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

Czech Technical University in Prague

- Prediction problem
- Empirical Risk Minimization
- Statistical consistency

The prediction problem

- $\mathcal{X}$ 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

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

Why we need learning ? . . . because we don't know $p(x, y)$

- We will address the problem when we can only collect examples $\{(x^1, y^1), (x^2, y^2), \ldots\}$ drawn from the i.i.d. random variables distributed according to the unknown $p(x, y)$.

Estimation of the risk by using test 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 $p(x, y)$.

- A a prediction strategy $h \colon \mathcal{X} \to \mathcal{Y}$ can be evaluated by the empirical risk computed on the test examples

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

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

Law of large numbers

- Arithmetic mean of the results of random trials will become 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* $\{z^1, \dots, z^l\} \in [a,b]^l$ *be realizations of independent random variables with the same expected value* μ. *Then for any* $\varepsilon > 0$ *it holds that*

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

Estimation of the risk by using test examples

- We are interested in the deviation $\left|R_{\mathcal{S}^l}(h) - R(h)\right|$ which equals to

$$\left|\frac{1}{l}\sum_{i=1}^{l}\ell(y^i,h(x^i)) - \mathbb{E}_{(x,y)\sim p}(\ell(y,h(x)))\right| = \left|\frac{1}{l}\sum_{i=1}^{l} z^i - \mu\right|$$

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

- According to the Hoeffding inequality, for any $\varepsilon > 0$ it holds that

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

i.e, probability of seeing a “bad test set” decreases exponentially fast with l.

Learning from examples by empirical risk minimization

- 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 $p(x, y)$.

- The expected risk $R(h)$ to be minimized is replaced by the empirical risk evaluated on training 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

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

where $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}} = \{h \colon \mathcal{X} \to \mathcal{Y}\}$ is a hypothesis space.

- The choice of $\mathcal{H}$ is of a key importance.

Error of the 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 learned from $\mathcal{T}^m$ has risk $R(h_m)$

Excess error measures how far is the learned predictor from the best one:

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

Remarks:

- The approximation error (not random) is determined by fixing $\mathcal{H}$
- The estimation error specifies how much we lose when learning from examples $\mathcal{T}^m$ instead of using the true $p(x, y)$.

Statistically consistent learning algorithm

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

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

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 the probability that the estimation error happens to be high can be pushed arbitrarily low if we have enough examples.
- Is the ERM algorithm statistically consistent ?

Example: ERM is not always statistically consistent

- 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 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.
- For unconstrained $\mathcal{H}$ the empirical risk $R_{\mathcal{T}^m}(h_m)$ may not be a good approximation of the true risk $R(h_m)$ even if m is arbitrary large.

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. *The ULLN satisfied for $\mathcal{H}$ implies the statistical consistency of ERM 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:

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

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

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

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

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

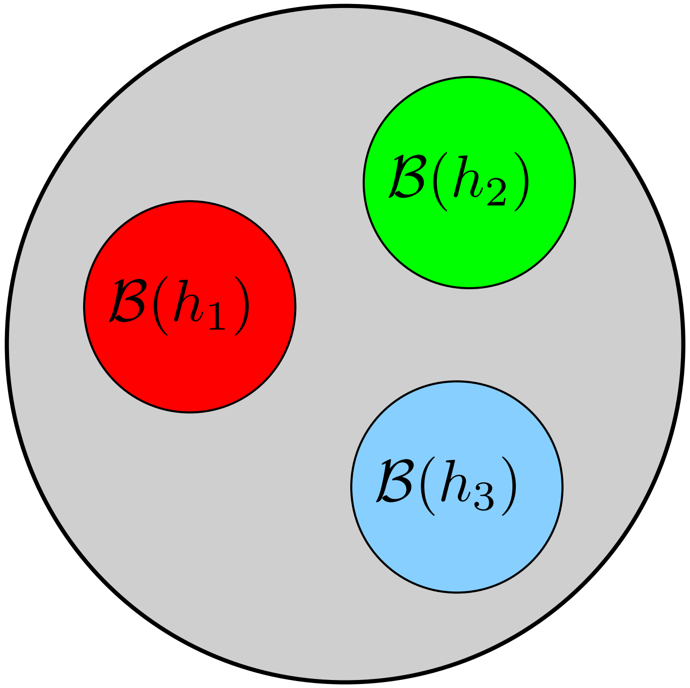
- We use the union bound to upper bound the probability of seeing a bad training set for at least one hypothesis from $h \in \mathcal{H}$

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

ULLN for finite hypothesis space

- Example: the union bound for three hypotheses

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



- The union bound is tight if the events are mutually exclusive as is the case 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

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

Theorem 3. *The finite hypothesis space satisfies the uniform law of large numbers.*

Confidence intervals for finite hypothesis space

- We have derived a bound valid for a finite $\mathcal{H}$:

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

- Denoting $\delta = 2|\mathcal{H}|e^{\frac{-2m\,\varepsilon^2}{(b-a)^2}}$ and solving for ε gives

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

- We see that for any $\delta > 0$, with probability at least $1-\delta$ it holds that

$$\max_{h\in\mathcal{H}} |R_{\mathcal{T}^m}(h) - R(h)| \leq \varepsilon(m,\delta)$$

and hence also $R(h) \in (R_{\mathcal{T}^m}(h) - \varepsilon(m,\delta,|\mathcal{H}|), R_{\mathcal{T}^m}(h) + \varepsilon(m,\delta,|\mathcal{H}|))$

Generalization bound for finite hypothesis space

Theorem 4. *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*

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

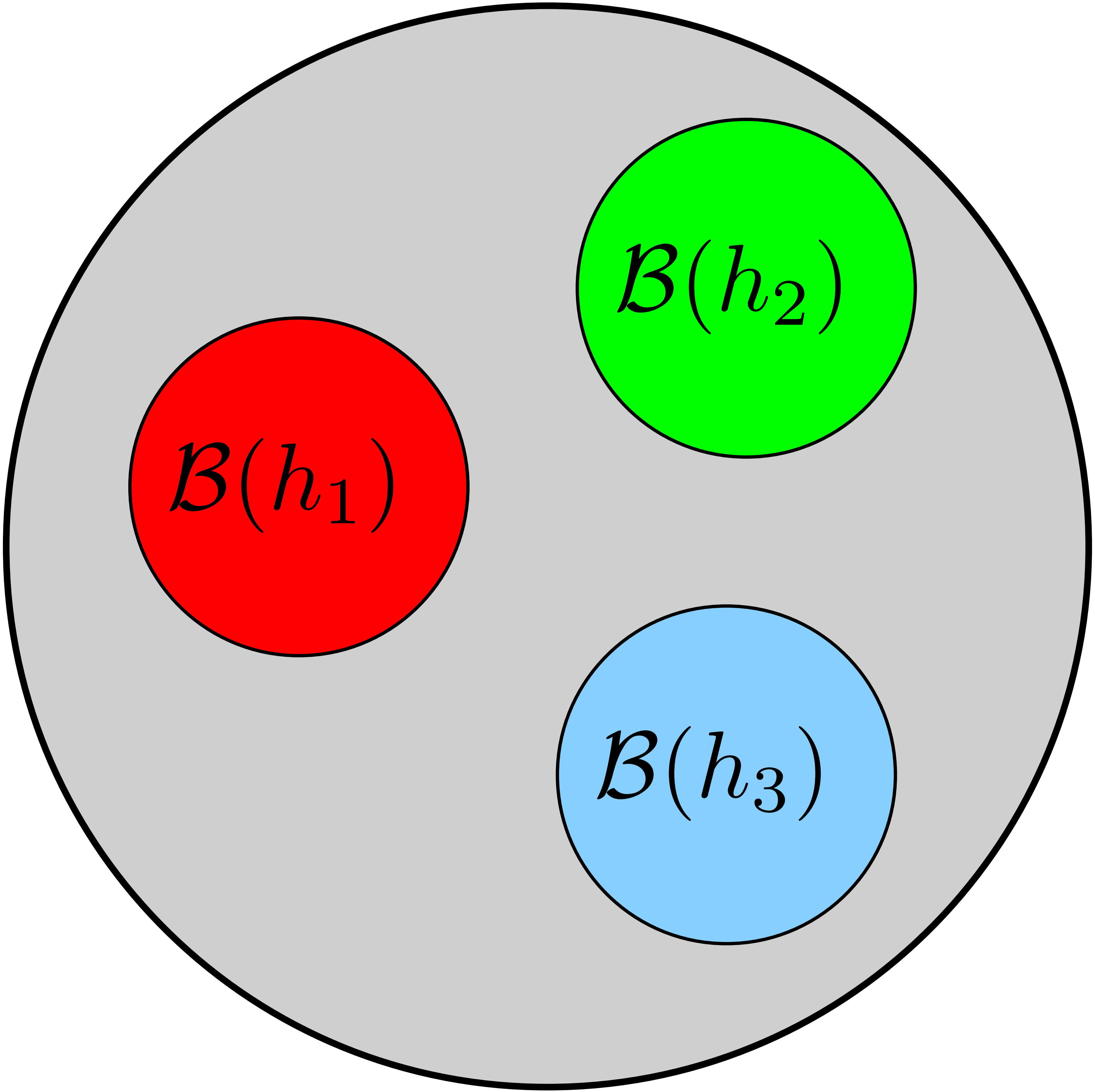
holds for any $h \in \mathcal{H}$ and any loss function $\ell \colon \mathcal{Y} \times \mathcal{Y} \to [a, b]$.

- The “worst-case” bound in Theorem 4 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


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