Exact results on high-dimensional linear regression via statistical physics

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

[1][1]London Institute for Mathematical Sciences, 35A South Street, London, W1K 2XF, United Kingdom.
[2]Dept of Mathematics, King's College London, The Strand, London WC2R 2LS, United Kingdom.
[3]Dept of Biophysics, Radboud University, 6525AJ Nijmegen, The Netherlands.
[4]Saddle Point Science Ltd., 35A South Street, London, W1K 2XF, United Kingdom.

(Dated: December 1, 2020)

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 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 a number of new exact results for linear regression in the high-dimensional regime.

PACS numbers: 02.50.Tt, 05.10.-a, 64.60.De

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 non-rigorous 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 [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.

In this paper we derive new 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 $\mathbf{t}\in\mathbb{R}^N$ and $\mathbf{Z}\in\mathbb{R}^{N\times d}$ are observed and the parameters $\boldsymbol{\theta}\in\mathbb{R}^d$ and $\sigma\in\mathbb{R}^+$ 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 prior $\mathcal{N}(\mathbf{0}, \eta^{-1}\mathbf{I}_d)$ 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 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 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 in order 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) The ML estimator $\hat{\sigma}^2_{\rm ML}$ of the noise parameter σ^2 is *self-averaging* as $(N,d)\to\infty$ (i.e. its variance is vanishing in this limit), for any distributions of $\mathbf{Z}$ and $\boldsymbol{\epsilon}$. We bound the likelihood of deviations of $\hat{\sigma}^2_{\rm ML}$ from its mean for Gaussian noise $\boldsymbol{\epsilon}$; (ii) If σ^2 is known and the distributions of $\mathbf{Z}$ and $\boldsymbol{\epsilon}$ are Gaussian, we compute the distribution of the MAP and ML estimators of $\boldsymbol{\theta}$; (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 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 *spectrum* of the 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 [9]; (v) If the true parameters $\boldsymbol{\theta}_0$ are independent *random* variables, 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 Appendix.

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 $(\mathbf{z}_i,t_i)\in\mathrm{I\!R}^{d+1}$, 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

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

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, expression (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 (2) for inference is to compute the *maximum a posteriori* (MAP) estimator

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

in which the so-called Bayesian likelihood function

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

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

$$\hat{\boldsymbol{\theta}}[\mathscr{D}]=\int\mathrm{d}\boldsymbol{\theta}\;P(\boldsymbol{\theta}|\mathscr{D})\boldsymbol{\theta} \tag{5}$$

(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

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

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

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

The estimators (3) and (5) are recovered from the average $\int\mathrm{d}\boldsymbol{\theta}\;P_\beta(\boldsymbol{\theta}|\mathscr{D})\boldsymbol{\theta}$ 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 (6) and the posterior (2) are identical, and that $\hat{\boldsymbol{\theta}}_{\mathrm{MAP}}[\mathscr{D}]=\lim_{\beta\to\infty}\int\mathrm{d}\boldsymbol{\theta}\;P_\beta(\boldsymbol{\theta}|\mathscr{D})\boldsymbol{\theta}$ by the Laplace argument [20]. We note that the interpretation of the MAP estimator (3) in the SP framework (6) is that $\hat{\boldsymbol{\theta}}_{\rm MAP}[\mathscr{D}]$ is the 'ground state' of the system.

The Kullback-Leibler (KL) 'distance' [21] between the distribution $P(t,\mathbf{z}|\boldsymbol{\theta})$ 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

$$D(\hat{P}[\mathscr{D}]||P_{\boldsymbol{\theta}}) = \int \mathrm{d}t\,\mathrm{d}\mathbf{z}\,\hat{P}(t,\mathbf{z}|\mathscr{D})\log\Big(\frac{\hat{P}(t,\mathbf{z}|\mathscr{D})}{P(t,\mathbf{z}|\boldsymbol{\theta})}\Big) \tag{8}$$

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

Finally, the KL distance (8) can also be used to define the difference $\Delta D(\boldsymbol{\theta},\boldsymbol{\theta}_0|\mathscr{D}) = D(\hat{P}[\mathscr{D}]||P_{\boldsymbol{\theta}}) - D(\hat{P}[\mathscr{D}]||P_{\boldsymbol{\theta}_0})$, where $\boldsymbol{\theta}_0$ are the true parameters responsible for the data, which served as a useful measure of over-fitting in ML inference [22], and was recently extended to MAP inference in generalized linear models [9]. Both latter studies used the SP framework, equivalent to (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].

Bayesian Linear Regression–In Bayesian linear regression (LR) with Gaussian priors, 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 (4) is given by

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

where $\eta \geq 0$ is the hyper-parameter for the prior $P(\boldsymbol{\theta})$. 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 also be written as

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

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 (6) is now

$$\begin{aligned} &P_\beta(\boldsymbol{\theta},\sigma^2|\mathscr{D}) \\ &\quad= \frac{P_\beta\big(\boldsymbol{\theta}|\,\sigma^2,\mathscr{D}\big)\mathrm{e}^{-\beta\left[F_{\beta,\sigma^2}[\mathscr{D}]+\frac{1}{2}N\log(2\pi\sigma^2)-\log P(\sigma^2)\right]}}{\int_0^\infty \mathrm{d}\tilde{\sigma}^2\,\mathrm{e}^{-\beta\left[F_{\beta,\tilde{\sigma}^2}[\mathscr{D}]+\frac{1}{2}N\log\tilde{\sigma}^2-\log P(\tilde{\sigma}^2)\right]}}, \end{aligned} \tag{11}$$

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

$$P_\beta(\boldsymbol{\theta}|\,\sigma^2,\mathscr{D}) = \mathcal{N}\big(\boldsymbol{\theta}|\,\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{t},\,\sigma^2\beta^{-1}\mathbf{J}_{\sigma^2\eta}^{-1}\big) \tag{12}$$

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

$$\begin{aligned} F_{\beta,\sigma^2}[\mathscr{D}] = \frac{d}{2\beta} + \frac{1}{2\sigma^2}\mathbf{t}^{\mathrm{T}}\big(\mathbf{I}_N - \mathbf{Z}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{Z}^T\big)\mathbf{t} \\ -\frac{1}{2\beta}\log|2\pi\mathrm{e}\sigma^2\beta^{-1}\mathbf{J}_{\sigma^2\eta}^{-1}|, \end{aligned} \tag{13}$$

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

$$\begin{aligned} F_\beta[\mathscr{D}] &= -\frac{1}{\beta}\log\int \mathrm{d}\boldsymbol{\theta}\,\mathrm{d}\sigma^2\,\mathrm{e}^{-\beta E\left(\boldsymbol{\theta},\sigma^2|\mathscr{D}\right)} \\ &= -\frac{1}{\beta}\log\int_0^\infty \mathrm{d}\sigma^2\mathrm{e}^{-\beta\left[F_{\beta,\sigma^2}[\mathscr{D}]+\frac{N}{2}\log(2\pi\sigma^2)-\log P(\sigma^2)\right]}. \end{aligned} \tag{14}$$

Note that if the noise parameter σ^2 is known, i.e. $P(\sigma^2)=\delta(\sigma^2-\sigma_0^2)$, then $F_\beta[\mathscr{D}]=F_{\beta,\sigma_0^2}[\mathscr{D}]-\frac{N}{2\beta}\log(2\pi\sigma_0^2)$ and $P_\beta(\boldsymbol{\theta},\sigma^2|\mathscr{D})=P_\beta(\boldsymbol{\theta}|\sigma_0^2,\mathscr{D})\delta(\sigma^2-\sigma_0^2)$. For $\beta\to\infty$ the free energy is via the Laplace argument given by $F_\infty[\mathscr{D}]=\min_{\boldsymbol{\theta},\sigma^2} E\left(\boldsymbol{\theta},\sigma^2|\mathscr{D}\right)$. $F_\infty[\mathscr{D}]$ is the ground state energy of (11). The ground state $\left\{\hat{\boldsymbol{\theta}}[\mathscr{D}],\hat{\sigma}^2[\mathscr{D}]\right\}=\operatorname{argmin}_{\boldsymbol{\theta},\sigma^2} E\left(\boldsymbol{\theta},\sigma^2|\mathscr{D}\right)$ is given by

$$\hat{\boldsymbol{\theta}}\,[\mathscr{D}]=\mathbf{J}^{-1}_{\sigma^2\eta}\mathbf{Z}^{\mathrm{T}}\mathbf{t}, \tag{15}$$

i.e. the mean of (12), and the solution of the equation

$$\sigma^2=\frac{1}{N}\left|\left|\mathbf{t}-\mathbf{Z}\hat{\boldsymbol{\theta}}\,[\mathscr{D}]\,\right|\right|^2+\frac{2\sigma^4}{N}\frac{\partial}{\partial\sigma^2}\log P(\sigma^2), \tag{16}$$

corresponding to the MAP estimators of the parameters [23]. From the second second line in (14) we infer

$$F_\infty[\mathscr{D}]=\min_{\sigma^2}\Big[F_{\infty,\sigma^2}[\mathscr{D}]+\frac{N\log(2\pi\sigma^2)}{2}-\log P(\sigma^2)\Big], \tag{17}$$

(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:

$$f_\beta[\mathscr{D}]=\min_{\sigma^2}\Big[\frac{F_{\beta,\sigma^2}\,[\mathscr{D}]}{N}+\frac{\log(2\pi\sigma^2)}{2}-\frac{\log P(\sigma^2)}{N}\Big]. \tag{18}$$

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

$$\begin{aligned}\int_0^\infty \mathrm{d}\boldsymbol{\theta}\,\mathrm{d}\sigma^2\;P_\beta(\boldsymbol{\theta},\sigma^2|\mathscr{D})\,\boldsymbol{\theta}&=\langle\mathbf{J}^{-1}_{\sigma^2\eta}\mathbf{Z}^{\mathrm{T}}\mathbf{t}\rangle_{\sigma^2}\\ \int_0^\infty \mathrm{d}\boldsymbol{\theta}\,\mathrm{d}\sigma^2\;P_\beta(\boldsymbol{\theta},\sigma^2|\mathscr{D})\,\sigma^2&=\langle\sigma^2\rangle_{\sigma^2},\end{aligned} \tag{19}$$

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

$$P_\beta(\sigma^2|\mathscr{D})=\frac{\mathrm{e}^{-\beta\left[F_{\beta,\sigma^2}[\mathscr{D}]+\frac{N}{2}\log(2\pi\sigma^2)-\log P(\sigma^2)\right]}}{\int_0^\infty\mathrm{d}\tilde{\sigma}^2\,\mathrm{e}^{-\beta\left[F_{\beta,\tilde{\sigma}^2}[\mathscr{D}]+\frac{N}{2}\log(2\pi\tilde{\sigma}^2)-\log P(\tilde{\sigma}^2)\right]}}. \tag{20}$$

For $(N,d)\to\infty$ this marginal is dominated by the solution of (18). The dominant value of $\boldsymbol{\theta}$ in (19) is (15), 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 (15) and (16) when $\beta\to\infty$:

$$\begin{aligned}\sigma^2=\frac{\beta}{(\beta-\zeta)}\frac{1}{N}\left|\left|\mathbf{t}-\mathbf{Z}\hat{\boldsymbol{\theta}}\,[\mathscr{D}]\,\right|\right|^2&-\frac{\sigma^4\eta}{(\beta-\zeta)}\frac{1}{N}\mathrm{Tr}\big[\mathbf{J}^{-1}_{\sigma^2\eta}\big]\\ &+\frac{2\sigma^4\beta}{(\beta-\zeta)N}\frac{\partial}{\partial\sigma^2}\log P(\sigma^2).\end{aligned} \tag{21}$$

The free energies (13) and (14) obey the Helmholtz free energy relations. In particular, with $E\,(\boldsymbol{\theta}|\mathscr{D})=E\left(\boldsymbol{\theta},\sigma^2|\mathscr{D}\right)-\frac{1}{2}N\log(2\pi\sigma^2)+\log P(\sigma^2)$ we get

$$F_{\beta,\sigma^2}\,[\mathscr{D}]=E_\beta[\mathscr{D}]-T\,\mathrm{S}_\beta[\mathscr{D}], \tag{22}$$

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

$$E_\beta[\mathscr{D}]=\int\mathrm{d}\boldsymbol{\theta}\;P_\beta(\boldsymbol{\theta}|\sigma^2,\mathscr{D})E(\boldsymbol{\theta}|\mathscr{D}), \tag{23}$$

and with the *differential entropy*

$$\mathrm{S}_\beta[\mathscr{D}]=-\int\mathrm{d}\boldsymbol{\theta}\;P_\beta(\boldsymbol{\theta}|\sigma^2,\mathscr{D})\log P_\beta(\boldsymbol{\theta}|\sigma^2,\mathscr{D}). \tag{24}$$

In the free energy (13) we have

$$\begin{aligned}E_\beta[\mathscr{D}]&=\frac{d}{2\beta}+\frac{1}{2\sigma^2}\mathbf{t}^{\mathrm{T}}\left(\mathbf{I}_N-\mathbf{Z}\mathbf{J}^{-1}_{\sigma^2\eta}\mathbf{Z}^{T}\right)\mathbf{t}\\ \mathrm{S}_\beta[\mathscr{D}]&=\frac{1}{2}\log|2\pi\mathrm{e}\sigma^2\beta^{-1}\mathbf{J}^{-1}_{\sigma^2\eta}|.\end{aligned} \tag{25}$$

Furthermore, the average energy can be written as $E_\beta[\mathscr{D}] = \frac{d}{2\beta} + \min_{\boldsymbol{\theta}} E(\boldsymbol{\theta}|\mathscr{D})$.

Distribution of estimators $\hat{\boldsymbol{\theta}}_{\rm MAP}$ and $\hat{\boldsymbol{\theta}}_{\rm ML}$ –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 $\mathcal{N}(\mathbf{0},\sigma_0^2\mathbf{I}_N)$, the distribution of the MAP estimator (15) is

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

For $\eta=0$, i.e. ML inference, and without averaging over $\mathbf{Z}$, this recovers Theorem 7.6b in [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 $[\mathbf{C}]_{k\ell}=N^{-1}\sum_{i=1}^N z_i(k)z_i(\ell)$ and $\mathbf{C}_{\zeta\sigma^2\eta}=\mathbf{C}+\zeta\sigma^2\eta\mathbf{I}$, so

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

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 $\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})$, the distribution of $\hat{\boldsymbol{\theta}}$ for any finite (N,d) is the Gaussian mixture

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

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 non-singular when $d\le N$. Note that (28) also represents the distribution of ‘ground states’ of (12).

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

$$\begin{aligned} P(\hat{\boldsymbol{\theta}}) = &\frac{\Gamma\left(\frac{N+1}{2}\right)}{\Gamma\left(\frac{N+1-d}{2}\right)}\left|\frac{(1-\zeta+1/N)\boldsymbol{\Sigma}}{\pi\,(N+1-d)\,\zeta\sigma_0^2}\right|^{\frac{1}{2}} \\ &\times\left[1+\frac{(1-\zeta+1/N)\boldsymbol{\Sigma}}{(N+1-d)\zeta\sigma_0^2}\Big(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\Big)\Big(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\Big)^{\rm T}\right]^{-\frac{N+1}{2}} \end{aligned} \tag{29}$$

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 (29). In the regime $(N,d)\to\infty$ one can recover from (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 $\hat{\boldsymbol{\theta}}_{\rm ML}$ is described by a Gaussian distribution [24, 25].

Statistical properties of the estimator $\hat{\sigma}^2_{\rm ML}$ –For $\eta=0$ the estimator (15) simplifies considerably to

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

giving us, via (16), the ML noise estimator

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

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

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

Hence for $(N,d)\to\infty$ the noise estimator (31) is independent of $\mathbf{Z}$ and self-averaging [26]. Furthermore, for finite (N,d) and $\delta>0$ the probability of finding an extreme value of $\hat{\sigma}^2_{\rm ML}$ is

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

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)

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

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

$$\begin{aligned}&\mathrm{Prob}\left[\hat{\sigma}^2_{\rm ML}\notin\big(\sigma_0^2(1-\zeta)-\delta,\sigma_0^2(1-\zeta)+\delta\big)\right]\\&\leq \mathrm{e}^{-\frac{1}{2}N\left[(1-\zeta)\log\left(\frac{1-\zeta}{(1-\zeta)-\delta/\sigma_0^2}\right)-\delta/\sigma_0^2\right]}\\&\quad+\mathrm{e}^{-\frac{1}{2}N\left[(1-\zeta)\log\left(\frac{1-\zeta}{(1-\zeta)+\delta/\sigma_0^2}\right)+\delta/\sigma_0^2\right]}\end{aligned}\tag{35}$$

For $\alpha=2\mathrm{i}a$ with $a\in\mathbb{R}$ the MGF (34) becomes the *characteristic function* (CF)

$$\Big\langle \mathrm{e}^{\mathrm{i}a\left|\left|\mathbf{t}-\mathbf{Z}\hat{\boldsymbol{\theta}}_{\rm ML}[\mathscr{D}]\right|\right|^2}\Big\rangle_{\mathscr{D}}=\big(1-\mathrm{i}a\,2\sigma_0^2\big)^{-\frac{1}{2}N(1-\zeta)}\tag{36}$$

of the random variable $\left|\left|\mathbf{t}-\mathbf{Z}\hat{\boldsymbol{\theta}}_{\rm ML}[\mathscr{D}]\right|\right|^2$. Note that (36) is the CF of the gamma distribution (see Theorem 7.6b in [1]), with mean $N\sigma_0^2(1-\zeta)$ and variance $N2\sigma_0^4(1-\zeta)$. Mean and variance of $\hat{\sigma}^2_{\rm ML}$ are $\sigma_0^2(1-\zeta)$ and $2\sigma_0^4\,(1-\zeta)/N$, respectively. For $\sigma_0=1$ we obtain that $N\hat{\sigma}^2_{\rm ML}$ is a chi-square distribution with $N(1-\zeta)$ degrees of freedom, as expected from Cochran's theorem [27].

Finally, it follows from (32) and (21) that the finite temperature ML noise estimator in the high-dimensional regime is given by $\hat{\sigma}^2_{\rm ML}=\frac{\beta}{\beta-\zeta}\sigma_0^2(1-\zeta)$. We observe that for $\beta=1$ we obtain unbiased estimation of σ^2. Note that these results also follow from the average free energy computed by the replica method [9].

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

$$\begin{aligned}\Big\langle \mathrm{e}^{\mathrm{i}\alpha||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}_{\rm ML}[\mathscr{D}]||^2}\Big\rangle_{\mathscr{D}}&=\int_0^\infty\mathrm{d}\omega\;\Gamma_{N+1-d}\,(\omega)\\&\quad\times\prod_{\ell=1}^d\left(1-\frac{\mathrm{i}\alpha2\zeta\sigma_0^2}{\omega(1-\zeta+N^{-1})\lambda_\ell(\mathbf{\Sigma})}\right)^{-\frac{1}{2}},\end{aligned}\tag{37}$$

with the gamma distribution $\Gamma_\nu\,(\omega)=\frac{\nu^{\nu/2}}{2^{\nu/2}\Gamma(\nu/2)}\omega^{\frac{\nu-2}{2}}\mathrm{e}^{-\frac{1}{2}\nu\omega}$ for $\nu>0$. The last term in (37) is the product of CFs of gamma distributions with the same 'shape' parameter 1/2, but different 'scale' parameters $2\zeta\sigma_0^2/\omega(1-\zeta+N^{-1})\lambda_\ell(\mathbf{\Sigma})$. From (37) one obtains mean and variance of the MSE:

$$\begin{aligned}\frac{1}{d}\big\langle||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}_{\rm ML}[\mathscr{D}]||^2\big\rangle_{\mathscr{D}}&=\frac{\zeta\sigma_0^2}{1-\zeta-N^{-1}}\frac{\mathrm{Tr}[\mathbf{\Sigma}^{-1}]}{d},\\\mathrm{Var}\Big(\frac{1}{d}||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}_{\rm ML}[\mathscr{D}]||^2\Big)&=\frac{2\zeta^2\sigma_0^4}{(1-\zeta)^2}\frac{\mathrm{Tr}[\mathbf{\Sigma}^{-2}]}{d^2}.\end{aligned}\tag{38}$$

The latter gives us the condition for self-averaging of the MSE, i.e. $\mathrm{Var}\big(\frac{1}{d}||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}_{\rm ML}[\mathscr{D}]||^2\big)\to0$ as $(N,d)\to\infty$.

We finally consider deviations of $\frac{1}{d}||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}_{\rm ML}[\mathscr{D}]||^2$ from the mean $\mu(\mathbf{\Sigma})$ given in (38). The probability of observing the event event $\frac{1}{d}||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}_{\rm ML}[\mathscr{D}]||^2\notin\big(\mu(\mathbf{\Sigma})-\delta,\mu(\mathbf{\Sigma})+\delta\big)$ for $\delta>0$, is bounded from above as follows

$$\begin{aligned}&\mathrm{Prob}\left[\frac{1}{d}||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}_{\rm ML}[\mathscr{D}]||^2\notin\left(\mu\left(\mathbf{\Sigma}\right)-\delta,\mu\left(\mathbf{\Sigma}\right)+\delta\right)\right]\\&\leq \mathrm{C}_-\mathrm{e}^{-N\Phi_-[\alpha,\mu(\lambda_d),\delta]}+\mathrm{C}_+\mathrm{e}^{-N\Phi_+[\alpha,\mu(\lambda_1),\delta]}.\end{aligned}\tag{39}$$

with some small $\alpha>0$ and positive constants $\mathrm{C}_\pm$. For the rate function $\Phi_-[\alpha,\mu(\lambda_d),\delta]$ to be positive for arbitrary small δ it is sufficient that $\mu(\lambda_d)\geq1$, 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 $\delta\in(0,\mu(\lambda_1))$.

Statistical properties of the free energy –We consider the free energy (13) 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

$$\begin{aligned}\big\langle F_{\beta,\sigma^2}[\mathscr{D}]\big\rangle_{\mathscr{D}}&=\frac{d}{2\beta}+\frac{1}{2\sigma^2}\boldsymbol{\theta}_0^{\rm T}\big\langle\big(\mathbf{J}-\mathbf{J}\mathbf{J}^{-1}_{\sigma^2\eta}\mathbf{J}\big)\big\rangle_{\mathbf{Z}}\boldsymbol{\theta}_0\\&\quad+\frac{\sigma_0^2}{2\sigma^2}\big(N-\big\langle\mathrm{Tr}[\mathbf{J}\mathbf{J}^{-1}_{\sigma^2\eta}]\big\rangle_{\mathbf{Z}}\big)\\&\quad-\frac{1}{2\beta}\big\langle\,\log|2\pi\mathrm{e}\sigma^2\beta^{-1}\mathbf{J}^{-1}_{\sigma^2\eta}|\big\rangle_{\mathbf{Z}}\end{aligned}\tag{40}$$

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

$$\begin{aligned}\mathrm{Var}\left(F_{\beta,\sigma^2}\left[\mathscr{D}\right]\right) &= \mathrm{Var}\left(E_\beta[\mathscr{D}]\right) + T^2\mathrm{Var}\left(\mathrm{S}_\beta[\mathscr{D}]\right)\\ &\quad - 2T\,\mathrm{Cov}\left(E_\beta[\mathscr{D}], \mathrm{S}_\beta[\mathscr{D}]\right).\end{aligned} \tag{41}$$

The full details on each term in (41) are found in the Appendix.

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 $z_i(\mu) = \mathcal{O}(1)$, (40) gives the average free energy density

$$\begin{aligned}\Big\langle \frac{F_{\beta,\sigma^2}\left[\mathscr{D}\right]}{N}\Big\rangle_{\mathscr{D}} &= \frac{1}{2}\frac{\sigma_0^2}{\sigma^2}(1-\zeta) + \frac{\zeta}{2\beta}\log\Big(\frac{\beta}{2\pi\sigma^2\zeta}\Big)\\ &\quad + \frac{\zeta}{2\beta}\int \mathrm{d}\lambda\;\rho_d(\lambda)\log(\lambda),\end{aligned} \tag{42}$$

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

$$\rho_d(\lambda|\mathbf{Z}) = \frac{1}{d}\sum_{\ell=1}^{d}\delta\big[\lambda - \lambda_\ell(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}/N)\big]. \tag{43}$$

The variance of free energy density is

$$\begin{aligned}\mathrm{Var}\Big(\frac{F_{\beta,\sigma^2}\left[\mathscr{D}\right]}{N}\Big) &= \mathrm{Var}\Big(\frac{E[\mathscr{D}]}{N}\Big) + T^2\mathrm{Var}\Big(\frac{\mathrm{S}(P[\mathscr{D}])}{N}\Big)\\ &= \frac{\zeta^2}{4\beta^2}\int \mathrm{d}\lambda\,\mathrm{d}\tilde{\lambda}\;C_d(\lambda,\tilde{\lambda})\log(\lambda)\log(\tilde{\lambda}) + \frac{\sigma_0^4(1-\zeta)}{2\sigma^4 N}\end{aligned} \tag{44}$$

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

Finally, if we use (42) in the free energy density (18) for $\eta = 0$, and assume Gaussian data with true population covariance matrix $\mathbf{\Sigma} = \mathbf{I}_d$, then for $\beta \in [\zeta,\infty)$ we find

$$\begin{aligned}\lim_{N\to\infty} f_\beta[\mathscr{D}] &= \frac{\beta-\zeta}{2\beta}\log\Big(\frac{2\pi\sigma_0^2\,(1-\zeta)}{\beta-\zeta}\Big) + \frac{\log(\beta)+1}{2}\\ &\quad - \frac{1}{2\beta}\Big(\zeta\log\zeta + (1-\zeta)\log(1-\zeta) + 2\zeta\Big),\end{aligned} \tag{45}$$

with the convention $0\log 0 = 0$. For $\beta \in (0,\zeta)$ we get $\lim_{N\to\infty} f_\beta\left[\mathscr{D}\right] = -\infty$. 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 $(N,d) \to \infty$ [28], with $a_\pm = (1\pm\sqrt{\zeta})^2$, the free energy density is self-averaging. Its values are plotted versus the temperature in Figure 1. Furthermore, the average free energy density (45) is identical to that of [9]. Since $\lim_{\beta\downarrow\zeta}\lim_{N\to\infty} f_\beta[\mathscr{D}]$ is finite, the system exhibits a zeroth-order phase transition [29] at $T = 1/\zeta$.

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 (40) over $\boldsymbol{\theta}_0$ becomes

$$\begin{aligned}\Big\langle\Big\langle \frac{F_{\beta,\sigma^2}\left[\mathscr{D}\right]}{N}\Big\rangle_{\mathscr{D}}\Big\rangle_{\boldsymbol{\theta}_0} &= \frac{\zeta}{2\beta} + \frac{S^2\zeta\eta}{2}\int\mathrm{d}\lambda\frac{\rho_d(\lambda)\lambda}{\lambda+\zeta\sigma^2\eta}\\ &\quad + \frac{\sigma_0^2}{2\sigma^2}\Big(1 - \zeta\int\mathrm{d}\lambda\frac{\rho_d(\lambda)\lambda}{\lambda+\zeta\sigma^2\eta}\Big)\\ &\quad + \frac{\zeta}{2\beta}\int\mathrm{d}\lambda\;\rho_d(\lambda)\log\big(\lambda+\zeta\sigma^2\eta\big)\\ &\quad - \frac{\zeta}{2\beta}\log\big(2\pi\mathrm{e}\sigma^2\beta^{-1}\zeta\big)\end{aligned} \tag{46}$$

Furthermore, using (41), we obtain, under the same assumptions, that $\mathrm{Var}(F_{\beta,\sigma^2}\left[\mathscr{D}\right]/N)$ is of the form (see section F 4 in the Appendix):

$$\mathrm{Var}\Big(\frac{F_{\beta,\sigma^2}[\mathscr{D}]}{N}\Big) = \int\mathrm{d}\lambda\,\mathrm{d}\tilde{\lambda}\,C_d(\lambda,\tilde{\lambda})\Phi(\lambda,\tilde{\lambda}) + \mathcal{O}\Big(\frac{1}{N}\Big). \tag{47}$$

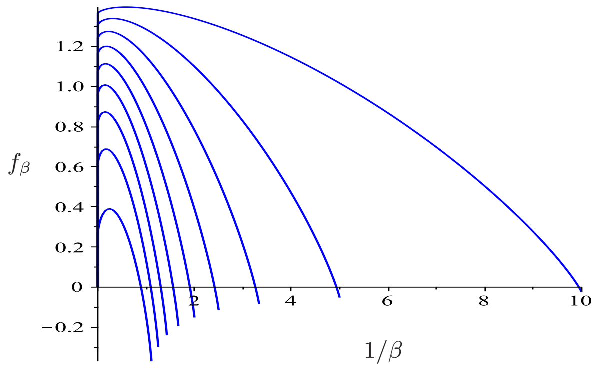

FIG. 1. Asymptotic free energy density $f_\beta = \lim_{N\to\infty} f_\beta[\mathscr{D}]$ of finite temperature ML inference as a function of temperature $T = 1/\beta$, plotted for $\zeta \in \{1/10, 2/10, \ldots, 9/10\}$ (from right to left) in the high-dimensional regime where $N, d \to \infty$ with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

Hence for $\eta > 0$ the 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 $C_d(\lambda, \lambda) \to 0$ as $(N, d) \to \infty$). This is the case e.g. for Gaussian data with covariance matrix $\mathbf{\Sigma} = \mathbf{I}_d$.

Summary and outlook–The above results emphasize that still much can be learned about high-dimensional Bayesian linear regression from exact calculations with standard methods. 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 (35), employing techniques used in [30]. 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 $\hat{\theta}_{\mathrm{MAP}}$ (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). 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.

ACKNOWLEDGMENTS

We would like to thank Dr Heather Battey and Sir David Cox for reading a first draft of this manuscript and valuable suggestions. AM is supported by Cancer Research UK (award C45074/A26553) and the UK's Medical Research Council (award MR/R014043/1). FA is supported by EIT Health ID 19277 (RGS@HOME) under H2020. FAL gratefully acknowledges financial support through a scholarship from Conacyt (Mexico).

Appendix A: Ingredients

We write $\mathbf{I}_N$ for $N \times N$ identity matrix. The data $\mathscr{D} = \{\mathbf{t}, \mathbf{Z}\}$, where $\mathbf{t} \in \mathbb{R}^N$ and $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_N)$ is the $N \times d$ matrix, is a set of observed 'input-output' pairs $\{(\mathbf{z}_1, t_1), \ldots, (\mathbf{z}_N, t_N)\}$ generated by the process

$$\mathbf{t} = \mathbf{Z}\boldsymbol{\theta}_0 + \boldsymbol{\epsilon}. \tag{A1}$$

The *true* parameter vector $\boldsymbol{\theta}_0 \in \mathbf{R}^d$ is unknown, and the vector $\boldsymbol{\epsilon} \in \mathbf{R}^N$ represents noise with mean $\mathbf{0}$ and covariance $\sigma_0^2\mathbf{I}_N$, with also the *true* noise parameter σ_0^2 unknown to us. The (empirical) covariance matrix of the input data is

$$\mathbf{J}[\mathbf{Z}] = \mathbf{Z}^{\mathrm{T}}\mathbf{Z}, \tag{A2}$$

where $[\mathbf{J}[\mathbf{Z}]]_{k\ell} = \sum_{i=1}^N z_i(k)z_i(\ell)$. To simplify notation we will sometimes omit the dependence on $\mathbf{Z}$ and write $\mathbf{J} \equiv \mathbf{J}[\mathbf{Z}]$. The *maximum a posteriori* estimator (MAP) of $\boldsymbol{\theta}_0$ in linear regression with Gaussian prior $\mathcal{N}(\mathbf{0}, \eta^{-1}\mathbf{I}_d)$ is

$$\hat{\boldsymbol{\theta}}\,[\mathscr{D}] = \mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{t}, \tag{A3}$$

where $\mathbf{J}_\eta = \mathbf{J} + \eta \mathbf{I}_d$. For $\eta = 0$ the above gives us the *maximum likelihood* (ML) estimator

$$\hat{\boldsymbol{\theta}}\left[\mathscr{D}\right] = \mathbf{J}^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{t} \tag{A4}$$

We are interested in the *high-dimensional* regime: $(N, d) \to (\infty, \infty)$ with fixed $\zeta = d/N > 0$, which we will write as $(N, d) \to \infty$ to simplify notation.

Appendix B: Distribution of $\hat{\boldsymbol{\theta}}$ estimator in MAP inference

Let us assume that the noise parameter σ^2 is *independent* of data $\mathscr{D}$, and that the noise $\boldsymbol{\epsilon}$ is sampled from the Gaussian distribution $\mathcal{N}(\mathbf{0}, \sigma_0^2 \mathbf{I}_N)$. The distribution of the MAP estimator (A3) can then be computed as follows

$$\begin{aligned}
P(\hat{\boldsymbol{\theta}}) &= \left\langle \delta\big(\hat{\boldsymbol{\theta}} - \mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{t}\big)\right\rangle_{\mathscr{D}} = \left\langle \delta\big(\hat{\boldsymbol{\theta}} - \mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{Z}^{\mathrm{T}}\left(\mathbf{Z}\boldsymbol{\theta}_0 + \boldsymbol{\epsilon}\right)\big)\right\rangle_{\mathscr{D}} \\
&= \left\langle\left\langle \delta\big(\hat{\boldsymbol{\theta}} - \mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{J}\boldsymbol{\theta}_0 - \mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{Z}^{\mathrm{T}}\boldsymbol{\epsilon}\big)\right\rangle_{\boldsymbol{\epsilon}}\right\rangle_{\mathbf{Z}} \\
&= \int \frac{\mathrm{d}\mathbf{x}}{(2\pi)^d} \mathrm{e}^{\mathrm{i}\mathbf{x}^T\hat{\boldsymbol{\theta}}} \left\langle \mathrm{e}^{-\mathrm{i}\mathbf{x}^T\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{J}\boldsymbol{\theta}_0} \left\langle \mathrm{e}^{-\mathrm{i}\mathbf{x}^T\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{Z}^{\mathrm{T}}\boldsymbol{\epsilon}}\right\rangle_{\boldsymbol{\epsilon}}\right\rangle_{\mathbf{Z}} \\
&= \int \frac{\mathrm{d}\mathbf{x}}{(2\pi)^d} \left\langle \mathrm{e}^{-\frac{1}{2}\sigma_0^2\mathbf{x}^T\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{J}\mathbf{x} + \mathrm{i}\mathbf{x}^T\left(\hat{\boldsymbol{\theta}} - \mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{J}\boldsymbol{\theta}_0\right)}\right\rangle_{\mathbf{Z}} \\
&= \frac{1}{(2\pi)^d}\left\langle \sqrt{(2\pi)^d \left|\left(\sigma_0^2\mathbf{J}_{\sigma^2\eta}^{-2}\mathbf{J}\right)^{-1}\right|} \int \mathrm{d}\mathbf{x}\, \mathcal{N}\big(\mathbf{x}|\mathbf{0}, \left(\sigma_0^2\mathbf{J}_{\sigma^2\eta}^{-2}\mathbf{J}\right)^{-1}\big) \mathrm{e}^{\mathrm{i}\mathbf{x}^T\left(\hat{\boldsymbol{\theta}} - \mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{J}\boldsymbol{\theta}_0\right)}\right\rangle_{\mathbf{Z}} \\
&= \left\langle \mathcal{N}\Big(\hat{\boldsymbol{\theta}}\,|\,\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{J}\boldsymbol{\theta}_0,\, \sigma_0^2\mathbf{J}_{\sigma^2\eta}^{-2}\mathbf{J}\Big)\right\rangle_{\mathbf{Z}}
\end{aligned} \tag{B1}$$

To take the limit $(N, d) \to \infty$, we rescale $z_i(\mu) \to z_i(\mu)/\sqrt{d}$ with $z_i(\mu) = \mathcal{O}(1)$. Now $\mathbf{J} = \mathbf{C}/\zeta$, where $\zeta = d/N$, $[\mathbf{C}]_{\mu\nu} = N^{-1}\sum_i z_i(\mu)z_i(\nu)$, and $\mathbf{J}_{\sigma^2\eta}^{-1} = \zeta\mathbf{C}_{\zeta\sigma^2\eta}^{-1}$ giving us the distribution

$$P(\hat{\boldsymbol{\theta}}) = \left\langle \mathcal{N}\Big(\hat{\boldsymbol{\theta}}\,|\,\mathbf{C}_{\zeta\sigma^2\eta}^{-1}\mathbf{C}\boldsymbol{\theta}_0,\, \zeta\sigma_0^2\mathbf{C}_{\zeta\sigma^2\eta}^{-2}\mathbf{C}\Big)\right\rangle_{\mathbf{Z}} \tag{B2}$$

with $\mathbf{C}_{\zeta\sigma^2\eta} \equiv \mathbf{C} + \zeta\sigma^2\eta\mathbf{I}$. Furthermore, since $\mathbf{C} \equiv \mathbf{C}[\mathbf{Z}]$ and $\mathbf{C}_{\zeta\sigma^2\eta}^{-1} \equiv \mathbf{C}_{\zeta\sigma^2\eta}^{-1}[\mathbf{C}[\mathbf{Z}]]$ we have that

$$\begin{aligned}
P(\hat{\boldsymbol{\theta}}) &= \left\langle \mathcal{N}\Big(\hat{\boldsymbol{\theta}}\,|\,\mathbf{C}_{\zeta\sigma^2\eta}^{-1}[\mathbf{C}[\mathbf{Z}]]\,\mathbf{C}[\mathbf{Z}]\boldsymbol{\theta}_0,\, \zeta\sigma_0^2\mathbf{C}_{\zeta\sigma^2\eta}^{-2}[\mathbf{C}[\mathbf{Z}]]\,\mathbf{C}[\mathbf{Z}]\Big)\right\rangle_{\mathbf{Z}} \\
&= \int \mathrm{d}\mathbf{C}\left\{\prod_{i=1}^N \int P(\mathbf{z}_i)\mathrm{d}\mathbf{z}_i\right\} \delta\big(\mathbf{C} - \mathbf{C}[\mathbf{Z}]\big)\mathcal{N}\Big(\hat{\boldsymbol{\theta}}\,|\,\mathbf{C}_{\zeta\sigma^2\eta}^{-1}[\mathbf{C}]\,\mathbf{C}\boldsymbol{\theta}_0,\, \zeta\sigma_0^2\mathbf{C}_{\zeta\sigma^2\eta}^{-2}[\mathbf{C}]\,\mathbf{C}\Big) \\
&= \int \mathrm{d}\mathbf{C}\int \frac{\mathrm{d}\hat{\mathbf{C}}}{(2\pi)^{d^2}}\left\{\prod_{i=1}^N \int P(\mathbf{z}_i)\mathrm{d}\mathbf{z}_i\right\} \exp\left[\mathrm{i}\mathrm{Tr}\left\{\hat{\mathbf{C}}\big(\mathbf{J} - \mathbf{J}[\mathbf{Z}]\big)\right\}\right] \mathcal{N}\Big(\hat{\boldsymbol{\theta}}\,|\,\mathbf{C}_{\zeta\sigma^2\eta}^{-1}[\mathbf{C}]\,\mathbf{C}\boldsymbol{\theta}_0,\, \zeta\sigma_0^2\mathbf{C}_{\zeta\sigma^2\eta}^{-2}[\mathbf{C}]\,\mathbf{C}\Big)
\end{aligned} \tag{B3}$$

Let us now consider the following average, assuming that $\mathbf{z}$ is sampled from the Gaussian distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$:

$$\begin{aligned}
\left\{\prod_{i=1}^N \int P(\mathbf{z}_i)\mathrm{d}\mathbf{z}_i\right\} \exp\left[-\mathrm{i}\mathrm{Tr}\left\{\hat{\mathbf{C}}\mathbf{J}[\mathbf{Z}]\right\}\right] &= \left\{\prod_{i=1}^N \int P(\mathbf{z}_i)\mathrm{d}\mathbf{z}_i\right\} \exp\left[-\mathrm{i}\frac{1}{N}\mathrm{Tr}\left\{\hat{\mathbf{C}}\sum_{i=1}^N \mathbf{z}_i\mathbf{z}_i^{\mathrm{T}}\right\}\right] \\
&= \left[\int P(\mathbf{z})\,\mathrm{d}\mathbf{z}\,\mathrm{e}^{-\mathrm{i}\frac{1}{N}\mathrm{Tr}\left\{\hat{\mathbf{C}}\,\mathbf{z}\mathbf{z}^{\mathrm{T}}\right\}}\right]^N = \left|\mathbf{I} + \frac{2\mathrm{i}}{N}\boldsymbol{\Sigma}\hat{\mathbf{C}}\right|^{-\frac{N}{2}}.
\end{aligned} \tag{B4}$$

This is the characteristic function of the Wishart distribution [31], defined by the density

$$\mathcal{W}\big(\mathbf{J}|\boldsymbol{\Sigma}/N, d, N\big) = \frac{|\boldsymbol{\Sigma}/N|^{-\frac{N}{2}}\,|\mathbf{J}|^{\frac{N-d-1}{2}}}{2^{\frac{Nd}{2}}\pi^{\frac{d(d-1)}{4}}\prod_{\ell=1}^d \Gamma\big(\frac{N+1-\ell}{2}\big)}\mathrm{e}^{-N\frac{1}{2}\mathrm{Tr}\left(\mathbf{J}\boldsymbol{\Sigma}^{-1}\right)}. \tag{B5}$$

The Wishart distribution is *singular* when $d > N$. Thus for Gaussian $\mathbf{z}$ the distribution (B1) is the Gaussian mixture

$$
\begin{aligned}
P(\hat{\boldsymbol{\theta}}) &= \int \mathrm{d}\mathbf{C}\; \mathcal{W}(\mathbf{C}|\boldsymbol{\Sigma}/N, d, N)\mathcal{N}\Big(\hat{\boldsymbol{\theta}}\,|\,\mathbf{C}^{-1}_{\zeta\sigma^2\eta}[\mathbf{C}]\,\mathbf{C}\boldsymbol{\theta}_0,\, \zeta\sigma_0^2\mathbf{C}^{-2}_{\zeta\sigma^2\eta}[\mathbf{C}]\,\mathbf{C}\Big) \\
&= \int \mathrm{d}\mathbf{C}\; \frac{|\boldsymbol{\Sigma}/N|^{-\frac{N}{2}}\,|\mathbf{C}|^{\frac{N-d-1}{2}}\;\mathrm{e}^{-N\frac{1}{2}\mathrm{Tr}\left(\mathbf{C}\boldsymbol{\Sigma}^{-1}\right)}}{2^{\frac{Nd}{2}}\pi^{\frac{d(d-1)}{4}}\prod_{\ell=1}^{d}\Gamma\left(\frac{N+1-\ell}{2}\right)}\, \frac{\mathrm{e}^{-\frac{1}{2\zeta\sigma_0^2}\left(\hat{\boldsymbol{\theta}}-\mathbf{C}^{-1}_{\zeta\sigma^2\eta}\mathbf{C}\boldsymbol{\theta}_0\right)^{\mathrm{T}}\left(\mathbf{C}^{-2}_{\zeta\sigma^2\eta}[\mathbf{C}]\,\mathbf{C}\right)^{-1}[\mathbf{C}]\left(\hat{\boldsymbol{\theta}}-\mathbf{C}^{-1}_{\zeta\sigma^2\eta}\mathbf{C}\boldsymbol{\theta}_0\right)}}{\left|2\pi\zeta\sigma_0^2\mathbf{C}^{-2}_{\zeta\sigma^2\eta}[\mathbf{C}]\,\mathbf{C}\right|^{\frac{1}{2}}}. \qquad \text{(B6)}
\end{aligned}
$$

We note that an alternative derivation of this result is provided in [9].

Appendix C: Distribution of $\hat{\theta}$ estimator in ML inference

Let us consider the following integral appearing in the distribution of MAP estimator (B6):

$$
\begin{aligned}
&\int \mathrm{d}\mathbf{C}\; |\mathbf{C}|^{\frac{N-d-1}{2}}\;\mathrm{e}^{-N\frac{1}{2}\mathrm{Tr}\left(\mathbf{C}\boldsymbol{\Sigma}^{-1}\right)}\, \frac{\mathrm{e}^{-\frac{1}{2\zeta\sigma_0^2}\left(\hat{\boldsymbol{\theta}}-\mathbf{C}^{-1}_{\zeta\sigma^2\eta}\mathbf{C}\boldsymbol{\theta}_0\right)^{\mathrm{T}}\left(\mathbf{C}^{-2}_{\zeta\sigma^2\eta}[\mathbf{C}]\,\mathbf{C}\right)^{-1}[\mathbf{C}]\left(\hat{\boldsymbol{\theta}}-\mathbf{C}^{-1}_{\zeta\sigma^2\eta}\mathbf{C}\boldsymbol{\theta}_0\right)}}{\left|2\pi\zeta\sigma_0^2\mathbf{C}^{-2}_{\zeta\sigma^2\eta}[\mathbf{C}]\,\mathbf{C}\right|^{\frac{1}{2}}} \\
&= \int \mathrm{d}\mathbf{C}\; |\mathbf{C}|^{\frac{N-d-1}{2}}\; \frac{\mathrm{e}^{-\frac{1}{2}\mathrm{Tr}\left\{\mathbf{C}\left[N\boldsymbol{\Sigma}^{-1}+\mathbf{C}^{-1}\frac{1}{\zeta\sigma_0^2}\left(\mathbf{C}^{-2}_{\zeta\sigma^2\eta}[\mathbf{C}]\,\mathbf{C}\right)^{-1}[\mathbf{C}]\left(\hat{\boldsymbol{\theta}}-\mathbf{C}^{-1}_{\zeta\sigma^2\eta}\mathbf{C}\boldsymbol{\theta}_0\right)\left(\hat{\boldsymbol{\theta}}-\mathbf{C}^{-1}_{\zeta\sigma^2\eta}\mathbf{C}\boldsymbol{\theta}_0\right)^{\mathrm{T}}\right]\right\}}}{\left|2\pi\zeta\sigma_0^2\mathbf{C}^{-2}_{\zeta\sigma^2\eta}[\mathbf{C}]\,\mathbf{C}\right|^{\frac{1}{2}}}
\end{aligned}
$$

For $\eta = 0$, i.e. ML inference, this integral simplifies to

$$
\int \mathrm{d}\mathbf{C}\; |\mathbf{C}|^{\frac{N-d-1}{2}}\; \frac{\mathrm{e}^{-\frac{1}{2}\mathrm{Tr}\left\{\mathbf{C}\left[N\boldsymbol{\Sigma}^{-1}+\frac{1}{\zeta\sigma_0^2}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)^{\mathrm{T}}\right]\right\}}}{|2\pi\zeta\sigma_0^2\mathbf{C}^{-1}|^{\frac{1}{2}}} = \left(\frac{1}{2\pi\zeta\sigma_0^2}\right)^{\frac{d}{2}} \int \mathrm{d}\mathbf{C}\; |\mathbf{C}|^{\frac{N-d}{2}}\;\mathrm{e}^{-\frac{1}{2}\mathrm{Tr}\left\{\mathbf{C}\left[N\boldsymbol{\Sigma}^{-1}+\frac{1}{\zeta\sigma_0^2}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)^{\mathrm{T}}\right]\right\}}
$$

and can be computed by using the normalization identity $\int \mathrm{d}\mathbf{J}\,\mathcal{W}\big(\mathbf{J}|\boldsymbol{\Sigma}, d, N\big) = 1$, from which one obtains

$$
\int \mathrm{d}\mathbf{C}\; |\mathbf{C}|^{\frac{N-d-1}{2}}\;\mathrm{e}^{-\frac{1}{2}\mathrm{Tr}\left(\mathbf{C}\boldsymbol{\Sigma}^{-1}\right)} = |\boldsymbol{\Sigma}|^{\frac{N}{2}}\, 2^{\frac{Nd}{2}}\pi^{\frac{d(d-1)}{4}}\prod_{\ell=1}^{d}\Gamma\left(\frac{N+1-\ell}{2}\right). \qquad \text{(C1)}
$$

Hence also

$$
\int \mathrm{d}\mathbf{C}\; |\mathbf{C}|^{\frac{N-d}{2}}\;\mathrm{e}^{-\frac{1}{2}\mathrm{Tr}\left(\mathbf{C}\boldsymbol{\Sigma}^{-1}\right)} = |\boldsymbol{\Sigma}|^{\frac{N+1}{2}}\, 2^{\frac{(N+1)d}{2}}\pi^{\frac{d(d-1)}{4}}\prod_{\ell=1}^{d}\Gamma\left(\frac{N+2-\ell}{2}\right) \qquad \text{(C2)}
$$

which gives us the result

$$
\int \mathrm{d}\mathbf{C}\; |\mathbf{C}|^{\frac{N-d}{2}}\;\mathrm{e}^{-\frac{1}{2}\mathrm{Tr}\left\{\mathbf{C}\left[N\boldsymbol{\Sigma}^{-1}+\frac{1}{\zeta\sigma_0^2}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)^{\mathrm{T}}\right]\right\}} = \left|N\boldsymbol{\Sigma}^{-1}+\frac{(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^{\mathrm{T}}}{\zeta\sigma_0^2}\right|^{-\frac{N+1}{2}} 2^{\frac{(N+1)d}{2}}\pi^{\frac{d(d-1)}{4}}\prod_{\ell=1}^{d}\Gamma\left(\frac{N+2-\ell}{2}\right). \qquad \text{(C3)}
$$

For $\eta = 0$ the distribution (B6) thereby becomes

$$
\begin{aligned}
P(\hat{\boldsymbol{\theta}}) &= \int \mathrm{d}\mathbf{C}\; \frac{|\boldsymbol{\Sigma}/N|^{-\frac{N}{2}}\,|\mathbf{C}|^{\frac{N-d-1}{2}}}{2^{\frac{Nd}{2}}\pi^{\frac{d(d-1)}{4}}\prod_{\ell=1}^{d}\Gamma\left(\frac{N+1-\ell}{2}\right)}\mathrm{e}^{-N\frac{1}{2}\mathrm{Tr}\left(\mathbf{C}\boldsymbol{\Sigma}^{-1}\right)}\, \frac{\mathrm{e}^{-\frac{1}{2}\frac{1}{\zeta\sigma_0^2}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)^{\mathrm{T}}\mathbf{C}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)}}{|2\pi\zeta\sigma_0^2\mathbf{C}^{-1}|^{\frac{1}{2}}}. \\
&= \frac{|\boldsymbol{\Sigma}/N|^{-\frac{N}{2}}}{2^{\frac{Nd}{2}}\pi^{\frac{d(d-1)}{4}}\prod_{\ell=1}^{d}\Gamma\left(\frac{N+1-\ell}{2}\right)} \int \mathrm{d}\mathbf{C}\; |\mathbf{C}|^{\frac{N-d-1}{2}}\;\mathrm{e}^{-\frac{1}{2}\mathrm{Tr}\left(\mathbf{C}N\boldsymbol{\Sigma}^{-1}\right)}\, \frac{\mathrm{e}^{-\frac{1}{2}\frac{1}{\zeta\sigma_0^2}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)^{\mathrm{T}}\mathbf{C}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)}}{|2\pi\zeta\sigma_0^2\mathbf{C}^{-1}|^{\frac{1}{2}}}. \\
&= \left(\frac{1}{\pi\zeta\sigma_0^2}\right)^{\frac{d}{2}}\prod_{\ell=1}^{d}\frac{\Gamma\left(\frac{N+2-\ell}{2}\right)}{\Gamma\left(\frac{N+1-\ell}{2}\right)}\,|\boldsymbol{\Sigma}/N|^{-\frac{N}{2}}\left|N\boldsymbol{\Sigma}^{-1}+\frac{1}{\zeta\sigma_0^2}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)^{\mathrm{T}}\right|^{-\frac{N+1}{2}} \\
&= \pi^{-\frac{d}{2}}\frac{\Gamma\left(\frac{N+1}{2}\right)}{\Gamma\left(\frac{N+1-d}{2}\right)}\left|\frac{\boldsymbol{\Sigma}}{\zeta\sigma_0^2 N}\right|^{\frac{1}{2}}\left(1+(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^{\mathrm{T}}\frac{\boldsymbol{\Sigma}}{\zeta\sigma_0^2 N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\right)^{-\frac{N+1}{2}}. \qquad \text{(C4)}
\end{aligned}
$$

The last line in above was obtained using the 'matrix determinant lemma'. Thus, after slight rearrangement,

$$P(\hat{\boldsymbol{\theta}}) = [\pi(N+1-d)]^{-\frac{d}{2}} \frac{\Gamma\left(\frac{N+1}{2}\right)}{\Gamma\left(\frac{N+1-d}{2}\right)} \left|\frac{(1-\zeta+1/N)\boldsymbol{\Sigma}}{\zeta\sigma_0^2}\right|^{\frac{1}{2}} \left(1+(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^{\mathrm{T}}\frac{(1-\zeta+1/N)\boldsymbol{\Sigma}}{\zeta\sigma_0^2(N+1-d)}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\right)^{-\frac{N+1}{2}}, \tag{C5}$$

which is the multivariate Student's t-distribution, with $N+1-d$ degrees of freedom, 'location' vector $\boldsymbol{\theta}_0$ and 'shape' matrix $\zeta\sigma_0^2\boldsymbol{\Sigma}^{-1}/(1-\zeta+1/N)$.

Appendix D: Statistical properties of $\hat{\sigma}^2$ estimator in ML inference

In ML inference the estimator of $\boldsymbol{\theta}$ is given by (A4) and the estimator of noise parameter σ^2 is given by the density

$$\begin{aligned}
\hat{\sigma}^2[\mathscr{D}] &= \frac{1}{N}\left|\left|\mathbf{t}-\mathbf{Z}\hat{\boldsymbol{\theta}}[\mathscr{D}]\right|\right|^2 = \frac{1}{N}\left|\left|\mathbf{t}-\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{t}\right|\right|^2 = \frac{1}{N}\left|\left|\left(\mathbf{I}_N-\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\right)\mathbf{t}\right|\right|^2\\
&= \frac{1}{N}\left|\left|\left(\mathbf{I}_N-\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\right)\left(\mathbf{Z}\boldsymbol{\theta}_0+\boldsymbol{\epsilon}\right)\right|\right|^2\\
&= \frac{1}{N}\left|\left|\left(\mathbf{I}_N-\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\right)\mathbf{Z}\boldsymbol{\theta}_0+\left(\mathbf{I}_N-\mathbf{Z}\,\mathbf{J}^{-1}[\mathbf{Z}]\mathbf{Z}^{\mathrm{T}}\right)\boldsymbol{\epsilon}\right|\right|^2\\
&= \frac{1}{N}\left|\left|\left(\mathbf{I}_N-\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\right)\boldsymbol{\epsilon}\right|\right|^2 = \frac{1}{N}\boldsymbol{\epsilon}^{\mathrm{T}}\left(\mathbf{I}_N-\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\right)^2\boldsymbol{\epsilon} = \frac{1}{N}\boldsymbol{\epsilon}^{\mathrm{T}}\left(\mathbf{I}_N-\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\right)\boldsymbol{\epsilon}
\end{aligned} \tag{D1}$$

In above we used $(\mathbf{I}_N-\mathbf{Z}\,(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}})\mathbf{Z}\boldsymbol{\theta}_0 = \mathbf{Z}\boldsymbol{\theta}_0-\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\boldsymbol{\theta}_0 = \mathbf{0}$ and $(\mathbf{I}_N-\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}})^2 = \mathbf{I}_N-\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}$, i.e. $\mathbf{I}_N-\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}$ is an *idempotent* matrix. For $\boldsymbol{\epsilon}$ sampled from *any* distribution with mean $\mathbf{0}$ and covariance $\sigma_0^2\mathbf{I}_N$, the average and variance of $\hat{\sigma}^2[\mathscr{D}]$ are (by Wick's theorem):

$$\left\langle\hat{\sigma}^2[\mathscr{D}]\right\rangle_{\boldsymbol{\epsilon}} = \frac{1}{N}\left\langle\boldsymbol{\epsilon}^{\mathrm{T}}\left(\mathbf{I}_N-\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\right)\boldsymbol{\epsilon}\right\rangle_{\boldsymbol{\epsilon}} = \frac{\sigma_0^2}{N}\mathrm{Tr}\left(\mathbf{I}_N-\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\right) = \sigma_0^2\left(1-\zeta\right) \tag{D2}$$

$$\left\langle\hat{\sigma}^4[\mathscr{D}]\right\rangle_{\boldsymbol{\epsilon}} - \left\langle\hat{\sigma}^2[\mathscr{D}]\right\rangle_{\boldsymbol{\epsilon}}^2 = \frac{2\sigma_0^4}{N}\left(1-\zeta\right) \tag{D3}$$

Next we are interested in the probability of event $\hat{\sigma}^2[\mathscr{D}] \notin \left(\sigma_0^2(1-\zeta)-\delta, \sigma_0^2(1-\zeta)+\delta\right)$. This is given by

$$\begin{aligned}
\mathrm{Prob}\left[\frac{1}{N}\sum_{i=1}^N\left(t_i-\hat{\boldsymbol{\theta}}[\mathscr{D}].\mathbf{z}_i\right)^2 \notin \left(\sigma_0^2(1-\zeta)-\delta, \sigma_0^2(1-\zeta)+\delta\right)\right] &= \mathrm{Prob}\left[\sum_{i=1}^N\left(t_i-\hat{\boldsymbol{\theta}}[\mathscr{D}].\mathbf{z}_i\right)^2 \leq N\left(\sigma_0^2(1-\zeta)-\delta\right)\right]\\
&+ \mathrm{Prob}\left[\sum_{i=1}^N\left(t_i-\hat{\boldsymbol{\theta}}[\mathscr{D}].\mathbf{z}_i\right)^2 \geq N\left(\sigma_0^2(1-\zeta)+\delta\right)\right].
\end{aligned} \tag{D4}$$

First, we consider the probability

$$\begin{aligned}
\mathrm{Prob}\left[\sum_{i=1}^N\left(t_i-\hat{\boldsymbol{\theta}}[\mathscr{D}].\mathbf{z}_i\right)^2 \geq N\left(\sigma_0^2(1-\zeta)+\delta\right)\right] &= \mathrm{Prob}\left[\mathrm{e}^{\frac{1}{2}\alpha\sum_{i=1}^N\left(t_i-\hat{\boldsymbol{\theta}}[\mathscr{D}].\mathbf{z}_i\right)^2} \geq \mathrm{e}^{\frac{1}{2}\alpha N\left(\sigma_0^2(1-\zeta)+\delta\right)}\right]\\
&\leq \left\langle\mathrm{e}^{\frac{1}{2}\alpha\sum_{i=1}^N\left(t_i-\hat{\boldsymbol{\theta}}[\mathscr{D}].\mathbf{z}_i\right)^2}\right\rangle_{\mathscr{D}}\mathrm{e}^{-\frac{1}{2}\alpha N\left(\sigma_0^2(1-\zeta)+\delta\right)},
\end{aligned} \tag{D5}$$

where i$\alpha > 0$ and we used Markov inequality to derive the upper bound. Let us assume that the distribution of noise is *Gaussian* and consider the moment-generating function

$$
\begin{aligned}
\left\langle \mathrm{e}^{\frac{1}{2}\alpha\sum_{i=1}^{N}\left(t_i-\hat{\boldsymbol{\theta}}[\mathscr{D}].\mathbf{z}_i\right)^2}\right\rangle_{\mathscr{D}} &= \left\langle \mathrm{e}^{\frac{1}{2}\alpha\boldsymbol{\epsilon}^{\mathrm{T}}\left(\mathbf{I}_N-\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\right)\boldsymbol{\epsilon}}\right\rangle_{\mathscr{D}} = \int \mathrm{d}\boldsymbol{\epsilon}\,\mathcal{N}\left(\boldsymbol{\epsilon}|\mathbf{0},\sigma_0^2\mathbf{I}_N\right)\left\langle \mathrm{e}^{\frac{1}{2}\alpha\boldsymbol{\epsilon}^{\mathrm{T}}\left(\mathbf{I}_N-\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\right)\boldsymbol{\epsilon}}\right\rangle_{\mathbf{Z}} \\
&= \frac{1}{\left(2\pi\sigma_0^2\right)^{N/2}}\left\langle \int \mathrm{d}\boldsymbol{\epsilon}\,\mathrm{e}^{-\frac{1}{2\sigma_0^2}\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\epsilon}+\frac{1}{2}\alpha\boldsymbol{\epsilon}^{\mathrm{T}}\left(\mathbf{I}_N-\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\right)\boldsymbol{\epsilon}}\right\rangle_{\mathbf{Z}} \\
&= \frac{1}{\left(2\pi\sigma_0^2\right)^{N/2}}\left\langle \int \mathrm{d}\boldsymbol{\epsilon}\,\mathrm{e}^{-\frac{1}{2}\boldsymbol{\epsilon}^{\mathrm{T}}\left[\mathbf{I}_N/\sigma_0^2-\alpha\left(\mathbf{I}_N-\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\right)\right]\boldsymbol{\epsilon}}\right\rangle_{\mathbf{Z}} \\
&= \frac{1}{\left(2\pi\sigma_0^2\right)^{N/2}}\left\langle \left|2\pi\left[\mathbf{I}_N/\sigma_0^2-\alpha\left(\mathbf{I}_N-\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\right)\right]^{-1}\right|^{\frac{1}{2}}\right\rangle_{\mathbf{Z}} \\
&= \left\langle \left|\mathbf{I}_N-\alpha\sigma_0^2\left(\mathbf{I}_N-\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\right)\right|^{-\frac{1}{2}}\right\rangle_{\mathbf{Z}} = \left\langle \left|\mathbf{I}_N(1-\alpha\sigma_0^2)+\alpha\sigma_0^2\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\right|^{-\frac{1}{2}}\right\rangle_{\mathbf{Z}} \\
&= \left(1-\alpha\sigma_0^2\right)^{-N/2}\left\langle \frac{1}{\left|\mathbf{I}_N+\frac{\alpha\sigma_0^2}{1-\alpha\sigma_0^2}\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\right|^{\frac{1}{2}}}\right\rangle_{\mathbf{Z}}
\end{aligned}
\tag{D6}
$$

Now $\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}$ is a *projection* matrix, and its eigenvalue are $\lambda_i \in \{0,1\}$, giving us

$$
\begin{aligned}
\left\langle \mathrm{e}^{\frac{1}{2}\alpha\sum_{i=1}^{N}\left(t_i-\hat{\boldsymbol{\theta}}[\mathscr{D}].\mathbf{z}_i\right)^2}\right\rangle_{\mathscr{D}} &= \left(1-\alpha\sigma_0^2\right)^{-N/2}\left\langle \frac{1}{\left|\mathbf{I}_N+\frac{\alpha\sigma_0^2}{1-\alpha\sigma_0^2}\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\right|^{\frac{1}{2}}}\right\rangle_{\mathbf{Z}} \\
&= \left(1-\alpha\sigma_0^2\right)^{-N/2}\left\langle \frac{1}{\left(\prod_{i=1}^{N}\left[1+\frac{\alpha\sigma_0^2}{1-\alpha\sigma_0^2}\lambda_i\left(\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\right)\right]\right)^{\frac{1}{2}}}\right\rangle_{\mathbf{Z}} \\
&= \left(1-\alpha\sigma_0^2\right)^{-N/2}\left\langle \mathrm{e}^{-\frac{N}{2}\sum_{\lambda}\frac{1}{N}\sum_{i=1}^{N}\delta_{\lambda;\lambda_i\left(\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\right)}\log\left[1+\frac{\alpha\sigma_0^2}{1-\alpha\sigma_0^2}\lambda\right]}\right\rangle_{\mathbf{Z}} \\
&= \left(1-\alpha\sigma_0^2\right)^{-N/2}\left\langle \mathrm{e}^{-\frac{N}{2}\log\left[1+\frac{\alpha\sigma_0^2}{1-\alpha\sigma_0^2}\right]\frac{1}{N}\sum_{i=1}^{N}\lambda_i\left(\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\right)}\right\rangle_{\mathbf{Z}} \\
&= \left(1-\alpha\sigma_0^2\right)^{-N/2}\left\langle \mathrm{e}^{-\frac{1}{2}\log\left(1+\frac{\alpha\sigma_0^2}{1-\alpha\sigma_0^2}\right)\mathrm{Tr}\left\{\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\right\}}\right\rangle_{\mathbf{Z}} \\
&= \left(1-\alpha\sigma_0^2\right)^{-N/2}\mathrm{e}^{-\frac{d}{2}\log\left(1+\frac{\alpha\sigma_0^2}{1-\alpha\sigma_0^2}\right)} \\
&= \mathrm{e}^{-\frac{N}{2}\left[\zeta\log\left(1+\frac{\alpha\sigma_0^2}{1-\alpha\sigma_0^2}\right)+\log\left(1-\alpha\sigma_0^2\right)\right]} = \mathrm{e}^{-\frac{N}{2}\left(1-\zeta\right)\log\left(1-\alpha\sigma_0^2\right)}.
\end{aligned}
\tag{D7}
$$

Hence

$$
\begin{aligned}
\mathrm{Prob}\left[\sum_{i=1}^{N}\left(t_i-\hat{\boldsymbol{\theta}}[\mathscr{D}].\mathbf{z}_i\right)^2 \geq N\left(\sigma_0^2(1-\zeta)+\delta\right)\right] &\leq \left\langle \mathrm{e}^{\frac{1}{2}\alpha\sum_{i=1}^{N}\left(t_i-\hat{\boldsymbol{\theta}}[\mathscr{D}].\mathbf{z}_i\right)^2}\right\rangle_{\mathscr{D}}\mathrm{e}^{-\frac{1}{2}\alpha N\left(\sigma_0^2(1-\zeta)+\delta\right)} \\
&\leq \mathrm{e}^{-\frac{1}{2}N\left(1-\zeta\right)\log\left(1-\alpha\sigma_0^2\right)}\mathrm{e}^{-\frac{1}{2}\alpha N\left(\sigma_0^2(1-\zeta)+\delta\right)} \\
&= \mathrm{e}^{-\frac{1}{2}N\left[\left(1-\zeta\right)\log\left(1-\alpha\sigma_0^2\right)+\alpha\left(\sigma_0^2(1-\zeta)+\delta\right)\right]} = \mathrm{e}^{-\frac{1}{2}N\Phi(\alpha)}.
\end{aligned}
\tag{D8}
$$

The *rate* function

$$
\Phi(\alpha) = \left(1-\zeta\right)\log\left(1-\alpha\sigma_0^2\right)+\alpha\left(\sigma_0^2(1-\zeta)+\delta\right)
\tag{D9}
$$

has a *maximum* at

$$
\alpha = \frac{\delta}{\sigma_0^2\left((1-\zeta)\sigma_0^2+\delta\right)},
\tag{D10}
$$

such that $\max_\alpha \Phi(\alpha) = \big(1-\zeta\big)\log\big(\frac{(1-\zeta)}{(1-\zeta)+\delta/\sigma_0^2}\big) + \delta/\sigma_0^2$, and hence

$$\mathrm{Prob}\left[\sum_{i=1}^{N}\big(t_i - \hat{\boldsymbol{\theta}}[\mathscr{D}].\mathbf{z}_i\big)^2 \geq N\big(\sigma_0^2(1-\zeta)+\delta\big)\right] \leq \mathrm{e}^{-\frac{1}{2}N\left(\left(1-\zeta\right)\log\left(\frac{(1-\zeta)}{(1-\zeta)+\delta/\sigma_0^2}\right)+\delta/\sigma_0^2\right)}. \tag{D11}$$

We note that in above expression the rate function $(1-\zeta)\log[(1-\zeta)/(1-\zeta+\delta/\sigma_0^2)] + \delta/\sigma_0^2$ vanishes when $\delta/\sigma_0^2 = 0$, and is a monotonically increasing function of δ/σ_0^2.

Second, for $\alpha > 0$ we consider the probability

$$\begin{aligned}\mathrm{Prob}\left[\sum_{i=1}^{N}\big(t_i - \hat{\boldsymbol{\theta}}[\mathscr{D}].\mathbf{z}_i\big)^2 \leq N\big(\sigma_0^2(1-\zeta)-\delta\big)\right] &= \mathrm{Prob}\left[\mathrm{e}^{-\frac{1}{2}\alpha\sum_{i=1}^{N}\left(t_i-\hat{\boldsymbol{\theta}}[\mathscr{D}].\mathbf{z}_i\right)^2} \geq \mathrm{e}^{-\frac{1}{2}\alpha N\left(\sigma_0^2(1-\zeta)-\delta\right)}\right] \\ &\leq \left\langle \mathrm{e}^{-\frac{1}{2}\alpha\sum_{i=1}^{N}\left(t_i-\hat{\boldsymbol{\theta}}[\mathscr{D}].\mathbf{z}_i\right)^2}\right\rangle_{\mathscr{D}} \mathrm{e}^{\frac{1}{2}\alpha N\left(\sigma_0^2(1-\zeta)-\delta\right)} \\ &= \mathrm{e}^{-\frac{N}{2}\left(1-\zeta\right)\log\left(1+\alpha\sigma_0^2\right)}\mathrm{e}^{\frac{1}{2}\alpha N\left(\sigma_0^2(1-\zeta)-\delta\right)} = \mathrm{e}^{-\frac{N}{2}\phi(\alpha)},\end{aligned} \tag{D12}$$

where

$$\phi(\alpha) = \big(1-\zeta\big)\log\big(1+\alpha\sigma_0^2\big) - \alpha\big(\sigma_0^2(1-\zeta)-\delta\big). \tag{D13}$$

Here we used the Markov inequality and the result (D7) with $\alpha \to -\alpha$. The rate function $\phi(\alpha)$ has a maximum at $\alpha = \delta/\sigma_0^2((1-\zeta)\sigma_0^2-\delta)$, such that $\max_\alpha \phi(\alpha) = (1-\zeta)\log[(1-\zeta)/(1-\zeta-\delta/\sigma_0^2)] - \delta/\sigma_0^2$, and hence

$$\mathrm{Prob}\left[\sum_{i=1}^{N}\big(t_i - \hat{\boldsymbol{\theta}}[\mathscr{D}].\mathbf{z}_i\big)^2 \leq N\big(\sigma_0^2(1-\zeta)-\delta\big)\right] \leq \mathrm{e}^{-\frac{N}{2}\left(\left(1-\zeta\right)\log\left(\frac{(1-\zeta)}{(1-\zeta)-\delta/\sigma_0^2}\right)-\delta/\sigma_0^2\right)}. \tag{D14}$$

Here the rate function $(1-\zeta)\log[(1-\zeta)/(1-\zeta-\delta/\sigma_0^2)] - \delta/\sigma_0^2$ vanishes when $\delta/\sigma_0^2 = 0$, and is a monotonic increasing function of δ/σ_0^2 when $\sigma_0^2\big(1-\zeta\big) > \delta$.

Finally, combining the inequalities (D11) and (D14) we obtain the inequality

$$\begin{aligned}\mathrm{Prob}\left[\hat{\sigma}^2[\mathscr{D}] \notin \big(\sigma_0^2(1-\zeta)-\delta, \sigma_0^2(1-\zeta)+\delta\big)\right] &\leq \mathrm{e}^{-\frac{1}{2}N\left[\left(\left(1-\zeta\right)\log\left(\frac{(1-\zeta)}{(1-\zeta)-\delta/\sigma_0^2}\right)-\delta/\sigma_0^2\right)\right]} \\ &+ \mathrm{e}^{-\frac{1}{2}N\left[\left(\left(1-\zeta\right)\log\left(\frac{(1-\zeta)}{(1-\zeta)+\delta/\sigma_0^2}\right)+\delta/\sigma_0^2\right)\right]}\end{aligned} \tag{D15}$$

which is valid for $\delta \in (0, \sigma_0^2(1-\zeta))$.

Appendix E: Statistical properties of MSE in ML inference

In this section we consider statistical properties of the *minimum square error* (MSE) $\frac{1}{d}||\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}[\mathscr{D}]||^2$, where $\boldsymbol{\theta}_0$ is the vector of the *true* parameters responsible for the data, and $\hat{\boldsymbol{\theta}}[\mathscr{D}]$ is the ML estimator (A4).

1. Moment generating function

Let us consider the moment generating function

$$\begin{aligned}
\Big\langle \mathrm{e}^{\frac{1}{2}\alpha||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}[\mathscr{D}]||^2}\Big\rangle_{\mathscr{D}} &= \int \mathrm{d}\hat{\boldsymbol{\theta}}\; P(\hat{\boldsymbol{\theta}})\,\mathrm{e}^{\frac{1}{2}\alpha||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}||^2} && \text{(E1)}\\
&= \big(\pi(N+1-d)\big)^{-\frac{d}{2}}\frac{\Gamma\big(\frac{N+1}{2}\big)}{\Gamma\big(\frac{N+1-d}{2}\big)}\left|\frac{(1-\zeta+1/N)\boldsymbol{\Sigma}}{\zeta\sigma_0^2}\right|^{\frac{1}{2}}\\
&\quad\times\int \mathrm{d}\hat{\boldsymbol{\theta}}\;\Big(1+\big(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\big)^{\mathrm{T}}\frac{1}{N+1-d}\frac{(1-\zeta+1/N)\boldsymbol{\Sigma}}{\zeta\sigma_0^2}\big(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\big)\Big)^{-\frac{N+1}{2}}\mathrm{e}^{\frac{1}{2}\alpha||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}||^2}\\
&= \big(\pi(N+1-d)\big)^{-\frac{d}{2}}\frac{\Gamma\big(\frac{N+1}{2}\big)}{\Gamma\big(\frac{N+1-d}{2}\big)}\left|\frac{(1-\zeta+1/N)\boldsymbol{\Sigma}}{\zeta\sigma_0^2}\right|^{\frac{1}{2}}\\
&\quad\times\int \mathrm{d}\hat{\boldsymbol{\theta}}\;\Big(1+\hat{\boldsymbol{\theta}}^{\mathrm{T}}\frac{1}{N+1-d}\frac{(1-\zeta+1/N)\boldsymbol{\Sigma}}{\zeta\sigma_0^2}\hat{\boldsymbol{\theta}}\Big)^{-\frac{N+1}{2}}\mathrm{e}^{\frac{1}{2}\alpha||\hat{\boldsymbol{\theta}}||^2}\\
&= \int \mathrm{d}\hat{\boldsymbol{\theta}}\int_0^\infty \mathrm{d}\omega\;\mathcal{N}\Big(\hat{\boldsymbol{\theta}}\,|\,\mathbf{0},\frac{\zeta\sigma_0^2}{\omega(1-\zeta+1/N)}\boldsymbol{\Sigma}^{-1}\Big)\Gamma_{N+1-d}\left(\omega\right)\mathrm{e}^{\frac{1}{2}\alpha||\hat{\boldsymbol{\theta}}||^2}\\
&= \int \mathrm{d}\hat{\boldsymbol{\theta}}\int_0^\infty \mathrm{d}\omega\;\frac{\mathrm{e}^{-\frac{1}{2}\hat{\boldsymbol{\theta}}^{\mathrm{T}}\left(\frac{\omega(1-\zeta+1/N)\boldsymbol{\Sigma}}{\zeta\sigma_0^2}-\alpha\mathbf{I}_d\right)\hat{\boldsymbol{\theta}}}}{\left|2\pi\frac{\zeta\sigma_0^2}{\omega(1-\zeta+1/N)}\boldsymbol{\Sigma}^{-1}\right|^{\frac{1}{2}}}\;\Gamma_{N+1-d}\left(\omega\right)\\
&= \int_0^\infty \mathrm{d}\omega\;\frac{\Gamma_{N+1-d}\left(\omega\right)}{\left|\frac{\zeta\sigma_0^2}{\omega(1-\zeta+1/N)}\boldsymbol{\Sigma}^{-1}\right|^{\frac{1}{2}}\left|\frac{\omega(1-\zeta+1/N)\boldsymbol{\Sigma}}{\zeta\sigma_0^2}-\alpha\mathbf{I}_d\right|^{\frac{1}{2}}}\\
&= \int_0^\infty \mathrm{d}\omega\;\frac{\Gamma_{N+1-d}\left(\omega\right)}{\left|\mathbf{I}_d-\alpha\frac{\zeta\sigma_0^2}{\omega(1-\zeta+1/N)}\boldsymbol{\Sigma}^{-1}\right|^{\frac{1}{2}}} = \int_0^\infty \mathrm{d}\omega\;\mathrm{e}^{-\frac{1}{2}\log\left|\mathbf{I}_d-\alpha\frac{\zeta\sigma_0^2}{\omega(1-\zeta+1/N)}\boldsymbol{\Sigma}^{-1}\right|}\Gamma_{N+1-d}\left(\omega\right)\\
&= \int_0^\infty \mathrm{d}\omega\;\mathrm{e}^{-\frac{1}{2}\sum_{\mu=1}^{d}\log\left(1-\frac{\alpha\zeta\sigma_0^2}{\omega(1-\zeta+1/N)\lambda_\mu(\boldsymbol{\Sigma})}\right)}\Gamma_{N+1-d}\left(\omega\right) && \text{(E2)}
\end{aligned}$$

where we encounter the *gamma* distribution, for $\nu > 0$,

$$\Gamma_\nu\left(\omega\right) = \frac{\nu^{\nu/2}}{2^{\nu/2}\Gamma(\nu/2)}\omega^{\frac{\nu-2}{2}}\mathrm{e}^{-\frac{1}{2}\nu\omega} \tag{E3}$$

We note that the above derivation was obtained using the mixture of Gaussians representation of multivariate Student t distribution [32]. Thus the moment generating function is given by

$$\begin{aligned}
\Big\langle \mathrm{e}^{\frac{1}{2}\alpha||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}[\mathscr{D}]||^2}\Big\rangle_{\mathscr{D}} &= \int_0^\infty \frac{\Gamma_{N+1-d}\left(\omega\right)}{\left|\mathbf{I}_d-\alpha\frac{\zeta\sigma_0^2}{\omega(1-\zeta+1/N)}\boldsymbol{\Sigma}^{-1}\right|^{\frac{1}{2}}}\,\mathrm{d}\omega && \text{(E4)}\\
&= \int_0^\infty \mathrm{d}\omega\;\Gamma_{N+1-d}\left(\omega\right)\prod_{\ell=1}^{d}\Big(1-\alpha\frac{\zeta\sigma_0^2}{\omega(1-\zeta+1/N)\lambda_\ell(\boldsymbol{\Sigma})}\Big)^{-\frac{1}{2}}
\end{aligned}$$

and, by the transformation $\alpha = 2\mathrm{i}a$ in the above, we also obtain the characteristic function

$$\Big\langle \mathrm{e}^{\mathrm{i}a||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}||^2}\Big\rangle_{\mathscr{D}} = \int_0^\infty \Gamma_{N+1-d}\left(\omega\right)\mathrm{d}\omega\prod_{\ell=1}^{d}\Big(1-\mathrm{i}a\frac{2\zeta\sigma_0^2}{\omega(1-\zeta+1/N)\lambda_\ell(\boldsymbol{\Sigma})}\Big)^{-\frac{1}{2}}. \tag{E5}$$

We note that the last term in above is the product of characteristic functions of gamma distributions. Each gamma distribution has the same 'shape' parameter 1/2 and different scale parameter $2\zeta\sigma_0^2/\omega(1-\zeta+1/N)\lambda_\ell(\boldsymbol{\Sigma})$.

a. The first two moments of the MSE

Let us now consider derivatives of the moment generating function (E4) upon replacing $\alpha \to \alpha/d$. The derivative with respect to α then gives us

$$\begin{aligned}
2\frac{\partial}{\partial\alpha}\left\langle \mathrm{e}^{\frac{1}{2d}\alpha||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}[\mathscr{D}]||^2}\right\rangle_{\mathscr{D}} &= 2\int_0^\infty \mathrm{d}\omega\ \Gamma_{N+1-d}(\omega)\,\frac{\partial}{\partial\alpha}\prod_{\ell=1}^{d}\big(1-\alpha\frac{\zeta\sigma_0^2}{\omega d(1-\zeta+1/N)\lambda_\ell(\boldsymbol{\Sigma})}\big)^{-\frac{1}{2}} \\
&= \int_0^\infty \mathrm{d}\omega\ \Gamma_{N+1-d}(\omega)\prod_{\ell=1}^{d}\big(1-\alpha\frac{\zeta\sigma_0^2}{\omega d(1-\zeta+1/N)\lambda_\ell(\boldsymbol{\Sigma})}\big)^{-\frac{1}{2}} \\
&\quad\times\frac{1}{d}\sum_{\ell=1}^{d}\big(1-\alpha\frac{\zeta\sigma_0^2}{\omega d(1-\zeta+1/N)\lambda_\ell(\boldsymbol{\Sigma})}\big)^{-1}\frac{\zeta\sigma_0^2}{\omega(1-\zeta+1/N)\lambda_\ell(\boldsymbol{\Sigma})}
\end{aligned}\tag{E6}$$

For $\alpha = 0$ this gives us the average

$$\begin{aligned}
\frac{1}{d}\left\langle||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}[\mathscr{D}]||^2\right\rangle_{\mathscr{D}} &= \frac{\zeta\sigma_0^2}{1-\zeta+1/N}\frac{1}{d}\mathrm{Tr}\left\{\boldsymbol{\Sigma}^{-1}\right\}\int_0^\infty \mathrm{d}\omega\ \Gamma_{N+1-d}(\omega)\,\omega^{-1} \\
&= \frac{\zeta\sigma_0^2}{1-\zeta+1/N}\frac{1-\zeta+1/N}{1-\zeta-1/N}\frac{1}{d}\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\right] = \frac{\zeta\sigma_0^2}{1-\zeta-1/N}\frac{1}{d}\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\right].
\end{aligned}\tag{E7}$$

Now we consider the second derivative with respect to α:

$$\begin{aligned}
4\frac{\partial^2}{\partial\alpha^2}\left\langle \mathrm{e}^{\frac{1}{2d}\alpha||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}[\mathscr{D}]||^2}\right\rangle_{\mathscr{D}} &= 4\int_0^\infty \mathrm{d}\omega\ \Gamma_{N+1-d}(\omega)\,\frac{\partial^2}{\partial\alpha^2}\prod_{\ell=1}^{d}\big(1-\alpha\frac{\zeta\sigma_0^2}{\omega d(1-\zeta+1/N)\lambda_\ell(\boldsymbol{\Sigma})}\big)^{-\frac{1}{2}} \\
&= 2\int_0^\infty \mathrm{d}\omega\ \Gamma_{N+1-d}(\omega)\,\frac{\partial}{\partial\alpha}\prod_{\ell=1}^{d}\big(1-\alpha\frac{\zeta\sigma_0^2}{\omega d(1-\zeta+1/N)\lambda_\ell(\boldsymbol{\Sigma})}\big)^{-\frac{1}{2}} \\
&\quad\times\frac{1}{d}\sum_{\ell=1}^{d}\big(1-\alpha\frac{\zeta\sigma_0^2}{\omega d(1-\zeta+1/N)\lambda_\ell(\boldsymbol{\Sigma})}\big)^{-1}\frac{\zeta\sigma_0^2}{\omega(1-\zeta+1/N)\lambda_\ell(\boldsymbol{\Sigma})} \\
&= \int_0^\infty \mathrm{d}\omega\ \Gamma_{N+1-d}(\omega)\Bigg\{2\frac{\partial}{\partial\alpha}\prod_{\ell=1}^{d}\big(1-\alpha\frac{\zeta\sigma_0^2}{\omega d(1-\zeta+1/N)\lambda_\ell(\boldsymbol{\Sigma})}\big)^{-\frac{1}{2}} \\
&\quad\times\frac{1}{d}\sum_{\ell=1}^{d}\big(1-\alpha\frac{\zeta\sigma_0^2}{\omega d(1-\zeta+1/N)\lambda_\ell(\boldsymbol{\Sigma})}\big)^{-1}\frac{\zeta\sigma_0^2}{\omega(1-\zeta+1/N)\lambda_\ell(\boldsymbol{\Sigma})} \\
&\quad+2\prod_{\ell=1}^{d}\big(1-\alpha\frac{\zeta\sigma_0^2}{\omega d(1-\zeta+1/N)\lambda_\ell(\boldsymbol{\Sigma})}\big)^{-\frac{1}{2}} \\
&\quad\times\frac{\partial}{\partial\alpha}\frac{1}{d}\sum_{\ell=1}^{d}\big(1-\alpha\frac{\zeta\sigma_0^2}{\omega d(1-\zeta+1/N)\lambda_\ell(\boldsymbol{\Sigma})}\big)^{-1}\frac{\zeta\sigma_0^2}{\omega(1-\zeta+1/N)\lambda_\ell(\boldsymbol{\Sigma})}\Bigg\} \\
&= \int_0^\infty \mathrm{d}\omega\ \Gamma_{N+1-d}(\omega)\prod_{\ell=1}^{d}\big(1-\alpha\frac{\zeta\sigma_0^2}{\omega d(1-\zeta+1/N)\lambda_\ell(\boldsymbol{\Sigma})}\big)^{-\frac{1}{2}} \\
&\quad\times\Bigg\{\left[\frac{1}{d}\sum_{\ell=1}^{d}\big(1-\alpha\frac{\zeta\sigma_0^2}{\omega d(1-\zeta+1/N)\lambda_\ell(\boldsymbol{\Sigma})}\big)^{-1}\frac{\zeta\sigma_0^2}{\omega(1-\zeta+1/N)\lambda_\ell(\boldsymbol{\Sigma})}\right]^2 \\
&\quad+\frac{2}{d^2}\sum_{\ell=1}^{d}\big(1-\alpha\frac{\zeta\sigma_0^2}{\omega d(1-\zeta+1/N)\lambda_\ell(\boldsymbol{\Sigma})}\big)^{-2}\left[\frac{\zeta\sigma_0^2}{\omega(1-\zeta+1/N)\lambda_\ell(\boldsymbol{\Sigma})}\right]^2\Bigg\}
\end{aligned}\tag{E8}$$

Evaluation at $\alpha=0$ gives us the second moment

$$\begin{aligned}\left\langle\frac{1}{d^2}||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}[\mathscr{D}]||^4\right\rangle_{\mathscr{D}} &= \big(\frac{\zeta\sigma_0^2}{(1-\zeta+1/N)}\big)^2\int_0^\infty \mathrm{d}\omega\ \Gamma_{N+1-d}\left(\omega\right)\omega^{-2}\left[\big(\frac{1}{d}\sum_{\ell=1}^{d}\frac{1}{\lambda_\ell(\boldsymbol{\Sigma})}\big)^2+\frac{2}{d^2}\sum_{\ell=1}^{d}\big(\frac{1}{\lambda_\ell(\boldsymbol{\Sigma})}\big)^2\right]\\
&= \big(\frac{\zeta\sigma_0^2}{(1-\zeta+1/N)}\big)^2\int_0^\infty \mathrm{d}\omega\ \Gamma_{N+1-d}\left(\omega\right)\omega^{-2}\left[\big(\frac{1}{d}\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\right]\big)^2+\frac{2}{d^2}\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-2}\right]\right]\\
&= \big(\frac{\zeta\sigma_0^2}{(1-\zeta+1/N)}\big)^2\frac{(1-\zeta+1/N)^2}{(1-\zeta-1/N)(1-\zeta-3/N)}\left[\big(\frac{1}{d}\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\right]\big)^2+\frac{2}{d^2}\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-2}\right]\right]\\
&= \frac{\zeta^2\sigma_0^4}{(1-\zeta-1/N)(1-\zeta-3/N)}\left[\big(\frac{1}{d}\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\right]\big)^2+\frac{2}{d^2}\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-2}\right]\right]\end{aligned}\tag{E9}$$

Now upon combining the mean (E7) and the second moment (E9) we obtain the variance of the random variable $\frac{1}{d}||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}[\mathscr{D}]||^2$ for $(N,d)\to\infty$:

$$\begin{aligned}\left\langle\frac{1}{d^2}||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}[\mathscr{D}]||^4\right\rangle_{\mathscr{D}}-\frac{1}{d^2}\left\langle||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}[\mathscr{D}]||^2\right\rangle_{\mathscr{D}}^2 &= \frac{\zeta^2\sigma_0^4}{(1-\zeta-1/N)(1-\zeta-3/N)}\left[\big(\frac{1}{d}\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\right]\big)^2+\frac{2}{d^2}\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-2}\right]\right]\\
&\quad-\frac{\zeta^2\sigma_0^4}{\big(1-\zeta-1/N\big)^2}\big(\frac{1}{d}\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\right]\big)^2\\
=\left[\frac{\zeta^2\sigma_0^4}{(1-\zeta-1/N)(1-\zeta-3/N)}-\frac{\zeta^2\sigma_0^4}{\big(1-\zeta-1/N\big)^2}\right]\frac{1}{d^2}\mathrm{Tr}^2\left[\boldsymbol{\Sigma}^{-1}\right] &+ \frac{\zeta^2\sigma_0^4}{(1-\zeta-1/N)(1-\zeta-3/N)}\frac{2}{d^2}\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-2}\right]\\
&= 2\big(\frac{\zeta\sigma_0^2}{1-\zeta}\big)^2\frac{1}{d^2}\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-2}\right].\end{aligned}\tag{E10}$$

b. Properties of the MGF for large (N,d)

Let us consider the moment generating function (E4) of the MSE for the covariance matrix $\boldsymbol{\Sigma}=\lambda\mathbf{I}_d$. The mean MSE $\langle\frac{1}{d}||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}[\mathscr{D}]||^2\rangle_{\mathscr{D}}$ is given by

$$\mu(\lambda)=\zeta\sigma_0^2/(1-\zeta)\lambda\tag{E11}$$

using equation (E7) for large (N,d). The MGF of the MSE is given by

$$\begin{aligned}\left\langle \mathrm{e}^{\frac{1}{2}\alpha||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}[\mathscr{D}]||^2}\right\rangle_{\mathscr{D}} &= \int_0^\infty\Gamma_{N+1-d}\left(\omega\right)\mathrm{d}\omega\frac{1}{\big(1-\alpha\frac{\zeta\sigma_0^2}{\omega(1-\zeta+1/N)\lambda}\big)^{\frac{d}{2}}}=\int_0^\infty\mathrm{d}\omega\ \Gamma_{N+1-d}\left(\omega\right)\big(\frac{\omega(1-\zeta+1/N)\lambda}{\omega(1-\zeta+1/N)\lambda-\alpha\zeta\sigma_0^2}\big)^{\frac{d}{2}}\\
&= \int_0^\infty\mathrm{d}\omega\ \Gamma_{N+1-d}\left(\omega\right)\Big(\frac{\omega}{\omega-\alpha\frac{\zeta\sigma_0^2}{(1-\zeta+1/N)\lambda}}\Big)^{\frac{d}{2}}=\int_0^\infty\mathrm{d}\omega\ \Gamma_{N+1-d}\left(\omega\right)\big(\frac{\omega}{\omega-\alpha\mu(\lambda)}\big)^{\frac{d}{2}},\end{aligned}\tag{E12}$$

where in the last line we assumed $(N,d)\to\infty$. Let us now consider the integral

$$\begin{aligned}\frac{1}{N}\log\int_0^\infty\mathrm{d}\omega\ \Gamma_{N+1-d}\left(\omega\right)\big(\frac{\omega}{\omega-\alpha}\big)^{\frac{d}{2}} &= \frac{1}{N}\log\frac{(N+1-d)^{(N+1-d)/2}}{2^{(N+1-d)/2}\Gamma((N+1-d)/2)}\\
&\quad+\frac{1}{N}\log\int_0^\infty\mathrm{d}\omega\ \omega^{\frac{N-d-1}{2}}\mathrm{e}^{-\frac{1}{2}(N-d+1)\omega}\big(\frac{\omega}{\omega-\alpha}\big)^{\frac{d}{2}}\\
&= \frac{1}{N}\log\big(\frac{1-\zeta}{2}N\big)^{\frac{1-\zeta}{2}N}/\,\Gamma\big(\frac{1-\zeta}{2}N\big)+\frac{1}{N}\log\int_0^\infty\mathrm{d}\omega\ \mathrm{e}^{N\left[\frac{1}{2}(1-\zeta)\log\omega-\frac{1}{2}(1-\zeta)\,\omega+\frac{\zeta}{2}\log\left(\frac{\omega}{\omega-\alpha}\right)\right]}\\
&= \frac{1-\zeta}{2}+\frac{1}{2N}\log\big(\frac{1-\zeta}{4\pi}N\big)+O\big(N^{-2}\big)+\frac{1}{2}\phi_-(\omega_0^-)+\frac{1}{2N}\log\big(\frac{4\pi}{N(-\phi_-''(\omega_0^-))}\big)+O\big(N^{-2}\big)\\
&= \frac{1-\zeta}{2}+\frac{1}{2}\phi_-(\omega_0^-)+\frac{1}{2N}\log\big(\frac{\zeta-1}{\phi_-''(\omega_0^-)}\big)+O\big(N^{-2}\big),\end{aligned}\tag{E13}$$

where $\omega_0^-=\mathrm{argmax}_{\omega\in(0,\infty)}\phi_-(\omega)$, with the function

$$\phi_-(\omega)=(1-\zeta)\log\omega-(1-\zeta)\,\omega+\zeta\log\big(\frac{\omega}{\omega-\alpha}\big)\tag{E14}$$

We note $\phi_-(\omega)$ has a maximum when the solution of

$$\phi'_-(\omega) = \frac{(1-\zeta)\omega^2 - (\alpha+1)(1-\zeta)\omega + \alpha}{(\alpha-\omega)\omega} = 0 \tag{E15}$$

satisfies the condition $\phi''_-(\omega) > 0$ given by the inequality $\omega^2\zeta - \big(\omega-\alpha\big)^2 < 0$. The latter is satisfied when $\omega \in \big(0,(1-\sqrt{\zeta})\alpha/(1-\zeta)\big)\cup\big((1+\sqrt{\zeta})\alpha/(1-\zeta),\infty\big)$ for $\zeta\in[0,1)$ and when $\omega\in\big(0,\alpha/2\big)$ for $\zeta=1$. However, the difference $\omega-\alpha$ for $\zeta\in[0,1)$ is negative on the interval $\big(0,(1-\sqrt{\zeta})\alpha/(1-\zeta)\big)$, so $\phi_-(\omega)$ is undefined. The same is also true for $\big(0,\alpha/2\big)$ at $\zeta=1$ thus leaving us only with the interval $\big((1+\sqrt{\zeta})\alpha/(1-\zeta),\infty\big)$ with $\zeta\in(0,1)$.

The equation (E15) has real solutions $\frac{1+\alpha}{2}\pm\sqrt{\big(\frac{1+\alpha}{2}\big)^2-\frac{\alpha}{(1-\zeta)}}$ when the inequality $\big(\frac{\alpha+1}{2}\big)^2/\alpha \geq 1/\big(1-\zeta\big)$ is satisfied. This holds when $\alpha\in\big(0,(1+\zeta-2\sqrt{\zeta})/(1-\zeta)\big)\cup\big((1+\zeta+2\sqrt{\zeta})/(1-\zeta),\infty\big)$, but for $\zeta\in(0,1)$ only one solution $\omega_0=\frac{1+\alpha}{2}+\sqrt{\big(\frac{1+\alpha}{2}\big)^2-\frac{\alpha}{(1-\zeta)}}$ belongs to the interval $\big((1+\sqrt{\zeta})\alpha)/(1-\zeta),\infty\big)$, when $\alpha\in\big(0,(1+\zeta-2\sqrt{\zeta})/(1-\zeta)\big)$, i.e. it corresponds to the maximum of $\phi_-(\omega)$.

We note that upon substituting $\alpha\to\alpha\mu(\lambda)$, the factor appearing in the integral (E12), we obtain

$$\omega_0^- = \frac{1+\alpha\mu(\lambda)}{2} + \sqrt{\big(\frac{1+\alpha\mu(\lambda)}{2}\big)^2 - \frac{\alpha\,\mu(\lambda)}{(1-\zeta)}} \tag{E16}$$

for $\alpha\ \in\big(0,\lambda(1+\zeta-2\sqrt{\zeta})/\zeta\sigma_0^2\big)$ and $\zeta\in(0,1)$. Now, using (E13), the moment generating function (E12) is for large N found to become

$$\Big\langle \mathrm{e}^{\frac{1}{2}\alpha||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}[\mathscr{D}]||^2}\Big\rangle_{\mathscr{D}} = \sqrt{\frac{\zeta-1}{\phi''_-(\omega_0^-)}}\,\mathrm{e}^{\frac{1}{2}N\left(1-\zeta+\phi_-(\omega_0^-)\right)+O\left(1/N\right)}. \tag{E17}$$

2. Deviations from the mean

We are interested in the probability of the event $\frac{1}{d}||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}[\mathscr{D}]||^2\notin\big(\mu(\lambda)-\delta,\mu(\lambda)+\delta\big)$. This is given by

$$\begin{aligned}\mathrm{Prob}\left[\frac{1}{d}||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}[\mathscr{D}]||^2\notin\big(\mu(\lambda)-\delta,\mu(\lambda)+\delta\big)\right] &= \mathrm{Prob}\left[||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}[\mathscr{D}]||^2\leq d\big(\mu(\lambda)-\delta\big)\right]\\ &\quad+\mathrm{Prob}\left[||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}[\mathscr{D}]||^2\geq d\big(\mu(\lambda)+\delta\big)\right].\end{aligned} \tag{E18}$$

First, for $\alpha>0$ we consider the probability

$$\begin{aligned}\mathrm{Prob}\left[||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}[\mathscr{D}]||^2\geq d\big(\mu(\lambda)+\delta\big)\right] &= \mathrm{Prob}\left[\mathrm{e}^{\frac{1}{2}\alpha||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}[\mathscr{D}]||^2}\geq \mathrm{e}^{\frac{1}{2}\alpha d\left(\mu(\lambda)+\delta\right)}\right]\leq\Big\langle \mathrm{e}^{\frac{1}{2}\alpha||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}[\mathscr{D}]||^2}\Big\rangle_{\mathscr{D}}\mathrm{e}^{-\frac{1}{2}\alpha d\left(\mu(\lambda)+\delta\right)}\\ &= \sqrt{\frac{\zeta-1}{\phi''_-(\omega_0^-)}}\,\mathrm{e}^{\frac{1}{2}N\left(1-\zeta+\phi_-(\omega_0^-)\right)+O\left(1/N\right)}\mathrm{e}^{-\frac{1}{2}\alpha\zeta N\left(\mu(\lambda)+\delta\right)}\\ &= \sqrt{\frac{\zeta-1}{\phi''_-(\omega_0^-)}}\,\mathrm{e}^{-\frac{1}{2}N\left[\zeta-1-\phi_-(\omega_0^-)+\alpha\zeta\left(\mu(\lambda)+\delta\right)\right]+O\left(1/N\right)}\end{aligned} \tag{E19}$$

From this it follows that for $N\to\infty$ we have

$$-\frac{2}{N}\log\mathrm{Prob}\left[||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}[\mathscr{D}]||^2\geq d\big(\mu(\lambda)+\delta\big)\right]\geq\zeta-1-\phi_-(\omega_0^-)+\alpha\,\zeta\big(\mu(\lambda)+\delta\big)+O\big(1/N\big) \tag{E20}$$

We seek an upper bound with respect to α, but it is not clear how to implement this analytically for any α. However, for small α the function (divided by ζ) appearing in the right-hand side of (E20) has the following Taylor expansion:

$$\big(\mu(\lambda)-1+\delta\big)\alpha+\frac{\zeta\big(\mu(\lambda)-1\big)^2-1}{2(1-\zeta)}\alpha^2+\frac{\big(\mu-1\big)^3\zeta^2+\big(3\,\mu^3-3\,\mu^2-3\,\mu+2\big)\zeta-1}{3\big(1-\zeta\big)^2}\alpha^3+O\big(\alpha^4\big), \tag{E21}$$

so if $\mu(\lambda)+\delta>1$, the first term in this expansion is positive and hence if $\alpha>0$ is sufficiently small then the RHS of (E20) is positive. We note that for $\mu(\lambda)\geq1$, where $\mu(\lambda)=\zeta\sigma_0^2/(1-\zeta)\lambda$, the value of $\delta>0$ can be made arbitrary small, but for $\mu<1$ the positivity of first term in (E21) is dependent on δ.

Second, for $\alpha > 0$ we consider the probability

$$\mathrm{Prob}\Big[||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}[\mathscr{D}]||^2 \le d\big(\mu(\lambda)-\delta\big)\Big] = \mathrm{Prob}\Big[\mathrm{e}^{-\frac{1}{2}\alpha||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}[\mathscr{D}]||^2} \ge \mathrm{e}^{-\frac{1}{2}\alpha d\big(\mu(\lambda)-\delta\big)}\Big] \le \Big\langle \mathrm{e}^{-\frac{1}{2}\alpha||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}[\mathscr{D}]||^2}\Big\rangle_{\mathscr{D}} \mathrm{e}^{\frac{1}{2}\alpha d\big(\mu(\lambda)-\delta\big)}$$
$$= \sqrt{\frac{\zeta-1}{\phi_+''(\omega_0^+)}}\,\mathrm{e}^{\frac{1}{2}N\big(1-\zeta+\phi_+(\omega_0^+)\big)+O\big(1/N\big)}\mathrm{e}^{\frac{1}{2}\alpha\zeta N\big(\mu(\lambda)-\delta\big)}$$
$$= \sqrt{\frac{\zeta-1}{\phi_+''(\omega_0^+)}}\,\mathrm{e}^{-\frac{1}{2}N\big[\zeta-1-\phi_+(\omega_0^+)-\alpha\zeta\big(\mu(\lambda)-\delta\big)\big]+O\big(1/N\big)}, \tag{E22}$$

where the function ϕ_+, defined as

$$\phi_+(\omega) = (1-\zeta)\log\omega - (1-\zeta)\,\omega + \zeta\log\big(\frac{\omega}{\omega+\alpha}\big), \tag{E23}$$

has a maximum at

$$\omega_0^+ = \frac{1}{2}\big(1-\alpha+\sqrt{\left(\alpha-1\right)^2+4\alpha/(1-\zeta)}\big). \tag{E24}$$

Now for $\alpha\to\alpha\mu(\lambda)$ we have in the exponential of (E22):

$$\zeta-1-\phi_+(\omega_0^+)-\alpha\zeta\big(\mu(\lambda)-\delta\big) = \delta\,\zeta\,\alpha - \frac{\mu^2(\lambda)\zeta}{2\big(1-\zeta\big)}\alpha^2 + \frac{\mu^3(\lambda)\zeta\,\left(\zeta+1\right)}{3\big(1-\zeta\big)^2}\alpha^3 - \frac{\mu^4(\lambda)\zeta\,\left(\zeta^2+3\,\zeta+1\right)}{4\big(1-\zeta\big)^3}\alpha^4 + O(\alpha^5). \tag{E25}$$

This is positive for sufficiently small α, and hence the lower bound in the inequality

$$-\frac{2}{N}\log\mathrm{Prob}\Big[||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}[\mathscr{D}]||^2 \le d\big(\mu(\lambda)-\delta\big)\Big] \ge \zeta-1-\phi_+(\omega_0^+)-\alpha\zeta\big(\mu(\lambda)-\delta\big)+O\big(1/N\big) \tag{E26}$$

is positive for any $\delta\in(0,\mu(\lambda))$ and sufficiently small α, when $N\to\infty$.

Now combining (E20) with (E26) allows us bound the probability (E18) as follows

$$\mathrm{Prob}\Big[\frac{1}{d}||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}[\mathscr{D}]||^2 \notin \big(\mu(\lambda)-\delta,\mu(\lambda)+\delta\big)\Big] \le \mathrm{C}_-\mathrm{e}^{-\frac{N}{2}\big[\zeta-1-\phi_-(\omega_0^-)+\alpha\,\zeta\,\big(\mu(\lambda)+\delta\big)\big]}$$
$$+\,\mathrm{C}_+\mathrm{e}^{-\frac{N}{2}\big[\zeta-1-\phi_+(\omega_0^+)-\alpha\zeta\big(\mu(\lambda)-\delta\big)\big]}. \tag{E27}$$

for some *constants* $\mathrm{C}_\pm$ and some sufficiently small $\alpha>0$. We note that for the first term in the above upper bound to vanish, as $N\to\infty$, for arbitrary small δ it is sufficient that $\mu(\lambda)\ge 1$, where $\mu(\lambda)=\zeta\sigma_0^2/(1-\zeta)\lambda$, but for $\mu(\lambda)<1$ the value of δ must be such that $\delta>1-\mu(\lambda)$. The second term in the upper bound is vanishing for any $\delta\in(0,\mu(\lambda))$.

Finally we consider deviations of $\frac{1}{d}||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}[\mathscr{D}]||^2$ from its mean $\mu\big(\boldsymbol{\Sigma}\big)=[\zeta\sigma_0^2/(1-\zeta-N^{-1})]\frac{1}{d}\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\right]$, derived in (E7). To this end we consider the probability of event $\frac{1}{d}||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}[\mathscr{D}]||^2\notin\big(\mu\big(\boldsymbol{\Sigma}\big)-\delta,\mu\big(\boldsymbol{\Sigma}\big)+\delta\big)$ given by the sum

$$\mathrm{Prob}\Big[||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}[\mathscr{D}]||^2 \le d\big(\mu\big(\boldsymbol{\Sigma}\big)-\delta\big)\Big] + \mathrm{Prob}\Big[||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}[\mathscr{D}]||^2 \ge d\big(\mu\big(\boldsymbol{\Sigma}\big)+\delta\big)\Big]$$
$$\le \Big\langle\mathrm{e}^{-\frac{1}{2}\alpha||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}[\mathscr{D}]||^2}\Big\rangle_{\mathscr{D}}\mathrm{e}^{\frac{1}{2}\alpha d\big(\mu\big(\boldsymbol{\Sigma}\big)-\delta\big)} + \Big\langle\mathrm{e}^{\frac{1}{2}\alpha||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}[\mathscr{D}]||^2}\Big\rangle_{\mathscr{D}}\mathrm{e}^{-\frac{1}{2}\alpha d\big(\mu\big(\boldsymbol{\Sigma}\big)+\delta\big)} \tag{E28}$$

for $\alpha>0$. Let us order the eigenvalues of $\boldsymbol{\Sigma}$ in a such a way that $\lambda_1(\boldsymbol{\Sigma})\le\lambda_2(\boldsymbol{\Sigma})\le\cdots\le\lambda_d(\boldsymbol{\Sigma})$ then, using (E4), for the moment generating functions in above we obtain

$$\Big\langle\mathrm{e}^{-\frac{1}{2}\alpha||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}[\mathscr{D}]||^2}\Big\rangle_{\mathscr{D}} = \int_0^\infty \Gamma_{N+1-d}\left(\omega\right)\mathrm{d}\omega\prod_{\ell=1}^{d}\big(1+\alpha\frac{\zeta\sigma_0^2}{\omega(1-\zeta+1/N)\lambda_\ell(\boldsymbol{\Sigma})}\big)^{-\frac{1}{2}}$$
$$\le \int_0^\infty \Gamma_{N+1-d}\left(\omega\right)\mathrm{d}\omega\big(1+\alpha\frac{\zeta\sigma_0^2}{\omega(1-\zeta+1/N)\lambda_1(\boldsymbol{\Sigma})}\big)^{-\frac{d}{2}} \tag{E29}$$

and

$$\Big\langle\mathrm{e}^{\frac{1}{2}\alpha||\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}[\mathscr{D}]||^2}\Big\rangle_{\mathscr{D}} = \int_0^\infty \Gamma_{N+1-d}\left(\omega\right)\mathrm{d}\omega\prod_{\ell=1}^{d}\big(1-\alpha\frac{\zeta\sigma_0^2}{\omega(1-\zeta+1/N)\lambda_\ell(\boldsymbol{\Sigma})}\big)^{-\frac{1}{2}}$$
$$\le \int_0^\infty \Gamma_{N+1-d}\left(\omega\right)\mathrm{d}\omega\big(1-\alpha\frac{\zeta\sigma_0^2}{\omega(1-\zeta+1/N)\lambda_d(\boldsymbol{\Sigma})}\big)^{-\frac{d}{2}}. \tag{E30}$$

Furthermore, by the inequalities $1/\lambda_d(\mathbf{\Sigma}) \le \frac{1}{d}\mathrm{Tr}\left[\mathbf{\Sigma}^{-1}\right] \le 1/\lambda_1(\mathbf{\Sigma})$, the mean obeys $\mu\big(\lambda_d(\mathbf{\Sigma})\big) \le \mu\big(\mathbf{\Sigma}\big) \le \mu\big(\lambda_1(\mathbf{\Sigma})\big)$, where $\mu\big(\lambda\big)=\zeta\sigma_0^2/(1-\zeta-N^{-1})\lambda$. The latter combined with the upper bounds in (E29) and (E30) gives us

$$\mathrm{Prob}\left[\frac{1}{d}\left|\left|\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}[\mathscr{D}]\right|\right|^2 \notin \big(\mu(\mathbf{\Sigma})-\delta, \mu(\mathbf{\Sigma})+\delta\big)\right] \le \int_0^\infty \mathrm{d}\omega\; \Gamma_{N+1-d}\left(\omega\right)\big(1+\frac{\alpha\zeta\sigma_0^2}{\omega(1-\zeta+1/N)\lambda_1(\mathbf{\Sigma})}\big)^{-\frac{d}{2}}\mathrm{e}^{\frac{1}{2}\alpha d\left(\mu\left(\lambda_1(\mathbf{\Sigma})\right)-\delta\right)}$$
$$+\int_0^\infty \mathrm{d}\omega\; \Gamma_{N+1-d}\left(\omega\right)\big(1-\frac{\alpha\zeta\sigma_0^2}{\omega(1-\zeta+1/N)\lambda_d(\mathbf{\Sigma})}\big)^{-\frac{d}{2}}\mathrm{e}^{-\frac{1}{2}\alpha d\left(\mu\left(\lambda_d(\mathbf{\Sigma})\right)+\delta\right)} \tag{E31}$$

Finally, using similar steps to those that led us earlier to (E27) give

$$\mathrm{Prob}\left[\frac{1}{d}\left|\left|\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}[\mathscr{D}]\right|\right|^2 \notin \big(\mu(\mathbf{\Sigma})-\delta, \mu(\mathbf{\Sigma})+\delta\big)\right] \le \mathrm{C}_-\mathrm{e}^{-N\Phi_-[\alpha,\mu(\lambda_d),\delta]}+\mathrm{C}_+\mathrm{e}^{-N\Phi_+[\alpha,\mu(\lambda_1),\delta]}, \tag{E32}$$

for some constants $\mathrm{C}_\pm$ and some sufficiently small $\alpha>0$. In the above we have defined the (rate) functions

$$\Phi_-[\alpha,\mu(\lambda_d),\delta]=\frac{1}{2}\Big[\zeta-1-\phi_-(\omega_0^-)+\alpha\,\zeta\big(\mu(\lambda_d(\mathbf{\Sigma}))+\delta\big)\Big], \tag{E33}$$
$$\Phi_+[\alpha,\mu(\lambda_1),\delta]=\frac{1}{2}\Big[\zeta-1-\phi_+(\omega_0^+)-\alpha\zeta\big(\mu\big(\lambda_1(\mathbf{\Sigma})\big)-\delta\big)\Big]. \tag{E34}$$

Here $\phi_-(\omega_0^-)$ is defined by (E14) and (E16) with $\mu(\lambda)$ replaced by $\mu(\lambda_d)$, and $\phi_+(\omega_0^+)$ is defined by (E23) and (E24) with α replaced by $\alpha\mu(\lambda_1)$. We note that for the first term in the upper bound (E32) to vanish, as $(N,d)\to\infty$, for arbitrary small δ, it is sufficient that $\mu(\lambda_d)\ge 1$, where $\mu(\lambda)=\zeta\sigma_0^2/(1-\zeta)\lambda$, but for $\mu(\lambda_d)<1$ for this to happen the δ must be such that $\delta>1-\mu(\lambda_d)$. The second term in the upper bound is vanishing for any $\delta\in(0,\mu(\lambda))$.

Appendix F: Statistical properties of free energy

In this section we consider statistical properties of the (conditional) free energy

$$F_{\beta,\sigma^2}\left[\mathscr{D}\right]=\frac{d}{2\beta}+\frac{1}{2\sigma^2}\mathbf{t}^{\mathrm{T}}\big(\mathbf{I}_N-\mathbf{Z}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{Z}^T\big)\mathbf{t}-\frac{1}{2\beta}\log\left|2\pi\mathrm{e}\sigma^2\beta^{-1}\mathbf{J}_{\sigma^2\eta}^{-1}\right| \tag{F1}$$

assuming that σ^2 is independent from data $\mathscr{D}$.

1. The average of free energy

Let us consider the average free energy

$$\begin{aligned}\big\langle F_{\beta,\sigma^2}\left[\mathscr{D}\right]\big\rangle_{\mathscr{D}} &= \frac{d}{2\beta}+\frac{1}{2\sigma^2}\left\langle \mathbf{t}^{\mathrm{T}}\big(\mathbf{I}_N-\mathbf{Z}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{Z}^T\big)\mathbf{t}\right\rangle_{\mathscr{D}}-\frac{1}{\beta}\frac{1}{2}\left\langle\log|2\pi\mathrm{e}\sigma^2\beta^{-1}\mathbf{J}_{\sigma^2\eta}^{-1}|\right\rangle_{\mathscr{D}}\\ &= \frac{d}{2\beta}+\frac{1}{2\sigma^2}\left\langle\left\langle \mathbf{t}^{\mathrm{T}}\big(\mathbf{I}_N-\mathbf{Z}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{Z}^T\big)\mathbf{t}\right\rangle_{\boldsymbol{\epsilon}}\right\rangle_{\mathbf{Z}}-\frac{1}{\beta}\frac{1}{2}\left\langle\log|2\pi\mathrm{e}\sigma^2\beta^{-1}\mathbf{J}_{\sigma^2\eta}^{-1}|\right\rangle_{\mathbf{Z}}\end{aligned} \tag{F2}$$

Now, assuming that the noise vector $\boldsymbol{\epsilon}$ has mean $\mathbf{0}$ and covariance $\sigma_0^2\mathbf{I}_N$, we can work out

$$\begin{aligned}\left\langle \mathbf{t}^{\mathrm{T}}\big(\mathbf{I}_N-\mathbf{Z}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{Z}^T\big)\mathbf{t}\right\rangle_{\boldsymbol{\epsilon}} &= \left\langle \mathrm{Tr}\left[\big(\mathbf{I}_N-\mathbf{Z}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{Z}^T\big)\mathbf{t}\mathbf{t}^{\mathrm{T}}\right]\right\rangle_{\boldsymbol{\epsilon}} = \mathrm{Tr}\left[\big(\mathbf{I}_N-\mathbf{Z}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{Z}^T\big)\left\langle\big(\mathbf{Z}\boldsymbol{\theta}_0+\boldsymbol{\epsilon}\big)\big(\mathbf{Z}\boldsymbol{\theta}_0+\boldsymbol{\epsilon}\big)^{\mathrm{T}}\right\rangle_{\boldsymbol{\epsilon}}\right]\\ &= \mathrm{Tr}\left[\big(\mathbf{I}_N-\mathbf{Z}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{Z}^T\big)\big(\mathbf{Z}\boldsymbol{\theta}_0\boldsymbol{\theta}_0^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}+2\mathbf{Z}\boldsymbol{\theta}_0\left\langle\boldsymbol{\epsilon}^{\mathrm{T}}\right\rangle_{\boldsymbol{\epsilon}}+\left\langle\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}\right\rangle_{\boldsymbol{\epsilon}}\big)\right]\\ &= \mathrm{Tr}\left[\big(\mathbf{I}_N-\mathbf{Z}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{Z}^T\big)\big(\mathbf{Z}\boldsymbol{\theta}_0\boldsymbol{\theta}_0^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}+\sigma_0^2\mathbf{I}_N\big)\right]\\ &= \boldsymbol{\theta}_0^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\big(\mathbf{I}_N-\mathbf{Z}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{Z}^T\big)\mathbf{Z}\boldsymbol{\theta}_0+\sigma_0^2\mathrm{Tr}\left[\mathbf{I}_N-\mathbf{Z}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{Z}^T\right]\\ &= \boldsymbol{\theta}_0^{\mathrm{T}}\big(\mathbf{J}-\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{J}\big)\boldsymbol{\theta}_0+\sigma_0^2\big(N-\mathrm{Tr}\left[\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}\right]\big),\end{aligned} \tag{F3}$$

and hence the average free energy is given by

$$\big\langle F_{\beta,\sigma^2}\left[\mathscr{D}\right]\big\rangle_{\mathscr{D}}=\frac{d}{2\beta}+\frac{1}{2\sigma^2}\boldsymbol{\theta}_0^{\mathrm{T}}\left\langle\big(\mathbf{J}-\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{J}\big)\right\rangle_{\mathbf{Z}}\boldsymbol{\theta}_0+\frac{\sigma_0^2}{2\sigma^2}\big(N-\left\langle\mathrm{Tr}\left[\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}\right]\right\rangle_{\mathbf{Z}}\big)-\frac{1}{2\beta}\Big\langle\log\left|2\pi\mathrm{e}\sigma^2\beta^{-1}\mathbf{J}_{\sigma^2\eta}^{-1}\right|\Big\rangle_{\mathbf{Z}}. \tag{F4}$$

2. The variance of free energy

We turn to the variance of the free energy $F_{\beta,\sigma^2}[\mathscr{D}]$. To this end we exploit the free energy equality $F = U - TS$, which gives us $\mathrm{Var}(F) = \mathrm{Var}(U - TS) = \mathrm{Var}(U) - 2T\mathrm{Cov}(U,S) + T^2\mathrm{Var}(S)$. The latter applied to (F1) leads to

$$\mathrm{Var}\big(F_{\beta,\sigma^2}[\mathscr{D}]\big) = \mathrm{Var}\big(E[\mathscr{D}]\big) + T^2\mathrm{Var}\big(\mathrm{S}[\mathscr{D}]\big) - 2T\,\mathrm{Cov}\big(E[\mathscr{D}],\mathrm{S}[\mathscr{D}]\big) \tag{F5}$$

Let us consider the energy variance

$$\begin{aligned}\mathrm{Var}\big(E[\mathscr{D}]\big) &= \mathrm{Var}\Big(\frac{d}{2\beta} + \frac{1}{2\sigma^2}\mathbf{t}^{\mathrm{T}}\big(\mathbf{I}_N - \mathbf{Z}\mathbf{J}^{-1}_{\sigma^2\eta}\mathbf{Z}^T\big)\mathbf{t}\Big) \\ &= \frac{1}{4\sigma^4}\mathrm{Var}\Big(\big(\mathbf{Z}\boldsymbol{\theta}_0 + \boldsymbol{\epsilon}\big)^{\mathrm{T}}\big(\mathbf{I}_N - \mathbf{Z}\mathbf{J}^{-1}_{\sigma^2\eta}\mathbf{Z}^T\big)\big(\mathbf{Z}\boldsymbol{\theta}_0 + \boldsymbol{\epsilon}\big)\Big)\end{aligned} \tag{F6}$$

If we define $\mathbf{v} = \mathbf{Z}\boldsymbol{\theta}_0$ and $\mathbf{A} = (\mathbf{I}_N - \mathbf{Z}\mathbf{J}^{-1}_{\sigma^2\eta}\mathbf{Z}^T)$ then the above is of the form

$$\begin{aligned}\mathrm{Var}\big((\mathbf{v}+\boldsymbol{\epsilon})^{\mathrm{T}}\mathbf{A}(\mathbf{v}+\boldsymbol{\epsilon})\big) &= \mathrm{Var}\big(\mathbf{v}^{\mathrm{T}}\mathbf{A}\mathbf{v} + 2\,\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{A}\mathbf{v} + \boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{A}\boldsymbol{\epsilon}\big) \\ &= \mathrm{Var}\big(\mathbf{v}^{\mathrm{T}}\mathbf{A}\mathbf{v}\big) + 4\,\mathrm{Var}\big(\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{A}\mathbf{v}\big) + \mathrm{Var}\big(\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{A}\boldsymbol{\epsilon}\big) \\ &\quad + 2\left[2\,\mathrm{Cov}\big(\mathbf{v}^{\mathrm{T}}\mathbf{A}\mathbf{v}, \boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{A}\mathbf{v}\big) + \mathrm{Cov}\big(\mathbf{v}^{\mathrm{T}}\mathbf{A}\mathbf{v}, \boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{A}\boldsymbol{\epsilon}\big) + 2\,\mathrm{Cov}\big(\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{A}\mathbf{v}, \boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{A}\boldsymbol{\epsilon}\big)\right].\end{aligned} \tag{F7}$$

In the following we will use the following identities:

$$\mathbf{Z}^{\mathrm{T}}\mathbf{A}\mathbf{Z} = \mathbf{J} - \mathbf{J}\mathbf{J}^{-1}_{\sigma^2\eta}\mathbf{J} \qquad \mathrm{Tr}[\mathbf{A}] = N - \mathrm{Tr}[\mathbf{J}\mathbf{J}^{-1}_{\sigma^2\eta}], \tag{F8}$$

$$\mathbf{Z}^{\mathrm{T}}\mathbf{A}^2\mathbf{Z} = \mathbf{J}\big(\mathbf{I}_d - \mathbf{J}^{-1}_{\sigma^2\eta}\mathbf{J}\big)^2 \qquad \mathrm{Tr}\left[\mathbf{A}^2\right] = N - 2\mathrm{Tr}[\mathbf{J}\mathbf{J}^{-1}_{\sigma^2\eta}] + \mathrm{Tr}[(\mathbf{J}\mathbf{J}^{-1}_{\sigma^2\eta})^2] \tag{F9}$$

We can now compute each term in (F7):

$$\begin{aligned}\mathrm{Var}\big(\mathbf{v}^{\mathrm{T}}\mathbf{A}\mathbf{v}\big) &= \mathrm{Var}\big(\boldsymbol{\theta}_0^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\big(\mathbf{I}_N - \mathbf{Z}\mathbf{J}^{-1}_{\sigma^2\eta}\mathbf{Z}^T\big)\mathbf{Z}\boldsymbol{\theta}_0\big) = \mathrm{Var}\big(\boldsymbol{\theta}_0^{\mathrm{T}}\big(\mathbf{J} - \mathbf{J}\mathbf{J}^{-1}_{\sigma^2\eta}\mathbf{J}\big)\boldsymbol{\theta}_0\big) \\ &= \Big\langle\big(\boldsymbol{\theta}_0^{\mathrm{T}}\big(\mathbf{J} - \mathbf{J}\mathbf{J}^{-1}_{\sigma^2\eta}\mathbf{J}\big)\boldsymbol{\theta}_0\big)^2\Big\rangle_{\mathbf{Z}} - \Big\langle\boldsymbol{\theta}_0^{\mathrm{T}}\big(\mathbf{J} - \mathbf{J}\mathbf{J}^{-1}_{\sigma^2\eta}\mathbf{J}\big)\boldsymbol{\theta}_0\Big\rangle_{\mathbf{Z}}^2\end{aligned} \tag{F10}$$

$$\begin{aligned}\mathrm{Var}\big(\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{A}\mathbf{v}\big) &= \Big\langle\Big\langle\big(\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{A}\mathbf{v}\big)^2\Big\rangle_{\boldsymbol{\epsilon}}\Big\rangle_{\mathbf{Z}} - \big\langle\big\langle\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{A}\mathbf{v}\big\rangle_{\boldsymbol{\epsilon}}\big\rangle_{\mathbf{Z}}^2 = \Big\langle\Big\langle\big(\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{A}\mathbf{v}\big)^2\Big\rangle_{\boldsymbol{\epsilon}}\Big\rangle_{\mathbf{Z}} = \big\langle\mathbf{v}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\left\langle\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}\right\rangle_{\boldsymbol{\epsilon}}\mathbf{A}\mathbf{v}\big\rangle_{\mathbf{Z}} \\ &= \sigma_0^2\left\langle\mathbf{v}^{\mathrm{T}}\mathbf{A}^2\mathbf{v}\right\rangle_{\mathbf{Z}} = \sigma_0^2\Big\langle\boldsymbol{\theta}_0^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\big(\mathbf{I}_N - \mathbf{Z}\mathbf{J}^{-1}_{\sigma^2\eta}\mathbf{Z}^T\big)^2\mathbf{Z}\boldsymbol{\theta}_0\Big\rangle_{\mathbf{Z}} \\ &= \sigma_0^2\boldsymbol{\theta}_0^{\mathrm{T}}\Big\langle\mathbf{J}\big(\mathbf{I}_d - \mathbf{J}^{-1}_{\sigma^2\eta}\mathbf{J}\big)^2\Big\rangle_{\mathbf{Z}}\boldsymbol{\theta}_0\end{aligned} \tag{F11}$$

$$\begin{aligned}\mathrm{Var}\big(\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{A}\boldsymbol{\epsilon}\big) &= \Big\langle\Big\langle\big(\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{A}\boldsymbol{\epsilon}\big)^2\Big\rangle_{\boldsymbol{\epsilon}}\Big\rangle_{\mathbf{Z}} - \big\langle\big\langle\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{A}\boldsymbol{\epsilon}\big\rangle_{\boldsymbol{\epsilon}}\big\rangle_{\mathbf{Z}}^2 = \sigma_0^4\big(\left\langle\mathrm{Tr}^2[\mathbf{A}]\right\rangle_{\mathbf{Z}} + 2\left\langle\mathrm{Tr}[\mathbf{A}^2]\right\rangle_{\mathbf{Z}} - \left\langle\mathrm{Tr}[\mathbf{A}]\right\rangle_{\mathbf{Z}}^2\big) \\ &= \sigma_0^4\Big[\Big\langle\big(N - \mathrm{Tr}[\mathbf{J}\mathbf{J}^{-1}_{\sigma^2\eta}]\big)^2\Big\rangle_{\mathbf{Z}} + 2\Big\langle N - 2\mathrm{Tr}[\mathbf{J}\mathbf{J}^{-1}_{\sigma^2\eta}] + \mathrm{Tr}[\big(\mathbf{J}\mathbf{J}^{-1}_{\sigma^2\eta}\big)^2]\Big\rangle_{\mathbf{Z}} - \Big\langle N - \mathrm{Tr}[\mathbf{J}\mathbf{J}^{-1}_{\sigma^2\eta}]\Big\rangle_{\mathbf{Z}}^2\Big]\end{aligned} \tag{F12}$$

$$\mathrm{Cov}\big(\mathbf{v}^{\mathrm{T}}\mathbf{A}\mathbf{v}, \boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{A}\mathbf{v}\big) = \big\langle\big\langle\mathbf{v}^{\mathrm{T}}\mathbf{A}\mathbf{v}\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{A}\mathbf{v}\big\rangle_{\boldsymbol{\epsilon}}\big\rangle_{\mathbf{Z}} - \big\langle\big\langle\mathbf{v}^{\mathrm{T}}\mathbf{A}\mathbf{v}\big\rangle_{\boldsymbol{\epsilon}}\big\rangle_{\mathbf{Z}}\big\langle\big\langle\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{A}\mathbf{v}\big\rangle_{\boldsymbol{\epsilon}}\big\rangle_{\mathbf{Z}} = 0 \tag{F13}$$

$$\begin{aligned}\mathrm{Cov}\big(\mathbf{v}^{\mathrm{T}}\mathbf{A}\mathbf{v}, \boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{A}\boldsymbol{\epsilon}\big) &= \big\langle\big\langle\mathbf{v}^{\mathrm{T}}\mathbf{A}\mathbf{v}\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{A}\boldsymbol{\epsilon}\big\rangle_{\boldsymbol{\epsilon}}\big\rangle_{\mathbf{Z}} - \big\langle\big\langle\mathbf{v}^{\mathrm{T}}\mathbf{A}\mathbf{v}\big\rangle_{\boldsymbol{\epsilon}}\big\rangle_{\mathbf{Z}}\big\langle\big\langle\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{A}\boldsymbol{\epsilon}\big\rangle_{\boldsymbol{\epsilon}}\big\rangle_{\mathbf{Z}} \\ &= \big\langle\mathbf{v}^{\mathrm{T}}\mathbf{A}\mathbf{v}\mathrm{Tr}\mathbf{A}\left\langle\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}\right\rangle_{\boldsymbol{\epsilon}}\big\rangle_{\mathbf{Z}} - \big\langle\mathbf{v}^{\mathrm{T}}\mathbf{A}\mathbf{v}\big\rangle_{\mathbf{Z}}\big\langle\big\langle\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{A}\boldsymbol{\epsilon}\big\rangle_{\boldsymbol{\epsilon}}\big\rangle_{\mathbf{Z}} \\ &= \sigma_0^2\Big[\Big\langle\boldsymbol{\theta}_0^{\mathrm{T}}\big(\mathbf{J} - \mathbf{J}\mathbf{J}^{-1}_{\sigma^2\eta}\mathbf{J}\big)\boldsymbol{\theta}_0\big(N - \mathrm{Tr}[\mathbf{J}\mathbf{J}^{-1}_{\sigma^2\eta}]\big)\Big\rangle_{\mathbf{Z}} - \Big\langle\boldsymbol{\theta}_0^{\mathrm{T}}\big(\mathbf{J} - \mathbf{J}\mathbf{J}^{-1}_{\sigma^2\eta}\mathbf{J}\big)\boldsymbol{\theta}_0\Big\rangle_{\mathbf{Z}}\Big\langle\big(N - \mathrm{Tr}[\mathbf{J}\mathbf{J}^{-1}_{\sigma^2\eta}]\big)\Big\rangle_{\mathbf{Z}}\Big]\end{aligned} \tag{F14}$$

$$\mathrm{Cov}\big(\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{A}\mathbf{v}, \boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{A}\boldsymbol{\epsilon}\big) = \big\langle\big\langle\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{A}\mathbf{v}\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{A}\boldsymbol{\epsilon}\big\rangle_{\boldsymbol{\epsilon}}\big\rangle_{\mathbf{Z}} - \big\langle\big\langle\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{A}\mathbf{v}\big\rangle_{\boldsymbol{\epsilon}}\big\rangle_{\mathbf{Z}}\big\langle\big\langle\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{A}\boldsymbol{\epsilon}\big\rangle_{\boldsymbol{\epsilon}}\big\rangle_{\mathbf{Z}} = 0 \tag{F15}$$

In deriving the above identities we also used the following result

$$\begin{aligned}\Big\langle\big(\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{A}\boldsymbol{\epsilon}\big)^2\Big\rangle_{\boldsymbol{\epsilon}} &= \sum_{i_1,\ldots,i_4}\left\langle\epsilon_{i_1}\epsilon_{i_2}\epsilon_{i_3}\epsilon_{i_4}\right\rangle_{\boldsymbol{\epsilon}}\mathbf{A}_{i_1i_2}\mathbf{A}_{i_3i_4} = \sigma_0^4\sum_{i_1,\ldots,i_4}\Big[\delta_{i_1,i_2}\delta_{i_3,i_4} + \delta_{i_1,i_3}\delta_{i_2,i_4} + \delta_{i_1,i_4}\delta_{i_2,i_3}\Big]\mathbf{A}_{i_1i_2}\mathbf{A}_{i_3i_4} \\ &= \sigma_0^4\sum_{i_1,i_2}\left\{\mathbf{A}_{i_1i_1}\mathbf{A}_{i_2i_2} + 2\mathbf{A}^2_{i_1i_2}\right\} = \sigma_0^4\big(\mathrm{Tr}^2[\mathbf{A}] + 2\mathrm{Tr}[\mathbf{A}^2]\big).\end{aligned} \tag{F16}$$

Using all of the above results in (F6) we obtain

$$
\begin{aligned}
4\sigma^4\mathrm{Var}\big(E[\mathscr{D}]\big) = & \Big\langle \big(\boldsymbol{\theta}_0^{\mathrm{T}}\big(\mathbf{J}-\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{J}\big)\boldsymbol{\theta}_0\big)^2\Big\rangle_{\mathbf{Z}} - \Big\langle \boldsymbol{\theta}_0^{\mathrm{T}}\big(\mathbf{J}-\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{J}\big)\boldsymbol{\theta}_0\Big\rangle_{\mathbf{Z}}^2 + 4\sigma_0^2\boldsymbol{\theta}_0^{\mathrm{T}}\Big\langle \mathbf{J}\big(\mathbf{I}_d-\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{J}\big)^2\Big\rangle_{\mathbf{Z}}\boldsymbol{\theta}_0 \\
& +\sigma_0^4\Bigg[\Big\langle\big(N-\mathrm{Tr}\left[\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}\right]\big)^2\Big\rangle_{\mathbf{Z}} + 2\Big\langle\big(N-2\mathrm{Tr}\left[\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}\right]+\mathrm{Tr}\left[\big(\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}\big)^2\right]\big)\Big\rangle_{\mathbf{Z}} - \Big\langle\big(N-\mathrm{Tr}\left[\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}\right]\big)\Big\rangle_{\mathbf{Z}}^2\Bigg] \\
& +2\sigma_0^2\Bigg(\Big\langle\boldsymbol{\theta}_0^{\mathrm{T}}\big(\mathbf{J}-\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{J}\big)\boldsymbol{\theta}_0\big(N-\mathrm{Tr}[\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}]\big)\Big\rangle_{\mathbf{Z}} - \Big\langle\boldsymbol{\theta}_0^{\mathrm{T}}\big(\mathbf{J}-\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{J}\big)\boldsymbol{\theta}_0\Big\rangle_{\mathbf{Z}}\Big\langle\big(N-\mathrm{Tr}[\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}]\big)\Big\rangle_{\mathbf{Z}}\Bigg).
\end{aligned}
\tag{F17}
$$

The entropy variance is given by

$$
\mathrm{Var}\big(\mathrm{S}[\mathscr{D}]\big) = \mathrm{Var}\big(\frac{1}{2}\log|2\pi\mathrm{e}\sigma^2\beta^{-1}\mathbf{J}_{\sigma^2\eta}^{-1}|\big) = \frac{1}{4}\Big\langle\log^2\left|\mathbf{J}_{\sigma^2\eta}^{-1}\right|\Big\rangle_{\mathbf{Z}} - \frac{1}{4}\Big\langle\log\left|\mathbf{J}_{\sigma^2\eta}^{-1}\right|\Big\rangle_{\mathbf{Z}}^2 \tag{F18}
$$

and the covariance is

$$
\begin{aligned}
\mathrm{Cov}\big(E[\mathscr{D}],\mathrm{S}[\mathscr{D}]\big) &= \mathrm{Cov}\big(\frac{d}{2\beta}+\frac{1}{2\sigma^2}\mathbf{t}^{\mathrm{T}}\big(\mathbf{I}_N-\mathbf{Z}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{Z}^T\big)\mathbf{t},\frac{d}{2}\log(2\pi\mathrm{e}\sigma^2\beta^{-1})+\frac{1}{2}\log\left|\mathbf{J}_{\sigma^2\eta}^{-1}\right|\big) \\
&= \frac{1}{4\sigma^2}\mathrm{Cov}\Big(\mathbf{t}^{\mathrm{T}}\big(\mathbf{I}_N-\mathbf{Z}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{Z}^T\big)\mathbf{t},\log\left|\mathbf{J}_{\sigma^2\eta}^{-1}\right|\Big) \\
&= \frac{1}{4\sigma^2}\Big\langle\Big\langle\mathbf{t}^{\mathrm{T}}\big(\mathbf{I}_N-\mathbf{Z}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{Z}^T\big)\mathbf{t}\Big\rangle_{\boldsymbol{\epsilon}}\log\left|\mathbf{J}_{\sigma^2\eta}^{-1}\right|\Big\rangle_{\mathbf{Z}} - \frac{1}{4\sigma^2}\Big\langle\Big\langle\mathbf{t}^{\mathrm{T}}\big(\mathbf{I}_N-\mathbf{Z}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{Z}^T\big)\mathbf{t}\Big\rangle_{\boldsymbol{\epsilon}}\Big\rangle_{\mathbf{Z}}\Big\langle\log\left|\mathbf{J}_{\sigma^2\eta}^{-1}\right|\Big\rangle_{\mathbf{Z}} \\
&= \frac{1}{4\sigma^2}\Big\langle\Big[\boldsymbol{\theta}_0^{\mathrm{T}}\big(\mathbf{J}-\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{J}\big)\boldsymbol{\theta}_0+\sigma_0^2\big(N-\mathrm{Tr}\left[\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}\right]\big)\Big]\log\left|\mathbf{J}_{\sigma^2\eta}^{-1}\right|\Big\rangle_{\mathbf{Z}} \\
&\quad - \frac{1}{4\sigma^2}\Big\langle\Big[\boldsymbol{\theta}_0^{\mathrm{T}}\big(\mathbf{J}-\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{J}\big)\boldsymbol{\theta}_0+\sigma_0^2\big(N-\mathrm{Tr}\left[\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}\right]\big)\Big]\Big\rangle_{\mathbf{Z}}\Big\langle\log\left|\mathbf{J}_{\sigma^2\eta}^{-1}\right|\Big\rangle_{\mathbf{Z}}
\end{aligned}
\tag{F19}
$$

Finally, using the results (F17), (F18) and (F19), the variance of free energy follows from equation (F5).

3. Free energy of ML inference

For $\eta = 0$ we simply have $\mathbf{J}_{\sigma^2\eta} = \mathbf{J}$ and the average free energy (F4) is given by

$$
\begin{aligned}
\big\langle F_{\beta,\sigma^2}[\mathscr{D}]\big\rangle_{\mathscr{D}} &= \frac{d}{2\beta}+\frac{1}{2\sigma^2}\boldsymbol{\theta}_0^{\mathrm{T}}\left\langle\big(\mathbf{J}-\mathbf{J}\mathbf{J}^{-1}\mathbf{J}\big)\right\rangle_{\mathbf{Z}}\boldsymbol{\theta}_0+\frac{\sigma_0^2}{2\sigma^2}\big(N-\left\langle\mathrm{Tr}\left[\mathbf{J}\mathbf{J}^{-1}\right]\right\rangle_{\mathbf{Z}}\big)-\frac{1}{2\beta}\left\langle\log|2\pi\mathrm{e}\sigma^2\beta^{-1}\mathbf{J}^{-1}|\right\rangle_{\mathbf{Z}} \\
&= \frac{d}{2\beta}+\frac{\sigma_0^2}{2\sigma^2}\big(N-d\big)-\frac{d}{2\beta}\log(2\pi\mathrm{e}\sigma^2)+\frac{d}{2\beta}\log(\beta)+\frac{1}{2\beta}\left\langle\log|\mathbf{J}|\right\rangle_{\mathbf{Z}}
\end{aligned}
\tag{F20}
$$

Once more we put $\mathbf{Z}\to\mathbf{Z}/\sqrt{d}$ where now $z_i(\mu)=\mathcal{O}(1)$ for all (i,μ), then we obtain the average free energy density

$$
\frac{1}{N}\left\langle F_{\beta,\sigma^2}[\mathscr{D}]\right\rangle_{\mathscr{D}} = \frac{1}{2}\frac{\sigma_0^2}{\sigma^2}\big(1-\zeta\big)+\frac{\zeta}{2\beta}\log\big(\frac{\beta}{2\pi\sigma^2\zeta}\big)+\frac{\zeta}{2\beta}\int\mathrm{d}\lambda\,\left\langle\rho_d(\lambda|\mathbf{Z})\right\rangle_{\mathbf{Z}}\log\lambda \tag{F21}
$$

with the density of eigenvalues of the $d\times d$ empirical covariance matrix $\mathbf{C}=\frac{1}{N}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}$,

$$
\rho_d(\lambda|\mathbf{Z}) = \frac{1}{d}\sum_{\ell=1}^{d}\delta\big(\lambda-\lambda_\ell(\mathbf{C})\big). \tag{F22}
$$

We can similarly compute the variance of (F1) for $\eta=0$ and $\mathbf{Z}\to\mathbf{Z}/\sqrt{d}$. Firstly, we consider the energy variance (F17) which is given by $\mathrm{Var}\big(E[\mathscr{D}]\big)=\frac{\sigma_0^4}{2\sigma^4}\big(N-d\big)$, from which we deduce

$$
\mathrm{Var}\big(\frac{E[\mathscr{D}]}{N}\big) = \frac{\sigma_0^4}{2\sigma^4}\frac{\big(1-\zeta\big)}{N}. \tag{F23}
$$

Secondly, the entropy variance (F18) is given by

$$
\begin{aligned}
\mathrm{Var}\big(\mathrm{S}[\mathscr{D}]\big) &= \frac{1}{4}\left\langle\log^2|\mathbf{J}|\right\rangle_{\mathbf{Z}}-\frac{1}{4}\left\langle\log|\mathbf{J}|\right\rangle_{\mathbf{Z}}^2 = \frac{d^2}{4}\Big[\Big\langle\big(\frac{1}{d}\sum_{\ell=1}^{d}\log\lambda_\ell\big)^2\Big\rangle_{\mathbf{Z}}-\Big\langle\frac{1}{d}\sum_{\ell=1}^{d}\log\lambda_\ell\Big\rangle_{\mathbf{Z}}^2\Big] \\
&= \frac{d^2}{4}\Big[\Big\langle\Big(\int\mathrm{d}\lambda\,\rho_d(\lambda|\mathbf{Z})\log\lambda\Big)^2\Big\rangle_{\mathbf{Z}}-\Big\langle\int\mathrm{d}\lambda\,\rho_d(\lambda|\mathbf{Z})\log\lambda\Big\rangle_{\mathbf{Z}}^2\Big]
\end{aligned}
\tag{F24}
$$

and hence

$$\mathrm{Var}\big(\frac{\mathrm{S}[\mathscr{D}]}{N}\big)=\frac{\zeta^2}{4}\Big[\Big\langle\big(\int\!\mathrm{d}\lambda\,\rho_d(\lambda|\mathbf{Z})\log\lambda\big)^2\Big\rangle_{\mathbf{Z}}-\Big\langle\int\!\mathrm{d}\lambda\,\rho_d(\lambda|\mathbf{Z})\log\lambda\Big\rangle_{\mathbf{Z}}^2\Big] \tag{F25}$$

Finally, we consider the covariance (F19) which gives us

$$\mathrm{Cov}\big(E[\mathscr{D}],\mathrm{S}[\mathscr{D}]\big)=\frac{\sigma_0^2}{4\sigma^2}(N-d)\left\langle\log\left|\mathbf{J}^{-1}_{\sigma^2\eta}\right|\right\rangle_{\mathbf{Z}}-\frac{\sigma_0^2}{4\sigma^2}(N-d)\left\langle\log\left|\mathbf{J}^{-1}_{\sigma^2\eta}\right|\right\rangle_{\mathbf{Z}}=0 \tag{F26}$$

Using above results in identity (F5) we obtain the desired variance of the free energy density:

$$\begin{aligned}\mathrm{Var}\big(\frac{F_{\beta,\sigma^2}[\mathscr{D}]}{N}\big)&=\mathrm{Var}\big(\frac{E[\mathscr{D}]}{N}\big)+T^2\mathrm{Var}\big(\frac{\mathrm{S}[\mathscr{D}]}{N}\big)\\&=\frac{\sigma_0^4}{2\sigma^4}\frac{1-\zeta}{N}+\frac{\zeta^2}{4\beta^2}\int\!\mathrm{d}\lambda\mathrm{d}\tilde{\lambda}\,\Big[\Big\langle\rho_d(\lambda|\mathbf{Z})\,\rho_d(\tilde{\lambda}|\mathbf{Z})\Big\rangle_{\mathbf{Z}}-\langle\rho_d(\lambda|\mathbf{Z})\rangle_{\mathbf{Z}}\,\langle\rho_d(\tilde{\lambda}|\mathbf{Z})\rangle_{\mathbf{Z}}\Big]\log(\lambda)\log(\tilde{\lambda})\end{aligned} \tag{F27}$$

4. Free energy of MAP inference

Let us assume that the true parameters $\boldsymbol{\theta}_0$ are drawn randomly, with mean $\mathbf{0}$ and covariance matrix $S^2\mathbf{I}_d$. We may then compute the average of (F4):

$$\begin{aligned}\big\langle\big\langle F_{\beta,\sigma^2}[\mathscr{D}]\big\rangle_{\mathscr{D}}\big\rangle_{\boldsymbol{\theta}_0}&=\frac{d}{2\beta}+\frac{1}{2\sigma^2}\left\langle\left\langle\boldsymbol{\theta}_0^{\mathrm{T}}(\mathbf{J}-\mathbf{J}\mathbf{J}^{-1}_{\sigma^2\eta}\mathbf{J})\boldsymbol{\theta}_0\right\rangle_{\boldsymbol{\theta}_0}\right\rangle_{\mathbf{Z}}+\frac{\sigma_0^2}{2\sigma^2}\big(N-\left\langle\mathrm{Tr}\left[\mathbf{J}\mathbf{J}^{-1}_{\sigma^2\eta}\right]\right\rangle_{\mathbf{Z}}\big)-\frac{1}{2\beta}\left\langle\log|2\pi\mathrm{e}\sigma^2\beta^{-1}\mathbf{J}^{-1}_{\sigma^2\eta}|\right\rangle_{\mathbf{Z}}\\&=\frac{d}{2\beta}+\frac{1}{2\sigma^2}\left\langle\mathrm{Tr}\left[(\mathbf{J}-\mathbf{J}\mathbf{J}^{-1}_{\sigma^2\eta}\mathbf{J})\left\langle\boldsymbol{\theta}_0\boldsymbol{\theta}_0^{\mathrm{T}}\right\rangle_{\boldsymbol{\theta}_0}\right]\right\rangle_{\mathbf{Z}}+\frac{\sigma_0^2}{2\sigma^2}\big(N-\left\langle\mathrm{Tr}\left[\mathbf{J}\mathbf{J}^{-1}_{\sigma^2\eta}\right]\right\rangle_{\mathbf{Z}}\big)-\frac{1}{2\beta}\left\langle\log|2\pi\mathrm{e}\sigma^2\beta^{-1}\mathbf{J}^{-1}_{\sigma^2\eta}|\right\rangle_{\mathbf{Z}}\\&=\frac{d}{2\beta}+\frac{S^2}{2\sigma^2}\left\langle\mathrm{Tr}\left[(\mathbf{J}-\mathbf{J}\mathbf{J}^{-1}_{\sigma^2\eta}\mathbf{J})\right]\right\rangle_{\mathbf{Z}}+\frac{\sigma_0^2}{2\sigma^2}\big(N-\left\langle\mathrm{Tr}\left[\mathbf{J}\mathbf{J}^{-1}_{\sigma^2\eta}\right]\right\rangle_{\mathbf{Z}}\big)-\frac{1}{2\beta}\left\langle\log|2\pi\mathrm{e}\sigma^2\beta^{-1}\mathbf{J}^{-1}_{\sigma^2\eta}|\right\rangle_{\mathbf{Z}}\\&=\frac{d}{2\beta}+\frac{S^2}{2\zeta\sigma^2}\left\langle\mathrm{Tr}\left[(\mathbf{C}-\mathbf{C}\mathbf{C}^{-1}_{\zeta\sigma^2\eta}\mathbf{C})\right]\right\rangle_{\mathbf{Z}}+\frac{\sigma_0^2}{2\sigma^2}\big(N-\left\langle\mathrm{Tr}\left[\mathbf{C}\mathbf{C}^{-1}_{\zeta\sigma^2\eta}\right]\right\rangle_{\mathbf{Z}}\big)-\frac{1}{2\beta}\left\langle\log|2\pi\mathrm{e}\sigma^2\beta^{-1}\zeta\mathbf{C}^{-1}_{\zeta\sigma^2\eta}|\right\rangle_{\mathbf{Z}}\end{aligned} \tag{F28}$$

In the last line we have set $\mathbf{Z}\to\mathbf{Z}/\sqrt{d}$, with $z_i(\mu)=\mathcal{O}(1)$ for all (i,μ), and used $\mathbf{J}=\mathbf{C}/\zeta$ and $\mathbf{C}^{-1}_{\sigma^2\eta}=\zeta\mathbf{C}^{-1}_{\zeta\sigma^2\eta}$, where $\mathbf{C}=\mathbf{Z}^{\mathrm{T}}\mathbf{Z}/N$. Let us consider the matrix product $\mathbf{C}\mathbf{C}_\eta^{-1}=\mathbf{C}(\mathbf{C}+\eta\mathbf{I}_d)^{-1}=((\mathbf{C}+\eta\mathbf{I}_d)\mathbf{C}^{-1})^{-1}=(\mathbf{I}_d+\eta\mathbf{C}^{-1})^{-1}$, giving

$$\begin{aligned}\mathrm{Tr}\left[\mathbf{C}\mathbf{C}_\eta^{-1}\right]&=\mathrm{Tr}\left[\big(\mathbf{I}_d+\eta\mathbf{C}^{-1}\big)^{-1}\right]=\sum_{\ell=1}^d 1/\lambda_\ell\big(\mathbf{I}_d+\eta\mathbf{C}^{-1}\big)=\sum_{\ell=1}^d 1/\big(1+\eta\lambda_\ell(\mathbf{C}^{-1})\big)\\&=\sum_{\ell=1}^d 1/\big(1+\eta\lambda_\ell^{-1}(\mathbf{C})\big)=\sum_{\ell=1}^d\lambda_\ell(\mathbf{C})/\big(\lambda_\ell(\mathbf{C})+\eta\big).\end{aligned} \tag{F29}$$

Similarly we can write

$$\mathbf{C}^2\mathbf{C}_\eta^{-1}=\mathbf{C}^2\big(\mathbf{C}+\eta\mathbf{I}_d\big)^{-1}=\mathbf{C}\big(\mathbf{I}_d+\eta\mathbf{C}^{-1}\big)^{-1}=\big((\mathbf{I}_d+\eta\mathbf{C}^{-1})\mathbf{C}^{-1}\big)^{-1}=\big(\mathbf{C}^{-1}+\eta\mathbf{C}^{-2}\big)^{-1} \tag{F30}$$

The matrices $\mathbf{C}^2$ and $\big(\mathbf{C}+\eta\mathbf{I}_d\big)^{-1}$ obviously commute, so

$$\mathrm{Tr}\left[\mathbf{C}\mathbf{C}_\eta^{-1}\mathbf{C}\right]=\mathrm{Tr}\left[\mathbf{C}^2\mathbf{C}_\eta^{-1}\right]=\sum_{\ell=1}^d\lambda_\ell(\mathbf{C}^2)/\lambda_\ell(\mathbf{C}_\eta)=\sum_{\ell=1}^d\lambda_\ell^2(\mathbf{C})/\lambda_\ell(\mathbf{C}_\eta). \tag{F31}$$

Now $\mathbf{C}_\eta=\mathbf{C}+\eta\mathbf{I}_d=\mathbf{C}(\mathbf{I}_d+\eta\mathbf{C}^{-1})=(\mathbf{I}_d+\eta\mathbf{C}^{-1})\mathbf{C}$, and hence

$$\lambda_\ell(\mathbf{C}_\eta)=\lambda_\ell(\mathbf{C})\lambda_\ell\big(\mathbf{I}_d+\eta\mathbf{C}^{-1}\big)=\lambda_\ell(\mathbf{C})\big(1+\eta\lambda_\ell(\mathbf{C}^{-1})\big)=\lambda_\ell(\mathbf{C})+\eta. \tag{F32}$$

Thus $\mathrm{Tr}\left[\mathbf{C}\mathbf{C}_\eta^{-1}\mathbf{C}\right]=\sum_{\ell=1}^d\lambda_\ell^2(\mathbf{C})/\big(\lambda_\ell(\mathbf{C})+\eta\big)$. Finally, we consider the inverse

$$\mathbf{C}_\eta^{-1}=\big(\mathbf{C}+\eta\mathbf{I}_d\big)^{-1}=\big(\mathbf{I}_d+\eta\mathbf{C}^{-1}\big)^{-1}\mathbf{C}^{-1}=\mathbf{C}^{-1}\big(\mathbf{I}_d+\eta\mathbf{C}^{-1}\big)^{-1}, \tag{F33}$$

The matrices $\mathbf{C}^{-1}$ and $\left(\mathbf{I}_d+\eta\mathbf{C}^{-1}\right)^{-1}$ obviously commute, and hence the ℓ-th eigenvalue of $\mathbf{C}_\eta^{-1}$ is given by

$$\lambda_\ell\big(\mathbf{J}_\eta^{-1}\big)=1/(\lambda_\ell(\mathbf{J})+\eta). \tag{F34}$$

Using the above results on the relevant matrices in (F28) allows us to compute the average free energy:

$$\begin{aligned}
\left\langle\left\langle\frac{1}{N}F_{\beta,\sigma^2}\left[\mathscr{D}\right]\right\rangle_{\mathscr{D}}\right\rangle_{\boldsymbol{\theta}_0} &= \frac{\zeta}{2\beta}+\frac{S^2}{2\zeta\sigma^2}\frac{1}{N}\left\langle\mathrm{Tr}\left[\left(\mathbf{J}-\mathbf{J}\mathbf{J}_{\zeta\sigma^2\eta}^{-1}\mathbf{J}\right)\right]\right\rangle_{\mathbf{Z}}+\frac{\sigma_0^2}{2\sigma^2}(1-\frac{1}{N}\left\langle\mathrm{Tr}\left[\mathbf{J}\mathbf{J}_{\zeta\sigma^2\eta}^{-1}\right]\right\rangle_{\mathbf{Z}})\\
&\quad-\frac{1}{2\beta}\frac{1}{N}\left\langle\log|2\pi\mathrm{e}\sigma^2\beta^{-1}\zeta\mathbf{J}_{\zeta\sigma^2\eta}^{-1}|\right\rangle_{\mathbf{Z}}\\
&=\frac{\zeta}{2\beta}+\frac{S^2}{2\zeta\sigma^2}\frac{d}{Nd}\sum_{\ell=1}^{d}\left\langle\lambda_\ell(\mathbf{J})-\frac{\lambda_\ell^2(\mathbf{J})}{\lambda_\ell(\mathbf{J})+\zeta\sigma^2\eta}\right\rangle_{\mathbf{Z}}+\frac{\sigma_0^2}{2\sigma^2}(1-\frac{d}{Nd}\left\langle\sum_{\ell=1}^{d}\frac{\lambda_\ell(\mathbf{J})}{\lambda_\ell(\mathbf{J})+\zeta\sigma^2\eta}\right\rangle_{\mathbf{Z}})\\
&\quad-\frac{\zeta}{2\beta}\log\left(2\pi\mathrm{e}\sigma^2\beta^{-1}\zeta\right)+\frac{\zeta}{2\beta}\frac{1}{d}\left\langle\sum_{\ell=1}^{d}\log\left(\lambda_\ell(\mathbf{J})+\zeta\sigma^2\eta\right)\right\rangle_{\mathbf{Z}}\\
&=\frac{\zeta}{2\beta}+\frac{S^2\zeta\eta}{2}\int\mathrm{d}\lambda\;\langle\rho_d(\lambda|\mathbf{Z})\rangle_{\mathbf{Z}}\frac{\lambda}{\lambda+\zeta\sigma^2\eta}+\frac{\sigma_0^2}{2\sigma^2}\Big(1-\zeta\int\mathrm{d}\lambda\;\langle\rho_d(\lambda|\mathbf{Z})\rangle_{\mathbf{Z}}\frac{\lambda}{\lambda+\zeta\sigma^2\eta}\Big)\\
&\quad-\frac{\zeta}{2\beta}\log\left(2\pi\mathrm{e}\sigma^2\beta^{-1}\zeta\right)+\frac{\zeta}{2\beta}\int\mathrm{d}\lambda\;\langle\rho_d(\lambda|\mathbf{Z})\rangle_{\mathbf{Z}}\log\left(\lambda+\zeta\sigma^2\eta\right)
\end{aligned} \tag{F35}$$

and hence the average free energy density is given by

$$\begin{aligned}
\left\langle\left\langle\frac{1}{N}F_{\beta,\sigma^2}\left[\mathscr{D}\right]\right\rangle_{\mathscr{D}}\right\rangle_{\boldsymbol{\theta}_0} &= \frac{\zeta}{2\beta}+\frac{S^2\zeta\eta}{2}\int\mathrm{d}\lambda\;\langle\rho_d(\lambda|\mathbf{Z})\rangle_{\mathbf{Z}}\frac{\lambda}{\lambda+\zeta\sigma^2\eta}+\frac{\sigma_0^2}{2\sigma^2}\Big(1-\zeta\int\mathrm{d}\lambda\;\langle\rho_d(\lambda|\mathbf{Z})\rangle_{\mathbf{Z}}\frac{\lambda}{\lambda+\zeta\sigma^2\eta}\Big)\\
&\quad-\frac{\zeta}{2\beta}\log\left(2\pi\mathrm{e}\sigma^2\beta^{-1}\zeta\right)+\frac{\zeta}{2\beta}\int\mathrm{d}\lambda\;\langle\rho_d(\lambda|\mathbf{Z})\rangle_{\mathbf{Z}}\log\left(\lambda+\zeta\sigma^2\eta\right)
\end{aligned} \tag{F36}$$

In the derivation of the above results we used the following simple eigenvalue identities

$$\lambda_\ell(\mathbf{C}_\eta^{-1})=\frac{1}{\lambda_\ell(\mathbf{C})+\eta},\qquad \lambda_\ell(\mathbf{C}\mathbf{C}_\eta^{-1})=\frac{\lambda_\ell(\mathbf{C})}{\lambda_\ell(\mathbf{C})+\eta},\qquad \lambda_\ell(\mathbf{C}-\mathbf{C}\mathbf{C}_\eta^{-1}\mathbf{C})=\frac{\eta\lambda_\ell(\mathbf{C})}{\lambda_\ell(\mathbf{C})+\eta}, \tag{F37}$$

from one also obtains

$$\mathrm{Tr}\left[\mathbf{J}-\mathbf{J}\mathbf{J}_\eta^{-1}\mathbf{J}\right]=d\eta\int\mathrm{d}\lambda\;\rho_d(\lambda|\mathbf{Z})\frac{\lambda}{\lambda+\eta},\qquad \mathrm{Tr}\left[\mathbf{J}\mathbf{J}_\eta^{-1}\right]=d\int\mathrm{d}\lambda\;\rho_d(\lambda|\mathbf{Z})\frac{\lambda}{\lambda+\eta} \tag{F38}$$

Finally, we compute the variance of the free energy (F1). First, we consider the energy variance (F17) which is

$$\begin{aligned}
4\sigma^4\mathrm{Var}\big(E[\mathscr{D}]\big) &= \left\langle\left\langle\left(\boldsymbol{\theta}_0^{\mathrm{T}}(\mathbf{J}-\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{J})\boldsymbol{\theta}_0\right)^2\right\rangle_{\mathbf{Z}}\right\rangle_{\boldsymbol{\theta}_0}-\left\langle\left\langle\boldsymbol{\theta}_0^{\mathrm{T}}(\mathbf{J}-\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{J})\boldsymbol{\theta}_0\right\rangle_{\boldsymbol{\theta}_0}\right\rangle_{\mathbf{Z}}^2+4\left\langle\sigma_0^2\boldsymbol{\theta}_0^{\mathrm{T}}\left\langle\mathbf{J}\left(\mathbf{I}_d-\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{J}\right)^2\right\rangle_{\mathbf{Z}}\boldsymbol{\theta}_0\right\rangle_{\boldsymbol{\theta}_0}\\
&\quad+\sigma_0^4\left[\left\langle\left(N-\mathrm{Tr}\left[\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}\right]\right)^2\right\rangle_{\mathbf{Z}}+2\left\langle\left(N-2\mathrm{Tr}\left[\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}\right]+\mathrm{Tr}\left[(\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1})^2\right]\right)\right\rangle_{\mathbf{Z}}-\left\langle\left(N-\mathrm{Tr}\left[\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}\right]\right)\right\rangle_{\mathbf{Z}}^2\right]\\
&\quad+2\sigma_0^2\left[\left\langle\left\langle\boldsymbol{\theta}_0^{\mathrm{T}}(\mathbf{J}-\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{J})\boldsymbol{\theta}_0\big(N-\mathrm{Tr}[\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}]\big)\right\rangle_{\boldsymbol{\theta}_0}\right\rangle_{\mathbf{Z}}-\left\langle\left\langle\boldsymbol{\theta}_0^{\mathrm{T}}(\mathbf{J}-\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{J})\boldsymbol{\theta}_0\right\rangle_{\boldsymbol{\theta}_0}\right\rangle_{\mathbf{Z}}\left\langle\big(N-\mathrm{Tr}[\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}]\big)\right\rangle_{\mathbf{Z}}\right]\\
&=S^4\left\langle\mathrm{Tr}^2\left[(\mathbf{J}-\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{J})\right]+2\mathrm{Tr}\left[(\mathbf{J}-\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{J})^2\right]\right\rangle_{\mathbf{Z}}-S^4\left\langle\mathrm{Tr}\left[\mathbf{J}-\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{J}\right]\right\rangle_{\mathbf{Z}}^2+4\sigma_0^2S^2\left\langle\mathrm{Tr}\left[\mathbf{J}\left(\mathbf{I}_d-\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{J}\right)^2\right]\right\rangle_{\mathbf{Z}}\\
&\quad+\sigma_0^4\left[\left\langle\left(N-\mathrm{Tr}\left[\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}\right]\right)^2\right\rangle_{\mathbf{Z}}+2\left\langle\left(N-2\mathrm{Tr}\left[\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}\right]+\mathrm{Tr}\left[(\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1})^2\right]\right)\right\rangle_{\mathbf{Z}}-\left\langle\left(N-\mathrm{Tr}\left[\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}\right]\right)\right\rangle_{\mathbf{Z}}^2\right]\\
&\quad+2\sigma_0^2S^2\Bigg(\left\langle\mathrm{Tr}\left[\mathbf{J}-\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{J}\right]\big(N-\mathrm{Tr}[\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}]\big)\right\rangle_{\mathbf{Z}}-\left\langle\mathrm{Tr}\left[\mathbf{J}-\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}\mathbf{J}\right]\right\rangle_{\mathbf{Z}}\left\langle\big(N-\mathrm{Tr}[\mathbf{J}\mathbf{J}_{\sigma^2\eta}^{-1}]\big)\right\rangle_{\mathbf{Z}}\Bigg)
\end{aligned} \tag{F39}$$

where we have used (F16). Hence for $\mathbf{Z}\to\mathbf{Z}/\sqrt{d}$ with $z_i(\mu)=\mathcal{O}(1)$, we have

$$
\begin{aligned}
4\sigma^4\mathrm{Var}\big(\frac{E[\mathscr{D}]}{N}\big) &= \frac{S^4}{\zeta^2N^2}\Big\langle \mathrm{Tr}^2\Big[\big(\mathbf{C}-\mathbf{C}\mathbf{C}^{-1}_{\zeta\sigma^2\eta}\mathbf{C}\big)\Big]+2\mathrm{Tr}\Big[\big(\mathbf{C}-\mathbf{C}\mathbf{C}^{-1}_{\zeta\sigma^2\eta}\mathbf{C}\big)^2\Big]\Big\rangle_{\mathbf{Z}}-\frac{S^4}{\zeta^2N^2}\Big\langle \mathrm{Tr}\Big[\mathbf{C}-\mathbf{C}\mathbf{C}^{-1}_{\zeta\sigma^2\eta}\mathbf{C}\Big]\Big\rangle^2_{\mathbf{Z}}\\
&\quad+\frac{4\sigma_0^2S^2}{\zeta N^2}\Big\langle \mathrm{Tr}\Big[\mathbf{C}\big(\mathbf{I}_d-\mathbf{C}^{-1}_{\zeta\sigma^2\eta}\mathbf{C}\big)^2\Big]\Big\rangle_{\mathbf{Z}}\\
&+\frac{\sigma_0^4}{N^2}\Bigg[\Big\langle\big(N-\mathrm{Tr}\Big[\mathbf{C}\mathbf{C}^{-1}_{\zeta\sigma^2\eta}\Big]\big)^2\Big\rangle_{\mathbf{Z}}+2\Big\langle\big(N-2\mathrm{Tr}\Big[\mathbf{C}\mathbf{C}^{-1}_{\zeta\sigma^2\eta}\Big]+\mathrm{Tr}\Big[\big(\mathbf{C}\mathbf{C}^{-1}_{\zeta\sigma^2\eta}\big)^2\Big]\big)\Big\rangle_{\mathbf{Z}}-\Big\langle\big(N-\mathrm{Tr}\Big[\mathbf{C}\mathbf{C}^{-1}_{\zeta\sigma^2\eta}\Big]\big)\Big\rangle^2_{\mathbf{Z}}\Bigg]\\
&+2\sigma_0^2S^2\frac{1}{\zeta N^2}\Bigg[\Big\langle\mathrm{Tr}\Big[\mathbf{C}-\mathbf{C}\mathbf{C}^{-1}_{\zeta\sigma^2\eta}\mathbf{C}\Big]\big(N-\mathrm{Tr}[\mathbf{C}\mathbf{C}^{-1}_{\zeta\sigma^2\eta}]\big)\Big\rangle_{\mathbf{Z}}-\Big\langle\mathrm{Tr}\Big[\mathbf{C}-\mathbf{C}\mathbf{C}^{-1}_{\zeta\sigma^2\eta}\mathbf{C}\Big]\Big\rangle_{\mathbf{Z}}\Big\langle\big(N-\mathrm{Tr}[\mathbf{C}\mathbf{C}^{-1}_{\zeta\sigma^2\eta}]\big)\Big\rangle_{\mathbf{Z}}\Bigg]\\
&=\frac{S^4}{\zeta^2N^2}\Bigg\langle\Big[d\zeta\sigma^2\eta\int\mathrm{d}\lambda\ \frac{\rho_d(\lambda|\mathbf{Z})\lambda}{\lambda+\zeta\sigma^2\eta}\Big]^2+2d(\zeta\sigma^2\eta)^2\int\mathrm{d}\lambda\ \rho_d(\lambda|\mathbf{Z})\Big(\frac{\lambda}{\lambda+\zeta\sigma^2\eta}\Big)^2\Bigg\rangle_{\mathbf{Z}}\\
&\quad-\frac{S^4}{\zeta^2N^2}\Big\langle d\zeta\sigma^2\eta\int\mathrm{d}\lambda\ \frac{\rho_d(\lambda|\mathbf{Z})\lambda}{\lambda+\zeta\sigma^2\eta}\Big\rangle^2_{\mathbf{Z}}+\frac{4\sigma_0^2S^2}{\zeta N^2}\Big\langle d(\zeta\sigma^2\eta)^2\int\mathrm{d}\lambda\ \frac{\rho_d(\lambda|\mathbf{Z})\lambda}{(\lambda+\zeta\sigma^2\eta)^2}\Big\rangle_{\mathbf{Z}}\\
&\quad+\sigma_0^4\Bigg[\Big\langle\Big(1-\frac{d}{N}\int\mathrm{d}\lambda\ \frac{\rho_d(\lambda|\mathbf{Z})\lambda}{\lambda+\zeta\sigma^2\eta}\Big)^2\Big\rangle_{\mathbf{Z}}-\Big\langle 1-\frac{d}{N}\int\mathrm{d}\lambda\ \frac{\rho_d(\lambda|\mathbf{Z})\lambda}{\lambda+\zeta\sigma^2\eta}\Big\rangle^2_{\mathbf{Z}}\\
&\qquad+\frac{2}{N^2}\Big\langle N-2d\int\mathrm{d}\lambda\ \frac{\rho_d(\lambda|\mathbf{Z})\lambda}{\lambda+\zeta\sigma^2\eta}+d\int\mathrm{d}\lambda\ \rho_d(\lambda|\mathbf{Z})\big(\frac{\lambda}{\lambda+\zeta\sigma^2\eta}\big)^2\Big\rangle_{\mathbf{Z}}\Bigg]\\
&\quad+2\sigma_0^2S^2\frac{1}{\zeta}\Bigg[\Big\langle\Big(\frac{d}{N}\zeta\sigma^2\eta\int\mathrm{d}\lambda\ \frac{\rho_d(\lambda|\mathbf{Z})\lambda}{\lambda+\zeta\sigma^2\eta}\Big)\Big(1-\frac{d}{N}\int\mathrm{d}\lambda\ \frac{\rho_d(\lambda|\mathbf{Z})\lambda}{\lambda+\zeta\sigma^2\eta}\Big)\Big\rangle_{\mathbf{Z}}\\
&\qquad-\Big\langle\frac{d}{N}\zeta\sigma^2\eta\int\mathrm{d}\lambda\ \frac{\rho_d(\lambda|\mathbf{Z})\lambda}{\lambda+\zeta\sigma^2\eta}\Big\rangle_{\mathbf{Z}}\Big\langle 1-\frac{d}{N}\int\mathrm{d}\lambda\ \frac{\rho_d(\lambda|\mathbf{Z})\lambda}{\lambda+\zeta\sigma^2\eta}\Big\rangle_{\mathbf{Z}}\Bigg]\\
&=\zeta^2\big(S^4\sigma^4\eta^2+\sigma_0^4-2\sigma_0^2S^2\sigma^2\eta\big)\Bigg[\Big\langle\Big(\int\mathrm{d}\lambda\ \frac{\rho_d(\lambda|\mathbf{Z})\lambda}{\lambda+\zeta\sigma^2\eta}\Big)^2\Big\rangle_{\mathbf{Z}}-\Big\langle\int\mathrm{d}\lambda\ \frac{\rho_d(\lambda|\mathbf{Z})\lambda}{\lambda+\zeta\sigma^2\eta}\Big\rangle^2_{\mathbf{Z}}\Bigg]+O(1/N) \qquad \text{(F40)}
\end{aligned}
$$

Hence

$$
\mathrm{Var}\big(\frac{E[\mathscr{D}]}{N}\big)=\frac{\zeta^2\big(S^4\sigma^4\eta^2+\sigma_0^4-2\sigma_0^2S^2\sigma^2\eta\big)}{4\sigma^4}\Bigg[\Big\langle\Big(\int\mathrm{d}\lambda\ \frac{\rho_d(\lambda|\mathbf{Z})\lambda}{\lambda+\zeta\sigma^2\eta}\mathrm{d}\lambda\Big)^2\Big\rangle_{\mathbf{Z}}-\Big\langle\int\mathrm{d}\lambda\ \frac{\rho_d(\lambda|\mathbf{Z})\lambda}{\lambda+\zeta\sigma^2\eta}\mathrm{d}\lambda\Big\rangle^2_{\mathbf{Z}}\Bigg]+\mathcal{O}(\frac{1}{N}). \qquad \text{(F41)}
$$

Furthermore, the covariance (F19) becomes

$$
\begin{aligned}
4\sigma^2\mathrm{Cov}\big(E[\mathscr{D}],\mathrm{S}[\mathscr{D}]\big)&=\mathrm{Cov}\Big(\mathbf{t}^{\mathrm{T}}\big(\mathbf{I}_N-\mathbf{Z}\mathbf{J}^{-1}_{\sigma^2\eta}\mathbf{Z}^T\big)\mathbf{t},\log\Big|\mathbf{J}^{-1}_{\sigma^2\eta}\Big|\Big)\\
&=\Big\langle\Big[\Big\langle\boldsymbol{\theta}_0^{\mathrm{T}}\big(\mathbf{J}-\mathbf{J}\mathbf{J}^{-1}_{\sigma^2\eta}\mathbf{J}\big)\boldsymbol{\theta}_0\Big\rangle_{\boldsymbol{\theta}_0}+\sigma_0^2\big(N-\mathrm{Tr}\Big[\mathbf{J}\mathbf{J}^{-1}_{\sigma^2\eta}\Big]\big)\Big]\log\Big|\mathbf{J}^{-1}_{\sigma^2\eta}\Big|\Big\rangle_{\mathbf{Z}}\\
&\quad-\Big\langle\Big[\Big\langle\boldsymbol{\theta}_0^{\mathrm{T}}\big(\mathbf{J}-\mathbf{J}\mathbf{J}^{-1}_{\sigma^2\eta}\mathbf{J}\big)\boldsymbol{\theta}_0\Big\rangle_{\boldsymbol{\theta}_0}+\sigma_0^2\big(N-\mathrm{Tr}\Big[\mathbf{J}\mathbf{J}^{-1}_{\sigma^2\eta}\Big]\big)\Big]\Big\rangle_{\mathbf{Z}}\Big\langle\log\Big|\mathbf{J}^{-1}_{\sigma^2\eta}\Big|\Big\rangle_{\mathbf{Z}}\\
&=\Big\langle\Big[S^2\mathrm{Tr}\Big[\mathbf{J}-\mathbf{J}\mathbf{J}^{-1}_{\sigma^2\eta}\mathbf{J}\Big]+\sigma_0^2\big(N-\mathrm{Tr}\Big[\mathbf{J}\mathbf{J}^{-1}_{\sigma^2\eta}\Big]\big)\Big]\log\Big|\mathbf{J}^{-1}_{\sigma^2\eta}\Big|\Big\rangle_{\mathbf{Z}}\\
&\quad-\Big\langle\Big[S^2\mathrm{Tr}\Big[\mathbf{J}-\mathbf{J}\mathbf{J}^{-1}_{\sigma^2\eta}\mathbf{J}\Big]+\sigma_0^2\big(N-\mathrm{Tr}\Big[\mathbf{J}\mathbf{J}^{-1}_{\sigma^2\eta}\Big]\big)\Big]\Big\rangle_{\mathbf{Z}}\Big\langle\log\Big|\mathbf{J}^{-1}_{\sigma^2\eta}\Big|\Big\rangle_{\mathbf{Z}} \qquad \text{(F42)}
\end{aligned}
$$

From this, upon setting $\mathbf{Z}\to\mathbf{Z}/\sqrt{d}$ with $z_i(\mu)=\mathcal{O}(1)$ for all (i,μ) then follows the result

$$
\begin{aligned}
4\sigma^2\mathrm{Cov}\big(E[\mathscr{D}]/N,\mathrm{S}[\mathscr{D}]/N\big) &= \Big\langle\Big[S^2\frac{1}{\zeta N}\mathrm{Tr}\left[\mathbf{J}-\mathbf{J}\mathbf{J}_{\zeta\sigma^2\eta}^{-1}\mathbf{J}\right]+\sigma_0^2\big(1-\frac{1}{N}\mathrm{Tr}\left[\mathbf{J}\mathbf{J}_{\zeta\sigma^2\eta}^{-1}\right]\big)\Big]\frac{1}{N}\log\left|\zeta\mathbf{J}_{\zeta\sigma^2\eta}^{-1}\right|\Big\rangle_{\mathbf{Z}}\\
&\quad-\Big\langle\Big[S^2\frac{1}{\zeta N}\mathrm{Tr}\left[\mathbf{J}-\mathbf{J}\mathbf{J}_{\zeta\sigma^2\eta}^{-1}\mathbf{J}\right]+\sigma_0^2\big(1-\frac{1}{N}\mathrm{Tr}\left[\mathbf{J}\mathbf{J}_{\zeta\sigma^2\eta}^{-1}\right]\big)\Big]\Big\rangle_{\mathbf{Z}}\frac{1}{N}\left\langle\log\left|\zeta\mathbf{J}_{\zeta\sigma^2\eta}^{-1}\right|\right\rangle_{\mathbf{Z}}\\
&=\zeta\int\mathrm{d}\lambda\mathrm{d}\tilde\lambda\ \left[\langle\rho_d(\lambda|\mathbf{Z})\rangle_{\mathbf{Z}}\langle\rho_d(\tilde\lambda|\mathbf{Z})\rangle_{\mathbf{Z}}-\langle\rho_d(\lambda|\mathbf{Z})\rho_d(\tilde\lambda|\mathbf{Z})\rangle_{\mathbf{Z}}\right]\\
&\qquad\times\Big[\frac{S^2\lambda\zeta\sigma^2\eta}{\lambda+\zeta\sigma^2\eta}+\sigma_0^2\big(1-\frac{\lambda\zeta}{\lambda+\zeta\sigma^2\eta}\big)\Big]\log\big(\tilde\lambda+\zeta\sigma^2\eta\big) \qquad\text{(F43)}
\end{aligned}
$$

Finally, the entropy variance (F18) for $\mathbf{Z}\to\mathbf{Z}/\sqrt{d}$ is given by

$$
\mathrm{Var}\big(\frac{\mathrm{S}[\mathscr{D}]}{N}\big)=\frac{\zeta^2}{4}\int\mathrm{d}\lambda\mathrm{d}\tilde\lambda\ \left[\langle\rho_d(\lambda|\mathbf{Z})\rho_d(\tilde\lambda|\mathbf{Z})\rangle_{\mathbf{Z}}-\langle\rho_d(\lambda|\mathbf{Z})\rangle_{\mathbf{Z}}\langle\rho_d(\tilde\lambda|\mathbf{Z})\rangle_{\mathbf{Z}}\right]\log(\lambda+\zeta\sigma^2\eta)\log(\tilde\lambda+\zeta\sigma^2\eta). \qquad\text{(F44)}
$$

Using all of the above results in (F5) we finally obtain the variance of the free energy density:

$$
\begin{aligned}
\mathrm{Var}\Big(\frac{F_{\beta,\sigma^2}[\mathscr{D}]}{N}\Big)&=\int\mathrm{d}\lambda\mathrm{d}\tilde\lambda\ \left[\langle\rho_d(\lambda|\mathbf{Z})\rho_d(\tilde\lambda|\mathbf{Z})\rangle_{\mathbf{Z}}-\langle\rho_d(\lambda|\mathbf{Z})\rangle_{\mathbf{Z}}\langle\rho_d(\tilde\lambda|\mathbf{Z})\rangle_{\mathbf{Z}}\right]\\
&\times\Bigg[\frac{\zeta^2}{4\sigma^4}\big(S^4\sigma^4\eta^2+\sigma_0^4-2\sigma_0^2S^2\sigma^2\eta\big)\frac{\lambda}{\lambda+\zeta\sigma^2\eta}\frac{\tilde\lambda}{\tilde\lambda+\zeta\sigma^2\eta}+\frac{1}{4}T^2\zeta^2\log\big(\lambda+\zeta\sigma^2\eta\big)\log\big(\tilde\lambda+\zeta\sigma^2\eta\big)\\
&\quad-\frac{T\zeta}{2\sigma^2}\Big(\frac{S^2\lambda\zeta\sigma^2\eta}{\lambda+\zeta\sigma^2\eta}+\sigma_0^2\Big(1-\frac{\lambda\zeta}{\lambda+\zeta\sigma^2\eta}\Big)\Big)\log\big(\tilde\lambda+\zeta\sigma^2\eta\big)\Bigg]+\mathcal{O}(1/N). \qquad\text{(F45)}
\end{aligned}
$$

Appendix G: Self-averaging of the estimator $\hat\sigma^2$

Let us consider the equation

$$
\sigma^2=\frac{\beta}{(\beta-\zeta)}\frac{1}{N}\left|\left|\mathbf{t}-\mathbf{Z}\hat{\boldsymbol{\theta}}[\mathscr{D}]\right|\right|^2-\frac{\sigma^4\eta}{(\beta-\zeta)}\frac{1}{N}\mathrm{Tr}\left[\mathbf{J}_{\sigma^2\eta}^{-1}\right]+\frac{2\sigma^4\beta}{(\beta-\zeta)N}\frac{\partial}{\partial\sigma^2}\log P(\sigma^2). \qquad\text{(G1)}
$$

Upon inserting the MAP estimator (A3) and the short-hand $\sigma^2=v$ we obtain

$$
v=\frac{\beta}{(\beta-\zeta)}\frac{1}{N}\mathbf{t}^{\mathrm{T}}\left(\mathbf{I}_N-\mathbf{Z}\mathbf{J}_{v\eta}^{-1}\mathbf{Z}^{\mathrm{T}}\right)^2\mathbf{t}-\frac{v^2\eta}{(\beta-\zeta)}\frac{1}{N}\mathrm{Tr}\left[\mathbf{J}_{v\eta}^{-1}\right]+\frac{2v^2\beta}{(\beta-\zeta)N}\frac{\partial}{\partial v}\log P(v). \qquad\text{(G2)}
$$

To solve this equation for v we define the following recursion, with the short-hand $\partial_v\equiv\frac{\partial}{\partial v}$:

$$
v_{t+1}=\frac{\beta}{(\beta-\zeta)}\frac{1}{N}\mathbf{t}^{\mathrm{T}}\left(\mathbf{I}_N-\mathbf{Z}\mathbf{J}_{v_t\eta}^{-1}\mathbf{Z}^{\mathrm{T}}\right)^2\mathbf{t}-\frac{v_t^2\eta}{(\beta-\zeta)}\frac{1}{N}\mathrm{Tr}\left[\mathbf{J}_{v_t\eta}^{-1}\right]+\frac{2v_t^2\beta}{(\beta-\zeta)N}\partial_v\log P(v)|_{v=v_t}, \qquad\text{(G3)}
$$

Since $\mathbf{t}=\mathbf{Z}\boldsymbol{\theta}_0+\boldsymbol{\epsilon}$ this recursion, of which the desired estimator is the fixed-point, has the general form

$$
v_{t+1}=\Psi\left[v_t|\mathbf{Z},\boldsymbol{\theta}_0,\boldsymbol{\epsilon}\right]. \qquad\text{(G4)}
$$

Thus for any choice of $\{\mathbf{Z},\boldsymbol{\theta}_0,\boldsymbol{\epsilon}\}$, i.e. which play the role of 'disorder', the function Ψ is a random non-linear operator acting on v_t. If the initial value v_0 is *independent* of the disorder, then the next value v_1 is independent from a *particular* realisation of disorder, i.e. v_1 is *self-averaging*, as soon as the operator Ψ is self-averaging, i.e. if

$$
\lim_{(N,d)\to\infty}\left\langle\Psi^2\left[v_0|\mathbf{Z},\boldsymbol{\theta}_0,\boldsymbol{\epsilon}\right]\right\rangle_{\mathbf{Z},\boldsymbol{\theta}_0,\boldsymbol{\epsilon}}-\left\langle\Psi^2\left[v_0|\mathbf{Z},\boldsymbol{\theta}_0,\boldsymbol{\epsilon}\right]\right\rangle^2_{\mathbf{Z},\boldsymbol{\theta}_0,\boldsymbol{\epsilon}}=0. \qquad\text{(G5)}
$$

By induction, all v_t with $t\geq1$ will then be self-averaging, and (G4) can for $(N,d)\to\infty$ be replaced by the following deterministic map, whose fixed-point will be the asymptotic estimator $\hat\sigma^2$ that is then guaranteed to be self-averaging:

$$
v_{t+1}=\left\langle\Psi\left[v_t|\mathbf{Z},\boldsymbol{\theta}_0,\boldsymbol{\epsilon}\right]\right\rangle_{\mathbf{Z},\boldsymbol{\theta}_0,\boldsymbol{\epsilon}} \qquad\text{(G6)}
$$

To prove the self-averaging property of Ψ we assume that the true parameters $\boldsymbol{\theta}_0$ and noise $\boldsymbol{\epsilon}$ have mean $\mathbf{0}$ and the covariance matrices $S^2\mathbf{I}_d$ and $\sigma_0^2\mathbf{I}_N$, respectively. Let us first consider the average

$$\begin{aligned}
\langle\Psi\left[v_0|\mathbf{Z},\boldsymbol{\theta}_0,\boldsymbol{\epsilon}\right]\rangle_{\mathbf{Z},\boldsymbol{\theta}_0,\boldsymbol{\epsilon}} &= \frac{\beta}{N(\beta-\zeta)}\left\langle \mathbf{t}^{\mathrm{T}}\left(\mathbf{I}_N-\mathbf{Z}\mathbf{J}_{v_0\eta}^{-1}\mathbf{Z}^{\mathrm{T}}\right)^2\mathbf{t}\right\rangle_{\mathbf{Z},\boldsymbol{\theta}_0,\boldsymbol{\epsilon}} - \frac{v_0^2\eta}{N(\beta-\zeta)}\left\langle \mathrm{Tr}\left[\mathbf{J}_{v_0\eta}^{-1}\right]\right\rangle_{\mathbf{Z}} + \frac{2v_0^2\beta}{(\beta-\zeta)N}\partial_v\log P(v)|_{v=v_0}\\
&= \frac{\beta\sigma_0^2}{N(\beta-\zeta)}\left\langle \mathrm{Tr}\left[\left(\mathbf{I}_N-\mathbf{Z}\mathbf{J}_{v_0\eta}^{-1}\mathbf{Z}^{\mathrm{T}}\right)^2\right]\right\rangle_{\mathbf{Z}} + \frac{\beta}{N(\beta-\zeta)}\left\langle \boldsymbol{\theta}_0^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\left(\mathbf{I}_N-\mathbf{Z}\mathbf{J}_{v_0\eta}^{-1}\mathbf{Z}^{\mathrm{T}}\right)^2\mathbf{Z}\boldsymbol{\theta}_0\right\rangle_{\mathbf{Z},\boldsymbol{\theta}_0}\\
&\quad - \frac{v_0^2\eta}{N(\beta-\zeta)}\left\langle \mathrm{Tr}\left[\mathbf{J}_{v_0\eta}^{-1}\right]\right\rangle_{\mathbf{Z}} + \frac{2v_0^2\beta}{(\beta-\zeta)N}\partial_v\log P(v)|_{v=v_0}\\
&= \frac{\beta}{N(\beta-\zeta)}\sigma_0^2\left\langle \mathrm{Tr}\left[\left(\mathbf{I}_N-\mathbf{Z}\mathbf{J}_{v_0\eta}^{-1}\mathbf{Z}^{\mathrm{T}}\right)^2\right]\right\rangle_{\mathbf{Z}} + \frac{\beta}{N(\beta-\zeta)}S^2\left\langle \mathrm{Tr}\left[\mathbf{J}\left(\mathbf{I}_d-\mathbf{J}_{v_0\eta}^{-1}\mathbf{J}\right)^2\right]\right\rangle_{\mathbf{Z}}\\
&\quad - \frac{v_0^2\eta}{N(\beta-\zeta)}\left\langle \mathrm{Tr}\left[\mathbf{J}_{v_0\eta}^{-1}\right]\right\rangle_{\mathbf{Z}} + \frac{2v_0^2\beta}{(\beta-\zeta)N}\partial_v\log P(v)|_{v=v_0}\\
&= \frac{\beta\sigma_0^2}{\beta-\zeta}\left(1-\zeta+\frac{1}{N}\left\langle \mathrm{Tr}\left[\left(\mathbf{I}_d-\mathbf{J}\mathbf{J}_{v_0\eta}^{-1}\right)^2\right]\right\rangle_{\mathbf{Z}}\right) + \frac{\beta S^2}{N(\beta-\zeta)}\left\langle \mathrm{Tr}\left[\mathbf{J}\left(\mathbf{I}_d-\mathbf{J}_{v_0\eta}^{-1}\mathbf{J}\right)^2\right]\right\rangle_{\mathbf{Z}}\\
&\quad - \frac{v_0^2\eta}{N(\beta-\zeta)}\left\langle \mathrm{Tr}\left[\mathbf{J}_{v_0\eta}^{-1}\right]\right\rangle_{\mathbf{Z}} + \frac{2v_0^2\beta}{(\beta-\zeta)N}\partial_v\log P(v)|_{v=v_0}
\end{aligned}\tag{G7}$$

We then make the usual substitution $\mathbf{Z}\to\mathbf{Z}/\sqrt{d}$, with $z_i(\mu)=\mathcal{O}(1)$ for all (i,μ), and we define the average $\rho_d(\lambda)=\langle\rho_d(\lambda|\mathbf{Z})\rangle_{\mathbf{Z}}$ of the eigenvalue density

$$\rho_d(\lambda|\mathbf{Z})=\frac{1}{d}\sum_{\ell=1}^{d}\delta\left[\lambda-\lambda_\ell(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}/N)\right].\tag{G8}$$

of the empirical covariance matrix. Then the above average becomes

$$\begin{aligned}
\langle\Psi\left[v_0|\mathbf{Z},\boldsymbol{\theta}_0,\boldsymbol{\epsilon}\right]\rangle_{\mathbf{Z},\boldsymbol{\theta}_0,\boldsymbol{\epsilon}} &= \frac{\beta\sigma_0^2}{\beta-\zeta}\left(1-\zeta+\zeta\int\!\mathrm{d}\lambda\;\rho_d(\lambda)\Big(\frac{\zeta v_0\eta}{\lambda+\zeta v_0\eta}\Big)^2\right) + \frac{\beta S^2}{\beta-\zeta}\int\!\mathrm{d}\lambda\;\rho_d(\lambda)\Big(\frac{\zeta v_0\eta}{\lambda+\zeta v_0\eta}\Big)^2\lambda\\
&\quad - \frac{v_0^2\eta\zeta^2}{\beta-\zeta}\int\!\mathrm{d}\lambda\;\frac{\rho_d(\lambda)}{\lambda+\zeta v_0\eta} + \frac{2v_0^2\beta}{(\beta-\zeta)N}\partial_v\log P(v)|_{v=v_0}
\end{aligned}\tag{G9}$$

Finally we turn to the variance

$$\begin{aligned}
\mathrm{Var}(\Psi\left[v_0|\mathbf{Z},\boldsymbol{\theta}_0,\boldsymbol{\epsilon}\right]) &= \left(\frac{\beta}{\beta-\zeta}\right)^2\frac{1}{N^2}\mathrm{Var}\left(\mathbf{t}^{\mathrm{T}}\left(\mathbf{I}_N-\mathbf{Z}\mathbf{J}_{v_0\eta}^{-1}\mathbf{Z}^{\mathrm{T}}\right)^2\mathbf{t}\right) + \frac{v_0^4\eta^2}{(\beta-\zeta)^2}\frac{1}{N^2}\mathrm{Var}\left(\mathrm{Tr}\left[\mathbf{J}_{v_0\eta}^{-1}\right]\right)\\
&\quad - \frac{2\beta v_0^2\eta}{(\beta-\zeta)^2}\frac{1}{N^2}\mathrm{Cov}\left(\mathbf{t}^{\mathrm{T}}\left(\mathbf{I}_N-\mathbf{Z}\mathbf{J}_{v_0\eta}^{-1}\mathbf{Z}^{\mathrm{T}}\right)^2\mathbf{t},\,\mathrm{Tr}\left[\mathbf{J}_{v_0\eta}^{-1}\right]\right).
\end{aligned}\tag{G10}$$

Computing in this expression the relevant averages over the random variables $\mathbf{Z}$, $\boldsymbol{\theta}_0$ and $\boldsymbol{\epsilon}$, with the familiar substitution $\mathbf{Z}\to\mathbf{Z}/\sqrt{d}$, gives us the following result

$$\begin{aligned}
\mathrm{Var}(\Psi\left[v_0|\mathbf{Z},\boldsymbol{\theta}_0,\boldsymbol{\epsilon}\right]) &= \int\!\mathrm{d}\lambda\mathrm{d}\tilde{\lambda}\;C_d(\lambda,\tilde{\lambda})\left\{\left(\frac{\beta}{\beta-\zeta}\right)^2\left(\frac{\zeta v_0\eta}{\lambda+\zeta v_0\eta}\right)^2\left(\frac{\zeta v_0\eta}{\tilde{\lambda}+\zeta v_0\eta}\right)^2\left[\sigma_0^4\zeta^2+S^4\lambda\tilde{\lambda}+2\sigma_0^2S^2\zeta\lambda\right]\right.\\
&\quad\left.+\left(\frac{v^2\zeta^2\eta}{\beta-\zeta}\right)^2\frac{1}{(\lambda+\zeta v_0\eta)\left(\tilde{\lambda}+\zeta v_0\eta\right)} - \frac{2\beta v_0^2\zeta^2\eta}{(\beta-\zeta)^2}\left(\frac{\zeta v_0\eta}{\lambda+\zeta v_0\eta}\right)^2\frac{\sigma_0^2\zeta+S^2\lambda}{\tilde{\lambda}+\zeta v_0\eta}\right\}\\
&\quad+\frac{2}{N}\left(\frac{\beta}{\beta-\zeta}\right)^2\int\!\mathrm{d}\lambda\;\rho_d(\lambda)\left\{\sigma_0^4\left[1-\zeta+\zeta\Big(\frac{\zeta v_0\eta}{\lambda+\zeta v_0\eta}\Big)^4\right]+\frac{S^4}{\zeta}\Big(\frac{\zeta v_0\eta}{\lambda+\zeta v_0\eta}\Big)^4\lambda^2\right\},
\end{aligned}\tag{G11}$$

with the correlation function $C_d(\lambda,\tilde{\lambda})=\left\langle\rho_d(\lambda|\mathbf{Z})\rho_d(\tilde{\lambda}|\mathbf{Z})\right\rangle_{\mathbf{Z}}-\left\langle\rho_d(\lambda|\mathbf{Z})\right\rangle_{\mathbf{Z}}\left\langle\rho_d(\tilde{\lambda}|\mathbf{Z})\right\rangle_{\mathbf{Z}}$. Clearly, if the spectrum $\rho_d(\lambda|\mathbf{Z})$ is self-averaging when $(N,d)\to\infty$, then the correlation function will vanish in this limit, and hence $\Psi\left[v_0|\mathbf{Z},\boldsymbol{\theta}_0,\boldsymbol{\epsilon}\right]$ will be self-averaging.

[1] A. C. Rencher and G. B. Schaalje, *Linear models in statistics* (John Wiley & Sons, 2008).

[2] P. J. Huber and E. M. Ronchetti, *Robust statistics* (John Wiley & Sons, 2009).
[3] M. Bayati and A. Montanari, IEEE Transactions on Information Theory **58**, 1997 (2011).
[4] N. El Karoui, D. Bean, P. J. Bickel, C. Lim, and B. Yu, Proceedings of the National Academy of Sciences **110**, 14557 (2013).
[5] N. E. Karoui and E. Purdom, arXiv preprint arXiv:1608.00696 (2016).
[6] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint* (Cambridge University Press, 2019).
[7] M. Advani and S. Ganguli, Physical Review X **6**, 031034 (2016).
[8] J. Barbier, F. Krzakala, N. Macris, L. Miolane, and L. Zdeborová, P. Natl. Acad. Sci. USA **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, Physical review **106**, 620 (1957).
[11] H. Nishimori, *Statistical physics of spin glasses and information processing: an introduction* (Clarendon Press, 2001).
[12] C. M. Bishop, *Pattern recognition and machine learning* (Springer, 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 Co Inc, 1987).
[14] M. Mézard and A. Montanari, *Information, physics, and computation* (Oxford University Press, 2009).
[15] D. Donoho and A. Montanari, Probability Theory and Related Fields **166**, 935 (2016).
[16] C. Gerbelot, A. Abbara, and F. Krzakala, arXiv preprint arXiv:2002.04372 (2020).
[17] J. Barbier, N. Macris, M. Dia, and F. Krzakala, IEEE Transactions on Information Theory **66**, 4270 (2020).
[18] A. Engel and C. Van den Broeck, *Statistical mechanics of learning* (Cambridge University Press, 2001).
[19] L. Zdeborová and F. Krzakala, Advances in Physics **65**, 453 (2016).
[20] N. G. De Bruijn, *Asymptotic methods in analysis* (New York: Dover, 1981).
[21] T. M. Cover and J. A. Thomas, *Elements of information theory* (John Wiley & Sons, 2012).
[22] A. C. C. Coolen, J. E. Barrett, P. Paga, and C. J. Perez-Vicente, Journal of Physics A: Mathematical and Theoretical **50**, 375001 (2017).
[23] If the inverse-χ^2 distribution is used as a prior for σ^2, then the MAP estimator for the latter is given by $\sigma^2 = (N+\nu+2)^{-1} + (N+\nu+2)^{-1}||\mathbf{t} - \mathbf{Z}\hat{\boldsymbol{\theta}}[\mathscr{D}]||^2$ which suggests that the hyper-parameter ν has to be *extensive* in order to remain relevant for large N.
[24] S. Nadarajah and S. Kotz, Acta Applicandae Mathematica **89**, 53 (2005).
[25] B. G. Kibria and A. H. Joarder, Journal of Statistical research **40**, 59 (2006).
[26] In statistics this phenomena would be referred to as having a fully concentrated measure.
[27] W. G. Cochran, Mathematical Proceedings of the Cambridge Philosophical Society **30**, 178–191 (1934).
[28] F. Götze and A. Tikhomirov, Bernoulli **10**, 503 (2004).
[29] P.-H. Chavanis, Phys. Rev. E. **65**, 056123 (2002).
[30] M. Rudelson and R. Vershynin, Electronic Communications in Probability **18** (2013).
[31] M. L. Eaton, *Multivariate statistics: a vector space approach.* (JOHN WILEY & SONS, 1983).
[32] M. Taboga, *Lectures on probability theory and mathematical statistics* (Independent Publishing Platform, 2017).