Exercices. Ecole Polytechnique. MAP670R-2022 Advanced Topics in Deep Learning.

Exercise 1 (Back-propagation) Assume that the layers of a MLP write for $0 \leq j < J$:

$$x_{j+1} = W_j \rho x_j = f_j(x_j, W_j)$$

so that $x_j \in \mathbb{R}^{n_j}$ and that $x_J \in \mathbb{R}^{n_J}$ is fed to $\ell : \mathbb{R}^{n_J} \to \mathbb{R}$. We write also $\Phi(x) = x_J$ the output of the MLP. Note this implies that $W_j \in \mathbb{R}^{n_j \times n_{j+1}}$. We write $\ell_j(x_j) = \ell(W_J\rho ... \rho W_j \rho x_j)$ and $\phi_j(x) = x_j$. We will write $(f_{j+1} \circ f_j)(x_j, W_j, W_{j+1}) \triangleq f_{j+1}(f_j(x_j, W_j), W_{j+1})$'.

1. Compute $\partial_{x_j} f_j(x_j, W_j)$ and $\partial_{W_j} f_j(x_j, W_j)$.
2. Verify that $\ell(\Phi(x)) = \ell_j \circ \phi_j(x)$
3. Compute $\nabla_{x_j} \ell_j$.
4. Deduce $\nabla_{W_j} \ell$.

We remind that $\{f(x), x \in \mathcal{X}\}$ is a centered Gaussian process if for any $x_1, ..., x_k \in \mathcal{X}$, $(f(x_1), ..., f(x_k))$ is a Gaussian variable with law $\mathcal{N}(0, \Sigma)$. In this case,

$$\Sigma_{ij} = K(x_i, x_j)\,,$$

where K is the covariance function of f. In the following, the Gaussian processes will be centered.

We will also write:

$$\Sigma_1(x, x') = \frac{1}{w_1} x^T x', \tag{1}$$

and also:

$$\Sigma_{j+1}(x, x') = \mathbb{E}_{(u,v)\sim\mathcal{N}(0, \begin{bmatrix} \Sigma_j(x,x) & \Sigma_j(x',x) \\ \Sigma_j(x,x') & \Sigma_j(x',x') \end{bmatrix})}[\rho(u)\rho(v)], \tag{2}$$

and:

$$\dot{\Sigma}_{j+1}(x, x') = \mathbb{E}_{(u,v)\sim\mathcal{N}(0, \begin{bmatrix} \Sigma_j(x,x) & \Sigma_j(x',x) \\ \Sigma_j(x,x') & \Sigma_j(x',x') \end{bmatrix})}[\dot{\rho}(u)\dot{\rho}(v)]]\,. \tag{3}$$

We will consider the following model:

$$\Phi(W; x) = W_J \frac{1}{\sqrt{w_J}} \rho W_{J-1} \frac{1}{\sqrt{w_{J-1}}} \rho ... \rho W_1 \frac{1}{\sqrt{w_1}} x\,, \tag{4}$$

and $W(0)$ is a Gaussian random initialization such that $\mathbb{E}[W(0)W(0)^T] = \mathbf{I}$. We assume that x, x' have non-negative entries.

Exercise 2 (The Neural Tangent Kernel) We introduce here:

$$K_{W(0)}(x, x') = \sum_{j=1}^{J} \partial_{W_j} \Phi(x; W(0)) \partial_{W_j} \Phi(x'; W(0))^T\,. \tag{5}$$

1. What is the shape of $K_{W(0)}$? and $\Sigma_j(x, x')$?
2. Show that:

$$K_{W(0)}(x, x') = \frac{1}{w_J} \rho \Phi_{J-1}(x) \rho \Phi_{J-1}(x')^T \tag{6}$$

$$+ \frac{1}{w_J} W_J [\partial\rho]_{\Phi_{J-1}} \partial_W \Phi_{J-1}(x) \partial_W \Phi_{J-1}(x')^T [\partial\rho]^T_{\Phi_{J-1}} W_J^T \tag{7}$$

3. Prove by induction that $F_{j,k}(x) \triangleq \lim_{w_j \to \infty} ... \lim_{w_2 \to \infty} \Phi_{j,k}(x)$ is a Gaussian process with kernel Σ_j for $j \leq J$ and $k \leq w_{j+1}$, and that the family $\{F_{j,k}\}_k$ is a family of independent Gaussian processes.

4. Show that we have the following limit:

$$\lim_{w_J \to \infty} ... \lim_{w_2 \to \infty} K_{W(0)}(x, x') = \sum_{j=1}^{J} \Sigma_j \dot{\Sigma}_{j+1} ... \dot{\Sigma}_J \mathbf{I} \,. \tag{8}$$

Observe that w_1 and w_{J+1} are constant.

Exercise 3 (The Neural Tangent Kernel dynamics) Now, we assume that the training dynamic is given for $t \in [0, T]$ by:

$$\frac{d}{dt} W(t) = -\lambda \partial_W \Phi(W(t))^T \nabla \mathcal{R}(\Phi(W(t))) \,, \tag{9}$$

for some step size $\lambda > 0$ given a posterio. We also assume that, as the layer grows: $\int_0^T \|\nabla \mathcal{R}(\Phi(W(t)))\| \, dt \leq C$ for some universal constant. For the sake of simplicity, we will also assume that $w_1 = w_{J+1} = 1$ and $w_j = w$ for $1 < j < J+1$ and that $|\rho'| \leq 1$.

1. Compute $\frac{d}{dt} W_j(t)$.
2. Let:

   $$u(t) = (\|W_1(t) - W_1(0)\| + \|W_1(0)\|, ..., \|W_J(t) - W_J(0)\| + \|W_J(0)\|).$$

   Show that for some $C' > 0$:

   $$\|\frac{d}{dt} W_j(t)\| \leq \frac{C'}{w^{(J-1)/2}} \|u(t)\|^{J-1} \|\nabla \mathcal{R}(\Phi(W(t)))\|$$

3. Show that for some $C'' > 0$:

   $$|\frac{d}{dt} \|u(t)\|^{2-J}| \leq \frac{C''}{w^{(J-1)/2}} \|\nabla \mathcal{R}(\Phi(W(t)))\|$$

4. Deduce that $\|W_j(t) - W_j(0)\| \to 0$.
5. Using a reasoning similar to Exercise 2, question 2, show that there is $C_J > 0$ such that:

   $$\|\partial_W \Phi(W(t); x) - \partial_W \Phi(W(0); x)\| \leq J(\sup_j \|W_j(t)\| + \|W_j(0)\|)^{J-1} \sup_j \|W_j(t) - W_j(0)\|$$

6. Show that $\|uu^T - vv^T\| \leq (\|u\| + \|v\|)\|u - v\|$. Prove that:

   $$\lim_{w \to \infty} \|K_{W(t)} - K_{W(0)}\| = 0 \,.$$

   Comment the result. (We will admit that $\int_0^T \|\nabla \mathcal{R}(\Phi(W(t)))\|$ bounded is satisfied for Gaussian entries and the MSE loss)