MSRI Summer School	Precept 3: Lazy training in deep learning
Instructor: J. Bruna	
TA: S. Jelassi	

In non-convex optimization, changes in the parametrization of a model, optimization algorithm or initialization can lead to a variety of models with very different properties. We have studied one implicit bias phenomenon during the last problem session in the case of gradient descent applied to underdetermined logistic regression. Today, we focus on a more general implicit bias that is called lazy training (Chizat and Bach, 2018). It corresponds to the case where the model behaves likes its linearization around the initialization.

Setting.

We want to minimize with gradient-based methods, the objective function $F\colon \mathbb{R}^p \to \mathbb{R}_+$ defined as

$$F(w) := R(h(w)),$$

where $\mathbb{R}^p$ is the parameter space, $\mathcal{F}$ is a Hilbert space, $h\colon \mathbb{R}^p \to \mathcal{F}$ is a smooth model and $R\colon \mathcal{F} \to \mathbb{R}_+$ is a smooth loss. With an initialization $w_0 \in \mathbb{R}^p$, the linearized model around w_0, $\bar{h}$ is defined as $\bar{h}(w_0) = h(w_0) + Dh(w_0)(w - w_0)$ and the corresponding objective $\bar{F}\colon \mathbb{R}^p \to \mathbb{R}_+$ as

$$\bar{F}(w) := R(\bar{h}(w)).$$

Assumption. h is differentiable with a locally Lipschitz differential Dh i.e. $\|Dh(w) - Dh(w')\| \leq L_{Dh}\|w - w'\|$ and R is differentiable with a Lipschitz gradient i.e. $\|\nabla R(z) - \nabla R(z')\| \leq L_R\|z - z'\|$.

Problem 1: When does lazy training occur?

Let's assume that we initialize the gradient-based method in a point $w_0 \in \mathbb{R}^p$ that is not a minimizer i.e. $F(w_0) > 0$ and not a critical point i.e. $\nabla F(w_0) \neq 0$. We consider a gradient descent step $w_1 := w_0 - \eta \nabla F(w_0)$ with a small stepsize $\eta > 0$.

1) Give an approximation to the relative change of the objective $\Delta(F) := \frac{|F(w_1) - F(w_0)|}{F(w_0)}$ in terms of $\nabla F(w_0)$, $F(w_0)$ and η.

2) Give an approximation to the relative change of the differential of h, $\Delta(Dh) := \frac{\|Dh(w_1) - Dh(w_0)\|}{\|Dh(w_0)\|}$ in terms of $\nabla F(w_0)$, $D^2h(w_0)$, $Dh(w_0)$ and η.

3) Lazy training refers to the case where the differential of h does not sensibly change while the loss drastically decreases. By using the previous questions, show that this corresponds to

$$\frac{\|\nabla F(w_0)\|}{F(w_0)} \gg \frac{\|D^2 h(w_0)\|}{\|Dh(w_0)\|}. \tag{1}$$

4) For the square loss $R(y) := \frac{1}{2}\|y - y^*\|^2$ for some $y^* \in \mathcal{F}$ this leads to the simpler criterion

$$\kappa_h(w_0) := \|h(w_0) - y^*\| \cdot \frac{\|D^2 h(w_0)\|}{\|Dh(w_0)\|^2} \ll 1. \tag{2}$$

By considering a scaling factor $\alpha > 0$, derive an expression for $\kappa_{\alpha h}(w_0)$. How should we choose α to favor lazy training?

Application to q-layers neural networks with homogeneous activations. Assume that $h(W_1, \dots, W_q) = W_q \sigma(W_{q-1}\sigma(W_{q-2} \dots \sigma(W_1 z)))$, where $W_1, \dots, W_q$ are the weight matrices and σ is a homogeneous activation function i.e. $\sigma(\lambda z) = \lambda\sigma(z)$, for $\lambda > 0$.

5) For $\lambda > 0$ and $W_0^{(1)}, \dots, W_0^{(q)} \in \mathbb{R}^p$, express $h(\lambda W_0^{(1)}, \dots, \lambda W_0^{(q)})$ as in terms of $h(W_0^{(1)}, \dots, W_0^{(q)})$.

6) Deduce an expression for $\kappa_h(\lambda W_0^{(1)}, \dots, \lambda W_0^{(q)})$. When does the lazy regime appears in these networks?

Problem 2: Analysis of lazy training dynamics

In this problem, we want to show that lazy training dynamics for the scaled objective $F_\alpha(w) := 1/\alpha^2 \cdot R(\alpha h(w))$ are close, when α is large, to those of the scaled objective for the linearized model $\bar{F}_\alpha(w) := 1/\alpha^2 \cdot R(\alpha \bar{h}(w))$, where $\bar{h}(w) := h(w_0) + Dh(w_0)(w - w_0)$.

With an initialization $w_0 \in \mathbb{R}^p$, the gradient flow of F_α is the path $(w_\alpha(t))_{t \geq 0}$ in the space of parameters $\mathbb{R}^p$ satisfying $w_\alpha(0) = 0$ and solves the ordinary differential equation (ODE)

$$w'_\alpha(t) = -\nabla F_\alpha(w_\alpha(t)) = -\frac{1}{\alpha} Dh(w_\alpha(t))^\top \nabla R(\alpha h(w_\alpha(t))). \tag{3}$$

Similarly, the gradient flow of $\bar{F}_\alpha$ is the path $(\bar{w}_\alpha(t))_{t \geq 0}$ satisfying $\bar{w}_\alpha(0) = 0$ and solving the ODE:

$$\bar{w}'_\alpha(t) = -\nabla \bar{F}_\alpha(\bar{w}_\alpha(t)) = -\frac{1}{\alpha} Dh(w_0)^\top \nabla R(\alpha \bar{h}(\bar{w}_\alpha(t))). \tag{4}$$

We assume that $h(w_0) = 0$. We would like to show that given a fixed time horizon $T > 0$, it holds that

$$\sup_{t \in [0,T]} \|\alpha h(w_\alpha(t)) - \alpha \bar{h}(\bar{w}_\alpha(t))\| = O(1/\alpha). \tag{5}$$

(5) leads to $\sup_{t\in[0,T]} \|w_\alpha - \bar{w}_\alpha(t)\| = O(1/\alpha^2)$ which shows that for a large α, the trajectory of the original model and the linearized model are similar.

1) For $T > 0$, show that $\int_0^T \|w'_\alpha(t)\|dt \leq \sqrt{T}\left(\int_0^T \|\nabla F_\alpha(w_\alpha(t))\|^2 dt\right)^{1/2}$.
2) By using the fact that $\frac{d}{dt}F_\alpha(w_\alpha(t)) = -\|\nabla F_\alpha(w_\alpha(t)\|^2$, show that $\sup_{t\in[0,T]} \|w_\alpha(t) - w(0)\| \leq (T \cdot (F_\alpha(w_\alpha(0)) - F_\alpha(w_\alpha(T))))^{1/2} \lesssim 1/\alpha$.
This implies that there exists $C > 0$ such that $\sup_{t\in[0,T]} \|\alpha h(w_\alpha(t)) - \alpha h(w_\alpha(0))\| \leq C$ and $\sup_{t\in[0,T]} \|\nabla R(\alpha h(w_\alpha(t)))\| \leq C$.
3) Let's define $\Delta(t) := \|\alpha h(w_\alpha(t)) - \alpha\bar{h}(w_\alpha(t))\|$. Remark that $\Delta(0) = 0$. Show that there exist two constants $C_1, C_2 > 0$ such that

$$\Delta'(t) \leq C_1/\alpha + C_2\Delta(t). \tag{6}$$

Solve the ordinary differential inequality (6) and deduce that $\sup_{t\in[0,T]} \Delta(t) \leq C_3/\alpha$, for some constant $C_3 > 0$. Conclude that (5) holds.
Hint: Use the ODEs corresponding to the dynamics of $\alpha h(w_\alpha(t))$ and $\alpha\bar{h}(\bar{w}_\alpha(t))$ given by

$$\frac{d}{dt}\alpha h(w_\alpha(t)) = -Dh(w_\alpha(t))Dh(w_\alpha(t))^\top \nabla R(\alpha h(w_\alpha(t)))$$
$$\frac{d}{dt}\alpha\bar{h}(\bar{w}_\alpha(t)) = -Dh(w(0))Dh(w(0))^\top \nabla R(\alpha\bar{h}(\bar{w}_\alpha(t))),$$

with initial conditions $\alpha h(w_\alpha(0)) = \alpha\bar{h}(\bar{w}_\alpha(0)) = \alpha h(w_0)$ and the fact that $\frac{d}{dt}\|g(t)\| \leq \|g'(t)\|$.

N.B: This exercise shows that during the optimization process, if we set α large enough, the iterates of the model follow those of the linearized model. Chizat and Bach, 2018 prove stronger results by quantifying the constants we introduced in this exercise and by giving bounds that are uniform in time. An interesting implication of (**??**) is that $\alpha h(w_\alpha(T))$ generalizes like $\alpha\bar{h}(w_\alpha(T))$ outside the training set for large α. In the context of neural networks, this establishes the connexion between lazy training and the Neural Tangent Kernel (NTK) (Jacot et al., 2018).

Problem 3: Interpolating between lazy and active training

In the previous exercises, we have established that choosing a large α when training a model leads to lazy training. Woodworth et al., 2019 show how the scale of initialization controls the transition between lazy and active training and affects generalization properties in multilayer homogeneous models. We study this phenomenon in this exercise.

Consider a training set $\{x^{(n)}, y^{(n)}\}_{n=1}^N$ and $x_n \in \mathbb{R}^d$. We want to fit a model that belongs to the class of linear (linear in x and not in w) functions over $\mathbb{R}^d$ with squared parametrization i.e.

$$f(w,x) = \sum_{i=1}^{d} (w_{+,i}^2 - w_{-,i}^2) x_i = \langle \beta_w, x \rangle \text{ where } w = \begin{bmatrix} w_+ \\ w_- \end{bmatrix} \in \mathbb{R}^{2d} \text{ and } \beta_w = w_+^2 - w_-^2, \quad (7)$$

where we use the notation z^2 for $z \in \mathbb{R}^d$ to denote the element-wise squaring. We are in the regression setting where we minimize the squared loss $L(w) = \sum_{n=1}^N \|f(w, x^{(n)}) - y^{(n)}\|^2$ and we choose to set $N \ll d$ which corresponds to the underdetermined case: there are many solutions to $X\beta = y$.

We would like to show that the minimizer we reach depend on the scale of initialization $\alpha \in \mathbb{R}_+$. We denote $w_\alpha(t)$ the dynamics obtained by the gradient flow

$$\dot{w}_\alpha(t) = -\nabla L(w_\alpha(t)), \quad (8)$$

with the initial condition $w_\alpha(0) = \alpha w_0$ for $w_0 = 1_{2d} \in \mathbb{R}^{2d}$. In particular, we focus on the predictor $\beta_\alpha(t) = \beta_{w_\alpha(t)} = w_{\alpha,+}^2(t) - w_{\alpha,-}^2(t)$.

The goal of this exercise is to characterize $\beta_\alpha(\infty)$ i.e. the limit of $\beta_\alpha(t)$ when $t \to \infty$ which is given by

$$\beta_\alpha(\infty) = \underset{\beta}{\operatorname{argmin}}\, Q_\alpha(\beta) \text{ s.t. } X\beta = y, \quad (9)$$

where $Q_\alpha(\beta) = \sum_{i=1}^d q(\beta_i/\alpha^2)$ and $q(z) = \int_0^z \arcsin(u/2)du$.

Part I: Computing $\beta_\alpha(\infty)$

1) Rewrite the gradient flow equation on $w_\alpha(t)$ as

$$\dot{w}_\alpha(t) = -2\tilde{X} r_\alpha(t) \circ w_\alpha(t), \quad (10)$$

where $\tilde{X} := [X \ -X]$, $r_\alpha(t) := 2(\tilde{X} w_\alpha(t)^2 - y)$ and $a \circ b$ denotes the elementwise product of a and b. Solve the ODE (10).

2) Deduce that $\beta_\alpha(t)$ is equal to

$$\beta_\alpha(t) = \alpha^2 \left(\exp\left(-4X^\top \int_0^t r_\alpha(s)ds \right) - \exp\left(4X^\top \int_0^t r_\alpha(s)ds \right) \right). \quad (11)$$

In what follows, we assume that there is some $\bar{r}_\alpha \in \mathbb{R}^n$, such that $\bar{r}_\alpha = \int_0^\infty r_\alpha(s)ds$. This ensures that $\beta_\alpha(\infty)$ exists.

3) Write the KKT conditions (stationarity and primal feasibility) of the convex program (9).
4) Show that $\beta_\alpha(\infty)$ satisfies the stationarity condition. The problem satisfies the strict saddle property (Ge et al., 2015) therefore gradient flow will converge to a zero-error solution i.e. $X\beta_\alpha(\infty) = y$. The primal feasibility being satisfied, this shows that $\beta_\alpha(\infty)$ is solution to (9).

Part II: Interpretation

5) When $\alpha \to \infty$, show that $Q_\alpha(\beta) \approx \|\beta\|_2^2$. *Hint: Around* $z = 0$, $q(z) = \frac{z^2}{4} + O(z^4)$.
6) When $\alpha \to 0$, show that $Q_\alpha(\beta) \approx \|\beta\|_1 + o(1)$.
7) How do you interepret these results?

References

Chizat, Lenaic and Francis Bach (2018). "A note on lazy training in supervised differentiable programming." In: *arXiv preprint arXiv:1812.07956*.

Ge, Rong, Furong Huang, Chi Jin, and Yang Yuan (2015). "Escaping from saddle pointsfffdfffdfff-donline stochastic gradient for tensor decomposition." In: *Conference on Learning Theory*, pp. 797–842.

Jacot, Arthur, Franck Gabriel, and Clément Hongler (2018). "Neural tangent kernel: Convergence and generalization in neural networks." In: *Advances in neural information processing systems*, pp. 8571–8580.

Woodworth, Blake, Suriya Gunasekar, Jason Lee, Daniel Soudry, and Srebro Nathan (2019). "Kernel and Deep Regimes in Overparametrized Models." In: *arXiv preprint arXiv:1906.05827*.