LINMA2471 - Optimization
Cours 5

S. Cheng, Q. Le, F. Delcourt

October 16, 2015

We've been investing in convex models and trying to find the best formulation of optimization problems. We now explore different methods that allow us to solve those problems and study their properties. In this part we focus on first order methods. Second order methods will be discussed further.

First order methods

1 Gradient Method

A main example of first order methods is the Gradient Method also known as the Steeped Descent Method. Many practical problems have constraints but let's consider for the moment that we have none.

Problem: $\min_{x \in \mathbb{R}^n} f(x)$

Algorithm:

Given x_0; k=0
Repeat
$x_{k+1} = x_k - \underbrace{h_k}_{\in \mathbb{R}} \underbrace{\nabla f(x_k)}_{\in \mathbb{R}^n}$
$k \leftarrow k+1$

h_k is called the step length and $-\nabla f(x_k)$ is called the direction.

1.1 Step length selection

1.1.1 h_k that minimize $f(x_k - h_k \nabla f(x_k))$

We have to solve for h_k at each step. Even if we only minimize one variable, it's still an iterative method and does not give directly h_k.

1.1.2 $h_k = \alpha$

Example 1 Let's consider x^2 and see what happens when we change h_k

$\boxed{h_k = 2}$ $\forall k$

Set x_0

$x_1 = x_0 - 2(2x_0) = -3x_0$

$x_2 = -3x_0 - 2(-6x_0) = 9x_0$

Instead of going to zero we go to $-3x_0$ or $9x_0$. Our step is clearly too large. This is called DIVERGING.

$\boxed{h_k = 1}$ $\forall k$

$x_1 = x_0 - 2x_0 = -x_0$

$x_2 == -x_0 - (-2x_0) = x_0$

This time, we are still too large and this is called CYCLING.

$\boxed{h_k = \frac{1}{2}} \quad \forall k$
$x_1 = x_0 - \frac{1}{2}(2x_0) = 0$
We have a FINITE CONVERGENCE.
$\boxed{h_k = \frac{1}{3}} \quad \forall k$
$x_1 = x_0 - \frac{1}{3}(2x_0) = \frac{1}{3}x_0$
$x_2 = \frac{1}{3}x_0 - \frac{1}{3}(2\frac{1}{3}x_0) = \frac{1}{9}x_0$
$\vdots$
$x_k = \frac{1}{3^k}x_0$

We get closer to x_0.

The simple gradient method is subjected to poor step length selection. However it still good to use a constant step length. We'll explicitly and numerically compute some value of α which guarantee a good behavior.

1.1.3 h_k satisfies some "dynamic" conditions (e.g.Wolfe condition)

In order to know what is the best step length, we need to know the function. We focus only on functions that are $C_L^{1,1}$, which means that $f \in \mathbb{C}^1$ and ∇f is Lipschitz with constant L. The Lipschitz condition is:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

In order words, if we take 2 points, their gradient can't be too different.

Example 2 Function that are Lipschitz with L=0? Linear function are $C_0^{1,1}$

Example 3 Standard quadratic function $x \to x^T Qx$ with no convexity assumption for now. The Lipschitz condition can be expressed as:

$$\|2Qx - 2Qy\| \leq L\|x - y\| \quad \forall x, y$$

$$2\|Q(x - y)\| \leq L\|x - y\| \quad \forall x, y$$

If $Q = Id$, it's clear that $L = 2$. When it comes to matrix like Q, we will use the spectral norm. From the matrix theory, we know that:

$$\|Qv\|_2 \leq \underbrace{\|Q\|_2}_{max|\lambda_i(Q)|} \|v\|_2$$

.

One can hence choose

$$L = 2\|Q\|_2$$

.

For more complicated function, we can try to bound the value of L and use its estimate.

As we can see, it's very easy to compute the constant for a quadratic function. But what can we do in a more general case?

Proposition 1 When f $\in C^2$ the Lipschitz constant is given by $L = \max_x \|\nabla^2 f(x)\|$.

This result is very useful for one variable functions but not for multi variables functions. The computation can be hard in that case.

Example 4 Let's apply propositon 1 with $f : x \to \sqrt{1+x^2}$

$$f'(x) = \frac{x}{\sqrt{1+x^2}}$$
$$f''(x) = \frac{\sqrt{1+x^2} - x\frac{x}{\sqrt{1+x^2}}}{1+x^2}$$

$$0 \leq \frac{1}{(1+x^2)(\sqrt{1+x^2})} \leq 1$$

The upperbound is clear and we also put the lowerbound because we want the absolute value of $f''(x)$ to be bounded. We will take $L = 1$.

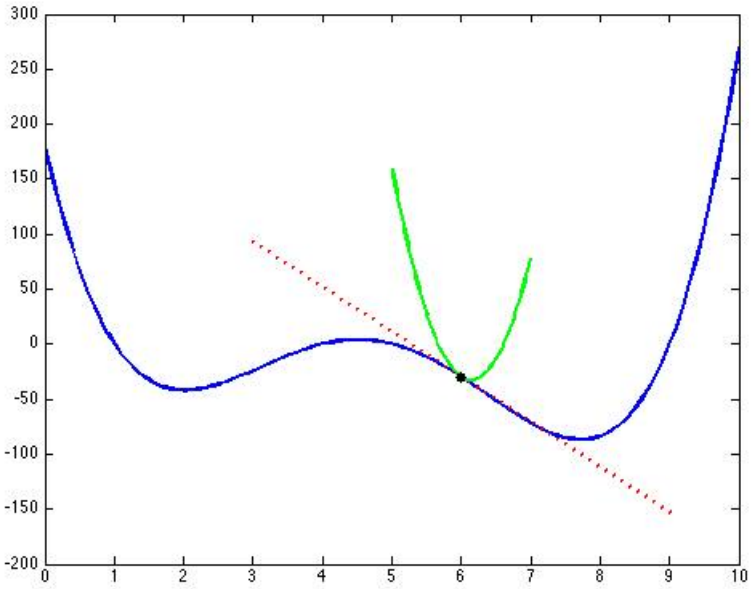

Figure 1: In blue: f of $C_L^{1,1}$, in red: tangeant to the function, in green: the QUB

What can we say about the minimum of a function based only on its first derivative? We are tempted to use the Taylor series which approximate the function by its tangent. Unfortunatly we can't say anything from this about the minima. One better idea is to construct an upper bound approximation of our function using the Lipschitz constant then try to minimize this bound. We are thus sure that our fuction will stay below this bound, as shown in figure 1.

Lemma 1 **Quadratic upper bound on $C_L^{1,1}$ function**
Let's $f \in C_L^{1,1}$. Then for any x we have that

$$\hat{f} : y \mapsto f(x) + \nabla f^T(x)(x-y) + \frac{L}{2}\|x-y\|$$

is an upper bound of $f(y)$.

Notice that $\hat{f}$ is a quadratic function of y. What is the minimum of the quadratic upper bound (QUB)? Our function is now convex and easy to minimize over y.

$$\nabla \hat{f}(y) = 0 \Leftrightarrow \nabla f(x) + \frac{L}{2} 2(x-y)(-1) = 0$$

$$\Rightarrow y^* = x - \frac{1}{L}\nabla f(x)$$

We can notice that the last equation is very similar to the gradient method. It is sensible to use $h_k = \frac{1}{L}$ in our method.

1.2 Analysis of the $h_k = \frac{1}{L}$ gradient method

After one step, taking y= x-$\frac{1}{L}$ ∇ f(x), the upper bound gives:

$$f(x) + \nabla f(x)^T[\frac{-1}{L}\nabla f(x)] + \frac{L}{2}\|\frac{-1}{L}\nabla f(x)\|^2 = f(x) - \frac{\|\nabla f(x)\|^2}{L} + \frac{1}{2L}\|\nabla f(x)\|^2$$
$$= f(x) - \underbrace{\frac{1}{2L}\|\nabla f(x)\|^2}_{\leq 0}$$

Looking at all iterations now:

$$f(x_{k+1}) - f(x_k) \leq -\frac{\|\nabla f(x)\|^2}{L}$$

(because the upper bound displays this decrease)

$$f(1) - f(0) \leq -\frac{\|\nabla f(x_0)\|^2}{L}$$
$$f(2) - f(1) \leq -\frac{\|\nabla f(x_1)\|^2}{L}$$
$$\vdots$$
$$f(N+1) - f(N) \leq -\frac{\|\nabla f(x_N)\|^2}{L}$$

summing the equations above, we get

$$f(N+1) - f(0) \leq -\frac{1}{L}\sum_{i=1}^{n} \|\nabla f(x_i)\|^2$$

so we deduce

$$L(f(0) - f(N+1)) \geq \sum_{i=1}^{n} \|\nabla f(x_i)\|^2 \geq (N+1) min \|\nabla f(x_i)\|^2$$

Hence

$$\min_{i=0,\ldots,N} \|\nabla f(x_i)\| \leq \sqrt{\frac{L(f(0) - f(N+1))}{N+1}}$$
$$\leq \sqrt{\frac{L(f(0) - f(x*))}{N+1}}$$

with x* minimum point of f.

Lemma 2 Assume in addition that f is convex (meaning $f \in \mathbb{F}_L^{1,1}$) [1], then we have:

$$f(N) - f(x*) \leq \frac{L\|x_0 - x*\|^2}{2N}$$

(assuming x* is one of the minimum points).

Example 5 Behavior of the gradient method

$$2x_2 + y_4 - 2y_2$$

we start from (1,0)

$$\nabla = \begin{pmatrix} 4x \\ 4y_3 - 4y \end{pmatrix}$$

$$\nabla^2 = \begin{pmatrix} 4x & 0 \\ 0 & 12y_2 - 4 \end{pmatrix}$$

[1] f is convex and $f \in C_L^{1,1}$

We have three stationary points (0,0),(0,1),(0,-1).

We observe that

$$\nabla(1,0) = \begin{pmatrix} 4 \\ 0 \end{pmatrix}$$

If we start from (1,0), we can only converge to (0,0) and we couldn't find all the points.