1. INTRODUCTION

In this paper, we are interested in understanding certain models of deep network architectures. We consider a regression task where $(X,Y) \in \mathbb{R}^{d_X} \times \mathbb{R}^{d_Y}$ is a random pair. As is standard, we consider a network with L hidden layers. The parameters of the network are defined via a choice of dimensions

$$d_0 := d_X, D_0, d_1, D_1, d_2, D_2, \ldots, d_L, D_L, d_{L+1} = d_Y,$$

of numbers of units for layers $1, \ldots, L$

$$\vec{N} := (N_1, \ldots, N_L) \in \mathbb{N}^L \backslash \{0\}.$$

We set $N_0 = N_{L+1} = 1$ as the number of units in the input and output. We also select functions:

$$\eta^{(\ell)} : \mathbb{R}^{d_\ell} \times \mathbb{R}^{D_\ell} \to \mathbb{R}^{d_{\ell+1}}, \sigma^{(\ell+1)} : \mathbb{R}^{d_{\ell+1}} \to \mathbb{R}^{d_{\ell+1}} \ (\ell = 0, \ldots, L).$$

We consider networks with L internal layers and full connections between layers.

This gives us a model with parameters:

$$\theta^{(\ell)}_{i_\ell, i_{\ell+1}} \, : \, \ell = 0, \ldots, L, \, (i_\ell, i_{\ell+1}) \in [N_\ell] \times [N_{\ell+1}]$$

each representing a connection between unit i_ℓ in layer ℓ and unit $i_{\ell+1}$ in layer $\ell+1$. Our total vector of parameters $\vec{\theta}_{\vec{N}}$ lives in dimension

$$p_{\vec{N}} := \sum_{\ell=0}^{L} N_\ell \, N_{\ell+1} D_\ell.$$

The function computed by our network is given by:

$$\widehat{y}_N = \widehat{y}_N : \mathbb{R}^{d_X} \times \mathbb{R}^{p_{\vec{N}}} \to \mathbb{R}^{d_Y}$$

that takes as input an element $x \in \mathbb{R}^{d_X}$ and a setting or parameters $\vec{\theta}_{\vec{N}}$ and produces a sequence of values as follows. For $1 \leq i_1 \leq N$,

$$a^{(1)}_{i_1}(x, \vec{\theta}_N) := \eta^{(0)}(x, \theta^{(0)}_{1,i_1}). \tag{1}$$

For $\ell = 1, \ldots, L$, $1 \leq i_{\ell+1} \leq N_{\ell+1}$:

$$z^{(\ell+1)}_{i_{\ell+1}}(x, \vec{\theta}_{\vec{N}}) := \frac{1}{N_\ell} \sum_{i_\ell=1}^{N_\ell} \eta^{(\ell)}(a^{\ell}_{i_\ell}(x, \vec{\theta}_{\vec{N}}), \theta^{(\ell)}_{i_\ell, i_{\ell+1}}); \tag{2}$$

$$a^{(\ell+1)}_{i_{\ell+1}}(x, \vec{\theta}_{\vec{N}}) := \sigma^{(\ell+1)}(z^{(\ell+1)}_{i_{\ell+1}}(x, \vec{\theta}_{\vec{N}})). \tag{3}$$

The output is $\widehat{y}(x, \vec{\theta}_{\vec{N}}) = a^{(L+1)}_1$.

Our model defines a very general version of a neural network with fully connected layers, where the internal units (with activations $a^{(\ell)}_{i_\ell}$) may have dimension greater than 1. This is convenient because it allows us to carry bias terms and consider more general versions of these units.

The derivatives of $\widehat{y}$ with respect to the $\theta^{(\ell)}_{i_\ell,i_{\ell+1}}$ can be computed via standard backpropagation. In what follows, we assume that all functions σ and eta are C^1-Fréchet differentiable. We let $D\sigma^{(\ell)}(z)$ denote the derivative of $\sigma^{(\ell)}$ and $D_a\eta^{(\ell)}(a,\theta)$, $D_\theta\eta^{(\ell)}(a,\theta)$ denote the partial derivatives of η with respect to the variables a and θ (respectively).

The derivative with respect to the weights $\theta^{(L)}_{i_L,1}$ is most easily computed. Omitting the $(x,\vec{\theta}_{\vec{N}})$ for simplicity,

$$\partial_{\theta^{(L)}_{i_L,1}}\widehat{y} = \frac{1}{N_\ell} D\sigma^{(L+1)}(z_1^{(L+1)}). \tag{4}$$

For other weights, we define:

$$M_{i_L}(x,\vec{\theta}_N) := D\sigma^{(L+1)}(z_1^{(L+1)})\, D_a\eta^{(L)}(a^{(L)}_{i_L},\theta^{(L)}_{i_L,1})\, D\sigma^{(L)}(z^{(L)}_{i_L})$$

and for $1 \le \ell \le L-1$, $1 \le i_\ell \le N$:

$$M_{i_\ell,i_{\ell+1},\dots,i_L}(x,\vec{\theta}_N) := M_{i_{\ell+1},\dots,i_L}(x,\vec{\theta}_N)\, D_a\eta^{(\ell)}(a^{(\ell)}_{i_\ell},\theta^{(\ell)}_{i_\ell,i_{\ell+1}})\, D\sigma^{(\ell)}(z^{(\ell)}_{i_\ell}).$$

Then the Fréchet derivatives of $\widehat{y}$ with respect to $\theta^{(\ell)}_{i_\ell,i_{\ell+1}}$ for $0 \le \ell \le L-1$ is

$$\partial_{\theta^{(L-1)}_{i_L,i_{L+1}}}\widehat{y}_N(x,\vec{\theta}_N) = \frac{1}{N_L N_{L-1}} M_{i_L}(x,\vec{\theta}_N)\, D_\theta\eta^{(L-1)}(a^{(L-1)}_{i_{L-1}},\theta^{(L-1)}_{i_{L-1},i_L})$$

and

$$\partial_{\theta^{(\ell)}_{i_\ell,i_{\ell+1}}}\widehat{y}_N(x,\vec{\theta}_N) = \frac{1}{\prod_{j=\ell}^{L} N_j}\sum_{i_{\ell+2},\dots,i_L=1}^{N} M_{i_{\ell+2},\dots,i_L}(x,\vec{\theta}_N)\, D_\theta\eta^{(\ell)}(a^{(\ell)}_{i_\ell}(x,\vec{\theta}_N),\theta^{(\ell)}_{i_\ell,i_{\ell+1}}).$$

If we define our population loss function $L_{\vec{N}} : \mathbb{R}^{p_{\vec{N}}} \to \mathbb{R}$ as:

$$L_{\vec{N}}(\vec{\theta}_{\vec{N}}) := \frac{1}{2}\mathbb{E}_{(X,Y)\sim P}\left[\|Y - \widehat{y}_{\vec{N}}(X,\vec{\theta}_{\vec{N}})\|^2\right],$$

then one can study an evolution

$$\frac{d\vec{\theta}_{\vec{N}}}{dt}(t) = -\alpha_{\vec{N}}(t)\nabla L_{\vec{N}}(\vec{\theta}_{\vec{N}}(t))$$

via the partial derivatives of $\widehat{y}$.

A few preliminary comments on these derivatives are in order. One of them is on *differing time scales across layers* at least when $N_1 = \cdots = N_L = N$. In this case formula (4) suggests the weights between layers L and $L-1$ and the weights between layers 0 and 1 move at rate N^{-1}, whereas other weights move at speed N^{-2}.

2. GOAL AND ASSUMPTIONS

Our goal is to analyse the evolution of the weight vector $\vec{\theta}_N(t)$ for $t \geq 0$ under a special setting of parameters $\vec{N} = (N, \ldots, N)$ with

$$\frac{d\vec{\theta}_N(t)}{dt} = -N^2 \alpha(t)\, \nabla L_N(\vec{\theta}_N(t)).$$

This implies in particular that

We will study this evolution under the following assumptions.

Assumption 1. *At time* 0*, the weights* $\theta^{(\ell)}_{i_\ell, i_{\ell+1}}(0) \in \mathbb{R}^{D_\ell}$ *(*$10 \leq \ell \leq L$*,* $(i_\ell, i_{\ell+1}) \in [N_\ell] \times [N_{\ell+1}]$*) are all independent. Moreover, there are probability laws* $\mu_0^{(\ell)}$ *over* $\mathbb{R}^{D_\ell}$ *(for* $0 \leq \ell \leq L$*) such that* $\theta^{(\ell)}_{i_\ell, i_{\ell+1}}(0) \sim \mu_0^{(\ell)}$ *for each* $0 \leq \ell \leq L$*,* $(i_\ell, i_{\ell+1}) \in [N_\ell] \times [N_{\ell+1}]$*. The distributions* $\mu_0^{(\ell)}$ *defined above all have bounded support contained in balls of radius* R *around the origins of their respective domains.*

Assumption 2. *There exists a* $C > 0$ *such that the Fréchet derivatives of the* $\sigma^{(\ell+1)}$ *and* $\eta^{(\ell)}$ *satisfy:*

$$\forall z \in \mathbb{R}^{d_{\ell+1}} \|D\sigma^{(\ell)}(z)\| \leq C$$

$$\forall (a, \theta) \in \mathbb{R}^{d_\ell} \times \mathbb{R}^{D_\ell} \,:\, \|D_a \eta^{(\ell)}(a, \theta)\| \leq C(1 + \|\theta\|) \text{ and } \|D_\theta \eta^{(\ell)}(a, \theta)\| \leq C(1 + \|a\|).$$

Moreover, $D\sigma^{(\ell)}(z)$

3. THE MEAN FIELD APPROXIMATION DISTRIBUTION

We now construct a mean-field approximating distribution for our problem. By this we mean that we will construct random continuous functions

$$\overline{a}$$