Statistical Mechanics of Optimal Convex Inference in High Dimensions

Madhu Advani* and Surya Ganguli†

Department of Applied Physics, Stanford University, Stanford, California 94305, USA

(Received 22 February 2016; revised manuscript received 7 June 2016; published 29 August 2016)

A fundamental problem in modern high-dimensional data analysis involves efficiently inferring a set of P unknown model parameters governing the relationship between the inputs and outputs of N noisy measurements. Various methods have been proposed to regress the outputs against the inputs to recover the P parameters. What are fundamental limits on the accuracy of regression, given finite signal-to-noise ratios, limited measurements, prior information, and computational tractability requirements? How can we optimally combine prior information with measurements to achieve these limits? Classical statistics gives incisive answers to these questions as the measurement density $\alpha = (N/P) \to \infty$. However, these classical results are not relevant to modern high-dimensional inference problems, which instead occur at finite α. We employ replica theory to answer these questions for a class of inference algorithms, known in the statistics literature as M-estimators. These algorithms attempt to recover the P model parameters by solving an optimization problem involving minimizing the sum of a loss function that penalizes deviations between the data and model predictions, and a regularizer that leverages prior information about model parameters. Widely cherished algorithms like maximum likelihood (ML) and maximum-*a posteriori* (MAP) inference arise as special cases of M-estimators. Our analysis uncovers fundamental limits on the inference accuracy of a subclass of M-estimators corresponding to computationally tractable convex optimization problems. These limits generalize classical statistical theorems like the Cramer-Rao bound to the high-dimensional setting with prior information. We further discover the optimal M-estimator for log-concave signal and noise distributions; we demonstrate that it can achieve our high-dimensional limits on inference accuracy, while ML and MAP cannot. Intriguingly, in high dimensions, these optimal algorithms become computationally simpler than ML and MAP while still outperforming them. For example, such optimal M-estimation algorithms can lead to as much as a 20% reduction in the amount of data to achieve the same performance relative to MAP. Moreover, we demonstrate a prediction of replica theory that no inference procedure whatsoever can outperform our optimal M-estimation procedure when signal and noise distributions are log-concave, by uncovering an equivalence between optimal M-estimation and optimal Bayesian inference in this setting. Our analysis also reveals insights into the nature of generalization and predictive power in high dimensions, information theoretic limits on compressed sensing, phase transitions in quadratic inference, and connections to central mathematical objects in convex optimization theory and random matrix theory.

DOI: 10.1103/PhysRevX.6.031034

Subject Areas: Complex Systems, Interdisciplinary Physics, Statistical Physics

I. INTRODUCTION

Remarkable advances in measurement technologies have thrust us squarely into the modern age of "big data," which yields the potential to revolutionize a variety of fields spanning the sciences, engineering, and humanities, including neuroscience [1,2], systems biology [3], health care [4], economics [5], social science [6], and history [7]. However, the advent of large-scale data sets presents severe statistical challenges that must be solved if we are to gain conceptual insights from such data.

A fundamental origin of the difficulty in analyzing many large-scale data sets lies in their high dimensionality [8–10]. For example, in classically designed experiments, we often measure a small number of P variables, chosen carefully ahead of time to test a specific hypothesis, and we take a large number of N measurements. Thus, the measurement density $\alpha = (N/P)$ is extremely large, and such data sets are *low* dimensional: They consist of a large number of N points in a low P dimensional space [Fig. 1(a)]. Much of the edifice of classical statistics operates within this low-dimensional, high-measurement density limit. Indeed, as reviewed below, as $\alpha \to \infty$, classical statistical theory gives us fundamental limits on the accuracy with which we can infer statistical models of

*msadvani@stanford.edu
†sganguli@stanford.edu

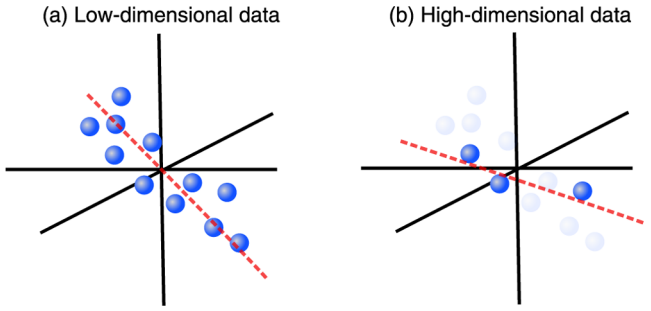

FIG. 1. A cartoon view of low-dimensional (a) versus high-dimensional (b) data. In the latter scenario, a finite measurement density, or ratio between data points and dimensions, leads to errors in inference.

such data, as well as the optimal statistical inference procedures to follow in order to achieve these limits.

In contrast to this classical scenario, our technological capacity for high-throughput measurements has led to a dramatic cultural shift in modern experimental design across many fields. We now often simultaneously measure many variables at once in advance of choosing any specific hypothesis to test. However, we may have limited time or resources to conduct such experiments, so we can only make a limited number of such simultaneous measurements. For example, through multielectrode recordings, we can simultaneously measure the activity $P = 1000$ neurons in mammalian circuits but only for $N = O(100)$ trials of any given trial type. Through microarrays, we can simultaneously measure the expression levels of $P = O(6000)$ genes in yeast but again in a limited number of $N = O(100)$ experimental conditions. Thus, while both N and P are large, the measurement density α is finite. Such data sets are *high* dimensional, in that they consist of a small number of points in a high-dimensional space [Fig. 1(b)], and it can be extremely challenging to detect regularities in such data [10]. Moreover, classical statistical theory gives no prescriptions for how to optimally analyze such data.

In our work, we focus on one of the most ubiquitous statistical inference procedures: regression, which attempts to find a linear relationship between a cloud of data points and another variable of interest. In order to study regression in the high-dimensional regime, we apply the technique of replica theory [11] from statistical physics. Indeed, replica theory has long played an important role in the analysis of high-dimensional statistical inference problems where the number of measurements or constraints is proportional to the number of unknowns, for example, in neural network memory capacity [12], perceptron learning theory [13,14], communication theory [15], compressed sensing [16–19], and most recently matrix factorization [20]. See also [10,21] for general reviews on replica theory in high-dimensional inference problems.

By applying replica theory to the central problem of high-dimensional regression, we obtain fundamental generalizations of statistical theorems dating back to the 1940s [22,23]. These theorems (reviewed below) place general limits on the accuracy of statistical inference through a set of procedures known as M-estimators (defined below, and see Refs. [24,25] for reviews) in a low-dimensional setting and reveal the optimal M-estimator (maximum likelihood estimation). We generalize these results to the high-dimensional setting with prior information, by (1) characterizing the performance of any convex regularized M-estimator on any high-dimensional regression problem, (2) finding the optimal convex M-estimator that achieves the best performance amongst all M-estimators, under the condition of log-concave signal and noise distributions, and (3) demonstrating that no inference algorithm whatsoever can outperform our optimal M-estimator in the setting where the prior distribution over parameters is known. Overall, our results reveal new optimal regression algorithms and quantitative insights into how the predictive power, or generalization capability, of a regression algorithm is related to its accuracy in separating signal from noise. Moreover, a variety of topics—including random matrix theory, compressed sensing, and fundamental objects in convex optimization theory, such as proximal mappings and Moreau envelopes—emerge naturally through our analysis. We give an intuitive summary of our results in the discussion section.

A. Statistical inference framework

To more concretely introduce this work, we give a precise definition of the inference problem we are studying. Formally, let $\mathbf{s}^0$ be an unknown P-dimensional vector governing the linear response of a system's scalar output y to a P-dimensional input $\mathbf{x}$ through the relation $y = \mathbf{x} \cdot \mathbf{s}^0 + \epsilon$, where ϵ denotes noise originating either from unobserved inputs or imperfect measurements. For example, in sensory neuroscience, y could reflect a linear approximation of the response of a single neuron to a sensory stimulus $\mathbf{x}$, so $\mathbf{s}^0$ is the neuron's receptive field. Alternatively, in genetic networks, y could reflect the linear response of one gene to the expression levels $\mathbf{x}$ of a set of P genes. Suppose we perform N measurements, indexed by $\mu = 1, \ldots, N$, in which we probe the system with an input $\mathbf{x}^\mu$ and record the resulting output y^μ. This yields a set of noisy measurements constraining the linear response vector $\mathbf{s}^0$ through the N equations $y^\mu = \mathbf{x}^\mu \cdot \mathbf{s}^0 + \epsilon^\mu$.

We assume the noise ϵ^μ and components s_i^0 are each drawn independently and identically distributed (i.i.d.) from a zero mean noise density $P_\epsilon(\epsilon)$ and a prior distribution $P_s(s)$. For convenience, below we define signal and noise energies in terms of the minus log probability of their respective distributions: $E_\epsilon = -\log P_\epsilon$ and $E_s = -\log P_s$. We further assume the experimental design of inputs is random: Input components $\mathbf{x}_i^\mu$ are drawn i.i.d. from a zero mean Gaussian with variance $1/P$, yielding inputs of expected norm 1. In many systems-identification applications, including, for example, in sensory

neuroscience, this random design would correspond to a white-noise stimulus. Now, given knowledge of the N input-output pairs $\{\mathbf{x}^\mu, y^\mu\}$, the noise density P_ϵ, and the prior information encoded in P_s, we would like to infer, in a computationally tractable manner, an estimate $\hat{\mathbf{s}}$ of the true response vector $\mathbf{s}^0$. A critical parameter governing inference performance is the ratio of the number of measurements N to the dimensionality P of the unknown model parameter $\mathbf{s}^0$, i.e., the measurement density $\alpha = (N/P)$.

The performance of any inference procedure can be characterized in several ways. Most simply, we would like to achieve a small, per-component mean-square error, $q_s = (1/P)\sum_{i=1}^{P}(\hat{\mathbf{s}}_i - \mathbf{s}_i^0)^2$, in inferring the true parameters, or signal $\mathbf{s}^0$. Alternatively, it is useful to note that any inference procedure yielding an estimate $\hat{\mathbf{s}}$ implicitly decomposes the measurement vector $\mathbf{y}$ into the sum of a signal component $\mathbf{X}\hat{\mathbf{s}}$ and a noise estimate $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \mathbf{X}\hat{\mathbf{s}}$. Here, $\mathbf{X}$ is an N-by-P matrix whose rows are the measurement vectors $\mathbf{x}^\mu$. Thus, an inference procedure corresponds to a particular separation of measurements into estimated signal and noise, $\mathbf{y} = \mathbf{X}\hat{\mathbf{s}} + \hat{\boldsymbol{\epsilon}}$, which will generically differ from the true decomposition, $\mathbf{y} = \mathbf{X}\mathbf{s}^0 + \boldsymbol{\epsilon}$. While q_s reflects the error in estimating signal, $q_\epsilon = \frac{1}{N}\sum_{\mu=1}^{N}(\hat{\boldsymbol{\epsilon}}_\mu - \boldsymbol{\epsilon}_\mu)^2$ reflects the error in estimating noise. Finally, one of the main performance measures of an inference procedure is its ability to generalize, or make predictions about, the measurement outcome y in response to a new randomly chosen input $\mathbf{x}$ not present in the training set $\{\mathbf{x}^\mu\}$. Given an estimate $\hat{\mathbf{s}}$, it can be used to make the prediction $\hat{y} = \mathbf{x}\cdot\hat{\mathbf{s}}$, and the average performance of this prediction is captured by the generalization error $\mathcal{E}^{\text{gen}} = \langle\langle (y - \hat{y})^2 \rangle\rangle$. Here, the double average $\langle\langle \cdot \rangle\rangle$ denotes an average over both the training data $\{\mathbf{x}^\mu, y^\mu\}$, which $\hat{\mathbf{s}}$ depends on, and the held-out testing data $\{\mathbf{x}, y\}$, which is necessarily independent of $\hat{\mathbf{s}}$. An alternate measure of performance is the average error in the ability of $\hat{\mathbf{s}}$ to simply predict the training data: $\mathcal{E}^{\text{train}} = (1/N)\sum_{\mu=1}^{N}(y^\mu - \mathbf{x}^\mu\cdot\hat{\mathbf{s}})^2 = (1/N)\sum_{\mu=1}^{N}\hat{\boldsymbol{\epsilon}}_\mu^2$. In general, $\mathcal{E}^{\text{train}} < \mathcal{E}^{\text{gen}}$, since through the process of inference, the learned parameters $\hat{\mathbf{s}}$ can acquire subtle correlations with the particular realization of training inputs $\{\mathbf{x}^\mu\}$ and noise $\{\epsilon^\mu\}$ so as to reduce $\mathcal{E}^{\text{train}}$. Situations where $\mathcal{E}^{\text{train}} \ll \mathcal{E}^{\text{gen}}$ correspond to inference procedures that overfit to the training data and do not exhibit predictive power by generalizing to new data.

Now, what inference procedures can achieve good performance in a computationally tractable manner? Regularized M-estimation (see Refs. [24,25] for reviews) yields a large family of computationally tractable estimation procedures in which $\hat{\mathbf{s}}$ is computed through the minimization

$$\hat{\mathbf{s}} = \underset{\mathbf{s}}{\text{argmin}}\left[\sum_{\mu=1}^{N}\rho(y^\mu - \mathbf{x}^\mu\cdot\mathbf{s}) + \sum_{i=1}^{P}\sigma(s_i)\right]. \quad (1)$$

Here, $\mathbf{s}$ is a candidate response vector, ρ is a loss function that penalizes deviations between actual measurements y^μ and expected measurements $\mathbf{x}^\mu\cdot\mathbf{s}$ under the candidate $\mathbf{s}$, and $\sigma(s)$ is a regularization function that exploits prior information about $\mathbf{s}^0$.

In the absence of such prior information, a widely used procedure is maximum likelihood (ML) inference,

$$\hat{\mathbf{s}}^{\text{ML}} = \underset{s}{\text{argmax}}\log P(\{y^\mu\}|\{\mathbf{x}^\mu\}, \mathbf{s}). \quad (2)$$

ML corresponds to noise energy minimization through the choice $\rho = E_\epsilon$ and $\sigma = 0$ in Eq. (1). Amongst all unbiased estimation procedures (in which $\langle\hat{\mathbf{s}}\rangle = \mathbf{s}^0$, where $\langle\cdot\rangle$ denotes an average over noise realizations), this energy minimization is optimal but only in the low-dimensional limit. Thus, amongst unbiased procedures, ML achieves the minimum mean-squared error (MMSE), when $\alpha \to \infty$, but not at finite α.

With prior knowledge, the Bayesian posterior mean achieves the MMSE estimate,

$$\hat{\mathbf{s}}^{\text{MMSE}} = \langle\mathbf{s}|\{y^\mu, \mathbf{x}^\mu\}\rangle = \int d\mathbf{s}\,\mathbf{s}P(\mathbf{s}|\{y^\mu, \mathbf{x}^\mu\}). \quad (3)$$

However, while no inference procedure can outperform high-dimensional Bayesian inference of the posterior mean, this procedure is not an M-estimator. It is also, in general, often computationally intractable because of the P-dimensional integral. However, as we discuss below in the related work section, it is thought that in the dense i.i.d. Gaussian measurement setting for $\mathbf{x}_i^\mu$ considered here, a good approximation to the integral can be obtained via efficient message-passing algorithms.

A widely used, generally more computationally tractable surrogate for the computation of the full posterior mean is maximum- *a posteriori* (MAP) inference,

$$\hat{\mathbf{s}}^{\text{MAP}} = \underset{s}{\text{argmax}}\log P(\mathbf{s}|\{y^\mu, \mathbf{x}^\mu\}), \quad (4)$$

which corresponds to noise and signal energy minimization through the choice $\rho = E_\epsilon$ and $\sigma = E_s$ in Eq. (1). MAP inference, by potentially introducing a nonzero bias (so that $\langle\hat{\mathbf{s}}\rangle \neq \mathbf{s}^0$), can outperform ML at finite α, but it is not, in general, optimal. However, the exploitation of prior information through a judicious, even if suboptimal, choice of σ can dramatically reduce estimation error. For example, the seminal advance of compressed sensing (CS) [26–28], as well as LASSO regression [29], uses $\rho = \frac{1}{2}\epsilon^2$ and $\sigma \propto |s|$. This choice can lead to accurate inference of sparse $\mathbf{s}^0$ even when $\alpha < 1$, where sparsity means that $P_s(s)$ assigns a small probability to nonzero values.

Despite the important and successful special cases of MAP inference, CS and LASSO, there is no general method to choose the best ρ and σ for inference. The

central questions we address in this work are as follows: (1) Given an estimation problem defined by the triplet of measurement density, noise, and prior (α, E_ϵ, E_s), and an estimation procedure defined by the loss and regularization pair (ρ, σ), what is the typical error q_s achieved for random inputs $\mathbf{x}^\mu$ and noise ϵ^μ? (2) What is the minimal achievable estimation error q^{opt} over all possible choices of convex procedures (ρ, σ)? (3) Which procedure (ρ^{opt}, σ^{opt}) achieves the minimal error q^{opt}, and under what conditions? (4) Are there simple universal relations between q_s and q_ϵ which measure the ability of an inference procedure to accurately separate signal and noise, $\mathcal{E}^{\text{train}}$ and $\mathcal{E}^{\text{gen}}$, which capture the predictive power of an inference procedure? (5) How does the performance q^{opt} of an optimal M-estimator compare to the best performance achievable by any algorithm, namely, that obtained by Bayesian MMSE inference? Our discussion section gives a summary of the answers we find to these questions.

B. Related work

For the special case of unregularized M-estimation ($\sigma = 0$), the error q_s and the form of the optimal loss function were characterized in a recent work [30], using mathematical arguments that are reminiscent of the cavity method in statistical physics. A closely related work [31] studied the same questions using a different technique known as approximate message passing (AMP), again assuming no regularization. By focusing on unregularized M-estimation, these works leave open the important question of how to exploit prior information about the signal distribution, which can often be essential for accurate inference in high dimensions. For example, the seminal advances of compressed sensing and LASSO reveal that simple choices of convex regularization can yield dramatic performance improvements in sparse signal recovery *even* at measurement densities less than 1. In contrast, the methods of Refs. [30,31] can be applied only in the case of measurement densities greater than 1 because of their focus on unregularized M-estimation. Here, motivated by the dramatic performance improvements enabled by even simple regularization choices, we focus on the fundamental question of how to optimally exploit prior information by choosing the best regularizer at any measurement density.

Also, in contrast to these works, we employ replica theory for our analysis. However, the techniques of AMP and replica theory are closely related. In particular, optimization problems of the form in Eq. (1) can be viewed as a graphical model [32] or a joint (zero-temperature) distribution over P variables with $N + P$ factors or constraints corresponding to each term in the sum. Belief propagation (BP) is a technique for finding the marginal distribution of a single variable in such a graphical model. BP is known to be exact on tree structured graphical models, and it often provides good approximate marginals on random sparse graphical models in which small numbers of variables interact with each other in each constraint [33,34]. In contrast, Eq. (1) corresponds to a dense graphical model in which all N variables interact in the measurement constraints due to the random Gaussian distribution of $\mathbf{x}_i^\mu$. AMP is an approximate version of BP designed to work well in such dense graphical models. It was proposed, for example, in Ref. [35] to study compressed sensing with Gaussian measurements. In such a dense Gaussian setting, the AMP algorithm was proven in Ref. [36] to yield the same answer as that obtained via a direct solution of the convex optimization problem. This result was extended in Ref. [37] from a Bayesian perspective.

A theoretical advantage of AMP is that its performance across iterations can be tracked using a set of state-evolution (SE) update equations. Remarkably, the fixed-point conditions of these SE equations often correspond to the self-consistency equations for the order parameters in replica theory (see, e.g., Refs. [19,34]), though there is no general theory that explains why this correspondence should always hold. However, it is fortunate that in our case, this correspondence does hold; in the very special case of zero regularization, our replica theory predictions for performance match those of Ref. [31], derived via state evolution, as well as those of Refs. [30,38], derived via cavitylike methods. For a general overview of replica theory, the cavity method, and message passing within the context of neural systems and high-dimensional data, see Ref. [10].

Interestingly, the Bayesian MMSE estimation algorithm (3) has also been studied from the perspective of both the replica method and AMP (see, e.g., Refs. [15,19,37]). Although it has not yet been rigorously proven, the AMP algorithms for Bayesian MMSE inference are conjectured to yield the same answer as direct integration in Eq. (3) in the high-dimensional data limit assuming Gaussian i.i.d. measurements x_i^μ (see Ref. [19] for a discussion). Such replica methods are widely accepted and have even been extended to analyze optimal matrix factorization [20]. Although Bayesian MMSE estimation is not the primary focus of this paper, we do compare the replica solution of Bayesian MMSE inference to the performance predicted by the optimal M-estimators we derive.

II. RESULTS

A. Review and formulation of classical scalar inference

Before considering the finite α regime, it is useful to review classical statistics in the $\alpha \to \infty$ limit, in the context of scalar estimation, where $P = 1$. In particular, we formulate these results in a suggestive manner that will aid in understanding the novel phenomena that emerge in modern, high-dimensional statistical inference, derived below. Here, for simplicity, we choose the scalar measurements $x^\mu = 1 \forall \mu$ in Eq. (1). Thus, we must estimate the scalar s^0 from $\alpha = N$ noisy measurements, $y^\mu = s^0 + \epsilon^\mu$.

With no regularization ($\sigma = 0$), for large N, $\hat{s}$ in Eq. (1) will be close to s^0, so Taylor expanding ρ about s^0 simply yields the asymptotic error (see Refs. [24,25], and Appendix A. 1 of Ref. [39])

$$q_s = \frac{1}{N}\frac{\langle\langle \rho'(\epsilon)^2 \rangle\rangle_\epsilon}{\langle\langle \rho''(\epsilon) \rangle\rangle_\epsilon^2}. \tag{5}$$

The Cramer-Rao (CR) bound is a fundamental information theoretic lower bound, at any N, on the error of *any* unbiased estimator $\hat{s}(\{y^\mu\})$ (obeying $\langle \hat{s} - s^0 \rangle_\epsilon = 0$):

$$q_s \geq \frac{1}{N}\frac{1}{J[\epsilon]}, \tag{6}$$

where $J[\epsilon]$ is the Fisher information from a single measurement y,

$$J[\epsilon] = \left\langle\left\langle \left(\frac{\partial}{\partial s^0}\log P(y|s^0)\right)^2 \right\rangle\right\rangle_y = \left\langle\left\langle \left(\frac{\partial}{\partial \epsilon}E_\epsilon\right)^2 \right\rangle\right\rangle_\epsilon. \tag{7}$$

The Fisher information measures the susceptibility of the output y to small changes in the parameter s^0. The higher this susceptibility, the lower the achievable error in Eq. (6). For finite N, it is not clear if there exists a loss function ρ whose performance saturates the CR bound. However, a central result in classical statistics states that as $N \to \infty$, the choice $\rho = E_\epsilon$ saturates Eq. (6), as can be seen by substituting $\rho = E_\epsilon$ in Eq. (5) (see Ref. [39], Appendix A. 2). Interestingly, at finite N the optimal equivariant estimator, in which a constant shift in the data results in the same shift in the estimator, is known. This estimator is an unbiased procedure known as Pitman estimation [40], which corresponds to $\hat{s}^{\mathrm{P}} = 1/[P(\{y^\mu\})] \int \mathrm{d}s s P(\{y^\mu\}|s)$. However, it is not an M-estimator, corresponding to any choice of ρ in Eq. (1).

It is also possible to perform more accurate inference with biased estimates by using knowledge of the true signal distribution $P(s^0)$. In particular, the posterior mean $\langle s|\{y^\mu\}\rangle = \int \mathrm{d}s s P(s|\{y^\mu\})$ achieves a minimal possible error q_s, amongst all inference procedures, biased or not, at any finite N. We compute this minimal q_s, in the limit of large N, via a saddle-point approximation to this Bayesian integral, yielding a mean-field theory (MFT) for low-dimensional Bayesian inference (see Ref. [39], Appendix A. 3), where the N measurements y^μ of s^0, corrupted by *non*-Gaussian noise ϵ^μ, can be replaced by a *single* measurement $y = s^0 + \sqrt{q_d}z$, corrupted by an effective Gaussian noise of variance

$$q_d = \frac{1}{NJ[\epsilon]}. \tag{8}$$

Here, z is a zero-mean unit-variance Gaussian variable. In our MFT, q_s is the MMSE error q_s^{MMSE} of this equivalent single-measurement, Gaussian noise inference problem:

$$q_s^{\mathrm{MMSE}}(q_d) = \langle\langle (s^0 - \langle s|y = s^0 + \sqrt{q_d}z\rangle)^2 \rangle\rangle_{s^0,z}. \tag{9}$$

We further prove a general lower bound on the asymptotic error,

$$q_s \geq \frac{1}{NJ[\epsilon] + J[s^0]}, \tag{10}$$

and demonstrate that this bound is tight when the signal and noise are Gaussian (see Ref. [39], Appendix A. 3). This bound is also known in the statistics literature as the Bayesian Cramer-Rao or Van-Trees inequality (see, e.g., Ref. [41]).

Thus, the classical theory of unbiased statistical inference as the measurement density $\alpha \to \infty$ reveals that ML achieves information theoretic limits on error (6). Moreover, an asymptotic analysis of Bayesian inference as $\alpha \to \infty$ [Eqs. (8)–(10)] reveals the extent to which biased procedures that optimally exploit prior information can circumvent such limits. Our work below constitutes a fundamental extension of these results to modern high-dimensional problems at finite measurement density.

B. Statistical mechanics framework

To understand the properties of the solution $\hat{\mathbf{s}}$ to Eq. (1), we define an energy function

$$E(\mathbf{s}) = \sum_{\mu=1}^{N} \rho(y^\mu - \mathbf{x}^\mu \cdot \mathbf{s}) + \sum_{i=1}^{P} \sigma(s_i), \tag{11}$$

yielding a Gibbs distribution $P_G(\mathbf{s}) = (1/Z)e^{-\beta E(\mathbf{s})}$ that freezes onto the solution of Eq. (1) in the zero-temperature $\beta \to \infty$ limit. In this statistical mechanics system, $\mathbf{x}^\mu$, ϵ^μ, and $\mathbf{s}^0$ play the role of quenched disorder, while the components of the candidate parameters $\mathbf{s}$ comprise thermal degrees of freedom. For large N and P, we expect self-averaging to occur: The properties of P_G for any typical realization of disorder coincide with the properties of P_G averaged over the disorder. Therefore, we compute the average free energy $-\beta\bar{F} \equiv \langle\langle \ln Z \rangle\rangle_{\mathbf{x}^\mu,\epsilon^\mu,\mathbf{s}^0}$ using the replica method [42]. We employ the replica symmetric (RS) approximation, which is effective for convex ρ and σ (see Ref. [39], Sec. II. 1 for details of our replica calculation). For a review of statistical mechanics methods applied to high-dimensional inference in diverse settings, see Ref. [10].

Central objects in optimization theory emerge naturally from our replica analysis, and the resulting MFT is most naturally described in terms of them. First is the proximal map $x \to \mathcal{P}_\lambda[f](x)$, where

$$\mathcal{P}_\lambda[f](x) = \underset{y}{\mathrm{argmin}}\left(\frac{(y-x)^2}{2\lambda} + f(y)\right). \tag{12}$$

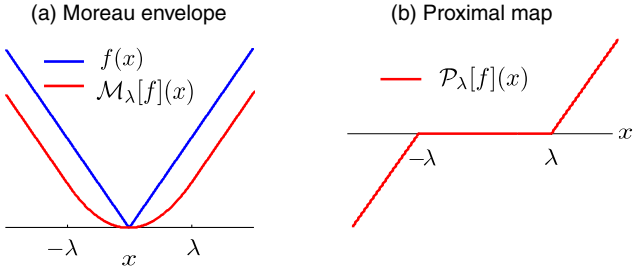

FIG. 2. (a) An example of a smooth, lower-bounding Moreau envelope $\mathcal{M}_\lambda[f](x)$ in Eq. (13) for $f(x) = |x|$. Explicitly, $\mathcal{M}_\lambda[f](x) = (x^2/2\lambda)$ for $|x| \le \lambda$, and $|x| - (\lambda/2)$ for $|x| \ge \lambda$. (b) The proximal map $\mathcal{P}_\lambda[f](x)$ in Eq. (12) for $f(x) = |x|$. Explicitly, $\mathcal{P}_\lambda[f](x) = 0$ for $|x| \le \lambda$, and $x - \mathrm{sign}(x)\lambda$ for $|x| \ge \lambda$. Thus, the proximal descent map $x \to \mathcal{P}_\lambda[f](x)$ moves x towards the minimum of $f(x)$.

This mapping is a proximal descent step that maps x to a new point that minimizes f while remaining proximal to x, as determined by a scale λ. The proximal map is closely related to the Moreau envelope of f, given by

$$\mathcal{M}_\lambda[f](x) = \min_y \left(\frac{(y-x)^2}{2\lambda} + f(y) \right). \tag{13}$$

$\mathcal{M}_\lambda[f]$ is a minimum convolution of $f(x)$ with a quadratic $x^2/2\lambda$, yielding a lower bound on f that is smoothed over a scale λ. See Figs. 2(a) and 2(b) for an example. The proximal map and Moreau envelope are related:

$$\mathcal{P}_\lambda[f](x) = x - \lambda \mathcal{M}'_\lambda[f](x), \tag{14}$$

where the prime denotes differentiation with respect to x. Thus, a proximal descent step on f can be viewed as a gradient descent step on $\mathcal{M}_\lambda[f]$ with step length λ. See Ref. [39], Appendix C. 1, and also Ref. [43] for a review of these topics.

Our replica analysis yields a pair of zero-temperature MFT distributions $P_{\mathrm{MF}}(s^0, \hat{s})$ and $P_{\mathrm{MF}}(\epsilon, \hat{\epsilon})$. The first describes the joint distribution of a single component $(s_i^0, \hat{s}_i)$ in Eq. (1), while the second describes the joint distribution of a noise component ϵ^μ and its estimate $\hat{\epsilon}^\mu \equiv y^\mu - \mathbf{x}^\mu \cdot \hat{\mathbf{s}}$. The MFT distributions can be described in terms of a pair of coupled scalar noise and signal estimation problems, depending on a set of RS order parameters $(q_s, q_d, \lambda_\rho, \lambda_\sigma)$. Here, q_s and q_d reflect the variance of additive Gaussian noise that corrupts the noise ϵ and signal s^0, respectively, yielding the measured variables

$$\epsilon_{q_s} = \epsilon + \sqrt{q_s} z_\epsilon, \qquad s^0_{q_d} = s^0 + \sqrt{q_d} z_s, \tag{15}$$

where z_ϵ and z_s are independent zero-mean unit-variance Gaussians. From these measurements, estimates $\hat{\epsilon}$ and $\hat{s}$ of the original noise ϵ and signal s^0 are obtained through proximal descent steps on the loss ρ and regularization σ:

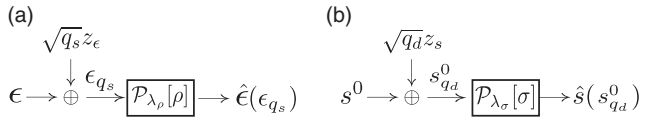

FIG. 3. A low-dimensional scalar MFT for high-dimensional inference. Diagrams (a) and (b) are schematic descriptions of Eqs. (15) and (16). They describe a pair of scalar statistical estimation problems, one for a noise variable ϵ, drawn from P_ϵ in (a), and the other for a signal variable s^0, drawn from P_s in (b). Each variable is corrupted by additive Gaussian noise, and from these noise-corrupted measurements, the original variables are estimated through proximal descent steps, yielding a noise estimate $\hat{\epsilon}$ in (a) and a signal estimate $\hat{s}$ in (b). The MFT distributions $P_{\mathrm{MF}}(\epsilon, \hat{\epsilon})$ and $P_{\mathrm{MF}}(s^0, \hat{s})$ are obtained by integrating out z_ϵ and z_s in (a) and (b), respectively. These joint MF distributions describe the joint distribution of pairs of single components $(\epsilon_\mu, \hat{\epsilon}_\mu)$ and $(s_i^0, \hat{s}_i)$ in Eq. (1), after integrating out all other elements of the quenched disorder in the training data and true signal.

$$\hat{\epsilon}(\epsilon_{q_s}) = \mathcal{P}_{\lambda_\rho}[\rho](\epsilon_{q_s}), \qquad \hat{s}(s^0_{q_d}) = \mathcal{P}_{\lambda_\sigma}[\sigma](s^0_{q_d}), \tag{16}$$

where λ_ρ and λ_σ reflect scale parameters. The joint MFT distributions are then obtained by integrating out z_ϵ and z_s. These MFT equations can be thought of as defining a pair of scalar estimation problems, one for the noise and one for the signal [see Figs. 3(a) and 3(b) for a schematic].

The order parameters obey self-consistency conditions that couple the performance of these scalar estimation problems:

$$q_d = \frac{\langle\langle \mathcal{M}'_{\lambda_\rho}[\rho](\epsilon_{q_s})^2 \rangle\rangle_{\epsilon_{q_s}}}{\alpha \langle\langle \mathcal{M}''_{\lambda_\rho}[\rho](\epsilon_{q_s}) \rangle\rangle^2_{\epsilon_{q_s}}}, \qquad q_s = \langle\langle (\hat{s} - s^0)^2 \rangle\rangle_{s^0_{q_d}}, \tag{17}$$

$$1 - \frac{1}{\alpha}\frac{\lambda_\rho}{\lambda_\sigma} = \langle\langle \hat{\epsilon}'(\epsilon_{q_s}) \rangle\rangle_{\epsilon_{q_s}}, \qquad \frac{\lambda_\rho}{\lambda_\sigma} = \langle\langle \hat{s}'(s^0_{q_d}) \rangle\rangle_{s^0_{q_d}}. \tag{18}$$

Here, $\langle\langle \cdot \rangle\rangle$ denotes averages over the quenched disorder in Eq. (15). The pair of MF distributions determine various measures of inference performance in Eq. (1). In particular, q_s predicts the typical per-component error of the learned model parameters, or signal $\hat{\mathbf{s}}$, while $q_\epsilon = \langle\langle (\hat{\epsilon} - \epsilon)^2 \rangle\rangle_{\epsilon_{q_s}}$ predicts the typical per-component error of the estimated noise. The model's prediction, or generalization error $\mathcal{E}^{\mathrm{gen}} = \langle\langle (y - \mathbf{x} \cdot \hat{\mathbf{s}})^2 \rangle\rangle$ on a new example $(\mathbf{x}, y)$ not present in the training set $\{\mathbf{x}^\mu, y^\mu\}$, can be obtained by substituting $y = \mathbf{x} \cdot \mathbf{s}^0 + \epsilon$ into $\mathcal{E}^{\mathrm{gen}}$. This yields the MFT prediction for the generalization error, $\mathcal{E}^{\mathrm{gen}} = \langle\langle (\epsilon_{q_s})^2 \rangle\rangle = q_s + \langle \epsilon^2 \rangle$. In contrast, the MFT prediction for the training error is simply $\mathcal{E}^{\mathrm{train}} = \langle\langle \hat{\epsilon}(\epsilon_{q_s})^2 \rangle\rangle$.

Because the proximal map is contractive, with Jacobian less than 1 [43], the MFT predicts, as expected, that $\mathcal{E}^{\mathrm{train}} < \mathcal{E}^{\mathrm{gen}}$. The reduced $\mathcal{E}^{\mathrm{train}}$ is due to the subtle

correlations that the learned parameters $\hat{\mathbf{s}}$ can acquire with the particular realization of training inputs $\{\mathbf{x}^\mu\}$ and noise $\{\epsilon^\mu\}$, through the optimization in Eq. (1). Remarkably, these subtle correlations are captured in the MFT simply through a proximal descent step in Eq. (16) on the cost ρ. This step contracts the variable ϵ_{q_s} controlling $\mathcal{E}^{\text{gen}}$ towards the minimum of ρ at the origin, leading to smaller $\mathcal{E}^{\text{train}}$. We explore many more consequences of this MFT below.

C. Inference without prior information

If we cannot exploit prior information, we simply choose $\sigma = 0$, which yields $\hat{s} = s^0_{q_d}$ in Eq. (16), so that the rhs of Eqs. (17) and (18) reduce to $q_s = q_d$ and $\lambda_\rho = \lambda_\sigma$. Then, replacing q_d with q_s on the lhs of Eq. (17), and comparing to Eq. (5), we see that the high-dimensional inference error is analogous to the low-dimensional one, with the number of measurements N replaced by the measurement density α, the cost $\rho(\cdot)$ replaced by its Moreau envelope $\mathcal{M}_{\lambda_\rho}[\rho](\cdot)$, and the noise ϵ further corrupted by additive Gaussian noise of variance q_s, with q_s and λ_ρ determined self-consistently through Eqs. (17) and (18).

As a simple example, consider the ubiquitous case of quadratic cost: $\rho(x) = \frac{1}{2}x^2$. Then the proximal map (16) is simply linear shrinkage to the origin, $\hat{\epsilon}(\epsilon_{q_s}) = [1/(1+\lambda_\rho)]\epsilon_{q_s}$, and Eqs. (17) and (18) are readily solved: $q_s = [1/(\alpha-1)]\langle\epsilon^2\rangle$, $\lambda_\rho = [1/(\alpha-1)]$, yielding $\mathcal{E}^{\text{gen}} = [\alpha/(\alpha-1)]\langle\epsilon^2\rangle$ and $\mathcal{E}^{\text{train}} = [(\alpha-1)/\alpha]\langle\epsilon^2\rangle$. Thus, as the measurement density approaches 1 from above, the errors in inferred parameters $\hat{\mathbf{s}}$ and $\mathcal{E}^{\text{gen}}$ diverge, while $\mathcal{E}^{\text{train}}$ vanishes, indicating severe overfitting.

Now, in the space of all convex costs ρ, for a given density α and noise energy E_ϵ, what is the minimum possible estimation error q^{opt}? By performing a functional minimization of q_s over ρ subject to the constraints (17) and (18) (see Ref. [39], Secs. 4.1 and 5.1 for details), we find that q^{opt} is the minimal solution to

$$q^{\text{opt}} = \frac{1}{\alpha}\frac{1}{J[\epsilon_{q^{\text{opt}}}]} \geq \frac{1}{(\alpha-1)J[\epsilon]}, \tag{19}$$

where the second inequality follows from the convolutional Fisher inequality (Ref. [39], Appendix B. 2). This result is the high-dimensional analog of the Cramer-Rao bound in Eq. (6). By the data-processing inequality for Fisher information, $J[\epsilon_{q^{\text{opt}}}] < J[\epsilon]$, indicating higher error in the high-dimensional setting [Eq. (19)] than the low-dimensional setting [Eq. (6)]. Thus, the price paid for even optimal high-dimensional inference at finite measurement density, relative to ML inference at infinite density, is increased error due to the presence of additional Gaussian noise with dimensionality-dependent variance q_s.

Now can this minimal error q^{opt} be achieved, and if so, which cost function ρ^{opt} achieves it? Constrained functional optimization over ρ yields the functional equation $\mathcal{M}_{q^{\text{opt}}}[\rho](x) = E_{\epsilon_{q^{\text{opt}}}}$ (see Ref. [39], Sec. V. 1 for details), which can be inverted (see Ref. [39], Appendix B. 2) to find

$$\rho^{\text{opt}}(x) = -\mathcal{M}_{q^{\text{opt}}}[-E_{\epsilon_{q^{\text{opt}}}}](x). \tag{20}$$

The validity of this equation under the RS assumption requires that ρ^{opt} be convex. Convexity of the noise energy

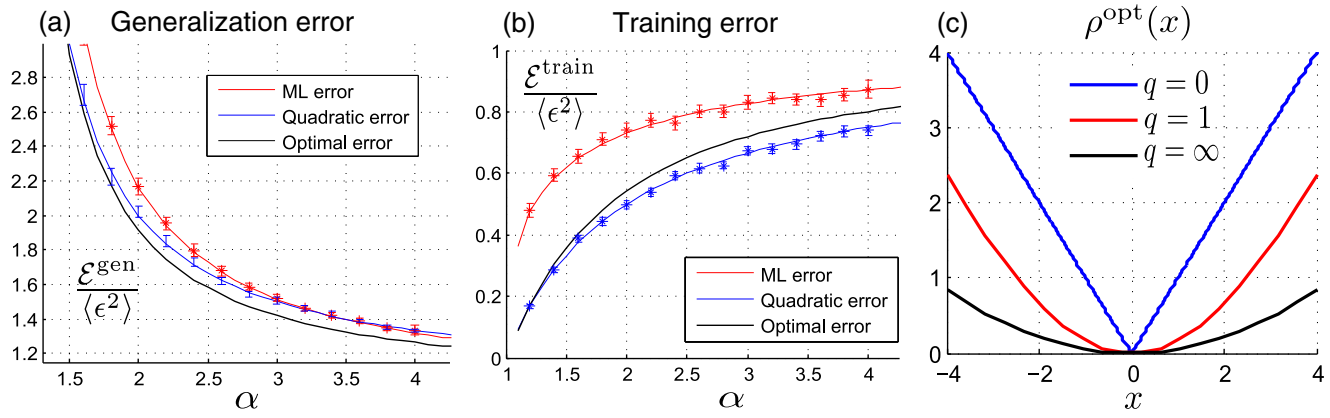

FIG. 4. Unregularized inference for Laplacian noise $E_\epsilon = |\epsilon|$. A comparison of the generalization error (a) and training error (b) of the optimal unregularized M-estimator (20) (black lines) with ML (red lines) and quadratic (blue lines) loss functions. Solid curves reflect theoretically derived predictions of performance. Error bars reflect performance obtained through numerical optimization of Eq. (1) using standard convex optimization solvers for finite-size problems (N and P vary, with $N = \alpha P$ and $\sqrt{NP} = 250$). The width of the error bars reflects standard deviation of performance across 100 different realizations of the quenched disorder. (c) The shape of the optimal loss function in Eq. (20) for high-dimensional inference as a function of the error or smoothing parameter q. As α varies from high to low measurement densities, q varies from low to high values, and the optimal loss function varies from the ML loss to quadratic. Intermediate versions of the optimal loss behave like a smoothed version of the ML loss, with increased smoothing as measurement density decreases (or dimensionality increases).

E_ϵ is sufficient to guarantee the convexity of $\rho^{\rm opt}$ (see Ref. [39], Appendix C. 3 for details), and so for this class of noise, Eq. (20) yields the optimal inference procedure.

In the classical $\alpha \to \infty$ limit, we expect $q^{\rm opt}$ to be small; indeed, to leading order in $1/\alpha$, Eq. (19) has the solution $q^{\rm opt} = [1/\alpha]\{1/J[\epsilon]\}$, while Eq. (20) reduces to $\rho^{\rm opt} = E_\epsilon$, recovering the optimality of ML and its performance [Eq. (6)] at infinite measurement density. In the high-dimensional $\alpha \to 1$ limit, $q^{\rm opt}$ diverges, so $\epsilon_{q^{\rm opt}}$ approaches a Gaussian with variance $\langle \epsilon^2 \rangle + q^{\rm opt}$, yielding $\rho^{\rm opt}(x) = (x^2/2)$ in Eq. (20). Thus, remarkably, at low measurement density, simple quadratic minimization, independent of the noise distribution, becomes an optimal inference procedure. As the measurement density decreases, $\rho^{\rm opt}$ interpolates between E_ϵ and a quadratic; in essence, $\rho^{\rm opt}$ at finite density α is a smoothed version of the ML choice $\rho = E_\epsilon$ where the amount of smoothing increases as the density decreases (or dimensionality increases). See Fig. 4 for an example of a family of optimal inference procedures, and their performance advantage relative to ML, for Laplacian noise ($E_\epsilon = |\epsilon|$).

These results are consistent with and provide a new statistical-mechanics-based derivation of results in Refs. [30,31,38], and they illustrate the severity of overfitting in the face of limited data.

D. Inference with prior information

We next explore how we can combat overfitting by optimally exploiting prior information about the distribution of the model parameters or signal s^0.

1. Optimal quadratic inference: A high SNR phase transition

To understand the MFT for regularized inference, it is useful to start with the oft-used quadratic loss and regularization: $\rho(x) = \frac{1}{2}x^2$ and $\sigma(x) = \frac{1}{2}\gamma x^2$. In this case, the proximal maps in Eq. (16) become linear and the RS equations (17) and (18) are readily solved (Ref. [39], Sec. III. 1). It is useful to express the results in terms of the fraction of unexplained variance $\bar{q}_s = [q_s/\langle s^2 \rangle]$ and the SNR $= \langle s^2 \rangle / \langle \epsilon^2 \rangle$. For quadratic inference, $\bar{q}_s$ depends on the signal and noise distributions only through the SNR. We find that in the strong regularization limit, $\gamma \to \infty$, $\bar{q}_s \to 1$, as the regularization pins the estimate $\hat{\mathbf{s}}$ to the origin, while in the weak regularization limit $\gamma \to 0$, $\bar{q}_s \to \{1/[{\rm SNR}(\alpha - 1)]\}$, recovering the unregularized case. There is an optimal intermediate value of the regularization weight, $\gamma = (1/{\rm SNR})$, leading to the highest fraction of variance explained. Thus, optimal quadratic inference obeys the principle that high-quality data, as measured by high SNR, requires weaker regularization. For this optimal γ, $\bar{q}_s$ arises as the solution to the set of simultaneous equations

$$q_d = \frac{\langle \epsilon^2 \rangle + q_s}{\alpha}, \qquad \frac{q_s}{\langle s^2 \rangle} = \frac{1}{1 + \frac{\langle s^2 \rangle}{q_d}}. \tag{21}$$

We denote the solution to these equations by $\bar{q}_s = \bar{q}_s^{\rm Quad}(\alpha, {\rm SNR})$. This function is simply the fraction of unexplained variance of optimal quadratic inference at a given measurement density and SNR, and an explicit expression is given by

$$\bar{q}_s^{\rm Quad} = \frac{1 - \alpha - \phi + \sqrt{(\phi + \alpha - 1)^2 + 4\phi}}{2}, \tag{22}$$

where $\phi = (1/{\rm SNR})$ (see Ref. [39], Sec. III. 2 for details).

This expression simplifies in several limits. At high SNR $\gg 1$,

$$\bar{q}_s^{\rm Quad} = \begin{cases} 1 - \alpha & \alpha < 1 \\ \frac{1}{\sqrt{\rm SNR}} & \alpha = 1 \\ \frac{1}{{\rm SNR}(\alpha - 1)} & \alpha > 1. \end{cases} \tag{23}$$

Thus, as a function of measurement density, the high SNR behavior of quadratic inference exhibits a phase transition at the critical density $\alpha_c = 1$. Below this density, in the undersampled regime, performance asymptotes to a finite error, independent of SNR. Above this density, in the oversampled regime, inference error decays with SNR as ${\rm SNR}^{-1}$. Surprisingly, at the critical density, the decay with SNR is slower, and it exhibits a universal decay exponent of $-\frac{1}{2}$, independent of the signal and noise distributions. This exponent, and its universality, is verified numerically in Fig. 5(a). Moreover, as $\alpha \to 1$, $\bar{q}_s^{\rm Quad}$ remains $O(1)$ at any finite SNR, unlike the unregularized case. Indeed, for $\alpha \ll 1$, $\bar{q}_s^{\rm Quad} = 1 - \alpha[{\rm SNR}/({\rm SNR} + 1)]$. Thus, quadratic regularization can tame the divergence of unregularized inference at low measurement density.

The phase transition behavior of optimal quadratic inference can be understood from the perspective of random matrix theory (RMT). In the special case of Eq. (1) when $\rho(x) = \frac{1}{2}x^2$ and $\sigma(x) = \frac{1}{2}(1/{\rm SNR})x^2$, the optimal estimate $\hat{\mathbf{s}}$ has the analytic solution

$$\hat{\mathbf{s}} = \left(\mathbf{X}^T\mathbf{X} + \frac{1}{\rm SNR}\mathbf{I} \right)^{-1} \mathbf{X}^T\mathbf{y}, \tag{24}$$

where $\mathbf{X}$ is an N-by-P measurement matrix whose N rows are the N measurement vectors $\mathbf{x}^\mu$ (see Ref. [39], Sec. III. 5, for more details). This analytic solution for $\hat{\mathbf{s}}$ enables a direct average over the noise $\boldsymbol{\epsilon}$ and true signal $\mathbf{s}^0$ in $\mathbf{y}$ to yield

$$\bar{q}_s^{\rm Quad} = \frac{1}{P}{\rm Tr}[\mathbf{I} + {\rm SNR}\,\mathbf{X}^T\mathbf{X}]^{-1}. \tag{25}$$

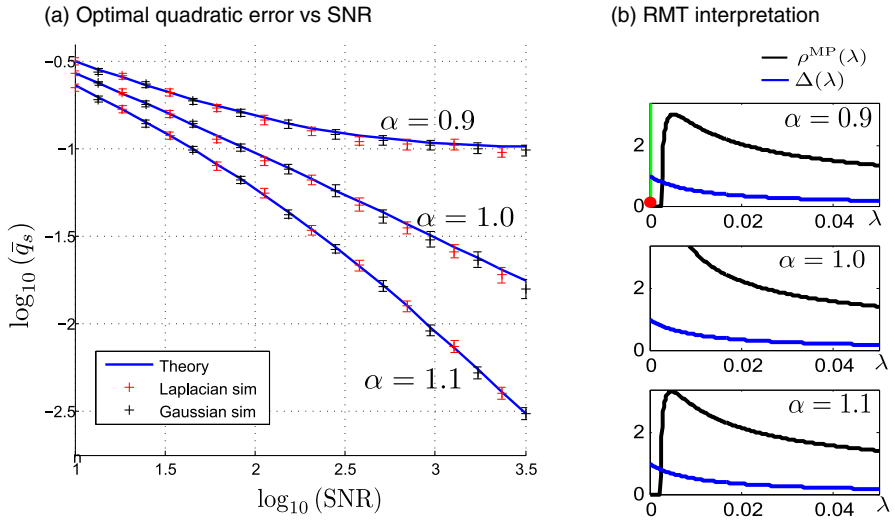

FIG. 5. A high SNR phase transition in optimal quadratic inference. (a) At large SNR, the MSE of optimal quadratic inference exhibits three distinct scaling regimes for $\alpha < 1$, $\alpha = 1$, and $\alpha > 1$ [see Eq. (23)], independent of the signal and noise distributions. For example, when $\alpha = 0.9 < 1$, $\bar{q}_s^{\mathrm{Quad}}$ approaches a constant, whereas when $\alpha = 1$ or $\alpha = 1.1 > 1$, $\bar{q}_s^{\mathrm{Quad}}$ approaches 0 as $\mathrm{SNR}^{-1/2}$ or SNR^{-1}, respectively. The theoretical curves (blue) match numerical experiments (error bars) for a finite-sized problems (N and P vary with $N = \alpha P$ and $\sqrt{NP} = 300$), where the error bars reflect the standard deviation across 80 trials using both signal and noise either Gaussian (black) or Laplacian (red) distributed. (b) The behavior of the MP density (black) in Eq. (26). For $\alpha \neq 1$, the nonzero continuous part of the density exhibits a gap at the origin, whereas for $\alpha = 1$, the gap vanishes and the distribution diverges at the origin. For $\alpha < 1$, there is an additional δ function at the origin (green bar) with weight $1 - \alpha$ (red dot). The blue curve shows the function $\Delta(\lambda) = (1 + \lambda \cdot \mathrm{SNR})^{-1}$ appearing in the integral for $\bar{q}_s^{\mathrm{Quad}}$ in Eq. (27), for the value $\mathrm{SNR} = 100$.

This expression can be reduced to an average over the eigenvalue distribution of the random measurement correlation matrix $\mathbf{X}^T\mathbf{X}$, which has the well-known Marcenko-Pasteur (MP) form [44]

$$\rho^{\mathrm{MP}}(\lambda) = \frac{1}{2\pi}\frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda} + \mathbf{1}_{\alpha<1}(1-\alpha)\delta(\lambda), \quad (26)$$

where the nonzero support of the density is restricted to the range $\lambda \in [\lambda_-, \lambda_+]$, with $\lambda_\pm = (\sqrt{\alpha} \pm 1)^2$. Also, $\mathbf{1}_{\alpha<1}$ is 1 when $\alpha < 1$ and 0 otherwise. Thus, at measurement densities $\alpha < 1$, the MP distribution has an additional delta function at the origin with weight $1 - \alpha$, reflecting the fact that the $P \times P$ measurement correlation matrix $\mathbf{X}^T\mathbf{X}$ is not full rank when $N < P$. In terms of $\rho^{\mathrm{MP}}(\lambda)$, Eq. (25) reduces to

$$\bar{q}_s^{\mathrm{Quad}} = \int \Delta(\lambda)\rho^{\mathrm{MP}}(\lambda)d\lambda, \quad (27)$$

where $\Delta(\lambda) = (1 + \lambda \cdot \mathrm{SNR})^{-1}$. Direct calculation reveals that expression (27) for $\bar{q}_s^{\mathrm{Quad}}(\alpha, \mathrm{SNR})$, derived via random matrix theory, is consistent with the expression (22), derived via our theory of high-dimensional statistical inference.

The expression for $\bar{q}_s^{\mathrm{Quad}}$ in Eq. (27) can now be used to elucidate the nature of the phase transition in Fig. 5(a). At high SNR, the function $\Delta(\lambda)$ remains $O(1)$ in a narrow regime of width $O(1/\mathrm{SNR})$ near the origin. However, when $\alpha < 1$, the left edge λ_- of the nonzero part of the MP density remains separated from the origin. Because of this eigenvalue density gap, the dominant contribution to the integral in Eq. (27) arises from the δ function at the origin, yielding $\bar{q}_s^{\mathrm{Quad}} \approx 1 - \alpha$ when $\alpha < 1$ [see Fig. 5(b), top]. When $\alpha > 1$, the δ function is absent, and the dominant contribution arises from the nonzero part of the MP density. This density has support over a range that is $O(\alpha)$ yielding $\bar{q}_s^{\mathrm{Quad}} = O(1/\mathrm{SNR}\alpha)$ [see Fig. 5(b), bottom]. Only when $\alpha = 1$ does the gap in the MP density vanish. In this case, near the origin, the density diverges as $\lambda^{-1/2}$ [see Fig. 5(b), middle]. At high SNR, because $\Delta(\lambda)$ induces an effective cutoff at $1/\mathrm{SNR}$, the integral in Eq. (27) can be approximated as $\int_0^{\mathrm{SNR}^{-1}} \lambda^{-1/2} d\lambda = O(\mathrm{SNR}^{-1/2})$.

Thus, the origin of the phase transition in Eq. (23) at the critical value $\alpha = 1$ arises from the vanishing of a gap in the MP distribution. Moreover, the universal decay exponent at the critical value of $\alpha = 1$ is related to the power-law behavior of the MP density near the origin at $\alpha = 1$. Remarkably, this highly nontrivial behavior is captured simply through the outcome of our replica analysis for optimal quadratic inference, encapsulated in the pair of equations in Eq. (21).

2. *The worst signal and noise distributions are Gaussian*

We note that this optimal quadratic inference procedure is optimal amongst all possible inference procedures, if and only if the signal and noise are Gaussian since, in that case,

it is equivalent to the Bayesian MMSE inference procedure. Moreover, we note that Gaussian signal and noise are, in some sense, the *worst* type of signal and noise distributions, in the space of all inference problems with a given SNR. To see this, consider a non-Gaussian signal and noise with a given SNR. The performance of optimal quadratic inference for this non-Gaussian signal and noise only depends on the pair of distributions through their SNR, and it is equivalent to the performance of optimal quadratic inference for Gaussian signal and noise at the same SNR. However, in the non-Gaussian case, a nonquadratic inference algorithm could potentially outperform the quadratic one but not in the Gaussian case since quadratic inference is already optimal in that case. Thus, in the space of inference problems of a given SNR, the worst-case performance of optimal inference occurs when both the signal and noise are Gaussian.

3. Optimal inference with non-Gaussian signal and noise

What is the optimal (nonquadratic) inference procedure in the face of non-Gaussian signal and noise? We address this by performing a functional minimization of q_s over *both* ρ and σ, subject to constraints (17) and (18), which yields (Ref. [39], Sec. V. 2),

$$\rho^{\rm opt}(x) = -\mathcal{M}_{q_s^{\rm opt}}[-E_{\epsilon_{q_s^{\rm opt}}}](x), \tag{28}$$

$$\sigma^{\rm opt}(x) = -\mathcal{M}_{q_d^{\rm opt}}[-E_{s_{q_d^{\rm opt}}}](x), \tag{29}$$

where $q_s^{\rm opt}$ and $q_d^{\rm opt}$ satisfy

$$q_d^{\rm opt} = \frac{1}{\alpha J[\epsilon_{q_s^{\rm opt}}]}, \qquad q_s^{\rm opt} = q_s^{\rm MMSE}(q_d^{\rm opt}), \tag{30}$$

and the function $q_s^{\rm MMSE}$ is defined in Eq. (9). Again, the validity of Eqs. (28) and (29) under the RS assumption requires convexity of $\rho^{\rm opt}$ and $\sigma^{\rm opt}$. Convexity of the signal and noise energies, E_s and E_ϵ, is sufficient to guarantee convexity of $\rho^{\rm opt}$ and $\sigma^{\rm opt}$ (see Ref. [39], Appendix C. 3, for details), and so for this class of signal and noise, with log concave distributions, Eqs. (28) and (29) yield an optimal inference procedure. However, by judicious applications of the Cauchy-Schwarz inequality, we prove (Ref. [39], Sec. IV. 2) that even for nonconvex E_s and E_ϵ, the inference error q_s for any convex procedure (ρ, σ) must exceed $q_s^{\rm opt}$ in Eq. (30). This result yields a fundamental limit on the performance of any convex inference procedure of the form (1) in high dimensions.

Intriguingly, by comparing the optimal achievable high-dimensional M-estimation performance $q_s^{\rm opt}$ in Eq. (30) to the asymptotic performance of low-dimensional scalar Bayesian inference in Eqs. (8) and (9), we find a striking parallel. In particular, $q_s^{\rm opt}$ corresponds to the low-dimensional asymptotic MMSE in a scalar estimation problem where the effective number of measurements $N = \alpha$ and the noise ϵ is further corrupted by additional Gaussian noise of variance $q_s^{\rm opt}$ ($\epsilon \to \epsilon + \sqrt{q_s^{\rm opt}} z$). The correction to the low-dimensional scalar asymptotics [Eq. (9)], valid only at large N, in the high-dimensional regime at finite measurement density α, is obtained by self-consistently solving for $q_s^{\rm opt}$ in Eq. (30). In essence, at finite measurement density, there is irreducible error in estimating the signal, $q_s^{\rm opt}$. This error contributes to the effective Gaussian noise $q_d^{\rm opt}$ in the scalar MFT estimation problem for the signal, shown in Fig. 3(b), where the proximal map becomes the Bayesian posterior mean map in the optimal case. On the other hand, this irreducible, extra Gaussian noise is absent in low dimensions [compare lhs of Eq. (30) to Eq. (8)]. This irreducible error $q_s^{\rm opt}$ can be found by self-consistently solving for it in the rhs of Eq. (30). Finally, as a simple point, we note that direct calculation reveals that Eq. (30) reduces to Eq. (21) when the signal and noise are both Gaussian distributed, as expected, since optimal quadratic inference is the best procedure for Gaussian signal and noise.

Furthermore, using the fact that the equalities in Eq. (30) become inequalities for nonoptimal procedures (see Ref. [39], Sec. IV.2), we can derive a high-dimensional analogue of Eq. (10) and prove a lower bound on the inference error q_s for any convex (ρ, σ):

$$q_s \geq \frac{1}{\alpha J[\epsilon_{q_s}] + J[s^0]}. \tag{31}$$

This result reflects a fundamental generalization of the high-dimensional CR bound (19), which includes information about the signal distribution P_s that can be optimally exploited by a regularizer σ. Since $J[\epsilon_{q_s}] < J[\epsilon]$, by the data-processing inequality for Fisher information, this high-dimensional lower bound is larger than the low-dimensional one [Eq. (10)] under the replacement $\alpha \to N$. Thus, as in the unregularized case [Eq. (19)], the price paid for even optimal high-dimensional regularized inference at finite measurement density, relative to scalar Bayesian inference at asymptotically infinite density, is increased error due to the presence of additional Gaussian noise with dimensionality-dependent variance $q_s^{\rm opt}$.

4. Optimal high-dimensional inference smoothly interpolates between MAP and quadratic inference

The optimal inference procedure in Eqs. (28) and (29) is a smoothed version of MAP inference [see Fig. 4(c) for an example of smoothing], where the MAP choices $\rho = E_\epsilon$ and $\sigma = E_s$ are smoothed over scales $q_s^{\rm opt}$ and $q_d^{\rm opt}$, respectively, to obtain $\rho^{\rm opt}$ and $\sigma^{\rm opt}$. As $\alpha \to \infty$, both $q_s^{\rm opt}$ and $q_d^{\rm opt}$ approach 0 at the same rate, implying $\rho^{\rm opt} \to E_\epsilon$ and $\sigma^{\rm opt} \to E_s$. Thus, at high measurement density,

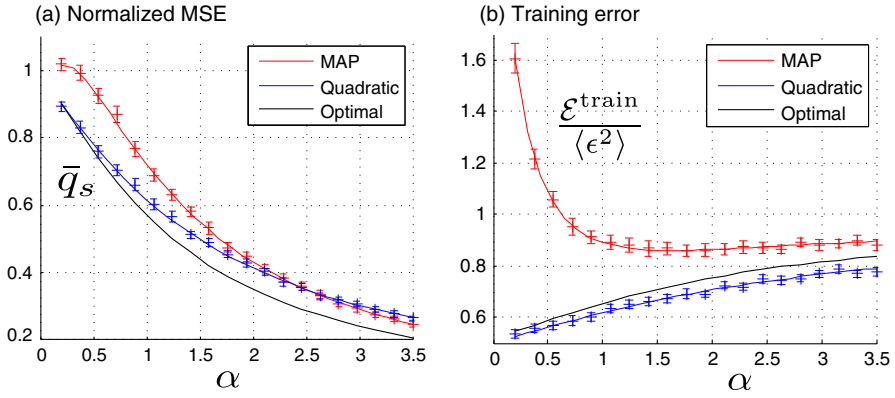

FIG. 6. Regularized inference for Laplacian noise and signal $E_\epsilon = |\epsilon|$, $E_s = |s^0|$. (a) The normalized MSE, or fraction of unexplained variance $\bar{q}_s$. (b) The training error. Each plot shows the respective performance of three different inference procedures: our optimal inference (28), (29) (black), MAP inference (red), and optimal quadratic inference (blue). The theoretical predictions (solid curves) match numerical simulations (error bars), which reflects the standard deviation calculated over 20 trials using a convex optimization solver for randomly generated, finite-sized data (with N and P varying while $N = \alpha P$ and $\sqrt{NP} = 250$). Note that optimal inference can significantly outperform common but suboptimal methods. For example, to achieve a fraction of unexplained variance of 0.4, optimal inference requires a measurement density of $\alpha \approx 1.7$, while quadratic and MAP inference require $\alpha \approx 2.1$ and $\alpha \approx 2.2$, respectively. This reflects a reduction of approximately 20% in the amount of required data.

MAP inference is the optimal M-estimator. This conclusion is intuitively reasonable because, at high measurement densities, the mode of the posterior distribution over the signal, returned by the MAP estimate, is typically close to the mean of the posterior distribution, which is the optimal MMSE estimate amongst all inference procedures.

Alternatively, as $\alpha \to 0$, $q_s^{\text{opt}} \to \langle s^2 \rangle$ from below, while q_d^{opt} diverges as $1/\alpha$. The divergence of q_d^{opt} implies that σ_{opt} in Eq. (29) approaches a quadratic. Thus, remarkably, at low measurement density, simple quadratic regularization, independent of the signal distribution, becomes an optimal inference procedure. Furthermore, in the low-density-plus-high-SNR limit, where $\langle \epsilon^2 \rangle \ll \langle s^2 \rangle$, ρ^{opt} also approaches a quadratic. Thus, overall, optimal high-dimensional inference at high SNR interpolates between MAP and quadratic inference as the measurement density decreases. In Fig. 6, we demonstrate, for Laplacian signal and noise, that optimal inference outperforms both MAP and quadratic inference at all α, approaching the former at large α and the latter at small α.

5. A relation between optimal high-dimensional inference and low-dimensional Bayesian inference

There is an interesting connection between optimal high-dimensional inference and low-dimensional scalar Bayesian inference. Indeed, when ρ and σ take their optimal forms in Eqs. (28) and (29), then the proximal descent steps in Eq. (16), which are used to estimate noise and signal in the pair of coupled estimation problems comprising the MFT [shown schematically in Figs. 3(a) and 3(b)] become optimal Bayesian estimators. In particular, for optimal ρ and σ, Eq. (16) becomes (see Ref. [39], Sec. V. 2)

$$\hat{\epsilon}(\epsilon_{q_s}) = \langle \epsilon | \epsilon_{q_s} \rangle, \qquad \hat{s}(s^0_{q_d}) = \langle s | s^0_{q_d} \rangle. \tag{32}$$

In essence, computation of the proximal map becomes computation of the posterior mean, which is the optimal, MMSE method for estimating signal and noise in the MFT scalar estimation problems. This gives an intuitive explanation for the form of ρ^{opt} and σ^{opt} in Eqs. (28) and (29): These are exactly the forms of loss and regularization required for the proximal descent estimates in Eq. (16) to become optimal posterior mean estimates in Eq. (32).

6. A relation between signal-noise separation and predictive power

Furthermore, there is an interesting connection between our ability to optimally estimate noise and signal, and the training and test error. In particular, just as our error q_s^{opt} in estimating the signal is given by Eqs. (30) and (9), our error in estimating the noise is given by $q_\epsilon^{\text{opt}} = \langle\!\langle (\hat{\epsilon} - \epsilon)^2 \rangle\!\rangle$, with $\hat{\epsilon}$ given in Eq. (32), yielding

$$q_\epsilon^{\text{opt}} = q_\epsilon^{\text{MMSE}}(q_s^{\text{opt}}) = \langle\!\langle (\epsilon - \langle \epsilon | \epsilon_{q_s^{\text{opt}}} \rangle)^2 \rangle\!\rangle. \tag{33}$$

In terms of these quantities, the generalization and training errors of the optimal M-estimator have very simple forms (see Ref. [39], Sec. V. 2):

$$\mathcal{E}^{\text{train}} = \langle \epsilon^2 \rangle - q_\epsilon^{\text{opt}}, \qquad \mathcal{E}^{\text{gen}} = \langle \epsilon^2 \rangle + q_s^{\text{opt}}. \tag{34}$$

This result leads to an intuitively appealing result: Inability to estimate the signal leads directly to increased generalization error, while inability to estimate the noise leads to *decreased* training error.

The reason for this latter effect is that if the optimal inference procedure cannot accurately separate signal from noise to correctly estimate the noise, then it mistakenly identifies noise in the training data as signal, and this noise is incorporated into the parameter estimate $\hat{\mathbf{s}}$. Thus, $\hat{\mathbf{s}}$ acquires correlations with the particular realization of noise in the training set so as to reduce training error. However, this reduced training error comes at the expense of increased generalization error, again due to mistaking noise for signal. The predicted decrease of training error and increase of generalization error for the optimal inference procedure as measurement density decreases is demonstrated in Fig. 6. Interestingly, this figure also demonstrates that training error need not decrease at low measurement density for suboptimal algorithms, like MAP.

Thus, in summary, the ability to correctly separate signal from noise to extract a model of the measurements $\mathbf{y}$ in Eq. (1) is intimately related to the predictive power of the extracted model $\hat{\mathbf{s}}$ in Eq. (1). Inability to estimate noise reduces training error, while inability to estimate signal increases generalization error. The combination is a hallmark of overfitting the learned model parameters to the training data, thereby leading to a loss of predictive power on new, held-out data.

E. No performance gap between optimal M-estimation and Bayesian MMSE inference

The improved performance of optimal inference via M-estimation, compared to either MAP or quadratic inference, demonstrated in Fig. 6(a) raises an important question: How does the performance of optimal M-estimation compare to the best performance achievable by any algorithm, namely, that obtained by Bayesian MMSE inference, described in Eq. (3)? To answer this question, we study the statistical mechanics of the energy function (11) at a finite, unit temperature $\beta = 1$, in contrast to the zero-temperature $\beta \to \infty$ limit that governs the performance of M-estimation. With $\beta = 1$, we further choose $\rho = -\log P_\epsilon$ and $\sigma = -\log P_s$ in Eq. (11) so that the corresponding Gibbs distribution is simply the posterior distribution over the signal:

$$P_G(\mathbf{s}) = \frac{1}{Z} e^{-\beta E(\mathbf{s})} = P(\mathbf{s}|\{y^\mu, \mathbf{x}^\mu\}). \tag{35}$$

Previous works have employed this statistical-mechanics-based method for studying Bayes optimal inference in the settings of compressed sensing [16,19] and matrix factorization [20].

We work out the replica theory for this finite-temperature statistical-mechanics problem in Ref. [39], Sec. V. 7. We work in the replica symmetric approximation at unit temperature. A sufficient, though not necessary, assumption guaranteeing the validity of the RS approximation is that ρ and σ are convex, or equivalently, the signal and noise distributions are log-concave. Indeed, as discussed above, this condition on signal and noise is sufficient to guarantee the validity of our optimal M-estimators. See, however, Refs. [19,20] for more general settings in which the RS assumption is valid for MMSE inference. In the setting of log-concave signal and noise, we discover an equivalence between MMSE inference and optimal M-estimation performance: Finite-temperature replica theory yields predictions for the corresponding replica symmetric order parameters identical to those provided by the zero-temperature replica theory for optimal M-estimation.

In particular, we find that the corresponding order parameters q_s^{Bayes} and q_d^{Bayes} in the finite-temperature replica theory satisfy precisely the same equations [Eq. (30)] that q_s^{opt} and q_d^{opt} satisfy in the zero-temperature theory for optimal M-estimation. This result implies an equivalence in performance between optimal M-estimation and Bayesian MMSE inference: $q_s^{\text{opt}} = q_s^{\text{Bayes}}$. This equivalence, in turn, implies that no algorithm whatsoever can outperform optimal convex M-estimation in the restricted scenario of log-concave signal and noise.

We note, however, that this equivalence between Bayes-optimal inference and optimal M-estimation is unlikely to hold in more general scenarios because a variety of non-log-concave signal distributions lead to hard MMSE inference problems that may not be solvable in polynomial time (see, e.g., Ref. [19]). Therefore, it is unlikely that a convex M-estimator that is solvable in polynomial time could match MMSE performance for such general distributions of signal and noise. However, even for the restricted setting of log-concave signal and noise, it is striking that two very different algorithms, namely, optimal M-estimation, solved via a convex optimization problem, and Bayesian inference, solved via a high-dimensional integral, yield identical performance.

Given the striking nature of this replica prediction, we test it numerically. It is computationally intractable to perform Bayes optimal MMSE inference by directly computing the high-dimensional integral in Eq. (3). However, in the asymptotic setting of high-dimensional, dense Gaussian measurements, with log-concave signal and noise distributions that we consider here, it is thought that an approximate message passing (AMP) procedure yields the same estimate for $\hat{\mathbf{s}}^{\text{MMSE}}$ obtained via the integral in Eq. (3) [37]. For the case of Laplacian signal and noise, we implemented this AMP procedure to numerically compute the optimal Bayes estimate $\hat{\mathbf{s}}^{\text{MMSE}}$ and compared its performance to the theoretical performance curve predicted by our zero-temperature replica theory for optimal M-estimation in Fig. 7, finding excellent agreement. Thus, this simulation provides numerical evidence for the replica prediction that the performance of optimal M-estimation is equivalent to Bayesian MMSE estimation in high dimensions.

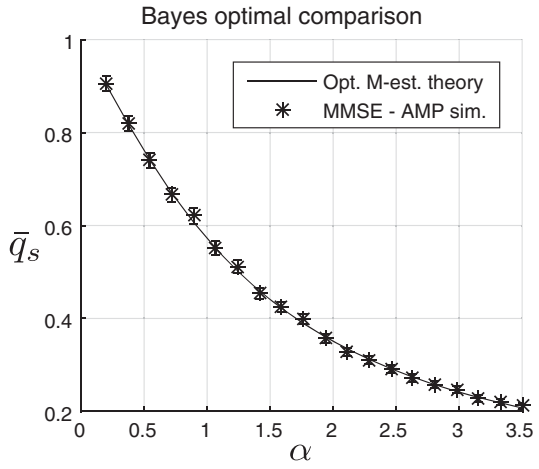

FIG. 7. A comparison between optimal M-estimation and Bayesian MMSE inference for the setting of Laplacian noise and signal ($E_\epsilon = |\epsilon|$, $E_s = |s^0|$, as also used in Fig. 6). We compare the normalized MSE, or fraction of unexplained variance $\bar{q}_s$ predicted by our theory of optimal regularized M-estimation (solid line), with simulations (error bars) of Bayes-optimal approximate message passing [37]. For our simulations, we randomly generated finite-size data (with N and P varying while $N = \alpha P$ and $\sqrt{NP} = 250$), and the error bars reflect standard deviations of message-passing performance calculated over 100 trials. We find an excellent match between optimal M-estimation theory and Bayesian AMP simulations.

F. Inference without noise

Motivated by compressed sensing, there has been a great deal of interest in understanding when and how we can perfectly infer the signal, so that $q_s = 0$, in the undersampled measurement regime $\alpha < 1$. This can only be done in the absence of noise ($\epsilon = 0$), but what properties must the signal distribution satisfy to guarantee such remarkable performance? In this special case of no noise, ϵ_{q_s} simply becomes a Gaussian variable with variance q_s, with Fisher information $J[\epsilon_{q_s}] = (1/q_s)$. Using this, and a relation between MMSE and Fisher information (Ref. [39], Appendix B. 4), the optimality formulas in Eq. (30) become

$$q_d^{\text{opt}} = \frac{q_s^{\text{opt}}}{\alpha} \qquad q_s^{\text{opt}} = q_d^{\text{opt}}\left(1 - q_d^{\text{opt}} J\left[s^0_{q_d^{\text{opt}}}\right]\right). \tag{36}$$

Partially eliminating q_d^{opt} yields

$$q_s^{\text{opt}} = \frac{\alpha(1-\alpha)}{J[s^0_{q_d^{\text{opt}}}]} \geq \frac{1-\alpha}{J[s^0]}. \tag{37}$$

Here, the inequality arises through an application of the convolutional Fisher inequality

$$\frac{1}{J[s^0_{q_d^{\text{opt}}}]} \geq \frac{1}{J[s^0]} + q_d^{\text{opt}}, \tag{38}$$

and then by fully eliminating q_d^{opt}.

Given that for any signal and noise distribution, we have proven that no convex inference procedure can achieve an error smaller than q_s^{opt}, Eq. (37) yields a general, sufficient, information theoretic condition for perfect recovery of the signal in the noiseless undersampled regime: The Fisher information of the signal distribution must diverge. This condition holds, for example, in sparse signal distributions that place finite probability mass at the origin. More generally, Eq. (37) yields a simple lower bound on noiseless, undersampled inference in terms of the measurement density and signal Fisher information. Moreover, in situations where the signal energy is convex, Eq. (29) remains the optimal inference procedure, while ρ^{opt} is replaced with a hard constraint enforcing optimization only over candidate signals $\mathbf{s}$ satisfying the noiseless measurement constraints $y^\mu = \mathbf{x}^\mu \cdot \hat{s}$.

III. DISCUSSION

In summary, our theoretical analyses, verified by simulations, yield a fundamental extension of time-honored results in low-dimensional classical statistics to the modern regime of high-dimensional inference, relevant in the current age of big data. In particular, we characterize the performance of any possible convex inference procedure for arbitrary signal and noise distributions [Eqs. (17) and (18)], we find fundamental information theoretic lower bounds on the error achievable by any convex procedure for arbitrary signal and noise [Eq. (31)], and we find the inference procedure that optimally exploits information about the signal and noise distributions, when their energies are convex [Eqs. (28) and (29)]. Moreover, we find a simple information theoretic condition for successful compressed sensing [Eq. (37)], or perfect inference without full measurement. These results generalize classical statistical results, based on Fisher information and the Cramer-Rao bound, that were discovered over 60 years ago. Intriguingly, there may be additional connections to classical statistical theorems that deserve further exploration in future work. One such theorem is the Rao-Blackwell theorem [45], proved in the 1950s, which demonstrates that any optimal estimator that achieves MMSE is a function of only the sufficient statistics of the noise distribution. Exploring relations between our work and extensions of this classical theorem that incorporate prior knowledge is an interesting future direction.

Moreover, our analysis uncovers several interesting surprises about the nature of optimal high-dimensional inference. In particular, we find that the optimal high-dimensional inference procedure is a smoothed version of ML in the unregularized case and a smoothed version of

MAP in the regularized case, where the amount of smoothing increases as the measurement density decreases or, equivalently, as the dimensionality increases. At low measurement densities and high dimensions, the optimal smoothed loss and regularization functions become simple quadratics [in the regularized case, this is provably true strictly at high SNR, but empirically, replacing the optimal loss with quadratic loss incurs very little performance decrement even at moderate SNR—Fig. 6(a)]. This observation reveals a fortuitous interplay between problem difficulty and algorithmic simplicity: At low measurement density, precisely when inference becomes statistically difficult, the optimal algorithm becomes computationally simple. Finally, we uncover phase transitions in the behavior of this simple quadratic inference algorithm, with a universal critical exponent in the decay of inference error with SNR at a critical measurement density [Eq. (23)].

Also, our analyses reveal several conceptual insights into the nature of overfitting and generalization in optimal high-dimensional inference through novel connections to scalar Bayesian inference in one dimension. This connection arises because of the nature of the mean-field theory of general high-dimensional inference, which can be expressed in terms of two coupled scalar estimation problems for the noise and signal, respectively (Fig. 3). In the optimal case, these scalar inference procedures based on proximal descent steps [Eq. (16)] become Bayesian inference procedures [Eq. (32)]. In particular, any inference algorithm implicitly decomposes the given measurements $y^\mu = \mathbf{x}^\mu \cdot \mathbf{s}^0 + \epsilon^\mu$ into a superposition of *estimated* signal and *estimated* noise: $y^\mu = \mathbf{x}^\mu \cdot \hat{\mathbf{s}} + \hat{\epsilon}^\mu$. The scalar Bayesian inference problems yield a MFT prediction for the error in estimating the signal (average per component L_2 discrepancy between $\mathbf{s}$ and $\hat{\mathbf{s}}$) and noise (average per component L_2 discrepancy between ϵ^μ and $\hat{\epsilon}^\mu$). Errors in inference arise because the noise ϵ^μ seeps into the estimated signal $\hat{\mathbf{s}}$. This inability to accurately separate signal and noise by even the optimal inference algorithm leads to divergent effects on the training and generalization error. The former decreases as the estimated signal $\hat{\mathbf{s}}$ acquires spurious correlations with the true noise ϵ^μ to explain the measurement outcomes y^μ. The latter increases because the noise in a held-out, previously unseen measurement outcome cannot possibly be correlated with the signal $\hat{\mathbf{s}}$ estimated from previously seen training data. Indeed, for the optimal inference algorithm, we find exceedingly simple quantitative relationships between inference errors of noise and signal, and high-dimensional training and generalization error [Eq. (34)]. This yields both quantitative and conceptual insight into the nature of overfitting in high dimensions, whereby training error can be far less than generalization error.

Finally, we also demonstrate a prediction of replica theory that no inference algorithm whatsoever can outperform our optimal M-estimator. We do so by deriving an equivalence between the replica prediction for the performance of the optimal M-estimator, derived using zero-temperature statistical mechanics, and the replica prediction for the performance of Bayesian MMSE inference, derived using unit-temperature statistical mechanics. This equivalence holds specifically when the signal and noise energies are convex or, equivalently, when their distributions are log-concave, and this excludes many interesting examples with nonconvex signal and noise energies in which MMSE inference is thought to be hard (not achievable in polynomial time). Even for this restricted class of log-concave signal and noise, this equivalence seems surprising since optimal M-estimation corresponds to solving an optimization problem, while Bayesian MMSE inference corresponds to solving an integration problem. Thus, at its heart, replica theory predicts a remarkable equivalence between optimization and integration. We provided numerical evidence for this prediction in Fig. 7. An understanding of this equivalence using rigorous, non-replica techniques constitutes an important direction for future work. We believe that proving the equivalence between these algorithms via approximate message-passing techniques may be a fruitful direction of approach.

Overall, our results illustrate the power of statistical-mechanics-based methods to generalize classical statistics to the modern regime of high-dimensional data analysis. We hope that these results will provide both firm theoretical guidance and practical algorithmic advantages in terms of both statistical and computational efficiency, to many fields spanning the ranges of science, engineering, and the humanities, as they all attempt to navigate the brave new world of big data.

ACKNOWLEDGMENTS

We thank Subhaneil Lahiri for useful discussions and also Alex Williams and Niru Maheswaranathan for comments on the manuscript. M. A. thanks the Stanford MBC and the Stanford Graduate Fellowship for support. S. G. thanks the Office of Naval Research, and the Burroughs Wellcome, Simons, Sloan, McKnight, and McDonnell Foundations for support.

Note added.—Upon completion of our work, we became aware of Ref. [46], which uses a different derivation technique to characterize the MSE of regularized M-estimation.

[1] T. J. Sejnowski, P. S. Churchland, and J. A. Movshon, *Putting Big Data to Good Use in Neuroscience*, Nat. Neurosci. **17**, 1440 (2014).

[2] S. Ganguli and H. Sompolinsky, *Compressed Sensing, Sparsity, and Dimensionality in Neuronal Information Processing and Data Analysis*, Annu. Rev. Neurosci. **35**, 485 (2012).

[3] R. Clarke, H. W. Ressom, A. Wang, J. Xuan, M. C. Liu, E. A. Gehan, and Y. Wang, *The Properties of High-Dimensional Data Spaces: Implications for Exploring Gene and Protein Expression Data*, Nat. Rev. Cancer **8**, 37 (2008).

[4] W. Raghupathi and V. Raghupathi, *Big Data Analytics in Healthcare: Promise and Potential*, Health Inf. Sci. Syst. **2**, 3 (2014).

[5] J. Fan, J. Lv, and L. Qi, *Sparse High Dimensional Models in Economics*, Ann. Rev. Econ. **3**, 291 (2011).

[6] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, *Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters*, Internet Math. **6**, 29 (2009).

[7] M. L. Jockers, *Macroanalysis: Digital Methods and Literary History* (University of Illinois Press, Champaign, IL, 2013).

[8] D. L. Donoho, *High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality*, *AMS Conference on Math Challenges of the 21st Century* (2000), pp. 1–32, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.329.3392.

[9] V. I. Serdobolskii, *Multivariate Statistical Analysis: A High-Dimensional Approach* (Springer Science & Business Media, Dordrecht, 2013), Vol. 41.

[10] M. Advani, S. Lahiri, and S. Ganguli, *Statistical Mechanics of Complex Neural Systems and High Dimensional Data*, J. Stat. Mech. (2013) P03014.

[11] M. Mezard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).

[12] D. J. Amit, H. Gutfreund, and H. Sompolinsky, *Statistical Mechanics of Neural Networks Near Saturation*, Ann. Phys. (N.Y.) **173**, 30 (1987).

[13] E. Gardner and B. Derrida, *Optimal Storage Properties of Neural Network Models*, J. Phys. A **21**, 271 (1988).

[14] E. Gardner, *The Space of Interactions in Neural Network Models*, J. Phys. A **21**, 257 (1988).

[15] D. Guo and S. Verdú, *Randomly Spread CDMA: Asymptotics via Statistical Physics*, IEEE Trans. Inf. Theory **51**, 1983 (2005).

[16] D. Gou, D. Baron, and S. Shamai, *A Single-Letter Characterization of Optimal Noisy Compressed Sensing*, in *Proceedings of the 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)* (IEEE, 2009), pp. 52–59, https://www.scholars.northwestern.edu/en/publications/a-single-letter-characterization-of-optimal-noisy-compressed-sens.

[17] S. Rangan, V. Goyal, and A. K. Fletcher, *Asymptotic Analysis of MAP Estimation via the Replica Method and Compressed Sensing*, in *Advances in Neural Information Processing Systems 22* (Curran Associates, Inc., Red Hook, 2009), pp. 1545–1553.

[18] S. Ganguli and H. Sompolinsky, *Statistical Mechanics of Compressed Sensing*, Phys. Rev. Lett. **104**, 188701 (2010).

[19] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová, *Probabilistic Reconstruction in Compressed Sensing: Algorithms, Phase Diagrams, and Threshold Achieving Matrices*, J. Stat. Mech. (2012) P08009.

[20] Y. Kabashima, F. Krzakala, M. Mézard, A. Sakata, and L. Zdeborová, *Phase Transitions and Sample Complexity in Bayes-Optimal Matrix Factorization*, IEEE Trans. Inf. Theory **62**, 4228 (2016).

[21] A. Engel and C. V. den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, England, 2001).

[22] H. Cramér, *Mathematical Methods of Statistics*, Princeton Mathematical Series (Princeton University Press, Princeton, 1946), Vol. 9.

[23] P. J. Huber, *Robust Regression: Asymptotics, Conjectures and Monte Carlo*, Ann. Stat. **1**, 821 (1973).

[24] A. W. van der Vaart, *Asymptotic Statistics* (Cambridge University Press, Cambridge, England, 2000), Vol. 3.

[25] P. Huber and E. Ronchetti, *Robust Statistics* (Wiley, New York, 2009).

[26] D. L. Donoho and M. Elad, *Optimally Sparse Representation in General (Non-orthogonal) Dictionaries via L1 Minimization*, Proc. Natl. Acad. Sci. U.S.A. **100**, 2197 (2003).

[27] E. Candes, J. Romberg, and T. Tao, *Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information.*, IEEE Trans. Inf. Theory **52**, 489 (2006).

[28] E. Candes and T. Tao, *Decoding by Linear Programming*, IEEE Trans. Inf. Theory **51**, 4203 (2005).

[29] R. Tibshirani, *Regression Shrinkage and Selection via the Lasso*, J. R. Stat. Soc. Ser. B **58**, 267 (1996).

[30] D. Bean, P. J. Bickel, N. El Karoui, and B. Yu, *Optimal M-estimation in High-Dimensional Regression*, Proc. Natl. Acad. Sci. U.S.A. **110**, 14563 (2013).

[31] D. Donoho and A. Montanari, *High Dimensional Robust M-estimation: Asymptotic Variance via Approximate Message Passing*, Probab. Theory Relat. Fields, doi:10.1007/s00440-015-0675-z (2013).

[32] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques* (MIT Press, Cambridge, MA, 2009).

[33] J. S. Yedidia, W. T. Freeman, and Y. Weiss, *Constructing Free-Energy Approximations and Generalized Belief Propagation Algorithms*, IEEE Trans. Inf. Theory **51**, 2282 (2005).

[34] M. Mezard and A. Montanari, *Information, Physics, and Computation* (Oxford University Press, New York, 2009).

[35] D. L. Donoho, A. Maleki, and A. Montanari, *Message-Passing Algorithms for Compressed Sensing*, Proc. Natl. Acad. Sci. U.S.A. **106**, 18914 (2009).

[36] M. Bayati and A. Montanari, *The Dynamics of Message Passing on Dense Graphs, with Applications to Compressed Sensing*, IEEE Trans. Inf. Theory **57**, 764 (2011).

[37] S. Rangan, *Generalized Approximate Message Passing for Estimation with Random Linear Mixing*, in *IEEE International Symposium on Information Theory Proceedings (ISIT), St. Petersburg, 2011* (IEEE, New York, 2011), pp. 2168–2172.

[38] N. El Karoui, *Asymptotic Behavior of Unregularized and Ridge-Regularized High-Dimensional Robust Regression Estimators: Rigorous Results*, arXiv:1311.2445.

[39] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevX.6.031034 for additional derivations.

[40] E. J. G. Pitman, *The Estimation of the Location and Scale Parameters of a Continuous Population of Any Given Form*, Biometrika **30**, 391 (1939).

[41] R. D. Gill and B. Y. Levit, *Applications of the van Trees Inequality: A Bayesian Cramér-Rao Bound*, Bernoulli **1**, 59 (1995).

[42] M. Mezard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).

[43] N. Parikh and S. Boyd, *Proximal Algorithms*, Found. Trends. Optimization **1**, 123 (2013).

[44] V. A. Marchenko and L. A. Pastur, *Distribution of Eigenvalues for Some Sets of Random Matrices*, Mat. Sb. **114**, 507 (1967).

[45] E. L. Lehmann and H. Scheffé, *Completeness, Similar Regions, and Unbiased Estimation: Part I*, Ind. J. Stat. **10**, 305 (1950).

[46] C. Thrampoulidis, E. Abbasi, and B. Hassibi, *Precise Error Analysis of Regularized M-estimators in High-Dimensions*, arXiv:1601.06233.