Fundamental Limits of Two-layer Autoencoders, and Achieving Them with Gradient Methods

.

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

Abstract

Autoencoders are a popular model in many branches of machine learning and lossy data compression. However, their fundamental limits, the performance of gradient methods and the features learnt during optimization remain poorly understood, even in the two-layer setting. In fact, earlier work has considered either linear autoencoders or specific training regimes (leading to vanishing or diverging compression rates). Our paper addresses this gap by focusing on non-linear two-layer autoencoders trained in the challenging proportional regime in which the input dimension scales linearly with the size of the representation. Our results characterize the minimizers of the population risk, and show that such minimizers are achieved by gradient methods; their structure is also unveiled, thus leading to a concise description of the features obtained via training. For the special case of a sign activation function, our analysis establishes the fundamental limits for the lossy compression of Gaussian sources via (shallow) autoencoders. Finally, while the results are proved for Gaussian data, numerical simulations on standard datasets display the universality of the theoretical predictions.

1. Introduction

Autoencoders represent a key building block in many branches of machine learning (Kingma & Welling, 2014; Rezende et al., 2014), including generative modeling (Bengio et al., 2013) and representation learning (Tschannen et al., 2018). Prompted by the fact that autoencoders learn succinct representations, neural autoencoding techniques have achieved remarkable success in lossy data compression, even outperforming classical methods, such as jpeg (Ballé et al., 2017; Theis et al., 2017; Agustsson et al., 2017). However, despite the large body of empirical work on neural autoencoders and compressors, basic theoretical questions remain poorly understood even in the shallow case:

What are the fundamental performance limits of autoencoders? Can we achieve such limits with gradient methods? What features does the optimization procedure learn?

Prior work has focused either on linear autoencoders (Baldi & Hornik, 1989; Kunin et al., 2019; Gidel et al., 2019), on the severely under-parameterized setting in which the input dimension is much larger than the number of neurons (Refinetti & Goldt, 2022), or on specific training regimes (lazy training (Nguyen et al., 2021) and mean-field regime with a polynomial number of neurons (Nguyen, 2021)), see Section 2. In contrast, in this paper we consider *non-linear* autoencoders trained in the *challenging proportional regime*, in which the number of inputs to compress scales linearly with the size of the representation. More specifically, we consider the prototypical model of a two-layer autoencoder

$$\hat{\boldsymbol{x}}(\boldsymbol{x}) := \hat{\boldsymbol{x}}(\boldsymbol{x}, \boldsymbol{A}, \boldsymbol{B}) = \boldsymbol{A}\sigma(\boldsymbol{B}\boldsymbol{x}). \tag{1}$$

Here, $\boldsymbol{x} \in \mathbb{R}^d$ is the input to compress, $\hat{\boldsymbol{x}} \in \mathbb{R}^n$ the reconstruction, $\boldsymbol{B} \in \mathbb{R}^{n\times d}$ the encoding matrix, and $\boldsymbol{A} \in \mathbb{R}^{d\times n}$ the decoding matrix; the activation $\sigma : \mathbb{R} \to \mathbb{R}$ is applied *element-wise*. We aim at minimizing the population risk

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

where the expectation is taken over the distribution of the input $\boldsymbol{x}$. Our focus is on Gaussian input data, i.e., $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$. When σ is the sign function, the encoder $\sigma(\boldsymbol{B}\boldsymbol{x})$ can be interpreted as a *compressor*, namely, it compresses the d-dimensional input signal into n bits. The problem (2) of compressing a Gaussian source with quadratic distortion has been studied in exquisite detail in the information theory literature (Cover & Thomas, 2006), and the optimal performance for general encoder/decoder pairs is known via the *rate-distortion* formalism which characterizes the lowest achievable distortion in terms of the rate $r = n/d$.

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

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

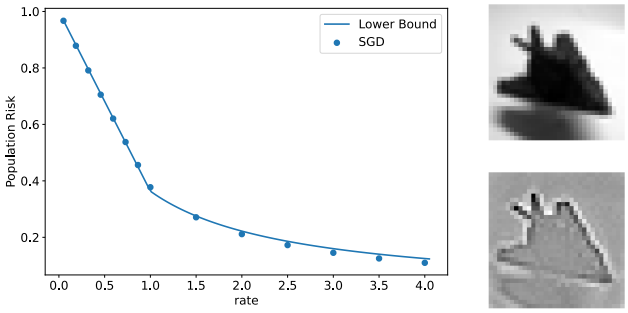

Figure 1. Compression ($\sigma \equiv \text{sign}$) of the CIFAR-10 "airplane" class with a two-layer autoencoder. The data is *whitened* so that $\boldsymbol{\Sigma} = \boldsymbol{I}$: on top, an example of a grayscale image; on the bottom, the corresponding whitening. The blue dots are the population risk obtained via SGD, and they agree well with the solid line corresponding to the lower bounds of Theorem 4.2 and Proposition 4.3.

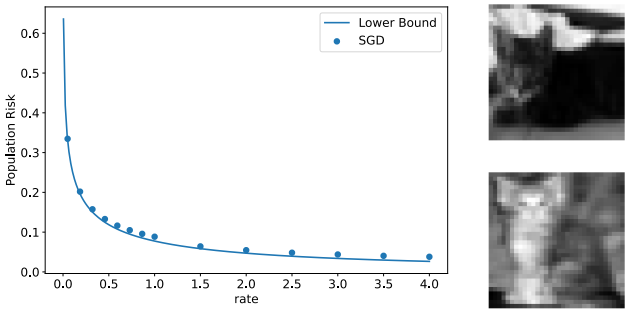

Figure 2. Compression ($\sigma \equiv \text{sign}$) of the CIFAR-10 "cat" class with a two-layer autoencoder. The data is *not whitened* ($\boldsymbol{\Sigma} \neq \boldsymbol{I}$). The blue dots are the SGD population risk, and they are close to the lower bound of Theorem 5.2.

Here, we focus on encoders and decoders that form the two-layer autoencoder (1): we study the fundamental limits of this learning problem, as well as the performance achieved by commonly used gradient descent methods.

Main contributions. Taken all together, our results show that, for two-layer autoencoders, gradient descent methods achieve a global minimizer of the population risk: this is rigorously proved in the isotropic case ($\boldsymbol{\Sigma} = \boldsymbol{I}$) and corroborated by numerical simulations for a general covariance $\boldsymbol{\Sigma}$. Furthermore, we unveil the structure of said minimizer: for $\boldsymbol{\Sigma} = \boldsymbol{I}$, the optimal decoder has unit singular values; for general covariance, the spectrum of the decoder exhibits the same block structure as $\boldsymbol{\Sigma}$, and it can be explicitly obtained from $\boldsymbol{\Sigma}$ via a water-filling criterion; in all cases, weight-tying is optimal, i.e., $\boldsymbol{A}$ is proportional to $\boldsymbol{B}^\top$. Specifically, our technical results can be summarized as follows.

- Section 4.1 characterizes the minimizers of the risk (2) for isotropic data: Theorem 4.2 provides a tight lower bound, which is achieved by the set (7) of weight-tied *orthogonal* matrices, when the compression rate $r = n/d \leq 1$; for $r > 1$, Propositions 4.3 and 4.4 give a lower bound, which is approached (as $d \to \infty$) by the set (12) of weight-tied *rotationally invariant* matrices.

- Section 4.2 shows that the above minimizers are reached by gradient descent methods for $r \leq 1$: Theorem 4.5 shows linear convergence of *gradient flow* for general initializations, under a weight-tying condition; Theorem 4.6 considers a Gaussian initialization and proves global convergence of the *projected gradient descent* algorithm, in which the encoder matrix $\boldsymbol{B}$ is optimized via a gradient method and the decoder matrix $\boldsymbol{A}$ is obtained directly via linear regression.

- Section 5 focuses on data with general covariance $\boldsymbol{\Sigma} \neq \boldsymbol{I}$. We observe that experimentally weight-tying is optimal and then derive the corresponding lower bound (see Theorem 5.2), which is also asymptotically achieved (as $d \to \infty$) by rotationally invariant matrices with a carefully designed spectrum (depending on $\boldsymbol{\Sigma}$), see Proposition 5.3.

When $\sigma \equiv \text{sign}$, our analysis characterizes the fundamental limits of the lossy compression of a Gaussian source via two-layer autoencoders. Remarkably, if we restrict to a certain class of *linear encoders* for compression, two-layer autoencoders achieve optimal performance (Tulino et al., 2013), which can be generally obtained via a message passing decoding algorithm (Rangan et al., 2019). However, for *general encoder/decoder pairs*, shallow autoencoders fail to meet the information-theoretic bound given by the rate-distortion curve, see Section 6.

Going beyond the Gaussian assumption on the data, we provide numerical validation to our theoretical predictions on standard datasets, both in the isotropic case (Figure 1) and for general covariance (Figure 2). Additional numerical results – together with the details of the experimental setting – are in Appendix I.

Proof techniques. The lower bound on the population risk of Theorem 4.2 comes from a sequence of relaxations of the objective function, which eventually allows to apply a trace inequality. For $r \geq 1$, Proposition 4.3 crucially exploits an inequality for the Hadamard product of PSD matrices (Khare, 2021), and the asymptotic achievability of Proposition 4.4 takes advantage of concentration-of-measure tools for orthogonal matrices. The key quantity in the analysis of gradient methods is the encoder Gram matrix at iteration t, i.e., $\boldsymbol{B}(t)\boldsymbol{B}(t)^\top$. In particular, for gradient flow (Theorem 4.5), due to the weight-tying condition, tracking $\log\det \boldsymbol{B}(t)\boldsymbol{B}(t)^\top$ leads to a quantitative convergence result. However, when the weights are not tied, this quantity does not appear to increase along the optimization trajectory. Thus, for projected gradient descent (Theorem 4.6), the idea is to decompose $\boldsymbol{B}(t)\boldsymbol{B}(t)^\top$ into *(i)* its value at the optimum (given by the identity), *(ii)* the contribution due

to the spectrum evolution (keeping the eigenbasis fixed), and *(iii)* the change in the eigenbasis. Via a sequence of careful approximations, we are able to show that the term *(iii)* vanishes. Hence, we can study explicitly the evolution of the spectrum and obtain the desired convergence.

2. Related Work

Theory of autoencoders. A popular line of work has focused on two-layer *linear* autoencoders: Oftadeh et al. (2020) analyze the loss landscape; Kunin et al. (2019) show that the minimizers of the regularized loss recover the principal components of the data and, notably, the corresponding autoencoder is weight-tied; Bao et al. (2020) prove that stochastic gradient descent – after a slight perturbation – escapes the saddles and eventually converges; Gidel et al. (2019) characterize the time-steps at which the network learns different sets of features. Rangamani et al. (2018); Nguyen et al. (2019) prove local convergence for weight-tied two-layer ReLU autoencoders. Nguyen et al. (2021) focus on the lazy training regime (Chizat et al., 2019; Jacot et al., 2018) and bound the over-parameterization needed for global convergence. Radhakrishnan et al. (2020) show that over-parameterized autoencoders learn solutions that are contractive around the training examples. The latent spaces of autoencoders are studied in (Jain et al., 2021), where it is shown that such latent spaces can be aligned by stretching along the left singular vectors of the data. More closely related to our work, Nguyen (2021) and (Refinetti & Goldt, 2022) track the gradient dynamics of *non-linear* two-layer autoencoders via the mean-field PDE and a system of ODEs, respectively. However, these analyses are restricted to diverging and vanishing rates: Nguyen (2021) considers weight-tied autoencoders with polynomially many neurons in the input dimension (so that $r \to \infty$); Refinetti & Goldt (2022) consider the other extreme regime in which the input dimension diverges (so that $r \to 0$).

Neural compression. In recent years, compressors based on neural networks have outperformed traditional schemes on real-world data in terms of minimizing distortion and producing visually pleasing reconstructions at reasonable complexity (Ballé et al., 2017; Theis et al., 2017; Agustsson et al., 2017; Ballé et al., 2021). These methods typically use an autoencoder architecture with quantization of the latent variables, which is trained over samples drawn from the source. More recently, other architectures such as attention or diffusion-based models have been incorporated into neural compressors (Cheng et al., 2020; Liu et al., 2019; Yang & Mandt, 2022; Theis et al., 2022), and improvements have been observed. We refer to Yang et al. (2022) for a detailed review on this topic. Given the remarkable success of neural compressors, it is imperative to understand the fundamental limits of compression using neural architectures. In this regard, Wagner & Ballé (2021) consider a highly-structured and low-dimensional random process, dubbed the *sawbridge*, and show numerically that the rate-distortion function is achieved by a compressor based on deep neural networks trained via stochastic gradient descent. In contrast, our work considers Gaussian sources, which are high-dimensional in nature, and provides the fundamental limits of compression for two-layer autoencoders. Our results also imply that two-layer autoencoders cannot achieve the rate-distortion limit on Gaussian data, see Section 6.

Additional related works on rate-distortion formalism and non-linear inverse problems are discussed in Appendix A.

3. Preliminaries

Notations. We use plain symbols for real numbers (e.g., a, b), bold symbols for vectors (e.g., $\boldsymbol{a}, \boldsymbol{b}$), and capitalized bold symbols for matrices (e.g., $\boldsymbol{A}, \boldsymbol{B}$). We let $[n] = \{1, \dots, n\}$, $\boldsymbol{I}$ be the identity matrix and $\mathbf{1}$ the column vector containing ones. Given a matrix $\boldsymbol{A}$, we denote its operator norm by $\|\boldsymbol{A}\|_{op}$ and its Frobenius norm by $\|\boldsymbol{A}\|_F$. Given two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of the same shape, we denote their element-wise (Hadamard/Schur) product by $\boldsymbol{A} \circ \boldsymbol{B}$ and the k-th element-wise power by $\boldsymbol{A}^{\circ k}$. We write $L^2(\mathbb{R}, \mu)$ for the space of L^2 integrable functions on $\mathbb{R}$ w.r.t. the standard Gaussian measure μ and $h_k(x)$ for the k-th normalized Hermite polynomial (see e.g. O'Donnell (2014)).

Setup. We consider the two-layer autoencoder (1) and aim at minimizing the population risk (2) for a given rate $r = n/d$. In particular, we provide tight lower bounds on the minimum of the population risk computed on Gaussian input data with covariance $\boldsymbol{\Sigma}$, i.e.,

$$\widehat{\mathcal{R}}(r) := \min_{\boldsymbol{A}, \boldsymbol{B}} \mathcal{R}(\boldsymbol{A}, \boldsymbol{B}). \tag{3}$$

In the isotropic case ($\boldsymbol{\Sigma} = \boldsymbol{I}$), our results hold for any odd activation $\sigma \in L^2(\mathbb{R}, \mu)$ after restricting the rows of the encoding matrix $\boldsymbol{B}$ to have unit norm. We remark that, when $\sigma(x) = \mathrm{sign}(x)$, the restriction is unnecessary since the activation is homogeneous.[1] We also note that restricting the norms of the rows of $\boldsymbol{B}$ prevents the model from entering the "linear" regime. In fact, when $\|\boldsymbol{B}\|_F \approx 0$, by linearizing the activation around zero, (1) reduces to the linear model $\hat{\boldsymbol{x}}(\boldsymbol{x}) \approx \boldsymbol{A}\boldsymbol{B}\boldsymbol{x}$, which exhibits a PCA-like behaviour. For general covariance $\boldsymbol{\Sigma}$, we consider odd homogeneous activations, which includes the sign function and monomials of arbitrary odd degree.

Any function $\sigma \in L^2(\mathbb{R}, \mu)$ can be expanded in terms of Hermite polynomials. This allows to perform Fourier analysis in the Gaussian space $L^2(\mathbb{R}, \mu)$, and it provides a natural tool because of the Gaussian assumption on the data. In

[1] We say that a function σ is homogeneous if there exists an integer k s.t. $\sigma(\alpha x) = \alpha^k \sigma(x)$ for all $\alpha \neq 0$.

particular, for odd σ, only odd Hermite polynomials occur, i.e.,

$$\sigma(x) = \sum_{\ell=0}^{\infty} c_{2\ell+1} h_{2\ell+1}(x), \tag{4}$$

where $\{c_\ell\}_{\ell\in\mathbb{N}}$ denote the Hermite coefficients of σ. We also consider the following auxiliary quantity

$$\widetilde{\mathcal{R}}(r) := \min_{\boldsymbol{A},\|(\boldsymbol{BD})_{i,:}\|_2=1} \mathcal{R}(\boldsymbol{A},\boldsymbol{B}), \tag{5}$$

that defines a minimum of the population risk for the autoencoder (1) with a certain norm constraint on the encoder weights $\boldsymbol{B}$. Here, $\boldsymbol{D}$ contains the square roots of the eigenvalues of $\boldsymbol{\Sigma}$ (i.e., $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{D}^2\boldsymbol{U}^\top$ for an orthogonal matrix $\boldsymbol{U}$), and $(\boldsymbol{BD})_{i,:}$ stands for the i-th row of the matrix $\boldsymbol{BD}$. A few remarks about the restricted population risk (5) are in order. First of all, if σ is homogeneous, the minimum of the restricted population risk (5) and of the unconstrained one (3) coincide (see Lemma 4.1 and Lemma 5.1). Thus, in this case, the analysis of $\widetilde{R}(r)$ directly provides results on the quantity of interest, i.e., $\widehat{\mathcal{R}}(r)$. The technical advantage of analysing (5) over (3) comes from fact that the expectation with respect to the Gaussian inputs, which arises in the constrained objective, can be explicitly computed via the reproducing property of Hermite polynomials (see, e.g., O'Donnell (2014)). To exploit this reproducing property, it is crucial that the inner products $\langle \boldsymbol{B}_{i,:}, \boldsymbol{x}\rangle$ have the same scale, which is ensured by picking $\|(\boldsymbol{BD})_{i,:}\|_2 = 1$. The sole dependence of the constraint on the spectrum $\boldsymbol{D}$ (and, not on a particular choice of $\boldsymbol{U}$) stems from the rotational invariance of the isotropic Gaussian distribution.

4. Main Results

In this section, we consider isotropic Gaussian data, i.e., $\boldsymbol{\Sigma} = \boldsymbol{D} = \boldsymbol{I}$. First, we derive a closed form expression for the population risk in Lemma 4.1. Then, in Theorem 4.2 we give a lower bound on the population risk for $r \leq 1$ and provide a complete characterization of the autoencoder parameters $(\boldsymbol{A},\boldsymbol{B})$ achieving it. Surprisingly, the minimizer exhibits a *weight-tying* structure and the corresponding matrices are *rotationally invariant*. Later, in Proposition 4.3 we derive an analogous lower bound for $r > 1$. While it is hard to characterize the minimizer structure explicitly for a finite input dimension d (and $r > 1$), we provide a sequence $\{(\boldsymbol{A}_d,\boldsymbol{B}_d)\}_{d\in\mathbb{N}}$ that meets the lower bound in the high-dimensional limit ($d\to\infty$), see Proposition 4.4. Notably, the elements of this sequence share the key features (weight-tying, rotational invariance) of the minimizers for $r \leq 1$. In Section 4.2 we describe gradient methods that provably achieve the optimal value of the population risk. Specifically, we consider gradient flow under a weight-tying constraint and projected (on the sphere) gradient descent with Gaussian initialization. The corresponding results are stated in Theorem 4.5 and Theorem 4.6.

We start by expanding σ in a Hermite series to obtain a closed-form expression for the population risk.

Lemma 4.1. *Consider any odd $\sigma \in L^2(\mathbb{R},\mu)$ and its Hermite expansion given by (4). Then, $\widetilde{\mathcal{R}}(r)$ is equivalent to*

$$\min_{\boldsymbol{A},\|\boldsymbol{B}_{i,:}\|_2=1} \frac{1}{d}\left(\mathrm{Tr}\left[\boldsymbol{A}^\top\boldsymbol{A}f(\boldsymbol{B}\boldsymbol{B}^\top)\right] - 2c_1\cdot\mathrm{Tr}\left[\boldsymbol{B}\boldsymbol{A}\right]\right)+1, \tag{6}$$

where $f(x) := \sum_{\ell=0}^{\infty}(c_{2\ell+1})^2 x^{2\ell+1}$ is applied elementwise. In particular, if $\sigma(x) = \mathrm{sign}(x)$, then $f(x) = c_1^2\cdot\arcsin(x)$ and $c_1 = \sqrt{2/\pi}$. Moreover, for any homogeneous σ, we have that $\widehat{\mathcal{R}}(r) = \widetilde{\mathcal{R}}(r)$.

The proof of the lemma above is contained in Appendix B. Note that, if $c_1 = 0$, it is easy to see that the minimum of $\widetilde{R}(r)$ equals 1 and it is attained when $\boldsymbol{A}^\top\boldsymbol{A}$ is the zero-matrix. Furthermore, if $\sum_{\ell=1}^{\infty}(c_{2\ell+1})^2 = 0$, then $\sigma(x) = c_1^2 x$ and we fall back into the simpler case of a linear autoencoder (Baldi & Hornik, 1989; Kunin et al., 2019; Gidel et al., 2019). Thus, for the rest of the section, we will assume that $c_1 \neq 0$ and $\sum_{\ell=1}^{\infty}(c_{2\ell+1})^2 \neq 0$.

4.1. Fundamental Limits: Lower Bound on Risk

We begin by providing a tight lower bound for $r \leq 1$, which is *uniquely* achieved on the set of *weight-tied* orthogonal matrices $\mathcal{H}_{n,d}$ defined as

$$\mathcal{H}_{n,d} := \left\{\widetilde{\boldsymbol{A}},\widetilde{\boldsymbol{B}}^\top\in\mathbb{R}^{d\times n} : \widetilde{\boldsymbol{A}} = \frac{c_1}{f(1)}\cdot\widetilde{\boldsymbol{B}}^\top, \widetilde{\boldsymbol{B}}\widetilde{\boldsymbol{B}}^\top = \boldsymbol{I}\right\}. \tag{7}$$

Theorem 4.2. *Consider any odd $\sigma \in L^2(\mathbb{R},\mu)$ and fix $r \leq 1$. Then, the following holds*

$$\widetilde{\mathcal{R}}(r) \geq \mathrm{LB}_{r\leq 1}(\boldsymbol{I}) := 1 - \frac{c_1^2}{f(1)}\cdot r,$$

and equality is achieved iff $(\boldsymbol{A},\boldsymbol{B}) \in \mathcal{H}_{n,d}$.

We note that the minimizers $\mathcal{H}_{n,d}$ of $\widetilde{\mathcal{R}}(r)$ do not directly correspond to the minimizers of the unconstrained population risk $\widehat{\mathcal{R}}(r)$, since in general $\widetilde{\mathcal{R}}(r) \neq \widehat{\mathcal{R}}(r)$. However, if σ is homogeneous, the "inverse" mapping can be readily obtained. For instance, when $\sigma(x) = \mathrm{sign}(x)$, rescaling the norms of the rows of $\boldsymbol{B}$ does not affect the compression, i.e., $\mathrm{sign}(\boldsymbol{B}\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{S}\boldsymbol{B}\boldsymbol{x})$ for any diagonal $\boldsymbol{S}$ with positive entries. Hence, to obtain a minimizer, it suffices that the rows of $\boldsymbol{B}$ form any set of orthogonal (not necessarily normalized) vectors. In contrast, note that $\boldsymbol{A}$ is still defined with respect to the row-normalized version of $\boldsymbol{B}$. Similar arguments hold for homogeneous activations.

We also note that the weight-tying structure (7) observed in the minimizers of the population risk is related to the early representation learning literature (Vincent et al., 2008; Hinton & Salakhutdinov, 2006).

We now provide a proof sketch for Theorem 4.2 and defer the full argument to Appendix C.1.

Proof sketch of Theorem 4.2. Using the series expansion of $f(\cdot)$, we can write

$$\begin{aligned}&\mathrm{Tr}\left[\boldsymbol{A}^\top \boldsymbol{A} f(\boldsymbol{B}\boldsymbol{B}^\top)\right] - 2c_1 \cdot \mathrm{Tr}\left[\boldsymbol{B}\boldsymbol{A}\right] \\ &= \sum_{\ell=0}^{\infty} c_{2\ell+1}^2 \left(\mathrm{Tr}\left[\boldsymbol{A}^\top \boldsymbol{A}\left(\boldsymbol{B}\boldsymbol{B}^\top\right)^{\circ 2\ell+1}\right] - 2\frac{c_1}{f(1)}\mathrm{Tr}\left[\boldsymbol{B}\boldsymbol{A}\right]\right).\end{aligned}$$

Thus, the minimization problem in Lemma 4.1 can be reduced to analysing each Hadamard power individually:

$$\min_{\boldsymbol{A}, \|\boldsymbol{B}_{i,:}\|_2=1} \mathrm{Tr}\left[\boldsymbol{A}^\top \boldsymbol{A}(\boldsymbol{B}\boldsymbol{B}^\top)^{\circ \ell}\right] - \frac{2c_1}{f(1)} \cdot \mathrm{Tr}\left[\boldsymbol{B}\boldsymbol{A}\right]. \quad (8)$$

The crux of the argument is to provide a suitable sequence of relaxations of (8). The first relaxation gives

$$\mathrm{Tr}\left[(\boldsymbol{A}^\top \boldsymbol{A} \circ \boldsymbol{Q})(\boldsymbol{B}\boldsymbol{B}^\top \circ \boldsymbol{Q})\right] - \frac{2c_1}{f(1)} \cdot \mathrm{Tr}\left[\boldsymbol{B}\boldsymbol{A}\right], \quad (9)$$

where $\boldsymbol{Q}$ is *any* PSD matrix with unit diagonal. Using the properties of the SVD of $\boldsymbol{Q}$, (9) can be further relaxed to

$$\sum_{i,j=1}^{n} \mathrm{Tr}\left[\boldsymbol{A}_j \boldsymbol{A}_j^\top \boldsymbol{B}_j \boldsymbol{B}_j^\top\right] - \frac{2c_1}{f(1)} \cdot \sum_{i=1}^{n} \mathrm{Tr}\left[\boldsymbol{B}_i \boldsymbol{A}_i\right], \quad (10)$$

where now $\boldsymbol{A}_i, \boldsymbol{B}_i^\top \in \mathbb{R}^{d\times n}$ are arbitrary matrices. The key observation is that

$$\sum_{i=1}^{n} \left\| \frac{c_1}{f(1)} \cdot \sqrt{\boldsymbol{X}}^{-1} \boldsymbol{A}_i^\top - \sqrt{\boldsymbol{X}} \boldsymbol{B}_i \right\|_F^2 = (10) + \frac{c_1^2}{(f(1))^2} \cdot n,$$

with $\boldsymbol{X} = \sum_{i=1}^n \boldsymbol{A}_i^\top \boldsymbol{A}_i$. As each relaxation lower bounds (8) and the Frobenius norm is positive, this argument leads to the lower bound on $\widetilde{R}(r)$. The fact that the lower bound is met for any $(\boldsymbol{A}, \boldsymbol{B}) \in \mathcal{H}_{n,d}$ can be verified via a direct calculation. The uniqueness follows by taking the intersection of the minimizers of (8) for different values of ℓ. □

Next, we move to the case $r > 1$.

Proposition 4.3. *Consider any odd $\sigma \in L^2(\mathbb{R}, \mu)$ and fix $r > 1$, then the following holds:*

$$\widetilde{\mathcal{R}}(r) \geq \mathrm{LB}_{r>1}(\boldsymbol{I}) := 1 - \frac{r}{r + \left(\frac{f(1)}{c_1^2} - 1\right)}.$$

The key difference with the proof of the lower bound in Theorem 4.2 is that the term $\mathrm{Tr}\left[\boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{B}\boldsymbol{B}^\top\right]$ requires a tighter estimate. This is due to the fact that the matrix $\boldsymbol{B}\boldsymbol{B}^\top$ is no longer full-rank when $r > 1$. We obtain the desired tighter bound by exploiting the following result by (Khare, 2021):

$$\boldsymbol{A}^\top \boldsymbol{A} \circ \boldsymbol{B}\boldsymbol{B}^\top \succeq \frac{1}{d} \cdot \mathrm{Diag}(\boldsymbol{B}\boldsymbol{A})\mathrm{Diag}(\boldsymbol{B}\boldsymbol{A})^\top, \quad (11)$$

where $\mathrm{Diag}(\boldsymbol{B}\boldsymbol{A})$ stands for the vector containing the diagonal entries of $\boldsymbol{B}\boldsymbol{A}$. The full argument is contained in Appendix C.2.1.

As for $r \leq 1$, the bound is met (here, in the limit $d \to \infty$) by considering weight-tied matrices:

$$\hat{\boldsymbol{B}}^\top = \sqrt{r} \cdot [\boldsymbol{I}_d, \boldsymbol{0}_{d,n-d}]\boldsymbol{U}^\top, \; \boldsymbol{b}_i = \frac{\hat{\boldsymbol{b}}_i}{\|\hat{\boldsymbol{b}}_i\|_2}, \; \boldsymbol{A} = \beta \boldsymbol{B}^\top, \quad (12)$$

where $\beta = \frac{c_1}{c_1^2 r + f(1) - c_1^2}$ and $\boldsymbol{U}$ is uniformly sampled from the group of rotation matrices. The idea behind the choice (12) is that, as $d \to \infty$, $(\boldsymbol{B}\boldsymbol{B}^\top)^{\circ 2\ell}$ for $\ell \geq 2$ is close to the identity matrix, and (11) is attained exactly. The formal statement is provided below and proved in Appendix C.2.2.

Proposition 4.4. *Consider any odd $\sigma \in L^2(\mathbb{R}, \mu)$ and fix $r > 1$. Let $\boldsymbol{A}, \boldsymbol{B}$ be defined as in (12). Then, for any $\epsilon > 0$ the following holds*

$$|\mathcal{R}(\boldsymbol{A}, \boldsymbol{B}) - \mathrm{LB}_{r>1}(\boldsymbol{I})| \leq C d^{-\frac{1}{2}+\epsilon},$$

with probability $1 - c/d^2$. Here, the constants c, C depend only on r and ϵ.

We note that all the arguments of this section directly apply to Gaussian data with a covariance matrix of the form $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 \boldsymbol{I}_{d-k} & \boldsymbol{0}_{(d-k)\times k} \\ \boldsymbol{0}_{k\times(d-k)} & \boldsymbol{0}_{k\times k} \end{bmatrix}$. For the details, see Appendix F.

4.2. Gradient Methods Achieve the Lower Bound

In this section, we discuss the achievability of the lower bound obtained in the previous section via gradient methods. We study two procedures which find the minimizer of the population risk $\mathcal{R}(\boldsymbol{A}, \boldsymbol{B})$ under the constraint $\|\boldsymbol{B}_{i,:}\|_2 = 1$. Namely, we analyse *(i) weight-tied gradient flow* on the sphere and *(ii)* its discrete version (with finite step size) *without* weight-tying, i.e., *projected gradient descent*.

The optimization objective in Lemma 4.1 is equivalent (up to a scaling independent of $(\boldsymbol{A}, \boldsymbol{B})$) to

$$\min_{\boldsymbol{A}, \|\boldsymbol{B}_{i,:}\|_2=1} \mathrm{Tr}\left[\boldsymbol{A}^\top \boldsymbol{A} \cdot f(\boldsymbol{B}\boldsymbol{B}^\top)\right] - 2 \cdot \mathrm{Tr}\left[\boldsymbol{B}\boldsymbol{A}\right], \quad (13)$$

where we have rescaled the function f by $1/c_1^2$. This follows from the fact that the multiplicative factor c_1 can be pushed inside $\boldsymbol{A}$. Note that such scaling does not affect the properties of gradient-based algorithms (modulo a constant change in their speed). Hence, without loss of generality, we will state and prove all our results for the problem (13).

Weight-tied gradient flow. We start by considering the weight-tied setting

$$\boldsymbol{A} = \beta \boldsymbol{B}^\top, \quad \beta \in \mathbb{R}. \quad (14)$$

This is motivated by the fact that the lower bounds on the population risk are approached by weight-tied matrices (see

Theorem 4.2 and Proposition 4.4). We keep the presentation brief and informal, and the formal setup is deferred to Appendix D. Note that for all $\boldsymbol{A}, \boldsymbol{B}$ under the weight-tying constraint (14), the optimal β^* in (13) can be found exactly. Thus, to optimize (13), we perform a (Riemannian) gradient flow on $\boldsymbol{B}$ with rows constrained to the unit sphere, where at each time t we pick the optimal $\beta^*(t)$:

$$\frac{\partial \boldsymbol{b}_i(t)}{\partial t} = -\boldsymbol{J}_i(t)\nabla_{\boldsymbol{b}_i}\Psi(\beta^*(t), \boldsymbol{B}(t)), \tag{15}$$

where $\boldsymbol{b}_i(t)$ stands for the i-th row of $\boldsymbol{B}(t)$ and $\boldsymbol{J}_i(t)$ projects the gradient $\nabla_{\boldsymbol{b}_i}\Psi(\beta(t), \boldsymbol{B}(t))$ of (13) under the weight-tying constraint on the unit sphere (see (64)-(65) in Appendix D for the exact expressions). This ensures that $\|\boldsymbol{b}_i(t)\|_2 = 1$ along the gradient flow trajectory.

Theorem 4.5 (Informal)**.** *For any $r \leq 1$, the gradient flow (15) initialized with full rank unit-row norm $\boldsymbol{B}$ converges to a global minimizer of (13), given by $\boldsymbol{B}\boldsymbol{B}^\top = \boldsymbol{I}$.*

Proof sketch of Theorem 4.5. It can be readily shown that the $\boldsymbol{B}$'s for which $\boldsymbol{B}\boldsymbol{B}^\top = \boldsymbol{I}$ are the unique minimizers of (13) under the weight-tying constraint and they satisfy the stationary point condition of the gradient flow (15). However, if $\boldsymbol{B}$ becomes not full-rank, such subspaces are never escaped by the gradient flow (15) (see Lemma D.2). Hence, the procedure would fail to converge to the global minimizer that has full-rank. We show that, under the full-rank initialization, this does not happen by lower bounding the time derivative of $\log\det(\boldsymbol{B}(t)\boldsymbol{B}(t)^\top)$ (see Lemma D.3), which vanishes uniquely at $\boldsymbol{B}\boldsymbol{B}^\top = \boldsymbol{I}$. This ensures that the solution of (15) will not saturate to a low-rank subspace. □

In Appendix D, we also provide a quantitative bound on the convergence time (see Lemma D.4). We remark that Theorem 4.5 holds for any d and for all full-rank initializations.

Projected gradient descent. We now move to the setting where the encoder and decoder weights are not weight-tied. In this case, we consider the commonly used Gaussian initialization and prove a result for sufficiently large d. The Gaussian initialization allows us to relax the requirement on f: we only need $c_2 = 0$, as opposed to the previous assumption that $c_{2\ell} = 0$ for any $\ell \in \mathbb{N}$ (see the statement of Lemma 4.1). Specifically, we consider the following algorithm to minimize (13):

$$\begin{aligned}&\boldsymbol{A}(t) = \boldsymbol{B}(t)^\top\left(f(\boldsymbol{B}(t)\boldsymbol{B}(t)^\top)\right)^{-1}\\ &\boldsymbol{B}'(t) := \boldsymbol{B}(t) - \eta\nabla_{\boldsymbol{B}(t)}, \boldsymbol{B}(t+1) := \mathrm{proj}(\boldsymbol{B}'(t)),\end{aligned} \tag{16}$$

where $\boldsymbol{A}(t)$ is the optimal matrix for a fixed $\boldsymbol{B}(t)$ and $\nabla_{\boldsymbol{B}(t)}$ (see (86) in Appendix E) is the projected gradient of the objective (13) with respect to $\boldsymbol{B}(t)$. Furthermore, $\mathrm{proj}(\boldsymbol{B}'(t))$ rescales all the rows to have unit norm. It will become apparent from the proof of Theorem 4.6 that the inversion in the definition of $\boldsymbol{A}(t)$ is indeed well defined. We remark that (16) can be viewed as the discrete counterpart of the Riemannian gradient flow for the weight-tied case (with the optimal $\boldsymbol{A}(t)$ in place of the weight-tying), where the application of $\mathrm{proj}(\cdot)$ keeps the rows of $\boldsymbol{B}(t)$ of unit norm. In the related literature, this procedure is often referred to as Riemannian gradient descent (see, e.g., Absil et al. (2009)). Alternatively, (16) may be viewed as coordinate descent (Wright, 2015) on the objective (13), where the step in $\boldsymbol{A}$ is performed exactly.

Our main result is that the projected gradient descent (16) converges to the global optimum of (13) for large enough d (with high probability). We give a sketch of the argument and defer the complete proof to Appendix E.

Theorem 4.6. *Consider the projected gradient descent (16) applied to the objective (13) for any f of the form $f(x) = x + \sum_{\ell=3} c_\ell^2 x^\ell$, where $\sum_{\ell=3} c_\ell^2 < \infty$. Initialize the algorithm with $\boldsymbol{B}(0)$ equal to a row-normalized Gaussian, i.e., $\boldsymbol{B}'_{i,j}(0) \sim \mathcal{N}(0, 1/d)$, $\boldsymbol{B}(0) = \mathrm{proj}(\boldsymbol{B}'(0))$. Let the step size η be $\Theta(1/\sqrt{d})$. Then, for any $r < 1$ and sufficiently large d, with probability at least $1 - Ce^{-cd}$, we have that $\boldsymbol{B}(t)\boldsymbol{B}(t)^\top$ converges to $\boldsymbol{I}$, which is the unique global optimum of (13). Moreover, defining $t = T/\eta$, we have the following bound on the rate of convergence*

$$\left\|\boldsymbol{B}(t)\boldsymbol{B}(t)^\top - \boldsymbol{I}\right\|_{op} \leq C(1-c)^T,$$

where $C > 0$ and $c \in (0, 1]$ are universal constants depending only on r and f.

Proof sketch of Theorem 4.6. Let $\boldsymbol{B}(0)\boldsymbol{B}(0)^\top = \boldsymbol{U}\boldsymbol{\Lambda}(0)\boldsymbol{U}^\top$ be the singular value decomposition (SVD) of the encoder Gram matrix. Then, the idea is to decompose $\boldsymbol{B}(t)\boldsymbol{B}(t)^\top$ at each step of the projected gradient descent dynamics as

$$\boldsymbol{B}(t)\boldsymbol{B}(t)^\top = \boldsymbol{I} + \boldsymbol{Z}(t) + \boldsymbol{X}(t), \tag{17}$$

where $\boldsymbol{Z}(t) = \boldsymbol{U}(\boldsymbol{\Lambda}(t) - \boldsymbol{I})\boldsymbol{U}^\top$. Here, $\boldsymbol{I}$ is the global optimum towards which we want to converge; $\boldsymbol{Z}(t)$ captures the evolution of the eigenvalues while keeping the eigenbasis fixed, as $\boldsymbol{U}$ comes from the SVD at initialization; and $\boldsymbol{X}(t)$ is the remaining error term capturing the change in the eigenbasis. The update on $\boldsymbol{\Lambda}(t)$ is given by $\boldsymbol{\Lambda}(t+1) = g(\boldsymbol{\Lambda}(t))$, where $g : \mathbb{R}^n \to \mathbb{R}^n$ admits an explicit expression. Hence, in light of this explicit expression, if we had $\boldsymbol{X}(t) \equiv 0$, then the desired convergence would follow from the analysis of the recursion for $\boldsymbol{\Lambda}(t)$ (see Lemma G.3).

The main technical difficulty lies in carefully controlling the error term $\boldsymbol{X}(t)$. In particular, we will show that $\boldsymbol{X}(t)$ decays for large enough d, which means that dynamics (17) is well approximated by $\boldsymbol{I} + \boldsymbol{Z}(t)$. The proof can be broken down in four steps. In the *first step*, we compute the leading order term of $\nabla_{\boldsymbol{B}(t)}$ (see Lemma E.2 and E.3). This simplifies the formula for $\nabla_{\boldsymbol{B}(t)}$, which can then be

expressed as an explicit nonlinear function of $\boldsymbol{Z}(t)$ and $\boldsymbol{X}(t)$. In the *second step*, we perform a Taylor expansion of $\nabla_{\boldsymbol{B}(t)}$, seen as a matrix-valued function in $\boldsymbol{Z}(t)$ and $\boldsymbol{X}(t)$ (see Lemma E.4). The intuition for this expansion comes from the fact that $\boldsymbol{X}(t)$ is a small quantity, and also $\|\boldsymbol{Z}(t)\|_{op} \to 0$ as $t \to \infty$. In the *third step*, we show that the norm of $\nabla_{\boldsymbol{B}(t)}$ vanishes sufficiently fast (see Lemma E.5), which implies that the projection step $\boldsymbol{B}(t+1) := \mathrm{proj}(\boldsymbol{B}'(t))$ has a negligible effect (see Lemma E.6). After doing these estimates, we finally obtain an explicit recursion for $\boldsymbol{X}(t)$. In the *fourth step*, we analyse this recursion (see Lemma E.7): first, we show that the error does not amplify too strongly (as in Gronwall's inequality); then, armed with this worst-case estimate, we can prove an exponential decay for $\boldsymbol{X}(t)$, which suffices to conclude the argument. □

Further discussions on the results in this section can be found in Appendix F.

5. Extension to General Covariance

In this section, we consider a Gaussian source with general covariance structure, i.e., $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{D}^2\boldsymbol{U}^\top$. Without loss of generality, the matrix $\boldsymbol{D}$ can be written as

$$\boldsymbol{D} = \mathrm{Diag}([\underbrace{D_1, \cdots, D_1}_{\times k_1} | \cdots | \underbrace{D_K, \cdots, D_K}_{\times k_K}]), \tag{18}$$

where $\sum_{i=1}^K k_i = d$, $k_i \geq 1$ and $D_i > D_{i+1} \geq 0$. We start by deriving a closed-form expression for the population risk, similar to Lemma 4.1. Its proof is given in Appendix B.

Lemma 5.1. *Let $\sigma \in L^2(\mathbb{R}, \mu)$ be an odd homogeneous activation, then $\widetilde{\mathcal{R}}(r)$ is equal to the minimum of*

$$\frac{1}{d}\left(\mathrm{Tr}\left[\boldsymbol{A}^\top \boldsymbol{A} f(\boldsymbol{B}\boldsymbol{B}^\top)\right] - 2c_1 \cdot \mathrm{Tr}\left[\boldsymbol{B}\boldsymbol{D}\boldsymbol{A}\right] + \mathrm{Tr}\left[\boldsymbol{D}^2\right]\right) \tag{19}$$

under the constraint $\|\boldsymbol{B}_{i,:}\|_2 = 1$. Moreover, $\widehat{\mathcal{R}}(r) = \widetilde{\mathcal{R}}(r)$.

Lemma 5.1 can be extended to any odd $\sigma \in L^2(\mathbb{R}, \mu)$ at the cost of losing the equivalence between $\widehat{\mathcal{R}}(r)$ and $\widetilde{\mathcal{R}}(r)$.

We restrict the theoretical analysis to proving a lower bound on (19) in the weight-tied setting (14). This lower bound is achieved via the choice of $\boldsymbol{A}, \boldsymbol{B}$ in Proposition 5.3, and we give numerical evidence (see Figure 6 in Appendix I) that gradient descent saturates the bound *without* the weight-tying constraint. Thus, we expect our lower bound to hold also for general (not necessarily weight-tied) matrices. The lower bound is given by the minimum of

$$\frac{1}{d}\left(\frac{g(1)}{n}\left(\sum_{i=1}^K \beta_i\right)^2 + \sum_{i=1}^K \left(c_1^2 \frac{\beta_i^2}{s_i} - 2c_1 D_i \beta_i + D_i^2\right)\right) \tag{20}$$

over all $\beta_i \geq 0$ and

$$0 \leq s_i \leq \min\{k_i, n\},\ 1 \leq \textstyle\sum_{i=1}^K s_i \leq \min\{d, n\}. \tag{21}$$

Here $g(x) := f(x) - c_1^2 x$, and we use the convention that $\frac{0^2}{0} = 0$ and $\frac{c}{0} = +\infty$ for $c > 0$. We can also explicitly characterize the optimal s_i, β_i. The optimal s_i are obtained via a *water-filling criterion*:

$$\boldsymbol{s} = [k_1, \cdots, k_{\mathrm{id}(n)-1}, \mathrm{res}(n), 0, \cdots, 0], \tag{22}$$

where $\boldsymbol{s} = [s_1, \cdots, s_k]$, $\mathrm{id}(n)$ denotes the first position at which $\min\{n, d\} - \sum_{i=1}^{\mathrm{id}(n)} k_i < 0$, and $\mathrm{res}(n) := \min\{n, d\} - \sum_{i=1}^{\mathrm{id}(n)-1} k_i$. The β_i can also be expressed explicitly in terms of f, s_i, D_i. This is summarized in the following theorem.

Theorem 5.2. *Consider the objective (19) under the weight-tied constraint (14). Then,*

$$(19) \geq \mathrm{LB}(\boldsymbol{D}) := \min_{s_i, \beta_i} (20), \tag{23}$$

where $\beta_i \geq 0$ and the s_i satisfy (21). Furthermore, the minimizers of (20) are the s_i obtained via the water-filling *criterion (22) and*

$$\beta_i = \begin{cases} \frac{s_i}{c_1} \cdot \left(\frac{\frac{g(1)}{c_1^2 n} \sum_{j=1}^{M^*} s_j \Delta_j + D_1}{\frac{g(1)}{c_1^2 n} \sum_{j=1}^{M^*} s_j + 1} - \Delta_i \right), & i \leq M^*, \\ 0, & i > M^*, \end{cases} \tag{24}$$

where $\Delta_j = D_1 - D_j$ and M^* is smallest index such that

$$\frac{g(1)}{c_1^2 n} \textstyle\sum_{j=1}^{M^*+1} s_j (D_{M^*+1} - D_j) + D_{M^*+1} \leq 0.$$

If no such index exists, then $M^* = K$.

We give a high-level overview of the proof below, and the complete argument is provided in Appendix H.

Proof sketch of Theorem 5.2. We first show that (23) holds. Consider the following block decomposition of $\boldsymbol{B}$ having the same block structure as $\boldsymbol{D}$:

$$\boldsymbol{B} = [\boldsymbol{\Gamma}_1 \boldsymbol{B}_1 | \cdots | \boldsymbol{\Gamma}_K \boldsymbol{B}_K], \tag{25}$$

where $\boldsymbol{B}_j \in \mathbb{R}^{n \times k_j}$ with $\|(\boldsymbol{B}_j)_{i,:}\|_2 = 1$ and $\{\boldsymbol{\Gamma}_j\}_{j=1}^K$ are diagonal matrices with $\sum_{j=1}^K \boldsymbol{\Gamma}_j^2 = \boldsymbol{I}$. Each $\boldsymbol{B}_i$ plays a similar role to the $\boldsymbol{B}$ in the isotropic case. The crucial bound for this step comes from Theorem A in Khare (2021):

$$(\boldsymbol{\Gamma}_i \boldsymbol{B}_i \boldsymbol{B}_i^\top \boldsymbol{\Gamma}_i)^{\circ 2} \succeq \frac{1}{s_i} \cdot \mathrm{Diag}(\boldsymbol{\Gamma}_i^2) \mathrm{Diag}(\boldsymbol{\Gamma}_i^2)^\top,$$

where $s_i = \mathrm{rank}(\boldsymbol{B}_i \boldsymbol{B}_i^\top)$. Now, ignoring the (PSD) cross-terms for $i \neq j$ we can proceed as in the proof of Proposition 4.3 to arrive at (20). It then remains to minimize (20), which is done using tools from convex analysis. □

Asymptotic achievability. We show that the lower bound in Theorem 5.2 can be asymptotically (i.e, as $d \to \infty$)

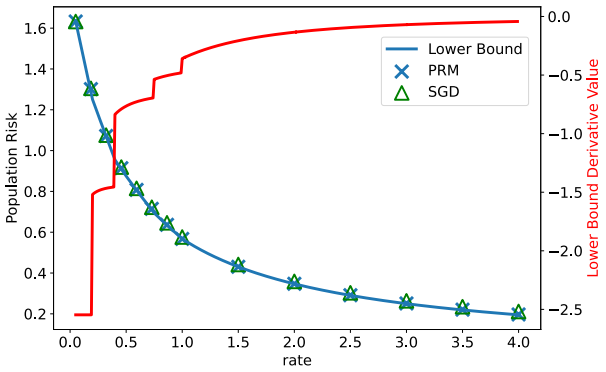

Figure 3. Performance comparison for the compression ($\sigma \equiv \text{sign}$) of an non-isotropic Gaussian source for $\boldsymbol{k} = (20, 20, 35, 25)$ and $(D_1, D_2, D_3, D_4) = (2, 1.5, 1, 0.8)$.

achieved by using the block form (25), after carefully picking $\boldsymbol{B}_i$ for each block. Specifically, first we generate a matrix $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ which is sampled uniformly from the group of orthogonal matrices. Next, we choose each $\boldsymbol{B}_i$ such that $\hat{\boldsymbol{B}}_i\hat{\boldsymbol{B}}_i^\top = \frac{n}{k_i}\boldsymbol{U}\boldsymbol{D}_i\boldsymbol{U}^\top$, where $\boldsymbol{D}_i$ is a diagonal matrix with

$$(\boldsymbol{D}_i)_{v,v} = \begin{cases} 1, & \text{if} \quad \sum_{j=1}^{i-1} k_j < v \le \sum_{j=1}^{i} k_j, \\ 0, & \text{otherwise}, \end{cases}$$

and the rows of $\boldsymbol{B}_i$ are given by $\boldsymbol{b}_i = \frac{\hat{\boldsymbol{b}}_i}{\|\hat{\boldsymbol{b}}_i\|_2}$. Furthermore, we pick $\boldsymbol{\Gamma}_i^2 = \frac{\gamma_i}{n}\boldsymbol{I}$ and $\boldsymbol{A} = \beta\boldsymbol{B}^\top$. The scalings γ_i and β are chosen such that $\beta_i := \beta\gamma_i$ are the minimizers of (20) for s_i as in (21). This is formalized in the following proposition.

Proposition 5.3. *Assume $\boldsymbol{A}, \boldsymbol{B}$ are constructed as described above and fix $r > 0$. Also assume that, for all i, $\frac{k_i}{n}$ converges to a strictly positive number as $d \to \infty$. Then, for any $\epsilon > 0$, with probability $1 - \frac{c}{d^2}$,*

$$|\mathcal{R}(\boldsymbol{A}, \boldsymbol{B}) - \text{LB}(\boldsymbol{D})| \le Cd^{-\frac{1}{2}+\epsilon},$$

where $\text{LB}(\boldsymbol{D})$ is defined in (23), and the constants c, C only depend on r, ϵ and $\lim_{d\to\infty}\frac{k_i}{n}$.

The proof of this lemma is similar to that of Proposition 4.4, and it is provided in Appendix H.

Taken together, Proposition 5.3 and Theorem 5.2 show that the optimal $\boldsymbol{B}$ exhibits the block structure (25), which agrees with the block structure (18) of the data covariance. The individual blocks are orthogonal in the sense that $\boldsymbol{B}_i^\top\boldsymbol{\Gamma}_i\boldsymbol{\Gamma}_j\boldsymbol{B}_j = \boldsymbol{0}$. Furthermore, we expect each block to have the same form as the minimizers in the isotropic case, up to some scaling. Such a structure is also confirmed by the numerical experiments: for instance, it is observed in the setting considered for Figure 6 in Appendix I.

Interpretation of the water-filling solution. To provide an intuitive illustration of the property of solutions in Proposition 5.3, consider the case of $K = 2$ and $k_1 = k_2$. In this case, the eigenvalues of $\boldsymbol{B}\boldsymbol{B}^\top$ will take only the two values λ_1, λ_2 corresponding to the two blocks. For rate $r \le k_1/d = 1/2$, we have $\lambda_2 = 0$ since *water-filling* implies that only the block corresponding to D_1 is utilized by $\boldsymbol{B}$. For rate $r > 1/2$, we have $\lambda_2 > 0$, as soon as

$$\left(1 + \frac{2c_1^2}{f(1) - c_1^2}r\right) D_2 > D_1.$$

The inequality can be obtained by evaluating explicitly the condition stated in Theorem 5.2 below Equation (24) in this special case. Furthermore, in the limit $r \to \infty$, we have that $\frac{\lambda_1}{\lambda_2} \to \frac{D_1}{D_2}$. This means that, for sufficiently large rates, the weight given by the encoder to each block is proportional to the corresponding eigenvalue of the data.

The *water-filling* behaviour can be observed in a setting with four blocks in Figure 3. Namely, at each of the rates $\{k_1/d, (k_1+k_2)/d, (k_1+k_2+k_3)/d, (k_1+k_2+k_3+k_4)/d\}$ that correspond to the earlier blocks being utilized to their full capacity, the derivative of the lower bound experiences a "jump" at which the next λ_i becomes positive.

6. Discussion

Population vs. empirical loss. All our results hold for the optimization of the population loss. Extending them to the empirical loss is an interesting direction for future research. One possible way forward is to exploit progress towards relating the landscape of empirical and population losses, see e.g. (Mei et al., 2018). We remark that, in the simulations of gradient descent, we always use the tempered straight-through estimator of the sign activation (see Appendix I for details). Thus, another promising direction is to show that, in the low-temperature regime (i.e., when the differentiable approximation of the sign becomes almost perfect), the gradient-based scheme converges to the minimizer of the population risk.

Optimality of two-layer autoencoders. This paper characterizes the minimizers of the expected ℓ_2 error incurred by two-layer autoencoders, and it shows that the minimum error is achieved, under certain conditions, by gradient-based algorithms. Thus, for the special case in which $\sigma \equiv \text{sign}$, a natural question is to what degree the model (1) is suitable for data compression. Let us fix the encoder to be a rotationally invariant matrix, i.e., $\boldsymbol{B} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^\top$ with $\boldsymbol{U}, \boldsymbol{V}$ independent and distributed according to the Haar measure and $\boldsymbol{\Lambda}$ having bounded entries. Then, the information-theoretically optimal reconstruction error can be computed via the replica method from statistical mechanics (Tulino et al., 2013) and, in a number of scenarios, it coincides with the error of a Vector Approximate Message Passing (VAMP) algorithm (Rangan et al., 2019; Schniter et al., 2016). Furthermore, it is also possible to optimize the spectrum $\boldsymbol{\Lambda}$ to minimize the error, which leads to the singular values of $\boldsymbol{B}$ being all

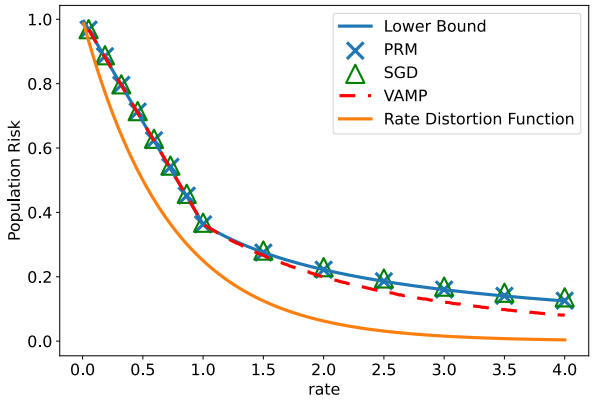

Figure 4. Performance comparison for the compression ($\sigma \equiv \text{sign}$) of an isotropic Gaussian source.

1 (Ma et al., 2021).[2] Surprisingly, for a compression rate $r \leq 1$, the optimal error found in (Ma et al., 2021) *coincides* with the minimizer of the population loss given by Theorem 4.2. Hence, two-layer autoencoders are optimal compressors under two conditions: *(i)* $r \leq 1$, and *(ii)* fixed encoder given by a rotationally invariant matrix. Both conditions are sufficient and also necessary. For $r > 1$, VAMP outperforms the two-layer autoencoder. Moreover, for a *general encoder/decoder pair*, the information-theoretically optimal reconstruction error is given by the rate-distortion function, which outperforms two-layer autoencoders for all $r > 0$. This picture is summarized in Figure 4: the blue curve represents the lower bound of Theorem 4.2 (for $r \leq 1$) and Proposition 4.3 (for $r > 1$), which is met by either running GD on the population risk (blue crosses) or SGD on samples taken from a isotropic Gaussian (green triangles) when $d = 100$;[3] this lower bound meets the performance of VAMP (red curve) if and only if $r \leq 1$; finally, the rate distortion function (orange curve) provides the best performance for all $r > 0$.

Universality of Gaussian predictions. Figure 4 (and also Figure 6 in Appendix I) show that gradient descent achieves the minimum of the population risk for the compression of Gaussian sources. Going beyond Gaussian inputs, to real-world datasets, Figures 1-2 (as well as those in Appendix I) show an excellent agreement between our predictions (using the empirical covariance of the data) and the performance of autoencoders trained on standard datasets (CIFAR-10, MNIST). As such, this agreement provides a clear indication of the universality of our predictions. In this regard, a flurry of recent research (see e.g. (Hastie et al., 2022; Hu & Lu, 2022; Loureiro et al., 2021; Goldt et al., 2022; Dudeja et al., 2022; Montanari & Saeed, 2022; Wang et al., 2022)) has proved that the Gaussian predictions actually hold in a much wider range of models. While none of the existing works exactly fits the setting considered in this paper, this gives yet another indication that our predictions should remain true more generally. The rigorous characterization of this universality is left for future work.

The choice of the activation function. The sign activation function constitutes an important special case of our analysis. However, our results hold for a broader class of activations. In particular, under the restriction that the rows of the encoder $\boldsymbol{B}$ lie on the unit sphere, all the results apply *for any odd activation*. The reason to fix the norm of the rows of the encoder is to prevent the network from entering the linear regime (e.g., by scaling $B \to \epsilon B$ and $A \to \frac{1}{\epsilon} A$). In fact, in the linear regime, perfect recovery can be achieved and this case has been well studied, see e.g. (Baldi & Hornik, 1989; Kunin et al., 2019; Gidel et al., 2019). We also note that, if the activation function is homogeneous, the restriction on the norm of the rows of $\boldsymbol{B}$ can be lifted, as the norm can be scaled out. Extending our analysis to activation functions that are not odd (e.g., ReLU) is an exciting avenue for future research. To achieve this goal, we expect that novel ideas will be needed, since our current approach relies on the fact that the Hermite expansion of the activation function (4) has only odd monomials.

[2] More specifically, Ma et al. (2021) consider an expectation propagation (EP) algorithm (Minka, 2001; Opper et al., 2005; Fletcher et al., 2016; He et al., 2017), which has been related to various forms of approximate message passing (Ma & Ping, 2017; Rangan et al., 2019).

[3] For further details on the experimental setup, see Appendix I.

Acknowledgements

Aleksandr Shevchenko, Kevin Kögler 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

Absil, P.-A., Mahony, R., and Sepulchre, R. Optimization algorithms on matrix manifolds. Princeton University Press, 2009.

Agustsson, E., Mentzer, F., Tschannen, M., Cavigelli, L., Timofte, R., Benini, L., and Gool, L. Soft-to-hard vector quantization for end-to-end learning compressible representations. In *NeurIPS*, 2017.

Arimoto, S. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, 18(1):14–20, 1972. doi: 10.1109/TIT.1972.1054753.

Baldi, P. and Hornik, K. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

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

Ballé, J., Chou, P. A., Minnen, D., Singh, S., Johnston, N., Agustsson, E., Hwang, S. J., and Toderici, G. Nonlinear transform coding. *IEEE Trans. on Special Topics in Signal Processing*, 15, 2021. URL https://arxiv.org/pdf/2007.03034.

Bao, X., Lucas, J., Sachdeva, S., and Grosse, R. B. Regularized linear autoencoders recover the principal components, eventually. In *NeurIPS*, 2020.

Bengio, Y., Yao, L., Alain, G., and Vincent, P. Generalized denoising auto-encoders as generative models. In *NeurIPS*, 2013.

Blahut, R. Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, 18(4):460–473, 1972. doi: 10.1109/TIT.1972.1054855.

Boufounos, P. T. and Baraniuk, R. G. 1-bit compressive sensing. In *2008 42nd Annual Conference on Information Sciences and Systems*, pp. 16–21. IEEE, 2008.

Boyd, S., Boyd, S. P., and Vandenberghe, L. *Convex optimization*. Cambridge University Press, 2004.

Candes, E. J., Strohmer, T., and Voroninski, V. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

Candes, E. J., Li, X., and Soltanolkotabi, M. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

Cheng, Z., Sun, H., Takeuchi, M., and Katto, J. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Conference on Computer Vision and Pattern Recognition*, 2020.

Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. In *NeurIPS*, 2019.

Ciliberti, S., Mézard, M., and Zecchina, R. Message-passing algorithms for non-linear nodes and data compression. *ComPlexUs*, 3(1-3):58–65, 2006.

Cover, T. M. and Thomas, J. A. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006. ISBN 0471241954.

del Pino, G. E. and Galaz, H. Statistical applications of the inverse gram matrix: A revisitation. *Brazilian Journal of Probability and Statistics*, pp. 177–196, 1995.

Dudeja, R., Sen, S., and Lu, Y. M. Spectral universality of regularized linear regression with nearly deterministic sensing matrices. *arXiv preprint arXiv:2208.02753*, 2022.

Fletcher, A., Sahraee-Ardakan, M., Rangan, S., and Schniter, P. Expectation consistent approximate inference: Generalizations and convergence. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pp. 190–194. IEEE, 2016.

Fletcher, A. K., Rangan, S., and Schniter, P. Inference in deep networks in high dimensions. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 1884–1888. IEEE, 2018.

Gidel, G., Bach, F., and Lacoste-Julien, S. Implicit regularization of discrete gradient dynamics in linear neural networks. In *NeurIPS*, 2019.

Goldt, S., Loureiro, B., Reeves, G., Krzakala, F., Mézard, M., and Zdeborová, L. The gaussian equivalence of generative models for learning with shallow neural networks. In *Mathematical and Scientific Machine Learning*, pp. 426–471. PMLR, 2022.

Gray, R. Vector quantization. *IEEE Assp Magazine*, 1(2): 4–29, 1984.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

He, H., Wen, C.-K., and Jin, S. Generalized expectation consistent signal recovery for nonlinear measurements. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pp. 2333–2337. IEEE, 2017.

Hinton, G. E. and Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science*, 313 (5786):504–507, 2006.

Hu, H. and Lu, Y. M. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 2022.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *NeurIPS*, 2018.

Jain, S., Radhakrishnan, A., and Uhler, C. A mechanism for producing aligned latent spaces with autoencoders. *arXiv preprint arXiv:2106.15456*, 2021.

Khare, A. Sharp nonzero lower bounds for the schur product theorem. *Proceedings of the American Mathematical Society*, 149(12):5049–5063, 2021.

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

Korada, S. B. and Urbanke, R. L. Polar codes are optimal for lossy source coding. *IEEE Transactions on Information Theory*, 56(4):1751–1768, 2010.

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

Lei, E., Hassani, H., and Bidokhti, S. S. Neural estimation of the rate-distortion function for massive datasets. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pp. 608–613. IEEE, 2022.

Liu, H., Chen, T., Guo, P., Shen, Q., Cao, X., Wang, Y., and Ma, Z. Non-local attention optimized deep image compression. *arXiv preprint arXiv:1904.09757*, 2019.

Loureiro, B., Gerbelot, C., Cui, H., Goldt, S., Krzakala, F., Mezard, M., and Zdeborová, L. Learning curves of generic features maps for realistic datasets with a teacher-student model. In *NeurIPS*, 2021.

Ma, J. and Ping, L. Orthogonal amp. *IEEE Access*, 5: 2020–2033, 2017.

Ma, J., Xu, J., and Maleki, A. Analysis of sensing spectral for signal recovery under a generalized linear model. In *NeurIPS*, 2021.

Matsumoto, N. and Mazumdar, A. Binary iterative hard thresholding converges with optimal number of measurements for 1-bit compressed sensing. *arXiv preprint arXiv:2207.03427*, 2022.

Meckes, E. S. *The random matrix theory of the classical compact groups*, volume 218. Cambridge University Press, 2019.

Mei, S., Bai, Y., and Montanari, A. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

Minka, T. P. Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pp. 362–369, 2001.

Montanari, A. and Saeed, B. N. Universality of empirical risk minimization. In *Conference on Learning Theory*, 2022.

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

Nguyen, T. V., Wong, R. K., and Hegde, C. On the dynamics of gradient descent for autoencoders. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Nguyen, T. V., Wong, R. K., and Hegde, C. Benefits of jointly training autoencoders: An improved neural tangent kernel analysis. *IEEE Transactions on Information Theory*, 67(7):4669–4692, 2021.

O'Donnell, R. *Analysis of boolean functions*. Cambridge University Press, 2014.

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

Opper, M., Winther, O., and Jordan, M. J. Expectation consistent approximate inference. *Journal of Machine Learning Research*, 6(12), 2005.

Radhakrishnan, A., Belkin, M., and Uhler, C. Overparameterized neural networks implement associative memory. *Proceedings of the National Academy of Sciences*, 117 (44):27162–27170, 2020.

Rangamani, A., Mukherjee, A., Basu, A., Arora, A., Ganapathi, T., Chin, S., and Tran, T. D. Sparse coding and autoencoders. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 36–40. IEEE, 2018.

Rangan, S., Schniter, P., and Fletcher, A. K. Vector approximate message passing. *IEEE Transactions on Information Theory*, 65(10):6664–6684, 2019.

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

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, 2014.

Santambrogio, F. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1):87–154, 2017.

Schniter, P., Rangan, S., and Fletcher, A. K. Vector approximate message passing for the generalized linear model. In *2016 50th Asilomar Conference on Signals, Systems and Computers*, pp. 1525–1529. IEEE, 2016.

Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

Shannon, C. E. Coding theorems for a discrete source with a fidelity criterion. *1959 IRE National Convention Record*, pp. 142–163, 1959.

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

Theis, L., Salimans, T., Hoffman, M. D., and Mentzer, F. Lossy compression with gaussian diffusion. *arXiv preprint arXiv:2206.08889*, 2022.

Tschannen, M., Bachem, O., and Lucic, M. Recent advances in autoencoder-based representation learning. *arXiv preprint arXiv:1812.05069*, 2018.

Tulino, A. M., Caire, G., Verdú, S., and Shamai, S. Support recovery with sparsely sampled free random matrices. *IEEE Transactions on Information Theory*, 59(7):4243–4271, 2013.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. Extracting and composing robust features with denoising autoencoders. In *International Conference on Machine Learning*, 2008.

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*, 304(1-3):45–68, 2000.

Wagner, A. B. and Ballé, J. Neural networks optimally compress the sawbridge. *2021 Data Compression Conference (DCC)*, pp. 143–152, 2021.

Wainwright, M. J., Maneva, E., and Martinian, E. Lossy source compression using low-density generator matrix codes: Analysis and algorithms. *IEEE Transactions on Information theory*, 56(3):1351–1368, 2010.

Wang, T., Zhong, X., and Fan, Z. Universality of approximate message passing algorithms and tensor networks. *arXiv preprint arXiv:2206.13037*, 2022.

Wright, S. J. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, 2015.

Yang, R. and Mandt, S. Lossy image compression with conditional diffusion models. *arXiv preprint arXiv:2209.06950*, 2022.

Yang, Y. and Mandt, S. Towards empirical sandwich bounds on the rate-distortion function. In *International Conference on Learning Representations*, 2021.

Yang, Y., Mandt, S., and Theis, L. An introduction to neural data compression. *arXiv preprint arXiv:2202.06533*, 2022.

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. Additional Related Works

Rate-distortion formalism. Lossy compression of stationary sources is a classical problem in information theory, and several approaches have been proposed, including vector quantization (Gray, 1984), or the usage of powerful channel codes (Korada & Urbanke, 2010; Ciliberti et al., 2006; Wainwright et al., 2010). The rate-distortion function characterizes the optimal trade-off between error and size of the representation for the compression of an i.i.d. source (Shannon, 1948; 1959; Cover & Thomas, 2006). However, computing the rate-distortion function is by itself a challenging task. The Blahut-Arimoto scheme (Blahut, 1972; Arimoto, 1972) provides a systematic approach, but it suffers from the issue of scalability (Lei et al., 2022). Consequently, to compute the rate-distortion of empirical datasets, approximate methods based on generative modeling have been proposed (Yang & Mandt, 2021; Lei et al., 2022).

Non-linear inverse problems. The task of estimating a signal $\boldsymbol{x}$ from non-linear measurements $\boldsymbol{y} = \sigma(\boldsymbol{B}\boldsymbol{x})$ has appeared in many areas, such as 1-bit compressed sensing where $\sigma(z) = \text{sign}(z)$ (Boufounos & Baraniuk, 2008), or phase retrieval where $\sigma(z) = |z|$ (Candes et al., 2013; 2015). While the focus of these problems is different from ours (e.g., compressed sensing has often an additional sparsity assumption), the ideas and proof techniques developed in this paper might be beneficial to characterize the fundamental limits and the performance of gradient-based methods for general inverse reconstruction tasks, see e.g. (Ma et al., 2021; Matsumoto & Mazumdar, 2022).

B. Closed Forms for the Population Risk

For the proofs of Lemmas 4.1 and 5.1 in the current section, we assume that the rows of $\boldsymbol{B}$ have non-zero norm, hence, in particular, they may be chosen to have unit norm. In the end of the section, we elaborate on why this assumption holds true.

Let us also mention that we call σ odd in L^2 sense. For this particular case, it means that $\sigma(x) = \sigma(-x)$ for $x \neq 0$ and $|\sigma(0)| < C$, where C is some universal constant. This concern is purely technical, since the main application of our results is *1-bit* compression. Namely, we do not set $\sigma(0) = \text{sgn}(0) = 0$. In fact, this would mean that the compressed sequence can take values in $\{-1, 0, 1\}$, which would not result in *1-bit* compression, but rather in $\log_2(3)$-*bits* compression. It is safe to ignore this technicality and intuitively assume that $\sigma(0) = 0$.

Proof of Lemma 4.1. Opening up the two-norm gives

$$\mathbb{E}\|\boldsymbol{x} - \boldsymbol{A}\sigma(\boldsymbol{B}\boldsymbol{x})\|_2^2 = \mathbb{E}\|\boldsymbol{x}\|_2^2 + \mathbb{E}\|\boldsymbol{A}\sigma(\boldsymbol{B}\boldsymbol{x})\|_2^2 - 2\mathbb{E}\langle \boldsymbol{x}, \boldsymbol{A}\sigma(\boldsymbol{B}\boldsymbol{x})\rangle. \tag{26}$$

Since $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, we get

$$\mathbb{E}\|\boldsymbol{x}\|_2^2 = d. \tag{27}$$

Let $\boldsymbol{B}^\top = [\boldsymbol{b}_1, \dots, \boldsymbol{b}_n] \in \mathbb{R}^{d\times n}$ and $\boldsymbol{A} = [\boldsymbol{a}_1, \dots, \boldsymbol{a}_n] \in \mathbb{R}^{d\times n}$, with $\|\boldsymbol{b}_i\|_2 = \|\boldsymbol{B}_{i,:}\| = 1$. Rewriting the second term in (26) gives

$$\mathbb{E}\|\boldsymbol{A}\sigma(\boldsymbol{B}\boldsymbol{x})\|_2^2 = \sum_{i,j=1}^{n} \langle \boldsymbol{a}_i, \boldsymbol{a}_j\rangle \cdot \mathbb{E}\left[\sigma(\langle \boldsymbol{b}_i, \boldsymbol{x}\rangle) \cdot \sigma(\langle \boldsymbol{b}_j, \boldsymbol{x}\rangle)\right]. \tag{28}$$

Using the reproducing property of Hermite coefficients (see, e.g., Chapter 11 in (O'Donnell, 2014)), since the random variables $\langle \boldsymbol{b}_i, \boldsymbol{x}\rangle$ and $\langle \boldsymbol{b}_j, \boldsymbol{x}\rangle$ are $\langle \boldsymbol{b}_i, \boldsymbol{b}_j\rangle$-correlated, we have

$$\mathbb{E}\left[h_{2\ell+1}(\langle \boldsymbol{b}_i, \boldsymbol{x}\rangle) \cdot h_{2\ell+1}(\langle \boldsymbol{b}_j, \boldsymbol{x}\rangle)\right] = \langle \boldsymbol{b}_i, \boldsymbol{b}_j\rangle^{2\ell+1}, \quad \mathbb{E}\left[h_{2\ell+1}(\langle \boldsymbol{b}_i, \boldsymbol{x}\rangle) \cdot h_{2k+1}(\langle \boldsymbol{b}_j, \boldsymbol{x}\rangle)\right] = 0,$$

for $k \neq \ell$. This gives that

$$\mathbb{E}\left[\sigma(\langle \boldsymbol{b}_i, \boldsymbol{x}\rangle) \cdot \sigma(\langle \boldsymbol{b}_j, \boldsymbol{x}\rangle)\right] = \sum_{\ell=0}^{\infty} (c_{2\ell+1})^2 \langle \boldsymbol{b}_i, \boldsymbol{b}_j\rangle^{2\ell+1} = f(\langle \boldsymbol{b}_i, \boldsymbol{b}_j\rangle),$$

and, hence, using (28) we arrive to

$$\mathbb{E}\|\boldsymbol{A}\sigma(\boldsymbol{B}\boldsymbol{x})\|_2^2 = \sum_{i,j=1}^{n} \langle \boldsymbol{a}_i, \boldsymbol{a}_j\rangle \cdot f(\langle \boldsymbol{b}_i, \boldsymbol{b}_j\rangle) = \text{Tr}\left[\boldsymbol{A}^\top \boldsymbol{A} \cdot f(\boldsymbol{B}\boldsymbol{B}^\top)\right]. \tag{29}$$

Rearranging the last term in (26) gives

$$\mathbb{E}\langle \boldsymbol{x}, \boldsymbol{A}\sigma(\boldsymbol{B}\boldsymbol{x})\rangle = \sum_{i=1}^{d}\sum_{j=1}^{n} a_j^i \cdot \mathbb{E}[x_i \sigma(\langle \boldsymbol{b}_j, \boldsymbol{x}\rangle)], \tag{30}$$

where a_j^i stands for the i-th coordinate of the vector $\boldsymbol{a}_j$ and x_i stands for the i-th coordinate of the vector $\boldsymbol{x}$. Let us now compute the inner expected value for each pair (i,j). Notice that the random variables $\langle \boldsymbol{b}_j, \boldsymbol{x}\rangle$ and x_i are jointly Gaussian with zero mean and covariance matrix $\widetilde{\boldsymbol{\Sigma}} \in \mathbb{R}^{2\times 2}$:

$$\widetilde{\boldsymbol{\Sigma}}_{21} = \widetilde{\boldsymbol{\Sigma}}_{12} = \mathbb{E} x_i \langle \boldsymbol{b}_j, \boldsymbol{x}\rangle = \mathbb{E} b_j^i x_i^2 = b_j^i, \quad \widetilde{\boldsymbol{\Sigma}}_{11} = \mathbb{E}\langle \boldsymbol{b}_j, \boldsymbol{x}\rangle^2 = \|\boldsymbol{b}_j\|_2^2 = 1, \quad \widetilde{\boldsymbol{\Sigma}}_{22} = \mathbb{E} x_i^2 = 1.$$

Hence, the random vectors $(\langle \boldsymbol{b}_j, \boldsymbol{x}\rangle, x_i)$ and

$$\left(y_1, b_j^i \cdot y_1 + \sqrt{1-(b_j^i)^2}\cdot y_2\right), \quad \text{with } (y_1, y_2) \sim \mathcal{N}(0, \boldsymbol{I})$$

are identically distributed. In this view, we obtain

$$\begin{aligned}\mathbb{E}[x_i \sigma(\langle \boldsymbol{b}_j, \boldsymbol{x}\rangle)] &= \mathbb{E}\left[\left(b_j^i \cdot y_1 + \sqrt{1-(b_j^i)^2}\cdot y_2\right)\sigma(y_1)\right] \\ &= b_j^i \cdot \mathbb{E}[y_1\sigma(y_1)] + \sqrt{1-(b_j^i)^2}\cdot \mathbb{E}[y_2]\cdot \mathbb{E}[\sigma(y_1)] = c_1 \cdot b_j^i,\end{aligned} \tag{31}$$

where we applied the reproducing property to conclude that $\mathbb{E}[y_1\sigma(y_1)] = c_1$. Consequently, by combining (30) and (31), we get that

$$\mathbb{E}\langle \boldsymbol{x}, \boldsymbol{A}\sigma(\boldsymbol{B}\boldsymbol{x})\rangle = c_1 \cdot \sum_{i=1}^{d}\sum_{j=1}^{n} a_j^i b_j^i = c_1 \cdot \mathrm{Tr}\left[\boldsymbol{B}\boldsymbol{A}\right]. \tag{32}$$

By combining (26), (27), (29) and (32), we obtain the desired expression for $\widetilde{R}(r)$.

Assume now that σ is homogeneous. Then, in (28) and (30), the norm of $\boldsymbol{b}_i$ can be pushed into the corresponding $\boldsymbol{a}_i$ and, hence, we obtain

$$\min_{\boldsymbol{A},\boldsymbol{B}} \mathbb{E}\|\boldsymbol{x} - \boldsymbol{A}\sigma(\boldsymbol{B}\boldsymbol{x})\|_2^2 = \min_{\boldsymbol{A}, \|\boldsymbol{B}_i\|_2 = 1} \mathbb{E}\|\boldsymbol{x} - \boldsymbol{A}\sigma(\boldsymbol{B}\boldsymbol{x})\|_2^2,$$

which proves that $\widehat{\mathcal{R}}(r) = \widetilde{\mathcal{R}}(r)$.

Finally, consider the case $\sigma(x) = \mathrm{sign}(x)$. Then, Grothendieck's identity (see, e.g., Lemma 3.6.6 in (Vershynin, 2018)) gives

$$\mathbb{E}\sigma(\langle \boldsymbol{b}_i, \boldsymbol{x}\rangle)\sigma(\langle \boldsymbol{b}_j, \boldsymbol{x}\rangle) = \frac{2}{\pi}\arcsin(\langle \boldsymbol{b}_i, \boldsymbol{b}_j\rangle) \Rightarrow f(x) = \frac{2}{\pi}\arcsin(x).$$

Recalling that the first Hermite coefficient of $\sigma(x) = \mathrm{sign}(x)$ is equal to $\sqrt{\frac{2}{\pi}}$ finishes the proof. □

Proof of Lemma 5.1. The proof of Lemma 5.1 follows from similar arguments as that of Lemma 4.1. Given this, we only explain the key differences. We first show that it is enough to consider $\boldsymbol{\Sigma} = \boldsymbol{D}^2$. Given the SVD $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{D}^2\boldsymbol{U}^\top$, we have $\boldsymbol{x} = \boldsymbol{U}\boldsymbol{D}\tilde{\boldsymbol{x}}$, where $\tilde{\boldsymbol{x}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Now, we can push the rotation $\boldsymbol{U}$ in $\boldsymbol{A}, \boldsymbol{B}$:

$$\|\boldsymbol{x} - \boldsymbol{A}\sigma(\boldsymbol{B}\boldsymbol{x})\|_2 = \left\|\boldsymbol{D}\tilde{\boldsymbol{x}} - \boldsymbol{U}^\top \boldsymbol{A}\sigma(\boldsymbol{B}\boldsymbol{U}\boldsymbol{D}\tilde{\boldsymbol{x}})\right\|_2.$$

Thus, after replacing $\boldsymbol{A}$ with $\boldsymbol{U}^\top\boldsymbol{A}$ and $\boldsymbol{B}$ with $\boldsymbol{B}\boldsymbol{U}$, we may assume that $\boldsymbol{x} = \boldsymbol{D}\tilde{\boldsymbol{x}}$.

We again open up the two-norm

$$\mathbb{E}\|\boldsymbol{x} - \boldsymbol{A}\sigma(\boldsymbol{B}\boldsymbol{x})\|_2^2 = \mathbb{E}\|\boldsymbol{x}\|_2^2 + \mathbb{E}\|\boldsymbol{A}\sigma(\boldsymbol{B}\boldsymbol{x})\|_2^2 - 2\mathbb{E}\langle \boldsymbol{x}, \boldsymbol{A}\sigma(\boldsymbol{B}\boldsymbol{x})\rangle. \tag{33}$$

For the first term, we clearly have

$$\mathbb{E}\|\boldsymbol{x}\|_2^2 = \mathrm{Tr}\left[\boldsymbol{D}^2\right].$$

Now, for the second term we write

$$\mathbb{E}\|\boldsymbol{A}\sigma(\boldsymbol{B}\boldsymbol{x})\|_2^2 = \mathbb{E}\|\boldsymbol{A}\sigma(\boldsymbol{B}\boldsymbol{D}\tilde{\boldsymbol{x}})\|_2^2,$$

where $\tilde{\boldsymbol{x}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Thus, as in the proof of Lemma 4.1, we have

$$\mathbb{E}\|\boldsymbol{A}\sigma(\boldsymbol{B}\boldsymbol{D}\tilde{\boldsymbol{x}})\|_2^2 = \mathrm{Tr}\left[\boldsymbol{A}^\top\boldsymbol{A}\cdot f(\boldsymbol{B}\boldsymbol{D}^2\boldsymbol{B}^\top)\right].$$

Similarly, for the last term we obtain

$$\mathbb{E}\langle \boldsymbol{x}, \boldsymbol{A}\sigma(\boldsymbol{B}\boldsymbol{x})\rangle = \mathbb{E}\langle \tilde{\boldsymbol{x}}, \boldsymbol{D}\boldsymbol{A}\sigma(\boldsymbol{B}\boldsymbol{D}\tilde{\boldsymbol{x}})\rangle = c_1\,\mathrm{Tr}\left[\boldsymbol{D}\boldsymbol{A}\boldsymbol{B}\boldsymbol{D}\right].$$

Finally, since σ is homogeneous, by abuse of notation we can replace $\boldsymbol{B}\boldsymbol{D}$ by any $\boldsymbol{B}$ with unit-norm rows. This follows from the fact that, similarly to the proof of Lemma 4.1 (namely, equations (28) and (30)), we have that

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{A}\sigma(\boldsymbol{B}\boldsymbol{D}\tilde{\boldsymbol{x}})\|_2^2 &= \sum_{i,j=1}^{n}\langle \boldsymbol{a}_i, \boldsymbol{a}_j\rangle\cdot\mathbb{E}\left[\sigma(\langle(\boldsymbol{B}\boldsymbol{D})_{i,:}, \tilde{\boldsymbol{x}}\rangle)\cdot\sigma(\langle(\boldsymbol{B}\boldsymbol{D})_{j,:}, \tilde{\boldsymbol{x}}\rangle)\right],\\
\mathbb{E}\langle \tilde{\boldsymbol{x}}, \boldsymbol{D}\boldsymbol{A}\sigma(\boldsymbol{B}\boldsymbol{D}\tilde{\boldsymbol{x}})\rangle &= \sum_{i=1}^{d}\sum_{j=1}^{n} a_j^i\cdot\mathbb{E}[(D_{i,i}\cdot\tilde{x}_i)\cdot\sigma(\langle(\boldsymbol{B}\boldsymbol{D})_{j,:}, \tilde{\boldsymbol{x}}\rangle)],
\end{aligned} \tag{34}$$

which, by homogeneity, readily gives that the norm of $(\boldsymbol{B}\boldsymbol{D})_{i,:}$ can be pushed into the corresponding $\boldsymbol{a}_i$.

As a result, the statement of Lemma 5.1 readily follows by comparing the terms. □

Rows of $\boldsymbol{B}$ are non-zero. We show that the assumption holds true by contradiction. Without loss of generality, assume that the first $n' \le n$ rows of $\boldsymbol{B}$ are zero vectors. Hence, from (34) we can see that the following holds:

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{A}\sigma(\boldsymbol{B}\boldsymbol{D}\tilde{\boldsymbol{x}})\|_2^2 &= \sum_{i,j=n'+1}^{n}\langle \boldsymbol{a}_i, \boldsymbol{a}_j\rangle\cdot\mathbb{E}\left[\sigma(\langle(\boldsymbol{B}\boldsymbol{D})_{i,:}, \tilde{\boldsymbol{x}}\rangle)\cdot\sigma(\langle(\boldsymbol{B}\boldsymbol{D})_{j,:}, \tilde{\boldsymbol{x}}\rangle)\right]\\
&+ \sum_{i\le n'\,\wedge\, j>n'}\langle \boldsymbol{a}_i, \boldsymbol{a}_j\rangle\cdot\mathbb{E}\left[\sigma(0)\cdot\sigma(\langle(\boldsymbol{B}\boldsymbol{D})_{j,:}, \tilde{\boldsymbol{x}}\rangle)\right]\\
&+ \sum_{i>n'\,\wedge\, j\le n'}\langle \boldsymbol{a}_i, \boldsymbol{a}_j\rangle\cdot\mathbb{E}\left[\sigma(0)\cdot\sigma(\langle(\boldsymbol{B}\boldsymbol{D})_{i,:}, \tilde{\boldsymbol{x}}\rangle)\right] + \sum_{i,j\le n'}\langle \boldsymbol{a}_i, \boldsymbol{a}_j\rangle\cdot\sigma(0)^2\\
&= \sum_{i,j=n'+1}^{n}\langle \boldsymbol{a}_i, \boldsymbol{a}_j\rangle\cdot\mathbb{E}\left[\sigma(\langle(\boldsymbol{B}\boldsymbol{D})_{i,:}, \tilde{\boldsymbol{x}}\rangle)\cdot\sigma(\langle(\boldsymbol{B}\boldsymbol{D})_{j,:}, \tilde{\boldsymbol{x}}\rangle)\right] + \sum_{i,j\le n'}\langle \boldsymbol{a}_i, \boldsymbol{a}_j\rangle\cdot\sigma(0)^2\\
&\ge \sum_{i,j=n'+1}^{n}\langle \boldsymbol{a}_i, \boldsymbol{a}_j\rangle\cdot\mathbb{E}\left[\sigma(\langle(\boldsymbol{B}\boldsymbol{D})_{i,:}, \tilde{\boldsymbol{x}}\rangle)\cdot\sigma(\langle(\boldsymbol{B}\boldsymbol{D})_{j,:}, \tilde{\boldsymbol{x}}\rangle)\right],
\end{aligned} \tag{35}$$

where in the fourth line we used that for $\tilde{\boldsymbol{x}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, as σ is odd, the following identity holds:

$$\mathbb{E}\left[\sigma(\langle(\boldsymbol{B}\boldsymbol{D})_{j,:}, \tilde{\boldsymbol{x}}\rangle)\right] = 0,$$

and the last inequality follows from the fact that for the Gram matrix $\boldsymbol{M}$ of the vectors $\{\boldsymbol{a}_i\}_{i=1}^{n'}$:

$$\sum_{i,j\le n'}\langle \boldsymbol{a}_i, \boldsymbol{a}_j\rangle\cdot\sigma(0)^2 = \sigma(0)^2\cdot\langle \boldsymbol{1}, \boldsymbol{M}\boldsymbol{1}\rangle \ge 0.$$

Similarly, one can verify that

$$\mathbb{E}\langle \tilde{\boldsymbol{x}}, \boldsymbol{D}\boldsymbol{A}\sigma(\boldsymbol{B}\boldsymbol{D}\tilde{\boldsymbol{x}})\rangle = \sum_{i=1}^{d}\sum_{j=n'+1}^{n} a_j^i\cdot\mathbb{E}[(D_{i,i}\cdot\tilde{x}_i)\cdot\sigma(\langle(\boldsymbol{B}\boldsymbol{D})_{j,:}, \tilde{\boldsymbol{x}}\rangle)]. \tag{36}$$

Combining (35) and (36), and recalling the population risk form in (33), we conclude that

$$\mathcal{R}(\boldsymbol{A}, \boldsymbol{B}) \ge \mathcal{R}(\boldsymbol{A}_{:,n'+1:}, \boldsymbol{B}_{n'+1:,:}),$$

where $\boldsymbol{A}_{:,n'+1:}$ and $\boldsymbol{B}_{n'+1:,:}$ are obtained by removing the zero columns/rows from $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively. This means that considering a matrix $\boldsymbol{B}$ with zero rows is equivalent to looking at a smaller rate $r' < r$. We show in Theorem 4.2, Proposition 4.3 and Theorem 5.2 that the population risk is monotone in the rate. Thus, having zero rows in $\boldsymbol{B}$ is clearly sub-optimal.

C. Proofs of Lower Bound on Loss (Section 4.1)

C.1. Case $r \leq 1$

C.1.1. Lower bound on $\widetilde{R}(r)$

Lemma C.1. *Let $\boldsymbol{A} = [\boldsymbol{a}_1, \dots, \boldsymbol{a}_n] \in \mathbb{R}^{d\times n}$ and $\boldsymbol{B}^\top = [\boldsymbol{b}_1, \dots, \boldsymbol{b}_n] \in \mathbb{R}^{d\times n}$, with $\|\boldsymbol{b}_i\|_2 = 1$ for $i \in [n]$. Let c_1 and $f(\cdot)$ be defined as per Lemma 4.1. Then, the following bound holds:*

$$\mathcal{L}_l(\boldsymbol{A}, \boldsymbol{B}) := \operatorname{Tr}\left[\boldsymbol{A}^\top \boldsymbol{A} \cdot (\boldsymbol{B}\boldsymbol{B}^\top)^{\circ(2\ell+1)}\right] - \frac{2c_1}{f(1)} \cdot \operatorname{Tr}\left[\boldsymbol{B}\boldsymbol{A}\right] \geq -\frac{c_1^2}{(f(1))^2} \cdot n. \tag{37}$$

Proof of Lemma C.1. For any symmetric $\boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{T} \in \mathbb{R}^{n\times n}$, a direct computation readily gives that

$$\operatorname{Tr}\left[\boldsymbol{P} \cdot (\boldsymbol{Q} \circ \boldsymbol{T})\right] = \operatorname{Tr}\left[(\boldsymbol{P} \circ \boldsymbol{Q}) \cdot \boldsymbol{T})\right]. \tag{38}$$

Thus, by taking $\boldsymbol{P} = \boldsymbol{A}^\top \boldsymbol{A}$, $\boldsymbol{Q} = (\boldsymbol{B}\boldsymbol{B}^\top)^{\circ\ell}$ and $\boldsymbol{T} = (\boldsymbol{B}\boldsymbol{B}^\top)^{\circ(\ell+1)}$, we obtain

$$\operatorname{Tr}\left[\boldsymbol{A}^\top \boldsymbol{A} \cdot (\boldsymbol{B}\boldsymbol{B}^\top)^{\circ(2\ell+1)}\right] = \operatorname{Tr}\left[(\boldsymbol{A}^\top \boldsymbol{A} \circ (\boldsymbol{B}\boldsymbol{B}^\top)^{\circ\ell}) \cdot (\boldsymbol{B}\boldsymbol{B}^\top \circ (\boldsymbol{B}\boldsymbol{B}^\top)^{\circ\ell})\right].$$

Note that $\boldsymbol{B}\boldsymbol{B}^\top$ is PSD and, therefore, $(\boldsymbol{B}\boldsymbol{B}^\top)^{\circ\ell}$ is also PSD by Schur product theorem. Furthermore, as the rows of B have unit norm, $(\boldsymbol{B}\boldsymbol{B}^\top)^{\circ\ell}$ has unit diagonal. As a result, if we show that, for any PSD matrix $\boldsymbol{Q}$ with unit diagonal entries,

$$\operatorname{Tr}\left[(\boldsymbol{A}^\top \boldsymbol{A} \circ \boldsymbol{Q}) \cdot (\boldsymbol{B}\boldsymbol{B}^\top \circ \boldsymbol{Q})\right] - \frac{2c_1}{f(1)} \cdot \operatorname{Tr}\left[\boldsymbol{B}\boldsymbol{A}\right] \geq -\frac{c_1^2}{(f(1))^2} \cdot n, \tag{39}$$

then the claim (37) immediately follows.

As $\boldsymbol{Q}$ is a PSD matrix with unit diagonal, it admits the following decomposition

$$\boldsymbol{Q} = \sum_{i=1}^n \boldsymbol{u}_i \boldsymbol{u}_i^\top, \quad \boldsymbol{D}_i = \operatorname{Diag}(\boldsymbol{u}_i), \quad \sum_{i=1}^n \boldsymbol{D}_i^2 = \boldsymbol{I}. \tag{40}$$

In this view, defining

$$\boldsymbol{A}_i = \boldsymbol{A}\boldsymbol{D}_i, \quad \boldsymbol{B}_i = \boldsymbol{D}_i \boldsymbol{B},$$

we can rewrite the LHS of (39) in a more convenient form for further analysis. In particular, for the second term we deduce the following

$$\operatorname{Tr}\left[\boldsymbol{B}\boldsymbol{A}\right] = \operatorname{Tr}\left[\boldsymbol{A}\boldsymbol{B}\right] = \operatorname{Tr}\left[\boldsymbol{A} \cdot \left(\sum_{i=1}^n \boldsymbol{D}_i^2\right) \cdot \boldsymbol{B}\right] = \sum_{i=1}^n \operatorname{Tr}\left[\boldsymbol{A} \cdot \boldsymbol{D}_i^2 \cdot \boldsymbol{B}\right] = \sum_{i=1}^n \operatorname{Tr}\left[(\boldsymbol{A}\boldsymbol{D}_i) \cdot (\boldsymbol{D}_i \boldsymbol{B})\right] = \sum_{i=1}^n \operatorname{Tr}\left[\boldsymbol{A}_i \boldsymbol{B}_i\right].$$

Let us now rearrange the first term of (39). Notice that

$$(\boldsymbol{A}^\top \boldsymbol{A} \circ \boldsymbol{Q})_{i,j} = \sum_{k=1}^n \langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle \cdot u_k^i u_k^j = \sum_{k=1}^n \langle \boldsymbol{a}_i \cdot u_k^i, \boldsymbol{a}_j \cdot u_k^j \rangle = \sum_{k=1}^n ((\boldsymbol{A}\boldsymbol{D}_k)^\top \cdot (\boldsymbol{A}\boldsymbol{D}_k))_{i,j} = \sum_{k=1}^n (\boldsymbol{A}_k^\top \boldsymbol{A}_k)_{i,j}.$$

In the same fashion we get

$$(\boldsymbol{B}\boldsymbol{B}^\top \circ \boldsymbol{Q})_{i,j} = \sum_{k=1}^n (\boldsymbol{B}_k \boldsymbol{B}_k^\top)_{i,j},$$

from which we deduce that

$$\operatorname{Tr}\left[(\boldsymbol{A}^\top\boldsymbol{A}\circ\boldsymbol{Q})\cdot(\boldsymbol{B}\boldsymbol{B}^\top\circ\boldsymbol{Q})\right]=\sum_{i,j=1}^{n}\operatorname{Tr}\left[\boldsymbol{A}_i^\top\boldsymbol{A}_i\boldsymbol{B}_j\boldsymbol{B}_j^\top\right].$$

Therefore, the proof of (39) can be obtained by proving that, for *any* matrices $\boldsymbol{A}_1,\dots,\boldsymbol{A}_n\in\mathbb{R}^{d\times n}$ and $\boldsymbol{B}_1,\dots,\boldsymbol{B}_n\in\mathbb{R}^{n\times d}$,

$$\sum_{i,j=1}^{n}\operatorname{Tr}\left[\boldsymbol{A}_i^\top\boldsymbol{A}_i\boldsymbol{B}_j\boldsymbol{B}_j^\top\right]-\frac{2c_1}{f(1)}\cdot\sum_{i=1}^{n}\operatorname{Tr}\left[\boldsymbol{A}_i\boldsymbol{B}_i\right]+\frac{c_1^2}{(f(1))^2}\operatorname{Tr}\left[\boldsymbol{I}\right]\geq 0. \tag{41}$$

To show the last claim, let us define the following matrices

$$\boldsymbol{X}=\sum_{i=1}^{n}\boldsymbol{A}_i^\top\boldsymbol{A}_i,\quad \boldsymbol{Y}=\sum_{i=1}^{n}\boldsymbol{B}_i\boldsymbol{B}_i^\top,\quad \boldsymbol{Z}=\sum_{i=1}^{n}\boldsymbol{B}_i\boldsymbol{A}_i,$$

which allows us to rewrite the statement of (41) as

$$\operatorname{Tr}\left[\boldsymbol{X}\boldsymbol{Y}-\frac{2c_1}{f(1)}\cdot\boldsymbol{Z}+\frac{c_1^2}{(f(1))^2}\cdot\boldsymbol{I}\right]\geq 0. \tag{42}$$

Note that $\boldsymbol{X}$ is PSD, hence it has a symmetric square root, which we denote by $\sqrt{\boldsymbol{X}}$. Using the continuity of the quantities involved in the LHS of (42), we can assume without loss of generality that $\boldsymbol{X}$ is invertible. In fact, the following quantities are continuous: trace, matrix product, matrix transpose. In addition, we can always introduce a small perturbation to $\boldsymbol{A}_i$'s which makes $\boldsymbol{X}$ full-rank. Thus, it suffices to show that (42) holds for $\boldsymbol{A}_i$'s such that $\boldsymbol{X}$ is invertible.

In this view, for any matrix $\boldsymbol{T}\in\mathbb{R}^{n\times n}$, we have

$$\begin{aligned}
0\leq\sum_{i=1}^{n}\left\|\frac{c_1}{f(1)}\cdot\boldsymbol{T}\boldsymbol{A}_i^\top-\sqrt{\boldsymbol{X}}\boldsymbol{B}_i\right\|_F^2&=\sum_{i=1}^{n}\operatorname{Tr}\left[\left(\frac{c_1}{f(1)}\cdot\boldsymbol{T}\boldsymbol{A}_i^\top-\sqrt{\boldsymbol{X}}\boldsymbol{B}_i\right)\cdot\left(\frac{c_1}{f(1)}\cdot\boldsymbol{A}_i\boldsymbol{T}^\top-\boldsymbol{B}_i^\top\sqrt{\boldsymbol{X}}\right)\right]\\
&=\sum_{i=1}^{n}\operatorname{Tr}\left[\frac{c_1^2}{(f(1))^2}\cdot\boldsymbol{T}\boldsymbol{A}_i^\top\boldsymbol{A}_i\boldsymbol{T}^\top-\frac{2c_1}{f(1)}\sqrt{\boldsymbol{X}}\boldsymbol{B}_i\boldsymbol{A}_i\boldsymbol{T}^\top+\boldsymbol{X}\boldsymbol{B}_i\boldsymbol{B}_i^\top\right]\\
&=\operatorname{Tr}\left[\frac{c_1^2}{(f(1))^2}\cdot\boldsymbol{T}\boldsymbol{X}\boldsymbol{T}^\top-\frac{2c_1}{f(1)}\sqrt{\boldsymbol{X}}\boldsymbol{Z}\boldsymbol{T}^\top+\boldsymbol{X}\boldsymbol{Y}\right],
\end{aligned} \tag{43}$$

where in the second line we used that $\operatorname{Tr}\left[\boldsymbol{M}\right]=\operatorname{Tr}\left[\boldsymbol{M}^\top\right]$ for any $\boldsymbol{M}$, and $\operatorname{Tr}\left[\boldsymbol{M}\boldsymbol{N}\right]=\operatorname{Tr}\left[\boldsymbol{N}\boldsymbol{M}\right]$ for any $\boldsymbol{M},\boldsymbol{N}$.

As $\boldsymbol{X}$ is invertible, its square root $\sqrt{\boldsymbol{X}}$ is invertible. As $\boldsymbol{X}$ is also PSD, its inverse, i.e., $\boldsymbol{X}^{-1}$, is PSD and, hence, it has a symmetric square root, i.e., $\sqrt{\boldsymbol{X}^{-1}}$. In this view, we get that

$$\sqrt{\boldsymbol{X}^{-1}}=(\sqrt{\boldsymbol{X}})^{-1}.$$

Thus, by picking $\boldsymbol{T}=(\sqrt{\boldsymbol{X}})^{-1}$, we obtain

$$\boldsymbol{T}^\top\boldsymbol{T}=\boldsymbol{T}^2=\boldsymbol{X}^{-1},\quad \boldsymbol{T}^\top\sqrt{\boldsymbol{X}}=\boldsymbol{T}\sqrt{\boldsymbol{X}}=\boldsymbol{I}.$$

Using these observations, we deduce that the RHS of (43) is equal to the LHS of (42), which concludes the proof. □

C.1.2. Matrices in $\mathcal{H}_{n,d}$ Are the Only Minimizers

Lemma C.2. *Let $\boldsymbol{A}\in\mathbb{R}^{d\times n}$ and $\boldsymbol{B}^\top=[\boldsymbol{b}_1,\dots,\boldsymbol{b}_n]\in\mathbb{R}^{d\times n}$, with $\|\boldsymbol{b}_i\|_2=1$ for $i\in[n]$. Let c_1 and $f(\cdot)$ be defined as per Lemma 4.1. Then, we have that the set of minimizers of*

$$\operatorname{Tr}\left[\boldsymbol{A}^\top\boldsymbol{A}\cdot f(\boldsymbol{B}\boldsymbol{B}^\top)\right]-2c_1\cdot\operatorname{Tr}\left[\boldsymbol{B}\boldsymbol{A}\right] \tag{44}$$

coincides with the set $\mathcal{H}_{n,d}$ of weight-tied orthogonal matrices .

Proof of Lemma C.2. A direct computation immediately shows that the lower bound (37) is achieved for all $\ell \in \mathbb{N}$ by matrices $(\boldsymbol{A}, \boldsymbol{B})$ that belong to the set $\mathcal{H}_{d,n}$. Define the sets of minimizers of (37) as follows

$$\mathcal{M}_\ell := \operatorname*{arg\,min}_{\boldsymbol{A},\boldsymbol{B}:\|\boldsymbol{b}_i\|_2=1} \mathcal{L}_\ell(\boldsymbol{A},\boldsymbol{B}) = \left\{ (\boldsymbol{A}_{\boldsymbol{B}}, \boldsymbol{B}) : \boldsymbol{A}_{\boldsymbol{B}} \in \operatorname*{arg\,min}_{\boldsymbol{A}} \mathcal{L}_\ell(\boldsymbol{A},\boldsymbol{B}),\ \boldsymbol{B} \in \operatorname*{arg\,min}_{\boldsymbol{B}:\|\boldsymbol{b}_i\|_2=1} \mathcal{L}_\ell(\boldsymbol{A}_{\boldsymbol{B}},\boldsymbol{B}) \right\}.$$

We will now show that

$$\bigcap_{l=0}^{\infty} \mathcal{M}_\ell = \mathcal{H}_{n,d}. \tag{45}$$

As the Taylor coefficients of $f(\cdot)$ are non-negative, (45) readily gives that the set of minimizers of (44) coincides with $\mathcal{H}_{n,d}$. Futher, recall that $c_1 \neq 0$ and $\sum_{\ell=1}^{\infty}(c_{2\ell+1})^2 \neq 0$ and, hence, (45) is the union of the linear term ($\ell = 0$) and at least one non-linear ($\ell > 0$) term.

We first prove that it is enough to consider the case $r = 1$. Thus, assume that the result holds for $n = d$ and consider now $n < d$. We have that, for any orthogonal matrix $\boldsymbol{O} \in \mathbb{R}^{d\times d}$,

$$\begin{aligned} \mathbb{E}_{\boldsymbol{x}} \|\boldsymbol{x} - \boldsymbol{A}\sigma(\boldsymbol{B}\boldsymbol{x})\|_2^2 &= \mathbb{E}_{\boldsymbol{x}} \|\boldsymbol{O}\boldsymbol{x} - \boldsymbol{A}\sigma(\boldsymbol{B}\boldsymbol{O}\boldsymbol{x})\|_2^2 \\ &= \mathbb{E}_{\boldsymbol{x}} \left\|\boldsymbol{x} - \boldsymbol{O}^\top\boldsymbol{A}\sigma(\boldsymbol{B}\boldsymbol{O}\boldsymbol{x})\right\|_2^2, \end{aligned} \tag{46}$$

where in the first step we have used the rotational invariance of $\boldsymbol{x}$, and in the second step we have multiplied the argument of the norm by the orthogonal matrix $\boldsymbol{O}^\top$. Thus, (46) gives that $(\boldsymbol{A}, \boldsymbol{B}) \in \mathcal{H}_{n,d}$ if and only if $(\boldsymbol{O}^\top\boldsymbol{A}, \boldsymbol{B}\boldsymbol{O}) \in \mathcal{H}_{n,d}$.

Let us write the SVD of $\boldsymbol{B}$ as $\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$, where $\boldsymbol{U} \in \mathbb{R}^{n\times n}, \boldsymbol{V} \in \mathbb{R}^{d\times d}$ are orthogonal matrices and $\boldsymbol{D} \in \mathbb{R}^{n\times d}$ is a (rectangular) diagonal matrix. Thus, by taking $\boldsymbol{O} = \boldsymbol{V}$, one can assume that $\boldsymbol{B}$ has the form $(\boldsymbol{B}_{1:n,1:n}, \boldsymbol{0}_{1:n,1:d-n})$, where $\boldsymbol{B}_{1:n,1:n}$ denotes the left $n \times n$ sub-matrix of $\boldsymbol{B}$ and $\boldsymbol{0}_{1:n,1:d-n}$ denotes a $n \times (d-n)$ matrix of 0's. We also write the decompositions $\boldsymbol{A} = ((\boldsymbol{A}_{1:n,1:n})^\top, (\boldsymbol{A}_{n+1:d,1:n})^\top)^\top$ and $\boldsymbol{x} = (\boldsymbol{x}_{1:n}, \boldsymbol{x}_{n+1:d})$, where $\boldsymbol{A}_{1:n,1:n}$ (resp. $\boldsymbol{A}_{n+1:d,1:n}$) denotes the top $n \times n$ (resp. bottom $(d-n) \times n$) sub-matrix of $\boldsymbol{A}$, and $\boldsymbol{x}_{1:n}$ (resp. $\boldsymbol{x}_{n+1:d}$) denotes the first n (resp. last $d-n$) components of $\boldsymbol{x}$. Hence, the objective (2) can be expressed (up to the constant multiplicative factor d^{-1}) as the sum of

$$\mathcal{R}_1(\boldsymbol{A},\boldsymbol{B}) = \mathbb{E}\left[\|\boldsymbol{x}_{1:n} - \boldsymbol{A}_{1:n,1:n}\sigma(\boldsymbol{B}_{1:n,1:n}\boldsymbol{x}_{1:n})\|^2\right]$$

and

$$\mathcal{R}_2(\boldsymbol{A},\boldsymbol{B}) = \mathbb{E}\left[\|\boldsymbol{x}_{n+1:d} - \boldsymbol{A}_{n+1:d,1:n}\sigma(\boldsymbol{B}_{1:n,1:n}\boldsymbol{x}_{1:n})\|^2\right].$$

As $\boldsymbol{x}_{n+1:d}$ has zero mean and it is independent from $\boldsymbol{x}_{1:n}$, we have that

$$\mathcal{R}_2(\boldsymbol{A},\boldsymbol{B}) = d - n + \mathbb{E}\left[\|\boldsymbol{A}_{n+1:d,1:n}\sigma(\boldsymbol{B}_{1:n,1:n}\boldsymbol{x}_{1:n})\|^2\right],$$

which is minimized by setting $\boldsymbol{A}_{n+1:d,1:n}$ to $\boldsymbol{0}$. Note that $\mathcal{R}_1$ depends only on $\boldsymbol{A}_{1:n,1:n}, \boldsymbol{B}_{1:n,1:n}$ (and not on $\boldsymbol{A}_{n+1:d,1:n}$), hence its minimizers are $(\boldsymbol{A}_{1:n,1:n}, \boldsymbol{B}_{1:n,1:n}) \in \mathcal{H}_{n,n}$ by our assumption on the $r = 1$ case. As a result, by using that $(\boldsymbol{A}, \boldsymbol{B}) \in \mathcal{H}_{d,n}$ if and only if $(\boldsymbol{O}^\top\boldsymbol{A}, \boldsymbol{B}\boldsymbol{O}) \in \mathcal{H}_{d,n}$, we conclude that all the minimizers of the desired objective have the form $\boldsymbol{O}((\boldsymbol{A}_{1:n,1:n})^\top, (\boldsymbol{0}_{1:n-d,1:n})^\top)^\top$ and $(\boldsymbol{B}_{1:n,1:n}, \boldsymbol{0}_{1:n,1:d-n})\boldsymbol{O}^\top$, i.e., they form the set $\mathcal{H}_{n,d}$ defined in (7).

It remains to prove the result for $r = 1$. First, consider $\ell = 0$. In this case, we have

$$\begin{aligned} \mathcal{L}_0(\boldsymbol{A},\boldsymbol{B}) &= \operatorname{Tr}\left[\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{B}\boldsymbol{B}^\top\right] - \frac{2c_1}{f(1)} \cdot \operatorname{Tr}\left[\boldsymbol{B}\boldsymbol{A}\right] \\ &= \operatorname{Tr}\left[\boldsymbol{B}^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{B}\right] - \frac{2c_1}{f(1)} \cdot \operatorname{Tr}\left[\boldsymbol{A}\boldsymbol{B}\right] \\ &= \|\boldsymbol{A}\boldsymbol{B}\|_F^2 - \frac{2c_1}{f(1)} \cdot \operatorname{Tr}\left[\boldsymbol{A}\boldsymbol{B}\right], \end{aligned} \tag{47}$$

where we have used that the trace is invariant under cyclic permutation. Notice that the minimizer of (47) is clearly $\boldsymbol{A}\boldsymbol{B} = \frac{c_1}{f(1)}\boldsymbol{I}_d$.

Consider some $\ell \geq 1$. As $\boldsymbol{A}\boldsymbol{B} = \frac{c_1}{f(1)}\boldsymbol{I}_d$ and $\boldsymbol{A}, \boldsymbol{B}$ are square matrices, $\boldsymbol{B}$ is invertible and $\boldsymbol{A}^\top \boldsymbol{A} = \frac{c_1^2}{(f(1))^2} \cdot (\boldsymbol{B}\boldsymbol{B}^\top)^{-1}$. Thus,

$$\begin{aligned}\mathcal{L}_\ell(\boldsymbol{A}, \boldsymbol{B}) &= \operatorname{Tr}\left[\boldsymbol{A}^\top \boldsymbol{A}(\boldsymbol{B}\boldsymbol{B}^\top)^{\circ(2\ell+1)}\right] - \frac{2c_1}{f(1)} \cdot \operatorname{Tr}\left[\boldsymbol{B}\boldsymbol{A}\right] \\ &= \frac{c_1^2}{(f(1))^2} \cdot \operatorname{Tr}\left[(\boldsymbol{B}\boldsymbol{B}^\top)^{-1}(\boldsymbol{B}\boldsymbol{B}^\top)^{\circ(2\ell+1)}\right] - \frac{2c_1^2}{(f(1))^2} \cdot n.\end{aligned} \tag{48}$$

Let $\boldsymbol{P} = \boldsymbol{B}\boldsymbol{B}^\top$. Note that $\boldsymbol{P}$ is symmetric and, hence, also its inverse is symmetric. Then, by using (38), we have that

$$\operatorname{Tr}\left[\boldsymbol{P}^{-1}\boldsymbol{P}^{\circ(2\ell+1)}\right] = \operatorname{Tr}\left[(\boldsymbol{P}^{-1} \circ \boldsymbol{P})\boldsymbol{P}^{\circ 2\ell}\right]. \tag{49}$$

An application of Theorem 5 in (Visick, 2000) gives that

$$\boldsymbol{P} \circ \boldsymbol{P}^{-1} \succeq \boldsymbol{I}, \tag{50}$$

where $\succeq$ denotes majorization in the PSD sense. We now show that $\boldsymbol{P} \circ \boldsymbol{P}^{-1} = \boldsymbol{I}$. To do so, suppose by contradiction that

$$\boldsymbol{P} \circ \boldsymbol{P}^{-1} = \boldsymbol{I} + \boldsymbol{R},$$

for some $\boldsymbol{R} \succeq \boldsymbol{0}$ such that $\boldsymbol{R} \neq \boldsymbol{0}$. Hence,

$$\operatorname{Tr}\left[(\boldsymbol{P}^{-1} \circ \boldsymbol{P})\boldsymbol{P}^{\circ 2\ell}\right] = \operatorname{Tr}\left[\boldsymbol{P}^{\circ 2\ell}\right] + \operatorname{Tr}\left[\boldsymbol{R}\boldsymbol{P}^{\circ 2\ell}\right] = n + \operatorname{Tr}\left[\boldsymbol{R}\boldsymbol{P}^{\circ 2\ell}\right], \tag{51}$$

where in the last equality we use that $\boldsymbol{P}$ (and, consequently, $\boldsymbol{P}^{\circ 2\ell}$) has unit diagonal. By the Schur product theorem, $\boldsymbol{P}^{\circ 2\ell} \succ \boldsymbol{0}$ and, hence, it admits a square root. Thus, we get

$$\operatorname{Tr}\left[\boldsymbol{R}\boldsymbol{P}^{\circ 2\ell}\right] = \operatorname{Tr}\left[\sqrt{\boldsymbol{P}^{\circ 2\ell}} \cdot \boldsymbol{R} \cdot \sqrt{\boldsymbol{P}^{\circ 2\ell}}\right].$$

It is easy to see that the matrix $\sqrt{\boldsymbol{P}^{\circ 2\ell}} \cdot \boldsymbol{R} \cdot \sqrt{\boldsymbol{P}^{\circ 2\ell}}$ is PSD and, thus,

$$\operatorname{Tr}\left[\sqrt{\boldsymbol{P}^{\circ 2\ell}} \cdot \boldsymbol{R} \cdot \sqrt{\boldsymbol{P}^{\circ 2\ell}}\right] \geq 0,$$

where the inequality is strict if and only if the corresponding matrix has only zero eigenvalues. However, for any non-zero $\boldsymbol{v} \in \mathbb{R}^n$, we have that

$$\boldsymbol{u}_{\boldsymbol{v}} := \sqrt{\boldsymbol{P}^{\circ 2\ell}} \cdot \boldsymbol{v} \neq 0,$$

since $\sqrt{\boldsymbol{P}^{\circ 2\ell}}$ is strictly positive definite (as $\boldsymbol{P}^{\circ 2\ell} \succ \boldsymbol{0}$) and, thus, it does not have 0 eigenvalues. Hence, if

$$\boldsymbol{v}^\top \cdot \sqrt{\boldsymbol{P}^{\circ 2\ell}} \cdot \boldsymbol{R} \cdot \sqrt{\boldsymbol{P}^{\circ 2\ell}} \cdot \boldsymbol{v} = \boldsymbol{u}_{\boldsymbol{v}}^\top \boldsymbol{R} \boldsymbol{u}_{\boldsymbol{v}} = 0,$$

then $\boldsymbol{u}_{\boldsymbol{v}} \neq \boldsymbol{0}$ is an eigenvector of $\boldsymbol{R}$ corresponding to a zero eigenvalue. In this view, if $\sqrt{\boldsymbol{P}^{\circ 2\ell}} \cdot \boldsymbol{R} \cdot \sqrt{\boldsymbol{P}^{\circ 2\ell}}$ has all zero eigenvalues, then all eigenvalues of $\boldsymbol{R}$ are zero. As $\boldsymbol{R}$ cannot be the zero matrix, by using (51), we conclude that

$$\operatorname{Tr}\left[(\boldsymbol{P}^{-1} \circ \boldsymbol{P})\boldsymbol{P}^{\circ 2\ell}\right] > n. \tag{52}$$

By combining (48), (49) and (52), we have that $\mathcal{L}_\ell(\boldsymbol{A}, \boldsymbol{B}) > -c_1^2 n/(f(1))^2$, which contradicts with the fact that $(\boldsymbol{A}, \boldsymbol{B})$ is a minimizer (since any $(\boldsymbol{A}', \boldsymbol{B}') \in \mathcal{H}_{n,d}$ achieves the value of $-c_1^2 n/(f(1))^2$). Therefore, we conclude that $\boldsymbol{P} \circ \boldsymbol{P}^{-1} = \boldsymbol{I}$.

At this point, we show that $\boldsymbol{P} \circ \boldsymbol{P}^{-1} = \boldsymbol{I}$ implies that $\boldsymbol{P} = \boldsymbol{I}$. Note that $\boldsymbol{P}$ is a Gram matrix, and let its basis be $\{\boldsymbol{b}_1, \cdots, \boldsymbol{b}_n\}$. Define

$$\boldsymbol{b}_i' = \boldsymbol{b}_i - \tilde{\boldsymbol{b}}_i,$$

where $\tilde{\boldsymbol{b}}_i$ is orthogonal projection of $\boldsymbol{b}_i$ onto the space spanned by $\{\boldsymbol{b}_j\}_{j \neq i}^n$. From a well-known result (see, for instance, Theorem 2.1 in (del Pino & Galaz, 1995)) we have that

$$\boldsymbol{P}_{ii}^{-1} = \frac{1}{\|\boldsymbol{b}_i'\|_2^2}. \tag{53}$$

Hence, we obtain that

$$\|\boldsymbol{b}_i'\|_2 \le \|\boldsymbol{b}_i\|_2 = 1, \tag{54}$$

where the inequality is sharp only if $\boldsymbol{b}_i$ is orthogonal to all $\{\boldsymbol{b}_j\}_{j\neq i}^n$. Then, from (53), we deduce

$$n = \mathrm{Tr}\left[\boldsymbol{I}\right] = \mathrm{Tr}\left[\boldsymbol{P}\circ\boldsymbol{P}^{-1}\right] = \sum_{i=1}^n \|\boldsymbol{b}_i\|_2^2 \cdot \frac{1}{\|\boldsymbol{b}_i'\|_2^2} = \sum_{i=1}^n \frac{1}{\|\boldsymbol{b}_i'\|_2^2}. \tag{55}$$

By combining (54) and (55), we conclude that $\{\boldsymbol{b}_i\}_{i\in[n]}$ form an orthonormal basis, and, hence, $\boldsymbol{P}=\boldsymbol{I}$. This means that (45) holds for $r=1$ since

$$(44) = \sum_{\ell=1}^{\infty}(c_{2\ell+1})^2\cdot\mathcal{L}_\ell(\boldsymbol{A},\boldsymbol{B}),$$

which concludes the proof. □

Proof of Theorem 4.2. It follows by combining the results of Lemma C.1 and C.2. □

C.2. Case $r>1$

C.2.1. Lower bound on $\widetilde{R}(r)$

Proof of Proposition 4.3. An application of Theorem A in (Khare, 2021) gives that

$$\mathrm{Tr}\left[\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{B}\boldsymbol{B}^\top\right] = \langle\mathbf{1},(\boldsymbol{A}^\top\boldsymbol{A}\circ\boldsymbol{B}\boldsymbol{B}^\top)\mathbf{1}\rangle \ge \frac{1}{d}\langle\mathbf{1},(\mathrm{Diag}(\boldsymbol{B}\boldsymbol{A})\mathrm{Diag}(\boldsymbol{B}\boldsymbol{A})^\top)\mathbf{1}\rangle = \frac{1}{d}\left(\mathrm{Tr}\left[\boldsymbol{B}\boldsymbol{A}\right]\right)^2,$$

where $\mathrm{Diag}(\boldsymbol{B}\boldsymbol{A})\in\mathbb{R}^n$ stands for the vector with entries corresponding to the diagonal of the matrix $\boldsymbol{B}\boldsymbol{A}$. Hence, we have

$$\mathrm{Tr}\left[\boldsymbol{A}^\top\boldsymbol{A}\cdot f(\boldsymbol{B}\boldsymbol{B}^\top)\right] - 2c_1\cdot\mathrm{Tr}\left[\boldsymbol{B}\boldsymbol{A}\right] \ge \frac{c_1^2}{d}\left(\mathrm{Tr}\left[\boldsymbol{B}\boldsymbol{A}\right]\right)^2 + \sum_{\ell=1}^{\infty}(c_{2\ell+1})^2\cdot\mathrm{Tr}\left[\boldsymbol{A}^\top\boldsymbol{A}\cdot(\boldsymbol{B}\boldsymbol{B}^\top)^{\circ 2\ell+1}\right] - 2c_1\cdot\mathrm{Tr}\left[\boldsymbol{B}\boldsymbol{A}\right]. \tag{56}$$

Define $\alpha := f(1)-c_1^2$. Then, for any $\beta\in[0,1]$, we can rewrite the RHS of (56) as

$$\left[\frac{c_1^2}{d}\left(\mathrm{Tr}\left[\boldsymbol{B}\boldsymbol{A}\right]\right)^2 - 2(1-\beta)c_1\cdot\mathrm{Tr}\left[\boldsymbol{B}\boldsymbol{A}\right]\right] + \sum_{\ell=1}^{\infty}(c_{2\ell+1})^2\cdot\left(\mathrm{Tr}\left[\boldsymbol{A}^\top\boldsymbol{A}\cdot(\boldsymbol{B}\boldsymbol{B}^\top)^{\circ 2\ell+1}\right] - \frac{2\beta c_1}{\alpha}\cdot\mathrm{Tr}\left[\boldsymbol{B}\boldsymbol{A}\right]\right). \tag{57}$$

The first term in (57) is a quadratic polynomial in $\mathrm{Tr}\left[\boldsymbol{B}\boldsymbol{A}\right]$. Hence, we have that

$$\left[\frac{c_1^2}{d}\left(\mathrm{Tr}\left[\boldsymbol{B}\boldsymbol{A}\right]\right)^2 - 2(1-\beta)c_1\cdot\mathrm{Tr}\left[\boldsymbol{B}\boldsymbol{A}\right]\right] \ge -d(1-\beta)^2. \tag{58}$$

Define $\boldsymbol{B}_e := [\boldsymbol{B},\mathbf{0}_{1:n,1:n-d}]$ and $\boldsymbol{A}_e^\top := [\boldsymbol{A}^\top,\mathbf{0}_{1:n,1:n-d}]$. One can readily verify that the traces in the second term of (57) remain unchanged if we replace $\boldsymbol{A}$ and $\boldsymbol{B}$ with $\boldsymbol{A}_e$ and $\boldsymbol{B}_e$, respectively. Note that $\boldsymbol{A}_e,\boldsymbol{B}_e$ are square matrices, hence we can apply Lemma C.1 (which readily generalizes to a different scaling in front of the second trace) to get

$$\sum_{\ell=1}^{\infty}(c_{2\ell+1})^2\cdot\left(\mathrm{Tr}\left[\boldsymbol{A}^\top\boldsymbol{A}\cdot(\boldsymbol{B}\boldsymbol{B}^\top)^{\circ 2\ell+1}\right] - \frac{2\beta c_1}{\alpha}\cdot\mathrm{Tr}\left[\boldsymbol{B}\boldsymbol{A}\right]\right) \ge -\sum_{\ell=1}^{\infty}(c_{2\ell+1})^2\cdot\frac{\beta^2c_1^2}{\alpha^2}n = -\frac{\beta^2c_1^2}{\alpha}n. \tag{59}$$

By combining (56), (57), (58) and (59), we obtain that

$$\frac{1}{d}\left(\mathrm{Tr}\left[\boldsymbol{A}^\top\boldsymbol{A}\cdot f(\boldsymbol{B}\boldsymbol{B}^\top)\right] - 2\cdot\mathrm{Tr}\left[\boldsymbol{A}\boldsymbol{B}\right]\right) + 1 \ge 1-(1-\beta)^2 - \frac{\beta^2c_1^2}{\alpha}r. \tag{60}$$

By taking $\beta=\alpha/(c_1^2r+\alpha)$ and re-arranging the RHS of (60), the desired result readily follows. □

C.2.2. Asymptotic Achievability of the Lower Bound

Lemma C.3. *Let $\boldsymbol{A}, \boldsymbol{B}$ be defined as in (12). Then, for any $\epsilon > 0$, we have that, with probability at least $1 - c/d^2$,*

$$\left|\left(\mathrm{Tr}\left[\boldsymbol{A}^\top \boldsymbol{A} f(\boldsymbol{B}\boldsymbol{B}^\top)\right] - 2c_1 \mathrm{Tr}\left[\boldsymbol{A}\boldsymbol{B}\right]\right) - \left(\beta^2 c_1^2 rn + \beta^2 \alpha n - 2c_1 \beta n\right)\right| \le C n^{\frac{1}{2}+\epsilon}.$$

Thus, choosing $\beta = \frac{c_1}{c_1^2 r + \alpha}$ the loss approaches $1 - \frac{r}{r + \frac{\alpha}{c_1^2}}$, i.e., with the same probability,

$$\left|\left(1 + \frac{1}{d}\left(\mathrm{Tr}\left[\boldsymbol{A}^\top \boldsymbol{A} f(\boldsymbol{B}\boldsymbol{B}^\top)\right] - 2c_1 \mathrm{Tr}\left[\boldsymbol{A}\boldsymbol{B}\right]\right)\right) - \left(1 - \frac{r}{r + \frac{\alpha}{c_1^2}}\right)\right| \le C d^{-\frac{1}{2}+\epsilon}.$$

Here, the constants c, C depend only on r and ϵ.

We start by proving the following.

Lemma C.4. *Let $\hat{\boldsymbol{B}}, \boldsymbol{B}$ be defined as in (12). Then, for any $\epsilon > 0$, we have that, with probability at least $1 - c/d^2$,*

$$\max_{i,j} \left| \frac{(\boldsymbol{B}\boldsymbol{B}^\top)_{i,j}}{(\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top)_{i,j}} - 1 \right| \le C n^{-\frac{1}{2}+\epsilon}.$$

Here, the constants c, C depend only on r and ϵ.

Proof. If $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ is sampled uniformly from $\mathbb{SO}(n)$, then it follows from rotational invariance that any fixed row or column is uniformly distributed on the n-dimensional sphere $\mathbb{S}^{n-1}$. Thus, any fixed row of $\boldsymbol{U}$ is distributed as $\boldsymbol{g}/\left\|\boldsymbol{g}\right\|_2$, where $\boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{I}/n)$. Now, it follows from the concentration of $\left\|\boldsymbol{g}\right\|_2$ (see e.g. Theorem 3.1.1 in (Vershynin, 2018)) that $\left\|\left\|\boldsymbol{g}\right\|_2 - 1\right\|_{\psi_2} \le C n^{-\frac{1}{2}}$, where $\|\cdot\|_{\psi_2}$ denotes the sub-Gaussian norm. Denote by $\boldsymbol{g}_d \in \mathbb{R}^d$ the first d components of $\boldsymbol{g}_d$. Then, by the same reasoning, it holds that $\left\|\sqrt{r}\left\|\boldsymbol{g}_d\right\|_2 - 1\right\|_{\psi_2} \le c d^{-\frac{1}{2}}$. Looking at the definition of $\hat{\boldsymbol{B}}$, we have that, for any fixed i, the distribution of its rows is given by $\hat{\boldsymbol{b}}_i \sim \sqrt{r}\boldsymbol{g}_d/\left\|\boldsymbol{g}\right\|_2$. Furthermore, for any pair of indices i, j, we have that

$$\frac{(\boldsymbol{B}\boldsymbol{B}^\top)_{i,j}}{(\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top)_{i,j}} = \frac{1}{\|\hat{\boldsymbol{b}}_i\|_2 \cdot \|\hat{\boldsymbol{b}}_j\|_2}.$$

Hence,

$$\mathbb{P}\left(\left|\frac{(\boldsymbol{B}\boldsymbol{B}^\top)_{i,j}}{(\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top)_{i,j}} - 1\right| \le n^{-\frac{1}{2}+\epsilon}\right) = \mathbb{P}\left(\left|\frac{1}{\|\hat{\boldsymbol{b}}_i\|_2 \cdot \|\hat{\boldsymbol{b}}_j\|_2} - 1\right| \le n^{-\frac{1}{2}+\epsilon}\right) \le C\exp\left(-\frac{d^\epsilon}{C}\right).$$

Now a simple union bound over all rows gives us

$$\mathbb{P}\left(\max_{i,j}\left|\frac{1}{\|\hat{\boldsymbol{b}}_i\|_2 \cdot \|\hat{\boldsymbol{b}}_j\|_2} - 1\right| \le n^{-\frac{1}{2}+\epsilon}\right) \le Cn\exp\left(-\frac{d^\epsilon}{C}\right) \le \frac{C}{d^2},$$

which implies the desired result. □

Next, we bound the traces of the terms $\boldsymbol{B}\boldsymbol{B}^\top(\boldsymbol{B}\boldsymbol{B}^\top)^{\circ(2\ell+1)}$. We start with the case $\ell = 0$.

Lemma C.5. *Let $\boldsymbol{B}$ be defined as in (12). Then, for any $\epsilon > 0$, with probability at least $1 - c/d^2$,*

$$\left|\mathrm{Tr}\left[\boldsymbol{B}\boldsymbol{B}^\top(\boldsymbol{B}\boldsymbol{B}^\top)\right] - rn\right| \le C d^{\frac{1}{2}+\epsilon}.$$

Here, the constants c, C depend only on r and ϵ.

Proof. Note that

$$\mathrm{Tr}\left[\boldsymbol{B}\boldsymbol{B}^\top(\boldsymbol{B}\boldsymbol{B}^\top)\right] = \sum_{i,j}\left((\boldsymbol{B}\boldsymbol{B}^\top)_{i,j}\right)^2 = \sum_{i,j}\left(\frac{\left((\boldsymbol{B}\boldsymbol{B}^\top)_{i,j}\right)^2}{\left((\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top)_{i,j}\right)^2} - 1\right)\left((\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top)_{i,j}\right)^2 + \mathrm{Tr}\left[\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top(\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top)\right].$$

Thus, an application of Lemma C.4 gives that, with probability at least $1 - c/d^2$,

$$\left|\mathrm{Tr}\left[\boldsymbol{B}\boldsymbol{B}^\top(\boldsymbol{B}\boldsymbol{B}^\top)\right] - \mathrm{Tr}\left[\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top(\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top)\right]\right| \le \mathrm{Tr}\left[\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top(\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top)\right] \cdot Cd^{-\frac{1}{2}+\epsilon}. \tag{61}$$

Since the trace is invariant under cyclic permutation, we readily have that

$$\mathrm{Tr}\left[\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top(\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top)\right] = rn. \tag{62}$$

By combining (61) and (62), the desired result follows. □

Finally, we consider the higher order terms for $\ell \ge 1$.

Lemma C.6. *Let $\boldsymbol{B}$ be defined as in (12). Then, for any $\epsilon > 0$, we have that, with probability at least $1 - c/d^2$,*

$$\sup_{\ell \ge 1}\left|\mathrm{Tr}\left[\boldsymbol{B}\boldsymbol{B}^\top(\boldsymbol{B}\boldsymbol{B}^\top)^{\circ(2\ell+1)}\right] - n\right| \le C\log^2 n.$$

Here, the constants c, C depend only on r and ϵ.

Proof. We first observe that

$$\mathrm{Tr}\left[\boldsymbol{B}\boldsymbol{B}^\top(\boldsymbol{B}\boldsymbol{B}^\top)^{\circ(2\ell+1)}\right] = \sum_{i,j}\left((\boldsymbol{B}\boldsymbol{B}^\top)_{i,j}\right)^{2\ell+2} = n + \sum_{i\neq j}\left((\boldsymbol{B}\boldsymbol{B}^\top)_{i,j}\right)^{2\ell+2}.$$

An application of Lemma C.4 gives that, with probability $1 - c/d^2$,

$$\sup_{\ell\ge 1}\sum_{i\neq j}\left((\boldsymbol{B}\boldsymbol{B}^\top)_{i,j}\right)^{2\ell+2} \le \sup_{\ell\ge 1}\sum_{i\neq j}\left((1+Cd^{-1/2+\epsilon})\cdot(\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top)_{i,j}\right)^{2\ell+2}. \tag{63}$$

Furthermore, by using the first part of Lemma G.2 with $\boldsymbol{A} = \hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top$, we have that, with probability at least $1 - 1/n^2$, the RHS of (63) is lower bounded by

$$\sup_{\ell\ge 1}\sum_{i\neq j}\left((1+Cd^{-1/2+\epsilon})\cdot C\sqrt{\frac{\log n}{n}}\right)^{2\ell+2} \le C\log^2 n,$$

which implies the desired result. □

At this point, we are ready to give the proof of Lemma C.3.

Proof of Lemma C.3. Recall that $\{(c_{2\ell+1})^2\}_{\ell=0}^\infty$ denote the Taylor coefficients of $f(x)$. By using that $\boldsymbol{A} = \beta\boldsymbol{B}^\top$, our objective becomes

$$\begin{aligned}
&\mathrm{Tr}\left[\boldsymbol{A}^\top\boldsymbol{A}f(\boldsymbol{B}\boldsymbol{B}^\top)\right] - 2c_1\mathrm{Tr}\left[\boldsymbol{A}\boldsymbol{B}\right] = \beta^2\mathrm{Tr}\left[\boldsymbol{B}\boldsymbol{B}^\top f(\boldsymbol{B}\boldsymbol{B}^\top)\right] - 2c_1\beta n\\
&\quad= \beta^2\sum_{\ell=0}^\infty(c_{2\ell+1})^2\mathrm{Tr}\left[\boldsymbol{B}\boldsymbol{B}^\top(\boldsymbol{B}\boldsymbol{B}^\top)^{\circ(2\ell+1)}\right] - 2c_1\beta n\\
&\quad= \beta^2c_1^2rn + \beta^2\sum_{\ell=1}^\infty(c_{2\ell+1})^2 n - 2c_1\beta n\\
&\qquad\qquad + \beta^2c_1^2\left(\mathrm{Tr}\left[\boldsymbol{B}\boldsymbol{B}^\top(\boldsymbol{B}\boldsymbol{B}^\top)\right] - rn\right) + \beta^2\sum_{\ell=1}^\infty(c_{2\ell+1})^2\left(\mathrm{Tr}\left[\boldsymbol{B}\boldsymbol{B}^\top(\boldsymbol{B}\boldsymbol{B}^\top)^{\circ(2\ell+1)}\right] - n\right).
\end{aligned}$$

Then, by bounding the last two terms with Lemma C.5 and Lemma C.6, the desired result follows. □

Proof of Proposition 4.4. The proof is a direct application of Lemma C.3. □

D. Global Convergence of Weight-tied Gradient Flow (Theorem 4.5)

We start by giving a formal recap of the weight-tied gradient flow considered in Section 4.2. Under the weight-tying constraint (14), the objective (13) has the following form

$$\begin{aligned}\Psi(\beta, \boldsymbol{B}) &:= \beta^2 \cdot \mathrm{Tr}\left[\boldsymbol{B}^\top \boldsymbol{B} \cdot f(\boldsymbol{B}\boldsymbol{B}^\top)\right] - 2\beta n \\ &= \beta^2 \cdot \sum_{i,j=1}^{n} \langle \boldsymbol{b}_i, \boldsymbol{b}_j\rangle \cdot f\left(\langle \boldsymbol{b}_i, \boldsymbol{b}_j\rangle\right) - 2\beta n,\end{aligned} \tag{64}$$

where $\|\boldsymbol{b}_i\|_2 = 1$ for all i. Note that the optimal β^* can be found exactly, since (64) is a quadratic polynomial in β. In this view, to optimize (64), we perform a gradient flow on $\{\boldsymbol{b}_i\}_{i=1}^n$, which are regarded as vectors on the unit sphere, and pick the optimal β^* at each time t. Formally,

$$\begin{aligned}\beta(t) &= \frac{n}{\sum_{i,j=1}^{n} \langle \boldsymbol{b}_i, \boldsymbol{b}_j\rangle \cdot f\left(\langle \boldsymbol{b}_i, \boldsymbol{b}_j\rangle\right)}, \\ \frac{\partial \boldsymbol{b}_i(t)}{\partial t} &= -\boldsymbol{J}_i(t) \nabla_{\boldsymbol{b}_i} \Psi(\beta(t), \boldsymbol{B}(t)),\end{aligned} \tag{65}$$

where $\boldsymbol{J}_i(t) := \boldsymbol{I} - \boldsymbol{b}_i(t)\boldsymbol{b}_i(t)^\top$ projects the gradient $\nabla_{\boldsymbol{b}_i}\Psi(\beta(t), \boldsymbol{B}(t))$ onto the tangent space at the point $\boldsymbol{b}_i(t)$ (see (69) for the closed form expression). This ensures that $\|\boldsymbol{b}_i(t)\|_2 = 1$ along the gradient flow trajectory. The described procedure can be viewed as Riemannian gradient flow, due to the projection of the gradient $\nabla_{\boldsymbol{b}_i}\Psi(\beta(t), \boldsymbol{B}(t))$ on the tangent space of the unit sphere. We now present the formal counterpart of Theorem 4.5.

Theorem D.1. *Fix $r \le 1$. Let $\boldsymbol{B}(t)$ be obtained via the gradient flow (65) applied to Ψ defined in (64). Let the initialization $\boldsymbol{B}(0)$ have unit-norm rows and $\mathrm{rank}(\boldsymbol{B}(0)) = n$. Then, as $t \to \infty$, $\boldsymbol{B}(t)\boldsymbol{B}(t)^\top$ converges to $\boldsymbol{I}$, which is the unique global optimum of (64). Moreover, define the residual*

$$\phi(t) = \mathrm{Tr}\left[(\boldsymbol{B}(t)\boldsymbol{B}(t)^\top - \boldsymbol{I}) \cdot f(\boldsymbol{B}(t)^\top \boldsymbol{B}(t))\right] \ge 0,$$

which vanishes at the minimizer, and let T be the first time such that $\phi(T) = \delta$. Then,

$$T \le -\mathbb{1}\{\phi(0) > nf(1)\} \cdot f(1) \cdot \log\det(\boldsymbol{B}(0)\boldsymbol{B}(0)^\top) - \mathbb{1}\{\delta \le nf(1)\} \cdot \frac{2f^2(1)}{\delta} \cdot \log\det(\boldsymbol{B}(0)\boldsymbol{B}(0)^\top). \tag{66}$$

In words, if the residual at initialization is bigger than $nf(1)$, then it takes at most constant time to reach the regime in which the convergence is linear in the precision δ. We also note that by choosing the optimal β^*, the function ϕ can be related to the objective (64) by $\Psi(\beta^*, \boldsymbol{B}(t)) = -\frac{n}{f(1)+\frac{\phi(t)}{n}}$. Hence, (66) gives a quantitative convergence in terms of the objective function as well.

We now are ready to present the proof of Theorem D.1. Let $\boldsymbol{B}^\top = [\boldsymbol{b}_1, \cdots, \boldsymbol{b}_n]$. Recall that, under the weight-tying (14), the objective in (13) can be re-written as

$$\beta^2 \cdot \sum_{i,j=1}^{n} \langle \boldsymbol{b}_i, \boldsymbol{b}_j\rangle \cdot f\left(\langle \boldsymbol{b}_i, \boldsymbol{b}_j\rangle\right) - 2\beta n. \tag{67}$$

By the definition in Theorem D.1, the residual $\phi(t)$ is given by

$$\phi(t) := \sum_{i\neq j}^{n} \langle \boldsymbol{b}_i, \boldsymbol{b}_j\rangle \cdot f\left(\langle \boldsymbol{b}_i, \boldsymbol{b}_j\rangle\right). \tag{68}$$

In this view, in accordance with (65), we study the following gradient flow:

$$\begin{cases} \frac{\partial \boldsymbol{b}_k(t)}{\partial t} = -\beta^2(t) \cdot \left[\boldsymbol{J}_k(t) \sum_{i\neq j} \boldsymbol{b}_j(t) \cdot g(\langle \boldsymbol{b}_k(t), \boldsymbol{b}_j(t)\rangle)\right], \\ \beta(t) = \frac{n}{nf(1) + \phi(t)}, \\ \|\boldsymbol{b}_k(0)\|_2 = 1, \end{cases} \tag{69}$$

where $g(x) := x \cdot f'(x) + f(x)$, and we have rescaled the time of the dynamics by a factor 2 to omit the factor 2 in front of $\beta^2(t)$. From here on, we will suppress the time notation when it is clear from the context, for the sake of simplicity. Note that one of the terms is absent in the summation, due to the fact that by definition of the operator $\boldsymbol{J}_k$:

$$\boldsymbol{J}_k \boldsymbol{b}_k = \boldsymbol{0}.$$

In addition, since $\boldsymbol{J}_k$ defines the projection of the gradient on the tangent space at the point $\boldsymbol{b}_k$ of the unit sphere, along the trajectory of the gradient flow (69) we have that $\|\boldsymbol{b}_k\|_2 = 1$.

The gradient flow (69) is well-defined (i.e., its solution exists and it is unique) when its RHS is Lipschitz continuous (see, for instance, (Santambrogio, 2017)). It suffices to check the Lipschitz continuity of $g(\cdot)$. Note that both $xf'(x)$ and $f(x)$ are Lipschitz continuous on any interval $[-1+\delta, 1-\delta]$ for some $\delta > 0$. Hence, the RHS of (69) is Lipschitz continuous, if

$$\max_{i \neq j} |\langle \boldsymbol{b}_i, \boldsymbol{b}_j \rangle| \leq 1 - \delta, \tag{70}$$

where δ is bounded away from 0 uniformly in t.

Recall that, by the assumption of Theorem D.1, we have that $\mathrm{rank}(\boldsymbol{B}(0)\boldsymbol{B}(0)^\top) = n$, hence $\det(\boldsymbol{B}(0)\boldsymbol{B}(0)^\top) \geq \varepsilon_1$ for some $\varepsilon_1 > 0$. Thus, from the result in Lemma D.3, we obtain that

$$\det(\boldsymbol{B}(t)\boldsymbol{B}(t)^\top) \geq \varepsilon_1. \tag{71}$$

Let $0 < \lambda_1 < \lambda_2 < \ldots < \lambda_n$ denote the eigenvalues of $\boldsymbol{B}(t)\boldsymbol{B}(t)^\top$ in increasing order. Then, (71) directly gives that

$$\lambda_1 \prod_{i=2}^{n} \lambda_i \geq \varepsilon_1 > 0.$$

Since $\boldsymbol{B}(t)\boldsymbol{B}(t)^\top$ has unit diagonal, we have that $\sum_{i=1}^n \lambda_i = n$. Hence, the smallest possible value of λ_1 during the gradient flow dynamics can be inferred from

$$\lambda_1 \geq \frac{\varepsilon_1}{\prod_{i=2}^n \lambda_i},$$

by picking the largest possible $\prod_{i=2}^n \lambda_i$ given the constraint $\sum_{i=2}^n \lambda_i \leq n$. This is achieved by taking

$$\lambda_i = \frac{n}{n-1}, \quad \forall i \in \{2, \cdots, n\},$$

which gives

$$\prod_{i=2}^{n} \lambda_i = \left(\frac{n}{n-1}\right)^{n-1} = \left(1 + \frac{1}{n-1}\right)^{n-1} \leq C,$$

where C is a universal constant, since the RHS converges from below to Euler's number as n increases. This proves that λ_1 is bounded away from zero uniformly in t. As a result, we can readily conclude that (70) holds. To see this last claim, consider a vector $\boldsymbol{v}$ which has 1 on position i and $-\mathrm{sign}\langle \boldsymbol{b}_i, \boldsymbol{b}_j \rangle$ on position j. Hence, we have that

$$2\lambda_1 = \lambda_1 \cdot \|\boldsymbol{v}\|_2^2 \leq \boldsymbol{v}^\top (\boldsymbol{B}(t)\boldsymbol{B}(t)^\top)\boldsymbol{v} = 2 - 2 \cdot |\langle \boldsymbol{b}_i, \boldsymbol{b}_j \rangle| \Rightarrow |\langle \boldsymbol{b}_i, \boldsymbol{b}_j \rangle| \leq 1 - \lambda_1.$$

Notice that

$$\phi(t) \leq (n^2 - n) f(1),$$

since $xf(x) \leq f(1)$ for $|x| \leq 1$. Hence, we have that $\beta(t) \geq \frac{1}{nf(1)} > 0$. In this view, along the trajectory of the gradient flow (69), the quantity $\phi(t)$ is *strictly* decreasing until convergence, by the property of gradient flow.

Lemma D.2 (Characterization of stationary points)**.** *Consider the gradient flow* (69)*. Then, the following holds:*

(A) Any orthogonal set of b_i is a stationary point and a global minimizer.

(B) The gradient flow (69) *never escapes any subspace spanned by a set of linearly dependent b_i. However, for each such subspace there exists a direction in which* (67) *can be improved.*

Proof of Lemma D.2. Recall that $\beta(t) > 0$ and $\{\boldsymbol{b}_i\}_{i=1}^n$. Then, the stationary point condition can be expressed as

$$\boldsymbol{J}_k \sum_{j \neq k} \boldsymbol{b}_j \cdot g\left(\langle \boldsymbol{b}_k, \boldsymbol{b}_j \rangle\right) = 0, \quad \forall k \in [n]. \tag{72}$$

Thus, any orthogonal set of vectors is clearly a stationary point by definition of $g(\cdot)$. Moreover, (67) is minimized *iff* $\boldsymbol{B}\boldsymbol{B}^\top = \boldsymbol{I}$ as $xf(x)$ is an even function since $f(\cdot)$ is odd.

Note that the kernel of the operator $\boldsymbol{J}_k$ is spanned by the vector $\boldsymbol{b}_k$. Thus, the condition (72) is equivalent to

$$\sum_{j \neq k} \boldsymbol{b}_j \cdot g\left(\langle \boldsymbol{b}_k, \boldsymbol{b}_j \rangle\right) = \gamma_k \cdot \boldsymbol{b}_k,$$

for some $\gamma_k \in \mathbb{R}$. One can readily verify that $g(x) = 0$ if and only if $x = 0$. Thus, either *(i)* $\boldsymbol{b}_k$ is orthogonal to $\boldsymbol{b}_j$ for all $j \neq k$ and $\gamma_k = 0$, or *(ii)* $\boldsymbol{b}_k$ lies in the span of $\{\boldsymbol{b}_j\}_{j \neq k}$. If condition *(i)* holds for all $k \in [n]$, then $\{\boldsymbol{b}_i\}_{i=1}^n$ form an orthogonal set of vectors and we fall in category (A). If condition *(ii)* holds for some $k \in [n]$, then we fall in category (B).

Now, let us show that, if $\{\boldsymbol{b}_i\}_{i=1}^n$ spans a sub-space of dimension smaller than n, then there is a direction along which the value of (67) can be improved. Since the $\{\boldsymbol{b}_i\}_{i=1}^n$ are linearly dependent, there exists $\boldsymbol{u}$ of unit norm such that

$$\langle \boldsymbol{u}, \boldsymbol{b}_j \rangle = 0, \quad \forall j \in [n]. \tag{73}$$

For some $k \in [n]$, consider the perturbation

$$\hat{\boldsymbol{b}}_k = \frac{1}{\sqrt{1+\lambda^2}} \cdot (\boldsymbol{b}_k + \lambda \cdot \boldsymbol{u}),$$

which has unit norm as $\langle \boldsymbol{b}_k, \boldsymbol{u} \rangle = 0$. Recall that (67) can be expressed as

$$\beta^2 \left(2 \cdot \sum_{j \neq k}^{n} \left\langle \hat{\boldsymbol{b}}_k, \boldsymbol{b}_j \right\rangle f\left(\left\langle \hat{\boldsymbol{b}}_k, \boldsymbol{b}_j \right\rangle \right) + \frac{\pi}{2} + \sum_{i,j \neq k}^{n} \langle \boldsymbol{b}_i, \boldsymbol{b}_j \rangle f\left(\langle \boldsymbol{b}_i, \boldsymbol{b}_j \rangle\right) \right) - 2\beta n. \tag{74}$$

Here, β is chosen to be the minimizer of the quantity (74) having fixed $\{\boldsymbol{b}_j\}_{j \neq k}$ and $\hat{\boldsymbol{b}}_k$. Thus, in order to prove that the population risk gets smaller by replacing $\boldsymbol{b}_k$ with $\hat{\boldsymbol{b}}_k$ for any $\lambda > 0$, it suffices to show that the following quantity

$$\sum_{j \neq k}^{n} \left\langle \hat{\boldsymbol{b}}_k, \boldsymbol{b}_j \right\rangle f\left(\left\langle \hat{\boldsymbol{b}}_k, \boldsymbol{b}_j \right\rangle \right), \tag{75}$$

is decreasing with λ. This last claim follows from the chain of inequalities below:

$$(75) = \frac{1}{\sqrt{1+\lambda^2}} \sum_{j \neq k} \langle \boldsymbol{b}_k, \boldsymbol{b}_j \rangle \cdot f\left(\frac{1}{\sqrt{1+\lambda^2}} \langle \boldsymbol{b}_k, \boldsymbol{b}_j \rangle \right) \tag{76}$$

$$= \frac{1}{\sqrt{1+\lambda^2}} \sum_{j \neq k} \langle \boldsymbol{b}_k, \boldsymbol{b}_j \rangle \cdot \sum_{\ell=0}^{\infty} \left(\frac{c_{2\ell+1}}{c_1} \right)^2 \cdot \left(\frac{1}{\sqrt{1+\lambda^2}} \right)^{2\ell+1} \cdot \langle \boldsymbol{b}_k, \boldsymbol{b}_j \rangle^{2\ell+1} \tag{77}$$

$$\leq \left(\frac{1}{\sqrt{1+\lambda^2}} \right)^2 \sum_{j \neq k} \langle \boldsymbol{b}_k, \boldsymbol{b}_j \rangle \cdot \sum_{\ell=0}^{\infty} \left(\frac{c_{2\ell+1}}{c_1} \right)^2 \cdot \langle \boldsymbol{b}_k, \boldsymbol{b}_j \rangle^{2\ell+1} \tag{78}$$

$$= \frac{1}{1+\lambda^2} \sum_{j \neq k} \langle \boldsymbol{b}_k, \boldsymbol{b}_j \rangle \cdot f\left(\langle \boldsymbol{b}_k, \boldsymbol{b}_j \rangle\right) < \sum_{j \neq k} \langle \boldsymbol{b}_k, \boldsymbol{b}_j \rangle \cdot f\left(\langle \boldsymbol{b}_k, \boldsymbol{b}_j \rangle\right), \tag{79}$$

where in the second line we substitute the Taylor expansion of $f(\cdot)$, the inequality in the third line uses that the coefficients $\{c_{2\ell+1}^2\}_{\ell=0}^{\infty}$ are all non-negative, and the last inequality follows from the fact that $\lambda > 0$.

Finally, we show that the gradient flow (69) does not escape the degenerate sub-space. If $\dim(\mathrm{span}(\{\boldsymbol{b}_i\}_{i=1}^n)) < n$, then there exists $\boldsymbol{u} \in \mathbb{R}^d$ such that (73) holds. By projecting the gradient expression (72) onto $\boldsymbol{u}$, we have

$$\left\langle \boldsymbol{u}, \boldsymbol{J}_k \sum_{j \neq k} \boldsymbol{b}_j \cdot g\left(\langle \boldsymbol{b}_k, \boldsymbol{b}_j \rangle\right) \right\rangle = 0.$$

Hence, for any $k \in [n]$, the directional derivative of $\boldsymbol{b}_k$ in the direction of $\boldsymbol{u}$ is equal to zero, and the gradient flow does not escape the low-rank sub-space, which concludes the proof. □

In next lemma we show that, if at initialization $\{\boldsymbol{b}_i\}_{i=1}^n$ spans a sub-space of dimension n, then it will never get stuck in a low-rank sub-space.

Lemma D.3 (Linearly independent $\{\boldsymbol{b}_i\}_{i=1}^n$ stay linearly independent)**.** *Consider the gradient flow (69) with full rank initialization, i.e.,* $\operatorname{rank}(\boldsymbol{B}(0)\boldsymbol{B}(0)^\top) = n$*. Then, the following holds*

$$\frac{\partial}{\partial t} \log\det(\boldsymbol{B}(t)\boldsymbol{B}(t)^\top) \geq 2\beta(t)^2 \cdot \phi(t) \geq 0,$$

where $\boldsymbol{B}(t)^\top = [\boldsymbol{b}_1(t), \cdots, \boldsymbol{b}_n(t)]$ *and* $\phi(t)$ *is defined in (68). In particular, this implies that* $\{\boldsymbol{b}_i\}_{i=1}^n$ *stay full-rank along the gradient flow trajectory.*

Proof of Lemma D.3. Applying the chain rule and using that the time derivative of $\boldsymbol{B}$ is given by the gradient flow (69) implies that

$$\frac{\partial}{\partial t} \log\det(\boldsymbol{B}\boldsymbol{B}^\top) = \operatorname{Tr}\left[(\boldsymbol{B}\boldsymbol{B}^\top)^{-1} \cdot \left(\frac{\partial \boldsymbol{B}}{\partial t} \cdot \boldsymbol{B}^\top + \boldsymbol{B} \cdot \frac{\partial \boldsymbol{B}^\top}{\partial t}\right)\right],$$

where

$$\frac{\partial \boldsymbol{b}_k}{\partial t} = -\beta(t)^2 \cdot \left(\boldsymbol{J}_k \sum_{j \neq k} \boldsymbol{b}_j \cdot g\left(\langle \boldsymbol{b}_k, \boldsymbol{b}_j\rangle\right)\right).$$

Let us compute the quantity

$$\left\langle \frac{\partial \boldsymbol{b}_k}{\partial t}, \boldsymbol{b}_\ell \right\rangle = \left(\frac{\partial \boldsymbol{B}}{\partial t} \cdot \boldsymbol{B}^\top\right)_{k,\ell}.$$

By definition of $\boldsymbol{J}_k$, we have that

$$\boldsymbol{J}_k \sum_{j \neq k} \boldsymbol{b}_j \cdot g\left(\langle \boldsymbol{b}_k, \boldsymbol{b}_j\rangle\right) = \sum_{j \neq k} \boldsymbol{b}_j \cdot g\left(\langle \boldsymbol{b}_k, \boldsymbol{b}_j\rangle\right) - \sum_{j \neq k} \boldsymbol{b}_k \cdot \langle \boldsymbol{b}_k, \boldsymbol{b}_j\rangle \cdot g\left(\langle \boldsymbol{b}_k, \boldsymbol{b}_j\rangle\right).$$

Note that

$$\left\langle \sum_{j \neq k} \boldsymbol{b}_k \cdot \langle \boldsymbol{b}_k, \boldsymbol{b}_j\rangle \cdot g\left(\langle \boldsymbol{b}_k, \boldsymbol{b}_j\rangle\right), \boldsymbol{b}_\ell \right\rangle = \left[\operatorname{Diag}\left[\mathbf{1}^\top((\boldsymbol{B}\boldsymbol{B}^\top - \boldsymbol{I}) \circ g(\boldsymbol{B}\boldsymbol{B}^\top))\right] \cdot \boldsymbol{B}\boldsymbol{B}^\top\right]_{k,\ell},$$

and that

$$\left\langle \sum_{j \neq k} \boldsymbol{b}_j \cdot g\left(\langle \boldsymbol{b}_k, \boldsymbol{b}_j\rangle\right), \boldsymbol{b}_\ell \right\rangle = \left[g(\boldsymbol{B}\boldsymbol{B}^\top) \cdot \boldsymbol{B}\boldsymbol{B}^\top\right]_{k,\ell} - g(1) \cdot [\boldsymbol{B}\boldsymbol{B}^\top]_{k,\ell}.$$

By combining these last four equations, we conclude that

$$\frac{\partial \boldsymbol{B}}{\partial t} \cdot \boldsymbol{B}^\top = -\beta(t)^2 \left(g(\boldsymbol{B}\boldsymbol{B}^\top) \cdot \boldsymbol{B}\boldsymbol{B}^\top - g(1) \cdot \boldsymbol{B}\boldsymbol{B}^\top - \operatorname{Diag}\left[\mathbf{1}^\top((\boldsymbol{B}\boldsymbol{B}^\top - \boldsymbol{I}) \circ g(\boldsymbol{B}\boldsymbol{B}^\top))\right] \cdot \boldsymbol{B}\boldsymbol{B}^\top\right).$$

Furthermore,

$$\boldsymbol{B} \cdot \frac{\partial \boldsymbol{B}^\top}{\partial t} = \left(\frac{\partial \boldsymbol{B}}{\partial t} \cdot \boldsymbol{B}^\top\right)^\top = -\beta(t)^2 \left(\boldsymbol{B}\boldsymbol{B}^\top \cdot g(\boldsymbol{B}\boldsymbol{B}^\top) - g(1) \cdot \boldsymbol{B}\boldsymbol{B}^\top - \boldsymbol{B}\boldsymbol{B}^\top \cdot \operatorname{Diag}\left[\mathbf{1}^\top((\boldsymbol{B}\boldsymbol{B}^\top - \boldsymbol{I}) \circ g(\boldsymbol{B}\boldsymbol{B}^\top))\right]\right).$$

Hence, by using the cyclic property of the trace, we get that

$$\begin{aligned}\frac{\partial}{\partial t} \log\det(\boldsymbol{B}\boldsymbol{B}^\top) &= 2\beta(t)^2 \cdot \operatorname{Tr}\left[\operatorname{Diag}\left[\mathbf{1}^\top((\boldsymbol{B}\boldsymbol{B}^\top - \boldsymbol{I}) \circ g(\boldsymbol{B}\boldsymbol{B}^\top))\right]\right] - 2\beta(t)^2 \cdot \operatorname{Tr}\left[g(\boldsymbol{B}\boldsymbol{B}^\top) - g(1) \cdot \boldsymbol{I}\right] \\ &= 2\beta(t)^2 \cdot \sum_{i \neq j}^{n} \langle \boldsymbol{b}_i, \boldsymbol{b}_j\rangle \cdot g\left(\langle \boldsymbol{b}_i, \boldsymbol{b}_j\rangle\right) + 0,\end{aligned}$$

Now, note that

$$xg(x) = x^2 f'(x) + xf(x) \geq 0,$$

since $x^2 f'(x)$ and $xf(x)$ are non-negative functions, which concludes the proof. □

The result of Lemma D.3 gives that $\det(\boldsymbol{B}\boldsymbol{B}^\top)$ is non-decreasing. Hence, if $\lambda_{\min}(\boldsymbol{B}\boldsymbol{B}^\top) > \delta > 0$ at initialization, then this quantity will be bounded away from zero during the gradient flow dynamics and the gradient flow will not get stuck in a low-rank solution. Therefore, by Lemma D.2, the gradient flow converges to a global minimum, in which the rows of $\boldsymbol{B}$ are orthogonal vectors with unit norm. The speed at which this happens is characterized by the next lemma.

Lemma D.4 (Rate of convergence)**.** *Consider the gradient flow* (69) *with full rank initialization, i.e.,* $\operatorname{rank}(\boldsymbol{B}(0)\boldsymbol{B}(0)^\top) = n$*. Let* T *be the time at which* $\phi(T)$ *hits the value* $\delta > 0$*. Then, the following holds*

$$T \leq -\det(\boldsymbol{B}(0)\boldsymbol{B}(0)^\top) \cdot \left(f(1) \cdot \mathbb{1}\{\phi(0) > n \cdot f(1)\} + \frac{2f^2(1)}{\delta} \cdot \mathbb{1}\{\delta \leq n \cdot f(1)\} \right). \tag{80}$$

Proof of Lemma D.4. For all t, we have that $\operatorname{Tr}\left[\boldsymbol{B}(t)\boldsymbol{B}(t)^\top\right] = n$, which implies that $\det(\boldsymbol{B}(t)\boldsymbol{B}(t)^\top) \leq 1$ and, as a consequence, that $\log\det(\boldsymbol{B}(t)\boldsymbol{B}(t)^\top) \leq 0$. From Lemma D.3, we know that

$$\frac{\partial}{\partial t} \log\det(\boldsymbol{B}(t)\boldsymbol{B}(t)^\top) \geq 2\beta(t)^2 \cdot \phi(t).$$

In this view, using the exact expression (69) for $\beta(t)$, we get

$$-\log\det(\boldsymbol{B}(0)\boldsymbol{B}(0)^\top) \geq \log\det(\boldsymbol{B}(t)\boldsymbol{B}(t)^\top) - \log\det(\boldsymbol{B}(t)\boldsymbol{B}(t)^\top) \geq \int_0^t \frac{2}{\left(f(1) + \frac{\phi(s)}{n}\right)^2} \cdot \phi(s)\mathrm{d}s. \tag{81}$$

Stage 1. Assume that $\phi(0) > n \cdot f(1)$, and let T_1 be such that $\phi(T_1) = n \cdot f(1)$. Recall that the function $\phi(t)$ is decreasing and note that $x/(1+x)^2$ is decreasing for $x \in [1, +\infty)$. In this view, we can lower bound the integrand in the RHS of (81) for all $t \leq T_1$ by

$$\frac{2 \cdot \phi(0)}{\left(f(1) + \frac{\phi(0)}{n}\right)^2} \geq \frac{2(n-1)}{nf(1)} \geq \frac{1}{f(1)}, \tag{82}$$

where the first inequality follows from the definition (68) of $\phi(\cdot)$, which readily implies that $\phi(0) \leq f(1) \cdot n(n-1)$. Hence, by combining (81) with the lower bound (82), we get

$$T_1 \leq -f(1) \cdot \log\det(\boldsymbol{B}(0)\boldsymbol{B}(0)^\top).$$

Stage 2. Assume that $\phi(0) \leq n \cdot f(1)$. Let $\delta \in (0, n \cdot f(1)]$ be the desired precision which should be reached during the gradient flow, and let T_2 be such that $\phi(T_2) = \delta$. As $\phi(t)$ is decreasing, we have that

$$\frac{1}{\left(f(1) + \frac{\phi(t)}{n}\right)^2} \geq \frac{1}{\left(f(1) + \frac{\phi(0)}{n}\right)^2} \geq \frac{1}{4f^2(1)}, \tag{83}$$

where in the last step we use that $\phi(0) \leq n \cdot f(1)$. Hence, by combining (81) with the lower bound (83), we get

$$-\log\det(\boldsymbol{B}(0)\boldsymbol{B}(0)^\top) \geq \frac{1}{2f^2(1)} \cdot T_2\delta,$$

which implies that

$$T_2 \leq -\frac{2f^2(1) \cdot \log\det(\boldsymbol{B}(0)\boldsymbol{B}(0)^\top)}{\delta}.$$

By combining the results of both stages, the desired result (80) readily follows. □

Proof of Theorem D.1. Theorem D.1 is a compilation of the results presented in current section. □

E. Global Convergence of Projected Gradient Descent (Theorem 4.6)

Recall from statement of Theorem 4.6 that

$$f(x) = x + \sum_{\ell=3}^{\infty} c_\ell^2 x^\ell,$$

with $\sum_{\ell=3}^{\infty} c_\ell^2 < \infty$. We also define $\alpha = \sum_{\ell=3}^{\infty} c_\ell^2$, and we assume that $\alpha > 0$. In fact, if $\alpha = 0$, then the algorithm trivially converges after one step. We denote by C, c uniform positive constants (depending only on r and α) the value of which might change from term to term. To make the notation lighter we will also but the time t as a subscript (for example $\boldsymbol{B}(t)$ becomes $\boldsymbol{B}_t$).

We analyze the following projected gradient descent procedure for minimizing the population risk

$$\sum_{i,j=1}^{n} \langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle \cdot f\left(\left\langle \frac{\boldsymbol{b}_i}{\|\boldsymbol{b}_i\|_2}, \frac{\boldsymbol{b}_j}{\|\boldsymbol{b}_j\|_2} \right\rangle \right) - 2 \sum_{i=1}^{n} \left\langle \boldsymbol{a}_i, \frac{\boldsymbol{b}_i}{\|\boldsymbol{b}_i\|_2} \right\rangle. \tag{84}$$

Given unit-norm initial $\{\boldsymbol{b}_i\}_{i\in[n]}$, at each step we pick the optimal value of $\boldsymbol{A}$ given $\boldsymbol{B}$

$$\boldsymbol{A}_t = \boldsymbol{B}_t^\top \left(f(\boldsymbol{B}_t \boldsymbol{B}_t^\top) \right)^{-1}. \tag{85}$$

Then, we update $\boldsymbol{B}_t$ with a gradient step and a projection on the sphere to keep the unit norm:

$$\boldsymbol{B}_t' := \boldsymbol{B}_t - \eta \nabla_{\boldsymbol{B}_t}, \quad \boldsymbol{B}_{t+1} := \mathrm{proj}(\boldsymbol{B}_t').$$

Here, the operator $\mathrm{proj}(\boldsymbol{M})$ normalizes the rows of $\boldsymbol{M}$ to be of unit norm and each row of $\nabla_{\boldsymbol{B}_t}$ is defined as the corresponding row of the gradient of $\boldsymbol{B}_t$, i.e.,

$$(\nabla_{\boldsymbol{B}_t})_{k,:} = \underbrace{-2\boldsymbol{J}_k \boldsymbol{a}_k + 2 \sum_{j\neq k} \langle \boldsymbol{a}_k, \boldsymbol{a}_j \rangle \boldsymbol{J}_k \boldsymbol{b}_j}_{:=\nabla^1_{\boldsymbol{B}_t} \quad \text{(part 1)}} + \underbrace{\sum_{l=3}^{\infty} \ell c_\ell^2 \sum_{j\neq k} \langle \boldsymbol{a}_k, \boldsymbol{a}_j \rangle \langle \boldsymbol{b}_k, \boldsymbol{b}_j \rangle^{l-1} \boldsymbol{J}_k \boldsymbol{b}_j}_{:=\nabla^2_{\boldsymbol{B}_t} \quad \text{(part 2)}}, \tag{86}$$

where $\boldsymbol{J}_k := \boldsymbol{I} - \boldsymbol{b}_k \boldsymbol{b}_k^\top$ and we have omitted the iteration number t on $\{\boldsymbol{a}_j, \boldsymbol{b}_j\}_{j\in[n]}$ to keep notation light. Note that in (86) the norms $\|\boldsymbol{b}_i\|_2$, $\|\boldsymbol{b}_j\|_2$ no longer appear as the projection step enforces $\|\boldsymbol{b}_i\|_2 = 1$. At each step of the projected gradient descent dynamics, we decompose $\boldsymbol{B}_t \boldsymbol{B}_t^\top$ as follows:

$$\boldsymbol{B}_t \boldsymbol{B}_t^\top = \boldsymbol{I} + \boldsymbol{Z}_t + \boldsymbol{X}_t, \tag{87}$$

where $\boldsymbol{B}_0 \boldsymbol{B}_0^\top = \boldsymbol{U} \boldsymbol{\Lambda}_0 \boldsymbol{U}^\top$, $\boldsymbol{Z}_t = \boldsymbol{U}(\boldsymbol{\Lambda}_t - \boldsymbol{I})\boldsymbol{U}^\top$ and $\boldsymbol{\Lambda}_{t+1} = g(\boldsymbol{\Lambda}_t)$ for some function $g : \mathbb{R}^n \to \mathbb{R}^n$ which defines the spectrum evolution. Here, $\boldsymbol{U}$ is an orthogonal matrix that importantly does not depend on t and $\boldsymbol{\Lambda}_t$ is the diagonal matrix containing the eigenvalues (i.e., $\boldsymbol{U}\boldsymbol{\Lambda}_t\boldsymbol{U}^\top$ is the SVD). We also define $\boldsymbol{X}_t^D := \mathrm{Diag}(\boldsymbol{X}_t)$ and $\boldsymbol{X}_t^O := \boldsymbol{X}_t - \boldsymbol{X}_t^D$.

For now we will make the following assumptions, which will be proved later in the argument. There exist universal constants $C, C_X > 0$ and $\delta \in (0,1)$ (depending only on r) such that, with probability at least $1 - Ce^{-cd}$,

$$\begin{aligned} &\inf_{t\geq 0} \lambda_{\min}(\boldsymbol{Z}_t) \geq -1 + \delta_r, \\ &\sup_{t\geq 0} \|\boldsymbol{Z}_t\|_{op} \leq C, \\ &\sup_{t\geq 0} \|\boldsymbol{X}_t\|_{op} \leq C_X \frac{\mathrm{poly}(\log d)}{\sqrt{d}}, \\ &\|\boldsymbol{\Lambda}_t - \boldsymbol{I}\|_{op} \leq C\, e^{-c\eta t}. \end{aligned} \tag{88}$$

Here, $\mathrm{poly}(\log d)$ is used to denote polynomial powers of $\log d$, i.e., $(\log d)^C$ for some universal constant C. In the assumptions (88), we specifically distinguish the constant C_X in the bound on $\|\boldsymbol{X}_t\|_{op}$ from the others. This important distinction between C and C_X will be apparent later to show that assumptions (88) indeed hold. Note also that, for sufficiently large d, (88) implies that

$$\sup_{t\geq 0} \|\boldsymbol{X}_t\|_{op} \leq 1. \tag{89}$$

We are now ready to give the proof Theorem 4.6. For the convenience of the reader we restate it here.

Theorem E.1. *Consider the projected gradient descent algorithm as described above applied to the objective (13) for any f of the form $f(x) = x + \sum_{\ell=3} c_\ell^2 x^\ell$, where $\sum_{\ell=3} c_\ell^2 < \infty$. Initialize the algorithm with $\boldsymbol{B}_0$ equal to a row-normalized Gaussian, i.e., $(\boldsymbol{B}_0')_{i,j} \sim \mathcal{N}(0, 1/d)$, $(\boldsymbol{B}_0)i, := \mathbf{Proj}_{\mathbb{S}^{d-1}}\left((\boldsymbol{B}_0')_{i,:}\right)$. Let the step size η be $\Theta(1/\sqrt{d})$. Then, for any $r < 1$, we have that at any time $t = T/\eta$, with probability at least $1 - Ce^{-cd}$,*

$$\left\|\boldsymbol{B}_t\boldsymbol{B}_t^\top - \boldsymbol{I}\right\|_{op} \le C(1-c)^T,$$

where $C > 0$ and $c \in (0, 1]$ are universal constants depending only on r and f.

Let $\boldsymbol{E}^t := \boldsymbol{E}(\boldsymbol{X}_t, \boldsymbol{Z}_t) \in \mathbb{R}^{n\times n}$ be a generic matrix whose operator norm is upper bounded by

$$\left\|\boldsymbol{E}^t\right\|_{op} \le C\left(\frac{\mathrm{poly}(\log d)}{\sqrt{d}} \cdot \|\boldsymbol{Z}_t\|_{op}^{1/2} + \|\boldsymbol{X}_t\|_{op}^2 + \|\boldsymbol{X}_t\|_{op}\|\boldsymbol{Z}_t\|_{op}^{1/2}\right). \tag{90}$$

We highlight that the constant in front of the upper-bound on the error term $\boldsymbol{E}^t$ is *independent* of C_X and t.

Lemma E.2 (Bound for the matrix inverse)**.** *Assume that (88) holds. Then, for all sufficiently large n, with probability at least $1 - 1/d^2$, jointly for all $t \ge 0$ and $\ell \ge 3$, the following bounds hold*

$$\|(\boldsymbol{B}_t\boldsymbol{B}_t^\top - \boldsymbol{I})^{\circ\ell}\|_{op} \le \|\boldsymbol{E}^t\|_{op}, \tag{91}$$

$$\|\left(f(\boldsymbol{B}_t\boldsymbol{B}_t^\top)\right)^{-1} - (\alpha\boldsymbol{I} + \boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1}\|_{op} \le \|\boldsymbol{E}^t\|_{op}, \tag{92}$$

where α was defined as $\alpha = \sum_{\ell=3}^\infty c_\ell^2$.

Proof of Lemma E.2. Note that, for any square matrices $\boldsymbol{R}, \boldsymbol{S} \in \mathbb{R}^{n\times n}$,

$$\|\boldsymbol{R} \circ \boldsymbol{S}\|_{op} \le \sqrt{n}\|\boldsymbol{S}\|_{op} \max_{i,j} |\boldsymbol{R}_{i,j}|. \tag{93}$$

Thus, for $\ell \ge 3$,

$$\begin{aligned}
\left\|(\boldsymbol{B}_t\boldsymbol{B}_t^\top - \boldsymbol{I})^{\circ\ell}\right\|_{op} &\le \sqrt{n}\left\|(\boldsymbol{B}_t\boldsymbol{B}_t^\top - \boldsymbol{I})^{\circ(\ell-3)}\right\|_{op} \max_{i,j} |((\boldsymbol{B}_t\boldsymbol{B}_t^\top - \boldsymbol{I})^{\circ 3})_{i,j}| \\
&= \sqrt{n}\left\|(\boldsymbol{B}_t\boldsymbol{B}_t^\top - \boldsymbol{I})^{\circ(\ell-3)}\right\|_{op} \max_{i\neq j} |((\boldsymbol{B}_t\boldsymbol{B}_t^\top - \boldsymbol{I})^{\circ 3})_{i,j}| \\
&= \sqrt{n}\left\|(\boldsymbol{B}_t\boldsymbol{B}_t^\top - \boldsymbol{I})^{\circ(\ell-3)}\right\|_{op} \max_{i\neq j} |((\boldsymbol{Z}_t + \boldsymbol{X}_t)^{\circ 3})_{i,j}|,
\end{aligned} \tag{94}$$

where in the first line we use (93), in the second line we use that $((\boldsymbol{B}_t\boldsymbol{B}_t^\top - \boldsymbol{I})^{\circ 3})_{i,i} = 0$ for $i \in [n]$ and in the third line we use the decomposition (87).

Let us bound the off-diagonal entries of $\boldsymbol{X}_t$ via (88) and the off-diagonal entries of $\boldsymbol{Z}_t$ via Lemma G.2. This gives that, with probability at least $1 - 1/d^2$, jointly for all $t \ge 0$,

$$\max_{i\neq j} |((\boldsymbol{Z}_t + \boldsymbol{X}_t)^{\circ 3})_{i,j}| \le (C + C_X)^3\left(\frac{\mathrm{poly}(\log d)}{d}\right)^{3/2}. \tag{95}$$

We will condition on this event (without explicitly mentioning it every time) for the reminder of the argument. By combining (94) and (95), we have that

$$\begin{aligned}
\left\|(\boldsymbol{B}_t\boldsymbol{B}_t^\top - \boldsymbol{I})^{\circ\ell}\right\|_{op} &\le \sqrt{n}\left[(C + C_X)^3\left(\frac{\mathrm{poly}(\log d)}{d}\right)^{3/2}\right]\left\|(\boldsymbol{B}_t\boldsymbol{B}_t^\top - \boldsymbol{I})^{\circ(\ell-3)}\right\|_{op} \\
&\le \left\|(\boldsymbol{B}_t\boldsymbol{B}_t^\top - \boldsymbol{I})^{\circ(\ell-3)}\right\|_{op}
\end{aligned} \tag{96}$$

where the last inequality holds for all sufficiently large n. Note that, for any square matrices R, S, an application of Theorem 1 in (Visick, 2000) gives that

$$\|\boldsymbol{R} \circ \boldsymbol{S}\|_{op} \le \|\boldsymbol{R}\|_{op}\|\boldsymbol{S}\|_{op}. \tag{97}$$

Hence,

$$\|(\boldsymbol{B}_t\boldsymbol{B}_t^\top - \boldsymbol{I})^{\circ\ell}\|_{op} \le \left\|(\boldsymbol{B}_t\boldsymbol{B}_t^\top - \boldsymbol{I})^{\circ(\ell-3)}\right\|_{op} \|(\boldsymbol{B}_t\boldsymbol{B}_t^\top - \boldsymbol{I})^{\circ 3}\|_{op}. \tag{98}$$

Now, by using again (97) and the assumptions (88), we have that, for $\ell \in [3]$,

$$\|(\boldsymbol{B}_t\boldsymbol{B}_t^\top - \boldsymbol{I})^{\circ\ell}\|_{op} \le C. \tag{99}$$

Thus, by combining (96) and (99), we obtain that $\left\|(\boldsymbol{B}_t\boldsymbol{B}_t^\top - \boldsymbol{I})^{\circ(\ell-3)}\right\|_{op}$ is uniformly bounded in ℓ, which together with (98) gives that

$$\|(\boldsymbol{B}_t\boldsymbol{B}_t^\top - \boldsymbol{I})^{\circ\ell}\|_{op} \le C\|(\boldsymbol{B}_t\boldsymbol{B}_t^\top - \boldsymbol{I})^{\circ 3}\|_{op}. \tag{100}$$

We remark here that C is independent of l and C_X. This means that it suffices to prove the claim (91) for $\ell = 3$.

To do so, define $\boldsymbol{H} := \mathbf{1}\mathbf{1}^\top - \boldsymbol{I}$, hence, since $\boldsymbol{B}_t\boldsymbol{B}_t^\top$ has unit diagonal, we have that

$$\begin{aligned}(\boldsymbol{B}_t\boldsymbol{B}_t^\top - \boldsymbol{I})^{\circ 3} &= (\boldsymbol{B}_t\boldsymbol{B}_t^\top - \boldsymbol{I})^{\circ 3} \circ \boldsymbol{H} = (\boldsymbol{U}(\boldsymbol{\Lambda}_t - \boldsymbol{I})\boldsymbol{U}^\top + \boldsymbol{X}_t^O + \boldsymbol{X}_t^D)^{\circ 3} \circ \boldsymbol{H} = (\boldsymbol{Z}_t \circ \boldsymbol{H} + \boldsymbol{X}_t^O \circ \boldsymbol{H} + \boldsymbol{X}_t^D \circ \boldsymbol{H})^{\circ 3} \\ &= (\boldsymbol{Z}_t \circ \boldsymbol{H} + \boldsymbol{X}_t^O)^{\circ 3} = (\boldsymbol{Z}_t \circ \boldsymbol{H})^{\circ 3} + 3(\boldsymbol{Z}_t \circ \boldsymbol{H})^{\circ 2} \circ \boldsymbol{X}_t^O + 3(\boldsymbol{Z}_t \circ \boldsymbol{H}) \circ (\boldsymbol{X}_t^O)^{\circ 2} + (\boldsymbol{X}_t^O)^{\circ 3}.\end{aligned}$$

Using again (97) and that, by Lemma G.1 for any $\boldsymbol{R} \in \mathbb{R}^{n\times n}$,

$$\|\boldsymbol{R} \circ \boldsymbol{H}\|_{op} = \|\boldsymbol{R} - \mathrm{diag}(\boldsymbol{R})\|_{op} \le C\|\boldsymbol{R}\|_{op},$$

we get

$$\begin{aligned}\|(\boldsymbol{B}_t\boldsymbol{B}_t^\top - \boldsymbol{I})^{\circ 3}\|_{op} &\le C\left(\|(\boldsymbol{Z}_t \circ \boldsymbol{H})^{\circ 3}\|_{op} + \|\boldsymbol{Z}_t\|_{op}^2\|\boldsymbol{X}_t^O\|_{op} + \|\boldsymbol{Z}_t\|_{op}\|\boldsymbol{X}_t^O\|_{op}^2 + \|\boldsymbol{X}_t^O\|_{op}^3\right) \\ &\le C\left(\|(\boldsymbol{Z}_t \circ \boldsymbol{H})^{\circ 3}\|_{op} + \|\boldsymbol{Z}_t\|_{op}^{1/2}\|\boldsymbol{X}_t^O\|_{op} + \|\boldsymbol{X}_t^O\|_{op}^2\right),\end{aligned} \tag{101}$$

where the second step holds since $\left\|\boldsymbol{X}_t^O\right\|_{op} \le 1$ and $\|\boldsymbol{Z}_t\|_{op} \le C$ by (88)-(89). Another application of (93) gives that

$$\begin{aligned}\|(\boldsymbol{Z}_t \circ \boldsymbol{H})^{\circ 3}\|_{op} = \|(\boldsymbol{Z}_t \circ \boldsymbol{H})^{\circ 2} \circ \boldsymbol{Z}_t\|_{op} &\le \sqrt{n} \cdot \max_{i\neq j} |(\boldsymbol{Z}_t)_{i,j}|^2 \cdot \|\boldsymbol{Z}_t\|_{op} \\ &\le C\frac{\log d}{\sqrt{d}} \cdot \|\boldsymbol{Z}_t\|_{op} \le C\frac{\log d}{\sqrt{d}} \cdot \|\boldsymbol{Z}_t\|_{op}^{1/2},\end{aligned} \tag{102}$$

where the second passage follows from Lemma G.2 and the last from $\|\boldsymbol{Z}_t\|_{op} \le C$. By combining (101) and (102), the proof of (91) for $\ell = 3$ is complete.

To prove (92), define the following quantity

$$\boldsymbol{Y} := \sum_{\ell=3}^{\infty} c_\ell^2 (\boldsymbol{B}_t\boldsymbol{B}_t^\top - \boldsymbol{I})^{\circ\ell}.$$

By definition of $f(\cdot)$ we have that

$$f(\boldsymbol{B}_t\boldsymbol{B}_t^\top) = \alpha\boldsymbol{I} + \boldsymbol{B}_t\boldsymbol{B}_t^\top + \boldsymbol{Y},$$

which implies that

$$\begin{aligned}\left(f(\boldsymbol{B}_t\boldsymbol{B}_t^\top)\right)^{-1} &= (\alpha\boldsymbol{I} + \boldsymbol{B}_t\boldsymbol{B}_t^\top + \boldsymbol{Y})^{-1} \\ &= (\boldsymbol{I} + \boldsymbol{Y}(\alpha\boldsymbol{I} + \boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1})^{-1}(\alpha\boldsymbol{I} + \boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1} \\ &= \left(\boldsymbol{I} + \sum_{k=1}^{\infty}(-1)^k(\boldsymbol{Y}(\alpha\boldsymbol{I} + \boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1})^k\right)(\alpha\boldsymbol{I} + \boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1}.\end{aligned} \tag{103}$$

By definition (90), we have that $\left\|\boldsymbol{E}^t\right\|_{op} \le 1/2$ under assumptions (88) for sufficiently large d. Hence, by the result (91) we have just proved, $\|(\boldsymbol{B}_t\boldsymbol{B}_t^\top - \boldsymbol{I})^{\circ\ell}\|_{op} \le 1/2$, which implies that $\sum_{\ell=3}^{\infty} c_\ell^2\|(\boldsymbol{B}_t\boldsymbol{B}_t^\top - \boldsymbol{I})^{\circ\ell}\|_{op} \le \alpha/2$. Thus, we have

$$\|\boldsymbol{Y}(\boldsymbol{B}_t\boldsymbol{B}_t^\top + \alpha\boldsymbol{I})^{-1}\|_{op} \le \|\boldsymbol{Y}\|_{op}\|(\boldsymbol{B}_t\boldsymbol{B}_t^\top + \alpha\boldsymbol{I})^{-1}\|_{op} \le \frac{\alpha}{2} \cdot \frac{1}{\alpha} \le \frac{1}{2}. \tag{104}$$

Therefore, we can conclude that

$$\begin{aligned}\|\big(f(\boldsymbol{B}_t\boldsymbol{B}_t^\top)\big)^{-1} - (\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1}\|_{op} &\le \|(\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1}\|_{op}\cdot\sum_{k=1}^{\infty}\|\boldsymbol{Y}(\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1}\|_{op}^k\\ &\le \frac{1}{\alpha}\cdot\frac{\|\boldsymbol{Y}(\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1}\|_{op}}{1-\|\boldsymbol{Y}(\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1}\|_{op}}\\ &\le \frac{2}{\alpha}\cdot\|\boldsymbol{Y}\|_{op}\|(\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1}\|_{op}\\ &\le \frac{2}{\alpha^2}\cdot\|\boldsymbol{Y}\|_{op},\end{aligned} \tag{105}$$

where the third inequality uses (104). By bounding $\|\boldsymbol{Y}\|_{op}$ via (91), the proof of (92) is complete. □

Lemma E.3 (Bound for the Schur product with $\boldsymbol{A}^\top\boldsymbol{A}$)**.** *Assume that* (88) *holds, and let* $\boldsymbol{A}_t$ *be given by* (85)*. Then, we have that, with probability at least* $1-1/d^2$*, jointly for all* $t\ge 0$ *and* $\ell\ge 2$*,*

$$\left\|\boldsymbol{A}_t^\top\boldsymbol{A}_t\circ(\boldsymbol{B}_t\boldsymbol{B}_t^\top-\boldsymbol{I})^{\circ\ell}\right\|_{op}\le\|\boldsymbol{E}^t\|_{op}. \tag{106}$$

Proof of Lemma E.3. We have that

$$\begin{aligned}\|\boldsymbol{A}_t^\top\boldsymbol{A}_t\circ(\boldsymbol{B}_t\boldsymbol{B}_t^\top-\boldsymbol{I})^{\circ\ell}\|_{op} &\le \|\boldsymbol{A}_t^\top\boldsymbol{A}_t\circ(\boldsymbol{B}_t\boldsymbol{B}_t^\top-\boldsymbol{I})^{\circ 2}\|_{op}\left\|(\boldsymbol{B}_t\boldsymbol{B}_t^\top-\boldsymbol{I})^{\circ(\ell-2)}\right\|_{op}\\ &\le C\|\boldsymbol{A}_t^\top\boldsymbol{A}_t\circ(\boldsymbol{B}_t\boldsymbol{B}_t^\top-\boldsymbol{I})^{\circ 2}\|_{op},\end{aligned} \tag{107}$$

where the first inequality uses (97) and the second inequality uses that $\left\|(\boldsymbol{B}_t\boldsymbol{B}_t^\top-\boldsymbol{I})^{\circ(\ell-2)}\right\|_{op}$ is uniformly bounded in l, which follows from (96) and (99).

Let us now focus on bounding the RHS of (107). An application of Lemma E.2 gives that

$$\big(f(\boldsymbol{B}_t\boldsymbol{B}_t^\top)\big)^{-1}=(\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1}+\boldsymbol{E}_1,$$

where

$$\|\boldsymbol{E}\|_{op}\le\left\|\boldsymbol{E}^t\right\|_{op}.$$

Hence, by using (85), we get that

$$\begin{aligned}\boldsymbol{A}_t^\top\boldsymbol{A}_t &= ((\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1}\boldsymbol{B}_t+\boldsymbol{E}_1^\top\boldsymbol{B}_t)(\boldsymbol{B}_t^\top(\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1}+\boldsymbol{B}_t^\top\boldsymbol{E}_1)\\ &= \boldsymbol{B}_t\boldsymbol{B}_t^\top(\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-2}+\boldsymbol{E}_1^\top\boldsymbol{B}_t\boldsymbol{B}_t^\top(\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1}+(\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1}\boldsymbol{B}_t\boldsymbol{B}_t^\top\boldsymbol{E}_1+\boldsymbol{E}_1^\top\boldsymbol{B}_t\boldsymbol{B}_t^\top\boldsymbol{E}_1,\end{aligned} \tag{108}$$

where we rearranged the first term in (108) using that $\boldsymbol{B}_t\boldsymbol{B}_t^\top$ and $(\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1}$ commute. By using the assumptions (88), we have that

$$\|\boldsymbol{B}_t\boldsymbol{B}_t^\top\|_{op}\le C,\qquad\|\boldsymbol{E}_1\|_{op}\le 1/2,\qquad\|(\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1}\|_{op}\le\frac{1}{\alpha}.$$

Hence, we can upper bound the operator norm of the last three terms in (108) as

$$\left\|\boldsymbol{E}_1^\top\boldsymbol{B}_t\boldsymbol{B}_t^\top(\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1}+(\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1}\boldsymbol{B}_t\boldsymbol{B}_t^\top\boldsymbol{E}_1+\boldsymbol{E}_1^\top\boldsymbol{B}_t\boldsymbol{B}_t^\top\boldsymbol{E}_1\right\|_{op}\le C\|\boldsymbol{E}_1\|_{op}. \tag{109}$$

Let us now take a closer look at the first term in (108). Recall that

$$\boldsymbol{B}_t\boldsymbol{B}_t^\top=\boldsymbol{U}\boldsymbol{\Lambda}_t\boldsymbol{U}^\top+\boldsymbol{X}_t.$$

As the operator norm is sub-multiplicative, we have that

$$\|\boldsymbol{X}_t\cdot(\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-2}\|_{op}\le C\|\boldsymbol{X}_t\|_{op}. \tag{110}$$

Furthermore,

$$\begin{aligned}\boldsymbol{U}\boldsymbol{\Lambda}_t\boldsymbol{U}^\top(\alpha\boldsymbol{I}+\boldsymbol{U}\boldsymbol{\Lambda}_t\boldsymbol{U}^\top+\boldsymbol{X}_t)^{-2} &= \boldsymbol{U}\boldsymbol{\Lambda}_t\boldsymbol{U}^\top\big((\boldsymbol{I}+\boldsymbol{X}_t(\alpha\boldsymbol{I}+\boldsymbol{U}\boldsymbol{\Lambda}_t\boldsymbol{U}^\top)^{-1})(\alpha\boldsymbol{I}+\boldsymbol{U}\boldsymbol{\Lambda}_t\boldsymbol{U}^\top)\big)^{-2}\\ &= \boldsymbol{U}\boldsymbol{\Lambda}_t\boldsymbol{U}^\top\boldsymbol{T}_1^{-1}\boldsymbol{T}_2^{-1}\boldsymbol{T}_1^{-1}\boldsymbol{T}_2^{-1},\end{aligned} \tag{111}$$

where we have defined

$$\boldsymbol{T}_1=\alpha\boldsymbol{I}+\boldsymbol{U}\boldsymbol{\Lambda}_t\boldsymbol{U}^\top,\qquad \boldsymbol{T}_2=\boldsymbol{I}+\boldsymbol{X}_t(\alpha\boldsymbol{I}+\boldsymbol{U}\boldsymbol{\Lambda}_t\boldsymbol{U}^\top)^{-1}.$$

By expanding $\boldsymbol{T}_2^{-1}$ as in (103)-(105), we get

$$\|\boldsymbol{T}_2^{-1}-\boldsymbol{I}\|_{op}\le C\|\boldsymbol{X}_t\|_{op},$$

or equivalently

$$\boldsymbol{T}_2^{-1}=\boldsymbol{I}+\boldsymbol{E}_2,$$

with $\|\boldsymbol{E}_2\|_{op}\le C\|\boldsymbol{X}_t\|_{op}$. In this view, looking at (111) we have

$$\boldsymbol{U}\boldsymbol{\Lambda}_t\boldsymbol{U}^\top\boldsymbol{T}_1^{-1}\boldsymbol{T}_2^{-1}\boldsymbol{T}_1^{-1}\boldsymbol{T}_2^{-1}=\boldsymbol{U}\boldsymbol{\Lambda}_t\boldsymbol{B}\boldsymbol{U}^\top\boldsymbol{T}_1^{-1}(\boldsymbol{I}+\boldsymbol{E}_2)\boldsymbol{T}_1^{-1}(\boldsymbol{I}+\boldsymbol{E}_2).$$

All the terms which involve $\boldsymbol{E}_2$ can be controlled. We provide the analysis for two terms of different nature, the rest follows from similar arguments. As $\|\boldsymbol{T}_1^{-1}\|_{op}\le 1/\alpha$ and $\|\boldsymbol{\Lambda}_t\|_{op}\le C$, we have that

$$\begin{aligned}\|\boldsymbol{U}\boldsymbol{\Lambda}_t\boldsymbol{U}^\top\boldsymbol{T}_1^{-1}\boldsymbol{E}_2\boldsymbol{T}_1^{-1}\boldsymbol{E}_2\|_{op}&\le\|\boldsymbol{T}_1^{-1}\|_{op}^2\|\boldsymbol{E}_2\|_{op}^2\le\frac{C}{\alpha^2}\|\boldsymbol{X}_t\|_{op}^2\le\frac{C}{\alpha^2}\|\boldsymbol{X}_t\|_{op},\\ \|\boldsymbol{U}\boldsymbol{\Lambda}_t\boldsymbol{U}^\top\boldsymbol{T}_1^{-1}\boldsymbol{I}\boldsymbol{T}_1^{-1}\boldsymbol{E}_2\|_{op}&\le\|\boldsymbol{T}_1^{-1}\|_{op}^2\|\boldsymbol{E}_2\|_{op}\le\frac{C}{\alpha^2}\|\boldsymbol{X}_t\|_{op},\end{aligned}$$

where we have also used that $\|\boldsymbol{X}_t\|_{op}$ is bounded via assumptions (88). Furthermore, a simple manipulation gives

$$\boldsymbol{U}\boldsymbol{\Lambda}_t\boldsymbol{U}^\top\boldsymbol{T}_1^{-2}=\boldsymbol{U}\boldsymbol{\Lambda}_t\boldsymbol{U}^\top(\alpha\boldsymbol{I}+\boldsymbol{U}\boldsymbol{\Lambda}_t\boldsymbol{U}^\top)^{-2}=\boldsymbol{U}\boldsymbol{\Lambda}_t(\alpha\boldsymbol{I}+\boldsymbol{\Lambda}_t)^{-2}\boldsymbol{U}^\top=\boldsymbol{U}\phi(\boldsymbol{\Lambda}_t)\boldsymbol{U}^\top,$$

where $\phi(x)=\frac{x}{(\alpha+x)^2}$. As a result,

$$\left\|\boldsymbol{U}\boldsymbol{\Lambda}_t\boldsymbol{U}^\top\boldsymbol{T}_1^{-1}\boldsymbol{T}_2^{-1}\boldsymbol{T}_1^{-1}\boldsymbol{T}_2^{-1}-\boldsymbol{U}\phi(\boldsymbol{\Lambda}_t)\boldsymbol{U}^\top\right\|_{op}\le C\left\|\boldsymbol{X}_t\right\|_{op},$$

which implies that

$$\|\boldsymbol{B}_t\boldsymbol{B}_t^\top(\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-2}-\boldsymbol{U}\phi(\boldsymbol{\Lambda}_t)\boldsymbol{U}^\top\|_{op}\le C\|\boldsymbol{X}_t\|_{op}. \tag{112}$$

By combining (108), (109) and (112), we have that

$$\|\boldsymbol{A}_t^\top\boldsymbol{A}_t-\boldsymbol{U}\phi(\boldsymbol{\Lambda}_t)\boldsymbol{U}^\top\|_{op}\le C\big(\|\boldsymbol{X}_t\|_{op}+\|\boldsymbol{E}_1\|_{op}\big). \tag{113}$$

At this point, we are ready to analyze the operator norm of $\|\boldsymbol{A}_t^\top\boldsymbol{A}_t\circ(\boldsymbol{B}_t\boldsymbol{B}_t^\top-\boldsymbol{I})^{\circ 2}\|_{op}$:

$$\begin{aligned}\boldsymbol{A}_t^\top\boldsymbol{A}_t\circ(\boldsymbol{B}_t\boldsymbol{B}_t^\top-\boldsymbol{I})^{\circ 2}&=(\boldsymbol{U}\phi(\boldsymbol{\Lambda}_t)\boldsymbol{U}^\top+\boldsymbol{E}_3)\circ(\boldsymbol{U}(\boldsymbol{\Lambda}_t-\boldsymbol{I})\boldsymbol{U}^\top+\boldsymbol{X}_t)^{\circ 2}\circ\boldsymbol{H}\\ &=(\boldsymbol{U}\phi(\boldsymbol{\Lambda}_t)\boldsymbol{U}^\top+\boldsymbol{E}_3)\circ((\boldsymbol{U}(\boldsymbol{\Lambda}_t-\boldsymbol{I})\boldsymbol{U}^\top)^{\circ 2}+\boldsymbol{X}_t^{\circ 2}+2(\boldsymbol{U}(\boldsymbol{\Lambda}_t-\boldsymbol{I})\boldsymbol{U}^\top)\circ\boldsymbol{X}_t)\circ\boldsymbol{H},\end{aligned} \tag{114}$$

where we have defined $\boldsymbol{H}:=\boldsymbol{1}\boldsymbol{1}^\top-\boldsymbol{I}$ and $\|\boldsymbol{E}_3\|_{op}\le C\big(\|\boldsymbol{X}_t\|_{op}+\|\boldsymbol{E}_1\|_{op}\big)$. We now decompose the quantity into three terms:

$$\boldsymbol{A}_t^\top\boldsymbol{A}_t\circ(\boldsymbol{B}_t\boldsymbol{B}_t^\top-\boldsymbol{I})^{\circ 2}=\boldsymbol{S}_1+\boldsymbol{S}_2+\boldsymbol{S}_3,$$

where

$$\begin{aligned}\boldsymbol{S}_1&=(\boldsymbol{U}\phi(\boldsymbol{\Lambda}_t)\boldsymbol{U}^\top\circ\boldsymbol{U}(\boldsymbol{\Lambda}_t-\boldsymbol{I})\boldsymbol{U}^\top\circ\boldsymbol{H})\circ\boldsymbol{U}(\boldsymbol{\Lambda}_t-\boldsymbol{I})\boldsymbol{U}^\top,\\ \boldsymbol{S}_2&=\boldsymbol{H}\circ\boldsymbol{E}_3\circ((\boldsymbol{U}(\boldsymbol{\Lambda}_t-\boldsymbol{I})\boldsymbol{U}^\top)^{\circ 2}+\boldsymbol{X}_t^{\circ 2}+2(\boldsymbol{U}(\boldsymbol{\Lambda}_t-\boldsymbol{I})\boldsymbol{U}^\top)\circ\boldsymbol{X}_t),\\ \boldsymbol{S}_3&=\boldsymbol{H}\circ\boldsymbol{U}\phi(\boldsymbol{\Lambda}_t)\boldsymbol{U}^\top\circ(\boldsymbol{X}_t^{\circ 2}+2(\boldsymbol{U}(\boldsymbol{\Lambda}_t-\boldsymbol{I})\boldsymbol{U}^\top)\circ\boldsymbol{X}_t).\end{aligned}$$

We proceed to bound each of these terms separately.

We start with $\boldsymbol{S}_1$. As $\phi(x)$ is differentiable for $x \geq 0$, the derivative of $\phi(x)$ is bounded for any compact interval $I \subseteq \mathbb{R}_+$. Hence, $\phi(x)$ is locally Lipschitz on I with Lipschitz constant $C_I > 0$, which implies that

$$|\phi(x) - \phi(1)| = \left|\phi(x) - \frac{1}{(1+\alpha)^2}\right| \leq C_I |x-1|.$$

By assumption (88), we have that $\boldsymbol{\Lambda}_t \succ 0$ and $\|\boldsymbol{\Lambda}_t\|_{op} \leq C$, hence

$$\left\| \boldsymbol{U}\phi(\boldsymbol{\Lambda}_t)\boldsymbol{U}^\top - \frac{1}{(1+\alpha)^2}\boldsymbol{I} \right\|_{op} \leq C_I \cdot \|\boldsymbol{Z}_t\|_{op}. \tag{115}$$

Hence, an application of Lemma G.2 gives that, with probability at least $1 - 1/d^2$,

$$\sup_{t \geq 0} m\left(\boldsymbol{U}\phi(\boldsymbol{\Lambda}_t)\boldsymbol{U}^\top - \frac{1}{(1+\alpha)^2}\boldsymbol{I} \right) \leq c\sqrt{\frac{\log d}{d}}, \tag{116}$$

where $c > 0$ is a universal constant. Another application of Lemma G.2 also gives that, with the same probability,

$$\sup_{t \geq 0} m\left(\boldsymbol{U}(\boldsymbol{\Lambda}_t - \boldsymbol{I})\boldsymbol{U}^\top \right) \leq c\sqrt{\frac{\log d}{d}}. \tag{117}$$

As a result, we obtain the bound

$$\|\boldsymbol{S}_1\|_{op} = \|([\boldsymbol{U}\phi(\boldsymbol{\Lambda}_t)\boldsymbol{U}^\top - 1/(1+\alpha)^2\boldsymbol{I}] \circ \boldsymbol{U}(\boldsymbol{\Lambda}_t - \boldsymbol{I})\boldsymbol{U}^\top \circ \boldsymbol{H}) \circ \boldsymbol{U}(\boldsymbol{\Lambda}_t - \boldsymbol{I})\boldsymbol{U}^\top\|_{op} \leq C\frac{\log d}{\sqrt{d}}\|\boldsymbol{Z}_t\|_{op}. \tag{118}$$

Here, the first equality is due to the fact that we are taking the Hadamard product with the matrix $\boldsymbol{H}$ which has 0 on the diagonal, hence we can add multiples of the identity to $\boldsymbol{U}\phi(\boldsymbol{\Lambda}_t)\boldsymbol{U}^\top$; and the second inequality uses (93) with $\boldsymbol{R} = [\boldsymbol{U}\phi(\boldsymbol{\Lambda}_t)\boldsymbol{U}^\top - 1/(1+\alpha)^2\boldsymbol{I}] \circ \boldsymbol{U}(\boldsymbol{\Lambda}_t - \boldsymbol{I})\boldsymbol{U}^\top \circ \boldsymbol{H}$ and $\boldsymbol{S} = \boldsymbol{U}(\boldsymbol{\Lambda}_t - \boldsymbol{I})\boldsymbol{U}^\top$ in combination with (116)-(117).

Next, we bound $\|\boldsymbol{S}_2\|_{op}$. We inspect the terms appearing in the expression for $\boldsymbol{S}_2$ one by one. First note that we can omit $\boldsymbol{H}$ in the expression since, by Lemma G.1 for any square matrix $\boldsymbol{R}$

$$\|\boldsymbol{R} \circ \boldsymbol{H}\|_{op} \leq C\|\boldsymbol{R}\|_{op}. \tag{119}$$

Hence, by using (97), we get

$$\begin{aligned}
&\|\boldsymbol{H} \circ \boldsymbol{E}_3 \circ ((\boldsymbol{U}(\boldsymbol{\Lambda}_t - \boldsymbol{I})\boldsymbol{U}^\top)^{\circ 2}\|_{op} \leq C\|\boldsymbol{E}_3\|_{op}\|\boldsymbol{Z}_t\|_{op}^2 \\
&\|\boldsymbol{H} \circ \boldsymbol{E}_3 \circ \boldsymbol{X}_t^{\circ 2}\|_{op} \leq C\|\boldsymbol{E}_3\|_{op}\|\boldsymbol{X}_t\|_{op}^2 \\
&\|\boldsymbol{H} \circ \boldsymbol{E}_3 \circ 2(\boldsymbol{U}(\boldsymbol{\Lambda}_t - \boldsymbol{I})\boldsymbol{U}^\top) \circ \boldsymbol{X}_t)\|_{op} \leq C\|\boldsymbol{E}_3\|_{op}\|\boldsymbol{X}_t\|_{op}\|\boldsymbol{Z}_t\|_{op},
\end{aligned}$$

which leads to the bound

$$\|\boldsymbol{S}_2\|_{op} \leq C\|\boldsymbol{E}_3\|_{op}\left(\|\boldsymbol{X}_t\|_{op}^2 + \|\boldsymbol{Z}_t\|_{op}^2 + \|\boldsymbol{X}_t\|_{op}\|\boldsymbol{Z}_t\|_{op}\right). \tag{120}$$

Finally, we bound $\|\boldsymbol{S}_3\|_{op}$. Consider the term

$$\|[\boldsymbol{H} \circ \boldsymbol{U}\phi(\boldsymbol{\Lambda}_t)\boldsymbol{U}^\top \circ 2(\boldsymbol{U}(\boldsymbol{\Lambda}_t - \boldsymbol{I})\boldsymbol{U}^\top] \circ \boldsymbol{X}_t\|_{op}.$$

Then, by using (119) and (115), we have

$$\|\boldsymbol{H} \circ \boldsymbol{U}\phi(\boldsymbol{\Lambda}_t)\boldsymbol{U}^\top\|_{op} = \left\|\boldsymbol{H} \circ [\boldsymbol{U}\phi(\boldsymbol{\Lambda}_t)\boldsymbol{U}^\top - \frac{1}{(1+\alpha)^2}\boldsymbol{I}]\right\|_{op} \leq C\left\|\boldsymbol{U}\phi(\boldsymbol{\Lambda}_t)\boldsymbol{U}^\top - \frac{1}{(1+\alpha)^2}\boldsymbol{I}\right\|_{op} \leq C\|\boldsymbol{Z}_t\|_{op}. \tag{121}$$

Hence, in conjunction with (97), we get

$$\|\boldsymbol{H} \circ \boldsymbol{U}\phi(\boldsymbol{\Lambda}_t)\boldsymbol{U}^\top \circ 2\boldsymbol{U}(\boldsymbol{\Lambda}_t - \boldsymbol{I})\boldsymbol{U}^\top\|_{op} \leq C \cdot \|\boldsymbol{Z}_t\|_{op}^2,$$

which invoking (97) one more time gives

$$\|[\boldsymbol{H}\circ\boldsymbol{U}\phi(\boldsymbol{\Lambda}_t)\boldsymbol{U}^\top\circ 2(\boldsymbol{U}(\boldsymbol{\Lambda}_t-\boldsymbol{I})\boldsymbol{U}^\top]\circ\boldsymbol{X}_t\|_{op}\le C\|\boldsymbol{Z}_t\|_{op}^2\|\boldsymbol{X}_t\|_{op}.$$

Furthermore, by combining (97) and (121), we get

$$\|[\boldsymbol{H}\circ\boldsymbol{\Lambda}\phi(\boldsymbol{\Lambda}_t)\boldsymbol{\Lambda}^\top]\circ\boldsymbol{X}_t^{\circ 2}\|_{op}\le C\|\boldsymbol{Z}_t\|_{op}\|\boldsymbol{X}_t\|_{op}^2.$$

Thus,

$$\|\boldsymbol{S}_3\|_{op}\le C(\|\boldsymbol{Z}_t\|_{op}^2\|\boldsymbol{X}_t\|_{op}+\|\boldsymbol{Z}_t\|_{op}\|\boldsymbol{X}_t\|_{op}^2). \tag{122}$$

Recall that, from assumptions (88)-(89), $\|\boldsymbol{X}_t\|_{op},\|\boldsymbol{Z}_t\|_{op}\le C$. Then, by combining the bounds in (118), (120) and (122), the desired result readily follows. □

By exploiting the above lemmas, we are able to make the following approximation for the gradient.

Lemma E.4 (Gradient approximation)**.** *Assume that* (88) *holds, and let* $\nabla_{\boldsymbol{B}_t}$ *be given by* (86)*. Further define* $\gamma=1+\alpha$ *and* $F(x)=\frac{1+x}{(\gamma+x)^2}$*. Then, for all sufficiently large* n*, with probability* $1-1/d^2$*, jointly for all* $t\ge 0$*,*

$$\left\|\frac{1}{2}\nabla_{\boldsymbol{B}_t}\boldsymbol{B}_t^\top+\alpha F(\boldsymbol{Z}_t)-\alpha\mathrm{Diag}\left(F(\boldsymbol{Z}_t)\right)(\boldsymbol{I}+\boldsymbol{Z}_t)-\frac{2\alpha}{\gamma^3}\boldsymbol{X}_t^O-\frac{\alpha}{\gamma^2}\boldsymbol{X}_t^D\right\|_{op}\le\|\boldsymbol{E}^t\|_{op}. \tag{123}$$

Proof of Lemma E.4. We start by showing that, with probability $1-1/d^2$, jointly for all $t\ge 0$,

$$\left\|\frac{1}{2}\nabla_{\boldsymbol{B}_t}+\alpha(\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-2}\boldsymbol{B}_t-\alpha\mathrm{Diag}\left((\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-2}(\boldsymbol{B}_t\boldsymbol{B}_t^\top)\right)\boldsymbol{B}_t\right\|_{op}\le\|\boldsymbol{E}^t\|_{op}. \tag{124}$$

Let us first consider the term $\nabla^1_{\boldsymbol{B}_t}$, which can be equivalently expressed as

$$\nabla^1_{\boldsymbol{B}_t}=2\big(-\boldsymbol{A}_t^\top+\mathrm{Diag}(\boldsymbol{B}_t\boldsymbol{A}_t)\boldsymbol{B}_t+\boldsymbol{T}\boldsymbol{B}_t-\mathrm{Diag}(\boldsymbol{T}(\boldsymbol{B}_t\boldsymbol{B}_t^\top))\boldsymbol{B}_t\big),$$

where $\boldsymbol{T}=\boldsymbol{A}_t^\top\boldsymbol{A}_t-\mathrm{Diag}(\boldsymbol{A}_t^\top\boldsymbol{A}_t)$. It is then easy to verify that

$$\frac{1}{2}\nabla^1_{\boldsymbol{B}_t}=-\boldsymbol{A}_t^\top+\boldsymbol{A}_t^\top\boldsymbol{A}_t\boldsymbol{B}_t+\mathrm{Diag}(\boldsymbol{B}_t\boldsymbol{A}_t)\boldsymbol{B}_t-\mathrm{Diag}(\boldsymbol{A}_t^\top\boldsymbol{A}_t\boldsymbol{B}_t\boldsymbol{B}_t^\top)\boldsymbol{B}_t. \tag{125}$$

Using Lemma E.2, we get

$$\boldsymbol{A}_t^\top\boldsymbol{A}_t=((\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1}+\boldsymbol{E}_1)\boldsymbol{B}_t\boldsymbol{B}_t^\top((\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1}+\boldsymbol{E}_1), \tag{126}$$

where $\|\boldsymbol{E}_1\|_{op}\le\|\boldsymbol{E}^t\|_{op}$. It follows from (88) that $\|\boldsymbol{B}_t\boldsymbol{B}_t^\top\|_{op}\le C$. Hence, using that $\boldsymbol{B}_t\boldsymbol{B}_t^\top$ and $(\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)$ commute in conjunction with $\|(\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1}\|_{op}\le 1/\alpha$ we get

$$\boldsymbol{A}_t^\top\boldsymbol{A}_t=\boldsymbol{B}_t\boldsymbol{B}_t^\top(\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-2}+\boldsymbol{E}_2, \tag{127}$$

where $\|\boldsymbol{E}_2\|_{op}\le\|\boldsymbol{E}^t\|_{op}$. Noting that $\frac{1}{\alpha+x}-\frac{\alpha}{(\alpha+x)^2}=\frac{x}{(\alpha+x)^2}$ and using the spectral theorem for the symmetric matrix $\boldsymbol{B}_t\boldsymbol{B}_t^\top$, we can further rewrite (127) as

$$\boldsymbol{A}_t^\top\boldsymbol{A}_t=(\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1}-\alpha(\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-2}+\boldsymbol{E}_2. \tag{128}$$

With similar arguments, by Lemma E.2, we can write

$$\boldsymbol{B}_t\boldsymbol{A}_t=\boldsymbol{B}_t\boldsymbol{B}_t^\top(\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1}+\boldsymbol{E}_3, \tag{129}$$

where $\|\boldsymbol{E}_3\|_{op}\le\|\boldsymbol{E}^t\|_{op}$. Noting that $1-\frac{\alpha}{\alpha+x}=\frac{x}{\alpha+x}$, again by the spectral theorem for $\boldsymbol{B}_t\boldsymbol{B}_t^\top$, we get

$$\boldsymbol{B}_t\boldsymbol{A}_t=\boldsymbol{I}-\alpha(\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1}+\boldsymbol{E}_3, \tag{130}$$

and, consequently, we obtain

$$\mathrm{Diag}(\boldsymbol{B}_t\boldsymbol{A}_t)\boldsymbol{B}_t = \boldsymbol{B}_t - \alpha\mathrm{Diag}((\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1})\boldsymbol{B}_t + \boldsymbol{E}_4, \tag{131}$$

where $\|\boldsymbol{E}_4\|_{op} \le \|\boldsymbol{E}^t\|_{op}$. Using (128) and $1-\frac{\alpha}{\alpha+x} = \frac{1}{x+\alpha}$, we get

$$\begin{aligned}\mathrm{Diag}(\boldsymbol{A}_t^\top\boldsymbol{A}_t\boldsymbol{B}_t\boldsymbol{B}_t^\top)\boldsymbol{B}_t &= \mathrm{Diag}((\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1}\boldsymbol{B}_t\boldsymbol{B}_t^\top)\boldsymbol{B}_t - \alpha\mathrm{Diag}((\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-2}\boldsymbol{B}_t\boldsymbol{B}_t^\top)\boldsymbol{B}_t + \boldsymbol{E}_5\\ &= \boldsymbol{B}_t - \alpha\mathrm{Diag}((\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1})\boldsymbol{B}_t - \alpha\mathrm{Diag}((\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-2}\boldsymbol{B}_t\boldsymbol{B}_t^\top)\boldsymbol{B}_t + \boldsymbol{E}_5,\end{aligned} \tag{132}$$

where $\|\boldsymbol{E}_5\|_{op} \le \|\boldsymbol{E}^t\|_{op}$.

With this in mind, we get back to (125). Combining the results of (128), (131) and (132) we get

$$\begin{aligned}\nabla^1_{\boldsymbol{B}_t} &= \underbrace{-(\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1}\boldsymbol{B}_t}_{-\boldsymbol{A}_t^\top} + \underbrace{(\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1}\boldsymbol{B}_t - \alpha(\alpha+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-2}\boldsymbol{B}_t}_{\boldsymbol{A}_t^\top\boldsymbol{A}_t\boldsymbol{B}_t} + \underbrace{\boldsymbol{B}_t - \alpha\mathrm{Diag}((\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1})\boldsymbol{B}_t}_{\mathrm{Diag}(\boldsymbol{B}_t\boldsymbol{A}_t)\boldsymbol{B}_t}\\ &\underbrace{-\boldsymbol{B}_t + \alpha\mathrm{Diag}((\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1})\boldsymbol{B}_t + \alpha\mathrm{Diag}((\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-2}\boldsymbol{B}_t\boldsymbol{B}_t^\top)\boldsymbol{B}_t}_{-\mathrm{Diag}(\boldsymbol{A}_t^\top\boldsymbol{A}_t\boldsymbol{B}_t\boldsymbol{B}_t^\top)\boldsymbol{B}_t} + \boldsymbol{E}_6\\ &= -\alpha(\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-2}\boldsymbol{B}_t + \alpha\mathrm{Diag}((\alpha\boldsymbol{I}+\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-2}\boldsymbol{B}_t\boldsymbol{B}_t^\top)\boldsymbol{B}_t + \boldsymbol{E}_6,\end{aligned} \tag{133}$$

where $\|\boldsymbol{E}_6\|_{op} \le \|\boldsymbol{E}^t\|_{op}$.

Let us now analyze the second part of the gradient which involves terms of the form below for $\ell \ge 3$:

$$\nabla^{2,k,\ell}_{\boldsymbol{B}_t} := c_\ell^2\cdot\ell\cdot\sum_{j\neq k}\langle\boldsymbol{a}_k,\boldsymbol{a}_j\rangle\langle\boldsymbol{b}_k,\boldsymbol{b}_j\rangle^{(\ell-1)}\boldsymbol{J}_k\boldsymbol{b}_j.$$

Now, from the fact that

$$\boldsymbol{J}_k = \boldsymbol{I} - \boldsymbol{b}_k\boldsymbol{b}_k^\top,$$

we can write

$$c_\ell^2\cdot\ell\cdot\sum_{j\neq k}\langle\boldsymbol{a}_k,\boldsymbol{a}_j\rangle\langle\boldsymbol{b}_k,\boldsymbol{b}_j\rangle^{(\ell-1)}\boldsymbol{J}_k\boldsymbol{b}_j = c_\ell^2\cdot\ell\cdot\sum_{j\neq k}\langle\boldsymbol{a}_k,\boldsymbol{a}_j\rangle\langle\boldsymbol{b}_k,\boldsymbol{b}_j\rangle^{(\ell-1)}(\boldsymbol{b}_j - \langle\boldsymbol{b}_k,\boldsymbol{b}_j\rangle\boldsymbol{b}_k). \tag{134}$$

The second term of the RHS gives the following contribution to the $\boldsymbol{B}_t$ update

$$\mathrm{Diag}(\boldsymbol{A}_t^\top\boldsymbol{A}_t(\boldsymbol{B}_t\boldsymbol{B}_t^\top - \boldsymbol{I})^{\circ\ell})\boldsymbol{B}_t.$$

By recalling that $\|\boldsymbol{A}_t^\top\boldsymbol{A}_t\|_{op} \le C$ and $\|\boldsymbol{B}_t\|_{op} \le C$, we have

$$\|\mathrm{Diag}(\boldsymbol{A}_t^\top\boldsymbol{A}_t(\boldsymbol{B}_t\boldsymbol{B}_t^\top - \boldsymbol{I})^{\circ\ell})\boldsymbol{B}_t\|_{op} \le C\|\boldsymbol{A}_t^\top\boldsymbol{A}_t(\boldsymbol{B}_t\boldsymbol{B}_t^\top - \boldsymbol{I})^{\circ\ell}\|_{op}\|\boldsymbol{B}_t\|_{op} \le C\|(\boldsymbol{B}_t\boldsymbol{B}_t^\top - \boldsymbol{I})^{\circ\ell}\|_{op}. \tag{135}$$

Now, for $\ell < 5$, we upper bound the RHS of (135) via Lemma E.2, which gives that

$$\|\mathrm{Diag}(\boldsymbol{A}_t^\top\boldsymbol{A}_t(\boldsymbol{B}_t\boldsymbol{B}_t^\top - \boldsymbol{I})^{\circ\ell})\boldsymbol{B}_t\|_{op} \le C\|\boldsymbol{E}^t\|_{op}. \tag{136}$$

Furthermore, if we follow passages analogous to (94)-(95) (the only difference being that we exchange the roles of the Hadamard powers 3 and $\ell-3$), we have that, with probability at least $1-1/d^2$, jointly for all $t\ge 0$ and $\ell \ge 5$,

$$\|\mathrm{Diag}(\boldsymbol{A}_t^\top\boldsymbol{A}_t(\boldsymbol{B}_t\boldsymbol{B}_t^\top - \boldsymbol{I})^{\circ\ell})\boldsymbol{B}_t\|_{op} \le C\sqrt{n}\|\boldsymbol{E}^t\|_{op}\left(\frac{\mathrm{poly}(\log d)}{d}\right)^{(\ell-3)/2} \le C\|\boldsymbol{E}^t\|_{op}\left(\frac{\mathrm{poly}(\log d)}{d}\right)^{(\ell-4)/2}, \tag{137}$$

for sufficiently large d.

Define the following quantity:

$$\boldsymbol{Y} = (\boldsymbol{A}_t^\top \boldsymbol{A}_t) \circ (\boldsymbol{B}_t \boldsymbol{B}_t^\top - \boldsymbol{I})^{\circ(\ell-1)}. \tag{138}$$

In this view, the first term in (134) can be written as $\boldsymbol{Y}\boldsymbol{B}_t$. For $l < 5$, by Lemma E.3 we have that $\|\boldsymbol{Y}\|_{op} \le \|\boldsymbol{E}^t\|_{op}$, hence $\|\boldsymbol{Y}\boldsymbol{B}_t\|_{op} \le C\|\boldsymbol{E}^t\|_{op}$ as $\|\boldsymbol{B}_t\|_{op} \le C$. Furthermore, with probability at least $1 - 1/d^2$, jointly for all $t \ge 0$ and $\ell \ge 5$, we have

$$\begin{aligned}
\|\boldsymbol{Y}\boldsymbol{B}_t\|_{op} \le C\|\boldsymbol{Y}\|_{op} &= C\sqrt{n}\|(\boldsymbol{A}_t^\top \boldsymbol{A}_t) \circ (\boldsymbol{B}_t \boldsymbol{B}_t^\top - \boldsymbol{I})^{\circ 2}\|_{op} \max_{i,j} |(\boldsymbol{B}_t \boldsymbol{B}_t^\top - \boldsymbol{I})_{i,j}|^{\ell-3} \\
&\le \sqrt{n}\|\boldsymbol{E}^t\|_{op} \max_{i,j} |(\boldsymbol{B}_t \boldsymbol{B}_t^\top - \boldsymbol{I})_{i,j}|^{\ell-3} \\
&\le \sqrt{n}\|\boldsymbol{E}^t\|_{op} \left[(C + C_X)^{\ell-3} \left(\frac{\mathrm{poly}(\log d)}{d} \right)^{(\ell-3)/2} \right] \\
&\le (C + C_X)^{\ell-3} \|\boldsymbol{E}^t\|_{op} \left(\frac{\mathrm{poly}(\log d)}{d} \right)^{(\ell-4)/2}.
\end{aligned} \tag{139}$$

Here, in the second line we use Lemma E.3; and in the third line we bound the off-diagonal entries of $\boldsymbol{X}_t$ via (88) and the off-diagonal entries of $\boldsymbol{Z}_t$ via Lemma G.2. Hence, by combining (137) and (139), we conclude that

$$\left\|\nabla^2_{\boldsymbol{B}_t}\right\|_{op} \le C\|\boldsymbol{E}^t\|_{op} + \|\boldsymbol{E}^t\|_{op} \sum_{\ell=5}^{\infty} (C + C_X)^{\ell-3} c_\ell^2 \, \ell \left(\frac{\mathrm{poly}(\log d)}{\sqrt{d}} \right)^{\ell-4} \le C\|\boldsymbol{E}^t\|_{op}, \tag{140}$$

where we used that the series $\sum_{\ell=5}^{\infty} (C + C_X)^{\ell-3} c_\ell^2 \, \ell \left(\frac{(\mathrm{poly}(\log d)}{\sqrt{d}} \right)^{\ell-4}$ converges to a finite value for all sufficiently large d, since $(C + C_X)\frac{\mathrm{poly}(\log d)}{\sqrt{d}} < 1$. This finishes the proof of (124).

We now further analyse the gradient in (124). Defining $F(x) = \frac{1+x}{(\gamma+x)^2}$, with $\gamma = 1 + \alpha$, we can write

$$\frac{1}{2}\nabla_{\boldsymbol{B}_t} \boldsymbol{B}_t^\top = -\alpha F(\boldsymbol{Z}_t + \boldsymbol{X}_t) + \alpha \mathrm{Diag}\left(F(\boldsymbol{Z}_t + \boldsymbol{X}_t)\right)) + \alpha \mathrm{Diag}\left(F(\boldsymbol{Z}_t + \boldsymbol{X}_t)\right)) (\boldsymbol{Z}_t + \boldsymbol{X}_t) + \boldsymbol{E}^t. \tag{141}$$

By a slight abuse of notation, we will denote by $F^{(l)}(0)$ the l-th derivative of the unidimensional function $F(x) = \frac{1+x}{(\gamma+x)^2}$ computed at $x = 0$. Here, $F(\boldsymbol{Z}_t + \boldsymbol{X}_t)$ is defined by the spectral theorem (note that indeed $\boldsymbol{Z}_t + \boldsymbol{X}_t = \boldsymbol{B}_t \boldsymbol{B}_t^\top - \boldsymbol{I}$ is symmetric).

We will now compute the error we incur if in (141) we replace $F(\boldsymbol{X}_t + \boldsymbol{Z}_t)$ by $F(\boldsymbol{Z}_t)$. We first consider the case when $\|\boldsymbol{Z}_t\|_{op} > \frac{\gamma}{3}$. In this case, we have that

$$\left\| F(\boldsymbol{Z}_t + \boldsymbol{X}_t) - F(\boldsymbol{Z}_t) - F^{(1)}(0)\boldsymbol{X}_t \right\|_{op} \le C \left\|\boldsymbol{X}_t\right\|_{op} \le C \left\|\boldsymbol{Z}_t\right\|_{op} \left\|\boldsymbol{X}_t\right\|_{op}. \tag{142}$$

Here, the second inequality trivially holds since $\|\boldsymbol{Z}_t\|_{op} > \frac{\gamma}{3}$. To prove the first inequality, let DF be the derivative of the matrix-valued function $F(\boldsymbol{M}) = (\boldsymbol{I} + \boldsymbol{M})(\gamma\boldsymbol{I} + \boldsymbol{M})^{-2}$. Then, by evaluating this derivative for $\boldsymbol{M} = \boldsymbol{Z}_t$ in the direction of $\boldsymbol{X}_t$, we obtain

$$DF(\boldsymbol{Z}_t)\,\boldsymbol{X}_t = -(\boldsymbol{I}+\boldsymbol{Z}_t)(\gamma\boldsymbol{I}+\boldsymbol{Z}_t)^{-1}\boldsymbol{X}_t(\gamma\boldsymbol{I}+\boldsymbol{Z}_t)^{-2} - (\boldsymbol{I}+\boldsymbol{Z}_t)(\gamma\boldsymbol{I}+\boldsymbol{Z}_t)^{-2}\boldsymbol{X}_t(\gamma\boldsymbol{I}+\boldsymbol{Z}_t)^{-1} + \boldsymbol{X}_t(\gamma\boldsymbol{I}+\boldsymbol{Z}_t)^{-2}. \tag{143}$$

To verify this expression we first note that the derivative of the function $G(\boldsymbol{M}) = \boldsymbol{M}^{-1}$ in the direction of $\boldsymbol{X}$ is given by $DG(\boldsymbol{M})\boldsymbol{X} = -\boldsymbol{M}^{-1}\boldsymbol{X}\boldsymbol{M}^{-1}$. Now, (143) easily follows from the product rule applied to $F(\boldsymbol{Z}) = (\boldsymbol{I} + \boldsymbol{Z})(\gamma\boldsymbol{I} + \boldsymbol{Z})^{-1}(\gamma\boldsymbol{I} + \boldsymbol{Z})^{-1}$. By the assumptions in (88), we have that $\boldsymbol{Z}_t, (\gamma\boldsymbol{I} + \boldsymbol{Z}_t)^{-1}$ are uniformly bounded, hence the map DF is uniformly bounded as well. This implies that

$$\|F(\boldsymbol{Z}_t + \boldsymbol{X}_t) - F(\boldsymbol{Z}_t)\|_{op} \le C \left\|\boldsymbol{X}_t\right\|_{op}.$$

As $\left\|F^{(1)}(0)\boldsymbol{X}_t\right\|_{op} \le C \left\|\boldsymbol{X}_t\right\|_{op}$, we readily obtain (142).

Now we consider the case where $\|\boldsymbol{Z}_t\|_{op} \leq \frac{\gamma}{3}$. First note that, by (88), $\|\boldsymbol{X}_t\|_{op} \leq \frac{\gamma}{3}$. Hence,

$$F(\boldsymbol{Z}_t + \boldsymbol{X}_t) = \sum_{\ell=0}^{\infty} F^{(\ell)}(0) \frac{(\boldsymbol{Z}_t + \boldsymbol{X}_t)^\ell}{\ell!}.$$

The series above converges absolutely since $F^{(\ell)}(0)$ scales as $\frac{\ell!}{\gamma^\ell}\text{poly}(\ell)$. To see this, first we note that, if $h(x) = \frac{1}{(\gamma+x)^2}$, then $h^{(\ell)}(0) = (-1)^\ell(\ell+1)!\frac{1}{\gamma^{\ell+2}}$. Thus, by the product rule, $F^{(\ell)}(0) = (-1)^\ell(\ell+1)!\frac{1}{\gamma^{\ell+2}} + (-1)^{\ell-1}\ell!\frac{1}{\gamma^{\ell+1}}$ which has the desired asymptotic behaviour. Expanding the brackets and applying the triangle inequality yields

$$\left\| F(\boldsymbol{Z}_t + \boldsymbol{X}_t) - \sum_{\ell=0}^{\infty} F^{(\ell)}(0)\frac{\boldsymbol{Z}_t^\ell}{\ell!} - F^{(1)}(0)\boldsymbol{X}_t \right\|_{op} \leq \sum_{\ell=2}^{\infty} F^{(\ell)}(0)\frac{\|\boldsymbol{X}_t\|_{op}^\ell}{\ell!} + \sum_{\ell=2}^{\infty} F^{(\ell)}(0)\frac{1}{\ell!}\sum_{i=1}^{\ell-1}\binom{\ell}{i}\|\boldsymbol{Z}_t\|_{op}^i \|\boldsymbol{X}_t\|_{op}^{\ell-i}.$$

As $\|\boldsymbol{Z}_t\|_{op}, \|\boldsymbol{X}_t\|_{op} \leq \frac{\gamma}{3}$, we have

$$\sum_{\ell=2}^{\infty} F^{(\ell)}(0)\frac{\|\boldsymbol{X}_t\|_{op}^\ell}{\ell!} \leq \|\boldsymbol{X}_t\|_{op}^2 \sum_{\ell=2}^{\infty} F^{(\ell)}(0)\left(\frac{\gamma}{3}\right)^{\ell-2}\frac{1}{\ell!} \leq C\|\boldsymbol{X}_t\|_{op}^2,$$

and

$$\sum_{\ell=2}^{\infty} F^{(\ell)}(0)\frac{1}{\ell!}\sum_{i=1}^{\ell-1}\binom{\ell}{i}\|\boldsymbol{Z}_t\|_{op}^i\|\boldsymbol{X}_t\|_{op}^{\ell-i} \leq \sum_{\ell=2}^{\infty} F^{(\ell)}(0)\frac{1}{\ell!}2^\ell\left(\frac{\gamma}{3}\right)^{\ell-2}\|\boldsymbol{Z}_t\|_{op}\|\boldsymbol{X}_t\|_{op} \leq C\|\boldsymbol{Z}_t\|_{op}\|\boldsymbol{X}_t\|_{op}.$$

By combining the last three expressions and using that

$$F(\boldsymbol{Z}_t) = \sum_{\ell=0}^{\infty} F^{(\ell)}(0)\frac{\boldsymbol{Z}_t^\ell}{\ell!},$$

we obtain

$$\left\|F(\boldsymbol{X}_t + \boldsymbol{Z}_t) - F(\boldsymbol{Z}_t) - F^{(1)}(0)\boldsymbol{X}_t\right\|_{op} \leq C\left(\|\boldsymbol{X}_t\|_{op}\|\boldsymbol{Z}_t\|_{op} + \|\boldsymbol{X}_t\|_{op}^2\right). \tag{144}$$

As the map DF is uniformly bounded, we have

$$\|F(\boldsymbol{Z}_t) - F(0)\boldsymbol{I}\|_{op} \leq C\|\boldsymbol{Z}_t\|_{op}. \tag{145}$$

By combining (144), (145) and (141), we obtain

$$\frac{1}{2}\nabla_{\boldsymbol{B}_t}\boldsymbol{B}_t^\top = -\alpha F(\boldsymbol{Z}_t) + \alpha\text{Diag}\left(F(\boldsymbol{Z}_t)\right)(\boldsymbol{I} + \boldsymbol{Z}_t) - \alpha F^{(1)}(0)\boldsymbol{X}_t + \alpha\text{Diag}\left(\boldsymbol{X}_t F^{(1)}(0)\right) + \alpha\boldsymbol{X}_t F(0) + \boldsymbol{E}^t. \tag{146}$$

Using that $F(0) = \frac{1}{\gamma^2}$ and $F^{(1)}(0) = \frac{1}{\gamma^2}(1 - \frac{2}{\gamma})$, we finally obtain

$$\frac{1}{2}\nabla_{\boldsymbol{B}_t}\boldsymbol{B}_t^\top = -\alpha F(\boldsymbol{Z}_t) + \alpha\text{Diag}\left(F(\boldsymbol{Z}_t)\right)(\boldsymbol{I} + \boldsymbol{Z}_t) + \frac{2\alpha}{\gamma^3}\boldsymbol{X}_t^O + \frac{\alpha}{\gamma^2}\boldsymbol{X}_t^D + \boldsymbol{E}^t, \tag{147}$$

which concludes the proof. □

Now let us return to the update equation of $\boldsymbol{B}_t\boldsymbol{B}_t^\top$ during the gradient step

$$\boldsymbol{B}_t'\boldsymbol{B}_t'^\top = (\boldsymbol{B}_t - \eta\nabla_{\boldsymbol{B}_t})(\boldsymbol{B}_t - \eta\nabla_{\boldsymbol{B}_t})^\top = \boldsymbol{B}_t\boldsymbol{B}_t^\top - \eta\cdot\nabla_{\boldsymbol{B}_t}\boldsymbol{B}_t^\top - \eta\cdot\boldsymbol{B}_t(\nabla_{\boldsymbol{B}_t})^\top + \eta^2\cdot\nabla_{\boldsymbol{B}_t}(\nabla_{\boldsymbol{B}_t})^\top. \tag{148}$$

Note that we can control the terms $\boldsymbol{B}_t(\nabla_{\boldsymbol{B}_t})^\top$ and $\nabla_{\boldsymbol{B}_t}\boldsymbol{B}_t^\top$ via Lemma E.4. In this view, it remains to argue that the contribution of the term $\eta^2\cdot\nabla_{\boldsymbol{B}_t}(\nabla_{\boldsymbol{B}_t})^\top$ and of the projection step are of order $\eta\left\|\boldsymbol{E}^t\right\|_{op}$. For convenience of the upcoming lemmas we define the following quantity:

$$\widetilde{\nabla}_{\boldsymbol{B}_t} := 2\left(-\alpha(\alpha\boldsymbol{I} + \boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-2}\boldsymbol{B}_t + \alpha\text{Diag}\left((\alpha\boldsymbol{I} + \boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-2}(\boldsymbol{B}_t\boldsymbol{B}_t^\top)\right)\boldsymbol{B}_t\right). \tag{149}$$

Lemma E.5. *Assume that (88) holds, and let $\nabla_{\boldsymbol{B}_t}$ be given by (86) with $\eta \le C/\sqrt{d}$. Then, for all sufficiently large n, with probability $1 - 1/d^2$, jointly for all $t \ge 0$:*

$$\eta^2 \left\|\nabla_{\boldsymbol{B}_t}(\nabla_{\boldsymbol{B}_t})^\top\right\|_{op} \le \eta \left\|\boldsymbol{E}^t\right\|_{op}.$$

Proof of Lemma E.5. We start by showing that

$$\|\widetilde{\nabla}_{\boldsymbol{B}_t}\|_{op} \le C(\|\boldsymbol{X}_t\|_{op} + \|\boldsymbol{Z}_t\|_{op}). \tag{150}$$

Recall that $\|\boldsymbol{B}_t\|_{op}, \|(\alpha\boldsymbol{I} + \boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-2}\|_{op} \le C$. Hence, the following chain of inequalities holds

$$\begin{aligned}
\|\widetilde{\nabla}_{\boldsymbol{B}_t}\|_{op} &\le \|\boldsymbol{B}_t\|_{op} \cdot \left\|-\alpha(\alpha\boldsymbol{I} + \boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-2} + \alpha\mathrm{Diag}\left((\alpha\boldsymbol{I} + \boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-2}(\boldsymbol{B}_t\boldsymbol{B}_t^\top)\right)\right\|_{op} \\
&\le C\left\|-\alpha(\alpha\boldsymbol{I} + \boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-2}(\boldsymbol{I} - \boldsymbol{B}_t\boldsymbol{B}_t^\top + \boldsymbol{B}_t\boldsymbol{B}_t^\top) + \alpha\mathrm{Diag}\left((\alpha\boldsymbol{I} + \boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-2}(\boldsymbol{B}_t\boldsymbol{B}_t^\top)\right)\right\|_{op} \\
&\le C\Big(\left\|(\alpha\boldsymbol{I} + \boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-2}(\boldsymbol{Z}_t + \boldsymbol{X}_t)\right\|_{op} \\
&\qquad + \left\|(\alpha\boldsymbol{I} + \boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-2}\boldsymbol{B}_t\boldsymbol{B}_t^\top - \mathrm{Diag}\left((\alpha\boldsymbol{I} + \boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-2}(\boldsymbol{B}_t\boldsymbol{B}_t^\top)\right)\right\|_{op}\Big) \\
&\le C\left(\|\boldsymbol{X}_t\|_{op} + \|\boldsymbol{Z}_t\|_{op} + \|F(\boldsymbol{X}_t + \boldsymbol{Z}_t) - \mathrm{Diag}(F(\boldsymbol{X}_t + \boldsymbol{Z}_t))\|_{op}\right),
\end{aligned} \tag{151}$$

where we recall the definition $F(x) = \frac{1+x}{(\gamma+x)^2}$, with $\gamma = 1 + \alpha$. By combining (144) and (145) (in the proof of Lemma E.4), we have

$$\|F(\boldsymbol{X}_t + \boldsymbol{Z}_t) - F(0)\boldsymbol{I}\|_{op} \le C(\|\boldsymbol{X}_t\|_{op} + \|\boldsymbol{Z}_t\|_{op}),$$

As $\|\mathrm{Diag}(\boldsymbol{M})\|_{op} \le C\|\boldsymbol{M}\|_{op}$ for any matrix $\boldsymbol{M}$, we also have that

$$\|\mathrm{Diag}(F(\boldsymbol{X}_t + \boldsymbol{Z}_t)) - F(0)\boldsymbol{I}\|_{op} \le C(\|\boldsymbol{X}_t\|_{op} + \|\boldsymbol{Z}_t\|_{op}).$$

Hence,

$$\|F(\boldsymbol{X}_t + \boldsymbol{Z}_t) - \mathrm{Diag}(F(\boldsymbol{X}_t + \boldsymbol{Z}_t))\|_{op} \le C(\|\boldsymbol{X}_t\|_{op} + \|\boldsymbol{Z}_t\|_{op}),$$

which finishes the proof of (150).

At this point, recall from (124) and (149) that

$$\left\|\nabla_{\boldsymbol{B}_t} - \widetilde{\nabla}_{\boldsymbol{B}_t}\right\|_{op} \le \left\|\boldsymbol{E}^t\right\|_{op}. \tag{152}$$

Thus,

$$\left\|\nabla_{\boldsymbol{B}_t}\nabla_{\boldsymbol{B}_t}^\top\right\|_{op} \le 2\left\|\widetilde{\nabla}_{\boldsymbol{B}_t}\boldsymbol{E}^t\right\|_{op} + \left\|\widetilde{\nabla}_{\boldsymbol{B}_t}(\widetilde{\nabla}_{\boldsymbol{B}_t})^\top\right\|_{op} + \left\|(\boldsymbol{E}^t)^2\right\|_{op}.$$

Recalling the previous bound on $\|\widetilde{\nabla}_{\boldsymbol{B}_t}\|_{op}$ in (150) and using the assumptions in (88), we get that

$$\left\|\widetilde{\nabla}_{\boldsymbol{B}_t}\boldsymbol{E}^t\right\|_{op},\ \left\|\boldsymbol{E}^t\right\|_{op}^2 \le C\|\boldsymbol{E}^t\|_{op},$$

and

$$\begin{aligned}
\eta^2\|\widetilde{\nabla}_{\boldsymbol{B}_t}\|_{op}^2 &\le C\eta(\|\boldsymbol{X}_t\|_{op}^2 + \|\boldsymbol{X}_t\|_{op}\|\boldsymbol{Z}_t\|_{op}) + C\eta^2\|\boldsymbol{Z}_t\|_{op}^2 \\
&\le C\eta\left(\frac{1}{\sqrt{d}}\|\boldsymbol{Z}_t\|_{op} + \|\boldsymbol{X}_t\|_{op}^2 + \|\boldsymbol{X}_t\|_{op}\|\boldsymbol{Z}_t\|_{op}\right) \le C\eta\left\|\boldsymbol{E}^t\right\|_{op},
\end{aligned} \tag{153}$$

where we have also used that $\eta \le C/\sqrt{d}$. This concludes the proof. □

The next lemma controls the contribution of the projection step.

Lemma E.6 (Projection step)**.** *Assume that* (88) *holds and* $\eta \le C/\sqrt{d}$*. Then, for all sufficiently large* n*, with probability* $1 - 1/d^2$*, jointly for all* $t \ge 0$*:*

$$\|\mathrm{proj}(\boldsymbol{B}_t') - \boldsymbol{B}_t'\|_{op} \le \eta \left\|\boldsymbol{E}^t\right\|_{op},$$

which implies that, by differentiability of the bilinear form,

$$\|\mathrm{proj}(\boldsymbol{B}_t')\mathrm{proj}(\boldsymbol{B}_t')^\top - \boldsymbol{B}_t'(\boldsymbol{B}_t')^\top\|_{op} \le \eta \left\|\boldsymbol{E}^t\right\|_{op}.$$

Proof of Lemma E.6. Recall that the objective (84) does not depend on the norm of $\{\boldsymbol{b}_i\}_{i=1}^n$, hence $(\nabla_{\boldsymbol{B}_t})_{i,:}$ is orthogonal to $(\boldsymbol{B}_t)_{i,:}$, which implies that

$$\mathrm{proj}_i(\boldsymbol{B}_t') = \frac{(\boldsymbol{B}_t)_{i,:} - \eta(\nabla_{\boldsymbol{B}_t})_{i,:}}{\sqrt{1+\eta^2\|(\nabla_{\boldsymbol{B}_t})_{i,:}\|^2}}.$$

Let us define

$$\boldsymbol{D}_t := \mathrm{Diag}\left(\frac{1}{\sqrt{1+\eta^2\|(\nabla_{\boldsymbol{B}_t})_{1,:}\|^2}}, \dots, \frac{1}{\sqrt{1+\eta^2\|(\nabla_{\boldsymbol{B}_t})_{n,:}\|^2}}\right).$$

Then, we obtain the following compact form:

$$\mathrm{proj}(\boldsymbol{B}_t') = \boldsymbol{D}_t(\boldsymbol{B}_t - \eta\nabla_{\boldsymbol{B}_t}) = \boldsymbol{D}_t\boldsymbol{B}_t'.$$

In this view, it remains to bound $\|\boldsymbol{D}_t - \boldsymbol{I}\|_{op}$. In more details, by (150) and (152), we have

$$\|\nabla_{\boldsymbol{B}_t}\|_{op} \le \|\widetilde{\nabla}_{\boldsymbol{B}_t}\|_{op} + \|\boldsymbol{E}^t\|_{op} \le C(\|\boldsymbol{X}_t\|_{op} + \|\boldsymbol{Z}_t\|_{op} + \|\boldsymbol{E}^t\|_{op}) \le C',$$

where $C' > 0$ is a universal constant (independent of C_X, n, d). Hence, by recalling that $\|\boldsymbol{B}_t\|_{op} \le C$ by assumption (88), we have

$$\|\mathrm{proj}(\boldsymbol{B}_t') - \boldsymbol{B}_t'\|_{op} = \|(\boldsymbol{D}_t - \boldsymbol{I})(\boldsymbol{B}_t - \eta\nabla_{\boldsymbol{B}_t})\|_{op} \le C\left\|\boldsymbol{D}_t - \boldsymbol{I}\right\|_{op}.$$

Note that function $1/\sqrt{1+x}$ is differentiable at 0, hence, we have that for small enough η (which follows from $\eta \le C/\sqrt{d}$):

$$\left|\frac{1}{\sqrt{1+\eta^2\|(\nabla_{\boldsymbol{B}_t})_{i,:}\|^2}} - 1\right| \le C\eta^2\|(\nabla_{\boldsymbol{B}_t})_{i,:}\|^2.$$

In this view, we have

$$\|\boldsymbol{D}_t - \boldsymbol{I}\|_{op} \le C\eta^2\|\nabla_{\boldsymbol{B}_t}\|_{op}^2 \le C\eta^2\|\widetilde{\nabla}_{\boldsymbol{B}_t}\|_{op}^2 + C\eta^2\|\widetilde{\nabla}_{\boldsymbol{B}_t}\|\|\boldsymbol{E}^t\|_{op} + C\eta^2\|\boldsymbol{E}^t\|^2.$$

Inspecting each term one by one and applying (150) in conjunction with $\eta \le C/\sqrt{d}$ gives that

$$\begin{aligned}
&\eta^2\left\|\boldsymbol{E}^t\right\|_{op}^2 \le C\eta\left\|\boldsymbol{E}^t\right\|_{op},\\
&\eta^2\|\widetilde{\nabla}_{\boldsymbol{B}_t}\|\|\boldsymbol{E}^t\|_{op} \le C\eta\|\boldsymbol{E}^t\|_{op},\\
&\eta^2\|\widetilde{\nabla}_{\boldsymbol{B}_t}\|_{op}^2 \le C\eta\left\|\boldsymbol{E}^t\right\|_{op},
\end{aligned}$$

where in the last step we have used (153). This concludes the proof. □

In this view, using (148) and Lemmas E.4, E.5 and E.6, we obtain

$$\begin{aligned}
\boldsymbol{I} + \boldsymbol{Z}_{t+1} + \boldsymbol{X}_{t+1} = \boldsymbol{B}_{t+1}\boldsymbol{B}_{t+1}^\top = \boldsymbol{I} + \boldsymbol{Z}_t + \boldsymbol{X}_t + 4\eta\alpha F(\boldsymbol{Z}_t) - 2\eta\alpha\mathrm{Diag}(F(\boldsymbol{Z}_t))(\boldsymbol{I} + \boldsymbol{Z}_t)\\
- 2\eta\alpha(\boldsymbol{I} + \boldsymbol{Z}_t)\mathrm{Diag}(F(\boldsymbol{Z}_t)) - \frac{8\alpha\eta}{\gamma^3}\boldsymbol{X}_t^O - \frac{4\alpha\eta}{\gamma^2}\boldsymbol{X}_t^D + \eta\boldsymbol{E}^t.
\end{aligned} \tag{154}$$

Furthermore, we have that

$$\begin{aligned}
\mathrm{Diag}(F(\boldsymbol{Z}_t))(\boldsymbol{I} + \boldsymbol{Z}_t) &= \left(\mathrm{Diag}(F(\boldsymbol{Z}_t) - F(0)\boldsymbol{I}) + F(0)\boldsymbol{I}\right)(\boldsymbol{I} + \boldsymbol{Z}_t)\\
&= \frac{1}{\gamma^2}(\boldsymbol{I} + \boldsymbol{Z}_t) + \left(\mathrm{Diag}(F(\boldsymbol{Z}_t) - F(0)\boldsymbol{I})\right)(\boldsymbol{I} + \boldsymbol{Z}_t)\\
&= \frac{1}{\gamma^2}(\boldsymbol{I} + \boldsymbol{Z}_t) + \left(\frac{1}{n}\mathrm{Tr}\left[F(\boldsymbol{Z}_t) - F(0)\boldsymbol{I}\right] + \boldsymbol{D}_t'\right)(\boldsymbol{I} + \boldsymbol{Z}_t),
\end{aligned} \tag{155}$$

where $\boldsymbol{D}_t'$ is a diagonal matrix such that, with probability at least $1-1/d^2$, its entries are upper bounded in modulus by $\frac{C\log d}{\sqrt{d}}\|\boldsymbol{Z}_t\|_{op}^{1/2}$. The last passage follows from Lemma G.2. Note that $\frac{1}{\gamma^2}(\boldsymbol{I}+\boldsymbol{Z}_t)=\frac{1}{n}\mathrm{Tr}\left[F(0)\boldsymbol{I}\right]$ and recall that $\|\boldsymbol{Z}_t\|_{op}\le C$. Hence, (155) implies that

$$\mathrm{Diag}(F(\boldsymbol{Z}_t))(\boldsymbol{I}+\boldsymbol{Z}_t)=\frac{1}{n}\mathrm{Tr}\left[F(\boldsymbol{Z}_t)\right](\boldsymbol{I}+\boldsymbol{Z}_t)+\boldsymbol{E}^t. \tag{156}$$

Similarly, we have that

$$(\boldsymbol{I}+\boldsymbol{Z}_t)\mathrm{Diag}(F(\boldsymbol{Z}_t))=\frac{1}{n}\mathrm{Tr}\left[F(\boldsymbol{Z}_t)\right](\boldsymbol{I}+\boldsymbol{Z}_t)+\boldsymbol{E}^t. \tag{157}$$

By combining (156)-(157) with (154) and using that $\boldsymbol{X}_t=\boldsymbol{X}_t^O+\boldsymbol{X}_t^D$, we get

$$\begin{aligned}\boldsymbol{Z}_{t+1}+\boldsymbol{X}_{t+1}=&\left(1-\frac{8\alpha}{\gamma^3}\eta\right)\boldsymbol{X}_t^O+\left(1-\frac{4\alpha}{\gamma^2}\eta\right)\boldsymbol{X}_t^D+\boldsymbol{Z}_t+4\eta\alpha F(\boldsymbol{Z}_t)\\&-4\eta\alpha\frac{1}{n}\mathrm{Tr}\left[F(\boldsymbol{Z}_t)\right](\boldsymbol{I}+\boldsymbol{Z}_t)+\eta\boldsymbol{E}^t.\end{aligned} \tag{158}$$

Hence, we can write the following system capturing the dynamics of the spectrum $\boldsymbol{Z}_t$ and of the errors $(\boldsymbol{X}_t^O,\boldsymbol{X}_t^D)$

$$\boldsymbol{Z}_{t+1}=\boldsymbol{Z}_t+4\eta\alpha F(\boldsymbol{Z}_t)-4\eta\alpha\frac{1}{n}\mathrm{Tr}\left[F(\boldsymbol{Z}_t)\right](\boldsymbol{I}+\boldsymbol{Z}_t), \tag{159}$$

$$\boldsymbol{X}_{t+1}^D=\left(1-\frac{4\alpha}{\gamma^2}\eta\right)\boldsymbol{X}_t^D+\eta\boldsymbol{E}^t, \tag{160}$$

$$\boldsymbol{X}_{t+1}^O=\left(1-\frac{8\alpha}{\gamma^3}\eta\right)\boldsymbol{X}_t^O+\eta\boldsymbol{E}^t. \tag{161}$$

Here, the operator norm of $\boldsymbol{E}^t$ is upper bounded as in (90), where we recall that the constant C is uniformly bounded in t.

In the view of (159), one can readily see that the updates on the spectrum of $\boldsymbol{Z}_t$ follow the one described in Lemma G.3 and, thus, converges exponentially. This means that the set of assumptions on $\boldsymbol{Z}_t$ in (88) is satisfied by suitably picking C.

Now it only remains to take care of $\boldsymbol{X}_t$. If we write $x_t^D=\left\|\boldsymbol{X}_t^D\right\|_{op}, x_t^O=\left\|\boldsymbol{X}_t^O\right\|_{op}, z_t=\|\boldsymbol{Z}_t\|_{op}^{1/2}$, then recalling the definition of $\boldsymbol{E}_t$ in (90), (160), (161) we have that

$$x_{t+1}^D\le\left(1-\frac{4\alpha}{\gamma^2}\eta\right)x_t^D+\eta C_D\left(\frac{\mathrm{poly}(\log d)}{\sqrt{d}}\cdot z_t+(x_t^D+x_t^O)^2+(x_t^D+x_t^O)z_t\right) \tag{162}$$

$$x_{t+1}^O\le\left(1-\frac{8\alpha}{\gamma^3}\eta\right)x_t^O+\eta C_O\left(\frac{\mathrm{poly}(\log d)}{\sqrt{d}}\cdot z_t+(x_t^D+x_t^O)^2+(x_t^D+x_t^O)z_t\right). \tag{163}$$

Since both of these recursive bounds are monotone in x_t^D, x_t^O, we can dominate them as follows. If we recursively define x_t by

$$x_{t+1}=\left(1-\eta\min\left\{\frac{4\alpha}{\gamma^2},\frac{8\alpha}{\gamma^3}\right\}\right)x_t+\eta\max\{C_D,C_O\}\left(\frac{\mathrm{poly}(\log d)}{\sqrt{d}}\cdot z_t+(x_t+x_t)^2+(x_t+x_t)z_t\right), \tag{164}$$

then by monotonicity $\max\{x_t^D,x_t^O\}\le x_t$. Thus, we only need to analyse the recursion (164), which we do in the following lemma. Note that the condition $z_t\le Ce^{-ct\eta}$ required by Lemma E.7 holds by (88).

Lemma E.7 (Error decay)**.** *Let $\{z_t\}_{t=0}^\infty$ be a non-negative exponentially decaying sequence, i.e., $z_t\le C_z e^{-\eta c_z t}$, and consider a non-negative sequence $\{x_t\}_{t=0}^\infty$ such that at each time-step t the following condition holds for $\eta=\Theta(1/\sqrt{d})$ and sufficiently large d:*

$$x_{t+1}=(1-\eta c_1)x_t+\eta C_2\cdot z_t\cdot x_t+\eta C_3 x_t^2+\eta C_4\cdot\frac{\mathrm{poly}(\log d)}{\sqrt{d}}\cdot z_t, \tag{165}$$

with $x_0=0$. Then, the following holds

$$x_t\le C\frac{\mathrm{poly}(\log d)}{\sqrt{d}}\cdot Te^{-cT}, \tag{166}$$

where $T=t\eta$.

Proof of Lemma E.7. We proceed in two parts. In the first part, we show that our recursion does not blow up in $t = K/\eta$ steps. In the second part, $z_t \le C_z \exp(-c_z K)$ will be small, which allows us to deduce (166).

Error does not blow up in finite time. Let $t = K/\eta$ where K is such that $K/\eta \in \mathbb{N}$. We start by analysing the simpler recursion

$$x_{t+1} = (1 - \eta c_1)x_t + \eta C_2 \cdot z_t \cdot x_t + \eta C_4 \cdot \frac{\mathrm{poly}(\log d)}{\sqrt{d}} \cdot z_t.$$

By hypothesis, $z_t \le C_z$. Hence, we arrive to

$$x_{t+1} = (1 - \eta c_1)x_t + \eta C_2 C_z \cdot x_t + \eta C_4 C_z \frac{\mathrm{poly}(\log d)}{\sqrt{d}}.$$

Writing $C_5 = C_2 C_z - c_1$, unrolling the recursion on the RHS and using $x_0 = 0$ gives

$$\begin{aligned}
x_{t+1} &= \eta C_4 C_z \frac{\mathrm{poly}(\log d)}{\sqrt{d}} \sum_{j=0}^{t} (1 + \eta C_5)^j \\
&\le \eta C_4 C_z \frac{\mathrm{poly}(\log d)}{\sqrt{d}} \sum_{j=0}^{K/\eta} e^{\eta C_5 j} \\
&= \eta C_4 C_z \frac{\mathrm{poly}(\log d)}{\sqrt{d}} \cdot e^{C_5 K} \sum_{j=0}^{K/\eta} e^{-C_5 \eta (t-j)} \\
&\le \eta C_4 C_z \frac{\mathrm{poly}(\log d)}{\sqrt{d}} \cdot \frac{e^{C_5 K}}{1 - e^{-\eta C_5}},
\end{aligned}$$

where the inequality holds for $t \le K/\eta$ and we have used $1 + x \le e^x$. For small enough η, we have that

$$\frac{\eta}{1 - e^{-C_5 \eta}} \le \frac{2}{C_5},$$

hence, for all $t \le K/\eta$,

$$x_{t+1} \le 2 \frac{\mathrm{poly}(\log d)}{\sqrt{d}} \frac{C_4 C_z}{C_5} \exp(C_5 K). \tag{167}$$

Let us now go back to our original recursion (165), which contains the term x_t^2. We claim that this recursion satisfies a bound like (167). Assume by contradiction that it exceeds the bound

$$x_t \le 4 \frac{\mathrm{poly}(\log d)}{\sqrt{d}} \frac{C_4 C_z}{C_5} \exp(C_5 K) \tag{168}$$

for the first time at step t'. Then, for all $t < t'$, (168) holds. Noting that $x_t^2 \le 4 \frac{\mathrm{poly}(\log d)}{\sqrt{d}} \frac{C_4 C_z}{C_5} \exp(C_5 K) x_t$ we define $C_5' = C_2 C_z + 4 C_3 \frac{\mathrm{poly}(\log d)}{\sqrt{d}} \frac{C_4 C_z}{C_5} \exp(C_5 K) - c_1$. By unrolling the recursion exactly as before, we obtain

$$x_{t+1} \le 2 \frac{\mathrm{poly}(\log d)}{\sqrt{d}} \frac{C_4 C_z}{C_5'} \exp(C_5' K) \le 3 \frac{\mathrm{poly}(\log d)}{\sqrt{d}} \frac{C_4 C_z}{C_5} \exp(C_5 K), \tag{169}$$

for d large enough. Here, the second inequality follows for large d, since it is clear from the definitions that $|C_5 - C_5'|$ vanishes for large d. This shows that we cannot violate (168), thus (169) holds for all $t \le K/\eta$.

Convergence of errors x_t to zero. We now choose K large enough so that

$$z_t = C_z e^{-\eta c_z t} < \frac{c_1}{2 C_2}, \quad \forall t \ge K/\eta.$$

Hence, the term corresponding to $\eta C_2 z_t x_t$ can be pushed inside the $(1-\eta c_1)x_t$ term. Consequently, we can equivalently study the following dynamics

$$x_{t+1} = (1-\eta c_1')x_t + \eta C_3 x_t^2 + \eta C_4 C_z \frac{\text{poly}(\log d)}{\sqrt{d}} e^{-\eta c_z t}, \tag{170}$$

where $c_1' = c_1/2$. Here, we initialize again at $t=0$, but now starting at

$$x_0 = C_6 \frac{\text{poly}(\log d)}{\sqrt{d}},$$

where $C_6 = 4\frac{C_4 C_z}{C_5}\exp(C_5 K)$, corresponding to the bound in (168). Rearranging we have

$$x_{t+1} = x_t + \eta\left(-c_1' x_t + C_3 x_t^2 + C_4 C_z \frac{\text{poly}(\log d)}{\sqrt{d}} e^{-\eta c_z t}\right). \tag{171}$$

As the last term inside the brackets vanishes when $d \to \infty$, we have two roots of the polynomial inside the brackets, corresponding to the fixed points of the iteration. The left root r_l scales as

$$r_l \le C_l \frac{\text{poly}(\log d)}{\sqrt{d}} e^{-\eta c_z t},$$

and the right root r_r as

$$r_r \ge \frac{c_1'}{C_3} - C\frac{\text{poly}(\log d)}{\sqrt{d}} e^{-\eta c_z t}.$$

In addition, it is easy to see that both roots are non-negative.

Next, we prove that $x_t \le C\frac{\text{poly}(\log d)}{\sqrt{d}}$ for all t. We will show this by contradiction. At initialization we have

$$x_0 = C_6 \frac{\text{poly}(\log d)}{\sqrt{d}}.$$

Choose A, B as follows:

$$A := \max\{C_l, C_6\}, \quad B = C_7 A.$$

We first note that, for small enough η and large enough d, we can choose C_7 such that $x_{\tilde{t}} \le A\frac{\text{poly}(\log d)}{\sqrt{d}}$ implies $x_{\tilde{t}+1} \le B\frac{\text{poly}(\log d)}{\sqrt{d}}$. We now show that $x_t \le B\frac{\text{poly}(\log d)}{\sqrt{d}}$ for all t. To do so, assume by contradiction that $x_{t+1} > B\frac{\text{poly}(\log d)}{\sqrt{d}}$. Then $x_t \in [A\frac{\text{poly}(\log d)}{\sqrt{d}}, B\frac{\text{poly}(\log d)}{\sqrt{d}}] \subseteq [r_l, r_r]$, thus

$$-c_1' x_t + C_3 x_t^2 + C_4 C_z \frac{\text{poly}(\log d)}{\sqrt{d}} e^{-\eta c_z t} < 0.$$

Hence, from (171) it follows that

$$x_{t+1} \le x_t \le B\frac{\text{poly}(\log d)}{\sqrt{d}},$$

which gives us the desired contradiction.

Thus, for all t,

$$x_t^2 \le B\frac{\text{poly}(\log d)}{\sqrt{d}} x_t.$$

This allows us to push the second term in (170) into the first one (for d large enough), which reduces the recursion to

$$x_{t+1} = (1-\eta c_1'')\, x_t + \eta C_4 C_z \frac{\text{poly}(\log d)}{\sqrt{d}} e^{-\eta c_z t},$$

where $c_1'' \geq c_1'/2$. By unrolling this last recursion and using $x_0 = C_6 \frac{\text{poly}(\log d)}{\sqrt{d}}$, we have that, for $t \geq 1$,

$$x_t = C_6 \frac{\text{poly}(\log d)}{\sqrt{d}} (1 - \eta c_1'')^t + \eta C_4 C_z \frac{\text{poly}(\log d)}{\sqrt{d}} \sum_{\ell=1}^{t} (1 - \eta c_1'')^{t-\ell} e^{-\eta c_z \ell} \tag{172}$$

$$\leq C_6 \frac{\text{poly}(\log d)}{\sqrt{d}} \exp(-\eta c_1'' t) + \eta C_4 C_z \frac{\text{poly}(\log d)}{\sqrt{d}} \sum_{\ell=1}^{t} e^{-\eta(c_z \ell + c_1''(t-\ell))}, \tag{173}$$

where the inequality follows from $1 - x \leq e^{-x}$. Since the term in the exponents of the sum is a linear function in ℓ, its maximum value is attained in the endpoints. Thus,

$$x_t \leq C_6 \frac{\text{poly}(\log d)}{\sqrt{d}} \exp(-\eta c_1'' t) + \eta C_4 C_z \frac{\text{poly}(\log d)}{\sqrt{d}} t \max\{e^{-\eta c_z t}, e^{-\eta c_1'' t}\},$$

which implies (166). □

By Lemma E.7 we know that

$$\|\boldsymbol{X}_t\|_{op} \leq \frac{C}{\sqrt{d}} \cdot T e^{-cT},$$

where C is independent of C_X by definition. Hence, we can pick C_X such that, for sufficiently large d, the assumptions on $\boldsymbol{X}_t$ in (88) are satisfied. With this in mind, we can use Lemma G.3 to bound the dynamics involving $\boldsymbol{Z}_t$ and Lemma E.7 to claim that the error $\boldsymbol{X}_t$ vanishes at least geometrically fast. This concludes the proof of Theorem E.1.

F. Discussion of Isotropic Gaussian Results

Degenerate isotropic Gaussian data. All the arguments of Section 4.1 directly apply for $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$, the only differences being the scaling of the term $\text{Tr}\,[\boldsymbol{B}\boldsymbol{A}]$ (which is additionally multiplied by σ) and the constant variance term σ^2 (in place of 1) in (6). Our results can be also easily extended to the case of degenerate isotropic Gaussian data, i.e., $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ with $\lambda_i(\boldsymbol{\Sigma}) = \sigma^2$ for $i \leq d - k$ and $\lambda_i(\boldsymbol{\Sigma}) = 0$ for $i > d - k$, where $\lambda_i(\boldsymbol{\Sigma})$ stands for the i-th eigenvalue of $\boldsymbol{\Sigma}$ in non-increasing order. In fact, by the rotational invariance of the Gaussian distribution, we can assume without loss of generality that $\boldsymbol{x} = [x_1, \cdots, x_{d-k}, 0, \cdots, 0]$, where $(x_i) \sim_{\text{i.i.d.}} \mathcal{N}(0, \sigma^2)$. Hence, by considering $\boldsymbol{A} \in \mathbb{R}^{(d-k)\times n}$ and $\boldsymbol{B} \in \mathbb{R}^{n \times (d-k)}$ and substituting d with $d - k$ where suitable, analogous results follow.

Scaling of the learning rate. Theorem 4.6 is stated for $\eta = \Theta(1/\sqrt{d})$, as this corresponds to the biggest learning rate for which our argument works (thus requiring the least amount of steps for convergence). The same result can be proved for $\eta = \Theta(d^{-\kappa})$ with $\kappa \geq 1/2$. The only piece of the proof affected by this change is the third part of Lemma G.2 (in particular, the chain of inequalities (187)), which continues to hold as long as η is polynomial in d^{-1}.

Assumptions on compression rate r. We expect an analog of Theorem 4.5 (see Theorem D.1 for the formal statement) to hold for $r > 1$, as long as d is *sufficiently large*. In fact, for a fixed d, it appears to be difficult to even characterize the global minimizer: the choice (12) approaches the lower bound $\text{LB}_{r>1}(\boldsymbol{I})$ only as $d \to \infty$, see Proposition 4.4. We also expect Theorem 4.6 to hold for $r \geq 1$. Here, an additional challenge is that the minimizer has non-zero off-diagonal entries. In combination with the lack of an exact characterization of the minimizer, this leads to an additional error term that would be difficult to control with the current tools. At the same time, the restriction $r < 1$ is likely to be an artifact of the proof as experimentally (see, for instance, Figure 4) the algorithm still converges to the global optimum for $r \geq 1$.

Gaussian initialization in Theorem 4.6. The Gaussian initialization ensures that, with high probability, the off-diagonal entries of $\boldsymbol{B}(t)\boldsymbol{B}(t)^\top$ are small. This allows us to approximate higher-order Hadamard powers of $\boldsymbol{B}(t)\boldsymbol{B}(t)^\top$ with $\boldsymbol{I}$. However, in experiments the Gaussian assumption seems to be unnecessary, and we expect the convergence result to hold for all (non-degenerate) initializations.

G. Auxiliary Results

Lemma G.1. *For any $\boldsymbol{R} \in \mathbb{R}^{n\times n}$ the following holds*

$$\|\boldsymbol{R} - \mathrm{diag}(\boldsymbol{R})\|_{op} \le C\,\|\boldsymbol{R}\|_{op}\,.$$

Proof. By definition of the operator norm we have that

$$\|\boldsymbol{R}\|_{op} = \sup_{\|\boldsymbol{x}\|_2=1} \|\boldsymbol{R}\boldsymbol{x}\|_2\,.$$

Note that by Cauchy-Schwarz, the following holds for $\|\boldsymbol{y}\|_2 = 1$:

$$\langle \boldsymbol{y}, \boldsymbol{R}\boldsymbol{x}\rangle \le \|\boldsymbol{R}\boldsymbol{x}\|_2\,,$$

and the inequality is met when $\boldsymbol{y}$ is aligned with $\boldsymbol{R}\boldsymbol{x}$. Hence, we get

$$\sup_{\|\boldsymbol{y}\|_2=1} \langle \boldsymbol{y}, \boldsymbol{R}\boldsymbol{x}\rangle = \|\boldsymbol{R}\boldsymbol{x}\|_2\,,$$

and, thus, the operator norm can be rewritten as

$$\|\boldsymbol{R}\|_{op} = \sup_{\|\boldsymbol{x}\|_2=1} \|\boldsymbol{R}\boldsymbol{x}\|_2 = \sup_{\|\boldsymbol{x}\|_2=\|\boldsymbol{y}\|_2=1} \langle \boldsymbol{y}, \boldsymbol{R}\boldsymbol{x}\rangle.$$

Note also that $\|\mathrm{diag}(\boldsymbol{R})\|_{op}$ is equal to the maximal diagonal element (in absolute value). Hence, by letting $\boldsymbol{e}_i$ be the i-th element of the canonical basis, we get

$$\|\mathrm{diag}(\boldsymbol{R})\|_{op} = \sup_i |\boldsymbol{R}_{i,i}| \le \sup_i |\langle \boldsymbol{e}_i, \boldsymbol{R}\boldsymbol{e}_i\rangle| \le \sup_{\|\boldsymbol{x}\|_2=\|\boldsymbol{y}\|_2=1} \langle \boldsymbol{y}, \boldsymbol{R}\boldsymbol{x}\rangle = \|\boldsymbol{R}\|_{op}\,.$$

In this view, an application of triangle inequality, i.e.,

$$\|\boldsymbol{R} - \mathrm{diag}(\boldsymbol{R})\|_{op} \le \|\boldsymbol{R}\|_{op} + \|\mathrm{diag}(\boldsymbol{R})\|_{op} \le 2\,\|\boldsymbol{R}\|_{op}\,,$$

finishes the proof. □

Lemma G.2. *Consider the matrix $\boldsymbol{A}_t = \boldsymbol{U}\boldsymbol{\Lambda}_t\boldsymbol{U}^\top$, where the matrix $\boldsymbol{U}$ is distributed according to the Haar measure and it is independent from the diagonal matrix $\boldsymbol{\Lambda}_t$. Further, assume that all the diagonal entries of $\boldsymbol{\Lambda}_t$ are bounded in absolute value by a constant. Then, the following results hold.*

1. *We have that, with probability at least $1 - 1/d^2$,*

$$\max_{i\neq j} |(\boldsymbol{A}_t)_{i,j}| \le c\sqrt{\frac{\log d}{d}}, \tag{174}$$

   for some absolute constant $c > 0$.

2. *Let $\boldsymbol{D}_t = diag(\boldsymbol{A}_t)$. Then,*

$$\boldsymbol{D}_t = \alpha\boldsymbol{I} + \boldsymbol{D}'_t,$$

   where

$$\alpha = \frac{1}{n}\mathrm{Tr}(\boldsymbol{\Lambda}_t),$$

   and $\boldsymbol{D}'_t$ is a diagonal matrix such that, with probability at least $1 - 1/d^2$,

$$\max_{i\in[n]} |(\boldsymbol{D}'_t)_{i,i}| \le c\frac{\log d}{\sqrt{d}}. \tag{175}$$

3. *Assume that, for all* $t \in \mathbb{N}$,

$$\|\boldsymbol{\Lambda}_t\|_{op} \leq Ce^{-c\eta t}, \tag{176}$$

where $c, C > 0$ *are absolute constants and* $\eta = \Theta(1/\sqrt{d})$. *Then, with probability at least* $1 - 1/d^2$,

$$\sup_{t\geq 0} \max_{i\neq j} |(\boldsymbol{A}_t)_{i,j}| \leq c\sqrt{\frac{\log d}{d}}, \tag{177}$$

$$\sup_{t\geq 0} \max_{i\in[n]} |(\boldsymbol{D}'_t)_{i,i}| \leq c\frac{\log d}{\sqrt{d}}. \tag{178}$$

Proof. We start by proving (174). Consider the metric measure space $(\mathbb{SO}(d), \|\cdot\|_F, \mathbb{P})$. Here, $\mathbb{SO}(d)$ denotes the special orthogonal group containing all $d \times d$ orthogonal matrices with determinant 1 (i.e., all rotation matrices), and $\mathbb{P}$ is the uniform probability measure on $\mathbb{SO}(d)$, i.e., the Haar measure. Given a diagonal matrix $\boldsymbol{\Lambda}_t$ and two indices $i, j \in [d]$, define $f : \mathbb{SO}(d) \to \mathbb{R}$ as

$$f(\boldsymbol{M}) = (\boldsymbol{M}\boldsymbol{\Lambda}_t\boldsymbol{M}^\top)_{i,j}. \tag{179}$$

Note that

$$\begin{aligned}
|f(\boldsymbol{M}) - f(\boldsymbol{M}')| &= |(\boldsymbol{M}\boldsymbol{\Lambda}_t\boldsymbol{M}^\top)_{i,j} - (\boldsymbol{M}'\boldsymbol{\Lambda}_t(\boldsymbol{M}')^\top)_{i,j}| \\
&\leq |(\boldsymbol{M}\boldsymbol{\Lambda}_t\boldsymbol{M}^\top)_{i,j} - (\boldsymbol{M}'\boldsymbol{\Lambda}_t\boldsymbol{M}^\top)_{i,j}| + |(\boldsymbol{M}'\boldsymbol{\Lambda}_t\boldsymbol{M}^\top)_{i,j} - (\boldsymbol{M}'\boldsymbol{\Lambda}_t(\boldsymbol{M}')^\top)_{i,j}| \\
&\leq |((\boldsymbol{M} - \boldsymbol{M}')\boldsymbol{\Lambda}_t\boldsymbol{M}^\top)_{i,j}| + |(\boldsymbol{M}'\boldsymbol{\Lambda}_t(\boldsymbol{M} - \boldsymbol{M}')^\top)_{i,j}| \\
&\leq \|(\boldsymbol{M} - \boldsymbol{M}')\boldsymbol{\Lambda}_t\boldsymbol{M}^\top\|_F + \|\boldsymbol{M}'\boldsymbol{\Lambda}_t(\boldsymbol{M} - \boldsymbol{M}')^\top\|_F \\
&\leq 2\|\boldsymbol{M} - \boldsymbol{M}'\|_F\|\boldsymbol{\Lambda}_t\|_{op}\|\boldsymbol{M}\|_{op} \leq 2\|\boldsymbol{M} - \boldsymbol{M}'\|_F\|\boldsymbol{\Lambda}_t\|_{op},
\end{aligned} \tag{180}$$

where in the fourth inequality we use that, for any two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $\|\boldsymbol{A}\boldsymbol{B}\|_F \leq \|\boldsymbol{A}\|_{op}\|\boldsymbol{B}\|_F$, and in the fifth inequality we use that $\|\boldsymbol{M}\|_{op} = 1$ as $\boldsymbol{M} \in \mathbb{SO}(d)$. Hence, f has Lipschitz constant upper bounded by $2\|\boldsymbol{\Lambda}_t\|_{op}$ and an application of Theorem 5.2.7 of (Vershynin, 2018) gives that

$$\mathbb{P}(|f(\boldsymbol{U}) - \mathbb{E}[f(\boldsymbol{U})]| \geq u) \leq 2\exp\left(-c_1\frac{d\boldsymbol{u}^2}{2\|\boldsymbol{\Lambda}_t\|_{op}}\right), \tag{181}$$

where c_1 is a universal constant.

Let $\boldsymbol{u}_i$ denote the i-th row of $\boldsymbol{U}$. Then,

$$f(\boldsymbol{U}) = \langle \boldsymbol{u}_i, \boldsymbol{\Lambda}_t\boldsymbol{u}_j\rangle. \tag{182}$$

Suppose that $i \neq j$. Since $\boldsymbol{U}$ is distributed according to the Haar measure, $\boldsymbol{u}_i$ is uniform on the unit sphere and $\boldsymbol{u}_j$ is uniformly distributed on the unit sphere in the orthogonal complement of $\boldsymbol{u}_i$ (see Section 1.2 of (Meckes, 2019)). Thus, $(\boldsymbol{u}_i, \boldsymbol{u}_j)$ has the same distribution as $(-\boldsymbol{u}_i, \boldsymbol{u}_j)$, which implies that, whenever $i \neq j$

$$\mathbb{E}[f(\boldsymbol{U})] = 0. \tag{183}$$

By combining (181)-(183) with a union bound over i, j, we have that

$$\mathbb{P}(\max_{i\neq j} |(\boldsymbol{U}\boldsymbol{\Lambda}_t\boldsymbol{U}^\top)_{i,j}| \geq u) \leq 2d^2\exp\left(-c_1\frac{du^2}{2\|\boldsymbol{\Lambda}_t\|_{op}}\right). \tag{184}$$

As $\|\boldsymbol{\Lambda}_t\|_{op}$ is upper bounded by a universal constant, the result (174) readily follows.

For the second part, note that

$$(\boldsymbol{D}_t)_{i,i} = \langle \boldsymbol{u}_i, \boldsymbol{\Lambda}_t\boldsymbol{u}_i\rangle. \tag{185}$$

Furthermore, the following chain of equalities hold

$$\mathbb{E}[(\boldsymbol{D}_t)_{i,i}] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[(\boldsymbol{D}_t)_{i,i}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (\boldsymbol{D}_t)_{i,i}\right] = \frac{1}{n}\mathrm{Tr}(\boldsymbol{D}_t), \tag{186}$$

where the first equality uses that the $\boldsymbol{u}_i$'s have the same (marginal) distribution, and the last term does not contain an expectation since $\mathrm{Tr}(\boldsymbol{D}_t) = \mathrm{Tr}(\boldsymbol{A}_t) = \sum_{i=1}^{d}(\boldsymbol{\Lambda}_t)_{i,i}$, which does not depend on $\boldsymbol{U}$. Therefore, by using (181) and by performing a union bound over $i \in [n]$, the result (175) follows.

For the third part, by performing a union bound over $t \geq 0$ in (184), we have that (177) holds with probability at least

$$\begin{aligned} 2\sum_{t=0}^{\infty}\exp\left(-c_1\frac{du^2}{2\|\boldsymbol{\Lambda}_t\|_{op}}\right) &\leq 2\sum_{t=0}^{\infty}\exp\left(-c_2\, d\, u^2\, e^{C\eta t}\right) \\ &\leq 2\sum_{t=0}^{\infty}\exp\left(-c_2\, d\, u^2\, e^{C\lfloor \eta t\rfloor}\right) \\ &\leq 2\left\lceil\frac{1}{\eta}\right\rceil\sum_{t=0}^{\infty}\exp\left(-c_2\, d\, u^2\, e^{Ct}\right) \\ &\leq C\sqrt{d}\sum_{t=0}^{\infty}\exp\left(-c_2\, d\, u^2\, e^{Ct}\right), \end{aligned} \tag{187}$$

where the first inequality follows from (176) and the last one from $\eta = \Theta(1/\sqrt{d})$. Choosing $u = c\frac{\log d}{\sqrt{d}}$ we can get that $b := \exp\left(-c_2\, d\, u^2\right) < 1$ and, hence, the following holds

$$\sum_{t=0}^{\infty}\exp\left(-c_2\, d\, u^2\right)^{e^{Ct}} \leq \sum_{t=0}^{\infty}\exp\left(-c_2\, d\, u^2\right)^{Ct+1} = \frac{b}{1-b^C} \leq \frac{1}{d^3},$$

where the first inequality uses that $e^t \geq 1 + t$ and the second inequality follows from the definition of b. This concludes the proof of (177). The proof of (178) uses an analogous union bound on $t \geq 0$. □

Lemma G.3. *Let $\lambda^0 = \{\lambda_1^0, \cdots, \lambda_n^0\}$ be a set of numbers in $\mathbb{R}$ such that*

$$\lambda^0_{min} := \min_{i\in[n]}\lambda_i^0 \geq \delta > 0, \quad \lambda^0_{max} := \max_{i\in[n]}\lambda_i^0 \leq M < +\infty, \quad \sum_{j=1}^{n}\lambda_j^0 = n.$$

Let the values $\{\lambda_i^t\}_{i=1}^n$ be updated according to the equation below

$$\lambda_i^{t+1} = \lambda_i^t + \eta\left(F(\lambda_i^t) - \lambda_i^t\cdot\frac{1}{n}\sum_{j=1}^{n}F(\lambda_j^t)\right) = G(\lambda_i^t, \lambda^t), \tag{188}$$

where $F(\cdot)$ is defined as per Lemma E.4, $\eta = \Theta\left(1/\sqrt{d}\right)$ and $\lambda^t := \{\lambda_1^t, \cdots, \lambda_n^t\}$. Then, for large enough d, we have

$$\left|\lambda_i^{t+1} - 1\right| \leq (1 - c\delta\cdot\eta)\left|\lambda_i^t - 1\right|$$

and thus after t iterations

$$\left|\lambda_i^t - 1\right| \leq \max\{(M-1), (1-\delta)\}\exp(-c\delta\cdot\eta t),$$

where $c, C > 0$ are constants.

Proof. We first show by induction that $\sum_{i=1}^{n}\lambda_i^t = n$ holds for all t. In fact,

$$\begin{aligned} \sum_{i=1}^{n}\lambda_i^{t+1} &= \sum_{i=1}^{n}\lambda_i + \eta\left(\sum_{i=1}^{n}F(\lambda_i^t) - \sum_{i=1}^{n}\lambda_i^t\cdot\frac{1}{n}\sum_{j=1}^{n}F(\lambda_j^t)\right) \\ &= n + \eta\left(\sum_{i=1}^{n}F(\lambda_i^t) - \sum_{j=1}^{n}F(\lambda_j^t)\right) = n. \end{aligned}$$

Now, we will show the convergence of λ_{min}^t and λ_{max}^t. To do so, we assume that $\lambda_{max}^t \leq M$ and $\lambda_{min}^t \geq \delta$ holds at time step t (we will verify this later). Define the function $g : \mathbb{R} \to \mathbb{R}$ as

$$g(x) := x + \eta \left(F(x) - x \cdot C \right). \tag{189}$$

By taking the derivative, we have that, for sufficiently large d,

$$g'(x) = 1 + \eta \left(F'(x) - C \right) > 0,$$

as $\|F'\|_\infty \leq C$. This implies that $g(\cdot)$ is a monotone increasing function, which gives that

$$\begin{aligned} \max_{i \in [n]} g(\lambda_i^t) &= g(\lambda_{max}^t), \\ \min_{i \in [n]} g(\lambda_i^t) &= g(\lambda_{min}^t). \end{aligned} \tag{190}$$

Note that the updates on λ_i^t in (188) have a common part for all $i \in [n]$, i.e.,

$$\left| \frac{1}{n} \sum_{j=1}^n F(\lambda_j^t) \right| \leq C,$$

where we used that $\|F\|_\infty \leq C$. In this view, by definition of g and (190), we have

$$\begin{aligned} \lambda_{max}^{t+1} &= G(\lambda_{max}^t, \lambda^t), \\ \lambda_{min}^{t+1} &= G(\lambda_{min}^t, \lambda^t), \end{aligned} \tag{191}$$

which means that the min/max value at the previous step are mapped to the min/max value at the next step of (188). Using that $\frac{1}{n} \sum_{i=1}^n \lambda_i^t = 1$ we can write

$$\begin{aligned} \lambda_i^{t+1} &= \lambda_i^t + \eta \left(\frac{1}{n} \sum_{j=1}^n \lambda_j^t \cdot F(\lambda_i^t) - \lambda_i^t \cdot \frac{1}{n} \sum_{j=1}^n F(\lambda_j^t) \right) \\ &= \lambda_i^t + \eta \left(\frac{1}{n} \sum_{j=1}^n \left[\frac{\lambda_j^t \lambda_i^t}{(\alpha + \lambda_i^t)^2} - \frac{\lambda_i^t \lambda_j^t}{(\alpha + \lambda_j^t)^2} \right] \right) \\ &= \lambda_i^t + \eta \left(\frac{1}{n} \sum_{j=1}^n \lambda_i^t \lambda_j^t \left(\frac{(2\alpha + \lambda_i^t + \lambda_j^t)(\lambda_j^t - \lambda_i^t)}{(\alpha + \lambda_i^t)^2 (\alpha + \lambda_j^t)^2} \right) \right). \end{aligned} \tag{192}$$

Recall that we assumed $\lambda_{max}^t \leq M$ and $\lambda_{min}^t \geq \delta$. In this view, we get the following bound

$$\lambda_{max}^t \lambda_j^t \left(\frac{(2\alpha + \lambda_{max}^t + \lambda_j^t)(\lambda_{max}^t - \lambda_j^t)}{(\alpha + \lambda_{max}^t)^2 (\alpha + \lambda_j^t)^2} \right) \geq (\lambda_{max}^t - \lambda_j^t) \cdot \frac{2\alpha\delta}{(\alpha + M)^4}, \tag{193}$$

which is justified as follows

$$\begin{aligned} \lambda_{max}^t \lambda_j^t \left(\frac{(2\alpha + \lambda_{max}^t + \lambda_j^t)(\lambda_{max}^t - \lambda_j^t)}{(\alpha + \lambda_{max}^t)^2 (\alpha + \lambda_j^t)^2} \right) &= (\lambda_{max}^t - \lambda_j^t) \cdot \left(\frac{(2\alpha + \lambda_{max}^t + \lambda_j^t) \lambda_{max}^t \lambda_j^t}{(\alpha + \lambda_{max}^t)^2 (\alpha + \lambda_j^t)^2} \right) \\ &\geq (\lambda_{max}^t - \lambda_j^t) \cdot \frac{2\alpha \cdot 1 \cdot \delta}{(\alpha + M)^2 (\alpha + M)^2}, \end{aligned}$$

where we used that $\lambda_{max}^t \geq 1$ since $\sum_{i=1}^n \lambda_i^t = n$. Hence, using the previous observation about mapping of extremes in (191) and the observation above, we get from (192) that

$$\lambda_{max}^{t+1} \leq \lambda_{max}^t - \eta \cdot \frac{1}{n} \sum_{j=1}^n \left[(\lambda_{max}^t - \lambda_j^t) \cdot \frac{2\alpha\delta}{(\alpha + M)^4} \right], \tag{194}$$

which leads to

$$\begin{aligned}\lambda_{max}^{t+1} - 1 &\le \lambda_{max}^t - 1 - \eta \cdot \frac{1}{n}\sum_{j=1}^n \left[(\lambda_{max}^t - \lambda_j^t) \cdot \frac{2\alpha\delta}{(\alpha+M)^4}\right] \\ &= \lambda_{max}^t - 1 - \eta \cdot \left[(\lambda_{max}^t - 1) \cdot \frac{2\alpha\delta}{(\alpha+M)^4}\right] \\ &= (\lambda_{max}^t - 1)\left(1 - \eta \cdot \frac{2\alpha\delta}{(\alpha+M)^4}\right) = (\lambda_{max}^t - 1)(1 - c\delta\cdot\eta),\end{aligned} \tag{195}$$

where we used that $\sum_{j=1}^n \lambda_j^t = n$ in the first equality. Hence, using that $\lambda_{max}^t \ge 1$ as $\sum_{j=1}^n \lambda_j^t = n$ we have

$$|\lambda_{max}^{t+1} - 1| = \lambda_{max}^{t+1} - 1 \le |\lambda_{max}^t - 1| \cdot (1 - c\delta\cdot\eta). \tag{196}$$

Similarly to the previous bound, we get that

$$\lambda_{min}^t \lambda_j^t \left(\frac{(2\alpha + \lambda_{min}^t + \lambda_j^t)(\lambda_{min}^t - \lambda_j^t)}{(\alpha+\lambda_{min}^t)^2(\alpha+\lambda_j^t)^2}\right) \le \lambda_j^t(\lambda_{min}^t - \lambda_j^t)\frac{2\alpha\delta}{(\alpha+M)^4},$$

since $\lambda_{min}^t \le \lambda_t$. Hence, using the previous observation about mapping of extremes in (191) and the observation above, we deduce from (192) that

$$\begin{aligned}\lambda_{min}^{t+1} - 1 &\ge (\lambda_{min}^t - 1) - \eta \cdot \frac{1}{n}\sum_{j=1}^n \left[\lambda_j^t(\lambda_{min}^t - \lambda_j^t)\frac{2\alpha\delta}{(\alpha+M)^4}\right] \\ &= (\lambda_{min}^t - 1) - \eta\cdot\lambda_{min}^t \cdot \frac{2\alpha\delta}{(\alpha+M)^4} + \eta\cdot\frac{2\alpha\delta}{(\alpha+M)^4}\cdot\frac{1}{n}\sum_{j=1}^t \left(\lambda_i^t\right)^2 \\ &\ge (\lambda_{min}^t - 1) - \eta\cdot(\lambda_{min}^t - 1)\cdot\frac{2\alpha\delta}{(\alpha+M)^4} \\ &= (\lambda_{min}^t - 1)\cdot(1 - c\delta\cdot\eta),\end{aligned} \tag{197}$$

where in the second inequality we used Jensen's inequality for x^2 as $\sum_{j=1}^n \lambda_j^t = n$. Hence, we get the following

$$|\lambda_{min}^{t+1} - 1| = 1 - \lambda_{min}^{t+1} \le |\lambda_{min}^t - 1| \cdot (1 - c\delta\cdot\eta), \tag{198}$$

since $\lambda_{min}^t \le 1$ as $\sum_{j=1}^n \lambda_j^t = n$.

In this view, the assumptions $\lambda_{max}^t \le M$ and $\lambda_{min}^t \ge \delta$ follow from (196) and (198) since the extremes are getting closer to one after each iteration. Recalling that by the assumption on initialization

$$\max_i |\lambda_i^0 - 1| \le \max\{(M-1), (1-\delta)\},$$

the claim follows. □

H. Proofs for General Covariance

Lemma H.1. *Assume that $\{\hat{\gamma}_i\}_{i\in[K]}, \{\hat{s}_i\}_{i\in[K]}$ minimize*

$$-\frac{\left(\sum_{i=1}^K D_i\gamma_i\right)^2}{\left(g(1)\cdot n + \sum_{i=1}^K \frac{\gamma_i^2}{s_i}\right)}. \tag{199}$$

Then, for any $i < j$, we must have $\hat{s}_i = \min\{\hat{s}_i + \hat{s}_j, k_i\}$.

Proof of Lemma H.1. Since the $\{\hat{\gamma}_i\}_{i\in[K]}, \{\hat{s}_i\}_{i\in[K]}$ are optimal, if we fix two indices $i < j$ the corresponding $\hat{\gamma}_i, \hat{\gamma}_j, \hat{s}_i, \hat{s}_j$ are optimal among all $\gamma_i, \gamma_j, s_i, s_j$ satisfying

$$\begin{cases} 0 < \gamma_i + \gamma_j = \gamma := \hat{\gamma}_i + \hat{\gamma}_j \leq n, \\ 0 < s_i + s_j = s := \hat{s}_i + \hat{s}_j \leq \min\{n, k_i + k_j\}. \end{cases} \tag{200}$$

Thus, we proceed by analysing the solution for two fixed indices under the constraints (200) (keeping all other $\hat{\gamma}_l, \hat{s}_l$ for $l \notin \{i, j\}$ fixed). Note that, for each fixed (γ_i, γ_j) satisfying the constraints (200), the following objective

$$\begin{aligned} &\frac{\gamma_i^2}{s_i} + \frac{\gamma_j^2}{s_j} \to \min_{s_i, s_j} \\ &\text{s.t.} \quad s_i \leq k_i,\ s_j \leq k_j,\ s_i + s_j = s \end{aligned} \tag{201}$$

is equivalent to finding optimal ranks for (199). Importantly, in (201) we consider continuous (s_i, s_j). This relaxation has the same minimum, since we will show that the optimal s_i, s_j have integer values. We may also assume that $\gamma_j > 0$ as otherwise clearly $s_i = \min\{s, k_i\}$ is optimal.

Since (201) is strictly convex (on the domain given by the constraints), we can find its unique minimizer by finding a solution to the KKT conditions:

$$-\frac{\gamma_i^2}{s_i^2} + (\lambda + \mu_i) = 0, \quad -\frac{\gamma_j^2}{s_j^2} + (\lambda + \mu_j) = 0, \quad \mu_i, \mu_j \geq 0, \quad \mu_i(s_i - k_i) = 0, \quad \mu_j(s_j - k_j) = 0, \quad s = s_i + s_j.$$

If $s_i = k_i$ or $s_j = 0$, then the claim is readily obtained. We will now prove that, if this is not the case, then we can find new $\widetilde{s}_i, \widetilde{s}_j, \widetilde{\gamma}_i, \widetilde{\gamma}_j$ which achieve a better value.

We first show that for $s_i < k_i, 0 < s_j < k_j$

$$\frac{\gamma_i^2}{s_i} + \frac{\gamma_j^2}{s_j} = \gamma_i \frac{\gamma}{s} + \gamma_j \frac{\gamma}{s} = \frac{\gamma^2}{s}. \tag{202}$$

Note that, in this case, $\mu_i = \mu_j = 0$, so the first two KKT conditions imply

$$\frac{\gamma_i}{s_i} = \sqrt{\lambda} = \frac{\gamma_j}{s_j}.$$

Thus, we have

$$\frac{\gamma_i}{s_i} = \frac{\gamma_j}{s_j} = \frac{\gamma_i + \gamma_j}{s_i + s_j} = \frac{\gamma}{s}, \tag{203}$$

from which (202) is immediate.

For the case $s_j = k_j$ and $s_i < k_i$, we have that $\mu_j \geq \mu_i = 0$, hence

$$\frac{\gamma_i}{s_i} = \sqrt{\lambda + \mu_i} \leq \sqrt{\lambda + \mu_j} = \frac{\gamma_j}{s_j}.$$

From the previous case, we know that without the constraints on k_i, k_j the optimal value in (201) is $\frac{\gamma^2}{s}$. Thus,

$$\frac{\gamma_i^2}{s_i} + \frac{\gamma_j^2}{s_j} \geq \frac{\gamma^2}{s}.$$

Now, for $\epsilon > 0$, define $\widetilde{s}_i = s_i + \epsilon, \widetilde{s}_j = s_j - \epsilon$. Note that, as $s_i < k_i$ and $s_j > 0$, we can choose ϵ small enough such that $0 < \widetilde{s}_i < k_i, 0 < \widetilde{s}_j < k_j$. At this point, let us simply choose $\widetilde{\gamma}_i, \widetilde{\gamma}_j$ such that

$$\frac{\widetilde{\gamma}_i}{\widetilde{s}_i} = \frac{\widetilde{\gamma}_j}{\widetilde{s}_j}$$

which as in (202), (203) implies that

$$\frac{\widetilde{\gamma}_i^2}{\widetilde{s}_i} + \frac{\widetilde{\gamma}_j^2}{\widetilde{s}_j} = \frac{\gamma^2}{s} \leq \frac{\gamma_i^2}{s_i} + \frac{\gamma_j^2}{s_j}. \tag{204}$$

We also have $\widetilde{\gamma}_i > \gamma_i$, as otherwise

$$\frac{\widetilde{\gamma}_i}{\widetilde{s}_i} < \frac{\gamma_i}{s_i} \leq \frac{\gamma_j}{s_j} < \frac{\widetilde{\gamma}_j}{\widetilde{s}_j}$$

would be a contradiction. This gives that

$$D_i\gamma_i + D_j\gamma_j < D_i\widetilde{\gamma}_i + D_j\widetilde{\gamma}_j,$$

which implies that our new choice achieves a lower value for (199), thus giving the desired contradiction.

□

Lemma H.2. *Assume that f, f_i are differentiable strictly convex functions on $\mathbb{R}$ such that*

$$f_i'(0) < f_j'(0) < 0,\ i < j, \quad \lim_{m_i \to +\infty} f_i'(m_i) = +\infty, \quad \lim_{m_i \to -\infty} f_i'(m_i) = -\infty, \tag{205}$$

and

$$f(0) = f'(0) = 0, \quad \lim_{m \to +\infty} f'(m) = +\infty. \tag{206}$$

Then, the objective given by

$$\min_{m_i \geq 0} f\left(m\right) + \sum_{i=1}^{K} f_i(m_i), \quad m = \sum_{i}^{K} m_i \tag{207}$$

has a unique minimizer. It is uniquely characterised by being of the form $(m_1, \ldots, m_M, 0, \ldots, 0)$ and satisfying

$$m = \sum_{i=1}^{M} \left(\left(-f_i'\right)^{-1} \circ f'\right)(m), \quad m_i = \left(\left(-f_i'\right)^{-1} \circ f'\right)(m) \geq 0, \quad f'(m) + f_i'(m_i) \geq 0, \quad i \in [M]. \tag{208}$$

Furthermore, it can be obtained via binary search by finding the largest index M, such that the corresponding m_i are all strictly positive.

While the assumptions of this theorem might seem technical, most of them can be relaxed. However, we note that all such assumptions are fulfilled by the setting being studied and relaxing them would come at the cost of the readability of the proof of Lemma H.2.

Proof of Lemma H.2. We start by showing that (207) has a unique minimizer. Recall that f and f_i are strictly convex functions, and, hence, their derivatives f' and f_i' are increasing. From (206), we also obtain that $\lim_{m\to+\infty} f'(m) = +\infty$. By monotonicity, we have $f_i'(m_i) \geq f_i'(0)$. Therefore,

$$\lim_{m \to +\infty} f'(m) + \sum_{i=1}^{K} f_i'(m_i) = +\infty,$$

and thus

$$\lim_{m \to +\infty} f(m) + \sum_{i=1}^{K} f_i(m_i) = +\infty.$$

As a consequence, the objective achieves its infimum. Therefore, as $f(m) + \sum_{i=1}^{K} f_i(m_i)$ is strictly convex, the minimum is unique.

Notice that Slater's condition is satisfied, since the feasible set of (207) has an interior point. Hence, $\{m_i\}_{i=1}^{K}$ is a unique minimizer of (207) if and only if it satisfies the following KKT conditions (for the "if and only if" statement, see for instance page 244 in Boyd et al. (2004)):

1. Stationary condition: $f'(m) + f_i'(m_i) - \lambda_i = 0$.

2. Primal feasibility: $m_i \geq 0$.
3. Complementary slackness: $\lambda_i m_i = 0$.
4. Dual feasibility: $\lambda_i \geq 0$.

In particular, the uniqueness of the minimizer implies that the KKT conditions have a unique solution. Thus, we only need to show that the m_i found by this procedure satisfy the above equations.

We now show that the active set $\mathcal{A} := \{i : m_i > 0\}$ for the optimal m_i is monotone, meaning that $\mathcal{A} = [M]$ for some $M \leq K$. We prove the statement by contradiction. Assume that there exists $m_i = 0$ and $m_j > 0$ where $i < j$. Recall that f_j' is strictly increasing, which by the ordering condition (205) implies that

$$f_i'(0) + f'\left(\sum_{\ell=1}^{K} m_\ell\right) < f_j'(m_j) + f'\left(\sum_{\ell=1}^{K} m_\ell\right).$$

Hence, taking some sufficiently small mass from m_j and redistributing it in m_i will decrease the objective value in (207), which concludes the proof.

Fix $M \leq K$. We now show that the solution of the following system of equations

$$f'(m) + f_i'(m_i) = 0, \quad \forall i \leq M \tag{209}$$

exists and unique. Note that this system comes from the 1. and 3. KKT conditions.

As f_i' is strictly monotone, its inverse exists and, hence, from (209) we get

$$m_i = (-f_i')^{-1}(f'(m)), \tag{210}$$

which gives

$$m = \sum_{i=1}^{M} (-f_i')^{-1}(f'(m)). \tag{211}$$

Let us argue the existence and uniqueness of the solution of equation (211) for a fixed M. Recall that f_i' is increasing and, thus, $-f_i'$ is decreasing. The inverse of a decreasing function is decreasing, hence $(-f_i')^{-1}$ is decreasing. Recalling that f' is increasing and that the composition of an increasing and a decreasing function is decreasing, it follows that $(-f_i')^{-1}(f'(m))$ is decreasing. By assumption $f_i'(0) < 0$ and f_i' is increasing such that $\lim_{m_i \to +\infty} f_i(m_i) = +\infty$, therefore the value $(-f_i')^{-1}(0)$ is well-defined and

$$(-f_i')^{-1}(0) > 0.$$

Thus, we have that

$$g_M(m) = \sum_{i=1}^{M} (-f_i')^{-1}(f'(m)) - m$$

is a strictly decreasing function with

$$\lim_{m \to +\infty} g_M(m) = -\infty, \quad g_M(0) > 0.$$

In this view, the solution of (211) exists and unique.

Next, we elaborate on why (210) is well-defined given the solution of (211). Note that, by our assumptions,

$$\lim_{m_i \to +\infty} f_i'(m_i) = +\infty, \quad \lim_{m_i \to -\infty} f_i'(m_i) = -\infty,$$

hence, the same holds for $(-f_i')^{-1}$, and, thus, due to continuity the quantity

$$(-f_i')^{-1}(x)$$

is well-defined for any $x \in \mathbb{R}$. Given this, we readily have that the solution of the system (209) exists and unique. Furthermore, this solution can be found using (211) and (210). Note also that (211) and (210) agree with (208).

We now show that the following procedure finds the optimal active set $\mathcal{A}^* = [M^*]$. Let $m_i(M),\ i \leq M$ be a solution of (209) for fixed value of $M \leq K$, and define $m(M) := \sum_{i=1}^{M} m_i(M)$. Using (211) and (210) find the smallest M such that the corresponding $m_M(M)$ is non-negative, then $M^* = M - 1$ if $M \geq 1$, otherwise, $m = m_i = 0,\ \forall i \in [K]$. If no such M was found, $M^* = [K]$. To show that the described procedure in fact gives the optimal active set $\mathcal{A}^* = [M^*]$, we need to prove that

1. If $M < M^*$, then $m_i(M) \geq 0$.
2. If $M > M^*$, then $m_M(M) \leq 0$.

Clearly, these two conditions imply that the active set of the minimizer is given by $[M^*]$, and it can be found via binary search.

We start by proving the first property. Note that, by the KKT conditions on the optimizer M^*, we have that

$$m_i(M^*) \geq 0.$$

First assume that $m(M) > m(M^*)$. By monotonicity, it follows from (210) that

$$m_i(M) < m_i(M^*),$$

but

$$m(M^*) = \sum_{i=1}^{M} m_i(M^*) + \sum_{i=M+1}^{M^*} m_i(M^*) \geq \sum_{i=1}^{M} m_i(M^*) > \sum_{i=1}^{M} m_i(M) = m(M),$$

where we have used that $m_i(M^*) \geq 0$, which is a contradiction. Thus, we have that $m(M) \leq m(M^*)$. Again, by (210) and monotonicity,

$$m_i(M) \geq m_i(M^*),$$

and, hence, all $m_i(M)$ are non-negative.

We finally argue the second property. We start by proving a weaker statement, i.e., there exists $i \geq M^* + 1$ such that $m_i(M) < 0$. Assume that $m(M) < m(M^*)$. By (210) and monotonicity

$$m_i(M) > m_i(M^*),$$

hence, the following holds:

$$m(M) = \sum_{i=1}^{M} m_i(M) = \sum_{i=1}^{M^*} m_i(M) + \sum_{i=M^*+1}^{M} m_i(M) > \sum_{i=1}^{M^*} m_i(M^*) + \sum_{i=M^*+1}^{M} m_i(M) = m(M^*) + \sum_{i=M^*+1}^{M} m_i(M),$$

which since $m(M) < m(M^*)$ implies that $\sum_{i=M^*+1}^{M} m_i(M)$ is a negative quantity. Thus, there exists $i \geq M^* + 1$ such that $m_i(M) < 0$. Assume now that $m(M) \geq m(M^*)$. Recall that only the minimizer satisfies the KKT conditions, thus

$$f'(m(M^*)) + f'_M(0) \geq 0,$$

which, as f' is increasing, implies that

$$f'(m(M)) + f'_M(0) \geq 0.$$

By construction of $m_M(M)$, we know that

$$f'(m(M)) + f'_M(m_M(M)) = 0,$$

thus, by monotonicity of f'_M we have $m_M(M) \leq 0$.

It remains to show that it suffices to check $m_M(M) \leq 0$ and not an arbitrary $m_i(M)$ for $i \geq M^* + 1$. Assume that $m_i(M) \leq 0$ for some $i \leq M$. Recall that by assumption

$$f'_i(0) < f'_M(0) < 0,$$

and by construction we have

$$f_i'(m_i(M)) = f_M'(m_M(M)) = -f'(m(M)).$$

Since f_i' is a decreasing function, we get that $-f'(m(M)) < f_i'(0)$. Recalling that $f_i'(0) < f_M'(0)$, we get $-f'(m(M)) < f_M'(0)$ and, hence, by monotonicity of f_M' we obtain that $m_M(M) \leq 0$, which concludes the proof. □

Lemma H.3. *The minimizer of* (20) *can be computed in* $\log(K)$ *steps via binary search by finding the smallest index* M^* *such that*

$$\frac{g(1)}{c_1^2 n} \sum_{j=1}^{M^*+1} s_j (D_{M^*+1} - D_j) + D_{M^*+1} \leq 0. \tag{212}$$

Then, the optimal active set has the form $\mathcal{A} = [M^*]$ *and corresponding non-zero* β_i *, for* $i \leq M^*$*, are computed as*

$$\beta_i = \frac{s_i}{c_1} \cdot \left(\frac{\frac{g(1)}{c_1^2 n} \sum_{j \in \mathcal{A}} s_j \Delta_j + D_1}{\frac{g(1)}{c_1^2 n} \sum_{j \in \mathcal{A}} s_j + 1} - \Delta_i \right), \tag{213}$$

where $\Delta_j = D_1 - D_j$.

Proof of Lemma H.3. By rescaling $g(x)$ as $\frac{g(x)}{c_1^2}$ and β_i as $c_1 \beta_i$, we may without loss of generality assume that $c_1 = 1$. From the results of Lemma H.2, by a direct computation, we get that for $\mathcal{A} = [M]$

$$\beta_j(M) = m_j(M) = s_j \cdot \left(\frac{\frac{g(1)}{n} \sum_{i=1}^{M} s_i \Delta_i + D_1}{\frac{g(1)}{n} \sum_{i=1}^{M} s_i + 1} - \Delta_j \right), \ \forall j \leq M,$$

thus, applying the described binary search procedure to find M^* such that $M^* + 1 = \min\left(\arg\min_M \mathbb{1}[m_M(M) > 0]\right)$ finishes the proof.

We now elaborate on the computations. For the compactness of the notation, we omit the dependence on active set in m_i's and m. We apply Lemma H.2 with

$$f(x) = \frac{g(1)}{n} \cdot x^2, \quad f_i(x) = \frac{x^2}{s_i} - 2D_i x,$$

which gives

$$f'(x) = \frac{2g(1)}{n} \cdot x, \quad f_i'(x) = \frac{2x}{s_i} - 2D_i.$$

Hence, we obtain that

$$(-f_i')^{-1}(x) = \frac{s_i \cdot (2D_i - x)}{2},$$

and, thus, by (211) we obtain

$$m = \sum_{i=1}^{M} (-f_i')^{-1}(f'(m)) = -f'(m) \cdot \sum_{i=1}^{M} \frac{s_i}{2} + \sum_{i=1}^{M} D_i s_i = -\frac{g(1)}{n} \cdot m \cdot \sum_{i=1}^{M} s_i + \sum_{i=1}^{M} D_i s_i.$$

In this view, we get

$$m = \frac{\sum_{i=1}^{M} D_i s_i}{\frac{g(1)}{n} \sum_{i=1}^{M} s_i + 1},$$

and, hence, since by (210) the following holds

$$m_j = (-f_j')^{-1}(f'(m)),$$

we get

$$m_j = s_j \cdot \frac{2D_j - f'(m)}{2} = s_j \cdot \frac{2D_j\left(\frac{g(1)}{n}\sum_{i=1}^{M} s_i + 1\right) - \frac{2g(1)}{n}\cdot\sum_{i=1}^{M} D_i s_i}{2\cdot\left(\frac{g(1)}{n}\sum_{i=1}^{M} s_i + 1\right)}$$
$$= s_j \cdot \frac{\frac{g(1)}{n}\sum_{i=1}^{M} D_j s_i + D_j - \frac{g(1)}{n}\sum_{i=1}^{M} D_i s_i + \frac{g(1)}{n}\sum_{i=1}^{M} D_1 s_i - \frac{g(1)}{n}\sum_{i=1}^{M} D_1 s_i - D_1 + D_1}{\frac{g(1)}{n}\sum_{i=1}^{M} s_i + 1} =$$
$$= s_j \cdot \left(\frac{\frac{g(1)}{n}\sum_{i=1}^{M} s_i\Delta_i + D_1}{\frac{g(1)}{n}\sum_{i=1}^{M} s_i + 1} - \Delta_j\right),$$

where $\Delta_j = D_1 - D_j$. It is easy to verify that the condition

$$\frac{g(1)}{n}\sum_{j=1}^{M^*+1} s_j(D_{M^*+1} - D_j) + D_{M^*+1} \le 0$$

described in the statement of the lemma is equivalent to $\beta_{M^*+1}(M^*+1) = m_{M^*+1}(M^*+1) \le 0$, which concludes the proof. □

Proof of Theorem 5.2. We start by showing how the lower bound reduces to the objective in (20). Consider the following block decomposition of $\boldsymbol{B}$ in accordance with $\boldsymbol{D}$ as in (25)

$$\boldsymbol{B} = [\boldsymbol{\Gamma}_1\boldsymbol{B}_1|\cdots|\boldsymbol{\Gamma}_K\boldsymbol{B}_K],$$

where $\boldsymbol{B}_j \in \mathbb{R}^{n\times k_j}$ with $\|(\boldsymbol{B}_j)_{i,:}\|_2 = 1$ and $\{\boldsymbol{\Gamma}_j\}_{j=1}^K$ are diagonal matrices.

Since we require $\|\boldsymbol{B}_{i,:}\|_2 = 1$, the $\boldsymbol{\Gamma}_i$ must satisfy

$$\sum_{j=1}^{K} \boldsymbol{\Gamma}_j^2 = \boldsymbol{I}. \tag{214}$$

Thus, up to a multiplicative factor $1/d$ and an additive term $\mathrm{Tr}\left[\boldsymbol{D}^2\right]$, the objective (19) can be written as:

$$\beta^2\left(\mathrm{Tr}\left[\boldsymbol{M}f(\boldsymbol{M})\right]\right) - 2c_1\beta\cdot\sum_{i=1}^{K} D_i\cdot\mathrm{Tr}\left[\boldsymbol{\Gamma}_i^2\right], \tag{215}$$

where $\boldsymbol{M} = \sum_{i=1}^K \boldsymbol{M}_i := \sum_{i=1}^K \boldsymbol{\Gamma}_i\boldsymbol{B}_i\boldsymbol{B}_i^\top\boldsymbol{\Gamma}_i$. Recall that $f(x) = c_1^2x + g(x)$, where g is the sum of odd monomials. Hence, we will be able to lower bound the terms in the first trace of (215) in a similar fashion to Proposition 4.4. Note that

$$\mathrm{Tr}\left[\boldsymbol{M}_i^2\right] = \langle\boldsymbol{1}, \boldsymbol{M}_i^{\circ 2}\boldsymbol{1}\rangle,$$

so applying Theorem A in Khare (2021) gives that

$$(\boldsymbol{\Gamma}_i\boldsymbol{B}_i\boldsymbol{B}_i^\top\boldsymbol{\Gamma}_i)^{\circ 2} \succeq \frac{1}{s_i}\cdot\mathrm{Diag}(\boldsymbol{\Gamma}_i^2)\mathrm{Diag}(\boldsymbol{\Gamma}_i^2)^\top,$$

where $s_i = \mathrm{rank}(\boldsymbol{B}_i\boldsymbol{B}_i^\top)$. Thus, we have the bound

$$\mathrm{Tr}\left[\boldsymbol{M}_i^2\right] \ge \frac{1}{s_i}\left(\mathrm{Tr}\left[\boldsymbol{\Gamma}_i^2\right]\right)^2$$

Since $xg(x) \ge 0$, we can lower bound the rest of the terms with the identity, i.e.,

$$\mathrm{Tr}\left[\boldsymbol{M}g(\boldsymbol{M})\right] = \langle\boldsymbol{1}, \boldsymbol{M}\circ g(\boldsymbol{M})\boldsymbol{1}\rangle \ge g(1)\cdot n$$

as $\mathrm{Diag}(\boldsymbol{M}) = \boldsymbol{I}$. Consequently, neglecting the cross-terms $\mathrm{Tr}\left[\boldsymbol{M}_i \boldsymbol{M}_j\right]$ (as the trace of the product of PSD matrices is non-negative) we arrive at

$$\mathrm{Tr}\left[\boldsymbol{M} f(\boldsymbol{M})\right] \geq g(1) \cdot n + c_1^2 \cdot \sum_{i=1}^{K} \frac{1}{s_i} \left(\mathrm{Tr}\left[\boldsymbol{\Gamma}_i^2\right]\right)^2.$$

Defining $\gamma_i := \mathrm{Tr}\left[\boldsymbol{\Gamma}_i^2\right] \geq 0$, we arrive at the following lower bound on (215):

$$\beta^2 \left(g(1) \cdot n + \sum_{i=1}^{K} \frac{\gamma_i^2}{s_i} \right) - 2\beta \cdot \sum_{i=1}^{K} D_i \gamma_i, \tag{216}$$

where, with an abuse of notation, we rescale $g(1) := g(1)/c_1^2$ and $\beta := c_1 \beta$. Now, by choosing $\beta_i := \beta \gamma_i$ and using that $\sum_{i=1}^{K} \gamma_i = n$ due to (214), the objective (216) is seen to be equivalent to (20). This shows that (19) $\geq \mathrm{LB}(\boldsymbol{D})$. We now give a brief outline of how one can obtain the optimal s_i and β_i for (20).

For finding the optimal s_i, it is more natural to still consider (216). Due to the block form (25), the s_i have to satisfy the constraints in (21). Note that (216) evaluated at the optimal β is equal to

$$(216) \geq -\frac{\left(\sum_{i=1}^{K} D_i \gamma_i\right)^2}{\left(g(1) \cdot n + \sum_{i=1}^{K} \frac{\gamma_i^2}{s_i}\right)}. \tag{217}$$

The optimal s_i for this objective are *water-filled*, i.e.,

$$\begin{cases} \boldsymbol{s} = [n, 0, \cdots, 0], & n \leq k_1, \\ \boldsymbol{s} = [k_1, k_2, \cdots, k_K], & d \leq n, \\ \boldsymbol{s} = [k_1, \cdots, k_{\mathrm{id}(n)-1}, \mathrm{res}(n), 0, \cdots, 0] & \text{otherwise}, \end{cases} \tag{218}$$

where $\boldsymbol{s} = [s_1, \cdots, s_k]$ and $\mathrm{id}(n)$ denotes the first position at which

$$\min\{n, d\} - \sum_{i=1}^{\mathrm{id}(n)} k_i < 0,$$

and

$$\mathrm{res}(n) = \min\{n, d\} - \sum_{i=1}^{\mathrm{id}(n)-1} k_i.$$

This follows directly from Lemma H.1. It only remains to show that the optimal β_i can be obtained via (24), which is done in Lemma H.3. This concludes the proof. □

Proof of Proposition 5.3. Except for terms of the form $\mathrm{Tr}\left[\boldsymbol{B}_i \boldsymbol{B}_i^\top \boldsymbol{B}_j \boldsymbol{B}_j^\top\right]$, all the other terms can be estimated as in the proof of Proposition 4.4. The only technical difference is that all the constants now depend on the ratios $\frac{k_i}{n}$.

We will show that, with probability at least $1 - c\exp\left(-cd^\epsilon\right)$, for all $i \neq j$,

$$\mathrm{Tr}\left[\boldsymbol{B}_i \boldsymbol{B}_i^\top \boldsymbol{B}_j \boldsymbol{B}_j^\top\right] \leq n^{\frac{1}{2}+\epsilon}. \tag{219}$$

Thus, by a simple union bound, we have that, with probability at least $1 - \frac{c}{d^2}$, this bound holds jointly for all pairs $\boldsymbol{B}_i, \boldsymbol{B}_j$. It follows as in the proof of Lemma C.3 that we can write

$$\boldsymbol{B}_i \boldsymbol{B}_i^\top = \boldsymbol{P}_i \boldsymbol{U} \boldsymbol{D}_i \boldsymbol{U}^\top \boldsymbol{P}_i,$$

where by abuse of notation we pushed the factor $\frac{n}{k_i}$ in $\boldsymbol{D}_i$ (which will only affect the constants c, C). Here, $\boldsymbol{P}_i$ is a diagonal matrix such that, for any $\epsilon > 0$, with probability at least $1 - c\exp\left(-cd^\epsilon\right)$, we have that

$$\left\|\boldsymbol{P}_i - \boldsymbol{I}\right\|_{op} \leq n^{-\frac{1}{2}+\epsilon}.$$

To see this, first observe that $\Theta : (\mathbb{R}^{n\times n})^4 \mapsto \mathbb{R}$ given by $\Theta(\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \boldsymbol{X}_4) = \mathrm{Tr}\left[\boldsymbol{X}_1\boldsymbol{U}\boldsymbol{D}_i\boldsymbol{U}^\top\boldsymbol{X}_2\boldsymbol{X}_3\boldsymbol{U}\boldsymbol{D}_j\boldsymbol{U}^\top\boldsymbol{X}_4\right]$ is differentiable (as it is the composition of the trace function with 4-linear form). Since by construction

$$\mathrm{Tr}\left[\boldsymbol{U}\boldsymbol{D}_i\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{D}_j\boldsymbol{U}^\top\right] = \mathrm{Tr}\left[\boldsymbol{0}\right] = 0,$$

this implies that, with probability at least $1 - \frac{c}{d^2}$,

$$\begin{aligned}
0 &\le \mathrm{Tr}\left[\boldsymbol{B}_i\boldsymbol{B}_i^\top\boldsymbol{B}_j\boldsymbol{B}_j^\top\right] \\
&= \mathrm{Tr}\left[\boldsymbol{P}_i\boldsymbol{U}\boldsymbol{D}_i\boldsymbol{U}^\top\boldsymbol{P}_i\boldsymbol{P}_j\boldsymbol{U}\boldsymbol{D}_j\boldsymbol{U}^\top\boldsymbol{P}_j\right] \\
&= \mathrm{Tr}\left[\boldsymbol{P}_i\boldsymbol{U}\boldsymbol{D}_i\boldsymbol{U}^\top\boldsymbol{P}_i\boldsymbol{P}_j\boldsymbol{U}\boldsymbol{D}_j\boldsymbol{U}^\top\boldsymbol{P}_j\right] - \mathrm{Tr}\left[\boldsymbol{U}\boldsymbol{D}_i\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{D}_j\boldsymbol{U}^\top\right] \\
&\le Cnn^{-\frac{1}{2}+\epsilon},
\end{aligned}$$

where in the last step we used that the derivative of the trace function is bounded by $n \cdot \|\cdot\|_{op}$. Thus, (219) holds.

By construction, the sum of all the cross terms is of the form

$$\sum_{i\neq j} \mathrm{Tr}\left[\boldsymbol{M}_i\boldsymbol{M}_j\right],$$

where $\boldsymbol{M}_i = \boldsymbol{\Gamma}_i\boldsymbol{B}_i\boldsymbol{B}_i^\top\boldsymbol{\Gamma}_i$, $\boldsymbol{\Gamma}_i^2 = \frac{\gamma_i}{n}\boldsymbol{I}$ and $\sum_{i=1}^K \gamma_i = n$. We have

$$\begin{aligned}
\left|\sum_{i\neq j} \mathrm{Tr}\left[\boldsymbol{M}_i\boldsymbol{M}_j\right]\right| &= \left|\sum_{i\neq j} \frac{\gamma_i\gamma_j}{n^2}\mathrm{Tr}\left[\boldsymbol{B}_i\boldsymbol{B}_i^\top\boldsymbol{B}_j\boldsymbol{B}_j^\top\right]\right| \\
&\le \sum_{i\neq j} \frac{\gamma_i\gamma_j}{n^2}\left|\mathrm{Tr}\left[\boldsymbol{B}_i\boldsymbol{B}_i^\top\boldsymbol{B}_j\boldsymbol{B}_j^\top\right]\right| \\
&\le C\sum_{i\neq j} \frac{\gamma_i\gamma_j}{n^2} n^{\frac{1}{2}+\epsilon} \\
&\le Cn^{\frac{1}{2}+\epsilon},
\end{aligned}$$

where in the third step we used a union bound on (219) and in the last step we used $\sum_{i=1}^K \frac{\gamma_i}{n} = 1$. □

I. Details of Experiments and Additional Numerical Results

We first describe the training details and the whitening procedure that is used to preprocess natural images for MNIST (Figure 8) and CIFAR-10 (Figures 1, 5 and 7). Next, we give some remarks about the experiments concerning VAMP (Figure 4) and about the discontinuous behaviour of the derivative of the lower bound highlighted in Figure 6. In addition, we present additional numerical experiments which cover extra classes of natural images.

Activation function and weight parameterization. Note that the derivative of the sign activation is zero almost everywhere (except one point, which is the origin). In this view, we cannot use conventional gradient-based algorithms to find the optimal set of parameters for an autoencoder with the sign activation. We tackle this issue by using a straight-through estimator (see, for instance, (Yin et al., 2019)) of the sign activation. During the forward pass the activations of the first layer are computed for $\sigma(x) = \mathrm{sign}(x)$, while during the backward pass $\sigma(x) = \tanh(x/\tau)$ is used. Here, the temperature parameter $\tau > 0$ controls how well the differentiable surrogate $\tanh(x/\tau)$ approximates $\mathrm{sign}(x)$, as

$$\lim_{\tau\to 0} \tanh(x/\tau) = \mathrm{sign}(x), \quad \forall x \in \mathbb{R}\setminus\{0\}.$$

More precisely, the differentiable approximation becomes more accurate for smaller values of τ. However, we also note that extremely small values of τ might cause numerical issues, since the derivative of the differentiable surrogate diverges at the origin as $\tau \to 0$. For the numerical experiments, we pick $\tau \in [0.01, 0.2]$, with the exact value depending on the specific setting.

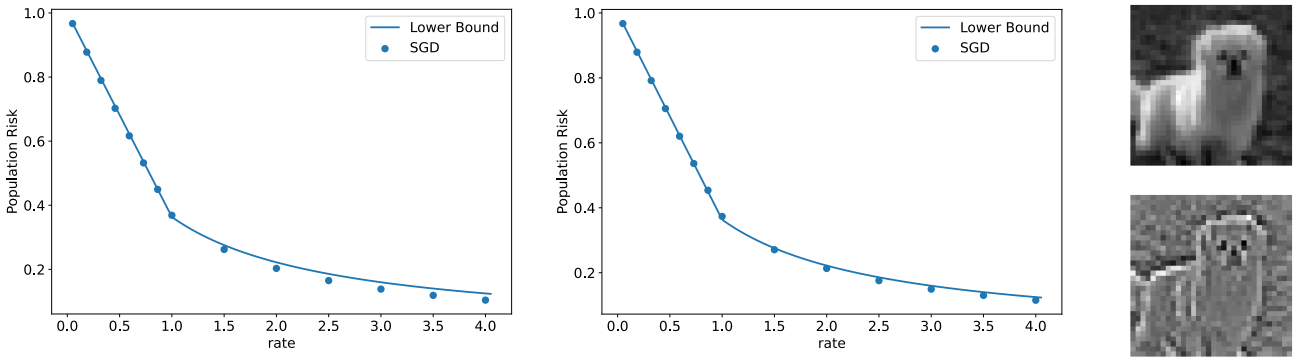

Figure 5. Compression ($\sigma \equiv \text{sign}$) of the CIFAR-10 "dog" class with a two-layer autoencoder. The data is *whitened* so that $\boldsymbol{\Sigma} = \boldsymbol{I}$: on top, an example of a grayscale image; on the bottom, the corresponding whitening. The blue dots are the population risk obtained via SGD, and they agree well with the solid line corresponding to the lower bounds of Theorem 4.2 and Proposition 4.3. Here, the effect of the number of augmentations used per image is shown. For the left plot each image was augmented 10 times, while for the right plot each image was augmented 15 times.

Note that the constraint on the encoder weights $\|\boldsymbol{B}_{i,:}\|_2 = 1$ can be enforced via a simple reparameterization that forces the rows of $\boldsymbol{B}$ to lie on the unit sphere $\mathbb{S}^{d-1}$. More precisely, we use the following classical differentiable reparameterization of $\boldsymbol{B}^\top = [\boldsymbol{b}_1, \cdots, \boldsymbol{b}_n]$, where $\boldsymbol{b}_i = \frac{\hat{\boldsymbol{b}}_i}{\|\hat{\boldsymbol{b}}_i\|_2}$, with $\{\hat{\boldsymbol{b}}_i\}_{i=1}^n$ being the trainable parameters. We note that it is not clear a priori whether we need to impose the constraints directly for the straight-through estimator, since during the forward pass we use the norm-agnostic sign function.

Augmentation and whitening. For the experiments on natural images, we augment the data of each class 15 times. This is done to emulate the optimization of the population risk, since the amount of initial data (approximately 5000 samples per class) leads to a gap between empirical and population risks, especially for high rates. The effect of the data augmentation is represented in Figure 5 for a whitened CIFAR-10 class. It can be seen that a mild amount of augmentation, i.e., $\times 10$ and $\times 15$, is already enough for our purposes, and the difference between the two plots is rather small. Notably, this amount of augmentation brings the dataset to the scale of the original data when all classes are considered (around 50000 training examples).

The whitening procedure used in the experiments concerning isotropic data is performed as follows: given the *centered* augmented data $\boldsymbol{X} \in \mathbb{R}^{\mathrm{n_{samples}} \times d}$, we compute its empirical covariance matrix given by

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

and then we multiply each input by the inverse square root of it, i.e.,

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

The resulting whitened images are represented in Figures 1, 5 and 8.

In the experiments concerning non-isotropic data (Figures 2 and 9), we center the data with the empirical mean and divide by a *scalar* empirical variance computed across all the pixels, which is the standard preprocessing procedure widely used for computer vision tasks.

VAMP experiments. For the VAMP experiments, we implement the State Evolution (SE) recursion which exactly characterizes the limiting performance of VAMP as $d \to \infty$, see (Schniter et al., 2016; Rangan et al., 2019) for an overview. We then plot the fixed point of said SE recursion. A concrete description for VAMP is provided by Algorithm 2 in (Fletcher et al., 2018), which however covers a more general multi-layer setting.

"Jumps" of the lower bound derivative. The derivative switch described in Figure 6 does not necessarily happen precisely at the point when the block is filled. A switch may occur at a later point since, even if $s_i > 0$, the corresponding optimal β_i may be 0. Intuitively, this phenomenon occurs in cases when it is still better to put more mass in the block where the rank is

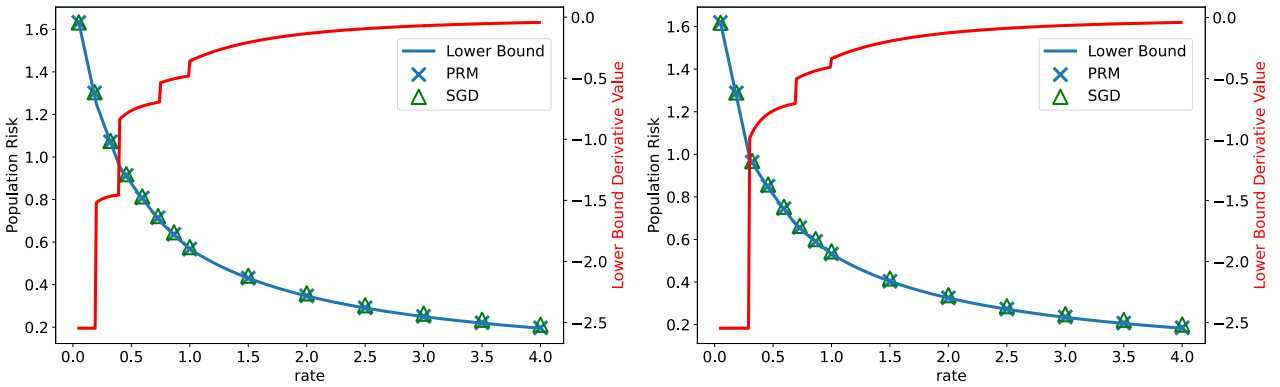

Figure 6. Compression ($\sigma \equiv \text{sign}$) of a non-isotropic Gaussian source, whose covariance matrix is obtained by taking $\boldsymbol{k} = (20, 20, 35, 25)$ and $(D_1, D_2, D_3, D_4) = (2, 1.5, 1, 0.8)$ for the left plot, and $\boldsymbol{k} = (30, 40, 30)$ and $(D_1, D_2, D_3) = (2, 1, 0.7)$ for the right plot. The blue crosses (Population Risk Minimizer, PRM) are obtained by optimizing (19) via GD. The green triangles are obtained by training an autoencoder via SGD on Gaussian samples with the given covariance structure. The red solid line plots the derivative of the population risk computed using a finite differences scheme. Note that the derivative jumps when the corresponding blocks are getting filled, although this may not happen in general, see Appendix I. A similar behavior can be observed in the isotropic case at $r = 1$, as there is only one block to fill (see Figure 4).

utilized to the fullest ($s_j = k_j$). This corresponds to the following condition on the derivatives of the objective (20):

$$\frac{\partial(20)}{\partial \beta_i}(0) > \frac{\partial(20)}{\partial \beta_j}(\beta_j^*),$$

where β_i^* stands for the optimal β_i and j denotes the first index at which $\beta_j^* > 0$. This behaviour occurs when the spectrum $\boldsymbol{D}$ has a large variation in scale, e.g.,

$$\boldsymbol{D} = [5, 0.02, 0.01].$$

In this case, the last components will be utilized for n significantly larger than k_1 ($n = k_1$ precisely characterizes the point where the rank of the first block of $\boldsymbol{B}$, i.e., $\boldsymbol{B}_1$, is the maximum possible). Note that, for this choice of $\boldsymbol{D}$, the plot of the derivative analogous to Figure 6 will not indicate such prominent "jumps". In fact, the contribution of the last components to the derivative value is less significant in comparison to the analogous quantity evaluated for the top-most eigenvalues.

Additional experimental data. We also provide additional numerical simulations, similar to those presented in the body of the paper. In particular, we provide more class variations for the natural data experiments (MNIST and CIFAR-10).

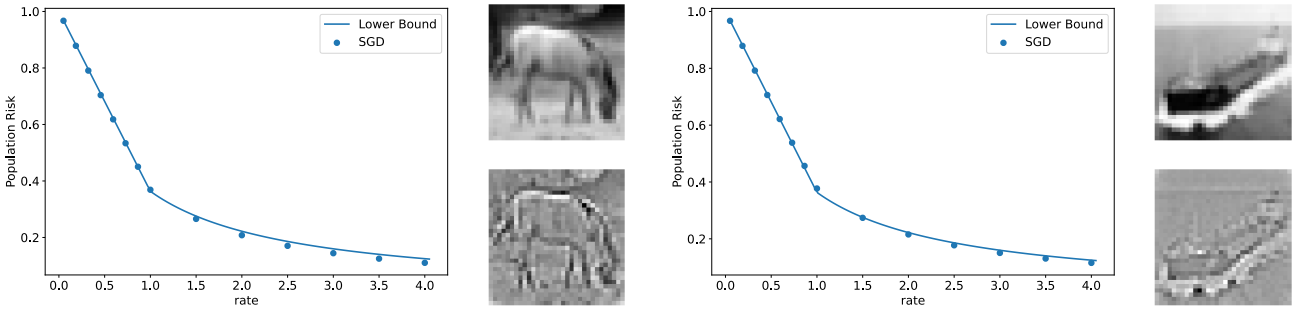

Figure 7. Compression ($\sigma \equiv \text{sign}$) of the CIFAR-10 "horse" class (left) and "ship" class (right) with a two-layer autoencoder. The data is *whitened* so that $\boldsymbol{\Sigma} = \boldsymbol{I}$: on top, an example of a grayscale image; on the bottom, the corresponding whitening. The blue dots are the population risk obtained via SGD, and they agree well with the solid line corresponding to the lower bounds of Theorem 4.2 and Proposition 4.3. Here, in both cases the amount of augmentations per image is equal to 15.

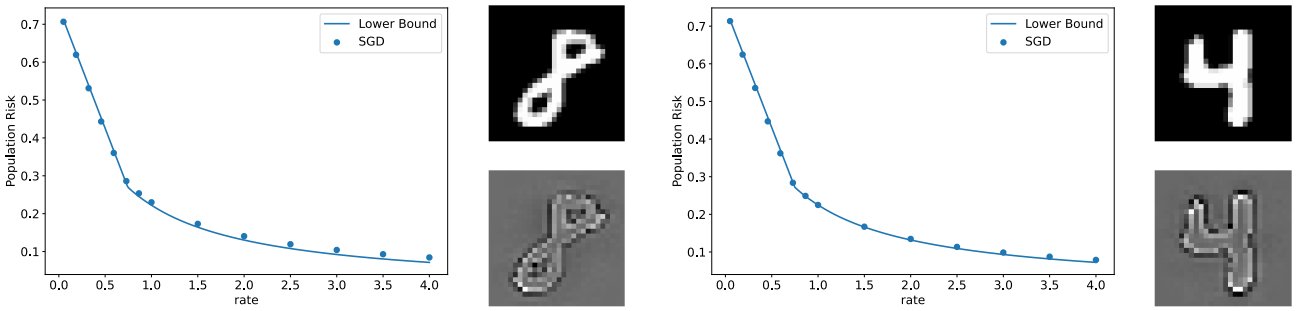

Figure 8. Compression ($\sigma \equiv \text{sign}$) of the MNIST "8" class (left) and "4" class (right) with a two-layer autoencoder. The data is *whitened* so that $\boldsymbol{\Sigma} = \boldsymbol{I}$: on top, an example of a grayscale image; on the bottom, the corresponding whitening. The blue dots are the population risk obtained via SGD, and they agree well with the solid line corresponding to the lower bounds of Theorem 4.2 and Proposition 4.3. Here, in both cases the amount of augmentations per image is equal to 10.

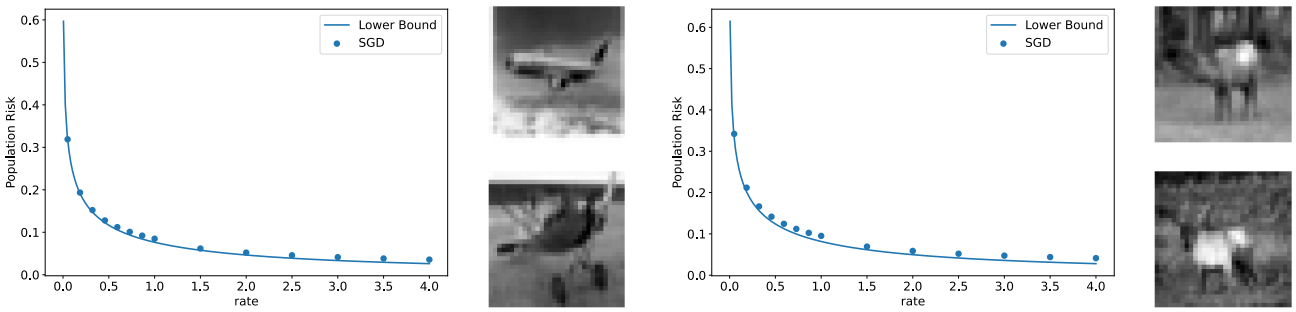

Figure 9. Compression ($\sigma \equiv \text{sign}$) of the CIFAR-10 "airplane" class (left) and "deer" class (right) with a two-layer autoencoder. The data is *not whitened* ($\boldsymbol{\Sigma} \neq \boldsymbol{I}$). The blue dots are the SGD population risk, and they are close to the lower bound of Theorem 5.2. Here, in both cases the amount of augmentations per image is equal to 15.