# Statistical Machine Learning (BE4M33SSU)
# Lecture 3: Empirical Risk Minimization II

Czech Technical University in Prague

# Linear classifier with minimal classification error

- $\mathcal{X}$ is a set of observations and $\mathcal{Y} = \{+1, -1\}$ is a set of hidden labels
- $\phi\colon \mathcal{X} \to \mathbb{R}^n$ is fixed feature map embedding observations from $\mathcal{X}$ to $\mathbb{R}^n$
- Task: we search for a linear classification strategy $h\colon \mathcal{X} \to \mathcal{Y}$

$$h(x; \boldsymbol{w}, b) = \operatorname{sign}(\langle \boldsymbol{w}, \boldsymbol{\phi}(x) \rangle + b) = \begin{cases} +1 & \text{if} \quad \langle \boldsymbol{w}, \boldsymbol{\phi}(x) \rangle + b \geq 0 \\ -1 & \text{if} \quad \langle \boldsymbol{w}, \boldsymbol{\phi}(x) \rangle + b < 0 \end{cases}$$

with minimal expected risk

$$R^{0/1}(h) = \mathbb{E}_{(x,y)\sim p}\Big(\ell^{0/1}(y, h(x))\Big) \quad \text{where} \quad \ell^{0/1}(y, y') = [\![ y \neq y' ]\!]$$

- We are given a set of training examples

$$\mathcal{T}^m = \{(x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, m\}$$

drawn from i.i.d. with the distribution $p(x, y)$.

# ERM learning for linear classifiers

- The Empirical Risk Minimization principle leads to solving

$$(\boldsymbol{w}^*, b^*) \in \operatorname*{Argmin}_{(\boldsymbol{w},b)\in(\mathbb{R}^n\times\mathbb{R})} R^{0/1}_{\mathcal{T}^m}(h(\cdot;\boldsymbol{w},b)) \qquad (1)$$

where the empirical risk is

$$R^{0/1}_{\mathcal{T}^m}(h(\cdot;\boldsymbol{w},b)) = \frac{1}{m}\sum_{i=1}^{m} [\![ y^i \neq h(x^i;\boldsymbol{w},b) ]\!]$$

In this lecture we address the following issues:

1. Algorithmic issues: In the general case there is no known algorithm solving the task (1) in time polynomial in $m$.

2. Is the ERM algorithm for hypothesis space containing linear classifiers statistically consistent? ... yes.

# Vapnik-Chervonenkis (VC) dimension

**Definition 1.** *Let* $\mathcal{H} \subseteq \{-1,+1\}^{\mathcal{X}}$ *and* $\{x^1,\dots,x^m\} \in \mathcal{X}^m$ *be a set of* $m$ *input observations. The set* $\{x^1,\dots,x^m\}$ *is said to be shattered by* $\mathcal{H}$ *if for all* $\boldsymbol{y} \in \{+1,-1\}^m$ *there exists* $h \in \mathcal{H}$ *such that* $h(x^i) = y^i$, $i \in \{1,\dots,m\}$.

**Definition 2.** *Let* $\mathcal{H} \subseteq \{-1,+1\}^{\mathcal{X}}$. *The Vapnik-Chervonenkis dimension of* $\mathcal{H}$ *is the cardinality of the largest set of points from* $\mathcal{X}$ *which can be shattered by* $\mathcal{H}$.

**Theorem 1.** *The VC-dimension of the hypothesis space of all linear classifiers operating in* $n$*-dimensional feature space* $\mathcal{H} = \{h(x;\boldsymbol{w},b) = \mathrm{sign}(\langle \boldsymbol{w}, \boldsymbol{\phi}(x)\rangle + b) \mid (\boldsymbol{w},b) \in (\mathbb{R}^n \times \mathbb{R})\}$ *is* $n+1$.

# Consistency of prediction with two classes and $0/1$-loss



**Theorem 2.** *Let $\mathcal{H} \subseteq \{+1,-1\}^{\mathcal{X}}$ be a hypothesis space with VC dimension $d < \infty$ and $\mathcal{T}^m = \{(x^1,y^1),\ldots,(x^m,y^m)\} \in (\mathcal{X}\times\mathcal{Y})^m$ a training set draw from i.i.d. rand vars with distribution $p(x,y)$. Then, for any $\varepsilon > 0$ it holds*

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}} \left| R^{0/1}(h) - R^{0/1}_{\mathcal{T}^m}(h)\right| \geq \varepsilon\right) \leq 4\left(\frac{2\,e\,m}{d}\right)^d e^{-\frac{m\,\varepsilon^2}{8}}\,.$$

**Corollary 1.** *Let $\mathcal{H} \subseteq \{+1,-1\}^{\mathcal{X}}$ be a hypothesis space with VC dimension $d < \infty$. Then ERM is statistically consistent in $\mathcal{H}$ w.r.t $\ell^{0/1}$ loss function.*

**Corollary 2.** *Let $\mathcal{H} \subseteq \{+1,-1\}^{\mathcal{X}}$ be a hypothesis space with VC dimension $d < \infty$. Then, for any $0 < \delta < 1$ the inequality*

$$R^{0/1}(h) \leq R^{0/1}_{\mathcal{T}^m}(h) + \sqrt{\frac{8\big(d\log\frac{2\,e\,m}{d} + \log\frac{4}{\delta}\big)}{m}}$$

*holds for any $h \in \mathcal{H}$ with probability $1-\delta$ at least.*

## Training linear classifier from separable examples

**Definition 3.** *The examples* $\mathcal{T}^m = \{(x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, m\}$ *are linearly separable w.r.t. feature map* $\boldsymbol{\phi}\colon \mathcal{X} \to \mathbb{R}^n$ *if there exists* $(\boldsymbol{w}, b) \in \mathbb{R}^{n+1}$ *such that*

$$y^i(\langle \boldsymbol{w}, \boldsymbol{\phi}(x^i)\rangle + b) > 0\,, \qquad i \in \{1, \ldots, m\} \tag{2}$$

- Implementation of the ERM for linearly separable examples $\mathcal{T}^m$ leads to solving (2) which yileds $h(x; \boldsymbol{w}, b)$ with $R^{0/1}_{\mathcal{T}^m}(h(\cdot; \boldsymbol{w}, b)) = 0$.
  Note that $y^i(\langle \boldsymbol{w}, \boldsymbol{\phi}(x^i)\rangle + b) > 0$ implies

$$h(x^i) = \mathrm{sign}(\langle \boldsymbol{w}, \boldsymbol{\phi}(x^i)\rangle + b) = y^i$$

- The linear programming task (2) can be solved by the Perceptron algorithm.

# Auxiliary prediction problem leading to tractable ERM



- $\mathcal{X}$, $\mathcal{Y} = \{+1, -1\}$ and $\boldsymbol{\phi}\colon \mathcal{X} \to \mathbb{R}^n$ defined as before.
- Auxiliary prediction problem: find a decision function $f\colon \mathcal{X} \to \mathbb{R}$ minimizing the expectation of the hinge loss $\psi\colon \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_+$:

$$R^\psi(f) = \mathbb{E}_{(x,y)\sim p}(\psi(y, f(x))) \quad \text{where} \quad \psi(y,t) = \max\{0, 1 - y\,t\}$$

- Assuming the hypothesis space which contains the linear functions

$$\mathcal{F} = \left\{ f(x) = \langle \boldsymbol{\phi}(x), \boldsymbol{w} \rangle + b \mid (\boldsymbol{w}, b) \in \mathbb{R}^{n+1} \right\}$$

the ERM principle leads to solving

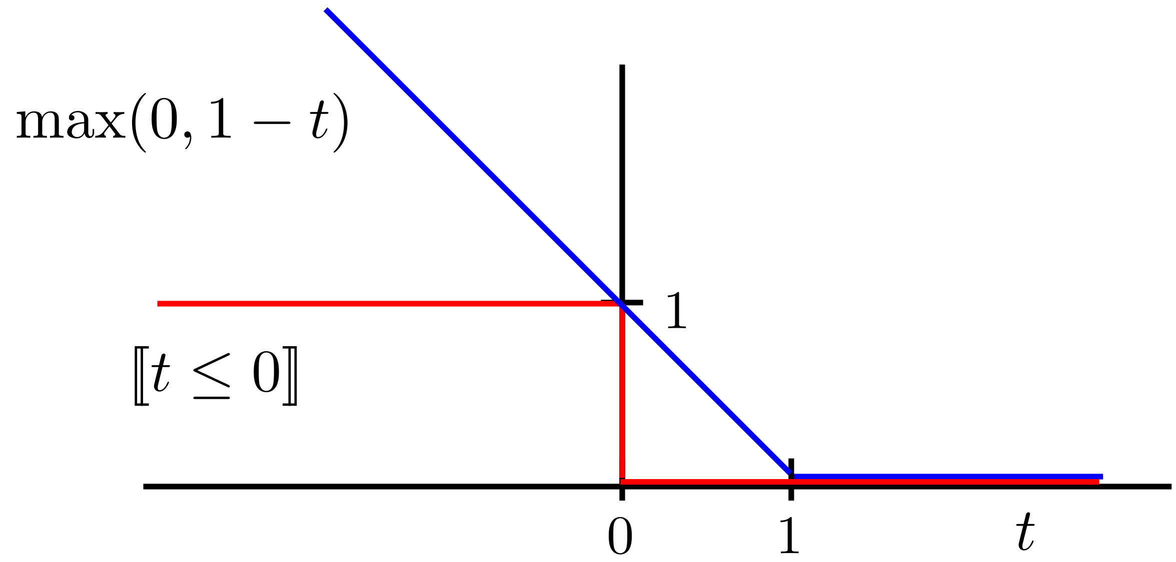
$$f^* = \operatorname*{Argmin}_{f \in \mathcal{F}} R^\psi_{\mathcal{T}^m}(f) \quad \text{where} \quad R^\psi_{\mathcal{T}^m}(f) = \frac{1}{m} \sum_{i=1}^m \psi(y^i, f(x^i))$$

- How is this task related to minimization of the classification error?

## The hinge-loss upper bounds the 0/1-loss

- The hinge-loss is an upper bound of the $0/1$-loss evaluated for the predictor $h(x) = \text{sign}(f(x))$:

$$\underbrace{[\![\text{sign}(f(x)) \neq y]\!]}_{\ell^{0/1}(y,f(x))} = [\![\, y\, f(x) \leq 0]\!] \leq \underbrace{\max\{0, 1 - y\ f(x)\}}_{\psi(y,f(x))}$$




- Therefore 0/1-risk of $h(x) = \text{sign}(f(x))$ is upper-bounded by $\psi$-risk:

$$R^{0/1}(\text{sign}(f)) \leq R^{\psi}(f) \qquad \text{for any} \qquad f\colon \mathcal{X} \to \mathbb{R}$$

## Excess error of $\psi$-risk upper bounds excess risk of 0/1-risk

- The best attainable 0/1-risk is $R^{0/1}_* = \inf_{h \in \mathcal{Y}^\mathcal{X}} R^{0/1}(h)$.
- The best attainable $\psi$-risk is $R^\psi_* = \inf_{f \in \mathbb{R}^\mathcal{X}} R^\psi(f)$
- The best predictor in $\mathcal{F}$ is $f_\mathcal{F} \in \operatorname{Argmin}_{f \in \mathcal{F}} R^\psi(f)$.

**Theorem 3.** *For any $f \colon \mathcal{X} \to \mathbb{R}$ the following inequality holds*

$$\underbrace{R^{0/1}(\operatorname{sign}(f)) - R^{0/1}_*}_{\substack{\textit{excess error} \\ \textit{of original task}}} \leq \underbrace{R^\psi(f) - R^\psi_*}_{\substack{\textit{excess error} \\ \textit{of auxiliary task}}}$$

**Corollary 3.** *Let $A' \colon \cup_{m=1}^\infty (\mathcal{X} \times \mathcal{Y}) \to \mathcal{F}$ be a learning algorithm statistically consistent in $\mathcal{F} \subseteq \mathbb{R}^\mathcal{X}$ w.r.t. $\psi$-risk. In addition, let $R^\psi(f_\mathcal{F}) = R^\psi_*$. Then, the learning algorithm $A(\mathcal{T}^m) = \operatorname{sign}(A'(\mathcal{T}^m))$ is statistically consistent in $\mathcal{H} = \{\operatorname{sign}(f) \mid f \in \mathcal{F}\}$ w.r.t. 0/1-risk.*

## Solving ERM problem of the auxiliary prediction task

- Let us consider a space of linear score functions with parameter vector inside a ball of radius $r$, that is,

$$\mathcal{F}_r = \left\{ f(x) = \langle \boldsymbol{\phi}(x), \boldsymbol{w} \rangle + b \mid (\boldsymbol{w}, b) \in \mathbb{R}^{n+1}, \|\boldsymbol{w}\| \leq r \right\}$$

- The ERM problem for $\psi(y,t) = \max\{0, 1 - y\, t\}$ loss reads

$$f^* = \operatorname*{Argmin}_{f \in \mathcal{F}_r} R^{\psi}_{\mathcal{T}^m}(f) \quad \text{where} \quad R^{\psi}_{\mathcal{T}^m}(f) = \frac{1}{m} \sum_{i=1}^{m} \psi(y^i, f(x^i))$$

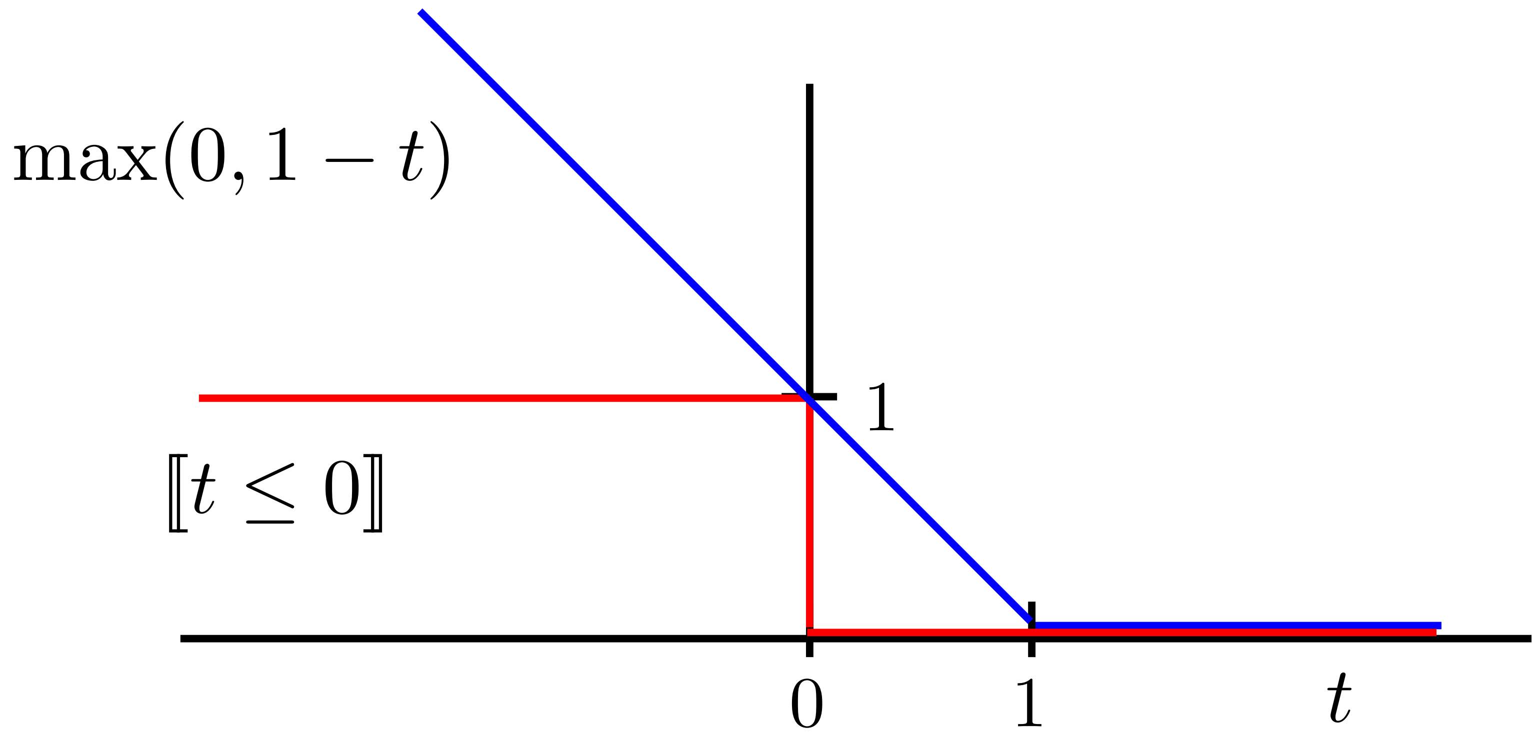
- The ERM problem is a convex unconstrained optimization task

$$(\boldsymbol{w}^*, b^*) = \operatorname*{argmin}_{\|\boldsymbol{w}\| \leq r, b \in \mathbb{R}} \left( \frac{1}{m} \sum_{i=1}^{m} \max\{0, 1 - y^i(\langle \boldsymbol{w}, \boldsymbol{\phi}(x^i) \rangle + b)\} \right)$$

# Summary

Topics covered in the lecture

- Linear classifier
- Vapnik-Chervonenkis dimension
- Consistency and generalization bound for two-class prediction and $0/1$-loss
- ERM problem for linear classifiers
- Auxiliary prediction problem ERM of which is tractable
- Excess error of the auxiliary problem upper bounds the excess error of the original problem



$\max(0, 1 - t)$
1
$[\![t \leq 0]\!]$
0
1
$t$