Exact results on high-dimensional linear regression via statistical physics

Alexander Mozeika,[1,*] Mansoor Sheikh,[2] Fabian Aguirre-Lopez,[3] Fabrizio Antenucci,[2] and Anthony C. C. Coolen[1,2,4]

[1]*London Institute for Mathematical Sciences, Royal Institution, 21 Albemarle Street, London W1S 4BS, United Kingdom*
[2]*Saddle Point Science Ltd., 10 Lincoln Street, York YO26 4YR, United Kingdom*
[3]*Université Paris-Saclay, CNRS, LPTMS, 91405, Orsay, France*
[4]*Department of Biophysics, Donders Institute, Radboud University, 6525AJ Nijmegen, The Netherlands*

(Received 23 November 2020; revised 1 April 2021; accepted 2 April 2021; published 29 April 2021)

It is clear that conventional statistical inference protocols need to be revised to deal correctly with the high-dimensional data that are now common. Most recent studies aimed at achieving this revision rely on powerful approximation techniques that call for rigorous results against which they can be tested. In this context, the simplest case of high-dimensional linear regression has acquired significant new relevance and attention. In this paper we use the statistical physics perspective on inference to derive several exact results for linear regression in the high-dimensional regime.

DOI: 10.1103/PhysRevE.103.042142

I. INTRODUCTION

The advent of modern high-dimensional data poses a significant challenge to statistical inference. The latter is understood well in the conventional regime of growing sample size with constant dimension. For high-dimensional data, where the dimension is of the same order as the sample size, the foundations of inference methods are still fragile, and even the simplest scenario of linear regression [1] has to be revised [2]. The study of linear regression (LR) in the high-dimensional regime has recently attracted significant attention in the mathematics [3–6] and statistical physics communities [7–9]. The statistical physics framework is naturally suited to deal with high-dimensional data.

While the connection between statistical physics and information theory was established a while ago by Jaynes [10], the approach has more recently been extended also to information processing [11] and machine learning [12]. In the statistical physics framework, the free energy encodes statistical properties of inference, akin to cumulant generating functions in statistics, but its direct computation via high-dimensional integrals is often difficult. This led to the development of several nonrigorous methods, such as the mean-field approximation, the replica trick and the cavity method [13]. Message passing in particular, which can be seen as algorithmic implementation of the latter [14], has emerged as an efficient analysis tool for statistical inference in high dimensions [15–17].

Most rigorous results on high-dimensional LR were obtained upon assuming uncorrelated data [4,8,15,17], possibly with sparsity of parameters [3,6]. Recently, correlations in sampling were analyzed in Ref. [16] for rotationally invariant data matrices. In all these studies, however, the parameters of the noise in the data were assumed *known*, unlike the standard statistical setting where they are usually inferred [1]. The exact prescription of the noise strength is unwelcome, since it is artificially removing an important source of overfitting in realistic applications of regression. In high-dimensional LR, inference protocols can mistake noise for signal, reflected in increased under-estimation of the noise and over-estimation of the magnitude of other model parameters (see Fig. 1).

In this paper we derive exact results for the high-dimensional regime of Bayesian LR which complement the aforementioned rigorous studies, using the statistical physics formulation of inference.

Statement of the problem and preview of results

We consider Bayesian inference of the LR model, $\mathbf{t} = \mathbf{Z}\boldsymbol{\theta} + \sigma\boldsymbol{\epsilon}$, where $\mathbf{t} \in \mathrm{IR}^N$ and $\mathbf{Z} \in \mathrm{IR}^{N\times d}$ are observed and the parameters $\boldsymbol{\theta} \in \mathrm{IR}^d$ and $\sigma \in \mathrm{IR}^+$ are to be inferred, with $\boldsymbol{\epsilon}$ denoting zero-average noise. We adopt a *teacher-student* scenario [18,19]: the teacher samples independently the rows of $\mathbf{Z}$ from some probability distribution $P(\mathbf{z})$ and then uses the LR model to obtain $\mathbf{t}$ with the *true* parameters $(\boldsymbol{\theta}_0, \sigma_0)$. We assume that the student then applies the Bayes formula to try to infer $(\boldsymbol{\theta}, \sigma)$ assuming a Gaussian[1] prior $\mathcal{N}(\boldsymbol{\theta}|\mathbf{0}, \eta^{-1}\mathbf{I}_d)$ for $\boldsymbol{\theta}$, and a generic prior $P(\sigma^2)$ for the noise parameter σ^2. Specifically, we do not consider the case where the observations are coming from an unknown source and/or where one needs to do model selection.

We map the LR inference problem onto a Gibbs-Boltzmann distribution with inverse 'temperature' β. This allows us to investigate properties of different inference protocols. In particular, maximum *a posteriori* (MAP) inference is obtained for $\beta \to \infty$ and $\eta > 0$, maximum likelihood (ML) inference for $\beta \to \infty$ and $\eta = 0$, and marginalization

[*] Corresponding author: alexander.mozeika@klc.ac.uk

[1] The distribution of a Gaussian (or Normal) random variable $\mathbf{x} \in \mathrm{IR}^d$, with mean $\boldsymbol{\mu} \in \mathrm{IR}^d$ and covariance $\boldsymbol{\Sigma} \in \mathrm{IR}^{d\times d}$, is given by the density $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}}{|2\pi\boldsymbol{\Sigma}|^{d/2}}$.

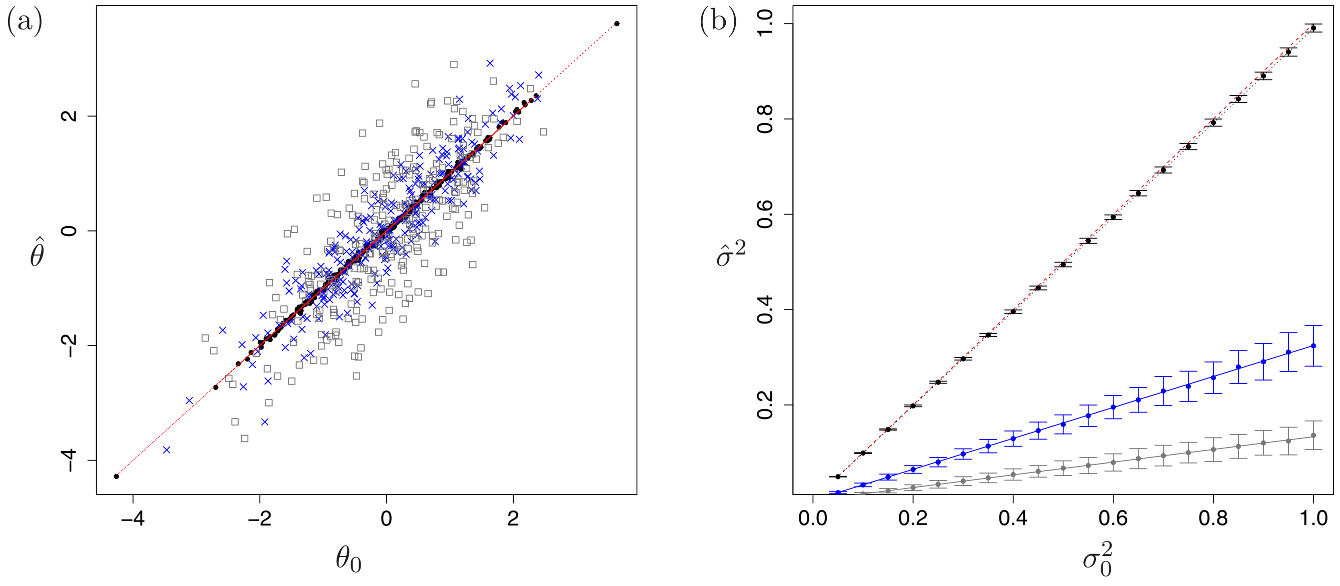

FIG. 1. High-dimensional phenomena in inference with the linear regression (LR) model in the teacher-student scenario (see text). Comparison between parameters inferred with maximum likelihood (ML) $(\hat{\boldsymbol{\theta}}, \hat{\sigma})$ and true values $(\boldsymbol{\theta}_0, \sigma_0)$. (a) Plot of the ordered set $\hat{\theta}(\theta_0) = \{[\hat{\theta}(1), \theta_0(1)], \ldots, [\hat{\theta}(d), \theta_0(d)]\}$ for $d/N \in \{0.01, 0.675, 0.867\}$, represented by symbols $\{\bullet, \times, \square\}$, with $N \in \{26000, 385, 300\}$. For each value of d/N the rows of $\mathbf{Z}$ were sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, $\boldsymbol{\epsilon}$ was sampled from $\mathcal{N}(\mathbf{0}, \sigma_0^2 \mathbf{I}_N)$, with $\sigma_0^2 = 0.1$, and $\boldsymbol{\theta}_0$ was sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. (b) Plot of $\hat{\sigma}^2$ versus σ_0^2, represented by points connected by lines, for $d/N \in \{0.01, 0.675, 0.867\}$ (top to bottom). Each point, together with ± one standard-deviation error-bars, represents an average over 250 samples. Note that in both plots the diagonal line corresponds to perfect inference.

inference for $\beta = 1$. We will refer to "ML (MAP) at finite temperature" for the case of $\eta = 0$ ($\eta > 0$) and β finite.

The *high-dimensional* regime is obtained for $(N, d) \to (\infty, \infty)$ with fixed ratio $\zeta = d/N \in (0, \infty)$. We will henceforth indicate this limit as $(N, d) \to \infty$, to simplify notation. Note that to keep $\mathbf{t}$ finite in the $(N, d) \to \infty$ limit, the matrix $\mathbf{Z}$ has to be replaced with $\mathbf{Z}/\sqrt{d}$ (unless of course we impose a suitable sparsity condition).

Within the above setting we obtain the following results: (i) If σ^2 is known and the distributions of $\mathbf{Z}$ and $\boldsymbol{\epsilon}$ are Gaussian, then we compute the distribution of the MAP and ML estimators of $\boldsymbol{\theta}$. (ii) The ML estimator $\hat{\sigma}^2_{\mathrm{ML}}$ of the noise parameter σ^2 is *self-averaging* as $(N, d) \to \infty$ (i.e., its variance is vanishing[2] in this limit), for any distributions of $\mathbf{Z}$ and $\boldsymbol{\epsilon}$. We bound the likelihood of deviations of $\hat{\sigma}^2_{\mathrm{ML}}$ from its mean for Gaussian noise $\boldsymbol{\epsilon}$. (iii) We compute the characteristic function of the *mean square error* $\frac{1}{d}||\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_{\mathrm{ML}}[\mathscr{D}]||^2$ for the ML estimator at finite (N, d), where $\boldsymbol{\theta}_0$ are the *true* parameters. (iv) We determine average and variance of the free energy, associated with Gibbs-Boltzmann distribution of Bayesian LR, of ML inference for the finite β and finite (N, d). The ML free-energy density is self-averaging as $(N, d) \to \infty$ if the eigenvalue *spectrum* of the empirical covariance matrix $\mathbf{Z}^{\mathrm{T}}\mathbf{Z}/N$ is self-averaging. For Gaussian $\boldsymbol{\epsilon}$ and $\mathbf{Z}$, we recover the results obtained by the replica method in Ref. [9]. (v) If the true parameters $\boldsymbol{\theta}_0$ are independent *random* variables, then we derive average and variance of the free energy of MAP inference for finite β and (N, d). The MAP free energy is shown to be self-averaging if the spectrum of $\mathbf{Z}^{\mathrm{T}}\mathbf{Z}/N$ is self-averaging as $(N, d) \to \infty$.

In the following subsections we describe how the above results were obtained, with full mathematical details relegated to the Supplemental Material [21].

II. STATISTICAL PHYSICS AND BAYESIAN INFERENCE

We assume that we observe a data sample of "input-output" pairs $\{(\mathbf{z}_1, t_1), \ldots, (\mathbf{z}_N, t_N)\}$, where $(\mathbf{z}_i, t_i) \in \mathrm{IR}^{d+1}$, generated randomly and independently from

$$P(t, \mathbf{z}|\boldsymbol{\Theta}) = P(t|\mathbf{z}, \boldsymbol{\Theta})P(\mathbf{z}), \tag{1}$$

with parameters $\boldsymbol{\Theta}$ that are unknown to us. If we assume a prior distribution $P(\boldsymbol{\Theta})$, then the distribution of $\boldsymbol{\Theta}$, given the data, follows from the Bayes formula

$$P(\boldsymbol{\Theta}|\mathscr{D}) = \frac{P(\boldsymbol{\Theta})\prod_{i=1}^N P(t_i|\mathbf{z}_i, \boldsymbol{\Theta})}{\int d\tilde{\boldsymbol{\Theta}} P(\tilde{\boldsymbol{\theta}})\{\prod_{i=1}^N P(t_i|\mathbf{z}_i, \tilde{\boldsymbol{\Theta}})\}}. \tag{2}$$

Here $\mathscr{D} = \{\mathbf{t}, \mathbf{Z}\}$, with $\mathbf{t} = (t_1, \ldots, t_N)$, and $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_N)$ is an $N \times d$ matrix. In Bayesian language, Eq. (2) is the posterior distribution of $\boldsymbol{\Theta}$, given the prior distribution $P(\boldsymbol{\Theta})$ and the observed data $\mathscr{D}$.

The simplest way to use Eq. (2) for inference is to compute the *maximum a posteriori* (MAP) estimator

$$\hat{\boldsymbol{\Theta}}_{\mathrm{MAP}}[\mathscr{D}] = \mathrm{argmin}_{\boldsymbol{\Theta}}\, E(\boldsymbol{\Theta}|\mathscr{D}), \tag{3}$$

[2] Here we adopt the definition from statistical physics of disordered systems which states that some density, such as free energy, average energy, etc., is self-averaging if its variance is vanishing in the thermodynamic limit [13]. In statistics this phenomena is also referred to as having a "fully concentrated measure" [6,20].

in which the so-called Bayesian likelihood function

$$E(\mathbf{\Theta}|\mathscr{D}) = -\sum_{i=1}^{N} \log P(t_i|\mathbf{z}_i, \mathbf{\Theta}) - \log P(\mathbf{\Theta}) \tag{4}$$

consists of a first term, the log-likelihood used also in *maximum likelihood* (ML) inference, and a second term, that acts as a regularizer. Bayesian inference can thus be seen as a generalization of MAP inference, and MAP inference generalizes ML inference.

The *square error* $\frac{1}{d}||\mathbf{\Theta}_0 - \hat{\mathbf{\Theta}}[\mathscr{D}]||^2$, with the Euclidean norm $||\cdots||$ and the *true* parameters $\mathbf{\Theta}_0$ underlying the data, is often used to quantify the quality of inference in Eq. (3). Its first moment is the *mean square error* (MSE) $\frac{1}{d}\langle\langle||\mathbf{\Theta}_0 - \hat{\mathbf{\Theta}}[\mathscr{D}]||^2\rangle_{\mathscr{D}}\rangle_{\mathbf{\Theta}_0}$. Furthermore, the posterior mean

$$\hat{\mathbf{\Theta}}[\mathscr{D}] = \int d\mathbf{\Theta}\; P(\mathbf{\Theta}|\mathscr{D})\mathbf{\Theta} \tag{5}$$

(the *marginalization* estimator) is the *minimum* MSE (MMSE) estimator in the Bayes optimal case, i.e., when prior distribution and model likelihood are known [19].

The above approaches to Bayesian inference can be unified conveniently in a single statistical physics (SP) formulation via the Gibbs-Boltzmann distribution

$$P_\beta(\mathbf{\Theta}|\mathscr{D}) = \frac{e^{-\beta E(\mathbf{\Theta}|\mathscr{D})}}{Z_\beta[\mathscr{D}]}, \tag{6}$$

with the normalization constant, or "partition function" $Z_\beta[\mathscr{D}] = \int d\mathbf{\Theta}\; e^{-\beta E(\mathbf{\Theta}|\mathscr{D})}$. For $\beta = 1$ this is the *evidence* term of Bayesian inference. In statistical physics language, Eq. (4) plays the role of "energy" in Eq. (6) and β is the (fictional) inverse temperature. The temperature can be interpreted as a noise amplitude in stochastic gradient descent minimization of $E(\mathbf{\Theta}|\mathscr{D})$ [9]. Properties of the system Eq. (6) follow upon evaluating the "free energy,"

$$F_\beta[\mathscr{D}] = -\frac{1}{\beta}\log Z_\beta[\mathscr{D}]. \tag{7}$$

The estimator Eqs. (3) and (5) are recovered from the average $\int d\mathbf{\Theta}\; P_\beta(\mathbf{\Theta}|\mathscr{D})\mathbf{\Theta}$ by taking the zero "temperature" limit $\beta \to \infty$, or by setting $\beta = 1$, respectively. This follows upon observing that for $\beta = 1$ the distribution Eq. (6) and the posterior Eq. (2) are identical, and that $\hat{\mathbf{\Theta}}_{\mathrm{MAP}}[\mathscr{D}] = \lim_{\beta\to\infty}\int d\mathbf{\Theta}\; P_\beta(\mathbf{\Theta}|\mathscr{D})\mathbf{\Theta}$ by the Laplace argument[3]. We note that the interpretation of the MAP estimator Eq. (3) in the SP framework Eq. (6) is that $\hat{\mathbf{\Theta}}_{\mathrm{MAP}}[\mathscr{D}]$ is the "ground state" of the system. Regarding the formulation Eq. (6) we stress that, even though the (inverse) temperatures $\beta = 1$ and $\beta \to \infty$ are the most common for inference, the benefits of the generic "thermal" noise are well known for optimization problems in general [23] and for Bayesian inference in particular [24].

The Kullback-Leibler (KL) "distance" [25] between the distribution $P(t, \mathbf{z}|\mathbf{\Theta})$ and its empirical counterpart $\hat{P}(t, \mathbf{z}|\mathscr{D}) = N^{-1}\sum_i \delta(t - t_i)\delta(\mathbf{z} - \mathbf{z}_i)$, given by

$$D(\hat{P}[\mathscr{D}]||P_{\mathbf{\Theta}}) = \int dt\, d\mathbf{z}\, \hat{P}(t, \mathbf{z}|\mathscr{D}) \log\left(\frac{\hat{P}(t, \mathbf{z}|\mathscr{D})}{P(t, \mathbf{z}|\mathbf{\Theta})}\right) \tag{8}$$

can also be used to obtain the ML estimator, via $\hat{\mathbf{\Theta}}_{\mathrm{ML}}[\mathscr{D}] = \mathrm{argmin}_{\mathbf{\Theta}} D(\hat{P}||P_{\mathbf{\Theta}})$. Furthermore, since $ND(\hat{P}[\mathscr{D}]||P_{\mathbf{\Theta}}) = E(\mathbf{\Theta}|\mathscr{D}) + \log P(\mathbf{\Theta}) - NS(\hat{P}[\mathscr{D}])$, where the last term, minus the Shannon entropy of $\hat{P}[\mathscr{D}]$, is independent of $\mathbf{\Theta}$, the MAP estimator can be obtained via $\hat{\mathbf{\Theta}}_{\mathrm{MAP}}[\mathscr{D}] = \mathrm{argmin}_{\mathbf{\Theta}}\{ND(\hat{P}||P_{\mathbf{\Theta}}) - \log P(\mathbf{\Theta})\}$.

Finally, the KL distance Eq. (8) can also be used to define the difference $\Delta D(\mathbf{\Theta}, \mathbf{\Theta}_0|\mathscr{D}) = D(\hat{P}[\mathscr{D}]||P_{\mathbf{\Theta}}) - D(\hat{P}[\mathscr{D}]||P_{\mathbf{\Theta}_0})$, where $\mathbf{\Theta}_0$ are the true parameters responsible for the data, which served as a useful measure of over-fitting in ML inference [26], and was recently extended to MAP inference in generalized linear models [9]. Both latter studies used the SP framework, equivalent to Eq. (7), to study *typical* (as opposed to *worst-case*) properties of inference in the *high-dimensional* regime via the average free-energy $\langle F_\beta[\mathscr{D}]\rangle_{\mathscr{D}}/N$ as computed by the replica method [13].

III. BAYESIAN LINEAR REGRESSION

In Bayesian linear regression (LR) with Gaussian priors, also called *ridge regression*, it is assumed that the data $(\mathbf{z}_i, t_i)$ are for all i sampled independently from the distribution $\mathcal{N}(t|\boldsymbol{\theta}.\mathbf{z}, \sigma^2)P(\mathbf{z})$, so the energy Eq. (4), with $\mathbf{\Theta} \equiv \{\boldsymbol{\theta}, \sigma^2\}$, is given by

$$E(\boldsymbol{\theta}, \sigma^2|\mathscr{D}) = \frac{1}{2\sigma^2}||\mathbf{t} - \mathbf{Z}\boldsymbol{\theta}||^2 + \frac{1}{2}\eta||\boldsymbol{\theta}||^2 + \frac{1}{2}N\log(2\pi\sigma^2) - \log P(\sigma^2), \tag{9}$$

where $\eta \geqslant 0$ is the hyper-parameter for the Gaussian prior $P(\boldsymbol{\theta})$ and $P(\sigma^2)$ is a generic prior. The true parameters of $\mathscr{D}$ are written as $\boldsymbol{\theta}_0$ and σ_0^2, i.e., we assume that $\mathbf{t} = \mathbf{Z}\boldsymbol{\theta}_0 + \boldsymbol{\epsilon}$ with the noise vector $\boldsymbol{\epsilon}$ being sampled from some distribution, e.g., the multivariate Gaussian $\mathcal{N}(\mathbf{0}, \sigma_0^2\mathbf{I}_N)$, with mean $\mathbf{0}$ and covariance $\sigma_0^2\mathbf{I}_N$.

The energy function can be expressed as

$$E(\boldsymbol{\theta}, \sigma^2|\mathscr{D}) = \frac{\left(\boldsymbol{\theta} - \mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{t}\right)^{\mathrm{T}}\mathbf{J}_{\sigma^2\eta}\left(\boldsymbol{\theta} - \mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{t}\right)}{2\sigma^2} + \frac{\mathbf{t}^{\mathrm{T}}\left(\mathbf{I}_N - \mathbf{Z}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{Z}^{T}\right)\mathbf{t}}{2\sigma^2} + \frac{N\log(2\pi\sigma^2) - 2\log P(\sigma^2)}{2}, \tag{10}$$

where we defined the $d \times d$ matrix $\mathbf{J} = \mathbf{Z}^{\mathrm{T}}\mathbf{Z}$, with elements $[\mathbf{J}]_{k\ell} = \sum_{i=1}^{N} z_i(k)z_i(\ell)$, and its "regularized" version $\mathbf{J}_{\sigma^2\eta} = \mathbf{J} + \sigma^2\,\eta\,\mathbf{I}_d$. The distribution Eq. (6) is now

$$P_\beta(\boldsymbol{\theta}, \sigma^2|\mathscr{D}) = \frac{P_\beta(\boldsymbol{\theta}|\,\sigma^2, \mathscr{D})e^{-\beta\left[F_{\beta,\sigma^2}[\mathscr{D}] + \frac{1}{2}N\log(2\pi\sigma^2) - \log P(\sigma^2)\right]}}{\int_0^\infty d\tilde{\sigma}^2\, e^{-\beta\left[F_{\beta,\tilde{\sigma}^2}[\mathscr{D}] + \frac{1}{2}N\log\tilde{\sigma}^2 - \log P(\tilde{\sigma}^2)\right]}},$$

where $P_\beta(\boldsymbol{\theta}|\sigma^2, \mathscr{D})$ is the Gaussian distribution

$$P_\beta(\boldsymbol{\theta}|\,\sigma^2, \mathscr{D}) = \mathcal{N}\left(\boldsymbol{\theta}\middle|\,\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{t},\; \sigma^2\beta^{-1}\mathbf{J}_{\sigma^2\eta}^{-1}\right), \tag{11}$$

[3] In this work we will mainly rely on the identities $\lim_{M\to\infty} -\frac{1}{M}\log\int d\mathbf{x}\, e^{-M\phi(\mathbf{x})} = \phi(\mathbf{x}_0)$, where $\mathbf{x}_0 = \mathrm{argmin}_{\mathbf{x}}\phi(\mathbf{x})$, and $\lim_{M\to\infty}\int d\mathbf{x}\frac{e^{-M\phi(\mathbf{x})}}{\int d\tilde{\mathbf{x}}\, e^{-M\phi(\tilde{\mathbf{x}})}}g(\mathbf{x}) = g(\mathbf{x}_0)$ for sufficiently smooth and well behaved functions ϕ, g of $\mathbf{x} \in \mathrm{IR}^p$ with $p = O(M^0)$ [22].

with mean $\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{t}$ and covariance $\sigma^2\beta^{-1}\mathbf{J}_{\sigma^2\eta}^{-1}$. We have also defined the *conditional* free energy,

$$F_{\beta,\sigma^2}[\mathscr{D}] = \frac{d}{2\beta} + \frac{1}{2\sigma^2}\mathbf{t}^{\mathrm{T}}\big(\mathbf{I}_N - \mathbf{Z}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{Z}^T\big)\mathbf{t} - \frac{1}{2\beta}\log\big|2\pi e\sigma^2\beta^{-1}\mathbf{J}_{\sigma^2\eta}^{-1}\big|, \tag{12}$$

while the full free energy associated with Eq. (11) is given by

$$\begin{aligned} F_\beta[\mathscr{D}] &= -\frac{1}{\beta}\log\int d\boldsymbol{\theta}\, d\sigma^2\, e^{-\beta E(\boldsymbol{\theta},\sigma^2|\mathscr{D})} \\ &= -\frac{1}{\beta}\log\int_0^\infty d\sigma^2 e^{-\beta\left[F_{\beta,\sigma^2}[\mathscr{D}]+\frac{N}{2}\log(2\pi\sigma^2)-\log P(\sigma^2)\right]}. \end{aligned} \tag{13}$$

For $\beta\to\infty$ the free energy is via the Laplace argument given by $F_\infty[\mathscr{D}] = \min_{\boldsymbol{\theta},\sigma^2} E(\boldsymbol{\theta},\sigma^2|\mathscr{D})$.

$F_\infty[\mathscr{D}]$ is the ground state energy of Eq. (11). The ground state $\{\hat{\boldsymbol{\theta}}[\mathscr{D}],\hat{\sigma}^2[\mathscr{D}]\} = \mathrm{argmin}_{\boldsymbol{\theta},\sigma^2}E(\boldsymbol{\theta},\sigma^2|\mathscr{D})$ is given by

$$\hat{\boldsymbol{\theta}}[\mathscr{D}] = \mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{t}, \tag{14}$$

i.e., the mean of Eq. (11), and the solution of the equation

$$\sigma^2 = \frac{1}{N}||\mathbf{t}-\mathbf{Z}\hat{\boldsymbol{\theta}}[\mathscr{D}]||^2 + \frac{2\sigma^4}{N}\frac{\partial}{\partial\sigma^2}\log P(\sigma^2), \tag{15}$$

corresponding to the MAP estimators of the parameters.[4] From the second line in Eq. (13) we infer

$$F_\infty[\mathscr{D}] = \min_{\sigma^2}\left[F_{\infty,\sigma^2}[\mathscr{D}] + \frac{N\log(2\pi\sigma^2)}{2} - \log P(\sigma^2)\right], \tag{16}$$

(again via the Laplace argument), as well as for $(N,d)\to\infty$ the free-energy density $f_\beta[\mathscr{D}] = \frac{1}{N}F_\beta[\mathscr{D}]$ at any β:

$$f_\beta[\mathscr{D}] = \min_{\sigma^2}\left[\frac{F_{\beta,\sigma^2}[\mathscr{D}]}{N} + \frac{\log(2\pi\sigma^2)}{2} - \frac{\log P(\sigma^2)}{N}\right]. \tag{17}$$

For $\beta = 1$ the distribution Eq. (11) can be used to compute the MMSE estimators of $\boldsymbol{\theta}$ and σ^2, given by the averages

$$\begin{aligned} &\int_0^\infty d\boldsymbol{\theta}\, d\sigma^2\, P_\beta(\boldsymbol{\theta},\sigma^2|\mathscr{D})\,\boldsymbol{\theta} = \big\langle \mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{t}\big\rangle_{\sigma^2}, \\ &\int_0^\infty d\boldsymbol{\theta}\, d\sigma^2\, P_\beta(\boldsymbol{\theta},\sigma^2|\mathscr{D})\,\sigma^2 = \langle\sigma^2\rangle_{\sigma^2}, \end{aligned} \tag{18}$$

where the short-hand $\langle\cdots\rangle_{\sigma^2}$ refers to averaging over the following marginal of the distribution Eq. (11):

$$P_\beta(\sigma^2|\mathscr{D}) = \frac{e^{-\beta\left[F_{\beta,\sigma^2}[\mathscr{D}]+\frac{N}{2}\log(2\pi\sigma^2)-\log P(\sigma^2)\right]}}{\int_0^\infty d\tilde{\sigma}^2\, e^{-\beta\left[F_{\beta,\tilde{\sigma}^2}[\mathscr{D}]+\frac{N}{2}\log(2\pi\tilde{\sigma}^2)-\log P(\tilde{\sigma}^2)\right]}}. \tag{19}$$

If the density $F_{\beta,\sigma^2}[\mathscr{D}]/N$ is self-averaging, then for $(N,d)\to\infty$ this marginal is dominated by the solution of Eq. (17). The dominant value of $\boldsymbol{\theta}$ in Eq. (18) is Eq. (14), but with σ^2 being the solution of the following equation, which for $\beta = 1$ gives the MMSE estimators, and which recovers the MAP estimators Eqs. (14) and (15) when $\beta\to\infty$:

$$\sigma^2 = \frac{\beta}{(\beta-\zeta)}\frac{1}{N}||\mathbf{t}-\mathbf{Z}\hat{\boldsymbol{\theta}}[\mathscr{D}]||^2 - \frac{\sigma^4\eta}{(\beta-\zeta)}\frac{1}{N}\mathrm{Tr}\big[\mathbf{J}_{\sigma^2\eta}^{-1}\big] + \frac{2\sigma^4\beta}{(\beta-\zeta)N}\frac{\partial}{\partial\sigma^2}\log P(\sigma^2). \tag{20}$$

The free energies Eqs. (12) and (13) obey the Helmholtz free-energy relations. In particular, with $E(\boldsymbol{\theta}|\mathscr{D}) = E(\boldsymbol{\theta},\sigma^2|\mathscr{D}) - \frac{1}{2}N\log(2\pi\sigma^2) + \log P(\sigma^2)$ we get

$$F_{\beta,\sigma^2}[\mathscr{D}] = E_\beta[\mathscr{D}] - T\,\mathrm{S}_\beta[\mathscr{D}], \tag{21}$$

where $T = 1/\beta$, with the average energy

$$E_\beta[\mathscr{D}] = \int d\boldsymbol{\theta}\, P_\beta(\boldsymbol{\theta}|\sigma^2,\mathscr{D})E(\boldsymbol{\theta}|\mathscr{D}), \tag{22}$$

and with the *differential entropy*

$$\mathrm{S}_\beta[\mathscr{D}] = -\int d\boldsymbol{\theta}\, P_\beta(\boldsymbol{\theta}|\sigma^2,\mathscr{D})\log P_\beta(\boldsymbol{\theta}|\sigma^2,\mathscr{D}). \tag{23}$$

In the free-energy Eq. (12) we have

$$\begin{aligned} E_\beta[\mathscr{D}] &= \frac{d}{2\beta} + \frac{1}{2\sigma^2}\mathbf{t}^{\mathrm{T}}\big(\mathbf{I}_N - \mathbf{Z}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{Z}^T\big)\mathbf{t}, \\ \mathrm{S}_\beta[\mathscr{D}] &= \frac{1}{2}\log\big|2\pi e\sigma^2\beta^{-1}\mathbf{J}_{\sigma^2\eta}^{-1}\big|. \end{aligned} \tag{24}$$

Furthermore, the average energy can be written as

$$E_\beta[\mathscr{D}] = \frac{d}{2\beta} + \min_{\boldsymbol{\theta}} E(\boldsymbol{\theta},\sigma^2|\mathscr{D}). \tag{25}$$

We stress that the formulation of LR Bayesian inference via a Gibbs-Boltzmann distribution is not new, see, e.g., Refs. [19,26]. However, unlike most previous works, here we also consider the case of unknown σ and we keep the temperature generic for most of the analysis, instead of limiting ourselves to the familiar cases $T\in\{0,1\}$. In the case in which the noise parameter σ^2 is known, i.e., $P(\sigma^2) = \delta(\sigma^2-\sigma_0^2)$, our free-energy expression reduces to $F_\beta[\mathscr{D}] = F_{\beta,\sigma_0^2}[\mathscr{D}] - \frac{N}{2\beta}\log(2\pi\sigma_0^2)$ and $P_\beta(\boldsymbol{\theta},\sigma^2|\mathscr{D}) = P_\beta(\boldsymbol{\theta}|\sigma_0^2,\mathscr{D})\delta(\sigma^2-\sigma_0^2)$.

A. Distribution of estimators $\hat{\boldsymbol{\theta}}_{\mathrm{MAP}}$ and $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$

If the noise parameter σ^2 is independent of the realization of the data $\mathscr{D}$, e.g., σ^2 is known or self-averaging as $(N,d)\to\infty$, and the noise $\boldsymbol{\epsilon}$ has Gaussian statistics $\mathcal{N}(\mathbf{0},\sigma_0^2\mathbf{I}_N)$, then the distribution of the MAP estimator Eq. (14) is

$$P(\hat{\boldsymbol{\theta}}) = \big\langle\mathcal{N}\big(\hat{\boldsymbol{\theta}}\big|\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{J}\boldsymbol{\theta}_0,\ \sigma_0^2\mathbf{J}_{\sigma^2\eta}^{-2}\mathbf{J}\big)\big\rangle_{\mathbf{Z}}. \tag{26}$$

For $\eta = 0$, i.e., ML inference, and without averaging over $\mathbf{Z}$, this recovers Theorem 7.6b in Ref. [1]. To probe the $(N,d)\to\infty$ regime we rescale $z_i(\mu)\to z_i(\mu)/\sqrt{d}$ with now $z_i(\mu) = \mathcal{O}(1)$. This gives $\mathbf{J} = \mathbf{C}/\zeta$ and $\mathbf{J}_{\sigma^2\eta} = \mathbf{C}_{\zeta\sigma^2\eta}/\zeta$, with the sample covariance matrix $[\mathbf{C}]_{\mu\nu} = N^{-1}\sum_{i=1}^N z_i(\mu)z_i(\nu)$ and $\mathbf{C}_{\zeta\sigma^2\eta} = \mathbf{C}+\zeta\sigma^2\eta\mathbf{I}$, so

$$P(\hat{\boldsymbol{\theta}}) = \big\langle\mathcal{N}\big(\hat{\boldsymbol{\theta}}\big|\mathbf{C}_{\zeta\sigma^2\eta}^{-1}\mathbf{C}\boldsymbol{\theta}_0,\ \zeta\sigma_0^2\mathbf{C}_{\zeta\sigma^2\eta}^{-2}\mathbf{C}\big)\big\rangle_{\mathbf{Z}}. \tag{27}$$

[4] If the inverse-χ^2 distribution is used as a prior for σ^2, then the MAP estimator for the latter is given by $\sigma^2 = \frac{1}{N+\nu+2} + \frac{1}{N+\nu+2}||\mathbf{t}-\mathbf{Z}\hat{\boldsymbol{\theta}}[\mathscr{D}]||^2$ which suggests that the hyper-parameter ν has to be *extensive* to remain relevant for large N.

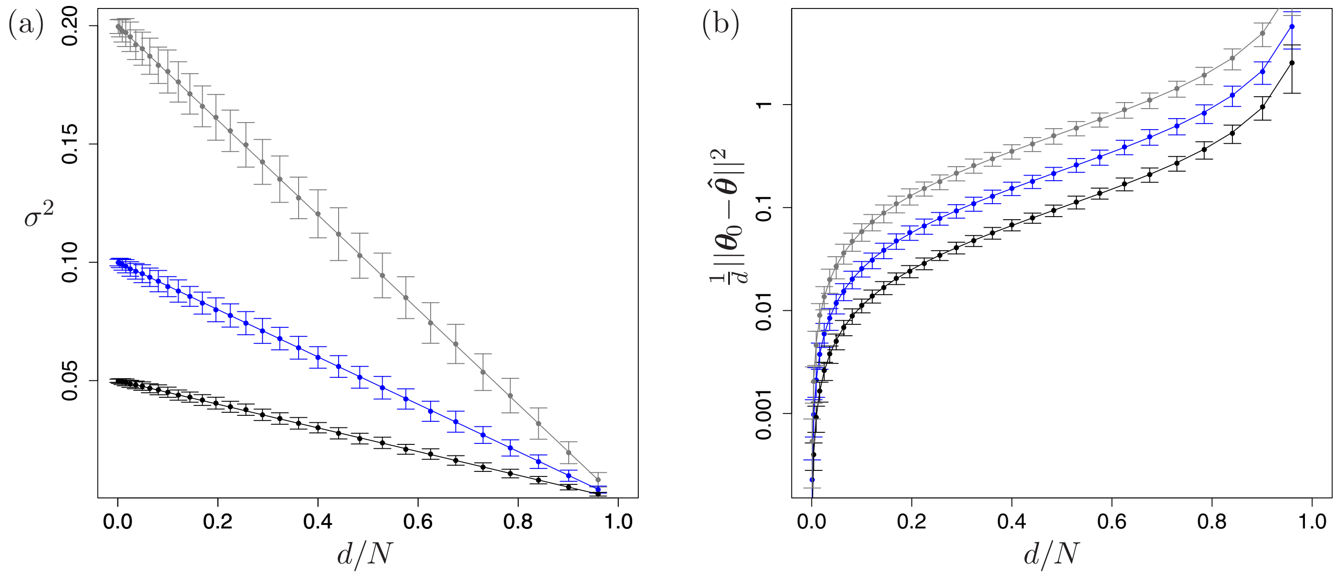

FIG. 2. Theoretical predictions for ML inference of the LR model $\mathbf{t}=\mathbf{Z}\boldsymbol{\theta}+\sigma\boldsymbol{\epsilon}$, where $\mathbf{t}\in\mathrm{IR}^N$ and $\mathbf{Z}\in\mathrm{IR}^{N\times d}$, in the high-dimensional regime $0<d/N<1$. For each sample the rows of $\mathbf{Z}$ were sampled from the Gaussian $\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})$. The covariance matrix $\boldsymbol{\Sigma}$ is such that $[\boldsymbol{\Sigma}]_{\nu,\nu}=1$, $[\boldsymbol{\Sigma}]_{\nu,\nu+1}=[\boldsymbol{\Sigma}]_{\nu+1,\nu}=\epsilon$ with $0\leqslant\epsilon<1$ for ν odd, and $[\boldsymbol{\Sigma}]_{\nu_1,\nu_2}=0$ for all other $\nu_1\neq\nu_2$. The density of eigenvalues of $\boldsymbol{\Sigma}$ is given exactly by $\rho(\lambda)=\frac{1}{2}\delta(\lambda-1-\epsilon)+\frac{1}{2}\delta(\lambda-1+\epsilon)$ for any even d. For each sample the noise vector $\boldsymbol{\epsilon}$ was sampled from $\mathcal{N}(\mathbf{0},\sigma_0^2\mathbf{I}_N)$. For each value of d/N the (true) parameter $\boldsymbol{\theta}_0$ was sampled from $\mathcal{N}(\mathbf{0},\mathbf{I}_d)$ only once. (a) Inferred noise parameter σ^2 as a function of d/N, where (d,N) pairs range from $(10,10^4)$ to (310,323), plotted for true value of noise $\sigma_0^2\in\{0.05,0.10,0.20\}$ (top to bottom). Solid lines are the averages predicted by the theory in Eq. (32) and symbols, with ± one standard-deviation error-bars, are empirical averages over the 250 samples of data $\{\mathbf{t},\mathbf{Z}\}$. Error-bars are consistent with the variance predicted by the theory in Eq. (32). (b) MSE as a function of d/N plotted for $\epsilon\in\{0,0.75,0.9\}$ (bottom to top) and $\sigma_0^2=0.1$. Solid lines correspond to the theoretical prediction $\frac{\zeta\sigma_0^2}{1-\zeta-N^{-1}}\frac{1}{1-\epsilon^2}$ for average MSE, computed via Eq. (39), and symbols, with ± one standard-deviation error-bars, are empirical averages over the 250 samples. Note the logarithmic scale of the vertical axis. Error-bars are consistent with the variance, $\frac{2\zeta^2\sigma_0^4}{(1-\zeta)^2}\frac{1+\epsilon^2}{d(1+\epsilon)^2(1-\epsilon)^2}$, as predicted by Eq. (39).

Furthermore, for a Gaussian sample with true covariance matrix $\boldsymbol{\Sigma}$, i.e., if each $\mathbf{z}_i$ in $\mathbf{Z}$ is drawn independently from $\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})$, then the distribution of $\hat{\boldsymbol{\theta}}$ for any finite (N,d) is the Gaussian mixture

$$P(\hat{\boldsymbol{\theta}})=\int d\mathbf{C}\;\mathcal{W}(\mathbf{C}|\boldsymbol{\Sigma}/N,d,N)\mathcal{N}\big(\hat{\boldsymbol{\theta}}\big|\mathbf{C}^{-1}_{\zeta\sigma^2\eta}[\mathbf{C}]\,\mathbf{C}\boldsymbol{\theta}_0,\ \zeta\sigma_0^2\mathbf{C}^{-2}_{\zeta\sigma^2\eta}[\mathbf{C}]\,\mathbf{C}\big). \tag{28}$$

The integral is over all symmetric positive definite $d\times d$ matrices, and $\mathcal{W}(\mathbf{C}|\boldsymbol{\Sigma}/N,d,N)$ is the Wishart distribution, which is nonsingular when $d\leqslant N$. Note that Eq. (28) also represents the distribution of "ground states" of Eq. (11).

For $\eta=0$ the distribution Eq. (28) becomes the multivariate Student's t-distribution with $N+1-d$ degrees of freedom:

$$P(\hat{\boldsymbol{\theta}})=\frac{\Gamma\big(\frac{N+1}{2}\big)}{\Gamma\big(\frac{N+1-d}{2}\big)}\left|\frac{(1-\zeta+1/N)\boldsymbol{\Sigma}}{\pi(N+1-d)\zeta\sigma_0^2}\right|^{\frac{1}{2}}\left[1+(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^{\mathrm{T}}\frac{(1-\zeta+1/N)\boldsymbol{\Sigma}}{(N+1-d)\zeta\sigma_0^2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\right]^{-\frac{N+1}{2}}. \tag{29}$$

The vector of true parameters $\boldsymbol{\theta}_0$ is the mode and $[\zeta\sigma_0^2/(1-\zeta-N^{-1})]\boldsymbol{\Sigma}^{-1}$ is the covariance matrix of Eq. (29). In the regime $(N,d)\to\infty$ one can recover from Eq. (29) the moments of the multivariate Gaussian suggested by the replica method [9]. In this regime one indeed finds that any *finite* subset of components of $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ is described by a Gaussian distribution [27,28].

B. Statistical properties of the estimator $\hat{\sigma}^2_{\mathrm{ML}}$

For $\eta=0$ the estimator Eq. (14) simplifies considerably to

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}}[\mathscr{D}]=(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{t}, \tag{30}$$

giving us, via Eq. (15), the ML noise estimator

$$\hat{\sigma}^2_{\mathrm{ML}}=\frac{1}{N}\boldsymbol{\epsilon}^{\mathrm{T}}(\mathbf{I}_N-\mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}})\boldsymbol{\epsilon}. \tag{31}$$

In particular, if the noise $\boldsymbol{\epsilon}$ originates from a distribution with mean $\mathbf{0}$ and covariance $\sigma_0^2\,\mathbf{I}_N$, then the mean and variance of $\hat{\sigma}^2_{\mathrm{ML}}$ are

$$\big\langle\hat{\sigma}^2_{\mathrm{ML}}\big\rangle_{\boldsymbol{\epsilon}}=\sigma_0^2(1-\zeta),\qquad \mathrm{Var}\big(\hat{\sigma}^2_{\mathrm{ML}}\big)=\frac{2\sigma_0^4}{N}(1-\zeta). \tag{32}$$

Hence, for $(N,d)\to\infty$ the noise estimator Eq. (31) is independent of $\mathbf{Z}$ and self-averaging (see Fig. 2). Furthermore, for finite (N,d) and $\delta>0$ the probability of finding an extreme value of $\hat{\sigma}^2_{\text{ML}} \notin \mathcal{I}_{\sigma_0,\delta}$, where $\mathcal{I}_{\sigma_0,\delta} \equiv (\sigma_0^2(1-\zeta)-\delta, \sigma_0^2(1-\zeta)+\delta)$, is

$$\begin{aligned} \text{Prob}\big[\hat{\sigma}^2_{\text{ML}} \notin \mathcal{I}_{\sigma_0,\delta}\big] &= \text{Prob}\big[N\hat{\sigma}^2_{\text{ML}} \leqslant N\big(\sigma_0^2(1-\zeta)-\delta\big)\big] \\ &\quad + \text{Prob}\big[N\hat{\sigma}^2_{\text{ML}} \geqslant N\big(\sigma_0^2(1-\zeta)+\delta\big)\big] \\ &\leqslant \big\langle e^{-\frac{1}{2}\alpha\left|\left|\mathbf{t}-\mathbf{Z}\hat{\boldsymbol{\theta}}_{\text{ML}}[\mathscr{D}]\right|\right|^2}\big\rangle_{\mathscr{D}}\, e^{\frac{1}{2}\alpha N(\sigma_0^2(1-\zeta)-\delta)} \\ &\quad + \big\langle e^{\frac{1}{2}\alpha||\mathbf{t}-\mathbf{Z}\hat{\boldsymbol{\theta}}_{\text{ML}}[\mathscr{D}]||^2}\big\rangle_{\mathscr{D}}\, e^{-\frac{1}{2}\alpha N(\sigma_0^2(1-\zeta)+\delta)}. \end{aligned} \tag{33}$$

Assuming that the noise $\boldsymbol{\epsilon}$ is Gaussian, described by $\mathcal{N}(\mathbf{0}, \sigma_0^2\mathbf{I}_N)$, the moment-generating function (MGF)

$$\big\langle e^{\frac{1}{2}\alpha\left|\left|\mathbf{t}-\mathbf{Z}\hat{\boldsymbol{\theta}}_{\text{ML}}[\mathscr{D}]\right|\right|^2}\big\rangle_{\mathscr{D}} = e^{-\frac{N}{2}(1-\zeta)\log(1-\alpha\sigma_0^2)} \tag{34}$$

is independent of $\mathbf{Z}$, allowing us to estimate the fluctuations of $\hat{\sigma}^2_{\text{ML}}$ for $\delta \in [0, \sigma_0^2(1-\zeta)]$ via the inequality

$$\text{Prob}\big[\hat{\sigma}^2_{\text{ML}} \notin \mathcal{I}_{\sigma_0,\delta}\big] \leqslant \sum_{s=\pm 1} e^{-\frac{1}{2}N[(1-\zeta)\log(\frac{1-\zeta}{1-\zeta+s\delta/\sigma_0^2})+s\delta/\sigma_0^2]}. \tag{35}$$

For $\alpha = 2ia$ with $a \in \text{IR}$, the MGF Eq. (34) becomes the *characteristic function* (CF)

$$\big\langle e^{ia||\mathbf{t}-\mathbf{Z}\hat{\boldsymbol{\theta}}_{\text{ML}}[\mathscr{D}]||^2}\big\rangle_{\mathscr{D}} = \big(1 - ia\,2\sigma_0^2\big)^{-\frac{1}{2}N(1-\zeta)} \tag{36}$$

of the random variable $||\mathbf{t}-\mathbf{Z}\hat{\boldsymbol{\theta}}_{\text{ML}}[\mathscr{D}]||^2$. Note that Eq. (36) is the CF of the gamma distribution (see Theorem 7.6b in Ref. [1]), with mean $N\sigma_0^2(1-\zeta)$ and variance $N2\sigma_0^4(1-\zeta)$. Mean and variance of $\hat{\sigma}^2_{\text{ML}}$ are $\sigma_0^2(1-\zeta)$ and $2\sigma_0^4(1-\zeta)/N$, respectively. For $\sigma_0 = 1$ we obtain that $N\hat{\sigma}^2_{\text{ML}}$ is a chi-square distribution with $N(1-\zeta)$ degrees of freedom, as expected from Cochran's theorem [29].

Finally, it follows from Eqs. (32) and (20) that the finite temperature ML noise estimator in the high-dimensional regime is given by

$$\hat{\sigma}^2_{\text{ML}} = \frac{\beta}{\beta-\zeta}\sigma_0^2(1-\zeta). \tag{37}$$

We observe that for $\beta = 1$ we obtain unbiased estimation of σ^2. The latter suggests that "thermal" noise, controlled by β, is beneficial for the ML inference of σ^2. However, that this is indeed the case, and that the value $\beta = 1$ is "special," is not *a priori* obvious for this model. Our development confirms the result obtained in evaluating the average free energy with the replica method [9].

C. Statistical properties of MSE in ML inference

Using the distribution Eq. (29) and with the eigenvalues $\lambda_1(\boldsymbol{\Sigma}) \leqslant \lambda_2(\boldsymbol{\Sigma}) \leqslant \cdots \leqslant \lambda_d(\boldsymbol{\Sigma})$ of the true (population) covariance matrix $\boldsymbol{\Sigma}$, the CF of the MSE, defined as $\frac{1}{d}||\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_{\text{ML}}[\mathscr{D}]||^2$ for finite (N, d), can be written as

$$\begin{aligned} \langle e^{i\alpha||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}_{\text{ML}}[\mathscr{D}]||^2}\rangle_{\mathscr{D}} &= \int_0^\infty d\omega\, \Gamma_{N+1-d}(\omega) \\ &\quad \times \prod_{\ell=1}^{d}\left(1 - \frac{i\alpha 2\zeta\sigma_0^2}{\omega(1-\zeta+N^{-1})\lambda_\ell(\boldsymbol{\Sigma})}\right)^{-\frac{1}{2}}, \end{aligned} \tag{38}$$

with the gamma distribution $\Gamma_\nu(\omega) = \frac{\nu^{\nu/2}}{2^{\nu/2}\Gamma(\nu/2)}\omega^{\frac{\nu-2}{2}}e^{-\frac{1}{2}\nu\omega}$ for $\nu > 0$. The last term in Eq. (38) is the product of CFs of gamma distributions with the same "shape" parameter $1/2$, but different "scale" parameters $2\zeta\sigma_0^2/\omega(1-\zeta+N^{-1})\lambda_\ell(\boldsymbol{\Sigma})$. From Eq. (38) one obtains mean and variance of MSE:

$$\begin{aligned} \mu(\boldsymbol{\Sigma}) &= \frac{1}{d}\langle||\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_{\text{ML}}[\mathscr{D}]||^2\rangle_{\mathscr{D}} = \frac{\zeta\sigma_0^2}{1-\zeta-N^{-1}}\frac{\text{Tr}[\boldsymbol{\Sigma}^{-1}]}{d}, \\ \text{Var}\left(\frac{1}{d}||\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_{\text{ML}}[\mathscr{D}]||^2\right) &= \frac{2\zeta^2\sigma_0^4}{(1-\zeta)^2}\frac{\text{Tr}[\boldsymbol{\Sigma}^{-2}]}{d^2}. \end{aligned} \tag{39}$$

The latter gives us the condition for self-averaging of the MSE, i.e., $\text{Var}(\frac{1}{d}||\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_{\text{ML}}[\mathscr{D}]||^2) \to 0$ as $(N, d) \to \infty$. We note that Eq. (39) suggests that MSE is dominated by the smallest eigenvalue of the true covariance $\boldsymbol{\Sigma}$ and hence it can grow with an increase in the covariate correlations (see Fig. 2).

We finally consider deviations of $\frac{1}{d}||\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_{\text{ML}}[\mathscr{D}]||^2$ from the mean $\mu(\boldsymbol{\Sigma})$ given in Eq. (39). The probability of observing the event event $\frac{1}{d}||\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_{\text{ML}}[\mathscr{D}]||^2 \notin \mathcal{I}_{\mu,\delta}$, where $\mathcal{I}_{\mu,\delta} \equiv (\mu(\boldsymbol{\Sigma})-\delta, \mu(\boldsymbol{\Sigma})+\delta)$ for $\delta > 0$, is bounded from above as follows:

$$\begin{aligned} &\text{Prob}\left[\frac{1}{d}||\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_{\text{ML}}[\mathscr{D}]||^2 \notin \mathcal{I}_{\mu,\delta}\right] \\ &\quad = \text{Prob}[||\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_{\text{ML}}[\mathscr{D}]||^2 \leqslant d(\mu(\boldsymbol{\Sigma})-\delta)] \\ &\qquad + \text{Prob}[||\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_{\text{ML}}[\mathscr{D}]||^2 \geqslant d(\mu(\boldsymbol{\Sigma})+\delta)] \\ &\qquad \leqslant \text{C}_- e^{-N\Phi_-[\alpha,\mu(\lambda_d),\delta]} + \text{C}_+ e^{-N\Phi_+[\alpha,\mu(\lambda_1),\delta]}, \end{aligned} \tag{40}$$

with some small $\alpha > 0$ and positive constants $\text{C}_\pm$. For the rate function $\Phi_-[\alpha, \mu(\lambda_d), \delta]$ to be positive for arbitrary small δ it is sufficient that $\mu(\lambda_d) \geqslant 1$, where $\mu(\lambda) = \zeta\sigma_0^2/(1-\zeta)\lambda$, while for $\mu(\lambda_d) < 1$ for this to happen the δ values must satisfy $\delta > 1 - \mu(\lambda_d)$. The rate function $\Phi_+[\alpha, \mu(\lambda_1), \delta]$ is positive for any $\delta \in [0, \mu(\lambda_1)]$.

D. Statistical properties of the free energy

We consider the free-energy Eq. (12) for finite inverse temperature β and finite (N, d). Assuming that the noise $\boldsymbol{\epsilon}$ has mean $\mathbf{0}$ and covariance $\sigma_0^2\,\mathbf{I}_N$, and that the parameter σ^2 is independent of $\mathscr{D}$, the average free energy is

$$\begin{aligned} \langle F_{\beta,\sigma^2}[\mathscr{D}]\rangle_{\mathscr{D}} &= \frac{d}{2\beta} + \frac{1}{2\sigma^2}\boldsymbol{\theta}_0^{\text{T}}\big\langle\big(\mathbf{J} - \mathbf{J}\mathbf{J}^{-1}_{\sigma^2\eta}\mathbf{J}\big)\big\rangle_{\mathbf{Z}}\boldsymbol{\theta}_0 \\ &\quad + \frac{\sigma_0^2}{2\sigma^2}\big(N - \big\langle\text{Tr}\big[\mathbf{J}\mathbf{J}^{-1}_{\sigma^2\eta}\big]\big\rangle_{\mathbf{Z}}\big) \\ &\quad - \frac{1}{2\beta}\big\langle\log\big|2\pi e\,\sigma^2\beta^{-1}\mathbf{J}^{-1}_{\sigma^2\eta}\big|\big\rangle_{\mathbf{Z}}. \end{aligned} \tag{41}$$

Under the same assumptions, the *variance* of $F_{\beta,\sigma^2}[\mathscr{D}]$ can be obtained by exploiting the Helmholtz free-energy representation $F_{\beta,\sigma^2}[\mathscr{D}] = E_\beta[\mathscr{D}] - T\text{S}_\beta[\mathscr{D}]$, giving us

$$\begin{aligned} \text{Var}(F_{\beta,\sigma^2}[\mathscr{D}]) &= \text{Var}(E_\beta[\mathscr{D}]) + T^2\text{Var}(\text{S}_\beta[\mathscr{D}]) \\ &\quad - 2T\,\text{Cov}(E_\beta[\mathscr{D}], \text{S}_\beta[\mathscr{D}]). \end{aligned} \tag{42}$$

The full details on each term in Eq. (42) are found in Ref. [21].

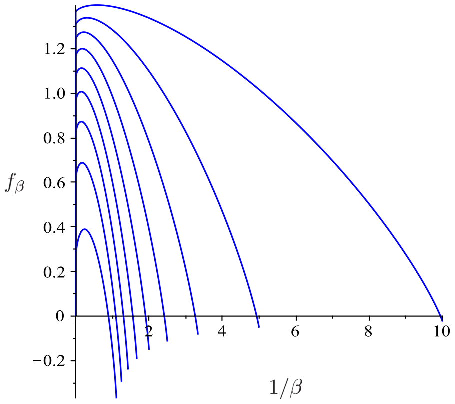

FIG. 3. Asymptotic free-energy density $f_\beta = \lim_{N\to\infty} f_\beta[\mathscr{D}]$ of finite temperature ML inference as a function of temperature $T = 1/\beta$, plotted for $\zeta \in \{1/10, 2/10, \ldots, 9/10\}$ (from right to left) in the high-dimensional regime where $N, d \to \infty$ with fixed ratio $\zeta = d/N$. For $\beta \to \infty$ it approaches the value $\frac{1}{2}\log[2\pi e\sigma_0^2(1-\zeta)]$. For $\beta \to \zeta$ it approaches $\frac{1}{2\zeta}[\zeta\log(1-\zeta) - \log(1-\zeta) - \zeta]$, and for $\beta \in (0, \zeta)$, i.e., in the high "temperature" region $T \in (1/\zeta, \infty)$, the free-energy density is $-\infty$. Here the true noise parameter is $\sigma_0^2 = 1$ and the true data covariance matrix is $\mathbf{I}_d$.

1. *Free energy of ML inference*

For $\eta = 0$ and after transforming $z_i(\mu) \to z_i(\mu)/\sqrt{d}$ for all (i, μ), with $z_i(\mu) = \mathcal{O}(1)$, Eq. (41) gives the average free-energy density

$$\left\langle \frac{F_{\beta,\sigma^2}[\mathscr{D}]}{N} \right\rangle_{\mathscr{D}} = \frac{1}{2}\frac{\sigma_0^2}{\sigma^2}(1-\zeta) + \frac{\zeta}{2\beta}\log\left(\frac{\beta}{2\pi\sigma^2\zeta}\right) + \frac{\zeta}{2\beta}\int d\lambda\, \rho_d(\lambda)\log(\lambda), \tag{43}$$

where we defined the average eigenvalue density $\rho_d(\lambda) = \langle \rho_d(\lambda|\mathbf{Z})\rangle_{\mathbf{Z}}$ of the empirical covariance matrix, with

$$\rho_d(\lambda|\mathbf{Z}) = \frac{1}{d}\sum_{\ell=1}^{d}\delta[\lambda - \lambda_\ell(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}/N)]. \tag{44}$$

The variance of free-energy density is

$$\mathrm{Var}\left(\frac{F_{\beta,\sigma^2}[\mathscr{D}]}{N}\right) = \mathrm{Var}\left(\frac{E[\mathscr{D}]}{N}\right) + T^2\mathrm{Var}\left(\frac{\mathrm{S}(P[\mathscr{D}])}{N}\right) = \frac{\zeta^2}{4\beta^2}\int d\lambda\, d\tilde{\lambda}\, C_d(\lambda, \tilde{\lambda})\log(\lambda)\log(\tilde{\lambda}) + \frac{\sigma_0^4(1-\zeta)}{2\sigma^4 N}, \tag{45}$$

where we defined the correlation function $C_d(\lambda, \tilde{\lambda}) = \langle \rho_d(\lambda|\mathbf{Z})\rho_d(\tilde{\lambda}|\mathbf{Z})\rangle_{\mathbf{Z}} - \langle \rho_d(\lambda|\mathbf{Z})\rangle_{\mathbf{Z}}\langle \rho_d(\tilde{\lambda}|\mathbf{Z})\rangle_{\mathbf{Z}}$. Clearly, if $\int d\lambda\, d\tilde{\lambda}\, C_d(\lambda, \tilde{\lambda}) f(\lambda, \tilde{\lambda}) \to 0$ as $(N, d) \to \infty$, for any smooth function $f(\lambda, \tilde{\lambda})$, then the free-energy density $f_\beta[\mathscr{D}] = F_\beta[\mathscr{D}]/N$ is self-averaging.

Finally, if we use Eq. (43) in the free-energy density Eq. (17) for $\eta = 0$, and assume Gaussian data with true population covariance matrix $\mathbf{\Sigma} = \mathbf{I}_d$, then we find

$$\lim_{N\to\infty} f_\beta[\mathscr{D}] = \begin{cases} \frac{\beta-\zeta}{2\beta}\log\left(\frac{2\pi\sigma_0^2(1-\zeta)}{\beta-\zeta}\right) + \frac{\log\beta+1}{2} - \frac{1}{2\beta}(\zeta\log\zeta + (1-\zeta)\log(1-\zeta) + 2\zeta) & \text{if } \beta > \zeta \\ -\infty & \text{if } \beta \in (0, \zeta) \end{cases}, \tag{46}$$

with the convention $0\log 0 = 0$. Since for $\lambda \in [a_-, a_+]$ and $0 < \zeta < 1$ the eigenvalue spectrum $\rho_d(\lambda|\mathbf{Z})$ converges to $(2\pi\lambda\zeta)^{-1}\sqrt{(\lambda - a_-)(a_+ - \lambda)}$ in a distributional sense as $(N, d) \to \infty$ [30], with $a_\pm = (1 \pm \sqrt{\zeta})^2$, the free-energy density is self-averaging. Its values are plotted versus the temperature in Fig. 3. Furthermore, the average free-energy density Eq. (46) is identical to that of Ref. [9]. Since $\lim_{\beta\downarrow\zeta}\lim_{N\to\infty} f_\beta[\mathscr{D}]$ is finite, the system exhibits a zeroth-order phase transition [31] at $T = 1/\zeta$.

2. *Free energy of MAP inference*

We next assume that the true parameters $\boldsymbol{\theta}_0$ are drawn at random, with mean $\mathbf{0}$ and covariance matrix $S^2\mathbf{I}_d$. As before we rescale $z_i(\mu) \to z_i(\mu)/\sqrt{d}$ where $z_i(\mu) = \mathcal{O}(1)$, and define $\mathbf{J} = \mathbf{C}/\zeta$ (so that $\mathbf{C} = \mathbf{Z}^{\mathrm{T}}\mathbf{Z}/N$) and $\mathbf{C}_{\zeta\sigma^2\eta} = \zeta\mathbf{J}_{\sigma^2\eta}$. Then the average of Eq. (41) over $\boldsymbol{\theta}_0$ becomes

$$\left\langle\left\langle \frac{F_{\beta,\sigma^2}[\mathscr{D}]}{N} \right\rangle_{\mathscr{D}}\right\rangle_{\boldsymbol{\theta}_0} = \frac{\zeta}{2\beta} + \frac{1}{2}\int d\lambda\, \rho_d(\lambda)\left[\frac{S^2\zeta\eta\lambda}{\lambda + \zeta\sigma^2\eta} + \frac{\sigma_0^2}{\sigma^2}\left[1 - \frac{\zeta\lambda}{\lambda + \zeta\sigma^2\eta}\right] + \frac{\zeta}{\beta}\log(\lambda + \zeta\sigma^2\eta)\right] - \frac{\zeta}{2\beta}\log(2\pi e\sigma^2\beta^{-1}\zeta). \tag{47}$$

Furthermore, using Eq. (42), we obtain, under the same assumptions, that $\mathrm{Var}(F_{\beta,\sigma^2}[\mathscr{D}]/N)$ is of the form [21]

$$\mathrm{Var}\left(\frac{F_{\beta,\sigma^2}[\mathscr{D}]}{N}\right) = \int d\lambda\, d\tilde{\lambda}\, C_d(\lambda,\tilde{\lambda})\Phi(\lambda,\tilde{\lambda}) + \mathcal{O}\left(\frac{1}{N}\right). \quad (48)$$

Hence, for $\eta > 0$ the conditional free energy is self-averaging with respect to the realization of the true parameter if the spectrum $\rho_d(\lambda|\mathbf{Z})$ is self-averaging (since then $C_d(\lambda,\tilde{\lambda}) \to 0$ as $(N,d) \to \infty$). The latter, under the same assumptions, is the condition for the MAP estimator Eq. (20) of the noise σ^2 to be self-averaging (see Ref. [21]) and hence the free-energy Eq. (17) is self-averaging if $\rho_d(\lambda|\mathbf{Z})$ is self-averaging. This is the case, e.g., for Gaussian data with covariance matrix $\mathbf{\Sigma} = \mathbf{I}_d$.

IV. DISCUSSION AND OUTLOOK

In this paper we derived exact results for the Bayesian model Eq. (2) of the linear regression $\mathbf{t} = \mathbf{Z}\boldsymbol{\theta} + \sigma\boldsymbol{\epsilon}$, where $\mathbf{t} \in \mathrm{IR}^N$ and $\mathbf{Z} \in \mathrm{IR}^{N\times d}$. Mapping this to a Gibbs-Boltzmann distribution Eq. (11), with finite (inverse) "temperature" β, allowed us to investigate properties of several inference protocols [1] for finite N (sample size), d (dimension) and in the (high-dimensional) limit $(N,d) \to \infty$. In particular, we studied statistical properties of free energy which is the main object of interest in statistical physics approaches to inference (see Ref. [9] and references therein).

If the noise strength σ^2 is known and the distributions of the data $\mathbf{Z}$ and the noise $\boldsymbol{\epsilon}$ are Gaussian, then the distribution of the MAP estimator $\hat{\boldsymbol{\theta}}_{\mathrm{MAP}}$ of $\boldsymbol{\theta}$ is the Gaussian mixture Eq. (28), for any finite (N,d). We used Eq. (28) to show that the distribution of ML estimator $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ is the Student's t-distribution Eq. (29). The consequence of this is that its marginal, the univariate Student distribution (which can be used in statistical hypothesis testing to calculate p-values), has "fat" tails for finite (N,d). However, any marginal of Eq. (29) that describes a finite number of components of $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ has a Gaussian form when $(N,d) \to \infty$.

Also, for any choice for the distributions of $\mathbf{Z}$ and $\boldsymbol{\epsilon}$, the ML estimator $\hat{\sigma}^2_{\mathrm{ML}}$ of the noise parameter σ^2 is *self-averaging*, i.e., its variance is vanishing as $(N,d) \to \infty$. Furthermore, deviations of $\hat{\sigma}^2_{\mathrm{ML}}$ from its mean, estimated by the bound in Eq. (35), are exponentially suppressed in (N,d) for Gaussian $\boldsymbol{\epsilon}$. As a consequence the inference of $\hat{\sigma}^2_{\mathrm{ML}}$ is almost deterministic even for moderate values of (N,d). This result is independent of $\mathbf{Z}$.

We used the distribution of the ML estimator Eq. (29) to derive the characteristic function of the MSE (38). The latter was used to derive the mean and variance Eq. (39), giving a condition for the MSE to be self-averaging as $(N,d) \to \infty$, and to estimate deviations of MSE from its mean, given by the bound Eq. (40), for finite (N,d). The result Eq. (39) suggests that the deviations of $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ from $\boldsymbol{\theta}_0$ can grow significantly with covariate correlations, proportional to $\mathrm{Tr}[\mathbf{\Sigma}^{-1}]$, thus leading to severe inefficacy in the inference of $\boldsymbol{\theta}$.

If we assume that the noise parameter σ^2 is known, then we obtain average Eq. (43) and variance Eq. (45) of the conditional free-energy density Eq. (12) of ML inference with *finite* temperature $T = 1/\beta$ and for finite (N,d). This result is independent of the distributions of the data $\mathbf{Z}$ and the noise $\boldsymbol{\epsilon}$. For finite T, the noise estimator $\hat{\sigma}^2_{\mathrm{ML}}$, given by Eq. (20) with $\eta = 0$, is self-averaging when $(N,d) \to \infty$. The same is true for the free-energy density Eq. (17) if the density of eigenvalue *spectrum* Eq. (44) of the covariance matrix $\mathbf{Z}^{\mathrm{T}}\mathbf{Z}/N$ is self-averaging. The latter is true if $\mathbf{Z}$ is sampled from a Gaussian with $\mathbf{\Sigma} = \mathbf{I}_d$. In that case, and upon assuming Gaussian noise $\boldsymbol{\epsilon}$, the free-energy density Eq. (17) recovers the result obtained by the replica method [9]. The ML estimator $\hat{\sigma}^2_{\mathrm{ML}}$ diverges at $\beta = \zeta$, and the free energy density Eq. (17) is *discontinuous* at this value of β. This is an instance where the presence of the thermal noise with finite generic β allows us to derive an interesting result. Another is Eq. (37) for the finite temperature ML noise estimator.

The additional assumption that the true parameters $\boldsymbol{\theta}_0$ are drawn at random, with mean $\mathbf{0}$ and covariance $S^2\mathbf{I}_d$, allows us to derive average Eq. (47) and variance Eq. (48) of the conditional free-energy Eq. (12) of MAP inference for finite T and finite (N,d). We also computed the variance of the MAP estimator Eq. (20) of the noise parameter σ^2. These results are again independent of the distributions of the data $\mathbf{Z}$ and the noise $\boldsymbol{\epsilon}$. We find that the free energy Eq. (17) is self-averaging if the spectrum of the empirical covariance matrix $\mathbf{Z}^{\mathrm{T}}\mathbf{Z}/N$ is self-averaging as $(N,d) \to \infty$.

The above results emphasize that still much can be learned about high-dimensional Bayesian linear regression from exact calculations with standard methods. While we present this as a minimal model of inference in the high-dimensional setting, linear regression is commonly used in many areas of research. For example, linear regression models are used extensively in the statistical analysis of genetic data. Genome-wide association studies, where the aim is to undercover effect sizes for each single nucleotide polymorphisms (SNPs), often use extremely high-dimensional datasets. Here the number N of individuals is $O(10^3)$ and the number d of SNPs is $O(10^6)$ with correlations occurring due to the phenomenon of genetic linkage. Another biological example is the analysis of gene expression data where due to nature of biological pathways involved [32], correlations pose a significant challenge in uncovering true associations in data [33].

Many questions remain still open and we hope that this paper may contribute to future work in this direction. Some results appear well within reach, such as the extension to sub-Gaussian noise for the argument that leads to Eq. (35), employing techniques used in Ref. [34]. Other results are less immediate but seem feasible, such as extending some of the ML results to MAP inference, starting from evaluation of the distribution of $\hat{\boldsymbol{\theta}}_{\mathrm{MAP}}$ Eq. (28) for $(N,d) \to \infty$. Another interesting line of work would be to try to extend our present results to generalized linear models (GLMs), a very similar distribution for the estimator $\hat{\boldsymbol{\theta}}_{\mathrm{MAP}}$ has already been conjectured through the use of the replica method, [9]. Other crucial investigations, such as a rigorous analytical study of the effect of model mismatch, appear instead to be still quite challenging with current techniques. Overall, we expect high-dimensional linear regression to serve as a starting point to tackle more realistic scenarios, which should include, among other things, correlation between data and noise and dimensional mismatch between the teacher and the student model.

ACKNOWLEDGMENTS

We thank Dr. Heather Battey and Sir David Cox for reading a first draft of this manuscript and valuable suggestions. A.M. is supported by Cancer Research UK (Award No. C45074/A26553) and the UK's Medical Research Council (Award No. MR/R014043/1). F.A. is supported by EIT Health ID 19277 (RGS@HOME) under H2020. FAL gratefully acknowledges financial support through a scholarship from Conacyt (Mexico).

[1] A. C. Rencher and G. B. Schaalje, *Linear Models in Statistics* (John Wiley & Sons, New York, NY, 2008).

[2] P. J. Huber and E. M. Ronchetti, *Robust Statistics* (John Wiley & Sons, New York, NY, 2009).

[3] M. Bayati and A. Montanari, IEEE Trans. Info. Theory **58**, 1997 (2011).

[4] N. El Karoui, D. Bean, P. J. Bickel, C. Lim, and B. Yu, Proc. Natl. Acad. Sci. U.S.A. **110**, 14557 (2013).

[5] N. E. Karoui and E. Purdom, arXiv:1608.00696 (2016).

[6] M. J. Wainwright, *High-dimensional Statistics: A Non-asymptotic Viewpoint* (Cambridge University Press, Cambridge, UK, 2019).

[7] M. Advani and S. Ganguli, Phys. Rev. X **6**, 031034 (2016).

[8] J. Barbier, F. Krzakala, N. Macris, L. Miolane, and L. Zdeborová, Proc. Natl. Acad. Sci. U.S.A. **116**, 5451 (2019).

[9] A. C. C. Coolen, M. Sheikh, A. Mozeika, F. A. Lopez, and F. Antenucci, J. Phys. A: Math. Theor. **53**, 365001 (2020).

[10] E. T. Jaynes, Phys. Rev. **106**, 620 (1957).

[11] H. Nishimori, *Statistical Physics of Spin Glasses and Information Processing: An Introduction* (Clarendon Press, Oxford, UK, 2001).

[12] C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer, Berlin, 2006).

[13] M. Mézard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications* (World Scientific Publishing, Singapore, 1987).

[14] M. Mézard and A. Montanari, *Information, Physics, and Computation* (Oxford University Press, Oxford, UK, 2009).

[15] D. Donoho and A. Montanari, Probab. Theory Relat. Fields **166**, 935 (2016).

[16] C. Gerbelot, A. Abbara, and F. Krzakala, arXiv:2002.04372 (2020).

[17] J. Barbier, N. Macris, M. Dia, and F. Krzakala, IEEE Trans. Inf. Theory **66**, 4270 (2020).

[18] A. Engel and C. Van den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, UK, 2001).

[19] L. Zdeborová and F. Krzakala, Adv. Phys. **65**, 453 (2016).

[20] R. Vershynin, *High-dimensional Probability: An Introduction with Applications in Data Science*, Vol. 47 (Cambridge University Press, Cambridge, UK, 2018).

[21] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.103.042142 for details of derivations and proofs, which includes Refs. [9,35,36].

[22] N. G. De Bruijn, *Asymptotic Methods in Analysis* (Dover, New York, NY, 1981).

[23] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, Science **220**, 671 (1983).

[24] Y. Iba, J. Phys. A: Math. Gen. **32**, 3875 (1999).

[25] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (John Wiley & Sons, New York, NY, 2012).

[26] A. C. C. Coolen, J. E. Barrett, P. Paga, and C. J. Perez-Vicente, J. Phys. A: Math. Theor. **50**, 375001 (2017).

[27] S. Nadarajah and S. Kotz, Acta Appl. Math. **89**, 53 (2005).

[28] B. G. Kibria and A. H. Joarder, J. Stat. Res. **40**, 59 (2006).

[29] W. G. Cochran, Math. Proc. Cambridge Philos. Soc. **30**, 178 (1934).

[30] F. Götze and A. Tikhomirov, Bernoulli **10**, 503 (2004).

[31] P.-H. Chavanis, Phys. Rev. E **65**, 056123 (2002).

[32] M. R. Segal, K. D. Dahlquist, and B. R. Conklin, J. Comput. Biol. **10**, 961 (2003).

[33] D. J. Balding, Nat. Rev. Genet. **7**, 781 (2006).

[34] M. Rudelson and R. Vershynin, Electron. Commun. Probab. **18**, 1 (2013).

[35] M. L. Eaton, *Multivariate Statistics: A Vector Space Approach* (John Wiley & Sons, New York, NY, 1983).

[36] M. Taboga, *Lectures on Probability Theory and Mathematical Statistics* (Independent Publishing Platform, California, 2017).