Machine Learning

Bayesian Decision Theory, Maximum Likelihood, and Regularized Empirical Risk Minimization

Prof. Matthias Hein

Machine Learning Group
Department of Mathematics and Computer Science
Saarland University, Saarbrücken, Germany

Lecture 5, 06.11.2013

Bayesian interpretation of loss functions

Regression: $y_i = f(x_i) + \varepsilon_i$,

Bayesian: Model for distribution $p(y|X = x, f) \rightarrow$ model for distribution of ε.

Likelihood: $p(y|X = x, f)$ models: how *likely* is the output y given the point x and the function value $f(x)$?

Maximum likelihood estimation:
Use the function f which maximizes the likelihood.

$$f^*(x) := \operatorname*{arg\,max}_{f \in \mathcal{F}} \mathbb{E}_{Y|X=x}\Big[p(Y|X = x, f)\Big]$$

Bayesian interpretation of loss functions II

Correspondence to loss function:

$$L(y, f(x)) = -\log p(y|X = x, f(x)) + c,$$

$\Rightarrow$ Maximizing the likelihood $p(y|X = x, f(x))$ is **equivalent** to minimizing the loss $L(y, f(x))$.

Example:

$$p(y|X = x, f(x)) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(y - f(x))^2}{2\sigma^2} \Big).$$

and so the corresponding loss function $L(y, f(x))$ is given as

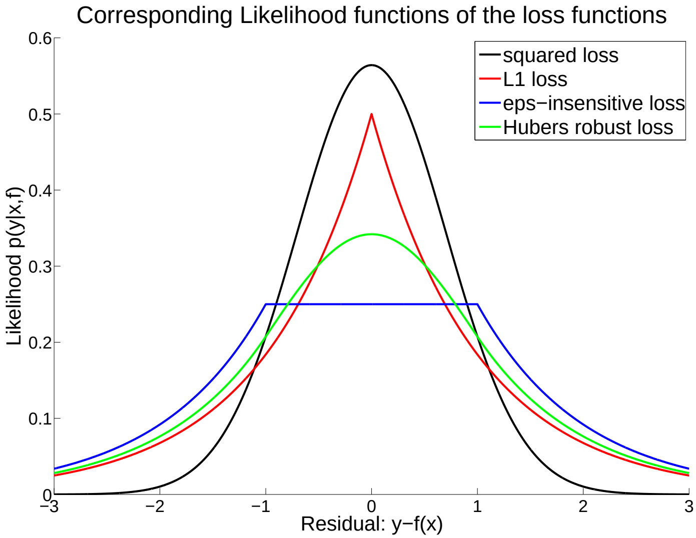
$$L(y, f(x)) = -\log p(y|X = x, f(x)) = \log(\sqrt{2\pi\sigma^2}) + \frac{\big(y - f(x)\big)^2}{2\sigma^2}.$$

Bayesian interpretation of loss functions III

With: $\varepsilon = y - f(x)$,

	$L(\lvert\varepsilon\rvert)$	$p(y\lvert x, f(x))$
squared loss	$\lvert\varepsilon\rvert^2$	$\frac{1}{\sqrt{\pi}} e^{-(y-f(x))^2}$,
L_1 - loss	$\lvert\varepsilon\rvert$	$\frac{1}{2} e^{-\lvert y-f(x)\rvert}$,
σ-insensitive	$(\lvert\varepsilon\rvert - \sigma)\mathbb{1}_{\lvert\varepsilon\rvert>\sigma}$	$\frac{1}{2+2\sigma} e^{-(\lvert y-f(x)\rvert-\varepsilon)\mathbb{1}_{\lvert y-f(x)\rvert>\sigma}}$,
Huber's robust loss	$\begin{cases} \frac{1}{2\sigma}\lvert\varepsilon\rvert^2 & \text{if } \lvert\varepsilon\rvert \leq \sigma \\ \lvert\varepsilon\rvert - \frac{\sigma}{2} & \text{if } \lvert\varepsilon\rvert > \sigma \end{cases}$	$\begin{cases} e^{-\frac{(y-f(x))^2}{2\sigma}} & \text{if } \lvert y-f(x)\rvert \leq \sigma \\ e^{-(\lvert y-f(x)\rvert - \frac{\sigma}{2})} & \text{if } \lvert y-f(x)\rvert > \sigma \end{cases}$

Bayesian interpretation of loss functions IV



Summary

- The optimal classifier for classification is the Bayes classifier. Extensions to cost-sensitive learning and the multi-class setting possible.
- Two schemes for solving multi-class problems: one-versus-all and one-versus-one.
- Discussion of the optimal function in regression (loss-dependent).
- Bayesian interpretation of loss functions $\Rightarrow$ Maximizing the likelihood is equivalent to minimizing the corresponding loss.

Empirical risk minimization

Problem: In order to compute the Bayes optimal learning rule we need to know the joint measure P on $\mathcal{X} \times \mathcal{Y}$,

but !

We do not know P but we have only the **training data** $(X_i, Y_i)_{i=1}^n$.

Idea: approximate the risk functional using the training data.

Empirical risk minimization II

Assumption: Training data $(X_i, Y_i)_{i=1}^n$ is an **i.i.d.** sample of the probability measure P on $\mathcal{X} \times \mathcal{Y}$.
i.i.d. = independently and identically distributed

- $(X_i, Y_i)_{i=1}^n$ are random variables,
- **independent**: joint density factorizes

$$p\big((x_1, y_1); (x_2, y_2); \ldots ; (x_n, y_n)\big) = \prod_{i=1}^{n} p_i(x_i, y_i).$$

- **identically distributed**:

$$p_i(x, y) = p_j(x, y), \qquad \forall i, j \in \{1, \ldots, n\}.$$

and $p(x, y)$ is the density of the data-generating measure P on $\mathcal{X} \times \mathcal{Y}$.

General Principle in Statistics

Statistics: Given an i.i.d. sample $(X_i)_{i=1}^n$, use the empirical measure

$$\mathrm{P}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x=X_i}$$

to approximate quantities of the data generating measure.

- **empirical mean**: $\mathbb{E}_{P_n}[X] = \frac{1}{n}\sum_{i=1}^n x\,\mathbb{1}_{x=X_i} = \frac{1}{n}\sum_{i=1}^n X_i$,
- **empirical variance**: $\mathrm{Var}[X] = \frac{1}{n}\sum_{i=1}^n \Big(X_i - \frac{1}{n}\sum_{j=1}^n X_j\Big)^2$,
- **empirical covariance**:
  $\mathrm{Cov}(X, Y) = \frac{1}{n}\sum_{i=1}^n X_i Y_i - \frac{1}{n}\sum_{i=1}^n X_i \,\frac{1}{n}\sum_{i=1}^n Y_i$,

$\boxed{\mathrm{P}_n \textbf{ approximates } \mathrm{P}}$

Empirical risk minimization III

Definition

Let $(X_i, Y_i)_{i=1}^n$ be an i.i.d. sample of P on $\mathcal{X} \times \mathcal{Y}$, which we call the **training sample**. The **empirical loss** is defined as

$$\mathbb{E}_{\mathrm{P}_n}[L(Y, f(X))] = \frac{1}{n}\sum_{i=1}^{n} L(Y_i, f(X_i)).$$

Given a class of functions $\mathcal{F}$, **empirical risk minimization** is defined as

$$f_n = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{\mathrm{P}_n}[L(Y, f(X))] = \arg\min_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} L(Y_i, f(X_i)),$$

where f_n is then learning rule based on the training sample.

Problems of Empirical risk minimization

- if the function class is too large it is likely that we overfit the training data,
- the mapping "data" to "learning rule" can be seen as an *inverse problem*.
  Definition of a **well-posed problem**:
  - a solution exists,
  - the solution is unique,
  - the solution depends continuously on the data.

  A problem which does not have one of these properties is called **ill-posed**. In particular the last two properties are most of the time not fulfilled in empirical risk minimization. In order to make problems well-posed one uses **regularization**.

Empirical risk minimization: Classification

Natural loss: 0-1-loss $L(y, f(x)) = \mathbb{1}_{y \neq f(x)}$.

Empirical risk minimization:
minimize the number of errors on the training set:

$$\frac{1}{n}\sum_{i=1}^{n} L(Y_i, f(X_i)) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{Y_i \neq f(X_i)}.$$

Problem:

- for several classes of functions, empirical risk minimization leads to NP-hard problems
  $\Longrightarrow$ use of convex margin-based loss functions.

Empirical risk minimization: Regression

Standard loss: squared loss $L(y, f(x)) = (y - f(x))^2$.

$$f_n = \mathop{\arg\min}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2$$

- $\mathcal{F} = \{f(x) = \langle w, x \rangle \mid w \in \mathbb{R}^d\}$, **linear least squares regression**.

Empirical risk minimization:

- function class $\mathcal{F}$ too large $\rightarrow$ overfitting,
- function class $\mathcal{F}$ too small $\rightarrow$ underfitting,

Idea: Use regularization together with a rather large function class $\mathcal{F}$.

Regularized empirical risk minimization

Definition

Let

- $(X_i, Y_i)_{i=1}^n$ be the training sample,
- $\mathcal{F}$ a fixed function class,
- $L(y, f(x))$ the loss function,
- $\Omega : \mathcal{F} \to \mathbb{R}_+$ the **regularization functional**.

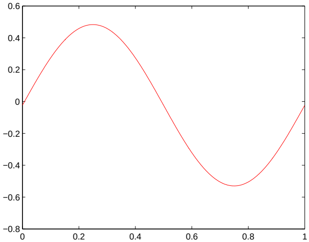
Then **regularized empirical risk minimization** is defined as

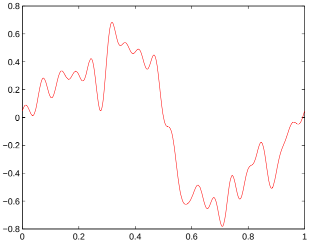
$$f_{n,\lambda} = \operatorname*{arg\,min}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(X_i)) + \lambda\,\Omega(f),$$

where $\lambda \in \mathbb{R}_+$ is called the **regularization parameter**.

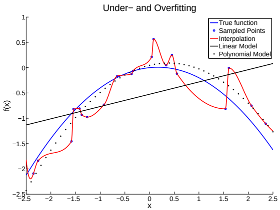
$\Longrightarrow$ This form of regularization is called **Tikhonov regularization**.

Trade-off between **fit of the data** and **complexity of the learning rule**.





Left: Relatively simple function, very smooth,
Right: Complex function, less smooth.



Regularized empirical risk minimization II

Equivalent formulation (Ivanov regularization):

Proposition

If the loss $L(y, f(x))$ and the regularization function $\Omega(f)$ are convex in f and the set $\{f \,|\, \Omega(f) < r\}$ is non-empty for every $r > 0$ and $\mathcal{F}$ is a convex set, then regularized empirical risk minimization is equivalent to the following problem:

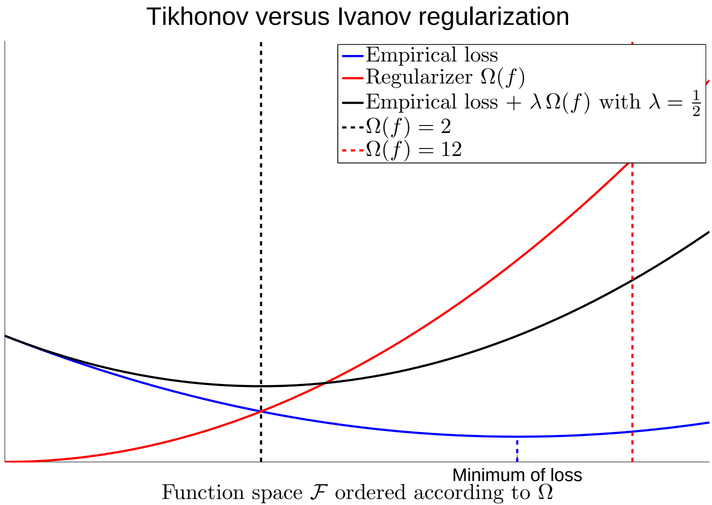
$$f_{n,r} = \underset{f \in \mathcal{F}, \quad \Omega(f) \leq r}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(X_i)),$$

in the sense that there exists for each r a corresponding λ such that $f_{n,r} = f_{n,\lambda}$.

Proof:

- use of duality in convex optimization,
- in the lecture notes: proof for finite-dimensional function classes.

Tikhonov versus Ivanov regularization



Occam's razor

General Principle: prefer less complex function, as measured by Ω, if they have the same loss.

"Occam's razor":

Pluralitas non est ponenda sine necessitas.' (Plurality should not be posited without necessity.),

or similarly:

"Having two competing theories which make exactly the same predictions, the one that is simpler is the better."

Regularized empirical risk minimization IV

Regularization parameter λ: controls trade-off between fit and complexity.

Limits: $\lambda \to 0$ and $\lambda \to \infty$.

$$\lambda \to 0 \qquad \underset{\underset{f \in \mathcal{F}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(X_i))}{\arg\min} \Omega(f),$$

$$\lambda \to \infty \qquad \underset{\{f \in \mathcal{F} \,|\, \Omega(f)=0\}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(X_i))$$

Example: $f : \mathbb{R}^d \to \mathbb{R}$,

$$\Omega(f) = \int_{\mathbb{R}^d} \sum_{i=1}^{d} \left(\frac{\partial f}{\partial x^i}\right)^2 dx = \int_{\mathbb{R}^d} \|\nabla f\|^2 \, dx$$

$$\Omega(f) = 0 \quad \Longleftrightarrow \quad \exists\, c \in \mathbb{R}, \text{ such that } f(x) = c, \; \forall x \in \mathbb{R}^d.$$

Regularized empirical risk minimization IV

Related principle: **Structural risk minimization** proposed by Vapnik.

- empirical risk minimization over nested function classes $\mathcal{F}_n$, such that $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$,
- as the size of the sample n increases one also allows more complex functions.

Example: start with the linear functions and then add polynomials of increasing order as n increases.

Relation I:

Empirical risk minimization
corresponds to
maximum likelihood estimation.

Relation II:

Regularized empirical risk minimization
corresponds to
maximum a posteriori estimation.

Maximum Likelihood Estimation

Problem: Given samples $x_1, \ldots, x_n$ identify the probability measure $p(x)$ which generated this sample.

General problem to difficult $\Longrightarrow$ parametric model of $p(x)$.

General Idea:

- parametric model of the data generating probability measure:

$$p(x \mid \theta) \qquad \text{(the likelihood)}.$$

- i.i.d. data $x_1, \ldots, x_n \implies p(x_1, \ldots, x_n \,|\theta) = \prod_{i=1}^n p(x_i \mid \theta)$.
- find parameter $\theta \in \Theta$ by maximizing the likelihood (resp. the log-likelihood)

$$\begin{aligned} \arg\max_{\theta \in \Theta} \prod_{i=1}^n p(x_i \mid \theta) &= \arg\max_{\theta \in \Theta} \log \Big(\prod_{i=1}^n p(x_i \mid \theta) \Big) \\ &= \arg\max_{\theta \in \Theta} \sum_{i=1}^n \log \big(p(x_i \mid \theta) \big) \end{aligned}$$

Maximum Likelihood Estimation - Example

Gaussian model:

$p(x|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, the variance σ^2 is assumed to be known.

Maximum likelihood estimation of μ:

$$\begin{aligned} \underset{\mu \in \mathbb{R}}{\arg\max} \sum_{i=1}^{n} \log\big(p(x_i \,|\, \mu)\big) &= \underset{\mu \in \mathbb{R}}{\arg\max} \sum_{i=1}^{n} \Big(-\frac{\log\big(2\pi\sigma^2\big)}{2} - \frac{(x_i - \mu)^2}{2\sigma^2} \Big) \\ &= \underset{\mu \in \mathbb{R}}{\arg\min} \sum_{i=1}^{n} (x_i - \mu)^2 \end{aligned}$$

The objective is convex in $\mu \Longrightarrow$ Every local minimum is a global minimum.

The mean parameter μ^* maximizing the likelihood is:

$$\mu^* = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Empirical Risk Minimization and Maximum Likelihood

Maximum Likelihood:

- model for the conditional distribution = the **likelihood**: $p(y|x, f)$,
  f denotes the parameter of the model,
  $\mathcal{F}$ is the set of parameters.
- i.i.d. sample of the data $D = (X_i, Y_i)_{i=1}^n$.

Definition

The **maximum likelihood** solution f_{ML} is then defined as

$$f_{ML} = \arg\max_{f \in \mathcal{F}} \mathrm{P}(D|f) = \arg\max_{f \in \mathcal{F}} \prod_{i=1}^{n} \mathrm{P}(Y_i|X_i, f),$$

$$
\begin{array}{lcl}
\text{Model for } p(y|x, f(x)) & \longrightarrow & L(y, f(x)) = -\log p(y|x, f(x)), \\
\text{Loss function } L(y, f(x)) & \xrightarrow[\text{correspondence}]{\text{generally no}} & p(y|x, f(x)) = e^{-L(y,f(x))}.
\end{array}
$$

Proposition

Given an i.i.d. training sample $(X_i, Y_i)_{i=1}^n$, a class of functions $\mathcal{F}$ and a likelihood $p(y|x, f)$, then the **maximum likelihood solution** *f_{ML} agrees with the solution of* **empirical risk minimization** *f_n for the loss function $L(y, f(x)) = -\log p(y|x, f)$.*

- output space $\mathcal{Y}$ is discrete: likelihood is probability $\mathrm{P}(y|x, f)$,
- output space $\mathcal{Y}$ is continuous: likelihood is density $p(y|x, f)$.

Proof: By assumption we know $L(y, f(x)) = -\log \mathrm{P}(y|x, f)$, then

$$\begin{aligned}
f_{ML} &= \operatorname*{arg\,max}_{f \in \mathcal{F}} \mathrm{P}(D|f) = \operatorname*{arg\,max}_{f \in \mathcal{F}} \prod_{i=1}^{n} \mathrm{P}(Y_i|X_i, f) \\
&= \operatorname*{arg\,max}_{f \in \mathcal{F}} \quad \log \Big[\prod_{i=1}^{n} \mathrm{P}(Y_i|X_i, f) \Big] \\
&= \operatorname*{arg\,max}_{f \in \mathcal{F}} \quad \sum_{i=1}^{n} \log \mathrm{P}(Y_i|X_i, f) \\
&= \operatorname*{arg\,min}_{f \in \mathcal{F}} - \sum_{i=1}^{n} \log \mathrm{P}(Y_i|X_i, f) \\
&= \operatorname*{arg\,min}_{f \in \mathcal{F}} \sum_{i=1}^{n} L(Y_i, f(X_i)) = f_n,
\end{aligned}$$

Maximum A Posteriori Estimation

Idea: integrate **prior belief** on the model parameter θ

Realization: θ is random, we have a **prior distribution** $p(\theta)$.

MAP Estimation:

- prior distribution $p(\theta)$
- using Bayes rule

$$p(\theta \,|\, x_1, \ldots, x_n) = \frac{p(x_1, \ldots, x_n \,|\, \theta) p(\theta)}{p(x_1, \ldots, x_n)} = \frac{p(x_1, \ldots, x_n \,|\, \theta) p(\theta)}{\int_\Theta p(x_1, \ldots, x_n \,|\, \theta) p(\theta) d\theta}.$$

  The denominator is called the **partition function**.
- find parameter θ by maximizing the a posteriori distribution $p(\theta \,|\, x_1, \ldots, x_n)$

$$\arg\max_{\theta \in \Theta} \prod_{i=1}^{n} p(\theta \,|\, x_1, \ldots, x_n) = \arg\max_{\theta \in \Theta} \log \Big(p(x_1, \ldots, x_n \,|\, \theta) p(\theta) \Big)$$

$$= \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log \big(p(x_i \,|\, \theta) \big) + \log \big(p(\theta) \big).$$

Maximum A Posteriori Estimation - Example

Gaussian model:

$p(x|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, the variance σ^2 is assumed to be known.

$p(\mu) = \frac{1}{\sqrt{2\pi\sigma_\mu^2}} e^{-\frac{(\mu-\mu_0)^2}{2\sigma_\mu^2}}$, the variance σ_μ^2 is assumed to be known.

MAP estimation of μ:

$$\begin{aligned}\arg\max_{\mu\in\mathbb{R}} p(\mu \,|\, x_1, \ldots, x_n) &= \arg\max_{\mu\in\mathbb{R}} \sum_{i=1}^n \log\big(p(x_i \,|\, \mu)\big) + \log\big(p(\mu)\big) \\ &= \arg\min_{\mu\in\mathbb{R}} \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 + \frac{1}{2\sigma_\mu^2}(\mu - \mu_0)^2\end{aligned}$$

The objective is convex in μ The MAP estimate of the mean parameter μ^* is:

$$\mu^* = \frac{1}{1 + \frac{\sigma^2}{n\sigma_\mu^2}} \frac{1}{n} \sum_{i=1}^n x_i \; + \; \frac{1}{1 + \frac{n\sigma_\mu^2}{\sigma^2}} \mu_0.$$

Ingredients:

- likelihood function $\mathrm{P}(y|x, f)$,
- set of parameters $\mathcal{F}$,
- **prior** $\mathrm{P}(f)$ over the function/parameter space $\mathcal{F}$.
  probability measure over the class of functions
  $\Longrightarrow$ which functions are more "likely" than others,
  $\Longrightarrow$ encodes **a priori** knowledge.

Maximum a posteriori estimation

Ingredients:

- likelihood function $\mathrm{P}(y|x, f)$,
- set of parameters $\mathcal{F}$,
- **prior** $\mathrm{P}(f)$ over the function/parameter space $\mathcal{F}$.
  probability measure over the class of functions
  $\Longrightarrow$ which functions are more "likely" than others,
  $\Longrightarrow$ encodes **a priori** knowledge.

Bayes theorem: Transform the prior and the likelihood into the **posterior probability** $\mathrm{P}(f|D)$,

$$\mathrm{P}(f|D) = \frac{\mathrm{P}(D|f)\mathrm{P}(f)}{\mathrm{P}(D)},$$

where $\mathrm{P}(D) = \int_{\mathcal{F}} \mathrm{P}(D|f)\mathrm{P}(f)df$.

MAP estimation II

Definition

The **maximum a posteriori** estimator for f is defined as

$$f_{MAP} = \arg\max_{f \in \mathcal{F}} \mathrm{P}(f|D) = \arg\max_{f \in \mathcal{F}} \prod_{i=1}^{n} \mathrm{P}(Y_i|X_i, f)\mathrm{P}(f),$$

where we have discarded $\mathrm{P}(D)$ since it is a constant.

Given a prior over functions $\mathrm{P}(f)$ we define the following regularization functional $\Omega(f)$,

$$\Omega(f) = -\log \mathrm{P}(f) \qquad \Longrightarrow \qquad \mathrm{P}(f) = e^{-\Omega(f)},$$

MAP estimation III

Proposition

The MAP estimator f_{MAP} agrees with the minimizer of $f_{\lambda=\frac{1}{n},n}$ of the regularized empirical risk minimization if

$$\begin{aligned} &\textbf{Loss function:} && L(y, f(x)) &&= -\log \mathrm{P}(y|x, f), \\ &\textbf{regularization functional:} && \Omega(f) &&= -\log \mathrm{P}(f). \end{aligned}$$

MAP estimation IV

Proof.

By assumption we know $L(y, f(x)) = -\log \mathrm{P}(y|x, f)$ and $\Omega(f) = -\log \mathrm{P}(f)$, then

$$\begin{aligned} f_{MAP} &= \arg\max_{f \in \mathcal{F}} \mathrm{P}(f|D) = \arg\max_{f \in \mathcal{F}} \prod_{i=1}^{n} \mathrm{P}(Y_i|X_i, f)\mathrm{P}(f) \\ &= \arg\max_{f \in \mathcal{F}} \sum_{i=1}^{n} \log \mathrm{P}(Y_i|X_i, f) + \log \mathrm{P}(f) \\ &= \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} L(Y_i, f(X_i)) + \Omega(f) = f_{n,\lambda=\frac{1}{n}}. \end{aligned}$$

where we have used that the logarithm is a strictly increasing function. □

Full Bayesian treatment

Posterior predictive distribution:

$$p(x|D) = \int_\Theta p(x, \theta \,|\, D)\, d\theta = \int_\Theta p(x \,|\, \theta, D)\, p(\theta \,|\, D)\, d\theta$$

- $p(x|D)$ is **not** the true data-generating distribution !
- if the posterior $p(\theta \,|\, D)$ is very peaked, this is roughly the same as $p(x \,|\, \theta_{\mathrm{MAP}})$

Learning setting:

$$p(y \,|\, x, D) = \int_\Theta p(y \,|\, x, \theta)\, p(\theta \,|\, D)\, d\theta$$

Bayesians consider the full distribution $p(\theta \,|\, D)$ against the point estimate in the MAP estimation.