Least Squares Problems

The Gauss-Newton method

Niclas Börlin

5DA001 Non-linear Optimization

Problem formulation

- A nonlinear least-squares problem is an optimization problem on 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

t_i	:	1	2	4	5	8
y_i	:	3	4	6	11	20

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

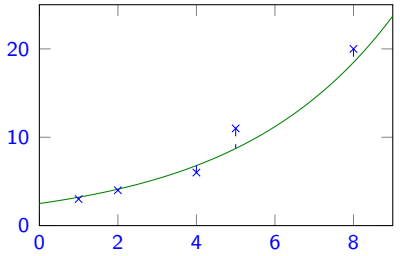
Nonlinear least-squares parameter estimation

- 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 $\Re^m$.

Geometric interpretation

- For the antelope data and model

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

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

Geometric interpretation

- 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.

Geometric interpretation

- Yet another alternative is to separate the observation vector

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

from the model function

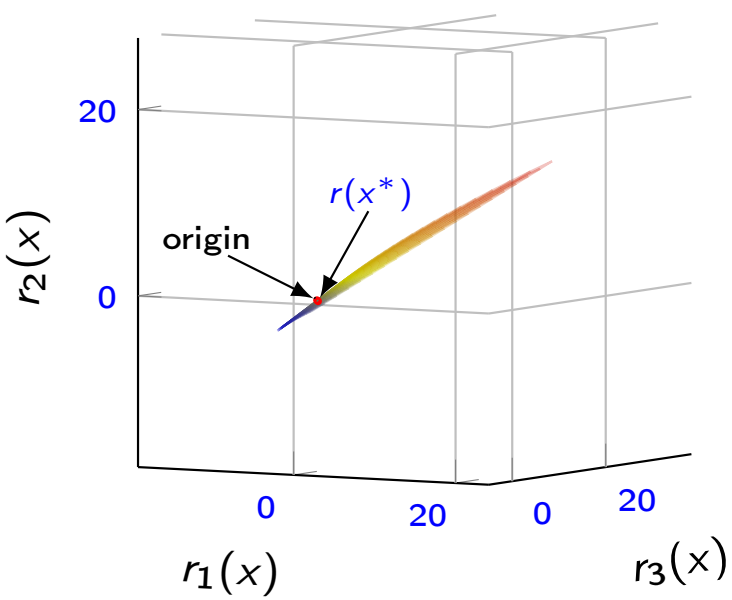
$$h(x) = \begin{bmatrix} g(t_1) \\ \vdots \\ g(t_m) \end{bmatrix} = \begin{bmatrix} x_1 e^{x_2 t_1} \\ \vdots \\ x_1 e^{x_2 t_m} \end{bmatrix}.$$

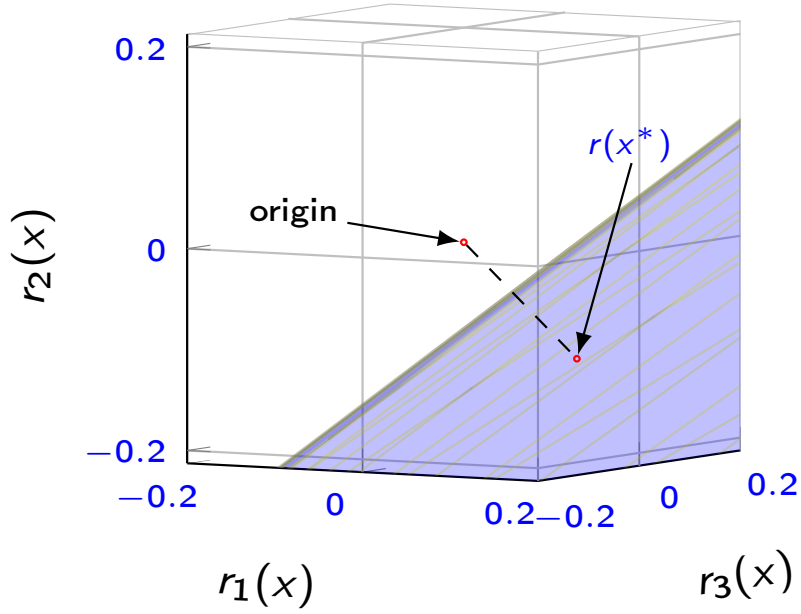
- Then,

$$\min_x \frac{1}{2}\|r(x)\|^2 = \min_x \frac{1}{2}\|h(x) - y\|^2$$

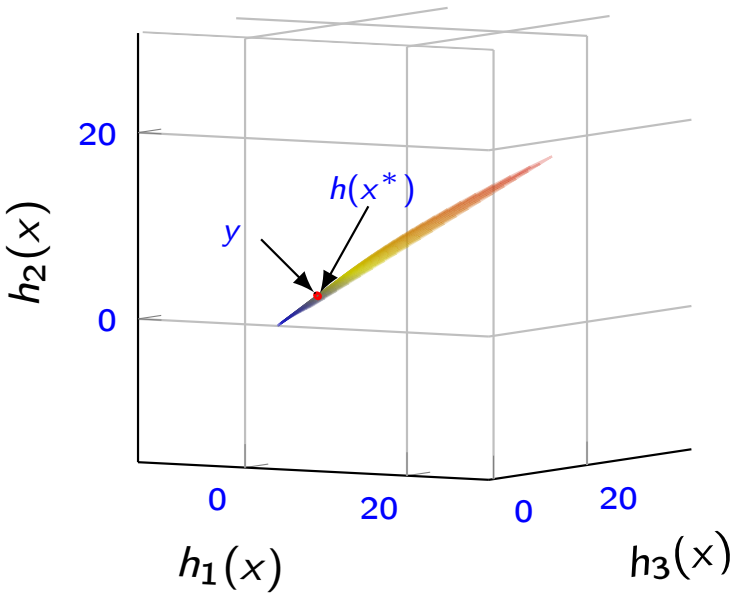
- With this formulation we are trying to find the point x^* in parameter space $\Re^n$ that corresponds to the point $h(x^*)$ on the model surface in observation space $\Re^m$ that is closest to the observations y.

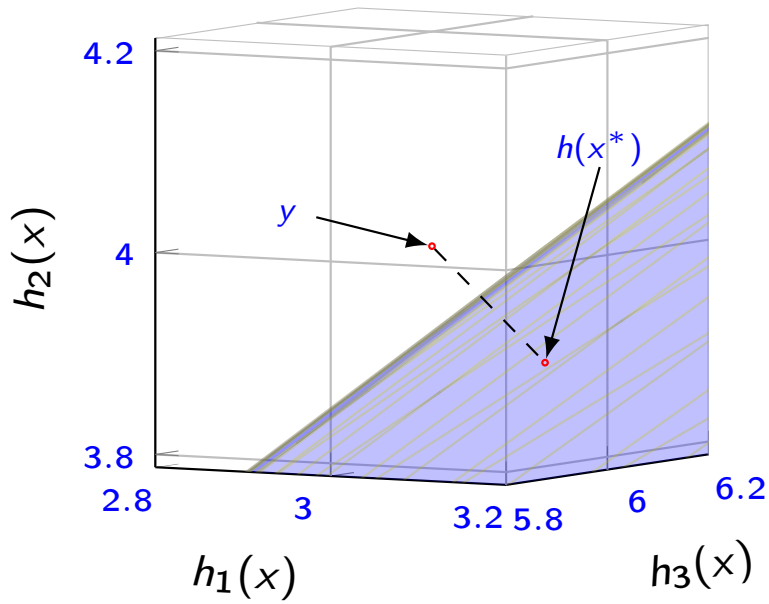
$r(x) = [r_1(x)\ r_2(x)\ r_3(x)]^T$ residual surface in $\Re^m$



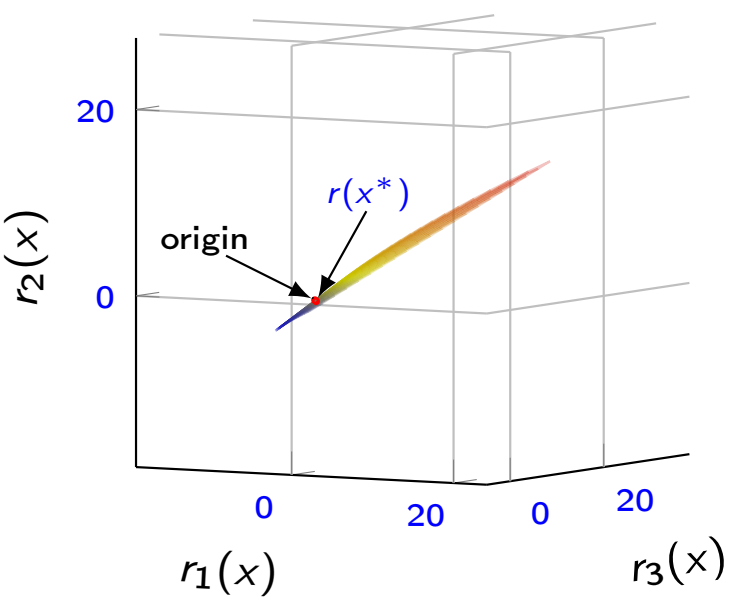


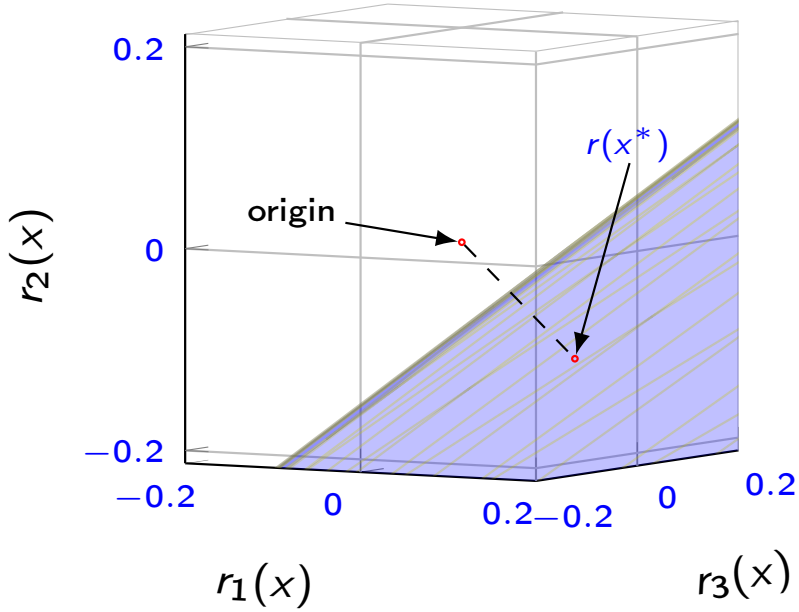
$h(x) = [h_1(x)\ h_2(x)\ h_3(x)]^T$ observation surface in $\Re^m$



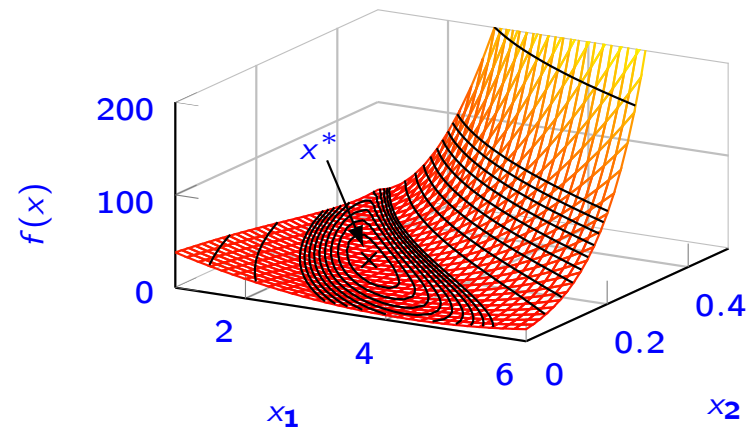


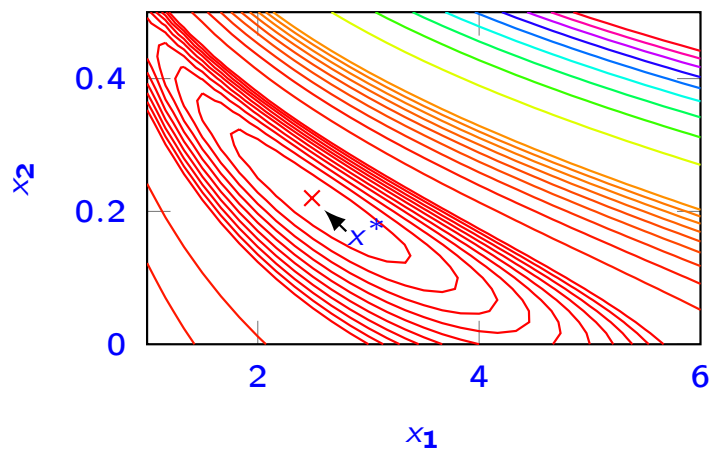
$r(x) = [r_1(x)\ r_2(x)\ r_3(x)]^T$ residual surface in $\Re^m$

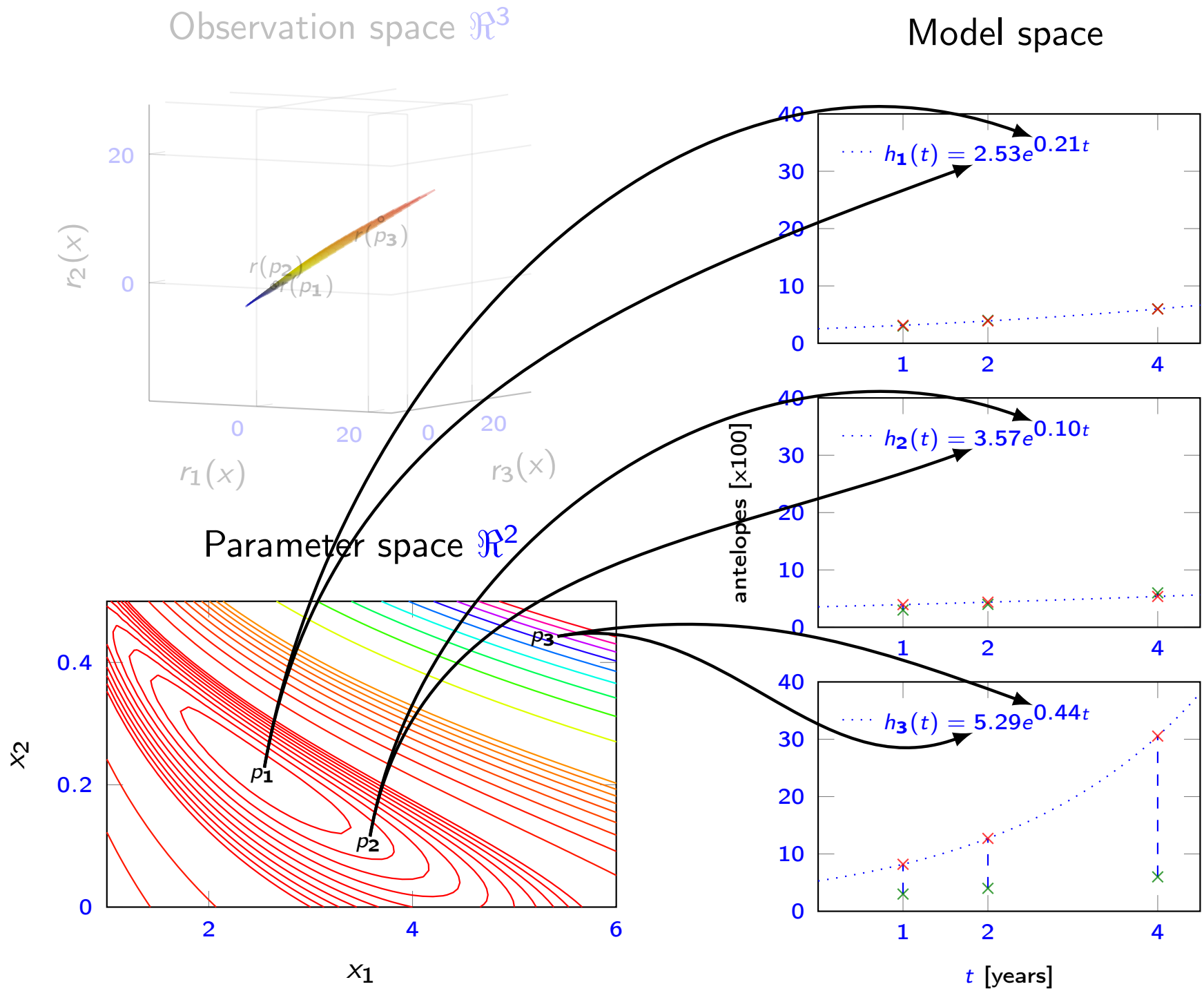


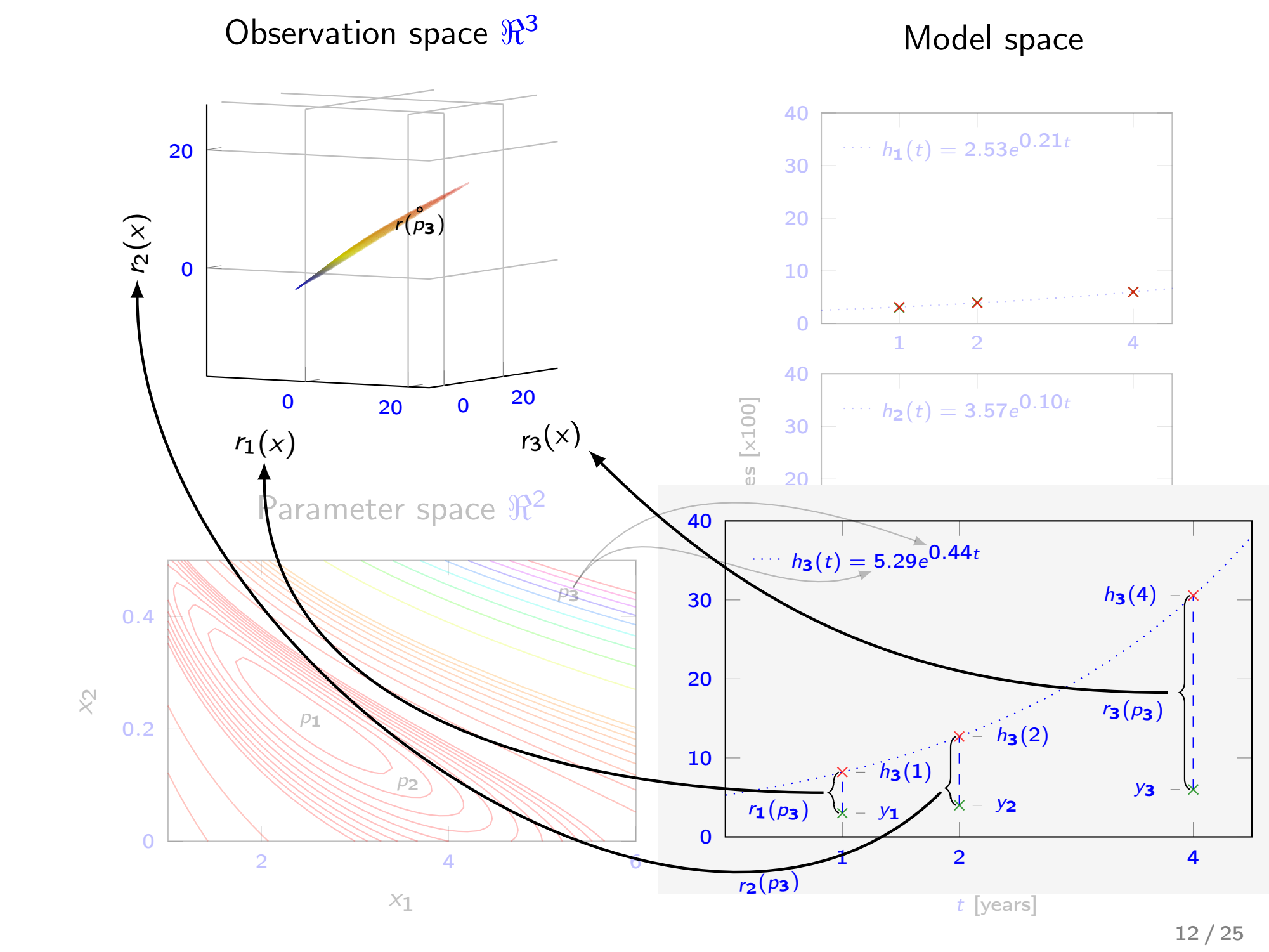


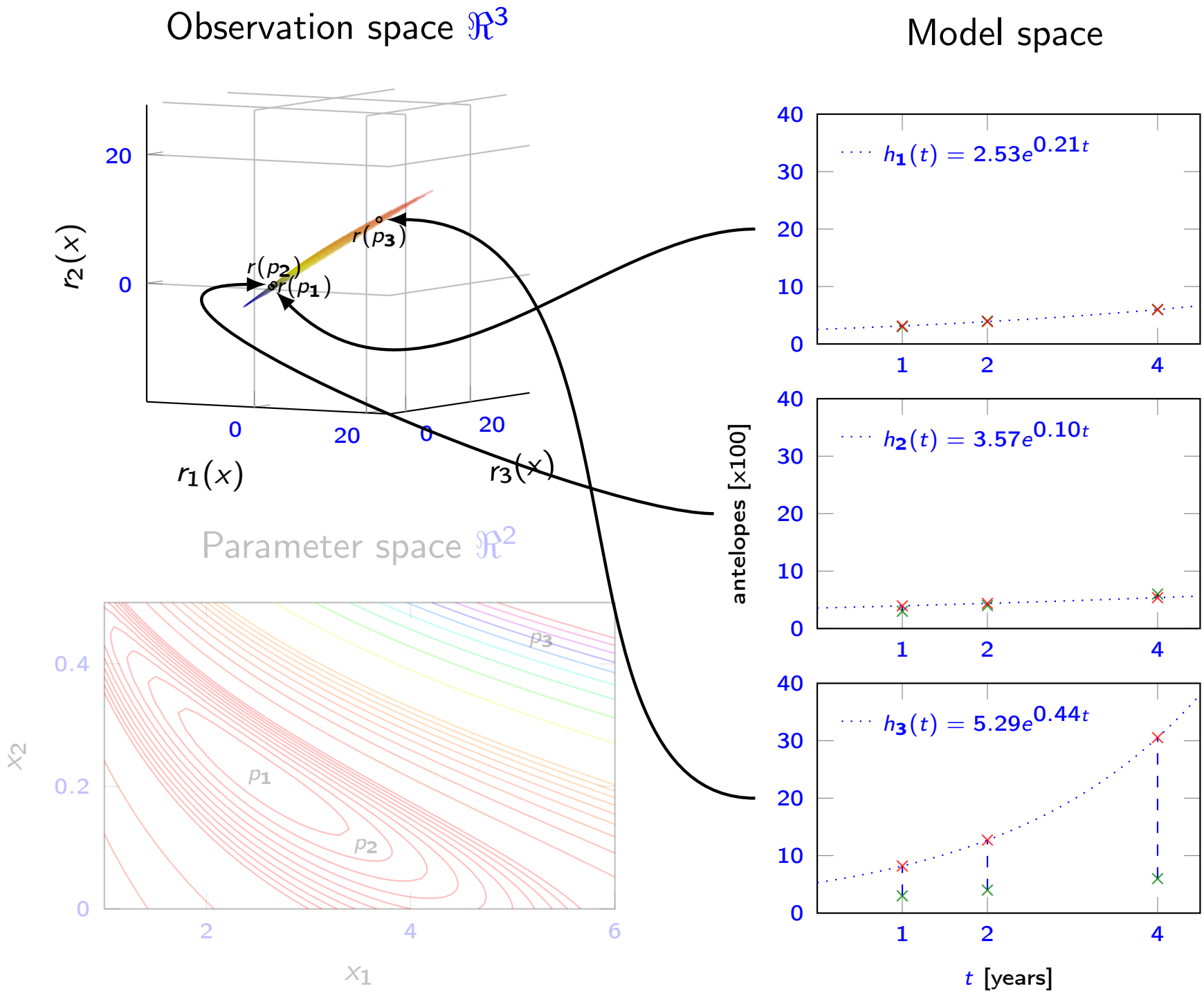
$f(x) = \frac{1}{2}\|r(x)\|^2$ as a function of $x = [x_1\ x_2]^T$ in $\Re^n$






Observation space $\Re^3$
Model space
$r_1(x)$
$r_2(x)$
$r_3(x)$
$r(p_1)$
$r(p_2)$
$r(p_3)$
Parameter space $\Re^2$
p_1
p_2
p_3
x_1
x_2
$h_1(t) = 2.53e^{0.21t}$
$h_2(t) = 3.57e^{0.10t}$
$h_3(t) = 5.29e^{0.44t}$
antelopes [x100]
t [years]


Observation space $\Re^3$
Model space
$r(p_3)$
$r_1(x)$
$r_2(x)$
$r_3(x)$
$h_1(t) = 2.53e^{0.21t}$
$h_2(t) = 3.57e^{0.10t}$
$h_3(t) = 5.29e^{0.44t}$
Parameter space $\Re^2$
p_1
p_2
p_3
x_1
x_2
t [years]
$h_3(4)$
$h_3(2)$
$h_3(1)$
$r_1(p_3)$
$r_2(p_3)$
$r_3(p_3)$
y_1
y_2
y_3


Observation space $\Re^3$
Model space
$r_2(x)$
$r_1(x)$
$r_3(x)$
$r(p_1)$
$r(p_2)$
$r(p_3)$
Parameter space $\Re^2$
x_1
x_2
p_1
p_2
p_3
$h_1(t) = 2.53e^{0.21t}$
$h_2(t) = 3.57e^{0.10t}$
$h_3(t) = 5.29e^{0.44t}$
antelopes [x100]
t [years]

Questions

- How would the dimension of the model/residual/observation/parameter spaces change if the number of observations were increased?
- How would the minimum value change if the number of observations were increased?
- What is the minimum number of observations that can yield a unique solution?
- How would the shape of the contour lines change if the time units was changed from years to decades?

The Gradient and Hessian structure

- The gradient $\nabla f(x)$ of

$$f(x) = \frac{1}{2}\|r(x)\|^2 = \frac{1}{2}r(x)^T r(x)$$

may be derived from the product rule

$$\nabla f(x) = \frac{1}{2}\nabla r(x) r(x) + \frac{1}{2}\left(r(x)^T \nabla r(x)^T\right)^T = J(x)^T r(x),$$

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

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

The Gradient and Hessian structure

- Using the product rule again on

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

the Hessian is

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

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

The Gradient, Jacobian, and Hessian

Example

For the antelope data and model

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

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

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

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.

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

The Gauss-Newton method

The Newton formulation

- 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 Gauss-Newton method

The linear least squares formulation

- Assume we approximate the residual function $r(x)$ with a linear Taylor function, i.e. the 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

$$p = (J_k^T J_k)^{-1} J_k^T (-r_k).$$

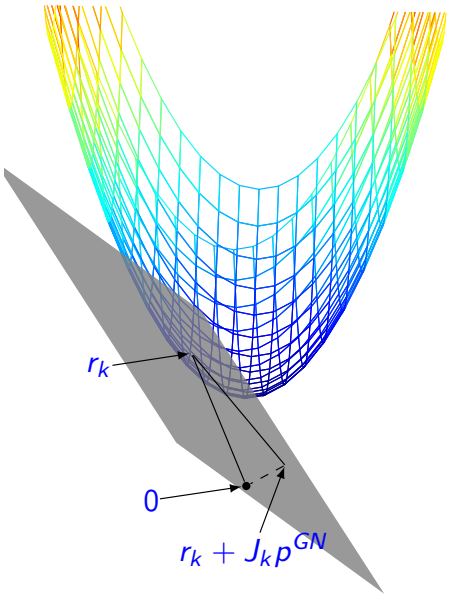
- Thus, the minimizer on the plane corresponds to the Gauss-Newton search direction.

The Gauss-Newton method

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

$$J_k p^{GN} = \underbrace{J_k (J_k^T J_k)^{-1} J_k^T}_{P_{\mathcal{R}(J_k)}} (-r_k).$$



The Gauss-Newton method

Geometrical interpretation of the search direction

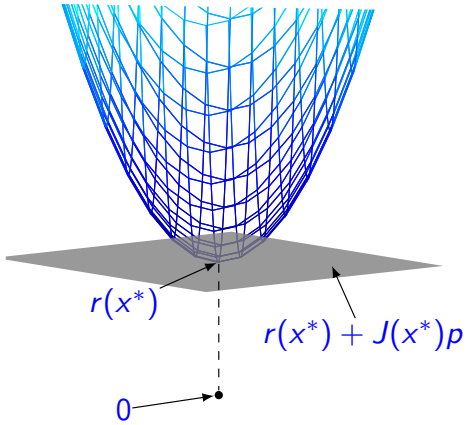
- The first order condition

$$\nabla f(x^*) = 0$$

corresponds to when

$$J(x^*)^T r(x^*) = 0,$$

i.e. when $r(x^*)$ is orthogonal to the tangent plane spanned by the columns of $J(x^*)$.



Geometric interpretation of the first order condition

Zero residual problems

- 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 $r(x^*) = 0$, 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 will not converge no matter how close to the solution we start.

Questions

- The modified Newton uses a scheme to modify the D matrix of the $LDL^T = \nabla^2 f(x_k)$ factorization to ensure a descent direction. Is a similar scheme needed when $\nabla^2 f(x_k)$ is approximated by $J(x)^T J(x)$?