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

On the Effect of Initialization: The Scaling Path of 2-Layer Neural Networks

Sebastian Neumayer* SEBASTIAN.NEUMAYER@EPFL.CH
Biomedical Imaging Group
École polytechnique fédérale de Lausanne
Lausane, CH-1015, Switzerland

Lénaïc Chizat LENAIC.CHIZAT@EPFL.CH
Chair of Dynamics of Learning Algorithms
École polytechnique fédérale de Lausanne
Lausane, CH-1015, Switzerland

Michael Unser MICHAEL.UNSER@EPFL.CH
Biomedical Imaging Group
École polytechnique fédérale de Lausanne
Lausane, CH-1015, Switzerland

Editor: Joan Bruna

Abstract

In supervised learning, the regularization path is sometimes used as a convenient theoretical proxy for the optimization path of gradient descent initialized from zero. In this paper, we study a modification of the regularization path for infinite-width 2-layer ReLU neural networks with nonzero initial distribution of the weights at different scales. By exploiting a link with unbalanced optimal-transport theory, we show that, despite the non-convexity of the 2-layer network training, this problem admits an infinite-dimensional convex counterpart. We formulate the corresponding functional-optimization problem and investigate its main properties. In particular, we show that, as the scale of the initialization ranges between 0 and $+\infty$, the associated path interpolates continuously between the so-called kernel and rich regimes. Numerical experiments confirm that, in our setting, the scaling path and the final states of the optimization path behave similarly, even beyond these extreme points.

Keywords: gradient-descent training, regularization path, neural tangent kernel, Γ-convergence, Hellinger–Kantorovich distance

1. Introduction

The mathematical theory of artificial neural networks (NNs) can be tackled from either a dynamic or a static viewpoint[1]. In the dynamic approach, one considers a NN in combination with a training algorithm. Then, one studies the statistical properties of the NN along, and at the end, of the training cycle. In the static approach, one studies NNs as a statistical hypothesis (or candidate) space, independently of any training routine. This space is typ-

∗. From 01.02.2024, the author is affiliated with TU Chemnitz and any future correspondence should be directed to sebastian.neumayer@mathematik.tu-chemnitz.de.

1. Matus Telgarsky, Deep learning theory lecture notes: https://mjt.cs.illinois.edu/dlt/

ically endowed with a norm (or, more generally, a metric) in parameter space, which acts as a regularizer (measure of complexity). Both approaches address distinct aspects. The dynamic approach studies the objects that are the most relevant to practice, but faces the difficulty that those are less tractable theoretically. Thus, much fewer statistical results are known compared to static approaches.

Let us consider a parametric model $\phi\colon \mathbb{R}^p \to \mathcal{F}$, where $\mathbb{R}^p$ is the space of parameters and $\mathcal{F}$ a space of functions, and let $\mathcal{L}\colon \mathcal{F} \to \mathbb{R}$ be an objective function such as the empirical risk. In the dynamic approach, a convenient object of study is the *optimization path* that results from a gradient flow. This path $(\boldsymbol{\theta}_t)_{t\geq 0}$ starts from a given initialization $\boldsymbol{\theta}_0 \in \mathbb{R}^p$ and solves

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\theta}_t = -\nabla\mathcal{L}(\phi(\boldsymbol{\theta}_t)), \tag{1}$$

as well as the associated path $\phi(\boldsymbol{\theta}_t)$ in function space. Many refinements are of course possible to make the model more realistic such as taking into account stochasticity (Li et al., 2019; Pesme et al., 2021), large stepsizes (Wang et al., 2022), or momentum (Su et al., 2014). As for static analyses, they often focus on the constrained path $\boldsymbol{\theta}^*_\delta = \arg\min_{\|\boldsymbol{\theta}\|\leq\delta} \mathcal{L}(\phi(\boldsymbol{\theta}))$ or the *regularization path*

$$\boldsymbol{\theta}^*_\lambda = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\phi(\boldsymbol{\theta})) + \lambda\|\boldsymbol{\theta}\|_2^2. \tag{2}$$

In the simple context of linear parameterizations, formalized as $\phi(\boldsymbol{\theta}) = \boldsymbol{\theta}^\top\Phi$ for some $\Phi \in \mathcal{F}^p$ with the initial parameter $\boldsymbol{\theta}_0 = \mathbf{0}$ for (1), the two approaches are tightly interconnected. More precisely, one creates a close link between the optimization (dynamic) and the regularization (static) paths (Suggala et al., 2018; Ali et al., 2020) by letting the tuning parameter take the form $\lambda = 1/(2t)$.

Scaling Path It is perhaps too optimistic to expect such a tight connection for nonlinear NNs. Indeed, this connection breaks, for example, in the cases studied in (Razin and Cohen, 2020; Woodworth et al., 2020). Still, if the regularization path was to preserve some of the key characteristics of optimization paths (such as certain asymptotic behaviors) this would make the static approach relevant to a better understanding of practical NNs.

For the rest of this work, we notate $\mathcal{L}$ as the empirical risk associated with samples $(\mathbf{x}_k, y_k) \in \mathbb{R}^d \times \mathbb{R}$, $k = 1, \dots, n$ and some loss function $L\colon \mathbb{R}^d \times \mathbb{R} \to [0, +\infty]$. To include the case of arbitrarily wide NNs, we replace the parameter space $\mathbb{R}^p$ with $\mathcal{P}_2(\mathbb{R}^p)$. Accordingly, we shall denote the parametrization function of our regression problem by $\phi[\mu] = \int_{\mathbb{R}^p} \phi(\mathbf{w})\,\mathrm{d}\mu(\mathbf{w})$.

A serious obstacle to the establishment of a link between (1) and (2) in the case of NNs is that it is inconvenient to initialize the optimization from $\mu_0 = \delta_0$ since this is often a stationary point of (1). As a remedy, one may instead initialize from the uniform distribution μ_0 on the sphere with radius α. Hence, we study a modification of the regularization path (2) that takes this nonzero initialization and the generalization to measures into account. More precisely, for fixed μ_0 and $\lambda > 0$, we define the *scaling path* with scale α as

$$\mu^*_\alpha = \arg\min_{\mu\in\mathcal{P}_2(\mathbb{R}^p)} \sum_{k=1}^{n} L\big(\phi[\mu](\mathbf{x}_k), y_k\big) + \lambda R_\alpha(\mu, \mu_0), \tag{3}$$

where the functional $R_\alpha(\mu, \mu_0) = (1+\alpha^2)W_2(\mu, S_{\alpha\#}\mu_0)^2$ with $S_\alpha(\mathbf{w}) = \alpha\mathbf{w}$ acts as our scale-dependent regularizer. A first important observation is that 2-layer ReLU NNs are covered within our framework if we choose

$$\phi_{\mathrm{ReLU}}(\mathbf{w}, \mathbf{x}) = w_1(\mathbf{w}_2^T\mathbf{x})^+, \qquad \mathbf{w} = (w_1, \mathbf{w}_2) \in \mathbb{R}\times\mathbb{R}^d\,. \tag{4}$$

Further, the problem in (3) is convex, which allows the use of standard optimization tools. In the context of 2-layer ReLU NNs with vanishing initialization scale α, we shall see in Section 3 that (3) reduces (up to rescaling of λ) to the regularization path (2).

For our analysis, it is actually more convenient to study (3) from a functional perspective by considering the resulting function $f = \phi[\mu]\colon \mathbb{R}^d \to \mathbb{R}$. As different μ can lead to the same network f, the regularizer $R_\alpha(\cdot, \mu_0)$ needs to be replaced by a complexity measure $N_\alpha(f, \mu_0)$ that takes the whole equivalence class into account. Essentially, $N_\alpha(f, \mu_0)$ describes the distance of a particular NN f to some initialization μ_0. With this notation, the scaling path (3) now takes the form

$$\operatorname*{arg\,min}_{f\in\mathrm{Net}_\phi(\mathbb{R}^d)} \sum_{k=1}^n L\big(f(\mathbf{x}_k), y_k\big) + \lambda N_\alpha(f, \mu_0), \tag{5}$$

where $\mathrm{Net}_\phi(\mathbb{R}^d)$ is a suitable space of functions. Our study of $N_\alpha(f, \mu_0)$ will reveal an interesting link between (3) and the theory of unbalanced optimal transport. Based on this link, we can prove our main result, namely, that the scaling path is Γ-convergent in some suitably chosen metric space. In particular, this implies that the family of minimizing networks $f_\alpha^* \in C(\mathbb{R}^d)$ depends continuously on α. For the limiting cases of vanishing and infinite scales α in (3), we get convergence to two well-known settings from the literature, which are discussed in the next paragraph.

Limits of the Scaling Path As the scale α in (3) vanishes ($S_{\alpha\#}\mu_0 \to \delta_0$), our problem turns into ℓ^2-weight regularization with

$$\operatorname*{arg\,min}_{\mu\in\mathcal{P}_2(\mathbb{R}^p)} \sum_{k=1}^n L\big(\phi[\mu](\mathbf{x}_k), y_k\big) + \lambda \int_{\mathbb{R}^p} \|\mathbf{w}\|^2 \,\mathrm{d}\mu(\mathbf{w}). \tag{6}$$

This problem was thoroughly investigated (Rosset et al., 2007; Bach, 2017; Parhi and Nowak, 2021; Neumayer and Unser, 2023), and is known to admit sparse solutions; namely, minimizers $\hat{\mu}$ that consist of few atoms $\delta_{\mathbf{w}_k}$. In the case of ReLU networks with ϕ_{ReLU}, these correspond to functions of the form $f_\mu = \sum_{k=1}^n w_{k,1}(\mathbf{w}_{k,2}^T\mathbf{x})^+$. Further, (6) leads to predictors with strong statistical properties, such as good adaptivity to anisotropic target functions. Following (Woodworth et al., 2020), we refer to this formulation as the "rich regime". This formulation is known to capture end-of-training behavior of the gradient flow of 2-layer NNs in certain contexts, such as with the logistic loss (Chizat and Bach, 2020; Lyu et al., 2021) or with a small initialization (Boursier et al., 2022).

In contexts such as large initialization with square loss, the training of NNs behaves instead according to the neural tangent kernel (NTK) theory (Jacot et al., 2018; Arora et al., 2019; Bietti and Mairal, 2019). There, the kernel in general form is given by

$$K(\mathbf{x}, \mathbf{x}') = \int_{\mathbb{R}^p} \big(\nabla_{\mathbf{w}}\phi(\mathbf{w}, \mathbf{x})\big)^T \nabla_{\mathbf{w}}\phi(\mathbf{w}, \mathbf{x}')\,\mathrm{d}\mu_0(\mathbf{w}) \tag{7}$$

$\mathcal{P}_2(\mathbb{R}^p)$	probability measures with finite second moments (equipped with 2-Wasserstein metric)
$\mathcal{M}^+(\mathbb{S}^{p-1})$	positive measures on the sphere $\mathbb{S}^{p-1}$
$\mathcal{H}_K$	reproducing kernel Hilbert space corresponding to K
$C(\mathbb{X})$	continuous functions on metric space $\mathbb{X}$ (equipped with supremum norm)
$L^2(\mathbb{R}^d, \mu_0)$	square integrable functions with norm weighted by μ_0
$W^{1,\infty}_{\mathrm{loc}}(\mathbb{R}^p)$	weakly differentiable functions with finite $W^{1,\infty}$-norm on compact sets

Table 1: Function and measure spaces.

and depends on the initial distribution $\mu_0 \in \mathcal{P}_2(\mathbb{R}^p)$. In this *kernel* (a.k.a. *lazy*) regime, the gradient flow in the large-time limit solves the associated kernel-ridge regression problem

$$\operatorname*{arg\,min}_{f \in \mathcal{H}_K} \sum_{k=1}^{n} L(f(\mathbf{x}_k), y_k) + \lambda \|f\|^2_{\mathcal{H}_K}, \tag{8}$$

which we identify as the limit of (3) as $\alpha \to \infty$. Note that the solution of this problem can be written in the form $\hat{f}(\mathbf{x}) = \sum_{k=1}^{n} K(\mathbf{x}, \mathbf{x}_k) c_i$ with $c_i \in \mathbb{R}$. Equivalently, we can also investigate the problem

$$\operatorname*{arg\,min}_{T \in L^2(\mathbb{R}^p, \mu_0)} \sum_{k=1}^{n} L\big(\big\langle T, \nabla_{\mathbf{w}} \phi(\cdot, \mathbf{x}_k) \big\rangle_{L^2(\mathbb{R}^p, \mu_0)}, y_k\big) + \lambda \|T\|^2_{L^2(\mathbb{R}^p, \mu_0)} \tag{9}$$

associated to the corresponding feature map (see (Berlinet and Thomas-Agnan, 2004) for details), which leads to the same solution in function space. The underlying feature map $\nabla_{\mathbf{w}} \phi(\cdot, \mathbf{x})$ is related to the Taylor expansion of the NN parameterization function ϕ around the initial parameters.

Outline To study the scaling path (3), we introduce and analyze the complexity measure $N(f, \mu_0)$ in Section 2. Based on the developed theory for $N(f, \mu_0)$, we investigate in Section 3 the associated family of functional-optimization problems (3). As our main result, we prove that the underlying family of functionals is Γ-convergent and that the rich regime (6) and the kernel regime (8) are the limits for $\alpha \to 0$ and $\alpha \to \infty$, respectively. Our theoretical results are illustrated in Section 4 by a simulation in which we compare the scaling path and the final state of the optimization path for several scales α. Finally, we draw conclusions in Section 5.

2. Infinite-Width Neural Networks

The function and measure spaces used throughout this manuscript are briefly introduced in Table 1. In abstract form, we can parameterize infinite-width NNs using probability measures with finite second-order moments $\mu \in \mathcal{P}_2(\mathbb{R}^p)$, where $\mathbb{R}^p$ is the parameter space, and a function $\phi \colon \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}$ that, in our case, satisfies the following properties.

- *2-homogeneity in w*: $\phi(r\mathbf{w}, \mathbf{x}) = r^2 \phi(\mathbf{w}, \mathbf{x})$ for all $(r, \mathbf{w}, \mathbf{x}) \in \mathbb{R}^+ \times \mathbb{R}^p \, x \times \mathbb{R}^d$.
- *Regularity in w*: for every $\mathbf{x} \in \mathbb{R}^d$, $\phi(\cdot, \mathbf{x}) \in W^{1,\infty}_{\mathrm{loc}}(\mathbb{R}^p)$ is twice continuously differentiable on an open cone $C_{\mathbf{x}} \subset \mathbb{R}^p$ with full Lebesgue measure λ, so that $\lambda(\mathbb{R}^p \setminus C_{\mathbf{x}}) = 0$, and $\|\nabla^2_{\mathbf{w}} \phi(\cdot, \mathbf{x})\|$ is uniformly bounded on $C_{\mathbf{x}}$.

- *Lipschitz regularity in* $\mathbf{x}$: For every $\mathbf{w} \in \mathbb{R}^p$, $\mathbf{x} \mapsto \phi(\mathbf{w}, \mathbf{x})$ is Lipschitz-continuous, and there exists a constant $C > 0$ such that $\mathrm{Lip}(\phi(\mathbf{w}, \cdot)) \leq C\|\mathbf{w}\|^2$.

Some of these conditions are similar to those of (Chizat and Bach, 2020). The first assumption implies that $\int_{\mathbb{R}^p} \phi(\mathbf{w}, \mathbf{x}) \,\mathrm{d}\mu(\mathbf{w}) < \infty$ for all $\mu \in \mathcal{P}_2(\mathbb{R}^p)$ and $\mathbf{x} \in \mathbb{R}^d$. The first two assumptions together imply that $\nabla_{\mathbf{w}}\phi(\cdot, \mathbf{x})$ is (positively) 1-homogeneous in its first variable on $C_{\mathbf{x}}$, in the sense that $\nabla_{\mathbf{w}}\phi(r\mathbf{w}, \mathbf{x}) = r\nabla_{\mathbf{w}}\phi(\mathbf{w}, \mathbf{x})$ for all $(r, \mathbf{w}) \in \mathbb{R}^+ \times C_{\mathbf{x}}$. Hence, $\phi(\cdot, \mathbf{x}) \in W^{1,\infty}_{\mathrm{loc}}(\mathbb{R}^p)$ also implies that $\nabla_{\mathbf{w}}\phi(\cdot, \mathbf{x}) \in L^2(\mathbb{R}^p, \mu)$ for all $\mu \in \mathcal{P}_2(\mathbb{R}^p)$ and $\mathbf{x} \in \mathbb{R}^d$. Using the function ϕ, we define an associated space of infinite width NNs as

$$\mathrm{Net}_\phi(\mathbb{R}^d) = \left\{ \int_{\mathbb{R}^p} \phi(\mathbf{w}, \cdot) \,\mathrm{d}\mu(\mathbf{w}) : \mu \in \mathcal{P}_2(\mathbb{R}^p) \right\}. \tag{10}$$

Note that the parameterization μ of a NN $f \in \mathrm{Net}_\phi(\mathbb{R}^d)$ is not necessarily unique.

Remark 1 *One can readily verify that 2-layer ReLU NNs with parameterization function* $\phi_{\mathrm{ReLU}}(\mathbf{w}, \mathbf{x}) = w_1(\mathbf{w}_2^T\mathbf{x})^+$, $\mathbf{w} = (w_1, \mathbf{w}_2) \in \mathbb{R}^{d+1}$, *fit into this abstract framework. Here,* w_1 *parameterizes the scalar output weights and* $\mathbf{w}_2$ *parameterizes the hidden layer. To allow for bias vectors, we can pad the input vector* $\mathbf{x}$ *with a* 1 *at the end and treat the biases as part of the weights* $\mathbf{w}$. *Given an atomic measure* $\mu = \sum_{k=1}^n \delta_{\mathbf{w}_k}$, *the associated finite width NN reads* $f_\mu = \sum_{k=1}^n w_{k,1}(\mathbf{w}_{k,2}^T\mathbf{x})^+$. *Finally, using the characteristic function* $\chi_{\mathbb{R}^+}$ *of the positive reals, the NTK of a 2-layer ReLU NN with initialization* μ_0 *is given by*

$$K_{\mathrm{ReLU}}(\mathbf{x}, \mathbf{x}') = \int_{\mathbb{R}^{d+1}} \Big((\mathbf{w}_2^T\mathbf{x})^+(\mathbf{w}_2^T\mathbf{x}')^+ + w_1^2\mathbf{x}^T\mathbf{x}'\chi_{\mathbb{R}^+}(\mathbf{w}_2^T\mathbf{x})\chi_{\mathbb{R}^+}(\mathbf{w}_2^T\mathbf{x}') \Big) \,\mathrm{d}\mu_0(\mathbf{w}). \tag{11}$$

Although 2-layer ReLU NNs are the most relevant choice of ϕ from a practical viewpoint, we prefer to carry out our analysis for this general class of functions ϕ. For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $f = \int_{\mathbb{R}^p} \phi(\mathbf{w}, \cdot) \,\mathrm{d}\mu(\mathbf{w}) \in \mathrm{Net}_\phi(\mathbb{R}^d)$, it holds that

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq C \int_{\mathbb{R}^p} \|\mathbf{w}\|^2 \,\mathrm{d}\mu(\mathbf{w}) \|\mathbf{x} - \mathbf{y}\|. \tag{12}$$

Hence, all functions in $\mathrm{Net}_\phi(\mathbb{R}^d)$ are Lipschitz-continuous. Therefore, $\mathrm{Net}_\phi(\mathbb{R}^d)$ is a subset of $C(\mathbb{R}^d)$. In Section 2.1, we construct a complexity measure that encodes the distance of a given NN to a reference parameterization $\mu_0 \in \mathcal{P}_2(\mathbb{R}^p)$, which could be, for example, the initialization of the NN before it is trained according to the gradient flow (1).

2.1 Measure of Complexity for Neural Networks

In the following, we rely heavily on optimal transport and, in particular, on the 2-Wasserstein metric W_2 (Ambrosio et al., 2005; Villani, 2009). Let $\mu_0 \in \mathcal{P}_2(\mathbb{R}^p)$ be a probability measure with polar disintegration $\mu_0(\mathrm{d}r, \mathrm{d}\boldsymbol{\theta}) = \mu_0(\mathrm{d}r|\boldsymbol{\theta})\hat{\mu}_0(\mathrm{d}\boldsymbol{\theta})$, where $\hat{\mu}_0 \in \mathcal{M}^+(\mathbb{S}^{p-1})$ and $\int_{\mathbb{R}^+} r^2\mu_0(\mathrm{d}r|\boldsymbol{\theta}) = 1$ for $\hat{\mu}_0$-a.e. $\boldsymbol{\theta} \in \mathbb{S}^{p-1}$. We then define the *complexity measure* $N(\cdot, \mu_0)\colon \mathrm{Net}_\phi(\mathbb{R}^d) \to \mathbb{R}^+$ as

$$N(f, \mu_0) = \inf_{\mu \in \mathcal{P}_2(\mathbb{R}^p)} \left\{ W_2^2(\mu, \mu_0) : f = \int_{\mathbb{R}^p} \phi(\mathbf{w}, \cdot) \,\mathrm{d}\mu(\mathbf{w}) \right\}. \tag{13}$$

Loosely speaking, $N(f, \mu_0)$ encodes by how much the parameter μ needs to move away from a reference measure μ_0 in order to realize the NN f. Using the Monge formulation of optimal transport, we obtain the upper bound

$$N(f, \mu_0) \leq \inf_{T \in L^2(\mathbb{R}^p, \mu_0)} \left\{ \|T - \mathrm{Id}\|^2_{L^2(\mu_0)} : f = \int_{\mathbb{R}^p} \phi(\mathbf{w}, \cdot) \, \mathrm{d}[T_{\#}\mu_0](\mathbf{w}) \right\}, \tag{14}$$

where $T_{\#}\mu_0 = \mu_0(T^{-1}(\cdot))$ denotes the push-forward measure of μ under T. Because optimal transport maps do not necessarily exist, the right-hand side of (14) is indeed an infimum rather than a minimum. However, the equality of (13) and (14) holds when μ_0 is absolutely continuous with respect to the Lebesgue measure λ. This relation turns out to be useful for the derivation of our main result in Section 3.

Remark 2 *Recently, the idea of using an optimal-transport-based complexity measure for NNs has also been pursued in (Chen et al., 2022, Section 5). In their simplest instance, where the underlying function space is isomorphic to $\mathbb{R}^d$ and σ_2 is the ReLU, they investigate the same NNs as we in Remark 1. Albeit closely related, their complexity measure $\gamma_2^\dagger$ differs from ours since transport plans π are supported on $\mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^d$ instead of $\mathbb{R}^{d+1} \times \mathbb{R}^{d+1}$ and the transport cost is only computed with respect to $\mathbf{w}_2 \in \mathbb{R}^d$. Based on this choice, they are able to derive Rademacher complexity bounds for $\gamma_2^\dagger$, which lead to a posteriori generalization error bounds for gradient-desecent-trained NNs depending on a notion of the length of the optimization path* (1)*. For our more specific 2-homogeneous setting, the aim is instead to study fine properties of the scaling path* (3) *as α varies.*

2.2 Properties of the Complexity Measure

First, we show that the complexity measure N satisfies a homogeneity property.

Lemma 3 (Homogeneity) *For all $f \in \mathrm{Net}_\phi(\mathbb{R}^d)$, $\alpha > 0$, and $\mu_0 \in \mathcal{P}_2(\mathbb{R}^p)$, it holds with $S_\alpha \colon \mathbb{R}^p \to \mathbb{R}^p$ given by $\mathbf{w} \mapsto \alpha \mathbf{w}$ that*

$$\alpha^2 N(f, \mu_0) = N(\alpha^2 f, S_{\alpha\#}\mu_0). \tag{15}$$

Proof Since $f = \int_{\mathbb{R}^p} \phi(\mathbf{w}, \cdot) \, \mathrm{d}\mu(\mathbf{w})$ and ϕ is 2-homogeneous, we have that

$$\int_{\mathbb{R}^p} \phi(\mathbf{w}, \cdot) \, \mathrm{d}[S_{\alpha\#}\mu](\mathbf{w}) = \int_{\mathbb{R}^p} \phi(\alpha\mathbf{w}, \cdot) \, \mathrm{d}\mu(\mathbf{w}) = \alpha^2 f. \tag{16}$$

The result then follows from $W_2(S_{\alpha\#}\mu, S_{\alpha\#}\mu_0)^2 = \alpha^2 W_2(\mu, \mu_0)^2$. ∎

The constraint in (13) has a very specific structure. Using the 2-homogeneous projection operator $\Pi_2 \colon \mathcal{P}_2(\mathbb{R}^p) \to \mathcal{M}^+(\mathbb{S}^{p-1})$ characterized by

$$\int_{\mathbb{S}^{p-1}} \varphi(\boldsymbol{\theta}) \, \mathrm{d}[\Pi_2(\mu)](\theta) = \int_{\mathbb{R}^p} \|\mathbf{w}\|^2 \varphi(\mathbf{w}/\|\mathbf{w}\|) \, \mathrm{d}\mu(\mathbf{w}) \tag{17}$$

for any $\varphi \in C(\mathbb{S}^{p-1})$, we rewrite (13) as

$$f = \int_{\mathbb{R}^p} \phi(\mathbf{w}, \cdot) \, \mathrm{d}\mu(\mathbf{w}) = \int_{\mathbb{S}^{p-1}} \phi(\boldsymbol{\theta}, \cdot) \, \mathrm{d}[\Pi_2(\mu)](\boldsymbol{\theta}). \tag{18}$$

Based on Π_2, we are now in the position to introduce the distance $\widehat{W}_2$ on $\mathcal{M}^+(\mathbb{S}^{p-1})$ known as the Hellinger–Kantorovich or the Wasserstein–Fisher–Rao distance (Liero et al., 2016; Kondratyev et al., 2016; Chizat et al., 2018). We consider the formulation introduced by (Liero et al., 2016), which is given for $\hat{\mu}_1, \hat{\mu}_2 \in \mathcal{M}^+(\mathbb{S}^{p-1})$ by

$$\widehat{W}_2^2(\hat{\mu}_1, \hat{\mu}_2) = \min_{\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^p)} \big\{ W_2^2(\mu_1, \mu_2) : \Pi_2(\mu_1) = \hat{\mu}_1, \Pi_2(\mu_2) = \hat{\mu}_2 \big\} \tag{19}$$

$$= \min_{\pi \in \mathcal{H}_{\leq}(\hat{\mu}_1, \hat{\mu}_2)} \int_{(\mathbb{R}^p)^2} \|\mathbf{w}_1 - \mathbf{w}_2\|^2 \,\mathrm{d}\pi(\mathbf{w}_1, \mathbf{w}_2) + \sum_{i=1}^{2} \big(\hat{\mu}_i - \Pi_2(\pi_i)\big)(\mathbb{S}^{p-1}) \tag{20}$$

$$= \min_{\pi \in \mathcal{M}^+((\mathbb{S}^{p-1})^2)} \sum_{i=1}^{2} \mathrm{KL}(\pi_i, \hat{\mu}_i) - 2 \int_{(\mathbb{S}^{p-1})^2} \log((\boldsymbol{\theta}_1^T \boldsymbol{\theta}_2)^+) \,\mathrm{d}\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2), \tag{21}$$

where $\mathcal{H}_{\leq}(\hat{\mu}_1, \hat{\mu}_2) = \{\pi \in \mathcal{P}_2((\mathbb{R}^p)^2) : \Pi_2(\pi_i) \leq \hat{\mu}_i\}$, and π_i denotes the respective marginal of the plan. It holds that $\widehat{W}_2$ is a metric on $\mathcal{M}^+(\mathbb{S}^{p-1})$, which metrizes the weak convergence (Liero et al., 2016, Thm. 3.6). Further, $\mathcal{M}^+(\mathbb{S}^{p-1})$ equipped with this metric is complete, and bounded sets are relatively compact. Finally, let us remark that

$$\Pi_2\big(\mu_0(\mathrm{d}r, \mathrm{d}\boldsymbol{\theta})\big) = \bigg(\int_{\mathbb{R}^+} r^2 \,\mathrm{d}\mu_0(r|\boldsymbol{\theta}) \bigg) \hat{\mu}_0(\mathrm{d}\boldsymbol{\theta}) = \hat{\mu}_0(\mathrm{d}\boldsymbol{\theta}). \tag{22}$$

Based on these observations, we derive an equivalent formulation for $N(f, \mu_0)$ under the assumption that $\mathrm{supp}(\hat{\mu}_0) \subset (\mathbb{S}^{p-1})$ covers a sufficiently large part of the space.

Proposition 4 (Compact-set formulation) *Let $\mu_0 \in \mathcal{P}_2(\mathbb{R}^p)$ such that the corresponding $\hat{\mu}_0 \in \mathcal{M}^+(\mathbb{S}^{p-1})$ satisfies*

$$\max_{\boldsymbol{\theta}_1 \in \mathbb{S}^{p-1}} \min_{\boldsymbol{\theta}_2 \in \mathrm{supp}(\hat{\mu}_0)} \mathrm{d}_{\mathbb{S}^{p-1}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) < \frac{\pi}{2}. \tag{23}$$

Then, any $\hat{\mu} \in \mathcal{M}^+(\mathbb{S}^{p-1})$ posseses a lift $\mu \in \mathcal{P}_2(\mathbb{R}^p)$ such that $\widehat{W}_2(\hat{\mu}, \hat{\mu}_0) = W_2(\mu, \mu_0)$. Further, it holds for any $f \in \mathrm{Net}_\phi(\mathbb{R}^d)$ that

$$N(f, \mu_0) = \inf_{\hat{\mu} \in \mathcal{M}^+(\mathbb{S}^{p-1})} \bigg\{ \widehat{W}_2^2(\hat{\mu}, \hat{\mu}_0) : f = \int_{\mathbb{S}^{p-1}} \phi(\boldsymbol{\theta}, \cdot) \,\mathrm{d}\hat{\mu}(\boldsymbol{\theta}) \bigg\}. \tag{24}$$

Proof First, recall that $\Pi_2(\mu_0) = \hat{\mu}_0$ due to (22). Based on some $\hat{\pi} \in \mathcal{M}^+((\mathbb{S}^{p-1})^2)$ minimizing $\widehat{W}_2^2(\hat{\mu}, \hat{\mu}_0)$ as in (21), we construct $\mu \in \mathcal{P}_2(\mathbb{R}^p)$ satisfying $\Pi_2(\mu) = \hat{\mu}$ and $\widehat{W}_2(\hat{\mu}, \hat{\mu}_0) = W_2(\mu, \mu_0)$. To this end, we make use of the Lebesgue decompositions $\hat{\mu} = \sigma\hat{\pi}_2 + \hat{\mu}^\perp$ and $\hat{\mu}_0 = \sigma_0\hat{\pi}_1 + \hat{\mu}_0^\perp$. By (Liero et al., 2018, Thm. 6.3b) and (23), we actually have that $\hat{\mu}^\perp = 0$.

Now, we define a measurable map $T_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2} \colon \mathbb{R}^2 \to \mathbb{R}^2$ via

$$T_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2}(r_1, r_2) = \begin{cases} \Big(\sqrt{\frac{\sigma(\boldsymbol{\theta}_1)}{\sigma_0(\boldsymbol{\theta}_2)}} r_1, r_2 \Big) & \text{if } \sigma_0(\boldsymbol{\theta}_2) > 0, \\ (r_1, r_2) & \text{else.} \end{cases} \tag{25}$$

Using $T_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2}$ and $\mu_0(\mathrm{d}r, \mathrm{d}\boldsymbol{\theta}) = \mu_0(\mathrm{d}r|\boldsymbol{\theta})\hat{\mu}_0(\mathrm{d}\boldsymbol{\theta})$, we define a lifted measure $\pi \in \mathcal{P}_2((\mathbb{R}^p)^2)$ via

$$\pi(\mathrm{d}r_1, \mathrm{d}\boldsymbol{\theta}_1, \mathrm{d}r_2, \mathrm{d}\boldsymbol{\theta}_2) = \sigma_0(\boldsymbol{\theta}_2) T_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \#} \big(\delta_{r_2}(\mathrm{d}r_1) \mu_0(\mathrm{d}r_2|\boldsymbol{\theta}_2) \big) \hat{\pi}(\mathrm{d}\boldsymbol{\theta}_1, \mathrm{d}\boldsymbol{\theta}_2). \tag{26}$$

First, observe that the marginal π_2 satisfies for any $\varphi \in C(\mathbb{R}^+ \times \mathbb{S}^{p-1})$ that

$$\begin{aligned}
&\int_{\mathbb{R}^+\times\mathbb{S}^{p-1}} \varphi(r_2, \boldsymbol{\theta}_2)\,\mathrm{d}\pi_2(r_2, \boldsymbol{\theta}_2)\\
&= \int_{(\mathbb{R}^+\times\mathbb{S}^{p-1})^2} \varphi(r_2, \boldsymbol{\theta}_2)\sigma_0(\boldsymbol{\theta}_2) T_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\#}\big(\delta_{r_2}(\mathrm{d}r_1)\mu_0(\mathrm{d}r_2|\boldsymbol{\theta}_2)\big)\hat{\pi}(\mathrm{d}\boldsymbol{\theta}_1, \mathrm{d}\boldsymbol{\theta}_2)\\
&= \int_{(\mathbb{R}^+\times\mathbb{S}^{p-1})^2} \varphi(r_2, \boldsymbol{\theta}_2)\sigma_0(\boldsymbol{\theta}_2)\delta_{r_2}(\mathrm{d}r_1)\mu_0(\mathrm{d}r_2|\boldsymbol{\theta}_2)\hat{\pi}(\mathrm{d}\boldsymbol{\theta}_1, \mathrm{d}\boldsymbol{\theta}_2)\\
&= \int_{\mathbb{R}^+\times\mathbb{S}^{p-1}} \varphi(r_2, \boldsymbol{\theta}_2)\sigma_0(\boldsymbol{\theta}_2)\mu_0(\mathrm{d}r_2|\boldsymbol{\theta}_2)\hat{\pi}_2(\mathrm{d}\boldsymbol{\theta}_2),
\end{aligned} \tag{27}$$

which implies that $\pi_2(\mathrm{d}r, \mathrm{d}\boldsymbol{\theta}) = \sigma_0(\boldsymbol{\theta})\mu_0(\mathrm{d}r|\boldsymbol{\theta})\hat{\pi}_2(\mathrm{d}\boldsymbol{\theta})$. Due to $\int_{\mathbb{R}} r^2\mu_0(\mathrm{d}r|\boldsymbol{\theta}) = 1$ for $\hat{\mu}_0$-a.e. $\boldsymbol{\theta} \in \mathbb{S}^{p-1}$, we further obtain that $\Pi_2(\pi_2) = \sigma_0\hat{\pi}_2$. Again by (Liero et al., 2018, Thm. 6.3b), there exists a Borel set $A \subset \mathrm{supp}(\pi_2)$ with $\pi_2(X \setminus A) = 0$ and $\sigma_0(\boldsymbol{\theta}) > 0$ for all $\boldsymbol{\theta} \in A$. Hence, we get for any $\varphi \in C(\mathbb{S}^{p-1})$ that

$$\begin{aligned}
&\int_{\mathbb{S}^{p-1}} \varphi(\boldsymbol{\theta}_1)\,\mathrm{d}[\Pi_2(\pi_1)](\boldsymbol{\theta}_1)\\
&= \int_{(\mathbb{R}^+\times\mathbb{S}^{p-1})^2} \varphi(\boldsymbol{\theta}_1) r_1^2\sigma_0(\boldsymbol{\theta}_2) T_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\#}\big(\delta_{r_2}(\mathrm{d}r_1)\mu_0(\mathrm{d}r_2|\boldsymbol{\theta}_2)\big)\hat{\pi}(\mathrm{d}\boldsymbol{\theta}_1, \mathrm{d}\boldsymbol{\theta}_2)\\
&= \int_{(\mathbb{R}^+\times\mathbb{S}^{p-1})\times(\mathbb{R}^+\times\{\boldsymbol{\theta}_2:\sigma_0(\boldsymbol{\theta}_2)>0\})} \varphi(\boldsymbol{\theta}_1) r_1^2\sigma(\boldsymbol{\theta}_1)\delta_{r_2}(\mathrm{d}r_1)\mu_0(\mathrm{d}r_2|\boldsymbol{\theta}_2)\hat{\pi}(\mathrm{d}\boldsymbol{\theta}_1, \mathrm{d}\boldsymbol{\theta}_2)\\
&= \int_{\mathbb{S}^{p-1}\times\{\boldsymbol{\theta}_2:\sigma_0(\boldsymbol{\theta}_2)>0\}} \varphi(\boldsymbol{\theta}_1)\sigma(\boldsymbol{\theta}_1)\,\mathrm{d}\hat{\pi}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\\
&= \int_{\mathbb{S}^{p-1}} \varphi(\boldsymbol{\theta}_1)\sigma(\boldsymbol{\theta}_1)\,\mathrm{d}\hat{\pi}_1(\boldsymbol{\theta}_1),
\end{aligned} \tag{28}$$

which implies that $\Pi_2(\pi_1) = \sigma\hat{\pi}_1$. Further, it holds that

$$\begin{aligned}
&\int_{(\mathbb{R}^+\times\mathbb{S}^{p-1})^2} r_1^2 + r_2^2 - 2r_1r_2\boldsymbol{\theta}_1^T\boldsymbol{\theta}_2\,\mathrm{d}\pi(r_1, \boldsymbol{\theta}_1, r_2, \boldsymbol{\theta}_2)\\
&= \int_{(\mathbb{S}^{p-1})^2} \sigma(\boldsymbol{\theta}_1) + \sigma_0(\boldsymbol{\theta}_2) - 2\sqrt{\sigma(\boldsymbol{\theta}_1)\sigma_0(\boldsymbol{\theta}_2)}\boldsymbol{\theta}_1^T\boldsymbol{\theta}_2\,\mathrm{d}\hat{\pi}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\\
&= \int_{(\mathbb{R}^+\times\mathbb{S}^{p-1})^2} r_1^2 + r_2^2 - 2r_1r_2\boldsymbol{\theta}_1^T\boldsymbol{\theta}_2\,\mathrm{d}\big[\big(\sigma(\boldsymbol{\theta}_1)^{1/2}, \boldsymbol{\theta}_1, \sigma_0(\boldsymbol{\theta}_2)^{1/2}, \boldsymbol{\theta}_2\big)_{\#}\hat{\pi}\big](r_1, \boldsymbol{\theta}_1, r_2, \boldsymbol{\theta}_2).
\end{aligned} \tag{29}$$

By (Liero et al., 2018, Thm. 7.20iii), this implies that π is optimal for $\widehat{W}_2(\hat{\mu}, \hat{\mu}_0)$ as in (20). Next, note that the measure

$$\pi^{\perp}(\mathrm{d}r_1, \mathrm{d}\boldsymbol{\theta}_1, \mathrm{d}r_2, \mathrm{d}\boldsymbol{\theta}_2) = \delta_0(\mathrm{d}r_1)\mu_0(\mathrm{d}r_2|\boldsymbol{\theta}_2)\hat{\mu}_0^{\perp}(\mathrm{d}\boldsymbol{\theta}_2) \tag{30}$$

satisfies $\Pi_2(\pi_1^{\perp}) = 0$, $\pi_2^{\perp}(\mathrm{d}r, \mathrm{d}\boldsymbol{\theta}) = \mu_0(\mathrm{d}r|\boldsymbol{\theta})\hat{\mu}_0^{\perp}(\mathrm{d}\boldsymbol{\theta})$, and $\Pi_2(\pi_2^{\perp}) = \hat{\mu}_0^{\perp}$. Since

$$\int_{(\mathbb{R}^+\times\mathbb{S}^{p-1})^2} r_1^2 + r_2^2 - 2r_1r_2\boldsymbol{\theta}_1^T\boldsymbol{\theta}_2\,\mathrm{d}\pi^{\perp}(r_1, \boldsymbol{\theta}_1, r_2, \boldsymbol{\theta}_2) = \hat{\mu}_0^{\perp}(\mathbb{S}^{p-1}), \tag{31}$$

we get that $\pi + \pi^\perp$ is an optimal plan for $\widehat{W}_2(\hat{\mu}, \hat{\mu}_0)$ as in (19) with the required properties.

For the second part, we conclude from (18) and (19) that

$$N(f, \mu_0) \geq \inf_{\hat{\mu} \in \mathcal{M}^+(\mathbb{S}^{p-1})} \left\{ \widehat{W}_2^2(\hat{\mu}, \hat{\mu}_0) : f = \int_{\mathbb{S}^{p-1}} \phi(\boldsymbol{\theta}, \cdot) \, \mathrm{d}\hat{\mu}(\boldsymbol{\theta}) \right\}. \tag{32}$$

Due to the established existence of lifts $\mu \in \mathcal{P}_2(\mathbb{R}^p)$, (32) is sharp and can be replaced by an equality. ∎

Using Proposition 4, which requires (23) to hold, we can prove the existence of minimizers for (13), namely, that the complexity measure is realized by some $\mu \in \mathcal{P}_2(\mathbb{R}^p)$.

Lemma 5 (Minimizing element) *Let $f \in \mathrm{Net}_\phi(\mathbb{R}^d)$ and $\mu_0 \in \mathcal{P}_2(\mathbb{R}^p)$ satisfy (23). Then, there exists $\mu \in \mathcal{P}_2(\mathbb{R}^p)$ with $N(f, \mu_0) = W_2(\mu, \mu_0)^2$ and $f = \int_{\mathbb{R}^p} \phi(\mathbf{w}, \cdot) \, \mathrm{d}\mu(\mathbf{w})$.*

Proof By Proposition 4, it suffices to show existence for (24) since optimal lifts to $\mathcal{P}_2(\mathbb{R}^p)$ do exist. Let $\{\hat{\mu}_k\}_{k \in \mathbb{N}} \subset \mathcal{M}^+(\mathbb{S}^{p-1})$ be a minimizing sequence. As any such sequence lies in a relatively compact set, we can extract a weakly convergent subsequence with limit $\hat{\mu} \in \mathcal{M}^+(\mathbb{S}^{p-1})$. Since $\widehat{W}_2^2(\cdot, \hat{\mu}_0)^2$ is weakly continuous and $\phi(\cdot, \mathbf{x}) \in C(\mathbb{S}^{p-1})$, we get, by definition of the weak convergence, that $\hat{\mu}$ is a minimizing element. ∎

To conclude this section, we prove some additional properties of N.

Lemma 6 (Variational properties) *The complexity measure N has the following properties.*

i) For any $f \in \mathrm{Net}_\phi(\mathbb{R}^d)$ and $\mu_0, \nu_0 \in \mathcal{P}_2(\mathbb{R}^p)$ it holds that

$$\left| \sqrt{N(f, \mu_0)} - \sqrt{N(f, \nu_0)} \right| \leq W_2(\nu_0, \mu_0). \tag{33}$$

ii) Let $\mu_0 \in \mathcal{P}_2(\mathbb{R}^p)$. For any $\{f_k\}_{k \in \mathbb{N}} \subset C(\mathbb{R}^d)$ with $f_k \to f \in C(\mathbb{R}^d)$ pointwise, it holds $N(f, \mu_0) \leq \liminf_{k \to \infty} N(f_k, \mu_0)$.

iii) For any $\mu_0 \in \mathcal{P}_2(\mathbb{R}^p)$, the functional $N(\cdot, \mu_0)$ is convex. If μ_0 is absolutely continuous with respect to the Lebesgue measure λ and satisfies (23), then $N(\cdot, \mu_0)$ is strictly convex.

iv) For all $f \in \mathrm{Net}_\phi(\mathbb{R}^d)$ and $\mathbf{x} \in \mathbb{R}^d$, it holds that

$$N(f, \delta_0) \geq \frac{f(\mathbf{x})}{\|\phi(\cdot, \mathbf{x})\|_{C(\mathbb{S}^{p-1})}}. \tag{34}$$

Proof i) Let $\mu \in \mathcal{P}_2(\mathbb{R}^p)$ with $f = \int_{\mathbb{R}^p} \phi(\mathbf{w}, \cdot) \, \mathrm{d}\mu(\mathbf{w})$. By definition of N, we get that

$$\sqrt{N(f, \mu_0)} \leq W_2(\mu, \mu_0) \leq W_2(\mu, \nu_0) + W_2(\nu_0, \mu_0). \tag{35}$$

Taking the infimum over all such μ, we get that $\sqrt{N(f, \mu_0)} \leq \sqrt{N(f, \nu_0)} + W_2(\nu_0, \mu_0)$ which, by symmetry of W_2, implies (33).

ii) First, we can assume that $\{N(f_k,\mu)\}_{k\in\mathbb{N}}$ has a bounded subsequence (the statement is trivial otherwise). Let $\{\hat{\mu}_k\}_{k\in\mathbb{N}} \subset \mathcal{M}^+(\mathbb{S}^{p-1})$ be a sequence with $f_k = \int_{\mathbb{S}^{p-1}} \phi(\boldsymbol{\theta},\cdot)\,\mathrm{d}\hat{\mu}_k(\boldsymbol{\theta})$ and $N(f_k,\mu_0) + 1/k \geq \widehat{W}_2^2(\hat{\mu}_k,\hat{\mu}_0)^2$. Hence, there exists a weakly convergent subsequence $\{\hat{\mu}_{k_j}\}_{j\in\mathbb{N}}$ with $\liminf_{k\to\infty} N(f_k,\mu_0) = \lim_{j\to\infty} \widehat{W}_2^2(\hat{\mu}_{k_j},\hat{\mu}_0)^2$. Since $\widehat{W}_2^2(\cdot,\hat{\mu}_0)^2$ is weakly continuous and $\phi(\cdot,\mathbf{x}) \in C(\mathbb{S}^{p-1})$, we further get that its limit $\hat{\mu} \in \mathcal{M}^+(\mathbb{S}^{p-1})$ satisfies that $\widehat{W}_2^2(\hat{\mu},\hat{\mu}_0)^2 \leq \liminf_{k\to\infty} N(f_k,\mu_0)$ and $f = \int_{\mathbb{S}^{p-1}} \phi(\boldsymbol{\theta},\cdot)\,\mathrm{d}\hat{\mu}(\boldsymbol{\theta})$. Hence, it holds that $N(f,\mu_0) \leq \liminf_{k\to\infty} N(f_k,\mu_0)$.

iii) Let $\lambda \in (0,1)$, $f_1, f_2 \in \mathrm{Net}_\phi(\mathbb{R}^d)$, and $\epsilon > 0$. Then, there exist $\mu_1,\mu_2 \in \mathcal{P}_2(\mathbb{R}^p)$ with $f_i = \int_{\mathbb{R}^p} \phi(\mathbf{w},\cdot)\,\mathrm{d}\mu_i(\mathbf{w})$ and $W_2^2(\mu_i,\mu_0) \leq N(f_i,\mu_0) + \epsilon$. Further, it is well-known that $W_2^2(\cdot,\mu_0)$ is convex. Consequently, we get that

$$\begin{aligned} N\big(\lambda f_1 + (1-\lambda) f_2, \mu_0\big) &\leq W_2^2\big(\lambda\mu_1 + (1-\lambda)\mu_2, \mu_0\big) \\ &\leq \lambda N(f_1,\mu_0) + (1-\lambda) N(f_2,\mu_0) + \epsilon. \end{aligned} \tag{36}$$

Convexity follows by taking $\epsilon \to 0$. If μ_0 is absolutely continuous with respect to λ and if (23) holds, then we can choose $\epsilon = 0$ and the result follows similarly as before due to the strict convexity of $W_2^2(\cdot,\mu_0)$ in this setting.

iv) Let $\mu \in \mathcal{P}_2(\mathbb{R}^p)$ satisfy $f(\mathbf{x}) = \int_{\mathbb{R}^p} \phi(\mathbf{w},\mathbf{x})\,\mathrm{d}\mu(\mathbf{w})$. Then, we estimate

$$f(\mathbf{x}) = \int_{\mathbb{R}^p} \phi(\mathbf{w}/\|\mathbf{w}\|_2, x)\|\mathbf{w}\|_2^2\,\mathrm{d}\mu(w) \leq \|\phi(\cdot,\mathbf{x})\|_{C(\mathbb{S}^{p-1})} \int_{\mathbb{R}^p} \|\mathbf{w}\|_2^2\,\mathrm{d}\mu(\mathbf{w}). \tag{37}$$

Hence, we conclude that $W_2(\mu,\delta_0)^2 \geq f(\mathbf{x})/\|\phi(\cdot,\mathbf{x})\|_{C(\mathbb{S}^{p-1})}$ and the claim follows. ∎

3. Interpolating Between the Rich and Kernel Regimes

As discussed in Section 1, it is known that in specific settings the gradient flow (1) converges to the rich regime (6) for small initializations and to the kernel regime (8) for large initializations. In this section, we show that the scaling path (3) interpolates continuously between these two endpoints as α varies from 0 to $+\infty$. To this end, we assume that we are given training samples $(\mathbf{x}_k, y_k) \in \mathbb{R}^d \times \mathbb{R}$, $k = 1,\dots,n$, such that $\nabla_{\mathbf{w}}\phi(\cdot,\mathbf{x}_k) \in L^2(\mathbb{R}^p,\mu_0)$, $k = 1,\dots,n$, are linearly independent. For the choice ϕ_{ReLU} from Remark 1, this is, for example, the case if the locations $\mathbf{x}_k$ of the training samples are distinct. Then, we can formulate a corresponding regularized learning problem

$$\operatorname*{arg\,min}_{f\in C(\mathbb{R}^d)} \sum_{k=1}^n L\big(f(\mathbf{x}_k), y_k\big) + \lambda(1+\alpha^2) N(f, S_{\alpha\#}\mu_0), \tag{38}$$

where $\alpha \in [0,\infty)$ is an interpolation parameter, $\lambda > 0$ is a regularization parameter, and the loss $L(\cdot,y_k)$ is proper, convex, and lower-semicontinuous for every $k = 1,\dots,n$.

Remark 7 *All of the results in this section remain true if we investigate*

$$\operatorname*{arg\,min}_{f\in C(\mathbb{R}^d)} \sum_{k=1}^n L\big(f(\mathbf{x}_k), y_k\big) \quad \textit{s.t. } (1+\alpha^2) N(f, S_{\alpha\#}\mu_0) \leq \delta \tag{39}$$

with L strictly convex. If L is only convex, then the uniqueness results do not hold.

For instance, Problem (38) includes classification problems with $y_k \in \{-1, 1\}$ and

$$\operatorname*{arg\,min}_{f \in C(\mathbb{R}^d)} N(f, S_{\alpha\#}\mu_0) \qquad \text{s.t. } y_k f(\mathbf{x}_k) \geq 1 \quad \forall k = 1, \dots, n, \tag{40}$$

as well as interpolation problems with $y_k \in \mathbb{R}$ and

$$\operatorname*{arg\,min}_{f \in C(\mathbb{R}^d)} N(f, S_{\alpha\#}\mu_0) \qquad \text{s.t. } f(\mathbf{x}_k) = y_k \quad \forall k = 1, \dots, n \tag{41}$$

as special cases. When L is the square loss, the interpolation problem (41) can be interpreted as the endpoint of the modified regularization path (3) as $\lambda \to 0$. Instead of (38), we can also investigate the *equivalent* parameter-space problems

$$\operatorname*{arg\,min}_{\mu \in \mathcal{P}_2(\mathbb{R}^p)} \sum_{k=1}^{n} L\left(\int_{\mathbb{R}^p} \phi(\mathbf{w}, \mathbf{x}_k)\, \mathrm{d}\mu(\mathbf{w}), y_k\right) + \lambda(1+\alpha^2) W_2^2(\mu, S_{\alpha\#}\mu_0) \tag{42}$$

and

$$\operatorname*{arg\,min}_{\hat{\mu} \in \mathcal{M}^+(\mathbb{S}^{p-1})} \sum_{k=1}^{n} L\left(\int_{\mathbb{S}^{p-1}} \phi(\boldsymbol{\theta}, \mathbf{x}_k)\, \mathrm{d}\hat{\mu}(\boldsymbol{\theta}), y_k\right) + \lambda(1+\alpha^2) \widehat{W}_2^2\big(\hat{\mu}, \alpha^2 \hat{\mu}_0\big). \tag{43}$$

These reformulations are essential to prove the continuity of the optimal solutions f_α^* for (38) with respect to α in Theorem 13. First, however, we establish the existence of minimizers for (38).

Lemma 8 *Assume that* (38) *is feasible. Then, there exists a minimizer. If, additionally, μ_0 is absolutely continuous with respect to the Lebesgue measure λ and satisfies* (23)*, then the minimizer is unique for $\alpha > 0$.*

Proof Let $\{f_k\}_{k\in\mathbb{N}}$ be a minimizing sequence, which implies that the corresponding sequence $\{N(f_k, S_{\alpha\#}\mu_0)\}_{k\in\mathbb{N}}$ is bounded. Similarly as in the proof of Lemma 6ii), we can extract a subsequence $\{f_{k_j}\}_{j\in\mathbb{N}}$ such that there exists a $\mu \in \mathcal{P}_2(\mathbb{R}^p)$ with $N(f_{k_j}, S_{\alpha\#}\mu_0) \to W_2^2(\mu, \mu_0)$ and $f_{k_j} \to f = \int_{\mathbb{R}^p} \phi(\mathbf{w}, \cdot)\, \mathrm{d}\mu(\mathbf{w})$ point-wise. However, this readily implies that $\sum_{l=1}^n L(f(\mathbf{x}_l), y_l) \leq \liminf_{j\to\infty} \sum_{l=1}^n L(f_{k_j}(\mathbf{x}_l), y_l)$ and further that $N(f, S_{\alpha\#}\mu_0) \leq \lim_{j\to\infty} N(f_{k_j}, S_{\alpha\#}\mu_0)$. Hence, we get that f is a minimizer. If the additional assumptions hold, uniqueness follows by strict convexity (see Lemma 6). ∎

Remark 9 *A similar statement also holds for the formulations* (42) *and* (43)*.*

Now, we investigate the behavior of the functional in (43) as α varies. To this end, we rely on the concept of Γ-convergence (see (Braides, 2002) for a detailed exposition). Let $\mathbb{X}$ be a topological space. Recall that $\{J_k\}_{k\in\mathbb{N}}$ with $J_k \colon \mathbb{X} \to [0, \infty]$ is said to Γ-converge to $J \colon \mathbb{X} \to [0, \infty]$ if the following two conditions are fulfilled for every $x \in \mathbb{X}$:

i) it holds that $J(x) \leq \liminf_{k\to\infty} J_k(x_k)$ whenever $x_k \to x$;

ii) there is a sequence $\{x_k\}_{k\in\mathbb{N}}$ with $x_k \to x$ and $\limsup_{k\to\infty} J_k(x_k) \leq J(x)$.

The importance of Γ-convergence is captured by Theorem 10. Recall that a family of functionals $J_k \colon \mathbb{X} \to \mathbb{R}$ is equicoercive if it is bounded from below by a coercive functional.

Theorem 10 (Theorem of Γ-convergence (Braides, 2002)) *Let $\{J_k\}_{k\in\mathbb{N}}$ be an equicoercive family of functionals $J_k \colon \mathbb{X} \to \mathbb{R}$. If J_k Γ-converges to J, then it holds that*

- *the optimal functional values converge* $\lim_{k\to\infty} \inf_{x\in\mathbb{X}} J_k(x) = \inf_{x\in\mathbb{X}} J(x)$;
- *all accumulation points of the minimizers of J_k are minimizers of J.*

Although, Theorem 10 and the next two paragraphs on Γ-convergence of the functional in (43) might appear quite abstract at first glance, they will ultimately enable us to prove continuity of the optimal solutions f_α^* for (38) with respect to α in our main Theorem 13.

Rich Regime Using Γ-convergence, we first investigate $\alpha \to \alpha_* \neq \infty$ and equip $\mathcal{M}^+(\mathbb{S}^{p-1})$ with the usual weak topology.

Proposition 11 *For $\alpha \to \alpha_* \neq \infty$, we have Γ-convergence of the functionals in*

$$\min_{\hat{\mu}\in\mathcal{M}^+(\mathbb{S}^{p-1})} \sum_{k=1}^{n} L\bigg(\int_{\mathbb{S}^{p-1}} \phi(\boldsymbol{\theta}, \mathbf{x}_k)\,\mathrm{d}\hat{\mu}(\boldsymbol{\theta}), y_k\bigg) + \lambda(1+\alpha^2)\widehat{W}_2^2\big(\hat{\mu}, \alpha^2\hat{\mu}_0\big). \tag{44}$$

Furthermore, the family of functionals in (44) *is equicoercive.*

Proof We first note that the functionals in (44) are equicoercive since

$$(1+\alpha^2)\widehat{W}_2^2\big(\hat{\mu}, \alpha^2\hat{\mu}_0\big) \geq \Big(\widehat{W}_2(\hat{\mu}, 0) - \widehat{W}_2\big(0, \alpha^2\hat{\mu}_0\big)\Big)^2 = \big(\widehat{W}_2(\hat{\mu}, 0) - \alpha\sqrt{\hat{\mu}_0(\mathbb{S}^{p-1})}\big)^2 \tag{45}$$

and L maps into $[0,\infty]$. For the lim inf inequality, let $\{\hat{\mu}_k\}_{k\in\mathbb{N}}$ and $\{\alpha_k\}_{k\in\mathbb{N}}$ be sequences with limits $\hat{\mu}$ and α_*, respectively. Since $\phi(\cdot, \mathbf{x}_l)$ is continuous, this directly implies that $\int_{\mathbb{S}^{p-1}} \phi(\boldsymbol{\theta}, \mathbf{x}_l)\,\mathrm{d}\hat{\mu}_k(\boldsymbol{\theta}) \to \int_{\mathbb{S}^{p-1}} \phi(\boldsymbol{\theta}, \mathbf{x}_l)\,\mathrm{d}\hat{\mu}(\boldsymbol{\theta})$ for all $l = 1, \ldots, n$. Then, since

$$\begin{aligned}(1+\alpha_k^2)\widehat{W}_2^2\big(\hat{\mu}_k, \alpha_k^2\hat{\mu}_0\big) &\geq (1+\alpha_k^2)\Big(\widehat{W}_2\big(\hat{\mu}_k, \alpha_*^2\hat{\mu}_0\big) - \widehat{W}_2\big(\alpha_*^2\hat{\mu}_0, \alpha_k^2\hat{\mu}_0\big)\Big)^2 &(46)\\ &\geq (1+\alpha_k^2)\Big(\widehat{W}_2\big(\hat{\mu}_k, \alpha_*^2\hat{\mu}_0\big) - \sqrt{|\alpha_*^2 - \alpha_k^2|\hat{\mu}_0(\mathbb{S}^{p-1})}\Big)^2, &(47)\end{aligned}$$

the claim follows by the continuity of $\widehat{W}_2$ and the lower-semicontinuity of $L(\cdot, y_l)$. Finally, the lim sup inequality follows if we let the recovery sequence be constant. ∎

Note that for $\alpha = 0$, problem (44) can be rewritten as

$$\min_{\hat{\mu}\in\mathcal{M}^+(\mathbb{S}^{p-1})} \sum_{k=1}^{n} L\bigg(\int_{\mathbb{S}^{p-1}} \phi(\boldsymbol{\theta}, \mathbf{x}_k)\,\mathrm{d}\hat{\mu}(\boldsymbol{\theta}), y_k\bigg) + \lambda\,\mathrm{TV}(\hat{\mu}). \tag{48}$$

Kernel Regime Next, we want to discuss the case $\alpha \to \infty$ and show that we approach the NTK problem (9) with feature maps $\mathbf{x} \mapsto \nabla_{\mathbf{w}}\phi(\cdot, \mathbf{x})$ if we reformulate (43) accordingly.

Proposition 12 *Let μ_0 be absolutely continuous with respect to the Lebesgue measure and $\int_{\mathbb{R}^p} \phi(\mathbf{w}, \cdot)\,\mathrm{d}\mu_0(\mathbf{w}) = 0$. Further, let $L(\cdot, y_k)$, $k = 1, \dots, n$, be either left- or right-continuous in every point of its domain. Then, for $\alpha \to \infty$, we have Γ-convergence of the functionals in*

$$\operatorname*{arg\,min}_{T \in L^2(\mathbb{R}^p, \mu_0)} \sum_{k=1}^{n} L\left(\int_{\mathbb{R}^p} \phi\big(\alpha\mathbf{w} + \alpha^{-1}T(\mathbf{w}), \mathbf{x}_k\big)\,\mathrm{d}\mu_0(\mathbf{w}), y_k\right) + \lambda\frac{\alpha^2+1}{\alpha^2}\|T\|^2_{L^2(\mathbb{R}^p,\mu_0)}, \tag{49}$$

which is a reformulation of (43) *using transport maps, to the one in*

$$\operatorname*{arg\,min}_{T \in L^2(\mathbb{R}^p, \mu_0)} \sum_{i=k}^{n} L\big(\big\langle T, \nabla_{\mathbf{w}}\phi(\cdot, \mathbf{x}_k)\big\rangle_{L^2(\mathbb{R}^p,\mu_0)}, y_k\big) + \lambda\|T\|^2_{L^2(\mathbb{R}^p,\mu_0)} \tag{50}$$

with respect to the weak topology in $L^2(\mathbb{R}^p, \mu_0)$. Further, the functionals in (49) *are equicoercive.*

Proof Equicoercivity of the functionals in (49) holds since $\|T\|^2_{L^2(\mathbb{R}^p,\mu_0)}$ is a lower bound for all of them. Due to the absolute continuity with respect to the Lesbegue measure, we can use the equivalent formulation (14) of the complexity measure in (13) to obtain

$$\begin{aligned}
&\min_{\mu \in \mathcal{P}_2(\mathbb{R}^p)} \sum_{k=1}^{n} L\left(\int_{\mathbb{R}^p} \phi(\mathbf{w}, \mathbf{x}_k)\,\mathrm{d}\mu(\mathbf{w}), y_k\right) + \lambda(1+\alpha^2)W_2^2(\mu, S_{\alpha\#}\mu_0)\\
&= \min_{T \in L^2(\mathbb{R}^p,\mu_0)} \sum_{k=1}^{n} L\left(\int_{\mathbb{R}^p} \phi\big(\mathbf{w}, \mathbf{x}_k\big)\,\mathrm{d}\big[(S_\alpha + \alpha^{-1}T)_{\#}\mu_0\big](\mathbf{w}), y_k\right) + \lambda\frac{\alpha^2+1}{\alpha^2}\|T\|^2_{L^2(\mathbb{R}^p,\mu_0)}\\
&= \min_{T \in L^2(\mathbb{R}^p,\mu_0)} \sum_{k=1}^{n} L\left(\int_{\mathbb{R}^p} \phi\big(\alpha\mathbf{w} + \alpha^{-1}T(\mathbf{w}), \mathbf{x}_k\big)\,\mathrm{d}\mu_0(\mathbf{w}), y_k\right) + \lambda\frac{\alpha^2+1}{\alpha^2}\|T\|^2_{L^2(\mathbb{R}^p,\mu_0)}.
\end{aligned} \tag{51}$$

Now, since ϕ is twice continuously differentiable on $C_{\mathbf{x}_k}$, we get, for any $\mathbf{w} \in C_{\mathbf{x}_k}$, that

$$\phi(\mathbf{w}+\mathbf{h}, \mathbf{x}_k) = \phi(\mathbf{w}, \mathbf{x}_k) + \big(\nabla_{\mathbf{w}}\phi(\mathbf{w}, \mathbf{x}_k)\big)^T\mathbf{h} + R(\mathbf{w}, \mathbf{h}, \mathbf{x}_k). \tag{52}$$

As $t \mapsto \phi(\mathbf{w} + t\mathbf{h}, \mathbf{x}_k)$ is absolutely continuous, the remainder $R(\mathbf{w}, \mathbf{h}, \mathbf{x}_k)$ can be estimated for any $\mathbf{w} \in C_{\mathbf{x}_k}$ and $\mathbf{h} \in \mathbb{R}^p$ by

$$|R(\mathbf{w}, \mathbf{h}, \mathbf{x}_k)| \le \max_{\tilde{\mathbf{w}} \in B(\mathbf{w}, \|\mathbf{h}\|)} \big\|\nabla_{\mathbf{w}}\phi(\tilde{\mathbf{w}}, \mathbf{x}_k) - \nabla_{\mathbf{w}}\phi(\mathbf{w}, \mathbf{x}_k)\big\|\|\mathbf{h}\|. \tag{53}$$

If additionally $\mathbf{h} \in B(\mathbf{w}, \epsilon)$, where the radius ϵ depends on $\mathbf{w}$ and $\mathbf{x}_i$, we can use differentiability to even get

$$|R(\mathbf{w}, \mathbf{h}, \mathbf{x}_i)| \le \sup_{\tilde{\mathbf{w}} \in B(\mathbf{w}, \epsilon)} \|\nabla^2_{\mathbf{w}}\phi(\tilde{\mathbf{w}}, \mathbf{x}_i)\|\|\mathbf{h}\|^2. \tag{54}$$

By defining the function

$$R_{\mathbf{x}_i,\alpha}(T) \coloneqq \int_{\mathbb{R}^p} R(\alpha\mathbf{w}, T(\mathbf{w})/\alpha, \mathbf{x}_i)\,\mathrm{d}\mu_0(\mathbf{w}), \tag{55}$$

we rewrite (49) for the following Γ-convergence discussion as

$$\min_{T\in L^2(\mathbb{R}^p,\mu_0)} \sum_{k=1}^{n} L\Big(\big\langle T, \nabla_{\mathbf{w}}\phi(\cdot,\mathbf{x}_k)\big\rangle_{L^2(\mathbb{R}^p,\mu_0)} + R_{\mathbf{x}_i,\alpha}(T), y_k\Big) + \lambda\frac{\alpha^2+1}{\alpha^2}\|T\|^2_{L^2(\mathbb{R}^p,\mu_0)}. \tag{56}$$

For the lim inf inequality of Γ-convergence, let $\{T_k\}_{k\in\mathbb{N}}$ and $\{\alpha_k\}_{k\in\mathbb{N}}$ be (weakly) convergent sequences with limits T and ∞, respectively. Since weakly convergent sequences are bounded, we get that $\|T_k\|_{L^2(\mathbb{R}^p,\mu_0)}/\alpha_k \to 0$. Hence, we can drop to a subsequence that satisfies $T_k(\mathbf{w})/\alpha_k \to 0$ for μ_0-a.e. $\mathbf{w}\in\mathbb{R}^p$, and there exists $g\in L^2(\mathbb{R}^p,\mu_0)$ with $|T_k(\mathbf{w})/\alpha_k| \le g(\mathbf{w})$ for μ_0-a.e. $\mathbf{w}\in\mathbb{R}^p$. Now, observe that

$$\begin{aligned}
|R_{\mathbf{x}_l,\alpha_k}(T_k)| &\le \int_{\mathbb{R}^p} \max_{\tilde{\mathbf{w}}\in B(\alpha_k\mathbf{w}, g(\mathbf{w}))} \|\nabla_{\mathbf{w}}\phi(\tilde{\mathbf{w}},\mathbf{x}_l) - \nabla_{\mathbf{w}}\phi(\alpha_k\mathbf{w},\mathbf{x}_l)\|\frac{\|T_k\|}{\alpha_k}\,\mathrm{d}\mu_0(\mathbf{w}) \\
&\le \left(\int_{\mathbb{R}^p} \max_{\tilde{\mathbf{w}}\in B(\mathbf{w},\frac{g(\mathbf{w})}{\alpha_k})} \|\nabla_{\mathbf{w}}\phi(\tilde{\mathbf{w}},\mathbf{x}_l) - \nabla_{\mathbf{w}}\phi(\mathbf{w},\mathbf{x}_l)\|^2\,\mathrm{d}\mu_0(\mathbf{w})\right)^{1/2} \|T_k\|_{L^2(\mathbb{R}^p,\mu_0)}.
\end{aligned} \tag{57}$$

Here, the integrand converges pointwise to 0 for every $\mathbf{w}\in C_{\mathbf{x}_l}$, and can be bounded by

$$\begin{aligned}
&\max_{\tilde{\mathbf{w}}\in B(\mathbf{w},g(\mathbf{w})/\alpha_k)} \|\nabla_{\mathbf{w}}\phi(\tilde{\mathbf{w}},\mathbf{x}_l) - \nabla_{\mathbf{w}}\phi(\mathbf{w},\mathbf{x}_l)\| \\
&\le \max_{\tilde{\mathbf{w}}\in B(\mathbf{w},g(\mathbf{w})/\alpha_k)} \|\tilde{\mathbf{w}}\|\left\|\nabla_{\mathbf{w}}\phi\Big(\frac{\tilde{\mathbf{w}}}{\|\tilde{\mathbf{w}}\|},\mathbf{x}_l\Big)\right\| + \|\mathbf{w}\|\left\|\nabla_{\mathbf{w}}\phi\Big(\frac{\mathbf{w}}{\|\mathbf{w}\|},\mathbf{x}_l\Big)\right\| \\
&\le \Big(2\|\mathbf{w}\| + \frac{g(\mathbf{w})}{\alpha_k}\Big) \max_{\tilde{\mathbf{w}}\in\mathbb{S}^p} \|\nabla_{\mathbf{w}}\phi(\tilde{\mathbf{w}},\mathbf{x}_l)\|.
\end{aligned} \tag{58}$$

From the dominated-convergence theorem, one has that $R_{\mathbf{x}_l,\alpha_k}(T_k)\to 0$. Given that $T_k \rightharpoonup T$, the lim inf inequality now follows as

$$\begin{aligned}
&\sum_{l=1}^{n}\big(\big\langle T, \nabla_{\mathbf{w}}\phi(\cdot,\mathbf{x}_l)\big\rangle_{L^2(\mu_0)}, y_l\big) + \lambda\|T\|^2_{L^2(\mathbb{R}^p,\mu_0)} \\
&\le \liminf_{k\to\infty} \sum_{l=1}^{n} L\big(\big\langle T_k, \nabla_{\mathbf{w}}\phi(\cdot,\mathbf{x}_l)\big\rangle_{L^2(\mathbb{R}^p,\mu_0)} + R_{\mathbf{x}_l,\alpha_k}(T_k), y_l\big) + \lambda\frac{\alpha_k^2+1}{\alpha_k^2}\|T_k\|^2_{L^2(\mathbb{R}^p,\mu_0)}.
\end{aligned} \tag{59}$$

For the lim sup inequality, we can assume that T has finite energy. Further, we use a dual basis $D_l\in L^2(\mathbb{R}^p,\mu_0)$, $l=1,\dots,n$, of the feature maps $\nabla_{\mathbf{w}}\phi(\cdot,\mathbf{x}_l)$. Then, we define $h(\mathbf{w}) \coloneqq |T(\mathbf{w})| + \sum_{l=1}^{n}|D_l(\mathbf{w})|$ and

$$M_{l,k} = \left(\int_{\mathbb{R}^p} \max_{\tilde{\mathbf{w}}\in B(\mathbf{w},h(\mathbf{w})/\alpha_k^2)} \|\nabla_{\mathbf{w}}\phi(\tilde{\mathbf{w}},\mathbf{x}_l) - \nabla_{\mathbf{w}}\phi(\mathbf{w},\mathbf{x}_l)\|^2\,\mathrm{d}\mu_0(\mathbf{w})\right)^{1/2} \|\mathbf{h}\|_{L^2(\mathbb{R}^p,\mu_0)}. \tag{60}$$

Now, set $s_l = 1$ if $L(\cdot, y_l)$ is right-continuous in $\langle T, \nabla_{\mathbf{w}}\phi(\cdot, \mathbf{x}_l)\rangle_{L^2(\mathbb{R}^p,\mu_0)}$ and $s_l = -1$ if it is left-continuous. Finally, we pick $T_k = T + \sum_l s_l M_{l,k} D_l$ as recovery sequence.

As in the first part of the proof, we can show that $M_{l,k} \to 0$ for $k \to \infty$. Hence, we can estimate as in (57) and obtain that $R_{\mathbf{x}_l,\alpha_k}(T_k) \to 0$ for $k \to \infty$. In the right-continuous case, it holds that

$$\big\langle T_k, \nabla_{\mathbf{w}}\phi(\cdot, \mathbf{x}_l)\big\rangle_{L^2(\mathbb{R}^p,\mu_0)} + R_{\mathbf{x}_l,\alpha_k}(T_k) \geq \big\langle T, \nabla_{\mathbf{w}}\phi(\cdot, \mathbf{x}_l)\big\rangle_{L^2(\mathbb{R}^p,\mu_0)}. \tag{61}$$

For the left-continuous case, we get that

$$\big\langle T_k, \nabla_{\mathbf{w}}\phi(\cdot, \mathbf{x}_l)\big\rangle_{L^2(\mathbb{R}^p,\mu_0)} + R_{\mathbf{x}_l,\alpha_k}(T_k) \leq \big\langle T, \nabla_{\mathbf{w}}\phi(\cdot, \mathbf{x}_l)\big\rangle_{L^2(\mathbb{R}^p,\mu_0)}. \tag{62}$$

Hence, we have for $l = 1, \ldots, n$ that $\langle T_k, \nabla_{\mathbf{w}}\phi(\cdot, \mathbf{x}_l)\rangle_{L^2(\mathbb{R}^p,\mu_0)} \to \langle T, \nabla_{\mathbf{w}}\phi(\cdot, \mathbf{x}_l)\rangle_{L^2(\mathbb{R}^p,\mu_0)}$ from the required direction, which concludes the proof. ■

Implications for $\mathrm{Net}_\phi(\mathbb{R}^d)$ Assume that μ_0 is absolutely continuous with respect to the Lebesgue measure. Observe that Proposition 11 and Theorem 10 directly imply that the family of measures $\hat{\mu}^*_\alpha$, $\alpha \in (0, \infty)$, determined by (43) is continuous in the $\widehat{W}_2$ metric. Further, for $\alpha \to 0$, these measures converge to some optimal solution $\hat{\mu}^*_0$ of (48). Finally, Proposition 12 and Theorem 10 imply that the the solutions T^*_α of (49) converge to an optimal solution of (50) in the weak L^2-topology. Equivalently, we can state these observations in terms of the optimal NNs as

$$f^*_\alpha = \int_{\mathbb{S}^{p-1}} \phi(\boldsymbol{\theta}, \cdot)\, \mathrm{d}\hat{\mu}^*_\alpha(\boldsymbol{\theta}) = \int_{\mathbb{R}^p} \phi\big(\mathbf{w}\theta + \alpha^{-1} T^*_\alpha(\mathbf{w}), \cdot\big)\, \mathrm{d}\mu_0(\mathbf{w}), \quad \alpha < \infty \tag{63}$$

and

$$f^*_\infty(\mathbf{x}) = \big\langle T^*_\infty, \nabla_{\mathbf{w}}\phi(\cdot, \mathbf{x})\big\rangle_{L^2(\mathbb{R}^p,\mu_0)}, \tag{64}$$

where μ^*_α, T^*_α, and T^*_∞ solve (43), (49), and (50), respectively.

Theorem 13 *Assume that μ_0 is absolutely continuous with respect to the Lebesgue measure. Then, the family $f^*_\alpha\colon \mathbb{R}^d \to \mathbb{R}$, $\alpha \in [0, \infty)$, of optimal solutions for (38) is continuous with respect to the uniform norm on any compact set $K \subset \mathbb{R}^d$. Further, for $\alpha \to \infty$, we have that $f^*_\alpha \to f^*_\infty$ pointwise.*

Proof Let $\alpha_k \to \alpha \in [0, \infty)$ with $\alpha_k \neq 0$. From (63) and the weak convergence of the $\hat{\mu}^*_{\alpha_k}$, we get that the sequence $f^*_{\alpha_k}$ is pointwise-convergent. Further, recall that all $f^*_{\alpha_k}$ are Lipschitz-continuous with constant $C\hat{\mu}^*_{\alpha_k}(\mathbb{S}^{p-1})$. Since weakly convergent sequences have bounded measures, the $f^*_{\alpha_k}$ are uniformly Lipschitz-continuous. Hence, we conclude that $\|f^*_{\alpha_k} - f^*_\alpha\|_{C(K)} \to 0$ for any compact set $K \subset \mathbb{R}^d$. For the case $\alpha_k \to \infty$, we have already shown in the proof of Proposition 12 that the sequence $f^*_{\alpha_k}$ converges pointwise to f^*_∞. ■

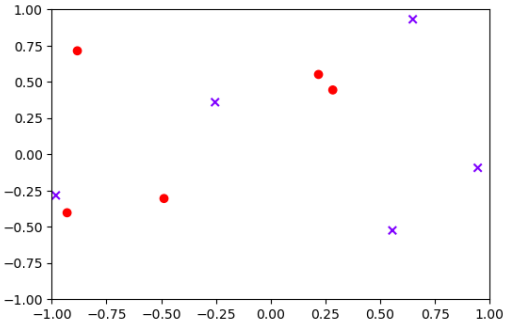

Figure 1: Data for Problem (66) (crosses corresponds to $y_k = -1$, circles to $y_k = 1$).

4. Path Comparison at Final States

To illustrate our theoretical observations, we investigate a 2D interpolation problem with samples $(\tilde{\mathbf{x}}_k, y_k) \in [-1,1]^2 \times \{-1,1\}$, $k = 1, \dots, 10$, which are depicted in Figure 1. Note that we have chosen to investigate an interpolation problem as it describes the end states of both the static and the dynamic paths. As discussed in Remark 1, we modify the x-component of these samples to $\mathbf{x}_k = (\tilde{\mathbf{x}}_k, 1) \in \mathbb{R}^3$ in order to use a 2-homogeneous infinite-width 2-layer ReLU NN model with parameterization function $\phi_{\mathrm{ReLU}} \colon \mathbb{R}^4 \times \mathbb{R}^3 \to \mathbb{R}$ given by

$$\phi_{\mathrm{ReLU}}(\mathbf{w}, \mathbf{x}) = w_1 \mathrm{ReLU}(\mathbf{w}_2^T \mathbf{x}), \tag{65}$$

where $\mathbf{w} = (w_1, \mathbf{w}_2) \in \mathbb{R} \times \mathbb{R}^3$. Now, our goal is to find a probability measure $\mu \in \mathcal{P}_2(\mathbb{R}^4)$ such that

$$\int_{\mathbb{R}^4} \phi_{\mathrm{ReLU}}(\mathbf{w}, \mathbf{x}_k) \, \mathrm{d}\mu(\mathbf{w}) = y_k, \qquad k = 1, \dots, n. \tag{66}$$

By the use of atomic measures, any finite-width 2-layer ReLU NN $\Psi(\mathbf{x}) = \sum_{k=1}^n w_k (\mathbf{v}_k^T \mathbf{x})^+$ with $w_k \in \mathbb{R}$ and $\mathbf{v}_k \in \mathbb{R}^3$ is covered by this formulation. Further, (66) can be recast as the search for a measure $\hat{\mu} \in \mathcal{M}^+(\mathbb{S}^3)$ (see (18)). But, even then, (66) is in general underdetermined and we need to employ some kind of explicit or implicit regularization in order to ensure *nice* solutions.

4.1 Scaling Path

First, we look into the solution of the variational problem (43) which, for the described interpolation setting, reads

$$\operatorname*{arg\,min}_{\hat{\mu} \in \mathcal{M}^+(\mathbb{S}^3)} \widehat{W}_2^2\big(\hat{\mu}, \alpha^2 \hat{\mu}_0\big) \quad \text{s.t.} \quad \int_{\mathbb{S}^3} \phi_{\mathrm{ReLU}}(\boldsymbol{\theta}, \mathbf{x}_k) \, \mathrm{d}\hat{\mu}(\boldsymbol{\theta}) = y_k, \quad k = 1, \dots, 10, \tag{67}$$

with $\boldsymbol{\theta} = (\theta_1, \boldsymbol{\theta}_2) \in \mathbb{S}^3 \subset \mathbb{R}^4$. The initialization $\hat{\mu}_0$ is chosen as the uniform measure on $P = \{\pm 1/\sqrt{2}\} \times \mathbb{S}^2/\sqrt{2} \subset \mathbb{S}^3$. The choice of a uniform measure on P instead of $\mathbb{S}^3$ is motivated, on the one hand, by the training dynamics and, on the other hand, by the initialization of the dynamic viewpoint based on the gradient flow (1) investigated in Section 4.2. To

compute $\widehat{W}_2^2$, we make use of the formulation (21). More precisely, this corresponds to the unbalanced optimal transport

$$\widehat{W}_2(\hat{\mu}_1, \hat{\mu}_2)^2 = \min_{\gamma \in \mathcal{M}^+(\mathbb{S}^3 \times \mathbb{S}^3)} \int_{\mathbb{S}^3 \times \mathbb{S}^3} c \, \mathrm{d}\gamma + \sum_{i=1}^{2} \mathrm{KL}(\pi_{i\sharp}\gamma, \hat{\mu}_i), \tag{68}$$

where

$$c(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \begin{cases} -2\log(\boldsymbol{\theta}_1^T \boldsymbol{\theta}_2) & \text{if } \boldsymbol{\theta}_1^T \boldsymbol{\theta}_2 > 0, \\ \infty & \text{else.} \end{cases} \tag{69}$$

Problem (67) is an infinite-dimensional convex-optimization problem. To make it computationally tractable, we discretize the spheres $\mathbb{S}^2$ in P using a Fibonacci grid $\mathbf{F}$ with 250^2 points (Swinbank and James Purser, 2006). The corresponding discrete version of P and the measure $\hat{\mu}_0$ are denoted by $\mathbf{P}$ and $\hat{\boldsymbol{\mu}}_0$, respectively. Additionally, we discretize the search space for μ based on the Fibonacci grid $\mathbf{F}$ as

$$\mathbf{Q} = \left\{ \frac{1}{7\sqrt{2}} \left(\pm k, \left(98 - k^2\right)^{1/2} \mathbf{x} \right) : 1 \leq k \leq 9, \mathbf{x} \in \mathbf{F} \right\} \subset \mathbb{S}^3. \tag{70}$$

The coarser discretization in the first coordinate of the set $\mathbf{Q}$ is motivated by the fact that $\phi_{\mathrm{ReLU}}((w_1, \mathbf{w}_2), \mathbf{x}) = \phi_{\mathrm{ReLU}}((w_1/|w_1|, \mathbf{w}_2|w_1|), \mathbf{x})$, namely, that the model is considerably over-parameterized. This choice also ensures that $\mathbf{P} \subset \mathbf{Q}$. Now, we obtain a discrete convex problem involving the unbalanced optimal-transport distance $\widehat{W}_2$, which is still computationally challenging due to its large size. Therefore, we resort to an entropy-regularized distance $\widehat{W}_{2,\epsilon}$ (see (Feydy et al., 2019)) instead of the original formulation (68). The divergence $\widehat{W}_{2,\epsilon}$ can be computed efficiently through the Sinkhorn algorithm, and its' gradients can be computed using algorithmic differentiation. For small regularization parameters such as $\epsilon = 1 \cdot 10^{-2}$, the approximation $\widehat{W}_{2,\epsilon}$ is reasonably close to the original $\widehat{W}_2$ distance (Feydy et al., 2019; Neumayer and Steidl, 2021). Finally, we arrive at the fully discrete problem

$$\operatorname*{arg\,min}_{\hat{\boldsymbol{\mu}} \in \mathcal{M}^+(\mathbf{Q})} \widehat{W}_{2,\epsilon}^2\left(\hat{\boldsymbol{\mu}}, \alpha^2 \hat{\boldsymbol{\mu}}_0\right) \quad \text{s.t.} \quad \int_{\mathbf{Q}} \phi_{\mathrm{ReLU}}(\boldsymbol{\theta}, \mathbf{x}_k) \, \mathrm{d}\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}) = y_k, \quad k = 1, \ldots, 10, \tag{71}$$

which amounts to the minimization of a differentiable convex objective subject to linear equality constraints. Such problems can be solved, for example, with the forward-backward splitting (Combettes and Wajs, 2005). To ensure fast convergence, we couple this method with a spectral step-size predictor and an Armijo linesearch to ensure convergence as detailed in (Goldstein et al., 2014). To evaluate $\widehat{W}_{2,\epsilon}^2\left(\cdot, \alpha^2 \hat{\boldsymbol{\mu}}_0\right)$ and its gradients, we make use of the *geomloss* package[2]. Our numerical results for various values of α (including the limiting cases $\alpha = 0$ and $\alpha = \infty$) are depicted in Figure 2. We clearly observe that a larger regularization scale α leads to smoother solutions. Additionally, we observe that the f_α^* converge visually for $\alpha \to 0$ and $\alpha \to \infty$, as predicted by Corollary 13. The corresponding functional values multiplied by the correct scaling $1 + \alpha^2$ can be found in Table 2. For the NTK setting, the optimal value corresponding to (50) is $2.83 \cdot 10^2$. Again, we observe convergence of $(1 + \alpha^2)\widehat{W}_{2,\epsilon}^2\left(\hat{\boldsymbol{\mu}}_\alpha^*, \alpha^2 \hat{\boldsymbol{\mu}}_0\right)$, as predicted by Propositions 11 and 12, and Theorem 10.

2. https://www.kernel-operations.io/geomloss/

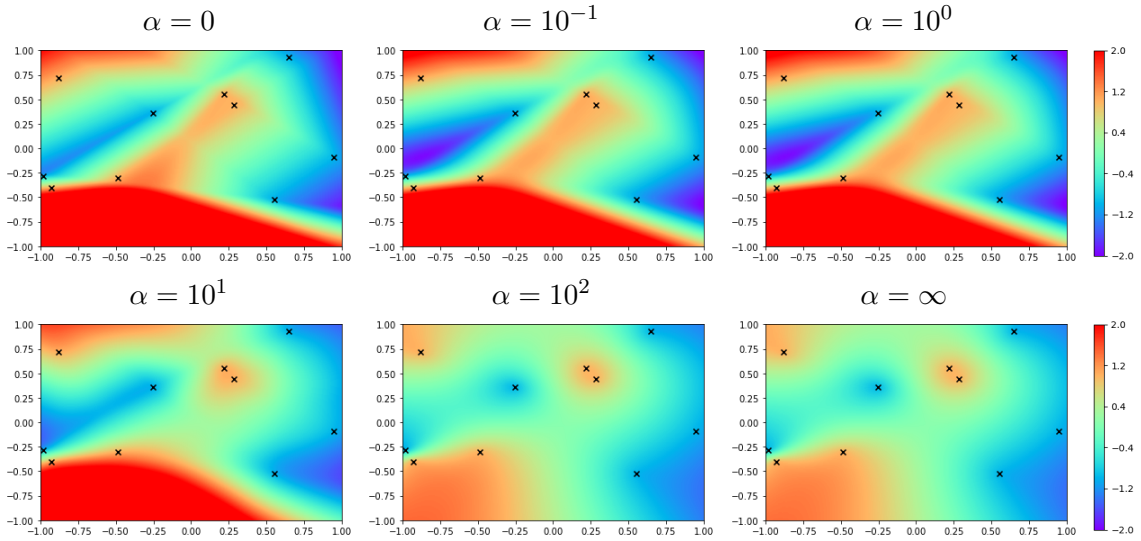

Figure 2: Solutions of (71) for several values of α. The range is clipped to $[-2, 2]$.

Remark 14 *In principal,* (66) *is still over-parameterized, even in the form* (18)*. Essentially, it suffices to consider* $\mathcal{M}^+(\{\pm 1/\sqrt{2}\} \times \mathbb{S}^2/\sqrt{2}) \subset \mathcal{M}^+(\mathbb{S}^3)$ *in* (18) *to realize any NN. This has the advantage that we only need to optimize over two 2D measures instead of a 3D one, which considerably reduces the computation time. Unfortunately, there is no theoretical guarantee that the optimal measures* $\hat{\boldsymbol{\mu}}_\alpha^*$ *must be supported on* $\mathbf{P}$*. However, we observed numerically that the assumption that* $\operatorname{supp}(\hat{\boldsymbol{\mu}}_\alpha^*) \subset P$ *leads essentially to the same results (Figure 2 and Figure 3). Therefore, we propose to replace* $\mathcal{M}^+(\mathbf{Q})$ *by* $\mathcal{M}^+(\mathbf{P})$ *in* (71) *to decrease the computational cost.*

4.2 Dynamic Viewpoint Based on Gradient Descent

Next, we illustrate the implicit regularizing effect of gradient descent training for the loss

$$\sum_{k=1}^{10} \left| \frac{1}{2 \cdot 250^2} \sum_{l=1}^{2 \cdot 250^2} \beta^2 \phi_{\mathrm{ReLU}}(\mathbf{w}_l, \mathbf{x}_k) - y_k \right|^2, \tag{72}$$

with $\mathbf{w}_l = (w_{1,l}, \mathbf{w}_{2,l}) \in \mathbb{R}^4$. To make a link with our approach in Section 4.1, the $\mathbf{w}_l$ are initialized as the points from $\mathbf{P}$. Depending on the initialization scale β in (72), gradient-descent training leads to very different results, as discussed in (Chizat et al., 2019; Woodworth et al., 2020). For all parameters β, we have chosen a sufficiently small stepsize and iterated gradient descent until convergence. The obtained empirical measure corresponding to the scale β is denoted by $\hat{\boldsymbol{\nu}}_\beta^*$.

A natural question is to investigate how the solutions induced by $\hat{\boldsymbol{\nu}}_\beta^*$ compare to the ones induced by $\hat{\boldsymbol{\mu}}_\alpha^*$. A visual comparison is provided in Figure 3. For larger values of α and β, the solutions corresponding to the same values are very similar. As predicted by our theory, the solutions induced by $\hat{\boldsymbol{\mu}}_\alpha^*$ indeed approach f_∞^* associated to the kernel formulation (50) for $\alpha \to \infty$. The same behavior was predicted for the solutions corresponding

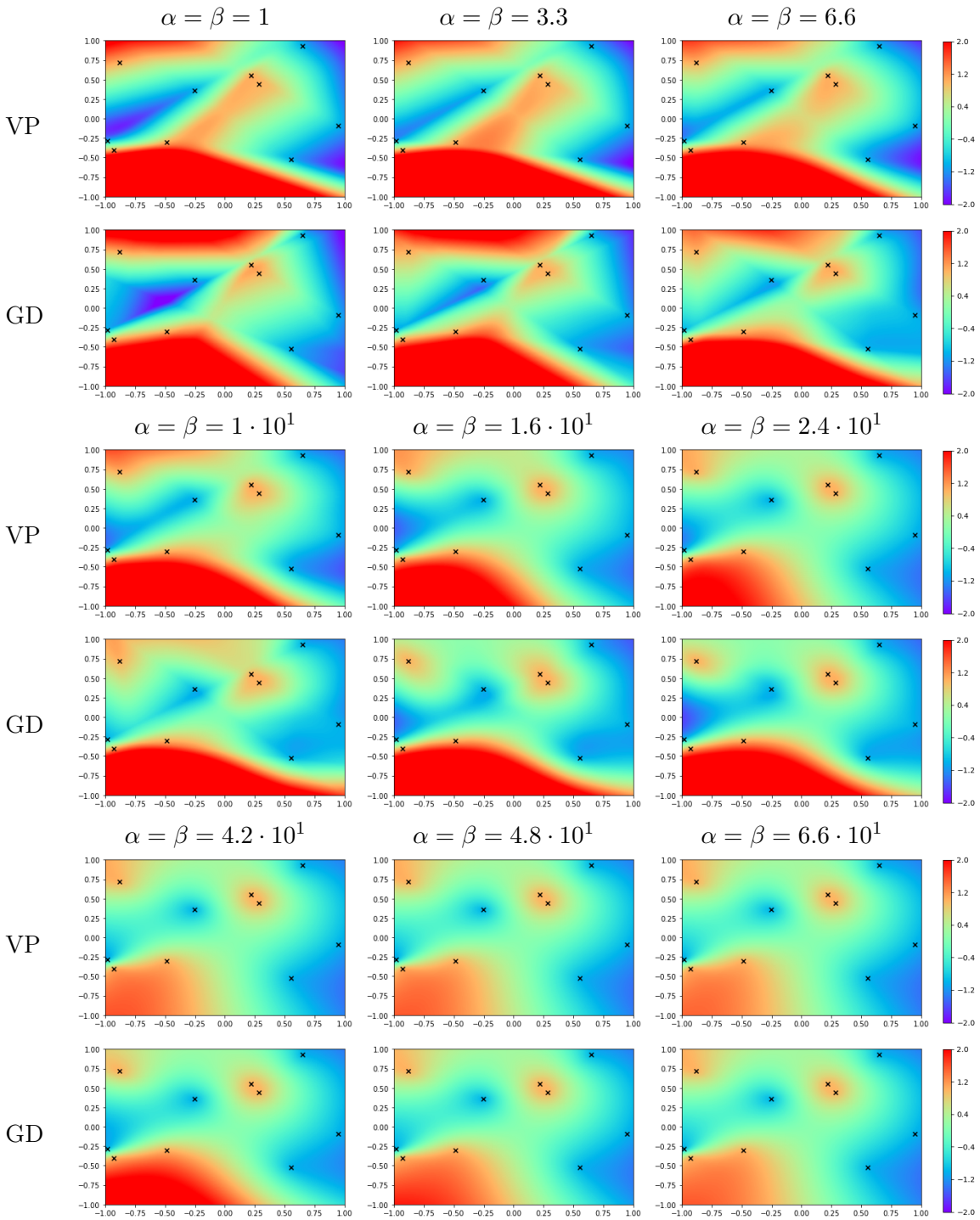

Figure 3: Optimization paths for gradient-descent training (GD) and the solutions of (71) (VP) restricted to $\mathcal{M}^+(\mathbf{P})$. The plots are clipped to $[-2, 2]$.

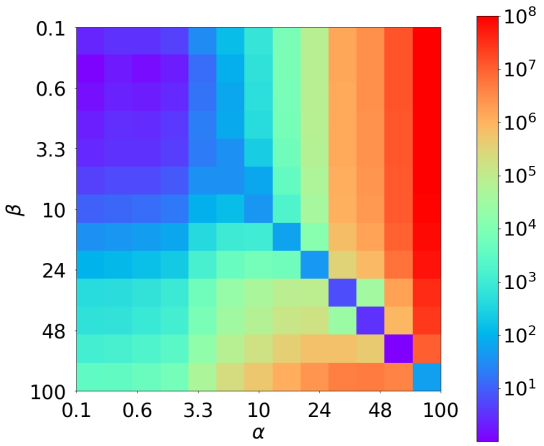

Figure 4: Heat map of $\widehat{W}^2_{2,\epsilon}(\hat{\boldsymbol{\nu}}^*_\beta, \alpha^2\hat{\boldsymbol{\mu}}_0) - \widehat{W}^2_{2,\epsilon}(\hat{\boldsymbol{\mu}}^*_\alpha, \alpha^2\hat{\boldsymbol{\mu}}_0)$.

to $\hat{\boldsymbol{\nu}}^*_\beta$ in (Chizat et al., 2019). Although the solutions start to differ for decreasing values of α and β, the path itself remains similar. The path becomes significantly different only for small values of α and β. However, for increasing width, the limits $\alpha \to 0$ and $\beta \to 0$ both lead to solutions of (48). Aside from this visual analysis, we can also examine the values of $\widehat{W}^2_{2,\epsilon}(\hat{\boldsymbol{\nu}}^*_\beta, \alpha^2\hat{\boldsymbol{\mu}}_0) - \widehat{W}^2_{2,\epsilon}(\hat{\boldsymbol{\mu}}^*_\alpha, \alpha^2\hat{\boldsymbol{\mu}}_0)$. A heat map is provided in Figure 4, and the exact values are given in Table 2. Although not necessarily contained in the optimization domain $\mathcal{M}^+(\mathbf{P})$ of (71), the gradient-descent-based solutions $\hat{\boldsymbol{\nu}}^*_\alpha$ usually have higher functional values than their variational counterparts $\hat{\boldsymbol{\mu}}^*_\alpha$. Moreover, the minimal value of $\widehat{W}^2_{2,\epsilon}(\hat{\boldsymbol{\nu}}^*_\beta, \alpha^2\hat{\boldsymbol{\mu}}_0) - \widehat{W}^2_{2,\epsilon}(\hat{\boldsymbol{\mu}}^*_\alpha, \alpha^2\hat{\boldsymbol{\mu}}_0)$ for large and fixed α is always obtained for $\beta = \alpha$. For smaller initialization scales α, the values are very close and $\beta = \alpha$ is close to being optimal.

5. Conclusions

In this paper, we have introduced the scaling path of a neural network. It involves the Hellinger–Kantorovich distance (a.k.a. Wasserstein–Fisher–Rao distance) and depends on an initialization scale. As main contribution, we have shown that the solutions of these paths depend continuously on the initialization scale, which makes them well-behaved objects amendable to further theoretical analyses. The relevance of the scaling path is demonstrated by a small-scale numerical example, in which we observed that the scaling path can be indeed qualitatively related to the training dynamics of gradient descent at large times, namely, the endpoint of the optimization path.

Acknowledgments

The research leading to these results was supported by the European Research Council (ERC) under European Union's Horizon 2020 (H2020), Grant Agreement - Project No 101020573 FunLearn.

α	$1 \cdot 10^{-1}$	$3.3 \cdot 10^{-1}$	$6.6 \cdot 10^{-1}$	1	3.3	6.6
Min. of (71)	$7.50 \cdot 10^{1}$	$7.44 \cdot 10^{1}$	$9.64 \cdot 10^{1}$	$1.26 \cdot 10^{2}$	$4.47 \cdot 10^{2}$	$7.73 \cdot 10^{2}$
$\beta = 1 \cdot 10^{-1}$	$\mathbf{9.36 \cdot 10^{1}}$	$9.87 \cdot 10^{1}$	$1.20 \cdot 10^{2}$	$1.58 \cdot 10^{2}$	$6.19 \cdot 10^{2}$	$1.41 \cdot 10^{3}$
$\beta = 3.3 \cdot 10^{-1}$	$8.44 \cdot 10^{1}$	$\mathbf{8.88 \cdot 10^{1}}$	$1.08 \cdot 10^{2}$	$1.41 \cdot 10^{2}$	$5.33 \cdot 10^{2}$	$1.17 \cdot 10^{3}$
$\beta = 6.6 \cdot 10^{-1}$	$8.73 \cdot 10^{1}$	$9.18 \cdot 10^{1}$	$\mathbf{1.11 \cdot 10^{2}}$	$1.45 \cdot 10^{2}$	$5.43 \cdot 10^{2}$	$1.12 \cdot 10^{3}$
$\beta = 1$	$9.18 \cdot 10^{1}$	$9.66 \cdot 10^{1}$	$1.17 \cdot 10^{2}$	$\mathbf{1.53 \cdot 10^{2}}$	$5.70 \cdot 10^{2}$	$1.12 \cdot 10^{3}$
$\beta = 3.3$	$9.45 \cdot 10^{1}$	$9.93 \cdot 10^{1}$	$1.20 \cdot 10^{2}$	$1.57 \cdot 10^{2}$	$\mathbf{5.66 \cdot 10^{2}}$	$9.68 \cdot 10^{2}$
$\beta = 6.6$	$1.07 \cdot 10^{2}$	$1.13 \cdot 10^{2}$	$1.37 \cdot 10^{2}$	$1.79 \cdot 10^{2}$	$6.43 \cdot 10^{2}$	$\mathbf{9.60 \cdot 10^{2}}$
$\beta = 1 \cdot 10^{1}$	$1.39 \cdot 10^{2}$	$1.47 \cdot 10^{2}$	$1.80 \cdot 10^{2}$	$2.37 \cdot 10^{2}$	$8.95 \cdot 10^{2}$	$1.41 \cdot 10^{3}$
$\beta = 1.6 \cdot 10^{1}$	$2.70 \cdot 10^{2}$	$2.88 \cdot 10^{2}$	$3.57 \cdot 10^{2}$	$4.77 \cdot 10^{2}$	$2.05 \cdot 10^{3}$	$4.36 \cdot 10^{3}$
$\beta = 2.4 \cdot 10^{1}$	$5.82 \cdot 10^{2}$	$6.27 \cdot 10^{2}$	$7.89 \cdot 10^{2}$	$1.06 \cdot 10^{3}$	$5.14 \cdot 10^{3}$	$1.36 \cdot 10^{4}$
$\beta = 4.2 \cdot 10^{1}$	$1.77 \cdot 10^{3}$	$1.92 \cdot 10^{3}$	$2.45 \cdot 10^{3}$	$3.36 \cdot 10^{3}$	$1.77 \cdot 10^{4}$	$5.57 \cdot 10^{4}$
$\beta = 4.8 \cdot 10^{1}$	$2.31 \cdot 10^{3}$	$2.51 \cdot 10^{3}$	$3.21 \cdot 10^{3}$	$4.41 \cdot 10^{3}$	$2.37 \cdot 10^{4}$	$7.62 \cdot 10^{4}$
$\beta = 6.6 \cdot 10^{1}$	$4.38 \cdot 10^{3}$	$4.78 \cdot 10^{3}$	$6.12 \cdot 10^{3}$	$8.44 \cdot 10^{3}$	$4.66 \cdot 10^{4}$	$1.56 \cdot 10^{5}$

α	$1 \cdot 10^{1}$	$1.6 \cdot 10^{1}$	$2.4 \cdot 10^{1}$	$4.2 \cdot 10^{1}$	$4.8 \cdot 10^{1}$	$6.6 \cdot 10^{1}$
Min. of (71)	$7.27 \cdot 10^{2}$	$4.60 \cdot 10^{2}$	$3.38 \cdot 10^{2}$	$2.88 \cdot 10^{2}$	$2.84 \cdot 10^{2}$	$2.79 \cdot 10^{2}$
$\beta = 1 \cdot 10^{-1}$	$3.39 \cdot 10^{3}$	$2.39 \cdot 10^{4}$	$1.63 \cdot 10^{5}$	$2.08 \cdot 10^{6}$	$3.74 \cdot 10^{6}$	$1.47 \cdot 10^{7}$
$\beta = 3.3 \cdot 10^{-1}$	$3.03 \cdot 10^{3}$	$2.38 \cdot 10^{4}$	$1.66 \cdot 10^{5}$	$2.10 \cdot 10^{6}$	$3.74 \cdot 10^{6}$	$1.48 \cdot 10^{7}$
$\beta = 6.6 \cdot 10^{-1}$	$2.69 \cdot 10^{3}$	$2.20 \cdot 10^{4}$	$1.59 \cdot 10^{5}$	$2.06 \cdot 10^{6}$	$3.71 \cdot 10^{6}$	$1.46 \cdot 10^{7}$
$\beta = 1$	$2.47 \cdot 10^{3}$	$2.04 \cdot 10^{4}$	$1.52 \cdot 10^{5}$	$2.02 \cdot 10^{6}$	$3.65 \cdot 10^{6}$	$1.45 \cdot 10^{7}$
$\beta = 3.3$	$1.79 \cdot 10^{3}$	$1.72 \cdot 10^{4}$	$1.40 \cdot 10^{5}$	$1.95 \cdot 10^{6}$	$3.55 \cdot 10^{6}$	$1.42 \cdot 10^{7}$
$\beta = 6.6$	$1.07 \cdot 10^{3}$	$1.22 \cdot 10^{4}$	$1.20 \cdot 10^{5}$	$1.83 \cdot 10^{6}$	$3.36 \cdot 10^{6}$	$1.37 \cdot 10^{7}$
$\beta = 1 \cdot 10^{1}$	$\mathbf{9.50 \cdot 10^{2}}$	$6.78 \cdot 10^{3}$	$9.27 \cdot 10^{4}$	$1.64 \cdot 10^{6}$	$3.06 \cdot 10^{6}$	$1.29 \cdot 10^{7}$
$\beta = 1.6 \cdot 10^{1}$	$4.37 \cdot 10^{3}$	$\mathbf{7.66 \cdot 10^{2}}$	$3.51 \cdot 10^{4}$	$1.16 \cdot 10^{6}$	$2.30 \cdot 10^{6}$	$1.07 \cdot 10^{7}$
$\beta = 2.4 \cdot 10^{1}$	$2.01 \cdot 10^{4}$	$1.70 \cdot 10^{4}$	$\mathbf{5.74 \cdot 10^{2}}$	$5.65 \cdot 10^{5}$	$1.31 \cdot 10^{6}$	$7.63 \cdot 10^{6}$
$\beta = 4.2 \cdot 10^{1}$	$1.03 \cdot 10^{5}$	$1.73 \cdot 10^{5}$	$1.86 \cdot 10^{5}$	$\mathbf{3.31 \cdot 10^{2}}$	$8.28 \cdot 10^{4}$	$2.50 \cdot 10^{6}$
$\beta = 4.8 \cdot 10^{1}$	$1.45 \cdot 10^{5}$	$2.62 \cdot 10^{5}$	$3.31 \cdot 10^{5}$	$6.36 \cdot 10^{4}$	$\mathbf{3.08 \cdot 10^{2}}$	$1.40 \cdot 10^{6}$
$\beta = 6.6 \cdot 10^{1}$	$3.16 \cdot 10^{5}$	$6.40 \cdot 10^{5}$	$1.01 \cdot 10^{6}$	$1.01 \cdot 10^{6}$	$7.44 \cdot 10^{5}$	$\mathbf{2.87 \cdot 10^{2}}$

Table 2: Values of $(1+\alpha^2)\widehat{W}_{2,\epsilon}^2(\hat{\boldsymbol{\nu}}_\beta^*, \alpha^2\hat{\boldsymbol{\mu}}_0)$ in terms of the scale β of gradient-descent training. The diagonal $\alpha = \beta$ is highlighted in bold.

References

Alnur Ali, Edgar Dobriban, and Ryan Tibshirani. The implicit regularization of stochastic gradient flow for least squares. In *International Conference on Machine Learning*, pages 233–244. PMLR, 2020.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Birkhäuser, Basel, 2005.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR, 2019.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, Boston, MA, 2004.

Alberto Bietti and Julien Mairal. On the inductive bias of neural tangent kernels. In *Advances in Neural Information Processing Systems*, volume 32, pages 12556–12567, 2019.

Etienne Boursier, Loucas Pillaud-Vivien, and Nicolas Flammarion. Gradient flow dynamics of shallow ReLU networks for square loss and orthogonal inputs. In *Advances in Neural Information Processing Systems*, volume 35, pages 20105–20118, 2022.

A. Braides. Γ-*Convergence for Beginners*. Oxford University Press, Oxford, 2002.

Zhengdao Chen, Eric Vanden-Eijnden, and Joan Bruna. A functional-space mean-field theory of partially-trained three-layer neural networks. *ArXiv:2210.16286*, 2022.

Lénaïc Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Proceedings of Machine Learning Research*, volume 125, pages 1305–1338. PMLR, 2020.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. An interpolating distance between optimal transport and Fisher–Rao metrics. *Foundations of Computational Mathematics*, 18:1–44, 2018.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 32, pages 2933–2943, 2019.

P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4:1168–1200, 2005.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shunichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *Proceedings of Machine Learning Research*, volume 89, pages 2681–2690. PMLR, 2019.

Tom Goldstein, Christoph Studer, and Richard Baraniuk. A field guide to forward-backward splitting with a FASTA implementation. *arXiv preprint arXiv:1411.3406*, 2014.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, pages 8580–8589, 2018.

Stanislav Kondratyev, Léonard Monsaingeon, and Dmitry Vorotnikov. A new optimal transport distance on the space of finite Radon measures. *Advances in Differential Equations*, 21(11/12):1117–1164, 2016.

Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and dynamics of stochastic gradient algorithms I: Mathematical foundations. *Journal of Machine Learning Research*, 20(1):1474–1520, 2019.

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal transport in competition with reaction: the Hellinger-Kantorovich distance and geodesic curves. *SIAM Journal on Mathematical Analysis*, 48(4):2869–2911, 2016.

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new Hellinger-Kantorovich distance between positive measures. *Inventiones Mathematicae*, 211(3):969–1117, 2018.

Kaifeng Lyu, Zhiyuan Li, Runzhe Wang, and Sanjeev Arora. Gradient descent on two-layer nets: Margin maximization and simplicity bias. In *Advances in Neural Information Processing Systems*, volume 34, pages 12978–12991, 2021.

Sebastian Neumayer and Gabriele Steidl. From optimal transport to discrepancy. In *Handbook of Mathematical Models and Algorithms in Computer Vision and Imaging*. Springer, 2021.

Sebastian Neumayer and Michael Unser. Explicit representations for Banach subspaces of Lizorkin distributions. *Analysis and Applications*, 21(5):1223–1250, 2023.

Rahul Parhi and Robert D. Nowak. Banach space representer theorems for neural networks and ridge splines. *Journal of Machine Learning Research*, 22(43):1–40, 2021.

Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of SGD for diagonal linear networks: a provable benefit of stochasticity. In *Advances in Neural Information Processing Systems*, volume 34, pages 29218–29230, 2021.

Noam Razin and Nadav Cohen. Implicit regularization in deep learning may not be explainable by norms. In *Advances in Neural Information Processing Systems*, volume 33, pages 21174–21187, 2020.

Saharon Rosset, Grzegorz Swirszcz, Nathan Srebro, and Ji Zhu. ℓ_1 regularization in infinite dimensional feature spaces. In *Conference on Learning Theory*, pages 544–558. Springer, 2007.

Weijie Su, Stephen Boyd, and Emmanuel Candés. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*, volume 27, pages 2510–2518, 2014.

Arun Suggala, Adarsh Prasad, and Pradeep K. Ravikumar. Connecting optimization and regularization paths. In *Advances in Neural Information Processing Systems*, volume 31, pages 10631–10641, 2018.

Richard Swinbank and R. James Purser. Fibonacci grids: A novel approach to global modelling. *Quarterly Journal of the Royal Meteorological Society*, 132(619):1769–1793, 2006.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der Mathematischen Wissenschaften*. Springer-Verlag, Berlin, 2009.

Yuqing Wang, Minshuo Chen, Tuo Zhao, and Molei Tao. Large learning rate tames homogeneity: Convergence and balancing effect. In *International Conference on Learning Representations*, 2022.

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Proceedings of Machine Learning Research*, volume 135, pages 3635–3673. PMLR, 2020.