Exercises: Implicit bias of gradient descent

Pierre Ablin

We let $X \in \mathbb{R}^{n\times p}$ and $y \in \mathbb{R}^n$ with $n < p$. We consider the problem

$$\min_w f(w) = \frac{1}{2}\|Xw - y\|^2.$$

We assume that X is of rank n, so that $XX^\top$ is invertible.

1 Implicit bias of gradient descent

1.1

We define $w^* = X^\top(XX^\top)^{-1}y$. Show that w^* is a minimizer of f. What is the set of minimizers of f? Demonstrate that w^* is the minimizer of minimum norm.

We expect gradient descent to converge towards a point in that set; but the quesion is which one? We will answer this in the next questions

1.2

Let w_t the iterations of gradient descent with step size $\eta \leq \frac{1}{L}$ starting from $w_0 = 0$ for this problem. What recursion does this sequence verify?

Show that for all t, there exists $u_t \in \mathbb{R}^n$ such that $w_t = X^\top u_t$. What is the update equation on u_t?

1.3

What is the limit of u_t? Demonstrate that

$$\lim_{t\to+\infty} w_t = w^*$$

Therefore, we have shown that gradient descent on f does not converge to a random minimizer, it actually chooses the minimizer of minimum norm ! This is an *implicit* bias: the norm minimization is never explicitely stated in the optimization problem.

2 Link between early stopping and regularization

We consider $w^*(\lambda)$ the ridge regression solution :

$$w^*(\lambda) = \arg\min \frac{1}{2}\|Xw - y\|^2 + \frac{\lambda}{2}\|w\|^2$$

In the following, we define $K = XX^\top$. We recall that K is invertible.

2.1

Show that $w^*(\lambda) = X^\top(K + \lambda I_n)^{-1}y$.

2.2

We consider the gradient flow equation for gradient descent on the least squares problem:

$$\dot{w}(t) = -X^\top(Xw(t) - y)$$

starting from $w(0) = 0$. Show that $w(t) = X^\top(I - \exp(-Kt))K^{-1}y$

2.3

What can you say as $t = 0$, as $t \to +\infty$? Same thing for λ.

2.4

Show that $w^*(\frac{1}{t}) \simeq w(t)$ as t gets close to 0. At which order in t is the previous equality true?

As a consequence, early stopping gradient descent at time t is roughly the same thing as solving the ridge regression problem with $\lambda = \frac{1}{t}$: early stopping introduces a form of regularization, which is great for underdetermined problems.