Non linear Least Squares

Lectures for PHD course on Numerical optimization

Enrico Bertolazzi

DIMS – Universitá di Trento

November 21 – December 14, 2011

Outline

- An important class on minimization problem when $\mathsf{f} : \mathbb{R}^n \mapsto \mathbb{R}$ is the nonlinear least squares and takes the form:

$$\mathsf{f}(\boldsymbol{x}) = \frac{1}{2}\sum_{i=1}^{m} F_i(\boldsymbol{x})^2, \qquad m \geq n$$

- When $n = m$ finding the minimum coincide to finding the solution of the non linear system $\mathbf{F}(\boldsymbol{x}) = \mathbf{0}$ where:

$$\mathbf{F}(\boldsymbol{x}) = \big(F_1(\boldsymbol{x}), F_2(\boldsymbol{x}), \ldots, F_n(\boldsymbol{x})\big)^T$$

- Thus, special methods developed for the solution of nonlinear least squares can be used for the solution of nonlinear systems, but not the converse if $m > n$.

Example

Consider the the following fitting model

$$M(\boldsymbol{x}, t) = x_3 \exp(x_1 t) + x_4 \exp(x_3 t)$$

which can be used to fit some data. The model depend on the parameters $\boldsymbol{x} = (x_1, x_2, x_3, x_4)^T$. If we have a number of points

$$(t_k, y_k)^T, \qquad k = 1, 2, \ldots, m$$

we want to find the parameters $\boldsymbol{x}$ such that $\frac{1}{2} \sum_{k=1}^{m} (M(\boldsymbol{x}, t_k) - y_k)^2$ is minimum. Defining

$$F_k(\boldsymbol{x}) = M(\boldsymbol{x}, t_k) - y_k, \qquad k = 1, 2, \ldots, m$$

then can be viewed as a non linear least squares problem.

- To solve nonlinear least squares problem, we can use any of the previously discussed method. For example BFGS or Newton method with globalization techniques.
- If for example we use Newton method we need to compute

$$\nabla^2 \mathsf{f}(\boldsymbol{x}) = \nabla^2 \frac{1}{2} \sum_{i=1}^{m} F_i(\boldsymbol{x})^2 = \frac{1}{2} \sum_{i=1}^{m} \nabla^2 F_i(\boldsymbol{x})^2$$

$$= \frac{1}{2} \sum_{i=1}^{m} \nabla (2F_i(\boldsymbol{x}) \nabla F_i(\boldsymbol{x}))^T$$

$$= \sum_{i=1}^{m} \nabla F_i(\boldsymbol{x})^T \nabla F_i(\boldsymbol{x}) + \sum_{i=1}^{m} F_i(\boldsymbol{x}) \nabla^2 F_i(\boldsymbol{x})$$

- If we define

$$\boldsymbol{J}(\boldsymbol{x}) = \begin{pmatrix} \nabla F_1(\boldsymbol{x}) \\ \nabla F_2(\boldsymbol{x}) \\ \vdots \\ \nabla F_m(\boldsymbol{x}) \end{pmatrix}$$

then we can write

$$\nabla^2 \mathsf{f}(\boldsymbol{x}) = \boldsymbol{J}(\boldsymbol{x})^T \boldsymbol{J}(\boldsymbol{x}) + \sum_{i=1}^{m} F_i(\boldsymbol{x}) \nabla^2 F_i(\boldsymbol{x})$$

- However, in practical problem normally $\boldsymbol{J}(\boldsymbol{x})$ is known, while $\nabla^2 F_i(\boldsymbol{x})$ is not known or impractical to compute.

- A common approximation is given by neglecting the terms $\nabla^2 F_i(\boldsymbol{x})$ obtaining,

$$\nabla^2 \mathsf{f}(\boldsymbol{x}) \approx \boldsymbol{J}(\boldsymbol{x})^T \boldsymbol{J}(\boldsymbol{x})$$

- This choice can be appropriate near the solution if $n = m$ in solving nonlinear system. In fact near the solution we have $F_i(\boldsymbol{x}) \approx 0$ so that the contribution of the neglected term is small.
- This choice is not good when near the minimum we have large residual (i.e. $\|\mathbf{F}(\boldsymbol{x})\|$ is large) because the contribution of $\nabla^2 F_i(\boldsymbol{x})$ cant be neglected.

From previous consideration applying Newton method to $\nabla f(\boldsymbol{x})^T = \mathbf{0}$, we have

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \nabla^2 f(\boldsymbol{x}_k)^{-1} \nabla f(\boldsymbol{x}_k)^T$$

and when $f(\boldsymbol{x}) = \frac{1}{2} \|\mathbf{F}(\boldsymbol{x})\|^2$:

$$\nabla f(\boldsymbol{x})^T = \boldsymbol{J}(\boldsymbol{x}) \mathbf{F}(\boldsymbol{x})$$

$$\nabla^2 f(\boldsymbol{x}) = \boldsymbol{J}(\boldsymbol{x})^T \boldsymbol{J}(\boldsymbol{x}) + \sum_{i=1}^{m} F_i(\boldsymbol{x}) \nabla^2 F_i(\boldsymbol{x}) \approx \boldsymbol{J}(\boldsymbol{x})^T \boldsymbol{J}(\boldsymbol{x})$$

And using the last approximation we obtain the Gauss-Newton algorithm.

Notice that the approximate Newton direction

$$\boldsymbol{d} = -\Big(\boldsymbol{J}(\boldsymbol{x})^T\boldsymbol{J}(\boldsymbol{x})\Big)^{-1}\boldsymbol{J}(\boldsymbol{x})\mathbf{F}(\boldsymbol{x}) \approx -\nabla^2 f(\boldsymbol{x})^{-1}\nabla f(\boldsymbol{x})^T$$

is a descent direction, in fact

$$\nabla f(\boldsymbol{x})\boldsymbol{d} = -\nabla f(\boldsymbol{x})\Big(\boldsymbol{J}(\boldsymbol{x})^T\boldsymbol{J}(\boldsymbol{x})\Big)^{-1}\nabla f(\boldsymbol{x})^T < 0$$

when $\boldsymbol{J}(\boldsymbol{x})$ is full rank.

Algorithm (Gauss-Newton algorithm)

$\boldsymbol{x}$ *assigned;*

$\boldsymbol{f} \leftarrow \mathbf{F}(\boldsymbol{x})$;

$\boldsymbol{J} \leftarrow \nabla \mathbf{F}(\boldsymbol{x})$

while $\left\| \boldsymbol{J}^T \boldsymbol{f} \right\| > \epsilon$ **do**

— *compute search direction*

$\boldsymbol{d} \leftarrow -(\boldsymbol{J}^T \boldsymbol{J})^{-1} \boldsymbol{J}^T \boldsymbol{f}$;

Approximate $\arg\min_{\alpha>0} \mathsf{f}(\boldsymbol{x} + \alpha \boldsymbol{d})$ *by linsearch;*

— *perform step*

$\boldsymbol{x} \leftarrow \boldsymbol{x} + \alpha \boldsymbol{d}$;

end while

Outline

The Levenberg–Marquardt Method

Levenberg (1944) and later Marquardt (1963) suggested to use a damped Gauss-Newton method:

$$\boldsymbol{d} = -\Big(\boldsymbol{J}(\boldsymbol{x})^T\boldsymbol{J}(\boldsymbol{x}) + \mu\boldsymbol{I}\Big)^{-1}\nabla f(\boldsymbol{x})^T, \qquad \nabla f(\boldsymbol{x})^T = \boldsymbol{J}(\boldsymbol{x})\mathbf{F}(\boldsymbol{x})$$

1. for all $\mu \geq 0$ is a descent direction, in fact

$$\nabla f(\boldsymbol{x})\boldsymbol{d} = -\nabla f(\boldsymbol{x})\Big(\boldsymbol{J}(\boldsymbol{x})^T\boldsymbol{J}(\boldsymbol{x}) + \mu\boldsymbol{I}\Big)^{-1}\nabla f(\boldsymbol{x})^T < 0$$

2. for large μ we have $\boldsymbol{d} \approx -\frac{1}{\mu}\nabla f(\boldsymbol{x})^T$ the gradient direction.
3. for small μ we have $\boldsymbol{d} \approx -(\boldsymbol{J}(\boldsymbol{x})^T\boldsymbol{J}(\boldsymbol{x}))^{-1}\nabla f(\boldsymbol{x})^T$ the Gauss-Newton direction

1. The choice of parameter μ affect both size and direction of the step
2. Levenberg–Marquardt becomes a method without line-search.
3. As for Trust region each step (approximately) solve the minimization of the model problem

$$\min\ m(\boldsymbol{x}+\boldsymbol{s}) = \mathsf{f}(\boldsymbol{x}) + \nabla \mathsf{f}(\boldsymbol{x})\boldsymbol{s} + \frac{1}{2}\boldsymbol{s}^T \boldsymbol{H}(\boldsymbol{x})\boldsymbol{s}$$

where $\boldsymbol{H}(\boldsymbol{x}) = \boldsymbol{J}(\boldsymbol{x})^T \boldsymbol{J}(\boldsymbol{x}) + \mu \boldsymbol{I}$ is symmetric and positive definite (SPD).

4. $\boldsymbol{H}(\boldsymbol{x})$ is SPD and the minimum is

$$\boldsymbol{s} = -\boldsymbol{H}(\boldsymbol{x})^{-1}\boldsymbol{g}(\boldsymbol{x}), \qquad \boldsymbol{g}(\boldsymbol{x}) = \nabla \mathsf{f}(\boldsymbol{x})^T$$

Algorithm (Generic LM algorithm)

$\boldsymbol{x}, \mu$ *assigned*; $\eta_1 = 0.25$; $\eta_2 = 0.75$; $\gamma_1 = 2$; $\gamma_2 = 1/3$;
$\boldsymbol{f} \leftarrow \mathbf{F}(\boldsymbol{x})$; $\boldsymbol{J} \leftarrow \nabla\mathbf{F}(\boldsymbol{x})$;
while $\|\boldsymbol{f}\| > \epsilon$ **do**

$$\boldsymbol{s} \leftarrow \arg\min\; m(\boldsymbol{x}+\boldsymbol{s}) = \frac{1}{2}\|\boldsymbol{f}\|^2 + \boldsymbol{f}^T\boldsymbol{s} + \frac{1}{2}(\boldsymbol{J}^T\boldsymbol{J} + \mu\boldsymbol{I})\boldsymbol{s};$$

$$pred \leftarrow m(\boldsymbol{x}+\boldsymbol{s}) - m(\boldsymbol{x});$$

$$ared \leftarrow \frac{1}{2}\|\mathbf{F}(\boldsymbol{x}+\boldsymbol{s})\|^2 - \frac{1}{2}\|\boldsymbol{f}\|^2;$$

$$r \leftarrow (ared/pred);$$

  if $r < \eta_1$ **then**
    $\boldsymbol{x} \leftarrow \boldsymbol{x}$; $\mu \leftarrow \gamma_1\mu$; — *reject step, enlarge* μ
  else
    $\boldsymbol{x} \leftarrow \boldsymbol{x} + \boldsymbol{s}$; — *accept step*
    if $r > \eta_2$ **then**
      $\mu \leftarrow \gamma_2\mu$; — *reduce* μ
    end if
  end if
end while

Let r the ratio of expected and actual reduction of a step a faster strategy for the μ update is the following

Algorithm (Generic LM algorithm)

if $r > 0$ **then**

$$\mu \leftarrow \mu \max \left\{ \frac{1}{3}, 1 - (2r-1)^3 \right\}$$

$\nu \leftarrow 2$

else

$\mu \leftarrow \mu\,\nu;$

$\nu \leftarrow 2\,\nu;$

end if

H.B. Nielsen
Damping Parameter in Marquardt's Method
IMM, DTU. Report IMM-REP-1999-05, 1999.
http://www.imm.dtu.dk/~hbn/publ/TR9905.ps

Outline

The Dog-Leg step

As for the Thrust Region method we have 2 searching direction:
One is the Gauss-Newton direction (when $\mu = 0$)

$$\boldsymbol{d}_{GN} = -\left(\boldsymbol{J}(\boldsymbol{x})^T \boldsymbol{J}(\boldsymbol{x})\right)^{-1} \nabla f(\boldsymbol{x})^T, \qquad \nabla f(\boldsymbol{x})^T = \boldsymbol{J}(\boldsymbol{x})\mathbf{F}(\boldsymbol{x})$$

and the gradient direction (when $\mu = \infty$)

$$\boldsymbol{d}_{SD} = -\nabla f(\boldsymbol{x})^T = -\boldsymbol{J}(\boldsymbol{x})^T \mathbf{F}(\boldsymbol{x}),$$

to be finished!

References

- J. Stoer and R. Bulirsch
  Introduction to numerical analysis
  Springer-Verlag, Texts in Applied Mathematics, **12**, 2002.
- J. E. Dennis, Jr. and Robert B. Schnabel
  Numerical Methods for Unconstrained Optimization and Nonlinear Equations
  SIAM, Classics in Applied Mathematics, **16**, 1996.