message-passing decoder and capacity-achieving sparse superposition codes. *IEEE Trans Inf Theory* 63:4894–4927.
38. Rush C, Greig A, Venkataramanan R (2017) Capacity-achieving sparse superposition codes via approximate message passing decoding. *IEEE Trans Inf Theory* 63:1476–1500.
39. Barbier J, Dia M, Macris N (2016) Threshold saturation of spatially coupled sparse superposition codes for all memoryless channels. *IEEE Information Theory Workshop (ITW)* (IEEE, Piscataway, NJ), pp 76–80.
40. Barbier J, Dia M, Macris N (2017) Universal sparse superposition codes with spatial coupling and GAMP decoding. arXiv:1707.04203. Preprint, posted July 13, 2017.
41. Mézard M (1989) The space of interactions in neural networks: Gardner's computation with the cavity method. *J Phys A Math Gen* 22:2181–2190.
42. Bolthausen E (2014) An iterative construction of solutions of the TAP equations for the Sherrington–Kirkpatrick model. *Commun Math Phys* 325:333–366.
43. Bayati M, Montanari A (2011) The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Trans Inf Theory* 57:764–785.
44. Bayati M, Lelarge M, Montanari A (2015) Universality in polytope phase transitions and message passing algorithms. *Ann Appl Probab* 25:753–822.
45. Javanmard A, Montanari A (2013) State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Inf Inference* 2:115–144.
46. Barbier J, Dia M, Macris N, Krzakala F (2016) The mutual information in random linear estimation. *54th Annual Allerton Conference on Communication, Control, and Computing*, eds Do M, Hovakimyan N (IEEE, Piscataway, NJ), pp 625–632.
47. Barbier J, Macris N, Dia M, Krzakala F (2017) Mutual information and optimality of approximate message-passing in random linear estimation. arXiv:1701.05823. Preprint, posted January 20, 2017.
48. Reeves G, Pfister HD (2016) The replica-symmetric prediction for compressed sensing with Gaussian matrices is exact. *IEEE International Symposium on Information Theory (ISIT)* (IEEE, Piscataway, NJ), pp 665–669.

49. Mézard M, Parisi G, Virasoro MA (1987) *Spin Glass Theory and Beyond* (World Scientific, Singapore).
50. Györgyi G (1990) First-order transition to perfect generalization in a neural network with binary synapses. *Phys Rev A* 41:7097–7100.
51. Sompolinsky H, Tishby N, Seung HS (1990) Learning from examples in large neural networks. *Phys Rev Lett* 65:1683–1686.
52. Barbier J, Macris N (2017) The adaptive interpolation method: A simple scheme to prove replica formulas in Bayesian inference. arXiv:1705.02780. Preprint, posted May 8, 2017.
53. Coja-Oghlan A, Krzakala F, Perkins W, Zdeborova L (2017) Information-theoretic thresholds from the cavity method. *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, eds Hatami H, McKenzie P, King V (Association for Computing Machinery, New York), pp 146–157.
54. Guo D, Shamai S, Verdú S (2005) Mutual information and minimum mean-square error in Gaussian channels. *IEEE Trans Inf Theory* 51:1261–1282.
55. Opper M, Haussler D (1991) Generalization performance of Bayes optimal classification algorithm for learning a perceptron. *Phys Rev Lett* 66:2677–2680.
56. Thouless DJ, Anderson PW, Palmer RG (1977) Solution of 'solvable model of a spin glass'. *Philos Mag* 35:593–601.
57. Kabashima Y (2008) Inference from correlated patterns: A unified theory for perceptron learning and linear vector channels. *J Phys Conf Ser* 95:012001.
58. Donoho DL, Javanmard A, Montanari A (2013) Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing. *IEEE Trans Inf Theory* 59:7434–7464.
59. Opper M, Winther O (1996) Mean field approach to Bayes learning in feed-forward neural networks. *Phys Rev Lett* 76:1964–1967.
60. Opper M, Winther O (2001) Tractable approximations for probabilistic models: The adaptive Thouless-Anderson-Palmer mean field approach. *Phys Rev Lett* 86: 3695–3699.
61. Mondelli M, Montanari A (2017) Fundamental limits of weak recovery with applications to phase retrieval. arXiv:1708.05932. Preprint, posted August 20, 2017.
62. Barbier J, Krzakala F, Macris N, Miolane L, Zdeborová L (2017) Data from "GeneralizedLinearModel2017." Available at https://github.com/sphinxteam/GeneralizedLinearModel2017. Deposited October 27, 2017.
63. Fletcher AK, Rangan S (2018) Iterative reconstruction of rank-one matrices in noise. *Inf Inference* 7:531–562.
64. Hansel D, Mato G, Meunier C (1992) Memorization without generalization in a multilayered neural network. *Europhys Lett* 20:471–476.
65. Pedregosa F, et al. (2011) Scikit-learn: Machine learning in Python. *J Machine Learn Res* 12:2825–2830.
66. Chollet F (2015) Keras. Available at https://github.com/fchollet/keras. Accessed February 12, 2019.
67. Wu Y, Verdú S (2010) Rényi information dimension: Fundamental limits of almost lossless analog compression. *IEEE Trans Inf Theory* 56:3721–3748.
68. Krzakala F, Mézard M, Sausset F, Sun Y, Zdeborová L (2012) Statistical-physics-based reconstruction in compressed sensing. *Phys Rev X* 2:021005.
69. Maleki A, Anitori L, Yang Z, Baraniuk RG (2013) Asymptotic analysis of complex lasso via complex approximate message passing (CAMP). *IEEE Trans Inf Theory* 59:4290–4308.
70. Soltanolkotabi M (2017) Structured signal recovery from quadratic measurements: Breaking sample complexity barriers via nonconvex optimization. arXiv:1702.06175. Preprint, posted February 20, 2017.
71. Rosenblatt F (1957) The perceptron, a perceiving and recognizing automaton (Cornell Aeronautical Laboratory, Buffalo, NY), Project Para Report 85-460-1.
72. Guerra F, Toninelli FL (2002) The thermodynamic limit in mean field spin glass models. *Commun Math Phys* 230:71–79.
73. Talagrand M (2010) *Mean Field Models for Spin Glasses: Volume I: Basic Examples* (Springer, Berlin), Vol 54.