1 17th of October 2018 — A. Frangioni

1.1 Optimization algorithms

Do you recall?

We are interested in finding the minimum of a function through an iterative procedure, such that we start form an initial guess x_0 and go on $x^i \rightsquigarrow x^{i+1}$. We want to move towards the otimum.

How to be sure we are in an optimum?

- (strong) $\{x^i\} \to x_*$: the whole sequence converges to an optimal solution;
- (weaker) all accumulation points of $\{x^i\}$ are optimal solutions;
- (weakest) at least one accumulation point of $\{x^i\}$ is optimal.

Such iterative process, can be held in two different forms:

LINE SEARCH: first choose $d^i \in \mathbb{R}^n$ (direction), then choose $\alpha^i \in \mathbb{R}$ (that we term stepsize or equivalently "learning rate") s.t. $x^{i+1} \leftarrow x^i + \alpha^i d^i$;

TRUST REGION: first choose α^i (trust radius), then choose d^i.

In both these alternatives, it is crucial to choose a proper model to approximate function f.

1.1.1 Gradient method for quadratic functions

The simplest model we can build is a linear one, namely $L^i(x) = L_{x^i}(x) = f(x^i) + \nabla f(x^i)(x - x^i)$ and find the direction according to the model.

We should not move too far from x^i, so we want $\alpha_i \to 0$ and then $d_i = argmin\{\lim_{t\to 0} \frac{f(x+td)}{t}\} = -\nabla f(x^i)$, aka the steepest descent direction.

Notice that a too short step is bad either, because the gain in the value of the function is very little.

At each step we want $\alpha^i \in \text{argmin}\{f(x^i + \alpha d^i) \ : \ \alpha \geq 0\}$.

Linear functions are unbounded below, so we would like to use a different family of functions. The easiest family of functions which are still more complex than linear ones are quadratic functions.

$$f(x) = \frac{1}{2} x^T Q x + qx$$

where $Q \succeq 0$ otherwise f is unbounded below.

The minimum here is the point in which the gradient ($\nabla f(x) = Qx + q$) is 0.

ALGORITHM 1.1 Pseudocode for quadratic functions local minimum detection.

```
1: procedure SDQ(f, x, ε)
2:     while (‖∇f(x)‖ > ε) do
3:         d ← −∇f(x);
4:         α ← ‖d‖² / (dᵀQd);
5:         x ← x + αd;
6:     end while
7: end procedure
```

For the time being we do not go into detail of how to chose the ε such that we can stop when the norm of the gradient is smallert than such a constant.

Let us see how to obtain the formula for α.

We are interested in computing the minimum of $\{f(x^i + \alpha d^i) : \alpha \geq 0\}$.

Let us do come algebra to describe better such an f:

$$\begin{aligned} f(x^i + \alpha d^i) &= \frac{1}{2}(x^i + \alpha d^i)^T Q(x^i + \alpha d^i) + q(x^i + \alpha d^i) \\ &= \cancel{\frac{1}{2}(x^i)^T Q x^i} + (x^i)^T Q(\alpha d^i) + \frac{1}{2}(d^i)^T Q d^i + \cancel{q x^i} + \alpha(q d^i) \\ &= [\frac{1}{2}(d^i)^T Q d^i]\alpha^2 + \alpha[(x^i Q + q)d^i] \\ &\overset{(1)}{=} [\frac{1}{2}(d^i)^T Q d^i]\alpha - \left\|d^i\right\| \end{aligned} \tag{1.1}$$

What if $(d^i)^T Q d^i = 0$? If Q is strictly positive definite this cannot happen, so the algorithm never breaks down.

Can we prove that the sequence of iterates is moving towards the optimum?

For this proof let us assume that $\varepsilon = 0$, hence the procedure will neve stop. We want to prove that the sequence $\{x^i\}$ is (or contains) a minimizing sequence.

First of all we can state that the sequence is monotone, so it has a limit for sure.

What we can prove is that the the point where the sequence is converging is a stationary point. $\lim_{i\to\infty} < \nabla f(x^i), \nabla f(x^{i+1}) >= 0 \overset{(*)}{=} < \nabla f(x), \nabla f(x) >$ and this holds because of the fundamental relationship $< \nabla f(x^i), \nabla f(x^{i+1}) >= 0$.

Notice that $\overset{(*)}{=}$ follows from the fact that the gradient is continuous.

How fast is the convergence?

In general, showing how fast $\| x^i - x_* \|$ decreases is more involved that showing how fast $f(x^i) - f_*$ decreases, but we do not know f_*.

We concentrate on computing $\lim_{i\to\infty} \frac{f(x^{i+1}) - f_*}{f(x^i) - f_*{}^p} = R$.

According to the values of p and R we get the following alternatives:

SUBLINEAR: $p = 1, R = 1$;

LINEAR: $p = 1, R < 1$;

SUPERLINEAR: $p = 1, R = 0$;

QUADRATIC: $p = 2, R > 0$.

Since the optimum is in $x^* = -Q^{-1}q$, we get that $f(x^*) = \frac{1}{2}q^TQ^{-1}QQ^{-1}q - q^TQ^{-1}q = \frac{1}{2}q^TQ^{-1}q - q^TQ^{-1}q = -\frac{1}{2}q^TQ^{-1}q$.

Let us use a nifty trick: let us define:

$$\begin{aligned}
\bar{f}(x) &= \frac{1}{2}(x - x_*)^TQ(x - x_*) \\
&= \frac{1}{2}x^TQx + \frac{1}{2}{x_*}^TQx_* - x^T(Qx) \\
&\overset{(2)}{=} \frac{1}{2}x^TQx + \frac{1}{2}Q^{-1}{q}^TQ(Q^{-1}q) + qx \\
&= \frac{1}{2}x^TQx + \frac{1}{2}q^T\cancel{Q^{-1}}\cancel{Q}Q^{-1}q + qx \\
&= \frac{1}{2}x^TQx + \frac{1}{2}q^TQ^{-1}q + qx \\
&= f(x) - f_*
\end{aligned} \tag{1.2}$$