TMA4180 Optimization: Least-Squares Problems

Elisabeth Köbis

NTNU, TMA4180, Optimering, Spring 2021

February 8th, 2021

Least-Squares Problems

Form of the objective function:

$$f(x) = \frac{1}{2} \sum_{j=1}^{m} r_j^2(x)$$

where each $r_j : \mathbb{R}^n \to \mathbb{R}$ is a real valued smooth function called *residual*. We assume that $m \geq n$.

Remark

- Least-squares problems arise in many areas of applications, and may in fact be the largest source of unconstrained optimization problems
- measure the discrepancy between the model and the observed behavior of the system.
- by minimizing f, we select values for the parameters that best match the model to the data.
- we show how to devise efficient, robust minimization algorithms by exploiting the special structure of the function f and its derivatives.

Least-squares problems are often easier to solve than general unconstrained minimization problems

Residual vector $r : \mathbb{R}^n \to \mathbb{R}^m$

$$r(x) := (r_1(x), \ldots, r_m(x))^T$$

Then $f = \frac{1}{2}||r(x)||_2^2$. The *Jacobian* ($m \times n$-matrix) is defined as

$$J(x) := \left[\frac{\partial r_j}{\partial x_i}\right]_{j=1,\ldots,m,i=1,\ldots,n} = \begin{pmatrix} \nabla r_1(x)^T \\ \vdots \\ \nabla r_m(x)^T \end{pmatrix}.$$

$$\nabla f(x) = \sum_{j=1}^{m} r_j(x) \nabla r_j(x) = J(x)^T r(x),$$

$$\underbrace{\nabla^2 f(x)}_{\text{“for free”}} = \sum_{j=1}^{m} \nabla r_j(x) \nabla r_j(x)^T + \sum_{j=1}^{m} r_j(x) \nabla^2 r_j(x)^T$$

$$= J(x)^T J(x) + \sum_{j=1}^{m} r_j(x) \nabla^2 r_j(x). \qquad (1)$$

Example

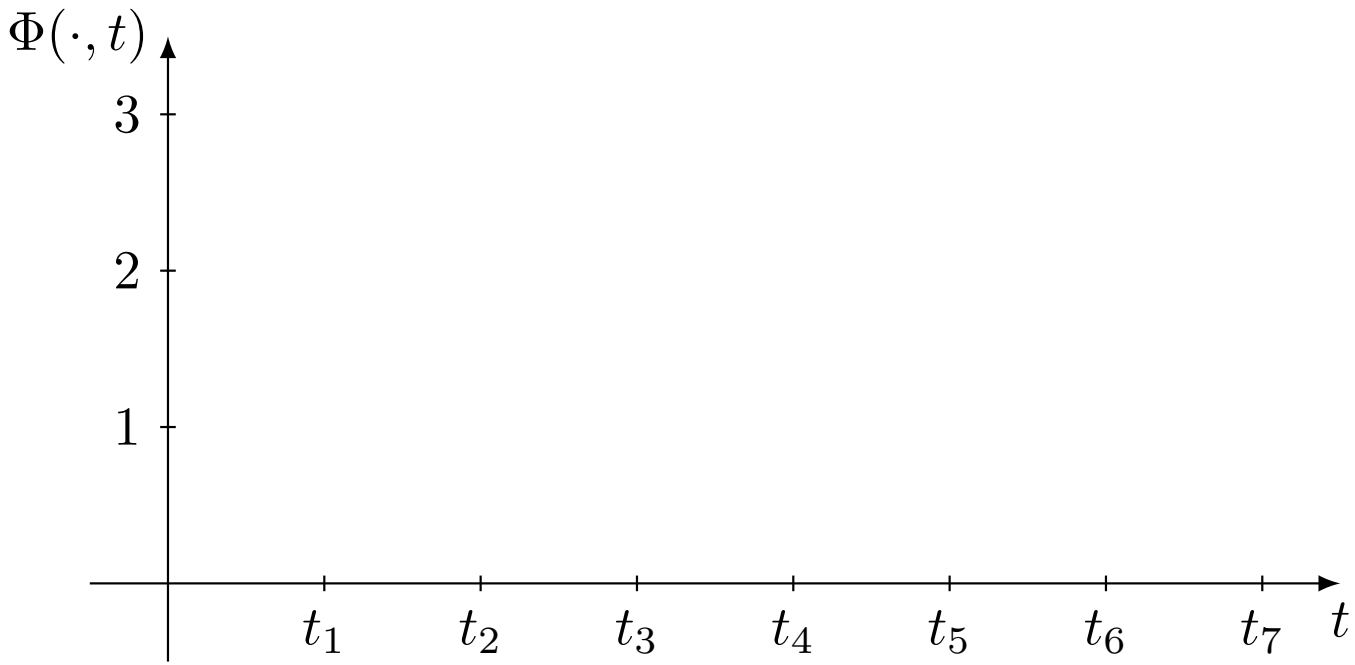
Study the effect of a certain medication on a patient: We draw blood samples at certain times after the patient takes a dose, and measure the concentration of the medication in each sample, tabulating the time t_j and concentration (observation) y_j for each sample. Based on our previous experience in such experiments, we find that the following function $\Phi(x;t)$ provides a good prediction of the concentration at time t, for appropriate values of the five-dimensional parameter vector $x = (x_1, x_2, x_3, x_4, x_5)$:

$$\Phi(x;t) = x_1 + tx_2 + t^2x_3 + x_4e^{-x_5t}$$

We choose the *parameter vector* x so that this model best agrees with our observation, in some sense. A good way to measure the difference between the predicted model values and the observations y_j is the following least-squares function:

$$\frac{1}{2}\sum_{j=1}^{m}(\underbrace{\Phi(x;t_j) - y_j}_{=r_j(x)})^2$$

Example



Having obtained x^*, we use $\Phi(x^*; t)$ to estimate the concentration of medication remaining in the patient's bloodstream at any time t.

Example

Fixed-regressor model

- assumes that the times t_j at which the blood samples are drawn are known to high accuracy, while the observations y_j may contain more or less random errors due to limitations of the equipment

Remark

In general data-fitting problems of the type just described, the ordinate t in the model $\Phi(x;t)$ could be a vector instead of a scalar. (In the example above, for instance, t could have two dimensions, with the first dimension representing the time since the drug was administered and the second dimension representing the weight of the patient. We could then use observations for an entire population of patients, not just a single patient, to obtain the "best" parameters for this model.)

Other ways of measuring the discrepancy between the model and the observations

$$f(x) := ||r(x)||_1 = \sum_{j=1}^{m} |\Phi(x; t_j) - y_j|$$
$$f(x) := ||r(x)||_\infty = \max_{j=1,\ldots,m} |\Phi(x; t_j) - y_j|$$

Algorithms For Nonlinear Least-Squares Problems

The Gauss-Newton Method

The standard Newton equations $\nabla^2 f(x_k)p = -\nabla f(x_k)$ are replaced by the following system to obtain the search direction p_k^{GN}:

$$J_k^T J_k p_k^{GN} = -J_k^T r_k.$$

Advantages

Using the approximation $\nabla^2 f(x_k) \approx J_k^T J_k$ saves the trouble of computing the individual residual Hessians $\nabla^2 r_j$ from (1). Often, the first term in (1) dominates the second one, such that $J_k^T J_k$ is a good approximation of $\nabla^2 f(x_k)$. In practice, many least-squares problems have small residuals at the solution, leading to rapid local convergence of Gauss-Newton.

p_k^{GN} is a descent direction:

$$(p_k^{GN})^T \nabla f_k = (p_k^{GN})^T J_k^T r_k = -(p_k^{GN})^T J_k^T J_k p_k^{GN} = -||J_k p_k^{GN}||_2^2 \leq 0.$$

The inequality is strict unless $J_k p_k^{GN} = 0$. If this is the case and J_k has full rank, then $J_k^T 0 = -J_k^T r_k = -\nabla f_k = 0$, that is, x_k is a stationary point.

Convergence of the Gauss-Newton Method

Theorem

Suppose each residual function r_j is Lipschitz continuously differentiable in a neighborhood $\mathcal{N}$ of the level set $\mathcal{L} := \{x | f(x) \leq f(x_0)\}$ which is assumed to be bounded, and that the Jacobians $J(x)$ satisfy the uniform full-rank condition $||J(x)z|| \geq \gamma||z||$ for some constant $\gamma > 0$ on $\mathcal{N}$. Then if the iterates x_k are generated by the Gauss-Newton method with step lengths α_k that satisfy the Wolfe conditions, we have

$$\lim_{k \to \infty} J_k^T r_k = 0.$$