Nonlinear Optimization

Least Squares Problems — The Gauss-Newton method

Niclas Börlin

Department of Computing Science

Umeå University
niclas.borlin@cs.umu.se

November 22, 2007

Problem formulation

- A nonlinear least-squares problem is an unconstrained optimization problem of the form

$$\min_{x \in \Re^n} f(x) = \frac{1}{2} \sum_{i=1}^{m} r_i(x)^2,$$

  where n is the number of variables.
- The objective function $f(x)$ is defined by m auxiliary *residual* functions $\{r_i(x)\}$. We will assume that $m \geq n$.
- The problem is called least-squares since we are minimizing the sum of squares of the residual functions.

Nonlinear least-squares parameter estimation

- A large class of optimization problems are the *non-linear least squares parameter estimation problems*.
- In a parameter estimation problem, the functions $r_i(x)$ represent the difference (residual) between a model function and a measured value. Study e.g. the data set

$$\begin{array}{rcccccc} t_i & : & 1 & 2 & 4 & 5 & 8 \\ y_i & : & 3 & 4 & 6 & 11 & 20 \end{array}$$

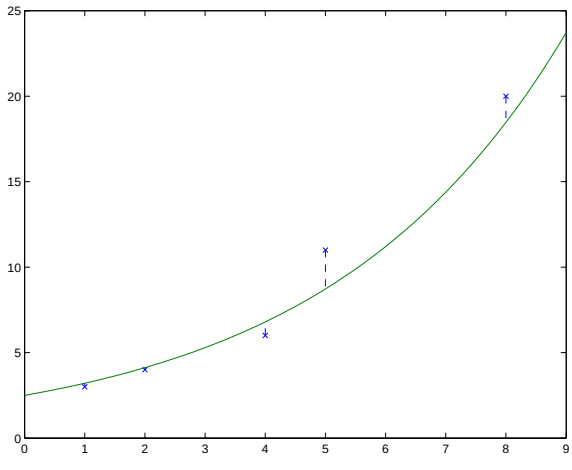
where t_i is the time in years and y_i is the size of antelope population (in hundreds).

- If we assume that the development of the population is exponential, the model function might be

$$g(t) = x_1 e^{x_2 t}$$

and the residuals

$$r_i(x) = g(t_i) - y_i = x_1 e^{x_2 t_i} - y_i.$$



- In standard least squares problems, the *vertical distance* (squared) between observations and a model function are minimized.

Geometric interpretation

- We will write the optimization problem as

$$\min_x f(x),$$

where

$$f(x) = \frac{1}{2}\sum_{i=1}^{m} r_i(x)^2 \equiv \frac{1}{2} r(x)^T r(x) \equiv \frac{1}{2}\|r(x)\|^2,$$

and r is a vector-valued function

$$r(x) = [r_1(x)\ r_2(x)\ \ldots\ r_m(x)]^T.$$

- For each value of x, the residual function value $r(x)$ may be interpreted as a point in "observation space" $\Re^m$.
- The residual function describes a (usually n-dimensional) surface in $\Re^m$.

- For the antelope data and model

$$f(x) = \frac{1}{2}\sum_{i=1}^{5}(x_1 e^{x_2 t_i} - y_i)^2 = \frac{1}{2} r(x)^T r(x),$$

$$r(x) = \begin{bmatrix} x_1 e^{x_2 t_1} - y_1 \\ x_1 e^{x_2 t_2} - y_2 \\ x_1 e^{x_2 t_3} - y_3 \\ x_1 e^{x_2 t_4} - y_4 \\ x_1 e^{x_2 t_5} - y_5 \end{bmatrix} = \begin{bmatrix} x_1 e^{1x_2} - 3 \\ x_1 e^{2x_2} - 4 \\ x_1 e^{4x_2} - 6 \\ x_1 e^{5x_2} - 11 \\ x_1 e^{8x_2} - 20 \end{bmatrix},$$

- Observe that

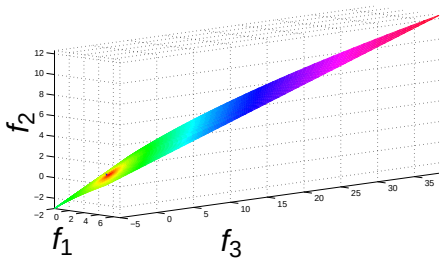
$$\min_x \frac{1}{2}\|r(x)\|^2$$

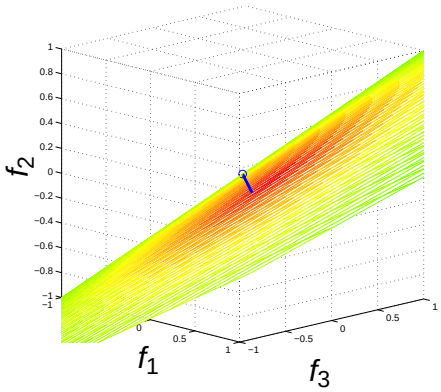
may be interpreted as

$$\min_x \frac{1}{2}\|r(x) - 0\|^2.$$

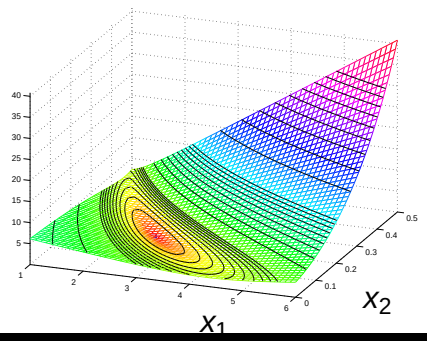
- Thus, a least squares problem may be interpreted as trying to find the point x^* in parameter space $\Re^n$ that corresponds to the point $r(x^*)$ in observation space $\Re^m$ that is *closest to the origin*.

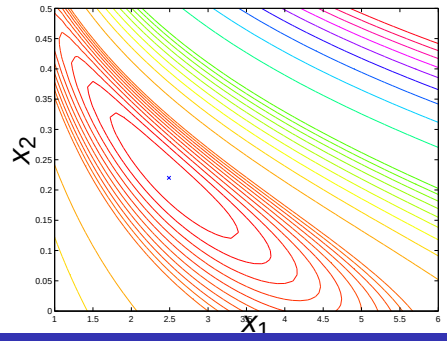
$r(x) = [r_1(x)\ r_2(x)\ r_3(x)]^T$ surface in $\Re^m$

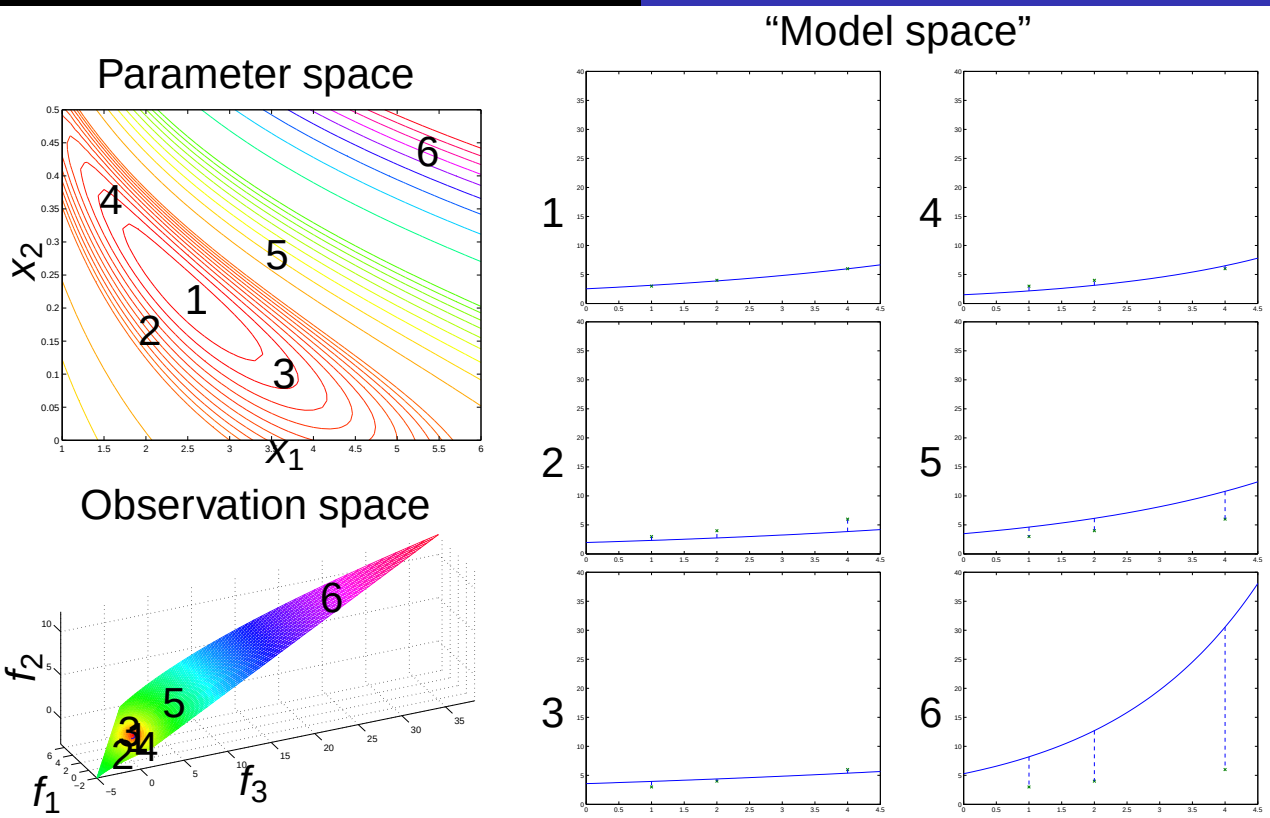




$f(x) = \frac{1}{2}\|r(x)\|^2$ as a function of $x = [x_1\ x_2]^T$ in $\Re^n$







Gradient and Hessian structure

- The gradient $\nabla f(x)$ may be derived from the chain rule

$$\nabla f(x) = \nabla r(x) r(x) = J(x)^T r(x),$$

where $J(x) = \nabla r(x)^T$ is the *Jacobian* of $r(x)$, i.e.

$$J(x) = \begin{bmatrix} \frac{\partial r_1(x)}{\partial x_1} & \cdots & \frac{\partial r_1(x)}{\partial x_n} \\ \frac{\partial r_2(x)}{\partial x_1} & \cdots & \frac{\partial r_2(x)}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial r_m(x)}{\partial x_1} & \cdots & \frac{\partial r_m(x)}{\partial x_n} \end{bmatrix}.$$

Gradient and Hessian structure

- Using the chain rule again, the Hessian is

$$\begin{aligned}\nabla^2 f(x) &= \nabla r(x)\nabla r(x)^T + \sum_{i=1}^m r_i(x)\nabla^2 r_i(x), \\ &= J(x)^T J(x) + Q(x).\end{aligned}$$

- Thus, the Hessian of a least-squares objective function is a sum of two terms; $J(x)^T J(x)$ with only first-order derivatives, and $Q(x)$ with second-order derivatives.

Gradient, Jacobian, and Hessian

For the antelope data and model

$$f(x) = \frac{1}{2}\sum_{i=1}^{5}(x_1 e^{x_2 t_i} - y_i)^2 = \frac{1}{2} r(x)^T r(x), \quad r(x) = \begin{bmatrix} x_1 e^{x_2 t_1} - y_1 \\ x_1 e^{x_2 t_2} - y_2 \\ x_1 e^{x_2 t_3} - y_3 \\ x_1 e^{x_2 t_4} - y_4 \\ x_1 e^{x_2 t_5} - y_5 \end{bmatrix} = \begin{bmatrix} x_1 e^{1x_2} - 3 \\ x_1 e^{2x_2} - 4 \\ x_1 e^{4x_2} - 6 \\ x_1 e^{5x_2} - 11 \\ x_1 e^{8x_2} - 20 \end{bmatrix},$$

$$\nabla f(x) = J(x)^T r(x), \quad J(x) = \begin{bmatrix} e^{x_2 t_1} & t_1 x_1 e^{x_2 t_1} \\ e^{x_2 t_2} & t_2 x_1 e^{x_2 t_2} \\ e^{x_2 t_3} & t_3 x_1 e^{x_2 t_3} \\ e^{x_2 t_4} & t_4 x_1 e^{x_2 t_4} \\ e^{x_2 t_5} & t_5 x_1 e^{x_2 t_5} \end{bmatrix} = \begin{bmatrix} e^{1x_2} & 1x_1 e^{1x_2} \\ e^{2x_2} & 2x_1 e^{2x_2} \\ e^{4x_2} & 4x_1 e^{4x_2} \\ e^{5x_2} & 5x_1 e^{5x_2} \\ e^{8x_2} & 8x_1 e^{8x_2} \end{bmatrix},$$

$$\nabla^2 f(x) = J(x)^T J(x) + Q(x), \quad Q(x) = \sum_{i=1}^{5}(x_1 e^{x_2 t_i} - y_i) \begin{bmatrix} 0 & t_i e^{x_2 t_i} \\ t_i e^{x_2 t_i} & x_1 t_i^2 e^{x_2 t_i} \end{bmatrix}$$

The Gauss-Newton method; the Newton formulation

- The Hessian is a sum of two components

$$\begin{aligned} \nabla^2 f(x) &= \nabla r(x)\nabla r(x)^T + \sum_{i=1}^m r_i(x)\nabla^2 r_i(x) \\ &= J(x)^T J(x) + Q(x). \end{aligned}$$

- If the problem has a zero residual, i.e. $r_i(x^*) = 0$, the term $Q(x)$ will be small close to the solution.
- A method that uses the approximation $Q(x) = 0$ is called the *Gauss-Newton* method and determines the search direction as the solution of the Newton equation

$$\nabla^2 f(x) p^N = -\nabla f(x)$$

with the Hessian approximated by $J(x)^T J(x)$, i.e.

$$J(x)^T J(x) p^{GN} = -J(x)^T r(x).$$

- If we assume that $J(x)$ has full rank, the Hessian approximation

$$J(x)^T J(x)$$

is positive definite and the Gauss-Newton search direction p^{GN} is a descent direction.

- Otherwise, $J(x)^T J(x)$ is non-invertible and the equation

$$J(x)^T J(x) p^{GN} = -J(x)^T r(x)$$

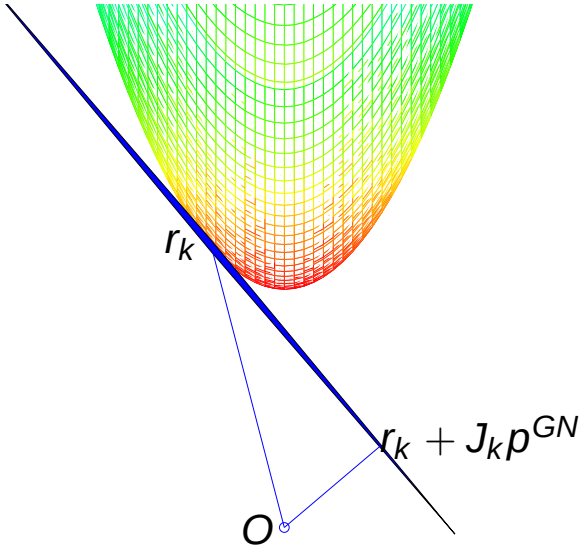
does not have a unique solution. In this case, the problem is said to be *under-determined* or *over-parameterized*.

The linear least squares formulation

- Assume we approximate the residual function $r(x)$ with a *linear* Taylor function, i.e. a plane
$$r(x_k + p) \approx r_k + J_k p.$$
- The minimizer on the plane is found by solving the linear least squares problem
$$\min_p \frac{1}{2}\|J_k p + r_k\|^2 = \min_p \frac{1}{2}\|J_k p - (-r_k)\|^2.$$
- The solution is given by the normal equations
$$J_k^T J_k p = -J_k^T r_k$$
or
$$p = (J_k^T J_k)^{-1} J_k^T (-r_k).$$
- Thus, the minimizer on the plane corresponds to the Gauss-Newton search direction!

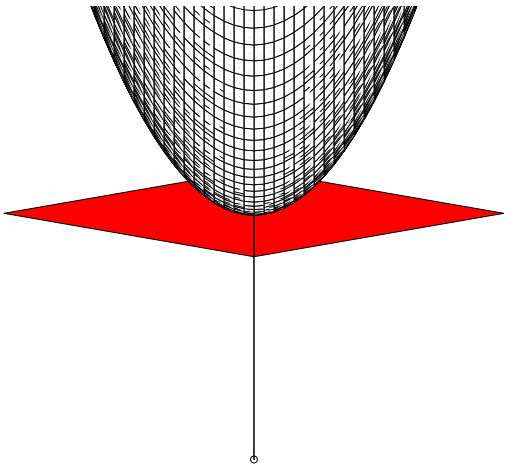
Geometrical interpretation of the search direction

- The linear approximation corresponds to a tangent plane to the surface $r(x)$ at $r_k = r(x_k)$.
- The point on the tangent plane closest to the origin is given by the projection of $-r_k$ onto the range space of J_k, since
$$J_k p^{GN} = \underbrace{J_k (J_k^T J_k)^{-1} J_k^T}_{P_{\mathcal{R}(J_k)}} (-r).$$



Geometric interpretation of the first order condition

- The first order condition $\nabla f(x^*) = 0$ corresponds to when $J(x^*)^T r(x^*) = 0$, i.e. $r(x^*)$ is orthogonal to the tangent plane spanned by the columns of $J(x^*)$.
- A special case is when $r(x^*) = 0 \Rightarrow f(x^*) = 0$.
- In this case the problem is said to have *zero residual* and the surface $r(x)$ intersects the origin.



Convergence for the Gauss-Newton method

- If $r(x^*) = 0$, then the approximation $Q(x) \approx 0$ is good and the Gauss-Newton method will behave like the Newton method close to the solution, i.e. converge quadratically if $J(x^*)$ has full rank.
- The advantage over the Newton method is that we do not need to calculate the second-order derivatives $\nabla^2 r_i(x)$.
- However, if any residual component $r_i(x^*)$ and/or the corresponding curvature $\nabla^2 r_i(x)$ is large, the approximation $Q(x) \approx 0$ will be poor, and the Gauss-Newton method will converge slower than the Newton method.
- For such problems, the Gauss-Newton method may not even be locally convergent, i.e. without a global strategy such as the line search, it wouldn't converge no matter how close to the solution we start.

Pertubation sensitivity

- If $J(x^*) = USV^T$ is the singular value decomposition of $J(x^*)$ with

$$U^T U = I_m, V^T V = I_n, S = \begin{bmatrix} S_0 \\ 0 \end{bmatrix}, S_0 = \begin{bmatrix} s_1 & & \\ & \ddots & \\ & & s_n \end{bmatrix}, s_1 \geq s_2 \geq \ldots \geq s_n \geq 0,$$

  the first order approximation $r(x^* + p) \approx r(x^*) + J(x^*)p$ becomes

$$r(x^*) + USV^T p = r(x^*) + u_1 s_1 v_1^T p + \ldots + u_n s_n v_n^T p.$$

- For small regions around x^*, the residual function $r(x^*)$ will change the most in the direction of v_1, and the change will be proportional to $s_1 u_1$, i.e. the residual values are most sensitive to changes in the v_1 direction.
- Similarly, the residual function will change the least in the direction of v_n, with the change proportional to $s_n u_n$. In the extreme case of $s_n = 0$, the residual will be constant in the direction of v_n and the solution x^* will not be unique.

- Since the search direction is calculated as

$$\begin{aligned} p &= (J^T J)^{-1}(J^T(-r)) = (VS^T U^T USV^T)^{-1} VS^T U^T(-r) \\ &= (VS_0^2 V^T)^{-1} VS^T U^T(-r) = VS_0^{-2} V^T VS^T U^T(-r) \\ &= V \begin{bmatrix} S_0^{-1} \\ 0 \end{bmatrix} U^T(-r) = \frac{v_1 u_1^T(-r)}{s_1} + \ldots + \frac{v_n u_n^T(-r)}{s_n}, \end{aligned}$$

  the opposite is true for the sensitivity of x^* as a function of r.
- The solution x^* is the most sensitive to pertubations of r in the direction of u_n and the change in x^* will be proportional to $\frac{1}{s_n} v_n$.

- For the antelope problem,

$$J(x^*) = \begin{bmatrix} 1.25 & 3.11 \\ 1.56 & 7.75 \\ 2.42 & 24.1 \end{bmatrix}, U = \begin{bmatrix} 0.13 & 0.76 & -0.63 \\ 0.31 & 0.58 & 0.76 \\ 0.94 & -0.29 & -0.16 \end{bmatrix},$$
$$V = \begin{bmatrix} 0.11 & -0.99 \\ 0.99 & 0.11 \end{bmatrix}, S_0 = \begin{bmatrix} 25 & \\ & 1.2 \end{bmatrix}.$$

- The solution is most sensitive to pertubation of the observations in the direction of $\begin{bmatrix} 0.76 & 0.58 & -0.29 \end{bmatrix}^T$, which will perturb the solution in the $\begin{bmatrix} -0.99 & 0.11 \end{bmatrix}^T$ direction.
- x_1 is the most sensitive variable, and it the most sensitive to the y_3 observation. Similarly, x_2 is the least sensitive variable, and the y_3 observation has the least effect on it.

Statistical interpretation

- If the residuals are interpreted statistically, i.e. we have a model

$$y_i = x_1 e^{x_2 t_i} + \varepsilon_i$$

and the errors ε_i are assumed to be independent and normally distributed $N(0, \sigma^2)$, our least squares estimation of the parameters will be the *maximum likelihood* estimators given our measurement y_i.

Variance of estimated parameters

- The variance for the estimated parameters are calculated from the *variance-covariance matrix*

$$D = \sigma^2 (\nabla^2 f(x^*))^{-1},$$

where each diagonal element d_{ii} correspond to the variance of the parameter x_i, and the off-diagonal element d_{ij} correspond to the covariance between parameters x_i and x_j.

- If σ^2 is unknown, it may be estimated by

$$\hat{\sigma}^2 = \frac{r(x^*)^T r(x^*)}{m - n},$$

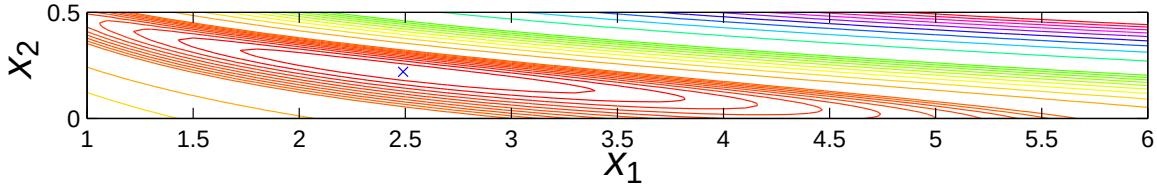
where m is the number of observations, and n is the number of parameters.

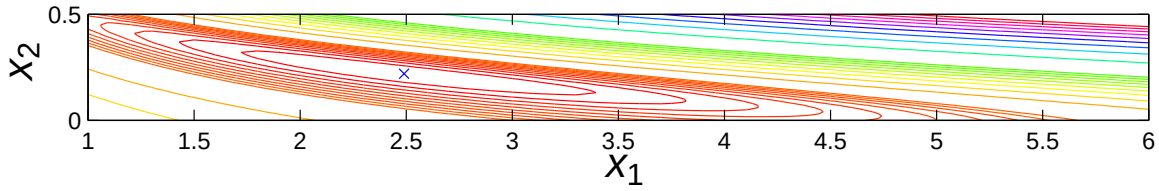
- A high variance means a high degree of uncertainty about a parameter. In this context, the inverse matrix

$$K = D^{-1} = \frac{1}{\sigma^2} \nabla^2 f(x^*),$$

is sometimes called the *information matrix*, since the higher the diagonal value k_{ii}, the more information we have about the parameter x_i.

- Since the information matrix is proportional to the hessian $H(x^*) = \nabla^2 f(x^*)$, strong curvature corresponds to high information, i.e. good localization of the parameter.





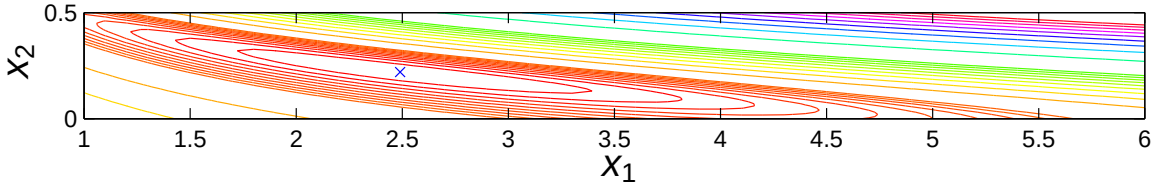
$$x^* = \begin{bmatrix} 2.49 \\ 0.22 \end{bmatrix}, r(x^*) = \begin{bmatrix} 0.11 \\ -0.13 \\ 0.03 \end{bmatrix}, J(x^*) = \begin{bmatrix} 1.25 & 3.11 \\ 1.56 & 7.75 \\ 2.42 & 24.1 \end{bmatrix}, J(x^*)^T J(x^*) = \begin{bmatrix} 9.84 & 74.3 \\ 74.3 & 651 \end{bmatrix},$$

$$Q(x^*) = \begin{bmatrix} 0 & 10^{-7} \\ 10^{-7} & 0.98 \end{bmatrix}, H(x^*) = J(x^*)^T J(x^*) + Q(x^*) = \begin{bmatrix} 9.84 & 74.3 \\ 74.3 & 652 \end{bmatrix},$$

$$H(x^*)^{-1} = \begin{bmatrix} 0.73 & -0.083 \\ -0.083 & 0.011 \end{bmatrix} = V \Lambda V^T, V = \begin{bmatrix} 0.99 & 0.11 \\ -0.11 & 0.99 \end{bmatrix}, \Lambda = \begin{bmatrix} 0.74 & 0 \\ 0 & 0.0015 \end{bmatrix},$$

- Thus, $\hat{\sigma} = \sqrt{r(x^*)^T r(x^*)/(3-2)} = 0.17$ (hecto-antelopes) and the standard deviation of x_1 is $\sqrt{0.73}\sigma = 0.14$ (hecto-antelopes) and of x_2 is $\sqrt{0.011}\sigma = 0.017$ (hecto-antelopes/year). With these units, the maximum uncertainty is in the direction of $0.99x_1 - 0.11x_2$.
- Note that the interpretation of the standard deviations is context-dependent, since it depends on e.g. the measurement units of each parameter.

- The approximations of the parameters and the covariances makes it possible to derive confidence limits, do hypothesis testing, etc.
- For linear problems, the covariance estimations are exact. For non-linear problems, the covariances are still exact, but the confidence limits are not, since the confidence regions are not ellipses.



- Furthermore, if the hessian is approximated by

$$\nabla^2 f(x^*) \approx J(x^*)^T J(x^*),$$

the covariances will only be first order approximations of the true covariances.

Weighted least squares

- If the observations errors are dependent and/or with different variances, *weighted least squares* should be used, i.e. the problem

$$\min_x r(x)^T W r(x),$$

should be solved.

- If the matrix Σ with elements σ_{ij}^2 contain the covariances between observations i and j, the optimal choice of W is

$$W = \Sigma^{-1},$$

and the solution of the weighted least squares problem is again the maximum likelyhood solution.

- The distance measure $r(x)^T \Sigma^{-1} r(x)$ is sometimes called the *Mahalanobis distance*.
- If the observations are independent, Σ and W will be diagonal matrices, and $w_i = 1/\sigma_i^2$. Thus the solution will rely more on "good" observations, since residuals with a corresponding small observation error will be weighted more heavily in the objective function.

Non-linear least squares problems
The Gauss-Newton method
Pertubation sensitivity
Statistical interpretation
Orthogonal regression
The error model
Variance of estimated parameters
Weighted least squares

- If we want to solve a weighted least squares problem, there are two equivalent solutions: Change the algorithm or change the residual and Jacobian function.
- A modified algorithm would solve the following equation

$$J^T W J p = -J^T W r.$$

- A modified residual/Jacobian would be

$$r_s(x) = R r(x), J_s(x) = R J(x),$$

where $R^T R = W$ is the Cholesky factorization of W. Such a factor R will always exist if W is positive semidefinite.

Ortogonal regression

- When we solve the problem

$$\min_x \frac{1}{2}\sum_{i=1}^{m} r_i(x)^2 = \frac{1}{2}\min_x r(x)^T r(x)$$

  where $r_i(x) = g(t_i) - y_i$ is the difference between our model and our measured values, we minimize the square of the *vertical* distance.
- In other contexts, e.g. if we can assume that we have errors also in the independent variable t_i, it may be appropriate to minimize the *orthogonal* distance between the model and the measurements instead.
- This may be formulated as that we solve the problem

$$\min_{x,\delta} f(x) = \sum_{i=1}^{m} r_i(x; t_i + \delta_i)^2 + \|\delta\|^2,$$

  where δ_i is the error in t_i and $r_i(x; t_i + \delta_i) = g_i(t_i + \delta_i) - y_i$.
- Problem minimizing the orthogonal distance between model and measurements are sometimes referred to as *orthogonal regression problems*.

- By reformulating the objective function, we may use algorithms for "conventional" non-linear least squares to solve orthogonal regression problems.
- For our example

$$y = g(t) = x_1 e^{x_2 t}$$

  we may introduce one point $(s_i, g(s_i))$ on the curve for each measurement (t_i, y_i).
- Defining the component function $r_i(x)$

$$r_i(x) = \begin{bmatrix} g(t_i) - y_i \\ s_i - t_i \end{bmatrix}, \text{ and } r(x) = \begin{bmatrix} r_1(x) \\ \vdots \\ r_m(x) \end{bmatrix},$$

  the least squares problem takes the following, standard, form:

$$\min_x r(x)^T r(x).$$