Journal of Machine Learning Research 25 (2024) 1-76 Submitted 4/23; Revised 4/24; Published 8/24

The Loss Landscape of Deep Linear Neural Networks: a Second-order Analysis

El Mehdi Achour ACHOUR@MATHC.RWTH-AACHEN.DE
Department of Mathematics
RWTH Aachen University
Aachen, Germany

François Malgouyres FRANCOIS.MALGOUYRES@MATH.UNIV-TOULOUSE.FR
Institut de Mathématiques de Toulouse ; UMR 5219
Université de Toulouse ; CNRS
UPS IMT F-31062 Toulouse Cedex 9, France

Sébastien Gerchinovitz SEBASTIEN.GERCHINOVITZ@IRT-SAINTEXUPERY.FR
Institut de Recherche Technologique Saint Exupéry, Toulouse, France

Editor: Martin Jaggi

Abstract

We study the optimization landscape of deep linear neural networks with square loss. It is known that, under weak assumptions, there are no spurious local minima and no local maxima. However, the existence and diversity of non-strict saddle points, which can play a role in first-order algorithms' dynamics, have only been lightly studied. We go a step further with a complete analysis of the optimization landscape at order 2. Among all critical points, we characterize global minimizers, strict saddle points, and non-strict saddle points. We enumerate all the associated critical values. The characterization is simple, involves conditions on the ranks of partial matrix products, and sheds some light on global convergence or implicit regularization that has been proved or observed when optimizing linear neural networks. In passing, we provide an explicit parameterization of the set of all global minimizers and exhibit large sets of strict and non-strict saddle points.

Keywords: Deep learning, landscape analysis, non-convex optimization, second-order geometry, strict saddle points, non-strict saddle points, global minimizers, implicit regularization

1. Introduction

Deep learning has been widely used recently due to its good empirical performances in image recognition, natural language processing, and speech recognition, among other fields. However, there is still a gap between theory and practice. One of the aspects that are partially missing in the picture is why gradient-based algorithms can achieve low training error despite a non-convex objective. Another partially open question is why they generalize well to unseen data despite many more parameters than the number of points in the training set, and how implicit regularization can help. One important research direction analyses the landscape of the empirical risk. In this paper, we characterize the local structures around critical points of the empirical risk, for deep linear neural networks with the square loss.

Before summarizing the related literature and our main contributions, we first recall definitions that will be key throughout the paper.

1.1 Reminder: Minimizers, Critical Points of Order 1 or 2, Strict and Non-strict Saddle Points

Let us recall the definitions of local structures of the landscape of the empirical risk, which are important from the statistical and optimization points of view.

For $\boldsymbol{w} \in \mathbb{R}^n$, denote by $\boldsymbol{w} \longmapsto L(\boldsymbol{w})$ the function we want to minimize. Assume that $\boldsymbol{w} \longmapsto L(\boldsymbol{w})$ is C^2, and denote by ∇L and $\nabla^2 L$ its gradient and its Hessian.[1] We also write $A \succeq 0$ to say that a matrix $A \in \mathbb{R}^{n \times n}$ is positive semi-definite. Recall the following four definitions, which are nested:

- $\boldsymbol{w^*}$ is a **global minimizer** if and only if $\forall \boldsymbol{w} \in \mathbb{R}^n,\ L(\boldsymbol{w}^*) \leq L(\boldsymbol{w})$.
- $\boldsymbol{w^*}$ is a **local minimizer** if and only if there exists a neighbourhood $\mathcal{O} \subset \mathbb{R}^n$ of $\boldsymbol{w^*}$ such that $\forall \boldsymbol{w} \in \mathcal{O},\ L(\boldsymbol{w}^*) \leq L(\boldsymbol{w})$.
- $\boldsymbol{w^*}$ is a **second-order critical point** if and only if $\nabla L(\boldsymbol{w}^*) = 0$ and $\nabla^2 L(\boldsymbol{w}^*) \succeq 0$. If, on the contrary, the Hessian has a negative eigenvalue, we say that the point has a negative curvature.
- $\boldsymbol{w^*}$ is a **first-order critical point** if and only if $\nabla L(\boldsymbol{w}^*) = 0$.

  We can also distinguish a specific type of first-order critical point: saddle points. As discussed below, they can be second-order critical points or not.[2]

- $\boldsymbol{w^*}$ is a **saddle point** if and only if it is a first-order critical point which is neither a local minimizer nor a local maximizer.
  - A saddle point $\boldsymbol{w^*}$ is **strict** if and only if it is not a second-order critical point (i.e., the Hessian $\nabla^2 L(\boldsymbol{w}^*)$ has a negative eigenvalue). Figure 2 gives an example.
  - A saddle point $\boldsymbol{w^*}$ is **non-strict** if and only if it is a second-order critical point. In that case, the Hessian $\nabla^2 L(\boldsymbol{w}^*)$ is positive semi-definite and has at least one eigenvalue equal to zero. Typically, in the direction of the corresponding eigenvectors, a higher-order term makes it a saddle point (e.g., $L(\boldsymbol{w}) = \sum_{i=1}^n w_i^3$ at $\boldsymbol{w}^* = 0$). Figure 1 gives an example.

1.2 On the Importance of a Landscape Analysis at Order 2

When the function we are trying to minimize is smooth, convex, and has a global minimizer, the gradient descent algorithm with a well-chosen learning rate converges to a first-order critical point, which is a global minimizer (Nesterov, 1998). However, in general, finding a global optimum of a non-convex function is an NP-complete problem (Murty and Kabadi, 1987); this is, in particular, the case for a simple 3-node neural network (Blum and Rivest, 1989). Despite that, when optimizing neural networks, the current practice is still to use gradient-based algorithms.

1. When the input parameter is not a vector, but, e.g., a sequence of matrices, the same definitions hold, where the gradient and the Hessian are computed with respect to the vectorized version of the input parameters.
2. Defining the index of a critical point as the number of negative eigenvalues of its Hessian, we can equivalently define strict saddle points as saddle points of index greater than or equal to 1. Similarly, non-strict saddle points are saddle points of index 0. Note that the latter are degenerate, i.e., their Hessian is singular.

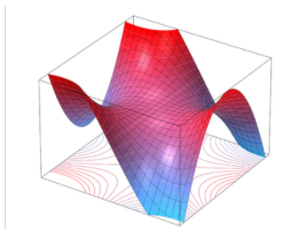

Figure 1: Example of a landscape with a plateau (non-strict saddle point).

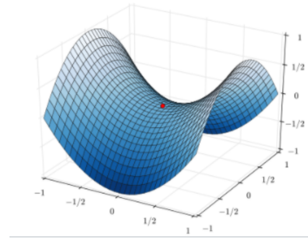

Figure 2: Example of a landscape with a strict saddle point at (0,0).

It has been known for decades that, even in the non-convex setting, for large classes of functions, gradient-based algorithms converge to a first-order critical point, in the sense that the iterates produced by the algorithm reach an arbitrary small gradient after a finite (polynomial) number of iterations (Nesterov, 1998). Recent works have shown that classical first-order algorithms escape strict saddle points (Lee et al., 2016, 2019). Well-chosen algorithms can be stopped at an output with arbitrarily small gradient and nearly-positive semi-definite Hessian in polynomial time (Jin et al., 2017, 2018; Daneshmand et al., 2018; Jin et al., 2021; Gadat and Gavra, 2022). Higher order algorithms, designed to escape strict saddle points, have been constructed and have a faster convergence (e.g., Adolphs et al., 2019; O'Neill and Wright, 2023). However, nothing prevents these algorithms to spend many epochs in the vicinity of non-strict saddle points. This results in a long plateau during training.

To see that this behavior actually occurs in practice, consider the simple experiment whose results are shown in Figures 3 and 4 (more details in Appendix G). For each run of this experiment, the parameters of a linear neural network of depth 5 are optimized to fit random input/output pairs. The discrepancy is measured with the square loss and we use the ADAM optimizer. Depending on the run, the algorithm is initialized in the vicinity either of a strict saddle point (in red) or a non-strict saddle point (in blue). The distance between the random initial iterate and the saddle point is purposely not negligible: it is fixed to around 10% of the norm of the saddle point. Figure 3 shows the typical loss evolution for both cases. We can see that ADAM rapidly escapes from the strict saddle point but needs many epochs to escape the plateau in the vicinity of the non-strict saddle point. Figure 4 shows that this observation generalizes to most runs. We compare the empirical distributions of a random time (called escape epoch) defined as the epoch at which the loss has significantly decreased from its initial value. When initialized in the vicinity of non-strict saddle points, the algorithm suffers from an often large escape epoch and might be stopped there, without the possibility to distinguish this non-strict saddle point from a global minimum. Improving the analysis beyond local minimizers and characterizing strict and non-strict saddle points are therefore key to understanding gradient descent dynamics and implicit regularization.

1.3 Related Works on Linear Networks

Despite the fact that they are rarely used to solve real-world applications[3], many recent works have focused on linear neural networks. These studies are motivated by the fact that the empirical risk of linear networks is nonconvex and shares similar properties with practical nonlinear neural networks.

3. They indeed compute a linear map between the input and output spaces.

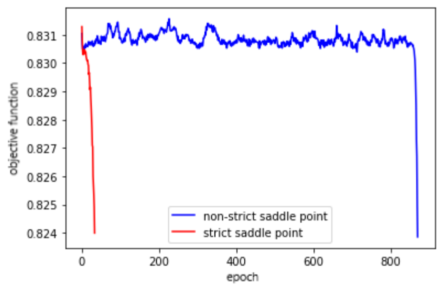

Figure 3: The loss function during the iterative process, when initialized around a strict saddle point (in red) or a non-strict saddle point (in blue).

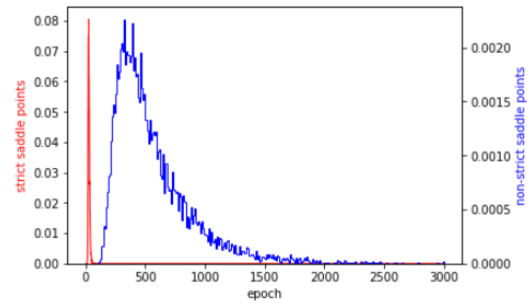

Figure 4: Histogram of escape epochs, when initialized around a strict (in red) or a non-strict saddle point (in blue). For clarity, the y-axis is endowed with two scales. The right axis corresponds to the blue curve and the left to the red one.

Indeed, as shown in Saxe et al. (2014), linear networks exhibit nonlinear learning phenomena similar to those seen during the optimization of nonlinear networks, including long plateaus followed by rapid transitions to lower error solutions. Also, the implicit regularization phenomena observed for nonlinear networks (Safran et al., 2022; Timor et al., 2023; Jacot, 2022; Marion et al., 2024; Belkin, 2021; Bartlett et al., 2021) occurs also for linear networks (see the paragraph on this topic below). Studying these phenomena for linear networks is a good starting point for rigorous work.

The study of linear neural networks can be divided into two categories. The first line of research studies the geometric landscape of the empirical risk. The second line studies the trajectory of gradient descent dynamics in linear networks. Our work falls into the first category.

Geometric landscape for linear networks: This first started with Baldi and Hornik (1989). They proved that for a 1-hidden layer linear network, under some conditions on the data matrices, and for the square loss, every local minimizer is a global minimizer. Kawaguchi (2016) later generalized and extended this result to deep linear neural networks under mild conditions and again proved that every local minimizer is a global minimizer (this part has been proved later by Lu and Kawaguchi (2017) with weaker assumptions on the data and simpler proofs). This author also proved that every other critical point is a saddle point, that for a 1-hidden layer linear network all saddle points are strict, while for deeper networks, there exist non-strict saddle points (Kawaguchi (2016) exhibits a space of non-strict saddle points where all but one weight matrix are equal to zero). Yun et al. (2018) gave a condition for a critical point to be either a global minimizer or a saddle point. Zhou and Liang (2018) removed all assumptions on the data and gave analytical forms for the critical points of the empirical risk. In the characterization, the weight matrices are defined recursively and can be found by solving equations; in particular, they gave a characterization of global minimizers. Nouiehed and Razaviyayn (2022) showed using assumptions only on the width of the layers that every local minimizer is a global minimizer. They prove that this assumption on the architecture is sharp in the sense that without it, and if we do not make assumptions on the data matrices as in previous works, then there exists a poor local minimizer. Zhu et al. (2020) used assumptions only on the input data matrix, to prove that for a 1-hidden layer linear network, every local minimizer is a global minimizer and every other critical point has a negative curvature. Laurent and von Brecht (2018) proved for different general convex losses that, under assumptions on the architecture, all local minima are global. Finally, Trager et al. (2020) and Mehta et al. (2021) used results from algebraic geometry to give other properties about critical points of linear networks.

Most of the previous works focus on local minimizers. None of these works provide simple necessary and sufficient conditions for a saddle point to be strict or not.[4] In particular, in the case of more than two hidden layers, only very specific examples of non-strict saddle points were described. Furthermore, global minimizers were characterized but not explicitly parameterized. See Section 3.4 for more details.

Gradient dynamics and implicit regularization for linear networks: In this line of research, authors study the dynamics of first-order algorithms for linear networks, which they sometimes combine with results about the loss landscape. Arora et al. (2019a) proved that gradient descent converges to a global minimum at a linear rate, under assumptions on the width of the layers, the initial iterate, and the loss at initialization. Other works also proved similar results with different assumptions (Eftekhari, 2020; Bartlett et al., 2018; Wu et al., 2019). However, as noted by Shamir (2019), these works consider strong assumptions on the loss at initialization. Indeed, Shamir (2019) gave a negative result on a deep linear network of width 1, by proving that for standard initializations, gradient descent can take exponential time to converge to the global minimizer. The author also provided empirical examples of the same phenomenon happening for larger widths. On the other hand, Du and Hu (2019) proved that if the layers are wide enough, convergence to a global minimimizer can be achieved in polynomial time using a classical data-independent random Gaussian initialization (known as Xavier initialization). The required minimum width of the network depends on the norm of a global minimizer of the linear regression problem. As we will see in Section 3.4 this global convergence result can be re-interpreted in terms of the loss landscape at order 2.

On a similar line of research, Chitour et al. (2023) proved using assumptions on the architecture of the network and the data matrices that gradient flow almost surely converges to a global minimizer for a 1-hidden layer linear network. Later, Bah et al. (2022) proved the same result under weaker assumptions. They also proved that, in deep linear networks, the gradient flow almost surely converges to global minimizers of the rank-constrained linear regression problem. This has been extended to gradient descent in Nguegnang et al. (2024). In Jacot et al. (2022), the authors conjecture that, for deep linear networks, the gradient flow initialized randomly in the vicinity of the origin, asymptotically exhibits a saddle-to-saddle dynamics, where the rank of the linear map increases at each new saddle.

This is related to another consequence of the landscape properties: implicit regularization. Arora et al. (2019b) showed that, for matrix recovery, deep linear networks converge to low-rank solutions even when all the hidden layers are of size larger than or equal to the input and output sizes. Razin and Cohen (2020) proved that, in deep matrix factorization, implicit regularization may not be explainable by norms, as all norms may go to infinity. They rather suggest seeing implicit regularization as a minimization of the rank. Saxe et al. (2019) and Gidel et al. (2019) proved with

4. By “simple”, we mean an easier-to-exploit condition than just looking at the smallest eigenvalue of the Hessian.

different assumptions on the data and a vanishing initialization that both gradient flow and discrete gradient dynamics sequentially learn solutions of a rank-constrained linear regression problem with a gradually increasing rank. Finally, Gissin et al. (2019) proved for a toy model that this incremental learning happens more often (with larger initialization), when the depth of the network increases. As we will see in Section 3.4, these results can be re-interpreted in the light of the landscape at order 2.

1.4 Summary of our Contributions

Our contributions on the optimization landscape of deep linear networks can be summarized as follows.

- We characterize the square loss landscape of deep linear networks at order 2 (see Theorem 7 and Figure 6). That is, under some classical and weak assumptions on the data, we characterize, among all first-order critical points, which are global minimizers, strict saddle points, and non-strict saddle points. The characterization is simple and involves conditions on the ranks of partial matrix products. To the best of our knowledge, this is the first simple, necessary and sufficient condition that differentiates strict saddle points from non-strict saddle points.

- Several results follow from the characterization: under the same assumptions,
  - we first immediately recover the fact that all saddle points are strict for one-hidden layer linear networks;
  - more importantly, for deeper networks, when proving that all cases considered in the characterization can indeed occur, we exhibit large sets of strict and non-strict saddle points (see Proposition 8 and its proof in Appendix B.8);
  - we show that the non-strict saddle points are associated with r_{max} plateau values of the empirical risk, where r_{max} is the size of the thinnest layer of the network (see Theorem 7). Typically these are values of the empirical risk that first-order algorithms can take for some time, as in Figure 3, and which might be confused with a global minimum.

- As a by-product of our analysis, we obtain explicit parameterizations of sets containing or included in the set of all first-order critical points (see Propositions 9 and 10). We also derive an explicit parameterization of the set of all global minimizers (see Proposition 11).

The above results are compared in details with previous works in Section 3.4. In particular, our second-order characterization sheds some light on two phenomena:

- Implicit regularization: we recover the fact that every non-strict saddle point corresponds to a global minimizer of the rank-constrained linear regression problem, as shown in (Bah et al., 2022, Proposition 35). Our characterization additionally shows that only a fraction of the critical points corresponding to rank-constrained solutions are non-strict saddle points. The others are strict saddle points. Given the differences in the behavior of first-order algorithms in the vicinity of strict and non-strict saddle points as illustrated on Figures 3 and 4, our results open new research directions related to the very nature of implicit regularization and its stability.

- Our characterization can also be useful to understand recent global convergence results in terms of the loss landscape at order 2. In particular, we show how to re-interpret a proof of

Du and Hu (2019) to see that gradient descent with Xavier initialization on wide enough deep linear networks meets no non-strict saddle points on its trajectory.

1.5 Outline of the Paper

The paper is organized as follows. We define the setting in Section 2 and state our results in Section 3. We prove our main result (Theorem 7) in Section 4. More precisely, we detail the proof structure and main arguments but defer all technical derivations to the appendix. We finally conclude our work in Section 5.

Most technical details can be found in the appendix, which is organized as follows. Section A contains additional notation and lemmas that will be useful in all subsequent sections. In Section B we provide proofs of propositions and lemmas related to first-order critical points, while Section C gathers the proofs for the parameterization of first-order critical points and global minimizers. Sections D, E, and F contain proofs corresponding to each subsection of Section 4. Finally, in Section G, we describe in more details the illustrative experiment underlying Figures 3 and 4.

2. Setting

In this section we formally define our setting (deep linear networks with square loss), set some notation, and describe our assumptions on the data.

Model and notation: We consider a fully-connected linear neural network of depth $H \geq 2$. The neural network consists of H layers and maps any input $x \in \mathbb{R}^{d_x}$ to an output $W_H \cdots W_1 x \in \mathbb{R}^{d_y}$, where $W_H \in \mathbb{R}^{d_y \times d_{H-1}}, \dots, W_h \in \mathbb{R}^{d_h \times d_{h-1}}, \dots, W_1 \in \mathbb{R}^{d_1 \times d_x}$, are the matrices associated with the H layers (d_h is the width of layer h). We set $d_H = d_y$ and $d_0 = d_x$. The input layer is of size d_x and the output layer is of size d_y. We also define the smallest width of the layers as $r_{max} = \min(d_H, \dots, d_0)$.[5] We denote the parameters of the model by $\mathbf{W} = (W_H, \dots, W_1)$.

Let $(x_i, y_i)_{i=1..m}$ with $x_i \in \mathbb{R}^{d_x}$ and $y_i \in \mathbb{R}^{d_y}$, be the training set that we gather column-wise in matrices $X \in \mathbb{R}^{d_x \times m}$ and $Y \in \mathbb{R}^{d_y \times m}$. We consider the empirical risk L defined by:

$$L(\mathbf{W}) = \sum_{i=1}^{m} \|W_H W_{H-1} \cdots W_2 W_1 x_i - y_i\|_2^2 = \|W_H \cdots W_1 X - Y\|^2 ,$$

where $\|.\|_2$ is the Euclidean norm and $\|.\|$ denotes the Frobenius norm of a matrix.
We set:

$$\Sigma_{XX} = \sum_{i=1}^{m} x_i x_i^T = XX^T \in \mathbb{R}^{d_x \times d_x} \quad , \quad \Sigma_{YY} = \sum_{i=1}^{m} y_i y_i^T = YY^T \in \mathbb{R}^{d_y \times d_y},$$

$$\Sigma_{XY} = \sum_{i=1}^{m} x_i y_i^T = XY^T \in \mathbb{R}^{d_x \times d_y} \quad , \quad \Sigma_{YX} = \sum_{i=1}^{m} y_i x_i^T = YX^T \in \mathbb{R}^{d_y \times d_x},$$

where, A^T denotes the transpose of A.

5. The notation r_{max} comes from the fact that it is the maximum possible rank of the product $W_H \cdots W_1$.

Assumption 1 *Throughout the article, we assume that* $d_y \leq d_x \leq m$*, that* Σ_{XX} *is invertible, and that* Σ_{XY} *is of full rank* d_y*. We define* $\Sigma^{1/2} = \Sigma_{YX}\Sigma_{XX}^{-1}X \in \mathbb{R}^{d_y \times m}$ *and* $\Sigma = \Sigma^{1/2}(\Sigma^{1/2})^T = \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY} \in \mathbb{R}^{d_y \times d_y}$*. We assume that the singular values of* $\Sigma^{1/2}$ *are all distinct (i.e., that* Σ *has* d_y *distinct eigenvalues).*

These assumptions are exactly the ones considered in Kawaguchi (2016). Note that we do not make any assumption on the width of the hidden layers. As noted by Baldi and Hornik (1989), full-rank matrices are dense, and deficient-rank matrices are of measure 0. In general, $m \geq d_x \geq d_y$, which is the classical learning regime, is essentially sufficient to have the other assumptions verified, due to the randomness of the data.
Let

$$\Sigma^{1/2} = U\Delta V^T \tag{1}$$

be a singular value decomposition of $\Sigma^{1/2}$, where $U \in \mathbb{R}^{d_y \times d_y}$ and $V \in \mathbb{R}^{m \times m}$ are orthogonal, and the diagonal elements of $\Delta \in \mathbb{R}^{d_y \times m}$ are in decreasing order.
Since $\Sigma = \Sigma^{1/2}(\Sigma^{1/2})^T$, Σ can be diagonalized as $\Sigma = U\Lambda U^T$ where $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_{d_y})$, with $\lambda_1 > \cdots > \lambda_{d_y} \geq 0$. Moreover, a consequence of Assumption 1 is that Σ is positive definite (see Lemma 20); therefore, we have $\lambda_{d_y} > 0$.

Additional notation: We list below some notation and conventions that will be used throughout the paper.
For all integers $a \leq b$, we denote by $[\![a, b]\!]$ the set of integers between a and b (including a and b). If $a > b$, $[\![a, b]\!]$ is the empty set (e.g. $[\![1, 0]\!] = \emptyset$).
If $\mathcal{S} = \emptyset$, then $\sum_{i \in \mathcal{S}} \lambda_i = 0$.
Given a matrix $A \in \mathbb{R}^{p \times q}$, $\text{col}(A)$, $\text{Ker}(A)$ and $\text{rk}(A)$, denote respectively the column space, the null space and the rank of A.
For a matrix $A \in \mathbb{R}^{p \times q}$, we write $A_i \in \mathbb{R}^p$ for the i-th column of A and $A_{\mathcal{J}} \in \mathbb{R}^{p \times |\mathcal{J}|}$ for the sub-matrix obtained by concatenating the column vectors A_i, for $i \in \mathcal{J}$. The identity matrix of size p will be denoted by I_p.
When we write $W_h \cdots W_{h'}$ for $h > h'$, the expression denotes the product of all W_j from $j = h$ to $j = h'$. To simplify later developments, we allow two additional cases: when $h = h'$, the expression simply denotes W_h, and when $h' = h + 1$, it stands for the identity matrix $I_{d_h} \in \mathbb{R}^{d_h \times d_h}$.

Considering submatrices of compatible sizes, we define a block matrix by one of the three following ways:

- $[A, B]$ is the horizontal concatenation of the matrices A and B;
- $\begin{bmatrix} G \\ H \end{bmatrix}$ is the vertical concatenation of G and H;
- $\begin{bmatrix} C & D \\ E & F \end{bmatrix}$ is a 2×2 block matrix.

By convention, in block matrices, some blocks can have 0 lines or 0 columns; this means that such blocks do not exist. However if we have a product between two matrices that have 0 as the

common size (the number of columns for the first matrix, of the lines for the second matrix), then their product equals a zero matrix, of the right size. More formally, if $A \in \mathbb{R}^{n\times 0}$ and $B \in \mathbb{R}^{0\times p}$, then, by convention, $AB = 0_{n\times p}$. Note that the product of block matrices is compatible with this convention (e.g., $[A\ ,\ B]\begin{bmatrix} C \\ D \end{bmatrix} = AC + BD$ is still true if $B \in \mathbb{R}^{n\times 0}$ and $D \in \mathbb{R}^{0\times p}$).
Further notation that are used in the appendix can be found at the beginning of Appendix A.

3. Main Results

In this section, we state the main results of this paper. We start with a necessary condition for being a first-order critical point of L (Proposition 1), to which we give a light reciprocal (Proposition 2). We then move to our main result (Theorem 7), which is a second-order classification of all first-order critical points. It distinguishes between global minimizers, strict saddle points and non-strict saddle points. Finally, the third result is a necessary parameterization for critical points (Proposition 9) and an explicit parameterization of all global minimizers (Proposition 11). These results are compared with previous works in Section 3.4. All the proofs can be found in Section 4 or in the appendix, where most technical derivations are deferred.

3.1 First-order Critical Points: Preliminary Results

In the next proposition, we restate in our framework a necessary condition for being a first-order critical point, which was already present in Baldi and Hornik (1989) and most of the papers in this line of research. This proposition will serve later to distinguish between different types of critical points.

Proposition 1 (Global map and critical values) *Suppose Assumption 1 in Section 2 holds true. Let $\mathbf{W} = (W_H, \ldots, W_1)$ be a first-order critical point of L and set $r = rk(W_H \cdots W_1) \in [\![0, r_{max}]\!]$. There exists a unique subset $\mathcal{S} \subset [\![1, d_y]\!]$ of size r such that:*

$$W_H \cdots W_1 = U_{\mathcal{S}} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1},$$

where U was defined in (1)*. We say that the critical point $\mathbf{W}$ is associated with $\mathcal{S}$. The associated critical value is*

$$L(\mathbf{W}) = \mathrm{tr}(\Sigma_{YY}) - \sum_{i\in\mathcal{S}} \lambda_i.$$

The proof can be found in Appendix B.2. The result is true even for $r = 0$, using the conventions from Section 2 (in this case, $\mathcal{S} = \emptyset$).
Note that $\Sigma_{YX}\Sigma_{XX}^{-1}$ corresponds to the solution of the classical linear regression problem. Therefore, we can see that for every critical point $\mathbf{W}$ of L, the product $W_H \cdots W_1$ is the projection of this least-squares estimator onto a subspace generated by a subset of the eigenvectors of Σ. Note that $\mathrm{tr}(\Sigma_{YY}) = \|Y\|^2$.

The following proposition is a light reciprocal to Proposition 1, by showing that all subsets $\mathcal{S}$ and the corresponding critical values $\mathrm{tr}(\Sigma_{YY}) - \sum_{i\in\mathcal{S}} \lambda_i$ are associated to an existing critical point. In particular, the largest critical value is reached for $\mathcal{S} = \emptyset$ and the smallest critical value for $\mathcal{S} = [\![1, r_{max}]\!]$.

Proposition 2 *Suppose Assumption 1 in Section 2 holds true. For any $\mathcal{S} \subset [\![1, d_y]\!]$ of size $r \in [\![0, r_{max}]\!]$, there exists a first-order critical point* **W** *associated with $\mathcal{S}$.*

The proof of Proposition 2 is deferred to Appendix B.6. The proof uses Proposition 10, which is proved in Appendix B.5, before Appendix B.6.

3.2 Second-order Classification of the Critical Points of L

The main result of this section is Theorem 7 below, where we classify all first-order critical points into global minimizers, strict saddle points and non-strict saddle points. To state Theorem 7 we first need to introduce some definitions.

Let $\mathbf{W} = (W_H, \ldots, W_1)$ be a first-order critical point of L. Below, we introduce the notions of complementary block, tightened pivot and tightened critical point that are key to the main results. Consider the sequence of H matrices $W_H, \ldots, W_2, W_1$ and connect them by plugging Σ_{XY} between W_1 and W_H so as to form a cycle as on Figure 5. Note that the dimensions of these matrices allow us to consider any product of consecutive matrices on this cycle, e.g., $W_H W_{H-1} W_{H-2}$ or $W_2 W_1 \Sigma_{XY} W_H$ (the matrix Σ_{XY} between W_1 and W_H is key here). Such products of consecutive matrices in the cycle are what we call "**blocks**". In the sequel, we call "**pivot**" any pair of indices $(i, j) \in [\![1, H]\!]$, with $i > j$, and we consider blocks around a pivot (i, j), as defined formally below.

Definition 3 (Complementary blocks) *Let* $\mathbf{W} = (W_H, \ldots, W_1)$ *be a first-order critical point of L. For any pivot $(i, j) \in [\![1, H]\!]$, $(i > j)$, we define the two complementary blocks to (i, j) as:*

$$W_{j-1} \cdots W_1 \Sigma_{XY} W_H \cdots W_{i+1} \qquad \textit{and} \qquad W_{i-1} \cdots W_{j+1}.$$

The general case is represented on Figure 5.
Note that, when $i = j + 1$, the second complementary block is $W_j W_{j+1}$, which using the convention in Section 2 is I_{d_j}. Similarly, if $i = H$ and $j = 1$, the first complementary block is Σ_{XY}. First we state a proposition about the ranks of the complementary blocks which is key to our analysis.

Proposition 4 *Suppose Assumption 1 in Section 2 holds true. Let* $\mathbf{W} = (W_H, \ldots, W_1)$ *be a first-order critical point of L and $r = \mathrm{rk}(W_H \cdots W_1)$. For any pivot (i, j), the rank of each of the two complementary blocks is larger than or equal to r.*

The proof is in Appendix B.7. The boundary case when at least one of the two ranks is equal to r plays a special role in the loss landscape at order 2.

Definition 5 (Tightened pivot) *Let* $\mathbf{W} = (W_H, \ldots, W_1)$ *be a first-order critical point of L and let $r = \mathrm{rk}(W_H \cdots W_1)$.*
We say that a pivot (i, j) is ***tightened*** *if and only if at least one of the two complementary blocks to (i, j) is of rank r.*

Definition 6 (Tightened critical point) *Let* $\mathbf{W} = (W_H, \ldots, W_1)$ *be a first-order critical point of L. We say that* **W** *is tightened if and only if every pivot (i, j) is tightened.*

When $H \geq 3$, note that a sufficient condition for a first-order critical point **W** to be tightened is the existence of three weight matrices W_{h_1}, W_{h_2} and W_{h_3} of rank $r = \mathrm{rk}(W_H \cdots W_1)$. This is

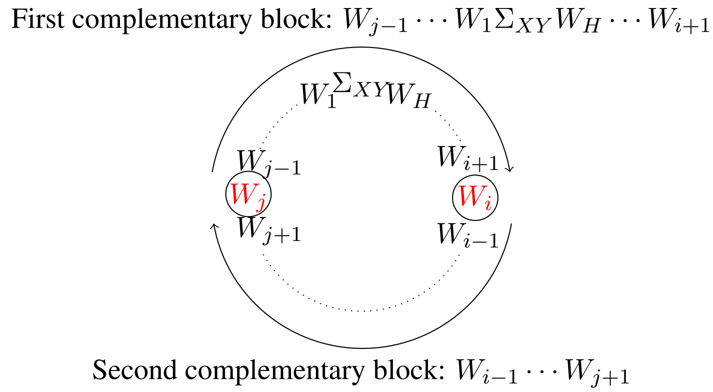

Figure 5: Complementary blocks to the pivot (i, j) .

a simple intuition on tightened critical points, that the reader can keep in mind when reading the article. A special case of this is when **W** is 0-balanced (Definition 1 in Arora et al. (2019a)), that is, when $W_{j+1}^T W_{j+1} = W_j W_j^T$ for all $j \in [\![1, H-1]\!]$. Indeed, in that case, the weight matrices W_j have equal ranks and $(W_H \cdots W_1)(W_H \cdots W_1)^T = W_H \cdots W_2 (W_1 W_1^T) W_2^T \cdots W_H^T = W_H \cdots W_2 (W_2^T W_2) W_2^T \cdots W_H^T = W_H \cdots W_3 (W_2 W_2^T)^2 W_3^T \cdots W_H^T = \ldots = (W_H W_H^T)^H$, so that $\mathrm{rk}(W_j) = \mathrm{rk}(W_H) = \mathrm{rk}(W_H \cdots W_1) = r$ for all $j \in [\![1, H]\!]$. Therefore, when $H \geq 3$, first-order critical points that are 0-balanced are tightened.

Note also that when $H = 2$, there is no tightened critical point with $r < r_{max}$, because the pivot $(2, 1)$ is not tightened (both complementary blocks Σ_{XY} and I_{d_1} are of full rank, which is larger than or equal to $r_{max} = \min\{d_y, d_1, d_x\}$).

We can now state our main theorem, which characterizes the nature of any first-order critical point **W** depending on the associated index set $\mathcal{S}$ and the tightening condition. The corresponding classification is illustrated on Figure 6. Note that Theorem 7 precisely differentiates between first-order critical points that are second-order critical points and those that are not. Combined with the fact that every first-order critical point is either a global minimizer or a saddle point (Kawaguchi, 2016), we can distinguish global minimizers, strict saddle points and non-strict saddle points. The main and most technical contribution is, in the case $\mathcal{S} = [\![1, r]\!]$, to distinguish between strict and non-strict saddle points.

We recall that $r_{max} = \min(d_H, \ldots, d_0)$ is the width of the thinnest layer, and that U corresponds to the eigenvectors of Σ (see (1)).

Theorem 7 (Classification of the critical points of L) *Suppose Assumption 1 in Section 2 holds true.*

Let **W** $= (W_H, \ldots, W_1)$ *be a first-order critical point of* L *and set* $r = \mathrm{rk}(W_H \cdots W_1) \leq r_{max}$. *Following Proposition 1, we consider the index set* $\mathcal{S}$ *associated with* **W**.

- *When* $r = r_{max}$*:*
  - *if* $\mathcal{S} = [\![1, r_{max}]\!]$*, then* **W** *is a global minimizer.*
  - *if* $\mathcal{S} \neq [\![1, r_{max}]\!]$*, then* **W** *is not a second-order critical point (***W** *is a strict saddle point).*
- *When* $r < r_{max}$*:* **W** *is a saddle point.*

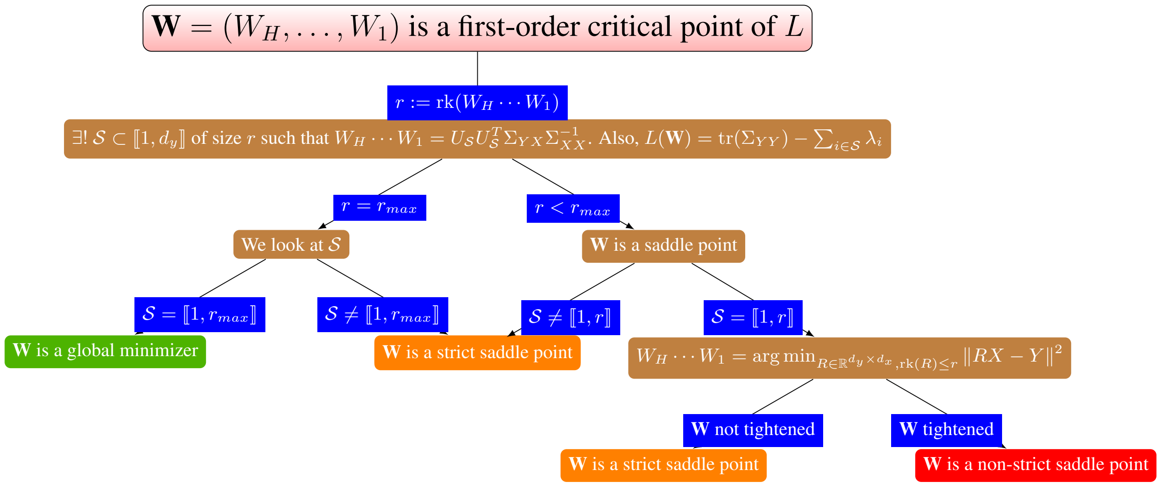

Figure 6: Second-order classification of the critical points of L.

- *if $\mathcal{S} \neq [\![1, r]\!]$, then $\mathbf{W}$ is not a second-order critical point ($\mathbf{W}$ is a strict saddle point).*
- *if $\mathcal{S} = [\![1, r]\!]$: we have $W_H \cdots W_1 = U_\mathcal{S} U_\mathcal{S}^T \Sigma_{YX} \Sigma_{XX}^{-1} \in \arg\min_{R \in \mathbb{R}^{d_y \times d_x}, \mathrm{rk}(R) \leq r} \|RX - Y\|^2$.*
  * *if $\mathbf{W}$ is not tightened, then $\mathbf{W}$ is not a second-order critical point ($\mathbf{W}$ is a strict saddle point).*
  * *if $\mathbf{W}$ is tightened, then $\mathbf{W}$ is a second-order critical point ($\mathbf{W}$ is a non-strict saddle point).*

The proof of Theorem 7 is given in Section 4, with most technical derivations deferred to the appendix. We now make several remarks. Note from the above that every non-strict saddle point corresponds to a global minimizer of the rank-constrained linear regression problem, as already shown by (Bah et al., 2022, Proposition 35).

The next proposition shows the existence of both tightened and non-tightened critical points for $H \geq 3$ (there are no tightened critical points when $H = 2$ and $r < r_{max}$). Combining this result with Proposition 2 indicates that all conclusions of Theorem 7 can be observed. In particular, as already established in Kawaguchi (2016), L is not a Morse function when $H \geq 3$.

Proposition 8 *Suppose Assumption 1 in Section 2 holds true. For $H \geq 3$, for every $\mathcal{S} = [\![1, r]\!]$ with $0 \leq r < r_{max}$, there exist both a tightened critical point and a non-tightened critical point associated with $\mathcal{S}$.*

The proof is postponed to Appendix B.8. It is constructive: we exhibit in the proof large sets of tightened and non-tightened critical points.

We can draw additional consequences from Theorem 7 and Propositions 2 and 8:

- For $H = 2$, for any $r < r_{max}$, there exist strict saddle points satisfying $W_H \cdots W_1 \in \arg\min_{R \in \mathbb{R}^{d_y \times d_x}, \mathrm{rk}(R) \leq r} \|RX - Y\|^2$.

- For $H \geq 3$, for any $r < r_{max}$, there exist both strict and non-strict saddle points satisfying $W_H \cdots W_1 \in \arg\min_{R \in \mathbb{R}^{d_y \times d_x}, \mathrm{rk}(R) \leq r} \|RX - Y\|^2$.

- In the special case $r = 0$, we have $\mathcal{S} = \emptyset$ and $\emptyset = [\![1, r]\!]$ by convention (see Section 2), so that $\mathcal{S} = [\![1, r]\!]$. In this case, Theorem 7 and Proposition 8 together imply that their exist both strict and non-strict saddle points $\mathbf{W}$ such that $W_H \cdots W_1 = 0$ when $H \geq 3$.

Finally, recall from a previous remark that, when $H \geq 3$, all first-order critical points that are 0-balanced are tightened. We know from earlier works (e.g., Arora et al. (2019a, 2018)) that the quantities $W_{j+1}^T W_{j+1} - W_j W_j^T$ are invariant under Gradient Flow. In particular, when we initialize the weight matrices such that these quantities are equal to zero (the so-called 0-balanced initialization), we have $W_{j+1}^T W_{j+1} - W_j W_j^T = 0$ for all $j \in [\![1, H-1]\!]$ along the whole trajectory of Gradient Flow. In that case, all eventually visited saddle points associated to some $\mathcal{S} = [\![1, r]\!]$ are 0-balanced, hence tightened (by the remark after Definition 6) and therefore non-strict (by Theorem 7).

3.3 Parameterization of First-order Critical Points and Global Minimizers

We now turn back to first-order critical points, and state all new related results. In our analysis, these results precede the proof of Theorem 7. The presentation has been reversed in Section 3 to highlight the main contribution of the article.
The next proposition provides an explicit parameterization of first-order critical points. Note that this is only a necessary condition.

Proposition 9 *Suppose Assumption 1 in Section 2 holds true. Let $\mathbf{W} = (W_H, \ldots, W_1)$ be a first-order critical point of L associated with $\mathcal{S}$ (cf Proposition 1), and let $Q = [\![1, d_y]\!] \backslash \mathcal{S}$. Then, there exist invertible matrices $D_{H-1} \in \mathbb{R}^{d_{H-1} \times d_{H-1}}, \ldots, D_1 \in \mathbb{R}^{d_1 \times d_1}$ and matrices $Z_H \in \mathbb{R}^{(d_y - r) \times (d_{H-1} - r)}$, $Z_1 \in \mathbb{R}^{(d_1 - r) \times d_x}$ and $Z_h \in \mathbb{R}^{(d_h - r) \times (d_{h-1} - r)}$ for $h \in [\![2, H-1]\!]$ such that if we denote $\widetilde{W}_H = W_H D_{H-1}$, $\widetilde{W}_1 = D_1^{-1} W_1$ and $\widetilde{W}_h = D_h^{-1} W_h D_{h-1}$, for all $h \in [\![2, H-1]\!]$, then we have*

$$\widetilde{W}_H = [U_{\mathcal{S}}, U_Q Z_H] \tag{2}$$

$$\widetilde{W}_1 = \begin{bmatrix} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} \\ Z_1 \end{bmatrix} \tag{3}$$

$$\widetilde{W}_h = \begin{bmatrix} I_r & 0 \\ 0 & Z_h \end{bmatrix} \quad \forall h \in [\![2, H-1]\!] \tag{4}$$

$$\widetilde{W}_H \cdots \widetilde{W}_2 = [U_{\mathcal{S}}, 0] \, . \tag{5}$$

The proposition is proved in Appendix C.1, and will be key to prove the last statement of Theorem 7. Next, we give a sufficient condition for any $\mathbf{W}$ satisfying (2), (3) and (4), to be a first-order critical point of L.

Proposition 10 *Suppose Assumption 1 in Section 2 holds true. Let $\mathcal{S} \subset [\![1, d_y]\!]$ of size $r \in [\![0, r_{max}]\!]$ and $Q = [\![1, d_y]\!] \setminus \mathcal{S}$. Let $D_{H-1} \in \mathbb{R}^{d_{H-1} \times d_{H-1}}, \ldots, D_1 \in \mathbb{R}^{d_1 \times d_1}$ be invertible matrices and let $Z_H \in \mathbb{R}^{(d_y - r) \times (d_{H-1} - r)}$, $Z_1 \in \mathbb{R}^{(d_1 - r) \times d_x}$ and $Z_h \in \mathbb{R}^{(d_h - r) \times (d_{h-1} - r)}$ for $h \in [\![2, H-1]\!]$. Let the* *parameter of the network* $\mathbf{W} = (W_H, \ldots, W_1)$ *be defined as follows:*

$$W_H = [U_{\mathcal{S}}, U_Q Z_H] D_{H-1}^{-1}$$
$$W_1 = D_1 \begin{bmatrix} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} \\ Z_1 \end{bmatrix}$$
$$W_h = D_h \begin{bmatrix} I_r & 0 \\ 0 & Z_h \end{bmatrix} D_{h-1}^{-1} \quad \forall h \in [\![2, H-1]\!] .$$

If $r = r_{max}$ *or if there exist* $h_1 \neq h_2$ *such that* $Z_{h_1} = 0$ *and* $Z_{h_2} = 0$*, then,* $\mathbf{W}$ *is a first-order critical point of* L *associated with* $\mathcal{S}$*.*

The proof of Proposition 10 is in Appendix B.5.
Note that, combining Propositions 9 and 10, we obtain an explicit parameterization of all critical points $\mathbf{W}$ with a global map $W_H \cdots W_1$ of maximum rank r_{max}. In particular, it yields the next proposition, which provides an explicit parameterization of all the global minimizers of L.

Proposition 11 (Parameterization of all global minimizers) *Suppose Assumption 1 in Section 2 holds true. Set* $\mathcal{S}_{max} = [\![1, r_{max}]\!]$ *and* $Q_{max} = [\![1, d_y]\!] \setminus \mathcal{S}_{max} = [\![r_{max} + 1, d_y]\!]$*.*
Then, $\mathbf{W} = (W_H, \ldots, W_1)$ *is a global minimizer of* L *if and only if there exist invertible matrices* $D_{H-1} \in \mathbb{R}^{d_{H-1} \times d_{H-1}}, \ldots, D_1 \in \mathbb{R}^{d_1 \times d_1}$*, and matrices* $Z_H \in \mathbb{R}^{(d_y - r_{max}) \times (d_{H-1} - r_{max})}$*,* $Z_h \in \mathbb{R}^{(d_h - r_{max}) \times (d_{h-1} - r_{max})}$ *for* $h \in [\![2, H-1]\!]$*, and* $Z_1 \in \mathbb{R}^{(d_1 - r_{max}) \times d_x}$ *such that:*

$$W_H = [U_{\mathcal{S}_{max}}, U_{Q_{max}} Z_H] D_{H-1}^{-1}$$
$$W_1 = D_1 \begin{bmatrix} U_{\mathcal{S}_{max}}^T \Sigma_{YX} \Sigma_{XX}^{-1} \\ Z_1 \end{bmatrix}$$
$$W_h = D_h \begin{bmatrix} I_{r_{max}} & 0 \\ 0 & Z_h \end{bmatrix} D_{h-1}^{-1} \qquad \forall h \in [\![2, H-1]\!] .$$

The proof is in Appendix C.2. See in particular a remark in the same appendix on how to interpret the above formulas precisely (some blocks Z_h have 0 lines or columns).

3.4 Comparison with the State-of-the-art

Next we further detail our contributions in light of earlier works.

Parameterization of global minimizers. To the best of our knowledge, Proposition 11 is the first explicit parameterization of the set of all global minimizers for deep linear networks and the square loss. For $H \geq 2$, it had been previously noted by Yun et al. (2018) that a critical point $\mathbf{W}$ is a global minimizer if and only if $\mathrm{rk}(W_H \cdots W_1) = r_{max}$ and $\mathrm{col}(W_H \cdots W_{d_{p+1}}) = \mathrm{col}(U_{\mathcal{S}_{max}})$, where $\mathcal{S}_{max} = [\![1, r_{max}]\!]$ and where p is any layer with the smallest width r_{max}. This is an implicit characterization.

Another previous work that characterized global minimizers is Zhou and Liang (2018), but their characterization is not explicit: the weight matrices are defined recursively and should satisfy some equations, while in Proposition 11 the weight matrices are given explicitly. The same remark holds for their characterization of first-order critical points.

Saddle points. Among saddle points, we give a characterization of those that are strict and those that are not.
Previously, for $H \geq 3$, it had been noted by Kawaguchi (2016) that $(0, \ldots, 0)$ is a non-strict saddle point. This result also follows from Theorem 1 since any critical point is tightened whenever at least 3 weight matrices are of rank $r = \mathrm{rk}(W_H \cdots W_1)$ (which is the case for $(0, \ldots, 0)$ with $r = 0$).

Also, Theorem 7 generalizes two results from Kawaguchi (2016) and Chitour et al. (2023) about sufficient conditions for strict saddle points. Indeed, it is proved in Kawaguchi (2016) that, if **W** is a saddle point such that $\mathrm{rk}(W_{H-1} \cdots W_2) = r_{max}$, then **W** is a strict saddle point. Chitour et al. (2023) proved under further assumptions on the data and the architecture that a sufficient condition for a saddle point to be strict is that $\mathrm{rk}(W_{H-1} \cdots W_2) > r = \mathrm{rk}(W_H \cdots W_1)$. Note that both results are special cases of Theorem 7, with the pivot $(H, 1)$. More precisely, assume that **W** is a saddle point such that either $\mathrm{rk}(W_{H-1} \cdots W_2) = r_{max} = r = \mathrm{rk}(W_H \cdots W_1)$ or $\mathrm{rk}(W_{H-1} \cdots W_2) > r = \mathrm{rk}(W_H \cdots W_1)$ (which includes both conditions above). Then, if $\mathcal{S} \neq [\![1, r]\!]$ (whether $r = r_{max}$ or not), by Theorem 7, **W** is a strict saddle point without any condition on **W**. But if $\mathcal{S} = [\![1, r]\!]$ with $r < r_{max}$, our assumption above implies that the pivot $(H, 1)$, and therefore **W**, is not tightened (recall that $\mathrm{rk}(\Sigma_{XY}) = d_y \geq r_{max} > r$). In any case, **W** is a strict saddle point.

Finally, Theorem 7 generalizes another result of Kawaguchi (2016) stating that all saddle points are strict for one-hidden layer linear networks. Indeed, let $H = 2$ and assume that we have a saddle point associated with $\mathcal{S} = [\![1, r]\!]$ for $r < r_{max}$ (the only case where we can expect to see non-strict saddle points, by Theorem 7). Since $H = 2$, there is only one pivot which is $(2, 1)$; this pivot is not tightened because the complementary blocks are I_{d_1} and Σ_{XY} and both are of rank larger than or equal to r_{max}. Therefore, by Theorem 7, when $H = 2$ (and under Assumption 1), all saddle points are strict.

Convergence to global minimizer: an example where gradient descent meets no non-strict saddle points. Some recent works on deep linear networks proved under assumptions on the data, the initialization, or the minimum width of the network, that gradient descent or variants converge to a global minimum in polynomial time (e.g., Arora et al., 2019a; Bartlett et al., 2018; Eftekhari, 2020; Du and Hu, 2019). Since for general non-convex functions, gradient descent may get stuck at a non-strict saddle point, and since non-strict saddle points exist for any linear neural network of depth $H \geq 3$, it seemed impossible to deduce convergence to a global minimum using landscape results only. Instead, papers such as Du and Hu (2019) chose to "directly analyze the trajectory generated by [...] gradient descent".

It turns out that our characterization of strict saddle points can help re-interpret such global convergence results. Consider for instance the work of Du and Hu (2019), who proved that with high probability gradient descent with Xavier initialization converges to a global minimum for any deep linear network which is wide enough. They analyze a network where all hidden layers have a width d_{hidden} at least proportional to the number H of layers and to other quantities depending on the data X, Y, the output dimension d_y, and the desired probability level. In their analysis, (Du and Hu, 2019, Section 7) prove that with high probability, a condition $\mathcal{B}(t)$ holds at every iteration t. Importantly, this condition implies that the point **W** output by gradient descent at iteration t cannot be a non-strict saddle point. Indeed, using our notation, the condition $\mathcal{B}(t)$ yields the lower-bound[6]

6. $\sigma_{\min}(W_H \cdots W_2)$ denotes the minimum singular value of $W_H \cdots W_2 \in \mathbb{R}^{d_y \times d_{\text{hidden}}}$, among $\min\{d_y, d_{\text{hidden}}\} = d_y$ singular values in total (Du and Hu 2019 assume that $d_{\text{hidden}} \geq d_y$).

$\sigma_{\min}(W_H \cdots W_2) \geq \frac{3}{4} d_{\text{hidden}}^{(H-1)/2} > 0$, which in particular entails that the matrix product $W_H \cdots W_2$ is of full rank $\min\{d_{\text{hidden}}, d_y\} \geq r_{max}$. Let us check that if **W** is a saddle point, then it is necessarily strict. By Theorem 7, either $r = \text{rk}(W_H \cdots W_1)$ is equal to r_{max}, in which case the saddle point **W** is indeed strict, or $r < r_{max}$, in which case the pivot $(H, 1)$ is not tightened (since the two blocks Σ_{XY} and $W_{H-1} \cdots W_2$ are of rank at least r_{max}), so that the saddle point **W** is strict, as previously claimed.

As a consequence, our characterization of strict saddle points in Theorem 7 helps re-interpret the analysis of (Du and Hu, 2019, Section 7): under Assumption 1, and for wide enough deep linear networks, gradient descent with Xavier initialization meets no non-strict saddle points on its trajectory.

Implicit regularization. Implicit regularization, in the context of linear networks, refers to statements showing that the iterates trajectory passes in the vicinity of critical points **W** such that $W_H \cdots W_1 = \arg\min_{R \in \mathbb{R}^{d_y \times d_x}, \text{rk}(R) \leq r} \|RX - Y\|^2$, for increasing $r \in [\![0, r_{max}]\!]$. In such settings, the gradient dynamics sequentially finds the best linear regression predictor in

$$\mathcal{D}_r = \{R \in \mathbb{R}^{d_y \times d_x} \ , \ \text{rk}(R) \leq r\},$$

for increasing r. The subset $\mathcal{D}_r \subset \mathbb{R}^{d_y \times d_x}$ is independent of X, Y and the network architecture, and plays the role of a regularization constraint in the function space.

In the parameter space however, as indicated in Theorem 7 and Proposition 8, there exists both non-strict and strict saddle points. As illustrated in Section 1.2, Figures 3 and 4, it takes more time to a first order algorithm to escape non-strict saddle points than strict ones. When $H \geq 3$, there exist two phenomenon: a 'light' implicit regularization, in the vicinity of strict saddle points, and a 'strong' implicit regularization in the vicinity of non-strict saddle points.

In Bah et al. (2022), the authors proved that gradient flow converges almost surely to a global minimizer or non-strict saddle points of L. The limit point corresponds to a global minimizer of the rank-constrained linear regression problem. In Theorem 7 and Proposition 8, we prove the existence and characterize such points, and in addition to non-strict saddle points we prove that some **W** leading to the solution of the rank-constrained linear regression problem are strict saddle points. Doing so, we characterize and drastically reduce the strong implicit regularization set.

In Gidel et al. (2019), the authors proved that for $H = 2$, for a vanishing initialization and a sufficiently small learning-rate, the gradient algorithm sequentially learns solutions of the rank-constrained linear regression problem with a gradually increasing rank. More precisely, the algorithm avoids all critical points associated with $\mathcal{S} \neq [\![1, r]\!]$, but comes close to a critical point associated with $\mathcal{S} = [\![1, r]\!]$, spends some time around it and decreases again. We know that for $H = 2$ all saddle points are strict and that the phenomenon described by the authors corresponds to a 'light implicit regularization'.

In Gissin et al. (2019), the authors proved for a toy linear network, that, for $H = 2$, the algorithms need an exponentially vanishing initialization for this incremental learning to occur, while for $H \geq 3$, a polynomially vanishing initialization is enough. This indicates that this incremental learning arises more frequently in deep networks. The difference might be explained by the 'strong' implicit regularization due to the existence of non-strict saddle points when $H \geq 3$.

Authors have put to evidence the rank related implicit regularization depicted in Theorem 7 and Proposition 8 for similar problems. In Arora et al. (2019b), the authors exhibit that for small initializations and learning-rate, for matrix recovery, deep matrix factorization favors solutions of low-rank. In the same context, the authors of Razin and Cohen (2020) state that, implicit regularization in deep matrix completion should be seen as a minimization of rank rather than norms.

3.5 Perspectives

Implicit regularization. From Theorem 7, we know that the critical points such that $W_H \cdots W_1 = \arg\min_{R \in \mathbb{R}^{d_y \times d_x}, \text{rk}(R) \leq r} \|RX - Y\|^2$ can be either strict saddle points or non-strict saddle points. From Proposition 8 we know that both cases exist. We know from the experiment described in Figures 3 and 4, that first order algorithms need more time to escape from the vicinity of non-strict saddle points than strict saddle points. There are two phenomena: 'light' and 'strong' implicit regularization. To the best of our knowledge, whether the saddle points approached by the iterates trajectory are strict or non-strict and the impact of this property on the implicit regularization phenomenon have not been studied.

Though this study goes beyond the scope of this paper, let us sketch the main trends that we can anticipate from our results. On one side, as explained above, we anticipate the number of iterations spent by a first-order algorithm in the vicinity of a non-strict saddle point to be larger than in the vicinity of a strict saddle point. Said differently, the 'size' of the flat region surrounding non-strict saddle points is larger than the one surrounding strict saddle points. On the other side, looking at the rank constraint in Definition 5 (which corresponds to the very last item of Theorem 7), we anticipate that there are much fewer non-strict saddle points than strict saddle points. 'Strong' implicit regularization therefore occurs at fewer locations. The influence of these two factors on the trajectory of the iterates depends on the initialization and the chosen algorithm.

Extent of 'flat regions'. Beyond the behavior of the objective function captured by the derivatives, it would be interesting to study the extent of the 'flat regions'. The goal would typically be to provide estimates of the time spent by a (stochastic) first order algorithm to escape the flat region. We observed in Figures 3 and 4, that the flat regions associated to non-strict saddle points are larger but it would be interesting to extend this empirical study and to study formal estimates of the 'size' of the flat regions.

Basins of attraction. Second order critical points can be limit points of gradient descent algorithms. Even worse, the basin of attraction of such points can be of positive Lebesgue measure. It would be interesting to exploit the tightness condition and the manifold of non-strict saddle points to prove that, as conjectured in Chitour et al. (2023) and Bah et al. (2022), the gradient descent algorithm almost surely converges to a global minimizer.

Generalizing the tightness condition. The tightness condition in the definitions 5 and 6 is for instance satisfied as soon as three factors are of rank r. It is adapted to linear networks. It would be interesting to generalize it to other problems such as matrix factorization, structured linear networks or tensor problems, sharing the same 'compositional structure'.

4. Proof of Theorem 7

The proof of Theorem 7 proceeds in several steps. In the end (see page 22), it will directly follow from Propositions 13, 14, 15 below and from Lemma 21 in Appendix A. In this section, we outline the overall proof structure and state the main intermediate results. We also provide proof sketches for these intermediate results, but defer many technical details to the appendix.

In our proofs, we will not compute the Hessian $\nabla^2 L(\mathbf{W})$ explicitly since this might be quite tedious. To show that a point $\mathbf{W}$ is (or is not) a second-order critical point of L, we will instead Taylor-expand $L(\mathbf{W} + t\mathbf{W}')$ along any direction $\mathbf{W}'$ and use the following lemma. Its proof follows directly from Taylor's theorem.

Lemma 12 (Characterization of first-order and second-order critical points) *Let* $\mathbf{W} = (W_H, \ldots, W_1)$. *Assume that, for all* $\mathbf{W}' = (W'_H, \ldots, W'_1)$, *the loss* $L(\mathbf{W} + t\mathbf{W}')$ *admits the following asymptotic expansion when* $t \to 0$*:*

$$L(\mathbf{W} + t\mathbf{W}') = L(\mathbf{W}) + c_1(\mathbf{W}, \mathbf{W}')t + c_2(\mathbf{W}, \mathbf{W}')t^2 + o(t^2). \tag{6}$$

Then:

- $\mathbf{W}$ *is a first-order critical point of* L *iff* $c_1(\mathbf{W}, \mathbf{W}') = 0$ *for all* $\mathbf{W}'$.
- $\mathbf{W}$ *is a second-order critical point of* L *iff* $c_1(\mathbf{W}, \mathbf{W}') = 0$ *and* $c_2(\mathbf{W}, \mathbf{W}') \geq 0$ *for all* $\mathbf{W}'$. *Therefore if for a first-order critical point* $\mathbf{W}$*, we can exhibit a direction* $\mathbf{W}'$ *such that* $c_2(\mathbf{W}, \mathbf{W}') < 0$*, then* $\mathbf{W}$ *is not a second-order critical point.*

We divide the proof of Theorem 7 into three parts. Recall that from Kawaguchi (2016), we know that all first-order critical points are either global minimizers or saddle points (that is, there is no local extrema apart from global minimizers). We refine this classification.

4.1 Global Minimizers and 'Simple' Strict Saddle Points

In this section, we start by identifying simple sufficient conditions on the support $\mathcal{S}$ associated to a first-order critical point $\mathbf{W}$ which guarantee that $\mathbf{W}$ is either a global minimizer or a strict saddle point. More subtle strict saddle points and non-strict saddle points will be addressed in Sections 4.2 and 4.3.

Proposition 13 *Suppose Assumption 1 in Section 2 holds true. Let* $\mathbf{W} = (W_H, \ldots, W_1)$ *be a first-order critical point of* L *associated with* $\mathcal{S}$ *and set* $r = \mathrm{rk}(W_H \cdots W_1) \leq r_{max}$.

- *When* $r = r_{max}$*:*
  - *if* $\mathcal{S} = [\![1, r_{max}]\!]$*, then* $\mathbf{W}$ *is a global minimizer.*
  - *if* $\mathcal{S} \neq [\![1, r_{max}]\!]$*, then* $\mathbf{W}$ *is not a second-order critical point (*$\mathbf{W}$ *is a strict saddle point).*
- *When* $r < r_{max}$*:* $\mathbf{W}$ *is a saddle point.*
  - *if* $\mathcal{S} \neq [\![1, r]\!]$*, then* $\mathbf{W}$ *is not a second-order critical point (*$\mathbf{W}$ *is a strict saddle point).*

The proof is postponed to Appendix D. To prove that $\mathbf{W}$ associated with $\mathcal{S} \neq [\![1, r]\!]$, $r \leq r_{max}$ is not a second-order critical point, we explicitly exhibit a direction $\mathbf{W}'$ such that the second-order coefficient $c_2(\mathbf{W}, \mathbf{W}')$ in the Taylor expansion of $L(\mathbf{W} + t\mathbf{W}')$ around $t = 0$, in (6), is negative. Using Lemma 12, we conclude that $\mathbf{W}$ is not a second-order critical point.
Recall from Proposition 1 that the loss at any first-order critical point is given by $tr(\Sigma_{YY}) - \sum_{i \in \mathcal{S}} \lambda_i$. The spirit of the proof is that critical points associated with $\mathcal{S} \neq [\![1, r]\!]$ capture a smaller singular value λ_j instead of a larger one λ_i with $i < j$. Thus, to see that the loss can be further decreased at order 2 (and is therefore not a second-order critical point by Lemma 12), a natural proof strategy is to perturb the singular vector corresponding to λ_j along the direction of the singular vector corresponding to λ_i. This part of the proof is an adaption of the proof of Baldi and Hornik (1989).

4.2 Strict Saddle Points Associated with $\mathcal{S} = [\![1, r]\!],\ r < r_{max}$

We now address situations that to our knowledge, have never been addressed, in the literature. We prove the following.

Proposition 14 *Suppose Assumption 1 in Section 2 holds true. Let* $\mathbf{W} = (W_H, \dots, W_1)$ *be a first-order critical point of L associated with* $\mathcal{S} = [\![1, r]\!]$*, with* $0 \le r < r_{max}$*.*
If $\mathbf{W}$ *is not tightened, then* $\mathbf{W}$ *is not a second-order critical point (*$\mathbf{W}$ *is a strict saddle point).*

We sketch the main arguments below. We will again construct a direction $\mathbf{W}'$ such that the second-order coefficient $c_2(\mathbf{W}, \mathbf{W}')$ in the asymptotic expansion of $L(\mathbf{W} + t\mathbf{W}')$ around $t = 0$, in (6), is negative.
More precisely, for a first-order critical point $\mathbf{W}$, for any $\beta \in \mathbb{R}$, we will consider a well-chosen $\mathbf{W}'_\beta$ such that $c_2(\mathbf{W}, \mathbf{W}'_\beta) = a\beta^2 + c\beta$ for some constants a, c (possibly depending on $\mathbf{W}$) such that $a \ge 0$ and $c \ne 0$. Taking

$$\beta = \begin{cases} -c & \text{if} \quad a = 0 \\ -\frac{c}{2a} & \text{if} \quad a > 0 \end{cases} \tag{7}$$

we obtain

$$c_2(\mathbf{W}, \mathbf{W}'_\beta) = \begin{cases} -c^2 & \text{if} \quad a = 0 \\ -\frac{c^2}{4a} & \text{if} \quad a > 0 \end{cases}$$

and therefore

$$c_2(\mathbf{W}, \mathbf{W}'_\beta) < 0.$$

Using Lemma 12, we can conclude that $\mathbf{W}$ is not a second-order critical point.
We now provide intuitions on how to choose $\mathbf{W}'$. Since $\mathbf{W}$ is not tightened, there exists a pivot (i, j), with $i > j$, which is not tightened. Depending on the values of i and j we will construct $\mathbf{W}'$ differently. However, the strategy for constructing $\mathbf{W}'$ is the same in all cases.
Recall again that from Proposition 1, at any first-order critical point $\mathbf{W}$, the value of the loss is given by $tr(\Sigma_{YY}) - \sum_{i\in\mathcal{S}} \lambda_i$. Contrary to the previous section, since $\mathcal{S} = [\![1, r]\!]$ there is no immediate way to decrease the loss (at order 2) without increasing the rank of the product of the weight matrices. Indeed, we have $W_H \cdots W_1 = U_\mathcal{S} U_\mathcal{S}^T \Sigma_{YX} \Sigma_{XX}^{-1} \in \arg\min_{\mathrm{rk}(R)\le r} \|RX - Y\|^2$.
Therefore, to be able to decrease the value of the loss, we need to perturb $\mathbf{W}$ in a way that the product of the perturbed parameter weight matrices becomes of rank strictly larger than r. Also, to prove that $\mathbf{W}$ is not a second-order critical point, we need to decrease the loss at order 2. This is possible when $\mathbf{W}$ is not tightened. For the non-tightened pivot (i, j), we choose a perturbation $\mathbf{W}'$ with all $W'_h = 0$ except for W'_i and W'_j. Furthermore, our construction of W'_i and W'_j depends on whether i and/or j are on the boundary $\{1, H\}$. This is due to the fact that H and 1 play a special role in the product of the perturbed weights $(W_H + tW'_H) \cdots (W_1 + tW'_1)$. This is why we distinguish the four cases below:

- 1st case: $i \in [\![2, H-1]\!]$ and $j = 1$. This case is treated in Appendix E.1.
- 2nd case: $i = H$ and $j = 1$. This case is treated in Appendix E.2.
- 3rd case: $i = H$ and $j \in [\![2, H-1]\!]$. This case is treated in Appendix E.3.
- 4th case: $i, j \in [\![2, H-1]\!]$ with $i > j$. This case is treated in Appendix E.4.

4.3 Non-strict Saddle Points

We now provide a sketch of the proof for the converse of Proposition 14, as stated in Proposition 15 below. All the proofs related to this section are deferred to Appendix F.

Proposition 15 *Suppose Assumption 1 in Section 2 holds true. Let* **W** $= (W_H, \ldots, W_1)$ *be a first-order critical point of* L *associated with* $\mathcal{S} = [\![1, r]\!]$, $0 \leq r < r_{max}$.
If **W** *is tightened, then* **W** *is a second-order critical point (***W** *is a non-strict saddle point).*

To prove Proposition 15, we first state a proposition which indicates that multiplications by invertible matrices do not change the nature of the critical point.

Lemma 16 *For all* $h \in [\![1, H-1]\!]$, *let* $D_h \in \mathbb{R}^{d_h \times d_h}$ *be an invertible matrix. We define* $\widetilde{W}_H = W_H D_{H-1}$, $\widetilde{W}_1 = D_1^{-1} W_1$ *and* $\widetilde{W}_h = D_h^{-1} W_h D_{h-1}$, *for all* $h \in [\![2, H-1]\!]$. *Then*

- **W** $= (W_H, \ldots, W_1)$ *is a first-order critical point of* L *if and only if* $\widetilde{\mathbf{W}} = (\widetilde{W}_H, \ldots, \widetilde{W}_1)$ *is a first-order critical point of* L.
- **W** $= (W_H, \ldots, W_1)$ *is a second-order critical point of* L *if and only if* $\widetilde{\mathbf{W}} = (\widetilde{W}_H, \ldots, \widetilde{W}_1)$ *is a second-order critical point of* L.

The lemma is proved in Appendix B.4.

Proposition 15 is then obtained using Proposition 9 (note that when **W** is tightened, $\widetilde{\mathbf{W}}$ is also tightened since the rank of a matrix does not change when multiplied by invertible matrices), by showing that $\widetilde{\mathbf{W}} = (\widetilde{W}_H, \ldots, \widetilde{W}_1)$ as given by Proposition 9 is a second-order critical point of L and using Lemma 16 to conclude that **W** is a second-order critical point. This is easier since $\widetilde{\mathbf{W}}$ has a simpler form.

More precisely, we have the following result, from which Proposition 15 follows (see Appendix F.2 for details).

Proposition 17 *Suppose Assumption 1 in Section 2 holds true. Let* **W** $= (W_H, \ldots, W_1)$ *be a first-order critical point of* L *associated with* $\mathcal{S} = [\![1, r]\!]$ *with* $0 \leq r < r_{max}$ *such that there exist matrices* $Z_H \in \mathbb{R}^{(d_y - r) \times (d_{H-1} - r)}$, $Z_1 \in \mathbb{R}^{(d_1 - r) \times d_x}$ *and* $Z_h \in \mathbb{R}^{(d_h - r) \times (d_{h-1} - r)}$ *for* $h \in [\![2, H-1]\!]$ *with*

$$W_H = [U_{\mathcal{S}}, U_Q Z_H] \tag{8}$$

$$W_1 = \begin{bmatrix} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} \\ Z_1 \end{bmatrix} \tag{9}$$

$$W_h = \begin{bmatrix} I_r & 0 \\ 0 & Z_h \end{bmatrix} \quad \forall h \in [\![2, H-1]\!] \tag{10}$$

$$W_H \cdots W_2 = [U_{\mathcal{S}}, 0], \tag{11}$$

where $Q = [\![1, d_y]\!] \setminus \mathcal{S}$.
If **W** *is tightened, then* **W** *is a second-order critical point of* L.

Proposition 17 is proved in details in Section F.1. We provide a proof sketch below.
We denote, for t in the neighborhood of 0, and $h \in [\![1, H]\!]$, $W_h(t) = W_h + tW'_h$ where $W'_h \in$

$\mathbb{R}^{d_h \times d_{h-1}}$ is arbitrary.
We define $\mathbf{W}(t) := (W_H(t), \dots, W_1(t))$ and $W(t) := W_H(t) \cdots W_1(t)$. As in the previous two sections, we use Lemma 12. However, this time, we show that the second-order coefficient $c_2(\mathbf{W}, \mathbf{W}')$ is non-negative for all directions $\mathbf{W}'$.
To compute the loss $\|W(t)X - Y\|^2$, we expand

$$\begin{aligned} W(t) &= W_H(t) \cdots W_1(t) \\ &= (W_H + tW'_H) \cdots (W_1 + tW'_1) \\ &= W_H \cdots W_1 + t \sum_{i=1}^{H} W_H \cdots W_{i+1} W'_i W_{i-1} \cdots W_1 \\ &\quad + t^2 \sum_{H \geq i > j \geq 1} W_H \cdots W_{i+1} W'_i W_{i-1} \cdots W_{j+1} W'_j W_{j-1} \cdots W_1 + o(t^2) \ . \end{aligned}$$

Therefore,

$$\begin{aligned} L(\mathbf{W}(t)) = \Bigg\| & W_H \cdots W_1 X - Y + t \sum_{i=1}^{H} W_H \cdots W_{i+1} W'_i W_{i-1} \cdots W_1 X \\ & + t^2 \sum_{H \geq i > j \geq 1} W_H \cdots W_{i+1} W'_i W_{i-1} \cdots W_{j+1} W'_j W_{j-1} \cdots W_1 X + o(t^2) \Bigg\|^2 \ . \end{aligned}$$

We can now easily calculate the second-order coefficient $c_2(\mathbf{W}, \mathbf{W}')$ in the Taylor expansion of $L(\mathbf{W}(t))$ around $t = 0$ (in (6)).
Recalling that $c_2(\mathbf{W}, \mathbf{W}')$ is such that $L(\mathbf{W}(t)) = L(\mathbf{W}) + c_2(\mathbf{W}, \mathbf{W}')t^2 + o(t^2)$ (since $\mathbf{W}$ is a first-order critical point), we have

$$\begin{aligned} c_2(\mathbf{W}, \mathbf{W}') = & \left\| \sum_{i=1}^{H} W_H \cdots W_{i+1} W'_i W_{i-1} \cdots W_1 X \right\|^2 \\ & + 2 \left\langle \sum_{H \geq i > j \geq 1} W_H \cdots W_{i+1} W'_i W_{i-1} \cdots W_{j+1} W'_j W_{j-1} \cdots W_1 X \ , \ W_H \cdots W_1 X - Y \right\rangle \ , \end{aligned}$$

where $\langle A, B \rangle = \mathrm{tr}(AB^T)$. In order to simplify the notation and equations, we define, for all $i \in [\![1, H]\!]$,

$$T_i = W_H \cdots W_{i+1} W'_i W_{i-1} \cdots W_1 X \ , \tag{12}$$

and for all $i, j \in [\![1, H]\!]$ with $i > j$:

$$T_{i,j} = \left\langle W_H \cdots W_{i+1} W'_i W_{i-1} \cdots W_{j+1} W'_j W_{j-1} \cdots W_1 X \ , \ W_H \cdots W_1 X - Y \right\rangle \ . \tag{13}$$

Then we set

$$FT = \left\| \sum_{i=1}^{H} T_i \right\|^2 \ , \tag{14}$$

and

$$ST = 2 \sum_{H \geq i > j \geq 1} T_{i,j} \ . \tag{15}$$

The coefficient becomes

$$c_2(\mathbf{W}, \mathbf{W}') = \left\| \sum_{i=1}^{H} T_i \right\|^2 + 2 \sum_{H \geq i > j \geq 1} T_{i,j} = FT + ST \ .$$

Using the fact that **W** is tightened, some weight products become simple (see Lemma 30) and we can simplify T_i and $T_{i,j}$ (see Lemmas 37 and 38 in Appendix F).
This allows us to establish that, for any $\mathbf{W}'$, there exist matrices A_2, A_3, A_4 and a non-negative scalar a_1 such that $FT = a_1 + \|A_2\|^2 + \|A_3\|^2 + \|A_4\|^2$ (see Appendix F.1.2) and $ST = -2 \langle A_3, A_4 \rangle$ (see Appendix F.1.3). Therefore

$$c_2(\mathbf{W}, \mathbf{W}') = FT + ST = a_1 + \|A_2\|^2 + \|A_3 - A_4\|^2 \geq 0 \ ,$$

and using Lemma 12 we conclude that **W** is a second-order critical point.

We are now in a position to prove Theorem 7 as a direct corollary from the above results.
Proof [Proof of Theorem 7] The classification into global minimizers, strict saddle points, and non-strict saddle points follows directly from Propositions 13, 14, and 15 above. As for the fact that

$$W_H \cdots W_1 = U_{\mathcal{S}} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} \in \underset{R \in \mathbb{R}^{d_y \times d_x}, \mathrm{rk}(R) \leq r}{\arg\min} \|RX - Y\|^2$$

when $\mathcal{S} = [\![1, r]\!]$, it follows from Proposition 1 above and from Lemma 21 in Appendix A. ■

5. Conclusion

We studied the optimization landscape of linear neural networks of arbitrary depth with the square loss. We first derived a necessary condition for being a first-order critical point by associating any of them with a set of eigenvectors of a data-dependent matrix. We then provided a complete characterization of the landscape at order 2 by distinguishing between global minimizers, strict saddle points, and non-strict saddle points. As a by-product of this analysis, we exhibited large sets of strict and non-strict saddle points and derived an explicit parameterization of all global minimizers. Our second-order characterization also sheds some light on the implicit regularization that may be induced by first-order algorithms, by proving that non-strict saddle points and some strict saddle points are among the global minimizers of the rank-constrained linear regression problem. It also helps re-interpret a recent convergence result, stating that gradient descent with Xavier initialization converges to a global minimum for any wide enough deep linear network.

Acknowledgments

Our work has benefited from the AI Interdisciplinary Institute ANITI. ANITI is funded by the French "Investing for the Future – PIA3" program under the Grant agreement n°ANR-19-PI3A-0004. The authors gratefully acknowledge the support of the DEEL project.[7] El Mehdi Achour acknowledges as well funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1481 – 442047500. We would like to thank the anonymous reviewer who made the connection between 0-balancedness and tightenedness.

Appendix A. Notation and Useful Properties

In this section, we define some additional notation and terminology that will be used through all subsequent appendices. We also state simple linear algebra facts (Section A.2), together with some properties about the Moore-Penrose inverse (Section A.3). Since most of the proofs rely on linear algebra, we recommend the unfamiliar reader to check classical textbooks.

Additional notation: If a matrix A has already a subscript like W_H for example, we denote by $(W_H)_{.,i}$ the i-th column and by $(W_H)_{.,J}$ the sub-matrix obtained by concatenating the column vectors $(W_H)_{.,i}$, for all $i \in \mathcal{J}$. Also $(W_H)_{i,.}$ denotes the i-th row of W_H and $(W_H)_{\mathcal{I},.}$ the sub-matrix obtained by concatenating the line vectors $(W_H)_{i,.}$, for all $i \in \mathcal{I}$. More generally $(W_H)_{\mathcal{I},\mathcal{J}}$ denotes the matrix W_H restricted to the index set $\mathcal{I} \times \mathcal{J}$. For instance, $(W_H)_{1:r,r+1:d_{H-1}} \in \mathbb{R}^{r\times(d_{H-1}-r)}$ is the matrix formed from W_H by keeping the rows from 1 to r and the columns from $r+1$ to d_{H-1}. The symbol $\delta_{i,j}$ denotes the Kronecker index which equal to 0 if $i \neq j$ and 1 if $i = j$.

Also, we define the partial gradients with respect to each weight matrix as follows.

A.1 Partial Gradients

Definition 18 (gradient and partial gradients of L) *Since the input $\mathbf{W} = (W_H, \ldots, W_1)$ of $L(\mathbf{W})$ is not a vector but a sequence of matrices, we define the gradient $\nabla L(\mathbf{W})$ of L at $\mathbf{W}$ with a similar format :*

$$\nabla L(\mathbf{W}) = (\nabla_{W_H} L(\mathbf{W}), \ldots, \nabla_{W_1} L(\mathbf{W})) \ ,$$

where each partial gradient $\nabla_{W_h} L(\mathbf{W}) \in \mathbb{R}^{d_h\times d_{h-1}}$ is the matrix whose entries are the partial derivatives $\frac{\partial L}{\partial (W_h)_{i,j}}$ for $i = 1, .., d_h$ and $j = 1, .., d_{h-1}$

The next lemma provides explicit formulas for the partial gradients of L. A proof can be found at the end of Yun et al. (2018).

Lemma 19 *Let $h \in [\![2, H-1]\!]$. The partial gradient of L with respect to W_h is:*

$$\nabla_{W_h} L(\mathbf{W}) = 2(W_H \cdots W_{h+1})^T (W_H \cdots W_1 \Sigma_{XX} - \Sigma_{YX})(W_{h-1} \cdots W_1)^T \ .$$

We also have the partial gradient with respect to W_H:

$$\nabla_{W_H} L(\mathbf{W}) = 2(W_H \cdots W_1 \Sigma_{XX} - \Sigma_{YX})(W_{H-1} \cdots W_1)^T \ .$$

Finally, the partial gradient with respect to W_1 is:

$$\nabla_{W_1} L(\mathbf{W}) = 2(W_H \cdots W_2)^T (W_H \cdots W_1 \Sigma_{XX} - \Sigma_{YX}) \ .$$

7. https://www.deel.ai/

A.2 Simple Linear Algebra Facts

Recall that $\Sigma^{1/2} = \Sigma_{YX}\Sigma_{XX}^{-1}X$ and $\Sigma = \Sigma^{1/2}(\Sigma^{1/2})^T = \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$. Recall also from (1) that $\Sigma^{1/2} = U\Delta V^T$ is a Singular Value Decomposition, where $U \in \mathbb{R}^{d_y \times d_y}$ and $V \in \mathbb{R}^{m\times m}$ are orthogonal matrices.

Lemma 20 *Suppose Assumption 1 in Section 2 holds true. Then Σ is invertible.*

Proof Given the definition of $\Sigma^{1/2}$, it is a standard fact of linear algebra that $\mathrm{rk}(\Sigma^{1/2}) = \mathrm{rk}(\Sigma_{YX}\Sigma_{XX}^{-1}X) \le \mathrm{rk}(\Sigma_{YX})$. On the other hand, $\mathrm{rk}(\Sigma^{1/2}) = \mathrm{rk}(\Sigma_{YX}\Sigma_{XX}^{-1}X) \ge \mathrm{rk}(\Sigma_{YX}\Sigma_{XX}^{-1}XX^T) = \mathrm{rk}(\Sigma_{YX})$ since $\Sigma_{XX} = XX^T$. Therefore $\mathrm{rk}(\Sigma^{1/2}) = \mathrm{rk}(\Sigma_{YX}) = d_y$ by Assumption 1. Finally, using another fact of linear algebra we have $\mathrm{rk}(\Sigma) = \mathrm{rk}(\Sigma^{1/2}(\Sigma^{1/2})^T) = \mathrm{rk}(\Sigma^{1/2})$, and therefore $\mathrm{rk}(\Sigma) = d_y$. Hence, Σ is invertible. ■

The next lemma is about global minimizers of the rank-constrained linear regression problem.

Lemma 21 *Suppose Assumption 1 in Section 2 holds true. Let $\mathcal{S} = [\![1, r]\!]$. We have*

$$U_{\mathcal{S}}U_{\mathcal{S}}^T\Sigma_{YX}\Sigma_{XX}^{-1} \in \underset{R\in\mathbb{R}^{d_y\times d_x},\mathrm{rk}(R)\le r}{\arg\min} \|RX - Y\|^2.$$

Proof A proof can be found in Yun et al. (2018). ■

We now present a lemma with elementary properties that we will use frequently and that are related to the orthogonality of U. The proof is straightforward.

Lemma 22 *We have the following properties related to the orthogonality of the matrix U:*

- *We have $I_{d_y} = UU^T = U^TU$.*
- *For any $i, j \in [\![1, d_y]\!]$, we have $U_i^TU_j = \delta_{i,j}$.*
- *For any $I, J \subset [\![1, d_y]\!]$ such that $I \cap J = \emptyset$, we have $U_I^TU_J = 0_{|I|\times|J|}$.*
- *For any $I, J \subset [\![1, d_y]\!]$ such that $I\cap J = \emptyset$ and $I\cup J = [\![1, d_y]\!]$, we have $I_{d_y} = U_IU_I^T + U_JU_J^T$.*
- *For any $J \subset [\![1, d_y]\!]$, we have $U_J^TU_J = I_{|J|}$ and $\mathrm{rk}(U_JU_J^T) = |J|$.*

Note that the same applies also to the other orthogonal matrix $V \in \mathbb{R}^{m\times m}$ appearing in the Singular Value Decomposition of $\Sigma^{1/2}$ (we only replace d_y by m).

Another useful lemma is the following:

Lemma 23 *For any $I, J \subset [\![1, d_y]\!]$ such that $I \cap J = \emptyset$, we have*

$$U_I^T\Sigma U_J = 0_{|I|\times|J|}.$$

In particular, for any $\mathcal{S} \subset [\![1, d_y]\!]$ and $Q = [\![1, d_y]\!] \setminus \mathcal{S}$, we have $U_{\mathcal{S}}^T\Sigma U_Q = 0$.

Proof We have, for any $k \in [\![1, d_y]\!]$, $\Sigma U_k = \lambda_kU_k$. Hence for $j \ne k$ we have $U_j^T\Sigma U_k = \lambda_kU_j^TU_k = 0$ since U is orthogonal. Therefore, if we take two disjoint sets $J = \{j_1, \dots, j_p\}, K = \{k_1, \dots, k_n\} \subset [\![1, d_y]\!]$, the coefficient in the position (l, m) of the matrix $U_J^T\Sigma U_K$ is equal to $U_{j_l}\Sigma U_{k_m}$ which is zero, since $j_l \ne k_m$. Therefore, $U_J^T\Sigma U_K = 0$. In particular, $U_{\mathcal{S}}^T\Sigma U_Q = 0$. ■

A.3 The Moore-Penrose Inverse and its Properties

The Moore-Penrose inverse is the most known and used generalized inverse[8]. It is defined as follows: For $A \in \mathbb{R}^{m\times n}$, the pseudo-inverse of A is defined as the matrix $A^+ \in \mathbb{R}^{n\times m}$ which satisfies the 4 following criteria known as the Moore-Penrose conditions:

1. $AA^+A = A$.
2. $A^+AA^+ = A^+$.
3. $(AA^+)^T = AA^+$.
4. $(A^+A)^T = A^+A$.

A^+ exists for any matrix A and is unique. We also have the following properties:

(i) $A^+ = \left(A^TA\right)^+ A^T$.

(ii) $\mathrm{rk}(A) = \mathrm{rk}(A^+) = \mathrm{rk}(AA^+) = \mathrm{rk}(A^+A)$.

(iii) If the linear system $Ax = b$ has any solutions, they are all given by

$$x = A^+b + (I - A^+A)w$$

for arbitrary vector w. This is equivalent to

$$x = A^+b + u$$

for arbitrary $u \in \mathrm{Ker}(A)$.

(iv) $P_A := AA^+$ is the orthogonal projection onto the range of A, and is therefore symmetric ($P_A^T = P_A$) (follows from 3) and idempotent ($P_A^2 = P_A$) (follows from 1).

(v) $I_n - A^+A$ is the orthogonal projector onto the kernel of A.

Appendix B. Propositions and Lemmas for First-order Critical Points

In this section, we prove all lemmas about first-order critical points. We start by stating some preliminary results.

B.1 Preliminaries

The following lemma gives a necessary condition for **W** to be a first-order critical point. It also provides the global map of the network, defined by $W_H \cdots W_1$. Finally, it states that the projection matrix P_K and Σ commute, where $K = W_H \cdots W_2$. This is key in the rest of the analysis.

Lemma 24 *Suppose Assumption 1 in Section 2 holds true. Let* $\mathbf{W} = (W_H, \ldots, W_1)$ *be a first-order critical point of* L*. We define* $K = W_H \cdots W_2$ *and* $W = W_H W_{H-1} \cdots W_1 = KW_1$*. Then, we have*

$$W_1 = K^+\Sigma_{YX}\Sigma_{XX}^{-1} + M\ ,$$

8. `en.wikipedia.org/wiki/Moore-Penrose_inverse`

where $M \in \mathbb{R}^{d_1 \times d_x}$ is such that $KM = 0$ and K^+ is the Moore-Penrose inverse of K (see Appendix A.3). As a consequence,

$$\begin{cases} W = P_K \Sigma_{YX} \Sigma_{XX}^{-1} \\ \mathrm{rk}(W) = \mathrm{rk}(P_K) = \mathrm{rk}(K) \end{cases}$$

where we recall that $P_K = KK^+ \in \mathbb{R}^{d_y \times d_y}$ is the matrix of the orthogonal projection onto the range of K. Finally,

$$\Sigma P_K = P_K \Sigma \ .$$

Note that $\Sigma_{YX}\Sigma_{XX}^{-1}$ is the global minimizer of the problem with one layer (i.e the classical linear regression problem). Therefore, the global map $W_H \cdots W_1$ of any first-order critical point of L is equal to the global minimizer of the linear regression projected onto the column space of K.

Proof Let $\mathbf{W} = (W_H, \ldots, W_1)$ be a first-order critical point of L. In particular, the partial gradients of L with respect to W_1 and W_H are equal to zero at $\mathbf{W}$. Using Lemma 19, this implies

$$\begin{cases} (W_H \cdots W_2)^T W_H \cdots W_1 \Sigma_{XX} = (W_H \cdots W_2)^T \Sigma_{YX} \\ W_H \cdots W_1 \Sigma_{XX} (W_{H-1} \cdots W_1)^T = \Sigma_{YX} (W_{H-1} \cdots W_1)^T \ . \end{cases}$$

We substitute in these equations $K = W_H W_{H-1} \cdots W_2$ and $W = W_H W_{H-1} \cdots W_1 = KW_1$. Using that Σ_{XX} is invertible, and multiplying the second equation on the right by W_H^T, we obtain that any critical point of L satisfies

$$\begin{cases} K^T K W_1 = K^T \Sigma_{YX} \Sigma_{XX}^{-1} \\ W \Sigma_{XX} W^T = \Sigma_{YX} W^T \ . \end{cases} \tag{16}$$

The first equation implies $W_1 = (K^T K)^+ K^T \Sigma_{YX} \Sigma_{XX}^{-1} + M$, where $M \in \mathbb{R}^{d_1 \times d_x}$ is such that $K^T K M = 0$ (see Property (iii) in the reminder on Moore-Penrose inverse in Appendix A.3).
We have $(K^T K)^+ K^T = K^+$(see Property (i) in Appendix A.3) and a standard fact of linear algebra is that $\mathrm{Ker}(K^T K) = \mathrm{Ker}(K)$.
Therefore, using these properties, we obtain $W_1 = K^+ \Sigma_{YX} \Sigma_{XX}^{-1} + M$, where $KM = 0$. This proves the first statement of the lemma. We then have,

$$W = KW_1 = KK^+ \Sigma_{YX} \Sigma_{XX}^{-1} + KM = P_K \Sigma_{YX} \Sigma_{XX}^{-1} \ . \tag{17}$$

where $P_K = KK^+$ is the orthogonal projection matrix onto the column space of K (see Appendix A.3). Using Assumption 1, we have that $\Sigma_{YX}\Sigma_{XX}^{-1}$ is of full row rank, hence

$$\mathrm{rk}(W) = \mathrm{rk}(P_K \Sigma_{YX} \Sigma_{XX}^{-1}) = \mathrm{rk}(P_K) = \mathrm{rk}(K) \ , \tag{18}$$

where the last equality comes from the property (ii) in Section A.3. Therefore, (16) and (18) prove the second statement of the lemma.
To prove that $\Sigma P_K = P_K \Sigma$, we remark that, using the second equation in (16), $\Sigma_{YX} W^T = W \Sigma_{XX} W^T$ and since $W \Sigma_{XX} W^T$ is symmetric and $(\Sigma_{YX})^T = \Sigma_{XY}$, we have

$$\Sigma_{YX} W^T = W \Sigma_{XY} \ .$$

Substituting the expression of W from (17), and since P_K and Σ_{XX}^{-1} are symmetric, we have

$$\Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY} P_K = P_K \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY} \ .$$

Using the definition of Σ, this can be rewritten as

$$\Sigma P_K = P_K \Sigma ,$$

which concludes the proof. ■

Lemma 25 *Suppose Assumption 1 in Section 2 holds true. Let* $\mathbf{W} = (W_H, \dots, W_1)$ *be a first-order critical point of* L*. We set* $K = W_H \cdots W_2$ *and* $r = \mathrm{rk}(W_H \cdots W_1)$*.*
There exists a unique subset $\mathcal{S} \subset [\![1, d_y]\!]$ *of size* r *such that:*

$$P_K = U \mathcal{I}^{\mathcal{S}} U^T = U_{\mathcal{S}} U_{\mathcal{S}}^T,$$

where $\mathcal{I}^{\mathcal{S}} \in \mathbb{R}^{d_y \times d_y}$ *is the diagonal matrix such that, for all* $i \in [\![1, d_y]\!]$*,* $(\mathcal{I}^{\mathcal{S}})_{i,i} = 1$ *if* $i \in \mathcal{S}$ *and* 0 *otherwise.*

Proof Let $\mathbf{W} = (W_H, \dots, W_1)$ be a first-order critical point of L. Using Lemma 24, we have $\Sigma P_K = P_K \Sigma$. Substituting the diagonalization of Σ from Section 2, this becomes $U \Lambda U^T P_K = P_K U \Lambda U^T$. Since U is orthogonal, multiplying by U^T on the left and by U on the right we obtain $\Lambda U^T P_K U = U^T P_K U \Lambda$. Hence, $U^T P_K U$ commutes with a diagonal matrix whose diagonal elements are all distinct. Therefore, $\Gamma := U^T P_K U$ is diagonal, and $P_K = U \Gamma U^T$ is a diagonalization of P_K. From Lemma 24, we also have $r = \mathrm{rk}(P_K)$. But, we know that $P_K = K K^+ \in \mathbb{R}^{d_y \times d_y}$ is the matrix of an orthogonal projection. Therefore, its eigenvalues are 1 with multiplicity r and 0 with multiplicity $d_y - r$.

Therefore, there exists an index set $\mathcal{S} \subset [\![1, d_y]\!]$ of size r such that $\Gamma = \mathcal{I}^{\mathcal{S}}$ where $\mathcal{I}^{\mathcal{S}} \in \mathbb{R}^{d_y \times d_y}$ is the diagonal matrix such that, for all $i \in [\![1, d_y]\!]$, $(\mathcal{I}^{\mathcal{S}})_{i,i} = 1$ if $i \in \mathcal{S}$ and 0 otherwise.
Therefore,

$$P_K = U \mathcal{I}^{\mathcal{S}} U^T = U \mathcal{I}^{\mathcal{S}} \mathcal{I}^{\mathcal{S}} U^T = U_{\mathcal{S}} U_{\mathcal{S}}^T.$$

If there exist $\mathcal{S}'$ such that $\Gamma = \mathcal{I}^{\mathcal{S}'}$, we get $P_K = U \mathcal{I}^{\mathcal{S}} U^T = U \mathcal{I}^{\mathcal{S}'} U^T$ which implies $\mathcal{I}^{\mathcal{S}} = \mathcal{I}^{\mathcal{S}'}$, hence $\mathcal{S} = \mathcal{S}'$. Therefore, $\mathcal{S}$ is unique. ■

B.2 Proof of Proposition 1

In this proof, we use Lemmas 24 and 25 stated and proved in the previous section.
Recall that $\lambda_1 > \dots > \lambda_{d_y}$ are the eigenvalues of $\Sigma = \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY} \in \mathbb{R}^{d_y \times d_y}$.
Let $\mathbf{W} = (W_H, \dots, W_1)$ be a first-order critical point of L. We set $K = W_H \cdots W_2$, $r = \mathrm{rk}(W_H \cdots W_1)$. Using Lemma 25, there exists a unique subset $\mathcal{S} \subset [\![1, d_y]\!]$ of size r such that:

$$P_K = U_{\mathcal{S}} U_{\mathcal{S}}^T.$$

Therefore, using Lemma 24,

$$W_H \cdots W_1 = P_K \Sigma_{YX} \Sigma_{XX}^{-1} = U_{\mathcal{S}} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1}.$$

This proves the first statement of Proposition 1.
To prove the second statement, notice that we have

$$\begin{aligned}
L(\mathbf{W}) &= \|WX - Y\|^2 \\
&= \|WX\|^2 - 2\langle WX\ ,\ Y\rangle + \|Y\|^2 \\
&= \operatorname{tr}(W\Sigma_{XX}W^T) - 2\operatorname{tr}(W\Sigma_{XY}) + \operatorname{tr}(\Sigma_{YY}) \\
&= \operatorname{tr}(U_{\mathcal{S}}U_{\mathcal{S}}^T\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XX}\Sigma_{XX}^{-1}\Sigma_{XY}U_{\mathcal{S}}U_{\mathcal{S}}^T) - 2\operatorname{tr}(U_{\mathcal{S}}U_{\mathcal{S}}^T\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}) + \operatorname{tr}(\Sigma_{YY}) \\
&= \operatorname{tr}(U_{\mathcal{S}}U_{\mathcal{S}}^T U_{\mathcal{S}}U_{\mathcal{S}}^T\Sigma) - 2\operatorname{tr}(U_{\mathcal{S}}U_{\mathcal{S}}^T\Sigma) + \operatorname{tr}(\Sigma_{YY})
\end{aligned}$$

Since $U_{\mathcal{S}}^T U_{\mathcal{S}} = I_r$ (see Lemma 22), using Lemma 25 and the fact that U diagonalizes Σ, this becomes

$$\begin{aligned}
L(\mathbf{W}) &= \operatorname{tr}(\Sigma_{YY}) - \operatorname{tr}(U_{\mathcal{S}}U_{\mathcal{S}}^T\Sigma) \\
&= \operatorname{tr}(\Sigma_{YY}) - \operatorname{tr}(U\mathcal{I}^{\mathcal{S}}U^T U\Lambda U^T) \\
&= \operatorname{tr}(\Sigma_{YY}) - \operatorname{tr}(\mathcal{I}^{\mathcal{S}}U^T U\Lambda U^T U) \\
&= \operatorname{tr}(\Sigma_{YY}) - \operatorname{tr}(\mathcal{I}^{\mathcal{S}}\Lambda) \\
&= \operatorname{tr}(\Sigma_{YY}) - \sum_{i\in\mathcal{S}}\lambda_i\ .
\end{aligned}$$

This proves the second and last statement of Proposition 1.

B.3 Lemma 26

In this section we state and prove a lemma about first-order critical points which will be useful in various proofs. This lemma gives a simpler form for $K = W_H \cdots W_2$ and W_1.

Lemma 26 *Suppose Assumption 1 in Section 2 holds true. Let $\mathbf{W} = (W_H, \ldots, W_1)$ be a first-order critical point of L associated with $\mathcal{S}$. We set $r = \operatorname{rk}(W_H \cdots W_1)$.*
Then there exists an invertible matrix $D \in \mathbb{R}^{d_1\times d_1}$, a matrix $M \in \mathbb{R}^{d_1\times d_x}$ satisfying $W_H \cdots W_2 M = 0$, such that:

$$K = W_H \cdots W_2 = \begin{bmatrix} U_{\mathcal{S}} & 0_{d_y\times(d_1-r)} \end{bmatrix} D$$

and

$$W_1 = D^{-1}\begin{bmatrix} U_{\mathcal{S}}^T\Sigma_{YX}\Sigma_{XX}^{-1} \\ 0_{(d_1-r)\times d_x} \end{bmatrix} + M\ .$$

Note that the result is still true when $r = 0$, provided that $U_\emptyset \in \mathbb{R}^{d_y\times 0}$.

To prove Lemma 26, we use Lemmas 24 and 25 stated and proved in the preliminaries of Appendix B.1. We will also need the following lemma

Lemma 27 *Let n be a positive integer and $\emptyset \neq \mathcal{S} \subset [\![1, d_y]\!]$ such that $n \geq r := |\mathcal{S}|$. Let $A \in \mathbb{R}^{d_y\times n}$ such that $AA^+ = U_{\mathcal{S}}U_{\mathcal{S}}^T$. Then there exists an invertible matrix $D \in \mathbb{R}^{n\times n}$ such that*

$$A = [U_{\mathcal{S}} \quad 0_{d_y\times(n-r)}]D$$

and

$$A^+ = D^{-1} \begin{bmatrix} U_{\mathcal{S}}^T \\ 0_{(n-r)\times d_y} \end{bmatrix} .$$

Proof [Proof of Lemma 27]

The matrix $I_n - A^+A$ is the orthogonal projection onto Ker(A) (see Appendix A.3), hence

$$\mathrm{rk}(I_n - A^+A) = \text{dim Ker}(A) = n - \mathrm{rk}(A)$$

But we have (see Property (ii) in Appendix A.3) $\mathrm{rk}(A^+A) = \mathrm{rk}(A) = \mathrm{rk}(AA^+)$ and, using Lemma 22, $\mathrm{rk}(A^+A) = \mathrm{rk}(U_{\mathcal{S}}U_{\mathcal{S}}^T) = r$. Therefore, $\mathrm{rk}(A) = r$ and

$$\mathrm{rk}(I_n - A^+A) = n - r.$$

Let $B \in \mathbb{R}^{n\times(n-r)}$ and $C \in \mathbb{R}^{(n-r)\times n}$ be such that $I_n - A^+A = BC$ (such matrices can be obtained by considering the Singular Value Decomposition of $I_n - A^+A$).

Denoting $D = \begin{bmatrix} U_{\mathcal{S}}^T A \\ C \end{bmatrix} \in \mathbb{R}^{n\times n}$, we have

$$[A^+U_{\mathcal{S}}\ ,\ B]D = [A^+U_{\mathcal{S}}\ ,\ B] \begin{bmatrix} U_{\mathcal{S}}^T A \\ C \end{bmatrix} = A^+U_{\mathcal{S}}U_{\mathcal{S}}^T A + BC = A^+AA^+A + I_n - A^+A\ .$$

Using Criteria 1 in Appendix A.3 we obtain

$$[A^+U_{\mathcal{S}}\ ,\ B]D = A^+A + I_n - A^+A = I_n\ .$$

Therefore, D is invertible and $D^{-1} = [A^+U_{\mathcal{S}}\ ,\ B]$. We have

$$[U_{\mathcal{S}}\ ,\ 0_{d_y\times(n-r)}]D = [U_{\mathcal{S}}\ ,\ 0_{d_y\times(n-r)}] \begin{bmatrix} U_{\mathcal{S}}^T A \\ C \end{bmatrix} = U_{\mathcal{S}}U_{\mathcal{S}}^T A = AA^+A = A\ ,$$

where the last equality follows from Criteria 1 in Appendix A.3. This proves the first equality of Lemma 27 Finally,

$$D^{-1} \begin{bmatrix} U_{\mathcal{S}}^T \\ 0_{(n-r)\times d_y} \end{bmatrix} = [A^+U_{\mathcal{S}}\ ,\ B] \begin{bmatrix} U_{\mathcal{S}}^T \\ 0_{(n-r)\times d_y} \end{bmatrix} = A^+U_{\mathcal{S}}U_{\mathcal{S}}^T = A^+AA^+ = A^+\ ,$$

where the last equality follows again from Criteria 2 in Appendix A.3. This concludes the proof of Lemma 27. ■

Now we prove Lemma 26.

Proof [Proof of Lemma 26]

Let $\mathbf{W} = (W_H, \ldots, W_1)$ be a first-order critical point of L associated with $\mathcal{S}$ and $r = \mathrm{rk}(W_H \cdots W_1)$.

Using Lemma 24, we have $r = \mathrm{rk}(W_H \cdots W_2)$.

If $r = 0$, the conclusion of Lemma 26 is trivial because of the convention $U_\emptyset \in \mathbb{R}^{d_y\times 0}$.

When $r \geq 1$, using Lemma 24 and Proposition 1, we have $W_H \cdots W_1 = P_K\Sigma_{YX}\Sigma_{XX}^{-1} =$

$U_{\mathcal{S}} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1}$. Since Σ_{YX} is of full row rank this implies $P_K = KK^+ = U_{\mathcal{S}} U_{\mathcal{S}}^T$. Therefore, we can apply Lemma 27 with $n = d_1$ and $A = K$ to conclude that there exists an invertible matrix $D \in \mathbb{R}^{d_1 \times d_1}$ such that

$$K = [U_{\mathcal{S}} \; , \; 0_{d_y \times (d_1 - r)}] D$$

which is the form of K in Lemma 26. Moreover, Lemma 27 also guarantees that

$$K^+ = D^{-1} \begin{bmatrix} U_{\mathcal{S}}^T \\ 0_{(d_1-r) \times d_y} \end{bmatrix} .$$

Using Lemma 24, we have $W_1 = K^+ \Sigma_{YX} \Sigma_{XX}^{-1} + M$ with $KM = 0$. Therefore, $W_1 = D^{-1} \begin{bmatrix} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} \\ 0_{(d_1-r) \times d_x} \end{bmatrix} + M$, with $KM = 0$. This concludes the proof of Lemma 26. ■

B.4 Proof of Lemma 16

For any $h \in [\![1, H-1]\!]$ let $D_h \in \mathbb{R}^{d_h \times d_h}$ be an invertible matrix. We define $\widetilde{\mathbf{W}} = (\widetilde{W}_H, \ldots, \widetilde{W}_1)$ by $\widetilde{W}_H = W_H D_{H-1}$, $\widetilde{W}_1 = D_1^{-1} W_1$ and $\widetilde{W}_h = D_h^{-1} W_h D_{h-1}$ for all $h \in [\![2, H-1]\!]$.
Assume that $\mathbf{W} = (W_H, \ldots, W_1)$ is a first-order critical point. Then using Lemma 19 this is equivalent to

$$\begin{cases} \nabla_{W_h} L(\mathbf{W}) = 2(W_H \cdots W_{h+1})^T (W_H \cdots W_1 \Sigma_{XX} - \Sigma_{YX})(W_{h-1} \cdots W_1)^T = 0 \quad \forall h \in [\![2, H-1]\!] \\ \nabla_{W_H} L(\mathbf{W}) = 2(W_H \cdots W_1 \Sigma_{XX} - \Sigma_{YX})(W_{H-1} \cdots W_1)^T = 0 \\ \nabla_{W_1} L(\mathbf{W}) = 2(W_H \cdots W_2)^T (W_H \cdots W_1 \Sigma_{XX} - \Sigma_{YX}) = 0 \; . \end{cases} \tag{19}$$

Using the definition of $\widetilde{\mathbf{W}}$ above, we have

$$\begin{cases} W_H \cdots W_1 = \widetilde{W}_H \cdots \widetilde{W}_1 \\ W_H \cdots W_{h+1} = \widetilde{W}_H \cdots \widetilde{W}_{h+1} D_h^{-1} \quad \forall h \in [\![1, H-1]\!] \\ W_{h-1} \cdots W_1 = D_{h-1} \widetilde{W}_{h-1} \cdots \widetilde{W}_1 \quad \forall h \in [\![2, H]\!] \; . \end{cases}$$

Therefore (19) is equivalent to

$$\begin{cases} (D_h^{-1})^T (\widetilde{W}_H \cdots \widetilde{W}_{h+1})^T (\widetilde{W}_H \cdots \widetilde{W}_1 \Sigma_{XX} - \Sigma_{YX})(\widetilde{W}_{h-1} \cdots \widetilde{W}_1)^T D_{h-1}^T = 0 \quad \forall h \in [\![2, H-1]\!] \\ (\widetilde{W}_H \cdots \widetilde{W}_1 \Sigma_{XX} - \Sigma_{YX})(\widetilde{W}_{H-1} \cdots \widetilde{W}_1)^T D_{H-1}^T = 0 \\ (D_1^{-1})^T (\widetilde{W}_H \cdots \widetilde{W}_2)^T (\widetilde{W}_H \cdots \widetilde{W}_1 \Sigma_{XX} - \Sigma_{YX}) = 0 \; . \end{cases}$$

This is equivalent to

$$\begin{cases} \nabla_{W_h} L(\widetilde{\mathbf{W}}) = 2(\widetilde{W}_H \cdots \widetilde{W}_{h+1})^T (\widetilde{W}_H \cdots \widetilde{W}_1 \Sigma_{XX} - \Sigma_{YX})(\widetilde{W}_{h-1} \cdots \widetilde{W}_1)^T = 0 \quad \forall h \in [\![2, H-1]\!] \\ \nabla_{W_H} L(\widetilde{\mathbf{W}}) = 2(\widetilde{W}_H \cdots \widetilde{W}_1 \Sigma_{XX} - \Sigma_{YX})(\widetilde{W}_{H-1} \cdots \widetilde{W}_1)^T = 0 \\ \nabla_{W_1} L(\widetilde{\mathbf{W}}) = 2(\widetilde{W}_H \cdots \widetilde{W}_2)^T (\widetilde{W}_H \cdots \widetilde{W}_1 \Sigma_{XX} - \Sigma_{YX}) = 0 \; . \end{cases}$$

which is equivalent to $\nabla_{W_h} L(\widetilde{\mathbf{W}}) = 0$, for all $h \in [\![1, H]\!]$. Therefore, $\mathbf{W}$ is a first-order critical point if and only if $\widetilde{\mathbf{W}}$ is a first-order critical point. This proves the first part of the proposition.
Now assume that $\mathbf{W} = (W_H, \dots, W_1)$ is a first-order critical point such that it is not a second-order critical point. Note that from the first part of the proof $\widetilde{\mathbf{W}} = (\widetilde{W_H}, \dots, \widetilde{W_1})$ is also a first-order critical point. Let us prove that $\widetilde{\mathbf{W}}$ is not a second-order critical point. Using Lemma 12, since $\mathbf{W}$ is not a second-order critical point, there exist $\mathbf{W}' = (W'_H, \dots, W'_1)$ such that, if we denote $\mathbf{W}(t) = \mathbf{W} + t\mathbf{W}'$, the second-order term of $L(\mathbf{W}(t))$ is strictly negative i.e $c_2(\mathbf{W}, \mathbf{W}') < 0$. We will prove that there exist $\widetilde{\mathbf{W}}'$ such that $c_2(\widetilde{\mathbf{W}}, \widetilde{\mathbf{W}}') < 0$ and, using again Lemma 12, we conclude.
As already said, we set $W_h(t) = W_h + tW'_h$, for all $h \in [\![1, H]\!]$. We denote

$$\begin{cases} \widetilde{W_H}(t) = \widetilde{W_H} + t\widetilde{W'_H} = \widetilde{W_H} + tW'_H D_{H-1} \\ \widetilde{W_1}(t) = \widetilde{W_1} + t\widetilde{W'_1} = \widetilde{W_1} + tD_1^{-1} W'_1 \\ \widetilde{W_h}(t) = \widetilde{W_h} + t\widetilde{W'_h} = \widetilde{W_h} + tD_h^{-1} W'_h D_{h-1} \quad \forall h \in [\![2, H-1]\!] \\ \widetilde{\mathbf{W}}' = (\widetilde{W'_H}, \dots, \widetilde{W'_1}) \; . \end{cases}$$

Hence, we have (where $\prod_{h=H-1}^{2} A_h$ should read as $A_{H-1} \cdots A_2$)

$$\begin{aligned} &\widetilde{W_H}(t) \cdots \widetilde{W_1}(t) \\ &= (W_H D_{H-1} + tW'_H D_{H-1}) \left(\prod_{h=H-1}^{2} (D_h^{-1} W_h D_{h-1} + tD_h^{-1} W'_h D_{h-1}) \right) (D_1^{-1} W_1 + tD_1^{-1} W'_1) \\ &= (W_H + tW'_H) \cdots (W_1 + tW'_1) \\ &= W_H(t) \cdots W_1(t) \; . \end{aligned}$$

Therefore, $L(\widetilde{\mathbf{W}}(t)) = L(\mathbf{W}(t))$ and

$$c_2(\widetilde{\mathbf{W}}, \widetilde{\mathbf{W}}') = c_2(\mathbf{W}, \mathbf{W}') \; .$$

Since by hypothesis $c_2(\mathbf{W}, \mathbf{W}') < 0$, we conclude that $c_2(\widetilde{\mathbf{W}}, \widetilde{\mathbf{W}}') < 0$. Hence $(\widetilde{W_H}, \dots, \widetilde{W_1})$ is not a second-order critical point.
We prove that if $\widetilde{\mathbf{W}}$ is not a second-order critical point then $\mathbf{W}$ is not a second-order critical point in the same way, by changing D_h with D_h^{-1} for all $h \in [\![1, H]\!]$. This proves the second part of the proposition and concludes the proof.

B.5 Proof of Proposition 10

Let $\mathcal{S} \subset [\![1, d_y]\!]$ of size $r \in [\![0, r_{max}]\!]$ and $Q = [\![1, d_y]\!] \setminus \mathcal{S}$. Let $Z_H \in \mathbb{R}^{(d_y - r) \times (d_{H-1} - r)}$, $Z_1 \in \mathbb{R}^{(d_1 - r) \times d_x}$ and $Z_h \in \mathbb{R}^{(d_h - r) \times (d_{h-1} - r)}$ for $h \in [\![2, H-1]\!]$. Let the parameter of the network $\mathbf{W} = (W_H, \dots, W_1)$ be defined as follows:

$$\begin{cases} W_H = [U_{\mathcal{S}}, U_Q Z_H] \\ W_1 = \begin{bmatrix} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} \\ Z_1 \end{bmatrix} \\ W_h = \begin{bmatrix} I_r & 0 \\ 0 & Z_h \end{bmatrix} \quad \forall h \in [\![2, H-1]\!] \; . \end{cases} \tag{20}$$

Note that the above definition of **W** does not involve the matrices $D_h \in \mathbb{R}^{d_h \times d_h}$. In fact, using Lemma 16, it suffices to prove that, when $r = r_{max}$ or there exist $h_1 \neq h_2$ such that $Z_{h_1} = 0$ and $Z_{h_2} = 0$, the **W** defined above is a first-order critical point to conclude that Proposition 10 holds .
We have

$$\begin{aligned} W_H \cdots W_1 &= [U_{\mathcal{S}}, U_Q Z_H] \begin{bmatrix} I_r & 0 \\ 0 & Z_{H-1} \end{bmatrix} \cdots \begin{bmatrix} I_r & 0 \\ 0 & Z_2 \end{bmatrix} \begin{bmatrix} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} \\ Z_1 \end{bmatrix} \\ &= U_{\mathcal{S}} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} + U_Q Z_H Z_{H-1} \cdots Z_2 Z_1 \end{aligned}$$

If there exists $h_1 \neq h_2$ such that $Z_{h_1} = 0$ and $Z_{h_2} = 0$, it immediately follows that $W_H \cdots W_1 = U_{\mathcal{S}} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1}$.
If $r = r_{max}$, then there exists $h \in [\![0, H]\!]$ such that $r = d_h$.

- If $r = d_H = d_y$, then $U_Q \in \mathbb{R}^{d_y \times 0}$ and $Z_H \in \mathbb{R}^{0 \times (d_{H-1} - r)}$, which, using conventions in Section 2, gives

$$U_Q Z_H = 0_{d_y \times (d_{H-1} - r)}. \tag{21}$$

  Therefore, $W_H \cdots W_1 = U_{\mathcal{S}} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1}$.

- If $r = d_0 = d_x$, then, since $d_x \geq d_y$, we have $r = d_y$, which we have already treated in the previous item.

- If $r = d_h$ for some $h \in [\![2, H-1]\!]$, then $Z_{h+1} \in \mathbb{R}^{(d_{h+1} - r) \times 0}$ and $Z_h \in \mathbb{R}^{0 \times (d_{h-1} - r)}$, which, using the conventions on Section 2, gives

$$Z_{h+1} Z_h = 0_{(d_{h+1} - r) \times (d_{h-1} - r)}. \tag{22}$$

  Therefore, $W_H \cdots W_1 = U_{\mathcal{S}} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1}$.

- If $r = d_1$, then $Z_2 \in \mathbb{R}^{(d_2 - r) \times 0}$ and $Z_1 \in \mathbb{R}^{0 \times d_x}$, which, using the conventions on Section 2, gives

$$Z_2 Z_1 = 0_{(d_2 - r) \times d_x}. \tag{23}$$

  Therefore, $W_H \cdots W_1 = U_{\mathcal{S}} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1}$.

Note that these results still hold if there is more than one layer with the minimum width.
Therefore, in all cases, when $r = r_{max}$ or there exist $h_1 \neq h_2$ such that $Z_{h_1} = 0$ and $Z_{h_2} = 0$ we have,

$$W_H \cdots W_1 = U_{\mathcal{S}} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} \ . \tag{24}$$

Let us prove that the gradient of L at **W** is equal to zero.
Recall that from Lemma 19 we have

$$\begin{aligned} \nabla_{W_h} L(\mathbf{W}) &= 2(W_H \cdots W_{h+1})^T (W_H \cdots W_1 \Sigma_{XX} - \Sigma_{YX})(W_{h-1} \cdots W_1)^T \quad \forall h \in [\![2, H-1]\!] \\ \nabla_{W_H} L(\mathbf{W}) &= 2(W_H \cdots W_1 \Sigma_{XX} - \Sigma_{YX})(W_{H-1} \cdots W_1)^T \\ \nabla_{W_1} L(\mathbf{W}) &= 2(W_H \cdots W_2)^T (W_H \cdots W_1 \Sigma_{XX} - \Sigma_{YX}) \ . \end{aligned}$$

Using (24) and Lemma 22, we have

$$\begin{aligned} W_H \cdots W_1 \Sigma_{XX} - \Sigma_{YX} &= U_{\mathcal{S}} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XX} - \Sigma_{YX} \\ &= (U_{\mathcal{S}} U_{\mathcal{S}}^T - I_{d_y}) \Sigma_{YX} \\ &= -U_Q U_Q^T \Sigma_{YX} \ . \end{aligned}$$

Also, using (20), for all $h \in [\![1, H-1]\!]$,

$$\begin{aligned} W_H \cdots W_{h+1} &= [U_{\mathcal{S}}, U_Q Z_H] \begin{bmatrix} I_r & 0 \\ 0 & Z_{H-1} \end{bmatrix} \cdots \begin{bmatrix} I_r & 0 \\ 0 & Z_{h+1} \end{bmatrix} \\ &= [U_{\mathcal{S}}, U_Q Z_H Z_{H-1} \cdots Z_{h+1}] \end{aligned}$$

and, for all $h \in [\![2, H]\!]$,

$$\begin{aligned} W_{h-1} \cdots W_1 &= \begin{bmatrix} I_r & 0 \\ 0 & Z_{h-1} \end{bmatrix} \cdots \begin{bmatrix} I_r & 0 \\ 0 & Z_2 \end{bmatrix} \begin{bmatrix} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} \\ Z_1 \end{bmatrix} \\ &= \begin{bmatrix} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} \\ Z_{h-1} \cdots Z_2 Z_1 \end{bmatrix} \ . \end{aligned}$$

We have, for all $h \in [\![2, H-1]\!]$,

$$\begin{aligned} \frac{1}{2} \left(\nabla_{W_h} L(\mathbf{W}) \right)^T &= (W_{h-1} \cdots W_1)(W_H \cdots W_1 \Sigma_{XX} - \Sigma_{YX})^T (W_H \cdots W_{h+1}) \\ &= - \begin{bmatrix} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} \\ Z_{h-1} \cdots Z_2 Z_1 \end{bmatrix} (U_Q U_Q^T \Sigma_{YX})^T [U_{\mathcal{S}}, U_Q Z_H Z_{H-1} \cdots Z_{h+1}] \\ &= - \begin{bmatrix} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} \\ Z_{h-1} \cdots Z_2 Z_1 \end{bmatrix} \Sigma_{XY} U_Q U_Q^T [U_{\mathcal{S}}, U_Q Z_H Z_{H-1} \cdots Z_{h+1}] \\ &= - \begin{bmatrix} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY} U_Q \\ Z_{h-1} \cdots Z_2 Z_1 \Sigma_{XY} U_Q \end{bmatrix} [U_Q^T U_{\mathcal{S}}, U_Q^T U_Q Z_H Z_{H-1} \cdots Z_{h+1}] \ . \end{aligned}$$

Using the definition of Σ, Lemma 22 and Lemma 23, we have

$$\begin{aligned} \frac{1}{2} \left(\nabla_{W_h} L(\mathbf{W}) \right)^T &= - \begin{bmatrix} U_{\mathcal{S}}^T \Sigma U_Q \\ Z_{h-1} \cdots Z_2 Z_1 \Sigma_{XY} U_Q \end{bmatrix} [0_{(d_y - r) \times r}, Z_H Z_{H-1} \cdots Z_{h+1}] \\ &= - \begin{bmatrix} 0_{r \times (d_y - r)} \\ Z_{h-1} \cdots Z_2 Z_1 \Sigma_{XY} U_Q \end{bmatrix} [0_{(d_y - r) \times r}, Z_H Z_{H-1} \cdots Z_{h+1}] \\ &= - \begin{bmatrix} 0_{r \times r} & 0_{r \times (d_h - r)} \\ 0_{(d_{h-1} - r) \times r} & Z_{h-1} \cdots Z_2 Z_1 \Sigma_{XY} U_Q Z_H Z_{H-1} \cdots Z_{h+1} \end{bmatrix} \ . \end{aligned}$$

Proceeding similarly, we obtain

$$\frac{1}{2} \left(\nabla_{W_H} L(\mathbf{W}) \right)^T = - \begin{bmatrix} 0_{r \times d_y} \\ Z_{H-1} \cdots Z_2 Z_1 \Sigma_{XY} U_Q U_Q^T \end{bmatrix}$$

and

$$\frac{1}{2} \left(\nabla_{W_1} L(\mathbf{W}) \right)^T = -[0_{d_x \times r} \ , \ \Sigma_{XY} U_Q Z_H Z_{H-1} \cdots Z_2] \ .$$

If there exists $h_1 \neq h_2$ such that $Z_{h_1} = 0$ and $Z_{h_2} = 0$, we can easily see that the gradient is equal to zero, i.e., **W** is a first-order critical point.
If $r = r_{max}$, then there exists $h' \in [\![1, H]\!]$ such that $r = d_{h'}$. Using the same arguments as above that yielded (21), (22) and (23), we have,

- For $h = 1$,
  - if $r = d_1$, we have $Z_2 \in \mathbb{R}^{(d_2-r)\times 0}$ and therefore $\Sigma_{XY} U_Q Z_H Z_{H-1} \cdots Z_2 \in \mathbb{R}^{d_x \times 0}$.
  - if $r = d_H$, then $U_Q Z_H = 0_{d_y \times (d_{H-1}-r)}$ and therefore $\Sigma_{XY} U_Q Z_H Z_{H-1} \cdots Z_2 = 0_{d_x \times (d_1 - r)}$.
  - if $r = d_{h'}$ for some $h' \in [\![2, H-1]\!]$, then $Z_{h'+1} Z_{h'} = 0_{(d_{h'+1}-r)\times(d_{h'-1}-r)}$ and therefore $\Sigma_{XY} U_Q Z_H Z_{H-1} \cdots Z_2 = 0_{d_x \times (d_1-r)}$.

  Hence, in all cases, $\nabla_{W_1} L(\mathbf{W}) = 0$.

- For $h = H$,
  - if $r = d_H = d_y$, then $U_Q U_Q^T = 0_{d_y \times d_y}$ and therefore $Z_{H-1} \cdots Z_2 Z_1 \Sigma_{XY} U_Q U_Q^T = 0_{(d_{H-1}-r)\times d_y}$.
  - if $r = d_{H-1}$, then $Z_{H-1} \in \mathbb{R}^{0 \times (d_{H-2}-r)}$ and therefore $Z_{H-1} \cdots Z_2 Z_1 \Sigma_{XY} U_Q U_Q^T \in \mathbb{R}^{0\times d_y}$.
  - if $r = d_{h'}$ for some $h' \in [\![2, H-2]\!]$, then $Z_{h'+1} Z_{h'} = 0_{(d_{h'+1}-r)\times(d_{h'-1}-r)}$ and therefore $Z_{H-1} \cdots Z_2 Z_1 \Sigma_{XY} U_Q U_Q^T = 0_{(d_{H-1}-r)\times d_y}$.
  - if $r = d_1$, then $Z_2 Z_1 = 0_{(d_2-r)\times d_x}$ and therefore $Z_{H-1} \cdots Z_2 Z_1 \Sigma_{XY} U_Q U_Q^T = 0_{(d_{H-1}-r)\times d_y}$.

  Hence, in all cases, $\nabla_{W_H} L(\mathbf{W}) = 0$.

- For $h \in [\![2, H-1]\!]$,
  - if $r = d_{h-1}$, then $Z_{h-1} \in \mathbb{R}^{0\times(d_{h-2}-r)}$ and therefore $Z_{h-1} \cdots Z_2 Z_1 \Sigma_{XY} U_Q Z_H Z_{H-1} \cdots Z_{h+1} \in \mathbb{R}^{0 \times (d_h - r)}$.
  - if $r = d_h$, then $Z_{h+1} \in \mathbb{R}^{(d_{h+1}-r)\times 0}$ and therefore $Z_{h-1} \cdots Z_2 Z_1 \Sigma_{XY} U_Q Z_H Z_{H-1} \cdots Z_{h+1} \in \mathbb{R}^{(d_{h-1}-r)\times 0}$.
  - if $r = d_H$, then $U_Q Z_H = 0_{d_y \times (d_{H-1}-r)}$ and therefore $Z_{h-1} \cdots Z_2 Z_1 \Sigma_{XY} U_Q Z_H Z_{H-1} \cdots Z_{h+1} = 0_{(d_{h-1}-r)\times(d_h-r)}$.
  - if $r = d_1$, then $Z_2 Z_1 = 0_{(d_2-r)\times d_x}$ and therefore $Z_{h-1} \cdots Z_2 Z_1 \Sigma_{XY} U_Q Z_H Z_{H-1} \cdots Z_{h+1} = 0_{(d_{h-1}-r)\times(d_h-r)}$.
  - if $r = d_{h'}$ for some $h' \in [\![2, H-1]\!] \setminus \{h, h-1\}$, then $Z_{h'+1} Z_{h'} = 0_{(d_{h'+1}-r)\times(d_{h'-1}-r)}$ and therefore
    $Z_{h-1} \cdots Z_2 Z_1 \Sigma_{XY} U_Q Z_H Z_{H-1} \cdots Z_{h+1} = 0_{(d_{h-1}-r)\times(d_h-r)}$.

  Hence, in all cases, $\nabla_{W_h} L(\mathbf{W}) = 0$.

Therefore, when $r = r_{max}$, **W** is also a first-order critical point of L.

B.6 Proof of Proposition 2

Let $\mathcal{S} \subset [\![1, d_y]\!]$ such that $|\mathcal{S}| = r \leq r_{max}$, and $Q = [\![1, d_y]\!] \setminus \mathcal{S}$.
We define $\mathbf{W} = (W_H, \ldots, W_1)$ by:

$$
\begin{aligned}
W_H &= [U_{\mathcal{S}}, 0_{d_y \times (d_{H-1}-r)}] \\
W_h &= \begin{bmatrix} I_r & 0_{r \times (d_{h-1}-r)} \\ 0_{(d_h-r)\times r} & 0_{(d_h-r)\times(d_{h-1}-r)} \end{bmatrix} \quad \forall h \in [\![2, H-1]\!] \\
W_1 &= \begin{bmatrix} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} \\ 0_{(d_1-r)\times d_x} \end{bmatrix},
\end{aligned}
$$

By Proposition 10, $\mathbf{W}$ is a first-order critical point of L. Moreover, we have $W_H \cdots W_1 = U_{\mathcal{S}} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1}$. Therefore, $\mathbf{W}$ is a first-order critical point associated with $\mathcal{S}$.

B.7 Proof of Proposition 4

Let $\mathbf{W} = (W_H, \ldots, W_1)$ be a first-order critical point and $r = \text{rk}(W_H \cdots W_1)$, using Proposition 1 there exists a unique $\mathcal{S} \subset [\![1, d_y]\!]$ of size r such that

$$W_H \cdots W_1 = U_{\mathcal{S}} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1},$$

which implies

$$W_H \cdots W_1 \Sigma_{XY} = U_{\mathcal{S}} U_{\mathcal{S}}^T \Sigma.$$

Let $i, j \in [\![1, H]\!]$ such that $i > j$. The complementary blocks are $W_{j-1} \cdots W_1 \Sigma_{XY} W_H \cdots W_{i+1}$ and $W_{i-1} \cdots W_{j+1}$.
Using Lemma 20 and $U_{\mathcal{S}}^T U_{\mathcal{S}} = I_r$, we have, for the second complementary block,

$$\text{rk}(W_{i-1} \cdots W_{j+1}) \geq \text{rk}(W_H \cdots W_1 \Sigma_{XY}) = \text{rk}(U_{\mathcal{S}} U_{\mathcal{S}}^T \Sigma) \geq \text{rk}(U_{\mathcal{S}}^T (U_{\mathcal{S}} U_{\mathcal{S}}^T \Sigma) \Sigma^{-1} U_{\mathcal{S}}) = \text{rk}(I_r) = r \ .$$

For the first complementary block, using the same arguments, we have

$$
\begin{aligned}
\text{rk}(W_{j-1} \cdots W_1 \Sigma_{XY} W_H \cdots W_{i+1}) &\geq \text{rk}(W_H \cdots W_1 \Sigma_{XY} W_H \cdots W_1 \Sigma_{XY}) \\
&= \text{rk}(U_{\mathcal{S}} U_{\mathcal{S}}^T \Sigma U_{\mathcal{S}} U_{\mathcal{S}}^T \Sigma) \\
&\geq \text{rk}\left(U_{\mathcal{S}}^T (U_{\mathcal{S}} U_{\mathcal{S}}^T \Sigma U_{\mathcal{S}} U_{\mathcal{S}}^T \Sigma) \Sigma^{-1} U_{\mathcal{S}}\right) \\
&= \text{rk}(U_{\mathcal{S}}^T \Sigma U_{\mathcal{S}}) \ .
\end{aligned}
$$

Recall that, from the diagonalization of Σ, we have $\Sigma U = U \Lambda$, hence, $\Sigma U_{\mathcal{S}} = U_{\mathcal{S}} \text{diag}((\lambda_s)_{s \in \mathcal{S}})$

$$
\begin{aligned}
\text{rk}(W_{j-1} \cdots W_1 \Sigma_{XY} W_H \cdots W_{i+1}) &\geq \text{rk}(U_{\mathcal{S}}^T U_{\mathcal{S}} \text{diag}((\lambda_s)_{s \in \mathcal{S}})) \\
&= \text{rk}(\text{diag}((\lambda_s)_{s \in \mathcal{S}})) \\
&= r \ .
\end{aligned}
$$

This concludes the proof.

B.8 Proof of Proposition 8

Let $H \geq 3$, $\mathcal{S} = [\![1, r]\!]$ with $0 \leq r < r_{max}$. We define **W** as follows:

$$\begin{cases} W_H = [U_{\mathcal{S}}, 0] \\ W_h = \begin{bmatrix} I_r & 0 \\ 0 & Z_h \end{bmatrix} \quad \text{for} \quad h \in [\![2, H-1]\!] \\ W_1 = \begin{bmatrix} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} \\ 0 \end{bmatrix}. \end{cases} \tag{25}$$

Using Proposition 10, **W** is a first-order critical point associated with $\mathcal{S}$. Let us show that depending on the choice of $(Z_h)_{h=2..H-1}$, **W** can be tightened or non-tightened.
Since $H \geq 3$, there exists $h \in [\![2, H-1]\!]$. If we choose $Z_{H-1}, \ldots, Z_2$ such that $Z_{H-1} \cdots Z_2 \neq 0$ (e.g. when only the top left entry of each Z_h is nonzero, which is possible since $r < r_{max} = \min(d_H, \ldots, d_0)$) then **W** is non-tightened. Indeed, the pivot $(H, 1)$ is non-tightened because $\text{rk}(\Sigma_{XY}) = d_y > r$ and $\text{rk}(W_{H-1} \cdots W_2) = \text{rk}\left(\begin{bmatrix} I_r & 0 \\ 0 & Z_{H-1} \cdots Z_2 \end{bmatrix}\right) > r$.
If we choose $Z_{H-1}, \ldots, Z_2$ such that $Z_{H-1} \cdots Z_2 = 0$ (e.g. $Z_2 = 0$), then **W** is tightened. Indeed, the pivot $(H, 1)$ is tightened because $W_{H-1} \cdots W_2 = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}$ is of rank r, and by construction we have $\text{rk}(W_H) = \text{rk}(W_1) = r$. Hence, all the other pivots are tightened because at least one of their complementary blocks includes W_H or W_1, and therefore, using Proposition 4, is of rank r. Therefore, **W** is tightened.

Appendix C. Parameterization of First-order Critical Points and Global Minimizers

In this section, we prove Propositions 9 and 11 that were stated in Section 3.3.

C.1 Proof of Proposition 9

Before proving Proposition 9, we introduce and prove two lemmas.

Lemma 28 *Let r be a nonnegative integer, and let n and p be two positive integers larger than or equal to r. Let $\mathcal{S} \subset [\![1, d_y]\!]$ of size r and let $Q = [\![1, d_y]\!] \setminus \mathcal{S}$. Let $A \in \mathbb{R}^{d_y \times n}$ and $B \in \mathbb{R}^{n \times p}$ be two matrices such that*

$$AB = [U_{\mathcal{S}}, 0] \, .$$

Then, there exist an invertible matrix $D \in \mathbb{R}^{n \times n}$ and two matrices $N \in \mathbb{R}^{(d_y - r) \times (n-r)}$ and $B_{DR} \in \mathbb{R}^{(n-r) \times (p-r)}$ such that

$$AD = [U_{\mathcal{S}}, U_Q N] \tag{26}$$

$$D^{-1} B = \begin{bmatrix} I_r & 0 \\ 0 & B_{DR} \end{bmatrix}. \tag{27}$$

In the proof below, we can easily see that the result still holds for $r = 0$ and $r = \min(d_y, n, p)$ with the conventions adopted in Section 2.

Proof Let n and p be non-negative integers such that $n, p \geq r$ and $A \in \mathbb{R}^{d_y \times n}$ and $B \in \mathbb{R}^{n \times p}$ such that

$$AB = [U_{\mathcal{S}}, 0]. \tag{28}$$

Recall that for any matrix C with n columns we write $C = [C_1, C_2, \ldots, C_n]$ where C_i represents the i-th column of C.
We have from (28),

$$A[B_1, B_2, \ldots, B_r] = U_{\mathcal{S}} \ . \tag{29}$$

Since the columns of U are linearly independent, we have

$$\mathrm{rk}(A[B_1, B_2, \ldots, B_r]) = \mathrm{rk}(U_{\mathcal{S}}) = r$$

and $\{B_1, \ldots, B_r\}$ are necessarily also linearly independent. Using the incomplete basis theorem, we complement $(B_1, \ldots, B_r)$ to form a basis $(B_1, \ldots, B_r, E_{r+1}, \ldots, E_n)$. We set $E = [B_1, \ldots, B_r, E_{r+1}, \ldots, E_n] \in \mathbb{R}^{n \times n}$. By construction, the matrix E is invertible.

We now set $A' = AE$ and $B' = E^{-1}B$. In particular $A'B' = AB$.
Also, note that

$$E \begin{bmatrix} I_r \\ 0 \end{bmatrix} = [B_1, \ldots, B_r] \ ,$$

so that

$$E^{-1}[B_1, \ldots, B_r] = \begin{bmatrix} I_r \\ 0 \end{bmatrix} \ .$$

Therefore, we can write

$$B' = E^{-1}B = \begin{bmatrix} I_r & B_{UR} \\ 0 & B_{DR} \end{bmatrix} \ , \tag{30}$$

with $B_{UR} \in \mathbb{R}^{r \times (p-r)}$ and $B_{DR} \in \mathbb{R}^{(n-r) \times (p-r)}$ such that

$$\begin{bmatrix} B_{UR} \\ B_{DR} \end{bmatrix} = E^{-1}[B_{r+1}, \ldots, B_p] \ .$$

We define $L \in \mathbb{R}^{r \times (n-r)}$ and $N \in \mathbb{R}^{(d_y - r) \times (n-r)}$ by $\begin{bmatrix} L \\ N \end{bmatrix} = [U_{\mathcal{S}}, U_Q]^{-1}[AE_{r+1}, \ldots, AE_n]$. We have

$$[AE_{r+1}, \ldots, AE_n] = [U_{\mathcal{S}}, U_Q] \begin{bmatrix} L \\ N \end{bmatrix} = U_{\mathcal{S}}L + U_Q N \ . \tag{31}$$

We also define the invertible matrix $F = \begin{bmatrix} I_r & L \\ 0 & I_{n-r} \end{bmatrix} \in \mathbb{R}^{n \times n}$. Using (29) and (31) we have

$$\begin{aligned} A' &= AE \\ &= A[B_1, \ldots, B_r, E_{r+1}, \ldots, E_n] \\ &= [U_{\mathcal{S}}, U_{\mathcal{S}}L + U_Q N] \\ &= [U_{\mathcal{S}}, U_Q N] \begin{bmatrix} I_r & L \\ 0 & I_{n-r} \end{bmatrix} \\ &= [U_{\mathcal{S}}, U_Q N] F \ . \end{aligned}$$

Therefore, defining the invertible matrix $D = EF^{-1} \in \mathbb{R}^{n\times n}$, we finally have

$$AD = AEF^{-1} = [U_{\mathcal{S}}, U_Q N]\ . \tag{32}$$

This proves (26).
We also have, using (30) and the definition of F

$$\begin{aligned} D^{-1}B &= FE^{-1}B \\ &= FB' \\ &= \begin{bmatrix} I_r & L \\ 0 & I_{n-r} \end{bmatrix} \begin{bmatrix} I_r & B_{UR} \\ 0 & B_{DR} \end{bmatrix} \\ &= \begin{bmatrix} I_r & B_{UR} + LB_{DR} \\ 0 & B_{DR} \end{bmatrix}\ . \end{aligned} \tag{33}$$

However, noticing that, since (28) holds,

$$(AD)(D^{-1}B) = AB = [U_{\mathcal{S}}, 0]\ ,$$

and using (32) and (33) we obtain

$$[U_{\mathcal{S}}, U_Q N] \begin{bmatrix} I_r & B_{UR} + LB_{DR} \\ 0 & B_{DR} \end{bmatrix} = [U_{\mathcal{S}}, 0]\ .$$

Therefore $U_{\mathcal{S}}(B_{UR}+LB_{DR})+U_Q N B_{DR} = 0$. Since $[U_{\mathcal{S}}, U_Q]$ is invertible we get $B_{UR}+LB_{DR} = 0$ and $NB_{DR} = 0$.
Finally, (33) becomes

$$D^{-1}B = \begin{bmatrix} I_r & 0 \\ 0 & B_{DR} \end{bmatrix}\ .$$

This proves (27) and concludes the proof. ■

The second lemma states that if the product of two factors takes the format of (27), then up to the product by an invertible matrix, the two factors have the same format. In the proof of Proposition 9, we will use this property several times to establish (4).

Lemma 29 *Let r, q, n and p be positive integers such that $r \leq \min(q, n, p)$. Let $B \in \mathbb{R}^{q\times n}$, $C \in \mathbb{R}^{n\times p}$ and $P \in \mathbb{R}^{(q-r)\times(p-r)}$ such that*

$$BC = \begin{bmatrix} I_r & 0 \\ 0 & P \end{bmatrix}\ .$$

Then, there exist an invertible matrix $D \in \mathbb{R}^{n\times n}$ and two matrices $B_{DR} \in \mathbb{R}^{(q-r)\times(n-r)}$ and $C_{DR} \in \mathbb{R}^{(n-r)\times(p-r)}$ such that

$$BD = \begin{bmatrix} I_r & 0 \\ 0 & B_{DR} \end{bmatrix} \tag{34}$$

$$D^{-1}C = \begin{bmatrix} I_r & 0 \\ 0 & C_{DR} \end{bmatrix}. \tag{35}$$

In the proof below, we can easily see that the result still holds for $r = 0$ and $r = \min(q, n, p)$ with the conventions adopted in Section 2.

Proof Let r, q, n and p be positive integers such that $r \leq \min(q, n, p)$. Let $B \in \mathbb{R}^{q\times n}$, $C \in \mathbb{R}^{n\times p}$ and $P \in \mathbb{R}^{(q-r)\times(p-r)}$ such that

$$BC = \begin{bmatrix} I_r & 0 \\ 0 & P \end{bmatrix} . \tag{36}$$

We have

$$B[C_1, C_2, \ldots, C_r] = \begin{bmatrix} I_r \\ 0 \end{bmatrix} . \tag{37}$$

Since the columns of $\begin{bmatrix} I_r \\ 0 \end{bmatrix}$ are linearly independent,

$$\mathrm{rk}(B[C_1, C_2, \ldots, C_r]) = r$$

and the vectors $C_1, \ldots, C_r$ are necessarily also linearly independent. Using the incomplete basis theorem, we complement $(C_1, \ldots, C_r)$ to form a basis $(C_1, \ldots, C_r, E_{r+1}, \ldots, E_n)$. We denote $E = [C_1, \ldots, C_r, E_{r+1}, \ldots, E_n] \in \mathbb{R}^{n\times n}$. By construction, the matrix E is invertible.

We now set $B' = BE$ and $C' = E^{-1}C$. In particular

$$B'C' = BC. \tag{38}$$

Also notice that

$$E \begin{bmatrix} I_r \\ 0 \end{bmatrix} = [C_1, \ldots, C_r] ,$$

so that

$$E^{-1}[C_1, \ldots, C_r] = \begin{bmatrix} I_r \\ 0 \end{bmatrix} .$$

Therefore, we can write

$$C' = E^{-1}C = \begin{bmatrix} I_r & C_{UR} \\ 0 & C_{DR} \end{bmatrix} , \tag{39}$$

where $C_{UR} \in \mathbb{R}^{r\times(p-r)}$ and $C_{DR} \in \mathbb{R}^{(n-r)\times(p-r)}$ are such that $\begin{bmatrix} C_{UR} \\ C_{DR} \end{bmatrix} = E^{-1}[C_{r+1}, \ldots, C_p]$.

Now notice that, using (37),

$$\begin{aligned} B' &= BE \\ &= B[C_1, \ldots, C_r, E_{r+1}, \ldots, E_n] \\ &= \begin{bmatrix} I_r & B_{UR} \\ 0 & B_{DR} \end{bmatrix} , \end{aligned} \tag{40}$$

where $B_{UR} \in \mathbb{R}^{r\times(n-r)}$ and $B_{DR} \in \mathbb{R}^{(q-r)\times(n-r)}$ are such that $\begin{bmatrix} B_{UR} \\ B_{DR} \end{bmatrix} = B[E_{r+1}, \ldots, E_n]$.

Plugging (40), (39) and (36) in the equality (38), we obtain

$$\begin{bmatrix} I_r & B_{UR} \\ 0 & B_{DR} \end{bmatrix} \begin{bmatrix} I_r & C_{UR} \\ 0 & C_{DR} \end{bmatrix} = \begin{bmatrix} I_r & 0 \\ 0 & P \end{bmatrix} ,$$

which yields

$$\begin{bmatrix} I_r & C_{UR}+B_{UR}C_{DR} \\ 0 & B_{DR}C_{DR} \end{bmatrix} = \begin{bmatrix} I_r & 0 \\ 0 & P \end{bmatrix} .$$

Therefore, $C_{UR} + B_{UR}C_{DR} = 0$ or, equivalently ,

$$C_{UR} = -B_{UR}C_{DR} . \tag{41}$$

Define $F = \begin{bmatrix} I_r & -B_{UR} \\ 0 & I_{n-r} \end{bmatrix}$. The matrix F is invertible. Moreover, using (39) and (41) we have

$$\begin{aligned} C' &= \begin{bmatrix} I_r & -B_{UR}C_{DR} \\ 0 & C_{DR} \end{bmatrix} \\ &= \begin{bmatrix} I_r & -B_{UR} \\ 0 & I_{n-r} \end{bmatrix} \begin{bmatrix} I_r & 0 \\ 0 & C_{DR} \end{bmatrix} \\ &= F \begin{bmatrix} I_r & 0 \\ 0 & C_{DR} \end{bmatrix} . \end{aligned}$$

Therefore, if we define $D = EF$, D is invertible and

$$D^{-1}C = F^{-1}E^{-1}C = F^{-1}C' = \begin{bmatrix} I_r & 0 \\ 0 & C_{DR} \end{bmatrix}.$$

This proves (35).
In order to prove (34), we remark that, using (40) and the definition of F, we also have

$$\begin{aligned} BD &= BEF \\ &= B'F \\ &= \begin{bmatrix} I_r & B_{UR} \\ 0 & B_{DR} \end{bmatrix} \begin{bmatrix} I_r & -B_{UR} \\ 0 & I_{n-r} \end{bmatrix} \\ &= \begin{bmatrix} I_r & 0 \\ 0 & B_{DR} \end{bmatrix} . \end{aligned}$$

This proves (34) and concludes the proof. ■

Now we prove Proposition 9.
Proof [Proof of Proposition 9]
Let $\mathbf{W} = (W_H, \dots, W_1)$ be a first-order critical point of L. Then using Lemma 26 there exist $D \in \mathbb{R}^{d_1 \times d_1}$ invertible and a matrix $M \in \mathbb{R}^{d_1 \times d_x}$ which satisfies $W_H \cdots W_2 M = 0$ such that

$$W_H \cdots W_2 = [U_{\mathcal{S}}\ ,\ 0]D \tag{42}$$

$$W_1 = D^{-1} \begin{bmatrix} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} \\ 0 \end{bmatrix} + M . \tag{43}$$

Denoting $D_1 = D^{-1}$ and using (42), we have $W_H \cdots W_2 D_1 = [U_{\mathcal{S}}, 0]$. Then applying Lemma 28 with $A = W_H$ and $B = W_{H-1} \cdots W_2 D_1$, there exist an invertible matrix $D_{H-1} \in \mathbb{R}^{d_{H-1} \times d_{H-1}}$, and matrices $Z_H \in \mathbb{R}^{(d_y-r)\times(d_{H-1}-r)}$ and $B_{DR} \in \mathbb{R}^{(d_{H-1}-r)\times(d_1-r)}$ such that

$$\widetilde{W}_H := W_H D_{H-1} = [U_{\mathcal{S}}, U_Q Z_H]$$
$$D_{H-1}^{-1} W_{H-1} \cdots W_2 D_1 = \begin{bmatrix} I_r & 0 \\ 0 & B_{DR} \end{bmatrix} . \tag{44}$$

The first equality proves (2).
Then applying Lemma 29 to (44) with $B = D_{H-1}^{-1} W_{H-1}$ and $C = W_{H-2} \cdots W_2 D_1$ we get the existence of an invertible matrix $D_{H-2} \in \mathbb{R}^{d_{H-2}\times d_{H-2}}$, $C_{DR} \in \mathbb{R}^{(d_{H-2}-r)\times(d_1-r)}$ and $Z_{H-1} \in \mathbb{R}^{(d_{H-1}-r)\times(d_{H-2}-r)}$ such that

$$\widetilde{W}_{H-1} := D_{H-1}^{-1} W_{H-1} D_{H-2} = \begin{bmatrix} I_r & 0 \\ 0 & Z_{H-1} \end{bmatrix} ,$$

and

$$D_{H-2}^{-1} W_{H-2} \cdots W_2 D_1 = \begin{bmatrix} I_r & 0 \\ 0 & C_{DR} \end{bmatrix} .$$

Reiterating the process by using Lemma 29 multiple times with $B = D_h^{-1} W_h$ and $C = W_{h-1} \cdots W_2 D_1$ for h decreasing from $H-2$ to 3, we can conclude that there exist invertible matrices $D_h \in \mathbb{R}^{d_h \times d_h}$ and matrices $Z_h \in \mathbb{R}^{(d_h - r)\times(d_{h-1}-r)}$, for $h \in [\![2, H-1]\!]$, such that

$$\widetilde{W}_h := D_h^{-1} W_h D_{h-1} = \begin{bmatrix} I_r & 0 \\ 0 & Z_h \end{bmatrix} \quad \forall h \in [\![2, H-1]\!] .$$

This entails (4).
We also have from (43) that $W_1 = D_1 \begin{bmatrix} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} \\ 0 \end{bmatrix} + M$ with $W_H \cdots W_2 M = 0$. Therefore,

$$D_1^{-1} W_1 = \begin{bmatrix} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} \\ 0 \end{bmatrix} + D_1^{-1} M .$$

Using (42), $D_1 = D^{-1}$ and $W_H \cdots W_2 M = 0$, we obtain

$$[U_{\mathcal{S}}\ ,\ 0] D_1^{-1} M = 0 .$$

Writing $D_1^{-1} M = \begin{bmatrix} Z_0 \\ Z_1 \end{bmatrix}$, where $Z_0 \in \mathbb{R}^{r \times d_x}$ and $Z_1 \in \mathbb{R}^{(d_1 - r)\times d_x}$, we have

$$\begin{aligned} 0 &= [U_{\mathcal{S}}\ ,\ 0] D_1^{-1} M \\ &= [U_{\mathcal{S}}\ ,\ 0] \begin{bmatrix} Z_0 \\ Z_1 \end{bmatrix} \\ &= U_{\mathcal{S}} Z_0 . \end{aligned}$$

Multiplying on the left by $U_{\mathcal{S}}^T$ we obtain

$$Z_0 = 0.$$

Therefore $D_1^{-1}M = \begin{bmatrix} 0 \\ Z_1 \end{bmatrix}$, which yields

$$\widetilde{W}_1 := D_1^{-1}W_1 = \begin{bmatrix} U_{\mathcal{S}}^T \Sigma_{YX}\Sigma_{XX}^{-1} \\ Z_1 \end{bmatrix} \ .$$

This proves (3).
Finally we have

$$\begin{aligned} \widetilde{W}_H \cdots \widetilde{W}_2 &= (W_H D_{H-1})(D_{H-1}^{-1} W_{H-1} D_{H-2}) \cdots (D_2^{-1} W_2 D_1) \\ &= W_H \cdots W_2 D_1 \\ &= [U_{\mathcal{S}} \ , \ 0] \ , \end{aligned}$$

where the last equality is due to (42) and $D_1 = D^{-1}$. This entails (5) and concludes the proof. ■

C.2 Proof of Proposition 11

We first make a comment about notational subtleties to help understand the statement of Proposition 11, and then prove the proposition.

Recall that $r_{max} = \min(d_H, \dots, d_0)$, and $d_x = d_0 \geq d_y = d_H$ by assumption. Therefore, in the statement of Proposition 11, some blocks Z_h have 0 lines or 0 columns, and thus do not exist. For example, depending on the value of r_{max}, we have

$$\begin{cases} W_H = U_{\mathcal{S}_{max}} D_{H-1}^{-1} & \text{if } r_{max} = d_{H-1} \\ W_1 = D_1 U_{\mathcal{S}_{max}}^T \Sigma_{YX}\Sigma_{XX}^{-1} & \text{if } r_{max} = d_1 \end{cases}$$

and for $h \in [\![2, H-1]\!]$

$$W_h = \begin{cases} D_h \begin{bmatrix} I_{r_{max}} & 0 \end{bmatrix} D_{h-1}^{-1} & \text{if } r_{max} = d_h < d_{h-1} \\ D_h \begin{bmatrix} I_{r_{max}} \\ 0 \end{bmatrix} D_{h-1}^{-1} & \text{if } r_{max} = d_{h-1} < d_h \\ D_h I_{r_{max}} D_{h-1}^{-1} & \text{if } r_{max} = d_h = d_{h-1} \end{cases}$$

Also, if $r_{max} = d_y$, then $Q_{max} = \emptyset$, hence $U_{Q_{max}} \in \mathbb{R}^{d_y \times 0}$ and $Z_H \in \mathbb{R}^{0 \times (d_{H-1} - r_{max})}$. Then, using the convention in Section 2, $U_{Q_{max}} Z_H = 0_{d_y \times (d_{H-1} - r_{max})}$, so that $W_H = [U_{\mathcal{S}_{max}}, 0_{d_y \times (d_{H-1} - r_{max})}] D_{H-1}^{-1} \in \mathbb{R}^{d_y \times d_{H-1}}$.

We are now ready to prove the proposition.

Proof [Proof of Proposition 11]

Let $\mathcal{S}_{max} = [\![1, r_{max}]\!]$. Let us first prove that $\mathbf{W}$ is a global minimizer of L if and only if $\mathbf{W}$ is a first-order critical point of L associated with $\mathcal{S}_{max}$. From Lemma 21, we have

$$U_{\mathcal{S}_{max}} U_{\mathcal{S}_{max}}^T \Sigma_{YX}\Sigma_{XX}^{-1} \in \mathop{\arg\min}_{\substack{R \in \mathbb{R}^{d_y \times d_x} \\ \text{rk}(R) \leq r_{max}}} \|RX - Y\|^2 \ .$$

Let $\mathbf{W}$ be a first-order critical point associated with $\mathcal{S}_{max}$ (note that from Proposition 2, such $\mathbf{W}$ exist). We have $W_H \cdots W_1 = U_{\mathcal{S}_{max}} U_{\mathcal{S}_{max}}^T \Sigma_{YX} \Sigma_{XX}^{-1}$, hence, for all $\mathbf{W}' = (W'_H, \ldots, W'_1)$, since $\mathrm{rk}(W'_H \cdots W'_1) \leq r_{max}$, we have

$$L(\mathbf{W}') \geq \min_{\substack{R \in \mathbb{R}^{d_y \times d_x} \\ \mathrm{rk}(R) \leq r_{max}}} \|RX - Y\|^2 = \|W_H \cdots W_1 X - Y\|^2 = L(\mathbf{W}) \ .$$

As a consequence, $\mathbf{W}$ is a global minimizer of L.
Conversely, if $\mathbf{W}$ is a global minimizer of L, then $\mathbf{W}$ is a first-order critical point of L. From Proposition 1, there exist $\mathcal{S} \subset [\![1, d_y]\!]$ of size $r \in [\![0, r_{max}]\!]$ such that $W_H \cdots W_1 = U_{\mathcal{S}} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1}$, and we have $L(\mathbf{W}) = \mathrm{tr}(\Sigma_{YY}) - \sum_{i \in \mathcal{S}} \lambda_i$. But we have from Assumption 1, $\lambda_1 > \ldots > \lambda_{d_y}$, and, since Σ is invertible (see Lemma 20), then $\lambda_{d_y} > 0$. Therefore, using Proposition 2, $\mathbf{W}$ is a global minimizer of L implies that $\mathcal{S} = [\![1, r_{max}]\!] = \mathcal{S}_{max}$. Hence, $\mathbf{W}$ is a global minimizer of L if and only if $\mathbf{W}$ is a first-order critical point of L associated with $\mathcal{S}_{max}$.
Let us now prove Proposition 11.
Let $\mathbf{W} = (W_H, \ldots, W_1)$ be a first-order critical point associated with $\mathcal{S}_{max} = [\![1, r_{max}]\!]$. Using Proposition 9, there exist invertible matrices $D_{H-1} \in \mathbb{R}^{d_{H-1} \times d_{H-1}}, \ldots, D_1 \in \mathbb{R}^{d_1 \times d_1}$, and matrices $Z_H \in \mathbb{R}^{(d_y - r_{max}) \times (d_{H-1} - r_{max})}$, $Z_h \in \mathbb{R}^{(d_h - r_{max}) \times (d_{h-1} - r_{max})}$ for $h \in [\![2, H-1]\!]$, and $Z_1 \in \mathbb{R}^{(d_1 - r_{max}) \times d_x}$ such that:

$$\begin{aligned} W_H &= [U_{\mathcal{S}_{max}}, U_{Q_{max}} Z_H] D_{H-1}^{-1} \\ W_1 &= D_1 \begin{bmatrix} U_{\mathcal{S}_{max}}^T \Sigma_{YX} \Sigma_{XX}^{-1} \\ Z_1 \end{bmatrix} \\ W_h &= D_h \begin{bmatrix} I_{r_{max}} & 0 \\ 0 & Z_h \end{bmatrix} D_{h-1}^{-1} \qquad \forall h \in [\![2, H-1]\!] \ . \end{aligned}$$

Conversely, consider matrices D_h, for $h \in [\![1, H-1]\!]$ and Z_h, for $h \in [\![1, H]\!]$ as in Proposition 11, and

$$\begin{aligned} W_H &= [U_{\mathcal{S}_{max}}, U_{Q_{max}} Z_H] D_{H-1}^{-1} \\ W_1 &= D_1 \begin{bmatrix} U_{\mathcal{S}_{max}}^T \Sigma_{YX} \Sigma_{XX}^{-1} \\ Z_1 \end{bmatrix} \\ W_h &= D_h \begin{bmatrix} I_{r_{max}} & 0 \\ 0 & Z_h \end{bmatrix} D_{h-1}^{-1} \qquad \forall h \in [\![2, H-1]\!] \ . \end{aligned}$$

Since $|\mathcal{S}_{max}| = r_{max}$, using Proposition 10, we have that $\mathbf{W}$ is a first-order critical point associated with $\mathcal{S}_{max}$. This concludes the proof. ■

Appendix D. Global Minimizers and Simple Strict Saddle Points (Proof of Proposition 13)

Recall that $r_{max} = \min(d_H, \ldots, d_0)$.

Let $\mathbf{W} = (W_H, \ldots, W_1)$ be a first-order critical point of L associated with $\mathcal{S}$ of size $r = \mathrm{rk}(W_H \cdots W_1) \leq r_{max}$.

Case 1: $\mathcal{S} = [\![1, r_{max}]\!] = \mathcal{S}_{max}$. In this case, using Lemma 21,

$$W_H \cdots W_1 = U_{\mathcal{S}_{max}} U_{\mathcal{S}_{max}}^T \Sigma_{YX} \Sigma_{XX}^{-1} \in \underset{\substack{R \in \mathbb{R}^{d_y \times d_x} \\ \text{rk}(R) \leq r_{max}}}{\arg\min} \|RX - Y\|^2 \ .$$

Moreover, for all $\mathbf{W}' = (W'_H, \ldots, W'_1)$, since $\text{rk}(W'_H \cdots W'_1) \leq r_{max}$, we have

$$L(\mathbf{W}') \geq \min_{\substack{R \in \mathbb{R}^{d_y \times d_x} \\ \text{rk}(R) \leq r_{max}}} \|RX - Y\|^2 = \|W_H \cdots W_1 X - Y\|^2 = L(\mathbf{W}) \ .$$

As a consequence, **W** is a global minimizer of L.

Case 2: In order to prove the two remaining statements, we assume that $\mathcal{S} \neq [\![1, r]\!]$ with $0 < r \leq r_{max}$, and show that **W** is not a second-order critical point.
To do this we will find $\mathbf{W}' = (W'_H, \ldots, W'_1)$ such that $c_2(\mathbf{W}, \mathbf{W}') < 0$ (see Lemma 12). More precisely, we find a linear trajectory of the form $W_h(t) = W_h + tW'_h$ such that the second-order coefficient of the asymptotic expansion of $L((W_h(t))_{h=1..H})$ around $t = 0$ is negative. This proves that **W** is not a second-order critical point.

Since $\mathcal{S} \neq [\![1, r]\!]$, and the eigenvalues $(\lambda_k)_{k \in [\![1, d_y]\!]}$ are distinct and in decreasing order (see Section 2), there exist $j \in \mathcal{S}$ and $i \notin \mathcal{S}$ such that

$$\lambda_i > \lambda_j \ . \tag{45}$$

We denote by $\mathcal{S} = \{i_1, \ldots, i_r\}$, hence there exists $g \in [\![1, r]\!]$ such that $j = i_g$.
Note that,

$$U_{\mathcal{S}} = U \sum_{k=1}^{r} E_{i_k, k}$$

where $E_{l,k} \in \mathbb{R}^{d_y \times r}$ is the matrix whose entries are all 0 except the one in position (l, k) which is equal to 1.
Denote by U_t the matrix formed by replacing in $U_{\mathcal{S}}$ the column corresponding to u_j by $u_j + tu_i$. More precisely, set

$$U_t = U_{\mathcal{S}} + tUE_{i,g} \ .$$

Set $V = UE_{i,g} \in \mathbb{R}^{d_y \times r}$ and

$$V_t = \sum_{k=1}^{r} E_{i_k, k} + tE_{i,g} \in \mathbb{R}^{d_y \times r}. \tag{46}$$

Hence we have

$$U_t = U_{\mathcal{S}} + tV = UV_t \ . \tag{47}$$

Considering $D \in \mathbb{R}^{d_1 \times d_1}$ as provided by Lemma 26, we set

$$\begin{cases} W'_1 = D^{-1} \begin{bmatrix} V^T \Sigma_{YX} \Sigma_{XX}^{-1} \\ 0_{(d_1 - r) \times d_x} \end{bmatrix} \\ W'_h = 0 \qquad \forall h \in [\![2, H-1]\!] \\ W'_H = V U_{\mathcal{S}}^T W_H \ . \end{cases}$$

and for all $h \in [\![1, H]\!]$, $W_h(t) = W_h + tW'_h$. Note that

$$W_H(t) = W_H + tW'_H = (I_{d_y} + tVU_{\mathcal{S}}^T)W_H \ ,$$

and therefore

$$K(t) := W_H(t) \cdots W_2(t) = (I_{d_y} + tVU_{\mathcal{S}}^T)K \ ,$$

where $K = W_H \cdots W_2$. Using Lemma 26, there exists $M \in \mathbb{R}^{d_1 \times d_x}$ satisfying $KM = 0$ such that

$$W_1 = D^{-1} \begin{bmatrix} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} \\ 0_{(d_1-r)\times d_x} \end{bmatrix} + M \ .$$

Hence,

$$W_1(t) = D^{-1} \begin{bmatrix} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} \\ 0_{(d_1-r)\times d_x} \end{bmatrix} + M + tW'_1 = D^{-1} \begin{bmatrix} (U_{\mathcal{S}}^T + tV^T) \Sigma_{YX} \Sigma_{XX}^{-1} \\ 0_{(d_1-r)\times d_x} \end{bmatrix} + M \ ,$$

where $M \in \mathbb{R}^{d_1 \times d_x}$ is such that $KM = 0$. Therefore

$$\begin{aligned} W_t :&= W_H(t) \cdots W_1(t) \\ &= K(t)W_1(t) \\ &= (I_{d_y} + tVU_{\mathcal{S}}^T) \left(KD^{-1} \begin{bmatrix} (U_{\mathcal{S}}^T + tV^T) \Sigma_{YX} \Sigma_{XX}^{-1} \\ 0_{(d_1-r)\times d_x} \end{bmatrix} + KM \right) \ . \end{aligned}$$

From Lemma 26, using that $KM = 0$ and $K = [U_{\mathcal{S}} \quad 0_{d_y \times (d_1-r)}]D$, this becomes

$$\begin{aligned} W_t &= (I_{d_y} + tVU_{\mathcal{S}}^T)[U_{\mathcal{S}} \quad 0_{d_y \times (d_1-r)}]DD^{-1} \begin{bmatrix} (U_{\mathcal{S}}^T + tV^T) \Sigma_{YX} \Sigma_{XX}^{-1} \\ 0_{(d_1-r)\times d_x} \end{bmatrix} \\ &= (I_{d_y} + tVU_{\mathcal{S}}^T) U_{\mathcal{S}} (U_{\mathcal{S}}^T + tV^T) \Sigma_{YX} \Sigma_{XX}^{-1} \ . \end{aligned}$$

Using that $U_{\mathcal{S}}^T U_{\mathcal{S}} = I_r$ (see Lemma 22), we obtain

$$W_t = (U_{\mathcal{S}} + tV)(U_{\mathcal{S}}^T + tV^T) \Sigma_{YX} \Sigma_{XX}^{-1} = U_t U_t^T \Sigma_{YX} \Sigma_{XX}^{-1} \ . \tag{48}$$

Recall that our goal is to show that the asymptotic expansion of (49) around $t = 0$ has a negative second-order coefficient. We calculate

$$\begin{aligned} L((W_h(t))_{h=1..H}) &= \|W_t X - Y\|^2 \\ &= \operatorname{tr}\left(W_t \Sigma_{XX} W_t^T\right) - 2\operatorname{tr}(W_t \Sigma_{XY}) + \operatorname{tr}(\Sigma_{YY}) \ . \end{aligned} \tag{49}$$

Let us simplify $\operatorname{tr}(W_t \Sigma_{XX} W_t^T)$ first. Using (48), we have

$$W_t \Sigma_{XX} W_t^T = U_t U_t^T \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XX} \Sigma_{XX}^{-1} \Sigma_{XY} U_t U_t^T = U_t U_t^T \Sigma U_t U_t^T \ .$$

Using (47), $U^T U = I_{d_y}$, $\Sigma = U \Lambda U^T$ and the cyclic property of the trace, we obtain

$$\operatorname{tr}\left(W_t \Sigma_{XX} W_t^T\right) = \operatorname{tr}\left(UV_t V_t^T U^T U \Lambda U^T U V_t V_t^T U^T\right) = \operatorname{tr}\left(V_t V_t^T \Lambda V_t V_t^T\right) = \operatorname{tr}\left(\left(V_t V_t^T\right)^2 \Lambda\right) \ .$$

We define $(\overline{E}_{k,l})_{k=1..d_y,l=1..d_y}$ the canonical basis of $\mathbb{R}^{d_y \times d_y}$. More precisely, $\overline{E}_{k,l} \in \mathbb{R}^{d_y \times d_y}$ has all its entries equal to 0, except a 1 at position (k,l). Note that for all $a, c \in [\![1, d_y]\!]$ and $b, d \in [\![1, r]\!]$

$$E_{a,b} E_{c,d}^T = \delta_{b,d} \overline{E}_{a,c} \ ,$$

where $\delta_{b,d}$ equals 1 if $b = d$ and 0 otherwise. Using the definition of V_t in (46) and $j = i_g$, for $g \in [\![1, r]\!]$, we have

$$\begin{aligned} V_t V_t^T &= \left(\sum_{k=1}^{r} E_{i_k,k} + t E_{i,g} \right) \left(\sum_{k'=1}^{r} E_{i_{k'},k'}^T + t E_{i,g}^T \right) \\ &= \left(\sum_{k=1}^{r} \overline{E}_{i_k,i_k} \right) + t \overline{E}_{i_g,i} + t \overline{E}_{i,i_g} + t^2 \overline{E}_{i,i} \\ &= \left(\sum_{k \in \mathcal{S}} \overline{E}_{k,k} \right) + t \overline{E}_{j,i} + t \overline{E}_{i,j} + t^2 \overline{E}_{i,i} \ . \end{aligned} \tag{50}$$

We also have for all $a, b, c, d \in [\![1, d_y]\!]$

$$\overline{E}_{a,b} \overline{E}_{c,d} = \delta_{b,c} \overline{E}_{a,d} \ .$$

Recalling that $j \in \mathcal{S}$ and $i \notin \mathcal{S}$, we obtain

$$\begin{aligned} (V_t V_t^T)^2 &= \left(\left(\sum_{k \in \mathcal{S}} \overline{E}_{k,k} \right) + t \overline{E}_{j,i} + t \overline{E}_{i,j} + t^2 \overline{E}_{i,i} \right) \left(\left(\sum_{k' \in \mathcal{S}} \overline{E}_{k',k'} \right) + t \overline{E}_{j,i} + t \overline{E}_{i,j} + t^2 \overline{E}_{i,i} \right) \\ &= \left(\left(\sum_{k \in \mathcal{S}} \overline{E}_{k,k} \right) + t \overline{E}_{j,i} + 0 + 0 \right) + (0 + 0 + t^2 \overline{E}_{j,j} + t^3 \overline{E}_{j,i}) \\ &\quad + (t \overline{E}_{i,j} + t^2 \overline{E}_{i,i} + 0 + 0) + (0 + 0 + t^3 \overline{E}_{i,j} + t^4 \overline{E}_{i,i}) \\ &= \left(\sum_{k \in \mathcal{S}} \overline{E}_{k,k} \right) + t^2(1+t^2) \overline{E}_{i,i} + t^2 \overline{E}_{j,j} + t(1+t^2) \overline{E}_{i,j} + t(1+t^2) \overline{E}_{j,i} \ . \end{aligned}$$

Finally, since for all $a, b \in [\![1, d_y]\!]$

$$\overline{E}_{a,b} \Lambda = \lambda_b \overline{E}_{a,b} \tag{51}$$

we have

$$\mathrm{tr}\left(W_t \Sigma_{XX} W_t^T \right) = \mathrm{tr}\left(\left(V_t V_t^T \right)^2 \Lambda \right) = \sum_{k \in \mathcal{S}} \lambda_k + t^2(1+t^2) \lambda_i + t^2 \lambda_j \ . \tag{52}$$

Coming back to (49), we calculate the other term $\mathrm{tr}(W_t \Sigma_{XY})$. Using (48), (47) and $\Sigma = U \Lambda U^T$, we obtain

$$\mathrm{tr}(W_t \Sigma_{XY}) = \mathrm{tr}(U_t U_t^T \Sigma) = \mathrm{tr}(U V_t V_t^T U^T U \Lambda U^T) = \mathrm{tr}(V_t V_t^T \Lambda) \ .$$

Combining with (50) and (51), we get

$$\mathrm{tr}(W_t\Sigma_{XY}) = \mathrm{tr}(V_tV_t^T\Lambda) = \sum_{k\in\mathcal{S}}\lambda_k + t^2\lambda_i \ . \tag{53}$$

Finally, substituting (52) and (53) in (49), we have

$$\begin{aligned} L((W_h(t))_{h=1..H}) &= \mathrm{tr}(\Sigma_{YY}) + \sum_{k\in\mathcal{S}}\lambda_k + t^2(1+t^2)\lambda_i + t^2\lambda_j - 2\sum_{k\in\mathcal{S}}\lambda_k - 2t^2\lambda_i \\ &= \mathrm{tr}(\Sigma_{YY}) - \sum_{k\in\mathcal{S}}\lambda_k + t^2(\lambda_j-\lambda_i) + \lambda_i t^4 \ . \end{aligned}$$

Using Proposition 1 and recalling (45), we finally get as $t \to 0$,

$$L((W_h(t))_{h=1..H}) = L(\mathbf{W}) + ct^2 + o(t^2) \quad \text{with} \quad c = \lambda_j - \lambda_i < 0 \ .$$

Therefore, we conclude from Lemma 12 that $\mathbf{W} = (W_H, \dots, W_1)$ is not a second-order critical point.

Appendix E. Strict Saddle Points with $\mathcal{S} = [\![1, r]\!],\ r < r_{max}$ (Proof of Proposition 14)

We refer the reader to Section 4.2, which introduces the 4 cases proved below. Recall that $\mathcal{S} = [\![1, r]\!]$ and we set $Q = [\![1, d_y]\!] \setminus \mathcal{S} = [\![r+1, d_y]\!]$.
In this section, for each vector space $\mathbb{R}^{d_h}$, we will denote by e_m the m-th element of the canonical basis of $\mathbb{R}^{d_h}$. That is, the entries of $e_m \in \mathbb{R}^{d_h}$ are all equal to 0 except for the m-th coordinate which is equal to 1. The size of e_m will not be ambiguous, once in context, so we do not include it in the notation.
Remark about $r = 0$: Using the conventions of Section 2, in this case we have $\mathcal{S} = \emptyset$ and $Q = [\![1, d_y]\!]$. Hence $U_\mathcal{S}$ is the matrix with no column, $U_Q = U$, and $U_\mathcal{S}U_\mathcal{S}^T = 0_{d_y\times d_y}$. For example, we still have $I_{d_y} = U_\mathcal{S}U_\mathcal{S}^T + U_QU_Q^T$. We can easily follow the proofs below with these conventions and see that the result still holds.

E.1 1st Case: $i \in [\![2, H-1]\!]$ and $j = 1$

In this case, the two complementary blocks are $\Sigma_{XY}W_H\cdots W_{i+1}$ and $W_{i-1}\cdots W_2$. Recall that $\mathcal{S} = [\![1, r]\!]$ and $r < r_{max} = \min(d_H, \dots, d_0)$. Note that $\mathrm{rk}(\Sigma_{XY}W_H\cdots W_{i+1}) = \mathrm{rk}(W_H\cdots W_{i+1})$ because Σ_{XY} is of full column rank (see Assumption 1, in Section 2) .
Since the pivot (i, j) is not tightened, using Proposition 4, we have

$$\begin{cases} \mathrm{rk}(W_H\cdots W_{i+1}) > r \\ \mathrm{rk}(W_{i-1}\cdots W_2) > r. \end{cases} \tag{54}$$

Let us first show that there exists $k \in [\![r+1, d_y]\!]$ and $l \in [\![1, d_i]\!]$ such that

$$U_k^T(W_H\cdots W_{i+1})_{.,l} \neq 0 \ . \tag{55}$$

Indeed, assume by contradiction that for all $k \in [\![r+1, d_y]\!]$ and $l \in [\![1, d_i]\!]$ we have

$$U_k^T(W_H\cdots W_{i+1})_{.,l} = 0.$$

Recalling that $Q = [\![1, d_y]\!] \setminus \mathcal{S} = [\![r+1, d_y]\!]$, we obtain $U_Q^T W_H \cdots W_{i+1} = 0$. Using from Lemma 22 that $I_{d_y} = U_\mathcal{S} U_\mathcal{S}^T + U_Q U_Q^T$, we have

$$\begin{aligned} W_H \cdots W_{i+1} &= (U_\mathcal{S} U_\mathcal{S}^T + U_Q U_Q^T) W_H \cdots W_{i+1} \\ &= U_\mathcal{S} U_\mathcal{S}^T W_H \cdots W_{i+1}. \end{aligned}$$

Therefore,

$$\mathrm{rk}(W_H \cdots W_{i+1}) = \mathrm{rk}(U_\mathcal{S} U_\mathcal{S}^T W_H \cdots W_{i+1}).$$

The latter is impossible since $\mathrm{rk}(U_\mathcal{S} U_\mathcal{S}^T W_H \cdots W_{i+1}) \leq |\mathcal{S}| = r$, which is not compatible with (54). Therefore (55) holds.

Since **W** is a first-order critical point, using Lemma 26, there exists an invertible matrix $D \in \mathbb{R}^{d_1 \times d_1}$ such that

$$W_H \cdots W_2 = [U_\mathcal{S}, 0_{d_y \times (d_1 - r)}] D \tag{56}$$

and since **W** is associated with $\mathcal{S}$, we have

$$W_H \cdots W_1 = U_\mathcal{S} U_\mathcal{S}^T \Sigma_{YX} \Sigma_{XX}^{-1} \ . \tag{57}$$

Using (54) and D invertible, we have $rk(W_{i-1} \cdots W_2 D^{-1}) = rk(W_{i-1} \cdots W_2) > r$. Hence there exists $g \in [\![r+1, d_1]\!]$ such that

$$(W_{i-1} \cdots W_2 D^{-1})_{.,g} \neq 0 \ .$$

Therefore, there exists $a \in \mathbb{R}^{d_{i-1}}$ such that

$$a^T (W_{i-1} \cdots W_2 D^{-1})_{.,g} = 1 \ . \tag{58}$$

Recall that k, l satisfy (55). We define $\mathbf{W}'_\beta = (W_H'^\beta, \ldots, W_1'^\beta)$ by

$$\begin{cases} W_i'^\beta = \beta W_i' = \beta e_l a^T \in \mathbb{R}^{d_i \times d_{i-1}}, \text{ where } e_l \in \mathbb{R}^{d_i} \\ W_1'^\beta = W_1' = D^{-1} e_g U_k^T \Sigma_{YX} \Sigma_{XX}^{-1} \in \mathbb{R}^{d_1 \times d_x}, \text{ where } e_g \in \mathbb{R}^{d_1} \\ W_h'^\beta = 0 \quad \forall h \in [\![2, H]\!] \setminus \{i\} \end{cases}$$

We set $\mathbf{W}^\beta(t) = (W_H^\beta(t), \ldots, W_1^\beta(t))$ such that $W_h^\beta(t) = W_h + t W_h'^\beta$ for $h \in [\![1, H]\!]$. We have

$$\begin{aligned} W^\beta(t) :&= W_H^\beta(t) \cdots W_1^\beta(t) \\ &= W_H \cdots W_{i+1} (W_i + t\beta W_i') W_{i-1} \cdots W_2 (W_1 + t W_1') \\ &= W_H \cdots W_1 + t(\beta W_H \cdots W_{i+1} W_i' W_{i-1} \cdots W_1 + W_H \cdots W_2 W_1') \\ &\quad + \beta t^2 W_H \cdots W_{i+1} W_i' W_{i-1} \cdots W_2 W_1' \ . \end{aligned}$$

Using (56) and (57), we obtain

$$\begin{aligned} W^\beta(t) &= U_\mathcal{S} U_\mathcal{S}^T \Sigma_{YX} \Sigma_{XX}^{-1} + t(\beta W_H \cdots W_{i+1} W_i' W_{i-1} \cdots W_1 + [U_\mathcal{S}, 0] D D^{-1} e_g U_k^T \Sigma_{YX} \Sigma_{XX}^{-1}) \\ &\quad + \beta t^2 (W_H \cdots W_{i+1})_{.,l} a^T (W_{i-1} \cdots W_2 D^{-1})_{.,g} U_k^T \Sigma_{YX} \Sigma_{XX}^{-1} \ . \end{aligned}$$

Using (58) and $g \in [\![r+1, d_1]\!]$, we have

$$W^\beta(t) = U_\mathcal{S} U_\mathcal{S}^T \Sigma_{YX}\Sigma_{XX}^{-1} + t\beta W_H \cdots W_{i+1} W_i' W_{i-1} \cdots W_1 + \beta t^2 (W_H \cdots W_{i+1})_{.,l} U_k^T \Sigma_{YX}\Sigma_{XX}^{-1} .$$

Denoting $N = W_H \cdots W_{i+1} W_i' W_{i-1} \cdots W_1$, we have

$$\begin{aligned} L(\mathbf{W}^\beta(t)) &= \|W^\beta(t)X - Y\|^2 \\ &= \|U_\mathcal{S} U_\mathcal{S}^T \Sigma_{YX}\Sigma_{XX}^{-1} X - Y + t\beta N X + \beta t^2 (W_H \cdots W_{i+1})_{.,l} U_k^T \Sigma_{YX}\Sigma_{XX}^{-1} X\|^2 . \end{aligned}$$

Expanding the square, the second-order term $c_2(\mathbf{W}, \mathbf{W}'_\beta)t^2$ has a coefficient equal to

$$\begin{aligned} c_2(\mathbf{W}, \mathbf{W}'_\beta) &= \beta^2 \|NX\|^2 + 2\beta \operatorname{tr}((W_H \cdots W_{i+1})_{.,l} U_k^T \Sigma_{YX}\Sigma_{XX}^{-1} X X^T \Sigma_{XX}^{-1} \Sigma_{XY} U_\mathcal{S} U_\mathcal{S}^T) \\ &\quad - 2\beta \operatorname{tr}((W_H \cdots W_{i+1})_{.,l} U_k^T \Sigma_{YX}\Sigma_{XX}^{-1} X Y^T) \\ &= \beta^2 \|NX\|^2 + 2\beta \operatorname{tr}((W_H \cdots W_{i+1})_{.,l} U_k^T \Sigma U_\mathcal{S} U_\mathcal{S}^T) - 2\beta \operatorname{tr}((W_H \cdots W_{i+1})_{.,l} U_k^T \Sigma) \\ &= \beta^2 \|NX\|^2 - 2\beta \lambda_k U_k^T (W_H \cdots W_{i+1})_{.,l} , \end{aligned}$$

where the last equality follows from Lemma 23 and $k \notin \mathcal{S}$, and $U^T\Sigma = \Lambda U^T$ and the cyclic property of the trace.
Using Lemma 20 and (55), we have $\lambda_k U_k^T (W_H \cdots W_{i+1})_{.,l} \neq 0$, hence we can choose β according to (7), such that $c_2(\mathbf{W}, \mathbf{W}'_\beta) < 0$. Therefore, **W** is not a second-order critical point.

E.2 2nd Case: $i = H$ and $j = 1$

In this case, the two complementary blocks are Σ_{XY} and $W_{H-1} \cdots W_2$. We follow again the same lines as above. Since the pivot (i, j) is not tightened, using Proposition 4, we have

$$\operatorname{rk}(W_{H-1} \cdots W_2) > r . \tag{59}$$

Again, since **W** is a first-order critical point, using Lemma 26, there exists an invertible matrix $D \in \mathbb{R}^{d_1 \times d_1}$ such that

$$W_H \cdots W_2 = [U_\mathcal{S}, 0_{d_y \times (d_1 - r)}]D \tag{60}$$

and since **W** is associated with $\mathcal{S}$, we have

$$W_H \cdots W_1 = U_\mathcal{S} U_\mathcal{S}^T \Sigma_{YX}\Sigma_{XX}^{-1} . \tag{61}$$

Using (59) and D invertible, we have $rk(W_{H-1} \cdots W_2 D^{-1}) = rk(W_{H-1} \cdots W_2) > r$. Hence there exists $g \in [\![r+1, d_1]\!]$ such that

$$(W_{i-1} \cdots W_2 D^{-1})_{.,g} \neq 0 .$$

Therefore, there exists $a \in \mathbb{R}^{d_{H-1}}$ such that

$$a^T (W_{H-1} \cdots W_2 D^{-1})_{.,g} = 1 . \tag{62}$$

We define $\mathbf{W}'_\beta = (W'^\beta_H, \dots, W'^\beta_1)$ by

$$\begin{cases} W'^\beta_H = \beta W'_H = \beta U_{r+1} a^T \in \mathbb{R}^{d_y \times d_{H-1}} \\ W'^\beta_1 = W'_1 = D^{-1} e_g U_{r+1}^T \Sigma_{YX} \Sigma_{XX}^{-1} \in \mathbb{R}^{d_1 \times d_x}, \text{ where } e_g \in \mathbb{R}^{d_1} \\ W'^\beta_h = 0 \quad \forall h \in [\![2, H-1]\!] \ . \end{cases}$$

We set $\mathbf{W}^\beta(t) = (W^\beta_H(t), \dots, W^\beta_1(t))$ such that $W^\beta_h(t) = W_h + tW'^\beta_h$, for all $h \in [\![1, H]\!]$. We have

$$\begin{aligned} W^\beta(t) :&= W^\beta_H(t) \cdots W^\beta_1(t) \\ &= (W_H + t\beta W'_H) W_{H-1} \cdots W_2 (W_1 + tW'_1) \\ &= W_H \cdots W_1 + t(\beta W'_H W_{H-1} \cdots W_1 + W_H \cdots W_2 W'_1) + \beta t^2 W'_H W_{H-1} \cdots W_2 W'_1 \ . \end{aligned}$$

Using (60) and (61), then (62) and $g \in [\![r+1, d_1]\!]$, we obtain

$$\begin{aligned} W^\beta(t) &= U_\mathcal{S} U_\mathcal{S}^T \Sigma_{YX} \Sigma_{XX}^{-1} + t(\beta W'_H W_{H-1} \cdots W_1 + [U_\mathcal{S}, 0] D D^{-1} e_g U_{r+1}^T \Sigma_{YX} \Sigma_{XX}^{-1}) \\ &\quad + \beta t^2 U_{r+1} a^T (W_{H-1} \cdots W_2 D^{-1})_{.,g} U_{r+1}^T \Sigma_{YX} \Sigma_{XX}^{-1} \\ &= U_\mathcal{S} U_\mathcal{S}^T \Sigma_{YX} \Sigma_{XX}^{-1} + t\beta W'_H W_{H-1} \cdots W_1 + \beta t^2 U_{r+1} U_{r+1}^T \Sigma_{YX} \Sigma_{XX}^{-1} \ . \end{aligned}$$

Denoting by $N = W'_H W_{H-1} \cdots W_1$, we have

$$\begin{aligned} L(\mathbf{W}^\beta(t)) &= \|W^\beta(t) X - Y\|^2 \\ &= \|U_\mathcal{S} U_\mathcal{S}^T \Sigma_{YX} \Sigma_{XX}^{-1} X - Y + t\beta N X + \beta t^2 U_{r+1} U_{r+1}^T \Sigma_{YX} \Sigma_{XX}^{-1} X\|^2 \ . \end{aligned}$$

As previously, expanding the square, we can see that the second-order coefficient $c_2(\mathbf{W}, \mathbf{W}'_\beta)$ of the polynomial $L(\mathbf{W}^\beta(t))$ is given by

$$\begin{aligned} c_2(\mathbf{W}, \mathbf{W}'_\beta) &= \beta^2 \|NX\|^2 + 2\beta \operatorname{tr}(U_{r+1} U_{r+1}^T \Sigma_{YX} \Sigma_{XX}^{-1} X X^T \Sigma_{XX}^{-1} \Sigma_{XY} U_\mathcal{S} U_\mathcal{S}^T) \\ &\quad - 2\beta \operatorname{tr}(U_{r+1} U_{r+1}^T \Sigma_{YX} \Sigma_{XX}^{-1} X Y^T) \\ &= \beta^2 \|NX\|^2 + 2\beta \operatorname{tr}(U_{r+1} U_{r+1}^T \Sigma U_\mathcal{S} U_\mathcal{S}^T) - 2\beta \operatorname{tr}(U_{r+1} U_{r+1}^T \Sigma) \ . \end{aligned}$$

Using the cyclic property of the trace, $U_\mathcal{S}^T U_{r+1} = 0$ (see Lemma 22), and $\Sigma U_{r+1} = \lambda_{r+1} U_{r+1}$, we obtain

$$\begin{aligned} c_2(\mathbf{W}, \mathbf{W}'_\beta) &= \beta^2 \|NX\|^2 - 2\beta \lambda_{r+1} U_{r+1}^T U_{r+1} \\ &= \beta^2 \|NX\|^2 - 2\beta \lambda_{r+1} \ . \end{aligned}$$

Using Lemma 20, we have $\lambda_{r+1} \neq 0$, hence we can choose β according to (7) such that $c_2(\mathbf{W}, \mathbf{W}'_\beta) < 0$. Therefore $\mathbf{W}$ is not a second-order critical point.

E.3 3rd Case: $i = H$ and $j \in [\![2, H-1]\!]$

In this case, the two complementary blocks are $W_{j-1} \cdots W_1 \Sigma_{XY}$ and $W_{H-1} \cdots W_{j+1}$. We follow again the same lines as above. Since the pivot (i, j) is not tightened, using Proposition 4, we have

$$\begin{cases} \operatorname{rk}(W_{H-1} \cdots W_{j+1}) > r \\ \operatorname{rk}(W_{j-1} \cdots W_1 \Sigma_{XY}) > r \ . \end{cases} \tag{63}$$

Let us first show that there exist $k \in [\![r+1, d_y]\!]$ and $l \in [\![1, d_{j-1}]\!]$ such that

$$(W_{j-1} \cdots W_1)_{l,.} \Sigma_{XY} U_k \neq 0 \; . \tag{64}$$

Indeed, assume by contradiction that for all $k \in [\![r+1, d_y]\!]$ and $l \in [\![1, d_{j-1}]\!]$ we have

$$(W_{j-1} \cdots W_1)_{l,.} \Sigma_{XY} U_k = 0.$$

Recalling that $Q = [\![1, d_y]\!] \setminus \mathcal{S} = [\![r+1, d_y]\!]$, we obtain $W_{j-1} \cdots W_1 \Sigma_{XY} U_Q = 0$, and using, from Lemma 22, that $I_{d_y} = U_{\mathcal{S}} U_{\mathcal{S}}^T + U_Q U_Q^T$, we have

$$\begin{aligned} W_{j-1} \cdots W_1 \Sigma_{XY} &= W_{j-1} \cdots W_1 \Sigma_{XY} (U_{\mathcal{S}} U_{\mathcal{S}}^T + U_Q U_Q^T) \\ &= W_{j-1} \cdots W_1 \Sigma_{XY} U_{\mathcal{S}} U_{\mathcal{S}}^T \; . \end{aligned}$$

Therefore,

$$\mathrm{rk}(W_{j-1} \cdots W_1 \Sigma_{XY}) = \mathrm{rk}(W_{j-1} \cdots W_1 \Sigma_{XY} U_{\mathcal{S}} U_{\mathcal{S}}^T).$$

The latter is impossible since $\mathrm{rk}(W_{j-1} \cdots W_1 \Sigma_{XY} U_{\mathcal{S}} U_{\mathcal{S}}^T) \leq |\mathcal{S}| = r$ is not compatible with (63). Therefore (64) holds.

We know that $\mathrm{rk}(W_H \cdots W_{j+1}) \geq \mathrm{rk}(W_H \cdots W_1) = r$. Therefore, depending on the value of $\mathrm{rk}(W_H \cdots W_{j+1})$, we distinguish two situations: either $\mathrm{rk}(W_H \cdots W_{j+1}) > r$ or $\mathrm{rk}(W_H \cdots W_{j+1}) = r$.

When $\mathrm{rk}(W_H \cdots W_{j+1}) > r$, since Σ_{XY} is of full column rank, we have $\mathrm{rk}(\Sigma_{XY} W_H \cdots W_{j+1}) = \mathrm{rk}(W_H \cdots W_{j+1}) > r$. Also, using (63), we have $\mathrm{rk}(W_{j-1} \cdots W_2) \geq \mathrm{rk}(W_{j-1} \cdots W_1 \Sigma_{XY}) > r$. Hence, in this case, the pivot $(j, 1)$ is not tightened either. We have already proved in Section E.1 (beware that the pivot is denoted $(i, 1)$, not $(j, 1)$, in Section E.1) that, when such a pivot is not tightened, **W** is not a second-order critical point. This concludes the proof in the case $\mathrm{rk}(W_H \cdots W_{j+1}) > r$.

In the rest of the section we assume that $\mathrm{rk}(W_H \cdots W_{j+1}) = r$.

Using (63), we have $\mathrm{rk}(W_{H-1} \cdots W_{j+1}) > r = \mathrm{rk}(W_H \cdots W_{j+1})$. Applying the rank-nullity theorem we obtain

$$\mathrm{Ker}(W_{H-1} \cdots W_{j+1}) \subsetneq \mathrm{Ker}(W_H \cdots W_{j+1}).$$

Therefore there exists $b \in \mathbb{R}^{d_j}$ such that

$$\begin{cases} b \in \mathrm{Ker}(W_H \cdots W_{j+1}) \\ b \notin \mathrm{Ker}(W_{H-1} \cdots W_{j+1}) \; . \end{cases} \tag{65}$$

Hence, there also exists $a \in \mathbb{R}^{d_{H-1}}$ such that

$$a^T W_{H-1} \cdots W_{j+1} b = 1 \; . \tag{66}$$

Recall that k, l satisfy (64). We define $\mathbf{W}'_\beta = (W'^{\beta}_H, \ldots, W'^{\beta}_1)$ by

$$\begin{cases} W'^{\beta}_H = \beta W'_H = \beta U_k a^T \in \mathbb{R}^{d_y \times d_{H-1}} \\ W'^{\beta}_j = W'_j = b e_l^T \in \mathbb{R}^{d_j \times d_{j-1}}, \text{ where } e_l \in \mathbb{R}^{d_{j-1}} \\ W'^{\beta}_h = 0 \quad \forall h \in [\![1, H]\!] \setminus \{i, j\} \end{cases}$$

We set $\mathbf{W}^\beta(t) = (W_H^\beta(t), \dots, W_1^\beta(t))$ such that $W_h^\beta(t) = W_h + tW_h'^\beta$ for $h \in [\![1, H]\!]$. We have

$$\begin{aligned}
W^\beta(t) :&= W_H^\beta(t) \cdots W_1^\beta(t) \\
&= (W_H + t\beta W_H')W_{H-1} \cdots W_{j+1}(W_j + tW_j')W_{j-1} \cdots W_1 \\
&= W_H \cdots W_1 + t(\beta W_H' W_{H-1} \cdots W_1 + W_H \cdots W_{j+1} W_j' W_{j-1} \cdots W_1) \\
&\quad + t^2 \beta W_H' \cdots W_{j+1} W_j' W_{j-1} \cdots W_1 \ .
\end{aligned}$$

Using Proposition 1 and the definition of $\mathbf{W}'_\beta$ above, we obtain

$$\begin{aligned}
W^\beta(t) &= U_\mathcal{S} U_\mathcal{S}^T \Sigma_{YX} \Sigma_{XX}^{-1} + t(\beta W_H' W_{H-1} \cdots W_1 + W_H \cdots W_{j+1} b e_l^T W_{j-1} \cdots W_1) \\
&\quad + \beta t^2 U_k a^T W_{H-1} \cdots W_{j+1} b (W_{j-1} \cdots W_1)_{l,.} \\
&= U_\mathcal{S} U_\mathcal{S}^T \Sigma_{YX} \Sigma_{XX}^{-1} + t\beta W_H' W_{H-1} \cdots W_1 + \beta t^2 U_k (W_{j-1} \cdots W_1)_{l,.} \ ,
\end{aligned}$$

where the last equality follows from (65) and (66) .
Denoting $N = W_H' W_{H-1} \cdots W_1$, we have

$$\begin{aligned}
L(\mathbf{W}^\beta(t)) &= \|W^\beta(t)X - Y\|^2 \\
&= \|U_\mathcal{S} U_\mathcal{S}^T \Sigma_{YX} \Sigma_{XX}^{-1} X - Y + t\beta N X + \beta t^2 U_k (W_{j-1} \cdots W_1)_{l,.} X\|^2 \ .
\end{aligned}$$

Using the cyclic property of the trace, and, since $k \notin \mathcal{S}$, $U_\mathcal{S}^T U_k = 0$, we get in this case a second-order coefficient equal to

$$\begin{aligned}
c_2(\mathbf{W}, \mathbf{W}'_\beta) &= \beta^2 \|NX\|^2 + 2\beta \operatorname{tr}\left(U_k (W_{j-1} \cdots W_1)_{l,.} X X^T \Sigma_{XX}^{-1} \Sigma_{XY} U_\mathcal{S} U_\mathcal{S}^T\right) \\
&\quad - 2\beta \operatorname{tr}(U_k (W_{j-1} \cdots W_1)_{l,.} \Sigma_{XY}) \\
&= \beta^2 \|NX\|^2 - 2\beta (W_{j-1} \cdots W_1)_{l,.} \Sigma_{XY} U_k \ .
\end{aligned}$$

Since from (64), $(W_{j-1} \cdots W_1)_{l,.} \Sigma_{XY} U_k \neq 0$, we can choose β according to (7), such that $c_2(\mathbf{W}, \mathbf{W}'_\beta) < 0$. Therefore **W** is not a second-order critical point.

E.4 4th Case: $i, j \in [\![2, H-1]\!]$, with $i > j$

In this case, the two complementary blocks are $W_{j-1} \cdots W_1 \Sigma_{XY} W_H \cdots W_{i+1}$ and $W_{i-1} \cdots W_{j+1}$. We follow again the same lines as above. Since the pivot (i, j) is not tightened, using Proposition 4, we have

$$\begin{cases} \operatorname{rk}(W_{i-1} \cdots W_{j+1}) > r \\ \operatorname{rk}(W_{j-1} \cdots W_1 \Sigma_{XY} W_H \cdots W_{i+1}) > r \ . \end{cases} \tag{67}$$

Let us first show that there exist $k \in [\![1, d_i]\!]$ and $l \in [\![1, d_{j-1}]\!]$ such that

$$(W_{j-1} \cdots W_1)_{l,.} \Sigma_{XY} U_Q U_Q^T (W_H \cdots W_{i+1})_{.,k} \neq 0 \ . \tag{68}$$

Indeed, assume by contradiction that, for all $k \in [\![1, d_i]\!]$ and $l \in [\![1, d_{j-1}]\!]$, we have

$$(W_{j-1} \cdots W_1)_{l,.} \Sigma_{XY} U_Q U_Q^T (W_H \cdots W_{i+1})_{.,k} = 0.$$

Then $W_{j-1}\cdots W_1\Sigma_{XY}U_QU_Q^TW_H\cdots W_{i+1}=0$, and so, using $I_{d_y}=U_\mathcal{S}U_\mathcal{S}^T+U_QU_Q^T$, we would have

$$\begin{aligned}W_{j-1}\cdots W_1\Sigma_{XY}W_H\cdots W_{i+1}&=W_{j-1}\cdots W_1\Sigma_{XY}I_{d_y}W_H\cdots W_{i+1}\\&=W_{j-1}\cdots W_1\Sigma_{XY}\left(U_\mathcal{S}U_\mathcal{S}^T+U_QU_Q^T\right)W_H\cdots W_{i+1}\\&=W_{j-1}\cdots W_1\Sigma_{XY}U_\mathcal{S}U_\mathcal{S}^TW_H\cdots W_{i+1}\ .\end{aligned}$$

Therefore,

$$\mathrm{rk}(W_{j-1}\cdots W_1\Sigma_{XY}W_H\cdots W_{i+1})=\mathrm{rk}(W_{j-1}\cdots W_1\Sigma_{XY}U_\mathcal{S}U_\mathcal{S}^TW_H\cdots W_{i+1}).$$

The latter is impossible since $\mathrm{rk}(W_{j-1}\cdots W_1\Sigma_{XY}U_\mathcal{S}U_\mathcal{S}^TW_H\cdots W_{i+1})\leq|\mathcal{S}|=r$ is not compatible with (67). Therefore (68) holds.

We know that $\mathrm{rk}(W_H\cdots W_{j+1})\geq\mathrm{rk}(W_H\cdots W_1)=r$. Therefore, depending on the value of $\mathrm{rk}(W_H\cdots W_{j+1})$, we distinguish two situations: either $\mathrm{rk}(W_H\cdots W_{j+1})>r$ or $\mathrm{rk}(W_H\cdots W_{j+1})=r$.
When $\mathrm{rk}(W_H\cdots W_{j+1})>r$, since Σ_{XY} is of full column rank, we have $\mathrm{rk}(\Sigma_{XY}W_H\cdots W_{j+1})=\mathrm{rk}(W_H\cdots W_{j+1})>r$. Also, using (67), we have $\mathrm{rk}(W_{j-1}\cdots W_2)\geq\mathrm{rk}(W_{j-1}\cdots W_1\Sigma_{XY}W_H\cdots W_{i+1})>r$. Hence, in this case, the pivot $(j,1)$ is not tightened either. We have already proved in Section E.1 (beware that the pivot is denoted $(i,1)$, not $(j,1)$, in Section E.1) that, when such a pivot is not tightened, **W** is not a second-order critical point. This concludes the proof when $\mathrm{rk}(W_H\cdots W_{j+1})>r$.
In the rest of the section we assume that $\mathrm{rk}(W_H\cdots W_{j+1})=r$.
Using (67), we have $\mathrm{rk}(W_{i-1}\cdots W_{j+1})>r=\mathrm{rk}(W_H\cdots W_{j+1})$. Applying the rank-nullity theorem, we obtain

$$\mathrm{Ker}(W_{i-1}\cdots W_{j+1})\subsetneq\mathrm{Ker}(W_H\cdots W_{j+1}).$$

Therefore there exists $b\in\mathbb{R}^{d_j}$ such that

$$\begin{cases}b\in\mathrm{Ker}(W_H\cdots W_{j+1})\\b\notin\mathrm{Ker}(W_{i-1}\cdots W_{j+1})\ .\end{cases}\tag{69}$$

Hence, there also exists $a\in\mathbb{R}^{d_{i-1}}$ such that

$$a^TW_{i-1}\cdots W_{j+1}b=1\ .\tag{70}$$

Recall that k,l satisfy (68). We define $\mathbf{W}'_\beta=(W_H'^\beta,\ldots,W_1'^\beta)$ by

$$\begin{cases}W_i'^\beta=\beta W_i'=\beta e_ka^T\in\mathbb{R}^{d_i\times d_{i-1}}\text{ where }e_k\in\mathbb{R}^{d_i}\\W_j'^\beta=W_j'=be_l^T\in\mathbb{R}^{d_j\times d_{j-1}}\text{ where }e_l\in\mathbb{R}^{d_{j-1}}\\W_h'^\beta=0\quad\forall h\in[\![1,H]\!]\setminus\{i,j\}\ .\end{cases}$$

We set $\mathbf{W}^\beta(t)=(W_H^\beta(t),\ldots,W_1^\beta(t))$ with $W_h^\beta(t)=W_h+tW_h'^\beta$ for all $h\in[\![1,H]\!]$. We have,

$$\begin{aligned}W^\beta(t):&=W_H^\beta(t)\cdots W_1^\beta(t)\\&=W_H\cdots W_{i+1}(W_i+t\beta W_i')W_{i-1}\cdots W_{j+1}(W_j+tW_j')W_{j-1}\cdots W_1\\&=W_H\cdots W_1+t(\beta W_H\cdots W_{i+1}W_i'W_{i-1}\cdots W_1+W_H\cdots W_{j+1}W_j'W_{j-1}\cdots W_1)\\&\quad+\beta t^2W_H\cdots W_{i+1}W_i'W_{i-1}\cdots W_{j+1}W_j'W_{j-1}\cdots W_1\ .\end{aligned}$$

Using Proposition 1 and the definition of $\mathbf{W}'_\beta$ above, we obtain

$$\begin{aligned} W^\beta(t) &= U_\mathcal{S}U_\mathcal{S}^T\Sigma_{YX}\Sigma_{XX}^{-1} + t(\beta W_H\cdots W_{i+1}W_i'W_{i-1}\cdots W_1 + W_H\cdots W_{j+1}be_l^TW_{j-1}\cdots W_1) \\ &\quad + \beta t^2(W_H\cdots W_{i+1})_{.,k}a^TW_{i-1}\cdots W_{j+1}b(W_{j-1}\cdots W_1)_{l,.} \\ &= U_\mathcal{S}U_\mathcal{S}^T\Sigma_{YX}\Sigma_{XX}^{-1} + t\beta W_H\cdots W_{i+1}W_i'W_{i-1}\cdots W_1 + \beta t^2(W_H\cdots W_{i+1})_{.,k}(W_{j-1}\cdots W_1)_{l,.}\ , \end{aligned}$$

where the last equality follows from (69) and (70) .
Denoting $N = W_H\cdots W_{i+1}W_i'W_{i-1}\cdots W_1$, we have

$$\begin{aligned} L(\mathbf{W}^\beta(t)) &= \|W^\beta(t)X - Y\|^2 \\ &= \|U_\mathcal{S}U_\mathcal{S}^T\Sigma_{YX}\Sigma_{XX}^{-1}X - Y + t\beta NX + \beta t^2(W_H\cdots W_{i+1})_{.,k}(W_{j-1}\cdots W_1)_{l,.}X\|^2\ . \end{aligned}$$

The second-order coefficient of $L(\mathbf{W}^\beta(t))$ is equal to

$$\begin{aligned} c_2(\mathbf{W},\mathbf{W}'_\beta) &= \beta^2\|NX\|^2 + 2\beta\,\mathrm{tr}\left((W_H\cdots W_{i+1})_{.,k}(W_{j-1}\cdots W_1)_{l,.}XX^T\Sigma_{XX}^{-1}\Sigma_{XY}U_\mathcal{S}U_\mathcal{S}^T\right) \\ &\quad - 2\beta\,\mathrm{tr}\left((W_H\cdots W_{i+1})_{.,k}(W_{j-1}\cdots W_1)_{l,.}\Sigma_{XY}\right) \\ &= \beta^2\|NX\|^2 + 2\beta\,\mathrm{tr}\left((W_H\cdots W_{i+1})_{.,k}(W_{j-1}\cdots W_1)_{l,.}\Sigma_{XY}(U_\mathcal{S}U_\mathcal{S}^T - I_{d_y})\right)\ . \end{aligned}$$

Using, from Lemma 22, that $U_\mathcal{S}U_\mathcal{S}^T - I_{d_y} = -U_QU_Q^T$, and then the cyclic property of the trace, we obtain

$$\begin{aligned} c_2(\mathbf{W},\mathbf{W}'_\beta) &= \beta^2\|NX\|^2 - 2\beta\,\mathrm{tr}\left((W_H\cdots W_{i+1})_{.,k}(W_{j-1}\cdots W_1)_{l,.}\Sigma_{XY}U_QU_Q^T\right) \\ &= \beta^2\|NX\|^2 - 2\beta(W_{j-1}\cdots W_1)_{l,.}\Sigma_{XY}U_QU_Q^T(W_H\cdots W_{i+1})_{.,k}\ . \end{aligned}$$

Since from (68), $(W_{j-1}\cdots W_1)_{l,.}\Sigma_{XY}U_QU_Q^T(W_H\cdots W_{i+1})_{.,k} \neq 0$, we can choose β according to (7) such that $c_2(\mathbf{W},\mathbf{W}'_\beta) < 0$. Therefore, $\mathbf{W}$ is not a second-order critical point.

Appendix F. Non-strict Saddle Points

In this section, we prove the results related to non-strict saddle points (see Section 4.3).

F.1 Proof of Proposition 17

To prove Proposition 17, we show that for any $\mathbf{W}'$, $c_2(\mathbf{W},\mathbf{W}') \geq 0$, which is equivalent to say (see Lemma 12) that $\mathbf{W}$ is a second-order critical point. We follow the proof strategy sketched in Section 4.3 after the statement of Proposition 17, and use the same notation introduced therein. Note that a first-order critical point can only be tightened if $H \geq 3$. Therefore, in all of this section we make the assumption $H \geq 3$. Recall that m is the number of examples in our sample, $\mathcal{S} = [\![1, r]\!]$, with $r < r_{max}$. We set $Q = [\![r+1, d_y]\!]$.
Recall also that

$$\Sigma^{1/2} = \Sigma_{YX}\Sigma_{XX}^{-1}X \in \mathbb{R}^{d_y\times m}.$$

and

$$\Sigma^{1/2} = U\Delta V^T$$

is a Singular Value Decomposition of $\Sigma^{1/2}$, where $\Delta \in \mathbb{R}^{d_y \times m}$ is such that $\Delta_{ii} = \sqrt{\lambda_i}$ for all $i \in [\![1, d_y]\!]$, and $(\lambda_i)_{i=1..d_y}$ are the eigenvalues of Σ.
We denote

$$\Delta^{(\mathcal{S})} = \mathrm{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_r}) \in \mathbb{R}^{r\times r} \tag{71}$$

and

$$\Delta^{(Q)} = \mathrm{diag}(\sqrt{\lambda_{r+1}}, \dots, \sqrt{\lambda_{d_y}}) \in \mathbb{R}^{(d_y-r)\times(d_y-r)} \ . \tag{72}$$

Recall that, from Section 4.3, $c_2(\mathbf{W}, \mathbf{W}') = FT + ST$.
In what follows, we are going to present a key lemma, then various quick technical lemmas, then we simplify the expressions of FT and ST and conclude the proof of Proposition 17. Then, we prove all the lemmas of Appendix F.1.
We present a lemma which uses that $\mathbf{W}$ is tightened to simplify some products of weight matrices and lighten further calculations. This is a key lemma as it introduces indices p and q which will be used multiple times in the proof.

Lemma 30 *Suppose Assumption 1 in Section 2 holds true. Let* $\mathbf{W} = (W_H, \dots, W_1)$ *be a first-order critical point of* L *verifying the hypotheses of Proposition 17, and* r*,* $\mathcal{S}$*,* Q*,* $(Z_h)_{h=1..H}$ *as in Proposition 17. If* $\mathbf{W}$ *is tightened, then, there exist* $p \in [\![3, H]\!]$ *and* $q \in [\![1, \min(p-1, H-2)]\!]$ *such that:*

$$\forall i \in [\![1, p-1]\!], \qquad W_H \cdots W_{i+1} = [U_{\mathcal{S}} \ , \ 0] \tag{73}$$

$$\forall i \in [\![p, H]\!], \qquad W_{i-1} \cdots W_2 = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} \tag{74}$$

$$\forall i \in [\![q+1, H]\!], \qquad Z_{i-1} \cdots Z_2 Z_1 \Sigma_{XY} U_Q = 0 \tag{75}$$

$$\forall i \in [\![1, q]\!], \qquad W_{H-1} \cdots W_{i+1} = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} \ . \tag{76}$$

The proof of Lemma 30 is in Appendix F.1.5.

F.1.1 USEFUL TECHNICAL LEMMAS

We now present technical lemmas which will be useful in Sections F.1.2, F.1.3 and F.1.4. In all of these Lemmas, we have $\mathcal{S} = [\![1, r]\!]$ and $Q = [\![r+1, d_y]\!]$, and Assumption 1 holds true.

Lemma 31 *We have*

$$\Sigma_{XY} U_Q = X V_Q \Delta^{(Q)} \ .$$

The proof of Lemma 31 is in Appendix F.1.6.

Lemma 32 *Let* n *be a positive integer. For any matrices* $A \in \mathbb{R}^{d_y \times n}$ *and* $B \in \mathbb{R}^{r \times n}$ *we have*

$$\|A + U_{\mathcal{S}} B\|^2 = \|U_{\mathcal{S}}^T A + B\|^2 + \|U_Q^T A\|^2 \ .$$

The proof of Lemma 32 is in Appendix F.1.7.

Lemma 33 *Let n be any positive integer. For any matrices $A \in \mathbb{R}^{n\times r}$ and $B \in \mathbb{R}^{n\times(d_y-r)}$ we have:*

$$\left\langle AU_{\mathcal{S}}^T \Sigma_{YX}\Sigma_{XX}^{-1} X \ , \ BV_Q^T \right\rangle = 0 \ .$$

The proof of Lemma 33 is in Appendix F.1.8.

Lemma 34 *Let n be any positive integer. Let $\mathbf{W} = (W_H, \dots, W_1)$ be a first-order critical point of L verifying the hypotheses of Proposition 17, and r, $\mathcal{S}$, Q, $(Z_h)_{h=1..H}$ as in Proposition 17. If W is tightened, then, for q as in Lemma 30, for any matrices $A \in \mathbb{R}^{n\times(d_q-r)}$ and $B \in \mathbb{R}^{n\times(d_y-r)}$, we have:*

$$\left\langle AZ_q \cdots Z_2 Z_1 X \ , \ BV_Q^T \right\rangle = 0 \ .$$

The proof of Lemma 34 is in Appendix F.1.9.

Lemma 35 *For any matrix $A \in \mathbb{R}^{(d_y-r)\times r}$ we have*

$$\|AU_{\mathcal{S}}^T \Sigma_{YX}\Sigma_{XX}^{-1} X\|^2 = \sum_{a=1}^{r} \sum_{b=r+1}^{d_y} (\lambda_a - \lambda_b)(A_{b-r,a})^2 + \|\Delta^{(Q)} A\|^2 \ .$$

The proof of Lemma 35 is in Appendix F.1.10.

Lemma 36 *Let $\mathbf{W} = (W_H, \dots, W_1)$ be a first-order critical point associated with $\mathcal{S}$. For any matrix $A \in \mathbb{R}^{d_y\times d_x}$, we have*

$$\langle AX \ , \ W_H \cdots W_1 X - Y\rangle = \left\langle A \ , \ -U_Q U_Q^T \Sigma_{YX} \right\rangle \ .$$

The proof of Lemma 36 is in Appendix F.1.11.

F.1.2 SIMPLIFYING FT

In this section and the next one, we simplify the expressions of FT and ST as defined in (14) and (15). In order to decompose $FT = a_1 + \|A_2\|^2 + \|A_3\|^2 + \|A_4\|^2$, with $a_1 \geq 0$, we first simplify the terms T_i, for $i \in [\![1, H]\!]$, defined in (12). Let us first consider $\mathbf{W}$ tightened satisfying the hypotheses of Proposition 17, and p and q defined as in Lemma 30. The simplification of T_i depends on the position of i with regard to 1 , q, p and H. We define $J_1 = [\![p, H-1]\!]$, $J_2 = [\![q+1, p-1]\!]$ and $J_3 = [\![2, q]\!]$.
Note that, according to the convention in Section 2, these sets could be empty.

- if $p = H$, $J_1 = \emptyset$
- if $q = p - 1$, $J_2 = \emptyset$
- if $q = 1$, $J_3 = \emptyset$.

Note also that $\{1\}, J_3, J_2, J_1, \{H\}$ are disjoint and $\{1\} \cup J_3 \cup J_2 \cup J_1 \cup \{H\} = [\![1, H]\!]$.
Depending on the position of i, we need to distinguish four cases, in order to simplify T_i.

Lemma 37 *Suppose Assumption 1 in Section 2 holds true. Let* $\mathbf{W} = (W_H, \ldots, W_1)$ *be a first-order critical point satisfying the hypotheses of Proposition 17, and* r, $\mathcal{S}$, Q, $(Z_h)_{h=1..H}$ *as in Proposition 17. Let* $i \in [\![1, H]\!]$. *For any* $\mathbf{W}' = (W'_H, \ldots, W'_1)$, *recall that , as defined in* (12)*,*

$$T_i = W_H \cdots W_{i+1} W'_i W_{i-1} \cdots W_1 X \ .$$

If $\mathbf{W}$ *is tightened, then, for* p *and* q *as defined in Lemma 30 and* J_1*,*J_2*,*J_3 *as defined above, we have*

- *For* $i = H$*:*

$$T_H = (W'_H)_{.,1:r} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} X \tag{77}$$

- *For* $i \in J_1$*:*

$$T_i = U_{\mathcal{S}} (W'_i)_{1:r,1:r} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} X + U_Q Z_H Z_{H-1} \cdots Z_{i+1} (W'_i)_{r+1:d_i,1:r} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} X \tag{78}$$

- *For* $i \in J_2 \cup J_3$*:*

$$T_i = U_{\mathcal{S}} (W'_i)_{1:r,1:r} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} X + U_{\mathcal{S}} (W'_i)_{1:r,r+1:d_{i-1}} Z_{i-1} \cdots Z_2 Z_1 X \tag{79}$$

- *For* $i = 1$*:*

$$T_1 = U_{\mathcal{S}} (W'_1)_{1:r,.} X \tag{80}$$

The proof of Lemma 37 is in Appendix F.1.12.
We now simplify FT. Substituting the formulas of Lemma 37 in (14) we have

$$\begin{aligned} FT &= \left\| \sum_{i=1}^{H} T_i \right\|^2 \\ &= \Big\| (W'_H)_{.,1:r} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} X \\ &\quad + \sum_{i \in J_1} \big(U_{\mathcal{S}} (W'_i)_{1:r,1:r} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} X + U_Q Z_H Z_{H-1} \cdots Z_{i+1} (W'_i)_{r+1:d_i,1:r} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} X \big) \\ &\quad + \sum_{i \in J_2 \cup J_3} \big(U_{\mathcal{S}} (W'_i)_{1:r,1:r} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} X + U_{\mathcal{S}} (W'_i)_{1:r,r+1:d_{i-1}} Z_{i-1} \cdots Z_2 Z_1 X \big) + U_{\mathcal{S}} (W'_1)_{1:r,.} X \Big\|^2 \ . \end{aligned}$$

FT can be identified with a term as $\|A + U_{\mathcal{S}} B\|^2$ if we take

$$A = (W'_H)_{.,1:r} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} X + \sum_{i \in J_1} U_Q Z_H Z_{H-1} \cdots Z_{i+1} (W'_i)_{r+1:d_i,1:r} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} X \ .$$

and

$$\begin{aligned} B &= \sum_{i \in J_1} (W'_i)_{1:r,1:r} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} X \\ &\quad + \sum_{i \in J_2 \cup J_3} \big((W'_i)_{1:r,1:r} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} X + (W'_i)_{1:r,r+1:d_{i-1}} Z_{i-1} \cdots Z_2 Z_1 X \big) + (W'_1)_{1:r,.} X \ . \end{aligned}$$

Applying Lemma 32, FT becomes:

$$\begin{aligned} FT &= \|U_{\mathcal{S}}^T A + B\|^2 + \|U_Q^T A\|^2 \\ &= \Bigg\| U_{\mathcal{S}}^T (W'_H)_{.,1:r} U_{\mathcal{S}}^T \Sigma_{YX}\Sigma_{XX}^{-1} X + \sum_{i\in J_1} U_{\mathcal{S}}^T U_Q Z_H Z_{H-1}\cdots Z_{i+1}(W'_i)_{r+1:d_i,1:r} U_{\mathcal{S}}^T \Sigma_{YX}\Sigma_{XX}^{-1} X \\ &\quad + \sum_{i\in J_1} (W'_i)_{1:r,1:r} U_{\mathcal{S}}^T \Sigma_{YX}\Sigma_{XX}^{-1} X \\ &\quad + \sum_{i\in J_2\cup J_3} \big((W'_i)_{1:r,1:r} U_{\mathcal{S}}^T \Sigma_{YX}\Sigma_{XX}^{-1} X + (W'_i)_{1:r,r+1:d_{i-1}} Z_{i-1}\cdots Z_2 Z_1 X\big) + (W'_1)_{1:r,.} X \Bigg\|^2 \\ &\quad + \Bigg\| U_Q^T (W'_H)_{.,1:r} U_{\mathcal{S}}^T \Sigma_{YX}\Sigma_{XX}^{-1} X + \sum_{i\in J_1} U_Q^T U_Q Z_H Z_{H-1}\cdots Z_{i+1}(W'_i)_{r+1:d_i,1:r} U_{\mathcal{S}}^T \Sigma_{YX}\Sigma_{XX}^{-1} X \Bigg\|^2 . \end{aligned}$$

Using Lemma 22, we have $U_{\mathcal{S}}^T U_Q = 0$ and $U_Q^T U_Q = I_{d_y - r}$, hence we can write

$$FT = FT_1 + FT_2 \ ,$$

where

$$\begin{aligned} FT_1 = \Bigg\| & U_{\mathcal{S}}^T (W'_H)_{.,1:r} U_{\mathcal{S}}^T \Sigma_{YX}\Sigma_{XX}^{-1} X + \sum_{i\in J_1} (W'_i)_{1:r,1:r} U_{\mathcal{S}}^T \Sigma_{YX}\Sigma_{XX}^{-1} X \\ & + \sum_{i\in J_2\cup J_3} \big((W'_i)_{1:r,1:r} U_{\mathcal{S}}^T \Sigma_{YX}\Sigma_{XX}^{-1} X + (W'_i)_{1:r,r+1:d_{i-1}} Z_{i-1}\cdots Z_2 Z_1 X\big) + (W'_1)_{1:r,.} X \Bigg\|^2 , \end{aligned}$$

and

$$FT_2 = \Bigg\| U_Q^T (W'_H)_{.,1:r} U_{\mathcal{S}}^T \Sigma_{YX}\Sigma_{XX}^{-1} X + \sum_{i\in J_1} Z_H Z_{H-1}\cdots Z_{i+1}(W'_i)_{r+1:d_i,1:r} U_{\mathcal{S}}^T \Sigma_{YX}\Sigma_{XX}^{-1} X \Bigg\|^2 .$$

Let us first simplify FT_1.
Recall that m is the number of examples in our sample, $V \in \mathbb{R}^{m\times m}$ is the orthogonal matrix defined in (1) and $Q = [\![r+1, d_y]\!]$. We set $\mathcal{S}' = \mathcal{S} \cup [\![d_y+1, m]\!] = [\![1, r]\!] \cup [\![d_y+1, m]\!]$ such that $\mathcal{S}' \cup Q = [\![1, m]\!]$.
Reordering the terms and, since V is orthogonal, using $I_m = VV^T = V_{\mathcal{S}'}V_{\mathcal{S}'}^T + V_Q V_Q^T$, we have

$$\begin{aligned} FT_1 = \Bigg\| & \Bigg(U_{\mathcal{S}}^T (W'_H)_{.,1:r} + \sum_{i\in J_1\cup J_2\cup J_3} (W'_i)_{1:r,1:r} \Bigg) U_{\mathcal{S}}^T \Sigma_{YX}\Sigma_{XX}^{-1} X \\ & + \sum_{i\in J_2} (W'_i)_{1:r,r+1:d_{i-1}} Z_{i-1}\cdots Z_2 Z_1 X \end{aligned}$$

$$+ \left(\sum_{i \in J_3} (W_i')_{1:r,r+1:d_{i-1}} Z_{i-1} \cdots Z_2 Z_1 X + (W_1')_{1:r,.} X \right) \left(V_{\mathcal{S}'} V_{\mathcal{S}'}^T + V_Q V_Q^T \right) \Bigg\|^2 .$$

Since for $i \in J_2$, we have $i - 1 \geq q$, we denote

$$N := \sum_{i \in J_2} (W_i')_{1:r,r+1:d_{i-1}} Z_{i-1} \cdots Z_{q+1} ,$$

Recall that, using the convention in Section 2, for $i - 1 = q$, we have $Z_{i-1} \cdots Z_{q+1} = I_{d_q - r}$.
We also denote

$$\begin{aligned}
M &:= \sum_{i \in J_3} (W_i')_{1:r,r+1:d_{i-1}} Z_{i-1} \cdots Z_2 Z_1 X V_Q + (W_1')_{1:r,.} X V_Q , \\
J &:= \sum_{i \in J_3} (W_i')_{1:r,r+1:d_{i-1}} Z_{i-1} \cdots Z_2 Z_1 X V_{\mathcal{S}'} + (W_1')_{1:r,.} X V_{\mathcal{S}'} , \\
L &:= U_{\mathcal{S}}^T (W_H')_{.,1:r} + \sum_{i \in J_1 \cup J_2 \cup J_3} (W_i')_{1:r,1:r} .
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
FT_1 &= \left\| L U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} X + N Z_q \cdots Z_2 Z_1 X + J V_{\mathcal{S}'}^T + M V_Q^T \right\|^2 \\
&= \left\| L U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} X + N Z_q \cdots Z_2 Z_1 X + J V_{\mathcal{S}'}^T \right\|^2 + \left\| M V_Q^T \right\|^2 \\
&\quad + 2 \left\langle L U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} X + N Z_q \cdots Z_2 Z_1 X + J V_{\mathcal{S}'}^T , \; M V_Q^T \right\rangle .
\end{aligned}$$

Using Lemma 33 and Lemma 34 and $V_Q^T V_{\mathcal{S}'} = 0$ (since V is orthogonal), the cross-product is equal to zero.
Noting also that since V is orthogonal $\|M V_Q^T\|^2 = \mathrm{tr}(M V_Q^T V_Q M^T) = \mathrm{tr}(M M^T) = \|M\|^2 = \|M^T\|^2$, we have

$$\begin{aligned}
FT_1 &= \left\| L U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} X + N Z_q \cdots Z_2 Z_1 X + J V_{\mathcal{S}'}^T \right\|^2 + \left\| M^T \right\|^2 \\
&= \|A_2\|^2 + \|A_4\|^2
\end{aligned}$$

where

$$\begin{aligned}
A_2 &:= L U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} X + N Z_q \cdots Z_2 Z_1 X + J V_{\mathcal{S}'}^T \\
&= U_{\mathcal{S}}^T (W_H')_{.,1:r} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} X + \sum_{i \in J_1} (W_i')_{1:r,1:r} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} X \\
&\quad + \sum_{i \in J_2} \left((W_i')_{1:r,1:r} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} X + (W_i')_{1:r,r+1:d_{i-1}} Z_{i-1} \cdots Z_2 Z_1 X \right) \\
&\quad + \sum_{i \in J_3} \left((W_i')_{1:r,1:r} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} X + (W_i')_{1:r,r+1:d_{i-1}} Z_{i-1} \cdots Z_2 Z_1 X V_{\mathcal{S}'} V_{\mathcal{S}'}^T \right) + (W_1')_{1:r,.} X V_{\mathcal{S}'} V_{\mathcal{S}'}^T
\end{aligned} \tag{81}$$

$$A_4 := M^T = \left(\sum_{i \in J_3} (W_i')_{1:r, r+1:d_{i-1}} Z_{i-1} \cdots Z_2 Z_1 X V_Q + (W_1')_{1:r,.} X V_Q \right)^T . \tag{82}$$

Let us now simplify FT_2.
We have $FT_2 = \left\| A U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} X \right\|^2$, with

$$A := U_Q^T (W_H')_{.,1:r} + \sum_{i \in J_1} Z_H Z_{H-1} \cdots Z_{i+1} (W_i')_{r+1:d_i, 1:r} \in \mathbb{R}^{(d_y - r) \times r} .$$

Hence, using Lemma 35, we have

$$\begin{aligned}
FT_2 &= \sum_{a=1}^{r} \sum_{b=r+1}^{d_y} (\lambda_a - \lambda_b)(A_{b-r,a})^2 + \|\Delta^{(Q)} A\|^2 \\
&= \sum_{a=1}^{r} \sum_{b=r+1}^{d_y} (\lambda_a - \lambda_b) \left(U_b^T (W_H')_{.,a} + \sum_{i \in J_1} (Z_H)_{b-r,.} Z_{H-1} \cdots Z_{i+1} (W_i')_{r+1:d_i, a} \right)^2 \\
&\quad + \left\| \Delta^{(Q)} \left(U_Q^T (W_H')_{.,1:r} + \sum_{i \in J_1} Z_H Z_{H-1} \cdots Z_{i+1} (W_i')_{r+1:d_i, 1:r} \right) \right\|^2 \\
&= a_1 + \|A_3\|^2 ,
\end{aligned}$$

where

$$a_1 := \sum_{a=1}^{r} \sum_{b=r+1}^{d_y} (\lambda_a - \lambda_b) \left(U_b^T (W_H')_{.,a} + \sum_{i \in J_1} (Z_H)_{b-r,.} Z_{H-1} \cdots Z_{i+1} (W_i')_{r+1:d_i, a} \right)^2 \tag{83}$$

$$A_3 := \Delta^{(Q)} \left(U_Q^T (W_H')_{.,1:r} + \sum_{i \in J_1} Z_H Z_{H-1} \cdots Z_{i+1} (W_i')_{r+1:d_i, 1:r} \right) \tag{84}$$

Finally,

$$\begin{aligned}
FT &= FT_1 + FT_2 \\
&= a_1 + \|A_2\|^2 + \|A_3\|^2 + \|A_4\|^2 ,
\end{aligned} \tag{85}$$

where a_1, A_2, A_3, A_4 are defined in (83), (81), (84), (82). Notice that, since $\lambda_1 > \cdots > \lambda_{d_y}$, we have

$$a_1 \geq 0 . \tag{86}$$

F.1.3 SIMPLIFYING ST

In this section, we prove that $ST = -2 \langle A_3, A_4 \rangle$, where ST, A_3 and A_4 are defined in (15), (84) and (82). In order to do so, we first state a lemma that simplifies the terms $T_{i,j}$ defined in (13). We remind that the sets J_1, J_2 and J_3 are defined at the beginning of Section F.1.2.

Lemma 38 *Suppose Assumption 1 in Section 2 holds true. Let* $\mathbf{W} = (W_H, \cdots, W_1)$ *be a first-order critical point satisfying the hypotheses of Proposition 17, and* r, $\mathcal{S}$, Q, $(Z_h)_{h=1..H}$ *defined as in Proposition 17. Let* $(i,j) \in [\![1,H]\!]^2$*, with* $i > j$*. For any* $\mathbf{W}' = (W'_H, \ldots, W'_1)$*, recall that , as defined in* (13)*,*

$$T_{i,j} = \left\langle W_H \cdots W_{i+1} W'_i W_{i-1} \cdots W_{j+1} W'_j W_{j-1} \cdots W_1 X \ , \ W_H \cdots W_1 X - Y \right\rangle \ .$$

If $\mathbf{W}$ *is tightened, then, for* p *and* q *as defined in Lemma 30 and* J_1, J_2, J_3 *as defined above, we have*

- *For* $i = H$*:*
  - *For* $j \in J_3$*:*
  $$T_{H,j} = -\left\langle \Delta^{(Q)} U_Q^T (W'_H)_{\cdot,1:r} \ , \ \left((W'_j)_{1:r,r+1:d_{j-1}} Z_{j-1} \cdots Z_2 Z_1 X V_Q\right)^T \right\rangle \ . \tag{87}$$
  - *For* $j = 1$*:*
  $$T_{H,1} = -\left\langle \Delta^{(Q)} U_Q^T (W'_H)_{\cdot,1:r} \ , \ \left((W'_1)_{1:r,\cdot} X V_Q\right)^T \right\rangle \ . \tag{88}$$
  - *For* $j \in J_1 \cup J_2$*:*
  $$T_{H,j} = 0 \ . \tag{89}$$
- *For* $i \in J_1$*:*
  - *For* $j \in J_3$*:*
  $$T_{i,j} = -\left\langle \Delta^{(Q)} Z_H Z_{H-1} \cdots Z_{i+1} (W'_i)_{r+1:d_i,1:r} \ , \ \left((W'_j)_{1:r,r+1:d_{j-1}} Z_{j-1} \cdots Z_2 Z_1 X V_Q\right)^T \right\rangle \ . \tag{90}$$
  - *For* $j = 1$*:*
  $$T_{i,1} = -\left\langle \Delta^{(Q)} Z_H Z_{H-1} \cdots Z_{i+1} (W'_i)_{r+1:d_i,1:r} \ , \ ((W'_1)_{1:r,\cdot} X V_Q)^T \right\rangle \ . \tag{91}$$
  - *For* $j \in J_1 \cup J_2$*:*
  $$T_{i,j} = 0 \ . \tag{92}$$
- *For* $i \in J_2 \cup J_3$*, for all* $j < i$*, we have*
$$T_{i,j} = 0 \ . \tag{93}$$

The proof of Lemma 38 is in Appendix F.1.13.
Let us now prove that $ST = -2\langle A_3, A_4 \rangle$. We remind that $[\![1,H]\!] = \{H\} \cup J_1 \cup J_2 \cup J_3 \cup \{1\}$ and separate the sum appearing in (15) accordingly.
We then substitute the formulas of Lemma 38 in (15) and obtain

$$ST = 2 \sum_{H \geq i > j \geq 1} T_{i,j}$$

$$
\begin{aligned}
&= 2\left(\sum_{j\in J_1\cup J_2} T_{H,j} + \sum_{j\in J_3} T_{H,j} + T_{H,1} + \sum_{i\in J_1}\sum_{\substack{j\in J_1\cup J_2,\\ j<i}} T_{i,j} + \sum_{i\in J_1}\sum_{j\in J_3} T_{i,j} + \sum_{i\in J_1} T_{i,1} + \sum_{i\in J_2\cup J_3}\sum_{j=1}^{i-1} T_{i,j}\right)\\
&= -2\sum_{j\in J_3}\left\langle \Delta^{(Q)} U_Q^T (W'_H)_{.,1:r}\ ,\ \left((W'_j)_{1:r,r+1:d_{j-1}} Z_{j-1}\cdots Z_2 Z_1 X V_Q\right)^T\right\rangle\\
&\quad - 2\left\langle \Delta^{(Q)} U_Q^T (W'_H)_{.,1:r}\ ,\ ((W'_1)_{1:r,.} X V_Q)^T\right\rangle\\
&\quad - 2\sum_{i\in J_1}\sum_{j\in J_3}\left\langle \Delta^{(Q)} Z_H Z_{H-1}\cdots Z_{i+1}(W'_i)_{r+1:d_i,1:r}\ ,\ \left((W'_j)_{1:r,r+1:d_{j-1}} Z_{j-1}\cdots Z_2 Z_1 X V_Q\right)^T\right\rangle\\
&\quad - 2\sum_{i\in J_1}\left\langle \Delta^{(Q)} Z_H Z_{H-1}\cdots Z_{i+1}(W'_i)_{r+1:d_i,1:r}\ ,\ ((W'_1)_{1:r,.} X V_Q)^T\right\rangle\\
&= -2\left\langle \Delta^{(Q)} U_Q^T (W'_H)_{.,1:r}\ ,\ \left(\sum_{j\in J_3} (W'_j)_{1:r,r+1:d_{j-1}} Z_{j-1}\cdots Z_2 Z_1 X V_Q + (W'_1)_{1:r,.} X V_Q\right)^T\right\rangle\\
&\quad - 2\left\langle \Delta^{(Q)} \sum_{i\in J_1} Z_H Z_{H-1}\cdots Z_{i+1}(W'_i)_{r+1:d_i,1:r}\ ,\right.\\
&\qquad\qquad \left.\left(\sum_{j\in J_3} (W'_j)_{1:r,r+1:d_{j-1}} Z_{j-1}\cdots Z_2 Z_1 X V_Q + (W'_1)_{1:r,.} X V_Q\right)^T\right\rangle\\
&= -2\left\langle \Delta^{(Q)}\left(U_Q^T (W'_H)_{.,1:r} + \sum_{i\in J_1} Z_H Z_{H-1}\cdots Z_{i+1}(W'_i)_{r+1:d_i,1:r}\right)\ ,\right.\\
&\qquad\qquad \left.\left(\sum_{j\in J_3} (W'_j)_{1:r,r+1:d_{j-1}} Z_{j-1}\cdots Z_2 Z_1 X V_Q + (W'_1)_{1:r,.} X V_Q\right)^T\right\rangle\\
&= -2\langle A_3, A_4\rangle\ , \qquad (94)
\end{aligned}
$$

where we remind that A_3 and A_4 are defined in (84) and (82).

F.1.4 CONCLUDING THE PROOF OF PROPOSITION 17

Using the simplifications (85) and (94) above, for any $\mathbf{W}$ satisfying the hypotheses of Proposition 17, if $\mathbf{W}$ is tightened, then for any $\mathbf{W}'$,

$$
\begin{aligned}
c_2(\mathbf{W},\mathbf{W}') &= FT + ST\\
&= a_1 + \|A_2\|^2 + \|A_3\|^2 + \|A_4\|^2 - 2\langle A_3, A_4\rangle\\
&= a_1 + \|A_2\|^2 + \|A_3 - A_4\|^2\ .
\end{aligned}
$$

Using (86), we find $c_2(\mathbf{W},\mathbf{W}') \geq 0$.
Therefore, $\mathbf{W} = (W_H, \ldots, W_1)$ is a second-order critical point.

F.1.5 PROOF OF LEMMA 30

First note that, for $r = 0$, we can easily follow the same proof and see that the result still holds with the conventions adopted in Section 2.
Let us prove (73).
Consider the pivot $(i, j) = (2, 1)$. Its complementary blocks are $\Sigma_{XY} W_H \cdots W_3$ and I_{d_1}. Since **W** is tightened and $\mathrm{rk}(I_{d_1}) = d_1 \geq r_{max} > r$, we have $\mathrm{rk}(\Sigma_{XY} W_H \cdots W_3) = r$. Since Σ_{XY} is full-column rank, we obtain $\mathrm{rk}(W_H \cdots W_3) = r$.
Let $p \in [\![3, H]\!]$ be the largest index such that

$$rk(W_H \cdots W_p) = r \ . \tag{95}$$

Using (8) and (10), we have $W_H \cdots W_p = [U_{\mathcal{S}} \ , \ U_Q Z_H Z_{H-1} \cdots Z_p]$.
Since $rk(W_H \cdots W_p) = r$ and since the columns of $U_Q Z_H Z_{H-1} \cdots Z_p$ are in the vector space spanned by the columns of U_Q (which are orthogonal to the columns of $U_{\mathcal{S}}$), (95) implies

$$Z_H Z_{H-1} \cdots Z_p = 0 \ .$$

Therefore,

$$W_H \cdots W_p = [U_{\mathcal{S}} \ , \ 0] \ .$$

Using (10), for all $i \in [\![1, p-1]\!]$,

$$\begin{aligned} W_H \cdots W_{i+1} &= (W_H \cdots W_p)(W_{p-1} \cdots W_{i+1}) \\ &= [U_{\mathcal{S}} \ , \ 0] \begin{bmatrix} I_r & 0 \\ 0 & Z_{p-1} \cdots Z_{i+1} \end{bmatrix} \\ &= [U_{\mathcal{S}} \ , \ 0] \ . \end{aligned}$$

This proves (73).
Let us prove (74).
We consider the pivot $(p, 1)$. Its complementary blocks are $\Sigma_{XY} W_H \cdots W_{p+1}$ and $W_{p-1} \cdots W_2$. We have, by definition of p, $\mathrm{rk}(W_H \cdots W_{p+1}) > r$. Therefore, since Σ_{XY} is full-column rank, we have $\mathrm{rk}(\Sigma_{XY} W_H \cdots W_{p+1}) = \mathrm{rk}(W_H \cdots W_{p+1}) > r$. Note that this holds both for $p = H$ and for $p < H$. Hence, since **W** is tightened, the second complementary block is of rank r, i.e.

$$\mathrm{rk}(W_{p-1} \cdots W_2) = r \ .$$

Using (10), we also have $W_{p-1} \cdots W_2 = \begin{bmatrix} I_r & 0 \\ 0 & Z_{p-1} \cdots Z_2 \end{bmatrix}$.
Then, since $\mathrm{rk}(W_{p-1} \cdots W_2) = r$, we have $Z_{p-1} \cdots Z_2 = 0$ and

$$W_{p-1} \cdots W_2 = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} .$$

Using (10) again, for all $i \in [\![p, H]\!]$,

$$W_{i-1} \cdots W_2 = (W_{i-1} \cdots W_p)(W_{p-1} \cdots W_2)$$

$$= \begin{bmatrix} I_r & 0 \\ 0 & Z_{i-1} \cdots Z_p \end{bmatrix} \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}$$
$$= \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} .$$

This proves (74).
Let us now prove (75).
Using Proposition 1, Lemma 20 and Lemma 22, we have

$$\text{rk}(W_{p-1} \cdots W_1 \Sigma_{XY}) \geq \text{rk}(W_H \cdots W_1 \Sigma_{XY}) = \text{rk}(U_{\mathcal{S}} U_{\mathcal{S}}^T \Sigma) \geq \text{rk}(U_{\mathcal{S}}^T (U_{\mathcal{S}} U_{\mathcal{S}}^T \Sigma) \Sigma^{-1} U_{\mathcal{S}}) = \text{rk}(I_r) = r .$$

Using (74) for $i = p$, we also have $\text{rk}(W_{p-1} \cdots W_1 \Sigma_{XY}) \leq \text{rk}(W_{p-1} \cdots W_2) = r$. Hence, $\text{rk}(W_{p-1} \cdots W_1 \Sigma_{XY}) = r$.
Notice that, considering the tightened pivot $(H, H-1)$, since $\text{rk}(I_{d_{H-1}}) = d_{H-1} \geq r_{max} > r$, we obtain $\text{rk}(W_{H-2} \cdots W_1 \Sigma_{XY}) = r$.
We consider $q \in [\![1, \min(p-1, H-2)]\!]$ the smallest index such that $rk(W_q \cdots W_1 \Sigma_{XY}) = r$.
Using (10) and (9), we have

$$W_q \cdots W_1 \Sigma_{XY} = \begin{bmatrix} U_{\mathcal{S}}^T \Sigma \\ Z_q \cdots Z_2 Z_1 \Sigma_{XY} \end{bmatrix}$$
$$= \begin{bmatrix} \lambda_1 U_1^T \\ \vdots \\ \lambda_r U_r^T \\ Z_q \cdots Z_2 Z_1 \Sigma_{XY} \end{bmatrix} .$$

Since $\text{rk}(W_q \cdots W_1 \Sigma_{XY}) = r$, every row of $Z_q \cdots Z_2 Z_1 \Sigma_{XY}$ lies in $\text{Vec}(U_1^T, \ldots, U_r^T)$, hence we have

$$Z_q \cdots Z_2 Z_1 \Sigma_{XY} U_Q = 0 .$$

Finally, we conclude that, for all $i \in [\![q+1, H]\!]$,

$$\begin{aligned} Z_{i-1} \cdots Z_2 Z_1 \Sigma_{XY} U_Q &= Z_{i-1} \cdots Z_{q+1} Z_q \cdots Z_2 Z_1 \Sigma_{XY} U_Q \\ &= Z_{i-1} \cdots Z_{q+1} 0 \\ &= 0 . \end{aligned}$$

This proves (75).
Let us now prove (76).
Consider the pivot (H, q). Its complementary blocks are $W_{q-1} \cdots W_1 \Sigma_{XY}$ and $W_{H-1} \cdots W_{q+1}$. We have, by definition of q, $\text{rk}(W_{q-1} \cdots W_1 \Sigma_{XY}) > r$. Hence, since **W** is tightened, the other complementary block is of rank r, i.e. $\text{rk}(W_{H-1} \cdots W_{q+1}) = r$. Using (10), we have

$$W_{H-1} \cdots W_{q+1} = \begin{bmatrix} I_r & 0 \\ 0 & Z_{H-1} \cdots Z_{q+1} \end{bmatrix} .$$

Therefore, $Z_{H-1} \cdots Z_{q+1} = 0$ and

$$W_{H-1} \cdots W_{q+1} = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} .$$

Finally, using (10), for all $i \in [\![1, q]\!]$,

$$\begin{aligned} W_{H-1} \cdots W_{i+1} &= W_{H-1} \cdots W_{q+1} W_q \cdots W_{i+1} \\ &= \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} I_r & 0 \\ 0 & Z_q \cdots Z_{i+1} \end{bmatrix} \\ &= \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} . \end{aligned}$$

This proves (76) and concludes the proof.

F.1.6 PROOF OF LEMMA 31

Recall that $\Sigma^{1/2} = \Sigma_{YX} \Sigma_{XX}^{-1} X$. We have

$$\begin{aligned} \Sigma_{XY} &= XY^T \\ &= XX^T (XX^T)^{-1} XY^T \\ &= X (\Sigma^{1/2})^T . \end{aligned}$$

Using (1), we obtain

$$\Sigma_{XY} = XV \Delta^T U^T,$$

and, since U is orthogonal, we have

$$\Sigma_{XY} U = XV \Delta^T.$$

Restricting the equality to the columns in Q, we obtain

$$\Sigma_{XY} U_Q = X V_Q \Delta^{(Q)} ,$$

where $\Delta^{(Q)}$ is defined in (72). This concludes the proof.

F.1.7 PROOF OF LEMMA 32

Let $A \in \mathbb{R}^{d_y \times n}$ and $B \in \mathbb{R}^{r \times n}$. We have

$$\begin{aligned} \|A + U_{\mathcal{S}} B\|^2 &= \|A\|^2 + \|U_{\mathcal{S}} B\|^2 + 2 \langle A \, , \, U_{\mathcal{S}} B \rangle \\ &= tr(A^T A) + tr(B^T U_{\mathcal{S}}^T U_{\mathcal{S}} B) + 2 \left\langle U_{\mathcal{S}}^T A \, , \, B \right\rangle . \end{aligned}$$

Using Lemma 22, this becomes

$$\begin{aligned} \|A + U_{\mathcal{S}} B\|^2 &= tr \left(A^T (U_{\mathcal{S}} U_{\mathcal{S}}^T + U_Q U_Q^T) A \right) + tr \left(B^T B \right) + 2 \left\langle U_{\mathcal{S}}^T A \, , \, B \right\rangle \\ &= tr(A^T U_Q U_Q^T A) + tr(A^T U_{\mathcal{S}} U_{\mathcal{S}}^T A) + tr(B^T B) + 2 \left\langle U_{\mathcal{S}}^T A \, , \, B \right\rangle \\ &= \|U_Q^T A\|^2 + \|U_{\mathcal{S}}^T A\|^2 + \|B\|^2 + 2 \left\langle U_{\mathcal{S}}^T A \, , \, B \right\rangle \\ &= \|U_Q^T A\|^2 + \|U_{\mathcal{S}}^T A + B\|^2 . \end{aligned}$$

F.1.8 PROOF OF LEMMA 33

Recall that $\Sigma^{1/2} = \Sigma_{YX}\Sigma_{XX}^{-1}X$ has a Singular Value Decomposition $\Sigma^{1/2} = U\Delta V^T$ (see (1)). Hence, we have $\Sigma^{1/2}V = U\Delta$ and therefore $\Sigma^{1/2}V_Q = U_Q\Delta^{(Q)}$, where $\Delta^{(Q)}$ is defined in (72). As a consequence,

$$\begin{aligned} U_{\mathcal{S}}^T\Sigma_{YX}\Sigma_{XX}^{-1}XV_Q &= U_{\mathcal{S}}^T\Sigma^{1/2}V_Q \\ &= U_{\mathcal{S}}^TU_Q\Delta^{(Q)} \\ &= 0 \ , \end{aligned}$$

where the last equality follows from Lemma 22. Finally, we obtain for any $A \in \mathbb{R}^{n\times r}$, $B \in \mathbb{R}^{n\times(d_y-r)}$

$$\begin{aligned} \left\langle AU_{\mathcal{S}}^T\Sigma_{YX}\Sigma_{XX}^{-1}X \ , \ BV_Q^T\right\rangle &= tr(AU_{\mathcal{S}}^T\Sigma_{YX}\Sigma_{XX}^{-1}XV_QB^T) \\ &= 0 \ . \end{aligned}$$

F.1.9 PROOF OF LEMMA 34

Using Lemma 31, we have $\Sigma_{XY}U_Q = XV_Q\Delta^{(Q)}$, then replacing this formula in (75) with $i = q+1$, we have

$$Z_q \cdots Z_2Z_1XV_Q\Delta^{(Q)} = 0 \ .$$

Since $\Delta^{(Q)}$ is diagonal and its diagonal elements are non-zero, it is invertible, hence

$$Z_q \cdots Z_2Z_1XV_Q = 0 \ .$$

Finally, for any matrices $A \in \mathbb{R}^{n\times(d_q-r)}$ and $B \in \mathbb{R}^{n\times(d_y-r)}$, we have

$$\begin{aligned} \left\langle AZ_q \cdots Z_2Z_1X \ , \ BV_Q^T\right\rangle &= tr(AZ_q \cdots Z_2Z_1XV_QB^T) \\ &= 0 \ . \end{aligned}$$

F.1.10 PROOF OF LEMMA 35

Recall that $\Delta^{(\mathcal{S})}$ is defined in (71) and $\Sigma = U\Lambda U^T$. Let $A \in \mathbb{R}^{(d_y-r)\times r}$, we have

$$\begin{aligned} \left\|AU_{\mathcal{S}}^T\Sigma_{YX}\Sigma_{XX}^{-1}X\right\|^2 &= tr(AU_{\mathcal{S}}^T\Sigma U_{\mathcal{S}}A^T) \\ &= tr(A \ \mathrm{diag}(\lambda_1, \dots, \lambda_r)A^T) \\ &= \left\|A\Delta^{(\mathcal{S})}\right\|^2 \\ &= \sum_{a=1}^{r}\sum_{b=r+1}^{d_y}\lambda_a(A_{b-r,a})^2 \\ &= \sum_{a=1}^{r}\sum_{b=r+1}^{d_y}(\lambda_a-\lambda_b)(A_{b-r,a})^2 + \sum_{a=1}^{r}\sum_{b=r+1}^{d_y}\lambda_b(A_{b-r,a})^2 \\ &= \sum_{a=1}^{r}\sum_{b=r+1}^{d_y}(\lambda_a-\lambda_b)(A_{b-r,a})^2 + \|\Delta^{(Q)}A\|^2 \ . \end{aligned}$$

F.1.11 PROOF OF LEMMA 36

Let $\mathbf{W} = (W_H, \ldots, W_1)$ be a first-order critical point associated with $\mathcal{S}$ verifying the hypotheses of Proposition 17 and let $A \in \mathbb{R}^{d_y \times d_x}$. Using (11), (9), and Lemma 22, we have

$$\begin{aligned}\langle AX\ ,\ W_H \cdots W_1 X - Y\rangle &= \left\langle A\ ,\ W_H \cdots W_1 X X^T - Y X^T\right\rangle \\ &= \left\langle A\ ,\ U_{\mathcal{S}} U_{\mathcal{S}}^T \Sigma_{YX}\Sigma_{XX}^{-1} X X^T - \Sigma_{YX}\right\rangle \\ &= \left\langle A\ ,\ U_{\mathcal{S}} U_{\mathcal{S}}^T \Sigma_{YX} - \Sigma_{YX}\right\rangle \\ &= \left\langle A\ ,\ -U_Q U_Q^T \Sigma_{YX}\right\rangle\ .\end{aligned}$$

F.1.12 PROOF OF LEMMA 37

Let $\mathbf{W} = (W_H, \cdots, W_1)$ be a tightened first-order critical point satisfying the hypotheses of Proposition 17, and r, $\mathcal{S}$, Q, $(Z_h)_{h=1..H}$ defined as in Proposition 17. Since $\mathbf{W}$ satisfies the hypotheses of Proposition 17, we are going to use all the equations (8), (9), (10) and (11) defined by these hypotheses and (73), (74), (75) and (76) of Lemma 30.
Let $\mathbf{W}' = (W'_H, \ldots, W'_1)$ and $i \in [\![1, H]\!]$. Recall that T_i is defined in (12) and $J_1 = [\![p, H-1]\!]$, $J_2 = [\![q+1, p-1]\!]$, $J_3 = [\![2, q]\!]$, where p and q are defined as in Lemma 30.

Consider the case $i = H$.
Substituting (74) and (9) in (12), we have

$$\begin{aligned}T_H &= W'_H (W_{H-1} \cdots W_2) W_1 X \\ &= W'_H \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U_{\mathcal{S}}^T \Sigma_{YX}\Sigma_{XX}^{-1} \\ Z_1 \end{bmatrix} X \\ &= W'_H \begin{bmatrix} U_{\mathcal{S}}^T \Sigma_{YX}\Sigma_{XX}^{-1} \\ 0 \end{bmatrix} X \\ &= (W'_H)_{.,1:r} U_{\mathcal{S}}^T \Sigma_{YX}\Sigma_{XX}^{-1} X\ .\end{aligned}$$

This proves (77).

Consider now the case $i \in J_1$.
Substituting (8), (10), (74) and (9), in (12), we have, for $i \in J_1$

$$\begin{aligned}T_i &= W_H (W_{H-1} \cdots W_{i+1}) W'_i (W_{i-1} \cdots W_2) W_1 X \\ &= [U_{\mathcal{S}}\ ,\ U_Q Z_H] \begin{bmatrix} I_r & 0 \\ 0 & Z_{H-1} \end{bmatrix} \cdots \begin{bmatrix} I_r & 0 \\ 0 & Z_{i+1} \end{bmatrix} W'_i \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U_{\mathcal{S}}^T \Sigma_{YX}\Sigma_{XX}^{-1} \\ Z_1 \end{bmatrix} X \\ &= [U_{\mathcal{S}}\ ,\ U_Q Z_H Z_{H-1} \cdots Z_{i+1}]\, W'_i \begin{bmatrix} U_{\mathcal{S}}^T \Sigma_{YX}\Sigma_{XX}^{-1} \\ 0 \end{bmatrix} X \\ &= [U_{\mathcal{S}}\ ,\ U_Q Z_H Z_{H-1} \cdots Z_{i+1}]\, (W'_i)_{.,1:r} U_{\mathcal{S}}^T \Sigma_{YX}\Sigma_{XX}^{-1} X \\ &= U_{\mathcal{S}} (W'_i)_{1:r,1:r} U_{\mathcal{S}}^T \Sigma_{YX}\Sigma_{XX}^{-1} X + U_Q Z_H Z_{H-1} \cdots Z_{i+1} (W'_i)_{r+1:d_i,1:r} U_{\mathcal{S}}^T \Sigma_{YX}\Sigma_{XX}^{-1} X\ .\end{aligned}$$

Note that the above calculations are still valid in the case $i = H-1$. In this case using the convention in Section 2, $W_{H-1} \cdots W_{i+1} = I_{d_{H-1}}$ and $Z_{H-1} \cdots Z_{i+1} = I_{d_{H-1}-r}$.

This proves (78).

Consider now the case $i \in J_2 \cup J_3 = [\![2, p-1]\!]$.
Substituting (73), (10) and (9), in (12), we have, for $i \in J_2 \cup J_3$,

$$\begin{aligned}
T_i &= (W_H \cdots W_{i+1})W_i'(W_{i-1} \cdots W_2)W_1 X \\
&= \begin{bmatrix} U_{\mathcal{S}}\ , \ 0 \end{bmatrix} W_i' \begin{bmatrix} I_r & 0 \\ 0 & Z_{i-1} \end{bmatrix} \cdots \begin{bmatrix} I_r & 0 \\ 0 & Z_2 \end{bmatrix} \begin{bmatrix} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} \\ Z_1 \end{bmatrix} X \\
&= U_{\mathcal{S}} (W_i')_{1:r,.} \begin{bmatrix} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} \\ Z_{i-1} \cdots Z_2 Z_1 \end{bmatrix} X \\
&= U_{\mathcal{S}} (W_i')_{1:r,1:r} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} X + U_{\mathcal{S}} (W_i')_{1:r,r+1:d_{i-1}} Z_{i-1} \cdots Z_2 Z_1 X \ .
\end{aligned}$$

Note that the above calculations are still valid in the case $i = 2$. In this case, using the conventions of Section 2, $W_{i-1} \cdots W_2 = I_{d_1}$ and $Z_{i-1} \cdots Z_2 = I_{d_2 - r}$.
This proves (79).

Consider finally the case $i = 1$.
Substituting (73) in (12), we have

$$\begin{aligned}
T_1 &= (W_H \cdots W_2) W_1' X \\
&= [U_{\mathcal{S}}\ , \ 0]\, W_1' X \\
&= U_{\mathcal{S}} (W_1')_{1:r,.} X \ .
\end{aligned}$$

This proves (80).
Note that, using the conventions of Section 2, the proof still holds for $r = 0$. In this case, $T_i = 0, \forall i$.
This concludes the proof.

F.1.13 PROOF OF LEMMA 38

Let $\mathbf{W} = (W_H, \cdots, W_1)$ be a tightened first-order critical point satisfying the hypotheses of Proposition 17, and r, $\mathcal{S}$, Q, $(Z_h)_{h=1..H}$ defined as in Proposition 17. Since $\mathbf{W}$ satisfies the hypotheses of Proposition 17, we are going to use all the equations (8), (9), (10) and (11) defined by these hypotheses and (73), (74), (75) and (76) of Lemma 30.
Let $\mathbf{W}' = (W_H', \ldots, W_1')$ and $(i, j) \in [\![1, H]\!]^2$, with $i > j$. Recall that $T_{i,j}$ is defined in (13) and $J_1 = [\![p, H-1]\!]$, $J_2 = [\![q+1, p-1]\!]$, $J_3 = [\![2, q]\!]$, where p and q are defined as in Lemma 30.

Consider the case $i \in \{H\} \cup J_1$ and $j \in J_1 \cup J_2$ with $i > j$.
Applying Lemma 36 to (13) and using (10) and (9), we obtain

$$\begin{aligned}
T_{i,j} &= \left\langle W_H \cdots W_{i+1} W_i' W_{i-1} \cdots W_{j+1} W_j' W_{j-1} \cdots W_1 X \ , \ W_H \cdots W_1 X - Y \right\rangle \\
&= \left\langle W_H \cdots W_{i+1} W_i' W_{i-1} \cdots W_{j+1} W_j' W_{j-1} \cdots W_1 \ , \ -U_Q U_Q^T \Sigma_{YX} \right\rangle \\
&= -tr\left(W_H \cdots W_{i+1} W_i' W_{i-1} \cdots W_{j+1} W_j' W_{j-1} \cdots W_1 \Sigma_{XY} U_Q U_Q^T \right) \\
&= -tr\left((W_H \cdots W_{i+1} W_i' W_{i-1} \cdots W_{j+1} W_j' \begin{bmatrix} U_{\mathcal{S}}^T \Sigma U_Q \\ Z_{j-1} \cdots Z_2 Z_1 \Sigma_{XY} U_Q \end{bmatrix} U_Q^T \right) \ .
\end{aligned}$$

Using Lemma 23 and since $j \geq q+1$, using (75), we obtain

$$T_{i,j} = 0.$$

This proves (89) and (92).

Consider now the case $i = H$ and $j \in J_3$.
Applying Lemma (36) to (13) and using (76), (10) and (9), we obtain

$$\begin{aligned}
T_{H,j} &= \left\langle W'_H W_{H-1} \cdots W_{j+1} W'_j W_{j-1} \cdots W_1 X \ , \ W_H \cdots W_1 X - Y \right\rangle \\
&= \left\langle W'_H W_{H-1} \cdots W_{j+1} W'_j W_{j-1} \cdots W_1 \ , \ -U_Q U_Q^T \Sigma_{YX} \right\rangle \\
&= \left\langle W'_H \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} W'_j \begin{bmatrix} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} \\ Z_{j-1} \cdots Z_2 Z_1 \end{bmatrix} \ , \ U_Q U_Q^T \Sigma_{YX} \right\rangle \\
&= -tr\left(W'_H \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} W'_j \begin{bmatrix} U_{\mathcal{S}}^T \Sigma U_Q U_Q^T \\ Z_{j-1} \cdots Z_2 Z_1 \Sigma_{XY} U_Q U_Q^T \end{bmatrix} \right) \ .
\end{aligned}$$

Using Lemma 31, Lemma 23 and the cyclic property of the trace, we have

$$\begin{aligned}
T_{H,j} &= -tr\left((W'_H)_{.,1:r} (W'_j)_{1:r,.} \begin{bmatrix} 0 \\ Z_{j-1} \cdots Z_2 Z_1 X V_Q \Delta^{(Q)} U_Q^T \end{bmatrix} \right) \\
&= -tr\left((W'_H)_{.,1:r} (W'_j)_{1:r,r+1:d_{j-1}} Z_{j-1} \cdots Z_2 Z_1 X V_Q \Delta^{(Q)} U_Q^T \right) \\
&= -tr\left(\Delta^{(Q)} U_Q^T (W'_H)_{.,1:r} (W'_j)_{1:r,r+1:d_{j-1}} Z_{j-1} \cdots Z_2 Z_1 X V_Q \right) \\
&= -\left\langle \Delta^{(Q)} U_Q^T (W'_H)_{.,1:r} \ , \ \left((W'_j)_{1:r,r+1:d_{j-1}} Z_{j-1} \cdots Z_2 Z_1 X V_Q \right)^T \right\rangle \ .
\end{aligned}$$

This proves (87).

Consider now the case $i = H$ and $j = 1$.
Applying Lemma 36 to (13) and using (76) and Lemma 31, we obtain

$$\begin{aligned}
T_{H,1} &= \left\langle W'_H W_{H-1} \cdots W_2 W'_1 X \ , \ W_H \cdots W_1 X - Y \right\rangle \\
&= \left\langle W'_H W_{H-1} \cdots W_2 W'_1 \ , \ -U_Q U_Q^T \Sigma_{YX} \right\rangle \\
&= -\left\langle W'_H \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} W'_1 \ , \ U_Q (X V_Q \Delta^{(Q)})^T \right\rangle \\
&= -\left\langle (W'_H)_{.,1:r} (W'_1)_{1:r,.} \ , \ U_Q \Delta^{(Q)} V_Q^T X^T \right\rangle \\
&= -\left\langle \Delta^{(Q)} U_Q^T (W'_H)_{.,1:r} \ , \ V_Q^T X^T \left((W'_1)_{1:r,.} \right)^T \right\rangle \\
&= -\left\langle \Delta^{(Q)} U_Q^T (W'_H)_{.,1:r} \ , \ \left((W'_1)_{1:r,.} X V_Q \right)^T \right\rangle \ .
\end{aligned}$$

This proves (88).

Consider now the case $i \in J_1$ and $j \in J_3$.
Applying Lemma 36 to (13) and using (8), (10) and (9) , we obtain

$$T_{i,j} = \left\langle W_H \cdots W_{i+1} W'_i W_{i-1} \cdots W_{j+1} W'_j W_{j-1} \cdots W_1 X \ , \ W_H \cdots W_1 X - Y \right\rangle$$

$$
\begin{aligned}
&= \left\langle W_H \cdots W_{i+1} W_i' W_{i-1} \cdots W_{j+1} W_j' W_{j-1} \cdots W_1 \ , \ -U_Q U_Q^T \Sigma_{YX} \right\rangle \\
&= -\left\langle [U_{\mathcal{S}} \ , \ U_Q Z_H Z_{H-1} \cdots Z_{i+1}] W_i' W_{i-1} \cdots W_{j+1} W_j' \begin{bmatrix} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1} \\ Z_{j-1} \cdots Z_2 Z_1 \end{bmatrix} \ , \ U_Q U_Q^T \Sigma_{XY} \right\rangle \\
&= -tr\left([U_{\mathcal{S}} \ , \ U_Q Z_H Z_{H-1} \cdots Z_{i+1}] W_i' W_{i-1} \cdots W_{j+1} W_j' \begin{bmatrix} U_{\mathcal{S}}^T \Sigma \\ Z_{j-1} \cdots Z_2 Z_1 \Sigma_{XY} \end{bmatrix} U_Q U_Q^T \right) \\
&= -tr\left(\left[U_Q^T U_{\mathcal{S}} \ , \ U_Q^T U_Q Z_H Z_{H-1} \cdots Z_{i+1}\right] W_i' W_{i-1} \cdots W_{j+1} W_j' \begin{bmatrix} U_{\mathcal{S}}^T \Sigma U_Q \\ Z_{j-1} \cdots Z_2 Z_1 \Sigma_{XY} U_Q \end{bmatrix} \right) \ .
\end{aligned}
$$

Using Lemma 22 and Lemma 23, we have

$$
\begin{aligned}
T_{i,j} &= -tr\left([0 \ , \ Z_H Z_{H-1} \cdots Z_{i+1}] W_i' W_{i-1} \cdots W_{j+1} W_j' \begin{bmatrix} 0 \\ Z_{j-1} \cdots Z_2 Z_1 \Sigma_{XY} U_Q \end{bmatrix} \right) \\
&= -tr\left(Z_H Z_{H-1} \cdots Z_{i+1} (W_i')_{r+1:d_i,.} W_{i-1} \cdots W_{j+1} (W_j')_{.,r+1:d_{j-1}} Z_{j-1} \cdots Z_2 Z_1 \Sigma_{XY} U_Q \right) \ .
\end{aligned} \tag{96}
$$

Here, since **W** is tightened, taking the tightened pivot (i,j) we have two possible cases: either $\mathrm{rk}(W_{i-1} \cdots W_{j+1}) = r$ or $\mathrm{rk}(W_{j-1} \cdots W_1 \Sigma_{XY} W_H \cdots W_{i+1}) = r$. We treat the two cases separately.
In the first case, using (10) we have

$$
\begin{aligned}
W_{i-1} \cdots W_{j+1} &= \begin{bmatrix} I_r & 0 \\ 0 & Z_{i-1} \end{bmatrix} \cdots \begin{bmatrix} I_r & 0 \\ 0 & Z_{j+1} \end{bmatrix} \\
&= \begin{bmatrix} I_r & 0 \\ 0 & Z_{i-1} \cdots Z_{j+1} \end{bmatrix} \ .
\end{aligned}
$$

Hence, $\mathrm{rk}(W_{i-1} \cdots W_{j+1}) = r$ implies $Z_{i-1} \cdots Z_{j+1} = 0$ and we conclude that

$$
W_{i-1} \cdots W_{j+1} = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} \ .
$$

Then, using this last equality, (96) becomes

$$
\begin{aligned}
T_{i,j} &= -tr\left(Z_H Z_{H-1} \cdots Z_{i+1} (W_i')_{r+1:d_i,.} \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} (W_j')_{.,r+1:d_{j-1}} Z_{j-1} \cdots Z_2 Z_1 \Sigma_{XY} U_Q \right) \\
&= -tr\left(Z_H Z_{H-1} \cdots Z_{i+1} (W_i')_{r+1:d_i,1:r} (W_j')_{1:r,r+1:d_{j-1}} Z_{j-1} \cdots Z_2 Z_1 \Sigma_{XY} U_Q \right) \ .
\end{aligned} \tag{97}
$$

In the second case, we have $\mathrm{rk}(W_{j-1} \cdots W_1 \Sigma_{XY} W_H \cdots W_{i+1}) = r$. Let us prove that (97) also holds in this case. Using (10), (9), (8), Lemma 23 and $\mathcal{S} = [\![1, r]\!]$, we have

$$
\begin{aligned}
&W_{j-1} \cdots W_1 \Sigma_{XY} W_H \cdots W_{i+1} \\
&= \begin{bmatrix} U_{\mathcal{S}}^T \Sigma \\ Z_{j-1} \cdots Z_2 Z_1 \Sigma_{XY} \end{bmatrix} [U_{\mathcal{S}} \ , \ U_Q Z_H Z_{H-1} \cdots Z_{i+1}] \\
&= \begin{bmatrix} U_{\mathcal{S}}^T \Sigma U_{\mathcal{S}} & U_{\mathcal{S}}^T \Sigma U_Q Z_H Z_{H-1} \cdots Z_{i+1} \\ Z_{j-1} \cdots Z_2 Z_1 \Sigma_{XY} U_{\mathcal{S}} & Z_{j-1} \cdots Z_2 Z_1 \Sigma_{XY} U_Q Z_H Z_{H-1} \cdots Z_{i+1} \end{bmatrix} \\
&= \begin{bmatrix} diag(\lambda_1, \cdots, \lambda_r) & 0 \\ Z_{j-1} \cdots Z_2 Z_1 \Sigma_{XY} U_{\mathcal{S}} & Z_{j-1} \cdots Z_2 Z_1 \Sigma_{XY} U_Q Z_H Z_{H-1} \cdots Z_{i+1} \end{bmatrix} \ .
\end{aligned}
$$

Therefore since $\mathrm{rk}(W_{j-1}\cdots W_1\Sigma_{XY}W_H\cdots W_{i+1}) = r$ and for all $i \in [\![1, r]\!]$, $\lambda_i \neq 0$, we must have

$$Z_{j-1}\cdots Z_2Z_1\Sigma_{XY}U_QZ_HZ_{H-1}\cdots Z_{i+1} = 0\,. \tag{98}$$

Using the above equation, and the cyclic property of the trace, (96) becomes

$$\begin{aligned}T_{i,j} &= -tr\left(Z_HZ_{H-1}\cdots Z_{i+1}(W'_i)_{r+1:d_i,.}W_{i-1}\cdots W_{j+1}(W'_j)_{.,r+1:d_{j-1}}Z_{j-1}\cdots Z_2Z_1\Sigma_{XY}U_Q\right)\\ &= -tr\left(Z_{j-1}\cdots Z_2Z_1\Sigma_{XY}U_QZ_HZ_{H-1}\cdots Z_{i+1}(W'_i)_{r+1:d_i,.}W_{i-1}\cdots W_{j+1}(W'_j)_{.,r+1:d_{j-1}}\right)\\ &= 0\,.\end{aligned}$$

We can use (98) again to write the equation $T_{i,j} = 0$ in the format of equation (97). Indeed, we have

$$\begin{aligned}&-tr\left(Z_HZ_{H-1}\cdots Z_{i+1}(W'_i)_{r+1:d_i,1:r}(W'_j)_{1:r,r+1:d_{j-1}}Z_{j-1}\cdots Z_2Z_1\Sigma_{XY}U_Q\right)\\ &= -tr\left(Z_{j-1}\cdots Z_2Z_1\Sigma_{XY}U_QZ_HZ_{H-1}\cdots Z_{i+1}(W'_i)_{r+1:d_i,1:r}(W'_j)_{1:r,r+1:d_{j-1}}\right)\\ &= 0\\ &= T_{i,j}\,.\end{aligned}$$

Therefore, in both cases we have

$$T_{i,j} = -tr\left(Z_HZ_{H-1}\cdots Z_{i+1}(W'_i)_{r+1:d_i,1:r}(W'_j)_{1:r,r+1:d_{j-1}}Z_{j-1}\cdots Z_2Z_1\Sigma_{XY}U_Q\right)\,.$$

Using Lemma 31, it becomes

$$\begin{aligned}T_{i,j} &= -tr\left(Z_HZ_{H-1}\cdots Z_{i+1}(W'_i)_{r+1:d_i,1:r}(W'_j)_{1:r,r+1:d_{j-1}}Z_{j-1}\cdots Z_2Z_1XV_Q\Delta^{(Q)}\right)\\ &= -tr\left(\Delta^{(Q)}Z_HZ_{H-1}\cdots Z_{i+1}(W'_i)_{r+1:d_i,1:r}(W'_j)_{1:r,r+1:d_{j-1}}Z_{j-1}\cdots Z_2Z_1XV_Q\right)\\ &= -\left\langle\Delta^{(Q)}Z_HZ_{H-1}\cdots Z_{i+1}(W'_i)_{r+1:d_i,1:r}\,,\ \left((W'_j)_{1:r,r+1:d_{j-1}}Z_{j-1}\cdots Z_2Z_1XV_Q\right)^T\right\rangle\,.\end{aligned}$$

This proves (90).

Consider now the case $i \in J_1$ and $j = 1$.
Using Lemma 36 to simplify (13), we have

$$\begin{aligned}T_{i,1} &= \left\langle W_H\cdots W_{i+1}W'_iW_{i-1}\cdots W_2W'_1X\,,\ W_H\cdots W_1X - Y\right\rangle\\ &= \left\langle W_H\cdots W_{i+1}W'_iW_{i-1}\cdots W_2W'_1\,,\ -U_QU_Q^T\Sigma_{YX}\right\rangle\,.\end{aligned}$$

Using Lemma 31 and substituting (8), (10), and since $i \geq p$, using (74) , this becomes

$$\begin{aligned}T_{i,1} &= -\left\langle[U_{\mathcal{S}}\,,\ U_QZ_HZ_{H-1}\cdots Z_{i+1}]\,W'_i\begin{bmatrix}I_r & 0\\ 0 & 0\end{bmatrix}W'_1\,,\ U_Q(XV_Q\Delta^{(Q)})^T\right\rangle\\ &= -\left\langle\Delta^{(Q)}\left[U_Q^TU_{\mathcal{S}}\,,\ U_Q^TU_QZ_HZ_{H-1}\cdots Z_{i+1}\right](W'_i)_{.,1:r}(W'_1)_{1:r,.}\,,\ (XV_Q)^T\right\rangle\,.\end{aligned}$$

Using Lemma 22, it becomes

$$T_{i,1} = -\left\langle\Delta^{(Q)}\left[0\,,\ Z_HZ_{H-1}\cdots Z_{i+1}\right](W'_i)_{.,1:r}(W'_1)_{1:r,.}\,,\ (XV_Q)^T\right\rangle$$

$$= -\left\langle \Delta^{(Q)} Z_H Z_{H-1} \cdots Z_{i+1} (W_i')_{r+1:d_i,1:r} \ , \ ((W_1')_{1:r,.} X V_Q)^T \right\rangle \ .$$

This proves (91).

Consider now the case $i \in J_2 \cup J_3 = [\![2, p-1]\!]$ **and** $j < i$.
Applying Lemma 36 to (13) and , since $i < p$, using (73), we obtain

$$\begin{aligned} T_{i,j} &= \langle W_H \cdots W_{i+1} W_i' W_{i-1} \cdots W_{j+1} W_j' W_{j-1} \cdots W_1 X \ , \ W_H \cdots W_1 X - Y \rangle \\ &= \langle W_H \cdots W_{i+1} W_i' W_{i-1} \cdots W_{j+1} W_j' W_{j-1} \cdots W_1 \ , \ -U_Q U_Q^T \Sigma_{YX} \rangle \\ &= -tr([U_\mathcal{S} \ , \ 0] W_i' W_{i-1} \cdots W_{j+1} W_j' W_{j-1} \cdots W_1 \Sigma_{XY} U_Q U_Q^T) \end{aligned}$$

The cyclic property of the trace and Lemma 22 lead to

$$\begin{aligned} T_{i,j} &= -tr([U_Q^T U_\mathcal{S} \ , \ 0] W_i' W_{i-1} \cdots W_{j+1} W_j' W_{j-1} \cdots W_1 \Sigma_{XY} U_Q) \\ &= 0 \ . \end{aligned}$$

This proves (93) and concludes the proof.
Note that, with the convention of Section 2, the proof still holds for $r = 0$. In this case, $T_{i,j} = 0, \forall i > j$.

F.2 Proof of Proposition 15

Let $\mathbf{W} = (W_H, \ldots, W_1)$ be a tightened first-order critical point associated with $\mathcal{S} = [\![1, r]\!]$ with $r < r_{max}$. Then, using Proposition 9 there exist invertible matrices $D_{H-1} \in \mathbb{R}^{d_{H-1} \times d_{H-1}}, \ldots, D_1 \in \mathbb{R}^{d_1 \times d_1}$ and matrices $Z_H \in \mathbb{R}^{(d_y - r) \times (d_{H-1} - r)}$, $Z_1 \in \mathbb{R}^{(d_1 - r) \times d_x}$ and $Z_h \in \mathbb{R}^{(d_h - r) \times (d_{h-1} - r)}$ for $h \in [\![2, H-1]\!]$ such that if we denote $\widetilde{W}_H = W_H D_{H-1}$, $\widetilde{W}_1 = D_1^{-1} W_1$ and $\widetilde{W}_h = D_h^{-1} W_h D_{h-1}$ for all $h \in [\![2, H-1]\!]$, and $\widetilde{\mathbf{W}} = (\widetilde{W}_H, \ldots, \widetilde{W}_1)$, then

$$\begin{aligned} \widetilde{W}_H &= [U_\mathcal{S}, U_Q Z_H] \\ \widetilde{W}_1 &= \begin{bmatrix} U_\mathcal{S}^T \Sigma_{YX} \Sigma_{XX}^{-1} \\ Z_1 \end{bmatrix} \\ \widetilde{W}_h &= \begin{bmatrix} I_r & 0 \\ 0 & Z_h \end{bmatrix} \quad \forall h \in [\![2, H-1]\!] \\ \widetilde{W}_H \cdots \widetilde{W}_2 &= [U_\mathcal{S}, 0] \ . \end{aligned}$$

Then, due to Lemma 16, and since $\mathbf{W}$ is a first-order critical point, we have that $\widetilde{\mathbf{W}}$ is a first-order critical point. We also have $\widetilde{W}_H \cdots \widetilde{W}_1 = W_H \cdots W_1$. Hence, according to Proposition 1 $\widetilde{\mathbf{W}}$ is also associated with $\mathcal{S}$.
Since $\mathbf{W}$ is tightened and multiplication by invertible matrices does not change the rank, $\widetilde{\mathbf{W}}$ is also tightened. Hence, $\widetilde{\mathbf{W}}$ satisfies the hypotheses of Proposition 17 and therefore is a second-order critical point. Finally, using Lemma 16, we conclude that $\mathbf{W}$ is a second-order critical point. Since $r < r_{max}$ and Σ is invertible (Lemma 20), using Proposition 1, we have

$$L(\mathbf{W}) = \mathrm{tr}(\Sigma_{YY}) - \sum_{i=1}^{r} \lambda_i > \mathrm{tr}(\Sigma_{YY}) - \sum_{i=1}^{r_{max}} \lambda_i \ .$$

Therefore, $\mathbf{W}$ is not a global minimizer, hence $\mathbf{W}$ is a non-strict saddle point.

Appendix G. A Simple Illustrative Experiment

Next we provide more details on the experiment whose results were plotted in Figures 3 and 4. The goal is to illustrate the behavior of the ADAM optimizer in the vicinity of strict or non-strict saddle points.

Experimental setting. We optimize a linear neural network starting in the vicinity either of a strict saddle point (10000 runs in total) or of a non-strict saddle point (10000 runs in total). For each run, the setting is the following:

- Network architecture: $d_x = 10$, $d_y = 4$, $H = 5$ and $d_4 = d_3 = d_2 = d_1 = 10$.

- Data construction: $m = 100$ i.i.d. data points $(x_1, y_1), \ldots, (x_m, y_m) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$ such that, for all $i = 1, \ldots, m$, the points x_i and y_i are drawn independently at random from the Gaussian distributions $\mathcal{N}(0,\ I_{d_x})$ and $\mathcal{N}(0,\ I_{d_y})$ respectively.

- Initial iterate: we define it as $(W_1, \ldots, W_H) = \mathbf{W}^{cp} + (V_1, \ldots, V_H)$, for a critical point $\mathbf{W}^{cp}$ (defined later) and a random perturbation $(V_1, \ldots, V_H)$ whose components $(V_h)_{i,j}$ are drawn independently from the distributions $\mathcal{N}(0, \sigma_h^2)$, with $\sigma_h = 0.1 \frac{\|W_h^{cp}\|_F}{\sqrt{d_{h-1} d_h}}$. The critical point $\mathbf{W}^{cp}$ is defined as in (25) in Appendix B.8, for $r = 2$ ($\mathcal{S} = \{1, 2\}$) and

  $$Z_h = \begin{cases} I_{d_h - 2} & \text{for all } h \in [\![2, 4]\!], \text{ for runs starting at a strict saddle point;} \\ 0_{(d_h-2)\times(d_h-2)} & \text{for all } h \in [\![2, 4]\!], \text{ for runs starting at a non-strict saddle point.} \end{cases}$$

  Since $d_4 = d_3 = d_2 = d_1$, note that the sizes of the above matrices Z_h are consistent with (25). As explained in Appendix B.8, when $Z_h = I_{d_h-2}$ for all $h \in [\![2, 4]\!]$, the critical point $\mathbf{W}^{cp}$ is non-tightened and therefore Theorem 7 guarantees that it is a strict saddle point. Similarly, when $Z_h = 0_{(d_h-2)\times(d_h-2)}$ for all $h \in [\![2, 4]\!]$, the critical point $\mathbf{W}^{cp}$ is tightened and Theorem 7 guarantees that it is a non-strict saddle point.

- Optimizer: we use the ADAM optimizer of the Keras library, with the default parameters.

Observations. Figure 3 in Section 1.2 shows the evolution of the loss along the optimization process for two representative runs (initialization near a strict or a non-strict saddle point). We can see that, when initialized in the vicinity of the strict saddle point, ADAM rapidly decreases below the initial value $L(\mathbf{W}^{cp})$. On the contrary, ADAM needs many epochs to exit the plateau at the critical value of the non-strict saddle point.

In order to assess the importance of this phenomenon, we repeated the above experiment 10000 times for both strict saddle points and non-strict saddle points. For each run, we define and compute the escape epoch as the first epoch such that $L(\mathbf{W}) < L(\mathbf{W}^{cp}) - \frac{\lambda_3}{2}$ (the average of the critical values associated with $\mathcal{S} = \{1, 2\}$ and $\mathcal{S}' = \{1, 2, 3\}$). On Figure 4 (Section 1.2) the histograms of the escape epoch are displayed separately for runs corresponding to strict saddle points (in red) or non-strict saddle points (in blue). We can see that, while ADAM quickly escapes from the vicinity of the strict saddle points, it takes many more epochs to escape from the vicinity of the non-strict saddle points. In the last case, the plateau can easily be confused with a global minimum.

References

Leonard Adolphs, Hadi Daneshmand, Aurelien Lucchi, and Thomas Hofmann. Local saddle point optimization: A curvature exploitation approach. In International Conference on Artificial Intelligence and Statistics, AISTATS'19, 2019.

Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In International Conference on Machine Learning, ICML'18, 2018.

Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. In International Conference on Learning Representations, ICLR'19, 2019a.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In Advances in Neural Information Processing Systems, NeurIPS'19, 2019b.

Bubacarr Bah, Holger Rauhut, Ulrich Terstiege, and Michael Westdickenberg. Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers. Information and Inference: A Journal of the IMA, 11(1):307–353, 2022.

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

Peter Bartlett, Dave Helmbold, and Philip Long. Gradient descent with identity initialization efficiently learns positive definite linear transformations by deep residual networks. In International Conference on Machine Learning, ICML'18, 2018.

Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. Acta Numerica, 30:87–201, 2021.

Mikhail Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. Acta Numerica, 30:203–248, 2021.

Avrim Blum and Ronald L Rivest. Training a 3-node neural network is np-complete. In Advances in Neural Information Processing Systems, NIPS'89, 1989.

Yacine Chitour, Zhenyu Liao, and Romain Couillet. A geometric approach of gradient descent algorithms in linear neural networks. Mathematical Control and Related Fields, 13(3):918–945, 2023.

Hadi Daneshmand, Jonas Kohler, Aurelien Lucchi, and Thomas Hofmann. Escaping saddles with stochastic gradients. In International Conference on Machine Learning, ICML'18, 2018.

Simon Du and Wei Hu. Width provably matters in optimization for deep linear neural networks. In International Conference on Machine Learning, ICML'19, 2019.

Armin Eftekhari. Training linear neural networks: Non-local convergence and complexity results. In International Conference on Machine Learning, ICML'20, 2020.

Sébastien Gadat and Ioana Gavra. Asymptotic study of stochastic adaptive algorithm in non-convex landscape. Journal of Machine Learning Research, 23(228):1–54, 2022.

Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. In Advances in Neural Information Processing Systems, NeurIPS'19, 2019.

Daniel Gissin, Shai Shalev-Shwartz, and Amit Daniely. The implicit bias of depth: How incremental learning drives generalization. In International Conference on Learning Representations, ICLR'19, 2019.

Arthur Jacot. Implicit bias of large depth networks: a notion of rank for nonlinear functions. In International Conference on Learning Representations, ICLR'22, 2022.

Arthur Jacot, François Ged, Berfin Şimşek, Clément Hongler, and Franck Gabriel. Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity. arXiv preprint arXiv:2104.05097, 2022.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In International Conference on Machine Learning, ICML'17, 2017.

Chi Jin, Praneeth Netrapalli, and Michael I Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. In Conference On Learning Theory ,COLT'18, 2018.

Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M. Kakade, and Michael I. Jordan. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. J. Association for Computing Machinery, 68(2), 2021.

Kenji Kawaguchi. Deep learning without poor local minima. Advances in Neural Information Processing Systems, NIPS'16, 2016.

Thomas Laurent and James von Brecht. Deep linear networks with arbitrary loss: All local minima are global. In International conference on machine learning, ICML'18, 2018.

Jason D. Lee, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In Conference on Learning Theory, COLT'16, 2016.

Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht. First-order methods almost always avoid strict saddle points. Mathematical programming, 176(1-2):311–337, 2019.

Haihao Lu and Kenji Kawaguchi. Depth creates no bad local minima. arXiv preprint arXiv:1702.08580, 2017.

Pierre Marion, Yu-Han Wu, Michael E. Sander, and Gérard Biau. Implicit regularization of deep residual networks towards neural odes. In International Conference on Learning Representations, ICLR'24, 2024.

Dhagash Mehta, Tianran Chen, Tingting Tang, and Jonathan Hauenstein. The loss surface of deep linear networks viewed through the algebraic geometry lens. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.

Katta G Murty and Santosh N Kabadi. Some np-complete problems in quadratic and nonlinear programming. Mathematical Programming, 39(2):117–129, 1987.

Yurii Nesterov. Introductory lectures on convex programming volume i: Basic course. Lecture notes, 1998.

Gabin Maxime Nguegnang, Holger Rauhut, and Ulrich Terstiege. Convergence of gradient descent for learning linear neural networks. Advances in Continuous and Discrete Models, 2024.

Maher Nouiehed and Meisam Razaviyayn. Learning deep models: Critical points and local openness. INFORMS Journal on Optimization, 4(2):148–173, 2022.

Michael J. O'Neill and Stephen J. Wright. A line-search descent algorithm for strict saddle functions with complexity guarantees. Journal of Machine Learning Research, 24(10):1–34, 2023.

Noam Razin and Nadav Cohen. Implicit regularization in deep learning may not be explainable by norms. In Advances in Neural Information Processing Systems, NeurIPS'20, 2020.

Itay Safran, Gal Vardi, and Jason D Lee. On the effective number of linear regions in shallow univariate relu networks: Convergence guarantees and implicit bias. Advances in Neural Information Processing Systems, NeurIPS'22, 2022.

Andrew M Saxe, James L Mcclelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural network. In In International Conference on Learning Representations, ICLR'14, 2014.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. Proceedings of the National Academy of Sciences, 116 (23):11537–11546, 2019.

Ohad Shamir. Exponential convergence time of gradient descent for one-dimensional deep linear neural networks. In Conference on Learning Theory, COLT'19, 2019.

Nadav Timor, Gal Vardi, and Ohad Shamir. Implicit regularization towards rank minimization in relu networks. In International Conference on Algorithmic Learning Theory, ALT'23, 2023.

Matthew Trager, Kathlén Kohn, and Joan Bruna. Pure and spurious critical points: a geometric study of linear networks. In International Conference on Learning Representations, ICLR'20, 2020.

Lei Wu, Qingcan Wang, and Chao Ma. Global convergence of gradient descent for deep linear residual networks. In Advances in Neural Information Processing Systems, NeurIPS'19, 2019.

Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. Global optimality conditions for deep neural networks. In International Conference on Learning Representations, ICLR'18, 2018.

Yi Zhou and Yingbin Liang. Critical points of linear neural networks: Analytical forms and landscape properties. In International Conference on Learning Representations, ICLR'18, 2018.

Zhihui Zhu, Daniel Soudry, Yonina C Eldar, and Michael B Wakin. The global optimization geometry of shallow linear neural networks. Journal of Mathematical Imaging and Vision, 62(3):279–292, 2020.