1 14th of November 2018 — A. Frangioni

Mantra

If you want better convergence, use a better model.

So far we chose the direction for the step as $d^i = -\nabla f(x^i)\,\|\nabla f(x^i)\|$, in particular this is the direction where the decrease of the function is maximum.

We introduced the **convergence argument** which says that in proximity of the stationary point of the linear model the norm of the gradient goes to 0. Formally, $\varphi'(0) = \frac{\partial f}{\partial d^i}(x^i) =< d^i, \nabla f(x^i) >= - < \nabla f(x^i), \nabla f(x^i) >= - \|\nabla f(x^i)\|$, which implies that $\|\nabla f(x)\| \to 0$ when $\varphi'(0) \to 0$.

Can we take another direction which isn't the opposite of the gradient and have that the same argument holds? Yes, for example if we choose as direction a rotation of the opposite of the gradient, the value of φ' is then the cosine of the angle of the rotation times the opposite of the norm fo the gradient. It's trivial to observe that there are infinite angles that could be chosen, so we have a lot of flexibility.

Note that the angle between d^i and the gradient shouldn't be too close to 90°, otherwise the cosine would get approximately 0.

Theorem 1.1 (Zoutendijk)**.** *Let $f \in C^1$, ∇f Lipschitz and f bounded below.*
If (A) ∩ (W') then $\sum_{i=1}^{\infty} \cos^2(\theta^i)\|\nabla f(x^i)\|^2 < \infty$.

If we have a positive infinite sequence and we have that the corresponding series converges to a number, than the limit of the sequence is going to 0 reasonably fast.

If we choose an angle which is bounded below then the norm of the gradient has to converge very fast. More formally,

Fact 1.2. $\cos(\theta^i) \geq \varepsilon > 0 \Longrightarrow \|\nabla f(x^i)\| \to 0$

Proof. From the observation above, we may recall that the n-th term of the Zoutendijk sum should be approximately 0 $\forall n > \tilde{n}$, which means that one between $\cos(\theta^i)$ and $\|\nabla f(x^i)\|^2$ should be zero.

The proof is obtained from the fact that it can't be the cosine, because it's bounded below. □

1.1 Taylor method

At this point we assume that the function is perfectly convex (Hessian strictly positive definite).

Theorem 1.3. *Let f be a function such that $\nabla^2 f(x^i) \succ 0$. Then the second order model $Q_{x^i}(y)$ admits a minimum.*

Proof.
For simplicity of notation let us forget about the index ^{i} of x^i. The Taylor expansion is the following:

$$\begin{aligned} Q_x(y) &= f(y) + < \nabla f(x), y - x > + \frac{1}{2}(y-x)^T \nabla^2 f(x)(y-x) \\ &= f(x) + < \nabla f(x), y > - < \nabla f(x), x > + \\ &+ \frac{1}{2}\left(y^T \nabla^2 f(x) y - 2x^T \nabla^2 f(x) y + x^T \nabla^2 f(x) y \right) \\ &\overset{*}{=} < \nabla f(x) + x^T \nabla^2 f(x), y > + \frac{1}{2} y^T \nabla^2 f(x) y \end{aligned} \tag{1.1}$$

Where $\overset{*}{=}$ is given by the fact that there are some contant terms.
$qy + \frac{1}{2} y^T Q y \Leftrightarrow q + Qy = 0$ so

$$\nabla f(x) - x^T \nabla^2 f(x) + \nabla^2 f(x) y = 0 \nabla^2 f(x) y = \nabla f^2(x) x - \nabla f(x) y = x - \left[\nabla^2 f(x)\right]^{-1} \nabla f(x)$$

□

Scrivere meglio la dim

Corollary 1.4. *Newton's direction is* $d^i \leftarrow -[\nabla^2 f(x^i)]^{-1} \nabla f(x^i)$ *(just* $\mathbb{R}^n$ *version).*

The direction d is obtained taking the opposite of the inverse of the Hessian times the gradient of the function

Observation 1.1. *We need the Hessian to be invertible, which isn't true in general.*

We need to solve a non linear equation, namely a system of equations. A way to circumvent this problem is putting the gradient to 0, so we can write the Taylor form of the gradient and solve a linear equation which is $\nabla f(x) \approx \nabla f(x^i) + \nabla^2 f(x^i)(x - x^i)$.

Theorem 1.5. *Let* f *be a function s.t.* $f \in C^3$, $\nabla f(x_*) = 0$ *and* $\nabla^2 f(x_*) \succ 0$. *Then* $\exists \mathcal{B}(x_*, r)$ *s.t.* $x^1 \in \mathcal{B}$, *which implies* $\{x^i\} \to x_*$ *quadratically.*

We may observe that the direction of the Newton method is good not only when we are close to the minimum, but also when we are far.

The scalar product should be negative and it is so, but we also need to ensure that the scalar product isn't too close to 0. When is it that this condition is satisfied? When the function is "reasonable".

Let us now introduce what we mean by "reasonable": a function f such that $uI \preceq \nabla^2 f \preceq LI$, which implies that the function is strongly convex, in other words that the eigenvalues of the Hessian don't get too close to zero (numerically speaking) .

Theorem 1.6. *Let* f *be a function that satisfies* $uI \preceq \nabla^2 f \preceq LI$ *and* $\cos(\theta^i) \leq -\lambda^n/\lambda^1 \leq -u/L$. *Then the method converges.*

Proof.

STEP 1 From the definition of d^i we have that:

$$d^i = -[\nabla^2 f(x^i)]^{-1} \nabla f(x^i)$$

or, equivalently:

$$\nabla^2 f(x^i) d^i = -\nabla f(x^i)$$

which implies:

$$d^i \nabla f(x^i) = -(d^i)^T \nabla^2 f(x^i) d^i \leq -\lambda^n \left\| d^i \right\|^2$$

STEP 2 Since we have that $\cos(\theta^i) = d^i \nabla f(x^i)(\|x^i\| \, \|\nabla f(x^i)\|) \leq \delta < 0$, we want to bound the norm of the gradient:

$$\left\| \nabla f(x^i) \right\| = \left\| \nabla^2 f(x^i) d^i \right\| \leq \left\| \nabla^2 f(x^i) \right\| \left\| d^i \right\| = \lambda^1 \left\| d^i \right\|$$

which implies:

$$\cos(\theta^i) \leq -\lambda^n / \lambda^1 \leq -u/L$$

□

Theorem 1.7. *Let f be a function that satisfies $uI \preceq \nabla^2 f \preceq LI$ and $\cos(\theta^i) \leq -\lambda^n/\lambda^1 \leq -u/L$. Then the method not only converges, but we also have that for some iteration onwards, $\alpha^i = 1$ always satisfy (A).*

Proof.

$$\begin{aligned} f(x^i + d^i) &= f(x^i) + d^i \nabla f(x^i) + \frac{1}{2}(d^i)[\nabla^2 f(x^i)] d^i + R(\left\| d^i \right\|) \\ &= f(x^i) - \nabla f(x^i)^T [\nabla^2 f(x^i)]^{-1} \nabla f(x^i) + \\ &+ \frac{1}{2} \nabla f(x^i)^T [\nabla^2 f(x^i)]^{-1} \nabla f(x^i) + R(\left\| d^i \right\|) \\ &= f(x^i) - \frac{1}{2} \nabla f(x^i)^T [\nabla^2 f(x^i)]^{-1} \nabla f(x^i) + R(\left\| d^i \right\|) \\ &= f(x^i) + \frac{1}{2} < \nabla f(x^i), d^i > + R(\left\| d^i \right\|) \end{aligned} \tag{1.2}$$

□

It can be proved that the convergence is superlinear.

If we start with a step size of 1, we end up in a situation in which the line search isn't computed when we are close to the minimum.

This works under the assumption that the eigenvalues are bounded both above and below (deriving from the bounds on the Hessian).

1.1.1 Interpretation of Newton method

Let us consider the Newton method from a different point of view. We can construct a different space where the gradient method coincides with the Newton method. Given R which is not singular, we make a variable change ($y = Rx$) — which is possible given that R is non singular — and we get $f_y(y) = \frac{1}{2}y^T I y + qR^{-1}y$, which has as an Hessian the identity metrix, which is the optimal matrix for convergence in Newton method.

Formally, the descending direction d_y is computed as follows: $d_y = -\nabla f_y(y) = -y - R^{-1}q$. Since we chose 1 as step size we obtain that $\nabla f_y(y+d_y) = \nabla f_y(y-y-R^{-1}q) = \nabla f_y(-R^{-1}q) = 0$.

It takes only one iteration, because all the eigenvalues are 1 so the the ratio between the greatest and the smallest is 1 and the subtraction is 0.

If we do the inverse operation ($x = R^{-1}y$) to the direction we get the direction in x variable.

Problem:

We made a lot of assumptions on the Hessian, without the ones there's no guarantee that we are moving on a descending direction. How can we relax these constraints?

1.1.2 Non convex case

If the Hessian isn't positive definite we may take something which is close to the Hessian, which has the bounding property we used before (eigenvalues bounded below and above). In truth, there is no reason to choose exactly the opposite of the inverse of the Hessian times the gradient for the direction, we may use another matrix which is striclty positive definite ad has more or less th same properties of the Hessian.

In particular, given an Hessian that has some negative eigenvalues we can sum to it a multiple of the identity matrix: $H^i = \nabla^2 f(x^i) + \varepsilon^i I \succ 0$. We iterate this procedure until the resulting matrix is strictly positive definite.

How to compute ε numerically? How much larger should ε be? We also want the smallest eigenvalue to be not too close to 0.

$\varepsilon = \max\{0,\ \delta - \lambda^n\}$ for "appropriately chosen smallish δ", in other words we want all the eigenvalues to be at least δ.

The reasons behind the choice of δ (not too small) are bothnumerical (any double $\leq$ `1e-16` "is 0") and algorithmical (if $\lambda^n(\nabla^2 f(x^i) + \varepsilon I)$ "very small" then the axes of $S(Q_{x^i}, \cdot)$ are "very elongated").

It can be proved that the ε we chose is the solution to an optimization problem: $\min\{\|H - \nabla^2 f(x^i)\| \ \ H \succeq \delta I\}$. The choice for δ in the code is 10^{-6}.

Observation 1.2. *Note that these contraints are important and we will get back to them later on in the course.*

Could we use a different norm instead of the 2-norm? Yes, for example we can use Frobenius norm, changing a bit the algorithm, but we still get convergence. We would need

to solve $\min\{\|H - \nabla^2 f(x^i)\|_F \; H \succeq \delta I\}$, which is performed in two steps:

1. compute spectral decomposition $\nabla^2 f(x^i) = H\Lambda H^T$
2. $H^i = H\bar{\Lambda}H^T$ with $\bar{\gamma}^i = \max\{\lambda^i, \delta\}$

In both cases, $\{x^i\} \to x_*$ with $\nabla^2 f(x_*) \succeq \delta I$, which implies $\varepsilon^i = 0$ and $H^i = \nabla^2 f(x^i)$ eventually. It holds also that we have superlinear convergence if "H^i looks like $\nabla^2 f(x^i)$ along d^i", formally $\lim_{i\to\infty} \|(H^i - \nabla^2 f(x^i))d^i\| \, \|d^i\| = 0$.

Computational complexity

We still need to compute eigenvalues, which takes $O(n^3)$, which is too much if we are in multidimensional spaces.

As a closing observation we may notice that Newton method is very fast to go to a local minimum and this may represent a problem, because it misses global minima.