IEOR 4601 - Recitation 1

.

1 Review on Linear Regression model.

We are given p observation outcomes $y_i \in \mathbb{R}$ and variables $x_i \in \mathbb{R}^n$, for $i = 1, ..., p$. We are looking for a prediction $f : \mathbb{R}^n \to \mathbb{R}$, such that $f(x_i) \approx y_i$. We consider the mean-square error

$$(MSE) \ = \ \frac{1}{p} \sum_{i=1}^{p} (y_i - f(x_i))^2.$$

We want the error MSE to be as small as possible, when we will consider new samples of $(x_\ell, y_\ell), \ell = p+1, ..., p+m$ (test set error) and on the set $(x_i, y_i), {}_1 = 1, ..., p$ (training set error). For example, we want to predict the willingness-to-pay of a customer as a function of the prices of the predict, the age of the customer, its income, etc. We already know past customers and if they bought or not, and we want to predict the willingness-to-pay of a given new customer. Looking for *any* function f is hard because the space of all functions $\mathbb{R}^n \to \mathbb{R}$ is very large (it has infinitely many dimensions). Therefore, we restrict our attention to the space of *affine* functions of the variables. Generally speaking, an affine map of $\mathbb{R}^n \to \mathbb{R}$ is a function f such that $f(x) = b^\top x + a$, where $a \in \mathbb{R}^n, b \in \mathbb{R}$. Therefore an affine function is only parametrized by $n+1$ components, which is far (actually infinitely) smaller than the space of all function from $\mathbb{R}^n$ to $\mathbb{R}$. As an example, you saw in class a model where the demand of a product is an affine (decreasing) function of the price of a product, ie, where

$$d(p) = a - b \cdot p.$$

The MSE then becomes, given past prices p_i (variable) and outcome d_i,

$$(MSE) \ = \ \frac{1}{p} \sum_{i=1}^{p} (d_i - (a - b \cdot p_i)^2.$$

In general, we want a model which is parametrized by as few components as possible; this was the very reason we introduced our affine model in the first place, in order to go from an infinite dimensional space (all function $\mathbb{R}^n \to \mathbb{R}$) to an $(n+1)$-dimensional space. Therefore, we want to penalize components who

are not zero. This can be done by adding a *regularization* term to our problem of minimizing the (MSE), ie, by considering errors in the flavor of

$$(MSE)_R \;=\; \frac{1}{p}\sum_{i=1}^{p}(d_i - (a - b\cdot p_i)^2 + \lambda\cdot\|(a,b)\|,$$

and we want to solve

$$\min_{(a,b)\in\mathbb{R}^+}\frac{1}{p}\sum_{i=1}^{p}(d_i - (a - b\cdot p_i)^2 + \lambda\cdot\|(a,b)\|, \tag{1}$$

where $\|\cdot\|$ can be any given norm. For $\|\cdot\| = \|\cdot\|_1$, we recover LASSO, which actually drops some components to zero, but for which the objective function becomes non-differentiable. For $\|\cdot\| = \|\cdot\|_2$, we recover Ridge regularization, for which the objective function is still differentiable (and becomes strongly convex depending of the parameter λ), but which only *reduces* the components, without setting them to zero.

2 Optimization.

In the class we will consider the following general optimization problem:

$$\min_{x\in\mathbb{X}} f(x) \tag{2}$$

where $f:\mathbb{R}^n\to\mathbb{R}$ is a *differentiable* function and $\mathbb{X}$ is a convex subset of $\mathbb{R}^n$. A function f is said to be differentiable if its *gradient* $\nabla f(x) = \left(\dfrac{\partial f}{\partial x_i}(x)\right)_{i\in[1,n]}$ exists for every point $x\in\mathbb{R}^n$. Note that some important function are non differentiable: for instance, $x\mapsto\|x\|_1$ is not differentiable at $x=0$ (in $n=1$ this is just the absolute value).

As an example you already saw the case where $n=1$ and $x=p$ represents the price of our product, and $f(p) = p\cdot(a-bp)$. The problem of fitting our model to the data is the optimization program (1).

The set of solution to problem (2) is called

$$\arg\min_{x\in\mathbb{X}} f(x),$$

and is defined by

$$x^*\in\arg\min_{x\in\mathbb{X}} f(x) \iff f(x^*) = \min_{x\in\mathbb{X}} f(x) \iff f(x^*)\le f(x), \forall x\in\mathbb{X}.$$

(one of) The main theorem of optimization is the following:

Theorem 2.1. *Let f a differentiable function and $\mathbb{X}$ a convex set. Then*

$$\arg\min_{x\in\mathbb{X}} f(x) \subseteq \{x\in int(\mathbb{X}) \mid \nabla f(x) = 0\}\bigcup\partial\mathbb{X},$$

where $\partial\mathbb{X} = cl(\mathbb{X})\backslash int(X)$ is the boundary or $\mathbb{X}$.

In small dimension the boundary of a set $\mathbb{X}$ is a simple object, and you can think of it as the line delimiting the frontier between the set $\mathbb{X}$ and $\mathbb{R}^n$. In dimension 1, the boundary of an interval $[a, b]$ is the set $\{a, b\}$. In dimension 2 for instance, the boundary of the disk $\{(x, y) \mid x^2 + y^2 \leq 1\}$ is the circle $\{(x, y) \mid x^2 + y^2 = 1\}$.

I emphasize Theorem 2.1 because we will see some examples in the class where the optimal solutions of the optimization program (2) are not attained at points where the gradient vanishes. A very simple example is the case where $n = 1, \mathbb{X} = [0, 1], f(x) = x$. For all $x \in [0, 1], f'(x) = 1$ and therefore there is no point that set the gradient (i.e., the derivative in this case) to zero. However, the minimum of the function is obviously attained at 0, which belongs to the boundary of $[0, 1]$, because $\partial([0, 1]) = \{0, 1\}$.

Take-away: Try to set the gradient to 0 when optimizing a problem in the form of (2). If there is no point that set the gradient to 0, that does not mean that program (2) has no solution, this only means that you need to consider the points at the border of $\mathbb{X}$.