Rechnernutzung in der Physik

Karlsruher Institut für Technologie
Wintersemester 2020/21

Teil – 1 “Computerbasierte Methoden der Statistik”

Roger Wolf (roger.wolf@cern.ch), Ralf Ulrich (ralf.ulrich@kit.edu)

Inhalt

Roger Wolf

1 Einführung und Grundlagen

Wahrscheinlichkeit, Statistik, Werkzeuge der statistischen Datenanalyse, …

2 Monte Carlo Methode als numerisches Hilfsmittel

Numerische Integration, Simulation komplexer Zusammenhänge, …

3 Parameterschätzung mit Hilfe der Maximum Likelihood Methode

Likelihood vs. Wahrscheinlichkeit, Maximum Likelihood als Optimierungproblem, …

4 Parameterschätzung mit Hilfe der χ^2-Methode

Ableitung aus Maximum Likelihood Methode, Optimierungsverfahren im allg., …

5 Hypothesentests in der modernen Physik

Begriffe des Hypothesentests, Beispiele, Anwendungen in der Physik, …

Ralf Ulrich

6 Kollaboratives Arbeiten und moderne Softwarewerkzeuge

7 High-Performance Computing: optimales Zusammenspiel von Hard- und Software

Literaturempfehlungen

- **Einführende Literatur zu Statistik und Numerik:**
  - G. Cowan, *Statistical data analysis*, Oxford (1997) (KIT-Bibliothek).
  - G. Bohm, G. Zech, *Einführung in Statistik und Messwertanalyse für Physiker*, DESY (2006) (eBook deutsch, eBook english).
  - V. Blobel, E. Lormann, *Statistische und numerische Methoden der Datenanalyse*, DESY (2012) (Webseite).
  - R. J. Barlow, *Statistics: A Guide to the use of statistical methods in the physical sciences*, Wiley (1989) (KIT-Bibliothek).
  - W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery, *Numerical recipes*, Cambridge Univ. Press (2007) (Webseite).
- Skriptensammlung von Prof. G. Quast (Link):

Zurück zu Folie 34

4 Parameterschätzung mit Hilfe der χ^2 Methode

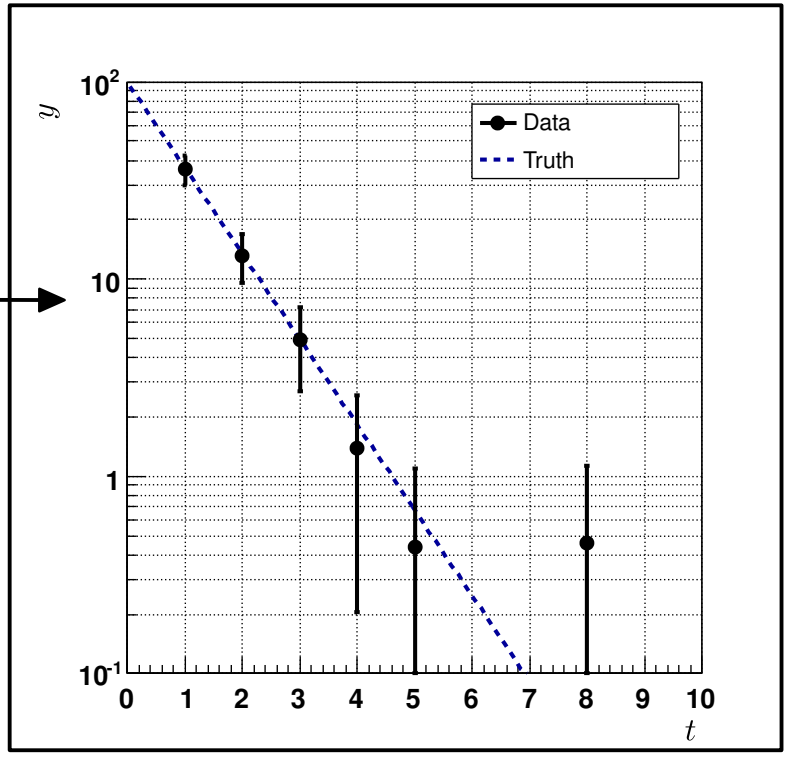
4.1 Maximum Likelihood vs. Least Square

Die Schätzfunktion der kleinsten Fehlerquadrate (Least Square, LS) entspricht der Maximum Likelihood (ML) Abschätzung für normalverteilte Zufallsgrößen.

Likelihood $\leftrightarrow \chi^2$-Methode

- Für die weitere Diskussion nutzen wir das folgende Beispiel:

  9 Messpunkte $\{y_i\}$ mit normalverteilter Wahrscheinlichkeitsdichte mit $\{(\mu_i, \sigma_i)\}$ die einer exponentiellen Verteilung mit fester Normierung folgen (siehe Bild rechts).



- **Likelihood Funktion**:

$$\mathcal{L}(\{y_i\},\theta) = \prod_{i=1}^{9} \frac{1}{\sqrt{2\,\pi\sigma_i^2}} e^{-(y_i-\mu_i(\theta))^2/2\sigma_i^2}$$

$$\ln(\mathcal{L}(\{y_i\},\theta)) = \underbrace{-\frac{1}{2}\sum_{i\leq 9}\frac{(y_i-\mu_i(\theta))^2}{\sigma_i^2}}_{\equiv z} + \underbrace{\sum_{i=1}^{9}\frac{1}{2}\ln\left(2\,\pi\sigma_i^2\right)}_{\text{const.}}$$

- Der quadratische Abstand der Messpunkte von der wahren Verteilung (z) folgt der Verteilung $\chi^2(x,n)\big|_{n=9}$, in unserem Fall mit $n = 9$ **Freiheitsgraden**.

- Die Likelihood wird maximal, wenn z minimal wird.

χ^2-Verteilung (Erinnerung)

Zurück zu Folie 17

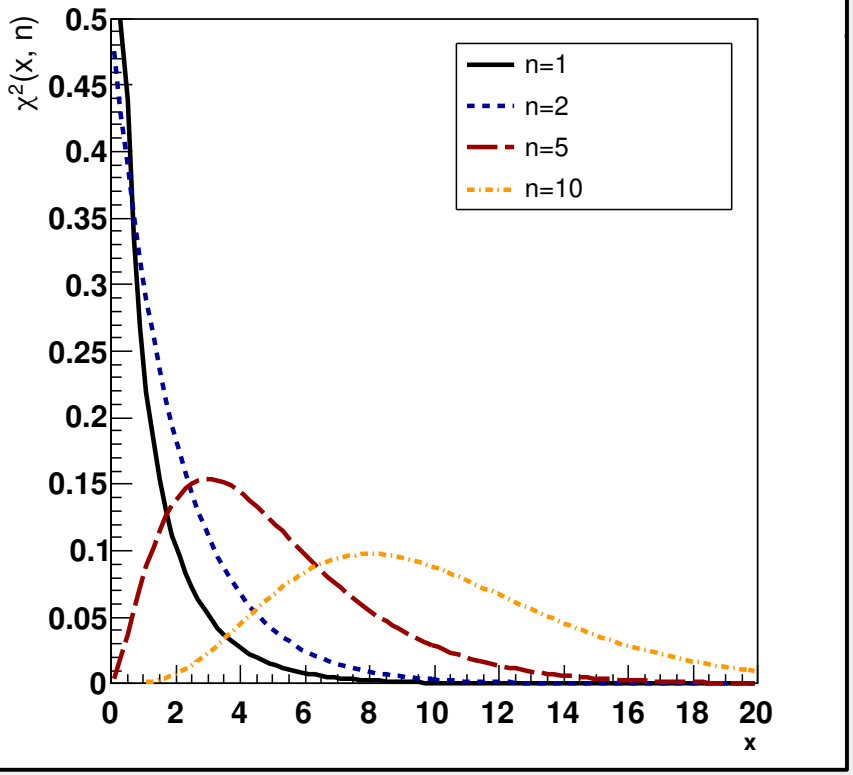
$$\chi^2(x,n) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}$$

mit:

$$\Gamma(x) = \int e^{-t} t^{x-1} \mathrm{d}t$$

$$E[x] = n \qquad \text{(Erwartungswert)}$$

$$\mathrm{var}[x] = 2n \qquad \text{(Varianz)}$$

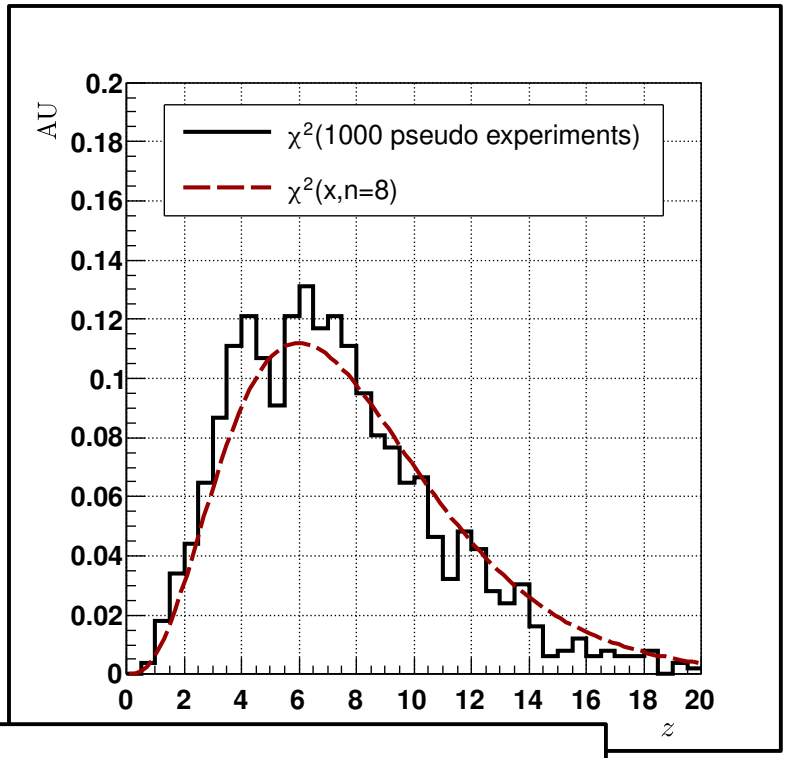


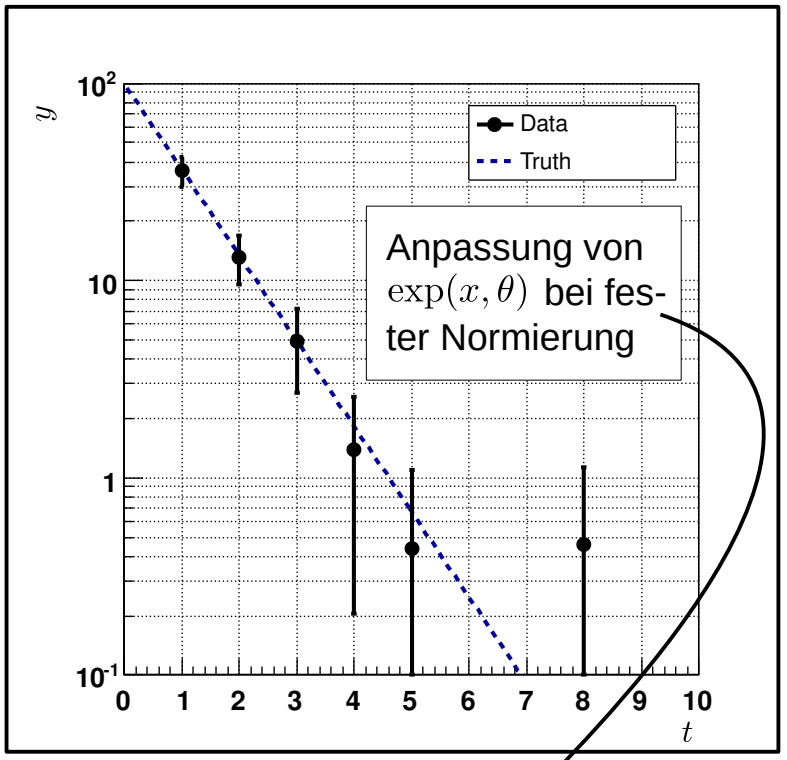
- **NB:**
  Die Summe der Quadrate von n normal-verteilten Zufallsgrößen x_i

$$z = \sum_{i=0}^{N} \frac{(x_i - \mu_i)^2}{\sigma_i^2}$$

ist χ^2-verteilt. Die Zahl n heißt **Freiheitsgrad**. Sie entspricht der Anzahl unabhängiger Normalverteilungen.

z, χ^2-Funktion und Freiheitsgrade bei Anpassungen





9 Messpunkte minus 1 freier Parameter → 8 Freiheitgrade.

NB: Für dieses Histogramm habe ich die 9 Datenpunkte 1000 mal mit einer zugrunde liegenden Normalverteilung mit entsprechender Standardabweichung neu erzeugt. Danach habe ich die Summe der quadratischen Abstände der Messwerte von der angepassten Verteilung bestimmt und in dieses Histogramm abgetragen.

Bei der Anpassung ist Folgendes zu beachten:

Der quadratische Abstand der Messpunkte von der angepassten Verteilung, $z(\hat{\theta})$, folgt $\chi^2(x, n-k)$, wobei k der Anzahl der anzupassenden Parameter entspricht. Ein Messwert legt jeweils einen Parameter der anzupassenden Funktion fest und ist daher nicht mehr von den anderen Messwerten unabhängig.

Definition χ^2-Schätzfunktion

Seien $\{y_i\}$ Einzelmessungen einer Messreihe, deren Erwartungswerte $\{\mu_i\}$ unbekannt sind. Die Varianzen $\{\sigma_i\}$ seien jedoch bekannt. Diese Einzelmessungen müssen nicht unabhängig sein, sondern können über eine bekannte Matrix V_{ij} korreliert sein. Die Funktion

$$\chi^2(\theta) = \sum_{i,j \leq n} (y_i - \mu_i(\theta))^\intercal \; V_{ij}^{-1} \; (y_j - \mu_j(\theta))$$

heißt Schätzfunktion der kleinsten Quadrate (Least Square, LS Schätzfunktion). Die Parameter $\hat{\theta}_{\mathrm{LS}}$, die $\chi^2(\theta)$ minimieren, heißen Schätzwerte der kleinsten Quadrate (LS Schätzwerte).

- Sind die Messwerte $\{y_i\}$ normalverteilt, dann ist die LS Schätzfunktion **zur ML Schätzfunktion äquivalent**.

- Die Martrix V_{ij} ist die Kovarianzmatrix (in der Basis) der Messwerte $\{y_i\}$. Im einfachsten Fall unkorrelierter Varianzen $\{\sigma_i\}$ hat sie die Form:

$$V_{ij} = \begin{pmatrix} \hat{\sigma}_1^2 & 0 & \cdots & 0 \\ 0 & \hat{\sigma}_2^2 & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \ldots & . \end{pmatrix}$$

Zurück zu Folie 9
Zurück zu Folie 11
Zurück zu Folie 14

Lineare LS Schätzfunktion

- Die LS Schätzfunktion kann i.A. numerisch minimiert werden. Wenn sie linear von den Parametern $\{\theta_j\}$ abhängt, d.h. wenn

$$\mu_i(\{\theta_j\}) = \sum_{j \leq n} A_{ij}\theta_j \; ; \qquad \vec{\mu}(\vec{\theta}) = A\,\vec{\theta}$$

gilt ,ist sie auch **analytisch** lösbar. Die $\{\mu_i\}$ hängen hier über die Matrix A linear von den $\{\theta_j\}$ ab. In Matrixschreibweise schreibt sich χ^2 als:

$$\chi^2(\vec{\theta}) = (\vec{y} - \vec{\mu})^\intercal V^{-1} (\vec{y} - \vec{\mu}) = (\vec{y} - A\vec{\theta})^\intercal V^{-1} (\vec{y} - A\vec{\theta})$$

- Um das Minimum zu finden, setzen wir:

$$\vec{\nabla}_\theta \chi^2(\vec{\theta}) = -2\left(A^\intercal V^{-1}\vec{y} - A^\intercal V^{-1} A\,\vec{\theta}\right) = 0$$

$$\hat{\vec{\theta}}_{\mathrm{LS}} = \underbrace{\left(A^\intercal V^{-1} A\right)^{-1} A^\intercal V^{-1}}_{\equiv B}\,\vec{y}$$

d.h. die LS Schätzwerte sind Linearkombinationen der ursprünglichen Messungen:

$$\hat{\theta}_{\mathrm{LS},k} = \sum_{i \leq n} B_{ik} y_i$$

Varianz von $\hat{\theta}_{\mathrm{LS}}$

- Es gibt zwei offensichtliche Methoden die Varianz von $\hat{\theta}_{\mathrm{LS}}$ auszurechnen:

- **Methode 1:** aus Fehlerfortpflanzung (d.h. Variablentransformation) von V.

$$\mathrm{cov}(\theta_i, \theta_j) = \boxed{B V B^\intercal} \equiv (A^\intercal V^{-1} A)^{-1}$$

In Komponentenschreibweise:

$$\mathrm{cov}(\theta_i, \theta_j) \equiv E\left[(\theta_i - E\left[\theta_i\right])\right] E\left[(\theta_j - E\left[\theta_j\right])\right] \; ;$$

$$\mathrm{cov}(y_i, y_j) \equiv E\left[(y_i - E\left[y_i\right])\right] E\left[(y_j - E\left[y_j\right])\right] \equiv V_{ij} \; ; \qquad \theta_i = B_{ij}\, y_j$$

In Matrixschreibweise:

$$\mathrm{cov}(\vec{\theta}, \vec{\theta}^\intercal) \equiv E\left[(B\vec{y} - E[B\vec{y}])\right] E\left[(B\vec{y} - E[B\vec{y}])^\intercal\right] = B\, E\left[(\vec{y} - E[\vec{y}])\right] E\left[(\vec{y} - E[\vec{y}])^\intercal\right]\, B^\intercal$$
$$= B\, V\, B^\intercal$$

Zurück zu Folie 11

Varianz von $\hat{\theta}_{\mathrm{LS}}$

- Es gibt zwei offensichtliche Methoden die Varianz von $\hat{\theta}_{\mathrm{LS}}$ auszurechnen:

- **Methode 1:** aus Fehlerfortpflanzung (d.h. Variablentransformation) von V.

$$\mathrm{cov}(\theta_i, \theta_j) = BVB^\intercal \equiv \boxed{(A^\intercal V^{-1} A)^{-1}}$$

$$\left((A^\intercal V^{-1} A)^{-1} A^\intercal V^{-1}\right) V \left((A^\intercal V^{-1} A)^{-1} A^\intercal V^{-1}\right)^\intercal =$$

$$(A^\intercal)^{-1} V A^{-1} A^\intercal \underbrace{V^{-1} V}_{\equiv 1} (V^{-1})^\intercal A (A^{-1})^\intercal V^\intercal A^{-1}$$

$$(A^\intercal)^{-1} V \underbrace{\underbrace{A^{-1} A^\intercal (V^{-1})^\intercal}_{\equiv X} \underbrace{A (A^{-1})^\intercal V^\intercal}_{\equiv X^{-1}}}_{\equiv 1} A^{-1} = (A^\intercal)^{-1} V A^{-1} = (A^\intercal V^{-1} A)^{-1}$$

Unter Verwendung von:

$(AB)^\intercal = B^\intercal A^\intercal$

$(A^{-1})^\intercal = (A^\intercal)^{-1}$

Varianz von $\hat{\theta}_{\mathrm{LS}}$

(vgl. mit VL-03 Folie 34)

- Es gibt zwei offensichtliche Methoden die Varianz von $\hat{\theta}_{\mathrm{LS}}$ auszurechnen:

- **Methode 2:** graphisch aus der Abweichung $\Delta\chi^2$ vom Minimum $\chi^2_{\mathrm{min}} = \chi^2(\hat{\theta}_{\mathrm{LS}})$.

$$\chi^2(\theta) \approx \underbrace{\chi^2(\hat{\theta}_{\mathrm{LS}})}_{\equiv \chi^2_{\mathrm{min}}} + \underbrace{\frac{1}{2}\left[\frac{\partial^2}{\partial\theta^2}\chi^2(\theta)\right]_{\theta=\hat{\theta}_{\mathrm{LS}}}}_{\hat{\sigma}_\theta^{-2}} \left(\theta - \hat{\theta}_{\mathrm{LS}}\right)^2$$

(vgl mit Folie 6)

$$\chi^2\left(\hat{\theta}_{\mathrm{LS}} \pm \hat{\sigma}_\theta\right) = \chi^2_{\mathrm{min}} + 1$$

d.h. die Variation aus dem Minimum um $\pm\hat{\sigma}$ bewirkt die Erhöhung von $\chi^2(\hat{\theta}_{\mathrm{ML}})$ um 1.

Beispiel in 1-dim

Varianz von $\hat{\theta}_{\mathrm{LS}}$

(vgl. mit VL-03 Folie 34)

- Es gibt zwei offensichtliche Methoden die Varianz von $\hat{\theta}_{\mathrm{LS}}$ auszurechnen:

- **Methode 2:** graphisch aus der Abweichung $\Delta\chi^2$ vom Minimum $\chi^2_{\mathrm{min}} = \chi^2(\hat{\theta}_{\mathrm{LS}})$.

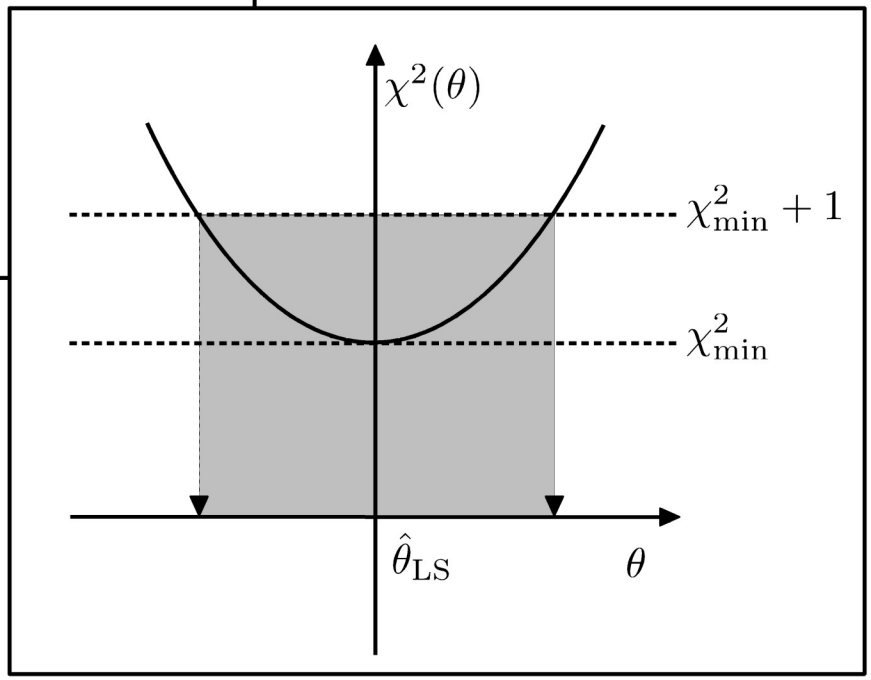
$$\chi^2(\theta) \approx \underbrace{\chi^2(\hat{\theta}_{\mathrm{LS}})}_{\equiv \chi^2_{\mathrm{min}}} + \underbrace{\frac{1}{2}\left[\frac{\partial^2}{\partial\theta^2}\chi^2(\theta)\right]_{\theta=\hat{\theta}_{\mathrm{LS}}}}_{\hat{\sigma}_\theta^{-2}} \left(\theta - \hat{\theta}_{\mathrm{LS}}\right)^2$$

(vgl mit Folie 6)

$$\chi^2\left(\hat{\theta}_{\mathrm{LS}} \pm \hat{\sigma}_\theta\right) = \chi^2_{\mathrm{min}} + 1$$

d.h. die Variation aus dem Minimum um $\pm\hat{\sigma}$ bewirkt die Erhöhung von $\chi^2(\hat{\theta}_{\mathrm{ML}})$ um 1.

Beispiel in 1-dim



Varianz von $\hat{\theta}_{\mathrm{LS}}$

(vgl. mit VL-03 Folie 34)

- Es gibt zwei offensichtliche Methoden die Varianz von $\hat{\theta}_{\mathrm{LS}}$ auszurechnen:

- **Methode 2:** graphisch aus der Abweichung $\Delta\chi^2$ vom Minimum $\chi^2_{\mathrm{min}} = \chi^2(\hat{\theta}_{\mathrm{LS}})$.

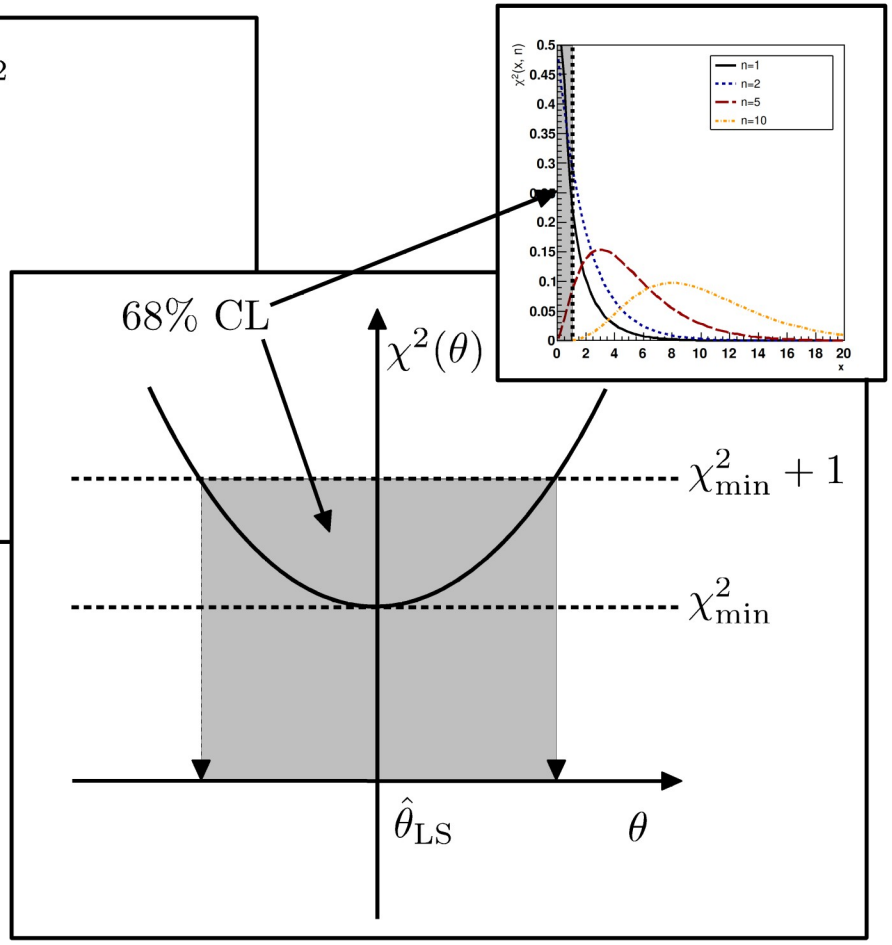
$$\chi^2(\theta) \approx \underbrace{\chi^2(\hat{\theta}_{\mathrm{LS}})}_{\equiv \chi^2_{\mathrm{min}}} + \underbrace{\frac{1}{2}\left[\frac{\partial^2}{\partial\theta^2}\chi^2(\theta)\right]_{\theta=\hat{\theta}_{\mathrm{LS}}}}_{\hat{\sigma}_\theta^{-2}} \left(\theta - \hat{\theta}_{\mathrm{LS}}\right)^2$$

(vgl mit Folie 6)

$$\chi^2\left(\hat{\theta}_{\mathrm{LS}} \pm \hat{\sigma}_\theta\right) = \chi^2_{\mathrm{min}} + 1$$

d.h. die Variation aus dem Minimum um $\pm\hat{\sigma}$ bewirkt die Erhöhung von $\chi^2(\hat{\theta}_{\mathrm{ML}})$ um 1.

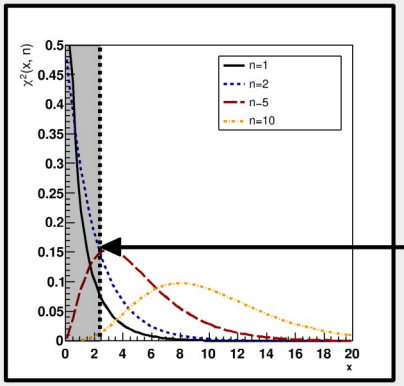
Beispiel in 1-dim



Ein lauffähiges RooT macro finden Sie hier.

Varianz von $\hat{\theta}_{\mathrm{LS}}$

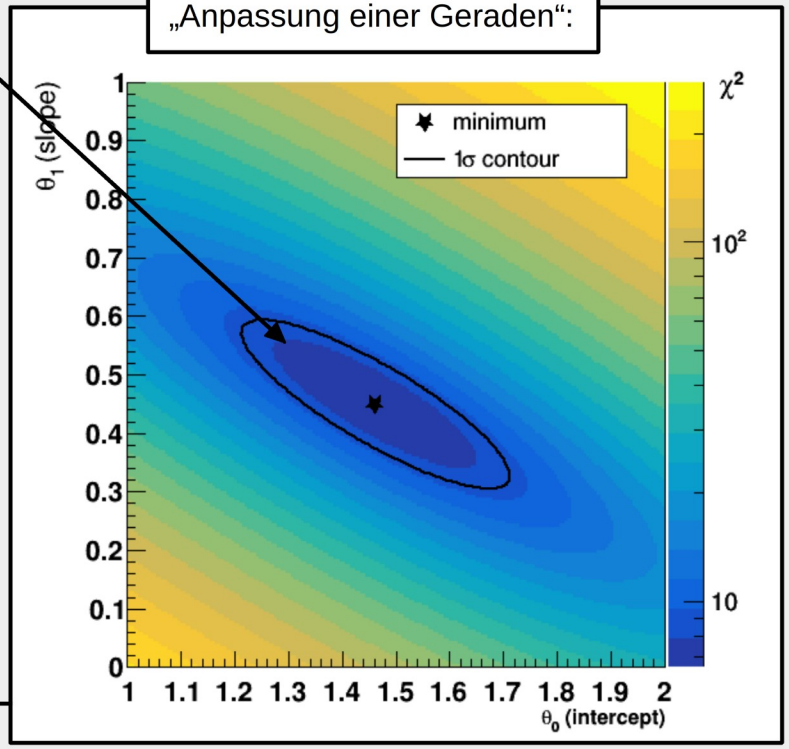
- Es gibt zwei offensichtliche Methoden die Varianz von $\hat{\theta}_{\mathrm{LS}}$ auszurechnen:
- **Methode 2:** graphisch aus der Abweichung $\Delta\chi^2$ vom Minimum $\chi^2_{\mathrm{min}} = \chi^2(\hat{\theta}_{\mathrm{LS}})$.



68% CL

Beispiel in 2-dim

„Anpassung einer Geraden“:



NB: für n unabhängige Parameter müssen Sie die Differenz $\Delta\chi^2$ zu χ^2_{min} als 68% Quantile einer $\chi^2(x,n)$ - Funktion mit n Freiheitsgraden bestimmen. Das können Sie z.B. unter diesem link tun. Im Folgenden sind einige Werte angegeben:

Dim.	$\Delta\chi^2(68\%\,\mathrm{CL})$	$\Delta\chi^2(95\%\,\mathrm{CL})$
1	1.0	3.8
2	2.3	6.0
3	3.5	7.8

Ein lauffähiges RooT macro finden Sie hier.

Zurück zu Folie 28

Vergleich mit der RCF-Ungleichung

(siehe VL-03 Folie 38)

- Für normalverteilte Zufallsgrößen gilt:

$$\mathrm{cov}[\theta_i, \theta_j] \geq \left[-\frac{\partial^2}{\partial\theta_i \partial\theta_j} \underbrace{\ln \mathcal{L}(\vec{\theta})}_{\equiv -\frac{1}{2}\chi^2(\vec{\theta})} \right]^{-1}_{\substack{\theta_i = \hat{\theta}_i \\ \theta_j = \hat{\theta}_j}} = A^\intercal V^{-1} A$$

$$\equiv A^\intercal V^{-1} A$$

vgl. mit Folie 8.

für normalverteilte Zufallsgrößen

$$\frac{\partial^2}{\partial\theta_i \partial\theta_j} \chi^2(\theta) = 2 A^\intercal V^{-1} A$$

vgl. mit Folie 6.

- Für normalverteilte Zufallsgrößen (und nur für diese!) ist die LS Abschätzung **effizient**.

4 Parameterschätzung mit Hilfe der χ^2 Methode

Zurück zu Folie 20
Zurück zu Folie 30
Zurück zu Folie 34

4.2 Beispiel für eine lineare LS Anpassung

Wir diskutieren auf den folgenden Folien das Beispiel einer linearen LS Anpassung einmal von Anfang bis Ende duch.

Anpassung einer Geraden an 5 Messpunkte

$$\mu_i(\{\theta_j\}) = \sum_{j \leq n} A_{ij}\theta_j$$

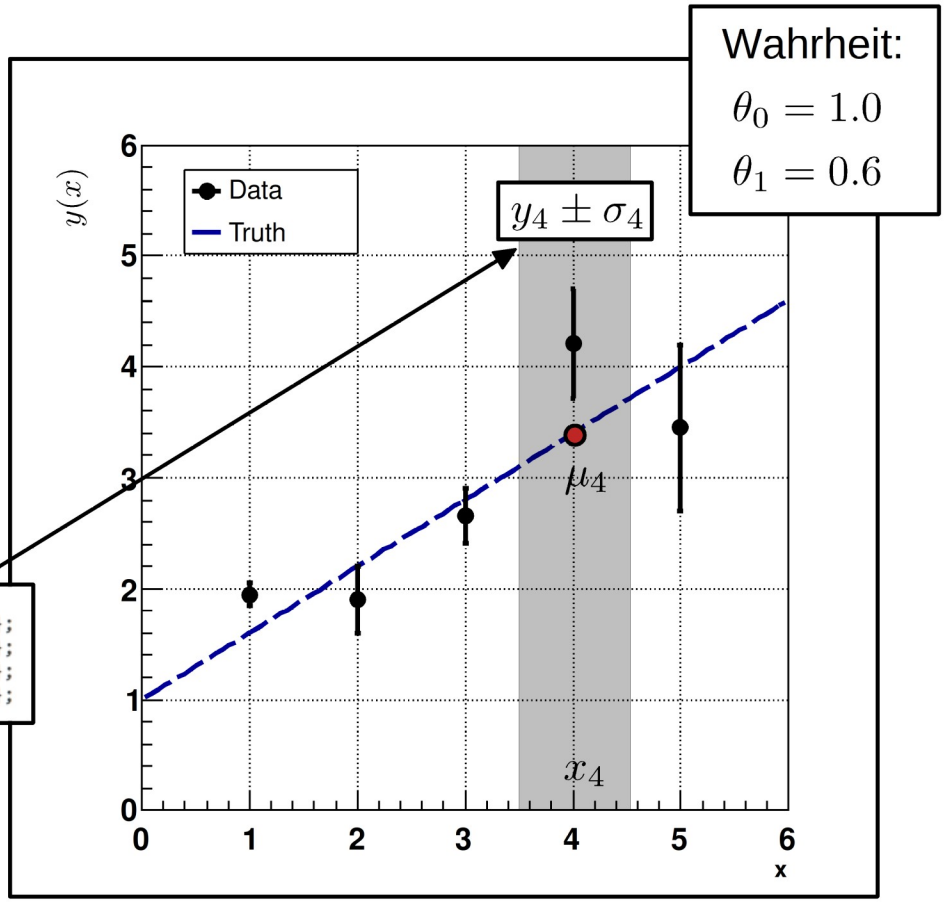
mit:

$$i = 1 \ldots 5\,; \quad j = 0, 1$$

- Fünf unkorrelierte Datenpunkte $\{y_i\}$ nach einer unbekannten linearen Funktion normalverteilt:

```
static const int LENGTH = 5;
static float XVALUES[] = {   1.00,    2.00,    3.00,    4.00,   5.00};
static float YVALUES[] = {1.94759, 1.90523, 2.65621, 4.20916, 3.44776};
static float XERRORS[] = {   0.00,    0.00,    0.00,    0.00,   0.00};
static float YERRORS[] = {   0.10,    0.30,    0.25,    0.50,   0.75};
```

- Anpassung einer Geraden mit y-Achsen-abschnitt θ_0 und Steigung θ_1.



Ein lauffähiges RooT macro finden Sie hier.

Algebra von Folie 6

$$\vec{\hat{\theta}}_{\mathrm{LS}} = \underbrace{\left(A^{\intercal}V^{-1}A\right)^{-1}A^{\intercal}V^{-1}}_{\equiv B}\vec{y}$$

– Allgemeine Lösung –

$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_5 \end{pmatrix}}_{\vec{y}} = \underbrace{\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_5 \end{pmatrix}}_{A} \cdot \underbrace{\begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix}}_{\vec{\theta}}$$

$$V^{-1} = \begin{pmatrix} 1/\hat{\sigma}_1^2 & 0 & \cdots & 0 \\ 0 & 1/\hat{\sigma}_2^2 & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \ldots & 1/\hat{\sigma}_5^2 \end{pmatrix}$$

– Formulierung des Problems –

$$A^{\intercal}V^{-1}A = \begin{pmatrix} \sum 1/\hat{\sigma}_i^2 & \sum x_i/\hat{\sigma}_i^2 \\ \sum x_i/\hat{\sigma}_i^2 & \sum x_i^2/\hat{\sigma}_i^2 \end{pmatrix}$$

$$\left(A^{\intercal}V^{-1}A\right)^{-1} \overset{(4)}{=} \frac{1}{\det(A^{\intercal}V^{-1}A)} \begin{pmatrix} \sum x_i^2/\hat{\sigma}_i^2 & -\sum x_i/\hat{\sigma}_i^2 \\ -\sum x_i/\hat{\sigma}_i^2 & \sum 1/\hat{\sigma}_i^2 \end{pmatrix} \equiv \mathrm{cov}[\theta_0, \theta_1]$$

$$A^{\intercal}V^{-1}\vec{y} = \begin{pmatrix} \sum \mu_i/\hat{\sigma}_i^2 \\ \sum \mu_i x_i/\hat{\sigma}_i^2 \end{pmatrix}$$

$$\begin{pmatrix} \hat{\theta}_0 \\ \hat{\theta}_1 \end{pmatrix} \overset{(4)}{=} \frac{1}{\det(A^{\intercal}V^{-1}A)} \begin{pmatrix} \left(\sum x_i^2/\hat{\sigma}_i^2\right)\left(\sum y_i/\hat{\sigma}_i^2\right) - \left(\sum x_i/\hat{\sigma}_i^2\right)\left(\sum y_i x_i/\hat{\sigma}_i^2\right) \\ \left(\sum 1/\hat{\sigma}_i^2\right)\left(\sum y_i x_i/\hat{\sigma}_i^2\right) - \left(\sum x_i/\hat{\sigma}_i\right)\left(\sum y_i/\hat{\sigma}_i^2\right) \end{pmatrix}$$

(4)
$$\det(A^{\intercal}V^{-1}A) = \left(\sum 1/\sigma_i^2\right)\left(\sum x_i^2/\sigma_i^2\right) - \left(\sum x_i/\sigma_i^2\right)^2$$

Algebra von Folie 6

$$\vec{\hat{\theta}}_{\mathrm{LS}} = \underbrace{\left(A^{\intercal}V^{-1}A\right)^{-1}A^{\intercal}V^{-1}}_{\equiv B}\vec{y}$$

– Allgemeine Lösung –

$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_5 \end{pmatrix}}_{\vec{y}} = \underbrace{\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_5 \end{pmatrix}}_{A} \cdot \underbrace{\begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix}}_{\vec{\theta}}$$

$$V^{-1} = \begin{pmatrix} 1/\hat{\sigma}_1^2 & 0 & \cdots & 0 \\ 0 & 1/\hat{\sigma}_2^2 & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \ldots & 1/\hat{\sigma}_5^2 \end{pmatrix}$$

– Vereinfachung des Problems –

$$A^{\intercal}V^{-1}A = \begin{pmatrix} \sum 1/\hat{\sigma}_i^2 & \sum x_i/\hat{\sigma}_i^2 \\ \sum x_i/\hat{\sigma}_i^2 & \sum x_i^2/\hat{\sigma}_i^2 \end{pmatrix}$$

$$\left(A^{\intercal}V^{-1}A\right)^{-1} \overset{(4)}{=} \frac{1}{\det(A^{\intercal}V^{-1}A)} \begin{pmatrix} \sum x_i^2/\hat{\sigma}_i^2 & -\sum x_i/\hat{\sigma}_i^2 \\ -\sum x_i/\hat{\sigma}_i^2 & \sum 1/\hat{\sigma}_i^2 \end{pmatrix} \equiv \mathrm{cov}[\theta_0, \theta_1]$$

$$A^{\intercal}V^{-1}\vec{y} = \begin{pmatrix} \sum \mu_i/\hat{\sigma}_i^2 \\ \sum \mu_i x_i/\hat{\sigma}_i^2 \end{pmatrix}$$

Steigung ($\hat{\theta}_1$) und y-Achsenabschnitt ($\hat{\theta}_0$) sind antikorreliert.

$$\begin{pmatrix} \hat{\theta}_0 \\ \hat{\theta}_1 \end{pmatrix} \overset{(4)}{=} \frac{1}{\det(A^{\intercal}V^{-1}A)} \begin{pmatrix} \left(\sum x_i^2/\hat{\sigma}_i^2\right)\left(\sum y_i/\hat{\sigma}_i^2\right) - \left(\sum x_i/\hat{\sigma}_i^2\right)\left(\sum y_i x_i/\hat{\sigma}_i^2\right) \\ \left(\sum 1/\hat{\sigma}_i^2\right)\left(\sum y_i x_i/\hat{\sigma}_i^2\right) - \left(\sum x_i/\hat{\sigma}_i\right)\left(\sum y_i/\hat{\sigma}_i^2\right) \end{pmatrix}$$

Priv. Doz. Dr. Roger Wolf
http://ekpwww.physik.uni-karlsruhe.de/~rwolf/

(4)
$$\det(A^{\intercal}V^{-1}A) = \left(\sum 1/\sigma_i^2\right)\left(\sum x_i^2/\sigma_i^2\right) - \left(\sum x_i/\sigma_i^2\right)^2$$

Zurück zu Folie 21

Lauffähiges C++ Beispiel

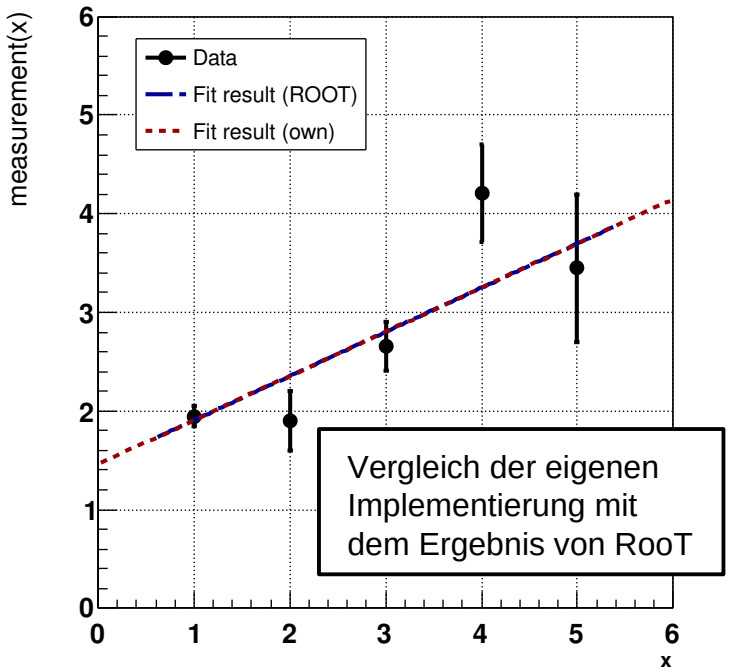
Ein lauffähiges RooT macro finden Sie hier.

```
std::pair<TMatrixD,TMatrixD> LLS(TMatrixD& A, TMatrixD& V, TMatrixD& X){
  TMatrixD U=TMatrixD(A,TMatrixD::kTransposeMult,TMatrixD(V,TMatrixD::kMult,A)).Invert();   /* (A^t*V*A)^(-1)       */
  TMatrixD B=TMatrixD(U,TMatrixD::kMult,TMatrixD(A,TMatrixD::kTransposeMult,V));           /* (A^t*V*A)^(-1)*A^t*V */
  TMatrixD T=TMatrixD(B,TMatrixD::kMult,X);                                                /* fit result           */
  return std::make_pair(T,U);
}

/* ------------------------------------------------------------------------------
 *  Anpassung einer Geraden (f(x)=a[0]+a[1]*x)
 *------------------------------------------------------------------------------*/
const int NPARAM=2;                              /* Koeffizientenmatrix          */
TMatrixD A(LENGTH,NPARAM);                       /*                              */
for(int i=0;i<LENGTH;++i){                       /*                              */
  for(int j=0;j<NPARAM;++j){                     /*                              */
    A(i,j)=(j==0?1.0:XVALUES[i]);                /*                              */
  }
}
// std::cout << "PRINTING A" << std::endl; print(A);
// inverse Korrelationsmatrix
TMatrixD V(LENGTH,LENGTH);                       /* inverse Korrelationsmatrix   */
for(int i=0;i<LENGTH;++i){                       /*                              */
  for(int j=0;j<LENGTH;++j){                     /*                              */
    V(i,j)=(i==j?1./(YERRORS[i]*YERRORS[i]):0.); /*                              */
  }
}
// std::cout << "PRINTING V" << std::endl; print(V);
TMatrixD Y(LENGTH,1);                            /* Vektor der Messwerte         */
for(int i=0;i<LENGTH;++i){                       /*                              */
  Y(i,0)=YVALUES[i];                             /*                              */
}
// std::cout << "PRINTING X" << std::endl; print(X);
std::pair<TMatrixD,TMatrixD> fit_result = LLS(A,V,Y);
TMatrixD T = fit_result.first;
TMatrixD U = fit_result.second;
```

LS Schätzwerte:	Wahrheit:
$\hat{\theta}_0 = 1.46 \pm 0.16$	$\theta_0 = 1.0$
$\hat{\theta}_1 = 0.45 \pm 0.09$	$\theta_1 = 0.6$
$\rho(\hat{\theta}_0, \hat{\theta}_1) = -0.85$	

$$\vec{\hat{\theta}}_{\mathrm{LS}} = \underbrace{\left(A^{\intercal} V^{-1} A\right)^{-1} A^{\intercal} V^{-1}}_{\equiv B} \vec{y}$$



Zurück zu Folie 21

Lauffähiges C++ Beispiel

Ein lauffähiges RooT macro finden Sie hier.

```
std::pair<TMatrixD,TMatrixD> LLS(TMatrixD& A, TMatrixD& V, TMatrixD& X){
  TMatrixD U=TMatrixD(A,TMatrixD::kTransposeMult,TMatrixD(V,TMatrixD::kMult,A)).Invert();   /* (A^t*V*A)^(-1)       */
  TMatrixD B=TMatrixD(U,TMatrixD::kMult,TMatrixD(A,TMatrixD::kTransposeMult,V));           /* (A^t*V*A)^(-1)*A^t*V */
  TMatrixD T=TMatrixD(B,TMatrixD::kMult,X);                                                /* fit result           */
  return std::make_pair(T,U
}

/* ------------------------
 *  Anpassung einer Geraden
 *------------------------
const int NPARAM=2;
TMatrixD A(LENGTH,NPARAM);
for(int i=0;i<LENGTH;++i){
  for(int j=0;j<NPARAM;++j)
    A(i,j)=(j==0?1.0:XVALUE
  }
}
// std::cout << "PRINTING A
// inverse Korrelationsmatr
TMatrixD V(LENGTH,LENGTH);
for(int i=0;i<LENGTH;++i){
  for(int j=0;j<LENGTH;++j)
    V(i,j)=(i==j?1./(YERROR
  }
}
// std::cout << "PRINTING V
TMatrixD Y(LENGTH,1);
for(int i=0;i<LENGTH;++i){
  Y(i,0)=YVALUES[i];
}
// std::cout << "PRINTING X
std::pair<TMatrixD,TMatrixD
TMatrixD T = fit_result.first;
TMatrixD U = fit_result.second;
```

Durch geschickte Wahl können die Parameter $\{\theta_j\}$ dekorreliert werden

$$y(x) = \theta_0 + \theta_1(x - \bar{x})$$

mit:

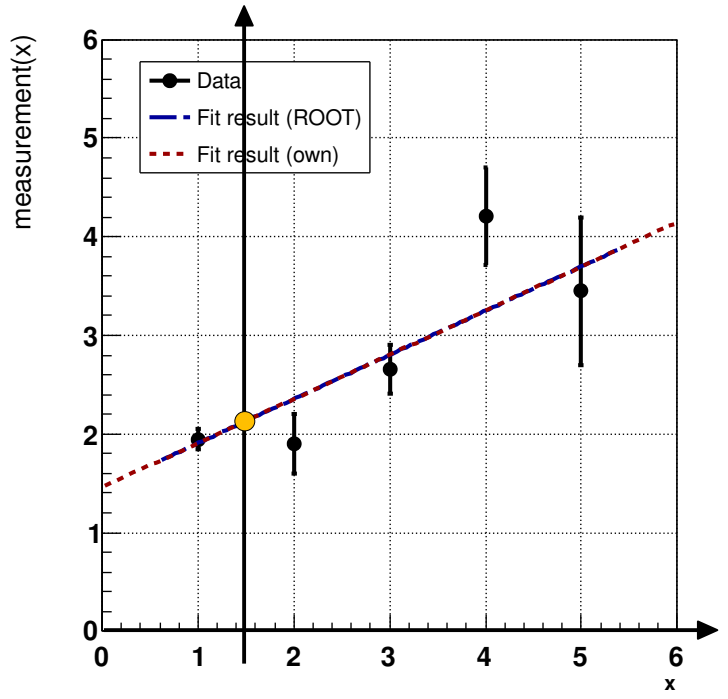
$$\bar{x} = \frac{\sum x_i/\sigma_i^2}{\sum 1/\sigma_i^2} = 1.47$$

führt auf

$$\rho(\hat{\theta}_0, \hat{\theta}_1) = 10^{-8}$$

LS Schätzwerte:	Wahrheit:
$\hat{\theta}_0 = 1.46 \pm 0.16$	$\theta_0 = 1.0$
$\hat{\theta}_1 = 0.45 \pm 0.09$	$\theta_1 = 0.6$
$\rho(\hat{\theta}_0, \hat{\theta}_1) = -0.85$	

$$\vec{\hat{\theta}}_{\mathrm{LS}} = \underbrace{\left(A^{\intercal} V^{-1} A\right)^{-1} A^{\intercal} V^{-1}}_{\equiv B} \vec{y}$$



Zurück zu Folie 21

Lauffähiges C++ Beispiel

Frage:
Ist die Anpassung einer Parabel auch ein Beispiel für ein lineares LS-Problem?

Ein lauffähiges RooT macro finden Sie hier.

```
std::pair<TMatrixD,TMatrixD> LLS(TMatrixD& A, TMatrixD& V, TMatrixD& X){
  TMatrixD U=TMatrixD(A,TMatrixD::kTransposeMult,TMatrixD(V,TMatrixD::kMult,A)).Invert();   /* (A^t*V*A)^(-1)       */
  TMatrixD B=TMatrixD(U,TMatrixD::kMult,TMatrixD(A,TMatrixD::kTransposeMult,V));           /* (A^t*V*A)^(-1)*A^t*V */
  TMatrixD T=TMatrixD(B,TMatrixD::kMult,X);                                                /* fit result           */
  return std::make_pair(T,U
}

/* ------------------------
 *  Anpassung einer Geraden
 *------------------------
const int NPARAM=2;
TMatrixD A(LENGTH,NPARAM);
for(int i=0;i<LENGTH;++i){
  for(int j=0;j<NPARAM;++j)
    A(i,j)=(j==0?1.0:XVALUE
  }
}
// std::cout << "PRINTING A
// inverse Korrelationsmatr
TMatrixD V(LENGTH,LENGTH);
for(int i=0;i<LENGTH;++i){
  for(int j=0;j<LENGTH;++j)
    V(i,j)=(i==j?1./(YERROR
  }
}
// std::cout << "PRINTING V
TMatrixD Y(LENGTH,1);
for(int i=0;i<LENGTH;++i){
  Y(i,0)=YVALUES[i];
}
// std::cout << "PRINTING X
std::pair<TMatrixD,TMatrixD
TMatrixD T = fit_result.first;
TMatrixD U = fit_result.second;
```

Durch geschickte Wahl können die Parameter $\{\theta_j\}$ dekorreliert werden

$$y(x) = \theta_0 + \theta_1(x - \bar{x})$$

mit:

$$\bar{x} = \frac{\sum x_i/\sigma_i^2}{\sum 1/\sigma_i^2} = 1.47$$

führt auf

$$\rho(\hat{\theta}_0, \hat{\theta}_1) = 10^{-8}$$

LS Schätzwerte:

$\hat{\theta}_0 = 1.46 \pm 0.16$

$\hat{\theta}_1 = 0.45 \pm 0.09$

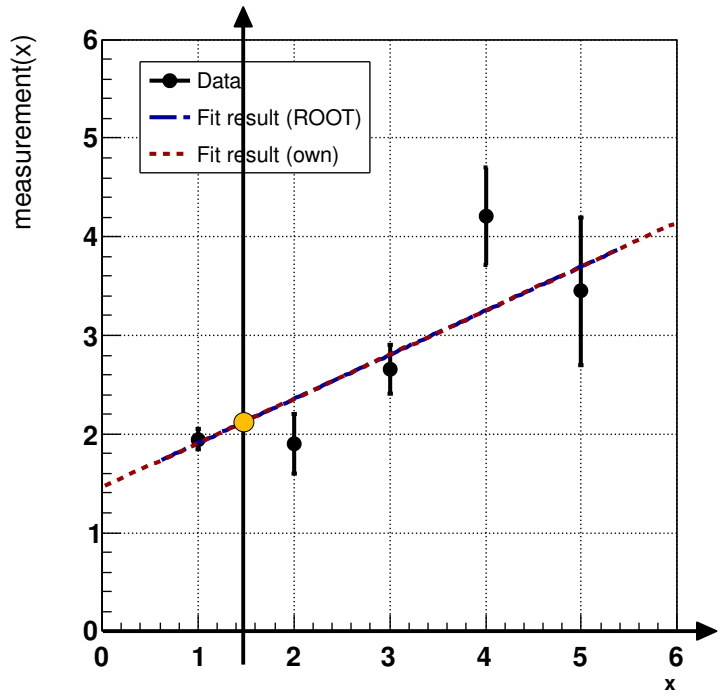
$\rho(\hat{\theta}_0, \hat{\theta}_1) = -0.85$

Wahrheit:

$\theta_0 = 1.0$

$\theta_1 = 0.6$

$$\vec{\hat{\theta}}_{\mathrm{LS}} = \underbrace{\left(A^\intercal V^{-1} A\right)^{-1} A^\intercal V^{-1}}_{\equiv B} \vec{y}$$



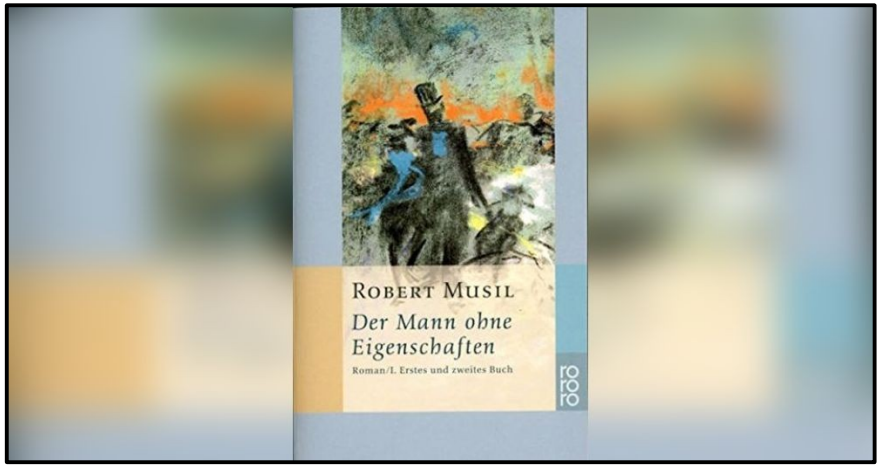
Priv. Doz. Dr. Roger Wolf
http://ekpwww.physik.uni-karlsruhe.de/~rwolf/

Zurück zu Folie 34

4 Parameterschätzung mit Hilfe der χ^2 Methode

4.3 Eigenschaften der LS Anpassung

Wir fassen im Folgenden einige Eigenschaften der LS Abschätzung zusammen.



Bedeutung von χ^2_{obs} für die LS-Schätzung

Für normalverteilte Zufallsgrößen folgt der quadratische Abstand der LS-Schätzwerte $\hat{y}_i\left(\{x_i\},\{\hat{\theta}_j\}\right)$ von den Messwerten $\{y_i\}$ für $j=1,\ldots,k$

$$\chi^2_{\text{obs}} = \sum_{i\leq n}\left(\hat{y}_i\left(\{x_i\},\hat{\theta}_j\right)-y_i\right)^2$$

einer $\chi^2(x,n-k)$-Verteilung mit $n-k$ Freiheitsgraden.

(siehe Folie 3)

- Dabei enspricht n der Anzahl der Messungen und k der Anzahl der Parameter zur Anpassung.

- Da der Erwartungswert von $\chi^2(x,n-k)$ $n-k$ ist, ist bei zugrundeliegender Normalverteilung der Messwerte

$$\frac{\chi^2_{\text{obs}}}{n-k}\approx 1$$

zu erwarten.

Bedeutung von $\chi^2_{\rm obs}$ für die LS-Schätzung

- Unter der Annahme, dass Ihren Messwerten eine Normalverteilung zugrundeliegt, können Sie die folgenden Schlüsse ziehen:

$$\frac{\chi^2_{\rm obs}}{n-k} \ll 1 \quad \Rightarrow \quad \text{Fehler zu groß abgeschätzt oder zu viele Parameter im Modell?}$$

$$\frac{\chi^2_{\rm obs}}{n-k} \gg 1 \quad \Rightarrow \quad \text{Fehler zu klein abgeschätzt oder Modell falsch.}$$

Verteilungsfreiheit

- Beachten Sie, dass für die LS-Schätzung nur die Werte ($\{y_i\}$) und Unsicherheiten ($\{\sigma_i\}$) der Messreihe angegeben werden. Im Gegensatz zur ML-Schätzung gibt es keine Aussage über die zugrundeliegende Wahrscheinlichkeitsdichte der Einzelmessungen!

- Man bezeichnet diese Eigenschaft als **Verteilungsfreiheit** der LS-Schätzung.

Verteilungsfreiheit – Demonstration

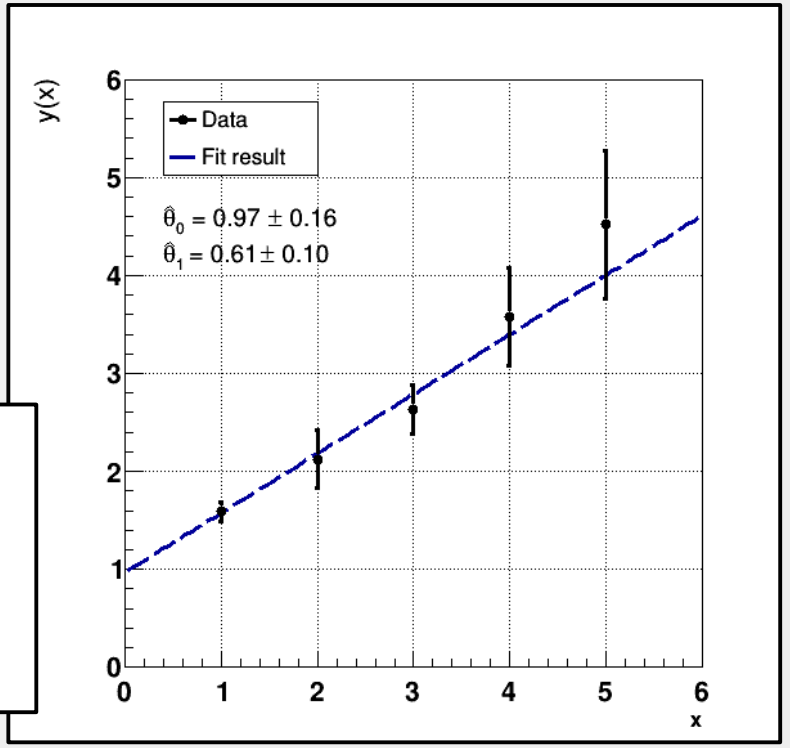
- Wir demonstrieren diese Eigenschaft des LS-Abschätzung anhand des Beispiels aus Abschnitt 4.2:

- Erinnerung:

Wahrheit:

$\theta_0 = 1.0$

$\theta_1 = 0.6$

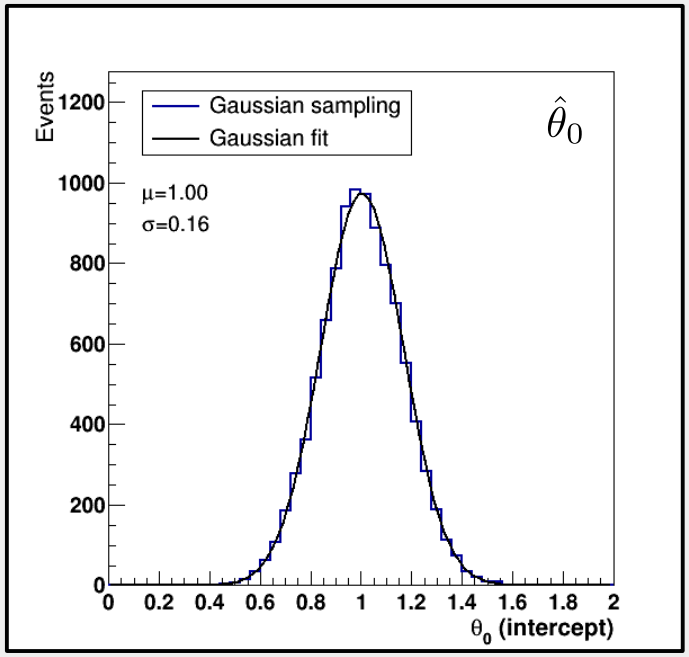


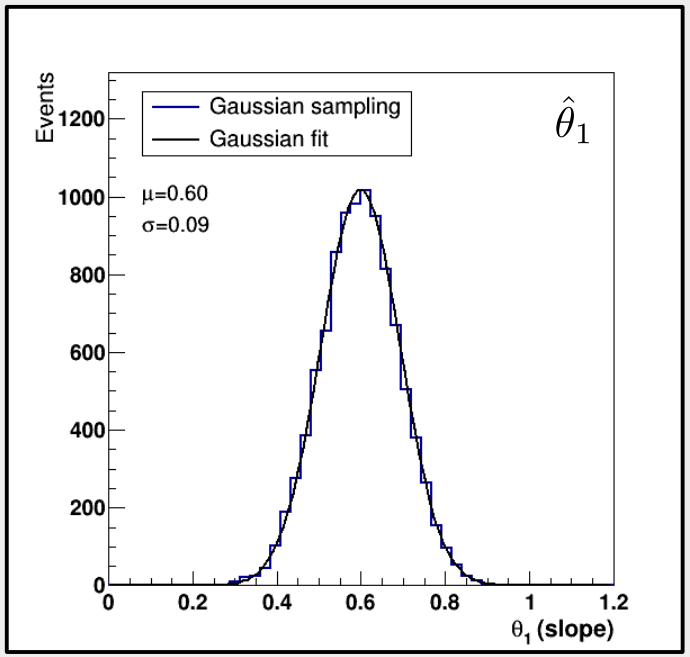
- Rechts sehen Sie einen möglichen Ausgang der Messung als Pseudoexperiment, wofür die Messwerte zufällig neuverteilt (randomisiert) wurden.

- Die zugrundeliegende Wahrscheinlichkeitsdichte für jede Einzelmessung war dabei vorgegeben durch $\varphi(x, y_i, \sigma_i)$.

Zugrundeliegende Normalverteilung

- Hier sehen Sie die Ausgänge der Messung nach 10k-facher Wiederholung des Pseudo-experiments, wobei jede Einzelmessung nach $\varphi(x, y_i, \sigma_i)$ *normalverteilt* ist:





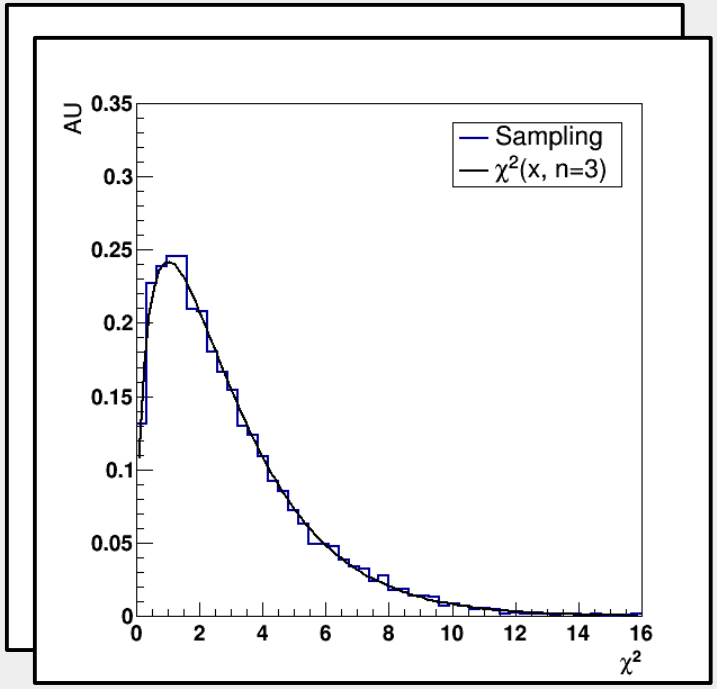
Ein lauffähiges RooT macro finden Sie hier.

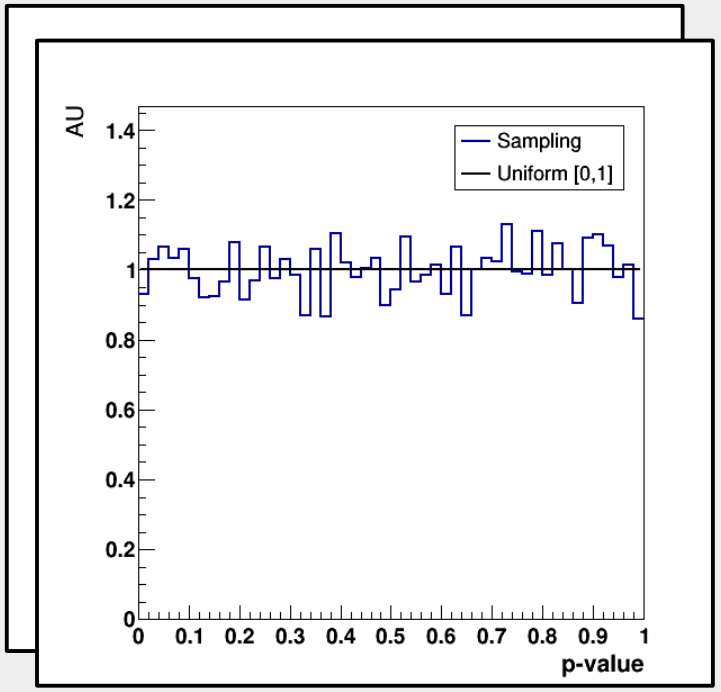
- Die Verteilungen der LS-Schätzungen $\hat{\theta}_0$ und $\hat{\theta}_1$ sind jeweils normalverteilt und unverzerrt.

- **NB**: Vergleichen Sie μ und σ mit Folie 15.

Zugrundeliegende Normalverteilung

- Hier sehen Sie die Ausgänge der Messung nach 10k-facher Wiederholung des Pseudo-experiments, wobei jede Einzelmessung nach $\varphi(x, y_i, \sigma_i)$ *normalverteilt* ist:





NB: AU steht für arbitary units

Der quadratische Abstand der LS-Schätzwerte von den wahren Werten

$$\chi^2_{\text{obs}} = \sum_{i \leq n} \left(\left(\hat{y}_i \left(\{x_i\}, \hat{\theta}_j \right) - y_i \right) / \sigma_{\hat{y}_i} \right)^2$$

ist nach $\chi^2(x, 3)$ verteilt.

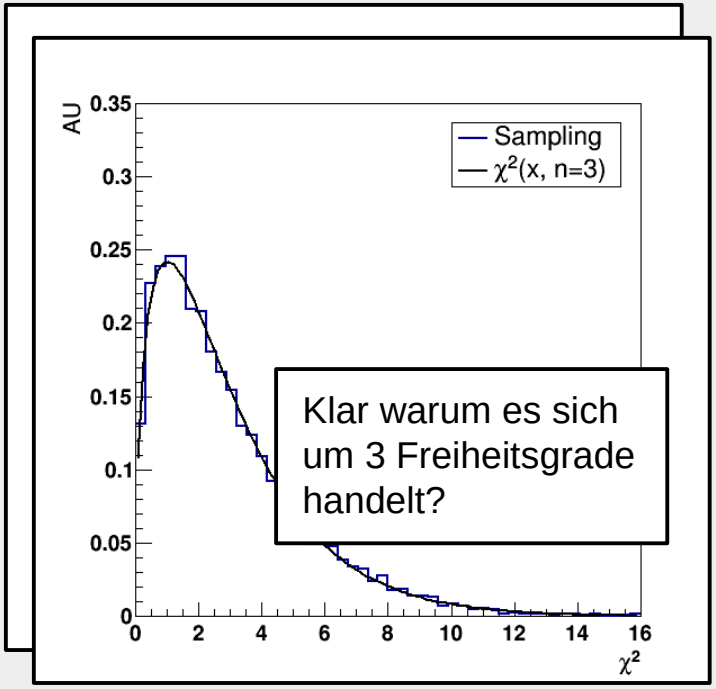
Der p-Wert

$$p = \int_{\chi^2_{\text{obs}}}^{\infty} \chi^2(x, 3)\, \mathrm{d}x$$

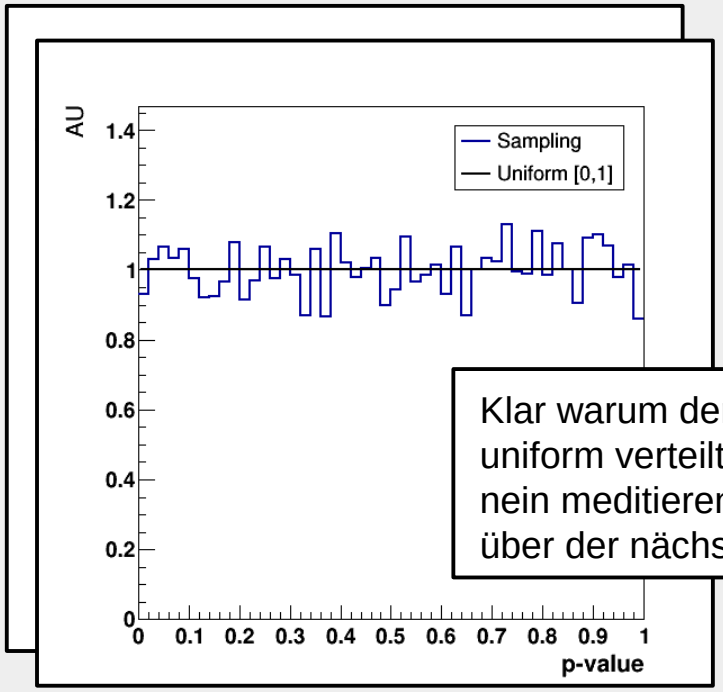
ist uniform verteilt.

Zugrundeliegende Normalverteilung

- Hier sehen Sie die Ausgänge der Messung nach 10k-facher Wiederholung des Pseudo-experiments, wobei jede Einzelmessung nach $\varphi(x, y_i, \sigma_i)$ *normalverteilt* ist:



Klar warum es sich um 3 Freiheitsgrade handelt?



Klar warum der p-Wert uniform verteilt ist? – Wenn nein meditieren Sie nochmal über der nächsten Folie.

NB: AU steht für arbitary units

Der quadratische Abstand der LS-Schätzwerte von den wahren Werten

$$\chi^2_{\text{obs}} = \sum_{i \leq n} \left(\left(\hat{y}_i \left(\{x_i\}, \hat{\theta}_j \right) - y_i \right) / \sigma_{\hat{y}_i} \right)^2$$

ist nach $\chi^2(x, 3)$ verteilt.

Der p-Wert

$$p = \int_{\chi^2_{\text{obs}}}^{\infty} \chi^2(x, 3)\, \mathrm{d}x$$

ist uniform verteilt.

Erinnerung p-Wert

- Wir erinnern hier nochmal durch indirekte Rechnung daran, dass der p-Wert einer zugrundeliegenden Verteilung gleichverteilt ist (hier am Beispiel der zugrundeliegenden $\chi^2(x,n)$ Verteilung):

$$P(\chi^2 \geq \chi^2_{\mathrm{obs}}) = \int_0^P 1 \,\mathrm{d}p' = \int_{\chi^2_{\mathrm{obs}}}^{\infty} \chi^2(x') \,\mathrm{d}x'$$

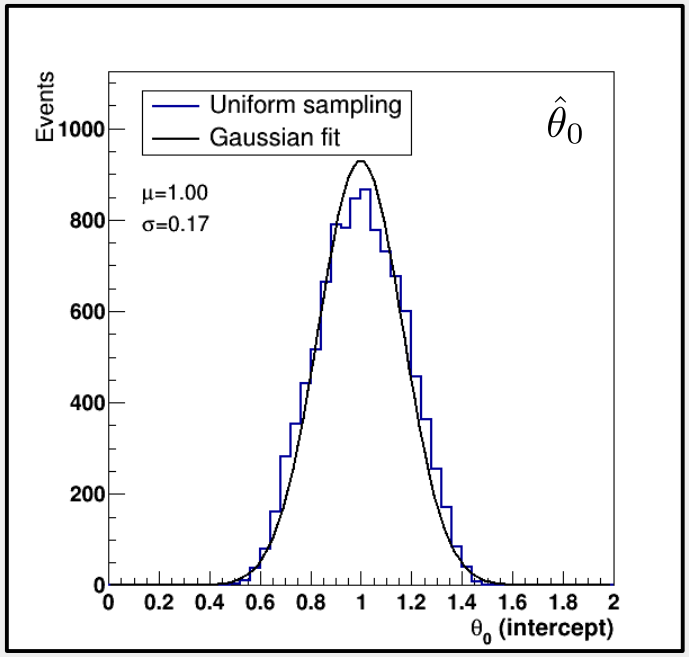
P gleichverteilt.

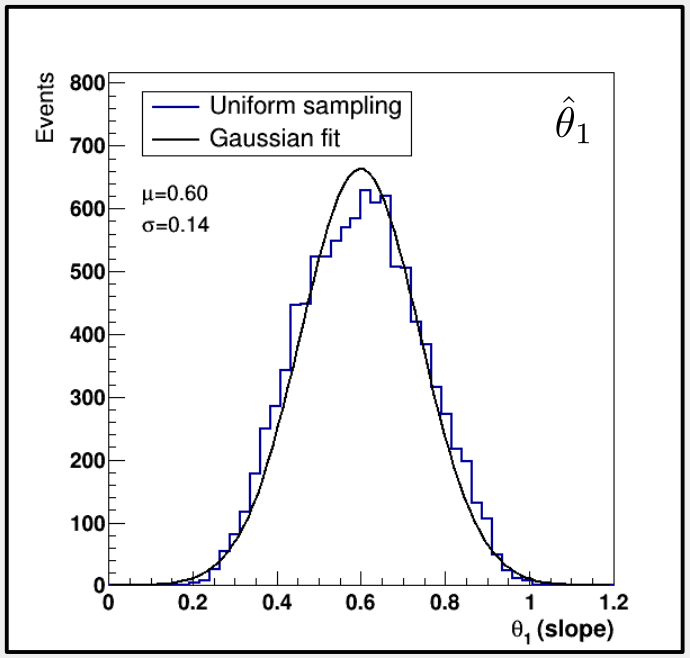
χ^2_{obs} wirklich nach $\chi^2(x,n)$ verteilt.

(Vergleichen Sie diese Rechnung mit VL-02 Folie 13)

Zugrundeliegende uniforme Verteilung

- Hier sehen Sie die Ausgänge der Messung nach 10k-facher Wiederholung des Pseudo-experiments, wobei jede Einzelmessung nach $U(x)$ *gleichverteilt* ist:



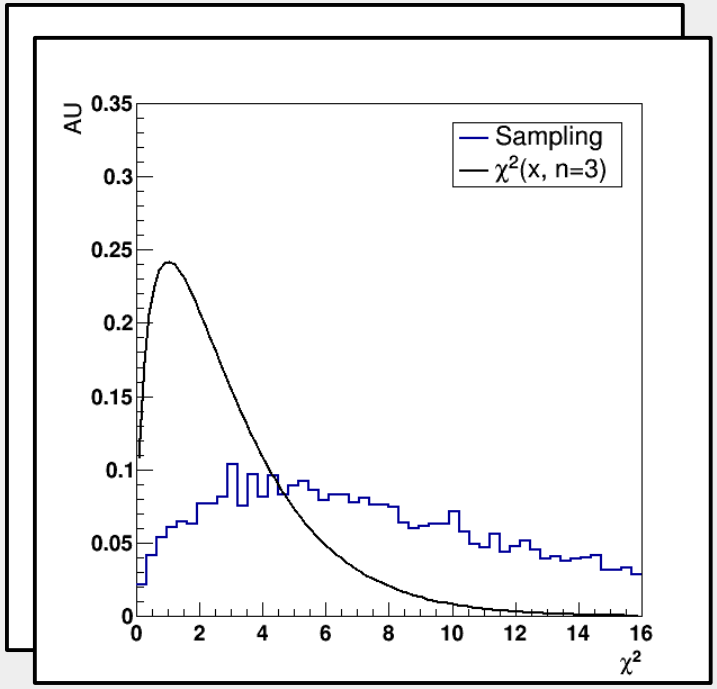


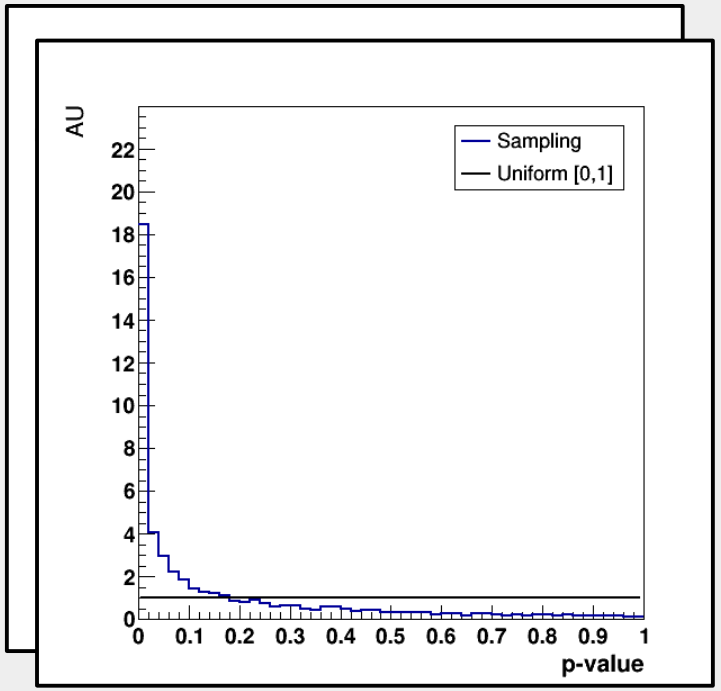
Ein lauffähiges RooT macro finden Sie hier.

- Die Verteilungen der LS-Schätzungen $\hat{\theta}_0$ und $\hat{\theta}_1$ sind jeweils normalverteilt und unverzerrt, aber nicht mehr effizient.

Zugrundeliegende uniforme Verteilung

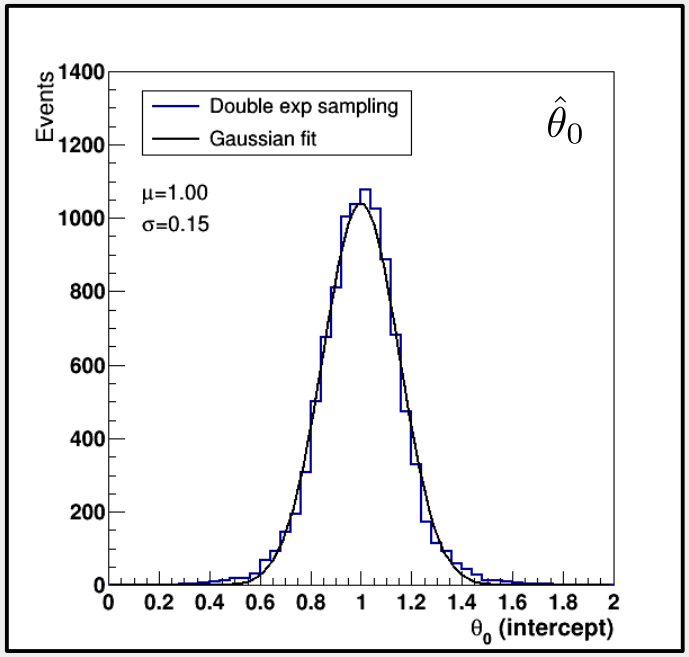
- Hier sehen Sie die Ausgänge der Messung nach 10k-facher Wiederholung des Pseudo-experiments, wobei jede Einzelmessung nach $U(x)$ *gleichverteilt* ist:

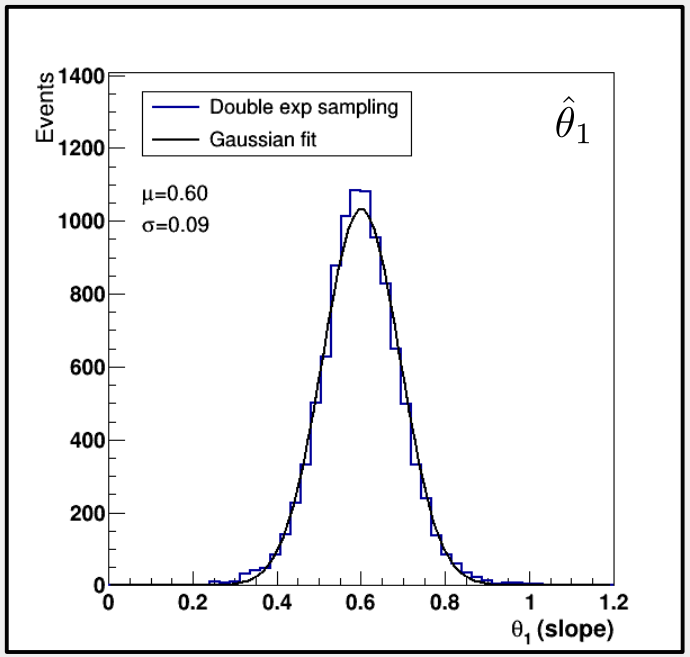




Zugrundeliegende exp Verteilung

- Hier sehen Sie die Ausgänge der Messung nach 10k-facher Wiederholung des Pseudo-experiments, wobei jede Einzelmessung beidseitig nach $\exp(x, \sigma_i/\sqrt{2})$ verteilt ist:



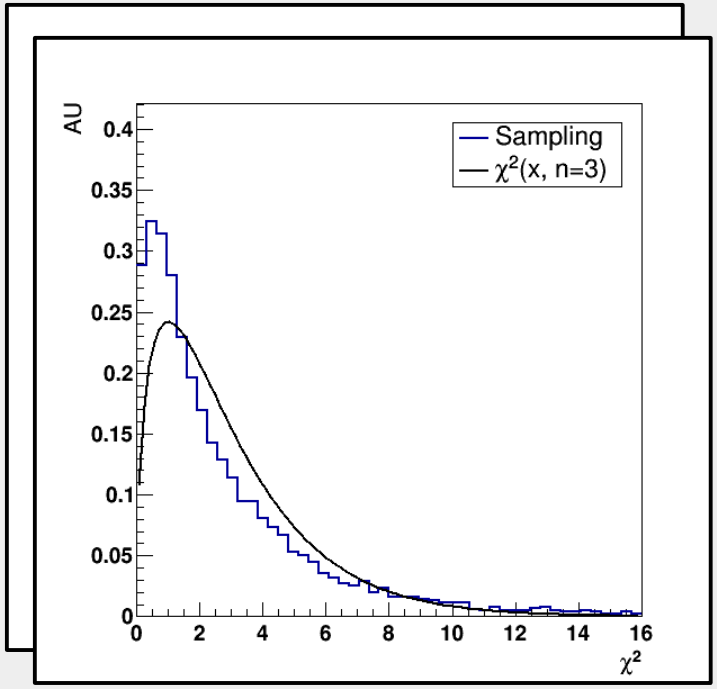


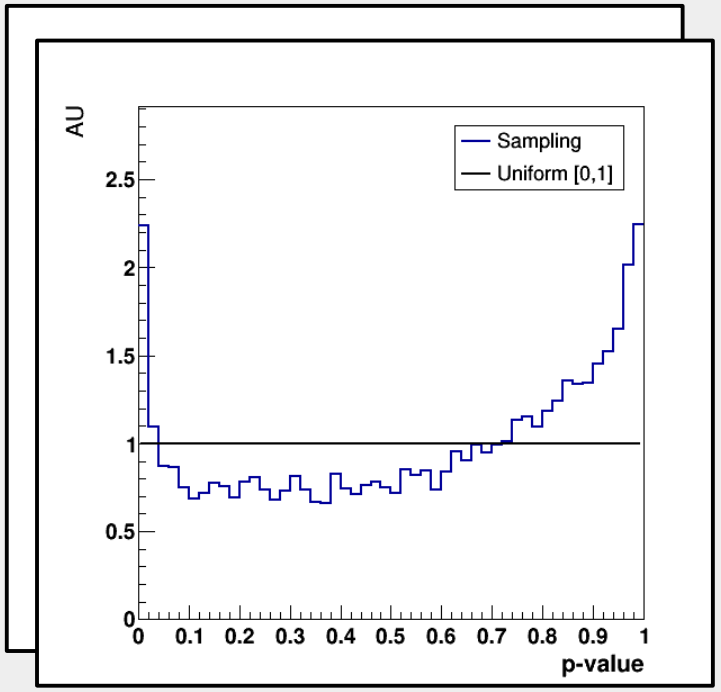
Ein lauffähiges RooT macro finden Sie hier.

- Die Verteilungen der LS-Schätzungen $\hat{\theta}_0$ und $\hat{\theta}_1$ sind jeweils normalverteilt und unverzerrt, aber nicht mehr effizient.

Zugrundeliegende exp Verteilung

- Hier sehen Sie die Ausgänge der Messung nach 10k-facher Wiederholung des Pseudo-experiments, wobei jede Einzelmessung beidseitig nach $\exp(x, \sigma_i/\sqrt{2})$ verteilt ist:





Verteilungsfreiheit – Zusammenfassung

- Unabhängig von der zugrundeliegenden Wahrscheinlichkeitsdichte für jede Einzelmessung sind die LS-Schätzwerte **normalverteilt** und (für symmetrische Wahrscheinlichkeitsdichten) **erwartungstreu**.

- Wie sich verschiedene zugrundeliegende Wahrscheinlichkeitsdichten äußern:
  - Die LS-Schätzung ist nicht mehr effizient (vgl. Folie 11).
  - χ^2_{obs} ist nicht nach $\chi^2(x,n)$ verteilt.
  - Der aus dem Integral der $\chi^2(x,n)$ Verteilung bestimmte p-Wert ist nicht uniform verteilt, wie dies der Fall wäre, wenn die LS-Schätzung einer $\chi^2(x,n)$ Verteilung folgen würde.

Pull-Verteilung

Für normalverteilte Messwerte folgt jeder LS-Schätzwert nach Standartisierung

$$Z_j = \frac{\hat{\theta}_j - E[\hat{\theta}_j]}{\sqrt{\mathrm{var}[\hat{\theta}_j]}} = \frac{\hat{\theta}_j - \theta_j}{\sqrt{\mathrm{var}[\hat{\theta}_j]}}$$

einer Standardnormalverteilung $\varphi(x, 0, 1)$.

- Man nennt die Verteilung der Werte für Z_j nach mehrfacher Durchführung von Pseudoexperimenten ***pull*-Verteilung**. Dabei werden bei bekannt vorausgesetzten Werten für θ_j die Werte von $\hat{\theta}_j$ und

  $$\sqrt{\mathrm{var}[\hat{\theta}_j]}$$

  für jeden Ausgang des Pseudoexperiments aus dem LS-Schätzwert und seiner Varianz bestimmt.

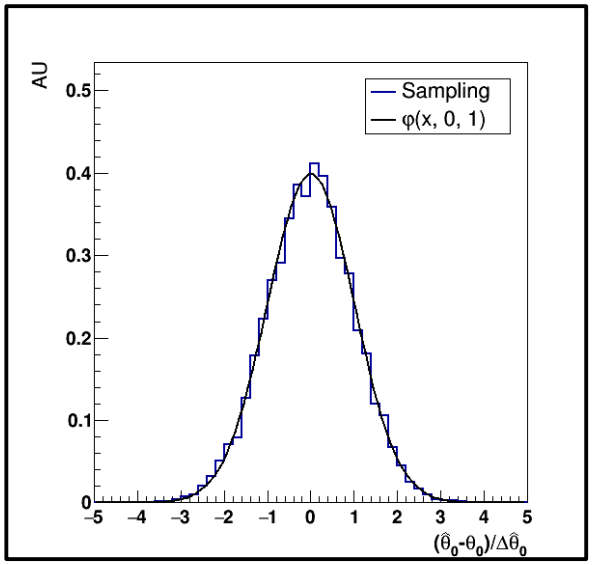
- Aus der *pull*-Verteilung wird z.B. üblicherweise die Erwartungstreue der Abschätzung abgeleitet.

Pull-Verteilung: Beispiel

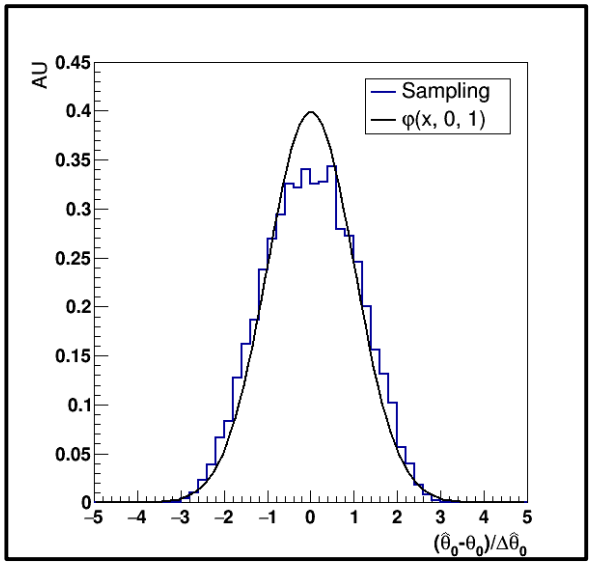
- Wir zeigen die *pull*-Verteilung anhand des Beispiels aus Abschnitt 4.2:

- Zugrundeliegende Wahrscheinlichkeitsdichte für Messwerte:

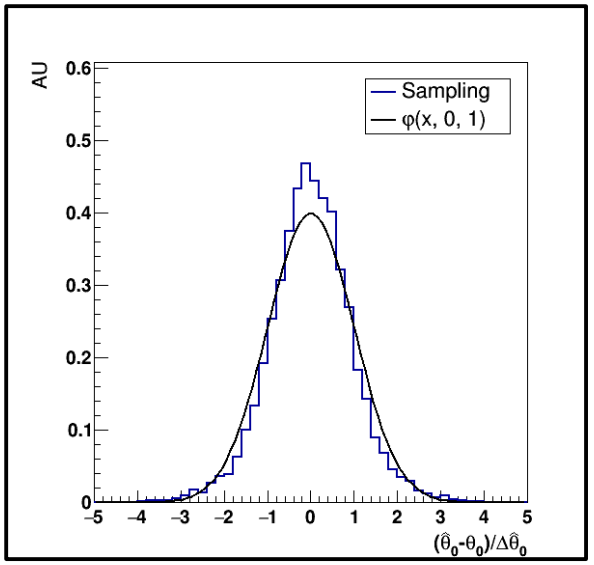
Normalverteilung:



Uniforme Verteilung:



Exponentielle Verteilung:



Ein lauffähiges RooT macro finden Sie hier.

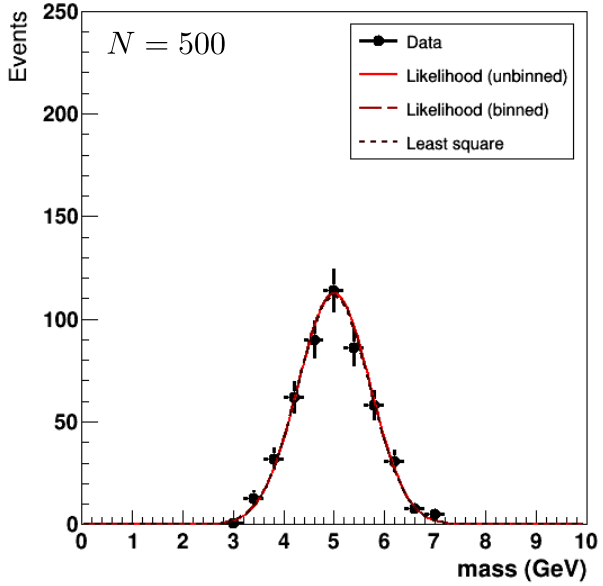
- Die LS-Schätzwerte sind zwar in allen Fällen erwartungstreu, ihre *pull*-Verteilungen sind jedoch nur im linken Fall standardnormalverteilt.

LS-Schätzung bei schwach populierten Histogramm-Bins

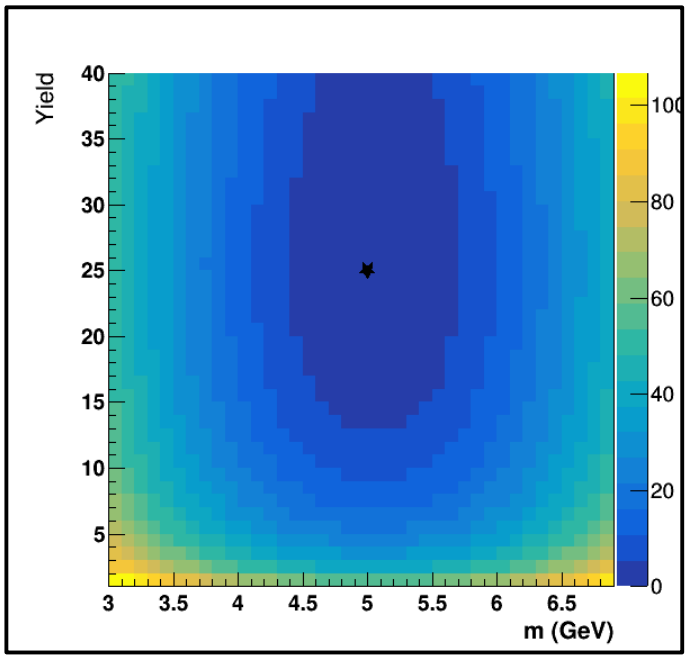
- Im Fall leerer oder schwach populierter Bins ist die LS-Schätzung, im Gegensatz zur ML-Schätzung, nicht mehr erwartungstreu. Wir demonstrieren dies anhand des folgenden Beispiels aus der Teilchenphysik:

Resonanz ohne Untergrund nach $\varphi(x, \mu = 5, \sigma = 1)$ verteilt. **N Ereignisse beobachtet**.

Abschätzung



Ein lauffähiges RooT macro finden Sie hier.



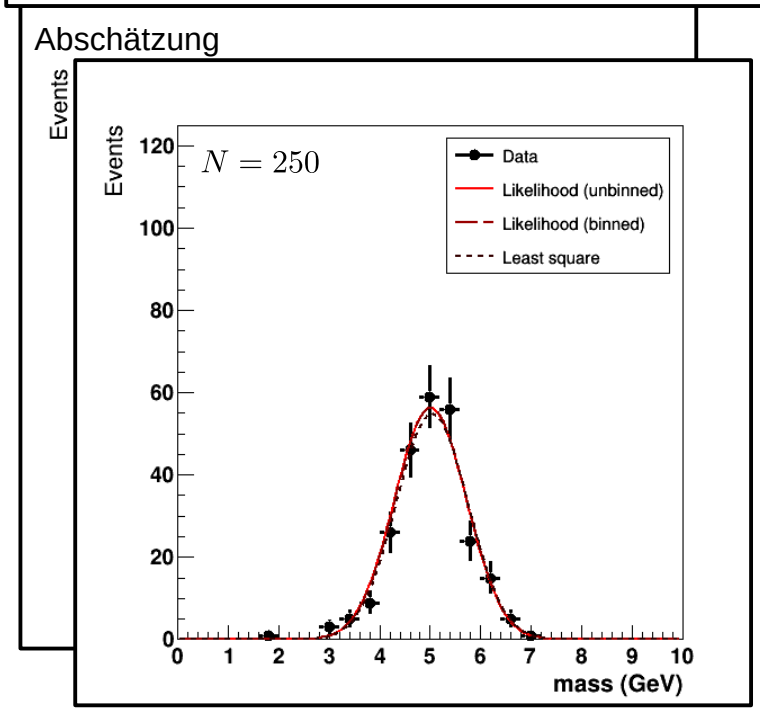
Ungebinnter Likelihood scan

LS-Schätzung bei schwach populierten Histogramm-Bins

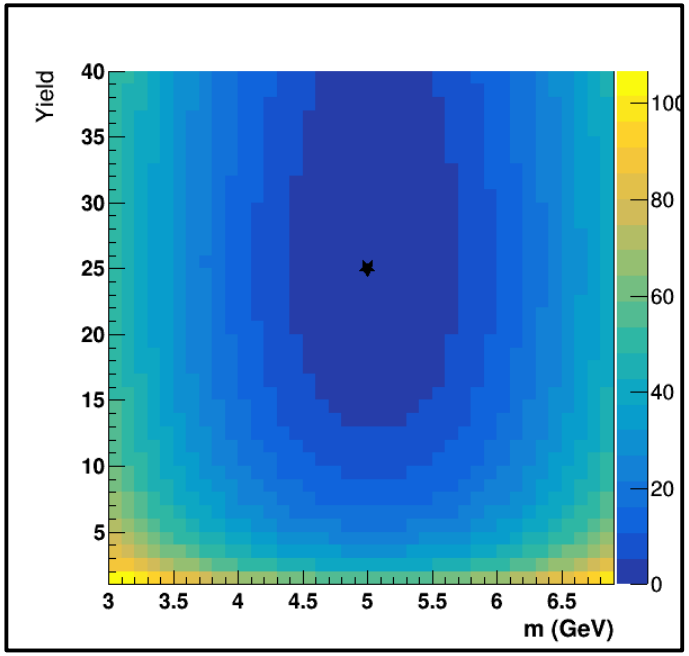
- Im Fall leerer oder schwach populierter Bins ist die LS-Schätzung, im Gegensatz zur ML-Schätzung, nicht mehr erwartungstreu. Wir demonstrieren dies anhand des folgenden Beispiels aus der Teilchenphysik:

Resonanz ohne Untergrund nach $\varphi(x, \mu = 5, \sigma = 1)$ verteilt. **N Ereignisse beobachtet**.

Abschätzung



Ein lauffähiges RooT macro finden Sie hier.



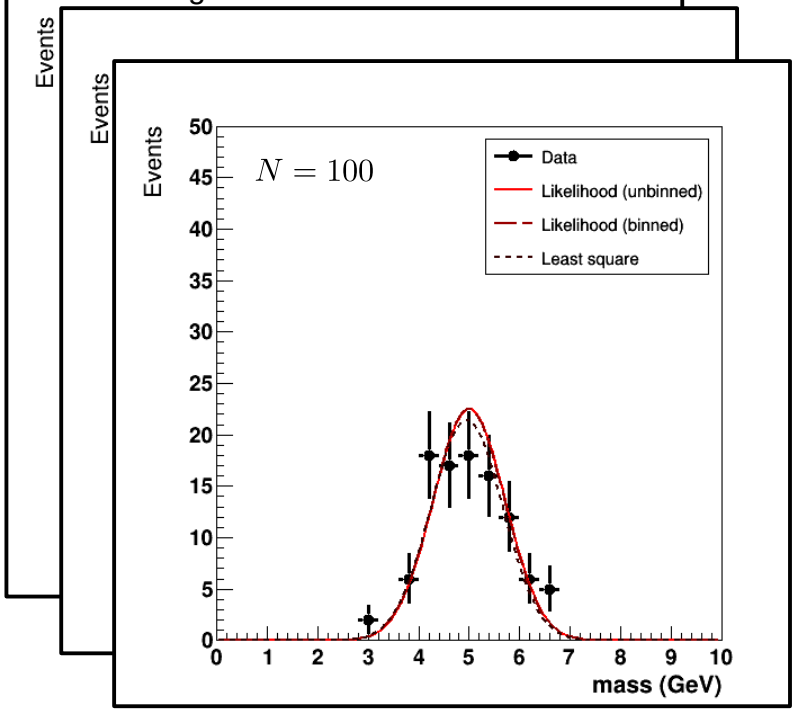
Ungebinnter Likelihood scan

LS-Schätzung bei schwach populierten Histogramm-Bins

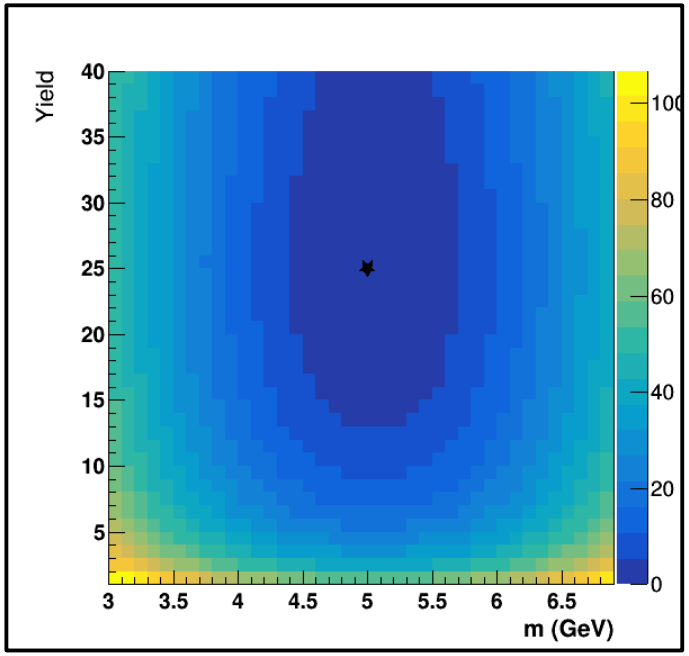
- Im Fall leerer oder schwach populierter Bins ist die LS-Schätzung, im Gegensatz zur ML-Schätzung, nicht mehr erwartungstreu. Wir demonstrieren dies anhand des folgenden Beispiels aus der Teilchenphysik:

Resonanz ohne Untergrund nach $\varphi(x, \mu = 5, \sigma = 1)$ verteilt. **N Ereignisse beobachtet**.

Abschätzung



Ein lauffähiges RooT macro finden Sie hier.



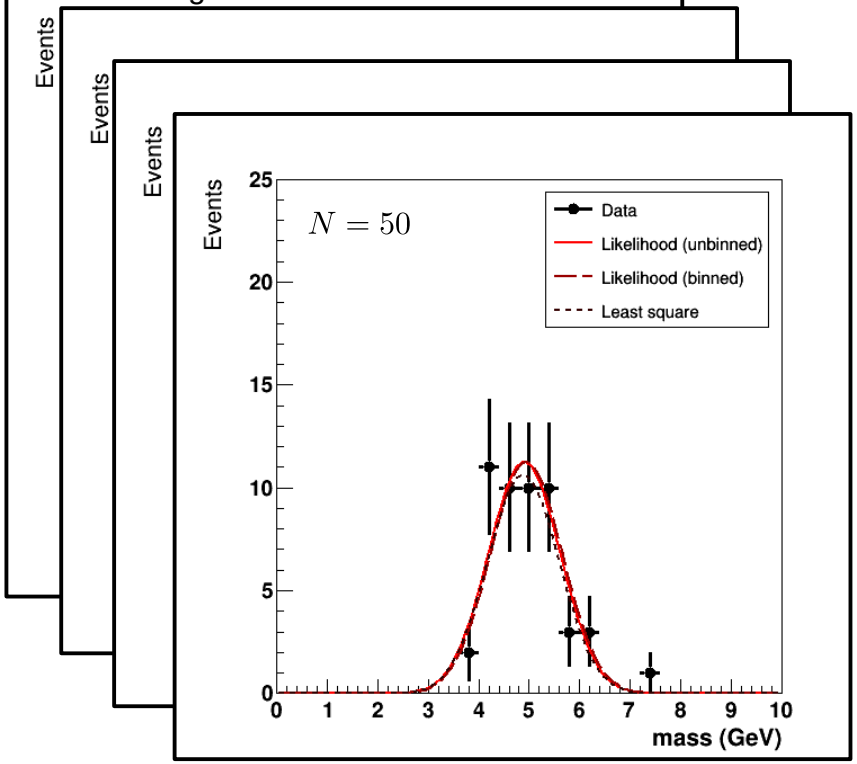
Ungebinnter Likelihood scan

LS-Schätzung bei schwach populierten Histogramm-Bins

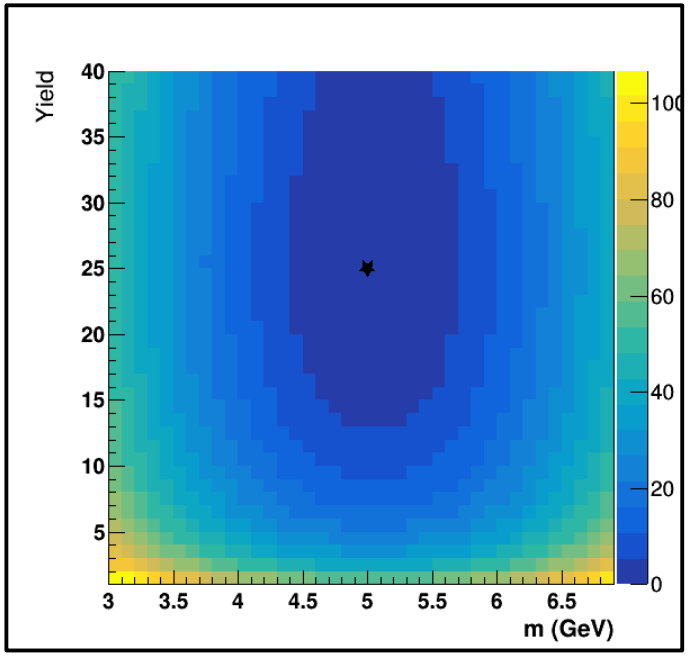
- Im Fall leerer oder schwach populierter Bins ist die LS-Schätzung, im Gegensatz zur ML-Schätzung, nicht mehr erwartungstreu. Wir demonstrieren dies anhand des folgenden Beispiels aus der Teilchenphysik:

Resonanz ohne Untergrund nach $\varphi(x, \mu = 5, \sigma = 1)$ verteilt. **N Ereignisse beobachtet**.

Abschätzung



Ein lauffähiges RooT macro finden Sie hier.



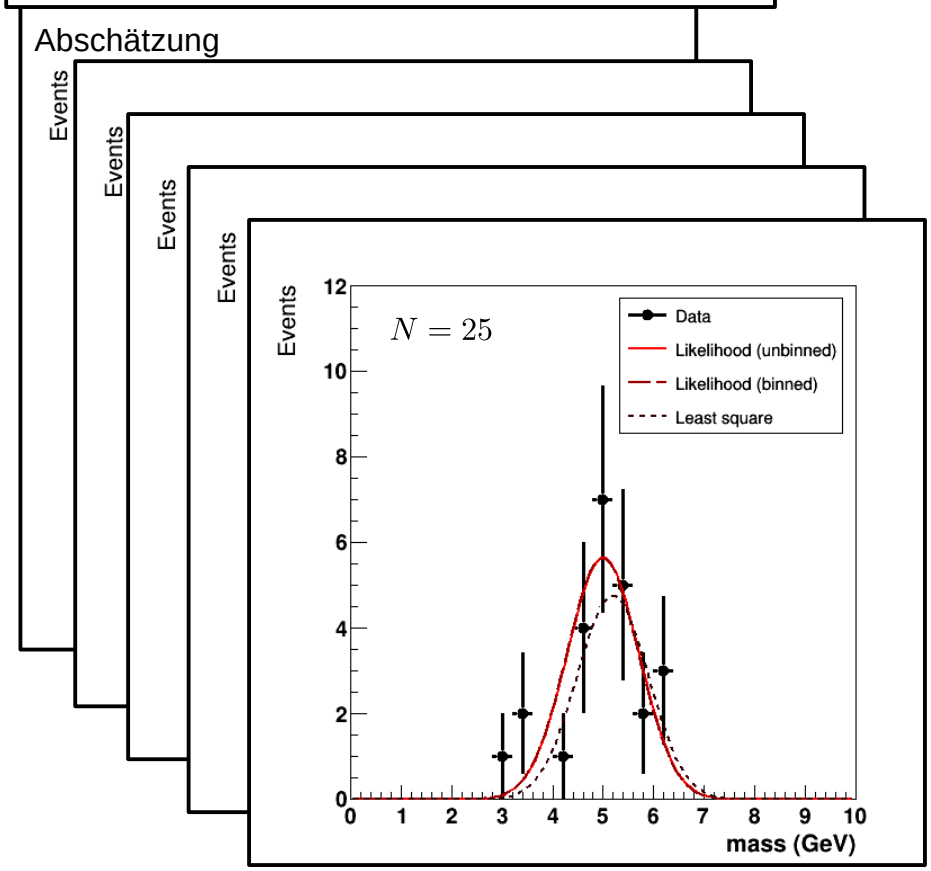
Ungebinnter Likelihood scan

LS-Schätzung bei schwach populierten Histogramm-Bins

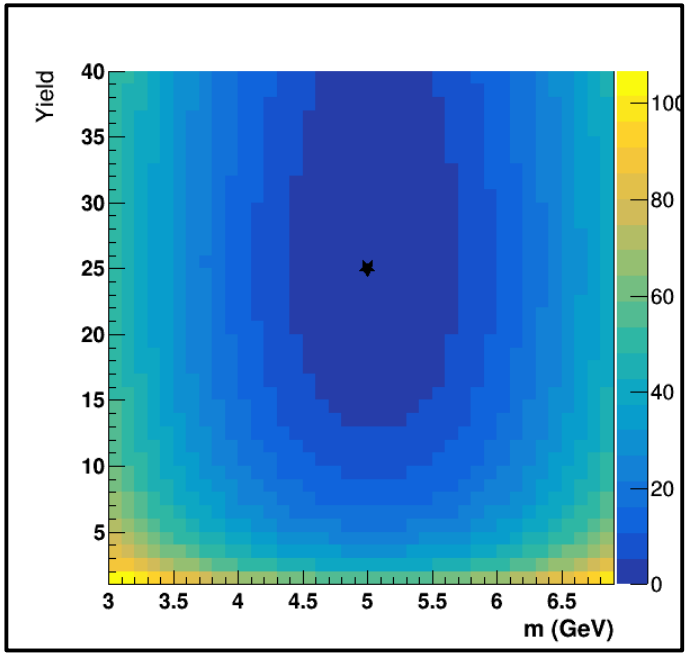
- Im Fall leerer oder schwach populierter Bins ist die LS-Schätzung, im Gegensatz zur ML-Schätzung, nicht mehr erwartungstreu. Wir demonstrieren dies anhand des folgenden Beispiels aus der Teilchenphysik:

Resonanz ohne Untergrund nach $\varphi(x, \mu = 5, \sigma = 1)$ verteilt. **N Ereignisse beobachtet**.

Abschätzung



Ein lauffähiges RooT macro finden Sie hier.



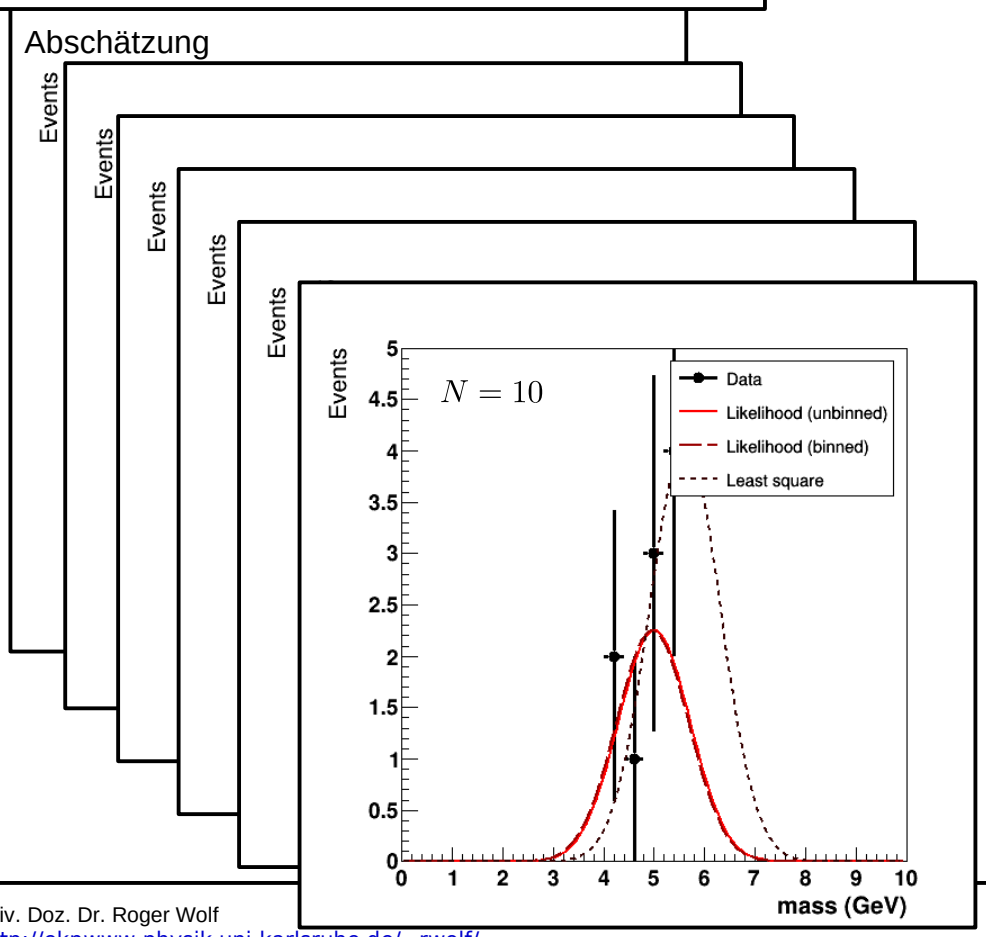
Ungebinnter Likelihood scan

LS-Schätzung bei schwach populierten Histogramm-Bins

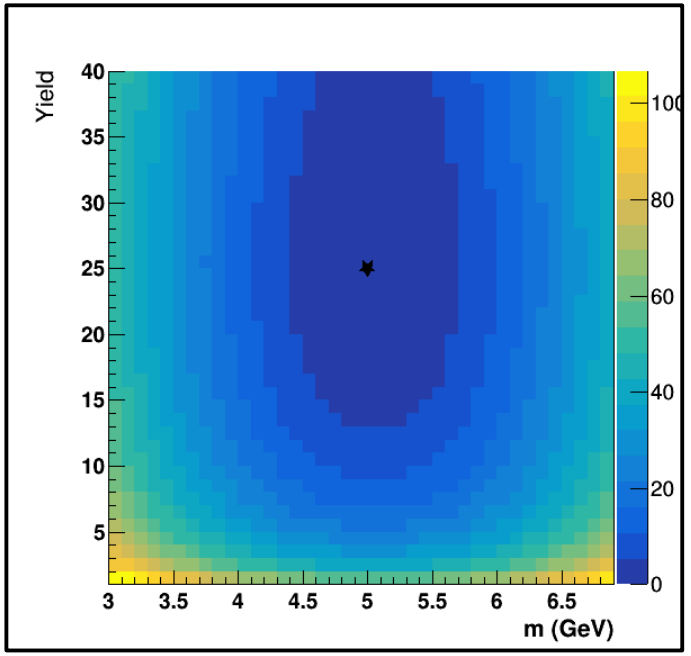
- Im Fall leerer oder schwach populierter Bins ist die LS-Schätzung, im Gegensatz zur ML-Schätzung, nicht mehr erwartungstreu. Wir demonstrieren dies anhand des folgenden Beispiels aus der Teilchenphysik:

Resonanz ohne Untergrund nach $\varphi(x, \mu = 5, \sigma = 1)$ verteilt. **N Ereignisse beobachtet**.

Abschätzung



Ein lauffähiges RooT macro finden Sie hier.



Ungebinnter Likelihood scan

LS- vs. ML-Schätzung

- Nutzen Sie die **ML-Schätzung**, wann immer möglich:
  - Sie setzt die Hypothese einer parametrisierten zugrundeliegenden Wahrscheinlichkeitsdichte voraus. Die Parameter werden maximiert.
  - Sie ist immer erwartungstreu.
  - Wenn es eine effiziente Schätzung des gesuchten Parameters gibt, ist die ML-Schätzung effizient, d.h. sie hat die geringste Varianz.
  - Die ML-Schätzung ist i.A. nicht analytisch lösbar. Heutzutage würden Sie in ernsthaften Studien zur Lösung statistischer Probleme ohnehin auf MC Methoden zurückgreifen.

LS- vs. ML-Schätzung – continued –

- Die **LS-Schätzung** benötigt weniger Voraussetzungen von Ihrer Seite: Nur die Varianz der Hypothese muss bekannt sein, die hypothetische zugrundeliegende Wahrscheinlichkeitsdichte ist die Normalverteilung:
  - Die LS-Schätzung bezieht ihre große Bedeutung daraus, dass sie für die große Klasse linearer Probleme analytisch lösbar ist.
  - Für eine größere Klasse an wohldefinierten Problemen (z.B. F-Test, T-Test, χ^2-Test) ist die Lösung ebenfalls bekannt.
  - Für normalverteilte Zufallsgrößen ist die LS Schätzung zur ML Schätzung äquivalent.
  - Für nicht-normalverteilte Zufallsgrößen und insbesondere für asymmetrische Unsicherheiten führen die LS und die ML Schätzung nicht zum gleichen Ergebnis.

Zusammenfassung

- Übergang von Maximum Likelihood zu LS-Schätzung.
- Lineare LS-Schätzung.
- Ausgewählte Eigenschaften der LS-Schätzung.

5 Optimierungsalgorithmen

5.1 Einführung und einfache Verfahren

Wir geben eine kurze Einführung und Übersicht über verschiedene Optimierungsalgorithmen.

Bedeutung von Optimierungsalgorithmen

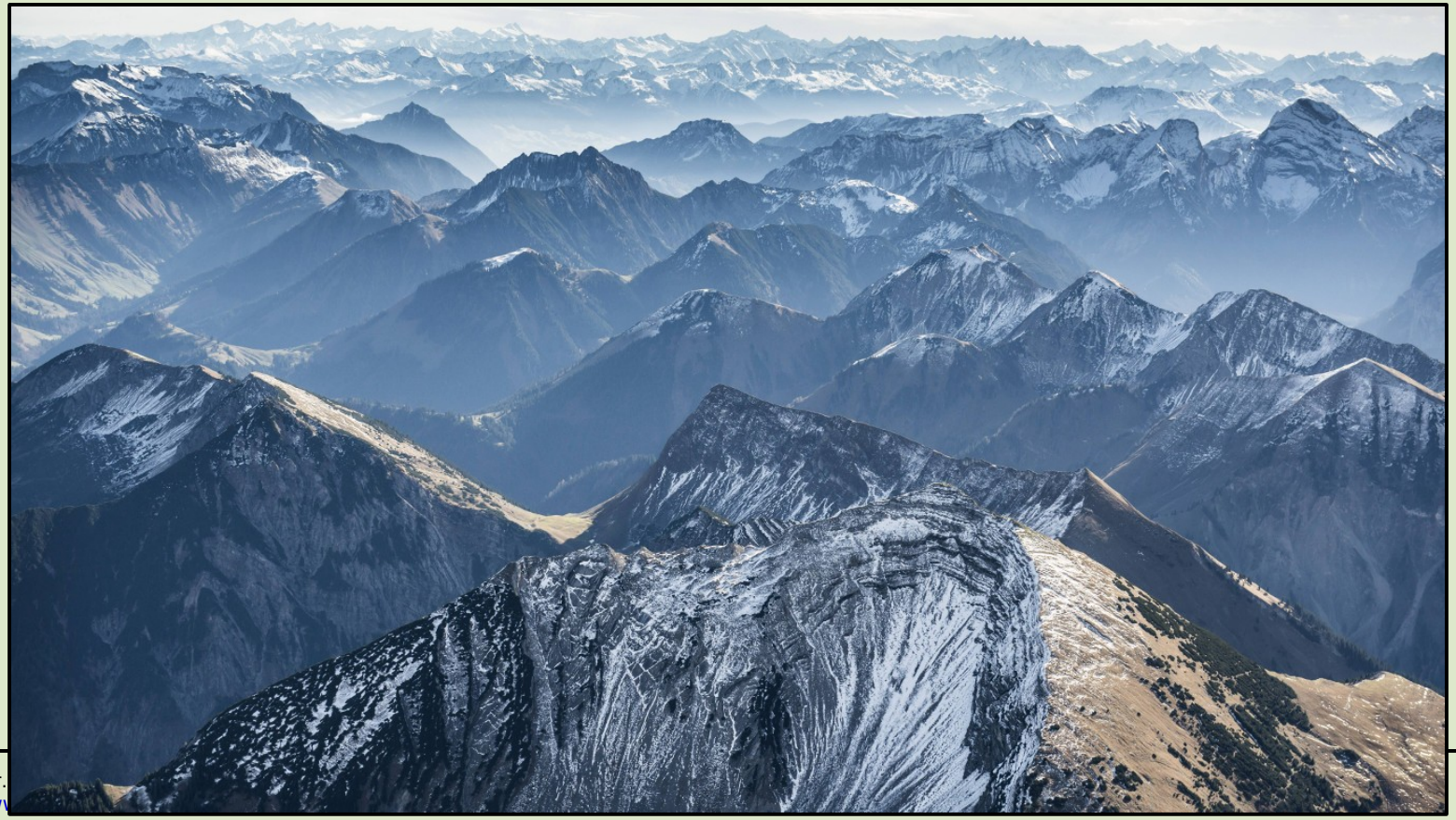
- Für einfache, stetige und differenzierbare Funktionen kennen Sie Optimierungsaufgaben aus Schule und Studium.

- Für allgemeine hochdimensionale Funktionen mit Sattelpunkten und Nebenextrema spielen **effiziente und sichere numerische Optimierungsalgorithmen** eine zentrale Rolle.

Bedeutung von Optimierungsalgorithmen

- Für einfache, stetige und differenzierbare Funktionen kennen Sie Optimierungsaufgaben aus Schule und Studium.

- Für allgemeine hochdimensionale Funktionen mit Sattelpunkten und Nebenextrema spielen **effiziente und sichere numerische Optimierungsalgorithmen** eine zentrale Rolle.



Bedeutung von Optimierungsalgorithmen

- Für einfache, stetige und differenzierbare Funktionen kennen Sie Optimierungsaufgaben aus Schule und Studium.

- Für allgemeine hochdimensionale Funktionen mit Sattelpunkten und Nebenextrema spielen **effiziente und sichere numerische Optimierungsalgorithmen** eine zentrale Rolle.

- Man kann Optimierungsalgorithmen grob einteilen nach:

  - Einfachen Verfahren, die auf den Funktionswerten selbst basieren (→ Rastersuchen).

  - Abstiegsverfahren, die auf Ableitungen basieren (→ Gradientenabstiegsverfahren, Newtonverfahren, …).

- Wir werden Ihnen an dieser Stelle einige Beispiele aus jeder der angegebenen Klassen vorstellen. Dabei formulieren wir das Optimierungsproblem als **Minimierungsproblem**.

Konvergenzrate

Sei $F(x)$ eine beliebige Funktion und $x_{\min}$ ein Minimum von $F(x)$. Für einen Algorithmus $\mathcal{A}$ mit der Aufgabe in einer Folge von Werten $\{x_k\}$ von einem beliebigen Startpunkt x_0 aus $x_{\min}$ zu finden, bezeichnen wir

$$R = \frac{F(x_{k+1}) - F(x_{\min})}{F(x_k) - F(x_{\min})}$$

als Konvergenzrate von $\mathcal{A}$. Für

$$\limsup_{k \to \infty} R < 1$$

bezeichnet man $\mathcal{A}$ als linear konvergent.

Konvergenzrate

Sei $F(x)$ eine beliebige Funktion und $x_{\min}$ ein Minimum von $F(x)$. Für einen Algorithmus $\mathcal{A}$ mit der Aufgabe in einer Folge von Werten $\{x_k\}$ von einem beliebigen Startpunkt x_0 aus $x_{\min}$ zu finden, bezeichnen wir

$$R = \frac{F(x_{k+1}) - F(x_{\min})}{F(x_k) - F(x_{\min})}$$

als Konvergenzrate von $\mathcal{A}$. Für

$$\limsup_{k\to\infty} R < 1$$

bezeichnet man $\mathcal{A}$ als linear konvergent.

- Ist ein Algorithmus linear konvergent bedeutet das, dass er tatsächlich exponentiell schnell auf das Minimum konvergiert:

$$|F(x_{k+1}) - F(x_{\min})| = c^k \, |F(x_0) - F(x_{\min})| \; , \quad c < 1$$

Rosenbrock-Funktion

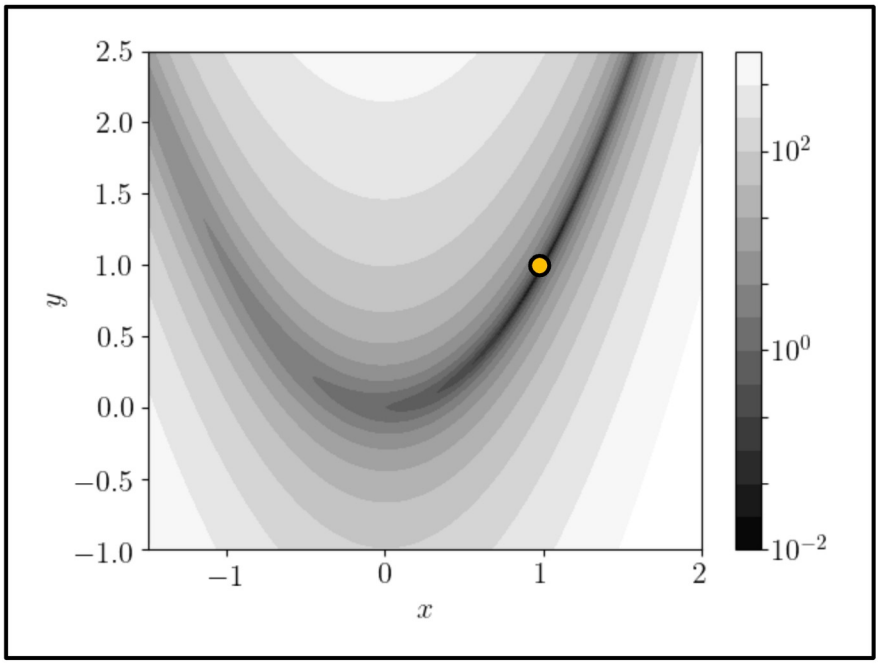
- Die Rosenbrock-Funktion wurde erstmals 1960 von Howard H. Rosenbrock als *benchmark* für Optimierungsalgorithmen eingeführt:

$$f(x,y) = (a-x)^2 + b\left(y-x^2\right)^2$$

- Die Funktion hat ein globales Minimum bei $(x,y) = (a,a^2)$, an dem sie den Wert $f(a,a^2) = 0$ annimmt.

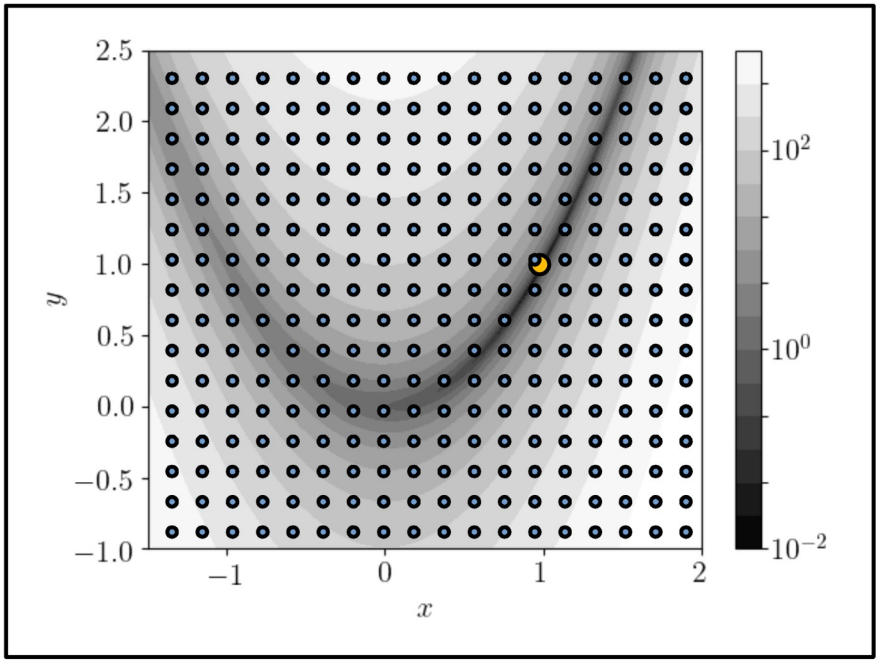
- Das Minimum liegt jedoch in einem langezogenen parabolischen Tal und ist daher numerisch schwer zu ermitteln.

- Rechts ist die Funktion für $a = 1,\ b = 100$ gezeigt.



Rasterverfahren

- Die einfachste Methode das Minimum zu ermitteln, besteht in einem *brute force* Rasterverfahren (engl. *grid search*, *range search*):
  - Auswertung der Funktion an k gleichverteilten Stützstellen. Das Raster kann ggf. sukzessive vereinfert werden.
  - In d Dimensionen k^d Rasterpunkte $\rightarrow$ für höherdimensionale Funktionen ungeeignet.



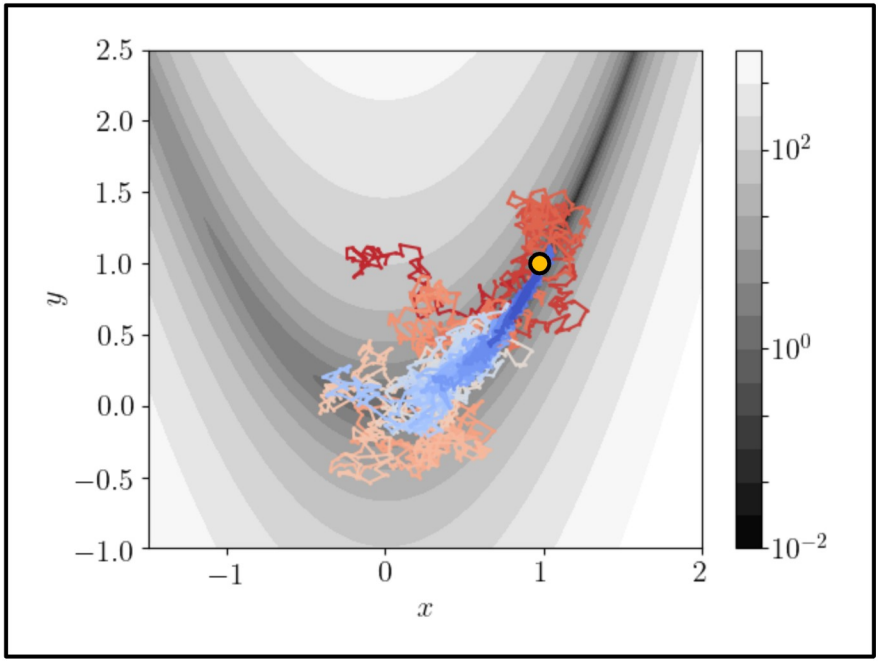
Simuliertes Abkühlen (engl. *simulated annealing*)

- Hierbei handelt es sich um ein Monte Carlo Verfahren zur Optimierung:
  - Vergleiche $f(x_i, y_i)$ mit $f(x_{i+1}, y_{i+1})$ an einem zufällig gewählten Punkt (x_{i+1}, y_{i+1}).
  - Ersetze (x_i, y_i) durch (x_{i+1}, y_{i+1}) wenn:
    - $f(x_{i+1}, y_{i+1}) \leq f(x_i, y_i)$.
    - Mit der Wahrscheinlichkeit

      $$\exp\left(-\frac{f(x_{i+1}, y_{i+1}) - f(x_i, y_i)}{T}\right)$$

      sonst.
    - Senke die „Temperatur“ T und wiederhole das Verfahren.

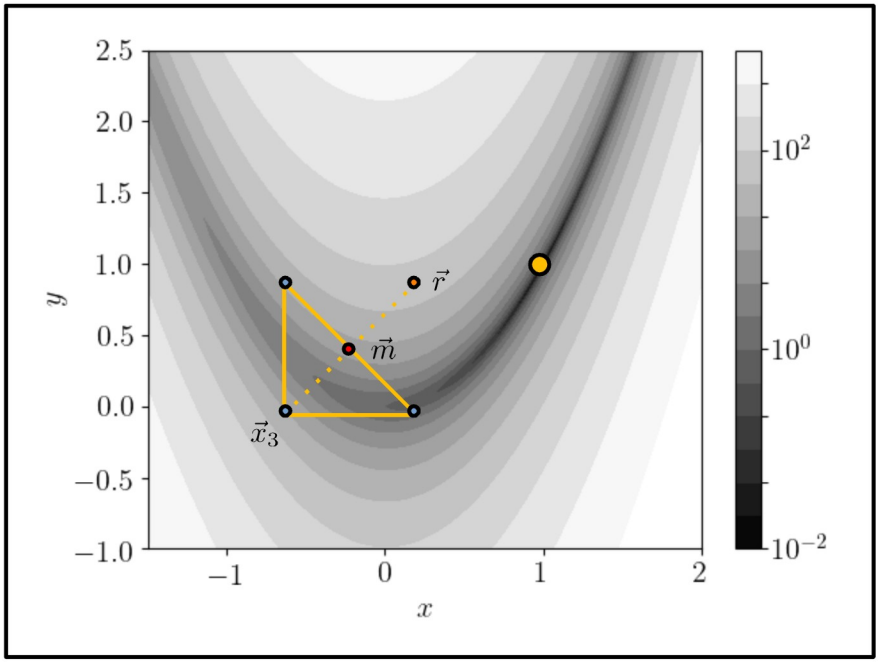
- Dadruch, dass der Algorithmus auch die Möglichkeit hat „nach oben zu steigen“, ist es möglich sich aus Nebenminima „heraus“ zu bewegen.



Minimum

Simplex Verfahren

- Eingeführt von John Nelder und Roger Mead 1975:
  - Starte mit einem Simplex aus $d+1$ Punkten in d Dimensionen; Werte die zu minimiernde Funktion an jedem Eckpunkt aus und sortiere aufsteigend nach Funktionswerten.
    $F(\vec{x}_1) < F(\vec{x}_2) < \ldots < F(\vec{x}_{d+1})$
  - Ermittele den Schwerpunkt der ersten d Punkte: $\vec{m} = \frac{1}{d}\sum_{i=1}^{d} \vec{x}_i \quad (*)$
  - Reflektiere $\vec{x}_{d+1}$ an $\vec{m}$:
    $\vec{r} = \vec{m} + \alpha\,(\vec{m} - \vec{x}_{d+1})$
  - Unterscheide mehrere Fälle:
    (1) $F(\vec{x}_1) \leq F(\vec{r}) \leq F(\vec{x}_d)$
    (2) $F(\vec{r}) \; < F(\vec{x}_1)$
    (3) $F(\vec{x}_d) < F(\vec{r})$
    (4) Ansonsten Kontraktion um $\vec{x}_1$

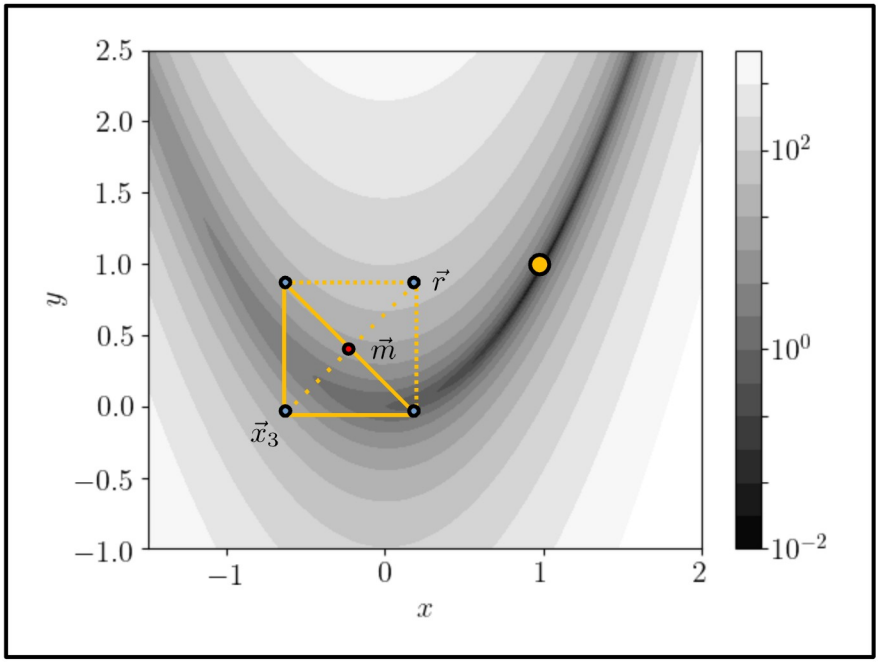


Minimum

Simplex Verfahren

- **Fallunterscheidung**:

(1) $F(\vec{x}_1) < F(\vec{r}) < F(\vec{x}_d)$:
Ersetze $\vec{x}_{d+1}$ durch $\vec{r}$; zurück zu (*).

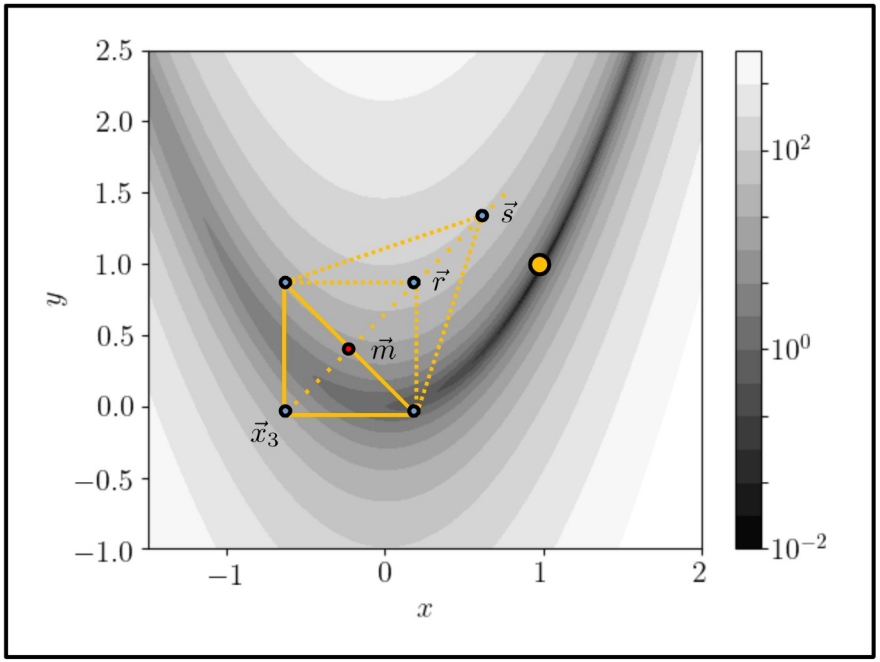


Minimum

Simplex Verfahren

- **Fallunterscheidung**:

(2) $F(\vec{r}) < F(\vec{x}_1)$: Gute Richtung $\rightarrow$ Streckung $(\vec{s} = \vec{m} + \beta\,(\vec{m} - \vec{x}_{d+1}))$
Ersetze $\vec{x}_{d+1}$ durch $\vec{s}$ wenn $F(\vec{s}) < F(\vec{r})$ und durch $\vec{r}$ sonst, zurück zu (*).

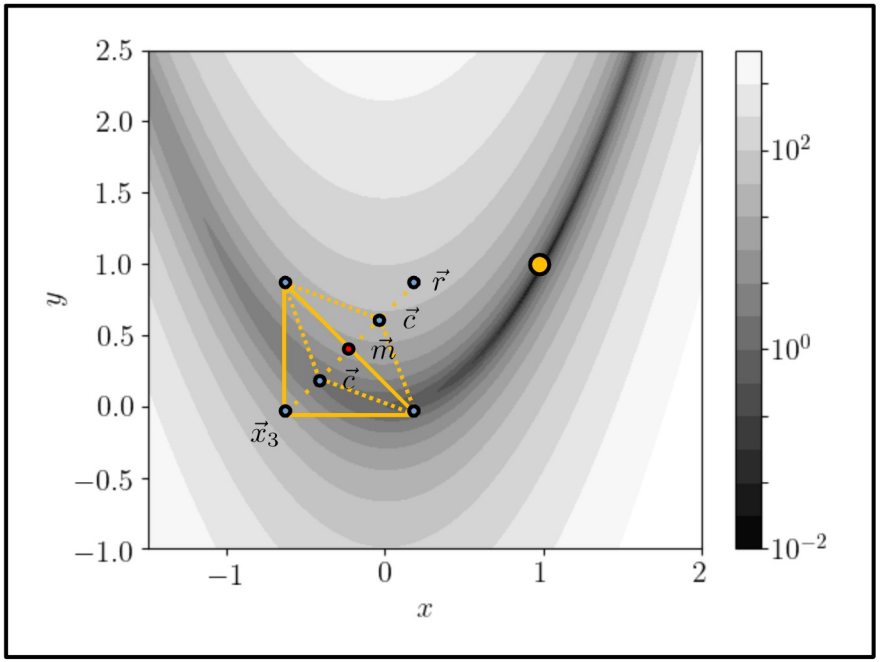


Minimum

Simplex Verfahren

- **Fallunterscheidung**:

(3) $F(\vec{x}_d) < F(\vec{r})$: Schlechte Richtung $\rightarrow$ Abflachung $(\vec{c} = \vec{m} + \gamma(\vec{m} - \vec{h}))$,
wobei $\vec{h} = \vec{x}_{d+1}$ wenn $F(\vec{x}_{d+1}) < F(\vec{r})$ und $\vec{r}$ sonst.
Ersetze $\vec{x}_{d+1}$ durch $\vec{c}$ wenn $F(\vec{c}) < F(\vec{x}_{d+1})$, zurück zu (*).



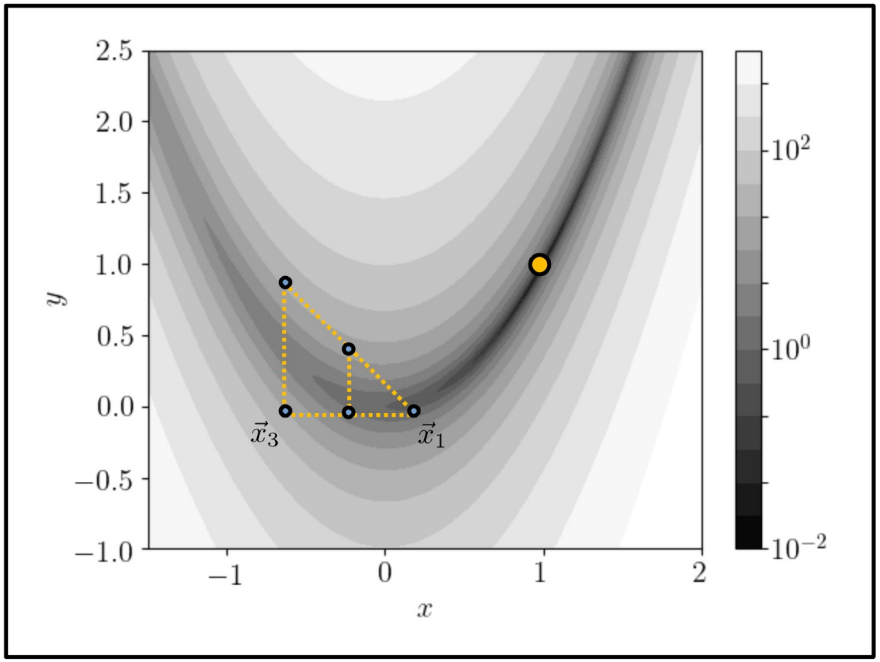
Minimum

Simplex Verfahren

- **Fallunterscheidung**:

(4) Ansonsten komprimiere den Simplex um $\vec{x}_1$

$\vec{x}_i = \vec{x}_1 + \sigma(\vec{x}_1 - \vec{x}_i) \quad \forall i = 2, \ldots d+1\,,$ zurück zu (*).



Minimum

Simplex Verfahren – Anmerkungen

- **Anmerkungen**:
  - Bei dem Verfahren ist darauf zu achten, dass die Startwerte linear unabhängig sind, so dass sich ein richtiger Simplex ergibt.
  - Typische Werte für die Parameter des Algorithmus sind:
    $\alpha = 1 \qquad \beta = 2 \qquad \gamma = {}^1\!/_2 \qquad \sigma = {}^1\!/_2$
  - Allgemein muss gelten:
    $0 < \alpha < \beta \qquad \gamma, \sigma \in (0, 1)$
- Der Algorithmus bewegt dann den Simplex auf ein (lokales) Minimum zu und schrumpft ihn um dieses Minimum zusammen, bis ein Abbruchkriterium erreicht wird.
- Eine beispielhafte Implementierung können Sie unter diesem link finden.

5 Optimierungsalgorithmen

5.2 Gradientenabstiegsverfahren

Gradientenverfahren machen neben den Funktionswerten selbst von den Ableitungen der zu minimierenden Funktion Gebrauch.

Gradientenabstiegsverfahren

- Das einfache Gradientenabstiegsverfahren hat die folgende Aktualisierungsregel:

$$x_{k+1} = x_k - \eta \frac{\mathrm{d}F(x_k)}{\mathrm{d}x}$$

$$\text{solange:}$$

$$|F(x_k) - F(x_{k-1})| > \epsilon$$

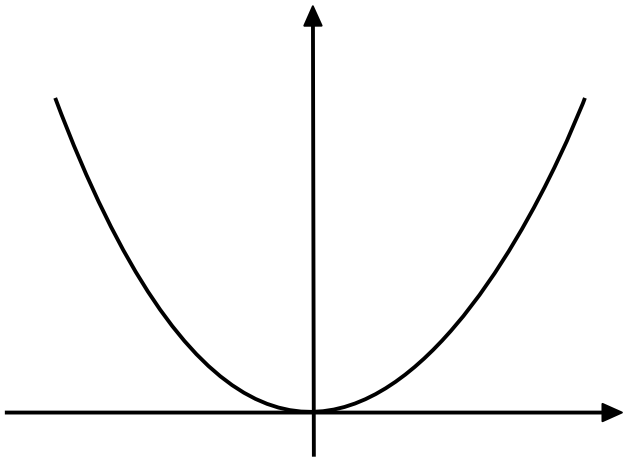
- Man bezeichnet $\eta \in \mathbb{R}$ als die **Schrittweite**.

Beispiel: Optimale Schrittweite

- Zur Diskussion der optimalen Schrittweite betrachten wir ein Polynom zweiter Ordnung:

$$F(x) = \frac{1}{2}ax^2 + bx + c$$

- Wie würden Sie die Schrittweite η wählen, um das Minimum der Funktion möglichst schnell zu erreichen und nach wieviel Schritten ist dies der Fall?

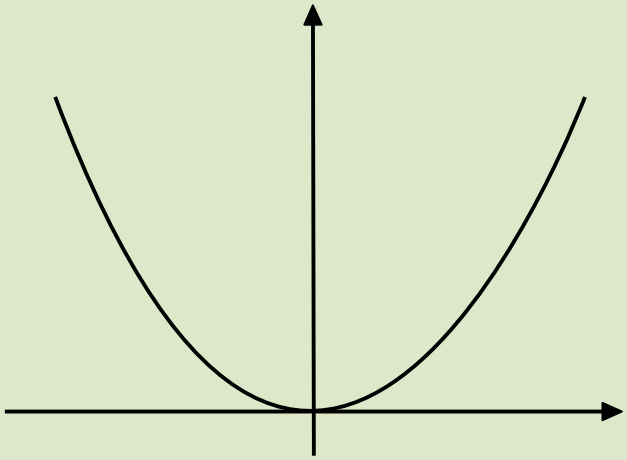


Beispiel: Optimale Schrittweite

- Zur Diskussion der optimalen Schrittweite betrachten wir ein Polynom zweiter Ordnung:

$$F(x) = \frac{1}{2}ax^2 + bx + c$$

- Wie würden Sie die Schrittweite η wählen, um das Minimum der Funktion möglichst schnell zu erreichen und nach wieviel Schritten ist dies der Fall?

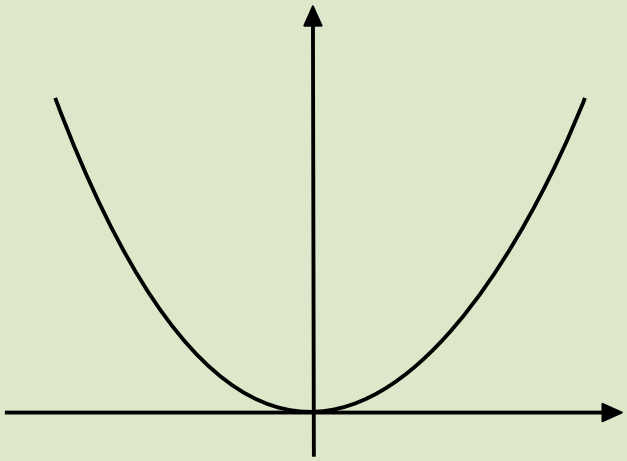


Beispiel: Optimale Schrittweite

- Zur Diskussion der optimalen Schrittweite betrachten wir ein Polynom zweiter Ordnung:

$$F(x) = \frac{1}{2}ax^2 + bx + c$$

- Wie würden Sie die Schrittweite η wählen, um das Minimum der Funktion möglichst schnell zu erreichen und nach wieviel Schritten ist dies der Fall?



- Taylorreihe um einen beliebigen Wert x_0:

$$F\,(x) = F(x_0) + F'(x_0)\,(x - x_0) + \frac{1}{2}F''(x_0)(x_0)\,(x - x_0)^2$$

$$F'(x) = F'(x_0) + F''(x_0)\,(x - x_0) = 0$$

$$x_{\min} = x_0 - \frac{F'(x_0)}{F''(x_0)}$$

D.h. Sie erreichen das Minimum einer Parabel immer mit einem Schritt, egal von wo Sie die Minimierung starten.

Vergleich mit Aktualisierungsregel:

$$x_1 = x_0 - \eta F'(x_0) \qquad \Rightarrow \qquad \eta_{\text{opt}} = \frac{1}{F''(x_0)} \quad \text{mit: } F''(x) = a \quad (\forall x \in \mathbb{R})$$

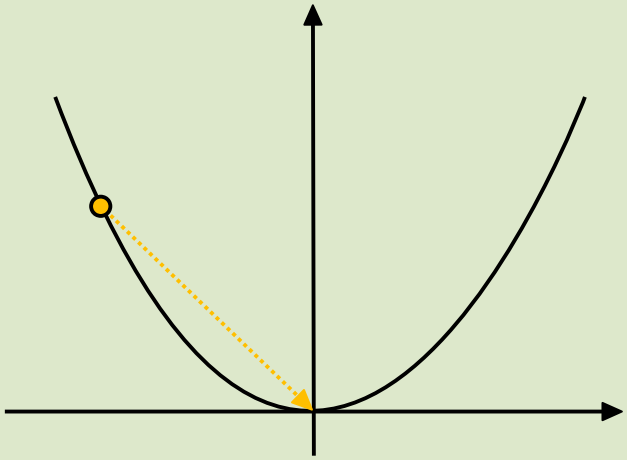
Diskussion Konvergenzverhalten

- Zur Diskussion der optimalen Schrittweite betrachten wir ein Polynom zweiter Ordnung:

$$F(x) = \frac{1}{2}ax^2 + bx + c$$

- Wir untersuchen im folgenden das Konvergenzverhalten für verschiedene Werte von η :

(1) $\eta = \dfrac{1}{F''(x)}$ } Konvergenz nach einem Schritt



Diskussion Konvergenzverhalten

- Zur Diskussion der optimalen Schrittweite betrachten wir ein Polynom zweiter Ordnung:

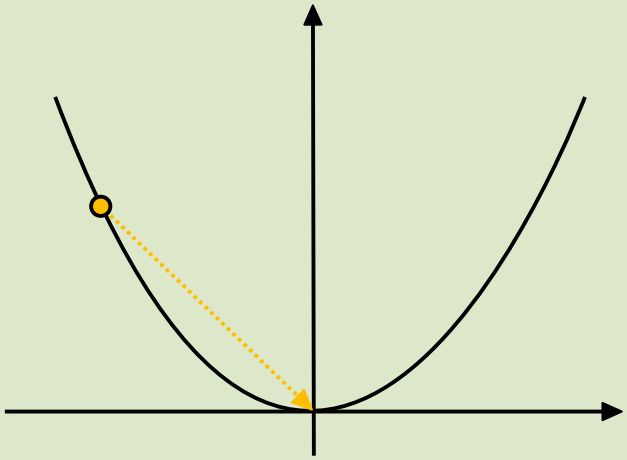
$$F(x) = \frac{1}{2}ax^2 + bx + c$$

- Wir untersuchen im folgenden das Konvergenzverhalten für verschiedene Werte von η :

$$(1) \quad \eta = \frac{1}{F''(x)}$$

$$(2) \quad \eta < \frac{1}{F''(x)}$$

} Konvergenz nach einem Schritt



Diskussion Konvergenzverhalten

- Zur Diskussion der optimalen Schrittweite betrachten wir ein Polynom zweiter Ordnung:

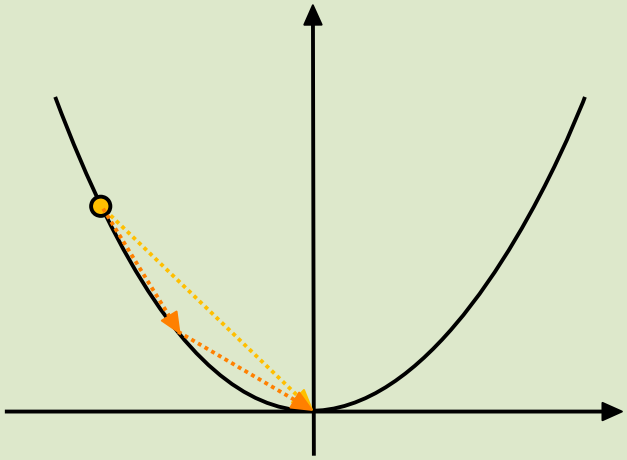
$$F(x) = \frac{1}{2}ax^2 + bx + c$$

- Wir untersuchen im folgenden das Konvergenzverhalten für verschiedene Werte von η :

$$(1) \quad \eta = \frac{1}{F''(x)}$$

$$(2) \quad \eta < \frac{1}{F''(x)}$$

} Konvergenz nach einem Schritt



Diskussion Konvergenzverhalten

- Zur Diskussion der optimalen Schrittweite betrachten wir ein Polynom zweiter Ordnung:

$$F(x) = \frac{1}{2}ax^2 + bx + c$$

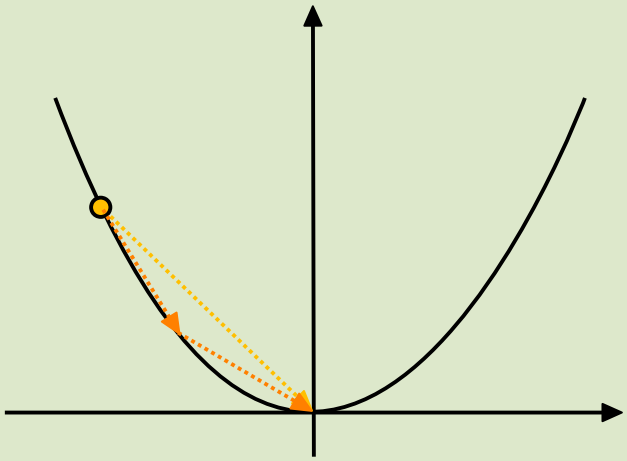
- Wir untersuchen im folgenden das Konvergenzverhalten für verschiedene Werte von η :

$$(1) \quad \eta = \frac{1}{F''(x)}$$

$$(2) \quad \eta < \frac{1}{F''(x)}$$

$$(3) \quad \eta > \frac{1}{F''(x)}$$

} Konvergenz nach einem Schritt



Diskussion Konvergenzverhalten

- Zur Diskussion der optimalen Schrittweite betrachten wir ein Polynom zweiter Ordnung:

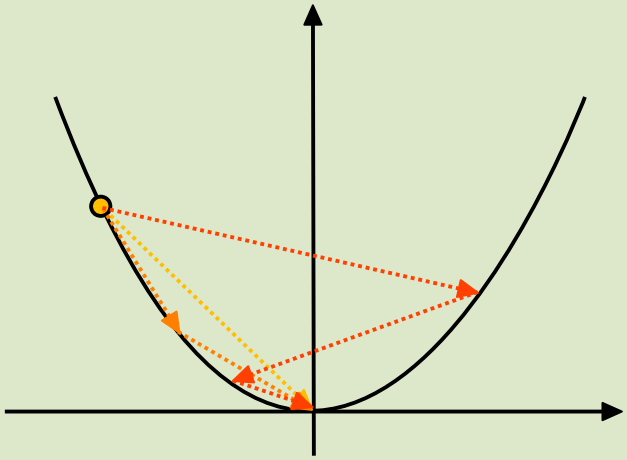
$$F(x) = \frac{1}{2}ax^2 + bx + c$$

- Wir untersuchen im folgenden das Konvergenzverhalten für verschiedene Werte von η :

(1) $\eta = \dfrac{1}{F''(x)}$ } Konvergenz nach einem Schritt

(2) $\eta < \dfrac{1}{F''(x)}$

(3) $\eta > \dfrac{1}{F''(x)}$ } Konvergenz nach mehr als einem Schritt (t.w. oszillierend



Diskussion Konvergenzverhalten

- Zur Diskussion der optimalen Schrittweite betrachten wir ein Polynom zweiter Ordnung:

$$F(x) = \frac{1}{2}ax^2 + bx + c$$

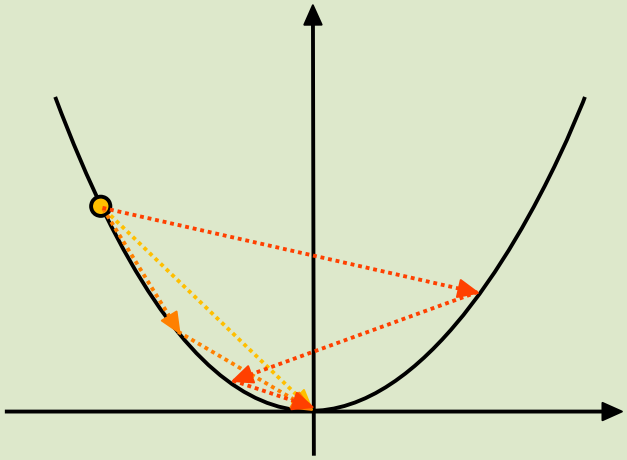
- Wir untersuchen im folgenden das Konvergenzverhalten für verschiedene Werte von η :

(1) $\eta = \dfrac{1}{F''(x)}$ } Konvergenz nach einem Schritt

(2) $\eta < \dfrac{1}{F''(x)}$ } Konvergenz nach mehr als einem Schritt (t.w. oszillierend

(3) $\eta > \dfrac{1}{F''(x)}$ }

(4) $\eta = 2\dfrac{1}{F''(x)}$

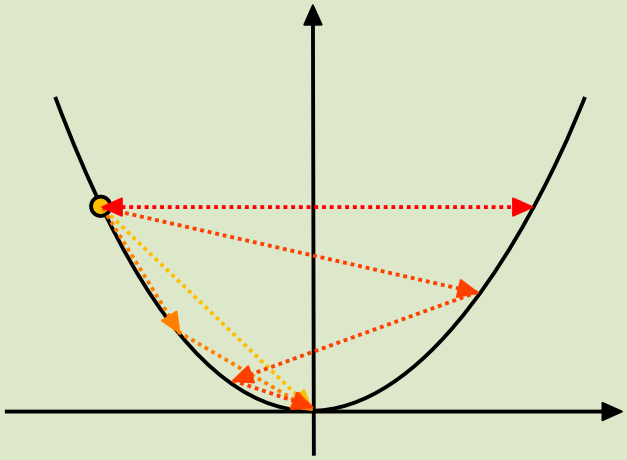


Diskussion Konvergenzverhalten

- Zur Diskussion der optimalen Schrittweite betrachten wir ein Polynom zweiter Ordnung:

$$F(x) = \frac{1}{2}ax^2 + bx + c$$

- Wir untersuchen im folgenden das Konvergenzverhalten für verschiedene Werte von η :



(1) $\eta = \frac{1}{F''(x)}$ } Konvergenz nach einem Schritt

(2) $\eta < \frac{1}{F''(x)}$

(3) $\eta > \frac{1}{F''(x)}$ } Konvergenz nach mehr als einem Schritt (t.w. oszillierend

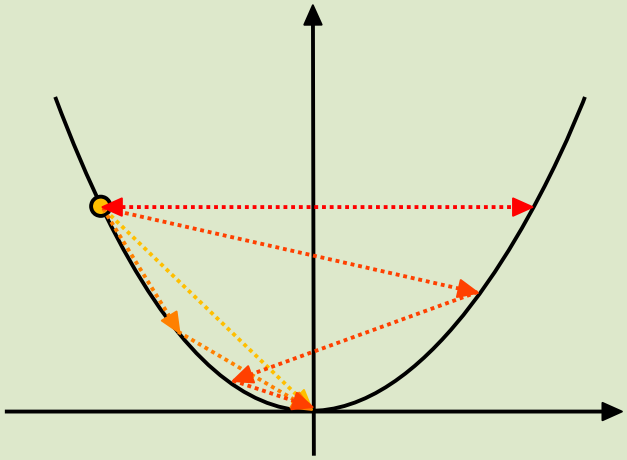
(4) $\eta = 2\frac{1}{F''(x)}$ } Stabile Oszillation

Diskussion Konvergenzverhalten

- Zur Diskussion der optimalen Schrittweite betrachten wir ein Polynom zweiter Ordnung:

$$F(x) = \frac{1}{2}ax^2 + bx + c$$

- Wir untersuchen im folgenden das Konvergenzverhalten für verschiedene Werte von η :



(1) $\eta = \frac{1}{F''(x)}$ } Konvergenz nach einem Schritt

(2) $\eta < \frac{1}{F''(x)}$ } Konvergenz nach mehr als einem Schritt (t.w. oszillierend

(3) $\eta > \frac{1}{F''(x)}$ }

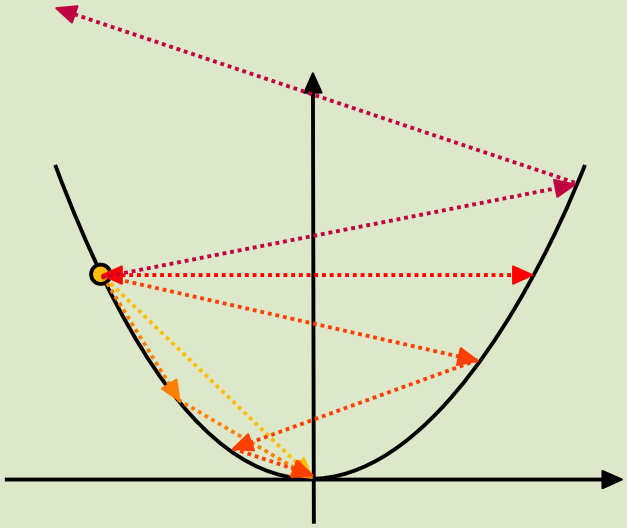
(4) $\eta = 2\frac{1}{F''(x)}$ } Stabile Oszillation

(5) $\eta > 2\frac{1}{F''(x)}$

Diskussion Konvergenzverhalten

- Zur Diskussion der optimalen Schrittweite betrachten wir ein Polynom zweiter Ordnung:

$$F(x) = \frac{1}{2}ax^2 + bx + c$$



- Wir untersuchen im folgenden das Konvergenzverhalten für verschiedene Werte von η :

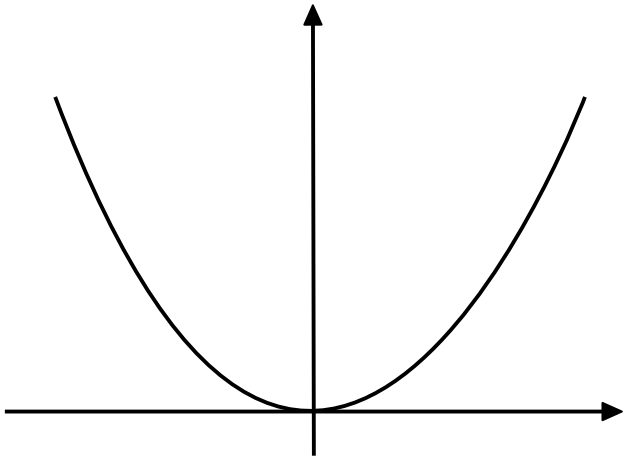
(1)	$\eta = \frac{1}{F''(x)}$	Konvergenz nach einem Schritt
(2)	$\eta < \frac{1}{F''(x)}$	Konvergenz nach mehr als einem Schritt (t.w. oszillierend
(3)	$\eta > \frac{1}{F''(x)}$	
(4)	$\eta = 2\frac{1}{F''(x)}$	Stabile Oszillation
(5)	$\eta > 2\frac{1}{F''(x)}$	Divergenz

Newtonverfahren

- Zur Diskussion der optimalen Schrittweite betrachten wir ein Polynom zweiter Ordnung:

$$F(x) = \frac{1}{2}ax^2 + bx + c$$

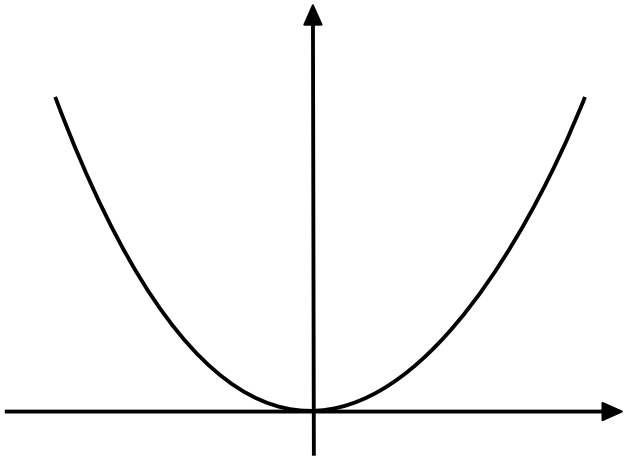
- Die Diskussion anhand von Polynomen zweiter Ordnung lässt sich auf beliebige zweifach differenzierbare Funktionen übertragen, solange man sie (nur) bis zur zweiten Ordnung in der Taylorreihe approximiert.



Newtonverfahren

- Zur Diskussion der optimalen Schrittweite betrachten wir ein Polynom zweiter Ordnung:

$$F(x) = \frac{1}{2}ax^2 + bx + c$$



- Die Diskussion anhand von Polynomen zweiter Ordnung lässt sich auf beliebige zweifach differenzierbare Funktionen übertragen, solange man sie (nur) bis zur zweiten Ordnung in der Taylorreihe approximiert.

- Die Aktualisierungsregel des Gradientenabstiegs nimmt dabei die folgende Form an:

$$x_{k+1} = x_k - \frac{F'(x_k)}{F''(x_k)}$$

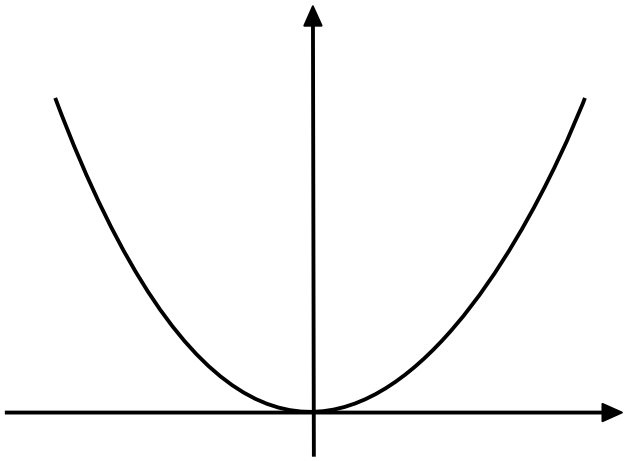
solange:

$$|F(x_k) - F(x_{k-1})| > \epsilon$$

Newtonverfahren

- Zur Diskussion der optimalen Schrittweite betrachten wir ein Polynom zweiter Ordnung:

$$F(x) = \frac{1}{2}ax^2 + bx + c$$



- Die Diskussion anhand von Polynomen zweiter Ordnung lässt sich auf beliebige zweifach differenzierbare Funktionen übertragen, solange man sie (nur) bis zur zweiten Ordnung in der Taylorreihe approximiert.

- Die Aktualisierungsregel des Gradientenabstiegs nimmt dabei die folgende Form an:

$$x_{k+1} = x_k - \frac{F'(x_k)}{F''(x_k)}$$

solange:

$$|F(x_k) - F(x_{k-1})| > \epsilon$$

Hierbei handelt es sich um das bekannte Newtonverfahren zur Nullstellenbestimmung angewandt auf $F'(x)$.

Newtonverfahren in d Dimensionen

- Ein allgemeines Polynom zweiter Ordnung in d Dimensionen hat die Form:

$$F(\vec{x}) = \frac{1}{2}\vec{x}A\vec{x}^{\intercal} + \vec{b}^{\intercal}\vec{x} + \vec{c}$$

- In der Aktualisierungsregel des Gradientenabstiegs wird dann die zweite Ableitung durch die **Hessematrix** ersetzt $H(\vec{x})$.

$$\vec{x}_{k+1} = \vec{x}_k - A^{-1}\vec{\nabla}F(x_k) = \vec{x}_k - H^{-1}(\vec{x}_k)\vec{\nabla}F(x_k)$$

solange:

$$|F(\vec{x}_k) - F(\vec{x}_{k-1})| > \epsilon$$

Gradientenabstieg in d Dimensionen

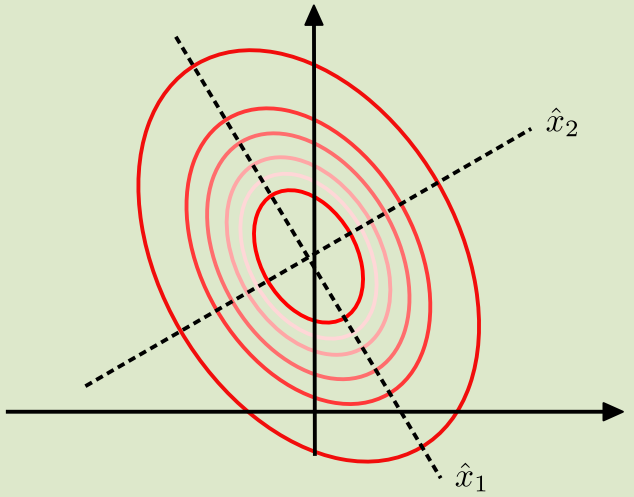
- In d Dimensionen läßt sich jede Parabel auf ihre Hauptachsen transformieren. Schnitte entlang der Hauptachsen führen wiederum auf Parabeln.

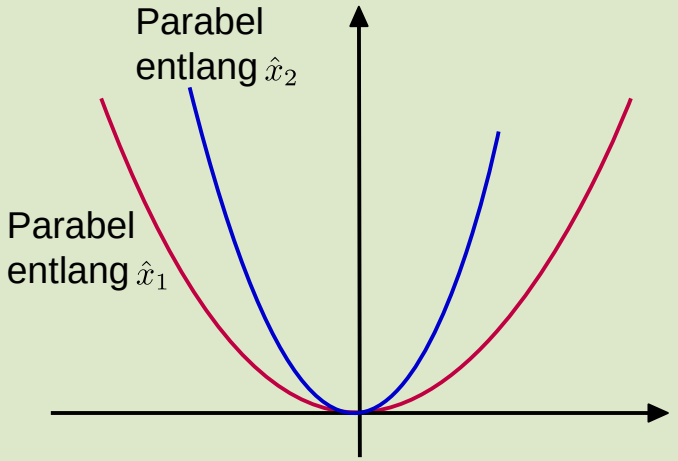
- **Problem**: Die Öffnungswinkel der Parabeln entlang der Hauptachsen – und damit verbunden die zweiten Richtungsableitungen entlang der Hauptachsen – sind i.a. unterschiedlich groß.

- Wie ist η zu wählen, um optimale und garantierte Konvergenz zu erreichen?

- Je höherdimensional der Definitionsbereich von $F(x)$ ist desto schwieriger lässt sich diese Frage im Rahmen des naiven Gradientenabstiegverfahrens beantworten.

- Konvergenz ist umso schwieriger zu erreichen je gößer die Konditionszahl der Hessematrix ist (d.h. je weiter die Eigenwerte der Hessematrix auseinander liegen).



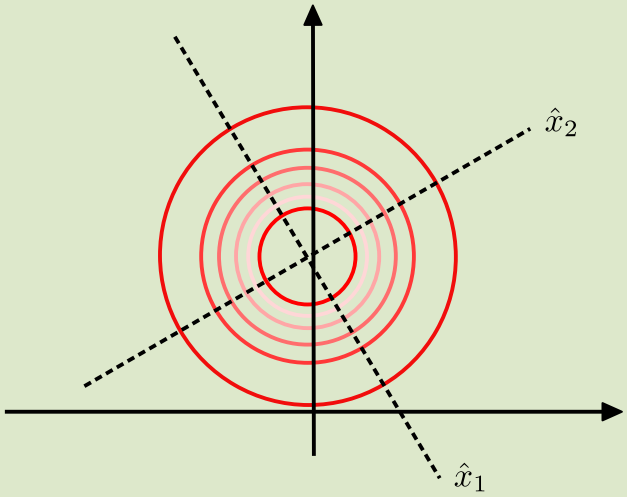


Gradientenabstieg in d Dimensionen

- In d Dimensionen läßt sich jede Parabel auf ihre Hauptachsen transformieren. Schnitte entlang der Hauptachsen führen wiederum auf Parabeln.

- **Problem**: Die Öffnungswinkel der Parabeln entlang der Hauptachsen – und damit verbunden die zweiten Richtungsableitungen entlang der Hauptachsen – sind i.a. unterschiedlich groß.

- Durch das Newtonverfahren wird das Problem automatisch gelöst. Die Multiplikation mit $H^{-1}(x)$ transformiert die Parabel auf eine Einheitsparabel.

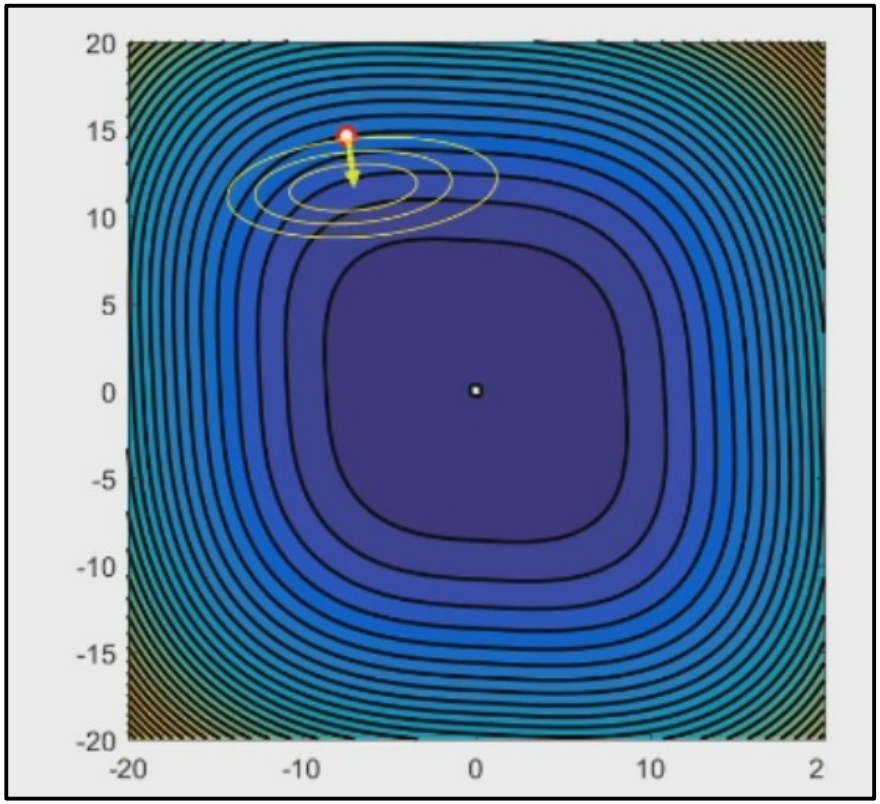


Sie können das auch durch die Aktualisierungsregel erkennen.

$$\vec{x}_{k+1} = \vec{x}_k - \underbrace{H^{-1}(\vec{x}_k)\vec{\nabla}F(x_k)}_{\text{Transformation des Gradienten mit } \eta \equiv 1}$$

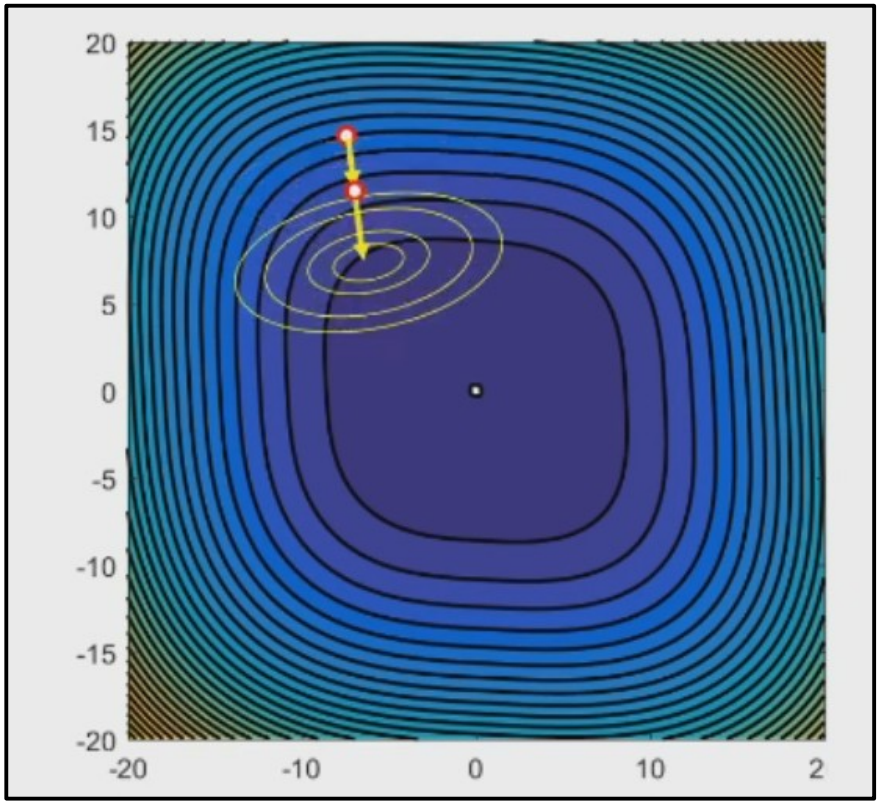
Newtonverfahren in 2 Dimensionen

- Die folgenden Folien veranschaulichen das Newtonverfahren in 2 Dimensionen:



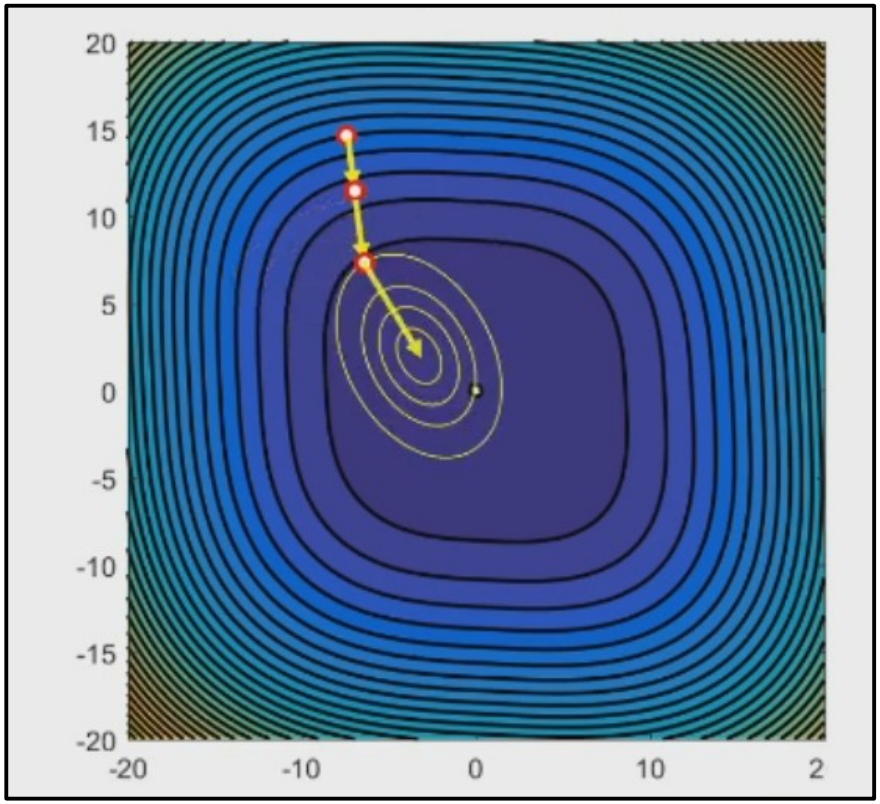
Newtonverfahren in 2 Dimensionen

- Die folgenden Folien veranschaulichen das Newtonverfahren in 2 Dimensionen:



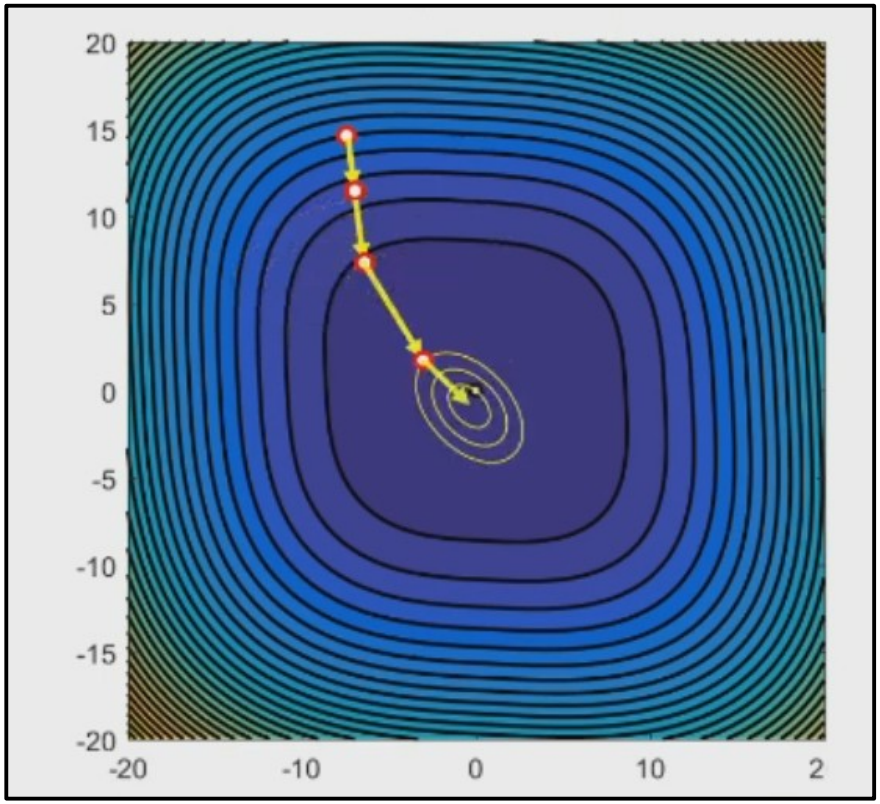
Newtonverfahren in 2 Dimensionen

- Die folgenden Folien veranschaulichen das Newtonverfahren in 2 Dimensionen:



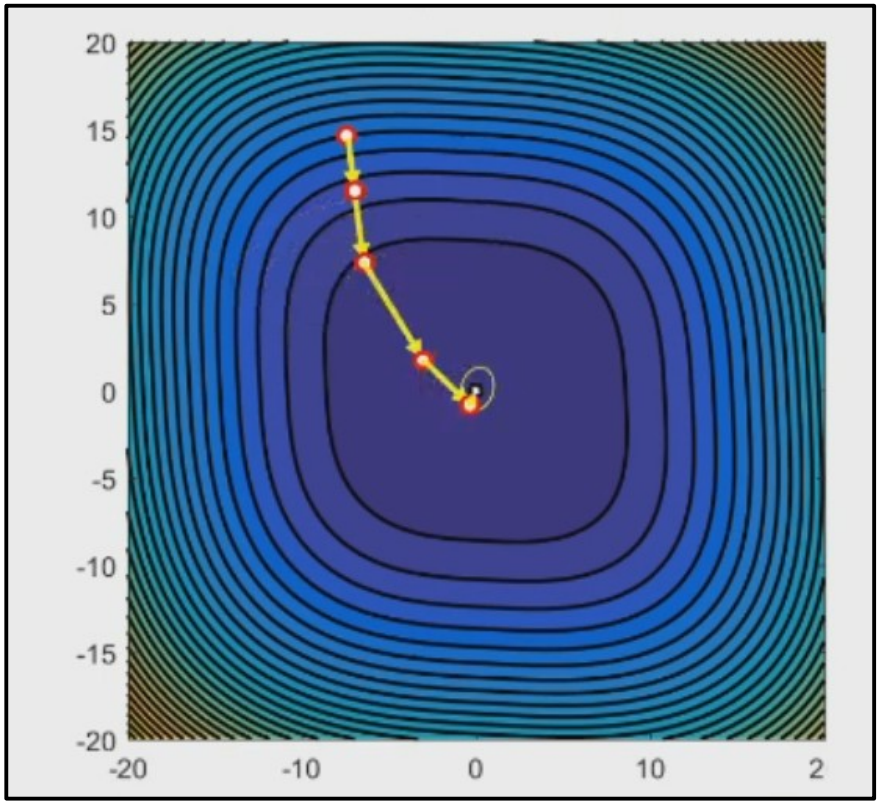
Newtonverfahren in 2 Dimensionen

- Die folgenden Folien veranschaulichen das Newtonverfahren in 2 Dimensionen:



Newtonverfahren in 2 Dimensionen

- Die folgenden Folien veranschaulichen das Newtonverfahren in 2 Dimensionen:



Newtonverfahren – Diskussion

- Das Newtonverfahren hat sehr gute Konvergenzeigenschaften für konvexe Funktionen.
- Konvergenz in ein globales Minimum ist jedoch nicht garantiert.
- Probleme bestehen bei nicht konvexen Funktionen, bei denen insbesondere in höheren Dimensionen $H(x)$ nicht garantiert **positiv definit** ist.
- Zur Optimierung bei Problemen sehr hoher Dimensionalität (wie z.B. im Bereich maschienellen Lernens) ist das Newtonverfahren ungeeignet, weil es nicht nur die Berechnung sondern auch die Inversion von $H(x)$ voraussetzt.

Adaptive Schrittweite

- Eine andere Möglichkeit Konvergenzprobleme zu vermeiden, besteht darin (ggf. mit großer) fester Schrittweite zu beginnen und diese sukzessive zu verringern.

- Der Beginn mit großer Schrittweite soll dabei verhindern, durch unglückliche Wahl der Anfangsparameter in Nebenminima „gefangen“ zu bleiben.

- Dabei sollten die folgenden Bedingungen an die Schrittweiten bestehen:

$$\sum_{k=1}^{\infty} \eta_k \to \infty$$

Es muss möglich sein, durch eine unendlich lange Folge den gesamten Definitionsbereich der Funktion abzudecken.

$$\sum_{k=1}^{\infty} \eta_k^2 < \infty$$

Die Reihe soll trotzdem konvergent sein und die Elemente der Folge immer kleiner werden.

Adaptive Schrittweite

- Eine andere Möglichkeit Konvergenzprobleme zu vermeiden, besteht darin (ggf. mit großer) fester Schrittweite zu beginnen und diese sukzessive zu verringern.

- Der Beginn mit großer Schrittweite soll dabei verhindern, durch unglückliche Wahl der Anfangsparameter in Nebenminima „gefangen“ zu bleiben.

- Etablierte Schrittweitenalgorithmen sind:

$$\eta_k = \frac{\eta_0}{k+1} \qquad \text{(Linear)}$$

$$\eta_k = \frac{\eta_0}{(k+1)^2} \qquad \text{(Quadratisch)}$$

$$\eta_k = \eta_0\, e^{-\beta\, k} \quad \beta > 0 \qquad \text{(Exponentiell)}$$

- Eine im Bereich maschienellen Lernens verwendete Methode ist z.B. mit einer festen Schrittweite zu beginnen, die nach einer bestimmten Anzahl an initialen Schritten sukessive reduziert wird.

Impulsverfahren (engl. *momentum methods*)

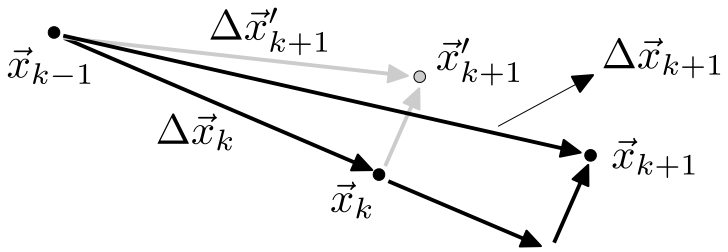
- Impulsverfahren beruhen auf der Annahme, dass die ursprüngliche Richtung, eine gute Wahl zum Auffinden des Minimums war.

- Man behält also ein „Gedächtnis“ dieser Information in der Aktualisierungregel des Gradientenabstiegs, z.B. durch (gewichtete) Mittelwertbildung:

$$\Delta\vec{x}_{k+1} = \beta\Delta\vec{x}_k - \eta\vec{\nabla}F(x_k)$$
$$\vec{x}_{k+1} = \vec{x}_k + \Delta\vec{x}_{k+1}$$

solange:

$$|F(\vec{x}_k) - F(\vec{x}_{k-1})| > \epsilon$$



Einfacher Gradientenabstieg (gestrichene Variablen) in **grau**.
Gradientenabstieg nach Impulsverfahren in **schwarz**.

- Dieses einfache Verfahren verzichtet auf die zweite Ableitung. Es unterdrückt schnelle, stark variierende Änderungen im Abstieg und somit auch Oszillationen.

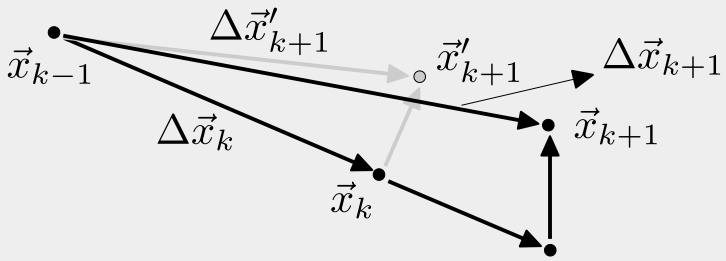
Nestorov's accelerated gradient (Y. Nesterov 1983)

- Eine Variation mit nachweisbar noch besserem Konvergenzverhalten besteht darin zuerst den Schritt aus dem vorherigen Abstieg durchzuführen und dort die Ableitung auszuwerten:

$$\Delta\vec{x}_{k+1} = \beta\Delta\vec{x}_k - \eta\vec{\nabla}F(x_k + \beta\Delta\vec{x}_k)$$
$$\vec{x}_{k+1} = \vec{x}_k + \Delta\vec{x}_{k+1}$$

solange:

$$|F(\vec{x}_k) - F(\vec{x}_{k-1})| > \epsilon$$



Einfacher Gradientenabstieg (gestrichene Variablen) in grau.
Gradientenabstieg nach Impulsverfahren in **schwarz**.

5 Optimierungsalgorithmen

5.3 Optimierung mit Nebenbedingungen

Wir diskutieren abschließen wie Nebenbedingungen in einem Optimierungsverfahren integriert werden können.

Zurück zu Folie 63

Einfache Beispiele für Nebenbedingung

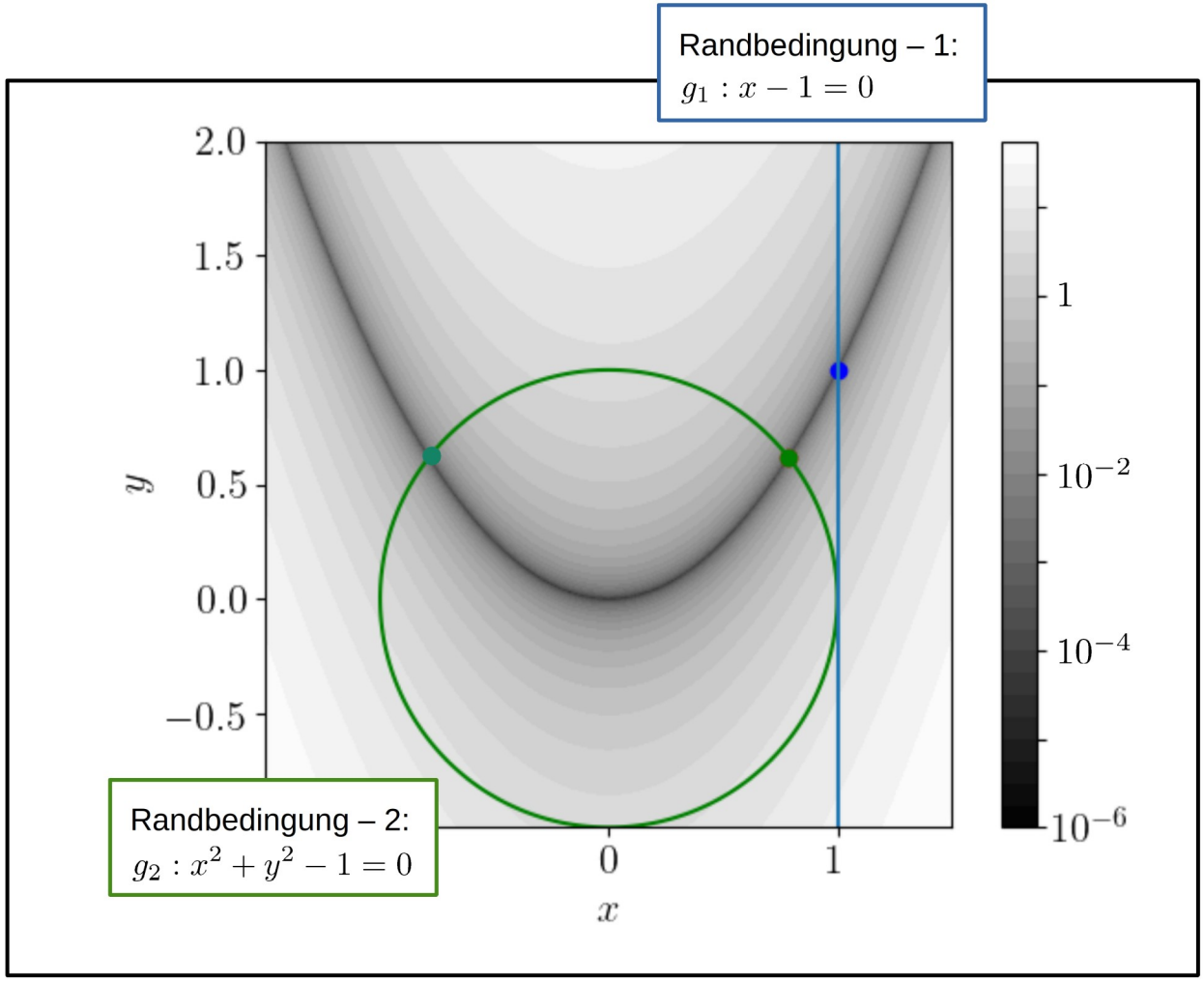
- Zu minimierende Funktion:

$$f(x,y) = \left(y - x^2\right)^2$$

- Parabelförmiges Tal von Minima (siehe Bild) mit:

$$f(x,y)|_{y=x^2} = 0$$

- Zwei Randbedingungen für Minimierung im Bild dargestellt.



Manifeste Randbedingungen

- Die klarste, wenn auch nicht immer einfachste, Möglichkeit Randbedingungen in den Minimierungsprozess zu implementieren, ist es diese in der zu minimierenden Funktion **manifest** zu machen.

- Wie würde das am Beispiel der vorherigen Folie für die Bedingungen – 1 und 2 aussehen?

Manifeste Randbedingungen

- Die klarste, wenn auch nicht immer einfachste, Möglichkeit Randbedingungen in den Minimierungsprozess zu implementieren, ist es diese in der zu minimierenden Funktion **manifest** zu machen.

- Wie würde das am Beispiel der vorherigen Folie für die Bedingungen – 1 und 2 aussehen?

$$f(x,y)|_{g_1} = (y-1)^2$$
$$f(x,y)|_{g_2} = (y-(1-y^2))^2$$

Manifeste Randbedingungen

- Die klarste, wenn auch nicht immer einfachste, Möglichkeit Randbedingungen in den Minimierungsprozess zu implementieren, ist es diese in der zu minimierenden Funktion **manifest** zu machen.

- Wie würde das am Beispiel der vorherigen Folie für die Bedingungen – 1 und 2 aussehen?

$$f(x,y)|_{g_1} = (y-1)^2$$
$$f(x,y)|_{g_2} = (y-(1-y^2))^2$$

- Sie können auf diese Weise auch Randbedingungen in Form von Ungleichungen in den Optimierungsprozess implementieren. Wie würden Sie die Bedingungen $a < x < b$ implementieren?

Manifeste Randbedingungen

- Die klarste, wenn auch nicht immer einfachste, Möglichkeit Randbedingungen in den Minimierungsprozess zu implementieren, ist es diese in der zu minimierenden Funktion **manifest** zu machen.

- Wie würde das am Beispiel der vorherigen Folie für die Bedingungen – 1 und 2 aussehen?

$$f(x,y)|_{g_1} = (y-1)^2$$
$$f(x,y)|_{g_2} = (y-(1-y^2))^2$$

- Sie können auf diese Weise auch Randbedingungen in Form von Ungleichungen in den Optimierungsprozess implementieren. Wie würden Sie die Bedingungen $a < x < b$ implementieren?

Antwort: durch geeignete Transformation $x \to x'$

$$x' = \arctan\left(\frac{2(x-a)}{b-a} - 1\right); \qquad x = \frac{\big(\sin(x')+1\big)b-a\big)}{2} + a$$

Methode der Lagrange Multiplikatoren

- Ein gängiges Verfahren beruht auf dem Satz der Lagrange-Multiplikatoren:
  - Addiere Randbedingungen als Gleichungen der Form $g_i(\vec{x})$ zu der ursprünglich zu minimierenden Funktion $f(\vec{x})$ und erzeuge so eine neue zu minimiernde Funktion:

$$F(\vec{x}, \{\lambda_i\}) = f(\vec{x}) + \sum_{i=1}^{n} \lambda_i g_i(\vec{i})$$

  - Durch die Ableitungen $\partial_{\lambda_i} F(\vec{x}, \{\lambda_i\})$ sind die Randbedingungen $\{g_i(\vec{x})\}$ automatisch erfüllt. Die $\{\lambda_i\}$ heißen **Lagrange-Multiplikatoren**.

Methode der Lagrange Multiplikatoren

- Ein gängiges Verfahren beruht auf dem Satz der Lagrange-Multiplikatoren:
  - Addiere Randbedingungen als Gleichungen der Form $g_i(\vec{x})$ zu der ursprünglich zu minimierenden Funktion $f(\vec{x})$ und erzeuge so eine neue zu minimiernde Funktion:

$$F(\vec{x}, \{\lambda_i\}) = f(\vec{x}) + \sum_{i=1}^{n} \lambda_i g_i(\vec{i})$$

  - Durch die Ableitungen $\partial_{\lambda_i} F(\vec{x}, \{\lambda_i\})$ sind die Randbedingungen $\{g_i(\vec{x})\}$ automatisch erfüllt. Die $\{\lambda_i\}$ heißen **Lagrange-Multiplikatoren**.

- Für das Beispiel von Folie 61 ergibt sich:

$$F(\vec{x}, \lambda)|_{g_1} = \left(y - x^2\right)^2 + \lambda\left(x - 1\right)$$
$$F(\vec{x}, \lambda)|_{g_2} = \left(y - x^2\right)^2 + \lambda\left(x^2 + y^2 - 1\right)$$

Berücksichtigung durch Strafterme

- Ein allgemeineres Vorgehen besteht darin, Strafterme (engl. *penalty term*) zu der zu minimierenden Funktion zuzufügen. Die neue, zu minimierende Funktion nimmt dann die folgende Form an:

$$\mathcal{P}(\vec{x}, \{\gamma_i\}) = f(\vec{x}) + \underbrace{\sum_{i=1}^{n} \gamma_i ||g_i(\vec{x})||^2}_{\text{penalty term}}$$

- Der *penalty term* kann dabei jede Form annehmen. Die Konvergenzeigenschaften können von der Wahl der $\{\gamma_i\}$ abhängen.

Berücksichtigung durch Strafterme

- Dieses Vorgehen erlaubt es Parameter mit Unsicherheiten mit vorgegebenem Prior in die Likelihood Funktion einzubringen (vgl. mit VL-03 Folien 45ff):

- **Beispiel:** bestimme das Minimum der NLL mit einem Störparameter θ der mit einer Unsicherheit von σ_θ bestimmt ist:

$$-\ln\left(\mathcal{L}(\vec{x},\theta)\right) = -\ln\left(\mathcal{L}_0(\vec{x},\theta)\right) + \frac{1}{2}\left(\frac{\theta-\theta_0}{\sigma_\theta}\right)^2$$

NLL mit Randbedingung

NLL ohne Randbedingung

Strafterm (*Gaußian constraint*)

Zusammenfassung

- Einführung in die Problematik und Diskussion einfacher Verfahren ohne Ableitungen.
  - Rasterverfahren, simuliertes Abkühlen, Simplex Verfahren.
- Diskussion komplexerer Verfahren it Berechnung von Ableitungen.
  - Einfacher Gradientenabstieg, Newtonverfahren, adaptive Schrittweiten, Impulsverfahren.
- Optimierung mit Randbedingungen.