Statistical Machine Learning (BE4M33SSU)
Lecture 2: Empirical Risk Minimization I

Czech Technical University in Prague

Prediction problem: the definition

- $\mathcal{X}$ is a set of input observations
- $\mathcal{Y}$ is a finite set of hidden labels
- $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is a realization of a random process with p.d.f. $p(x, y)$
- A prediction strategy $h \colon \mathcal{X} \to \mathcal{Y}$
- A loss function $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ penalizes a single prediction
- We want to find a precition strategy with the minimal expected risk

$$R(h) = \int \sum_{y \in \mathcal{Y}} \ell(y, h(x))\, p(x, y)\, \mathrm{d}x = \mathbb{E}_{(x,y) \sim p}\Big(\ell(y, h(x))\Big)$$

Prediction problem: an example

- Assignment:
  - $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{+1, -1\}$, $\ell(y, y') = \begin{cases} 0 & \text{if} \quad y = y' \\ 1 & \text{if} \quad y \neq y' \end{cases}$
  - $p(x, y) = p(y) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu_y)^2}$, $y \in \mathcal{Y}$.
- Since $p(x, y)$ is known the solution of the prediction problem is easy:
  - $h(x) = \operatorname{argmax}_{y \in \mathcal{Y}} p(x, y) = \begin{cases} +1 & \text{if} \quad x \geq \theta \\ -1 & \text{if} \quad x < \theta \end{cases}$
  - $R(h) = \int_{-\infty}^{\theta} p(x, 1) \mathrm{d}x + \int_{\theta}^{\infty} p(x, 2) \mathrm{d}x$
- We will try to solve the problem using only a set of examples

$$\{(x^1, y^1), (x^2, y^2), \ldots\}$$

sampled from i.i.d. rand vars distributed according to unknown $p(x, y)$.

Estimation of the expected risk from examples

- We are given a set of test examples

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

  which are drawn from i.i.d. random variables with distribution $p(x, y)$.

- Given prediction strategy $h \colon \mathcal{X} \to \mathcal{Y}$, we can compute the empirical risk

$$R_{\mathcal{S}^l}(h) = \frac{1}{l} \sum_{i=1}^{l} \ell(y^i, h(x^i))$$

- Is the empirical risk $R_{\mathcal{S}^l}(h)$ a good approximation of the true expected risk $R(h)$?

- Note that the empirical risk $R_{\mathcal{S}^l}(h)$ is a random number.

Law of large numbers

- Arithmetic mean of the results of random trials gets closer to the expected value as more trials are performed.
- Example: The expected value of a single roll of a fair die is

$$\frac{1+2+3+4+5+6}{6} = 3.5$$

According to the LLA, the arithmetic mean of a large number of rolls is likely to be close to 3.5 .

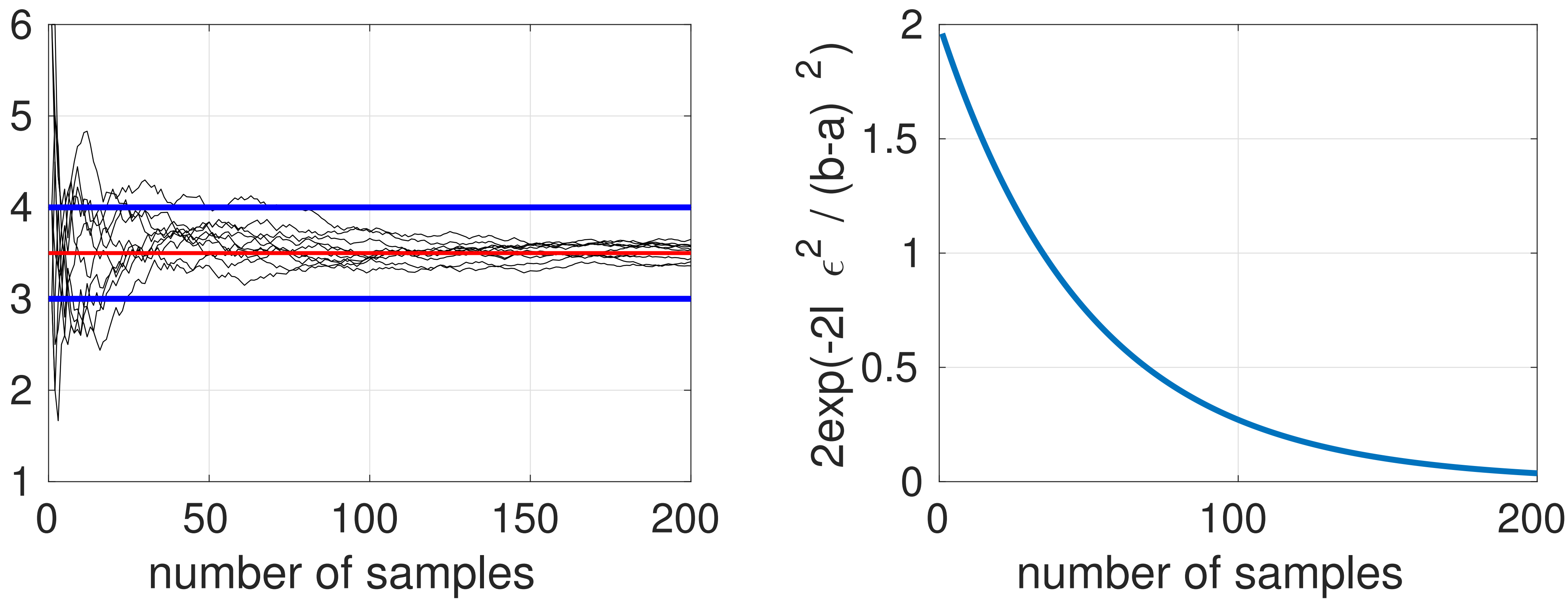
Theorem 1. *(Hoeffding inequality) Let* $\{z^1, \ldots, z^l\} \in [a,b]^l$ *be realizations of independent random variables with the same expected value* μ*. Then for any* $\varepsilon > 0$ *it holds that*

$$\mathbb{P}\left(\left| \frac{1}{l} \sum_{i=1}^{l} z^i - \mu \right| \geq \varepsilon \right) \leq 2e^{-\frac{2l\,\varepsilon^2}{(b-a)^2}}$$

Law of large numbers: example

- Rolling a die: $\mu = 3.5$, $\hat{\mu}_l = \frac{1}{l}\sum_{i=1}^{l} z_i$, $z_i \in [1,6]$, $\varepsilon = 0.5$.

$$\mathbb{P}\left(\left| \hat{\mu}_l - \mu \right| \geq \varepsilon \right) \leq 2e^{-\frac{2l\,\varepsilon^2}{(b-a)^2}}$$



Confidence interval

- Let $\hat{\mu}_l = \frac{1}{l}\sum_{i=1}^{l} z^i$ be the arithmetic average computed from $\{z^1,\ldots,z^l\} \in [a,b]^l$ sampled from rand vars with expected value μ.
- For which ε is μ in interval $(\hat{\mu}_l - \varepsilon, \hat{\mu}_l + \varepsilon)$ with probability at least γ ?

  Using the Hoeffding inequality we can write:

$$\mathbb{P}\Big(|\hat{\mu}_l - \mu| < \varepsilon\Big) = 1 - \mathbb{P}\Big(|\hat{\mu}_l - \mu| \geq \varepsilon\Big) \leq 1 - 2e^{-\frac{2\,l\,\varepsilon^2}{(b-a)^2}} = \gamma$$

  and solving the last equality for ε yields

$$\varepsilon = |b-a|\sqrt{\frac{\log(2) - \log(1-\gamma)}{2\,l}}$$

- Similarly, for fixed ε and γ we can get the minimal number of samples

$$l = \frac{\log(2) - \log(1-\gamma)}{2\,\varepsilon^2}\,(b-a)^2$$

  such that μ is in $(\hat{\mu}_l - \varepsilon, \hat{\mu}_l + \varepsilon)$ with probability at least γ.

Estimation of the expected risk from examples

- Given test examples $\mathcal{S}^l = \{(x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, l\}$, predictor $h\colon \mathcal{X} \to \mathcal{Y}$ and loss $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, we estimate the predictor's risk $R(h) = \mathbb{E}_{(x,y)\sim p}(\ell(y, h(x)))$ by $R_{\mathcal{S}^l}(h) = \frac{1}{l}\sum_{i=1}^{l} \ell(y^i, h(x^i))$.

- For fixed strategy h, the numbers $z^i = \ell(y^i, h(x^i)) \in [\ell_{\min}, \ell_{\max}]$, $i \in \{1, \ldots, l\}$, are realizations of i.i.d. random variables with the expected value $\mu = R(h)$.

- According to the Hoeffding inequality, for any $\varepsilon > 0$ the probability of seeing a "bad test set" can be bound by

$$\mathbb{P}\left(\left|R_{\mathcal{S}^l}(h) - R(h)\right| \geq \varepsilon\right) \leq 2e^{-\frac{2l\,\varepsilon^2}{(\ell_{\min}-\ell_{\max})^2}}$$

where by "bad test set" we mean that our empirical estimate deviates from the true risk by ε at least.

Learning algorithm

- The goal is to find the prediction rule $h\colon \mathcal{X} \to \mathcal{Y}$ minimizing $R(h)$ in the case when $p(x,y)$ is unknown.

- We are given a training set of examples

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

  drawn from i.i.d. random variables distributed according to $p(x,y)$.

- Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}} = \{h\colon \mathcal{X} \to \mathcal{Y}\}$ be a hypothesis space.

- The algorithm $A\colon \cup_{m=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{H}$ selects hypothesis $h_m = A(\mathcal{T}^m)$ based on training examples $\mathcal{T}^m$.

Learning by Empirical Risk Minimization

- The expected risk $R(h)$, i.e. the true but unknown objective, is replaced by the empirical risk computed from examples

$$R_{\mathcal{T}^m}(h) = \frac{1}{m}\sum_{i=1}^{m} \ell(y^i, h(x^i))$$

- The ERM learning algorithm returns h_m such that

$$h_m \in \operatorname*{Argmin}_{h\in\mathcal{H}} R_{\mathcal{T}^m}(h) \tag{1}$$

- Depending on the choice of $\mathcal{H}$, ℓ and algorithm solving (1) we get individual instances, e.g.: Support Vector Machines, Linear Regression, Logistic regression, Neural Networks learned by back-propagation, AdaBoost, ...

Example: ERM does not always work

- Let $\mathcal{X} = [a, b] \subset \mathbb{R}$, $\mathcal{Y} = \{+1, -1\}$, $\ell(y, y') = [\![y \neq y']\!]$, $p(x \mid y = +1)$ and $p(x \mid y = -1)$ be uniform distributions on $\mathcal{X}$ and $p(y = +1) = 0.8$.
- The optimal strategy is $h(x) = +1$ with the Bayes risk $R^* = 0.2$.
- Consider a "cheating" learning algorithm which for a given training set $\mathcal{T}^m = \{(x^1, y^1), \ldots, (x^m, y^m)\}$ returns strategy

$$h_m(x) = \begin{cases} y^j & \text{if } x = x^j \text{ for some } j \in \{1, \ldots, m\} \\ -1 & \text{otherwise} \end{cases}$$

- The empirical risk is $R_{\mathcal{T}^m}(h_m) = 0$ with probability 1 for any m.
- The expected risk is $R(h_m) = 0.8$ for any m.
- In case of unconstrained $\mathcal{H}$ we have no guarantee that the empirical risk $R_{\mathcal{T}^m}(h_m)$ is a good approximation of the true risk $R(h_m)$ regardless the number of examples m.

Errors characterizing a learning algorithm

- The best attainable (Bayes) risk is $R^* = \inf_{h \in \mathcal{Y}^\mathcal{X}} R(h)$
- The best predictor in $\mathcal{H}$ is $h_\mathcal{H} \in \operatorname{Argmin}_{h \in \mathcal{H}} R(h)$
- The predictor $h_m = A(\mathcal{T}_m)$ learned from $\mathcal{T}^m$ has risk $R(h_m)$

Excess error measures deviation of the learned predictor from the best one:

$$\underbrace{\Big(R(h_m) - R^*\Big)}_{\text{excess error}} = \underbrace{\Big(R(h_m) - R(h_\mathcal{H})\Big)}_{\text{estimation error}} + \underbrace{\Big(R(h_\mathcal{H}) - R^*\Big)}_{\text{approximation error}}$$

Questions:

- Which of the quantities are random and which are not?
- What cases the errors?
- How do the errors depend on $\mathcal{H}$ and the number of examples m?

Statistically consistent learning algorithm

Definition 1. *The algorithm* $A\colon \cup_{m=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^m \rightarrow \mathcal{H}$ *is statistically consistent in* $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ *if for any* $p(x, y)$ *and* $\varepsilon > 0$ *it holds that*

$$\lim_{m\rightarrow\infty} \mathbb{P}\bigg(R(h_m) - R(h_{\mathcal{H}}) \geq \varepsilon \bigg) = 0$$

where $h_m = A(\mathcal{T}^m)$ *is the hypothesis returned by the algorithm* A *for training set* $\mathcal{T}^m$ *generated from* $p(x, y)$*.*

- The statistically consistent means that we can make the estimation error arbitrarily small if we have enough examples.
- Is the ERM algorithm statistically consistent ?

Uniform Law of Large Numbers

Definition 2. *The hypothesis space $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ satisfies the uniform law of large numbers if for all $\varepsilon > 0$ it holds that*

$$\lim_{m\to\infty} \mathbb{P}\left(\sup_{h\in\mathcal{H}} \left| R(h) - R_{\mathcal{T}^m}(h) \right| \geq \varepsilon \right) = 0$$

- ULLN says that the probability of seeing a "bad training set" for at least one hypothesis from $\mathcal{H}$ can be made arbitrarily low if we have enough examples.

Theorem 2. *If $\mathcal{H}$ satisfies ULLN then ERM is statistically consistent in $\mathcal{H}$.*

Proof: ULLN implies consistency of ERM

For fixed $\mathcal{T}^m$ and $h_m \in \operatorname{Argmin}_{h\in\mathcal{H}} R_{\mathcal{T}^m}(h)$ we have:

$$
\begin{aligned}
R(h_m) - R(h_{\mathcal{H}}) &= \Big(R(h_m) - R_{\mathcal{T}^m}(h_m)\Big) + \Big(R_{\mathcal{T}^m}(h_m) - R(h_{\mathcal{H}})\Big)\\
&\le \Big(R(h_m) - R_{\mathcal{T}^m}(h_m)\Big) + \Big(R_{\mathcal{T}^m}(h_{\mathcal{H}}) - R(h_{\mathcal{H}})\Big)\\
&\le 2\sup_{h\in\mathcal{H}} \Big|R(h) - R_{\mathcal{T}^m}(h)\Big|
\end{aligned}
$$

Therefore $\varepsilon \le R(h_m) - R(h_{\mathcal{H}})$ implies $\frac{\varepsilon}{2} \le \sup_{h\in\mathcal{H}} \Big|R(h) - R_{\mathcal{T}^m}(h)\Big|$ and

$$
\mathbb{P}\Big(R(h_m) - R(h_{\mathcal{H}}) \ge \varepsilon\Big) \le \mathbb{P}\Big(\sup_{h\in\mathcal{H}} \Big|R(h) - R_{\mathcal{T}^m}(h)\Big| \ge \frac{\varepsilon}{2}\Big)
$$

so if converges the RHS to zero (ULLN) so does the LHS (estimation error).

ULLN for finite hypothesis space

- Let us assume a finite hypothesis space $\mathcal{H} = \{h_1, \ldots, h_K\}$.
- We define the set of all "bad" training sets for a hypothesis $h \in \mathcal{H}$ as

$$\mathcal{B}(h) = \left\{ \mathcal{T}^m \in (\mathcal{X} \times \mathcal{Y})^m \,\middle|\, \left| R_{\mathcal{T}^m}(h) - R(h) \right| \geq \varepsilon \right\}$$

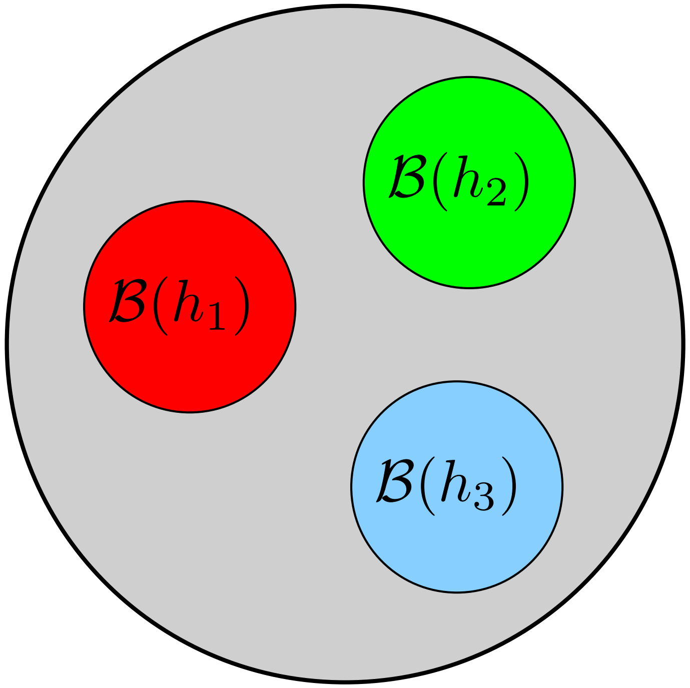
- We use the union bound to upper bound the probability of seeing a bad training set for at least one hypothesis from $h \in \mathcal{H}$

$$\begin{aligned} &\mathbb{P}\Big(\max_{h \in \mathcal{H}} |R_{\mathcal{T}^m}(h) - R(h)| \geq \varepsilon \Big) \\ &\qquad = \mathbb{P}\Big(\mathcal{T}^m \in \mathcal{B}(h_1) \bigvee \mathcal{T}^m \in \mathcal{B}(h_2) \bigvee \cdots \bigvee \mathcal{T}^m \in \mathcal{B}(h_K) \Big) \\ &\qquad\qquad \leq \sum_{h \in \mathcal{H}} \mathbb{P}(\mathcal{T}^m \in \mathcal{B}(h)) \end{aligned}$$

ULLN for finite hypothesis space

- Example: the union bound for three hypotheses

$$\mathbb{P}\Big(\mathcal{T}^m \in \mathcal{B}(h_1) \bigvee \mathcal{T}^m \in \mathcal{B}(h_2) \bigvee \mathcal{T}^m \in \mathcal{B}(h_3)\Big) \leq \sum_{i=1}^{3} \mathbb{P}(\mathcal{T}^m \in \mathcal{B}(h_i))$$



- The union bound is tight if the events are mutually exclusive (i.e. each $\mathcal{T}^m$ is bad for one hypothesis at most) as is shown in the figure.

ULLN for finite hypothesis space

- Combining the union bound with the Hoeffding inequality yields

$$\mathbb{P}\Big(\max_{h\in\mathcal{H}} |R_{\mathcal{T}^m}(h) - R(h)| \geq \varepsilon\Big) \leq \sum_{h\in\mathcal{H}} \mathbb{P}(\mathcal{T}^m \in \mathcal{B}(h)) \leq 2|\mathcal{H}| e^{-\frac{2m\,\varepsilon^2}{(b-a)^2}}$$

- Therefore we see that

$$\lim_{m\to\infty} \mathbb{P}\Big(\max_{h\in\mathcal{H}} |R_{\mathcal{T}^m}(h) - R(h)| \geq \varepsilon\Big) = 0$$

Corollary 1. *The finite hypothesis space satisfies the uniform law of large numbers.*

Confidence intervals for finite hypothesis space

- We have generalized the Hoeffding inequality for a finite hypothesis space $\mathcal{H}$:

$$\mathbb{P}\Big(\max_{h \in \mathcal{H}} |R_{\mathcal{T}^m}(h) - R(h)| \geq \varepsilon \Big) \leq 2|\mathcal{H}| e^{-\frac{2m\,\varepsilon^2}{(b-a)^2}}$$

- For which ε is $R(h)$ in the interval $(R_{\mathcal{T}^m}(h) - \varepsilon, R_{\mathcal{T}^m}(h) + \varepsilon)$ with the probability $1 - \delta$ at least, regardless what $h \in \mathcal{H}$ we consider ?

$$\begin{aligned} \mathbb{P}\Big(\max_{h \in \mathcal{H}} |R_{\mathcal{T}^m}(h) - R(h)| < \varepsilon \Big) &= 1 - \mathbb{P}\Big(\max_{h \in \mathcal{H}} |R_{\mathcal{T}^m}(h) - R(h)| \geq \varepsilon \Big) \\ &\leq 1 - 2|\mathcal{H}| e^{-\frac{2m\,\varepsilon^2}{(b-a)^2}} = 1 - \delta \end{aligned}$$

and solving the last equality for ε yields

$$\varepsilon = (b-a)\sqrt{\frac{\log 2|\mathcal{H}| + \log \frac{1}{\delta}}{2m}}$$

Generalization bound for finite hypothesis space

Theorem 3. *Let $\mathcal{H}$ be a finite hypothesis space 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. random variables with distribution $p(x, y)$. Then, for any $0 < \delta < 1$, with probability at least $1 - \delta$ the inequality*

$$R(h) \leq R_{\mathcal{T}^m}(h) + (b - a)\sqrt{\frac{\log 2|\mathcal{H}| + \log \frac{1}{\delta}}{2m}}$$

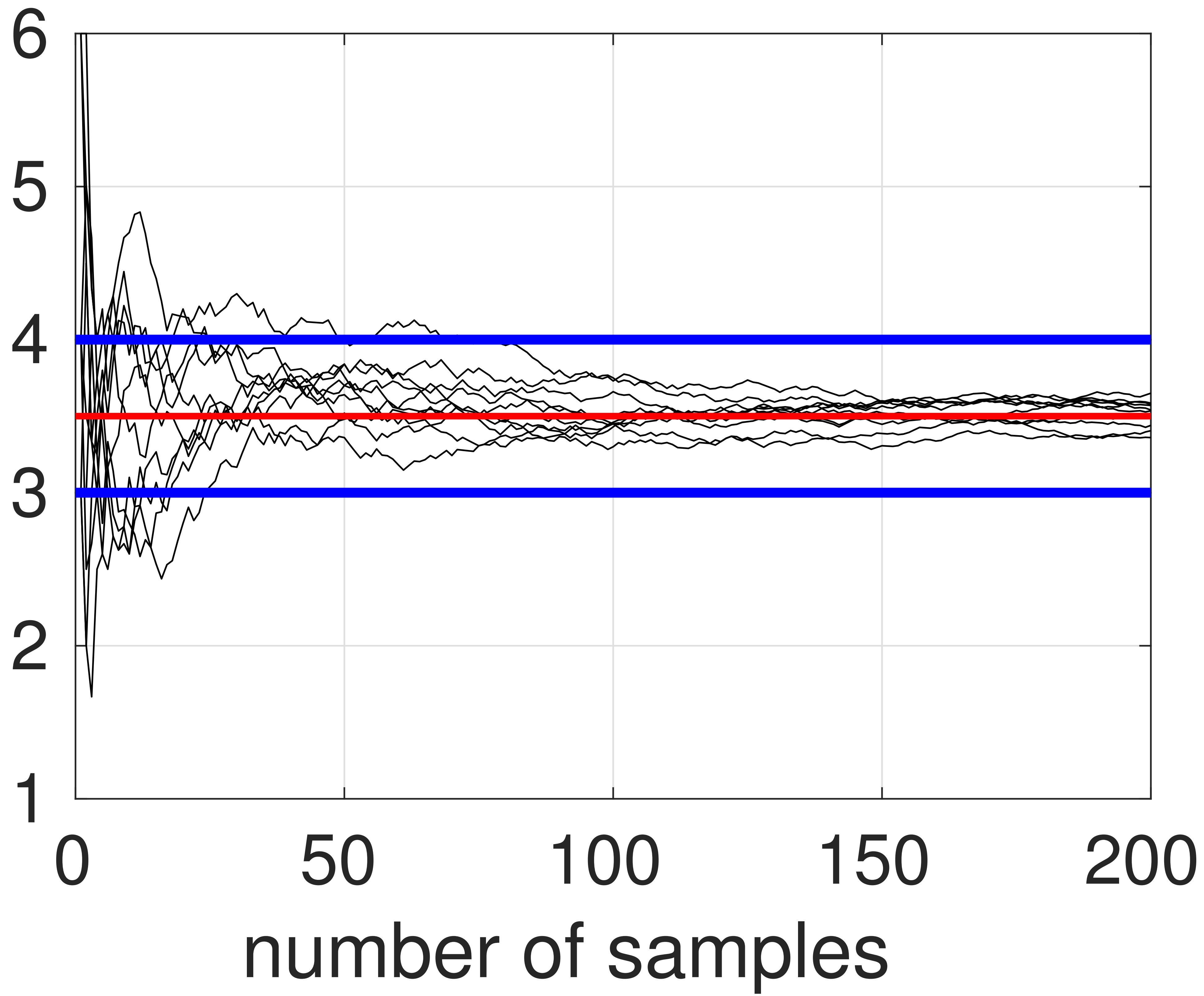
holds for any $h \in \mathcal{H}$ and any loss function $\ell \colon \mathcal{Y} \times \mathcal{Y} \to [a, b]$.

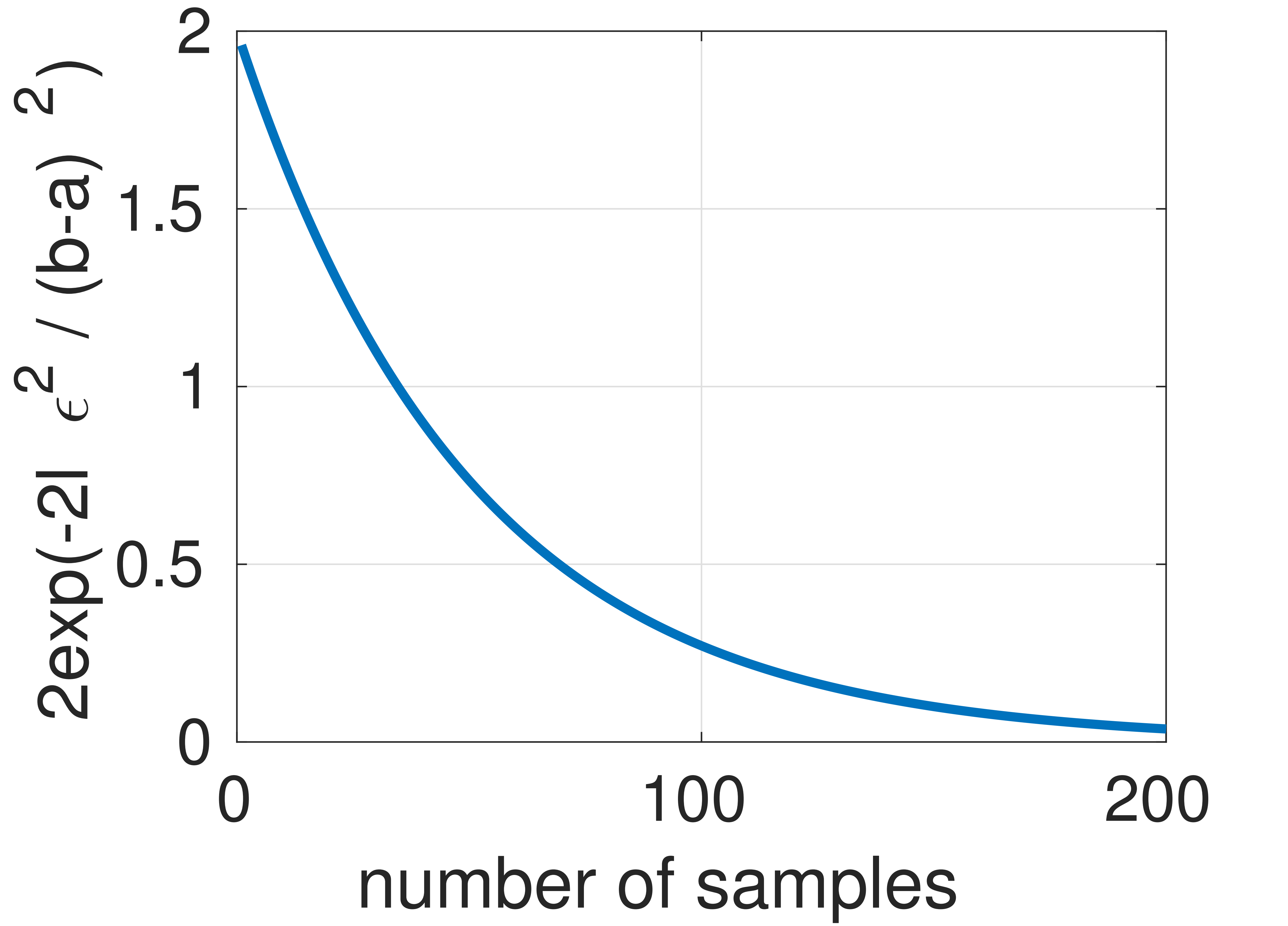
- The "worst-case" bound in Theorem 3 holds for any $h \in \mathcal{H}$, in particular, for the ERM algorithm which minimizes the first term.
- The second term suggests that we have to use $\mathcal{H}$ with appropriate cardinality (complexity); e.g. if m is small and $|\mathcal{H}|$ is high we can overfit.

Summary

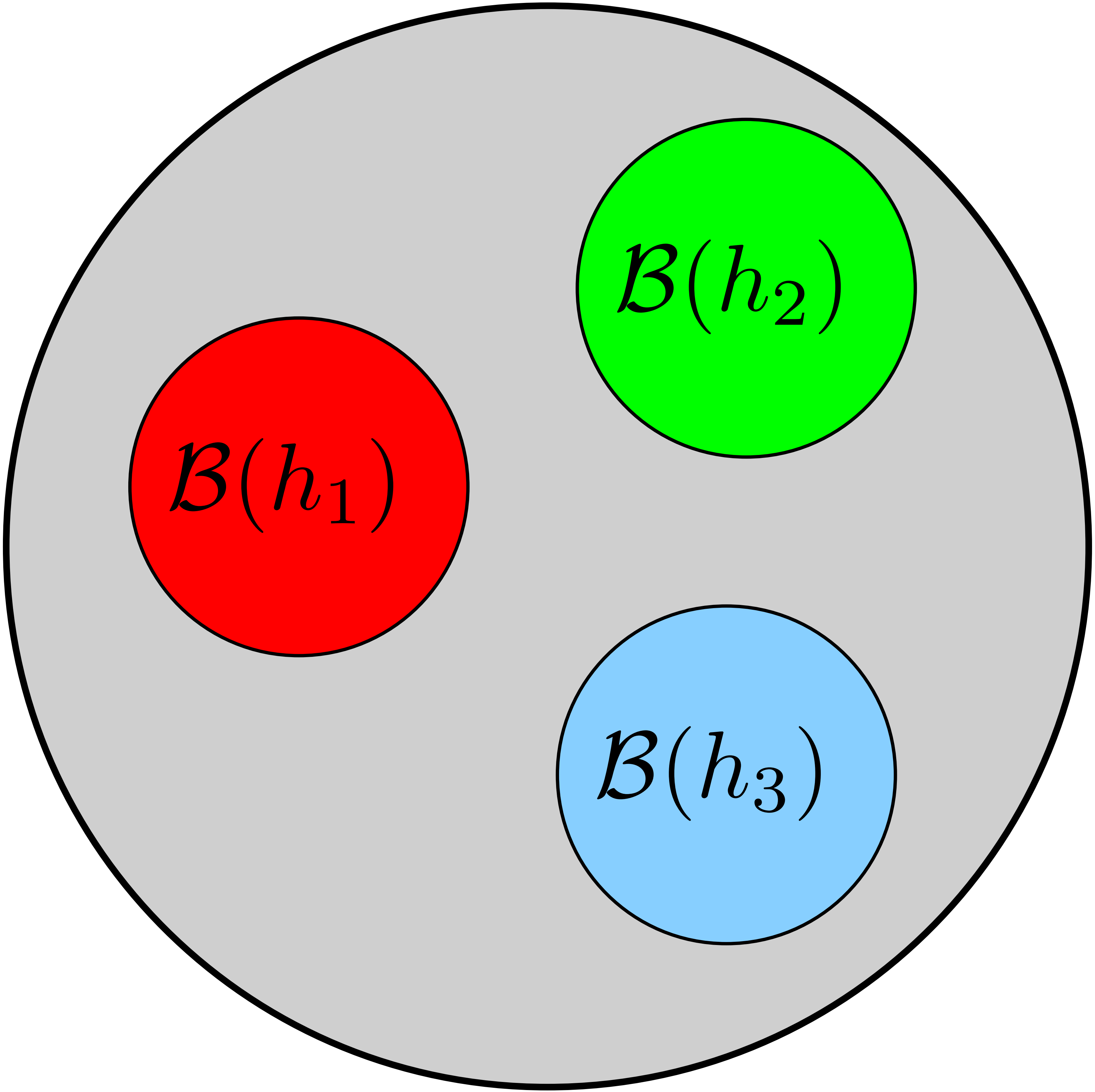
Topics covered in the lecture:

- Prediction problem
- Test risk and its justification by the law of large numbers
- Empirical Risk Minimization
- Excess error $=$ estimation error $+$ approximation error
- Statistical consistency of learning algorithm
- Uniform law of large numbers
- Generalization bound for finite hypothesis space


6
5
4
3
2
1
0
50
100
150
200
number of samples



2
1.5
1
0.5
0
0
100
200
number of samples
2exp(-2l ϵ^2 / (b-a)^2)


$\mathcal{B}(h_2)$
$\mathcal{B}(h_1)$
$\mathcal{B}(h_3)$