Frequentist Statistics and Bayesian Statistics

Volker Tresp
Winter 2023-2024

Preamble

From Frequencies to Parameters

- From a statistical analysis, we might have derived

$$P(\mathit{Flever} = 1 | \mathit{Flue} = 1) = \theta_{\mathit{Flue}=1}$$

$$P(\mathit{Flever} = 1 | \mathit{Flue} = 1) = \theta_{\mathit{Flue}=0}$$

- With notation $\mathit{Flue} = \mathit{flue}$, $\mathit{flue} \in \{0, 1\}$, we might form $\theta_{\mathit{flue}} = w_0 + w_1 \mathit{flue}$
- Then, one writes more concisely $P_{\mathbf{w}}(\mathit{Flever} = \mathit{fever} | \mathit{Flue} = \mathit{flue})$

Classical Parameterized Statistics

- With many inputs, the number of θ-parameters grows exponentially $(\theta_{x_1,\ldots,x_n})$, whereas the dimensions of $\mathbf{w}$ might only grow linearly $\theta_{x_1,\ldots,x_n} = w_0 + \sum_i w_i x_i$
- Thus estimating $\mathbf{w}$ from data is more data efficient than estimating the θ
- In classical statistics one treats $\mathbf{w}$ as a parameter vector to be estimated $(\hat{\mathbf{w}})$
- We have $P_{\hat{\mathbf{w}}}(\mathit{Flever} = \mathit{fever} | \mathit{Flue} = \mathit{flue})$
- Samples are often treated as i.i.d.

Bayesian Statistics

- In Bayesian statistics, one conditions, $P(\mathit{Flever} = \mathit{fever} | \mathit{Flue} = \mathit{flue}, \mathbf{w})$ and one treats $\mathbf{w}$ as a vector of random variables (just like the other random variables)
- In a Bayesian approach we can ask the expert about the dependency
- As a prior, it might be reasonable to ask the expert to specify her/his prior belief as a $\theta_{prior} \pm \epsilon$ with some tolerance
- It is more difficult to ask the expert to specify her/his prior belief as $w_{prior} \pm \epsilon$
- As we will see, we will often assume that a priori $\mathbf{w} \sim \mathcal{N}(0, \alpha^2 I)$

Bayesian Statistics (cont'd)

- Before we see any data, the apriori prediction then is

$$P(\textit{fever}|\textit{flue}) = \int P(\textit{fever}|\textit{flue}, \mathbf{w})P(\mathbf{w})d\mathbf{w}$$

  $P(\mathbf{w})$ is the *prior* distribution

- After we have observed the data

$$P(\textit{fever}|\textit{flue}, D) = \int P(\textit{fever}|\textit{flue}, \mathbf{w})P(\mathbf{w}|D)d\mathbf{w}$$

  $P(\mathbf{w}|D)$ is the *a posterori* distribution

- The i.i.d requirement becomes the requirement for "an exchangeable sequence of random variables"

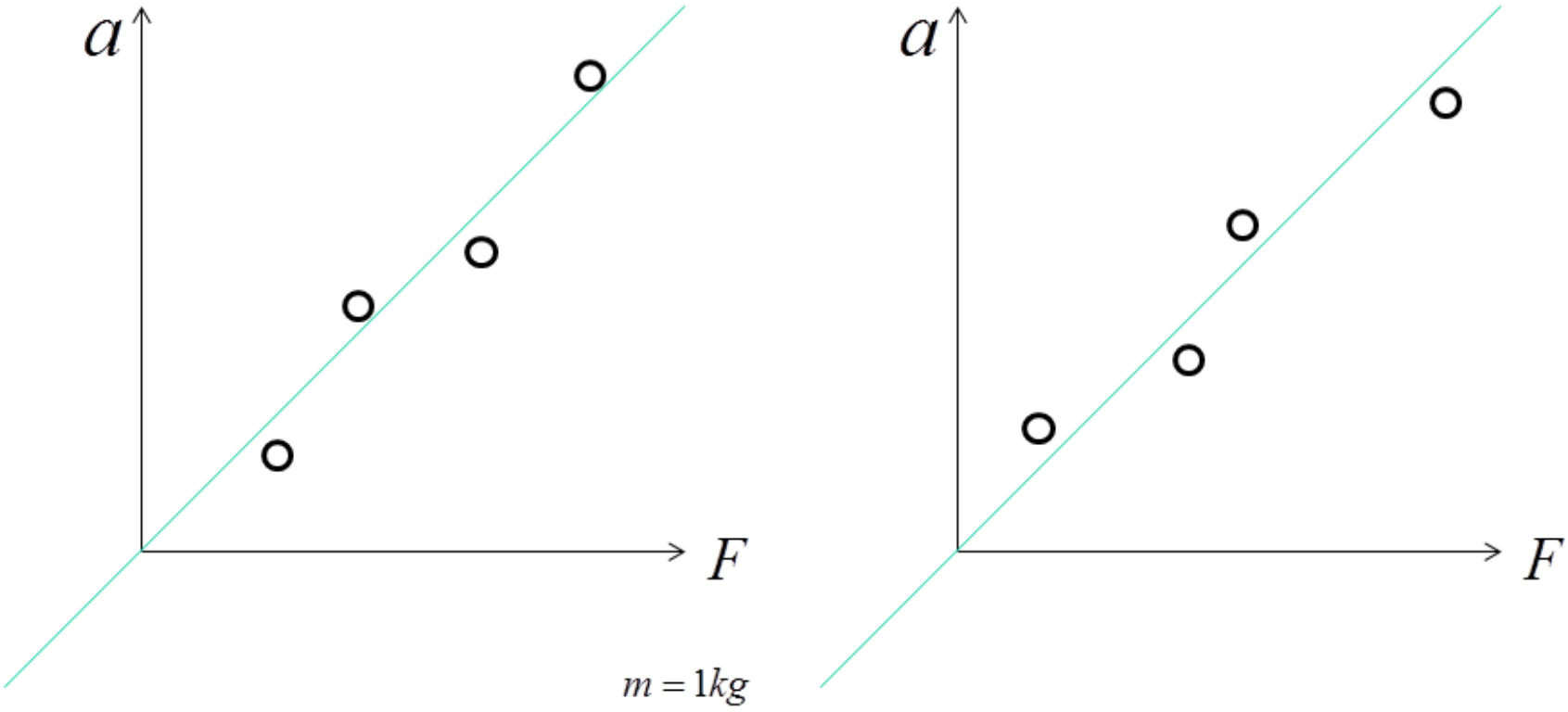
Frequentist Statistics

Approach

- Natural science attempts to find regularities and rules in nature

$$F = ma$$

- The laws are valid under idealized conditions. Example: Fall of a point object without air friction, with velocities much smaller than the speed of light
- There might be measurement errors, but there is an underlying true (simple) dependency
- This motivates the frequentist statistics: *derivation of probabilistic statements under repeatable experiments under identical conditions*



Repeated experiments with an underlying linear dependency

Basic Terms

- Thus a statistical analysis requires a precise description of the experiment. For example, the details on who gets which medication (randomized?)
- A **statistical unit** is an object, on which measurements are executed (attributes are registered). Could be a person. A statistical unit defines a row in the data matrix, the attributes define the columns
- The population is the conceptual set of all statistical units about which we want to perform statistical inference. Example: diabetics
- For the analysis, only a sample is available (training data). Often it is assumes that the sample is a random subset of the population

Population

- A population can be finite, infinite, or hypothetic
- Example: all people who vote in an election

Typical Assumption

- The sample D is a random subset of the population
- For each statistical unit i in the sample, we determine the attributes (features) $\mathbf{x}_i$
- Assuming a random sample, we can write (in a finite sample, we would assume sampling with replacement) with $P(\cdot)$ known

$$P(D) = P(\mathbf{x}_1, ..., \mathbf{x}_N) = \prod_{i=1}^{N} P(\mathbf{x}_i)$$

- The probability that I sample N units with attributes $\mathbf{x}_1, ..., \mathbf{x}_N$ is the product of the probabilities of observing individual units with their individual attributes

Modelling

- $P(\mathbf{x})$ is unknown
- Assumption in parametric modelling: The data has been generated by a probability distribution $P_{\mathbf{w}}(\mathbf{x})$, which is parameterized by the parameter vector $\mathbf{w}$. For example, we might assume a Gaussian distribution with unknown mean and variance.
- Thus we assume that for at least one parameter vector $\mathbf{w}$

$$P_{\mathbf{w}}(\mathbf{x}) \approx P(\mathbf{x})$$

- The goal is to estimate the parameter vector

Example: a Person's Height

- We assume that the height x is Gaussian distributed with unknown mean and variance

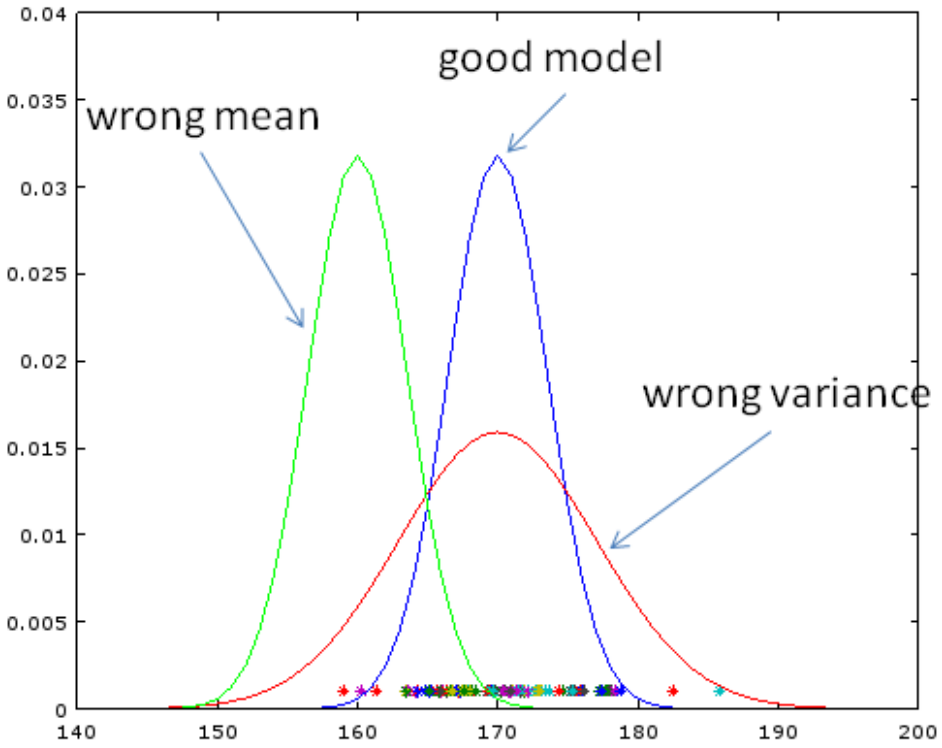
$$P_{\mathbf{w}}(x) = \mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

  with $\mathbf{w} = (\mu, \sigma)^T$

- Thus we get

$$P_{\mathbf{w}}(x_1, ..., x_N) = \prod_{i=1}^{N} P_{\mathbf{w}}(x_i) = \prod_{i=1}^{N} \mathcal{N}(x_i; \mu, \sigma^2)$$

$$= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2\right)$$



How can we define and find the best parameters?

Maximum Likelihood

- We consider the probability of the observed data as a function of the parameters. This is the likelihood-function, where we assume that data points where generated independently

$$L(\mathbf{w}) = P_{\mathbf{w}}(\mathbf{x}_1, ..., \mathbf{x}_N) = \prod_{i=1}^{N} P_{\mathbf{w}}(\mathbf{x}_i)$$

- It is often more convenient to work with the log-likelihood,

$$l(\mathbf{w}) = \log L(\mathbf{w}) = \sum_{i=1}^{N} \log P_{\mathbf{w}}(\mathbf{x}_i)$$

- The maximum likelihood (ML) estimator is given by

$$\hat{\mathbf{w}}_{ml} \doteq \arg\max(l(\mathbf{w}))$$

- This means: in the family of distributions under considerations, the ML estimator is the one which explains the data the best

ML-estimator for Person's Height

- The ML estimators are

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

and

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})^2$$

ML-Estimator for a Linear Model

- We are interested in the conditional $P(y|\mathbf{x})$; let' assume that the true dependency is linear, but we only have available noisy target measurements

$$y_i = \mathbf{x}_i^T \mathbf{w} + \epsilon_i$$

- Let's further assume that the noise is Gaussian distributed

$$P(\epsilon_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\epsilon_i^2\right)$$

- It follows that

$$P_{\mathbf{w}}(y_i|\mathbf{x}_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mathbf{x}_i^T \mathbf{w})^2\right)$$

- It is easier to deal with the log

$$\log P_{\mathbf{w}}(y_i|\mathbf{x}_i) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y_i - \mathbf{x}_i^T \mathbf{w})^2$$

ML Estimator

- The log-likelihood function is then

$$l = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \mathbf{x}_i^T\mathbf{w})^2$$

- The first term does not depend on $\mathbf{w}$. Under the assumption of independent additive noise, the ML estimator is the same as the LS estimator

$$\hat{\mathbf{w}}_{ml} \doteq \arg\max(l(\mathbf{w})) = \hat{\mathbf{w}}_{LS}$$

Since, $\hat{\mathbf{w}}_{ml} = \arg\max[-\sum_i(y_i - \mathbf{x}_i^T\mathbf{w})^2]$ and $\hat{\mathbf{w}}_{ls} = \arg\min[\sum_i(y_i - \mathbf{x}_i^T\mathbf{w})^2]$

Analysis of Estimators

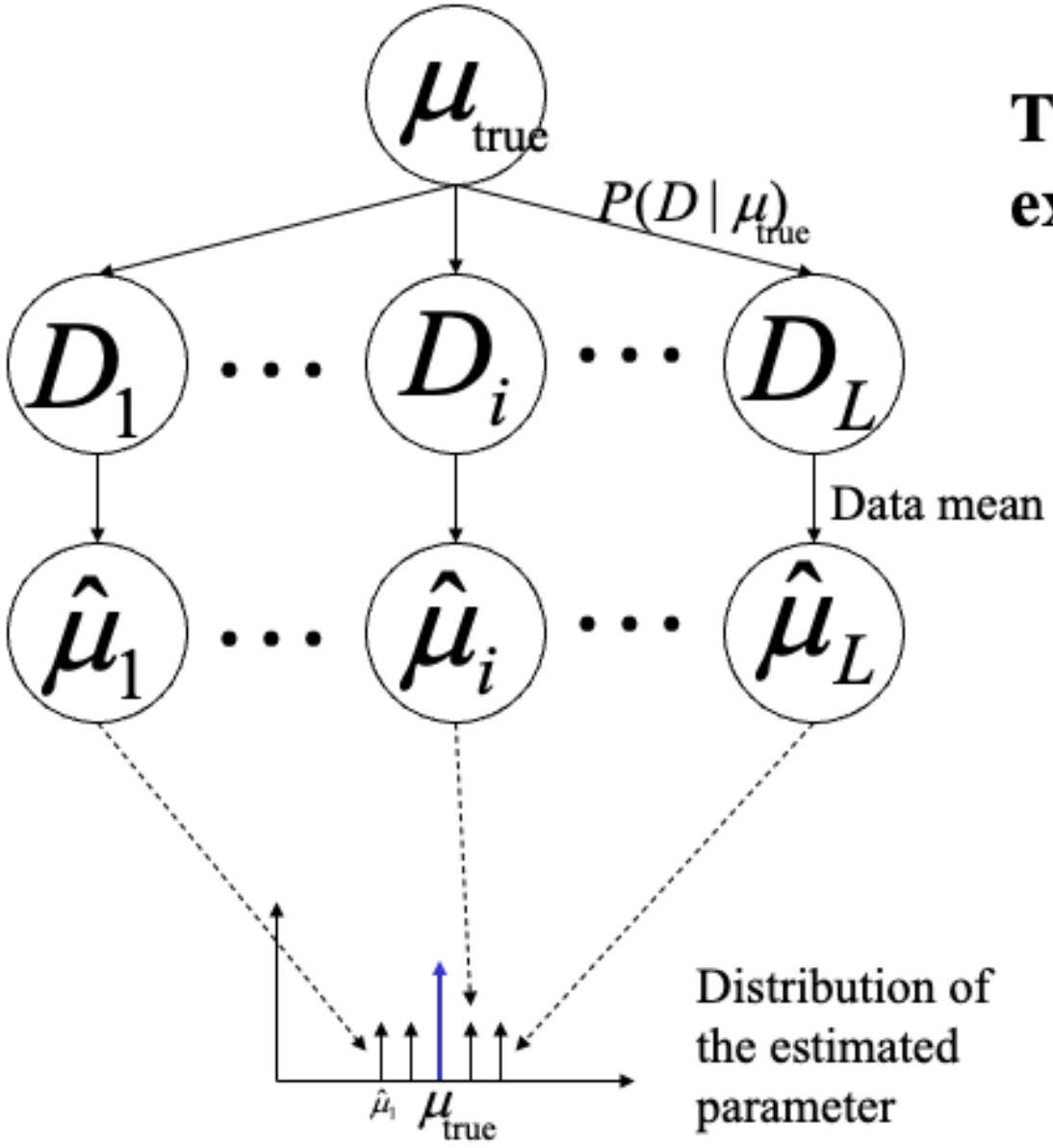
- Certainly the ML estimator makes sense (best fit). But how certain are we about the estimates. Maybe there are parameter values that would give us almost the same likelihood?
- To analyse the ML estimate we do the following thought experiment (see next slide)
- Let μ be the unknown but fixed parameter
- In addition to the available sample we are able to generate additional samples $D_1, D_2, \ldots D_L$, $L \to \infty$, each of size N
- For each of these D_i, we estimate the parameter and obtain $\widehat{\mu}_i$ (for example, using the ML-estimator)

Analysis of Estimators (cont'd)

- We analyse the distribution of the estimated parameter
- In the example, we get for the mean person height (with known σ^2)

$$P_\mu(\widehat{\mu} - \mu) = \mathcal{N}\left(\widehat{\mu} - \mu; 0, \frac{\sigma^2}{N}\right)$$

- The interpretation of probability here is: averaged of all $D_1, D_2, \ldots D_L$
- We can calculate this distribution of the difference between estimated and true parameter without knowing any particular data set (although I need σ^2)
- Assuming, we estimate $\widehat{\mu}$ from the available sample, we can answer the question: how probable is it to measure $\widehat{\mu}$ if the true value is $\mu = 175cm$?



The frequentist experiment

Distribution of the estimated parameter

$$P(\hat{\mu} \mid \mu)_{true} \propto \mathbf{N}\left(\mu_{true}, \sigma^2/N\right)$$

Bias of an Estimator

- The difference between the true parameter and the expected value of the parameter estimate (averaged over many data sets of size N) is called the bias

$$\text{Bias}[\widehat{w}] = E_D(\widehat{w}) - w_{true}$$

Here,

$$E_D(\widehat{w}) = \lim_{L\to\infty} \frac{1}{L} \sum_{i=1}^{L} \widehat{w}_{D_i}$$

In the example, the bias is zero for the mean

The ML-Estimator can be Biased with finite Data

- The ML-estimator can be biased with finite data

$$\hat{\sigma}^2_{ml} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \hat{\mu})^2$$

$$\hat{\sigma}^2_{unbiased} = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x}_i - \hat{\mu})^2$$

Variance of an Estimator

- The variance indicates how much an estimator varies around its mean

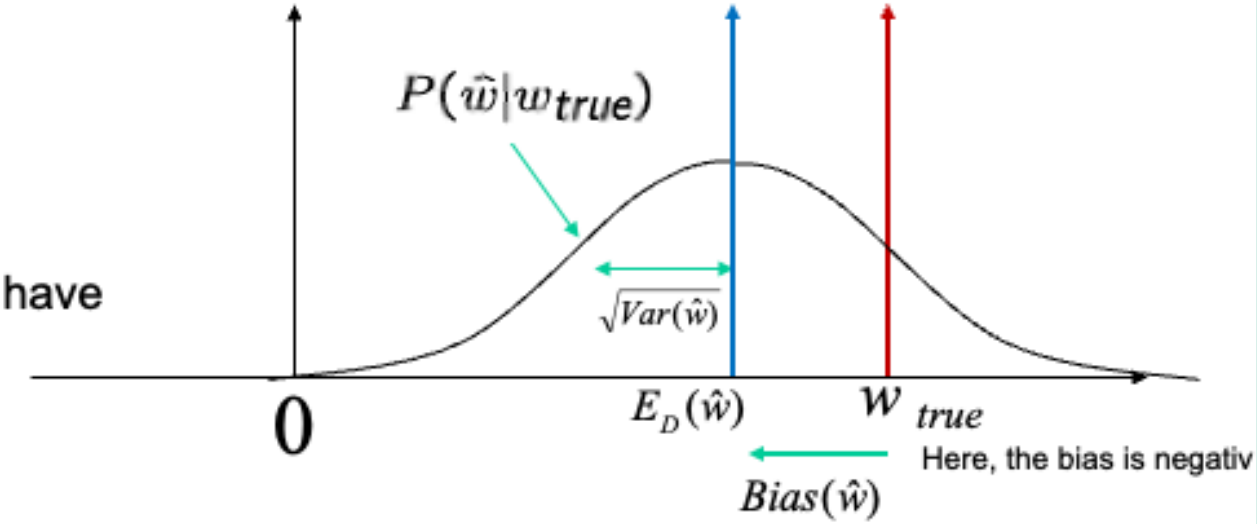
$$\mathrm{Var}[\hat{w}] = E_D\left(\hat{w} - E_D(\hat{w})\right)^2$$

$$\mathrm{Var}[\hat{w}] = \lim_{L\to\infty} \frac{1}{L} \sum_{i=1}^{L} \left(\hat{w}_{D_i} - E_D(\hat{w})\right)^2$$

- In the example: $\mathrm{Var}[\hat{w}] = \sigma^2/N$

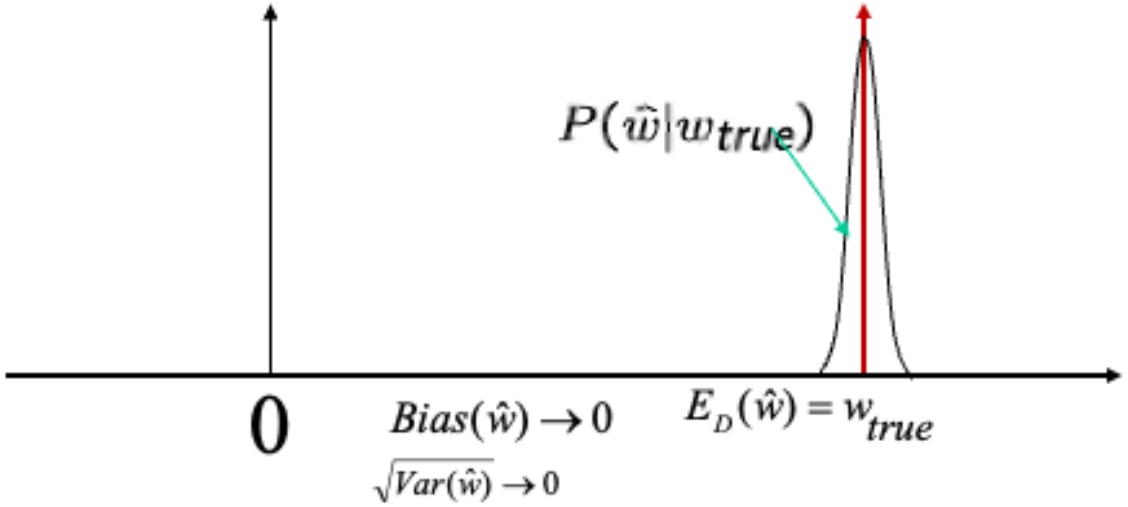
For finite N

The ML estimator can have a finite bias



For $N \to \infty$

The ML estimator is unbiased



Expected Error

- The expected mean squared error evaluates the deviation of the estimator from the **true parameter**

$$\mathsf{MSE}[\hat{w}] = E_D\left(\hat{w} - w_{true}\right)^2$$

$$\mathsf{MSE}[\hat{w}] = \lim_{L\to\infty} \frac{1}{L} \sum_{i=1}^{L} \left(\hat{w}_{D_i} - w_{true}\right)^2$$

- The expected mean squared error is the sum of the variance and the square of the bias

$$\mathsf{MSE}[\hat{w}] = \mathsf{Var}[\hat{w}] + \mathsf{Bias}^2[\hat{w}]$$

Expected Error (cont'd)

Proof:

$$\mathsf{MSE}[\hat{w}] = E_D\left(\hat{w} - w_{true}\right)^2 = E_D\left[(\hat{w} - E_D(\hat{w})) - (w_{true} - E_D(\hat{w}))\right]^2$$

$$= E_D\left(\hat{w} - E_D(\hat{w})\right)^2 + E_D\left(w_{true} - E_D(\hat{w})\right)^2$$

$$-2E_D\left[(\hat{w} - E_D(\hat{w}))(w_{true} - E_D(\hat{w}))\right] = \mathsf{Var}[\hat{w}] + \mathsf{Bias}^2[\hat{w}] + 0$$

The cross term is zero since

$$E_D\left[(\hat{w} - E_D(\hat{w}))(w_{true} - E_D(\hat{w}))\right] =$$

$$\left(w_{true} - E_D(\hat{w})\right)\, E_D(\hat{w} - E_D(\hat{w})) = 0$$

Desirable Properties of Estimators

- An estimator is unbiased, if $\mathsf{Bias}[\widehat{w}] = 0$
- An estimator is asymptotically unbiased, if $\mathsf{Bias}[\widehat{w}] = 0$, for $N \to \infty$
- An estimator is MSE consistent, if we have

$$\mathsf{MSE}[\widehat{w}]_{N\to\infty} \to 0$$

- An estimator $\widehat{w}$ ist MSE-efficient, if

$$\mathsf{MSE}[\widehat{w}](Estimator) \leq \mathsf{MSE}[\widehat{w}](Estimator') \quad \forall Estimator'$$

Properties of the ML-Estimator

- The ML-estimator has many desirable properties:
  - The ML-estimator is asymptotically $N \to \infty$ **unbiased** (although with a finite sample size it might be biases)
  - Maybe surprisingly, the ML estimator is asymptotically ($N \to \infty$) MSE-efficient among all unbiased estimators
  - Asymptotically, the estimator is Gaussian distributed, even when the noise is not!
- The analysis generalises to a parameter vector

Estimating the Variance via Bootstrap

- In particular for complex models it might be difficult to derive the sampling distribution, for example the distribution of the ML parameter estimate
- Recall that ideally we would have many training sets of the same size available, fit the model, and observe the distribution of the parameter estimates
- Proxies for the new data sets of the same size N can be generated surprisingly simple: A new data set can be generated by sampling N times from the original data with replacement

Classical Statistical Inference

- For hypothesis testing and the derivation of error bounds, please consult your favorite statistics book.

Discussion: ML

- The likelihood can be calculated even for complex models models (e.g., models with latent variables)
- With the assumption that the data haven been generated independently, the log-likelihood is the sum over the log likelihoods of individual data points

$$l(\mathbf{w}) = \sum_{i=1}^{N} \log P(y_i|\mathbf{w})$$

- Thus a log-likelihood defines a cost function (cross-entropy cost function)

$$cost_i(\mathbf{w}) = -\log P(y_i|\mathbf{w})$$

Discussion: ML (cont'd)

- The necessity to emulate the data generating process leads to interesting problem specific models
- A certain problem: One needs to assume that the true model is (approximately) in the class of the models under considerations.
- With finite data, the ML estimator can lead to over fitting: more complex models will have a higher likelihood
- The frequentist statistics has a strong focus in the analysis of the properties of parameter estimates

Violations of IID

- The following decomposition assumes that the data points are independent and identically distributed (IID, or i.i.d.)

$$L(\mathbf{w}) = P_{\mathbf{w}}(\mathbf{x}_1, ..., \mathbf{x}_N) = \prod_{i=1}^{N} P_{\mathbf{w}}(\mathbf{x}_i)$$

- Statistical analysis under IID is well studied
- For more complex sampling situations, as in time-series modelling of for graph data, the i.i.d. principle can often not be applied, but one can still define a likelihood for the observed data and one can obtain an ML estimate
- The generalization to new data is often nontrivial and is case specific
- Examples: a social network model where new individuals become known; the generalization of a social network, developed for one university, to another university

Bayesian Statistics

The Bayesian Approach

- In a frequentist setting, the parameters are fixed but unknown and the data are generated by a random process
- In a Bayesian approach, also the parameters have been generated by a random process
- This means we need an *a priori* distribution

$$P(\mathbf{w})$$

- The we obtain a complete probabilistic model

$$P(\mathbf{w})P(D|\mathbf{w})$$

- ... and can calculate the posterior parameter distribution using Bayes' formula as

$$P(\mathbf{w}|D) = \frac{P(D|\mathbf{w})P(\mathbf{w})}{P(D)}$$

The Prior

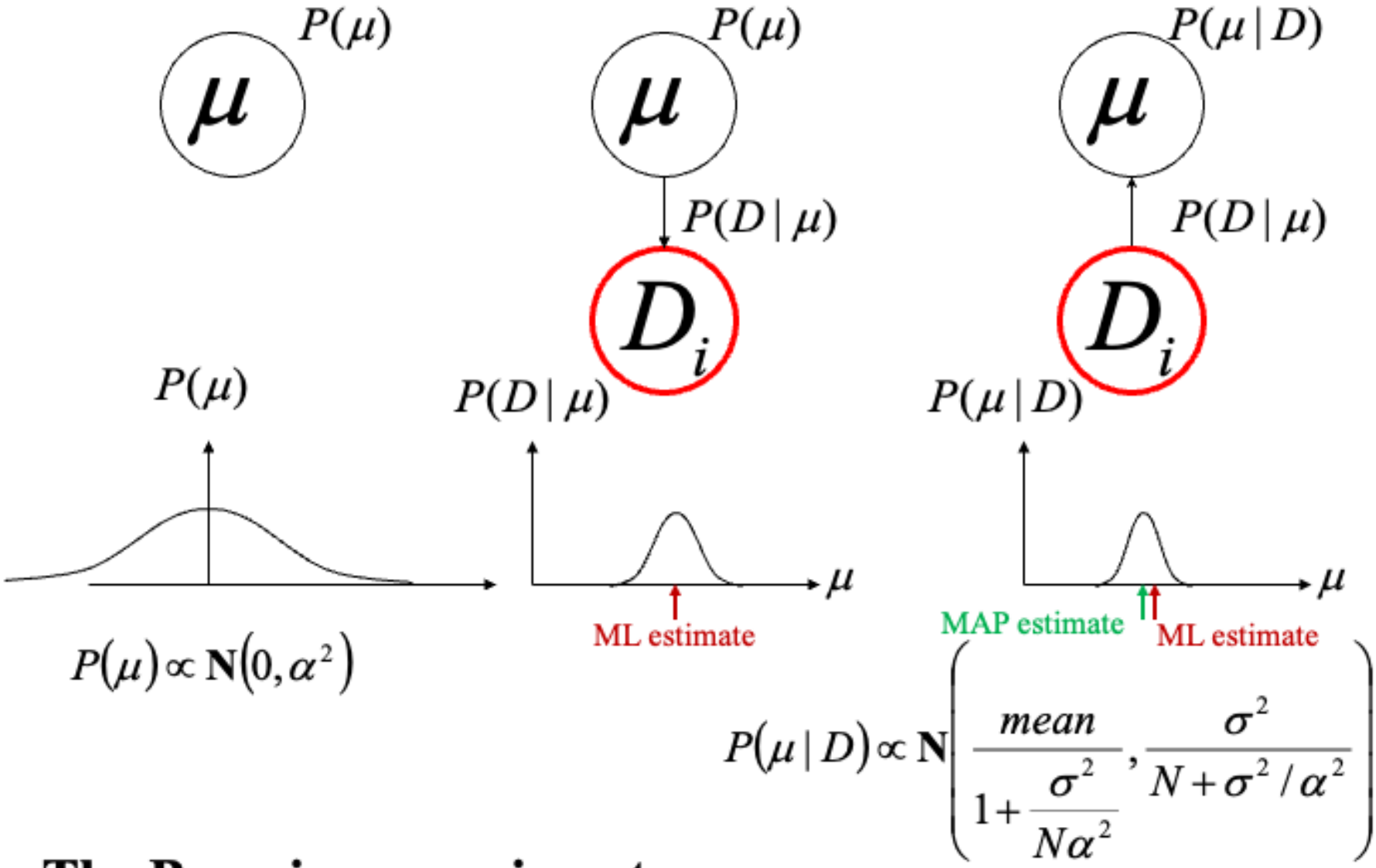
- Does it make sense to assume a personal $P(\mathbf{w})$?
- Cox (1946): If one is willing to assign numbers to ones personal beliefs, then one arrives, under few consistent conditions, at the Bayesian formalism

The Bayesian Experiment

- In contrast to the frequentist experiment, we only work with the actual data D and do not need to assume that additional hypothetical data sets can be generated
- One assume that the true parameter μ has been generated from the prior distribution $P(\mu)$ in one experiment. In the example: $P(\mu) = \mathcal{N}(\mu; 0, \alpha^2)$
- The data are generated from $P(D|\mu)$, in the example $P(D|\mu) = \prod_i \mathcal{N}(x_i; \mu, \sigma^2)$
- Applying Bayes' formula I get the *a posteriori* distribution

$$P(\mu|D) = \frac{P(D|\mu)P(\mu)}{P(D)} = \mathcal{N}\left(\mu; \frac{mean}{1 + \frac{\sigma^2}{N\alpha^2}}, \frac{\sigma^2}{N + \sigma^2/\alpha^2}\right)$$

with $mean = 1/N \sum_{i=1}^{N} x_i$



The Bayesian experiment

Analysis

- The Bayesian approach gives you the complete a posteriori parameter distribution
- One can derive a maximum *a posteriori* estimator as,

$$\hat{\mathbf{w}}_{map} \doteq \arg\max(P(\mathbf{w}|D))$$

In the example,

$$\hat{\mu}_{MAP} = \frac{mean}{1 + \frac{\sigma^2}{N\alpha^2}}$$

- Note, that the MAP estimator converges to the ML estimator, for $N \to \infty$

Our Favorite Example: Linear Regression

- Assume, that the true dependency is linear but that we only measure noisy target data

$$y_i = \mathbf{x}_i^T \mathbf{w} + \epsilon_i$$

We get (same as in the frequentist approach)

$$P(y_i|\mathbf{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mathbf{x}_i^T \mathbf{w})^2\right)$$

Linear Regression: a priori Assumption

- A convenient *a priori* assumption is that

$$P(\mathbf{w}) = (2\pi\alpha^2)^{-M/2} \exp\left(-\frac{1}{2\alpha^2}\sum_{i=0}^{M-1} w_i^2\right)$$

- We give smaller parameters a higher *a priori* probability
- Ockhams razor: simple explanations should be preferred
- We will assume that the hyperparameters σ^2 and α^2 are known. If they are unknown, one can define prior distributions for those. The analysis becomes more involved

Linear Regression: the a posteriori Distribution

- Using the likelihood-function and the prior parameter distribution, we can apply Bayes' formula and obtain the *a posteriori* distribution

$$P(\mathbf{w}|D) = \frac{P(\mathbf{w})P(D|\mathbf{w})}{P(D)}$$

Linear Regression: Calculating the a posteriori Distribution

$$P(\mathbf{w}|D) = \frac{P(\mathbf{w})P(D|\mathbf{w})}{P(D)} \propto \exp\left(-\frac{1}{2\alpha^2}\sum_{j=0}^{M-1} w_i^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \mathbf{x}_i^T\mathbf{w})^2\right)$$

This can be written as

$$P(\mathbf{w}|D) = \mathcal{N}\left(\mathbf{w}; \mathbf{w}_{map}, cov(\mathbf{w}|D)\right)$$

With

$$\mathbf{w}_{map} = \left(\mathbf{X}^T\mathbf{X} + \frac{\sigma^2}{\alpha^2}I\right)^{-1}\mathbf{X}^T\mathbf{y}$$

and covariance

$$cov(\mathbf{w}|D) = \sigma^2\left(\mathbf{X}^T\mathbf{X} + \frac{\sigma^2}{\alpha^2}I\right)^{-1}$$

Linear Regression: the MAP estimate and the PLS-solution

- The most probable parameter value, after observing the data, is (the maximum *a posteriori* (MAP) estimate)

$$\hat{\mathbf{w}}_{map} \doteq \arg\max(P(\mathbf{w}|D)) = \hat{\mathbf{w}}_{Pen}$$

with $\lambda = \frac{\sigma^2}{\alpha^2}$.

- One sees that despite different experimental assumptions the frequentist ML estimate and the Bayesian MAP estimate are very similar. The ML estimate corresponds to the LS-solution and the MAP estimate corresponds to the PLS solution

Bayesian Prediction with Linear Regression

- An important difference between is prediction. In a frequentist solution one substitutes the parameter estimate $\widehat{y}_i = \mathbf{x}_i^T \mathbf{w}_{ml}$, and one can calculate the variance in the prediction. In a Bayesian approach one applies the rules of probability and marginalizes (integrates over) the parameters

- With

$$P(y, \mathbf{w}|\mathbf{x}, D) = P(\mathbf{w}|D)P(y|\mathbf{w}, \mathbf{x})$$

it follows that

$$P(y|\mathbf{x}, D) = \int P(y|\mathbf{w}, \mathbf{x})P(\mathbf{w}|D)d\mathbf{w}$$

Predictive Distribution for a Linear Model

- The *a posteriori* predictive distribution becomes

$$P(y|\mathbf{x}, D) = \int P(y|\mathbf{w}, \mathbf{x}) P(\mathbf{w}|D) d\mathbf{w}$$

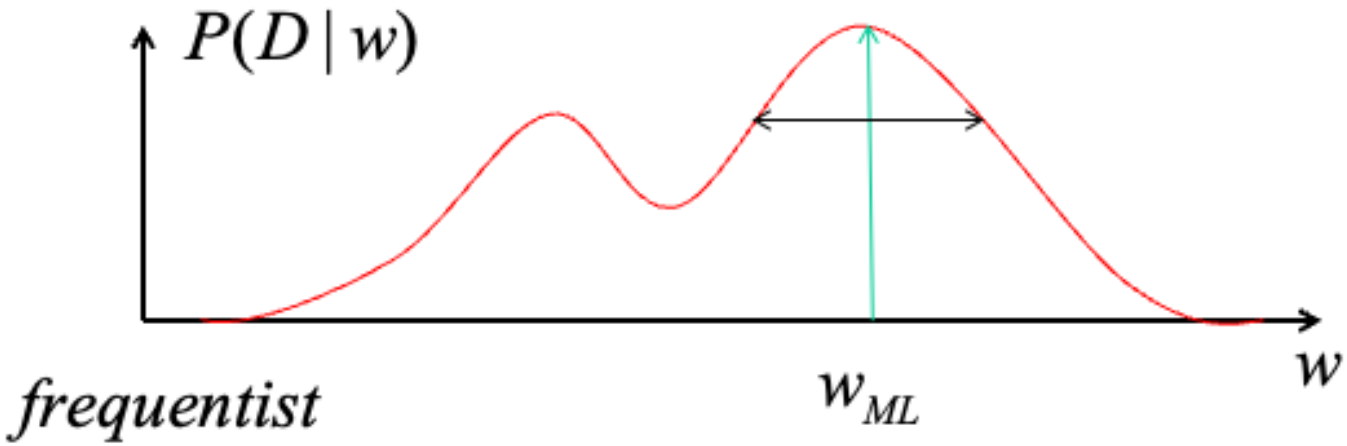
$$= \mathcal{N}\left(y;\ \mathbf{x}^T \widehat{\mathbf{w}}_{map},\ \mathbf{x}^T cov(\mathbf{w}|D)\ \mathbf{x} + \sigma^2\right)$$

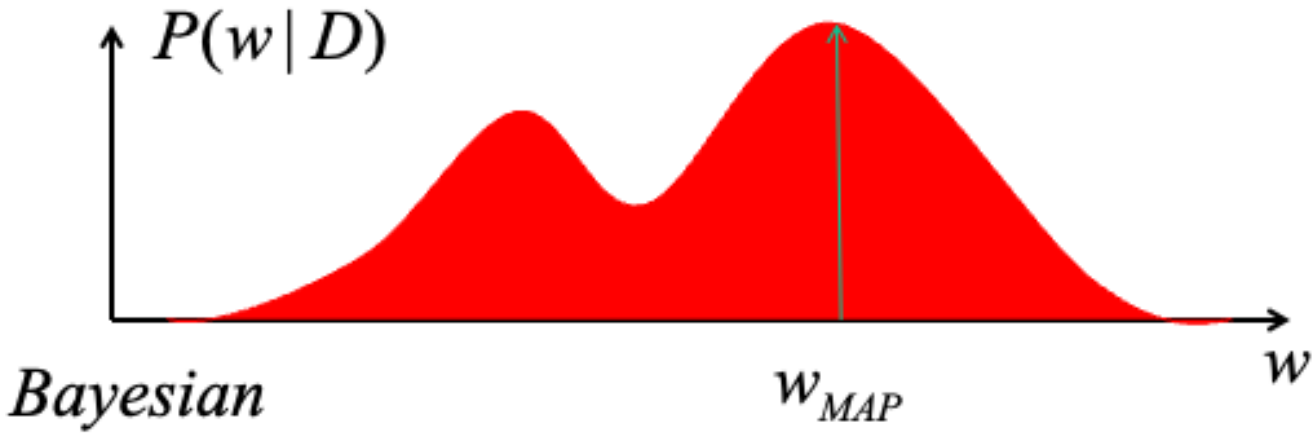
  and is Gaussian distributed with mean $\mathbf{x}^T \widehat{\mathbf{w}}_{map}$ and variance $\mathbf{x}^T cov(\mathbf{w}|D)\mathbf{x} + \sigma^2$

- The variance on the prediction considers both the noise on the prediction as well as the uncertainty in the parameters (by integrating over possible values)

- This is an essential advantage of the Bayesian approach: one considers all plausible parameter values and, e.g., one can also consider all local optima in the integral

- This is also the main technical challenge: for the Bayesian solution complex integrals need to be solved or approximated


P(D | w)
w
frequentist
w_ML


P(w | D)
w
Bayesian
w_MAP

Discussion: the Bayesian Solution

- Personal belief is formulated as a probability distribution
- Consistent approach for various kinds of modeling uncertainty
- For basic distributions (Gaussian, Poisson, Dirichlet, ...) which belong to the *exponential family of distributions*, closed form solutions for the complete Bayesian approach are available!
- For more complex models, a predictive analysis leads to integrals which often cannot be solved analytically
- Special approximations: Monte-Carlo integration, evidence framework
- The simplest approximation is

$$P(y|\mathbf{x}, D) = \int P(y|\mathbf{w}, \mathbf{x}) P(\mathbf{w}|D) d\mathbf{w} \approx P(y|\mathbf{x}, \mathbf{w}_{map})$$

which means that one uses a MAP point estimate