Contents

Contributors

René Vidal *Mathematical Institute for Data Science, Department of Biomedical Engineering, Johns Hopkins University, Clark 302B, 3400 N. Charles Street, Baltimore MD 21218, USA*

Zhihui Zhu *Department of Electrical & Computer Engineering, University of Denver, Engineering and Computer Science Building 126C, 2155 E. Wesley Avenue, Denver CO 80208, USA*

Benjamin D Haeffele *Mathematical Institute for Data Science, Department of Biomedical Engineering, Johns Hopkins University, 3400 N. Charles Street, Baltimore MD 21218, USA*

1

Optimization Landscape of Neural Networks

Abstract: Many tasks in machine learning involve solving a convex optimization problem, which significantly facilitates the analysis of properties of the resulting algorithms such as optimality, robustness and generalization. An important challenge in training neural networks is that the associated optimization problem is non-convex, which complicates the analysis because global optima can be difficult to characterize and the optimization landscape can also include spurious local minima and saddle points. As a consequence, different algorithms might find different weights depending on initialization, parameter tuning, etc. Despite this challenge, existing algorithms routinely converge to good solutions in practice, which suggests that the landscape might be simpler than expected, at least for certain classes of networks.

This chapter summarizes recent advances on the analysis of the optimization landscape of neural network training. We first review classical results for linear networks trained with the squared loss and without regularization. Such results show that under certain conditions on the input-output data spurious local minima are guaranteed not to exist, i.e. critical points are either saddle points or global minima. Moreover, the globally optimal weights can be found by factorizing certain matrices obtained from the input-output covariance matrices. We then review recent results for deep networks with parallel structure, positively homogeneous network mapping and regularization, and trained with a convex loss. Such results show that the non-convex objective on the weights can be lower-bounded by a convex objective on the network mapping. Moreover, when the network is sufficiently wide, local minima of the non-convex objective that satisfy a certain condition yield global minima of both the non-convex and convex objectives, and that there is always a non-increasing path to a global minimizer from any initialization.

[a] From *Elliptic Cohomology: Geometry, Applications, and Higher Chromatic Analogues*, edited by Haynes R. Miller and Douglas C. Ravenel

Keywords: deep learning, non-convex optimization

1.1 Introduction

Many machine learning tasks involve solving an optimization problem of the form

$$\min_{\mathbf{W}} \mathcal{L}(\Phi(\mathbf{X}, \mathbf{W}), \mathbf{Y}) + \lambda\Theta(\mathbf{W}). \tag{1.1}$$

For example, in the case of classification, $\mathcal{L}(\Phi(\mathbf{X}, \mathbf{W}), \mathbf{Y})$ is a *loss function* that measures the agreement between the true matrix of labels, $\mathbf{Y}$, and the matrix of predicted labels, $\Phi(\mathbf{X}, \mathbf{W})$, where $\mathbf{X}$ is the input data matrix, $\mathbf{W}$ represents the classifier parameters, $\Theta(\mathbf{W})$ is a *regularization function* designed to prevent overfitting[1], and $\lambda > 0$ is a parameter that controls the trade-off between the loss function and the regularization function. Another example is regression, where the setting is essentially the same, except that $\mathbf{Y}$ is typically continued valued, while in classification $\mathbf{Y}$ is categorical.

Some machine learning problems, such as linear regression, support vector machines, ℓ_1 minimization, and nuclear norm minimization, involve solving a *convex optimization problem*, where both the loss and the regularization functions are assumed to be convex functions of $\mathbf{W}$. For example, in linear regression with the squared loss and Tikhonov regularization, we have[2]

$$\mathcal{L}(\Phi(\mathbf{X}, \mathbf{W}), \mathbf{Y}) = \|\mathbf{Y} - \mathbf{W}^\top\mathbf{X}\|_F^2 \qquad \text{and} \qquad \Theta(\mathbf{W}) = \|\mathbf{W}\|_F^2. \tag{1.2}$$

When the optimization problem is convex, non-global local minima and saddle points are guaranteed not to exist, which significantly facilitates the analysis of optimization algorithms, especially the study of their convergence to a global minimizer. In addition, convexity allows one to analyze properties of the resulting machine learning algorithm, such as robustness and generalization, without having to worry about the particulars of the optimization method, such as initialization, step size (learning rate), etc., as the global optima are easily characterized and many optimization schemes exist which provide guaranteed convergence to a global minimizer.[3]

Unfortunately, many other machine learning problems – particularly those that seek to learn an appropriate representation of features directly from the

[1] Note that Θ could also depend on $\mathbf{X}$, but we will omit this for notational simplicity.

[2] The squared loss between vectors $\mathbf{y}$ and $\mathbf{z}$ is $\mathcal{L}(\mathbf{y}, \mathbf{z}) = \|\mathbf{y} - \mathbf{z}\|_2^2$. Here we consider a dataset $(\mathbf{X}, \mathbf{Y})$ with m training examples arranged as columns of a matrix $(\mathbf{X}, \mathbf{Y}) = \left([\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}], [\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(m)}]\right)$. The sum of the squared losses over the training examples $\sum_{i=1}^m \|\mathbf{y}^{(i)} - \mathbf{z}^{(i)}\|_2^2$ becomes the Frobenius norm $\|\mathbf{Y} - \mathbf{Z}\|_F^2$.

[3] For convex learning problems, the convergence of the optimization method to a global minimum does depend on initialization and parameter tuning, but the analysis of generalization does not.

data – with principal component analysis (PCA), nonnegative matrix factorization, sparse dictionary learning, tensor factorization and deep learning being well-known examples – involve solving a *non-convex optimization problem* of the form:

$$\min_{\{\mathbf{W}^{[l]}\}_{l=1}^{L}} \mathcal{L}(\Phi(\mathbf{X}, \mathbf{W}^{[1]}, \ldots, \mathbf{W}^{[L]}), \mathbf{Y}) + \lambda\Theta(\mathbf{W}^{[1]}, \ldots, \mathbf{W}^{[L]}), \tag{1.3}$$

where Φ is an arbitrary, convexity destroying mapping. In PCA, for example, the goal is to factorize a given data matrix $\mathbf{Y}$ as the product of two matrices $\mathbf{W}^{[1]}$ and $\mathbf{W}^{[2]}$, subject to the constraint that the columns of $\mathbf{W}^{[1]}$ are orthonormal. In this case, $\Phi(\mathbf{X}, \mathbf{W}^{[1]}, \mathbf{W}^{[2]}) = \mathbf{W}^{[1]}\mathbf{W}^{[2]\top}$ and Θ enforces the orthogonality constraints $\mathbf{W}^{[1]\top}\mathbf{W}^{[1]} = \mathbf{I}$, both of which make the optimization problem non-convex. Similarly, in deep neural network training, the output of the network is typically generated by applying an alternating series of linear and non-linear functions to the input data:

$$\Phi(\mathbf{X}, \mathbf{W}^{[1]}, \ldots, \mathbf{W}^{[L]}) = \psi_L(\mathbf{W}^{[L]}\psi_{L-1}(\mathbf{W}^{[L-1]} \cdots \psi_2(\mathbf{W}^{[2]}\psi_1(\mathbf{W}^{[1]}\mathbf{X})) \cdots)), \tag{1.4}$$

where each $\mathbf{W}^{[l]}$ is an appropriately sized matrix that contains the connection weights between layers $l-1$ and l of the network, and the $\psi_l(\cdot)$ functions apply some form of non-linearity after each matrix multiplication, e.g., a sigmoid function, rectified linear unit (ReLU), max-pooling.[4]

For a very small number of non-convex problems, e.g., PCA, one is fortunate, and a global minimizer can be found in closed form. For other problems, e.g., ℓ_0 minimization, rank minimization, and low-rank matrix completion, one can replace the non-convex objective by a convex surrogate and show that under certain conditions the solutions to both problems are the same.[5] In most cases, however, the optimal solutions cannot be computed in closed form, and a good convex surrogate may not be easy to find. This presents significant challenges to existing optimization algorithms – including (but certainly not limited to) alternating minimization, gradient descent, stochastic gradient descent, block coordinate descent, back-propagation, and quasi-Newton methods – which are typically only guaranteed to converge to a critical point of the objective function (Mairal et al., 2010; Rumelhart et al., 1988; Wright and Nocedal, 1999; Xu and Yin, 2013). As the set of critical points for non-convex problems includes not only global minima, but also spurious (non-global) local minima, local maxima, saddle points and sad-

[4] Here we have shown the linear operations to be simple matrix multiplications to simplify notation, but this easily generalizes to other linear operators (e.g., convolution) and affine operators (i.e., using bias terms).

[5] See e.g., Donoho (2006); Candès and Tao (2010) for the relationships between ℓ_0 and ℓ_1 minimization.

dle plateaus, as illustrated in Figure 1.1, the non-convexity of the problem leaves the model somewhat ill-posed in the sense that it is not just the model formulation that is important but also implementation details, such as how the model is initialized and particulars of the optimization algorithm, which can have a significant impact on the performance of the model.

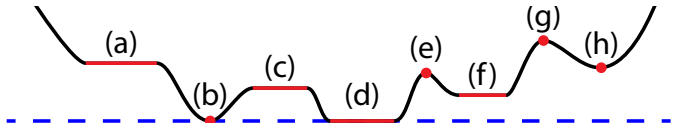

Figure 1.1 Example critical points of a non-convex function (shown in red). (a,c) Plateaus. (b,d) Global minima. (e,g) Local maxima. (f,h) Local minima.

Despite these challenges, optimization methods that combine Backpropagation (Werbos, 1974) with variants of Stochastic Gradient Descent (Robbins and Monro, 1951), such as Nesterov Accelerated Gradient (Nesterov, 1983), Adam (Kingma and Ba, 2014) and Adagrad (Duchi et al., 2017), appear to routinely yield good solutions for training deep networks. Recent work attempting to understand this phenomenon can be broadly classified along three main themes:

(i) *Benign optimization landscape*: While the optimization problem in (1.3) is not convex for deep network training, there are certain classes of networks for which there are no spurious local minima (Baldi and Hornik, 1989; Kawaguchi, 2016; Haeffele and Vidal, 2017), local minima concentrate near the global optimum (Choromanska et al., 2015), or critical points are more likely to be saddle points rather than spurious local minima Dauphin et al. (2014). A similar benign landscape has also been observed for non-convex problems arising in phase retrieval (Sun et al., 2016), dictionary learning (Sun et al., 2017) and blind deconvolution (Zhang et al., 2018).

(ii) *Optimization dynamics lead to global optima*: In addition to the study of the landscape of the learning objective, there has also been work focused on how specific algorithms (largely gradient descent-based) perform when optimizing neural networks. For example, Gori and Tesi (1991, 1992) show that gradient descent generally finds a global minimizer for linearly separable data. More generally, work has also shown that if the optimization landscape satisfies the *strict saddle property* (where the Hessian evaluated at every saddle point has a sufficiently negative eigenvalue) then gradient descent (and many other first-order descent techniques) is guaranteed to converge to a local minimum and not get stuck in saddle points

(Ge et al., 2015; Lee et al., 2019). Using these results, it has been shown that gradient descent converges to a global minimum (Kawaguchi, 2016; Nouiehed and Razaviyayn, 2018; Zhu et al., 2019) for linear neural networks that satisfy the strict saddle conditions. Unfortunately, however, the strict saddle property does not typically hold for non-linear neural networks. Nevertheless, several recent studies have shown that if the network is sufficiently large, then under certain conditions gradient descent will converge at a linear rate to global minimizers. However, the necessary conditions are potentially quite strict. Moreover, it is unclear if such results can be generalized to other formulations that include regularization on the network parameters (Du et al., 2019; Allen-Zhu et al., 2019).

(iii) *Implicit bias of the optimization algorithm*: Another possible explanation for the observed success of deep learning is that the optimization algorithm either explores only a subset of the landscape (depending on properties of the data or initialization of the algorithm) or automatically induces a regularizer which avoids spurious local minima. For example, Gunasekar et al. (2017, 2018a,b) show that gradient descent applied to certain classes of linear networks automatically induces a bias towards solutions that minimize a certain norm. Further, Arora et al. (2019) extend this idea to deep linear models are argue that depth in linear networks trained with gradient descent induces a low-rank regularization through the dynamics gradient descent. Further, other optimization techniques such as dropout (Srivastava et al., 2014), which adds stochastic noise by randomly setting the output of neurons to zero during training, have been shown to induce low-rank structures in the solution (Cavazza et al., 2018; Mianjy et al., 2018; Pal et al., 2020).

This chapter concentrates on the first theme by presenting an overview of the study of the optimization landscape of neural network training. In Section 1.3 we study the landscape of linear networks. Specifically, in Section 1.3.1 we review classical results from Baldi and Hornik (1989) for single-hidden layer linear networks trained using the squared loss, which show that under certain conditions on the network width and the input-output data every critical point is either a global minimum or a saddle point, as well as recent results from (Nouiehed and Razaviyayn, 2018; Zhu et al., 2019) which show that all saddle points are strict (i.e., at least one eigenvalue of the Hessian is negative). Moreover, (Baldi and Hornik, 1989; Nouiehed and Razaviyayn, 2018; Zhu et al., 2019) also show that the globally optimal weights can be found by factorizing certain matrix obtained from the input-output covariance matrices. Then, in Section 1.3.2 we review the work of

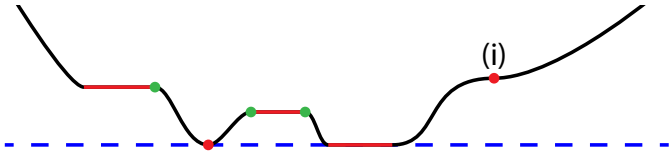

Figure 1.2 Guaranteed properties of the proposed framework. Starting from any initialization, a non-increasing path exists to a global minimizer. Starting from points on a plateau, a simple "sliding" method exists to find the edge of the plateau (green points).

Kawaguchi (2016), which extends these results to networks of any depth and width by showing that critical points are also either global minima or saddle points. In addition, Kawaguchi (2016) shows that saddle points of networks with one hidden layer are strict, but networks with two or more layers can have "bad" (non-strict) saddle points.

In Section 1.4 we study the landscape of nonlinear networks. Specifically, we review recent results from Haeffele and Vidal (2017, 2019) that study conditions under which the optimization landscape for the non-convex optimization problem in (1.3) is such that *all critical points are either global minimizers or saddle points/plateaus*, as shown in Figure 1.2. Their results show that if the network size is large enough and the functions Φ and Θ are *sums of positively homogeneous functions of the same degree*, then a monotonically decreasing path to a global minimizer exists from every point.

1.2 Basics of Statistical Learning

Let $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$ be a pair of random variables drawn from an unknown distribution $\mathbb{P}_{(\mathbf{x},\mathbf{y})}$, where $\mathcal{X}$ is called the input space and $\mathcal{Y}$ the output space. Assume we wish to predict $\mathbf{y}$ from an observation about $\mathbf{x}$ by finding a hypothesis $\hat{f} \in \mathcal{Y}^{\mathcal{X}}$, i.e. $\hat{f} : \mathcal{X} \to \mathcal{Y}$, that minimizes the *expected loss* or *risk*

$$\min_{f \in \mathcal{F}} \big[\mathcal{R}(f) \doteq \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathcal{L}(f(\mathbf{x}), \mathbf{y})]\big]. \tag{1.5}$$

Here $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$ is the space of hypotheses (e.g., the space of linear functions or the space of measurable functions from $\mathcal{X}$ to $\mathcal{Y}$) and $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to [0, \infty]$ is a loss function, where $\mathcal{L}(f(\mathbf{x}), \mathbf{y})$ gives the cost of predicting $\mathbf{y}$ as $f(\mathbf{x})$ (e.g., the zero-one loss $1_{\mathbf{y} \neq f(\mathbf{x})}$ for classification or the squared loss $\|\mathbf{y} - f(\mathbf{x})\|_2^2$ for regression). The smallest expected risk $\mathcal{R}(\hat{f})$ is called the *Bayes error*.

Since $\mathbb{P}_{(\mathbf{x},\mathbf{y})}$ is unknown, $\hat{f}$ and $\mathcal{R}(\hat{f})$ cannot be computed. Instead, we assume we are given a training set $\mathcal{S} = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^{m}$ of i.i.d. samples from

$\mathbb{P}_{(\mathbf{x},\mathbf{y})}$ and seek to find a hypothesis $\hat{f}_{\mathcal{F},\mathcal{S}}$ that minimizes the *empirical risk*

$$\min_{f\in\mathcal{F}} \big[\mathcal{R}_{\mathcal{S}}(f) \doteq \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}(f(\mathbf{x}^{(i)}), \mathbf{y}^{(i)})\big]. \tag{1.6}$$

Since the objective functions in (1.5) and (1.6) are different, a priori there is no guarantee that $\hat{f}_{\mathcal{F},\mathcal{S}}$ or its risk, $\mathcal{R}(\hat{f}_{\mathcal{F},\mathcal{S}})$, will be close to $\hat{f}$ or $\mathcal{R}(\hat{f})$, respectively. This leads to the question of *generalization*, which seeks to understand the performance of $\hat{f}_{\mathcal{F},\mathcal{S}}$ not only on the training set $\mathcal{S}$, but on the entire population. In principle, we could use the error $\mathcal{R}(\hat{f}_{\mathcal{F},\mathcal{S}}) - \mathcal{R}(\hat{f})$ to assess the quality of $\hat{f}_{\mathcal{F},\mathcal{S}}$. However, since $\hat{f}_{\mathcal{F},\mathcal{S}}$ depends on the data $\mathcal{S}$, $\hat{f}_{\mathcal{F},\mathcal{S}}$ is a random function and $\mathcal{R}(\hat{f}_{\mathcal{F},\mathcal{S}})$ is a random variable. While we could use the expectation of $\mathcal{R}(\hat{f}_{\mathcal{F},\mathcal{S}})$ with respect to the data, $\mathbb{E}_{\mathcal{S}}[\mathcal{R}(\hat{f}_{\mathcal{F},\mathcal{S}}) - \mathcal{R}(\hat{f})]$, or verify if $\hat{f}_{\mathcal{F},\mathcal{S}}$ is *universally consistent*, i.e. check if

$$\lim_{m\to\infty} \mathcal{R}(\hat{f}_{\mathcal{F},\mathcal{S}}) = \mathcal{R}(\hat{f}) \qquad \text{almost surely,} \tag{1.7}$$

both approaches are difficult to implement because $\mathbb{P}_{(\mathbf{x},\mathbf{y})}$ is unknown.

To address this issue, a common practice is to decompose the error as

$$\begin{aligned}\mathcal{R}(\hat{f}_{\mathcal{F},\mathcal{S}}) - \mathcal{R}(\hat{f}) = \big(\mathcal{R}(\hat{f}_{\mathcal{F},\mathcal{S}}) - \mathcal{R}_{\mathcal{S}}(\hat{f}_{\mathcal{F},\mathcal{S}})\big) + \big(\mathcal{R}_{\mathcal{S}}(\hat{f}_{\mathcal{F},\mathcal{S}}) - \mathcal{R}_{\mathcal{S}}(\hat{f})\big) \\ + \big(\mathcal{R}_{\mathcal{S}}(\hat{f}) - \mathcal{R}(\hat{f})\big)\end{aligned} \tag{1.8}$$

and use the fact that the second term is nonpositive and the third term has zero expectation to arrive at an upper bound on the expected error

$$\begin{aligned}\mathbb{E}_{\mathcal{S}}[\mathcal{R}(\hat{f}_{\mathcal{F},\mathcal{S}}) - \mathcal{R}(\hat{f})] &\le \mathbb{E}_{\mathcal{S}}[\mathcal{R}(\hat{f}_{\mathcal{F},\mathcal{S}}) - \mathcal{R}_{\mathcal{S}}(\hat{f}_{\mathcal{F},\mathcal{S}})] \\ &\le \mathbb{E}_{\mathcal{S}}[\sup_{f\in\mathcal{F}} \mathcal{R}(f) - \mathcal{R}_{\mathcal{S}}(f)].\end{aligned} \tag{1.9}$$

As the bound on the right hand side may not be easily computable, a typical approach is to derive an easier to compute upper bound, say $\Theta(f)$, and then solve the *regularized empirical risk minimization* problem

$$\min_{f\in\mathcal{F}} \mathcal{R}_{\mathcal{S}}(f) + \Theta(f). \tag{1.10}$$

In other words, rather than minimizing the empirical risk, $\mathcal{R}_{\mathcal{S}}(f)$, we usually minimize the regularized empirical risk, $\mathcal{R}_{\mathcal{S}}(f) + \Theta(f)$, in the hope of controlling the error $\mathbb{E}_{\mathcal{S}}[\mathcal{R}(\hat{f}_{\mathcal{F},\mathcal{S}}) - \mathcal{R}(\hat{f})]$.

Therefore, this chapter will focus on understanding the landscape of the optimization problem in (1.10), although we will also make connections with the optimization problem in (1.5) whenever possible (e.g., for single-hidden layer linear networks trained with the squared loss). We refer the reader to other chapters in this book for a study of the generalization properties.

1.3 Optimization Landscape of Linear Networks

In this section we study the landscape of linear networks trained using the squared loss. In Section 1.3.1 we show that under certain conditions every critical point of a single-hidden layer linear network is either a global minima or a strict saddle point, and that the globally optimal weights can be obtained using linear-algebraic methods. In Section 1.3.2 we show that critical points of linear networks with more than two layers are either a global minima or a saddle point, but saddle points may not be strict.

1.3.1 Single-Hidden Layer Linear Networks with Squared Loss and Fixed Size Regularization

Let us first consider the case of linear networks with n_0 inputs, n_2 outputs, and a single hidden layer with n_1 neurons. In this case, the hypothesis space $\mathcal{F}$ can be parametrized by the network weights $(\mathbf{W}^{[1]}, \mathbf{W}^{[2]}) = (\mathbf{U}, \mathbf{V})$ as[6]

$$\mathcal{F} = \{f \in \mathcal{Y}^{\mathcal{X}} : f(\mathbf{x}) = \mathbf{U}\mathbf{V}^\top \mathbf{x}, \text{ where } \mathbf{U} \in \mathbb{R}^{n_2 \times n_1} \text{ and } \mathbf{V} \in \mathbb{R}^{n_0 \times n_1}\}. \quad (1.11)$$

In this section, we study the optimization landscape for single-hidden layer linear networks trained using the squared loss, $\mathcal{L}(\mathbf{z}, \mathbf{y}) = \|\mathbf{y} - \mathbf{z}\|_2^2$. No regularization on the network weights is assumed, except that the network size n_1 is assumed to be known and sufficiently small relative to the input-output dimensions, i.e., $n_1 \leq \min\{n_0, n_2\}$. Under these assumptions, the problem of minimizing the expected risk reduces to[7]

$$\min_{\mathbf{U},\mathbf{V}} \big[\mathcal{R}(\mathbf{U}, \mathbf{V}) \doteq \mathbb{E}_{\mathbf{x},\mathbf{y}}[\|\mathbf{y} - \mathbf{U}\mathbf{V}^\top \mathbf{x}\|_2^2]\big]. \quad (1.12)$$

Letting $\mathbf{\Sigma_{xx}} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top] \in \mathbb{R}^{n_0 \times n_0}$, $\mathbf{\Sigma_{xy}} = \mathbb{E}[\mathbf{x}\mathbf{y}^\top] \in \mathbb{R}^{n_0 \times n_2}$, $\mathbf{\Sigma_{yx}} = \mathbb{E}[\mathbf{y}\mathbf{x}^\top] = \mathbf{\Sigma_{xy}}^\top \in \mathbb{R}^{n_2 \times n_0}$, and $\mathbf{\Sigma_{yy}} = \mathbb{E}[\mathbf{y}\mathbf{y}^\top] \in \mathbb{R}^{n_2 \times n_2}$, the expected risk can be rewritten as:

$$\mathcal{R}(\mathbf{U}, \mathbf{V}) = \text{trace}(\mathbf{\Sigma_{yy}} - 2\mathbf{\Sigma_{yx}}\mathbf{V}\mathbf{U}^\top + \mathbf{U}\mathbf{V}^\top \mathbf{\Sigma_{xx}}\mathbf{V}\mathbf{U}^\top). \quad (1.13)$$

Consider now the problem of minimizing the empirical risk

$$\min_{\mathbf{U},\mathbf{V}} \Big[\mathcal{R}_{\mathcal{S}}(\mathbf{U}, \mathbf{V}) = \frac{1}{m}\sum_{i=1}^{m} \|\mathbf{y}^{(i)} - \mathbf{U}\mathbf{V}^\top \mathbf{x}^{(i)}\|_2^2 = \frac{1}{m}\|\mathbf{Y} - \mathbf{U}\mathbf{V}^\top \mathbf{X}\|_F^2\Big], \quad (1.14)$$

where $\mathcal{S} = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^m$ is the training set, and $\mathbf{X} = [\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}]$ and

[6] For simplicity of notation, if we only have two groups of parameters we will use $(\mathbf{U}, \mathbf{V})$ rather than $(\mathbf{W}^{[1]}, \mathbf{W}^{[2]})$.

[7] With an abuse of notation, we will write the risk as a function of the network weights, i.e., $\mathcal{R}(\mathbf{U}, \mathbf{V})$, rather than as a function of the input-output map, i.e., $\mathcal{R}(f)$.

$\mathbf{Y} = [\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(m)}]$ are the input and output data matrices. It is easy to see that the empirical risk $\mathcal{R}_{\mathcal{S}}(\mathbf{U}, \mathbf{V})$ is equal to $\mathcal{R}(\mathbf{U}, \mathbf{V})$, except that the covariance matrices $\mathbf{\Sigma_{xx}}$, $\mathbf{\Sigma_{xy}}$ and $\mathbf{\Sigma_{yy}}$ need to be substituted by their empirical estimates $\frac{1}{m}\mathbf{X}\mathbf{X}^\top$, $\frac{1}{m}\mathbf{X}\mathbf{Y}^\top$ and $\frac{1}{m}\mathbf{Y}\mathbf{Y}^\top$, respectively. Therefore, in this case the analysis of the optimization landscape for both the expected and empirical risk can be done by analyzing the landscape of $\mathcal{R}(\mathbf{U}, \mathbf{V})$.

To motivate the analysis of the landscape of $\mathcal{R}(\mathbf{U}, \mathbf{V})$, let us first analyze the landscape of the risk as a function of the product of the weights, i.e., $\mathbf{Z} = \mathbf{U}\mathbf{V}^\top$, which is given by $\mathcal{R}(\mathbf{Z}) = \text{trace}(\mathbf{\Sigma_{yy}} - 2\mathbf{\Sigma_{yx}}\mathbf{Z}^\top + \mathbf{Z}\mathbf{\Sigma_{xx}}\mathbf{Z}^\top)$. When there is no constraint on $\mathbf{Z}$ (e.g., when $\mathbf{U}$ and $\mathbf{V}$ are full column rank), the risk is a convex function of $\mathbf{Z}$ and the first order condition for optimality is given by $\mathbf{Z}\mathbf{\Sigma_{xx}} = \mathbf{\Sigma_{yx}}$. Thus, if $\mathbf{\Sigma_{xx}}$ is invertible, the global minimum is unique and is given by $\mathbf{Z}^* = \mathbf{U}^*\mathbf{V}^{*\top} = \mathbf{\Sigma_{yx}}\mathbf{\Sigma_{xx}}^{-1}$. Of course, this provides a characterization of the optimal $\mathbf{Z}$, but not of the optimal $\mathbf{U}$ and $\mathbf{V}$. The challenge in characterizing the landscape of $\mathcal{R}(\mathbf{U}, \mathbf{V})$ is hence to understand the effect of the low rank constraint $n_1 \leq \min\{n_0, n_2\}$ or to consider the possibility that critical points for $\mathbf{U}$ or $\mathbf{V}$ might not be low-rank.

The following lemma characterizes properties of the critical points of $\mathcal{R}(\mathbf{U}, \mathbf{V})$. The original statements and proofs for these results can be found in Baldi and Hornik (1989). Here we provide a unified treatment for both the expected and empirical risk, as well as alternative derivations.

Lemma 1.1 *If $(\mathbf{U}, \mathbf{V})$ is a critical point of $\mathcal{R}$, then*

$$\mathbf{U}\mathbf{V}^\top\mathbf{\Sigma_{xx}}\mathbf{V} = \mathbf{\Sigma_{yx}}\mathbf{V} \quad \textit{and} \quad \mathbf{\Sigma_{xx}}\mathbf{V}\mathbf{U}^\top\mathbf{U} = \mathbf{\Sigma_{xy}}\mathbf{U}. \tag{1.15}$$

Moreover, if $\mathbf{\Sigma_{xx}}$ is invertible, then the following three properties hold.

(i) *If $\mathbf{V}$ is full column rank, then $\mathbf{U} = \mathbf{\Sigma_{yx}}\mathbf{V}(\mathbf{V}^\top\mathbf{\Sigma_{xx}}\mathbf{V})^{-1}$.*
(ii) *If $\mathbf{U}$ is full column rank, then $\mathbf{V} = \mathbf{\Sigma_{xx}}^{-1}\mathbf{\Sigma_{xy}}\mathbf{U}(\mathbf{U}^\top\mathbf{U})^{-1}$.*
(iii) *Let $\mathbf{\Sigma} = \mathbf{\Sigma_{yx}}\mathbf{\Sigma_{xx}}^{-1}\mathbf{\Sigma_{xy}}$. If $\mathbf{U}$ is full column rank and $\mathbf{P_U} = \mathbf{U}(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top$, then $\mathbf{\Sigma}\mathbf{P_U} = (\mathbf{\Sigma}\mathbf{P_U})^\top = \mathbf{P_U}\mathbf{\Sigma}$.*

Proof The gradient of $\mathcal{R}$ w.r.t. $\mathbf{U}$ is given by

$$\frac{\partial \mathcal{R}}{\partial \mathbf{U}} = -2(\mathbf{\Sigma_{yx}} - \mathbf{U}\mathbf{V}^\top\mathbf{\Sigma_{xx}})\mathbf{V} = \mathbf{0} \implies \mathbf{U}\mathbf{V}^\top\mathbf{\Sigma_{xx}}\mathbf{V} = \mathbf{\Sigma_{yx}}\mathbf{V}. \tag{1.16}$$

Therefore, when $\mathbf{\Sigma_{xx}}$ is invertible and $\mathbf{V}$ is full column rank we have

$$\mathbf{U} = \mathbf{\Sigma_{yx}}\mathbf{V}(\mathbf{V}^\top\mathbf{\Sigma_{xx}}\mathbf{V})^{-1}. \tag{1.17}$$

as claimed in (i). On the other hand, the gradient of $\mathcal{R}$ w.r.t. $\mathbf{V}$ is given by

$$\frac{\partial \mathcal{R}}{\partial \mathbf{V}} = -2(\mathbf{\Sigma_{xy}} - \mathbf{\Sigma_{xx}}\mathbf{V}\mathbf{U}^\top)\mathbf{U} = \mathbf{0} \implies \mathbf{\Sigma_{xx}}\mathbf{V}\mathbf{U}^\top\mathbf{U} = \mathbf{\Sigma_{xy}}\mathbf{U}. \tag{1.18}$$

Therefore, when $\mathbf{\Sigma_{xx}}$ is invertible and $\mathbf{U}$ is full column rank we have

$$\mathbf{V} = \mathbf{\Sigma_{xx}^{-1}\Sigma_{xy}U}(\mathbf{U}^\top\mathbf{U})^{-1}, \tag{1.19}$$

as claimed in (ii). Moreover, notice that

$$\mathbf{UV}^\top = \mathbf{U}(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top\mathbf{\Sigma_{yx}\Sigma_{xx}^{-1}} = \mathbf{P_U\Sigma_{yx}\Sigma_{xx}^{-1}}. \tag{1.20}$$

Combining this with the first equation in (1.15) we obtain

$$\mathbf{UV}^\top\mathbf{\Sigma_{xx}VU}^\top = \mathbf{\Sigma_{yx}VU}^\top \tag{1.21}$$

$$\mathbf{P_U\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xx}\Sigma_{xx}^{-1}\Sigma_{xy}P_U} = \mathbf{\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}P_U} \tag{1.22}$$

$$\mathbf{P_U\Sigma P_U} = \mathbf{\Sigma P_U}. \tag{1.23}$$

As a consequence, $\mathbf{\Sigma P_U} = (\mathbf{\Sigma P_U})^\top = \mathbf{P_U\Sigma}$, as claimed in (iii). □

Baldi and Hornik (1989) used these properties to show that, under certain conditions, the expected loss has a unique global minimum (up to an equivalence) and that all other critical points are saddle points. Recently, Nouiehed and Razaviyayn (2018); Zhu et al. (2019) extended such results to show that all saddle points are strict. Recall that a critical point is a strict saddle if the Hessian evaluated at this point has a strictly negative eigenvalue, indicating that not only it is not a local minimum, but also the objective function has a negative curvature at this point. The following theorem characterizes the landscape of the risk functional for single-hidden layer linear networks.

Theorem 1.2 *Assume $\mathbf{\Sigma_{xx}}$ is invertible and $\mathbf{\Sigma} = \mathbf{\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}}$ is full rank with n_2 distinct eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_{n_2}$. Let $\mathbf{\Sigma} = \mathbf{Q\Lambda Q}^\top$ the eigendecomposition of $\mathbf{\Sigma}$, where the columns of $\mathbf{Q} \in \mathbb{R}^{n_2\times n_2}$ contain the corresponding eigenvectors. Let $\mathbf{Q}_\mathcal{J}$ denote the submatrix of $\mathbf{Q}$ whose columns are indexed by $\mathcal{J}$. Then, the following holds.*

- *If $\mathbf{U}$ is full column rank, the set of critical points of $\mathcal{R}(\mathbf{U},\mathbf{V})$ is given by*

$$\mathbf{U} = \mathbf{Q}_\mathcal{J}\mathbf{C} \quad \textit{and} \quad \mathbf{V} = \mathbf{\Sigma_{xx}^{-1}\Sigma_{xy}Q}_\mathcal{J}\mathbf{C}^{-\top}, \tag{1.24}$$

  where $\mathcal{J}$ is an ordered subset of $[n_2]$ of cardinality n_1, i.e. $\mathcal{J} \subset [n_2]$ and $|\mathcal{J}| = n_1$, and $\mathbf{C} \in \mathbb{R}^{n_1\times n_1}$ is an arbitrary invertible matrix.
- *If $\mathbf{U}$ is full column rank, then critical points with $\mathcal{J} \neq [n_1]$ are strict saddles, i.e., the Hessian evaluated at this point has a strictly negative eigenvalue, while critical points with $\mathcal{J} = [n_1]$ are global minima. Specifi-*

cally, the set of global minima $(\mathbf{U}, \mathbf{V})$ *of the risk* $\mathcal{R}$ *is given by*

$$\begin{aligned} \mathbf{U} &= \mathbf{Q}_{1:n_1}\mathbf{C}, \\ \mathbf{V} &= \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\boldsymbol{\Sigma}_{\mathbf{xy}}\mathbf{Q}_{1:n_1}\mathbf{C}^{-\top}, \\ \mathbf{U}\mathbf{V}^\top &= \mathbf{Q}_{1:n_1}\mathbf{Q}_{1:n_1}^\top\boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}, \end{aligned} \tag{1.25}$$

where $\mathbf{C}$ *is an arbitrary invertible matrix.*

- *If* $\mathbf{U}$ *is rank deficient, then any critical point is a strict saddle.*

Proof The proof of part (i) is based on Baldi and Hornik (1989), while the proofs of parts (ii) and (iii) are based on Nouiehed and Razaviyayn (2018) and Zhu et al. (2019). For part (i), note that

$$\mathbf{P}_{\mathbf{Q}^\top\mathbf{U}} = \mathbf{Q}^\top\mathbf{U}(\mathbf{U}^\top\mathbf{Q}\mathbf{Q}^\top\mathbf{U})^{-1}\mathbf{U}^\top\mathbf{Q} = \mathbf{Q}^\top\mathbf{U}(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top\mathbf{Q} = \mathbf{Q}^\top\mathbf{P}_{\mathbf{U}}\mathbf{Q},$$

which together with Lemma 1.1 (iii) gives

$$\mathbf{P}_{\mathbf{Q}^\top\mathbf{U}}\boldsymbol{\Lambda} = \mathbf{Q}^\top\mathbf{P}_{\mathbf{U}}\mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^\top\mathbf{Q} = \mathbf{Q}^\top\mathbf{P}_{\mathbf{U}}\boldsymbol{\Sigma}\mathbf{Q} = \mathbf{Q}^\top\boldsymbol{\Sigma}\mathbf{P}_{\mathbf{U}}\mathbf{Q} = \boldsymbol{\Lambda}\mathbf{P}_{\mathbf{Q}^\top\mathbf{U}}.$$

Since $\boldsymbol{\Lambda}$ is a diagonal matrix with diagonal entries $\lambda_1 > \lambda_2 > \cdots > \lambda_{n_2} > 0$, it follows that $\mathbf{P}_{\mathbf{Q}^\top\mathbf{U}}$ is also a diagonal matrix. Notice that $\mathbf{P}_{\mathbf{Q}^\top\mathbf{U}}$ is an orthogonal projector of rank n_1, i.e., it has n_1 eigenvalues 1 and $n_2 - n_1$ eigenvalues 0. Therefore, $\mathbf{P}_{\mathbf{Q}^\top\mathbf{U}} = \mathbf{I}_{\mathcal{J}}\mathbf{I}_{\mathcal{J}}^\top$, where $\mathcal{J} \subset [n_2]$ is an ordered subset of $[n_2]$ with cardinality $|\mathcal{J}| = n_1$. Here, we denote by $\mathbf{I}_{\mathcal{J}}$ the submatrix of the identity matrix $\mathbf{I}$ obtained by only keeping the columns indexed by $\mathcal{J}$. It follows that

$$\mathbf{P}_{\mathbf{U}} = \mathbf{Q}\mathbf{P}_{\mathbf{Q}^\top\mathbf{U}}\mathbf{Q}^\top = \mathbf{Q}\mathbf{I}_{\mathcal{J}}\mathbf{I}_{\mathcal{J}}^\top\mathbf{Q}^\top = \mathbf{Q}_{\mathcal{J}}\mathbf{Q}_{\mathcal{J}}^\top,$$

which implies that $\mathbf{U}$ and $\mathbf{Q}_{\mathcal{J}}$ have the same column spaces. Thus, there exists an invertible $n_1 \times n_1$ matrix $\mathbf{C}$ such that $\mathbf{U} = \mathbf{Q}_{\mathcal{J}}\mathbf{C}$. Now according to Lemma 1.1 (ii), we have $\mathbf{V} = \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\boldsymbol{\Sigma}_{\mathbf{xy}}\mathbf{Q}_{\mathcal{J}}\mathbf{C}^{-\top}$.

We now prove the first statement in part (ii), i.e., for any $\mathcal{J} \neq [n_1]$, the corresponding critical point has strictly negative curvature. Towards that

goal, standard computations give the Hessian quadrature form[8]

$$\nabla^2\mathcal{R}(\mathbf{U},\mathbf{V})[\Delta,\Delta] = \|(\mathbf{U}\Delta_{\mathbf{V}}^\top+\Delta_{\mathbf{U}}\mathbf{V}^\top)\mathbf{\Sigma}_{\mathbf{xx}}^{1/2}\|_F^2+2\langle\Delta_{\mathbf{U}}\Delta_{\mathbf{V}}^\top,\mathbf{U}\mathbf{V}^\top\mathbf{\Sigma}_{\mathbf{xx}}-\mathbf{\Sigma}_{\mathbf{yx}}\rangle. \tag{1.26}$$

for any $\Delta = (\Delta_{\mathbf{U}}, \Delta_{\mathbf{V}}) \in \mathbb{R}^{n_2\times n_1}\times\mathbb{R}^{n_0\times n_1}$.

Since $\mathcal{J}\neq[n_1]$, there exists $k\le n_1$ such that $k\notin\mathcal{J}$. Let J be the largest element of $\mathcal{J}$, and choose $\Delta_{\mathbf{U}} = \mathbf{q}_k\mathbf{e}_{n_1}^\top\mathbf{C}$ and $\Delta_{\mathbf{V}} = \mathbf{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{\Sigma}_{\mathbf{xy}}\mathbf{q}_k\mathbf{e}_{n_1}^\top\mathbf{C}^{-\top}$, where $\mathbf{q}_k$ is the k-th column of $\mathbf{Q}$ and $\mathbf{e}_{n_1}$ is the n_1-th standard basis vector of appropriate dimension, i.e., all entries of $\mathbf{e}_{n_1}\in\mathbb{R}^{n_1}$ are zero except for the last entry which is equal to 1. The first term in (1.26) reduces to

$$\begin{aligned}
\|(\mathbf{U}\Delta_{\mathbf{V}}^\top+\Delta_{\mathbf{U}}\mathbf{V}^\top)\mathbf{\Sigma}_{\mathbf{xx}}^{1/2}\|_F^2 &= \|\mathbf{Q}_{\mathcal{J}}\mathbf{e}_{n_1}\mathbf{q}_k^\top\mathbf{\Sigma}_{\mathbf{yx}}\mathbf{\Sigma}_{\mathbf{xx}}^{-1/2}+\mathbf{q}_k\mathbf{e}_{n_1}^\top\mathbf{Q}_{\mathcal{J}}^\top\mathbf{\Sigma}_{\mathbf{yx}}\mathbf{\Sigma}_{\mathbf{xx}}^{-1/2}\|_F^2\\
&= \|\mathbf{q}_J\mathbf{q}_k^\top\mathbf{\Sigma}_{\mathbf{yx}}\mathbf{\Sigma}_{\mathbf{xx}}^{-1/2}\|_F^2+\|\mathbf{q}_k\mathbf{q}_J^\top\mathbf{\Sigma}_{\mathbf{yx}}\mathbf{\Sigma}_{\mathbf{xx}}^{-1/2}\|_F^2\\
&\le \|\mathbf{q}_k^\top\mathbf{\Sigma}_{\mathbf{yx}}\mathbf{\Sigma}_{\mathbf{xx}}^{-1/2}\|_2^2+\|\mathbf{q}_J^\top\mathbf{\Sigma}_{\mathbf{yx}}\mathbf{\Sigma}_{\mathbf{xx}}^{-1/2}\|_2^2\\
&= \lambda_k+\lambda_J.
\end{aligned}$$

Similarly, the second term in (1.26) can be computed as

$$\begin{aligned}
\langle\Delta_{\mathbf{U}}\Delta_{\mathbf{V}}^\top,\mathbf{U}\mathbf{V}^\top\mathbf{\Sigma}_{\mathbf{xx}}-\mathbf{\Sigma}_{\mathbf{yx}}\rangle &= \langle\mathbf{q}_k\mathbf{q}_k^\top\mathbf{\Sigma}_{\mathbf{yx}}\mathbf{\Sigma}_{\mathbf{xx}}^{-1},\mathbf{Q}_{\mathcal{J}}\mathbf{Q}_{\mathcal{J}}^\top\mathbf{\Sigma}_{\mathbf{yx}}-\mathbf{\Sigma}_{\mathbf{yx}}\rangle\\
&= -\langle\mathbf{q}_k\mathbf{q}_k^\top\mathbf{\Sigma}_{\mathbf{yx}}\mathbf{\Sigma}_{\mathbf{xx}}^{-1},\mathbf{\Sigma}_{\mathbf{yx}}\rangle = -\lambda_k.
\end{aligned}$$

Therefore, since $k\le n_1$ and $J>n_1$, we have

$$\nabla^2\mathcal{R}(\mathbf{U},\mathbf{V})[\Delta,\Delta] = \lambda_k+\lambda_J-2\lambda_k = \lambda_J-\lambda_k<0.$$

As a consequence, all critical points with $\mathcal{J}\neq[n_1]$ are strict saddles, and all critical points with $\mathcal{J}=[n_1]$ are a global minimum.

To show part (iii), notice that when $\mathrm{rank}(\mathbf{U})<n_1$, there exists a non-zero vector $\mathbf{a}\in\mathbb{R}^{n_1}$ such that $\mathbf{U}\mathbf{a}=0$. Since $\mathbf{\Sigma}_{\mathbf{xx}}$ and $\mathbf{\Sigma}=\mathbf{\Sigma}_{\mathbf{yx}}\mathbf{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{\Sigma}_{\mathbf{xy}}$ are assumed to be invertible, $\mathbf{\Sigma}_{\mathbf{xy}}$ must be full column rank n_2, hence $n_2\le n_0$. Since the rank of $\mathbf{U}\mathbf{V}^\top\mathbf{\Sigma}_{\mathbf{xx}}$ is at most the rank of $\mathbf{U}$ and $\mathbf{\Sigma}_{\mathbf{yx}}$ has rank n_2, we know $\mathbf{U}\mathbf{V}^\top\mathbf{\Sigma}_{\mathbf{xx}}-\mathbf{\Sigma}_{\mathbf{yx}}\neq0$. Without loss of generality, we assume its

[8] For a scalar function $f(\mathbf{W}):\mathbb{R}^{m\times n}\to\mathbb{R}$, its Hessian $\nabla^2 f(\mathbf{W})$ is a 4D tensor, or an $(mn)\times(mn)$ matrix if we vectorize the variable $\mathbf{W}$. An alternative way to represent the Hessian is by a bilinear form defined via $[\nabla^2 f(\mathbf{W})](\mathbf{A},\mathbf{B}) = \sum_{i,j,k,l}\frac{\partial^2 f(W)}{\partial W_{ij}\partial W_{kl}}A_{ij}B_{kl}$ for any $\mathbf{A},\mathbf{B}\in\mathbb{R}^{m\times n}$. When $n=1$, i.e., $\nabla^2 f(\mathbf{W})\in\mathbb{R}^{m\times m}$ and $\mathbf{A}$ and $\mathbf{B}$ are vectors, this bilinear form reduces to the standard matrix-vector multiplication $[\nabla^2 f(\mathbf{W})](\mathbf{A},\mathbf{B}) = \mathbf{A}^\top\nabla^2 f(\mathbf{W})\mathbf{B}$. Thus, the bilinear form allows us to represent the Hessian in a simple way even when the variable is a matrix. Also, without explicitly computing the eigenvalues of $\nabla^2 f(\mathbf{W})$, we know it has a strictly negative eigenvalue if we can find a direction $\Delta\in\mathbb{R}^{m\times n}$ such that $[\nabla^2 f(\mathbf{W})](\Delta,\Delta)<0$. Note that this quadratic form appears naturally in the Taylor expansion $f(\mathbf{W}+\Delta) = f(\mathbf{W})+\langle\nabla f(\mathbf{W}),\Delta\rangle+\frac{1}{2}[\nabla^2 f(\mathbf{W})](\Delta,\Delta)+\cdots$, which indeed provides a simple but very useful trick to compute $[\nabla^2 f(\mathbf{W})](\Delta,\Delta)$ as long as $f(\mathbf{W}+\Delta)$ can be easily expanded. For example, when $f(\mathbf{W}) = \frac{1}{2}\|\mathbf{Y}-\mathbf{W}\|_F^2$, we have $f(\mathbf{W}+\Delta) = \frac{1}{2}\|\mathbf{Y}-\mathbf{W}-\Delta\|_F^2 = \frac{1}{2}\|\mathbf{Y}-\mathbf{W}\|_F^2+\langle\mathbf{W}-\mathbf{Y},\Delta\rangle+\frac{1}{2}\|\Delta\|_F^2$, which implies that $[\nabla^2 f(\mathbf{W})](\Delta,\Delta) = \|\Delta\|_F^2$.

(i,j)-th entry is non-zero, and we choose $\Delta_{\mathbf{U}} = \mathbf{e}_i\mathbf{a}^\top$ and $\Delta_{\mathbf{V}} = \alpha\tilde{\mathbf{e}}_j\mathbf{a}^\top$, where $\mathbf{e}_i \in \mathbb{R}^{n_2}$ and $\tilde{\mathbf{e}}_j \in \mathbb{R}^{n_0}$ are the standard basis vectors, whose entries are equal to zero except for their i-th and j-th elements, respectively. With this, we compute the first term in (1.26) as

$$\|(\mathbf{U}\Delta_{\mathbf{V}}^\top + \Delta_{\mathbf{U}}\mathbf{V}^\top)\boldsymbol{\Sigma}_{\mathbf{xx}}^{1/2}\|_F^2 = \|(\alpha\mathbf{U}\mathbf{a}\tilde{\mathbf{e}}_j^\top + \mathbf{e}_i\mathbf{a}^\top\mathbf{V}^\top)\boldsymbol{\Sigma}_{\mathbf{xx}}^{1/2}\|_F^2 = \|\mathbf{e}_i\mathbf{a}^\top\mathbf{V}^\top\boldsymbol{\Sigma}_{\mathbf{xx}}^{1/2}\|_F^2$$

and the second term in (1.26) by

$$\begin{aligned}\langle\Delta_{\mathbf{U}}\Delta_{\mathbf{V}}^\top, \mathbf{U}\mathbf{V}^\top\boldsymbol{\Sigma}_{\mathbf{xx}} - \boldsymbol{\Sigma}_{\mathbf{yx}}\rangle &= \langle\alpha\|\mathbf{a}\|^2\mathbf{e}_i\tilde{\mathbf{e}}_j^\top, \mathbf{U}\mathbf{V}^\top\boldsymbol{\Sigma}_{\mathbf{xx}} - \boldsymbol{\Sigma}_{\mathbf{yx}}\rangle \\ &= \alpha\|\mathbf{a}\|^2(\mathbf{U}\mathbf{V}^\top\boldsymbol{\Sigma}_{\mathbf{xx}} - \boldsymbol{\Sigma}_{\mathbf{yx}})_{ij},\end{aligned}$$

from which it follows that

$$\nabla^2\mathcal{R}(\mathbf{U},\mathbf{V})[\Delta,\Delta] = \|\mathbf{e}_i\mathbf{a}^\top\mathbf{V}^\top\boldsymbol{\Sigma}_{\mathbf{xx}}^{1/2}\|_F^2 + 2\alpha\|\mathbf{a}\|^2(\mathbf{U}\mathbf{V}^\top\boldsymbol{\Sigma}_{\mathbf{xx}} - \boldsymbol{\Sigma}_{\mathbf{yx}})_{ij},$$

where the right hand side can always be negative by choosing appropriate α. Thus, $\mathcal{R}$ has negative curvature when $\mathbf{U}$ is rank deficient.

□

Theorem 1.2 implies that, under certain conditions on the input-output covariance matrices, $\boldsymbol{\Sigma}_{\mathbf{xx}}$, $\boldsymbol{\Sigma}_{\mathbf{yx}}$ and $\boldsymbol{\Sigma}_{\mathbf{yy}}$, both the expected and empirical risk of a single-hidden layer linear neural network with the squared loss have no spurious local minima and the saddle points are strict. This benign geometry ensures that a number of local search algorithms (such as gradient descent) converge to a global minimum when training a single-hidden layer linear neural network (Ge et al., 2015; Lee et al., 2019).

But what if the conditions in Theorem 1.2 are violated? When $\boldsymbol{\Sigma}_{\mathbf{xx}}$ is invertible but $\boldsymbol{\Sigma}$ is rank deficient, a more sophisticated analysis shows that one can still characterize the set of critical points with full column rank $\mathbf{U}$ as in (1.24) but with a slightly different form (Zhu et al., 2019). Then, following the sequence of arguments in the proof of Theorem 1.2, one can show that a critical point $(\mathbf{U},\mathbf{V})$ that is not a global minimum has strictly negative curvature by finding a direction (which depends on whether $\mathbf{U}$ is full column rank) such that the corresponding Hessian quadrature form is strictly negative.

When $\boldsymbol{\Sigma}_{\mathbf{xx}}$ is also not invertible, it seems difficult to characterize all the critical points as in (1.24). Nevertheless, by exploiting the *local openness* property of the risk $\mathcal{R}(\mathbf{U},\mathbf{V})$, Nouiehed and Razaviyayn (2018) show that any local minimum is a global minimum for all possible input-output covariance matrices, $\boldsymbol{\Sigma}_{\mathbf{xx}}$, $\boldsymbol{\Sigma}_{\mathbf{yx}}$ and $\boldsymbol{\Sigma}_{\mathbf{yy}}$. We summarize these results in the following Theorem, but we refer to Nouiehed and Razaviyayn (2018); Zhu et al. (2019) for the full proof.

Theorem 1.3 (Nouiehed and Razaviyayn (2018); Zhu et al. (2019)) *Any local minimum of $\mathcal{R}$ is a global minimum. Moreover, if $\mathbf{\Sigma_{xx}}$ is invertible, then any critical point of $\mathcal{R}$ that is not a global minimum is a strict saddle.*

1.3.2 Deep Linear Networks with Squared Loss

In this section we extend the analysis of the optimization landscape of single-hidden layer linear networks trained using the unregularized squared loss to deep linear networks. We consider a network with dimensions $n_0, n_1, \ldots, n_L$, where n_0 is the input dimension, n_L is the output dimension, $n_1, \ldots, n_{L-1}$ are the hidden layer dimensions, $\mathbf{W}^{[l]} \in \mathbb{R}^{n_l \times n_{l-1}}$ is the matrix of weights between layers $l-1$ and l, and L is the number of weight layers. The hypothesis space $\mathcal{F}$ can be parametrized in terms of the network weights $\mathbf{W} = \{\mathbf{W}^{[l]}\}_{l=1}^{L}$ as

$$\mathcal{F} = \{f \in \mathcal{Y}^{\mathcal{X}} : f(\mathbf{x}) = \mathbf{W}^{[L]}\mathbf{W}^{[L-1]} \cdots \mathbf{W}^{[1]}\mathbf{x}, \text{ where } \mathbf{W}^{[l]} \in \mathbb{R}^{n_l \times n_{l-1}}\}. \quad (1.27)$$

Therefore, the problem of minimizing the expected risk becomes

$$\min_{\{\mathbf{W}^{[l]}\}_{l=1}^{L}} \big[\mathcal{R}(\mathbf{W}) \doteq \mathbb{E}_{\mathbf{x},\mathbf{y}}[\|\mathbf{y} - \mathbf{W}^{[L]}\mathbf{W}^{[L-1]} \cdots \mathbf{W}^{[1]}\mathbf{x}\|_2^2]\big]. \quad (1.28)$$

Similar to the single-hidden layer case in (1.13), the expected risk can be rewritten as:

$$\mathcal{R}(\mathbf{W}) = \operatorname{trace}(\mathbf{\Sigma_{yy}} - 2\mathbf{\Sigma_{yx}}\mathbf{W}_{1:L}^{\top} + \mathbf{W}_{1:L}\mathbf{\Sigma_{xx}}\mathbf{W}_{1:L}^{\top}), \quad (1.29)$$

where $\mathbf{W}_{L:1} = \mathbf{W}^{[L]}\mathbf{W}^{[L-1]} \cdots \mathbf{W}^{[1]}$. Also similar to the single-hidden layer case in (1.14), the problem of minimizing the empirical risk

$$\min_{\{\mathbf{W}^{[l]}\}_{l=1}^{L}} \big[\mathcal{R}_{\mathcal{S}}(\mathbf{W}) \doteq \frac{1}{m}\|\mathbf{Y} - \mathbf{W}_{L:1}\mathbf{X}\|_F^2\big], \quad (1.30)$$

where $\mathbf{X}$ and $\mathbf{Y}$ are the input and output data matrices, is equivalent to minimizing $\mathcal{R}(\mathbf{W})$, except that $\mathbf{\Sigma_{xx}}$, $\mathbf{\Sigma_{xy}}$ and $\mathbf{\Sigma_{yy}}$ need to be substituted by their empirical estimates $\frac{1}{m}\mathbf{X}\mathbf{X}^{\top}$, $\frac{1}{m}\mathbf{X}\mathbf{Y}^{\top}$ and $\frac{1}{m}\mathbf{Y}\mathbf{Y}^{\top}$, respectively. Thus, the analysis of the optimization landscape of both the expected and empirical risk can be done by analyzing the landscape of $\mathcal{R}(\mathbf{W})$. As discussed in Section 1.3.1, when there is no constraint on $\mathbf{W}_{L:1}$ (e.g., when the dimensions of the hidden layers is sufficiently large) and $\mathbf{\Sigma_{xx}}$ is invertible, the optimal input-output weight matrix is given by $\mathbf{W}_{L:1}^{*} = \mathbf{\Sigma_{yx}}\mathbf{\Sigma_{xx}}^{-1}$. However, this result does not provide a characterization of the optimal weight matrices $\mathbf{W}^{[l]}$ for each layer. Thus, the challenge in characterizing the landscape of $\mathcal{R}(\mathbf{W})$ is to understand the effect of the low-rank constraint $n_l \leq \min\{n_0, n_L\}$ or to consider the possibility that critical points for $\mathbf{W}^{[l]}$ might not be low-rank.

Recent work by Kawaguchi (2016) provides a formal analysis of the optimization landscape of $\mathcal{R}(\mathbf{W})$. In particular, by using a purely deterministic approach that exploits both first-order and second-order information at critical points (as also used in the proof of Theorem 1.2), Kawaguchi (2016) characterizes the following properties of critical points of $\mathcal{R}(\mathbf{W})$.

Theorem 1.4 (Kawaguchi (2016)) *Assume that $\mathbf{\Sigma_{xx}}$ and $\mathbf{\Sigma_{xy}}$ are of full rank with $n_L \leq n_0$ and that $\mathbf{\Sigma} = \mathbf{\Sigma_{yx}}\mathbf{\Sigma_{xx}^{-1}}\mathbf{\Sigma_{xy}}$ is of full rank with n_L distinct eigenvalues. Then $\mathcal{R}(\mathbf{W})$ has the following properties:*

- *Any local minimum is a global minimum.*
- *Every critical point that is not a global minimum is a saddle point.*
- *A saddle point $\mathbf{W}$ such that rank$(\mathbf{W}^{[L-1]} \cdots \mathbf{W}^{[2]}) = \min_{1 \leq l \leq L-1} n_l$ is strict, i.e., the Hessian of $\mathcal{R}$ at $\mathbf{W}$ has a strictly negative eigenvalue.*
- *A saddle point $\mathbf{W}$ such that rank$(\mathbf{W}^{[L-1]} \cdots \mathbf{W}^{[2]}) < \min_{1 \leq l \leq L-1} n_l$ may not be strict, i.e., the Hessian at $\mathbf{W}$ may not have any negative eigenvalue.*

On the one hand, similar to Theorem 1.2 for one-hidden layer linear networks, Theorem 1.4 guarantees that under similar conditions on $\mathbf{\Sigma_{xx}}$ and $\mathbf{\Sigma_{xy}}$, any local minimum of the risk is a global minimum. On the other hand, unlike the results for single-hidden layer linear network where every saddle point is strict, Theorem 1.4 shows that networks with two or more hidden layers can have "bad" (non-strict) saddle points, which are also referred to as degenerate saddle points or higher order saddle points, since the first and second order derivatives cannot distinguish them from local optima. To illustrate why depth introduces degenerate saddle points, consider the simplest case where $L = 3$ and $n_0 = n_1 = n_2 = n_3 = 1$. In this case, the risk becomes

$$\mathcal{R}(w^{[1]}, w^{[2]}, w^{[3]}) = \sigma_{\mathbf{yy}} - 2\sigma_{\mathbf{yx}} w^{[1]} w^{[2]} w^{[3]} + \sigma_{\mathbf{xx}} (w^{[1]} w^{[2]} w^{[3]})^2. \quad (1.31)$$

By computing the gradient (derivatives) and Hessian (second-order derivatives), it is easy to see that $(0, 0, 0)$ is a critical point, but the Hessian is the zero matrix, which has no negative eigenvalue. This also holds true for general deep linear networks. Intuitively, when the network has more layers, the objective function tends to be flatter at the origin, making the origin a higher-order saddle. This is similar to the fact that 0 is a critical point of both functions $(1 - u^2)^2$ and $(1 - u^3)^2$: the former has negative second-order derivative at 0, while the later has second-order derivative 0 at this point.

We end up the discussion for deep linear networks by noting that there is recent work that improves upon Theorem 1.4, mostly with weaker conditions to guarantee the absence of spurious local minima. For example, to show that any local minimum is a global minimum, Lu and Kawaguchi (2017)

require $\mathbf{\Sigma_{xx}}$ and $\mathbf{\Sigma_{xy}}$ to be full rank, while Laurent and Brecht (2018) only require that the size of the hidden layers be bigger than or equal to the input or output dimensions, i.e., $n_1, \ldots, n_{L-1} \geq \min\{n_0, n_L\}$, which could potentially help guide the design of network architectures.

1.4 Optimization Landscape of Nonlinear Networks

In this section, we review recent work by Haeffele and Vidal (2015, 2017) on the analysis of the landscape of a class of nonlinear networks with positively homogeneous activation functions, such as Rectified Linear Units (ReLU), max-pooling, etc. Critical to the analysis tools that are employed is to consider networks regularized by a function that is also positively homogeneous of the same degree as the network mapping. These results apply to a class of deep networks whose output is formed as the sum of the outputs of multiple positively homogeneous subnetworks connected in parallel (see right panel in Figure 1.3), where the architecture of a subnetwork can be arbitrary provided the overall mapping of the subnetwork is a positively homogeneous function of the network parameters. Specifically, we show that when the network is sufficiently wide then a path to a global minimizer always exists from any initialization (i.e., local minima which require one to increase the objective to escape are guaranteed not to exist).

As a motivating example, before considering the case of deep positively homogeneous networks, in Section 1.4.1 we revisit the case of shallow linear networks discussed in Section 1.3.1, as this simple particular case conveys the key insights behind the more general cases discussed in Section 1.4.2. The primary difference with the case of shallow linear networks discussed in Section 1.3.1 is that, rather than fixing the number of columns in $(\mathbf{U}, \mathbf{V})$ *a priori*, we constrain the hypothesis space using Tykhonov regularization on $(\mathbf{U}, \mathbf{V})$ while allowing the number of columns in $(\mathbf{U}, \mathbf{V})$ to be variable. As we will see, the Tykhonov regularization results in promoting low-rank solutions even though we do not place an explicit constraint on the number of columns in $(\mathbf{U}, \mathbf{V})$. The extension of these results to deep positively homogeneous networks will highlight the importance of using similar explicit regularization to constrain the overall size of the network.

1.4.1 Motivating Example

Consider the empirical risk minimization problem in (1.14) for a single-hidden layer linear network with n_0 inputs, n_1 hidden neurons, and n_2 out-

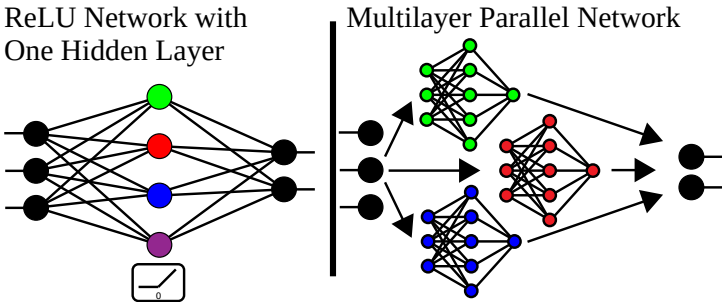

Figure 1.3 Example networks. (Left panel) ReLU network with a single hidden layer with the mapping Φ_{n_1} described by the equation in (1.44) with ($n_1 = 4$). Each color corresponds to one element of the elemental mapping $\phi(\mathbf{X}, \mathbf{W}_i^{[1]}, \mathbf{W}_i^{[2]})$. The colored hidden units have rectifying non-linearities, while the black units are linear. (Right panel) Multilayer ReLU network with 3 fully connected parallel subnetworks ($r = 3$), where each color corresponds to the subnetwork described the elemental mapping $\phi(\mathbf{X}, \mathbf{W}_i^{[1]}, \mathbf{W}_i^{[2]}, \mathbf{W}_i^{[3]}, \mathbf{W}_i^{[4]})$.

puts, which is equivalent to:

$$\min_{\mathbf{U},\mathbf{V}} \tfrac{1}{2}\|\mathbf{Y} - \mathbf{U}\mathbf{V}^\top \mathbf{X}\|_F^2, \tag{1.32}$$

where $\mathbf{U} \in \mathbb{R}^{n_2 \times n_1}$ and $\mathbf{V} \in \mathbb{R}^{n_0 \times n_1}$ are the network weights. We showed in Section 1.3.1 that, under certain conditions, this problem has no spurious local minima and all saddle points are strict. In particular, we assumed that the number of hidden neurons is fixed and limited by the input and output dimensions, specifically, $n_1 \leq \min\{n_0, n_2\}$.

In what follows, we relax this constraint on the network size and optimize over both the network size and weights. Arguably, this requires some form of regularization on the network weights that allows us to control the growth of the network size. A commonly used regularizer is weight decay, $\Theta(\mathbf{U}, \mathbf{V}) = \|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2$, also known as Tykhonov regularization. Here, we use weight decay to regularize (1.32) in the particular case where $\mathbf{X} = \mathbf{I}$ for simplicity of presentation,[9]

$$\min_{n_1 \in \mathbb{N}^+} \min_{\substack{\mathbf{U} \in \mathbb{R}^{n_2 \times n_1} \\ \mathbf{V} \in \mathbb{R}^{n_0 \times n_1}}} \frac{1}{2}\|\mathbf{Y} - \mathbf{U}\mathbf{V}^\top\|_F^2 + \frac{\lambda}{2}(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2), \tag{1.33}$$

where $\lambda > 0$ is a regularization parameter.

There are several reasons for considering this particular case. First, (1.33) can be understood as a matrix factorization problem, where given a matrix $\mathbf{Y} \in \mathbb{R}^{n_2 \times n_0}$, the goal is to factorize it as $\mathbf{Y} \approx \mathbf{U}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{R}^{n_2 \times n_1}$

[9] Note that with this simplification the problem becomes an unsupervised learning problem (matrix factorization) instead of the original supervised learning problem (linear network training).

and $\mathbf{V} \in \mathbb{R}^{n_0 \times n_1}$. Second, it is known that weight decay is closely connected to the nuclear norm of the product of the factorized matrices $\mathbf{Z} = \mathbf{U}\mathbf{V}^\top$, where recall that the nuclear norm $\|\mathbf{Z}\|_*$ is the sum of the singular values of a matrix $\mathbf{Z}$, via the so-called variational form of the nuclear norm (Srebro et al., 2004):

$$\|\mathbf{Z}\|_* = \min_{n_1 \in \mathbb{N}^+} \min_{\substack{\mathbf{U},\mathbf{V}: \\ \mathbf{U}\mathbf{V}^\top = \mathbf{Z}}} \frac{1}{2}(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2), \tag{1.34}$$

where the above equation states that given a matrix $\mathbf{Z}$, if one considers all possible factorizations of $\mathbf{Z}$ into $\mathbf{U}\mathbf{V}^\top$ then finding a factorization $\mathbf{Z} = \mathbf{U}\mathbf{V}^\top$ which minimizes the Tykhonov regularization on $(\mathbf{U}, \mathbf{V})$ will be equal to the nuclear norm of $\mathbf{Z}$. Third, recall that the nuclear norm is a convex relaxation of the rank of the matrix, which is known to encourage low-rank solutions. As a result, this will allow us to control the network width, n_1, and ensure capacity control via matrix factorization techniques without placing explicit constraints on n_1. Indeed, due to the variational definition of the nuclear norm in (1.34), the objective in (1.33) will be closely related to a *convex* optimization problem with nuclear norm regularization:

$$\min_{\mathbf{Z}} \frac{1}{2}\|\mathbf{Y} - \mathbf{Z}\|_F^2 + \lambda\|\mathbf{Z}\|_*. \tag{1.35}$$

The strong similarity between (1.34) and the regularizer in (1.33) suggests looking at the *convex* problem in (1.35), whose solution can be found in closed form from the SVD of $\mathbf{Y}$. Specifically, if $\mathbf{Y} = \mathbf{U}_\mathbf{Y}\mathbf{\Sigma}_\mathbf{Y}\mathbf{V}_\mathbf{Y}^\top$ is the SVD of $\mathbf{Y}$, then the global minimizer of (1.35) is given by the singular value thresholding operator $\mathbf{Z} = \mathcal{D}_\lambda(\mathbf{Y}) = \mathbf{U}_\mathbf{Y}(\mathbf{\Sigma}_\mathbf{Y} - \lambda\mathbf{I})_+\mathbf{V}_\mathbf{Y}^\top$, where the singular vectors of $\mathbf{Y}$ (columns of $\mathbf{U}_\mathbf{Y}$ and $\mathbf{V}_\mathbf{Y}$) are maintained, while the singular values of $\mathbf{Y}$ (diagonal entries of $\mathbf{\Sigma}_\mathbf{Y}$) are shrunk by λ and then thresholded at zero, i.e., $a_+ = \max(a, 0)$.

But how do these results for the convex problem in (1.35) relate to solutions to the non-convex problem in (1.33)? First, observe that (1.34) implies that the convex problem provides a global lower bound for the non-convex problem. Specifically, for any $(\mathbf{U}, \mathbf{V}, \mathbf{Z})$ such that $\mathbf{Z} = \mathbf{U}\mathbf{V}^\top$ (1.34) implies

$$\frac{1}{2}\|\mathbf{Y} - \mathbf{Z}\|_F^2 + \lambda\|\mathbf{Z}\|_* \leq \frac{1}{2}\|\mathbf{Y} - \mathbf{U}\mathbf{V}^\top\|_F^2 + \frac{\lambda}{2}(\|\mathbf{U}\|_F^2 + \|\mathbf{V}^2\|_F^2). \tag{1.36}$$

Additionally, this lower bound is always tight once n_1 becomes sufficiently large in the sense that for any $\mathbf{Z}$ one can find a $(\mathbf{U}, \mathbf{V})$ such that $\mathbf{Z} = \mathbf{U}\mathbf{V}^\top$ and the inequality above will be an equality. As a result, a global minimum $(\mathbf{U}, \mathbf{V})$ of the non-convex problem (1.33) gives a global minimum $\mathbf{Z} = \mathbf{U}\mathbf{V}^\top$ for the convex problem (1.35), and due to the global lower-bound in (1.36)

this further implies that we have a global minimum of both the convex and non-convex problems, as we show next.

Theorem 1.5 *Let* $\mathbf{u}_i$ *and* $\mathbf{v}_i$ *be the ith columns of* $\mathbf{U}$ *and* $\mathbf{V}$*, respectively. If* $(\mathbf{U}, \mathbf{V}, n_1)$ *is a local minimum of* (1.33) *such that for some* $i \in [n_1]$ *we have* $\mathbf{u}_i = \mathbf{0}$ *and* $\mathbf{v}_i = \mathbf{0}$*, then (i)* $\mathbf{Z} = \mathbf{U}\mathbf{V}^\top$ *is a global minimum of* (1.35)*, and (ii)* $(\mathbf{U}, \mathbf{V}, n_1)$ *is a global minimum of* (1.33) *and* (1.34)*.*

Proof Recall the Fenchel conjugate of a function Ω is defined as $\Omega^*(\mathbf{Q}) = \sup_{\mathbf{Z}} \langle \mathbf{Q}, \mathbf{Z} \rangle - \Omega(\mathbf{Z})$, leading to Fenchel's inequality $\langle \mathbf{Q}, \mathbf{Z} \rangle \leq \Omega(\mathbf{Z}) + \Omega^*(\mathbf{Q})$. Also recall the subgradient of a convex function Ω at $\mathbf{Z}$ is defined as $\partial\Omega(\mathbf{Z}) = \{\mathbf{Q} : \Omega(\bar{\mathbf{Z}}) \geq \Omega(\mathbf{Z}) + \langle \mathbf{Q}, \bar{\mathbf{Z}} - \mathbf{Z} \rangle, \forall \bar{\mathbf{Z}}\} = \{\mathbf{Q} : \langle \mathbf{Q}, \mathbf{Z} \rangle = \Omega(\mathbf{Z}) + \Omega^*(\mathbf{Q})\}$. Applying this to the nuclear norm $\Omega(\mathbf{Z}) = \|\mathbf{Z}\|_*$ and using (1.34), we obtain

$$\begin{aligned}
\Omega^*(\mathbf{Q}) &= \sup_{\mathbf{Z}} \langle \mathbf{Q}, \mathbf{Z} \rangle - \|\mathbf{Z}\|_* = \sup_{n_1 \in \mathbb{N}^+} \sup_{\tilde{\mathbf{U}}, \tilde{\mathbf{V}}} \langle \mathbf{Q}, \tilde{\mathbf{U}}\tilde{\mathbf{V}}^\top \rangle - \frac{1}{2}(\|\tilde{\mathbf{U}}\|_F^2 + \|\tilde{\mathbf{V}}\|_F^2) \\
&= \sup_{n_1 \in \mathbb{N}^+} \sup_{\tilde{\mathbf{U}}, \tilde{\mathbf{V}}} \sum_{i=1}^{n_1} \left(\tilde{\mathbf{u}}_i^\top \mathbf{Q} \tilde{\mathbf{v}}_i - \frac{1}{2}(\|\tilde{\mathbf{u}}_i\|_2^2 + \|\tilde{\mathbf{v}}_i\|_2^2) \right) \\
&= \begin{cases} 0 & \text{if} \quad \mathbf{u}^\top \mathbf{Q} \mathbf{v} \leq \frac{1}{2}(\|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2) \quad \forall (\mathbf{u}, \mathbf{v}) \\ \infty & \text{else,} \end{cases}
\end{aligned} \tag{1.37}$$

which then implies

$$\partial \|\mathbf{Z}\|_* = \{\mathbf{Q} \colon \langle \mathbf{Q}, \mathbf{Z} \rangle = \|\mathbf{Z}\|_* \text{ and } \mathbf{u}^\top \mathbf{Q} \mathbf{v} \leq \tfrac{1}{2}(\|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2) \ \forall (\mathbf{u}, \mathbf{v})\}. \tag{1.38}$$

To show (i), we need to show that $0 \in \mathbf{Z} - \mathbf{Y} + \lambda \partial \|\mathbf{Z}\|_*$ or $\mathbf{Y} - \mathbf{Z} \in \lambda \partial \|\mathbf{Z}\|_*$. Let us first show that $\mathbf{Q} = \frac{\mathbf{Y} - \mathbf{Z}}{\lambda}$ satisfies the inequality in (1.38). Assume without loss of generality that the last column of $\mathbf{U}$ and $\mathbf{V}$ are zero, choose any $\mathbf{u} \in \mathbb{R}^{n_2}$, $\mathbf{v} \in \mathbb{R}^{n_0}$, and $\epsilon > 0$ and let $\mathbf{U}_\epsilon = \mathbf{U} + \epsilon^{1/2} \mathbf{u} \mathbf{e}_{n_1}^\top$ and $\mathbf{V}_\epsilon = \mathbf{V} + \epsilon^{1/2} \mathbf{v} \mathbf{e}_{n_1}^\top$ so that $\mathbf{Z}_\epsilon = \mathbf{U}_\epsilon \mathbf{V}_\epsilon^\top = \mathbf{U}\mathbf{V}^\top + \epsilon \mathbf{u}\mathbf{v}^\top = \mathbf{Z} + \epsilon \mathbf{u}\mathbf{v}^\top$. Since $(\mathbf{U}, \mathbf{V}, n_1)$ is a local minimum of (1.33), for all $(\mathbf{u}, \mathbf{v})$ there exists $\delta > 0$ such that for all $\epsilon \in (0, \delta)$ we have:

$$\begin{aligned}
\frac{1}{2}\|\mathbf{Y} - \mathbf{Z}_\epsilon\|_F^2 + \frac{\lambda}{2}(\|\mathbf{U}_\epsilon\|_F^2 + \|\mathbf{V}_\epsilon\|_F^2) - \frac{1}{2}\|\mathbf{Y} - \mathbf{Z}\|_F^2 - \frac{\lambda}{2}(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2) &\geq 0 \\
\frac{1}{2}(-2\langle \mathbf{Y}, \mathbf{Z}_\epsilon - \mathbf{Z} \rangle + \langle \mathbf{Z}_\epsilon + \mathbf{Z}, \mathbf{Z}_\epsilon - \mathbf{Z} \rangle) + \frac{\lambda}{2}(\|\mathbf{U}_\epsilon\|_F^2 - \|\mathbf{U}\|_F^2 + \|\mathbf{V}_\epsilon\|_F^2 - \|\mathbf{V}\|_F^2) &\geq 0 \\
\frac{\epsilon}{2}(-2\langle \mathbf{Y}, \mathbf{u}\mathbf{v}^\top \rangle + \langle 2\mathbf{Z} + \epsilon \mathbf{u}\mathbf{v}^\top, \mathbf{u}\mathbf{v}^\top \rangle) + \frac{\lambda \epsilon}{2}(\|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2) &\geq 0 \\
\mathbf{u}^\top (\mathbf{Z} - \mathbf{Y}) \mathbf{v} + \frac{\epsilon}{2}\|\mathbf{u}\|_2^2 \|\mathbf{v}\|_2^2 + \frac{\lambda}{2}(\|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2) &\geq 0.
\end{aligned}$$

Letting $\epsilon \searrow 0$, gives $\mathbf{u}^\top \frac{(\mathbf{Y} - \mathbf{Z})}{\lambda} \mathbf{v} \leq \frac{1}{2}(\|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2)$, $\forall (\mathbf{u}, \mathbf{v})$ as claimed.

Let us now show that $\mathbf{Q} = \frac{\mathbf{Y}-\mathbf{Z}}{\lambda}$ satisfies the equality in (1.38). Because the inequality in (1.38) holds, we know that $\Omega^*(\mathbf{Q}) = 0$, which together with Fenchel's inequality gives $\langle \mathbf{Q}, \mathbf{Z} \rangle \leq \|\mathbf{Z}\|_*$. Then, since $\mathbf{Z} = \mathbf{U}\mathbf{V}^\top$, it follows from (1.34) that $\|\mathbf{Z}\|_* \leq \frac{1}{2}(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2)$. Therefore, to show that $\langle \mathbf{Q}, \mathbf{Z} \rangle = \|\mathbf{Z}\|_*$ it suffices to show that $\langle \mathbf{Q}, \mathbf{Z} \rangle = \frac{1}{2}(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2)$. For this particular problem it is possible to show that this equality is satisfied simply by considering the first-order optimality conditions, which must be satisfied since $(\mathbf{U}, \mathbf{V}, n_1)$ is a local minimum:

$$-(\mathbf{Y} - \mathbf{U}\mathbf{V}^\top)\mathbf{V} + \lambda \mathbf{U} = 0 \quad \text{and} \quad -(\mathbf{Y} - \mathbf{U}\mathbf{V}^\top)^\top \mathbf{U} + \lambda \mathbf{V} = 0. \tag{1.39}$$

It follows that

$$\mathbf{U}^\top(\mathbf{Y} - \mathbf{U}\mathbf{V}^\top)\mathbf{V} = \lambda \mathbf{U}^\top \mathbf{U} \quad \text{and} \quad \mathbf{U}^\top(\mathbf{Y} - \mathbf{U}\mathbf{V}^\top)\mathbf{V}^\top = \lambda \mathbf{V}^\top \mathbf{V}. \tag{1.40}$$

Summing, taking the trace and dividing by λ gives the desired result

$$\langle \tfrac{\mathbf{Y}-\mathbf{U}\mathbf{V}^\top}{\lambda}, \mathbf{U}\mathbf{V}^\top \rangle = \frac{1}{2}(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2) = \|\mathbf{Z}\|_*. \tag{1.41}$$

As a consequence, $\mathbf{W} \in \partial \|\mathbf{Z}\|_*$ and hence $\mathbf{Z} = \mathbf{U}\mathbf{V}^\top$ is a global minimum of the convex problem in (1.35), thus concluding the proof of (i).

As an alternative to the above approach, to develop an intuition for more general results, we will also provide an alternative proof of the equality in (1.41) without relying on the objective being differentiable w.r.t. $(\mathbf{U}, \mathbf{V})$ and only requiring the loss function to be differentiable w.r.t. $\mathbf{Z} = \mathbf{U}\mathbf{V}^\top$. In particular, let $\mathbf{U}_\tau = (1+\tau)^{1/2}\mathbf{U}$, $\mathbf{V}_\tau = (1+\tau)^{1/2}\mathbf{V}$, and $\mathbf{Z}_\tau = \mathbf{U}_\tau \mathbf{V}_\tau^\top = (1+\tau)\mathbf{U}\mathbf{V}^\top = (1+\tau)\mathbf{Z}$. Again since $(\mathbf{U}, \mathbf{V}, n_1)$ is a local minimum for $\tau > 0$ sufficiently small we have:

$$\begin{aligned}
\frac{1}{2}\|\mathbf{Y} - \mathbf{Z}_\tau\|_F^2 + \frac{\lambda}{2}(\|\mathbf{U}_\tau\|_F^2 + \|\mathbf{V}_\tau\|_F^2) - \frac{1}{2}\|\mathbf{Y} - \mathbf{Z}\|_F^2 - \frac{\lambda}{2}(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2) &\geq 0 \\
\frac{1}{2}\|\mathbf{Y} - (1+\tau)\mathbf{Z}\|_F^2 - \frac{1}{2}\|\mathbf{Y} - \mathbf{Z}\|_F^2 + \frac{\lambda}{2}(\tau\|\mathbf{U}\|_F^2 + \tau\|\mathbf{V}\|_F^2) &\geq 0 \\
\frac{1}{\tau}\left(\frac{1}{2}\|\mathbf{Y} - \mathbf{Z} - \tau\mathbf{Z}\|_F^2 - \frac{1}{2}\|\mathbf{Y} - \mathbf{Z}\|_F^2\right) &\geq -\frac{\lambda}{2}(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2).
\end{aligned}$$

Taking the limit $\tau \searrow 0$ (where recall that the above limit on the l.h.s. is the directional derivative of the loss in the direction $\mathbf{Z}$) gives:

$$\langle \mathbf{Z} - \mathbf{Y}, \mathbf{Z} \rangle \geq -\frac{\lambda}{2}(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2) \Longrightarrow \langle \tfrac{\mathbf{Y}-\mathbf{U}\mathbf{V}^\top}{\lambda}, \mathbf{U}\mathbf{V}^\top \rangle \leq \frac{1}{2}(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2).$$

If we then let $\bar{\mathbf{U}}_\tau = (1-\tau)^{1/2}\mathbf{U}$, $\bar{\mathbf{V}}_\tau = (1-\tau)^{1/2}\mathbf{V}$, and $\bar{\mathbf{Z}}_\tau = \bar{\mathbf{U}}_\tau \bar{\mathbf{V}}_\tau^\top = (1-\tau)\mathbf{Z}$, then by repeating an identical set of arguments as before we obtain

the opposite inequality:

$$\frac{1}{2}\|\mathbf{Y}-\bar{\mathbf{Z}}_\tau\|_F^2+\frac{\lambda}{2}(\|\bar{\mathbf{U}}_\tau\|_F^2+\|\bar{\mathbf{V}}_\tau\|_F^2)-\frac{1}{2}\|\mathbf{Y}-\mathbf{Z}\|_F^2-\frac{\lambda}{2}(\|\mathbf{U}\|_F^2+\|\mathbf{V}\|_F^2)\geq 0$$

$$\frac{1}{2}\|\mathbf{Y}-(1-\tau)\mathbf{Z}\|_F^2-\frac{1}{2}\|\mathbf{Y}-\mathbf{Z}\|_F^2-\frac{\lambda}{2}(\tau\|\mathbf{U}\|_F^2+\tau\|\mathbf{V}\|_F^2)\geq 0$$

$$\frac{1}{\tau}\left(\frac{1}{2}\|\mathbf{Y}-\mathbf{Z}+\tau\mathbf{Z}\|_F^2-\frac{1}{2}\|\mathbf{Y}-\mathbf{Z}\|_F^2\right)\geq\frac{\lambda}{2}(\|\mathbf{U}\|_F^2+\|\mathbf{V}\|_F^2)$$

$$\text{taking the limit } \tau\searrow 0 \implies$$

$$\langle\mathbf{Z}-\mathbf{Y},-\mathbf{Z}\rangle\geq\frac{\lambda}{2}(\|\mathbf{U}\|_F^2+\|\mathbf{V}\|_F^2)\Longrightarrow\langle\tfrac{\mathbf{Y}-\mathbf{U}\mathbf{V}^\top}{\lambda},\mathbf{U}\mathbf{V}^\top\rangle\geq\frac{1}{2}(\|\mathbf{U}\|_F^2+\|\mathbf{V}\|_F^2).$$

As a result, we have again shown that (1.41) must be true without relying on the differentiability of the objective w.r.t. $(\mathbf{U},\mathbf{V})$, only the differentiablity of the loss function w.r.t. $\mathbf{Z}$ when we take the limits as $\tau\searrow 0$.

Finally, to see that claim (ii) is true, observe the equality on the right hand side of (1.41) implies that $(\mathbf{U},\mathbf{V},n_1)$ is an optimal factorization, i.e., a global minimum of (1.34). Finally, since the convex problem in (1.35) is a global lower bound for the non-convex problem in (1.33) and $\mathbf{Z}=\mathbf{U}\mathbf{V}^\top$ is a global minimum of the convex problem, it follows that $(\mathbf{U},\mathbf{V},n_1)$ must be a global minimum of the non-convex problem.

□

In summary, we have shown that the non-convex problem matrix factorization problem in $(\mathbf{U},\mathbf{V})$ admits a global lower bound in the product space $\mathbf{Z}=\mathbf{U}\mathbf{V}^\top$. Moreover, the lower bound is a convex function of $\mathbf{Z}$, and the global minima agree, i.e., if $(\mathbf{U},\mathbf{V},n_1)$ is a global minimum of the non-convex problem, then $\mathbf{U}\mathbf{V}^\top$ is a global minimum of the convex problem. In addition, Theorem 1.5 provides a characterization of local minima of the non-convex problem which are also global: local minima with one column of $\mathbf{U}$ and the corresponding column of $\mathbf{V}$ being zero. Such a statement can be easily extended to local minima $(\mathbf{U},\mathbf{V},n_1)$ that are rank deficient, i.e., there exists $\mathbf{e}\neq 0$ such that $\mathbf{U}\mathbf{e}=0$ and $\mathbf{V}\mathbf{e}=0$, since the only part of the proof that depends on columns of $\mathbf{U}$ and $\mathbf{V}$ being zero is the definition of $\mathbf{U}_\epsilon$ and $\mathbf{V}_\epsilon$, which can be readily replaced by $\mathbf{U}_\epsilon=\mathbf{U}+\epsilon^{1/2}\mathbf{u}\mathbf{e}^\top$ and $\mathbf{V}_\epsilon=\mathbf{V}+\epsilon^{1/2}\mathbf{v}\mathbf{e}^\top$ with $\|\mathbf{e}\|_2=1$. In addition, observe that the proof of Theorem 1.5 relies only on the following sufficient and necessary conditions for global optimality of any $(\mathbf{U},\mathbf{V},n_1)$.

Corollary 1.6 *$(\mathbf{U},\mathbf{V},n_1)$ is a global minimum of (1.33) if and only if it satisfies the following conditions*

(i) $\langle\mathbf{Y}-\mathbf{U}\mathbf{V}^\top,\mathbf{U}\mathbf{V}^\top\rangle=\frac{\lambda}{2}(\|\mathbf{U}\|_F^2+\|\mathbf{V}\|_F^2)$.

(ii) $\mathbf{u}^\top(\mathbf{Y} - \mathbf{U}\mathbf{V}^\top)\mathbf{v} \leq \frac{\lambda}{2}(\|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2)$ *for all* $(\mathbf{u}, \mathbf{v})$.

Recall that the global minimum of the convex problem (1.35) is given by the singular value thresholding of $\mathbf{Y}$, $\mathbf{Z} = \mathcal{D}_\lambda(\mathbf{Y}) = \mathbf{U}_\mathbf{Y}(\mathbf{\Sigma}_\mathbf{Y} - \lambda\mathbf{I})_+\mathbf{V}_\mathbf{Y}^\top$, where $\mathbf{Y} = \mathbf{U}_\mathbf{Y}\mathbf{\Sigma}_\mathbf{Y}\mathbf{V}_\mathbf{Y}^\top$ is the SVD of $\mathbf{Y}$. It follows that a global minimum of (1.33) can be obtained as $\mathbf{U} = \mathbf{U}_\mathbf{Y}(\mathbf{\Sigma}_\mathbf{Y} - \lambda\mathbf{I})_+^{1/2}$ and $\mathbf{V} = \mathbf{V}_\mathbf{Y}(\mathbf{\Sigma}_\mathbf{Y} - \lambda\mathbf{I})_+^{1/2}$.

In practice, while a globally optimal solution to (1.35) can be found using linear algebraic techniques, computing the SVD of $\mathbf{Y}$ is highly inefficient for large matrices $\mathbf{Y}$. Therefore, we may still be interested in solving (1.35) using, e.g., gradient descent. In this case, we may be interested in using Corollary 1.6 to check if a global minimum has been found. Observe from the proof of Theorem 1.5 that condition (i) is satisfied by any first-order point, and that optimization methods are often guaranteed to converge to first order points. Therefore, the important condition to check is condition (ii). It can be shown that condition (ii) is equivalent to $\sigma_{\max}(\mathbf{Y} - \mathbf{U}\mathbf{V}^\top) \leq \lambda$, which involves computing only the largest singular value of a matrix. Now, what if condition (ii) is violated? In this case, one might wonder if condition (ii) may be used to escape the non-global local minimum. Indeed, if condition (ii) is violated, then there exists $(\mathbf{u}, \mathbf{v})$ such that $\mathbf{u}^\top(\mathbf{Y} - \mathbf{U}\mathbf{V}^\top)\mathbf{v} > \frac{\lambda}{2}(\|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2)$. Then, it follows from the proof of Theorem 1.5 that if we choose $\mathbf{U}_\epsilon = [\mathbf{U}\ \epsilon^{1/2}\mathbf{u}]$ and $\mathbf{V}_\epsilon = [\mathbf{V}\ \epsilon^{1/2}\mathbf{v}]$ for ϵ small enough, then we can reduce the objective. This suggest an algorithm for minimizing (1.33) which consists of the following two steps

(i) For a fixed n_1, use a local descent strategy to minimize (1.33) with respect to $\mathbf{U}$ and $\mathbf{V}$ until convergence to a first-order point.

(ii) Check if condition (ii) is satisfied, which is equivalent to solving the following optimization problem (called the polar problem)

$$\max_{\mathbf{u},\mathbf{v}} \frac{\mathbf{u}^\top(\mathbf{Y} - \mathbf{U}\mathbf{V}^\top)\mathbf{v}}{\|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2} \leq \frac{\lambda}{2} \iff \max_{\mathbf{u},\mathbf{v}} \frac{\mathbf{u}^\top(\mathbf{Y} - \mathbf{U}\mathbf{V}^\top)\mathbf{v}}{\|\mathbf{u}\|_2\|\mathbf{v}\|_2} \leq \lambda. \tag{1.42}$$

If the condition holds, then a global minimum has been found. Otherwise, let $(\mathbf{u}, \mathbf{v})$ be a solution to the polar problem,[10] augment $\mathbf{U}$ and $\mathbf{V}$ with one additional column as $\mathbf{U}_\epsilon = [\mathbf{U}\ \epsilon^{1/2}\mathbf{u}]$ and $\mathbf{V}_\epsilon = [\mathbf{V}\ \epsilon^{1/2}\mathbf{v}]$ for some $\epsilon > 0$, and go to (i).

We refer the reader to Haeffele and Vidal (2019) for a more precise and detailed description of this meta-algorithm.

[10] Note that a solution to the polar problem is given by the left and right singular vectors of $\mathbf{Y} - \mathbf{U}\mathbf{V}^\top$ associated with its largest singular value.

1.4.2 Positively Homogeneous Networks

The above discussion on matrix factorization can be extended to neural networks with one hidden layer by properly adjusting the definitions of the maps Φ and Θ. In the above matrix factorization example (returning to the use of $\mathbf{W}$ to notate the model parameters), Φ and Θ can be re-written as

$$\Phi(\mathbf{X}, \mathbf{W}^{[1]}, \mathbf{W}^{[2]}) = \mathbf{W}^{[2]}(\mathbf{W}^{[1]})^\top = \sum_{i=1}^{n_1} \mathbf{w}_i^{[2]}(\mathbf{w}_i^{[1]})^\top, \qquad \text{and} \tag{1.43}$$

$$\Theta(\mathbf{W}^{[1]}, \mathbf{W}^{[2]}) = \frac{1}{2}(\|\mathbf{W}^{[1]}\|_F^2 + \|\mathbf{W}^{[2]}\|_F^2) = \sum_{i=1}^{n_1} \frac{1}{2}(\|\mathbf{w}_i^{[1]}\|_2^2 + \|\mathbf{w}_i^{[2]}\|_2^2),$$

where $\mathbf{w}_i^{[1]}$ and $\mathbf{w}_i^{[2]}$ are the ith columns of $\mathbf{W}^{[1]}$ and $\mathbf{W}^{[2]}$, respectively. A key observation is that the map Φ and regularization Θ decompose as sums of functions over the columns of $\mathbf{W}^{[1]}$ and $\mathbf{W}^{[2]}$. Further, these functions are both positively homogeneous[11] with degree 2.

Turning to single-hidden-layer neural networks, if we again let n_1 denote the number of neurons in the hidden layer this motivates the following more general definitions for Φ and Θ:

$$\begin{aligned} \Phi_{n_1}(\mathbf{X}, \mathbf{W}^{[1]}, \mathbf{W}^{[2]}) &= \sum_{i=1}^{n_1} \phi(\mathbf{X}, \mathbf{w}_i^{[1]}, \mathbf{w}_i^{[2]}), \qquad \text{and} \\ \Theta_{n_1}(\mathbf{W}^{[1]}, \mathbf{W}^{[2]}) &= \sum_{i=1}^{n_1} \theta(\mathbf{w}_i^{[1]}, \mathbf{w}_i^{[2]}), \end{aligned} \tag{1.44}$$

where $\phi(\mathbf{X}, \mathbf{w}^{[1]}, \mathbf{w}^{[2]})$ and $\theta(\mathbf{w}^{[1]}, \mathbf{w}^{[2]})$ are functions which are both positively homogeneous of the same degree $p > 0$ w.r.t. $(\mathbf{w}^{[1]}, \mathbf{w}^{[2]})$. Clearly, $\phi(\mathbf{X}, \mathbf{w}^{[1]}, \mathbf{w}^{[2]}) = \mathbf{w}^{[1]}(\mathbf{w}^{[2]})^\top$ and $\theta(\mathbf{w}^{[1]}, \mathbf{w}^{[2]}) = \frac{1}{2}(\|\mathbf{w}^{[1]}\|_2^2 + \|\mathbf{w}^{[2]}\|_2^2)$ satisfy this property, with $p = 2$. But notice that it is also satisfied, for example, by the map $\phi(\mathbf{X}, \mathbf{w}^{[1]}, \mathbf{w}^{[2]}) = \mathbf{w}^{[2]}\text{ReLU}((\mathbf{w}^1)^\top \mathbf{X})$, where recall $\text{ReLU}(z) = \max(z, 0)$ is a ReLU applied to each entry of $(\mathbf{w}^{[1]})^\top \mathbf{X}$. The fundamental observation is that both linear transformations and ReLU nonlinearities[12] are positively homogeneous functions, and so the composition of such functions is also positively homogeneous. With these definitions, it is easy to see that the output of a two-layer neural network with ReLU nonlinearity on the hidden units, such as the one illustrated in the left panel of Figure 1.3, can be expressed by the map Φ in (1.44).

[11] Recall that a function f is said to be **positively homogeneous with degree-p** if $\forall \alpha \geq 0$ one has $f(\alpha x) = \alpha^p f(x)$.

[12] Notice that many other neural network operators such as max-pooling, leaky ReLUs, raising to a polynomial power, and convolution are also positively homogeneous.

This same approach can be generalized beyond single-hidden-layer networks to the more general multi-layer parallel network shown in the right panel of Fig. 1.3 by considering more general ϕ and θ functions. In particular, we define the mapping of the multi-layer parallel network and its corresponding regularization function as the sum of the corresponding mappings and regularization functions for r parallel subnetworks with identical architectures but possibly different weights. Specifically, we define the mapping of the multi-layer parallel network and its regularization function as

$$\begin{aligned} \Phi_r(\mathbf{X}, \mathbf{W}^{[1]}, \dots, \mathbf{W}^{[L]}) &= \sum_{i=1}^{r} \phi(\mathbf{X}, \mathbf{W}_i^{[1]}, \dots, \mathbf{W}_i^{[L]}), \qquad \text{and} \\ \Theta_r(\mathbf{W}^{[1]}, \dots, \mathbf{W}^{[L]}) &= \sum_{i=1}^{r} \theta(\mathbf{W}_i^{[1]}, \dots, \mathbf{W}_i^{[L]}), \end{aligned} \tag{1.45}$$

where $\mathbf{W}_i^{[l]}$ denotes the weight parameters for the lth layer of the ith subnetwork, $\mathbf{W}^{[l]} = \{\mathbf{W}_i^{[l]}\}_{i=1}^{r}$ is the set of weight parameters for the lth layer of all r subnetworks, and the network mapping $\phi(\mathbf{X}, \mathbf{W}_i^{[1]}, \dots, \mathbf{W}_i^{[L]})$ and regularization function $\theta(\mathbf{W}_i^{[1]}, \dots, \mathbf{W}_i^{[L]})$ are positively homogeneous functions of degree $p > 0$ on the weights of the ith subnetwork $(\mathbf{W}_i^{[1]}, \dots, \mathbf{W}_i^{[L]})$.[13] Therefore, we can write the training problem for a network which consists of the sum of parallel subnetworks (where we also search over the number of subnetworks, r) as:

$$\min_{r \in \mathbb{N}^+} \min_{\mathbf{W}^{[1]}, \dots, \mathbf{W}^{[L]}} \mathcal{L}(\mathbf{Y}, \Phi_r(\mathbf{X}, \mathbf{W}^{[1]}, \dots, \mathbf{W}^{[L]})) + \lambda \Theta_r(\mathbf{W}^{[1]}, \dots, \mathbf{W}^{[L]}). \tag{1.46}$$

Note that this problem is typically non-convex due to the mapping Φ_r regardless of the choice of the loss and regularization functions, $\mathcal{L}$ and Θ, respectively. Therefore, to analyze this non-convex problem, we define a generalization of the variational form of the nuclear norm in (1.34) for neural networks which consist of the sum of parallel subnetworks as:

$$\Omega_{\phi,\theta}(\mathbf{Z}) = \min_{r \in \mathbb{N}^+} \min_{\mathbf{W}^{[1]}, \dots, \mathbf{W}^{[L]} : \Phi_r(\mathbf{X}, \mathbf{W}^{[1]}, \dots, \mathbf{W}^{[L]}) = \mathbf{Z}} \Theta_r(\mathbf{W}^{[1]}, \dots, \mathbf{W}^{[L]}), \tag{1.47}$$

with the additional condition that $\Omega_{\phi,\theta}(\mathbf{Z}) = \infty$ if $\Phi_r(\mathbf{X}, \mathbf{W}^{[1]}, \dots, \mathbf{W}^{[L]}) \neq \mathbf{Z}$ for all $(\mathbf{W}^{[1]}, \dots, \mathbf{W}^{[L]}, r)$. The intuition behind the above problem is that, given an output $\mathbf{Z}$ generated by the network for some input $\mathbf{X}$, we wish to find the number of subnetworks (or the number of hidden-units in the single-hidden-layer case) and weights $(\mathbf{W}^{[1]}, \dots, \mathbf{W}^{[L]})$ that produce the output $\mathbf{Z}$.

[13] Note that θ could additionally depend on $\mathbf{X}$, but we omit that here for notational simplicity.

Then, among all possible choices of sizes and weights, we prefer those that minimize $\Theta_r(\mathbf{W}^{[1]}, \ldots, \mathbf{W}^{[L]})$.

Note that the function $\Omega_{\phi,\theta}$ is completely specified once one chooses a ϕ and θ function in (1.45), so for $\Omega_{\phi,\theta}$ to be well-posed it is required that ϕ and θ satisfy the following conditions:[14]

Definition 1.7 We will say that (ϕ, θ) are a **nondegenerate pair** if for any set of weights for one subnetwork $(\bar{\mathbf{W}}^{[1]}, \ldots, \bar{\mathbf{W}}^{[L]})$ the functions satisfy the following three conditions:

(i) Both ϕ and θ are positively homogeneous functions of the weights with the same degree $p > 0$: $\phi(\mathbf{X}, \alpha\bar{\mathbf{W}}^{[1]}, \ldots, \alpha\bar{\mathbf{W}}^{[L]}) = \alpha^p \phi(\mathbf{X}, \bar{\mathbf{W}}^{[1]}, \ldots, \bar{\mathbf{W}}^{[L]})$ and $\theta(\alpha\bar{\mathbf{W}}^{[1]}, \ldots, \alpha\bar{\mathbf{W}}^{[L]}) = \alpha^p \theta(\bar{\mathbf{W}}^{[1]}, \ldots, \bar{\mathbf{W}}^{[L]})$ for all $\alpha \geq 0$.
(ii) θ is positive semi-definite: $\theta(\bar{\mathbf{W}}^{[1]}, \ldots, \bar{\mathbf{W}}^{[L]}) \geq 0$.
(iii) The set $\{\phi(\mathbf{X}, \bar{\mathbf{W}}^{[1]}, \ldots, \bar{\mathbf{W}}^{[L]}) : \theta(\bar{\mathbf{W}}^{[1]}, \ldots, \bar{\mathbf{W}}^{[L]}) \leq 1\}$ is compact.

As a concrete example, choosing $\phi(\mathbf{X}, \mathbf{w}^{[1]}, \mathbf{w}^{[2]}) = \mathbf{w}^{[2]}\text{ReLU}((\mathbf{w}^{[1]})^\top \mathbf{X})$ as above and $\theta(\mathbf{w}^{[1]}, \mathbf{w}^{[2]}) = \frac{1}{2}(\|\mathbf{w}^{[1]}\|_2^2 + \|\mathbf{w}^{[2]}\|_2^2)$ satisfies the above requirements and corresponds to a single-hidden-layer fully-connected network (as $\Phi_{n_1}(\mathbf{X}, \mathbf{W}^{[1]}, \mathbf{W}^{[2]}) = \mathbf{W}^{[2]}\text{ReLU}((\mathbf{W}^{[1]})^\top \mathbf{X})$) with ℓ_2 weight decay on the parameters. From these preliminaries, one can show that $\Omega_{\phi,\theta}$ satisfies the following properties:

Proposition 1.8 (Haeffele and Vidal (2015, 2017)) *Given a nondegenerate pair of functions (ϕ, θ) then $\Omega_{\phi,\theta}(\mathbf{Z})$ has the following properties:*

(i) *Positive definite:* $\Omega_{\phi,\theta}(\mathbf{Z}) > 0, \forall \mathbf{Z} \neq \mathbf{0}$ *and* $\Omega_{\phi,\theta}(\mathbf{0}) = 0$.
(ii) *Positively homogeneous with degree 1:* $\Omega_{\phi,\theta}(\alpha\mathbf{Z}) = \alpha\Omega_{\phi,\theta}(\mathbf{Z}),\ \forall \alpha \geq 0, \forall \mathbf{Z}$.
(iii) *Triangle inequality:* $\Omega_{\phi,\theta}(\mathbf{Q} + \mathbf{Z}) \leq \Omega_{\phi,\theta}(\mathbf{Q}) + \Omega_{\phi,\theta}(\mathbf{Z}),\ \forall \mathbf{Q}, \mathbf{Z}$.
(iv) *Convex with respect to* $\mathbf{Z}$.
(v) *The infimum in* (1.47) *can be achieved with* $r \leq \text{card}(\mathbf{Z})$, $\forall \mathbf{Z}$.

Further, if for any choice of weights for one subnetwork, $\bar{\mathbf{W}}^{[1]}, \ldots, \bar{\mathbf{W}}^{[L]}$, *there exists a vector* $\mathbf{s} = \{-1, 1\}^L$ *such that* $\phi(\mathbf{X}, s_1\bar{\mathbf{W}}^{[1]}, \ldots, s_L\bar{\mathbf{W}}^{[L]}) = -\phi(\mathbf{X}, \bar{\mathbf{W}}^{[1]}, \ldots, \bar{\mathbf{W}}^{[L]})$ *and* $\theta(s_1\bar{\mathbf{W}}^{[1]}, \ldots, s_L\bar{\mathbf{W}}^{[L]}) = \theta(\bar{\mathbf{W}}^{[1]}, \ldots, \bar{\mathbf{W}}^{[L]})$, *then* $\Omega_{\phi,\theta}(\mathbf{Z})$ *is also a norm on* $\mathbf{Z}$.

Note that regardless of whether $\Omega_{\phi,\theta}$ is a norm or not, we always have that $\Omega_{\phi,\theta}(\mathbf{Z})$ is convex on $\mathbf{Z}$. Therefore, if the loss $\mathcal{L}$ is convex on $\mathbf{Z}$, so is the problem

$$\min_{\mathbf{Z}} \mathcal{L}(\mathbf{Y}, \mathbf{Z}) + \lambda\Omega_{\phi,\theta}(\mathbf{Z}). \tag{1.48}$$

[14] The first two conditions are typically easy to verify, while the third condition is needed to avoid trivial situations such as $\Omega_{\phi,\theta}(\mathbf{Z}) = 0,\ \forall \mathbf{Z}$.

Further, just as in the previous matrix factorization example we have that for any $(\mathbf{Z}, \mathbf{W}^{[1]}, \ldots, \mathbf{W}^{[L]})$ such that $\mathbf{Z} = \Phi_r(\mathbf{X}, \mathbf{W}^{[1]}, \ldots, \mathbf{W}^{[L]})$ the following global lower-bound exists:

$$\mathcal{L}(\mathbf{Y}, \mathbf{Z}) + \lambda \Omega_{\phi,\theta}(\mathbf{Z}) \leq \mathcal{L}(\mathbf{Y}, \Phi_r(\mathbf{X}, \mathbf{W}^{[1]}, \ldots, \mathbf{W}^{[L]})) + \lambda \Theta_r(\mathbf{W}^{[1]}, \ldots, \mathbf{W}^{[L]}), \tag{1.49}$$

and the lower-bound is tight in the sense that for any $\mathbf{Z}$ such that $\Omega_{\phi,\theta}(\mathbf{Z}) \neq \infty$ there exists $(\mathbf{W}^{[1]}, \ldots, \mathbf{W}^{[L]}, r)$ such that $\mathbf{Z} = \Phi_r(\mathbf{X}, \mathbf{W}^{[1]}, \ldots, \mathbf{W}^{[L]})$ and the above inequality becomes an equality. As a result, using a very similar analysis as was used to prove Theorem 1.5 one can show the following result:

Theorem 1.9 (Haeffele and Vidal (2015, 2017)) *Given a nondegenerate pair of functions (ϕ, θ), if $(\mathbf{W}^{[1]}, \ldots, \mathbf{W}^{[L]}, r)$ is a local minimum of* (1.46) *such that for some $i \in [r]$ we have $\mathbf{W}_i^{[1]} = 0, \ldots, \mathbf{W}_i^{[L]} = 0$, then (i) $\mathbf{Z} = \Phi_r(\mathbf{X}, \mathbf{W}^{[1]}, \ldots, \mathbf{W}^{[L]})$ is a global minimum of* (1.48) *and (ii) $(\mathbf{W}^{[1]}, \ldots, \mathbf{W}^{[L]}, r)$ is a global minimum of* (1.46) *and* (1.47).

We refer to Haeffele and Vidal (2015, 2017) for the full proof, but it closely follows the sequences of arguments from the proof of Theorem 1.5. In particular, we note that the key property which is needed to generalize the proof of Theorem 1.5 is that ϕ and θ are positively-homogeneous functions of the same degree.

Additionally, building on the discussion of the meta-algorithm from section 1.4.1, it can also be shown (in combination with Theorem 1.9) that if the network is sufficiently large (as measured by the number of subnetworks, r) then there always exists a path from any initialization to a global minimizer which does not require one to increase the value of the objective.

Theorem 1.10 (Haeffele and Vidal (2015, 2017)) *Given a nondegenerate pair of functions (ϕ, θ), let $|\phi|$ denote the number of elements in the output of the function ϕ. Then if $r > |\phi|$ for the following optimization problem:*

$$\min_{\mathbf{W}^{[1]}, \ldots, \mathbf{W}^{[L]}} \mathcal{L}(\mathbf{Y}, \Phi_r(\mathbf{X}, \mathbf{W}^{[1]}, \ldots, \mathbf{W}^{[L]})) + \lambda \Theta_r(\mathbf{W}^{[1]}, \ldots, \mathbf{W}^{[L]}) \tag{1.50}$$

a non-increasing path to a global minimizer will always exist from any initialization.

Again we refer to Haeffele and Vidal (2015, 2017) for the complete proof, but the key idea is that once one arrives at a local minimum either the condition of Theorem 1.9 is satisfied or if not the outputs of the subnetworks will be linearly dependent. As a result, by positive homogeneity one can traverse a flat surface of the objective landscape until arriving at a point which does satisfy the condition of Theorem 1.9. From there the point is

either a local minimum (and hence a global minimum from the Theorem) or a descent direction must exist.

1.5 Conclusions

We have studied the optimization landscape of neural network training for two classes of networks: linear networks trained with the squared loss and without regularization, and positively homogeneous networks with parallel structure trained with a convex loss and positively homogeneous regularization. In the first case, we derived conditions on the input-output covariance matrices under which all critical points are either global minimizers or saddle points. In the second case, we showed that when the networks is sufficiently wide, the non-convex objective on the weights can be lower-bounded by a convex objective on the network mapping, and derived conditions under which local minima of the non-convex objective yield global minima of both objectives. Future avenues for research include extending the results presented here to other classes of deep architectures. In particular, current results are limited to parallel architectures whose size is measured by the number of parallel subnetworks of fixed depth and width. This motivates extending the framework to cases in which both the depth and width of the network are varied. Moreover, the landscape of the objective is only one of the ingredients for explaining the role of optimization in deep learning. As discussed in the introduction, other ingredients are to develop efficient algorithms for finding a global minimum and to study the implicit regularization and generalization performance of such algorithms. We refer the reader to other chapters in this book for recent results on these fascinating subjects.

Acknowledgments

The authors acknowledge partial support by the NSF-Simons Research Collaborations on the Mathematical and Scientific Foundations of Deep Learning (NSF grant 2031985), the NSF HDR TRIPODS Institute for the Foundations of Graph and Deep Learning (NSF grant 1934979), NSF grants 1704458 and 2008460, and ARO grant MURI W911NF-17-1-0304.

Bibliography

Allen-Zhu, Zeyuan, Li, Yuanzhi, and Song, Zhao. 2019. A convergence theory for deep learning via over-parameterization. Pages 242–252 of: *International Conference on Machine Learning*.

Arora, Sanjeev, Cohen, Nadav, Hu, Wei, and Luo, Yuping. 2019. Implicit regularization in deep matrix factorization. Pages 7413–7424 of: *Neural Information Processing Systems*.

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

Candès, E., and Tao, T. 2010. The Power of Convex Relaxation: Near-Optimal Matrix Completion. *IEEE Transactions on Information Theory*, **56**(5), 2053–2080.

Cavazza, Jacopo, Haeffele, Benjamin D, Lane, Connor, Morerio, Pietro, Murino, Vittorio, and Vidal, Rene. 2018. Dropout as a Low-Rank Regularizer for Matrix Factorization. Pages 435–444 of: *International Conference on Artificial Intelligence and Statistics*, vol. 84.

Choromanska, Anna, Henaff, MIkael, Mathieu, Michael, Ben Arous, Gerard, and LeCun, Yann. 2015. The Loss Surfaces of Multilayer Networks. Pages 192–204 of: *International Conference on Artificial Intelligence and Statistics*.

Dauphin, Yann N, Pascanu, Razvan, Gulcehre, Caglar, Cho, Kyunghyun, Ganguli, Surya, and Bengio, Yoshua. 2014. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. Pages 2933–2941 of: *Neural Information Processing Systems*.

Donoho, David L. 2006. For Most Large Underdetermined Systems of Linear Equations the Minimal ℓ^1-norm Solution is also the Sparsest Solution. *Communications on Pure and Applied Mathematics*, **59**(6), 797–829.

Du, Simon, Lee, Jason, Li, Haochuan, Wang, Liwei, and Zhai, Xiyu. 2019. Gradient descent finds global minima of deep neural networks. Pages 1675–1685 of: *International Conference on Machine Learning*.

Duchi, J., Hazan, E., and Singer, Y. 2017. Adaptive Subgradient Methods of Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, **12**, 2121–2159.

Ge, Rong, Huang, Furong, Jin, Chi, and Yuan, Yang. 2015. Escaping from saddle points—online stochastic gradient for tensor decomposition. Pages 797–842 of: *Conference on Learning Theory*.

Gori, M., and Tesi, A. 1991. Backpropagation converges for multi-layered networks and linearly-separable patterns. Page 896 of: *International Joint Conference on Neural Networks*, vol. 2. IEEE.

Gori, M., and Tesi, A. 1992. On the problem of local minima in backpropaga-

tion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **14**(1), 76–86.

Gunasekar, Suriya, Woodworth, Blake E, Bhojanapalli, Srinadh, Neyshabur, Behnam, and Srebro, Nati. 2017. Implicit regularization in matrix factorization. Pages 6151–6159 of: *Neural Information Processing Systems*.

Gunasekar, Suriya, Lee, Jason, Soudry, Daniel, and Srebro, Nathan. 2018a. Characterizing implicit bias in terms of optimization geometry. In: *International Conference on Machine Learning*.

Gunasekar, Suriya, Lee, Jason D, Soudry, Daniel, and Srebro, Nati. 2018b. Implicit bias of gradient descent on linear convolutional networks. Pages 9461–9471 of: *Advances in Neural Information Processing Systems*.

Haeffele, Benjamin D, and Vidal, René. 2015. Global Optimality in Tensor Factorization, Deep Learning, and Beyond. *arXiv preprint arXiv:1506.07540*, **abs/1506.07540**.

Haeffele, Benjamin D, and Vidal, Rene. 2017. Global Optimality in Neural Network Training. Pages 7331–7339 of: *IEEE Conference on Computer Vision and Pattern Recognition*.

Haeffele, Benjamin D, and Vidal, René. 2019. Structured low-rank matrix factorization: Global optimality, algorithms, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **42**(6), 1468–1482.

Kawaguchi, Kenji. 2016. Deep learning without poor local minima. Pages 586–594 of: *Neural Information Processing Systems*.

Kingma, Diederik, and Ba, Jimmy. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Laurent, Thomas, and Brecht, James. 2018. Deep Linear Networks with Arbitrary Loss: All Local Minima Are Global. Pages 2902–2907 of: *International Conference on Machine Learning*.

Lee, Jason D, Panageas, Ioannis, Piliouras, Georgios, Simchowitz, Max, Jordan, Michael I, and Recht, Benjamin. 2019. First-order methods almost always avoid strict saddle points. *Mathematical Programming*, 1–27.

Lu, Haihao, and Kawaguchi, Kenji. 2017. Depth creates no bad local minima. *arXiv preprint arXiv:1702.08580*.

Mairal, Julien, Bach, Francis, Ponce, Jean, and Sapiro, Guillermo. 2010. Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research*, **11**, 19–60.

Mianjy, Poorya, Arora, Raman, and Vidal, René. 2018. On the implicit bias of dropout. In: *International Conference on Machine Learning*.

Nesterov, Y. 1983. A Method of Solving a Convex Programming Problem with Convergence Rate $O(1/k^2)$. *Soviet Mathematics Doklady*, **27**(2), 372–376.

Nouiehed, Maher, and Razaviyayn, Meisam. 2018. Learning deep models: Critical points and local openness. *arXiv preprint arXiv:1803.02968*.

Pal, Ambar, Lane, Connor, Vidal, René, and Haeffele, Benjamin D. 2020. On the regularization properties of structured dropout. Pages 7671–7679 of: *IEEE Conference on Computer Vision and Pattern Recognition*.

Robbins, H., and Monro, S. 1951. A stochastic approximation method. *Annals of Mathematical Statistics*, **22**, 400–407.

Rumelhart, David E, Hinton, Geoffrey E, and Williams, Ronald J. 1988. Learning representations by back-propagating errors. *Cognitive Modeling*, **5**.

Srebro, Nathan, Rennie, Jason DM, and Jaakkola, Tommi S. 2004. Maximum-Margin Matrix Factorization. Pages 1329–1336 of: *Neural Information Processing Systems*, vol. 17.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, **15**(1), 1929–1958.

Sun, J., Qu, Q., and Wright, J. 2017. Complete Dictionary Recovery over the Sphere I: Overview and the Geometric Picture. *IEEE Transactions on Information Theory*, **63**(2), 853–884.

Sun, Ju, Qu, Qing, and Wright, John. 2016. A Geometric Analysis of Phase Retrieval. *Forthcoming*.

Werbos, PJ. 1974. *Beyond regression: New tools for predictions and analysis in the behavioral science. Cambridge, MA, itd.* Ph.D. thesis, Harvard University.

Wright, Stephen J, and Nocedal, Jorge. 1999. *Numerical Optimization*. Vol. 2. Springer New York.

Xu, Yangyang, and Yin, Wotao. 2013. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on Imaging Sciences*, **6**(3), 1758–1789.

Zhang, Yuqian, Kuo, Han-Wen, and Wright, John. 2018. Structured Local Minima in Sparse Blind Deconvolution. Pages 2322–2331 of: *Advances in Neural Information Processing Systems*.

Zhu, Zhihui, Soudry, Daniel, Eldar, Yonina C., and Wakin, Michael B. 2019. The Global Optimization Geometry of Shallow Linear Neural Networks. *Journal of Mathematical Imaging and Vision*, May.