LEARNING DEEP LINEAR NEURAL NETWORKS: RIEMANNIAN GRADIENT FLOWS AND CONVERGENCE TO GLOBAL MINIMIZERS

BUBACARR BAH[1], HOLGER RAUHUT[2], ULRICH TERSTIEGE[2], AND MICHAEL WESTDICKENBERG[3]

ABSTRACT. We study the convergence of gradient flows related to learning deep linear neural networks (where the activation function is the identity map) from data. In this case, the composition of the network layers amounts to simply multiplying the weight matrices of all layers together, resulting in an overparameterized problem. The gradient flow with respect to these factors can be re-interpreted as a Riemannian gradient flow on the manifold of rank-r matrices endowed with a suitable Riemannian metric. We show that the flow always converges to a critical point of the underlying functional. Moreover, we establish that, for almost all initializations, the flow converges to a global minimum on the manifold of rank k matrices for some $k \leq r$.

1. INTRODUCTION

Deep learning [11] forms the basis of remarkable breakthroughs in many areas of machine learning. Nevertheless, its inner workings are not yet well-understood and mathematical theory of deep learning is still in its infancy. Training a neural networks amounts to solving a suitable optimization problem, where one tries to minimize the discrepancy between the predictions of the model and the data. One important open question concerns the convergence of commonly used gradient descent and stochastic gradient descent algorithms to the (global) minimizers of the corresponding objective functionals. Understanding this problem for general nonlinear deep neural networks seems to be very involved. In this paper, we study the convergence properties of gradient flows for learning deep *linear* neural networks from data. While the class of linear neural networks may be not be rich enough for many machine learning tasks, it is nevertheless instructive and still a non-trivial task to understand the convergence properties of gradient descent algorithms. Linearity here means that the activation functions in each layer are just the identity map, so that the weight matrices of all layers are multiplied together. This results in an overparameterized problem.

Our analysis builds on previous works on optimization aspects for learning linear networks [23, 13, 5, 4, 10, 22]. In [4] the gradient flow for weight matrices of all network layers is analyzed and an equation for the flow of their product is derived. The article [4] then establishes local convergence for initial points close enough to the (global) minimum. In [10] it is shown that under suitable conditions the flow converges to a critical point for any initial point. We contribute to this line of work in the following ways:

- We show that in the balanced case (see definition 1) the evolution of the product of all network layer matrices can be re-interpreted as a Riemannian gradient flow on the manifold of matrices of some fixed rank (see Corollary 17). This is remarkable because it is shown in [4] that the flow of this product cannot be interpreted as a standard gradient flow with respect to some functional. Our result is possible because we use a non-trivial Riemannian metric.
- We show in Theorem 5 that the flow always converges to a critical point of the loss functional L^N, see (2). This results applies under significantly more general assumptions than the mentioned result of [10].

[1]AFRICAN INSTITUTE FOR MATHEMATICAL SCIENCES (AIMS) SOUTH AFRICA, & STELLENBOSCH UNIVERSITY, 6 MELROSE ROAD, MUIZENBERG, CAPE TOWN 7945, SOUTH AFRICA

[2]CHAIR FOR MATHEMATICS OF INFORMATION PROCESSING, RWTH AACHEN UNIVERSITY, PONTDRIESCH 10, 52062 AACHEN, GERMANY

[3]INSTITUTE FOR MATHEMATICS, RWTH AACHEN UNIVERSITY, TEMPLERGRABEN 55, 52062 AACHEN, GERMANY

E-mail addresses: bubacarr@aims.ac.za, rauhut@mathc.rwth-aachen.de, terstiege@mathc.rwth-aachen.de, mwest@instmath.rwth-aachen.de.

Date: October 15, 2020.

- We show that the flow converges to the global optimum of L^1, see (4), restricted to the manifold of rank k matrices for almost all initializations (Theorem 39), where the rank may be anything between 0 and r (the smallest of the involved matrix dimensions). In the case of two layers, we show in the same theorem that for almost all initial conditions, the flow converges to a global optimum of L^2, see (2). Our result in the case of two layers again applies under significantly more general conditions than a similar result in [10]. For the proof, we extend an abstract result in [16] that shows that strict saddle points of the functional are avoided almost surely. Moreover, we give an analysis of the critical points and saddle points of L^1 and L^N, which generalizes and refines results of [13, 22].

We believe that our results shed new light on global convergence of gradient flows (and thereby on gradient descent algorithms) for learning neural network. We expect that the insights will be useful for extending them to learning *nonlinear* neural networks.

Structure. This article is structured as follows. Section 2 describes the setup of gradient flows for learning linear neural networks and collects some basic results. Section 3 shows convergence of the flow to a critical point of the functional. Section 4 provides the interpretation as Riemannian gradient flow on the manifold of rank-r matrices. For the special case of a linear autoencoder with two coupled layers and balanced initial points, Section 5 shows convergence of the flow to a global optimum for almost all starting points by building on [23]. Section 6 extends this result to general linear networks with an arbitrary number of (non-coupled) layers by first extending an abstract result in [16] that first order methods avoid strict saddle points almost surely to gradient flows and then analyzing the strict saddle point property for our functional under consideration. Section 7 illustrates our findings with numerical experiments. Appendices A and B contain detailed proofs of Propositions 10 and 11; while Appendices C and D collect additional results on flows on manifolds and on the autoencoder case with two (non-coupled) layers, respectively.

2. Gradient flows for learning linear networks

Suppose we are given data points $x_1, \ldots, x_m \in \mathbb{R}^{d_x}$ and label points $y_1, \ldots, y_m \in \mathbb{R}^{d_y}$. The learning task consists in finding a map f such that $f(x_j) \approx y_j$. In deep learning, candidate maps are given by deep neural networks of the form

$$f(x) = f_{W_1, \ldots, W_N, b_1, \ldots, b_N}(x) = g_N \circ g_{N-1} \circ \cdots \circ g_1(x),$$

where each layer is of the form $g_j(z) = \sigma(W_j z + b_j)$ with matrices W_j and vectors b_j and an activation function $\sigma : \mathbb{R} \to \mathbb{R}$ that acts componentwise. The parameters $W_1, \ldots, W_N, b_1, \ldots, b_N$ are commonly learned from the data via empirical risk minimization. Given a suitable loss function $\ell : \mathbb{R}^{d_y} \times \mathbb{R}^{d_y} \to \mathbb{R}$, one considers the optimization problem

$$\min_{W_1, \ldots, W_N, b_1, \ldots, b_N} \sum_{j=1}^{m} \ell(f_{W_1, \ldots, W_N, b_1, \ldots, b_N}(x_j), y_j).$$

In this article, we are interested in understanding the convergence behavior of the gradient flow (as simplification of gradient descent) for the minimization of this functional. Since providing such understanding for the general case seems to be hard, we concentrate on the special case of linear networks (with $b_j = 0$ for all j) and the ℓ_2-loss $\ell(z, y) = \|y - z\|_2^2 / 2$ in this article, i.e., the network takes the form

$$f(x) = W_N \cdot W_{N-1} \cdots W_1 x, \quad \text{for } N \geq 2,$$

where $W_j \in \mathbb{R}^{d_j \times d_{j-1}}$ for $d_0 = d_x$, $d_N = d_y$ and $d_1, \ldots, d_{N-1} \in \mathbb{N}$. Clearly, $f(x) = Wx$ with the factorization

$$W = W_N \cdots W_1, \tag{1}$$

which can be viewed as an overparameterization of the matrix W. Note that the factorization imposes a rank constraint as the rank of W is at most $r = \min\{d_0, d_1, \ldots, d_N\}$. The ℓ_2-loss leads to the functional

$$L^N(W_1, \ldots, W_N) = \frac{1}{2} \sum_{j=1}^m \|y_j - W_N \cdots W_1 x_j\|_2^2 = \frac{1}{2} \|Y - W_N \cdots W_1 X\|_F^2 \tag{2}$$

where $X \in \mathbb{R}^{d_x \times m}$ is the matrix with columns $x_1, \ldots, x_m$ and $Y \in \mathbb{R}^{d_y \times m}$ the matrix with columns $y_1, \ldots, y_m$. Here $\|\cdot\|_F$ denotes the Frobenius norm induced by the inner product $\langle A, B \rangle_F := \operatorname{tr}(AB^T)$.

Empirical risk minimization is the optimization problem

$$\min_{W_1, \ldots, W_N} L^N(W_1, \ldots, W_N), \quad \text{where } W_j \in \mathbb{R}^{d_j \times d_{j-1}}, \ j = 1, \ldots, N. \tag{3}$$

For $W \in \mathbb{R}^{d_y \times d_x}$, we further introduce the functional

$$L^1(W) := \frac{1}{2} \|Y - WX\|_F^2. \tag{4}$$

Since the rank of $W = W_N \cdots W_1$ is at most $r = \min\{d_0, d_1, \ldots, d_N\}$, minimization of L^N is closely related to the minimization of L^1 restricted to the set of matrices of rank at most r, but the optimization of L^N does not require to formulate this constraint explicitly. However, L^N is not jointly convex in $W_1, \ldots, W_N$ so that understanding the behavior of corresponding optimization algorithms is not trivial.

The case of an autoencoder [11, Chapter 14], studied in detail below, refers to the situation where $Y = X$. Here one tries to find for W a projection onto a subspace of dimension r that best approximates the data, i.e., $Wx_\ell \approx x_\ell$ for $\ell = 1, \ldots, m$. This task is relevant for unsupervised learning and only the rank deficient case, where $r := \min_{i=0,\ldots,N} d_i < m$ is of interest then, as otherwise one could simply set $W = I_{d_x}$ and there would be nothing to learn.

The gradient of L^1 is given as

$$\nabla_W L^1(W) = WXX^T - YX^T.$$

For given initial values $W_j(0)$, $j \in \{1, \ldots, N\}$, we consider the system of gradient flows

$$\dot{W}_j = -\nabla_{W_j} L^N(W_1, \ldots, W_N). \tag{5}$$

Our aim is to investigate when this system converges to an optimal solution, i.e., one that is minimizing our optimization problem (3). For $W = W_N \cdots W_1$ we also want to understand the behavior of $W(t)$ as t tends to infinity. Clearly, the gradient flow is a continuous version of gradient descent algorithms used in practice and has the advantage that its analysis does not require discussing step sizes etc. We postpone the extension of our results to gradient descent algorithms to later contributions.

Definition 1. Borrowing notation from [4], for $W_j \in \mathbb{R}^{d_j \times d_{j-1}}$, $j = 1, \ldots, N$, we say that $W_1, \ldots, W_N$ are 0-*balanced* or simply *balanced* if

$$W_{j+1}^T W_{j+1} = W_j W_j^T \text{ for } j = 1, \ldots, N-1.$$

We say that the flow (5) has balanced initial conditions if $W_1(0), \ldots, W_N(0)$ are balanced.

The following lemma summarizes basic properties of the flow which are well known; see [5, 4, 10].

Lemma 2. *With the notation above, the following holds:*

(1) For $j \in \{1, \ldots, N\}$,

$$\nabla_{W_j} L^N(W_1, \ldots, W_N) = W_{j+1}^T \cdots W_N^T \nabla_W L^1(W_N \cdots W_1) W_1^T \cdots W_{j-1}^T.$$

(2) Assume the $W_j(t)$ satisfy (5). Then $W = W_N \cdots W_1$ satisfies

$$\frac{dW(t)}{dt} = -\sum_{j=1}^N W_N \cdots W_{j+1} W_{j+1}^T \cdots W_N^T \nabla_W L^1(W) W_1^T \cdots W_{j-1}^T W_{j-1} \cdots W_1. \tag{6}$$

(3) *For all $j = 1, \ldots, N-1$ and all $t \geq 0$ we have that*

$$\frac{d}{dt}\left(W_{j+1}^T(t)W_{j+1}(t)\right) = \frac{d}{dt}\left(W_j(t)W_j^T(t)\right).$$

In particular, the differences

$$W_{j+1}^T(t)W_{j+1}(t) - W_j(t)W_j^T(t), \quad j = 1, \ldots, N-1,$$

and the differences

$$\|W_j(t)\|_F^2 - \|W_i(t)\|_F^2, \quad i, j = 1, \ldots, N,$$

are all constant in time.

(4) *If $W_1(0), \ldots, W_N(0)$ are* balanced, *then*

$$W_{j+1}^T(t)W_{j+1}(t) = W_j(t)W_j^T(t)$$

for all $j \in \{1, \ldots, N-1\}$ and $t \geq 0$, and

$$R(t) := \frac{dW(t)}{dt} + \sum_{j=1}^N (W(t)W(t)^T)^{\frac{N-j}{N}} \nabla_W L^1(W)(W(t)^T W(t))^{\frac{j-1}{N}} = 0. \tag{7}$$

Here and the sequel, by the p-th root of a symmetric and positive semidefinite matrix we mean the principal p-th root, i.e. the p-th root is symmetric and positive semidefinite again. A concrete reference for the statements of the lemma is [5, Theorem 1] together with its proof. For point (3), see also [10, Lemma 1].

Definition 3. *For $W, Z \in \mathbb{R}^{d_y \times d_x}$ and $N \geq 2$ let*

$$\mathcal{A}_W(Z) = \sum_{j=1}^N (WW^T)^{\frac{N-j}{N}} \cdot Z \cdot (W^T W)^{\frac{j-1}{N}}. \tag{8}$$

Thus, if the $W_j(0)$ are balanced (see Definition 1), then

$$\frac{dW(t)}{dt} = -\mathcal{A}_{W(t)}\left(\nabla_W L^1(W(t))\right). \tag{9}$$

We will write this as a gradient flow with respect to a suitable Riemannian metric in Section 4.

3. Convergence of the gradient flow

In this section we will show that the gradient flow always converges to a critical point of L^N, also called an equilibrium point in the following, provided that XX^T has full rank. We do not assume balancedness of the initial data. A similar statement was shown in [10, Proposition 1] and similarly as in loc. cit., our proof is based on Lojasiewicz's Theorem and uses the fact that for $j = 1, \ldots, N-1$ the terms $W_{j+1}^T(t)W_{j+1}(t) - W_j(t)W_j^T(t)$ are constant (cf. Lemma 2), but the technical exposition differs and we do not need the assumptions $d_y \leq d_x$ and $d_y \leq r = \min\{d_1, \ldots, d_{N-1}\}$ made in [10], which, for instance, exclude the autoencoder case and imply that the set $\mathcal{M}_r$ of all admissible matrices appearing as a product $W = W_N \cdots W_1$, i.e., the variety of matrices of rank at most r, coincides with the vector space $\mathbb{R}^{d_x \times d_y}$. Let us first recall the following corollary of Lojasiewicz's Inequality; see [1, 17, 10, 14, 21].

Theorem 4. *If $f: \mathbb{R}^n \to \mathbb{R}$ is analytic and the curve $t \mapsto x(t) \in \mathbb{R}^n$, $t \in [0, \infty)$, is bounded and a solution of the gradient flow equation $\dot{x}(t) = -\nabla f(x(t))$, then $x(t)$ converges to a critical point of f as $t \to \infty$.*

This result, sometimes called Lojasiewicz's Theorem, follows from Theorem 2.2 in [1], for example (see also Theorem 1 in [10]). Indeed it is shown in [1] that under our assumptions $x(t)$ converges to a limit point x^*. By continuity, it follows that also the time derivative $\dot{x}(t) = -\nabla f(x(t))$ converges to a limit point $z := -\nabla f(x^*)$. Then $z = 0$, i.e., x^* is a critical point of f. Indeed, if z had a component $z_k \neq 0$ then for t_0 large enough we would have $|\dot{x}_k(t) - z_k| \leq \frac{|z_k|}{2}$ for all $t \geq t_0$ and hence for $t_2 \geq t_1 \geq t_0$ we would have $|x_k(t_2) - x_k(t_1)| = |\int_{t_1}^{t_2} \dot{x}_k(t)dt| \geq (t_2 - t_1)\frac{|z_k|}{2}$, contradicting the convergence of x_k.

Theorem 5. *Assume XX^T has full rank. Then the flows $W_i(t)$ defined by (5) and $W(t)$ given by (6) are defined and bounded for all $t \geq 0$ and $(W_1, \ldots, W_N)$ converges to a critical point of L^N as $t \to \infty$.*

Proof. Note that the right-hand sides of (5) and (6) are continuous functions so existence of solutions locally in time follows from the Cauchy-Peano theorem. In order to show that the solutions exist for all times and to be able to apply Lojasiewicz's Theorem, we want to show that the $\|W_i(t)\|_F$ are bounded. We will first show that the flow $W(t)$ given by (6) remains bounded for all t. We observe that for all $t \geq 0$ for which $W(t)$ is defined we have $L^1(W(t)) \leq L^1(W(0))$. To see this, note that

$$\frac{d}{dt} L^1(W(t)) = \frac{d}{dt} L^N(W_1(t), \ldots, W_N(t)) = \sum_{i=1}^N D_{W_i} L^N((W_1(t), \ldots, W_N(t))\dot{W}_i(t)$$

$$= -\sum_{i=1}^N \|\nabla_{W_i} L^N((W_1(t), \ldots, W_N(t))\|_F^2 \leq 0.$$

Here the notation D_{W_i} denotes the directional derivative w.r.t. W_i. Hence, for any $t \geq 0$ we have

$$\|W(t)\|_F = \|W(t)XX^T(XX^T)^{-1}\|_F \leq \|W(t)X\|_F \|X^T(XX^T)^{-1}\|_F = \|W(t)X - Y + Y\|_F \|X^T(XX^T)^{-1}\|_F$$

$$\leq (\|W(t)X - Y\|_F + \|Y\|_F) \|X^T(XX^T)^{-1}\|_F = \left(\sqrt{2L^1(W(t))} + \|Y\|_F\right) \|X^T(XX^T)^{-1}\|_F$$

$$\leq \left(\sqrt{2L^1(W(0))} + \|Y\|_F\right) \|X^T(XX^T)^{-1}\|_F.$$

In particular, $\|W(t)\|_F$ is bounded. Recall that the Frobenius norm is submultiplicative.

Next, in order to show the boundedness of the $\|W_i(t)\|_F$, we show the following claim: For any $i \in \{1, \ldots, N\}$, we have

$$\|W_i(t)\|_F \leq C_i \|W(t)\|_F^{1/N} + \widetilde{C}_i, \tag{10}$$

for all $t \geq 0$ (for which the $W_i(t)$ and hence also $W(t)$ are defined). Here C_i and $\widetilde{C}_i$ are suitable positive constants depending only on the initial conditions.

Before we prove the claim, we introduce the following notation.

Definition 6. Suppose we are given a set of (real valued) matrices $\{X_i, i \in I\}$, where I is a finite set. A polynomial P in the matrices X_i, $i \in I$, with matrix coefficients is a (finite) sum of terms of the form

$$A_1 X_{i_1} A_2 X_{i_2} \cdots A_n X_{i_n} A_{n+1}. \tag{11}$$

The A_j are the matrix coefficients of the monomial (11) (where the dimensions of the A_j have to be such that the product (11) as well as the sum of all the terms of the form (11) in the polynomial P are well defined). The degree of the polynomial P is the maximal value of n in the summands of the above form (11) defining P (where $n = 0$ is also allowed).

In the following, the constants are allowed to depend on the dimensions d_i and the initial matrices $W_i(0)$. We will suppress the argument t.

To prove the claim, we observe that

$$WW^T = W_N \cdots W_1 W_1^T \cdots W_N^T.$$

Replacing $W_1 W_1^T$ by $W_2^T W_2 + A_{12}$, where A_{12} is a constant matrix (see Lemma 2 (3)), we obtain

$$WW^T = W_N \cdots W_3 W_2 W_2^T W_2 W_2^T W_3^T \cdots W_N^T + W_N \cdots W_2 A_{12} W_2^T \cdots W_N^T.$$

We now replace $W_2 W_2^T$ by $W_3^T W_3 + A_{23}$ and, proceeding in this manner, we finally arrive at

$$WW^T = (W_N W_N^T)^N + P(W_2, \ldots, W_N, W_2^T, \ldots, W_N^T), \tag{12}$$

where $P(W_2, \ldots, W_N, W_2^T, \ldots, W_N^T)$ is a polynomial in $W_2, \ldots, W_N, W_2^T, \ldots, W_N^T$ (with matrix coefficients) whose degree is at most $2N - 2$.

In the following, we denote by σ_N the maximal singular value of W_N. Thus

$$\sigma_N^{2N} \leq \|(W_N W_N^T)^N\|_F \leq \|WW^T\|_F + \|P(W_2, \ldots, W_N, W_2^T, \ldots, W_N^T)\|_F. \tag{13}$$

Since $\|W_N\|_F^2$ and $\|W_i\|_F^2$ differ only by a constant (depending on i), there are suitable constants a_i and b_i such that $\|W_i\|_F \leq a_i\sigma_N + b_i$ for all $i \in \{1, \ldots, N\}$. It follows that

$$\|P(W_2, \ldots, W_N, W_2^T, \ldots, W_N^T)\|_F \leq P_N(\sigma_N),$$

where P_N is a polynomial in one variable of degree at most $2N - 2$. Since the degree of P_N is strictly smaller than $2N$, there exists a constant C, which depends on the coefficients of P_N, such that $|P_N(x)| \leq \frac{1}{2}x^{2N} + C$ for all $x \geq 0$. Hence we obtain from (13)

$$\sigma_N^{2N} \leq B_N\|WW^T\|_F + \widetilde{B}_N, \tag{14}$$

and therefore also

$$\sigma_N \leq B_N'\|W\|_F^{1/N} + \widetilde{B}_N', \tag{15}$$

for suitable positive constants $B_N, \widetilde{B}_N, B_N', \widetilde{B}_N'$ (we can choose $B_N = 2$ by the discussion above). Since $\|W_i\|_F \leq a_i\sigma_N + b_i$, estimate (10) for $\|W_i\|_F$ follows.

The fact that all the $\|W_i\|_F$ are bounded now follows from the fact that $\|W\|_F$ is bounded as shown above together with estimate (10). This ensures the existence of solutions $W_i(t)$ (and hence $W(t)$) for all $t \geq 0$. The convergence of $(W_1, \ldots, W_N)$ to an equilibrium point (i.e., a critical point of L^N) now follows from Lojasiewicz's Theorem 4. $\qquad\square$

4. Riemannian gradient flows

Recall that in order to define a gradient flow, it is necessary to also specify the local geometry of the space. More precisely, suppose that a C^2 manifold $\mathcal{M}$ is given, on which a C^2-function $x \mapsto E(x) \in \mathbb{R}$ is defined for all $x \in \mathcal{M}$. Then the differential $dE(x)$ of E at the point x is a *co-tangent* vector, i.e., a linear map from the tangent space $T_x\mathcal{M}$ to $\mathbb{R}$. On the other hand, the derivative along any curve $t \mapsto \gamma(t) \in \mathcal{M}$ is a *tangent* vector. If now g_x denotes a Riemannian metric on $\mathcal{M}$ at x, then it is possible to associate to the differential $dE(x)$ a unique tangent vector $\nabla E(x)$, called the *gradient* of E at x, that satisfies

$$dE(x)v =: g_x(\nabla E(x), v) \quad \text{for all tangent vectors } v \in T_x\mathcal{M}.$$

It is the tangent vector $\nabla E(x)$ that enters in the definition of gradient flow $\dot{\gamma}(t) = -\nabla E(\gamma(t))$.

In this section, we are interested in minimizing the functional L^N introduced in (2) over the family of all matrices $W_1, \ldots, W_N$. This can be accomplished by considering the long-time limit of the gradient flow of L^N. Alternatively, we observe that we can equivalently lump all matrices together in the product $W := W_N \cdots W_1$ and minimize the functional L^1 defined in (4) over the set of all matrices W having this product form.

We consider the manifold $\mathcal{M}_k$ of real $d_y \times d_x$ matrices of rank $k \leq \min\{d_x, d_y\}$. We regard $\mathcal{M}_k$ as a submanifold of the manifold of all real $d_y \times d_x$ matrices, from which we inherit the structure of a differentiable manifold for $\mathcal{M}_k$. We denote by $T_W(\mathcal{M}_k)$ the tangential space of $\mathcal{M}_k$ at the point $W \in \mathcal{M}_k$. We have

$$T_W(\mathcal{M}_k) = \{WA + BW : A \in \mathbb{R}^{d_x \times d_x}, B \in \mathbb{R}^{d_y \times d_y}\}; \tag{16}$$

see [12, Proposition 4.1]. We will need the following result on the orthogonal projection onto the tangent space, which is well-known, cf. [3, Equation 9]. Below the notions *self-adjoint*, *positive definite*, and *orthogonal complement* are understood with respect to the Frobenius scalar product, which we denote by $\langle\ ,\ \rangle_F$. Recall that $\langle A, B\rangle_F = \text{tr}(AB^T)$.

Lemma 7. *Let $W \in \mathcal{M}_k$ with full singular value decomposition $W = USV^T$ and reduced singular decomposition $W = \bar{U}\Sigma\bar{V}^T$, where $U \in \mathbb{R}^{d_y \times d_y}$ and $V \in \mathbb{R}^{d_x \times d_x}$ are orthogonal and $\bar{U} \in \mathbb{R}^{d_y \times k}$ and $\bar{V} \in \mathbb{R}^{d_x \times k}$ are submatrices consisting of the first k columns of U and V, respectively. Let $Q_U = \bar{U}\bar{U}^T = UP_kU^T$ denote the orthogonal projection onto the range of $\bar{U}$, where $P_k = \text{diag}(1, \ldots, 1, 0, \ldots, 0)$ is the diagonal matrix with k ones on the diagonal, and likewise define $Q_V = \bar{V}\bar{V}^T = VP_kV^T$. Then the orthogonal projection $P_W : \mathbb{R}^{d_y \times d_x} \to T_W(\mathcal{M}_k)$ onto the tangent space at W is given by*

$$P_W(Z) = Q_UZ + ZQ_V - Q_UZQ_V \quad \text{for } Z \in \mathbb{R}^{d_y \times d_x}.$$

Proof. For convenience, we give a proof. For a matrix $Z = WA + BW \in T_W(\mathcal{M}_k)$, a simple computation using $\bar{U}^T\bar{U} = I_k = \bar{V}^T\bar{V}$ gives

$$P_W(Z) = \bar{U}\bar{U}^T(\bar{U}\Sigma\bar{V}^TA + B\bar{U}\Sigma\bar{V}^T) + (\bar{U}\Sigma\bar{V}^TA + B\bar{U}\Sigma\bar{V}^T)\bar{V}\bar{V}^T - \bar{U}\bar{U}^T(\bar{U}\Sigma\bar{V}^TA + B\bar{U}\Sigma\bar{V}^T)\bar{V}\bar{V}^T$$

$$= \bar{U}\Sigma\bar{V}^TA + B\bar{U}\Sigma\bar{V}^T = Z.$$

Moreover, for an arbitrary $Z \in \mathbb{R}^{d_y \times d_x}$ we have

$$P_W(Z) = \bar{U}\bar{U}^TZ(I_{d_x} - Q_V) + Z\bar{V}\bar{V}^T = \bar{U}\Sigma\bar{V}^T\bar{V}\Sigma^{-1}\bar{U}^TZ(I_{d_x} - Q_V) + Z\bar{V}\Sigma^{-1}\bar{U}^T\bar{U}\Sigma\bar{V}^T$$

$$= W\bar{V}\Sigma^{-1}\bar{U}^TZ(I_{d_x} - Q_V) + Z\bar{V}\Sigma^{-1}\bar{U}^TW$$

so that $P_W(Z) \in T_W(\mathcal{M}_k)$. We conclude that $P_W^2 = P_W$. Moreover, it is easy to verify that $\langle P_W(Z), Y\rangle_F = \langle Z, P_W(Y)\rangle_F$ for all $Z, Y \in \mathbb{R}^{d_y \times d_x}$ so that P_W is self-adjoint. Altogether, this proves the claim. $\square$

Inspired by [9], we use the operator $\mathcal{A}_W$ to define a Riemannian metric on $\mathcal{M}_k$. The following lemma can be seen as an extension of (a part of) claim 1 in the proof of [5, Theorem 1], where positive semidefiniteness of $\mathcal{A}_W$ is established.

Lemma 8. *For any given $W \in \mathbb{R}^{d_y \times d_x}$ let k be the rank of W, so that $W \in \mathcal{M}_k$. Let $N \geq 2$. Then the map $\mathcal{A}_W : \mathbb{R}^{d_y \times d_x} \to \mathbb{R}^{d_y \times d_x}$ defined in (8) is a self-adjoint endomorphism. Its image is $T_W(\mathcal{M}_k)$ and its kernel is (consequently) the orthogonal complement $T_W(\mathcal{M}_k)^\perp$ of $T_W(\mathcal{M}_k)$. The restriction of $\mathcal{A}_W$ to arguments $Z \in T_W(\mathcal{M}_k)$ defines a self-adjoint and positive definite endomorphism*

$$\bar{\mathcal{A}}_W : T_W(\mathcal{M}_k) \to T_W(\mathcal{M}_k).$$

In particular, $\bar{\mathcal{A}}_W$ is invertible and the inverse $\bar{\mathcal{A}}_W^{-1}$ is self-adjoint and positive definite as well.

Proof. We split the proof into four steps.

Step 1. It is clear that $\mathcal{A}_W$ defines an endomorphism of $\mathbb{R}^{d_y \times d_x}$. To see that it is self-adjoint, we calculate, for $Z_1, Z_2 \in \mathbb{R}^{d_y \times d_x}$,

$$\langle\mathcal{A}_W(Z_1), Z_2\rangle_F = \mathrm{tr}\left(\sum_{j=1}^N (WW^T)^{\frac{N-j}{N}} Z_1 (W^TW)^{\frac{j-1}{N}} Z_2^T\right) = \mathrm{tr}\left(\sum_{j=1}^N Z_1 (W^TW)^{\frac{j-1}{N}} Z_2^T (WW^T)^{\frac{N-j}{N}}\right)$$

$$= \mathrm{tr}\left(Z_1 \mathcal{A}_W(Z_2)^T\right) = \langle Z_1, \mathcal{A}_W(Z_2)\rangle_F.$$

We conclude that $\mathcal{A}_W$ is indeed self-adjoint.

Step 2. Next we show that the image of $\mathcal{A}_W$ lies in $T_W(\mathcal{M}_k)$; see (16). Let $W = \bar{U}\Sigma\bar{V}^T$ be a (reduced) singular value decomposition of W in the following form: $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_k) \in \mathbb{R}^{k \times k}$ is the diagonal matrix containing the non-zero singular values of W and the columns of $\bar{U} \in \mathbb{R}^{d_y \times k}$ and $\bar{V} \in \mathbb{R}^{d_x \times k}$ are orthonormal, so that $\bar{U}^T\bar{U} = I_k = \bar{V}^T\bar{V}$. For any index $1 \leq j < N$, we observe the identity

$$(WW^T)^{\frac{N-j}{N}} = \bar{U}\Sigma^2{}^{\frac{N-j}{N}}\bar{U}^T = \bar{U}\Sigma\bar{V}^T\bar{V}\Sigma^2{}^{\frac{N-j}{N}-1}\bar{U}^T = W(\bar{V}\Sigma^2{}^{\frac{N-j}{N}-1}\bar{U}^T). \tag{17}$$

Note that the second factor on the right-hand side of (17) is well-defined even though the exponent $2\frac{N-j}{N} - 1$ may be negative because the diagonal entries of Σ are all positive. Similarly, for $1 < j \leq N$ we find

$$(W^TW)^{\frac{j-1}{N}} = \bar{V}\Sigma^2{}^{\frac{j-1}{N}}\bar{V}^T = \bar{V}\Sigma^2{}^{\frac{j-1}{N}-1}(\bar{U}^T\bar{U})\Sigma\bar{V}^T = (\bar{V}\Sigma^2{}^{\frac{j-1}{N}-1}\bar{U}^T)W.$$

We observe that every term in the sum (8) is of the form WA or of the form BW for suitable $A \in \mathbb{R}^{d_x \times d_x}$ or $B \in \mathbb{R}^{d_y \times d_y}$. Hence $\mathcal{A}_W(Z) \in T_W(\mathcal{M}_k)$ for any $Z \in \mathbb{R}^{d_y \times d_x}$. It follows that the restriction of $\mathcal{A}_W$ to $T_W(\mathcal{M}_k)$, denoted by $\bar{\mathcal{A}}_W$, is a self-adjoint endomorphism. To prove that $\bar{\mathcal{A}}_W$ is injective, it therefore suffices to show that all eigenvalues are non-zero. Since $T_W(\mathcal{M}_k)$ is a finite-dimensional vector space, injectivity of $\bar{\mathcal{A}}_W$ then implies bijectivity.

Step 3. We show that $\bar{\mathcal{A}}_W$ is positive definite. For $Z \in T_W(\mathcal{M}_k)$, we need to establish that $\langle\mathcal{A}_W(Z), Z\rangle_F > 0$ if $Z \neq 0$. We will first show that for all $j \in \{1, \ldots, N\}$

$$\mathrm{tr}\left((WW^T)^{\frac{N-j}{N}} Z(W^TW)^{\frac{j-1}{N}} Z^T\right) \geq 0.$$

Let again $W = \bar{U}\Sigma\bar{V}^T$ be a (reduced) singular value decomposition of W as in Step 2. If $j = 1$, then

$$\text{tr}\left((WW^T)^{\frac{N-1}{N}}ZZ^T\right) = \text{tr}\left((\bar{U}\Sigma^{2\frac{N-1}{N}}\bar{U}^T)ZZ^T\right) = \text{tr}(R_1 R_1^T) \geq 0,$$

where $R_1 := \Sigma^{\frac{N-1}{N}}\bar{U}^T Z$. Similarly, for $j = N$ we get

$$\text{tr}\left(Z(W^T W)^{\frac{N-1}{N}}Z^T\right) = \text{tr}\left(Z(\bar{V}\Sigma^{2\frac{N-1}{N}}\bar{V}^T)Z^T\right) = \text{tr}(R_N R_N^T) \geq 0,$$

where $R_N := Z\bar{V}\Sigma^{\frac{N-1}{N}}$. Finally, if $1 < j < N$, then

$$\text{tr}\left((WW^T)^{\frac{N-j}{N}}Z(W^T W)^{\frac{j-1}{N}}Z^T\right) = \text{tr}\left((\bar{U}\Sigma^{2\frac{N-j}{N}}\bar{U}^T)Z(\bar{V}\Sigma^{2\frac{j-1}{N}}\bar{V}^T)Z^T\right) = \text{tr}(R_j R_j^T) \geq 0,$$

where $R_j := \Sigma^{\frac{N-j}{N}}\bar{U}^T Z\bar{V}\Sigma^{\frac{j-1}{N}}$. It follows that $\langle \mathcal{A}_W(Z), Z\rangle_F \geq 0$ for all $Z \in T_W(\mathcal{M}_k)$.

Suppose now that $\langle \mathcal{A}_W(Z), Z\rangle_F = 0$. Then $\|R_j\|_F^2 = \text{tr}(R_j R_j^T) = 0$, thus $R_j = 0$ for every $j \in \{1, \ldots, N\}$. Since $\Sigma \in \mathbb{R}^{k \times k}$ is invertible this implies for $j = 1$ that $\bar{U}^T Z = 0$ and for $j = N$ that $Z\bar{V} = 0$. By Lemma 7 we have

$$Z = P_W(Z) = \bar{U}\bar{U}^T Z + Z\bar{V}\bar{V}^T - \bar{U}\bar{U}^T Z\bar{V}\bar{V}^T = 0.$$

This proves that $\bar{\mathcal{A}}_W$ is strictly positive definite, therefore injective (bijective) as a map from $T_W(\mathcal{M}_k)$ to itself.

Step 4. It remains to prove that the kernel of $\mathcal{A}_W$ is the orthogonal complement of $T_W(\mathcal{M}_k)$. This follows from the general fact that for any self-adjoint endomorphism f of an inner product space, the kernel of f is the orthogonal complement of the image of the adjoint of f. $\qquad\square$

Definition 9. We introduce a Riemannian metric[1] g on the manifold $\mathcal{M}_k$ (for $k \leq \min\{d_x, d_y\}$) by

$$g_W(Z_1, Z_2) := \langle \bar{\mathcal{A}}_W^{-1}(Z_1), Z_2\rangle_F \tag{18}$$

for any $W \in \mathcal{M}_k$ and for all tangent vectors $Z_1, Z_2 \in T_W(\mathcal{M}_k)$.

By Lemma 8, the map g_W is well defined and defines indeed a scalar product on $T_W(\mathcal{M}_k)$. We provide explicit expressions for this scalar product in the next result.

Proposition 10. *For $N \geq 2$, the metric g on $\mathcal{M}_k$ defined in (18) satisfies*

$$g_W(Z_1, Z_2) = \frac{\sin(\pi/N)}{\pi}\int_0^\infty \text{tr}\left((tI_{d_y} + WW^T)^{-1}Z_1(tI_{d_x} + W^T W)^{-1}Z_2^T\right)t^{1/N}dt \tag{19}$$

$$= \frac{1}{N\Gamma(1 - 1/N)}\int_0^\infty \int_0^t \text{tr}\left(e^{-sWW^T}Z_1 e^{-(t-s)W^T W}Z_2^T\right)ds\, t^{-(1/N+1)}dt \tag{20}$$

for all $W \in \mathcal{M}_k$ and $Z_1, Z_2 \in T_W(\mathcal{M}_k)$, where Γ denotes the Gamma function.

In the case $N = 2$, we additionally have

$$g_W(Z_1, Z_2) = \int_0^\infty \text{tr}\left(e^{-t(WW^T)^{\frac{1}{2}}}Z_1 e^{-t(W^T W)^{\frac{1}{2}}}Z_2^T\right)dt. \tag{21}$$

Proof. The proof is postponed to Appendix A. $\qquad\square$

The next result states that the Riemannian metric is continuously differentiable as a function of $W \in \mathcal{M}_k$.

Proposition 11. *The metric g on $\mathcal{M}_k$ given by (18) is of class C^1.*

Proof. The proof uses the representation (19). The main step consists in showing that the directional derivates with respect to W of the corresponding integrand remain integrable so that Lebesgue's dominated convergence theorem can be applied to interchange integration and differentiation. The lengthy details are postponed to Appendix B. $\qquad\square$

---

[1]Our use of the term *Riemannian metric* does not imply any smoothness properties. Nevertheless, we will show in Proposition 11 that our metric g is of class C^1.

For any differentiable function $f\colon \mathbb{R}^{d_y \times d_x} \to \mathbb{R}$, any $W \in \mathcal{M}_k \subset \mathbb{R}^{d_y \times d_x}$, and any $Z \in T_W(\mathcal{M}_k)$, we have

$$g_W\big(\mathcal{A}_W(\nabla f(W)), Z\big) = \left\langle \bar{\mathcal{A}}_W^{-1}\big(\mathcal{A}_W(\nabla f(W))\big), Z \right\rangle_F = \langle \nabla f(W), Z \rangle_F = Df(W)[Z],$$

where Df denotes the differential of f (which can be computed from the derivative with respect to W). Note here that by Lemma 8, the two quantities $\bar{\mathcal{A}}_W^{-1}(\mathcal{A}_W(\nabla f(W)))$ and $\nabla f(W)$ differ only by an element in $T_W(\mathcal{M}_k)^\perp$, which is perpendicular to Z with respect to the Frobenius norm, as noticed above. This allows us to identify $\mathcal{A}_W(\nabla f(W))$ with the gradient of f with respect to the new metric g. We write

$$\mathcal{A}_W(\nabla f(W)) =: \nabla^g f(W). \tag{22}$$

In particular, we have for all $Z \in T_W(\mathcal{M}_k)$ that $g_W(\nabla^g f(W), Z) = Df(W)[Z]$. Let now $k \leq \min\{d_0, \ldots, d_N\}$ and recall that, in the balanced case, the evolution of the product $W = W_N \cdots W_1$ is given by (9).

We note that the solutions $W_1(t), \ldots, W_N(t)$ of the gradient flow (5) of L^N are unique (given initial values), since (5) obviously satisfies a local Lipschitz condition. Therefore the tuple $W_1(t), \ldots, W_N(t)$ gives rise to a well defined product $W(t) = W_N(t) \cdots W_1(t)$ which in the balanced case solves equation (9). However, due to the appearance of N-th roots in (9), it is unclear at the moment whether there are also other solutions of (9). The next proposition shows that in the balanced case (and for XX^T of full rank) the solution $W(t) = W_N(t) \cdots W_1(t)$ of (9) stays in $\mathcal{M}_k$ for all finite times t provided that $W(0) \in \mathcal{M}_k$.

Proposition 12. *Assume that XX^T has full rank and suppose that $W_1(t), \ldots, W_N(t)$ are solutions of the gradient flow (5) of L^N with balanced initial values $W_j(0)$ and define the product $W(t) := W_N(t) \cdots W_1(t)$. If $W(0)$ is contained in $\mathcal{M}_k$ for some $k \leq \min\{d_0, \ldots, d_N\}$ then $W(t)$ is contained in $\mathcal{M}_k$ for all $t \geq 0$.*

Proof. It follows from Theorem 5 that for any given $t_0 \in \mathbb{R}$ and initial values $W_1(t_0), \ldots, W_N(t_0)$, a (unique) solution $W_1(t), \ldots, W_N(t)$ of (5) is defined for all $t \geq t_0$. By the Picard–Lindelöf theorem, there also exists $\varepsilon > 0$ such that the solution $W_1(t), \ldots, W_N(t)$ is defined and unique on $(t_0 - \varepsilon, 0]$, hence on $(t_0 - \varepsilon, \infty)$.

Since the initial values $W_j(0)$ are balanced, for any $t \geq 0$ the matrices $W_1(t), \ldots, W_N(t)$ are balanced as well, cf. Lemma 2. It follows that for any $t \geq 0$, we have

$$W(t)W(t)^T = (W_N(t)W_N(t)^T)^N \quad \text{and} \quad W(t)^T W(t) = (W_1(t)^T W_1(t))^N,$$

compare also the proof of Theorem 1 in [5]. Both equations are directly verified for $N = 2$ and easily follow by induction for any $N \geq 2$.

Let now $P(t) = W_1(t)^T W_1(t)$ and $Q(t) = W_N(t)W_N(t)^T$. It follows that $P(t) = (W(t)^T W(t))^{1/N}$ and $Q(t) = (W(t)W(t)^T)^{1/N}$ for all $t \in [0, \infty)$. Using $\nabla_W L^1(W) = WXX^T - YX^T$ together with the explicit form of the gradient flow (5) for W_1 and W_N given by Lemma 2, point (1), and substituting $P = (W^T W)^{1/N}$ and $Q = (WW^T)^{1/N}$ in the flow equation (7) for W, we obtain the following system of differential equations for P, Q, W.

$$
\begin{aligned}
\dot{P} &= -W^T(WXX^T - YX^T) - (WXX^T - YX^T)^T W, \\
\dot{Q} &= -(WXX^T - YX^T)W^T - W(WXX^T - YX^T)^T, \\
\dot{W} &= -\sum_{j=1}^{N} Q^{N-j}(WXX^T - YX^T)P^{j-1}.
\end{aligned}
\tag{23}
$$

Since the right hand side of the system (23) is locally Lipschitz continuous in P, Q, W, it follows in particular that $W(t)$ (and also $P(t)$ and $Q(t)$) is uniquely determined by any initial values $P(t_0), Q(t_0), W(t_0)$.

Assume now that the claim of the proposition does not hold. Then there are $t_0, t_1 \in [0, \infty)$ with $\operatorname{rank}(W(t_1)) > \operatorname{rank}(W(t_0))$. Since $W(t) = W_N(t) \cdots W_1(t)$, it follows that

$$\min(d_0, \ldots, d_N) \geq \operatorname{rank}(W(t_1)) > \operatorname{rank}(W(t_0)).$$

We define $\ell = \operatorname{rank}(W(t_0))$ and distinguish the cases $\ell = 0$ and $\ell > 0$.

Case 1. $\ell = 0$. Then $W(t_0) = 0$ and hence also $W_1(t_0) = 0$. Due to balancedness it follows that $W_i(t_0) = 0$ for all $i \in \{1, \ldots, N\}$. But then it follows that $W_i(t) = 0$ for all $t \in [0, \infty)$ and for all $i \in \{1, \ldots, N\}$, hence also $W(t) = 0$ for all $t \in [0, \infty)$, so the rank of W is constant.

Case 2. $\ell > 0$. We assume first that $t_1 > t_0$ and will discuss the case $t_0 > t_1$ below.

We replace the first hidden layer (which has size d_1) by a new hidden layer of size ℓ (all other layer sizes remain as before) and define new initial values $\tilde{W}_1, \ldots, \tilde{W}_N$ (at t_0) for our new layer sizes in such a way that $\tilde{W}_1, \ldots, \tilde{W}_N$ are balanced and $\tilde{W} := \tilde{W}_N \cdots \tilde{W}_1 = W(t_0)$ and $\tilde{P} := \tilde{W}_1^T \tilde{W}_1 = W_1(t_0)^T W_1(t_0)$ and $\tilde{Q} := \tilde{W}_N \tilde{W}_N^T = W_N(t_0) W_N(t_0)^T$. For $t \in [t_0, \infty)$, let $\tilde{W}_1(t), \ldots, \tilde{W}_N(t)$ be the corresponding solutions of the gradient flow (5) for the new layer sizes with initial values at t_0 given by $\tilde{W}_1(t_0) = \tilde{W}_1, \ldots, \tilde{W}_N(t_0) = \tilde{W}_N$. Similarly, let $\tilde{W}(t) = \tilde{W}_N(t) \cdots \tilde{W}_1(t)$. Assuming that we can construct $\tilde{W}_1, \ldots, \tilde{W}_N$ as above, it follows in particular that $\tilde{W}(t) = W(t)$ for all $t \in [t_0, \infty)$, since, as discussed before, $W(t)$ is uniquely determined by $P(t_0) = \tilde{P}, Q(t_0) = \tilde{Q}, W(t_0) = \tilde{W}$ for all $t \in [0, \infty)$. But our new minimal layer size is ℓ, so it follows that the product $\tilde{W}(t) = \tilde{W}_N(t) \cdots \tilde{W}_1(t)$ has rank at most ℓ for any $t \in [t_0, \infty)$. In particular, $\operatorname{rank}(W(t_1)) = \operatorname{rank}(\tilde{W}(t_1)) \leq \ell$. This contradicts our assumption $\operatorname{rank}(W(t_1)) > \operatorname{rank}(W(t_0)) = \ell$.

Assume now that $t_0 > t_1$. Here we cannot directly argue as above since *backward in time* we only have *local* existence of solutions of (5). However, since the set $\{W \in \mathbb{R}^{d_y \times d_x} : \operatorname{rank}(W) < \operatorname{rank}(W(t_1))\}$ is closed in $\mathbb{R}^{d_y \times d_x}$, it follows that the set $\{t \geq t_1 : \operatorname{rank}(W(t)) < \operatorname{rank}(W(t_1))\}$ has a minimum τ_0, which is larger than t_1. Then for any $\varepsilon > 0$, there is a $\tau_1 \in (\tau_0 - \varepsilon, \tau_0)$ with $\operatorname{rank}(W(\tau_1)) > \operatorname{rank}(W(\tau_0))$.

Now replace t_0 by τ_0, define as before $\ell = \operatorname{rank}(W(t_0)) = \operatorname{rank}(W(\tau_0))$ and assume that we can construct $\tilde{W}_1, \ldots, \tilde{W}_N$ as above. Then for some $\varepsilon > 0$, the flows $\tilde{W}_1(t), \ldots, \tilde{W}_N(t)$ solving the gradient flow (5) for the new layer sizes and with initial values at t_0 given by $\tilde{W}_1(t_0) = \tilde{W}_1, \ldots, \tilde{W}_N(t_0) = \tilde{W}_N$ are defined (and balanced) on the interval $(t_0 - \varepsilon, \infty)$. Next we replace t_1 by a suitable $\tau_1 \in (\tau_0 - \varepsilon, \tau_0)$ with $\operatorname{rank}(W(\tau_1)) > \operatorname{rank}(W(\tau_0)) = \ell$. Now we can argue as above: On the one hand, the rank of $\tilde{W}(t_1) = \tilde{W}_N(t_1) \cdots \tilde{W}_1(t_1)$ is at most ℓ, on the other hand, we have $\tilde{W}(t_1) = W(t_1)$, so the rank of $W(t_1)$ is also at most ℓ, giving the desired contradiction.

It remains to construct $\tilde{W}_1, \ldots, \tilde{W}_N$ as announced. First, we introduce some notation. Let $\tilde{d}_1 = \ell$ and for $j \in \{0, \ldots, N\} \setminus \{1\}$ let $\tilde{d}_j = d_j$. (Thus the $\tilde{d}_j$ are our new layer sizes.) Given integers $a, b \geq \ell$ and $c_1, \ldots, c_\ell \in \mathbb{R}$, we denote by $S_{a,b}(c_1, \ldots, c_\ell) \in \mathbb{R}^{a \times b}$ the $a \times b$ diagonal matrix whose first ℓ diagonal entries are $c_1, \ldots, c_\ell$ and whose remaining entries are all equal to 0.

Now write $W := W(t_0) = USV^T$, where $U \in O(d_y) = O(d_N)$ and $V \in O(d_x) = O(d_0)$ and $S = S_{d_N, d_0}(\sigma_1, \ldots, \sigma_\ell)$, where $\sigma_1 \geq \ldots \geq \sigma_\ell > 0$. Let $W_N = W_N(t_0)$ and $W_1 = W_1(t_0)$. Then since $W^T W = (W_1^T W_1)^N$, we can write $W_1 = U_1 S_{d_1, d_0}(\sigma_1^{1/N}, \ldots, \sigma_\ell^{1/N}) V^T$ for some $U_1 \in O(d_1)$. Similarly, since $W W^T = (W_N W_N^T)^N$, we have $W_N = U S_{d_N, d_{N-1}}(\sigma_1^{1/N}, \ldots, \sigma_\ell^{1/N}) V_N^T$ for some $V_N \in O(d_{N-1})$. Define now $\tilde{W}_1 = S_{\tilde{d}_1, \tilde{d}_0}(\sigma_1^{1/N}, \ldots, \sigma_\ell^{1/N}) V^T$ and $\tilde{W}_N = U S_{\tilde{d}_N, \tilde{d}_{N-1}}(\sigma_1^{1/N}, \ldots, \sigma_\ell^{1/N})$ and, for $j \in \{2, \ldots, N-1\}$, $\tilde{W}_j = S_{\tilde{d}_j, \tilde{d}_{j-1}}(\sigma_1^{1/N}, \ldots, \sigma_\ell^{1/N})$. Note that this construction is possible since $\min(\tilde{d}_0, \ldots, \tilde{d}_N) = \ell$. (Compare [4, Section 3.3] for a similar construction of balanced initial conditions.) Then obviously, the $\tilde{W}_i$ are indeed balanced, and we have $\tilde{W}_1^T \tilde{W}_1 = W_1^T W_1$ and $\tilde{W}_N \tilde{W}_N^T = W_N W_N^T$ and $\tilde{W}_N \cdots \tilde{W}_1 = W$. This ends the proof. $\qquad\square$

Remark 13. *Assume again the situation of Proposition 12. Then in the limit $t \to \infty$, the rank of W still cannot increase, i.e., if $W(0)$ has rank k then the rank of $\lim\limits_{t \to \infty} W(t)$ is at most k. This follows from Proposition 12 together with the fact that the set of matrices of rank at most k is closed in $\mathbb{R}^{d_y \times d_x}$. However, it can happen that the rank of $\lim\limits_{t \to \infty} W(t)$ is strictly smaller than k, see Remark 42 for an explicit example.*

Remark 14. *Proposition 12 may fail if the initial values $W_j(0)$, $j = 1, \ldots, N$, are not balanced, i.e., the rank of $W(t)$ may then drop or increase in finite time. An example for such behaviour can be easily given in the case $N = 2$, $d_0 = d_1 = d_2 = 1$, $X = Y = 1$. Choosing $W_1(0) = 0$ and $W_2(0) = 1$ gives $W(0) = 0$. Moreover $\frac{d}{dt} W_1(0) = W_2(0) = 1$ and $\frac{d}{dt} W_2(0) = W_1(0) = 0$. This means that for $t \neq 0$ and $|t|$ sufficiently small we have $W_1(t) \neq 0$ and $W_2(t) \neq 0$, hence $\operatorname{rank} W(t) = 1$. In other words, for small enough $\varepsilon > 0$, when moving with t from $-\epsilon$ to 0 the rank of $W(t)$ drops from 1 to 0 and when continuing from $t = 0$ to $t = \epsilon$ the rank increases again to 1.*

The statements of Lemma 15 and Corollary 16 below are probably well known, but we include them here for completeness.

Lemma 15. *Let $\mathcal{M}$ be a C^2-manifold which carries a Riemannian metric g of class C^1 and let $L : \mathcal{M} \to \mathbb{R}$ be a C^2-map. Then $-\nabla^g(L)$ is a C^1-vector field.*

Proof. In local coordinates, we have $-\nabla^g(L) = -\sum_{i,j} g^{i,j} \frac{\partial L}{\partial x_i} \frac{\partial}{\partial x_j}$, compare [6, Lemma 4.3]. Since by assumption the matrix with entries $g_{i,j}$ is C^1, also the inverse matrix $(g^{i,j})_{i,j}$ is C^1. Since also by assumption L is a C^2-map, the partial derivatives $\frac{\partial L}{\partial x_i}$ are C^1. It follows that $-\nabla^g(L)$ is indeed a C^1-vector field. $\square$

Corollary 16. *In the situation of Lemma 15, for any $x_0 \in \mathcal{M}$, there is a unique maximal integral curve $\phi : J \to M$ with $\phi(0) = x_0$ and*

$$\dot{\phi}(t) = -\nabla^g(L(\phi(t))) \; \forall t \in J.$$

Here maximal means that the interval J is the maximal open interval containing 0 with this property.

Proof. This follows from Lemma 15 together with Theorem 44 in appendix C. For the existence of J, see also appendix C or directly [15, Section IV, §2]. $\square$

Corollary 17. *Suppose that XX^T has full rank and that $W_1(t), \ldots, W_N(t)$ are solutions of the gradient flow (5) of L^N, with initial values $W_j(0)$ that are balanced; recall Definition 1. Define the product $W(t) := W_N(t) \cdots W_1(t)$. If $W(0)$ is contained in $\mathcal{M}_k$ (i.e., has rank k), then $W(t)$ solves for all $t \in [0, \infty)$ the gradient flow equation*

$$\dot{W} = -\nabla^g L^1(W) \tag{24}$$

on $\mathcal{M}_k$, where ∇^g denotes the Riemannian gradient of L^1 with respect to the metric g on $\mathcal{M}_k$ defined in (18). Further this is the only solution of (24) in $\mathcal{M}_k$.

Proof. Proposition 12 shows that $W(t) \in \mathcal{M}_k$ for all $t \in [0, \infty)$. Lemma 8 and the discussion below Proposition 11 show that $W(t)$ solves indeed equation (24) with the particular choice of g as in (18) as the metric. (Note that then (24) is a reformulation of (9).) By Corollary 16 there are no other solutions in $\mathcal{M}_k$. $\square$

Theorem 18. *Assume that XX^T has full rank and let $N \geq 2$. Then for any initialization $W(0) \in \mathbb{R}^{d_y \times d_x}$, denoting by k the rank of $W(0)$, there is a uniquely defined flow $W(t)$ on $\mathcal{M}_k$ for $t \in [0, \infty)$ which satisfies (24).*

Proof. Any $W(0) \in \mathcal{M}_k$ can be written as a product $W(0) = W_N(0) \cdots W_1(0)$ for suitable balanced $W_i(0) \in \mathbb{R}^{d_i \times d_{i-1}}$, $i = 1, \ldots N$, where $d_0 = d_x$ and $d_N = d_y$ and the remaining d_i (i.e. for $i \in \{1, \ldots, N-1\}$) are arbitrary integers greater than or equal to k; compare the proof of Proposition 12 or [4, Section 3.3]. If $W_1(t), \ldots, W_N(t)$ satisfy equation (5) with initial values $W_1(0), \ldots, W_N(0)$, then $W(t)$ solves (24) on $\mathcal{M}_k$; see Corollary 17. Since the $W_i(t)$ are defined for all $t \geq 0$ by Theorem 5, the claim follows (see again Corollary 17 for the fact that W is well defined, i.e., there are no other solutions on $\mathcal{M}_k$). $\square$

Remark 19. *Our Riemannian metric is (in the limit $N \to \infty$) similar to the Bogoliubov inner product of quantum statistical mechanics (when replacing $\mathcal{A}_W^{-1}$ with $\mathcal{A}_W$), which is defined on the manifold of positive definite matrices; see [9].*

5. LINEAR AUTOENCODERS WITH ONE HIDDEN LAYER

In this section we consider linear autoencoders with one hidden layer in the symmetric case, i.e., we assume $Y = X$ and $N = 2$ and we impose that $W_2 = W_1^T$. The nonsymmetric case with one hidden layer will be discussed in Appendix D.

For $V := W_2 = W_1^T \in \mathbb{R}^{d \times r}$ (where we write d for $d_x = d_y$ and r for d_1), let

$$E(V) = L^2(V^T, V) = \frac{1}{2} \|X - VV^T X\|_F^2.$$

We consider the gradient flow:

$$\dot{V} = -\nabla E(V), \quad V(0) = V_0, \tag{25}$$

where we assume that $V_0^T V_0 = I_r$. Computing the gradient of E gives

$$\nabla E(V) = -(I_d - VV^T) XX^T V - XX^T (I_d - VV^T) V.$$

Thus the gradient flow for V is given by

$$\dot{V} = (I_d - VV^T)XX^TV + XX^T(I_d - VV^T)V, \qquad V(0) = V_0, \ V_0^TV_0 = I_r. \tag{26}$$

This can be analyzed using results by Helmke, Moore, and Yan on Oja's flow [23].

Theorem 20. *(1) The flow (26) has a unique solution on the interval $[0, \infty)$.*
  (2) $V(t)^TV(t) = I_r$ for all $t \geq 0$.
  (3) The limit $\overline{V} = \lim_{t \to \infty} V(t)$ exists and it is an equilibrium.
  (4) The convergence is exponential: There are positive constants c_1, c_2 such that

$$\|V(t) - \overline{V}\|_F \leq c_1 e^{-c_2 t}$$

  for all $t \geq 0$.
  (5) The equilibrium points of the flow (26) are precisely the matrices of the form

$$\overline{V} = (v_1 | \ldots | v_r)\,Q,$$

  where $v_1, \ldots, v_r$ are orthonormal eigenvectors of XX^T and Q is an orthogonal $r \times r$-matrix.

Proof. In [23] it is shown that Oja's flow given by

$$\dot{V} = (I_d - VV^T)XX^TV$$

satisfies all the claims in the proposition provided that $V(0)^TV(0) = I_r$. In particular, by [23, Corollary 2.2], all $V(t)$ in any solution of Oja's flow with $V(0)^TV(0) = I_r$ fulfill $V(t)^TV(t) = I_r$. It follows that under the initial condition $V(0)^TV(0) = I_r$ the flow (26) is identical to Oja's flow because the term $XX^T(I_d - VV^T)V$ then vanishes for all t if V is a solution to Oja's flow.

Hence, (2) follows from [23, Corollary 2.2]. In [23, Theorem 2.1] an existence and uniqueness result on $[0, \infty)$ is shown for Oja's flow and thus implies (1). Statements (3) and (4) follow from [23, Theorem 3.1] (which states that the solution to Oja's flow exponentially converges to an equilibrium point). Point (5) follows from [23, Corollary 4.1] (which shows that the equilibrium points V of Oja's flow satisfying $V^TV = I_r$ are of the claimed form). $\qquad\square$

Remark 21. *Choosing $v_1, \ldots, v_r$ orthonormal eigenvectors corresponding to the largest r eigenvalues of XX^T, we obtain (for varying Q) precisely the possible solutions for the matrix V in the PCA-problem.*

In order to make this more precise and to see this claim, we recall the PCA-Theorem, cf. [18]. Given: $x_1, \ldots, x_m \in \mathbb{R}^d$ and $1 \leq r \leq d$, we consider the following problem: Find $v_1, \ldots, v_r \in \mathbb{R}^d$ orthonormal and $h_1, \ldots, h_m \in \mathbb{R}^r$ such that

$$\mathcal{L}(V; h_1 \ldots, h_m) := \frac{1}{m}\sum_i \|x_i - Vh_i\|_2^2 \tag{27}$$

is minimal. (Here $V = (v_1 | \ldots | v_r) \in \mathbb{R}^{d \times r}$.)

Theorem 22 (PCA-Theorem [18]). *A minimizer of (27) is obtained by choosing $v_1, \ldots, v_r$ as orthonormal eigenvectors corresponding to the r largest eigenvalues of $\sum_i x_i x_i^T = XX^T$ and $h_i = V^T x_i$.*
  The other possible solutions for V are of the form $V = (v_1 | \ldots | v_r)\,Q$, where $v_1, \ldots, v_r$ are chosen as above and Q is an orthogonal $r \times r$-matrix. Again $h_i = V^T x_i$.

Let $\lambda_1 \geq \ldots \geq \lambda_d$ be the eigenvalues of XX^T and let $v_1, \ldots, v_d$ be corresponding orthonormal eigenvectors.

Theorem 23. *Assume that XX^T has full rank and that $\lambda_r > \lambda_{r+1}$. Then $\lim_{t \to \infty} V(t) = (v_1 | \ldots | v_r)\,Q$ for some orthogonal Q if and only if $V_0^T(v_1 | \ldots | v_r)$ has rank r.*

Proof. This follows from [23, Theorem 5.1] (where an analogous statement for Oja's flow is made) together with [23, Corollary 2.1]. $\qquad\square$

Corollary 24. *Under the assumptions of Theorem 23, for almost all initial conditions (w.r.t. the Lebesgue measure), the flow converges to an optimal equilibrium, i.e., one of the form $V = (v_1 | \ldots | v_r)\,Q$ in the notation of Theorem 23.*

Proof. This follows from Theorem 23, cf. also the analogous [23, Corollary 5.1]. □

In Section 6 we extend this result to autoencoders with $N > 2$ layers using a more abstract approach. The following theorem shows that the optimal equilibria are the only stable equilibria:

Theorem 25. *Assume* $V = (v_{i_1} | \dots | v_{i_r}) Q$, *where the orthonormal eigenvectors* $v_{i_1}, \dots, v_{i_r}$ *are not eigenvectors corresponding to the largest* r *eigenvalues of* XX^T. *Then in any neighborhood of* V *there is a matrix* $\widetilde{V}$ *with* $E(\widetilde{V}) < E(V)$ *(and* $\widetilde{V}^T \widetilde{V} = I_r$).

Proof. Let v_{i_j} be one of the eigenvectors $v_{i_1}, \dots, v_{i_r}$ whose eigenvalue does not belong to the r largest eigenvalues of XX^T. Let v be an eigenvector of XX^T of unit length which is orthogonal to the eigenvectors $v_{i_1}, \dots, v_{i_r}$ and whose eigenvalue λ belongs to the r largest eigenvalues of XX^T. Now for any $\varepsilon \in [0,1]$ consider $v_{i_j}(\varepsilon) := \varepsilon v + \sqrt{1 - \varepsilon^2} v_{i_j}$. Then $V(\varepsilon) := (v_{i_1} | \dots | v_{i_j}(\varepsilon) | \dots | v_{i_r}) Q$ satisfies $E(V(\varepsilon)) < E(V)$ for $\varepsilon \in (0,1]$ and $V(\varepsilon)^T V(\varepsilon) = I_r$. To see that indeed $E(V(\varepsilon)) < E(V)$, we compute $E(V) = \frac{1}{2} \|X - VV^T X\|_F^2 = \frac{1}{2} \operatorname{tr}(XX^T - XX^T VV^T)$ and $E(V(\varepsilon)) = \frac{1}{2} \operatorname{tr}(XX^T - XX^T V(\varepsilon)V(\varepsilon)^T)$. Writing $XX^T v_{i_k} = \lambda_{i_k} v_{i_k}$, we note that $\operatorname{tr}(XX^T VV^T) = \sum_{k=1}^r \lambda_{i_k}$ and $\operatorname{tr}(XX^T V(\varepsilon)V(\varepsilon)^T) = \varepsilon^2 \lambda + (1 - \varepsilon^2)\lambda_{i_j} + \sum_{k=1, k \neq j}^r \lambda_{i_k}$. Since $\lambda > \lambda_{i_j}$, the claim follows. □

6. Avoiding saddle points

In Section 3 we have proven convergence of the gradient flow (5) a to critical point of L^N. (Together with Proposition 34 below, this also implies that the product W converges to a critical point of L^1 restricted to $\mathcal{M}_k$ for some $k \leq r$.) Since we will remain in a saddle point forever if the initial point is a saddle point, the best we can hope for is convergence to global optima for almost all initial points (as in Corollary 24 for the particular autoencoder case with $N = 2$).

We will indeed establish such a result for both L^N and L^1 restricted to $\mathcal{M}_r$ in the autoencoder case. We note, however, that we can only ensure that the limit corresponds to an optimal point for L^1 restricted to $\mathcal{M}_k$ for some $k \leq r$ for almost all initialization. We conjecture $k = r$ (for almost all initializations), but this remains open for now.

We proceed by showing a general result on the avoidance of saddle points by extending the main result of [16] from gradient descent to gradient flows. A crucial ingredient is the notion of a strict saddle point. The application of the general abstract result to our scenario then requires to analyze the saddle points.

6.1. **Strict saddle points.** We start with the definition of a strict saddle point of a function on the Euclidean space $\mathbb{R}^d$.

Definition 26. Let $f : \Omega \to \mathbb{R}$ be a twice continuously differentiable function on an open domain $\Omega \subset \mathbb{R}^d$. A critical point $x_0 \in \Omega$ is called a strict saddle point if the Hessian $Hf(x_0)$ has a negative eigenvalue.

Intuitively, the function f possesses a direction of descent at a strict saddle point. Note that our definition also includes local maxima, which does not pose problems for our purposes.

Let us extend the notion of strict saddle points to functions on Riemannian manifolds $(\mathcal{M}, g)$. To this end, we first introduce the Riemannian Hessian of a C^2-function f on $\mathcal{M}$. Denoting by ∇ be the Riemannian connection (Levi-Civita connection) on $(\mathcal{M}, g)$ the Riemannian Hessian of f at $x \in \mathcal{M}$ is the linear mapping $\operatorname{Hess} f(x) : T_x \mathcal{M} \to T_x \mathcal{M}$ defined by

$$\operatorname{Hess}^g f(x)[\xi] := \nabla_\xi \nabla^g f.$$

Of course, if $(\mathcal{M}, g)$ is Euclidean, then this definition can be identified with the standard definition of the Hessian. Moreover, if $x \in \mathcal{M}$ is a critical point of f, i.e., $\nabla^g f(x) = 0$, then the Hessian $\operatorname{Hess}^g f(x)$ is independent of the choice of the connection. Below, we will need the following chain type rule for curves γ on $\mathcal{M}$, see e.g. [19, Eq. (3.1)],

$$\frac{d^2}{dt^2} f(\gamma(t)) = g\left(\dot{\gamma}(t), \operatorname{Hess}^g f(\gamma(t))[\dot{\gamma}(t)]\right) + g\left(\frac{D}{dt}\dot{\gamma}(t), \nabla^g f(\gamma(t))\right), \tag{28}$$

where $\frac{D}{dt}\dot{\gamma}(t)$ is related to the Riemannian connection that is used to define the Hessian, see [2, Section 5.4]. We refer to [2] for more details on the Riemannian Hessian.

Definition 27. Let $(\mathcal{M}, g)$ be a Riemannian manifold with Levi-Civita connection ∇ and let $f : \mathcal{M} \to \mathbb{R}$ be a twice continuously differentiable function. A critical point $x_0 \in \mathcal{M}$, i.e., $\nabla^g f(x_0) = 0$ is called a strict saddle point if Hess $f(x)$ has a negative eigenvalue. We denote the set of all strict saddles of f by $\mathcal{X} = \mathcal{X}(f)$. We say that f has the strict saddle point property, if all critical points of f that are not local minima are strict saddle points.

Note that our definition of strict saddle points includes local maxima, which is fine for our purposes.

6.2. **Flows avoid strict saddle points almost surely.** We now prove a general result that gradient flows on a Riemannian manifold $(\mathcal{M}, g)$ for functions with the strict saddle point property avoid saddle point for almost all initial values. This result extends the main result of [16] from time discrete systems to continuous flows and should be of independent interest.

For a twice continuously differentiable function $L : \mathcal{M} \to \mathbb{R}$, we consider the Riemannian gradient flow

$$\frac{d}{dt} \phi(t) = -\nabla^g L(\phi(t)), \quad \phi(0) = x_0 \in \mathcal{M}, \tag{29}$$

where ∇^g denotes the Riemannian gradient. When emphasizing the dependence on x_0, we write

$$\psi_t(x_0) = \phi(t), \tag{30}$$

where $\phi(t)$ is the solution to (29) with initial condition x_0.

Sets of measure zero on $\mathcal{M}$ (as used in the next theorem) can be defined using push forwards of the Lebesgue measure on charts of the manifold $\mathcal{M}$.

Theorem 28. *Let $L : \mathcal{M} \to \mathbb{R}$ be a C^2-function on a second countable finite dimensional Riemannian manifold $(\mathcal{M}, g)$, where we assume that $\mathcal{M}$ is of class C^2 as a manifold and the metric g is of class C^1. Assume that $\psi_t(x_0)$ exists for all $x_0 \in \mathcal{M}$ and all $t \in [0, \infty)$. Then the set*

$$\mathcal{S}_L := \{x_0 \in \mathcal{M} : \lim_{t \to \infty} \psi_t(x_0) \in \mathcal{X} = \mathcal{X}(L)\}$$

of initial points such that the corresponding flow converges to a strict saddle point of L has measure zero.

The proof of this relies on the following result for iteration maps (e.g., gradient descent iterations) shown in [16].

Theorem 29. *Let $h : \mathcal{M} \to \mathcal{M}$ be a continuously differentiable function on a second countable differentiable finite-dimensional manifold such that $\det(Dh(x)) \neq 0$ for all $x \in \mathcal{M}$ (in particular, h is a local C^1 diffeomorphism). Let*

$$\mathcal{A}_h^* = \{x \in \mathcal{M} : h(x) = x, \max_j |\lambda_j(Dh(x))| > 1\},$$

where $\lambda_j(Dh(x))$ denote the eigenvalues of $Dh(x)$, and consider sequences with initial point $x_0 \in \mathcal{M}$, $x_k = h(x_{k-1})$, $k \in \mathbb{N}$. Then the set $\{x_0 \in \mathcal{M} : \lim_{k \to \infty} x_k \in \mathcal{A}_h^*\}$ has measure zero.

Proof of Theorem 28. By Lemma 15 and Theorem 45 in appendix C, the map

$$h \colon \mathcal{M} \to \mathcal{M}, \ x_0 \mapsto \psi_T(x_0)$$

defines a diffeomorphism of $\mathcal{M}$ onto an open subset of $\mathcal{M}$. In particular, $Dh = D\psi_T$ is non-singular, i.e. $\det(Dh(x)) \neq 0$ for all $x \in \mathcal{M}$.

Because of the semigroup property $\psi_{t+s}(x_0) = \psi_t(\psi_s(x_0))$ the sequence $x_k = \psi_{kT}(x_0)$, $k \in \mathbb{N}$, satisfies $x_k = h(x_{k-1})$ and $\lim_{t \to \infty} \psi_t(x_0) \in \mathcal{X}$ implies $\lim_{k \to \infty} x_k \in \mathcal{X}$.

By Theorem 29 the set

$$\{x_0 \in \mathcal{M} : \lim_{k \to \infty} \psi_{kT}(x_0) \in \mathcal{A}_{\psi_T}^*\}$$

has measure zero. We need to show that if $\bar{x}$ is a strict saddle point of L, then $\bar{x} \in \mathcal{A}_{\psi_T}^*$ for suitable (i.e., sufficiently small) $T > 0$. We will work with a sequence of parameters $T = \frac{1}{n}$ with $n \in \mathbb{N}$.

Let $\bar{x} \in \mathcal{X}(L)$ be a strict saddle point of L. If we choose local coordinates giving rise to an orthonormal basis with respect to the Riemannian metric at $\bar{x}$, then it follows from (29) that, for all $n \in \mathbb{N}$,

$$D\psi_{1/n}(\bar{x}) = I - \frac{1}{n} \text{Hess}^g L(\bar{x}) + o(1/n),$$

where $\lim_{t\to 0} o(t)/t = 0$. Compare also [6, Lemma 4.4] for the fact that we can identify here the differential of $\nabla^g L(\bar{x})$ with $\mathrm{Hess}^g\, L(\bar{x})$. (More precisely, it is shown in loc. cit. that the the differential of $\nabla^g L(\bar{x})$ coincides with the matrix $(\frac{\partial^2 L(\bar{x})}{\partial x_i \partial x_j})_{i,j}$ at the critical point $\bar{x}$, if we assume that the local coordinates give rise to an orthonormal basis at this point. Using again that $\bar{x}$ is a critical point, we see that this matrix is the Riemannian Hessian at $\bar{x}$ in our local coordinates.) Since $\bar{x}$ is a strict saddle point of L, the matrix $\mathrm{Hess}^g\, L(\bar{x})$ has at least one strictly negative eigenvalue. It follows that there exists $N \in \mathbb{N}$ such that for all $n \geq N$ the differential $D\psi_{1/n}(\bar{x})$ has an eigenvalue larger than 1. Hence $\bar{x} \in \mathcal{A}^*_{\psi_{1/n}}$ and

$$\{x_0 \in \mathcal{M} : \lim_{t\to\infty} \psi_t(x_0) = \bar{x}\} \subset \{x_0 \in \mathcal{M} : \lim_{k\to\infty} \psi_{k/n}(x_0) = \bar{x}\} \subset \{x_0 \in \mathcal{M} : \lim_{t\to\infty} \psi_{k/n} \in \mathcal{A}^*_{\psi_{1/n}}\}$$

for all $n \geq N$. It follows that

$$\{x_0 \in \mathcal{M} : \lim_{t\to\infty} \psi_t(x_0) \in \mathfrak{X}(L)\} \subset \bigcup_{n\in\mathbb{N}} \{x_0 \in \mathcal{M} : \lim_{k\to\infty} \psi_{k/n}(x_0) \in \mathcal{A}^*_{\psi_{1/n}}\}.$$

The set on the right hand side is a countable union of null sets and therefore has measure zero. This implies the claim of the theorem and the proof is completed. $\square$

Remark 30. *The proof of Theorem 29 uses the center and stable manifold theorem, see, e.g., [20, Chapter 5, Theorem III.7]. If the absolute eigenvalues of $Dh(x)$ are all different from 1, i.e., if all eigenvalues of the Hessian $\mathrm{Hess}^g\, f(x)$ are different from 0 at a saddle point x, then slightly stronger conclusions may be drawn, including the speed at which the flow moves away from saddle points. We will not elaborate on this point here.*

6.3. **The strict saddle point property for L^1 on $\mathcal{M}_r$.** In this section we establish the strict saddle point property of L^1 on $\mathcal{M}_k$ by showing that the Riemannian Hessian $\mathrm{Hess}\, L^1$ at all critical points that are not a global minimizer has a strictly negative eigenvalue. We assume that XX^T has full rank $d_x = d_0$ and start with an analysis of the critical points. We first recall the following result of Kawaguchi [13].

Theorem 31. [13, Theorem 2.3] *Assume that XX^T and XY^T are of full rank with $d_y \leq d_x$ and that the matrix $YX^T(XX^T)^{-1}XY^T$ has d_y distinct eigenvalues. Let r be the minimum of the d_i. Then the loss function $L^N(W_1, \ldots, W_N)$ has the following properties.*

(1) It is non-convex and non-concave.
(2) Every local minimum is a global minimum.
(3) Every critical point that is not a global minimum is a saddle point.
(4) If $W_{N-1} \cdots W_2$ has rank r then the Hessian at any saddle point has at least one negative eigenvalue.

Below we will remove the assumption that XY^T has full rank and that $YX^T(XX^T)^{-1}XY^T$ has distinct eigenvalues. Moreover, we will give more precise information on the strict saddle points.

The following matrix, which is completely determined by the given matrices X, Y that define L^1 and L^N (see (2) and (4)), will play a central role in our discussion. We define

$$Q := YX^T(XX^T)^{-\frac{1}{2}} \tag{31}$$

and let $q := \mathrm{rank}(Q)$ be its rank. We will use a reduced singular value decomposition

$$Q = U\Sigma V^T = \sum_{i=1}^{q} \sigma_i u_i v_i^T,$$

of Q, where $\sigma_1 \geq \ldots \geq \sigma_q > 0$ are the singular values of Q and $U \in \mathbb{R}^{d_y \times q}$, $V \in \mathbb{R}^{d_x \times q}$ have orthonormal columns $u_1, \ldots, u_q$ and $v_1, \ldots, v_q$, respectively. Clearly, it holds $q \leq n := \min\{d_x, d_y\}$.

Let $k \leq n$ and let g be an arbitrary Riemannian metric on the manifold $\mathcal{M}_k$ of all matrices in $\mathbb{R}^{d_y \times d_x}$ of rank k, for example it could be the metric induced by the standard metric on $\mathbb{R}^{d_y \times d_x}$ or the metric g introduced in Section 4 for some number of layers N.

The next statement is similar in spirit to Kawaguchi's result, Theorem 31, and follows from [22].

Proposition 32. *Let Q be defined by (31) and $q = \mathrm{rank}(Q)$.*

(1) *The critical points of L^1 on $\mathcal{M}_k$ are precisely the matrices of the form*

$$W = \sum_{j \in J} \sigma_j u_j v_j^T (XX^T)^{-\frac{1}{2}}, \tag{32}$$

where $J \subseteq \{1, \ldots, q\}$ consists of precisely k elements. Consequently, if $k > q$, then no such subset J can exist and therefore L^1 restricted to $\mathcal{M}_k$ cannot have any critical points.

(2) *If W is a critical point of L (so that W has the form (32)), then*

$$L^1(W) = \frac{1}{2} \left(\operatorname{tr}(YY^T) - \sum_{j \in J} \sigma_j^2 \right).$$

It follows that the critical point W is a global minimizer of L^1 on $\mathcal{M}_k$ if and only if

$$\{\sigma_j : j \in J\} = \{\sigma_1, \ldots, \sigma_k\},$$

i.e., the set J picks precisely the k largest singular values of Q. In particular, if $k = q$, then there cannot be any saddle points. Recall that there are no critical points if $k > q$ because of (1).

Proof. For $X = I$ see the proof of [22, Theorem 28]. To obtain the general case we observe that

$$L^1(W) = \frac{1}{2}\|WX - Y\|_F^2 = \frac{1}{2}\|W(XX^T)^{\frac{1}{2}} - YX^T(XX^T)^{-\frac{1}{2}}\|_F^2 + C = \frac{1}{2}\|W(XX^T)^{\frac{1}{2}} - Q\|_F^2 + C,$$

where $C := \frac{1}{2}\|Y\|_F^2 - \frac{1}{2}\|Q\|_F^2$ does not depend on W. Since XX^T has full rank, the map $W \mapsto W(XX^T)^{\frac{1}{2}}$ is invertible (on any $\mathcal{M}_k$). Therefore the critical points of the map $W \mapsto \frac{1}{2}\|W(XX^T)^{\frac{1}{2}} - Q\|_F^2$ restricted to $\mathcal{M}_k$ are just the critical points of the map $W \mapsto \frac{1}{2}\|W - Q\|_F^2$ (restricted to $\mathcal{M}_k$) multiplied by $(XX^T)^{-\frac{1}{2}}$. Now we substitute the results of [22, Theorem 28] on the critical points of the map $W \mapsto \frac{1}{2}\|W - Q\|_F^2$ restricted to $\mathcal{M}_k$ (which are just as claimed here in the case $X = I$) and we obtain the claim of the proposition. $\square$

Proposition 33. *The function L^1 on $\mathcal{M}_k$ for $k \leq n$ satisfies the strict saddle point property. More precisely, all critical points of L^1 on $\mathcal{M}_k$ except for the global minimizers are strict saddle points.*

Proof. If $k \geq q = \operatorname{rank}(Q)$ then there are no saddle points by Proposition 32 so that the statement holds trivially. Therefore, we assume $k < q$ from now on. By Proposition 32, it is enough to show that the Riemannian Hessian of L^1 has a negative eigenvalue at any point of the form

$$W = \sum_{j \in J} \sigma_j u_j v_j^T (XX^T)^{-\frac{1}{2}},$$

where $J \subseteq \{1, \ldots, q\}$ consists of precisely k elements and has the property that there is a $j_0 \in J$ with $\sigma_{j_0} < \sigma_k$. Thus there is also a $\sigma_{j_1} \in \{\sigma_1, \ldots, \sigma_k\}$ with $\sigma_{j_1} > \sigma_{j_0}$ and $j_1 \notin J$. We define for $t \in (-1, 1)$:

$$u_{j_0}(t) = tu_{j_1} + \sqrt{1 - t^2}u_{j_0} \text{ and } v_{j_0}(t) = tv_{j_1} + \sqrt{1 - t^2}v_{j_0}.$$

Now consider the curve $\gamma : (-1, 1) \to \mathcal{M}_k$ given by

$$\gamma(t) = \left(\sigma_{j_0} u_{j_0}(t) v_{j_0}(t)^T + \sum_{j \in J, j \neq j_0} \sigma_j u_j v_j^T \right) (XX^T)^{-\frac{1}{2}}.$$

Obviously we have $\gamma(0) = W$. We claim that it is enough to show that

$$\frac{d^2}{dt^2} L^1(\gamma(t)) \bigg|_{t=0} < 0.$$

Indeed, by (28) it holds (for any Riemannian metric g) that

$$\frac{d^2}{dt^2} L^1(\gamma(t)) = g\left(\dot{\gamma}(t), \operatorname{Hess}^g L^1(\gamma(t))\dot{\gamma}(t)\right) + g\left(\frac{D}{dt}\dot{\gamma}(t), \nabla^g L^1(\gamma(t))\right),$$

and since $\nabla^g L^1(\gamma(0)) = \nabla^g L^1(W) = 0$, it follows that $g\left(\dot{\gamma}(0), \operatorname{Hess}^g L^1(W)\dot{\gamma}(0)\right) < 0$ if $\frac{d^2}{dt^2} L^1(\gamma(t)) \big|_{t=0} < 0$ and hence that $\operatorname{Hess}^g L^1(W)$ has a negative eigenvalue in this case. (Note that $\operatorname{Hess}^g L^1(W)$ is self-adjoint

with respect to the scalar product g on $T_W(\mathcal{M}_k)$ and that it cannot be positive semidefinite (wrt. g) if $g\left(\dot\gamma(0), \mathrm{Hess}^g\, L^1(W)\dot\gamma(0)\right) < 0$, hence it has a negative eigenvalue in this case.)

We note that

$$L^1(\gamma(t)) = \frac{1}{2}\|\gamma(t)X - Y\|_F^2 = \frac{1}{2}\,\mathrm{tr}(\gamma(t)^T\gamma(t)XX^T - 2\gamma(t)XY^T + YY^T). \tag{33}$$

We compute

$$\left(\sigma_{j_0}v_{j_0}(t)u_{j_0}(t)^T + \sum_{j\in J, j\neq j_0}\sigma_j v_j u_j^T\right)\left(\sigma_{j_0}u_{j_0}(t)v_{j_0}(t)^T + \sum_{j\in J, j\neq j_0}\sigma_j u_j v_j^T\right)$$
$$= \sum_{j\in J\setminus\{j_0\}}\sigma_j^2 v_j v_j^T + \sigma_{j_0}^2 v_{j_0}(t)v_{j_0}(t)^T$$

so that

$$\mathrm{tr}(\gamma(t)^T\gamma(t)XX^T) = \mathrm{tr}\left((XX^T)^{-\frac{1}{2}}\left(\sum_{j\in J\setminus\{j_0\}}\sigma_j^2 v_j v_j^T + \sigma_{j_0}^2 v_{j_0}(t)v_{j_0}(t)^T\right)(XX^T)^{-\frac{1}{2}}XX^T\right)$$
$$= \sum_{j\in J}\sigma_j^2.$$

In particular, this expression is independent of t. Further,

$$\mathrm{tr}(-2\gamma(t)XY^T) = -2\,\mathrm{tr}\left(\left(\sigma_{j_0}u_{j_0}(t)v_{j_0}(t)^T + \sum_{j\in J, j\neq j_0}\sigma_j u_j v_j^T\right)(XX^T)^{-\frac{1}{2}}XY^T\right)$$
$$= -2\,\mathrm{tr}\left(\left(\sigma_{j_0}u_{j_0}(t)v_{j_0}(t)^T + \sum_{j\in J, j\neq j_0}\sigma_j u_j v_j^T\right)Q^T\right)$$
$$= -2\,\mathrm{tr}\left(\left(\sigma_{j_0}u_{j_0}(t)v_{j_0}(t)^T + \sum_{j\in J, j\neq j_0}\sigma_j u_j v_j^T\right)\sum_{j=1}^{q}\sigma_j v_j u_j^T\right)$$
$$= -2\,\mathrm{tr}\left(\sigma_{j_0}u_{j_0}(t)v_{j_0}(t)^T(\sigma_{j_0}v_{j_0}u_{j_0}^T + \sigma_{j_1}v_{j_1}u_{j_1}^T)\right) - 2\sum_{j\in J, j\neq j_0}\sigma_j^2$$
$$= -2(\sigma_{j_0}^2(1-t^2) + t^2\sigma_{j_0}\sigma_{j_1}) - 2\sum_{j\in J, j\neq j_0}\sigma_j^2$$
$$= 2t^2\sigma_{j_0}(\sigma_{j_0} - \sigma_{j_1}) - 2\sum_{j\in J}\sigma_j^2.$$

Together with equation (33) it follows that

$$\left.\frac{d^2}{dt^2}L^1(\gamma(t))\right|_{t=0} = 2\sigma_{j_0}(\sigma_{j_0} - \sigma_{j_1}) < 0.$$

This concludes the proof. $\qquad\square$

We note that a construction similar to the curve constructed in the preceding proof is considered in the proof of [22, Theorem 28]. However, it is not discussed there that this implies strictness of the saddle points.

6.4. **Strict saddle points of L^N.** Before discussing the strict saddle point property, let us first investigate the relation of the critical points of L^N and the ones of L^1 restricted to $\mathcal{M}_r$, where

$$r = \min\{d_0, d_1, \ldots, d_N\}.$$

Throughout this section we assume that XX^T has full rank.

Proposition 34. (a) Let $(W_1, \ldots, W_N)$ be a critical point of L^N. Define $W = W_N\cdots W_1$ and let $k := \mathrm{rank}(W) \leq r$. Then W is a critical point of L^1 restricted to $\mathcal{M}_k$.

(b) *Let W be a critical point of L^1 restricted to $\mathcal{M}_k$ for some $k \leq r$. Then there exists a tuple $(W_1, \ldots, W_N)$ with $W_N \cdots W_1 = W$ that is a critical point of L^N.*

Proof. For (a), let $Z \in T_W(\mathcal{M}_k)$ be arbitrary, i.e., $Z = WA + BW$ for some matrices $A \in \mathbb{R}^{d_x \times d_x}$ and $B \in \mathbb{R}^{d_y \times d_y}$. It suffices to show that for a curve $\gamma : \mathbb{R} \to \mathcal{M}_k$ with $\gamma(0) = W$ and $\dot{\gamma}(0) = Z$ that $\frac{d}{dt} L^1(\gamma(t))\big|_{t=0} = 0$. We choose the curve

$$\gamma(t) = (W_N + tV_n) \cdot W_{N-1} \cdots W_2 \cdot (W_1 + tV_1), \tag{34}$$

where $V_1 = W_1 A$ and $V_N = BW_N$. Then, indeed $\gamma(0) = W_N \cdots W_1 = W$ and $\dot{\gamma}(0) = W_N W_{N-1} \cdots W_1 A + BW_N \cdots W_1 = Z$. Next, observe that

$$\begin{aligned} \frac{d}{dt} L^1(\gamma(t))\bigg|_{t=0} &= \frac{d}{dt} L^N(W_1 + tV_1, W_2, \ldots, W_{N-1}, W_N + tV_N)\bigg|_{t=0} \\ &= \langle \nabla L^N(W_1, \ldots, W_N), (V_1, 0, \ldots, 0, V_N) \rangle = 0, \end{aligned}$$

since $(W_1, \ldots, W_N)$ is a critical point of L^N. Since Z was arbitrary, this shows (a).

For (b) we first note that by Lemma 2, for a point $(W_1, \ldots, W_N)$ to be a critical point of L^N, it suffices that

$$(WXX^T - YX^T)W_1^T = 0 \quad \text{and} \quad W_N^T(WXX^T - YX^T) = 0. \tag{35}$$

This is equivalent to

$$WXX^T W_1^T = Q(XX^T)^{\frac{1}{2}} W_1^T \quad \text{and} \quad W_N^T WXX^T = W_N^T Q(XX^T)^{\frac{1}{2}}. \tag{36}$$

Since W is a critical point of L^1 restricted to $\mathcal{M}_k$, we can write

$$W = \sum_{j \in J} \sigma_j u_j v_j^T (XX^T)^{-\frac{1}{2}},$$

where $J \subseteq \{1, \ldots, q\}$ consists of k elements, see Proposition 32. We write $J = \{j_{i_1}, \ldots, j_{i_k}\}$ to enumerate the elements in J. For $i, l \in \mathbb{N}$ with $i \leq l$ we denote by $e_i^{(l)}$ the i-th standard unit vector of dimension l (i.e., it has l entries, the i-th entry is 1 and all other entries are 0). Now we define

$$W_1 := \sum_{i=1}^{k} e_i^{(d_1)} v_{j_i}^T (XX^T)^{-\frac{1}{2}},$$

$$W_l := \sum_{i=1}^{k} e_i^{(d_l)} (e_i^{(d_{l-1})})^T \quad \text{for } l = 2, \ldots, N-1,$$

$$W_N := \sum_{i=1}^{k} \sigma_{j_i} u_{j_i} (e_i^{(d_{N-1})})^T.$$

Since $k \leq r$ this is well defined and one easily checks that

$$W_N \cdots W_1 = \sum_{j \in J} \sigma_j u_j v_j^T (XX^T)^{-\frac{1}{2}} = W$$

and that the conditions in 36 are fulfilled (recall that $Q = \sum_{i=1}^{q} \sigma_i u_i v_i^T$). $\square$

Let us now analyze the Hessian of L^N in critical points.

Proposition 35. *Let $(W_1, \ldots, W_N)$ be a critical point of L^N such that $W = W_N \cdots W_1$ has $\mathrm{rank}(W) = k$. If W is not a global optimum of L^1 on $\mathcal{M}_k$ then $(W_1, \ldots, W_N)$ is a strict saddle point of L^N.*

Proof. Since $(W_1, \ldots, W_N)$ is a critical point of L^N, the matrix $W = W_N \cdots W_1$ is a critical point of L^1 restricted to $\mathcal{M}_k$, by Proposition 34 (a). Since W is not a global optimum of L^1 on $\mathcal{M}_k$ it must be a strict saddle point of L^1 on $\mathcal{M}_k$, by Proposition 33. Therefore, there exists $Z \in T_W(\mathcal{M}_k)$ such that (for some

Riemannian metric g) it holds $g(\text{Hess}^g L^1(W)Z, Z) < 0$. Write $Z = WA + BW$ and choose again the curve (34) with $V_1 = W_1 A$ and $V_N = BW_N$. Then

$$\frac{d^2}{dt^2}L^1(\gamma(t))\Big|_{t=0} = g_W(\text{Hess}^g L^1(W)Z, Z) < 0.$$

On the other hand

$$0 > \frac{d^2}{dt^2}L^1(\gamma(t))\Big|_{t=0} = \frac{d^2}{dt^2}L^N(W_1 + tV_1, W_2, \ldots, W_{N-1}, W_N + tV_N)\Big|_{t=0}$$
$$= \langle \text{Hess}\, L^N(W)(V_1, 0, \ldots, 0, V_N), (V_1, 0, \ldots, 0, V_N)\rangle,$$

which implies that $\text{Hess}\, L^N(W)$ is not positive semidefinite, i.e., has a negative eigenvalue. In other words, $(W_1, \ldots, W_N)$ is a strict saddle point. $\qquad\square$

We note that the global minimizers of L^1 restricted to $\mathcal{M}_k$ for some $k < r$ are not covered by the above proposition, i.e., the proposition does not identify the corresponding tuples $(W_1, \ldots, W_N)$ (such that the product $W = W_N \cdots W_1$ is a global minimizer of L^1 restricted to $\mathcal{M}_k$) as strict saddle points of L^N. (In the language of [22] such points are called spurious local minima and they may lead to saddle points of L^N, see also Propositions 6 and 7 in [22].) The above proposition does not exclude that such points correspond to non-strict saddle points of L^N. In fact, in the special case of $k = 0$, the point $(0, \ldots, 0)$ is indeed not a strict saddle point if $N \geq 3$ as shown in the next result, which extends [13, Corollary 2.4] to the situation that XX^T does not necessarily need to have distinct eigenvalues.

Proposition 36. *If $XY^T \neq 0$, the point $(0, \ldots, 0)$ is a saddle point of L^N, which is strict if $N = 2$ and not strict if $N \geq 3$.*

Remark 37. *Note that if $XY^T = 0$, then $(0, \ldots, 0)$ is a global minimum of L^N.*

Proof. For convenience, we give a different proof than the one in [13, Corollary 2.4]. It is easy to see that $\nabla_{W_j} L^N(0, \ldots, 0) = 0$ for every $j = 1, \ldots, N$ so that $(0, \ldots, 0)$ is a critical point of L^N. Consider a tuple $(V_1, \ldots, V_N)$ of matrices, set $Z = V_N \cdots V_1$ and

$$\gamma(t) = (tV_N) \cdot (tV_{N-1}) \cdots (tV_1) = t^N Z.$$

Note that by (33)

$$L^N(tV_1, \ldots, tV_N)) = L^1(\gamma(t)) = \frac{1}{2}\text{tr}(\gamma(t)^T \gamma(t) XX^T - 2\gamma(t)XY^T + YY^T)$$
$$= \frac{1}{2}t^{2N}\text{tr}(Z^T ZXX^T) - t^N \text{tr}(ZXY^T) + \frac{1}{2}\text{tr}(YY^T).$$

Hence,

$$\frac{d^2}{dt^2}L^N(tV_1, \ldots, tV_N)) = \frac{d^2}{dt^2}L^1(\gamma(t)) = N(2N-1)t^{2N-2}\text{tr}(Z^T ZXX^T) - N(N-1)t^{N-2}\text{tr}(ZXY^T).$$

Note that $\text{tr}(Z^T ZXX^T) \geq 0$. Recall also that $N \geq 2$. Since $XY^T \neq 0$, there clearly exist matrices $(V_1, \ldots, V_N)$ such $\text{tr}(ZXY^T) > 0$ for $Z = V_N \cdots V_1$, so that $L^N(tV_1, \ldots, tV_N) < L^N(0, \ldots, 0)$ for small enough t. Hence, $(0, \ldots, 0)$ is not a local minimum, but a saddle point. Moreover,

$$\langle \text{Hess}\, L^N(0, \ldots, 0)(V_1, \ldots, V_N), (V_1, \ldots, V_N)\rangle = \frac{d^2}{dt^2}L^1(\gamma(t))\Big|_{t=0}$$
$$= \begin{cases} -2\,\text{tr}(ZXY^T) & \text{if } N = 2 \\ 0 & \text{if } N \geq 3 \end{cases}$$

If $N = 2$, we can find matrices V_1, V_2 such that $\text{tr}(ZXY^T) > 0$ for $Z = V_2V_1$ so that $(0,0)$ is a strict saddle for $N = 2$. If $N \geq 3$ it follows that $\text{Hess}\, L^N(0, \ldots, 0) = 0$ so that $(0, \ldots, 0)$ is not a strict saddle. $\qquad\square$

In the case $N = 2$, the following result implies that all local minima of L^2 are global and all saddle points of L^2 are strict.

Proposition 38. *Let $N = 2$ and $k < \min\{r, q\}$, where $r = \min\{d_0, d_1, d_2\}$ and $q = \mathrm{rank}(Q)$. Let W be a global minimum of L^1 restricted to $\mathcal{M}_k$, i.e., $W = \sum_{j \in J} \sigma_j u_j v_j^T (XX^T)^{-\frac{1}{2}}$, where $|J| = k$ and $\{\sigma_j : j \in J\} = \{\sigma_1, \ldots, \sigma_k\}$. Then any critical point $(W_1, W_2) \in \mathbb{R}^{d_1 \times d_0} \times \mathbb{R}^{d_2 \times d_1}$ such that $W_2 \cdot W_1 = W$ is a strict saddle point of L^2.*

Proof. For $\kappa \in \mathbb{R} \setminus \{0\}$ and $u \in \mathbb{R}^{d_2}$, $v \in \mathbb{R}^{d_1}$, $w \in \mathbb{R}^{d_0}$ with $u^T u = 1$ and $v^T v = 1$ we define the curve

$$\gamma(t) = (W_2 + t\kappa uv^T) \cdot (W_1 + t\kappa^{-1}vw^T) = W + t(\kappa uv^T W_1 + \kappa^{-1}W_2 vw^T) + t^2 uw^T.$$

Then by (33)

$$L^2(W_1 + t\kappa^{-1}vw^T, W_2 + t\kappa uv^T) = L^1(\gamma(t)) = \frac{1}{2}\mathrm{tr}(\gamma(t)^T \gamma(t) XX^T - 2\gamma(t)XY^T + YY^T).$$

We compute

$$\gamma(t)^T \gamma(t) = t^2 \left(W_1^T vu^T W_2 vw^T + (W_1^T vu^T W_2 vw^T)^T + \kappa^2 W_1^T vv^T W_1 + \kappa^{-2} wv^T W_2^T W_2 vw^T + wu^T W \right.$$
$$\left. + W^T uw^T \right) + \text{ terms which are not of order } t^2.$$

It follows that

$$\frac{d^2}{dt^2} L^1(\gamma(t))\bigg|_{t=0} = \mathrm{tr}\left((W_1^T vu^T W_2 vw^T + (W_1^T vu^T W_2 vw^T)^T + \kappa^2 W_1^T vv^T W_1 \right.$$
$$\left. + \kappa^{-2} wv^T W_2^T W_2 vw^T + wu^T W + W^T uw^T)XX^T - 2uw^T XY^T \right).$$

Let us now choose the vectors u, v, w. Note that since (W_1, W_2) is a critical point of L^2, we have $W_2^T(WXX^T - YX^T) = 0$ by Lemma 2, point 1, and hence

$$W_2^T \left(\sum_{j \in J} \sigma_j u_j v_j^T - \sum_{j=1}^{q} \sigma_j u_j v_j^T \right)(XX^T)^{\frac{1}{2}} = 0.$$

Since XX^T has full rank it follows that for any $j_0 \in \{1, \ldots, q\} \setminus J$ we have $W_2^T u_{j_0} = 0$. Since $k < q$ such a j_0 exists. Thus we may choose $j_0 \in \{1, \ldots, q\} \setminus J$ and define $u = u_{j_0}$ and $w = (XX^T)^{-\frac{1}{2}} v_{j_0}$.

If the kernel of W_1^T is trivial then $d_1 \le d_0$ and W_1 has rank d_1. It follows that then the kernel of W_2 cannot be trivial since otherwise W_2 would be injective and the rank of $W = W_2 W_1$ would be d_1. But the rank of W is $k < r \le d_1$. Hence we may choose v as follows: We choose v to be an element of the kernel of W_1^T with $\|v\|_2 = 1$ if such a v exists and otherwise we choose v to be an element of the kernel of W_2 with $\|v\|_2 = 1$.

With these choices for u, v, w we have $W_1^T vu^T W_2 vw^T = 0$ and $W^T uw^T = W_1^T W_2^T u_{j_0} w^T = 0$ so that

$$\frac{d^2}{dt^2} L^1(\gamma(t))\bigg|_{t=0} = \mathrm{tr}\left((\kappa^2 W_1^T vv^T W_1 + \kappa^{-2} wv^T W_2^T W_2 vw^T)XX^T - 2uw^T XY^T \right),$$

where at least one of the terms $W_1^T vv^T W_1$ and $wv^T W_2^T W_2 vw^T$ vanishes. We have

$$\mathrm{tr}(uw^T XY^T) = w^T XY^T u = v_{j_0}^T (XX^T)^{-\frac{1}{2}} XY^T u_{j_0} = v_{j_0}^T Q^T u_{j_0} = v_{j_0}^T \sum_{j=1}^{q} \sigma_j v_j u_j^T u_{j_0} = \sigma_{j_0}.$$

Hence

$$\frac{d^2}{dt^2} L^1(\gamma(t))\bigg|_{t=0} = \kappa^2 \mathrm{tr}\left(W_1^T vv^T W_1 XX^T \right) + \kappa^{-2} \mathrm{tr}\left(wv^T W_2^T W_2 vw^T XX^T \right) - 2\sigma_{j_0}.$$

Since $\sigma_{j_0} > 0$ and since at least one of the terms $W_1^T vv^T W_1 XX^T$ and $wv^T W_2^T W_2 vw^T XX^T$ vanishes, we can always choose $\kappa > 0$ such that $\frac{d^2}{dt^2} L^1(\gamma(t))\big|_{t=0} < 0$.

As in the proof of Propositon 33, this shows that (W_1, W_2) is a strict saddle point. $\qquad\square$

6.5. **Convergence to global minimizers.** We now state the main result of this article about convergence to global minimizers. Part (b) of Theorem 39 for two layers generalizes a result in [10, Section 4], where it is assumed that $d_x \geq d_y$ and $d_y \leq \min\{d_1, \ldots, d_{N-1}\}$ on top of some mild technical assumptions on matrices formed with X and Y (see Assumptions 3 and 4 of [10]).

Theorem 39. *Assume that XX^T has full rank, let $q = \operatorname{rank}(Q)$, $r = \min\{d_0, \ldots, d_N\}$ and let $\bar{r} := \min\{q, r\}$.*

(a) *For almost all initial values $W_1(0), \ldots, W_N(0)$, the flow (5) converges to a critical point $(W_1, \ldots, W_N)$ of L^N such that $W := W_N \cdots W_1$ is a global minimizer of L^1 on the manifold $\mathcal{M}_k$ of matrices in $\mathbb{R}^{d_N \times d_0}$ of rank $k := \operatorname{rank}(W)$, where k lies between 0 and $\bar{r}$ and depends on the initialization.*

(b) *For $N = 2$, for almost all initial values $W_1(0), \ldots, W_N(0)$, the flow (5) converges to a global minimizer of L^N on $\mathbb{R}^{d_0 \times d_1} \times \ldots \times \mathbb{R}^{d_{N-1} \times d_N}$.*

By *"for almost all initial values"* we mean that there exists a set N with Lebesgue measure zero in $\mathbb{R}^{d_0 \times d_1} \times \ldots \times \mathbb{R}^{d_{N-1} \times d_N}$ such that the statement holds for all initial values outside N.

Proof. By Theorem 5, under the flow (5), the curve $(W_1(t), \ldots, W_N(t))$ converges to some critical point $(W_1, \ldots, W_N)$ of L^N. Let k be the rank of $W := W_N \cdots W_1$. Then $k \leq r$, by construction. But also $k \leq q$ because, if $(W_1, \ldots, W_N)$ is a critical point of L^N, then W as above is a critical point of L^1 restricted to $\mathcal{M}_k$, by Proposition 34 (a). But we know that there are no critical points of L^1 in $\mathcal{M}_k$ with rank larger than q because of Proposition 32 (a). This proves that $0 \leq k \leq \bar{r}$.

If W is not a global minimizer of L^1 restricted to $\mathcal{M}_k$, then $(W_1, \ldots, W_N)$ must be strict a saddle point of L^N because of Proposition 35. By Theorem 28 only a negligible set of initial values $W_1(0), \ldots, W_N(0)$ converges to a strict saddle point of L^N. All other initial values therefore converge to a limit point $(W_1, \ldots, W_N)$ for which $W = W_N \cdots W_1$ is a global minimizer of L^1 restricted to $\mathcal{M}_k$. This proves part (a) of Theorem 39. Note that W being a minimizer of L^1 restricted to $\mathcal{M}_k$ does not imply that the corresponding matrix tuple $(W_1, \ldots, W_N)$ is a minimizer of L^N. This happens only if the rank of W is as large as possible, i.e., if $k = r$.

In the case $N = 2$, Proposition 38 shows that if the limit (W_1, W_2) has the property that $W = W_2 W_1$ is a global minimizer of L^1 in $\mathcal{M}_k$ but $k < \bar{r}$, then (W_1, W_2) is a strict saddle point of L^2. But we already know that the set of initial values $W_1(0), W_2(0)$ that converge to a strict saddle point of L^2 is negligible. We conclude that generically the solution of (5) converges to a limit (W_1, W_2) for which $W = W_2 W_1$ is a global minimizer of L^1 on $\mathcal{M}_k$ with $k = \bar{r}$, which implies that (W_1, W_2) is a global minimizer of L^N. $\square$

Balanced initial values $(W_1(0), \ldots, W_N(0))$ are of special interest as they give rise to a Riemannian gradient flow on $\mathcal{M}_k$. Unfortunately, Theorem 39 does not allow to make conclusions about the set of balanced initial values because this is a set of Lebesgue measure zero in $\mathbb{R}^{d_1 \times d_0} \times \cdots \times \mathbb{R}^{d_N \times d_{N-1}}$. We are nevertheless able to derive the following convergence result by applying Theorem 28 to the Riemannian gradient flow on $\mathcal{M}_k$.

Theorem 40. *Assume that XX^T has full rank and let $N \geq 2$. Then for almost all initializations $W(0) \in \mathbb{R}^{d_y \times d_x}$ on $\mathcal{M}_k$, the flow $W(t)$ on $\mathcal{M}_k$ solving (24) (cf. Theorem 18) converges to a global minimum of L^1 restricted to $\mathcal{M}_k$ or to a critical point on some $\mathcal{M}_\ell$, where $\ell < k$. Note that for $k > \operatorname{rank}(Q)$ there is no global minimum of L^1 on $\mathcal{M}_k$ so that then the second option applies. Here "for almost all $W(0)$" means for all $W(0)$ up to a set of measure zero.*

Proof. As in the proof of Theorem 18 we write $W(t) = W_N(t) \cdots W_1(t)$, where the $W_i(t) \in \mathbb{R}^{d_i \times d_{i-1}}$ (for some $d_i \geq k$ with $d_0 = d_x$ and $d_N = d_y$) are balanced and solve (5). Since the tuple $(W_1(t), \ldots, W_N(t))$ satisfies (5), it converges to a critical point of L^N by Theorem 5, hence the flow $W(t)$ converges for all initial values to some W that is a critical point of L^1 on some $\mathcal{M}_\ell$, where $\ell \leq k$; see Proposition 34. Since by Proposition 33 all critical points of L^1 on $\mathcal{M}_k$ except for the global minimizers are strict saddle points, the claim follows from Theorem 28, whose assumptions are satisfied by Proposition 11 and Theorem 18. $\square$

The reason why we cannot choose $k = \bar{r}$ in Theorem 39 (a), i.e., state that the flow for $N \geq 3$ converges to the global minimum of L^1 on $\mathcal{M}_{\bar{r}}$ for almost all initializations is that not all saddle points of L^N are necessarily strict for $N \geq 3$. Nevertheless, we conjecture a more precise version of the previous result in the spirit of Theorem 23. Part (a) below is a strengthened version of the overfitting conjecture in [10], where additional assumptions are made.

Conjecture 41. *Assume that XX^T has full rank.*

  (a) *The statement in Theorem 39 (b) also holds for $N > 2$.*

  (b) *Consider the autoencoder case $X = Y$ and let $d = d_0 = d_N$. Let $r = \min_{i=1,\ldots,N} d_i$. Let $\lambda_1 \geq \ldots \geq \lambda_d$ be the eigenvalues of XX^T and let $u_1, \ldots, u_d$ be corresponding orthonormal eigenvectors. Let U_r be the matrix with columns $u_1, \ldots, u_r$. Assume that $\lambda_r > \lambda_{r+1}$. Assume further that $W(0)U_r$ has rank r and that for all $i \in \{1, \ldots, r\}$ we have*

$$u_i^T W(0) u_i > 0, \tag{37}$$

  where $W(t) = W_N(t) \cdots W_1(t)$. Then $W(t)$ converges to $\sum_{i=1}^r u_i u_i^T$.

  (c) *In Theorem 40, if $k \leq \mathrm{rank}(Q)$, convergence to a critical point on some $\mathcal{M}_\ell$, where $\ell < k$, happens only for a set of initial values that has measure zero.*

Remark 42. *Without the condition that $u_i^T W(0) u_i > 0$ for all $i \in \{1, \ldots, r\}$, the above conjecture (b) is wrong.*

Proof. Indeed, in the autoencoder case with $N = 2$ and $r = 1$ with $W_1(0) = u_1^T$ and $W_2(0) = -u_1$ (which is a balanced starting condition and $W(0)U_1$ has obviously rank 1), we show that W_1, W_2 and W all converge to the zero-matrix of their respective size. Write $W_1 = (\alpha_1, \ldots, \alpha_d)$ and $W_2 = (\beta_1, \ldots, \beta_d)^T$. We may assume that XX^T is a diagonal matrix with entries $\lambda_1 \geq \ldots \geq \lambda_d > 0$. (In particular, the u_i are given by the standard unit vectors $u_i = e_i$.) Then the system (5), see also (48), becomes

$$\dot{\alpha}_j = -\lambda_j \alpha_j \sum_{i=1}^d \beta_i^2 + \lambda_j \beta_j, \quad \alpha_j(0) = \delta_{j1},$$
$$\dot{\beta}_j = -\beta_j \sum_{i=1}^d \lambda_i \alpha_i^2 + \lambda_j \alpha_j, \quad \beta_j(0) = -\delta_{j1}. \tag{38}$$

This system is solved by the following functions:

$$\alpha_1(t) = \frac{1}{\sqrt{2e^{2\lambda_1 t} - 1}}, \quad \alpha_j(t) = 0 \text{ for all } j \geq 2,$$
$$\beta_1(t) = \frac{-1}{\sqrt{2e^{2\lambda_1 t} - 1}}, \quad \beta_j(t) = 0 \text{ for all } j \geq 2. \tag{39}$$

Obviously, all α_j and β_j converge to 0 as t tends to infinity. From this the claim follows. (By Theorem 49, this equilibrium is not stable, so this behavior may not be obvious in numerical simulations.) □

7. NUMERICAL RESULTS

We numerically study the convergence of gradient flows in the linear supervised learning setting as a proof of concept of the convergence results presented above in both the general supervised learning case and the special case of autoencoders. Moreover, in the autoencoder case the experiments also computationally explore the conjecture (Conjecture 41) of the manuscript.

7.1. Autoencoder case. We study the gradient flow (5) in the autoencoder setting, where $Y = X \in \mathbb{R}^{d_x \times m}$ in (3) for different dimensions of X (i.e., d_x and m) and different values of the number N of layers, where we typically use $N \in \{2, 5, 10, 20\}$. A Runge-Kutta method (RK4) is used to solve the gradient flow differential equation with appropriate step sizes $t_n = t_0 + nh$ for large n and $h \in (0, 1)$. The experiments fall into two categories based on initial conditions of the gradient flow: a) *balanced* – where the balanced conditions are satisfied; and b) *non-balanced* – where the balanced conditions are not satisfied. Under a) we investigate the general case in these balanced conditions where condition (37) of Conjecture 41 is satisfied, but also a special case were the balanced conditions are satisfied but condition (37) of Conjecture 41 is not satisfied.

The results in summary, considering $W = W_N \cdots W_1$ as the limiting solution of the gradient flow, that is $W = \lim_{t \to \infty} W(t)$, where $W(t) = W_N(t) \cdots W_1(t)$: We show that with balanced initial conditions, the solutions of the gradient flow converges to $U_r U_r^T$, where the columns of U_r are the r eigenvectors corresponding to the r largest eigenvalues of XX^T. The convergence rates decrease with an increase in either d or N or both. We see similar results for the non-balanced case.

7.1.1. *Balanced initial conditions.* In this section and Section 7.1.2 the data matrix $X \in \mathbb{R}^{d_x \times m}$ is generated with columns drawn i.i.d. from a Gaussian distribution, i.e., $x_i \sim \mathcal{N}(0, \sigma^2 I_{d_x})$, where $\sigma = 1/\sqrt{d_x}$. Random realization of X with sizes $d_x = d$ and $m = 3d$ are varied to investigate different dimensions of the input data, i.e., with $2N \le d \le 20N$. For each fixed d, the dimensions d_j of the $W_j \in \mathbb{R}^{d_j \times d_{j-1}}$ for $j = 1, \dots, N$ are selected as follows: We set $d_1 = r = [d/2]$, where $[\cdot]$ rounds to the nearest integer, and put $d_j = [r + (d-r)(j-1)/(N-1)]$, $j = 2, \dots, N$ (generating an integer "grid" of numbers between $d_1 = r$ and $d_N = d_x = d$).

In the first set of experiments, we consider a general case of the balanced initial conditions, precisely $W_{j+1}^T(0)W_{j+1}(0) = W_j(0)W_j^T(0)$, $j = 1, \dots, N-1$, where condition (37) of Conjecture 41 is satisfied. The dimensions of the W_j and their initializations are as follows. Recall, $W_j \in \mathbb{R}^{d_j \times d_{j-1}}$ for $j = 1, \dots, N$ where $d_N = d_0 = d_x = d$ and $d_1 = r$ is the rank of $W = W_N \cdots W_1$. We randomly generate $d_j \times d_j$ orthogonal matrices V_j and then form $W_j(0) = V_j I_{d_j d_1} U_{j-1}^T$ for $j = 1, \dots, N$, where $U_j \in \mathbb{R}^{d_j \times d_1}$ is composed of the d_1 columns of V_j, and I_{ab} is the (rectangular) $a \times b$ identity matrix. For all the values of N and the different ranks of W considered, Figure 1 shows that the limit of $W(t)$ as $t \to \infty$ is $U_r U_r^T$, where the columns of U_r are r eigenvectors of XX^T corresponding to the largest r eigenvalues of XX^T. This agrees with the theoretical results obtained for the autoencoder setting.
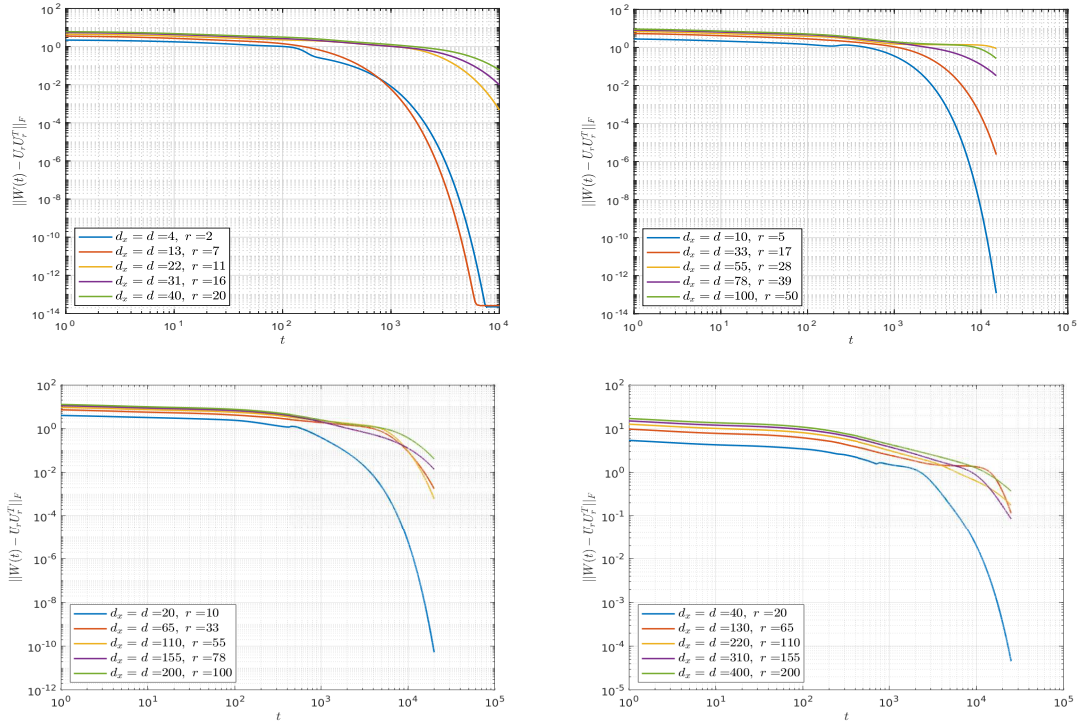

FIGURE 1. Convergence of solutions for the general balanced case. Error between $W(t)$ and $U_r U_r^T$ for different r and d values. *Top left panel*: $N = 2$; *top right panel*: $N = 5$; *bottom left panel*: $N = 10$; *bottom right panel*: $N = 20$.

In addition, when $W(t)$ converges to $U_r U_r^T$ then $\|X - W(t)X\|_F$ converges to $\sqrt{\sum_{i>r} \sigma_i^2}$. This is also tested and confirmed for $N = 2, 5, 10, 20$, but for the purpose of saving space we show results for $N = 2$ and $N = 20$ in Figure 2. This depicts convergence of the functional $L^1(W(t))$ to the optimal error, which is the square-root of the sum of the tail eigenvalues of XX^T of order greater than r. Moreover, in the
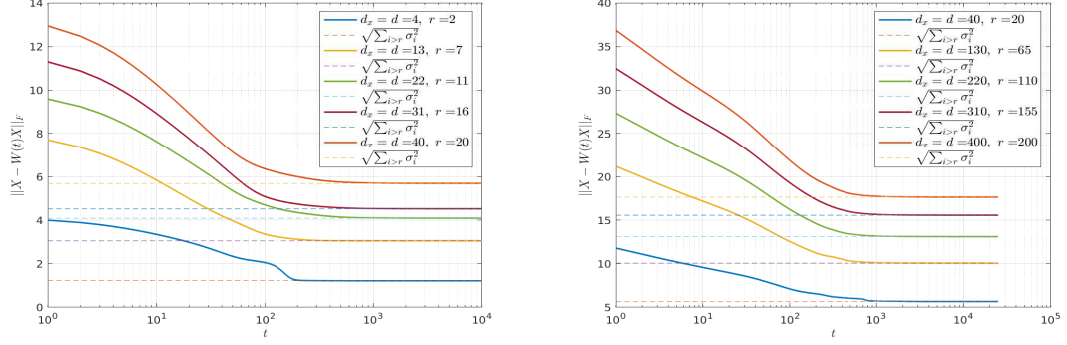
FIGURE 2. Convergence of solutions for the general balanced case. Errors between X and $W(t)X$ for different r and d values. *Left panel:* $N = 2$; *right panel:* $N = 20$.
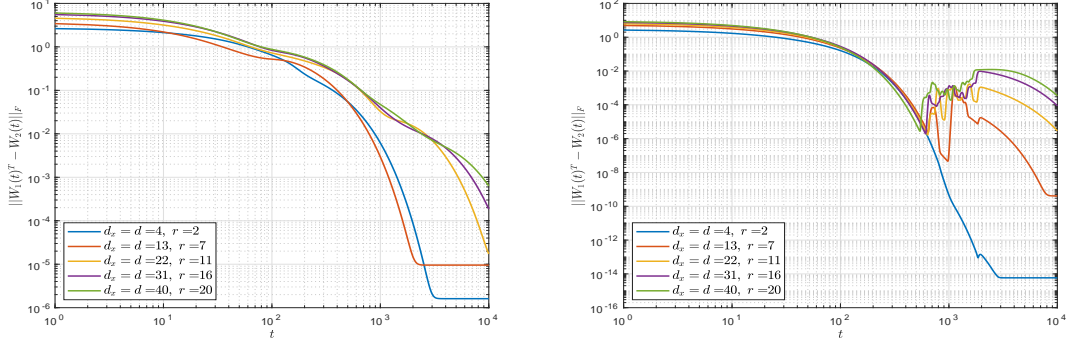

FIGURE 3. Difference between $W_1(t)$ and $W_2(t)^T$ in the $N = 2$ settings, for *left panel:* general balanced case; *right panel:* special balanced case.

autoencoder setting when $N = 2$ we showed in Lemma 46 that the optimal solutions are $W_2 = W_1^T$. This is also confirmed in the numerics as can be seen in the left panel plot of Figure 3.
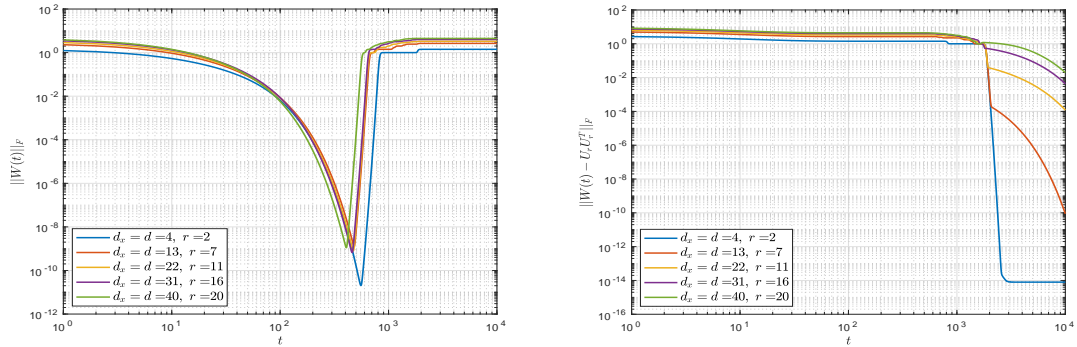
FIGURE 4. In the special balanced case, *left panel:* norm of $W(t)$; *right panel:* errors between $W(t)$ and $U_r U_r^T$ for different r and d values.

In the second set of experiments, we attempt to test Conjecture 41 by constructing pathological examples, where we have balanced initial conditions, but $W(0)$ violates condition (37) of Conjecture 41. Precisely, in the case $N = 2$ we take $W_1(0) = V_r^T$ and $W_2(0) = -V_r$, where the columns of V_r are the top r eigenvectors of XX^T. Such $W(0)$ clearly violates the condition of the conjecture $u_i^T W(0) u_i > 0$ for all $i \in [r]$.

The hypothesis is that in such a setting the solution will not converge to the optimal solution proposed in Conjecture 41. Remark 42 showed that in such a case the solution should converge to 0, that is $\lim_{t \to \infty} W(t) = 0$. This can be seen in the left panel plot of Figure 4. The dip in the left panel shows that $W(t)$ is approaching zero in a first phase. However, probably due to numerical errors the flow escapes the equilibrium point at zero. In fact, zero is an unstable point (a strict saddle point), so that, numerically, the flow will hardly converge to zero. The right panel plot of Figure 4 shows very slow convergence to $U_r U_r^T$. Moreover, the limiting solutions (despite slow convergence) satisfy $W_2 = W_1^T$ as shown in the right plot of Figure 3.

7.1.2. *Non-balanced initial conditions.* For $W_j(0)$, $j = 1, \ldots, N$, we randomly generate Gaussian matrices. The two plots in Figure 5 and the left panel plot of Figure 6 show that $W(t)$ converges to $U_r U_r^T$. As in the balanced case we can confirm that $\|X - W(t)X\|_F$ converges to $\sqrt{\sum_{i>r} \sigma_i^2}$. On the other hand, for $N = 2$ in this case we see that $W_2(t)$ does not converge to $W_1(t)^T$ in contrast to the balanced case, as can be seen in the right panel plot of Figure 6.
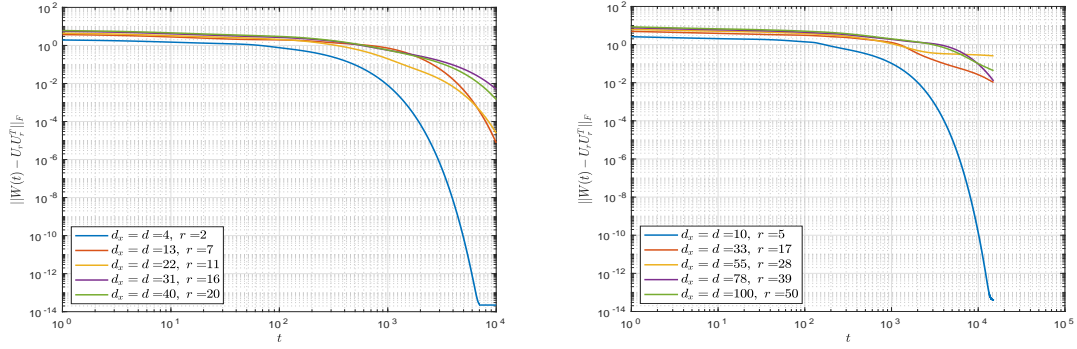

FIGURE 5. Non-balanced case convergence of solutions of the gradient flow. Errors between $W(t)$ and $U_r U_r^T$ for different r and d values for *left panel:* $N = 2$, *right panel:* $N = 5$.
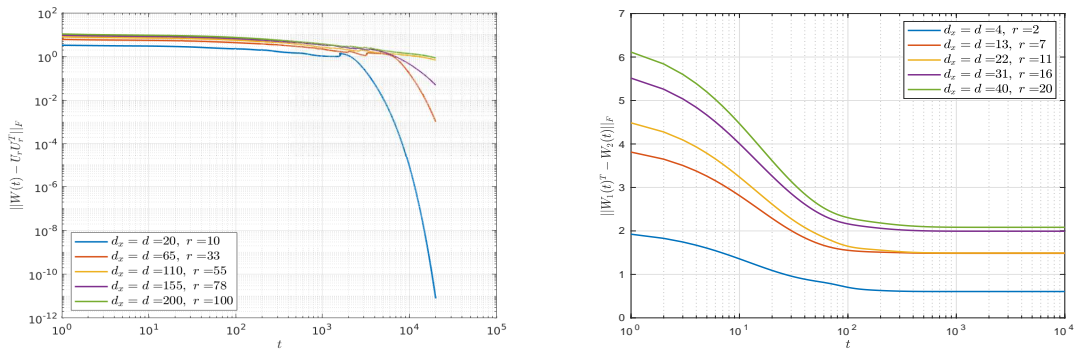

FIGURE 6. Non-balanced case convergence. *Left panel:* Errors between $W(t)$ and $U_r U_r^T$ for different r and d values for $N = 10$. *Right panel:* Errors between $W_2(t)$ and $W_1(t)^T$.

7.1.3. *Convergence rates.* Here the data matrix $X \in \mathbb{R}^{d_x \times m}$ is generated with columns drawn i.i.d.from a Gaussian distribution, i.e., $x_i \sim \mathcal{N}(0, \sigma^2 I_{d_x})$, where $\sigma = 1/\sqrt{d_x}$. Random realization of X with two different values for d_x (as in above $m = 3d$) and different r, the rank of $W(t)$, are used. For each fixed d, the dimensions d_j of the $W_j \in \mathbb{R}^{d_j \times d_{j-1}}$ are selected using an arbitrarily chosen r and setting $d_j = [r + (d-r)(j-1)/(N-1)]$ for $j = 1, \ldots, N$. The value of r is stated in the caption of the figures. The experiments show very rapid convergence of the solutions but also the dependence of the convergence rate on N, d_x, and r. We investigate this for different values of N, d_x and r, in both the balanced and non-balanced cases. Convergence plots for the balanced initial conditions are shown in Figure 7, depicting smooth convergence. Similarly, we have convergence rates of the non-balanced case in Figure 8. These plots also show a slightly faster convergence for the balanced case than for the non-balanced case.
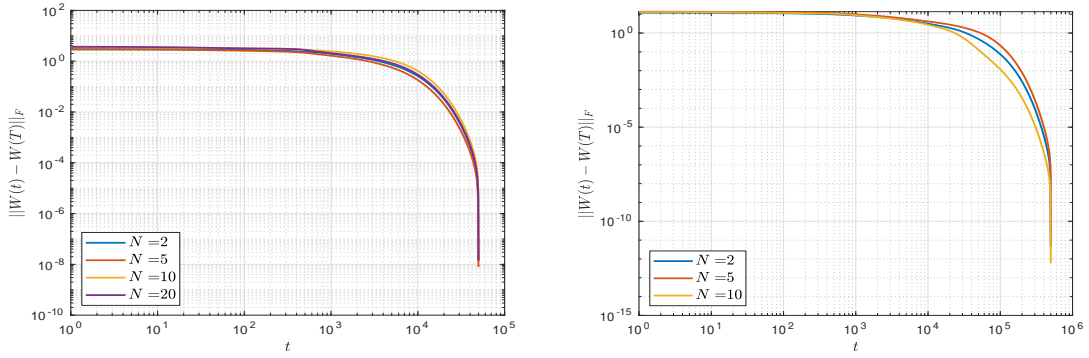

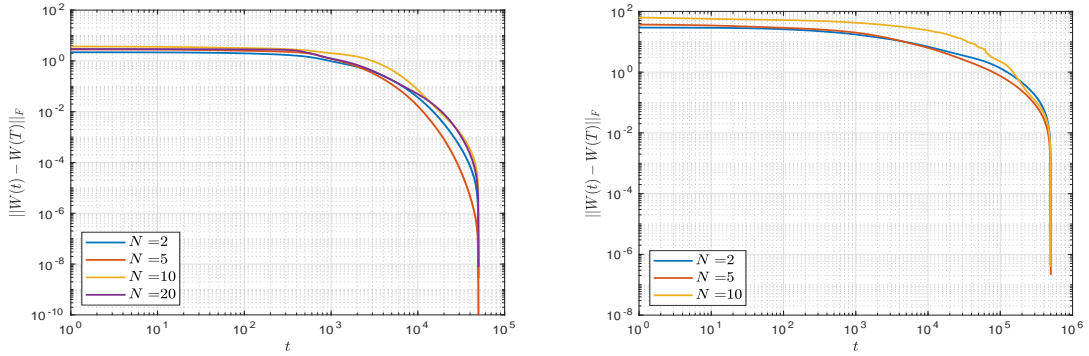FIGURE 7. Convergence rates of solutions of the gradient flow in the autoencoder case with balanced initial conditions – errors between $W(t)$ and $W(T)$ for different N values, where T is the final time. Dimensions *Left panel*: $d_x = 20$, $r = 1$; *Right panel*: $d_x = 200$, $r = 10$.

FIGURE 8. Convergence rates of solutions of the gradient flow in the autoencoder case with non-balanced initial conditions – errors between $W(t)$ and $W(T)$ for different N values, where T is the final time. Dimensions *Left panel*: $d_x = 20$, $r = 1$; *Right panel*: $d_x = 200$, $r = 10$.

7.2. **General supervised learning case.** Experiments were also conducted to test the results in the general supervised learning setting to support theoretical results in Theorem 31 and Propositions 32 and 33. We show results for $N = 2, 5, 10, 20$, and two sets of values for d_x and r (rank of $W(t)$ and $\widetilde{W}$, the true parameters). The data matrix X is generated as in the autoencoder case and $Y = \widetilde{W}X$, where

$\widetilde{W} = \widetilde{W}_N \cdots \widetilde{W}_1$, with $\widetilde{W}_j \in \mathbb{R}^{d_j \times d_{j-1}}$ for $j = 1, \ldots, N$ with $d_N = d_0 = d_x = d$ and $d_1 = r$ is the rank of $\widetilde{W}$. The entries of $\widetilde{W}_j$ are randomly generated independently from a Gaussian distribution with standard deviation $\sigma = 1/\sqrt{d_j}$. The dimensions $d_j \times d_{j-1}$ of the W_j for $j = 1, \ldots, N$, are again selected respectively in an integer grid, i.e., $d_j = [r + (d_x - r)(j-1)/(N-1)]$, where r is arbitrarily fixed. The initial conditions are generated as was done in the autoencoder case. We investigate the convergence rates for the balanced and non-balanced initial conditions of the gradient flows. The results of the experiments are plotted in Figures 9 and 10. In these plots k is the rank of $Q \in \mathbb{R}^{d_y \times d_x}$ defined in (31), and $Q = U_k \Sigma_k V_k$ is the (reduced) singular value decomposition, i.e., $U_k \in \mathbb{R}^{d_x \times k}$ and $V_k \in \mathbb{R}^{d_y \times k}$ have orthonormal columns and $\Sigma_k \in \mathbb{R}^{k \times k}$ is a diagonal matrix containing the non-zero singular values of Q.
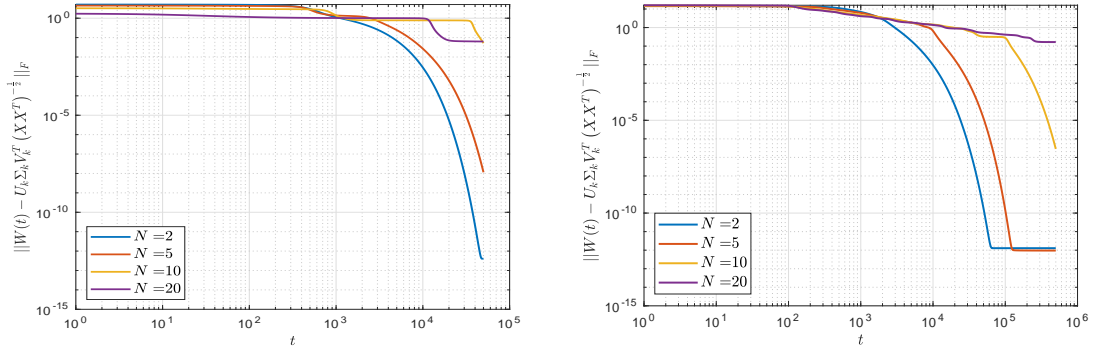

FIGURE 9. Convergence rates of solutions of the gradient flow of the general supervised learning problem depicted by convergence to W in (1) of Proposition 32 with balanced initial conditions for *left panel: $d_x = 20$, $r = 2$; right panel: $d_x = 200$, $r = 20$.*

With balanced initial conditions the plots of Figure 9 show convergence rates of the flow to W in (1) of Proposition 32. With non-balanced initial conditions the plots of Figure 10 show convergence rates to W in (1) of Proposition 32. These results show rapid convergence of the flow and the dependence of the convergence rate on N, r and d_x with either balanced or non-balanced initial conditions. Note that W in (1) of Proposition 32 is the same as the true parameters $\widetilde{W}$. This can be seen by comparing the left panel plot of Figure 9 to the left panel plot of Figure 11 and the left panel plot of Figure 10 to the right panel plot of Figure 11.
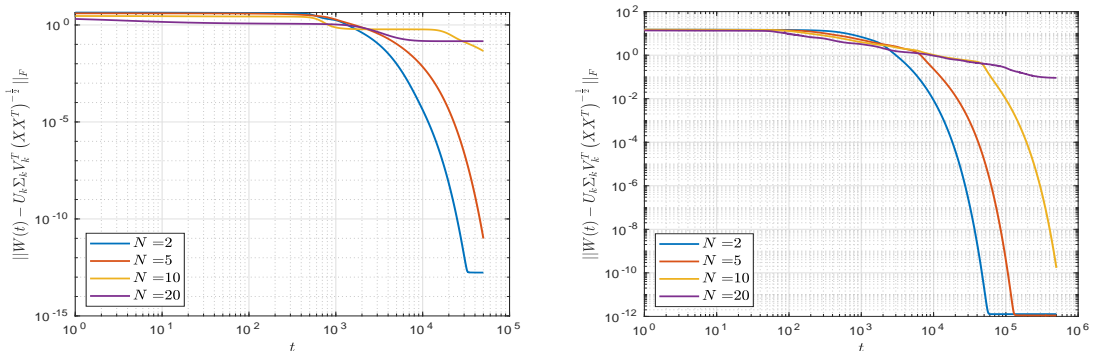

FIGURE 10. Convergence rates of solutions of the gradient flow of the general supervised learning problem depicted by convergence to W in (1) of Proposition 32 with non-balanced initial conditions for *left panel: $d_x = 20$, $r = 2$; right panel: $d_x = 200$, $r = 20$.*

Convergence is slower for larger N, and it seems not to depend on the initial conditions, balanced or non-balanced, see the plots of Figures 9 and 10. Equivalently, this can be seen from the error of the supervised learning loss shown in the plots of Figure 12 for balanced initial conditions. There is much stronger dependence on N in this setting than in the autoencoder setting.
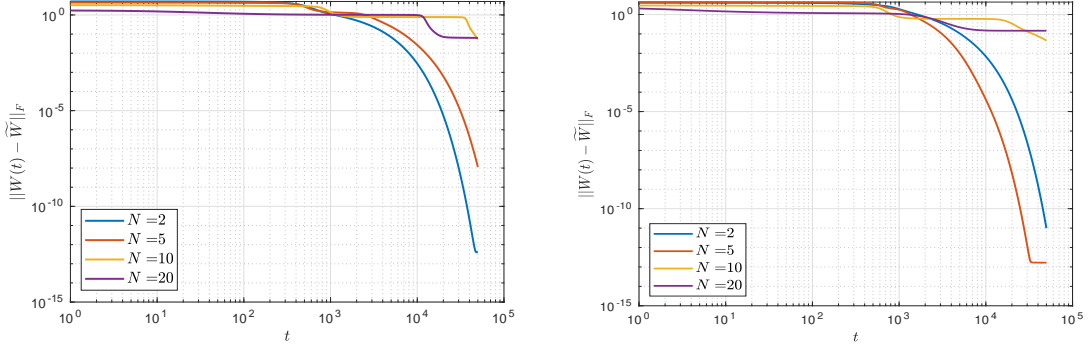

FIGURE 11. Convergence to the true parameters $\widetilde{W}$ for ($d_x = 20$, $r = 2$) with *left panel:* balanced initial conditions; *right panel:* non-balanced initial conditions.
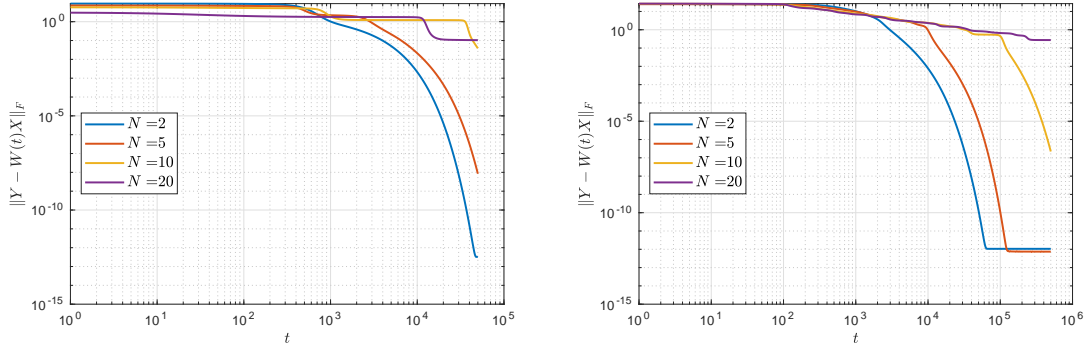

FIGURE 12. General supervised learning errors with balanced initial conditions for dimensions *left panel:* $d_x = 20$, $r = 2$; *right panel:* $d_x = 200$, $r = 20$.

7.3. **Conclusion.** To conclude the numerical section we summarise our results as follows. In the autoencoder case we confirmed that the solutions of the gradient flow converges to $U_r U_r^T$, while in the general supervised learning case we confirmed convergence of the flow to W in (1) of Proposition 32. Such convergence occurs with either balanced or non-balanced initial conditions albeit a slight faster convergence in the balanced than in the non-balanced. Secondly, in the autoencoder case we numerically confirmed the hypothesis of Conjecture 41 and that $W_2(t) = W_1(t)^T$ as claimed for $N = 2$ with balanced initial conditions, which does not necessarily hold with non-balanced initial conditions. Moreover, in both the autoencoder and the general supervised learning setting we see that as the size (N, d_x, r) of the problem instance increases the convergence rates decrease. In the autoencoder case we saw stronger dependence in d_x and r than in the general supervised learning case. On the other hand the dependence on N seems to be stronger in the general supervised learning case than in the autoencoder case.

References

[1] P. A. Absil, R. Mahony, and B. Andrews. Convergence of the iterates of descent methods for analytic cost functions. *SIAM J. Optim.* 16(2), pp. 531-547, 2005.

[2] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

[3] P.-A. Absil, I.V. Oseledets. Low-rank retractions: a survey and new results. *Comput. Optim. Appl.* 62: 5–29, 2015.

[4] S. Arora, N. Cohen, N. Golowich, and W. Hu. A convergence analysis of gradient descent for deep linear neural networks, ICLR, 2019. (arXiv:1810.02281).

[5] S. Arora, N. Cohen, and E. Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. *Preprint arXiv:1802.06509*, 2018.

[6] A. Banyaga, D. Hurtubise. *Lectures on Morse homology*. Springer Science and Business Media, 2004.

[7] R. Bhatia. *Matrix Analysis*, volume 169. Springer, 1997.

[8] R. Bhatia, M. Uchiyama. The operator equation $\sum_{i=0}^{n} A^{n-i} X B^i = Y$. *Expo. Math.* 27:251–255, 2009.

[9] E. Carlen and J. Maas. An analog of the 2-Wasserstein metric in non-commutative probability under which the fermionic Fokker-Planck equation is gradient flow for the entropy. *Comm. Math. Phys.*, 331, 2012.

[10] Y. Chitour, Z. Liao, and R. Couillet. A geometric approach of gradient descent algorithms in neural networks. *Preprint*, arXiv:1811.03568, 2018.

[11] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[12] U. Helmke and M. A. Shayman. Critical points of matrix least squares distance functions. *Lin. Alg. Appl.*, 215:1–19, 1995.

[13] K. Kawaguchi. Deep learning without poor local minima. *Advances in Neural Information Processing Systems 29*, pages 586–594, 2016.

[14] K. Kurdyka, T. Mostowski, and A. Parusinski. Proof of the gradient conjecture of R. Thom. Ann. of Math. (2), 152, pp. 763–792, 2000.

[15] S. Lang. Fundamentals of Differential Geometry. Springer, 1999.

[16] J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan, and B. Recht. First-order methods almost always avoid strict saddle points. *Math. Program.*, 176(1):311–337, 2019.

[17] S. Lojasiewicz. Sur les trajectoires du gradient d'une fonction analytique. *Seminari di geometria*, 1983:115–117, 1984.

[18] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT Press, Cambridge, Mass. [u.a.], 2013.

[19] F. Otto and M. Westdickenberg. Eulerian calculus for the contraction in the Wasserstein distance. *SIAM J. Math. Analysis*, 37:1227–1255, 2005.

[20] M. Shub. *Global Stability of Dynamical Systems*. Springer, 1986.

[21] L. Simon. *Theorems on Regularity and Singularity of Energy Minimizing Maps*. Birkhäuser, 1996.

[22] M. Trager, K. Kohn, J. Bruna, *Pure and spurious critical points: a geometric study of linear networks*. Preprint arXiv:1910.01671, 2019.

[23] W. Yan, U. Helmke, and J. B. Moore. Global analysis of Oja's flow for neural networks. *IEEE Trans. Neural Netw.*, 5(5):674–683, 1994.

Appendix A. Proof of Proposition 10

The proof is based on the next lemma, which follows from [8].

Lemma 43. *Let $A \in \mathbb{R}^{m \times m}$, $B \in \mathbb{R}^{n \times n}$, be positive definite matrices and $Y \in \mathbb{R}^{m \times n}$. Then, for $p \in \mathbb{N}$, $p \geq 2$, the solution $X \in \mathbb{R}^{m \times n}$ of the matrix equation*

$$A^{p-1} X + A^{p-2} X B + \cdots + A X B^{p-2} + X B^{p-1} = Y \tag{40}$$

satisfies

$$X = \frac{\sin(\pi/p)}{\pi} \int_0^\infty (t I_m + A^p)^{-1} Y (t I_n + B^p)^{-1} t^{1/p} dt \tag{41}$$

$$= \frac{1}{p \Gamma(1 - 1/p)} \int_0^\infty \int_0^t e^{-s A^p} Y e^{-(t-s) B^p} ds \, t^{-(1+1/p)} dt. \tag{42}$$

Proof. The first formula (41) is shown for $n = m$ in [8], for matrices with eigenvalues in $\{z \in \mathbb{C} : z \neq 0, -\pi/p < \arg z < \pi/p\}$. Positive definite matrices clearly have their eigenvalues in this set. Formula (41) extends to squares A, B of possibly different dimensions m, n. In fact, [8] first proves (41) for $A = B$, see [8, eq. (8)] and then extends the solution by the "Berberian trick" which introduces the block matrices

$$\widetilde{A} = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}, \quad \widetilde{Y} = \begin{pmatrix} 0 & Y \\ 0 & 0 \end{pmatrix}, \quad \widetilde{X} = \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix}.$$

If $\widetilde{X}$ solves $\sum_{i=1}^{p} \widetilde{A}^{p-i} \widetilde{X} \widetilde{A}^{i-1} = \widetilde{Y}$, then the submatrix $X = X_{12}$ solves (40) and one obtains (41). This argument works for general m, n so that (41) holds under the conditions of the lemma.

In the case that $A = B$, [8, eq. (15)] implies (42). The general case of possibly different A, B is then established again by the Berberian trick as above. $\qquad\square$

Proof of Proposition 10. We aim at applying Lemma 43 for the matrices $A = (WW^T)^{1/N}$ and $B = (W^T W)^{1/N}$ in order to obtain a formula for $\bar{\mathcal{A}}_W^{-1}$. Unfortunately, in the rank deficient case, these matrices are only positive semi-definite and not positive definite. We will overcome this problem by using an approximation argument. For $u > 0$, the matrices $(uI_{d_y} + WW^T)^{1/N}$ and $(uI_{d_x} + W^T W)^{1/N}$ are positive definite and hence, the linear operator

$$\mathcal{A}_{W,u} : \mathbb{R}^{d_y \times d_x} \to \mathbb{R}^{d_y \times d_x}, \quad \mathcal{A}_{W,u}(Z) = \sum_{j=1}^{N} (uI_{d_y} + WW^T)^{\frac{N-j}{N}} Z (uI_{d_x} + W^T W)^{\frac{j-1}{N}}$$

is invertible by Lemma 43 with inverse $\mathcal{A}_{W,u}^{-1} : \mathbb{R}^{d_y \times d_x} \to \mathbb{R}^{d_y \times d_x}$,

$$\mathcal{A}_{W,u}^{-1}(Z) = \frac{\sin(\pi/N)}{\pi} \int_0^\infty \left((t+u)I_{d_y} + WW^T \right)^{-1} Z \left((t+u)I_{d_x} + W^T W \right)^{-1} t^{1/N} dt. \qquad (43)$$

Furthermore, $\mathcal{A}_{W,u}$ maps $T_W(\mathcal{M}_k)$ into $T_W(\mathcal{M}_k)$ for all $u \geq 0$. Indeed, let $W = U\Sigma V^T$ be the full singular value decomposition of W with U, V being (square) orthogonal matrices and $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)$. If $Z = WA + BW \in T_W(\mathcal{M}_k)$ then for $\alpha, \beta \geq 0$,

$$\begin{aligned}
&(uI_{d_y} + WW^T)^\alpha Z (uI_{d_x} + W^T W)^\beta \\
&\quad = U \, \text{diag}((u+\sigma_1^2)^\alpha, \ldots, (u+\sigma_k^2)^\alpha, u^\alpha, \ldots, u^\alpha)) U^T U \Sigma V^T A (uI_{d_x} + W^T W)^\beta \\
&\qquad + (uI_{d_y} + WW^T)^\alpha B U \Sigma V^T V \, \text{diag}((u+\sigma_1^2)^\beta, \ldots, (u+\sigma_k^2)^\beta, u^\beta, \ldots, u^\beta)) V^T \\
&\quad = U\Sigma \, \text{diag}((u+\sigma_1^2)^\alpha, \ldots, (u+\sigma_k^2)^\alpha, 0, \ldots, 0) V^T A (uI_{d_x} + W^T W)^\beta \\
&\qquad + (uI_{d_y} + WW^T)^\alpha B U \Sigma \, \text{diag}((u+\sigma_1^2)^\beta, \ldots, (u+\sigma_k^2)^\beta, 0, \ldots, 0) V^T \\
&\quad = W V \, \text{diag}((u+\sigma_1^2)^\alpha, \ldots, (u+\sigma_k^2)^\alpha, 0, \ldots, 0) V^T A (uI_{d_x} + W^T W)^\beta \\
&\qquad + (uI_{d_y} + WW^T)^\alpha B U \, \text{diag}((u+\sigma_1^2)^\beta, \ldots, (u+\sigma_k^2)^\beta, 0, \ldots, 0) U^T W.
\end{aligned}$$

The last expression is clearly an element of $T_W(\mathcal{M}_k)$ and by the formula for $\mathcal{A}_{W,u}$ this implies that this operator maps the tangent space $T_W(\mathcal{M}_k)$ into itself. Let us denote $\bar{\mathcal{A}}_{W,u} : T_W(\mathcal{M}_k) \to T_W(\mathcal{M}_k)$ the corresponding restriction and by $\bar{\mathcal{A}}_{W,u}^{-1} : T_W(\mathcal{M}_k) \to T_W(\mathcal{M}_k)$ the restriction of the inverse map to $T_W(\mathcal{M}_k)$. Clearly, (43) still holds for the restriction $\bar{\mathcal{A}}_{W,u}^{-1}$. Lemma 8 implies that the restrictions $\bar{\mathcal{A}}_{W,u}$ and $\bar{\mathcal{A}}_{W,u}^{-1}$ are both well-defined also for $u = 0$ with $\bar{\mathcal{A}}_{W,0} = \bar{\mathcal{A}}_W$ and $\bar{\mathcal{A}}_{W,0}^{-1} = \bar{\mathcal{A}}_W^{-1}$. Moreover, the map $u \mapsto \bar{\mathcal{A}}_{W,u}$ is continuous in $u \geq 0$ and, hence, also $u \mapsto \bar{\mathcal{A}}_{W,u}^{-1}$ is continuous in $u \geq 0$. We claim that also the right hand side of (43) with $Z \in T_W(\mathcal{M}_k)$ is well-defined and continuous for all $u \geq 0$, which will give an inversion formula for $\bar{\mathcal{A}}_W^{-1}$ by setting $u = 0$.

In order to show continuity of the right hand side of (43) in u, we investigate uniform integrability of the integrand for $Z \in T_W(\mathcal{M}_k)$. By Lemma 7 we can write $Z = P_W(Z) = UP_k U^T Z(I_{d_x} - VP_k V^T) + ZVP_k V^T$, where P_k is the diagonal matrix from the lemma and $W = U\Sigma V^T$ is the (full) singular value decomposition of W. In particular, the matrix $\Sigma \in \mathbb{R}^{d_y \times d_x}$ has the singular values $\sigma_1 \geq \ldots \geq \sigma_k > 0$ on the diagonal, with all other entries equal to zero. For simplicity we write $E = I_{d_x} - VP_k V^T$. Denoting

$$\Lambda_{t,u} := ((t+u)I_{d_y} + \Sigma\Sigma^T)^{-1} = \text{diag}\left(\frac{1}{t+u+\sigma_1^2}, \ldots, \frac{1}{t+u+\sigma_k^2}, \frac{1}{t+u}, \ldots, \frac{1}{t+u} \right) \in \mathbb{R}^{d_y \times d_y},$$

$$\widetilde{\Lambda}_{t,u} := ((t+u)I_{d_x} + \Sigma^T\Sigma)^{-1} = \text{diag}\left(\frac{1}{t+u+\sigma_1^2}, \ldots, \frac{1}{t+u+\sigma_k^2}, \frac{1}{t+u}, \ldots, \frac{1}{t+u} \right) \in \mathbb{R}^{d_x \times d_x},$$

$$K_{t,u} := ((t+u)I_{d_y} + \Sigma\Sigma^T)^{-1} P_k = \text{diag}\left(\frac{1}{t+u+\sigma_1^2}, \ldots, \frac{1}{t+u+\sigma_k^2}, 0, \ldots, 0 \right) \in \mathbb{R}^{d_y \times d_y},$$

$$\widetilde{K}_{t,u} := P_k((t+u)I_{d_x} + \Sigma^T\Sigma)^{-1} = \mathrm{diag}\left(\frac{1}{t+u+\sigma_1^2},\ldots,\frac{1}{t+u+\sigma_k^2},0,\ldots,0\right) \in \mathbb{R}^{d_x \times d_x}$$

we have

$$\left((t+u)I_{d_y} + WW^T\right)^{-1} Z \left((t+u)I_{d_x} + W^TW\right)^{-1} = UK_{t,u}U^T ZEV\widetilde{\Lambda}_{t,u}V^T + U\Lambda_{t,u}U^T ZV\widetilde{K}_{t,u}V^T.$$

Taking the spectral norm gives

$$\| \left((t+u)I_{d_y} + WW^T\right)^{-1} Z \left((t+u)I_{d_x} + W^TW\right)^{-1} \|_{2\to2} \le 2\|Z\|_{2\to2}\sigma_k^{-2}(t+u)^{-1} \le 2\|Z\|_{2\to2}\sigma_k^{-2}t^{-1}.$$

This estimate will be good enough for $t \to 0$, for any $u > 0$. For $t \to \infty$ we need a second estimate

$$\| \left((t+u)I_{d_y} + WW^T\right)^{-1} Z \left((t+u)I_{d_x} + W^TW\right)^{-1} \|_{2\to2}$$
$$\le \| \left((t+u)I_{d_y} + WW^T\right)^{-1} \|_{2\to2}\|Z\|_{2\to2}\| \left((t+u)I_{d_x} + W^TW\right)^{-1} \|_{2\to2}$$
$$\le \|Z\|_{2\to2}(t+u)^{-2} \le \|Z\|_{2\to2}t^{-2},$$

which holds uniformly in $u > 0$ and follows from the fact that WW^T and W^TW are positive semidefinite.

Altogether, for $Z \in T_W(\mathcal{M}_k)$ the integrand in (43) satisfies

$$\| \left((t+u)I_{d_y} + WW^T\right)^{-1} Z \left((t+u)I_{d_x} + W^TW\right)^{-1} t^{1/N}\|_{2\to2} \le \|Z\|_{2\to2}\min\{\sigma_k^{-2}t^{-1+1/N}, t^{-2+1/N}\}.$$

The latter function is integrable over $t \in (0,\infty)$ since $N \ge 2$, and hence, for all $u \ge 0$, the integrand in (43) is uniformly dominated by an integrable function. By Lebesgue's dominated convergence theorem and continuity of $u \mapsto \left((t+u)I_{d_y} + WW^T\right)^{-1} Z \left((t+u)I_{d_x} + W^TW\right)^{-1} t^{1/N}$ for all $t \in (0,\infty)$, the function $u \mapsto \bar{\mathcal{A}}_{W,u}^{-1}(Z)$ is continuous for all $Z \in T_W(\mathcal{M}_k)$. Altogether, we showed that

$$\bar{\mathcal{A}}_W^{-1}(Z) = \frac{\sin(\pi/N)}{\pi}\int_0^\infty \left(tI_{d_y} + WW^T\right)^{-1} Z \left(tI_{d_x} + W^TW\right)^{-1} t^{1/N}dt, \quad Z \in T_W(\mathcal{M}_k),$$

and this implies (19).

A similar argument based on (42) proves (20). In the case $N = 2$, it can be shown as in [7, Theorem VII.2.3] and with the approximation argument above that

$$\bar{\mathcal{A}}_W^{-1}(Z) = \int_0^\infty e^{-t(WW^T)^{\frac12}} Z e^{-t(W^TW)^{\frac12}}\, dt,$$

which implies (21). $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

APPENDIX B. PROOF OF PROPOSITION 11

Given a system $\{\gamma_1,\ldots,\gamma_n\}$ of sufficiently smooth local coordinates on $\mathcal{M}_k$, where n is the dimension of $\mathcal{M}_k$, the vectors $\xi_j(W) := \frac{\partial}{\partial\gamma_j}(W)$, $j = 1,\ldots,n$, form a basis of the tangent space $T_W(\mathcal{M}_k)$. We need to show that the maps

$$W \mapsto F_{i,j}(W) := g_W\left(\xi_i(W),\xi_j(W)\right)$$

are continuously differentiable for all i,j. Note that the vector fields $\xi_j(W)$ are smooth in W. We consider the representation (19) and introduce the functions

$$K_{i,j,t}(W) := \mathrm{tr}\left((t+WW^T)^{-1}\xi_i(W)(t+W^TW)^{-1}\xi_j^T(W)\right),$$

where we write $(t+WW^T)$ for $(tI_{d_y} + WW^T)$ and likewise $(t+W^TW) = (tI_{d_x} + W^TW)$. For a function f we denote the differential of f at A applied to Y by $Df(A)[Y]$. Denoting $\phi_t(W) = (t+WW^T)^{-1}$ and $\psi_t(W) = (t+W^TW)^{-1}$ the product rule gives, for $W \in \mathcal{M}_k$ and $Y \in T_W(\mathcal{M}_k)$,

$$DK_{i,j,t}(W)[Y] = \mathrm{tr}(D\phi_t(W)[Y]\xi_i(W)\psi_t(W)\xi_j^T(W)) + \mathrm{tr}(\phi_t(W)D\xi_i(W)[Y]\psi_t(W)\xi_j^T(W)) \qquad (44)$$
$$+ \mathrm{tr}(\phi_t(W)\xi_i(W)D\psi_t(W)[Y]\xi_j^T(W)) + \mathrm{tr}(\phi_t(W)\xi_i(W)\psi_t(W)D\xi_j^T(W)[Y]). \qquad (45)$$

The differential of the function $\phi(A) = A^{-1}$ satisfies $D\phi(A)[Y] = -A^{-1}YA^{-1}$ so that

$$D\phi_t(W)[Y] = -(t+WW^T)^{-1}(WY^T + YW^T)(t+WW^T)^{-1},$$
$$D\psi_t(W)[Y] = -(t+W^TW)^{-1}(W^TY + Y^TW)(t+W^TW)^{-1}.$$

Let $W = U\Sigma V^T$ be the (full) singular value decomposition of W, i.e., $U \in \mathbb{R}^{d_y \times d_y}$, $V \in \mathbb{R}^{d_x \times d_x}$ with $U^T U = I_{d_y}$, $V^T V = I_{d_x}$ and $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0) \in \mathbb{R}^{d_y \times d_x}$ with $\sigma_1 \geq \cdots \geq \sigma_k > 0$. The first term on the right hand side of (44) satisfies

$$\operatorname{tr}((D\phi_t(W)[Y]\xi_i(W)\psi_t(W)\xi_j^T(W))$$
$$= -\operatorname{tr}\left((t + WW^T)^{-1}(WY^T + YW^T)(t + WW^T)^{-1}\xi_i(W)(t + W^T W)^{-1}\xi_j^T(W)\right).$$

Note that $(t + WW^T)^{-1}(WY^T + YW^T)(t + WW^T)^{-1} = Q + Q^T$ with

$$Q = (t + WW^T)^{-1}WY^T(t + WW^T)^{-1}$$
$$= U \operatorname{diag}\left(\frac{1}{t + \sigma_1^2}, \ldots, \frac{1}{t + \sigma_k^2}, \frac{1}{t}, \ldots, \frac{1}{t}\right) U^T (U\Sigma V^T Y^T) U \operatorname{diag}\left(\frac{1}{t + \sigma_1^2}, \ldots, \frac{1}{t + \sigma_k^2}, \frac{1}{t}, \ldots, \frac{1}{t}\right) U^T$$
$$= U \operatorname{diag}\left(\frac{\sigma_1}{t + \sigma_1^2}, \ldots, \frac{\sigma_k}{t + \sigma_k^2}, 0, \ldots, 0\right) V^T Y^T U \operatorname{diag}\left(\frac{1}{t + \sigma_1^2}, \ldots, \frac{1}{t + \sigma_k^2}, \frac{1}{t}, \ldots, \frac{1}{t}\right) U^T.$$

By Lemma 7 it holds

$$\xi_i(W) = P_W(\xi_i(W)) = U P_k U^T \xi_i(W)(I_{d_x} - V P_k V^T) + \xi_i(W) V P_k V^T = \zeta_i^1(W) + \zeta_i^2(W). \qquad (46)$$

where $P_k = \operatorname{diag}(1, \ldots, 1, 0, \ldots, 0)$ (with k ones on the diagonal), $\zeta_i^1(W) = U P_k U^T \xi_i(W)(I_{d_x} - V P_k V^T)$ and $\zeta_i^2(W) = \xi_i(W) V P_k V^T$. Note that also U and V are functions of W, which may be non-unique, but in this case, we just fix one choice. Then we have

$$\xi_i(W)(t + W^T W)^{-1}\xi_j^T(W) = \overbrace{\zeta_i^1(W)(t + W^T W)^{-1}(\zeta_j^1(W))^T}^{=:E_1} + \overbrace{\zeta_i^1(W)(t + W^T W)^{-1}(\zeta_j^2(W))^T}^{=:E_2}$$
$$+ \underbrace{\zeta_i^2(W)(t + W^T W)^{-1}(\zeta_j^1(W))^T}_{=:E_3} + \underbrace{\zeta_i^2(W)(t + W^T W)^{-1}(\zeta_j^2(W))^T}_{=:E_4}.$$

Using cyclicity of the trace, we obtain

$$\operatorname{tr}(QE_1) = \operatorname{tr}\left(\operatorname{diag}\left(\frac{\sigma_1}{t + \sigma_1^2}, \ldots, \frac{\sigma_k}{t + \sigma_k^2}, 0, \ldots, 0\right) V^T Y^T U \operatorname{diag}\left(\frac{1}{t + \sigma_1^2}, \ldots, \frac{1}{t + \sigma_k^2}, 0, \ldots, 0\right)\right.$$
$$\left. \times U^T \xi_i(W)(I_{d_x} - V P_k V^T) V \operatorname{diag}\left(\frac{1}{t + \sigma_1^2}, \ldots, \frac{1}{t + \sigma_k^2}, \frac{1}{t}, \ldots, \frac{1}{t}\right) V^T(I_{d_x} - V P_k V^T)\xi_j^T(W)U\right),$$

$$\operatorname{tr}(QE_2) = \operatorname{tr}\left(\operatorname{diag}\left(\frac{\sigma_1}{t + \sigma_1^2}, \ldots, \frac{\sigma_k}{t + \sigma_k^2}, 0, \ldots, 0\right) V^T Y^T U \operatorname{diag}\left(\frac{1}{t + \sigma_1^2}, \ldots, \frac{1}{t + \sigma_k^2}, 0, \ldots, 0\right)\right.$$
$$\left. \times U^T \xi_i(W)(I_{d_x} - V P_k V^T) V \operatorname{diag}\left(\frac{1}{t + \sigma_1^2}, \ldots, \frac{1}{t + \sigma_k^2}, 0, \ldots, 0\right) V^T \xi_j^T(W)U\right),$$

$$\operatorname{tr}(QE_3) = \operatorname{tr}\left(\operatorname{diag}\left(\frac{\sigma_1}{t + \sigma_1^2}, \ldots, \frac{\sigma_k}{t + \sigma_k^2}, 0, \ldots, 0\right) V^T Y^T U \operatorname{diag}\left(\frac{1}{t + \sigma_1^2}, \ldots, \frac{1}{t + \sigma_k^2}, \frac{1}{t}, \ldots, \frac{1}{t}\right)\right.$$
$$\left. \times U^T \xi_i(W) V \operatorname{diag}\left(\frac{1}{t + \sigma_1^2}, \ldots, \frac{1}{t + \sigma_k^2}, 0, \ldots, 0\right) V^T(I_{d_x} - V P_k V^T)\xi_j^T(W)U\right),$$

$$\operatorname{tr}(QE_4) = \operatorname{tr}\left(\operatorname{diag}\left(\frac{\sigma_1}{t + \sigma_1^2}, \ldots, \frac{\sigma_k}{t + \sigma_k^2}, 0, \ldots, 0\right) V^T Y^T U \operatorname{diag}\left(\frac{1}{t + \sigma_1^2}, \ldots, \frac{1}{t + \sigma_k^2}, \frac{1}{t}, \ldots, \frac{1}{t}\right)\right.$$
$$\left. \times U^T \xi_i(W) V \operatorname{diag}\left(\frac{1}{t + \sigma_1^2}, \ldots, \frac{1}{t + \sigma_k^2}, 0, \ldots, 0\right) V^T \xi_j^T(W)U\right).$$

Note that $1/(t + \sigma_i^2) \leq \min\{1/t, 1/\sigma_i^2\}$ for all $t > 0$. Using Cauchy-Schwarz inequality for the Frobenius inner product, the fact that $\|AB\|_F \leq \|A\|_{2\to 2}\|B\|_F$, and unitarity of U and V, we obtain

$$|\operatorname{tr}(QE_\ell)| \leq \|Y\|_F \|\xi_i(W)\|_F \|\xi_j(W)\|_F \sigma_k^{-3} t^{-1}, \quad \ell = 1, 2, 3, 4.$$

By continuity of ξ_i and ξ_j, it follows that there exists a neighborhood $\mathcal{U} \subset \mathcal{M}_k$ around a fixed $W_0 \in \mathcal{M}_k$ (in which $\sigma_k(W) \geq c > 0$ for some $c > 0$ and all $W \in \mathcal{U}$) and a constant $C > 0$ (depending only on the

neighborhood) such that

$$|\operatorname{tr}(Q\xi_i(W)(t+W^TW)^{-1}\xi_j^T(W))| \le C\|Y\|_F t^{-1} \quad \text{for all } t > 0 \text{ and } W \in \mathcal{U}.$$

In the same way, one shows the above inequality for Q replaced by Q^T and hence

$$|\operatorname{tr}((D\phi_t(W)[Y]\xi_i(W)\psi_t(W)\xi_j^T(W))| \le 2C\|Y\|_F t^{-1} \quad \text{for all } t > 0 \text{ and } W \in \mathcal{U}.$$

Moreover, by the Cauchy-Schwarz inequality for the Frobenius inner product and since WW^T as well as W^TW are positive semidefinite, it holds

$$\begin{aligned}
&|\operatorname{tr}((D\phi_t(W)[Y]\xi_i(W)\psi_t(W)\xi_j^T(W))| \\
&\le \|(t+WW^T)^{-1}(WY^T+YW^T)(t+WW^T)^{-1}\|_F\|\xi_i(W)(t+W^TW)^{-1}\xi_j^T(W)\|_F \\
&\le \|(t+WW^T)^{-1}\|_{2\to2}^2\|(t+W^TW)^{-1}\|_{2\to2}\|WY^T+YW^T\|_F\|\xi_i(W)\|_F\|\xi_j(W)\|_F \\
&\le 2t^{-3}\|WY^T\|_F\|\xi_i(W)\|_F\|\xi_j(W)\|_F.
\end{aligned} \tag{47}$$

Altogether, it holds, for a suitable constant $C_1 > 0$,

$$|\operatorname{tr}((D\phi_t(W)[Y]\xi_i(W)\psi_t(W)\xi_j^T(W))| \le C_1\|Y\|_F \min\{t^{-1}, t^{-3}\} \quad \text{for all } t > 0 \text{ and } W \in \mathcal{U}.$$

Let us now consider the second term on the right hand side of (44). As in (47), we obtain

$$|\operatorname{tr}(\phi_t(W)D\xi_i(W)[Y]\psi_t(W)\xi_j^T(W))| \le t^{-2}\|D\xi_i(W)[Y]\|_F\|\xi_j^T(W)\|_F.$$

By Lemma 7 we can write $\xi_j(W) = P_W(\xi_j(W)) = \zeta_j^1(W) + \zeta_j^2(W)$ with $\zeta_j^1(W) = UP_kU^T\xi_j(W)(I_{d_x} - VP_kV^T)$ and $\zeta_j^2(W) = \xi_j(W)VP_kV^T$ as in (46). With $E = I_{d_x} - VP_kV^T$ this gives

$$\begin{aligned}
&\operatorname{tr}(\phi_t(W)D\xi_i(W)[Y]\psi_t(W)\xi_j^T(W)) \\
&= \operatorname{tr}\left((t+WW^T)^{-1}D\xi_i(W)[Y](t+W^TW)^{-1}(\zeta_j^1(W)+\zeta_j^2(W))^T\right) \\
&= \operatorname{tr}\left(\operatorname{diag}\left(\frac{1}{t+\sigma_1^2},\ldots,\frac{1}{t+\sigma_k^2},0,\ldots,0\right)U^TD\xi_i(W)[Y]V\operatorname{diag}\left(\frac{1}{t+\sigma_1^2},\ldots,\frac{1}{t+\sigma_k^2},\frac{1}{t},\ldots,\frac{1}{t}\right)V^TE\xi_j^T(W)U\right) \\
&\quad + \operatorname{tr}\left(\operatorname{diag}\left(\frac{1}{t+\sigma_1^2},\ldots,\frac{1}{t+\sigma_k^2},\frac{1}{t},\ldots,\frac{1}{t}\right)U^TD\xi_i(W)[Y]V\operatorname{diag}\left(\frac{1}{t+\sigma_1^2},\ldots,\frac{1}{t+\sigma_k^2},0,\ldots,0\right)V^T\xi_j^T(W)U\right).
\end{aligned}$$

By the Cauchy-Schwarz inequality for the Frobenius inner product it follows that

$$|\operatorname{tr}(\phi_t(W)D\xi_i(W)[Y]\psi_t(W)\xi_j^T(W))| \le 2\|D\xi_i(W)[Y]\|_F\|\xi_j(W)\|_F\sigma_k^{-2}t^{-1}.$$

Since the ξ_i are continuously differentiable there exists $C_2 > 0$ such that

$$|\operatorname{tr}(\phi_t(W)D\xi_i(W)[Y]\psi_t(W)\xi_j^T(W))| \le C_2\|Y\|_F \min\{t^{-1}, t^{-2}\} \quad \text{for all } t > 0 \text{ and } W \in \mathcal{U}.$$

The terms in (45) can be bounded in the same way as the ones in (44), hence

$$|DK_{i,j,t}(W)[Y]| \le C'\|Y\|_F \min\{t^{-1}, t^{-2}\} \quad \text{for all } t > 0 \text{ and } W \in \mathcal{U}.$$

It follows that $\int_0^\infty |DK_{i,j,t}(W)[Y]|t^{1/N}dt$ exists and is uniformly bounded in $W \in \mathcal{U}$. By Lebesgue's dominated convergence theorem, it follows that we can interchange integration and differentiation, and hence, all directional derivatives of $F_{i,j}$ at W in the direction of Y exist and are continuous since $DK_{i,j,t}(W)[Y]$ is continuous in W for all $Y \in T_W(\mathcal{M}_k)$. Hence, by (19), $F_{i,j}$ is (totally) continuously differentiable for all i, j with

$$DF_{i,j}(W)[Y] = \frac{\sin(\pi/N)}{\pi}\int_0^\infty DK_{i,j,t}(W)[Y]t^{1/N}dt.$$

Hence, the metric is of class C^1 as claimed.

APPENDIX C. SOME RESULTS ON FLOWS ON MANIFOLDS

Here we summarize some notions and results on flows on manifolds that can be found in [15, Chapter IV,§2] to which we also refer for more details.

Let $p \geq 2$ be an integer or $p = \infty$, let $\mathcal{M}$ be a (finite dimensional) C^p-manifold and let ξ be a vector field of class C^{p-1} on $\mathcal{M}$. An *integral curve* for ξ with initial condition $x_0 \in \mathcal{M}$ is a C^{p-1}-curve

$$\gamma : J \to \mathcal{M},$$

where J is an open intervall containing 0 such that

$$\dot{\gamma}(t) = \xi(\gamma(t)) \quad \forall t \in J \quad \text{and} \quad \gamma(0) = x_0.$$

Theorem 44 (Theorem 2.1 in Chapter IV,§2 in [15]). *If $\gamma_1 : J_1 \to \mathcal{M}$ and $\gamma_2 : J_2 \to \mathcal{M}$ are integral curves for ξ with the same initial condition then $\gamma_1 = \gamma_2$ on $J_1 \cap J_2$.*

Let $D(\xi) \subseteq \mathbb{R} \times \mathcal{M}$ be the set of all pairs (t, x_0) such that for any $x_0 \in \mathcal{M}$ the set

$$J(x_0) := \{t \in \mathbb{R} \mid (t, x_0) \in D(\xi)\}$$

is the maximal open existence interval of an integral curve for ξ with initial condition x_0. For any $x_0 \in \mathcal{M}$, this interval is non-empty (locally one can argue as in the case $\mathcal{M} = \mathbb{R}^n$).

A *global flow* for ξ is a mapping

$$\alpha : D(\xi) \to \mathcal{M}$$

such that for all $x_0 \in \mathcal{M}$ the map $t \mapsto \alpha(t, x_0)$ for $t \in J(x_0)$ is an integral curve for ξ with initial condition x_0, i.e. $\alpha(0, x_0) = x_0$ and $\dot{\alpha}(t, x_0) = \xi(\alpha(t, x_0))$ for any $t \in J(x_0)$. Note that by Theorem 44 there is only one such mapping α. For $t \in \mathbb{R}$ let

$$D_t(\xi) := \{x \in \mathcal{M} \mid (t, x) \in D(\xi)\}$$

and define the map $\alpha_t \colon D_t(\xi) \to \mathcal{M}$ by $\alpha_t(x) = \alpha(t, x)$.

Theorem 45 (Theorems 2.6 and 2.9 in Chapter IV,§2 in [15]).

(1) *The set $D(\xi)$ is open in $\mathbb{R} \times \mathcal{M}$ and α is a C^{p-1}-morphism.*

(2) *For any $t \in \mathbb{R}$, the set $D_t(\xi)$ is open in $\mathcal{M}$ and (for $D_t(\xi)$ non-empty) α_t defines a diffeomorphism of $D_t(\xi)$ onto an open subset of $\mathcal{M}$ (namely $\alpha_t(D_t(\xi)) = D_{-t}(\xi)$) and $\alpha_t^{-1} = \alpha_{-t}$).*

APPENDIX D. THE NON-SYMMETRIC AUTOENCODER CASE FOR $N = 2$

Here we consider the optimization problem (3) with $N = 2$ and the additional constraint that $Y = X$, but we do not assume that $W_2 = W_1^T$. We also assume balanced starting conditions, i.e., $W_2(0)^T W_2(0) = W_1(0)W_1(0)^T$. In this special case, we can use a more direct approach than in Section 6 to establish some additional explicit statements below.

We write again d for $d_x = d_y$ and r for d_1. The equations for the flow here are:

$$\begin{aligned}
\dot{W}_1 &= -W_2^T W_2 W_1 X X^T + W_2^T X X^T, \\
\dot{W}_2 &= -W_2 W_1 X X^T W_1^T + X X^T W_1^T.
\end{aligned} \tag{48}$$

Next we analyze the equilibrium points of the flow (48) and of the product $W = W_2 W_1$ again assuming balanced initial conditions. We begin by exploring the equilibrium points of the flow (48) by setting the expressions in (48) equal to zero:

$$\begin{aligned}
-W_2^T W_2 W_1 X X^T + W_2^T X X^T &= 0, \\
-W_2 W_1 X X^T W_1^T + X X^T W_1^T &= 0.
\end{aligned} \tag{49}$$

If $W_2 \in \mathbb{R}^{d \times r}$ is the zero matrix then (since $X X^T$ has full rank) it follows that (49) is solved if and only if W_1 is the $r \times d$ zero-matrix, hence W is the $d \times d$ zero-matrix. The following lemma characterizes the non-trivial solutions. (The second part of the lemma is a special case of Proposition 32 below.)

Lemma 46. *The balanced nonzero solutions (i.e. solutions with $W_2 \neq 0$) of (49) are precisely the matrices of the form*

$$W_2 = UV^T, \quad W_1 = W_2^T = VU^T, \quad W = W_2 W_1 = UU^T, \tag{50}$$

where $U \in \mathbb{R}^{d \times k}$ for some $1 \leq k \leq r$ and where the columns of U are orthonormal eigenvectors of XX^T and $V \in \mathbb{R}^{r \times k}$ has orthonormal columns.

In particular, the equilibrium points for $W = W_2 W_1$ are precisely the matrices of the form

$$W = \sum_{j=1}^{k} u_j u_j^T, \tag{51}$$

where $k \in \{1, \ldots, r\}$ and $u_1, \ldots, u_k$ (the columns of U above) are orthonormal eigenvectors of XX^T.

Remark 47. *Note that $W = 0$ is also an equilibrium point of (48) which formally corresponds to taking $k = 0$ in (51).*

Proof. Since $W_2 \neq 0$, the rank k of W_2 is at least 1. The balancedness condition $W_2^T W_2 = W_1 W_1^T$ implies that W_1 and W_2 have the same singular values. Since XX^T has full rank, the first equation of (49) yields $W_2^T = W_2^T W_2 W_1$. Again due to balancedness, this shows that $W_2^T = W_1 W_1^T W_1$. It follows that all positive singular values of W_1 and of W_2 are equal to 1 and that $W_2 = W_1^T$. The second equation of (49) thus gives the equation

$$(I_d - W_2 W_2^T)XX^T W_2 = 0. \tag{52}$$

(The equilibrium points of full rank r could now be obtained using [23, Propositon 4.1] again, but we are interested in all solutions here.) Since the positive singular values of W_2 are all equal to 1, it follows that we can write

$$W_2 W_2^T = \sum_{i=1}^{k} u_i u_i^T,$$

where the u_i are orthonormal. We extend the system $u_1, \ldots, u_k$ to an orthonormal basis $u_1, \ldots, u_d$ of $\mathbb{R}^d$. From (52) we obtain $(I_d - W_2 W_2^T)XX^T W_2 W_2^T = 0$, hence $\sum_{j=k+1}^{d} u_j u_j^T XX^T \sum_{i=1}^{k} u_i u_i^T = 0$ and consequently

$$\sum_{j=k+1}^{d} \sum_{i=1}^{k} (u_j^T XX^T u_i) u_j u_i^T = 0.$$

It follows that for all $j \in \{k+1, \ldots, d\}$ and for all $i \in \{1, \ldots, k\}$ we have $u_j^T XX^T u_i = 0$. This in turn implies that XX^T maps the span of $u_1, \ldots, u_k$ into itself and also maps the span of $u_{k+1}, \ldots, u_d$ into itself. This implies that we can choose $u_1, \ldots, u_d$ as orthonormal eigenvectors of XX^T. Thus we can indeed write the (reduced) singular value decomposition of W_2 as $W_2 = UV^T$, where the columns $u_1, \ldots, u_k$ of U are orthonormal eigenvectors of XX^T and where V is as in the statement of the lemma. Since $W_1 = W_2^T$ and $W = W_2 W_1$, it follows that $W_1 = VU^T$ and $W = UU^T$ as claimed. Altogether, we have shown that the equations (50) are necessary for having a balanced solution of (49). One easily checks that W_1, W_2 defined by (50) also satisfy (49), which shows sufficiency. This completes the proof. $\square$

Corollary 48. *Consider a linear autoencoder with one hidden layer of size r with balanced initial conditions and assume that XX^T has eigenvalues $\lambda_1 \geq \ldots \geq \lambda_d > 0$ and corresponding orthonormal eigenvectors $u_1, \ldots, u_d$.*

(1) The flow $W(t)$ always converges to an equilibrium point of the form $W = \sum_{j \in J_W} u_j u_j^T$, where J_W is a (possibly empty) subset of $\{1, \ldots, d\}$ of at most r elements.

(2) The flow $W_2(t)$ converges to $UV^T =: W_2$, where the columns of U are the $u_j, j \in J_W$ and $V \in \mathbb{R}^{r \times k}$ has orthonormal columns ($k = |J_W|$) . Furthermore $W_1(t)$ converges to W_2^T.

(3) If $L^1(W(0)) < \frac{1}{2} \sum_{i=r, i \neq r+1}^{d} \lambda_i$ then $W(t)$ converges to the optimal equilibrium $W = \sum_{j=1}^{r} u_j u_j^T$.

(4) If $\lambda_r > \lambda_{r+1}$, then there is an open neighbourhood of the optimal equilibrium point in which we have convergence of the flow $W(t)$ to the optimal equilibrium point.

Proof. The first and the second point follow from Lemma 46 together with Theorem 5. (Note that if (W_1, W_2) is an equilibrium point to which the flow converges then W_1, W_2 are balanced since we assume that the flow has balanced initial conditions.) To prove the third point, note that the loss of an equilibrium point $W = \sum_{j \in J_W} u_j u_j^T$ is given by $L^1(W) = \frac{1}{2} \sum_{i \in K_W} \lambda_i$, where $K_W = \{1, \ldots, d\} \setminus J_W$. This sum is minimal for $J_W = \{1, \ldots, r\}$. Among the remaining possible J_W, the value of $L^1(W)$ is minimal for $J_W = \{1, \ldots, r+1\} \setminus \{r\}$, i.e., $K_W = \{r, \ldots, d\} \setminus \{r+1\}$. Since the value of $L^1(W(t))$ monotonically decreases as t increases (as follows e.g. from equation (24)), the claim now follows from the first point. The last point follows from the third point. □

The following result is an analogue to Theorem 25.

Theorem 49. *If $k \leq r$ and $u_1, \ldots, u_k$ are orthonormal eigenvectors of XX^T which do not form a system of eigenvectors to the r largest eigenvalues of XX^T (in particular for $k < r$), in any neighborhood of the equilibrium point $W = \sum_{j=1}^k u_j u_j^T$ there is some $\widetilde{W}$ of rank at most r for which $L^1(\widetilde{W}) < L^1(W)$. In particular, the equilibrium in W is non-stable.*

Proof. If $k < r$ and $W = \sum_{j=1}^k u_j u_j^T$ for orthonormal eigenvectors u_j of XX^T then for any additional eigenvector u_{k+1} orthonormal to the u_j and for any $\varepsilon > 0$, we can choose $\widetilde{W} = W + \varepsilon u_{k+1} u_{k+1}^T$ to obtain $L^1(\widetilde{W}) < L^1(W)$. Let now $k = r$. This case can be treated analogously to the proof of Theorem 25: let u_i be one of the eigenvectors $u_1, \ldots, u_r$ whose eigenvalue does not belong to the r largest eigenvalues of XX^T. Let v be an eigenvector of XX^T of unit length which is orthogonal to the eigenvectors $u_1, \ldots, u_r$ and whose eigenvalue belongs to the r largest eigenvalues of XX^T. Now for any $\varepsilon \in [0, 1]$ consider $u_i(\varepsilon) := \varepsilon v + \sqrt{1 - \varepsilon^2} u_i$. Then $W(\varepsilon) := u_i(\varepsilon) u_i(\varepsilon)^T + \sum_{j=1, j \neq i}^r u_j u_j^T$ satisfies $L^1(W(\varepsilon)) < L^1(W)$ for $\varepsilon \in (0, 1]$. From this the claim follows. □

Remark 50. *With the notation*

$$V = \begin{pmatrix} W_1^T \\ W_2 \end{pmatrix} \in \mathbb{R}^{2d \times r} \; \text{ and } \; C = XX^T \in \mathbb{R}^{d \times d}$$

and assuming that C has full rank, the flow (48) can be written as the following Riccati-type-like ODE.

$$\dot{V} = \left(I_{2d} + \begin{pmatrix} -C & 0 \\ 0 & 0 \end{pmatrix} VV^T \begin{pmatrix} 0 & 0 \\ C^{-1} & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & -I_d \end{pmatrix} VV^T \begin{pmatrix} 0 & I_d \\ 0 & 0 \end{pmatrix} \right) \begin{pmatrix} 0 & C \\ C & 0 \end{pmatrix} V. \tag{53}$$