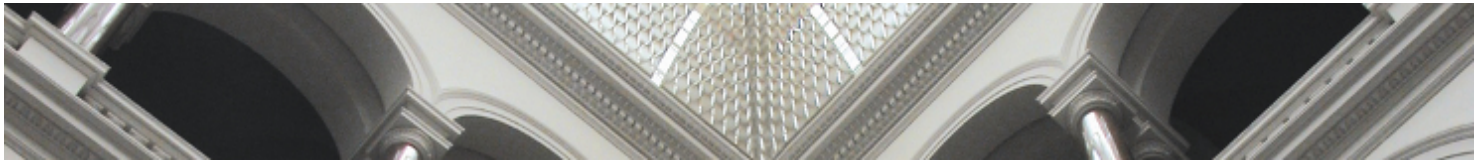
Technische Universität Berlin


Bayesian Learning

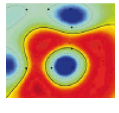
Sergej Dogadov
s.dogadov@tu-berlin.de
Technische Universität Berlin - Machine Learning Group
Beginners Workshop Machine Learning 2019

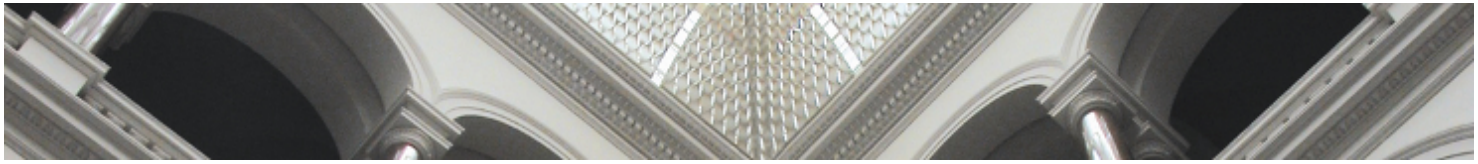


Technische Universität Berlin

Agenda

- First Part
  - Frequentist vs. Bayesian Approaches
  - Introduction to Probability Theory
  - Bayes' Theorem
  - Naive Bayes Classifier
  - Linear Regression
- Second Part
  - Bayesian Linear Regression
  - EM Algorithm
  - Variational Linear Regression
  - Variational Inference

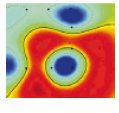


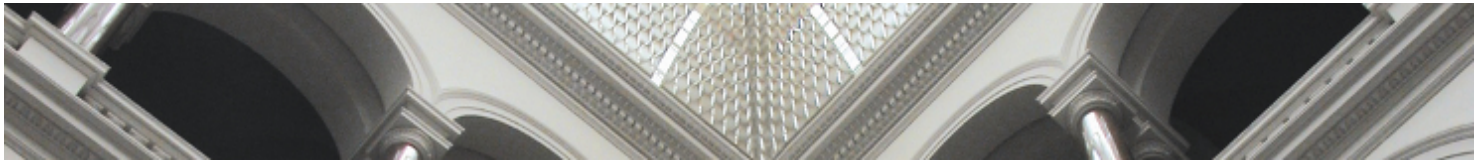


Technische Universität Berlin

Frequentist vs. Bayesian Approaches

Search for a missing base phone





Technische Universität Berlin

Search for a missing base phone

Frequentist Approach

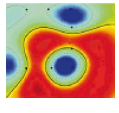
- Hear the phone beeping
- Identify an area only by the beeping sound
- Search for the phone

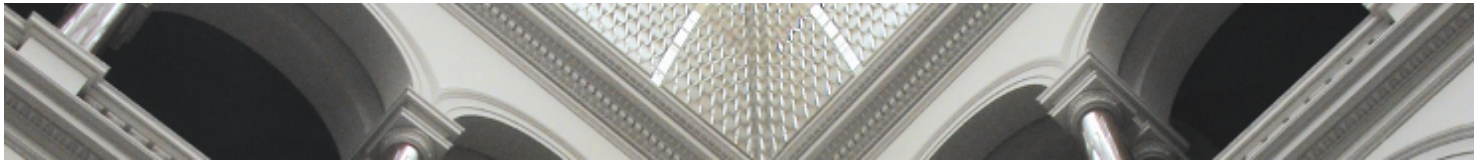
Bayesian Approach

- Hear the phone beeping
- Identify an area by the sound and the locations where I could misplace the phone in the past

Bayesian approach incorporates the prior knowledge into the model in order to enhance the inference over the model's latent variables. In this case the location of the missing phone.

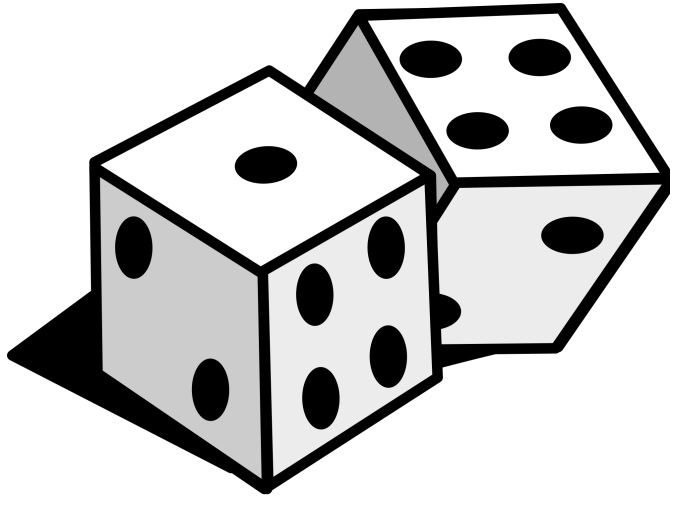
Example from: https://stats.stackexchange.com/questions/22/bayesian-and-frequentist-reasoning-in-plain-english

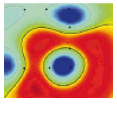


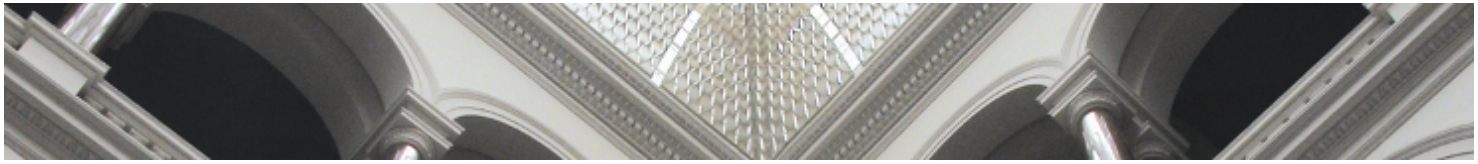


Technische Universität Berlin

Introduction to Probability Theory







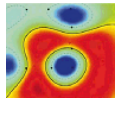
Technische Universität Berlin

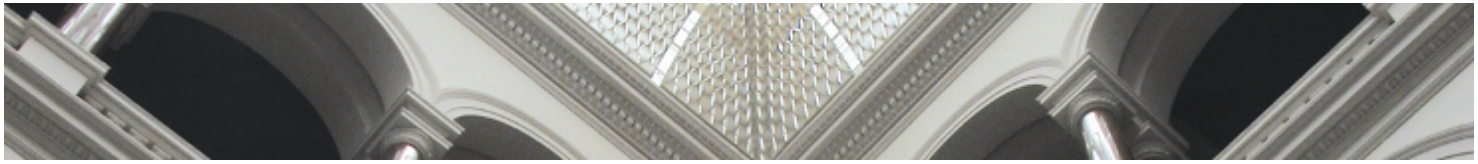
Probability Theory for reasoning under uncertainty

In many settings, we must try to understand what is going on in a system when we have imperfect or incomplete information.

- Two reasons why we might reason under uncertainty:
  1. *laziness* (modeling every detail of a complex system is costly)
  2. *ignorance* (we may not completely understand the system)

- Probabilities quantify uncertainty regarding the occurrence of events.

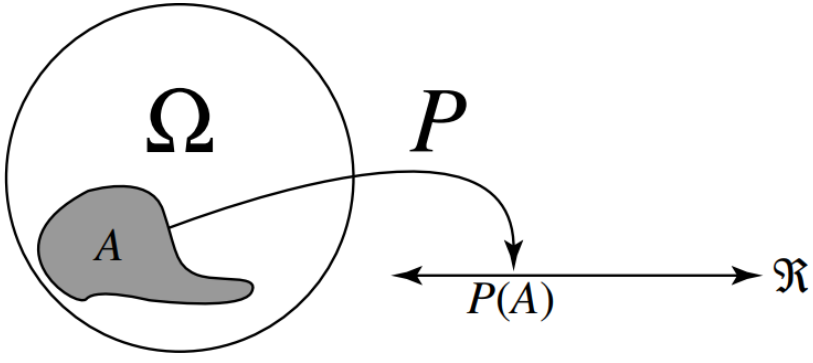




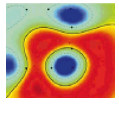
Technische Universität Berlin

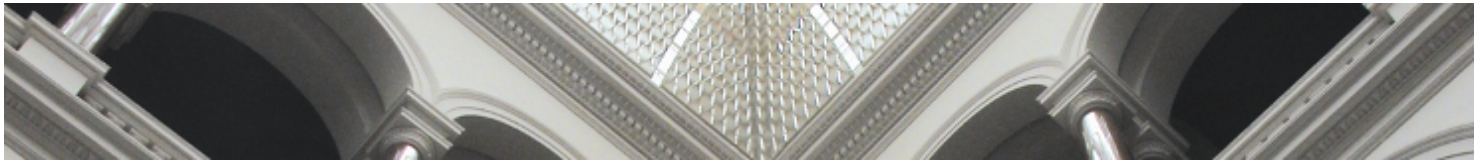
Probability spaces

- A probability space represents our uncertainty regarding an experiment.
- It has two parts:
  1. the sample space Ω is a set of *outcomes* and
  2. the probability measure P, which is a real function of the subsets of Ω.



- A set of outcomes $A \subseteq \Omega$ is called an event. $P(A)$ represents how likely it is that the experiment's actual outcome will be a member of A.

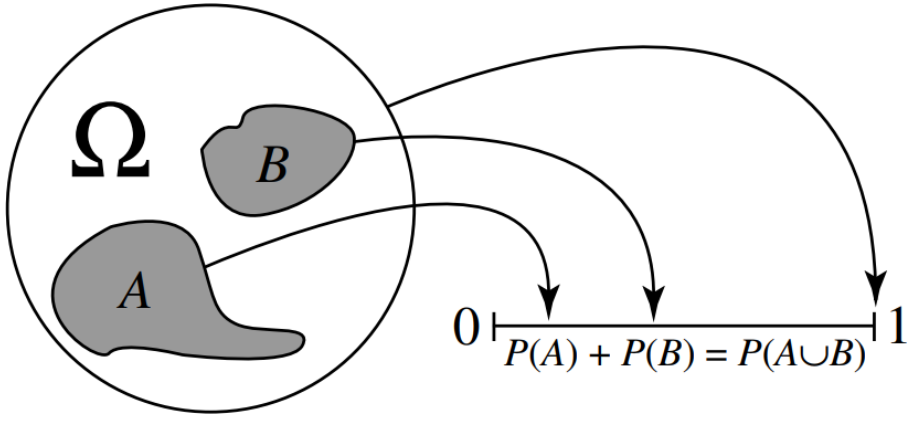


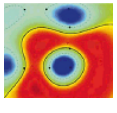


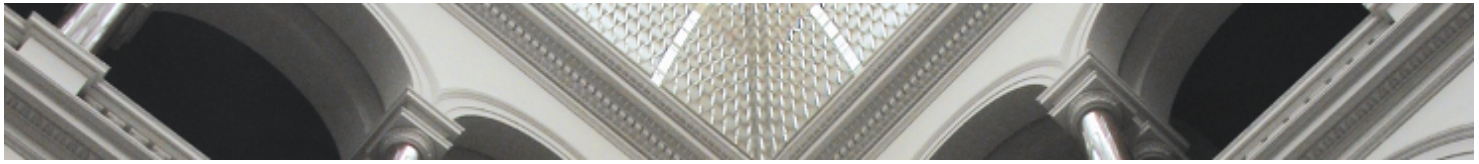
Technische Universität Berlin

The three axioms of Probability Theory

1. $P(A) \geq 0$ for all events A
2. $P(\Omega) = 1$
3. $P(A \cup B) = P(A) + P(B)$ for disjoint events A and B







Technische Universität Berlin

Basic Formulas for Probabilities

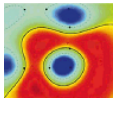
Product Rule: probability $p(A \cap B)$ of a conjunction of two events A and B:

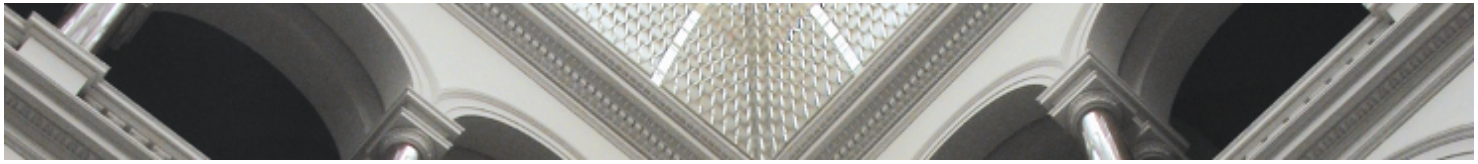
$$p(A \cap B) = p(A|B)p(B) = p(B|A)p(A)$$

The probability that A and B both happen is the probability that A happens times the probability that B happens, given A has occurred.

Sum Rule: probability $p(A \cup B)$ of a disjunction of two events A and B:

$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$





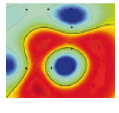
Technische Universität Berlin

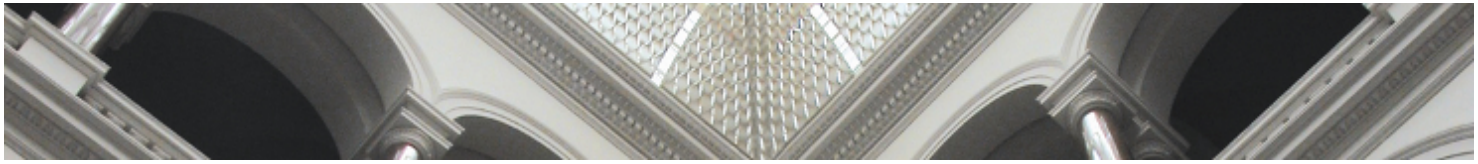
Bayes' Theorem

– **Bayes' Theorem** is a simple mathematical formula that has revolutionized how we understand and deal with uncertainty.

– It gives an answer to the question:

When we encounter new evidence, how much should it change our confidence in a belief?



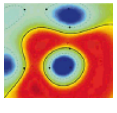


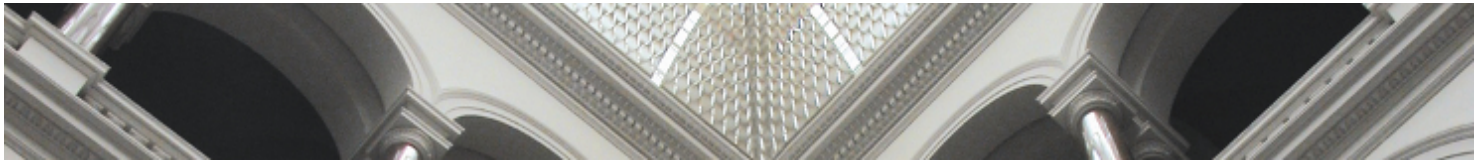
Technische Universität Berlin

Bayes' Theorem

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

- $p(A|B)$ is the probability of event A being true given that event B is true.
- $p(B|A)$ is the probability of event B being true given that event A is true.
- $p(A)$ and $p(B)$ are the probabilities of events A and B, where the events don't impact each other.



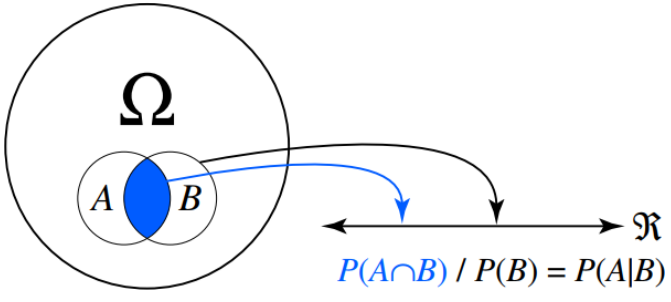


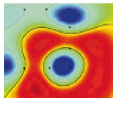
Technische Universität Berlin

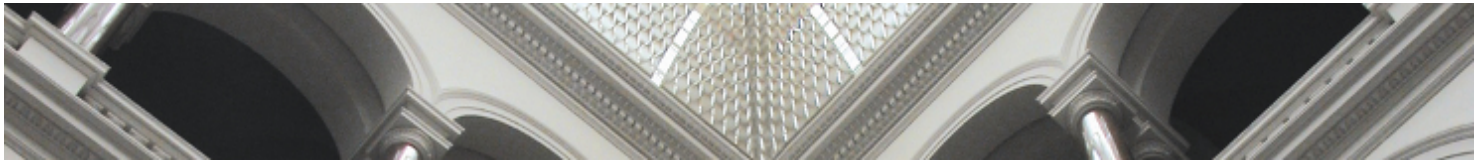
Bayes' Theorem

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} = \frac{p(B \cap A)}{p(B)}$$

- $P(A)$ is called the a *priori* (or *prior*) probability of A and $P(A|B)$ is called the a *posteriori* probability of A given B.
- $p(B|A)$ is the **conditional** probability or **likelihood**, is the degree of belief in B, given that the proposition A is true.







Technische Universität Berlin

Bayes' Theorem

In plain English, using Bayesian probability terminology, the Bayes rule can be written as

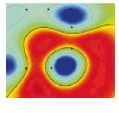
$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

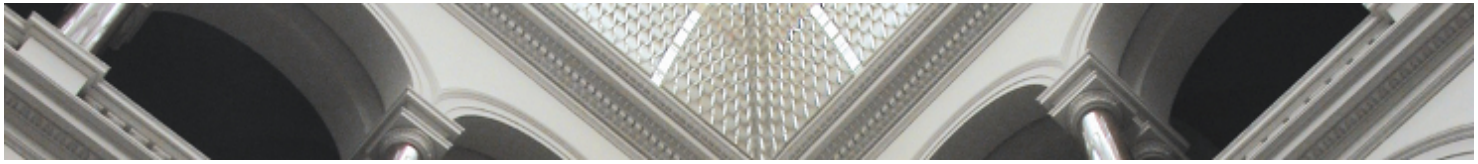
in other words

$$\textbf{posterior} = \frac{\textbf{likelihood} \times \textbf{prior}}{\textbf{evidence}}$$

or

$$\textbf{posterior} \propto \textbf{likelihood} \times \textbf{prior}$$





Technische Universität Berlin

Theorem of Total Probability

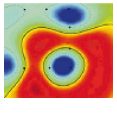
If events $B_1, \ldots, B_n$ are mutually exclusive with $\sum_{i=1}^{n} p(B_i) = 1$ (discrete case)

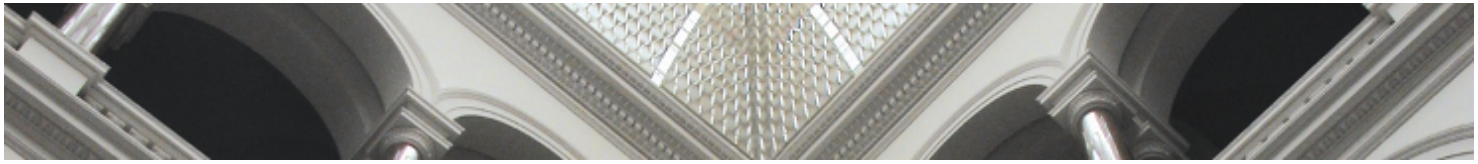
$$p(A) = \sum_{i=1}^{n} p(A|B_i)p(B_i)$$

and $\int_{\Omega} p(B)dB = 1$ (continuous case)

$$p(A) = \int_{\Omega} p(A|B)p(B)dB$$

also known as **marginalization**.





Technische Universität Berlin

Theorem of Total Probability

Suppose that two factories supply light bulbs to the market. Factory X's bulbs work for over 5000 hours in 99% of cases, whereas factory Y's bulbs work for over 5000 hours in 95% of cases. It is known that factory X supplies 60% of the total bulbs available and Y supplies 40% of the total bulbs available.

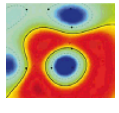
What is the chance that a purchased bulb will work for longer than 5000 hours?

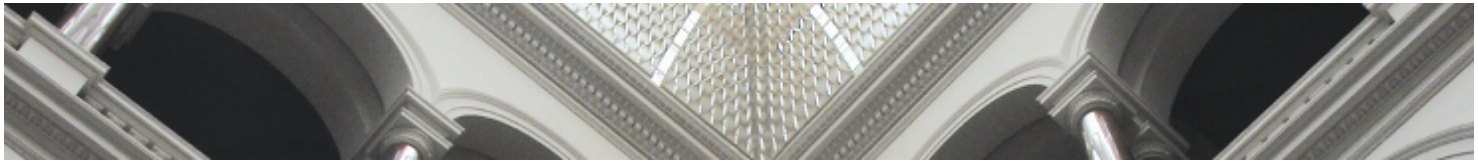
Applying the Theorem of Total Probability, we have:

$$P(A) = P(A|B_X)P(B_X) + P(A|B_Y)P(B_Y) = \tag{1}$$

$$\tag{2}$$

$$\frac{99}{100} \cdot \frac{6}{10} + \frac{95}{100} \cdot \frac{4}{10} = \frac{594 + 380}{1000} = \frac{974}{1000} \tag{3}$$



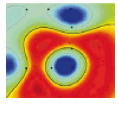


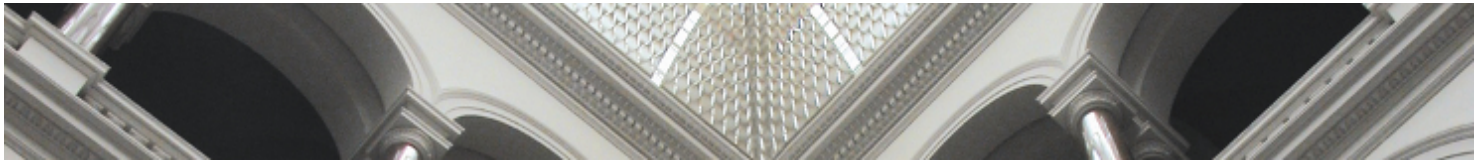
Technische Universität Berlin

Naïve Bayes Classifier

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

Thomas Bayes
1702 - 1761

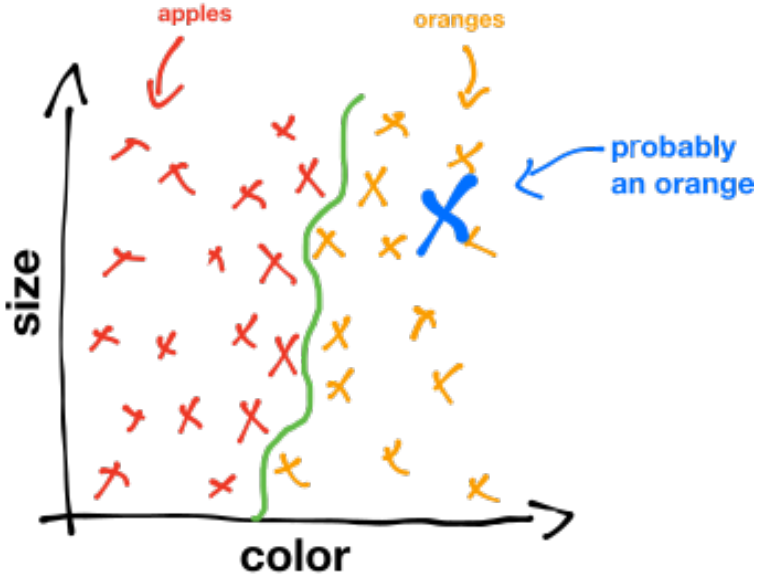


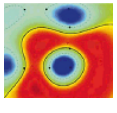


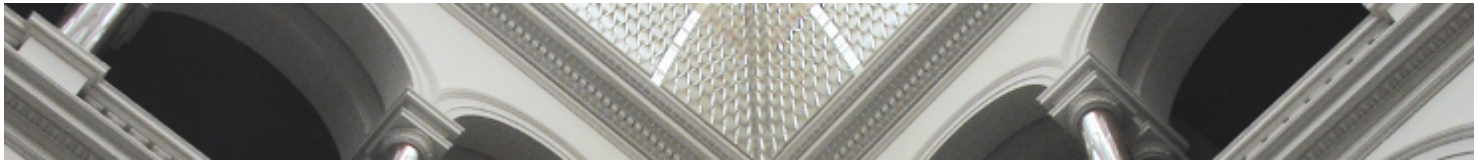
Technische Universität Berlin

Naive Bayes Classifier

Naive Bayes classifier is a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features.







Technische Universität Berlin

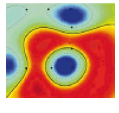
Probabilistic Model

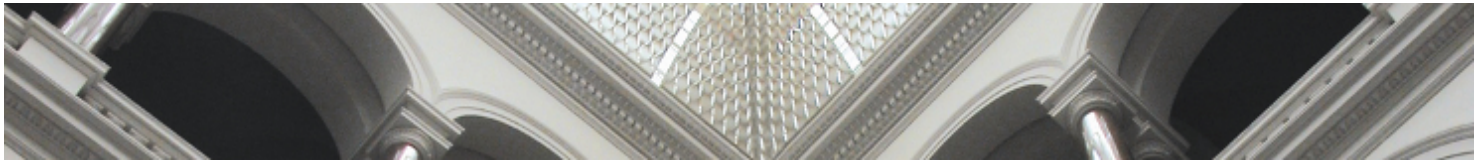
Naive Bayes can be seen a conditional probability model, where each instance $\boldsymbol{x} = (x_1, \ldots, x_n)$ with some n features needs to be classified.

The classifier assigns probabilities for each of K outcomes or **classes** C_k.

$$p(C_k|\boldsymbol{x}) = \frac{p(x_1, \cdots, x_n|C_k)p(C_k)}{p(\boldsymbol{x})} \propto p(x_1, \cdots, x_n, C_k)$$

The denominator does not depend on C and the values of the features x_i are given, so that the denominator is effectively constant.





Technische Universität Berlin

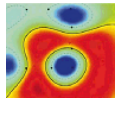
Probabilistic model

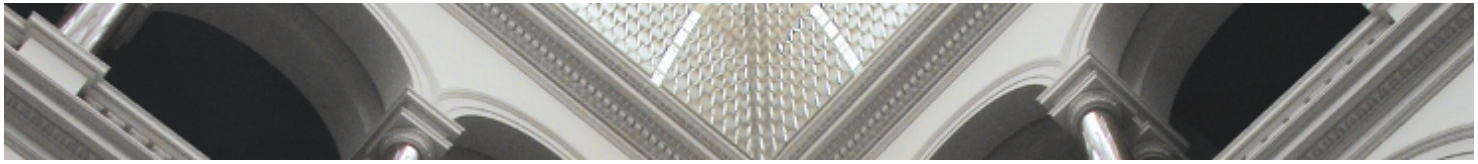
The **"naive"** assumption states that each feature x_i is conditionally independent of every other feature x_j for $i \neq j$, given the category C_k.

The joint model can be expressed as

$$p(C_k|\boldsymbol{x}) \propto p(x_1, \cdots, x_n, C_k) = p(C_k)p(x_1|C_k)p(x_2|C_k)\cdots p(x_n|C_k) =$$

$$p(C_k)\prod_{i=1}^{n} p(x_i|C_k)$$





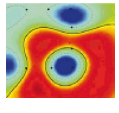
Technische Universität Berlin

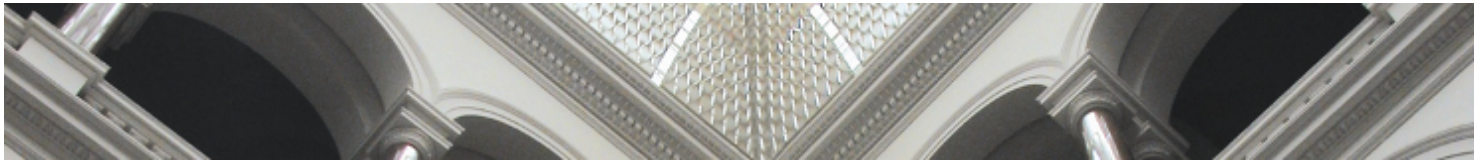
Naive Bayes Classifier

The naive Bayes classifier is the function that assigns a class label $\hat{y} = C_k$ for some k which is most probable.

The rule is also known as the **maximum a posteriori** (MAP).

$$\hat{y} = \underset{k \in \{1,\dots,K\}}{\operatorname{argmax}} \; p(C_k) \prod_{i=1}^{n} p(x_i | C_k)$$

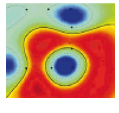


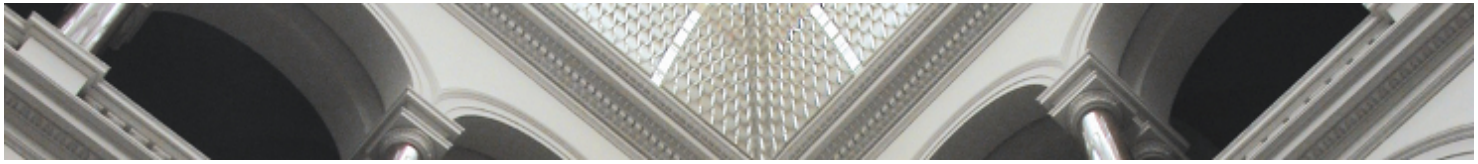


Technische Universität Berlin

Naive Bayes Classifier

Let's define a spam classifier using Bayes Theorem which can tell whether a given message is spam or not.





Technische Universität Berlin

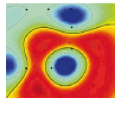
Naive Bayes Classifier

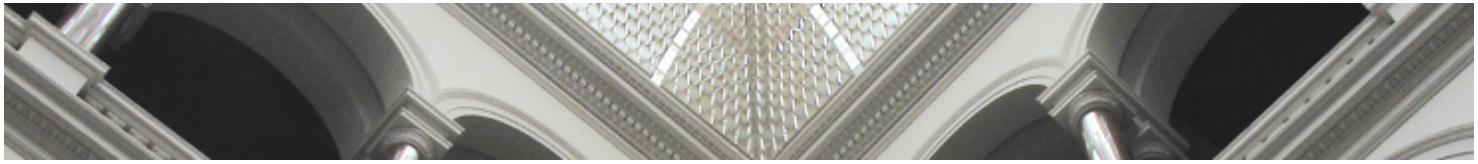
We have two classes $C_1 = \text{spam}$ and $C_2 = \text{ham}$. And each instance $\boldsymbol{x}$ is a message with n words as features.

$$\boldsymbol{x} = (w_1, w_2, ..., w_n)$$

In order to classify a message we need to determine which is greater

$$p(C = spam|w_1, w_2, ..., w_n) \quad \text{or} \quad p(C = ham|w_1, w_2, ..., w_n)$$





Technische Universität Berlin

Naive Bayes Classifier

From the Bayes rule we know that the posterior can be expressed as

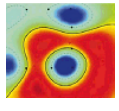
$$p(C = spam|w_1, w_2, ..., w_n) \propto p(C = spam) \prod_{i=1}^{n} p(w_i|C = spam)$$

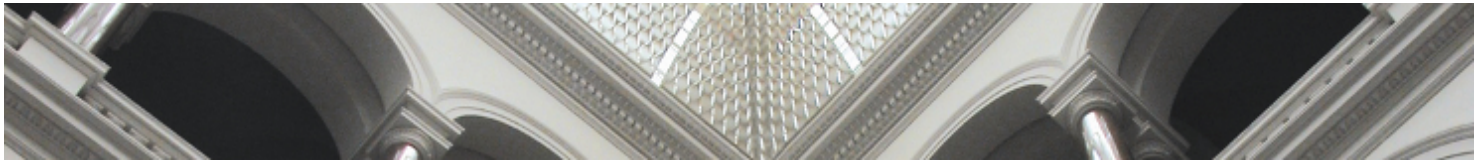
Where

$$p(w_i|C = \text{spam}) = \frac{\text{Total number of occurernces of } w_i \text{ in spam messages}}{\text{Total number of words in spam messages}}$$

and

$$p(C = \text{spam}) = \frac{\text{Total number of spam messages}}{\text{Total number of messages}}.$$





Technische Universität Berlin

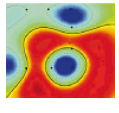
Naive Bayes Classifier

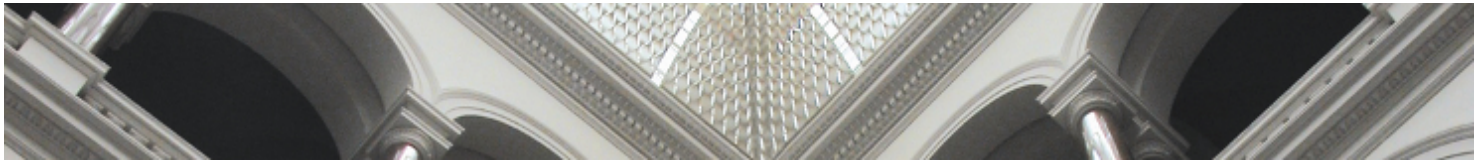
How to deal with rare or non seen words ?

The 'trick' to avoid zero counts is so called *Laplace* smoothing.

Let V be the set of words in the training set.

$$p(w_i|C = \text{spam}) = \frac{\text{Total number of occurernces of } w_i \text{ in spam messages} + 1}{\text{Total number of words in spam messages} + |V| + 1} \quad (4)$$

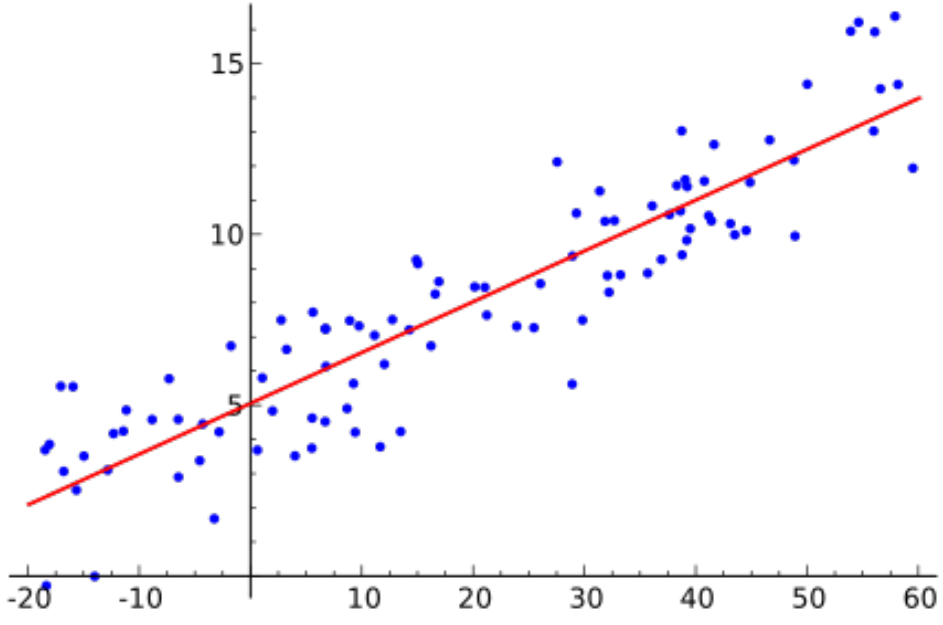


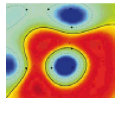


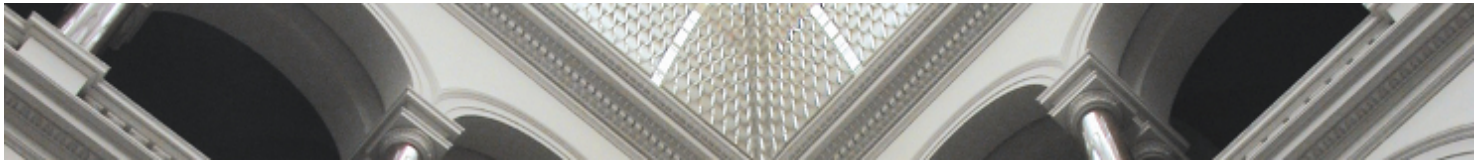
Technische Universität Berlin

Bayesian Learning

Example of the Linear Regression model

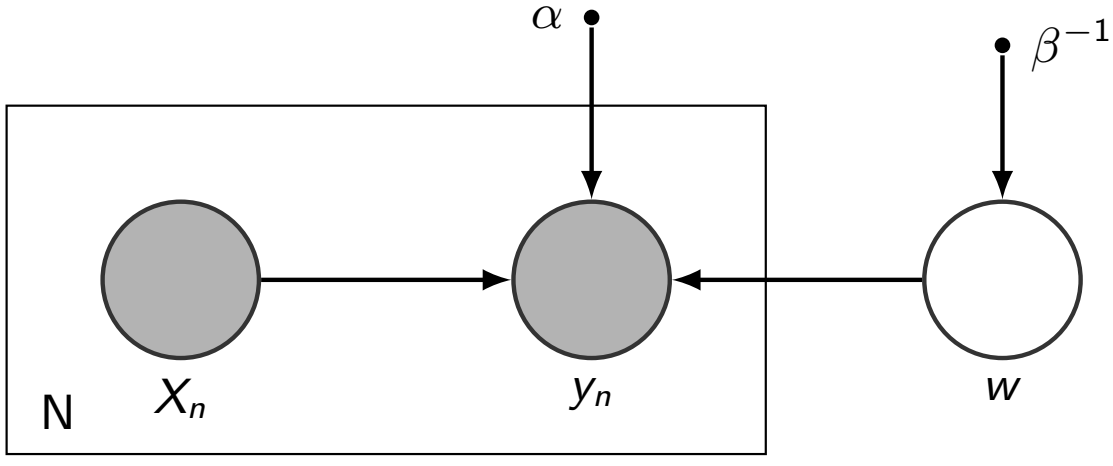






Technische Universität Berlin

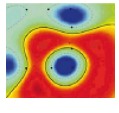
Standard Linear Regression model (plate notation)

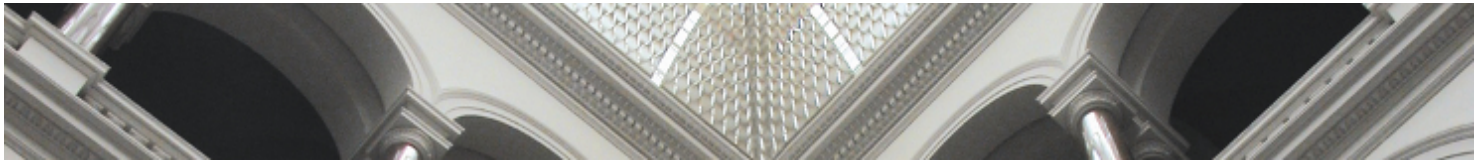


For each data point $x_i \in \mathbb{R}^{(1,k)}$, $\quad i = \overline{1..N}$

$$y_i = x_i w + \epsilon_i$$

where $w \in \mathbb{R}^{(k,1)}$ is a regression vector and $\epsilon_i = \mathcal{N}(\epsilon_i|0, \alpha)$ is an independent zero-mean Gaussian noise with variance $\alpha \in \mathbb{R}^+$.





Technische Universität Berlin

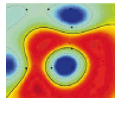
Prior of the regression vector

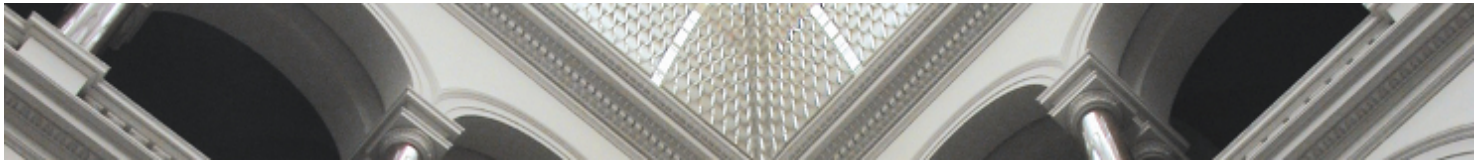
The prior over the regression vector w can be expressed with the precision parameter β as Gaussian with spherical covariance matrix:

$$p(w|\beta) = \mathcal{N}(w|\mu_0 = 0, \Sigma_0 = \beta^{-1}\mathbb{I}_k)$$

$k-$dimensional identity matrix:

$$\mathbb{I}_k = \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_k \end{bmatrix}, \quad \text{where} \quad d_i = 1, \quad i = \overline{1..k}$$





Technische Universität Berlin

Multivariate Gaussian distribution

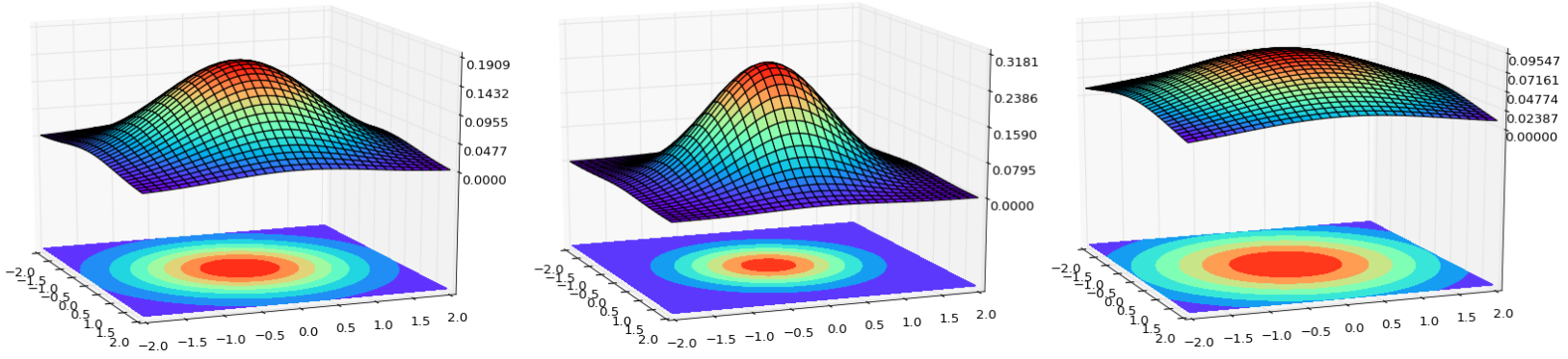
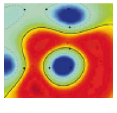


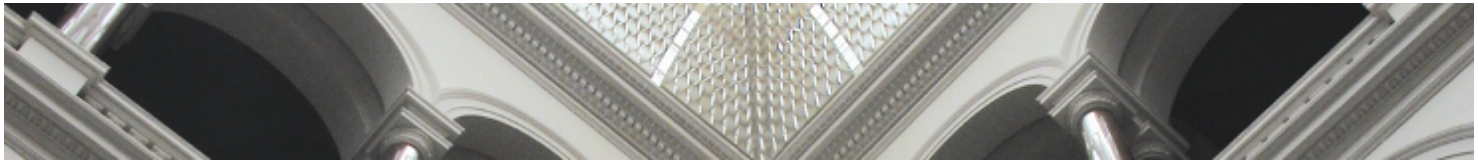
Figure: (left) $\beta = 1$, (middle) $\beta = 1.6$, (right) $\beta = 0.5$

PDF:

$$\mathcal{N}(x|\mu, \Sigma) = |2\pi\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

$$\ln \mathcal{N}(x|\mu, \Sigma) = -\frac{1}{2}\left(\ln 2\pi + \ln|\Sigma| + (x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

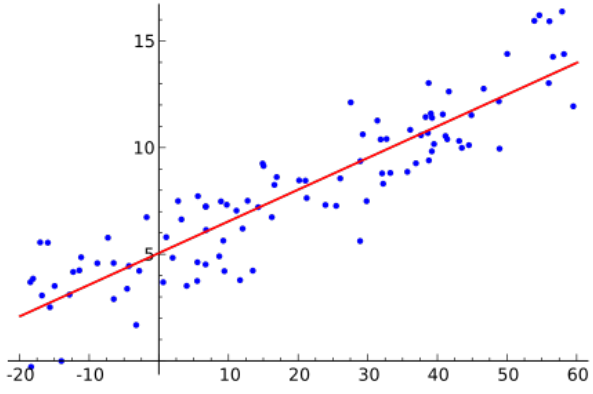




Technische Universität Berlin

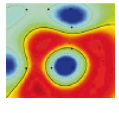
Linear Regression generative model

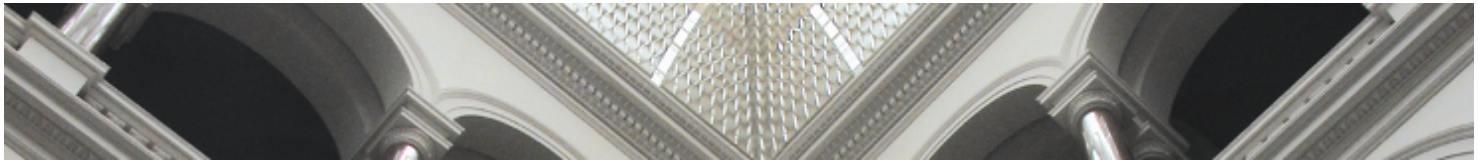
Given $\boldsymbol{X} \in \mathbb{R}^{(N,k)}$ data points with $\mathcal{E} \sim \mathcal{N}(\mathcal{E}|0, \sigma^2 \mathbb{I}_N) \in \mathbb{R}^N$ noise.



Generative model for w as a given vector:

$$y = \boldsymbol{X}w + \mathcal{E} \quad \text{or} \quad y \sim \mathcal{N}_N(y|\boldsymbol{X}w, \sigma^2 \mathbb{I}_N) \in \mathbb{R}^N$$





Maximum a Posteriori

Task: to find a most probable value of w based on the observed data y and $\boldsymbol{X}$.

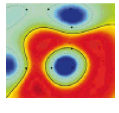
From Bayes rule we know that the posterior is proportional to likelihood times prior:

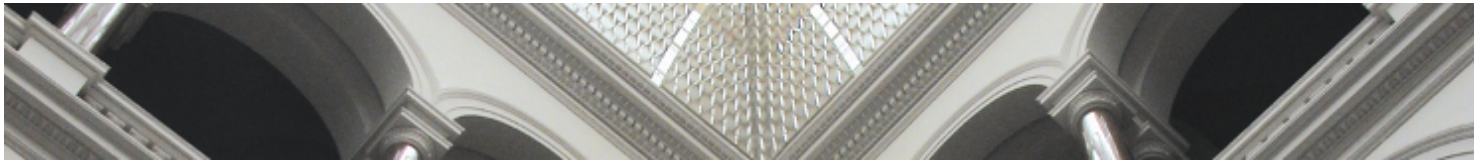
$$p(w|y, \boldsymbol{X}, \alpha, \beta) \propto p(y|\boldsymbol{X}, w, \alpha)p(w|\beta)$$

The log of the posterior can be written as following:

$$\mathcal{L}(w, \alpha, \beta) = \ln p(y|\boldsymbol{X}, w, \alpha) + \ln p(w|\beta) + const$$

$$\mathcal{L}(w, \alpha, \beta) = \ln \mathcal{N}(y|\boldsymbol{X}w, \alpha) + \ln \mathcal{N}(w|0, \beta^{-1}\mathbb{I}_k) + const$$





Technische Universität Berlin

Apply log of the Gaussian distribution expression:

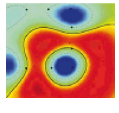
$$\mathcal{L}(w, \alpha, \beta) = -\frac{1}{2}\left(N \ln \alpha + \alpha^{-1}(y - \boldsymbol{X}w)^\top (y - \boldsymbol{X}w) - k \ln \beta + \beta w^\top w\right) + const$$

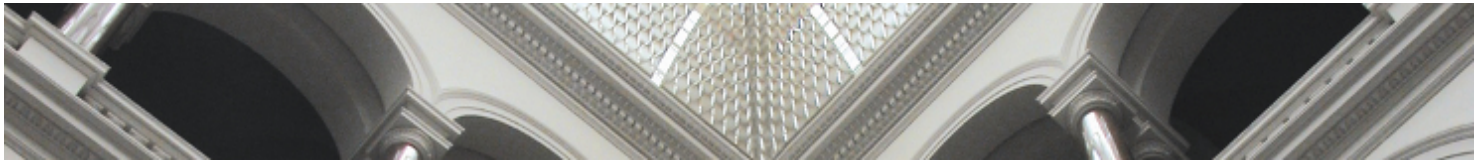
Maximum of the function for w can be found as following:

$$\frac{\partial \mathcal{L}(w, \alpha, \beta)}{\partial w} = \left(\alpha^{-1}\boldsymbol{X}^\top (y - \boldsymbol{X}w) - \beta w\right) = 0$$

$$\alpha^{-1}\boldsymbol{X}^\top y = (\alpha^{-1}\boldsymbol{X}^\top \boldsymbol{X} + \beta \mathbb{I}_k)w$$

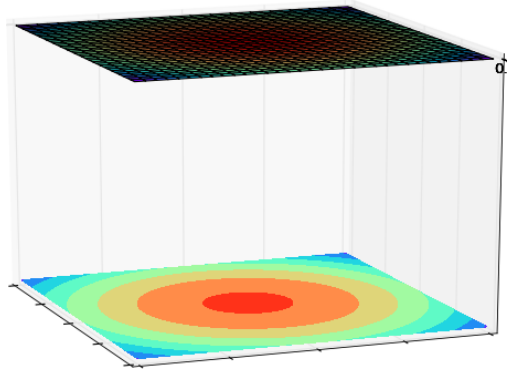
$$\boxed{w^* = \left(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \mathbb{I}_k\right)^{-1} \boldsymbol{X}^\top y, \quad \lambda = \alpha\beta}$$





Technische Universität Berlin

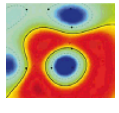
Broad prior

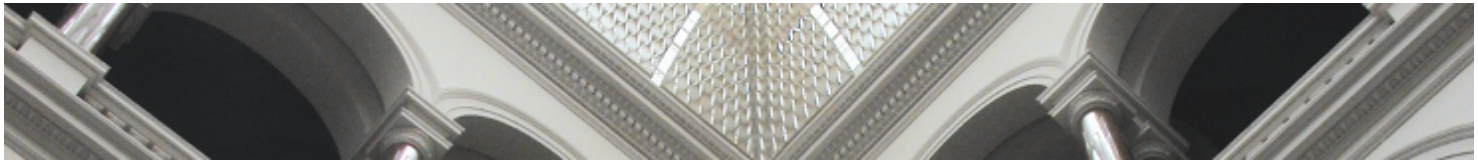


In case we put a broad prior ($\beta \to 0$) for the regression coefficients the *MAP* solution is

$$\boxed{w^* = \boldsymbol{X}^\dagger y, \quad \text{where}}$$

the quantity $\boldsymbol{X}^\dagger \equiv \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top$ is known as **Moore-Penrose pseudo-inverse** of matrix $\boldsymbol{X} \in \mathbb{R}^{(N,k)}$.





Technische Universität Berlin

Moore-Penrose pseudo-inverse

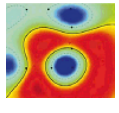
Properties:

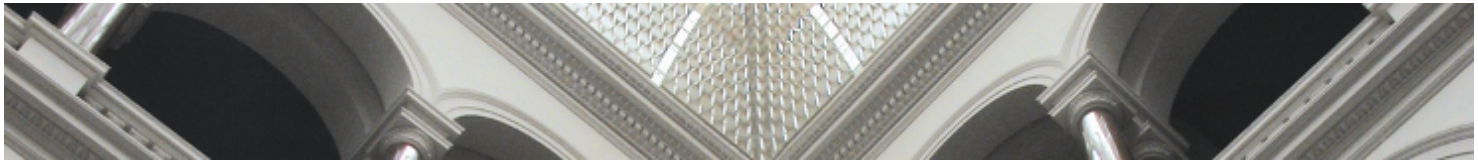
$$\boldsymbol{X}\boldsymbol{X}^{\dagger}\boldsymbol{X} = \boldsymbol{X}$$

if there is an inverse of $\boldsymbol{X}$, then

$$\boldsymbol{X}\boldsymbol{X}^{\dagger}\boldsymbol{X}\boldsymbol{X}^{-1} = \boldsymbol{X}\boldsymbol{X}^{-1} \quad \Rightarrow \boldsymbol{X}\boldsymbol{X}^{\dagger} = \mathbb{I}$$

$\boldsymbol{X}^{\dagger}$ is a *TRUE* inverse.



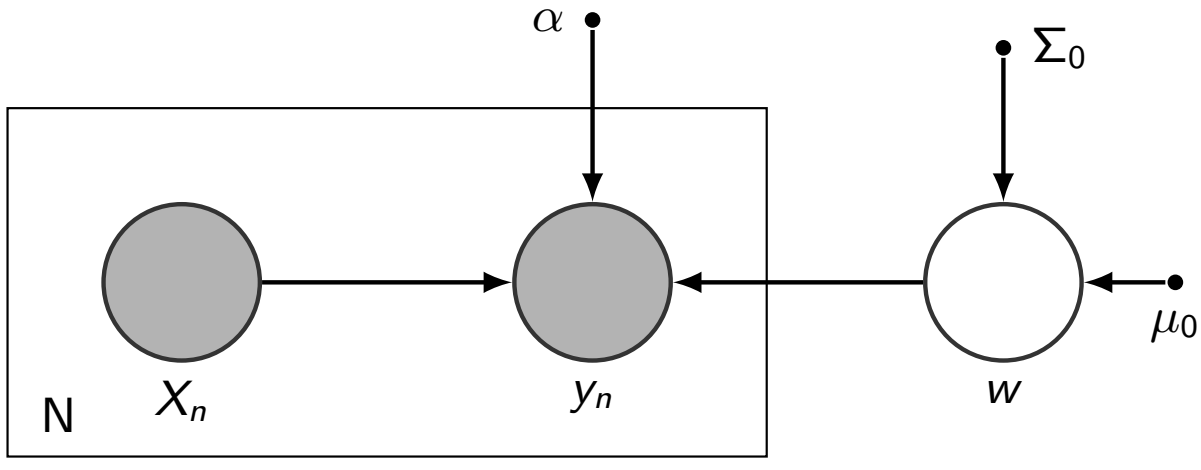


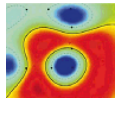
Technische Universität Berlin

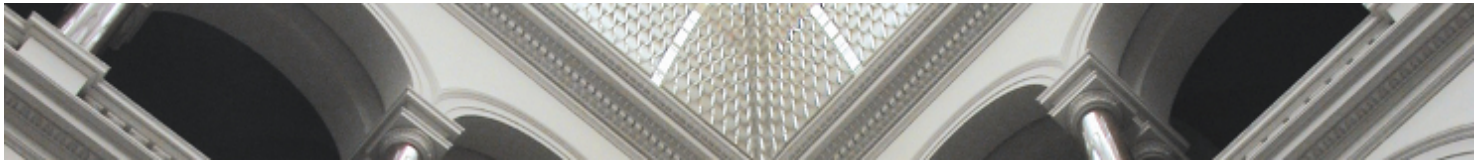
Linear Regression model with mean and covariance prior

Assume multivariate Gaussian prior for w with mean $\mu_0 \in \mathbb{R}^k$ and covariance matrix $\Sigma_0 \in \mathbb{R}^{(k,k)}$.

$$p(w|\mu_0, \Sigma_0) = \mathcal{N}(w|\mu_0, \Sigma_0)$$







Technische Universität Berlin

Posterior distribution with mean and covariance

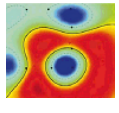
Task: to find a form of the posterior.

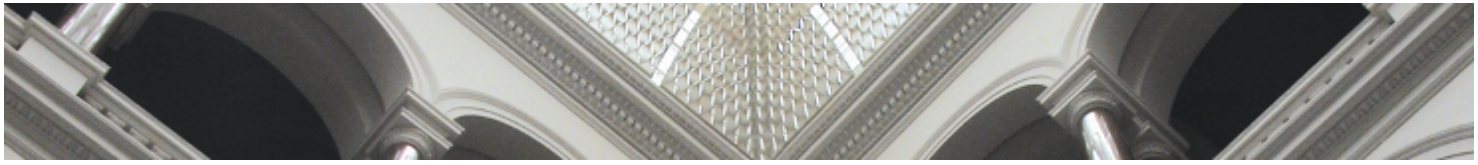
In this case the log posterior has the following form:

$$\mathcal{L}(w, \alpha, \mu_0, \Sigma_0) = \ln \mathcal{N}(y|\boldsymbol{X}w, \alpha) + \ln \mathcal{N}(w|\mu_0, \Sigma_0) + const$$

Apply log of the Gaussian distribution expression:

$$\mathcal{L}(w, \alpha, \mu_0, \Sigma_0) = -\frac{1}{2}\Big(N \ln \alpha + \alpha^{-1}(y - \boldsymbol{X}w)^\top (y - \boldsymbol{X}w) + \ln |\Sigma_0| + (w - \mu_0)^\top \Sigma_0^{-1}(w - \mu_0)\Big) + const$$





Technische Universität Berlin

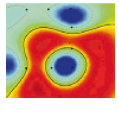
Open brackets and consider the terms depending only on w.

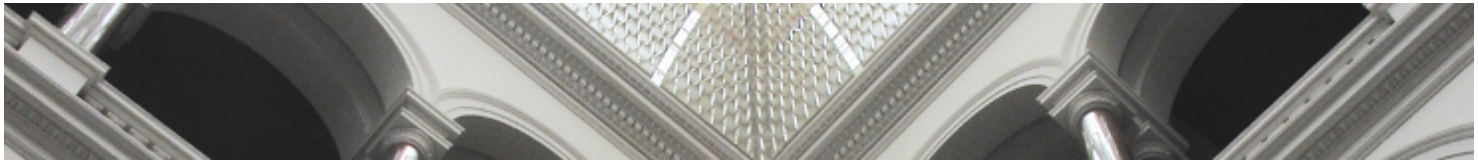
$$\mathcal{L}(w) = -\frac{1}{2}\Big(-2\alpha^{-1} w^\top \boldsymbol{X}^\top y + \alpha^{-1} w^\top \boldsymbol{X}^\top \boldsymbol{X} w - 2 w^\top \mu_0 \Sigma_0^{-1} + w^\top \Sigma_0^{-1} w \Big) + const$$

f(w) = wˆTAw - 2wˆTb = (w-Aˆ-1b)ˆTA(w-Aˆ-1b)+const

Combine terms within $w^\top ... w$ and $w^\top$ = mean = sumˆ-1

$$\mathcal{L}(w) = -\frac{1}{2}\Big(w^\top (\overbrace{\alpha^{-1}\boldsymbol{X}^\top \boldsymbol{X} + \Sigma_0^{-1}}^{A}) w - 2 w^\top (\overbrace{\alpha^{-1}\boldsymbol{X}^\top y + \Sigma_0^{-1}\mu_0}^{b}) \Big) + const$$





Technische Universität Berlin

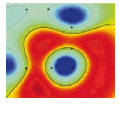
Completing square

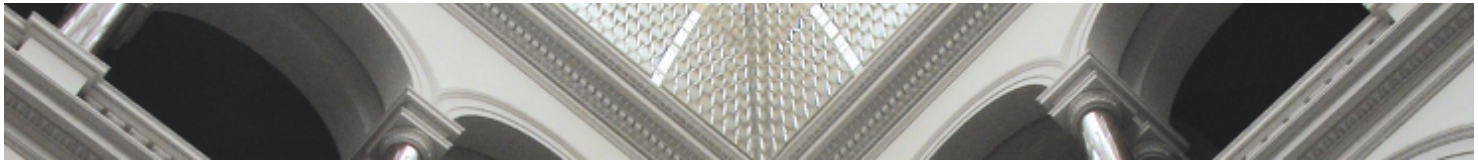
Mathematical trick useful for building quadric forms.

$$w^\top A w - 2 w^\top b = (w - A^{-1} b)^\top A (w - A^{-1} b) - b^\top A^{-1} b$$

Proof.

$$\begin{aligned}(w - A^{-1} b)^\top A (w - A^{-1} b) =& w^\top A w - 2 w^\top A^{-1} A b + b^\top A^{-1} A A^{-1} b = \\ & w^\top A w - 2 w^\top b + b^\top A^{-1} b\end{aligned}$$





Technische Universität Berlin

Gaussian posterior distribution

After completing the square the posteriors has a form of Gaussian multivariate distribution with mean μ_N and Σ_N.

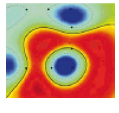
w ~ q(mean,sum) = N (wlmean=A^-1b,sum=A^-1)

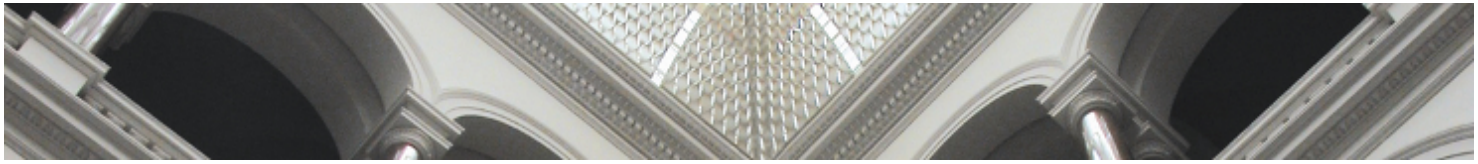
$$p(w|y, \boldsymbol{X}, \alpha, \mu_0, \Sigma_0) = \mathcal{N}(w|\mu_N, \Sigma_N)$$

$$\boxed{\Sigma_N = \left(\alpha^{-1}\boldsymbol{X}^\top\boldsymbol{X} + \Sigma_0^{-1}\right)^{-1}} \qquad \boxed{\mu_N = \Sigma_N(\alpha^{-1}\boldsymbol{X}^\top y + \Sigma_0^{-1}\mu_0)}$$

or

$$\boxed{\Sigma_N = \left(\boldsymbol{X}^\top\boldsymbol{X} + \alpha\Sigma_0^{-1}\right)^{-1}} \qquad \boxed{\mu_N = \Sigma_N(\boldsymbol{X}^\top y + \alpha\Sigma_0^{-1}\mu_0)}$$





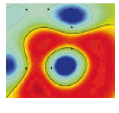
Technische Universität Berlin

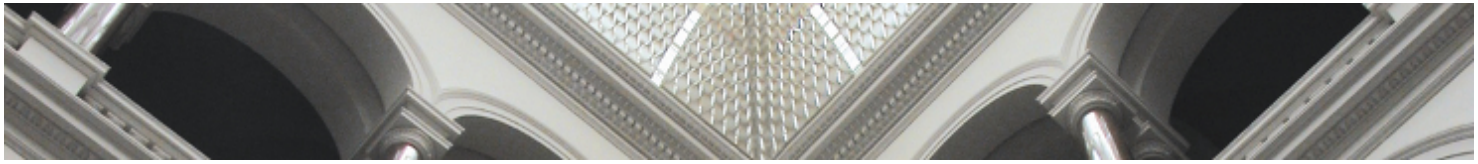
Properties of the Gaussian Posterior

- Because the posterior is Gaussian its mode coincides with its mean. Thus the maximum posterior weight is given by $w_{MAP} = \mu_N$
- If we consider infinitely broad prior $\Sigma_0 = \beta^{-1}\mathbb{I}, \quad \beta \to 0$, the mean of the posterior reduces to the MAP solution without regularization term.

$$\mu_N = \boldsymbol{X}^{\dagger} y$$

- If $N = 0$, posterior reverts to the prior.
- If data points arrive sequentially, then posterior to any stage acts as prior distribution for subsequent data points.





Technische Universität Berlin

Sequential Bayesian Learning example

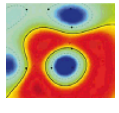
We can illustrate the sequential update of a posterior in a simple example.

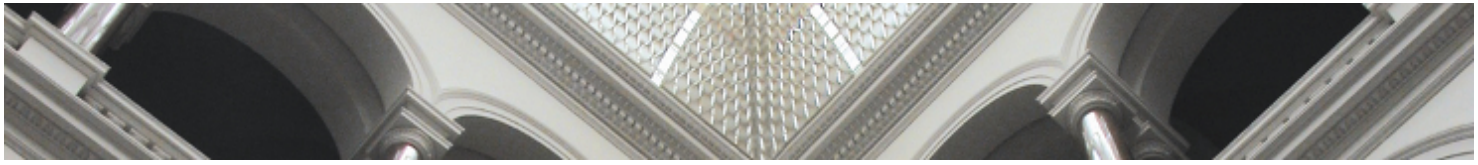
- Consider a linear model of the form $y(x, a) = a_0 + a_1 x$ where x is a single input variable and $a_0 = -0.3 \quad \text{and} \quad a_1 = 0.5$

- We can generate the synthetic data by choosing the values of x from univariate distribution $x_n \sim \mathcal{U}(x| -1, 1)$ then evaluate $y(x_n, w)$

- Finally add some Gaussian noise with standard deviation $\alpha = 0.2$ to obtain the target value y_n

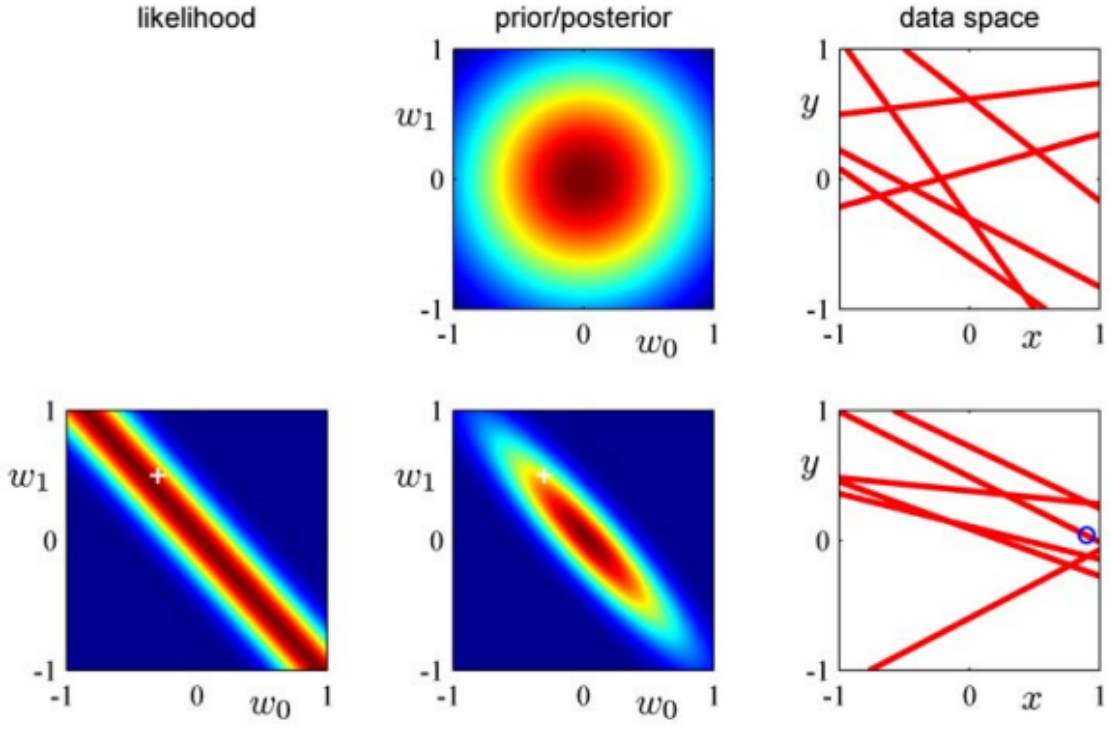
Example from: Christopher M. Bishop "*Pattern Recognition and Machine Learning*" P. 155

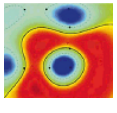


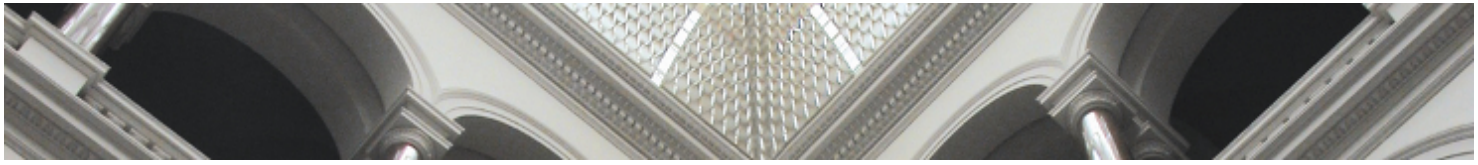


Technische Universität Berlin

Illustration of sequential Bayesian I

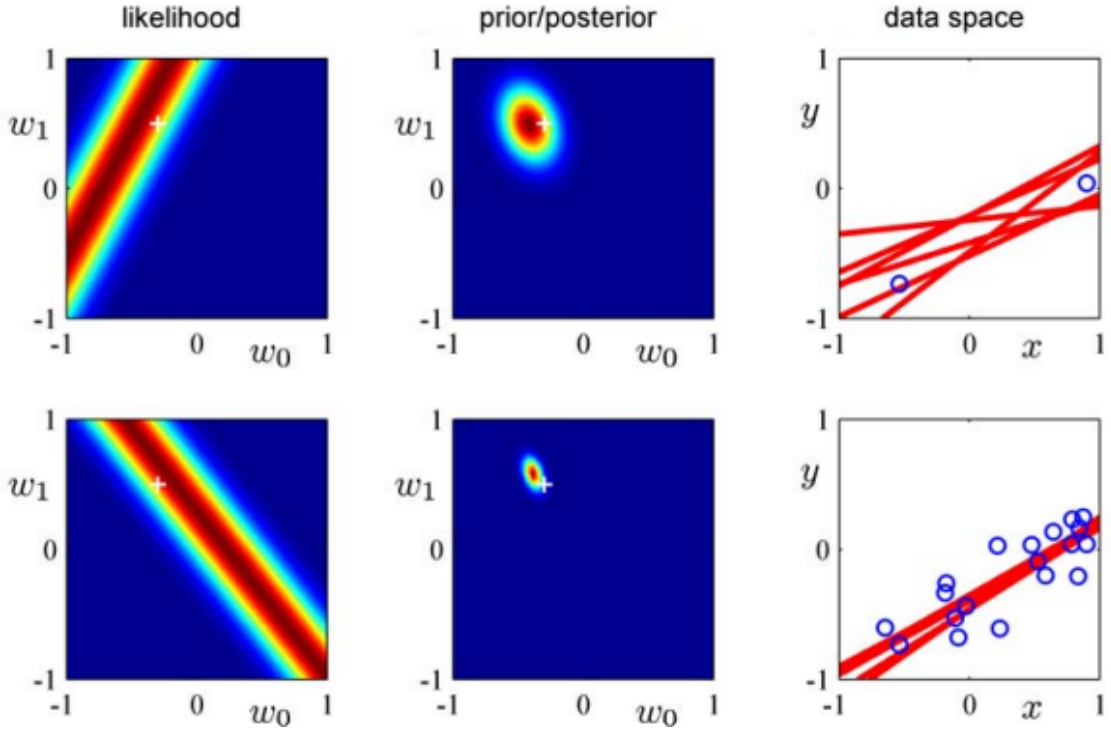


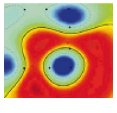


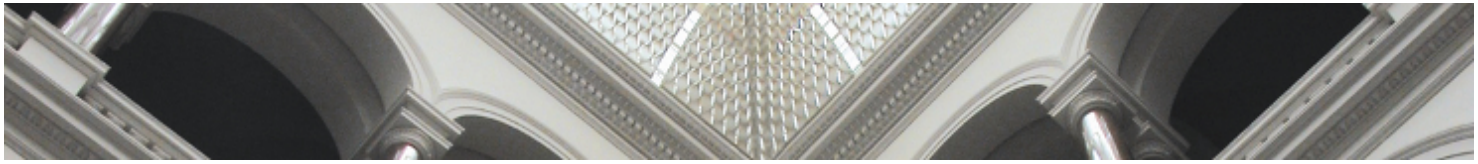


Technische Universität Berlin

Illustration of sequential Bayesian II

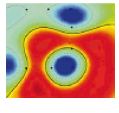


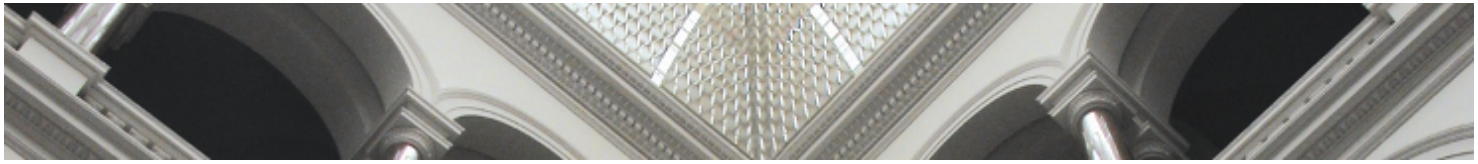




Technische Universität Berlin

End of the first part.





Technische Universität Berlin

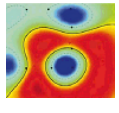
Expectation Maximization motivation

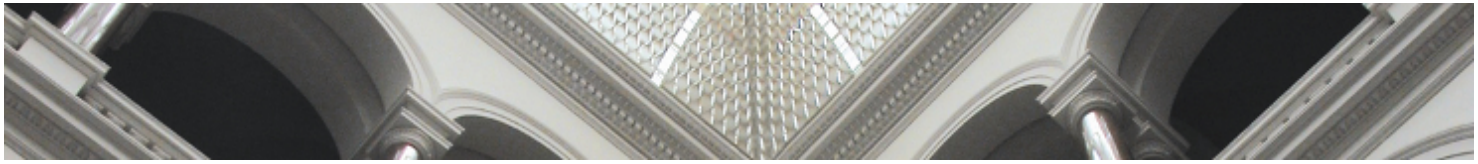
Question: How to optimize the model parameters ?

Find the maximum of the log-likelihood function w.r.t the model parameters and the expected value of w.

Update the model variable and repeat the procedure till the convergence (*Expectation maximization* algorithm)

$$\mathbb{E}[\mathcal{L}(w, \alpha, \mu_0, \Sigma_0)] = -\frac{1}{2}\Big(N \ln \alpha + \alpha^{-1}\mathbb{E}\Big[(y - \boldsymbol{X}w)^\top (y - \boldsymbol{X}w)\Big] + \ln |\Sigma_0| + \mathbb{E}\Big[(w - \mu_0)^\top \Sigma_0^{-1}(w - \mu_0)\Big]\Big) + const$$





Technische Universität Berlin

Some math...

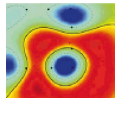
Consider terms depending on the noise factor α.

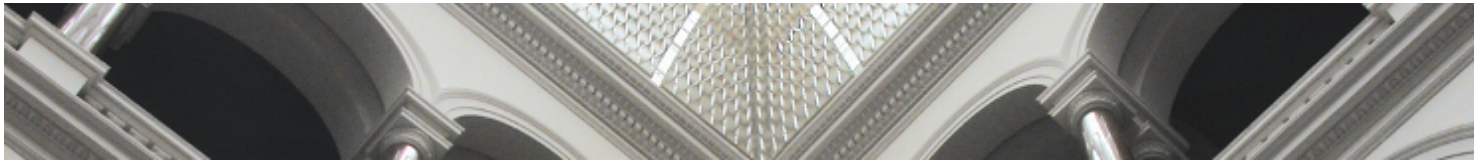
$$\mathcal{L}(\alpha) = -\frac{1}{2}\left(N \ln \alpha + \alpha^{-1}\mathbb{E}\left[(y - \boldsymbol{X}w)^\top (y - \boldsymbol{X}w)\right]\right) + const$$

$$\frac{\partial \mathcal{L}(\alpha)}{\partial \alpha} = -\frac{1}{2}\left(N\alpha^{-1} - \alpha^{-2}\mathbb{E}\left[(y - \boldsymbol{X}w)^\top (y - \boldsymbol{X}w)\right]\right) = 0$$

$$N\alpha^{-1} = \alpha^{-2}\mathbb{E}\left[(y - \boldsymbol{X}w)^\top (y - \boldsymbol{X}w)\right] \quad | \quad * \alpha^2$$

$$\hat{\alpha} = \frac{y^\top y - 2y^\top \boldsymbol{X}\mathbb{E}[w] + \mathrm{Tr}\left(\mathbb{E}[ww^\top]\boldsymbol{X}^\top \boldsymbol{X}\right)}{N}$$





Technische Universität Berlin

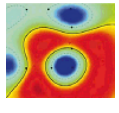
Noise variance optimization α

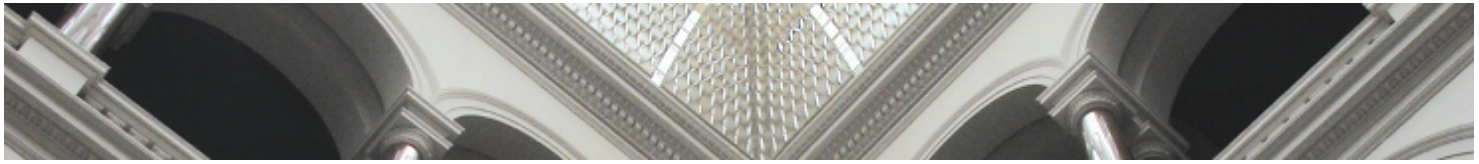
Because the posterior is Gaussian with mean μ_N and covariance Σ_N the first and the second moment are

$$\mathbb{E}[w] = \mu_N, \quad \mathbb{E}[ww^\top] = \mu_N \mu_N^\top + \Sigma_N$$

The updated noise variance value is then

$$\hat{\alpha} = \frac{N}{y^\top y - 2y^\top \boldsymbol{X}\mu_N + \mathrm{Tr}\left((\mu_N \mu_N^\top + \Sigma_N)\boldsymbol{X}^\top \boldsymbol{X}\right)}$$





Technische Universität Berlin

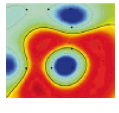
Mean of the prior distribution optimization μ_0

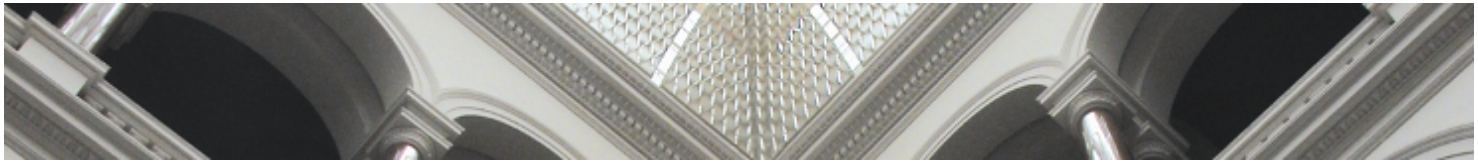
Consider terms depending on μ_0.

$$\mathbb{E}[\mathcal{L}(\mu_0, \Sigma_0)] = -\frac{1}{2}\left(\ln |\Sigma_0| + \mathbb{E}\left[(w - \mu_0)^\top \Sigma_0^{-1} (w - \mu_0) \right] \right) + const$$

$$\frac{\partial \mathbb{E}[\mathcal{L}(\mu_0)]}{\partial \mu_0} = -\mu_0 \Sigma_0^{-1} + \mathbb{E}[w]\Sigma_0^{-1} = 0$$

$$\boxed{\hat{\mu}_0 = \mathbb{E}[w] = \mu_N}$$





Technische Universität Berlin

Covariance of the prior distribution optimization Σ_0

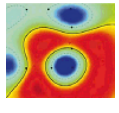
Consider terms depending on Σ_0.

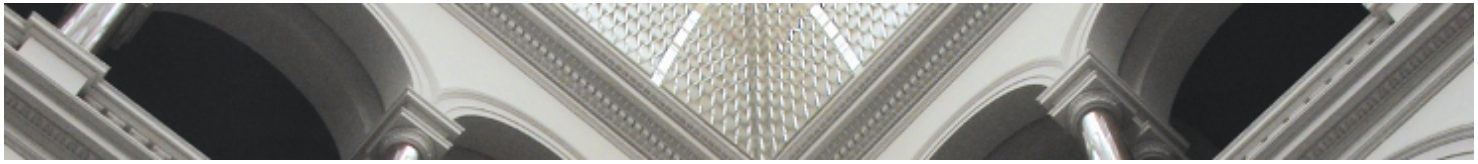
$$\frac{\partial \mathbb{E}[\mathcal{L}(\Sigma_0)]}{\partial \Sigma_0} = \Sigma_0^{-1} - \left(\mathbb{E}[ww^\top] - 2\mathbb{E}[w]\hat{\mu}_0 + \hat{\mu}_0\hat{\mu}_0^\top\right)\Sigma_0^{-2} = 0$$

$$\hat{\Sigma}_0 = \mathbb{E}[ww^\top] - 2\mathbb{E}[w]\hat{\mu}_0 + \hat{\mu}_0\hat{\mu}_0^\top$$

$$\hat{\Sigma}_0 = \mu_N\mu_N^\top + \Sigma_N - 2\mu_N\mu_N^\top + \mu_N\mu_N^\top = \Sigma_N$$

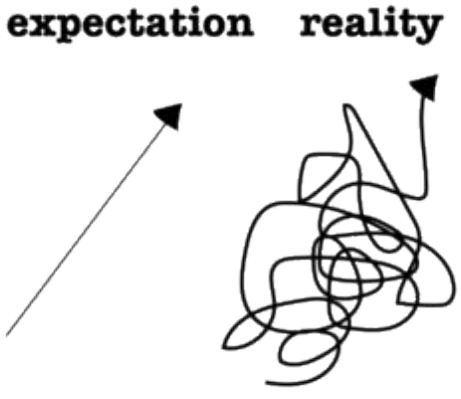
$$\boxed{\hat{\Sigma}_0 = \Sigma_N}$$

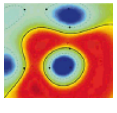


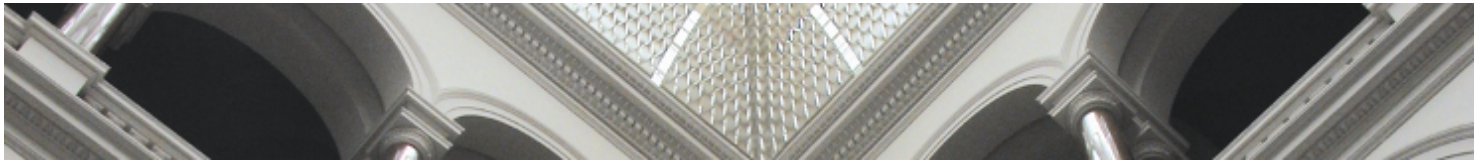


Technische Universität Berlin

The EM Algorithm in General







Technische Universität Berlin

Theory

The *expectation maximization* algorithm (EM algorithm) is a general technique for finding maximum likelihood solutions for probabilistic models having latent variables.

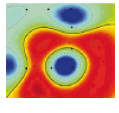
- Consider a probabilistic model with all observed variables $\boldsymbol{X}$ and the hidden variables $\boldsymbol{Z}$.
- The joint distribution $p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ is conditioned on the model parameters $\boldsymbol{\theta}$.

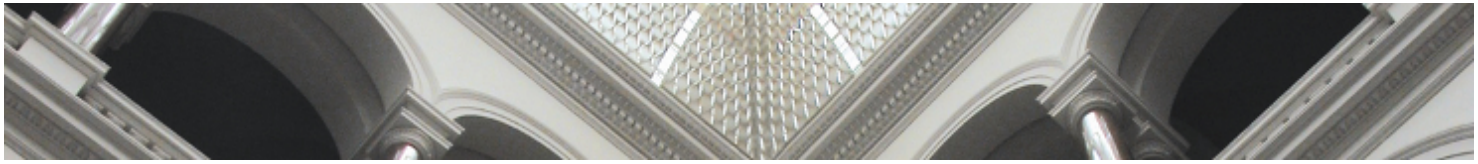
Goal: to maximize the log likelihood functions is given by:

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \ln \left\{ \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})} \right\}$$

because of

$$p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) = p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}) p(\boldsymbol{X}|\boldsymbol{\theta}) \quad \text{or} \quad p(\boldsymbol{X}|\boldsymbol{\theta}) = \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})}$$





Technische Universität Berlin

Theory

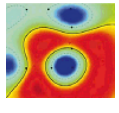
Multiply and divide by an arbitrary distribution over $\boldsymbol{Z}$ called $q(\boldsymbol{Z})$.

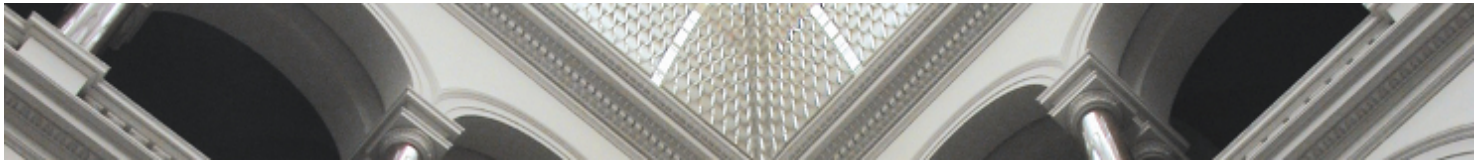
$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \ln \frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})q(\boldsymbol{Z})}{q(\boldsymbol{Z})p(\boldsymbol{Z},\boldsymbol{X}|\boldsymbol{\theta})} = \ln \frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} - \ln \frac{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}{q(\boldsymbol{Z})}$$

Integrate out an arbitrary distribution $q(\boldsymbol{Z})$

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \underbrace{\underbrace{\int_Z q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} d\boldsymbol{Z}}_{\mathcal{L}(q,\boldsymbol{\theta})}}_{\text{lower bound}} + \underbrace{\int_Z -q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}{q(\boldsymbol{Z})} d\boldsymbol{Z}}_{\mathbb{D}_{KL}(q||p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}$$

$$\boxed{\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \mathcal{L}(q,\boldsymbol{\theta}) + \mathbb{D}_{KL}(q||p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}$$

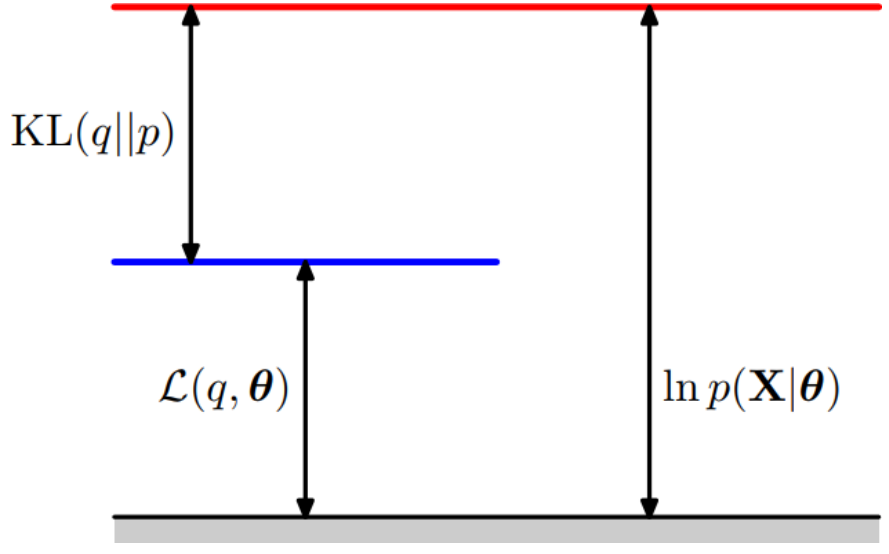




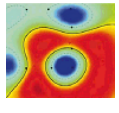
Technische Universität Berlin

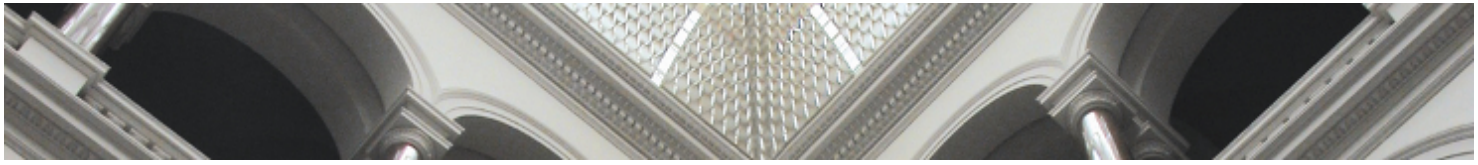
Lower bound on the log likelihood function $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathbb{D}_{KL}(q||p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$$



For any arbitrary distribution $q(\boldsymbol{Z})$ the KL divergence $\mathbb{D}_{KL}(q||p) \geq 0$.

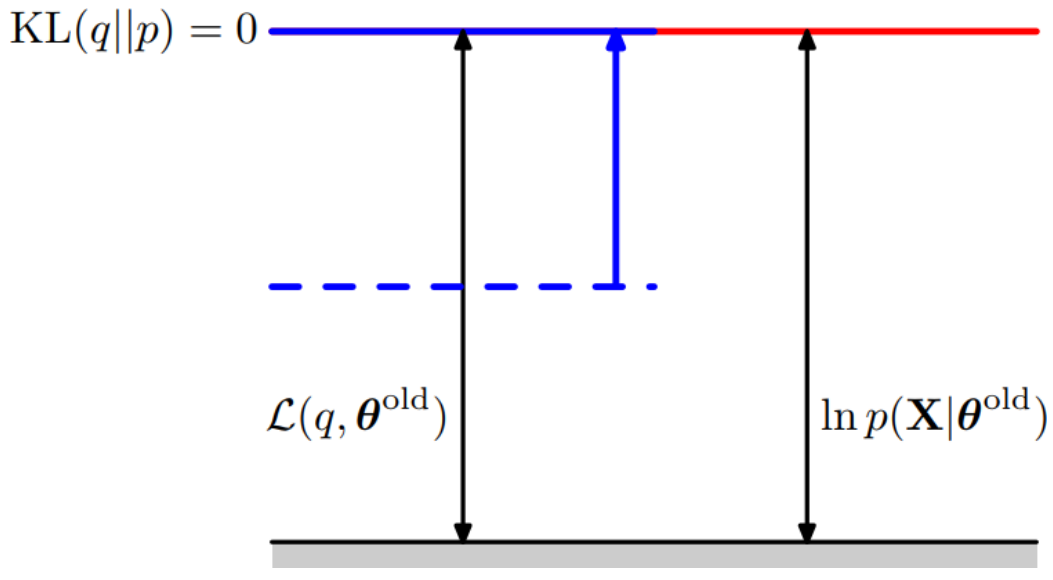




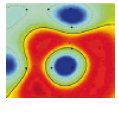
Technische Universität Berlin

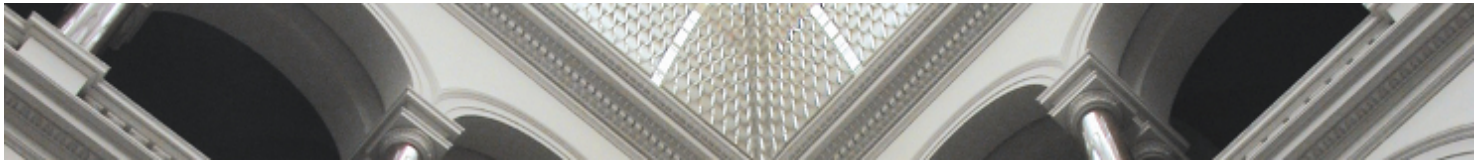
E-step

The EM algorithm is a two-stage iterative optimization technique for finding maximum likelihood solutions.



E-step: the lower bound $\mathcal{L}(q, \theta^{old})$ is maximized w.r.t. $q(\boldsymbol{Z})$.
It's reached when KL term vanishes, in other words when $q(\boldsymbol{Z})$ is equal the posterior distribution $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta^{old}})$ for the fixed model parameters θ^{old}.

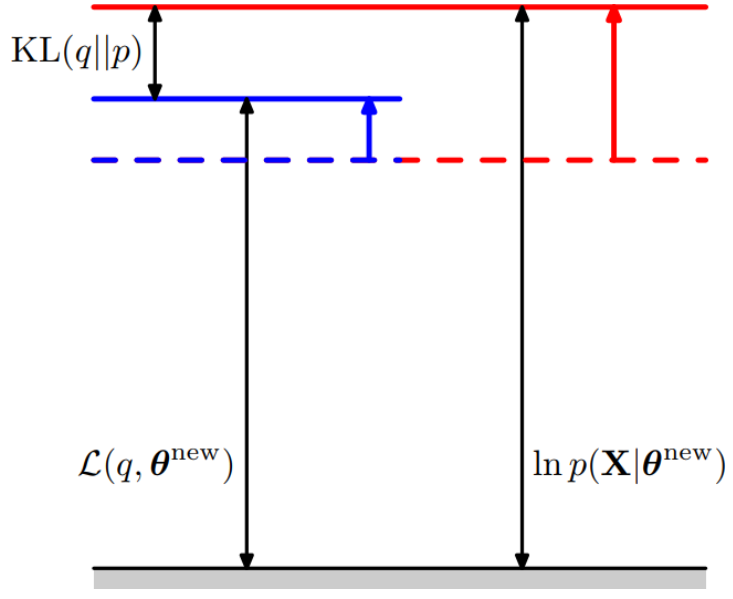




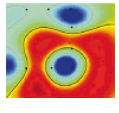
Technische Universität Berlin

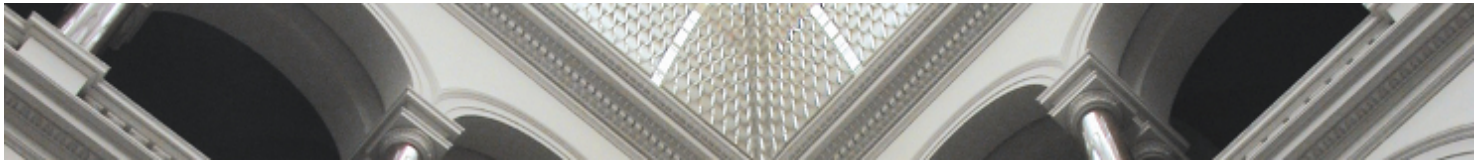
M-step

In the M-step the distribution $q(\boldsymbol{Z})$ is fixed and the lower bound $\mathcal{L}(\boldsymbol{Z}, \boldsymbol{\theta})$ is optimized w.r.t. $\boldsymbol{\theta}$ to give some new value $\boldsymbol{\theta}^{new}$.



Because the KL divergence is non negative, the log likelihood $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ increases at least as much as the lower bound does.

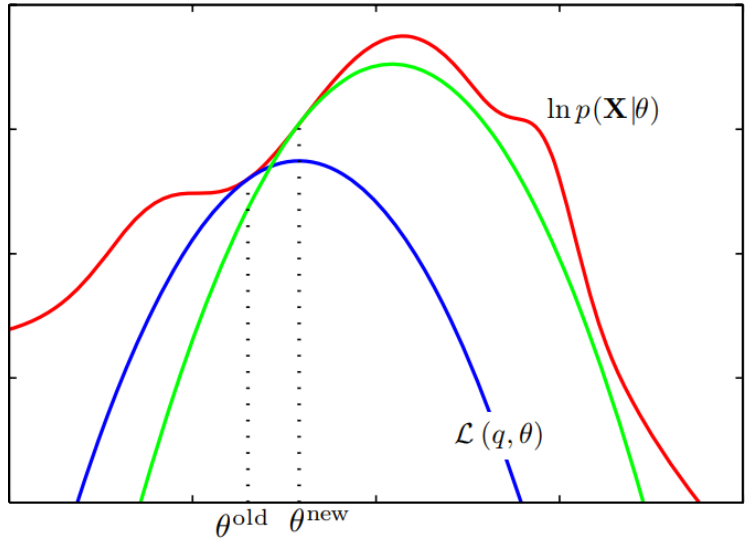




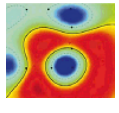
Technische Universität Berlin

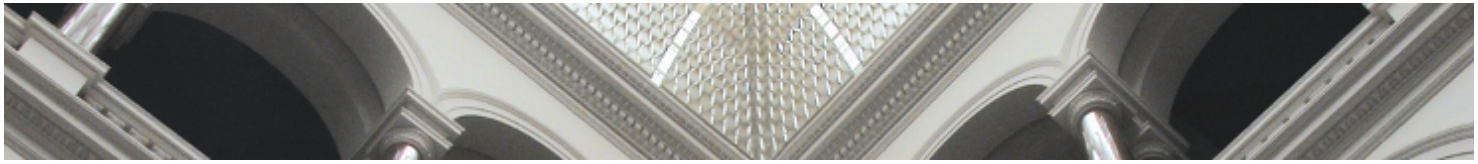
EM algorithm in the parameter space

The red curve is the log likelihood function whose value we wish to maximize.
In the first **E-step** we evaluate the posterior distribution over latent variables.
The blue line is a lower bound $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$ equals the log likelihood at $\boldsymbol{\theta}^{old}$.



In the **M-step**, the bound is maximized giving the value $\boldsymbol{\theta}^{new}$, which gives a larger value of log likelihood than $\boldsymbol{\theta}^{old}$. The subsequent E step then constructs a bound that is tangential at $\boldsymbol{\theta}^{new}$ as shown by the green curve.





Technische Universität Berlin

The EM Algorithm overview

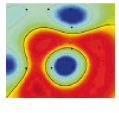
Given a joint distribution $p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ over observed variables $\boldsymbol{X}$ and latent variables $\boldsymbol{Z}$, governed by model parameters $\boldsymbol{\theta}$.

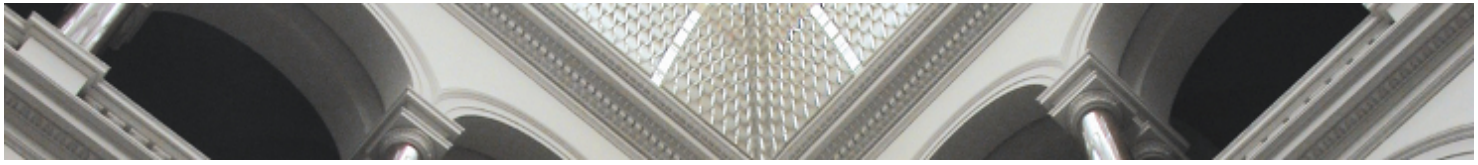
Goal: to maximize the joint likelihood function $p(\boldsymbol{X}|\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$

1. Choose an initial values for model parameters $\boldsymbol{\theta}_{old}$
2. **E-step**: Evaluate posterior $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}_{old})$
3. **M-step**: Evaluate $\boldsymbol{\theta}_{new}$ given by

$$\boldsymbol{\theta}_{new} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\theta}_{old}), \text{ where } \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\theta}_{old}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}_{old}) \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$$

4. Iterate till convergence either of the log likelihood or the parameter values by setting $\boldsymbol{\theta}_{old} \leftarrow \boldsymbol{\theta}_{old}$

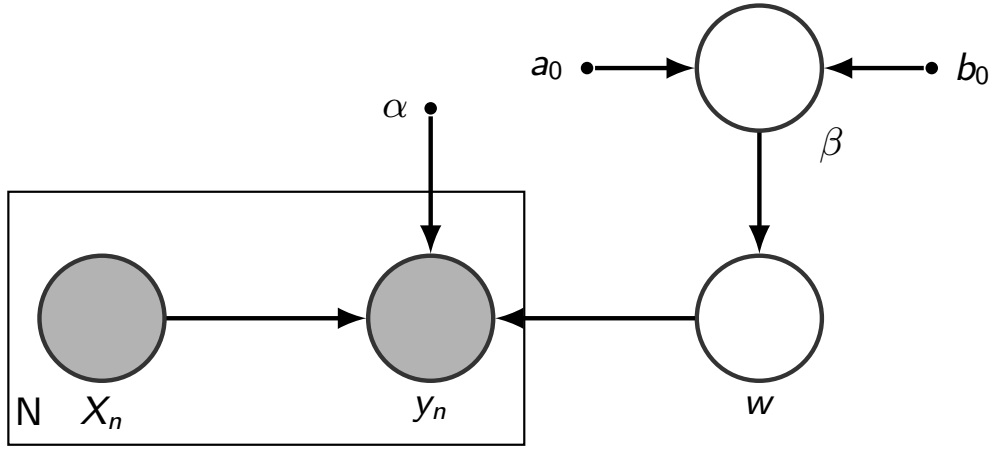




Technische Universität Berlin

Fully Bayesian Linear model

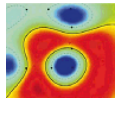
Consider a prior distribution not only over the regression vector w but also over its precision parameter β.

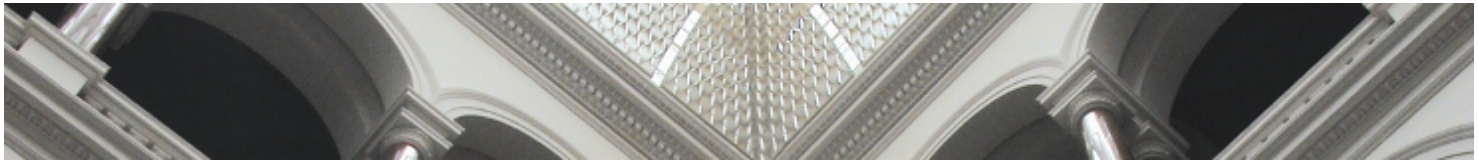


$$p(\beta|a_0, b_0) = \mathcal{G}(\beta|a_0, b_0) \quad \text{(Gamma distribution)}$$

The joint distribution of all variables is given by:

$$p(y, \boldsymbol{w}, \beta|\boldsymbol{X}, \alpha) = p(y|\boldsymbol{w})p(w|\beta)p(\beta)$$



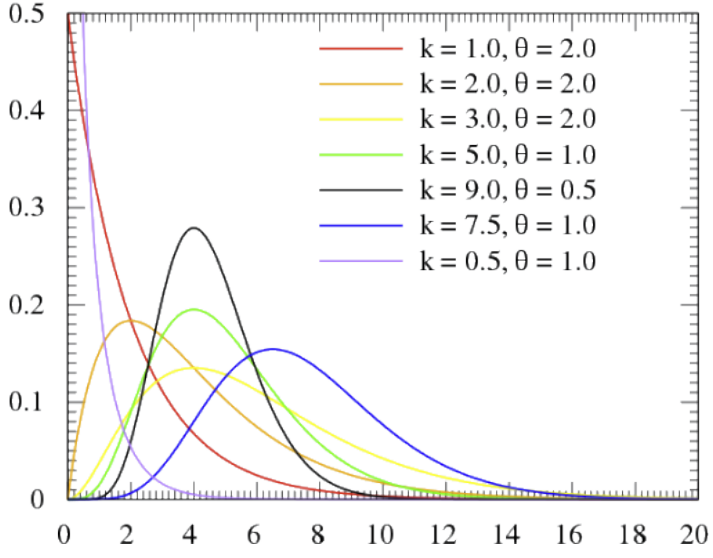


Technische Universität Berlin

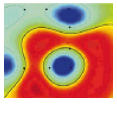
Gamma distribution

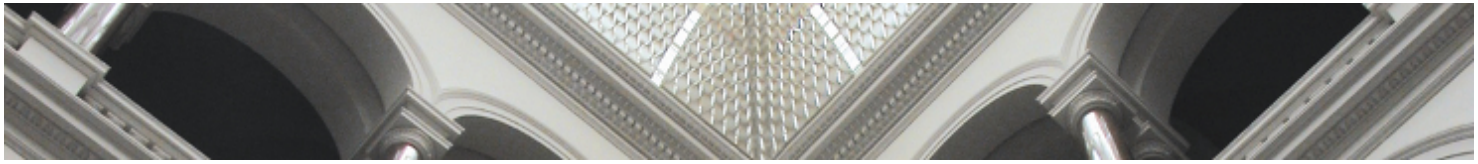
$$\mathcal{G}(x|a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp^{-bx}$$

Gamma PDF with $k = a$ and $\theta = 1/b$



$$\mathbb{E}[x] = \frac{a}{b}, \quad \mathbb{E}[\ln x] = \psi(a) - \ln b$$





Technische Universität Berlin

Non tractable posterior

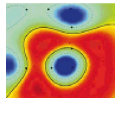
The joint distribution of all variables is

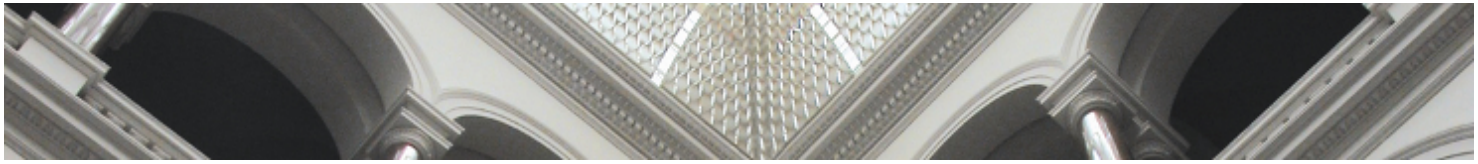
$$p(y, w, \beta|\boldsymbol{X}, \alpha) = \mathcal{N}(y|\boldsymbol{X}w, \alpha^{-1})\mathcal{N}_k(w|0, \beta^{-1}\mathbb{I}_k)\mathcal{G}(\beta|a_0, b_0)$$

The posterior distribution $q(w, \beta|y, X, a, b, \alpha)$ for the fully Bayesian Linear model unfortunately has no closed form solution and therefore intractable.

Goal: to find an approximation to the posterior $p(w, \beta|y, \boldsymbol{X}, a, b, \alpha)$.

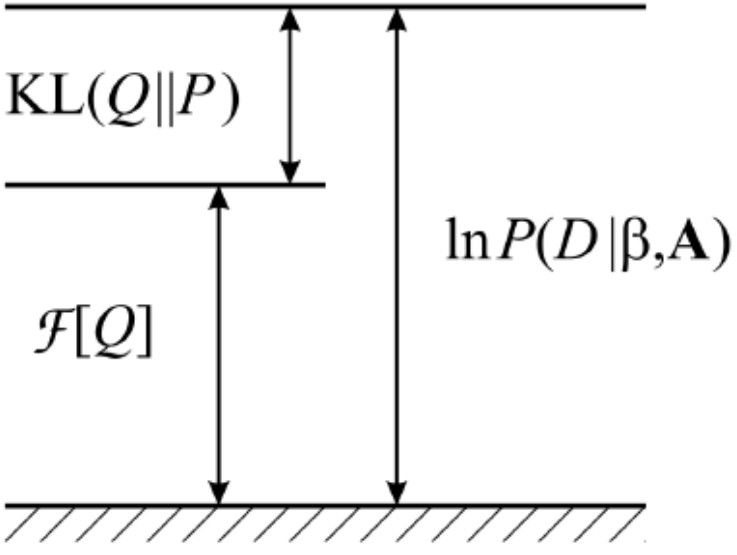
In order to find the approximation we can employ the *Variational Inference* framework.

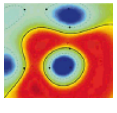


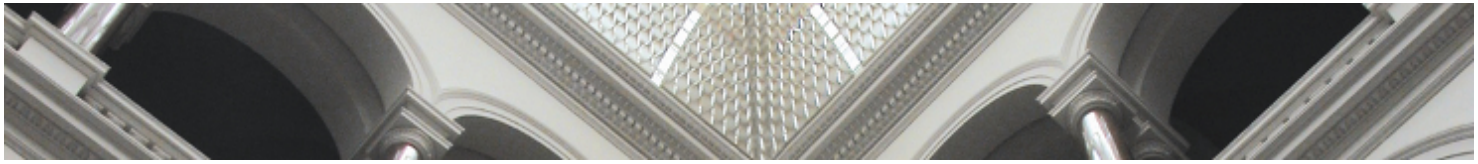


Technische Universität Berlin

Variational Inference







Technische Universität Berlin

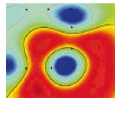
Suppose we have a fully Bayesian model in which all parameters are given prior distributions.

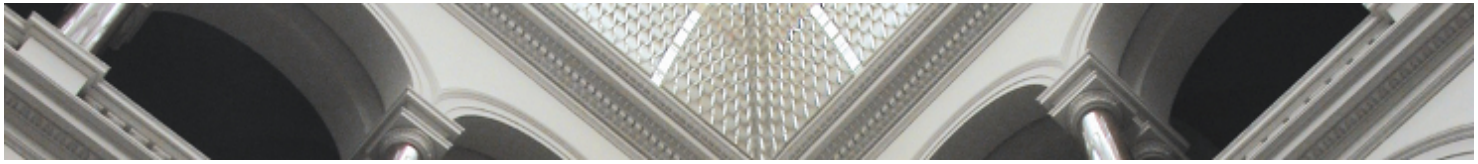
- The set of all observed variables denoted by $\boldsymbol{X}$ and hidden variables by $\boldsymbol{Z}$.
- Our probabilistic model specifies the joint distribution $p(\boldsymbol{X}, \boldsymbol{Z})$, and our goal is to find an approximation for the posterior distribution $p(\boldsymbol{Z}|\boldsymbol{X})$.

Similar to EM Algorithm we can decompose the log marginal probability $p(\boldsymbol{X})$ as following:

$$p(\boldsymbol{X}) = \mathcal{L}(q) + \mathbb{D}_{KL}(p||q)$$

This differs from our discussion of EM only in that the parameter vector $\boldsymbol{\theta}$ no longer appears, because the parameters are now stochastic variables and are absorbed into $\boldsymbol{Z}$.





Technische Universität Berlin

Factorized distributions

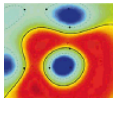
The EM Algorithm cannot be applied because the form of the posterior distribution is intractable.

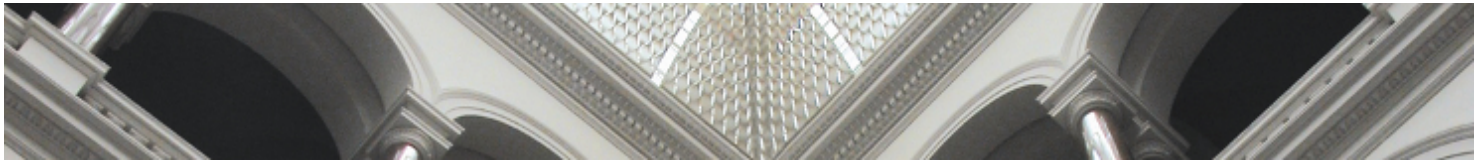
Solution: to partition the elements of $\boldsymbol{Z}$ into disjoint groups that we denote by $\boldsymbol{Z}_i$ where $i = 1, \cdots, M$ so that

$$q(\boldsymbol{Z}) = \prod_{i=1}^{M} q(\boldsymbol{Z}_i) \quad \textit{mean field assumpsion}$$

Optimal solutions for the approximated posterior has following form:

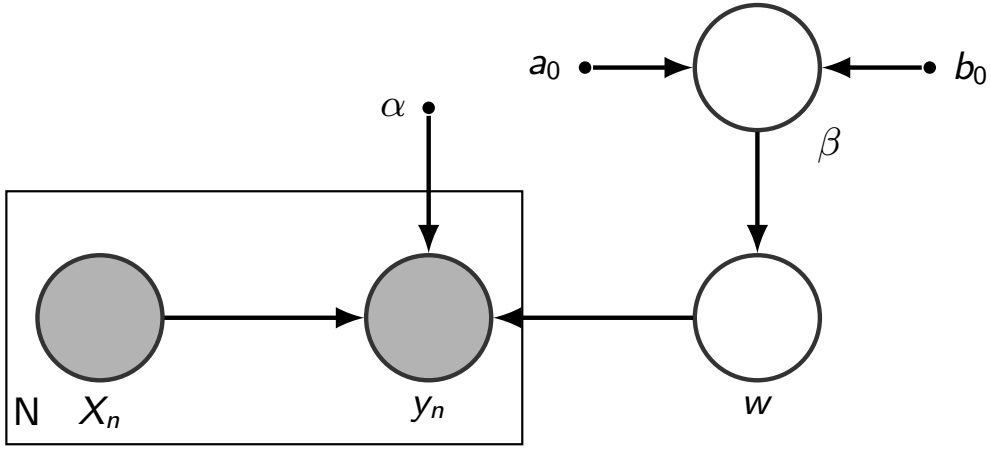
$$\boxed{\ln q_j^*(\boldsymbol{Z}_j) = \mathbb{E}_{i \neq j}\Big[\ln p(\boldsymbol{X}, \boldsymbol{Z})\Big] + const}$$





Technische Universität Berlin

Fully Bayesian Linear model solution



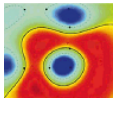
The joint distribution of all variables is

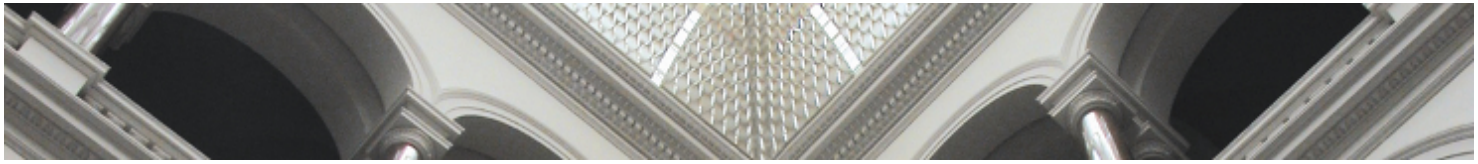
$$p(y, \boldsymbol{w}, \beta | \boldsymbol{X}, \alpha, a_0, b_0) = p(y | w, \boldsymbol{X}, \alpha) p(w | \beta) p(\beta | a_0, b_0)$$

$$q^*(\beta) = \mathbb{E}_w[\ln p(w|\beta)] + \ln p(\beta | a_0, b_0) + const$$

and

$$q^*(w) = \ln p(y | w, \boldsymbol{X}, \alpha) + \ln \mathbb{E}_\beta[p(w|\beta)] + const$$





Technische Universität Berlin

Approximate posterior for precision β

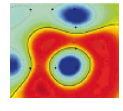
$$\ln q^*(\beta) = \mathbb{E}_w[\ln \mathcal{N}_k(w|0, \beta^{-1}\mathbb{I}_k)] + \ln \mathcal{G}(\beta|a_0, b_0) + const$$

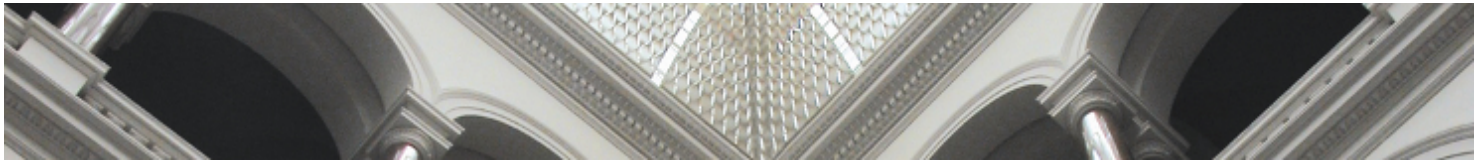
$$\ln q^*(\beta) = \frac{k}{2}\ln\beta - \frac{\beta}{2}\mathbb{E}[w^\top w] + (a_0 - 1)\ln\beta - b_0\beta + const$$

$$\ln q^*(\beta) = (a_0 + \frac{k}{2} - 1)\ln\beta - (b_0 + \frac{\mathbb{E}[w^\top w]}{2})\beta + const$$

$$q^*(\beta) = \mathcal{G}(\beta|a_N, b_N), \quad \text{where}$$

$$\boxed{a_N = a_0 + \frac{k}{2}, \quad b_N = b_0 + \frac{\mathbb{E}[w^\top w]}{2}}$$





Technische Universität Berlin

Approximate posterior for weight vector w

$$q^*(w) = \mathcal{N}_k(w|\mu_N, \Sigma_N), \quad \text{where}$$

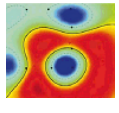
$$\boxed{\mu_N = \alpha \Sigma_N \boldsymbol{X}^\top y, \quad \Sigma_N = \left(\mathbb{E}[\beta]\mathbb{I}_k + \alpha \boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}}$$

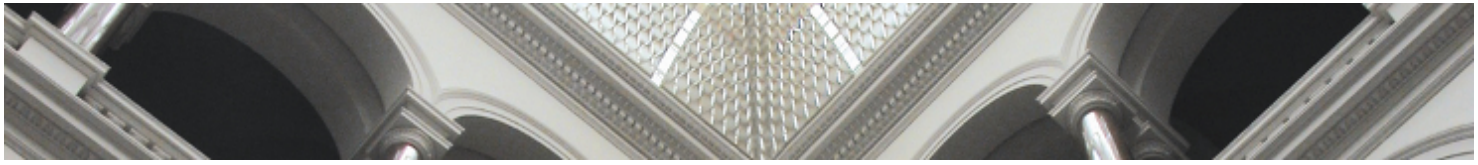
Fist moment of the Gamma distribution:

$$\mathbb{E}[\beta] = \frac{a_N}{b_N}$$

Fist moment of the Gaussian distribution:

$$\mathbb{E}[w^\top w] = \mu_N^\top \mu_N + \Sigma_N$$





Technische Universität Berlin

THANK YOU!