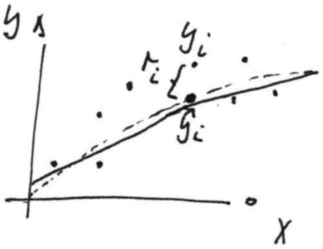
Parameter estimation & optimization

- estimation of parameters appering in models / functions
- efficient use of data for mathematical modelling
- simplest model linear regression



$$y_i = p_0 + p_1 x_i + \varepsilon_i \qquad \varepsilon_i \in \mathcal{N}(0, \sigma^2)$$

Intuition:

p_0, p_1 so wählen dass $y(x) = \hat{p}_0 + p_1 x$ möglichst nah an den Daten liegt d.h. Abstand minimieren

- let $\vec{p}$ be the parameter vector ~~containing the [illegible]~~ of parameters to be optimized

  m: is number of parameters $p_1, \dots, p_m$

- let $\vec{z}$ be the ~~measurement~~ vector which is the "ideal" output of the system to be modeled

  the system in the noise-free case is described by a vector function $\vec{f}$ which relates $\vec{z}$ to $\vec{p}$ ~~and~~ that

$$\vec{f}(\vec{p}, \vec{z}) = 0$$

- In practice measurements $\vec{y}$ are only available for system output $\vec{z}$ with noise

$$\vec{y} = \vec{z} + \vec{\varepsilon}$$

- We take multiple measurements of the system $\{y_i\}$ $i = 1, \dots, n$ and want to estimate $\vec{p}$ using $\{y_i\}$

  → due to noise $\vec{f}(\vec{p}, y_i) = 0$ is not valid anymore

~~the~~ Solution: We write cost function or objective function $\mathcal{F}$ describing the error between measurement and system output for given parameters

$$\mathcal{F}(\vec{p}, y_1, \dots y_n)$$

and minimize the cost

if there are no constraints on $\vec{p}$ and function $\mathcal{F}$ has first and second order partial derivatives, necessary conditions for a minimum are

$$\frac{\partial F}{\partial \vec{p}} = 0 \qquad \text{and} \qquad \frac{\partial^2 F}{\partial^2 \vec{p}^2} > 0$$

Least-square optimization

minimizes the error of sum of squares

$$\min \; \mathcal{F} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

with $\hat{y}_i = \vec{f}(x_i, \hat{\vec{p}})$ model predictions for specific x_i using the parameters

$$\mathcal{F}(\hat{\vec{p}}) = \sum^{n} (y_i - f(x_i, \hat{p}))^2$$

if the errors are normally distributed
the least square estimates are also the maximum likelihood estimates

How to find $\hat{\vec{p}}$ which minimizes $\mathcal{F}$:

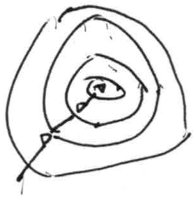
1. Finding an analytical solution by
   - differentiating $\mathcal{F}$ with respect to $p_1, \dots, p_m$
   - setting partial derivatives zero $\frac{\partial \mathcal{F}}{\partial p_i} = 0$
   - solving the resulting m normal equations
   - only working for very few nonlinear models

2 Numerical solutions
   - try different values for parameters $\vec{p}_g$
   - Calculate $\mathcal{F}(\vec{p}_g)$ and work towards smaller $\mathcal{F}$;

3 main procedures (sensitivity based approaches)

- Steepest descent method (gradient descent)

  Searches minimum $\mathcal{F}$ by iteratively determining the direction in which the parameter estimate should change



  $\mathcal{F}(\vec{p})$ is defined and differentiable

  $\mathcal{F}(\vec{p})$ decreases fastest in point $\vec{p} = \vec{a}$ if one goes from $\vec{a}$ in the direction of negative gradient of $\mathcal{F}$ at $\vec{a}$

  $$-\nabla \mathcal{F}(\vec{a})$$

  $$\vec{a}_{k+1} = \vec{a}_k - \gamma \nabla \mathcal{F}(a_k)$$

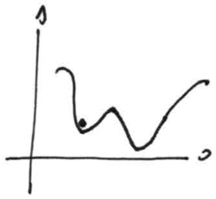
  γ is the step size (allowed to change)

Gauss-Newton method (GNA) or linearization
- uses taylor series expansion to approximate the nonlinear model with linear terms.

  terms are used in linear regression to come up with new terms

(Levenberg-)Marquardt algorithm or damped least squares
interpolates between Gauss Newton & steepest descent algorithm)

Starting values:
- all iterative procedures require starting values
  - risk of local minima (multiple start methods; particle swarms)



Important

- the simpler the model, the better the behavior in the estimation process..
- Over parametrization often leads to convergence problems
  - .. may have multiple solutions
  - . high correlation between parameter estimates

Parametrization can have large influences

$$y_i = \frac{\beta_0 x_i}{\beta_1 x_i + \beta_2} + \varepsilon_i \quad \text{vs.} \quad y_i = \frac{x_i}{c_0 x_i + c_1}$$

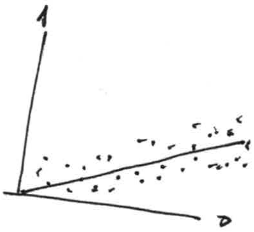
Practical example:

linear regression

$$y_i = \beta_1 \underbrace{x_i}_{\hat{}}^{\;z_i} + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2)$$

$$F(\beta_1) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\min \sum_{i=1}^{n} (y_i - \beta_1 x_i)^2$$



~~$\frac{\partial F}{\partial \beta_1} = \sum_{i=1}^{n} 2(y_i - \beta_1 x_i) \cdot (-x_i) * 0$~~

~~$\sum_{i=1}^{n} (-2 x_i y_i + 2 \hat{\beta}_1 x_i^2) = 0$~~

~~$\sum x_i y_i = \sum \hat{\beta}_1 x_i^2 = \hat{\beta}_1 \sum x_i^2$~~

~~$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2}$~~

~~$\frac{\partial^2 F}{\partial^2 \beta} = 2 \sum_i x_i^2 > 0$~~

$$\beta_0 = \langle \bar{y} \rangle - b_1 \langle x \rangle$$

$$\beta_1 = \frac{\sum_{i=1}^{n} (x_i - \langle x \rangle)(y_i - \langle y \rangle)}{\sum_{i=1}^{n} (x_i - \langle x \rangle)^2}$$

$$= \frac{\mathrm{COV}(x, y)}{\mathrm{VAR}(x)}$$

From where come the ~~derivations~~ derivatives from (for sensitivity based methods)

$\frac{\partial F}{\partial p_i}$:

- often finite differences, but very numerical intensive

$$\frac{\partial F}{\partial p_i}(\vec{a}) \approx \frac{F(\vec{a}) - F(\vec{a} + \Delta p_i)}{p_i - \Delta p_i}$$

- based on sensitivity equations analytical; e.g. use ODE's + F and perform derivatives (symbolic math)

- Stochastic approaches:
  - simulated annealing
  - particle swarm – population of candidate solutions
- Bayes approaches: posterior distributions
  maximum likelihood
  + Gibbs sampling / MCMC sampling
- Identifiability: Profile likelihood
- local vs global maximum: Waterfall plots