MSRI Summer School	PRECEPT 2: MYSTERIES OF GRADIENT DESCENT
Instructor: J. Bruna	
TA: S. Jelassi	

Problem 1: Implicit bias of gradient descent in Logistic Regression

The goal of this exercise is to understand the implicit bias of gradient descent in a simple setting. Soudry et al., 2018 consider solving an underdetermined logistic regression problem on a linearly separable dataset by using gradient descent.

More formally, we consider a classification problem where we are given a dataset $\{x_n, y_n\}_{n=1}^N$ with $x_n \in \mathbb{R}^d$ and $y_n \in \{-1, 1\}$. We assume that the dataset is linearly separable i.e. $\exists w_* \in \mathbb{R}^d$ such that $\forall n : w_*^\top x_n > 0$. The problem is also assumed to be underdetermined i.e. $N < d$ which implies that there exist several hyperplanes that seperate the data. We aim at minimizing an empirical loss of the form

$$\min_{w \in \mathbb{R}^d} \mathcal{L}(w) := \sum_{n=1}^{N} \exp\left(-y_n \cdot w^\top x_n\right). \tag{1}$$

Remark that here we use the exp-loss (and not logistic regression) for simplicity. The results that we prove below also hold for the logistic regression. We also assume for simplicity that all the labels are positive: $y_n = 1$ for all $n \in [N]$– this is true without loss of generality, since we can always redefine $y_n x_n$ as x_n.

Part I: Convergence to the global minimum

1) Write the gradient descent update for the problem (1).

2) Show that there are no finite critical points $\bar{w} \in \mathbb{R}^d$ for which $\nabla\mathcal{L}(\bar{w}) = 0$.

Hint: Give an expression to $w_*^\top \nabla\mathcal{L}(w)$ for all $w \in \mathbb{R}^d$ and show that it cannot be equal to zero.

3) Determine the limit of $\nabla\mathcal{L}(w(t))$ when $t \to \infty$.

Hint: Use the fact that gradient descent on a smooth loss with an appropriate stepsize is always guaranteed to converge to a critical point.

4) From 2) and 3), deduce that the iterates of GD $w(t)$ satisfy $\lim_{t\to\infty} \|w(t)\| = \infty$.

5) Prove that gradient descent converges to the global minimum i.e. $\lim_{t\to\infty} \mathcal{L}(w(t)) \to 0$.

Part II: Implicit bias of Gradient Descent

In part I, we proved that norm of the predictor is not minimized, since it grows to infinity.

However, for prediction in a classification problem, only the direction of the predictor, the normalized $w(t)/\|w(t)\|$ matters. The goal of this part is to understand how $w(t)/\|w(t)\|$ behaves as $t \to \infty$.

In what follows, for the gradient descent iterate $w(t)$, we define $r(t) := w(t) - \hat{w}\log(t) - \tilde{w}$ where $\hat{w}$ is the L_2-max margin error (the solution to the hard margin SVM) i.e.

$$w_\infty = \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \|w\|^2 \ \text{ s.t. } \ w^\top x_n \geq 1. \tag{2}$$

and $\tilde{w}$ is a vector which satisfies

$$\sum_{n \in \mathcal{S}} \exp(-\tilde{w}^\top x_n) x_n = \hat{w}, \tag{3}$$

where $\mathcal{S} = \operatorname{argmin}_n \hat{w}^\top x_n$ (i.e. the set of vectors x_n such that $\hat{w}^\top x_n = 1$) are the support vectors associated to $\hat{w}$. For simplicity, we assume that such $\tilde{w}$ exists here. The goal of this part is to show that $\|r(t)\|$ is bounded.

6) Show that $r(t)$ satisfies the ODE:

$$\dot{r}(t) = -\nabla \mathcal{L}(w(t)) - \frac{1}{t}\hat{w}. \tag{4}$$

7) Deduce that

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\|r(t)\|^2 = &\sum_{n \in \mathcal{S}} \exp\left(-\log(t)\hat{w}^\top x_n - \tilde{w}^\top x_n - x_n^\top r(t)\right) x_n^\top r(t) - \frac{1}{t}\hat{w}^\top r(t) \\ &+ \sum_{n \notin \mathcal{S}} \exp\left(-\log(t)\hat{w}^\top x_n - \tilde{w}^\top x_n - x_n^\top r(t)\right) x_n^\top r(t). \end{aligned} \tag{5}$$

8) By using the fact that $\mathcal{S}$ is the set of vectors x_n such that $\hat{w}^\top x_n = 1$ and the definition of (3), show that the first term of (5) is equal to

$$\frac{1}{t}\sum_{n \in \mathcal{S}} \exp(-\hat{w}^\top x_n)(\exp(-x_n^\top r(t)) - 1) x_n^\top r(t).$$

What is the sign of this latter quantity? *Hint:* $\forall z$, $z(e^{-z} - 1) \leq 0$.

9) Using the fact that $ze^{-z} \leq 1$ and defining $\theta = \operatorname{argmin}_{n \notin \mathcal{S}} x_n^\top w > 1$, show that the second term of (5) is upper bounded by $\frac{1}{t^\theta}\sum_{n \notin \mathcal{S}} \exp(-\tilde{w}^\top x_n)$.

10) By plugging the results from 8) and 9) in (5), prove that there exist $C > 0$ such that for all $t_1 > 0$ and for all $t > t_1$, $\|r(t)\|^2 - \|r(t_1)\|^2 \leq C' < \infty$.

This proves that $\|r(t)\|$ is bounded which implies that $\rho(t) = r(t) + \tilde{w}$ is bounded. Therefore, we have proved that $w(t) = \hat{w}\log(t) + \rho(t)$ which implies that $\lim_{t \to \infty} \frac{w(t)}{\|w(t)\|} = \frac{\hat{w}}{\|\hat{w}\|}$.

Problem 2: Gradient Descent Only Converges to Minimizers

The goal of this exercise is to show in simple non-convex examples that gradient descent avoids saddle points and converge to local minimizers (Lee et al., 2016).

Part I: Non-convex quadratic function

In this part, we consider a non-convex quadratic $f(x) = \frac{1}{2}x^\top Hx$ where $H = \text{diag}(\lambda_1, \dots, \lambda_n)$ with $\lambda_1, \dots, \lambda_p > 0$ and $\lambda_{p+1}, \dots, \lambda_n < 0$.

1) Determine the critical points of f. Are they local minima, local maxima or saddle points?

2) Show that the Gradient Descent update on f when initialized in a point $x_0 \in \mathbb{R}^n$ and using stepsize $\eta > 0$ is

$$x_{k+1} = \sum_{i=1}^{n} (1 - \eta\lambda_i)^{k+1} \langle e_i, x_0 \rangle e_i,$$

where $\{e_i\}_{i=1}^n$ is the canonical basis in $\mathbb{R}^n$.

Gradient descent is guaranteed to converge with constant stepsize when $0 < \eta < \frac{2}{L}$ where $L = \max_i |\lambda_i|$. We assume here that $\eta < 1/L$.

3) For $i \in [n]$, compare $(1 - \eta\lambda_i)$ and 1.

4) What is the behavior of x_k when gradient descent is initialized in $x_0 \in E_s := \text{span}(e_1, \dots, e_p)$?

5) What happens if x_0 has a component outside E_s?

Part II: Non-convex non-quadratic function

Now, we consider the function $f(x, y) = \frac{1}{2}x^2 + \frac{1}{4}y^4 - \frac{1}{2}y^2$.

6) Determine the gradient mapping i.e. $g(x, y) = [x, y]^\top - \eta\nabla f(x, y)$.

7) Determine the critical points of f. Show that there is one saddle point and determine the nature of the other critical points.

8) If we choose $x_0 = [x, 0]^\top$ for some $x \in \mathbb{R}$, to what critical point does gradient descent converge?

9) By looking at the Hessian evaluated in the saddle point, make a connection between its eigenvalues and x_0 as defined in question 8).

N.B: In these two examples, we showed that the attractive set of saddle points of a function is spanned by E_s the set of eigenvectors corresponding to positive eigenvalues of the Hessian. If we choose our initial point at random, it can be showed that the probability of that point landing in E_s is zero. This gives some intuition on why gradient descent avoids saddle points and converges to minimizers. A proof in the general case can be found in (ibid.).

Problem 3: Overparametrization in ℓ_p regression

The goal of this exercise is to show that for linear regression with ℓ_p loss, a slight overparametrization can have an effect on the optimization (Arora et al., 2018). Consider the objective for a scalar linear regression problem with ℓ_p loss

$$L(w) = \mathbb{E}_{x,y\sim S}\left[\frac{1}{p}(x^\top w - y)^p\right], \tag{6}$$

where $x \in \mathbb{R}^d$ are examples, $y \in \mathbb{R}$ are labels and S is a training set and $w \in \mathbb{R}^d$ is a learned parameter vector. We make a simple overparametrization that consists in replacing the parameter vector w by a vector $w_1 \in \mathbb{R}^d$ times a scalar $w_2 \in \mathbb{R}$

$$L(w_1, w_2) = \mathbb{E}_{x,y\sim S}\left[\frac{1}{p}(x^\top w_1 w_2 - y)^p\right]. \tag{7}$$

Remark that the overparametrization does not affect the expressiveness of the linear model.
1) Write the gradient descent updates over $L(w_1, w_2)$ with a stepsize η. For convenience, we will use the notation $\nabla_{w(t)}$, $\nabla_{w_1^{(t)}}$ and $\nabla_{w_2^{(t)}}$ for the gradient of $L(w(t))$ with respect to $w(t)$, the gradient of $L(w_1(t), w_2(t))$ with respect to $w_1^{(t)}$ and the gradient of $L(w_1(t), w_2(t))$ with respect to $w_2^{(t)}$.
2) Using the previous question and assuming that η is small enough, show that

$$w^{(t+1)} = w^{(t)} - \rho^{(t)}\nabla_{w^{(t)}} - \gamma^{(t)}w^{(t)}, \tag{8}$$

where $\rho^{(t)} := \eta(w_2^{(t)})^2 \in \mathbb{R}$, $\gamma^{(t)} := \eta(w_2^{(t)})^{-1}\nabla_{w_2^{(t)}}$.
We assume that w_1 and w_2 are initialized near zero implying that w is also initialized near zero.
3) Show that there exist constants $\mu^{(t,\tau)} \in \mathbb{R}$ such that

$$w^{(t+1)} = w^{(t)} - \rho^{(t)}\nabla_{w^{(t)}} - \sum_{\tau=1}^{t-1}\mu^{(t,\tau)}\nabla_{w^{(\tau)}}. \tag{9}$$

4) Interpret the result obtained in equation (9).

N.B: We have seen in this exercise how overparametrization influences the optimization update in the case of the ℓ_p regression. ibid. also focus on overparametrization in the case of linear networks and show that depth acts as a preconditioner in the gradient descent update which may accelerate convergence.

References

Arora, Sanjeev, Nadav Cohen, and Elad Hazan (2018). "On the optimization of deep networks: Implicit acceleration by overparameterization." In: *arXiv preprint arXiv:1802.06509*.

Lee, Jason D, Max Simchowitz, Michael I Jordan, and Benjamin Recht (2016). "Gradient descent converges to minimizers." In: *arXiv preprint arXiv:1602.04915*.

Soudry, Daniel, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro (2018). "The implicit bias of gradient descent on separable data." In: *The Journal of Machine Learning Research* 19.1, pp. 2822–2878.