Lernen à la Bayes

—Von der Bildrekonstruktion zur inversen Quantentheorie —

Teil 2: Regressionsmodelle (Verallgemeinerte freie Felder)

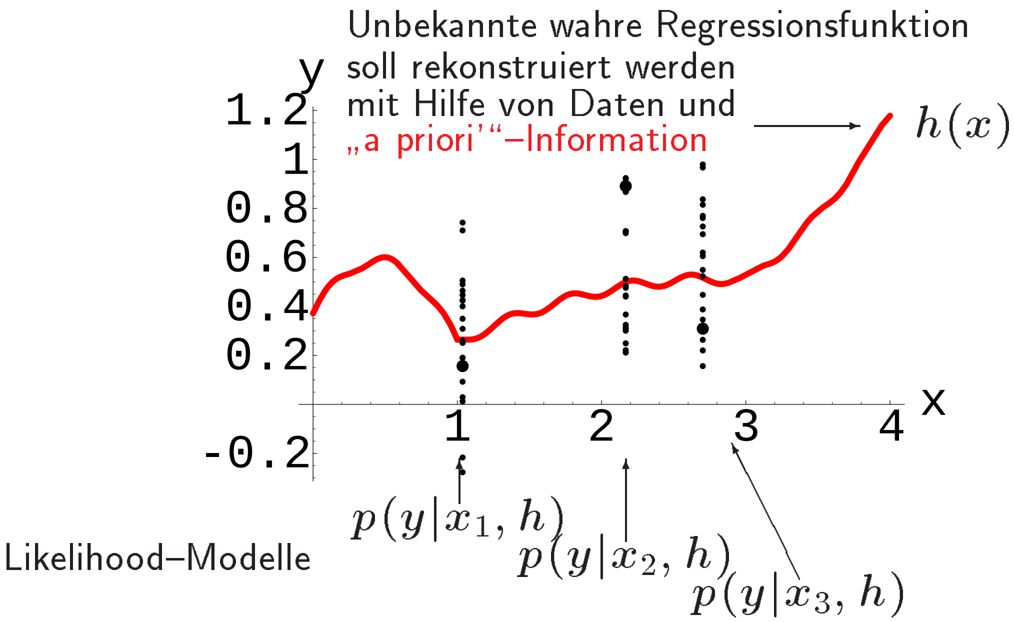
Jörg Lemm

Institut für Theoretische Physik
Universität Münster

lemm@uni-muenster.de

http://pauli.uni-muenster.de/~lemm

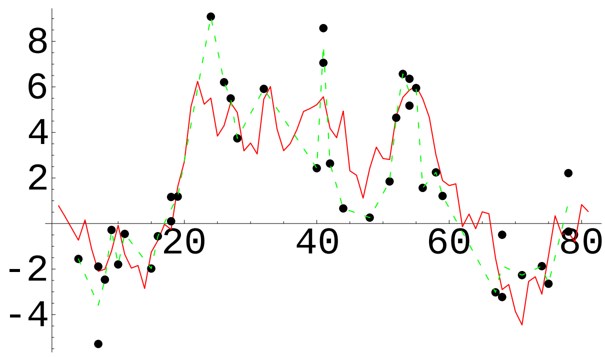
Probabilistisches Messmodell (Regression)

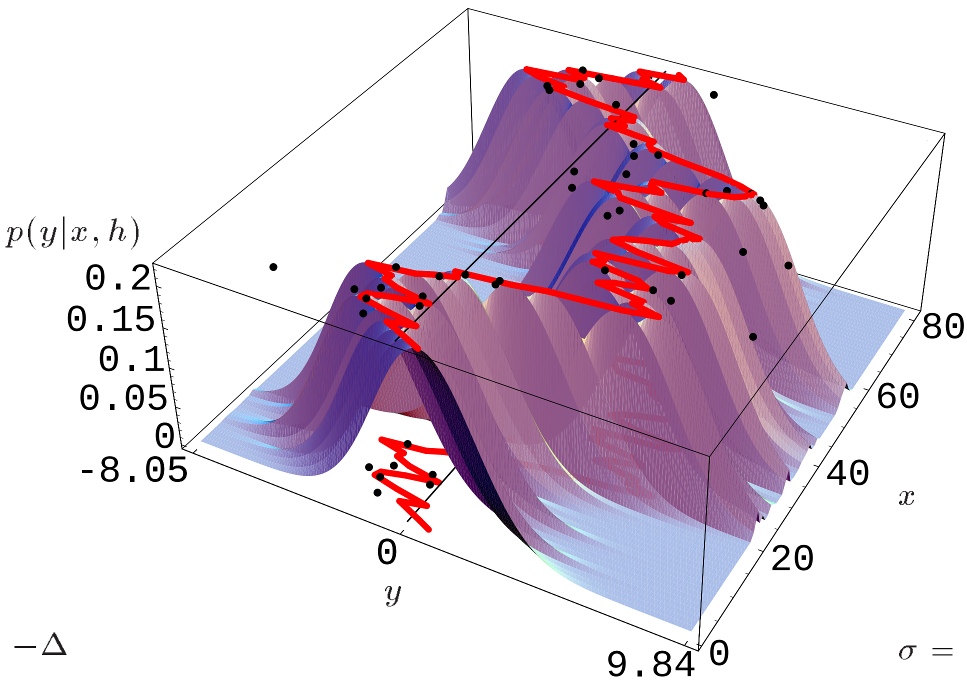


Likelihoodfunktion für h bei gegebenen (unabhängigen) Daten $D = \{(x_i, y_i)|1 \le i \le n\}$

$$\prod_{i=1}^{n} p(y_i|x_i, h)$$

Likelihoodmodell der (Gauß'schen) Regression



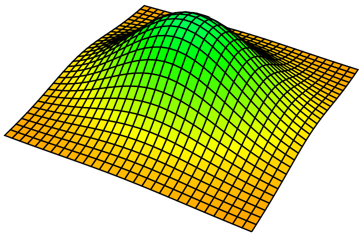


$K_0 = -\Delta$

$\sigma = 1.8$

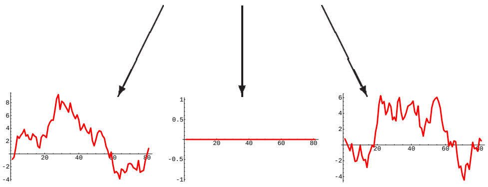
Bayes'sches Lernmodell (Regression)

„A priori“–Verteilung $p(h)$

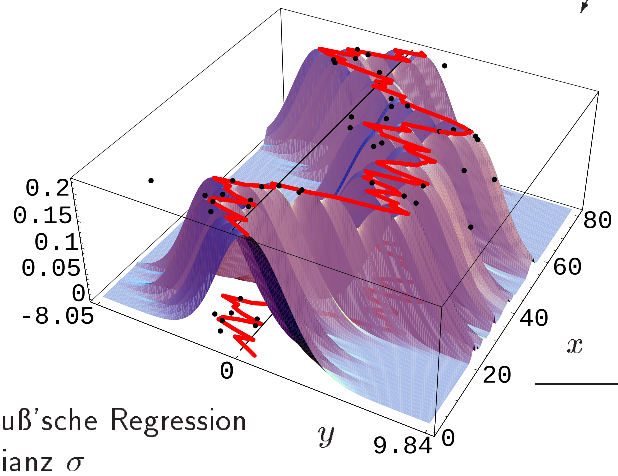


Stochastischer Prozess
z.B. Gauß –Prozess mit
Kovarianz $C = K_0^{-1}$

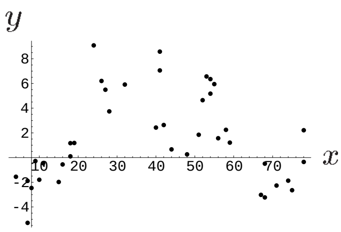
Beispiele von
Realisierungen einer
Regressionsfunktion $h(x)$



Likelihood–Modell $p(y|x, h)$

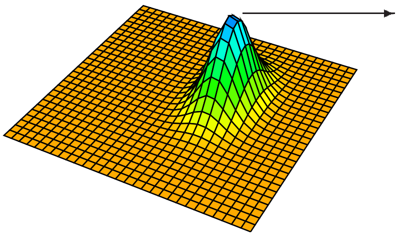


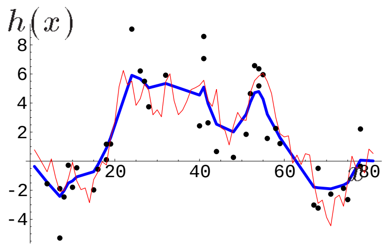
Eine mögliche
Realisierung von
(Trainings–)Daten D



z.B. Gauß'sche Regression
mit Varianz σ

„A posteriori“–Verteilung $p(h|D)$





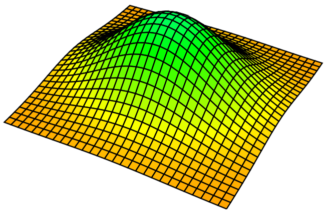
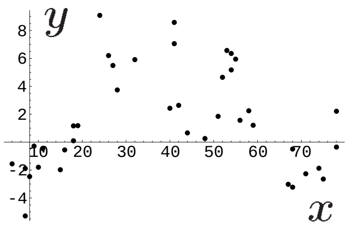
Maximum–Posterior–Lösung

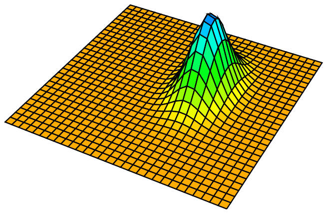
Posterior $p(h|D) \propto p(h) \prod_{i=1}^{n} p(y_i|x_i, h)$

Lernen à la Bayes

Prior × Likelihood ∝ Posterior

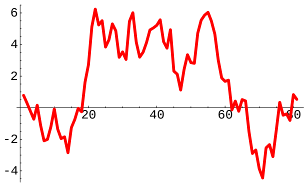
$$p(h) \times \prod_{i=1}^{n} p(y_i|x_i,h) \propto p(h|D)$$


$\times$


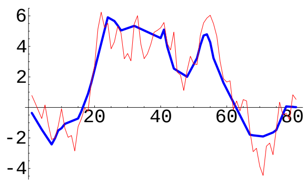
$\propto$


Zufallsauswahl
eines Zustandes

Messen

$h(x) =$


Unbekannter „wahrer Zustand der Natur“

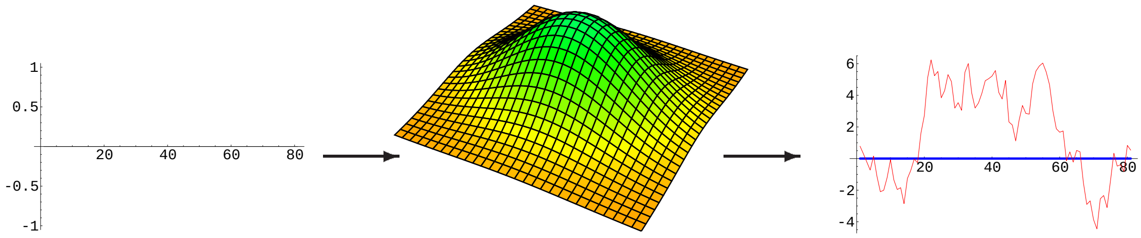


Maximum–Posterior–Lösung

Vorhersage neuer Daten

Iteratives Lernen

„A priori"–Verteilung $p(h)$

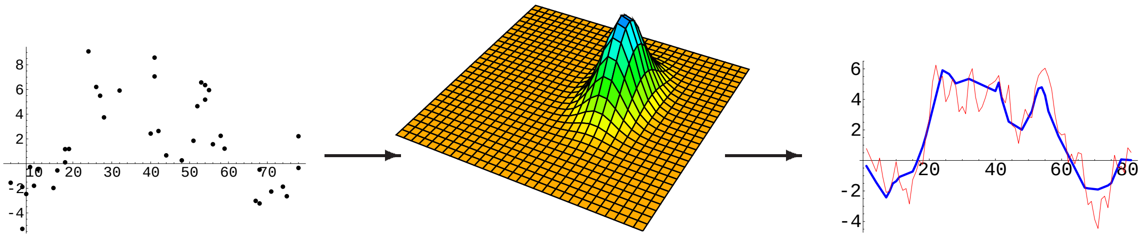


keine Daten

Maximum–Posterior–Lösung (blau) vs. Original (rot)

„A posteriori"–Verteilung $p(h|D_1) = \frac{p(D_1|h)p(h)}{p(D_1)}$

(spielt die Rolle einer neuen „a priori"–Verteilung)

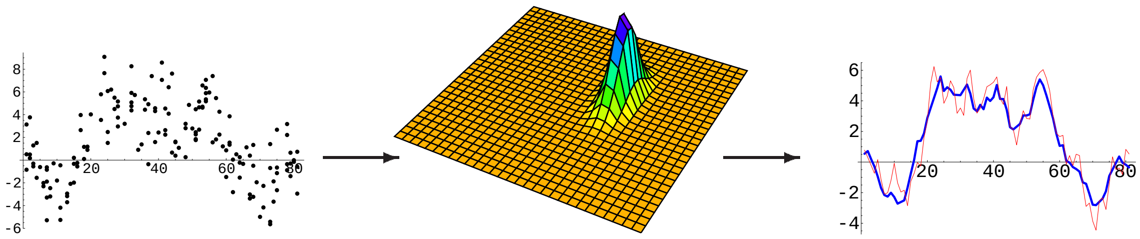


Daten

Maximum–Posterior–Lösung (blau) vs. Original (rot)

neue „a posteriori"–Verteilung $p(h|D_1, D_2) = \frac{p(D_1,D_2|h)p(h)}{p(D_1,D2)}$

(spielt die Rolle einer neuen „a priori"–Verteilung)



neue Daten

Maximum–Posterior–Lösung (blau) vs. Original (rot)

usw.

Wahrscheinlichkeiten und Energien

Eine beliebige Wahrscheinlichkeit kann, in Analogie zur statistischen Mechanik, geschrieben werden in der Form

$$p(x) = \frac{\exp[-\beta \overbrace{E(x)}^{\text{Energie}}]}{Z}$$

mit Normierungskonstante (Zustandssumme)

$$Z = \int dx\, e^{-\beta E(x)}$$

Vorteile:

- Produkte von Wahrscheinlichkeiten werden zu Summen von Energien
- Zur Maximierung von $p(x)$ kann auf Normierung verzichtet werden
- Die gewohnten Notationen der statistischen Mechanik können verwendet werden (erzeugende Funktionen, Störungstheorie, Gibbs–Verteilungen)

Maximieren einer Wahrscheinlichkeit $p(x)$ bedeutet Minimieren der Energie $E(x)$. Technisch äquivalent ist das Minimieren einer entsprechenden Fehlerfunktion $E(x)$.

Gauß'sche Regression

$$\underbrace{p(h|D)}_{\text{Posterior}} \propto \underbrace{\left(\prod_i p(y_i|x_i,h)\right)}_{\text{Likelihood}} \underbrace{p(h)}_{\text{Prior}}$$

$$\propto \exp\Bigg(-\underbrace{\frac{1}{2\sigma^2}\sum_i |y_i - h(x_i)|^2}_{\substack{\text{Trainingsfehler} \\ \text{bzw. „Trainings–Energie“}}} - \overbrace{\underbrace{\frac{\lambda}{2}\int dx \left|\frac{dh(x)}{dx}\right|^2}_{\substack{\text{Regularisierungsterm} \\ \text{bzw. „Prior–Energie“}}}}^{\text{z.B. Glattheit}}\Bigg)$$

(analog zu kinet. Energie)

Prior–Energie

$$\frac{\lambda}{2}\int dx \left|\frac{dh(x)}{dx}\right|^2 \underbrace{=}_{\text{R.B.}} \frac{\lambda}{2}\langle h\,|-\Delta|\,h\rangle$$

Gauß–Prior

Mittelwert(sfunktion)

Inverse Kovarianz

$$p(h) \propto \exp\left(-\frac{1}{2}\langle h - t_0\,|\mathbf{K}_0|\,h - t_0\rangle\right)$$

Trainings–Energie

$$\frac{1}{2\sigma^2}\sum_i |y_i - h(x_i)|^2 = \frac{1}{2}\langle h - t_D\,|K_D|\,h - t_D\rangle$$

wobei $\mathbf{K}_D(x,x') = \delta(x-x')\sum_i \delta(x-x_i)/\sigma^2$ und $t_D(x) = \frac{1}{n_x}\sum_i y_i\,\delta(x-x_i)$

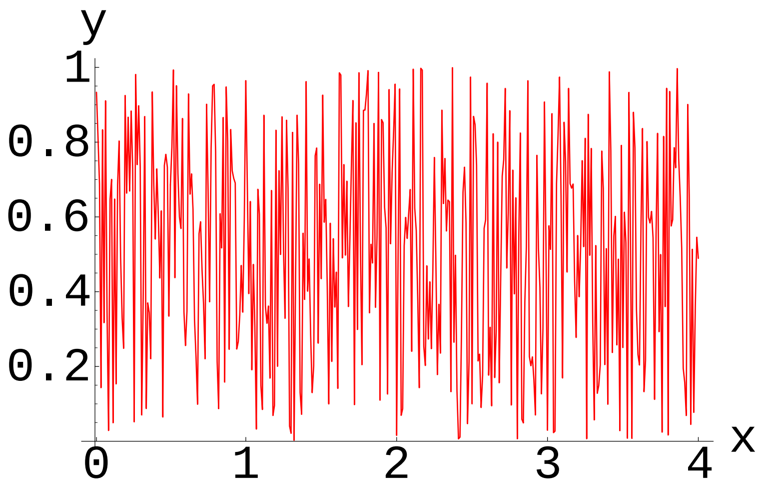
Posterior–Energie für Regressionsprobleme (verallg. freies Feld)

$$-\ln p(h|D) = \frac{1}{2}\sum_k \langle h - t_k|\mathbf{K}_k|h - t_k\rangle + c = \frac{1}{2}\langle h - t|\mathbf{K}|h - t\rangle + c'$$

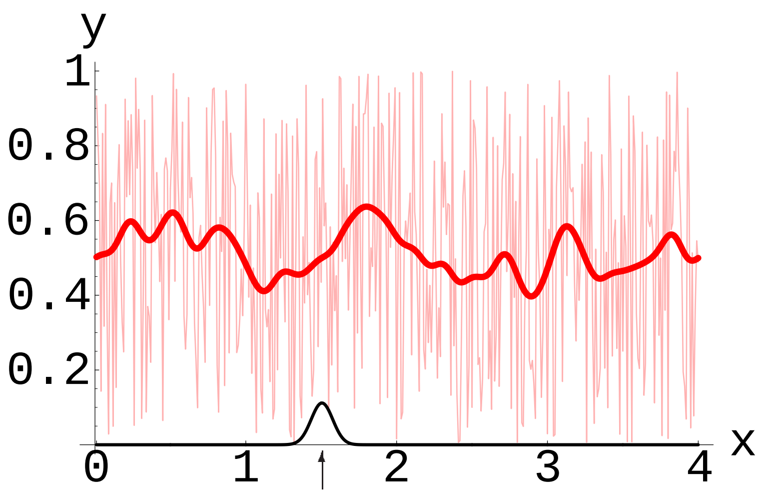
mit $\mathbf{K} = \sum_k \mathbf{K}_k$ und $t = \mathbf{K}^{-1}\sum_k \mathbf{K_k} t_k = h^*$.

Warum gibt es so viele glatte Funktionen?

Rein zufällige Regressionsfunktion:
$\longrightarrow$ kein Verallgemeinern/Lernen möglich!

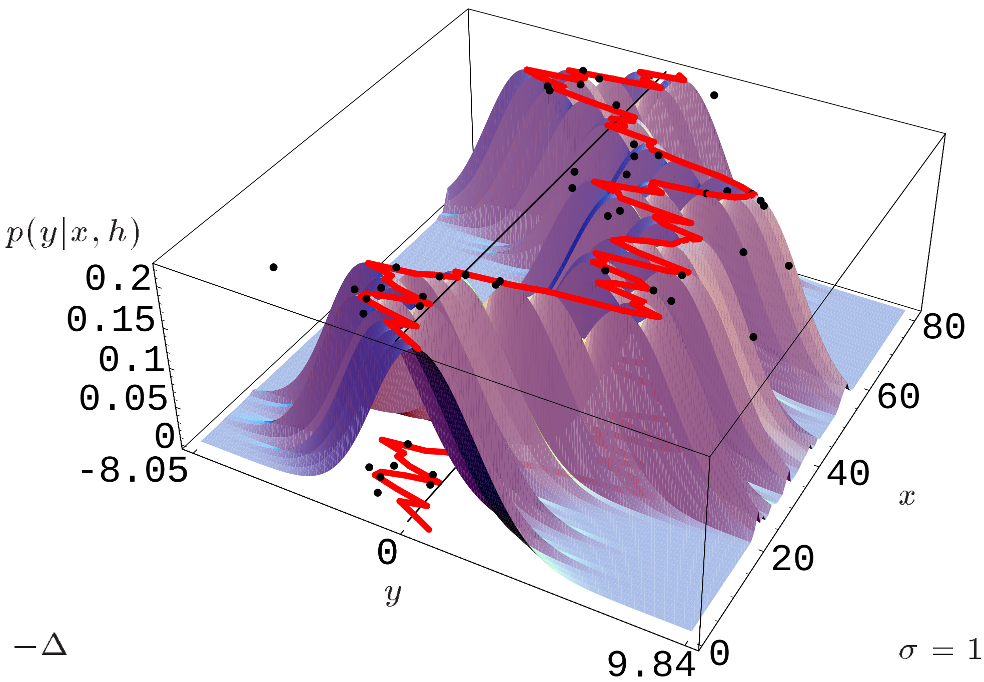


Realistisches Messgerät mit Input–Rauschen bzw. Input–Mittelung $\longrightarrow$ Verallgemeinern/Lernen möglich

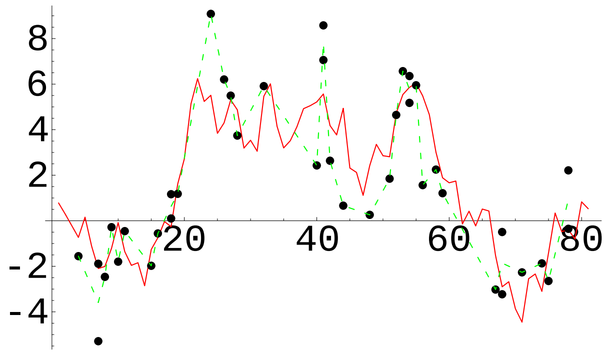


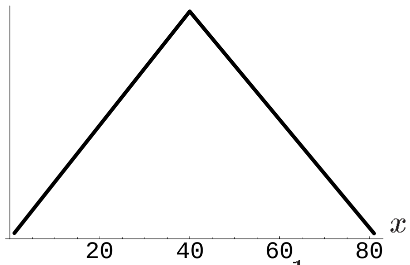
„A priori"–Information ist hierbei implementiert durch „a posteriori"–Kontrolle, d.h. zum Zeitpunkt des Testens einer Theorie.

Rekonstruktion stetiger Funktionen

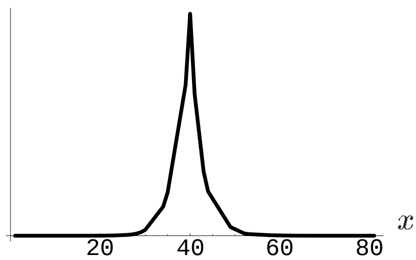


$K_0 = -\Delta$ $\sigma = 1.8$



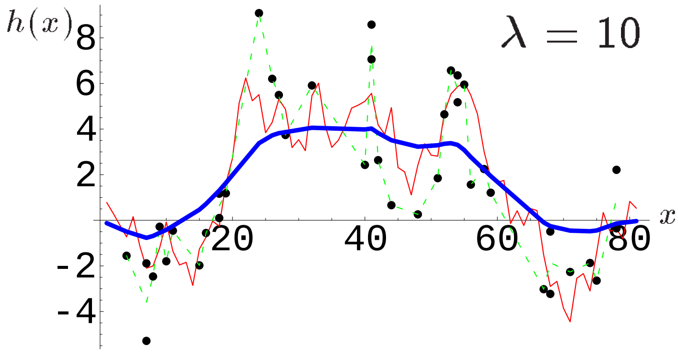


Kovarianz $K_0^{-1}(x, x_0)$

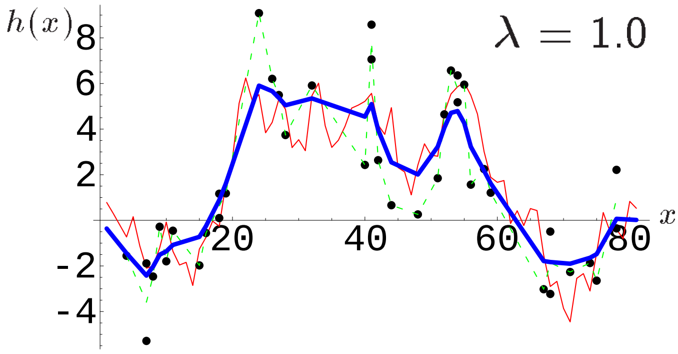


Äquival. Kern $K^{-1}(x, x_0)$

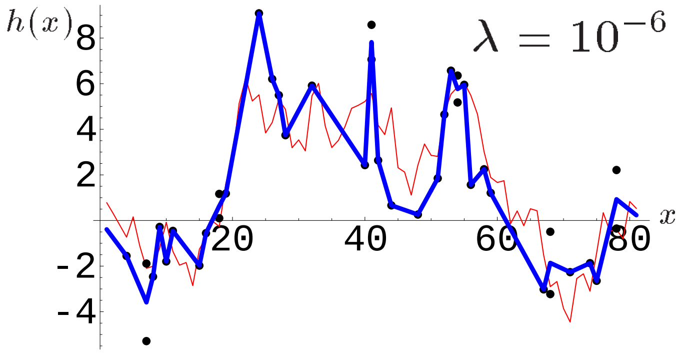
Rekonstruktion stetiger Funktionen



zu stark geglättet

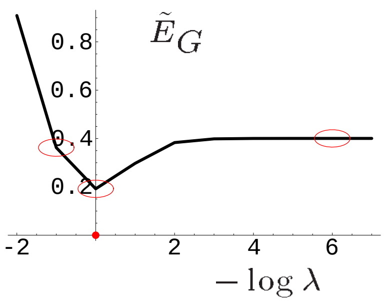


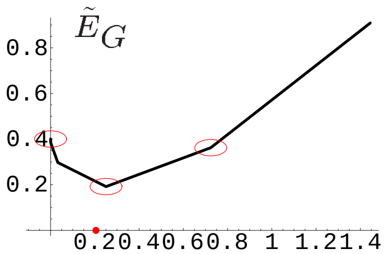
optimal



"overfitting"
(stückweise lineare Interpolation)

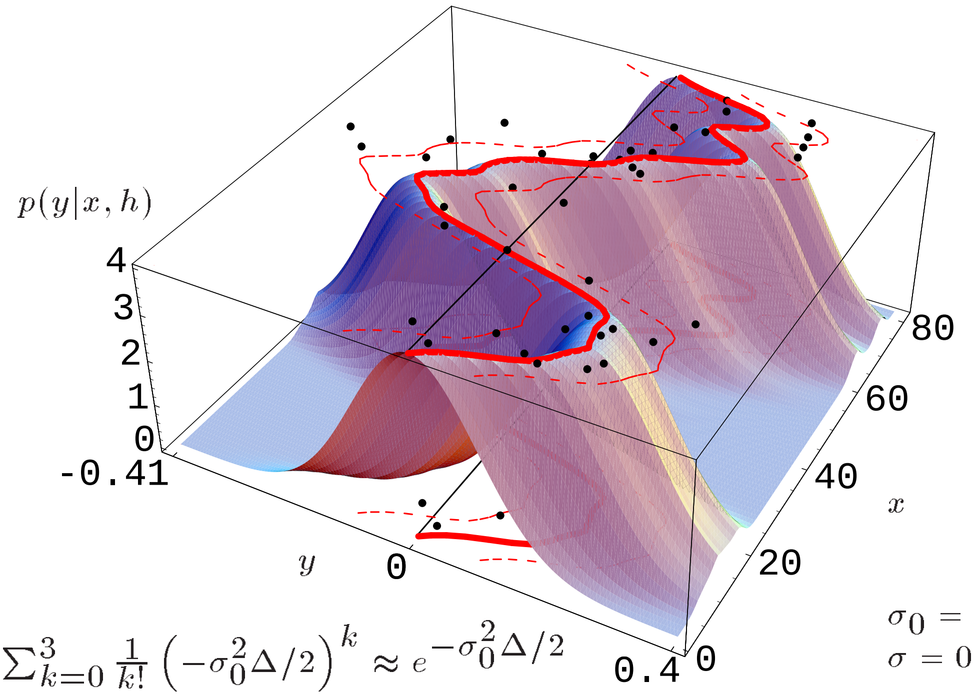
REL. GENERALISIERUNGSFEHLER $\tilde{E}_G = (E_G - \sigma^2)/(2\sigma^2)$





REL. TRAININGSFEHLER $E_T/(2n\sigma^2)$

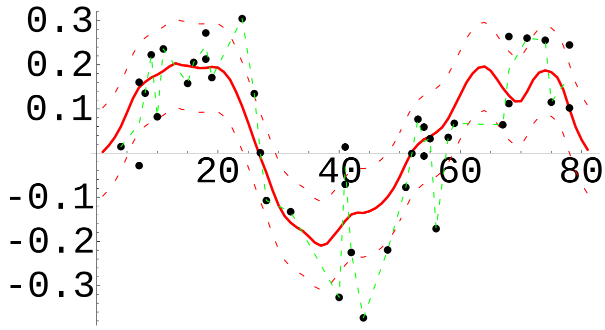
Rekonstruktion glatter Funktionen

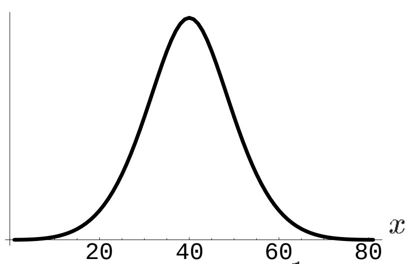


$$K_0 = \sum_{k=0}^{3} \frac{1}{k!} \left(-\sigma_0^2 \Delta/2\right)^k \approx e^{-\sigma_0^2 \Delta/2}$$

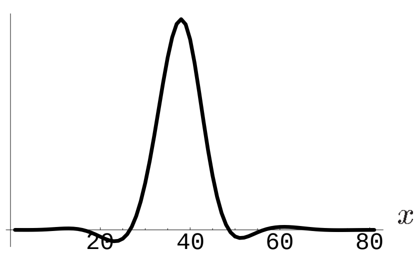
$\sigma_0 = 10$

$\sigma = 0.1$



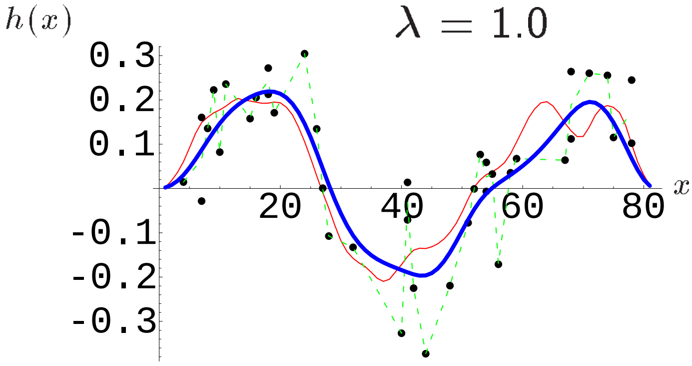


Kovarianz $K_0^{-1}(x, x_0)$

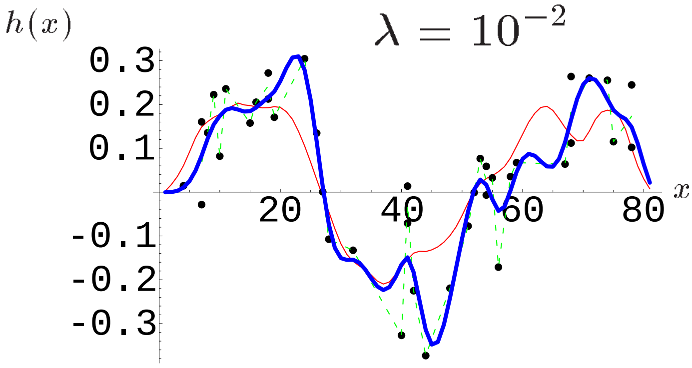


Äquival. Kern $K^{-1}(x, x_0)$

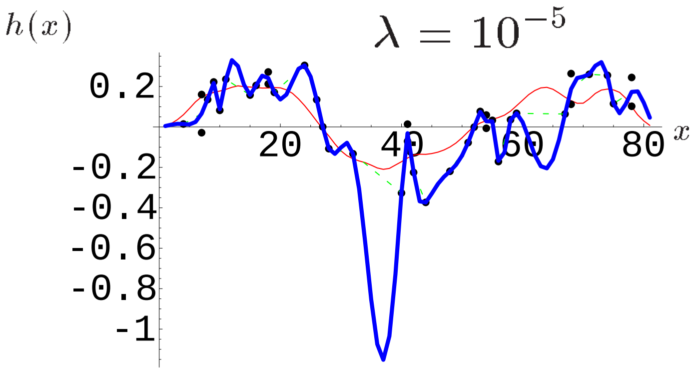
Rekonstruktion glatter Funktionen



optimal

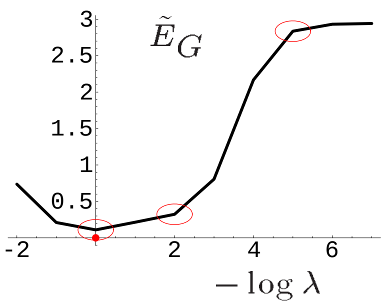


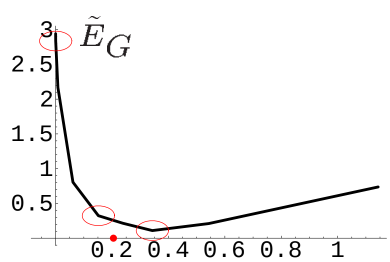
leichtes “overfitting”



starkes “overfitting”

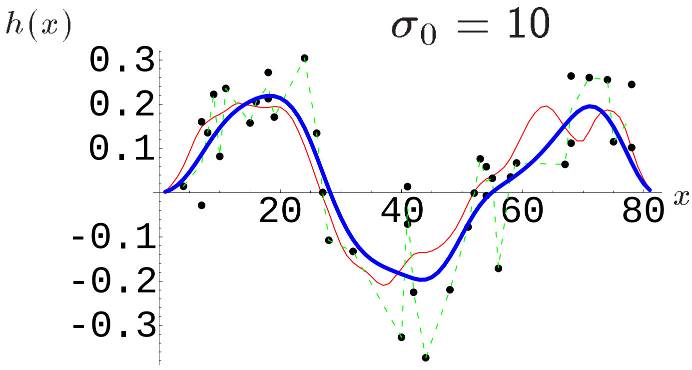
REL. GENERALISIERUNGSFEHLER $\tilde{E}_G = (E_G - \sigma^2)/(2\sigma^2)$



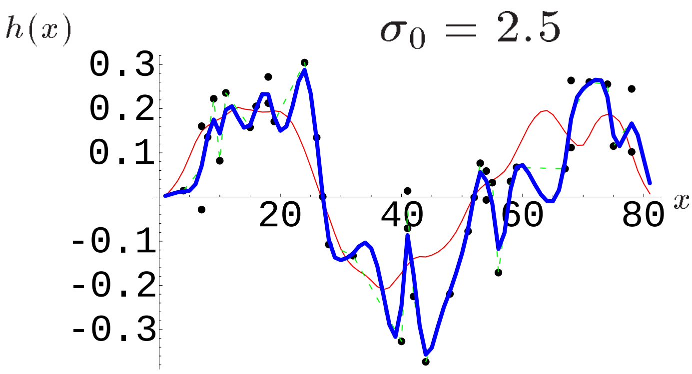


REL. TRAININGSFEHLER $E_T/(2n\sigma^2)$

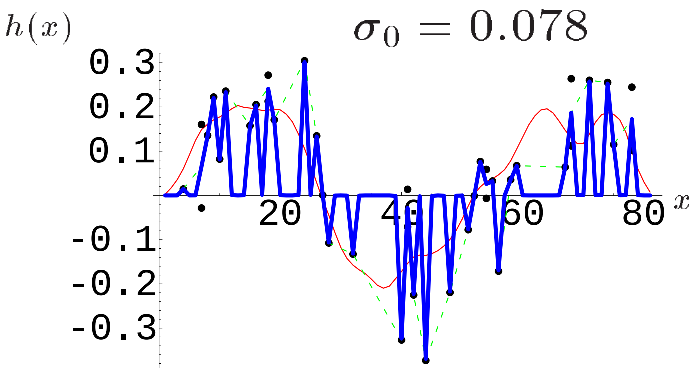
Rekonstruktion glatter Funktionen



Superposition von
breiten Gauß –Glocken

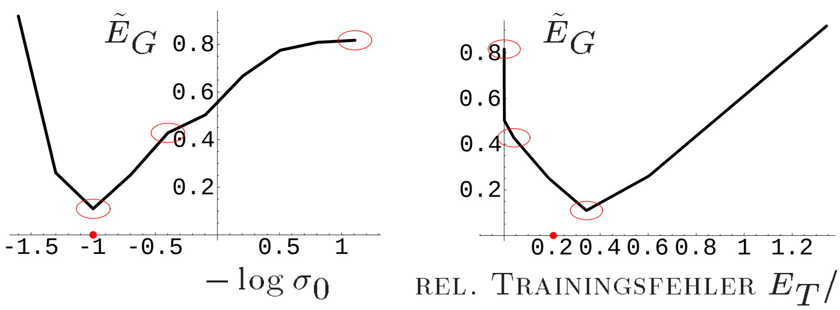


Superposition von
engen Gauß –Glocken



δ-Peaks

REL. GENERALISIERUNGSFEHLER $\tilde{E}_G = (E_G - \sigma^2)/(2\sigma^2)$



REL. TRAININGSFEHLER $E_T/(2n\sigma^2)$

Kovarianzen und approximative Symmetrien

Für eine Funktion $h(x)$, symmetrisch unter der Symmetrie-transformation $\mathbf{S}$ gilt

$$\mathbf{S}h = h \Rightarrow (\mathbf{I} - \mathbf{S})h = 0$$

$\Rightarrow$ Priorenergie zur Implementierung einer approximativen Symmetrie $\mathbf{S}$ durch einen Gauß–Prozess mit Kovarianz $\mathbf{K}_S^{-1}$

$$\frac{1}{2}\langle h\,|\mathbf{K}_S|\,h\rangle \quad \text{wobei} \quad \mathbf{K}_S = (\mathbf{I} - \mathbf{S})^T\,(\mathbf{I} - \mathbf{S})$$

Für eine Lie–Gruppe $\mathbf{S} = e^{\theta\,\mathbf{s}}$ findet man für eine approximative Symmetrie unter der infinitesimalen Transformation $\mathbf{K}_s = \mathbf{s}^T\mathbf{s}$.

Beispiele:

Approximative Symmetrie unter infinitesimalen Translationen

$$\mathbf{s} = \frac{d}{dx} \Rightarrow \mathbf{K}_s = -\Delta \quad \text{(neg. Laplaceoperator)}$$

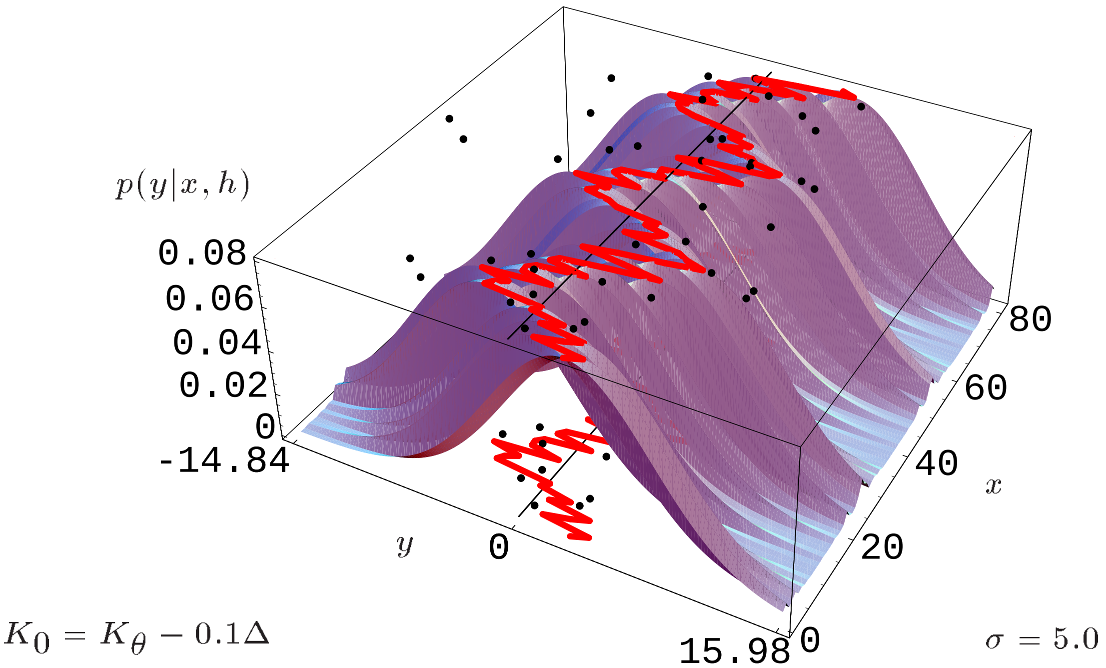
Approximative Periodizität

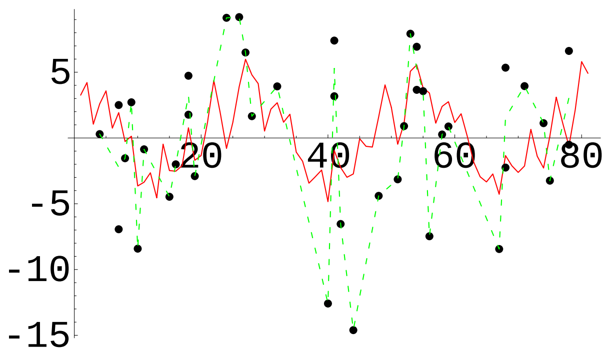
$$\mathbf{S}h(x) = h(x+\theta) \Rightarrow \mathbf{K}_S = -\Delta_\theta$$

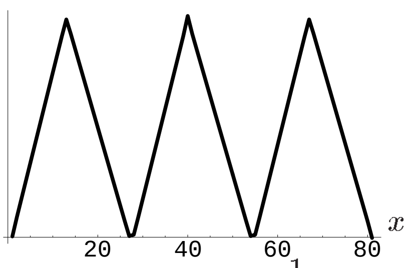
wobei (z.B. auf dem Gitter für eine Periodenlänge von $\theta = 2$)

$$-\Delta_\theta = \begin{pmatrix} 2 & 0 & -1 & 0 & -1 & 0 \\ 0 & 2 & 0 & -1 & 0 & -1 \\ -1 & 0 & 2 & 0 & -1 & 0 \\ 0 & -1 & 0 & 2 & 0 & -1 \\ -1 & 0 & -1 & 0 & 2 & 0 \\ 0 & -1 & 0 & -1 & 0 & 2 \end{pmatrix}$$

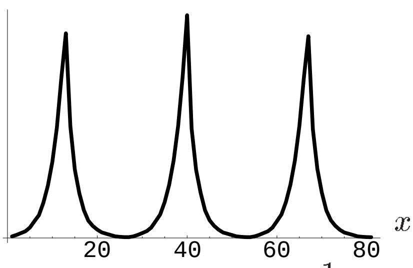
Rekonstruktion stetiger und approximativ periodischer Funktionen





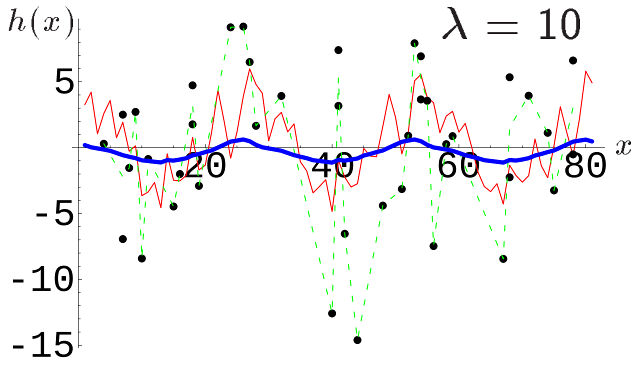


Kovarianz $K_0^{-1}(x, x_0)$

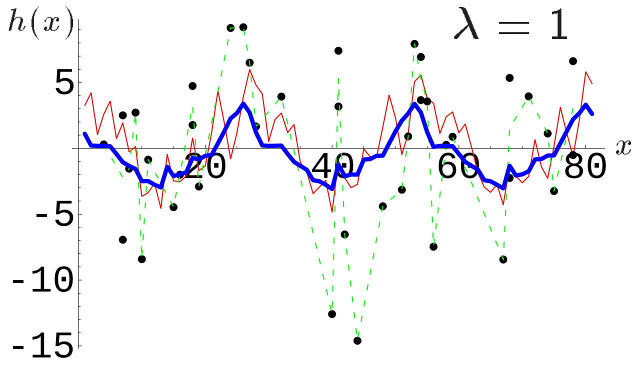


Äquival. Kern $K^{-1}(x, x_0)$

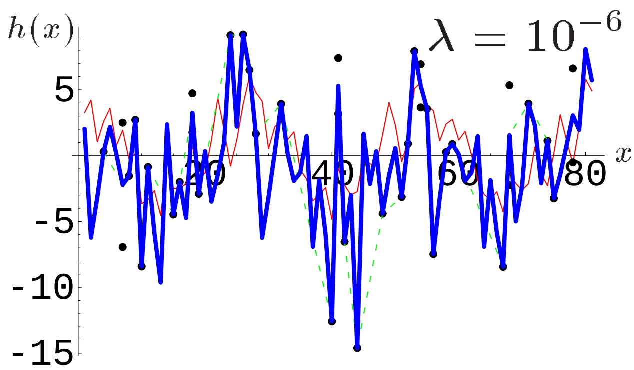
Rekonstruktion stetiger und approximativ periodischer Funktionen



zu glatt

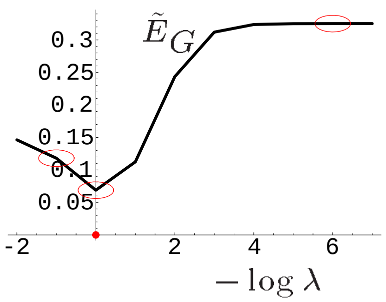


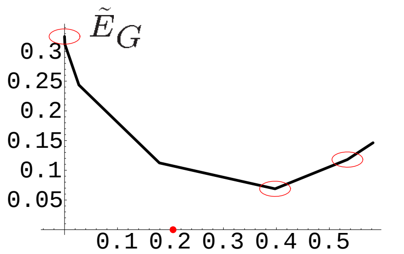
optimal



"overfitting"

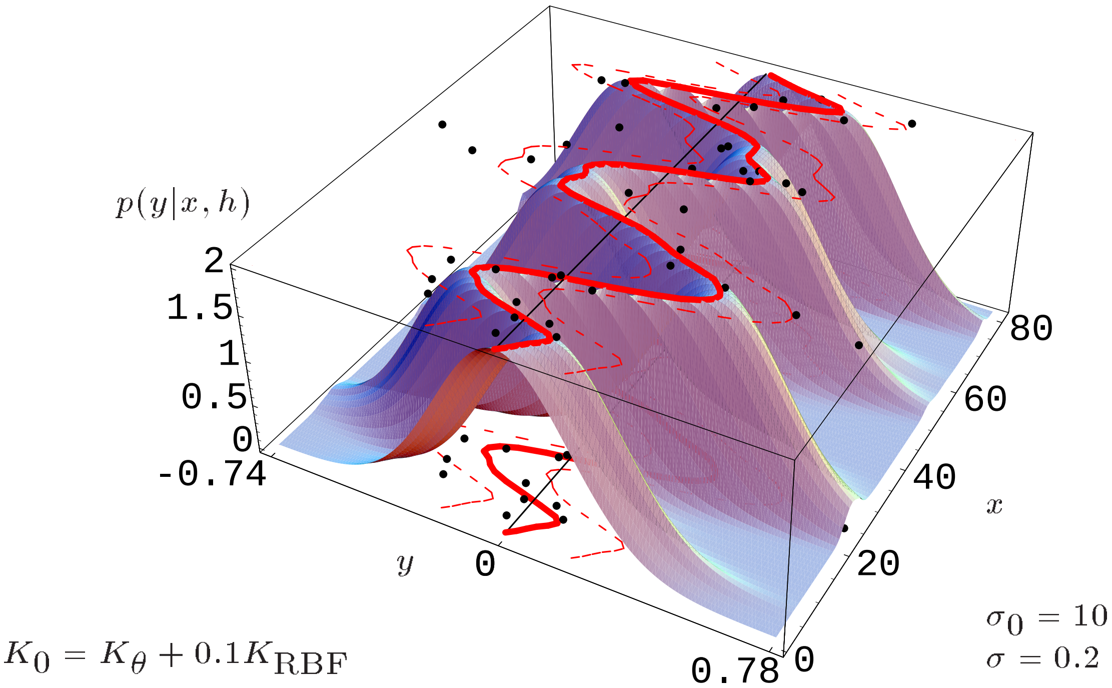
rel. Generalisierungsfehler $\tilde{E}_G = (E_G - \sigma^2)/(2\sigma^2)$

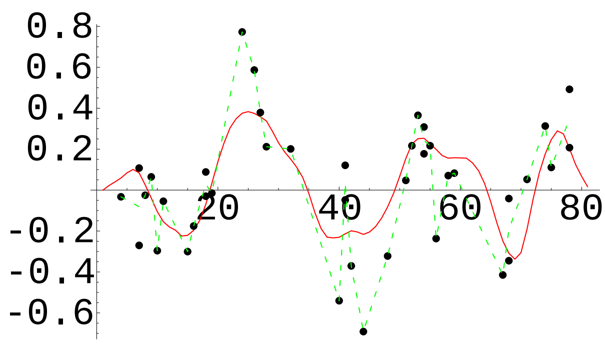


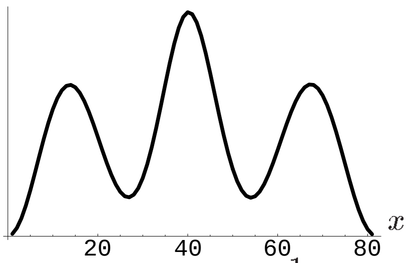


rel. Trainingsfehler $E_T/(2n\sigma^2)$

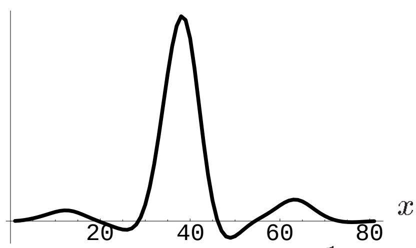
Rekonstruktion glatter und approximativ periodischer Funktionen





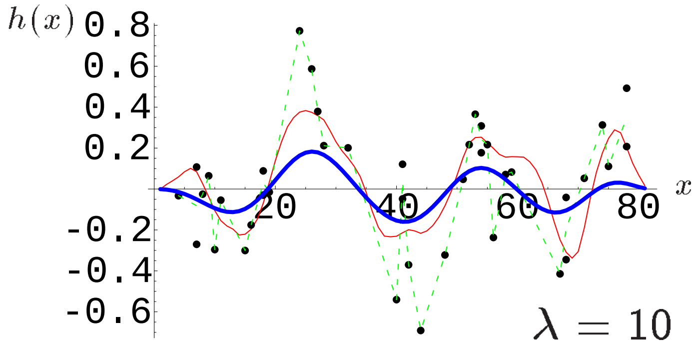


Kovarianz $K_0^{-1}(x, x_0)$

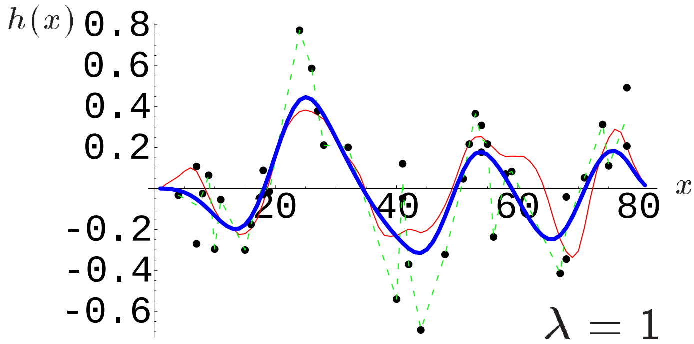


Äquival. Kern $K^{-1}(x, x_0)$

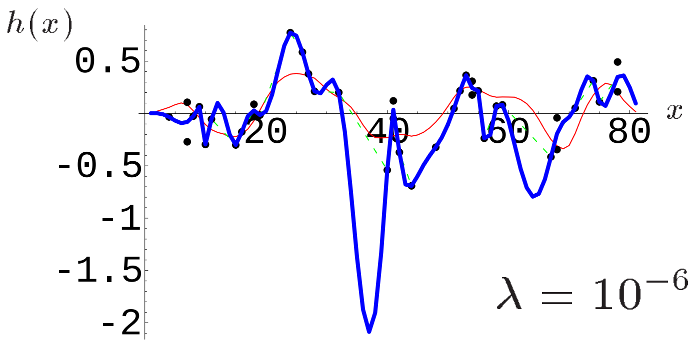
Rekonstruktion glatter und approximativ periodischer Funktionen



zu glatt

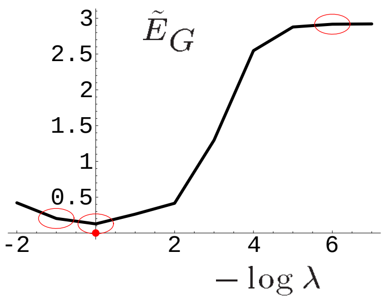


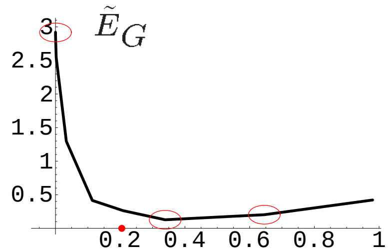
optimal



"overfitting"

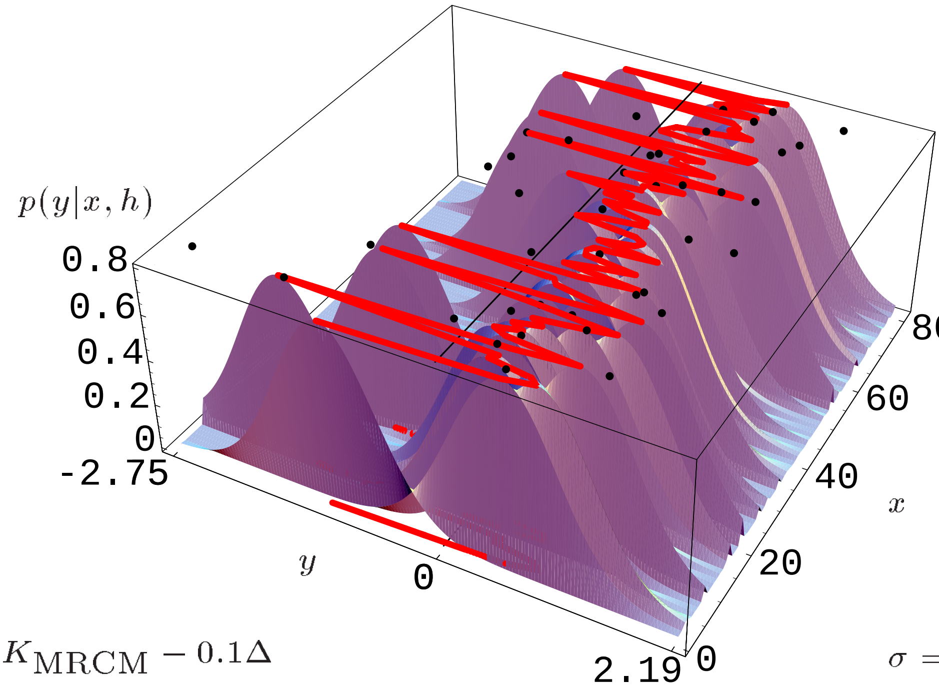
REL. GENERALISIERUNGSFEHLER $\tilde{E}_G = (E_G - \sigma^2)/(2\sigma^2)$





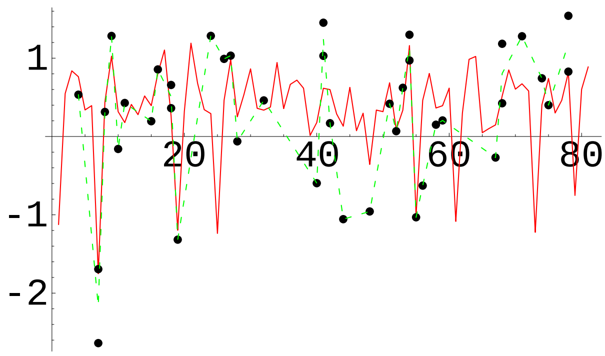
REL. TRAININGSFEHLER $E_T/(2n\sigma^2)$

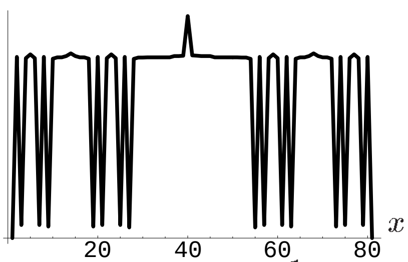
Rekonstruktion fraktaler Funktionen



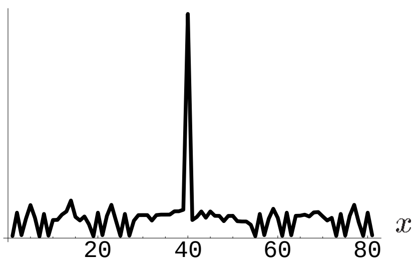
$K_0 = K_{\mathrm{MRCM}} - 0.1\Delta$

$\sigma = 0.5$



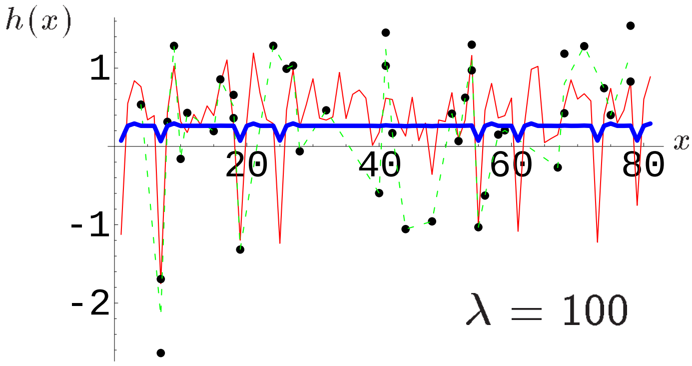


Kovarianz $K_0^{-1}(x, x_0)$

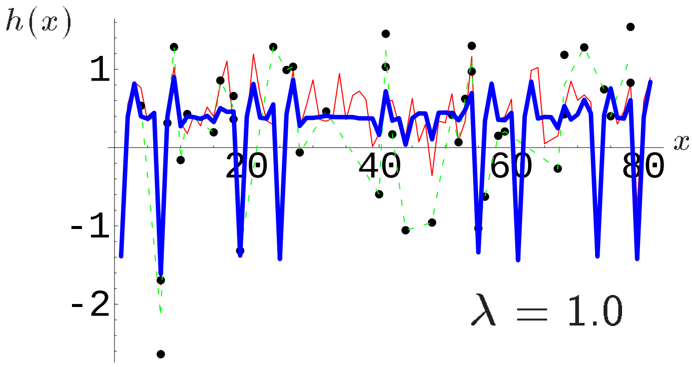


Äquival. Kern $K^{-1}(x, x_0)$

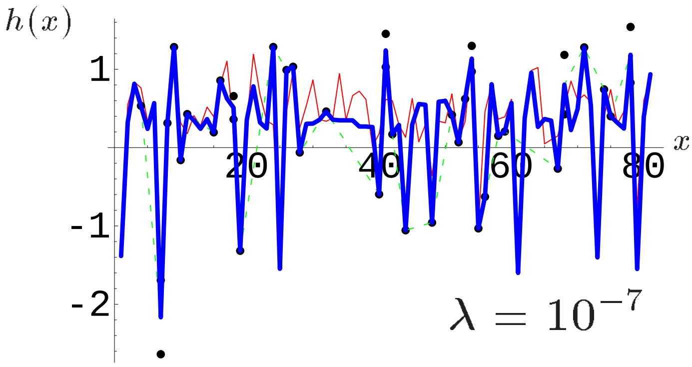
Rekonstruktion fraktaler Funktionen



zu glatt

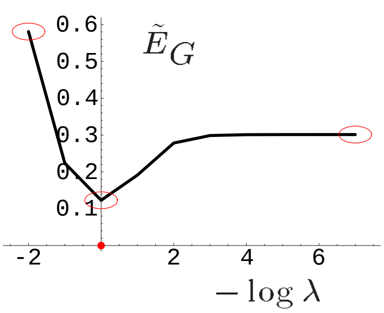


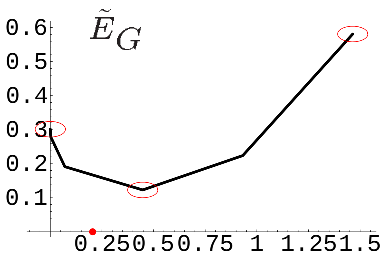
optimal



"overfitting"

rel. Generalisierungsfehler $\tilde{E}_G = (E_G - \sigma^2)/(2\sigma^2)$





rel. Trainingsfehler $E_T/(2n\sigma^2)$