Lecture 3: Empirical Risk Minimization

Isabel Valera

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

25.04.2023

Outline

1. Bibliography
2. Statistical Learning
3. Empirical Risk Minimization
4. Regularized ERM
5. Bayesian Interpretation
6. Summary

Main references

- "Learning with Kernels." Schölkopf & Smola, MPIT Press 1998 – Chapters 3 & 4.
- "Principles of Risk Minimization for Learning Theory." Vapnik, NeurIPS 1991.
- "The nature of Statistical Learning Theory (2nd edition)." Vapnik, Springer 1999 – Chapters 1 & 4 (for a learning theory perspective).

Outline

1. Bibliography
2. Statistical Learning
3. Empirical Risk Minimization
4. Regularized ERM
5. Bayesian Interpretation
6. Summary

Motivation

- Bayesian decision theory allows us to make optimal decisions under uncertainty.
- So far we have only considered the uncertainty that comes from the stochastic nature of the problem, i.e., we only accounted for the fact that the outcomes $y \in \mathcal{Y}$ are **non-deterministic** given the features x.
- In other words, we have assumed that the **probability measure** P on $\mathcal{X} \times \mathcal{Y}$ is known.
- However, in practice we do not have access to the probability measure, but instead we only observe training data generated from such a probability measure.
- **Can we still make "optimal" decisions? In other words, can we still learn a stable (under small changes in the training data) function that maps features into outputs, i.e.,** $y = f(x)$**?**

Statistical Learning Problems

Goal: Reason on the outcome value y for given features x.
Assumption: Only an **independently and identically distributed (i.i.d.)** sample $(x_i, y_i)_{i=1}^n$ (*training data*) of the probability measure P on $\mathcal{X} \times \mathcal{Y}$ is available. Thus,

- **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**:

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

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

Discriminative versus generative learning

We usually distinguish between two main types of approaches to solve supervised statistical learning problems:

- **Generative Learning:** Estimate the joint distribution $p(x, y)$ (inference) and then use Bayes rule to compute the conditional probability $p(y|x)$.
- **Discriminative Learning:** Directly approximate the conditional distribution $p(y|x)$.

Usually estimating $p(x, y)$ is a harder poblem than approximating the conditional probability $p(y|x)$, thus many methods adopt a discriminative approach. The main advantage of generative learning is that it allows you to sample new (synthetic data) as well as handle uncertainties on the obverved features (e.g., to handle missing values or outliers).

General Principle in Statistics

Statistics: Let X be a random variable with **unknown** probability measure P and $(x_i)_{i=1}^n$ an i.i.d. sample from P, we use the empirical measure

$$P_n(X_n) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_n = x_i}$$

to approximate quantities of the data generating measure.[1] For example,

- **Empirical mean**: $\mathbb{E}[X] \approx \mathbb{E}_{P_n}[X_n] = \frac{1}{n} \sum_{i=1}^n x_i \, \mathbb{1}_{X_n = x_i} = \frac{1}{n} \sum_{i=1}^n x_i.$
- **Empirical variance**: $\mathrm{Var}[X] \approx \mathrm{Var}[X_n] = \frac{1}{n} \sum_{i=1}^n \left(x_i - \frac{1}{n} \sum_{j=1}^n x_j \right)^2.$
- **Empirical covariance**: Assume now that $X = (X_1, X_2)$ then: $\mathrm{Cov}(X_1, X_2) \approx \frac{1}{n} \sum_{i=1}^n x_{1i} x_{2i} - \frac{1}{n} \sum_{i=1}^n x_{1i} \, \frac{1}{n} \sum_{i=1}^n x_{2i}.$

$$\boxed{P_n \textbf{ approximates } P}$$

[1] Essentially we use the law of large numbers, e.g., $lim_{n \to \infty} \mathbb{E}_{P_n}[X_n] = \mathbb{E}[X]$

Outline

1. Bibliography
2. Statistical Learning
3. Empirical Risk Minimization
4. Regularized ERM
5. Bayesian Interpretation
6. Summary

Empirical risk minimization

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

$$\mathbb{E}_{P_n}[L(Y, f(X))] = \frac{1}{n} \sum_{i=1}^n L(y_i, f(Xx_i)).$$

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

$$f_n = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{P_n}[L(Y, f(X))] = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n L(y_i, f(x_i)),$$

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

Binary classification

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

Empirical risk minimization: we aim at finding the classifier $\hat{y}(X)$ that minimizes the number of errors on the training set, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n} L(y_i,\hat{y}(x_i)) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{y_i\neq\hat{y}(x_i)}.$$

Problem: For several classes of functions, minimizing the above empirical risk leads to NP-hard (nondeterministic polynomial time) problems. *Solution:* Minimize a **convex margin-based loss function** $L:\Re\times\Re\to\Re_+$ (refer to Lecture 3) as

$$f_n = \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i))$$

and classify a feature (vector) x using the classifier $\hat{y}(x) = \operatorname{sign} f_n(x)$.

Regression

Standard loss used in practice: squared loss $L(Y, f(X)) = (Y - f(X))^2$.

$$f_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

- Example: $X \in \mathbb{R}^d$ and $\mathcal{F} = \{f(X = \mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle \mid \mathbf{w} \in \mathbb{R}^d\}$, **linear least squares regression** (Lecture 5).

Outline

1. Bibliography
2. Statistical Learning
3. Empirical Risk Minimization
4. Regularized ERM
5. Bayesian Interpretation
6. Summary

Problems of ERM

- **Problems of ERM:**
  - function class $\mathcal{F}$ too large → overfitting.
  - function class $\mathcal{F}$ too small → underfitting.
- The mapping "data" to "learning/decision rule" can be seen as an *inverse problem* (i.e., we want to determine the function parameters that produce the data). A **well-posed problem** fulfills that:
  - 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**.

Solution: Assume a large function class $\mathcal{F}$ and use regularization.

Illustration of over/under-fitting I

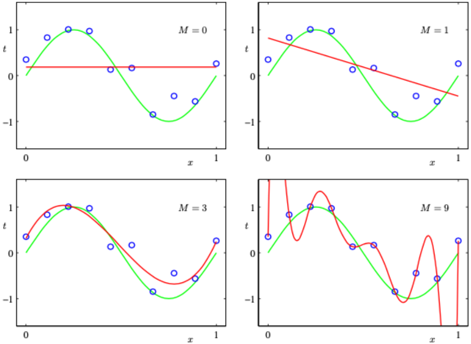

Figure: Image from Bishop. Here M corresponds to the order of the polynomial used to fit the data. The blue circles correspond to the training data, and the gree and red curves correspond respectively to the functions, respectively, used to generate the data and fitted minimizing the (empirical) squared loss over training data.

Illustration of over/under-fitting II

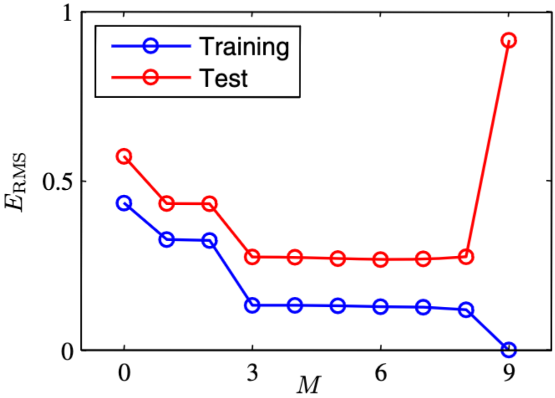

Figure: Image from Bishop. The ‘Training’ and ‘Test’ curves show the root mean square error evaluated on two different i.i.d. samples from the probability measure P, being the training sample the one used to learn the regression function minimizing the empirical squared loss.

Regularized empirical risk minimization I

Definition (Tikhonov regularization)

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

$$f_{n,\lambda} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda \Omega(f),$$

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

Observation: the regularization parameter λ trades-off between **fit of the data** and **complexity of the learning rule**.

Regularized empirical risk minimization II

Proposition (Ivanov regularization)

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:*

$$f_{n,r} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) \tag{1}$$

$$\textit{subject to } \Omega(f) \leq r \tag{2}$$

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 (refer to Block III).

Regularized empirical risk minimization III

Regularization parameter λ: controls the trade-off between data fitting and model complexity, i.e., controls over/under-fitting.

Limits:

$\lambda \to 0$: selects the least complex function among the "optimal" ones (note that $\lambda \neq 0$), i.e.,

$$\arg\min_{f \in \mathcal{F}^*} \Omega(f), \text{ with } \mathcal{F}^* = \{f \in \mathcal{F} | \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i))\}.$$

$\lambda \to \infty$: considers only functions of zero complexity, $\Omega(f) = 0$, and selects the function which has the smallest loss, i.e.,

$$\arg\min_{f \in \mathcal{F}^*} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) \text{ with } \mathcal{F}^* = \{f \in \mathcal{F} | \Omega(f) = 0\}.$$

Example: $f : \mathbb{R}^d \to \mathbb{R}$, and $\Omega(f) = \sup_{\mathbf{x} \in \mathbb{R}^d} \max_{i=1,\dots,d} \left| \frac{\partial f}{\partial x^i}(\mathbf{x}) \right|$

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

Illustration of RERM I

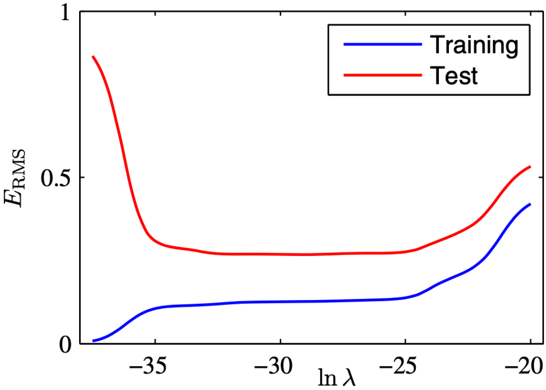

Figure: Image 1.8 from Bishop. Here, we consider $f(x) = \sum_{m=0}^{M} w_m x^m$ with $M = 9$ and solve the RERM problem asumming the squared loss and regularizer $\Omega(f) = \|\mathbf{w}\|^2 = \sum_{m=0}^{M} w_m^2$. Here, λ controls the effective complexity (as lambda grows, the coefficients w_m take values closer to zero) and hence allow us to control over/under-fitting.

Structural risk minimization

Structural risk minimization proposed by Vapnik considers:

- empirical risk minimization over nested function classes $\mathcal{F}_n$, such that $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$, and
- 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 the number of observations n increases.

Occam's razor

"Occam's razor" (William of Ockham, 1287-1347):

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

Or similarly:

"The simplest explanation is usually the right one."

In ERM: Between two functions with same loss (risk), select the least complex one measured by Ω.
In ML: Between two ML models with similar performance, select the simpler one (Block II - performance metrics and model selection).

Outline

1. Bibliography
2. Statistical Learning
3. Empirical Risk Minimization
4. Regularized ERM
5. Bayesian Interpretation
6. Summary

Bayesian Interpretation of (R)ERM

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 (MLE)

Problem: Given i.i.d. samples $x_1, \ldots, x_n$ from an unknown probability density $p(x)$, we aim at estimating $p(x)$.

MLE Solution:

1. Assume a parametric model of the data generating probability density $p(x \,|\, \theta)$ (i.e., a likelihood model), such that we can evaluate the likelihood (which is a function of the parameters) as:

$$p(x_1, \ldots, x_n \,|\theta) = \prod_{i=1}^{n} p(x_i \,|\, \theta).$$

2. Find parameter $\theta \in \Theta$ by maximizing the likelihood (resp. the log-likelihood), i.e.,

$$\begin{aligned}\arg\max_{\theta \in \Theta} \prod_{i=1}^{n} p(x_i \,|\, \theta) &= \arg\max_{\theta \in \Theta} \, \log\left(\prod_{i=1}^{n} p(x_i \,|\, \theta)\right) \\ &= \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log\left(p(x_i \,|\, \theta)\right)\end{aligned}$$

Example - Gaussian model

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

Maximum likelihood estimation of μ:

$$\begin{aligned} \arg\max_{\mu\in\mathbb{R}} \sum_{i=1}^{n} \log\left(p(x_i \mid \mu)\right) &= \arg\max_{\mu\in\mathbb{R}} \sum_{i=1}^{n} \left(- \frac{\log\left(2\pi\sigma^2\right)}{2} - \frac{(x_i-\mu)^2}{2\sigma^2} \right) \\ &= \arg\min_{\mu\in\mathbb{R}} \sum_{i=1}^{n} (x_i-\mu)^2 \end{aligned}$$

The mean parameter μ^* maximizing the likelihood is (*Exercise*-Proof!):

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

Observation: the log-likelihood is convex with respect to μ, thus there exist a unique global minimum (Block III).

ERM vs MLE I

Given a training dataset $D = (x_i, y_i)_{i=1}^n$. Next, we aim to approximate the conditional distribution (**likelihood**) $p(Y|X, f)$, where f denotes the (parameters of the) model, and $\mathcal{F}$ the family of considered functions (characterized by its set of parameters).

Definition

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

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

where $D = (x_i, y_i)_{i=1}^n$ denotes the training data.

Note: $P(D|f) = \prod_{i=1}^n P(Y = y_i, X = x_i|f) = \prod_{i=1}^n P(Y = y_i|x_i, f)P(X = x_i)$

ERM vs MLE II

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* $L(Y,f(X)) = -\log p(Y|X,f)$.

Observations:

- For a given likelihood function, $p(Y|X,f(X))$ we can define an associated loss function as $L(Y,f(X)) = -\log p(Y|X,f(X))$.
- An arbitrary loss function $L(Y,f(X))$ does in general not correspond to a likelihood of the form $p(Y|X,f(X)) \simeq e^{-L(Y,f(X))}$.

Note:

- output space $\mathcal{Y}$ is discrete: likelihood is a probability (mass function) $P(Y|X,f)$,
- output space $\mathcal{Y}$ is continuous: likelihood is (probability) density (function) $p(Y|X,f)$.

ERM vs MLE III

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

$$
\begin{aligned}
f_{ML} &= \arg\max_{f\in\mathcal{F}} P(D|f) = \arg\max_{f\in\mathcal{F}} \prod_{i=1}^{n} P(y_i|x_i, f) \\
&= \arg\max_{f\in\mathcal{F}} \quad \log\left[\prod_{i=1}^{n} P(y_i|x_i, f)\right] \\
&= \arg\max_{f\in\mathcal{F}} \quad \sum_{i=1}^{n} \log P(y_i|x_i, f) \\
&= \arg\min_{f\in\mathcal{F}} -\sum_{i=1}^{n} \log P(y_i|x_i, f) \\
&= \arg\min_{f\in\mathcal{F}} \sum_{i=1}^{n} L(y_i, f(x_i)) = f_n,
\end{aligned}
$$

Maximum A Posteriori (MAP) Estimation

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

MAP Estimation:

1. Treat the model parameters θ as a random variable, and assume a prior distribution $p(\theta)$ which accounts for prior belief.
2. Use Bayes rule to obtain the posterior of the parameters as:

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

   The denominator is called the partition function (or evidence).

3. Find parameters θ by maximizing the posterior distribution, i.e.,

$$\begin{aligned} \arg\max_{\theta \in \Theta} p(\theta \,|\, x_1, \ldots, x_n) &= \arg\max_{\theta \in \Theta} \; \log \Big(p(x_1, \ldots, x_n \,|\, \theta) p(\theta) \Big) \\ &= \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log \big(p(x_i \,|\, \theta) \big) + \log \big(p(\theta) \big). \end{aligned}$$

Example - Gaussian model

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

Gaussian prior (on the mean parameter): $p(\mu) = \frac{1}{\sqrt{2\pi\sigma_\mu^2}} e^{-\frac{(\mu-\mu_0)^2}{2\sigma_\mu^2}}$ (with known parameters).

MAP estimation of μ:

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

The MAP estimate μ_{MAP} of the mean parameter is (note that the objective is convex in μ):

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

Regularized ERM and MAP estimation I

Assumption: Data samples $D = (x_i, y_i)_{i=1}^n$ are conditionally independent given f and the inputs are independent of f.

Definition

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

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

where we have discarded $P(D)$ and $P(X)$ since they are constant w.r.t. changes to f.

Note: Bayes theorem states that $P(f|D) = \frac{P(D|f)P(f)}{P(D)}$, where $P(D) = \int_{\mathcal{F}} P(D|f)P(f)df$.

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

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

Regularized ERM and MAP estimation II

Proposition

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

Loss function: $L(Y, f(X)) = -\log P(Y|X, f),$

Regularization functional: $\Omega(f) = -\log P(f).$

Regularized ERM and MAP estimation III

Proof.

By assumption we know $L(Y, f(X)) = -\log P(Y|X, f)$ and $\Omega(f) = -\log P(f)$, then

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

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

Full Bayesian treatment

So far, we have just computed a point estimate of the ML model f, e.g., the ML or the MAP estimate of f. However, we could instead adopt a fully **Bayesian approach**, which treats the model (parameters) f as a random variable, i.e.:

$$p(Y \mid X, D) = \int_{f \in \mathcal{F}} p(Y \mid X, f)\, p(f \mid D)\, df$$

where $p(f \mid D)$ denotes the posterior (distribution) of f (given the training data), and $p(Y \mid X, f)$ corresponds to the predictive posterior, i.e., the probability of Y given X and the training data D, after integrating out the model (parameters) f.

Observations:

- $p(Y \mid X, D)$ is **not** the true label-generating distribution!
- if the posterior $p(f \mid D)$ is very peaked, this is roughly the same as $p(Y \mid X, f_{\mathrm{MAP}})$.

Outline

1. Bibliography
2. Statistical Learning
3. Empirical Risk Minimization
4. Regularized ERM
5. Bayesian Interpretation
6. Summary

Summary

- ERM provides an approach to solve learning problems (and thus make decisions under uncertainty) when the probability measure P on the feature space $\mathcal{X}$ and the outcome (output) space $\mathcal{Y}$ is unknown, but we only observe an i.i.d sample of that measure, i.e., **training data**.
- The key idea of ERM is to find a function that minimizes a given loss evaluated empirically (i.e., using training data). However, if the family of considered functions is too restrictive, we may end up **under-fitting** the data. Otherwise, if the family of functions is very expressive (e.g., a neural network) we may **over-fit** the training data, and thus obtain a function that generalizes poorly for new unseen data (e.g., test data).
- **Regularized ERM** provides a framework to learn assuming a flexible function class, while mitigating overfitting via a regularization term that penalizes model (function) complexity.
- Relation between ERM and regularized ERM and, respectively, ML and MAP estimation.
- Block II is about learning regression and classification functions!