Bilinear Sequence Regression: A Model for Learning from Long Sequences of High-Dimensional Tokens

Vittorio Erba,[1] Emanuele Troiani,[1] Luca Biggio,[1,2] Antoine Maillard,[3] and Lenka Zdeborová[1]

[1]*Statistical Physics of Computation Laboratory, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland*

[2]*Department of Computing Sciences, Università Bocconi, Milan, Italy*

[3]*Department of Mathematics, ETH Zürich, Switzerland*

(Received 30 October 2024; revised 17 March 2025; accepted 8 May 2025; published 16 June 2025)

Current progress in artificial intelligence is centered around so-called large language models that consist of neural networks processing long sequences of high-dimensional vectors called tokens. Statistical physics provides powerful tools to study the functioning of learning with neural networks and has played a recognized role in the development of modern machine learning. The statistical physics approach relies on simplified and analytically tractable models of data. However, simple tractable models for long sequences of high-dimensional tokens are largely underexplored. Inspired by the crucial role models such as the single-layer teacher-student perceptron (also known as generalized linear regression) played in the theory of fully connected neural networks, in this paper, we introduce and study the *bilinear sequence regression* (BSR) as one of the most basic models for sequences of tokens. We note that modern architectures naturally subsume the BSR model due to the skip connections. Building on recent methodological progress, we compute the Bayes-optimal generalization error for the model in the limit of long sequences of high-dimensional tokens and provide a message-passing algorithm that matches this performance. We quantify the improvement that optimal learning brings with respect to vectorizing the sequence of tokens and learning via simple linear regression. We also unveil surprising properties of the gradient descent algorithms in the BSR model.

DOI: 10.1103/l4p2-vrxt

Subject Areas: Interdisciplinary Physics, Statistical Physics

I. INTRODUCTION

A. Motivation

1. Deep learning and the statistical physics approach to understand it

We are witnessing unprecedented progress in artificial intelligence, largely thanks to advances in learning with large multilayer neural networks, commonly referred to as deep learning [1]. Milestones such as the classification of images from the ImageNet dataset [2,3] or superhuman performance in the game of Go [4] used deep neural networks based on combinations of fully connected and convolutional layers that map rather high-dimensional vectors into vectors of (in general) different, but still high, dimension. While deep learning is undeniably successful in practical tasks, the underlying theoretical mechanisms behind its functioning remain covered with open questions. This led to an abundance of theoretical works aiming to explain the behavior of deep neural networks that are observed in practice, such as the lack of overfitting in overparametrized neural networks, the principles thanks to which gradient-based training dynamics reach configurations with good generalization performance while many other configurations of equally good training loss exist and lead to bad generalization, or theoretically grounded principles leading to the choice of the best-performing architecture, algorithms, and hyperparameters, for a given dataset and task. A subfield of the theory of deep learning stems from statistical physics, a scientific field that is particularly well suited to come up with models that are solvable in the high-dimensional limit and to provide insights into the above questions [5]. This line of work was initiated decades ago, with, e.g., Refs. [6–10], and regained broad interest in the past decade with a number of influential works, e.g., Refs. [11–18]. One of the instrumental models in this line of work is the *teacher-student* model, where one investigates whether a neural network can learn from data that were generated by a teacher neural network whose weights are not known to the student neural network. This teacher-student setting was introduced in Ref. [9] and used broadly, including in most of the above-cited works.

2. Sequence modeling takes the lead

The landscape of research in deep learning reshaped considerably with the rise of transformer architectures [19] and subsequent large language models (LLMs), sometimes also called foundation models, such as the GPT family [20–22], leading to the well-known chatbot ChatGPT by OpenAI [22,23] that took the field of AI and the whole high-tech industry by storm. Transformers are types of deep neural networks that stand behind this progress. They are composed of combinations of fully connected layers and, crucially, so-called attention layers. While a fully connected layer maps vectors into vectors, an attention layer maps sequences of vectors (called tokens) into sequences of vectors (tokens). In language modeling, a token would typically be associated with a word, and each such token is mapped (embedded) into a relatively high-dimensional (typically, around a 1000-dimensional) vector. The sequence then corresponds to a text composed of words or tokens and is also long, corresponding to the number of words in a text. It is fair to say that the impressive performance of current LLMs was not anticipated by many. The functioning of LLMs is surrounded by even more theoretical open questions—including the emergence of capabilities [24], the neural scaling laws [25], or in-context learning [26].

In our opinion, the most fundamental theoretical question underlying transformers is *why is it advantageous to present the data as long sequences of high-dimensional tokens?* More specifically, why is it advantageous for network architecture to act differently in token space and embedding space? Indeed, if one vectorized the data into a single large vector and used a fully connected architecture, the universal approximation theorem [27,28] would still imply that a generic set of functions can be represented this way. There must be a computational advantage in presenting the data as sequences of tokens that the transformer architecture exploits. This advantage may be related to the underlying structure of the data, the reduction of the number of trainable parameters, or the flexibility with respect to sequence length. However, the precise reasons are not understood, and the advantage with respect to learning from the vectorized data is not quantified theoretically.

One can anticipate that, also in this context of learning from sequences of tokens, the statistical physics approach, based on simplified models that capture some of the intriguing properties and behaviors, will help to clarify some of the key questions surrounding LLMs, transformers, and learning with attention layers. Indeed, works in this direction started appearing in the past couple of years. From those we are aware of, several build interesting simplified models and then investigate the training of the corresponding toy-transformer architecture numerically or phenomenologically [29–31]. Others study the propagation of a signal through a trained transformer [32,33]. So far, only a handful of works have been able to analyze the training of a toy transformer analytically. Concerning works that analyze the training of a neural network for data consisting of sequences of tokens, Ref. [34] analyzes an attention layer learning Gaussian data generated by a model where the sequence length L is large, but the token dimension d is small. The authors of Ref. [35] analyze minimizers of the training loss and corresponding phase transitions for a teacher-student-like data model where the token dimension d is large, but the length of the sequence is small $L = O(1)$. Finally, the authors of Ref. [36] consider a very interesting case of in-context learning linear regression where both the token dimension d and the sequence length L (corresponding to the number of samples given in each context in Ref. [36]) are large, but they analyze only a linear attention layer, which can be seen as a special case of ridge regression, limiting the generalizability of their approach to address a broader set of questions. In practical settings, such as the GPT architectures, both the length of the sequence L and the embedding dimension d are large, typically in thousands [37]. It is, thus, critical to build a theoretically analyzable model, where the thermodynamic (or high-dimensional) limit corresponds to both the length L and the embedding dimension d going to infinity.

Another motivation of this paper stems from the following lines. While complex nonlinear neural networks perform extremely well in practice, the deep learning revolution has exposed many fundamental questions even in basic statistical methods like linear regression. Indeed, describing training procedures in nonconvex optimization problems is highly nontrivial even in simple single-layer neural networks with nonlinear outputs, such as in the dynamics of gradient descent for phase retrieval problems [38,39]. Another example of a phenomenon that can be understood already in linear regression (or its slight variations) is double descent [40], which has fundamentally changed our understanding of overfitting and the bias-variance trade-off, while it has been theoretically explained within the framework of linear regression [40,41]. It is clear from these examples that such basic models as linear regression have been extremely useful in clarifying the properties of modern deep learning. Therefore, the question arises: *What is the basic model, analogous to perceptron or generalized linear regression, for sequences of tokens?* In this paper, we introduce such a model and initiate its study, thus providing a rich theoretical playground to tackle questions about learning from long sequences of high-dimensional data.

B. Definition of the bilinear sequence regression model

Motivated by the above questions, the present paper introduces a prototypical analytically solvable model for supervised learning from long sequences of high-dimensional tokens, which we name the bilinear sequence regression (BSR) model. We consider a supervised regression task on a dataset of n input-output pairs $\mathcal{D}_n = \{(X^\mu, y^\mu)\}_{\mu=1}^n$ with $X^\mu \in \mathbb{R}^{L\times d}$ being the inputs consisting

of a sequence of length L of d-dimensional tokens and $y \in \mathbb{R}$ the labels. We focus here on regression, as opposed to the more common next-token prediction, because several theoretical studies of neural networks focused on linear regression; we are, thus, able to leverage the insights gained in the linear regression literature. Following the statistical physics studies of the teacher-student setting to analyze learning with fully connected neural networks, we draw each component of the input data X^{μ}_{ij} independently from a Gaussian distribution of zero mean and unit variance. The labels are then generated through the following (teacher) model:

$$y^{\mu} \sim P_{\text{out}}\left(\cdot \middle| \frac{1}{\sqrt{dLr}} \sum_{a,i=1}^{L,d} X^{\mu}_{ai} \sum_{\gamma=1}^{r} U^{*}_{i\gamma} V^{*}_{\gamma a}\right), \tag{1}$$

where $U^{*} \in \mathbb{R}^{d\times r}$, $V^{*} \in \mathbb{R}^{r\times L}$, and their components are taken independent identically distributed (i.i.d.) from a standard Gaussian and P_{out} is a probabilistic scalar output channel. The parameter r will be called the *width* of the model. Given the dataset $\mathcal{D}_n$, the task is then to learn a function $f : X \in \mathbb{R}^{L\times d} \to y \in \mathbb{R}$ and obtain a good performance on a test set.

We consider a Bayesian setting, in which the learner knows the architecture of Eq. (1), i.e., the form of P_{out} and the distributions of U^{*} and V^{*}. In this context, the main task is to recover the values of U^{*} and V^{*}. It is well known that the optimal performance for this task is reached by the so-called *Bayes-optimal* estimator, which corresponds to the mean of the posterior distribution, as we describe in more detail in Sec. II.

We think of the inputs $X^{\mu} \in \mathbb{R}^{L\times d}$ as sequences of tokens and the outputs y^{μ} as labels. As a concrete example, each row of X^{μ} can be thought of as a vector embedding of a word, so that X^{μ} represents a text in some language and y^{μ} may be a sentiment score associated with the text, categorizing its meaning, for example, as uplifting or depressing. As said above, data in the form of sequences of tokens are ubiquitous in modern machine learning, including natural language datasets (where tokens are words and sequences are phrases) or biological datasets (where tokens are amino acids and sequences are proteins), yet our understanding of the performance of learning algorithms on such data is scarce. In this sense, the model (1) provides a benchmark dataset where the inputs X^{μ} are unstructured (random) and the function from $X \in \mathbb{R}^{L\times d}$ to $y \in \mathbb{R}$ is parametrized by ground-truth latent variables U^{*} and V^{*} in a bilinear form way which is among the simplest functional forms one can posit when inputs X^{μ} are sequences of tokens.

What we perceive as a key property of a simple sequence model is that in the BSR the factor V^{*} acts on the sequence elements while U^{*} acts separately on the embedding dimensions of the data. This is mimicking the way transformers process data (see Sec. I C for a review). Indeed, a key aspect of the dot-product attention layer is that it transforms the representation in the embedding space in one way (via the key, query, and value matrices) and the representation in the sequence space in another way (via the attention matrix). Notice also that the attention layer treats elements of the sequence in a permutationally invariant manner (unless positional embeddings are explicitly provided) and does so as well for the elements of the embedding. All this is reproduced in our model—the permutational invariance between elements of the sequence as well as the transformation of the embedding representation (via the matrix U^{*} in our model) being different from the transformation of the sequence representation (via the matrix V^{*} in our model).

In this paper, we show that Eq. (1) is a useful toy model for supervised learning over sequential data in a similar way as the teacher-student perceptron for nonsequential data, which is widely studied in the statistical physics literature. The key question, then, is how neural networks learn on such a dataset in order to be able to predict labels on previously unseen inputs. Additionally, how does the performance depend on the architecture of the network and the algorithm used to learn the data?

In the present paper, we address the following questions:

(Q1) What is the performance of the Bayes-optimal estimator learning from a given number of samples n of data generated by the model (1) in the limit of d and L large, proportionally to each other? We consider the full range of possible values for the width parameter r, with a specific focus on the regimes where r is either proportional to d and L or remains of the order of $\mathcal{O}(1)$, as $L, d \to \infty$.

(Q2) Can this Bayes-optimal performance be reached by efficient algorithms, and which ones?

(Q3) Does the Bayes-optimal performance present sharp thresholds (phase transitions) in performance as a function of the number of samples? If yes, at which sample complexities?

(Q4) How does the Bayes-optimal performance compare to the performance of linear regression on the vectorized data?

(Q5) What is the performance of gradient descent minimizing a loss that uses an *Ansatz* for the function from X to y that matches Eq. (1), and how does it depend on the parameters of the model, the learning rate, and the initialization of the algorithm? Unlike in linear regression, where the most natural loss is convex, the bilinear nature of the present model leads to a nonconvex optimization problem with multiple minima. It is a hard endeavor to understand the behavior of gradient descent in such cases. This question is, hence, addressed numerically.

The present paper answers all these questions. In particular, (Q1) is answered analytically in Sec. II C, (Q2) via the generalized approximate message passing-rotationally invariant estimator (GAMP-RIE) algorithm in Sec. II D,

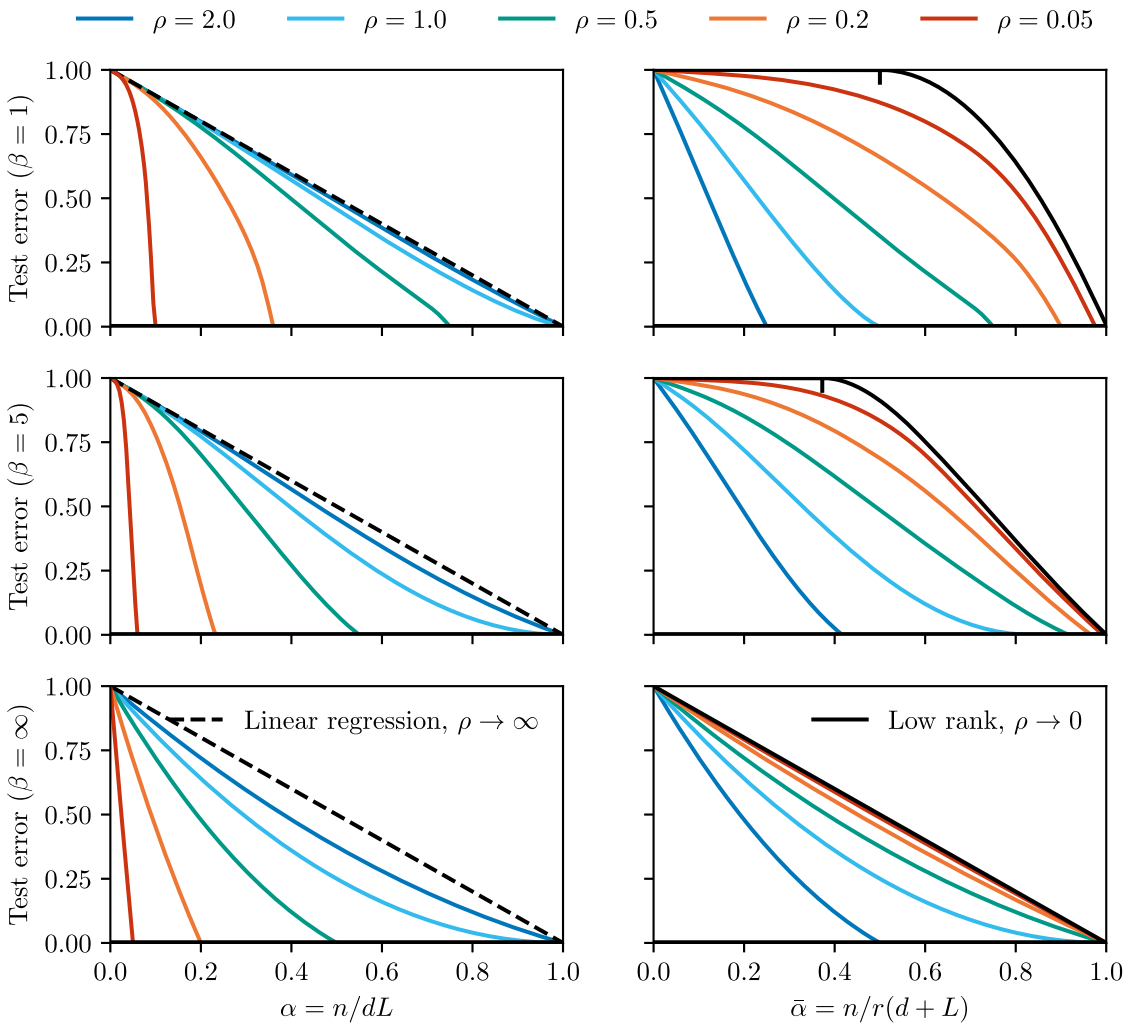

FIG. 1. Bayes-optimal test error for the BSR model with noiseless output channel ($\Delta = 0$) as a function of the sample ratio $\alpha = n/(dL)$ (left column) and of the low-width sample ratio $\bar{\alpha} = n/[r(d + L)]$ (right column). We plot a different value of the aspect ratio $\beta = \max(d, L)/\min(d, L) = 1, 5, +\infty$ for each row from top to bottom and in each panel compare several values of the width ratio $\rho = r/\min(d, L) = 0.05$, 0.1, 0.2, 0.5, 1, 2 (colored solid lines). In the left column, we also plot for comparison the performance of optimally regularized linear regression (in this case, $\lambda \to 0^+$) on the vectorized data (it does not depend on ρ and β) in the black dashed line, which corresponds also to the BO error for $\rho \to \infty$. We observe that the BO test error is always better than the linear regression test error and that it gets better and better as ρ decreases: The more structure in the distribution of the signal, i.e., the lower the width, the better one can estimate it. We also observe that the BO test error vanishes at a finite value of α, the so-called strong recovery threshold, and that this threshold is smaller than one for $\rho < 1$. In this regime, there are values of α for which the BO estimator achieves zero test error, while the linear regression estimator has a nonzero test error. The middle and bottom show the same overall phenomenology as β increases from 1 to infinity. The right column shows the same curves as a function of the low-width sample ratio $\bar{\alpha} = n/[r(d + L)]$, comparing with the already known low-width BO test error (solid black line) [42]. We observe a clear convergence to the low-width error curve as $\rho \to 0$, but we highlight that, e.g., at $\beta = 1$, the test error of the BO estimator is still quantitatively better than its low-width counterpart already at $\rho = 0.05$. Notice also that for $\rho \to 0$ the BO estimator has a weak recovery threshold at $\bar{\alpha}_{\text{weak}} = (1 + \Delta)\sqrt{\beta}/(1 + \beta)$, i.e., below it has the same performance as the zero estimator $\hat{S}_{\text{zero}}(\mathcal{D}) = 0$. As soon as $\rho > 0$, the weak recovery threshold disappears, allowing for better-than-trivial performance at all values of $\bar{\alpha}$. The weak recovery threshold is marked by a vertical black marker: Notice that for $\beta \to \infty$ the weak recovery threshold is at zero. The Bayes-optimal curves are plotted using Result 2 and Eq. (24) for extensive width and Previous Result 4 for intensive width. Linear regression is plotted using Previous Result 6.

(Q3) in the discussion in Sec. III B, (Q4) in Sec. III C and Figs. 1 and 2, and (Q5) numerically in Sec. IV.

The numerical code used to produce all presented experiments is available [43].

C. Bilinear sequence regression as the backbone of a transformer and a one-layer MLP mixer

In this section, we describe the connection between the bilinear sequence regression (BSR) model (1) and modern

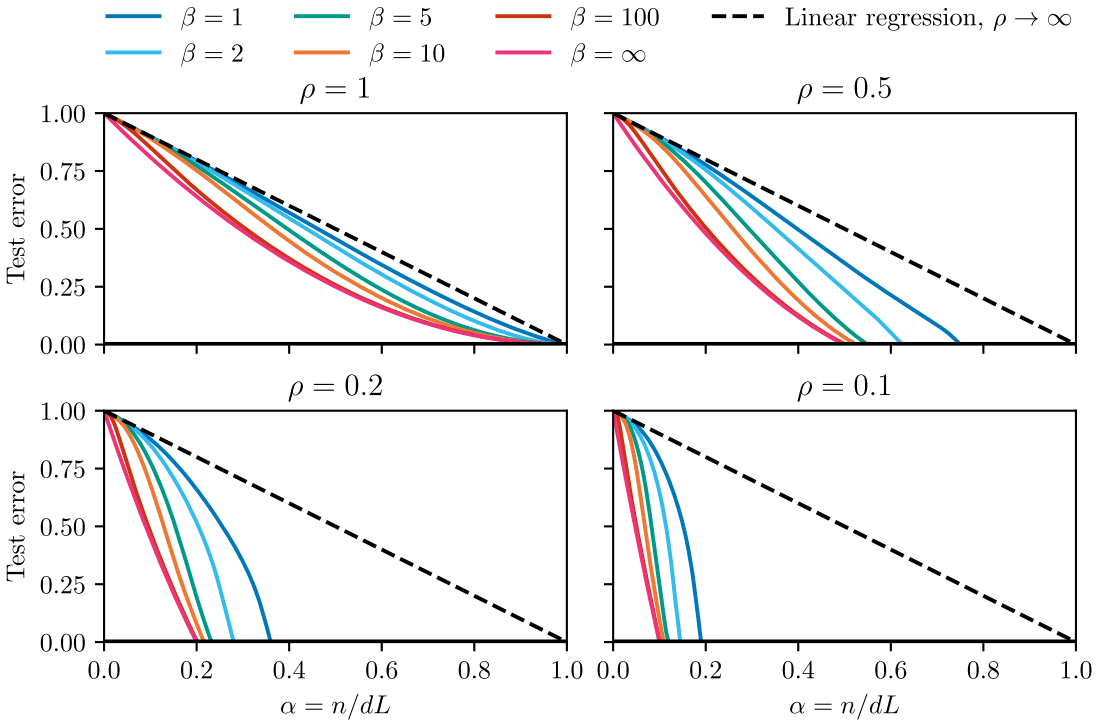

FIG. 2. Bayes-optimal test error for the BSR model with noiseless output channel ($\Delta = 0$) as a function of the sample ratio $\alpha = n/(dL)$. We plot a different value of the width ratio $\rho = r/\min(d, L) = 0.05, 0.5, 1, 2$ in each panel, and several values of the aspect ratio $\beta = \max(d, L)/\min(d, L) = 1, 5, +\infty$ (colored solid lines) in all panels. The black dashed line is the performance of optimally regularized linear regression on the vectorized data. Again, we observe that the more structured signals (larger β and smaller ρ), the better the achieved test errors. The Bayes-optimal curves are plotted using Result 2 and Eq. (24). Linear regression is plotted using Previous Result 6.

neural network architectures that achieve state-of-the-art performance in vision and natural language processing tasks.

The BSR model can be viewed as a simple, one-layer instance of the so-called MLP-Mixer architecture (where MLP stands for multilayer perceptron) [44]. MLP-Mixers operate on token sequences by alternately applying two types of MLPs: one applied *token-wise*—mixing the embedding dimensions independently for each token—and one applied *dimension-wise*—mixing information across tokens independently for each embedding coordinate. These operations are then repeated across multiple layers.

The BSR model mirrors this structure in a minimal form: the matrix (U) serves as the weights of a single-layer MLP applied token-wise, while the matrix (V) corresponds to the weights of a single-layer MLP applied embedding-coordinate-wise. In this sense, BSR is the simplest variant of an MLP-Mixer applied to token sequences, analogous to how generalized linear regression is the simplest variant of an MLP applied to vectors.

Another way to motivate the form of the bilinear sequence regression model (1) is as the bare backbone of a prototypical transformer architecture [19] designed for a supervised regression task, i.e. where the output is a continuous scalar. Let us first describe the key components of such architectures. This part can also serve to readers not familiar with these types of neural networks, to get a concise account of their key ingredients.

We consider the following prototypical architecture for a transformer that acts on sequences of tokens $X^\mu \in \mathbb{R}^{L\times d}$ and maps them to scalar labels y^μ. A toy model for a transformer would typically include a first linear embedding layer, followed by an attention layer, followed by a fully connected layer with omnipresent skip connections, followed by a final linear readout. Written mathematically, the embedding layer implements the linear mapping

$$f_{\text{embedding}}\colon \mathbb{R}^{L\times d} \to \mathbb{R}^{L\times d'} \quad \text{such that}$$
$$Z_{a\gamma} \coloneqq [f_{\text{embedding}}(X)]_{a\gamma} = \sum_{i=1}^{d} X_{ai} U_{i\gamma}, \quad (2)$$

with learnable weights $U \in \mathbb{R}^{d\times d'}$. We remark that the embedding layer usually serves to reduce the dimensionality, i.e., $d' < d$. This is because, in language data, d would correspond to the size of the dictionary, which is typically much larger than the embedding dimension. The attention layer with a skip connection implements the mapping

$$f_{\text{attention}}\colon \mathbb{R}^{L\times d'} \to \mathbb{R}^{L\times d'} \quad \text{such that } Z'_{a\gamma}$$
$$\coloneqq [f_{\text{attention}}(Z)]_{a\gamma}$$
$$= Z_{a\gamma} + \sum_{b=1}^{L} A_{ab}(Z)\left(\sum_{j=1}^{d'} Z_{bj}(w_V)_{j\gamma}\right), \quad (3)$$

where $w_V \in \mathbb{R}^{d' \times d'}$ is a learnable matrix called *value* and the dot-product attention is defined as

$$A_{ab}(X) = \mathrm{softmax}\left[\sum_{i=1}^{d'}\left(\sum_{\gamma=1}^{d'} Z_{a\gamma}(w_Q)_{\gamma i}\right)\left(\sum_{\gamma=1}^{d'} Z_{b\gamma}(w_K)_{\gamma i}\right)\right], \tag{4}$$

where $w_Q, w_K \in \mathbb{R}^{d' \times d'}$ are the learnable *query* and *key* matrices, respectively, and the softmax function maps rows of matrices into rows of normalized probabilities as

$$\mathrm{softmax}(M_{ab}) := \frac{e^{M_{ab}}}{\sum_{c=1}^{L} e^{M_{ac}}}. \tag{5}$$

A crucial aspect of the attention layer is the fact that it acts separately in sequence and embedding space. The information along the embedding dimension is processed through the key, query, and value matrices, while the attention matrix itself acts only on the sequence dimension. The subsequent one-hidden-layer fully connected network with a skip connection across the nonlinearity implements the following mapping:

$f_{\mathrm{MLP}}: \mathbb{R}^{L \times d'} \to \mathbb{R}^{L \times d'}$ such that

$$Z''_{a\gamma} := f_{\mathrm{MLP}}(Z')_{a\gamma} = Z'_{a\gamma} + \sigma\left(\sum_{j=1}^{d'} Z'_{aj}(w_F)_{j\gamma}\right), \tag{6}$$

where $w_F \in \mathbb{R}^{r \times d'}$ is a learnable matrix of weights and MLP stands for multilayer perceptron. Finally, for regression tasks, the natural last mapping is a readout:

$f_{\mathrm{readout}}: \mathbb{R}^{L \times d'} \to \mathbb{R}$ such that

$$y = f_{\mathrm{readout}}(Z'') = \sum_{a=1}^{L}\sum_{\gamma=1}^{d'} Z''_{a\gamma} V_{\gamma a}, \tag{7}$$

where $V \in \mathbb{R}^{d' \times L}$ is a learnable matrix.

A realization that is key to motivate the bilinear sequence regression model (1) is that the skip-only part of this transformer architecture [essentially setting $\sigma = 0$ and $A = 0$ in Eqs. (3) and (6)] reduces to

$$y = \sum_{a,i=1}^{L,d} X_{ai} \sum_{\gamma=1}^{d'} U_{i\gamma} V_{\gamma a}. \tag{8}$$

We are now considering how this transformer architecture aims to learn from the data produced by the BSR model (1). Notice that if we set the width of the embedding layer d' to be the width of the BSR model $d' = r$, the skip-only part gives an architecture that matches the bilinear sequence regression model (1) [with $P_{\mathrm{out}}(y|y') = \delta(y - y')$ a noiseless output channel, which is our main focus when applying our results]. We think of the skip-only part as the bare backbone of the architecture, i.e., the transformer stripped of the attention and fully connected layers. We stress that the skip-only part, thus, acts as a student model that matches the architecture of the teacher (1).

The case where $d' \neq r$ is also of interest, particularly when $d' > r$, which corresponds to the student model (8) being overparametrized relative to the target (teacher) function (1). A detailed analysis of this very rich setting is deferred to future work, as we focus here on the more easily analyzable Bayes-optimal case $d' = r$.

We also note that transformers such as the ones considered in Refs. [19,20] actually possess more features than the ones presented above: For example, they use attention and MLP layers multiple times, they use positional encoding to represent the ordering of the sequences, and the attention layer has multiple so-called "heads," meaning the size of the value matrix w_V is $(d'/h) \times (d'/h)$, with h the number of heads and the outputs Z' are concatenated together from all the heads to get back to dimension d'. The architecture presented above should be thought of as simplified.

As a matter of fact, the above rationale is valid for any model with skip connections (not only a transformer) designed to process sequences of tokens. The BSR would be the backbone of more general sequence models with skip connections designed for regression.

D. Related works

As far as we know, the bilinear sequence regression model (1) was not yet studied in the context of sequence modeling, neither as a model for synthetic data nor as an analytically tractable model for learning. It was, however, studied in the literature under the umbrella of *matrix sensing* [42,45,46] in the context of signal processing where U^* and V^* represent a hidden signal that is to be recovered.

We note here that other models where the regression parameters are matrix valued and having structure corresponding to a product of two matrices have been considered in the literature, e.g., Refs. [47–49]. None of them is exactly the same as the BSR or the matrix sensing. A more conceptual difference between such works and ours is that the models are usually proposed as *Ansätze* for functions that are then fitted to the data. Their focus is on regression tasks on real data, whereas we view the BSR model as a synthetic probabilistic model to generate data that is theoretically tractable, allowing us to compute the Bayes-optimal performance and to propose a matching algorithm. This is in line with the statistical physics works on fully connected feedforward architectures such as the perceptron and its multilayer version, for which the main contribution of statistical physics was the analysis of the optimal generalization error, in an influential line of work [9,10].

Going back to the related work on matrix sensing where each input matrix $X^\mu \in \mathbb{R}^{L \times d}$ is seen as a random linear projection operator, the task becomes to retrieve the (sometimes sparse) signals U^* and V^* from the projections. The well-known line of work represented by Refs. [45,46]

proposes and analyzes an algorithm based on a convex relaxation of the problem, where the nuclear norm (defined as the sum of singular values) of the matrix $S^* = U^* V^*/\sqrt{r}$ is minimized. Similarly to other convex relaxations, this algorithm, however, reaches suboptimal performance with respect to the Bayes-optimal estimator. Another line of work considered the matrix sensing problem solved via gradient descent in an overparametrized setting: Their findings suggested that gradient descent with infinitesimal initialization could have implicit regularization toward the minimum nuclear norm [50,51].

The Bayes-optimal performance for matrix sensing, which is an instance of model (1), was addressed in Ref. [42] using approximate message-passing algorithms and their state evolution. Their results provide an exact characterization of the Bayes-optimal performance in the low-width limit where $L, d \to \infty$ with $L/d = \mathcal{O}(1)$, $n/d = \mathcal{O}(1)$, and crucially the width remains of the order of $r = \mathcal{O}(1)$. Reference [42] also claims to provide an asymptotically exact characterization of the case where the width r is extensive, i.e., $r/d = \mathcal{O}(1)$, but this claim is based on incorrect assumptions, as was later found in the closely related problem of extensive-rank (the width parameter r plays the role of the rank) matrix factorization in the line of work [52–58]. The characterization of the Bayes-optimal performance for the extensive width $r/d = \mathcal{O}(1)$ thus remained open: This is the first main technical contribution of the present paper, together with the proposition of an approximate message-passing algorithm that reaches, in the studied cases, the optimal performance in polynomial time.

Another model studied in the literature that is technically related to ours would correspond to the width $r = 1$, with each sample X^μ being a symmetric matrix and $U = V^T$. Such a model has been studied both for random labels y [59–61] and with labels generated by a teacher model [62].

On a technical level, our work builds upon the works in Refs. [55,63]. More specifically, we extend the recent analysis of Ref. [63], that treats a model that can technically be seen as a symmetric version of the BSR, where one imposes $U = V^T$ and X_{ai} is a symmetric matrix. Our derivation further relies on the optimal performance in the denoising of extensive-rank nonsymmetric matrices [55].

Regarding the dynamics of gradient descent in matrix sensing, a symmetric version where $U^* = V^*$ has been considered in Ref. [64]. There, the authors show that all minima of the natural square loss for the problem achieve perfect reconstruction in the high-dimensional limit as long as the number of samples is of the order of $n \geq C \cdot dr$, where r is the width, but their approach does not have access to the optimal constant $C > 0$. A similar model has also been studied in Ref. [65], where the authors show that flat minima of a natural loss generalize well if $n > Cr(d + L)$, but again they do not have access to the optimal constant C. In comparison, we consider generic values of the width, in particular, in Sec. IV the extensive one $r = \Theta(d)$, and show by combining the Bayes-optimal analysis and numerical evidence on gradient descent that an averaged version of gradient descent generalizes perfectly as soon as this is information-theoretically possible, i.e., as soon as the Bayes-optimal estimator generalizes perfectly. We provide both the scaling $n = \mathcal{O}(dL)$ of this threshold, as well as the constant (see Result 5). Finally, let us mention that the dynamics of alternate minimization in a related model has been considered in Ref. [66].

II. MAIN TECHNICAL RESULTS

A. Generalized bilinear sequence regression model

From a technical point of view, it is advantageous to think about the BSR model (1) in a slightly more general way. We consider the model

$$y^\mu \sim P_{\text{out}}(\cdot|h^\mu) \quad \text{with} \quad h^\mu = \frac{1}{\sqrt{Ld}} \sum_{a,i=1}^{L,d} X^\mu_{ai} S^*_{ia}, \tag{9}$$

where $S^* \in \mathbb{R}^{d\times L}$ is a hidden or latent weight matrix, $X^\mu \in \mathbb{R}^{L\times d}$ for $\mu = 1, \ldots, n$ are input sequences composed by L tokens, each a vector in dimension d, and $y^\mu \in \mathbb{R}$ for $\mu = 1, \ldots, n$ are the associated scalar labels.

We assume the data to be Gaussian, i.e., $X^\mu_{ai} \sim \mathcal{N}(0, 1)$ independently for each value of (μ, a, i). We believe that this assumption can be relaxed within the context of Gaussian universality results (see, e.g., Refs. [67,68]) without altering the main points of our analysis: We will pursue this generalization in future works. The labels y^μ are generated through a possibly probabilistic output channel P_{out}, conditioned on the value of the scalar preactivations h^μ.

The BSR model of Eq. (1) corresponds to the specific case of factorized Gaussian prior on S^*, i.e.,

$$S^*_{ia} = \frac{1}{\sqrt{r}} \sum_{\gamma=1}^{r} U^*_{i\gamma} V^*_{\gamma a}, \tag{10}$$

with $U^* \in \mathbb{R}^{d\times r}$ and $V^* \in \mathbb{R}^{r\times L}$ matrices of i.i.d. standard Gaussian entries. This distribution introduces nontrivial dependencies between each entry of S^*, coupling the token and embedding dimensions.

Notice that the formulation of the model with structured S^* in Eq. (9) is strictly more general than the BSR and includes the factorized form of $S^\star$ in Eq. (1) as a special case. In particular, all our results on Bayes-optimal learning in Secs. II and III apply directly to the BSR (1) and take into account fully its factorized structure (through the prior on S^*).

We consider the model (9) in the high-dimensional setting, where $L, d \to \infty$ with fixed ratio $L = \Theta(d)$. In particular, we define

$$\beta := \frac{\max(L, d)}{\min(L, d)} \geq 1,$$

which remains finite as $L, d \to \infty$. β measures the aspect ratio of the matrices X^μ and S^*, irrespective of which among L and d is bigger. In general, the scaling for the number of samples n in the high-dimensional limit, i.e., the number of samples needed to at least partially retrieve the signal S^*, depends on the choice of its distribution (and it usually scales with the total amount of unknowns included in S^*).

The main novel results of this paper relate to the so-called *extensive-width* limit where the width r is also proportional to the dimensionality. We, thus, define a width-related parameter

$$\rho \coloneqq \frac{r}{\min(d, L)}$$

that will remain finite in the high-dimensional limit of the model. Note that the rank of the matrix S^* is constrained to be at most r, and, generically, when the width r is extensive, the rank of the matrix S^* is also extensive. We, thus, call the corresponding limit the extensive-width or extensive-rank case. For $\rho \to \infty$, we expect by the central limit theorem that the distribution of S^* approaches the one of a matrix with i.i.d. standard Gaussian entries. In the regime where r scales linearly with d, L, we see that the correct sample scale to observe nontrivial retrieval of the ground-truth signal is $n = \mathcal{O}(dL)$. We define α as

$$\alpha \coloneqq \frac{n}{dL}$$

with $\alpha > 0$ finite in the high-dimensional limit. The low-width limit where ρ is small down to the width $r = \mathcal{O}(1)$ is considered for comparison in Sec. II E, building on the result of Ref. [42].

We stress that our main technical results apply to a much wider class of distributions for S^* (priors), more specifically, rotationally invariant distributions P_0, as long they admit a well-defined limiting spectral density in the high-dimensional limit. Without loss of generality, we assume that P_0 is normalized as

$$Q_* \coloneqq \lim_{d,L\to\infty} \mathbb{E}_{S\sim P_0} \frac{1}{dL} \sum_{a,i=1}^{L,d} (S^*_{ia})^2 = 1. \tag{11}$$

In informal terms, this ensures that the entries of $S^* \sim P_0$ are on average of the order of $\mathcal{O}(1)$ in the high-dimensional limit. Rotational invariance means that $P_0(S) = P_0(O_1 S O_2)$ for any pair of rotation matrices O_1, O_2 in dimensions, respectively, d and L, while having a well-defined limiting (symmetrized) spectral density means that

$$\lim_{d,L\to\infty} \frac{1}{2\ \min(d, L)} \sum_{i=1}^{\min(d,L)} [\delta[x - \sigma_i(S)] + \delta[x + \sigma_i(S)]] = \mu_S(x), \tag{12}$$

for some density $\mu_S(x)$, where $\sigma_i(S)$ are the singular values of $S/\sqrt[4]{dL}$ [the normalization ensures that $\sigma_i(S)$ remains of the order of $\mathcal{O}(1)$ in the high-dimensional limit].

B. Bayes-optimal estimation

We are interested in predicting the information-theoretical limits for retrieving the hidden parameters S^* from a typical dataset $\mathcal{D}_n = \{(X^\mu, y^\mu\}_{\mu=1}^n$. To achieve this goal, we study the Bayes-optimal (BO) estimator and its performance. We are interested in two performance metrics. The first one is the averaged *test error* (also called generalization error), which is the error obtained when predicting the label on a newly sampled $(\underline{X}, \underline{y})$ data pair, and is defined as

$$E_{\text{gen}}(\hat{S}) \coloneqq \mathbb{E}_{(\underline{X},\underline{y})} \mathbb{E}_{S^*} \mathbb{E}_{\mathcal{D}|S^*} \times \left(\underline{y} - P_{\text{out}}\left(\frac{1}{\sqrt{Ld}} \sum_{ai} \underline{X}_{ai} \hat{S}(\mathcal{D})^\top_{ai} \right) \right)^2, \tag{13}$$

where $\hat{S}$ is a generic estimator [a function mapping the dataset $\mathcal{D}$ to a candidate set of weights $\hat{S}(\mathcal{D})$] and by $P_{\text{out}}(x)$ we mean the (possibly random) output of the output channel conditioned on x. Our second metric is the averaged *estimation error*, which is the discrepancy between the estimated weights and the true weights, and is defined as

$$E_{\text{est}}(\hat{S}) \coloneqq \frac{1}{dL} \mathbb{E}_{S^*} \mathbb{E}_{\mathcal{D}|S^*} \|S^* - \hat{S}(\mathcal{D})\|^2. \tag{14}$$

The BO estimator with respect to either one of the two performance metrics is defined as the estimator function $\hat{S}: \mathcal{D} \to \hat{S}(\mathcal{D})$ minimizing the respective metric. It is a very classical result that the BO estimator with respect to the test error is given by

$$\hat{S}_{\text{BO,gen}}(\mathcal{D}) = \mathbb{E}_{S\sim P(S|\mathcal{D})} \mathbb{E}_{(\underline{X},\underline{y})|S} [\underline{y}\, \underline{X}], \tag{15}$$

where $(\underline{X}, \underline{y})$ is a newly sampled input-output pair conditioned on a set of weights S and $P(S|\mathcal{D})$ is the posterior distribution, i.e., the probability that a candidate signal S has been used to generate the dataset $\mathcal{D}$, which can be expressed (through Bayes' theorem) using the prior distribution and the output channel:

$$P(S|\mathcal{D}) \propto P_0(S) \prod_\mu P_{\text{out}}\left(y^\mu \middle| \frac{1}{\sqrt{Ld}} \sum_{ai} X_{ai} S^\top_{ai} \right). \tag{16}$$

The BO estimator with respect to the estimation error instead equals the posterior mean, i.e.,

$$\hat{S}_{\text{BO,est}}(\mathcal{D}) = \mathbb{E}_{S\sim P(S|\mathcal{D})}[S]. \tag{17}$$

Both of these expressions are standard (see, for example, Ref. [69]) and can be recovered by taking the derivative with respect to the estimator $\hat{S}$ in the errors' definitions and setting it equal to zero.

We remark that for generic P_{out} the two BO estimators may differ. However, in the case of Gaussian inputs X and Gaussian label noise [i.e., Gaussian output channel $P_{\text{out}}(\cdot|h) = N(h, \Delta)$, which includes noiseless observations], one can show that

$$E_{\text{gen}}(\hat{S}) = E_{\text{est}}(\hat{S}) + \Delta, \tag{18}$$

meaning that the estimation and test BO estimators coincide and that the respective values of the errors differ only by a constant additive shift quantifying the amount of label noise. In the following, we focus on this special case, and for this reason we here restrict our analysis to the BO estimator with respect to the estimation error. Finally, we remark that the BO estimation error is also called minimal mean square error (MMSE) in the literature.

We recall that the estimation error of any estimator $\hat{S}$ is given by

$$E_{\text{est}}(\hat{S}) = \frac{1}{dL}\mathbb{E}_{S^*}\mathbb{E}_{\mathcal{D}|S^*}\|S^* - \hat{S}(\mathcal{D})\|^2 = 1 + q(\hat{S}) - 2m(\hat{S}), \tag{19}$$

where we define (using conventions originating in statistical physics [70]) the average "magnetization" $m(\hat{S})$ and "overlap" $q(\hat{S})$ of the estimator $\hat{S}$ as

$$\begin{aligned} m(\hat{S}) &\coloneqq \frac{1}{dL}\mathbb{E}_{S^*}\mathbb{E}_{\mathcal{D}|S^*}\text{Tr}[(S^*)^T\hat{S}(\mathcal{D})] \quad \text{and} \\ q(\hat{S}) &\coloneqq \frac{1}{dL}\mathbb{E}_{S^*}\mathbb{E}_{\mathcal{D}|S^*}\text{Tr}[\hat{S}^T(\mathcal{D})\hat{S}(\mathcal{D})], \end{aligned} \tag{20}$$

respectively, and use that the prior is normalized to have self-overlap $Q_* = 1$; see Eq. (11). For the BO estimator, Nishimori's identities [70] imply $m_{\text{BO}} = q_{\text{BO}}$, from which we obtain

$$\text{MMSE} = E_{\text{est}}(\hat{S}_{\text{BO}}) = 1 - q_{\text{BO}}, \tag{21}$$

and, moreover, q_{BO} reduces to the overlap between two independent samples of the posterior distribution $P(S|\mathcal{D})$, i.e.,

$$q_{\text{BO}} = \frac{1}{dL}\mathbb{E}_{\mathcal{D}}\mathbb{E}_{S_1,S_2\sim P(S|\mathcal{D})}\text{Tr}(S_1^T S_2). \tag{22}$$

We stress that, in the case of a generic non-Gaussian P_{out}, the BO test error can differ from the BO estimation error. Yet, it can still be computed as a function of the same order parameter q_{BO}, which is the quantity for which we provide a precise asymptotic analytical treatment in the following sections.

Finally, we notice that the estimation error for the Gibbs sampler of the posterior, i.e., the expected estimation error of a uniform sample of the posterior, satisfies

$$\begin{aligned} &\mathbb{E}_{S^*}\mathbb{E}_{\mathcal{D}|S^*}\mathbb{E}_{S\sim P(S|\mathcal{D})}\frac{1}{dL}\|S^* - S\|^2 \\ &= 1 - 2\frac{1}{dL}\mathbb{E}_{S^*}\mathbb{E}_{\mathcal{D}|S^*}\mathbb{E}_{S\sim P(S|\mathcal{D})}\text{Tr}[(S^*)^T S] \\ &\quad + \frac{1}{dL}\mathbb{E}_{S^*}\mathbb{E}_{\mathcal{D}|S^*}\mathbb{E}_{S\sim P(S|\mathcal{D})}\text{Tr}(S^T S) \\ &= 1 - 2m_{\text{BO}} + \frac{1}{dL}\mathbb{E}_{S^*}\mathbb{E}_{\mathcal{D}|S^*}\mathbb{E}_{S\sim P(S|\mathcal{D})}\text{Tr}(S^T S) \\ &= 2(1 - q_{\text{BO}}), \end{aligned} \tag{23}$$

where in the last step we use Nishimori's identities, stating that a sample from the posterior is statistically equivalent to the ground truth. Notice the crucial difference between the BO estimator and a sample of the posterior: In the overlap term $q(S)$, the BO estimator uses the overlap between two independent samples of the posterior, while the Gibbs sampler uses the self-overlap of a single sample of the posterior. This implies that the estimation error of the Gibbs sampler, on average, is twice the BO estimation error.

C. Optimal error for extensive-width BSR

In this section, we present our novel results concerning the asymptotic characterisation of the BO estimator and of the associated optimal estimation error in the high-dimensional limit for all output channels. These results provide the answer to (Q1) posed in the Introduction. We start with the general case of arbitrary rotationally invariant priors on S and follow up with the results for the extensive-width BSR model as a consequence.

Result 1 (MMSE for generalized BSR)—Consider any rotationally invariant prior $P_0(S)$ such that the empirical symmetrized singular value density of $\underline{S} = S/\sqrt[4]{dL}$ with $S \sim P_0$ converges to a well-defined probability distribution for $d \to \infty$ with $\beta = \{[\max(L,d)]/[\min(L,d)]\} \geq 1$ finite. Call $\hat{\mu}_{\underline{Y}}$ the symmetrized singular value density of $\underline{Y} = \underline{S} + \sqrt{\delta}\underline{Z}$, where $\underline{Z}$ is a matrix of i.i.d. Gaussian entries $N(0, 1/\sqrt[4]{dL})$ and $\delta > 0$. Define the sample ratio as $\alpha = n/(dL)$.

Then, $\lim_{d\to\infty}\text{MMSE} = 1 - q$, where $(q, \hat{q}) \in \mathbb{R}^2$ are a solution to the nonlinear system of equations

$$\begin{aligned} q &= 1 - \frac{1}{\hat{q}} + \frac{2}{\beta^{3/2}\hat{q}^2}\int dx\hat{\mu}_{\underline{Y}}(x)\left[\frac{(\beta-1)^2}{2x^2} + \frac{2\pi^2}{3}\hat{\mu}_{\underline{Y}}(x)^2\right], \\ \hat{q} &= \frac{\alpha}{q}\int Dzdy\frac{(\partial_z I_{\text{out}}(z,y;q))^2}{I_{\text{out}}(z,y;q)}, \end{aligned} \tag{24}$$

where

$$I_{\text{out}}(z,y;q) \coloneqq \int \frac{dhd\hat{h}}{2\pi} P_{\text{out}}\left(y|h\right) \times \exp\left(-\frac{1-q}{2}\hat{h}^2 + (\sqrt{q}z+h)i\hat{h}\right). \quad (25)$$

The integral over y is intended over the image of P_{out}. The dashed integral is regularized as specified in Appendix A in Ref. [55]. In the case where Eq. (24) admits multiple solutions, one should pick the one maximizing an associated free entropy, whose expression we provide in Appendix B; see Eq. (B49). Notice that for Gaussian output channels $P_{\text{out}}(\cdot|h) = N(h,\Delta)$, the equation for $\hat{q}$ simplifies to

$$\hat{q} = \frac{\alpha}{\Delta + 1 - q}. \quad (26)$$

Result 1 can be obtained by performing a nonrigorous (hence the phrasing "result" rather than theorem) but exact computation based on replica theory, detailed in Appendix B. One writes the partition function associated to the posterior distribution, computes it through replica theory, and obtains a saddle-point characterization for the overlap order parameter, i.e., Eq. (24). The main technical novelty in Result 1 is that, for rotationally invariant priors with extensive rank, a standard factorization passage of the computation fails (contrary to what has been claimed by Ref. [42], where the authors' analysis is not correct in the extensive-rank regime; see the references cited in the Introduction on the matter). However, we can now overcome this difficulty by adapting recent results for Bayes optimal extensive-rank matrix denoising [53,55].

The main difficulty in solving Eq. (24) is the computation of $\hat{\mu}_{\underline{Y}}(x)$, which for generic rotationally invariant priors is *a priori* nontrivial. For the factorized Gaussian prior on S, corresponding to the BSR model, this difficulty can be overcome. Details on how to compute efficiently $\hat{\mu}_{\underline{Y}}(x)$ in this specific case can be found in Section III. 3 and Appendix F in Ref. [55], adapting previous work by Ref. [71]. The explicit formula for the MMSE for the BSR model is then as follows.

Result 2 (MMSE for the bilinear sequence regression model, consequence of Result 1)— For the bilinear sequence regression model, $\text{MMSE} = 1 - q$, where q satisfies Eq. (24). The spectral density $\hat{\mu}_{\underline{Y}}(x)$ to use in Eq. (24) is characterized as follows. Define the Stieltjes transform of $\hat{\mu}_{\underline{Y}}(x)$ as

$$g_{\underline{Y}}(z) \coloneqq \int dx \frac{\hat{\mu}_{\underline{Y}}(x)}{z-x}. \quad (27)$$

Then,

$$g_{\underline{Y}}(z) = z g_{\underline{Y}^2}(z^2), \quad (28)$$

where $g_{\underline{Y}^2}$ is the Stieltjes transform of the asymptotic spectral density of $\underline{Y}\underline{Y}^T$ (if $d \le L$) or of $\underline{Y}^T\underline{Y}$ (if $d > L$). Moreover, $g_{\underline{Y}^2}(z^2)$ is the root with largest imaginary part of the quartic polynomial $\sum_{a=0}^{4} a_k X^k$, where

$$\begin{aligned} a_0 &= -\psi^3, \\ a_1 &= \psi\{\zeta(\psi-\phi) + \psi[\eta(\phi-\psi) + \psi z^2)]\}, \\ a_2 &= -\zeta^2(\phi-\psi)^2 + \zeta[\eta(\phi-\psi)^2 + \psi z^2(2\phi-\psi)] \\ &\quad - \eta\psi^2 z^2\phi, \\ a_3 &= -\zeta z^2\phi(2\zeta\psi - 2\zeta\phi - 2\eta\psi + 2\eta\phi + \psi z^2), \\ a_4 &= \zeta z^4\phi^2(\eta-\zeta), \end{aligned} \quad (29)$$

where $\phi = \rho/\beta$, $\psi = \rho$, $\eta = (1+\delta)\sqrt{\beta}$, and $\zeta = \sqrt{\beta}$. Finally, the symmetrized singular value density can be recovered as

$$\hat{\mu}_{\underline{Y}}(x) = \frac{1}{\pi}\lim_{\epsilon\to 0^+} \text{Im}\, g_{\underline{Y}}(x - i\epsilon). \quad (30)$$

We start by highlighting that in this result the factorized nature of the prior (10) enters explicitly through the parameters ϕ and ψ, both dependent on the width parameter ρ of the BSR model. Even more, the full form of the spectral density $\hat{\mu}_{\underline{Y}}(x)$ actually depends on the choice of prior (10), in the sense that a different prior with the same width parameter ρ, but different overall structure, will have, in general, a different spectral density.

Notice also that, whenever $\rho < 1$, the symmetrized singular value density of $\underline{S}$ has a delta contribution at the origin with mass $(1-\rho)$, while the nontrivial bulks of the distribution (positively and negatively supported) each have mass $\rho/2$. Moreover, while the asymptotic spectral distribution of $\underline{S}$ is singular due to the delta peak, the one of its noisy version $\underline{Y} = \underline{S} + \sqrt{\delta}\underline{Z}$ can be shown to possess a smooth density for all values of $\delta > 0$ [72].

Finally, in the limit of large $\beta \gg 1$ (i.e., when $L \gg d$ or $L \ll d$) and of factorized Gaussian priors [see Eq. (10)], we are able to simplify Result 1 significantly, bypassing the computation of the nontrivial limiting spectral density. We present this result in Appendix C 4.

D. Message-passing algorithm

For general rotationally invariant priors P_0, we are also able to derive an algorithm that, in the high-dimensional limit, achieves the MMSE (unless computationally hard phases arise, which we have not observed in the present problem; see the discussion in the remainder of the section), giving an efficient implementation of the posterior mean, an often intractable problem. This provides the answer to question (Q2) posed in the Introduction. The algorithm we present is a variant of the well-known generalized approximate message-passing (GAMP) algorithm [73], with an

additional matrix denoising step, similarly to the algorithm designed in Ref. [63].

Result 3 (GAMP-RIE for rotationally invariant priors)— Consider the same setting as in Result 1. Define

(i) $g_{\rm out}$ as

$$g_{\rm out}(y,\omega,V) := \frac{1}{V}\frac{\int dz(z-\omega)e^{-(z-\omega)^2/2V}P_{\rm out}(y|z)}{\int dz e^{-(z-\omega)^2/2V}P_{\rm out}(y|z)}, \tag{31}$$

which reduces to $g_{\rm out}(y,\omega,V) = (y-\omega)/(\Delta+V)$ for the Gaussian label noise output channel with variance Δ [i.e., $P_{\rm out}(\cdot|h) = \mathcal{N}(h,\Delta)$];

(ii) $f_{\rm RIE}(\cdot,\delta)$ as the BO rectangular matrix Gaussian denoiser (Results 1 and 2 in Ref. [55]) with noise-to-signal ratio δ and $\mathrm{MMSE}_{\rm denoising}(\delta)$ the corresponding MMSE. Explicitly, if $R = U\Lambda V$ is the singular value decomposition of a matrix R, with $\Lambda = \mathrm{diag}(\lambda_1,\ldots,\lambda_t)$, the denoiser acts on each separate singular value as

$$f_{\rm RIE}(R = U\Lambda V,\delta) = U\mathrm{diag}\left(\lambda_i - \frac{2\delta}{\sqrt{\beta}}\left[\frac{\beta-1}{2\lambda_i} + \fint dx\frac{\hat{\mu}_{\underline{Y}}(x)}{\lambda_i - x}\right]\right)_{i=1,\ldots,t} V, \tag{32}$$

with $\hat{\mu}_{\underline{Y}}$ as defined in Result 1, and

$$\mathrm{MMSE}_{\rm denoising}(\delta) = \delta - \frac{\delta^2}{\sqrt{\beta}}\left[\frac{(\beta-1)^2}{\beta}\fint dx\frac{\hat{\mu}_{\underline{Y}}(x)}{x^2} + \frac{4\pi^2}{3\beta}\fint dx\hat{\mu}_{\underline{Y}}(x)^3\right] \tag{33}$$

is the associated BO mean square error. The dashed integral is regularized as specified in Appendix A in Ref. [55].

Then, Algorithm 1 achieves at convergence an overlap $q = [1/(dL)]\mathbb{E}_{S^*}\mathbb{E}_{\mathcal{D}|S^*}\mathrm{Tr}[\hat{S}^T(\mathcal{D})\hat{S}(\mathcal{D})]$ satisfying Eq. (24) in the high-dimensional limit. For the specific case of the BSR model, we can analytically compute the spectral density $\hat{\mu}_{\underline{Y}}$, as detailed in Result 2.

Notice that Algorithm 1 depends explicitly on the width parameter ρ and on the full factorized form of the prior of the BSR model through the spectral density $\hat{\mu}_{\underline{Y}}$, as detailed in Result 2.

Let us now justify the last claim of Result 3, concerning the performance achieved by Algorithm 1. It builds on the connection between approximate-message-passing algorithms and replica theory, which is a well-established result in the theory of generalized linear models [70]. This connection stems from the fact that AMP algorithms can be tracked in high dimension by an iterative update equation for the order parameters, such as the overlap q and its conjugate parameter $\hat{q}$, called state evolution [74]. One can show that the state evolution equations for Algorithm 1 are

Algorithm 1. GAMP-RIE for the BSR model with extensive width.

Result: An estimator $\hat{S}_{\rm AMP}$
Input: Dataset $\{(X^\mu, y^\mu)\}_{\mu=1}^n$;
Initialize $S_{t=0} \sim P_0$, set $\underline{S}_{t=0} = S_{t=0}/\sqrt[4]{dL}$ and $c_{t=0} = 1, \omega_{t=0} = 1\times[1,\ldots,1]^T \in \mathbb{R}^N, V_{t=0} = 1$;
Rescale $\tilde{X} = X/\sqrt[4]{Ld}$;
while not converging **do**
• *Estimation of the variance and mean of* $\mathrm{Tr}(\underline{S}_t^T\tilde{X}^\mu)$;
$V_t = c_t$ and $\omega_t^\mu = \mathrm{Tr}(\underline{S}_t^T\tilde{X}^\mu) - g_{\rm out}(y^\mu,\omega_{t-1}^\mu,V_{t-1})V_t$;
• *Variance and mean of* $\underline{S}$ *estimated from the channel observations*
$A_t = \frac{\alpha}{n}\sum_{\mu=1}^n g_{\rm out}(y^\mu,\omega_t^\mu,V_t)^2$ and $R_t = \underline{S}_t + \frac{1}{A_t\sqrt{dL}}\sum_{\mu=1}^n g_{\rm out}(y^\mu,\omega_t^\mu,V_t)\tilde{X}^\mu$;
• *Update of the estimation of* $S^\star$ *with the prior information*;
$\underline{S}_{t+1} = f_{\rm RIE}(R_t,\frac{1}{A_t})$ and $c_{t+1} = \mathrm{MMSE}_{\rm denoising}\left(\frac{1}{A_t}\right)$;
$t = t+1$;
end
Rescale $\hat{S}_{\rm AMP} = \sqrt[4]{Ld}\underline{S}_t$.

$$\hat{q}_t = \frac{\alpha}{q_t}\int Dzdy\,\frac{(\partial_z I_{\text{out}}(z,y;q_t))^2}{I_{\text{out}}(z,y;q_t)},$$
$$q_{t+1} = 1 - \frac{1}{\hat{q}_t} + \frac{2}{\beta^{3/2}\hat{q}_t^2}\int dx\hat{\mu}_{\underline{Y}_t}(x)\left[\frac{(\beta-1)^2}{2x^2} + \frac{2\pi^2}{3}\hat{\mu}_{\underline{Y}_t}(x)^2\right] = 1 - \text{MMSE}_{\text{denoising}}\left(\frac{1}{\hat{q}_t}\right), \tag{34}$$

where $\underline{Y}_t = \underline{S} + \sqrt{1/\hat{q}_t}\underline{Z}$, the last equality is justified in Eq. (B46), and

$$q^t = \text{Tr}(\underline{S}_*^T\underline{S}_t). \tag{35}$$

This is just a particular iterative scheme for Eq. (24), justifying our claim that Algorithm 1 satisfies Eq. (24) at convergence.

To derive the state evolution equation (34), one can follow directly Secs. 6.3 and 6.4 in Ref. [70] (see also Ref. [63] for a symmetric version of the same GAMP-RIE algorithm). One considers the relaxed-BP algorithm (Algorithm 1 in Ref. [70]) with the substitution $f_a \to f_{\text{RIE}}$ and $f_v \to \text{MMSE}_{\text{denoising}}$—notice that in Ref. [70] these functions are applied coordinatewise on R_t, while here they are applied directly to the full matrix, as in Ref. [70] they consider factorized distributions P_0. Then, one can follow independently Sec. VI. 3. 2 in Ref. [70] to derive the GAMP-RIE algorithm (our Algorithm 1) as an asymptotic approximation of the r-BP algorithm and Sec. VI. 4. 1 in Ref. [70] to derive the state evolution equations for the GAMP-RIE algorithm from the r-BP algorithm. Both derivations can be followed step by step with the mentioned substitutions. The first of the state evolution equations is then found directly. For the second state evolution equation, one notices that

$$R_t \overset{d}{=} \underline{S}_* + \frac{1}{A_t}\underline{Z} \tag{36}$$

in distribution, where $\underline{S}_*$ is the ground truth and $\underline{Z}$ is an i.i.d. Gaussian noise [both normalized to have $\mathcal{O}(1)$ singular values]. Thus, $\underline{S}_{t+1}$ is the BO estimate of $\underline{S}_*$, and the associated MSE satisfies $c_{t+1} = \text{MMSE}_{\text{denoising}}(1/A_t) = 1 - q^{t+1}$, where q^{t+1} is the overlap between the iterate $\hat{S}_{t+1}$ and the ground truth $\underline{S}_*$, leading to the second, nontrivial, state evolution equation. A more general treatment of state evolution for GAMP with nonseparable denoisers is given in Refs. [75,76].

We stress that, depending on the choice of prior and output channel, Algorithm 1 may not achieve the BO performance when initialized at zero overlap with the ground truth. This is usually referred to as a computational-to-statistical gap [70,73,74], denoting a region where we have MMSE < 1 (i.e., the BO estimator retrieves some information about the ground truth), while the AMP algorithm is stuck at a larger, possibly trivial MMSE. This can happen if among the solutions of Eq. (24) there are multiple local maximizers of the associated free entropy (B49). In that case, GAMP-RIE will find the local maximizer with smallest value q, while the BO performance will be given by the global maximizer. A gap arises if the local maximizer with smallest value q is not the global maximizer.

For our case study, i.e., the BSR model with Gaussian label noise, we do not observe any such gap. In Appendix A, we verify that the free entropy (B49) has a unique local maximum and that our solver of Eq. (24) finds that maximum, for a selection of values of β, ρ, and α. While not an analytical justification, our observations do not provide any hint to the existence of such a computational-to-statistical gap, allowing us to conjecture that no hard phase is present for this specific choice of prior. Other rotationally invariant priors may still, however, exhibit computational gaps, but we leave such a study for future work.

We provide numerical experiments on the GAMP-RIE algorithm in Sec. III A, Fig. 3. Notice that to improve the convergence of the algorithm we use a damped version, where the iteration for the variance and mean of $\text{Tr}(\underline{S}_t^T\tilde{X}^\mu)$, i.e., V_t and ω_t, becomes (e.g., for V) $V_t = (1-\gamma)c_t + \gamma V_{t-1}$ for some damping factor $0 < \gamma < 1$. We also needed to fine-tune the scale of the initialization $c_{t=0} = \zeta, \omega_{t=0} = \zeta \times [1,\ldots,1]^T \in \mathbb{R}^N, V_{t=0} = \zeta$ to $\zeta \approx 20$ to obtain satisfactory results. We discuss how we tuned γ and ζ in Appendix A.

Notice that Algorithm 1 provides the Bayes-optimal versions to the class of matrix denoisers considered in Ref. [77], in which the authors approximate the RIE denoiser with different suboptimal denoisers.

E. MMSE in the low-width case

We recall here the results of Ref. [42] for the MMSE in the low-width case with our notations. In Appendix B, we rederive the result of Ref. [42] in the more general case of correlated low-width priors. In Appendix C 3, we also derive the large β limit of this result.

Previous Result 4 (MMSE for low-width BSR model [42])—Consider the factorized Gaussian prior on S (10) in the low-width high-dimensional limit, i.e., $d \to \infty$ with $\beta = \max(d,L)/\min(d,L)$ fixed and $r = \mathcal{O}(1)$. Define the sample ratio as

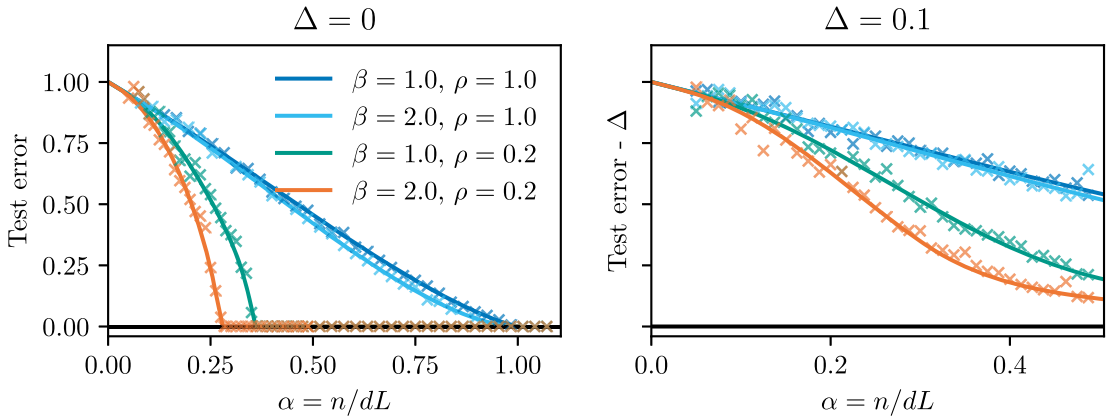

FIG. 3. Comparison between the BO test error and the test error of GAMP-RIE (Algorithm 1) for two choices of the aspect ratio $\beta = \max(d, L)/\min(d, L) = 1, 2$ and the width ratio $\rho = r/\min(d, L) = 0.2, 1$ in both the noiseless $\Delta = 0$ (left) and noisy $\Delta = 0.1$ (right) cases. Solid lines are the theoretical prediction from Eq. (24). The crosses represent numerical experiments for the test error measured after iterating GAMP-RIE until convergence, on instances of size $\min(d, L) = 100$, with initialization from the prior distribution. Each point is a run over a single realization of the data and ground truth. The Bayes-optimal curves are plotted using Result 2 and Eq. (24). The performance of GAMP-RIE is given by Eq. (14) applied to the output of Algorithm 1.

$$\bar{\alpha} = \frac{n}{r(d+L)} = \frac{\beta}{\rho(1+\beta)}\alpha. \tag{37}$$

Then, MMSE $= 1 - q$, where q is a solution to the nonlinear system of equations

$$\begin{aligned} q &= g_1 g_2, \\ g_1 &= \frac{(\beta+1)^2\hat{q}^2 - \beta}{(\beta+1)\hat{q}(\beta\hat{q}+\hat{q}+1)}, \\ g_2 &= \frac{(\beta+1)^2\hat{q}^2 - \beta}{(\beta+1)\hat{q}(\beta\hat{q}+\hat{q}+\beta)}, \\ \hat{q} &= \frac{\bar{\alpha}}{q}\int Dzdy \frac{(\partial_z I_{\text{out}}(z, y; q))^2}{I_{\text{out}}(z, y; q)}, \end{aligned} \tag{38}$$

with the same I_{out} as in Result 1.

Notice importantly that the result depends on the number of samples n and the width r only through the ratio $\bar{\alpha}$. Consequently, the dependence of the MMSE on the width r is very simple. Compared to this, the MMSE in the extensive-width regime depends on the widths in a richer way.

Notice that Result 1 (for the extensive-width case) and Previous Result 4 (for finite width) apply to different scalings for the number of samples n. In particular, one can see that nonzero overlap in the low-width case happens on a sample scale $n = \mathcal{O}(rd)$, much smaller than the scale $n = \mathcal{O}(d^2)$ in the rotationally invariant, extensive-width case.

Previous Result 4 is derived in the strictly intensive width regime $r = \mathcal{O}(1)$. The extension to all subextensive widths $r \ll d$ may be technically nontrivial (see Refs. [78–80] for related discussion in another model), but it turns out that Previous Result 4 can be recovered as the limit $\rho \to 0$ [recall $r = \rho \min(d, L)$] of Result 1 (with an appropriately rescaled sample ratio); see Fig. 1.

We also note here that optimal algorithms based on message passing for low-width priors have also been derived and discussed in Ref. [42].

Let us remark that our Result 2 completes the analysis of all scaling regimes for the BSR model. It turns out that only three nontrivial scaling regimes are present as a function of the width scaling, i.e., $r \ll \mathcal{O}(d)$, $r = \mathcal{O}(d)$, and $r \gg \mathcal{O}(d)$. Indeed, Fig. 2 shows that in the extensive-width regime $r = \mathcal{O}(d)$, when $r/d \to 0$, the performance reduces to that of the $r \ll \mathcal{O}(d)$ regime. On the other hand, we show in Appendix C 5 that in the extensive-width regime $r = \mathcal{O}(d)$, when $r/d \to +\infty$, the performance reduces to that of linear regression, which is the Bayes-optimal estimator for i.i.d. Gaussian priors (to which the factorized BSR prior converges in the large width limit).

III. CONSEQUENCES OF MAIN RESULTS

In this section, we study the Bayes-optimal test error, i.e., the minimum theoretically achievable test error, for the BSR model and Gaussian label noise, as a function of the parameters $\beta = \max(d, L)/\min(d, L)$ (the sequence length to embedding dimension ratio), $\rho = r/\min(d, L)$ (the width ratio—note that a lower ρ corresponds to a more structured prior), $\alpha = n/(dL)$ (the number of samples ratio), and Δ (the label noise). We identify the corresponding phase transitions in the performance, thus answering question (Q3) from the Introduction.

We also compare the BO estimator to some other baseline algorithms. Importantly, in Sec. III C, we compare with the linear regression on the vectorized sequence of tokens, thus quantifying the advantage of an estimator specializing in sequences of tokens over basic linear regression, which answers question (Q4) from the Introduction.

Again, notice that the factorized nature of the prior (10) enters explicitly in all our results, typically through the dependence on the width parameter ρ of the BSR model.

A. General phenomenology of the Bayes-optimal performance

Figures 1 and 2 show the BO test error as a function of the sample complexity α, for several values of β and width ρ, and in the noiseless setting $\Delta = 0$. We observe that lower values of ρ and larger values of β (i.e., larger structure in the prior) lead to lower test error. We also observe that the BO estimator can achieve zero test error at a finite value $\alpha_{\rm BO} < 1$, defining a threshold called the *strong recovery threshold* that we discuss in detail in Sec. III B.

We provide analogous data for $\Delta > 0$ in Appendix A. The overall phenomenology is similar, with the important difference that we observe no strong recovery at finite α, and the test error in the noisy case to be a continuous and differentiable function of the parameters.

In Fig. 1, right column, we highlight the convergence of the extensive-width test error to the low-width result for $\rho \to 0$, after appropriately rescaling the sample ratio α to $\bar{\alpha} = \{\beta/[\rho(1+\beta)]\}\alpha$. We observe quantitative differences from the low-width result for width ratios as low as 0.05, stressing the fact that the extensive-width analysis is relevant in finite-size applications, where ρ may be small but not strictly vanishing.

Figure 3 shows numerical experiments on GAMP-RIE with $\max(d, L) = 100$ for $\beta = 1, 2$, $\rho = 0.2, 1$, and $\Delta = 0$, 0.1, comparing it with the MMSE obtained by solving Eq. (24). We observe a very nice agreement already at these moderate system sizes.

Result 2, combined with Previous Result 4, provide the full picture for the Bayes-optimal test error in the BSR model, completely solving question (Q1). Additionally, Figs. 1 and 2 showcase the phenomenology for the noiseless observation channel. We have also answered positively to question (Q2), as the GAMP-RIE algorithm efficiently achieves the BO error in the high-dimensional limit.

B. Strong and weak recovery thresholds

In the noiseless output channel, we can provide an explicit characterization of the strong recovery threshold, i.e., the value of $\alpha_{\rm BO}$ such that, for all $\alpha > \alpha_{\rm BO}$, zero test error is achieved.

Result 5 (BO strong recovery threshold)—Consider the same setting as in Result 1, and specify it to the BSR model (10) and noiseless output channel. Then, in the high-dimensional limit, the strong recovery threshold satisfies

$$\alpha_{\rm BO} = \begin{cases} \frac{\rho}{\beta}(1+\beta-\rho) & 0 < \rho < 1, \\ 1 & \rho \geq 1. \end{cases} \tag{39}$$

The derivation of the threshold is performed in Appendix C 2 and involves expanding Result 1 in the limit $q \to 1^-$ and $\hat{q} \to +\infty$. We recall that, for nonzero label noise Δ, the strong recovery threshold is at infinity.

For comparison, the strong recovery threshold in the low-width limit equals $\lim_{\rho\to 0} \alpha_{\rm BO} = 0$, and in the more appropriate low-width sample scaling

$$\lim_{\rho\to 0} \bar{\alpha}_{\rm BO} = \lim_{\rho\to 0} \frac{\beta}{\rho(1+\beta)} \alpha_{\rm BO} = 1, \tag{40}$$

which can be derived either by taking the limit of Eq. (39) or independently by taking Previous Result 4 and solving the corresponding equations, as we do in Appendix C 1.

We remark that the strong recovery threshold can be guessed (but not properly justified) also through a counting argument where we compare the number of observations with the number of degrees of freedom (see also Ref. [77]). The spectrum of S^* accounts for a number $\mathcal{O}(d)$ of degrees of freedom. The singular vectors of $S^\star$ are a set of r orthonormal vectors in dimension d and one in dimension L. It is known that the set of $1 \leq r \leq d$ orthonormal vectors in dimension d, as a manifold (called the Stiefel manifold), has dimension [81]

$$\dim(r, d) = dr - \frac{r(r+1)}{2}. \tag{41}$$

Thus, by a dimensional argument, the number of samples needed to learn such bases for $r \leq \min(d, L)$ (i.e., $\rho \leq 1$) should equal

$$\begin{aligned} n &= dr - \frac{r(r+1)}{2} + Lr - \frac{r(r+1)}{2} + \mathcal{O}(d) \\ &= (d+L)r - r(r+1) + \mathcal{O}(d) \sim \frac{\rho}{\beta}(1+\beta-\rho)dL. \end{aligned} \tag{42}$$

This counting argument recovers the analytically derived threshold (39) and hints to the fact that this threshold will be universal to a larger subset of rotationally invariant priors with rank constraint, not limited to the BSR model for which our derivation of Eq. (39) holds.

We now turn to the weak recovery threshold. Recall that $\alpha_{\rm weak}$ is the largest α such that the performance of the BO estimator is the same as the performance of randomly sampling the prior. In the extensive-width case, the weak recovery threshold $\alpha_{\rm weak}$ is trivial. In other words, $\alpha_{\rm weak} = 0$ for $\rho > 0$, and for any $\alpha > 0$ nontrivial recovery is achieved. Instead, in the low-width case, a nontrivial weak recovery threshold arises (also for positive label noise Δ) at

$$\bar{\alpha}_{\rm weak} = (1+\Delta)\frac{\sqrt{\beta}}{1+\beta}. \tag{43}$$

We derive this threshold in Appendix C 1.

This section gives a clear answer to question (Q3). For noiseless output channels, the BO test error has a sharp threshold, corresponding to a second-order phase transition, between a region of positive error (at a small number of samples) and a region of zero error (at a large number of samples). Result 5 pinpoints the sample complexity α of the transition analytically. This, together with previous results for the low-width case, provides a full picture of the transition in the noiseless BSR model.

Notice that our result for the BO strong recovery threshold agrees with lower bounds discussed in Appendix 4.3 in Ref. [77] and originally shown in Ref. [82] in a related context of matrix denoising. This, along with Algorithm 1, settles the question posed by Ref. [77] whether an AMP algorithm can reach the naive dimensional lower bound for the strong recovery threshold. We answer positively and provide an AMP which not only achieves optimal recovery threshold, but that is also optimal at all sampling ratios α.

C. Comparison with linear regression on the vectorized data

As a crucial baseline motivating this work, we consider here the performance of linear regression performed on the vectorized input data. In the context of learning sequences of tokens, this amounts to flattening the data matrix $X_{ij} \in \mathbb{R}^{L\times d}$ into an Ld-dimensional vector, thus losing the semantic separation between token space and embedding space. The performance of such a procedure is quantified below.

Comparing this baseline to the Bayes-optimal performance of the BSR model quantifies the gain one can get when performing learning using a specialized sequence model as opposed to vectorizing the data and using fully connected neural networks (of which linear regression is the simplest example), effectively discarding some prior information.

Previous Result 6 (performance of ridge regression for Gaussian output channels and arbitrary priors)—Consider the ridge regression estimator

$$\hat{S}_{\text{ridge}}(\mathcal{D}) = \arg\min_{S}\left[\frac{1}{2}\sum_{\mu}(y_\mu - (Ld)^{-1/2}\text{Tr}(S^T X^\mu))^2 + \frac{\lambda}{2}\text{Tr}(S^T S)\right], \tag{44}$$

and a dataset $\mathcal{D}$ generated by Eq. (9) with Gaussian label output channel with variance Δ and arbitrary prior P_0 normalized such that $Q_* = 1$. Define the sample ratio as $\alpha = n/(dL)$. Then, the optimal value of the regularization is $\lambda_{\text{opt}} = \Delta$, and the mean square estimation error of the optimally regularized ridge regression estimator equals (in the large-dimensional limit)

$$\text{MSE}^{\text{ridge}}_{\alpha,\Delta} = \frac{1+\alpha+\Delta-\sqrt{(\alpha+\Delta+1)^2-4\alpha}}{2}, \tag{45}$$

which reduces to $\text{MSE}^{\text{ridge}}_{\alpha,\Delta=0} = \max(1-\alpha, 0)$ in the noiseless case.

The analysis of empirical risk minimizers for convex losses, and, in particular, for ridge regression, is standard; see Ref. [17], for example, for a very generic derivation. We point out here only that the prior, arbitrarily complicated and with extensive width, enters this performance only through its second moment $Q_* = 1$ due to the choice of ℓ_2 regularization. This can also be seen directly by the explicit solution of the ridge regression problem, which notably depends only on the second-order statistics of the data and labels.

Figure 1 shows the BO test error as a function of α for several values of β and ρ and compares with the performance of linear regression (which is independent on ρ and β in the scaling we chose). We see that, in all cases, larger values of β and smaller ρ lead to more significant gains in using the prior-aware BO estimator compared with the simple ridge estimator, in both noiseless ($\Delta = 0$) and noisy ($\Delta > 0$; see Appendix A) cases.

Notice that in the noiseless case we see that $\text{MSE}^{\text{ridge}}_{\alpha,\Delta=0} = 0$ at $\alpha = 1$, recovering the trivial fact that an invertible linear system of p equations in p unknowns has a unique solution. More generally, the strong recovery threshold for ridge regression with noiseless data is given by $\alpha_{\text{ridge}} = 1$. Whenever $\rho < 1$, there exists a full range of sample ratios $\alpha_{\text{BO}} < \alpha < \alpha_{\text{ridge}}$ where the BO estimator achieves zero test error, while the ridge estimator does not.

As $\rho \to \infty$, S converges to a matrix with i.i.d. standard Gaussian entries: The problem is effectively vectorized, as there remain no correlations between the token and embedding dimensions L and d. In this limit, we expect that optimally regularized ridge regression will achieve the BO performance, as the loss and regularization choice matches the distribution on S and the generative process for the labels. This is indeed the case, as we show in Appendix C 5 by explicitly computing the large-ρ limit of Eq. (24) for the BSR model.

This answers question (Q4) from the Introduction: Vectorizing data and learning it with linear regression is suboptimal for any finite ρ. For $0 < \rho < 1$, the suboptimality is particularly striking, as there exists a full region of sample ratio $\alpha_{\text{BO}} < \alpha < 1$ in which linear regression has nonzero test error, while the BO estimator achieves zero error.

D. Comparison with a minimal nuclear norm estimator

As discussed in the Introduction, previous works explored algorithms based on nuclear norm minimization to solve the matrix sensing and denoising problem [45,46,82]. It is instructive to compare the performance of this algorithm to the optimal estimator. The *minimal nuclear norm estimator* (MNNE) is defined as

$$S_{\mathrm{MNNE}} \coloneqq \arg\ \min \|S\|_{\mathrm{nuc}} = \arg\ \min \mathrm{Tr}\left(\sqrt{SS^T}\right)$$

$$\text{such that } y^{\mu} = \frac{1}{\sqrt{Ld}} \sum_{a,i=1}^{L,d} X^{\mu}_{ai} S_{ia}, \tag{46}$$

where we recall that the nuclear norm is just the sum of the singular values of a given matrix. This algorithm is the convex relaxation of the minimum rank estimator, where one seeks a matrix with minimal rank fitting the dataset. It has been observed [46,82] that the MNNE can achieve zero estimation error, provided that the ground-truth matrix S^* has constrained rank (i.e., $0 < \rho < 1$) and that the number of samples n is large enough. The authors are also able to characterize the corresponding strong recovery threshold α_{MNNE} and provide an explicit asymptotic value, which we report here for completeness.

Previous Result 7 (strong recovery threshold for the MNNE [46,82])—Consider the Marčenko-Pastur distribution defined by

$$p_{\gamma}(t) \coloneqq \frac{1}{2\pi\gamma t}\sqrt{(\gamma_+ - t)(t-\gamma_-)} \cdot \mathbf{1}_{[\gamma_-,\gamma_+]}(t), \tag{47}$$

where $\gamma_{\pm} = (1 \pm \sqrt{\gamma})^2$, and define its complementary incomplete moments as

$$P_{\gamma}(x;k) \coloneqq \int_x^{\gamma_+} t^k p_{\gamma}(t) dt. \tag{48}$$

Let

$$\mathbf{M}(\Lambda;\rho,\beta) \coloneqq \rho(1+\beta-\rho) + (\beta-\rho)\left[\rho\Lambda^2 + (1-\rho)(P_{\gamma}(\Lambda^2;1) - 2\Lambda P_{\gamma}\left(\Lambda^2;\frac{1}{2}\right) + \Lambda^2 P_{\gamma}(\Lambda^2;0))\right], \tag{49}$$

for $0 < \rho < 1$ and $\beta \geq 1$, with

$$\gamma = \frac{1-\rho}{\beta-\rho}. \tag{50}$$

Then, the strong recovery threshold of the MNNE satisfies

$$\alpha_{\mathrm{MNNE}} = \min_{0 \leq \Lambda \leq \gamma_+(\rho,\beta)} \mathbf{M}(\Lambda;\rho,\tilde{\beta}). \tag{51}$$

The minimum can be computed numerically by solving the zero-derivative condition $d\mathbf{M}/d\Lambda = 0$ on the interval $(0,\gamma_+)$, i.e.,

$$P_{\gamma}\left(\Lambda^2;\frac{1}{2}\right) - \Lambda \cdot P_{\gamma}(\Lambda^2;0) = \frac{\Lambda\rho}{1-\rho}. \tag{52}$$

This can be done by a bisection algorithm.

In Fig. 4, we plot the theoretical prediction of α_{MNNE} from Refs. [46,82] and compare it with the BO threshold α_{BO} that we derive in Sec. III B for different values of β. We observe that for all values of $0 < \rho < 1$ and $\beta \geq 1$ the two thresholds are different, and, in particular, $\alpha_{\mathrm{BO}} < \alpha_{\mathrm{MNNE}}$. This highlights an intrinsic suboptimality of the MNNE, which for an extended range of values $\alpha_{\mathrm{BO}} < \alpha < \alpha_{\mathrm{MNNE}}$ fails to achieve the BO performance.

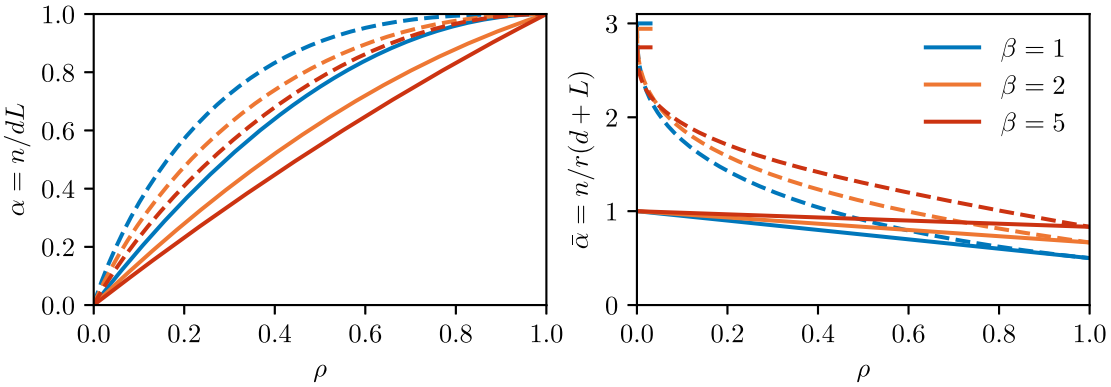

FIG. 4. Comparison between the BO strong recovery threshold (Result 5, solid lines) and the MNNE strong recovery threshold (Previous Result 7, dashed lines) for $\beta = \max(d,L)/\min(d,L) = 1, 2, 10$ as a function of the width ratio $\rho = r/\min(d,L)$. On the left, we plot the strong recovery thresholds in the scaling $\alpha = n/(dL)$, natural in the extensive-width case $\rho > 0$. On the right, we plot the same data in the low-width sample scaling $\bar{\alpha} = n/[r(d+L)]$, highlighting the strong suboptimality of MNNE at low ranks and widths. The colored markers on the vertical axis highlight the finite $\rho \to 0$ limit of the strong recovery threshold of MNNE, as given in Eq. (53).

In the low-rank regime $\rho \to 0$, Ref. [46] provides the following asymptotic value for the strong recovery threshold of the MNNE:

$$\bar{\alpha}_{\mathrm{MNNE}} = \lim_{\rho\to 0} \frac{\beta}{\rho(1+\beta)} \alpha_{\mathrm{MNNE}} = 2\left(1 + \frac{\sqrt{\beta}}{1+\beta}\right), \quad (53)$$

while for the BO strong recovery threshold we have $\lim_{\rho\to 0} \bar{\alpha}_{\mathrm{BO}} = 1$. We, thus, see that also in the low-rank limit (with the appropriately rescaled sample ratio) the MNNE recovery threshold remains suboptimal. This is akin to what happens in the compressed sensing problem when we compare the Bayes-optimal performance to the performance of the convex relaxation via L_1 regularization [83]. In Appendix A, we provide numerical experiments comparing the performance of the MNNE estimator with the BO test error, with GAMP-RIE, and with the prediction for the strong recovery threshold (Previous Result 7).

IV. BEHAVIOR OF GRADIENT DESCENT

Arguably, the most interesting algorithm to study in the context of the BSR model is *gradient descent* (GD), since its variants are the driving horse of state-of-the-art applications of machine learning. We consider here the BSR model with Gaussian additive noise channel and assume the width parameter r is known. For this case, the most natural choice of loss function is

$$\mathcal{L}(U,V) = \frac{1}{4}\sum_{\mu=1}^{n}\left(y^{\mu} - \frac{1}{\sqrt{Ldr}}\sum_{a,i=1}^{L,d} X^{\mu}_{ai}\sum_{j=1}^{r} U_{ij}V_{ja}\right)^2, \quad (54)$$

where $U \in \mathbb{R}^{d\times r}$ and $V \in \mathbb{R}^{r\times L}$. The loss is then minimized over the factors U and V using the following gradient descent iterations:

$$U^{t+1} = U^t - \eta\nabla_U\mathcal{L}(U^t,V^t) \text{ and } V^{t+1} = V^t - \eta\nabla_V\mathcal{L}(U^t,V^t) \quad (55)$$

with $\eta > 0$ being the learning rate. Unlike in linear regression, the loss (54) is nonconvex, and, thus, keeping particular attention to the initialization, learning rate, and the stopping criterion is required to properly understand the properties of the GD estimator. In general nonconvex settings, the generalization performance of the GD algorithms is mostly a widely open question that is actively studied.

In this section, we initiate the understanding the performance of the GD in the BSR model and investigate how the choice of initialization and learning rate influence the performance of the algorithm, thus providing some answers to (Q5) from the Introduction. We argue that the BSR model is a simple yet very interesting model to further the understanding of the broad set of questions behind the functioning of the GD algorithm. Without exhaustively mapping the possible choices of initialization, learning rate, and stopping time, we identify two remarkable properties and discuss them further below:

(i) *GD can reach the Bayes-optimal performance*—For the noiseless BSR model, we find that, for well-chosen learning rate and factors initialized in the prior distribution, GD behaves *as if* it was sampling uniformly the space of global minimizers, and an averaged version of GD [defined in Eq. (56)] reaches the Bayes-optimal generalization error. We stress that there is no *a priori* reason for GD being able to sample the minimizers; this observations is, thus, very surprising. When noise is present, the behavior is more complex, and GD does not seem to sample the minimizers anymore. Numerical evidence is given in Fig. 5. Note that similar properties were observed for a related model in Ref. [63].

(ii) *Implicit regularization of GD, but not with respect to the minimum nuclear norm*—We find that many choices of the learning rate and initialization lead to a generalization performance that is better than the one of a randomly chosen global minimizer. An interesting existing line of work proposed that in some settings the implicit regularization may be related to the nuclear norm [50]. We, thus, ask whether this would be the case in the BSR model. Our numerical investigation suggests that in the BSR model, even with small learning rate and small amplitude initialization, GD does not minimize the nuclear norm. This is evidenced in Fig. 6.

More work is needed to fully characterize the behavior of the GD algorithm in the BSR model, and this characterization is a prerequisite for further understanding of learning dynamics in more complex sequence models.

A. GD and the Bayes-optimal performance

In Fig. 5, we initialize both U and V as i.i.d. Gaussian matrices, with each entry having mean zero and unit variance, i.e., from the same distribution as in the BSR model (1). Figure 5, left, shows that GD run on the BSR model without noise reaches very small training loss if the learning rate η is properly tuned and a test error equal to the one of a uniformly sampled global minimum of the loss [this is usually called a Gibbs sampler for the posterior distribution and has test error equal to twice the BO test error as we show in Eq. (23)]. This prompts us to speculate that runs of GD with independent initialization may be close to sampling the space of global minimizers of the loss (as a Gibbs sampler would do). Notice that GD can converge only to the boundary of the set of global minimizers and that in high dimension the uniform measure over such set is plausibly concentrated on the boundary, provided that the set of global minimizers is not pathological. This observation, plus the numerical observation

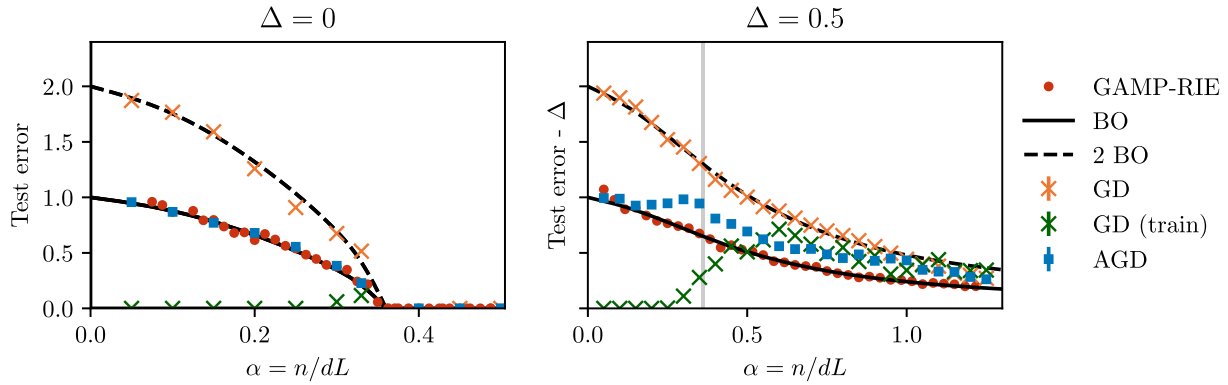

FIG. 5. Comparison between the test error achieved by GD and AGD initialised in the prior and of the BO test error for $\beta = 1$, $\rho = 0.2$, and $\Delta = 0, 0.5$ (left and right, respectively). In the noisy case, we depict the test error minus the variance of the noise Δ. Solid lines are the BO test error; dashed lines are twice the BO test error corresponding to the error of the Gibbs sampler. Orange crosses are numerical experiments for the test errors at the end of the run of GD for $d = L = 100$, maximum number of steps $\tau = 50\,000$, and runs are averaged over 16 instances of the data. Blue squares are numerical experiments for the test errors at the end of the run of AGD (averaged over 32 initial conditions), and they are averaged over two instances of the data (eight in the right up to $\alpha = 0.6$). The error bars denoting standard error on the mean are negligible. In both cases, a fine-tuned value of the learning rate $\eta(\alpha)$ must be used, dependent on the sample ratio α. We provide the values used to generate this plot in Appendix A. The green crosses mark the value of the training loss at the end of the training for GD. The gray vertical line in the right-hand side marks where the number of samples equals the number of degrees of freedom. Finally, red dots are numerical experiments for GAMP-RIE, with a single random instance of $d = L = 100$. We observe that, in the noiseless case $\Delta = 0$ (left), GD achieves a test error compatible with the error of the Gibbs sampler and that AGD achieves a test error compatible with the BO test error. Instead, for $\Delta = 0.5$ (right), we observe that AGD does not reach the BO error, and, moreover, it trivializes (namely, all differently initialized runs of GD converge to the same estimator) for α large enough, roughly around $\alpha \approx 1$ here. We show qualitatively similar comparisons at $\beta = 2$ in Appendix A. The Bayes-optimal curves are plotted using Result 2 and Eq. (24). The performance of GAMP-RIE, GD, and AGD are given by Eq. (13) applied to the output of the respective algorithm.

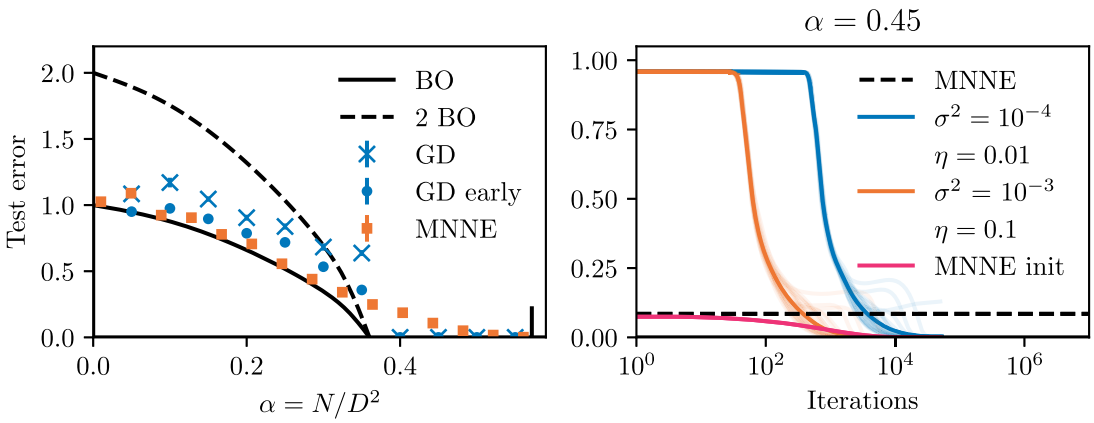

FIG. 6. Left: comparison between the test error achieved by GD initialized with small norm (blue crosses, $d = L = 50$, maximum iterations $T = 10^4$, learning rate $\eta = 0.2$, initialization norm $\sigma^2 = 10^{-4}$) and its early stopped version (blue dots) averaged over 16 instances, the test error of the MNNE (orange squares, $d = L = 50$) and of the BO test error (solid black line; dashed lines are twice the BO error corresponding to the error of the Gibbs sampler) for $\beta = 1$, $\rho = 0.2$, and $\Delta = 0$. We observe that the MNNE performs slightly better than GD with small initialization up to roughly the BO strong recovery thresholds, while for larger α GD becomes better and notably has a better strong recovery threshold than MNNE. Right: comparison between the MNNE and several runs of GD with small initialization (maximum iterations $\tau = 20\,000$), all on the same instance of the data and ground truth with $d = L = 50$, $\rho = 0.2$, and $\Delta = 0$. We perform this comparison at $\alpha = 0.45$, where the left suggests that GD will achieve zero error, while MNNE does not. We run GD from two different initialization magnitudes $\sigma^2 = 10^{-3}, 10^{-4}$ (orange and blue curves, respectively; thick curves mark the average), and we also run GD from the MNNE initialisation (see the main text for a precise definition). We observe that in all cases GD outperforms MNNE at convergence. The Bayes-optimal curves are plotted using Result 2 and Eq. (24). The performance of MNNE and GD are given by Eq. (13) applied to the output of the respective algorithm.

that the test error matches the Gibbs sampler, justifies our speculation.

The Bayes-optimal estimator in the noiseless case is given by averaging over the global minimizers of the loss. Given our hypothesis above that GD samples uniformly the set of global minimizers, we are prompted to average J GD runs to construct a novel estimator. If our hypothesis is correct, then this averaged estimator should achieve close to BO performance. Thus, we sample J different pairs of initial matrices (U_j^0, V_j^0) and iterate GD for τ iterations obtaining (U_j^τ, V_j^τ) to compute the estimator of S as

$$\hat{S}_{\mathrm{GD,avg}} = \frac{1}{J}\sum_{j=1}^{J} \frac{U_j^\tau V_j^\tau}{\sqrt{r}}. \tag{56}$$

We call this the *averaged GD* (AGD) algorithm. Figure 5 shows that this AGD estimator indeed reaches the Bayes-optimal test error. This observation holds for different values of β.

Figure 5, right, shows experiments in the same setting but with label noise $\Delta = 0.5$. We observe here a different phenomenology, in particular, with AGD not reaching the BO test error. We notice also that the training error equals zero before the naive interpolation threshold where the number of samples equals the number of degree of freedom ($\alpha = \alpha_{\mathrm{BO}}$ in the BSR model). At the interpolation threshold, AGD has a trace of what may be an interpolation peak, which does not appear in the simple GD. More work is needed to fully understand whether there is a way to tune the parameters of the AGD algorithm to reach the BO error also in the noisy case.

B. Implicit regularization of GD, comparison to the minimum nuclear norm

The data reported in Fig. 5 depend strongly on the choice of the initialization and the learning rate. We consider initialization where the component of matrices U^0 and V^0 are still i.i.d. Gaussian random variables of zero mean but this time with variance σ^2. Figure 5 is for $\sigma^2 = 1$, but initializing with small σ^2 is considered more interesting, one reason being that the BO estimator at $\alpha = 0$ is simply 0. In Fig. 6, left, we show that GD initialized with small norm $\sigma^2 = 10^{-3}$ or 10^{-4} reaches a test error which is slightly larger than the BO one but still significantly smaller than initializing at $\sigma^2 = 1$ (where nonaveraged GD reaches the Gibbs sampling error). When GD leads to a better test error than a random global minimizer, the machine learning literature often refers to so-called *implicit regularization* of gradient descent [50,51,84,85]. What we observe in Fig. 6 is a clear sign of implicit regularization. In this case, the role of the learning rate is less influential, as long as it is kept small enough (see Fig. 11 in Appendix A). We also find that for small initialization early stopping can be advantageous, as also reported in the figure.

It is natural to compare GD with small initialization to the MNNE, as it was suggested that in some settings (in particular, overparametrized ones) gradient flow with vanishing norm at initialization has an implicit bias toward minimizing the nuclear norm [50]. Later work then questioned this conjecture and disproved it in a constructed special case [51]. In the BSR model, we find numerically that GD starting with small initialization does not go to the minimizer corresponding to the smallest nuclear norm. Evidence for this is reported in Fig. 6. In particular, the most striking difference between GD and the MNNE is at values of sample complexity α slightly above the BO strong recovery threshold but below the MNNE strong recovery threshold, where we see that GD already reaches strong recovery and the MNNE does not.

In Fig. 6, right, we show a numerical experiment on a single instance of the data and ground truth, comparing the test error during training for GD with small initialization and small learning rate ($\sigma^2 = 10^{-3}, 10^{-4}$) and GD initialized in the MNNE solution with the MNNE error. To initialize GD in the MNNE solution, we consider the singular value decomposition of the MNNE $S_{\mathrm{MNNE}} = U_{\mathrm{MNNE}} D_{\mathrm{MNNE}} V_{\mathrm{MNNE}}$, and we take $U^0 = U_{\mathrm{MNNE}}\sqrt{\tilde{D}_{\mathrm{MNNE}}}$ and similar for V, where $\tilde{D}_{\mathrm{MNNE}}$ is the truncated version of D_{MNNE} to the leading r singular values. We plot this comparison at $\alpha = 0.45$, where we observe GD to achieve zero error and MNNE to not achieve zero error from the averaged comparison in Fig. 6, left. We clearly see that at a single instance GD outperforms the MNNE estimator and that the MNNE estimator is not even a stable minimum of the GD landscape.

V. DISCUSSION AND FUTURE DIRECTIONS

We introduced the bilinear sequence regression as a basic model for learning from long sequences of high-dimensional tokens. In this paper, we addressed and answered questions (Q1)–(Q4) posed in the Introduction. Our analysis involved techniques of statistical physics that are, in general, not mathematically rigorous, and the rigorous establishment of our results is a clear avenue for future work. Concerning question (Q5) about gradient descent, we investigated the behavior of this algorithm numerically, and it is clear that, already for this rather simple model, there is a very rich behavior of gradient descent that can be observed. Our experiments reveal some of the intriguing properties of GD in the BSR model. A clear avenue for future work is to analyze the behavior of gradient descent in the large size limit, e.g., via dynamical mean field theory [86], and aim to explain our numerical observation theoretically.

The need for a detailed understanding of the properties of the optimizer is also emphasized by a set of experiments that we performed using the toy transformer architecture presented in Sec. I C on the data from the BSR model. We observed that the performance is comparable to what we report for gradient descent in Sec. IV. The transformer

model seems able to figure out that the attention part of the architecture is not useful and only the skip connection is. However, again the performance depended strongly on the optimizer used, and, thus, it seems to us that attempts to quantify the cases where the attention layer is advantageous need to be preceded by a more complete understanding of the gradient descent and its variants.

Since sequence models are behind the recent progress of artificial intelligence, having a basic model for studying learning from sequences of tokens opens the avenue to address many of the questions underlying these systems. For this, future work will need to generalize the BSR model in several directions.

(i) *Structured input data and more general tasks*—So far, we considered Gaussian i.i.d. input data X and a task defined by Eq. (1). One should generalize the model to add nontrivial structure in the input X to mimic correlations present, e.g., in natural language. A step toward this that is technically achievable may be done in a similar manner as in the hidden manifold model in Ref. [15] or in the general Gaussian covariance model treated in Ref. [17]. For sequences of token, the correlation can be added both between different tokens and among the different embedding dimensions, two cases which are of interest.

(ii) *Learning with other architectures (in particular, those involving attention layers)*—In this paper, we start with the analysis of the Bayes-optimal performance for data generated by the BSR model. We then consider gradient descent for the model (8) that matches the BSR model, $d' = r$. The next step would be to study in detail the mismatched case where the learning model uses a different width than the data generative model $d' \neq r$. The landscape of the corresponding loss (54) is of interest as well as the behavior of the GD algorithm. Future work should also study models that include attention layers and clarify how their use is related to the structure in the data or task.

(iii) *Training algorithms*—In this work, we studied the Bayes-optimal estimation and a related message-passing algorithm. Understanding the behavior and properties of gradient descent theoretically is more challenging and is left for future work. Gradient descent should also be analyzed in more general, e.g., overparametrized $d' > r$ settings. Studying the behavior and properties of other algorithms such as stochastic gradient descent, variants with momentum, and adaptive learning rates are also of interest. Investigation of an algorithm that would eventually lead to better learning than the currently existing variant of gradient descent would be immensely important.

Of course, the investigation of the above three directions cannot be done separately, because the right architecture will depend on the structure of the data and the task and on the fact that the used training algorithm needs to run efficiently. The three ingredients—the data or task structure; the architecture or estimator; and the algorithm—interplay in ways that need to be understood better. The present paper initiates this study for sequence models and opens a natural program for future research along these directions, in terms of both more realistic models and avenues for developing the methodological toolbox to study learning in high dimensions.

ACKNOWLEDGMENTS

We would like to thank Florent Krzakala for numerous discussions related to this work, Jason Lee for pointing out the connection to the works on implicit regularization of gradient descent and Christian Keup for early discussions. We acknowledge funding from the Swiss National Science Foundation grants SNFS SMArtNet (Grant No. 212049) and TMPFP2-210012.

APPENDIX A: ADDITIONAL PLOTS

In this section, we provide additional details on the plots we presented in the main text, as well as additional plots for different values of the parameters.

(i) In Fig. 7, we verify for $\beta = 1, 2$ and $\rho = 0.5, 1$ that the free entropy as a function of q is maximized at a unique nontrivial $q < 1$ solution for all values of $\alpha < \alpha_{\mathrm{BO}}$. We cannot check all values of alpha. We check a selection and claim that by regularity of the free entropy no hard phase (i.e., no maximality of the trivial solution of the state evolution equations $q = 1$) is expected up to the largest value of alpha we checked, which is numerically very close to the predicted strong recovery threshold for all values of β and ρ shown.

(ii) Figures 8 and 9 are the noisy versions, with $\Delta = 0.1$, of the main text Figs. 2 and 1, respectively.

(iii) In Fig. 3, left, for $\alpha < \alpha_{\mathrm{BO}}$ we use $\zeta = 20$ and $\gamma = 0.5$, while for $\alpha > \alpha_{\mathrm{BO}}$ we use $\zeta = 20$ and $\gamma = 0.8$. In Fig. 3, right, we use $\zeta = 20$ and $\gamma = 0.5$. In all cases, the tolerance used to determine convergence is $\epsilon = 10^{-6}$ on the MSE between successive iterates. Around the transition α_{BO} in the noiseless case, more iterates are needed, and for this reason we use a lower tolerance $\epsilon = 10^{-8}$.

(iv) In Fig. 12, we plot the MSE of the MNNE as a function of α obtained through finite-size numerical simulations [$\max(d, L) = 20,\ 50$ using CVXPY [87,88]] for $\beta = 1, 2$ and $\rho = 0.2, 0.5$. We observe that, somewhat surprisingly, for $0 < \alpha \lessapprox \alpha_{\mathrm{BO}}$ the performance of the MNNE is quantitatively very close to the BO performance, while for $\alpha_{\mathrm{BO}} < \alpha < \alpha_{\mathrm{MNNE}}$ the MNNE performs sizably worse than the BO estimator. A theoretical prediction of the MSE of

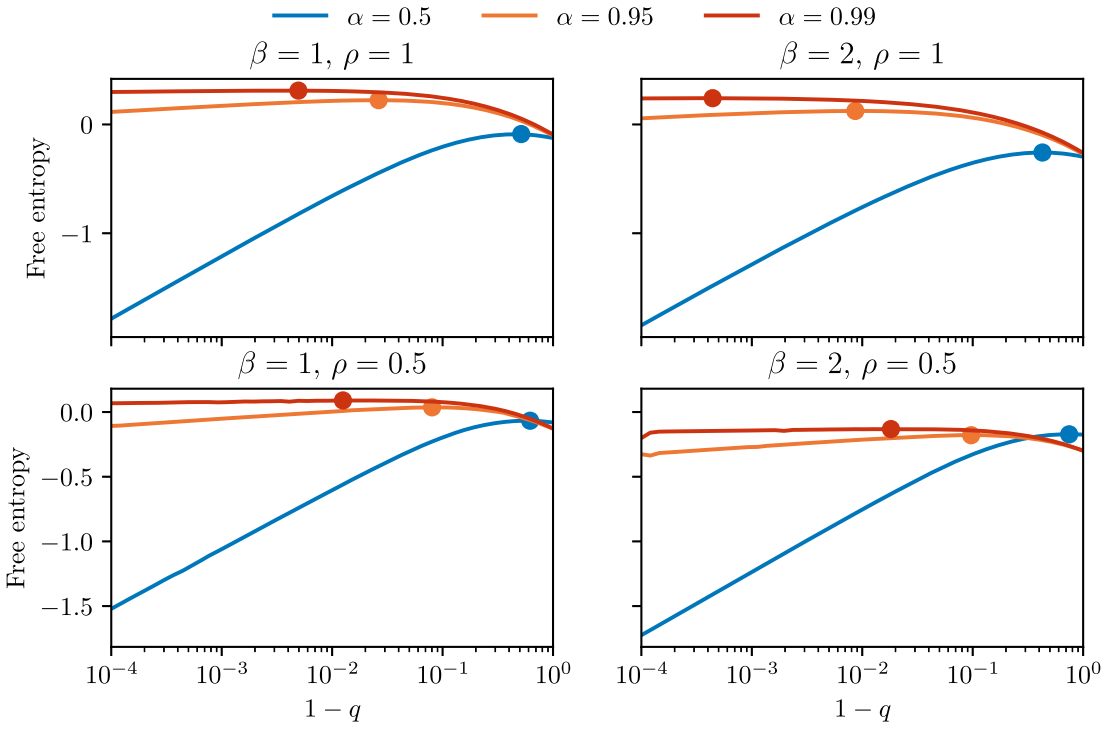

FIG. 7. Free entropy (B49) for $\beta = 1, 2$ and $\rho = 0.5, 1$ as a function of q. We highlight that the free entropy has only a single maximum in $0 < q < 1$ in all these cases.

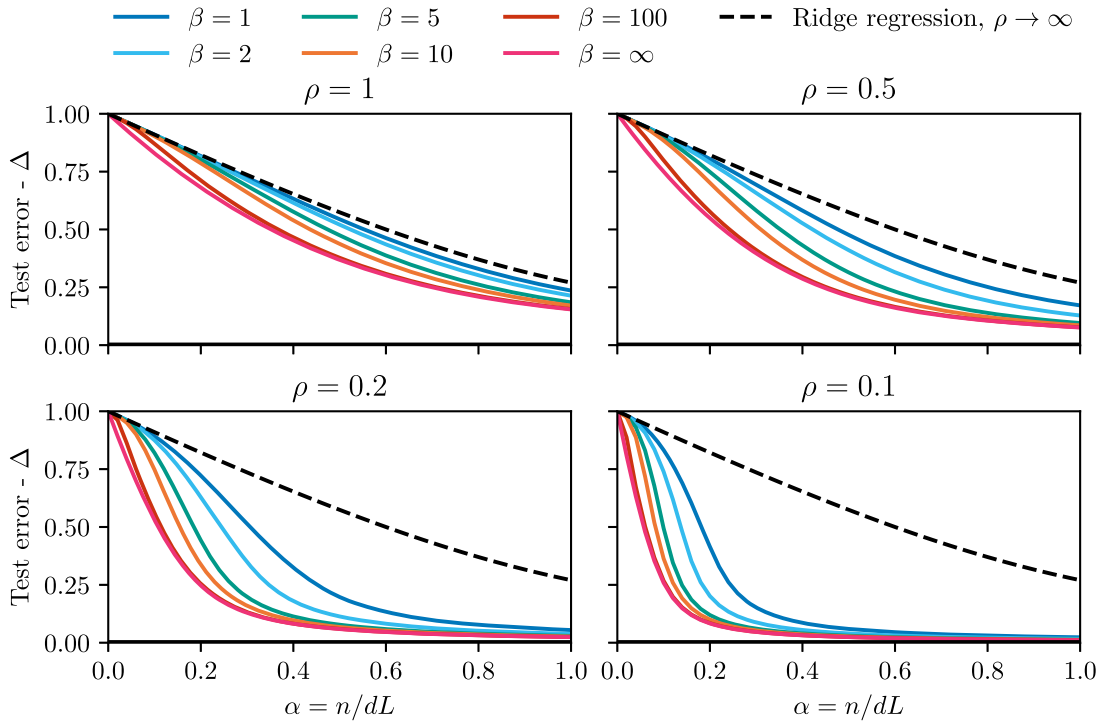

FIG. 8. The same as Fig. 2, only with $\Delta = 0.1$.

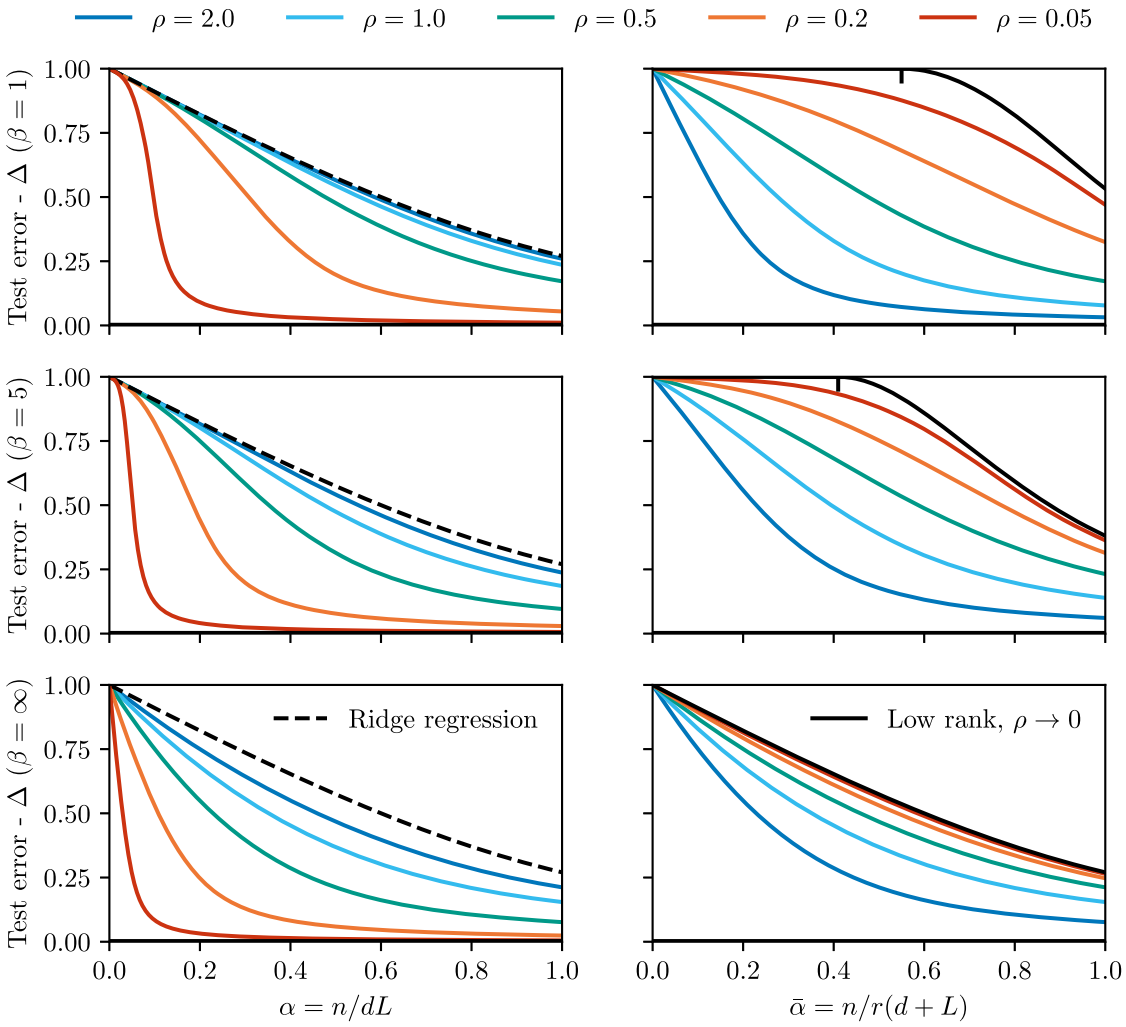

FIG. 9. The same as Fig. 1, only with $\Delta = 0.1$.

the MNNE in the high-dimensional limit is, as far as we know, not readily available.

(v) Figure 10 is the rectangular version $\beta = 2$ of Fig. 5, and Table III indicates the learning rates used.

(vi) Tables I and II list all learning rates used to produce Fig. 5.

(vii) In Fig. 11 we probe the effect of the learning rate on the test error. We can see that if we initialize with small norm, we need the learning rate to be small enough for

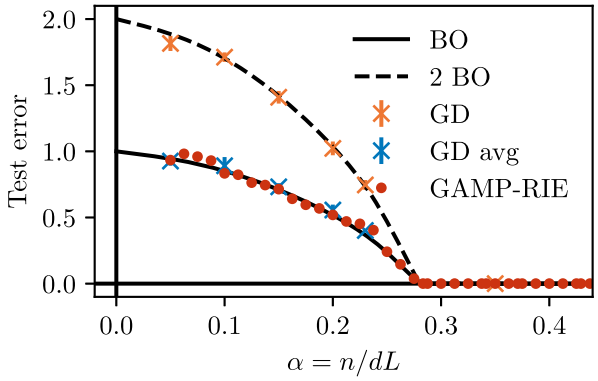

FIG. 10. The same as Fig. 5, only with $\beta = 2$.

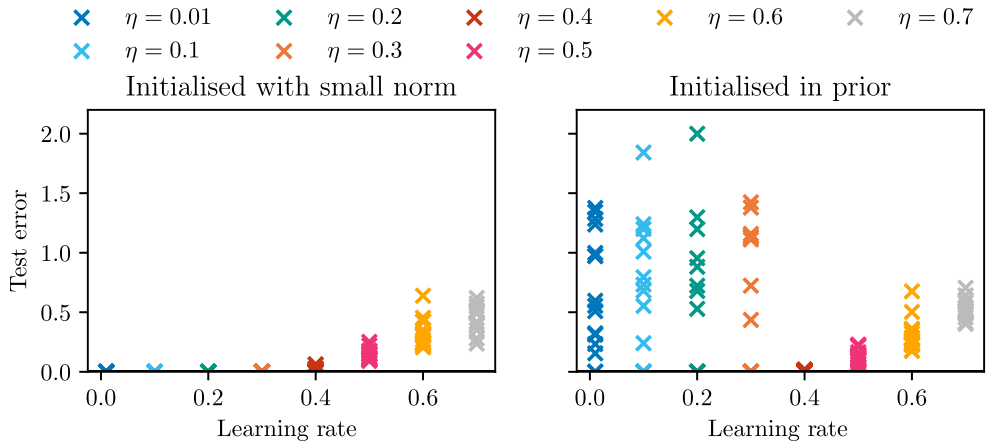

FIG. 11. Test error as a function of the learning rate for GD initialized with small norm (left) or in the prior (right). Here, $D = 50$, $\beta = 1$, $\rho = 0.2$, and $\alpha = 0.44$. In this regime, we expect GAMP-RIE to achieve zero error. We can see how choosing the right learning rate influences the performance of the model rather drastically.

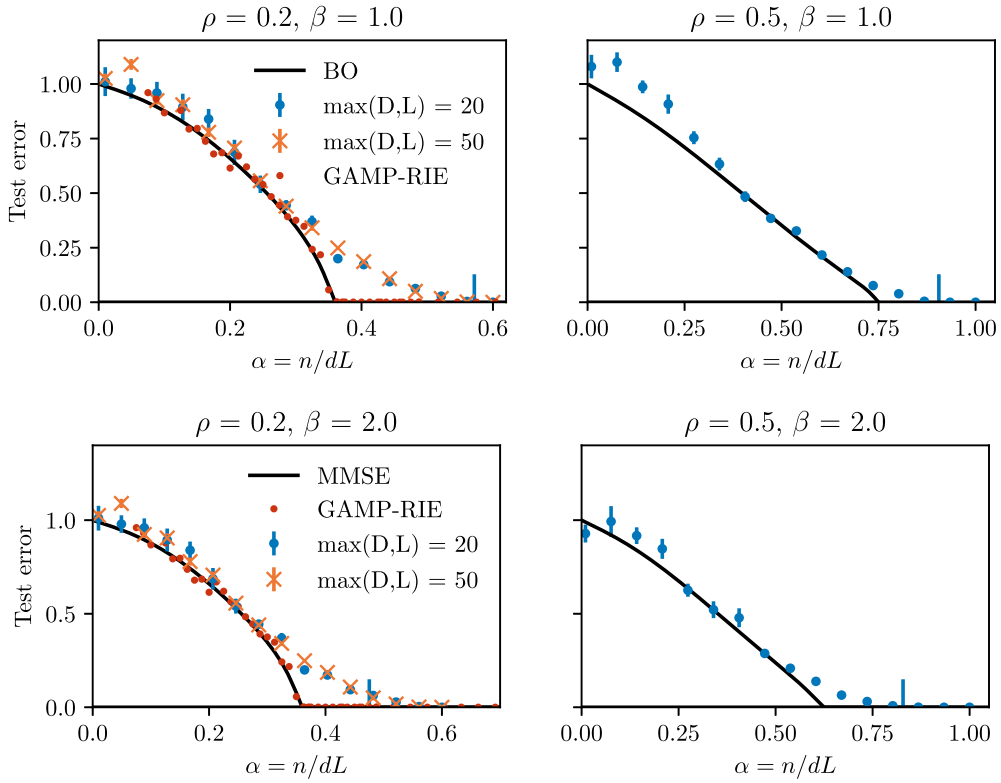

FIG. 12. Comparison between of the test error of the MNNE estimator and the BO test error for $\rho = r/\min(d, L) = 0.2$, 0.5 and $\beta = \max(d, L)/\min(d, L) = 1$. Dots are numerical experiments for MNNE at $\max(d, L) = 20$, 50 averaged over 16 independent instances of the data and ground truth, and error bars denote the standard error on the mean. Solid lines denote the BO test error, dots and crosses the predicted asymptotic strong recovery threshold for the MNNE estimator (Previous Result 7), and red dots the performance of AMP (same parameters as in Fig. 3). We observe two qualitatively different behaviors. For $\alpha \lessapprox \alpha_{\mathrm{BO}}$, the MNNE estimator achieves a test error very close to the BO one. For $\alpha_{\mathrm{BO}} \lessapprox \alpha < \alpha_{\mathrm{MNNE}}$ instead, the BO error is precisely zero, while the MNNE error is nonzero. Finally, our numerical experiments are compatible with the theoretical prediction for the strong recovery threshold of MNNE (Previous Result 7). We show qualitatively similar comparisons at $\beta = 2$ in the second row.

TABLE I. Learning rate as a function of sample complexity for Fig. 5, right.

Sample ratio $\alpha = n/(dL)$	0.05	0.1	0.15	0.2	0.25	0.3–0.45	0.5–0.6
Learning rate η	0.7	0.75	0.65	0.58	0.53	0.5	0.45
Sample ratio $\alpha = n/(dL)$	0.65	0.7	0.75–0.8	0.85–0.9	0.95–1.1	1.15–1.25	
Learning rate η	0.4	0.35	0.3	0.25	0.2	0.15	

TABLE II. Learning rate as a function of sample complexity for Fig. 5, left.

Sample ratio $\alpha = n/(dL)$	0.05	0.1–0.33	0.45–0.5
Learning rate η	0.5	0.7	0.3

TABLE III. Learning rate as a function of sample complexity for Fig. 10.

Sample ratio $\alpha = n/(dL)$	0.05	0.1	0.15	0.2	0.23	0.35
Learning rate η	0.85	0.75	0.7	0.65	0.65	0.25

GD to perform well, while if we initialize in the prior, we need to carefully tune our parameters.

APPENDIX B: REPLICA COMPUTATION FOR THE BAYES-OPTIMAL CASE

In this appendix, we derive Result 1, and we rederive Previous Result 4 under more general priors, through the replica method. We study both the intensive and extensive-width BSR model using a common framework. We then specialize to the extensive case in Appendix B 10 and to the intensive case in Appendix B 11.

1. A word on scalings

In the following derivation, we consider the sample ratio $\bar{\alpha} = n[r(d+L)]$, where the number of samples scales proportionally to the number of unknown scalars in the ground-truth signal. This allows to treat the low-width case $r \ll d$ and the extensive-width case $r = O(d)$ within a unique framework. This choice dictates also the overall scaling for the free entropy (defined below) to be $\mathcal{O}[r(d+L)]$. In the extensive-width case, one recovers the results in term of the ratio $\alpha = n/(dL)$ through the rescaling

$$\bar{\alpha} = \frac{n}{r(d+L)} = \frac{\beta}{\rho(1+\beta)}\alpha. \tag{B1}$$

The scaling $\bar{\alpha}$ makes sense only for factorized priors, where r exists. Our computation holds also for nonfactorized but rotationally invariant priors. To obtain the associated results set $\rho = 1$.

2. Preliminaries

We start by recalling the definition of the posterior distribution and the associated partition function. The posterior distribution $P(S|\mathcal{D})$ is the probability that, given an observed dataset $\mathcal{D}_n = \{(X^\mu, y^\mu)\}_{\mu=1}^n$, the dataset has been generated from a given set of weights S. By the Bayes rule, we have

$$\begin{aligned} P(S|\mathcal{D}_n) &= \frac{P(\mathcal{D}_n|S)P_0(S)}{P(\mathcal{D}_n)} \\ &= \frac{P_0(S)}{P(\mathcal{D}_n)}\prod_{\mu=1}^{n} P_{\text{out}}\left(y^\mu\middle|\frac{1}{\sqrt{dL}}\text{Tr}(S^T X^\mu)\right), \end{aligned} \tag{B2}$$

where $P(\mathcal{D}_n)$ is interpreted as the normalization factor for the distribution (as it is independent on $\mathcal{D}_n$) and for this reason is thought of as a partition function. As usual in the statistical mechanics of disordered systems and in its applications in inference, we expect the free entropy $\Phi_{\mathcal{D}} = \{1/[r(d+L)]\}\log P(\mathcal{D}_n)$ to concentrate both with respect to the variable S and the quenched disorder $\mathcal{D}$ in the high-dimensional limit. For this reason, we study the averaged free entropy $\Phi = \mathbb{E}_{\mathcal{D}}\Phi_{\mathcal{D}}$, and we do so using the replica method.

3. Replica trick

The first step is to study the integer moments of the partition function $\mathbb{E}_{\mathcal{D}}P(\mathcal{D}_n)^u$, from which the averaged free entropy can be recovered (using a carefully chosen analytic continuation in u) as

$$\begin{aligned} \Phi &= \frac{1}{r(d+L)}\mathbb{E}_{\mathcal{D}}\log P(\mathcal{D}_n) \\ &= \frac{1}{r(d+L)}\lim_{u\to 0}\frac{\mathbb{E}_{\mathcal{D}}P(\mathcal{D}_n)^u - 1}{u}. \end{aligned} \tag{B3}$$

The averaged replicated partition function is (we call S_0 the ground truth S_* to highlight the fact that it behaves identically to other replicas)

$$\begin{aligned}
\mathbb{E}P(\mathcal{D}_n)^u &= \mathbb{E}_{X,y,S_0}\int\prod_{a=1}^{u}dS_aP_0(S_a)\prod_{\mu=1}^{n}P_{\text{out}}\left(y^\mu\middle|\frac{1}{\sqrt{dL}}\text{Tr}(S_a^TX^\mu)\right)\\
&= \mathbb{E}_X\mathbb{E}_{S_0}\mathbb{E}_{y|X,S_0}\int\prod_{a=1}^{u}dS_aP_0(S_a)\prod_{\mu=1}^{n}P_{\text{out}}\left(y^\mu\middle|\frac{1}{\sqrt{dL}}\text{Tr}(S_a^TX^\mu)\right)\\
&= \mathbb{E}_X\int dS_0P_0(S_0)\prod_{\mu=1}^{n}dy^\mu P_{\text{out}}\left(y^\mu\middle|\frac{1}{\sqrt{dL}}\text{Tr}(X^\mu S_0)\right)\int\prod_{a=1}^{u}dS_aP_0(S_a)\prod_{\mu=1}^{n}P_{\text{out}}\left(y^\mu\middle|\frac{1}{\sqrt{dL}}\text{Tr}(S_a^TX^\mu)\right)\\
&= \int\prod_{a=0}^{n}\prod_{\mu=1}^{n}dS_aP_0(S_a)\mathbb{E}_{X^\mu}\int dy^\mu P_{\text{out}}\left(y^\mu\middle|\frac{1}{\sqrt{dL}}\text{Tr}(S_a^TX^\mu)\right)\\
&= \int\prod_{a=0}^{n}\prod_{\mu=1}^{n}dS_aP_0(S_a)\mathbb{E}_{X^\mu}\int dy^\mu\frac{dh^{\mu a}d\hat{h}^{\mu a}}{(2\pi)^{Nr}}P_{\text{out}}\left(y^\mu|h^{\mu a}\right)\exp\left[i\hat{h}^{\mu a}\left(h^{\mu a}-\frac{1}{\sqrt{dL}}\text{Tr}(S_a^TX^\mu)\right)\right]\\
&= \int\prod_{a=0}^{n}\prod_{\mu=1}^{n}dS_aP_0(S_a)\mathbb{E}_{X^\mu}\int dy^\mu d\hat{h}^{\mu a}\hat{P}_{\text{out}}\left(y^\mu,\hat{h}^{\mu a}\right)\exp\left(-i\frac{1}{\sqrt{dL}}\hat{h}^{\mu a}\text{Tr}(S_a^TX^\mu)\right),
\end{aligned}\tag{B4}$$

where we define

$$\hat{P}_{\text{out}}(y,\hat{h}) = \int\frac{dh}{2\pi}P_{\text{out}}(y|h)\exp[i\hat{h}h].\tag{B5}$$

For the Gaussian noise channel $P_{\text{out}}(y|h) = N(y;h,\Delta)$, we have

$$\hat{P}_{\text{out}}(y,\hat{h}) = \int\frac{dh}{2\pi\sqrt{2\pi\Delta}}e^{-[(y-h)^2/2\Delta]+i\hat{h}h} = e^{i\hat{h}y}\int\frac{dx}{2\pi\sqrt{2\pi\Delta}}e^{-(x^2/2\Delta)-i\hat{h}x} = \frac{1}{2\pi}e^{i\hat{h}y-\Delta/2\hat{h}^2},\tag{B6}$$

which reduces to the noiseless case for $\Delta = 0$. Notice also that by normalization $\int dyP_{\text{out}}(y|h) = 1$, implying

$$\int dy\hat{P}_{\text{out}}(y,\hat{h}) = \int dy\frac{dh}{2\pi}P_{\text{out}}(y|h)e^{i\hat{h}h} = \int\frac{dh}{2\pi}e^{i\hat{h}h} = \delta(\hat{h}).\tag{B7}$$

4. Disorder average

We can now perform the average over the data X^μ. We have

$$\begin{aligned}
\mathbb{E}_{X^\mu}\exp\left(-i\frac{1}{\sqrt{dL}}\sum_{a=0}^{u}\hat{h}^{\mu a}\text{Tr}(X^\mu S_a)\right) &= \mathbb{E}_{X^\mu}\exp\left(-i\frac{1}{\sqrt{dL}}\sum_{i=1}^{d}\sum_{j=1}^{L}X_{ij}^\mu\sum_{a=0}^{u}\hat{h}^{\mu a}S_{a,ij}\right)\\
&= \exp\left(-\frac{1}{2dL}\sum_{a,b=0}^{u}\hat{h}^{\mu a}\hat{h}^{\mu b}\sum_{i=1}^{d}\sum_{j=1}^{L}S_{a,ij}S_{b,ij}\right)\\
&= \exp\left(-\frac{1}{2}\sum_{a,b=0}^{u}\hat{h}^{\mu a}\hat{h}^{\mu b}Q(S_a,S_b)\right),
\end{aligned}\tag{B8}$$

where we introduce the overlaps

$$Q(S_a,S_b) = \frac{1}{dL}\sum_{i=1}^{d}\sum_{j=1}^{L}S_{a,ij}S_{b,ij} = \frac{1}{dL}\text{Tr}(S_a^TS_b).\tag{B9}$$

This allows us to rewrite the replicated partition function as

$$\begin{aligned}\mathbb{E}P(\mathcal{D}_n)^u &= \int dS_a dy^\mu d\hat{h}^{\mu a}\left[\prod_a P_0(S_a)\prod_{\mu a}\hat{P}_{\text{out}}\left(y^\mu|\hat{h}^{\mu a}\right)\right]\exp\left(-\frac{1}{2}\sum_{\mu ab}\hat{h}^{\mu a}\hat{h}^{\mu b}Q(S_a,S_b)\right)\\ &= \int \frac{dQ_{ab}d\hat{Q}_{ab}}{(2\pi)^{u(u+1)/2}}dS_a dy^\mu d\hat{h}^{\mu a}\left[\prod_a P_0(S_a)\prod_{\mu a}\hat{P}_{\text{out}}\left(y^\mu|\hat{h}^{\mu a}\right)\right]\\ &\quad\times\exp\left(-\frac{1}{2}\sum_{\mu ab}\hat{h}^{\mu a}\hat{h}^{\mu b}Q_{ab}+ir(d+L)\sum_{a\le b}Q_{ab}\hat{Q}_{ab}-i\frac{r(d+L)}{dL}\sum_{a\le b}\hat{Q}_{ab}\text{Tr}(S_a^TS_b)\right)\\ &= \int \frac{dQ_{ab}d\hat{Q}_{ab}}{(2\pi)^{u(u+1)/2}}\exp\left(ir(d+L)\sum_{a\le b}Q_{ab}\hat{Q}_{ab}\right)\left[\int dyd\hat{h}^a\hat{P}_{\text{out}}(y|\hat{h}^a)\exp\left(-\frac{1}{2}\sum_{ab}\hat{h}^a\hat{h}^bQ_{ab}\right)\right]^n\\ &\quad\times\left[\int dS_aP_0(S_a)\exp\left(-i\frac{r(d+L)}{dL}\sum_{a\le b}\hat{Q}_{ab}\text{Tr}(S_a^TS_b)\right)\right].\end{aligned}\tag{B10}$$

5. Replica symmetric *Ansatz*

It is well known that, in the BO case, the replica symmetric *Ansatz* is correct due to Nishimori's identities [70]. Then, we can take $Q_{00}=Q_0$, $Q_{0a}=m$, $Q_{aa}=Q$, and $Q_{ab}=q$ and $i\hat{Q}_{00}=\hat{Q}_0$, $i\hat{Q}_{0a}=-\hat{m}$, $i\hat{Q}_{aa}=\hat{Q}$, and $i\hat{Q}_{ab}=-\hat{q}$. Nishimori's identities additionally imply $m=q$ and $Q_0=Q$, and similarly for the hat variables. Using these simplifications, we can perform the following rewritings. The algebraic term becomes

$$i\sum_{a\le b}Q_{ab}\hat{Q}_{ab}=(u+1)Q\hat{Q}-\frac{u(u+1)}{2}q\hat{q}\underset{u\to 0}{=}(1+u)Q\hat{Q}-\frac{u}{2}q\hat{q}+O(u^2).\tag{B11}$$

The output channel term can be treated by a standard decoupling trick involving an Hubbard-Stratonovich transformation. One has first the rewriting

$$\sum_{ab}\hat{h}^a\hat{h}^bQ_{ab}=Q\sum_a(\hat{h}^a)^2+q\sum_{a\ne b}\hat{h}^a\hat{h}^b=(Q-q)\sum_a(\hat{h}^a)^2+q\left(\sum_a\hat{h}^a\right)^2\tag{B12}$$

and then the Hubbard-Stratonovich decoupling

$$\exp\left(-\frac{q}{2}\left(\sum_a\hat{h}^a\right)^2\right)\propto\int Dz\exp\left(\sqrt{q}z\sum_a i\hat{h}^a\right),\tag{B13}$$

where Dz denotes integration against a standard Gaussian measure, from which

$$\begin{aligned}\int dyd\hat{h}^a\hat{P}_{\text{out}}(y|\hat{h}^a)\exp\left(-\frac{1}{2}\sum_{ab}\hat{h}^a\hat{h}^bQ_{ab}\right) &= \int Dzdy\left[\int d\hat{h}\hat{P}_{\text{out}}(y|\hat{h})\exp\left(-\frac{Q-q}{2}\hat{h}^2+\sqrt{q}zi\hat{h}\right)\right]^{u+1}\\ &= \int Dzdy I_{\text{out}}(z,y)^{u+1}\\ &\underset{u\to 0}{=}\int Dzdy I_{\text{out}}(z,y)+u\int Dzdy I_{\text{out}}(z,y)\log I_{\text{out}}(z,y)+O(u^2),\end{aligned}\tag{B14}$$

where

$$I_{\text{out}}(z,y)=\int d\hat{h}\hat{P}_{\text{out}}(y|\hat{h})\exp\left(-\frac{Q-q}{2}\hat{h}^2+\sqrt{q}zi\hat{h}\right).\tag{B15}$$

A similar procedure can be repeated for the prior term. We start by

$$-i\sum_{a\le b}\hat{Q}_{ab}\mathrm{Tr}(S_a^T S_b) = -\hat{Q}\sum_a \mathrm{Tr}(S_a^T S_a) + \hat{q}\sum_{a<b}\mathrm{Tr}(S_a^T S_b)$$
$$= -\left(\hat{Q}+\frac{\hat{q}}{2}\right)\sum_a \mathrm{Tr}(S_a^T S_a) + \frac{\hat{q}}{2}\sum_{ab}\mathrm{Tr}(S_a^T S_b), \tag{B16}$$

and then

$$\exp\left(\frac{\hat{q}}{2}\mathrm{Tr}\left(\left(\sum_a S_a\right)^T\left(\sum_b S_b\right)\right)\right) = \int DY \exp\left(-\frac{1}{2}\mathrm{Tr}(Y^T Y) + \sqrt{\hat{q}}\mathrm{Tr}\left(Y^T\left(\sum_a S_a\right)\right)\right), \tag{B17}$$

where Y is a $d\times L$ matrix with standard Gaussian entries. This gives

$$\int dS^a P_0(S^a)\exp\left(-i\frac{r(d+L)}{dL}\sum_{a\le b}\hat{Q}_{ab}\mathrm{Tr}(S_a^T S_b)\right)$$
$$= \int DY\left[\int dS P_0(S)\exp\left(-\frac{r(d+L)}{dL}\left(\hat{Q}+\frac{\hat{q}}{2}\right)\mathrm{Tr}(S^T S) + \sqrt{\frac{r(d+L)}{dL}}\sqrt{\hat{q}}\mathrm{Tr}(Y^T S)\right)\right]^{u+1}$$
$$= \int DY I_0(Y)^{u+1}$$
$$\underset{u\to 0}{=} \int DY I_0(Y) + u\int DY I_0(Y)\log I_0(Y) + O(u^2), \tag{B18}$$

where

$$I_0(Y) = \int dS P_0(S)\exp\left(-\frac{r(d+L)}{dL}\left(\hat{Q}+\frac{\hat{q}}{2}\right)\mathrm{Tr}(S^T S) + \sqrt{\frac{r(d+L)}{dL}}\sqrt{\hat{q}}\mathrm{Tr}(Y^T S)\right). \tag{B19}$$

6. Zero replicas: Q and $\hat{Q}$ equations

For $u=0$, we need to recover the trivial result $\mathbb{E}P(\mathcal{D}_n)^0 = 1$. The replicated partition function in that case equals

$$\mathbb{E}P(\mathcal{D}_n)^0 = \int \frac{dQ_{ab}d\hat{Q}_{ab}}{(2\pi)^{u(u+1)/2}}\exp\left(r(d+L)Q\hat{Q} + n\log\int Dz dy I_{\mathrm{out}}(z,y) + \log\int DY I_0(Y)\right), \tag{B20}$$

where [using Eq. (B7)]

$$\int Dz dy I_{\mathrm{out}}(z,y) = \int Dz dy d\hat{h}\hat{P}_{\mathrm{out}}(y|\hat{h})\exp\left(-\frac{Q-q}{2}\hat{h}^2 + \sqrt{q}z i\hat{h}\right) = 1 \tag{B21}$$

and

$$\int DY I_0(Y) = \int DY dS P_0(S)\exp\left(-\frac{r(d+L)}{dL}\left(\hat{Q}+\frac{\hat{q}}{2}\right)\mathrm{Tr}(S^T S) + \sqrt{\frac{r(d+L)}{dL}}\sqrt{\hat{q}}\mathrm{Tr}(Y^T S)\right)$$
$$= \int dS P_0(S)\exp\left(-\frac{r(d+L)}{dL}\left(\hat{Q}+\frac{\hat{q}}{2}\right)\mathrm{Tr}(S^T S) + \frac{r(d+L)}{dL}\frac{\hat{q}}{2}\mathrm{Tr}(S^T S)\right)$$
$$= \int dS P_0(S)\exp\left(-\frac{r(d+L)}{dL}\hat{Q}\mathrm{Tr}(S^T S)\right). \tag{B22}$$

Taking a saddle-point approximation on the scale $r(d+L)$ gives two equations for Q and $\hat{Q}$, namely, $\hat{Q}=0$ and

$$Q = \frac{1}{r(d+L)} \frac{\int dS P_0(S) \exp\left(-\hat{Q}\mathrm{Tr}(S^T S)\right) \frac{r(d+L)}{dL} \mathrm{Tr}(S^T S)}{\int dS P_0(S) \exp\left(-\hat{Q}\mathrm{Tr}(S^T S)\right)} \underset{\hat{Q}=0}{=} \frac{1}{dL} \int dS P_0(S) \mathrm{Tr}(S^T S) = Q_*. \tag{B23}$$

This, in turn, implies that at the saddle point

$$\int DY I_0(Y) = \int dS P_0(S) = 1. \tag{B24}$$

7. Free entropy

At the first nontrivial order in $u \to 0$, the replicated partition function reads (using the above results for Q and $\hat{Q}$ and dropping irrelevant constant factors)

$$\mathbb{E}P(\mathcal{D}_n)^u \propto \int dq d\hat{q} \exp\left[ur(d+L)\left(-\frac{1}{2}q\hat{q} + \bar{\alpha}\log\int DY I_0(Y)\log I_0(Y) + \frac{1}{r(d+L)}\log\int Dz dy I_{\mathrm{out}}(z,y)\log I_{\mathrm{out}}(z,y)\right)\right], \tag{B25}$$

from which we get

$$\Phi = extr_{q,\hat{q}}\left(-\frac{1}{2}q\hat{q} + \frac{1}{r(d+L)}\int DY I_0(Y)\log I_0(Y) + \bar{\alpha}\int Dz dy I_{\mathrm{out}}(z,y)\log I_{\mathrm{out}}(z,y)\right). \tag{B26}$$

8. Equation with respect to q

Taking the derivative with respect to q gives

$$\begin{aligned}\hat{q} &= 2\bar{\alpha}\partial_q \int Dz dy I_{\mathrm{out}}(z,y)\log I_{\mathrm{out}}(z,y) \\ &= 2\bar{\alpha}\int Dz dy(1+\log I_{\mathrm{out}}(z,y))\partial_q I_{\mathrm{out}}(z,y).\end{aligned} \tag{B27}$$

Now, using Eq. (19) in Ref. [42], we get

$$\begin{aligned}\partial_q I_{\mathrm{out}}(z,y) &= \partial_q \int dh \mathcal{N}(h;\sqrt{q}z;Q_*-q)P_{\mathrm{out}}\left(y|h\right) \\ &= \frac{e^{z^2/2}}{2q}\partial_z\left[e^{-z^2/2}\partial_z\int dh\mathcal{N}(h;\sqrt{q}z;Q_*-q)P_{\mathrm{out}}\left(y|h\right)\right] \\ &= \frac{e^{z^2/2}}{2q}\partial_z[e^{-z^2/2}\partial_z I_{\mathrm{out}}(z,y)],\end{aligned} \tag{B28}$$

giving

$$\begin{aligned}\hat{q} &= 2\bar{\alpha}\int Dzdy(1+\log I_{\text{out}}(z,y))\partial_q I_{\text{out}}(z,y)\\ &= -2\bar{\alpha}\int Dzdy(1+\log I_{\text{out}}(z,y))\frac{e^{z^2/2}}{2q}\partial_z\left[e^{-z^2/2}\partial_z I_{\text{out}}(z,y)\right]\\ &= -\frac{\bar{\alpha}}{q\sqrt{2\pi}}\int dzdy(1+\log I_{\text{out}}(z,y))\partial_z\left[e^{-z^2/2}\partial_z I_{\text{out}}(z,y)\right]\\ &= \frac{\bar{\alpha}}{q\sqrt{2\pi}}\int dzdy\partial_z(1+\log I_{\text{out}}(z,y))\left[e^{-z^2/2}\partial_z I_{\text{out}}(z,y)\right]\\ &= \frac{\bar{\alpha}}{q}\int Dzdy\frac{(\partial_z I_{\text{out}}(z,y))^2}{I_{\text{out}}(z,y)},\end{aligned} \tag{B29}$$

which matches Eqs. (60) and (76) in Ref. [42] modulo a different definition of the sample ratio $\bar{\alpha}$. For the Gaussian noise channel, we have

$$\begin{aligned}I_{\text{out}}(z,y) &= \int\frac{d\hat{h}}{2\pi}\exp\left(-\frac{\Delta+Q-q}{2}\hat{h}^2+(\sqrt{q}z+y)i\hat{h}\right)\\ &= \frac{1}{\sqrt{2\pi(\Delta+Q-q)}}\exp\left(-\frac{(\sqrt{q}z+y)^2}{2(\Delta+Q-q)}\right),\end{aligned} \tag{B30}$$

from which

$$\begin{aligned}\hat{q} &= \frac{\bar{\alpha}}{q}\int Dzdy\frac{(\partial_z I_{\text{out}}(z,y))^2}{I_{\text{out}}(z,y)}\\ &= \bar{\alpha}\int Dzdy\frac{\frac{(\sqrt{q}z+y)^2}{(\Delta+Q-q)^2}}{\sqrt{2\pi(\Delta+Q-q)}}\exp\left(-\frac{(\sqrt{q}z+y)^2}{2(\Delta+Q-q)}\right)\\ &= \bar{\alpha}\int dt\frac{\frac{t^2}{(\Delta+Q-q)^2}}{\sqrt{2\pi(\Delta+Q-q)}}\exp\left(-\frac{t^2}{2(\Delta+Q-q)}\right)\\ &= \frac{\bar{\alpha}}{\Delta+Q_*-q}.\end{aligned} \tag{B31}$$

We also have, for the free entropy term and Gaussian output channel,

$$\begin{aligned}\int Dzdy I_{\text{out}}(z,y)\log I_{\text{out}}(z,y) &= \int Dzdy\frac{\exp\left(-\frac{(\sqrt{q}z+y)^2}{2(\Delta+Q-q)}\right)}{\sqrt{2\pi(\Delta+Q-q)}}\log\frac{\exp\left(-\frac{(\sqrt{q}z+y)^2}{2(\Delta+Q-q)}\right)}{\sqrt{2\pi(\Delta+Q-q)}}\\ &= \int Dt\log\frac{\exp\left(-\frac{t^2}{2}\right)}{\sqrt{2\pi(\Delta+Q-q)}}\\ &= \int Dt\left(-\frac{t^2}{2}-\frac{1}{2}\log\left(2\pi(\Delta+Q-q)\right)\right)\\ &= -\frac{1}{2}-\frac{1}{2}\log\left(2\pi(\Delta+Q-q)\right)\\ &= -\frac{1}{2}\log\left(\Delta+Q-q\right)+\cdots,\end{aligned} \tag{B32}$$

where we change variable $y\to(y+\sqrt{q}z)/\sqrt{\Delta+Q-q}=t$ and keep only the order-parameter-dependent terms in the last passage.

9. Equation with respect to $\hat{q}$ and a denoising subproblem

The nontrivial part of the free entropy, involving $\hat{q}$, is

$$\frac{1}{r(d+L)}\int DY I_0(Y)\log I_0(Y), \tag{B33}$$

where (recall that $\beta = L/d$)

$$I_0(Y) = \int dS P_0(S)\exp\left(-\frac{r(d+L)\hat{q}}{dL}\frac{1}{2}\mathrm{Tr}(S^T S) + \sqrt{\frac{r(d+L)\hat{q}}{dL}}\mathrm{Tr}(Y^T S)\right). \tag{B34}$$

We recognize that this quantity is proportional to the entropy $H(Y) = -\mathbb{E}_Y \log P(Y)$ of the observation Y in the denoising problem

$$Y = \sqrt{\frac{r(d+L)\hat{q}}{dL}}S_* + Z \tag{B35}$$

with prior $S \sim P_0$ and additive i.i.d. standard Gaussian noise Z, by the identification $I_0(Y) = P(Y)$ [notice that we show in Eq. (B24) that I_0 is properly normalized]. Thus, $\hat{q}$ plays the role of a rescaled signal-to-noise ratio. Moreover, we have that by the I-MMSE theorem the derivative with respect to $\hat{q}$ of the denoising observation entropy will be related to the MMSE of the same denoising problem.

Thus, to proceed we need to solve the asymptotics of this denoising free entropy for the prior (10) and, in particular, of the associated MMSE. This requires different approaches based on the scaling of r. If $r \ll d$, the computation is already in the literature [42], but we reproduce it here for convenience and generalize it to correlated priors. If $r = O(d)$, the computation for the denoiser has been carried out in Ref. [55], but its application to our problem is novel.

10. Equation with respect to $\hat{q}$ in the extensive-width case

In the regime where $r = \rho \min(L, d)$ with constant ρ, we use the results from Ref. [55] (which more generally apply to all rotationally invariant priors). Again, in this subsection, we assume that $d \le L$, understanding that the general case is retrieved by substituting $d \to \min(L, d)$. It is best to rescale the various quantities to match those in Ref. [55] to easily adapt their results. We define

$$\delta(\hat{q}) = \frac{\beta}{\rho(1+\beta)\hat{q}}. \tag{B36}$$

We want to compute

$$\mathcal{I} = \frac{1}{\rho(1+\beta)d^2}\int DY I_0(Y)\log I_0(Y), \tag{B37}$$

where

$$I_0(Y) = \int dS P_0(S)\exp\left(-\frac{1}{2\delta}\mathrm{Tr}(S^T S) + \sqrt{\frac{1}{\delta}}\mathrm{Tr}(Y^T S)\right). \tag{B38}$$

We perform the change of variable $Y = \underline{Y}\sqrt[4]{Ld}/\sqrt{\delta}$, $S = \underline{S}\sqrt[4]{Ld}$, and $S_* = \underline{S}_*\sqrt[4]{Ld}$ and get

$$\begin{aligned}\mathcal{I} = {} & \frac{1}{\rho(1+\beta)d^2}\int d\underline{Y} d\underline{S}_* \underline{P}_0(\underline{S}_*)(2\pi\delta/\sqrt{Ld})^{-Ld/2}\exp\left(-\frac{\sqrt{Ld}}{2\delta}\mathrm{Tr}[(\underline{Y}-\underline{S}_*)^T(\underline{Y}-\underline{S}_*)]\right)\\ & \times \log\int d\underline{S}\underline{P}_0(\underline{S})(2\pi\delta/\sqrt{Ld})^{-Ld/2}\exp\left(-\frac{\sqrt{Ld}}{2\delta}\mathrm{Tr}[(\underline{Y}-\underline{S})^T(\underline{Y}-\underline{S})]\right)\\ & + \frac{\sqrt{\beta}}{2\rho(1+\beta)\delta d}\int d\underline{Y} d\underline{S}_* \underline{P}_0(\underline{S}_*)N(\underline{Y};\underline{S}_*,\delta/\sqrt{Ld})\mathrm{Tr}(\underline{Y}^T\underline{Y}) + \frac{1}{\rho(1+\beta)d^2}\frac{Ld}{2}\log\frac{2\pi\delta}{\sqrt{Ld}},\end{aligned} \tag{B39}$$

where $\underline{P}_0$ is the rescaled prior, still normalized and such that its samples have $\mathcal{O}(1)$ spectral density. Notice that after the rescaling we have the denoising problem $\underline{Y} = \underline{S} + \sqrt{\delta(\hat{q})}\underline{Z}$. The second term can be simplified to

$$\frac{\sqrt{\beta}}{2\rho(1+\beta)\delta d}\int d\underline{Y}d\underline{S}_*\underline{P}_0(\underline{S}_*)N(\underline{Y};\underline{S}_*,\delta/\sqrt{Ld})\mathrm{Tr}(\underline{Y}^T\underline{Y}) = \frac{\sqrt{\beta}}{2\rho(1+\beta)\delta d}\int d\underline{S}_*\underline{P}_0(\underline{S}_*)\left(\delta\sqrt{Ld}+\mathrm{Tr}(\underline{S}_*^T\underline{S}_*)\right) = \frac{\beta(\delta+Q_*)}{2\rho(1+\beta)\delta}, \tag{B40}$$

so that

$$\begin{aligned}\mathcal{I} = &\frac{\beta(\delta+Q_*)}{2\rho(1+\beta)\delta}+\frac{\beta}{2\rho(1+\beta)}\log\frac{2\pi\delta}{\sqrt{Ld}}\\ &+\frac{1}{\rho(1+\beta)d^2}\int d\underline{Y}d\underline{S}_*\underline{P}_0(\underline{S}_*)(2\pi\delta/\sqrt[4]{Ld})^{-Ld/2}\exp\left(-\frac{\sqrt{Ld}}{2\delta}\mathrm{Tr}[(\underline{Y}-\underline{S}_*)^T(\underline{Y}-\underline{S}_*)]\right)\\ &\times\log\int d\underline{S}\underline{P}_0(\underline{S})\exp\left(-\frac{\sqrt{Ld}}{2\delta}\mathrm{Tr}[(\underline{Y}-\underline{S})^T(\underline{Y}-\underline{S})]\right).\end{aligned} \tag{B41}$$

We now recognize that the third term is proportional to Eq. (15) in Ref. [55], which is the free entropy of the denoising problem with noise-to-signal ratio $\sqrt{\delta}$. We just need to be careful of a factor β, as in Ref. [55] the free entropy is normalized by $1/Ld$ and not by $1/d^2$. Thus, we can use directly Eq. (18) in Ref. [55] to get (one needs to set $\beta=\beta$)

$$\mathcal{I} = \frac{\beta(\delta+Q_*)}{2\rho(1+\beta)\delta}+\frac{\beta}{2\rho(1+\beta)}\log\frac{2\pi\delta}{\sqrt[4]{Ld}}+\frac{\beta}{\rho(1+\beta)}\left[\mathrm{const}(\beta,\underline{P}_0)-\frac{1}{\beta}\Sigma[\hat{\mu}_{\underline{Y}}]-\frac{\beta-1}{\beta}\Lambda[\hat{\mu}_{\underline{Y}}]\right], \tag{B42}$$

where $\hat{\mu}_{\underline{Y}}$ is the symmetrized singular value density of a matrix distributed as $\underline{Y}$ in the high-dimensional limit and

$$\Sigma[\hat{\mu}_{\underline{Y}}] = \iint dxdy\hat{\mu}_{\underline{Y}}(x)\hat{\mu}_{\underline{Y}}(y)\log|x-y| \quad\text{and}\quad \Lambda[\hat{\mu}_{\underline{Y}}] = \int dx\hat{\mu}_{\underline{Y}}(x)\log|x| \tag{B43}$$

regularized as in Appendix A in Ref. [55]. Thus, translating this back into a function of $\hat{q}$, we get, dropping all $\hat{q}$ independent terms,

$$\mathcal{I} = \mathrm{const}+\frac{Q_*\hat{q}}{2}-\frac{\beta}{2\rho(1+\beta)}\log(\hat{q})-\frac{\beta}{\rho(1+\beta)}\left[\frac{1}{\beta}\Sigma[\hat{\mu}_{\underline{S}+\sqrt{\delta(\hat{q})}\underline{Z}}]+\frac{\beta-1}{\beta}\Lambda[\hat{\mu}_{\underline{S}+\sqrt{\delta(\hat{q})}\underline{Z}}]\right]. \tag{B44}$$

Then, taking the derivative with respect to $\hat{q}$, we obtain the second state equation

$$\begin{aligned}q &= 2\partial_{\hat{q}}\mathcal{I}\\ &= Q_*-\frac{\beta}{\rho(1+\beta)\hat{q}}+\frac{2\beta^2}{\rho^2(1+\beta)^2\hat{q}^2}\partial_\delta\left[\frac{1}{\beta}\Sigma[\hat{\mu}_{\underline{S}+\sqrt{\delta}\underline{Z}}]+\frac{\beta-1}{\beta}\Lambda[\hat{\mu}_{\underline{S}+\sqrt{\delta}\underline{Z}}]\right]_{\delta=\beta/\rho(1+\beta)\hat{q}}\\ &= Q_*-\delta(\hat{q})+\delta(\hat{q})^2\int dx\hat{\mu}_{\underline{S}+\sqrt{\delta(\hat{q})}\underline{Z}}(x)\left[\frac{(\beta-1)^2}{\beta^{3/2}x^2}+\frac{4\pi^2}{3\beta^{3/2}}\hat{\mu}_{\underline{S}+\sqrt{\delta(\hat{q})}\underline{Z}}(x)^2\right],\end{aligned} \tag{B45}$$

where the derivative of the spectral term was computed using Eqs. (35), (70), and (71) in Ref. [55].

We also notice that

$$\partial_\delta\left[\frac{1}{\beta}\Sigma[\hat{\mu}_{\underline{Y}}]+\frac{\beta-1}{\beta}\Lambda[\hat{\mu}_{\underline{Y}}]\right] = -\partial_\delta\Phi_{\mathrm{denoising}}(\delta) = \frac{\delta-\mathrm{MMSE}_{\mathrm{denoising}}(\delta)}{2\delta^2}, \tag{B46}$$

where we use Eq. (14) in Ref. [55], so that the state equation can be rewritten as

$$q = Q_* - \delta(\hat{q}) + \delta(\hat{q})^2 \frac{\delta(\hat{q}) - \mathrm{MMSE}_{\mathrm{denoising}}[\delta(\hat{q})]}{\delta(\hat{q})^2}$$
$$= Q_* - \mathrm{MMSE}_{\mathrm{denoising}}[\delta(\hat{q})]. \tag{B47}$$

The only nontrivial ingredient needed to solve the state equations is the symmetrized singular value density of $\underline{Y}$. This can be computed numerically efficiently, as detailed in Sec. III. 3 and Appendix F in Ref. [55], setting $R_1 = \beta$, $R_2 = \rho$ and $\Delta = \delta(\hat{q})$ for the factorized Gaussian prior. For generic rotationally invariant priors, this spectral density may be difficult to compute accurately.

It is also useful to write the associated free entropy. One gets, discarding all terms that are not dependent on the order parameters (and using $Q = Q_* = 1$),

$$\Phi \approx -\frac{\bar{\alpha}}{2}\log(\Delta + 1 - q) + \frac{\hat{q}(1-q)}{2} - \frac{\beta \log(\hat{q})}{2\rho(1+\beta)} - \frac{\beta}{\rho(1+\beta)}\left[\frac{1}{\beta}\Sigma[\hat{\mu}_{\underline{S}+\sqrt{\delta(\hat{q})}\underline{Z}}] + \frac{\beta-1}{\beta}\Lambda[\hat{\mu}_{\underline{S}+\sqrt{\delta(\hat{q})}\underline{Z}}]\right]. \tag{B48}$$

Notice also that by substituting back $\bar{\alpha} = \{\beta/[\rho(1+\beta)]\}\alpha$ and rescaling $\hat{q} \to \hat{q}\{\beta/[\rho(1+\beta)]\}$ we get

$$\frac{\rho(1+\beta)}{\beta}\Phi \approx -\frac{\alpha}{2}\log(\Delta + 1 - q) + \frac{\hat{q}(1-q)}{2} - \frac{\log(\hat{q})}{2} - \left[\frac{1}{\beta}\Sigma[\hat{\mu}_{\underline{S}+\sqrt{1/\hat{q}}\underline{Z}}] + \frac{\beta-1}{\beta}\Lambda[\hat{\mu}_{\underline{S}+\sqrt{1/\hat{q}}\underline{Z}}]\right]. \tag{B49}$$

In the case of Gaussian label noise, we get $\hat{q} = \alpha/(1 - q + \Delta)$ so that (again discarding all terms not dependent on the order parameter $\hat{q}$, the only one left now)

$$\frac{\rho(1+\beta)}{\beta}\Phi \approx \frac{\alpha-1}{2}\log(\hat{q}) - \frac{\hat{q}\Delta}{2} - \left[\frac{1}{\beta}\Sigma[\hat{\mu}_{\underline{S}+\sqrt{1/\hat{q}}\underline{Z}}] + \frac{\beta-1}{\beta}\Lambda[\hat{\mu}_{\underline{S}+\sqrt{1/\hat{q}}\underline{Z}}]\right]. \tag{B50}$$

11. Equation with respect to $\hat{q}$ in the low-width case

In the regime where $r \ll d$, we rederive results from matrix compressed sensing [42]. We assume without loss of generality $d \le L$. Not assuming this is equivalent to substituting $d \to \min(L, d)$ everywhere in this subsection. We want to compute

$$\mathcal{I} = \frac{1}{r(d+L)}\int DY I_0(Y)\log I_0(Y), \tag{B51}$$

where (recall that $\beta = L/d$)

$$I_0(Y) = \int dS P_0(S)\exp\left(-\frac{(1+\beta)\hat{q}}{\beta}\frac{r}{d}\frac{1}{2}\mathrm{Tr}(S^T S) + \sqrt{\frac{(1+\beta)\hat{q}}{\beta}\frac{r}{d}}\mathrm{Tr}(Y^T S)\right). \tag{B52}$$

We consider factorized priors $S = AB^T/\sqrt{r}$ with $A \in \mathbb{R}^{L\times r}$ and $B \in \mathbb{R}^{d\times r}$, where the priors on A and B are row-factorized but arbitrarily correlated along the width r through distributions G_A and G_B. A special case is that of fully factorized prior, already treated in Ref. [42], among which one finds the i.i.d. Gaussian priors in which we are interested in the main text. To start, we have

$$\mu_2(0) = \frac{1}{rdL}\mathbb{E}_{S\sim P_0}\mathrm{Tr}(S^T S)$$
$$= \frac{1}{rdL}\mathbb{E}_{A,B}\sum_{ijkl} A_{ik}B_{jk}A_{il}B_{jl}$$
$$= \frac{1}{r}\sum_{kl}\mathbb{E}_A[A_{\cdot k}A_{\cdot l}]\mathbb{E}_B[B_{\cdot k}B_{\cdot l}]$$
$$= \frac{1}{r}\sum_{kl}\Sigma^A_{kl}\Sigma^B_{kl}, \tag{B53}$$

where $\Sigma^{A,B}$ are the column covariances of A and B in $\mathbb{R}^r$. To compute $\mathcal{I}$ and deal with the logarithm, we need to replicate again. We replicate s times, so that we need to compute $\mathcal{I} = \lim_{s\to 0}(\mathcal{I}_s - 1)/s$ with

$$\begin{aligned}\mathcal{I}_s &= \int DY I_0(Y)^{s+1} \\ &= \int \left(\prod_a dS_a P_0(S_a)\right) \exp\left(-\frac{(1+\beta)\hat{q}}{\beta}\frac{1}{2}\frac{r}{d}\sum_a \mathrm{Tr}(S_a^T S_a)\right) \int DY \exp\left(\sqrt{\frac{(1+\beta)\hat{q}}{\beta}}\frac{r}{d}\sum_{ij} Y_{ij}\sum_a S_{ij}^a\right) \\ &= \int \left(\prod_a dS_a P_0(S_a)\right) \exp\left(\frac{r}{d}\frac{(1+\beta)\hat{q}}{\beta}\sum_{a<b}\mathrm{Tr}(S_a^T S_b)\right). \end{aligned} \tag{B54}$$

Then, we use the prior

$$dS_a P_0(S_a) = dA_a dB_a P_A(A_a) P_B(B_a) \tag{B55}$$

to get

$$\mathrm{Tr}(S_a^T S_b) = \sum_{ij} S_{ij}^a S_{ij}^b = \frac{1}{r}\sum_{ijkl} A_{ik}^a B_{jk}^a A_{il}^b B_{jl}^b = \frac{1}{r}\sum_{kl}\left(\sum_i A_{ik}^a A_{il}^b\right)\left(\sum_j B_{jk}^a B_{jl}^b\right), \tag{B56}$$

where the terms we highlighted inside the parentheses are summed over an extensive coordinate i, j and have two free indices k, l that run up to r, which are intensive quantities. Thus, we can introduce two overlaps (call them g instead of q to avoid confusion with the overlaps of the original problem)

$$\begin{aligned} g_{ab,kl}^A &= \frac{1}{d}\sum_{i=1}^d A_{ik}^a A_{il}^b, \\ g_{ab,kl}^B &= \frac{1}{L}\sum_{j=1}^L B_{jk}^a B_{jl}^b, \end{aligned} \tag{B57}$$

giving

$$\frac{r}{d}\mathrm{Tr}(S_a^T S_b) = \beta d \sum_{kl} g_{ab,kl}^A(A) g_{ab,kl}^B(B). \tag{B58}$$

Thus, we have

$$\begin{aligned}\mathcal{I}_s &= \int \left(\prod_a dS_a P_0(S_a)\right) \exp\left(\frac{r}{d}\frac{(1+\beta)\hat{q}}{\beta}\sum_{a<b}\mathrm{Tr}(S_a^T S_b)\right) \\ &= \int \left(\prod_{a<b,kl} dg_{ab,kl}^A dg_{ab,kl}^B d\hat{g}_{ab,kl}^A d\hat{g}_{ab,kl}^B\right) \\ &\quad \times \exp\left(d(1+\beta)\hat{q}\sum_{a<b}\sum_{kl} g_{ab,kl}^A g_{ab,kl}^B - id\sum_{a<b,kl} g_{ab,kl}^A \hat{g}_{ab,kl}^A - id\beta\sum_{a<b,kl} g_{ab,kl}^B \hat{g}_{ab,kl}^B\right) \\ &\quad \times \int \left(\prod_a dA_a dB_a P_A(A_a) P_B(B_a)\right) \exp\left(i\sum_{a<b,kl}\hat{g}_{ab,kl}^A \sum_{i=1}^d A_{ik}^a A_{il}^b + i\sum_{a<b,kl}\hat{g}_{ab,kl}^B \sum_{i=1}^L B_{ik}^a B_{il}^b\right). \end{aligned} \tag{B59}$$

We take notice that for $s = 0$ the argument of the exp vanishes, and $\mathcal{I}_{s=0} = 1$ as it should. Now we can perform the replica symmetric *Ansatz*, with no diagonal as that simplified away and using Nishimori's identities to avoid having to single out the zeroth replica, to get

$$\mathcal{I}_s = \int \left(\prod_{kl} dg^A_{kl} dg^B_{kl} d\hat{g}^A_{kl} d\hat{g}^B_{kl}\right) \times \exp\left(d(1+\beta)\hat{q}\frac{s(s+1)}{2}\sum_{kl} g^A_{kl} g^B_{kl} - d\frac{s(s+1)}{2}\sum_{kl} g^A_{kl}\hat{g}^A_{kl} - d\beta\frac{s(s+1)}{2}\sum_{kl} g^B_{kl}\hat{g}^B_{kl}\right) \times \int \left(\prod_a dA_a dB_a P_A(A_a) P_B(B_a)\right) \exp\left(\sum_{kl}\hat{g}^A_{kl}\sum_i \sum_{a<b} A^a_{ik} A^b_{il} + \sum_{kl}\hat{g}^B_{kl}\sum_i\sum_{a<b} B^a_{ik} B^b_{il}\right). \tag{B60}$$

Now we use the assumption that the priors on A, B are row-factorized, call the prior on a row G (it is a distribution over $\mathbb{R}^r$, so intensive), and get

$$\mathcal{I}_s = \int \left(\prod_{kl} dg^A_{kl} dg^B_{kl} d\hat{g}^A_{kl} d\hat{g}^B_{kl}\right) \times \exp\left(d(1+\beta)\hat{q}\frac{s(s+1)}{2}\sum_{kl} g^A_{kl} g^B_{kl} - d\frac{s(s+1)}{2}\sum_{kl} g^A_{kl}\hat{g}^A_{kl} - d\beta\frac{s(s+1)}{2}\sum_{kl} g^B_{kl}\hat{g}^B_{kl}\right) \times \left[\int \left(\prod_a da_a G_A(a_a)\right) \exp\left(\sum_{kl}\hat{g}^A_{kl}\sum_{a<b} a^a_k a^b_l\right)\right]^d \times \left[\int \left(\prod_a db_a G_B(b_a)\right) \exp\left(\sum_{kl}\hat{g}^B_{kl}\sum_{a<b} b^a_k b^b_l\right)\right]^{\beta d}, \tag{B61}$$

where a and b are rows of the respective matrices. Finally, we need to decouple the replicas. We use

$$\begin{aligned}\exp\left(\sum_{kl}\hat{g}^B_{kl}\sum_{a<b} b^a_k b^b_l\right) &= \exp\left(\sum_{kl}\hat{g}^B_{kl}\left[\frac{1}{2}\sum_{a,b} - \frac{1}{2}\sum_a\right] b^a_k b^b_l\right) \\ &= \exp\left(-\frac{1}{2}\sum_{kl}\hat{g}^B_{kl}\sum_a b^a_k b^a_l\right)\exp\left(\frac{1}{2}\sum_{kl}\left(\sum_a b^a_k\right)\hat{g}^B_{kl}\left(\sum_a b^a_l\right)\right) \\ &= \exp\left(-\frac{1}{2}\sum_{kl}\hat{g}^B_{kl}\sum_a b^a_k b^a_l\right)\int dz\mathcal{N}(z,0,\hat{g}^B_{kl})\exp\left(\sum_k z_k\sum_a b^a_k\right) \\ &= \int dz\mathcal{N}(z,0,\hat{g}^B_{kl})\left[\exp\left(-\frac{1}{2}\sum_{kl}\hat{g}^B_{kl} b_k b_l + \sum_k z_k b_k\right)\right]^{s+1},\end{aligned} \tag{B62}$$

where $\mathcal{N}(x,\mu,\Sigma)$ to denote the Gaussian density with given mean μ and covariance Σ, so that

$$\mathcal{I}_s = \int \left(\prod_{kl} dg^A_{kl} dg^B_{kl} d\hat{g}^A_{kl} d\hat{g}^B_{kl}\right) \times \exp\left(d(1+\beta)\hat{q}\frac{s(s+1)}{2}\sum_{kl} g^A_{kl} g^B_{kl} - d\frac{s(s+1)}{2}\sum_{kl} g^A_{kl}\hat{g}^A_{kl} - d\beta\frac{s(s+1)}{2}\sum_{kl} g^B_{kl}\hat{g}^B_{kl}\right) \times \left[\int dz\mathcal{N}(z,0,\hat{g}^A_{kl})\left[\int da G_A(a)\exp\left(-\frac{1}{2}\sum\nolimits_{kl}\hat{g}^A_{kl} a_k a_l + \sum\nolimits_k z_k a_k\right)\right]^{s+1}\right]^d \times \left[\int dz\mathcal{N}(z,0,\hat{g}^B_{kl})\left[\int db G_B(b)\exp\left(-\frac{1}{2}\sum_{kl}\hat{g}^B_{kl} b_k b_l + \sum_k z_k b_k\right)\right]^{s+1}\right]^{d\beta}. \tag{B63}$$

Now we take $s \to 0$ and get

$$\begin{aligned}
&\int dz\mathcal{N}(z,0,\hat{g}^A_{kl})\left[\int daG_A(a)\exp\left(-\frac{1}{2}\sum_{kl}\hat{g}^A_{kl}a_ka_l+\sum_k z_ka_k\right)\right]^{s+1}\\
&=\int dz\mathcal{N}(z,0,\hat{g}^A_{kl})\left[\int daG_A(a)\exp\left(-\frac{1}{2}\sum_{kl}\hat{g}^A_{kl}a_ka_l+\sum_k z_ka_k\right)\right]\\
&\quad\times\left[1+s\log\int daG_A(a)\exp\left(-\frac{1}{2}\sum_{kl}\hat{g}^A_{kl}a_ka_l+\sum_k z_ka_k\right)+\cdots\right]\\
&=1+s\int dz\mathcal{N}(z,0,\hat{g}^A_{kl})\left[\int daG_A(a)\exp\left(-\frac{1}{2}\sum_{kl}\hat{g}^A_{kl}a_ka_l+\sum_k z_ka_k\right)\right]\\
&\quad\times\log\left[\int daG_A(a)\exp\left(-\frac{1}{2}\sum_{kl}\hat{g}^A_{kl}a_ka_l+\sum_k z_ka_k\right)\right]+\cdots\\
&=1+s\Phi_A(\hat{g}^A)+\cdots,
\end{aligned}\tag{B64}$$

where $\Phi_A(\hat{g}^A)$ is related to an r-dimensional vector denoising problem with Gaussian noise and prior G_A. More importantly, it is just an r-dimensional integral and $r \ll d$. Thus, we get

$$\begin{aligned}
\mathcal{I}_s=\int\left(\prod_{kl}dg^A_{kl}dg^B_{kl}d\hat{g}^A_{kl}d\hat{g}^B_{kl}\right)\exp sd\Bigg(&\frac{\hat{q}(1+\beta)}{2}\sum_{kl}g^A_{kl}g^B_{kl}-\frac{1}{2}\sum_{kl}g^A_{kl}\hat{g}^A_{kl}-\frac{\beta}{2}\sum_{kl}g^B_{kl}\hat{g}^B_{kl}\\
&+\Phi_A(\hat{g}^A)+\beta\Phi_B(\hat{g}^B)+\cdots\Bigg)\Bigg),
\end{aligned}\tag{B65}$$

from which

$$\mathcal{I}=\frac{1}{r(1+\beta)}\operatorname*{extr}_{g^A,g^B,\hat{g}^A,\hat{g}^B}\left[\frac{\hat{q}(1+\beta)}{2}\sum_{kl}g^A_{kl}g^B_{kl}-\frac{1}{2}\sum_{kl}g^A_{kl}\hat{g}^A_{kl}-\frac{\beta}{2}\sum_{kl}g^B_{kl}\hat{g}^B_{kl}+\Phi_{G^A}(\hat{g}^A_{kl})+\beta\Phi_{G^B}(\hat{g}^B_{kl})\right],\tag{B66}$$

where for any (symmetric, positive semidefinite) $r\times r$ matrix $\hat{g}$ we define

$$\begin{aligned}
\Phi_G(\hat{g})=&\int dz\mathcal{N}(z,0,\hat{g})\left[\int daG(a)\exp\left(-\frac{1}{2}\sum_{kl}\hat{g}_{kl}a_ka_l+\sum_k z_ka_k\right)\right]\\
&\times\log\left[\int daG(a)\exp\left(-\frac{1}{2}\sum_{kl}\hat{g}_{kl}a_ka_l+\sum_k z_ka_k\right)\right].
\end{aligned}\tag{B67}$$

The extremization conditions for $\mathcal{I}$ lead to the equations

$$\begin{aligned}
\hat{g}^A_{kl}&=\hat{q}(1+\beta)g^B_{kl},\\
\hat{g}^B_{kl}&=\hat{q}\frac{1+\beta}{\beta}g^A_{kl},\\
g^A_{kl}&=2\partial_{\hat{g}^A_{kl}}\Phi_A(\hat{g}^A),\\
g^B_{kl}&=2\partial_{\hat{g}^B_{kl}}\Phi_B(\hat{g}^B).
\end{aligned}\tag{B68}$$

At the extremizer of $\mathcal{I}$, the state equation for the overlap of the original problem reads

$$q=\frac{1}{r}\sum_{kl}g^A_{kl}g^B_{kl}.\tag{B69}$$

In the special case of factorized priors, Ref. [42] showed that one can take the order parameters g and $\hat{g}$ to be diagonal and with all diagonal elements equal. In that case, one can show by explicit computation that

$$\Phi(\hat{g}) = r\int Dt da_0 G(a_0)\log\left[\int daG(a)\exp\left(-\frac{1}{2}\hat{g}a^2 + (\sqrt{\hat{g}}t + \hat{g}a_0)a\right)\right],$$
$$2\partial_{\hat{g}}\Phi_G(\hat{g}) = \frac{r}{\sqrt{\hat{g}}}\int dz\frac{(\int daG(a)a\mathcal{N}(a, z/\sqrt{\hat{g}}, 1/\hat{g}))^2}{\int daG(a)\mathcal{N}(a, z/\sqrt{\hat{g}}, 1/\hat{g})}, \tag{B70}$$

where now all g and $\hat{g}$ are scalars and which leads to Eqs. (76)–(78) in Ref. [42]. Notice that in the extremization conditions an additional factor r comes out due to the sums for k, l in $\mathcal{I}$ trivializing. Finally, for Gaussian priors one can solve all integrals in closed form, obtaining the equations

$$\begin{aligned} q &= g^A g^B, \\ \hat{g}^A_{kl} &= \hat{q}(1+\beta)g^B_{kl}, \\ \hat{g}^B_{kl} &= \hat{q}\frac{1+\beta}{\beta}g^A_{kl}, \\ g^A_{kl} &= \frac{\hat{g}^A}{1+\hat{g}^A}, \\ g^B_{kl} &= \frac{\hat{g}^B}{1+\hat{g}^B}, \end{aligned} \tag{B71}$$

which can be solved explicitly to

$$\begin{aligned} g^A &= \frac{(\beta+1)^2\hat{q}^2 - \beta}{(\beta+1)\hat{q}(\beta\hat{q}+\hat{q}+1)}, \\ g^B &= \frac{(\beta+1)^2\hat{q}^2 - \beta}{(\beta+1)\hat{q}(\beta\hat{q}+\hat{q}+\beta)}, \\ q &= g^A g^B. \end{aligned} \tag{B72}$$

Notice that q does not depend on r in this case, so that all low-rank problems have the same MMSE. This is due to fact that the priors over the factors are i.i.d.

APPENDIX C: CONSEQUENCES OF THE MAIN RESULTS

In this appendix, we derive various consequences of Result 1 and Previous Result 4. In particular,

(i) In Appendix C 1, we derive the weak and strong recovery thresholds for the BSR model in the intensive width regime, as presented in Sec. III B.
(ii) In Appendix C 2, we derive the strong recovery threshold for the BSR model in the extensive-width regime, as presented in Result 5.
(iii) In Appendix C 3, we derive the large β limit of the BO error for the BSR model in the intensive width regime.
(iv) In Appendix C 4, we derive the large β limit of the BO error for the BSR model in the extensive-width regime.
(v) In Appendix C 5, we derive the large ρ limit of the BO error for the BSR model and show that it equals the error of optimally regularized ridge regression.

1. Weak and strong recovery thresholds in the intensive width BSR model without noise

The state equations are Eqs. (B31) and (B72), i.e.,

$$\begin{aligned} g^A &= \frac{(\beta+1)^2\hat{q}^2 - \beta}{(\beta+1)\hat{q}(\beta\hat{q}+\hat{q}+1)}, \\ g^B &= \frac{(\beta+1)^2\hat{q}^2 - \beta}{(\beta+1)\hat{q}(\beta\hat{q}+\hat{q}+\beta)}, \\ q &= g^A g^B, \\ \hat{q} &= \frac{\bar{\alpha}}{1-q}. \end{aligned} \tag{C1}$$

The strong recovery threshold $\bar{\alpha}_{\mathrm{BO}}$ is such that

$$\bar{\alpha}_{\mathrm{BO}} = \lim_{\hat{q}\to\infty}\hat{q}[1-q(\hat{q})] = 1, \tag{C2}$$

as can be seen by explicitly computing the limit. In the main text scaling $n = \alpha d L$, this translates to

$$\alpha_{\mathrm{BO}} = \frac{\rho(1+\beta)}{\beta} = 0, \tag{C3}$$

as for $r \ll d$ then $\rho \to 0$. This highlights the importance of the scaling $r(d+L)$ to study both intensive and extensive width in a common scaling.

The weak recovery threshold, i.e., the value $\bar{\alpha}_{\mathrm{weak}}$ at which $q = 0$, can be found by imposing $q = 0$, which gives either $g^A = 0$ or $g^B = 0$, implying $(\beta+1)^2\hat{q}^2 - \beta = 0$. Combined with the equation for $\hat{q}$, this gives

$$\bar{\alpha}_{\mathrm{weak}} = (1+\Delta)\hat{q} = (1+\Delta)\frac{\sqrt{\beta}}{1+\beta}. \tag{C4}$$

Notice that the weak threshold is nontrivial also in the noiseless case.

2. Strong recovery threshold in the extensive width BSR model without noise

We have MMSE $= 1 - q$ with the state equations (B31) and (B47), i.e.,

$$\delta = \frac{\beta}{\rho(1+\beta)}\frac{1-q}{\bar{\alpha}},$$
$$q = 1-\delta+\delta^2\int dx\hat{\mu}_{\underline{S}+\sqrt{\delta}\underline{Z}}(x) \times\left[\frac{(\beta-1)^2}{\beta^{3/2}x^2}+\frac{4\pi^2}{3\beta^{3/2}}\hat{\mu}_{\underline{S}+\sqrt{\delta}\underline{Z}}(x)^2\right]. \quad \text{(C5)}$$

Strong recovery happens for $\bar{\alpha}_{\mathrm{BO}}$ such that MMSE $= 0$, i.e., for $q \to 1$ and $\delta \to 0$. Rearranging the equations, we have

$$\bar{\alpha}_{\mathrm{BO}} = \lim_{\delta\to 0}\frac{\beta}{\rho(1+\beta)}\frac{1-q(\delta)}{\delta} = \frac{\beta}{\rho(1+\beta)} - \frac{\beta}{\rho(1+\beta)}\lim_{\delta\to 0}\delta\int dx\hat{\mu}_{\underline{S}+\sqrt{\delta}\underline{Z}}(x) \times\left[\frac{(\beta-1)^2}{\beta^{3/2}x^2}+\frac{4\pi^2}{3\beta^{3/2}}\hat{\mu}_{\underline{S}+\sqrt{\delta}\underline{Z}}(x)^2\right]. \quad \text{(C6)}$$

Thus, we need to study the second equation in the limit of $\delta \to 0$ and specifically look whether the integral terms develop divergencies at $\delta \to 0$.

Notice that, assuming that the last limit-integral term is finite as we verify later, $q = 1$, $\delta = 0$ is a solution of these equations for all values of $\bar{\alpha}$. This is kind of a trivial solution, and we expect that it is not the only solution for low enough values of $\bar{\alpha}$. Thus, to find the strong recovery threshold, which can be seen as the bifurcation point at which the nontrivial, low-$\bar{\alpha}$ solution of the equations merges with the trivial one, we assume that $q < 1$ and $\delta > 0$ and take the limit $q \to 1^-$ and $\delta \to 0^+$, effectively moving along the nontrivial solution toward the bifurcation point.

Intuitively, for $\delta \to 0$, the spectral density $\hat{\mu}_{\underline{S}+\sqrt{\delta}\underline{Z}}(x)$ will be composed either by a single bulk [if the width parameter does not constrain the rank of $\underline{S}$, i.e., if $\rho \geq \min(1,\beta)$] or by two bulks [if the width parameter constrains the rank of $\underline{S}$, i.e., $0 < \rho < \min(1,\beta)$], one gapped away from zero and the other close to zero. This is due to the fact that the spectrum of $\underline{S}$ is either composed by a bulk gapped away from zero or diverging at zero or by a bulk gapped away from zero and an additional delta accounting for the rank deficiency. Gaussian noise with vanishingly small variance alters this picture only perturbatively.

Then, if no rank deficiency is present [$0 < \rho < \min(1,\beta)$] and assuming that ungapped bulks are not problematic, the integrals will have no divergence and

$$\bar{\alpha}_c = \frac{\beta}{\rho(1+\beta)}. \quad \text{(C7)}$$

Otherwise, the first integral will develop a divergence due to the interplay of the noised delta peak of the spectrum of $\underline{S}$ and the $1/x^2$ factor in the integral, leading to a nontrivial $\bar{\alpha}_c$.

To follow this intuition, we perform the change of variable $x = \sqrt{\delta}z$ in the integrals, so that

$$\bar{\alpha}_{\mathrm{BO}} = \frac{\beta}{\rho(1+\beta)} - \frac{\beta}{\rho(1+\beta)}\lim_{\delta\to 0}\int dz\sqrt{\delta}\hat{\mu}_{\underline{S}+\sqrt{\delta}\underline{Z}}(\sqrt{\delta}z) \times\left[\frac{(\beta-1)^2}{\beta^{3/2}z^2}+\frac{4\pi^2}{3\beta^{3/2}}(\sqrt{\delta}\hat{\mu}_{\underline{S}+\sqrt{\delta}\underline{Z}}(\sqrt{\delta}z))^2\right]. \quad \text{(C8)}$$

The scaling $\sqrt{\delta}$ can be guessed, as it allows one to identify an expression which depends on δ only through the rescaled density $\sqrt{\delta}\hat{\mu}_{\underline{S}+\sqrt{\delta}\underline{Z}}(\sqrt{\delta}z)$, which we now study.

As shown in Ref. [71] (but we follow the notations and definitions of Ref. [55]), the Stieltjes transform of $\underline{Y} = \underline{S}+\sqrt{\delta}\underline{Z}$ satisfies

$$g_{\underline{Y}}(x) = z g_{\underline{Y}\underline{Y}^T}(x^2), \quad \text{(C9)}$$

where $g_{\underline{Y}\underline{Y}^T}(x^2)$ is a root of the polynomial $p(G) = \sum_{a=0}^{4} a_k(x^2,\delta)G^k$. We plug in the equation $x = \sqrt{\delta}z$, take the scaling *Ansatz* $G = H/\delta$, and expand everything at leading order for $\delta \to 0$, obtaining (after simplifying an overall factor of δ)

$$H(z^2) = \frac{\sqrt{\beta z^4 - 2\sqrt{\beta}z^2(\beta-2\rho+1)+(\beta-1)^2}+\sqrt{\beta}z^2-\beta+1}{2z^2} \quad \text{(C10)}$$

(the other solutions have either the wrong sign in front of the square root or no square root). Here, we define $\rho = \rho/\min(1,\beta)$ for simplicity. Now, recall that μ is the discontinuity at branch cuts over the real axes of H. Thus, only the square root term will contribute and only when its argument is negative. The roots of the argument of the square roots are

$$z_\pm = \frac{\pm 2\sqrt{(\rho-1)(\rho-\beta)}+\beta-2\rho+1}{\sqrt{\beta}}, \quad \text{(C11)}$$

giving the distribution

$$f(z) = \frac{1}{\pi}\lim_{\epsilon\to 0^+}\mathrm{Im}\left[(x-i\epsilon)H((x-i\epsilon)^2)\right] = \frac{\sqrt{\beta(z_+-z^2)(z^2-z_-)}}{2\pi z}. \quad \text{(C12)}$$

We can check using *Mathematica* that

$$
\begin{aligned}
2\int_{z_-}^{z_+} dz f(z) &= (1-\rho),\\
2\int_{z_-}^{z_+} dz f(z)^3 &= \frac{3\sqrt{\beta}(1-\rho)^2}{4\pi^2},\\
2\int_{z_-}^{z_+} dz f(z) z^{-2} &= \frac{\sqrt{\beta}(1-\rho)}{\beta-1}.
\end{aligned}
\tag{C13}
$$

The normalization is correct, as we are in a scaling limit where only the noisy version of the $(1-\rho)\delta_0$ contribution to the non-noisy measure. Notice that all this derivation holds only for $0<\rho<1$, as $0<z_-<z_+$ holds only in this case. For $\rho \geq 1$, the square root never develops a branch cut, so that the function $f(z)$ is identically zero.

Thus, we get for $\rho \geq 1$

$$
\bar{\alpha}_{\mathrm{BO}} = \frac{\beta}{\rho(1+\beta)}, \tag{C14}
$$

as expected, and for $0<\rho<1$

$$
\begin{aligned}
\bar{\alpha}_{\mathrm{BO}} &= \frac{\beta}{\rho(1+\beta)} - \frac{\beta}{\rho(1+\beta)}\frac{(1-\rho)(\beta-\rho)}{\beta}\\
&= \frac{\beta}{\rho(1+\beta)} - \frac{(\rho-1)(\rho-\beta)}{\rho(1+\beta)} = 1-\frac{\rho}{1+\beta}.
\end{aligned}
\tag{C15}
$$

For $\rho \to 0$, the threshold reduces to $\bar{\alpha}_c = 1$, as found explicitly in the low-width case [Eq. (C2)].

In the main text scaling $n = \alpha d L$, this translates to

$$
\alpha_{\mathrm{BO}} = \begin{cases} \frac{\rho}{\beta}(1+\beta-\rho) & 0<\rho<1,\\ 1 & \rho>1. \end{cases} \tag{C16}
$$

3. Large β limit for the intensive width BSR model

The state equations are Eqs. (B31) and (B72), i.e.,

$$
\begin{aligned}
g^A &= \frac{(\beta+1)^2\hat{q}^2-\beta}{(\beta+1)\hat{q}(\beta\hat{q}+\hat{q}+1)},\\
g^B &= \frac{(\beta+1)^2\hat{q}^2-\beta}{(\beta+1)\hat{q}(\beta\hat{q}+\hat{q}+\beta)},\\
q &= g^A g^B,\\
\hat{q} &= \frac{\bar{\alpha}}{1-q+\Delta}.
\end{aligned}
\tag{C17}
$$

For large β, assuming that $\hat{q}$ remains finite, we have

$$
\begin{aligned}
g^A &\sim \frac{\beta^2\hat{q}^2}{\beta\hat{q}(\beta\hat{q})} = 1,\\
g^B &\sim \frac{\beta^2\hat{q}^2}{\beta\hat{q}(\beta\hat{q}+\beta)} = \frac{\hat{q}}{1+\hat{q}},\\
q &\sim \frac{\hat{q}}{1+\hat{q}},\\
\hat{q} &= \frac{\bar{\alpha}}{1-q+\Delta}.
\end{aligned}
\tag{C18}
$$

The equations can be solved explicitly to

$$
q = \frac{1}{2}\left(1+\Delta+\bar{\alpha}-\sqrt{(\bar{\alpha}-1)^2+2(\bar{\alpha}+1)\Delta+\Delta^2}\right), \tag{C19}
$$

which reduces to

$$
q = \min(1,\bar{\alpha}), \tag{C20}
$$

in the noiseless case $\Delta = 0$.

4. Large β limit for the extensive-width BSR model

We have the state equations (B31) and (B47), i.e.,

$$
\begin{aligned}
\delta &= \frac{1-q+\Delta}{\alpha},\\
q &= 1-\delta+\delta^2\int dx \hat{\mu}_{\underline{S}+\sqrt{\delta}\underline{Z}}(x)\\
&\quad\times\left[\frac{(\beta-1)^2}{\beta^{3/2}x^2}+\frac{4\pi^2}{3\beta^{3/2}}\hat{\mu}_{\underline{S}+\sqrt{\delta}\underline{Z}}(x)^2\right],
\end{aligned}
\tag{C21}
$$

where we rescale α and $q = 1/\delta$ to match the scaling $\alpha = n/(dL)$ of the main text with respect to Appendix B.

We show below that

$$
f(w) = \lim_{\beta\to\infty} \sqrt[4]{\beta}\hat{\mu}_{\underline{S}+\sqrt{\delta}\underline{Z}}\left(\sqrt[4]{\beta}w\right) \tag{C22}
$$

for a finite and compactly supported function $f(w)$ fully independent on β. This implies that

$$
\begin{aligned}
q &= 1-\delta+\delta^2\int dx\hat{\mu}_{\underline{S}+\sqrt{\delta}\underline{Z}}(x)\left[\frac{(\beta-1)^2}{\beta^{3/2}x^2}+\frac{4\pi^2}{3\beta^{3/2}}\hat{\mu}_{\underline{S}+\sqrt{\delta}\underline{Z}}(x)^2\right]\\
&\sim 1-\delta+\delta^2\int dw\sqrt[4]{\beta}\hat{\mu}_{\underline{S}+\sqrt{\delta}\underline{Z}}\left(\sqrt[4]{\beta}w\right)\left[\frac{1}{w^2}+\frac{4\pi^2}{3\beta^2}\sqrt{\beta}\hat{\mu}_{\underline{S}+\sqrt{\delta}\underline{Z}}\left(\sqrt[4]{\beta}w\right)^2\right]\\
&\sim 1-\delta+\delta^2\int dw f(w)\left[\frac{1}{w^2}+\frac{4\pi^2}{3\beta^2}f(w)^2\right]\\
&\sim 1-\delta+\delta^2\int dw\frac{f(w)}{w^2},
\end{aligned}
\tag{C23}
$$

where we change variable $x=\sqrt[4]{\beta}w$ and use that

$$
\int dw f(w)^3<+\infty,
\tag{C24}
$$

as $f(w)$ is finite and compactly supported. We now need to compute f and show that it is indeed well behaved.

Recall that the Stieltjes transform of $\hat{\mu}_{\underline{S}+\sqrt{\delta}\underline{Z}}$ satisfies [55,71]

$$
g_{\underline{Y}}(z)=zG_{\underline{Y}^T\underline{Y}}(z^2)
\tag{C25}
$$

so that

$$
\sqrt[4]{\beta}\beta g_{\underline{Y}}(z)\left(\sqrt[4]{\beta}\beta w\right)=w\sqrt{\beta}G_{\underline{Y}^T\underline{Y}}\left(\sqrt{\beta}w^2\right).
\tag{C26}
$$

Now, recall that $G_{\underline{Y}^T\underline{Y}}(z)$ satisfies a quartic polynomial equation. By rescaling $z=\sqrt[4]{\beta}w$ and $G=H/\sqrt{\beta}$ and considering the leading order in β, one obtains the simplified quadratic equation

$$
H^2(\delta-w^2)+H[1+\rho w^2-(1+\delta)\rho]-\rho=0,
\tag{C27}
$$

which solves to

$$
wH(w)=w\frac{\sqrt{(-\delta\rho+\rho w^2-\rho+1)^2+4\rho(\delta-w^2)}+\delta\rho-\rho w^2+\rho-1}{2(\delta-w^2)},
\tag{C28}
$$

leading by the standard inversion formula to the associated measure

$$
f(w)=\frac{\rho w}{2\pi(w^2-\delta)}\sqrt{(b_+-w^2)(w^2-b_-)},
\tag{C29}
$$

supported on

$$
b_-<w^2<b_+
\tag{C30}
$$

and on the symmetric interval, where we define

$$
b_\pm=1+\delta+\frac{1}{\rho}\pm\frac{2}{\sqrt{\rho}}.
\tag{C31}
$$

This confirms that f is finite and compactly supported, as $t\le b_-<b_+$ for all δ and ρ. We can also verify that

$$\int dw f(w) = 2\frac{\rho}{2\pi}\int_{\sqrt{b_-}}^{\sqrt{b_+}} dww\frac{\sqrt{(b_+ - w^2)(w^2 - b_-)}}{w^2 - \delta}$$
$$= \frac{\rho}{2\pi}\int_{b_-}^{b_+} dt\frac{\sqrt{(b_+ - t)(t - b_-)}}{t - \delta}$$
$$= \frac{\rho}{2\pi}\begin{cases} 2\pi & \text{if } \rho < 1, \\ \frac{2\pi}{\rho} & \text{otherwise,} \end{cases}$$
$$= \min(\rho, 1), \tag{C32}$$

where we use $t = w^2$. This is the correct normalization for the bulk of the distribution, excluding the rank-deficiency-induced spike in zero of mass $\max(0, 1-\rho)$.

Now, we need to compute

$$\int dw\frac{f(w)}{w^2} = 2\frac{\rho}{2\pi}\int_{\sqrt{b_-}}^{\sqrt{b_+}} dww\frac{\sqrt{(b_+ - w^2)(w^2 - b_-)}}{w^2(w^2 - \delta)}$$
$$= \frac{\rho}{2\pi}\int_{b_-}^{b_+} dt\frac{\sqrt{(b_+ - t)(t - b_-)}}{t(t - \delta)}$$
$$= \frac{\sqrt{1 + \rho(\rho - 2 + 2\delta + (2+\delta)\delta\rho)} - \delta\rho + \rho - 1}{2\delta}, \tag{C33}$$

where we use $t = w^2$. Thus, we obtain the equations

$$\delta = \frac{1 - q + \Delta}{\alpha},$$
$$q = 1 - \delta + \delta\frac{\sqrt{1 + \rho(\rho - 2 + 2\delta + (2+\delta)\delta\rho)} - \delta\rho + \rho - 1}{2}, \tag{C34}$$

to be solved for q.

5. Large ρ limit for the extensive-width BSR model

We have the state equations (B31) and (B47), i.e.,

$$\delta = \frac{1 - q + \Delta}{\alpha},$$
$$q = 1 - \delta + \delta^2\int dx\hat{\mu}_{\underline{S} + \sqrt{\delta}\underline{Z}}(x) \times \left[\frac{(\beta - 1)^2}{\beta^{3/2}x^2} + \frac{4\pi^2}{3\beta^{3/2}}\hat{\mu}_{\underline{S} + \sqrt{\delta}\underline{Z}}(x)^2\right], \tag{C35}$$

where we rescale α and $q = 1/\delta$ to match the scaling $\alpha = n/(dL)$ of the main text with respect to Appendix B.

We want to compute the $\rho \to \infty$ limit of the equations. Recall that the Stieltjes transform of $\hat{\mu}_{\underline{S} + \sqrt{\delta}\underline{Z}}$ satisfies [55,71]

$$g_{\underline{Y}}(z) = zG_{\underline{Y}^T\underline{Y}}(z^2) \tag{C36}$$

and that $G_{\underline{Y}^T\underline{Y}}(z)$ satisfies a quartic polynomial equation. One obtains the simplified quadratic equation in the large ρ limit:

$$G^2\frac{z^2(1+\delta)}{\sqrt{\beta}} + G\left(z^2 - \frac{(1+\delta)(\beta - 1)}{\sqrt{\beta}}\right) - 1 = 0, \tag{C37}$$

which solves to

$$zG(z^2) = z\frac{\sqrt{\beta}\left(\sqrt{-\frac{2(\beta+1)(\delta+1)z^2}{\sqrt{\beta}} + \frac{(\beta-1)^2(\delta+1)^2}{\beta} + z^4} + z^2\right) - \beta(\delta + 1) + \delta + 1}{2(\delta + 1)z^2}, \tag{C38}$$

leading by the standard inversion formula to the associated measure

$$f(z) = \frac{\sqrt{\beta}}{1 + \delta}\frac{\sqrt{(b_+ - z^2)(z^2 - b_-)}}{2\pi z}, \tag{C39}$$

supported on

$$b_- < z^2 < b_+ \tag{C40}$$

and on the symmetric interval, where we define

$$b_\pm = (\sqrt{\beta} \pm 1)^2\frac{1 + \delta}{\sqrt{\beta}}. \tag{C41}$$

We can verify that

$$2\int_{b_-}^{b_+} dzf(z) = 1. \tag{C42}$$

This is the correct normalization for the bulk of the distribution. We also have

$$2\int_{b_-}^{b_+} dz f(z)\frac{1}{z^2} = \frac{\sqrt{\beta}}{(\beta-1)(\delta+1)} \tag{C43}$$

and

$$2\int_{b_-}^{b_+} dz f(z)^3 = \frac{3\sqrt{\beta}}{4\pi^2(1+\delta)}, \tag{C44}$$

from which one gets the equations

$$\delta = \frac{1-q+\Delta}{\alpha}, \qquad q = \frac{1}{1+\delta}, \tag{C45}$$

whose solution gives

$$\frac{1+\alpha+\Delta-\sqrt{(\alpha+\Delta+1)^2-4\alpha}}{2}. \tag{C46}$$

This coincides with the overlap achieved by optimally regularized ridge regression Previous Result 6.

[1] Y. LeCun, Y. Bengio, and G. Hinton, *Deep learning*, Nature (London) **521**, 436 (2015).

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, *ImageNet: A large-scale hierarchical image database*, in *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, New York, 2009), pp. 248–255.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, *Imagenet classification with deep convolutional neural networks*, Adv. Neural Inf. Process. Syst. **25** (2012).

[4] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, *Mastering the game of go with deep neural networks and tree search*, Nature (London) **529**, 484 (2016).

[5] L. Zdeborová, *Understanding deep learning is also a job for physicists*, Nat. Phys. **16**, 602 (2020).

[6] J. J. Hopfield, *Neural networks and physical systems with emergent collective computational abilities*, Proc. Natl. Acad. Sci. U.S.A. **79**, 2554 (1982).

[7] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, *A learning algorithm for Boltzmann machines*, Cogn. Sci. **9**, 147 (1985).

[8] E. Gardner and B. Derrida, *Optimal storage properties of neural network models*, J. Phys. A **21**, 271 (1988).

[9] E. Gardner and B. Derrida, *Three unfinished works on the optimal storage capacity of networks*, J. Phys. A **22**, 1983 (1989).

[10] H. S. Seung, H. Sompolinsky, and N. Tishby, *Statistical mechanics of learning from examples*, Phys. Rev. A **45**, 6056 (1992).

[11] A. M. Saxe, J. L. McClelland, and S. Ganguli, *Exact solutions to the nonlinear dynamics of learning in deep linear neural networks*, *International Conference on Learning Representations* (2014).

[12] C. Baldassi, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina, *Subdominant dense clusters allow for simple learning and high computational performance in neural networks with discrete synapses*, Phys. Rev. Lett. **115**, 128101 (2015).

[13] C. Baldassi, C. Borgs, J. T. Chayes, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina, *Unreasonable effectiveness of learning neural networks: From accessible states and robust ensembles to basic algorithmic schemes*, Proc. Natl. Acad. Sci. U.S.A. **113**, E7655 (2016).

[14] J. Barbier, F. Krzakala, N. Macris, L. Miolane, and L. Zdeborová, *Optimal errors and phase transitions in high-dimensional generalized linear models*, Proc. Natl. Acad. Sci. U.S.A. **116**, 5451 (2019).

[15] S. Goldt, M. Mézard, F. Krzakala, and L. Zdeborová, *Modeling the influence of data structure on learning in neural networks: The hidden manifold model*, Phys. Rev. X **10**, 041044 (2020).

[16] M. S. Advani and A. M. Saxe, *High-dimensional dynamics of generalization error in neural networks*, Neural Netw. **132**, 428 (2020).

[17] B. Loureiro, C. Gerbelot, H. Cui, S. Goldt, F. Krzakala, M. Mezard, and L. Zdeborová, *Learning curves of generic features maps for realistic datasets with a teacher-student model*, Adv. Neural Inf. Process. Syst. **34**, 18137 (2021).

[18] B. Sorscher, R. Geirhos, S. Shekhar, S. Ganguli, and A. Morcos, *Beyond neural scaling laws: Beating power law scaling via data pruning*, Adv. Neural Inf. Process. Syst. **35**, 19523 (2022).

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, *Attention is all you need*, Adv. Neural Inf. Process. Syst. **30** (2017).

[20] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, *Language models are unsupervised multitask learners*, OpenAI blog **1**, 9 (2019).

[21] B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal *et al.*, *Language models are few-shot learners*, arXiv:2005.14165.

[22] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, *GPT-4 technical report*, arXiv:2303.08774.

[23] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg *et al.*, *Sparks of artificial general intelligence: Early experiments with GPT-4*, arXiv:2303.12712.

[24] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus, *Emergent abilities of large language models*, Trans. Mach. Learn. Res. (2022).

[25] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, *Scaling laws for neural language models*, arXiv:2001.08361.

[26] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, *Language models are few-shot learners*, Adv. Neural Inf. Process. Syst. **33**, 1877 (2020).

[27] G. Cybenko, *Approximation by superpositions of a sigmoidal function*, Math. Control Signals Syst. **2**, 303 (1989).

[28] K. Hornik, M. Stinchcombe, and H. White, *Multilayer feedforward networks are universal approximators*, Neural Netw. **2**, 359 (1989).

[29] A. Raventós, M. Paul, F. Chen, and S. Ganguli, *Pretraining task diversity and the emergence of non-Bayesian in-context learning for regression*, Adv. Neural Inf. Process. Syst. **36**, 14228 (2023).

[30] F. Cagnetta and M. Wyart, *Towards a theory of how the structure of language is acquired by deep neural networks*, arXiv:2406.00048.

[31] F. Behrens, L. Biggio, and L. Zdeborová, *Understanding counting in small transformers: The interplay between attention and feed-forward layers*, arXiv:2407.11542.

[32] B. Geshkovski, C. Letrouit, Y. Polyanskiy, and P. Rigollet, *A mathematical perspective on transformers*, arXiv:2312.10794.

[33] A. Cowsik, T. Nebabu, X.-L. Qi, and S. Ganguli, *Geometric dynamics of signal propagation predict trainability of transformers*, arXiv:2403.02579.

[34] R. Rende, F. Gerace, A. Laio, and S. Goldt, *Mapping of attention mechanisms to a generalized Potts model*, Phys. Rev. Res. **6**, 023057 (2024).

[35] H. Cui, F. Behrens, F. Krzakala, and L. Zdeborová, *A phase transition between positional and semantic learning in a solvable model of dot-product attention*, Adv. Neural Inf. Process. Syst. **37**, 36342 (2024).

[36] Y. M. Lu, M. I. Letey, J. A. Zavatone-Veth, A. Maiti, and C. Pehlevan, *Asymptotic theory of in-context learning by linear attention*, arXiv:2405.11751.

[37] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, *The Llama 3 herd of models*, arXiv:2407.21783.

[38] F. Mignacco, P. Urbani, and L. Zdeborová, *Stochasticity helps to navigate rough landscapes: Comparing gradient-descent-based algorithms in the phase retrieval problem*, Mach. Learn. **2**, 035029 (2021).

[39] S. Sarao Mannelli, G. Biroli, C. Cammarota, F. Krzakala, P. Urbani, and L. Zdeborová, *Complex dynamics in simple neural networks: Understanding gradient flow in phase retrieval*, Adv. Neural Inf. Process. Syst. **33**, 3265 (2020).

[40] M. Belkin, D. Hsu, S. Ma, and S. Mandal, *Reconciling modern machine-learning practice and the classical bias–variance trade-off*, Proc. Natl. Acad. Sci. U.S.A. **116**, 15849 (2019).

[41] F. Gerace, B. Loureiro, F. Krzakala, M. Mézard, and L. Zdeborová, *Generalisation error in learning with random features and the hidden manifold model*, in *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research Vol. 119 (PMLR, Cambridge, MA, 2020), pp. 3452–3462.

[42] C. Schülke, P. Schniter, and L. Zdeborová, *Phase diagram of matrix compressed sensing*, Phys. Rev. E **94**, 062136 (2016).

[43] https://github.com/SPOC-group/bilinear-sequence-regression.

[44] I. Tolstikhin *et al.*, *Mlp-mixer: An all-mlp architecture for vision*, Adv. Neural Inf. Process. Syst. **34**, 24261 (2021).

[45] B. Recht, M. Fazel, and P. A. Parrilo, *Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization*, SIAM Rev. **52**, 471 (2010).

[46] D. L. Donoho, M. Gavish, and A. Montanari, *The phase transition of matrix recovery from Gaussian measurements matches the minimax MSE of matrix denoising*, Proc. Natl. Acad. Sci. U.S.A. **110**, 8405 (2013).

[47] C. Giraud, *Low rank multivariate regression*, Electron. J. Stat. **5**, 775 (2011).

[48] P. D. Hoff, *Multilinear tensor regression for longitudinal relational data*, Ann. Appl. Stat. **9**, 1169 (2015).

[49] E. Y. Chen and J. Fan, *Statistical inference for high-dimensional matrix-variate factor models*, J. Am. Stat. Assoc. **118**, 1038 (2023).

[50] S. Gunasekar, B. E. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro, *Implicit regularization in matrix factorization*, Adv. Neural Inf. Process. Syst. **30** (2017).

[51] Z. Li, Y. Luo, and K. Lyu, *Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning*, in *Proceedings of the International Conference on Learning Representations* (2021).

[52] H. C. Schmidt, *Statistical physics of sparse and dense models in optimization and inference*, Ph.D. thesis, Université Paris Saclay (COmUE), 2018.

[53] A. Maillard, F. Krzakala, M. Mézard, and L. Zdeborová, *Perturbative construction of mean-field equations in extensive-rank matrix factorization and denoising*, J. Stat. Mech. (2022) 083301.

[54] J. Barbier and N. Macris, *Statistical limits of dictionary learning: Random matrix theory and the spectral replica method*, Phys. Rev. E **106**, 024136 (2022).

[55] E. Troiani, V. Erba, F. Krzakala, A. Maillard, and L. Zdeborová, *Optimal denoising of rotationally invariant rectangular matrices*, in *Mathematical and Scientific Machine Learning*, Proceedings of Machine Learning Research Vol. 190, edited by B. Dong, Q. Li, L. Wang, and Z.-Q. J. Xu (PMLR, Cambridge, MA, 2022), pp. 97–112.

[56] G. Semerjian, *Matrix denoising: Bayes-optimal estimators via low-degree polynomials*, J. Stat. Phys. **191**, 139 (2024).

[57] F. Camilli and M. Mézard, *Matrix factorization with neural networks*, Phys. Rev. E **107**, 064308 (2023).

[58] F. Pourkamali and N. Macris, *Rectangular rotational invariant estimator for general additive noise matrices*, in *Proceedings of the 2023 IEEE International Symposium on Information Theory (ISIT)* (IEEE, New York, 2023), pp. 2081–2086.

[59] Y. V. Fyodorov, *A spin glass model for reconstructing nonlinearly encrypted signals corrupted by noise*, J. Stat. Phys. **175**, 789 (2019).

[60] P. J. Kamali and P. Urbani, *Dynamical mean field theory for models of confluent tissues and beyond*, SciPost Phys. **15**, 219 (2023).

[61] A. Montanari and E. Subag, *Solving overparametrized systems of random equations: I. Model and algorithms for approximate solutions*, arXiv:2306.13326.

[62] P. J. Kamali and P. Urbani, *Stochastic gradient descent outperforms gradient descent in recovering a high-dimensional signal in a glassy energy landscape*, arXiv:2309.04788.

[63] A. Maillard, E. Troiani, S. Martin, F. Krzakala, and L. Zdeborová, *Bayes-optimal learning of an extensive-width neural network from quadratically many samples*, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2024), Vol. 37, pp. 82085–82132, https://proceedings.neurips.cc/paper_files/paper/2024/file/

953e742190ca02fc8f9f710052f2fead-Paper-Conference.pdf.
[64] S. Bhojanapalli, B. Neyshabur, and N. Srebro, *Global optimality of local search for low rank matrix recovery*, Adv. Neural Inf. Process. Syst. **29** (2016).
[65] L. Ding, D. Drusvyatskiy, M. Fazel, and Z. Harchaoui, *Flat minima generalize for low-rank matrix recovery*, Inf. Inference **13**, iaae009 (2024).
[66] K. Okajima and T. Takahashi, *Asymptotic dynamics of alternating minimization for bilinear regression*, arXiv:2402.04751.
[67] H. Hu and Y. M. Lu, *Universality laws for high-dimensional learning with random features*, IEEE Trans. Inf. Theory **69**, 1932 (2022).
[68] Z. Wang, E. Nichani, and J. D. Lee, *Learning hierarchical polynomials with three-layer neural networks*, in *Proceedings of the Twelfth International Conference on Learning Representations* (2024).
[69] T. M. Cover and J. A. Thomas, *Information theory and statistics*, Elem. Inf. Theor. **1**, 279 (1991).
[70] L. Zdeborová and F. Krzakala, *Statistical physics of inference: Thresholds and algorithms*, Adv. Phys. **65**, 453 (2016).
[71] J. Pennington and P. Worah, *Nonlinear random matrix theory for deep learning*, Adv. Neural Inf. Process. Syst. **30** (2017).
[72] P. Biane, *On the free convolution with a semi-circular distribution*, Indiana Univ. Math. J. **46**, 705 (1997).
[73] S. Rangan, *Generalized approximate message passing for estimation with random linear mixing*, in *Proceedings of the 2011 IEEE International Symposium on Information Theory Proceedings* (IEEE, New York, 2011), pp. 2168–2172.
[74] D. L. Donoho, A. Maleki, and A. Montanari, *Message-passing algorithms for compressed sensing*, Proc. Natl. Acad. Sci. U.S.A. **106**, 18914 (2009).
[75] R. Berthier, A. Montanari, and P.-M. Nguyen, *State evolution for approximate message passing with non-separable functions*, Inf. Inference **9**, 33 (2020).
[76] C. Gerbelot and R. Berthier, *Graph-based approximate message passing iterations*, Inf. Inference **12**, 2562 (2023).
[77] E. Romanov and M. Gavish, *Near-optimal matrix recovery from random linear measurements*, Proc. Natl. Acad. Sci. U.S.A. **115**, 7200 (2018).
[78] J. Barbier, J. Ko, and A. A. Rahman, *Information-theoretic limits for sublinear-rank symmetric matrix factorization*, in *Proceedings of the International Zurich Seminar on Information and Communication (IZS 2024)* (ETH Zürich, 2024), p. 16.
[79] J. Barbier, J. Ko, and A. A. Rahman, *A multiscale cavity method for sublinear-rank symmetric matrix factorization*, arXiv:2403.07189.
[80] F. Pourkamali, J. Barbier, and N. Macris, *Matrix inference in growing rank regimes*, IEEE Trans. Inf. Theory (to be published).
[81] U. Helmke and J. B. Moore, *Optimization and Dynamical Systems* (Springer Science & Business Media, New York, 2012).
[82] D. Donoho and M. Gavish, *Minimax risk of matrix denoising by singular value thresholding*, Ann. Stat. **42**, 2413 (2014).
[83] F. Krzakala, M. Mézard, F. Sausset, Y. F. Sun, and L. Zdeborová, *Statistical-physics-based reconstruction in compressed sensing*, Phys. Rev. X **2**, 021005 (2012).
[84] B. Neyshabur, R. Tomioka, and N. Srebro, *In search of the real inductive bias: On the role of implicit regularization in deep learning*, arXiv:1412.6614.
[85] S. Arora, N. Cohen, W. Hu, and Y. Luo, *Implicit regularization in deep matrix factorization*, Adv. Neural Inf. Process. Syst. **32** (2019).
[86] F. Mignacco, F. Krzakala, P. Urbani, and L. Zdeborová, *Dynamical mean-field theory for stochastic gradient descent in Gaussian mixture classification*, Adv. Neural Inf. Process. Syst. **33**, 9540 (2020).
[87] S. Diamond and S. Boyd, *CVXPY: A Python-embedded modeling language for convex optimization*, J. Mach. Learn. Res. **17**, 1 (2016).
[88] S. D. Akshay Agrawal, Robin Verschueren, and S. Boyd, *A rewriting system for convex optimization problems*, J. Control Decis. **5**, 42 (2018).