Bayesian Methods

— From parametric and nonparametric regression, classification, and density estimation to inverse quantum theory—

Jörg Lemm

Institut für Theoretische Physik
Universität Münster

Email: lemm@uni-muenster.de

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

Bayesian Statistics and Physics

- Physics needs methods for constructing general laws from a finite number of observational data (empirical learning, induction problem)
- Methods of Bayesian statistics are closely related to statistical mechanics (Statistical field theory $\rightarrow$ Bayesian field theory)

Computer–based empirical learning: Some examples

Empirical Learning = constructing general laws from a finite number of observational data

- 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, ...

Closely related to inverse (ill-posed) problems, like

- X-ray computer tomography
- infering the earth subsurface properties from reflection seismic data
- analysis of observed spectra
- reconstruction of noisy images, ...

General problem classes:

1. classification, pattern or object recognition
(integer variables)
2. regression (reconstructing the mean function for distribution of known form)
3. density estimation (most general, unrestricted case)

Inverse problems of quantum theory

Direct problem:

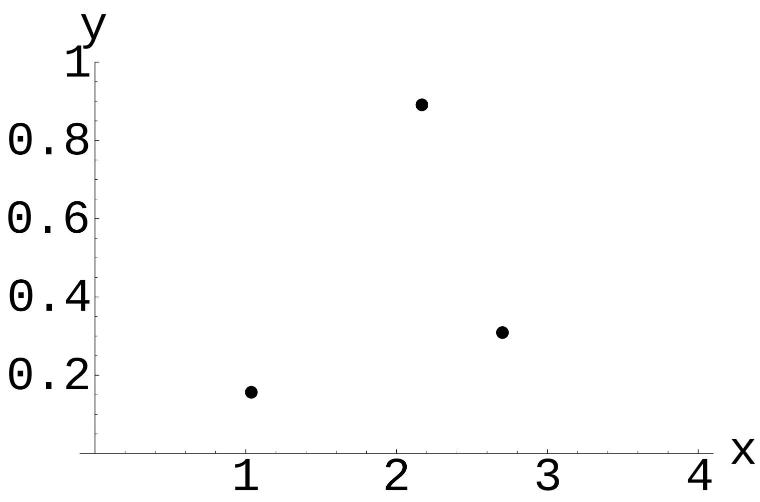
		Description of:
Hamiltonian		quantum statistical systems
$H = T + V$	$\rightarrow$	time–dependent quantum systems
given		many–body systems

Inverse problem:

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

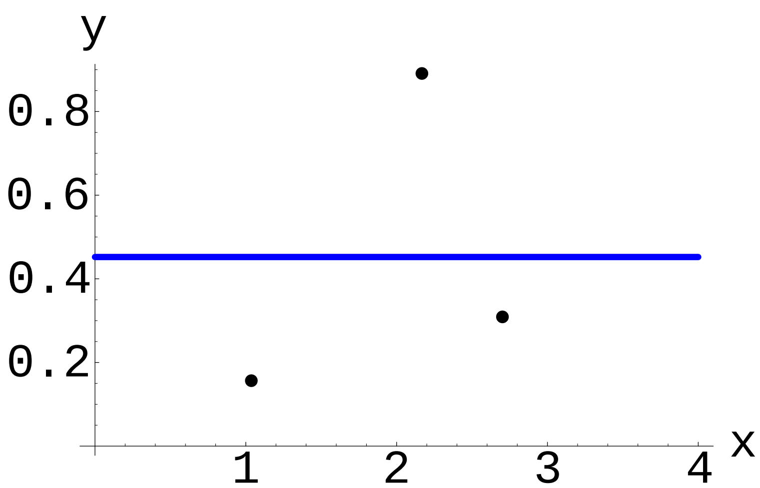
This requires an inverse quantum theory.

Three data points



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

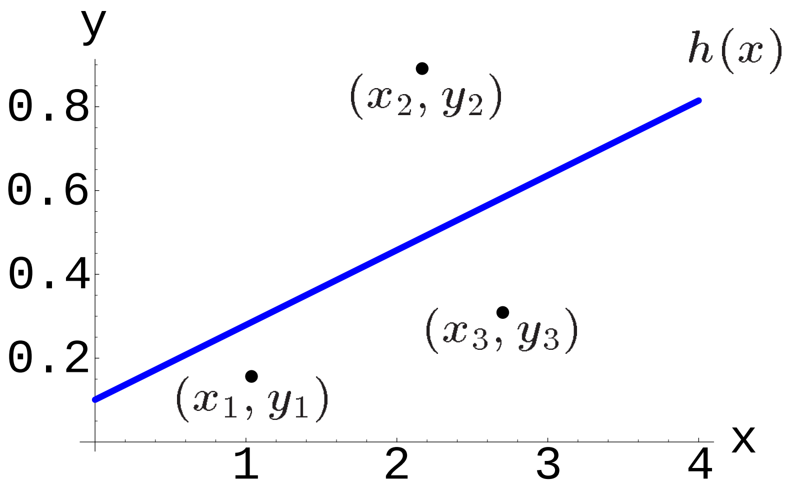
Fitting a constant



Ignoring the x variable:

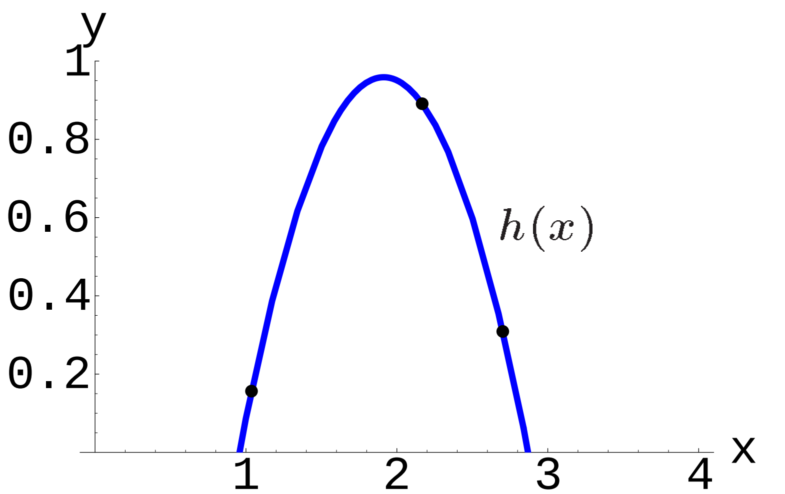
$$h(x) = a$$

Linear regression



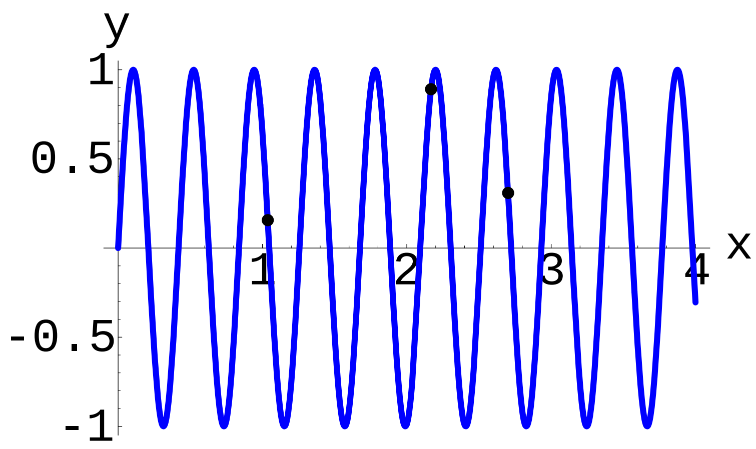
$$h(x) = a + bx$$

Lagrange's formula



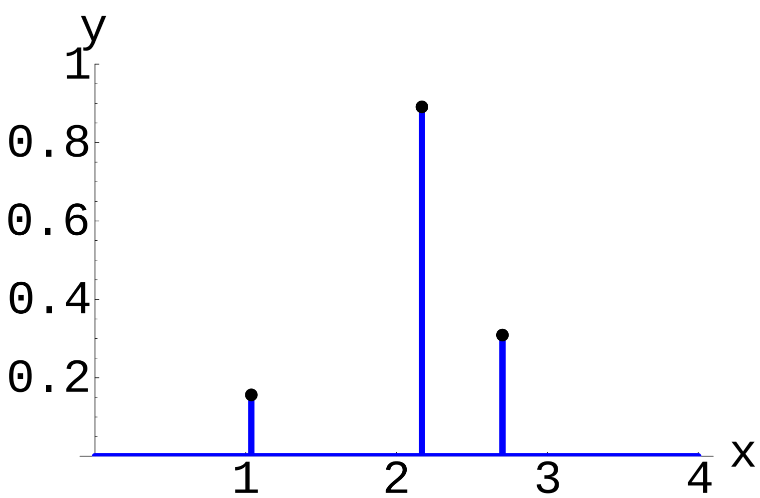
$$\begin{aligned} h(x) &= a + bx + cx^2 \\ &= y_1 \frac{(x - x_2)(x - x_3)}{(x_1 - x_2)(x_1 - x_3)} \\ &\quad + y_2 \frac{(x - x_1)(x - x_3)}{(x_2 - x_1)(x_2 - x_3)} \\ &\quad + y_3 \frac{(x - x_1)(x - x_2)}{(x_3 - x_1)(x_3 - x_2)} \end{aligned}$$

Reconstruction with 1 free parameter



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

Yet another possibility



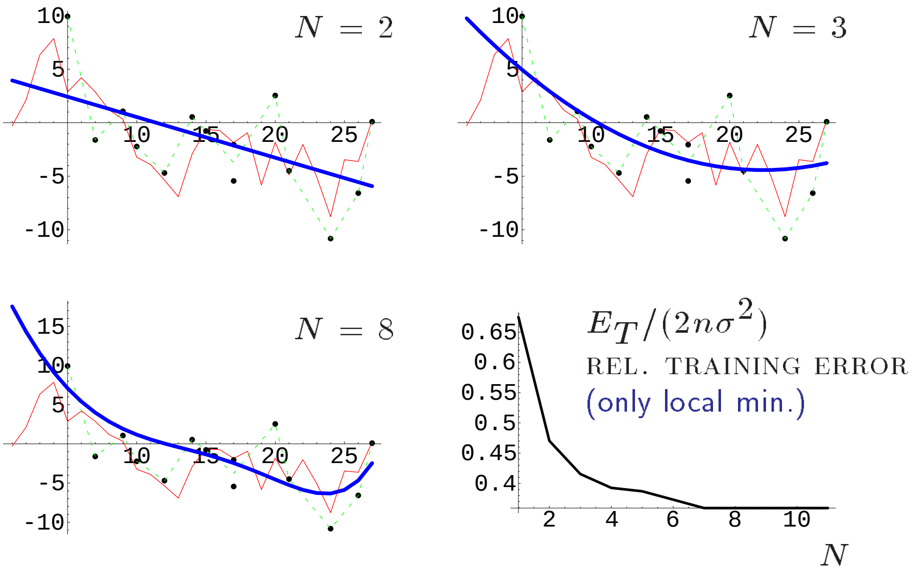
$$h(x) = \sum_{i=1}^{n} y_i \delta(x - x_i)$$

Parametric reconstruction

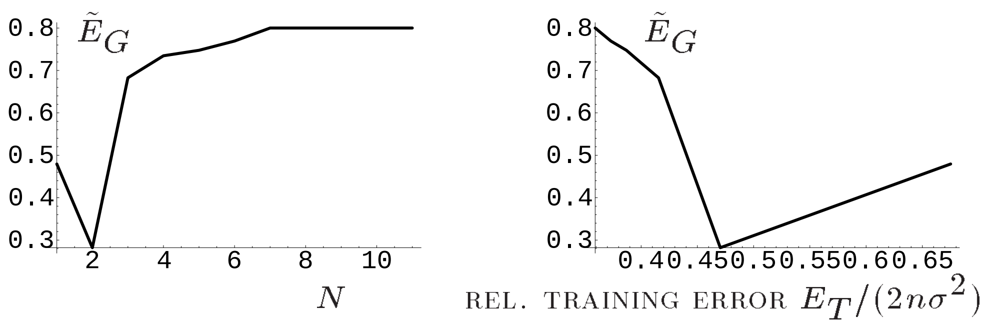
Minimize for training data $D = \{(x_i, y_i) | 1 \leq i \leq n\}$ the quadratic error (there may by many local minima!):

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

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



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



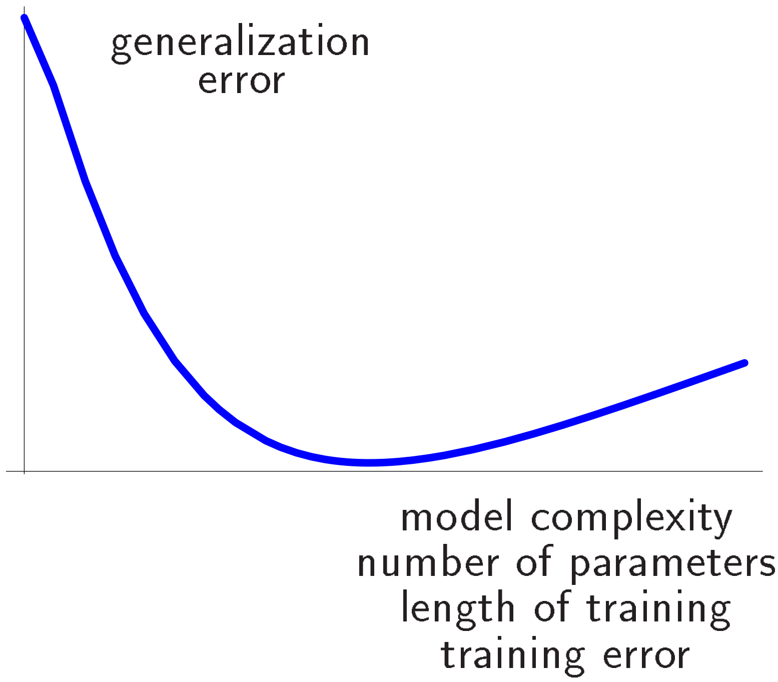
Training error $E_T = \sum_i |y_i - h(x_i)|^2$

Generalization error

$E_G = \int dx\, dy\, p(x) p(y|x, h_{\text{true}})\, |y - h(x)|^2$

Overfitting

Adding more degrees of freedoms to a learning model reduces the training error, e.g.
$E_T = \sum_{i=1}^{n} (y_i - h(x_i))^2$,
but does not reduce necessarily the generalization error, like for instance
$E_G = \int dx\, dy\, p(x)p(y|x, h_{\text{true}})\, (y - h(x))^2$
which measures also the error for non–training data



Crossvalidation–like techniques:

1. Use training set to train models with different complexity
2. Use (an independent) test set to select optimal model

3 Use (still another independent) validation set to guess the true expected generalization error

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 :

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

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

$$4 \leq h(x_1+1) \leq 6$$
$$3 \leq h(x_1+2) \leq 7$$

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")

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

or, in words

$$\underbrace{\text{Posterior}}_{\text{`inverse probability'}} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

Prediction of new data (y,x) with

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

Laplace

Pierre-Simon Laplace
23 mars 1749 - 5 mars 1827
Theorie Analytique des Probabilités (1812)

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

Some basic formulas of probability theory

OR:

$$p(A+B) = p(A) + p(B) - p(A, B)$$

AND:

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

Certainty: $p(A) = 1$, A is certainly true

Falsity: $p(A) = 0$, A is certainly false

Constraints: (non–negativity and normalization)

$$p(A) \geq 0 \quad \text{and} \quad \sum_i p(A_i) = 1$$

for exclusive and exhaustive A_i

Marginalization: ("extending the conversation")

$$p(A) = \sum_i p(A, B_i) = \sum_i p(A|B_i)p(B_i)$$

Graphical models

Graphical representation of the independence structure of a joint probability model

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

$$\begin{tikzcd} A \arrow[r] & B \end{tikzcd} \qquad \begin{tikzcd} A & B \arrow[l] \end{tikzcd}$$

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

$$\begin{tikzcd} & B \arrow[dr] & \\ A \arrow[ur] \arrow[rr] & & C \end{tikzcd}$$

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

$$\begin{tikzcd} & B \arrow[dr] & \\ A \arrow[rr] & & C \end{tikzcd}$$

(Pearl, Lauritzen, Jordan, . . .)

Basic variables

1. **hidden variables** (not directly measurable)

$$h \in \mathcal{H}$$

   ($model\ states,\ possible\ state\ of\ Nature$)

2. **visible variables**
   a. **independent variables** (describing kinds of measurement)

$$x \in \mathcal{X}$$

   b. **dependent variables** (responses)

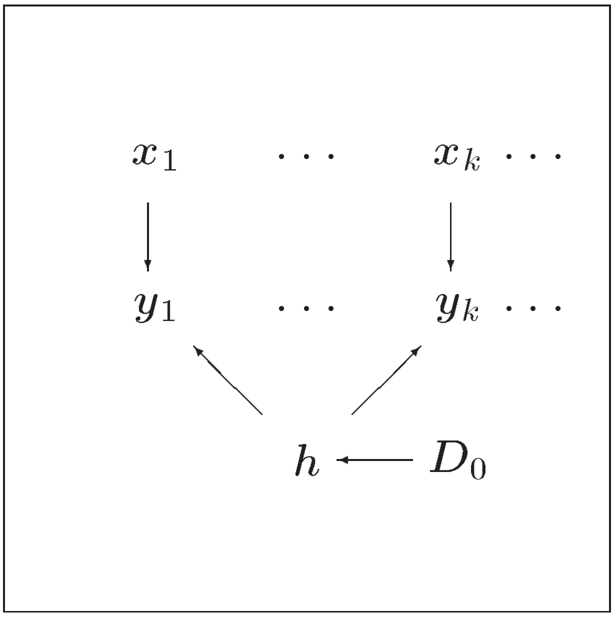
$$y \in \mathcal{Y}$$

   c. **fixed variables** ($a\ priori$ information)

$$D_0$$

   (determining prior probability $p(h|D_0)$)
   If D_0 is considered constant it is sufficient to write $p(h)$.

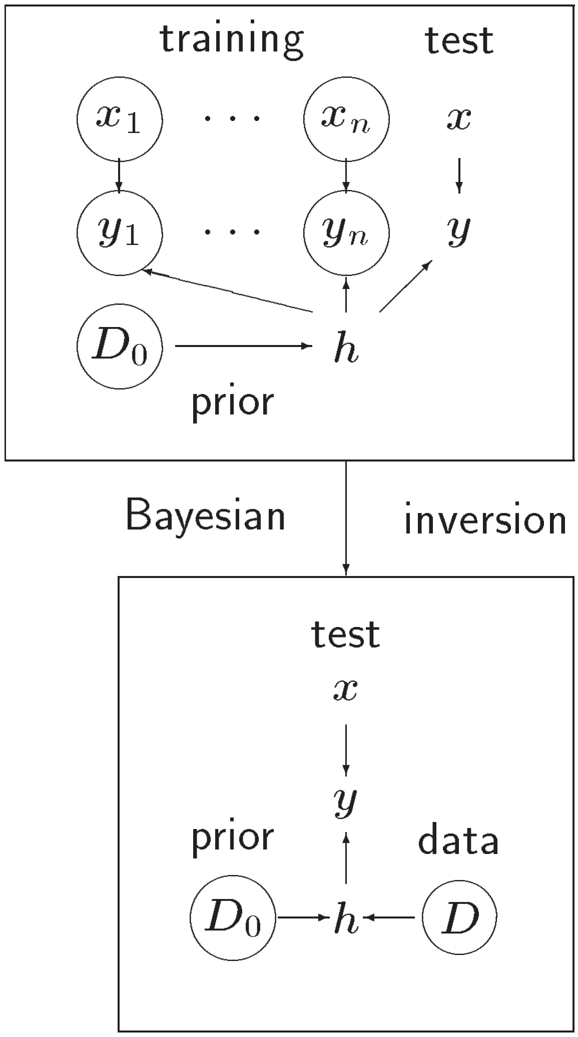
Basic model



Graphical representation of a probabilistic model with independent data, factorizing according to

$$p(x, y, h|D_0) = p(h|D_0) \prod_k p(x_k) p(y_k|x_k, h).$$

Learning

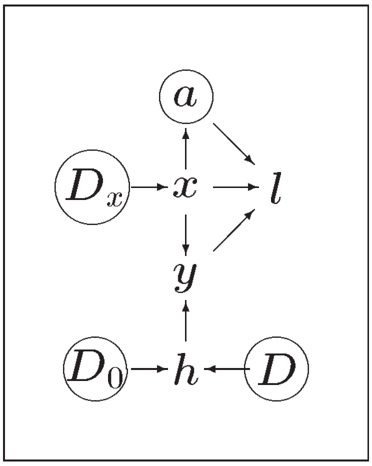


Given training data $D = \{(x_i, y_i)|1 \leq i \leq n\}$ with likelihood $p(y|x, h)$ and *a priori* information $p(h)$ calculate posterior $p(h|D)$ to obtain predictive density $p(y|x, D)$

Bayesian Decision Theory

Assume we have to choose in situation x between possible **actions** $a(x) \in \mathcal{A}$ and are able to define a **loss function** (loss suffered if action a is chosen in situation x and outcome y occurs) $l(x, y, a(x))$
E.g. for density estimation: $-\ln p(y|x, a)$ for regression: $|y - a(x)|^2$ or $|y - a(x)|$, for classification $1 - \delta_{y,a(x)}$.

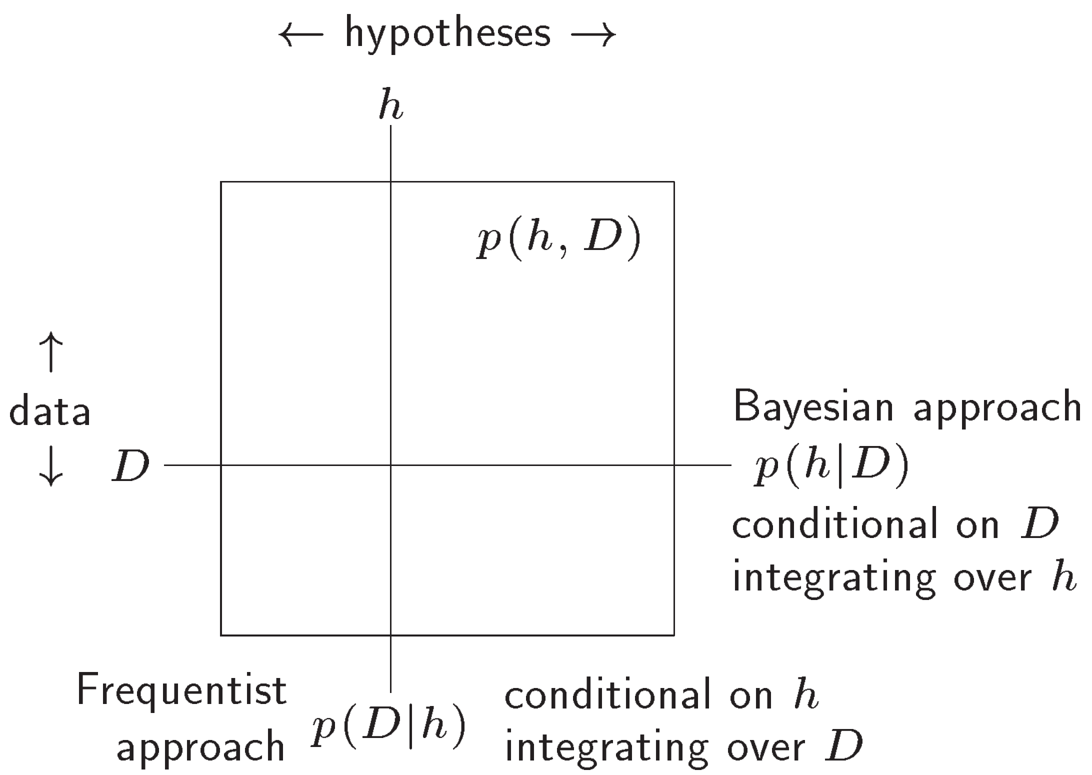
Aim: Minimize functional of $p(l|a, D)$, e.g. **expected risk**
$r(a, D) = \int dl\; l\; p(l|a, D)$
$= \int dh\, dx\, dy\, l(x, y, a(x))\, p(x)\, p(y|x, h)\, p(h|D)$



deterministic loss $p(l|x, y, a) = \delta(l - l(x, y, a))$
deterministic action $p(a|x) = \delta(a - a(x))$
loss posterior under action a: $p(l|a, D) =$
$\int dh\, dx\, dy\, p(l|x, y, a(x))\, p(x|D_x)\, p(y|x, h)\, p(h|D)$

expected risk (= expected loss) $r(a, h) =$
$\int dh\, dx\, dy\, l(x, y, a(x))\, p(x|D_x)\, p(y|x, h)\, p(h|D)$

Bayesian vs. Frequentist Approach



Frequentist approaches characterize a specific hypothesis h_0, calculating for example the probability to observe more extreme data values than the observed D given h_0 (P-values). Frequentist data space integrals $\int dD$ factorize for independent data where $p(D|h_0) = \prod p(x_i)p(y_i|x_i, h_0)$.

Bayesian approaches focus on solving the problem of making the best decision in a specific situation, i.e., for given observational data D. The Bayesian integral $\int dh$ over all hypotheses does usually not factorize because $p(h|D) \neq \prod_i p(h|x_i, y_i)$. The integral can be very high dimensional, for nonparametric models it is even a functional integral. Some low dimensional Bayesian models can be calculated analytically, high dimensional integrals can be tackled by Monte Carlo integration methods. A typical approximation for complex models is the maximum a posteriori approximation (MAP).