# Optimization Methods In Machine Learning

## Lecture 3: Empirical Risk Minimization and Structure Risk Minimization

Professor Katya Scheinberg

Lehigh University

Spring 2016

# Outline

Empirical Risk Minimization (ERM)
- Incorrect Argument for ERM
- Post-hoc Guarantees
- Relative Learning Guarantees

Structure Risk Minimization (SRM)
- Overfitting
- Structure Risk Minimization
- Choosing a Complexity Level

This lecture is taken from a short course at UT Austin taught by N. Srebro and K. Scheinberg in 2011.

# Outline

# Empirical Risk Minimization (ERM)

## Incorrect Argument for ERM

- Recall examples of complexity of hypothesis class from the previous lecture:
  - $\mathcal{H}_b = \{$"predictors based only on month and day of birthdate"$\}$, $|\mathcal{H}_b| = 2^{365}$.
  - $\mathcal{H}_n = \{$"predictors based only on nationality"$\}$, $|\mathcal{H}_n| = 2^n, n$ is the number of countries.
  - $\mathcal{H}_h = \{$"predictors based only on short or long hair"$\}$, $|\mathcal{H}_h| = 2^2$.
  - $\mathcal{H}_p = \{$"predictors based only on last four digits of phone number"$\}$, $|\mathcal{H}_p| = 2^{10000}$.

# Empirical Risk Minimization (ERM)

## Correcting Argument for justifying ERM

- Recall our re-interpretation of Hoeffding's inequality:

$$|R_{01}(h) - \hat{R}_{s,01}(h)| \leq \sqrt{\frac{\log \frac{2}{\delta}}{2m}} \text{ with probability at least } (1-\delta).$$

  Then recall from our previous lecture that

$$\begin{aligned} R_{01}(\hat{h}) \leq \hat{R}_{s,01}(\hat{h}) + \sqrt{\frac{\log \frac{2}{\delta}}{2m}} &\leq \hat{R}_{s,01}(h^*) + \sqrt{\frac{\log \frac{2}{\delta}}{2m}} \\ &\leq R_{01}(h^*) + 2\sqrt{\frac{\log \frac{2}{\delta}}{2m}} \end{aligned}$$

- What is flawed in the above argument?
  Reason: the problem is that Hoeffding's inequality holds for each $h$ separately but does **NOT** hold for all $h \in \mathcal{H}$. We neglected the complexity of the hypothesis class in our previous argument.

# Empirical Risk Minimization (ERM)

## Correcting Argument for justifying ERM

- Recall the knowledge of the union bound from probability theory

### Theorem (Boole's inequality, also known as the union bound)

*For a countable set of different events $A_1, A_2, \dots$, we have*

$$\mathbb{P}(\bigcup_i A_i) \leq \sum_i \mathbb{P}(A_i).$$

- Correct the flaw by using the union bound, we should have the following inequality hold,

### Theorem

$$\mathbb{P}\{\forall h \in \mathcal{H}, |\hat{R}_{s,01}(h) - R_{01}(h)| \leq \sqrt{\frac{\log|\mathcal{H}| + \log\frac{2}{\delta_{bad}}}{2m}}\} \geq 1 - \delta_{bad}.$$

# Empirical Risk Minimization (ERM)

## Correcting Argument for justifying ERM

**Proof** Considering the union bound, let event $A_i =$ "hypothesis $h_i$ looks misleadingly good", i.e. "$h_i$ is cheating, explicitly written out as $|R_{01}(h_i) - \hat{R}_{s,01}(h_i)| \geq \epsilon$. Then $\bigcup_{i \in \mathcal{H}} A_i =$ "at least one hypothesis from $\mathcal{H}$ is cheating". So the following inequality holds,

$$
\begin{aligned}
&\mathbb{P}(\text{all hypotheses from } \mathcal{H} \text{ "behave well"}) \\
&= 1 - \mathbb{P}(\text{at least one hypothesis from } \mathcal{H} \text{ is cheating}) \\
&= 1 - \mathbb{P}(\bigcup_{i \in \mathcal{H}} A_i) \geq 1 - \sum_{i \in \mathcal{H}} \mathbb{P}(A_i) \\
&= 1 - \sum_{i \in \mathcal{H}} \mathbb{P}\{|R_{01}(h_i) - \hat{R}_{s,01}(h_i)| \geq \epsilon\}
\end{aligned}
$$

# Empirical Risk Minimization (ERM)

## Correcting Argument for justifying ERM

- Event "all hypotheses from $\mathcal{H}$ 'behave well' " can be explicitly written as "$\forall h \in \mathcal{H}, |\hat{R}_{s,01}(h) - R_{01}(h)| \leq \epsilon$. Then

$$\mathbb{P}\{\forall h \in \mathcal{H}, |\hat{R}_{s,01}(h) - R_{01}(h)| \leq \epsilon\}$$
$$\geq 1 - \sum_{i \in \mathcal{H}} \mathbb{P}\{|R_{01}(h_i) - \hat{R}_{s,01}(h_i)| \geq \epsilon\}$$

Recall from our previous lecture $\mathbb{P}\{|R_{01}(h) - \hat{R}_{s,01}(h)| \geq \epsilon\} \leq 2e^{-2\epsilon^2 m}$ for any $h \in \mathcal{H}$, the above inequality is equivalent to

$$\mathbb{P}\{\forall h \in \mathcal{H}, |\hat{R}_{s,01}(h) - R_{01}(h)| \leq \epsilon\} \geq 1 - 2|\mathcal{H}|e^{-2\epsilon^2 m}$$

# Empirical Risk Minimization (ERM)

Example

- By letting $\delta_{bad} = 2|\mathcal{H}|e^{-2\epsilon^2 m}$, we have $\epsilon = \sqrt{\frac{\log|\mathcal{H}|+\log\frac{2}{\delta_{bad}}}{2m}}$. Thus,

$$\mathbb{P}\{\forall h \in \mathcal{H}, |\hat{R}_{s,01}(h) - R_{01}(h)| \leq \sqrt{\frac{\log|\mathcal{H}| + \log\frac{2}{\delta_{bad}}}{2m}}\} \geq 1 - \delta_{bad}$$

Proof is finished here.

## Example (3.1)

Suppose we have an hypothesis class of size $|\mathcal{H}| = 10000$, and we want the probability of there being some "possibly misleading" hypothesis in $\mathcal{H}$ (such that its sample date set performance differs from its source distribution performance by more than $\epsilon = 0.01$) to be no higher than $\delta_{bad} = e^{-7}$. How many samples do we need?

**Solution:** $\epsilon = 0.01 \leq sqrt\frac{\log 10000+\log 2/e^{-7}}{2m} \Longrightarrow m \geq \frac{2}{17}\cdot 10^4 \approx 1200.$

# Empirical Risk Minimization (ERM)

- Similar as previous lectures, let $h^* := \arg\min_{h\in\mathcal{H}} R_{01}(h)$, $\hat{h} := \arg\min_{h\in\mathcal{H}} \hat{R}_{s,01}(h)$. From the inequality

$$\mathbb{P}\{\forall h \in \mathcal{H}, |\hat{R}_{s,01}(h) - R_{01}(h)| \leq \sqrt{\frac{\log|\mathcal{H}| + \log\frac{2}{\delta_{bad}}}{2m}}\} \geq 1 - \delta_{bad},$$

we have: with probability at least $(1 - \delta_{bad})$,

$$\begin{aligned} R_{01}(\hat{h}) &\leq \hat{R}_{s,01}(\hat{h}) + \sqrt{\frac{\log|\mathcal{H}| + \log\frac{2}{\delta_{bad}}}{2m}} \\ &\leq \hat{R}_{s,01}(h^*) + \sqrt{\frac{\log|\mathcal{H}| + \log\frac{2}{\delta_{bad}}}{2m}} \\ &\leq R_{01}(h^*) + 2\sqrt{\frac{\log|\mathcal{H}| + \log\frac{2}{\delta_{bad}}}{2m}}. \end{aligned}$$

# Empirical Risk Minimization (ERM)

## Post-hoc Guarantees

- With probability at least $(1 - \delta_{bad})$, we will have gotten a sample set realization $s$ for which the **post-hoc generalization guarantee** holds:

$$R_{01}(\hat{h}) \leq \hat{R}_{s,01}(\hat{h}) + \sqrt{\frac{\log|\mathcal{H}| + \log\frac{2}{\delta_{bad}}}{2m}}.$$

# Empirical Risk Minimization (ERM)

## Relative Learning Guarantees

- With probability at least $(1-\delta_{bad})$, we will have the **(relative) learning guarantee**

$$R_{01}(\hat{h}) \leq R_{01}(h^*) + 2\sqrt{\frac{\log|\mathcal{H}| + \log\frac{2}{\delta_{bad}}}{2m}}.$$

# Empirical Risk Minimization (ERM)

Examples

## Example (3.2)

Suppose for hypothesis class $\mathcal{H}$, $h(x) = \{x_i \geq \theta, i = 1, 2\}$. (Assume the machine is 64-bit.)

- For $\theta$, there $2^{64}$ possible values, we have 2 hypotheses for each $\theta$ in 2-dimensional space.
  So the complexity of hypothesis class
  $$|\mathcal{H}| = 2^{64} + 2^{64} = 2^{65} \Rightarrow \log |\mathcal{H}| \approx 45.$$

## Example (3.3)

$h(x) = \{w^T x + b \geq 0\}$, $\mathcal{H} = \{h_{w,b} | w \in \mathbb{R}^d, b \in \mathbb{R}\}$. (Assume the machine is 64-bit.)

- Similarly, the complexity of hypothesis class $(D := d + 1)$
  $$|\mathcal{H}| = 2^{64D} + 2^{64D} = 2^{65D} \Rightarrow \log |\mathcal{H}| \approx 45D(\textit{linear!}).$$

# Outline

# Structure Risk Minimization (SRM)

Overfitting

## Example (3.4)

$x \in \mathbb{R}, y = \{+1, -1\}$. Given a series of data samples $(x_i, y_i)$.

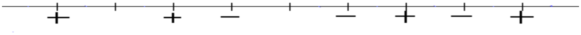


Figure: Given data samples

$\mathcal{H} = \{x \to \text{sign}[\sin(wx + \theta)] \; | w \in \mathbb{R}, \theta \in \mathbb{R}\}$ can overfit any data when $w$ is sufficiently large.

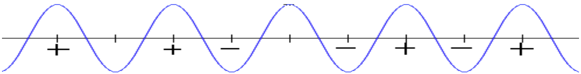


Figure: Example of overfitting

# Structure Risk Minimization (SRM)

- Recall from our previous slide of relative learning guarantee, with probability at least $(1-\delta_{bad})$,

$$R_{01}(\hat{h}) \leq R_{01}(h^*) + 2\sqrt{\frac{\log|\mathcal{H}| + \log\frac{2}{\delta_{bad}}}{2m}},$$

where $R_{01}(h^*)$ is called approximation error and $R_{01}(\hat{h}) - R_{01}(h^*)$ is the estimation error which is bounded by $2\sqrt{\frac{\log|\mathcal{H}|+\log\frac{2}{\delta_{bad}}}{2m}}$.

# Structure Risk Minimization (SRM)

- If we are optimizing $R_{01}(h_k^*)(h_k^* \in \mathcal{H}_k)$ over a hierarchy of hypothesis classes $\mathcal{H}_0 \subseteq \mathcal{H}_5 \subseteq \mathcal{H}_{10} \subseteq \mathcal{H}_{15} \subseteq \dots$, $R_{01}(h_k^*)$ will get smaller as more complexity parameters can make better model by fitting the data.

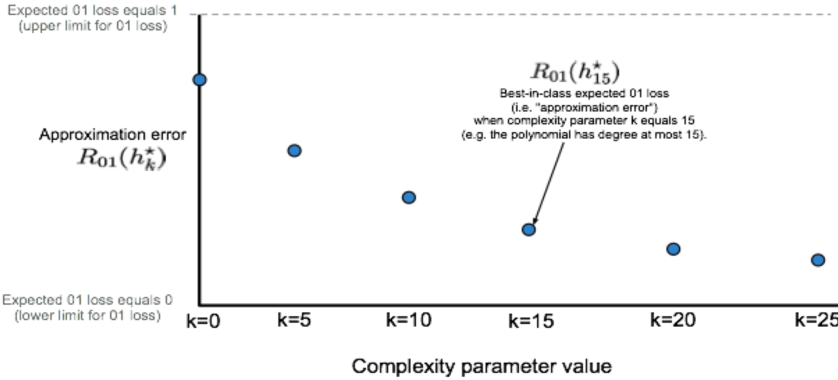


Figure: Optimizing $R_{01}(h_k^*)$ over a hierarchy of classes

# Structure Risk Minimization (SRM)

- However, the bound of estimation error $2\sqrt{\frac{\log|\mathcal{H}|+\log\frac{2}{\delta_{bad}}}{2m}}$ is increasing as complexity parameter value increases. SRM principle aims to balance the model's complexity against its success at fitting the data.

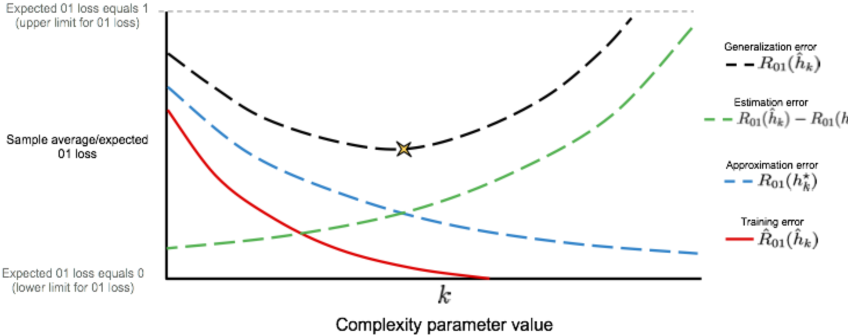


Figure: Behaviors of different errors over a hierarchy of classes

# Structure Risk Minimization (SRM)

## Choosing a Complexity Level

- There is a trade-off between the approximation error and the estimation error, so how do we know which class of hypothesis is better?
  1. Sample again! And test the predictors on the new sample, compare their behaviors.
  2. Cross-validation: Partition a sample set into complementary subsets, perform analysis on one subset (training set) and validate the analysis on the other subset (testing set). Can do this multiple times with different partitions.