Implicit Bias of Gradient Descent for Wide Two-layer Neural Networks Trained with the Logistic Loss

Lénaïc Chizat   LENAIC.CHIZAT@UNIVERSITE-PARIS-SACLAY.FR
Laboratoire de Mathématiques d'Orsay, CNRS, Université Paris-Saclay, France

Francis Bach   FRANCIS.BACH@INRIA.FR
INRIA, ENS, PSL Research University, Paris, France

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

Neural networks trained to minimize the logistic (a.k.a. cross-entropy) loss with gradient-based methods are observed to perform well in many supervised classification tasks. Towards understanding this phenomenon, we analyze the training and generalization behavior of infinitely wide two-layer neural networks with homogeneous activations. We show that the limits of the gradient flow on exponentially tailed losses can be fully characterized as a max-margin classifier in a certain non-Hilbertian space of functions. In presence of hidden low-dimensional structures, the resulting margin is independent of the ambient dimension, which leads to strong generalization bounds. In contrast, training only the output layer implicitly solves a kernel support vector machine, which a priori does not enjoy such an adaptivity. Our analysis of training is non-quantitative in terms of running time but we prove computational guarantees in simplified settings by showing equivalences with online mirror descent. Finally, numerical experiments suggest that our analysis describes well the practical behavior of two-layer neural networks with ReLU activations and confirm the statistical benefits of this implicit bias.

1. Introduction

Artificial neural networks are successfully used in a variety of difficult supervised classification tasks, but the mechanisms behind their performance remain unclear. The situation is particularly intriguing when the number of parameters of these models exceeds by far the number of input data points and they are trained with gradient-based methods until zero training error, without any explicit regularization. In this case, the training algorithm induces an *implicit bias*: among the many classifiers which overfit on the training set, it selects a specific one which often turns out to perform well on the test set. In this paper, we study the implicit bias of wide neural networks with two layers (i.e., with a single hidden-layer) trained with gradient descent on the logistic loss, or any loss with an exponential tail. Our analysis lies at the intersection of two lines of research that study (i) the implicit bias of gradient methods, and (ii) the training dynamics of wide neural networks.

Implicit bias of gradient methods. Soudry et al. (2018) show that for linearly separable data, training a linear classifier with gradient descent on the logistic loss, or any loss with an exponential tail, implicitly leads to a max-margin linear classifier for the ℓ_2-norm. This result together with results in the boosting literature (Telgarsky, 2013) have led to a fruitful line of research. Fine analyses of convergence rates have been carried out by Nacson et al. (2019b); Ji and Telgarsky (2019b, 2018), and extensions to other gradient-based algorithms and to factored parameterizations

are considered by Gunasekar et al. (2018a). Linear neural networks have been studied by Gunasekar et al. (2018b); Ji and Telgarsky (2019a); Nacson et al. (2019a), and some properties in general non-convex cases are given by Xu et al. (2018). Closer to the present paper, Lyu and Li (2019) show that for homogeneous neural networks the training trajectory converges in direction to a critical point of some nonconvex max-margin problem. In the present work, we improve this result for the two-layer case: we characterize the learnt classifier as the solution of a *convex* max-margin problem. Importantly, this characterization is precise enough to enable a statistical analysis (see Section 6).

Dynamics of infinitely-wide neural networks. This fine characterization is made possible by looking at the infinite width limit of two-layer neural networks. This strategy has been used in several works to obtain insights on their statistical properties (Bengio et al., 2006; Bach, 2017a) or training behavior (Nitanda and Suzuki, 2017; Rotskoff and Vanden-Eijnden, 2018; Chizat and Bach, 2018; Mei et al., 2018; Sirignano and Spiliopoulos, 2019), which can be described by a Wasserstein gradient flow (Ambrosio et al., 2008). In particular, Chizat and Bach (2018) show that if the loss is convex, if the initialization is "diverse enough", and if the gradient flow of the objective converges, then its limit is a global minimizer. This result does not apply in our context because the gradient flow diverges, which turns out to be beneficial for the analysis of the implicit bias that we propose.

A general drawback of those *mean-field* analyses is that they are mostly non-quantitative, both in terms of number of neurons and number of iterations. While some works have shown quantitative results by modifying the dynamics (Mei et al., 2019; Wei et al., 2019; Chizat, 2019), we do not take this path in order to stay close to the way neural networks are used in practice and because our numerical experiments suggest that those modifications are not necessary to obtain a good practical behavior. Finally, we stress that our analysis does not take place in the *lazy training* regime (Chizat et al., 2019) which consists of training dynamics that can be analyzed in a perturbative regime around the initialization (see, e.g., Li and Liang, 2018; Jacot et al., 2018; Du et al., 2019). Lazy training is another kind of implicit bias that amounts to training a linear model and does not lead to adaptivity results as those shown in Section 6 (see Figure 3 for an illustration in our context).

1.1. Organization and contributions

After preliminaries on wide neural networks in Section 2, we make the following contributions :

- In Section 3, we show that for a class of two-layer neural networks and for losses with an exponential tail, the classifier learnt by the non-convex gradient flow is a max-margin classifier for a certain functional norm known as the variation norm.

- When fixing the "directions" of the neurons (Section 4), or when only training the output layer (Section 5), we show that the dynamics implicitly performs *online mirror ascent* on a sequence of smooth-margin objectives and thus naturally maximizes the margin. This leads to convergence guarantees in $O(\log(t)/\sqrt{t})$ in situations where no rate was previously known.

- In Section 6, we study the margins of those classifiers and prove dimension-independent generalization bounds for classification in presence of hidden linear structures.

- We perform numerical experiments in Section 7 for two-layer ReLU neural networks which confirm the statistical efficiency of this implicit bias in a high-dimensional setting.

In summary, we show that training two-layer ReLU neural networks implicitly solves a problem with strong statistical benefits. We stress however that the runtime of the algorithm is still unknown.

1.2. Notation

We denote by $\mathcal{M}(\mathbb{R}^p)$ (resp. $\mathcal{M}_+(\mathbb{R}^p)$) the set of signed (resp. nonnegative) finite Borel measures on $\mathbb{R}^p$ and by $\mathcal{P}(\mathbb{R}^p)$ (resp. $\mathcal{P}_2(\mathbb{R}^p)$) the set of probability measures (resp. with finite second moment). The set $\Delta^{m-1} = \{p \in \mathbb{R}^m_+ \; ; \; \mathbf{1}^\top p = 1\}$ is the simplex.

2. Preliminaries on infinitely wide two-layer networks

2.1. 2-homogeneous neural networks

We consider a binary classification problem with a training set $(x_i, y_i)_{i \in [n]}$ of n pairs of observations with $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$ and prediction functions of the form

$$h_m(\mathbf{w}, x) = \frac{1}{m} \sum_{j=1}^m \phi(w_j, x), \tag{1}$$

where $m \geq 1$ is the number of units and $\mathbf{w} = (w_j)_{j \in [m]} \in (\mathbb{R}^p)^m$ are the trainable parameters. This setting covers two-layer neural networks where m is the size of the hidden layer. In this paper, we are interested in the over-parameterized regime where m is large, and the prefactor $1/m$ is needed to obtain a non-degenerate limit. We refer to ϕ as a *feature function*, and we focus on the case where ϕ is 2-*homogeneous* and *balanced*:

(A1) The function ϕ is (positively) 2-*homogeneous* in its first variable, i.e., $\phi(rw, x) = r^2 \phi(w, x)$ for all $(r, w, x) \in \mathbb{R}_+ \times \mathbb{R}^p \times \mathbb{R}^d$ and it is *balanced*, which means that there is a map $T : \mathbb{S}^{p-1} \to \mathbb{S}^{p-1}$ such that for all $\theta \in \mathbb{S}^{p-1}$, $\phi(T(\theta), \cdot) = -\phi(\theta, \cdot)$.

Here are examples of models which satisfy (A1):

- *ReLU networks.* A two-layer neural network with the rectified linear unit (ReLU) activation function is obtained by setting $\phi(w, x) = b(a^\top(x, 1))_+$ where $w = (a, b) \in \mathbb{R}^{d+2}$. This is the motivating example of this article. It is not differentiable but is covered by Theorem 5.

- *S-ReLU networks.* With the function $\phi(w, x) = \epsilon (a^\top(x, 1))^2_+$ where $w = (a, \epsilon) \in \mathbb{R}^{d+1} \times \{-1, 1\}$, we recover the same hypothesis class than two-layer neural networks with squared ReLU activation. This function is differentiable and rigorously covered by all theorems[1].

2.2. Parameterizing with a measure

The particular structure of two-layer neural networks allows for an alternative description of the predictor function. For $\mu \in \mathcal{P}_2(\mathbb{R}^p)$, we define

$$h(\mu, x) = \int_{\mathbb{R}^p} \phi(w, x) \, \mathrm{d}\mu(w). \tag{2}$$

Finite width networks as in Eq. (1) are recovered when μ is a discrete measure with m atoms.

---

1. Our arguments can indeed be applied to any situation where the parameter space can be factored as $\mathbb{R}_+ \times \Theta$ where Θ is a compact Riemannian manifold without boundary, see Chizat (2019). For clarity, we limit ourselves to a parameter space $\mathbb{R}^p$ (which corresponds to $\Theta = \mathbb{S}^{p-1}$) while for S-ReLU, this would correspond to $\Theta = \mathbb{S}^{p-1} \times \{-1, 1\}$.

The representation in Eq. (2) can be reduced to a *convex neural network* parameterized by an unnormalized measure (Bengio et al., 2006), as follows. We define the 2-homogenous projection operator, $\Pi_2 : \mathcal{P}_2(\mathbb{R}^p) \to \mathcal{M}_+(\mathbb{S}^{p-1})$ characterized by the property that, for any $\varphi \in \mathcal{C}(\mathbb{S}^{p-1})$, it holds

$$\int_{\mathbb{S}^{p-1}} \varphi(\theta) \, \mathrm{d}[\Pi_2(\mu)](\theta) = \int_{\mathbb{R}^p} \|w\|^2 \varphi(w/\|w\|) \, \mathrm{d}\mu(w), \tag{3}$$

where the last integrand is extended by continuity at $w = 0$. This operator projects the mass of μ on the unit sphere after re-weighting it by the squared distance to the origin. Seing $\Pi_2(\mu)$ as a measure on $\mathbb{R}^p$ supported on the sphere, it holds by construction $h(\mu, \cdot) = h(\Pi_2(\mu), \cdot)$ for all $\mu \in \mathcal{P}_2(\mathbb{R}^p)$. Note that the restriction to *nonnegative* measures, which is not present in convex neural networks, does not change the expressivity of the model thanks to the assumption that ϕ is balanced in (A1).

2.3. Max-margins and functional norms

Given the training set $(x_i, y_i)_{i \in [n]}$, the margin of a predictor $f : \mathbb{R}^d \to \mathbb{R}$ is given by $\min_{i \in [n]} y_i f(x_i)$. When the margin is strictly positive, the predictor makes no error on the training set and its value is typically seen as the worst confidence of the predictor. Max-margin predictors are those that maximize the margin over a certain set of functions. When this set of functions is given by a unit ball for a certain norm $\|\cdot\|$, they solve

$$\max_{\|f\| \le 1} \min_{i \in [n]} y_i f(x_i).$$

In this paper we deal with two notions of norms, that in turn define two types of max-margin classifiers. We refer to Bach (2017a) for a more detailed presentation.

Variation norm. Given a feature function ϕ satisfying (A1), we consider the space $\mathcal{F}_1$ of functions that can be written as $f(x) = \int_{\mathbb{S}^{p-1}} \phi(\theta, x) \, \mathrm{d}\nu(\theta)$, where $\nu \in \mathcal{M}_+(\mathbb{S}^{p-1})$ has finite mass (since ϕ is balanced, we could equivalently take $\nu \in \mathcal{M}(\mathbb{S}^{p-1})$). The infimum of $\nu(\mathbb{S}^{p-1})$ over all such decompositions defines a norm $\|f\|_{\mathcal{F}_1}$, sometimes called the *variation norm* on $\mathcal{F}_1$ (Kurková and Sanguineti, 2001). The $\mathcal{F}_1$-max-margin of the training set is denoted γ_1 and given by

$$\gamma_1 \coloneqq \max_{\|f\|_{\mathcal{F}_1} \le 1} \min_{i \in [n]} y_i f(x_i) = \max_{\substack{\nu \in \mathcal{M}_+(\mathbb{S}^{p-1}) \\ \nu(\mathbb{S}^{p-1}) \le 1}} \min_{i \in [n]} y_i \int_{\mathbb{S}^{p-1}} \phi(\theta, x) \, \mathrm{d}\nu(\theta). \tag{4}$$

For ReLU networks, the variation norm defined above does not *a priori* coincide with the variation norm as defined by Bengio et al. (2006) and Bach (2017a) where the feature function is instead $\tilde{\phi}(\tilde{\theta}, x) = (a \cdot x + b)_+$ for $\tilde{\theta} = (a, b) \in \mathbb{R}^d \times \mathbb{R}$. Still, using ϕ or $\tilde{\phi}$ leads to norms which are equal up to a factor 2 (see Neyshabur et al. (2014) or Appendix B). See Savarese et al. (2019); Ongie et al. (2019) for analytical descriptions of the space $\mathcal{F}_1$ for ReLU networks.

RKHS norm. Considering more specifically a two-layer neural network with activation function $\sigma : \mathbb{R} \to \mathbb{R}$, we can define another norm and function space, which leads to a reproducing kernel Hilbert space (RKHS). Let $\tau \in \mathcal{P}(\mathbb{S}^{p-1})$ be the uniform measure on the sphere $\mathbb{S}^{p-1}$ where $p = d + 1$ and define $\mathcal{F}_2$ as the space of functions of the form $f(x) = \int_{\mathbb{S}^{p-1}} \sigma(b + c^\top x) g(b, c) \, \mathrm{d}\tau(b, c)$, for some square-integrable function $g \in L^2(\tau)$. The infimum of $\|g\|_{L^2(\tau)} = \left(\int |g(b, c)|^2 \, \mathrm{d}\tau(b, c) \right)^{\frac{1}{2}}$ over such decompositions defines a norm $\|f\|_{\mathcal{F}_2}$. It is shown by Bach (2017a) that $\mathcal{F}_2$ is a RKHS.

The $\mathcal{F}_2$-max-margin of the training set is denoted γ_2 and given by

$$\gamma_2 := \max_{\|f\|_{\mathcal{F}_2} \leq 1} \min_{i \in [n]} y_i f(x_i) = \max_{\|g\|_{L^2(\tau)} \leq 1} \min_{i \in [n]} y_i \int_{\mathbb{S}^{p-1}} \sigma(b + c^\top x_i) g(b, c) \, \mathrm{d}\tau(b, c). \tag{5}$$

This is a separable kernel support vector machine problem.

Statistical and computational properties. In Section 6, we will show that the margin γ_1 can be large even in high dimension when the dataset has hidden low dimensional structure, which leads to strong generalization guarantees, which is *a priori* not true for γ_2. While $\mathcal{F}_2$-max-margin classifiers can be found with convex optimization techniques (such as training only the output layer, as shown in Section 5), it is not clear *a priori* how to find $\mathcal{F}_1$-max-margin classifiers. In the next section, we show that training an over-parameterized two-layer neural network precisely does that.

2.4. Training dynamics in the infinite width limit

Assumptions. Given a loss function $\ell : \mathbb{R} \to \mathbb{R}_+$, we define the empirical risk associated to a predictor $h_m(\mathbf{w}, \cdot)$ of the form Eq. (1) as $\frac{1}{n} \sum_{i=1}^n \ell(-y_i h_m(\mathbf{w}, x_i))$. Our analysis of the training dynamics relies on the following assumptions on the loss.

(A2) The loss ℓ is differentiable with a locally Lipschitz-continuous gradient. It has an *exponential tail* in the sense that $\ell(u) \sim \ell'(u) \sim \exp(u)$ as $u \to -\infty$, it is strictly increasing and there exists $c > 0$ such that $\ell'(u) \geq c$ for $u \geq 0$.

The main examples are the logistic loss $\ell(u) = \log(1 + \exp(u))$ and exponential loss $\ell(u) = \exp(u)$. Note that for our main result Theorem 3, we do not assume convexity of the loss since only the tail behavior matters. We make the following assumptions on the feature function, in addition to (A1).

(A3) The family $(\phi(\cdot, x_i))_{i \in [n]}$ is linearly independent and for $i \in [n]$, the function $\phi(\cdot, x_i)$ is differentiable with a Lipschitz-continuous gradient and subanalytic (i.e., its graph is locally the linear projection of a bounded semianalytic set).

Let us comment these assumptions. Requiring that the family is linearly independent is equivalent to requiring that arbitrary labels can be fitted on the training input $(x_i)_{i \in [n]}$ within our hypothesis class $\{x \mapsto \int \phi(\theta, x) \, \mathrm{d}\nu(\theta) \; ; \; \nu \in \mathcal{M}_+(\mathbb{S}^{p-1})\}$. This assumption is satisfied by ReLU and S-ReLU networks (Bach, 2017a) as soon as $x_i \neq x_{i'}, \forall i \neq i'$. The differentiability assumption is the most undesirable one because it excludes ReLU networks (but not S-ReLU networks). Although the training dynamic could potentially be defined without this assumption (Lyu and Li, 2019), the proof of Theorem 3 relies on it. Finally, subanalyticity is a mild assumption required in a technical proof step that invokes Sard's lemma. Functions defined by piecewise polynomials are subanalytic and thus both ReLU and S-ReLU networks satisfy it, see Bolte et al. (2006) for a definition.

Gradient flow of the smooth-margin objective. In order to obtain simpler proofs we consider maximizing minus the logarithm of the empirical risk, instead of the empirical risk itself. This allows to directly interpret the training dynamics as maximizing a *smooth-margin* and leads to the same continuous time dynamics up to time reparameterization. We define the function $S : \mathbb{R}^n \to \mathbb{R}$ as

$$S(u) = -\log\left(\frac{1}{n} \sum_{i=1}^n \ell(-u_i)\right). \tag{6}$$

When ℓ is the exponential, this function is known as the soft-min or the free energy and is concave (Mézard and Montanari, 2009). Notice however that for now we do not assume that ℓ is the exponential but only (A2), in order to cover the case of the logistic function.

With a model of the form in Eq. (1), this leads to an objective function $F_m : (\mathbb{R}^p)^m \to \mathbb{R}$ on the vector of parameters $\mathbf{w} = (w_j)_{j \in [m]}$ defined as $F_m(\mathbf{w}) = S(\hat{h}_m(\mathbf{w}))$, where we have denoted $\hat{h}_m(\mathbf{w}) = (y_i \, h_m(\mathbf{w}, x_i))_{i \in [n]}$. We consider a (potentially random) initialization $\mathbf{w}(0) \in (\mathbb{R}^p)^m$ and the (ascending) gradient flow of this objective function, which is a differentiable path $(\mathbf{w}(t))_{t \geq 0}$ starting from $\mathbf{w}(0)$ and such that for all $t \geq 0$,

$$\frac{d}{dt}\mathbf{w}(t) = m \nabla F_m(\mathbf{w}(t)). \tag{7}$$

Up to the gradient sign, this gradient flow is an approximation of gradient descent (Gautschi, 1997; Scieur et al., 2017) and stochastic gradient descent (SGD) (Kushner and Yin, 2003, Thm. 2.1) with small step sizes.[2] Classical results guarantee that under Assumption (A2-3), this gradient flow is uniquely well defined.

Wasserstein gradient flow. Taking the point of view presented in Section 2.2, we may interpret the training dynamics as a path $\mu_{t,m} = \frac{1}{m}\sum_{j=1}^{m} \delta_{w_j(t)}$ in $\mathcal{P}_2(\mathbb{R}^p)$. As we now explain, it turns out that this dynamics is a gradient flow for a function defined on $\mathcal{P}_2(\mathbb{R}^p)$, which allows to seamlessly take the limit $m \to \infty$. Let F be the functional on $\mathcal{P}_2(\mathbb{R}^p)$ defined as

$$F(\mu) = S(\hat{h}(\mu)),$$

where similarly as above we define $\hat{h} : \mathcal{P}_2(\mathbb{R}^p) \to \mathbb{R}^n$ as $\hat{h}(\mu) = (y_i \, h(\mu, x_i))_{i \in [n]}$, and let F'_μ be its Fréchet derivative at μ, which is represented by the function $F'_\mu(w) = \sum_{i=1}^{n} y_i \phi(w, x_i) \nabla_i S(\hat{h}(\mu))$. Let us give a definition of Wasserstein gradient flow (tailored to our smooth setting), which will be connected to the training dynamics of Eq. (7) in Theorem 2.

Definition 1 (Wasserstein gradient flow) *A Wasserstein gradient flow for the functional F is a path $(\mu_t)_{t \geq 0}$ such that there exists a flow $X : \mathbb{R}_+ \times \mathbb{R}^p \to \mathbb{R}^p$ satisfying $\mu_t = (X_t)_\# \mu_0$ (where $X_t(\cdot) = X(t, \cdot)$), $X(0, \cdot) = X_0 = \mathrm{id}_{\mathbb{R}^p}$ and for all $(t, w) \in \mathbb{R}_+ \times \mathbb{R}^p$,*

$$\frac{d}{dt}X(t, w) = \nabla F'_{\mu_t}(X(t, w)). \tag{8}$$

It can be directly checked that when μ_0 is discrete, we recover the training dynamics defined in Eq. (7). In this case, $w_j(t) = X(t, w_j(0))$ is the position (in parameter space) at time t of the hidden unit initialized with parameters $w_j(0)$. The following theorem shows that Wasserstein gradient flows characterize the training dynamics of infinitely wide two-layer neural networks. It is an application of Chizat and Bach (2018, Thm. 2.6), see details in Appendix C (hereafter, by convergence in $\mathcal{P}_2$, we mean weak convergence and convergence of the second moments (Ambrosio et al., 2008)).

Theorem 2 (Infinite width limit of training) *Under (A1-3), if the sequence $(w_j(0))_{j \in \mathbb{N}_*}$ is such that $\mu_{0,m}$ converges in $\mathcal{P}_2(\mathbb{R}^p)$ to μ_0, then $\mu_{t,m}$ converges in $\mathcal{P}_2(\mathbb{R}^p)$ to the unique Wasserstein gradient flow of F starting from μ_0. The convergence is uniform on bounded time intervals.*

---

2. Although Theorem 3 below could be extended to discrete time analysis, this would be of little interest since the result is so far purely qualitative. In simpler settings, we study discrete time dynamics in Sections 4 and 5.

This limit can be made quantitative using the geodesic convexity estimates of Chizat and Bach (2018), and the stability results of Ambrosio et al. (2008, Thm. 11.2.1) but with an exponential dependency in time. In the different setting of the square loss, error estimates for SGD have been derived by Mei et al. (2018, 2019). This limit dynamics covers, but is not limited to, the lazy training dynamics studied by Li and Liang (2018); Jacot et al. (2018); Du et al. (2019) which here corresponds to a short time analysis when the initialization has a large variance (see Figure 3 in Section 7).

3. Main result: implicit bias of gradient flow

We are now in position to state the main theorem of this paper, which characterizes the implicit bias of training infinitely wide two-layer neural networks with a loss with an exponential tail.

Theorem 3 (Implicit bias) *Under* (A1-3), *assume that* $\Pi_2(\mu_0)$ *has full support on* $\mathbb{S}^{p-1}$. *If* $\nabla S(\hat{h}(\mu_t))$ *converges and* $\bar{\nu}_t = \Pi_2(\mu_t)/([\Pi_2(\mu_t)](\mathbb{S}^{p-1}))$ *converges weakly to some* $\bar{\nu}_\infty$, *then this limit* $\bar{\nu}_\infty$ *is a maximizer for the* $\mathcal{F}_1$-*max-margin problem in Eq.* (4).

We can make the following observations:

- The strength of this result is that the limit $\bar{\nu}_\infty$ of a *non-convex* dynamics is a *global* minimizer of Eq. (4). Its proof relies, among other things, on a compatibility between the optimality conditions and the gradient flow dynamics, which is specific to the 2-homogeneous case.

- It is an open question to prove that $\nabla S(\hat{h}(\mu_t))$ and $\bar{\nu}_t$ converge is this setting. Note that the unnormalized measure ν_t does *not* converge, so the global convergence result from Chizat and Bach (2018) (which has a similar assumption regarding the existence of a limit) does not apply.

- Unlike in the convex case (Soudry et al., 2018), the dynamics does not completely forget where it started from. For instance, when initialized with a Dirac measure, the Wasserstein gradient flow can only converge to a Dirac measure, which is typically not a global minimizer.

Together, Theorems 2 and 3 give asymptotic guarantees for training finite width neural networks.

Corollary 4 *Under the assumptions of Theorem 3, assume that the sequence* $(w_j(0))_{j \in \mathbb{N}_*}$ *is such that* $\mu_{0,m}$ *converges in* $\mathcal{P}_2(\mathbb{R}^p)$ *to* μ_0. *Then, denoting* $\bar{\nu}_{m,t} = \Pi_2(\mu_{m,t})/[\Pi_2(\mu_{m,t})](\mathbb{S}^{p-1})$, *it holds*

$$\lim_{m,t\to\infty} \left(\min_{i\in[n]} y_i \int \phi(\theta, x_i) \, \mathrm{d}\bar{\nu}_{m,t} \right) = \gamma_1.$$

Typically, the sequence of initial neurons parameters $(w_j(0))_{j \in \mathbb{N}_*}$ are sampled from a measure μ_0 that satisfies the support condition of Theorem 3, such as a Gaussian distribution. Note that limits in t and m can be interchanged so the convergence is not conditioned on a particular scaling.

Dealing with ReLU.   Let us now state a result which is a first step towards covering the case of ReLU networks, in spite of their non-differentiability[3]. Although its assumption (*) is arguably too strong, we state it in order to point to technical open questions, see details in Appendix H. For an input distribution $\rho \in \mathcal{P}(\mathbb{R}^d)$ with a bounded density and bounded support $\mathcal{X}$, and $y : \mathcal{X} \to \{-1, 1\}$ continuous, we consider the population objective

$$F(\mu) = -\log \Big[\int_{\mathcal{X}} \exp \Big(-y(x)h(\mu, x) \Big) \, \mathrm{d}\rho(x) \Big]. \tag{9}$$

---

3. Added in revision. A similar idea for the initialization of ReLU was proposed independently in Wojtowytsch (2020).

Theorem 5 *There exists $(\mu_t)_{t \geq 0}$ a Wasserstein gradient flow of the objective Eq. (9) with $\mu_0 = \mathcal{U}(\mathbb{S}^d) \otimes \mathcal{U}(\{-1, 1\})$, i.e., input (resp. output) weights uniformly distributed on the sphere (resp. on $\{-1, 1\}$). If $\nabla S[\hat{h}(\mu_t)]$ converges weakly in $\mathcal{P}(\mathcal{X})$, if $\bar{\nu}_t = \Pi_2(\mu_t)/([\Pi_2(\mu_t)](\mathbb{S}^{p-1}))$ converges weakly in $\mathcal{P}(\mathbb{S}^{p-1})$ and if (*) F'_{μ_t} converges in $\mathcal{C}^1_{loc}$ to some F' that satisfies the Morse-Sard property (see details in Appendix H), then $h(\bar{\nu}_\infty, \cdot)$ is a maximizer for $\max_{\|f\|_{\mathcal{F}_1} \leq 1} \min_{x \in \mathcal{X}} y(x) f(x)$.*

4. Insights on the convergence rate and choice of step-size

While making Corollary 4 quantitative in terms of number of neurons and the number of iterations is left as an open question, it is of practical importance to better understand the effect of the choice of step-size. In this section, we look at a simplified dynamics where the direction of each parameter $w_j(t)$ is fixed after initialization and only its magnitude evolves. A complete discrete-time analysis is possible in this case, using tools from convex analysis.

We consider a model of the form of Eq. (1) but with $w_j(t)$ written as $r_j(t)\theta_j$, where $r_j(t) \in \mathbb{R}_+$ is trained and $\theta_j \in \mathbb{S}^{p-1}$ is fixed at initialization. Plugging this model into the soft-min loss (6) yields an objective function $F_m : \mathbb{R}^m_+ \to \mathbb{R}$ defined as

$$F_m(r) = -\log\left(\frac{1}{n}\sum_{i=1}^n \exp\left(-\frac{1}{m}\sum_{j=1}^m z_{i,j} r_j^2\right)\right),$$

where $z_{i,j} = y_i \phi(\theta_j, x_i)$ are the signed fixed features. We focus on the exponential loss $\ell = \exp$ in this section and the next one for simplicity. We study the gradient ascent dynamics with initialization $r(0) \in \mathbb{R}^m_+$ and sequence of step-sizes $(\eta(t))_{t \in \mathbb{N}}$,

$$r(t+1) = r(t) + \eta(t) m \nabla F_m(r(t)).$$

This dynamics is studied by Gunasekar et al. (2018a) where it is shown to converge to a max ℓ_1-margin classifier without a rate. In the next proposition, we prove convergence of the best iterate to maximizers at an asymptotic rate $\log(t)/\sqrt{t}$, by exploiting an analogy with online mirror ascent.

Proposition 6 *Let $a_j(t) = r_j(t)^2/m$ for $j \in [m]$, $\beta(t) = \|a(t)\|_1$ and $\bar{a}(t) = a(t)/\beta(t)$. For the step-sizes $\eta(t) = 1/(16\|z\|_\infty \sqrt{t+1})$ and a uniform initialization $r(0) \propto \mathbf{1}$, it holds*

$$\max_{0 \leq s \leq t-1} \min_{i \in [n]} z_i^\top \bar{a}(s) \geq \gamma_1^{(m)} - \frac{\|z\|_\infty}{\sqrt{t}}(8\log(m) + \log(t) + 1) - \frac{4B\log n}{\sqrt{t}}.$$

where $\gamma_1^{(m)} := \max_{a \in \Delta^{m-1}} \min_{i \in [n]} z_i^\top a$ and $B := \sum_{s=0}^\infty \frac{1}{\beta(s)\sqrt{s+1}} < \infty$ when $\gamma_1^{(m)} > 0$.

In the proof of Lemma 17, it can be seen that our bound on B grows to ∞ as $\gamma_1^{(m)}$ goes to zero.

Proof idea. To prove Proposition 6, we consider the family of *smooth-margin* functions

$$G_\beta(a) = -\frac{1}{\beta}\log\left(\frac{1}{n}\sum_{i=1}^n \exp\left(-\beta\sum_{j=1}^m z_{i,j} a_j\right)\right),$$

and we show that $\bar{a}(t)$ approximately follows *online mirror ascent* for the sequence of concave functions $G_{\beta(t)}$ in the simplex Δ^{m-1} with step-sizes $\eta(t)$. It then only remains to apply classical bounds for mirror descent and use the fact that $|\min_{i \in [n]} z_i^\top a - G_\beta(a)| \leq \log(n)/\beta$. This algorithm thus implicitly performs online optimization on the regularization path. It is also analogous to smoothing techniques in non-smooth optimization (Nesterov, 2005).

Continuous limit. Using the notations from Section 2, the dynamics $\sum_{j=1}^m \bar{a}_j(t)\delta_{\theta_j}$ solves

$$\gamma_1^{(m)} := \max_{\substack{\nu \in \mathcal{M}_+(\mathbb{S}^{p-1}) \\ \nu(\mathbb{S}^{p-1}) \leq 1}} \min_{i \in [n]} y_i \int_{\mathbb{S}^{p-1}} \phi(\theta, x_i)\, \mathrm{d}\nu(\theta) \quad \text{subject to} \quad \nu \text{ supported on } \{\theta_j\}_{j \in [m]}.$$

When $\frac{1}{m}\sum_{j=1}^m \delta_{\theta_j}$ converges to the uniform measure on the sphere, we thus recover the same implicit bias as in Theorem 3 and $\gamma_1^{(m)} \to \gamma_1$ (note that the logarithmic dependency in m in Proposition 6 could be removed with a slightly finer analysis as done in Chizat (2019)). While functions in $\mathcal{F}_1$ may be well-approximated with a small number of neurons (Bach, 2017a; Jones, 1992), this is not anymore true if the positions $\{\theta_j\}_{j \in [m]}$ of those neurons are fixed a priori (see Barron (1993) for exponential lower bounds in a similar setting). In Theorem 3, positions are allowed to vary during training: this makes its setting more challenging but also much more relevant.

5. Training only the output layer

For two-layer neural networks, it is instructive to compare the implicit bias of training both layers (as in Section 3) versus that of training only the output layer, the input layer being initialized randomly and fixed. This model gives the objective function $F : \mathbb{R}^m \to \mathbb{R}$ defined as

$$F(r) = -\log\left(\frac{1}{n}\sum_{i=1}^n \exp\left(-\frac{1}{m}\sum_{j=1}^m z_{i,j}r_j\right)\right),$$

where $z_{i,j} = y_i\sigma(b_j + x_i^\top c_j)$ is the signed output of neuron j for the training point i and $\sigma : \mathbb{R} \to \mathbb{R}$ is the non-linearity, e.g., $\sigma(u) = \max\{0, u\}$ for ReLU networks. We study the gradient ascent dynamics with initialization $r(0) \in \mathbb{R}^m$ and sequence of step-sizes $(\eta(t))_{t \in \mathbb{N}}$:

$$r(t+t) = r(t) + \eta(t)m\nabla F(r(t)).$$

Soudry et al. (2018) show that for a step-size of order $1/\sqrt{t}$, this dynamics converges in $O(\log(t)/\sqrt{t})$ to a max ℓ_2-margin classifier. Next, we show that it converges in $O(1/\sqrt{t})$ for larger, non-vanishing step-sizes, with a different proof technique. The fact that the algorithm converges at essentially the same speed for very different step-sizes shows an advantageous self-regularizing property.

Proposition 7 *Let $a(t) = r(t)/m$, $\beta(t) = \max\{1, \max_{0 \leq s \leq t} \sqrt{m}\|a(t)\|_2\}$ and $\bar{a}(t) = a(t)/\beta(t)$. Assume $\gamma_2^{(m)} := \max_{\sqrt{m}\|a\|_2 \leq 1} \min_{i \in [n]} z_i^\top a > 0$. For the step-sizes $\eta(t) = \beta(t)\sqrt{2}/(\|z\|_\infty\sqrt{t+1})$ and initialization $r(0) = 0$, it holds*

$$\max_{0 \leq s \leq t-1} \min_{i \in [n]} z_i^\top \bar{a}(s) \geq \gamma_2^{(m)} - \frac{\|z\|_\infty}{\sqrt{t}}\left(2\sqrt{2} + \frac{\sqrt{3}\log n}{\gamma_2^{(m)}}\right).$$

Proof idea. Similarly to the proof of Proposition 6, we show that $\bar{a}(t)$ follows an *online projected gradient ascent* for the sequence of functions $G_{\beta(t)}$ in the ball $\{a \in \mathbb{R}^m \ ; \ \|a\|_2 \leq 1/\sqrt{m}\}$ and with step-sizes $\eta(t)/(m\beta(t))$. From there, we use standard optimization results and prove that $\beta(t) \to \infty$ to conclude. Note that a different reduction to mirror descent for this dynamics was also exhibited by Ji and Telgarsky (2019b) and used to derive tight convergence rates but with a much smaller step-size than in Proposition 7 and with a different Bregman divergence.

Random features for kernel max-margin classifier. Using the notations from Section 2, the dynamics $(r_j)_{j\in[m]}$ converges to a solution to

$$\gamma_2^{(m)} = \max_{g\in L^2(\tau_m)} \min_{i\in[n]} y_i \int g(b,c)\sigma(b+x_i^\top c)\,\mathrm{d}\tau_m(b,c) \quad \text{subject to} \quad \|g\|_{L^2(d\tau_m)} \leq 1,$$

where $\tau_m = \frac{1}{m}\sum_{j=1}^m \delta_{(b_j,c_j)}$. Typically, the input layer parameters are sampled from a distribution $\tau \in \mathbb{R}^{1+d}$, which corresponds to a random feature approximation for the $\mathcal{F}_2$-max-margin problem of Eq. (5) and we have $\gamma_2^{(m)} \to \gamma_2$. In stark contrast to the space $\mathcal{F}_1$, functions in $\mathcal{F}_2$ can be well approximated with few random features even in high dimension. See Rahimi and Recht (2008); Bach (2017b) for a analysis of the number of features needed for an approximation with error ε, typically of order $1/\varepsilon^2$.

6. Dimension independent generalization bounds

In this section, we give arguments showing the favorable statistical properties of the bias exhibited in Theorem 3 for ReLU networks. We propose to measure the complexity of the dataset $S_n = (x_i, y_i)_{i=1}^n$ with the following *projected interclass distance* defined, for $r \in [d]$, as

$$\Delta_r(S_n) := \sup_P \left\{ \inf_{y_i \neq y_{i'}} \|P(x_i) - P(x_{i'})\|_2 \; ; \; P \text{ is a rank-}r \text{ orthogonal projection} \right\}. \quad (10)$$

For each dimension r, it looks for the r-dimensional subspace which maximizes the distance between the two classes. Interclass distance often appears in the statistical analysis of classification problems (see, e.g., Li and Liang, 2018) often complemented with "clustered data" assumptions. Our definition is designed to capture the fact that if $\Delta_r \approx \Delta_d$ for $r \ll d$, then there is a hidden structure which can be exploited for statistical efficiency.

Theorem 8 (Generalization bound) *For any $\epsilon \in (0,1)$ and $r \in [d]$, there exist $C(r), C_\epsilon(r) > 0$ such that the following holds. If $(x,y) \sim \mathbb{P}$ is such that for some $R > 0$ and $\Delta_r(\mathbb{P}) \leq C(r)$, it holds $\Delta_r(S_n) \leq \Delta_r(\mathbb{P})$ and $\|x\|_2 \leq R$ almost surely, then it holds with probability at least $1 - \delta$ over the choice of i.i.d. samples $S_n = (x_i, y_i)_{i=1}^n$, for f the $\mathcal{F}_1$-max-margin classifier on S_n,*

$$\mathbb{P}[yf(x) < 0] \leq \frac{C_\epsilon(r)}{\sqrt{n}} \left(\frac{R}{\Delta_r(\mathbb{P})}\right)^{\frac{r+3}{2-\epsilon}} + \sqrt{\frac{\log(B)}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

where $B = \log_2(4(R+1)C_2(r)) + (r+2)\log_2(R/\Delta_r(\mathbb{P}))$. The same bound applies to the $\mathcal{F}_2$-max-margin classifier for $r = d$.

Proof idea. We first lower-bound the margins γ_1 and γ_2 in terms of $\Delta_r(S_n)$ and then apply margin-based generalization bounds (Koltchinskii and Panchenko, 2002) and bounds on the Rademacher complexity of the unit ball of $\mathcal{F}_1$ and $\mathcal{F}_2$.

The rate $n^{-1/2}$ is suboptimal (an exponential decay of the test error is possible in this context (Pillaud-Vivien et al., 2018)), but this statement is a strong non-asymptotic bound: for the $\mathcal{F}_1$-max-margin classifier, d does not appear in the exponent of the ratio $R/\Delta_r(\mathbb{P})$, which characterizes the difficulty of the problem. Related generalization bounds are given by Wei et al. (2019), where a factor d improvement for the $\mathcal{F}_1$ versus $\mathcal{F}_2$-max-margin classifier is shown on a specific example. Also Montanari et al. (2019) prove generalization bounds for linear max-margin classifiers.
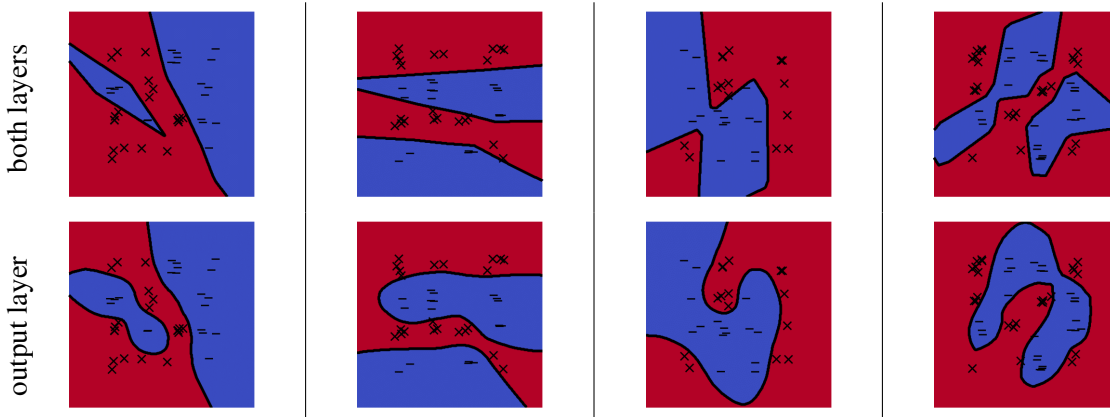
Figure 1: Comparison of the implicit bias of training (top) both layers versus (bottom) the output layer for ReLU networks with $d = 2$ and for 4 different random training sets.

7. Numerical experiments

In this section, we consider a large ReLU network with $m = 1000$ hidden units, and compare the implicit bias and statistical performances of training both layers – which leads to a max margin classifier in $\mathcal{F}_1$ – versus the output layer – which leads to max margin classifier in $\mathcal{F}_2$. The experiments are reproducible with the Julia code that can be found online[4].

Setting. Our data distribution is supported on $[-1/2, 1/2]^d$ and is generated as follows. In dimension $d = 2$, the distribution of input variables is a mixture of k^2 uniform distributions on disks of radius $1/(3k - 1)$ on a uniform 2-dimensional grid with step $3/(3k - 1)$, see Figure 2(a) for an illustration with $k = 3$. In dimension larger than 2, all other coordinates follow a uniform distribution on $[-1/2, 1/2]$. Each cluster is then randomly assigned a class in $\{-1, +1\}$. For such distributions, the parameters appearing in Theorem 8 satisfy $\Delta_2(\mathbb{P}) \geq 1/(3k - 1)$ and $R \leq \sqrt{d}$.

Low dimensional illustrations. Figure 1 illustrates the differences in the implicit biases when $d = 2$. It represents a sampled training set and the resulting decision boundary between the two classes for 4 examples. The $\mathcal{F}_1$- max-margin classifier is non-smooth and piecewise affine, which comes from the fact that the mass constraint in Eq. (4) favors sparse solutions. In contrast, the max-margin classifier in $\mathcal{F}_2$ has a smooth decision boundary, which is typical of learning in a RKHS.

Performance. In higher dimensions, we observe the superiority of training both layers by plotting the test error versus m or d on Figure 2(b) and 2(c). We ran 20 independent experiments with $k = 3$ and show with a thick line the average of the test error $\mathbb{P}(yf(x) < 0)$ after training. Note that R grows as $\sqrt{d}$ so the dependency in d observed in Figure 2(c) is not in contradiction with Theorem 8. Finally, Figure 2(d) illustrates Corollary 4 and shows the $\mathcal{F}_1$-margin after training both layers. For each m, we ran 30 experiments using fresh random samples from the same data distribution.

Two implicit biases in one dynamics. In Figure 3, we illustrate for $d = 2$ a case where *two* different kinds of implicit biases show up in a single dynamics (t is the number of iterations with a

---

4. https://github.com/lchizat/2020-implicit-bias-wide-2NN

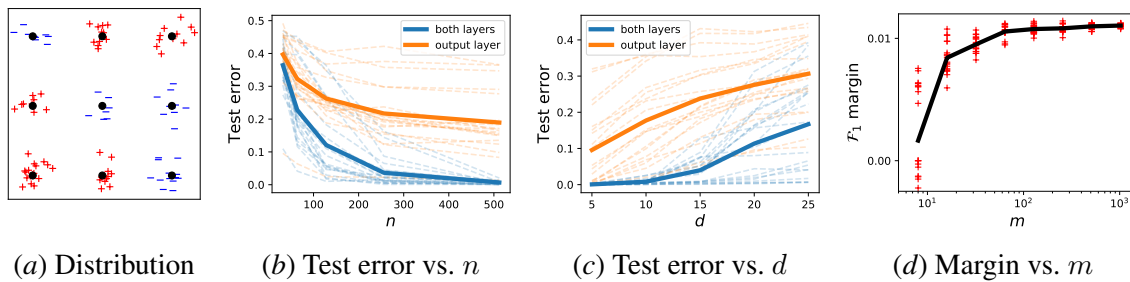(a) Distribution   (b) Test error vs. n (c) Test error vs. d (d) Margin vs. m

Figure 2: (a) Projection of the data distribution on the two first dimensions, (b) test error as a function of n with $d = 15$, (c) test error as a function of d with $n = 256$ (d) $\mathcal{F}_1$-margin at convergence as a function of m when training both layers with $n = 256$, $d = 15$.

(a) $t = 10^2$ (b) $t = 5 * 10^2$ (c) $t = 5 * 10^3$ (d) $t = 3 * 10^4$ (e) $t = 6 * 10^4$ (f) $t = 3.10^5$
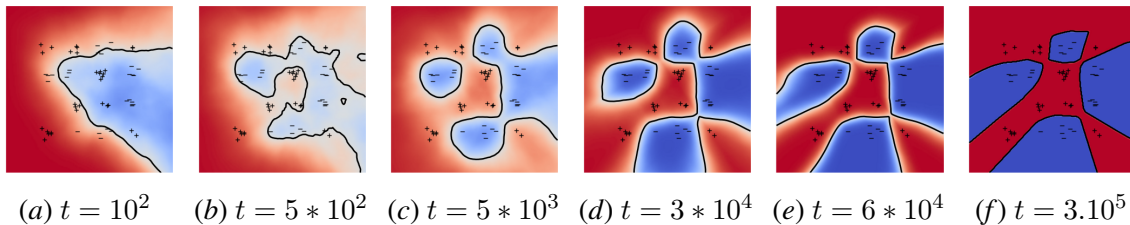
Figure 3: Dynamics of the classifier while training both layers for an initialization with a large variance and a small initial step-size. The classifier first approaches the max-margin classifier for the *tangent kernel* (Jacot et al., 2018) (c) and eventually converges to the $\mathcal{F}_1$-max-margin (f).

constant step-size). We initialize the ReLU network with a large variance ($\mathcal{N}(0, 40^2)$). The model is at first in the *lazy regime* (Chizat et al., 2019) and follows closely the dynamics of its linearization around initialization, which converges to the max-margin classifier for the *tangent kernel* (Jacot et al., 2018). It then converges to the $\mathcal{F}_1$-max-margin classifier as suggested by Theorem 3. In order to observe this intermediate implicit bias, one needs an initial step-size inversely proportional to the scale of the initialization (Chizat et al., 2019).

8. Conclusion

We have shown that for wide two-layer ReLU neural networks, training both layers or only the output layer leads to very different implicit biases. When training both layers, the classifier converges to a max-margin classifier for a non-Hilbertian norm, which enjoys favorable statistical properties. Interestingly, this problem does not seem to be directly solvable with known convex methods in high dimension. Proving complexity guarantees for this non-convex gradient flow is an important open question for future work. In particular, even for infinite width, continuous time dynamics as in Theorem 3, it is still unknown whether a convergence rate can be given under reasonable conditions.

Acknowledgements

References

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: in Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media, 2008.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017a.

Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *The Journal of Machine Learning Research*, 18(1):714–751, 2017b.

Andrew R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Yoshua Bengio, Nicolas Le Roux, Pascal Vincent, Olivier Delalleau, and Patrice Marcotte. Convex neural networks. In *Advances in Neural Information Processing Systems*, pages 123–130, 2006.

Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. A nonsmooth Morse–Sard theorem for subanalytic functions. *Journal of Mathematical Analysis and Applications*, 321(2):729–740, 2006.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Lénaïc Chizat. Sparse optimization on measures with over-parameterized gradient descent. *arXiv preprint arXiv:1907.10300*, 2019.

Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, 2018.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, 2019.

Simon S. Du, Wei Hu, and Jason D. Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In *Advances in Neural Information Processing Systems*, pages 384–395, 2018.

Simon S. Du, Xiyu Zhai, Barnabás Póczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019.

John C. Duchi, Peter L. Bartlett, and Martin J. Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012.

Walter Gautschi. *Numerical Analysis*. Springer Science & Business Media, 1997.

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

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

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, pages 8571–8580, 2018.

Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. In *International Conference on Learning Representations*, 2019a.

Ziwei Ji and Matus Telgarsky. A refined primal-dual analysis of the implicit bias. *arXiv preprint arXiv:1906.04540*, 2019b.

Lee K. Jones. A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *The Annals of Statistics*, 20(1):608–613, 1992.

Vladimir Koltchinskii and Dmitry Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 2002.

Stanislav Kondratyev and Dmitry Vorotnikov. Spherical Hellinger–Kantorovich gradient flows. *SIAM Journal on Mathematical Analysis*, 51(3):2053–2084, 2019.

Vera Kurková and Marcello Sanguineti. Bounds on rates of variable-basis and neural-network approximation. *IEEE Transactions on Information Theory*, 47(6):2659–2665, 2001.

Harold Kushner and G. George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*, volume 35. Springer Science & Business Media, 2003.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pages 8157–8166, 2018.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

Stefania Maniglia. Probabilistic representation and uniqueness results for measure-valued solutions of transport equations. *Journal de Mathématiques Pures et Appliquées*, 87(6):601–626, 2007.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. *arXiv preprint arXiv:1902.06015*, 2019.

Marc Mézard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

Mor Shpigel Nacson, Suriya Gunasekar, Jason Lee, Nathan Srebro, and Daniel Soudry. Lexicographic and depth-sensitive margins in homogeneous and non-homogeneous deep models. *arXiv preprint arXiv:1905.07325*, 2019a.

Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3420–3428, 2019b.

Yurii Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

Atsushi Nitanda and Taiji Suzuki. Stochastic particle gradient descent for infinite ensembles. *arXiv preprint arXiv:1712.05438*, 2017.

Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro. A function space view of bounded norm infinite width ReLU nets: The multivariate case. *arXiv preprint arXiv:1910.01635*, 2019.

Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Exponential convergence of testing error for stochastic gradient methods. *Proceedings of Machine Learning Research vol*, 75:1–47, 2018.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pages 1177–1184, 2008.

Grant Rotskoff, Samy Jelassi, Joan Bruna, and Eric Vanden-Eijnden. Neuron birth-death dynamics accelerates gradient descent and converges asymptotically. In *International Conference on Machine Learning*, pages 5508–5517, 2019.

Grant M. Rotskoff and Eric Vanden-Eijnden. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. In *Advances in Neural Information Processing Systems*, 2018.

Pedro Savarese, Itay Evron, Daniel Soudry, and Nathan Srebro. How do infinite width bounded norm networks look in function space? *arXiv preprint arXiv:1902.05040*, 2019.

Damien Scieur, Vincent Roulet, Francis Bach, and Alexandre d'Aspremont. Integration methods and optimization algorithms. In *Advances in Neural Information Processing Systems*, pages 1109–1118, 2017.

Maurice Sion. On general minimax theorems. *Pacific Journal of mathematics*, 8(1):171–176, 1958.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 2019.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1): 2822–2878, 2018.

Matus Telgarsky. Margins, shrinkage, and boosting. In *International Conference on Machine Learning*, pages 307–315, 2013.

Colin Wei, Jason Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets v.s. their induced kernel. In *Advances in Neural Information Processing Systems*, 2019.

Stephan Wojtowytsch. On the convergence of gradient descent training for two-layer relu-networks in the mean field regime. *arXiv preprint arXiv:2005.13530*, 2020.

Tengyu Xu, Yi Zhou, Kaiyi Ji, and Yingbin Liang. When will gradient methods converge to max-margin classifier under ReLU models? *arXiv preprint arXiv:1806.04339*, 2018.

Appendix A. Organization of the appendix

- In Appendix B, we prove the equivalence between our definition of the variation norm in Section 2.3 with the one that is used in the literature on convex neural networks.

- In Appendix C, we discuss properties of the Wasserstein gradient flow and justify Theorem 2.

- In Appendix D, we prove our main theorem Theorem 3 and its corollary.

- In Appendix E, we prove Proposition 6 on the convergence rate with fixed "positions".

- In Appendix F, we prove Proposition 7 on the convergence rate when training the output layer.

- In Appendix G, we prove Theorem 8 on the margins and generalization performance.

- In Appendiz H, we prove Theorem 5 which covers the case of ReLU networks.

Appendix B. Equivalence of two variation norms

Let us introduce, for ReLU networks, the variation norm introduced in Section 2.3 and the different definition from the literature (Bengio et al., 2006; Bach, 2017a). We will show that they are equal up to a factor 2. This is a known result (Neyshabur et al., 2014), and we provide here a natural proof using the measure theoretic formalism, for the sake of completeness. We stress that the analogous equivalence would fail for the RKHS norms, i.e., such a modification of the feature function could lead to different functional spaces.

Consider the feature functions $\phi(\theta, z) = c(a \cdot z)_+$ where $\theta = (a, c) \in \mathbb{S}^{p-1}$ and $z = (x, 1)$ (here we see $\mathbb{S}^{p-1}$ as a subset of $\mathbb{R}^{p-1} \times \mathbb{R}$) and $\tilde{\phi}(a, z) = (a \cdot z)_+$ where $a \in \mathbb{S}^{p-2}$. For a function $f : \mathbb{S}^{p-2} \to \mathbb{R}$, we consider two norms

$$\|f\|_{\mathcal{F}_1} = \inf \left\{ \nu(\mathbb{S}^{p-1}) \ ; \ f(z) = \int \phi(\theta, z) \, \mathrm{d}\nu(\theta), \ \nu \in \mathcal{M}_+(\mathbb{S}^{p-1}) \right\}$$

and

$$\|f\|_{\tilde{\mathcal{F}}_1} = \inf \left\{ |\tilde{\nu}|(\mathbb{S}^{p-2}) \ ; \ f(z) = \int \tilde{\phi}(a, z) \, \mathrm{d}\tilde{\nu}(a), \ \nu \in \mathcal{M}(\mathbb{S}^{p-2}) \right\},$$

and the associated functional spaces $\mathcal{F}_1$ and $\tilde{\mathcal{F}}_1$ where these two norms are finite. Classical weak compactness arguments guarantee that the infimum defining these two norms is attained.

Proposition 9 *It holds $\mathcal{F}_1 = \tilde{\mathcal{F}}_1$ and for all $f \in \mathcal{F}_1$, it holds $\|f\|_{\mathcal{F}_1} = 2\|f\|_{\tilde{\mathcal{F}}_1}$. Moreover, any measure that reaches the infimum for $\| \cdot \|_{\mathcal{F}_1}$ is concentrated on the set*

$$\left\{ (a, c) \in \mathbb{R}^{p-1} \times \mathbb{R} \ ; \ \|a\| = |c| = 1/\sqrt{2} \right\}.$$

An interesting consequence of this result is that empirical risk minimization with the commonly used *weight decay* regularization and the total variation regularization used by Bengio et al. (2006); Bach (2017a) (the path-norm) are equivalent.

Proof We give a constructive proof where we explicitly build a minimizer for each norm given a minimizer for the other norm. Let us start with a measure $\nu \in \mathcal{M}_+(\mathbb{S}^{p-1})$ such that $f(x) =$

$\int \phi(\theta, x) \, \mathrm{d}\mu(\theta)$. We define the linear operator $\Pi : \mathcal{M}_+(\mathbb{S}^{p-1}) \to \mathcal{M}(\mathbb{S}^{p-2})$ where $\Pi(\nu)$ is characterized by

$$\int \varphi \, \mathrm{d}\Pi(\nu) = \int c\|a\|\varphi(a/\|a\|) \, \mathrm{d}\nu((a,c)),$$

where, as usual, the integrand is extended by continuity at $a = 0$. By construction, it holds $\forall z \in \mathbb{R}^{p-2}$,

$$\int \tilde{\phi}(a, z) \, \mathrm{d}\Pi(\nu)(a) = \int c\|a\|\tilde{\phi}(a/\|a\|, z) \, \mathrm{d}\nu((a,c)) = \int \phi((a,c), z) \, \mathrm{d}\nu((a,c)) = f(z).$$

As for the total variation norm $\|\Pi(\nu)\| := |\Pi(\nu)|(\mathbb{S}^{p-2})$ of $\Pi(\nu)$, it can be bounded as follows. In the definition of $\Pi(\nu)$, we may restrict the integral over $\{c > 0\}$, which defines a measure $\Pi_+(\nu) \in \mathcal{M}_+(\mathbb{S}^{p-2})$. Similarly restricting the integral over $\{c < 0\}$ an taking the opposite gives another measure $\Pi_-(\nu) \in \mathcal{M}_+(\mathbb{S}^{p-2})$. It holds $\Pi(\nu) = \Pi_+(\nu) - \Pi_-(\nu)$ and thus $\|\Pi(\nu)\| \leq \|\Pi_+(\nu)\| + \|\Pi_-(\nu)\|$. Moreover, by integrating against $\varphi = 1$, it holds

$$\|\Pi_+(\nu)\| = \int 1 \, \mathrm{d}\Pi_+(\nu) = \int_{c>0} c\|a\| \, \mathrm{d}\nu((a,c)) \leq \frac{1}{2} \int_{c>0} \mathrm{d}\nu((a,c)),$$

since $c\|a\| \leq (\|a\|^2 + c^2)/2 = 1/2$ for $(a, c) \in \mathbb{S}^{p-1}$. Using a similar bound for $\|\Pi_-(\nu)\|$, we get that

$$\|\Pi(\nu)\| \leq \frac{1}{2} \int_{c>0} \mathrm{d}\nu((a,c)) + \frac{1}{2} \int_{c<0} \mathrm{d}\nu((a,c)) \leq \frac{1}{2}\nu(\mathbb{S}^{p-1}).$$

Finally, tracking the equality cases, it holds $\|\Pi(\nu)\| = \frac{1}{2}\nu(\mathbb{S}^{p-1})$ if and only if ν is concentrated on the set given in Proposition 9, which is the intersection of the sphere with the set of points satisfying $2|c|\|a\| = \|a\|^2 + |c|^2$.

Conversely, let $\nu \in \mathcal{M}(\mathbb{S}^{p-2})$ and consider its Jordan decomposition $\nu = \nu_+ - \nu_-$ into two nonnegative measures, which is such that $\|\nu\| = \|\nu_+\| + \|\nu_-\|$. We define two maps $T^+, T^- : \mathbb{S}^{p-2} \to \mathbb{S}^{p-1}$ as $T^+(a) = (a,1)/\sqrt{2}$ and $T^-(a) = (a,-1)/\sqrt{2}$. Now, define the linear map $T : \mathcal{M}(\mathbb{S}^{p-2}) \to \mathcal{M}_+(\mathcal{S}^{p-1})$ as

$$T(\nu) = 2(T_\#^+\nu_+ + T_\#^-\nu_-).$$

Since pushforwards preserve the mass of nonnegative measures, it holds $\|T(\nu)\| = 2(\|\nu_+\| + \|\nu_-\|) = 2\|\nu\|$. Moreover, using the definition of pushforward measures, we have

$$\int \phi(\theta, z) \, \mathrm{d}T(\nu)(\theta) = 2 \int \phi((a,1)/\sqrt{2}, z) \, \mathrm{d}\nu_+(a) + 2 \int \phi((a,-1)/\sqrt{2}, z) \, \mathrm{d}\nu_-(a)$$

$$= \int \tilde{\phi}(a, z) \, \mathrm{d}\nu_+(a) - \int \tilde{\phi}(a, z) \, \mathrm{d}\nu_-(a) = \int \tilde{\phi}(a, z) \, \mathrm{d}\nu(a) = f(z).$$

To sum up, for any feasible measure ν for the definition of $\|f\|_{\mathcal{F}_1}$, we have built a measure $\Pi(\nu)$ that is feasible for $\|f\|_{\tilde{\mathcal{F}}_1}$ with a norm divided by at most 2. Conversely, for any feasible measure ν for the definition of $\|\tilde{f}\|_{\tilde{\mathcal{F}}_1}$, we have built a measure $T(\nu)$ that is feasible for $\|f\|_{\mathcal{F}_1}$ with a norm multiplied exactly by 2. This concludes the proof. ∎

Appendix C. Details on Wasserstein gradient flows

C.1. Alternative formulations of the Wasserstein gradient flow

- (Divergence form) It can be shown that a Wasserstein gradient flow as defined in Definition 1 satisfies, in the sense of distributions, the following partial differential equation (Ambrosio et al., 2008)

$$\partial_t \mu_t = -\mathrm{div}(\nabla F'_{\mu_t} \mu_t).$$

- (Projected representation) If we look at the projected trajectory $\nu_t = \Pi_2(\mu_t)$, it can be shown that it solves the following dynamic, which is known as Wasserstein-Fisher-Rao or Hellinger-Kantorovich gradient flow of the functional $J : \mathcal{M}_+(\mathbb{S}^{p-1})$ satisfying $J(\Pi_2(\mu)) = F(\mu)$. In equation,

$$\partial_t \nu_t = -\mathrm{div}(\nabla J'_{\nu_t} \nu_t) + 4J'_{\nu_t} \nu_t, \tag{11}$$

where $J'_\nu(\theta) = \sum_{i=1}^n \nabla_i S(\hat{h}(\nu)) y_i \phi(\theta, x_i)$ is defined on the sphere, see Chizat (2019). Note that there is also a Lagrangian representation for this projected dynamics (Maniglia, 2007) .

- (Renormalized dynamics) It can also be seen with a direct computation that the normalized dynamics $\bar{\nu}_t = \nu_t/\|\nu_t\|$ satisfies the following equation

$$\partial_t \bar{\nu}_t = -\mathrm{div}(\nabla J'_{\nu_t} \bar{\nu}_t) + 4\left(J'_{\nu_t} - \int J'_{\nu_t} \, d\bar{\nu}_t\right)\bar{\nu}_t. \tag{12}$$

When the driving potential is $J'_{\bar{\nu}_t}$ (instead of J'_{ν_t}), this dynamics is known as the spherical Wasserstein-Fisher-Rao or spherical Hellinger Kantorovich gradient flow (Kondratyev and Vorotnikov, 2019) and was considered by Rotskoff et al. (2019) for neural networks training.

C.2. Proof of Theorem 2

We just need to prove that the assumptions of Chizat and Bach (2018, Theorem 2.6) are satisfied and justify that non-compactly supported initialization are also allowed.

Checking regularity assumptions. With the notations used by Chizat and Bach (2018, Assumptions 2.1), the Hilbert space $\mathcal{F}$ is $\mathbb{R}^n$, the domain "Ω" is $\mathbb{R}^p$, the risk "R" is the smooth-margin S, the function "Φ" is here $\Phi(\theta) = (y_i\phi(\theta, x_i))_{i\in[n]}$, there is no regularization, and the family of nested sets "Ω_r" are the closed balls of radius r in $\mathbb{R}^p$. We can directly check that:

- under Assumption (A2), $S : \mathbb{R}^n \to \mathbb{R}$ is differentiable and its gradient given in coordinates by $\nabla_i S(u) = \ell'(-u_i)/(\sum_{i'} \ell(-u_{i'}))$ is Lipschitz continuous and bounded on superlevel sets;

- under Assumption (A3), the function Φ is differentiable with a locally Lipschitz continuous gradient. Moreover its gradient has at most a linear growth by 2-homogeneity (this verifies assumptions from Chizat and Bach (2018, Assumptions 2.1-(iii)-(c))).

Removing the compact support assumption. Chizat and Bach (2018, Theorem 2.6) only allows an initialization in some "Ω_r", which means here that μ_0 should be compactly supported. However, when Φ is positively 2-homogeneous, this condition can be relaxed (Chizat, 2019, Appendix C.1) because the dynamics is entirely characterized by its projection on the sphere (through Π_2). Indeed, for any $\mu \in \mathcal{P}_2(\mathbb{R}^p)$ distinct from δ_0, there exists a compactly supported $\tilde{\mu} \in \mathcal{P}_2(\mathbb{R}^p)$ such that

$\Pi_2(\mu) = \Pi_2(\tilde{\mu})$. Since we have existence and uniqueness for the Wasserstein gradient flow starting from $\tilde{\mu}$, we get existence and uniqueness from the Wasserstein gradient flow $(\mu_t)_t$ starting fom μ (since it is entirely determined by the velocity field $\nabla F'_{\mu_t}$ which is itself determined by the projection of the dynamics $\Pi_2(\mu_t)$. Remark that with similar arguments, we can adapt the proof of Chizat and Bach (2018, Theorem 2.6) to any initialization $\mu \in \mathcal{P}_2(\mathbb{R}^p)$.

Appendix D. Appendix to Section 3: main theorem

D.1. Proof of the main theorem

Let us restate Theorem 3 and give its proof. We recall that $\nu_t := \Pi_2(\mu_t)$ and $\bar{\nu}_t := \nu_t / (\nu_t(\mathbb{S}^{p-1}))$.

Theorem 10 *Under* (A1-3)*, assume that $\Pi_2(\mu_0)$ has full support on $\mathbb{S}^{p-1}$, that $\nabla S(h(\mu_t))$ converges in $\mathbb{R}^n$ and that $\bar{\nu}_t$ converges weakly towards some $\bar{\nu}_\infty$. Then $\bar{\nu}_\infty$ is a maximizer of Eq. (4).*

Proof By Lemma 14 the limit $p(\infty) \in \mathbb{R}^n$ of $p(t) := \nabla S(\hat{h}(\mu_t))$ is non-zero. This implies that J'_{ν_t}, the restriction of F'_{ν_t} to the sphere, converges in $\mathcal{C}^1(\mathbb{S}^{p-1})$ (i.e. the function and its gradient converge uniformly) to a function $J'_\infty : \theta \mapsto \sum_{i=1}^n p_i(\infty) y_i \phi(\theta, x_i)$ which is non-zero because the family $(\phi(\cdot, x_1), \dots, \phi(\cdot, x_n))$ is linearly independent by (A3). Since ϕ is balanced by (A1), we have $M := \max_{\theta \in \mathbb{S}^{p-1}} J'_\infty(\theta) > 0$. The rest of the proof is divided into 3 steps.

Step 1: mass grows unbounded. In a first step, we prove that $\nu_t(\mathbb{S}^{p-1}) \to \infty$. Assume that J'_∞ is not constant (the other case will be considered later), and let $v \in]0, M/8[$ be such that $M - v$ is a regular value of J'_∞, i.e., be such that $\|\nabla J'_\infty\|$ does not vanish on the $M - v$ level-set of J'_∞. Such a v is guaranteed to exist thanks to the fact that $\phi(\cdot, x_i)$ is subanalytic (which implies that J'_∞, which is a finite sum of such $\phi(\cdot, x_i)$, is also subanalytic) and that the sphere is a subanalytic set, and then applying (Bolte et al., 2006, Thm. 14). Note that such admissible v are dense in the range of J'_∞, which will be useful in Step. 3. Let $K_v = (J'_\infty)^{-1}([M - v, M]) \subset \mathbb{S}^{p-1}$ be the corresponding super-level set. By the regular value theorem, the boundary ∂K_v of K_v is a differentiable orientable compact submanifold of $\mathbb{S}^{p-1}$ and is orthogonal to $\nabla J'_\infty$. By construction, it holds for all $\theta \in K_v$, $J'_\infty(\theta) \geq M - v$ and, for some $u > 0$, by the regular value property, $\nabla J'_\infty(\theta) \cdot \boldsymbol{n}_\theta \geq u$ for all $\theta \in \partial K_v$ where $\boldsymbol{n}_\theta$ is the unit normal vector to ∂K_v at θ pointing inwards. Since J'_{ν_t} converges in $\mathcal{C}^1(\mathbb{S}^{p-1})$ towards J'_∞, there exists $t_0 > 0$ such that for all $t \geq t_0$, $\|J'_{\nu_t} - J'_\infty\|_{\mathcal{C}^1(\mathbb{S}^{p-1})} \leq \min\{v, u/2\}$ and thus

$$\forall \theta \in K_v, \quad J'_{\nu_t}(\theta) \geq m - 2v \qquad \text{and} \qquad \forall \theta \in \partial K_v, \quad \nabla J'_{\nu_t}(\theta) \cdot \boldsymbol{n}_\theta \geq u/2.$$

This second property guarantees that no mass leaves K_v due to the divergence term in Eq. (11) for $t \geq t_0$ (equivalently, due to the fact that $\mathbb{R}_+ K_v$ is a positively invariant set of the flow X for $t \geq t_0$). Thus, taking into account the reaction/growth term in Eq. (11), it holds for $t \geq t_0$,

$$\frac{\mathrm{d}}{\mathrm{d}t} \nu_t(K_v) \geq 4 \int_{K_v} J'_{\nu_t} \, \mathrm{d}\nu_t \geq 4(M - 2v)\nu_t(K_v).$$

It follows by Grönwall's lemma that $\nu_t(K_v) \geq \exp(4(M - 2v)t)\nu_{t_0}(K_v)$ for $t \geq t_0$. On the other hand, ν_{t_0} has full support on $\mathbb{S}^{p-1}$ since it can be written as the pushforward of a rescaled version of ν_0 by a diffeomorphism, see Maniglia (2007, Eq. (1.3)) (this is the only place where the assumption on the support is needed). Thus $\nu_{t_0}(K_v) > 0$ and it follows that $\nu_t(\mathbb{S}^{p-1}) \to \infty$. To deal with the case where J'_∞ is constant and is equal to $M > 0$, we can directly take $K = \mathbb{S}^{p-1}$ to show that

$\nu_t(\mathbb{S}^{p-1}) \to \infty$. In the rest of the proof, we show that $(\bar{\nu}_\infty, p(\infty))$ satisfy the optimality conditions of Eq. (4) given by Proposition 12, which we refer to as the *complementary slackness* conditions.

Step 2: complementary slackness (I). We first show that $\min_i \hat{h}_i \to \infty$. Using the property of gradient flows and previously established estimates, we have for $T > 0$,

$$F(\mu_T) - F(\mu_0) = \int_0^T \|\nabla F'_{\mu_t}\|^2 \, \mathrm{d}\mu_t \geq \int_{t_0}^T \int_{K_v} 4|J'_{\nu_t}|^2 \, \mathrm{d}\nu_t \, \mathrm{d}t \geq 4(M - 2v)^2 \int_{t_0}^T \nu_t(K_v) \, \mathrm{d}t \to \infty.$$

Thus $F(\mu_t) \to \infty$ which implies that for all $i \in [n]$, $\ell(-\hat{h}_i(\mu_t)) \to 0$ and thus $\hat{h}_i(\mu_t) \to \infty$. Applying Lemma 15, it follows that $p_{i_0}(\infty) = 0$ for all $i_0 \in \arg\min_i \hat{h}_i(\nu_\infty)$.

Step 3: complementary slackness (II). We now show that $\bar{\nu}_\infty$ is concentrated on $(J'_\infty)^{-1}(M)$, where $\bar{\nu}_t = \nu_t/\|\nu_t\| \in \mathcal{P}(\mathbb{S}^{p-1})$ is the normalized path and $\bar{\nu}_\infty$ its limit. This is immediate if J'_∞ is constant. Otherwise, assuming that $M - 4v$ is also a regular value of J'_∞ and taking a potentially smaller u (which can always be achieved by perturbing v if needed, since regular values are dense as mentioned in Step. 1), it holds for $t \geq t_0$,

$$\frac{\mathrm{d}}{\mathrm{d}t} \nu_t(\mathbb{S}^{p-1} \setminus K_{4v}) \leq 4 \int_{\mathbb{S}^{p-1} \setminus K_{4v}} J'_{\nu_t} \, \mathrm{d}\nu_t \leq 4(m - 3v)\nu_t(K_{4v})$$

using the fact that no mass enters into $\mathbb{S}^{p-1} \setminus K_{4v}$ due to the divergence term in Eq. (11) for $t \geq t_0$. Comparing the rate of growth of the mass in K_v and in $\mathbb{S}^{p-1} \setminus K_{4v}$, we get that $\bar{\nu}_\infty(\mathbb{S}^{p-1} \setminus K_{4v}) \leq \lim_{t\to\infty} \bar{\nu}_t(\mathbb{S}^{p-1} \setminus K_{4v}) = 0$ since $\mathbb{S}^{p-1} \setminus K_{4v}$ is open and by the properties of weak convergence of measures (Portmanteau Theorem). Since this holds for v arbitrarily close to 0, it follows that $\bar{\nu}_\infty$ is concentrated on $(J'_\infty)^{-1}(M)$.

Step 4: conclusion. We have proved the two complementary slackness properties, so by Proposition 12, the pair $(\bar{\nu}_\infty, p(\infty))$ satisfies the optimality conditions, which concludes the proof. ∎

D.2. Proof of Corollary 4

Corollary 11 *Under the assumptions of Theorem 3, assume that the sequence $(w_j(0))_{j\in\mathbb{N}_*}$ is such that $\mu_{0,m}$ converges in $\mathcal{P}_2(\mathbb{R}^p)$ to μ_0. Then, denoting $\bar{\nu}_{m,t} = \Pi_2(\mu_{m,t})/[\Pi_2(\mu_{m,t})](\mathbb{S}^{p-1})$, it holds*

$$\lim_{m,t\to\infty} \left(\min_{i\in[n]} y_i \int \phi(\theta, x_i) \, \mathrm{d}\bar{\nu}_{m,t} \right) = \gamma_1.$$

Proof The fact that

$$\lim_{t\to\infty} \lim_{m\to\infty} \left(\min_{i\in[n]} y_i \int \phi(\theta, x_i) \, \mathrm{d}\bar{\nu}_{m,t}(\theta) \right) = \gamma_1$$

is obtained by combining Theorem 2 with Theorem 3 because $\int \phi(\theta, \cdot)\bar{\nu}_{m,t}$ depends continuously on $\mu_{t,m}$ in $\mathcal{P}_2(\mathbb{R}^p)$ (endowed with the Wasserstein distance W_2), so we only need to prove that limits can be interchanged. We detail the proof for $\ell = \exp$, noticing that we only use the asymptotic behavior of ℓ and ℓ' so it extends to any loss satisfying (A2). Intuitively, in order to prove the other limit, we need to show that if for some (t_0, m_0) the classifier is close to the max-margin classifier, then this remains true for (t, m_0), $t \geq t_0$. First, for $\beta = \|\nu_t\| > 0$, it holds at time t (considering β fixed):

$$\frac{d}{dt} S_\beta(\hat{h}(\bar{\nu}_t)) = \int \|\nabla J'_{\nu_t}\|^2 \, \mathrm{d}\bar{\nu}_t + 4\left(\int |J'_{\nu_t}|^2 \, \mathrm{d}\bar{\nu}_t - \left(\int |J'_{\nu_t}| \, \mathrm{d}\bar{\nu}_t \right)^2 \right) \geq 0.$$

On the other hand, direct computations using the fact that $\nabla S(u) \in \Delta^{n-1}$ leads to

$$\partial_\beta S_\beta(u) = -\frac{1}{\beta^2} \sum_{i=1}^n \nabla_i S(u) \log(n \nabla_i S(u)) \geq -\frac{n}{\beta^2}.$$

Combining both gives the total derivative

$$\frac{\mathrm{d}}{\mathrm{d}t} S_{\|\nu_t\|}(\hat{h}(\bar{\nu}_t)) \geq -n \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{\|\nu_t\|}\right) \qquad \Rightarrow \qquad S_{\|\nu_t\|}(\hat{h}(\bar{\nu}_t)) \geq S_{\|\nu_{t_0}\|}(\hat{h}(\bar{\nu}_{t_0})) - \frac{n}{\|\nu_{t_0}\|}.$$

From the first limit above, for any $\epsilon > 0$, there exists (t_0, m_0) such that for all $m \geq m_0$, $\min_{i \in [n]} y_i \int \phi(\theta, x_i) \, \mathrm{d}\bar{\nu}_{m,t_0}(\theta) \geq \gamma_1 - \epsilon/3$. We have shown in the proof of Theorem 3 that $\|\nu_t\| \to \infty$ so choosing t_0 and m_0 potentially larger if needed, since $\|\nu_{t,m}\| \to \|\nu_t\|$ it also holds $\|\nu_{t_0,m}\| \geq 3n/\epsilon$ for $m \geq m_0$. By the inequality above, we thus have for all $t \geq t_0$ and $m \geq m_0$

$$S_{\|\nu_{t,m}\|}(\hat{h}(\bar{\nu}_{t,m})) \geq S_{\|\nu_{t_0,m}\|}(\hat{h}(\bar{\nu}_{t_0,m})) - \epsilon/3.$$

Moreover, by Lemma 13, we have $|S_{\|\nu_t\|}(u) - \min_i u| \leq \epsilon/3$ for $\|\nu_t\|$ large enough, uniformly for u in a compact set. Hence for all $m \geq m_0$,

$$\lim_{t \to \infty} S_{\|\nu_{t,m}\|}(\hat{h}(\bar{\nu}_{t,m})) \geq S_{\|\nu_{t_0,m}\|}(\hat{h}(\bar{\nu}_{t_0,m})) - \epsilon/3 \geq \gamma_1 - \epsilon.$$

It remains to show that for all $C > 0$, there exists m_0 such that if $m \geq m_0$ then $\liminf_{t \to \infty} \|\nu_{t,m}\| \geq C$, so that we deduce from the above

$$\lim_{t \to \infty} \min_{i \in [n]} \hat{h}(\bar{\nu}_{t,m})[i] \geq \gamma_1 - 2\epsilon.$$

Since ϵ is arbitrary, it would follow that $\lim_{m \to \infty} \lim_{t \to \infty} \min_{i \in [n]} \hat{h}(\bar{\nu}_{\infty,t}) \geq \gamma_1$ which is our claim.

To see this, it is sufficient to notice that since $F(\nu_{t,m})$ is increasing in t for all m, and $S(u) - \log(n) \leq \min_i u$, we have that for all $C > 0$, there exists t_0, m_0 such that $\min_i \hat{h}(\nu_{t,m})[i] \geq C$ for all $t \geq t_0$ and $m \geq m_0$. ∎

D.3. Intermediate results

This section contains intermediate results used in the proof of Theorem 3.

Proposition 12 (Optimality conditions) *The maximization problem* (4) *admits global maximizers $\nu^\star \in \mathcal{M}_+(\mathbb{S}^{p-1})$. Moreover, a measure $\nu^\star \in \mathcal{M}_+(\mathbb{S}^{p-1})$ is a global maximizer of* (4) *if and only if $\nu^\star \in \mathcal{P}(\mathbb{S}^{p-1})$ and there exists $p^\star \in \Delta^{n-1})$ such that (i)* $\mathrm{spt}\, \nu^\star \subset \arg\max_\theta \sum_i p_i y_i \phi(\theta, x_i)$ *and (ii)* $\mathrm{spt}\, p \subset \arg\min_i y_i \int \phi(\theta, x_i) \, \mathrm{d}\nu^\star(\theta)$.

Proof By minimax duality (Sion, 1958), we can rewrite Eq (4) as the minimax problem

$$\sup_{\nu \in \mathcal{P}(\Theta)} \inf_{p \in \Delta^{n-1}} \sum_{i=1}^n p_i y_i \int \phi(\theta, x_i) \, \mathrm{d}\nu(\theta) = \inf_{p \in \Delta^{n-1}} \sup_{\nu \in \mathcal{P}(\Theta)} \sum_{i=1}^n p_i y_i \int \phi(\theta, x_i) \, \mathrm{d}\nu(\theta),$$

and it admits (at least) a saddle point $(\nu^\star, p^\star)$. Moreover, the optimality conditions are necessary and sufficient for the right-hand side to equal the left-hand side. ∎

Let us now prove some useful properties of the function

$$S_\beta(u) = -\frac{1}{\beta} \log \left(\frac{1}{n} \sum_{i=1}^n \ell(-\beta u_i) \right), \tag{13}$$

which is a soft-min function under assumption (A2), as shown in the next lemma.

Lemma 13 *If $\ell(u) \sim \exp(u)$ as $u \to \infty$, and if $\bar{u} \in (\mathbb{R}_+^*)^n$, then*

$$\lim_{\beta \to \infty} S_\beta(\bar{u}) = \min_{i \in [n]} \bar{u}.$$

Proof Let $m := \min_{i \in [n]} \bar{u}_i$. We have $\exp\left(-\beta\left(S_\beta(\bar{u}) - m\right)\right) = \frac{1}{n} \sum_{i=1}^n \ell(-\beta\bar{u}_i) \exp(\beta m)$, where each term satisfies, in the large β regime,

$$\ell(-\beta\bar{u}_i) \exp(\beta m) \sim \exp(-\beta(\bar{u}_i - m)) \to \begin{cases} 1 & \text{if } i \in \arg\min_i \bar{u}_i, \\ 0 & \text{otherwise.} \end{cases}$$

As a consequence, $\exp\left(-\beta\left(S_\beta(\bar{u}) - m\right)\right) \to \frac{1}{n} \# \arg\min_i \bar{u}_i \in \,]0, 1]$. Thus, $S_\beta(\bar{u}) \to m$. ∎

The next lemma is immediate when $\ell = \exp$, but worth detailing for the general case.

Lemma 14 *Under Assumption (A2), let $u(t)$ be a sequence such that $S(u(t))$ is lower bounded and $\nabla S(u(t))$ converges. Then $\lim_{t \to \infty} \nabla S(u(t)) \neq 0$.*

Proof We analyze separately two cases, whether there is $i_0 \in [n]$ such that $u_{i_0}(t)$ is upper bounded or not. In the first case, we have

$$\nabla_{i_0} S(u(t)) = \frac{\ell'(-u_{i_0}(t))}{\sum_i \ell(-u_i(t))} = \frac{1}{n} \ell'(-u_{i_0}(t)) \exp(S(u(t))),$$

which is uniformly lower bounded by a positive constant under (A2) and due to the lower bound on $S(u(t))$ hence $\lim_{t \to \infty} \nabla S_{i_0}(u(t)) \neq 0$. In the other case, up to taking a subsequence (which does not change the limit), we can assume that $u_i(t) \to \infty$ for all $i \in [n]$. Then using the equivalent of ℓ and ℓ' at $-\infty$, we have that $\lim_{t \to \infty} \sum_{i \in [n]} \nabla S(u(t)) = 1$ which is sufficient to conclude. ∎

The next lemma is adapted from Gunasekar et al. (2018a, Lemma 8) and exploits the fact that the gradient of S is a soft-argmax.

Lemma 15 (Convergence of soft-argmin) *Let $\beta : \mathbb{R}_+ \to \mathbb{R}_+$ and $\bar{u} : \mathbb{R}_+ \to \mathbb{R}^n$ be such that $\beta(t) \to \infty$ and $\bar{u}(t) \to \bar{u}(\infty) \in (\mathbb{R}_+^*)^n$ as $t \to \infty$. If $\ell(-u) \sim \ell'(-u) \sim \exp(-u)$ as $u \to \infty$, then for any $i_0 \notin \arg\min_i \bar{u}_i(\infty)$, as $t \to \infty$, it holds $\nabla_{i_0} S_{\beta(t)}(\bar{u}(t)) \to 0$.*

Proof Taking any $\gamma \in \,]\min_i \bar{u}_i(\infty), \bar{u}_{i_0}(\infty)[$, we have

$$\nabla_{i_0} S_{\beta(t)}(\bar{u}(t)) = \frac{\ell'(-\beta\bar{u}_{i_0}(t))}{\sum_{i=1}^n \ell(-\beta\bar{u}_i(t))} = \frac{\ell'(-\beta\bar{u}_{i_0}(t)) \exp(\beta\gamma)}{\sum_{i=1}^n \ell(-\beta\bar{u}_i(t)) \exp(\beta\gamma)}.$$

By the assumption on ℓ', the numerator is equivalent to $\exp(-\beta(\bar{u}_{i_0}(t) - \gamma))$ and goes to 0 as $t \to \infty$. Also, by the assumption on ℓ, each of the term in the denominator is equivalent to $\exp(-\beta(\bar{u}_i(t) - \gamma))$ which goes to ∞ for $i \in \arg\min_i \bar{u}_i(\infty)$, hence the conclusion. ∎

Appendix E. Appendix for Section 4

Let us define

$$G_\beta(a) = -\frac{1}{\beta} \log \left(\frac{1}{n} \sum_{i=1}^{n} \exp \left(-\beta \sum_{j=1}^{m} z_{i,j} a_j \right) \right),$$

which satisfies

$$\min_{i \in [n]} z_i^\top a \leq G_\beta(a) \leq \min_{i \in [n]} z_i^\top a + \frac{\log n}{\beta}. \tag{14}$$

Let us recall Proposition 6 and prove it.

Proposition 16 *Let $a_j(t) = r_j(t)^2/m$ for $j \in [m]$, $\beta(t) = \|a(t)\|_1$ and $\bar{a}(t) = a(t)/\beta(t)$. For the step-sizes $\eta(t) = 1/(16\|z\|_\infty \sqrt{t+1})$ and a uniform initialization $r(0) \propto \mathbf{1}$, it holds*

$$\max_{0 \leq s \leq t-1} \min_i z_i^\top \bar{a}(s) \geq \gamma_1^{(m)} - \frac{\|z\|_\infty}{\sqrt{t}} (8 \log(m) + \log(t) + 1) - \frac{4 \log n}{\sqrt{t}} \sum_{s=0}^{t-1} \frac{1}{\beta(s)\sqrt{s+1}}. \tag{15}$$

where $\gamma_1^{(m)} := \max_{a \in \Delta^{m-1}} \min_{i \in [n]} z_i^\top a$ and the last sum is uniformly bounded in t as soon as $\gamma_1^{(m)} > 0$.

Proof Let us first prove that the normalized dynamics $\bar{a}(t)$ satisfies the (perturbed) *online mirror ascent* recursion (on the simplex, with the entropy mirror map):

$$\begin{cases} b_j(t+1) = \bar{a}_j(t) \left(1 + 4\eta(t)\nabla_j G_{\beta(t)}(\bar{a}(t)) + 4\eta(t)^2 |\nabla_j G_{\beta(t)}(\bar{a}(t))|^2 \right), \ \forall j \in [m] \\ \bar{a}(t+1) = b(t+1)/\|b(t+1)\|_1. \end{cases}$$

We mention that it is perturbed because for the plain *online mirror ascent*, the multiplicative term in the first line would be $\exp(4\eta(t)\nabla G_{\beta(t)}(\bar{a}(t))$, so we have second order corrections in $\eta(t)$.

Since $\nabla_j F(r(t)) = (2/m)r_j(t)\nabla G_1(a(t))$, it follows

$$\begin{aligned} a_j(t+1) &= (r_j(t) + 2\eta(t)r_j(t)\nabla_j G_1(a(t)))^2 \\ &= r_j(t)^2 + 4\eta(t)r_j(t)^2 \nabla_j G_1(a(t)) + 4\eta(t)^2 r_j(t)^2 |\nabla_j G_1(a(t))|^2 \\ &= a_j(t) \left(1 + 4\eta(t)\nabla_j G_1(a(t)) + 4\eta(t)^2 |\nabla_j G_1(a(t)|^2 \right). \end{aligned}$$

Now using the fact that $\nabla G_1(a) = \nabla G_{\|a\|_1}(a/\|a\|_1)$, it follows

$$\bar{a}_j(t+1) = \frac{\|a(t)\|_1}{\|a(t+1)\|_1} \bar{a}_j(t) \left(1 + 4\eta(t)\nabla_j G_{\beta(t)}(\bar{a}(t)) + 4\eta(t)^2 |\nabla_j G_{\beta(t)}(\bar{a}(t)|^2 \right),$$

hence the iterations. Let us rewrite these iterations using the framework of Bregman divergences (see Bubeck (2015, Chap. 4) for details). For $a, b \in \mathbb{R}_+^m$, let $\phi(a) = \sum_{i=1}^m a_j \log(a_j) - a_j$, let $D(a, b) = \phi(a) - \phi(b) - \nabla\phi(b)^\top(a - b)$ and let $\Pi(a) = a/\|a\|_1 = \arg\min_{b \in \Delta^{m-1}} D(b, a)$ be the Bregman projection on the simplex for the divergence D. With $g(s) = \nabla G_{\beta(s)}(\bar{a}(s))$, we have

$$\begin{cases} \nabla\phi(b(t+1)) = \nabla\phi(\bar{a}(t)) + \eta(t)g(t) + \eta_t^2 e(t) \\ \bar{a}(t+1) = \Pi(b(t+1)), \end{cases}$$

which are the online mirror ascent updates, with a second-order error term $e(t)_j = \log(1 + 4\eta(s)g_j(t) + 4\eta(t)^2g_j(t)^2) - 4\eta(t)g_j(t)$. Notice that $\forall t, \|g(t)\|_\infty \leq \|z\|_\infty$ so if we assume that $\eta(t) \leq 1/(16\|z\|_\infty)$ we have $\eta(t)\|g(t)\|_\infty \leq 1/16$. Using the inequality $|\log(1+u) - u| \leq u^2$ for $|u| \leq 1/2$, we get by applying it with $u_j = 4\eta(t)g_j(t) + 4\eta(t)^2g_j(t)^2$,

$$\|e(s)\|_\infty \leq (4\eta(t)\|z\|_\infty + 4\eta(t)^2\|z\|_\infty^2)^2 + 4\eta(t)^2\|z\|_\infty^2 \leq 23\eta(t)^2\|z\|_\infty^2.$$

We now follow the usual proof of mirror ascent from Bubeck (2015, Thm. 4.2) (or Beck and Teboulle (2003) for the variable step-size case) and including this error term leads to, for all $\bar{a}^* \in \Delta^{m-1}$,

$$4\eta(t)g(t)^\top(\bar{a}^* - \bar{a}(t)) \leq D(\bar{a}^*, \bar{a}(t)) - D(\bar{a}^*, \bar{a}(t+1)) + 24\eta(t)^2\|z\|_\infty^2.$$

We get a telescopic sum and using the concavity of each G_β,

$$S(t) := \sum_{s=0}^{t-1} \eta(s)(G_{\beta(s)}(\bar{a}^*) - G_{\beta(s)}(\bar{a}(s))) \leq \frac{1}{4}D(\bar{a}^*, \bar{a}(0)) + 6\|z\|_\infty^2 \sum_{s=0}^{t-1} \eta(s)^2.$$

With our choice of initialization, $D(\bar{a}^*, \bar{a}(0)) \leq \log(m)$. Let us choose $\eta(t) = \tau/\sqrt{t+1}$. Using the inequalities

$$\sum_{s=0}^{t-1} \frac{1}{\sqrt{s+1}} \geq \int_1^{t+1} \frac{ds}{\sqrt{s}} = 2\sqrt{t+1} - 2.$$

and

$$\sum_{s=0}^{t-1} \left(\frac{1}{\sqrt{s+1}}\right)^2 = \sum_{s=1}^{t} \frac{1}{s} \leq 1 + \int_{s=1}^{t} \frac{ds}{s} = 1 + \log(t).$$

It follows that for all $t \geq 1$,

$$S(t) := \frac{\sum_{s=0}^{t-1} \eta(s)(G_{\beta(s)}(\bar{a}^*) - G_{\beta(s)}(\bar{a}(s)))}{\sum_{s=0}^{t-1} \eta(s)} \leq \frac{\log(n)/4 + 6\tau^2\|z\|_\infty^2(1 + \log(t))}{2\tau(\sqrt{t+1} - 1)}.$$

In particular, with the choice $\tau = 1/(16\|z\|_\infty)$, we get

$$S(t) \leq \frac{\|z\|_\infty}{\sqrt{t+1} - 1}\left(2\log(m) + (1 + \log(t))/4\right) \leq \frac{\|z\|_\infty}{\sqrt{t}}(8\log(m) + \log(t) + 1).$$

where we used $\sqrt{t}/4 \leq \sqrt{t+1} - 1$ to simplify the expression. Finally, using inequality (14), we have

$$G_{\beta(s)}(\bar{a}^*) - G_{\beta(s)}(\bar{a}(s)) \geq \min_i z_i^\top \bar{a}^* - \min_i z_i^\top \bar{a}(s) - \frac{\log n}{\beta(s)}.$$

Taking the weighted sum gives

$$\gamma_1^{(m)} - \max_{0 \leq s \leq t-1} \min_i z_i^\top \bar{a}(s) \leq S(t) + \frac{\sum_{s=0}^{t-1} \eta(s)(\log n)/\beta(s)}{\sum_{s=0}^{t-1} \eta(s)}$$

$$\leq \frac{\|z\|_\infty}{\sqrt{t}}(8\log(m) + \log(t) + 1) + \frac{4\log n}{\sqrt{t}}\sum_{s=0}^{t-1} \frac{1}{\beta(s)\sqrt{s+1}}.$$

25

The conclusion follows by Lemma 17. ∎

In the next result, we show that the norm of the iterates grows to $+\infty$ and that $\sum_{s=0}^{t-1} \frac{1}{\beta(s)\sqrt{s+1}}$ is finite. For simplicity, we do not track the constants.

Lemma 17 *Under the assumptions of Proposition 6, we have that $\beta(t) \to \infty$ and $\sum_{s=0}^{t-1} \frac{1}{\beta(s)\sqrt{s+1}}$ is bounded uniformly in t.*

Proof For this result, we look at a different online mirror ascent dynamics. We consider $\alpha > 0$ (to be chosen appropriately later) and define $\tilde{\beta}(t) = \max\{1, \max_{0 \le s \le t}\{\beta(s)/\alpha\}\}$ and the iterates $\tilde{a}(t) = a(t)/\tilde{\beta}(t)$. With the same arguments than in the proof of Proposition 7, it can be seen that those iterates satisfy the recursion, with $g(t) = \nabla G_{\tilde{\beta}(t)}(\tilde{a}(t))$,

$$\begin{cases} b_j(t+1) = \tilde{a}_j(t)(1 + 4\eta(t)g_j(t) + 4\eta(t)^2 g_j(t)^2). \\ \tilde{a}(t+1) = b(t+1)/\|b(t+1)\|. \end{cases}$$

These are (perturbed) online mirror ascent iterates for the sequence of losses $G_{\tilde{\beta}(t)}$, step-sizes η_t on the set $\alpha B_+^1 = \{\tilde{a} \in \mathbb{R}^m_+ ; \sum_j \tilde{a}_j \le \alpha\}$. Note that the entropy $\phi(s) = s \log s - s + 1$ is $1/\alpha$ strongly convex with respect to $\|\cdot\|_1$ on this set, and that $\|g\|_\infty$ is bounded uniformly in α. We have the usual mirror descent bound (Chap. 4 Bubeck, 2015) with $\bar{a}^*$ the ℓ_1-max-margin solution,

$$\eta(t)g(s)^\top(\alpha\bar{a}^* - \tilde{a}(t)) \le H(\alpha\bar{a}^*, \tilde{a}(t+1)) - H(\alpha\bar{a}^*, \tilde{a}(t)) + \alpha C\eta(t)^2$$

where C only depend on $\|z\|_\infty$. By summing we get

$$S_\alpha(t) := \frac{\sum_{s=0}^{t-1}\eta(s)g(s)^\top(\alpha\bar{a}^* - \tilde{a}(s))}{\sum_{s=0}^{t-1}\eta(s)} \le \frac{H(\alpha\bar{a}^*, \tilde{a}(0)) + C\alpha\sum_{s=0}^{t-1}\eta(s)^2}{\sum_{s=0}^{t-1}\eta(s)}$$

$$\lesssim \frac{\alpha\log(\alpha) + \log(t)}{\sqrt{t+1} - 1},$$

where we only track the dependency in t and α. On the other hand, using inequality (14),

$$S_\alpha(t) \ge \frac{\sum_{s=0}^{t-1}\eta(s)(G_{\tilde{\beta}(s)}(\alpha\bar{a}^*) - G_{\tilde{\beta}(s)}(\tilde{a}(s)))}{\sum_{s=0}^{t-1}\eta(s)}$$

$$\ge \alpha\gamma_1^{(m)} - \log(n)\Big(\frac{\sum_{s=0}^{t-1}\eta(s)/\tilde{\beta}(s)}{\sum_{s=0}^{t-1}\eta(s)}\Big) - \gamma_1^{(m)}\Big(\frac{\sum_{s=0}^{t-1}\eta(s)\|\tilde{a}(s)\|_1}{\sum_{s=0}^{t-1}\eta(s)}\Big).$$

Thus, using the fact that $\|\tilde{a}(s)\|_1 \le \|a(s)\| = \beta(s)$, and that $\tilde{\beta}(s) \ge 1$, it follows

$$\frac{\sum_{s=0}^{t-1}\eta(s)\beta(s)}{\sum_{s=0}^{t-1}\eta(s)} \ge \alpha - \frac{\log n}{\gamma_1^{(m)}} - S_\alpha(t).$$

As a consequence

$$\sum_{s=0}^{t-1}\frac{\beta(s)}{\sqrt{s+1}} \gtrsim \sqrt{t+1}\Big(\alpha - \frac{\alpha\log(\alpha) + \log(t)}{\sqrt{t+1} - 1}\Big).$$

Taking for instance $\alpha = \sqrt{t}$ shows that $\sum_{s=0}^{t-1} \frac{\beta(s)}{\sqrt{s+1}} \gtrsim t - \log(t)$ and thus $\beta(t) \to \infty$. By Lemma 18 then $\beta(t)$ is increasing for $t \geq t_0$ and grows to ∞ at a super-polynomial rate and the conclusion follows. $\blacksquare$

We now prove the asymptotic rate of growth of the norm of the iterates.

Lemma 18 *If $\eta(t) \asymp 1/\sqrt{t+1}$ and $\beta(t) \to \infty$, then $\beta(t)$ is increasing for t large enough and*

$$\log(\beta(t)) \gtrsim \min\{1, \gamma_1^{(m)}\}\sqrt{t}.$$

Proof We have for all t,

$$\frac{\beta(t+1)}{\beta(t)} = 1 + \eta(t)\nabla S(Za(t))^\top Z\bar{a}(t) + O(1/(t+1)).$$

By Eq. (15) (which holds irrespective of this lemma), we know that $\beta(t) \to \infty$ implies $\bar{a}(t) \to \bar{a}^*$ where $\bar{a}^*$ is the ℓ_1-max-margin solution. Since $\nabla S(Za(t)) \in \Delta^{n-1}$, it holds for t large enough

$$\frac{\beta(t+1)}{\beta(t)} \geq 1 + \frac{1}{2}\gamma_1^{(m)}\eta(t) + O(1/(t+1)).$$

Taking the logarithm and summing, we get

$$\log(\beta(t)) - \log(\beta(0)) \geq \sum_{s=0}^{t-1}\left(\frac{1}{2}\gamma_1^{(m)}\eta(s) + O(1/s)\right) = \frac{1}{2}\gamma_1^{(m)}\sum_{s=0}^{t-1}\eta(s) + O(\log(t)).$$

The result follows since $\sum_{s=0}^{t-1}(t+1)^{-1/2} \geq 2(\sqrt{t-1} - 1)$. $\blacksquare$

Appendix F. Appendix to Section 5

Let us recall Proposition 7 and prove it.

Proposition 19 *Let $a(t) = r(t)/m$, $\beta(t) = \max\{1, \max_{0 \leq s \leq t} \sqrt{m}\|a(t)\|_2\}$ and $\bar{a}(t) = a(t)/\beta(t)$ and assume that $\gamma_2^{(m)}$ is positive. For the step-sizes $\eta(t) = \beta(t)\sqrt{2}/(\|z\|_\infty\sqrt{t+1})$ and initialization $r(0) = 0$, it holds*

$$\max_{0 \leq s \leq t-1} \min_{i \in [n]} z_i^\top \bar{a}(s) \geq \gamma_2^{(m)} - \frac{\|z\|_\infty}{\sqrt{t}}\left(2\sqrt{2} + \frac{\sqrt{3}\log n}{\gamma_2^{(m)}}\right).$$

Proof Using the fact that $a(t) = r(t)/m$ and $m\nabla F(r) = \nabla G(r/m)$ we have

$$a(t+1) = a(t) + \frac{\eta(t)}{m}\nabla G(a(t)).$$

It follows that

$$\frac{a(t+1)}{\beta(t)} = \bar{a}(t) + \frac{\eta(t)}{m\beta(t)}\nabla G_{\beta(t)}(\bar{a}(t)) =: b(t+1).$$

and thus

$$\bar{a}(t+1) = \frac{\beta(t)}{\beta(t+1)}\frac{a(t+1)}{\beta(t)} = \frac{\beta(t)}{\beta(t+1)}b(t+1).$$

Finally, since $\beta(t)\max\{1, \sqrt{m}\|b(t+1)\|_2\} = \max\{\beta(t), \sqrt{m}\|a(t+1)\|_2\} = \beta(t+1)$ it follows that $\bar{a}(t+1) = b(t+1)/\max\{1, \sqrt{m}\|b(t+1)\|_2\}$. Thus $\bar{a}(t)$ follows the iterations $\bar{a}(0) = 0$ and

$$\begin{cases} b(t+1) = \bar{a}(t) + \dfrac{\eta(t)}{m\beta(t)}\nabla G_{\beta(t)}(\bar{a}(t)) \\ \bar{a}(t+1) = b(t+1)/\max\{1, \sqrt{m}\|b(t+1)\|_2\}. \end{cases}$$

These are *online projected gradient ascent* iterations on the set $\{a \in \mathbb{R}^m \ ; \ \sqrt{m}\|a\|_2 \le 1\}$ and for the sequence of functions $G_{\beta(t)}$, with step-size $\eta(t)/(m\beta(t))$. Using the fact that the Lipschitz constant of G_β if upper bounded by $\max_{i\in[n]}\|z_i\|_2 \le \sqrt{m}\|z\|_\infty$ and the diameter of the constraint set is $2/\sqrt{m}$, we have the classical bound, with the step-size $\eta(t)/(m\beta(t)) = \sqrt{2}/(m\|z\|_\infty\sqrt{t+1})$,

$$\frac{1}{t}\sum_{s=0}^{t-1}\left(G_{\beta(s)}(\bar{a}^*) - G_{\beta(s)}(\bar{a}(s))\right) \le \frac{1}{t}DL\sqrt{2t} = \frac{2\sqrt{2}\|z\|_\infty}{\sqrt{t}}.$$

Now using the bound of Eq. (14), it follows

$$\max_{0\le s\le t-1}\min_{i\in[n]}z_i^\top\bar{a}(s) \ge \gamma_2^{(m)} - \frac{\log n}{t}\sum_{s=0}^{t-1}\frac{1}{\beta(s)} - \frac{2\sqrt{2}\|z\|_\infty}{\sqrt{t}}.$$

Let us now look at the evolution of $\beta(t)$. Let $\bar{a}^*$ be a max ℓ_2-margin solution and remark that

$$a(t+1)^\top\bar{a}^* - a(t)^\top\bar{a}^* = \eta(t)\nabla G(a(t))^\top\bar{a}^* \ge \eta(t)\gamma_2^{(m)}$$

since $\nabla G(a(t)) \in \Delta^{m-1}$. It follows that $\beta(t) \ge \sqrt{m}\|a(t)\|_2 \ge m\gamma_2^{(m)}\sum_{s=0}^{s-1}\eta(s)$. Using the fact that $\beta(t) \ge C\sum_{s=0}^{s-1}\beta(s)/\sqrt{s+1}$ with $C = \gamma_2^{(m)}\sqrt{2}/\|z\|_\infty$ and the bound $\sum_{s=0}^{s-1}1/\sqrt{s+1} \ge 2(\sqrt{t-1}-1)$, it follows that $\beta(t) \ge 2C\sqrt{t+1}/\sqrt{6}$ and thus

$$\sum_{s=0}^{t-1}\frac{1}{\beta(s)} \le \frac{\sqrt{6}}{2C}\sum_{s=0}^{t-1}\frac{1}{\sqrt{s+1}} \le \frac{\sqrt{6t}}{C}.$$

Plugging into the previous bound gives the conclusion. Note that we did not attempt to make the lower bound on $\beta(t)$ tight. ∎

Appendix G. Appendix to Section 6: generalization bounds

With the notations of Section 6, let us first lower bound the margins in $\mathcal{F}_1$ and in $\mathcal{F}_2$.

Lemma 20 *Assume that $\|x_i\|_2 \le R$ for $i \in [n]$. For any $\epsilon \in (0,1)$ and $r \in [d]$, there exists $C(r), C_\epsilon(r) > 0$ such that*

$$\gamma_2 \ge \min\left\{C(d), C_\epsilon(d)\left(\frac{\Delta_d(S_n)}{R}\right)^{\frac{d+3}{2-\epsilon}}\right\} \quad and \quad \gamma_1 \ge \min_{r\in[d]}\min\left\{C(r), C_\epsilon(r)\left(\frac{\Delta_r(S_n)}{R}\right)^{\frac{r+3}{2-\epsilon}}\right\}.$$

Proof Let $\text{dist}_{\mathcal{S}}$ be the distance function to a set $\mathcal{S}$, i.e. $\text{dist}_{\mathcal{S}}(x) = \inf_{\tilde{x} \in \mathcal{S}} \|x - \tilde{x}\|_2$, which is 1-Lipschitz, and let $D_{\pm} = \{x_i \ ; \ y_i = \pm 1\}$. For P_r a projection that achieves the supremum in Eq. (10) (which exists by compactness of Grassmannians and continuity of the objective), we consider the following function

$$f_r(x) = 2 \max \left(0, 1 - \frac{2 \, \text{dist}_{P_r(D_+)}(P_r(x))}{\Delta_r(S_n)} \right) - 2 \max \left(0, 1 - \frac{2 \, \text{dist}_{P_r(D_-)}(P_r(x))}{\Delta_r(S_n)} \right).$$

This function is $4/\Delta_r(S_n)$-Lipschitz continuous, satisfies $\|f\|_{\infty} \leq 2$ and $y_i f(x_i) = 2$ for all $i \in [n]$. Let us first consider the case $r = d$. Using the approximation results of Lipschitz functions in $\mathcal{F}_2$ from Bach (2017a, Prop. 6), we know that if $N > 0$ is larger than a constant independent of $\Delta_d(S_n)$ and satisfies

$$C(d)\eta(N/\eta)^{-2/(d+1)} \log(N/\eta) \leq 1,$$

where $\eta = \max\{2, 4R/\Delta_d(S_n)\} = 4R/\Delta_d(S_n)$, then there exists $\hat{f}$ such that $\|\hat{f}\|_{\mathcal{F}_2} \leq N$ and $\sup_{\|x\|_2 \leq R} |\hat{f}(x) - f_d(x)| \leq 1$. Since $\hat{f}/N$ is feasible for the $\mathcal{F}_2$-max-margin problem Eq. (5), this shows that $\gamma_2 \leq 1/N$ and it remains to estimate how large N must be. In the next computations, the dimension dependent constant $C(d)$ might change from line to line. Using the bound $\log(u) \leq C(\epsilon, d)u^{\epsilon/(d+1)}$ for $\epsilon > 0$, we obtain the stronger condition on N:

$$C(\epsilon, d)\eta(N/\eta)^{(\epsilon-2)/(d+1)} \leq 1 \ \Leftrightarrow \ N \geq C(\epsilon, d)\eta^{(d+3-\epsilon)/(2-\epsilon)} \leq C(\epsilon, d) \left(\frac{\Delta_d(S_n)}{R} \right)^{(d+3)/(2-\epsilon)}.$$

This gives the bound on γ_2. For the bound on γ_1, it follows from the fact that for all $r \in [d]$, $\mathcal{F}_1$ contains the functions of the form $f \circ P_r$ where f belongs to the space $\mathcal{F}_2$ over $\mathbb{R}^r$ and $\|f \circ P_r\|_{\mathcal{F}_1} \leq \|f\|_{\mathcal{F}_2}$, see arguments and details in Bach (2017a, Section 4.5). ■

Theorem 21 (Generalization bound) *For any $\epsilon \in (0,1)$ and $r \in [d]$, there exist $C(r), C_\epsilon(r) > 0$ such that the following holds. If $(x, y) \sim \mathbb{P}$ is such that for some $R > 0$ and $0 < \Delta_r(\mathbb{P}) \leq C(r)$, it holds $\Delta_r(S_n) \leq \Delta_r(\mathbb{P})$ and $\|x\|_2 \leq R$ almost surely, then for f the $\mathcal{F}_1$-max-margin classifier, it holds with probability at least $1 - \delta$ over the choice of i.i.d. samples $S_n = (x_i, y_i)_{i=1}^n$,*

$$\mathbb{P}[yf(x) < 0] \leq \frac{C_\epsilon(r)}{\sqrt{n}} \left(\frac{R}{\Delta_r(\mathbb{P})} \right)^{\frac{r+3}{2-\epsilon}} + \sqrt{\frac{\log(B)}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}},$$

where $B = \log_2(4(R+1)C_2(r)) + (r+2)\log_2(R/\Delta_r(\mathbb{P}))$. The same bound holds for the $\mathcal{F}_2$-max-margin classifier for $r = d$.

Proof This is a direct application of the margin-based generalization bounds of Theorem 22, using that for any $f \in \mathcal{F}_1$ with $\|f\|_{\mathcal{F}_1} \leq 1$,

$$\sup_{\|x\|_2 \leq R} f(x) \leq \sup_{\|x\|_2 \leq R, \theta \in \mathbb{S}^{p-1}} \phi(\theta, x) \leq R + 1,$$

and the same holds in $\mathcal{F}_2$ since for $g \in \mathcal{F}_2$ it holds $g \in \mathcal{F}_1$ and $\|g\|_{\mathcal{F}_1} \leq \|g\|_{\mathcal{F}_2}$ by Jensen's inequality. We also use the Rademacher complexity bound $\text{Rad}_n(B_2) \leq \text{Rad}_n(B_1) \leq \frac{1}{\sqrt{n}}$. where B_i is the unit ball in $\mathcal{F}_i$. This can be found for instance in Bach (2017a, Prop. 7). ■

In the next theorem, $\mathcal{F}$ refers to a hypothesis class, $\text{Rad}_n(\mathcal{F})$ to its Rademacher complexity and γ to its margin over the training set (see the cited reference for definitions).

Theorem 22 (Koltchinskii and Panchenko (2002)) *Assume that $\forall f \in \mathcal{F}$ we have $\sup_x |f(x)| \leq C$. Then, with probability at least $1 - \delta$ over the sample, for all margins $\gamma > 0$ and all $f \in \mathcal{F}$ we have*

$$\mathbb{P}[yf(x) < 0] \leq 4\frac{\mathrm{Rad}_n(\mathcal{F})}{\gamma} + \sqrt{\frac{\log(\log_2 \frac{4C}{\gamma})}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

Appendix H. Proof for the ReLU case

In this appendix, we detail how to rigorously cover the case of ReLU networks, i.e. models as in Eq. (2) with a feature function of the form

$$\phi(w, x) = b\,(a^\top(x, 1))_+,$$

where $w = (a, b) \in \mathbb{R}^{d+1} \times \mathbb{R}$ (thus $w \in \mathbb{R}^p$ with $p = d + 2$) and $(u)_+ = \max\{0, u\}$. The difficulty with that case is that this function is not differentiable in w: (i) when $a = 0$ for all x, or (ii) whenever $a^\top(x, 1) = 0$. We resolve these two sources of non-differentiability as follows: for (i), we consider a specific initialization that guarantees that a does not vanish along the dynamics and for (ii), we consider an input distribution ρ without atoms, in contrast to the empirical distribution $\rho = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ that is implicitly used in the rest of the paper (as in Eq. (6)).

Assumption on the input. We assume that the input distribution has the following properties:

(A4) The input distribution $\rho \in \mathcal{P}(\mathbb{R}^d)$ has a bounded density and a compact support denoted $\mathcal{X}$. The labels are given by a continuous deterministic function $y : \mathcal{X} \to \mathcal{Y} = \{-1, 1\}$.

The assumption on ρ excludes discrete measures. Also, the continuity assumption on y means that the sets where $y = +1$ and where $y = -1$ are disconnected, and this implies that they are at a positive distance from each other since the level-sets $\{y = 1\}$ and $\{y = -1\}$ are both compact and with empty intersection. This distribution ρ could be for instance the population distribution of the input, or also could be obtained by taking the expectation over small perturbations of the input training set, which is a well-known smoothing technique (e.g., Duchi et al. (2012)). This leads to the definition of the *population smooth-margin*

$$S(f) = -\log\left(\int_{\mathcal{X}} \exp(-f(x))\,\mathrm{d}\rho(x)\right)$$

defined for $f \in \mathcal{C}(\mathcal{X})$, where $\mathcal{C}(\mathcal{X})$ is the space of continuous and real-valued functions on $\mathcal{X}$ endowed with the supremum norm. Let us give some facts about this function.

Lemma 23 *The function $S : \mathcal{C}(\mathcal{X}) \to \mathbb{R}$ is Fréchet differentiable with a gradient at $f \in \mathcal{C}(\mathcal{X})$ given by $\nabla S[f] \in \mathcal{P}(\mathcal{X})$:*

$$\nabla S[f](\mathrm{d}x) = \frac{\exp(-f(x))\rho(\mathrm{d}x)}{\int_{\mathcal{X}} \exp(-f(x'))\,\mathrm{d}\rho(x')}.$$

Also, as a function $\mathcal{C}(\mathcal{X}) \to L^1(\rho)$, the function $f \mapsto \frac{\mathrm{d}\nabla S[f]}{\mathrm{d}\rho}$ is Lipschitz continuous on bounded sets in $\mathcal{C}(\mathcal{X})$.

Proof Let $\ell = \exp$. Since ℓ' is locally Lipschitz continuous, for any $f, \tilde{f} \in \mathcal{C}(\mathcal{X})$ there exists $L > 0$ such that $|\ell(-(f(x) + \epsilon \tilde{f}(x))) - \ell(-f(x)) + \epsilon \ell'(-f(x)) \tilde{f}(x)| \leq (L^2/2)\epsilon^2 \|\tilde{f}\|_\infty$. Thus when $\epsilon \to 0$, we have $\sup_{x \in \mathcal{X}} |\epsilon^{-1}(\ell(-(f(x) + \epsilon \tilde{f}(x))) - \ell(-f(x))) + \ell'(-f(x))\tilde{f}(x)| \to 0$ which shows that $f \mapsto \int_{\mathcal{X}} \exp(-f) \, d\rho$ is Fréchet differentiable with gradient $-\exp(-f) \, d\rho \in \mathcal{M}(\mathcal{X})$. The differentiability of S follows by composition with $-\log$. The Lipschitz continuity of $d\nabla S / d\rho$ on bounded sets is a consequence of the Lipschitz continuity of $\exp$ and the existence of positive lower bounds for $\exp$ on bounded intervals. ∎

Lemma 24 (Convergence of soft-argmin) *Let $\rho \in \mathcal{P}(\mathbb{R}^d)$ have a compact support $\mathcal{X}$. Let $\beta_t > 0$ and $f_t \in \mathcal{C}(\mathcal{X})$ be sequences such that $\beta_t \to \infty$ and $f_t \to f_\infty \in \mathcal{C}(\mathcal{X})$ (uniformly) as $t \to \infty$. If $\lambda_t = \nabla S[\beta_t f_t]$ converges weakly to some $\lambda \in \mathcal{P}(\mathcal{X})$, then $\mathrm{spt}(\lambda) \subset \arg\min_{x \in \mathcal{X}} f_\infty(x)$.*

Proof Let $x^*, x_0 \in \mathcal{X}$ be such that $x^* \in \arg\min_{x \in \mathcal{X}} f_\infty$ and $x_0 \notin \arg\min f_\infty$. By continuity of f_∞ and uniform convergence, there exists $\epsilon, \delta, t_0 > 0$ such that $\forall t \geq t_0$, it holds $f_t(x) \geq f(x^*) + \epsilon$ for all $x \in B_\delta(x_0)$ and $f_t(x) \leq f(x^*) + \epsilon/2$ for all $x \in B_\delta(x^*)$, where $B_\delta(x)$ is the open ball of radius δ centered at x. It follows

$$\lambda_t(B_\delta(x_0)) = \frac{\int_{B_\delta(x_0)} \exp(-\beta_t f_t(x)) \, d\rho(x)}{\int_{\mathcal{X}} \exp(-\beta_t f_t(x)) \, d\rho(x)} \leq \frac{\exp(-\beta(f(x^*) + \epsilon))\rho(B_\delta(x_0))}{\exp(-\beta(f(x^*) + \epsilon/2))\rho(B_\delta(x^*))} \to 0.$$

This shows that $x_0 \notin \mathrm{spt}(\lambda)$. Since this is true for any $x_0 \notin \arg\min f_\infty$, it follows that $\mathrm{spt}(\lambda) \subset \arg\min_{x \in \mathcal{X}} f_\infty(x)$. ∎

Now, we define the objective

$$F(\mu) = S(\hat{h}(\mu))$$

where $\hat{h}(\mu) \in \mathcal{C}(\mathcal{X})$ is the function $x \mapsto y(x)h(\mu, x)$. Seing F as a function on unnormalized measures, it admits a Fréchet differential at $\mu \in \mathcal{P}_2(\mathbb{R}^p)$, represented by $F'_\mu \in \mathcal{C}(\mathbb{R}^p)$ given by

$$F'_\mu(w) = \int_{\mathcal{X}} y(x)\phi(w, x) \, d(\nabla S[\hat{h}(\mu)])(x).$$

Our next step is to gather some regularity properties of F'_μ. Let us consider the sets

$$D := \{(a, b) \in \mathbb{R}^{p-1} \times \mathbb{R} \; ; \; \|a\| = |b|\} \qquad \text{and} \qquad \mathbb{S}_\pm := \mathbb{S}_+ \cup \mathbb{S}_- := \mathbb{S}^{p-1} \cap D. \qquad (16)$$

We endow $\mathbb{S}_\pm$ with its Riemannian geometry inherited from the sphere (it is a disconnected manifold with two connected components $\mathbb{S}_+$ and $\mathbb{S}_-$) and let us denote J'_μ the restriction of F'_μ to $\mathbb{S}_\pm$.

Lemma 25 *Under Assumption (A4) for all $\mu \in \mathcal{P}_2(\mathbb{R}^p)$, the function F'_μ is differentiable on $\{(a, b) \in \mathbb{R}^p \; ; \; a \neq 0\}$. Also let $A \subset \mathcal{P}_2(\mathbb{R}^p)$ be a set of measures with uniformly bounded second moments. Then $\nabla J'_\mu$ is Lipschitz continuous on $\mathbb{S}_\pm$, uniformly on A and the function $J' : \mathcal{P}_2(\mathbb{R}^p) \to \mathcal{C}^1(\mathbb{S}_\pm)$ is Lipschitz continuous on A.*

Proof Let us denote $\lambda_\mu = \nabla S[\hat{h}(\mu)]$, let $w, \bar{w} \in \mathbb{R}^p$ where $w = (a, b)$ with $a \neq 0$ and $V \subset \mathbb{R}^p$ be a convex bounded open set that contains w and $\bar{w}$. Since the function $w \mapsto y(x)\phi(w, x)$ is Lipschitz on V, uniformly in $x \in \mathcal{X}$, there exists a constant $L > 0$ such that $\epsilon^{-1}|\phi(w + \epsilon \bar{w}, x) - \phi(w, \epsilon)| \leq L$ for

ϵ small enough. Moreover, if $x \notin H_a := \{x \in \mathcal{X} \; ; \; a^\top x = 0\}$, then $\epsilon^{-1}(\phi(w + \epsilon\bar{w}, x) - \phi(w, \epsilon)) \to \nabla_w\phi(w, x)^\top\bar{w}$ as $\epsilon \to 0$. Since $\rho(H_a) = 0$ it follows by the dominated convergence theorem that

$$\lim_{\epsilon \to 0} \int_{\mathcal{X}} |\epsilon^{-1}y(x)(\phi(w + \epsilon\bar{w}, x) - \phi(w, \epsilon)) - y(x)\nabla_w\phi(w, x)^\top\bar{w}| \, d\lambda_\mu(x) = 0,$$

which shows that F'_μ is differentiable with the gradient $w \mapsto \int_{\mathcal{X}} y(x)\nabla_w\phi(w, x) \, d\lambda_\mu(x)$ if $a \neq 0$.

Let us now show that $\nabla F'_\mu$ is Lipschitz continuous on $\mathbb{S}_\pm$, which implies the same for $\nabla J'_\mu$. Since this is immediate for the component ∇_b so let us focus on ∇_a. For $(a, b), (\bar{a}, b) \in \mathbb{S}_+$, defining $H_{a,\bar{a}} = \{x \in \mathcal{X} \; ; \; (a^\top x)^0_+ \neq (\bar{a}^\top x)^0_+\}$, we have

$$\|\nabla_a F'_\mu(a, b) - \nabla_a F'_\mu(\bar{a}, b)\| \leq 2^{-1/2} \sup_{x \in \mathcal{X}} \|(x, 1)\| \lambda_\mu(H_{a,\bar{a}}).$$

But since the Lebesgue measure of $\{x \in \mathcal{X} \; ; \; (x, 1) \in H_{a,\bar{a}}\}$ is bounded by a constant times $\text{dist}(a, \bar{a})$ and λ_μ has a bounded density with respect to the Lebesgue measure, it follows that for some constant L, $\lambda_\mu(H_{a,\bar{a}}) \leq L \, \text{dist}(a, \bar{a})$. Moreover this constant L is uniform over measures λ_μ with a bounded density with respect to ρ. But if $A \subset \mathcal{P}_2(\mathbb{R}^p)$ is such that for some $C > 0$, $\int \|w\|^2_2 \, d\mu < C$ for all $\mu \in A$, then $\{\lambda_\mu \; ; \; \mu \in A\} \subset \mathcal{P}(\mathcal{X})$ is a set of measures with uniformly bounded densities with respect to ρ, so L is uniform over A.

For the last claim, notice that $\hat{h} : \mathcal{P}_2(\mathbb{R}^p) \to \mathcal{C}(\mathcal{X})$ is Lipschitz continuous and that $(d\nabla S/d\rho) : \mathcal{C}(\mathcal{X}) \to L^1(\rho)$ is Lipschitz continuous, by Lemma 23. It follows that $\mu \mapsto (d\lambda_\mu/d\rho)$ is Lipschitz continuous as a function $\mathcal{P}_2(\mathbb{R}^p) \to L^1(\rho)$. Finally, $M := \sup_{w \in \mathbb{S}_\pm} \sup_{x \in \mathcal{X}}\{|\phi(w, x)|, \|\nabla\phi(w, x)\|\}$ is finite (here $\sup_{x \in \mathcal{X}}$ is an essential supremum w.r.t. ρ), so it follows that $J' : \mathcal{P}_2(\mathbb{R}^p) \to \mathcal{C}^1(\mathbb{S}_\pm)$ is Lipschitz continuous because $\|J'_\mu - J'_{\tilde{\mu}}\|_{\mathcal{C}^1} \leq M\|(d\lambda_\mu/d\rho) - (d\lambda_{\tilde{\mu}}/d\rho)\|_{L^1(\rho)}$. ∎

Existence of a Wasserstein gradient flow. After these preliminaries, we are in position to show the existence of Wasserstein gradient flows for certain initializations. In this section, we will not attempt to prove uniqueness of the dynamics.

Lemma 26 *Under Assumption* (A4)*, let $\mu_0 \in \mathcal{P}_2(\mathbb{R}^{d+2})$ satisfying $\text{spt}(\mu_0) \subset D$. Then there exists a well-defined Wasserstein gradient flow of F starting from μ_0. It satisfies $\text{spt}(\mu_t) \subset D$ for $t \geq 0$.*

Proof We will essentially follow the proof of (Chizat and Bach, 2018, Thm. 2.6), by highlighting the points that need to be adapted. Without loss of generality (the trivial case $\mu_0 = \delta_0$ put aside), we assume that $\text{spt}(\mu_0)$ is included in a compact set $K \subset D$ that does not contain 0, since the general case can be treated as in Theorem 2 by exploiting the 2-homogeneity of F'_μ for all $\mu \in \mathcal{P}_2(\mathbb{R}^p)$.

Let $\mu_{0,m}$ be a sequence of empirical measures concentrated in K that converges weakly to μ_0. For all m, the Wasserstein gradient flow equation Eq. (8) boils down to an ordinary differential equation which existence is guaranteed by Peano existence theorem and Lemma 25, at least on some maximal interval $[0, t_0[$. If $t_0 < \infty$, the growth of $\nabla F'_{\mu_{m,t}}$ for $t \in [0, t_0[$ is such that Grönwall type arguments guarantee that the atoms of $\mu_{m,t}$ are uniformly bounded (also bounded away from 0) and converge as $t \to t_0$, thus defining a measure μ_{m,t_0}. We denote $X(t, w) = (A_t(w), B_t(w)) \in \mathbb{R}^{p-1} \times \mathbb{R}$ where X is the flow from Eq. (8). It holds, denoting $\tilde{x} = (x, 1)$ and expanding Eq. (8):

$$\frac{d}{dt}\frac{1}{2}\|A_t(w)\|^2 = A_t(w)^\top \int B_t(w)(A_t(w)^\top\tilde{x})^0_+\tilde{x} \, d[\nabla S(\hat{h}(\mu_{t,m}))](x),$$

$$\frac{d}{dt}\frac{1}{2}|B_t(w)|^2 = B_t(w) \int (A_t(w)^\top\tilde{x})_+ \, d[\nabla S(\hat{h}(\mu_{t,m}))](x).$$

These quantities being equal, this shows that $\forall w \in \mathbb{R}^p$, $\|A_t(w)\|_2^2 - |b_t(w)|^2$ is constant on $[0, t_0[$, as long as $A_t(w) \neq 0$ on this interval. It follows that whenever $w \in D$, we have $\|A_t(w)\| = |B_t(w)|$ for all $t \in [0, t_0[$. Since $\mu_{t,m} = (X_t)_\# \mu_{0,m}$, we have that if $\mathrm{spt}(\mu_{m,0}) \subset D$ then $\mathrm{spt}(\mu_{m,t}) \subset D$ for all $t \in [0, t_0[$ and since D is closed, we have $\mathrm{spt}(\mu_{t_0,m}) \subset D$ which contradicts the maximality of t_0 (the trajectory could be extended by Peano existence theorem). Thus $t_0 = \infty$. The principle behind the initialization on D is illustrated on Figure 4 and is related to well-known properties of homogeneous models (Du et al., 2018).

Figure 4: 2-d slice of the vector field $\nabla F'_{\mu_t}$ in the ReLU case (illustration). Initializing on the invariant set D allows to avoid $a = 0$.

So far, we have proved that for all m, there exists a Wasserstein gradient flow $\mu_{t,m}$ starting from $\mu_{0,m}$, defined on $[0, \infty[$ and such that $\mathrm{spt}(\mu_{t,m}) \subset D$ for $t \geq 0$. The rest of the proof follows that in (Chizat and Bach, 2018, Thm. 2.6) where we extract a (weak) limit curve μ_t by compactness and show that it satisfies the Wasserstein gradient flow equation Eq. (8). The only technical point is Step.(iii) in that proof, which is taken care of by the last claim of Lemma 25. Note that here we do not prove uniqueness of the Wasserstein gradient flow, only its existence. ∎
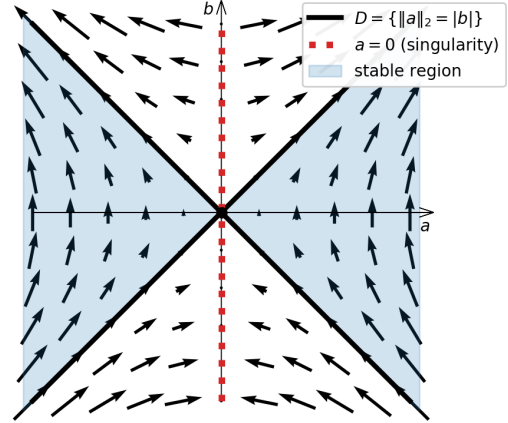
Implicit bias for ReLU networks. Let us restate Theorem 5 in a slightly more general form and making Assumption (*) explicit. We recall that $\nu_t := \Pi_2(\mu_t)$ and $\bar{\nu}_t := \nu_t/(\nu_t(\mathbb{S}^{p-1}))$.

Theorem 27 *Let* $\mu_0 \in \mathcal{P}_2(\mathbb{R}^p)$ *be such that* $\mathrm{spt}(\Pi_2(\mu_0)) = \mathbb{S}_\pm$ *and assume* (A4). *Then there exists* $(\mu_t)_{t \geq 0}$ *a Wasserstein gradient flow of the objective Eq. (9) with initialization* μ_0. *Moreover,*

- *if* $\bar{\nu}_t = \Pi_2(\mu_t)/([\Pi_2(\mu_t)](\mathbb{S}^{p-1}))$ *converges weakly to some* $\bar{\nu}_\infty \in \mathcal{P}(\mathbb{S}^{p-1})$,

- *if* $\nabla S[\hat{h}(\mu_t)]$ *converges weakly to some* $\lambda_\infty \in \mathcal{P}(\mathcal{X})$, *and*

- *(*) if* J'_{μ_t} *converges in* $\mathcal{C}^1(\mathbb{S}_\pm)$ *to some* J'_∞ *that satisfies the Morse-Sard property: the set of critical values of* J'_∞ *(i.e.,* $v \in \mathbb{R}$ *such that there exists* $\theta \in \mathbb{S}_\pm$ *such that* $J'_\infty(\theta) = v$ *and* $\nabla J'_\infty(\theta) = 0$*) has Lebesgue measure zero,*

then $h(\bar{\nu}_\infty, \cdot)$ *is a maximizer for the* $\mathcal{F}_1$*-max-margin problem* $\max_{\|f\|_{\mathcal{F}_1} \leq 1} \min_{x \in \mathrm{spt}(\rho)} y(x)f(x)$.

Before proving this theorem, let us discuss its assumptions. First, the assumption on the initialization given here is satisfied by μ_0 given in Theorem 5 (which is an example given for the sake of concreteness). The conditions in the two first bullets are similar to those of Theorem 3 and just require the uniqueness of limits of some sequences that live in compact spaces. In Assumption (*), the fact that the Morse-Sard property holds for J'_∞ was already required in Chizat and Bach (2018) and it is an open question to guarantee that this property holds in this context (where J'_∞ is potentially an infinite sum of subanalytic functions, instead of a finite sum as in Theorem 5). Finally, the most undesirable assumption is perhaps the convergence of J'_{μ_t} to J'_∞ in $\mathcal{C}^1(\mathbb{S}_\pm)$. The fact that it

converges in $\mathcal{C}(\mathbb{S}_{\pm})$ can be shown a priori, so the assumption is really on the uniform convergence of the gradient. In particular, it requires J'_∞ to be continuously differentiable, which is for instance not true if λ_∞ is a discrete measure.

Proof Since $\mathrm{spt}(\mu_0) \subset D$, the existence of a Wasserstein gradient flow is proved in Lemma 26. Let $\lambda_t = \nabla S[h(\mu_t)] \in \mathcal{P}(\mathcal{X})$ which, as assumed, converges weakly to λ_∞. Since $y(x)\lambda_\infty(\,\mathrm{d}x)$, seen as a signed measure on $\mathcal{X}$ is non-zero, it follows by the contraposition of Lemma 28 that J'_∞ is not identically 0 on $\mathbb{S}_+$. Moreover, since J'_∞ takes opposite values on $\mathbb{S}_+$ and $\mathbb{S}_-$, we have that $M := \max_{w\in\mathbb{S}_\pm} J'_\infty(w) > 0$. For the rest of the proof, we just comment on how to adapt the arguments of Theorem 3 to this case.

Step 1. The argument of Theorem 3 goes through if we replace $\mathbb{S}^{p-1}$ by the set $\mathbb{S}_\pm$, in particular thanks to Assumption (*). Lemma 25 and the positive 2-homogeneity guarantee that the restriction of the flow X to D is a diffeomorphism.

Step 2. Since we focus on the exponential loss, we can directly apply Lemma 24, which gives $\mathrm{spt}(\lambda_\infty) \subset \arg\min_{x\in\mathrm{spt}(\rho)} y(x)h(\bar{\nu}_\infty, x)$.

Steps 3. The argument of Theorem 3 goes through, again replacing $\mathbb{S}^{p-1}$ by the set $\mathbb{S}_\pm$.

Steps 4. We conclude with the optimality conditions in Proposition 29 and the structure of the minimizers given in Proposition 9. ∎

Lemma 28 *Assume that $\mathcal{X} \subset \mathbb{R}^d$ is compact. If $\lambda \in \mathcal{M}(\mathcal{X})$ satisfies $\int (a^\top(x,1))_+\,\mathrm{d}\lambda(x) = 0$ for all $a \in \mathbb{S}^d$, then $\lambda = 0$.*

Proof Consider the measure $\tilde{\lambda} \in \mathcal{M}(\mathbb{S}^d)$ that is such that $\int \varphi((x,1))\,\mathrm{d}\lambda(x) = \int \varphi(z)\,\mathrm{d}\tilde{\lambda}(z)$ for all positively homogeneous functions $\varphi \in \mathcal{C}(\mathbb{R}^{d+1})$. By construction, $\tilde{\lambda}$ is concentrated on the set S_α of points in $\mathbb{S}^d$ for which the last coordinate is larger than α for some $\alpha > 0$. Let us show that $\tilde{\lambda} = 0$, which implies that $\lambda = 0$ and thus the claim.

Let $\mathrm{d}\tau \in \mathcal{P}(\mathbb{S}^d)$ be the uniform distribution on the sphere. It is known (Bach, 2017a) that the set of functions $\{f_g : a \mapsto \int (a^\top z)_+ g(z)\,\mathrm{d}\tau(z); \; g \in L^2(\tau)\}$ is dense in $\mathcal{C}(S_\alpha)$. Since for $g \in L^2(\tau)$ it holds, by Fubini,

$$\int f_g\,\mathrm{d}\tilde{\lambda} = \int \mathrm{d}\tilde{\lambda}(a) \int \mathrm{d}\tau(z)(a^\top z)_+ g(z) = \int \mathrm{d}\tau(z) \int \mathrm{d}\tilde{\lambda}(a)(a^\top z)_+ g(z) = 0,$$

it follows that $\tilde{\lambda} = 0$ and thus $\lambda = 0$. ∎

Finally, let us state the optimality conditions of the optimization problem mentionned in Theorem 5 (it is an application of minimax duality (Sion, 1958), just like Proposition 29). For $\mathcal{X} \subset \mathbb{R}^d$ compact, let

$$\gamma_1 := \max_{\|f\|_{\mathcal{F}_1}\le 1}\min_{x\in\mathcal{X}} y(x)f(x) = \max_{\nu\in\mathcal{P}(\mathbb{S}^{p-1})}\min_{\lambda\in\mathcal{P}(\mathcal{X})}\int_\mathcal{X}\int_{\mathbb{S}^{p-1}} y(x)\phi(\theta,x)\,\mathrm{d}\nu(\theta)\,\mathrm{d}\lambda(x). \quad (17)$$

Proposition 29 (Optimality conditions - ReLu) *The maximization problem (17) admits global maximizers $\nu^\star \in \mathcal{P}(\mathbb{S}^{p-1})$. Moreover, a measure $\nu^\star \in \mathcal{P}(\mathbb{S}^{p-1})$ is a global maximizer of (4) if and only if $\nu^\star \in \mathcal{P}(\mathbb{S}^{p-1})$ and there exists $\lambda^\star \in \mathcal{P}(\mathcal{X})$ such that*

$$\mathrm{spt}\,\nu^\star \subset \arg\max_{\theta\in\mathbb{S}^{p-1}}\int_\mathcal{X} y(x)\phi(\theta,x)\,\mathrm{d}\lambda^*(x) \quad and \quad \mathrm{spt}\,\lambda^\star \subset \arg\min_{x\in\mathcal{X}}\int_{\mathbb{S}^{p-1}} y(x)\phi(\theta,x)\,\mathrm{d}\nu^\star(\theta).$$