Learning à la Bayes

— From image reconstruction to inverse quantum theory —

Part 1: Introduction

Jörg Lemm

Institut für Theoretische Physik
Universität Münster

lemm@uni-muenster.de

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

Computer–based empirical learning: Some examples

Empirical Learning = Ableitung allgemeiner Gesetzmäßigkeiten aus (einer endlichen Anzahl von) empirischen Beobachtungsdaten.

- NETtalk (Sejnowski, Rosenberg 1987)
- ZIP code recognition (LeCun 1989)
- car driving (Pomerleau 1989)
- optical character recognition, speech recognition,
- face detection
- medical diagnosis
- time series prediction (e.g., financial time series)
- credit risk evaluation, $\cdots$

Problem classes:

1. classification, pattern or object recognition
2. regression
3. density estimation

Inverse problems of quantum theory

Direct problem:

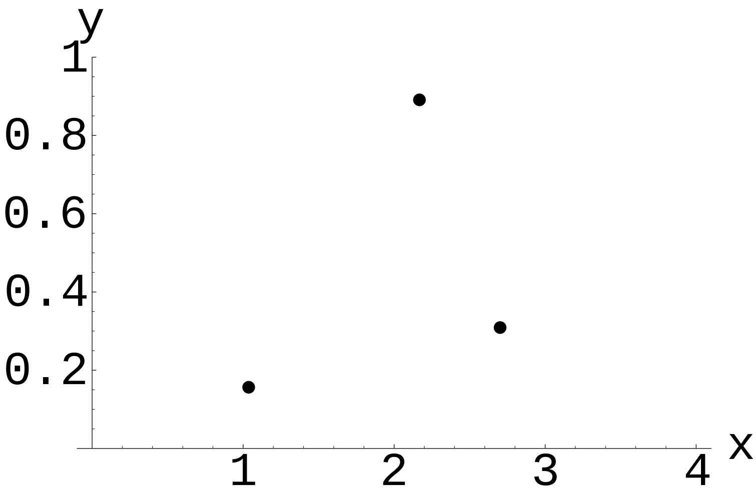
		Description of:
Hamiltonian		quantum statistical systems
$H = T + V$	$\to$	time–dependent quantum systems
given		many–body systems

Inverse problem:

Measuring:		
quantum statistical systems		Which potential
time–dependent quantum systems	$\to$	V
many–body systems		is the right one?

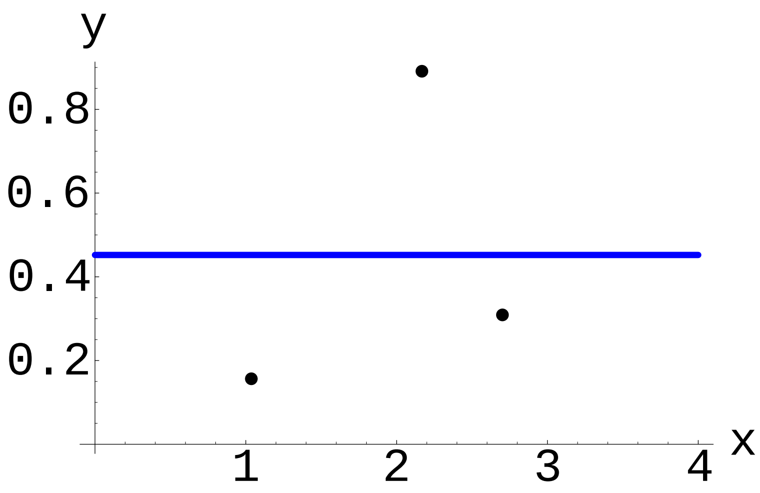
This requires an inverse quantum theory.

Three data points



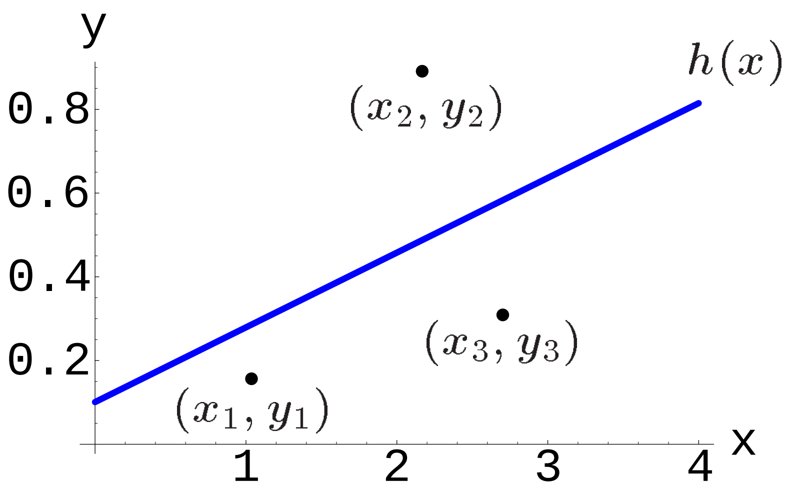
data: $D = \{(x_i, y_i) | 1 \leq i \leq n\}$

Fitting a constant



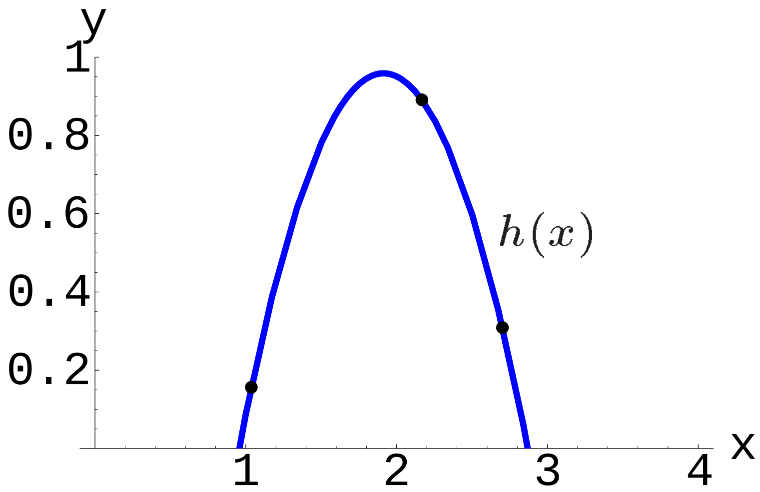
$$h(x) = a$$

Linear regression



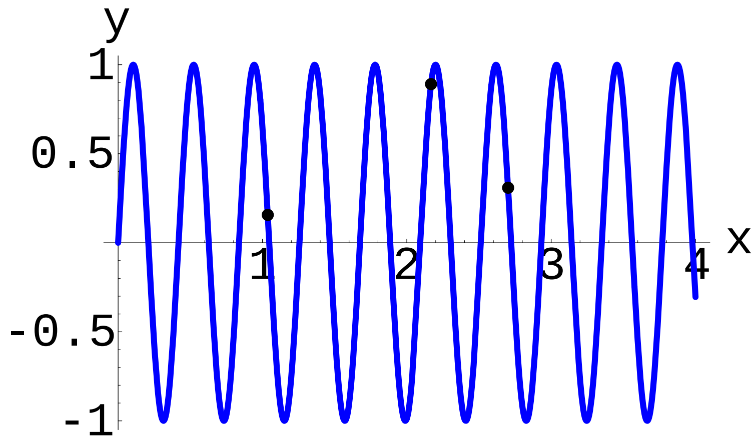
$$h(x) = a + bx$$

Lagrange's formula



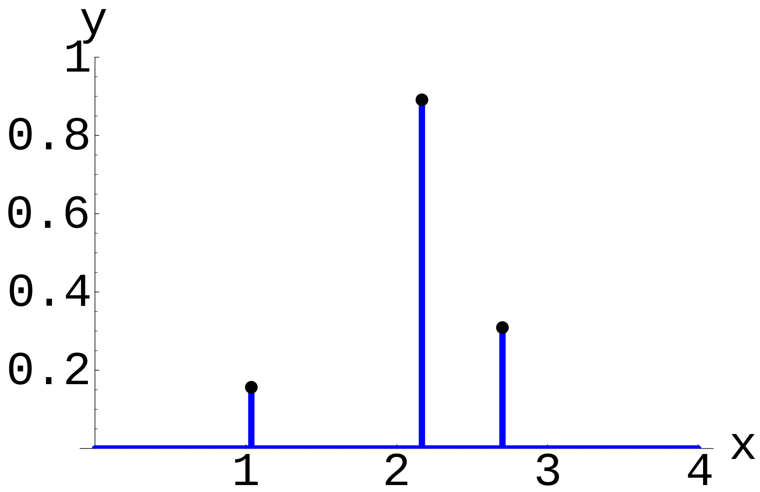
$$
\begin{aligned}
h(x) &= a + bx + cx^2 \\
&= y_1 \frac{(x - x_2)(x - x_3)}{(x_1 - x_2)(x_1 - x_3)} \\
&\quad + y_2 \frac{(x - x_1)(x - x_3)}{(x_2 - x_1)(x_2 - x_3)} \\
&\quad + y_3 \frac{(x - x_1)(x - x_2)}{(x_3 - x_1)(x_3 - x_2)}
\end{aligned}
$$

Reconstruction with 1 free parameter



$$h(x) = \sin(ax)$$

Another possibility



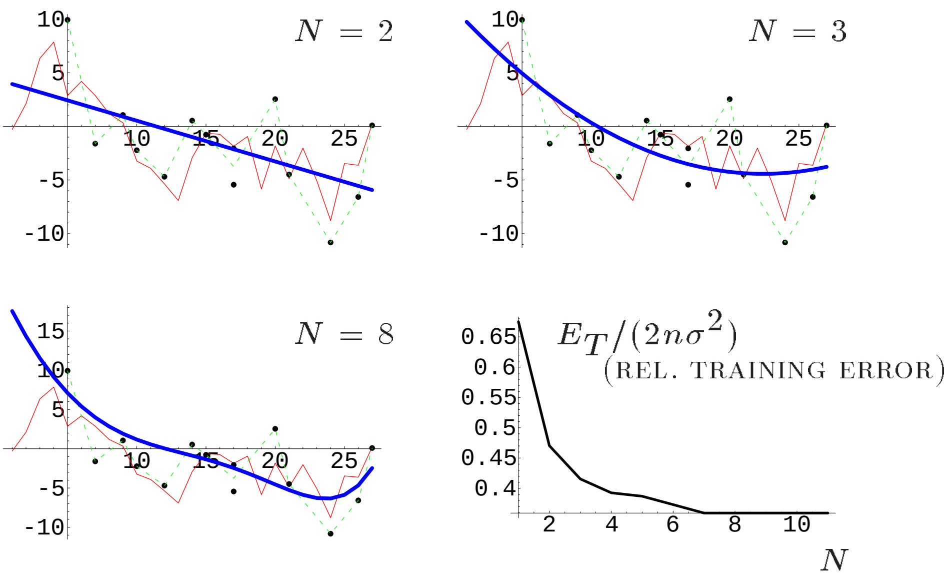
$$h(x) = \sum_{i=1}^{n} y_i \delta(x - x_i)$$

Parametric reconstruction

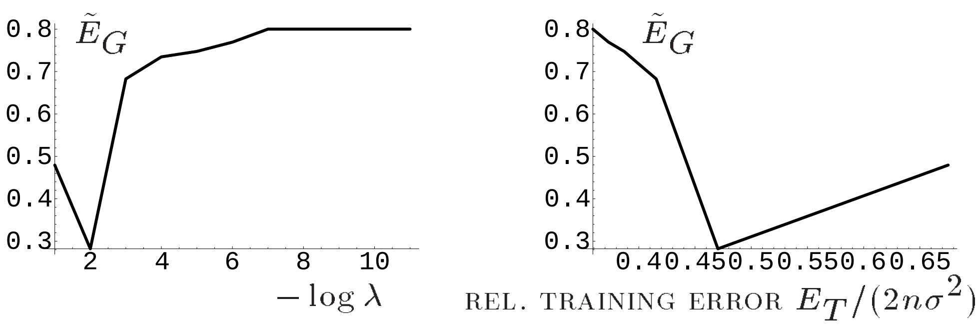
Minimize for training data $D = \{(x_i, y_i) | 1 \leq i \leq n\}$ the quadratic error (local minima):

$$\mathrm{argmin}_{\{\xi_k | 1 \leq k \leq N\}} \sum_{i=1}^{n} (y_i - h(x_i; \{\xi_k\}))^2$$

$$\text{e.g. : polynomial} \quad h(x; \{\xi_k\}) = \sum_{k=1}^{N} \xi_k x^{k-1}$$



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



What is empirical learning?

Given: Training data (e.g., time series):

$$h(x_1) = y_1 = 5 \quad , h(x_2) = y_2 = 7, \cdots$$

Task: Prediction of (future) test data

$$h(x_3) = ? \quad , h(x_4) = ?, \cdots$$

This requires to know dependencies between training and test data (a priori information, rules), like for example, smoothness :

$$|h(x) - h(x+1)| \le 1$$

Learning = Generalization of training data based on available a priori information:

$$\begin{aligned} 4 \le h(x_1+1) \le 6 \\ 3 \le h(x_1+2) \le 7 \end{aligned}$$

Number of different x values often very large, e.g., infinite $\Rightarrow$ (induction) problem:

How to measure a priori information?

Predictions based on specific a priori information can be checked with new data, e.g.:

$h(x_1+1) = 6$ agrees with above prior
$h(x_1+2) = 8$ contradicts above prior

Bayes' Theorem

Reverend Thomas Bayes 1702-1761

„Essay Towards Solving a Problem in the Doctrine of Chances“ (1763), published posthumously in the Philosophical Transactions of the Royal Society of London.

From the definition of a conditional probability

$$p(A, B) = p(A)p(B|A) = p(B)p(A|B)$$

follows for $A = D$ ("data") and $B = h$ ("hypotheses")

$$\underbrace{p(h|D)}_{\text{unknown}} = \frac{\overbrace{p(D|h)}^{\text{known}}\, p(h)}{p(D)} \qquad \text{(Bayes' Theorem)}$$

or, in words

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

Prediction of new data (y, x) with

$$p(y|x, D) = \int dh\, p(y|x, h)\, p(h|D) \underset{\text{MAP}}{\approx} p(y|x, h^*)$$

A fair coin?

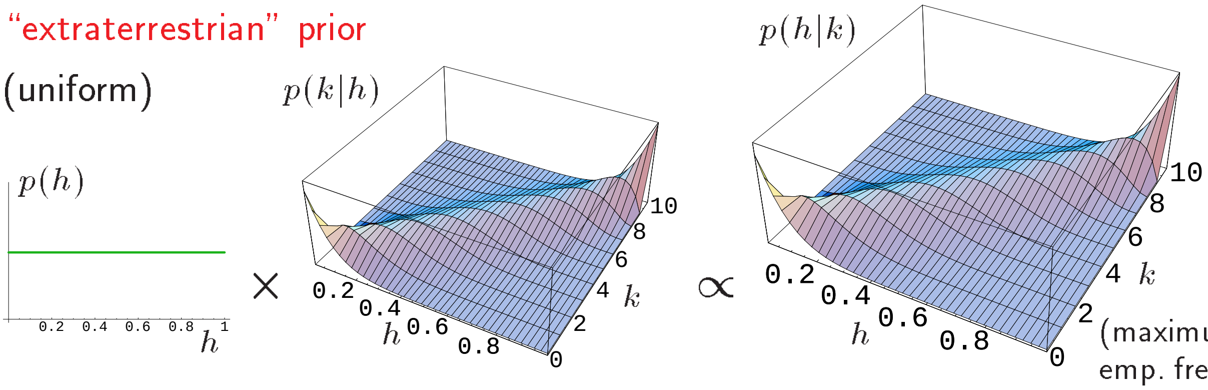
$h = p(K)$ probability to find "head"
k denotes the number of "heads" found in $n = 10$ trials (corresponds to data D).

Likelihood: $p(D|h) = p(k|h) = \binom{n}{k} h^k (1-h)^{(n-k)}$

prior $p(h)$ $\times$ likelihood $p(D|h)$ $\propto$ posterior $p(h|D)$

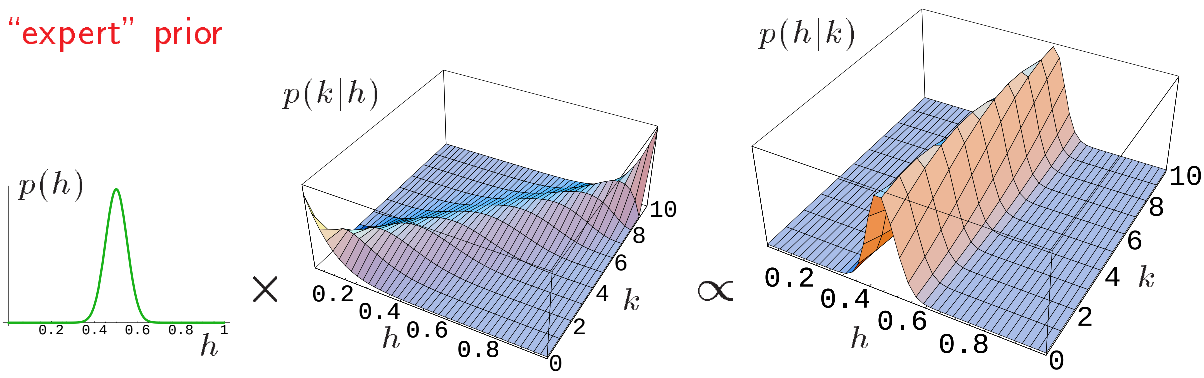
"extraterrestrian" prior

(uniform)

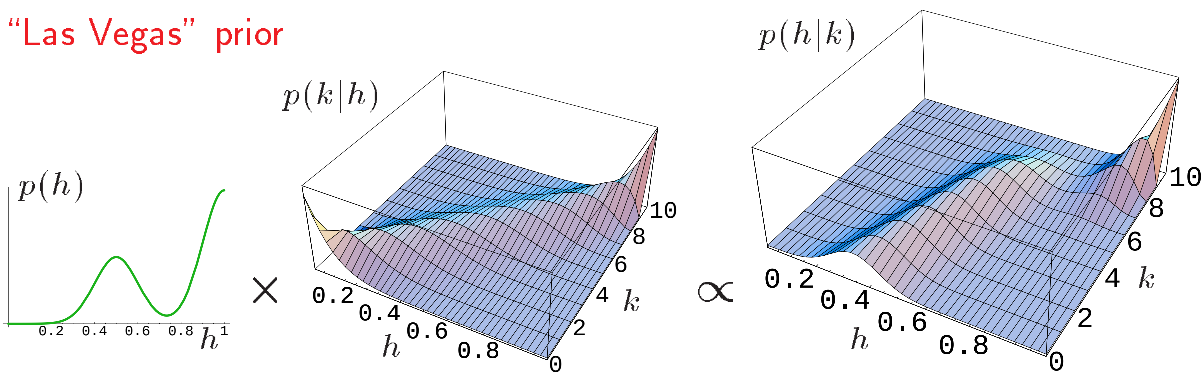


(maximum at emp. frequency)

"expert" prior



"Las Vegas" prior



Likelihood and prior models

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

likelihood model	prior model
describes	
measuring process	generalization
determined by	
parameter	hyperparameter
examples are	variants are
regression classification density estimation inverse quantum theory	Gaussian processs mixture of Gaussian priors Non–quadratic potentials hard constraints