On the Optimization Properties of Optimal-Transport Based Generative Adversarial Linear Networks

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

DOCTOR RERUM NATURALIUM
(Dr. rer. nat.)

im Fachgebiet

Informatik

Vorgelegt

von M. Sc. Pierre Bréchet

Die Annahme der Dissertation wurde empfohlen von:

1. Prof. Dr. Richard Y. Zhang (University of Illinois Urbana-Champaing)
2. Prof. Dr. Shayan Mukherjee (Universität Leipzig)

Die Verleihung des akademischen Grades erfolgt mit Bestehen
der Verteidigung am 18.12.2025 mit dem Gesamtprädikat magna cum laude.

Abstract

This thesis is concerned with the analysis of the optimization problem in a class of generative models called Generative Adversarial Networks (GANs). GANs are successful generative models used to learn to mimic a given sampled distribution, and can be linked to Optimal Transport (OT) by the loss they rely on. OT is a reliable means to quantitatively compare two given probability distributions. The overall optimization problem is challenging as it involves a minimax optimization, in a generally highly non-convex-concave framework. In order to alleviate the analysis issues raised by this setting, several means can be employed. This thesis explores some of them in order to study a simplified OT-based generative model.

After introducing relevant mathematical tools in Chapter 2, we present in Chapter 3 the general framework of GANs, as well as the different simplifications that can be employed in order to derive analytical guarantees on the optimization system. Two examples, one discrete and one continuous, are investigated in order to illustrate how the analytical framework can be used in very simplified settings.

A particularly useful simplification is to reduce the minimax problem to a minimization problem only, by considering the maximization problem as being optimally solved at each time. In OT-based GAN, this typically boils down to analysing a type of distance between probability distributions, the Wasserstein distance. The minimization problem corresponds to finding a distribution that minimizes this distance to a given, target distribution. Chapters 4 and 5 leverage this simplification and study the landscape and convergence properties of a particular, linear OT-based GAN. In this case, the optimization problem can be cast as an overparametrized matrix factorization problem, combining a deep linear neural network with a covariance matrix parametrizations.

The landscape of the resulting loss, based on the so-called Bures-Wasserstein distance between the covariance matrices, is investigated in Chapter 4. More precisely, the critical points of the loss are derived, under the appropriate parametrizations. Given a rank constraint on the model covariance matrix, we make explicit the global minimizer to the matrix approximation problem. We find that they are are identical to the minimizers when considering a Euclidean distance between the covariance matrices. The Hessian of the loss at the critical points is derived and numerically investigated, in order to understand what differences the Euclidean and the Bures-Wasserstein distances may introduce in the optimization.

In Chapter 5, we analyze the gradient flow and gradient descent optimization schemes of this deep linear neural network under the Bures-Wasserstein loss. Convergence to

minimizers of the loss is proven, under additional assumptions on the initialization of the scheme. The rank constraint on the approximation matrix appears here naturally as a bottleneck in the network representing the model covariance matrix.

Complementary to the main discussion about GANs, the thesis also includes a chapter on the optimization landscape in a supervised learning setting (Chapter 6 on mode connectivity). Its conclusions may be linked to the GAN discussion in the simplified, limited setting of a minimization-only optimization scheme.

Authorship

This thesis is based on several published joint works as detailed below.

Chapters 2 and 3 are previously unpublished.

Chapters 4 and 5 are based on the work [BPA+23] published at the International Conference on Machine Learning (ICML) conference in collaboration with Katerina Papagianouli, Jing An, and Guido Montúfar. Sections 4.5 and 4.6 do not appear in the publication, and are my own work. Section 5.4.2 is a newly rewritten section which contains corrections to parts of the analysis of gradient descent in the aforementioned publication.

Chapter 6 has been published as [NBM21] at the Neural Information Processing Systems (NeurIPS) conference in collaboration with Quynh Nguyen and Marco Mondelli. In this work, I was only responsible for the numerical experiments (Sections 6.6 and 6.B).

Acknowledgments

This thesis would not have been possible without my advisor, Guido Montúfar. I would like to thank him for his support in all circumstances, the fruitful discussions, the vision he had on so many topics. He made everything run smoothly, even when we were geographically separated. I would also like to thank the members of my Thesis Advisory Committee, Jürgen Jost, Jörg Lehnert, and my TAC mentor André Uschmajew, who gave me valuable insights on my working topics. I would also like to particularly thank Katerina Papagiannouli for her help, both professional and emotional. The bond we have developed allowed me to overcome the most challenging times. I am very grateful to the people I had to chance to be closely working with. They all enabled me to extend my research abilities: Quyhn Nguyen, Marco Mondelli, and Jing An. The distance separating us was did not prevent us from solving interesting problems.

My office mates made me enjoy my research journey day by day: Johannes Müller, Celia Hacker, Hanna Tseran, and Friedrich Wicke, The institute in which I had the chance to work for this thesis provided me with tremendous resources and the people working there are the most caring I could imagine. I would like to especially thank the librarians for their always helpful knowledge and the coffee they provide, the IT crew for the precious technical support in times of need, as well as Katharina Matschke at the institute and Ricarda Lohmann at the university, for always being available and willing to help for administrative matters.

And, last but not least, I would like to thank my partner, my friends, and my family, who suffered me whole along and provided me the emotional support

Funding acknowledgments This thesis has been funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no 757983). The last year has been funded by the University Leipzig.

Contents

Abstract i

1 Introduction 1
Notations . . . 4

2 Mathematical framework 7
2.1 General analysis framework . . . 7
2.2 Optimal transport primer . . . 19

3 Optimization of general Generative Adversarial Networks 27
3.1 Introduction to the GAN framework . . . 27
3.2 Optimization procedure . . . 34
3.3 Optimal discriminator hypothesis . . . 37
3.4 Examples . . . 40
3.5 Conclusion . . . 47

4 Bures-Wasserstein loss landscape 49
4.1 Motivation . . . 50
4.2 Wasserstein generative linear networks . . . 53
4.3 Expression of the critical points . . . 59
4.4 Proofs of section 4.3 . . . 62
4.5 BW and the Eckart-Young-Mirsky theorem . . . 72
4.6 Expression and empirical evaluation of the Hessians . . . 78
4.7 Conclusion . . . 89

5 Convergence of a linear Bures-Wasserstein GAN 91
5.1 Previous work . . . 91
5.2 Gradient learning in deep linear networks . . . 93
5.3 Boundedness of the Bures-Wasserstein loss . . . 96
5.4 Convergence analysis . . . 100
5.5 Proofs of convergence results in Section 5.4 . . . 105
5.6 Experimental evaluation of the convergence rate . . . 124
5.7 Conclusion . . . 126

6 Mode connectivity 129
6.1 Introduction . . . 129
6.2 Related Work . . . 132

6.3 Preliminaries . . . 132
6.4 Main Result: Statement and Discussion . . . 134
6.5 Proof of Theorem 6.1 . . . 138
6.6 Numerical Experiments . . . 140
6.7 Concluding Remarks . . . 142
6.A Missing Proofs . . . 143
6.B Additional Numerical Experiments . . . 145
6.C Training Details . . . 146

Conclusion 151

Index 153

Bibliography 155

1 Introduction

We provide here an overview of the different topics, motivations, and goals of the thesis.

The main topic of this thesis is the study of Generative Adversarial Networks (GANs) — an instance of implicit generative modelling — and their related optimization properties. Given samples from an otherwise unknown data distribution, a GAN's goal is to learn how to generate samples that are indistinguishable from the original distribution. Therefore, at the heart of such a learning system lies a means of comparing two probability distributions. A celebrated mathematical tool for doing so is Optimal Transport (OT), which leverages a ground cost from the sample space onto the corresponding space of distributions. The presentation of the core concepts and theorems related to OT is given in Chapter 2. The computation of OT quantities is involved, and often requires adequate relaxations or simplifications. Moreover, the optimization of a GAN is challenging to properly analyze, as it typically involves a minimization-maximization problem of a highly non-convex-non-concave objective.

Chapter 3 provides a general introduction on GANs, and the motivations for different wide-spread losses. It also summarizes the usual strategies used to alleviate the inherent optimization challenges mentioned above. The optimal discriminator hypothesis, for instance, reduces the minimization-maximization system of a GAN to a minimization-only problem, which is more amenable to analyze. Under this hypothesis, regularity assumptions on the objective can be spelled out in order to enable the convergence of a gradient-descent algorithm to converge towards a critical point of the loss. Nonetheless, a toy example presented later in this chapter illustrates how these conditions can be violated even in a very simple setting. The study of such an example also aims at gaining intuition on how to analyze more complex models.

Given the intricacy of a GAN optimization scheme, and the OT computational challenges, we adopt simplifying assumptions in the following two chapters in order to analyse an OT-based GAN. First, we focus on the generator network's analysis, by assuming the optimal discriminator hypothesis. Second, we consider a linear architecture for the generator which, together with a suitable latent distribution, reduces the Wasserstein loss expression to a closed-form computation: the Bures-Wasserstein distance (BW) between covariance matrices. Together, these assumptions lead to the study of a Bures-Wasserstein linear GAN. Since it is defined on covariance matrices, the BW loss requires a covariance (or Burer-Monteiro) parametrization, which is seamlessly encoded by the linear GAN we consider. This parametrization comes on top of the one defining a deep linear network. Although dramatically simpler to analyse than a regular GAN, this learning system is

still worth of study, since it combines a non-Euclidean loss landscape, a rank-constrained matrix approximation problem, a covariance parametrization, and an overparametrized optimization scheme. We split its examination in two parts.

In Chapter 4, we study the landscape of the BW loss; namely, we give an expression of the critical points of the loss (restricted to a fixed-rank set of matrices), and inspect some of their properties.

Since the differential of the BW loss is not continuous everywhere, we have to also consider a regularized version of the loss, of which we also derive the critical points of a given rank. Interestingly, the rank-restricted critical points of the (original) BW loss coincide with the critical points found when, instead, a loss derived from the Euclidean geometry of the space of matrices — the Frobenius, or "squared error" loss — is considered. In particular, the (rank-restricted) minimizers for the two losses are the same. Since the Frobenius loss is expressed as (the square of) an unitary invariant norm, its rank-constrained minimizers' expression can be derived by the Eckart-Young-Mirsky theorem. In the case of the BW loss on the other hand, a direct application of the theorem is not possible. We therefore investigate in more detail from where this correspondence comes, and find out its origin in the (Euclidean) ground cost leveraged by the Wasserstein distance. Although the two landscapes share their critical points, the Hessian of the losses at these loci are different. We derive the Hessian matrix expression for the BW loss (in all parameter, function, and covariance spaces), and experimentally evaluate its difference with the Frobenius loss. The derivations also allows to express the Hessian of any loss relying on the covariance (or Burer-Monteiro) parametrization.

Once the landscape of the BW loss has been inspected, in Chapter 5, we turn to the dynamical analysis of the corresponding learning system. More specifically, we consider first-order optimization algorithms on the parameters of the network (gradient flow and gradient descent) and derive convergence guarantees towards minimizers of the BW loss. In order to limit the number of parameters that correspond to a unique matrix (due to the symmetry of the deep-linear network parametrization), assumptions are made, depending on the learning algorithm considered.

The deep-linear-network parametrization has many symmetries, and many different weights correspond to the same matrix. We restrict this by making assumptions, invariant during training, on the weights. We obtain two different types of results, corresponding to two different sets of hypothesis. First, for the gradient flow, we require an (arguably strong) assumption on the weights, which implies that the parametrized covariance matrix remains strictly positive throughout training. In this setting, we obtain an exponential rate of convergence towards the global minimizer of the original BW loss. Second, for the gradient descent, we only assume a bounded initialization, but have to consider in turn the regularized loss to ensure the well-posedness of the optimization problem. Convergence is shown to a minimizer of the considered loss, which is global over the set of same-or-less rank matrices. Experimental evaluations back the theoretical exponential

rate of convergence towards the global minimizer, and its accelerated dependency on the depth of the network.

Due to the different symmetries (of the deep linear network and covariance parametrizations), the actual critical points in the parameter space are infinite. In many other over-parametrized context, the learning problem has many more than one solution. Chapter 6 constructively exhibits a path connecting any two parameters achieving a small loss, in the supervised learning framework. Numerical computations run on a multitude of networks and different classification tasks show that very often, stochastic gradient descent finds such low-loss-achieving parameters.

Notations

The following notations will be used throughout the text.

E, F, G	normed (Banach) vector spaces; 8, 10
$\mathscr{L}(E, F)$	set of bounded linear operators between E and F; 11
$\operatorname{dom} \varphi$	effective domain of $\varphi\colon E \to (-\infty, \infty]$; 8
(E, E^*)	topologically paired vector spaces; 9
φ^*	convex conjugate of φ; 9
$\partial\varphi(\bar{x})$	subdifferential of φ at $\bar{x} \in E$; 10
$\|A\|$	operator norm for $A\colon E \to F$; 11
E'	topological dual of E; 11
$\langle \cdot, \cdot \rangle_{E^*,E}$	duality pairing between E and E^*; 11, 19
$\mathscr{C}(X)$	real-valued continuous functions on X; 11, 11
class $\mathscr{C}^k$, for $k \in \mathbb{N}_0$	real-valued functions that have a kth derivative; 58
$\mathscr{C}_0(X)$	continuous function vanishing at infinity on X; 11
$\mathscr{M}(X)$	signed measures on X; 12
$\mathscr{M}_+(X)$	positive measures on X; 12
$\mathscr{P}(X)$	probability measures on X; 19
$\mathscr{P}_2(X)$	square integrable probability measures on X; 24, 75
$L^1(\mu)$	μ-integrable functions, for $\mu \in \mathscr{M}(X)$; 12
$\cong$	isomorphism equivalence relation; 13
$\mathrm{p}_i\colon X_1 \times X_2 \to X_i$	Cartesian projection; 14
$\bar{\mathrm{p}}_{\mathcal{C}}\colon \mathcal{C} \subseteq \mathbb{R}^n \to \mathbb{R}^n$	orthogonal projection onto $\mathcal{C}$; 75
A^*	adjoint operator of $A \in \mathscr{L}(E, F)$; 14
$T_{\#}\mu$	push-forward operator of $\mu \in \mathscr{M}(X)$; 14, 32, 75
$\chi[P]$	convex indicator function of the boolean value P; 17
$\Gamma(\mu, \nu)$	admissible transport plans between μ and ν; 18
$\mathcal{U}_{L^1}(c)$	admissible function variables in L^1 with cost c; 20
$\mathcal{U}_{\mathscr{C}_0}(c)$	admissible function variables in $\mathscr{C}_0$ with cost c; 18
u^c	c-transform of u; 21
$u^{(p)}$	p-transform of u 22
$\operatorname{Lip}(u)$	Lipschitz constant of u; 23, 41
$\operatorname{Lip}_1(X)$	functions on X with Lipschitz constant upper bounded by 1; 23
$\mathscr{N}(\boldsymbol{m}, \Sigma)$	Gaussian distribution with mean $\boldsymbol{m}$ and covariance matrix Σ; 24, 54
$\mathbb{S}_+(n)$	symmetric positive semi-definite matrices of size n; 24, 52
$\Theta \subseteq \mathbb{R}^p$, $\Psi \subseteq \mathbb{R}^q$	parameter spaces; 27
$Z \subset \mathbb{R}^{d_z}$	latent space; 27
$X \subset \mathbb{R}^{d_x}$	sample space; 27
$\omega \in \mathscr{P}(Z)$	latent distribution; 28
$\mu_\theta \in \mathscr{P}(X)$	generated distribution with parameter $\theta \in \Theta$; 32
$\nu \in \mathscr{P}(X)$	target distribution; 32
δ_x	Dirac distribution at x; 28

Φ_μ	optimal discriminator for μ; 39
$\mathbb{N}$	set of positive integers; 40
$\mathbb{N}_0$	$= \mathbb{N} \cup \{0\}$
$[a:b]$	$= \{i \in \mathbb{N} \mid a \leqslant i \leqslant b\}$; 104
$[n]$	$= [1:n]$, set of integers between 1 and n; 40, 55
$\Delta_{n-1} \equiv \mathscr{P}([n])$	simplex; 40
$\mathbb{1}[P]$	indicator for the set or condition P; 41
$\bar{\delta}$	discrete metric on $[n]^2$; 41
$\mathscr{U}(X)$	uniform distribution on X, measurable set of finite measure; 42
1_n	$= (1, \dots, 1)^\top \in \mathbb{R}^n$; 46
δ_{ij}	Kronecker delta; 46
$\mathbb{M}(n,m)$	set of real matrices of size $n \times m$
$\mathbb{M}(n)$	$\equiv \mathbb{M}(n,n)$; 83, 74
$\mathcal{M}((\leqslant)k;n,m)$	the set of real $n \times m$ matrices of rank $(\leqslant)k$, for $k \in [\min(n,m)]$; 59
$\mathcal{M}((\leqslant)k)$	$\equiv \mathcal{M}((\leqslant)k;n,m)$; 59
$\mathcal{M}_*$	$\equiv \mathcal{M}(\min(n,m))$; 96
$\mathbb{S}(n)$	set of symmetric real matrices of size n; 52
$\mathbb{S}_{++}(n)\,(\mathbb{S}_+(n))$	set of symmetric (semi-)positive real matrices of size n; 24, 52
$\mathbb{O}(n)$	set of real orthogonal matrices of size n; 100, 100
$\langle A, B\rangle$	$= \mathrm{tr}(A^\top B)$, for $(A,B) \in \mathbb{M}(n,m)^2$; 87, 59
I_n	identity matrix of size n
$\mathrm{T}_x X$	tangent space of a smooth manifold X at $x \in X$; 62
$\mathrm{d}f(x)[v]$	differential at $x \in X$ in the direction $v \in \mathrm{T}_x X$ of $f\colon X \to Y$; 62
$\mathrm{Crit}(f)$	$= \{x \in X \mid \mathrm{d}f(x) = 0\}$, set of critical points of $f\colon X \to Y$; 59
K_n	commutation matrix, of size $n^2 \times n^2$; 80
$A \odot B$	Hadamard product between $A \in \mathbb{M}(n,m)$ and $B \in \mathbb{M}(n,m)$; 84, 63
$\sigma_{\{\min,\max\}}(A)$	extremal singular values of a matrix A; 52
$\lambda_{\{\min,\max\}}(A)$	extremal eigenvalues of a symmetric matrix A; 85
$\|A\| = \sigma_{\max}(A)$	spectral norm for $A \in \mathbb{M}(n,m)$; 95
$\mathbb{S}_{++}(n;\varepsilon)$	$= \{A \in \mathbb{S}_{++}(n) \mid \lambda_{\min}(A) \geqslant \varepsilon\}$

2 Mathematical framework

In this thesis, we will study some instances from the general problem of *generative modelling*, powered by neural network approximations. This field has attracted much of attention in the last decade, as it combines challenging theoretical aspects with powerful applications.

Several mathematical tools lie at the heart of the optimization analysis. We summarize them in this chapter. The general (convex) functional analysis are presented in Section 2.1. The goal of this section will be to introduce the celebrated Fenchel-Rockafellar duality theorem (Theorem 2.27), which formalizes the dual formulation of the optimization problem. Next, a general introduction to optimal transport is given in Section 2.2. The goal of this section will be to present the primal and dual optimal transport optimization problems, in the general and in a specific case.

2.1 General analysis framework

In this section, we summarize several key mathematical notions on which the thesis' derivations are based. They mostly originate from convex functional analysis on general Banach spaces. Indeed, it is in this context that the different optimization problems will be cast and analysed.

The aim of this section is threefold. First, in Section 2.1.1, different notions of convex analysis are presented. They will be used throughout the text. Second, in Section 2.1.2 are presented the duality mechanics and tools from functional analysis that will be used in order to cast a given problem in its dual expression. Third, the optimal transport problem is introduced and formalized in Section 2.2, and the tools from the previous sections are leveraged to derive the duality between the primal and dual problems.

2.1.1 Convex analysis, framework of the analysis, definitions

In this section, we recall the principal notions from convex analysis that will be useful for our later discussions and presentations. The interested reader can refer to [Roc70; RW98; HL01] for the Euclidean case, and [Roc67; Zal02; KZ05; ABM14] for the normed vector space case.

In what follows, $(E, \|\cdot\|_E)$ is a normed vector space, except otherwise indicated. When possible, the subscript in the norm notation will be dropped. Some results could be extended to a more general settings, where E is simply a Hausdorff and locally compact topological space. Nevertheless, as we will not be interested in non-metric spaces, we will state the different results in the metric-space case, which implies separability and local compactness

For instance, E could be the set of real-valued continuous functions on some compact metric space $X \subset \mathbb{R}^n$ with the usual metric topology, endowed with the supremum norm, i.e. $\forall \varphi \in E, \|\varphi\| = \sup\{|\varphi(x)| \mid x \in X\}$.

We will allow the real functions to take the value ∞, assuming the usual algebraic rules on $(-\infty, \infty]$. This leads to the notion of effective domain for an extended function.

Definition 2.1 (*Effective domain of a funtion*). Let E be a vector space. The *effective domain* of a function $\varphi\colon E \to (-\infty, \infty]$ is defined as

$$\operatorname{dom}\varphi := \{x \in E \mid \varphi(x) < \infty\}.$$

The effective domain of a function describes the set of *feasible solutions* for the function.

A subset of functions, named *proper*, will be of great importance.

Definition 2.2 (*Proper function*). A *proper* function is a function $\varphi\colon E \to (-\infty, \infty]$ that fulfils $\operatorname{dom}\varphi \neq \emptyset$.

A proper function on E is thus equivalently a real-valued function defined on a non-empty set $C \subseteq E$ that is then extended to ∞ on $E \setminus C$. This will allow to restrict the set of feasible solutions for an optimization problem, by simply setting the value of the functional to ∞ outside the constraint set.

Convex functions play a central role in optimization. We next give the two basic definitions of convexity.

Definition 2.3 (*Convex set*). A set $C \subseteq E$ is said to be *convex* if

$$\forall (x, y) \in C^2, \forall \lambda \in (0, 1), \quad \lambda x + (1 - \lambda) y \in C.$$

We now recall the definition of a convex function on some (normed) vector space E.

Definition 2.4 (*Convex function*). A (generally defined) function $\varphi\colon E \to (-\infty, \infty]$ is said to be *convex* if its epigraph $\operatorname{epi}\varphi := \{(x, t) \in E \times \mathbb{R} \mid \varphi(x) \leqslant t\}$ is a convex set of $E \times \mathbb{R}$.

A slightly more general property than continuity is often required in convex optimization; namely, the *lower semicontinuity*. It can be defined in the context of any topological space E, but we will use it only in the case where E is a normed vector space.

Definition 2.5 (*Lower semicontinuity*). A function $\varphi\colon E \to (-\infty, \infty]$ is said to be *lower semicontinuous* (l.s.c.) if, for every $t \in \mathbb{R}$, the set $\varphi^{-1}((-\infty, t]) = \{x \in E \mid \varphi(x) \leqslant t\}$ is closed.

Remark 2.6. The function defined by Definition 2.5 is also sometimes called *closed*.

A central notion will be heavily leveraged throughout the chapter. It pairs two topological vector spaces with a bilinear form, and ensures compatibility between the topology on the spaces and the duality introduced by the bilinear form.

Definition 2.7 (*Topologically paired vector spaces*). Let F and G be two real topological vector spaces. Given a bilinear form $(x, y) \in F \times G \mapsto \langle x, y\rangle \in \mathbb{R}$, we say that F and G are *topologically paired* if the maps $(x \mapsto \langle x, y\rangle) \in F'$ for y through G and $(y \mapsto \langle x, y\rangle) \in G'$ for x through F set up all possible continuous linear forms on F and G.

Remark 2.8. In the following, given a locally convex, separable topological space E, we shall use the notation E^* to denote any locally convex, separable topologically vector space that is paired with E, instead of the usual algebraic dual of E (i.e. the space of all linear forms on E), which will be denoted E', see Definition 2.15. The reason for this choice is to not restrict the following exposition to the pair (E, E'), and have a slightly more general and comprehensive framework.

More details on topological pairing can be found in, e.g., [Bou87, Chapter IV].

The next definition introduces the convex conjugate for functions defined on a general topological vector space.

Definition 2.9 (*Convex conjugate*). Let E and E^* be two topologically paired vector spaces. Denote by $\langle \cdot, \cdot \rangle$ the duality pairing between E and E^*. Given any proper function $\varphi\colon E \to (-\infty, \infty]$ (i.e. $\varphi \not\equiv \infty$), its *convex conjugate* (sometimes also called *Legendre transform*) φ^* is defined as

$$\begin{array}{rccl} \varphi^*\colon & E^* & \longrightarrow & (-\infty, +\infty], \\ & x^* & \longmapsto & \varphi^*(x^*) = \sup\{\langle x^*, x\rangle - \varphi(x) \mid x \in E\}. \end{array}$$

The convex conjugate is always a proper convex function. Moreover, $\varphi^{**} = \varphi$ if and only if φ is a l.s.c. convex function, which is leveraged in several derivations of convex analysis.

Proposition 2.10 (Fenchel-Moreau [Zal02, Theorem 2.3.3]). *Let $\varphi\colon E \to (-\infty, \infty]$ be a l.s.c. proper convex function. Then, $\varphi^{**} = \varphi$.*

A last notion, central in convex optimization, will be used in Chapter 3: the subdifferential.

Definition 2.11 (*Subdifferential*). Let (E, E^*) be two topologically paired vector spaces. Given a function $\varphi\colon E \to \mathbb{R}$, its *subdifferential* at $\bar{x} \in E$ is

$$\partial\varphi(\bar{x}) = \begin{cases} \{x^* \in E^* \mid \forall x \in E,\ \varphi(x) \geqslant \varphi(\bar{x}) + \langle x^*, x - \bar{x}\rangle\} & \text{if } \bar{x} \in \operatorname{dom}\varphi, \\ \emptyset & \text{else.} \end{cases}$$

Notably, from Proposition 2.10, one can easily derive the next proposition.

Proposition 2.12 *Let (E, E^*) be two topologically paired vector spaces, and let $\varphi\colon E \to (-\infty, \infty]$ be a l.s.c. convex function. Then, for all $(\bar{x}, \bar{x}^*) \in E \times E^*$, there is the following equivalence*

$$\bar{x}^* \in \partial\varphi(\bar{x}) \quad \Longleftrightarrow \quad \bar{x} \in \partial\varphi^*(\bar{x}^*) \quad \Longleftrightarrow \quad \langle \bar{x}^*, \bar{x}\rangle = \varphi(\bar{x}) + \varphi^*(\bar{x}^*).$$

Now, we have the core convex analysis definitions and properties that will be required throughout the thesis. In the next section, an overview of the functional analysis tools will be introduce.

2.1.2 Functional analysis

The aim of this section is to present the Fenchel-Rockafellar duality theorem (Theorem 2.27), which bridges the optimization between two dually connected problems. First, the Riesz-Markov-Kakutani representation theorem (Theorem 2.21) is required, in order to dually represent bounded linear functionals on (the closure of) compactly supported continuous functions.

With the additional assumption that the topological normed vector space E is complete (with respect to the strong topology induced by its norm), $(E, \|\cdot\|)$ becomes a *Banach space*, which definition is recalled next.

Definition 2.13 (*Banach space*). A normed vector space $(E, \|\cdot\|)$ is said to be *Banach* if it is complete, for the metric defined by the norm.

All of our analysis will take place on Banach spaces. For instance, the space of (bounded) real-valued continuous functions on a finite-dimensional, compact vector space $X \subseteq \mathbb{R}^n$, denoted by $\mathscr{C}(X)$, endowed with the supremum norm $\|\varphi\|_\infty := \sup\{|\varphi(x)| \mid x \in X\}$ is a well-known Banach space. Notice that this norm is relevant for continuous functions only if X is compact.

In the following chapters, we will deal with vector spaces that are not compact (for instance $\mathbb{R}^n$). In order to accommodate it, we need to restrict the set of functions that are considered; namely, we will deal with continuous functions that vanish at infinity.

Let X be separable, locally compact space. Denote by $\mathscr{C}(X)$ the set of real-valued continuous functions on X, endowed with the supremum norm. Denote further by $\mathscr{C}_0(X)$ real-valued continuous functions on X vanishing at infinity. More specifically,

$$\mathscr{C}_0(X) = \{f \in \mathscr{C}(X) \mid \forall \varepsilon > 0,\, \exists K \subset X,\, K \text{ compact},\, \|f|_{X\setminus K}\| < \varepsilon\}. \tag{2.1}$$

Note the following inclusion $\mathscr{C}_0(X) \subseteq \mathscr{C}(X)$, with equality if and only if X is compact.

As we will see shortly, the set of *bounded linear functionals* on $\mathscr{C}_0(X)$ is isomorphic to the set of *signed measures* on X. This is the key correspondence that is leveraged in the next section, and the rest of the present section is devoted to it.

Definition 2.14 (*Bounded linear operator*). A linear operator $A\colon E \to F$ between two Banach spaces $(E, \|\cdot\|_E)$ and $(F, \|\cdot\|_F)$ is said to be *bounded* if its operator norm is finite, i.e., if

$$\|A\| = \sup\{\|Ax\|_F \mid x \in E,\, \|x\|_E \leqslant 1\} < \infty. \tag{2.2}$$

Notation The set of all bounded linear operators between E and F is denoted by $\mathscr{L}(E, F)$.

Recall that a linear operator $A\colon E \to F$ is continuous if and only if it is bounded. We will therefore employ the terms bounded and continuous interchangeably for linear operators.

Definition 2.15 (*Topological dual*). Given a Banach space $(E, \|\cdot\|)$, its *topological dual* E' is the set of all bounded linear functionals defined on E; i.e.,

$$E' \equiv \mathscr{L}(E, \mathbb{R}) = \{x'\colon E \to \mathbb{R} \mid x' \text{ is linear and bounded}\}.$$

Remark 2.16. We adopt the notation E' instead of the arguably more widespread E^* for the topological dual in order to distinguish it later on from a generic vector space that is topologically paired with E.

The duality pairing between E and E' is denoted by $\langle\cdot,\cdot\rangle_{E'\times E}$; i.e., for all $x' \in E'$, we can equivalently write the action of x' on $x \in E$ as $x'(x)$ or $\langle x', x\rangle_{E'\times E}$. The subscript $E' \times E$ will be dropped when there is no confusion on the current spaces that are considered.

Remark 2.17. The spaces E and E' in Definition 2.15 are topologically paired vector spaces, in the sense of Definition 2.7.

This notion of duality is key in the expression of different optimization problems, as the topological dual can be identified with a space in which the optimization is more feasible as the first one, as will be stated by Theorem 2.21.

The set of finite signed measures on X is central for the following presentation, as it will be isomorphic to $\mathscr{C}_0(X)'$ (the set of bounded linear operators on $\mathscr{C}_0(X)$). We introduce here succinctly, by defining the finite signed measures on a set X. More details on measure theory for our purpose can be found in e.g. [Rud13].

Definition 2.18 (*Signed* (*positive*) *measure*). Let $\mathcal{A}$ be a σ-algebra on a set X. Then, a *signed* (resp. *positive*) measure μ is a function on $\mathcal{A}$ whose range is in $[-\infty, \infty]$ (resp. $[0, \infty]$) such that, for every countable disjoint union (or partition) of $A \in \mathcal{A}$ (i.e. with $A = \bigcup_{i\in\mathbb{N}} A_i$ and $i \neq j \implies A_i \cap A_j = \emptyset$), we have

$$\sum_{i\in\mathbb{N}} |\mu(A_i)| < \infty \quad \text{and} \quad \mu(A) = \sum_{i\in\mathbb{N}} \mu(A_i).$$

The set of signed (resp. positive) measures on X is denoted by $\mathscr{M}(X)$ (resp. $\mathscr{M}_+(X)$). For $\mu \in \mathscr{M}(X)$, the set of μ-integrable functions is denoted by $\mathrm{L}^1(\mu)$.

Definition 2.19 (*Total variation*). The total variation of a signed measure μ in a measurable space $(X, \mathcal{A})$ is the positive measure $|\mu|$ defined as

$$\forall A \in \mathcal{A}, \quad |\mu|(A) = \sup \left\{ \sum_{i\in\mathbb{N}} |\mu(A_i)| \;\middle|\; \{A_i\}_{i\in\mathbb{N}} \text{ is a partition of } A \right\}$$

By [Rud13, Theorem 6.4], if μ is a signed measure on X, then $|\mu|(X) < \infty$. Therefore, *any signed measure is finite*.

Remark 2.20. The presentation in [Rud13] considers complex functions, whereas we are confined to the real case. Therefore, the (more general) complex measures are here replaced with signed measures, sometimes also called real measures.

The next theorem finally states, for X a locally compact separable vector space, the duality between the set of continuous functions on X vanishing at infinity $\mathscr{C}_0(X)$ (2.1) and $\mathscr{M}(X)$, the set of signed measures on X (Definition 2.18). This is a restriction of [Rud13, Theorem 6.19] to the real case.

Theorem 2.21 (Riesz-Markov-Kakutani representation theorem for signed measures). *Let X be a locally compact separable space. Endow $\mathscr{C}_0(X)$ (2.1) with the supremum norm, so that $(\mathscr{C}_0(X), \|\cdot\|)$ is a Banach space. Then, every element its topological dual $\mathscr{C}_0(X)'$ is uniquely represented as an element of $\mathscr{M}(X)$; i.e., for any $\mathcal{F} \in \mathscr{C}_0(X)'$, there exits a unique $\mu \in \mathscr{M}(X)$ such that*

$$\forall f \in \mathscr{C}_0(X), \quad \mathcal{F}(f) = \langle \mu, f \rangle = \int_X f \, \mathrm{d}\mu.$$

Moreover, $\mathscr{C}_0(X)'$ is isomorphic to $\mathscr{M}(X)$, endowed with the norm $\|\mu\| := |\mu|(X)$, with $|\mu|$ the positive total variation measure defined by Definition 2.19. We write it $\mathscr{C}_0(X)' \cong \mathscr{M}(X)$; this means notably that $\mathscr{C}_0(X)$ and $\mathscr{M}(X)$ are topologically paired in the sense of Definition 2.7.

Remark 2.22. In the case where X is also compact, $\mathscr{C}_0(X) = \mathscr{C}(X)$, and the statement of Theorem 2.21 can be adapted accordingly.

Remark 2.23. Theorem 2.21 implies that $\mathscr{C}_0(X)$ and $\mathscr{M}(X)$ are topologically paired topological vector spaces. This will allow us to translate an optimization problem on $\mathscr{M}(X)$ to a problem on the set $\mathscr{C}_0(X)$ — or even $\mathscr{C}(X)$ when X is compact.

In the next example, we start to introduce the central spaces that will be relevant for the thesis analysis. This example will be run through the whole section, and will serve as basis for the next one.

Example 2.1 Let X_1 and X_2 be two locally compact separable spaces. Let $E = \mathscr{C}_0(X_1) \times \mathscr{C}_0(X_2)$ and $F = \mathscr{C}_0(X_1 \times X_2)$, both endowed with the supremum norm. Note that, for $(u, v) \in E$, $\|(u, v)\|_\infty = \max\{\sup\{|u(x_1)| \mid x_1 \in X_1\}, \sup\{|v(x_2)| \mid x_2 \in X_2\}\}$.

Then, by Theorem 2.21, we can identify their dual spaces as $E' \cong \mathscr{M}(X_1) \times \mathscr{M}(X_2)$ and $F' \cong \mathscr{M}(X_1 \times X_2)$. Notice that, for all $(u, v) \in E$ and all $(\alpha, \beta) \in E'$,

$$\langle (\alpha, \beta), (u, v) \rangle_{E' \times E} = \langle \alpha, u \rangle_{\mathscr{M}(X_1) \times \mathscr{C}_0(X_1)} + \langle \beta, v \rangle_{\mathscr{M}(X_2) \times \mathscr{C}_0(X_2)} = \int u \, \mathrm{d}\alpha + \int v \, \mathrm{d}\beta.$$

$\Diamond$

Remark 2.24. When, for some $(m, n) \in \mathbb{N}^2$, $X_1 \subseteq \mathbb{R}^m$ and $X_2 \subseteq \mathbb{R}^n$, the topological assumptions on X_1 and X_2 in Example 2.1 are satisfied. This will typically be the framework of the thesis.

With Theorem 2.21, we can identify $\mathscr{C}_0(X) \cong \mathscr{M}(X)$ for X a locally compact separable space. A last key theorem will conclude this section, namely the Fenchel-Rockafellar duality theorem. It will make a correspondence between the primal and dual optimization problems.

In order to present it, the notion of adjoint operator is next introduced.

Definition 2.25 (*Adjoint operator*). Given two Banach spaces E and F, if $A \in \mathscr{L}(E, F)$, its *adjoint operator* is $A^* \in \mathscr{L}(F', E')$ such that

$$\forall x \in E,\, \forall y' \in F', \quad \langle y', Ax\rangle_{F'\times F} = \langle A^* y', x\rangle_{E'\times E}$$

We now define the push-forward operation by an operator $T\colon X_1 \to X_2$. In the next example, we see that the adjoint of the pull-back operation is the push-forward operator, which is defined below.

Recall that a Borel set $(X, \mathcal{B})$ is a measurable space and $\mathcal{B}$ is the Borel σ-algebra on a topological space X; i.e., the σ-algebra generated by the open sets of X.

Definition 2.26 (*Push-foward operation*). Let $(X_1, \mathcal{B}_1)$ and $(X_2, \mathcal{B}_2)$ be two Borel sets. Let $T\colon X_1 \to X_2$ be a Borel map, and let μ be a Borel measure on X_1. Then, the *push-forward* of μ by T, denoted by $T_{\#}\mu$, is the Borel measure on X_2 such that

$$\forall B \in \mathcal{B}_2, \quad T_{\#}\mu(B) = \mu(T^{-1}(B)). \tag{2.3}$$

Equivalently,

$$\forall g\colon X_2 \to \mathbb{R},\, g \text{ Borel function}, \quad \int g \,\mathrm{d}T_{\#}\mu = \int g \circ T \,\mathrm{d}\mu.$$

Example 2.2 (*Example 2.1 continued*). Let $(E, \|\cdot\|_E)$ and $(F, \|\cdot\|_F)$ be the Banach spaces defined in Example 2.1.

Let $A \in \mathscr{L}(E, F)$ be the linear operator defined as

$$\begin{array}{rccl} A\colon & E & \longrightarrow & F, \\ & (u, v) & \longmapsto & u \circ \mathrm{p}_1 + v \circ \mathrm{p}_2, \end{array} \tag{2.4}$$

where, for $i \in \{1, 2\}$,

$$\begin{array}{rccl} \mathrm{p}_i\colon & X_1 \times X_2 & \longrightarrow & X_i, \\ & (x_1, x_2) & \longmapsto & x_i, \end{array} \tag{2.5}$$

is the projection onto X_i. Notice that A is continuous by continuity of the projection applications.

Then, the adjoint of A is

$$\begin{array}{rccl} A^*\colon & F' & \longrightarrow & E', \\ & \gamma & \longmapsto & (\mathrm{p}_{1\#}\gamma, \mathrm{p}_{2\#}\gamma). \end{array} \tag{2.6}$$

Indeed, for all $(u, v) \in E$ and all $\gamma \in F'$, letting $A^*\gamma = (\alpha, \beta) \in E'$, we have

$$\begin{aligned} & \langle \gamma, A(u,v) \rangle_{F' \times F} = \langle A^*\gamma, (u,v) \rangle_{E' \times E} \\ \iff & \int u \circ \mathrm{p}_1 \, \mathrm{d}\gamma + \int v \circ \mathrm{p}_2 \, \mathrm{d}\gamma = \int u \, \mathrm{d}\alpha + \int v \, \mathrm{d}\beta \\ \iff & \begin{cases} \alpha = \mathrm{p}_{1\#}\gamma \\ \beta = \mathrm{p}_{2\#}\gamma \end{cases}. \end{aligned}$$

This is no surprise, as the adjoint of the push-forward operator is the pull-back. The case we have at hand is simply the product space variant of this fact. ◊

We have presented the analysis tools that will be essential to our discussion. In the next section, the notion of *convex duality* will be detailed.

2.1.3 Optimization problem

We are now ready to present the general convex optimization framework that we will consider, and the duality herein.

Given

i) two Banach spaces E and F,
ii) two proper, lower semicontinuous convex functions $\varphi \colon E \to (-\infty, \infty]$ and $\psi \colon F \to (-\infty, \infty]$, and
iii) a bounded linear operator $A \in \mathscr{L}(E, F)$,

we are interested in the minimization problem

$$\mathrm{minimize} \, \{\varphi(x) + \psi(Ax) \mid x \in E\}.$$

Notation Throughout the text, the following convention will be followed. The operator "$\mathrm{minimize}$" will refer to the minimization problem, whereas the term "$\inf$" will refer to the value of the solution. If the bound is attained, we will also write "$\min$". Similarly, the terms "$\mathrm{maximize}$", "$\sup$" and "$\max$" will be used for maximization problems.

Since ψ is proper, l.s.c. and convex, by Fenchel-Moreau theorem (Proposition 2.10),

$\psi = \psi^{**}$. Therefore, one always has

$$\begin{aligned}
\inf_{x\in E} \varphi(x) + \psi(Ax) &= \inf_{x\in E} \varphi(x) + \psi^{**}(Ax) \\
&= \inf_{x\in E} \varphi(x) + \sup_{y'\in F'} \langle y', Ax\rangle - \psi^*(y') \\
&= \inf_{x\in E} \sup_{y'\in F'} \left\{\langle A^*y', x\rangle + \varphi(x) - \psi^*(y')\right\} \\
&\geqslant \sup_{y'\in F'} \inf_{x\in E} \left\{\langle A^*y', x\rangle + \varphi(x)\right\} - \psi^*(y') \qquad (2.7) \\
&= \sup_{y'\in F'} -\varphi^*(-A^*y') - \psi^*(y'),
\end{aligned}$$

where inequality (2.7) (called *weak duality*) is simply due to the inequality

$$\sup_y \inf_x f(x,y) \leqslant \inf_x \sup_y f(x,y),$$

true for all $f\colon X \times Y \to \mathbb{R}$, and the last equality holds by definition of φ^*. Strong duality is obtained when one has equality in (2.7). It is the case under additional (mild) assumptions on the problem, e.g. when X and Y are compact convex sets, and f is convex-concave. This is a special instance of a minimax theorem, of which a collection can be found in e.g. [Sim95].

We now state the main theorem of the section. It gives notably sufficient conditions for which we have equality in Equation (2.7), i.e., it allows to swap the "$\inf\sup$" problem to a "$\sup\inf$" in saddle optimization problem.

Theorem 2.27 (Fenchel-Rockafellar, [Roc67, Theorems 1 and 3]). *Let $\varphi\colon E \to (-\infty, \infty]$ and $\psi\colon F \to (-\infty, \infty]$ be two l.s.c, proper convex functions on two Banach spaces E and F, and let $A \in \mathscr{L}(E, F)$ be a continuous (bounded) linear operator from E to F. Assume that there exists $x_0 \in E$ such that φ is finite at x_0 and ψ is finite and continuous at Ax_0. Then,*

$$\begin{aligned}
&\inf\{\varphi(x) + \psi(Ax) \mid x \in E\} = \max\{-\varphi^*(-A^*y') - \psi^*(y') \mid y' \in F'\} \\
\iff &\min\{\varphi^*(-A^*y') + \psi^*(y') \mid y' \in F'\} = \sup\{-\varphi(x) - \psi(Ax) \mid x \in E\}
\end{aligned}$$

Remark 2.28. The statement in [Roc67] is more general, in the sense that it considers two pairs of topologically paired vector spaces (E, E^*) and (F, F^*) as defined by Definition 2.7, instead of the topological dual of Banach spaces E and F as we do: in our case, the duality is given by the algebraic dual space of a Banach space.

Note that the maximum is attained for the dual problem as part of the theorem statement. A simple proof of this theorem (in the case where $E = F$, $A = \mathrm{id}$) can be found in [Vil03, Theorem 1.9; Bre11, Theorem 1.12]. The general case proof can be found in [Roc67].

As per Theorem 2.21, the duality pairing at our disposition identifies, for X a locally compact separable space, the dual of $\mathscr{C}_0(X)$ with $\mathscr{M}(X)$.

For $P\colon E \to \{0,1\}$ a boolean function on E such that $P^{-1}(\{1\})$ is a convex set, we denote by

$$\chi[P(x)] = \begin{cases} 0 & \text{if } P(x) = 1, \\ \infty & \text{else} \end{cases} \tag{2.8}$$

the *characteristic function* of the condition. Notice that it is a convex function.

Continuing the running example study, we will therefore be able to apply this theorem with $E = \mathscr{C}_0(X_1)\times\mathscr{C}_0(X_2)$, $F = \mathscr{C}_0(X_1\times X_2)$, $A = (\circ\mathrm{p}_1 + \circ\mathrm{p}_2)$, $\varphi\colon (u,v) \mapsto -\langle\mu,u\rangle - \langle\nu,v\rangle$, and $\psi\colon w \mapsto \chi[w \leqslant c \text{ a.e.}]$

Example 2.3 (*Example 2.2 continued*). Let X_1 and X_2 be defined as in Example 2.1. Let μ and ν be two probability measures on X_1 and X_2, and let $c\colon X_1 \times X_2 \to [0,\infty)$ be a proper, lower semicontinuous function. For any $w \in \mathscr{C}_0(X_1 \times X_2)$, we will write "$w \leqslant c$ a.e." as a shorthand for the condition "for μ-almost all $x_1 \in X_1$ and ν-almost all $x_2 \in X_2$, $w(x_1,x_2) \leqslant c(x_1,x_2)$".

With the same spaces as in Example 2.2, i.e. $E = \mathscr{C}_0(X_1) \times \mathscr{C}_0(X_2)$, $E' \cong \mathscr{M}(X_1) \times \mathscr{M}(X_2)$, $F = \mathscr{C}_0(X_1 \times X_2)$, and $F' \cong \mathscr{M}(X_1 \times X_2)$, define the convex functions φ and ψ as

$$\begin{aligned} \varphi\colon \quad \mathscr{C}_0(X_1)\times\mathscr{C}_0(X_2) &\longrightarrow (-\infty,\infty], \\ (u,v) &\longmapsto -\langle\mu,u\rangle - \langle\nu,v\rangle, \end{aligned} \tag{2.9}$$

$$\begin{aligned} \psi\colon \quad \mathscr{C}_0(X_1\times X_2) &\longrightarrow (-\infty,\infty], \\ w &\longmapsto \chi[w \leqslant c \text{ a.e.}] = \begin{cases} 0 & \text{if } w \leqslant c \text{ a.e.} \\ \infty & \text{else} \end{cases}. \end{aligned} \tag{2.10}$$

Then, by Definition 2.9,

$$\begin{aligned} \varphi^*\colon \quad \mathscr{M}(X_1)\times\mathscr{M}(X_2) &\longrightarrow (-\infty,\infty], \\ (\alpha,\beta) &\longmapsto \sup\{\langle\alpha,u\rangle + \langle\beta,v\rangle - \varphi(u,v) \mid (u,v) \in \mathscr{C}_0(X_1)\times\mathscr{C}_0(X_2)\} \\ &= \sup\{\langle\alpha+\mu,u\rangle + \langle\beta+\nu,v\rangle \mid (u,v) \in \mathscr{C}_0(X_1)\times\mathscr{C}_0(X_2)\} \\ &= \chi[\alpha = -\mu, \beta = -\nu], \end{aligned}$$

$$\begin{aligned} \psi^*\colon \quad \mathscr{M}(X_1\times X_2) &\longrightarrow (-\infty,\infty], \\ \gamma &\longmapsto \sup\{\langle\gamma,w\rangle - \chi[w \leqslant c \text{ a.e.}] \mid w \in \mathscr{C}_0(X_1\times X_2)\} \\ &= \sup\{\langle\gamma,w\rangle \mid w \in \mathscr{C}_0(X_1\times X_2),\ w \leqslant c \text{ a.e.}\} \\ &= \begin{cases} \int c\,\mathrm{d}\gamma & \text{if } \gamma \in \mathscr{M}_+(X_1\times X_2) \text{ and } c \in \mathrm{L}^1(\gamma), \\ \infty & \text{else} \end{cases}. \end{aligned}$$

The last equality is justified by the fact that, for $\gamma \in \mathscr{M}(X_1 \times X_2)$ and $c \in \mathrm{L}^1(\gamma)$, for any $\varepsilon > 0$, one can find $w_\varepsilon \in \mathscr{C}_0(X_1 \times X_2)$ and $w_\varepsilon \leqslant c$ a.e., such that $\langle\gamma, w_\varepsilon\rangle \geqslant \int c\,\mathrm{d}\gamma - \varepsilon$ since $|\gamma|(X_1 \times X_2) < \infty$.

Recall that the operator $A \in \mathscr{L}(E, F)$ (2.4) and its adjoint A^* (2.6) are defined as

$$\begin{aligned} A\colon \quad \mathscr{C}_0(X_1) \times \mathscr{C}_0(X_2) &\longrightarrow \mathscr{C}_0(X_1 \times X_2), \\ (u, v) &\longmapsto u \circ \mathrm{p}_1 + v \circ \mathrm{p}_2, \\ A^*\colon \quad \mathscr{M}(X_1 \times X_2) &\longrightarrow \mathscr{M}(X_1) \times \mathscr{M}(X_2), \\ \gamma &\longmapsto (\mathrm{p}_{1\#}\gamma, \mathrm{p}_{2\#}\gamma). \end{aligned}$$

Define the two (convex) sets

$$\Gamma(\mu, \nu) := \big\{\gamma \in \mathscr{M}_+(X_1 \times X_2) \mid \mathrm{p}_{1\#}\gamma = \mu,\ \mathrm{p}_{2\#}\gamma = \nu\big\}, \tag{2.11}$$

$$\mathcal{U}_{\mathscr{C}_0}(c) := \{(u, v) \in \mathscr{C}_0(X_1) \times \mathscr{C}_0(X_2) \mid u \circ \mathrm{p}_1 + v \circ \mathrm{p}_2 \leqslant c \text{ a.e.}\}. \tag{2.12}$$

Then, one computes, for all $\gamma \in \mathscr{M}(X_1 \times X_2)$ and all $(u, v) \in \mathscr{C}_0(X_1) \times \mathscr{C}_0(X_2)$,

$$\begin{aligned} \varphi^*(-A^*\gamma) + \psi^*(\gamma) &= \chi\big[-\mathrm{p}_{1\#}\gamma = -\mu, -\mathrm{p}_{2\#}\gamma = -\nu\big] + \int c \,\mathrm{d}\gamma \\ &\quad + \chi[\gamma \in \mathscr{M}_+(X_1 \times X_2)] + \chi\big[c \in \mathrm{L}^1(\gamma)\big] \\ &= \int c \,\mathrm{d}\gamma + \chi[\gamma \in \Gamma(\mu, \nu)], \end{aligned} \tag{2.13}$$

and

$$\begin{aligned} -\varphi(u, v) - \psi(A(u, v)) &= \langle \mu, u \rangle + \langle \nu, v \rangle - \chi[u \circ \mathrm{p}_1 + v \circ \mathrm{p}_2 \leqslant c \text{ a.e.}] \\ &= \langle \mu, u \rangle + \langle \nu, v \rangle - \chi[(u, v) \in \mathcal{U}_{\mathscr{C}_0}(c)]. \end{aligned}$$

With a slight abuse of notation, we write in (2.13) the quantity $\int c \,\mathrm{d}\gamma$ even when $c \notin \mathrm{L}^1(\gamma)$, in which case the integral conventionally evaluates to ∞.

We claim now that there exists $x_0 \in E$ such that φ is finite at x_0 and ψ is finite and continuous at Ax_0.

Indeed, φ (2.9) is always finite. In the case of ψ (2.10), it is finite and continuous on $\operatorname{int} \operatorname{dom} \psi = \{w \in \mathscr{C}_0(X_1 \times X_2) \mid w < c\}$. Since $c \geqslant 0$, ψ will be necessarily continuous and finite on $\{w \in \mathscr{C}_0(X_1 \times X_2) \mid w < 0\}$ for any c. By taking $x_0 = (u_0, v_0) \in \mathscr{C}_0(X_1, (-\infty, 0)) \times \mathscr{C}_0(X_2, (-\infty, 0))$ a couple of negative continuous functions that vanish at infinity, $x_0 \in E$ and $w_0 = Ax_0 = u_0 \circ \mathrm{p}_1 + v_0 \circ \mathrm{p}_2 < 0 \leqslant c$, so that ψ is finite and continuous at w_0.

Therefore, we can invoke Theorem 2.27 and obtain

$$\min\left\{\int c \,\mathrm{d}\gamma \mid \gamma \in \Gamma(\mu, \nu)\right\} = \sup\{\langle \mu, u \rangle + \langle \nu, v \rangle \mid (u, v) \in \mathcal{U}_{\mathscr{C}_0}(c)\}, \tag{2.14}$$

This duality will be central in the derivation of the next section.

Remark 2.29. In light of the various definitions, it is the left-hand side of (2.14) that should be named the dual problem, and the right-hand size the primal one. Nonetheless, the opposite nomenclature is often found in the literature. We shall therefore refer to the left-hand side of (2.14) as the *primal* problem, and to the right-hand side as the *dual* problem.

$\Diamond$

We have now the core mathematical tools required to present the optimal transport derivations.

2.2 Optimal transport primer

In this section, we expose the different key concepts of optimal transport that will be relevant throughout the thesis. Optimal transport has a long history, that is usually traced back to Gaspard Monge's formulation [Mon81]. Many recent works have exposed and explored optimal transport theory, see e.g. [Vil03; AGS05; Vil09; San15]. We will provide here a condensed, contemporary formulation of the problem, through the lens of convex analysis and duality concepts introduced in Section 2.1.

2.2.1 General setting

In its general and historical form, optimal transport aims at finding a way to "move" (or *transport*) a given source to a target distribution, with minimal cumulated cost. The (ground) cost at play is defined on the product space where the distributions are defined, and optimal transport leverages it to quantitatively compare the distributions themselves. The presentation that follows is stated in modern probabilistic parlance. It focuses on the main features of optimal transport that will be leveraged in the following chapters.

Let $(X, \|\cdot\|)$ be any complete, separable metric space. Such a space is sometimes called a *Polish* space.

Recall that, given a Polish space X, $\mathscr{M}(X)$ (resp. $\mathscr{M}_+(X)$) denotes the set of finite signed (resp. positive) measures on X, as defined by Definition 2.18.

Denote by $\langle \cdot, \cdot \rangle$ the duality pairing between $\mathscr{C}_0(X)$ and $\mathscr{M}(X)$, as per Theorem 2.21:

$$\forall (\mu, f) \in \mathscr{M}(X) \times \mathscr{C}_0(X), \quad \langle \mu, f \rangle \equiv \int f \, \mathrm{d}\mu.$$

For X a Polish space, denote $\mathscr{P}(X) = \{\mu \in \mathscr{M}_+(X) \mid \mu(X) = 1\}$ the set of probability measures on X.

The transport has to be performed between two sets. Let therefore X_1 and X_2 be two Polish spaces. Let $\mu \in \mathscr{P}(X_1)$ and $\nu \in \mathscr{P}(X_2)$ be two probability measures on these spaces. The distributions μ and ν are respectively considered the source and the target distributions.

Further, let $c\colon X_1 \times X_2 \to [0, \infty)$ be a proper, lower semicontinuous cost function. The optimal transport problem aims at solving the following constrained minimization problem

$$\text{minimize} \left\{ \int c \, \mathrm{d}\gamma \mid \gamma \in \Gamma(\mu, \nu) \right\}, \tag{2.15}$$

where $\Gamma(\mu, \nu)$ is the set of feasible *plans* (or *couplings*) between μ and ν, i.e.

$$\begin{aligned} \Gamma(\mu, \nu) &= \{\gamma \in \mathscr{M}_+(X_1 \times X_2) \mid \text{for all measurable sets } A \subset X_1 \text{ and } B \subset X_2, \\ &\qquad \gamma(A \times X_1) = \mu(A) \text{ and } \gamma(X_2 \times B) = \nu(B)\}, \\ &= \{\gamma \in \mathscr{M}_+(X_1 \times X_2) \mid \mathrm{p}_{1\#}\gamma = \mu, \, \mathrm{p}_{2\#}\gamma = \nu\}. \end{aligned}$$

Recall that, for $i \in \{1, 2\}$, p_i is the projection onto X_i (Equation (2.5)), and $\cdot_\#$ is the push-forward operation (Definition 2.26).

The set $\Gamma(\mu, \nu)$ being compact, one can show that (2.15) always admits a minimizer (which was also a result of Theorem 2.27).

Define, for $\gamma \in \mathscr{P}(X_1 \times X_2)$ and for $(u, v) \in \mathrm{L}^1(\mu) \times \mathrm{L}^1(\nu)$, the functionals I and J as

$$I(\gamma) = \int c \, \mathrm{d}\gamma, \tag{2.16}$$

$$J(u, v) = \int u \, \mathrm{d}\mu + \int v \, \mathrm{d}\nu. \tag{2.17}$$

They evaluate a transport plan (I) or a pair of dual variables (J).

Denote further the set of feasible dual variables

$$\begin{aligned} \mathcal{U}_{\mathrm{L}^1}(c) = &\{(u, v) \in \mathrm{L}^1(\mu) \times \mathrm{L}^1(\nu) \mid u(x_1) + v(x_2) \leqslant c(x_1, x_2) \\ &\qquad \text{for } \mu\text{-almost all } x_1 \in X_1 \text{ and } \nu\text{-almost all } x_2 \in X_2\}. \end{aligned} \tag{2.18}$$

Lemma 2.30 (Weak duality, [Vil03, Proposition 1.5]). *One always has*

$$\sup \{J(u, v) \mid (u, v) \in \mathcal{U}_{\mathrm{L}^1}(c)\} \leqslant \inf \{I(\gamma) \mid \gamma \in \Gamma(\mu, \nu)\}. \tag{2.19}$$

Proof. Indeed, let $(u, v) \in \mathcal{U}_{\mathrm{L}^1}(c)$ and $\gamma \in \Gamma(\mu, \nu)$. Then,

$$J(u, v) = \int u \, \mathrm{d}\mu + \int v \, \mathrm{d}\nu = \int (u \circ \mathrm{p}_1 + v \circ \mathrm{p}_2) \, \mathrm{d}\gamma.$$

Since $(u, v) \in \mathcal{U}_{\mathrm{L}^1}(c)$ and $\gamma \in \Gamma(\mu, \nu)$, $u \circ \mathrm{p}_1 + v \circ \mathrm{p}_2 \leqslant c$ γ-a.e., then

$$J(u, v) \leqslant \int c \, \mathrm{d}\gamma = I(\gamma).$$

This is true for all $(u, v) \in \mathcal{U}_{\mathrm{L}^1}(c)$ and all $\gamma \in \Gamma(\mu, \nu)$. Taking the supremum over $\mathcal{U}_{\mathrm{L}^1}(c)$ on the left-hand side and the infimum over $\Gamma(\mu, \nu)$ on the right-hand side yields the desired inequality (2.19). □

Now, we connect the analysis to the running example in Sections 2.1.2 and 2.1.3 that shows how the inequality becomes an equality, and how the class of functions can be reduced. Therefore, as in Example 2.3, let

$$\mathcal{U}_{\mathscr{C}_0}(c) = \{(u, v) \in \mathscr{C}_0(X_1) \times \mathscr{C}_0(X_2) \mid u \circ \mathrm{p}_1 + v \circ \mathrm{p}_2 \leqslant c\}. \tag{2.20}$$

Example 2.3 shows that

$$\min \{I(\gamma) \mid \gamma \in \Gamma(\mu, \nu)\} = \sup \{J(u, v) \mid (u, v) \in \mathcal{U}_{\mathscr{C}_0}(c)\} \tag{2.21}$$

Remark 2.31. Under the condition that c is real-valued and pointwise upper-bounded by integrable functions, [Vil09, Theorem 5.10.(*iii*)] asserts that the right-hand size of (2.21) also attains its bound.

Since $\mathcal{U}_{\mathscr{C}_0}(c) \subset \mathcal{U}_{\mathrm{L}^1}(c)$, one always has

$$\sup \{J(u, v) \mid (u, v) \in \mathcal{U}_{\mathscr{C}_0}(c)\} \leqslant \sup \{J(u, v) \mid (u, v) \in \mathcal{U}_{\mathrm{L}^1}(c)\}.$$

Therefore, by (2.21) and (2.19), one has

$$\sup \{J(u, v) \mid (u, v) \in \mathcal{U}_{\mathscr{C}_0}(c)\} = \sup \{J(u, v) \mid (u, v) \in \mathcal{U}_{\mathrm{L}^1}(c)\}, \tag{2.22}$$

and one can interchangeably consider the sets $\mathcal{U}_{\mathscr{C}_0}(c)$ or $\mathcal{U}_{\mathrm{L}^1}(c)$. We shall loosely use the notation $\mathcal{U}(c)$ to refer to one or the other set, when no distinction has to be made.

c-transform

As a last remark for the general case, we recall (see e.g. [Vil09]) that the problem

$$\text{maximize} \left\{ \int u \, \mathrm{d}\mu + \int v \, \mathrm{d}\nu \;\middle|\; (u, v) \in \mathcal{U}(c) \right\}$$

is equivalently solved with the *c-transform* of the variable u, either in $\mathscr{C}_0(X_1)$ or in $\mathrm{L}^1(\mu)$, defined as

$$\begin{array}{rccl} u^c \colon & X_2 & \longrightarrow & \mathbb{R}, \\ & x_2 & \longmapsto & u^c(x_2) = \inf \{c(x_1, x_2) - u(x_1) \mid x_1 \in X_1\}, \end{array}$$

in the sense that

$$\sup\{J(u,v) \mid (u,v) \in \mathcal{U}(c)\} = \sup\left\{\int u\,\mathrm{d}\mu + \int u^c\,\mathrm{d}\nu \;\middle|\; u \in \mathscr{C}_0(X_1)\right\} \tag{2.23a}$$

$$= \sup\left\{\int u\,\mathrm{d}\mu + \int u^c\,\mathrm{d}\nu \;\middle|\; u \in \mathrm{L}^1(\mu)\right\}. \tag{2.23b}$$

In addition, the supremum is attained on the left-hand side if and only if it is attained on the right-hand side.

This is useful in order to reduce the number of variables to one, instead of two functions. It simplifies the problem when the c-transform of u has a straightforward expression.

When $X_1 = X_2 \equiv X$ and, $\forall (x,x') \in X^2$, $c(x,x') = \|x - x'\|^p$ for some norm $\|\cdot\|\colon X \to [0,\infty)$ and some $p \in [1,\infty)$, we shall denote, for any $u\colon X \to \mathbb{R}$, the *p-transform* of u as

$$\begin{array}{rccl} u^{(p)}\colon & X & \longrightarrow & \mathbb{R}, \\ & x' & \longmapsto & \inf\{\|x' - x\|^p - u(x) \mid x \in X\}. \end{array} \tag{2.24}$$

This concludes the general case presentation. Some special cases are worthy of a more detailed attention.

2.2.2 Special cases

In this section, we present some specializations of the general problem (2.21).

$\boldsymbol{X_1}$ or $\boldsymbol{X_2}$ is compact

When, for some $i \in \{1,2\}$, the set X_i is compact, then $\mathscr{C}_0(X_i) = \mathscr{C}(X_i)$. This has the following practical application. Say X_1 is compact. The supremum in (2.23a) is then taken over all $u \in \mathscr{C}(X_1)$. One can use a neural network to parametrize functions approximating such a continuous function.

$\boldsymbol{X_1 = X_2}$ and $\boldsymbol{c = d^p}$

The case where $X_1 = X_2 = X$ and c is the p-power, for $p \in [1,\infty)$ of any distance d on X — i.e. $\forall (x,x') \in X^2$, $c(x,x') = d(x,x')^p$ — is of particular interest. It is the topic of, e.g., [Vil09, Chapter 6], from which we pick the most salient features.

Let

$$\mathscr{P}_p(X) := \left\{\mu \in \mathscr{P}(X) \;\middle|\; \int d(x,x_0)^p\,\mathrm{d}\mu(x) < \infty\right\}$$

be the space of probability distributions with finite p-moment for some, hence any, $x_0 \in X$.

Let $(\mu, \nu) \in \mathscr{P}_p(X)^2$ be two probability distributions with finite p-moment. Under this setting, (2.15) takes the form

$$\text{minimize} \left\{ \int d(x,x')^p \, \mathrm{d}\gamma(x,x') \;\middle|\; \gamma \in \Gamma(\mu,\nu) \right\}. \tag{2.25}$$

By setting

$$\mathcal{W}_p(\mu,\nu) := \left[\inf \left\{ \int d(x,x')^p \, \mathrm{d}\gamma(x,x') \;\middle|\; \gamma \in \Gamma(\mu,\nu) \right\} \right]^{1/p}, \tag{2.26}$$

it defines a distance (called *p-Wasserstein distance*) on $\mathscr{P}_p(X)$. What's more, it metrizes the weak convergence in $\mathscr{P}_p(X)$ [Vil09, Theorem 6.9].

In the particular case where d is distance derived from the norm $\|\cdot\|$ on X (i.e. $\forall (x,x') \in X^2$, $c(x,x') = \|x - x'\|^p$), we will denote the set of admissible function variables $\mathcal{U}_{\mathrm{L}^1}(p)$ and $\mathcal{U}_{\mathscr{C}_0}(p)$.

The next two special cases will be an Euclidean specialization of this settings, with $X_1 = X_2 = \mathbb{R}^n \equiv X, \forall (x,x') \in X^2$, $d(x,x') = \|x - x'\|$ and $p = 1$ and $p = 2$, respectively.

$\boldsymbol{X_1 = X_2 = \mathbb{R}^n}$ and $\boldsymbol{c(x,x') = \|x - x'\|}$

In the case where $X_1 = X_2 = \mathbb{R}^n \equiv X$ and $\forall (x,x') \in X^2$, $c(x,x') = \|x - x'\|$, the Wasserstein distance enjoys a useful characterization. Indeed, in this case, u is c-concave (i.e. there exists a function v such that $u = v^{(1)}$ as defined by (2.24)) if and only if $\mathrm{Lip}(u) \leqslant 1$, where, for $u \colon X \to (-\infty, \infty]$,

$$\mathrm{Lip}(u) = \sup \left\{ \frac{|u(x') - u(x)|}{\|x' - x\|} \;\middle|\; (x,x') \in X^2,\, x' \neq x \right\} \in [0, \infty]. \tag{2.27}$$

In addition, if $\mathrm{Lip}(u) \leqslant 1$, then $u^{(1)} = -u$ [PC18, Proposition 6.1].

Define

$$\mathrm{Lip}_1(\mathbb{R}^n) := \{ u \in \mathscr{C}(\mathbb{R}^n) \mid \mathrm{Lip}(u) \leqslant 1 \}. \tag{2.28}$$

The set of function variables can then be reduced to $\mathscr{C}_0(\mathbb{R}^n) \cap \mathrm{Lip}_1(\mathbb{R}^n)$, so that

$$\mathcal{W}_1(\mu,\nu) = \sup \{ \langle \mu, u \rangle - \langle \nu, u \rangle \mid u \in \mathscr{C}_0(\mathbb{R}^n) \cap \mathrm{Lip}_1(\mathbb{R}^n) \}. \tag{2.29}$$

Equation (2.29) is called the Kantorovich-Rubinstein formula.

$$\boldsymbol{X_1 = X_2 = \mathbb{R}^n} \text{ and } \boldsymbol{c(x, x') = \|x - x'\|^2}$$

As presented above, with $(X, \|\cdot\|)$ a metric space, the 2-Wasserstein distance is a distance on the space of quadratically integrable probability measures $\mathscr{P}_2(X) := \{\mu \in \mathscr{P}(X) \mid \int \|x\|^2 \,\mathrm{d}\mu(x) < \infty\}$, or equivalently, the set of probability measures on X with finite variance. The distance $\mathcal{W}_2$ is sometimes also called the *Fréchet distance* between two distributions.

It enjoys particular geometrical properties. Notably, it solves a dynamical displacement problem between two distributions, under a continuity equation constraint [Vil03, Thereom 8.1]. This is known as the *Benamou-Brenier* formulation.

For $n \in \mathbb{N}$, let $\mathbb{S}_+(n)$ be the set of symmetric positive semi-definite (or covariance) matrices of size $n \times n$. For $i \in \{1, 2\}$, let $(\boldsymbol{m}_i, \Sigma_i) \in \mathbb{R}^n \times \mathbb{S}_+(n)$ be mean vectors and covariance matrices. In the case where $\mu = \mathcal{N}(\boldsymbol{m}_1, \Sigma_1)$ and $\nu = \mathcal{N}(\boldsymbol{m}_2, \Sigma_2)$ are two Gaussian distributions, the 2-Wasserstein distance between μ and ν has the following closed-form expression [DL82].

Proposition 2.32 *For $i \in \{1, 2\}$, let $(\boldsymbol{m}_i, \Sigma_i) \in \mathbb{R}^n \times \mathbb{S}_+(n)$. If $\mu = \mathcal{N}(\boldsymbol{m}_1, \Sigma_1)$ and $\nu = \mathcal{N}(\boldsymbol{m}_2, \Sigma_2)$, then*

$$\mathcal{W}_2^2(\mu, \nu) = \|\boldsymbol{m}_1 - \boldsymbol{m}_2\|^2 + \operatorname{tr}\left(\Sigma_1 + \Sigma_2 - 2\left[\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right]^{1/2}\right). \tag{2.30}$$

The trace term in (2.30) is called the (squared) *Bures-Wasserstein distance* between the covariance matrices Σ_1 and Σ_2.

Definition 2.33 (*Bures-Wasserstein distance on $\mathbb{S}_+(n)$*). Let $\Sigma_1 \in \mathbb{S}_+(n)$ and $\Sigma_2 \in \mathbb{S}_+(n)$ be two symmetric positive semi-definite matrices. The Bures-Wasserstein distance between Σ_1 and Σ_2 is

$$\mathcal{B}^2(\Sigma_1, \Sigma_2) = \operatorname{tr}\left(\Sigma_1 + \Sigma_2 - 2\left[\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right]^{1/2}\right). \tag{2.31}$$

It is indeed a distance on $\mathbb{S}_+(n)$. We refer to [BJL19] for more details on this metric. It will be central in Chapters 5 and 4.

Lastly, as we have seen previously, the squared 2-Wasserstein distance between two distributions $\mu \in \mathscr{P}_2(\mathbb{R}^n)$ and $\nu \in \mathscr{P}_2(\mathbb{R}^n)$ has the following expression:

$$\begin{aligned}
\mathcal{W}_2^2(\mu, \nu) &= \sup\{\langle \mu, u\rangle + \langle \nu, v\rangle \mid (u, v) \in \mathcal{U}_{\mathscr{C}_0}(2)\} \\
&= \sup\left\{\langle \mu, u\rangle + \langle \nu, u^{(2)}\rangle \mid u \in \mathscr{C}_0(X_1)\right\} \\
&= \sup\left\{\int u \,\mathrm{d}\mu + \int v \,\mathrm{d}\nu \mid (u, v) \in \mathcal{U}_{\mathrm{L}^1}(2)\right\} \\
&= \sup\left\{\int u \,\mathrm{d}\mu + \int u^{(2)} \,\mathrm{d}\nu \mid u \in \mathrm{L}^1(\mu)\right\}
\end{aligned} \tag{2.32}$$

where

$$\mathcal{U}_{\mathscr{C}_0}(2) = \Big\{(u,v) \in \mathscr{C}_0(\mathbb{R}^n) \times \mathscr{C}_0(\mathbb{R}^n) \mid \forall (x,x') \in (\mathbb{R}^n)^2,\, u(x) + v(x') \leqslant \|x - x'\|^2\Big\},$$
$$\mathcal{U}_{\mathrm{L}^1}(2) = \big\{(u,v) \in \mathrm{L}^1(\mu) \times \mathrm{L}^1(\nu) \mid \text{for } \mu\text{-almost all } x \text{ and } \nu\text{-almost all } x',$$
$$u(x) + v(x') \leqslant \|x - x'\|^2\big\},$$

and $u^{(2)}$ is defined by (2.24). Therefore, the dual variables u and v are required to be either in $\mathcal{U}_{\mathscr{C}_0}(2)$ or $\mathcal{U}_{\mathrm{L}^1}(2)$. This will be specifically useful when finding a lower-bound for $\mathcal{W}_2(\mu,\nu)$, as done in Lemma 5.9.

3 Optimization of general Generative Adversarial Networks

In this chapter we introduce the general Generative Adversarial Network (GAN) framework and highlight the key challenges that this learning scheme entails. We first give a review of the different GAN formulations as they appeared in the literature and give insight on which problems they aimed to solve. This includes two main origins for the GAN formulation, a probabilistic one (Section 3.1.1) and an optimal transportation one (Section 3.1.2). Then, the general optimization scheme used in order to train such a GAN is introduced in Section 3.2, before presenting a simplified analysis framework and the optimal discriminator hypothesis in Section 3.3. We finish this chapter by studying a simple example in Section 3.4 to illustrate the key optimization challenges faced by GANs. This will allow to portray the analytical tools exposed throughout this chapter.

3.1 Introduction to the GAN framework

Generally speaking, GANs [GPM+14] are neural networks aiming at producing samples as if they were drawn from a target distribution. They are optimized in an adversarial fashion, were a *discriminator* and a *generator* networks work jointly in achieving learning. Due to this joint optimization, the stability and convergence properties of such learning systems is more delicate to analyze compared to a minimization problem only, as usually found in supervised learning for instance. We will present the general GAN optimization paradigm and its difficulties.

In this section, $\Theta \subseteq \mathbb{R}^p$ and $\Psi \subseteq \mathbb{R}^q$ are the parameter spaces of the generator and the discriminator, $Z \subseteq \mathbb{R}^{d_z}$ is a latent space, and $X \subseteq \mathbb{R}^{d_x}$ is the sample space. The *generator* g is a parametric function with parameter $\theta \in \Theta$ which takes points z in the latent space to points $x = g_\theta(z)$ in the sample space:

$$\begin{array}{rccl} g\colon & Z \times \Theta & \longrightarrow & X, \\ & (z, \theta) & \longmapsto & g(z;\theta) \equiv g_\theta(z), \end{array} \tag{3.1}$$

Further, the *discriminator* f is also a parametric function with parameter $\psi \in \Psi$, which takes points x in the sample space and map them to real numbers:

$$\begin{array}{rccl} f\colon & X \times \Psi & \longrightarrow & \mathbb{R}, \\ & (x, \psi) & \longmapsto & f(x;\psi) \equiv f_\psi(x). \end{array} \tag{3.2}$$

Assume a *target distribution* $\nu \in \mathscr{P}(X)$ is given. A GAN has to learn how to generate samples as if they were drawn from ν.

Given $N \in \mathbb{N}$ target samples $\{x_i\}_{i\in[N]}$, one can write

$$\hat{\nu} := \frac{1}{N} \sum_{i\in[N]} \delta_{x_i} \tag{3.3}$$

the discrete, target distribution formed by the target samples, where $\delta_x \in \mathscr{P}(X)$ is the Dirac measure at $x \in X$.

With $\omega \in \mathscr{P}(Z)$ a probability distribution on the latent space, given a parameter $\theta \in \Theta$, the generator defines a distribution μ_θ that can be written with the push-forward operator as

$$\mu_\theta = (g_\theta)_{\#}\omega. \tag{3.4}$$

It simply means that, to sample from μ_θ, one samples ω and applies g_θ to the resulting latent samples in Z in order to produce samples in X.

Even when ω has a density (i.e. is absolutely continuous with respect to the Lebesgue measure), the push-forward measure μ_θ might not be continuous. For instance, if the dimension of X is higher than its of Z, the manifold of generated images will have Lebesgue measure 0 in X. This has been linked to instabilities in the training of GANs [AB17], as in this case, a given statistical divergence between the two distributions usually has singularities.

In practice, one samples a minibatch from ω and evaluates the generator at the latent codes, so that the effective dataset on which the discriminator is trained consists of a mix of samples from the target distribution and samples from the generator. This will be particularly relevant in Section 3.1.1. When distinction has to be made between ω and its sampled version, we shall write $\hat{\omega}$ for the sampled distribution, and likewise $\hat{\mu}_\theta$ for the distribution sampled from μ_θ.

Some works, such as [FG15] for the rate of convergence of a Wasserstein estimator, or [FZF+20] for the effect of a minibatch approach in estimating the Wasserstein distance, study the approximation error suffered by sampling a population in order to compute a Wasserstein-based quantity. Nonetheless, the general effect of using minibatch to approximate objectives as expected values will not be part of our concerns. Therefore, we will not usually make a distinction between ν and $\hat{\nu}$, assuming that we always sample the target distribution, nor between μ_θ and $\hat{\mu}_\theta$. Either way, ω, μ_θ, and ν might be continuous or discrete, depending on the context.

With this broad setup in mind, we now present the two motivations at the core of many GAN systems: a probabilistic (Section 3.1.1) and an optimal transport (Section 3.1.2) one. They share the objective to compare probabilistic distributions on a given sample space, but differ in the way they tackle this issue and in their interpretations. We begin by presenting the (historically first) probabilistic motivation for GANs.

3.1.1 Probabilistic motivation

Originally, the discriminator and the generator would optimize an objective which stemmed from a probabilistic perspective as a cross-entropy loss between *generated* (or "fake") and *target* (or "real") samples. Indeed, the original GAN loss is derived from a binary classification problem: the discriminator is then typically trained to encode the probability of a given sample to be real.

Let $\theta \in \Theta$ be a generator parameter, and let $\mathcal{D}(\theta) = \{x_i\}_{i\in[N]}$ be a collection of samples in X, some coming from a target dataset and some generated by a generator with parameter θ.

Let $\mathcal{D}_\mathrm{t} = \{x \in \mathcal{D} \mid x \text{ is a target}\}$ be the target samples and let $\mathcal{D}_\mathrm{g}(\theta) = \{x \in \mathcal{D} \mid x \text{ is generated by } \theta\}$ be the generated samples, so that $\mathcal{D}(\theta) = \mathcal{D}_\mathrm{t} \cup \mathcal{D}_\mathrm{g}(\theta)$. The dependency of $\mathcal{D}$ or $\mathcal{D}_\mathrm{g}$ on θ will be implicit when possible to lighten the notation.

Let $I_\mathrm{t} := \{i \in [N] \mid x_i \in \mathcal{D}_\mathrm{t}\}$ and $I_\mathrm{g} := \{i \in [N] \mid x_i \in \mathcal{D}_\mathrm{g}(\theta)\}$ be the index sets associated with the real and generated samples, and let $N_\mathrm{t} := |\mathcal{D}_\mathrm{t}|$ and $N_\mathrm{g} := |\mathcal{D}_\mathrm{g}|$ be the respective number of target and generated samples.

Let g be a generator defined as in (3.1). There exist by definition N_g fixed latent codes $\{z_j \in Z \mid j \in [N_\mathrm{g}]\}$ and a parameter $\theta \in \Theta$ such that

$$\mathcal{D}_\mathrm{g}(\theta) = \{g_\theta(z_j) \mid j \in [N_\mathrm{g}]\}.$$

Let then $\hat{\omega} := \dfrac{1}{N_\mathrm{g}} \displaystyle\sum_{j\in[N_\mathrm{g}]} \delta_{z_j}$ be the fixed latent distribution and let

$$\hat{\mu}_\theta := \frac{1}{N_\mathrm{g}} \sum_{j\in[N_\mathrm{g}]} \delta_{g_\theta(z_j)} \tag{3.5}$$

be the distribution of $\mathcal{D}_\mathrm{g}(\theta)$. Note that we still have $\hat{\mu}_\theta = (g_\theta)_{\#}\omega$, as presented in Section 3.1.

Let $\hat{\nu} = N_\mathrm{t}^{-1} \sum_{i\in I_\mathrm{t}} \delta_{x_i}$ be the uniform distribution on the target samples in $\mathcal{D}_\mathrm{t}$.

A discriminator f defined as in (3.2) can parametrize the probability that a point $x \in X$ is from the target distribution. Usually, we do not want to constrain the values of f and thus, to obtain valid probabilities bound between 0 and 1, we use a sigmoid function: $\sigma\colon \mathbb{R} \to (0,1);\ t \mapsto \sigma(t) := (1 + e^{-t})^{-1}$, so that we can set, for $x \in \mathcal{D}$,

$$\mathbb{P}(x \in \mathcal{D}_\mathrm{t} \mid \psi) = \sigma(f_\psi(x)).$$

The likelihood to observe an independent and identically distributed (i.i.d.) dataset $\mathcal{D}$ composed of N_t target and N_g generated samples under this model is then

$$\mathbb{P}(\mathcal{D} \mid \psi) = \prod_{i\in I_\mathrm{t}} \sigma(f_\psi(x_i)) \prod_{i\in I_\mathrm{g}} (1 - \sigma(f_\psi(x_i))) \tag{3.6}$$

Hence, given $(\theta, \psi) \in \Theta \times \Psi$, the log-likelihood of the dataset $\mathcal{D}(\theta)$ under the model f_ψ is

$$\begin{aligned}
\mathrm{LL}(\theta, \psi) &= \log \mathbb{P}(\mathcal{D}(\theta) \,|\, \psi) \\
&= \sum_{i \in I_\mathrm{t}} \log \sigma(f_\psi(x_i)) + \sum_{i \in I_\mathrm{g}} \log(1 - \sigma(f_\psi(x_i))) \\
&= N_\mathrm{t} \int \log \sigma(f_\psi(x)) \,\mathrm{d}\hat{\nu}(x) + N_\mathrm{g} \int \log(1 - \sigma(f_\psi(x))) \,\mathrm{d}\hat{\mu}_\theta(x) \\
&= N_\mathrm{t} \int \log \sigma(f_\psi(x)) \,\mathrm{d}\hat{\nu}(x) + N_\mathrm{g} \int \log(1 - \sigma(f_\psi(g_\theta(z)))) \,\mathrm{d}\hat{\omega}(z) \\
&= -N_\mathrm{t} \int \log(1 + \exp(-f_\psi(x))) \,\mathrm{d}\hat{\nu}(x) - N_\mathrm{g} \int \log(1 + \exp(f_\psi(g_\theta(z)))) \,\mathrm{d}\hat{\omega}(z).
\end{aligned}$$

Usually, N_g is taken to equal N_t, so that the problem is balanced.

Hence, maximizing the log-likelihood $\mathrm{LL}(\theta, \psi)$ of the i.i.d. dataset $\mathcal{D}$ given $\theta \in \Theta$ is written

$$\begin{aligned}
\mathrm{ML}(\theta) &= \text{maximize} \,\{\mathrm{LL}(\theta, \psi) \mid \psi \in \Psi\} \\
&= \text{maximize} \left\{ - \int \log(1 + \exp(-f_\psi(x))) \,\mathrm{d}\hat{\nu}(x) \right. \\
&\qquad\qquad \left. - \int \log(1 + \exp(f_\psi(g_\theta(z)))) \,\mathrm{d}\hat{\omega}(z) \;\middle|\; \psi \in \Psi \right\}.
\end{aligned} \tag{3.7}$$

This problem on $\psi \in \Psi$ (sometimes called the *inner problem*), is what the discriminator is dedicated to solve. Then, the generator parameters $\theta \in \Theta$ are optimized so as to *minimize* $\mathrm{ML}(\theta)$ (then called the *outer problem*):

$$\text{minimize} \,\{\mathrm{ML}(\theta) \mid \theta \in \Theta\}. \tag{3.8}$$

This update rule can be shown to be theoretically equivalent to minimizing a Jensen-Shannon (JS) divergence between the generated and the target distributions, under the assumption that the discriminator solves the problem (3.7) perfectly [GPM+14]. This is what is called the *perfect* (or *optimal*) discriminator hypothesis, and will be the main topic of Section 3.3.

In the seminal work on GANs [GPM+14], the generator and discriminator networks optimize two different losses. Rather than minimizing $\mathbb{E}_{z \sim \hat{\omega}}[-\log(1 + \exp(f_{\psi^*}(g_\theta(z))))]$ in (3.7), where $\psi^* \in \operatorname{argmax} \{\mathrm{LL}(\theta, \psi) \mid \psi \in \Psi\}$, the generator is trained to maximize $\mathbb{E}_{z \sim \hat{\omega}}[-\log(1 + \exp(-f_{\psi^*}(g_\theta(z))))]$. This technique of training was introduced to prevent the generator loss from saturating, as it was observed that $\exp(f_{\psi^*}(g_\theta(z)))$ vanishes

early in training (when the discriminator can be almost certain of which samples are generated). This will be referred to as the "non-saturating GAN", since it prevents the generator loss (and hence the gradients for g) from vanishing early in training.

The non-saturating loss is the standard update scheme when training GANs. The associated divergence actually minimized (under the optimality assumption) was expressed a few years later: under the same assumption of a perfect (optimal) discriminator, the non-saturating GAN loss amounts for the generator to minimizing a divergence mixing a JS and an inverted KL divergence between the two distributions [AB17, Theorem 2.5].

Note that, more generally, f-divergences can be considered when training GANs [NCT16], still under the perfect discriminator hypothesis. The properties of the divergences are more or less adapted to comparing the generated to the target distribution. For instance, the Jensen-Shannon divergence quickly saturates when the support of the distributions disagree, whereas the KL divergence is theoretically infinite when the supports of the distributions mismatch [AB17] — which is especially the case at the beginning of the training. The mechanics of the f-divergences offer a wide range of losses to pick from when training a GAN.

Other architectures or losses have been considered in order to train GANs. One worthy of mention is the conditional GAN (c-GAN) [MO14], where both the generator and the discriminator take an additional input (i.e. the class of the sample generated), and the generator also generates this input in addition to the sample. This type of GAN is not unsupervised anymore, as it requires access to labels for the samples. Nonetheless, it allows to condition the generation of the samples on a specific class or type of output one would like to obtain.

3.1.2 Optimal Transport motivation

As point out pointed out in [AB17], when the two distributions we are comparing either are discontinuous, or do not share their support, then the Jensen-Shannon divergence computed between the two maxes out. This leads to a theoretical issue in the training of GANs, as, in this case, a perfect discriminator (i.e. perfectly separating the generated and the target samples) will exist and have vanishing gradients on the samples.

To alleviate these issues, other means of comparing the generated distribution to the target one have been considered in order to train a GAN. One of them, stemming from optimal transport theory, will be the main topic of this section. The Wasserstein distance, as defined in Section 2.2.2, has been investigated as a means to compare the generated distribution μ_θ to the target distribution ν [ACB17; GAA+17; HRU+17; RST18].

This distance is in general costly to compute directly, but its dual expression can be leveraged to approximate it with a discriminator network. For instance, as seen in

Section 2.2.2, when $X = \mathbb{R}^n$, the dual expression of the 1-Wasserstein distance between two probability measures $\mu \in \mathscr{P}(X)$ and $\nu \in \mathscr{P}(X)$ is

$$\mathcal{W}_1(\mu, \nu) = \sup \left\{ \int f \, \mathrm{d}(\mu - \nu) \mid f \in \mathscr{C}_0(X) \cap \mathrm{Lip}_1(X) \right\}, \tag{3.9}$$

where $\mathscr{C}_0(X)$ are the continuous functions on X vanishing at infinity (Equation (2.1)) and $\mathrm{Lip}_1(X) = \{f \in \mathscr{C}(X) \mid \mathrm{Lip}(f) \leqslant 1\}$ (Equation (2.28)).

In the case where X is compact (e.g. $X = [0,1]^{h \times w \times c}$ is the space of images of dimensions $h \times w \times c$), then $\mathscr{C}_0(X) = \mathscr{C}(X)$ and therefore f can be represented by a neural network. The Lipschitz constraint has to be enforced explicitly. The correct and efficient way to do it is still an open question [SV18; LRC19; FRH+19; QMW20; KKM+21].

Assume now that X is compact. Let $\omega \in \mathscr{P}(Z)$ be a latent distribution, and let $g \colon Z \times \Theta \to X$ and $f \colon X \times \Psi \to \mathbb{R}$ be a generator and a discriminator networks, as introduced by Equations (3.1) and (3.2).

Denote $\mu_\theta = (g_\theta)_{\#}\omega$ the (continuous) generated distribution with parameter $\theta \in \Theta$ and $\nu = \frac{1}{N} \sum_{i \in [N]} \delta_{x_i}$ the (discrete) target distribution. Define further, for $(\theta, \psi) \in \Theta \times \Psi$,

$$\mathcal{L}_{\mathcal{W}_1}(\theta, \psi) := \int f_\psi(x) \, \mathrm{d}\mu_\theta(x) - \int f_\psi(y) \, \mathrm{d}\nu(y). \tag{3.10}$$

Denote $\Psi_{\mathrm{Lip}_1} := \{\psi \in \mathbb{R}^q \mid \mathrm{Lip}(f_\psi) \leqslant 1\}$. Then, due to the universal approximation theorem[Fun89], there exists an architecture for the discriminator so that we have, for $\theta \in \Theta$,

$$\sup \{\mathcal{L}_{\mathcal{W}_1}(\theta, \psi) \mid \psi \in \Psi_{\mathrm{Lip}_1}\} \approx \mathcal{W}_1(\mu_\theta, \nu),$$

in some sense. We keep the statement loose as we will not be concerned with the approximation error of the resulting system. It can be noted that, although originally motivated by the computation of a Wasserstein distance between the two distributions, the ability of a Wasserstein GAN to approximate the actual distance does not seem to corroborate with the generation quality of the model [MMG19; SEK+21].

The constraint $\psi \in \Psi_{\mathrm{Lip}_1}$ is difficult to enforce for neural networks. Several approaches exists as to successfully do it, from the weight clipping in [ACB17] to spectral normalization in [MKK+18].

A popular approach consists in regularizing the norm of the gradient of the network, as a function of the sample input. Indeed, for any differentiable functions $f \colon X \to \mathbb{R}$, the Lipschitz constraint is equivalent to

$$\mathrm{Lip}(f) \leqslant 1 \iff \forall x \in \mathrm{dom}(f), \, \|\nabla f(x)\|_2 \leqslant 1.$$

Remark 3.1. One can assume that f is only continuous with a *finite number of non-differentiable* points since, from the Rademacher's theorem, a Lipschitz function on an open of $\mathbb{R}^n$ is almost everywhere differentiable.

Therefore, regularizing the gradient's norm will constrain the Lipschitz constant of the function. This paradigm was first motivated in [GAA+17] and has known many variations since, especially on the way to sample the inputs in X on which the gradient has to be constrained.

Wasserstein GANs have received a great amount of attention in the hope and promise of actually computing a distance between μ_θ and ν, that will then serve as a loss for the generator. Nonetheless, some caution should be observed when considering this interpretation, as expressed in subsequent works.

In [SEK+21], the authors show experimental evidence that the 1-Wasserstein distance with pixelwise Euclidean ground cost does *not* correlate with our perception of distance between natural images. Notably, a batch of centroid images (in the case of a face dataset, CelebA) has a lower mean distance to real images (target) than a batch of real face images to other real faces, which mean that a perfect generator would suffer from a higher 1-Wasserstein distance relative to a generator that produces only centroids of the dataset. The authors' hypothesis for this undesired phenomenon is the use of the pixelwise Euclidean distance as the ground cost for the Wasserstein distance.

Possible remedies to this would be to use a Wasserstein ground metric as envisaged in [DLL+19b], more robust to perceptual modifications on the image, use a (pretrained) feature extraction to compute the ground cost, or try to learn the distance between images, based on an adversarial cost as in [SZR+18]. In their work, the authors achieve adversarially optimizing a network to learn a cost between images. The learned network effectively extracts features from the images into a lower-dimension space, and perform a cost between features there. All of these methods aim at using a more sensible ground cost than an Euclidean distance in the space of images, as it is obvious that images cannot properly be compared by a pixel-wise comparison.

3.1.3 General optimization framework for GANs

We now abstract the optimization problem and consider a general discriminator and generator losses, to focus on the optimization procedure itself. This will allow us to understand why, in general, the optimization problem of a GAN is involved to analyse.

As before, let Θ and Ψ denote the parameters spaces of respectively the generator and the discriminator networks, such that $\Theta \subseteq \mathbb{R}^p$ and $\Psi \subseteq \mathbb{R}^q$. Let $\mathcal{L}_\mathrm{g} : \Theta \times \Psi \to \mathbb{R}$ and $\mathcal{L}_\mathrm{d} : \Theta \times \Psi \to \mathbb{R}$ be objective functions for respectively the generator and the discriminator, considered as functions on the trainable parameters. Training a GAN therefore means generally finding parameters θ^* and ψ^* that is a *Nash equilibrium* — i.e., a solution where no party can improve their outcome without deteriorating the other's — for the system,

$$\begin{cases} \text{maximize } \{\mathcal{L}_\mathrm{d}(\theta, \psi) \mid \psi \in \Psi\} & \text{(3.11a)} \\ \text{minimize } \{\mathcal{L}_\mathrm{g}(\theta, \psi) \mid \theta \in \Theta\} & \text{(3.11b)} \end{cases}$$

In the case where $\mathcal{L}_{\mathrm{d}} = \mathcal{L}_{\mathrm{g}}$, one denotes simply $\mathcal{L} = \mathcal{L}_{\mathrm{d}} = \mathcal{L}_{\mathrm{g}}$, and System (3.11) can be written

$$\text{minimize} \left\{ \text{maximize} \left\{ \mathcal{L}(\theta, \psi) \mid \psi \in \Psi \right\} \mid \theta \in \Theta \right\} \tag{3.12}$$

This double optimization procedure is one of the reasons why a GAN is difficult to train, even in the simplest setting [MNG17; MGN18].

Two main approaches can be considered when analysing (3.12). First, the *inner problem* $\sup \{\mathcal{L}(\theta, \psi) \mid \psi \in \Psi\}$ can be assumed to be solved exactly at every step. This is known as the *optimal* (or *perfect*) *discriminator hypothesis*. It leads to a greatly simplified expression, and will be presented more in depth in Section 3.3. Second, the nested optimization in (3.12) can be studied directly in some simple cases, leading to a dynamical analysis of the system. This will be presented next, in Section 3.2.

3.2 Optimization procedure

The optimization problem solved by a GAN is by nature difficult. It relies on a *minimax* expression, which *i*) is ensured to have a unique solution only if the objective function is convex-concave — which is not the case for a typical GAN — and *ii*) requires to leverage two networks and gradient descent-ascent schemes. Several approaches have been followed in order to study such optimization configurations. One of the first simplification possible is to make the objective function convex-concave, and even bilinear in the parameters (e.g. $\mathcal{L}(\theta, \psi) = \langle \theta, \psi \rangle$). In this setting, the Jacobian of the problem can be easily computed, and classical analysis results can help classifying the stationary points into stable or unstable points [MNG17; MGN18]. Further, the optimization scheme can be adapted in order to yield the desired convergence properties. This is typically done by the addition of a regularizer to the original objective, as advised by [MGN18; CMF20], or by considering a different algorithm altogether to perform the parameter update. Even though the main algorithm to solve this problem is gradient descent-ascent (GDA), other techniques have been considered, such as optimistic mirror descent [DIS+18], variance reduction mechanism [YKH20], or with a proximal-GDA algorithm [CZX+20] for instance.

Lastly, the inner loop can be assumed to be solved perfectly at each iteration, leading to an overall minimisation problem. This will be one of the main feature in Chapter 5. Although this method has flaws (as experimental results tend to show that the inner problem is far from being solved by a GAN), this approach has the merit to considerably simplify the expression of the optimization and thus allows to conclude on specific convergence properties.

More recently, the random feature model in the context of WGAN was investigated in order to show convergence of the WGAN optimization procedure in the case of a

very large two-layer neural networks GAN: in [NYK+21], the authors study a WGAN with two-layer random feature model for both the generator and the discriminator. The number of weights in the generator if either infinite or finite, the number of weights for the discriminator is finite. Under mild assumption on the activation functions (i.e. sigmoidal) and the initialization of the discriminator, they find that there are no spurious critical points for the generator within a ball which can be made arbitrarily large as the number of parameters for the generator N_{g} grows to infinity. The regularization on the loss is the square norm of the discriminator parameters, and *ad hoc* learning rates make the stochastic gradient ascent-descent update on the mixed loss equivalent to a stochastic descent on the loss for the generator only. The property on the spurious critical points is then proven for the loss on the generator only. Classical tools used are universal approximation of functions (infinite and finite cases) and random feature learning.

Local convergence to equilibrium can be shown under some assumptions for the loss function $\mathcal{L}$ and the behaviour of the optimal parameter solutions θ^* and ψ^*. In [NK17], the authors study gradient flow evolution of the parameters under a general loss

$$\forall(\theta,\psi)\in\mathbb{R}^p\times\mathbb{R}^q,\quad \mathcal{L}(\theta,\psi)=\int \ell(-f(g(z;\theta);\psi))\,\mathrm{d}\omega(z)+\int f(x;\psi)\,\mathrm{d}\nu(x),\qquad (3.13)$$

where $\ell\colon\mathbb{R}\to\mathbb{R}$ can take different values in order to represent different types of GANs. For instance, taking $\ell(t)=-\log(1+\exp(-t))$ recovers the original (negative log-likelihood) GAN formulation as written in (3.7):

$$\mathcal{L}_{\mathrm{CE}}(\theta,\psi)=-\int \log(1+\exp(-f_\psi(x)))\,\mathrm{d}\nu(x)-\int \log(1+\exp(f_\psi(g_\theta(z))))\,\mathrm{d}\omega(z),\quad (3.14)$$

whereas taking $\ell(t)=-t$ recovers the *unconstrained* WGAN formulation — constrained by setting $\Psi=\Psi_{\mathrm{Lip}_1}=\{\psi\in\mathbb{R}^q\mid \mathrm{Lip}(f_\psi)\leqslant 1\}$, distinction not considered by the authors. Under the assumption that ℓ is concave, local convergence to equilibrium points is shown under gradient descent-ascent update rule. Furthermore, the work [NK17] adds a regularizer on the objective, very similar to [MNG17], in order to help convergence of the optimization scheme.

Different optimizers can be considered in order to solve (3.11). For instance, extra-gradient computes an extra step to evaluate the gradient at a better location. Another example is the lookahead scheme, which computes a backward step at a constant interval [ZLB+19; HK22]

Regularizer in [RLN+17] is defined on the sample space (similar to the gradient penalty in [GAA+17]). Regularizer in [MNG17] is a gradient on the parameter space.

3.2.1 Optimization algorithms

Gradient Descent-Ascent Given $\{\eta_{\mathrm{g}}^{(k)}\}_{k\in\mathbb{N}_0}$ and $\{\eta_{\mathrm{d}}^{(k)}\}_{k\in\mathbb{N}_0}$ two sequences of positive learning rates, and the initializations $\theta^{(0)}\in\Theta$ and $\psi^{(0)}\in\Psi$, the most common gradient descent-ascent rules for (3.11) are

$$\forall k \in \mathbb{N}_0, \quad \begin{cases} \psi^{(k+1)} = \mathrm{p}_\Psi \left(\psi^{(k)} + \eta_{\mathrm{d}}^{(k)} \nabla_\psi \mathcal{L}_{\mathrm{d}}(\psi^{(k)}, \psi^{(k)}) \right) \\ \theta^{(k+1)} = \mathrm{p}_\Theta \left(\theta^{(k)} - \eta_{\mathrm{g}}^{(k)} \nabla_\theta \mathcal{L}_{\mathrm{g}}(\theta^{(k)}, \psi^{(k)}) \right) \end{cases} \tag{SimGDA}$$

for the simultaneous gradient descent-ascent and

$$\forall k \in \mathbb{N}_0, \quad \begin{cases} \psi^{(k+1)} = \mathrm{p}_\Psi \left(\psi^{(k)} + \eta_{\mathrm{d}}^{(k)} \nabla_\psi \mathcal{L}_{\mathrm{d}}(\psi^{(k)}, \psi^{(k)}) \right) \\ \theta^{(k+1)} = \mathrm{p}_\Theta \left(\theta^{(k)} - \eta_{\mathrm{g}}^{(k)} \nabla_\theta \mathcal{L}_{\mathrm{g}}(\theta^{(k)}, \psi^{(k+1)}) \right) \end{cases} \tag{AltGDA}$$

for the alternating gradient descent-ascent, where for some set $S \subseteq \mathbb{R}^n$, $\mathrm{p}_S \colon \mathbb{R}^n \to S$ is some projection operator. The projection steps are usually unnecessary since, in most implementations, $\Theta = \mathbb{R}^p$ and $\Psi = \mathbb{R}^q$.

Note that, if $\mathrm{p}_\Theta = \mathrm{id}_{\mathbb{R}^p}$ and $\mathrm{p}_\Psi = \mathrm{id}_{\mathbb{R}^q}$, and further $\forall k \in \mathbb{N}_0$, $\eta_{\mathrm{d}}^{(k)} = \eta_{\mathrm{g}}^{(k)} \equiv \eta^{(k)}$, Update (SimGDA) can be condensed into a single operation on the concatenated parameters (θ, ψ). In this case, at each step $k \in \mathbb{N}_0$, the overall rule can be written with an *update operator* $F_{\eta^{(k)}}$ defined as

$$\begin{aligned} F_{\eta^{(k)}} \colon \Theta \times \Psi &\longrightarrow \Theta \times \Psi \\ \begin{pmatrix} \theta \\ \psi \end{pmatrix} &\longmapsto \begin{pmatrix} \theta \\ \psi \end{pmatrix} + \eta^{(k)} \begin{pmatrix} -\nabla_\theta \mathcal{L}_{\mathrm{g}}(\theta, \psi) \\ \nabla_\psi \mathcal{L}_{\mathrm{d}}(\theta, \psi) \end{pmatrix}, \end{aligned}$$

so that, for all $k \in \mathbb{N}_0$,

$$(\theta^{(k+1)}, \psi^{(k+1)}) = F_{\eta^{(k)}}(\theta^{(k)}, \psi^{(k)}),$$

and the optimization can be studied with the tools of fixed-point operators [MNG17; MGN18].

For Update (AltGDA), the operator is the combination of the two updates, i.e. $F_{(\eta_{\mathrm{d}}, \eta_{\mathrm{g}})} = F_{\mathrm{g}, \eta_{\mathrm{g}}} \circ F_{\mathrm{d}, \eta_{\mathrm{d}}}$, with $F_{\mathrm{d}, \eta_{\mathrm{d}}} = \mathrm{id}_{\Theta \times \Psi} + \eta_{\mathrm{d}}(0, \nabla_\psi \mathcal{L}_{\mathrm{d}})$ and $F_{\mathrm{g}, \eta_{\mathrm{g}}} = \mathrm{id}_{\Theta \times \Psi} + \eta_{\mathrm{g}}(-\nabla_\theta \mathcal{L}_{\mathrm{g}}, 0)$, which can be used to analyse the alternating update scheme, as done in [MGN18].

3.2.2 Dynamical analysis

The nested problem (3.12) can be analyzed through the lens of dynamical system analysis.

For instance, the work of [NK17] studies the continuous time setting. In this case, the system is given by

$$(\dot{\theta}(t), \dot{\psi}(t)) = v(\theta(t), \psi(t)), \tag{3.15}$$

where $v \colon \Theta \times \Psi \to \Theta \times \Psi$ is the vector field $v = (-\nabla_\theta \mathcal{L}_{\mathrm{g}}, \nabla_\psi \mathcal{L}_{\mathrm{d}})$.

The authors show that, under some realisation condition and the condition that ℓ in Equation (3.13) is strictly concave (ruling out unregularized WGAN), the continuous

system (3.15) is locally exponentially fast convergent to local minima. The regularization they suggest guarantees *any* GAN with a concave ℓ to have exponential stability around the equilibrium and relies on computing the gradient of the loss with respect to the generator parameters θ.

The works of [MNG17; MGN18] focus on the discrete time setting and come up with different regularizers to help convergence.

In [MNG17], the authors develop the "consensus optimization" regularization, based on the squared norm of the vector field v in (3.15), i.e. $R(\theta, \psi) = \frac{1}{2}\|v(\theta, \psi)\|^2$. With $\lambda \in (0, \infty)$ a regularization strength, the flow (3.15) is then performed on the modified vector field $w(\theta, \psi) = v(\theta, \psi) - \lambda \nabla R(\theta, \psi)$. This flow update can be discretized with a step size $\eta \in (0, \infty)$, and the authors show that, for η small enough, the iterates on w converge to a local Nash-equilibrium of the problem, at least linearly.

In [MGN18], the regularisation is based on the gradient of the squared norm of the discriminator with respect to the input parameter $\int \|f(x; \psi)\|^2 \, \mathrm{d}\alpha(x)$, sampled either from the generated ($\alpha = \mu_\theta$) or the target ($\alpha = \nu$) distribution. Using this regularization, and under regularity assumptions on the generated samples, the authors then show at least linear local convergence of Updates (SimGDA) and (AltGDA) to optimal values.

This approach of regularization is fruitful in helping GANs to converge, and is widely used by practitioners, as in [KLA18] for instance. A study of different regularizers can be found in [KLZ+19]. Nonetheless, the exact analysis of a dynamical system such as in System (3.11) is highly intricate. Therefore, a strong hypothesis is often made to be able to derive theoretical results on GANs; namely, the optimal discriminator hypothesis, as discussed next.

3.3 Optimal discriminator hypothesis

One of the techniques for studying GANs is to assume that the inner problem (maximisation problem) is solved perfectly at each time. This is referred to as the *perfect* (or *optimal*) *discriminator hypothesis*. This also implicitly assumes that the discriminator representation capacity is large enough to represent any continuous function $f : X \to \mathbb{R}$, or at least, any optimizer of the inner problem.

This hypothesis amounts to solving, for a given $\theta \in \Theta$,

$$\psi^*(\theta) \in \operatorname{argmax} \{\mathcal{L}_\mathrm{d}(\theta, \psi) \mid \psi \in \Psi\} \tag{3.16}$$

in closed-form (the maximum exists when $\psi \mapsto \mathcal{L}_\mathrm{d}(\theta, \psi)$ is concave for all $\theta \in \Theta$ and Ψ is closed). Then, there only remains a minimisation problem in $\theta \in \Theta$:

$$\operatorname{minimize} \{\mathcal{L}_\mathrm{g}(\theta, \psi^*(\theta)) \mid \theta \in \Theta\}. \tag{3.17}$$

Doing this allows to express the loss as a statistical divergence, or even sometimes a distance, between the two distributions that are being compared, i.e., μ_θ and ν in Equations (3.4) and (3.3).

For instance, in [GPM+14], the authors show that, under the optimal discriminator hypothesis, the overall optimization scheme using $\mathcal{L}_\mathrm{d} = \mathcal{L}_\mathrm{g} = \mathcal{L}$ is equivalent, for the generator, to minimizing the Jensen-Shannon divergence between μ_θ and ν in the formulation of the cross-entropy GAN loss (3.14). Different examples of the actual divergence minimized by a GAN under the perfect discriminator hypothesis are given in [CMF20, Appendix B].

Assuming a perfect discriminator goes on par with an additional level of abstraction: consider the loss of the generator as a loss on the probability measures parametrized by the generator. In this case, the loss $\mathcal{L}_\mathrm{g}$ can be defined on Θ only, as it does *not* depend on the discriminator parameters anymore.

Furthermore, given a generator $g\colon Z \times \Theta \to X$, its loss $\mathcal{L}_\mathrm{g}$ can be abstractly defined on $\mathscr{P}(X)$, as g parametrizes a family of distributions as follows. Given a generator g as in (3.1) and a latent measure $\omega \in \mathscr{P}(Z)$, the family of generated distribution parametrized by $\theta \in \Theta$ is given by

$$\{\mu_\theta = g_{\theta\#}\omega \mid \theta \in \Theta\}. \tag{3.18}$$

Then, given a convex cost on measures $J\colon \mathscr{M}(X) \to (-\infty, \infty]$, learning the generator of a GAN *under the perfect discriminator hypothesis* boils down to finding the minimizer

$$\theta^* \in \operatorname{argmin}\{J(\mu_\theta) \mid \theta \in \Theta\}. \tag{3.19}$$

This has been, from the first paper on GAN [GPM+14] on, a means to study the convergence properties of the otherwise involved optimization problem. In [CMF20], the authors embrace the definition of the function on $\mathscr{M}(X)$ in order to leverage dual convexity results. We follow their presentation in the following.

Assume that X is a compact set. In order to restrict the domain of J to probability measures, set $J(\mu) = \infty$ for $\mu \in \mathscr{M}(X) \setminus \mathscr{P}(X)$. Since J is convex, it is equal to its biconjugate J^{**} and, since X is assumed compact, $\mathscr{C}(X)' \cong \mathscr{M}(X)$ as per Theorem 2.21, so that

$$\forall \mu \in \mathscr{P}(X),\ J(\mu) = J^{**}(\mu) = \sup\{\langle \mu, \varphi\rangle - J^*(\varphi) \mid \varphi \in \mathscr{C}(X)\}. \tag{3.20}$$

This expression gives way to evaluating J via $\varphi \in \mathscr{C}(X)$ a continuous function, and J^* the convex conjugate of J (see Definition 2.9).

From this, the general minimax optimization scheme for a GAN can be understood as the following nested problem

$$\inf\{J(\mu_\theta) \mid \theta \in \Theta\} = \inf\{\sup\{\langle \mu_\theta, \varphi\rangle - J^*(\varphi) \mid \varphi \in \mathscr{C}(X)\} \mid \theta \in \Theta\} \tag{3.21}$$

Under the *optimal discriminator hypothesis*, the supremum in Equation (3.20) is assumed to be achieved, and the optimal discriminator $\Phi_\mu \in \mathscr{C}(X)$ is the function that achieves the supremum, so that $J(\mu) = \langle \mu, \Phi_\mu \rangle - J^*(\Phi_\mu)$. This is equivalent to $\Phi_\mu \in \partial J(\mu)$ and $\mu \in \partial J^*(\Phi_\mu)$, as stated by Proposition 2.12. In this context, the dual variable $\varphi \in \mathscr{C}(X)$ is the discriminator, and the loss dual J^* is the discriminator loss. One can then reason on the optimal discriminator Φ_μ, and assume that the inner problem is well-enough solved in order to validate this hypothesis. This is presented in the following part.

GAN regularity conditions

In [CMF20], some conditions on the discriminator and the generator of a GAN are stated in order for the optimization problem Equation (3.21) to be L-Lipschitz [CMF20, Theorem 1], and hence ensure convergence of the gradient descent algorithm on the generator loss to a stationary point. The conditions on the discriminator are stated for the optimal discriminator Φ_μ achieving the supremum in Equation (3.21). With the notations previously introduced, we report here these regularity conditions.

Recall that, for $L \in (0, \infty)$, we say that a function f is L-Lipschitz if $\mathrm{Lip}(f) \leqslant L$, where $\mathrm{Lip}(f)$ is defined by Equation (2.27).

Regularity Conditions (RC, [CMF20, Theorem 1])**.**

(D1) (*Lipschitz continuity*) $x \mapsto \Phi_\mu(x)$ is α-Lipschitz.

(D2) (*Lipschitz smoothness I*) $x \mapsto \nabla_x \Phi_\mu(x)$ is β_1-Lipschitz.

(D3) (*Lipschitz smoothness II*) $\mu \mapsto \nabla_x \Phi_\mu(x)$ is β_2-Lipschitz with respect to the 1-Wasserstein distance, i.e.
$\exists \beta_2 \in [0, \infty), \forall x \in X, \forall (\mu, \mu') \in \mathscr{M}(X)^2, \quad \|\nabla_x \Phi_\mu(x) - \nabla_x \Phi_{\mu'}(x)\|_2 \leqslant \beta_2 \mathcal{W}_1(\mu, \mu')$.

(G1) (*Lipschitz parametrization*) $\theta \mapsto g_\theta(z)$ is A-Lipschitz in expectation, i.e.
$\exists A \in [0, \infty), \forall (\theta, \theta') \in \Theta^2, \int \|g_\theta(z) - g_{\theta'}(z)\|_2 \, \mathrm{d}\omega(z) \leqslant A \|\theta - \theta'\|_2$.

(G2) (*Lipschitz-smooth parametrization*) $\theta \mapsto \partial_\theta g_\theta(z)$ is B-Lipschitz in expectation, i.e.
$\exists B \in [0, \infty), \forall (\theta, \theta') \in \Theta^2, \int \|\partial_\theta g_\theta(z) - \partial_{\theta'} g_{\theta'}(z)\|_2 \, \mathrm{d}\omega(z) \leqslant B \|\theta - \theta'\|_2$.

Notice how Conditions (G1) and (G2) do not depend on the loss considered, but rather on the architecture of the generator network alone, whereas Conditions (D1) to (D3) are dependent on the actual functional used through the expression of the optimal discriminator Φ_μ achieving the supremum in Equation (3.20).

The authors show that, under those conditions, the mapping $\theta \mapsto J(\mu_\theta)$ is L-smooth, with $L = \alpha B + A^2(\beta_1 + \beta_2)$. In this case, the overall minimisation problem (3.21)

admits a stationary point, to which gradient descent with an adequate learning rate can converge.

They also study different types of GANs, and check whether they satisfy Conditions (D1) to (D3) or not, summarizing their findings in [CMF20, Table 2]. A question mark is left in the case of WGANs for Condition (D3). They also leave aside the study of Conditions (G1) and (G2).

In the next section, we take a closer look at Condition (D3) in a specific WGAN setting, and show that it is violated when a particular loss is considered, and show that Conditions (G1) and (G2) are satisfied for a shallow linear network.

3.4 Examples

We provide in this section an in-depth study of a specific example where 1-Wasserstein distance enjoys a simple expression on the one hand, and the study of a shallow linear generator on the other hand.

3.4.1 Discrete-GAN with discrete metric

The objective of this section is to complete [CMF20, Table 2], where Condition (D3) is unknown to be true for a Wasserstein GAN. In the discrete setting, and with the discrete ground metric, we are able to answer the question by the negative.

Notation In order to lighten the subscript notations, we shall indistinctly treat vectors (resp. matrices) as functions on the discrete set $[n] := \{1, \dots, n\}$ (resp. $[n]^2$).

Assume the sample space X to be discrete, so that $X = [n]$ for some $n \in \mathbb{N}$. The set of distributions $\mathscr{P}(X)$ is then the simplex $\Delta_{n-1} := \{\mu \in [0,\infty)^n \mid \sum_{i\in[n]} \mu(i) = 1\}$.

Let $d\colon X \times X \to [0,\infty]$ be any ground metric on X, so that (X,d) is a metric space. This metric will define a graph structure on X.

Indeed, given (X,d), a weight function w can be defined on X^2 as

$$\begin{array}{rccl} w\colon & X\times X & \longrightarrow & [0,\infty), \\ & (i,j) & \longmapsto & w(i,j) = \begin{cases} d(i,j)^{-2} & \text{if } d(i,j) \in (0,\infty) \\ 0 & \text{else} \end{cases}. \end{array} \tag{3.22}$$

This can be interpreted as a weighted graph $G = (X, E, w)$ with vertices X and where the set of edges E is simply $E = \{(i,j) \in X^2 \mid w(i,j) > 0\}$.

This setting has already been considered in previous works. For instance, the optimization problem of minimizing the Wasserstein distance to independence models with discrete sample space has been studied in [ÇJM+20]. The optimization problem of minimizing optimal transport distance to a variety has been studied by [ÇJM+19].

Given a discrete metric space (X, d) and its corresponding weight function w (3.22), the *discrete gradient* of a function $\Phi\colon X \to \mathbb{R}$ is defined as

$$\begin{array}{rccl} \nabla\Phi\colon & X \times X & \longrightarrow & \mathbb{R}, \\ & (i,j) & \longmapsto & \nabla\Phi(i,j) = \sqrt{w(i,j)}(\Phi(i) - \Phi(j)) \end{array} \tag{3.23}$$

This definition leverages the ground metric d through w so that the closer nodes are, the more they influence each other.

We denote by $\nabla\Phi(i) = \nabla\Phi(i,\cdot)$ the gradient at a particular node $i \in X$ corresponding to all contributions leaving i, i.e. $\nabla\Phi(i) = (\nabla\Phi(i,j))_{j=1}^{n}$.

Further, the *discrete Lipschitz constant* of $\Phi\colon X \to \mathbb{R}$ is defined as

$$\operatorname{Lip}(\Phi) := \sup_{\substack{(i,j)\in X^2, \\ i\neq j}} \sqrt{w(i,j)}|\Phi(i) - \Phi(j)| = \sup_{(i,j)\in X^2} |\nabla\Phi(i,j)| = \|\nabla\Phi\|_\infty. \tag{3.24}$$

Analogously to (2.28), we denote $\operatorname{Lip}_1(X) = \{\Phi\colon X \to \mathbb{R} \mid \operatorname{Lip}(\Phi) \leqslant 1\}$.

From now on, consider the discrete metric

$$\begin{array}{rccl} \bar{\delta}\colon & X^2 & \longrightarrow & \{0,1\}, \\ & (i,j) & \longmapsto & \bar{\delta}(i,j) := \mathbb{1}[i \neq j]. \end{array} \tag{3.25}$$

Let $\nu \in \Delta_{n-1}$ be a target measure. Let the convex functional in (3.19) be

$$\begin{array}{rccl} J\colon & \Delta_{n-1} & \longrightarrow & \mathbb{R}, \\ & \mu & \longmapsto & \mathcal{W}_{\bar{\delta},1}(\mu,\nu), \end{array} \tag{3.26}$$

where $\mathcal{W}_{\bar{\delta},1}$ is the 1-Wasserstein distance with the discrete metric $\bar{\delta}$ (3.9): for $\mu \in \Delta_{n-1}$ and $\nu \in \Delta_{n-1}$ two distributions on X, one has

$$\begin{aligned} \mathcal{W}_{\bar{\delta},1}(\mu,\nu) &= \sup\left\{ \int u \,\mathrm{d}(\mu-\nu) \;\middle|\; u \in \operatorname{Lip}_1(X) \right\} \\ &= \sup\left\{ \sum_{i\in[n]} u(i)(\mu(i) - \nu(i)) \;\middle|\; u \in \operatorname{Lip}_1(X) \right\}. \end{aligned} \tag{3.27}$$

We now leverage the simple expression of the 1-Wasserstein distance on Δ_{n-1} in the case the ground metric is $\bar{\delta}$.

Lemma 3.2 *Let $\mu \in \Delta_{n-1}$ and $\nu \in \Delta_{n-1}$ be two distributions. The 1-Wasserstein distance with ground metric $\bar{\delta}$ (3.25) between μ and ν is $\mathcal{W}_{\bar{\delta},1}(\mu,\nu) = \frac{1}{2}\|\mu - \nu\|_1$.*

Proof. For $(\mu, \nu) \in (\Delta_{n-1})^2$, consider the dual expression of the 1-Wasserstein distance (3.27):

$$\mathcal{W}_{\bar{\delta},1}(\mu,\nu) = \sup\left\{ \sum_{i\in[n]} u(i)(\mu(i)-\nu(i)) \;\middle|\; u \in \mathrm{Lip}_1(X) \right\}. \tag{3.28}$$

Under the discrete metric $\bar{\delta}$, $\mathrm{Lip}_1(X) = \{u \in \mathbb{R}^n \mid \forall (i,j) \in [n]^2, |u(i)-u(j)| \leqslant 1\}$. In order to attain the maximum in (3.28), one has to set $u(i) = \frac{1}{2}$ when $\mu(i) - \nu(i) \geqslant 0$ and $u(i) = -\frac{1}{2}$ when $\mu(i) - \nu(i) \leqslant 0$, resulting in $\mathcal{W}_{\bar{\delta},1}(\mu,\nu) = \frac{1}{2}\|\mu-\nu\|_1$ as claimed. □

The functional J (3.26) is therefore equivalently written, for $\nu \in \Delta_{n-1}$ a target distribution, as

$$\forall \mu \in \Delta_{n-1}, \quad J(\mu) = \frac{1}{2}\|\mu - \nu\|_1.$$

In this setting, Condition (D3) is not satisfied. We will show it by providing a counter-example.

Let $n = 3$, $\Theta = [0,1]$, and let h be a one dimensional model in the simplex Δ_2 defined as

$$\begin{aligned} h\colon \quad [0,1] &\longrightarrow \Delta_2, \\ \theta &\longmapsto \begin{pmatrix} \theta^2 \\ 2\theta(1-\theta) \\ (1-\theta)^2 \end{pmatrix}. \end{aligned} \tag{3.29}$$

Denote $\mathcal{M} = \mathrm{img}\,(h)$ the manifold of the model h (3.29) (see Figure 3.2).

Let $Z = [0,1]$, and let $\omega = \mathscr{U}(Z)$ be the uniform distribution on Z. A parametric generator associated to h (i.e. such that, for all $\theta \in [0,1]$, $h(\theta) = (g_\theta)_{\#}\omega$) is

$$\begin{aligned} g\colon \quad Z \times \Theta &\longrightarrow \{1,2,3\}, \\ (z,\theta) &\longmapsto g(z,\theta) \equiv g_\theta(z) = \begin{cases} 1 & \text{if } z \in [0,\theta^2) \\ 2 & \text{if } z \in [\theta^2, \theta(2-\theta)) \\ 3 & \text{else.} \end{cases} \end{aligned} \tag{3.30}$$

It corresponds to the inverse cumulative distribution function of the random variable $X \sim h(\theta)$. The function g (3.30) is discontinuous (to the left) in both of its argument, at the points (θ, z) for which $z \in \{\theta^2, \theta(2-\theta)\}$.

The functional J (3.26) can be considered as a function $J\colon \mathbb{R}^3 \to \mathbb{R}$. With the usual topology, its gradient is discontinuous at the measures μ for which there exists some $i \in [3]$ such that $\mu(i) = \nu(i)$.

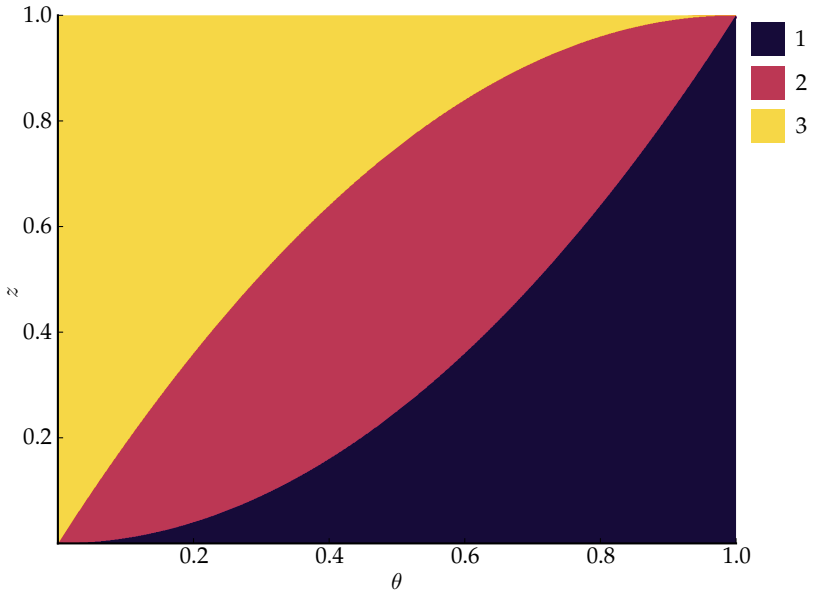


Figure 3.1: Output of g (3.30) on its domain $Z \times \Theta = [0,1] \times [0,1]$.

We appeal to its subdifferential instead: since J is convex in μ, its subdifferential ∂J — as defined by Definition 2.11 — exists and is equal to, for $\mu \in \Delta_2$,

$$\partial_\mu J(\mu) = \left\{ \Phi_\mu \in \mathbb{R}^3 : \forall i \in [3], \begin{cases} \Phi_\mu(i) = \frac{1}{2}\operatorname{sgn}(\mu(i) - \nu(i)) & \text{if } \mu(i) \neq \nu(i) \\ \Phi_\mu(i) \in [-\frac{1}{2}, \frac{1}{2}] & \text{else} \end{cases} \right\}. \quad (3.31)$$

One checks that this also corresponds to the set of optimal discriminators $\{u^* \in \mathrm{Lip}_1(X) \mid J(\mu) = \sum_{i\in[3]} u^*(i)(\mu(i) - \nu(i))\}$ (in accordance with [CBG19, Theorem 2]).

Given one $\mu \in \Delta_2$, an optimal discriminator at μ is therefore $\Phi_\mu \in \partial_\mu J(\mu)$ (3.31). Since $\Phi_\mu \in \mathrm{Lip}_1(X)$, Condition (D1) is satisfied with $\alpha = 1$.

By the definition of the discrete gradient (3.23), and since for all $i \in [3]$, $\Phi_\mu(i) \in [-\frac{1}{2}, \frac{1}{2}]$, one has

$$\begin{aligned} & \forall (k,l) \in [3]^2, \quad \nabla\Phi_\mu(k,l) \in [-1,1] \\ \implies \quad & \forall (i,j) \in [3]^2, \quad \|\nabla\Phi_\mu(i) - \nabla\Phi_\mu(j)\|_2 \leqslant 2\sqrt{3}. \end{aligned}$$

Therefore, the function $i \mapsto \nabla\Phi_\mu(i)$ is $2\sqrt{3}$-Lipschitz, and Condition (D2) is satisfied with $\beta_1 = 2\sqrt{3}$.

The next lemma states that J does not satisfy Condition (D3), expressed in the discrete setting.

Lemma 3.3 *Condition (D3) is not satisfied for the functional* $J\colon \Delta_2 \ni \mu \mapsto J(\mu) = \mathcal{W}_{\bar{\delta},1}(\mu,\nu) = \frac{1}{2}\|\mu-\nu\|_1$.

Proof. Recall that Condition (D3) is

$$\exists \beta_2 \in [0,\infty),\ \forall x \in X,\ \forall (\mu,\mu') \in \mathscr{M}(X)^2,\ \|\nabla_x \Phi_\mu(x) - \nabla_x \Phi_{\mu'}(x)\|_2 \leqslant \beta_2 \mathcal{W}_1(\mu,\mu'). \tag{3.32}$$

The proof consists in building a counter-example to (3.32); that is to say, for any $\nu \in \Delta_2$, and for any $\beta_2 \in [0,\infty)$, find $(\mu^+,\mu^-) \in \mathcal{M}^2$ and a point $i \in X$ such that, $\|\nabla\Phi_{\mu^+}(i) - \nabla\Phi_{\mu^-}(i)\| > \beta_2 \mathcal{W}_{\bar{\delta},1}(\mu^+,\mu^-)$. This amounts to finding (μ^-,μ^+) such that $\mathcal{W}_{\bar{\delta},1}(\mu^-,\mu^+)$ is arbitrarily small and upper bounded by $\|\nabla\Phi_{\mu^+}(i) - \nabla\Phi_{\mu^-}(i)\|$.

The discontinuity of $\mu \mapsto \Phi_\mu(i)$ for some $i \in X$ will be key in finding such distributions. Intuitively, for suitable (μ^+,μ^-) as close to each other as liked (in the sense of $\mathcal{W}_{\bar{\delta},1}$), $\|\nabla\Phi_{\mu^+}(i) - \nabla\Phi_{\mu^-}(i)\|$ will stay constant. We construct such (μ^+,μ^-) as presented next.

Let $\nu \in \Delta_2$, and let $\varepsilon \in (0,\infty)$. Consider $(\mu^+,\mu^-) \in \mathcal{M}^2$ such that $\mu^+(1) - \nu(1) = \varepsilon$, $\mu^-(1) - \nu(1) = -\varepsilon$, $\mu^\pm(2) - \nu(2) < 0, \mu^\pm(3) - \nu(3) > 0$ (see Figure 3.2).

Then, the optimal discriminators $\Phi_+ \equiv \Phi_{\mu^+}$ and $\Phi_- \equiv \Phi_{\mu^-}$ are, by Equation (3.31),

$$\Phi_+ = \frac{1}{2}\begin{pmatrix}1\\1\\-1\end{pmatrix} \text{ and } \Phi_- = \frac{1}{2}\begin{pmatrix}1\\-1\\-1\end{pmatrix}.$$

Hence $\nabla\Phi_+ = \begin{pmatrix}0&1&0\\-1&0&-1\\0&1&0\end{pmatrix}$ and $\nabla\Phi_- = \begin{pmatrix}0&0&-1\\0&0&-1\\1&1&0\end{pmatrix}$, so that the left hand side of (3.32) becomes, at the point $i=1$,

$$\|\nabla\Phi_+(1) - \nabla\Phi_-(1)\| = \|\begin{pmatrix}0&1&1\end{pmatrix}\|_2 = \sqrt{2}. \tag{3.33}$$

This remains true for any $\varepsilon \in (0,\infty)$.

It was shown by [ÇJM+19] that the optimal $\mu^* \in \mathcal{M}$ minimizing the Wasserstein distance to ν, $\mathcal{W}(\mathcal{M},\nu) := \inf\{\mathcal{W}(\mu,\nu) \mid \mu \in \mathcal{M}\}$, lies at the intersection of two cells in Figure 3.2. These cells split Δ_2 in regions where the optimal discriminator Φ_μ is constant.

Denoting $\theta^* \in \Theta = [0,1]$ the optimal parameter such that $h(\theta^*) = \mu^*$, one can construct two sequences μ^+ and μ^- with values in Δ_2 such that

$$\forall k \in \mathbb{N}, \begin{cases} (\mu^+)_k = h(\theta^* + k^{-1}), \\ (\mu^-)_k = h(\theta^* - k^{-1}), \end{cases}$$

and the requirements expressed earlier hold for those sequences.

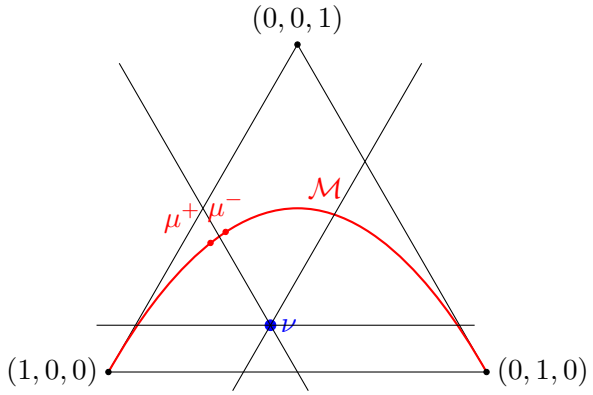


Figure 3.2: The model $\mathcal{M} = \operatorname{img} h$, as defined in Equation (3.29), on Δ_2. The target is $\nu \in \Delta_2$. The lines intersecting at ν depict the distributions μ such that $\mu(i) = \nu(i)$ for some $i \in [3]$. The two distributions μ^+ and μ^- are the ones defined in the proof of Lemma 3.3.

By the continuity of h, we have $(\mu^+)_k \xrightarrow[k\to\infty]{} \mu^*$ and $(\mu^-)_k \xrightarrow[k\to\infty]{} \mu^*$, so that

$$\mathcal{W}_{\bar{\delta},1}\left((\mu^+)_k, (\mu^-)_k\right) \xrightarrow[k\to\infty]{} 0$$

for any ground metric that is a distance on X, whereas the norm of the gradient at $i = 1$ stays constant, as per Equation (3.33)

$$\forall k \in \mathbb{N},\ \|\nabla\Phi_{(\mu^+)_k}(1) - \nabla\Phi_{(\mu^-)_k}(1)\|_2 = \sqrt{2},$$

which violates the condition (3.32). □

Notice that, due to the discrete nature of g (3.30), Condition (G1) is readily satisfied. The discontinuity of g would require a more detailed definition of the operator $\partial_\theta g_\theta$ that appears in Condition (G2).

Instead, we show in the next section an instance where the generator satisfies Conditions (G1) and (G2) in the continuous setting.

3.4.2 Shallow linear generator

In this last section, we study the smoothness assumptions on the generator (Conditions (G1) and (G2)) for a shallow linear generator architecture.

Notation (*Einstein summation convention*). Since the generator is a map $g\colon Z \times \Theta \to X$, its derivative with respect to the parameters is a tensor. We will adopt the Einstein notation convention in order to compute the tensor expressions. We will use the convention of

writing all indices as subscripts. A summation is implicitly performed for every index that is repeated once. For instance, the matrix-matrix product between two matrices A and B with compatible sizes produces a matrix C so that $C_{ij} = A_{ik}B_{kj}$, where the sizes are derived from context and a sum over the index k is performed.

Let X be a sample space, such that $X \subseteq \mathbb{R}^{d_\mathrm{x}}$, and let Z be the latent space, such that $Z \subseteq \mathbb{R}^{d_\mathrm{z}}$. We will consider the case of a shallow linear generator; i.e., $\Theta = \mathbb{M}(d_\mathrm{x}, d_\mathrm{z}) \times \mathbb{R}^{d_\mathrm{x}}$, and, for $\theta = (W, b) \in \Theta$, $g_\theta \colon Z \to X;\ z \mapsto g_\theta(z) = Wz + b$.

Assuming $\omega \in \mathscr{P}(X)$ has finite first moment (i.e. $\int \|z\|_2 \,\mathrm{d}\omega(z) < \infty$), Conditions (G1) and (G2) are satisfied for a shallow linear generator $g_\theta(z) = Wz + b$. Indeed, in this case, for any $(\theta, \theta') = ((W, b), (W', b')) \in \Theta^2$, one has

$$\begin{aligned}\forall z \in Z,\ \|g_\theta(z) - g_{\theta'}(z)\|_2 &= \|Wz + b - W'z - b'\|_2 \\ &\leqslant \|(W - W')z\|_2 + \|b - b'\|_2.\end{aligned}$$

Therefore, denoting $a = \int \|z\|_2 \,\mathrm{d}\omega(z) < \infty$,

$$\begin{aligned}\int \|g_\theta(z) - g_{\theta'}(z)\|_2 \,\mathrm{d}\omega(z) &\leqslant \|W - W'\|_2 \int \|z\|_2 \,\mathrm{d}\omega(z) + \|b - b'\|_2 \\ &\leqslant a\|W - W'\|_2 + \|b - b'\|_2,\end{aligned}$$

so that Condition (G1) is satisfied.

Further, with $1_n := (1, \dots, 1)^\top \in \mathbb{R}^n$, the differential of g_θ at $z \in Z$ with respect to $\theta \in \Theta$ is $\partial_\theta g_\theta(z) = \{\frac{\partial g_\theta(z)}{\partial W}, \frac{\partial g_\theta(z)}{\partial b}\} = \{\frac{\partial g_\theta(z)}{\partial W}, 1_n\}$. This is a three-dimensional tensor. Let $x = Wz + b$ and denote $\delta_{ij} = \mathbb{1}[i = j]$ the Kronecker delta. Then, using Einstein summation convention, one has

$$\begin{aligned}\frac{\partial x_p}{\partial W_{qr}} &= \frac{\partial}{\partial W_{qr}}(W_{pi}z_i + b_p) \\ &= \delta_{pq}\delta_{ri}z_i \\ &= \delta_{pq}z_r, \\ \frac{\partial x_p}{\partial b_q} &= \delta_{pq}.\end{aligned}$$

Therefore, $\partial_{\theta_1} g_{\theta_1}(z) = \partial_{\theta_2} g_{\theta_2}(z)$ and Condition (G2) is trivially satisfied.

Understanding how Conditions (G1) and (G2) are satisfied in the case of a more complex generator architecture (e.g. deep linear, shallow non-linear) is an obvious goal for future endeavors.

3.5 Conclusion

Throughout the chapter, we have introduced the general GAN optimization framework and showed why it is challenging to study it. Two approaches are used to simplify it, namely the optimal discriminator hypothesis and the simplification of the networks themselves. The next chapter, Chapter 5, will rely on those assumptions in order to derive convergence guarantees of a gradient flow and gradient descent schemes.

Although some analysis was possible in the discrete case on the one hand (Section 3.4.1) and for a shallow linear generator on the other hand (Section 3.4.2), a general GAN is more challenging to analyse. Obtaining mathematical proofs of convergence for a more complex GAN is intricate, and a general theory for its minimax optimization scheme remains unknown.

Crucially, the problem becomes non convex-concave as soon as the architecture of the generator becomes deep, with several layers of neurons. This non-convex-concavity introduces the most challenging aspect of the general optimization scheme.

As we will investigate in the next two chapters, the assumption introduced in Section 3.3 will be useful in order to reduce the problem complexity and allow the analysis in the specific of a Gaussian distribution generation problem.

4 Bures-Wasserstein loss landscape

Some of this chapter is based on the work published as [BPA+23]. *PB's main contributions are the study of the properties of the Bures-Wasserstein loss differential, the expression of the critical points, the investigation with the connection to the Eckart-Young-Mirsky theorem, the statement of correspondence between the different parametrized spaces, and the expression and empirical evaluation of the Hessians.*

As we have seen in the previous chapter, a generic GAN optimization is challenging as it involves two competing objectives, and the resulting minimization-maximization problem is generally not convex-concave.

The next two chapters are devoted to the analysis of a simple GAN setting under the optimal discriminator hypothesis, which effectively removes one layer of optimization and reduces the overall problem to a simpler minimization problem.

In this chapter, we introduce the Bures-Wasserstein deep linear network learning problem (Section 4.2) and analyse the optimization landscape of the corresponding learning problem. Namely, we derive the expression of the critical points for the Bures-Wasserstein loss under the covariance matrix (or Burer-Monteiro) parametrization, and inspect some of their properties. Such an analysis is required in order to conclude on the overall convergence to the global minimizer of the loss by gradient descent algorithms, which will be performed in Chapter 5.

The problem at hand is challenging from different points of view. First, the Bures-Wasserstein loss is not everywhere differentiable, leading to the introduction and analysis of its regularization. Second, the chaining of different parametrizations inflates the number of possible degeneracies and makes the correspondence between the different spaces non-trivial. We will cover this point between the covariance and the function space. Last, the critical parameters are not all minimizers, but we will show that the loss on the function space satisfies the so-called *strict-saddle-points property*, which states that its critical points are either global minimizers or strict saddle points. This allows to conclude on the nature of the critical parameters that correspond to a critical function, as stated in Proposition 4.12.

We will derive a general expression for the first and second-order derivative informations, under both a deep linear network and a Burer-Monteiro (covariance) parametrizations.

The chapter is organised as follows. We introduce the learning problem in Section 4.2. In Section 4.3, we provide the expression of the critical points for both the original and the regularized Bures-Wasserstein loss. In Section 4.5, we draw the link between the resulting

critical points obtained and the celebrated Eckart-Young-Mirsky theorem, stated in the case of a squared-norm type of loss. In Section 4.6, we give the expression and numerical evaluations of the Hessian of the Frobenius and the Bures-Wasserstein losses, in the parameter space. This gives insight on the different behavior running a gradient-descent type of algorithm to minimize such a loss.

4.1 Motivation

In this chapter, we investigate generative deep linear networks and their optimization using the Bures-Wasserstein distance. More precisely, we consider the problem of learning a target Gaussian distribution using a deep linear neural network. The family of distributions generated by our model is again a family of Gaussian distributions. The optimization is driven by the minimization of the Bures-Wasserstein distance between the target and generated distributions, or, more exactly, their covariances.

This problem is relevant in two ways. First, it pertains to the optimization of deep linear networks for a type of loss that is qualitatively different from the well-studied and very particular squared error loss. Second, it can be regarded as a simplified yet instructive instance of the parameter optimization problem in generative networks, specifically Wasserstein generative adversarial networks, which are currently not as well understood as discriminative models.

The optimization landscapes and the properties of parameter optimization procedures for neural networks are among the most puzzling and actively studied topics in theoretical deep learning [see, e.g. MMN18; LZB22]. Deep linear networks, i.e. neural networks having the identity as activation function, serve as simplified models for such investigations [BH89; Kaw16; TKB20; KMM+22; BRT+21]. The study of linear networks has guided the development of several useful notions and intuitions in the theoretical analysis of neural networks, from the absence of bad local minima to the role of parametrization and overparametrization in gradient optimization [ACH18; ACG+19; ACH+19].

Many previous works have focused on discriminative or autoregressive settings and have emphasized the squared Frobenius error loss. Although this loss is indeed a popular choice in regression tasks, it interacts in a very special way with the particular geometry of linear networks [TKB20]. The behavior of linear networks optimized with different losses has also been considered in several works [LB18; LK17; TKB20] but is less well understood.

Notably, [DSF+21] observed experimentally that the Bures-Wasserstein metric reduces the infamous problem of mode collapse in GANs. In particular, the authors reported improvements in mode coverage and generation quality by adding the Bures metric to the objective function of a GAN. Our work casts light on the theoretical properties

of Bures-Wasserstein metric as a loss function to train deep linear generative neural networks, by studying a specific 2-Wasserstein GAN model.

We report next the most relevant previous works on the topics of optimization of deep linear networks and the Bures-Wasserstein distance considered as a training loss.

Overview of previous optimization landscape results First, we give a quick overview of the literature on the topic, which mostly focuses on the squared (Frobenius) error loss.

Indeed, the optimization landscape of linear networks has been studied in numerous works for the squared error loss. The pioneering work of [BH89] focused on the two-layer case, and showed that there is a single minimum (up to a trivial parametrization symmetry) and all other critical points are saddle points. [Kaw16] obtained corresponding results for deep linear networks and showed the existence of bad saddles (with no negative Hessian eigenvalues) in parameter space for networks with more than three layers. [LK17] showed that if the loss is such that any local minimizer in parameter space can be perturbed to an equally good minimizer with full-rank factor matrices, then all local minimizers in parameter space are local minimizers in function space. [YSJ18] found sets of parameters in which any critical point is a global minimizer, and any outside critical point is a saddle point. We also mention other works that study critical points for different types of neural network architectures, such as deep linear residual networks [HM17] and deep linear convolutional networks [KMM+22; KMS+23].

There are also several results for different losses. [LB18] showed that for deep linear nets with no bottlenecks all local minima are global for arbitrary convex differentiable losses. [TKB20] found that for linear networks with arbitrarily rank-constrained function space, the squared error loss is special in the sense that it ensures the non-existence of non-global local minima. However, for arbitrary convex losses, non-global local minimizers, when they exist, are always pure, meaning that they correspond to local minimizers in function space.

We also mention that [FFG+20] characterized the optimal solution of a 2-Wasserstein GAN with a rank-k linear generator as the k-PCA solution. We will obtain an analogous result in our particular parametrization, albeit with an additional exhaustive description of the critical points.

Low rank matrix approximation The function (or representation) space of a linear network corresponds to $n \times m$ matrices, for some $(n, m) \in \mathbb{N}^2$, of rank at most $\underline{d}$, the smallest width of the network. Hence, the optimization of a linear network in its function space is closely related to the problem of approximating a given data matrix by a low-rank matrix. When the approximation error is measured in Frobenius norm, [EY36] characterized the optimal bounded-rank approximation of a given matrix in terms of its singular value decomposition. [Mir60] obtained the same characterization for the more

general case of unitary invariant matrix norms, which include the Euclidean operator norm and the Schatten-p norms. There are further generalizations to certain weighted norms [RZ79; DL17]. However, for general norms the problem is known to be difficult [SWZ17; GV18; GS19]. We extend the result to the Bures-Wasserstein loss, and explain why the connection holds even though the Bures-Wasserstein distance cannot be expressed as the norm of some matrix.

Bures-Wasserstein distance The Bures-Wasserstein distance was introduced by [Bur69] to study Hermitian operators in quantum information, particularly density matrices. It induces a metric on the space of positive semi-definite matrices, and corresponds to the 2-Wasserstein distance between two centered Gaussian distributions [BJL19]. Wasserstein distances have several useful properties, e.g. they remain well defined between disjointly supported measures and have duality formulations [Vil03] that allow for practical implementations. As described in Chapter 3, this makes them good candidates and indeed popular choices for learning generative models, with a well-known case being the Wasserstein Generative Adversarial Networks (WGANs) [ACB17]. While the 1-Wasserstein distance has been most commonly used in this context, the Bures-Wasserstein distance has also attracted interest, e.g. in the works of [MC18; CMR+20; MGM21], and has also appeared in the context of linear quadratic Wasserstein generative adversarial networks [FFG+20].

A 2-Wasserstein GAN is a minimum 2-Wasserstein distance estimator expressed in Kantorovich duality (see details in Section 2.2.2). This model can serve as a platform to develop the theory particularly when the inner problem can be solved in closed-form. Such a formula is available when comparing pairs of Gaussian distributions, in particular centered Gaussians, which corresponds precisely to the Bures-Wasserstein distance between the corresponding covariance matrices. Strikingly, even in this simple case, the properties of the corresponding optimization problem are not well understood; we aim to address this in the present work. In this chapter, we first focus on the landscape associated to this loss, to then analyse the optimization and convergence properties of the corresponding learning algorithm in Chapter 5.

Notation For $(n, m) \in \mathbb{N}^2$, we will denote $\mathbb{M}(n, m)$ the set of real matrices of n rows and m columns, and write $\mathbb{M}(n) \equiv \mathbb{M}(n, n)$. We denote further $\mathbb{S}(n)$ the set of real symmetric matrices of size n, and $\mathbb{S}_{++}(n)$ (resp. $\mathbb{S}_{+}(n)$) the set of real symmetric positive (semi-)definite matrices of size n. We may denote $A \in \mathbb{S}_{++}(n) \iff A \succ 0$ (resp. $A \in \mathbb{S}_{+}(n) \iff A \succcurlyeq 0$). The set $\mathbb{S}_{+}(n)$ will also be referred to as the *covariance space*. For $A \in \mathbb{S}_{+}(n)$, denote $A^{1/2}$ the (principal) square root of A, i.e. such that $A^{1/2} \in \mathbb{S}_{+}(n)$ and $A = A^{1/2}A^{1/2}$. Moreover, for any $A \in \mathbb{S}(n)$, its extremal eigenvalues are denoted $\lambda_{\{\min,\max\}}(A)$. For any matrix $X \in \mathbb{M}(n, m)$, we write its singular (or spectral) norm as $\|X\|$, and its extremal singular values as $\sigma_{\{\min,\max\}}(X)$. Hence, $\|X\| = \sigma_{\max}(X) = \lambda_{\max}^{1/2}(XX^\top)$. For two matrices $(X, Y) \in \mathbb{M}(n, m)^2$, the Frobenius scalar product is denoted $\langle X, Y \rangle := \operatorname{tr}(XY^\top)$, and its associated norm is $\|X\|_{\mathrm{F}} = \sqrt{\operatorname{tr} XX^\top}$. For two matrices $(A, B) \in \mathbb{S}(n)^2$,

denote $A \succ B \iff A - B \succ 0$ (resp. $A \succcurlyeq B \iff A - B \succcurlyeq 0$). For any discrete set $\mathcal{J} \subseteq [n]$, we write $|\mathcal{J}|$ its cardinality.

The next section introduces the learning problem of a linear Wasserstein GAN.

4.2 Wasserstein generative linear networks

In this section, we introduce the Bures-Wasserstein loss and some of its properties. We will also give its connection to a Gaussian linear GAN problem. Then, we explicitly give the different parametrizations that will be used in the learning problem. On top of the deep linear network parametrization μ already given in (4.3), the so-called *Burer-Monteiro* parametrization will be introduced, in order to parametrize a covariance matrix (hence its alternative name, *covariance* parametrization). The presentation of the 2-Wasserstein distance has already been carried out in Section 2.2.2.

We first introduce the general framework of overparametrized linear neural networks.

4.2.1 Overparametrized linear networks

In the following, we consider a bias-free linear network with $d_0 \in \mathbb{N}$ inputs and $N \in \mathbb{N}$ layers of widths $(d_1, \dots, d_N) \in \mathbb{N}^N$. This is a model of linear functions of the form

$$\begin{array}{rcl} \mathbb{R}^{d_0} & \longrightarrow & \mathbb{R}^{d_N}, \\ x & \longmapsto & W_N \cdots W_1 x, \end{array} \tag{4.1}$$

parametrized by the weight matrices $W_j \in \mathbb{M}(d_j, d_{j-1})$, for all $j \in [N]$, where $\mathbb{M}(d_j, d_{j-1})$ denotes the set of all real matrices with d_j rows and d_{j-1} columns. We will denote the tuple of weight matrices by $\theta = (W_1, \dots, W_N)$ and the space of all such tuples by

$$\Theta = \bigtimes_{j=1}^{N} \mathbb{M}(d_j, d_{j-1}). \tag{4.2}$$

This is the *parameter space* of our model. To slightly lighten the notation, we will also denote the input and output dimensions by $m \equiv d_0$ and $n \equiv d_N$, respectively. Given $\theta \in \Theta$, we may write $W_\theta := W_N \cdots W_1$ for the corresponding end-to-end matrix. Its dependency on θ is then made explicit. For any $1 \leqslant i \leqslant j \leqslant N$, we will also write $W_{j:i} := W_j \cdots W_i$ for the matrix product of layer i up to j. We note that the represented function $(x, \theta) \mapsto W_\theta x$ is linear in the network input x, but the parametrization is *not* linear in the parameters θ. We denote the network's *parametrization map* by

$$\begin{array}{rrcl} \mu\colon & \Theta & \longrightarrow & \mathbb{M}(d_N, d_0), \\ & \theta = (W_1, \dots, W_N) & \longmapsto & W_{N:1} = W_N \cdots W_1. \end{array} \tag{4.3}$$

The *function space* of the network is the set of linear functions it can represent. This corresponds to the set of possible end-to-end matrices, which are the $n \times m$ matrices of rank at most $\underline{d} := \min\{d_0, \dots, d_N\}$. When $\underline{d} = \min\{d_0, d_N\}$, the function space is a vector space. Otherwise, when there is a bottleneck (i.e. when $\underline{d} < \min\{d_0, d_N\}$), it is a non-convex subset of $\mathbb{M}(d_N, d_0)$ determined by polynomial constraints, namely the vanishing of the $(\underline{d}+1) \times (\underline{d}+1)$ determinants minors.

4.2.2 Linear Wasserstein GAN

In this subsection, we draw the connection between the introduced distance and the linear Wasserstein GANs for generating a Gaussian distribution. For simplicity, we present the problem without regularization, i.e. $\tau = 0$ in (4.7).

As per Equation (2.30), the squared distance on $\mathbb{S}_+(n)$ defined in (4.5) corresponds to the squared 2-Wasserstein distance between two zero-centered Gaussians. It can be used as a loss for training models of Gaussian distributions, in particular generative linear networks. Recall that a zero-centered Gaussian distribution is completely specified by its covariance matrix. Given a bias-free linear network (4.1) and a latent Gaussian distribution $\mathcal{N}(0, \mathrm{I}_m)$ (with $m \equiv d_0 \in \mathbb{N}$ the latent space dimension), a linear network generator with parameter $\theta \in \Theta$ computes a push-forward of the latent distribution, which is again a Gaussian distribution. If $Z \sim \mathcal{N}(0, \mathrm{I}_m)$ and $X = W_\theta Z$, then

$$X \sim \mathcal{N}(0, W_\theta W_\theta^\top) =: \nu_\theta.$$

We will denote the map that takes the end-to-end matrix W_θ to the covariance matrix $W_\theta W_\theta^\top$ by

$$\begin{aligned} \pi\colon \quad \mathbb{M}(n,m) &\longrightarrow \mathbb{S}_+(n), \\ W &\longmapsto WW^\top. \end{aligned} \tag{4.4}$$

This parametrization is known in the literature as the *Burer-Monteiro* parametrization [BM05; LLZ21; LKB22].

Given a target distribution $\nu_0 = \mathcal{N}(0, \Sigma_0)$ (or simply a covariance matrix Σ_0, which may be a sample covariance matrix), one can select the best θ by minimizing $\mathcal{B}^2(W_\theta W_\theta^\top, \Sigma_0) = \mathcal{W}_2^2(\nu_\theta, \nu_0)$. This falls under the framework exposed in Section 3.1.2, with the loss being the 2-Wasserstein distance between the two distributions.

4.2.3 The Bures-Wasserstein loss

The Bures-Wasserstein (BW) distance is defined over $\mathbb{S}_+(n)^2$. We recall its definition and some key properties of its gradient.

Definition 4.1 (*Bures-Wasserstein distance, Definition 2.33*). Given two symmetric positive semi-definite matrices $(\Sigma_0, \Sigma) \in \mathbb{S}_+(n)^2$, the squared Bures-Wasserstein distance between Σ_0 and Σ is defined as

$$\mathcal{B}^2(\Sigma, \Sigma_0) = \operatorname{tr}\left(\Sigma + \Sigma_0 - 2\left[\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2}\right]^{1/2}\right), \tag{4.5}$$

and the mapping $\Sigma \mapsto \mathcal{B}^2(\Sigma, \Sigma_0)$ is referred to as the squared Bures-Wasserstein distance *to* Σ_0, or, since Σ_0 will be fixed, simply the *Bures-Wasserstein loss*.

In the remaining, we will focus on the case where the target Σ_0 is definite.

So, let $\Sigma_0 \in \mathbb{S}_{++}(n)$ be fixed. [KSS21, Lemma A.3] shows that the matrix square root is differentiable on the set of positive definite matrices $\mathbb{S}_{++}(n)$. Therefore, one can differentiate the squared Bures-Wasserstein distance to Σ_0 at $\Sigma \in \mathbb{S}_{++}(n)$.

However, the squared Bures-Wasserstein distance to Σ_0 is *not* differentiable at all $n \times n$ matrices. Indeed, given some $\Sigma \in \mathbb{S}_+(n)$, if we let $\Gamma Q \Gamma^\top$ be a spectral decomposition of $\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2}$, Equation (4.5) can then be written as

$$\mathcal{B}^2(\Sigma, \Sigma_0) = \|\Sigma^{1/2}\|_{\mathrm{F}}^2 + \|\Sigma_0^{1/2}\|_{\mathrm{F}}^2 - 2\operatorname{tr} Q^{1/2}. \tag{4.6}$$

Due to the square root on Q, the map $\Sigma \mapsto \mathcal{B}^2(\Sigma, \Sigma_0)$ is not differentiable when the number of positive eigenvalues of Q — i.e., the rank of $\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2}$ — changes.

More specifically, while one can compute, for each $k \in [n]$, the gradient of the mapping $\Sigma \mapsto \mathcal{B}^2(\Sigma, \Sigma_0)$ over the set of matrices of rank k, its norm blows up if the matrix changes rank. We provide the gradient of the BW loss restricted to the set of fixed-rank matrices in Lemma 4.15 and refer the interested reader to [BJL19] for further details on the Bures-Wasserstein distance.

4.2.4 Regularized distance

As noted previously, given $\Sigma_0 \in \mathbb{S}_{++}(n)$, the Bures-Wasserstein loss $L\colon \Sigma \mapsto \mathcal{B}^2(\Sigma, \Sigma_0)$ (4.5) is not differentiable on $\mathbb{S}_+(n)$.

One way to circumvent this difficulty is to regularize by adding a small perturbation $\tau \mathrm{I}_n$, for $\tau \in (0, \infty)$. This ensures that $\Sigma_\tau := \Sigma + \tau \mathrm{I}_n$ is non-singular, which in turn makes the BW loss differentiable at Σ_τ. This leads to the loss

$$\forall \Sigma \in \mathbb{S}_+(n), \quad \mathcal{B}_\tau^2(\Sigma, \Sigma_0) = \operatorname{tr}\left(\Sigma + \Sigma_0 - 2\left[\Sigma_0^{1/2}(\Sigma + \tau \mathrm{I}_n)\Sigma_0^{1/2}\right]^{1/2}\right) + \tau n. \tag{4.7}$$

We can upper-bound the deviation from the original loss, showing a pointwise convergence of $\mathcal{B}_\tau^2$ to $\mathcal{B}^2$ as τ vanishes, as per the following lemma.

Lemma 4.2 *Let $\Sigma \in \mathbb{S}_+(n)$ and let $\tau \in (0,\infty)$. Assume that $\operatorname{rank}(\Sigma_0) = n$. Then, with $\mathcal{B}^2$ given by (4.5) and $\mathcal{B}^2_\tau$ given by (4.7), we have*

$$\left|\mathcal{B}^2_\tau(\Sigma,\Sigma_0) - \mathcal{B}^2(\Sigma,\Sigma_0)\right| \leqslant n\sqrt{\tau}\left(\sqrt{\tau} + \frac{2\lambda_{\max}(\Sigma_0^{1/2})}{\lambda_{\min}(\Sigma_0^{1/2})}\right), \tag{4.8}$$

with $\lambda_{\{\min,\max\}}(\Sigma_0^{1/2})$ the extremal eigenvalues of $\Sigma_0^{1/2}$.

We observe that the upper bound (4.8) is tight in τ in the sense that it goes to zero as τ goes to zero.

Proof. For $\Sigma \in \mathbb{S}_+(n)$ and $\tau \in (0,\infty)$, let $\Sigma_\tau = \Sigma + \tau \mathrm{I}_n$. In view of Equations (4.5) and (4.7), the difference between the perturbative and the original loss at Σ is given by

$$\begin{aligned}
\left|\mathcal{B}^2_\tau(\Sigma,\Sigma_0) - \mathcal{B}^2(\Sigma,\Sigma_0)\right| &= \left|\mathcal{B}^2(\Sigma_\tau,\Sigma_0) - \mathcal{B}^2(\Sigma,\Sigma_0)\right| \\
&= \left|\tau n - 2\operatorname{tr}\left(\left(\Sigma_0^{1/2}\Sigma_\tau\Sigma_0^{1/2}\right)^{1/2} - \left(\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2}\right)^{1/2}\right)\right| \\
&\leqslant \tau n + 2\left|\operatorname{tr}\left(\left(\Sigma_0^{1/2}\Sigma_\tau\Sigma_0^{1/2}\right)^{1/2} - \left(\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2}\right)^{1/2}\right)\right|.
\end{aligned} \tag{4.9}$$

Let $A := \Sigma_0^{1/2}\Sigma\Sigma_0^{1/2}$ and $A_\tau := \Sigma_0^{1/2}\Sigma_\tau\Sigma_0^{1/2} = A + \tau\Sigma_0$. We aim to bound

$$\left|\operatorname{tr}\left(A_\tau^{1/2} - A^{1/2}\right)\right| = \left|\operatorname{tr}\left((A+\tau\Sigma_0)^{1/2} - A^{1/2}\right)\right|.$$

We bound the absolute value of the trace of a matrix $X \in \mathbb{M}(n,m)$ by its spectral norm (as $|\operatorname{tr}(X)| \leqslant n\|X\|$), which can be bounded in turn.

First, assume that Σ (hence A) is full rank, so that $A \succcurlyeq \lambda_{\min}(A)\mathrm{I}_n \succ 0$. We then use the perturbation inequality from [Sch92, Lemma 2.2] and find that

$$\|(A+\tau\Sigma_0)^{1/2} - A^{1/2}\| \leqslant \frac{1}{\sqrt{\lambda_{\min}(A+\tau\Sigma_0)} + \sqrt{\lambda_{\min}(A)}}\|\tau\Sigma_0^{1/2}\|.$$

Since

$$\begin{aligned}
\lambda_{\min}(A+\tau\Sigma_0) &\geqslant \lambda_{\min}(A) + \lambda_{\min}(\tau\Sigma_0), \\
\lambda_{\min}(A) &\geqslant \lambda_{\min}(\Sigma)\lambda_{\min}(\Sigma_0), \\
\text{and} \qquad \forall (a,b) \in [0,\infty)^2, \quad \sqrt{a} + \sqrt{b} &\geqslant \sqrt{a+b},
\end{aligned}$$

one has

$$\|(A+\tau\Sigma_0)^{1/2} - A^{1/2}\| \leqslant \frac{\tau\|\Sigma_0^{1/2}\|}{\lambda_{\min}(\Sigma_0^{1/2})\sqrt{\tau + 2\lambda_{\min}(\Sigma)}} \leqslant \sqrt{\tau}\frac{\lambda_{\max}(\Sigma_0^{1/2})}{\lambda_{\min}(\Sigma_0^{1/2})},$$

so that

$$\left|\operatorname{tr}\left((A+\tau\Sigma_0)^{1/2}-A^{1/2}\right)\right| \leqslant n\sqrt{\tau}\frac{\lambda_{\max}(\Sigma_0^{1/2})}{\lambda_{\min}(\Sigma_0^{1/2})}.$$

Plugging it into (4.9) yields the bound

$$|L(\Sigma_\tau)-L(\Sigma)| \leqslant n\sqrt{\tau}\left(\sqrt{\tau}+\frac{2\lambda_{\max}(\Sigma_0^{1/2})}{\lambda_{\min}(\Sigma_0^{1/2})}\right). \tag{4.10}$$

Second, in the case where Σ has a rank deficiency, the bound found in (4.10) still holds, as it does not depend on Σ. This completes the proof. □

4.2.5 Loss in covariance, function, and parameter spaces

The two parametrizations μ (4.3) and π (4.4) introduced in the previous sections lead to consider the loss (4.5) on three different spaces; namely, the covariance, function, or parameter space.

- Firstly, given a target covariance matrix $\Sigma_0 \in \mathbb{S}_{++}(n)$, we define the loss over covariance matrices as

$$\begin{aligned} L\colon \ \mathbb{S}_+(n) &\longrightarrow \mathbb{R}, \\ \Sigma &\longmapsto L(\Sigma)=\mathcal{B}^2(\Sigma,\Sigma_0)=\operatorname{tr}\left(\Sigma+\Sigma_0-2\left[\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2}\right]^{1/2}\right). \end{aligned} \tag{4.11a}$$

  This loss is convex on $\mathbb{S}_+(n)$.

- Secondly, given $\pi\colon \mathbb{M}(n,m)\to\mathbb{S}_+(n)$ (4.4), we define the loss in the function space, i.e. over end-to-end matrices $W\in\mathbb{M}(n,m)$, as $L^1\colon W\mapsto L\circ\pi(W)$. This is explicitly given by

$$\begin{aligned} L^1\colon \ \mathbb{M}(n,m) &\longrightarrow \mathbb{R}, \\ W &\longmapsto L^1(W)=\operatorname{tr}\left(WW^\top+\Sigma_0-2\left[\Sigma_0^{1/2}WW^\top\Sigma_0^{1/2}\right]^{1/2}\right). \end{aligned} \tag{4.11b}$$

  This loss is *not* convex, which can be seen even in the scalar case.

- Lastly, for a tuple of weight matrices $\theta=(W_1,\dots,W_N)$, we compose L^1 (4.11b) with the deep linear network parametrization map $\mu\colon\Theta\to\mathbb{M}(n,m)$ (4.3) to define the loss in the parameter space as

$$\begin{aligned} L^N\colon \ \Theta &\longrightarrow \mathbb{R}, \\ \theta &\longmapsto L^1\circ\mu(\theta)=L\circ\pi\circ\mu(\theta). \end{aligned} \tag{4.11c}$$

  Observe that this is, again, a non-convex loss.

Thus, for $\theta \in \Theta$, $L^N(\theta) = L^1(\mu(\theta)) = L(\pi(\mu(\theta)) = \mathcal{B}^2(\pi \circ \mu(\theta), \Sigma_0)$. While a typical gradient update lies on the parameter space Θ, viewing the problem in the covariance space is useful since then the objective function is convex, even if it may be subject to non-convex constraints. One of our goals is to translate properties between L, L^1, and L^N.

Smooth perturbative loss As mentioned before, the Bures-Wasserstein loss is non-smooth at covariance matrices with vanishing eigenvalues. As a result, the usual analysis tools to prove uniqueness and convergence of the gradient flow do not apply here. To tackle this issue, (4.7) introduces a smooth perturbative version of the loss. We give here the explicit expression of this regularized loss on the different spaces.

Given a regularization strength $\tau \in (0, \infty)$, consider the perturbation map

$$\begin{array}{rccc} \varphi_\tau \colon & \mathbb{S}_+(n) & \longrightarrow & \mathbb{S}_{++}(n), \\ & \Sigma & \longmapsto & \Sigma + \tau \mathrm{I}_n. \end{array} \tag{4.12}$$

Then, the perturbative loss in the covariance space is defined as

$$\begin{array}{rccc} L_\tau \colon & \mathbb{S}_+(n) & \longrightarrow & \mathbb{R}, \\ & \Sigma & \longmapsto & L \circ \varphi_\tau(\Sigma), \end{array} \tag{4.13a}$$

and similarly, the perturbative loss in the function space is

$$\begin{array}{rccc} L^1_\tau \colon & \mathbb{M}(n,m) & \longrightarrow & \mathbb{R}, \\ & W & \longmapsto & L_\tau \circ \pi(W). \end{array}$$

More explicitly, we have for $W \in \mathbb{M}(n,m)$,

$$L^1_\tau(W) = \operatorname{tr}\Big(WW^\top + \tau \mathrm{I}_n + \Sigma_0 - 2(\Sigma_0^{1/2}(WW^\top + \tau \mathrm{I}_n)\Sigma_0^{1/2})^{1/2}\Big). \tag{4.13b}$$

This function is differentiable and allows us to apply usual convergence arguments for the gradient flow. Likewise,

$$\begin{array}{rccc} L^N_\tau \colon & \Theta & \longrightarrow & \mathbb{R}, \\ & \theta & \longmapsto & L_\tau \circ \pi \circ \mu(\theta) \end{array} \tag{4.13c}$$

is well-defined and of class $\mathscr{C}^2$ on Θ. Moreover, by composition, it is analytic on Θ.

With this, the different domains and parametrizations for a 2-Wasserstein deep linear GAN are fully introduced. We are now ready to provide the expression of the critical points of the Bures-Wasserstein loss.

4.3 Expression of the critical points

In this section, we characterize the critical points of the Bures-Wasserstein loss restricted to matrices of a given rank. The proofs of results in this section are given in Section 4.4.

Let $(n, m) \in \mathbb{N}^2$. For $k \in [\min(n, m)]$, we denote $\mathcal{M}(k)$ the manifold of rank-k matrices of size $n \times m$:

$$\mathcal{M}(k) \equiv \mathcal{M}(k; n, m) := \{W \in \mathbb{M}(n, m) \mid \operatorname{rank} W = k\}. \tag{4.14}$$

Similarly, we denote $\mathcal{M}(\leqslant k) \equiv \mathcal{M}(\leqslant k; n, m)$ the set of $n \times m$ matrices of rank at most k. It is not a smooth (sub)manifold.

The manifold $\mathcal{M}(k)$ is an embedded submanifold of the linear space $(\mathbb{M}(n, m), \langle \cdot, \cdot \rangle_{\mathrm{F}})$, with codimension $(n - k)(m - k)$ [HS95, Proposition 4.5; Bou23, §2.6]. Given a function $f : \mathbb{M}(n, m) \to \mathbb{R}$, its *restriction* on $\mathcal{M}(k)$ is denoted by $f|_{\mathcal{M}(k)} : \mathcal{M}(k) \ni W \mapsto f(W)$. A function f may not be differentiable everywhere on $\mathbb{R}^{n \times m}$ but still have a restriction on $\mathcal{M}(k)$ that is differentiable.

Definition 4.3 Let $\mathcal{M} \subset \mathbb{M}(n, m)$ be a smooth manifold. Let $f : \mathbb{M}(n, m) \to \mathbb{R}$ be any function such that its restriction on $\mathcal{M}$ is differentiable. A point $W \in \mathcal{M}$ is said to be a *critical point of* $f|_{\mathcal{M}}$ if the differential of $f|_{\mathcal{M}}$ at W is the zero function, i.e. $\mathrm{d}f|_{\mathcal{M}}(W) = 0$.

The set of all critical points of $f|_{\mathcal{M}}$ is denoted by $\operatorname{Crit}(f|_{\mathcal{M}})$.

Remark 4.4. Definition 4.3 can be equivalently stated with the gradient of f at $W \in \mathcal{M}$ in lieu of the differential of f at W. We will use either of those notions in the derivations, using the fact that the gradient of $f|_{\mathcal{M}}$ at $W \in \mathcal{M}$, denoted $\nabla f|_{\mathcal{M}}(W)$, is the unique vector on the tangent space $\mathrm{T}_W\mathcal{M}$ such that, for all $X \in \mathrm{T}_W\mathcal{M}$, $\mathrm{d}f|_{\mathcal{M}}(W)[X] = \langle \nabla f|_{\mathcal{M}}(W), X \rangle$.

4.3.1 Critical points of L^1 over $\mathcal{M}(k)$

As noted above, the loss L (4.11a), along with its different parametrizations, is not differentiable over all $\mathbb{S}_+(n)$. More specifically, it is not differentiable whenever the matrix changes rank.

For $n \in \mathbb{N}$ and $k \in [n]$, we denote $\mathcal{J}_k \subset [n]$ any subset of k elements from $[n]$, i.e., with $k = |\mathcal{J}_k|$.

Given a matrix $A \in \mathbb{M}(n, n)$ and a set $\mathcal{J}_k \subseteq [n]$, we denote by $A_{\mathcal{J}_k} \in \mathbb{M}(n, k)$ the sub-matrix of A consisting of the columns with index in $\mathcal{J}_k$. If the matrix A is diagonal, we let $\bar{A}_{\mathcal{J}_k} \in \mathbb{M}(k)$ be the diagonal sub-matrix which extracts the rows *and* columns with index in $\mathcal{J}_k$.

The following result characterizes the critical points of the loss L^1 (4.11b) in function space. It can be regarded as a type of Eckart-Young-Mirsky result for the case of the Bures-Wasserstein loss, to which the connection will be made clearer in Section 4.5.

Theorem 4.5 (Critical points of L^1). *Assume $\Sigma_0 \in \mathbb{S}_{++}(n)$ has n distinct, positive eigenvalues. Let $\Omega\Lambda\Omega^\top$ be a spectral decomposition of Σ_0 (so $\Omega \in \mathbb{O}(n)$), with eigenvalues ordered decreasingly. Let $k \in [\min\{n, m\}]$. A matrix $W^* \in \mathcal{M}(k)$ is a critical point of $L^1|_{\mathcal{M}(k)}$ if and only if $W^* = \Omega_{\mathcal{J}_k}\bar{\Lambda}_{\mathcal{J}_k}^{1/2}V^\top$ for some $\mathcal{J}_k \subseteq [n]$ with $|\mathcal{J}_k| = k$ and $V \in \mathbb{M}(m, k)$ with $V^\top V = \mathrm{I}_k$. Moreover, the minimum over $\mathcal{M}(\leqslant k)$ is attained precisely when $\mathcal{J}_k = [k]$. In particular, $\inf\{L^1(W) \mid W \in \mathcal{M}(k)\} = \min\{L^1(W) \mid W \in \mathcal{M}(k)\}$ and $\min\{L^1(W) \mid W \in \mathcal{M}(k)\} = \min\{L^1(W) \mid W \in \mathcal{M}(\leqslant k)\}$.*

Remark 4.6. Notice that the number of critical points is in reality infinite, and the $\binom{n}{k}$ critical points choices are only up to right multiplication by an arbitrary orthonormal matrix (the trivial symmetry of $W \mapsto WW^\top$).

The proof relies on evaluating the zeros of the gradient $\nabla L^1|_{\mathcal{M}(k)}$ (see Lemma 4.15). Then evaluating the loss at these critical points allows us to identify which of them attain the minimum.

Remark 4.7. Interestingly, the critical points and the minimizer of L^1 characterized in the above result agree with those of the squared error loss [EY36; Mir60]. Nonetheless, we observe that (4.5) is only defined for positive semi-definite matrices. Hence the notion of unitary invariance considered by [Mir60] only makes sense for left and right multiplication by the same matrix. Moreover, $\mathcal{B}$ (4.5) is not a norm in the sense that there is no function $B\colon \mathbb{S}_+(n) \to \mathbb{R}$ such that $\mathcal{B}(\Sigma, \Sigma_0) = B(\Sigma - \Sigma_0)$, and hence it does not fall into the framework of [Mir60]. We offer more details about this in Section 4.5.

Remark 4.8. Under the Burer-Monteiro parametrization π (4.29), a point in the function space $W \in \mathbb{M}(n, m)$, modulo the right multiplication by an orthogonal matrix, is a critical point for L^1 if and only if the associated covariance matrix $\pi(W) = WW^\top$ is a critical point for L.

4.3.2 Critical points of the perturbative loss

For the critical points of the perturbative loss L^1_τ (4.13b), we obtain the following results.

Theorem 4.9 (Critical points of L^1_τ). *Assume Σ_0 has n distinct, positive eigenvalues. Let $\Omega\Lambda\Omega^\top$ be a spectral decomposition of Σ_0, with eigenvalues ordered decreasingly. A point $W^* \in \mathcal{M}(k)$ is a critical point of $L^1_\tau|_{\mathcal{M}(k)}$ if and only if $W^* = \Omega_{\mathcal{J}_k}(\bar{\Lambda}_{\mathcal{J}_k} - \tau\mathrm{I}_k)^{1/2}V^\top$ for*

some $\mathcal{J}_k \subseteq [n]$ with $|\mathcal{J}_k| = k$ and $V \in \mathbb{M}(n,k)$ with $V^\top V = \mathrm{I}_k$. Moreover, the value at such a point is $L_\tau^1(W^*) = \sum_{j \notin \mathcal{J}_k} (\sqrt{\lambda_j} - \sqrt{\tau})^2$. The minimum over $\mathcal{M}(\leqslant k)$ is therefore attained precisely when $\mathcal{J}_k = [k]$. In particular, $\inf\left\{L_\tau^1(W) \mid W \in \mathcal{M}(k)\right\} = \min\left\{L_\tau^1(W) \colon W \in \mathcal{M}(k)\right\}$ *and* $\min\left\{L_\tau^1(W) \mid W \in \mathcal{M}(k)\right\} = \min\left\{L_\tau^1(W) \mid W \in \mathcal{M}(\leqslant k)\right\}$.

Remark 4.10. The above characterization of the critical points imposes an upper bound on τ if we require to have all $\binom{n}{k}$ different critical points for L_τ when $\Sigma_0 \in \mathbb{S}_{++}(n)$. Indeed, for a given Σ^* to be a critical point of L_τ, according to Theorem 4.9, one must have $\bar{\Lambda}_{\mathcal{J}_k} - \tau \mathrm{I}_k \succcurlyeq 0$ for all $\mathcal{J}_k \subseteq [n]$ with $|\mathcal{J}_k| = k$. Therefore, it imposes that $\tau \in [0, \lambda_n)$, where $\lambda_n = \min\{\lambda_j \mid j \in [n]\}$.

In order to link the critical points in the parameter space Θ to the critical points in the function space $\mathbb{M}(n,m)$, we appeal to the correspondence drawn by [TKB20, Propositions 6 and 7]. For the Bures-Wasserstein loss, this allows to conclude the following.

Proposition 4.11 (Critical points in parameter space are critical points in function space). *Assume a full-rank target Σ_0 with spectral decomposition $\Sigma_0 = \Omega\Lambda\Omega^\top$ and distinct eigenvalues $\lambda_1 > \cdots > \lambda_n > 0$ ordered decreasingly. Let $\tau \in (0, \lambda_n]$. If $\theta^* \in \mathrm{Crit}(L_\tau^N)$, then $W^* = \mu(\theta^*)$ is a critical point of the loss $L_\tau^1|_{\mathcal{M}(k)}$, where $k = \operatorname{rank} W^*$. Moreover, when $k = \underline{d} = \min_{i \in [N]}\{d_i\}$, θ is a local minimizer of the loss L_τ^N if and only if $W^* = \mu(\theta^*)$ is a local minimizer, and therefore the global minimizer, of $L_\tau^1|_{\mathcal{M}(\underline{d})}$. In this case, $\Sigma_\tau^* = W^* W^{*\top} + \tau \mathrm{I}_n$ is the τ-best $\underline{d}$-rank approximation of the target in the covariance space, in the sense that $\Sigma_\tau^* = \Omega \begin{pmatrix} \Lambda_{[\underline{d}]} & \\ & \tau \end{pmatrix} \Omega^\top$.*

Proposition 4.11 ensures that, under the assumption that the solution of the gradient flow is a (local) minimizer in the parameter space and has the highest possible rank $\underline{d}$ for the given network architecture, the solution in the covariance space is the best $\underline{d}$-rank approximation of the target in the sense of the τ-smoothed Bures-Wasserstein distance. The fact that any local minimizer of $L_\tau^1|_{\mathcal{M}(\underline{d})}$ is indeed a global minimizer is not immediate (since neither the loss L_τ^1 nor the set $\mathcal{M}(\underline{d})$ are convex), but can be shown as we do in Lemma 4.22.

From the fact that L_τ^1 satisfies the strict saddle point property Lemma 4.22, we further cast [BRT+21, Proposition 6.9] in our setting.

Proposition 4.12 *Let $\overrightarrow{W} = (W_1, \ldots, W_N)$ be a critical point of L_τ^N such that $W = W_{N:1}$ has $\operatorname{rank}(W) = k$. If W is not a global optimum of L_τ^1 on $\mathcal{M}_k$, then $\overrightarrow{W}$ is a strict saddle point of L_τ^N.*

The proof of Proposition 4.12 is given in Section 4.4.2.

4.4 Proofs of section 4.3

In this section, we provide the proofs of the statements about the critical points of the loss functions in function space, $L^1|_{\mathcal{M}(k)}$ and $L^1_\tau|_{\mathcal{M}(k)}$. We characterize the critical points and show that all saddles are strict.

4.4.1 Critical points of $L^1|_{\mathcal{M}(k)}$

First, the loss L^1 is expressed on the manifolds $\mathcal{M}(k)$ (Lemma 4.14), where it is differentiable (Lemma 4.15). Then, necessary conditions (Lemma 4.17) on the critical points can be expressed, leading to the first part of Theorem 4.5. The second part of Theorem 4.5 is then proven by evaluating the loss at the critical points found, and ranking them.

Recall Definition 4.3 of the critical points of a function restricted to a manifold. Computing the differential of the restriction $L^1|_{\mathcal{M}(k)}$ will allow to characterize the different critical points.

Definition 4.13 (*Gradient*). Given an embedded manifold $\mathcal{M}$ and a function f with a differentiable restriction $f|_{\mathcal{M}}$, the gradient of $f|_{\mathcal{M}}$ at $x \in \mathcal{M}$ is the (unique) element of the tangent space $\mathrm{T}_x\mathcal{M}$ such that, for all $v \in \mathrm{T}_x\mathcal{M}$, $\mathrm{d}f|_{\mathcal{M}}(x)[v] = \langle \nabla f|_{\mathcal{M}}(x), v\rangle$.

We begin by expressing the loss $L^1|_{\mathcal{M}(k)}$ with the Singular Value Decomposition (SVD) of $\Sigma_0^{1/2}W$.

Lemma 4.14 *Let $USV^\top = \Sigma_0^{1/2}W$ be a thin SVD of $\Sigma_0^{1/2}W$, so that $U \in \mathbb{M}(n,k)$, $V \in \mathbb{M}(m,k)$, $U^\top U = V^\top V = \mathrm{I}_k$, $S = \mathrm{Diag}(s_1, \dots, s_k) \in \mathbb{M}(k)$, where $k = \operatorname{rank}\Sigma_0^{1/2}W = \operatorname{rank} W$. The loss L^1 from* (4.11b) *on $\mathcal{M}(k)$ can be expressed as*

$$L^1|_{\mathcal{M}(k)}(W) = \|W\|_{\mathrm{F}}^2 + \|\Sigma_0^{1/2}\|_{\mathrm{F}}^2 - 2\operatorname{tr} S. \tag{4.15}$$

Proof. If $USV^\top = \Sigma_0^{1/2}W$ is a thin SVD of $\Sigma_0^{1/2}W$, then $\left(\Sigma_0^{1/2}W(\Sigma_0^{1/2}W)^\top\right)^{1/2} = USU^\top$. Therefore, the expression of the loss L^1 given by (4.11b) can be written as

$$L^1|_{\mathcal{M}(k)}(W) = \operatorname{tr} WW^\top + \operatorname{tr}\Sigma_0 - 2\operatorname{tr} USU^\top = \|W\|_{\mathrm{F}}^2 + \|\Sigma_0^{1/2}\|_{\mathrm{F}}^2 - 2\operatorname{tr} S,$$

as claimed. □

With this description at hand, we now give the gradient of $L^1|_{\mathcal{M}(k)}$.

Lemma 4.15 (Gradient of $L^1|_{\mathcal{M}(k)}$). *Let $(n, m) \in \mathbb{N}^2$, and let $k \in [\min\{n, m\}]$. The loss $L^1|_{\mathcal{M}(k)}$ (as given by (4.15)) is twice continuously differentiable on $\mathcal{M}(k)$. With $W \in \mathcal{M}(k)$ and $USV^\top = \Sigma_0^{1/2} W$ a thin SVD of $\Sigma_0^{1/2} W$, its gradient is*

$$\nabla L^1|_{\mathcal{M}(k)}(W) = 2W - 2\Sigma_0^{1/2} UV^\top. \tag{4.16}$$

In order to prove Lemma 4.15, we need the differential expression for the singular values appearing in the SVD of a matrix.

Lemma 4.16 (Differential of the SVD). *Let $k \in [\min\{n, m\}]$ and let $X \in \mathcal{M}(k)$ be a matrix with $\operatorname{rank} X = k$. Let $USV^\top = X$ be a thin SVD of X, with $U \in \mathbb{M}(n, k)$, $S \in \mathbb{M}(k)$, $V \in \mathbb{M}(m, k)$, S diagonal and $U^\top U = V^\top V = \mathrm{I}_k$. Then, the differential $\mathrm{d}S$ is*

$$\mathrm{d}S = \mathrm{I}_k \odot (U^\top \mathrm{d}X V),$$

where $A \odot B$ denotes the Hadamard product between A and B.

Proof. Let $USV^\top = X$ be a decomposition as given in the lemma statement. The differential rules ensure that

$$\mathrm{d}X = \mathrm{d}U SV^\top + U\,\mathrm{d}S V^\top + US\,\mathrm{d}V^\top.$$

This implies that

$$\begin{aligned} U^\top \mathrm{d}X V &= U^\top \mathrm{d}U SV^\top V + U^\top U\,\mathrm{d}S V^\top V + U^\top US\,\mathrm{d}V^\top V \\ &= U^\top \mathrm{d}U S + \mathrm{d}S + S\,\mathrm{d}V^\top V \\ \implies \qquad \mathrm{d}S &= U^\top \mathrm{d}X V - U^\top \mathrm{d}U S - S\,\mathrm{d}V^\top V. \end{aligned}$$

Since $U^\top U = \mathrm{I}_k$, $(\mathrm{d}U)^\top U + U^\top(\mathrm{d}U) = 0$, and $A := U^\top \mathrm{d}U = -(\mathrm{d}U)^\top U = -A^\top$. Likewise, $B := V^\top \mathrm{d}V$ is also antisymmetric. The matrices A and B being antisymmetric, their diagonals are null; hence so are the diagonals of AS and SB, i.e. $\mathrm{I}_k \odot (AS) = \mathrm{I}_k \odot (SB) = 0$. Since S is constrained to be diagonal, $\mathrm{d}S$ must also be diagonal, i.e. $\mathrm{I}_k \odot \mathrm{d}S = \mathrm{d}S$. Therefore,

$$\mathrm{d}S = \mathrm{I}_k \odot (U^\top \mathrm{d}X V)$$

as was claimed. □

Now that the differential of the singular values is available, we are ready to prove Lemma 4.15.

Proof of Lemma 4.15. For $W \in \mathcal{M}(k)$, let $USV^\top = \Sigma_0^{1/2} W$ be a thin SVD of $\Sigma_0^{1/2} W =: X$. Lemma 4.14 ensures that

$$L^1|_{\mathcal{M}(k)}(W) = \|W\|_{\mathrm{F}}^2 + \|\Sigma_0^{1/2}\|_{\mathrm{F}}^2 - 2 \operatorname{tr} S. \tag{4.17}$$

According to Lemma 4.16, the matrix S is differentiable and has differential $\mathrm{d}S = \mathrm{I}_k \odot (U^\top \mathrm{d}X V)$. Therefore, the loss $L^1|_{\mathcal{M}(k)}$ is differentiable. With the fact that $\mathrm{d} \operatorname{tr} S = \operatorname{tr} \mathrm{d}S$ [see, e.g. MN19, Chap. 8, Eq. 18], we can compute

$$\begin{aligned}\mathrm{d} \operatorname{tr} S &= \operatorname{tr} \mathrm{d}S = \operatorname{tr} (\mathrm{I}_k \odot (U^\top \mathrm{d}X V)) = \operatorname{tr} (U^\top \mathrm{d}X V) \\ &= \langle UV^\top, \mathrm{d}X \rangle = \langle UV^\top, \Sigma_0^{1/2} \mathrm{d}W \rangle = \langle \Sigma_0^{1/2} UV^\top, \mathrm{d}W \rangle.\end{aligned}$$

Moreover, $\mathrm{d}\|W\|_{\mathrm{F}}^2 = 2\langle W, \mathrm{d}W \rangle$, and so

$$\mathrm{d}L^1|_{\mathcal{M}(k)}(W) = \mathrm{d}\|W\|_{\mathrm{F}}^2 - 2\,\mathrm{d}\operatorname{tr} S = 2\langle W - \Sigma_0^{1/2} UV^\top, \mathrm{d}W \rangle,$$

and

$$\nabla L^1|_{\mathcal{M}(k)}(W) = 2(W - \Sigma_0^{1/2} UV^\top).$$

Since the matrices U and V are continuously differentiable on $\mathcal{M}(k)$, $\nabla L^1|_{\mathcal{M}(k)}(W) = 2(W - \Sigma_0^{1/2} UV^\top)$ is again continuously differentiable, and $L^1|_{\mathcal{M}(k)}$ is twice continuously differentiable. □

We are now ready to give the proof of Theorem 4.5. We divide the proof into necessary and sufficient conditions for a point to be a critical point of $L^1|_{\mathcal{M}(k)}$.

Lemma 4.17 (Necessary condition on the critical points of $L^1|_{\mathcal{M}(k)}$). *Assume Σ_0 has n distinct eigenvalues. Let $W^* \in \mathcal{M}(k)$ be a critical point of $L^1|_{\mathcal{M}(k)}$. Then, with $U^* S^* V^{*\top} = \Sigma_0^{1/2} W^*$ a thin SVD of $\Sigma_0^{1/2} W^*$, and $\Omega \Lambda \Omega^\top = \Sigma_0$ a spectral decomposition of Σ_0 (i.e. with $\Omega \in \mathbb{O}(n)$), there exists $\mathcal{J}_k \subseteq [n]$, such that $S^* = \bar{\Lambda}_{\mathcal{J}_k}$ and $U^* = \Omega_{\mathcal{J}_k}$.*

Proof. Let W^* be as in the lemma statement. Since $W^* \in \mathcal{M}(k)$, and $U^* S^* V^{*\top} = \Sigma_0^{1/2} W^*$ is a thin SVD of $\Sigma_0^{1/2} W^*$, this means that $S^* \in \mathbb{M}(k)$. Then,

$$\begin{aligned}\nabla L^1(W^*) = 0 \implies & W^* = \Sigma_0^{1/2} U^* V^{*\top}, && \text{by (4.16)} \\ \implies & \Sigma_0^{1/2} W^* = \Sigma_0 U^* V^{*\top} \\ \implies & U^* S^* V^{*\top} = \Sigma_0 U^* V^{*\top} \\ \implies & S^* = U^{*\top} \Sigma_0 U^*, && U^{*\top} U^* = \mathrm{I}_k,\ V^{*\top} V^* = \mathrm{I}_k.\end{aligned}$$

Therefore, $U^{*\top} \Sigma_0 U^*$ must be diagonal; and since U^* is semi-orthogonal, this is the case if and only if the vectors in U^* are eigenvectors for Σ_0, by uniqueness of the spectral

decomposition of Σ_0. Therefore, there exist $j_1, \dots, j_k$ indices between 1 and n such that $U^* = \begin{pmatrix} \omega_{j_1} & \cdots & \omega_{j_k} \end{pmatrix} = \Omega_{\mathcal{J}_k}$, in which case

$$S^* = {\Omega_{\mathcal{J}_k}}^\top \Sigma_0 \Omega_{\mathcal{J}_k} = \begin{pmatrix} \lambda_{j_1} & & \\ & \ddots & \\ & & \lambda_{j_k} \end{pmatrix} = \bar{\Lambda}_{\mathcal{J}_k}.$$

□

Now we are ready to prove the first part of Theorem 4.5.

Proof of Theorem 4.5, first part. Consider the expression for the gradient of $L^1|_{\mathcal{M}(k)}$ given in (4.16). The necessary condition follows from Lemma 4.17, since

$$\begin{aligned} \nabla L^1|_{\mathcal{M}(k)}(W^*) = 0 \implies \Sigma_0^{1/2} W^* &= \Omega_{\mathcal{J}_k} \bar{\Lambda}_{\mathcal{J}_k} V^\top \\ \implies \qquad W^* &= \Sigma_0^{-1/2} \Omega_{\mathcal{J}_k} \bar{\Lambda}_{\mathcal{J}_k} V^\top \\ &= \Omega \Lambda^{-1/2} \Omega^\top \Omega_{\mathcal{J}_k} \bar{\Lambda}_{\mathcal{J}_k} V^\top \\ &= \Omega \Lambda^{-1/2} \Lambda_{\mathcal{J}_k} V^\top \\ &= \Omega \Lambda^{1/2} P_{\mathcal{J}_k} V^\top \\ &= \Omega P_{\mathcal{J}_k} \bar{\Lambda}_{\mathcal{J}_k}^{1/2} V^\top \\ &= \Omega_{\mathcal{J}_k} \bar{\Lambda}_{\mathcal{J}_k}^{1/2} V^\top, \end{aligned}$$

which corresponds to the necessary condition in Theorem 4.5.

The sufficient condition can be verified as follows. With $W^* = \Omega_{\mathcal{J}_k} \bar{\Lambda}_{\mathcal{J}_k}^{1/2} V^\top$, one has $\Sigma_0^{1/2} W^* = \Omega \Lambda^{1/2} \Omega^\top \Omega_{\mathcal{J}_k} \bar{\Lambda}_{\mathcal{J}_k}^{1/2} V^\top = \Omega_{\mathcal{J}_k} \bar{\Lambda}_{\mathcal{J}_k} V^\top$, and, as this is a correct thin SVD of $\Sigma_0^{1/2} W^*$, Lemma 4.15 gives

$$\nabla L^1|_{\mathcal{M}(k)}(W^*) = 2(W^* - \Sigma_0^{1/2} \Omega_{\mathcal{J}_k} V^\top).$$

Further,

$$\Sigma_0^{1/2} \Omega_{\mathcal{J}_k} = \Omega \Lambda^{1/2} \Omega^\top \Omega_{\mathcal{J}_k} = \Omega \Lambda^{1/2} P_{\mathcal{J}_k} = \Omega P_{\mathcal{J}_k} \bar{\Lambda}_{\mathcal{J}_k}^{1/2} = \Omega_{\mathcal{J}_k} \bar{\Lambda}_{\mathcal{J}_k}^{1/2}.$$

Hence

$$\nabla L^1|_{\mathcal{M}(k)}(W^*) = 2(W^* - \Sigma_0^{1/2} \Omega_{\mathcal{J}_k} V^\top) = 2(\Omega_{\mathcal{J}_k} \bar{\Lambda}_{\mathcal{J}_k}^{1/2} V^\top - \Omega_{\mathcal{J}_k} \bar{\Lambda}_{\mathcal{J}_k}^{1/2} V^\top) = 0,$$

and the sufficient condition is verified. □

Now, the loss can be evaluated at the critical points in order to identify its minimizers.

Corollary 4.18 (Value of L^1 at the critical points). *The value of the loss L^1 at a critical point $W^* = \Omega_{\mathcal{J}_k} \bar{\Lambda}_{\mathcal{J}_k}^{1/2} V^\top \in \mathcal{M}(k)$ is $L^1(W^*) = \operatorname{tr} \Lambda - \operatorname{tr} \bar{\Lambda}_{\mathcal{J}_k} = \sum_{i \notin \mathcal{J}_k} \lambda_i$.*

Proof. Let $k \in [\min\{n, m\}]$, and let W^* be a critical point of $L^1|_{\mathcal{M}(k)}$. From Theorem 4.5, with $\Sigma_0 = \Omega \Lambda \Omega^\top$ a spectral decomposition of Σ_0, there exists a set $\mathcal{J}_k$ and a semi-orthogonal matrix $V \in \mathbb{M}(n, k)$ such that $W^* = \Omega_{\mathcal{J}_k} \bar{\Lambda}_{\mathcal{J}_k}^{1/2} V^\top$. One can then compute the value of the loss at W^*:

$$\begin{aligned} L^1(W^*) &= \operatorname{tr} W^* W^{*\top} + \operatorname{tr} \Sigma_0 - 2 \operatorname{tr} \left(\left[\left(\Sigma_0^{1/2} W^* \right) \left(\Sigma_0^{1/2} W^* \right)^\top \right]^{1/2} \right) \\ &= \operatorname{tr} \Omega_{\mathcal{J}_k} \bar{\Lambda}_{\mathcal{J}_k} \Omega_{\mathcal{J}_k} + \operatorname{tr} \Lambda - 2 \operatorname{tr} \left(\Omega_{\mathcal{J}_k} \bar{\Lambda}_{\mathcal{J}_k}^2 \Omega_{\mathcal{J}_k}^\top \right)^{1/2} \\ &= \operatorname{tr} \bar{\Lambda}_{\mathcal{J}_k} + \operatorname{tr} \Lambda - 2 \operatorname{tr} \bar{\Lambda}_{\mathcal{J}_k} \\ &= \operatorname{tr} \Lambda - \operatorname{tr} \bar{\Lambda}_{\mathcal{J}_k}. \end{aligned}$$

□

We now have all ingredients needed to prove the second part of Theorem 4.5.

Proof of Theorem 4.5, second part. The first part of the statement is readily implied by Corollary 4.18, as the eigenvalues are in decreasing order. The second part is implied by the fact that the minimum $L^1|_{\mathcal{M}(k)}$ is indeed achieved for any $k \in [n]$ (by selecting the k largest eigenvalues of Σ_0) and the optimal value of the loss L_k^* is smaller when considering more eigenvalues, i.e. $\min_{\mathcal{M}(k)} L^1 \leqslant \min_{\mathcal{M}(<k)} L^1$. □

Next we show that only one point per set $\mathcal{M}(k)$ is a minimizer of the loss $L^1|_{\mathcal{M}(k)}$ and all other points are (strict) saddle points. We recall the definition of a strict saddle point: a point where there exists a descent direction.

Definition 4.19 (*Strict saddle point*). A critical point x of a function f is said to be a *strict saddle point* if the Hessian of f at x has a strict negative eigenvalue. If all critical points of f are either a strict saddle point or a global minimizer, the we say that f satisfies the *strict saddle point property*.

If the gradient flow can be expressed on a manifold, with a Riemannian gradient corresponding to a given metric, there is an equivalent definition of those saddle points, which will be handy to use. We refer to [BRT+21, §6.1] for the details.

Proposition 4.20 *The loss $L^1|_{\mathcal{M}(k)}$ satisfies the strict saddle point property.*

Proof. Let $\Sigma_0 = \Omega\Lambda\Omega^\top$ be the spectral decomposition of Σ_0 with decreasing eigenvalues. For $k \in \mathbb{N}$, according to Theorem 4.5, W^* is a critical point of $L^1|_{\mathcal{M}(k)}$ if and only if there exists $\mathcal{J}_k \subset [n]$, such that $W^* = \Omega_{\mathcal{J}_k}\Lambda_{\mathcal{J}_k}^{1/2}V^\top$, with any $V \in \mathbb{M}(m,k)$ so that $V^\top V = \mathrm{I}_k$. If $\mathcal{J}_k = [k]$, W^* is a global minimum of $L^1|_{\mathcal{M}(k)}$, as shown in Corollary 4.18, and the proposition holds.

Assume $\mathcal{J}_k \neq [k]$, then there exists $j_0 \in \mathcal{J}_k$ such that $\lambda_{j_0} < \lambda_k$, and there exists $j_1 \notin \mathcal{J}_k$ but $j_1 \in [k]$ such that $\lambda_{j_1} > \lambda_{j_0}$. We will show that W^* is a strict saddle point of $L^1|_{\mathcal{M}(k)}$.

The critical point W^* can equivalently be expressed as

$$W^* = \Sigma_0^{-1/2} \sum_{j \in \mathcal{J}_k} \lambda_j \omega_j v_j^\top, \tag{4.18}$$

where ω_j, v_j are corresponding orthonormal vectors in Ω and V, and λ_j are eigenvalues in Λ.

For $t \in (-1,1)$, we define

$$\omega_{j_0}(t) = t\omega_{j_1} + \sqrt{1-t^2}\omega_{j_0}$$

and the curve $\gamma\colon (-1,1) \mapsto \mathcal{M}(k)$. We look at the perturbed matrix

$$\gamma(t) = \Sigma_0^{-1/2}\Big(\lambda_{j_0}\omega_{j_0}(t)v_{j_0}^\top + \sum_{j \in \mathcal{J}\setminus\{j_0\}} \lambda_j\omega_j v_j^\top\Big).$$

Note that $\gamma(0) = W$. Recall $L^1(W) = \operatorname{tr}\big(WW^\top + \Sigma_0 - 2(\Sigma_0^{1/2}WW^\top\Sigma_0^{1/2})^{1/2}\big)$. By [BRT+21, §6.1], it is enough to show that

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

We check it term by term,

$$\begin{aligned}
\operatorname{tr}\Big(\gamma(t)\gamma(t)^\top\Big) &= \operatorname{tr}\Big(\Sigma_0^{-1/2}\big(\lambda_{j_0}\omega_{j_0}(t)v_{j_0}^\top + \sum_{j\in\mathcal{J}\setminus\{j_0\}}\lambda_j\omega_j v_j^\top\big)\cdot \\
&\qquad \big(\lambda_{j_0}\omega_{j_0}(t)v_{j_0}^\top + \sum_{j\in\mathcal{J}\setminus\{j_0\}}\lambda_j\omega_j v_j^\top\big)^\top\Sigma_0^{-1/2}\Big) \\
&= \operatorname{tr}\Big(\Sigma_0^{-1}\big(\lambda_{j_0}^2\omega_{j_0}(t)\omega_{j_0}(t)^\top + \sum_{j\in\mathcal{J}\setminus\{j_0\}}\lambda_j^2\omega_j\omega_j^\top\big)\Big) \\
&= \operatorname{tr}\Big(\big(\sum_{1\leqslant j\leqslant n}\lambda_j^{-1}\omega_j\omega_j^\top\big)\big(\lambda_{j_0}^2\omega_{j_0}(t)\omega_{j_0}(t)^\top + \sum_{j\in\mathcal{J}\setminus\{j_0\}}\lambda_j^2\omega_j\omega_j^\top\big)\Big) \\
&= \frac{\lambda_{j_0}^2}{\lambda_{j_1}}t^2 + \lambda_{j_0}(1-t^2) + \sum_{j\in\mathcal{J}\setminus\{j_0\}}\lambda_j^2,
\end{aligned}$$

and

$$\begin{aligned}
&\operatorname{tr}\Big(\big(\Sigma_0^{1/2}\gamma(t)\gamma^\top(t)\Sigma_0^{1/2}\big)^{1/2}\Big)\\
&\quad= \operatorname{tr}\Big(\big((\lambda_{j_0}\omega_{j_0}(t)v_{j_0}^\top + \sum_{j\in\mathcal{J}\setminus\{j_0\}}\lambda_j\omega_j v_j^\top\big)\big(\lambda_{j_0}\omega_{j_0}(t)v_{j_0}^\top + \sum_{j\in\mathcal{J}\setminus\{j_0\}}\lambda_j\omega_j v_j^\top\big)^\top\big)^{1/2}\Big)\\
&\quad= \operatorname{tr}\Big(\big(\lambda_{j_0}^2\omega_{j_0}(t)\omega_{j_0}(t)^\top + \sum_{j\in\mathcal{J}\setminus\{j_0\}}\lambda_j^2\omega_j\omega_j^\top\big)^{1/2}\Big)\\
&\quad= \operatorname{tr}\Big(\big(t^2\lambda_{j_0}^2\omega_{j_1}\omega_{j_1}^\top + (1-t^2)\lambda_{j_0}^2 u_{j_0}\omega_{j_0}^\top + \sum_{j\in\mathcal{J}\setminus\{j_0\}}\lambda_j^2\omega_j\omega_j^\top\big)^{1/2}\Big)\\
&\quad= t|\lambda_{j_0}| + \sqrt{1-t^2}|\lambda_{j_0}| + \sum_{j\in\mathcal{J}\setminus\{j_0\}}|\lambda_j|.
\end{aligned}$$

Thus, since $\lambda_{j_1} > \lambda_{j_0}$,

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}L^1(\gamma(t))\Big|_{t=0} = 2(\lambda_{j_0}^2\lambda_{j_1}^{-1} - \lambda_{j_0}) - |\lambda_{j_0}| < 0.$$

This completes the proof. □

4.4.2 Critical points of the perturbative loss $L_\tau^1|_{\mathcal{M}_k}$

In this section, we provide the derivations for Section 4.3.2. The structure of reasoning is similar to the one found in the proof of Theorem 4.5: first the gradient of L_τ^1 is computed, then the critical points are characterized and ordered.

Lemma 4.21 (Gradient of L_τ^1). *The loss L_τ^1 has the following gradient*

$$\forall W \in \mathbb{M}(n,m), \quad \nabla L_\tau^1(W) = 2\big(W - \Sigma_0^{1/2}\big[\Sigma_0^{1/2}(WW^\top + \tau \mathrm{I}_n)\Sigma_0^{1/2}\big]^{-1/2}\Sigma_0^{1/2}W\big). \tag{4.19}$$

Proof. This results comes from the chain rule for the loss $L_\tau^1(W) = L\circ\varphi_\tau\circ\pi(W)$. With $\Sigma = \pi(W) = WW^\top$ and $\Sigma_\tau = \varphi_\tau(\Sigma) = \Sigma + \tau\mathrm{I}_n$, and since $\mathrm{d}\pi(W)[Z] = WZ^\top + ZW^\top$ and $\mathrm{d}\varphi_\tau(\Sigma) = \mathrm{id}$, one has

$$\begin{aligned}
\mathrm{d}L_\tau^1(W)[Z] &= \mathrm{d}(L\circ\varphi_\tau\circ\pi)(W)[Z]\\
&= \mathrm{d}L(\Sigma_\tau)\Big[\mathrm{d}\varphi_\tau(\Sigma)\Big[\mathrm{d}\pi(W)[Z]\Big]\Big]\\
&= \mathrm{d}L(\Sigma_\tau)[WZ^\top + ZW^\top]\\
\langle\nabla L_\tau^1(W), Z\rangle &= \langle\nabla L(\Sigma_\tau), WZ^\top + ZW^\top\rangle\\
\iff \qquad \nabla L_\tau^1(W) &= (\nabla L(\Sigma_\tau) + \nabla L(\Sigma_\tau)^\top)W\\
&= 2(W - \Sigma_0^{1/2}[\Sigma_0^{1/2}\Sigma_\tau\Sigma_0^{1/2}]^{-1/2}\Sigma_0^{1/2}W).
\end{aligned}$$

□

With the expression of the gradient of L^1_τ available, Theorem 4.9 can be proven.

Proof of Theorem 4.9. The eigenvectors of $WW^\top + \tau$ are the same as the one of $WW^\top$, and their corresponding eigenvalues are shifted by τ. Therefore, the expression of the critical points in the original loss can be adapted, so that the modified critical points have the same left singular vectors and shifted singular values. This leads to having $W^* = \Omega_{\mathcal{J}_k}(\bar{\Lambda}_{\mathcal{J}_k} - \tau \mathrm{I}_k)^{1/2} V_\parallel^\top = \begin{pmatrix}\Omega_{\mathcal{J}_k} & \mathbf{0}_{n\times n-k}\end{pmatrix} \begin{pmatrix}(\bar{\Lambda}_{\mathcal{J}_k} - \tau \mathrm{I}_k)^{1/2} & \\ & \mathbf{0}_{n-k\times m-k}\end{pmatrix} \begin{pmatrix}V_\parallel & V_\perp\end{pmatrix}^\top$, with $V = \begin{pmatrix}V_\parallel & V_\perp\end{pmatrix} \in \mathbb{R}^{m\times m}$, such such that $V^\top V = VV^\top = \mathrm{I}_m$. In the following, we will make sure that $\nabla L^1_\tau(W^*) = 0$.

Indeed, assume without loss of generality that $\Omega = \begin{pmatrix}\Omega_{\mathcal{J}_k} & \Omega_{\mathcal{J}_k{}^c}\end{pmatrix}$ (and $\Lambda = \begin{pmatrix}\bar{\Lambda}_{\mathcal{J}_k} & \\ & \bar{\Lambda}_{\mathcal{J}_k{}^c}\end{pmatrix}$), where $\mathcal{J}_k{}^c := [n] \setminus \mathcal{J}_k$ for $\mathcal{J}_k \subseteq [n]$. Then,

$$
\begin{aligned}
W^*W^{*\top} &= \Omega_{\mathcal{J}_k}(\bar{\Lambda}_{\mathcal{J}_k} - \tau \mathrm{I}_k)\Omega_{\mathcal{J}_k}^\top \\
&= \Omega \begin{pmatrix}\bar{\Lambda}_{\mathcal{J}_k} - \tau \mathrm{I}_k & \\ & \mathbf{0}_{n-k\times n-k}\end{pmatrix} \Omega^\top, \\
\Sigma^*_\tau := W^*W^{*\top} + \tau \mathrm{I}_n &= W^*W^{*\top} + \tau \Omega\Omega^\top \\
&= \Omega \begin{pmatrix}\bar{\Lambda}_{\mathcal{J}_k} & \\ & \tau \mathrm{I}_{n-k}\end{pmatrix} \Omega^\top, \\
\Sigma_0^{1/2}\Sigma^*_\tau\Sigma_0^{1/2} &= \Omega\Lambda^{1/2}\Omega^\top\Omega \begin{pmatrix}\bar{\Lambda}_{\mathcal{J}_k} & \\ & \tau \mathrm{I}_{n-k}\end{pmatrix} \Omega^\top\Omega\Lambda^{1/2}\Omega^\top \\
&= \Omega \begin{pmatrix}\bar{\Lambda}^2_{\mathcal{J}_k} & \\ & \tau\bar{\Lambda}_{\mathcal{J}_k{}^c}\end{pmatrix} \Omega^\top, \\
\Sigma_0^{1/2}(\Sigma_0^{1/2}\Sigma^*_\tau\Sigma_0^{1/2})^{-1/2}\Sigma_0^{1/2} &= \Omega\Lambda^{1/2}\Omega^\top\Omega \begin{pmatrix}\bar{\Lambda}^{-1}_{\mathcal{J}_k} & \\ & (\tau\bar{\Lambda}_{\mathcal{J}_k{}^c})^{-1/2}\end{pmatrix} \Omega^\top\Omega\Lambda^{1/2}\Omega^\top \\
&= \Omega \begin{pmatrix}\mathrm{I}_k & \\ & \tau^{-1/2}\bar{\Lambda}^{1/2}_{\mathcal{J}_k{}^c}\end{pmatrix} \Omega^\top,
\end{aligned}
$$

and

$$
\mathrm{I}_n - \Sigma_0^{1/2}(\Sigma_0^{1/2}\Sigma^*_\tau\Sigma_0^{1/2})^{-1/2}\Sigma_0^{1/2} = \Omega \begin{pmatrix}\mathbf{0}_{k\times k} & \\ & \mathrm{I}_{n-k} - \tau^{-1/2}\bar{\Lambda}^{1/2}_{\mathcal{J}_k{}^c}\end{pmatrix} \Omega^\top.
$$

Since $\Omega^\top \Omega_{\mathcal{J}_k} = \mathrm{I}_{\mathcal{J}_k} = \begin{pmatrix} A \\ \mathbf{0}_{n-k\times k} \end{pmatrix}$, with $A \in \mathbb{M}(k)$, the gradient evaluates to

$$\begin{aligned}
\nabla L^1_\tau(W^*) &= 2(\mathrm{I}_n - \Sigma_0^{1/2}(\Sigma_0^{1/2}\Sigma^*_\tau\Sigma_0^{1/2})^{-1/2}\Sigma_0^{1/2})\Omega_{\mathcal{J}_k}(\bar\Lambda_{\mathcal{J}_k} - \tau \mathrm{I}_k)^{1/2}V_\parallel^\top \\
&= 2\Omega \begin{pmatrix} \mathbf{0}_{k\times k} & \\ & \mathrm{I}_{n-k} - \tau^{-1/2}\bar\Lambda^{1/2}_{\mathcal{J}_k{}^c} \end{pmatrix} \begin{pmatrix} A \\ \mathbf{0}_{n-k\times k} \end{pmatrix} (\bar\Lambda_{\mathcal{J}_k} - \tau \mathrm{I}_k)^{1/2}V_\parallel^\top \\
&= 0.
\end{aligned}$$

For such a critical point $W^* = \Omega_{\mathcal{J}_k}(\bar\Lambda_{\mathcal{J}_k} - \tau \mathrm{I}_k)^{1/2}V^\top$, with regularized covariance $\Sigma^*_\tau = W^*W^{*\top} + \tau \mathrm{I}_n$, the value of the loss is

$$\begin{aligned}
L^1_\tau(W^*) &= \operatorname{tr}\Sigma^*_\tau + \operatorname{tr}\Sigma_0 - 2\operatorname{tr}(\Sigma_0^{1/2}\Sigma^*_\tau\Sigma_0^{1/2})^{1/2} \\
&= \sum_{j\in\mathcal{J}_k}\lambda_j + \tau(n-k) + \sum_{j\in\mathcal{J}_k}\lambda_j + \sum_{j\in\mathcal{J}_k{}^c}\lambda_j - 2(\sum_{j\in\mathcal{J}_k}\lambda_j + \sum_{j\in\mathcal{J}_k{}^c}\sqrt{\tau\lambda_j}) \\
&= \sum_{j\in\mathcal{J}_k{}^c}\lambda_j + \tau - 2\sqrt{\tau\lambda_j} \\
&= \sum_{j\in\mathcal{J}_k{}^c}(\sqrt{\lambda_j} - \sqrt{\tau})^2,
\end{aligned}$$

which is uniquely minimized of $\mathcal{J}_k$ for $\mathcal{J}_k = [k]$ when the eigenvalues of Σ_0 are distinct and in descending order.

Moreover, as in the unregularized case, we have the increasing sequence of minimizers $\min\{L^1_\tau(W) \mid W \in \mathcal{M}(k)\} \leqslant \min\{L^1_\tau(W) \mid W \in \mathcal{M}(< k)\}$ which, together with the identity $\mathcal{M}(\leqslant k) = \mathcal{M}(k) \cup \mathcal{M}(< k)$, implies that $\min\{L^1_\tau \mid \mathcal{M}(\leqslant k)\} = \min\{L^1_\tau \mid \mathcal{M}(k)\}$. □

The loss L^1_τ satisfies the strict-saddle point property in a similar fashion as Proposition 4.20 for L^1.

Lemma 4.22 *The loss $L^1_\tau|_{\mathcal{M}(k)}$ satisfies the strict saddle point property.*

Proof of Lemma 4.22. The proof of Proposition 4.20 can be adapted, with the expression of the critical points (4.18) replaced by, if $\Sigma_0 = \Omega\Lambda\Omega^\top$ and $V \in \mathbb{M}(n,k)$ is any semi-orthogonal matrix, $W^* = (\Sigma_0 - \tau \mathrm{I}_n)^{-1/2}\sum_{j\in\mathcal{J}_k}(\lambda_j - \tau)\omega_j v_j^\top$. □

We are now ready to prove Proposition 4.11.

Proof of Proposition 4.11. The fact that $W^* = \mu(\theta^*)$ is a critical point of $L^1_\tau|_{\mathcal{M}(k)}$ (with $k = \operatorname{rank} W^*$) if and only if θ^* is a critical point for L^N_τ comes from [TKB20, Proposition 7],

and the fact that, when $k = \underline{d} = \min_i\{d_i\}$, W^* is a local minimizer of $L^1_\tau|_{\mathcal{M}(\underline{d})}$ if and only if θ^* is a local minimizer of L^N_τ comes from [TKB20, Proposition 6], since L^1_τ is smooth.

The additional fact that any local minimizer of rank $\underline{d}$ of $L^1_\tau|_{\mathcal{M}(\underline{d})}$ is a global minimizer of $L^1_\tau|_{\mathcal{M}(\underline{d})}$ comes from Lemma 4.22: $L^1_\tau|_{\mathcal{M}(\underline{d})}$ satisfies the strict saddle point property, therefore, the only critical points of $L^1_\tau|_{\mathcal{M}(\underline{d})}$ are strict saddle points and the global minimizer.

Now, the expression of such a global minimizer is given by Theorem 4.9: with $\Sigma_0 = \Omega\Lambda\Omega^\top$ a spectral decomposition of Σ_0 in descending order of the eigenvalues, there exists $V \in \mathbb{O}(m)$ orthogonal, such that $W^* = \Omega_{[\underline{d}]}(\bar{\Lambda}_{[\underline{d}]} - \tau \mathrm{I}_{\underline{d}})^{1/2} V^\top_{[\underline{d}]}$, and $\Sigma^*_\tau = W^* W^{*\top} + \tau \mathrm{I}_n = \Omega \begin{pmatrix} \Lambda_{[\underline{d}]} & \\ & \tau \end{pmatrix} \Omega^\top$. □

From [TKB20], we have (since L^1_τ is smooth):

Proposition 4.23 ([TKB20, Proposition 7]). *Let $\overrightarrow{W} \in \Theta$ and let $W = \mu(\overrightarrow{W})$ be its image by μ (4.3) so that $k = \operatorname{rank} W$. Then, W is a critical point of L^1_τ restricted to $\mathcal{M}_k$ if and only if $\overrightarrow{W}$ is a critical point of L^N_τ.*

With this proposition, we can easily show Proposition 4.12.

Proof of Proposition 4.12. The proof of [BRT+21, Proposition 6.9] is taken as is. We repeat it for the reader's convenience.

Recall that, for $k \in \mathbb{N}$, the tangent space of $\mathcal{M}_k \subseteq \mathbb{M}(n, m)$ at $W \in \mathcal{M}_k$ is given by

$$\mathrm{T}_W(\mathcal{M}_k) = \{WA + BW \mid A \in \mathbb{M}(m, m),\ B \in \mathbb{M}(n, n)\}, \tag{4.20}$$

see e.g. [HS95, Proposition 4.1].

Since $\overrightarrow{W}$ is a critical point of L^N_τ, with $W = \mu(\overrightarrow{W})$ and $k = \operatorname{rank} W$, W is a critical point of $L^1_\tau|_{\mathcal{M}_k}$ by Proposition 4.23.

Since W is a not a global minimizer of $L^1_\tau|_{\mathcal{M}_k}$, it must be a strict saddle point by Lemma 4.22. Therefore, there exists $Z \in \mathrm{T}_W(\mathcal{M}_k)$ such that, for some Riemannian metric g as defined by [BRT+21, Definition 4.1], it holds $g(\mathrm{Hess}^g\, L^1_\tau(W)Z, Z) < 0$, where $\mathrm{Hess}^g\, f(x)$ denotes the Riemannian Hessian of f at x, as defined in [BRT+21, §6.1]. We endow $\mathcal{M}_k$ with a this Riemannian manifold structure and reason with it, to deduce properties on L^N_τ, independently on the metric g considered.

Write $Z = WA + BW$ and choose the curve

$$\gamma(t) = (W_N + tV_N) \cdot W_{N-1} \cdots W_2 \cdot (W_1 + tV_1), \tag{4.21}$$

where $V_1 = W_1 A$ and $V_N = BW_N$. We have by construction $\gamma(0) = W$ and $\dot{\gamma}(0) = Z$.

Then,

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} L_\tau^1(\gamma(t))\bigg|_{t=0} = g_W(\mathrm{Hess}^g\, L_\tau^1(W)Z, Z) < 0$$

and

$$\begin{aligned} 0 > \frac{\mathrm{d}^2}{\mathrm{d}t^2} L_\tau^1(\gamma(t))\bigg|_{t=0} &= \frac{\mathrm{d}^2}{\mathrm{d}t^2} L_\tau^N(W_1 + tV_1, W_2, \dots, W_N + tV_N)\bigg|_{t=0} \\ &= \left\langle \mathrm{Hess}\, L_\tau^N(\overrightarrow{W})(V_1, 0, \dots, 0, V_N), (V_1, 0, \dots, 0, V_N) \right\rangle, \end{aligned}$$

which shows that $\mathrm{Hess}\, L_\tau^N(\overrightarrow{W})$ is not positive semidefinite, i.e. has a negative eigenvalue. This Hessian is the Euclidean Hessian of L_τ^N and does not depend on the metric g used. This means that $\overrightarrow{W}$ is a strict saddle point.

□

4.5 BW and the Eckart-Young-Mirsky theorem

In this section, we provide further background on the Bures-Wasserstein distance. First, we show that, except in some particular cases (Lemma 4.24), the Bures-Wasserstein distance between two covariance matrices is not translation invariant (Lemma 4.25). This implies that it cannot be expressed as the norm (let alone unitary) of a difference between two matrices. In this section, we provide an explanation as to why the minimizers found in Theorem 4.5 are the same as the one found when using the squared Frobenius norm between Σ and Σ_0. First, we remark the following.

Lemma 4.24 *In the case Σ_0 and Σ commute, the Bures-Wasserstein distance reduces to the Hellinger distance:*

$$\Sigma_0\Sigma = \Sigma\Sigma_0 \quad \Longrightarrow \quad \mathcal{B}^2(\Sigma, \Sigma_0) = \|\Sigma^{1/2} - \Sigma_0^{1/2}\|_{\mathrm{F}}^2.$$

Proof. This follows from the fact that, if Σ and Σ_0 commute, so do $\Sigma^{1/2}$ and $\Sigma_0^{1/2}$, so that $\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2} = (\Sigma_0^{1/2}\Sigma^{1/2})^2$ and

$$\begin{aligned} \mathrm{tr}\left((\Sigma^{1/2})^2 + (\Sigma_0^{1/2})^2 - 2(\Sigma_0^{1/2}\Sigma^{1/2})\right) &= \mathrm{tr}\left((\Sigma^{1/2} - \Sigma_0^{1/2})(\Sigma^{1/2} - \Sigma_0^{1/2})^\top\right) \\ &= \|\Sigma^{1/2} - \Sigma_0^{1/2}\|_{\mathrm{F}}^2, \end{aligned}$$

as claimed. □

From this, one remarks that the problem of minimizing the BW distance between covariance matrices that commute falls under the framework of the Eckart-Young-Mirsky theorem. In this case, if the optimization variable is the square root $\Sigma^{1/2} = (WW^\top)^{1/2}$, we obtain a formulation in terms of the squared error loss. Nonetheless, in the case where Σ and Σ_0 do *not* commute, we do not have such a correspondence, as in general, the BW distance is not translation invariant, neither when considered as a function of (Σ, Σ_0) nor when considered as a function of $(\Sigma^{1/2}, \Sigma_0^{1/2})$. This is stated in the next two lemmas.

Lemma 4.25 (BW is not translation invariant). *There exist positive semidefinite matrices $(\Sigma, \Sigma_0) \in \mathbb{S}_+(n)^2$ and a translation $T \in \mathbb{S}_+(n)$, such that $\mathcal{B}^2(\Sigma+T, \Sigma_0+T) \neq \mathcal{B}^2(\Sigma, \Sigma_0)$. The same statement also holds for the loss $\mathcal{E}$ defined on the matrix square roots, $\mathcal{E}(\Sigma^{1/2}, \Sigma_0^{1/2}) := \mathcal{B}^2(\Sigma, \Sigma_0)$.*

Proof. For the first part of the statement, taking

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \Sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}, \quad T = \begin{pmatrix} t & 0 \\ 0 & t \end{pmatrix}, \quad t > 0,$$

then $\mathcal{B}^2(\Sigma+T, \Sigma_0+T) - \mathcal{B}^2(\Sigma, \Sigma_0) = \left(\sqrt{2+t} - \sqrt{1+t}\right)^2 - \left(\sqrt{2}-1\right)^2$, which is non-zero.

For the second part of the statement, if

$$\Sigma_0^{1/2} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}, \quad \Sigma^{1/2} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}, \quad T = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

and $\mathcal{E} \colon \mathbb{S}_+(n)^2 \to \mathbb{R}$ is such that, for all $(\Sigma, \Sigma_0^{1/2}) \in \mathbb{S}_+(n)^2$, $\mathcal{E}(\Sigma^{1/2}, \Sigma_0^{1/2}) = \mathcal{B}^2(\Sigma, \Sigma_0)$, one computes

$$\begin{aligned} \mathcal{E}(\Sigma^{1/2}, \Sigma_0^{1/2}) =& \|\Sigma^{1/2}\|_{\mathrm{F}}^2 + \|\Sigma_0^{1/2}\|_{\mathrm{F}}^2 - 2\,\mathrm{tr}\left(\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2}\right)^{1/2} \\ =& 12 - 2\,\mathrm{tr}\begin{pmatrix} 2 & 6 \\ 6 & 20 \end{pmatrix}^{1/2} \end{aligned}$$

and

$$\begin{aligned} \mathcal{E}(\Sigma^{1/2}+T, \Sigma_0^{1/2}+T) =& \|\Sigma^{1/2}+T\|_{\mathrm{F}}^2 + \|\Sigma_0^{1/2}+T\|_{\mathrm{F}}^2 \\ & - 2\,\mathrm{tr}\left((\Sigma_0^{1/2}+T)(\Sigma^{1/2}+T)(\Sigma^{1/2}+T)(\Sigma_0^{1/2}+T)\right)^{1/2} \\ =& 28 - 2\,\mathrm{tr}\begin{pmatrix} 20 & 30 \\ 30 & 90 \end{pmatrix}^{1/2}, \end{aligned}$$

which gives the difference $\mathcal{E}(\Sigma^{1/2}+T, \Sigma_0^{1/2}+T) - \mathcal{E}(\Sigma^{1/2}, \Sigma_0^{1/2}) \approx 0.121229 \neq 0$. □

Lemma 4.25 implies that in general one cannot express the Bures-Wasserstein distance (neither on the covariance nor on their square roots) as a norm of a difference (otherwise, the loss would be translation invariant). This hinders a direct application of the Eckart-Young-Mirsky theorem, where, for some unitary invariant norm $\|\cdot\|_*$, the problem is cast as $\min\{\|A - X\|_* \mid X \in \mathcal{M}(\leqslant k; n)\}$ with a fixed $A \in \mathbb{M}(n)$.

Nonetheless, there is a close link between the Bures-Wasserstein distance and the (squared) Euclidean distance. This is best seen through the definition of the 2-Wasserstein distance between two zero-centered Gaussian distributions, as we will present next. We follow here an approach inspired by [FFG+20, Theorem 1], for which we provide details in order to show a link between the minimization of the Bures-Wasserstein distance over rank-constrained covariance matrices and the Eckart-Young-Mirsky theorem (or k-PCA).

Let $n \in \mathbb{N}$ and let $k \in [n]$. The set of rank-k positive semi-definite matrices is denoted by $\mathbb{S}_+(k; n)$. Given a target matrix $B \in \mathbb{S}_{++}(n)$, we are interested in the following rank-constrained minimization problem

$$\inf\left\{\mathcal{B}^2(A, B) \mid A \in \mathbb{S}_+(k; n)\right\}. \tag{4.22}$$

For any measure $\alpha \in \mathscr{M}(\mathbb{R}^n)$, denote $\operatorname{supp}(\alpha)$ its support, i.e. $\alpha(X) = 0$ for $X \subseteq \mathbb{R}^n \setminus \operatorname{supp}(\alpha)$. The following lemma is a well known connection between covariance matrices and the support of the corresponding Gaussian probability distributions.

Lemma 4.26 *Let $A \in \mathbb{S}_+(k; n)$ and let $\alpha = \mathcal{N}(0, A)$. Then, the support of α is equal to the column space of A; i.e.,*

$$\operatorname{supp}(\alpha) = \operatorname{span}(A).$$

For $k \in [n]$, denote the set of linear subspaces of $\mathbb{R}^n$ of dimension k by $\mathfrak{L}(\mathbb{R}^n, k)$, and, for $\mathcal{C} \in \mathfrak{L}(\mathbb{R}^n, k)$, denote by $\mathcal{N}(\mathcal{C}) := \{\mathcal{N}(0, S) \mid S \in \mathbb{S}_+(k; n), \operatorname{span}(S) = \mathcal{C}\}$ the set of all Gaussian distributions with mean 0 and support $\mathcal{C}$.

Lemma 4.26 allows to translate the problem (4.22) to a problem on linear subspaces of fixed dimension. Indeed, with $\alpha = \mathcal{N}(0, A)$ and $\beta = \mathcal{N}(0, B)$, we know that $\mathcal{B}^2(A, B) = \mathcal{W}_2^2(\alpha, \beta)$. Therefore, we can split the optimization problem as

$$\inf\left\{\mathcal{B}^2(A, B) \mid A \in \mathbb{S}_+(k; n)\right\} = \inf\left\{\inf\left\{\mathcal{W}_2^2(\alpha, \beta) \mid \alpha \in \mathcal{N}(\mathcal{C})\right\} \mid \mathcal{C} \in \mathfrak{L}(\mathbb{R}^n, k)\right\}. \tag{4.23}$$

Solving (4.22) is therefore equivalent to solving the right-hand side of (4.23), which is split in two parts:

1. For a given linear subset $\mathcal{C}$ of dimension k, find the Gaussian distribution α^* that minimizes the 2-Wasserstein distance to β. Lemma 4.27 below states that α^* is the projection of β onto $\mathcal{C}$.

2. Then, find the subset $\mathcal{C}$ of required dimension that minimizes the variance of the projection of β onto the orthogonal complement of $\mathcal{C}$; or, equivalently, find $\mathcal{C}$ that maximizes the variance of the projection of β onto $\mathcal{C}$. The solution to this problem is the k-PCA decomposition of β, as stated in Lemma 4.28.

We solve the sub-problem 1 first. Recall that $\mathscr{P}_2(\mathbb{R}^n) = \{\beta \in \mathscr{P}(\mathbb{R}^n) \mid \int \|x\|^2 \,\mathrm{d}\beta(x) < \infty\}$ is the set of square integrable probability measures (i.e. the set of probability measures with finite variance) on $\mathbb{R}^n$. Given any $\beta \in \mathscr{P}_2(\mathbb{R}^n)$, we are interested in solving $\inf\{\mathcal{W}_2^2(\alpha,\beta) \mid \alpha \in \mathscr{N}(\mathcal{C})\}$. The next lemma gives the solution this problem in α. For any given linear subspace $\mathcal{C} \subseteq \mathbb{R}^n$, denote $\bar{\mathrm{p}}_{\mathcal{C}}$ the orthogonal projection onto $\mathcal{C}$.

Lemma 4.27 *Let $\beta \in \mathscr{P}_2(\mathbb{R}^n)$. One has*

$$\inf\,\{\mathcal{W}_2^2(\alpha,\beta) \mid \alpha \in \mathscr{N}(\mathcal{C})\} = \min\,\{\mathcal{W}_2^2(\alpha,\beta) \mid \alpha \in \mathscr{N}(\mathcal{C})\},$$

and the distribution α^* that achieves the minimum for a given $\mathcal{C}$ is the orthogonal projection of β onto $\mathcal{C}$: $\alpha^* = (\bar{\mathrm{p}}_{\mathcal{C}})_{\#}\beta = \beta_{\mathcal{C}}$.

Proof. Denote the admissible set of couplings with given marginals by $\Gamma(\alpha,\beta) = \{\gamma \in \mathscr{P}_2(\mathbb{R}^n \times \mathbb{R}^n) \mid \gamma_1 = \alpha,\, \gamma_2 = \beta\}$, with γ_i the marginal along the ith variable, so that

$$\mathcal{W}_2^2(\alpha,\beta) = \inf\left\{\int \|x-y\|^2 \,\mathrm{d}\gamma(x,y) \mid \gamma \in \Gamma(\alpha,\beta)\right\}.$$

Then, for any given linear subspace $\mathcal{C} \subseteq \mathbb{R}^n$, denote $\bar{\mathrm{p}}_{\mathcal{C}}$ the orthogonal projection onto $\mathcal{C}$. Define $\mu_{\mathcal{C}} := (\bar{\mathrm{p}}_{\mathcal{C}})_{\#}\mu$ for any $\mu \in \mathscr{P}_2(\mathbb{R}^n)$, and likewise $\gamma_{\mathcal{C}\times\mathcal{C}} := (\bar{\mathrm{p}}_{\mathcal{C}\times\mathcal{C}})_{\#}\gamma$, for $\gamma \in \mathscr{P}_2(\mathbb{R}^n \times \mathbb{R}^n)$, where, for $(\mathcal{X},\mathcal{Y})$ two (measurable) spaces, the push-forward $T_{\#}\mu \in \mathscr{P}_2(\mathcal{Y})$ of a measure $\mu \in \mathscr{P}_2(\mathcal{X})$ by an operator $T\colon \mathcal{X} \to \mathcal{Y}$ is such that, for any measurable set $\mathcal{S} \subseteq \mathcal{Y}$, $T_{\#}\mu(\mathcal{S}) = \mu(T^{-1}(\mathcal{S}))$ (see Definition 2.26).

Assume $\mathrm{supp}(\alpha) = \mathcal{C}$ (i.e. $\alpha = \alpha_{\mathcal{C}}$). Then, if $\gamma \in \Gamma(\alpha,\beta)$, for all $(x,y) \in \mathrm{supp}(\gamma) = \mathrm{supp}(\alpha) \times \mathrm{supp}(\beta)$, $x = \bar{\mathrm{p}}_{\mathcal{C}}(x)$ and

$$\begin{aligned}
\mathcal{W}_2^2(\alpha,\beta) &= \inf\left\{\int \|x-y\|^2 \,\mathrm{d}\gamma(x,y) \;\middle|\; \gamma \in \Gamma(\alpha,\beta)\right\} \\
&= \inf\left\{\int \|x - (\bar{\mathrm{p}}_{\mathcal{C}}(y) + \bar{\mathrm{p}}_{\mathcal{C}^\perp}(y))\|^2 \,\mathrm{d}\gamma(x,y) \;\middle|\; \gamma \in \Gamma(\alpha,\beta)\right\} \\
&= \inf\left\{\int \|x-y\|^2 \,\mathrm{d}\gamma_{\mathcal{C}\times\mathcal{C}}(x,y) \mid \gamma \in \Gamma(\alpha,\beta)\right\} + \int \|y\|^2 \,\mathrm{d}\beta_{\mathcal{C}^\perp}(y). \qquad (4.24)
\end{aligned}$$

Thus, for a given $\mathcal{C} \in \mathfrak{L}(\mathbb{R}^n, k)$, one has

$$\begin{aligned}
&\inf\left\{\mathcal{W}_2^2(\alpha,\beta) \mid \alpha \in \mathscr{N}(\mathcal{C})\right\} \\
&= \inf\left\{\inf\left\{\int \|x-y\|^2 \,\mathrm{d}\gamma_{\mathcal{C}\times\mathcal{C}}(x,y) \;\middle|\; \gamma \in \Gamma(\alpha,\beta)\right\} \;\middle|\; \alpha \in \mathscr{N}(\mathcal{C})\right\} + \int \|y\|^2 \,\mathrm{d}\beta_{\mathcal{C}^\perp}(y) \\
&= \inf\left\{\inf\left\{\int \|x-y\|^2 \,\mathrm{d}\gamma(x,y) \;\middle|\; \gamma \in \Gamma(\alpha,\beta_{\mathcal{C}})\right\} \;\middle|\; \alpha \in \mathscr{N}(\mathcal{C})\right\} + \int \|y\|^2 \,\mathrm{d}\beta_{\mathcal{C}^\perp}(y) \\
&= \inf\left\{\mathcal{W}_2^2(\alpha,\beta_{\mathcal{C}}) \mid \alpha \in \mathscr{N}(\mathcal{C})\right\} + \int \|y\|^2 \,\mathrm{d}\beta_{\mathcal{C}^\perp}(y).
\end{aligned}$$

Since $\beta_{\mathcal{C}} \in \mathscr{N}(C)$, the solution is attained for $\alpha^* = \beta_{\mathcal{C}}$ as $\mathcal{W}_2^2(\beta_{\mathcal{C}},\beta_{\mathcal{C}}) = 0$. □

In the light of Lemma 4.27, the problem (4.23) is equivalent to

$$\begin{aligned}
\inf\left\{\inf\left\{\mathcal{W}_2^2(\alpha,\beta) \mid \alpha \in \mathscr{N}(\mathcal{C})\right\} \mid \mathcal{C} \in \mathfrak{L}(\mathbb{R}^n,k)\right\} &= \inf\left\{\int \|y\|^2 \,\mathrm{d}\beta_{\mathcal{C}^\perp}(y) \mid \mathcal{C} \in \mathfrak{L}(\mathbb{R}^n,k)\right\} \\
&= \sup\left\{\int \|y\|^2 \,\mathrm{d}\beta_{\mathcal{C}}(y) \mid \mathcal{C} \in \mathfrak{L}(\mathbb{R}^n,k)\right\}.
\end{aligned}$$

Therefore, the problem boils down to finding the linear subspace $\mathcal{C}$ which maximizes the variance of the target when projected onto $\mathcal{C}$. The solution to this sub-problem 2, also known as k-PCA, is given in the next lemma, of which we provide a proof for convenience.

Lemma 4.28 *Let $\Omega\Lambda\Omega^\top = B$ be a spectral decomposition of $B \in \mathbb{S}_{++}(n)$, with the eigenvalues in Λ ranked in decreasing order. Let $k \in [n]$, and let $\Omega =: \begin{pmatrix}\Omega_{[k]} & \Omega_\perp\end{pmatrix}$ be such that $\Omega_{[k]} \in \mathbb{M}(n,k)$ are the eigenvectors corresponding to the k highest eigenvalues of B. Denote $\beta = \mathscr{N}(0,B)$ and for a linear subspace $\mathcal{C}$, denote by $\beta_{\mathcal{C}} = (\bar{\mathrm{p}}_{\mathcal{C}})_{\#}\beta$ the orthogonal projection of β onto $\mathcal{C}$. Then,*

$$\begin{aligned}
\sup\left\{\int \|y\|^2 \,\mathrm{d}\beta_{\mathcal{C}}(y) \mid \mathcal{C} \in \mathfrak{L}(\mathbb{R}^n,k)\right\} &= \max\left\{\int \|y\|^2 \,\mathrm{d}\beta_{\mathcal{C}}(y) \mid \mathcal{C} \in \mathfrak{L}(\mathbb{R}^n,k)\right\} \\
&= \int \|y\|^2 \,\mathrm{d}\beta_{\mathcal{C}^*}(y),
\end{aligned}$$

where $\mathcal{C}^* = \operatorname{span}\Omega_{[k]}$.

Proof. Recall that $\beta_{\mathcal{C}} = (\bar{\mathrm{p}}_{\mathcal{C}})_{\#}\beta$, where $\bar{\mathrm{p}}_{\mathcal{C}}$ is the orthogonal projection onto any $\mathcal{C} \in \mathfrak{L}(\mathbb{R}^n,k)$. Then,

$$\begin{aligned}
\int \|y\|^2 \,\mathrm{d}\beta_{\mathcal{C}}(y) &= \int \|\bar{\mathrm{p}}_{\mathcal{C}}(y)\|^2 \,\mathrm{d}\beta(y) = \int y_{\mathcal{C}}^\top y_{\mathcal{C}} \,\mathrm{d}\beta(y) = \int \operatorname{tr}(y_{\mathcal{C}}^\top y_{\mathcal{C}}) \,\mathrm{d}\beta(y) \\
&= \int \operatorname{tr}(y_{\mathcal{C}} y_{\mathcal{C}}^\top) \,\mathrm{d}\beta(y) = \operatorname{tr}\Big(\int y_{\mathcal{C}} y_{\mathcal{C}}^\top \,\mathrm{d}\beta(y)\Big), \qquad (4.25)
\end{aligned}$$

where the notation for $y_{\mathcal{C}} = \bar{\mathrm{p}}_{\mathcal{C}}(y)$ is used.

For any $\mathcal{C} \in \mathfrak{L}(\mathbb{R}^n, k)$, one can find an orthonormal basis of $\mathcal{C}$ of size k, which means there is equivalence between the following two statements:

(*i*) $\mathcal{C} \in \mathfrak{L}(\mathbb{R}^n, k)$.

(*ii*) $\exists C \in \mathbb{M}(n,k)\colon C^\top C = \mathrm{I}_k,\ \mathcal{C} = \operatorname{span} C$.

With such a C spanning $\mathcal{C}$, the orthogonal projection onto $\mathcal{C}$ can be written $\bar{\mathrm{p}}_{\mathcal{C}} = CC^\top$, and

$$\int y_{\mathcal{C}} y_{\mathcal{C}}^\top \,\mathrm{d}\beta(y) = \int CC^\top yy^\top CC^\top \,\mathrm{d}\beta(y) = CC^\top \Big(\int yy^\top \,\mathrm{d}\beta(y) \Big) CC^\top = CC^\top BCC^\top,$$

so that (4.25) becomes

$$\operatorname{tr}\Big(\int y_{\mathcal{C}} y_{\mathcal{C}}^\top \,\mathrm{d}\beta(y) \Big) = \operatorname{tr}(CC^\top BCC^\top) = \operatorname{tr}(C^\top BC)$$

Therefore, with $\beta = \mathscr{N}(0, B)$ and $\beta_{\mathcal{C}} = (\bar{\mathrm{p}}_{\mathcal{C}})_{\#}\beta$ for any $\mathcal{C} \in \mathfrak{L}(\mathbb{R}^n, k)$, the following equivalence holds

$$\sup\left\{ \int \|y\|^2 \,\mathrm{d}\beta_{\mathcal{C}}(y) \mid \mathcal{C} \in \mathfrak{L}(\mathbb{R}^n, k) \right\} = \sup\left\{ \operatorname{tr}(C^\top BC) \mid C \in \mathbb{M}(n,k),\ C^\top C = \mathrm{I}_k \right\}. \tag{4.26}$$

Let $\Omega\Lambda\Omega^\top = \sum_{i=1}^n \lambda_i \omega_i \omega_i^\top$ be a spectral decomposition of B with decreasing eigenvalues $\lambda_1 \geqslant \cdots \geqslant \lambda_n$.

For any $C \in \mathbb{M}(n,k)$ such that $C^\top C = \mathrm{I}_k$, we compute

$$\operatorname{tr}(C^\top BC) = \sum_{i=1}^n \lambda_i \operatorname{tr}(C^\top \omega_i \omega_i^\top C) = \sum_{i=1}^n \lambda_i \operatorname{tr}(\omega_i^\top CC^\top \omega_i) = \sum_{i=1}^n \lambda_i \langle \omega_i, CC^\top \omega_i \rangle. \tag{4.27}$$

For each $i \in [n]$, by orthogonal decomposition $\omega_i = CC^\top \omega_i + (\mathrm{I}_n - CC^\top)\omega_i$, one has that $\langle \omega_i, CC^\top \omega_i \rangle \leqslant 1$, with equality if and only if $CC^\top \omega_i = \omega_i$, i.e., if and only if $\omega_i \in \operatorname{span} C$. The sum (4.27) is therefore maximized for $C = \Omega_{[k]}$.

Therefore, the supremum in (4.26) is attained for $C = \Omega_{[k]} \iff \mathcal{C} = \operatorname{span} \Omega_{[k]}$, concluding the proof. □

Thus, the solution to (4.23) can be given as follows.

Proposition 4.29 *Let $k \in [n]$, and let $\begin{pmatrix} \Omega_{[k]} & \Omega_\perp \end{pmatrix} \begin{pmatrix} \Lambda_{[k]} & \\ & \Lambda_\perp \end{pmatrix} \begin{pmatrix} \Omega_{[k]}^\top \\ \Omega_\perp^\top \end{pmatrix} = B$ be a spectral decomposition of $B \in \mathbb{S}_{++}(n)$, with the k largest eigenvalues in $\Lambda_{[k]}$. Using the same notations as before, the problem (4.23) is solved for $\mathcal{C} = \operatorname{span}(\Omega_{[k]})$. In this case, $\alpha^* = \mathscr{N}(0, B|_k)$, where $B|_k = \Omega_{[k]} \Lambda_{[k]} \Omega_{[k]}^\top$.*

This recovers the Eckart-Young-Mirsky theorem [Mir60] describing the set of solutions to $\min\{\|A-B\|_* \mid A \in \mathcal{M}(\leqslant k; n)\}$ for some unitary invariant norm $\|\cdot\|_*$.

Proof. Lemma 4.28 already shows that the supremum is obtained for $\mathcal{C} = \operatorname{span}\Omega_{[k]}$. In this case, the optimal α^* (the projection of β onto $\mathcal{C} = \operatorname{span}\Omega_{[k]}$ according to Lemma 4.27) has covariance matrix

$$\begin{aligned} A^* &= \int xx^\top \,\mathrm{d}\alpha^*(x) = \int xx^\top \,\mathrm{d}(\bar{\mathrm{p}}_{\mathcal{C}})_{\#}\beta(x) \\ &= \int x_{\mathcal{C}} x_{\mathcal{C}}^\top \,\mathrm{d}\beta(x) \\ &= \Omega_{[k]}\Omega_{[k]}^\top \Sigma \Omega_{[k]}\Omega_{[k]}^\top \\ &= \Omega_{[k]}\Lambda_{[k]}\Omega_{[k]}^\top \\ &= B|_k. \end{aligned}$$

□

The connexion to the Eckart-Young-Mirsky theorem therefore lies in the specific properties of the ground distance $\|\cdot\|_2$ used to compute the 2-Wasserstein distance. Nevertheless, even though the critical points and the minimizers are shared between the Frobenius and the Bures-Wasserstein losses, the Hessian of the losses at the critical points are different. This will be investigated in the next section.

4.6 Expression and empirical evaluation of the Hessians

From the previous section, we see that the critical points are identical between the two losses, Frobenius or Bures-Wasserstein. One natural question that arises is to ask how the two losses geometrically differ. A way to study it is to look at their second-order derivative information. Indeed, the Hessian of the loss at the critical points will dictate how the gradient flow will escape the saddle points (i.e. the critical points that are not the global minimizer), as both losses share the strict-saddle-points property.

Therefore, in this section, we conduct experiments evaluating the Hessian of both the Frobenius and the (regularized) Bures-Wasserstein losses.

We will heavily rely on the chain rule for the second-order differential, which we recall next.

Lemma 4.30 (Chain rule for second-order differential, [MN19, Theorem 6.9]). *Let $f\colon R \to S$ and $g\colon S \to T$ be two differentiable functions on open sets, such that $h = g \circ f\colon R \to T$ is always well defined. Then, given two directions u, v, the second-order differential of h at c is*

$$\mathrm{d}^2 h(c)[u,v] = \mathrm{d}^2 g(f(c))\big[\,\mathrm{d}f(c)[u], \mathrm{d}f(c)[v]\big] + \mathrm{d}g(f(c))[\mathrm{d}^2 f(c)[u,v]]. \tag{4.28}$$

We first discuss the Burer-Monteiro parametrization and relate the differential of a given loss in function and covariance space.

4.6.1 Second-order differential of the Burer-Monteiro parametrization

In this section, we compute second-order derivatives of a *generic differentiable function* $f\colon \mathbb{S}_+(n) \to \mathbb{R}$ under the parametrization $f \circ \pi$, where

$$\begin{aligned} \pi\colon \quad \mathbb{M}(n,m) &\longrightarrow \mathbb{S}_+(n), \\ W &\longmapsto WW^\top, \end{aligned} \tag{4.29}$$

is the covariance (or Burer-Monteiro) parametrization (4.4).

Proposition 4.31 (Second-order differential chain rule for the Burer-Monteiro parametrization). *Let π (4.29) be the Burer-Monteiro parametrization of a positive semi-definite matrix, and let $W \in \mathbb{M}(n,m)$, $Z \in \mathbb{M}(n,m)$, and $\Sigma = \pi(W) = WW^\top$. Then, for any twice-differentiable function $f\colon \mathbb{S}_+(n) \to \mathbb{R}$, the second-order differential of $f \circ \pi$ at W in the direction Z can be expressed as:*

$$\begin{aligned} \mathrm{d}^2(f\circ\pi)(W)[Z] = \mathrm{d}^2 f(\Sigma)[ZW^\top] + \mathrm{d}^2 f(\Sigma)[WZ^\top] + 2\,\mathrm{d}^2 f(\Sigma)[WZ^\top, ZW^\top] \\ + 2\,\mathrm{d}f(\Sigma)[ZZ^\top]. \end{aligned} \tag{4.30}$$

Proof. Recall the chain rule for the second-order differential from Lemma 4.30. It states that

$$\mathrm{d}^2(f\circ\pi)(W)[Z] = \mathrm{d}^2 f(\pi(W))[\mathrm{d}\pi(W)[Z]] + \mathrm{d}f(\pi(W))[\mathrm{d}^2\pi(W)[Z]].$$

Since $\mathrm{d}\pi(W)[Z] = ZW^\top + WZ^\top$ and $\mathrm{d}^2\pi(W)[Z] = 2ZZ^\top$, and with $\Sigma = \pi(W)$, one further has

$$\begin{aligned} \mathrm{d}^2(f\circ\pi)(W)[Z] &= \mathrm{d}^2 f(\Sigma)[ZW^\top + WZ^\top] + \mathrm{d}f(\Sigma)[2ZZ^\top] \\ &= \mathrm{d}^2 f(\Sigma)[ZW^\top] + \mathrm{d}^2 f(\Sigma)[WZ^\top] + 2\,\mathrm{d}^2 f(\Sigma)[ZW^\top + WZ^\top] \\ &\quad + 2\,\mathrm{d}f(\Sigma)[ZZ^\top], \end{aligned}$$

where the last inequality comes from the bilinearity of $\mathrm{d}^2 f(\Sigma)$ and the linearity of $\mathrm{d}f(\Sigma)$. □

Now, we turn to the expression of the Hessian matrix for the function $f \circ \pi$ at some point $W \in \mathbb{M}(n,m)$. Recall that, by definition, this is the only symmetric matrix of size $nm \times nm$, denoted by $\nabla^2(f\circ\pi)(W)$, such that, for all $Z \in \mathbb{M}(n,m)$, $\mathrm{d}^2(f\circ\pi)(W)[Z] = (\operatorname{vec} Z)^\top [\nabla^2(f\circ\pi)(W)] \operatorname{vec} Z$.

Corollary 4.32 *With the same notations as in Proposition 4.31, the Hessian of the loss can then be identified as*

$$\begin{aligned}\nabla^2(f\circ\pi)(W) = ((W^\top\otimes \mathrm{I}_n)\mathrm{K}_n + W^\top\otimes \mathrm{I}_n)\nabla^2 f(\Sigma)(\mathrm{K}_n(W\otimes \mathrm{I}_n) + W\otimes \mathrm{I}_n)\\ + 2\mathrm{I}_m\otimes\nabla f(\Sigma).\end{aligned} \tag{4.31}$$

where $\mathrm{K}_n \in \mathbb{M}(n^2)$ *is such that, for all* $X \in \mathbb{M}(n)$, $\mathrm{K}_n \operatorname{vec} X^\top = \operatorname{vec} X$.

Proof. In order to prove the relation (4.31), we rely on the identity $\operatorname{vec} ABC = (C^\top \otimes A)\operatorname{vec}(B)$ which holds for any matrices of compatible shapes. For more details about the matrix computations, we refer to [MN19]. Using the expression of the second-order differential (4.30), it implies that, for $(W, Z) \in \mathbb{M}(n,m)^2$, one has

$$\begin{aligned}\mathrm{d}^2 f(\Sigma)[ZW^\top] &= (\operatorname{vec} ZW^\top)^\top \nabla^2 f(\Sigma)\operatorname{vec} ZW^\top\\ &= (W\otimes \mathrm{I}_n \operatorname{vec} Z)^\top \nabla^2 f(\Sigma)(W\otimes \mathrm{I}_n)\operatorname{vec} Z\\ &= (\operatorname{vec} Z)^\top (W^\top\otimes \mathrm{I}_n)\nabla^2 f(\Sigma)(W\otimes \mathrm{I}_n)\operatorname{vec} Z,\\ \mathrm{d}^2 f(\Sigma)[WZ^\top] &= (\operatorname{vec}(ZW^\top)^\top)^\top \nabla^2 f(\Sigma)\operatorname{vec}((ZW^\top)^\top)\\ &= (\mathrm{K}_n \operatorname{vec} ZW^\top)^\top \nabla^2 f(\Sigma)(\mathrm{K}_n \operatorname{vec} ZW^\top)\\ &= (\operatorname{vec} Z)^\top (W^\top\otimes \mathrm{I}_n)\mathrm{K}_n\nabla^2 f(\Sigma)\mathrm{K}_n(W\otimes \mathrm{I}_n)\operatorname{vec} Z,\end{aligned}$$

where $\mathrm{K}_n \in \mathbb{M}(n^2)$ is such that, for all $X \in \mathbb{M}(n)$, $\mathrm{K}_n \operatorname{vec} X^\top = \operatorname{vec} X$, and

$$\begin{aligned}\mathrm{d}^2 f(\Sigma)[WZ^\top, ZW^\top] &= (\operatorname{vec} Z)^\top (W^\top\otimes \mathrm{I}_n)\mathrm{K}_n\nabla^2 f(\Sigma)(W\otimes \mathrm{I}_n)\operatorname{vec} Z\\ &= (\operatorname{vec} Z)^\top (W^\top\otimes \mathrm{I}_n)\nabla^2 f(\Sigma)\mathrm{K}_n(W\otimes \mathrm{I}_n)\operatorname{vec} Z.\end{aligned}$$

Moreover,

$$\mathrm{d}f(\Sigma)[ZZ^\top] = \langle \nabla f(\Sigma), ZZ^\top\rangle = \operatorname{tr}\nabla f(\Sigma)ZZ^\top = (\operatorname{vec} Z)^\top(\mathrm{I}_m\otimes\nabla f(\Sigma))\operatorname{vec} Z,$$

where we have used the identity $\operatorname{tr} ABCD = (\operatorname{vec} D^\top)^\top (C^\top\otimes A)\operatorname{vec} B$ for the last equality. Adding the different terms according to (4.30) and factorizing, we get the desired expression

$$\nabla^2(f\circ\pi)(W)[Z] = (W^\top\otimes\mathrm{I}_n + (W^\top\otimes\mathrm{I}_n)\mathrm{K}_n)\nabla^2 f(\Sigma)(W\otimes\mathrm{I}_n + \mathrm{K}_n(W\otimes\mathrm{I}_n)) + 2(\mathrm{I}_m\otimes\nabla f(\Sigma)).$$

□

Therefore, in order to compute $\nabla^2(f\circ\pi)(W)[Z]$, one only needs to know $\nabla^2 f(\Sigma)$ and $\nabla f(\Sigma)$.

4.6.2 Frobenius loss

Define

$$\begin{array}{rccl} L_{\mathrm{F}}\colon & \mathbb{S}_+(n) & \longrightarrow & \mathbb{R}, \\ & \Sigma & \longmapsto & L_{\mathrm{F}}(\Sigma) = \frac{1}{2}\|\Sigma - \Sigma_0\|_{\mathrm{F}}^2, \end{array}$$

and denote L_{F}^1 the Frobenius loss defined on $W \in \mathbb{M}(n, m)$, and L_{F}^N the Frobenius norm defined on the parameters $\theta \in \Theta$ (4.2).

Since L_{F}^1 admits a Burer-Monteiro factorization, we only need to recall the gradient and Hessian matrix of L_{F} in order to compute their counterparts on the function space.

Lemma 4.33 *The first-order differential of L_{F} at $\Sigma \in \mathbb{S}_+(n)$ is $\mathrm{d}L_{\mathrm{F}}(\Sigma)[X] = \langle \Sigma - \Sigma_0, X \rangle$, and its gradient is $\nabla L_{\mathrm{F}}(\Sigma) = \Sigma - \Sigma_0$.*

Lemma 4.34 *The second-order differential of L_{F} at $\Sigma \in \mathbb{S}_+(n)$ in the direction $X \in \mathbb{M}(n)$ is*

$$\mathrm{d}^2 L_{\mathrm{F}}(\Sigma)[X] = \operatorname{tr} XX^\top,$$

and its Hessian matrix at Σ is $\nabla^2 L_{\mathrm{F}}(\Sigma) = \mathrm{I}_n \otimes \mathrm{I}_n$.

We can then give the Hessian matrix of the loss L_{F}^1.

Corollary 4.35 *The Hessian matrix of L_{F}^1 at W is given by*

$$\nabla^2 L_{\mathrm{F}}^1(W) = 2\Big(W^\top W \otimes \mathrm{I}_n + \mathrm{K}_n(W \otimes W^\top) + \mathrm{I}_m \otimes (\Sigma - \Sigma_0)\Big)$$

Proof. This is a direct consequence of the previous lemma and Corollary 4.32. □

4.6.3 Bures-Wasserstein loss

We provide here the first- and second-order differential properties of the Bures-Wasserstein loss. We give our notation for the differential that will be relevant for this chapter and the next one.

Notation (*Differential*). We denote the differential of f at X in the direction H by $\mathrm{d}f(X)[H]$. Sometimes, with $Y = f(X)$, the shorthand notation $\mathrm{d}Y$ is preferred, and it is assumed that the direction H is a small perturbation $\mathrm{d}X$ around X. For instance, if $Y = f(X) = XX^\top$, then $\mathrm{d}Y = \mathrm{d}X X^\top + X\,\mathrm{d}X^\top$ is one way to write $\mathrm{d}f(X)[H] = HX^\top + XH^\top$.

Lemma 4.36 (Differential of L). *Let $\Sigma \in \mathbb{S}_{++}(n)$. The first-order differential of L (4.11a) at Σ in the direction $X \in \mathbb{S}_{++}(n)$ is*

$$\mathrm{d}L(\Sigma)[X] = \operatorname{tr}\left(X - \Sigma_0^{1/2}\left[\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2}\right]^{-1/2}\Sigma_0^{1/2}X\right).$$

Proof. We will use the fact that, for $A \in \mathbb{S}_{++}(n)$, $\mathrm{d}\operatorname{tr}(A^{1/2}) = \frac{1}{2}\operatorname{tr}(A^{-1/2}\,\mathrm{d}A)$. By the differential calculus rules, for $\Sigma \in \mathbb{S}_{++}(n)$,

$$\begin{aligned}
\mathrm{d}L(\Sigma) &= \mathrm{d}\operatorname{tr}(\Sigma + \Sigma_0 - 2(\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2})^{1/2}) \\
&= \operatorname{tr}\mathrm{d}\Sigma - 2\operatorname{tr}\mathrm{d}[(\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2})^{1/2}] \\
&= \operatorname{tr}(\mathrm{d}\Sigma - (\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2})^{-1/2}\,\mathrm{d}(\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2})) \\
&= \operatorname{tr}(\mathrm{d}\Sigma - \Sigma_0^{1/2}(\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2})^{-1/2}\Sigma_0^{1/2}\,\mathrm{d}\Sigma).
\end{aligned}$$

□

Corollary 4.37 (Gradient of L). *Let $\Sigma \in \mathbb{S}_{++}(n)$. The gradient of L (4.11a) at Σ is*

$$\nabla L(\Sigma) = \mathrm{I}_n - \Sigma_0^{1/2}[\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2}]^{-1/2}\Sigma_0^{1/2}.$$

We now turn to the expression of the Hessian matrix for the Bures-Wasserstein loss. Recall that

$$\forall \Sigma \in \mathbb{S}_+(n), \quad L_\tau(\Sigma) = \mathcal{B}(\Sigma_\tau, \Sigma)^2$$

is the τ-regularized Bures-Wasserstein loss.

Again given $\theta \in \Theta$, we only need to derive the expressions of $\nabla^2 L_\tau(\Sigma)$ with $\Sigma = \pi \circ \mu(\theta)$ in order to obtain $\nabla^2 L_\tau^N(\theta)$.

The second-order derivative of L on $\mathbb{S}_{++}(n)$ is given next.

Lemma 4.38 (Second-order differential of L, [KSS21, Lemma A.6]). *Let $\Sigma \in \mathbb{S}_{++}(n)$. Given that $\Sigma \succ 0$, the loss L given by (4.11a) is twice continuously differentiable at Σ. Let $\Gamma Q \Gamma^\top = \Sigma_0^{1/2}\Sigma\Sigma_0^{1/2}$ be a spectral decomposition of $\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2}$, with $Q = \operatorname{Diag}(q_1, \dots, q_n)$. For $Y \in \mathbb{S}_{++}(n)$, define $\Delta(Y) \in \mathbb{S}(n)$ to be the matrix with elements $\Delta(Y)_{ij} = (\sqrt{q_i} + \sqrt{q_j})^{-1}(\Gamma^\top\Sigma_0^{1/2}Y\Sigma_0^{1/2}\Gamma)_{ij}$. Let $\mathcal{G}$ be the linear operator defined as*

$$\begin{aligned}
\mathcal{G}\colon \quad \mathbb{S}_{++}(n) &\longrightarrow \mathbb{S}(n) \\
Y &\longmapsto \mathcal{G}(Y) = \Sigma_0^{1/2}\Gamma Q^{-1/2}\Delta(Y)Q^{-1/2}\Gamma^\top\Sigma_0^{1/2}.
\end{aligned} \tag{4.32}$$

Then, the second-order differential of L is given by

$$\forall (X, Y) \in \mathbb{S}_{++}(n)^2, \quad \mathrm{d}^2 L(\Sigma)[X, Y] = \langle X, \mathcal{G}(Y)\rangle.$$

Proof. For completeness, we provide a proof of the statement (slightly) different from the one by [KSS21]. The first-order differential for the loss L evaluated on the matrix $\Sigma \succ 0$ is given by Lemma 4.36: for $X \in \mathbb{S}_{++}(n)$,

$$\begin{aligned}\mathrm{d}L(\Sigma)[X] &= \operatorname{tr}(X - \Sigma_0^{1/2}(\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2})^{-1/2}\Sigma_0^{1/2}X) \\ &= \langle \mathrm{I}_n - \Sigma_0^{1/2}(\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2})^{-1/2}\Sigma_0^{1/2}, X\rangle.\end{aligned}$$

Let $\mathrm{GL}(n) = \{A \in \mathbb{M}(n) \mid \det A \neq 0\}$, and let $f\colon \mathrm{GL}(n) \ni F \mapsto F^{-1}$; then f is differentiable with differential $\mathrm{d}f(F)[X] = -F^{-1}XF^{-1}$ [MN19, Theorem 8.3]. Let $g\colon \mathbb{S}_{++}(n) \ni A \mapsto A^{1/2}$ be the matrix square root. The function g is differentiable on $\mathbb{S}_{++}(n)$, and its differential can be computed as follows [KSS21, Lemma A.1].

Let $A \in \mathbb{S}_{++}(n)$, and let $\Gamma Q \Gamma^\top$ be its spectral decomposition, with $Q = \operatorname{Diag}(q_1, \dots, q_n)$. For $X \in \mathbb{S}(n)$, define $\delta(X) \in \mathbb{M}(n)$ to be the matrix with elements $\delta(X)_{ij} := (\sqrt{q_i} + \sqrt{q_j})^{-1}(\Gamma^\top X \Gamma)_{ij}$. Then, the differential of g at A in the direction X is $\mathrm{d}g(A)[X] = \Gamma\delta(X)\Gamma^\top$. Further, the chain rule on the differentials gives

$$\mathrm{d}(f\circ g)(A)[X] = \mathrm{d}f(g(A))[\mathrm{d}g(A)[X]] = -A^{-1/2}\,\mathrm{d}g(A)[X]A^{-1/2} = -A^{-1/2}\Gamma\delta(X)\Gamma^\top A^{-1/2},$$

and, with $A = \Sigma_0^{1/2}\Sigma\Sigma_0^{1/2}$,

$$\begin{aligned}\mathrm{d}^2L(\Sigma)[X,Y] &= \mathrm{d}(\Sigma \mapsto \mathrm{d}L(\Sigma)[X])[Y] \\ &= \mathrm{d}(\operatorname{tr}(X - (\Sigma_0^{1/2}(\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2})^{-1/2}\Sigma_0^{1/2}X)))[Y] \\ &= -\operatorname{tr}(\Sigma_0^{1/2}(\mathrm{d}(\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2})^{-1/2}[Y])\Sigma_0^{1/2}X) \\ &= -\operatorname{tr}(\Sigma_0^{1/2}(-A^{-1/2}\Gamma\delta(\mathrm{d}(\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2})[Y])\Gamma^\top A^{-1/2})\Sigma_0^{1/2}X) \\ &= \operatorname{tr}(\Sigma_0^{1/2}\Gamma Q^{-1/2}\delta(\Sigma_0^{1/2}Y\Sigma_0^{1/2})Q^{-1/2}\Gamma^\top\Sigma_0^{1/2}X) \\ &= \langle X, \mathcal{G}(Y)\rangle\end{aligned}$$

with

$$\mathcal{G}(Y) = \Sigma_0^{1/2}\Gamma Q^{-1/2}\Delta(Y)Q^{-1/2}\Gamma^\top\Sigma_0^{1/2}$$

and

$$\Delta(Y)_{ij} = \delta(\Sigma_0^{1/2}Y\Sigma_0^{1/2})_{ij} = (\sqrt{q_i} + \sqrt{q_j})^{-1}(\Gamma^\top\Sigma_0^{1/2}Y\Sigma_0^{1/2}\Gamma)_{ij}.$$

□

Given the definition of the operator $\mathcal{G}$ in (4.32), and Δ in (4.38), we first need to express the Hadamard product of two matrices $(A, B) \in \mathbb{M}(n)^2$, denoted by $A \odot B$, as a linear operation. This is done in the next lemma.

Lemma 4.39 (Hadamard product as matrix multiplications). *For any $(A, M) \in \mathbb{M}(n)^2$ with $\begin{pmatrix} u_1 & \cdots & u_n \end{pmatrix} \operatorname{Diag}(\sigma_1, \ldots, \sigma_n) \begin{pmatrix} v_1 & \cdots & v_n \end{pmatrix}^\top$ a SVD of A, and, for $k \in [n]$, letting $E_k = \operatorname{Diag}(u_k)$ and $F_k = \operatorname{Diag}(v_k)$, one has*

$$A \odot M = \left(\sum_{k \in [n]} \sigma_k u_k \otimes v_k \right) \odot M = \sum_{k \in [n]} \sigma_k E_k M F_k.$$

Then, the Hessian of the loss L can be computed as follows.

Proposition 4.40 *Let $\Pi \operatorname{Diag}(q_1, \ldots, q_n) \Pi^\top$ be a spectral decomposition of $\Sigma_0^{1/2} \Sigma \Sigma_0^{1/2}$, and let $\begin{pmatrix} u_1 & \cdots & u_n \end{pmatrix} \operatorname{Diag}(\sigma_1, \ldots, \sigma_n) \begin{pmatrix} u_1 & \cdots & u_n \end{pmatrix}^\top = \left((\sqrt{q_i} + \sqrt{q_j})^{-1} \right)_{(i,j) \in [n]^2} =: P$ be a spectral decomposition of the (symmetric) matrix P. Furthermore, for $k \in [n]$, let $E_k = \operatorname{Diag}(u_k)$ and let $B_k = \Sigma_0^{1/2} \Pi Q^{-1/2} E_k \Pi^\top \Sigma_0^{1/2} = B_k^\top$. The Hessian matrix for the Bures-Wasserstein loss L at $\Sigma \in \mathbb{S}_{++}(n)$ is*

$$\nabla^2 L(\Sigma) = \sum_{k=1}^{n} \sigma_k B_k^\top \otimes B_k.$$

Proof. This follows from the expression of the second-order differential given in Lemmas 4.38 and 4.39. □

Remark 4.41. Lemmas 4.36 and 4.38 are stated on $\mathbb{S}_{++}(n)$. In the general setting, when $\Sigma \in \mathbb{S}_+(n)$, the parametrization φ_τ (4.12) for $\tau \in (0, \infty)$ ensures that $\Sigma_\tau = \varphi_\tau(\Sigma) = \Sigma + \tau \mathrm{I}_n$ belongs to $\mathbb{S}_{++}(n)$. The results of Lemmas 4.36 and 4.38 can therefore be stated with Σ_τ in place of Σ; or equivalently, for the regularized loss L_τ (4.13a) on $\mathbb{S}_+(n)$.

The loss L_τ^1 admits the parametrization $L_\tau^1 = L_\tau \circ \pi$, its Hessian matrix can therefore be computed using Corollary 4.32 and Lemma 4.38. The Hessian matrix of the loss $L_\tau^N = L_\tau^1 \circ \mu$ can then be computed using Corollary 4.43 located in Section 4.6.5.

4.6.4 Condition number of the Hessian

We evaluate the Hessian for both losses in the setting $n = m = d = 20$, and $N = 3$. We evaluate the Hessians of L_τ^N and L_{F}^N (on the parameter space) according to the discussion in Sections 4.6.2 and 4.6.3, together with Corollary 4.43.

Let $H \in \mathbb{S}(n)$ be any symmetric matrix. We define the condition number for H as

$$\kappa(H) := \begin{cases} \dfrac{|\lambda|_{\max}(H)}{|\lambda|_{\min}(H)} & \text{if } |\lambda|_{\min}(H) \neq 0, \\ \infty & \text{else.} \end{cases} \tag{4.33}$$

where $|\lambda|_{\{\max,\min\}}(H)$ are the extremal eigenvalues of H in absolute value. Note that $\kappa(H) \in [1, \infty]$.

The condition number $\kappa(H)$ characterizes the condition of the matrix H: the closer it is to 1, the closer are the eigenvalues of H to another, and the better H is conditioned. In our case, H will be a Hessian matrix at a given point $\theta \in \Theta$. The better $H(\theta)$ will be conditioned, the better gradient-descent algorithms will perform at θ.

The most relevant points to inspect the Hessian are the saddle points of the loss. The condition of the Hessian at these points dictates how quickly the iterations of gradient-descent algorithm will escape them, given that there exists an eigen-direction with negative eigenvalue.

The saddle points for the optimization have a closed-form. Specifically, on the function space, they are given by the spectral decomposition of the target: if $\Sigma_0 = \Omega\Lambda\Omega^\top$ and the eigenvalues are in decreasing order, then for $i \in [n]$,

$$W^*_{i,\mathrm{F}} = \Omega_{[n-i]}\bar{\Lambda}^{1/2}_{[n-i]}V^\top_{[n-i]} \tag{4.34a}$$

is the critical point of L^1_{F} of index i, and

$$W^*_{i,\tau} = \Omega_{[n-i]}(\bar{\Lambda}_{[n-i]} - \tau \mathrm{I}_{n-i})^{1/2}V^\top_{[n-i]} \tag{4.34b}$$

is a critical point of L^1_τ of index i according to Theorem 4.9. The corresponding points in the covariance spaces are simply the images through π of $W^*_{i,\mathrm{F}}$ and $W^*_{i,\tau}$.

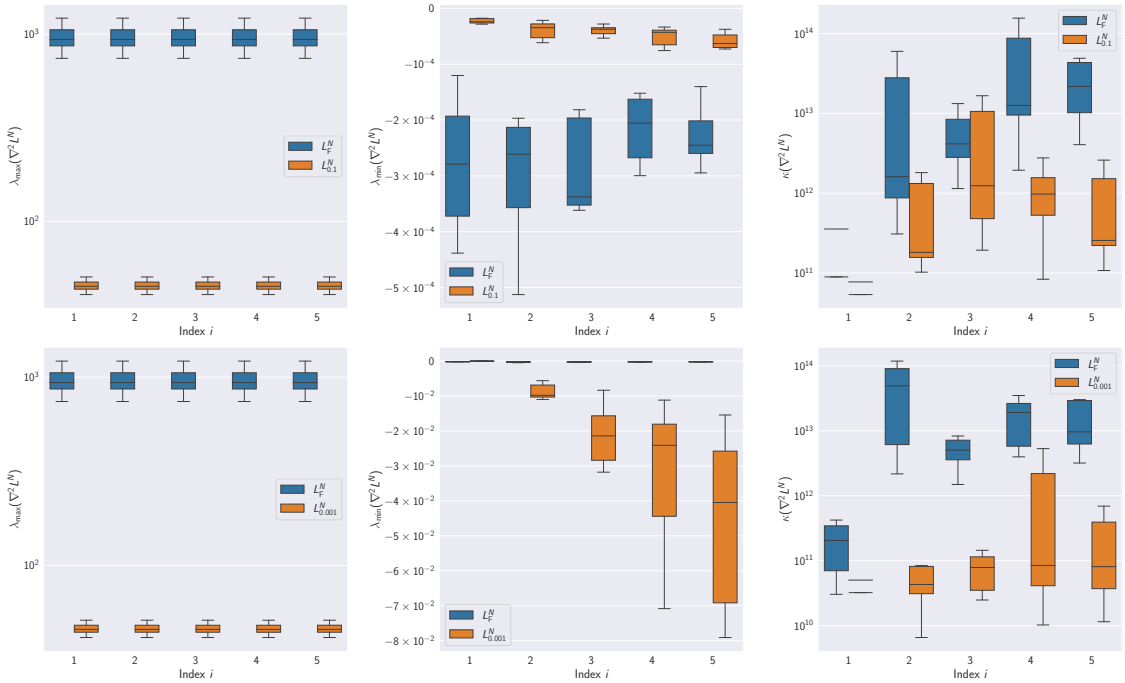
There are infinitely many points in the parameter spaces corresponding to each of the saddle points in the function (or covariance) space, but only one that satisfies the balancedness property. This is the one we select.

In Figure 4.1 are reported, for $\tau \in \{0.1, 0.001\}$, the different spectral quantities quantities $\lambda_{\max}$, $\lambda_{\min}$, and κ (4.33) at the five first local minimizers of the losses with decreasing rank (starting with the full-rank one) — i.e. for $i \in [5]$. The distribution is computed with seven different random targets Σ_0.

In Figures 4.1a and 4.1b are shown the extremal eigenvalues of the Hessians. There is an order of magnitude difference between the two losses for their maximal eigenvalues which seems independent of τ, whereas the minimal eigenvalues for the Hessians of L^N_τ become smaller as τ is decreased. In the case where $\tau = 0.001$, we also observe that the saddle point with larger index i have smaller eigenvalues.

In Figure 4.1c, the condition number κ is reported for the different losses at the different saddle points. In all cases, we note that the Hessians of the Bures-Wasserstein loss enjoy a condition number that is smaller by an order of magnitude or more. This hints at the fact that the use of the Bures-Wasserstein loss could lead to a better training dynamic.

A more systematic evaluation could be beneficial to conclude on the advantages of one or the other loss. We leave it to future work.



(a) $\lambda_{\max}(\nabla^2 L^N)$ (b) $\lambda_{\min}(\nabla^2 L^N)$ * (c) $\kappa(\nabla^2 L^N)$

*Due to the proximity of the values with 0, the scale of the plots is linear for $\lambda_{\min}$.

Figure 4.1: Spectral quantities for the Hessian matrices $\nabla^2 L_{\mathrm{F}}^N$ (blue) and $\nabla^2 L_\tau^N$ (orange) at five different critical points, with $\tau = 0.1$ (top) and $\tau = 0.001$ (bottom). The abscissa i refers to the index of the critical point $W^*_{i,\{\mathrm{F},\tau\}}$ (4.34). The boxplot representation is computed with seven different targets Σ_0.

4.6.5 General results for linear networks

This section deals with general properties of linear networks and their second-order differential in parameter space. We provide results that hold for any twice-differentiable loss $\mathcal{L}^1$ on $\mathbb{M}(n,m)$ and its parametrization $\mathcal{L}^N = \mathcal{L}^1 \circ \mu$ on Θ, where μ is given by (4.3). These results have a long history in the linear neural networks literature [BH89; Kaw16; ACH18; ACG+19; CLC22; BRT+21]. By convention, the product of matrices $W_{q:p} := W_q W_{q-1} \cdots W_p$ is equal to I_{d_q} when $q < p$.

In the case of a twice-differentiable loss $\mathcal{L}^1$ and the parametrization $\mathcal{L}^N = \mathcal{L}^1 \circ \mu$, one can express the second-order differential as follows.

Lemma 4.42 (Second-order differential). *Let $(\overrightarrow{U}, \overrightarrow{V}) \in \Theta^2$ be two parameters, such that $\overrightarrow{U} = (U_1, \ldots, U_N)$ and $\overrightarrow{V} = (V_1, \ldots, V_N)$. The second-order differential of the loss $\mathcal{L}^N = \mathcal{L}^1 \circ \mu$*

at $\overrightarrow{W} = (W_1, \dots, W_N) \in \Theta$ *is*

$$\begin{aligned} \mathrm{d}^2\mathcal{L}^N(\overrightarrow{W})[\overrightarrow{U}, \overrightarrow{V}] &= \sum_{i=1}^{N} \sum_{j \neq i} \langle U_i, W_{i+1}^\top \cdots V_j^\top \cdots W_N^\top \nabla \mathcal{L}^1(W) (W_{i-1:1})^\top \rangle \\ &+ \sum_{i=1}^{N} \sum_{j=1}^{N} \mathrm{vec}(U_i)^\top \Big([W_{i-1:1} \otimes (W_{N:i+1})^\top] \cdot \nabla^2 \mathcal{L}^1(W) \cdot [(W_{j-1:1})^\top \otimes W_{N:j+1}] \Big) \mathrm{vec}(V_j), \end{aligned} \tag{4.35}$$

where $\langle A, B \rangle = \operatorname{tr} AB^\top$ *for two matrices of compatible sizes,* $A \otimes B$ *is the Kronecker product between* A *and* B*, and* $\nabla^2 \mathcal{L}^1(W) \in \mathbb{M}(nm, nm)$ *is the matrix such that,* $\forall (U, V) \in (\mathbb{M}(n, m))^2$, $\mathrm{d}^2\mathcal{L}^1(W)[U, V] = \mathrm{vec}(U)^\top \nabla^2 \mathcal{L}^1(W) \, \mathrm{vec}(V)$*.*

Proof. The second-order differential for the parametrization μ is, for two parameters $(\overrightarrow{U}, \overrightarrow{V}) \in \Theta \times \Theta$,

$$\begin{aligned} \mathrm{d}^2\mu(\overrightarrow{W})[\overrightarrow{U}, \overrightarrow{V}] &= \mathrm{d}(\overrightarrow{W} \mapsto \mathrm{d}\mu(\overrightarrow{W})[\overrightarrow{U}])[\overrightarrow{V}] = \{\mathrm{d}\mu(\overrightarrow{W} + \overrightarrow{V})[\overrightarrow{U}] - \mathrm{d}\mu(\overrightarrow{W})[\overrightarrow{U}]\}\big|_{\text{lin}} \\ &= \Big\{ \sum_{i=1}^{N} (W_N + V_N) \cdots U_i \cdots (W_1 + V_1) - \sum_{i=1}^{N} W_N \cdots U_i \cdots W_1 \Big\}\Big|_{\text{lin}} \\ &= \sum_{i=1}^{N} \sum_{j \neq i}^{N} W_N \cdots V_j \cdots U_i \cdots W_1. \end{aligned}$$

Here, $f(\overrightarrow{U}, \overrightarrow{V})|_{\text{lin}}$ refers to the linear part of f with respect to *each* U_i and V_j.

Now, from the chain rule for second-order differentials (4.28),

$$
\begin{aligned}
&\mathrm{d}^2\mathcal{L}^N(\overrightarrow{W})[\overrightarrow{U}, \overrightarrow{V}] \\
&\quad= \mathrm{d}^2(\mathcal{L}^1 \circ \mu)(\overrightarrow{W})[\overrightarrow{U}, \overrightarrow{V}] \\
&\quad= \mathrm{d}^2\mathcal{L}^1(W)\Big[\mathrm{d}\mu(\overrightarrow{W})[\overrightarrow{U}], \mathrm{d}\mu(\overrightarrow{W})[\overrightarrow{V}]\Big] + \mathrm{d}\mathcal{L}^1(W)\Big[\mathrm{d}^2\mu(\overrightarrow{W})[\overrightarrow{U}, \overrightarrow{V}]\Big] \\
&\quad= \sum_{i=1}^{N}\sum_{j=1}^{N} \mathrm{d}^2\mathcal{L}^1(W)[W_N \cdots U_i \cdots W_1,\, W_N \cdots V_j \cdots W_1] \\
&\qquad\quad + \sum_{i=1}^{N}\sum_{j\neq i} \mathrm{d}\mathcal{L}^1(W)[W_N \cdots V_j \cdots U_i \cdots W_1] \\
&\quad= \sum_{i=1}^{N}\sum_{j=1}^{N} \operatorname{vec}(W_N \cdots U_i \cdots W_1)^\top \nabla^2\mathcal{L}^1(W) \operatorname{vec}(W_N \cdots V_j \cdots W_1) \\
&\qquad\quad + \sum_{i=1}^{N}\sum_{j\neq i} \langle \nabla\mathcal{L}^1(W), W_N \cdots V_j \cdots U_i \cdots W_1\rangle \\
&\quad= \sum_{i=1}^{N}\sum_{j=1}^{N} \Big((W_{j-1:1})^\top \otimes (W_{N:i+1}) \operatorname{vec}(U_i)\Big)^\top \nabla^2\mathcal{L}^1(W)\Big((W_{j-1:1})^\top \otimes (W_{N:i+1})\Big) \operatorname{vec}(V_j) \\
&\qquad\quad + \sum_{i=1}^{N}\sum_{j\neq i} \langle W_{i+1}^\top \cdots V_j^\top \cdots W_N^\top \nabla\mathcal{L}^1(W) W_1^\top \cdots W_{i-1}^\top, U_i\rangle \\
&\quad= \sum_{i=1}^{N}\sum_{j=1}^{N} \operatorname{vec}(U_i)^\top \Big((W_{j-1:1}) \otimes (W_{N:i+1})^\top\Big) \nabla^2\mathcal{L}^1(W)\Big((W_{j-1:1})^\top \otimes (W_{N:i+1})\Big) \operatorname{vec}(V_j) \\
&\qquad\quad + \sum_{i=1}^{N}\sum_{j\neq i} \langle W_{i+1}^\top \cdots V_j^\top \cdots W_N^\top \nabla\mathcal{L}^1(W) W_1^\top \cdots W_{i-1}^\top, U_i\rangle,
\end{aligned}
$$

which concludes the proof. □

Corollary 4.43 (Hessian of the Loss). *The Hessian of $\mathcal{L}^N$ at $\overrightarrow{W} \in \Theta = \bigtimes_{i\in[N]} \mathbb{M}(d_i, d_{i-1})$, denoted by $\nabla^2\mathcal{L}^N(\overrightarrow{W})$, can be represented as a $d_\theta \times d_\theta$ matrix, where $d_\theta = \sum_{i\in[N]} d_i d_{i-1}$. It is a block matrix with blocks corresponding to different layers. For $(i,j) \in [N]^2$, each block $\nabla^2_{W_i,W_j}\mathcal{L}^N(\overrightarrow{W})$ has dimension $d_i d_{i-1} \times d_j d_{j-1}$, and corresponds to the differential $\mathrm{d}^2\mathcal{L}^N(\overrightarrow{W})[\overrightarrow{U}_i, \overrightarrow{U}_j]$, where $\overrightarrow{U}_i = (0, \ldots, 0, U_i, 0, \ldots, 0) \in \Theta$ for some $U_i \in \mathbb{M}(d_i, d_{i-1})$. The diagonal block elements are*

$$
\nabla^2_{W_i}\mathcal{L}^N(\overrightarrow{W}) = [W_{i-1:1} \otimes (W_{N:i+1})^\top] \cdot \nabla^2\mathcal{L}^1(W) \cdot [(W_{i-1:1})^\top \otimes W_{N:i+1}], \tag{4.36}
$$

and the off-diagonal blocks are

$$\begin{aligned}\nabla^2_{W_i,W_j}\mathcal{L}^N(\overrightarrow{W}) = [W_{i-1:1}\otimes(W_{N:i+1})^\top]\cdot\nabla^2\mathcal{L}^1(W)\cdot[(W_{j-1:1})^\top\otimes W_{N:j+1}] \\ + \Big[(W_{i-1:1}\nabla\mathcal{L}^1(W)^\top W_{N:j+1})\otimes(W_{j-1:i+1})^\top\Big]\mathrm{K}_{d_jd_{j-1}},\end{aligned} \tag{4.37}$$

where K_{pq} *is the* pq*-commutation matrix (for* $X\in\mathbb{M}(p,q)$, $\mathrm{K}_{pq}\operatorname{vec}X=\operatorname{vec}X^\top$*).*

Proof. The evaluation of the second-order differential $\mathrm{d}^2L^N(\overrightarrow{W})[\overrightarrow{U},\overrightarrow{V}]$ given in (4.35) at $[\overrightarrow{U}_i,\overrightarrow{U}_i]$ readily provides the diagonal blocks of the Hessian. For the off-diagonal blocks, the expression

$$\langle U_i, W_{i+1}^\top\cdots U_j^\top\cdots W_N^\top\nabla\mathcal{L}^1(W)W_1^\top\cdots W_{i-1}\rangle$$

can be transformed into

$$\begin{aligned}&\operatorname{vec}(U_i)^\top\operatorname{vec}(W_{i+1}^\top\cdots U_j^\top\cdots W_N^\top\nabla\mathcal{L}^1(W)(W_{i-1:1})^\top)\\ &=\operatorname{vec}(U_i)^\top\Big[(W_{i-1:1}\nabla\mathcal{L}^1(W)W_{N:j+1})\otimes(W_{j-1:i+1})^\top\Big]\operatorname{vec}(U_j^\top)\\ &=\operatorname{vec}(U_i)^\top\Big[(W_{i-1:1}\nabla\mathcal{L}^1(W)W_{N:j+1})\otimes(W_{j-1:i+1})^\top\Big]\mathrm{K}_{d_jd_{j-1}}\operatorname{vec}(U_j),\end{aligned}$$

proving (4.37). □

4.7 Conclusion

In this chapter, we studied the loss landscape of a deep linear network under the Bures-Wasserstein loss. We characterized the critical points and minimizers of this loss in function space, over the set of matrices of fixed rank k, for both the original and perturbative losses. The Hessian of the loss at the critical points is empirically investigated, with a derivation to compute it in closed-form. Even though the expression of the critical points are similar between the Frobenius and the Bures-Wasserstein losses, the Hessian spectral properties at these points for the two losses are experimentally different.

From the experiments, the Bures-Wasserstein loss appears to enjoy better conditioned Hessian matrices at the saddle points of the optimization.

Our results contribute to the ongoing efforts to better characterize the optimization problems and understanding the loss landscape that arise in learning with deep neural networks beyond the commonly considered discriminative settings with the square loss.

In future work, it would be interesting to refine our characterization of critical points of the Bures-Wasserstein loss in the parameter space. In addition, a more elaborate comparison between the Frobenius and Bures-Wasserstein loss would enlighten their

respective benefits or pitfalls. Moreover, we have limited the study to the discrete case, with finite-dimensional covariance matrices. This analysis could be enriched by studying Gaussian processes instead, and investigating if and how corresponding learning systems benefit from a Bures-Wasserstein-type loss.

5 Convergence of a linear Bures-Wasserstein GAN

This chapter is based on work published as [BPA+23]. *PB's main contributions are the derivations for the bounds of the second-order derivative of the loss, the derivation of convergence results for the gradient flow scheme and gradient descent schemes, as well as the design and implementation of the empirical experiments.*

In this chapter, we derive *convergence guarantees* of the optimization procedure defined for the learning problem introduced in Chapter 4, for both the gradient flow and gradient descent schemes. Two simplifying assumptions make the analysis amenable: *i*) the discriminator will be assumed to be perfect at all times, and *ii*) the generator will have a bias-free deep linear architecture, i.e. a neural network with identity function as activation. Together, these assumptions will lead to study the minimization problem of the *Bures-Wasserstein distance* between two centered Gaussian distributions, as introduced in Section 4.2. Moreover, a set of additional hypothesis will be assumed on the weights initialization, depending on the analyzed learning algorithm. Strong assumptions will allow the quantitative analysis of the convergence under gradient flow, and a milder ones will give rise to a qualitative convergence result in the case of gradient descent.

The chapter is organised as follows. We provide first in Section 4.1 a literature review and a thorough motivation of the problem at hand. Then, in Section 5.2, we recall different properties and usual assumptions on linear neural networks, and present the two learning algorithms that will be analyzed, as well as their corresponding weights hypothesis. In Section 5.4, the dynamical analysis is performed for both the gradient flow and gradient descent algorithms. Finally, empirical illustration of the model is offered in Section 5.6, which corroborate the theoretical findings. Proofs and further developments are gathered in Sections 5.3 and 5.5.

5.1 Previous work

Most of the previous work on the convergence study of overparametrized network considers either the squared Frobenius loss or a differentiable convex loss on the function (matrix) space.

Optimization dynamics of deep linear networks [SMG14] studied the learning dynamics of deep linear networks under different types of initial conditions. [ACH+19] obtained a closed-form expression for the parametrization along time in a deep linear network for the squared error loss. Notably, the authors found that solutions with a lower rank are preferred as the depth of the network increases. [ACH18] derived several invariances of the flow and compared the dynamics in parameter and function spaces. For the squared error loss, [ACG+19] proved linear convergence of gradient descent for linear networks without bottlenecks, with weights initialized to fulfill two assumptions — approximate balancedness and so that the end-to-end matrix is close in some sense to the solution. We frame our discussion by similar assumptions. Under the balancedness assumption for the initial weights, [BRT+21] showed that for deep linear neural networks, the gradient flow of the squared error loss can be cast as a Riemannian gradient flow in the function space, and as such converges to a critical point which is a global minimizer on the manifold of fixed rank matrices of a given rank. More recently, [NRT21] extended this convergence analysis to the (full-batch) gradient descent algorithm.

As a last note, a detailed analysis of the dynamics in the case of shallow linear networks with the squared error loss was conducted by [TFH+21; MTV+21]. The authors use symmetric and asymmetric factorization of a shallow linear network to study its convergence dynamics. The role of the "imbalancedness" of the weights was also remarked in those works.

Bures-Wasserstein distance The Bures-Wasserstein distance has been of particular interest due to its geometrical properties when dealing with positive (semi-)definite matrices. [CMR+20] studied the convergence of gradient descent algorithms for the Bures-Wasserstein barycenter, proving linear rates of convergence. In contrast to our work, they considered a Polyak-Łojasiewicz inequality derived from optimal transport theory to circumvent the non-geodesic convexity of the barycenter. In the same vein, [MC18] exploited optimal transport theory to optimize the distance between two elliptical distributions. To avoid rank deficiency, they perturbed the diagonal elements of the covariance matrix by a small parameter. We will follow the same approach in order to make the loss everywhere differentiable, although in some cases, the assumptions on the network parameters force the generated covariance matrix to be full-rank.

Overall, previous works have mostly focused on the use of the (squared) Frobenius norm, which enjoys specific geometrical properties that are not shared with many other losses. For matrix approximation, a more general category of loss is the unitary invariant norms, which are invariant under unitary transformations of the matrix. For instance, the shape of the minimizers of a unitary invariant distance to a target can be expressed with the Singular Value Decomposition (SVD) of the target [Mir60].

This chapter will tackle the issues of deriving convergence guarantees of a learning system in the case a Bures-Wasserstein loss is considered.

Recall the framework presented in Section 4.2. We detail next the different learning algorithm that will be analysed in this chapter.

5.2 Gradient learning in deep linear networks

Let $N \in \mathbb{N}$, and let $(d_0, \dots, d_N) \in \mathbb{N}^{N+1}$ be $N+1$ dimensions. The parameter space of a linear network of depth N is given by $\Theta = \bigtimes_{j=1}^{N} \mathbb{M}(d_j, d_{j-1})$. Recall the deep linear network parametrization μ (4.3):

$$\begin{array}{rccc} \mu\colon & \Theta & \longrightarrow & \mathbb{M}(d_N, d_0), \\ & \theta = (W_1, \dots, W_N) & \longmapsto & W_{N:1} = W_N \cdots W_1. \end{array} \tag{5.1}$$

We recall a few results on the gradient dynamics of linear networks for general differentiable losses, which have been established in previous works with focus on the squared error loss [Kaw16; BRT+21; CLC22; ACH18].

For the remainder of this section, let $\mathcal{L}^1\colon \mathbb{M}(n,m) \to \mathbb{R}$ be *any* differentiable loss function, and let

$$\begin{array}{rccc} \mathcal{L}^N\colon & \Theta & \longrightarrow & \mathbb{R}, \\ & \theta & \longmapsto & \mathcal{L}^1 \circ \mu(\theta), \end{array} \tag{5.2}$$

be defined through the parametrization μ (5.1). Notice that $\mathcal{L}^N$ thus defined is differentiable.

The gradient of $\mathcal{L}^N$ depends on the gradient of $\mathcal{L}^1$ as

$$\nabla_{W_j} \mathcal{L}^N(\theta) = W_{j+1}^\top \cdots W_N^\top \nabla \mathcal{L}^1(\mu(\theta)) W_1^\top \cdots W_{j-1}^\top. \tag{5.3}$$

5.2.1 Descent trajectories

Gradient flow

Let $\theta^{(0)} \in \Theta$ be an initial datum. The gradient flow of $\theta\colon [0,\infty) \to \Theta$ under a loss $\mathcal{L}^N$ (5.2) is defined by

$$\begin{cases} \theta(0) = \theta^{(0)}, \\ \dfrac{\mathrm{d}\theta(t)}{\mathrm{d}t} = -\nabla \mathcal{L}^N(\theta(t)). \end{cases} \tag{5.4a}$$

$$\Longleftrightarrow$$

$$\forall j \in [N], \quad \begin{cases} W_j(0) = W_j^{(0)}, \\ \dfrac{\mathrm{d}W_j(t)}{\mathrm{d}t} = -\nabla_{W_j} \mathcal{L}^N(W_1(t), \dots, W_N(t)). \end{cases} \tag{5.4b}$$

This governs the evolution of the parameters in the continuous time setting.

Gradient descent

In the discrete-time setting, with a small abuse of notation, we consider the discrete path $\theta\colon \mathbb{N}_0 \to \Theta$. Given a stepsize $\eta \in (0, \infty)$ and the initial datum $\theta^{(0)} \in \Theta$, the gradient descent iterations are defined as

$$\begin{cases} \theta(0) = \theta^{(0)}, \\ \forall k \in \mathbb{N}_0, \quad \theta(k+1) = \theta(k) - \eta \nabla \mathcal{L}^N(\theta(k)). \end{cases} \tag{5.5}$$

We conclude this section by recalling one important property of the gradient flow for deep linear networks that we will leverage in our analysis.

Lemma 5.1 (Gradient flow, [BRT+21, Lemma 2.1]). *For any differentiable loss $\mathcal{L}^1$ and its parametrization $\mathcal{L}^N$ (5.2), if $[0, \infty) \ni t \mapsto \theta(t) = (W_1(t), \dots, W_N(t)) \in \Theta$ follows the flow (5.4), then the product $W_\theta(t) = \mu(\theta(t))$ satisfies, for all $t \in [0, \infty)$,*

$$\frac{\mathrm{d}W_\theta(t)}{\mathrm{d}t} = -\sum_{j=1}^{N} W_{N:j+1}(t) W_{N:j+1}^\top(t) \nabla \mathcal{L}^1(W_\theta(t)) W_{j-1:1}^\top(t) W_{j-1:1}(t) \tag{5.6}$$

The conservation quantity described in Lemma 5.1 for the gradient flow (5.4) is key in numerous analyses.

5.2.2 Balancedness assumptions

Gradient flow

As it turns out, the gradient flow dynamics (5.4) preserves the difference of the Gramians of subsequent layer weight matrices, which are thus invariants of the gradient flow; i.e.

$$\forall t \in [0, \infty), \quad \frac{\mathrm{d}}{\mathrm{d}t}(W_{j+1}(t)^\top W_{j+1}(t)) = \frac{\mathrm{d}}{\mathrm{d}t}(W_j(t) W_j(t)^\top). \tag{5.7}$$

The notion of *balancedness* for the weights of linear networks was first introduced by [Fuk98] in the shallow case and generalized to the deep case by [DHL18]. This is useful as it removes the redundancy of the parametrization when investigating the dynamics in function space and has been considered in numerous works.

Definition 5.2 (*Balanced weights*). The weights $(W_1, \dots, W_N) \in \Theta$ are said to be *balanced* if, for all $j \in [N-1]$, $W_j W_j^\top = W_{j+1}^\top W_{j+1}$.

Assuming balanced initial weights, if the flow of each W_j is defined and bounded, then the rank of the end-to-end matrix W remains constant during a finite-time training [BRT+21, Proposition 4.4]. Moreover, the products $W_{N:1} W_{N:1}^\top$ and $W_{N:1}^\top W_{N:1}$ can be written in a concise manner; namely, $W_{N:1} W_{N:1}^\top = W_N \cdots W_1 W_1^\top \cdots W_N^\top = (W_N W_N^\top)^N$ and $W_{N:1}^\top W_{N:1} = (W_1^\top W_1)^N$, which simplifies computations.

Gradient descent

Under gradient descent (5.5), the invariant (5.7) does not hold exactly. A slightly relaxed notion has emerged in the literature to accommodate the deviation from the infinitely small step size of the gradient flow; namely, the *approximate balancedness*, or δ*-balancedness*. It exists under different formulations, mostly depending on the type of norm used to define it. We follow the presentation of [NRT21] for the gradient descent analysis, and therefore borrow their definition based on the spectral norm of the weights.

Under a gradient update (5.5) of the weights, the balancedness property is not conserved during training. A relaxation of the property, named δ-balancedness, is required in order to carry out the convergence analysis. Following [NRT21], we state the δ-balancedness property in terms of the spectral norm of the weight matrices (in contrast with [ACG+19] where the property is defined in terms of the Frobenius norm). Recall that, for $A \in \mathbb{M}(n, m)$, we denote $\|A\|$ the spectral norm of A; i.e., $\|A\| = \sigma_{\max}(A)$.

Definition 5.3 (*Approximate balancedness,* [*NRT21, Definition 2.3*]). The tuple $\overrightarrow{W} = (W_1, \dots, W_N) \in \Theta$ is said to have *balancedness constant* $\delta \in [0, \infty)$, or to be δ-balanced, if

$$\forall j \in [N-1], \quad \|W_{j+1}^\top W_{j+1} - W_j W_j^\top\| \leqslant \delta. \tag{5.8}$$

Remark 5.4. Sometimes, the approximate balancedness property is stated with the Frobenius norm in (5.8). As, for any matrix A, $\|A\| \leqslant \|A\|_{\mathrm{F}}$, the notion of approximate balancedness given by Definition 5.3 is more permissive than when considering the Frobenius norm.

This general framework for deep linear network is relevant for any differentiable loss $\mathcal{L}^1$ and its parametrization $\mathcal{L}^N$. We will specialize the results in the case we pick a particular loss, namely, the (regularized) Bures-Wasserstein loss introduced in Section 4.2.5. This loss will have the particularity of not being convex on $\mathbb{M}(n, m)$, in contrast with the usual Frobenius loss. We recall lastly the different losses, introduced in Section 4.2.5, that we consider.

5.2.3 Bures-Wasserstein losses

The covariance (or Burer-Monteiro) parametrization π (4.4) is given by

$$\begin{array}{rccc} \pi\colon & \mathbb{M}(n,m) & \longrightarrow & \mathbb{S}_+(n) \\ & W & \longmapsto & \pi(W) = WW^\top. \end{array} \tag{5.9}$$

Under π (5.9) and μ (5.1), recall the different parametrizations of the Bures-Wasserstein loss:

$$
\begin{aligned}
L\colon \quad \mathbb{S}_+(n) &\longrightarrow \mathbb{R} && \text{(5.10a)}\\
\Sigma &\longmapsto L(\Sigma) = \operatorname{tr}\left(\Sigma + \Sigma_0 - 2\left[\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2}\right]^{1/2}\right),\\
L^1\colon \quad \mathbb{M}(n,m) &\longrightarrow \mathbb{R} && \text{(5.10b)}\\
W &\longmapsto L^1(W) = L\circ\pi(W),\\
L^N\colon \quad \Theta &\longrightarrow \mathbb{R} && \text{(5.10c)}\\
\theta &\longmapsto L^N(\theta) = L\circ\pi\circ\mu(\theta).
\end{aligned}
$$

As done in Section 4.2.5, recall the regularized (or perturbative) losses:

$$
\begin{aligned}
L_\tau\colon \quad \mathbb{S}_+(n) &\longrightarrow \mathbb{R} && \text{(5.11a)}\\
\Sigma &\longmapsto L_\tau(\Sigma) = \operatorname{tr}\left(\Sigma + \tau\mathrm{I}_n + \Sigma_0 - 2\left[\Sigma_0^{1/2}\left(\Sigma+\tau\mathrm{I}_n\right)\Sigma_0^{1/2}\right]^{1/2}\right),\\
L_\tau^1\colon \quad \mathbb{M}(n,m) &\longrightarrow \mathbb{R} && \text{(5.11b)}\\
W &\longmapsto L_\tau^1(W) = L_\tau\circ\pi(W),\\
L_\tau^N\colon \quad \Theta &\longrightarrow \mathbb{R} && \text{(5.11c)}\\
\theta &\longmapsto L_\tau^N(\theta) = L_\tau\circ\pi\circ\mu(\theta).
\end{aligned}
$$

Remark 5.5. The perturbative loss (5.11b), as well as the original loss on fixed-rank matrices, are differentiable. Many results of [BRT+21] can be carried over for these differentiable Bures-Wasserstein losses. For example, the uniform boundedness at any time $t \in [0,\infty)$ of the end-to-end matrix holds, as $\|W(t)\| \leqslant \sqrt{2L^1(W(0)) + \operatorname{tr}\Sigma_0}$ (see Corollary 5.11). Similar observations may apply for L^1 if the matrix $WW^\top$ remains positive definite throughout training, in which case the gradient flow remains well-defined and the loss is monotonically decreasing. We expand on this in Section 5.5.4.

We first provide useful bounds on the Bures-Wasserstein loss.

5.3 Boundedness of the Bures-Wasserstein loss

In this section, we will derive the results about the (smooth) Bures-Wasserstein loss required for the convergence analysis. The main result of the section is Proposition 5.12, stating the convergence of the gradient flow on the parameter space on the perturbative loss to a critical point of the loss.

Our goal is to show convergence to a critical point of $L_\tau^N = L_\tau\circ\pi\circ\mu$ (5.10c) under the gradient flow update of the parameters (Equation (5.4)). We first show that the BW loss L^1 restricted to the matrices W of full row-rank $\mathcal{M}_*$ satisfies the so-called Łojasiewicz inequality (meaning there exist constants $(p,q) \in (0,\infty)^2$ such that, for all $W \in \mathcal{M}_*$ in a neighbourhood of a critical point $W^* \in \mathcal{M}_*$, $\|\nabla L^1(W)\| > p\|L^1(W) - L^1(W^*)\|^q$).

Lemma 5.6 *For any $W \in \mathcal{M}_*$ (such that $WW^\top \in \mathbb{S}_{++}(n)$), and for the loss L^1 defined in* (5.10b)*, we have*

$$\|\nabla_W L^1(W)\|_{\mathrm{F}}^2 = 4L^1(W). \tag{5.12}$$

Proof. This equality can be obtained by direct computation. Since

$$\nabla L^1(W) = 2W - 2\Sigma_0^{1/2}(\Sigma_0^{1/2}WW^\top\Sigma_0^{1/2})^{-1/2}\Sigma_0^{1/2}W,$$

we have

$$\begin{aligned}\|\nabla_W L^1(W)\|_{\mathrm{F}}^2 &= 4\,\mathrm{tr}\Big(\big(W - \Sigma_0^{1/2}(\Sigma_0^{1/2}WW^\top\Sigma_0^{1/2})^{-1/2}\Sigma_0^{1/2}W\big)\cdot \\ &\qquad \big(W^\top - W^\top\Sigma_0^{1/2}(\Sigma_0^{1/2}WW^\top\Sigma_0^{1/2})^{-1/2}\Sigma_0^{1/2}\big)\Big) \\ &= 4\,\mathrm{tr}(WW^\top) - 4\,\mathrm{tr}\Big(WW^\top\Sigma_0^{1/2}(\Sigma_0^{1/2}WW^\top\Sigma_0^{1/2})^{-1/2}\Sigma_0^{1/2}\Big) \\ &\quad - 4\,\mathrm{tr}\Big(\Sigma_0^{1/2}(\Sigma_0^{1/2}WW^\top\Sigma_0^{1/2})^{-1/2}\Sigma_0^{1/2}WW^\top\Big) + 4\,\mathrm{tr}(\Sigma_0).\end{aligned}$$

Note that the two middle terms above are the same, and they can be simplified as

$$\begin{aligned}\mathrm{tr}\Big(WW^\top\Sigma_0^{1/2}(\Sigma_0^{1/2}WW^\top\Sigma_0^{1/2})^{-1/2}\Sigma_0^{1/2}\Big) &= \mathrm{tr}\Big(\Sigma_0^{1/2}(\Sigma_0^{1/2}WW^\top\Sigma_0^{1/2})^{-1/2}\Sigma_0^{1/2}WW^\top\Big) \\ &= \mathrm{tr}\Big((\Sigma_0^{1/2}WW^\top\Sigma_0^{1/2})^{1/2}\Big).\end{aligned}$$

Combining all the terms together, we get the equality (5.12). □

A key property for deep-linear-neural networks is the following, which ensures that the gradient flow (5.4) converges to a critical point of $\mathcal{L}^N$. Namely, if the trajectory $t \mapsto \theta(t)$ remains bounded for all $t \in [0,\infty)$, and if $\mathcal{L}^1$ is an analytic function (i.e. locally given by a power series), then (5.4) converges to a critical point of $\mathcal{L}^N$, i.e. a point $\theta^* \in \Theta$ so that $\nabla\mathcal{L}^N(\theta^*) = 0$. This is stated in the next theorem.

Theorem 5.7 (Gradient flow converges to a critical point of $\mathcal{L}^N$). *Let $\mathcal{L}^1$ be analytic and suppose the trajectory $[0,\infty) \ni t \mapsto \mu(\theta(t))$ remains bounded under the gradient flow evolution $\dot\theta = -\nabla[\mathcal{L}^1 \circ \mu](\theta)$. Then, the flows of $W_i(t)$ and $W(t)$ given by* (5.4) *and* (5.6) *are defined and bounded for all $t \in [0,\infty)$, and $\theta(t) = (W_1(t),\dots,W_N(t))$ converges to a critical point of $\mathcal{L}^N = \mathcal{L}^1 \circ \mu$ as $t \to \infty$.*

Proof. This result is proven by [BRT+21, Theorem 3.2] for the squared error loss, but it can be stated for an arbitrary analytic loss. It relies on the Łojasiewicz argument for the convergence of gradient flows [AMA05, Theorem 2.2], and the fact that each of the weights W_i is bounded in norm as long as the end-to-end product $\mu(\theta) = W_{N:1}$ is. This last claim is proven by [BRT+21, Theorem 3.2] and does not depend on the particular loss, as long as it is differentiable (so that the gradient flow is well defined). □

The boundedness of $\|W\|_{\mathrm{F}}$ can be shown depending on the loss that is considered. For example, it holds for the regularized loss L_τ^1 as we discuss next. For the loss L_τ^1 introduced in (5.11b), one can indeed bound the norm of W throughout training as stated in Lemma 5.10 below. Since the loss L_τ^1 is analytic, one immediately gets the following result. We give a simple test to show the boundedness of a trajectory under (5.4), using the decrease of the loss along training.

Lemma 5.8 *Let $\mathcal{L}^1 \colon \mathbb{M}(n,m) \to \mathbb{R}$ be a given loss differentiable loss and let $\mu \colon \Theta \to \mathbb{M}(n,m)$ (5.1) be the linear network parametrization. Let $\theta \colon [0,\infty) \to \Theta$ be a path on the parameter space under gradient flow (5.4) of $\mathcal{L}^N = \mathcal{L}^1 \circ \mu$. Denote $W = \mu \circ \theta \colon [0,\infty) \to \mathbb{M}(n,m)$ the corresponding trajectory in the function space. Assume that there exists an increasing function $f \colon \mathbb{R} \to \mathbb{R}$ such that, for any $t \in [0,\infty)$, one has $\|W(t)\|_{\mathrm{F}} \leqslant f(\mathcal{L}^1(W(t)))$. Then, the trajectory $W \colon [0,\infty)$ is bounded; i.e., $\forall t \in [0,\infty)$, $\|W(t)\|_{\mathrm{F}} < \infty$.*

Proof. Under gradient flow, for any $t \in [0,\infty)$, $\mathcal{L}^1(W(t)) \leqslant \mathcal{L}^1(W(0))$. Indeed, writing the chain rule and the gradient flow (5.4b),

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}^1(W(t)) &= \sum_j \left\langle \nabla_{W_j}\mathcal{L}^N(W_1(t),\dots,W_N(t)), \frac{\mathrm{d}W_j(t)}{\mathrm{d}t} \right\rangle \\ &= -\sum_j \left\| \nabla_{W_j}\mathcal{L}^N(W_1,\dots,W_N) \right\|_{\mathrm{F}}^2 \leqslant 0. \end{aligned} \tag{5.13}$$

Therefore, for any $t \in [0,\infty)$, $\mathcal{L}^1(W(t)) \leqslant \mathcal{L}^1(W(0))$. Now, let $f \colon \mathbb{R} \to \mathbb{R}$ be an increasing function, so that $f(\mathcal{L}^1(W(t))) \leqslant f(\mathcal{L}^1(W(0)))$. If, for any $t \in [0,\infty)$, $\|W(t)\|_{\mathrm{F}} \leqslant f(\mathcal{L}^1(W(t)))$, then $\|W(t)\|_{\mathrm{F}} \leqslant f(\mathcal{L}^1(W(t))) \leqslant f(\mathcal{L}^1(W(0)))$ is bounded. □

The assumption in Lemma 5.8 is satisfied for a couple of losses, including the squared error loss [BRT+21] and the L_τ^1 loss, as shown in Lemma 5.10 below. This allows us to consider losses that "grow with the weights", so that the end-to-end matrix is bounded when the loss converges to zero. We now show the boundedness of the weights when considering the Bures-Wasserstein loss (5.10b).

Lemma 5.9 (Boundedness for the BW loss L). *Given a target $\Sigma_0 \in \mathbb{S}_{++}(n)$, and for any $\Sigma \in \mathbb{S}_+(n)$, the loss L (4.11a) at Σ is lower-bounded by $\frac{1}{2}\operatorname{tr}\Sigma - \operatorname{tr}\Sigma_0$.*

Proof. For any $\nu \in \mathscr{M}(X)$, recall that $\mathrm{L}^1(\nu)$ denotes the set of ν-integrable functions on X (Definition 2.18).

By the dual expression of the 2-Wasserstein distance using the 2-transform mechanics presented in Equation (2.32), we have

$$L(\Sigma) = \mathcal{W}_2^2(\nu_0,\nu_\theta) = \sup\left\{ \int u \,\mathrm{d}\nu_\theta + \int u^{(2)} \,\mathrm{d}\nu_0 \mid u \in \mathrm{L}^1(\nu_\theta) \right\},$$

where $u^{(2)}$ is the 2-transform of u defined, for any $u \in \mathrm{L}^1(\nu_\theta)$, as

$$\begin{aligned} u^{(2)}\colon \ \mathbb{R}^d &\longrightarrow \mathbb{R}, \\ x' &\longmapsto \inf\left\{\|x - x'\|^2 - u(x) \mid x \in \mathbb{R}^d\right\}. \end{aligned}$$

Choosing $v\colon x \mapsto \frac{1}{2}\|x\|^2$, we have $v \in \mathrm{L}^1(\nu_\theta)$ and its 2-transform is $v^{(2)}\colon x' \mapsto -\|x'\|^2$. Therefore, we get

$$L(\Sigma) = \mathcal{W}_2^2(\nu_0, \nu_\theta) \geqslant \int \frac{1}{2}\|x\|^2 \,\mathrm{d}\nu_\theta(x) - \int \|x'\|^2 \,\mathrm{d}\nu_0(x') = \frac{1}{2}\operatorname{tr}\Sigma - \operatorname{tr}\Sigma_0 \tag{5.14}$$

as claimed. □

Lemma 5.9 will next be used in order to bound the norm of the end-to-end matrix.

Lemma 5.10 (Boundedness for the loss L_τ^1). *The norm of the end-to-end matrix W is upper-bounded when using the loss L_τ^1* (5.11b).

Proof. With $\varphi_\tau(\Sigma) = \Sigma + \tau \mathrm{I}_n$ and $\pi(W) = WW^\top$, the loss L_τ^1 satisfies

$$\begin{aligned} & L_\tau^1(W) = L(\varphi_\tau(\pi(W))) \geqslant \frac{1}{2}\operatorname{tr}\left(\pi(W) + \tau \mathrm{I}_n\right) - \operatorname{tr}\Sigma_0 = \frac{1}{2}\operatorname{tr} WW^\top - \operatorname{tr}\Sigma_0 + \frac{n}{2}\tau \\ \implies \quad & \sqrt{2L_\tau^1(W) + 2\operatorname{tr}\Sigma_0 - n\tau} \geqslant \|W\|_{\mathrm{F}}. \end{aligned}$$

Therefore, there exists an increasing function f such that $\|W\|_{\mathrm{F}} \leqslant f(L_\tau^1(W))$. Since the loss decreases under gradient flow, one has

$$\|W(t)\|_{\mathrm{F}} \leqslant \sqrt{2L_\tau^1(W(0)) + 2\operatorname{tr}\Sigma_0 - n\tau}, \tag{5.15}$$

and the boundedness of $t \mapsto W(t)$ is shown. □

Taking $\tau = 0$ in Lemma 5.10 gives the next corollary.

Corollary 5.11 *For the Bures-Wasserstein loss L, if $WW^\top$ is positive definite, so that the loss is differentiable, then, the norm of the end-to-end matrix $W(t) = \mu(\theta(t))$ is uniformly bounded throughout the flow:*

$$\forall t \in [0, \infty), \quad \|W(t)\|_{\mathrm{F}} \leqslant \sqrt{2L^1(W(0)) + 2\operatorname{tr}\Sigma_0}, \tag{5.16}$$

Proposition 5.12 *The gradient flow* (5.4) *on the perturbative loss* (5.11c) *converges to a critical point θ^* of L_τ^N.*

Proof. This is a direct consequence of Theorem 5.7 and Lemma 5.10. □

This property of the gradient flow is necessary in order to prove the convergence of the training to a minimizer of L_τ^1. At first glance, there is no immediate reason to expect that the critical points of $L_\tau^N = L_\tau^1 \circ \mu$ correspond to critical points of L_τ^1, since the parametrization μ could introduce superfluous critical points for L_τ^N. This last aspect led [TKB20] to distinguish between the *pure* and *spurious* critical points of a linear network; i.e. points that are critical for both $\mathcal{L}^N$ and $\mathcal{L}^1$, and those that are critical only for $\mathcal{L}^N$, and study conditions under which spurious local minima can be excluded.

We turn now to the convergence analysis under gradient flow and gradient descent of the (regularized) Bures-Wasserstein loss.

5.4 Convergence analysis

Equipped with the bounds derived in the previous section, we are able to analyze the convergence properties of the training algorithms presented in Section 5.2 for the Bures-Wasserstein loss (5.10c). The gradient flow convergence will require a full-rank assumption on the covariance matrix.

5.4.1 Gradient flow convergence

The Bures-Wasserstein distance can be viewed through the lens of the Procrustes metric [DKZ09; MPZ19]. In fact, it can be obtained by the following minimization problem.

Lemma 5.13 ([BJL19, Theorem 1]). *For $(\Sigma, \Sigma_0) \in \mathbb{S}_+(n)^2$,*

$$\mathcal{B}^2(\Sigma, \Sigma_0) = \min\left\{\|\Sigma^{1/2} - \Sigma_0^{1/2} U\|_{\mathrm{F}}^2 \mid U \in \mathbb{O}(n)\right\}, \tag{5.17}$$

where $\mathbb{O}(n)$ denotes the set of $n \times n$ orthogonal matrices. Moreover, any minimizer $\bar{U}$ is the orthogonal element of a polar decomposition of $\Sigma^{1/2}\Sigma_0^{1/2}$: there exists $WSV^\top$ a SVD of $\Sigma^{1/2}\Sigma_0^{1/2}$ so that $\bar{U} = WV^\top$.

We emphasize that in the above description of the Bures-Wasserstein distance, a minimizer $\bar{U}$ (is not unique and) depends on Σ, so that $\mathcal{B}^2$ fundamentally differs from a squared Frobenius norm. Moreover, the square root on Σ can lead to singularities when differentiating the loss. Nonetheless, based on (5.17), we can formulate the following deficiency margin concept to avoid such singularities.

Definition 5.14 (*Modified deficiency margin* (*MDM*)). Given a target matrix $\Sigma_0 \in \mathbb{S}_{++}(n)$ and a positive constant $c \in (0, \infty)$, we say that $\Sigma \in \mathbb{S}_+(n)$ has a c-modified-deficiency margin (c-MDM) with respect to Σ_0 if

$$\min\left\{\|\Sigma^{1/2} - \Sigma_0^{1/2} U\|_{\mathrm{F}} \mid U \in \mathbb{O}(n)\right\} \leqslant \sigma_{\min}(\Sigma_0^{1/2}) - c. \tag{5.18}$$

Recall the parametrizations μ (5.1) and π (5.9) . With a slight abuse of terminology, we will say that $W \in \mathbb{M}(n, m)$ (resp. $\theta \in \Theta$) has a modified-deficiency margin if $\pi(W)$ (resp. $\pi(\mu(\theta))$) does. The deficiency margin idea can be traced back to [ACG+19].

If we initialize the parameters so that Σ is close to the target Σ_0, then (5.18) holds trivially. In fact, if the initial value $W(0)$ satisfies the modified-deficiency margin condition, then the least singular value of $W(t)$ remains bounded below by c for all $t \in [0, \infty)$, along a gradient flow trajectory with decreasing L^N. This is stated by the next lemma.

Lemma 5.15 *Let $L^N \colon \Theta \to \mathbb{R}$ be the BW loss* (5.10c)*. Let $\theta \colon [0, \infty) \to \Theta$ be the path in parameter space under gradient flow* (5.4) *of L^N, and denote $\Sigma = \pi \circ \mu \circ \theta \colon [0, \infty) \to \mathbb{S}_+(n)$ the corresponding covariance trajectory. Suppose $\Sigma(0)$ has a modified deficiency margin c with respect to Σ_0. Then,*

$$\forall t \in [0, \infty), \quad \lambda_{\min}\Big(\Sigma^{1/2}(t)\Big) \geqslant c. \tag{5.19}$$

The proof of this and all results in this section are provided in Section 5.5. We note that, while the modified deficiency margin assumption is sufficient for Lemma 5.15 to hold, it is by no means necessary. We will assume that the modified deficiency margin assumption holds for simplicity of exposition, but the gradient flow analysis in the next paragraph only requires the less restrictive Lemma 5.15 to hold.

By Lemma 5.15, if initialized with a modified deficiency margin $c \in (0, \infty)$, the path in the covariance space $\Sigma = \pi \circ \mu \circ \theta \colon [0, \infty) \to \mathbb{S}_+(n)$, corresponding to a given parameter path $\theta \colon [0, \infty) \to \Theta$ under the gradient flow of L^N, will stay inside

$$\mathbb{S}_{++}(n; c^2) := \{A \in \mathbb{S}_{++}(n) \mid A \succcurlyeq c^2 \mathrm{I}_n\}.$$

This in turn makes L strongly-convex along the trajectory Σ, as stated next.

Lemma 5.16 *Let $\theta \colon [0, \infty) \to \Theta$ be a parameter trajectory and let $\Sigma = \pi \circ \mu \circ \theta \colon [0, \infty) \to \mathbb{S}_+(n)$ be the corresponding path in the covariance space. Assume that there exists $\varepsilon \in (0, \infty)$ such that, for all $t \in [0, \infty)$, $\Sigma(t) \succcurlyeq \varepsilon \mathrm{I}_n$ (i.e., $\Sigma \colon [0, \infty) \to \mathbb{S}_{++}(n; \varepsilon)$). Let*

$$K_\varepsilon := \frac{\lambda_{\min}^{1/2}(\Sigma_0)}{2\varepsilon^{3/2}}. \tag{5.20}$$

Then, for all $t \in [0, \infty)$, the Hessian operator $\mathcal{G}(t)$ of the loss L at $\Sigma(t)$, as defined by (4.32)*, satisfies*

$$\lambda_{\min}(\mathcal{G}(t)) \geqslant K_\varepsilon.$$

This is proven in Section 5.5.2. The spectrum of the loss' Hessian at positive matrices is therefore lower-bounded. Together with Lemma 5.15, this has the following direct implication.

Corollary 5.17 *Let $\theta\colon [0,\infty) \to \Theta$ be a parameter path under the gradient flow* (5.4) *of the loss L^N* (5.10c)*, and let $c \in (0,\infty)$. If $\theta(0)$ has a c-MDM* (*Definition 5.14*)*, the loss L* (5.10a) *is strongly convex along the trajectory $\Sigma = \pi \circ \mu \circ \theta\colon [0,\infty) \to \mathbb{S}_{++}(n; c^2)$ with constant $K_{c^2} = \frac{\lambda_{\min}^{1/2}(\Sigma_0)}{2c^3}$.*

Equipped with the strong convexity of the loss L, under the c-MDM assumption, given by Corollary 5.17, we are ready to show that the gradient flow (5.4) of the loss L^N (5.10c) has convergence rate $O(e^{-K'_{c,N}K_c t})$ to the global minimizer of L, where K_c is the constant from the Hessian bound given by Lemma 5.16, and $K'_{c,N}$ only depends on the modified-margin deficiency c and the depth of the network N.

The following theorem is the main result of this part.

Theorem 5.18 (Convergence of gradient flow for the BW loss). *Let $\theta\colon [0,\infty) \to \Theta$ be a path in the parameter space under the gradient flow* (5.4) *for the loss L^N* (5.10c)*, and let $\Sigma = \pi \circ \mu \circ \theta\colon [0,\infty) \to \mathbb{S}_+(n)$ be the corresponding trajectory in the covariance space. Let $c \in (0,\infty)$. Assume both balancedness* (*Definition 5.2*) *and c-modified deficiency margin* (*Definition 5.14*) *for $\theta(0)$. Let $\Delta^* := \Sigma(0) - \Sigma^*$ be the distance from the initialization to the minimizer $\Sigma^* := \operatorname{argmin}\{L(\Sigma) \mid \Sigma \in \mathbb{S}_{++}(n)\} = \Sigma_0$. Then, one has the following convergence rate*

$$L(\Sigma(t)) - L(\Sigma^*) \leqslant e^{-4\lambda_{\min}^{1/2}(\Sigma_0) N c^{\frac{N-2}{N}} t} \Delta^*. \tag{5.21}$$

Remark 5.19. Theorem 5.18 shows that the depth of the network can accelerate the convergence of the gradient algorithms. Indeed, $c^{\frac{N-2}{N}} \sim c$ as $N \to \infty$ in (5.21), so that $e^{-4\lambda_{\min}^{1/2}(\Sigma_0) N c^{\frac{N-2}{N}} t} \sim e^{-4\lambda_{\min}^{1/2}(\Sigma_0) N c t}$ as $N \to \infty$. We test this asymptotic linear dependency of the rate in c and $\lambda_{\min}^{1/2}(\Sigma_0)$ experimentally in Section 5.6, as well as the predicted convergence acceleration as N grows.

Now, as we also study the regularized BW loss, we can state the convergence theorem if we consider the gradient flow on the regularized loss L_τ^N instead of the original one L^N. In this case, the regularized covariance matrix belongs to $\mathbb{S}_{++}(n;\tau)$, and the statement can be adapted as required.

Corollary 5.20 (Convergence of gradient flow for the smooth loss). *Let $\Delta_\tau^* := \Sigma_\tau(0) - \Sigma_\tau^*$ be the distance from the initialization to the minimizer $\Sigma_\tau^* := \operatorname{argmin}\{L_\tau(\Sigma) \mid \Sigma \in \mathbb{S}_{++}(n)\} = \Sigma_0 - \tau \mathrm{I}_n$. Assume both balancedness* (*Definition 5.2*) *and the modified deficiency margin* (*Definition 5.14*)*. Then, the gradient flow $\dot\theta(t) = -\nabla L_\tau^N(\theta(t))$ converges as*

$$L(\Sigma_\tau(t)) - L(\Sigma_\tau^*) \leqslant e^{-8Nc^{\frac{2(2N-1)}{N}} K_\tau t} \Delta_\tau^*, \tag{5.22}$$

where $K_\tau = \dfrac{\lambda_{\min}^{1/2}(\Sigma_0)}{2\tau^{3/2}}$ is the strong convexity parameter from Lemma 5.16.

This concludes the continuous-time-setting analysis of the BW loss minimization.

5.4.2 Convergence of gradient descent for the BW loss

We turn to the study of the gradient-descent (5.5) optimization for the BW loss. We adapt the theorems of [NRT21], where the analysis is performed in the square-loss-regression setting, to our specific loss. We show similar results in our generative setting. In particular, we also obtain an almost sure convergence of the gradient descent iterates to a critical point of the regularized Bures-Wasserstein loss. Furthermore, when assuming a bounded initialization, we show that this minimizer is a global minimizer for the loss restricted to a rank-bounded set of matrices. We do not adapt the statement in the shallow case (which affirms that, under suitable initialization, the minimizer found by gradient descent is a global minimizer across all possibly parametrized matrices). We leave the exact derivation of the rate of convergence for future work. We cover the case of a general $N \geqslant 2$. A more detailed analysis of the case $N = 2$ is also left for future work.

In order to prove convergence, we require the loss to be analytic. Due to the matrix square root in L^N (5.10c), the BW loss is *not* analytic everywhere. Moreover, we will not assume that the initialization of the weights satisfy the MDM property (Definition 5.14), in contrast to what was done in Section 5.4.1. Therefore, we will derive the convergence results in this section on the regularized loss L^N_τ (5.11c), for some $\tau \in (0, \infty)$. Moreover, the perfect balancedness property is not conserved by a gradient descent update. As discussed in Section 5.2.2, we rely in this case on the approximate balancedness (or δ-balancedness) of the weights instead, as defined by Definition 5.3.

All of the proofs of this section can be found in Section 5.5.5.

We first state the convergence of gradient descent towards a critical point of the loss.

Theorem 5.21 *Let $\overrightarrow{W} : \mathbb{N}_0 \to \Theta$ be a path in the parameter space under the gradient descent iterations (5.5) of L^N_τ (5.11c), with stepsize $\eta \in (0, \infty)$. Suppose that the initialization $\overrightarrow{W}(0)$ of the iterations has balancedness property constant $\alpha\delta$ for some $(\delta, \alpha) \in (0, \infty) \times [0, 1)$. Assume further that the stepsize η satisfies*

$$\eta \leqslant \frac{2(1-\alpha)\delta}{4L^N_\tau(\overrightarrow{W}(0)) + (1-\alpha)\delta B_\delta} \tag{5.23}$$

where

$$B_\delta := eNK_\delta^{N-1}\left(\frac{\|\Sigma_0\|\left(eK_\delta^N + \tau\right)eK_\delta^N}{\tau^2} + 4\sqrt{n}\left[1 + \frac{\sqrt{\tau\lambda_{\min}(\Sigma_0)}\lambda_{\max}(\Sigma_0)}{\tau\lambda_{\min}(\Sigma_0)}\right]\right), \tag{5.24a}$$

$$K_\delta := M^{2/N} + (N+1)\delta^2, \tag{5.24b}$$

$$M := \sqrt{2\left(L^N_\tau(\overrightarrow{W}(0)) + \|\Sigma_0^{1/2}\|_{\mathrm{F}}^2\right)}. \tag{5.24c}$$

Then, the sequence $\left(\overrightarrow{W}(k)\right)_{k\in\mathbb{N}_0}$ *converges to a critical point of* L_τ^N.

This theorem only ensures that the iterates converge to a critical point of the loss. With an additional assumption on the initial point, we can refine the statement and show the convergence to a global minimizer from almost every initialization. In this regard, we set a bounded region in the parameter space $\mathcal{B} \subsetneq \Theta$ and let

$$\delta_{\mathcal{B}} = \sup_{\overrightarrow{W}\in\mathcal{B}} \max_{j\in[N-1]} \left\| W_{j+1}^\top W_{j+1} - W_j W_j^\top \right\|, \tag{5.25a}$$

$$L_{\mathcal{B}} = \sup_{\overrightarrow{W}\in\mathcal{B}} L_\tau^N(\overrightarrow{W}), \tag{5.25b}$$

$$M_{\mathcal{B}} = \sqrt{2\left(L_{\mathcal{B}} + \|\Sigma_0^{1/2}\|_{\mathrm{F}}^2\right)}. \tag{5.25c}$$

The next theorem ensures convergence towards specific critical points; namely, minimizers of the loss over some rank-constrained-matrices sets.

Theorem 5.22 *Let* $\mathcal{B} \subsetneq \Theta = \bigtimes_{j\in[N]} \mathbb{M}(d_j, d_{j-1})$ *be a bounded set, and let* $(\delta_{\mathcal{B}}, L_{\mathcal{B}}, M_{\mathcal{B}})$ *be defined as in* (5.25)*, so that* $\delta_{\mathcal{B}} \leqslant \alpha\delta$ *for some* $(\delta, \alpha) \in (0,\infty)\times[0,\infty)$*. Let* $\underline{d} = \min_{j\in[0:N]}\{d_j\}$*. Let* $\eta \in (0,\infty)$ *be a stepsize such that*

$$\eta \leqslant \frac{2(1-\alpha)\delta}{4L_{\mathcal{B}} + (1-\alpha)\delta B_\delta},$$

where

$$B_\delta := eNK_\delta^{N-1}\left(\frac{\|\Sigma_0\|\left(eK_\delta^N + \tau\right)eK_\delta^N}{\tau^2} + 4\sqrt{n}\left[1 + \frac{\sqrt{\tau\lambda_{\min}(\Sigma_0)}\lambda_{\max}(\Sigma_0)}{\tau\lambda_{\min}(\Sigma_0)}\right]\right),$$

$$K_\delta := M_{\mathcal{B}}^{2/N} + (N+1)\delta^2.$$

Consider the path $\overrightarrow{W}\colon \mathbb{N}_0 \to \Theta$ *under gradient descent* (5.5) *of the loss* L_τ^N *with stepsize* η*. Then, for almost all initializations* $\overrightarrow{W}(0) \in \mathcal{B}$*, the iterates converge to a critical points* $\overrightarrow{W}^*$ *of* L_τ^N *such that* $W^* = \mu(\overrightarrow{W}^*)$ *is a global minimum of* L_τ^1 (5.11b) *on the manifold* $\mathcal{M}_k \subseteq \mathbb{M}(d_N, d_0)$ *of matrices of rank* $k = \operatorname{rank}(W^*) \in [0:\underline{d}]$.

In the function and covariance spaces, the rank of the critical point we converge to is upper-bounded by the width of the smallest layer in the network. It is left to future work to understand what exactly the value for this rank is; notably, if the highest possible rank is achieved or not.

5.5 Proofs of convergence results in Section 5.4

In this section, we provide the proofs of the gradient flow (Section 5.5.4) and the gradient descent (Section 5.5.5) statements (Theorems 5.18, 5.21 and 5.22). They mostly rely on bounds of the Hessian of the loss (Section 5.5.2). First, we provide the proof of Lemma 5.15.

5.5.1 Proof of Lemma 5.15

We start by proving Lemma 5.15, stating that with the modified-margin deficiency assumption on the initial weights, the parametrized covariance does not degenerate along the gradient flow (5.4) training algorithm. In fact, its minimal eigenvalue does not degrade along any path $W\colon [0,\infty) \to \mathbb{M}(n,m)$ such that $L^1 \circ W\colon [0,\infty) \to \mathbb{R}$ is decreasing, which is in particular the case when gradient flow or gradient descent updates are considered, as per (5.13).

Proof of Lemma 5.15. Let $\theta\colon [0,\infty)$ be the path under gradient flow (5.4) of the loss L^N (5.10c). Let $W = \mu \circ \theta\colon [0,\infty) \to \mathbb{M}(n,m)$ and $\Sigma = \pi \circ W\colon [0,\infty) \to \mathbb{S}_+(n)$ be the corresponding trajectories in the function and covariance spaces. Denote $\bar{U}(t) \in \operatorname{argmin}\left\{ \|\Sigma^{1/2}(t) - \Sigma_0^{1/2} U\|_{\mathrm{F}}^2 \mid U \in \mathbb{O}(n) \right\}$ for each $t \in [0,\infty)$.

Since, in this setting, $L^1 \circ W\colon [0,\infty) \to \mathbb{R}$ is decreasing (Lemma 5.8), $L^1(W(t)) \leqslant L^1(W(0))$ for all $t \in [0,\infty)$. Therefore, we have

$$
\begin{aligned}
\lambda_{\min}\Big(\Sigma^{1/2}(t)\Big) &= \lambda_{\min}\Big(\Sigma^{1/2}(t) - \Sigma_0^{1/2}\bar{U}(t) + \Sigma_0^{1/2}\bar{U}(t)\Big) \\
&\geqslant \lambda_{\min}\Big(\Sigma_0^{1/2}\bar{U}(t)\Big) - \lambda_{\max}\Big(\Sigma^{1/2}(t) - \Sigma_0^{1/2}\bar{U}(t)\Big) \\
&\geqslant \lambda_{\min}\Big(\Sigma_0^{1/2}\bar{U}(t)\Big) - \Big\|\Sigma^{1/2}(t) - \Sigma_0^{1/2}\bar{U}(t)\Big\|_{\mathrm{F}} \\
&= \lambda_{\min}\Big(\Sigma_0^{1/2}\bar{U}(t)\Big) - \sqrt{L^1(W(t))} \\
&\geqslant \lambda_{\min}\Big(\Sigma_0^{1/2}\bar{U}(t)\Big) - \sqrt{L^1(W(0))} \\
&= \lambda_{\min}\Big(\Sigma_0^{1/2}\bar{U}(t)\Big) - \Big\|\Sigma^{1/2}(0) - \Sigma_0^{1/2}\bar{U}(0)\Big\|_{\mathrm{F}} \\
&\geqslant \lambda_{\min}\Big(\Sigma_0^{1/2}\bar{U}(t)\Big) - \lambda_{\min}\Big(\Sigma_0^{1/2}\Big) + c = c.
\end{aligned}
$$

The cancellation in the last equality works due to the fact that the multiplication with an arbitrary unitary matrix does not change singular values. □

5.5.2 Bounds on the Hessian of L

The aim of this section is to provide the bounds relative to the Hessian operator $\mathcal{G}$ (4.32) of the BW loss L (5.10a).

Definition 5.23 (*Eigenvalues of matrix operators*). Let $\mathcal{F}\colon \mathbb{M}(p,q) \to \mathbb{M}(p,q)$ be a linear operator. Then, its extremal eigenvalues $\lambda_{\{\min,\max\}}(\mathcal{F})$ are defined as

$$\lambda_{\min}(\mathcal{F}) := \inf\{\langle U, \mathcal{F}(U)\rangle \mid U \in \mathbb{M}(p,q),\ \|U\|_{\mathrm{F}} = 1\}, \tag{5.26a}$$
$$\lambda_{\max}(\mathcal{F}) := \sup\{\langle U, \mathcal{F}(U)\rangle \mid U \in \mathbb{M}(p,q),\ \|U\|_{\mathrm{F}} = 1\}. \tag{5.26b}$$

One can use the bounds of [KSS21, Lemma A.3] in order to bound the Hessian of the loss.

Lemma 5.24 (Bounds on the second-order differential, [KSS21, Lemma A.3]). *Let $\Sigma \in \mathbb{S}_{++}(n)$, and let $\mathcal{G}$ be defined as in* (4.32). *Then, for all $X \in \mathbb{S}_{++}(n)$, the following bounds hold*

$$\langle X, \mathcal{G}(X)\rangle \geqslant \frac{\lambda_{\min}^{1/2}(\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2})}{2}\|\Sigma^{-1/2}X\Sigma^{-1/2}\|_{\mathrm{F}}^2, \tag{5.27a}$$
$$\langle X, \mathcal{G}(X)\rangle \leqslant \frac{\lambda_{\max}^{1/2}(\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2})}{2}\|\Sigma^{-1/2}X\Sigma^{-1/2}\|_{\mathrm{F}}^2. \tag{5.27b}$$
$$\tag{5.27c}$$

Equations (5.27a) and (5.27b) in turn bound the extremal eigenvalues, as defined in Definition 5.23, of the Hessian operator $\mathcal{G}$ of L at $\Sigma \in \mathbb{S}_{++}(n)$. The bound requires that the eigenvalues of Σ are bounded away from zero. Therefore, for any $\varepsilon \in (0,\infty)$, let

$$\mathbb{S}_{++}(n;\varepsilon) := \{A \in \mathbb{S}_{++}(n) \mid A \succcurlyeq \varepsilon \mathrm{I}_n\} \tag{5.28}$$

be the set of ε-strictly positive definite matrices.

For $c \in (0,\infty)$, under the c-modified-deficiency margin assumption (Definition 5.14) at initialization, the parametrized covariance matrix will belong to $\mathbb{S}_{++}(n;c^2)$ along the gradient flow trajectory, as per Lemma 5.15.

Lemma 5.25 (Spectral bounds of the Hessian on $\mathbb{S}_{++}(n;\varepsilon)$). *Let $\varepsilon \in (0,\infty)$, and let $\Sigma \in \mathbb{S}_{++}(n;\varepsilon)$. Let $\mathcal{G}$ be the Hessian operator of L at Σ, as defined by* (4.32). *Then, its extremal eigenvalues are bounded as*

$$\frac{\sqrt{\varepsilon\lambda_{\min}(\Sigma_0)}}{2\varepsilon^2} \leqslant \lambda_{\min}(\mathcal{G}) \leqslant \lambda_{\max}(\mathcal{G}) \leqslant \frac{\sqrt{C\lambda_{\max}(\Sigma_0)}}{2\varepsilon^2}, \tag{5.29}$$

where $C := 2(L(\Sigma) + \mathrm{tr}(\Sigma_0))$.

Proof. We first provide the proof for the minimal eigenvalue. By Equation (5.26a), the maximal eigenvalue of $\mathcal{G}$ is defined by

$$\lambda_{\min}(\mathcal{G}) = \inf \{\langle X, \mathcal{G}(X)\rangle \mid X \in \mathbb{S}_{++}(n),\ \|X\|_{\mathrm{F}} = 1\}.$$

From the upper-bound of $\langle X, \mathcal{G}(X)\rangle$ in (5.27a), one has

$$\begin{aligned}
\inf_{X:\|X\|_{\mathrm{F}}=1} \langle X, \mathcal{G}(X)\rangle &\geqslant \inf_{X:\|X\|_{\mathrm{F}}=1} \frac{\lambda_{\min}^{1/2}(\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2})}{2} \|\Sigma^{-1/2} X \Sigma^{-1/2}\|_{\mathrm{F}}^2 \\
&= \frac{\lambda_{\min}^{1/2}(\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2})}{2} \inf_{X:\|X\|_{\mathrm{F}}=1} \|\Sigma^{-1/2} X \Sigma^{-1/2}\|_{\mathrm{F}}^2 \\
&= \frac{\lambda_{\min}^{1/2}(\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2})}{2} \inf_{X:\|X\|_{\mathrm{F}}=1} \|\Sigma^{-1} X\|_{\mathrm{F}}^2 \\
&= \frac{\lambda_{\min}^{1/2}(\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2})}{2} \lambda_{\min}^2(\Sigma^{-1}) \\
&\geqslant \frac{\lambda_{\min}^{1/2}(\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2})}{2\varepsilon^2}.
\end{aligned} \tag{5.30}$$

The last inequality is implied by the fact that

$$\Sigma \in \mathbb{S}_{++}(n;\varepsilon) \implies \lambda_{\min}(\Sigma^{-1}) = \lambda_{\max}^{-1}(\Sigma) > \varepsilon^{-1}.$$

Now, for any positive definite matrices $(A, B) \in (\mathbb{S}_{++}(n))^2$ with *increasing* eigenvalues and for any $k \in [n]$, we know that

$$\lambda_k(A)\lambda_1(B) \leqslant \lambda_k(AB) = \lambda_k(A^{1/2}BA^{1/2}) \leqslant \lambda_k(A)\lambda_n(B). \tag{5.31}$$

Taking $k = 1$ and $(A, B) = (\Sigma_0, \Sigma)$ provides the bound

$$\lambda_{\min}^{1/2}(\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2}) \geqslant \lambda_{\min}^{1/2}(\Sigma_0)\lambda_{\min}^{1/2}(\Sigma)$$

Moreover, $\lambda_{\min}(\Sigma) \geqslant \varepsilon$, so that (5.30) becomes

$$\lambda_{\min}(\mathcal{G}) \geqslant \frac{\sqrt{\varepsilon\lambda_{\min}(\Sigma_0)}}{2\varepsilon^2} = \frac{\lambda_{\min}^{1/2}(\Sigma_0)}{2\varepsilon^{3/2}}.$$

The proof for the minimal eigenvalue is similar and follows from the bound (5.27b). One derives

$$\sup_{X:\|X\|_F=1} \langle X, \mathcal{G}(X)\rangle \leqslant \frac{\lambda_{\max}(\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2})}{2\varepsilon^2} \tag{5.32}$$

Now, by choosing $(A, B) = (\Sigma_0, \Sigma)$ and $k = n$ in (5.31), and taking the square root, one gets the bound

$$\lambda_{\max}^{1/2}(\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2}) \leqslant \lambda_{\max}^{1/2}(\Sigma)\lambda_{\max}^{1/2}(\Sigma_0).$$

Moreover, $\lambda_{\max}(\Sigma) \leqslant \operatorname{tr}\Sigma$, and from Lemma 5.9, we know that $\operatorname{tr}\Sigma \leqslant 2(L(\Sigma) - L(\Sigma_0)) =: C$. Plugging it into (5.32), we obtain

$$\lambda_{\max}(\mathcal{G}) \leqslant \frac{\sqrt{C\lambda_{\max}(\Sigma_0)}}{2\varepsilon^2},$$

finishing the proof. □

Corollary 5.26 *Let $\varepsilon \in (0,\infty)$. The loss L is K_ε-strongly convex on $\mathbb{S}_{++}(n;\varepsilon)$, with $K_\varepsilon = \frac{\lambda_{\min}^{1/2}(\Sigma_0)}{2\varepsilon^{3/2}}$*

Proof. Let $\varepsilon \in (0,\infty)$. We show that $\mathbb{S}_{++}(n;\varepsilon)$ is convex. Let A and B in $\mathbb{S}_{++}(n,\varepsilon)$, and let $\alpha \in (0,1)$. For any $x \in \mathbb{R}^n \setminus \{0\}$, $\langle x, (\alpha A + (1-\alpha)B)x\rangle = \alpha\langle x, Ax\rangle + (1-\alpha)\langle x, Bx\rangle > 0$, so that $\alpha A + (1-\alpha)B \in \mathbb{S}_{++}(n)$. Moreover, $\lambda_{\min}(\alpha A + (1-\alpha)B) = \inf\{\langle x, (\alpha A + (1-\alpha)B)x\rangle \mid \|x\| = 1\} = \inf\{\alpha\langle x, Ax\rangle + (1-\alpha)\langle x, Bx\rangle \mid \|x\| = 1\} \geqslant \alpha \inf\{\langle x, Ax\rangle \mid \|x\| = 1\} + (1-\alpha)\inf\{\langle x, Bx\rangle \mid \|x\| = 1\} = \alpha\lambda_{\min}(A) + (1-\alpha)\lambda_{\min}(B) \geqslant \varepsilon$. Therefore, $\alpha A + (1-\alpha)B \in \mathbb{S}_{++}(n,\varepsilon)$, and $\mathbb{S}_{++}(n,\varepsilon)$ is a convex set.

On $\mathbb{S}_{++}(n,\varepsilon)$, the Hessian of the loss L has its spectrum lower-bounded by $K_\varepsilon = \frac{\sqrt{\varepsilon\lambda_{\min}(\Sigma_0)}}{2\varepsilon^2}$, as stated in Lemma 5.25. Hence, L is K_ε-strongly convex on $\mathbb{S}_{++}(n;\varepsilon)$. □

The lower-bound derived in Lemma 5.25 allows proof Lemma 5.16.

Proof of Lemma 5.16. Under the c-MDM hypothesis for $\theta(0)$, Lemma 5.15 ensures that $\Sigma(t) \in \mathbb{S}_{++}(n;c^2)$ for all $t \in [0,\infty)$. Therefore, by Lemma 5.25, for all $t \in [0,\infty)$, $\lambda_{\min}(\mathcal{G}(t)) \geqslant K_{c^2} = \frac{\lambda_{\min}^{1/2}(\Sigma_0)}{2c^3}$. □

Remark 5.27. For $\tau \in (0,\infty)$ and $\Sigma \in \mathbb{S}_+(n)$, the associated regularized covariance matrix $\Sigma_\tau = \Sigma + \tau \mathrm{I}_n$ belongs to $\mathbb{S}_{++}(n;\tau)$. Therefore, Lemma 5.25 and corollary 5.26 can be stated for L_τ on $\mathbb{S}_+(n)$, with ε replaced by τ.

The strong convexity of the loss on positive matrices will be the cornerstone for proving the upper-bound on the rate of convergence of gradient flow in Section 5.5.4. Before that, we show that the loss L_τ^1 is Lipschitz smooth.

5.5.3 Hessian of L_τ^1

In this section, we provide the expressions on the Hessian of the Bures-Wasserstein loss in the deep-linear-network parametrization L_τ^1 (5.11b) on $\mathbb{M}(n,m)$.

Recall the second-order differential of L on $\mathbb{S}_{++}(n)$ given in Lemma 4.38. With the computation rule in Lemma 4.30, we are able to give the second-order differential of $L^1 = L \circ \pi$.

Since the BW loss is only continuously differentiable on $\mathbb{S}_{++}(n)$, we will consider the regularization introduced in Section 4.2.5 and eq. (5.11).

Let $\tau \in (0,\infty)$. The Lemma 4.38 can equivalently be stated for L at $\Sigma_\tau = \Sigma + \tau \mathrm{I}_n \in \mathbb{S}_{++}(n)$, or for L_τ on $\mathbb{S}_+(n)$. In the latter case, we will write the associated Hessian operator $\mathcal{G}_\tau$.

Lemma 5.28 (Second-order differential of L_τ^1). *Let $W \in \mathbb{M}(n,m)$. For any $(U,V) \in \mathbb{M}(n,m)^2$, the second-order differential of L_τ^1 at W in the directions U and V is*

$$\mathrm{d}^2 L_\tau^1(W)[U,V] = \langle U, \mathcal{H}_\tau(V)\rangle,$$

where

$$\mathcal{H}_\tau(V) = 2(\mathcal{G}_\tau(VW^\top + WV^\top)W + (\mathrm{I}_n - \Sigma_0^{1/2}(\Sigma_0^{1/2}\Sigma_\tau\Sigma_0^{1/2})^{-1/2}\Sigma_0^{1/2})V), \tag{5.33}$$

and $\mathcal{G}_\tau$ is the Hessian operator (4.32) *associated to the loss L_τ.*

Proof. Applying the formula (4.28) to $L_\tau^1 = L_\tau \circ \pi$ gives, with $\Sigma = \pi(W)$ and $\mathrm{d}^2\pi(W)[U,V] = UV^\top + VU^\top$,

$$\begin{aligned}
\mathrm{d}^2 L_\tau^1(W)[U,V] &= \mathrm{d}^2 L_\tau(\Sigma)[\mathrm{d}\pi(W)[U], \mathrm{d}\pi(W)[V]] + \mathrm{d}L_\tau(\Sigma)[\mathrm{d}^2\pi(W)[U,V]] \\
&= \langle UW^\top + WU^\top, \mathcal{G}_\tau(VW^\top + WV^\top)\rangle + \operatorname{tr}(UV^\top + VU^\top) \\
&\quad - \operatorname{tr}\Sigma_0^{1/2}(\Sigma_0^{1/2}\Sigma_\tau\Sigma_0^{1/2})^{-1}\Sigma_0^{1/2}(UV^\top + VU^\top) \\
&= 2\langle U, \mathcal{G}_\tau(VW^\top + WV^\top)W + V - \Sigma_0^{1/2}(\Sigma_0^{1/2}\Sigma_\tau\Sigma_0^{1/2})^{-1}\Sigma_0^{1/2}V\rangle \\
&= \langle U, \mathcal{H}_\tau(V)\rangle,
\end{aligned}$$

where we used the symmetry of $\Sigma_0^{1/2}(\Sigma_0^{1/2}\Sigma_\tau\Sigma_0^{1/2})^{-1/2}\Sigma_0$ to simplify the expression. □

With the expression given by Lemma 5.28 and the bounds from Lemma 5.24, we bound next the spectrum of $\mathcal{H}_\tau$, the Hessian operator of L_τ^1.

Lemma 5.29 (Spectral bound of $\mathcal{H}_\tau$). *Let $\mathcal{H}_\tau$ be defined as in* (5.33). *The maximal eigenvalue for the Hessian of L_τ^1 respects the following bound*

$$\lambda_{\max}(\mathcal{H}_\tau) \leqslant \lambda_{\max}^{1/2}(\Sigma_0^{1/2}\Sigma_\tau\Sigma_0^{1/2})\frac{2C^2}{\tau^2} + 2(1 - \lambda_{\min}(\Sigma_0^{1/2}(\Sigma_0^{1/2}\Sigma_\tau\Sigma_0^{1/2})^{-1/2}\Sigma_0^{1/2})) \tag{5.34}$$

Proof. From (5.27b), one has for any $X \in \mathbb{S}_{++}(n)$,

$$\langle X, \mathcal{G}_\tau(X)\rangle \leqslant \frac{\lambda_{\max}^{1/2}(\Sigma_0^{1/2}\Sigma_\tau\Sigma_0^{1/2})}{2}\|\Sigma_\tau^{-1/2}X\Sigma_\tau^{-1/2}\|_{\mathrm{F}}^2.$$

Let $U \in \mathbb{M}(n,m)$. With $X(U) = UW^\top + WU^\top$, the bound becomes

$$\begin{aligned}
&\langle UW^\top + WU^\top, \mathcal{G}_\tau(X(U))\rangle \leqslant \frac{\lambda_{\max}^{1/2}(\Sigma_0^{1/2}\Sigma_\tau\Sigma_0^{1/2})}{2}\|\Sigma_\tau^{-1/2}X(U)\Sigma_\tau^{-1/2}\|_{\mathrm{F}}^2 \\
\iff \quad &2\langle UW^\top, \mathcal{G}_\tau(X(U))\rangle \leqslant \frac{\lambda_{\max}^{1/2}(\Sigma_0^{1/2}\Sigma_\tau\Sigma_0^{1/2})}{2}\|\Sigma_\tau^{-1/2}X(U)\Sigma_\tau^{-1/2}\|_{\mathrm{F}}^2 \\
\iff \quad &2\langle U, \mathcal{G}_\tau(X(U))W\rangle \leqslant \frac{\lambda_{\max}^{1/2}(\Sigma_0^{1/2}\Sigma_\tau\Sigma_0^{1/2})}{2}\|\Sigma_\tau^{-1/2}X(U)\Sigma_\tau^{-1/2}\|_{\mathrm{F}}^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\langle U, \mathcal{H}_\tau(U)\rangle &= 2\langle U, \mathcal{G}_\tau(X(U))W + (\mathrm{I}_n - \Sigma_0^{1/2}(\Sigma_0^{1/2}\Sigma_\tau\Sigma_0^{1/2})^{-1/2}\Sigma_0^{1/2})U\rangle \qquad (5.35)\\
&\leqslant \frac{\lambda_{\max}^{1/2}(\Sigma_0^{1/2}\Sigma_\tau\Sigma_0^{1/2})}{2}\|\Sigma_\tau^{-1/2}X(U)\Sigma_\tau^{-1/2}\|_{\mathrm{F}}^2 \\
&\quad + 2\langle U, (\mathrm{I}_n - \Sigma_0^{1/2}(\Sigma_0^{1/2}\Sigma_\tau\Sigma_0^{1/2})^{-1/2}\Sigma_0^{1/2})U\rangle.
\end{aligned}$$

We proceed by bounding each of the summands.

First, consider the term $\|\Sigma_\tau^{-1/2}X(U)\Sigma_\tau^{-1/2}\|_{\mathrm{F}}^2 = \|\Sigma_\tau^{-1}X(U)\|_{\mathrm{F}}^2$. If U is such that $\|U\|_{\mathrm{F}} = 1$, then $\|X(U)\|_{\mathrm{F}}^2 = \|UW^\top + WU^\top\|^2 \leqslant 4\|W\|_{\mathrm{F}}^2$ by the triangle inequality and the sub-multiplicativity of $\|\cdot\|_{\mathrm{F}}$. From Equation (5.15), we know that $\|W\|_{\mathrm{F}} \leqslant C$ for some constant C.

Therefore, $\|U\|_{\mathrm{F}} = 1 \implies \|X(U)\| \leqslant 2C$ and

$$\begin{aligned}
&\sup\left\{\|\Sigma_\tau^{-1}X(U)\|_{\mathrm{F}}^2 \mid U \in \mathbb{M}(n,m),\ \|U\|_{\mathrm{F}} = 1\right\} \\
&\quad\leqslant \sup\left\{\|\Sigma_\tau^{-1}X(U)\|_{\mathrm{F}}^2 \mid X \in \mathbb{S}(n),\ \|X\|_{\mathrm{F}} \leqslant 2C\right\} \\
&\quad\leqslant \sup\left\{4C^2\|\Sigma_\tau^{-1}X\|_{\mathrm{F}}^2 \mid X \in \mathbb{S}(n),\ \|X\|_{\mathrm{F}} = 1\right\} \\
&\quad\leqslant 4C^2\lambda_{\max}^2(\Sigma_\tau^{-1}) = \frac{4C^2}{\tau^2}.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\sup\left\{\frac{\lambda_{\max}^{1/2}(\Sigma_0^{1/2}\Sigma_\tau\Sigma_0^{1/2})}{2}\|\Sigma_\tau^{-1/2}X(U)\Sigma_\tau^{-1/2}\|_{\mathrm{F}}^2 \mid U \in \mathbb{M}(n,m),\ \|U\|_{\mathrm{F}} = 1\right\} \\
&\leqslant \lambda_{\max}^{1/2}(\Sigma_0^{1/2}\Sigma_\tau\Sigma_0^{1/2})\frac{2C^2}{\tau^2}.
\end{aligned}$$

The second summation in (5.35) can be bounded as

$$\begin{aligned}
&\sup\left\{2\langle U, (\mathrm{I}_n - \Sigma_0^{1/2}(\Sigma_0^{1/2}\Sigma_\tau\Sigma_0^{1/2})^{-1/2}\Sigma_0^{1/2})U\rangle \mid U \in \mathbb{M}(n,m),\ \|U\|_{\mathrm{F}} = 1\right\} \\
&= 2\lambda_{\max}(\mathrm{I}_n - \Sigma_0^{1/2}(\Sigma_0^{1/2}\Sigma_\tau\Sigma_0^{1/2})^{-1/2}\Sigma_0^{1/2}) \\
&= 2(1 - \lambda_{\min}(\Sigma_0^{1/2}(\Sigma_0^{1/2}\Sigma_\tau\Sigma_0^{1/2})^{-1/2}\Sigma_0^{1/2})).
\end{aligned}$$

□

Corollary 5.30 (Lipschitz-smoothness of L_τ^1). *For $\tau \in (0,\infty)$, the loss $W \mapsto L_\tau^1(W)$ is Lipschitz smooth.*

Proof. This directly follows from the boundedness of the Hessian showed previously and the convexity of L_τ^1 using Taylor approximation. □

Once the Lipschitz-smoothness of the loss has been proven, one can turn to showing that the rank is preserved under balanced initial conditions.

Proposition 5.31 ([BRT+21, Proposition 4.4]). *Let $\mathcal{L}^1 \colon \mathbb{M}(n,m) \to \mathbb{R}$ be a Lipschitz smooth function (i.e., a differentiable function with Lipschitz gradient). Let $t_0 \in [0,\infty)$. Suppose that $W_1(t), \dots, W_N(t)$ are solutions of the gradient flow (5.4) of L^N with balanced initial values $W_j(t_0)$ and define the product $W(t) = \mu(\theta(t)) = W_N(t)\cdots W_1(t)$. If $W(t_0)$ is contained in $\mathcal{M}(k)$ for some $k \in \mathbb{N}_0$, then $W(t)$ is contained in $\mathcal{M}(k)$ for all $t \in [t_0,\infty)$.*

Remark 5.32. The loss L_τ^1 satisfies the conditions of Proposition 5.31; therefore, the flow on L_τ^1 remains in the manifold $\mathcal{M}(k)$ for all $t \in [t_0,\infty)$ if $W(t_0) \in \mathcal{M}(k)$ for some $t_0 \in [0,\infty)$.

5.5.4 Proof of gradient flow convergence

We first prove here the convergence of the gradient flow of L^N to a parameter corresponding to a covariance matrix that is a global minimizer of L_τ, under the assumptions of balanced weights (Definition 5.2) and modified deficiency margin (Definition 5.14). Then, in Corollary 5.20, we state the theorem for the regularized loss, under the same assumptions on the weights.

Proof of Theorem 5.18. Let $c \in (0,\infty)$, and let $\theta \colon [0,\infty) \to \Theta$ be a path under gradient flow (5.4) of L^N (5.10c). The idea of the proof is to transfer the strong convexity property from L on $\mathbb{S}_{++}(n;c^2)$ to the evolution of the parameters. Indeed, assuming that $\theta(0)$ has

c-MDM, Lemma 5.15 ensures that the parametrized covariances $\Sigma = \pi \circ \mu \circ \theta$ remain inside $\mathbb{S}_{++}(n; c^2)$ onto which, by Lemma 5.16, the loss L is K_{c^2}-strongly-convex with

$$K_{c^2} = \frac{\lambda_{\min}^{1/2}(\Sigma_0)}{2c^3}. \tag{5.36}$$

Hence, for any $t \in (0, \infty)$, the following inequality holds

$$L(\Sigma(t)) - L(\Sigma^*) \leqslant \frac{1}{2K_{c^2}} \|\nabla L(\Sigma(t))\|_{\mathrm{F}}^2,$$

The dependency on t will be implied in the following in order to lighten the notation. Rearranging the terms in the above equation, we have

$$-\|\nabla L(\Sigma)\|_{\mathrm{F}}^2 \leqslant -2K_{c^2} \left(L(\Sigma) - L(\Sigma^*)\right). \tag{5.37}$$

On the covariance space, for the regularized loss, the gradient flow is written

$$\begin{aligned}
\frac{\mathrm{d}L(\Sigma)}{\mathrm{d}t} &= \langle \nabla L(\Sigma), \frac{\mathrm{d}}{\mathrm{d}t}\Sigma(t) \rangle \\
&= \langle \nabla_\Sigma L(\Sigma), W \frac{\mathrm{d}W}{\mathrm{d}t}^\top + \frac{\mathrm{d}W}{\mathrm{d}t} W^\top \rangle \\
&= 2 \langle \nabla_\Sigma L(\Sigma) W, \frac{\mathrm{d}W}{\mathrm{d}t} \rangle.
\end{aligned}$$

The expression of $\frac{\mathrm{d}W}{\mathrm{d}t}$ is given by Lemma 5.1:

$$\frac{\mathrm{d}W}{\mathrm{d}t} = -\sum_{\ell=1}^{N} W_{N:\ell+1} W_{N:\ell+1}^\top \nabla L^1(W) W_{\ell-1:1}^\top W_{\ell-1:1}.$$

Since $\nabla L^1(W) = 2\nabla L(\Sigma) W$, and from the balancedness assumption we have $W_{N:\ell+1} W_{N:\ell+1}^\top = (WW^\top)^{\frac{N-\ell}{N}}$ and $W_{\ell-1:1}^\top W_{\ell-1:1} = (W^\top W)^{\frac{\ell-1}{N}}$, we get

$$\frac{\mathrm{d}L(\Sigma(t))}{\mathrm{d}t} = -4 \sum_{\ell=1}^{N} \left\langle \nabla L(\Sigma) W, \left[WW^\top\right]^{\frac{N-\ell}{N}} \nabla L(\Sigma) W \left[W^\top W\right]^{\frac{\ell-1}{N}} \right\rangle.$$

Now, let $USV^\top = W$ be a (thin) SVD of W, so that $WW^\top = US^2U^\top$ and $W^\top W = VS^2V^\top$. For one layer $\ell \in [N]$, we then have

$$\begin{aligned}
&\left\langle \nabla L(\Sigma) W, \left[WW^\top\right]^{\frac{N-\ell}{N}} \nabla L(\Sigma) W \left[W^\top W\right]^{\frac{\ell-1}{N}} \right\rangle \\
&= \mathrm{tr}\left(\nabla L(\Sigma) W \left[W^\top W\right]^{\frac{\ell-1}{N}} W^\top \nabla L(\Sigma) \left[WW^\top\right]^{\frac{N-\ell}{N}} \right) \\
&= \mathrm{tr}\left(\nabla L(\Sigma) USV^\top VS^{\frac{2(\ell-1)}{N}} V^\top VSU^\top \nabla L(\Sigma) US^{\frac{2(N-\ell)}{N}} U^\top \right) \\
&= \mathrm{tr}\left(U^\top \nabla L(\Sigma) US^{\frac{2(N+\ell-1)}{N}} U^\top \nabla L(\Sigma) US^{\frac{2(N-\ell)}{N}} \right) \\
&= \left\langle U^\top \nabla L(\Sigma) US^{\frac{2(N+\ell-1)}{N}}, S^{\frac{2(N-\ell)}{N}} U^\top \nabla L(\Sigma) U \right\rangle.
\end{aligned}$$

Let $X := U^\top \nabla L(\Sigma) U$, $D := S^{\frac{2(N+\ell-1)}{N}}$, and $E := S^{\frac{2(N-\ell)}{N}}$. We evaluate $\langle XD, EX\rangle$ for the diagonal D, E as

$$\langle XD, EX\rangle = \sum_{i,j} X_{i,j} D_j X_{i,j} E_i = \sum_{i,j} E_i D_j X_{i,\ell}^2.$$

Since $E_i = s_i^{\frac{2(N-\ell)}{N}}$ and $D_j = s_j^{\frac{2(N+\ell-1)}{N}}$, and due to the modified-margin deficiency assumption, for all $(i,j) \in [k]^2$, we have $E_i \geqslant c^{\frac{2(N-\ell)}{N}}$ and $D_j \geqslant c^{\frac{2(N+\ell-1)}{N}}$, so that

$$\langle XD, EX\rangle \geqslant c^{\frac{2(2N-1)}{N}} \sum_{i,j} X_{i,j}^2 = c^{\frac{2(2N-1)}{N}} \|X\|_{\mathrm{F}}^2. \tag{5.38}$$

Since $X = U^\top \nabla L(\Sigma) U^\top$, we have that $\|X\|_{\mathrm{F}}^2 = \|\nabla L(\Sigma)\|_{\mathrm{F}}^2$, so that in total

$$\frac{\mathrm{d}}{\mathrm{d}t} L(\Sigma(t)) \leqslant -4 \sum_{\ell=1}^{N} c^{\frac{2(2N-1)}{N}} \|\nabla L(\Sigma)\|_{\mathrm{F}}^2 = -4Nc^{\frac{2(2N-1)}{N}} \|\nabla L(\Sigma)\|_{\mathrm{F}}^2.$$

From the strong convexity of L (5.37), we get the bound

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} L(\Sigma(t)) &\leqslant -8Nc^{\frac{2(2N-1)}{N}} K_{c^2} (L(\Sigma) - L(\Sigma^*)) \\ \implies \quad \frac{1}{L(\Sigma(t)) - L(\Sigma^*)} \frac{\mathrm{d}}{\mathrm{d}t} (L(\Sigma(t)) - L(\Sigma^*)) &\leqslant -8Nc^{\frac{2(2N-1)}{N}} K_{c^2}. \end{aligned}$$

Now, by integrating both sides from 0 to t,

$$\ln\left(\frac{L(\Sigma(t)) - L(\Sigma^*)}{L(\Sigma(0)) - L(\Sigma^*)}\right) \leqslant -8Nc^{\frac{2(2N-1)}{N}} K_{c^2} t.$$

Let $\Delta^* = \Sigma(0) - \Sigma^*$ be the distance from the initialization to optimality. Finally we get the desired exponential rate

$$L(\Sigma(t)) - L(\Sigma^*) \leqslant e^{-8Nc^{\frac{2(2N-1)}{N}} K_{c^2} t} \Delta^* = e^{-4\lambda_{\min}^{1/2}(\Sigma_0) Nc^{\frac{N-2}{N}} t} \Delta^*,$$

where we used the expression of K_{c^2} (5.36) to simplify the right-hand-side term. This concludes the proof. □

5.5.5 Proofs of gradient descent convergence

We now turn to the analysis of the finite-step-size gradient descent optimization of the BW loss. We first recall and slightly relax the bound for $\|W\|$ in terms of $L_\tau^1(W)$ provided by Lemma 5.10, adapting the bound given in [NRT21, Lemma 3.1] in the regression case with quadratic loss. In the rest of the section, let $\tau \in (0, \infty)$ be the regularization strength.

Lemma 5.33 (Lemma 5.10). *Let $\Sigma_0 \in \mathbb{S}_{++}(n)$, and let $W \in \mathbb{M}(n,m)$. Then,*

$$\|W\| \leqslant \sqrt{2\left(L_\tau^1(W) + \|\Sigma_0^{1/2}\|_{\mathrm{F}}^2\right) - n\tau} \leqslant \sqrt{2\left(L_\tau^1(W) + \|\Sigma_0^{1/2}\|_{\mathrm{F}}^2\right)}.$$

Consequently, if for $k \in \mathbb{N}_0$, $L_\tau^1(W(k)) \leqslant L_\tau^1(W(0)) = L_\tau^N(\overrightarrow{W}(0))$, then

$$\|W(k)\| \leqslant \sqrt{2\left(L_\tau^N(\overrightarrow{W}(0)) + \|\Sigma_0^{1/2}\|_{\mathrm{F}}^2\right)}.$$

The next lemma is taken as-is from [NRT21], and allow to bound the individual weight matrices (spectral) norm as a function of the end-to-end matrix norm, in the case the parameter $\overrightarrow{W}$ is δ-balanced.

Lemma 5.34 ([NRT21, Proposition 3.2]). *Let $\overrightarrow{W} = (W_1, \dots, W_N) \in \Theta$ be with balancedness constant $\delta \in [0,\infty)$ and let $W = W_N \cdots W_1$. Then,*

$$\forall j \in [N], \quad \|W_j\|^2 \leqslant \|W\|^{2/N} + (N+1)^2\delta. \tag{5.39}$$

The next proposition is key in obtaining the Theorems 5.21 and 5.22. This is an adaptation of [NRT21, Proposition 3.4]. Once proven, the other results from [NRT21] will follow almost immediately. .

Proposition 5.35 (Strong descent for gradient descent). *Let $\overrightarrow{W} : \mathbb{N}_0 \to \Theta$ be a gradient descent path for the loss* (5.10c)*, and let (B_δ, K_δ, M) be defined by* (5.24)*. Assume that $\overrightarrow{W}(0)$ has balancedness constant $\alpha\delta$ for $\delta \in (0,\infty)$ and $\alpha \in [0,1)$. Assume the stepsize η satisfies*

$$\eta \leqslant \min\left\{\frac{2(1-\sigma)}{B_\delta}, \frac{\sigma(1-\alpha)\delta}{2L_\tau^N(\overrightarrow{W}(0))}\right\}, \tag{5.40}$$

for some $\sigma \in (0,1)$. Then, the gradient iterates defined by (5.5) *satisfy, for all $k \in \mathbb{N}_0$,*

(1) *$\overrightarrow{W}(k)$ has balancedness constant δ,*
(2) *$L_\tau^N(\overrightarrow{W}(k)) \leqslant L_\tau^N(\overrightarrow{W}(0))$,*
(3) *$\forall j \in [N], \quad \|W_j(k)\|^2 \leqslant K_\delta = M^{\frac{N}{2}} + (N+1)^2\delta$,*
(4) *$L_\tau^N(\overrightarrow{W}(k)) - L_\tau^N(\overrightarrow{W}(k+1)) \geqslant \sigma\eta\left\|\nabla L_\tau^N(\overrightarrow{W}(k))\right\|_{\mathrm{F}}^2$.*

Remark 5.36. *a*) The first condition on η mostly ensures that the norm of the weights during training remain bounded (statement (3)) and that the strong-descent condition holds (statement (4)) (B_δ will be shown to be a lower bound on the Hessian's norm of the loss), whereas the second condition ensures that the δ-balancedness property is preserved during training (statement (1)), as will be shown in the proof.

b) Taking $\sigma = \dfrac{4L_\tau^N(\overrightarrow{W}(0))}{4L_\tau^N(\overrightarrow{W}(0)) + (1-\alpha)\delta B_\delta}$ reduces condition (5.40) to condition (5.23).

Proof of Proposition 5.35. We follow the proof of [NRT21, Proposition 3.4]. Essentially, only the bound B_δ of the Hessian for L_τ^N is different in our case, and statement (4) will require most of the effort. The proof of the statement (1) is identical to the but will be nevertheless reproduced for the reader's convenience.

The statements (1) to (3) are proved by induction on $k \in \mathbb{N}_0$. We will show that statement (4) holds for k if statement (3) holds for k.

For $k = 0$, statement (1) holds by assumption on $\overrightarrow{W}(0)$, and statement (2) holds trivially. Further, Statement (3) holds by Lemmas 5.34 and 5.33.

Assume now that statements (1) to (3) hold for some $k \in \mathbb{N}_0$. We will show the different statements hold for $k+1$ in the order (2), (1), and finally (3).

Let us start from the gradient descent of the loss with respect to each layer

$$\begin{aligned} W_j(k+1) &= W_j(k) - \eta \nabla_{W_j} L_\tau^N(W_1(k), \cdots W_n(k)) \\ &= W_j(k) - \eta W_{N:j+1}^\top(k) \nabla_W L_\tau^1(W(k)) W_{j-1:1}^\top(k), \quad 1 \leqslant j \leqslant N, \end{aligned} \tag{5.41}$$

with the boundary conditions $W_{1:0}(k) = \mathrm{I}_{d_0}$ and $W_{N+1:N}(k) = \mathrm{I}_{d_N}$ for all $k \in \mathbb{N}_0$.

With the notations $\overrightarrow{W} = (W_1, W_2, \cdots, W_N)$ and

$$\nabla L_\tau^N(\overrightarrow{W}) = \begin{pmatrix} \nabla_{W_1} L_\tau^N(\overrightarrow{W}) \\ \vdots \\ \nabla_{W_N} L_\tau^N(\overrightarrow{W}) \end{pmatrix},$$

we consider to write the Taylor expansion in the form

$$\begin{aligned} L_\tau^N(\overrightarrow{W}(k+1)) = L_\tau^N(\overrightarrow{W}(k)) &+ \Big\langle \nabla L_\tau^N(\overrightarrow{W}(k)), \overrightarrow{W}(k+1) - \overrightarrow{W}(k) \Big\rangle \\ &+ \frac{1}{2} \Big\langle (\overrightarrow{W}(k+1) - \overrightarrow{W}(k))^\top \nabla^2 L_\tau^N(\overrightarrow{A_\xi}(k)), \overrightarrow{W}(k+1) - \overrightarrow{W}(k) \Big\rangle, \end{aligned} \tag{5.42}$$

with

$$\overrightarrow{A_\xi}(k) = \overrightarrow{W}(k) + \xi(\overrightarrow{W}(k+1) - \overrightarrow{W}(k)), \quad \text{for some } \xi \in [0,1].$$

Under the gradient update, $\overrightarrow{W}(k+1) - \overrightarrow{W}(k) = -\eta \nabla L_\tau^N(\overrightarrow{W})$. Therefore,

$$\begin{aligned} L_\tau^N(\overrightarrow{W}(k+1)) &= L_\tau^N(\overrightarrow{W}(k)) - \eta \|\nabla L_\tau^N(\overrightarrow{W}(k))\|_{\mathrm{F}}^2 \\ &\quad + \frac{\eta^2}{2} \langle \nabla L_\tau^N(\overrightarrow{W}(k)), \nabla^2 L_\tau^N(\overrightarrow{A_\xi}(k)) \nabla L_\tau^N(\overrightarrow{W}(k)) \rangle \\ \implies L_\tau^N(\overrightarrow{W}(k)) - L_\tau^N(\overrightarrow{W}(k+1)) &\geqslant \Big(1 - \frac{\eta}{2} \|\nabla^2 L_\tau^N(\overrightarrow{A_\xi})\|_{\mathrm{F}\to\mathrm{F}}\Big) \eta \|\nabla L_\tau^N(\overrightarrow{W}(k))\|_{\mathrm{F}}^2 \end{aligned} \tag{5.43}$$

By induction hypothesis of statement (2) at step k and Lemma 5.10,

$$\|W(k)\| \leqslant \|W(k)\|_{\mathrm{F}} \leqslant \sqrt{2\Big(L_\tau^1(W(k)) + \|\Sigma_0^{1/2}\|_{\mathrm{F}}^2\Big) - n\tau} \leqslant \sqrt{2\Big(L_\tau^1(W(k)(0)) + \|\Sigma_0^{1/2}\|_{\mathrm{F}}^2\Big)} \overset{(5.24c)}{=} M.$$

From Lemma 5.34, under the δ-balancedness induction hypothesis at step k,

$$\forall j \in [N], \quad \|W(k)_j\|^2 \leqslant \|W(k)\|^{2/N} + (N+1)\delta^2 \leqslant M^{2/N} + (N+1)\delta^2 = K_\delta \tag{5.44}$$

Moreover, for all $j \in [N]$, since $\xi \in [0,1]$,

$$\begin{aligned}
\|A_{\xi,j}(k)\| &= \|W_j(k) + \xi(W_j(k+1) - W_j(k))\| \\
&\leqslant \|W_j(k)\| + \|W_j(k+1) - W_j(k)\| \\
&= \|W_j(k)\| + \eta\|W_{N:j+1}^\top(k)\nabla L_\tau^1(W(k))W_{j-1:1}^\top(k)\| \\
&\leqslant K_\delta^{1/2} + \eta K_\delta^{\frac{N-1}{2}}\|\nabla L_\tau^1(W(k))\|.
\end{aligned}$$

From (5.12), $\|\nabla L_\tau^1(W(k))\| = 2\sqrt{L_\tau^N(\overrightarrow{W}(k))} \leqslant 2\sqrt{L_\tau^N(\overrightarrow{W}(0))}$ by induction hypothesis (2). Hence,

$$\|A_{\xi,i}(k)\| \leqslant K_\delta^{1/2} + 2\eta K_\delta^{\frac{N-1}{2}}\sqrt{L_\tau^N(\overrightarrow{W}(0))}.$$

We will make use of the following lemma in order to bound $\|A_{\xi,i}(k)\|$ further.

Lemma 5.37 *For $\eta < \dfrac{2}{B_\delta}$, one has*

$$\|A_{\xi,i}(k)\| \leqslant \left(1 + \frac{1}{2N}\right) K_\delta^{1/2} := \mathrm{A}_\delta \tag{5.45}$$

Proof. From the definition of (B_δ, K_δ, M) (5.24), $B_\delta \geqslant 4e\sqrt{n}NK_\delta^{N-1} \geqslant 8NK_\delta^{N-1}$ since $n \geqslant 1$. Then, if $\eta < \dfrac{2}{B_\delta}$, one has

$$\begin{aligned}
2\eta K_\delta^{\frac{N-1}{2}}\sqrt{L_\tau^N(\overrightarrow{W}(0))} \leqslant \frac{4K_\delta^{\frac{N-1}{2}}\sqrt{L_\tau^N(\overrightarrow{W}(0))}}{B_\delta} &\leqslant \frac{K_\delta^{\frac{N-1}{2}}\sqrt{\frac{M^2}{2} - \|\Sigma_0^{1/2}\|_{\mathrm{F}}^2}}{2NK_\delta^{N-1}} \\
&\leqslant \frac{K_\delta^{\frac{N-1}{2}}M}{2NK_\delta^{N-1}} \\
&= \frac{K_\delta^{\frac{N-1}{2}}\Big[K_\delta - (N+1)^2\delta\Big]^{N/2}}{2NK_\delta^{N-1}} \\
&\leqslant \frac{K_\delta^{\frac{N-1}{2}}K_\delta^{\frac{N}{2}}}{2NK_\delta^{N-1}} = \frac{1}{2N}K_\delta^{\frac{1}{2}},
\end{aligned}$$

so that

$$\|A_{\xi,i}(k)\| \leqslant K_\delta^{\frac{1}{2}} + 2\eta K_\delta^{\frac{N-1}{2}} \sqrt{L_\tau^N(\overrightarrow{W}(0))} \leqslant K_\delta^{\frac{1}{2}} + \frac{1}{2N} K_\delta^{\frac{1}{2}} = \left(1 + \frac{1}{2N}\right) K_\delta^{\frac{1}{2}}.$$

□

We next show that $\left\|\nabla^2 L_\tau^N\right\|_{\mathrm{F}\to\mathrm{F}} \leqslant B_\delta$.

The second-order differential of the loss L_τ^N can be upper-bounded using the chain-rule Lemma 4.30. With $(\overrightarrow{W}, \overrightarrow{X}) \in \Theta^2$, $W = \mu(\overrightarrow{W})$, and $\Sigma = \pi \circ \mu(\overrightarrow{W})$, one has

$$\begin{aligned}
\mathrm{d}\mu(\overrightarrow{W})[\overrightarrow{X}] &= \sum_{i\in[N]} W_N \cdots X_i \cdots W_1 =: W_{\overrightarrow{X}}, \\
\mathrm{d}^2\mu(\overrightarrow{W})[\overrightarrow{X}] &= \sum_{i\in[N]} \sum_{j\in[N]\setminus\{i\}} W_N \cdots X_j \cdots X_i \cdots W_1 =: W_{\overrightarrow{X},\overrightarrow{X}}, \\
\mathrm{d}\pi(W)[X] &= WX^\top + XW^\top, \\
\mathrm{d}^2\pi(W)[X] &= 2XX^\top, \\
\mathrm{d}(\pi\circ\mu)(\overrightarrow{W})[\overrightarrow{X}] &= \mathrm{d}\pi(W)[\mathrm{d}\mu(\overrightarrow{W})[\overrightarrow{X}]] \\
&= W_{\overrightarrow{X}} W^\top + W W_{\overrightarrow{X}}^\top, \\
\mathrm{d}^2(\pi\circ\mu)(\overrightarrow{W})[\overrightarrow{X}] &= \mathrm{d}^2\pi(W)[\mathrm{d}\mu(\overrightarrow{W})[\overrightarrow{X}]] + \mathrm{d}\pi(W)[\mathrm{d}^2\mu(\overrightarrow{W})[\overrightarrow{X}]] \\
&= 2W_{\overrightarrow{X}} W_{\overrightarrow{X}}^\top + W W_{\overrightarrow{X},\overrightarrow{X}}^\top + W_{\overrightarrow{X},\overrightarrow{X}} W^\top, \\
\mathrm{d}^2(L_\tau\circ\pi\circ\mu)(\overrightarrow{W})[\overrightarrow{X}] &= \mathrm{d}^2 L_\tau(\Sigma)[\mathrm{d}(\pi\circ\mu)(\overrightarrow{W})[\overrightarrow{X}]] + \mathrm{d}L_\tau(\Sigma)[\mathrm{d}^2(\pi\circ\mu)(\overrightarrow{W})[\overrightarrow{X}]], \\
&= \mathrm{d}^2 L_\tau(\Sigma)[W_{\overrightarrow{X}} W^\top + W W_{\overrightarrow{X}}^\top] + 2\,\mathrm{d}L_\tau(\Sigma)[W_{\overrightarrow{X}} W_{\overrightarrow{X}}^\top] \\
&\quad + \mathrm{d}L_\tau(\Sigma)[W_{\overrightarrow{X},\overrightarrow{X}} W^\top + W W_{\overrightarrow{X},\overrightarrow{X}}^\top]
\end{aligned}$$

Given any $(\overrightarrow{W}, \overrightarrow{X}) \in \Theta^2$, we denote $W_{\overrightarrow{X}} = \sum_{i\in[N]} W_N \cdots X_i \cdots W_1$ and $W_{\overrightarrow{X},\overrightarrow{X}} = \sum_{i\in[N]} \sum_{j\in[N]\setminus\{i\}} W_N \cdots X_j \cdots X_i \cdots W_1$.

Then, with the operator $\mathcal{G}_\tau$ defined by (4.32) and its bound (5.27b)

$$\begin{aligned}
\mathrm{d}^2 L_\tau^N(\overrightarrow{W})[\overrightarrow{X},\overrightarrow{X}] &= \mathrm{d}^2 L_\tau(\Sigma)[W_{\overrightarrow{X}} W^\top + W W_{\overrightarrow{X}}^\top] + 2\,\mathrm{d}L_\tau(\Sigma)[W_{\overrightarrow{X}} W_{\overrightarrow{X}}^\top] \\
&\quad + \mathrm{d}L_\tau(\Sigma)[W_{\overrightarrow{X},\overrightarrow{X}} W^\top + W W_{\overrightarrow{X},\overrightarrow{X}}^\top] \\
&= \left\langle W_{\overrightarrow{X}} W^\top + W W_{\overrightarrow{X}}^\top, \mathcal{G}_\tau(W_{\overrightarrow{X}} W^\top + W W_{\overrightarrow{X}^\top}) \right\rangle \\
&\quad + 2\left\langle \nabla L_\tau(\Sigma), W_{\overrightarrow{X}} W_{\overrightarrow{X}}^\top \right\rangle + \left\langle \nabla L_\tau(\Sigma), W_{\overrightarrow{X},\overrightarrow{X}} W^\top + W W_{\overrightarrow{X},\overrightarrow{X}}^\top \right\rangle \\
\implies \left|\mathrm{d}^2 L_\tau^N(\overrightarrow{W})[\overrightarrow{X}]\right| &\leqslant \frac{\lambda_{\max}^{1/2}(\Sigma_0^{1/2}\Sigma_\tau\Sigma_0^{1/2})}{2} \|\Sigma_\tau^{-1}(W_{\overrightarrow{X}} W^\top + W W_{\overrightarrow{X}}^\top)\|_{\mathrm{F}}^2 \\
&\quad + \|\nabla L_\tau(\Sigma)\|_{\mathrm{F}} \left(2\|W_{\overrightarrow{X}} W_{\overrightarrow{X}}^T\|_{\mathrm{F}} + \|W_{\overrightarrow{X},\overrightarrow{X}} W^\top + W W_{\overrightarrow{X},\overrightarrow{X}}^\top\|_{\mathrm{F}}\right)
\end{aligned}$$

We use the triangular inequality, the fact that, for $A \in \mathbb{M}(n,m)$ and $B \in \mathbb{M}(m,k)$, one has $\|AB\|_{\mathrm{F}} \leqslant \|A\|\|B\|_{\mathrm{F}}$, and the sub-multiplicativity of $\|\cdot\|$ in order to write that

$$\|A_{\overrightarrow{X}}\|_{\mathrm{F}} = \left\| \sum_{i\in[N]} A_N \cdots X_i \cdots A_1 \right\|_{\mathrm{F}} \tag{5.46}$$

$$\leqslant \sum_{i\in[N]} \left(\prod_{j\in[N]\setminus\{i\}} \|A_j\| \right) \|X_i\|_{\mathrm{F}}, \tag{5.47}$$

$$\leqslant \mathrm{A}_\delta^{N-1} \sum_{i\in[N]} \|X_i\|_{\mathrm{F}} \tag{5.48}$$

$$\implies \|A_{\overrightarrow{X}}\|_{\mathrm{F}}^2 \leqslant \mathrm{A}_\delta^{2(N-1)} \|\overrightarrow{X}\|_{\mathrm{F}}^2 \leqslant N\mathrm{A}_\delta^{2(N-1)} \|\overrightarrow{X}\|_{\mathrm{F}}^2, \tag{5.49}$$

and

$$\begin{aligned}
\|A_{\overrightarrow{X},\overrightarrow{X}}\|_{\mathrm{F}} &= \left\| \sum_{i\in[N]} \sum_{j\in[N]\setminus\{i\}} A_N \cdots X_\jmath \cdots X_\imath \cdots A_1 \right\|_{\mathrm{F}}, \quad (\imath,\jmath) = (\min(i,j), \max(i,j)) \\
&\leqslant \sum_{i\in[N]} \sum_{j\in[N]\setminus\{i\}} \|A_N \cdots X_\jmath \cdots X_\imath \cdots A_1\|_{\mathrm{F}} \\
&\leqslant \sum_{i\in[N]} \sum_{j\in[N]\setminus\{i\}} \|X_i\|_{\mathrm{F}} \|X_j\|_{\mathrm{F}} \left(\prod_{k\in[N]\setminus\{i,j\}} \|A_k\| \right) \\
&\leqslant \mathrm{A}_\delta^{N-2} \left(\sum_{i\in[N]} \|X_i\|_{\mathrm{F}} \right)^2 \\
&\leqslant \mathrm{A}_\delta^{N-2} \|\overrightarrow{X}\|_{\mathrm{F}}^2 \leqslant N\mathrm{A}_\delta^{N-2} \|\overrightarrow{X}\|_{\mathrm{F}}^2,
\end{aligned}$$

were we have used the condition on the stepsize η (5.40) together with Lemma 5.37.

Moreover, $\lambda_{\max}\left(\Sigma_0^{1/2}\left[A_\xi A_\xi^\top + \tau \mathrm{I}_n\right]\Sigma_0^{1/2}\right) \leqslant \|\Sigma_0\|\|A_\xi A_\xi^\top + \tau \mathrm{I}_n\| \leqslant \|\Sigma_0\|\left(\|A_\xi\|^2 + \tau\right)$

and $\|A_\xi\| \leqslant \mathrm{A}_\delta^N$. Hence,

$$\begin{aligned}
&\left|\mathrm{d}^2 L_\tau^N(\overrightarrow{A_\xi}(k))[\overrightarrow{X}(k)]\right| \\
&\leqslant \lambda_{\max}\left(\Sigma_0^{1/2}\left[A_\xi A_\xi^\top + \tau \mathrm{I}_n\right]\Sigma_0^{1/2}\right)\left\|\left(A_\xi A_\xi^\top + \tau \mathrm{I}_n\right)^{-1}\right\|^2 \|A_\xi\|^2 \left\|A_{\overrightarrow{X}}\right\|_{\mathrm{F}}^2 \\
&\quad + 2\left\|\nabla L_\tau(A_\xi A_\xi^\top)\right\|_{\mathrm{F}}\left(\left\|A_{\overrightarrow{X}}\right\|_{\mathrm{F}}^2 + \|A_\xi\|\left\|A_{\overrightarrow{X},\overrightarrow{X}}\right\|_{\mathrm{F}}\right) \\
&\leqslant \frac{\|\Sigma_0\|\left(\|A_\xi\|^2+\tau\right)\|A_\xi\|^2}{\lambda_{\min}^2(A_\xi A_\xi^\top + \tau \mathrm{I}_n)} N \mathrm{A}_\delta^{2(N-1)}\left\|\overrightarrow{X}\right\|_{\mathrm{F}}^2 \\
&\quad + 4N\mathrm{A}_\delta^{2(N-1)}\left\|\nabla L_\tau(A_\xi A_\xi^\top)\right\|_{\mathrm{F}}\left\|\overrightarrow{X}\right\|_{\mathrm{F}}^2
\end{aligned} \tag{5.50}$$

From the definition of A_δ (5.45) and the fact that $\left(1+\frac{1}{2N}\right)^{2(N-1)} \leqslant \left(1+\frac{1}{2N}\right)^{2N} \leqslant e$, we have

$$\mathrm{A}_\delta^{2(N-1)} = \left(1+\frac{1}{2N}\right)^{2(N-1)} K_\delta^{N-1} \leqslant e K_\delta^{N-1}, \tag{5.51a}$$

$$\text{and} \qquad \mathrm{A}_\delta^{2N} \leqslant e K_\delta^N. \tag{5.51b}$$

We now bound $\|\nabla L_\tau(A_\xi A_\xi^\top)\|_{\mathrm{F}}$. From Corollary 4.37, we know that $\nabla L_\tau(A_\xi A_\xi^\top) = \mathrm{I}_n - \Sigma_0^{1/2}\left[\Sigma_0^{1/2}\left(A_\xi A_\xi^\top + \tau \mathrm{I}_n\right)\Sigma_0^{1/2}\right]^{-1/2}\Sigma_0^{1/2}$. Then,

$$\begin{aligned}
\|\nabla L_\tau(A_\xi A_\xi^\top)\|_{\mathrm{F}} &\leqslant \sqrt{n}\|\nabla L_\tau(A_\xi A_\xi^\top)\| \\
&= \sqrt{n}\left\|\mathrm{I}_n - \Sigma_0^{1/2}\left[\Sigma_0^{1/2}\left(A_\xi A_\xi^\top + \tau \mathrm{I}_n\right)\Sigma_0^{1/2}\right]^{-1/2}\Sigma_0^{1/2}\right\| \\
&\leqslant \sqrt{n}\left[1 + \lambda_{\max}\left(\Sigma_0^{1/2}\left[\Sigma_0^{1/2}\left(A_\xi A_\xi^\top + \tau \mathrm{I}_n\right)\Sigma_0^{1/2}\right]^{-1/2}\Sigma_0^{1/2}\right)\right]
\end{aligned}$$

Further,

$$\begin{aligned}
&\lambda_{\max}\left(\Sigma_0^{1/2}\left[\Sigma_0^{1/2}\left(A_\xi A_\xi^\top+\tau \mathrm{I}_n\right)\Sigma_0^{1/2}\right]^{-1/2}\Sigma_0^{1/2}\right)\\
&\quad\leqslant \lambda_{\max}\left(\Sigma_0\right)\lambda_{\max}\left(\left[\Sigma_0^{1/2}\left(A_\xi A_\xi^\top+\tau \mathrm{I}_n\right)\Sigma_0^{1/2}\right]^{-1/2}\right)\\
&\quad= \lambda_{\max}(\Sigma_0)\lambda_{\min}^{-1/2}\left(\Sigma_0^{1/2}\left(A_\xi A_\xi^\top+\tau \mathrm{I}_n\right)\Sigma_0^{1/2}\right)\\
&\quad\leqslant \lambda_{\max}(\Sigma_0)\lambda_{\min}^{-1}\left(\Sigma_0^{1/2}\right)\lambda_{\min}^{-1/2}\left(A_\xi A_\xi^\top+\tau \mathrm{I}_n\right)\\
&\quad= \frac{\lambda_{\max}(\Sigma_0)}{\lambda_{\min}(\Sigma_0^{1/2})\lambda_{\min}^{1/2}(A_\xi A_\xi^\top+\tau \mathrm{I}_n)}\\
&\quad\leqslant \frac{\lambda_{\max}(\Sigma_0)}{\lambda_{\min}(\Sigma_0^{1/2})\sqrt{\tau}}\\
&\quad= \frac{\sqrt{\tau\lambda_{\min}(\Sigma_0)}\lambda_{\max}(\Sigma_0)}{\tau\lambda_{\min}(\Sigma_0)}
\end{aligned}$$

hence the bound

$$\|\nabla L_\tau(A_\xi A_\xi^\top)\|_{\mathrm{F}} \leqslant \sqrt{n}\left[1+\frac{\sqrt{\tau\lambda_{\min}(\Sigma_0)}\lambda_{\max}(\Sigma_0)}{\tau\lambda_{\min}(\Sigma_0)}\right] \tag{5.52}$$

With (5.51) and (5.52), the bound in (5.50) consequently becomes

$$\begin{aligned}
&\left|\mathrm{d}^2 L_\tau^N(\overrightarrow{A_\xi})[\overrightarrow{X}]\right|\\
&\leqslant eNK_\delta^{N-1}\frac{\|\Sigma_0\|\left(eK_\delta^N+\tau\right)eK_\delta^N}{\tau^2}\left\|\overrightarrow{X}\right\|_{\mathrm{F}}^2\\
&\qquad+4e\sqrt{n}NK_\delta^{N-1}\left[1+\frac{\sqrt{\tau\lambda_{\min}(\Sigma_0)}\lambda_{\max}(\Sigma_0)}{\tau\lambda_{\min}(\Sigma_0)}\right]\left\|\overrightarrow{X}\right\|_{\mathrm{F}}^2\\
&= eNK_\delta^{N-1}\left(\frac{\|\Sigma_0\|\left(eK_\delta^N+\tau\right)eK_\delta^N}{\tau^2}+4\sqrt{n}\left[1+\frac{\sqrt{\tau\lambda_{\min}(\Sigma_0)}\lambda_{\max}(\Sigma_0)}{\tau\lambda_{\min}(\Sigma_0)}\right]\right)\|\overrightarrow{X}\|_{\mathrm{F}}^2
\end{aligned}$$

so that

$$\begin{aligned}
&\left\|\nabla^2 L_\tau^N(\overrightarrow{A_\xi})\right\|_{\mathrm{F}\to\mathrm{F}}\\
&\leqslant eNK_\delta^{N-1}\left(\frac{\|\Sigma_0\|\left(eK_\delta^N+\tau\right)eK_\delta^N}{\tau^2}+4\sqrt{n}\left[1+\frac{\sqrt{\tau\lambda_{\min}(\Sigma_0)}\lambda_{\max}(\Sigma_0)}{\tau\lambda_{\min}(\Sigma_0)}\right]\right)\\
&\overset{(5.24a)}{=} B_\delta.
\end{aligned} \tag{5.53}$$

From (5.43), this leads to the inequality

$$L_\tau^N(\overrightarrow{W}(k)) - L_\tau^N(\overrightarrow{W}(k+1)) \geqslant \left(1 - \frac{1}{2}\eta B_\delta\right)\eta\left\|\nabla L_\tau^N(\overrightarrow{W}(k))\right\|_{\mathrm{F}}^2$$
$$\geqslant \sigma\eta\left\|\nabla L_\tau^N(\overrightarrow{W}(k))\right\|_{\mathrm{F}}^2 \geqslant 0, \tag{5.54}$$

where we use the condition on η (5.40). This shows statement (2) for $k+1$. This also shows that, if statements (2) and (3) hold for k, then statement (4) holds for k as well.

We now turn to statement (1). Therefore, we consider the δ-balancedness condition at the step size $k+1$: for all $j \in [N-1]$,

$$\begin{aligned}
&\left\|W_{j+1}^\top(k+1)W_{j+1}(k+1) - W_j(k+1)W_j^\top(k+1)\right\| \\
&= \Bigg\| \left[W_{j+1}(k) - \eta\nabla_{W_{j+1}(k)}L_\tau^N(\overrightarrow{W}(k))\right]^\top \left[W_{j+1}(k) - \eta\nabla_{W_{j+1}(k)}L_\tau^N(\overrightarrow{W}(k))\right] \\
&\quad - \left[W_j(k) - \eta\nabla_{W_j(k)}L_\tau^N(\overrightarrow{W}(k))\right]\left[W_j(k) - \eta\nabla_{W_j(k)}L_\tau^N(\overrightarrow{W}(k))\right]^\top \Bigg\| \\
&= \Bigg\| W_{j+1}^\top(k)W_{j+1}(k) - W_j(k)W_j^\top(k) \\
&\quad + \eta\Big[-\nabla_{W_{j+1}}^\top L_\tau^N(\overrightarrow{W}(k))W_{j+1}(k) + \nabla_{W_j}L_\tau^N(\overrightarrow{W}(k))W_j^\top(k) \\
&\qquad - W_{j+1}^\top(k)\nabla_{W_{j+1}}L_\tau^N(\overrightarrow{W}(k)) + W_j(k)\nabla_{W_j}^\top L_\tau^N(\overrightarrow{W}(k))\Big] \\
&\quad + \eta^2\left[\nabla_{W_{j+1}}^\top L_\tau^N(\overrightarrow{W}(k))\nabla_{W_{j+1}}L_\tau^N(\overrightarrow{W}(k)) - \nabla_{W_j}L_\tau^N(\overrightarrow{W}(k))\nabla_{W_j}^\top L_\tau^N(\overrightarrow{W}(k))\right] \Bigg\| \\
&\leqslant \left\|W_{j+1}^\top(k)W_{j+1}(k) - W_j(k)W_j^\top(k)\right\| \\
&\quad + \eta^2\left(\left\|\nabla_{W_{j+1}}L_\tau^N(\overrightarrow{W}(k))\right\|^2 + \left\|\nabla_{W_j}L_\tau^N(\overrightarrow{W}(k))\right\|^2\right)
\end{aligned}$$

which by recursion on the time-step k, implies that, for all layer $j \in [N-1]$,

$$\begin{aligned}
&\left\|W_{j+1}^\top(k+1)W_{j+1}(k+1) - W_j(k+1)W_j^\top(k+1)\right\| \\
&\quad \leqslant \left\|W_{j+1}^\top(0)W_{j+1}(0) - W_j(0)W_j^\top(0)\right\| \\
&\qquad + \eta^2\left(\sum_{\ell=0}^k\left\|\nabla_{W_{j+1}}L_\tau^N(\overrightarrow{W}(\ell))\right\|^2 + \left\|\nabla_{W_j}L_\tau^N(\overrightarrow{W}(\ell))\right\|^2\right) \\
&\quad \leqslant \alpha\delta + 2\eta^2\sum_{\ell=0}^k\left\|\nabla L_\tau^N(\overrightarrow{W}(\ell))\right\|^2
\end{aligned} \tag{5.55}$$

since $\overrightarrow{W}(0)$ has $\alpha\delta$ balancedness by assumption and $\left\|\nabla L_\tau^N(\overrightarrow{W}(t))\right\|^2 \geqslant \max_{j\in[N]}\left\|\nabla_{W_j} L_\tau^N(\overrightarrow{W}(t))\right\|^2 \geqslant \frac{1}{2}\left(\left\|\nabla_{W_j} L_\tau^N(\overrightarrow{W}(t))\right\|^2 + \left\|\nabla_{W_{j+1}} L_\tau^N \overrightarrow{W}(t)\right\|^2\right)$ for all $j \in [N-1]$ and all $t \in [k]$.

By (5.54), one further has

$$L_\tau^N(\overrightarrow{W}(0)) - L_\tau^N(\overrightarrow{W}(k+1)) = \sum_{t=0}^{k} L_\tau^N(\overrightarrow{W}(t)) - L_\tau^N(\overrightarrow{W}(t+1)) \geqslant \sum_{t=0}^{k} \sigma\eta\left\|\nabla L_\tau^N(\overrightarrow{W}(t))\right\|_{\mathrm{F}}^2,$$

so that (5.55) becomes

$$\begin{aligned}\left\|W_{j+1}^\top(k+1)W_{j+1}(k+1) - W_j(k+1)W_j^\top(k+1)\right\| &\leqslant \alpha\delta + \frac{2}{\sigma}\eta\left(L_\tau^N(\overrightarrow{W}(0)) - L_\tau^N(\overrightarrow{W}(k+1))\right)\\ &\leqslant \alpha\delta + \frac{2\eta L_\tau^N(\overrightarrow{W}(0))}{\sigma} \leqslant \delta.\end{aligned}$$

where we have used the condition $\eta \leqslant \dfrac{\sigma(1-\alpha)\delta}{2L_\tau^N(\overrightarrow{W}(0))}$ (5.40). This shows the induction for statement (1).

It remains the show statement (3) for $k+1$. As we have shown statements (1) and (2) for $k+1$ already (i.e. $\overrightarrow{W}(k+1)$ is δ-balanced and $L^N(\overrightarrow{W}(k+1)) \leqslant L^N(\overrightarrow{W}(0)$, Lemmas 5.33 and 5.34, allow to conclude:

$$\|W_j(k+1)\| \leqslant \|W(k+1)\|^{2/N} + (N+1)\delta^2 \leqslant \sqrt{2(L^N(\overrightarrow{W}(0)) + \|\Sigma_0^{1/2}\|_{\mathrm{F}}^2)} + (N+1)\delta^2 = K_\delta,$$

and statement (3) holds at step $k+1$.

As remarked previously, since for all $k \in \mathbb{N}_0$, statements (2) and (3) hold, statement (4) holds as well.

This concludes the proof. □

In order to show the convergence of the iterates to a critical point of the loss, we rely of the strong-descent conditions [AMA05], as does [NRT21].

Definition 5.38 (*Strong-descent conditions, [AMA05, Definition 3.1]*). Let $f\colon \mathbb{R}^q \to \mathbb{R}$ be a differentiable function. We say that a sequence $(x_k)_{k\in\mathbb{N}_0}$ of elements in $\mathbb{R}^q$ satisfies the *strong-descent conditions* if there exist $\sigma \in (0,\infty)$ and $k_0 \in \mathbb{N}_0$ such that, for all $k \geqslant k_0$,

$$f(x_k) - f(x_{k+1}) \geqslant \sigma\|f(x_k)\|\|x_{k+1} - x_k\| \tag{5.56a}$$

and

$$f(x_{k+1}) = f(x_k) \implies x_{k+1} = x_k. \tag{5.56b}$$

We also rely on the next theorem to conclude, which is an extension of Łojasiewicz' theorem to a discrete gradient descent scheme.

Theorem 5.39 ([AMA05, Theorem 3.2]). *Let $f\colon \mathbb{R}^q \to \mathbb{R}$ be an analytic function. Let the sequence $(x_k)_{k\in\mathbb{N}_0}$ satisfy the strong-descent condition (Definition 5.38). Then, either $\lim_{k\to\infty} \|x_k\| = \infty$, or there exists a single element $x^* \in \mathbb{R}^q$ such that*

$$\lim_{k\to\infty} x_k = x^*.$$

Proof of Theorem 5.21. The proof follows exactly the one of [NRT21, Theorem 2.4]. We provide here the main idea of the proof. Let $\overrightarrow{W}\colon \mathbb{N}_0 \to \Theta$ be a path in parameter space for the gradient descent iterations of the loss L^N_τ, initialized such that $\overrightarrow{W}(0)$ is $\alpha\delta$ balanced for $(\alpha,\delta) \in [0,\infty)\times(0,\infty)$, and let η be such that condition (5.23) is satisfied. As per Remark 5.36.b), the parameter σ in Proposition 5.35 can be chosen such that condition (5.40) reduces to condition (5.23).

We first show that the sequence $(\overrightarrow{W}(k))_{k\in\mathbb{N}_0}$ admits a limit, and then show that this limit is a critical point for the loss L^N_τ.

For any $k \in \mathbb{N}_0$, the gradient update can be written $\overrightarrow{W}(k+1) - \overrightarrow{W}(k) = \eta\nabla L^N_\tau(\overrightarrow{W}(k))$. Then, Proposition 5.35.(4) ensures that

$$L^N_\tau(\overrightarrow{W}(k)) - L^N_\tau(\overrightarrow{W}(k+1)) \geqslant \sigma\left\|\nabla L^N_\tau(\overrightarrow{W}(k))\right\|_{\mathrm{F}}\left\|\overrightarrow{W}(k+1) - \overrightarrow{W}(k)\right\|_{\mathrm{F}}, \tag{5.57}$$

which implies that the first strong-descent condition (5.56a) is satisfied. The second condition (5.56b) is also satisfied, as $L^N_\tau(\overrightarrow{W}(k+1)) = L^N_\tau(\overrightarrow{W}(k)) \implies \|\nabla L^N_\tau(\overrightarrow{W}(k))\|_{\mathrm{F}}\|\overrightarrow{W}(k+1) - \overrightarrow{W}(k)\|_{\mathrm{F}} = 0$, implying that either $\overrightarrow{W}(k) = \overrightarrow{W}(k+1)$ or $\nabla L^N_\tau(\overrightarrow{W}(k)) = 0$, in which case we also have $\overrightarrow{W}(k) = \overrightarrow{W}(k+1)$ by the gradient descent update. Therefore, the sequence $(\overrightarrow{W}(k))_{k\in\mathbb{N}_0}$ satisfies the strong-descent conditions (5.56).

Further, the sequence $(\overrightarrow{W}(k))_{k\in\mathbb{N}_0}$ is bounded (by Proposition 5.35.(2)) and L^N_τ is analytic, which means by Theorem 5.39 that the iterates converge to a unique element in Θ.

Let $\overrightarrow{W}^*$ be this limit. We now show that $\overrightarrow{W}^*$ is a critical point for L^N_τ, i.e., that $\nabla L^N_\tau(\overrightarrow{W}^*) = 0$. As ∇L^N_τ is continuous, we know that $\nabla L^N_\tau(\overrightarrow{W}^*) = \lim_{k\to\infty} \nabla L^N_\tau(\overrightarrow{W}(k))$.

By applying repeatedly Proposition 5.35.(4), one obtains that, for all $k \in \mathbb{N}_0$,

$$L^N_\tau(\overrightarrow{W}(0)) - L^N_\tau(\overrightarrow{W}(k+1)) \geqslant \sigma\eta\sum_{i=0}^{k}\|\nabla L^N_\tau(\overrightarrow{W}(i))\|^2_{\mathrm{F}} \tag{5.58}$$

The limit as $k \to \infty$ on the left-hand size of (5.58) exists since the sequence $(L^N_\tau(\overrightarrow{W}(k)))_{k\in\mathbb{N}_0}$ is decreasing and lower bounded by 0. Hence, the series

$\sum_{i=0}^{\infty} \|\nabla L_\tau^N(\overrightarrow{W}(i))\|_{\mathrm{F}}^2$ converges, implying that $\lim_{k\to\infty} \|\nabla L_\tau^N(\overrightarrow{W}(k))\|_{\mathrm{F}}^2 = \|\nabla L_\tau^N(\overrightarrow{W}^*)\|_{\mathrm{F}}^2 = 0$. Therefore, $\overrightarrow{W}^*$ is a critical point for L_τ^N. □

Since, for $\tau \in (0, \infty)$, the loss L_τ^N is analytic, and since Proposition 5.35 holds, the proof of Theorem 5.21 is exactly the same as the proof of [NRT21, Theorem 2.4]

Further, with a constant step size, [LPP+19, Corollary 2] can be used in order to conclude that we do not end up in a strict saddle point of the loss, if the initialization is generic.

Proposition 5.40 ([LPP+19, Corollary 2]). *Let $f \in \mathscr{C}^2(\mathbb{R}^q)$, and assume that there exists $m \in (0, \infty)$ such that $\nabla^2 f(x) < m$ for all $x \in \mathbb{R}^q$. Consider the iterations $x\colon \mathbb{N}_0 \to \mathbb{R}^q$ of the gradient descent algorithm (5.5). Then, for $\eta < \frac{1}{m}$, the set of initializations $x(0) \in \mathbb{R}^q$ such that $(x(k))_{k\in\mathbb{N}_0}$ converges to a strict saddle point of f has measure zero.*

We prove Theorem 5.22 by slightly adapting [NRT21, Theorem 2.6] to our setting.

Proof of Theorem 5.22. For $\tau \in (0, \infty)$, the loss L_τ^N is twice-continuously differentiable. Due to the definition of $(\delta_{\mathcal{B}}, L_{\mathcal{B}}, M_{\mathcal{B}})$ (5.25) together with condition on the stepsize (5.40), the conditions of Theorem 5.21 are satisfied for each initialization $\overrightarrow{W}(0) \in \mathcal{B}$. Hence, $(\overrightarrow{W}(k))_{k\in\mathbb{N}_0}$ converges to a critical point of L_τ^N for all $\overrightarrow{W}(0) \in \mathcal{B}$. By Proposition 5.40, the convergence of gradient descent with initial value in $\mathcal{B}$ and with stepsize η to a strict saddle point only occurs for subset of $\mathcal{B}$ that has measure zero.

For the rest of the proof, we have to show that a critical point of L_τ^N that is not a strict saddle point will correspond to a global minimizer of the loss L_τ^1 restricted to a given manifold $\mathcal{M}_k$. This corresponds essentially to the same line of argument of [BRT+21, Theorem 6.12], from which we repeat the main points transposed in our setting.

Let us denote by $\overrightarrow{W}^* = (W_1^*, \dots, W_N^*)$ the limit of $(\overrightarrow{W}(k))_{k\in\mathbb{N}_0}$, $W^* = W_{N:1}^*$, and $k = \operatorname{rank} W^*$. Recall that $\underline{d} = \min_{i\in[n]}\{d_i\}$. Then, $k \leqslant \underline{d}$ and W^* is a critical point of L_τ^1 restricted to the manifold $\mathcal{M}_k$ by Proposition 4.23. If W is *not* a global minimizer of L_τ^1 restricted to $\mathcal{M}_k$, then W is a strict saddle point of L_τ^N by Proposition 4.12, which has measure zero, which finishes the proof.

□

5.6 Experimental evaluation of the convergence rate

We conduct numerical experiments[†] to illustrate our theoretical result Theorem 5.18. With a learning rate chosen to approximate a gradient flow setting, we observe empirically

[†]The source code for the experiments can be found at https://github.com/brechetp/BW-linear-networks.

(Figure 5.1) the linear dependency of the asymptotic rate of convergence as a function of the depth of the network N and the minimum singular value square root of the target $\sigma_{\min}(\Sigma_0^{1/2})$.

5.6.1 Setup

The target covariance matrix is sampled as $\Sigma_0 := \Omega\Lambda\Omega^\top$, where $\Omega \in \mathbb{O}(n)$ is a random orthogonal matrix, and the eigenvalues in Λ follow a Zipf distribution, $\lambda_j \propto j^{-1}$ for $j \in [n]$. The input data dimension is set to be $n = 20$. We vary the minimum eigenvalue for the target $\lambda_{\min}$ and set $\lambda_j = n \cdot j^{-1} \cdot \lambda_{\min}$ for $j \in [n]$. We consider constant width networks with $d_i = n = m = 20$, for each $i \in [N]$.

To fulfill the modified deficiency margin assumption (Definition 5.14), we initialize the parameters close to the target Σ_0. If $\Sigma_0 = \Omega\Lambda\Omega^\top$, then the weights are initialized in a way such that the initial covariance matrix is $\Sigma(0) = \Sigma_0 + A$, with A being a random perturbation. More precisely, we choose $A = \Gamma D \Gamma^\top$, where Γ is a random orthogonal matrix, and D is a diagonal matrix with small eigenvalues — so that the overall distance between the initialization and the target is bounded by $\sigma_{\min} - c$, for some $c \in (0, \infty)$. With this initialization and the balancedness protocol explained by [ACG+19], the network satisfies both the balancedness and modified deficiency margin assumptions. In this case we expect Theorem 5.18 to hold for small step sizes. We estimate the asymptotic linear convergence rate numerically.

5.6.2 Result

We compute the rate of convergence as follows. First, the network (initialized as detailed above) is trained with a small enough learning rate $\eta \in (0, \infty)$. Then, we compute $\log\left(L(\Sigma(t)) - L(\Sigma^*)\right)$. Theorem 5.18 states that this should be a linear function of the time t. Therefore, a linear regression is performed, and the slope taken as the empirical rate of convergence. According to Theorem 5.18, this rate should be linear in the depth N and linear in the strong convexity parameter K, which suggests that it could be linear in $\sigma_{\min}(\Sigma_0^{1/2})$. Hence, we compute the empirical rate of convergence for varying depths and $\sigma_{\min}(\Sigma_0^{1/2})$, reported in Figure 5.1. In Figure 5.1a the linear dependence in the depth is clearly visible, and Figure 5.1b indicates a linear dependence in $\sigma_{\min}$ too. Our Theorem 5.18 only provides an upper bound on the convergence rate and hence we compare with the empirical rates. The results suggest that this is indeed the actual behavior in practice.

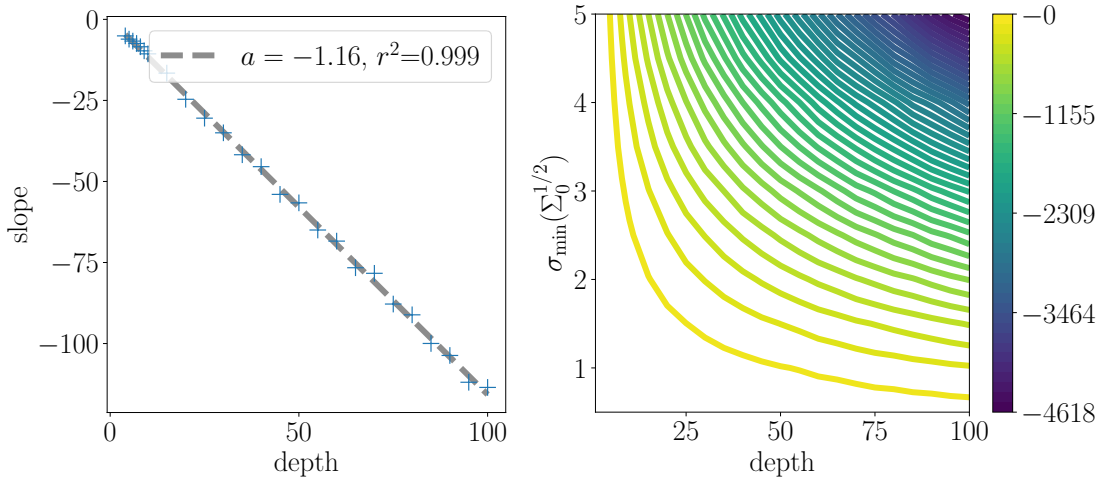


Figure 5.1: Logarithm of the linear convergence rate as a function of the depth N and the minimum singular value $\sigma_{\min}(\Sigma_0^{1/2})$. In (a), $\sigma_{\min}(\Sigma_0^{1/2}) \approx 0.7078$ is fixed; convergence rate and its linear regression as a function of the depth N. In (b), both the depth and $\sigma_{\min}$ are varying, and the rate is shown in a contour plot. The hyperbolas indicate a linear dependency on both the depth N of the network and the minimum singular value of the target square root $\sigma_{\min}(\Sigma_0^{1/2})$, which is coherent with the upper-bound given by Theorem 5.18.

5.7 Conclusion

In this chapter, we studied the training of generative linear neural networks using the Bures-Wasserstein distance. In both the continuous and discrete time settings, we derived convergence guarantees of the gradient algorithms towards minimizers of the (regularized) Bures-Wasserstein loss. Depending on the learning algorithm, a different set of initialization assumptions are made. For the gradient flow algorithm, exact balancedness of the weights and modified deficiency margin are assumed, which leads to an exponential upper-bound on the rate of convergence for the original loss. For the gradient descent algorithm, approximate balancedness and (simply) bounded initialization are assumed. These milder assumptions lead to showing convergence of the algorithm towards a minimizer for the loss, which is global if the loss on the function space is restricted to a bounded-rank set of matrices.

The results we obtained contribute to the ongoing efforts to better characterize the optimization problems that arise in learning with deep neural networks beyond the commonly considered discriminative settings with the square loss.

Many questions remain open. Notably, it would be interesting to relax the different as-

sumptions that were made to analyze the linear model. Arguably, the strongest of these assumptions is the modified deficiency margin condition that we invoked in order to establish our convergence results for the gradient flow, as this constrains the parametrization to be of full rank. Other assumptions include the balancedness of the weights, or the definitiveness of the target matrix. Moreover, deriving (an upper-bound of) the rate of convergence in the case of a gradient-descent update would also be an interesting line of study.

Although the linear-network framework is attractive since it is relatively easy to analyze, the analysis of neural networks with activation functions would be manifestly worth of study in the future. In addition, relaxing the optimal discriminator hypothesis would enable analyzing a more realistic GAN model, as a discriminator network would then be required to optimize as well.

This completes the analysis of a Bures-Wasserstein linear GAN.

6 Mode connectivity

This chapter is the work published as [NBM21]. *PB has performed all the experiments and redacted Sections 6.6, 6.B and 6.C.*

The question of how and why the phenomenon of mode connectivity occurs in training deep neural networks has gained remarkable attention in the research community. From a theoretical perspective, two possible explanations have been proposed: (i) the loss function has connected sublevel sets, and (ii) the solutions found by stochastic gradient descent are dropout stable. While these explanations provide insights into the phenomenon, their assumptions are *not* always satisfied in practice. In particular, the first approach requires the network to have one layer with order of N neurons (N being the number of training samples), while the second one requires the loss to be almost invariant after removing half of the neurons at each layer (up to some rescaling of the remaining ones). In this work, we improve both conditions by exploiting the quality of the features at every intermediate layer together with a *milder* over-parameterization requirement. More specifically, we show that: (i) under generic assumptions on the features of intermediate layers, it suffices that the last two hidden layers have order of $\sqrt{N}$ neurons, and (ii) if subsets of features at each layer are linearly separable, then almost no over-parameterization is needed to show the connectivity. Our experiments confirm that the proposed condition ensures the connectivity of solutions found by stochastic gradient descent, even in settings where the previous requirements do not hold.

6.1 Introduction

The aim of this work is to provide further theoretical insights into the mode connectivity phenomenon which has recently attracted some considerable attention from the research community (i.e. training a neural network via stochastic gradient descent (SGD) from different random initializations often leads to solutions that could be connected by a path of low loss) [AY20; DVS+18; FJ19; FB17; GIP+18]. Existing theoretical insights from optimization landscape analysis suggest that, for over-parameterized networks, the loss function has connected sublevel sets, and hence the solutions found by SGD are also connected [Ngu19; SS16; VBB19]. However, in order for such a result to hold, the network needs to have one wide layer with at least $N+1$ neurons (N being the number of training samples) [Ngu21]. This bound is also known to be tight for 2-layer networks. Here, we would like to understand the above phenomenon for deep and mildly over-parameterized networks (i.e. we aim to prove weaker guarantees under milder conditions). The idea

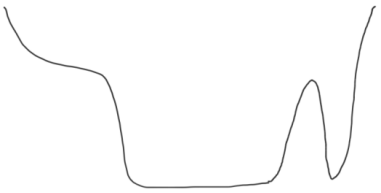


Figure 6.1: A function with disconnected sublevel sets and two connected sets of minima.

is that, even when the sublevel sets are disconnected, SGD may only concentrate in a (large) region with connected minima, see e.g. Figure 6.1. This leads to the question that whether one can characterize the properties of the solutions found by SGD which capture the mode connectivity phenomenon, for mildly over-parameterized networks. Along this line, a recent work [KWL+19] proposed the so-called "dropout stability" condition, that is, removing half the neurons from every hidden layer leads to almost no change in the value of the loss. As shown in the prior works (also in our experiments in Section 6.6), this condition holds for very wide networks or when dropout is used as a regularizer during training, but it does not hold under other practical settings.

The aim of this work is to provide a new characterization of SGD solutions which capture their connectivity property, and to improve the previous assumptions (width $\Omega(N)$, and dropout stability) for networks of practical sizes, learned without the support of dropout training. As a remark, although our work is motivated by the original phenomenon concerning mainly the solutions found by SGD, all the theoretical results are still valid for other pairs of solutions in parameter space.

Main contributions. Our main Theorem 6.1 provides an upper bound of the loss along a path connecting two given solutions in parameter space. This bound is characterized by the loss of L *optimized and sparse* sub-networks, L being the network depth. These L sub-networks connect subsets of at least half the neurons at each layer to the output (see Figure 6.2), and their weights can be optimized independently of the weights of the original network, which leads to a significant reduction of the loss. This is the key difference to the previously proposed dropout stability property [KWL+19]. Next, we specialize our theorem to different settings to demonstrate the improvement over prior works:

- Corollary 6.2 shows that our condition is provably milder than dropout stability [KWL+19]. In fact, dropout stability implies that the loss does not change much when we remove at least half of the neurons at each layer and rescale the remaining ones. Here, we allow to optimize again all the weights of the sub-networks connecting the subsets of features to the output.
- As the weights of the sub-networks can be optimized, we establish a formal connection between mode connectivity and memorization capacity of neural nets. In

particular, Corollary 6.3 proves that, if the last two hidden layers have $\Omega(\sqrt{N})$ neurons, then under some additional mild conditions on the features of intermediate layers, any two given solutions can be connected by a path of low loss. This improves the width condition of prior works [Ngu19; Ngu21; VBB19] from $\Omega(N)$ to $\Omega(\sqrt{N})$ at the cost of ensuring a weaker guarantee (i.e. the connectivity of sublevel sets proved in prior works implies the connectivity of any two pairs of solutions, but in this paper we only show the connectivity for certain pairs).

- Corollary 6.4 proves that, for a classification task, if subsets of features at each layer are linearly separable, then no over-parameterization is required to prove the connectivity.

The message of this paper is that, in order to ensure the connectivity, a smooth trade-off between feature quality and over-parameterization suffices. The two extremes of this trade-off are captured by Corollary 6.3 and 6.4 mentioned above. More generally, at each layer l, we pick a subset $\mathcal{I}_l$ containing at least half the neurons. If the features indexed by $\mathcal{I}_l$ are good enough, in the sense that there exists a small subnetwork ξ_l achieving small loss when trained on these features as inputs, then all layers $k > l$ in the original network do not need to be larger than twice the widths of the subnet ξ_l. Hence, the better the features at lower layers, the less over-parameterized the higher layers need to be.

Finally, our numerical results show that, if the learning task is sufficiently simple (e.g. MNIST) and the network is sufficiently over-parameterized, then dropout stability holds and the SGD solutions can be connected as in [KWL+19]. However, when the learning task is more difficult (e.g. CIFAR-10), or for networks of moderate sizes, dropout stability does not hold and the path of [KWL+19] exhibits high loss. In this challenging setting, we are still able to find low-loss paths connecting the SGD solutions.

Proof techniques. We remark that, while previous work [Ngu19; Ngu21; VBB19] focused on the connectivity of all the sublevel sets (i.e. a property of the global landscape), here we are only interested in the connectivity of a particular subset of solutions, especially those that are discovered by SGD. This leads to several differences in the proofs. First, previous analysis relies on the fact that the first hidden layer has at least $N + 1$ neurons [Ngu21], and thus the training inputs can essentially be mapped to a feature space where they become linearly independent. This property allows the network to realize arbitrary outputs at each layer, which leads to the connectivity of sublevel sets. This property is no longer true for the networks considered here, where the layer widths are allowed to be *sublinear* in N. Second, the analysis of sublevel sets often do not take into account the bias of SGD, whereas here we focus more on understanding the properties of these solutions which lead to their connectivity. More closely related is the work by [KWL+19], with which we provide a detailed discussion in Section 6.4.

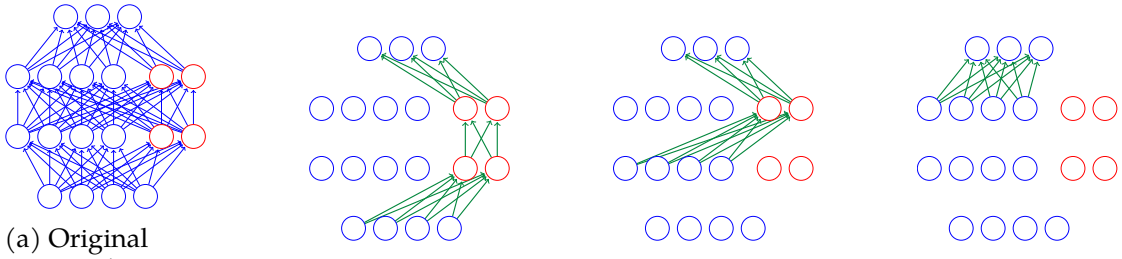
6.2 Related Work

Theoretical research on the loss surface of neural networks can be divided into two main categories: one that studies the global optimality of local minima, see e.g. [CHM+15; FA00; FYM+19; GGS+20; HV17; HM17; Kaw16; KB19; LB18; LDS18; LSL+18a; LSL+18b; LK17; NH17; SS18; SH18; WLL18; XLS17; YSJ18; ZSS19; ZL18], and the other that studies landscape connectivity. This work falls into the latter category. Specifically, the works of [FB17; SS16] suggest that the loss function of a two-layer network becomes more and more connected as the number of neurons increases. In [VBB19] and [Ngu19], it is proved that networks with one sufficiently wide layer have no spurious valleys and connected sublevel sets, respectively. All these works require networks with at least one layer of order N neurons, whereas Corollary 6.3 requires widths of order $\sqrt{N}$ neurons. In [BSI+19], the authors study the connectivity of *equivalent* minima (i.e. representing the same network function) which arise by permuting neurons from the same layer. A similar property is investigated in [HWT20] for all local minima lying in a cell or activation region (i.e. those sharing the same activation patterns across all the hidden neurons and training samples). In contrast, here we consider a more general setting: Theorem 6.1 provides upper bounds on the loss when connecting two *arbitrary* points in parameter space. In particular, these points need not be local minima, nor share the same activation patterns, or even represent the same function. This allows us to study the mode connectivity phenomenon of SGD solutions [KWL+19], which are typically not captured by the previous assumptions. Other works discuss lower bounds on over-parameterization to ensure the existence of descent paths in shallow nets [SSG19], the satisfiability of the dropout stability condition in the mean field regime [SM20].

Memorization capacity is an old and well-studied topic [Bau88; Cov65]. Networks with $\Omega(N)$ parameters are known to memorize N arbitrary samples, and this type of results has been studied for a variety of settings, see [BHL+19; BEL+20; GWZ19; MZ20; Ver20; YSJ19]. Nevertheless, how memorization capacity is related to the mode connectivity phenomenon remains unclear, especially in the absence of a layer of width $\Omega(N)$. Our work draws a formal connection between the two, which opens up new directions for future works.

6.3 Preliminaries

Given two nonnegative integers i, j, let $[i, j] = \{i, \dots, j\}$ and $[i] = \{1, 2, \dots, i\}$. Given a set $\mathcal{I}$, we denote by $|\mathcal{I}|$ its cardinality. We consider an L-layer neural network with weight matrices $W_l \in \mathbb{R}^{n_{l-1} \times n_l}$ for $l \in [L]$, and bias vectors $b_l \in \mathbb{R}^{n_l}$ for $l \in [L-1]$. Here, n_0 is the input dimension, n_L the output dimension, and $n_1, \dots, n_{L-1}$ the widths of all the hidden layers. Let θ be the set of all the network parameters. Given an input $x \in \mathbb{R}^{n_0}$, the feature vector $f_l \in \mathbb{R}^{n_l}$ at layer l is defined as



(a) Original net θ' (b) Subnet ξ_0 (c) Subnet ξ_1 (d) Subnet ξ_2

Figure 6.2: An illustration of the sub-networks appearing in our main Theorem 6.1. The blue neurons correspond to the subset $\mathcal{I}_l$ in the theorem, and its cardinality is at least half the layer width. Each sub-network ξ_l receives as inputs the blue neurons from layer l, and pass it through the red ones at the higher layers, all the way to the output of the network. All the weights of the sub-networks (in green) can be re-optimized to minimize the upper bound in the theorem. This is shown via the infimum taken over ξ_l on the RHS of Equation (6.4).

$$f_l(\theta, x) = \begin{cases} x, & l = 0, \\ \sigma(W_l^T f_{l-1}(\theta, x) + b_l), & l \in [L-1], \\ W_L^T f_{L-1}(\theta, x), & l = L, \end{cases} \tag{6.1}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is some activation function. Let Ω be the parameter space of the network. For a data distribution $(x, y) \sim \mathcal{D}$, the population loss $\Phi : \Omega \to \mathbb{R}$ is defined as

$$\Phi(\theta) = \mathbb{E}_{\mathcal{D}}[\Psi(f_L(\theta, x), y)], \tag{6.2}$$

where $\Psi(\hat{y}, y)$ is assumed to be convex with respect to the first argument. Typical examples of Ψ which satisfy this assumption include the standard square loss and cross-entropy loss. By letting $\mathcal{D}$ be the uniform distribution over a fixed set of N samples, the loss Φ reduces to the empirical loss.

In the following, we frequently encounter situations where we are given a layer $l \in [0, L-1]$, and we want to study a subnetwork from layer l to the output layer L. The widths and the inputs to this subnet are defined as follows. Let $\mathcal{I}_p \subseteq [n_p]$ be a subset of neurons at layer p, for $p \in [l, L-1]$. By convention, let $\mathcal{I}_0 = [n_0]$. Let $f_{p,\mathcal{I}_p} \in \mathbb{R}^{|\mathcal{I}_p|}$ be obtained from f_p by taking the neurons indexed by $\mathcal{I}_p$. The layer widths of this subnet (associated to l) are given by $m_l^{(l)} = |\mathcal{I}_l|$, $m_p^{(l)} = n_p - |\mathcal{I}_p|$ for $p \in [l+1, L-1]$, and $m_L^{(l)} = n_L$. The associated weight matrices and bias vectors are $\{U_p^{(l)}\}_{p=l+1}^{L}$ and $\{v_p^{(l)}\}_{p=l+1}^{L-1}$, where $U_p^{(l)} \in \mathbb{R}^{m_{p-1}^{(l)} \times m_p^{(l)}}$ and $v_p^{(l)} \in \mathbb{R}^{m_p^{(l)}}$. Let $\xi_l = ((U_p^{(l)})_{p=l+1}^{L}, (v_p^{(l)})_{p=l+1}^{L-1})$, see Figure 6.2 for an illustration. Given an input $z \in \mathbb{R}^{m_l^{(l)}}$ to this subnetwork, the corresponding output, denoted by $h_L(\xi_l, z) \in \mathbb{R}^{n_L}$, is recursively defined for $p \in [l, L]$ as

$$h_p(\xi_l, z) = \begin{cases} z, & p = l, \\ \sigma\left((U_p^{(l)})^T h_{p-1}(\xi_l, z) + v_p^{(l)}\right), & p \in [l+1, L-1], \\ (U_L^{(l)})^T h_{L-1}(\xi_l, z), & p = L. \end{cases} \tag{6.3}$$

6.4 Main Result: Statement and Discussion

We now state our main result, which is proved in Section 6.5.

Theorem 6.1 *Fix any $\theta_0, \theta_1 \in \Omega$ and any $\epsilon > 0$. Then, there exists a continuous piecewise linear (PWL) path $\theta : [0,1] \to \Omega$ such that $\theta(0) = \theta_0, \theta(1) = \theta_1$, and the following holds:*

$$\max_{t\in[0,1]} \Phi(\theta(t)) \leqslant \max_{\theta' \in \{\theta_0, \theta_1\}} \max \left(\Phi(\theta'), \min_{\substack{\mathcal{I}_1 \subseteq [n_1], \ldots, \mathcal{I}_{L-1} \subseteq [n_{L-1}] \\ \text{s.t. } |\mathcal{I}_1| \geqslant \frac{n_1}{2}, \ldots, |\mathcal{I}_{L-1}| \geqslant \frac{n_{L-1}}{2}}} \max_{l \in [0, L-1]} \inf_{\xi_l} \mathbb{E}_{\mathcal{D}}[\Psi\left(h_L\left(\xi_l, f_{l,\mathcal{I}_l}(\theta', x)\right), y\right)] + \epsilon \right) \tag{6.4}$$

where $h_L\left(\xi_l, f_{l,\mathcal{I}_l}(\theta', x)\right)$ is the output of the subnetwork from layer l to layer L as defined in (6.3), fed with input $f_{l,\mathcal{I}_l}(\theta', x)$, and ξ_l contains the parameters of this subnetwork.

We remark that the second maximum on the RHS of (6.4) is intended to be taken between the two components in the open bracket (i.e., $\Phi(\theta')$ and the $\min \max \inf$).

An illustration of the subnets ξ_l's appearing in the theorem can be found in Figure 6.2. In words, for layer $l \in [L-1]$, we pick a subset $\mathcal{I}_l$ containing at least half the hidden neurons in that layer. Then, the loss over the PWL path is upper bounded by the loss at L points that are formed by connecting the subset of features (indexed by $\mathcal{I}_l$) at layer l to the output via the subnetwork defined in (6.3). The key idea is that we can construct low-loss paths connecting these sparse points, and the sub-networks guarantee a sufficient amount of sparsity: they have $|\mathcal{I}_l|$ neurons at layer l, and $n_p - |\mathcal{I}_p|$ neurons at layer $p \in [l+1, L-1]$. The additive factor ϵ in the RHS of (6.4) serves to deal with the case in which the $\inf$ is not attained (as for the cross-entropy loss). We note that Theorem 6.1 also holds for deep nets without biases, which can be proved by following the same argument as in Section 6.5.

Equivalent formulation. Before discussing the implications of Theorem 6.1, we provide an equivalent way to write the "$\min \max \inf$ " on the RHS of (6.4). This characterization will be useful when evaluating numerically our bound in Section 6.6. First, note that the subnetwork ξ_l in (6.4) is fully determined by $\mathcal{I}_l$ and the cardinalities of $\mathcal{I}_{l+1}, \ldots, \mathcal{I}_{L-1}$. To this end, let $K_l = |\mathcal{I}_l|, l \in [L-1]$, and

$$Q_l = \inf_{\xi_l} \mathbb{E}_{\mathcal{D}}[\Psi\left(h_L\left(\xi_l, f_{l,\mathcal{I}_l}(\theta', x)\right), y\right)], \quad l \in [0, L-1]. \tag{6.5}$$

Then, clearly Q_l is fully determined by the variables $\{\theta', \mathcal{I}_l, K_{l+1}, \dots, K_{L-1}\}$, and thus it does not depend on the actual content of $\mathcal{I}_k$'s for $k \neq l$. Based on this observation, one can easily argue that

$$\min_{\substack{\mathcal{I}_1 \subseteq [n_1], \dots, \mathcal{I}_{L-1} \subseteq [n_{L-1}] \\ \text{s.t. } |\mathcal{I}_1| \geqslant \frac{n_1}{2}, \dots, |\mathcal{I}_{L-1}| \geqslant \frac{n_{L-1}}{2}}} \max_{l \in [0, L-1]} Q_l = \min_{\frac{n_1}{2} \leqslant K_1 \leqslant n_1, \dots, \frac{n_{L-1}}{2} \leqslant K_{L-1} \leqslant n_{L-1}} \max_{l \in [0, L-1]} \min_{\substack{\mathcal{I}_l \subseteq [n_l] \\ |\mathcal{I}_l| = K_l}} Q_l. \tag{6.6}$$

To see why (6.6) holds, we first write the minimum over the sets $\{\mathcal{I}_l\}_{l \in [L-1]}$ on the LHS as a minimum over their cardinalities $\{K_l\}_{l \in [L-1]}$ followed by a minimum over all the possible choices of subsets with fixed cardinalities. From the above discussion, once $K_{l+1}, \dots, K_{L-1}$ are fixed, the quantity Q_l is entirely independent of $\mathcal{I}_k$ for all $k \neq l$. Thus, one can swap the maximum over $l \in [0, L-1]$ with the minimum over the sets, and the result readily follows.

A condition milder than dropout stability. Theorem 1 of [KWL+19] can be deduced as a corollary of Theorem 6.1 stated for networks without biases. Let us first recall the definition of ε-dropout stability from [KWL+19]. We say that some point $\theta \in \Omega$ is ε-dropout stable if, for all $l \in [L-1]$, there exists a subset of *at most* $\lfloor n_p/2 \rfloor$ hidden neurons in each of the layers $p \in [l, L-1]$ with the following property: after rescaling the outputs of these hidden units by some factor r and setting the outputs of the remaining units to zero, we obtain a parameter $\theta_{l,\mathrm{d}}$ such that $\Phi(\theta_{l,\mathrm{d}}) \leqslant \Phi(\theta) + \varepsilon$. The proof of the following corollary is postponed to Appendix 6.A.1.

Corollary 6.2 *Fix any $\theta_0, \theta_1 \in \Omega$ and any $\epsilon > 0$. Assume that θ_0 and θ_1 are ε-dropout stable. Then, there exists a continuous PWL path $\theta : [0,1] \to \Omega$ such that $\theta(0) = \theta_0, \theta(1) = \theta_1$, and it holds:*

$$\max_{t \in [0,1]} \Phi(\theta(t)) \leqslant \max\left(\Phi(\theta_0),\ \Phi(\theta_1)\right) + \epsilon + \varepsilon. \tag{6.7}$$

We highlight that requiring dropout stability to ensure the connectivity of solutions is a much stronger assumption than what is needed in Theorem 6.1. In fact, in order to have that the loss along the path connecting θ_0 and θ_1 is ε-close to the loss at the extremes, Theorem 1 of [KWL+19] requires that θ_0 and θ_1 are ε-dropout stable. On the contrary, our main result requires that, for $\theta' \in \{\theta_0, \theta_1\}$,

$$\inf_{\xi_l} \mathbb{E}_{\mathcal{D}}[\Psi\left(h_L\left(\xi_l, f_{l,\mathcal{I}_l}(\theta', x)\right), y\right)] \tag{6.8}$$

is no bigger than $\Phi(\theta')$. The key difference with respect to [KWL+19] is that in (6.8) we are not restricted to rescaling the weights of the original network (as in the dropout stability assumption), but we are allowed to optimize the weights of the subnetwork. This makes a significant difference in practical settings. In Section 6.6, we provide examples of networks trained on MNIST and CIFAR-10 such that dropout stability is not satisfied,

but our mild condition still holds. Thus, even if the path provided by [KWL+19] leads to a large loss, we are still able to connect the SGD solutions via a low-loss path.

Connections with memorization capacity. The fact that we can choose arbitrary weights to approach the $\inf$ in (6.8) provides a link between the connectivity of solutions and the memorization capacity of neural nets. Consider a setting in which there are N data samples $\{(x_j, y_j)\}_{j\in[N]}$ and we wish to minimize the empirical training loss, i.e. $\mathcal{D}$ is the uniform distribution over these samples. By exploiting recent progress on the finite-sample expressivity of neural networks [BEL+20; YSJ19], one deduces the corollary below, whose proof is deferred to Appendix 6.A.2. We recall that a set of points $\{x_i\}_{i\in[N]}$ in $\mathbb{R}^d$ is said to be in generic position if any hyperplane in $\mathbb{R}^d$ contains at most d points.

Corollary 6.3 *Fix any $\theta_0, \theta_1 \in \Omega$ and any $\epsilon > 0$, and let σ be ReLU. Let $\{(x_j, y_j)\}_{j\in[N]}$ be N data samples, where $y_j \in [-1,1]^{n_L}$ for $j \in [N]$ and the vectors $\{x_j\}_{j\in[N]}$ are distinct. Let $\mathcal{D}$ be the uniform distribution over these samples. For $i \in \{0,1\}$, assume that there exist subsets of neurons $\{\mathcal{I}_l^{(i)}\}_{l\in[L-1]}$ such that the following holds:*

(A1) $|\mathcal{I}_{L-1}^{(i)}| \geqslant \frac{n_{L-1}}{2}$ *and* $4\left\lfloor \frac{n_{L-2}}{8} \right\rfloor \left\lfloor \frac{n_{L-1} - |\mathcal{I}_{L-1}^{(i)}|}{4n_L} \right\rfloor \geqslant N.$

(A1-b) For some $\varepsilon \geqslant 0$,

$$\inf_{W \in \mathbb{R}^{|\mathcal{I}_{L-1}^{(i)}| \times n_L}} \mathbb{E}_{\mathcal{D}}[\Psi(W^T f_{L-1,\mathcal{I}_{L-1}^{(i)}}(\theta_i, x), y)] \leqslant \Phi(\theta_i) + \varepsilon. \tag{6.9}$$

(A2) $|\mathcal{I}_{L-2}^{(i)}| = \lceil \frac{n_{L-2}}{2} \rceil$*, and the feature vectors $\{f_{L-2,\mathcal{I}_{L-2}^{(i)}}(\theta_i, x_j)\}_{j\in[N]}$ are in generic position.*

(A3) For all $l \in [L-3]$: $|\mathcal{I}_l^{(i)}| = n_l - 1$, and the feature vectors $\{f_{l,\mathcal{I}_l^{(i)}}(\theta_i, x_j)\}_{j\in[N]}$ are distinct.

Then, there exists a continuous PWL path $\theta : [0,1] \to \Omega$ s.t. $\theta(0) = \theta_0, \theta(1) = \theta_1$, and it holds:

$$\max_{t\in[0,1]} \Phi(\theta(t)) \leqslant \max\left(\Phi(\theta_0),\ \Phi(\theta_1)\right) + \epsilon + \varepsilon. \tag{6.10}$$

This result should be regarded as a proof of concept. Improved bounds on memorization capacity* would immediately lead to stronger results about connectivity. Let us now discuss in more detail the assumptions needed in Corollary 6.3. Assumption (A2) ensures sufficient "diversity" among the features at the second last hidden layer: the vectors $\{f_{L-2,\mathcal{I}_{L-2}^{(i)}}(\theta_i, x_j)\}_{j\in[N]}$ need to be in generic position, which is rather standard in the literature about the memorization capacity of neural networks, see e.g. [Bau88]. Assumption (A3) requires a minimal condition that the feature vectors $\{f_{l,\mathcal{I}_l^{(i)}}(\theta_i, x_j)\}_{j\in[N]}$ are distinct for $l \in [L-3]$. Assumption (A1-b) involves the quality of the features at the last hidden layer: it requires that, by removing $n_{L-1} - |\mathcal{I}_{L-1}^{(i)}|$ neurons and picking

*E.g., less parameters in the interpolating network or less assumptions on the input data.

suitable new outcoming weights for the remaining ones, we do not increase the loss by more than ε. Assumption (A1) involves the over-parameterization of the last two hidden layers of the network: it requires that $n_{L-2}(n_{L-1} - |\mathcal{I}_{L-1}^{(i)}|)$ is at least roughly $8Nn_L$. Pick $|\mathcal{I}_{L-1}^{(i)}| = (1-\delta)n_{L-1}$, and assume that n_L is a constant independent of N. Then, assumption (A1-b) holds if the loss remains stable after removing a δ-fraction of the neurons, and assumption (A1) holds if $n_{L-2}n_{L-1} = \Omega(N)$. We remark that there is no requirement on the widths of the remaining layers.

It is shown in [Ngu21] that, for deep nets, if the first hidden layer has $N+1$ neurons then one gets connected sublevel sets. This bound is tight as a 2-layer network of width N can have disconnected sublevel sets. Here, we use properties of the features of the solutions we wish to connect in order to relax the requirement on the over-parameterization. In particular, if the features at the last two hidden layers are good enough (i.e. removing a small number of neurons at layer $L-1$ does not increase the loss much, and a subset of features at layer $L-2$ are in generic position), then $\Omega(\sqrt{N})$ neurons in each of these two layers are sufficient to ensure the connectivity. More generally, the fact that we are free to choose the sizes of the sets $\mathcal{I}_1, \dots, \mathcal{I}_{L-1}$ in the RHS of (6.4) allows one to trade-off between the quality of the features at a given layer and the amount of over-parameterization in the layers above it.

Linearly separable features. One extreme example of this trade-off is when the features are linearly separable (i.e. the features are highly representative of the labels). More formally, consider a classification task where one is given the labels $y_j \in \{0,1\}^{n_L}$, and the data samples are linearly separable. Recall that, a set of data samples $\{(x_j, y_j)\}_{j\in[N]}$ is called "linearly separable" if there exist n_L hyperplanes, each separating the samples of one class from the other classes perfectly. Pick two points such that, by removing n_L neurons per layer, their features remain linearly separable at each layer. Then, these points can be connected by a low-loss path (without any extra over-parameterization requirement). This is formalized in the following corollary, whose proof appears in Appendix 6.A.3.

Corollary 6.4 *Fix any $\theta_0, \theta_1 \in \Omega$ and any $\epsilon > 0$. Assume that $n_l \geqslant 2n_L$ for $l \in [L-1]$ and that there exists an interval $[\alpha, \beta]$ s.t. σ is strictly monotonic on $[\alpha, \beta]$. Let $\{(x_j, y_j)\}_{j\in[N]}$ be N linearly separable data samples, where $y_j \in \{0,1\}^{n_L}$ for $j \in [N]$. Let $\mathcal{D}$ be the uniform distribution over these samples, and let Ψ be the standard cross-entropy loss. Assume that, for $i \in \{0,1\}$, there exists subsets $\{\mathcal{I}_l^{(i)}\}_{l\in[L-1]}$ such that $\frac{n_l}{2} \leqslant |\mathcal{I}_l^{(i)}| \leqslant n_l - n_L$ and the set of features $\{f_{l,\mathcal{I}_l^{(i)}}(\theta_i, x_j), y_j\}_{j\in[N]}$ is linearly separable, for all $l \in [0, L-1]$. Then, there exists a continuous PWL path $\theta : [0,1] \to \Omega$ such that $\theta(0) = \theta_0, \theta(1) = \theta_1$, and the following holds:*

$$\max_{t\in[0,1]} \Phi(\theta(t)) \leqslant \max\left(\Phi(\theta_0),\ \Phi(\theta_1)\right) + \epsilon. \tag{6.11}$$

We remark that the result of Corollary 6.4 is not restricted to the cross-entropy loss, but it holds for any loss function Ψ whose infimum is 0 when the input is linearly separable.

Furthermore, the assumption on the activation function being monotonic in an interval is very mild, and it serves to avoid degenerate cases (e.g. constant activation function).

6.5 Proof of Theorem 6.1

In the following, the "incoming weights" of a neuron refer to both the incoming weight vector and the bias of that neuron, and the "outcoming weights" of a neuron simply refer to its outcoming weight vector. We call a neuron "inactive" if its incoming weights are zero, and we call a neuron "pure zero" if both its incoming and outcoming weights are zero. Clearly, a pure zero neuron is inactive. Whenever we say "set some parameter w from a to b", we simply mean to change the current value of w from a to b by following a direct line segment: $w(\lambda) = (1-\lambda)a + \lambda b$, where $\lambda \in [0,1]$. First, we state and prove a useful claim. Then, we present the proof of Theorem 6.1.

Claim 6.5 *Consider two distinct hidden neurons p and q on the same layer, and assume that p is a pure zero neuron. Then, one can swap p and q by a continuous PWL path along which the network output is invariant.*

Proof of Claim 6.5. Let p_{in} and p_{out} denote the incoming resp. outcoming weights of p. Similarly, we denote by q_{in} and q_{out} the incoming and outcoming weights of q. We change p and q continuously by applying the following paths sequentially (after each step, we reset the value of p and q to the ending point of the path): (i) $p_{\text{in}}(\lambda) = (1-\lambda)p_{\text{in}} + \lambda q_{\text{in}}$, (ii) $[p_{\text{out}}(\lambda), q_{\text{out}}(\lambda)] = [\lambda q_{\text{out}}, (1-\lambda)q_{\text{out}}]$, and (iii) $q_{\text{in}}(\lambda) = (1-\lambda)q_{\text{in}}$. For all these paths, λ goes from 0 to 1. Note that path (i) does not change the network output, since the outcoming weights of p are zero. At the end of the first path, p and q have the same incoming weights (i.e. they are identical neurons). Furthermore, path (ii) does not change the network output, since the total contribution of p and q to the next layer is constant along the path. Finally, also path (iii) does not change the network output, since the outcoming weights of q are zero at the end of path (ii). After performing these three steps, p and q are swapped, i.e. q becomes pure zero while p has the original weights of q. □

Proof of Theorem 6.1. Fix the subsets $\{\mathcal{I}_l\}_{l\in[L-1]}$, where $\mathcal{I}_l \subseteq [n_l]$, $|\mathcal{I}_l| \geqslant n_l/2$, and let $\bar{\mathcal{I}}_l$ be the complement of $\mathcal{I}_l$. We show that there is a continuous PWL path from θ_0 to some other point for which all the neurons in $\{\bar{\mathcal{I}}_l\}_{l\in[L-1]}$ are pure zero, and the loss along the path is at most

$$\max\left(\Phi(\theta_0), \max_{l\in[0,L-1]} \inf_{\xi_l} \mathbb{E}_{\mathcal{D}}[\Psi\left(h_L\left(\xi_l, f_{l,\mathcal{I}_l}(\theta_0, x)\right), y\right)] + \epsilon\right). \tag{6.12}$$

To prove the claim, we start from θ_0 and apply a sparsification procedure that makes the neurons in $\bar{\mathcal{I}}_{L-1}, \ldots, \bar{\mathcal{I}}_1$ become pure zero, one layer at a time from layer $L-1$ down to layer 1. First, we construct a path that makes the outcoming weights of the neurons

in $\bar{\mathcal{I}}_{L-1}$ become zero by changing only the weights in the last layer. Note that Ψ is convex in the weights of the last layer, hence Φ is also convex in those weights. Therefore, the loss along this path is at most

$$\max\left(\Phi(\theta_0), \inf_{\xi_{L-1}} \mathbb{E}_{\mathcal{D}}[\Psi\left(h_L\left(\xi_{L-1}, f_{L-1,\mathcal{I}_{L-1}}(\theta_0, x)\right), y\right)] + \epsilon\right).$$

Once the outcoming weights of a neuron are all zero, the incoming weights can also be set to zero without changing the loss. Thus, the neurons in $\bar{\mathcal{I}}_{L-1}$ can be made pure zero.

Now, we assume via induction that we have already made all the neurons in $\{\bar{\mathcal{I}}_{l+1}, \ldots, \bar{\mathcal{I}}_{L-1}\}$ pure zero, and we want to make the neurons in $\bar{\mathcal{I}}_l$ pure zero. It suffices to make all the inter-connections between $\bar{\mathcal{I}}_l$ and $\mathcal{I}_{l+1}$ become zero, because the connections between $\bar{\mathcal{I}}_l$ and $\bar{\mathcal{I}}_{l+1}$ are already zero due to the inactivity of $\bar{\mathcal{I}}_{l+1}$. To do this, one first observes that the pure-zero state of the neurons in $\{\bar{\mathcal{I}}_{l+1}, \ldots, \bar{\mathcal{I}}_{L-1}\}$ separates the original network from layer $l+1$ to layer $L-1$ into two parallel subnetworks: one with neurons in $\{\mathcal{I}_p\}_{p=l+1}^{L-1}$, and the other with pure zero neurons. Now, set the incoming weights of the neurons in $\{\bar{\mathcal{I}}_{l+1}, \ldots, \bar{\mathcal{I}}_{L-1}\}$ to the corresponding values given by the subnetwork ξ_*, where ξ_* satisfies the following

$$\mathbb{E}_{\mathcal{D}}[\Psi\left(h_L\left(\xi_*, f_{l,\mathcal{I}_l}(\theta_0, x)\right), y\right)] \leqslant \inf_{\xi_l} \mathbb{E}_{\mathcal{D}}[\Psi\left(h_L\left(\xi_l, f_{l,\mathcal{I}_l}(\theta_0, x)\right), y\right)] + \epsilon. \tag{6.13}$$

Note that a subnetwork ξ_* which satisfies (6.13) always exists: if the infimum on the RHS of (6.13) is actually a minimum, then we can pick ξ_* to be one of the minimizers. Otherwise, there exists a ξ_* whose loss approaches the infimum with arbitrarily small error. In the above step, only the incoming weights that are relevant to the subnetwork ξ_* are reset, while the other ones are kept unchanged. In particular, this step does not affect the outputs of the neurons in the other subsets $\{\mathcal{I}_{l+1}, \ldots, \mathcal{I}_{L-1}\}$ as well as the output of the network. Next, let us set all the outcoming weights in $\bar{\mathcal{I}}_{L-1}$ to the corresponding value given by ξ_*, and all the outcoming weights in $\mathcal{I}_{L-1}$ to zero (in one step). By the convexity of Ψ w.r.t. the output weights, we have that the loss along this path is upper bounded by the loss at the extremes of the path, which is in turn upper bounded by (6.12). Now, since the outcoming weights of $\mathcal{I}_{L-1}$ are zero, it follows that the parallel subnetwork associated to $\{\mathcal{I}_{l+1}, \ldots, \mathcal{I}_{L-1}\}$ does not contribute to the network output. Thus, we can change the incoming weights of all the neurons in this subnetwork to any arbitrary values without affecting the loss. Hence, we turn all these neurons into inactive and then pure zero, without changing the loss. As a result, all the inter-connections between $\bar{\mathcal{I}}_l$ and $\mathcal{I}_{l+1}$ will become zero, as wanted from above. At this point, note that the weights of the neurons in $\{\bar{\mathcal{I}}_{l+1}, \ldots, \bar{\mathcal{I}}_{L-1}\}$ are those of ξ_*. Thus, in order to finish the induction proof, we need to make them become pure zero again. This can be done by swapping, at each layer $p \in [l+1, L-1]$, one pure zero neuron in $\mathcal{I}_p$ with one neuron in $\bar{\mathcal{I}}_p$. As $|\mathcal{I}_p| \geqslant \frac{n_p}{2} \geqslant |\bar{\mathcal{I}}_p|$, this is always possible, and the network output is invariant by Claim 6.5.

The sparsification procedure described above can be performed for any subsets $\{\mathcal{I}_l\}_{l\in[L-1]}$ s.t. $|\mathcal{I}_l| \geqslant \frac{n_l}{2}$ for $l \in [L-1]$. Thus, we can take the minimum in (6.12) over such

subsets. Let $\{\mathcal{I}_l^*\}_{l\in[L-1]}$ be the optimal subsets for θ_0. Let us apply the same sparsification procedure to θ_1. We can assume w.l.o.g. that the optimal subsets for θ_1 and θ_0 are the same. In fact, if this is not the case, we use Claim 6.5 and swap the pure zero neurons with other ones on the same layer in order to satisfy this property. At this point, both θ_0 and θ_1 have been sparsified, and they have the same set of pure zero neurons at each layer, indexed by $\{\bar{\mathcal{I}}_l^*\}_{l\in[L-1]}$.

Finally, we connect the sparsified versions of θ_0 and θ_1. Let A be the subnetwork consisting of all the hidden neurons in $\{\bar{\mathcal{I}}_l^*\}_{l\in[L-1]}$, and B the subnetwork consisting of all the remaining neurons. Both A and B share the same input and output layers as the original network. Note also that the hidden layers of B have larger widths than the corresponding layers of A. Since A and B are decoupled, one can turn one on and the other off by setting the appropriate output weights to zero, and then update the incoming weights of the other subnetwork to any desired value without affecting the loss. In this way, one can make all the incoming weights (restricted to each parallel subnetwork) of θ_0 equal to the corresponding values of θ_1. Finally, one uses the convexity of the loss Ψ again to equalize the output weights of both θ_0 and θ_1. Hence, the loss incurred to connect the sparsified versions of θ_0 and θ_1 is no bigger than

$$\max_{\theta'\in\{\theta_0,\theta_1\}} \inf_{\xi_0} \mathbb{E}_{\mathcal{D}}[\Psi\left(h_L\left(\xi_0, f_0(\theta', x)\right), y\right)] + \epsilon.$$

By taking the maximum of the loss along the paths considered in this proof, we get the bound (6.4). □

6.6 Numerical Experiments

We compare the losses along the path of Theorem 6.1 and the one in [KWL+19] which our theoretical analysis has improved upon. Implementation details and a link to the code are given in Appendix 6.C.

Datasets and architectures. We consider MNIST [LBB+01] and CIFAR-10 [Kri09] datasets. As a pre-processing stage, we perform a scaling into $[-1, 1]$ for MNIST, and data whitening for CIFAR-10. For MNIST, we consider a fully connected network (FCN) with 10 hidden layers ($L = 11$), each of width 245. For CIFAR-10, we consider two architectures: (i) an FCN with 5 hidden layers ($L = 6$), each of width 500, and (ii) the VGG-11 network [SZ15]. All architectures have ReLU activations. We train each network by standard SGD with cross-entropy loss, batch size 100 and no explicit regularizers. All the networks are trained beyond the point where the dataset is perfectly separated, and no early stopping mechanism is used. For each trained network, we perform the following two experiments.

Experiment (A): we evaluate the RHS of (6.4), with θ' being the trained model, by using the equivalent characterization (6.6). First, we fix the integers $K_l = \lceil n_l/2 \rceil$. Then, for

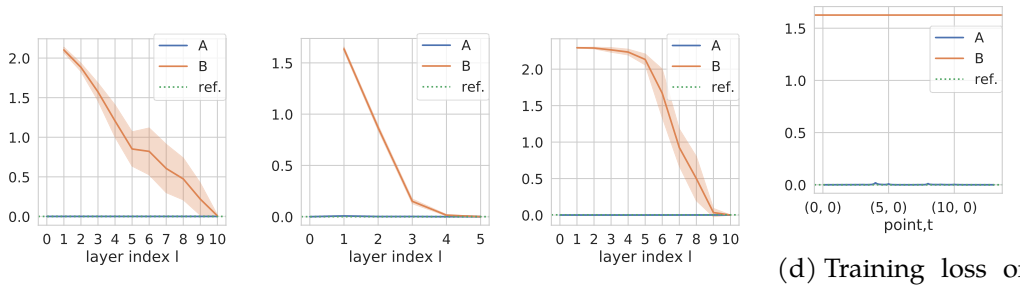
each $l \in [0, L-1]$, we compute the quantity $\tilde{Q}_l$, defined as:

$$\tilde{Q}_l := \min_{\substack{\mathcal{I}_l \subseteq [n_l] \\ |\mathcal{I}_l| = K_l}} Q_l = \min_{\substack{\mathcal{I}_l \subseteq [n_l] \\ |\mathcal{I}_l| = K_l}} \inf_{\xi_l} \mathbb{E}_{\mathcal{D}}[\Psi\left(h_L\left(\xi_l, f_{l,\mathcal{I}_l}(\theta', x)\right), y\right)]. \tag{6.14}$$

We upper bound the $\inf$ in (6.14) by running an independent SGD routine, and upper bound the $\min$ there by sampling the subsets $\mathcal{I}_l$ uniformly at random for $T_A = 20$ times and by picking the best training loss resulting from the $\inf$. The resulting bounds on $\tilde{Q}_l$ are shown by the blue curves in Figure 6.3. Note that by taking the maximum of these values for $l \in [0, L-1]$, one obtains an upper bound on (6.6), which is exactly the $\min \max \inf$ term in Theorem 6.1. As Theorem 6.1 requires at least two different points θ' to connect, in Figure 6.3 we run our experiments multiple times (i.e. by training multiple network θ' above), and we report the mean together with the 95% confidence interval. The performance of these original models θ''s is shown by the horizontal dotted green line in Figure 6.3 (used as a reference). As one can see, the blue curve is always close to the reference curve, which means that our upper bound as given in Theorem 6.1 for connecting these θ' solutions is always close to their original training losses, and thus all these solutions are approximately connected.

Experiment (B): we test the dropout stability of the trained model as defined in [KWL+19]. For each $l \in [L-1]$, we proceed as follows: (i) randomly remove half of the neurons at all layers $p \in [l, L-1]$, and (ii) optimize the scaling factor of the output of the remaining network by minimizing the training loss. This last optimization is performed in our experiments via the Scipy method `minimize_scalar(method="brent")`. For each $l \in [L-1]$, this procedure (i.e. do (i), and then (ii)) is repeated $T_B = 200$ times, and we pick the best result based on the training loss. We have chosen the value $T_B = 200$ by observing that, if we increase T_B from 100 to 200, the change in the minimum loss is negligible. The results here are shown by the orange curves in Figure 6.3. One observes that dropout stability tends to be satisfied at the layers near the output of the network, while dropping the neurons from the first few layers increases the loss significantly. This is consistent with the findings of [KWL+19], where it is reported that dropout stability does not hold for the first few layers of the VGG-11 architecture without dropout training. If one instead uses dropout during training, [KWL+19] show that the trained model satisfies the dropout stability property. In Experiment (A), we have shown that, even when dropout is *not* used as an explicit regularizer during training (and consequently dropout stability may not hold), we can still ensure the existence of low-loss connecting paths. This is because our bound in Theorem 6.1 allows the optimization of the weights of the subnetwork ξ_l.

Figure 6.3d illustrates the paths between two SGD solutions obtained from Experiment (A) and (B) for the FCN with $L = 6$ trained on CIFAR-10. While the path (A) constructed via Theorem 6.1 has loss close to zero (and close to the loss of the reference model), the path of [KWL+19] experiences a large loss. Additional results in Appendix 6.B show similar patterns for the test error (see Figure 6.6) and for different dropout ratios (see Figures 6.7 and 6.8). Furthermore, the log-scale plot in Figure 6.5 of Appendix 6.B allows one to zoom into the detailed difference between our bound and the reference.



(a) FCN MNIST (b) FCN CIFAR-10 (c) VGG CIFAR-10 (d) Training loss on path

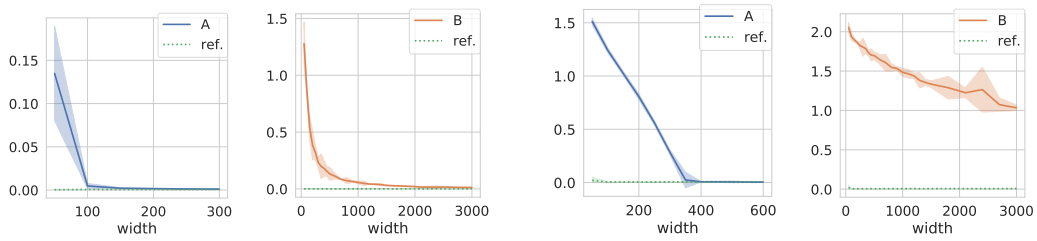
Figure 6.3: a–c: Training loss for Experiments (A) and (B). The dotted green line is the originally trained model θ', used as a reference. Mean and 95% confidence interval are shown for 10 different trained networks θ' in a-b, and for 5 different θ' in c. The plot d shows the loss along the paths of Experiments (A) and (B) between a fixed pair of solutions for the FCN trained on CIFAR-10.

In Figure 6.4, we perform Experiments (A)-(B) for an FCN with 2 hidden layers of equal width, trained on MNIST and CIFAR-10. Here, we vary the width, and report the maximum of the bounds over all the layers. As expected, the loss gets smaller as width increases. However, the bound provided by Theorem 6.1 is close to 0 already for networks of moderate sizes: about 200 neurons for MNIST, and 500 for CIFAR-10. These networks widths roughly correspond to $\sqrt{N}$, as predicted by Corollary 6.3. On the contrary, much larger widths are necessary for dropout stability to hold: 2000 neurons for MNIST, and for CIFAR-10, networks with 3000 neurons are still quite far from being dropout stable.

Taking all these experimental results together, we have demonstrated that, for simple datasets (e.g. MNIST) and for large enough widths, dropout stability holds. However, for more difficult learning tasks (e.g. CIFAR-10), and for networks of moderate widths or larger depths, there is a strong evidence that dropout stability is not satisfied. Yet, our main theorem provides connecting paths with low loss, even when the over-parameterization of the trained networks is moderate.

6.7 Concluding Remarks

We provide a new characterization for SGD solutions (or more generally, for any two points in parameter space) to be connected by low-loss paths. Numerical experiments demonstrate that the proposed sufficient condition is satisfied by trained networks, also in settings where the requirements of the previous works do not hold. Our results and analysis lead directly to the following open question: *Is having (more than two) hidden layers with $\Omega(\sqrt{N})$ neurons in general sufficient to guarantee the connectivity of all the sublevel sets?* Although width N is both necessary and sufficient for two-layer nets [Ngu21], our Corollary 6.3 gives hope that this bound may be not tight for deeper nets, and sublinear



(a) MNIST (b) CIFAR-10

Figure 6.4: Training loss (maximum over all layers) for Experiments (A) and (B) for an FCN with two hidden layers of equal albeit varying widths. The dotted line is the originally trained model. Mean and 95% confidence interval are shown for 3 different solutions θ'.

widths there might suffice. Answering this question would help us better understand the role of depth on shaping the geometry of neural net loss function. It is also interesting to study: (i) Why does SGD converge to solutions fulfilling the trade-off we describe here between feature quality and over-parameterization? (ii) Does SGD provably converge to a global optimum for such networks? This last question was studied, e.g. in [ALS19; DLL+19a; NM20; OS20; ZCZ+20] for regression tasks, but the current results still require at least one layer with $\Omega(N)$ neurons. The analysis here suggests that SGD may explore a well-behaved region of the loss landscape even for networks with sublinear layer widths.

6.A Missing Proofs

6.A.1 Proof of Corollary 6.2

Proof. Let $\theta' \in \{\theta_0, \theta_1\}$. Then, as θ' is ε-dropout stable, we obtain that, for every $l \in [L-1]$, there exist subsets of indices $\{\mathcal{I}_p^{l,\mathrm{d}}\}_{p=l}^{L-1}$, with $\mathcal{I}_p^{l,\mathrm{d}} \subseteq [n_p]$ and $|\mathcal{I}_p^{l,\mathrm{d}}| \leqslant \lfloor n_p/2 \rfloor$, such that the following property holds: after rescaling the outputs of the hidden units indexed by $\{\mathcal{I}_p^{l,\mathrm{d}}\}_{p=l}^{L-1}$ and setting the outputs of the remaining units to zero, we obtain $\theta'_{l,\mathrm{d}}$ satisfying $\Phi(\theta'_{l,\mathrm{d}}) \leqslant \Phi(\theta') + \varepsilon$. Now, one observes that, by using Theorem 6.1, it suffices to show that there exists a choice of $\{\mathcal{I}_j\}_{j\in[L-1]}$ with $\mathcal{I}_j \subseteq [n_j]$ and $|\mathcal{I}_j| \geqslant n_j/2$ such that, for $l \in [0, L-1]$, it holds

$$\inf_{\xi_l} \mathbb{E}_{\mathcal{D}}[\Psi\left(h_L\left(\xi_l, f_{l,\mathcal{I}_l}(\theta', x)\right), y\right)] \leqslant \Phi(\theta'_{l,\mathrm{d}}), \tag{6.15}$$

with $\mathcal{I}_0 = [n_0]$.

For $l \in [L-1]$, pick $\mathcal{I}_l$ such that $\mathcal{I}_l \supseteq \mathcal{I}_l^{l,\mathrm{d}}$ and $|\mathcal{I}_l| = \lceil n_l/2 \rceil$. This is possible since $|\mathcal{I}_l^{l,\mathrm{d}}| \leqslant \lfloor n_l/2 \rfloor$. Note that since $|\mathcal{I}_l| = \lceil n_l/2 \rceil$, the requirement that $|\mathcal{I}_l| \geqslant n_l/2$

from (6.4) is fulfilled. As $\mathcal{I}_l \supseteq \mathcal{I}_l^{l,\mathrm{d}}$, $f_{l,\mathcal{I}_l}(\theta', x)$ contains all the non-zero elements of the feature vector $f_l(\theta'_{l,\mathrm{d}}, x)$ (i.e. after setting the outputs of the remaining neurons to zero as described above). Furthermore, recall that ξ_l contains the parameters of a subnetwork from layer l to L with layer widths $m_l^{(l)} = |\mathcal{I}_l|$, $m_p^{(l)} = n_p - |\mathcal{I}_p| = \lfloor n_p/2 \rfloor$ for $p \in [l+1, L-1]$ and $m_L^{(l)} = n_L$. Thus, since $\theta'_{l,\mathrm{d}}$ contains at most $\lfloor n_p/2 \rfloor$ non-zero neurons at layer $p \in [l+1, L-1]$, it follows that the network computed at $\theta'_{l,\mathrm{d}}$ can be implemented by the subnetwork. In other words, there exists ξ_l such that $h_L\left(\xi_l, f_{l,\mathcal{I}_l}(\theta', x)\right) = f_L(\theta'_{l,\mathrm{d}}, x)$, for all x. Hence, (6.15) holds and the desired claim follows. □

6.A.2 Proof of Corollary 6.3

Proof. We show that, for $i \in \{0,1\}$ and for $l \in [0, L-1]$,

$$\inf_{\xi_l} \mathbb{E}_{\mathcal{D}}\left[\Psi\left(h_L\left(\xi_l, f_{l,\mathcal{I}_l^{(i)}}(\theta_i, x)\right), y\right)\right] \leqslant \Phi(\theta_i) + \varepsilon, \tag{6.16}$$

with $\mathcal{I}_0^{(i)} = [n_0]$. When $l = L-1$, (6.16) is equivalent to (6.9), which holds by assumption (A1-b). When $l = L-2$, we note that $\{f_{L-2,\mathcal{I}_{L-2}^{(i)}}(\theta_i, x_j)\}_{j\in[N]}$ are in generic position by assumption (A2), and assumption (A1) holds. Thus, (6.16) follows by applying Proposition 4 of [BEL+20] to each of the n_L outputs of the network. When $l \in [0, L-3]$, we note that $\{f_{l,\mathcal{I}_l^{(i)}}(\theta_i, x_j)\}_{j\in[N]}$ are distinct by assumption (A3), the samples $\{x_j\}_{j\in[N]}$ are distinct by assumption, and $y_j \in [-1,1]^{n_L}$ for $j \in [N]$. Thus, by using again assumption (A1), we can apply Corollary A.1 of [YSJ19] and (6.16) readily follows.[†] Finally, one can easily check that $|\mathcal{I}_l^{(i)}| \geqslant n_l/2$ for all $l \in [L-1], i \in \{0,1\}$, which satisfies the requirement in (6.4). Thus, the desired claim follows from Theorem 6.1. □

6.A.3 Proof of Corollary 6.4

Proof. We show that, for $i \in \{0,1\}$ and for $l \in [0, L-1]$,

$$\inf_{\xi_l} \mathbb{E}_{\mathcal{D}}\left[\Psi\left(h_L\left(\xi_l, f_{l,\mathcal{I}_l^{(i)}}(\theta_i, x)\right), y\right)\right] = 0, \tag{6.17}$$

with $\mathcal{I}_0^{(i)} = [n_0]$. Recall that, by hypothesis, $\{f_{l,\mathcal{I}_l^{(i)}}(\theta_i, x_j), y_j\}_{j\in[N]}$ are linearly separable for $l \in [0, L-1]$. For $k \in [n_L]$, denote by $(y_j)_k$ the k-th component of the vector $y_j \in \{0,1\}^{n_L}$.

[†]Note that we apply Corollary A.1 of [YSJ19] with $m = 1$ and $l_1 = L-2$. In this setting, $[m-1]$ and $[l_m + 2 : L-1]$ are the empty set, hence the 1st, 3rd and 4th conditions of Corollary A.1 hold. To verify the 2nd condition, note that $r_1 = 0$, $d_{L-2} = n_{L-2} - |\mathcal{I}_{L-2}^{(i)}| = \lfloor \frac{n_{L-2}}{2} \rfloor$, $d_{L-1} = n_{L-1} - |\mathcal{I}_{L-1}^{(i)}|$, and $d_y = n_L$. Thus, the 2nd condition follows from assumption (A1), and the application of Corollary A.1 is justified.

First, consider the case $l = L - 1$. Then, the problem in (6.17) is a convex learning problem with linearly separable inputs, namely $\{f_{L-1,\mathcal{I}_{L-1}^{(i)}}(\theta_i, x_j), y_j\}_{j\in[N]}$. Since Ψ is the cross-entropy loss, (6.17) holds immediately.

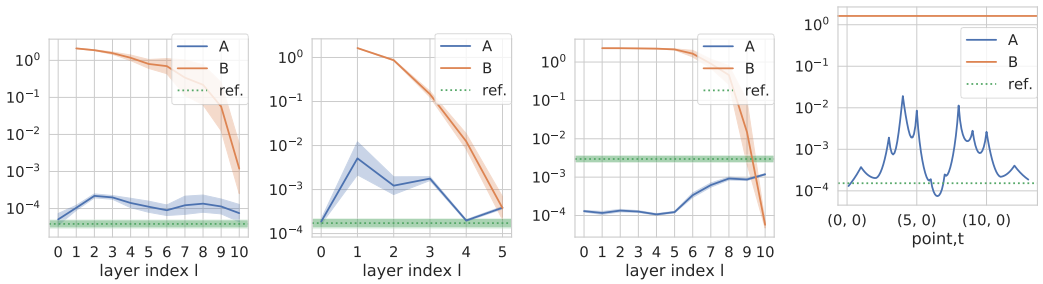
Consider now the case $l \in [0, L-2]$. Then, there exists $\gamma \in (\alpha, \beta)$, a set of n_L weights $w_1, \dots, w_{n_L} \in \mathbb{R}^{|\mathcal{I}_l^{(i)}|}$ and a set of n_L biases $c_1, \dots, c_{n_L} \in \mathbb{R}$ such that the following holds: if $(y_j)_k = 1$, then $\langle f_{l,\mathcal{I}_l^{(i)}}(\theta_i, x_j), w_k\rangle + c_k \in (\gamma, \beta)$; otherwise, $\langle f_{l,\mathcal{I}_l^{(i)}}(\theta_i, x_j), w_k\rangle + c_k \in (\alpha, \gamma)$. In words, this means that the pre-activation output of neuron k at layer $l+1$ maps all the samples of class k into (γ, β) and the remaining samples into (α, γ). Let us assume w.l.o.g. that σ is strictly monotonically increasing on (α, β). Then, we have that, if $(y_j)_k = 1$, $\sigma(\langle f_{l,\mathcal{I}_l^{(i)}}(\theta_i, x_j), w_k\rangle + c_k) \in (\sigma(\gamma), \sigma(\beta))$; otherwise, $\sigma(\langle f_{l,\mathcal{I}_l^{(i)}}(\theta_i, x_j), w_k\rangle + c_k) \in (\sigma(\alpha), \sigma(\gamma))$. Hence, at layer $l+1$, we have n_L neurons, each perfectly separating the samples of one class from the others. This shows that the set of features formed by these neurons is linearly separable. By repeating this argument, we can choose the weights of n_L neurons from the next layers $l+2, \dots, L-1$ such that this linear separability property is maintained. Consequently, the subnetwork implemented by ξ_l (which has at least n_L neurons at each of its hidden layers) can perfectly separate the features $\{f_{l,\mathcal{I}_l^{(i)}}(\theta_i, x_j), y_j\}_{j\in[N]}$. Thus, (6.17) holds, and the desired claim follows from Theorem 6.1. □

6.B Additional Numerical Experiments

Training loss in logarithmic scale. Figure 6.5 plots the loss reported in Figure 6.3 on a logarithmic y-axis. This allows one to see more clearly the difference between our bound (A) and the reference model. We note that this difference remains small across all layers of the network and experiments. Likewise, Figure 6.9 plots the loss reported in Figure 6.4 on a logarithmic y-axis. The logarithmic scale of the plot magnifies the drop in the metric around a certain threshold, about 100 neurons for MNIST and 400 for CIFAR-10.

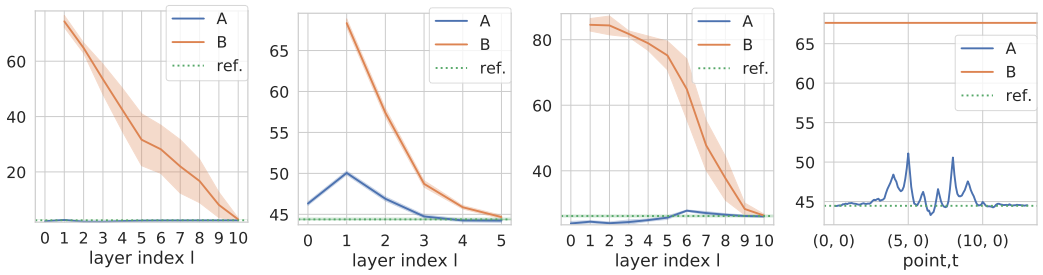
Test error. Figure 6.6 shows the corresponding test errors for the experiment in Figure 6.3. One can observe a similar behavior that our bound (A) continues to remain close to the reference model, while the test error (B) coming from the dropout stability assumption is large at the first layers.

Different dropout ratios. In Figures 6.7 and 6.8, we perform Experiments (A) and (B) again with other values of K_l's – the cardinalities of the subsets of neurons $\mathcal{I}_l$'s. Intuitively, in our proofs K_l can be seen as the number of neurons that we choose to keep at each layer, and thus it can also be defined via a certain dropout ratio. For simplicity, we fix this ratio to be the same at every layer, say $p \in [0, 1]$. Let $K_l = \lceil (1-p)n_l \rceil$, for $l \in [L-1]$. Then, one notes that the results in Figures 6.3 and 6.4 have been obtained by taking $p = \frac{1}{2}$. Now, we repeat these experiments for $p \in \{\frac{1}{3}, \frac{1}{4}\}$, and report the results in Figures 6.7 and 6.8, respectively. As for Experiment (A), this means that, while more neurons are



(a) FCN MNIST (b) FCN CIFAR-10 (c) VGG CIFAR-10 (d) Training loss on path

Figure 6.5: Results of Figure 6.3 plotted with a logarithmic y-axis.



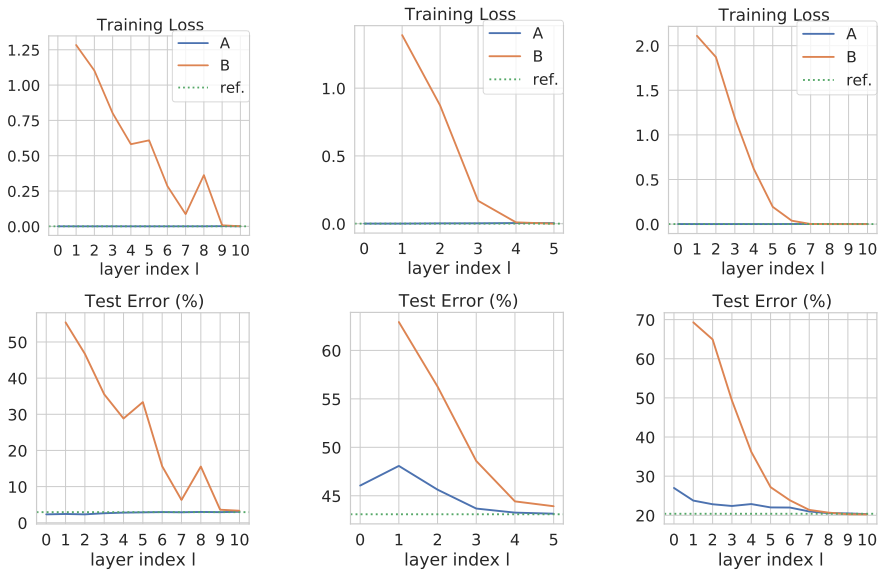
(a) FCN MNIST (b) FCN CIFAR-10 (c) VGG CIFAR-10 (d) Test error on path

Figure 6.6: Test error (in %) for the experiment in Figure 6.3.

preserved at each hidden layer (due to smaller p), the capacities of the sub-networks ξ_l's from Theorem 6.1 are smaller. As for Experiment (B), since more neurons are kept at each layer, there is a better chance that the dropout stability assumption is satisfied. Let us remark that the results of [KWL+19] show the existence of a low-loss path for $p \geqslant \frac{1}{2}$, whereas our Theorem 6.1 also applies to $p \leqslant \frac{1}{2}$. The plots show that although experiment (B) benefits from a lower p, there is still a significant performance gap with the originally trained model, especially at the lower layers. Meanwhile, this gap is consistently smaller for (A) across different values of p. This is on par with the case $p = \frac{1}{2}$ as reported in Section 6.6.

6.C Training Details

All models were trained from scratch on MNIST and CIFAR-10. The hyperparameters for training the original networks θ' (O) and for Experiment (A) can be found in Table 6.1. They were manually chosen in such a way that SGD achieves a reasonable convergence speed. All experiments were run on Nvidia GPUs Tesla V100. The implementation can



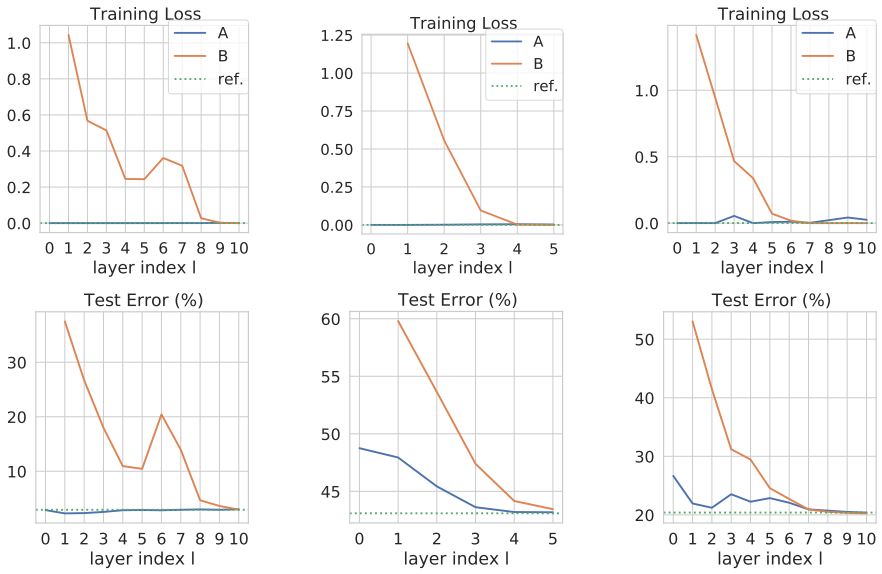
(a) FCN MNIST (b) FCN CIFAR-10 (c) VGG CIFAR-10

Figure 6.7: Dropout ratio $p = \frac{1}{3}$. Training loss (top) and test error (bottom) for Experiments (A) and (B). The dotted green line is the originally trained model, used as a reference. We perform the experiments for only one trained model θ', and thus there is no confidence interval.

be found on Github[‡]. The original models are trained until 0 training error is reached. For Experiment (A), training is done for at least 100 epochs, and then it is stopped when 0 error or a maximum of 400 epochs is reached. Moreover, sub-networks ξ_l with larger depth (i.e. starting at a lower l) required more care and a lower learning rate than shallow ones. For VGG, we employ a learning rate scheduler. As reported in the table, we tweak the scheduler for certain values of l (giving the deepest models) in the case $p = \frac{1}{4}$. For the varying-width experiment (Figure 6.4), the same learning rate is used for different widths to allow a comparison of results. The maximal learning rate that makes every model converge is kept.

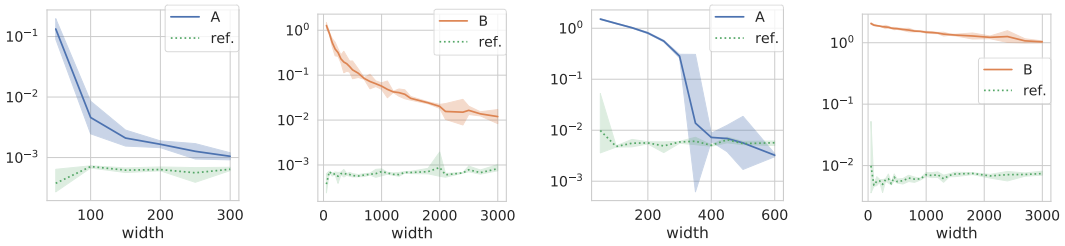
Pruning for VGG-11. As in previous work [KWL+19], for CNN architectures like VGG-11, each convolutional feature map associated with a filter/kernel can be treated as a single neuron of a fully connected architecture. In this way, Experiment (B) can be done similarly to the case of fully connected networks. For Experiment (A), this means that the sub-networks ξ_l's consist of a small fraction of the convolutional filters from each layer of the original network. All the max-pooling layers, as part of the original VGG-11 architecture, are kept unchanged.

[‡] https://github.com/modeconnectivity/modeconnectivity



(a) FCN MNIST (b) FCN CIFAR-10 (c) VGG CIFAR-10

Figure 6.8: Dropout ratio $p = \frac{1}{4}$. Training loss (top) and test error (bottom) for Experiments (A) and (B). The dotted green line is the originally trained model, used as a reference. We perform the experiments for only one trained model θ', and thus there is no confidence interval.



(a) MNIST (b) CIFAR-10

Figure 6.9: Results of Figure 6.4 plotted with a logarithmic y-axis. Mean and 95% confidence intervals are depicted for three different trained networks θ'.

Model	(O-11)	(O-3)	(A-11)	(A-3)
learning rate	10^{-3}	$5 \cdot 10^{-3}$	$2 \cdot 10^{-3}$– $5 \cdot 10^{-3}$	$5 \cdot 10^{-3}$
momentum	0.95	—	—	—

(a) FCN MNIST

Model	(O-6)	(O-3)	(A-6)	(A-3), $l = 0$	(A-3), $l = 1, 2$
learning rate	10^{-3}	$5 \cdot 10^{-4}$	10^{-3} – $5 \cdot 10^{-3}$	10^{-3}	$5 \cdot 10^{-3}$
momentum	0.95	—	—	—	—

(b) FCN CIFAR-10

Model	(O)	(A)†	(A), $p = \frac{1}{4}, l = 0$	(A), $p = \frac{1}{4}, l = 1$
learning rate	$5 \cdot 10^{-3}$	—	—	—
momentum	0.9	—	—	—
weight decay	$5 \cdot 10^{-4}$	—	—	—
scheduler	$\gamma = 0.5$ steps = 10	$\gamma = 0.5$ steps = 20	$\gamma = 0.1$ steps = 250	$\gamma = 0.9$ mode = "plateau"

† All of the cases except $p = \frac{1}{4}, l \in \{0, 1\}$

(c) VGG CIFAR-10

Table 6.1: Hyperparameters of the SGD optimizer for training the original model (O) and the sub-networks of Experiment (A). For each $n \in \{3, 6, 11\}$, the models (O-n) and (A-n) refer to those sub-networks that have total number of layers n (i.e. $n \in \{6, 11\}$ in Figure 6.3, and $n = 3$ in Figure 6.4). The index l refers to the subnetwork ξ_l. Dashes stand for duplicate values from the column on the left.

Conclusion

In this thesis, we have investigated different aspects of generative modelling, specifically Generative Adversarial Networks (GANs) which are based on Optimal Transport (OT) quantities.

The core mathematical framework underlying the expression of OT was first presented in Chapter 2. Then, in Chapter 3, we presented different formulations of a general GAN loss, along with typical approaches and hypothesis used to analyse the resulting optimization system. Focusing on the optimal discriminator hypothesis, we recalled sufficient regularity assumptions a GAN should satisfy in order for the overall optimization problem to converge to a critical point of the loss. We illustrated them with toy examples, and showed that, even in this simple case, at least one of the condition was not satisfied. Considering more intricate examples should be the topic of future work.

Broadening the field of study, in Chapter 4, we introduced the Bures-Wasserstein (BW) linear GAN. It is one of the few instances where a closed-form expression of the Wasserstein distance is known and can be leveraged. Moreover, the deep linear architecture of the generator encodes a deep matrix factorization of a covariance matrix. Together, the resulting learning system possesses attractive characteristics: it unites a rank-constrained matrix approximation problem, a non-Euclidean type of loss, and an over-parametrized optimization scheme. The expression of the critical point for the BW loss restricted to a fixed-rand set of matrices was derived. (Theorems 4.5 and 4.9). Furthermore, their correspondence in the different spaces (covariance, function, and parameter spaces) was drawn (Propositions 4.11 and 4.12). Interestingly, the resulting critical points of the BW loss are identical to the critical points found when considering the Frobenius loss.

Even if the Frobenius and BW losses share the expression of the critical points, the Hessian at these critical points are different. As it influences the optimization procedure, we derived the Hessian matrices expression (under the parametrization considered) and experimentally investigated their condition numbers. Although the findings are preliminary, they might pave the way to better understanding the fundamental differences between the two losses.

As the problem studied can be cast as an instance of a rank-constrained (covariance) matrix approximation, one may legitimately ask what its connection to the celebrated Eckart-Young-Mirsky theorem is. Nonetheless, an application of this theorem is not possible in this case, as the BW distance is not a norm on covariance matrices. We investigated why the minimizers of the approximation problem were the same for the two losses, and pinpointed the origin of this equivalence to the ground cost used in the

sample space on which the Wasserstein distance is based. A generalization of the results in the case another ground distance is considered to compute the BW loss is left to future endeavors.

The use of this specific loss raises challenges in the analytical study of the resulting optimization scheme. We have tackled some of them, with the help of appropriate hypothesis on the initialization of the weights of the network. In particular, we showed exponential convergence to a global minimizer of the original loss under gradient flow on the one hand (Theorem 5.18), and convergence for the regularized loss to a minimizer that achieves minimum loss over a set of rank-constrained matrices under gradient descent on the other hand (Theorem 5.22). Experimental evaluations corroborate the theoretical statement, and indeed show a convergence acceleration correlated with the depth of the network considered.

Again, some geometrical aspects of the landscape are shared with the squared norm loss. For instance, the strict saddle point property is shared between the two losses, so that gradient descent algorithms, when generically initialized, are almost sure to converge to local minimizers. What's more, the local minimizer turns out to be a global if we restrict the search-space to a set of rank-constrained matrices. As a future line of work, relaxing or even discarding the optimal discriminator hypothesis could be envisaged. This would, for instance, imply considering a second network in order to model a discriminator. More generally, relaxing the assumptions (balancedness of the weights, margin deficiency) of the study are naturally possible extensions to the present work.

Moreover, chain of parametrizations (deep linear network and covariance) produces an infinite number of critical points in the parameter and function spaces that all correspond to a single critical point in the covariance space. Even when a bijection between the parameter space and the function space is obtained, thanks to the balancedness property of the parameters, there is still an infinity of functions (matrices) that map to the same covariance space. Therefore, the convergence of the gradient-descent algorithms is only ensured to find the unique covariance matrix solution, but can do so with any corresponding end-to-end matrix. Understanding if one is favored over another is a question that we leave to future research.

This question is related to the last chapter of the thesis, Chapter 6, where experimental evidence suggests that gradient descent algorithms naturally tend to converge toward basins of low-error solutions. The work is framed in a general classification setting, and an manifest direction of research is to extend it to a generative setting. Notice nonetheless that the study carried out in Chapters 4 and 5 essentially reduces the GAN optimization problem to a minimization-only problem, and thus, falls under the framework of Chapter 6. The detailed connection between the infinitely-many solutions to the over-parametrized matrix approximation and the basins of good solutions would also probably be a fruitful path to pursue.

Index

2-Wasserstein GAN, 52

adjoint operator, 14

Banach space, 10
Benamou-Brenier formulation, 24
Borel set, 14
Burer-Monteiro parametrization, 54
Bures-Wasserstein distance, 24, 55

c-transform, 21
characteristic function, 17
continuous piecewise linear path, 134
convex conjugate, 9
convex set, 8
coupling, 20

δ-balancedness, 95

effective domain, 8

Fréchet distance, 24
function
    convex, 8
    lower semicontinuous (l.s.c.), 9
    proper, 8

inner problem, 30

Legendre transform, 9
linear network
    balanced weights, 94
    function space, 54
    parameter space, 53
    parametrization, 53
linear networks, 53
linear Wasserstein GAN, 54

modified-deficiency margin, 100

Nash equilibrium, 33

optimal discriminator hypothesis, 39
outer problem, 30

p-transform, 22
p-Wasserstein distance, 23
perturbative Bures-Wasserstein loss, 58
Polish space, 19
population loss, 133
push-forward, 14

spectral norm, 95
subdifferential, 10

topological dual, 11
total variation, 12

vector space
    topologically paired, 9

weak duality, 16

Bibliography

[AB17] M. Arjovsky and L. Bottou. "Towards Principled Methods for Training Generative Adversarial Networks". In: *Proceedings of the International Conference on Learning Representations (ICLR)*. Jan. 17, 2017.

[ABM14] H. Attouch, G. Buttazzo, and G. Michaille. *Variational Analysis in Sobolev and BV Spaces: Applications to PDEs and Optimization*. Second edition. MOS-SIAM Series on Optimization 17. Philadelphia: Society for Industrial and Applied Mathematics : Mathematical Optimization Society, 2014. 793 pp.

[ACB17] M. Arjovsky, S. Chintala, and L. Bottou. "Wasserstein Generative Adversarial Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. PMLR, 2017, pp. 214–223.

[ACG+19] S. Arora, N. Cohen, N. Golowich, and W. Hu. "A Convergence Analysis of Gradient Descent for Deep Linear Neural Networks". In: *International Conference on Learning Representations*. 2019.

[ACH+19] S. Arora, N. Cohen, W. Hu, and Y. Luo. "Implicit Regularization in Deep Matrix Factorization". In: *Advances in Neural Information Processing Systems*. Vol. 32. 2019.

[ACH18] S. Arora, N. Cohen, and E. Hazan. "On the Optimization of Deep Networks: Implicit Acceleration by Overparameterization". In: *Proceedings of the 35th International Conference on Machine Learning*. 2018, pp. 244–253.

[AGS05] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics ETH Zürich. Boston: Birkhäuser, 2005. 333 pp.

[ALS19] Z. Allen-Zhu, Y. Li, and Z. Song. "A Convergence Theory for Deep Learning via Over-Parameterization". In: *International Conference on Machine Learning (ICML)*. 2019, pp. 242–252.

[AMA05] P. A. Absil, R. Mahony, and B. Andrews. "Convergence of the Iterates of Descent Methods for Analytic Cost Functions". In: *SIAM Journal on Optimization* 16.2 (2005), pp. 531–547.

[AY20] I. Anokhin and D. Yarotsky. "Low-Loss Connection of Weight Vectors: Distribution-Based Approaches". In: *Proceedings of the 37th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, Nov. 21, 2020, pp. 335–344.

[Bau88] E. B. Baum. "On the Capabilities of Multilayer Perceptrons". In: *Journal of Complexity* 4 (1988).

[BEL+20] S. Bubeck, R. Eldan, Y. T. Lee, and D. Mikulincer. "Network Size and Weights Size for Memorization with Two-Layers Neural Networks". Nov. 3, 2020.

[BH89] P. Baldi and K. Hornik. "Neural Networks and Principal Component Analysis: Learning from Examples without Local Minima". In: *Neural Networks* 2.1 (1989), pp. 53–58.

[BHL+19] P. Bartlett, N. Harvey, C. Liaw, and A. Mehrabian. "Nearly-Tight VC-dimension and Pseudodimension Bounds for Piecewise Linear Neural Networks." In: *Journal of Machine Learning Research (JMLR)* 20.63 (2019), pp. 1–17.

[BJL19] R. Bhatia, T. Jain, and Y. Lim. "On the Bures–Wasserstein Distance between Positive Definite Matrices". In: *Expositiones Mathematicae* 37.2 (June 2019), pp. 165–191.

[BM05] S. Burer and R. D. Monteiro. "Local Minima and Convergence in Low-Rank Semidefinite Programming". In: *Mathematical Programming* 103.3 (2005), pp. 427–444.

[Bou23] N. Boumal. *An Introduction to Optimization on Smooth Manifolds*. 1st ed. Cambridge University Press, Mar. 16, 2023.

[Bou87] N. Bourbaki. *Topological Vector Spaces*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1987.

[BPA+23] P. Bréchet, K. Papagiannouli, J. An, and G. Montufar. "Critical Points and Convergence Analysis of Generative Deep Linear Networks Trained with Bures-Wasserstein Loss". In: *Proceedings of the 40th International Conference on Machine Learning*. Apr. 24, 2023.

[Bre11] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. New York, NY: Springer New York, 2011.

[BRT+21] B. Bah, H. Rauhut, U. Terstiege, and M. Westdickenberg. "Learning Deep Linear Neural Networks: Riemannian Gradient Flows and Convergence to Global Minimizers". In: *Information and Inference: A Journal of the IMA* 11.1 (2021), pp. 307–353.

[BSI+19] J. Brea, B. Simsek, B. Illing, and W. Gerstner. *Weight-Space Symmetry in Deep Networks Gives Rise to Permutation Saddles, Connected by Equal-Loss Valleys across the Loss Landscape*. 2019. Pre-published.

[Bur69] D. Bures. "An Extension of Kakutani's Theorem on Infinite Product Measures to the Tensor Product of Semifinite w*-Algebras". In: *Transactions of the American Mathematical Society* 135 (1969), pp. 199–212.

[CBG19] C. Chu, J. Blanchet, and P. Glynn. "Probability Functional Descent: A Unifying Perspective on GANs, Variational Inference, and Reinforcement Learning". In: International Conference on Learning Representations, ICLR. 2019.

[CHM+15] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun. "The Loss Surfaces of Multilayer Networks". In: *International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2015, pp. 192–204.

[ÇJM+19] T. Ö. Çelik, A. Jamneshan, G. Montúfar, B. Sturmfels, and L. Venturello. "Optimal Transport to a Variety". Sept. 25, 2019.

[ÇJM+20] T. Ö. Çelik, A. Jamneshan, G. Montúfar, B. Sturmfels, and L. Venturello. "Wasserstein Distance to Independence Models". Mar. 14, 2020.

[CLC22] Y. Chitour, Z. Liao, and R. Couillet. "A Geometric Approach of Gradient Descent Algorithms in Linear Neural Networks". In: *Mathematical Control and Related Fields* (2022).

[CMF20] C. Chu, K. Minami, and K. Fukumizu. "Smoothness and Stability in GANs." In: *8th International Conference on Learning Representations*. International Conference on Learning Representations. 2020.

[CMR+20] S. Chewi, T. Maunu, P. Rigollet, and A. J. Stromme. "Gradient Descent Algorithms for Bures-Wasserstein Barycenters". In: (Jan. 6, 2020).

[Cov65] T. M. Cover. "Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition". In: *IEEE Transactions on Electronic Computers* 3 (1965), pp. 326–334.

[CZX+20] Z. Chen, Y. Zhou, T. Xu, and Y. Liang. "Proximal Gradient Descent-Ascent: Variable Convergence under KŁ Geometry". In: International Conference on Learning Representations. Sept. 28, 2020.

[DHL18] S. S. Du, W. Hu, and J. D. Lee. "Algorithmic Regularization in Learning Deep Homogeneous Models: Layers Are Automatically Balanced". In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc., 2018.

[DIS+18] C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng. "Training GANs with Optimism". In: International Conference on Learning Representations. Feb. 15, 2018.

[DKZ09] I. L. Dryden, A. Koloydenko, and D. Zhou. "Non-Euclidean Statistics for Covariance Matrices with Applications to Diffusion Tensor Imaging". In: *The Annals of Applied Statistics* 3.3 (2009), pp. 1102–1123.

[DL17] A. Dutta and X. Li. "On a Problem of Weighted Low-Rank Approximation of Matrices". In: *SIAM Journal on Matrix Analysis and Applications* 38.2 (2017), pp. 530–553.

[DL82] D. C. Dowson and B. V. Landau. "The Fréchet Distance between Multivariate Normal Distributions". In: *Journal of Multivariate Analysis* 12.3 (1982), pp. 450–455.

[DLL+19a] S. S. Du, J. D. Lee, H. Li, L. Wang, and X. Zhai. "Gradient Descent Finds Global Minima of Deep Neural Networks". May 28, 2019.

[DLL+19b] Y. Dukler, W. Li, A. Lin, and G. Montufar. "Wasserstein of Wasserstein Loss for Learning Generative Models". In: *Proceedings of the 36th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, May 24, 2019, pp. 1716–1725.

[DSF+21] H. De Meulemeester, J. Schreurs, M. Fanuel, B. De Moor, and J. A. K. Suykens. "The Bures Metric for Generative Adversarial Networks". In: *Machine Learning and Knowledge Discovery in Databases. Research Track*. Ed. by N. Oliver, F. Pérez-Cruz, S. Kramer, J. Read, and J. A. Lozano. Vol. 12976. Cham: Springer International Publishing, 2021, pp. 52–66.

[DVS+18] F. Draxler, K. Veschgini, M. Salmhofer, and F. Hamprecht. "Essentially No Barriers in Neural Network Energy Landscape". In: *Proceedings of the 35th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, July 3, 2018, pp. 1309–1318.

[EY36] C. Eckart and G. Young. "The Approximation of One Matrix by Another of Lower Rank". In: *Psychometrika* 1.3 (1936), pp. 211–218.

[FA00] K. Fukumizu and S.-i. Amari. "Local Minima and Plateaus in Hierarchical Structures of Multilayer Perceptrons". In: *Neural Networks* (2000), pp. 317–327.

[FB17] C. D. Freeman and J. Bruna. "Topology and Geometry of Half-Rectified Network Optimization". June 1, 2017.

[FFG+20] S. Feizi, F. Farnia, T. Ginart, and D. Tse. "Understanding GANs in the LQG Setting: Formulation, Generalization and Stability". In: *IEEE Journal on Selected Areas in Information Theory* 1.1 (2020), pp. 304–311.

[FG15] N. Fournier and A. Guillin. "On the Rate of Convergence in Wasserstein Distance of the Empirical Measure". In: *Probability Theory and Related Fields* 162.3-4 (Aug. 2015), pp. 707–738.

[FJ19] S. Fort and S. Jastrzebski. "Large Scale Structure of Neural Network Loss Landscapes". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 602. Red Hook, NY, USA: Curran Associates Inc., Dec. 8, 2019, pp. 6709–6717.

[FRH+19] M. Fazlyab, A. Robey, H. Hassani, M. Morari, and G. Pappas. "Efficient and Accurate Estimation of Lipschitz Constants for Deep Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019.

[Fuk98] K. Fukumizu. "Effect of Batch Learning in Multilayer Neural Networks". In: *In Proceedings of ICONIP'98*. 1998.

[Fun89] K.-I. Funahashi. "On the Approximate Realization of Continuous Mappings by Neural Networks". In: *Neural Networks* 2.3 (Jan. 1989), pp. 183–192.

[FYM+19] K. Fukumizu, S. Yamaguchi, Y.-i. Mototake, and M. Tanaka. "Semi-Flat Minima and Saddle Points by Embedding Neural Networks to Overparameterization". In: *Neural Information Processing Systems (NeurIPS)*. 2019.

[FZF+20] K. Fatras, Y. Zine, R. Flamary, R. Gribonval, and N. Courty. "Learning with Minibatch Wasserstein : Asymptotic and Gradient Properties". Mar. 9, 2020.

[GAA+17] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. "Improved Training of Wasserstein GANs". Mar. 31, 2017.

[GGS+20] M. Goldblum, J. Geiping, A. Schwarzschild, M. Moeller, and T. Goldstein. "Truth or Backpropaganda? An Empirical Investigation of Deep Learning Theory". In: *International Conference on Learning Representations (ICLR)*. 2020.

[GIP+18] T. Garipov, P. Izmailov, D. Podoprikhin, D. Vetrov, and A. G. Wilson. "Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs". Oct. 30, 2018.

[GPM+14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems*. Vol. 27. 2014.

[GS19] N. Gillis and Y. Shitov. "Low-Rank Matrix Approximation in the Infinity Norm". In: *Linear Algebra and its Applications* 581 (2019), pp. 367–382.

[GV18] N. Gillis and S. A. Vavasis. "On the Complexity of Robust PCA and 1-Norm Low-Rank Matrix Approximation". In: *Mathematics of Operations Research* 43.4 (2018), pp. 1072–1084.

[GWZ19] R. Ge, R. Wang, and H. Zhao. *Mildly Overparametrized Neural Nets Can Memorize Training Data Efficiently*. 2019. Pre-published.

[HK22] J. Ha and G. Kim. "On Convergence of Lookahead in Smooth Games". In: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. International Conference on Artificial Intelligence and Statistics. PMLR, May 3, 2022, pp. 4659–4684.

[HL01] J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001.

[HM17] M. Hardt and T. Ma. "Identity Matters in Deep Learning". In: *International Conference on Learning Representations*. 2017.

[HRU+17] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium". June 26, 2017.

[HS95] U. Helmke and M. A. Shayman. "Critical Points of Matrix Least Squares Distance Functions". In: *Linear Algebra and its Applications* 215 (Jan. 1995), pp. 1–19.

[HV17] B. Haeffele and R. Vidal. "Global Optimality in Neural Network Training". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

[HWT20] F. He, B. Wang, and D. Tao. "Piecewise Linear Activations Substantially Shape the Loss Surfaces of Neural Networks". In: *International Conference on Learning Representations (ICLR)*. 2020.

[Kaw16] K. Kawaguchi. "Deep Learning without Poor Local Minima". In: (2016), p. 24.

[KB19] K. Kawaguchi and Y. Bengio. "Depth with Nonlinearity Creates No Bad Local Minima in ResNets". In: *Neural Networks* (2019), pp. 167–174.

[KKM+21] D. Kwon, Y. Kim, G. Montúfar, and I. Yang. *Training Wasserstein GANs without Gradient Penalties*. Oct. 27, 2021. Pre-published.

[KLA18] T. Karras, S. Laine, and T. Aila. "A Style-Based Generator Architecture for Generative Adversarial Networks". In: Computer Vision and Pattern Recognition (CVPR). Dec. 12, 2018.

[KLZ+19] K. Kurach, M. Lucic, X. Zhai, M. Michalski, and S. Gelly. "A Large-Scale Study on Regularization and Normalization in GANs". May 14, 2019.

[KMM+22] K. Kohn, T. Merkh, G. Montúfar, and M. Trager. "Geometry of Linear Convolutional Networks". In: *SIAM Journal on Applied Algebra and Geometry* 6.3 (2022), pp. 368–406.

[KMS+23] K. Kohn, G. Montúfar, V. Shahverdi, and M. Trager. *Function Space and Critical Points of Linear Convolutional Networks*. 2023. Pre-published.

[Kri09] A. Krizhevsky. *Learning Multiple Layers of Features from Tiny Images*. University of Toronto, 2009, p. 60.

[KSS21] A. Kroshnin, V. Spokoiny, and A. Suvorikova. "Statistical Inference for Bures–Wasserstein Barycenters". In: *The Annals of Applied Probability* 31.3 (2021).

[KWL+19] R. Kuditipudi, X. Wang, H. Lee, Y. Zhang, Z. Li, W. Hu, R. Ge, and S. Arora. "Explaining Landscape Connectivity of Low-Cost Solutions for Multilayer Nets". In: *Neural Information Processing Systems (NeurIPS)*. 2019, pp. 14601–14610.

[KZ05] A. Kurdila and M. Zabarankin. *Convex Functional Analysis*. Systems & Control. Basel ; Boston: Birkhauser Verlag, 2005. 228 pp.

[LB18] T. Laurent and J. Brecht. "Deep Linear Networks with Arbitrary Loss: All Local Minima Are Global". In: *International Conference on Machine Learning (ICML)*. 2018, pp. 2902–2907.

[LBB+01] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-Based Learning Applied to Document Recognition". In: *Intelligent Signal Processing*. Ed. by S. Haykin and B. Kosko. IEEE Press, 2001, pp. 306–351.

[LDS18] D. Li, T. Ding, and R. Sun. *Over-Parameterized Deep Neural Networks Have No Strict Local Minima for Any Continuous Activations*. 2018. preprint.

[LK17] H. Lu and K. Kawaguchi. *Depth Creates No Bad Local Minima*. 2017. Pre-published.

[LKB22] E. Levin, J. Kileel, and N. Boumal. *The Effect of Smooth Parametrizations on Nonconvex Optimization Landscapes*. 2022. Pre-published.

[LLZ21] Y. Luo, X. Li, and A. R. Zhang. *On Geometric Connections of Embedded and Quotient Geometries in Riemannian Fixed-rank Matrix Optimization*. Oct. 22, 2021. Pre-published.

[LPP+19] J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan, and B. Recht. "First-Order Methods Almost Always Avoid Strict Saddle Points". In: *Mathematical Programming* 176.1-2 (July 2019), pp. 311–337.

[LRC19] F. Latorre, P. Rolland, and V. Cevher. "Lipschitz Constant Estimation of Neural Networks via Sparse Polynomial Optimization". In: International Conference on Learning Representations. Sept. 25, 2019.

[LSL+18a] S. Liang, R. Sun, J. D. Lee, and R. Srikant. "Adding One Neuron Can Eliminate All Bad Local Minima". In: *Neural Information Processing Systems (NeurIPS)*. 2018, pp. 4350–4360.

[LSL+18b] S. Liang, R. Sun, Y. Li, and R. Srikant. "Understanding the Loss Surface of Neural Networks for Binary Classification". In: *International Conference on Machine Learning (ICML)*. 2018, pp. 2840–2849.

[LZB22] C. Liu, L. Zhu, and M. Belkin. "Loss Landscapes and Optimization in Over-Parameterized Non-Linear Systems and Neural networks". In: *Applied and Computational Harmonic Analysis* 59 (2022), pp. 85–116.

[MC18] B. Muzellec and M. Cuturi. "Generalizing Point Embeddings Using the Wasserstein Space of Elliptical Distributions". In: *Advances in Neural Information Processing Systems*. 2018.

[MGM21] A. Mallasto, A. Gerolin, and H. Q. Minh. "Entropy-Regularized 2-Wasserstein Distance between Gaussian Measures". In: *Information Geometry* (Aug. 16, 2021).

[MGN18] L. Mescheder, A. Geiger, and S. Nowozin. "Which Training Methods for GANs Do Actually Converge?" 2018.

[Mir60] L. Mirsky. "Symmetric Gauge Functions and Unitary Invariant Norms". In: *The Quarterly Journal of Mathematics* 11.1 (1960), pp. 50–59.

[MKK+18] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. "Spectral Normalization for Generative Adversarial Networks". In: *ICLR 2018 : International Conference on Learning Representations 2018*. 2018.

[MMG19] A. Mallasto, G. Montúfar, and A. Gerolin. *How Well Do WGANs Estimate the Wasserstein Metric?* Oct. 9, 2019. Pre-published.

[MMN18] S. Mei, A. Montanari, and P. Nguyen. "A Mean Field View of the Landscape of Two-Layer Neural Networks". In: *Proceedings of the National Academy of Sciences* 115.33 (2018), E7665–E7671.

[MN19] J. R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Third edition. Wiley Series in Probability and Statistics. Wiley, 2019.

[MNG17] L. Mescheder, S. Nowozin, and A. Geiger. "The Numerics of GANs". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017.

[MO14] M. Mirza and S. Osindero. "Conditional Generative Adversarial Nets." In: *CoRR* abs/1411.1784 (2014).

[Mon81] G. Monge. "Mémoire Sur La Théorie Des Déblais et de Remblais". In: *Histoire de l'Académie Royale Des Sciences de Paris, Avec Les Mémoires de Mathématique et de Physique Pour La Même Année*. 1781.

[MPZ19] V. Masarotto, V. Panaretos, and Y. Zemel. "Procrustes Metrics on Covariance Operators and Optimal Transportation of Gaussian Processes". In: *Sankhya A. The Indian Journal of Statistics* 81.1 (2019), pp. 172–213.

[MTV+21] H. Min, S. Tarmoun, R. Vidal, and E. Mallada. "On the Explicit Role of Initialization on the Convergence and Implicit Bias of Overparametrized Linear Networks". In: *Proceedings of the 38th International Conference on Machine Learning*. 2021, pp. 7760–7768.

[MZ20] A. Montanari and Y. Zhong. *The Interpolation Phase Transition in Neural Networks: Memorization and Generalization under Lazy Training*. 2020. Pre-published.

[NBM21] Q. N. Nguyen, P. Bréchet, and M. Mondelli. "When Are Solutions Connected in Deep Networks?" In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., 2021, pp. 20956–20969.

[NCT16] S. Nowozin, B. Cseke, and R. Tomioka. "F-GAN: Training Generative Neural Samplers Using Variational Divergence Minimization." In: *Advances in Neural Information Processing Systems*. 2016, pp. 271–279.

[Ngu19] Q. Nguyen. "On Connected Sublevel Sets in Deep Learning". In: *Proceedings of the 36th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, May 24, 2019, pp. 4790–4799.

[Ngu21] Q. Nguyen. *A Note on Connectivity of Sublevel Sets in Deep Learning*. Jan. 21, 2021. Pre-published.

[NH17] Q. Nguyen and M. Hein. "The Loss Surface of Deep and Wide Neural Networks". In: *International Conference on Machine Learning (ICML)*. 2017, pp. 2603–2612.

[NK17] V. Nagarajan and J. Z. Kolter. "Gradient Descent GAN Optimization Is Locally Stable". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017.

[NM20] Q. Nguyen and M. Mondelli. "Global Convergence of Deep Networks with One Wide Layer Followed by Pyramidal Topology". In: *Neural Information Processing Systems (NeurIPS)*. 2020.

[NRT21] G. M. Nguegnang, H. Rauhut, and U. Terstiege. *Convergence of Gradient Descent for Learning Linear Neural Networks*. 2021. Pre-published.

[NYK+21] A. No, T. Yoon, S. Kwon, and E. K. Ryu. "WGAN with an Infinitely Wide Generator Has No Spurious Stationary Points". June 9, 2021.

[OS20] S. Oymak and M. Soltanolkotabi. "Towards Moderate Overparameterization: Global Convergence Guarantees for Training Shallow Neural Networks". In: *IEEE Journal on Selected Areas in Information Theory* (2020).

[PC18] G. Peyré and M. Cuturi. *Computational Optimal Transport*. Mar. 1, 2018.

[QMW20] Y. Qin, N. Mitra, and P. Wonka. "How Does Lipschitz Regularization Influence GAN Training?" In: *Computer Vision – ECCV 2020*. Ed. by A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm. Cham: Springer International Publishing, 2020, pp. 310–326.

[RLN+17] K. Roth, A. Lucchi, S. Nowozin, and T. Hofmann. "Stabilizing Training of Generative Adversarial Networks through Regularization". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017.

[Roc67] R. T. Rockafellar. "Duality and Stability in Extremum Problems". In: 21.1 (1967).

[Roc70] R. T. Rockafellar. "Convex Analysis". In: (1970).

[RST18] P. K. Rubenstein, B. Schoelkopf, and I. Tolstikhin. *Learning Disentangled Representations with Wasserstein Auto-Encoders*. Feb. 12, 2018. Pre-published.

[Rud13] W. Rudin. *Real and Complex Analysis*. 3. ed., internat. ed., [Nachdr.] McGraw-Hill International Editions Mathematics Series. New York, NY: McGraw-Hill, 2013. 416 pp.

[RW98] R. T. Rockafellar and R. J. B. Wets. *Variational Analysis*. Red. by M. Berger et al. 3rd ed. Vol. 317. Grundlehren Der Mathematischen Wissenschaften. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998.

[RZ79] G. Ruben and S. Zamir. "Lower Rank Approximation of Matrices by Least Squares with Any Choice of Weights". In: *Technometrics : a journal of statistics for the physical, chemical, and engineering sciences* 21.4 (1979), pp. 489–498.

[San15] F. Santambrogio. *Optimal Transport for Applied Mathematicians*. 2015. 1-376.

[Sch92] B. A. Schmitt. "Perturbation Bounds for Matrix Square Roots and Pythagorean Sums". In: *Linear Algebra and its Applications* 174 (Sept. 1, 1992), pp. 215–227.

[SEK+21] J. Stanczuk, C. Etmann, L. M. Kreusser, and C.-B. Schönlieb. "Wasserstein GANs Work Because They Fail (to Approximate the Wasserstein Distance)". Mar. 5, 2021.

[SH18] D. Soudry and E. Hoffer. *Exponentially Vanishing Sub-Optimal Local Minima in Multilayer Neural Networks*. 2018. Pre-published.

[Sim95] S. Simons. "Minimax Theorems and Their Proofs". In: *Minimax and Applications*. Ed. by D.-Z. Du and P. M. Pardalos. Red. by P. Pardalos and R. Horst. Vol. 4. Boston, MA: Springer US, 1995, pp. 1–23.

[SM20] A. Shevchenko and M. Mondelli. "Landscape Connectivity and Dropout Stability of SGD Solutions for Over-parameterized Neural Networks". July 23, 2020.

[SMG14] A. M. Saxe, J. L. McClelland, and S. Ganguli. "Exact Solutions to the Nonlinear Dynamics of Learning in Deep Linear Neural Networks". In: *ICLR*. 2014.

[SS16] I. Safran and O. Shamir. "On the Quality of the Initial Basin in Overspecified Neural Networks". In: *International Conference on Machine Learning (ICML)*. 2016, pp. 774–782.

[SS18] I. Safran and O. Shamir. "Spurious Local Minima Are Common in Two-Layer ReLU Neural Networks". In: *International Conference on Machine Learning (ICML)*. 2018.

[SSG19] A. Sharifnassab, S. Salehkaleybar, and S. J. Golestani. "Bounds on Over-Parameterization for Guaranteed Existence of Descent Paths in Shallow ReLU Networks". In: International Conference on Learning Representations. Sept. 25, 2019.

[SV18] K. Scaman and A. Virmaux. "Lipschitz Regularity of Deep Neural Networks: Analysis and Efficient Estimation". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS'18. Red Hook, NY, USA: Curran Associates Inc., Dec. 3, 2018, pp. 3839–3848.

[SWZ17] Z. Song, D. P. Woodruff, and P. Zhong. "Low Rank Approximation with Entrywise L1-Norm Error". In: *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. STOC 2017. New York, NY, USA: Association for Computing Machinery, 2017, pp. 688–701.

[SZ15] K. Simonyan and A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *International Conference on Learning Representations (ICLR)*. 2015.

[SZR+18] T. Salimans, H. Zhang, A. Radford, and D. Metaxas. "Improving GANs Using Optimal Transport". Mar. 14, 2018.

[TFH+21] S. Tarmoun, G. Franca, B. D. Haeffele, and R. Vidal. "Understanding the Dynamics of Gradient Flow in Overparameterized Linear Models". In: *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 2021, pp. 10153–10161.

[TKB20] M. Trager, K. Kohn, and J. Bruna. "Pure and Spurious Critical Points: A Geometric Study of Linear Networks". In: *International Conference on Learning Representations*. 2020.

[VBB19] L. Venturi, A. S. Bandeira, and J. Bruna. "Spurious Valleys in One-hidden-layer Neural Network Optimization Landscapes". In: *Journal of Machine Learning Research* 20.133 (2019), pp. 1–34.

[Ver20] R. Vershynin. "Memory Capacity of Neural Networks with Threshold and Rectified Linear Unit Activations". In: *SIAM Journal on Mathematics of Data Science* 2.4 (2020), pp. 1004–1033.

[Vil03] C. Villani. *Topics in Optimal Transportation*. Graduate Studies in Mathematics. American Mathematical Society, 2003.

[Vil09] C. Villani. *Optimal Transport: Old and New*. Grundlehren Der Mathematischen Wissenschaften 338. Berlin: Springer, 2009. 973 pp.

[WLL18] C. Wu, J. Luo, and J. D. Lee. "No Spurious Local Minima in a Two Hidden Unit ReLU Network". In: *International Conference on Learning Representations (ICLR) Workshop*. 2018.

[XLS17] B. Xie, Y. Liang, and L. Song. "Diverse Neural Network Learns True Target Functions". In: *International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2017, pp. 1216–1224.

[YKH20] J. Yang, N. Kiyavash, and N. He. "Global Convergence and Variance Reduction for a Class of Nonconvex-Nonconcave Minimax Problems". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 1153–1165.

[YSJ18] C. Yun, S. Sra, and A. Jadbabaie. "Global Optimality Conditions for Deep Neural Networks". In: *International Conference on Learning Representations (ICLR)*. 2018.

[YSJ19] C. Yun, S. Sra, and A. Jadbabaie. "Small ReLU Networks Are Powerful Memorizers: A Tight Analysis of Memorization Capacity". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019.

[Zal02] C. Zalinescu. *Convex Analysis in General Vector Spaces*. River Edge, N.J. ; London: World Scientific, 2002. 367 pp.

[ZCZ+20] D. Zou, Y. Cao, D. Zhou, and Q. Gu. "Gradient Descent Optimizes Over-Parameterized Deep ReLU Networks". In: *Machine Learning* 109.3 (2020), pp. 467–492.

[ZL18] Y. Zhou and Y. Liang. "Critical Points of Linear Neural Networks: Analytical Forms and Landscape Properties". In: *International Conference on Learning Representations (ICLR)*. 2018.

[ZLB+19] M. Zhang, J. Lucas, J. Ba, and G. E. Hinton. "Lookahead Optimizer: k Steps Forward, 1 Step Back". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019.

[ZSS19] H. Zhang, J. Shao, and R. Salakhutdinov. "Deep Neural Networks with Multi-Branch Architectures AreIntrinsically Less Non-Convex". In: *International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2019.

Bibliographische Daten

Titel	On the Optimization Properties of Optimal-Transport Based Generative Adversarial Linear Networks
Autor	Bréchet, Pierre
Verlag	Universität Leipzig, Dissertation, 2024
Referenzen	151

Selbständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialen oder erbrachten Dienstleistungen als solche gekennzeichnet.

Leipzig, den 11. November 2024

_____

Pierre Vincent Serge Bréchet

Daten zum Autor

Vornamen, Name	Pierre Vincent Serge, Bréchet
Geburtsdatum	01. März 1994
Geburtsort	Cincinnati, Ohio, USA
09.2013 – 07.2019	Dipl. Ing. Télécom ParisTech und EURECOM
10.2015 – 07.2019	M. Sc. Informatik, Technische Universität München
Seit 09.2020	Doktorand der Informatik, MPI MiS