Compression of Structured Data with Autoencoders: Provable Benefit of Nonlinearities and Depth

Kevin Kögler [*1] Alexander Shevchenko [*1] Hamed Hassani [2] Marco Mondelli [1]

Abstract

Autoencoders are a prominent model in many empirical branches of machine learning and lossy data compression. However, basic theoretical questions remain unanswered even in a shallow two-layer setting. In particular, to what degree does a shallow autoencoder capture the structure of the underlying data distribution? For the prototypical case of the 1-bit compression of *sparse* Gaussian data, we prove that gradient descent converges to a solution that completely disregards the sparse structure of the input. Namely, the performance of the algorithm is the same as if it was compressing a Gaussian source – with no sparsity. For general data distributions, we give evidence of a phase transition phenomenon in the shape of the gradient descent minimizer, as a function of the data sparsity: below the critical sparsity level, the minimizer is a rotation taken uniformly at random (just like in the compression of non-sparse data); above the critical sparsity, the minimizer is the identity (up to a permutation). Finally, by exploiting a connection with approximate message passing algorithms, we show how to improve upon Gaussian performance for the compression of sparse data: adding a denoising function to a shallow architecture already reduces the loss provably, and a suitable multi-layer decoder leads to a further improvement. We validate our findings on image datasets, such as CIFAR-10 and MNIST.

1. Introduction

Autoencoders have achieved remarkable performance in many machine learning areas, such as generative modeling (Kingma & Welling, 2014), inverse problems (Peng et al., 2020) and data compression (Ballé et al., 2017; Theis et al., 2017; Agustsson et al., 2017). Motivated by this practical success, an active area of research is aimed at theoretically analyzing the performance of autoencoders to understand the quality and dynamics of representation learning when these architectures are trained with gradient methods.

Formally, consider the encoding of $\boldsymbol{x} \in \mathbb{R}^d$ given by

$$\boldsymbol{z} = \sigma(\boldsymbol{B}\boldsymbol{x}), \quad \boldsymbol{B} \in \mathbb{R}^{n \times d}, \quad \boldsymbol{z} \in \mathbb{R}^n, \tag{1}$$

where the non-linear activation $\sigma(\cdot)$ is applied component-wise. The ratio $r = n/d$ is referred to as the compression rate, which in the case of 1-bit compressed sensing corresponds to the number of bits per input dimension. For a shallow (two-layer) autoencoder, the decoding consists of a single linear transformation $\boldsymbol{A} \in \mathbb{R}^{d \times n}$:

$$\hat{\boldsymbol{x}}_{\boldsymbol{\Theta}}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{z} = \boldsymbol{A}\sigma(\boldsymbol{B}\boldsymbol{x}). \tag{2}$$

The optimal set of parameters $\boldsymbol{\Theta} = \{\boldsymbol{A}, \boldsymbol{B}\}$ minimizes the mean-squared error (MSE)

$$\mathcal{R}(\boldsymbol{\Theta}) := d^{-1}\mathbb{E}\left[\|\boldsymbol{x} - \hat{\boldsymbol{x}}_{\boldsymbol{\Theta}}(\boldsymbol{x})\|_2^2\right], \tag{3}$$

where the expectation is taken over the data distribution $\boldsymbol{x}$. The model described in (2) is a natural extension of *linear* autoencoders ($\sigma(x) = \alpha \cdot x$ for some $\alpha \neq 0$), which were thoroughly studied over the past years (Kunin et al., 2019; Gidel et al., 2019; Bao et al., 2020). In an effort to go beyond the linear setting, a number of recent works have considered the *non-linear* model (2). Specifically, Refinetti & Goldt (2022); Nguyen (2021) study the training dynamics under specific scaling regimes of the input dimension d and the number of neurons n, which lead to either vanishing or diverging compression rates. Shevchenko et al. (2023) focus on the proportional regime in which d and n grow at the same speed, but their analysis relies heavily on Gaussian data assumptions. In contrast with Gaussian data that lacks any particular structure, real data often exhibits rich structural properties. For instance, images are inherently

[*]Equal contribution [1]ISTA, Klosterneuburg, Austria [2]Department of Electrical and Systems Engineering, University of Pennsylvania, USA. Correspondence to: Kevin Kögler <kevin.koegler@ist.ac.at>, Alexander Shevchenko <alex.shevchenko@ist.ac.at>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

sparse (in, e.g., Wavelet or FFT domain), and this property has been exploited by various compression schemes such as jpeg. In this view, it is paramount to go beyond the analysis of unstructured Gaussian data and address the following fundamental questions:

Does gradient descent training of the two-layer autoencoder (2) *capture the structure in the data? How does increasing the expressivity of the decoder impact the performance?*

To address these questions, we consider the compression of structured data via the non-linear autoencoder (1) with $\sigma \equiv \mathrm{sign}$ (1-bit compressed sensing, (Boufounos & Baraniuk, 2008)) and show how the data structure is captured by the architecture of the decoder. Let us explain the choice of $\sigma \equiv \mathrm{sign}$. Apart from the connection to classical information and coding theory (Cover & Thomas, 2006), its scale invariance prevents the model from entering the *linear regime*. Namely, if $\sigma(\cdot)$ has a well-defined non-vanishing derivative at zero, by picking an encoding matrix $\boldsymbol{B}$ s.t. $\|\boldsymbol{B}\|_{op} \ll 1$, one can linearize the model, i.e., $\hat{\boldsymbol{x}}_{\boldsymbol{\Theta}}(\boldsymbol{x}) \approx \sigma'(0) \cdot \boldsymbol{A}\boldsymbol{B}\boldsymbol{x}$, which results in PCA-like behaviour (Refinetti & Goldt, 2022). Thus, sign is a natural candidate to tackle the non-linear setting of interest in applications and, in fact, hard-thresholding activations are common in large-scale models (Van Den Oord et al., 2017).

Our main contributions can be summarized as follows:

- Theorem 4.1 proves that the *linear decoder* in (2) may be *unable to exploit the sparsity* in the data: when $\boldsymbol{x}$ has a Bernoulli-Gaussian (or "sparse Gaussian") distribution, both the gradient descent solution and the MSE coincide with those obtained for the compression of purely Gaussian data (with no sparsity).
- Going beyond Gaussian data, we give evidence of the emergence of a *phase transition* in the structure of the optimal matrices $\boldsymbol{A}, \boldsymbol{B}$ in (2), as the sparsity level $p \in (0, 1)$ varies: Proposition 4.2 locates the critical value of p such that the minimizer stops being a random rotation (as for purely Gaussian data), and it becomes the identity (up a permutation); numerical simulations for gradient descent corroborate this phenomenology and display a "staircase" behavior of the loss function.
- While for the compression of sparse Gaussian data the linear decoder in (2) does not capture the sparsity, we show in Section 5 that increasing the expressivity of the decoder improves upon Gaussian performance. First, we post-process the output of (2), i.e., we consider

$$\hat{\boldsymbol{x}}_{\boldsymbol{\Theta}}(\boldsymbol{x}) = f(\boldsymbol{A}\boldsymbol{z}) = f(\boldsymbol{A}\mathrm{sign}(\boldsymbol{B}\boldsymbol{x})), \qquad (4)$$

  where f is applied component-wise, and we prove that a suitable f leads to a smaller MSE. In other words, adding a *nonlinearity* to the linear decoder in (2) provably helps. Finally, we further improve the performance by increasing the *depth* and using a multi-layer decoder. Our analysis leverages a connection between multi-layer autoencoders and the iterates of the RI-GAMP algorithm proposed by Venkataramanan et al. (2022), which may be of independent interest.

Experiments on syntethic data confirm our findings, and similar phenomena are displayed when running gradient descent to compress CIFAR-10/MNIST images. Taken together, our results show that, for the compression of structured data, a more expressive decoding architecture provably improves performance. This is in sharp contrast with the compression of unstructured, Gaussian data where, as discussed in Section 6 of (Shevchenko et al., 2023), multiple decoding layers do not help.

2. Related work

Theoretical results for autoencoders. The practical success of autoencoders has spurred a flurry of theoretical research, started with the analysis of linear autoencoders: Kunin et al. (2019) indicate a PCA-like behaviour of the minimizers of the L_2-regularized loss; Bao et al. (2020) provide evidence that the convergence to the minimizer is slow due to ill-conditioning, which worsens as the dimension of the latent space increases; Oftadeh et al. (2020) study the geometry of the loss landscape; Gidel et al. (2019) quantify the time-steps of the training dynamics at which deep linear networks recover features of increasing complexity. More recently, the focus has shifted towards *non-linear* autoencoders. Refinetti & Goldt (2022) characterize the training dynamics via a system of ODEs when the compression rate r is vanishing. Nguyen (2021) takes a mean-field view that requires a polynomial growth of the number of neurons n in the input dimension d, which results in a diverging compression rate. Cui & Zdeborová (2023) use tools from statistical physics to predict the MSE of denoising a Gaussian mixture via a two-layer autoencoder with a skip connection. Shevchenko et al. (2023) consider the compression of Gaussian data with a two-layer autoencoder when the compression rate r is fixed and show that gradient descent methods achieve a minimizer of the MSE.

Incremental learning and staircases in the training dynamics. Phenomena similar to the staircase behavior of the loss function that we exhibit in Figure 2 have drawn significant attention. For parity learning, the line of works (Abbe et al., 2021; 2022; 2023a) shows that parities are recovered in a sequential fashion with increasing complexity. A similar behaviour is observed in transformers with diagonal weight matrices at small initialization (Abbe et al., 2023b): gradient descent progressively learns a solution of increasing rank. For a single index model, Berthier et al. (2023) show a separation of time-scales at which the training dynamics follows an alternating pattern of plateaus and rapid decreases in the loss. Evidence of incremental learning in deep linear networks is provided by Berthier (2023); Pesme & Flammarion (2023); Simon et al. (2023); Jacot et al. (2021); Milanesi et al. (2021). The recent work by Székely et al. (2023) shows that the cumulants of the data distribution are learnt sequentially, revealing a sample complexity gap between neural networks and random features.

Approximate Message Passing (AMP). AMP refers to a family of iterative algorithms developed for a variety of statistical inference problems (Feng et al., 2022). Such problems include the recovery of a signal $\boldsymbol{x}$ from observations $\boldsymbol{z}$ of the form in (1), namely, a Generalized Linear Model (McCullagh & Nelder, 1989), when the encoder matrix $\boldsymbol{B}$ is Gaussian (Rangan, 2011; Mondelli & Venkataramanan, 2022) or rotationally-invariant (Rangan et al., 2019; Schniter et al., 2016; Ma & Ping, 2017; Takeuchi, 2019). Of particular interest for our work is the RI-GAMP algorithm by Venkataramanan et al. (2022). In fact, RI-GAMP enjoys a computational graph structure that can be mapped to a suitable neural network, and it approaches the information-theoretically optimal MSE. The optimal MSE was computed via the replica method by Takeda et al. (2006); Tulino et al. (2013), and these predictions were rigorously confirmed for the high-temperature regime by Li et al. (2023).

3. Preliminaries

Notation. We use plain symbols a, b for scalars, bold symbols $\boldsymbol{a}, \boldsymbol{b}$ for vectors, and capitalized bold symbols $\boldsymbol{A}, \boldsymbol{B}$ for matrices. Given a vector $\boldsymbol{a}$, its ℓ_2-norm is $\|\boldsymbol{a}\|_2$. Given a matrix $\boldsymbol{A}$, its operator norm is $\|\boldsymbol{A}\|_{op}$. We denote a unidimensional Gaussian distribution with mean μ and variance σ^2 by $\mathcal{N}(\mu, \sigma^2)$. We use the shorthand $\hat{\boldsymbol{x}}$ for $\hat{\boldsymbol{x}}_{\boldsymbol{\Theta}}$. Unless specified otherwise, function are applied *component-wise* to vector/matrix-valued inputs. We denote by $C, c > 0$ universal constants, which are independent of n, d.

Data distribution and MSE. For $p \in (0, 1]$, a sparse Gaussian distribution $\mathrm{SG}_1(p)$ is equal to $\mathcal{N}(0, 1/p)$ with probability p and is 0 otherwise. The scaling of the variance of the Gaussian component ensures a unit second moment for all p. We use the notation $\boldsymbol{x} \sim \mathrm{SG}_d(p)$ to denote a vector with i.i.d. components distributed according to $\mathrm{SG}_1(p)$. Decreasing p makes $\boldsymbol{x} \sim \mathrm{SG}_d(p)$ more sparse: for $p = 1$ one recovers the isotropic Gaussian, i.e., $\mathrm{SG}_d(1) \equiv \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, while $p = 0$ implies that $\boldsymbol{x} = \boldsymbol{0}$.

Shevchenko et al. (2023) consider Gaussian data $\boldsymbol{x} \sim \mathrm{SG}_d(1)$ and the two-layer autoencoder with linear decoder in (2). Their analysis shows that, for a compression rate $r \leq 1$, the MSE obtained by minimizing (3) over $\boldsymbol{\Theta} = \{\boldsymbol{A}, \boldsymbol{B}\}$ is given by

$$\mathcal{R}_{\mathrm{Gauss}} := 1 - \frac{2}{\pi} \cdot r. \tag{5}$$

The set of minimizers $(\boldsymbol{A}, \boldsymbol{B})$ has a weight-tied orthogonal structure, i.e., $\boldsymbol{B}\boldsymbol{B}^\top = \boldsymbol{I}$ and $\boldsymbol{A} \propto \boldsymbol{B}^\top$, and gradient-based optimization schemes reach a global minimum.

4. Limitations of a linear decoding layer

Our main technical result is that a two-layer autoencoder with a single linear decoding layer does not capture the sparse structure of the data. Specifically, we consider the autoencoder in (2) with Gaussian data $\boldsymbol{x} \sim \mathrm{SG}_d(p)$ trained via gradient descent. We show that, when n, d are both large (holding the compression rate $r = n/d$ fixed), the trajectory of the algorithm is the same as that obtained from the compression of non-sparse data, i.e., $\boldsymbol{x} \sim \mathrm{SG}_{\mathrm{d}}(1) \equiv \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. As a consequence, the minimizer has a weight-tied orthogonal structure ($\boldsymbol{B}\boldsymbol{B}^\top = \boldsymbol{I}$, $\boldsymbol{A} \propto \boldsymbol{B}^\top$), and the MSE at convergence is given by $\mathcal{R}_{\mathrm{Gauss}}$ as defined in (5).

We now go into the details. Since the optimization objective is convex in $\boldsymbol{A}$, we consider the following alternating minimization version of Riemannian gradient descent:

$$\begin{aligned} \boldsymbol{A}(t+1) &= \arg\min_{\boldsymbol{A}} \mathcal{R}(\boldsymbol{A}, \boldsymbol{B}(t)), \\ \boldsymbol{B}(t+1) &:= \mathrm{proj}\left(\boldsymbol{B}(t) - \eta\left(\nabla_{\boldsymbol{B}(t)} + \boldsymbol{G}(t)\right)\right). \end{aligned} \tag{6}$$

In fact, due to the convexity in $\boldsymbol{A}$ of the MSE $\mathcal{R}(\cdot, \cdot)$ in (3), we can compute in closed form $\arg\min_{\boldsymbol{A}} \mathcal{R}(\boldsymbol{A}, \boldsymbol{B}(t))$. Here, Riemannian refers to the space of matrices with unit-norm rows, $\nabla_{\boldsymbol{B}(t)}$ is a shorthand for the gradient $\nabla_{\boldsymbol{B}(t)}\mathcal{R}(\boldsymbol{A}(t), \boldsymbol{B}(t))$, and proj normalizes the rows of a matrix to have unit norm. The projection step (and, hence, the Riemannian nature of the optimization) is due to the scale-invariance of sign, and it ensures numerical stability. The term $\boldsymbol{G}(t)$ corresponds to Gaussian noise of arbitrarily small variance, which acts as a (probabilistic) smoothing for the discontinuity of sign at 0 and, therefore, implies that the gradient is well-defined along the trajectory of the algorithm. (Note that $\boldsymbol{G}(t)$ is not needed in experiments, as we use a straight-through estimator, see Appendix C.2).

Theorem 4.1 (Gradient descent does not capture the sparsity)**.** *Consider the gradient descent algorithm in* (6) *with* $\boldsymbol{x} \sim \mathrm{SG}_d(p)$ *and* $(\boldsymbol{G}(t))_{i,j} \sim \mathcal{N}(0, \sigma^2)$*, where* $d^{-\gamma_g} \leq \sigma \leq C/d$ *for some fixed* $1 < \gamma_g < \infty$*. Initialize the algorithm with* $\boldsymbol{B}(0)$ *equal to a row-normalized Gaussian, i.e.,* $\boldsymbol{B}'_{i,j}(0) \sim \mathcal{N}(0, 1/d)$*,* $\boldsymbol{B}(0) = \mathrm{proj}(\boldsymbol{B}'(0))$*, and let* $\boldsymbol{B}(0) = \boldsymbol{U}\boldsymbol{S}(0)\boldsymbol{V}^\top$ *be its SVD. Let the step size* η *be* $\Theta(1/\sqrt{d})$*. Then, for any fixed* $r < 1$ *and* $T_{\max} \in (0, \infty)$*, with probability at least* $1 - Cd^{-3/2}$*, the following holds for all* $t \leq T_{\max}/\eta$

$$\begin{aligned} &\boldsymbol{B}(t) = \boldsymbol{U}\boldsymbol{S}(t)\boldsymbol{V}^\top + \boldsymbol{R}(t), \\ &\left\|\boldsymbol{S}(t)\boldsymbol{S}(t)^\top - \boldsymbol{I}\right\|_{op} \leq C\exp\left(-c\eta t\right), \\ &\lim_{d\to\infty} \sup_{t\in[0, T_{\max}/\eta]} \|\boldsymbol{R}(t)\|_{op} = 0, \end{aligned} \tag{7}$$

where C, c are universal constants depending only on p, r and $T_{\max}$. Moreover, we have that, almost surely,

$$\lim_{t\to\infty}\lim_{d\to\infty}\mathcal{R}(\boldsymbol{A}(t),\boldsymbol{B}(t))=\mathcal{R}_{\mathrm{Gauss}}, \tag{8}$$

$$\lim_{d\to\infty}\sup_{t\in[0,T_{\max}/\eta]}\|\boldsymbol{B}(t)-\boldsymbol{B}_{\mathrm{Gauss}}(t)\|_{op}=0, \tag{9}$$

where $\mathcal{R}_{\mathrm{Gauss}}$ is defined in (5) and $\boldsymbol{B}_{\mathrm{Gauss}}(t)$ is obtained by running (6) with $\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})$.

In words, (7) gives a precise characterization of the gradient descent trajectory: throughout the dynamics, the eigenbasis of $\boldsymbol{B}(t)$ does not change significantly (i.e., it remains close to that of $\boldsymbol{B}(0)$) and, as t grows, all the singular values of $\boldsymbol{B}(t)$ approach 1. As a consequence, (8) gives that, at convergence, the MSE achieved by (6) with $\boldsymbol{x}\sim\mathrm{SG}_d(p)$ approaches $\mathcal{R}_{\mathrm{Gauss}}$, which corresponds to the compression of standard Gaussian data $\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})$. In fact, a stronger result holds: (9) gives that the whole trajectory of (6) for $\boldsymbol{x}\sim\mathrm{SG}_d(p)$ is the same as that obtained for $\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})$. The fact that the autoencoder model in (2) is not able to exploit the signal structure is quite surprising, especially since information-theoretically (see Appendix C.1) a sparse Gaussian source is more suitable for compression than its non-sparse counterpart.

Proof sketch. The sparse Gaussian distribution can be seen as the component-wise product between a standard Gaussian vector and a mask $\boldsymbol{m}\in\{0,1\}^d$ with i.i.d. Bernoulli(p) entries. The key idea is to approximate the randomness in the mask $\boldsymbol{m}$ with its average, which heuristically corresponds to having again Gaussian data. This is done formally by pushing the mask into the network parameters, and then using high-dimensional concentration tools to bound the deviation from the average.

We now go into the details. As a starting point, Lemma A.1 shows that, up to an error exponentially small in d, instead of the MSE in (3) we can consider the objective

$$\mathbb{E}_{\boldsymbol{m}}\left[\mathrm{Tr}\left[\boldsymbol{A}^\top\boldsymbol{A}\cdot\arcsin(\hat{\boldsymbol{B}}_{\boldsymbol{m}}\hat{\boldsymbol{B}}_{\boldsymbol{m}}^\top)\right]-\frac{2}{\sqrt{p}}\cdot\mathrm{Tr}\left[\boldsymbol{A}\hat{\boldsymbol{B}}_{\boldsymbol{m}}\right]\right]. \tag{10}$$

Here, $\boldsymbol{B}_{\boldsymbol{m}}$ denotes a masked version of $\boldsymbol{B}$, i.e., the columns of $\boldsymbol{B}$ are set to zero according to the Bernoulli mask $\boldsymbol{m}$, and $\hat{\boldsymbol{B}}_{\boldsymbol{m}}$ is obtained by normalizing the rows of $\boldsymbol{B}_{\boldsymbol{m}}$, i.e., $(\hat{\boldsymbol{B}}_{\boldsymbol{m}})_{i,:}=(\boldsymbol{B}_{\boldsymbol{m}})_{i,:}/\|(\boldsymbol{B}_{\boldsymbol{m}})_{i,:}\|_2$.

Next, we provide a number of concentration bounds for quantities to which the Bernoulli mask $\boldsymbol{m}$ is applied. We start with random vectors (Lemma A.7), random matrices (Lemmas A.8 and A.9), and quantities that appear when optimizing the objective (10) via gradient descent (Lemma A.11). We note that both the largest entry and the operator norm of the error matrix have to be controlled. Then, we take care of the row normalization in the definition of $\hat{\boldsymbol{B}}_{\boldsymbol{m}}$. To do so, Lemma A.12 is a general result showing that

$$\mathbb{E}_{\boldsymbol{m}}F\left(\hat{\boldsymbol{B}}_{\boldsymbol{m}}\right)\approx\mathbb{E}_{\boldsymbol{m}}F\left(\frac{1}{\sqrt{p}}\boldsymbol{B}_{\boldsymbol{m}}\right), \tag{11}$$

for a class of sufficiently regular matrix-valued functions F. In words, (11) gives that, on average over $\boldsymbol{m}$, the row normalization can be replaced by the multiplication with $1/\sqrt{p}$. This result is instantiated in Lemma A.13 for three choices of F useful for the analysis of gradient descent.

Armed with these technical tools, we are able to remove the effect of the masking from the gradient descent dynamics. First, Lemma A.14 focuses on the optimization of the matrix $\boldsymbol{A}$, which has a closed form due to the convexity of the objective (10) in $\boldsymbol{A}$. Next, Lemmas A.15 and A.16 estimate the gradient $\nabla_{\boldsymbol{B}(t)}$ as

$$\left\|\nabla_{\boldsymbol{B}(t)}-\boldsymbol{U}\tilde{\boldsymbol{S}}(t)\boldsymbol{V}^\top\right\|_{op}\le C(T_{\max})\cdot\frac{\log^{10}(d)}{\sqrt{d}},$$

where $\boldsymbol{U},\boldsymbol{V}$ come from the SVD of $\boldsymbol{B}(0)$, $\boldsymbol{S}(t)$ is a diagonal matrix containing the singular values of $\boldsymbol{B}(t)$, and $\tilde{\boldsymbol{S}}(t)=G(\boldsymbol{S}(t))$ for a deterministic function G. This shows that, up to the leading order in the approximation, the singular vectors of $\boldsymbol{B}(t)$ are fixed along the gradient trajectory. Crucially, the function G does not depend on the sparsity p of the data. Thus, for any $p\in(0,1)$, the gradient update for the masked objective (10) is close to the update for the same objective without the masking (i.e., corresponding to the compression of Gaussian data with $p=1$).

Finally, Lemma A.19 derives an a-priori Grönwall-type estimate, which bootstraps the bounds to the whole gradient descent trajectory (6) and concludes the proof. The complete argument is deferred to Appendix A. □

Beyond Gaussian data: Phase transitions, staircases in the learning dynamics, and image data. For general i.i.d. distributions of the data $\boldsymbol{x}$, we empirically observe that the minimizers of the model in (2) found by stochastic gradient descent (SGD) either *(i)* coincide with those obtained for standard Gaussian data, or *(ii)* are equivalent to (suitably subsampled) permutations of the identity. Up to a permutation of the neurons, these two candidates can be expressed as:

$$\hat{\boldsymbol{x}}_{\mathrm{Haar}}(\boldsymbol{x})=\alpha_{\mathrm{Haar}}\cdot\boldsymbol{B}^\top\mathrm{sign}(\boldsymbol{B}\boldsymbol{x}), \tag{12}$$

$$\hat{\boldsymbol{x}}_{\mathrm{Id}}(\boldsymbol{x})=\alpha_{\mathrm{Id}}\cdot\begin{bmatrix}\boldsymbol{I}_n\\ \boldsymbol{0}_{(d-n)\times n}\end{bmatrix}\mathrm{sign}([\boldsymbol{I}_n,\boldsymbol{0}_{n\times(d-n)}]\boldsymbol{x}),$$

where $\boldsymbol{B}$ is obtained by subsampling a Haar matrix (i.e., a matrix taken uniformly from the group of rotations), $\boldsymbol{0}_{(d-n)\times n}$ is a $(d-n)\times n$ matrix of zeros, and $(\alpha_{\mathrm{Haar}},\alpha_{\mathrm{Id}})$ are scalar coefficients. The losses of these two candidates can be expressed in a closed form as derived below.

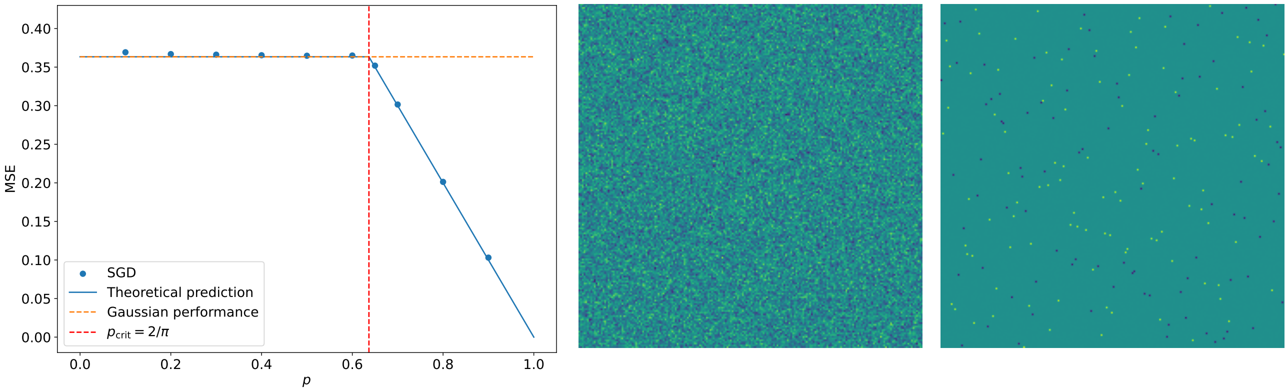

Figure 1. Compression of sparse Rademacher data via the two-layer autoencoder in (2). We set $d = 200$ and $r = 1$. *Left.* MSE achieved by SGD at convergence, as a function of the sparsity level p. The empirical values (dots) match our theoretical prediction (blue line): for $p < p_{\text{crit}}$, the loss is equal to the value obtained for Gaussian data, i.e., $\mathcal{R}_{\text{Gauss}} = 1 - 2r/\pi$; for $p \geq p_{\text{crit}}$, the loss is smaller, and it is equal to $1 - r \cdot (\mathbb{E}|x_1|)^2 = 1 - r \cdot p$. *Center.* Encoder matrix $\boldsymbol{B}$ at convergence of SGD when $p = 0.3 < p_{\text{crit}}$: the matrix is a random rotation. *Right.* Encoder matrix $\boldsymbol{B}$ at convergence of SGD when $p = 0.7 \geq p_{\text{crit}}$: the negative sign in part of the entries of $\boldsymbol{B}$ is cancelled by the corresponding sign in the entries of $\boldsymbol{A}$; hence, $\boldsymbol{B}$ is equivalent to a permutation of the identity.

Proposition 4.2 (Candidate comparison)**.** *Let $r \leq 1$ and let $\boldsymbol{x}$ have i.i.d. components with zero mean and unit variance. Then, we have that, almost surely, the MSE of $\hat{\boldsymbol{x}}_{\text{Haar}}(\cdot)$ coincides with the Gaussian performance $\mathcal{R}_{\text{Gauss}}$ in (5), i.e.,*

$$\min_{\alpha_{\text{Haar}} \in \mathbb{R}} \lim_{d \to \infty} \frac{1}{d} \cdot \mathbb{E}_{\boldsymbol{x}} \left[\|\hat{\boldsymbol{x}}_{\text{Haar}}(\boldsymbol{x}) - \boldsymbol{x}\|_2^2 \right] = 1 - \frac{2}{\pi} \cdot r \,. \quad (13)$$

Furthermore, we have that, for any d,

$$\min_{\alpha_{\text{Id}} \in \mathbb{R}} \frac{1}{d} \cdot \mathbb{E}_{\boldsymbol{x}} \left[\|\hat{\boldsymbol{x}}_{\text{Id}}(\boldsymbol{x}) - \boldsymbol{x}\|_2^2 \right] = 1 - r \cdot (\mathbb{E}|x_1|)^2 \,, \quad (14)$$

where x_1 is the first component of $\boldsymbol{x}$. This implies that $\hat{\boldsymbol{x}}_{\text{Id}}(\cdot)$ is superior to $\hat{\boldsymbol{x}}_{\text{Haar}}(\cdot)$ in terms of MSE whenever

$$\mathbb{E}|x_1| > \sqrt{2/\pi} = \mathbb{E}_{g \sim \mathcal{N}(0,1)}|g|. \quad (15)$$

The MSE of $\hat{\boldsymbol{x}}_{\text{Id}}(\cdot)$ in (14) is obtained via a direct calculation. To evaluate the MSE of $\hat{\boldsymbol{x}}_{\text{Haar}}(\cdot)$ in (13), we relate this estimator to the first iterate of the RI-GAMP algorithm proposed by Venkataramanan et al. (2022). Then, the high dimensional limit of $\|\boldsymbol{B}^\top \text{sign}(\boldsymbol{B}\boldsymbol{x}) - \boldsymbol{x}\|_2^2$ follows from the state evolution analysis of RI-GAMP. A similar strategy will be used also in Section 5 to analyze different decoding architectures. The complete proof is in Appendix B.1.

As mentioned above, our numerical results lead us to conjecture that SGD recovers either of the candidates in (12), depending on which achieves a smaller loss. Specifically, if condition (15) is met, the SGD predictor converges to $\hat{\boldsymbol{x}}_{\text{Id}}(\cdot)$ and improves upon the Gaussian loss $\mathcal{R}_{\text{Gauss}}$; otherwise, it converges to $\hat{\boldsymbol{x}}_{\text{Haar}}(\cdot)$ and its MSE is equal to $\mathcal{R}_{\text{Gauss}}$.

For sparse Gaussian data, condition (15) is never satisfied, as $\mathbb{E}_{x_1 \sim \text{SG}_1(p)}|x_1| = \sqrt{2p/\pi} \leq \sqrt{2/\pi}$. In fact, as proved in Theorem 4.1, the SGD solution approaches $\hat{\boldsymbol{x}}_{\text{Haar}}(\cdot)$ and its MSE matches $\mathcal{R}_{\text{Gauss}}$.

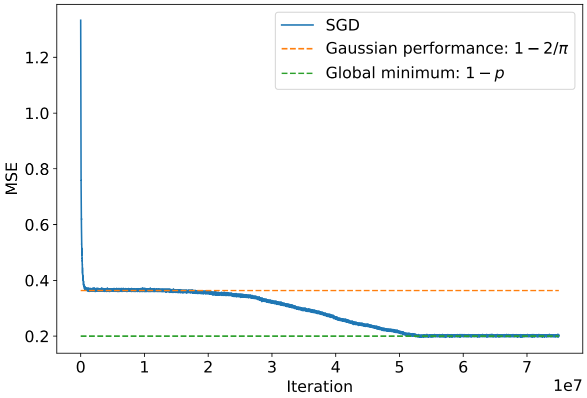

Figure 2. Compression of sparse Rademacher data via the two-layer autoencoder in (2). We set $d = 200$, $r = 1$ and $p = 0.8$. The MSE is plotted as a function of the number of iterations and, as $p > p_{\text{crit}}$, it displays a staircase behavior.

For sparse Rademacher data[1], condition (15) reduces to $p > p_{\text{crit}} := 2/\pi \approx 0.64$, and Figure 1 shows a *phase transition* in the structure of the minimizers found by SGD:

- For $p < p_{\text{crit}}$, SGD converges to a solution s.t. $\boldsymbol{B}$ is a uniform rotation (central heatmap) and the MSE is close to $\mathcal{R}_{\text{Gauss}} = 1 - 2r/\pi$, see (13).
- For $p > p_{\text{crit}}$, SGD converges to a solution s.t. $\boldsymbol{B}$ is equivalent to a permutation of the identity (right heatmap) and the MSE is close to $1 - r \cdot (\mathbb{E}|x_1|)^2 = 1 - r \cdot p$, see (14). In both cases, $\boldsymbol{A} \propto \boldsymbol{B}^\top$.

[1] Each i.i.d. component is equal to 0 w.p. $1 - p$ and to $\pm 1/\sqrt{p}$ w.p. $p/2$, which ensures a unit second moment for all $p \in [0, 1]$.

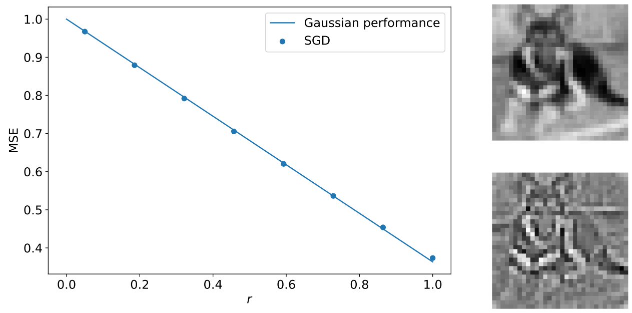

Figure 3. Compression of masked and whitened CIFAR-10 images of the class "dog" via the two-layer autoencoder in (2). First, the data is whitened so that it has identity covariance (as in the setting of Theorem 4.1). Then, the data is masked by setting each pixel independently to 0 with probability $p = 0.7$. An example of an original image is on the top right, and the corresponding masked and whitened image is on the bottom right. The SGD loss at convergence (dots) matches the solid line, which corresponds to the prediction in (5) for the compression of standard Gaussian data (with no sparsity).

If there is an improvement upon $\mathcal{R}_{\mathrm{Gauss}}$ (i.e., $p > p_{\mathrm{crit}}$), the SGD dynamics exhibits a *staircase* behavior. This phenomenon is displayed in Figure 2, which plots the error as a function of the number of SGD iterations for $p = 0.8 > p_{\mathrm{crit}}$: first, the MSE rapidly converges to $\mathcal{R}_{\mathrm{Gauss}}$; then, there is a plateau; finally, the global minimum $1 - r \cdot p$ is reached. We also remark that, as p approaches p_{crit}, the time needed by SGD to escape the plateau increases. A possible explanation is that, as p decreases, the noise due to masking increases, which increases the variance of the gradient. This makes it harder for $\boldsymbol{B}$ to find a direction towards a permutation of the identity (i.e., the global minimum). Additional evidence of both the phase transition and the staircase behavior of SGD is in Appendix C.3, where Figure 10 considers Rademacher data and Figures 11-12 data coming from a sparse mixture of Gaussians.

The proof strategy of Theorem 4.1 could be useful to track SGD until it reaches the plateau. However, characterizing the time-scale needed to escape the plateau likely requires new tools, and it provides an exciting research direction.

Finally, Figure 3 shows that our theory predicts well the behavior of the compression of *CIFAR-10 images* via the two-layer autoencoder in (2). We let x_1 be the empirical distribution of the image pixels after whitening and masking[2], and we verify that condition (15) does not hold. Then, as expected, the autoencoder is unable to capture the structure coming from masking part of the pixels, and the loss at the end of SGD training equals $\mathcal{R}_{\mathrm{Gauss}}$. Similar results hold for MNIST, see Figure 17 in Appendix C.4.

5. Provable benefit of nonlinearities and depth

In this section, we prove that more expressive decoders than the linear one in (2) capture the sparsity of the data and, therefore, improve upon the Gaussian loss $\mathcal{R}_{\mathrm{Gauss}}$.

5.1. Provable benefit of nonlinearities

First, we apply a nonlinearity at the output of the linear decoding layer, as in (4). The ResNet-like denoising architecture analyzed in (Cui & Zdeborová, 2023) suggests a suitable choice of the non-linearity. The corresponding denoising network has the following form:

$$\boldsymbol{x} \cdot \alpha + \boldsymbol{\Theta}_1 \cdot \tanh(\boldsymbol{\Theta}_2 \cdot \boldsymbol{x}). \qquad (16)$$

To map (16) to a scalar denoising function, we fix $\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2 \propto \boldsymbol{I}$ (or a row-subsampled version of an identity matrix for $r < 1$). Specifically, we take

$$f(x) = \alpha_1 \cdot x + \alpha_2 \cdot \tanh(\alpha_3 \cdot x), \qquad (17)$$

and run SGD on the weight matrices $(\boldsymbol{A}, \boldsymbol{B})$ and on the trainable parameters $(\alpha_1, \alpha_2, \alpha_3)$ in f. Figure 4 shows that, at convergence, the minimizers have the same weight-tied orthogonal structure as obtained for Gaussian data ($\boldsymbol{B}\boldsymbol{B}^\top = \boldsymbol{I}$, $\boldsymbol{A} \propto \boldsymbol{B}^\top$), see the left plot. However, in sharp contrast with Gaussian data, the loss is *smaller* than $\mathcal{R}_{\mathrm{Gauss}}$, see the blue dots on the right plot and compare them with the orange dashed curve. This empirical evidence motivates us to analyze the performance of autoencoders of the form (4), where $\boldsymbol{B}$ is obtained by subsampling a Haar matrix of appropriate dimensions and $\boldsymbol{A} = \boldsymbol{B}^\top$.

Proposition 5.1 (MSE characterization)**.** *Let $r \leq 1$ and $\boldsymbol{x}$ have i.i.d. components with zero mean and unit variance. Consider the autoencoder $\hat{\boldsymbol{x}}(\boldsymbol{x})$ in (4), where $\boldsymbol{B}$ is obtained by subsampling a Haar matrix, $\boldsymbol{A} = \boldsymbol{B}^\top$, and f is a Lipschitz function. Then, we have that, almost surely,*

$$\lim_{d \to \infty} \frac{1}{d} \cdot \mathbb{E}_{\boldsymbol{x}} \|\boldsymbol{x} - \hat{\boldsymbol{x}}(\boldsymbol{x})\|_2^2 = \mathbb{E}_{x_1, g} |x_1 - f(\mu x_1 + \sigma g)|_2^2, \qquad (18)$$

where x_1 is the first entry of $\boldsymbol{x}$, $g \sim \mathcal{N}(0,1)$ and independent of x_1, and the parameters (μ, σ) are given by

$$\mu = r \cdot \sqrt{\frac{2}{\pi}}, \quad \sigma^2 = r\left(1 - r \cdot \frac{2}{\pi}\right) > 0. \qquad (19)$$

Proposition 5.1 is a generalization of Proposition 4.2, which corresponds to taking a linear f. The idea is to relate $f(\boldsymbol{B}^\top \mathrm{sign}(\boldsymbol{B}\boldsymbol{x}))$ to the first iterate of a suitable RI-GAMP algorithm, so that the characterization in (18) follows from state evolution. The details are in Appendix B.2.

Armed with Proposition 5.1, one can readily establish the function f that minimizes the MSE for large d. This in fact corresponds to the f that minimizes the RHS of (18), i.e.,

$$f^*(y) = \mathbb{E}[x_1 | \mu x_1 + \sigma g = y], \qquad (20)$$

[2] The whitening makes the data have isotropic covariance, as required by our theory; the masking makes the data sparse.

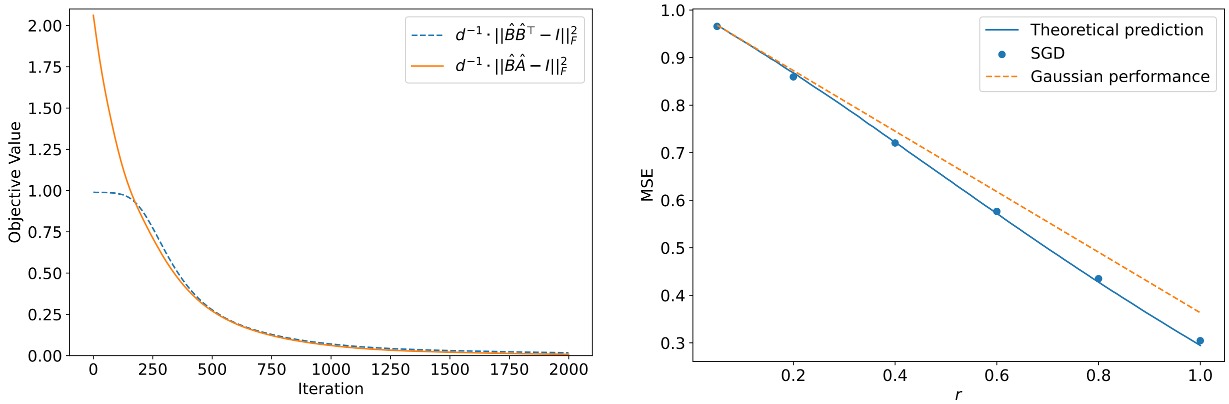

Figure 4. Compression of sparse Gaussian data via the autoencoder in (4), where f has the form in (17) and its parameters $(\alpha_1, \alpha_2, \alpha_3)$ are optimized via SGD. We set $d = 100$ and $p = 0.4$. *Left.* Distance between $\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top$, $\hat{\boldsymbol{B}}\hat{\boldsymbol{A}}$ and the identity, as a function of the number of iterations, where $\hat{\boldsymbol{B}}$, $\hat{\boldsymbol{A}}$ denote the row-normalized versions of $\boldsymbol{B}$, $\boldsymbol{A}$. $\|\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top - \boldsymbol{I}\|_F$ and $\|\hat{\boldsymbol{B}}\hat{\boldsymbol{A}} - \boldsymbol{I}\|_F$ decrease and tend to 0, meaning that (up to a rescaling of the rows) $\boldsymbol{B}\boldsymbol{A}$ and $\boldsymbol{B}\boldsymbol{B}^\top$ approach the identity. Here, we take $r = 1$. *Right.* MSE achieved by SGD at convergence, as a function of the compression rate r. The empirical values (dots) match the characterization of Proposition 5.1 for $f = f^*$ in (20) (blue line), and they outperform the MSE (5) obtained by compressing standard Gaussian data (orange dashed line).

as long as the latter is Lipschitz (so that Proposition 5.1 can be applied). Sufficient conditions for f^* to be Lipschitz are that either *(i)* x_1 has a log-concave density, or *(ii)* there exist independent random variables u_0, v_0 s.t. u_0 is Gaussian, v_0 is compactly supported and x_1 is equal in distribution to $u_0 + v_0$, see Lemma 3.8 of (Feng et al., 2022). The expression of f^* for distributions of x_1 considered in the experiments (sparse Gaussian, Laplace, and Rademacher) is derived in Appendix B.4.

The blue curve in the right plot of Figure 4 evaluates the RHS of (18) for the optimal $f = f^*$, when $x_1 \sim \mathrm{SG}_1(p)$. Two observations are in order:

1. The blue curve matches the blue dots, obtained by optimizing via SGD the matrices $\boldsymbol{A}, \boldsymbol{B}$ and f in the parametric form (17). This means that the SGD performance is accurately tracked by plugging the optimal function (20) into the prediction of Proposition 5.1.

2. The blue curve improves upon the Gaussian loss $\mathcal{R}_{\mathrm{Gauss}}$ (orange dashed line). This means that, while the two-layer autoencoder in (2) is stuck at the MSE in orange (as proved by Theorem 4.1), by incorporating a nonlinearity, the autoencoder in (4) does better. In fact, as shown in Figure 19 in Appendix C.6, the MSE achieved by the autoencoder in (4) with the optimal choice of f (namely, the RHS of (18) with $f = f^*$) is strictly lower than $\mathcal{R}_{\mathrm{Gauss}}$ for any $p \in (0, 1)$.

Beyond Gaussian data: Phase transitions, staircases in the learning dynamics, and image data. For general data $\boldsymbol{x}$ with i.i.d. zero-mean unit-variance components, the autoencoder in (4) displays a behavior similar to that described in Section 4 for the autoencoder in (2): the SGD minimizers of the weight matrices $\boldsymbol{A}, \boldsymbol{B}$ either exhibit a weight-tied orthogonal structure ($\boldsymbol{B}\boldsymbol{B}^\top = \boldsymbol{I}$, $\boldsymbol{A} \propto \boldsymbol{B}^\top$), or come from permutations of the identity. This leads to a *phase transition* in the structure of the minimizer (and in the MSE expression), as the sparsity p varies. To quantify the critical value of p at which the minimizer changes, one can compare the MSE when $\boldsymbol{B}$ is subsampled *(i)* from a Haar matrix, and *(ii)* from the identity. The former is readily obtained from Proposition 5.1 where f is given by (20), and the latter is given by the result below, which is proved in Appendix B.3.

Proposition 5.2. *Let $\boldsymbol{x}$ have i.i.d. components with zero mean, unit variance and a symmetric distribution (i.e., the law of x_1 is the same as that of $-x_1$). Define $\hat{\boldsymbol{x}}_{\mathrm{Id}}(\boldsymbol{x})$ as in* (12)*, and fix $r \leq 1$. Then, we have that, for any d,*

$$\min_f \frac{1}{d} \cdot \mathbb{E}_{\boldsymbol{x}} \left[\|f(\hat{\boldsymbol{x}}_{\mathrm{Id}}(\boldsymbol{x})) - \boldsymbol{x}\|_2^2 \right] = 1 - r \cdot (\mathbb{E}|x_1|)^2. \quad (21)$$

Figure 6 displays the phase transition for the compression of sparse Rademacher data:

- For $p < \tilde{p}_{\mathrm{crit}} \approx 0.67$, SGD converges to a solution with MSE given by the RHS of (18) with $f = f^*$. Furthermore, $\boldsymbol{B}$ is a uniform rotation (see the central heatmap in Figure 21 of Appendix C.7).
- For $p > \tilde{p}_{\mathrm{crit}}$, SGD converges to a solution with MSE given by the RHS of (21). Furthermore, $\boldsymbol{B}$ is equivalent to a permutation of the identity (see the right heatmap in Figure 21 of Appendix C.7).

By comparing the blue dots/curve with the orange dashed line in Figure 6, we also conclude that, for all p, the MSE of the autoencoder in (4) improves upon the Gaussian performance $\mathcal{R}_{\mathrm{Gauss}}$. This is in contrast with the behavior of

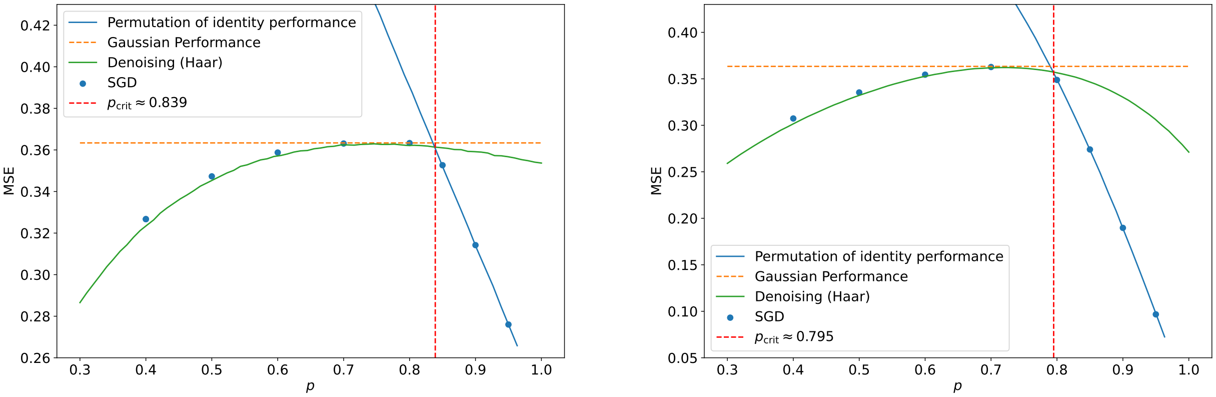

Figure 5. Compression of sparse Beta (*left*) and sparse Gaussian mixture (*right*) via the autoencoder in (4). We set $d = 200$ and $r = 1$. The MSE achieved by SGD at convergence is plotted as a function of the sparsity level p. The empirical values (blue dots) match our theoretical prediction (green/blue lines). For $p < \tilde{p}_{\mathrm{crit}}$, the MSE is given by Proposition 5.1 for $\boldsymbol{B}$ sampled from the Haar distribution; for $p \geq \tilde{p}_{\mathrm{crit}}$, the MSE is given by Proposition 5.2 for $\boldsymbol{B}$ equal to the identity.

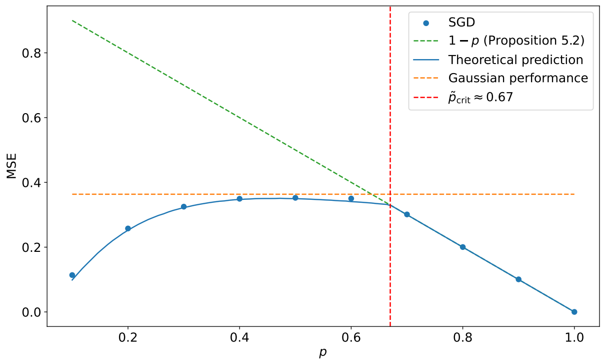

Figure 6. Compression of sparse Rademacher data via the autoencoder in (4). We set $d = 200$ and $r = 1$. The MSE achieved by SGD at convergence is plotted as a function of the sparsity level p. The empirical values (blue dots) match our theoretical prediction (blue line). For $p < \tilde{p}_{\mathrm{crit}}$, the MSE is given by Proposition 5.1 for $\boldsymbol{B}$ sampled from the Haar distribution; for $p \geq \tilde{p}_{\mathrm{crit}}$, the MSE is given by Proposition 5.2 for $\boldsymbol{B}$ equal to the identity.

the autoencoder in (2) which remains stuck at $\mathcal{R}_{\mathrm{Gauss}}$ for $p < 2/\pi$ (see Figure 1), and it demonstrates the benefit of adding the nonlinearity f.

For $p > \tilde{p}_{\mathrm{crit}}$, the learning dynamics exhibits again a *staircase* behavior in which the MSE first gets stuck at the value given by the RHS of (18) with $f = f^*$, and then reaches the optimal value of $1 - r \cdot (\mathbb{E}|x_1|)^2$. This is reported for $p = 0.9 > \tilde{p}_{\mathrm{crit}} \approx 0.67$ in Figure 22 of Appendix C.7.

Indeed this behaviour is persistent among i.i.d. data. In Figure 5, we showcase two additional data distributions, a sparse Beta and a sparse Gaussian mixture, see (114) and (112) in Appendix C.3 for details. We observe the same behaviour as in the Rademacher case, which is consistent with the heuristic described above.

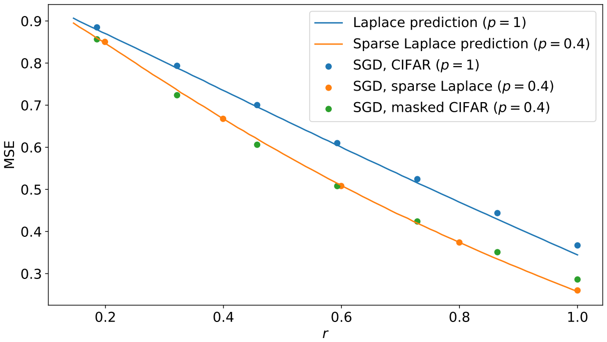

Figure 7. Compression of masked and whitened CIFAR-10 images of the class “dog” via the autoencoder in (4). We plot the MSE as a function of the compression rate r. Dots are obtained by training the decoder matrix $\boldsymbol{A}$ and the parameters $(\alpha_1, \alpha_2, \alpha_3)$ via SGD on masked ($p = 0.4$, green) or original ($p = 1$, blue) CIFAR-10 images. Continuous lines refer to the predictions of Proposition 5.1 for the optimal $f = f^*$ in (20), where x_1 has a Laplace distribution ($p = 1$, blue) or a sparse Laplace distribution ($p = 0.4$, orange). These curves match well the corresponding values obtained via SGD. Orange dots are obtained by training the matrices $\boldsymbol{A}, \boldsymbol{B}$ and the parameters $(\alpha_1, \alpha_2, \alpha_3)$ via SGD when $\boldsymbol{x}$ has i.i.d. sparse Laplace entries with $p = 0.4$.

Finally, Figure 7 shows that the key features we unveiled for the autoencoder in (4) are still present when compressing *sparse CIFAR-10 data*. The empirical distribution of the image pixels after whitening is well approximated by a Laplace random variable (see Figure 18 in Appendix C.5), thus we denote by x_1 the corresponding sparse Laplace distribution (see (109) in Appendix B.4 for a formal definition). The encoder matrix $\boldsymbol{B}$ is obtained by subsampling a Haar matrix, and it is fixed; the decoder matrix $\boldsymbol{A}$ and the parameters $(\alpha_1, \alpha_2, \alpha_3)$ in the definition (17) of f are obtained via SGD training. Two observations are in order:

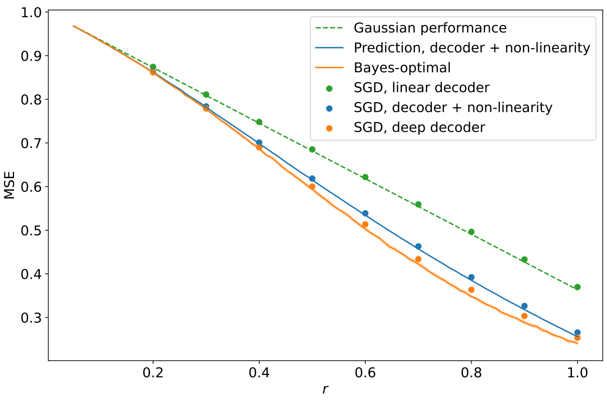

Figure 8. Compression of sparse Gaussian data $\boldsymbol{x} \sim \mathrm{SG}_d(p)$ for $p = 0.3$ and $d = 500$. We plot the MSE as a function of the compression rate r for various autoencoder architectures. The architecture in (23) (orange dots) outperforms the autoencoders in (2) (green dots) and in (4) (blue dots), and it approaches the Bayes-optimal MSE (orange line).

1. The autoencoder in (4) captures the sparsity: the MSE achieved on sparse data ($p = 0.4$, green dots) is lower than the MSE on non-sparse data ($p = 1$, blue dots).

2. For both values of p, the SGD performance matches the RHS of (18) (continuous lines) with $f = f^*$. As expected, this MSE is smaller than $1 - r \cdot (\mathbb{E}|x_1|)^2$, and it coincides with that obtained for compressing synthetic data with i.i.d. Laplace entries (orange dots).

5.2. Provable benefit of depth

We conclude by showing that the MSE can be further reduced by considering a multi-layer decoder. Our design of the decoding architecture is inspired by the RI-GAMP algorithm (Venkataramanan et al., 2022), which iteratively estimates $\boldsymbol{x}$ from an observation of the form $\sigma(\boldsymbol{B}\boldsymbol{x})$ via

$$\boldsymbol{x}^t = \boldsymbol{B}^\top \hat{\boldsymbol{z}}^t - \sum_{i=1}^{t-1} \beta_{t,i} \hat{\boldsymbol{x}}^i, \quad \hat{\boldsymbol{x}}^t = f_t(\boldsymbol{x}^1, \cdots, \boldsymbol{x}^t), \quad (22)$$
$$\boldsymbol{z}^t = \boldsymbol{B}\hat{\boldsymbol{x}}^t - \sum_{i=1}^{t} \alpha_{t,i} \hat{\boldsymbol{z}}^i, \quad \hat{\boldsymbol{z}}^{t+1} = g_t(\boldsymbol{z}^1, \cdots, \boldsymbol{z}^t, \hat{\boldsymbol{z}}^1).$$

Here, f_t, g_t are Lipschitz and applied component-wise, and the initialization is $\hat{\boldsymbol{z}}^1 = \mathrm{sign}(\boldsymbol{B}\boldsymbol{x})$. The coefficients $\{\beta_{t,i}\}$ and $\{\alpha_{t,i}\}$ are chosen so that, under suitable assumptions on $\boldsymbol{B}$,[3] the empirical distribution of the iterates is tracked via a low-dimensional recursion, known as *state evolution*. This in turn allows to evaluate the MSE $\lim_{d\to\infty} \frac{1}{d}\|\boldsymbol{x} - \hat{\boldsymbol{x}}^t\|_2^2$.

The results of Proposition 4.2 and 5.1 follow from relating the autoencoders in (2)-(4) to RI-GAMP iterates in (22). More generally, $\hat{\boldsymbol{x}}^t$ is obtained by multiplications with $\boldsymbol{B}, \boldsymbol{B}^\top$, linear combinations of previous iterates, and component-wise applications of Lipschitz functions. As such, it can be expressed via a multi-layer decoder with residual connections. The numerical results in (Venkataramanan et al., 2022) show that taking f_t, g_t as posterior means (as in (20)) leads to Bayes-optimal performance, having fixed the encoder matrix $\boldsymbol{B}$. Thus, this provides a proof-of-concept of the optimality of multi-layer decoders.

In fact, Figure 8 shows that an architecture with three decoding layers is already near-optimal when $\boldsymbol{x} \sim \mathrm{SG}_d(p)$. The decoder output is $\hat{\boldsymbol{x}}^2$ computed as (see also the block diagram in Figure 25 in Appendix C.8)

$$\begin{aligned}
&\hat{\boldsymbol{z}}^1 = \mathrm{sign}(\boldsymbol{B}\boldsymbol{x}), \quad \boldsymbol{x}^1 = \boldsymbol{W}_1 \hat{\boldsymbol{z}}^1, \quad \hat{\boldsymbol{x}}^1 = f_1(\boldsymbol{x}^1), \\
&\hat{\boldsymbol{z}}^2 = g_1(\boldsymbol{V}_1 \hat{\boldsymbol{x}}^1 \oplus_1 \hat{\boldsymbol{z}}^1), \qquad (23)\\
&\boldsymbol{x}^2 = \hat{\boldsymbol{x}}^1 \oplus_2 \boldsymbol{W}_2 \hat{\boldsymbol{z}}^2, \quad \hat{\boldsymbol{x}}^2 = f_2(\boldsymbol{x}^1 \oplus_3 \boldsymbol{x}^2).
\end{aligned}$$

Here, $f_1(\cdot), f_2(\cdot), g_1(\cdot)$ are trainable parametric functions of the form in (17) and, for $i \in \{1, 2, 3\}$, $\boldsymbol{a} \oplus_i \boldsymbol{b} = \beta_i \boldsymbol{a} + \gamma_i \boldsymbol{b}$, where $\{\beta_i, \gamma_i\}$ are also trained. The plot demonstrates the benefit of employing more expressive decoders:

1. The green dots are obtained via SGD training of the autoencoder in (2) and, as proved in Theorem 4.1, they match the Gaussian performance $\mathcal{R}_{\mathrm{Gauss}}$.

2. The blue dots are obtained via SGD training of the autoencoder in (4) and they match the prediction of Proposition 5.1 with $f = f^*$ in (20).

3. The orange dots are obtained by using the decoder in (23) where $\boldsymbol{W}_1 = \boldsymbol{W}_2 = \boldsymbol{B}^\top$, $\boldsymbol{V}_1 = \boldsymbol{B}$ are subsampled Haar matrices and the parameters in the functions $f_1, f_2, g_1, \{\oplus_i\}_{i=1}^3$ are trained via SGD. Similar results are obtained by training also $\boldsymbol{W}_1, \boldsymbol{W}_2, \boldsymbol{V}_1$, although at the cost of a slower convergence.

In summary, the architecture in (23) improves upon those in (2)-(4), and it approaches the orange curve which gives the Bayes-optimal MSE achievable by fixing a rotationally invariant encoder matrix $\boldsymbol{B}$ (Ma et al., 2021). Additional details are deferred to Appendix C.8.

We also note that considering a deep fully-connected decoder in place of the architecture in (23) does not improve upon the autoencoder in (4). In fact, while sufficiently wide and deep models have high expressivity, their SGD training is notoriously difficult, due to e.g. vanishing/exploding gradients (Glorot & Bengio, 2010; He et al., 2016).

Impact statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

[3] $\boldsymbol{B}$ has to be bi-rotationally invariant in law, namely, the matrices appearing in its SVD are sampled from the Haar distribution.

Acknowledgements

Kevin Kögler, Alexander Shevchenko and Marco Mondelli are supported by the 2019 Lopez-Loreta Prize. Hamed Hassani acknowledges the support by the NSF CIF award (1910056) and the NSF Institute for CORE Emerging Methods in Data Science (EnCORE).

References

Abbe, E., Boix-Adsera, E., Brennan, M. S., Bresler, G., and Nagaraj, D. The staircase property: How hierarchical structure can guide deep learning. *Advances in Neural Information Processing Systems*, 2021.

Abbe, E., Adsera, E. B., and Misiakiewicz, T. The merged-staircase property: a necessary and nearly sufficient condition for SGD learning of sparse functions on two-layer neural networks. *Conference on Learning Theory*, 2022.

Abbe, E., Adsera, E. B., and Misiakiewicz, T. SGD learning on neural networks: leap complexity and saddle-to-saddle dynamics. *Conference on Learning Theory*, 2023a.

Abbe, E., Bengio, S., Boix-Adserà, E., Littwin, E., and Susskind, J. M. Transformers learn through gradual rank increase. *Advances in Neural Information Processing Systems*, 2023b.

Agustsson, E., Mentzer, F., Tschannen, M., Cavigelli, L., Timofte, R., Benini, L., and Gool, L. V. Soft-to-hard vector quantization for end-to-end learning compressible representations. *Advances in Neural Information Processing Systems*, 2017.

Arimoto, S. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, 1972.

Ballé, J., Laparra, V., and Simoncelli, E. P. End-to-end optimized image compression. *International Conference on Learning Representations*, 2017.

Bao, X., Lucas, J., Sachdeva, S., and Grosse, R. B. Regularized linear autoencoders recover the principal components, eventually. *Advances in Information Processing Systems*, 2020.

Berthier, R. Incremental learning in diagonal linear networks. *Journal of Machine Learning Research*, 2023.

Berthier, R., Montanari, A., and Zhou, K. Learning time-scales in two-layers neural networks. *arXiv preprint arXiv:2303.00055*, 2023.

Boufounos, P. T. and Baraniuk, R. G. 1-bit compressive sensing. *Conference on Information Sciences and Systems*, 2008.

Cover, T. M. and Thomas, J. A. *Elements of Information Theory*. 2006.

Cui, H. and Zdeborová, L. High-dimensional asymptotics of denoising autoencoders. *Advances in Neural Information Processing Systems*, 2023.

Dytso, A., Poor, H. V., and Shitz, S. S. A general derivative identity for the conditional mean estimator in gaussian noise and some applications. *IEEE International Symposium on Information Theory*, 2020.

Feng, O. Y., Venkataramanan, R., Rush, C., Samworth, R. J., et al. A unifying tutorial on approximate message passing. *Foundations and Trends® in Machine Learning*, 2022.

Gidel, G., Bach, F., and Lacoste-Julien, S. Implicit regularization of discrete gradient dynamics in linear neural networks. *Advances in Neural Information Processing Systems*, 2019.

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. *International Conference on Artificial Intelligence and Statistics*, 2010.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Jacot, A., Ged, F., Şimşek, B., Hongler, C., and Gabriel, F. Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity. *arXiv preprint arXiv:2106.15933*, 2021.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *International Conference on Learning Representations*, 2014.

Kunin, D., Bloom, J., Goeva, A., and Seed, C. Loss landscapes of regularized linear autoencoders. *International Conference on Machine Learning*, 2019.

Li, Y., Fan, Z., Sen, S., and Wu, Y. Random linear estimation with rotationally-invariant designs: Asymptotics at high temperature. *IEEE Transactions on Information Theory*, 2023.

Ma, J. and Ping, L. Orthogonal AMP. *IEEE Access*, 2017.

Ma, J., Xu, J., and Maleki, A. Analysis of sensing spectral for signal recovery under a generalized linear model. *Advances in Neural Information Processing Systems*, 2021.

McCullagh, P. and Nelder, J. A. *Generalized linear models*. Monographs on Statistics and Applied Probability. 1989.

Milanesi, P., Kadri, H., Ayache, S., and Artières, T. Implicit regularization in deep tensor factorization. *IEEE International Joint Conference on Neural Networks*, 2021.

Mondelli, M. and Venkataramanan, R. Approximate message passing with spectral initialization for generalized linear models. *Journal of Statistical Mechanics: Theory and Experiment*, 2022.

Nguyen, P.-M. Analysis of feature learning in weight-tied autoencoders via the mean field lens. *arXiv preprint arXiv:2102.08373*, 2021.

Oftadeh, R., Shen, J., Wang, Z., and Shell, D. Eliminating the invariance on the loss landscape of linear autoencoders. *International Conference on Machine Learning*, 2020.

Peng, P., Jalali, S., and Yuan, X. Solving inverse problems via auto-encoders. *IEEE Journal on Selected Areas in Information Theory*, 2020.

Pesme, S. and Flammarion, N. Saddle-to-saddle dynamics in diagonal linear networks. *arXiv preprint arXiv:2304.00488*, 2023.

Rangan, S. Generalized approximate message passing for estimation with random linear mixing. *IEEE International Symposium on Information Theory*, 2011.

Rangan, S., Schniter, P., and Fletcher, A. K. Vector approximate message passing. *IEEE Transactions on Information Theory*, 2019.

Refinetti, M. and Goldt, S. The dynamics of representation learning in shallow, non-linear autoencoders. *International Conference on Machine Learning*, 2022.

Schniter, P., Rangan, S., and Fletcher, A. K. Vector approximate message passing for the generalized linear model. In *Asilomar Conference on Signals, Systems and Computers*, 2016.

Shevchenko, A., Kögler, K., Hassani, H., and Mondelli, M. Fundamental limits of two-layer autoencoders, and achieving them with gradient methods. *International Conference on Machine Learning*, 2023.

Simon, J. B., Knutins, M., Ziyin, L., Geisz, D., Fetterman, A. J., and Albrecht, J. On the stepwise nature of self-supervised learning. *arXiv preprint arXiv:2303.15438*, 2023.

Székely, E., Bardone, L., Gerace, F., and Goldt, S. Learning from higher-order statistics, efficiently: hypothesis tests, random features, and neural networks. *arXiv preprint arXiv:2312.14922*, 2023.

Takeda, K., Uda, S., and Kabashima, Y. Analysis of CDMA systems that are characterized by eigenvalue spectrum. *Europhysics Letters*, 2006.

Takeuchi, K. Rigorous dynamics of expectation-propagation-based signal recovery from unitarily invariant measurements. *IEEE Transactions on Information Theory*, 2019.

Theis, L., Shi, W., Cunningham, A., and Huszár, F. Lossy image compression with compressive autoencoders. *International Conference on Learning Representations*, 2017.

Tulino, A. M., Caire, G., Verdú, S., and Shamai, S. Support recovery with sparsely sampled free random matrices. *IEEE Transactions on Information Theory*, 2013.

Van Den Oord, A., Vinyals, O., et al. Neural discrete representation learning. *Advances in Neural Information Processing Systems*, 2017.

Venkataramanan, R., Kögler, K., and Mondelli, M. Estimation in rotationally invariant generalized linear models via approximate message passing. *International Conference on Machine Learning*, 2022.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*. 2018.

Visick, G. A quantitative version of the observation that the hadamard product is a principal submatrix of the kronecker product. *Linear Algebra and Its Applications*, 2000.

Winkelbauer, A. Moments and absolute moments of the normal distribution. *arXiv preprint arXiv:1209.4340*, 2012.

Yin, P., Lyu, J., Zhang, S., Osher, S., Qi, Y., and Xin, J. Understanding straight-through estimator in training activation quantized neural nets. *arXiv preprint arXiv:1903.05662*, 2019.

A. Proof of Theorem 4.1

A.1. Additional notation

Given two matrices $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ of the same shape, their element-wise Schur product is $\boldsymbol{M}_1 \circ \boldsymbol{M}_2$ and the ℓ-th element-wise power is $\boldsymbol{M}_1^{\circ \ell}$. The same notation is adopted for the element-wise product of vectors, i.e., $\boldsymbol{v} \circ \boldsymbol{u}$. By convention, if $\boldsymbol{B}_{i,:}$ is a vector of zeroes, its normalization $\hat{\boldsymbol{B}}_{i,:}$ is also a vector of zeroes. We fix the evaluation of $\operatorname{sign}(\cdot)$ at the origin to be a Rademacher random variable, i.e., $\operatorname{sign}(0)$ takes values in the set $\{-1, 1\}$ with equal probability. Note that this is a technical assumption with no bearing on the proof of the result.

For a matrix $\boldsymbol{B}$, we denote its i-th row by $\boldsymbol{b}_i = \boldsymbol{B}_{i,:}$, the exception being that by convention $\boldsymbol{a}_k$ denotes the k-th column of $\boldsymbol{A}$. $\boldsymbol{B}_{\boldsymbol{m}}$ is the masked version of a matrix $\boldsymbol{B}$, where masking is defined as $\bar{\boldsymbol{b}}_i = \boldsymbol{b}_i \circ \boldsymbol{m}$ and $\boldsymbol{m}$ has i.i.d. Bernoulli(p) components. For convenience of notation, we define $\bar{\boldsymbol{B}} = \boldsymbol{B}_{\boldsymbol{m}}$ *only* for the matrix $\boldsymbol{B}$. By convention, masking has priority over transposing, i.e., $\boldsymbol{B}_{\boldsymbol{m}}^\top = (\boldsymbol{B}_{\boldsymbol{m}})^\top$. For $\boldsymbol{B}$ (and only $\boldsymbol{B}$), we define $\hat{\boldsymbol{B}} = \hat{\boldsymbol{B}}_{\boldsymbol{m}} = \hat{\boldsymbol{D}} \bar{\boldsymbol{B}}_{\boldsymbol{m}}$, where $\hat{\boldsymbol{D}}$ is a diagonal matrix with entries $\hat{\boldsymbol{D}}_{i,i} = 1/\left\|\bar{\boldsymbol{b}}_i\right\|$, as the masked and re-normalized version of $\boldsymbol{B}$. We define $\|\boldsymbol{B}\|_{\max} := \max_{i,j} |\boldsymbol{B}_{i,j}|$.

We use the following convention for the constants. All constants are independent of d including those that are dependent on the quantities $p, r, f(x) = \arcsin(x), \alpha = f(1) - 1$ and the dependence on these quantities will be suppressed most of the time. Uppercase constants like C, C_X, C_R should be thought of as being much larger than 1, whereas lowercase constants should be thought of as being smaller than 1.

For a vector, the norm $\|\cdot\|$ without subscript always refers to the 2-norm $\|\cdot\|_2$. Unless stated otherwise, we consider the space of matrices $\mathcal{M}_{n\times d}$ to be endowed with $\|\cdot\|_{op}$. For a matrix $\boldsymbol{R}$, we denote by $O(\boldsymbol{R})$ a matrix of the same dimensions as $\boldsymbol{R}$ with $\|O(\boldsymbol{R})\|_{op} \le C\|\boldsymbol{R}\|_{op}$. This is a way to extend the big O notation to matrices. Similarly, we will use the notation O_{max} which functions as O except that $\|\cdot\|_{op}$ is replaced by $\|\cdot\|_{max}$. We will often use that $n = O(d)$, since $r = \frac{n}{d}$ is fixed.

A.2. Outline

The start of our analysis is the following lemma.

Lemma A.1. *Let $\mathcal{R}(\cdot,\cdot)$ be the MSE defined in (3), with $\boldsymbol{x} \sim \mathrm{SG}_d(p)$. Assume that all entries of $\boldsymbol{B}$ are not zero. Then, up to a multiplicative scaling and an additive constant, $\mathcal{R}(\boldsymbol{A},\boldsymbol{B})$ is given by*

$$\mathbb{E}_{\boldsymbol{m}\neq \boldsymbol{0}}\left[\operatorname{Tr}\left[\boldsymbol{A}^\top \boldsymbol{A} f(\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top)\right] - \frac{2}{\sqrt{p}}\operatorname{Tr}\left[\boldsymbol{A}\hat{\boldsymbol{B}}\right]\right] + O\left((1-p)^d \|\boldsymbol{A}\|_{op}^2\right), \tag{24}$$

where $f = \arcsin$ is applied component-wise *and the second term on the RHS is independent of $\boldsymbol{B}$.*

Proof. For any $\boldsymbol{m} \neq \boldsymbol{0}$ we can fix $\boldsymbol{m}$ and apply Lemma 4.1 in (Shevchenko et al., 2023). The second term on the RHS corresponds to $\boldsymbol{m} = \boldsymbol{0}$. □

We now briefly elaborate on some technical details. First, by convention, all expectations over $\boldsymbol{m}$ are understood to be over $\boldsymbol{m} \neq \boldsymbol{0}$. Second, as the last term on the RHS in (24) does not depend on $\boldsymbol{B}$, it suffices to take the gradient of the objective without it. Lastly, the term $O\left((1-p)^d \|\boldsymbol{A}\|_{op}^2\right)$ has a negligible effect when running the gradient descent algorithm in (6). In fact, a by-product of our analysis is that $\boldsymbol{A}$ has bounded norm throughout the training trajectory, see Lemma A.14. Hence, the quantity $O\left((1-p)^d \|\boldsymbol{A}\|_{op}^2\right)$ is exponentially small in d and, therefore, it can be incorporated in the error of order $C\frac{\mathrm{poly}(\log(\mathrm{d}))}{\sqrt{d}}$ being tracked during the recursion.

As a result, we can consider the objective

$$\mathbb{E}_{\boldsymbol{m}}\left[\operatorname{Tr}\left[\boldsymbol{A}^\top \boldsymbol{A} f(\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top)\right] - \frac{2}{\sqrt{p}}\operatorname{Tr}\left[\boldsymbol{A}\hat{\boldsymbol{B}}\right]\right], \tag{25}$$

where $\boldsymbol{m} \in \{0,1\}^d$ denotes a mask with i.i.d. Bernoulli(p) entries, $\hat{\boldsymbol{b}}_i = \boldsymbol{m} \circ \boldsymbol{b}_i / \|\boldsymbol{m} \circ \boldsymbol{b}_i\|_2$ and $(1-p)$ is the sparsity. Thus, the Riemannian gradient descent algorithm (6) applied to the objective (25) can be rewritten as

$$\begin{aligned} \boldsymbol{A}(t) &= \frac{1}{\sqrt{p}}\mathbb{E}_{\boldsymbol{m}}\hat{\boldsymbol{B}}(t)^\top \left(\mathbb{E}_{\boldsymbol{m}} f(\hat{\boldsymbol{B}}(t)\hat{\boldsymbol{B}}(t)^\top)\right)^{-1}, \\ \boldsymbol{B}'(t) &:= \boldsymbol{B}(t) - \eta\left(\nabla_{\boldsymbol{B}(t)} + \boldsymbol{G}(t)\right), \boldsymbol{B}(t+1) := \operatorname{proj}(\boldsymbol{B}'(t)), \end{aligned} \tag{26}$$

where $\boldsymbol{A}(t)$ is the optimal matrix for a fixed $\boldsymbol{B}(t)$, $\nabla_{\boldsymbol{B}(t)}$ is defined below in (30) and $(\boldsymbol{G}(t))_{i,j} \sim \mathcal{N}(0, \sigma^2)$ with $d^{-\gamma_g} \leq \sigma \leq C/d$ for some fixed $1 < \gamma_g < \infty$.

The goal of this Appendix is to show the following theorem.

Theorem A.2. *Consider the gradient descent* (26) *with $\boldsymbol{x} \sim \mathrm{SG}_d(p)$. Initialize the algorithm with $\boldsymbol{B}(0)$ equal to a row-normalized Gaussian, i.e., $\boldsymbol{B}'_{i,j}(0) \sim \mathcal{N}(0, 1/d)$, $\boldsymbol{B}(0) = \mathrm{proj}(\boldsymbol{B}'(0))$, and let $\boldsymbol{B}(0) = \boldsymbol{U}\boldsymbol{S}(0)\boldsymbol{V}^\top$ be its SVD. Let the step size η be $\Theta(1/\sqrt{d})$. Then, for any fixed $r < 1$ and $T_{\max} \in (0, \infty)$, with probability at least $1 - Cd^{-3/2}$, the following holds for all $t \leq T_{\max}/\eta$*

$$\begin{aligned} &\boldsymbol{B}(t) = \boldsymbol{U}\boldsymbol{S}(t)\boldsymbol{V}^\top + \boldsymbol{R}(t), \\ &\left\|\boldsymbol{S}(t)\boldsymbol{S}(t)^\top - \boldsymbol{I}\right\|_{op} \leq C \exp\left(-c\eta t\right), \\ &\lim_{d\to\infty} \sup_{t\in[0,T_{\max}/\eta]} \left\|\boldsymbol{R}(t)\right\|_{op} = 0, \end{aligned} \tag{27}$$

where C, c are universal constants depending only on p, r and $T_{\max}$. Moreover, we have that, almost surely,

$$\begin{aligned} &\lim_{t\to\infty} \lim_{d\to\infty} \mathcal{R}(\boldsymbol{A}(t), \boldsymbol{B}(t)) = \mathcal{R}_{\mathrm{Gauss}}, \\ &\lim_{d\to\infty} \sup_{t\in[0,T_{\max}/\eta]} \left\|\boldsymbol{B}(t) - \boldsymbol{B}_{\mathrm{Gauss}}(t)\right\|_{op} = 0, \end{aligned} \tag{28}$$

where $\mathcal{R}_{\mathrm{Gauss}}$ is defined in (5) *and $\boldsymbol{B}_{\mathrm{Gauss}}(t)$ is obtained by running* (26) *with $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.*

Let us provide a high-level overview of the proof strategy. Using high-dimensional probability tools, we will show that with high probability

$$\begin{aligned} &\boldsymbol{B}(t) = \boldsymbol{X}(t) + \boldsymbol{R}(t), \\ &\left\|\boldsymbol{X}(t)\right\|_{op} \leq C_X, \quad \left\|\boldsymbol{X}(t)\right\|_{max} \leq C_X \frac{\log(d)}{\sqrt{d}}, \quad \boldsymbol{X}(t) = \boldsymbol{U}\boldsymbol{S}(t)\boldsymbol{V}, \text{ with } \boldsymbol{U}, \boldsymbol{V} \text{ Haar}, \\ &\left\|\boldsymbol{R}(t)\right\|_{op} \leq C_R \frac{\log^{\alpha_R}(d)}{\sqrt{d}}, \\ &\boldsymbol{A}(t) = \boldsymbol{X}(t)^\top \left(\boldsymbol{X}(t)\boldsymbol{X}(t)^\top + \alpha \boldsymbol{I}\right)^{-1} + O\left(C_X^{10} \frac{\log^{10}(d)}{\sqrt{d}}\right) + O\left(\boldsymbol{R}\right), \end{aligned} \tag{29}$$

where $\alpha = f(1) - 1 = \arcsin(1) - 1$. This implies that the gradient in (30) concentrates to the Gaussian one, namely, to the gradient obtained for $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Then, an a-priori Grönwall-type inequality will extend these bounds for all times $t \in [0, T_{\max}]$. It is essential that the constants C_X, C_R in (29) can be chosen to only depend on $T_{\max}$, as otherwise the gradient dynamics could diverge in finite time. Thus, it is crucial that in all our lemmas we keep track of these constants explicitly and that in the error estimates they do not depend on each other. While the analysis is quite technical, the high-level idea is simple: if each term that depends on $\boldsymbol{m}$ were replaced by its mean, then we would immediately recover the Gaussian case $p = 1$ which was studied in (Shevchenko et al., 2023). By showing that each of the terms concentrates well enough, we can make this intuition rigorous. The main technical difficulty lies in controlling the additional error terms, which requires a more nuanced approach compared to the Gaussian case considered in (Shevchenko et al., 2023).

The rest of this appendix is structured as follows. Section A.3 contains a collection of auxiliary results that are simple applications of standard results. In Section A.4, we develop our high-dimensional concentration tools. In Section A.5, we use these tools to show that $\boldsymbol{A}, \hat{\boldsymbol{B}}, \nabla_{\boldsymbol{B}}$ all concentrate. Finally in Section A.6, we combine these approximations with an a-priori Grönwall bound in Lemma A.19, which allows us to bound the difference between the gradient trajectory and that obtained with Gaussian data.

A.3. Auxiliary results

A straightforward computation gives:

Lemma A.3 (Gradient formulas)**.** *The derivative of* (25) *w.r.t. $\boldsymbol{B}$ is given by*

$$(\nabla_{\boldsymbol{B}})_{k,:} = \mathbb{E}_{\boldsymbol{m}} \left[-2\frac{1}{\sqrt{p}} \boldsymbol{m} \circ \hat{\boldsymbol{J}}_k \boldsymbol{a}_k + 2\sum_{\ell=1}^{\infty} \ell c_\ell^2 \sum_{j\neq k} \langle \boldsymbol{a}_k, \boldsymbol{a}_j \rangle \langle \hat{\boldsymbol{b}}_k, \hat{\boldsymbol{b}}_j \rangle^{\ell-1} \hat{\boldsymbol{J}}_k \hat{\boldsymbol{b}}_j\right], \tag{30}$$

where $\hat{\boldsymbol{J}}_k = \frac{1}{\|\bar{\boldsymbol{b}}_k\|}\left(\boldsymbol{I} - \hat{\boldsymbol{b}}_k\hat{\boldsymbol{b}}_k^\top\right)$, $\boldsymbol{a}_k = \boldsymbol{A}_{:,k}$ *and* c_ℓ^2 *are the Taylor coefficients of* $\arcsin(x)$.

We will make extensive use of the following linear algebra results.

Lemma A.4 (Linear algebra results)**.** *The following results hold:*

1. $\left\|\bar{\boldsymbol{B}}\right\|_{op} \le \|\boldsymbol{B}\|_{op}$.
2. *For any* $\boldsymbol{M} \in \mathbb{R}^{n\times d}$, *we have* $\|\boldsymbol{M}\|_{op} \le \sqrt{n}\max_k \|\boldsymbol{M}_{k,:}\|$. *In particular,* $\left\|\hat{\boldsymbol{B}}\right\|_{op} \le \sqrt{n}$.
3. *For square matrices* $\boldsymbol{M}_1, \boldsymbol{M}_2 \in \mathbb{R}^{n\times n}$,
$$\|\boldsymbol{M}_1 \circ \boldsymbol{M}_2\|_{op} \le \sqrt{n}\,\|\boldsymbol{M}_1\|_{op}\,\|\boldsymbol{M}_2\|_{max}\,. \tag{31}$$
4. *For any square matrix* $\boldsymbol{M}$, *we have* $\left\|(\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I})\circ\boldsymbol{M}\right\|_{op} \le 2\,\|\boldsymbol{M}\|_{op}$.
5. *For any square matrix* $\boldsymbol{M} \in \mathbb{R}^{n\times n}$, *we have* $\|\boldsymbol{M}\|_{op} \le n\,\|\boldsymbol{M}\|_{max}$.

Proof. 1. This follows directly from the variational characterization of the operator norm, i.e.,

$$\left\|\bar{\boldsymbol{B}}\right\|_{op} = \sup_{\|\boldsymbol{u}\|\le 1}\left\|\bar{\boldsymbol{B}}\boldsymbol{u}\right\| = \sup_{\|\boldsymbol{u}\|\le 1}\|\boldsymbol{B}(\boldsymbol{m}\circ\boldsymbol{u})\| \le \sup_{\|\boldsymbol{u}\|\le 1}\|\boldsymbol{B}\boldsymbol{u}\|\,,$$

where the last step follows from $\|\boldsymbol{m}\circ\boldsymbol{u}\| \le \|\boldsymbol{u}\|$.

2. For $\|\boldsymbol{v}\| = 1$, we have

$$\|\boldsymbol{M}\boldsymbol{v}\| = \sqrt{\sum_{k=1}^n \langle \boldsymbol{M}_{k,:}, \boldsymbol{v}\rangle^2} \le \sqrt{\sum_{k=1}^n \|\boldsymbol{M}_{k,:}\|^2} \le \sqrt{n\max_k \|\boldsymbol{M}_{k,:}\|^2} = \sqrt{n}\max_k \|\boldsymbol{M}_{k,:}\|\,.$$

3. For a unit vector $\boldsymbol{e}_i$, we have

$$\|\boldsymbol{M}_1\circ\boldsymbol{M}_2\boldsymbol{e}_i\| \le \|\boldsymbol{M}_2\|_{max}\,\|\boldsymbol{M}_1\boldsymbol{e}_i\| \le \|\boldsymbol{M}_2\|_{max}\,\|\boldsymbol{M}_1\|_{op}\,.$$

For a general vector $\boldsymbol{v}$, we can use the triangle inequality to obtain

$$\|\boldsymbol{M}_1\circ\boldsymbol{M}_2\boldsymbol{v}\| \le \sum_i |v_i|\,\|\boldsymbol{M}_2\|_{max}\,\|\boldsymbol{M}_1\boldsymbol{e}_i\| \le \|\boldsymbol{M}_2\|_{max}\,\|\boldsymbol{M}_1\|_{op}\sum_i |v_i|\,.$$

By using

$$\sum_i |v_i| \le \sqrt{n}\,\|\boldsymbol{v}\|\,,$$

we obtain the desired bound.

4. Note that $(\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I})\circ\boldsymbol{M} = \boldsymbol{M} - \mathrm{Diag}\,(\boldsymbol{M})$, so

$$\left\|(\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I})\circ\boldsymbol{M}\right\|_{op} \le \|\boldsymbol{M}\|_{op} + \|\mathrm{Diag}\,(\boldsymbol{M})\|_{op} \le 2\,\|\boldsymbol{M}\|_{op}\,,$$

where we have used that

$$\|\mathrm{Diag}\,(\boldsymbol{M})\|_{op} = \max_i |\boldsymbol{M}_{i,i}| \le \|\boldsymbol{M}\|_{op}\,.$$

5. Note that $\|\boldsymbol{M}_{k,:}\| \le \sqrt{n}\,\|\boldsymbol{M}\|_{max}$. Thus, the result follows from the point 2. above.

□

Lemma A.5. *Denote by c_ℓ^2 the ℓ-th Taylor coefficient of the function $\arcsin(x)$. Then, for $x \in [0, 1)$, $\ell_0 \in \mathbb{N}_+$, we have*

$$\sum_{\ell=\ell_0}^{\infty} \ell c_\ell^2 x^{\ell-1} \leq C(\ell_0) x^{\ell_0 - 1} \frac{1}{\sqrt{1-x}}. \tag{32}$$

Proof. Recall that

$$\frac{d}{dx} \arcsin(x) = \frac{1}{\sqrt{1-x}} = \sum_{\ell=1}^{\infty} \ell c_\ell^2 x^{\ell-1},$$

with

$$c_{2k} = 0, \qquad c_{2k+1}^2 = \frac{(2k)!}{4^k (k!)^2 (2k+1)}.$$

By Stirling's approximation we have

$$c_{2k+1}^2 = \Theta\left(\frac{1}{k^{\frac{3}{2}}}\right),$$

which implies that, for odd $\ell, \ell_0 > 0$

$$\frac{\ell c_\ell^2}{(\ell + \ell_0 - 1) c_{\ell+\ell_0-1}^2} \leq C \frac{\ell (\ell + \ell_0 - 1)^{\frac{3}{2}}}{(\ell + \ell_0 - 1) \ell^{\frac{3}{2}}} \leq C(\ell_0).$$

Thus we have

$$\begin{aligned} \sum_{\ell=\ell_0}^{\infty} \ell c_\ell^2 x^{\ell-1} = x^{\ell_0-1} \sum_{\ell=\ell_0}^{\infty} \ell c_\ell^2 x^{\ell-\ell_0} &= x^{\ell_0-1} \sum_{\ell=1}^{\infty} (\ell + \ell_0 - 1) c_{\ell+\ell_0-1}^2 x^{\ell-1} \\ &\leq x^{\ell_0-1} \sum_{\ell=1}^{\infty} C(\ell_0) \ell c_\ell^2 x^{\ell-1} = C(\ell_0) x^{\ell_0-1} \frac{1}{\sqrt{1-x}}, \end{aligned}$$

which finishes the proof. □

Lemma A.6. *Assume that $\boldsymbol{B} = \boldsymbol{X} + \boldsymbol{R}$ with $\|\boldsymbol{X}\|_{op} \leq C_X$, $\|\boldsymbol{X}\|_{max} \leq C_X \frac{\log(d)}{\sqrt{d}}$, $\|\boldsymbol{R}\|_{op} \leq C_R \frac{\log^{\alpha_R}(d)}{\sqrt{(d)}}$. Then, for large enough d, we have*

$$\|\boldsymbol{b}_i\|_4^2 \leq C C_X^2 \frac{\log(d)}{\sqrt{d}}. \tag{33}$$

Proof. By Hölder, we have

$$\|\boldsymbol{r}_i\|_4 \leq \|\boldsymbol{r}_i\|_2^{\frac{1}{2}} \|\boldsymbol{r}_i\|_\infty^{\frac{1}{2}} \leq \|\boldsymbol{R}\|_{op}^{\frac{1}{2}} \|\boldsymbol{R}\|_{op}^{\frac{1}{2}} = \|\boldsymbol{R}\|_{op},$$

and

$$\|\boldsymbol{x}_i\|_4 \leq \|\boldsymbol{x}_i\|_2^{\frac{1}{2}} \|\boldsymbol{x}_i\|_\infty^{\frac{1}{2}} \leq \|\boldsymbol{X}\|_{op}^{\frac{1}{2}} \|\boldsymbol{X}\|_{max}^{\frac{1}{2}},$$

so

$$\|\boldsymbol{b}_i\|_4^2 \leq (\|\boldsymbol{x}_i\|_4 + \|\boldsymbol{r}_i\|_4)^2 \leq C(\|\boldsymbol{x}_i\|_4^2 + \|\boldsymbol{r}_i\|_4^2) \leq C C_X^2 \frac{\log(d)}{\sqrt{d}} + C C_R^2 \frac{(\log^{\alpha_R}(d))^2}{d},$$

which implies (33). □

Lemma A.7 (Concentration of $\hat{\boldsymbol{D}}$). *For $\boldsymbol{b} \in \mathbb{R}^d$, $\|\boldsymbol{b}\| = 1$ and $\boldsymbol{m} \sim \mathrm{Bern}(p)$ i.i.d., we have*

$$\mathbb{P}\left(\left|\|\boldsymbol{m} \circ \boldsymbol{b}\|_2^2 - p \|\boldsymbol{b}\|_2^2\right| > \lambda\right) \leq C \exp\left(-c \frac{\lambda^2}{\|\boldsymbol{b}\|_4^4}\right), \tag{34}$$

which implies

$$\mathbb{P}\left(\left|\|\boldsymbol{m} \circ \boldsymbol{b}\|_2 - \sqrt{p} \|\boldsymbol{b}\|_2\right| > \lambda\right) \leq C \exp\left(-c \frac{\lambda^2}{\|\boldsymbol{b}\|_4^4}\right). \tag{35}$$

Proof. Equation (34) is an immediate consequence of Hoeffding's inequality applied to the random variables $(\boldsymbol{m}_i - p)b_i^2$ (cf. Theorem 2.6.2 in (Vershynin, 2018)).

To obtain (35) we note that $\left|\|\boldsymbol{m} \circ \boldsymbol{b}\|_2 - \sqrt{p}\,\|\boldsymbol{b}\|_2\right| > \lambda$ implies

$$\left|\|\boldsymbol{m} \circ \boldsymbol{b}\|_2^2 - p\,\|\boldsymbol{b}\|_2^2\right| = |\|\boldsymbol{m} \circ \boldsymbol{b}\|_2 - \sqrt{p}\,\|\boldsymbol{b}\|_2|\,|\|\boldsymbol{m} \circ \boldsymbol{b}\|_2 + \sqrt{p}\,\|\boldsymbol{b}\|_2| > \sqrt{p}\,\|\boldsymbol{b}\|_2\,\lambda.$$

Since by assumption $\|\boldsymbol{b}\| = 1$, the above implies (35). □

Lemma A.8. *Let $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ be a Haar matrix and $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ an independent random matrix with $\|\boldsymbol{M}\|_{op} \le C_M$. Then, for $\boldsymbol{Y} := \boldsymbol{U}\boldsymbol{M}\boldsymbol{U}^\top$, we have*

$$\mathbb{P}\left(\left\|\boldsymbol{Y} - \frac{1}{n}\mathrm{Tr}\left[\boldsymbol{Y}\right]\boldsymbol{I}\right\|_{max} > \lambda\right) \le Cd^2\mathrm{exp}\left(-\frac{c}{C_M^2}d\lambda^2\right). \tag{36}$$

If instead we have $\boldsymbol{M} \in \mathbb{R}^{n\times d}$ with $\frac{n}{d} \le C$ and $\boldsymbol{Y} := \boldsymbol{U}\boldsymbol{M}$, then

$$\mathbb{P}\left(\|\boldsymbol{Y}\|_{max} > \lambda\right) \le Cd^2\mathrm{exp}\left(-\frac{c}{C_M^2}d\lambda^2\right). \tag{37}$$

Proof. We first fix $\boldsymbol{M}$, and note that since $\boldsymbol{M}$ and $\boldsymbol{U}$ are independent, the distribution of $\boldsymbol{U}$ does not change if we condition on $\boldsymbol{M}$. For both inequalities, for fixed $\boldsymbol{M}$, the map $(SO_n, \|\cdot\|_F) \to (\mathcal{M}_{n\times d}, \|\cdot\|_{op})$, $\boldsymbol{U} \to \boldsymbol{Y}$ is Lipschitz as it is a bounded (bi-)linear form on a bounded set. The composition with the projection on the (i,j)-th component of a matrix is also Lipschitz, so we can apply Theorem 5.2.7 in (Vershynin, 2018) to obtain that $\boldsymbol{Y}_{i,j}$ is subgaussian in $\boldsymbol{U}$ with subgaussian norm $\frac{CC_M}{\sqrt{d}}$. Formally this means that

$$\mathbb{P}_{\boldsymbol{U}}\left(|\boldsymbol{Y}_{i,j} - \mathbb{E}\boldsymbol{Y}_{i,j}| > \lambda|\boldsymbol{M}\right) \le C\mathrm{exp}\left(-\frac{c}{C_M^2}d\lambda^2\right).$$

Since the RHS is independent of $\boldsymbol{M}$ (i.e., it only depends on C_M) we have

$$\begin{aligned}\mathbb{P}\left(|\boldsymbol{Y}_{i,j} - \mathbb{E}\boldsymbol{Y}_{i,j}| > \lambda\right) &= \mathbb{P}_{\boldsymbol{M}}\left(\mathbb{P}_{\boldsymbol{U},\boldsymbol{V}}\left(|\boldsymbol{Y}_{i,j} - \mathbb{E}\boldsymbol{Y}_{i,j}| > \lambda|\boldsymbol{M}\right)\right)\\ &\le \mathbb{P}_{\boldsymbol{M}}\left(C\mathrm{exp}\left(-\frac{c}{C_M^2 d}\lambda^2\right)\right)\\ &= C\mathrm{exp}\left(-\frac{c}{C_M^2 d}\lambda^2\right).\end{aligned}$$

Now, (36) follows by noting that $\mathbb{E}\boldsymbol{U}\boldsymbol{M}\boldsymbol{U}^\top = \frac{1}{n}\mathrm{Tr}\left[\boldsymbol{U}\boldsymbol{M}\boldsymbol{U}^\top\right]\boldsymbol{I}$ and using a simple union bound over (i,j). The proof of (37) is the same, with the only difference being $\mathbb{E}\boldsymbol{Y}_{i,j} = 0$. □

Lemma A.9. *Let $\boldsymbol{U} \in \mathbb{R}^{n\times n}, \boldsymbol{V} \in \mathbb{R}^{d\times d}$ be Haar matrices, and $\boldsymbol{S}_1, \boldsymbol{S}_2^\top \in \mathbb{R}^{n\times d}$ be deterministic diagonal matrices. Define $\bar{\boldsymbol{M}} = \boldsymbol{S}_1(\boldsymbol{V}^\top)_{\boldsymbol{m}}(\boldsymbol{V}^\top)_{\boldsymbol{m}}^\top\boldsymbol{S}_2, \bar{\boldsymbol{Y}} = \boldsymbol{U}\bar{\boldsymbol{M}}\boldsymbol{U}^\top, \boldsymbol{Y} = p\boldsymbol{U}\boldsymbol{S}_1\boldsymbol{S}_2\boldsymbol{U}^\top$, $C_M := \|\boldsymbol{S}_1\|_{op}\|\boldsymbol{S}_2\|_{op}$. Then, for any $\gamma > 0$ and $d > d_0(\gamma)$, we have with probability at least $1 - C/d^2$ (in $\boldsymbol{U}, \boldsymbol{V}$)*

$$\mathbb{P}_{\boldsymbol{m}}\left(\left\|\bar{\boldsymbol{Y}} - \frac{1}{n}\mathrm{Tr}\left[\boldsymbol{Y}\right]\boldsymbol{I}\right\|_{max} > CC_M\frac{\log^3(d)}{\sqrt{d}}\middle|\boldsymbol{U},\boldsymbol{V}\right) \le C\frac{1}{d^\gamma}. \tag{38}$$

Proof. In the first step we will show that with probability at least $1 - C/d^2$ in $\boldsymbol{U}, \boldsymbol{V}$

$$\mathbb{P}_{\boldsymbol{m}}\left(\left\|\bar{\boldsymbol{Y}} - \frac{1}{n}\mathrm{Tr}\left[\bar{\boldsymbol{Y}}\right]\boldsymbol{I}\right\|_{max} > C_M\frac{\log^3(d)}{\sqrt{d}}\middle|\boldsymbol{U},\boldsymbol{V}\right) \le C\frac{1}{d^\gamma}. \tag{39}$$

The key observation is that

$$\begin{aligned}\underset{\boldsymbol{U},\boldsymbol{V}}{\mathbb{P}}\left(\underset{\boldsymbol{m}}{\mathbb{P}}\left(\left\|\bar{\boldsymbol{Y}}-\frac{1}{n}\text{Tr}\left[\bar{\boldsymbol{Y}}\right]\boldsymbol{I}\right\|_{max}>C_M\lambda\middle|\boldsymbol{U},\boldsymbol{V}\right)>\alpha\right)&\le\frac{1}{\alpha}\mathbb{E}_{\boldsymbol{U},\boldsymbol{V}}\underset{\boldsymbol{m}}{\mathbb{P}}\left(\left\|\bar{\boldsymbol{Y}}-\frac{1}{n}\text{Tr}\left[\bar{\boldsymbol{Y}}\right]\boldsymbol{I}\right\|_{max}>C_M\lambda\middle|\boldsymbol{U},\boldsymbol{V}\right)\\&=\frac{1}{\alpha}\mathbb{P}\left(\left\|\bar{\boldsymbol{Y}}-\frac{1}{n}\text{Tr}\left[\bar{\boldsymbol{Y}}\right]\boldsymbol{I}\right\|_{max}>C_M\lambda\right)\\&\le C\frac{1}{\alpha}d^2\text{exp}\left(-\frac{c}{C_M^2}dC_M^2\lambda^2\right)\\&=C\frac{1}{\alpha}d^2\text{exp}\left(-cd\lambda^2\right),\end{aligned}$$

where the first passage follows from Markov's inequality and the last inequality is due to Lemma A.8 and $C_M=\|\boldsymbol{S}_1\|_{op}\|\boldsymbol{S}_2\|_{op}\ge\|\bar{\boldsymbol{M}}\|_{op}$.

By choosing $\alpha=\frac{1}{d^\gamma}$ and $\lambda=\frac{\log^3(d)}{\sqrt{d}}$, we obtain that, with probability at least $1-C/d^2$ in $\boldsymbol{U},\boldsymbol{V}$,

$$\underset{\boldsymbol{m}}{\mathbb{P}}\left(\left\|\bar{\boldsymbol{Y}}-\frac{1}{n}\text{Tr}\left[\bar{\boldsymbol{Y}}\right]\boldsymbol{I}\right\|_{max}>C_M\frac{\log^3(d)}{\sqrt{d}}\middle|\boldsymbol{U},\boldsymbol{V}\right)\le\frac{1}{d^\gamma}. \tag{40}$$

Now, in the second step, we will show that, with probability at least $1-C/d^2$ over $\boldsymbol{V}$,

$$\underset{\boldsymbol{m}}{\mathbb{P}}\left(\left\|\frac{1}{n}\text{Tr}\left[\bar{\boldsymbol{Y}}\right]\boldsymbol{I}-\frac{1}{n}\text{Tr}\left[\boldsymbol{Y}\right]\boldsymbol{I}\right\|_{max}>C_M\frac{\log^3(d)}{\sqrt{d}}\middle|\boldsymbol{U},\boldsymbol{V}\right)\le\frac{1}{d^\gamma}. \tag{41}$$

First, note that by Lemma A.8 with probability at least $1-C/d^2$ (in $\boldsymbol{V}$) we have that $\|\boldsymbol{V}\|_{max}\le C\frac{\log(d)}{\sqrt{d}}$, so $\|\boldsymbol{V}_{:,i}\|_4^4\le C\frac{\log^4(d)}{d}$. By Lemma A.7, we have that

$$\mathbb{P}\left(\left|\left((\boldsymbol{V}^\top)_{\boldsymbol{m}}(\boldsymbol{V}^\top)_{\boldsymbol{m}}^\top\right)_{i,i}-p\right|>\lambda\right)\le C\text{exp}\left(-c\frac{d\lambda^2}{\log^4(d)}\right).$$

Choosing $\lambda=\frac{\log^3(d)}{\sqrt{d}}$ we obtain for large d

$$\mathbb{P}\left(\left|\left((\boldsymbol{V}^\top)_{\boldsymbol{m}}(\boldsymbol{V}^\top)_{\boldsymbol{m}}^\top\right)_{i,i}-p\right|>\frac{\log^3(d)}{\sqrt{d}}\right)\le C\frac{1}{d^{\gamma+1}}.$$

By a simple union bound we obtain

$$\mathbb{P}\left(\left\|\text{Diag}\left((\boldsymbol{V}^\top)_{\boldsymbol{m}}(\boldsymbol{V}^\top)_{\boldsymbol{m}}^\top\right)-p\boldsymbol{I}\right\|_{op}>\frac{\log^3(d)}{\sqrt{d}}\right)\le C\frac{1}{d^\gamma}.$$

Note that since $\boldsymbol{S}_1,\boldsymbol{S}_2$ are diagonal we have

$$\text{Tr}\left[\boldsymbol{S}_1(\boldsymbol{V}^\top)_{\boldsymbol{m}}(\boldsymbol{V}^\top)_{\boldsymbol{m}}^\top\boldsymbol{S}_2\right]=\text{Tr}\left[\boldsymbol{S}_1\text{Diag}\left((\boldsymbol{V}^\top)_{\boldsymbol{m}}(\boldsymbol{V}^\top)_{\boldsymbol{m}}^\top\right)\boldsymbol{S}_2\right],$$

so $\left\|\text{Diag}\left((\boldsymbol{V}^\top)_{\boldsymbol{m}}(\boldsymbol{V}^\top)_{\boldsymbol{m}}^\top\right)-p\boldsymbol{I}\right\|_{op}\le\frac{\log^3(d)}{\sqrt{d}}$ implies

$$\begin{aligned}\left|\text{Tr}\left[\boldsymbol{S}_1(\boldsymbol{V}^\top)_{\boldsymbol{m}}(\boldsymbol{V}^\top)_{\boldsymbol{m}}^\top\boldsymbol{S}_2\right]-p\text{Tr}\left[\boldsymbol{S}_1\boldsymbol{S}_2\right]\right|&=\left|\text{Tr}\left[\boldsymbol{S}_1\left(\text{Diag}\left((\boldsymbol{V}^\top)_{\boldsymbol{m}}(\boldsymbol{V}^\top)_{\boldsymbol{m}}^\top\right)-p\boldsymbol{I}\right)\boldsymbol{S}_2\right]\right|\\&\le n\left\|\boldsymbol{S}_1\left(\text{Diag}\left((\boldsymbol{V}^\top)_{\boldsymbol{m}}(\boldsymbol{V}^\top)_{\boldsymbol{m}}^\top\right)-p\boldsymbol{I}\right)\boldsymbol{S}_2\right\|_{op}\\&\le C_Mn\frac{\log^3(d)}{\sqrt{d}}.\end{aligned}$$

Thus,

$$\mathbb{P}\left(\left|\frac{1}{n}\mathrm{Tr}\left[\boldsymbol{S}_1(\boldsymbol{V}^\top)_{\boldsymbol{m}}(\boldsymbol{V}^\top)_{\boldsymbol{m}}^\top\boldsymbol{S}_2\right]-p\frac{1}{n}\mathrm{Tr}\left[\boldsymbol{S}_1\boldsymbol{S}_2\right]\right|>C_M\frac{\log^3(d)}{\sqrt{d}}\right)\le C\frac{1}{d^\gamma}.$$

Noting that, by definition, $\mathrm{Tr}\left[\bar{\boldsymbol{Y}}\right]=\mathrm{Tr}\left[\boldsymbol{S}_1(\boldsymbol{V}^\top)_{\boldsymbol{m}}(\boldsymbol{V}^\top)_{\boldsymbol{m}}^\top\boldsymbol{S}_2\right]$ and $\mathrm{Tr}\left[\boldsymbol{Y}\right]=p\mathrm{Tr}\left[\boldsymbol{S}_1\boldsymbol{S}_2\right]$, we obtain (41).

Finally, combining (40) and (41) finishes the proof. □

Lemma A.10. *Let $\boldsymbol{B}\in\mathbb{R}^{n\times d}$ be an arbitrary matrix, $\boldsymbol{G}\in\mathbb{R}^{n\times d}$ with $\boldsymbol{G}_{ij}\sim\mathcal{N}(0,\sigma^2)$, and assume $n=O(d)$. Then, for any $\delta>0$, we have*

$$\mathbb{P}_{\boldsymbol{G}}\left(\min_{i,j}|\boldsymbol{B}_{ij}+\boldsymbol{G}_{ij}|\le\delta\right)\le Cd^2\frac{\delta}{\sigma}.$$

Proof. By the scale invariance of the problem, we may assume that $\sigma=1$. Let g be a standard normal variable and $b\in\mathbb{R}$. Then,

$$\mathbb{P}\left(|b+g|\le\delta\right)=\mathbb{P}\left(g\in[b-\delta,b+\delta]\right)\le C\delta,$$

where the second step holds since the pdf of g is bounded by a universal constant. The result of the lemma now follows by a simple union bound over all (i,j) (and using $n=O(d)$). □

A.4. Concentration tools

In this section, we provide the matrix concentration results needed for the proof. We recall that we use the shorthand notation $\bar{\boldsymbol{B}}=\boldsymbol{B}_{\boldsymbol{m}}$ and $\hat{\boldsymbol{B}}=\hat{\boldsymbol{B}}_{\boldsymbol{m}}=\hat{\boldsymbol{D}}\bar{\boldsymbol{B}}_{\boldsymbol{m}}, \hat{\boldsymbol{D}}_{i,i}=1/\left\|\bar{\boldsymbol{b}}_i\right\|$ *only* for the matrix $\boldsymbol{B}$. Here, the masking $\boldsymbol{B}_{\boldsymbol{m}}$ was defined as $(\boldsymbol{B}_{\boldsymbol{m}})_{i,j}=\boldsymbol{B}_{i,j}\boldsymbol{m}_j$.

Lemma A.11. *Let $\boldsymbol{B}=\boldsymbol{X}+\boldsymbol{R}=\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top+\boldsymbol{R}$, $\boldsymbol{A}=\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top+\alpha\boldsymbol{I})^{-1}+O\left(\boldsymbol{R}\right)+O\left(C_X^7\frac{\log^{10}(d)}{\sqrt{d}}\right)$, where $\boldsymbol{U},\boldsymbol{V}$ are Haar matrices, $\left\|\boldsymbol{R}\right\|_{op}=o(1)$, $\boldsymbol{S}$ is a diagonal matrix s.t. $\left\|\boldsymbol{S}\right\|_{op}\le C_X$, $\frac{1}{n}\mathrm{Tr}\left[\boldsymbol{S}\boldsymbol{S}^\top\right]=1$, and $\alpha>0$ fixed. Then, for any $\gamma>0,d>d_0(\gamma)$, with probability at least $1-C/d^2$ in $\boldsymbol{U},\boldsymbol{V}$ and at least $1-C/d^\gamma$ in $\boldsymbol{m}$,*

1. $\left\|\boldsymbol{B}\right\|_{max}\le C_X\frac{\log(d)}{\sqrt{d}}+O\left(\left\|\boldsymbol{R}\right\|_{op}\right)$.
2. $\mathrm{Diag}\left(\left(\boldsymbol{B}\boldsymbol{B}^\top+\alpha\boldsymbol{I}\right)^{-2}\boldsymbol{B}\boldsymbol{B}^\top\right)=\frac{1}{n}\mathrm{Tr}\left[\left(\boldsymbol{B}\boldsymbol{B}^\top+\alpha\boldsymbol{I}\right)^{-2}\boldsymbol{B}\boldsymbol{B}^\top\right]\boldsymbol{I}+O\left(\frac{\log(d)}{\sqrt{d}}\right)$.
3. $\left\|\boldsymbol{A}^\top\boldsymbol{A}-\frac{1}{n}\mathrm{Tr}\left[\left(\boldsymbol{X}\boldsymbol{X}^\top+\alpha\boldsymbol{I}\right)^{-2}\boldsymbol{X}\boldsymbol{X}^\top\right]\boldsymbol{I}\right\|_{max}\le C_X^7\frac{\log^{10}(d)}{\sqrt{d}}+O\left(\left\|\boldsymbol{R}\right\|_{op}\right)$.
4. $\frac{1}{p}\bar{\boldsymbol{B}}\bar{\boldsymbol{B}}^\top-\frac{1}{n}\mathrm{Tr}\left[\boldsymbol{B}\boldsymbol{B}^\top\right]\boldsymbol{I}=O_{max}(CC_X^2\frac{\log^3(d)}{\sqrt{d}})+O\left(C_X\left\|\boldsymbol{R}\right\|_{op}\right)$.
5. $\mathrm{Diag}\left(\frac{1}{p}\bar{\boldsymbol{B}}\boldsymbol{A}\right)=\frac{1}{n}\mathrm{Tr}\left[\boldsymbol{B}\boldsymbol{A}\right]\boldsymbol{I}+O\left(C_X^8\frac{\log^{10}(d)}{\sqrt{d}}\right)+O\left(C_X\boldsymbol{R}\right)$.
6. $\mathrm{Diag}\left(\frac{1}{p}\boldsymbol{A}^\top\boldsymbol{A}\bar{\boldsymbol{B}}\bar{\boldsymbol{B}}^\top\right)=\frac{1}{n}\mathrm{Tr}\left[\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{B}\boldsymbol{B}^\top\right]\boldsymbol{I}+O\left(C_X^9\frac{\log^{10}(d)}{\sqrt{d}}\right)+O\left(C_X^2\boldsymbol{R}\right)$.

Proof. The $O\left(\boldsymbol{R}\right),O\left(\left\|\boldsymbol{R}\right\|_{op}\right)$ terms are always extracted by using that the LHS is a continuous function in $\boldsymbol{R}$ (w.r.t. $\left\|\cdot\right\|_{op}$). We carry this out explicitly for the first item and skip the details for the other items.

1. By Lemma A.8, with probability at least $1-C/d^\gamma$,

$$\left\|\boldsymbol{B}\right\|_{max}=\left\|\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top+\boldsymbol{R}\right\|_{max}\le\left\|\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top\right\|_{max}+\left\|\boldsymbol{R}\right\|_{max}\le C_X\frac{\log(d)}{\sqrt{d}}+\left\|\boldsymbol{R}\right\|_{op}=C_X\frac{\log(d)}{\sqrt{d}}+O\left(\left\|\boldsymbol{R}\right\|_{op}\right),$$

where we have used (37) with $\lambda=C_X\frac{\log(d)}{\sqrt{d}}$ in the third step.

2. This follows from (36) with $\lambda = \frac{\log(d)}{\sqrt{d}}$ by noting that for any matrix $\boldsymbol{B}$ we have $\left\| \left(\boldsymbol{B}\boldsymbol{B}^\top + \alpha \boldsymbol{I} \right)^{-2} \boldsymbol{B}\boldsymbol{B}^\top \right\|_{op} \leq \frac{1}{\alpha}$.

3. As in the previous item, we have $\left\| \left(\boldsymbol{X}\boldsymbol{X}^\top + \alpha \boldsymbol{I} \right)^{-2} \boldsymbol{X}\boldsymbol{X}^\top \right\|_{op} \leq \frac{1}{\alpha}$ so the result follows again from (36) with $\lambda = \frac{\log(d)}{\sqrt{d}}$.

4. First note that by 1. in Lemma A.4 we have $\left\| \bar{\boldsymbol{B}} \right\|_{op} \leq \left\| \boldsymbol{B} \right\|_{op} \leq CC_X$, where we have used the assumption $O\left(\left\| \boldsymbol{R} \right\|_{op} \right) = o(1)$. Thus, we have

$$\bar{\boldsymbol{B}}\bar{\boldsymbol{B}}^\top = \boldsymbol{U}\boldsymbol{S}\left(\boldsymbol{V}^\top \right)_{\boldsymbol{m}} \left(\boldsymbol{V}^\top \right)_{\boldsymbol{m}}^\top \boldsymbol{S}\boldsymbol{U}^\top + O\left(C_X \boldsymbol{R} \right),$$

so the result follows from Lemma A.9.

5. Note that $\left\| \boldsymbol{X}^\top \left(\boldsymbol{X}\boldsymbol{X}^\top + \alpha \boldsymbol{I} \right)^{-2} \right\|_{op} \leq \frac{1}{2\sqrt{\alpha}}$ and by Lemma A.4 we have $\left\| \bar{\boldsymbol{B}} \right\|_{op} \leq CC_X$. Thus, we have

$$\bar{\boldsymbol{B}}\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\left(\boldsymbol{V}^\top \right)_{\boldsymbol{m}} \left(\boldsymbol{V}^\top \right)_{\boldsymbol{m}}^\top \tilde{\boldsymbol{S}}\boldsymbol{U}^\top + O\left(C_X \boldsymbol{R} \right),$$

where $\tilde{\boldsymbol{S}}$ is a diagonal matrix, so the result follows from Lemma A.9.

6. By using again that $\left\| \boldsymbol{X}^\top \left(\boldsymbol{X}\boldsymbol{X}^\top + \alpha \boldsymbol{I} \right)^{-2} \right\|_{op} \leq \frac{1}{2\sqrt{\alpha}}$ and $\left\| \bar{\boldsymbol{B}} \right\|_{op} \leq CC_X$, we have

$$\boldsymbol{A}^\top \boldsymbol{A}\bar{\boldsymbol{B}}\bar{\boldsymbol{B}}^\top = \boldsymbol{U}\tilde{\boldsymbol{S}}^2\boldsymbol{S}\left(\boldsymbol{V}^\top \right)_{\boldsymbol{m}} \left(\boldsymbol{V}^\top \right)_{\boldsymbol{m}}^\top \boldsymbol{S}\boldsymbol{U}^\top + O\left(C_X^2 \boldsymbol{R} \right),$$

where $\tilde{\boldsymbol{S}}$ is a diagonal matrix, so the result follows from Lemma A.9.

□

Lemma A.12 (Master concentration for $\hat{\boldsymbol{B}}$)**.** *Consider a fixed $\boldsymbol{B} = \boldsymbol{X} + \boldsymbol{R}$ with unit norm rows and $\left\| \boldsymbol{X} \right\|_{op} \leq C_X$, $\left\| \boldsymbol{X} \right\|_{max} \leq C_X \frac{\log(d)}{\sqrt{d}}$, $\left\| \boldsymbol{R} \right\|_{op} \leq C_R \frac{\log^{\alpha_R}(d)}{\sqrt{d}}$. Let $2C_X \frac{1}{\sqrt{p}} > C_b > (1+c)C_X \frac{1}{\sqrt{p}} > 0$, for some small constant $c > 0$. Let $F : \mathcal{B}_{\sqrt{d}}(\boldsymbol{0}) \cup (\Omega \cap \mathcal{B}_{C_b}(\boldsymbol{0})) \subset \mathcal{M}_{n \times d} \to \mathcal{M}_{\tilde{n} \times \tilde{d}}$, for arbitrary $\tilde{n}, \tilde{d}$. Assume that Ω is s.t. with probability at least $1 - Cd^{-k_F/2 - 1/2}$ in $\boldsymbol{m}$ we have for $\hat{\boldsymbol{D}}^b_{i,i} = \min\{\frac{1}{\left\| \bar{\boldsymbol{b}}_i \right\|}, \frac{C_b}{C_X}\}$ that $\hat{\boldsymbol{D}}^b \bar{\boldsymbol{B}}, \frac{1}{\sqrt{p}}\bar{\boldsymbol{B}} \in \Omega$. Further assume that F satisfies the following properties:*

1. *$\left\| F(\boldsymbol{M}) \right\|_{op} \leq C_F d^{\frac{k_F}{2}}$ for every $\boldsymbol{M} = \hat{\boldsymbol{B}}$;*

2. *$\left\| F(\boldsymbol{M}) \right\|_{op} \leq C_F$ for every $\boldsymbol{M} \in \Omega \cap \mathcal{B}_{C_b}(\boldsymbol{0})$;*

3. *F is Lipschitz with constant $C_{F'}$ on $\Omega \cap \mathcal{B}_{C_b}(\boldsymbol{0})$ (w.r.t. $\left\| \cdot \right\|_{op}$ on both spaces).*

Then, for large enough $d > d_0(C_F, C_{F'}, C_b, C_R)$,

$$\left\| \mathbb{E}_{\boldsymbol{m}} F(\hat{\boldsymbol{B}}) - \mathbb{E}_{\boldsymbol{m}} \mathbb{1}_{\{\hat{\boldsymbol{D}}^b \bar{\boldsymbol{B}}, \frac{1}{\sqrt{p}}\bar{\boldsymbol{B}} \in \Omega\}} F(\frac{1}{\sqrt{p}}\bar{\boldsymbol{B}}) \right\|_{op} \leq CC_{F'} C_X^3 \frac{\log^3(d)}{\sqrt{d}}, \tag{42}$$

where crucially the RHS is independent of C_R.

Proof. Define $\mathbb{1}_{\bar{\Omega}} := \mathbb{1}_{\{\hat{\boldsymbol{D}}^b\bar{\boldsymbol{B}}, \frac{1}{\sqrt{p}}\bar{\boldsymbol{B}}\in\Omega\}}$ and $\mathbb{1}_{\bar{\Omega}^c} := 1 - \mathbb{1}_{\bar{\Omega}}$. We will actually show the slightly stronger statement

$$\mathbb{E}_{\boldsymbol{m}} \left\| F(\hat{\boldsymbol{B}}) - \mathbb{1}_{\bar{\Omega}} F\left(\frac{1}{\sqrt{p}}\bar{\boldsymbol{B}}\right)\right\|_{op} \leq C C_{F'} C_X^3 \frac{\log^3(d)}{\sqrt{d}}, \tag{43}$$

which immediately implies (42). First, by a simple triangle estimate we have

$$\mathbb{E}_{\boldsymbol{m}} \left\| F(\hat{\boldsymbol{B}}) - \mathbb{1}_{\bar{\Omega}} F\left(\frac{1}{\sqrt{p}}\bar{\boldsymbol{B}}\right)\right\|_{op} \leq \mathbb{E}_{\boldsymbol{m}} \left\| \mathbb{1}_{\bar{\Omega}} \left(F(\hat{\boldsymbol{B}}) - F\left(\frac{1}{\sqrt{p}}\bar{\boldsymbol{B}}\right)\right)\right\|_{op} + \mathbb{E}_{\boldsymbol{m}} \left\| \mathbb{1}_{\bar{\Omega}^c} F(\hat{\boldsymbol{B}})\right\|_{op}.$$

By our assumptions on Ω and assumption 1. we have

$$\mathbb{E}_{\boldsymbol{m}} \left\| \mathbb{1}_{\bar{\Omega}^c} F(\hat{\boldsymbol{B}})\right\|_{op} \leq C C_F d^{k_F/2} d^{-k_F/2-1/2} = 2C_F \frac{1}{\sqrt{d}},$$

so w.l.o.g. we may assume that $\mathbb{1}_{\bar{\Omega}} \equiv 1$.

We now show that we can truncate $\hat{\boldsymbol{D}}$ to $\hat{\boldsymbol{D}}^b$ by applying the truncation function $\min\{x, \frac{C_b}{C_X}\}$ to each entry. Note that by definition of C_b we have $\frac{1+c}{\sqrt{p}} \leq \frac{C_b}{C_X} \leq \frac{2}{\sqrt{p}}$. By 2. in Lemma A.4, we have that $\left\|\hat{\boldsymbol{B}}\right\|_{op}, \left\|\hat{\boldsymbol{D}}^b\bar{\boldsymbol{B}}\right\|_{op} \leq \sqrt{n}$. Thus, by assumption, we obtain

$$\mathbb{E}_{\boldsymbol{m}} \left\| F(\hat{\boldsymbol{B}}) - F(\hat{\boldsymbol{D}}^b\bar{\boldsymbol{B}})\right\|_{op} \leq C_F d^{k_F/2} \mathbb{P}\left(\left\|\hat{\boldsymbol{D}}\right\|_{op} > \frac{C_b}{C_X}\right). \tag{44}$$

We also have the trivial bound

$$\|\boldsymbol{b}_i\|_4^4 \leq d\left(C_X \frac{\log(d)}{\sqrt{d}} + C_R \frac{\log^{\alpha_R}(d)}{\sqrt{d}}\right)^4 \leq \frac{1}{\sqrt{d}}.$$

By a simple union bound, it follows from Lemma A.7 that, for large enough d,

$$\begin{aligned}\mathbb{P}\left(\left\|\hat{\boldsymbol{D}}\right\|_{op} > \frac{C_b}{C_X}\right) &\leq Cd \cdot \exp\left(-c\left(\frac{C_X}{C_b} - \sqrt{p}\right)^2 \sqrt{d}\right) \\ &\leq C \cdot \frac{1}{d^{1+k_F/2}},\end{aligned}$$

where we have used that $\frac{1}{\|\bar{\boldsymbol{b}}\|} \geq \frac{C_b}{C_X}$ implies $\left\|\bar{\boldsymbol{b}}\right\| \leq \frac{C_X}{C_b} \leq (1-c)\sqrt{p}$ (here, c can indeed be treated as a universal constant by assumption). Together with (44), we have

$$\mathbb{E}_{\boldsymbol{m}} \left\| F(\hat{\boldsymbol{B}}) - F(\hat{\boldsymbol{D}}^b\bar{\boldsymbol{B}})\right\|_{op} \leq C_F \frac{1}{d}.$$

We now need to bound

$$\mathbb{E}_{\boldsymbol{m}} \left\| F(\hat{\boldsymbol{D}}^b\bar{\boldsymbol{B}}) - F\left(\frac{1}{\sqrt{p}}\bar{\boldsymbol{B}}\right)\right\|_{op}.$$

Since by 3. in the assumptions

$$\mathbb{E}_{\boldsymbol{m}} \left\| F(\hat{\boldsymbol{D}}^b\bar{\boldsymbol{B}}) - F\left(\frac{1}{\sqrt{p}}\bar{\boldsymbol{B}}\right)\right\|_{op} \leq C_{F'} \mathbb{E}_{\boldsymbol{m}} \left\| \hat{\boldsymbol{D}}^b\bar{\boldsymbol{B}} - \frac{1}{\sqrt{p}}\bar{\boldsymbol{B}}\right\|_{op},$$

we will only need to show concentration for $\hat{\boldsymbol{D}}^b$.

Recall that $\hat{\boldsymbol{D}}^b_{i,i} = \min\{\left\|\bar{\boldsymbol{b}}_i\right\|_2^{-1}, C_b/C_X\}$. We now show that $\left|\hat{\boldsymbol{D}}^b_{i,i} - \frac{1}{\sqrt{p}}\right| > \lambda$ implies $\left|\left\|\bar{\boldsymbol{b}}_i\right\|_2 - \sqrt{p}\right| > c\lambda$. To do so, we distinguish two cases. If $\left\|\bar{\boldsymbol{b}}_i\right\| \geq \frac{C_X}{C_b}$, we have

$$\left|\hat{\boldsymbol{D}}^b_{i,i} - \frac{1}{\sqrt{p}}\right| = \left|\frac{\left\|\bar{\boldsymbol{b}}_i\right\| - \sqrt{p}}{\left\|\bar{\boldsymbol{b}}_i\right\|\sqrt{p}}\right|.$$

Thus, $\left|\hat{\boldsymbol{D}}^b_{i,i} - \frac{1}{\sqrt{p}}\right| > \lambda$ implies $\left|\|\bar{\boldsymbol{b}}_i\|_2 - \sqrt{p}\right| > \|\bar{\boldsymbol{b}}_i\| \sqrt{p}\lambda \geq \sqrt{p}\frac{C_X}{C_b}\lambda = c\lambda$.

Next, if $\|\bar{\boldsymbol{b}}_i\| \leq \frac{C_X}{C_b} < \sqrt{p}$, then $\left|\hat{\boldsymbol{D}}^b_{i,i} - \frac{1}{\sqrt{p}}\right| = \frac{C_b}{C_X} - \frac{1}{\sqrt{p}}$ so necessarily $\lambda \leq \frac{C_b}{C_X} - \frac{1}{\sqrt{p}}$. But then

$$\left|\|\bar{\boldsymbol{b}}\|_2 - \sqrt{p}\right| \geq \left|\frac{C_X}{C_b} - \sqrt{p}\right| \geq c\lambda,$$

where the last step is just the previous case for $\|\bar{\boldsymbol{b}}_i\| = \frac{C_X}{C_b}$.

This completes the proof that $\left|\hat{\boldsymbol{D}}^b_{i,i} - \frac{1}{\sqrt{p}}\right| > \lambda$ implies $\left|\|\bar{\boldsymbol{b}}\|_2 - \sqrt{p}\right| > c\lambda$. Now, by Lemma A.6 we have $\|\boldsymbol{b}_i\|_4^2 \leq CC_X^2\frac{\log(d)}{\sqrt{d}}$. So, we can use Lemma A.7 and a union bound to obtain

$$\mathbb{P}\left(\left\|\hat{\boldsymbol{D}}^b - \frac{1}{\sqrt{p}}\boldsymbol{I}\right\|_{op} > \lambda\right) \leq Cd\cdot\exp\left(-c\frac{d\lambda^2}{C_X^4\log(d)^2}\right).$$

For $\lambda = C_X^2\frac{\log^3(d)}{\sqrt{d}}$ and large enough d, we can bound the RHS by d^{-2}. By 1. in Lemma A.4, we have that, for large d, $\|\bar{\boldsymbol{B}}\|_{op} \leq \|\boldsymbol{B}\|_{op} \leq C_X + C_R\frac{\log^3(d)}{\sqrt{d}} \leq 2C_X$. Note also that $\hat{\boldsymbol{D}}^b\bar{\boldsymbol{B}} - \frac{1}{\sqrt{p}}\bar{\boldsymbol{B}} = \left(\hat{\boldsymbol{D}}^b - \frac{1}{\sqrt{p}}\boldsymbol{I}\right)\bar{\boldsymbol{B}}$. Hence,

$$\mathbb{P}\left(\left\|\hat{\boldsymbol{D}}^b\bar{\boldsymbol{B}} - \frac{1}{\sqrt{p}}\bar{\boldsymbol{B}}\right\|_{op} > 2C_X^3\frac{\log^3(d)}{\sqrt{d}}\right) \leq \mathbb{P}\left(\left\|\hat{\boldsymbol{D}}^b - \frac{1}{\sqrt{p}}\boldsymbol{I}\right\|_{op} > C_X^2\frac{\log^3(d)}{\sqrt{d}}\right) \leq \frac{1}{d^2}.$$

Thus, by fixing $\lambda = 2C_X^3\frac{\log^3(d)}{\sqrt{d}}$ and using that $\hat{\boldsymbol{D}}^b$ is bounded, we conclude

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{m}}\left\|\hat{\boldsymbol{D}}^b\bar{\boldsymbol{B}} - \frac{1}{\sqrt{p}}\bar{\boldsymbol{B}}\right\|_{op} &\leq \lambda\mathbb{P}\left(\left\|\hat{\boldsymbol{D}}^b\bar{\boldsymbol{B}} - \frac{1}{\sqrt{p}}\bar{\boldsymbol{B}}\right\|_{op} \leq \lambda\right) + \max_{\boldsymbol{m}}\left\|\hat{\boldsymbol{D}}^b\bar{\boldsymbol{B}} - \frac{1}{\sqrt{p}}\bar{\boldsymbol{B}}\right\|_{op}\mathbb{P}\left(\left\|\hat{\boldsymbol{D}}^b\bar{\boldsymbol{B}} - \frac{1}{\sqrt{p}}\bar{\boldsymbol{B}}\right\|_{op} > \lambda\right) \\
&\leq \lambda + CC_X\mathbb{P}\left(\left\|\hat{\boldsymbol{D}}^b\bar{\boldsymbol{B}} - \frac{1}{\sqrt{p}}\bar{\boldsymbol{B}}\right\|_{op} > \lambda\right) \\
&\leq CC_X^3\frac{\log^3(d)}{\sqrt{d}} + CC_X\frac{1}{d^2} \\
&\leq CC_X^3\frac{\log^3(d)}{\sqrt{d}}.
\end{aligned}$$

□

Lemma A.13 (Explicit approximations)**.** *Assume that $\mathbb{P}_{\boldsymbol{m}}\left(\bar{\boldsymbol{B}} \in \Omega\right) \geq 1 - C\frac{1}{d^2}$, where*

$$\begin{aligned}
\Omega = \{&\boldsymbol{M}|(\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I})\circ(\boldsymbol{M}\boldsymbol{M}^\top) = \boldsymbol{Y} + \boldsymbol{Z}, \|\boldsymbol{Y}\|_{max} \leq CC_X^2\frac{\log^3(d)}{\sqrt{d}}, \|\boldsymbol{Z}\|_{op} \leq CC_XC_R\frac{\log^{\alpha_R}(d)}{\sqrt{d}}\} \\
\subset \{&\boldsymbol{M}|\left\|(\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I})\circ\boldsymbol{M}\boldsymbol{M}^\top\right\|_{max} \leq C(C_X^2 + C_XC_R)\frac{\log^{\alpha_R}(d))}{\sqrt{d}}\}.
\end{aligned}$$

Then, with probability at least $1 - C\frac{1}{d^2}$ in $\boldsymbol{U}, \boldsymbol{V}$, the following functions satisfy the assumption of Lemma A.12 (with $C_F, C_{F'}$ independent of d)

1. $F(\boldsymbol{B}) = -\boldsymbol{A}^\top + \frac{1}{\sqrt{p}}\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{B} + \frac{1}{p}\mathrm{Diag}\left(\boldsymbol{B}\boldsymbol{A}\right)\boldsymbol{B} - \frac{1}{\sqrt{p}}\mathrm{Diag}(\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{B}\boldsymbol{B}^\top)\boldsymbol{B}$,
2. $F(\boldsymbol{B}) = \sum_{\ell\geq3} c_\ell^2(\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I})\circ(\boldsymbol{B}\boldsymbol{B}^\top)^{\circ\ell}, \sum_{\ell\geq3} c_\ell^2 < \infty$,
3. $F(\boldsymbol{B}) = \sum_{\ell\geq3} \ell c_\ell^2\left((\boldsymbol{A}^\top\boldsymbol{A})\circ(\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I})\circ(\boldsymbol{B}\boldsymbol{B}^\top)^{\circ(\ell-1)} - \mathrm{Diag}\left(\boldsymbol{A}^\top\boldsymbol{A}(\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I})\circ(\boldsymbol{B}\boldsymbol{B}^\top)^{\circ\ell}\right)\right)\boldsymbol{B}$,
   $\sum_{\ell\geq3} c_\ell^2 < \infty$,

where for 3. we need to additionally assume that the conclusion of Lemma A.14 holds. (Note that this is not an issue as the proof of Lemma A.14 uses only 2. in the current lemma).

Proof. First note that, for fixed C_b, C_X, p, by scaling the constant in the definition of Ω by $\frac{1}{p}\frac{C_b^2}{C_X^2}$, we may w.l.o.g. assume that $\mathbb{P}\left(\boldsymbol{D}^b\bar{\boldsymbol{B}}, \frac{1}{\sqrt{p}}\bar{\boldsymbol{B}} \in \Omega\right) \geq 1 - C\frac{1}{d^2}$.

We now show the claim for each of the functions separately.

1. The first function is the sum of of multi-linear functions and thus is polynomially bounded and Lipschitz on bounded sets. Since we have a dimension independent bound on $\boldsymbol{A}$, the bounds are dimension-independent as well.

2. We will show condition 1., 2. and 3. in Lemma A.12 separately. For 1., note that since $\left|(\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top)_{i,j}\right| \leq 1$ we have $\left|F(\hat{\boldsymbol{B}})_{i,j}\right| \leq \sum_\ell c_\ell^2 = C < \infty$. Thus from 5. in Lemma A.4, we obtain $\left\|F(\hat{\boldsymbol{B}})\right\|_{op} \leq Cd$ as desired.

   Next we show condition 2. We have the following estimate for $\boldsymbol{M} \in \Omega$ and $\ell \geq 3$

$$\begin{aligned}\ell\left\|(\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I}) \circ \left(\boldsymbol{M}\boldsymbol{M}^\top\right)^{\circ\ell}\right\|_{op} &\leq n\ell\left\|(\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I}) \circ \left(\boldsymbol{M}\boldsymbol{M}^\top\right)^{\circ\ell}\right\|_{max} \\ &\leq n\ell\left(C(C_M^2 + C_M C_R)\frac{\log^{\alpha_R}(d)}{\sqrt{d}}\right)^\ell \\ &\leq C(C_M^2 + C_M C_R)^3\frac{(\log^{\alpha_R}(d))^3}{\sqrt{d}} \\ &\leq C\frac{1}{d^{\frac{1}{4}}},\end{aligned} \tag{45}$$

   where the first step follows from 5. in Lemma A.4. Since the above bound is independent of ℓ it also holds for F since $\sum_{\ell\geq 3} c_\ell^2 < \infty$. This gives us the desired bound for 2.

   To show 3. we write $F(\boldsymbol{M}) = \sum_\ell c_\ell^2 F_2^\ell(F_1(\boldsymbol{M}))$ where $F_1(\boldsymbol{M}) := (\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I}) \circ (\boldsymbol{M}\boldsymbol{M}^\top)$ and $F_2^\ell(\boldsymbol{Q}) = \boldsymbol{Q}^{\circ\ell}$. We will show that F_1, F_2^ℓ are Lipschitz. By 4. in Lemma A.4, we have that $\left\|(\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I}) \circ \boldsymbol{Q}\right\|_{op} \leq 2\left\|\boldsymbol{Q}\right\|_{op}$. Thus, for $\boldsymbol{B}_1, \boldsymbol{B}_2 \in \mathcal{B}_{C_b}(\boldsymbol{0})$,

$$\begin{aligned}\left\|F_1(\boldsymbol{B}_1) - F_1(\boldsymbol{B}_2)\right\|_{op} &= \left\|(\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I}) \circ (\boldsymbol{B}_1\boldsymbol{B}_1^\top - \boldsymbol{B}_2\boldsymbol{B}_2^\top)\right\|_{op} \\ &\leq 2\left\|\boldsymbol{B}_1\boldsymbol{B}_1^\top - \boldsymbol{B}_2\boldsymbol{B}_2^\top\right\|_{op} \\ &= 2\left\|\boldsymbol{B}_1(\boldsymbol{B}_1^\top - \boldsymbol{B}_2^\top) + (\boldsymbol{B}_1 - \boldsymbol{B}_2)\boldsymbol{B}_2^\top\right\|_{op} \\ &\leq 4C_b\left\|\boldsymbol{B}_1 - \boldsymbol{B}_2\right\|_{op}.\end{aligned}$$

   Hence, F_1 is Lipschitz with constant $4C_b$.

   For F_2^ℓ, we will show that $\left\|DF_2^\ell(\boldsymbol{Q})\right\| \leq Cd^{-\frac{1}{4}}$ if $\boldsymbol{Q} = (\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I}) \circ (\boldsymbol{M}\boldsymbol{M}^\top), \boldsymbol{M} \in \Omega \cap \mathcal{B}_{C_b}(\boldsymbol{0})$. Note that $\left\|\boldsymbol{Q}\right\|_{max} \leq C(C_M^2 + C_M C_R)\frac{\log^{\alpha_R}(d)}{\sqrt{d}}$ and $\left\|\boldsymbol{Q}\right\|_{op} \leq 2C_b^2$. Furthermore, F_2^ℓ is a symmetric ℓ-linear function, so the derivative in the direction $\boldsymbol{Z}$ is given by $DF_2^\ell(\boldsymbol{Q})\boldsymbol{Z} = \ell\boldsymbol{Z} \circ \boldsymbol{Q}^{\circ\ell-1}$. From 3. in Lemma A.4 we have

$$\begin{aligned}\left\|\ell\boldsymbol{Z} \circ \boldsymbol{Q}^{\circ\ell-1}\right\|_{op} &\leq \ell\sqrt{n}\left\|\boldsymbol{Z}\right\|_{op}\left(C(C_M^2 + C_M C_R)\frac{\log^{\alpha_R}(d)}{\sqrt{d}}\right)^{\ell-1} \\ &\leq C(C_M^2 + C_M C_R)^2\left\|\boldsymbol{Z}\right\|_{op}\frac{(\log^{\alpha_R}(d))^2}{\sqrt{d}} \\ &\leq C\left\|\boldsymbol{Z}\right\|_{op}\frac{1}{d^{\frac{1}{4}}},\end{aligned} \tag{46}$$

so $\left\|DF_2^\ell(F_1(\boldsymbol{B}))\right\| \le Cd^{-\frac{1}{4}}$. Now, note that since

$$\{\boldsymbol{Q} | \left\|(\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I}) \circ \boldsymbol{Q}\right\|_{max} \le C(C_M^2 + C_M C_R)\frac{\log^{\alpha_R}(d)}{\sqrt{d}}, \mathrm{Diag}\,(\boldsymbol{Q}) = 0\}$$

is convex, the line segment between any two points lies in the set, so a bound on the derivative implies that the $\boldsymbol{Q}^{\circ\ell}$ is Lipschitz with the same constant. Multiplying the two Lipschitz constants of F_1, F_2^ℓ we obtain that their composition is Lipschitz with constant $4CC_b d^{-\frac{1}{4}}$.

Since none of the bounds depends on ℓ, this immediately implies that $F(\boldsymbol{M}) = \sum_\ell c_\ell^2 F_2^\ell(F_1(\boldsymbol{M}))$ is Lipschitz as well, up to an additional constant $\sum_{\ell\ge 3} c_\ell^2$.

3. Again we will first show that condition 1. holds for $\boldsymbol{B} = \hat{\boldsymbol{B}}$. First note that we can write (as in Lemma A.15 below)

$$(F(\hat{\boldsymbol{B}}))_{k,:} = \sum_{\ell=3}^{\infty} \ell c_\ell^2 \sum_{j\ne k} \langle \boldsymbol{a}_k, \boldsymbol{a}_j\rangle \langle \hat{\boldsymbol{b}}_k, \hat{\boldsymbol{b}}_j\rangle^{\ell-1} \hat{\boldsymbol{J}}_k' \hat{\boldsymbol{b}}_j,$$

where $\hat{\boldsymbol{J}}_k' = \boldsymbol{I} - \hat{\boldsymbol{b}}_k \hat{\boldsymbol{b}}_k^\top$. Observe that

$$\left\|\hat{\boldsymbol{J}}_k' \hat{\boldsymbol{b}}_j\right\|^2 = \left\|\hat{\boldsymbol{b}}_j - \hat{\boldsymbol{b}}_k \langle \hat{\boldsymbol{b}}_k, \hat{\boldsymbol{b}}_j\rangle\right\|^2 = 1 - \langle \hat{\boldsymbol{b}}_k, \hat{\boldsymbol{b}}_j\rangle^2 \le 2\left(1 - \left|\langle \hat{\boldsymbol{b}}_k, \hat{\boldsymbol{b}}_j\rangle\right|\right). \tag{47}$$

Thus, using Lemma A.5, we have

$$\begin{aligned}\left\|\sum_{\ell=3}^{\infty} \ell c_\ell^2 \sum_{j\ne k} \langle \boldsymbol{a}_k, \boldsymbol{a}_j\rangle \langle \hat{\boldsymbol{b}}_k, \hat{\boldsymbol{b}}_j\rangle^{\ell-1} \hat{\boldsymbol{J}}_k' \hat{\boldsymbol{b}}_j\right\| &\le C \left\|(\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I}) \circ \boldsymbol{A}^\top \boldsymbol{A}\right\|_{max} \sum_{\ell=3}^{\infty} \ell c_\ell^2 \sum_{j\ne k} \left|\langle \hat{\boldsymbol{b}}_k, \hat{\boldsymbol{b}}_j\rangle^{\ell-1}\right| \left\|\hat{\boldsymbol{J}}_k' \hat{\boldsymbol{b}}_j\right\| \\ &\le C \left\|(\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I}) \circ \boldsymbol{A}^\top \boldsymbol{A}\right\|_{max} \left\|\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top\right\|_{max}^2 \sum_{j\ne k} \frac{1}{\sqrt{1 - \left|\langle \hat{\boldsymbol{b}}_k, \hat{\boldsymbol{b}}_j\rangle\right|}} \left\|\hat{\boldsymbol{J}}_k' \hat{\boldsymbol{b}}_j\right\| \\ &\le Cn \left\|(\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I}) \circ \boldsymbol{A}^\top \boldsymbol{A}\right\|_{max} \left\|\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top\right\|_{max}^2 \\ &\le Cd,\end{aligned}$$

where we have used (47) in the third step . Using 2. in Lemma A.4, the above implies

$$\left\|F(\hat{\boldsymbol{B}})\right\|_{op} \le CC_R d^{\frac{3}{2}}.$$

This shows that condition 1. in Lemma A.12 holds for $\hat{\boldsymbol{B}}$.

To show the rest of the conditions, we may now assume that $\boldsymbol{B} \in \Omega \cap \mathcal{B}_{C_b}(\boldsymbol{0})$. Note that, if we show that F is Lipschitz on this set, condition 2. holds since

$$\|F(\boldsymbol{B})\|_{op} = \|F(\boldsymbol{B}) - F(\boldsymbol{0})\|_{op} \le C_F' \|\boldsymbol{B}\|_{op}, \tag{48}$$

where we have used $F(\boldsymbol{0}) = \boldsymbol{0}$. Thus we only need to show the third condition.

Similarly to the previous case in 3., we define $F(\boldsymbol{M}) = F_2(F_1(\boldsymbol{M}))\boldsymbol{M}$ where $F_1(\boldsymbol{M}) := (\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I}) \circ (\boldsymbol{M}\boldsymbol{M}^\top)$ and

$$F_2(\boldsymbol{Q}) = \sum_{\ell\ge 3} \ell c_\ell^2 \left((\boldsymbol{A}^\top \boldsymbol{A}) \circ \boldsymbol{Q}^{\circ(\ell-1)} - \mathrm{Diag}\left(\boldsymbol{A}^\top \boldsymbol{A} \boldsymbol{Q}^{\circ\ell}\right)\right).$$

Note that it is enough to show that $F_2(F_1(\boldsymbol{M}))$ is Lipschitz, as then by (48) and $\boldsymbol{M} \in \mathcal{B}_{C_b}(\boldsymbol{0})$, F is the product of two bounded Lipschitz functions, and thus Lipschitz. As for the previous function, we have that F_1 is Lipschitz with constant $4C_b$. We will now derive a uniform bound for all $\ell \ge 3$. Define

$$F_2^\ell(\boldsymbol{Q}) := \ell c_\ell^2 \left((\boldsymbol{A}^\top \boldsymbol{A}) \circ \boldsymbol{Q}^{\circ(\ell-1)} - \mathrm{Diag}\left(\boldsymbol{A}^\top \boldsymbol{A} \boldsymbol{Q}^{\circ\ell}\right)\right).$$

As in the previous case, since $\boldsymbol{Q}^{\circ \ell}$ is a symmetric ℓ-linear function, we have

$$DF_2^\ell(\boldsymbol{Q})\boldsymbol{Z} = \ell c_\ell^2 \left((\ell-1)(\boldsymbol{A}^\top \boldsymbol{A}) \circ \boldsymbol{Q}^{\circ(\ell-2)} \circ \boldsymbol{Z} - \ell \mathrm{Diag}\left(\boldsymbol{A}^\top \boldsymbol{A} \boldsymbol{Q}^{\circ(\ell-1)} \circ \boldsymbol{Z} \right) \right).$$

Recall we assume the conclusion of Lemma A.14 to hold, so

$$\boldsymbol{A} = \boldsymbol{X}^\top \left(\boldsymbol{X}\boldsymbol{X}^\top + \alpha \boldsymbol{I} \right)^{-1} + O\left(\boldsymbol{R}\right) + O\left(C_X^7 \frac{\log^{10}(d)}{\sqrt{d}} \right) = O_{max}\left(\frac{\log(d)}{\sqrt{d}} \right) + O\left(C_X^7 C_R \frac{\log^{\alpha_R}(d)}{\sqrt{d}} \right), \quad (49)$$

where we have used Lemma A.8 in the second step. Similarly, we have

$$\boldsymbol{A}^\top \boldsymbol{A} = O_{max}\left(\frac{\log(d)}{\sqrt{d}} \right) + O\left(C_X^7 C_R \frac{\log^{\alpha_R}(d)}{\sqrt{d}} \right). \quad (50)$$

Thus, by using that $\|\boldsymbol{R} \circ \boldsymbol{S}\|_{op} \le \|\boldsymbol{R}\|_{op} \|\boldsymbol{S}\|_{op}$ for any square matrices $\boldsymbol{R}, \boldsymbol{S}$ (see Theorem 1 in (Visick, 2000)), we obtain

$$\left\| (\boldsymbol{A}^\top \boldsymbol{A}) \circ \boldsymbol{Q}^{\circ(\ell-2)} \circ \boldsymbol{Z} \right\|_{op} \le \left\| O_{max}\left(\frac{\log(d)}{\sqrt{d}} \right) \circ \boldsymbol{Q}^{\circ(\ell-2)} \circ \boldsymbol{Z} \right\|_{op} + O\left(C_X^7 C_R \frac{\log^{\alpha_R}(d)}{\sqrt{d}} \right) \left\| \boldsymbol{Q}^{\circ(\ell-2)} \circ \boldsymbol{Z} \right\|_{op},$$

Using the same estimate as in (46), 3. in Lemma A.4 and recalling that $\ell \ge 3$,

$$\begin{aligned}
&\left\| DF_2^\ell(\boldsymbol{Q})\boldsymbol{Z} \right\|_{op} \le \\
&\le C\ell^2 c_\ell^2 \left(\sqrt{n} \|\boldsymbol{Z}\|_{op} \frac{\log(d)}{\sqrt{d}} \|\boldsymbol{Q}\|_{max}^{\ell-2} + C_X^7 C_R \sqrt{n} \|\boldsymbol{Z}\|_{op} \frac{\log^{\alpha_R}(d)}{\sqrt{d}} \|\boldsymbol{Q}\|_{max}^{\ell-2} + \sqrt{n} \|\boldsymbol{Z}\|_{op} \|\boldsymbol{Q}\|_{max}^{\ell-1} \right) \\
&\le C\ell^2 \sqrt{d} \left(\left(C_X^7 C_R \frac{\log^{\alpha_R}(d)}{\sqrt{d}} \right) \left(C(C_X^2 + C_X C_R) \frac{\log^{\alpha_R}(d)}{\sqrt{d}} \right)^{\ell-2} + \left(C(C_X^2 + C_X C_R) \frac{\log^{\alpha_R}(d)}{\sqrt{d}} \right)^{\ell-1} \right) \|\boldsymbol{Z}\|_{op} \\
&\le C d^{-\frac{\ell-2}{4}} \|\boldsymbol{Z}\|_{op}.
\end{aligned} \quad (51)$$

Since $\{\boldsymbol{Q} | \, \|\boldsymbol{Q}\|_{max} \le C(C_M^2 + C_M C_R) \frac{\log^3(d)}{\sqrt{d}}, \mathrm{Diag}\left(\boldsymbol{Q}\right) = 0\}$ is convex, the line segment between any two points lies in the set, so a bound on the derivative implies that the F_2^ℓ is Lipschitz with the same constant. As $F_2 = \sum_{\ell \ge 3} F_2^\ell$, we have that F_2 is Lipschitz with constant $\sum_{\ell \ge 3} C d^{-\frac{\ell-2}{4}} \le C d^{-\frac{1}{4}}$. Finally the composition $F(\boldsymbol{B}) = F_2(F_1(\boldsymbol{B}))$ is Lipschitz with constant $CC_b d^{-\frac{1}{4}}$, so condition 3. holds, which concludes the proof.

□

A.5. Concentration of the gradient

Lemma A.14 (Error analysis of $\boldsymbol{A}$)**.** *Assume that $\boldsymbol{B} = \boldsymbol{X} + \boldsymbol{R}$ with unit norm rows, $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top, \boldsymbol{U}, \boldsymbol{V}$ Haar, $\|\boldsymbol{X}\|_{op} \le C_X$, $\|\boldsymbol{X}\|_{max} \le C_X \frac{\log(d)}{\sqrt{d}}$, $\|\boldsymbol{R}\|_{op} \le C_R \frac{\log^{\alpha_R}(d)}{\sqrt{d}}$. Then, for $d > d_0(C_R)$ with probability at least $1 - C\frac{1}{d^2}$ (in $\boldsymbol{U}, \boldsymbol{V}$) we have*

$$\begin{aligned}
\boldsymbol{A} = \frac{1}{\sqrt{p}} \mathbb{E}_{\boldsymbol{m}} \hat{\boldsymbol{B}}^\top \left(\mathbb{E}_{\boldsymbol{m}} f\left(\hat{\boldsymbol{B}} \hat{\boldsymbol{B}}^\top \right) \right)^{-1} &= \boldsymbol{B}^\top \left(\boldsymbol{B}\boldsymbol{B}^\top + \alpha \boldsymbol{I} \right)^{-1} + O\left(C_X^7 \frac{\log^{10}(d)}{\sqrt{d}} \right) \\
&= \boldsymbol{X}^\top \left(\boldsymbol{X}\boldsymbol{X}^\top + \alpha \boldsymbol{I} \right)^{-1} + O\left(\boldsymbol{R}\right) + O\left(C_X^7 \frac{\log^{10}(d)}{\sqrt{d}} \right).
\end{aligned} \quad (52)$$

Proof. By a straightforward application of Lemma A.12, we have

$$\begin{aligned}
\frac{1}{\sqrt{p}} \mathbb{E}_{\boldsymbol{m}} \hat{\boldsymbol{B}} &= \mathbb{E}_{\boldsymbol{m}} \frac{1}{p} \bar{\boldsymbol{B}} + O\left(C_X^3 \frac{\log^3(d)}{\sqrt{d}} \right) \\
&= \boldsymbol{B} + O\left(C_X^3 \frac{\log^3(d)}{\sqrt{d}} \right).
\end{aligned} \quad (53)$$

Next, we will estimate $\mathbb{E}_{\boldsymbol{m}} f\left(\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top\right)$. Recall that $f(x) = \sum_\ell c_\ell^2 x^\ell$. As $f(1) < \infty$, we can define $\alpha = \sum_{\ell \geq 3} c_\ell^2$. As $c_1 = 1$, we have

$$f(\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top) := \hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top + \sum_{\ell \geq 3} c_\ell^2 \left(\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top\right)^{\circ \ell} = \hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top + \alpha \boldsymbol{I} + \sum_{\ell \geq 3} c_\ell^2 \left(\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top - \boldsymbol{I}\right)^{\circ \ell}.$$

Let $\Omega = \{\boldsymbol{M} | (\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I}) \circ (\boldsymbol{M}\boldsymbol{M}^\top) = \boldsymbol{Y} + \boldsymbol{Z}, \|\boldsymbol{Y}\|_{max} \leq CC_X^2 \frac{\log^3(d)}{\sqrt{d}}, \|\boldsymbol{Z}\|_{op} \leq CC_X C_R \frac{\log^{\alpha_R}(d)}{\sqrt{d}}\} \subset \{\boldsymbol{M} | \left\|(\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I}) \circ \boldsymbol{M}\boldsymbol{M}^\top\right\|_{max} \leq C(C_X^2 + C_X C_R) \frac{\log^{\alpha_R}(d))}{\sqrt{d}}\}$, then by using 4. in lemma A.11 with $\gamma = 2$ we have $\mathbb{P}_{\boldsymbol{m}}\left(\bar{\boldsymbol{B}} \in \Omega\right) \geq 1 - C\frac{1}{d^2}$ (with probability at least $1 - C/d^2$ in $\boldsymbol{U}, \boldsymbol{V}$). By noting that $\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top$ satisfies the assumptions of Lemma A.12 and using 2. in Lemma A.13 we have

$$\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top + \sum_{\ell \geq 3} c_\ell^2 \left(\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top\right)^{\circ \ell} = \frac{1}{p}\bar{\boldsymbol{B}}\bar{\boldsymbol{B}}^\top + \alpha \boldsymbol{I} + \mathbb{1}_{\bar{\boldsymbol{B}} \in \Omega} \sum_{\ell \geq 3} c_\ell^2 (\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I}) \circ \left(\frac{1}{p}\bar{\boldsymbol{B}}\bar{\boldsymbol{B}}^\top\right)^{\circ \ell} + O\left(CC_X^3 \frac{\log^3(d)}{\sqrt{d}}\right). \tag{54}$$

By linearity, we have $\mathbb{E}_{\boldsymbol{m}} \frac{1}{p}\bar{\boldsymbol{B}}\bar{\boldsymbol{B}}^\top = \boldsymbol{B}\boldsymbol{B}^\top$. We will now show that

$$\mathbb{E}_{\boldsymbol{m}} \sum_{\ell \geq 3} \mathbb{1}_{\bar{\boldsymbol{B}} \in \Omega} c_\ell^2 (\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I}) \circ \left(\frac{1}{p}\bar{\boldsymbol{B}}\bar{\boldsymbol{B}}^\top\right)^{\circ \ell} = O\left(C_X^6 \frac{\log^{10}(d)}{\sqrt{d}}\right). \tag{55}$$

For now, let $(\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I}) \circ \bar{\boldsymbol{B}}\bar{\boldsymbol{B}}^\top = \boldsymbol{Y} + \boldsymbol{Z}$, $\|\boldsymbol{Y}\|_{max} \leq CC_X^2 \frac{\log^3(d)}{\sqrt{d}}$, $\|\boldsymbol{Z}\|_{op} \leq CC_X C_R \frac{\log^{\alpha_R}(d)}{\sqrt{d}}$, as in the definition of Ω above. By the definition of Ω, we have that, for $\bar{\boldsymbol{B}} \in \Omega$ and $\ell \geq 3$,

$$\left(\frac{1}{p}(\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I}) \circ \bar{\boldsymbol{B}}\bar{\boldsymbol{B}}^\top\right)^{\circ \ell} = \frac{1}{p^\ell}\boldsymbol{Y}^{\circ \ell} + \frac{1}{p^\ell}\ell \boldsymbol{Y}^{\circ(\ell-1)} \circ \left(\boldsymbol{Z} + e^\ell O\left(\boldsymbol{Z}^2\right)\right). \tag{56}$$

Thus, by 3. in Lemma A.4, we have that for $\ell \geq 3$

$$\frac{1}{p^\ell}\left\|\boldsymbol{Y}^{\circ \ell}\right\|_{op} \leq C\frac{1}{p^\ell}\sqrt{d}C_X^2 \left(C_X^2 \frac{\log^3(d)}{\sqrt{d}}\right)^{\ell-1} \leq CC_X^6 \frac{\log^{10}(d)}{\sqrt{d}}$$

and

$$\frac{1}{p^\ell}e^\ell \ell \left\|\boldsymbol{Y}^{\circ(\ell-1)} \circ \left(\boldsymbol{Z} + e^\ell O\left(\boldsymbol{Z}^2\right)\right)\right\|_{op} \leq C\frac{1}{p^\ell}e^\ell \ell \sqrt{d} C_X C_R \frac{\log^{\alpha_R}(d)}{\sqrt{d}} \left(C_X^2 \frac{\log^3(d)}{\sqrt{d}}\right)^{\ell-1} \leq d^{-3/4}.$$

Thus, we can further estimate (56) by

$$\mathbb{E}_{\boldsymbol{m}} \left\|\mathbb{1}_{\bar{\boldsymbol{B}} \in \Omega}(\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I}) \circ \left(\frac{1}{p}\bar{\boldsymbol{B}}\bar{\boldsymbol{B}}^\top\right)^{\circ \ell}\right\|_{op} \leq CC_X^6 \frac{\log^{10}(d)}{\sqrt{d}}. \tag{57}$$

Since the bound is independent of ℓ, this shows (55).

Combining (54) and (55) we now have

$$\mathbb{E}_{\boldsymbol{m}} f(\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top) = \boldsymbol{B}\boldsymbol{B}^\top + \alpha \boldsymbol{I} + O\left(C_X^6 \frac{\log^{10}(d)}{\sqrt{d}}\right). \tag{58}$$

From (58) it also immediately follows that

$$\left(\mathbb{E}_{\boldsymbol{m}} f\left(\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top\right)\right)^{-1} = \left(\boldsymbol{B}\boldsymbol{B}^\top + \alpha \boldsymbol{I}\right)^{-1} + O\left(C_X^6 \frac{\log^{10}(d)}{\sqrt{d}}\right), \tag{59}$$

since for any psd matrix $\boldsymbol{X} > \alpha \boldsymbol{I}$, the map from $(\mathcal{M}_{n,n}, \|\cdot\|_{op}) \to (\mathcal{M}_{n,n}, \|\cdot\|_{op})$, $\boldsymbol{R} \to (\boldsymbol{X} + \boldsymbol{R})^{-1}$ is locally continuously differentiable at 0. Combining (53) and (59) yields the first equality in (52). To see the second equality in (52), it suffices to use the fact that the function $\boldsymbol{B} \to \boldsymbol{B}^\top \left(\boldsymbol{B}\boldsymbol{B}^\top + \alpha \boldsymbol{I}\right)$ is Lipschitz on bounded sets w.r.t $\|\cdot\|_{op}$. □

Lemma A.15 (Gradient concentration, Part 1)**.** *Assume that* $\boldsymbol{B} = \boldsymbol{X} + \boldsymbol{R}$ *with unit norm rows,* $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top, \boldsymbol{U}, \boldsymbol{V}$ *Haar,* $\|\boldsymbol{X}\|_{op} \le C_X$*,* $\|\boldsymbol{X}\|_{max} \le CC_X \frac{\log(d)}{\sqrt{d}}$*,* $\|\boldsymbol{R}\|_{op} \le C_R \frac{\log^{\alpha_R}(d)}{\sqrt{d}}, \|\boldsymbol{A}\|_{op} \le C$*. Further assume that* $\min_{i,j} |\boldsymbol{B}_{i,j}| \ge \delta > d^{-\gamma_\delta}$ *for some* $\gamma_\delta > 0$*. Then, for* $d > d_0(C_X, C_R, \gamma_\delta)$ *with probability at least* $1 - C\frac{1}{d^2}$ *in* $\boldsymbol{U}, \boldsymbol{V}$*, the gradient of* (25) *w.r.t.* $\boldsymbol{B}$ *can be written as*

$$\nabla_{\boldsymbol{B}} = \mathbb{E}_{\boldsymbol{m}} \nabla^1_{\hat{\boldsymbol{B}}} + \sum_{\ell=3}^{\infty} \ell c_\ell^2 \mathbb{E}_{\boldsymbol{m}} \nabla^\ell_{\hat{\boldsymbol{B}}} + O\left(C_X^7 \frac{\log^2(d)}{\sqrt{d}}\right), \tag{60}$$

where

$$\begin{aligned} \frac{1}{2}(\nabla^1_{\hat{\boldsymbol{B}}})_{k,:} &= -\boldsymbol{a}_k + \frac{1}{p}\langle \boldsymbol{a}_k, \hat{\boldsymbol{b}}_k \rangle \hat{\boldsymbol{b}}_k + \frac{1}{\sqrt{p}} \sum_j \langle \boldsymbol{a}_k, \boldsymbol{a}_j \rangle \hat{\boldsymbol{J}}'_k \hat{\boldsymbol{b}}_j, \\ \frac{1}{2}\nabla^1_{\hat{\boldsymbol{B}}} &= -\boldsymbol{A}^\top + \frac{1}{\sqrt{p}} \boldsymbol{A}^\top \boldsymbol{A} \hat{\boldsymbol{B}} + \frac{1}{p} \mathrm{Diag}(\hat{\boldsymbol{B}}\boldsymbol{A})\hat{\boldsymbol{B}} - \frac{1}{\sqrt{p}} \mathrm{Diag}(\boldsymbol{A}^\top \boldsymbol{A} \hat{\boldsymbol{B}} \hat{\boldsymbol{B}}^\top) \hat{\boldsymbol{B}}, \end{aligned} \tag{61}$$

$$\begin{aligned} \frac{1}{2}(\nabla^\ell_{\hat{\boldsymbol{B}}})_{k,:} &= \frac{1}{\sqrt{p}} \sum_j \langle \boldsymbol{a}_k, \boldsymbol{a}_j \rangle \langle \hat{\boldsymbol{b}}_k, \hat{\boldsymbol{b}}_j \rangle^{\ell-1} \hat{\boldsymbol{J}}'_k \hat{\boldsymbol{b}}_j, \\ \frac{1}{2}\nabla^\ell_{\hat{\boldsymbol{B}}} &= \frac{1}{\sqrt{p}} (\boldsymbol{A}^\top \boldsymbol{A}) \circ (\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top - \boldsymbol{I})^{\circ(\ell-1)} \hat{\boldsymbol{B}} - \frac{1}{\sqrt{p}} \mathrm{Diag}(\boldsymbol{A}^\top \boldsymbol{A} (\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top - \boldsymbol{I})^{\circ \ell}) \hat{\boldsymbol{B}}, \end{aligned} \tag{62}$$

and

$$\hat{\boldsymbol{J}}'_k := \frac{1}{\sqrt{p}} \left(\boldsymbol{I} - \hat{\boldsymbol{b}}_k \hat{\boldsymbol{b}}_k^\top \right).$$

Proof. Recall from (30) that the gradient is given by

$$(\nabla_{\boldsymbol{B}})_{k,:} = \mathbb{E}_{\boldsymbol{m}} \left[-2 \frac{1}{\sqrt{p}} \boldsymbol{m} \circ \hat{\boldsymbol{J}}_k \boldsymbol{a}_k \right] + 2 \sum_{\ell=1}^{\infty} \ell c_\ell^2 \sum_{j \ne k} \langle \boldsymbol{a}_k, \boldsymbol{a}_j \rangle \langle \hat{\boldsymbol{b}}_k, \hat{\boldsymbol{b}}_j \rangle^{\ell-1} \hat{\boldsymbol{J}}_k \hat{\boldsymbol{b}}_j, \tag{63}$$

where $\hat{\boldsymbol{J}}_k = \frac{1}{\|\bar{\boldsymbol{b}}_k\|} \left(\boldsymbol{I} - \hat{\boldsymbol{b}}_k \hat{\boldsymbol{b}}_k^\top \right)$.

We will approximate $\hat{\boldsymbol{J}}_k$ by $\hat{\boldsymbol{J}}'_k$. This will make the gradient have the same functional form (for fixed $\boldsymbol{m}$) as in the Gaussian case. This follows from the fact that the gradient inside the expectation is the same as the gradient of the Gaussian objective (86) in (Shevchenko et al., 2023) evaluated at $\boldsymbol{B} = \bar{\boldsymbol{B}}$. We denote the new gradient with $\hat{\boldsymbol{J}}_k$ replaced by $\hat{\boldsymbol{J}}'_k$ as $\nabla'_{\boldsymbol{B}}$. We proceed by decomposing the error $\|(\nabla_{\boldsymbol{B}})_{k,:} - (\nabla'_{\boldsymbol{B}})_{k,:}\|$ into multiple parts and analysing them individually.

First, we need to decompose the error. Combining Lemma A.7 and Lemma A.6, we have

$$\begin{aligned} \mathbb{P}_{\boldsymbol{m}} \left(\left| \|\bar{\boldsymbol{b}}_k\| - \sqrt{p} \right| > C_X^2 \frac{\log^2(d)}{\sqrt{d}} \right) &= \mathbb{P}_{\boldsymbol{m}} \left(\left| \|\bar{\boldsymbol{b}}_k\|^2 - p \right| > C_X^2 (\|\bar{\boldsymbol{b}}_k\| + \sqrt{p})^{-1} \frac{\log^2(d)}{\sqrt{d}} \right) \\ &\le \mathbb{P}_{\boldsymbol{m}} \left(\left| \|\bar{\boldsymbol{b}}_k\|^2 - p \right| > C_X^2 \frac{1}{1 + \sqrt{p}} \frac{\log^2(d)}{\sqrt{d}} \right) \\ &\le C \frac{1}{d^\gamma}. \end{aligned} \tag{64}$$

Denoting by A the event that $\left| \|\bar{\boldsymbol{b}}_k\| - \sqrt{p} \right| > C_X^2 \frac{\log^2(d)}{\sqrt{d}}$ jointly for all k, we have

$$\begin{aligned} \mathbb{E}_{\boldsymbol{m}} (\nabla_{\boldsymbol{B}})_{k,:} - (\nabla'_{\boldsymbol{B}})_{k,:} &= \mathbb{E}_{\boldsymbol{m}} \left[\mathbf{1}_A ((\nabla_{\boldsymbol{B}})_{k,:} - (\nabla'_{\boldsymbol{B}})_{k,:}) \right] + \mathbb{E}_{\boldsymbol{m}} \left[\mathbf{1}_{A^c} (\nabla_{\boldsymbol{B}})_{k,:} - (\nabla'_{\boldsymbol{B}})_{k,:} \right] \\ &= \mathbb{E}_{\boldsymbol{m}} \left[\mathbf{1}_A ((\nabla_{\boldsymbol{B}})_{k,:} - (\nabla'_{\boldsymbol{B}})_{k,:}) \right] \\ &\quad + \mathbb{E}_{\boldsymbol{m}} \left[-2 \frac{1}{\sqrt{p}} \boldsymbol{m} \circ \epsilon^k_{\boldsymbol{m}} \hat{\boldsymbol{J}}'_k \boldsymbol{a}_k + 2 \sum_{\ell=1}^{\infty} \ell c_\ell^2 \sum_{j \ne k} \langle \boldsymbol{a}_k, \boldsymbol{a}_j \rangle \langle \hat{\boldsymbol{b}}_k, \hat{\boldsymbol{b}}_j \rangle^{\ell-1} \epsilon^k_{\boldsymbol{m}} \hat{\boldsymbol{J}}'_k \hat{\boldsymbol{b}}_j \right] \\ &=: (\nabla^1_{err})_{k,:} + (\nabla^2_{err})_{k,:}, \end{aligned} \tag{65}$$

where $\left|\epsilon^k_{\boldsymbol{m}}\right| \le C_X^2 \frac{\log^2(d)}{\sqrt{d}}$ and $\nabla^1_{err}, \nabla^2_{err}$ are the matrices corresponding to the first and second expectation, respectively. Using this notation, proving the lemma is equivalent to showing that

$$\nabla^1_{err} + \nabla^2_{err} = O\left(C_X^7 \frac{\log^2(d)}{\sqrt{d}}\right). \tag{66}$$

We will start with bounding $\left\|(\nabla^1_{err})_{k,:}\right\|_{op}$. By the definition of $(\nabla_{\boldsymbol{B}})_{k,:} - (\nabla'_{\boldsymbol{B}})_{k,:})$, we have the following simple bound:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{m}} \left\|\mathbf{1}_A((\nabla_{\boldsymbol{B}})_{k,:} - (\nabla'_{\boldsymbol{B}})_{k,:})\right\| \le C\frac{1}{d^\gamma} \max_{\boldsymbol{m}} \Bigg\| &-2\frac{1}{\sqrt{p}}\boldsymbol{m} \circ (\hat{\boldsymbol{J}}_k - \hat{\boldsymbol{J}}'_k)\boldsymbol{a}_k \\ &+ 2\sum_{l=1}^{\infty} \ell c_\ell^2 \sum_{j \neq k} \langle \boldsymbol{a}_k, \boldsymbol{a}_j \rangle \langle \hat{\boldsymbol{b}}_k, \hat{\boldsymbol{b}}_j \rangle^{\ell-1} (\hat{\boldsymbol{J}}_k - \hat{\boldsymbol{J}}'_k)\hat{\boldsymbol{b}}_j \Bigg\|.\end{aligned} \tag{67}$$

Note that

$$\left\|\hat{\boldsymbol{J}}_k - \hat{\boldsymbol{J}}'_k\right\|_{op} = \left\|\left(\frac{1}{\|\bar{\boldsymbol{b}}_k\|} - \frac{1}{\sqrt{p}}\right)\left(\boldsymbol{I} - \hat{\boldsymbol{b}}_k \hat{\boldsymbol{b}}_k^\top\right)\right\|_{op} \le \left(\frac{\sqrt{p}}{\delta} + 1\right)\left\|\hat{\boldsymbol{J}}'_k\right\|_{op}. \tag{68}$$

Furthermore, since by definition $\left\|\hat{\boldsymbol{b}}_j\right\| = \left\|\hat{\boldsymbol{b}}_k\right\| = 1$,

$$p\left\|\hat{\boldsymbol{J}}'_k \hat{\boldsymbol{b}}_j\right\|^2 = \left\|\hat{\boldsymbol{b}}_j - \hat{\boldsymbol{b}}_k \langle \hat{\boldsymbol{b}}_k, \hat{\boldsymbol{b}}_j \rangle\right\|^2 = 1 - \langle \hat{\boldsymbol{b}}_k, \hat{\boldsymbol{b}}_j \rangle^2 \le 2\left(1 - \left|\langle \hat{\boldsymbol{b}}_k, \hat{\boldsymbol{b}}_j \rangle\right|\right). \tag{69}$$

We clearly have

$$\left\|\boldsymbol{m} \circ (\hat{\boldsymbol{J}}_k - \hat{\boldsymbol{J}}'_k)\boldsymbol{a}_k\right\| \le \left(\frac{\sqrt{p}}{\delta} + 1\right)\left\|\hat{\boldsymbol{J}}'_k\right\|_{op} \|\boldsymbol{a}_k\| \le C(1 + \frac{1}{\delta}), \tag{70}$$

where we have used (68) and the fact that masking reduces the norm.

By Lemma A.5, we have

$$\begin{aligned}\left\|\sum_{\ell=1}^{\infty} \ell c_\ell^2 \sum_{j \neq k} \langle \boldsymbol{a}_k, \boldsymbol{a}_j \rangle \langle \hat{\boldsymbol{b}}_k, \hat{\boldsymbol{b}}_j \rangle^{\ell-1} (\hat{\boldsymbol{J}}_k - \hat{\boldsymbol{J}}'_k)\hat{\boldsymbol{b}}_j\right\| &\le C(1 + \frac{1}{\delta}) \left\|(\mathbf{1}\mathbf{1}^\top - \boldsymbol{I}) \circ \boldsymbol{A}^\top \boldsymbol{A}\right\|_{max} \sum_{\ell=1}^{\infty} \ell c_\ell^2 \sum_{j \neq k} \left|\langle \hat{\boldsymbol{b}}_k, \hat{\boldsymbol{b}}_j \rangle^{\ell-1}\right| \left\|\hat{\boldsymbol{J}}'_k \hat{\boldsymbol{b}}_j\right\| \\ &\le C(1 + \frac{1}{\delta}) \|\boldsymbol{A}\|^2_{op} \sum_{j \neq k} \frac{1}{\sqrt{1 - \left|\langle \hat{\boldsymbol{b}}_k, \hat{\boldsymbol{b}}_j \rangle\right|}} \left\|\hat{\boldsymbol{J}}'_k \hat{\boldsymbol{b}}_j\right\| \\ &\le C(1 + \frac{1}{\delta})d,\end{aligned} \tag{71}$$

where the last step follows from (69). Now combining (70) and (71) we can bound the RHS of (67) by

$$C\frac{1}{d^\gamma} \max_{\boldsymbol{m}} \left\|-2\frac{1}{\sqrt{p}}\boldsymbol{m} \circ (\hat{\boldsymbol{J}}_k - \hat{\boldsymbol{J}}'_k)\boldsymbol{a}_k + 2\sum_{l=1}^{\infty} \ell c_\ell^2 \sum_{j \neq k} \langle \boldsymbol{a}_k, \boldsymbol{a}_j \rangle \langle \hat{\boldsymbol{b}}_k, \hat{\boldsymbol{b}}_j \rangle^{l-1} (\hat{\boldsymbol{J}}_k - \hat{\boldsymbol{J}}'_k)\hat{\boldsymbol{b}}_j\right\| \le CC_X^2(1 + \frac{1}{\delta})d^{-(\gamma-1)}. \tag{72}$$

From this and 2. in Lemma A.4, it follows that

$$\left\|\nabla^1_{err}\right\|_{op} \le C(1 + \frac{1}{\delta})d^{-(\gamma-1)}\sqrt{d}, \tag{73}$$

Now by choosing $\gamma = 3 + \gamma_\delta$ the RHS is of of order than $O\left(d^{-3/2}\right)$, which finishes bounding $\left\|\nabla^1_{err}\right\|_{op}$.

For $\left\|\nabla^2_{err}\right\|_{op}$ we need a more nuanced approach. We will break this term in three different parts, $\nabla^2_{err} = -\frac{2}{\sqrt{p}}\boldsymbol{M}_1 + 2\boldsymbol{M}_2 + 2\boldsymbol{M}_3$, in (74), (76), (78) below. First we consider

$$(\boldsymbol{M}_1)_{k,:} := \boldsymbol{m} \circ \epsilon^k_{\boldsymbol{m}} \hat{\boldsymbol{J}}'_k \boldsymbol{a}_k = \epsilon^k_{\boldsymbol{m}} \boldsymbol{m} \circ (\boldsymbol{a}_k - \hat{\boldsymbol{b}}_k \langle \hat{\boldsymbol{b}}_k, \boldsymbol{a}_k \rangle), \tag{74}$$

so defining $(\boldsymbol{D}_\epsilon)_{k,k} := \epsilon_{\boldsymbol{m}}^k$ can write

$$\boldsymbol{M}_1 = \boldsymbol{D}_\epsilon(\boldsymbol{A}_{\boldsymbol{m}} - \mathrm{Diag}\left(\hat{\boldsymbol{B}}\boldsymbol{A}\right)\hat{\boldsymbol{B}}_{\boldsymbol{m}}).$$

By 1. in Lemma A.4, we can bound

$$\left\|\boldsymbol{M}_1\right\|_{op} \leq \left\|\boldsymbol{D}_\epsilon\right\|_{op}\left(\left\|\boldsymbol{A}\right\|_{op} + \left\|\hat{\boldsymbol{B}}\right\|_{op}^2 \left\|\boldsymbol{A}\right\|_{op}\right) \leq CC_X^2\frac{\log^2(d)}{\sqrt{d}}\left(1 + \left\|\hat{\boldsymbol{B}}\right\|_{op}^2\right).$$

By Lemma A.12, we have

$$\mathbb{E}_{\boldsymbol{m}}\left\|\hat{\boldsymbol{B}}\right\|_{op}^2 \leq \frac{1}{p}\left\|\bar{\boldsymbol{B}}\right\|_{op}^2 + CC_X^3\frac{\log^3(d)}{\sqrt{d}} \leq CC_X^2 + CC_X^3\frac{\log^3(d)}{\sqrt{d}},$$

which gives us

$$\mathbb{E}_{\boldsymbol{m}}\left\|\boldsymbol{M}_1\right\|_{op} \leq CC_X^7\frac{\log^2(d)}{\sqrt{d}}. \tag{75}$$

Next, we consider the term

$$(\boldsymbol{M}_2)_{k,:} := \epsilon_{\boldsymbol{m}}^k \sum_j \langle \boldsymbol{a}_k, \boldsymbol{a}_j\rangle \hat{\boldsymbol{J}}'_k\hat{\boldsymbol{b}}_j + 3c_3^2\langle \boldsymbol{a}_k, \boldsymbol{a}_j\rangle\langle\hat{\boldsymbol{b}}_k, \hat{\boldsymbol{b}}_j\rangle^2\hat{\boldsymbol{J}}'_k\hat{\boldsymbol{b}}_j, \tag{76}$$

which we can write as

$$\boldsymbol{M}_2 = \boldsymbol{D}_\epsilon\left(\boldsymbol{A}^\top\boldsymbol{A}\hat{\boldsymbol{B}} - \mathrm{Diag}\left(\boldsymbol{A}^\top\boldsymbol{A}\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top\right)\hat{\boldsymbol{B}} + 3c_3^2(\boldsymbol{A}^\top\boldsymbol{A})\circ(\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top - \boldsymbol{I})^{\circ 2}\hat{\boldsymbol{B}} - 3c_3^2\mathrm{Diag}(\boldsymbol{A}^\top\boldsymbol{A}(\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top - \boldsymbol{I})^{\circ 3})\hat{\boldsymbol{B}}\right).$$

One can verify that the RHS satisfies the assumption of Lemma A.12. Hence, the same reasoning as for $\boldsymbol{M}_1$ gives that

$$\mathbb{E}_{\boldsymbol{m}}\left\|\boldsymbol{M}_2\right\|_{op} \leq CC_X^7\frac{\log^2(d)}{\sqrt{d}}. \tag{77}$$

Lastly, define

$$(\boldsymbol{M}_3)_{k,:} = \epsilon_{\boldsymbol{m}}^k\sum_{\ell=5}^\infty \ell c_\ell^2\sum_j\langle \boldsymbol{a}_k, \boldsymbol{a}_j\rangle\langle\hat{\boldsymbol{b}}_k, \hat{\boldsymbol{b}}_j\rangle^{\ell-1}\hat{\boldsymbol{J}}'_k\hat{\boldsymbol{b}}_j. \tag{78}$$

Using Lemma A.5, we have

$$\begin{aligned}
\left\|\epsilon_{\boldsymbol{m}}^k\sum_{\ell=5}^\infty \ell c_\ell^2\sum_{j\neq k}\langle \boldsymbol{a}_k, \boldsymbol{a}_j\rangle\langle\hat{\boldsymbol{b}}_k, \hat{\boldsymbol{b}}_j\rangle^{\ell-1}\hat{\boldsymbol{J}}'_k\hat{\boldsymbol{b}}_j\right\| &\leq CC_X^2\frac{\log^2(d)}{\sqrt{d}}\left\|(\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I})\circ\boldsymbol{A}^\top\boldsymbol{A}\right\|_{max}\sum_{j\neq k}\sum_{\ell=5}^\infty \ell c_\ell^2\left|\langle\hat{\boldsymbol{b}}_k, \hat{\boldsymbol{b}}_j\rangle^{\ell-1}\right|\left\|\hat{\boldsymbol{J}}'_k\hat{\boldsymbol{b}}_j\right\| \\
&\leq CC_X^2\frac{\log^2(d)}{\sqrt{d}}\left\|(\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I})\circ\boldsymbol{A}^\top\boldsymbol{A}\right\|_{max}\sum_{j\neq k}\frac{\langle\hat{\boldsymbol{b}}_k, \hat{\boldsymbol{b}}_j\rangle^4}{\sqrt{1 - \left|\langle\hat{\boldsymbol{b}}_k, \hat{\boldsymbol{b}}_j\rangle\right|}}\left\|\hat{\boldsymbol{J}}'_k\hat{\boldsymbol{b}}_j\right\| \\
&\leq CC_X^2\frac{\log^2(d)}{\sqrt{d}}d\left\|(\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I})\circ\boldsymbol{A}^\top\boldsymbol{A}\right\|_{max}\left\|(\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I})\circ\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top\right\|_{max}^4.
\end{aligned} \tag{79}$$

Note that

$$\left\|(\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I})\circ\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top\right\|_{max}^4 = \left\|(\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I})\circ\hat{\boldsymbol{D}}\bar{\boldsymbol{B}}\bar{\boldsymbol{B}}^\top\hat{\boldsymbol{D}}\right\|_{max}^4 \leq \min\{\left\|\hat{\boldsymbol{D}}\right\|_{op}^8\left\|(\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I})\circ\bar{\boldsymbol{B}}\bar{\boldsymbol{B}}^\top\right\|_{max}^4, 1\}.$$

Thus, by using 4. in Lemma A.11, with probability at least $1 - C\frac{1}{d^2}$ in $\boldsymbol{U}, \boldsymbol{V}$, we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{m}}\left\|(\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I})\circ\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top\right\|_{max}^4 &\leq \mathbb{P}\left(\left\|\hat{\boldsymbol{D}}\right\|_{op} > \frac{2}{\sqrt{p}}\right) + \left(\frac{2}{\sqrt{p}}\right)^8\left(CC_X^2\frac{\log^3(d)}{\sqrt{d}} + O\left(C_X\left\|\boldsymbol{R}\right\|_{op}\right)\right)^4 \\
&\leq Cd^{-\gamma} + C\left(C_X^2\frac{\log^3(d)}{\sqrt{d}} + C_XC_R\frac{\log^3(d)}{\sqrt{d}}\right)^4 \\
&\leq C\frac{1}{d^{\frac{3}{2}}}.
\end{aligned} \tag{80}$$

Next, note that under the assumptions of the current lemma we can apply both Lemma A.14 and 3. in Lemma A.11 to obtain

$$\left\|(\mathbf{1}\mathbf{1}^\top - \boldsymbol{I}) \circ \boldsymbol{A}^\top \boldsymbol{A}\right\|_{max} \leq C_X^7 \frac{\log^{10}(d)}{\sqrt{d}} + C_R \frac{\log^{\alpha_R}(d)}{\sqrt{d}}. \tag{81}$$

Combining (79), (80), (81), we obtain

$$\mathbb{E}_{\boldsymbol{m}} \left\|(\boldsymbol{M}_3)_{k,:}\right\|^2 \leq C C_X^4 \frac{\log(d)^4}{d} d^2 (C_X^7 \log^{10}(d) + C_R \log^{\alpha_R}(d))^2 \frac{1}{d} \frac{1}{d^3} \leq C \frac{1}{d^{\frac{5}{2}}}.$$

This now gives us

$$\mathbb{E}_{\boldsymbol{m}} \left\|\boldsymbol{M}_3\right\|_{op} \leq \mathbb{E}_{\boldsymbol{m}} \sqrt{\sum_k \left\|(\boldsymbol{M}_3)_{k,:}\right\|_{op}^2} \leq \sqrt{\sum_k \mathbb{E}_{\boldsymbol{m}} \left\|(\boldsymbol{M}_3)_{k,:}\right\|_{op}^2} \leq C \frac{1}{d^{\frac{3}{4}}}, \tag{82}$$

where we have used Jensen's inequality in the second step. Finally combining (73), (75), (77), (82) we can conclude. □

Lemma A.16 (Gradient concentration, Part 2)**.** *Assume we have $\boldsymbol{B} = \boldsymbol{X} + \boldsymbol{R}$ with unit norm rows, $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top, \boldsymbol{U}, \boldsymbol{V}$ Haar, $\operatorname{Tr}\left[\boldsymbol{S}\boldsymbol{S}^\top\right] = n, \|\boldsymbol{X}\|_{op} \leq C_X$, $\|\boldsymbol{X}\|_{max} \leq C C_X \frac{\log(d)}{\sqrt{d}}$, $\|\boldsymbol{R}\|_{op} \leq C_R \frac{\log^{\alpha_R}(d)}{\sqrt{d}}$. Further assume that $\min_{i,j} |\boldsymbol{B}_{i,j}| \geq \delta > d^{-\gamma_\delta}$ for some $\gamma_\delta > 0$. Then, for $d > d_0(C_X, C_R, \gamma_\delta)$ with probability at least $1 - C\frac{1}{d^2}$ in $\boldsymbol{U}, \boldsymbol{V}$*

$$\frac{1}{2}\nabla_{\boldsymbol{B}} = -\alpha \left(\boldsymbol{B}\boldsymbol{B}^\top + \alpha \boldsymbol{I}\right)^{-2} \boldsymbol{B} + \alpha \frac{1}{n} \operatorname{Tr}\left[\left(\boldsymbol{B}\boldsymbol{B}^\top + \alpha \boldsymbol{I}\right)^{-2} \boldsymbol{B}\boldsymbol{B}^\top\right] \boldsymbol{B} + O\left(C_X^{10} \frac{\log^{10}(d)}{\sqrt{d}}\right) \tag{83}$$

$$= -\alpha \left(\boldsymbol{X}\boldsymbol{X}^\top + \alpha \boldsymbol{I}\right)^{-2} \boldsymbol{X} + \alpha \frac{1}{n} \operatorname{Tr}\left[\left(\boldsymbol{X}\boldsymbol{X}^\top + \alpha \boldsymbol{I}\right)^{-2} \boldsymbol{X}\boldsymbol{X}^\top\right] \boldsymbol{X} + O\left(\boldsymbol{R}\right) + O\left(C_X^{10} \frac{\log^{10}(d)}{\sqrt{d}}\right), \tag{84}$$

where $\nabla_{\boldsymbol{B}}$ was defined in (30)*.*

Proof. By Lemma A.15, we may assume that, up to an error of order $O\left(C_X^7 \frac{\log^2(d)}{\sqrt{d}}\right)$, the gradient is given by (60), (61) and (62).

We will start by analysing the first part of the gradient in (61) which we restate here for convenience:

$$\frac{1}{2}\nabla^1_{\hat{\boldsymbol{B}}} = -\boldsymbol{A}^\top + \frac{1}{\sqrt{p}} \boldsymbol{A}^\top \boldsymbol{A} \hat{\boldsymbol{B}} + \frac{1}{p} \operatorname{Diag}(\hat{\boldsymbol{B}} \boldsymbol{A}) \hat{\boldsymbol{B}} - \frac{1}{\sqrt{p}} \operatorname{Diag}(\boldsymbol{A}^\top \boldsymbol{A} \hat{\boldsymbol{B}} \hat{\boldsymbol{B}}^\top) \hat{\boldsymbol{B}}. \tag{85}$$

By Lemma (A.14), we have with probability at least $1 - C\frac{1}{d^2}$ in $\boldsymbol{U}, \boldsymbol{V}$

$$\boldsymbol{A} = \boldsymbol{B}^\top \left(\boldsymbol{B}\boldsymbol{B}^\top + \alpha \boldsymbol{I}\right)^{-1} + O\left(C_X^7 \frac{\log^{10}(d)}{\sqrt{d}}\right)$$

$$= \boldsymbol{X}^\top \left(\boldsymbol{X}\boldsymbol{X}^\top + \alpha \boldsymbol{I}\right)^{-1} + O\left(C_X^7 \frac{\log^{10}(d)}{\sqrt{d}}\right) + O\left(\boldsymbol{R}\right),$$

where the expectation over $\boldsymbol{m}$ has not been taken yet. Using 1. in Lemma A.13, we see that the RHS in (85) satisfies the assumptions of Lemma A.12 (noting that Ω is the entire space for 1.), so we have

$$\mathbb{E}_{\boldsymbol{m}} \frac{1}{2}\nabla^1_{\hat{\boldsymbol{B}}} = \mathbb{E}_{\boldsymbol{m}} - \boldsymbol{A}^\top + \frac{1}{p} \boldsymbol{A}^\top \boldsymbol{A} \bar{\boldsymbol{B}} + \frac{1}{p} \operatorname{Diag}(\frac{1}{p} \bar{\boldsymbol{B}} \boldsymbol{A}) \bar{\boldsymbol{B}} - \frac{1}{p} \operatorname{Diag}(\boldsymbol{A}^\top \boldsymbol{A} \frac{1}{p} \bar{\boldsymbol{B}} \bar{\boldsymbol{B}}^\top) \bar{\boldsymbol{B}} + O\left(C_X^{10} \frac{\log^{10}(d)}{\sqrt{d}}\right).$$

We now estimate $\mathbb{E}_{\boldsymbol{m}} \frac{1}{2}\nabla^1_{\hat{\boldsymbol{B}}}$. We clearly have

$$\mathbb{E}_{\boldsymbol{m}} - \boldsymbol{A}^\top + \frac{1}{p} \boldsymbol{A}^\top \boldsymbol{A} \bar{\boldsymbol{B}} = -\boldsymbol{A}^\top + \boldsymbol{A}^\top \boldsymbol{A} \boldsymbol{B}.$$

For the third term we have by 5. in Lemma A.11 that, with probability at least $1 - C\frac{1}{d^2}$ in $\boldsymbol{m}$,

$$\operatorname{Diag}(\frac{1}{p}\bar{\boldsymbol{B}}\boldsymbol{A}) = \beta\boldsymbol{I} + O\left(C_X^8\frac{\log^{10}(d)}{\sqrt{d}}\right) + O\left(C_X\boldsymbol{R}\right),$$

where $\beta = \frac{1}{n}\operatorname{Tr}\left[\boldsymbol{B}\boldsymbol{A}\right]$, which implies that

$$\mathbb{E}_{\boldsymbol{m}}\frac{1}{p}\operatorname{Diag}(\frac{1}{p}\bar{\boldsymbol{B}}\boldsymbol{A})\bar{\boldsymbol{B}} = \beta\boldsymbol{B} + O\left(C_X^9\frac{\log^{10}(d)}{\sqrt{d}}\right) + O\left(C_X^2\boldsymbol{R}\right).$$

By exactly the same argument, we can use 6. in Lemma A.11 and obtain

$$\mathbb{E}_{\boldsymbol{m}}\operatorname{Diag}(\boldsymbol{A}^\top\boldsymbol{A}\frac{1}{p}\bar{\boldsymbol{B}}\bar{\boldsymbol{B}}^\top)\bar{\boldsymbol{B}} = \tilde{\beta}\boldsymbol{B} + O\left(C_X^{10}\frac{\log^{10}(d)}{\sqrt{d}}\right) + O\left(C_X^3\boldsymbol{R}\right),$$

where $\tilde{\beta} = \frac{1}{n}\operatorname{Tr}\left[\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{B}\boldsymbol{B}^\top\right]$.

In total we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{m}}\frac{1}{2}\nabla^1_{\hat{\boldsymbol{B}}} &= -\boldsymbol{A}^\top + \boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{B} + \beta\boldsymbol{B} - \tilde{\beta}\boldsymbol{B} + O\left(C_X^{10}\frac{\log^{10}(d)}{\sqrt{d}}\right) + O\left(C_X^3\boldsymbol{R}\right)\\ &= \left(-\boldsymbol{A}^\top + \boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{X} + \beta\boldsymbol{X} - \tilde{\beta}\boldsymbol{X}\right) + O\left(C_X^{10}\frac{\log^{10}(d)}{\sqrt{d}}\right) + O\left(C_X^3\boldsymbol{R}\right).\end{aligned}$$

Plugging in $\boldsymbol{A} = \boldsymbol{B}^\top\left(\boldsymbol{B}\boldsymbol{B}^\top + \alpha\boldsymbol{I}\right)^{-1} + O\left(C_X^7\frac{\log^{10}(d)}{\sqrt{d}}\right)$ in the second term and $\boldsymbol{A} = \boldsymbol{X}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top + \alpha\boldsymbol{I}\right)^{-1} + O\left(C_X^7\frac{\log^{10}(d)}{\sqrt{d}}\right) + O\left(\boldsymbol{R}\right)$ in the third term, we obtain the leading order terms for (83).

To see that this also implies (84) note that $\beta = \frac{1}{n}\operatorname{Tr}\left[\boldsymbol{B}\boldsymbol{A}\right] = \frac{1}{n}\operatorname{Tr}\left[\boldsymbol{X}\boldsymbol{A}\right] + O\left(\boldsymbol{R}\right)$ and $\tilde{\beta} = \frac{1}{n}\operatorname{Tr}\left[\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{B}\boldsymbol{B}^\top\right] = \frac{1}{n}\operatorname{Tr}\left[\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{X}\boldsymbol{X}^\top\right] + O\left(C_X\boldsymbol{R}\right)$.

It remains to show that the higher order terms are small. Here we will not need to distinguish between the two approximations of $\boldsymbol{A}$. The remaining part of the gradient in (62) is given by

$$\sum_{\ell\geq 3} c_\ell^2\ell\nabla^\ell_{\hat{\boldsymbol{B}}},$$

where

$$\frac{1}{2}\nabla^\ell_{\hat{\boldsymbol{B}}} = \frac{1}{\sqrt{p}}(\boldsymbol{A}^\top\boldsymbol{A})\circ(\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top - \boldsymbol{I})^{\circ(\ell-1)}\hat{\boldsymbol{B}} - \frac{1}{\sqrt{p}}\operatorname{Diag}(\boldsymbol{A}^\top\boldsymbol{A}(\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top - \boldsymbol{I})^{\circ\ell})\hat{\boldsymbol{B}}. \tag{86}$$

Let $\Omega = \{\boldsymbol{M}|(\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I})\circ(\boldsymbol{M}\boldsymbol{M}^\top) = \boldsymbol{Y} + \boldsymbol{Z}, \|\boldsymbol{Y}\|_{max} \leq CC_X^2\frac{\log^3(d)}{\sqrt{d}}, \|\boldsymbol{Z}\|_{op} \leq CC_XC_R\frac{\log^{\alpha_R}(d)}{\sqrt{d}}\}$. Then, by 4. in Lemma A.11, $\mathbb{P}\left(\bar{\boldsymbol{B}}\in\Omega\right) \geq 1 - C\frac{1}{d^2}$. Thus, by Lemma A.13, we have

$$\begin{aligned}\frac{1}{2}\sum_{\ell\geq 3}c_\ell^2\ell\nabla^\ell_{\hat{\boldsymbol{B}}} = &\frac{1}{\sqrt{p}}\sum_{\ell\geq 3}c_\ell^2\ell\mathbb{1}_{\{\bar{\boldsymbol{B}}\in\Omega\}}\\ &\cdot\left((\boldsymbol{A}^\top\boldsymbol{A})\circ\left(\left(\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I}\right)\circ\frac{1}{p}\bar{\boldsymbol{B}}\bar{\boldsymbol{B}}^\top\right)^{\circ(\ell-1)} - \operatorname{Diag}\left(\boldsymbol{A}^\top\boldsymbol{A}\left(\left(\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I}\right)\circ\frac{1}{p}\bar{\boldsymbol{B}}\bar{\boldsymbol{B}}^\top\right)^{\circ\ell}\right)\right)\frac{1}{\sqrt{p}}\bar{\boldsymbol{B}}\\ &+ O\left(C_X^3\frac{\log^3(d)}{\sqrt{d}}\right).\end{aligned} \tag{87}$$

We will now individually bound the different terms. In the following we always assume $\ell \geq 3$. We first analyse the term

$$(\boldsymbol{A}^\top\boldsymbol{A})\circ\left(\left(\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I}\right)\circ\frac{1}{p}\bar{\boldsymbol{B}}\bar{\boldsymbol{B}}^\top\right)^{\circ(\ell-1)}.$$

We had previously derived the following in (56)

$$\left(\frac{1}{p}(\mathbf{1}\mathbf{1}^\top - \boldsymbol{I}) \circ \bar{\boldsymbol{B}}\bar{\boldsymbol{B}}^\top\right)^{\circ\ell-1} = \frac{1}{p^{\ell-1}}\boldsymbol{Y}^{\circ(\ell-1)} + \frac{1}{p^{\ell-1}}\ell\boldsymbol{Y}^{\circ(\ell-2)} \circ \left(\boldsymbol{Z} + e^\ell O\left(\boldsymbol{Z}^2\right)\right). \tag{88}$$

Thus, as in 3. of Lemma A.13 we obtain from Lemma A.14

$$\boldsymbol{A}^\top\boldsymbol{A} = O_{max}\left(\frac{\log(d)}{\sqrt{d}}\right) + O\left(C_X^7 C_R \frac{\log^{\alpha_R}(d)}{\sqrt{d}}\right),$$

so

$$\begin{aligned}\left\|(\boldsymbol{A}^\top\boldsymbol{A}) \circ \left(\frac{1}{p}(\mathbf{1}\mathbf{1}^\top - \boldsymbol{I}) \circ \bar{\boldsymbol{B}}\bar{\boldsymbol{B}}^\top\right)^{\circ(\ell-1)}\right\|_{op} &\le \left\|O_{max}\left(\frac{\log(d)}{\sqrt{d}}\right) \circ \left(\frac{1}{p}(\mathbf{1}\mathbf{1}^\top - \boldsymbol{I}) \circ \bar{\boldsymbol{B}}\bar{\boldsymbol{B}}^\top\right)^{\circ(\ell-1)}\right\|_{op} \\ &+ \left\|O\left(C_X^7 C_R \frac{\log^{\alpha_R}(d)}{\sqrt{d}}\right) \circ \left(\frac{1}{p}(\mathbf{1}\mathbf{1}^\top - \boldsymbol{I}) \circ \bar{\boldsymbol{B}}\bar{\boldsymbol{B}}^\top\right)^{\circ(\ell-1)}\right\|_{op}.\end{aligned} \tag{89}$$

Plugging in (88) and using 3. and 5. in Lemma A.4 we obtain

$$\begin{aligned}&\ell\left\|O_{max}\left(\frac{\log(d)}{\sqrt{d}}\right) \circ \left(\frac{1}{p}(\mathbf{1}\mathbf{1}^\top - \boldsymbol{I}) \circ \bar{\boldsymbol{B}}\bar{\boldsymbol{B}}^\top\right)^{\circ(\ell-1)}\right\|_{op} \le \\ &\le C\frac{\ell}{p^{\ell-1}}\left(n\frac{\log(d)}{\sqrt{d}}\left\|\boldsymbol{Y}^{\circ(\ell-1)}\right\|_{max} + \ell e^\ell \sqrt{n}\left\|\boldsymbol{Z}\right\|_{op}\frac{\log(d)}{\sqrt{d}}\left\|\boldsymbol{Y}^{\circ(\ell-2)}\right\|_{max}\right) \\ &\le C\frac{\ell}{p^{\ell-1}}\left(C_X^{2(\ell-1)}\frac{\log^{1+3(\ell-1)}(d)}{d^{(\ell-2)/2}} + CC_X^{1+2(\ell-2)}C_R e^\ell \ell \frac{\log^{\alpha_R}(d)\log^{1+3(\ell-2)}(d)}{d^{(\ell-1)/2}}\right) \\ &\le CC_X^4\frac{\log^7(d)}{\sqrt{d}}.\end{aligned} \tag{90}$$

Similarly, we have

$$\begin{aligned}&\ell\left\|O\left(C_X^7 C_R \frac{\log^{\alpha_R}(d)}{\sqrt{d}}\right) \circ \left(\frac{1}{p}(\mathbf{1}\mathbf{1}^\top - \boldsymbol{I}) \circ \bar{\boldsymbol{B}}\bar{\boldsymbol{B}}^\top\right)^{\circ(\ell-1)}\right\|_{op} \le \\ &\le C\frac{\ell}{p^{\ell-1}}\left(\sqrt{n}C_X^7 C_R\frac{\log^{\alpha_R}(d)}{\sqrt{d}}\left\|\boldsymbol{Y}^{\circ(\ell-1)}\right\|_{max} + \ell\sqrt{n}C_X^7 C_R\frac{\log^{\alpha_R}(d)}{\sqrt{d}}\left\|\boldsymbol{Z} \circ \boldsymbol{Y}^{\circ(\ell-2)}\right\|_{max}\right) \\ &\le C\frac{\ell}{p^{\ell-1}}\left(C_X^{7+2(\ell-1)}C_R\frac{\log^{\alpha_R}(d)\log^{3(\ell-1)}(d)}{d^{(\ell-1)/2}} + CC_X^{7+2(\ell-1)}C_R\ell\frac{(\log^{\alpha_R}(d))^2\log^{3(\ell-1)}(d)}{d^{(\ell-1)/2}}\right) \\ &\le \frac{1}{\sqrt{d}}.\end{aligned} \tag{91}$$

Next we have

$$\left\|\mathrm{Diag}\left(\boldsymbol{A}^\top\boldsymbol{A}\left((\mathbf{1}\mathbf{1}^\top - \boldsymbol{I}) \circ \frac{1}{p}\bar{\boldsymbol{B}}\bar{\boldsymbol{B}}^\top\right)^{\circ\ell}\right)\right\|_{op} \le C\left\|\left((\mathbf{1}\mathbf{1}^\top - \boldsymbol{I}) \circ \frac{1}{p}\bar{\boldsymbol{B}}\bar{\boldsymbol{B}}^\top\right)^{\circ\ell}\right\|_{op}.$$

Now exactly as in the proof of (57) we obtain

$$C\ell\left\|\left((\mathbf{1}\mathbf{1}^\top - \boldsymbol{I}) \circ \frac{1}{p}\bar{\boldsymbol{B}}\bar{\boldsymbol{B}}^\top\right)^{\circ\ell}\right\|_{op} \le CC_X^6\frac{\log^{10}(d)}{\sqrt{d}}. \tag{92}$$

(Note that when writing out the proof, the ℓ factor is trivially absorbed for $\ell \ge 4$.)

Finally, we can combine (90), (91), (92) to obtain that the RHS of (87) is of order $O(C_X^7 \frac{\log^{10}(d)}{\sqrt{d}})$, where we get an extra C_X from bounding the operator norm of $\bar{\boldsymbol{B}}$. Thus, using that $\sum_{\ell \geq 3} c_\ell^2 < \infty$, we conclude

$$\sum_{\ell \geq 3} c_\ell^2 \ell \nabla_{\hat{\boldsymbol{B}}}^\ell = O\left(C_X^7 \frac{\log^{10}(d)}{\sqrt{d}} \right),$$

which finishes the proof.

□

A.6. GD-analysis and reduction to Gaussian

To simplify the notation we will push the time dependence in the subscript, i.e. $\boldsymbol{B}_t = \boldsymbol{B}(t)$.

Theorem A.17 (Gaussian recursion)**.** *If the entries $(\boldsymbol{B}_0')_{i,j} \sim \mathcal{N}(0, \frac{1}{d})$ are i.i.d., $\boldsymbol{B}_0 = \mathrm{proj}(\boldsymbol{B}_0')$ and*

$$\nabla_{\boldsymbol{B}} \boldsymbol{B}_t^\top = -\alpha \left(\boldsymbol{B}_t \boldsymbol{B}_t^\top + \alpha \boldsymbol{I} \right)^{-2} \boldsymbol{B}_t \boldsymbol{B}_t^\top + \alpha \mathrm{Diag}\left(\left(\boldsymbol{B}_t \boldsymbol{B}_t^\top + \alpha \boldsymbol{I} \right)^{-2} \boldsymbol{B}_t \boldsymbol{B}_t^\top \right) \boldsymbol{B}_t \boldsymbol{B}_t^\top + \tilde{\boldsymbol{E}}_t, \tag{93}$$

$$\boldsymbol{B}_t \boldsymbol{B}_t^\top = \boldsymbol{I} + \boldsymbol{Z}_t + \boldsymbol{E}_t, \tag{94}$$

with $\boldsymbol{Z}_t = \boldsymbol{U}(\boldsymbol{\Lambda}_t - \boldsymbol{I})\boldsymbol{U}^\top$, $\boldsymbol{U}$ a Haar matrix and

$$\left\| \tilde{\boldsymbol{E}}^t \right\|_{op} \leq C_E \left(\frac{\mathrm{poly}_E(\log(\mathrm{d}))}{\sqrt{d}} \cdot \|\boldsymbol{Z}_t\|_{op}^{1/2} + \|\boldsymbol{E}_t\|_{op}^2 + \|\boldsymbol{E}_t\|_{op} \|\boldsymbol{Z}_t\|_{op}^{1/2} \right).$$

Consider the GD-min algorithm in (26) without noise ($\boldsymbol{G}_t = \boldsymbol{0}$ for all t) and on the Gaussian objective (i.e., $p = 1$). Pick a learning rate $\eta = C/\sqrt{d}$. Then, with probability at least $1 - C\exp(-cd)$, we have that, jointly for all $t \geq 0$,

$$\begin{aligned} \|\boldsymbol{E}_t\|_{op} &\leq C_E e^{-c\eta t} . \frac{\mathrm{poly}_E(\log(\mathrm{d}))}{\sqrt{d}}, \\ \|\boldsymbol{Z}_t\|_{op} &\leq C_Z e^{-c\eta t}. \end{aligned} \tag{95}$$

Proof. The claim follows from the analysis in Appendix E of (Shevchenko et al., 2023). First, note that here $\tilde{\boldsymbol{E}}_t$ and $\boldsymbol{E}_t$ respectively correspond to $\boldsymbol{E}_t$ and $\boldsymbol{X}_t$ in (90) in (Shevchenko et al., 2023). Then, the assumptions of our theorem correspond to the conclusion of Lemma E.4 and Lemma E.5. The projection step is handled in Lemma E.6 and the recursion is analysed in Lemma E.7. □

Lemma A.18 (Reduction to Gaussian recursion)**.** *Fix $t_{\max} = T_{\max}/\eta, T_{\max} \in (0, \infty)$, let $(\boldsymbol{B}_0')_{i,j} \sim \mathcal{N}(0, \frac{1}{\sqrt{d}})$ i.i.d., $\boldsymbol{B}_0 = \mathrm{proj}(\boldsymbol{B}_0')$ and assume that for $t \leq t_{\max}$ we have*

$$\nabla_{\boldsymbol{B}} \boldsymbol{B}_t^\top + \boldsymbol{G}_t = -\alpha \left(\boldsymbol{B}_t \boldsymbol{B}_t^\top + \alpha \boldsymbol{I} \right)^{-2} \boldsymbol{B}_t \boldsymbol{B}_t^\top + \alpha \mathrm{Diag}\left(\left(\boldsymbol{B}_t \boldsymbol{B}_t^\top + \alpha \boldsymbol{I} \right)^{-2} \boldsymbol{B}_t \boldsymbol{B}_t^\top \right) \boldsymbol{B}_t \boldsymbol{B}_t^\top + O\left(C_R(T_{\max}) \frac{\log^{\alpha_R}(d)}{\sqrt{d}} \right). \tag{96}$$

Consider the GD-min algorithm in (26) for any $p \in (0, 1)$. Pick a learning rate $\eta = C/\sqrt{d}$. Then, with probability at least $1 - C\exp(-cd)$, we have that, jointly for all $t \geq 0$, (94) holds with

$$\begin{aligned} \|\boldsymbol{E}_t\|_{op} &\leq C_E e^{-c\eta t} . \frac{\mathrm{poly}_E(\log(\mathrm{d}))}{\sqrt{d}}, \\ \|\boldsymbol{Z}_t\|_{op} &\leq C_Z e^{-c\eta t}, \end{aligned} \tag{97}$$

where crucially $C_E, \mathrm{poly}_E, C_Z$ are independent of d, and C_Z is independent of $T_{\max}$.

Proof. The claim follows from Theorem A.17 after showing that

$$C_R(T_{\max}) \frac{\log^{\alpha_R}(d)}{\sqrt{d}} \leq C_E \left(\frac{\mathrm{poly}_E(\log(\mathrm{d}))}{\sqrt{d}} \cdot \|\boldsymbol{Z}_t\|_{op}^{1/2} + \|\boldsymbol{E}_t\|_{op}^2 + \|\boldsymbol{E}_t\|_{op} \|\boldsymbol{Z}_t\|_{op}^{1/2} \right), \tag{98}$$

where $\|\boldsymbol{E}_t\|_{op}, \|\boldsymbol{Z}_t\|_{op}$ satisfy (95). Now, (98) trivially holds if $\|\boldsymbol{Z}_t\|_{op} \geq c_Z(T_{\max}) > 0$, where $c_Z(T_{\max})$ is independent of d.

It remains to show the lower bound on $\|\boldsymbol{Z}_t\|_{op}$. This can be readily seen by analyzing the deterministic recursion of the spectrum of $\boldsymbol{Z}$ as in Lemma G.3 in (Shevchenko et al., 2023). First, for sufficiently large d, η gets arbitrarily small, hence we can approximate such discrete recursion with its continouous analogue. Next, we linearize the continuous evolution since $\boldsymbol{Z}$ is small (otherwise we already have the desired lower bound). Since the coefficient of the linearization is strictly negative (and, hence, bounded away from 0), we readily have that $\|\boldsymbol{Z}_t\|_{op}$ cannot reach 0 in finite time. □

For technical reasons, we need the following lemma that shows that the spectrum of $\boldsymbol{B}$ a priori cannot grow faster than exponentially in the effective time of the dynamics. The proof is a non-tight analog of the analysis done in Lemma E.7 and G.3 in (Shevchenko et al., 2023) for $\boldsymbol{B}$ instead of $\boldsymbol{B}\boldsymbol{B}^\top$.

Lemma A.19 (Spectrum evolution of $\boldsymbol{B}$)**.** *Consider the GD-min algorithm in* (26) *for any* $p \in (0, 1)$. *Pick a learning rate* $\eta = C/\sqrt{d}$. *Under the gradient approximation given in* (84) *with* $C_X(t) := \exp(C\eta t)\|\boldsymbol{B}_0\|_{op}$, *we have that, for* $t \leq t_{\max}$ *and* $d > d_0(C_X(t_{\max}))$,

$$\boldsymbol{B}_t = \boldsymbol{X}_t + \boldsymbol{R}_t,$$

where $\boldsymbol{X}_t$ *has the same singular vectors as* $\boldsymbol{B}_0$*,*

$$\|\boldsymbol{X}_t\|_{op} \leq C_X(t), \quad \text{and} \quad \|\boldsymbol{R}_t\|_{op} \leq CC_X^7(t_{\max})\exp(CT_{\max})\frac{\log^{10}(d)}{\sqrt{d}},$$

with probability at least $1 - C(\eta t_{\max})\frac{1}{d^{3/2}}$.

Proof. Consider the recursion where the gradient is given below:

$$\frac{1}{2}\nabla_{\boldsymbol{B}} := -\alpha\left(\boldsymbol{X}\boldsymbol{X}^\top + \alpha\boldsymbol{I}\right)^{-2}\boldsymbol{X} + \alpha\frac{1}{n}\text{Tr}\left[\left(\boldsymbol{X}\boldsymbol{X}^\top + \alpha\boldsymbol{I}\right)^{-2}\boldsymbol{X}\boldsymbol{X}^\top\right]\boldsymbol{X}. \tag{99}$$

It is evident that this recursion only updates the singular values s^i of $\boldsymbol{B}$ as the RHS has the same singular vectors as $\boldsymbol{B}$. Furthermore, the update equation for the s^i is given by

$$s^i_{t+1} = s^i_t - \eta\left(-\alpha\frac{s^i}{(\alpha + (s^i_t)^2)^2} + s^i_t\frac{1}{n}\sum_{i=1}^n\frac{(s^i_t)^2}{(\alpha + (s^i_t)^2)^2}\right).$$

Note that

$$\left|s^i_{t+1} - s^i_t\right| \leq C\eta\left|s^i_t\right|.$$

Thus, letting $b_t = \|\boldsymbol{B}_t\|_{op}$, the above implies that

$$b_{t+1} \leq (1 + C\eta)b_t, \tag{100}$$

which by monotonicity gives that $b_t \leq (1 + C\eta)^t b_0$. Hence, if the recursion of the gradient was actually given by (99), the claim would immediately follow.

Now, the recursion of the gradient is given by (84). Thus, to deal with the error, we can follow the strategy of Lemma E.7 in (Shevchenko et al., 2023). In particular, denoting $r_t := \|\boldsymbol{R}_t\|_{op}, \epsilon_d := C_X^{10}\frac{\log^{10}(d)}{\sqrt{d}}$, the evolution of the error is given by

$$r_{t+1} \leq (1 + C_1\eta)r_t + C_2\epsilon_d.$$

By monotonicity, this recursion is upper bounded by the solution of

$$r_{t+1} = (1 + C_1\eta)r_t + C_2\epsilon_d.$$

Since the recursion is initialized with $r_0 = 0$, we can unroll it as

$$
\begin{aligned}
r_{t+1} &= C_2 \sum_{i=1}^{t} (1 + C_1\eta)^i \eta\epsilon_d \\
&\le C_2 \sum_{i=1}^{t} \exp\left(C_1\eta i\right)\eta\epsilon_d \\
&= C_2 \exp\left(C_1\eta t\right) \sum_{i=1}^{t} \exp\left(-C_1\eta(t-i)\right)\eta\epsilon_d \\
&\le C_2 \exp\left(C_1\eta t\right) \frac{1}{1-\exp\left(-C_1\eta\right)} \eta\epsilon_d,
\end{aligned}
$$

where we have used $1 + x \le \exp\left(x\right)$. For small enough η, we have $\frac{1}{1-\exp(-C_1\eta)} \le \frac{1}{C_1\eta}$. Hence,

$$
r_{t+1} \le \frac{C_2}{C_1} \exp\left(C_1\eta t\right)\epsilon_d, \tag{101}
$$

which gives that $\left\|\boldsymbol{R}_t\right\|_{op} \le C_X^{10}\frac{C_2}{C_1}\exp\left(C_1\eta t\right)\frac{\log^{10}(d)}{\sqrt{d}}$, as required. Hence, by (84), b_t is upper bounded by the solution to the recursion

$$
b_{t+1} = (1 + C_1\eta) b_t + C_2\eta \exp\left(C_3\eta t\right)\epsilon_d.
$$

As $\operatorname{Tr}\left[\boldsymbol{B}\boldsymbol{B}^\top\right] = n$, we have that $b_t \ge 1$. Thus,

$$
b_{t+1} \le \left(1 + \left(C_1 + C_2\exp\left(C_3\eta t\right)\epsilon_d\right)\eta\right) b_t \le \left(1 + \left(C_1 + C_2\exp\left(C_3 T_{\max}\right)\epsilon_d\right)\eta\right) b_t.
$$

Taking a sufficiently large d gives that $C_2\exp\left(C_3 T_{\max}\right)\epsilon_d \le C_1$, which leads to

$$
b_{t+1} \le (1 + 2C_1\eta) b_t.
$$

Using again monotonicity and $1 + x \le \exp\left(x\right)$, we conclude that $C_X := \left\|\boldsymbol{B}_{t_{\max}}\right\|_{op} \le \exp\left(2C_1\eta t\right)\left\|\boldsymbol{B}_0\right\|_{op}$.

This proves the claim of the lemma for a gradient recursion given exactly by (84). We note that the GD-min algorithm in (26) has two additional steps: *(i)* adding noise $\boldsymbol{G}_t$ at each step, and *(ii)* the projections step, which normalizes the rows of $\boldsymbol{B}_t$ after the gradient update.

As for *(i)*, let $\boldsymbol{G}$ be an $n \times d$ matrix with i.i.d. $\mathcal{N}(0, \sigma^2)$ entries. Then, by Theorem 4.4.5 in (Vershynin, 2018) (with $t = \sqrt{d}$), we have that $\left\|\boldsymbol{G}\right\|_{op} \le C\sigma(\sqrt{n} + \sqrt{d}) \le C\sigma\sqrt{d}$ with probability at least $1 - C\frac{1}{d^2}$. Recall that in (26) we assume that $\sigma \le C\frac{1}{d}$. Hence, the additional error from the noise is of higher order than all the other error terms and can be neglected. By a union bound over $T_{\max}/\eta$ steps, the above bound holds for all time steps with probability at least $1 - C\frac{1}{d^{3/2}}$.

As for *(ii)*, a straightforward analysis shows that $\left\|\operatorname{proj}(\boldsymbol{B}') - \boldsymbol{B}'\right\|_{op} \le C\eta^2\left\|\nabla_{\boldsymbol{B}} + \boldsymbol{G}\right\|_{op}^2$, which is also of higher order. We skip the details here and refer to Lemma E.6 in (Shevchenko et al., 2023). This concludes the proof. □

We are now ready to give the proof of Theorem A.2 by combining the previous results and carrying out an induction over the time steps.

Proof of Theorem A.2. We fix $p \in (0, 1)$ and $t_{\max} = T_{\max}/\eta, T_{\max} \in (0, \infty)$. We want to show that the assumptions of Lemma A.18 are satisfied, as the conclusion of Lemma A.18 is precisely (27).

By Lemma A.19, we have that, with probability at least $1 - C(T_{\max})\frac{1}{d^{3/2}}$, for all $t \le t_{\max}$, $C_X(t) := \exp\left(C\eta t\right)\left\|\boldsymbol{B}_0\right\|_{op}$, and $C_R(t) = CC_X^{10}(t_{\max})\exp\left(C\eta t\right)\frac{\log^{\alpha_R}(d)}{\sqrt{d}}$. Furthermore, by choosing $\delta = d^{-(4+\gamma_g)}$, we can apply Lemma A.10 for each step so that, with probability at least $1 - C\frac{1}{d^{3/2}}$, $\min_{i,j}\left|\boldsymbol{B}'_{ij}\right| \ge 2\delta$. Note that the projection step does not change the scale of any entry by more than a factor that converges to 1 as d grows large (see Lemma E.6 in (Shevchenko et al., 2023) for details), so in particular $\min_{i,j}\left|\boldsymbol{B}'_{ij}\right| \ge 2\delta$ implies $\min_{i,j}\left|\boldsymbol{B}_{ij}\right| \ge \delta$. This gives that, with probability at least $1 - C(T_{\max})\frac{1}{d^{3/2}}$, for all $t \le t_{\max}$, the assumptions of Lemma A.16 are satisfied, hence (83) and (84) hold.

By 2. in Lemma A.11, at each step with probability $1 - C\frac{1}{d^2}$, we have that

$$\mathrm{Diag}\left(\left(\boldsymbol{B}\boldsymbol{B}^\top + \alpha\boldsymbol{I}\right)^{-2}\boldsymbol{B}\boldsymbol{B}^\top\right) = \frac{1}{n}\mathrm{Tr}\left[\left(\boldsymbol{B}\boldsymbol{B}^\top + \alpha\boldsymbol{I}\right)^{-2}\boldsymbol{B}\boldsymbol{B}^\top\right]\boldsymbol{I} + C\exp\left(CT_{\max}\right)\left\|\boldsymbol{B}_0\right\|_{op}\frac{\log(d)}{\sqrt{d}},$$

so this holds jointly for all $t \le t_{\max}$ with probability at least $1 - C\frac{1}{d^{\frac{3}{2}}}$ Combining this with (83), we conclude that, with probability at least $1 - C(T_{\max})\frac{1}{d^{\frac{3}{2}}}$, the assumptions of lemma A.18 hold, which immediately implies

$$\lim_{d\to\infty} \sup_{t\in[0,t_{\max}]} \|\boldsymbol{R}_t\| = 0$$

and

$$\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{I}\right\|_{op} \le C\exp\left(-cT_{\max}\right).$$

This proves (27).

To prove (28), we note that the combination of (52), (53) and (58) gives

$$\mathbb{E}_{\boldsymbol{m}}\left[\mathrm{Tr}\left[\boldsymbol{A}^\top\boldsymbol{A}f(\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top)\right] - \frac{2}{\sqrt{p}}\mathrm{Tr}\left[\boldsymbol{A}\hat{\boldsymbol{B}}\right]\right] = \mathrm{Tr}\left[\boldsymbol{A}^\top\boldsymbol{A}f(\boldsymbol{B}\boldsymbol{B}^\top)\right] - 2\mathrm{Tr}\left[\boldsymbol{A}\boldsymbol{B}^\top\right] + O\left(C(T_{\max})d\frac{\mathrm{poly}(\log(d))}{\sqrt{d}}\right). \tag{102}$$

Since (25) and (3) differ only by a constant and a factor $1/d$, the above implies that, for any $p \in (0,1)$, (3) is close to the Gaussian objective up to an error $C(T_{\max})\frac{\mathrm{poly}(\log(d))}{\sqrt{d}}$. The fact that the evolution of $\boldsymbol{B}$ matches the Gaussian case is also clear, since the gradient approximation in Lemma A.16 coincides with the Gaussian recursion in Theorem A.17. □

B. MSE characterizations

B.1. Proof of Proposition 4.2

Denote by $\boldsymbol{x}^1$ the first iterate of the RI-GAMP algorithm (Venkataramanan et al., 2022), as in (22). Then, by taking σ to be the sign, one can readily verify that

$$\boldsymbol{x}^1 = \boldsymbol{B}^\top\mathrm{sign}(\boldsymbol{B}\boldsymbol{x}).$$

Note that $\boldsymbol{B}$ is bi-rotationally invariant in law and, as $\boldsymbol{x}$ has i.i.d. components, its empirical distribution converges in Wasserstein-2 distance to a random variable whose law is that of the first component of $\boldsymbol{x}$, denoted by x_1. Therefore, the assumptions of Theorem 3.1 in (Venkataramanan et al., 2022) are satisfied. Hence, for any ψ pseudo-Lipschitz of order 2,[4] we have that, almost surely,

$$\lim_{d\to\infty}\frac{1}{d}\sum_{i=1}^{d}\psi((\boldsymbol{x}^1)_i,(\boldsymbol{x})_i) = \mathbb{E}[\psi(\mu x_1 + \sigma g, x_1)],$$

where $g \sim \mathcal{N}(0,1)$ is independent of x_1 and the state evolution parameters (μ, σ) for $r \le 1$ can be computed as

$$\mu = r\cdot\sqrt{\frac{2\kappa_2}{\pi}} = r\cdot\sqrt{\frac{2}{\pi}}, \quad \sigma^2 = r\cdot\left(\kappa_2 + \kappa_4\cdot\frac{2}{\pi\kappa_2}\right) = r\cdot\left(1 - r\cdot\frac{2}{\pi}\right), \tag{103}$$

that is equation (11) in (Venkataramanan et al., 2022). Here, $\{\kappa_{2k}\}_{k\in\mathbb{N}}$ denote the rectangular free cumulants of the constant random variable equal to 1 (since all the singular values of $\boldsymbol{B}$ are equal to 1 by assumption). Noting that $\psi(x,y) = (x - \alpha\cdot y)^2$ is pseudo-Lipschitz of order 2, we get that, almost surely,

$$\lim_{d\to\infty}\frac{1}{d}\cdot\|\boldsymbol{x} - \alpha\cdot\boldsymbol{B}^\top\mathrm{sign}(\boldsymbol{B}^\top\boldsymbol{x})\|_2^2 = \mathbb{E}_{x_1,g}[|x_1 - \alpha(\mu x_1 + \sigma g)|_2^2],$$

which implies that

$$\lim_{d\to\infty}\frac{1}{d}\cdot\mathbb{E}_{\boldsymbol{x}}\|\boldsymbol{x} - \alpha\cdot\boldsymbol{B}^\top\mathrm{sign}(\boldsymbol{B}^\top\boldsymbol{x})\|_2^2 = \mathbb{E}_{x_1,g}[|x_1 - \alpha(\mu x_1 + \sigma g)|_2^2].$$

[4] We recall that $\psi : \mathbb{R}^2 \to \mathbb{R}$ is pseudo-Lipschitz of order 2 if, for all $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^2$, $|\psi(\boldsymbol{a}) - \psi(\boldsymbol{b})| \le L\|\boldsymbol{a} - \boldsymbol{b}\|_2(1 + \|\boldsymbol{a}\|_2 + \|\boldsymbol{b}\|_2)$ for some constant $L > 0$.

By expanding the RHS of the last equation and using that x_1 has unit second moment by assumption, we get

$$\begin{aligned}\mathbb{E}_{x_1,g}[|x_1-\alpha(\mu x_1+\sigma g)|_2^2] &= (1-\alpha\mu)^2\cdot\mathbb{E}[x_1^2]+\alpha^2\sigma^2\cdot\mathbb{E}[g^2]=(1-\alpha\mu)^2+\alpha^2\sigma^2\\ &= 1-2\alpha\mu+\alpha^2(\mu^2+\sigma^2)=1-2\alpha\cdot r\sqrt{\frac{2}{\pi}}+\alpha^2 r.\end{aligned}$$

Thus, by minimizing over α, we have

$$\min_\alpha \mathbb{E}_{x_1,g}[|x-\alpha(\mu x+\sigma g)|_2^2]=1-\frac{2}{\pi}\cdot r,$$

which concludes the proof of (13).

To prove (14), a direct calculation gives

$$\begin{aligned}\frac{1}{d}\cdot\mathbb{E}_{\boldsymbol{x}}\left\|\boldsymbol{x}-\alpha\cdot\begin{bmatrix}\boldsymbol{I}_n\\ \boldsymbol{0}_{(d-n)\times n}\end{bmatrix}\mathrm{sign}([\boldsymbol{I}_n,\boldsymbol{0}_{n\times(d-n)}]\boldsymbol{x})\right\|_2^2 &= 1-r+r\cdot\mathbb{E}[(x_1-\alpha\mathrm{sign}(x_1))^2]\\ &= 1-r+r\cdot(\mathbb{E}[x_1^2]-2\alpha\cdot\mathbb{E}[|x_1|]+\alpha^2\cdot\mathbb{E}[\mathrm{sign}^2(x_1)])\\ &= 1-r+r\cdot(1-2\alpha\cdot\mathbb{E}[|x_1|]+\alpha^2)\\ &= 1+r\cdot(\alpha^2-2\alpha\cdot\mathbb{E}[|x_1|]).\end{aligned}$$

The RHS is minimized by $\alpha=\mathbb{E}[|x_1|]$, which gives

$$\min_\alpha\frac{1}{d}\cdot\mathbb{E}_{\boldsymbol{x}}\left\|\boldsymbol{x}-\alpha\cdot\begin{bmatrix}\boldsymbol{I}_n\\ \boldsymbol{0}_{(d-n)\times n}\end{bmatrix}\mathrm{sign}([\boldsymbol{I}_n,\boldsymbol{0}_{n\times(d-n)}]\boldsymbol{x})\right\|_2^2=1-r\cdot(\mathbb{E}[|x_1|])^2,$$

and the proof is complete. □

B.2. Proof of Proposition 5.1

Let $\hat{\boldsymbol{x}}^1$ be an iterate of the RI-GAMP algorithm (Venkataramanan et al., 2022), as in (22). Then, by taking σ to be the sign and $f_t=f$, one can readily verify that

$$\hat{\boldsymbol{x}}^1=f(\boldsymbol{B}^\top\mathrm{sign}(\boldsymbol{B}\boldsymbol{x})),$$

which is exactly the form of the autoencoder in (4) that we wish to analyze. Thus, as f is Lipschitz, the assumptions of Theorem 3.1 in (Venkataramanan et al., 2022) are satisfied and, following the same passages as in the proof of Proposition 4.2, we have

$$\lim_{d\to\infty}\frac{1}{d}\cdot\mathbb{E}_{\boldsymbol{x}}\|\boldsymbol{x}-f(\boldsymbol{B}^\top\mathrm{sign}(\boldsymbol{B}\boldsymbol{x}))\|_2^2=\mathbb{E}_{x_1,g}[|x_1-f(\mu x_1+\sigma g)|_2^2], \tag{104}$$

where x_1 is the first entry of $\boldsymbol{x}$, $g\sim\mathcal{N}(0,1)$ is independent of x_1, and (μ,σ) are given by (103) (which coincides with (19)). This concludes the proof. □

B.3. Proof of Proposition 5.2

A direct calculation gives

$$\frac{1}{d}\cdot\mathbb{E}_{\boldsymbol{x}}\left\|\boldsymbol{x}-f\left(\begin{bmatrix}\boldsymbol{I}_n\\ \boldsymbol{0}_{(d-n)\times n}\end{bmatrix}\mathrm{sign}([\boldsymbol{I}_n,\boldsymbol{0}_{n\times(d-n)}]\boldsymbol{x})\right)\right\|_2^2=(1-r)\cdot\mathbb{E}\left[(x_1-f(0))^2\right]+r\cdot\mathbb{E}\left[(x_1-f(\mathrm{sign}(x_1)))^2\right], \tag{105}$$

where x_1 is the first entry of $\boldsymbol{x}$. The first term in (105) is minimized when $f(0)=\mathbb{E}[x]=0$. Hence, we obtain that, at the optimum,

$$(1-r)\cdot\mathbb{E}\left[(x_1-f(0))^2\right]=1-r,$$

as $\mathbb{E}[x^2]=1$. As for the second term in (105), we rewrite

$$\mathbb{E}\left[(x_1-f(\mathrm{sign}(x_1)))^2\right]=\mu_{x_1}(\{0\})\cdot\frac{1}{2}\cdot(f(1)^2+f(-1)^2)+\mathbb{E}[\mathbb{1}_{x_1>0}(x_1-f(1))^2]+\mathbb{E}[\mathbb{1}_{x_1<0}(x_1-f(-1))^2], \tag{106}$$

where μ_{x_1} stands for the measure that corresponds to the distribution of x_1, and we use that $\mathrm{sign}(0)$ is a Rademacher random variable by convention. As the distribution of x_1 is the same as that of $-x_1$, (106) is minimized by taking $f(1) = -f(-1)$. Thus, we have that

$$\min_f (106) = \min_{u\in\mathbb{R}} \mathbb{E}[(x_1 - u \cdot \mathrm{sign}(x_1))^2].$$

The RHS of this last expression can be further rewritten as

$$\min_{u\in\mathbb{R}} \mathbb{E}[(x_1 - u \cdot \mathrm{sign}(x_1))^2] = \mathbb{E}[x_1^2] + \min_{u\in\mathbb{R}} \left\{u^2 - 2u \cdot \mathbb{E}|x_1|\right\} = 1 - (\mathbb{E}|x_1|)^2,$$

which concludes the proof. □

B.4. Computation of f^*

Sparse Gaussian. Using Bayes rule, the conditional expectation can be expressed as

$$\mathbb{E}[x|\mu x + \sigma g = y] = \frac{\mathbb{E}_x\left[x \cdot P(\mu x + \sigma g = y|x)\right]}{\mathbb{E}_x\left[P(\mu x + \sigma g = y|x)\right]} = \frac{\mathbb{E}_x\left[x \cdot P(\mu x + \sigma g = y|x)\right]}{P(\mu x + \sigma g = y)}. \tag{107}$$

Given that $x \sim \mathrm{SG}_1(p)$, with probability p we have that $\mu x + \sigma g \sim \mathcal{N}(0, \mu^2/p + \sigma^2)$ as $x \sim \mathcal{N}(0, 1/p)$, and with probability $(1-p)$ we have that $x = 0$, and, hence, $\mu x + \sigma g = \sigma g \sim \mathcal{N}(0, \sigma^2)$. Combining gives

$$P(\mu x + \sigma g = y) = p \cdot \frac{\sqrt{p}}{\sqrt{2\pi(\mu^2 + p\sigma^2)}} \cdot \exp\left(-\frac{py^2}{2(\mu^2 + p\sigma^2)}\right) + (1-p) \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{y^2}{2\sigma^2}\right).$$

Note that due to sparsity, we have that

$$\mathbb{E}_x\left[x \cdot P(\mu x + \sigma g = y|x)\right] = p \cdot \mathbb{E}_{x\sim\mathcal{N}(0,1/p)}\left[x \cdot P(\mu x + \sigma g = y|x)\right], \tag{108}$$

and, in this case, we conclude that

$$\mu x + \sigma g | x \sim \mathcal{N}(\mu x, \sigma^2).$$

Thus, the RHS of (108) is a Gaussian integral, which is straight-forward to calculate by "completing a square". The computation gives

$$\mathbb{E}_{x\sim\mathcal{N}(0,1/p)}\left[x \cdot P(\mu x + \sigma g = y|x)\right] = \sqrt{\frac{p}{2\pi}} \cdot \mu y \cdot \exp\left(-\frac{py^2}{2(\mu^2 + p\sigma^2)}\right) \cdot \frac{1}{(\mu^2 + p\sigma^2)^{3/2}}.$$

Note that, when $p = 1$, i.e., $\boldsymbol{x}$ is an isotropic Gaussian vector, f^* is just a rescaling by a constant factor, i.e., $f^*(y) = \mathrm{const}(\mu, \sigma) \cdot y$.

Sparse Laplace. The sparse Laplace distribution with sparsity level $(1-p)$ has the following law

$$(1-p) \cdot \delta_0 + p \cdot \sqrt{\frac{p}{2}} \cdot \exp\left(-\sqrt{2p} \cdot |x|\right), \tag{109}$$

where δ_0 stands for the delta distribution centered at 0. The scaling for different p is chosen to ensure a unit second moment.

First, we derive the expression for the conditional expectation for $p = 1$. For $p \neq 1$ we elaborate later how a simple change of variables allows to obtain closed-form expressions of the corresponding expectations via the case $p = 1$. For $p = 1$, the denominator in (107) is equivalent to

$$\int_{\mathbb{R}} p(x)p(\mu x + \sigma g = y|x)\mathrm{d}x = \frac{1}{\sqrt{4\pi\sigma^2}} \int_{\mathbb{R}} \exp\left(-\sqrt{2} \cdot |x|\right) \exp\left(-\frac{(y - \mu x)^2}{2\sigma^2}\right) \mathrm{d}x. \tag{110}$$

By considering two cases, i.e., $x < 0$ and $x \geq 0$, for the limits of integration and for each of them "completing a square", we obtain

$$\int_{\mathbb{R}_+} \exp\left(-\sqrt{2} \cdot x\right) \exp\left(-\frac{(y - \mu x)^2}{2\sigma^2}\right) \mathrm{d}x = \left(1 + \mathrm{erf}\left(\frac{\sqrt{2}\mu y - 2\sigma^2}{2\mu\sigma}\right)\right) \cdot \exp\left(\frac{\sigma^2 - \sqrt{2}\mu y}{\mu^2}\right) \cdot \sqrt{\frac{\pi}{2}} \cdot \frac{\sigma}{\mu},$$

$$\int_{\mathbb{R}_-} \exp\left(\sqrt{2} \cdot x\right) \exp\left(-\frac{(y - \mu x)^2}{2\sigma^2}\right) \mathrm{d}x = \mathrm{erfc}\left(\frac{\sqrt{2}\mu y + 2\sigma^2}{2\mu\sigma}\right) \cdot \exp\left(\frac{\sigma^2 + \sqrt{2}\mu y}{\mu^2}\right) \cdot \sqrt{\frac{\pi}{2}} \cdot \frac{\sigma}{\mu},$$

where $\mathrm{erf}(\cdot)$ stands for the Gaussian error function, and $\mathrm{erfc}(\cdot)$ for its complement. For the case of $p \neq 1$, we get that the RHS of (110) becomes

$$(1-p) \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{y^2}{2\sigma^2}\right) + p \cdot \sqrt{\frac{p}{4\pi\sigma^2}} \cdot \int_{\mathbb{R}} \exp\left(-\sqrt{2p} \cdot |x|\right) \exp\left(-\frac{(y-\mu x)^2}{2\sigma^2}\right) \mathrm{d}x.$$

The change in normalization constant of the second term is then trivial. For the integral itself, consider the change of variables $\tilde{x} = x \cdot \sqrt{p}$:

$$\begin{aligned}\int_{\mathbb{R}} \exp\left(-\sqrt{2p} \cdot |x|\right) \exp\left(-\frac{(y-\mu x)^2}{2\sigma^2}\right) \mathrm{d}x &= \frac{1}{\sqrt{p}} \cdot \int_{\mathbb{R}} \exp\left(-\sqrt{2} \cdot |\tilde{x}|\right) \exp\left(-\frac{(y - \frac{\mu}{\sqrt{p}} \cdot \tilde{x})^2}{2\sigma^2}\right) \mathrm{d}\tilde{x} \\ &= \frac{1}{\sqrt{p}} \cdot \int_{\mathbb{R}} \exp\left(-\sqrt{2} \cdot |\tilde{x}|\right) \exp\left(-\frac{(y - \tilde{\mu} \cdot \tilde{x})^2}{2\sigma^2}\right) \mathrm{d}\tilde{x},\end{aligned}$$

which is exactly the previous integral in (110) but with $\tilde{\mu} = \mu/\sqrt{p}$ and an additional scaling factor in front.

Consider the numerator of (107) for $p = 1$. For this case, the computation reduces to evaluating:

$$\int_{\mathbb{R}} x \cdot p(x) p(\mu x + \sigma g = y|x) \mathrm{d}x = \frac{1}{\sqrt{4\pi\sigma^2}} \int_{\mathbb{R}} x \cdot \exp\left(-\sqrt{2} \cdot |x|\right) \exp\left(-\frac{(y-\mu x)^2}{2\sigma^2}\right) \mathrm{d}x. \tag{111}$$

Reducing to cases again and "completing a square" gives

$$\begin{aligned}&\int_{\mathbb{R}_+} x \cdot \exp\left(-\sqrt{2} \cdot x\right) \exp\left(-\frac{(y-\mu x)^2}{2\sigma^2}\right) \mathrm{d}x \\ &= \exp\left(-\frac{y^2}{2\sigma^2}\right) \cdot \left[\frac{\sigma^2}{\mu^2} + \frac{\sqrt{\pi}\sigma \cdot (\sqrt{2}\mu y - 2\sigma^2) \cdot e^{\frac{(\mu y - \sqrt{2}\sigma^2)^2}{2\mu^2\sigma^2}} \cdot \left(1 + \mathrm{erf}\left(\frac{y}{\sqrt{2}\sigma} - \frac{\sigma}{\mu}\right)\right)}{2\mu^3}\right], \\ &\int_{\mathbb{R}_-} x \cdot \exp\left(\sqrt{2} \cdot x\right) \exp\left(-\frac{(y-\mu x)^2}{2\sigma^2}\right) \mathrm{d}x \\ &= \exp\left(-\frac{y^2}{2\sigma^2}\right) \cdot \left[-\frac{\sigma^2}{\mu^2} + \frac{\sqrt{\pi}\sigma \cdot (\sqrt{2}\mu y + 2\sigma^2) \cdot e^{\frac{(\mu y + \sqrt{2}\sigma^2)^2}{2\mu^2\sigma^2}} \cdot \mathrm{erfc}\left(\frac{y}{\sqrt{2}\sigma} + \frac{\sigma}{\mu}\right)}{2\mu^3}\right].\end{aligned}$$

The derivation for the case $p \neq 1$ can be obtained analogously, by noting that (111) in this case is written as

$$p \cdot \sqrt{\frac{p}{4\pi\sigma^2}} \cdot \int_{\mathbb{R}} x \cdot \exp\left(-\sqrt{2p} \cdot |x|\right) \exp\left(-\frac{(y-\mu x)^2}{2\sigma^2}\right) \mathrm{d}x.$$

Sparse Rademacher. The sparse Rademacher distribution with sparsity level $(1-p)$ has the following law

$$(1-p) \cdot \delta_0 + \frac{p}{2} \cdot \left(\delta_{1/\sqrt{p}} + \delta_{-1/\sqrt{p}}\right).$$

The denominator in (107) reduces to

$$(1-p) \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{y^2}{2\sigma^2}\right) + \frac{p}{2} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \left[\exp\left(-\frac{(y-\mu/\sqrt{p})^2}{2\sigma^2}\right) + \exp\left(-\frac{(y+\mu/\sqrt{p})^2}{2\sigma^2}\right)\right].$$

Moreover, it is easy to see that the enumerator of (107) reduces to

$$\frac{\sqrt{p}}{2} \cdot \left[\exp\left(-\frac{(y-\mu/\sqrt{p})^2}{2\sigma^2}\right) - \exp\left(-\frac{(y+\mu/\sqrt{p})^2}{2\sigma^2}\right)\right].$$

Numerical denoising. For the sparse Beta mixture in (114), the Gaussian mixture with variable aspect ratio in (115) and the sparse Gaussian mixture in (112), it is cumbersome to get a closed form expression for the optimal denoiser in (20), in order to compute the performance given the Haar design in (18). We, thus, employ a typical binning in conjunction with Monte-Carlo to estimate the value of the conditional expectation in (20).

C. Experimental details and additional numerical results

C.1. Sparse Gaussian: rate-distortion function

For sparse Gaussian data, one can compute *rate-distortion function*, which is the information-theoretically optimal MSE that can be achieved for a given compression rate. This is done via the Blahut-Arimoto algorithm (Arimoto, 1972) in Figure 9. We observe that as sparsity increase, the optimal MSE decreases, so the data is easier to compress.

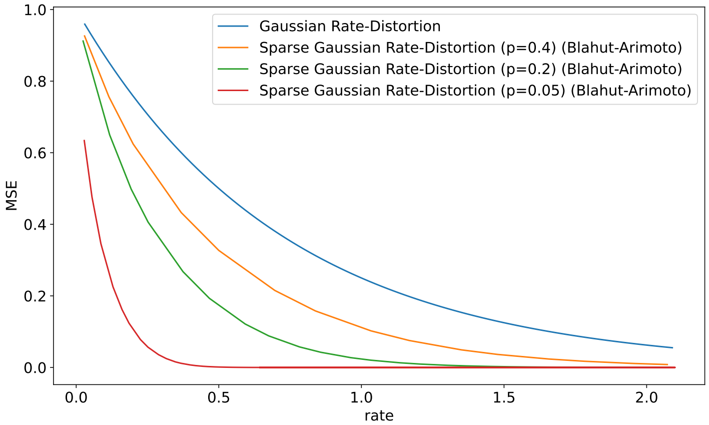

Figure 9. Numerical computation of the rate-distortion function for a sparse Gaussian source via the Blahut-Arimoto algorithm. We plot the optimal MSE against the rate r for different values of sparsity p.

C.2. Numerical setup

Activation function and reparameterization of the weight matrix $\boldsymbol{B}$. Since the sign activation has derivative zero almost everywhere, it is not directly suited for gradient-based optimization. To overcome this issue for SGD training of the models described in the main body, we use a "straight-through" (see for example (Yin et al., 2019)) approximation of it. In details, during the forward pass the activation of the network $\sigma(\cdot)$ is treated as a sign activation. However, during the backward pass (gradient computation) the derivatives are computed as if instead of $\sigma(\cdot)$ its relaxed version is used, namely, the tempered hyperbolic tangent:

$$\sigma_\tau(x) = \tanh\left(\frac{x}{\tau}\right).$$

We also note that such approximation is pointwise consistent except zero:

$$\lim_{\tau\to 0} = \sigma(x), \quad \forall\, \boldsymbol{x} \in \mathbb{R} \setminus \{0\}.$$

For the experiments we fix the temperature τ to the value of 0.1. Refining the approximation further, i.e., making τ smaller, does not affect the end result, but it makes numerics a bit less stable due to the increased variance of the derivative.

To ensure consistency of the "straight-though" approximation, we enforce the condition $\boldsymbol{B}_{i,:} \in \mathbb{S}^{d-1}$ via a simple differentiable reparameterization. Let $\boldsymbol{B} \in \mathbb{R}^{n\times d}$ be trainable network parameters, then

$$\hat{\boldsymbol{B}}_{i,:} = \frac{\boldsymbol{B}_{i,:}}{\|\boldsymbol{B}_{i,:}\|_2}.$$

It should be noted that it is not clear whether this constraint is necessary, since during the forward pass we use directly $\sigma(\cdot)$, which is agnostic to the row scaling of $\boldsymbol{B}$.

Augmentation and whitening. For the natural image experiments in Figures 3, 7 and 17, we use data augmentation to bring the amount of images per class to the initial dataset scale. This step is crucial to simulate the minimization of the population risk and not the empirical one, when the number of samples per class is insufficient. We augment each image 15 times for CIFAR-10 data and 10 times for MNIST data. We note that the described amount of augmentation is sufficient: increasing it further does not change the results of the numerical experiments and only increases computational cost.

The whitening procedure corresponds to the matrix multiplication of each image by the inverse square root of the empirical covariance of the data. This is done to ensure that the data is isotropic (to be closer to the i.i.d. data assumption needed for the theoretical analysis). More formally, let $\boldsymbol{X} \in \mathbb{R}^{\mathrm{n}_{\mathrm{samples}} \times d}$ be the augmented data that is centered, i.e., the data mean is subtracted. Its empirical covariance is then given by

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{\mathrm{n}_{\mathrm{samples}} - 1} \cdot \sum_{i=1}^{\mathrm{n}_{\mathrm{samples}}} \boldsymbol{X}_{i,:} \boldsymbol{X}_{i,:}^{\top}.$$

In this view, the whitened data $\hat{\boldsymbol{X}} \in \mathbb{R}^{\mathrm{n}_{\mathrm{samples}} \times d}$ is obtained from the initial data $\boldsymbol{X}$ as follows

$$\hat{\boldsymbol{X}}_{i,:} = \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}} \boldsymbol{X}_{i,:},$$

where $\boldsymbol{X}_{i,:}$ defines the i-th data sample.

C.3. Phase transition and staircase in the learning dynamics for the autoencoder in (2)

First, we provide an additional numerical simulation similar to the one in Figure 2 for the case of non-sparse Rademacher data, i.e., $p = 1$. Since condition (15) holds, we expect the minimizer to be a permutation of the identity, and the corresponding SGD dynamics to experience a staircase behaviour, as discussed in Section 4. Namely, the SGD algorithm first finds a random rotation that achieves Gaussian performance (indicated by the orange dashed line). Next, it searches a direction towards a sparse solution given by a permutation of the identity, and the corresponding loss remains at the plateau. Finally, the correct direction is found, and SGD quickly converges to the optimal solution.

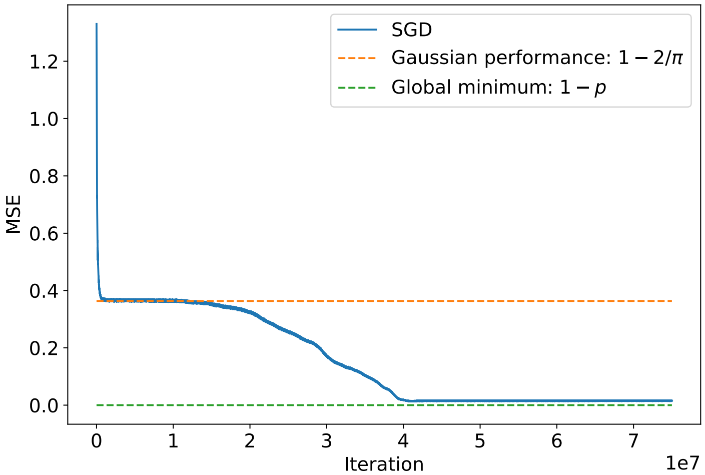

Figure 10. Compression of Rademacher data ($p = 1$) via the autoencoder in (2). We set $d = 200$ and $r = 1$. The MSE is plotted as a function of the number of iterations, and it displays a staircase behavior.

Sparse Gaussian mixture. Next, we consider the compression of $\boldsymbol{x}$ with i.i.d. components distributed according to the following sparse mixture of Gaussians:

$$x_i \sim p \cdot \left(\frac{1}{2} \cdot \mathcal{N}\left(1, \frac{1-p}{p}\right) + \frac{1}{2} \cdot \mathcal{N}\left(-1, \frac{1-p}{p}\right) \right) + (1-p) \cdot \delta_0. \tag{112}$$

It is easy to verify that $\mathbb{E}[x_i^2] = 1$. In order to compute the transition point we need to access the first absolute moment of x_i, i.e., $\mathbb{E}|x_i|$. Using the result in (Winkelbauer, 2012), we are able to claim that

$$\mathbb{E}_{x\sim\mathcal{N}(\pm 1,\sigma^2)}|x| = \sigma\sqrt{\frac{2}{\pi}} \cdot \Phi\left(-\frac{1}{2}, \frac{1}{2}, -\frac{1}{2\sigma^2}\right), \tag{113}$$

where $\Phi(a, b, c)$ stands for Kummer's confluent hypergeometric function:

$$\Phi(a, b, c) = \sum_{n=1}^{\infty} \frac{a^{\overline{n}}}{b^{\overline{n}}} \cdot \frac{c^n}{n!},$$

with $x^{\overline{n}}$ denoting the rising factorial, i.e.,

$$x^{\overline{n}} = z \cdot (z+1) \cdot \dots \cdot (z+n-1), \quad n \in \mathbb{N}_0.$$

We use `scipy.special.hyp1f1` to evaluate numerically $\Phi\left(-\frac{1}{2}, \frac{1}{2}, -\frac{1}{2\sigma^2}\right)$, where $\sigma^2 = (1-p)/p$. Likewise, to find p_{crit} at which $\mathbb{E}|x_i| = \sqrt{\frac{2}{\pi}}$ we rely on numerics. The results are presented in Figure 11.

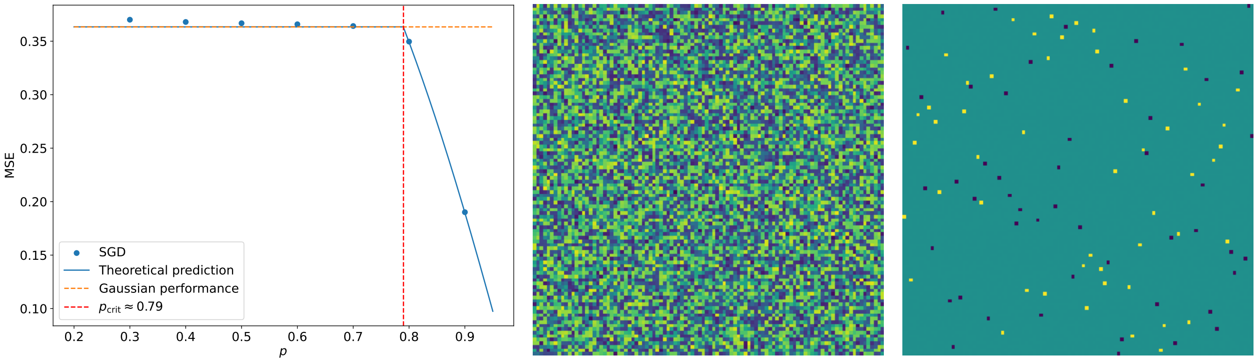

Figure 11. Compression of data whose distribution is given by a sparse mixture of Gaussians via the autoencoder in (2). We set $d = 100$ and $r = 1$. *Left.* MSE achieved by SGD at convergence, as a function of the sparsity level p. The empirical values (dots) match our theoretical prediction (blue line): for $p < p_{\text{crit}}$, the loss is equal to the value obtained for Gaussian data, i.e., $1 - 2r/\pi$; for $p > p_{\text{crit}}$, the loss is smaller, and it is equal to $1 - r \cdot (\mathbb{E}|x_1|)^2$. *Center.* Encoder matrix $\boldsymbol{B}$ at convergence of SGD when $p = 0.6 < p_{\text{crit}}$: the matrix is a random rotation. *Right.* Encoder matrix $\boldsymbol{B}$ at convergence of SGD when $p = 0.9 \geq p_{\text{crit}}$. The negative sign in part of the entries of $\boldsymbol{B}$ is cancelled by the corresponding sign in the entries of $\boldsymbol{A}$. Hence, $\boldsymbol{B}$ is equivalent to a permutation of the identity.

We remark that the first absolute moment can always be estimated via Monte-Carlo sampling if a functional expression such as (113) is out of reach. We also note that the behaviour of the predicted curve after the transition point p_{crit} can be arbitrary. In particular, it is not always linear like in the case of sparse Rademacher data in Figure 1. For instance, in the case of the sparse Gaussian mixture of Figure 11, the shape is clearly of non-linear nature.

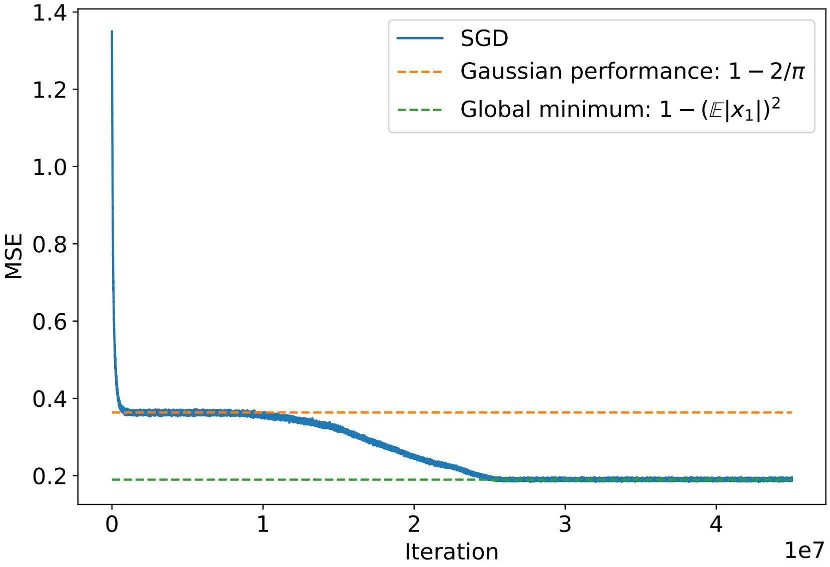

Figure 12. Compression of data whose distribution is given by a sparse mixture of Gaussians via the autoencoder in (2). We set $d = 100$, $r = 1$, and $p = 0.9$. The MSE is plotted as a function of the number of iterations and, as $p > p_{\text{crit}}$, it displays a staircase behavior.

In Figure 12, we provide an experiment similar to that of Figure 2, but for the compression of a sparse mixture of Gaussians with $p = 0.9$ at $r = 1$. We can clearly see that Figure 12 again indicates the emergent staircase behaviour of the SGD loss for $p > p_{\text{crit}}$.

Sparse Beta mixture. Next, we consider the compression of $\boldsymbol{x}$ with i.i.d. components distributed according to a sparse mixture of Beta distributions with sparsity $(1-p)$. The mixture is defined via the following sampling procedure:

$$\begin{aligned}&\hat{x}_i \sim \mathrm{Beta}(2,5),\\&\hat{x}_i \mapsto \mathrm{scale}\cdot\hat{m}_i\cdot\hat{x}_i, \quad \hat{m}_i \sim \mathrm{Rademacher}(0.5), \quad \text{(zero mean)},\end{aligned}$$

where scale is such that $\mathrm{Var}(\hat{x}_i) = 1$. The final step of sampling is the addition of sparsity:

$$x_i = \frac{1}{\sqrt{p}}\cdot\hat{x}_i\cdot m_i, \quad m_i \sim \mathrm{Bernoulli}(p), \tag{114}$$

where the $1/\sqrt{p}$ factor ensures $\mathrm{Var}(x_i) = 1$.

In this case, there is a phase transition at $p_{\text{crit}} \approx 0.835$: for $p < p_{\text{crit}}$, the condition in (15) is not satisfied and GD converges to Haar weights giving the Gaussian MSE; for $p > p_{\text{crit}}$, the condition in (15) is satisfied and GD converges to a sub-sampled permutation of the identity, which improves upon the Gaussian MSE. This is reported in Figure 13. To estimate the first absolute moment, i.e., $\mathbb{E}|x_1|$, we use a Monte-Carlo estimate over 10^7 samples.

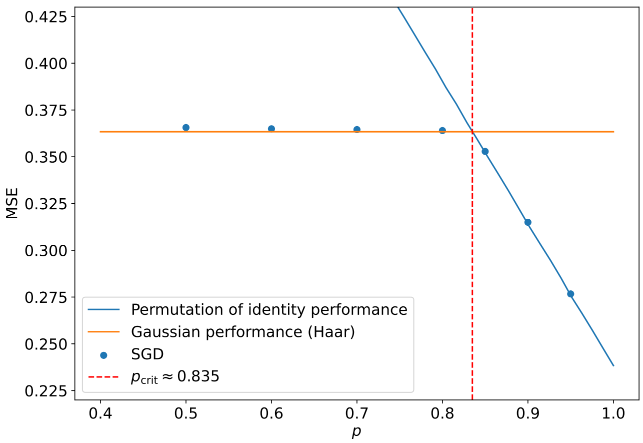

Figure 13. Compression of data whose distribution is given by a sparse mixture of Beta distributions via the autoencoder in (2). We set $d = 100$ and $r = 1$. We plot the MSE achieved by SGD at convergence, as a function of the sparsity level p. The empirical values (dots) match our theoretical prediction (blue/orange lines): for $p < p_{\text{crit}}$, the loss is equal to the value obtained for Gaussian data, i.e., $1 - 2r/\pi$; for $p > p_{\text{crit}}$, the loss is smaller, and it is equal to $1 - r \cdot (\mathbb{E}|x_1|)^2$.

Gaussian mixture with variable aspect ratio. Next, we consider the compression of $\boldsymbol{x}$ with i.i.d. components distributed according to a Gaussian mixture with varying aspect ratio γ. The mixture is defined via the following sampling procedure:

$$\begin{aligned} &\mu = 1, \quad \sigma = \mu \cdot \gamma \\ &x_i = \frac{1}{\sqrt{\mu^2 + \sigma^2}} \cdot m_i \cdot \hat{x}_i, \quad \hat{x}_i \sim \mathcal{N}(\mu, \sigma^2), \quad \hat{m}_i \sim \text{Rademacher}(0.5). \end{aligned} \tag{115}$$

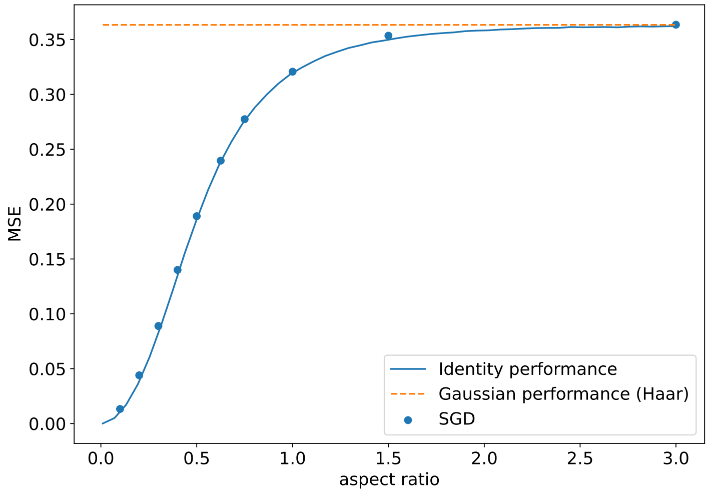

Figure 14. Compression of data whose distribution is given by a (non-sparse) mixture of Gaussians via the autoencoder in (2). We set $d = 200$ and $r = 1$. We plot the MSE achieved by SGD at convergence, as a function of the aspect ratio γ. The empirical values (dots) match our theoretical prediction (blue line): $1 - r \cdot (\mathbb{E}|x_1|)^2$, which corresponds to the minimizer given by a permutation of the identity.

Condition (15) is satisfied for all levels of γ and, as conjectured, SGD converges to a sub-sampled permutation of the identity, which improves upon the Gaussian MSE. This is reported in Figure 14.

Furthermore, the training loss exhibits a staircase behaviour: first the MSE rapidly converges to the Gaussian MSE (corresponding to Haar weights); then, there is a plateau; finally, the global minimum (corresponding to the permutation of identity weights) is reached. This is reported in Figure 15.

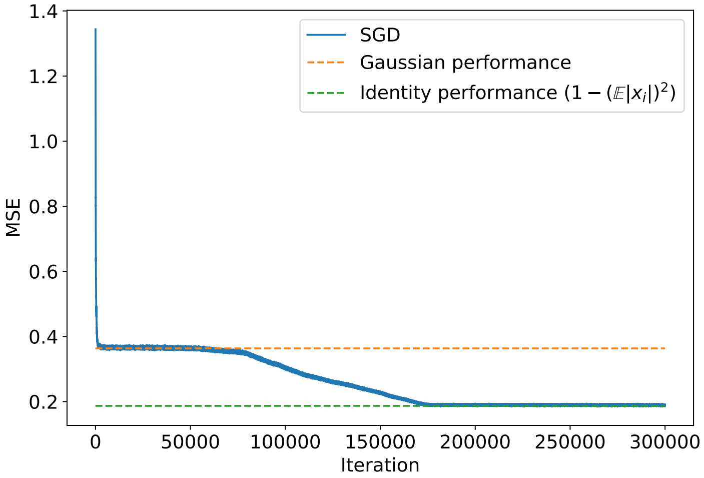

Figure 15. Compression of data whose distribution is given by (non-sparse) mixture of Gaussians via the autoencoder in (2). We set $d = 200$, $r = 1$, and aspect ratio $\gamma = 0.5$. The MSE is plotted as a function of the number of iterations and, as condition (15) is satisfied, it displays a staircase behavior.

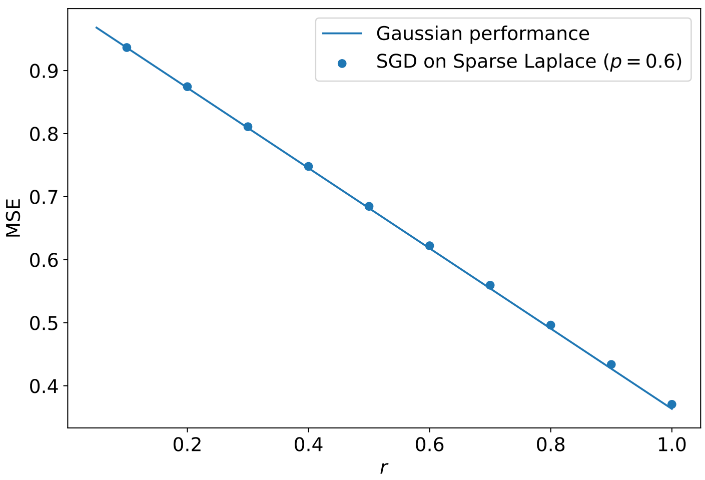

Figure 16. Compression of data whose distribution is given by a sparse Laplace distribution via the autoencoder in (2). We set $d = 400$ and $p = 0.6$. The MSE is plotted as a function of compression rate r and, as condition (15) is never satisfied, it displays Gaussian performance for all rates $r \leq 1$.

Sparse Laplace distribution. Next, we consider the compression of $\boldsymbol{x}$ with i.i.d. components distributed according to a sparse Laplace distribution (see (109)). In this case, the condition (15) is never met regardless of the sparsity level p. In other words, SGD will always converge to the Haar minimizer. We report the corresponding numerical values in Figure 16 for different compression rates r, $d = 400$ and $p = 0.6$.

C.4. MNIST experiment

In this subsection, we provide additional numerical evidence complementing the results presented in Figure 3. Namely, we provide a similar evaluation on Bernoulli-masked whitened MNIST data. As for the experiment in Figure 3, the sparsity level p is set to 0.7.

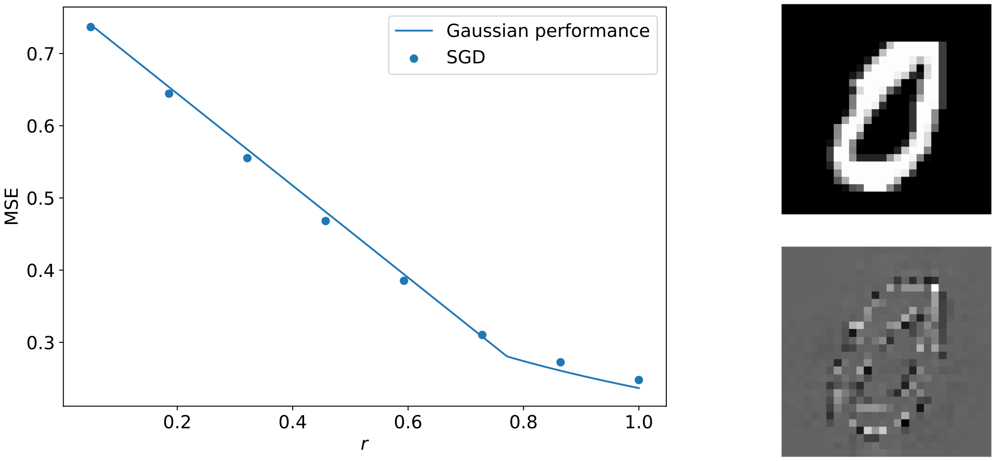

Figure 17. Compression of masked and whitened MNIST images that correspond to digit "zero" via the two-layer autoencoder in (2). First, the data is whitened so that it has identity covariance (as in the setting of Theorem 4.1). Then, the data is masked by setting each pixel independently to 0 with probability $p = 0.7$. An example of an original image is on the top right, and the corresponding masked and whitened image is on the bottom right. The SGD loss at convergence (dots) matches the solid line, which corresponds to the prediction in (5) for the compression of standard Gaussian data (with no sparsity).

Note that the eigen-decomposition of the covariance of MNIST data has zero eigenvalues. In this case, we need to apply the lower bound from (Shevchenko et al., 2023) that accounts for a degenerate spectrum. The corresponding result is stated in Theorem 5.2 of (Shevchenko et al., 2023). In particular, the number of zero eigenvalues n_0 is equal to 179, which means that at the value of the compression rate r given by

$$r = \frac{d - n_0}{d} = \frac{28^2 - 179}{28^2} \approx 0.77$$

the derivative of the lower bound experiences a jump-like behavior, as described in (Shevchenko et al., 2023).

C.5. CIFAR-10: Laplace approximation of pixel distribution

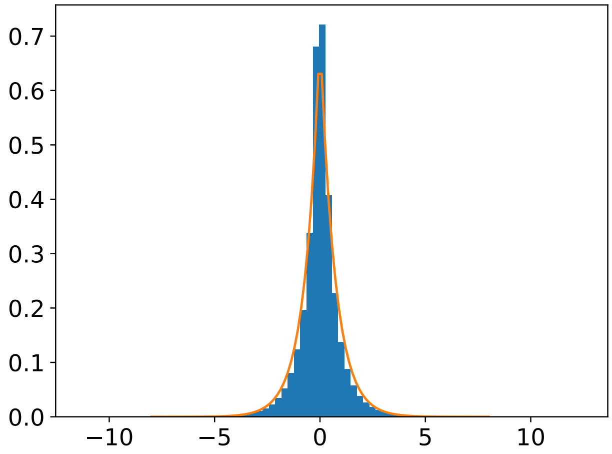

Figure 18. Empirical distribution of whitened CIFAR-10 image pixels (blue histogram), and its approximation via a Laplace distribution with unit second moment (orange curve).

Figure 18 demonstrates the quality of the Laplace approximation for whitened CIFAR-10 images. Namely, we note that the empirical distribution of the image pixels after whitening is well approximated by a Laplace random variable with unit second moment.

C.6. Provable benefit of nonlinearities for the compression of sparse Gaussian data

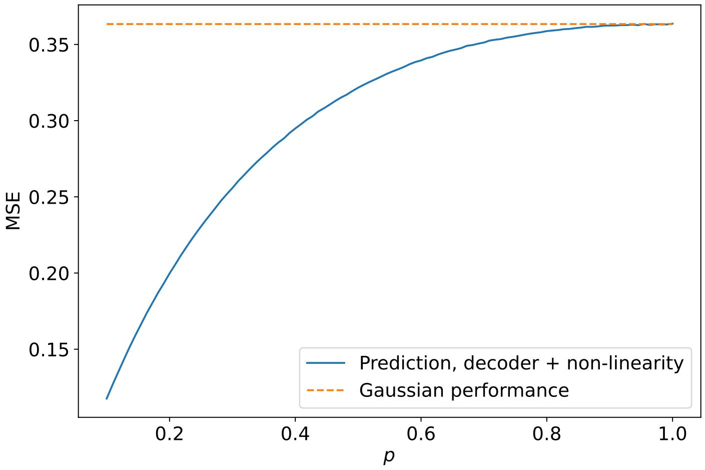

Figure 19. Compression of sparse Gaussian data. We set $r = 1$. The solid blue line corresponds to the MSE in (18) with $f = f^*$ (defined in (20)), for different values of p; the dashed orange line corresponds to the Gaussian performance in (5), which is achieved by the autoencoder in (2).

Figure 19 considers the compression of sparse Gaussian data, and it shows that the MSE achieved by the autoencoder in (4) with the optimal choice of f (namely, the RHS of (18) with $f = f^*$) is strictly lower than the MSE (5) achieved by the autoencoder in (2), for any sparsity level $p \in (0, 1)$. The conditional expectation $\mathbb{E}[x_1|\mu x_1 + \sigma g]$ (cf. the definition of f^* in (20)) is computed numerically via a Monte-Carlo approximation.

C.7. Phase transition and staircase in the learning dynamics for the autoencoder in (4)

Sparse Rademacher data. For sparse Rademacher data, the optimal f^* given by (20) is computed explicitly in Appendix B.4 and plotted in Figure 20. We note that functions of the form in (17) are unable to approximate f^* well. Thus, in the experiments we use a different parametric function for f given by the following mixture of hyperbolic tangents:

$$f(x) = \mathbb{1}_{x \geq 0} \cdot (\gamma_1 \cdot \tanh(\varepsilon_1 \cdot x - \alpha_1) + \beta_1) + \mathbb{1}_{x<0} \cdot (\gamma_2 \cdot \tanh(\varepsilon_2 \cdot x - \alpha_2) + \beta_2). \tag{116}$$

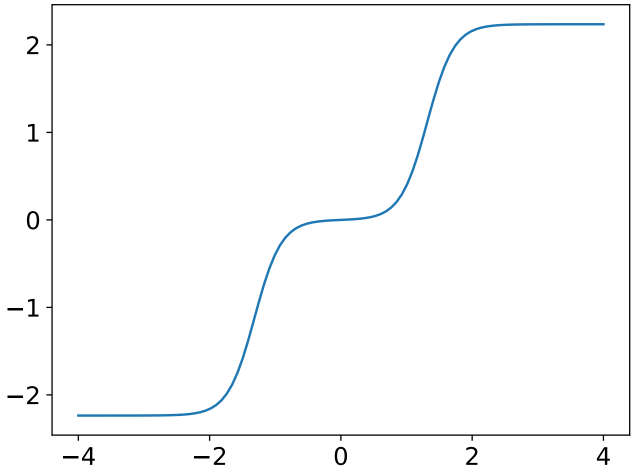

Figure 20. Optimal f^* in (20) when x_1 is a sparse Rademacher random variable. We set $r = 1$ and $p = 0.2$.

The numerical evaluation of the autoencoder in (4) with f of the form in (116) for the compression of sparse Rademacher data is provided in Figure 21. We set $r = 1$ and $d = 200$. The solid blue line corresponds to the prediction of Proposition

5.1, obtained for random Haar $\boldsymbol{B}$; the solid orange line corresponds to the prediction of Proposition 5.2, obtained for $\boldsymbol{B}$ equal to the identity. The blue dots correspond to the performance of SGD, and they exhibit the transition in the learnt $\boldsymbol{B}$ from a random Haar matrix ($p < p_{\rm crit}$) to a permutation of the identity ($p > p_{\rm crit}$). The critical value $p_{\rm crit}$ is obtained from the intersection between the blue curve and the orange curve. For all values of p, the autoencoder in (4) outperforms the Gaussian MSE (5) (green dashed line) and, hence, it is able to exploit the structure in the data.

For $p > p_{\rm crit}$, the staircase behavior of the SGD training dynamics is presented in Figure 22.

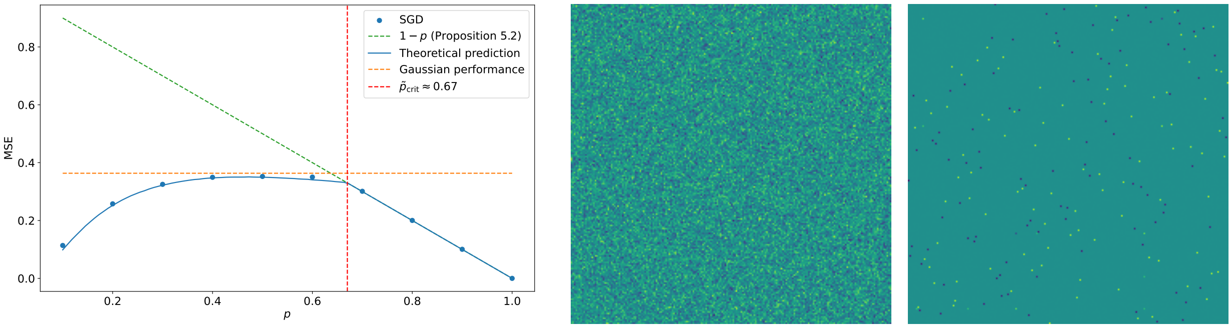

Figure 21. Compression of sparse Rademacher data via the autoencoder in (4) with f of the form in (116). We set $d = 200$ and $r = 1$. *Left.* MSE achieved by SGD at convergence, as a function of the sparsity level p. The empirical values (dots) match our theoretical prediction (blue line). For $p < p_{\rm crit}$, the loss is given by Proposition 5.1 for $\boldsymbol{B}$ sampled from the Haar distribution; for $p \geq p_{\rm crit}$, the loss is given by Proposition 5.2 for $\boldsymbol{B}$ equal to the identity. *Center.* Encoder matrix $\boldsymbol{B}$ at convergence of SGD when $p = 0.3 < p_{\rm crit}$: the matrix is a random rotation. *Right.* Encoder matrix $\boldsymbol{B}$ at convergence of SGD when $p = 0.7 \geq p_{\rm crit}$. The negative sign in part of the entries of $\boldsymbol{B}$ is cancelled by the corresponding sign in the entries of $\boldsymbol{A}$. Hence, $\boldsymbol{B}$ is equivalent to a permutation of the identity.

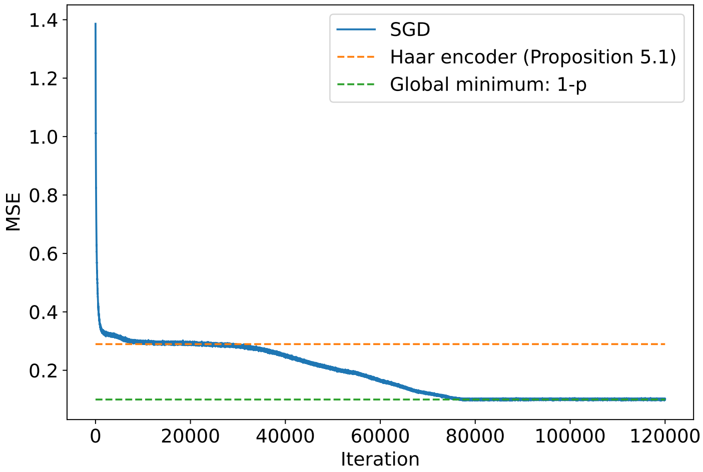

Figure 22. Compression of sparse Rademacher data via the autoencoder in (4). We set $d = 200$, $r = 1$, and $p = 0.9$. The MSE is plotted as a function of the number of iterations and, as $p > p_{\rm crit}$, it displays a staircase behavior.

Sparse Laplace. The numerical evaluation of the autoencoder in (4) for data which comes from a sparse Laplace distribution (see (109)) is illustrated in Figure 23 for $d = 512$ and $r = 1$. As predicted by our theory, in this case, regardless of the sparsity level p, SGD converges to the minimizer which corresponds to an orthogonal matrix $\boldsymbol{B}$. In fact, the MSE value for the Haar design in Proposition 5.1 (orange curve) is always superior to the corresponding value achieved by a permutation of identity in Proposition 5.2 (solid blue line). As discussed in Section B.4, in order to obtain values for the solid orange curve, we use a numerical estimate for the conditional expectation (20).

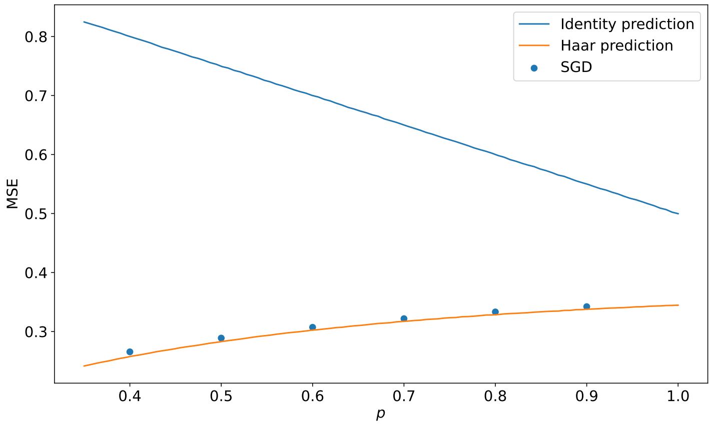

Figure 23. Compression of data whose distribution is given by a sparse Laplace prior via the autoencoder in (2). We set $d = 512$ and $r = 1$. We plot the MSE achieved by SGD at convergence, as a function of the sparsity level p. The empirical values (dots) match our theoretical prediction (solid orange line) and the loss is given by Proposition 5.1 for $\boldsymbol{B}$ sampled from the Haar distribution.

Gaussian mixture with aspect ratio. The numerical evaluation of the autoencoder in (4) for data which comes from a sparse Gaussian mixture (see (115)) is illustrated in Figure 24 for $d = 200$ and $r = 1$. As predicted by our theory, in this case, regardless of the aspect ratio γ, SGD converges to the minimizer which corresponds to a matrix $\boldsymbol{B}$ given by a permutation of the identity. In fact, the MSE value for the Haar design in Proposition 5.1 (dashed orange curve) is always inferior to the corresponding value achieved by a permutation of identity in Proposition 5.2 (solid blue curve). As discussed in Section B.4, in order to obtain values for the dashed orange curve, we use a numerical estimate for the conditional expectation (20).

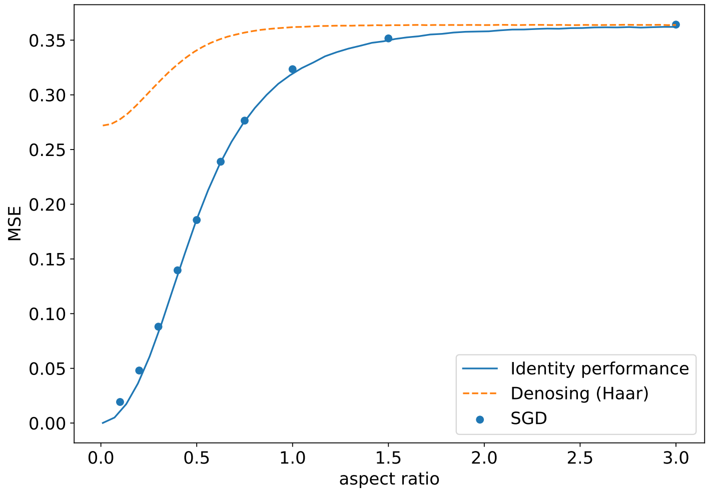

Figure 24. Compression of data whose distribution is given by a (non-sparse) Gaussian mixture with aspect ratio via the autoencoder in (2). We set $d = 200$ and $r = 1$. We plot the MSE achieved by SGD at convergence, as a function of the sparsity level p. The empirical values (dots) match our theoretical prediction (blue line) and the loss is given by Proposition 5.2 for $\boldsymbol{B}$ equal to a permutation of identity.

C.8. Discussion on multi-layer decoder

First, let us elaborate on some design points for the network in (23). The merging operations $\oplus_2$ and $\oplus_3$ play the role of the correction terms $-\sum_{i=1}^{t-1} \beta_{t,i} \hat{\boldsymbol{x}}^i$ and $-\sum_{i=1}^{t} \alpha_{t,i} \hat{\boldsymbol{z}}^i$ in the RI-GAMP iterates in (22). Furthermore, the composition of $\oplus_3$ and $f_2(\cdot)$ in $\hat{\boldsymbol{x}}_2$ approximates taking the posterior mean in (22). We note that the network (23) can be generalized to emulate more RI-GAMP iterations, at the cost of additional layers and skip connections (induced by the merging operations $\oplus_k$).

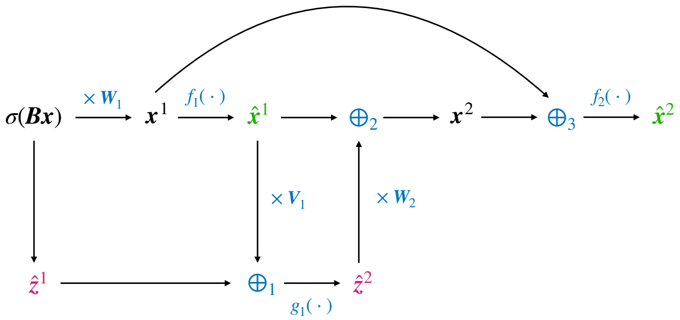

Figure 25. Block diagram of the decoder in (23).

In the rest of this appendix, we discuss how to obtain the orange curve in the right plot of Figure 8, which corresponds to the Bayes-optimal MSE when $\boldsymbol{B}$ is sampled from the Haar distribution. This optimal MSE is achieved by the fixed point of the VAMP algorithm proposed in (Rangan et al., 2019). Thus, we implement the state evolution recursion from (Rangan et al., 2019), in order to evaluate the fixed point. As the specific setting considered here ($\boldsymbol{x} \sim \mathrm{SG}_d(p)$, $\boldsymbol{B}$ a Haar matrix, and a generalized linear model with sign activation) is not considered in (Rangan et al., 2019), we provide explicit expressions for the recursion leading to the desired MSE.

First state evolution function - $\mathcal{E}_1(\gamma_1)$. We start with the state evolution function that is equal to the following expected value of the derivative of the conditional expectation

$$\mathcal{E}_1(\gamma_1) = \mathbb{E}_{R_1}\left[\frac{\partial}{\partial R_1}\mathbb{E}[X|R_1 = X + P]\right], \quad X \sim \mathrm{SG}_1(p), \quad P \sim \mathcal{N}(0, \gamma_1^{-1}). \tag{117}$$

For completeness, we note that the quantity

$$\frac{\partial}{\partial R_1}\mathbb{E}[X|R_1 = X + P]$$

is in fact the conditional variance $\mathrm{Var}[X|R_1 = X + P]$ up to a scaling (Dytso et al., 2020), which is related to the optimal MSE.

Modulo the scalings, the computation of $\mathbb{E}[X|R_1 = X + P]$ is similar to the computation performed in Section B.4. For brevity, we just state the final result:

$$\mathbb{E}[X|R_1 = X + P] = \frac{p \cdot \frac{R_1}{\sqrt{2\pi p^{-1}}} \cdot \exp\left(-\frac{pR_1^2}{2(p\gamma_1^{-1}+1)}\right) \cdot \frac{1}{(p\gamma_1^{-1}+1)^{3/2}}}{p \cdot \frac{1}{\sqrt{2\pi(p^{-1}+\gamma_1^{-1})}} \cdot \exp\left(-\frac{pR_1^2}{2(p\gamma_1^{-1}+1)}\right) + (1-p) \cdot \frac{1}{\sqrt{2\pi\gamma_1^{-1}}} \cdot \exp\left(-\frac{R_1^2}{2\gamma_1^{-1}}\right)} := \frac{E(R_1)}{p(R_1)}. \tag{118}$$

Taking the partial derivative in R_1 and substituting in (117) yields:

$$\begin{aligned}\mathcal{E}_1(\gamma_1) &= \gamma_1^{-1}\int_{\mathbb{R}} \frac{\partial}{\partial R_1}\mathbb{E}[X|R_1 = X + P] \cdot p(R_1)\mathrm{d}R_1 = \gamma_1^{-1}\int_{\mathbb{R}} \frac{E'(R_1)p(R_1) - E(R_1)p'(R_1)}{p^2(R_1)}p(R_1)\mathrm{d}R_1 \\ &= \gamma_1^{-1}\int_{\mathbb{R}}\left(E'(R_1) - E(R_1)\cdot\frac{\partial}{\partial R_1}\log p(R_1)\right)\mathrm{d}R_1.\end{aligned} \tag{119}$$

We can readily verify that

$$\int_{\mathbb{R}} E'(R_1)\mathrm{d}R_1 = \lim_{\mathrm{ext}\to\infty} E(R_1)\Big|_{-\mathrm{ext}}^{+\mathrm{ext}} = 0.$$

An integration by parts for the remaining term in (119) gives:

$$\mathcal{E}_1(\gamma_1) = \gamma_1^{-1}\lim_{\mathrm{ext}\to\infty} E(R_1)\log p(R_1)\Big|_{-\mathrm{ext}}^{+\mathrm{ext}} - \gamma_1^{-1}\int_{\mathbb{R}} E'(R_1)\log p(R_1)\mathrm{d}R_1 = -\gamma_1^{-1}\int_{\mathbb{R}} E'(R_1)\log p(R_1)\mathrm{d}R_1. \tag{120}$$

The RHS of (120) is then evaluated via numerical integration. For completeness, the derivative $E'(R_1)$ has the following form:

$$\begin{aligned} E'(R_1) = p \cdot \frac{1}{\sqrt{2\pi p^{-1}}} \cdot \exp\left(-\frac{pR_1^2}{2(p\gamma_1^{-1}+1)}\right) \cdot \frac{1}{(p\gamma_1^{-1}+1)^{3/2}} \\ - p^2 \cdot \frac{R_1^2}{\sqrt{2\pi p^{-1}}} \cdot \exp\left(-\frac{pR_1^2}{2(p\gamma_1^{-1}+1)}\right) \cdot \frac{1}{(p\gamma_1^{-1}+1)^{5/2}}. \end{aligned}$$

Second state evolution function - $\mathcal{E}_2(\tau_2, \gamma_2)$. This function is defined in terms of spectrum of $\boldsymbol{B}^\top \boldsymbol{B} \in \mathbb{R}^{d\times d}$. Namely, for $r \le 1$, the distribution of the eigenvalues of $\boldsymbol{B}^\top \boldsymbol{B}$ obeys the following law

$$\rho_S = r \cdot \delta_1 + (1-r) \cdot \delta_0.$$

The state evolution function $\mathcal{E}_2(\tau_2, \gamma_2)$ is then defined as follows

$$\mathcal{E}_2(\tau_2, \gamma_2) := \mathbb{E}_{S\sim\rho_S}\left[\frac{1}{\tau_2 \cdot S^2 + \gamma_2}\right] = r \cdot \frac{1}{\tau_2+\gamma_2} + (1-r)\cdot\frac{1}{\gamma_2}.$$

Third state evolution function - $\mathcal{B}_2(\tau_2, \gamma_2)$. The computation is similar to the case of the second state evolution function. Namely, the third state evolution function is defined as follows:

$$\mathcal{B}_2(\tau_2, \gamma_2) = \frac{1}{r} \cdot \mathbb{E}_{S\sim\rho_S}\left[\frac{\tau_2 S^2}{\tau_2 S^2 + \gamma_2}\right] = \frac{1}{r}\cdot r \cdot \frac{\tau_2}{\tau_2+\gamma_2} = \frac{\tau_2}{\tau_2+\gamma_2}.$$

Fourth state evolution function - $\mathcal{B}_1(\tau_1)$. The last state evolution function is defined similarly to $\mathcal{E}_1(\gamma_1)$, namely

$$\mathcal{B}_1(\tau_1) = \mathbb{E}_{P_1,Y}\left[\frac{\partial}{\partial P_1}\mathbb{E}[Z|P_1,Y]\right]. \tag{121}$$

Here, $Z \sim \mathcal{N}(0,1)$ has variance one (since the spectrum of $\boldsymbol{B}$ has unit variance), $Y = \operatorname{sign}(Z)$ and $P_1 = b\cdot Z + a \cdot G$, where $G \sim \mathcal{N}(0,1)$ is independent of Z and

$$b = 1 - \tau_1^{-1}, \quad a = \sqrt{b\cdot(1-b)}.$$

The outer expectation in (121) is estimated via Monte-Carlo. We now compute the conditional expectation. First note that the following decomposition (depending on the sign of Y) holds:

$$\mathbb{E}[Z|P_1,Y] = \mathbb{E}[Z'|P_1'], \tag{122}$$

where $Z' = \mathbb{1}_{ZY\ge 0}\cdot Z$ and $P_1' = b \cdot Z' + a\cdot G$. Using Bayes formula, we get that

$$\mathbb{E}[Z'|P_1'] = \frac{\int_{ZY\ge0} Z \exp\left(-\frac{Z^2}{2}\right)\exp\left(-\frac{(P_1-bZ)^2}{2a^2}\right)\mathrm{d}Z}{\int_{ZY\ge0} \exp\left(-\frac{Z^2}{2}\right)\exp\left(-\frac{(P_1-bZ)^2}{2a^2}\right)\mathrm{d}Z}. \tag{123}$$

Completing the square in the exponents gives

$$\frac{Z^2a^2 + (P_1 - bZ)^2}{2a^2} = \frac{bZ^2 - 2bZP_1 + P_1^2}{2b(1-b)} = \frac{(Z-P_1)^2}{2(1-b)} + \frac{P_1^2}{2b},$$

which after substitution in (123) results in

$$\mathbb{E}[Z'|P_1'] = \frac{\int_{ZY\ge0} Z \exp\left(-\frac{(Z-P_1)^2}{2\tau_1^{-1}}\right)\mathrm{d}Z}{\int_{ZY\ge0} \exp\left(-\frac{(Z-P_1)^2}{2\tau_1^{-1}}\right)\mathrm{d}Z}. \tag{124}$$

Note that the denominator of (124) is easy to access via the standard Gaussian CDF $\Psi(\cdot)$ as follows

$$\begin{aligned}\frac{1}{\sqrt{2\pi\tau_1^{-1}}}\int_{ZY\geq 0}\exp\left(-\frac{(Z-P_1)^2}{2\tau_1^{-1}}\right)\mathrm{d}Z &= \mathbb{1}_{Y\geq 0}\cdot\left[1-\Psi\left(-\frac{P_1}{\tau_1^{-1/2}}\right)\right]+\mathbb{1}_{Y<0}\cdot\Psi\left(-\frac{P_1}{\tau_1^{-1/2}}\right)\\ &=\Psi\left(\frac{YP_1}{\tau_1^{-1/2}}\right),\end{aligned}\tag{125}$$

where for the last equality we use that $\Psi(x)=1-\Psi(-x)$ and $Y\in\{-1,+1\}$. For the numerator of (124), we get

$$\frac{1}{\sqrt{2\pi\tau_1^{-1}}}\int\mathbb{1}_{YZ\geq 0}\cdot Z\exp\left(-\frac{(Z-P_1)^2}{2\tau_1^{-1}}\right)\mathrm{d}Z.\tag{126}$$

Let us denote the PDF of $\mathcal{N}(\mu,\sigma^2)$ by ρ_{μ,σ^2}, and use the shorthand $\rho(\cdot)$ for $\rho_{0,1}(\cdot)$. Note that $\rho_{x,\sigma^2}(0)=\sigma^{-1}\rho(x/\sigma)$. Then, by Stein's identity, we have

$$\mathbb{E}\left[\mathbb{1}_{YZ\geq 0}\cdot(Z-P_1)\right]=\tau_1^{-1}\cdot\mathbb{E}[Y\cdot\delta_0(Z)]=Y\tau_1^{-1}\cdot\rho_{P_1,\tau_1^{-1}}(0)=Y\tau_1^{-1/2}\cdot\rho\left(\frac{P_1}{\tau_1^{-1/2}}\right),$$

as the weak derivative of $\mathbb{1}_{YZ\geq 0}$ is well-defined and equal to $Y\cdot\delta_0(Z)$. Noting that similarly to (125)

$$\mathbb{E}\left[\mathbb{1}_{YZ\geq 0}\cdot P_1\right]=P_1\cdot\Psi\left(\frac{YP_1}{\tau_1^{-1/2}}\right),$$

we conclude that

$$(126)=P_1\cdot\Psi\left(\frac{YP_1}{\tau_1^{-1/2}}\right)+Y\tau_1^{-1/2}\cdot\rho\left(\frac{P_1}{\tau_1^{-1/2}}\right).\tag{127}$$

Combining the results gives

$$\mathbb{E}[Z_1'|P_1']=\frac{P_1\cdot\Psi\left(\frac{YP_1}{\tau_1^{-1/2}}\right)+Y\tau_1^{-1/2}\cdot\rho\left(\frac{P_1}{\tau_1^{-1/2}}\right)}{\Psi\left(\frac{YP_1}{\tau_1^{-1/2}}\right)}=P_1+Y\tau_1^{-1/2}\cdot\frac{\rho\left(\frac{P_1}{\tau_1^{-1/2}}\right)}{\Psi\left(\frac{YP_1}{\tau_1^{-1/2}}\right)}.\tag{128}$$

It now remains to take the derivative in P_1. We get that

$$\mathcal{B}_1(\tau_1)=1-\frac{YP_1\sqrt{\tau_1}\cdot\rho\left(\frac{P_1}{\tau_1^{-1/2}}\right)\cdot\Psi\left(\frac{YP_1}{\tau_1^{-1/2}}\right)+\rho\left(\frac{P_1}{\tau_1^{-1/2}}\right)^2}{\Psi\left(\frac{YP_1}{\tau_1^{-1/2}}\right)^2},\tag{129}$$

where we used that $Y^2=1$ and that $\frac{\partial}{\partial x}\Psi(x)=\rho(x)$.

State evolution recursion. At this point, we are ready to present the state evolution recursion, which reads

$$\begin{aligned}\gamma_{2,k}&=\gamma_{1,k}\cdot\frac{1-\mathcal{E}_1(\gamma_{1,k})}{\mathcal{E}_1(\gamma_{1,k})},\\ \tau_{2,k}&=\tau_{1,k}\cdot\frac{1-\mathcal{B}_1(\tau_{1,k})}{\mathcal{B}_1(\tau_{1,k})},\\ \gamma_{1,k+1}&=\gamma_{2,k}\cdot\frac{1-\mathcal{E}_2(\tau_{2,k},\gamma_{2,k})}{\mathcal{E}_2(\tau_{2,k},\gamma_{2,k})}=\gamma_{2,k}\cdot\frac{r\cdot\tau_{2,k}}{(1-r)\cdot\tau_{2,k}+\gamma_{2,k}},\\ \tau_{1,k+1}&=\tau_{2,k}\cdot\frac{1-\mathcal{B}_2(\tau_{2,k},\gamma_{2,k})}{\mathcal{B}_2(\tau_{2,k},\gamma_{2,k})}=\gamma_{2,k}.\end{aligned}\tag{130}$$

The initialization $\gamma_{1,0}$ and $\tau_{1,0}$ can be set to a small strictly positive number. For the experiments, we choose the value of 10^{-6}.

MSE from the state evolution parameter $\gamma_{1,k+1}$. The MSE after k steps of the recursion can be accessed via the function previously computed in (118). Namely, let $\boldsymbol{x} \sim \mathrm{SG}_d(p)$ and $\boldsymbol{r}_1 = \boldsymbol{x} + \boldsymbol{p}$, where $\boldsymbol{p}$ has i.i.d. entries with distribution $\mathcal{N}(0, \gamma_{1,k+1}^{-1})$. Define

$$g(\boldsymbol{r}_1) = \mathbb{E}[\boldsymbol{x} | \boldsymbol{r}_1 = \boldsymbol{x} + \boldsymbol{p}].$$

By the tower property of the conditional expectation, we claim that the following holds

$$\mathbb{E}[\mathbb{E}[X|Y] \cdot X] = \mathbb{E}[\mathbb{E}[\mathbb{E}[X|Y] \cdot X|Y]] = \mathbb{E}\left[(\mathbb{E}[X|Y])^2\right],$$

where we use that $\mathbb{E}[X|Y]$ is measurable w.r.t. Y. Thus, we have that

$$\mathbb{E}\langle g(\boldsymbol{r}_1), \boldsymbol{x}\rangle = d \cdot \mathbb{E}\left[(g(\boldsymbol{r}_1)_1)^2\right],$$

where $g(\boldsymbol{r}_1)_1$ denotes the first entry of the vector $g(\boldsymbol{r}_1)$. Finally, the desired MSE after k steps of the recursion is equal to

$$d^{-1} \cdot \mathbb{E}\|\boldsymbol{x} - g(\boldsymbol{r}_1)\|_2^2 = 1 - \mathbb{E}\left[(g(\boldsymbol{r}_1)_1)^2\right]. \tag{131}$$

We evaluate (131) for k large enough, so that the MSE has converged. For the experiment in Figure 8, we use $k = 15$, as for $k \geq 15$ the MSE value in (131) is stable.