Statistical Machine Learning

https://cvml.ist.ac.at/courses/SML_W20

Christoph Lampert

IST AUSTRIA

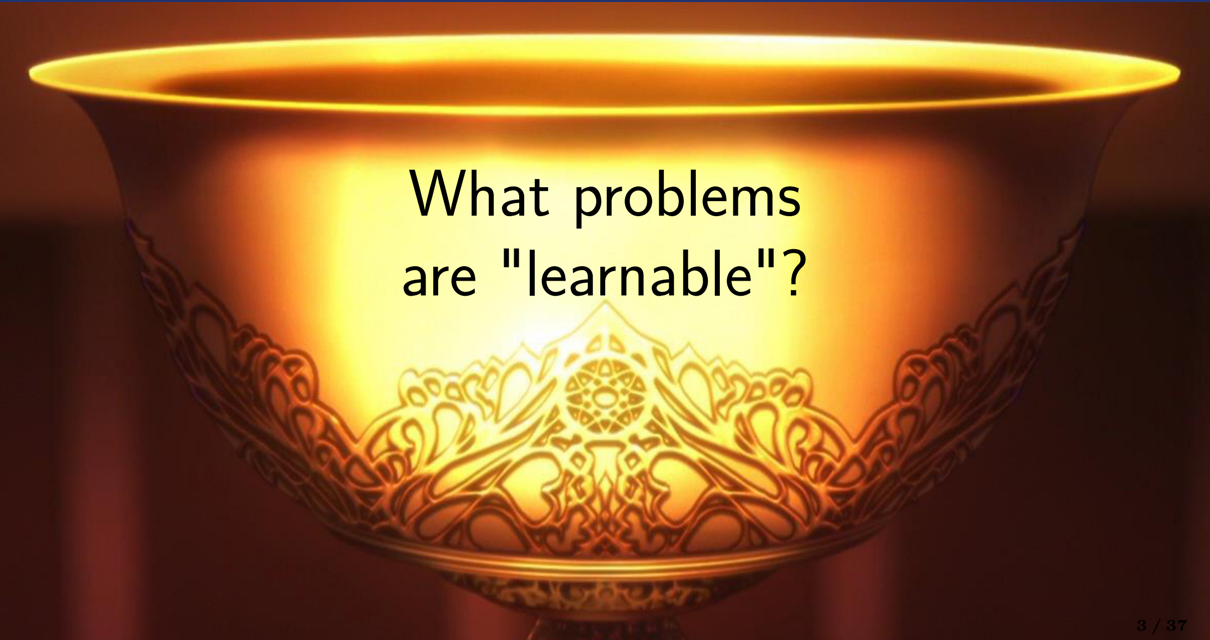
Institute of Science and Technology

Fall Semester 2020/2021

Lecture 7

Overview (tentative)

Date		no.	Topic
Oct 05	Mon	1	A Hands-On Introduction
Oct 07	Wed	2	Bayesian Decision Theory, Generative Probabilistic Models
Oct 12	Mon	3	Discriminative Probabilistic Models
Oct 14	Wed	4	Maximum Margin Classifiers, Generalized Linear Models
Oct 19	Mon	5	Estimators; Overfitting/Underfitting, Regularization, Model Selection
Oct 21	Wed	6	Bias/Fairness, Domain Adaptation
Oct 26	Mon	-	no lecture (public holiday)
Oct 28	Wed	7	Learning Theory I, Concentration of Measure
Nov 02	Mon	8	Learning Theory II
Nov 04	Wed	9	Deep Learning I
Nov 09	Mon	10	Deep Learning II
Nov 11	Wed	11	Unsupervised Learning
Nov 16	Mon	12	project presentations
Nov 18	Wed	13	buffer



What problems are "learnable"?

PAC Learning Scenario

- input set $\mathcal{X}$, label set $\mathcal{Y} = \{\pm 1\}$, loss $\ell(y, y') = [\![y \neq y']\!]$, data distribution $p(x, y)$
  for now: assume **deterministic labels**, $y = f(x)$ for some unknown $f : \mathcal{X} \to \mathcal{Y}$

- training set $\mathcal{D}_m = \{(x_1, y_1), \dots, (x_m, y_m)\} \overset{i.i.d.}{\sim} p(x, y)$
- hypothesis set $\mathcal{H} \subseteq \{h : \mathcal{X} \to \mathcal{Y}\}$, e.g. "all linear classifiers in $\mathbb{R}^d$"
  for now: assume **realizability**, i.e. the true labeling function, f, lies in $\mathcal{H}$

Quantity of interest:

- risk $\mathcal{R}(h) \;=\; \mathbb{E}_{(x,y)\sim p(x,y)}\, \ell(\, y, h(x)\,) \;=\; \Pr_{x\sim p(x)}\{\, f(x) \neq h(x)\,\}$

"Learning" becomes "search with limited information":

- We know: there is at least one $h \in \mathcal{H}$ that fulfills $\mathcal{R}(h) = 0$.
- Questions: Can we find such h from $\mathcal{D}_m$? If yes, how large does m have to be?
- Answer: that depends on $\mathcal{H}$ (and pretty much nothing else)

Example (Learning a threshold)

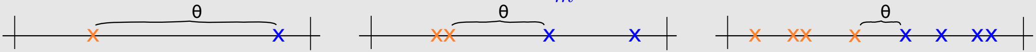
- $\mathcal{X} = [0,1], \quad \mathcal{Y} = \{\pm 1\}, \quad \ell(y,y') = [\![y \neq y']\!]$
- true labeling function $f^*(x) = \operatorname{sign}(x - \theta^*)$ for some $\theta^* \in [0,1]$
- data distribution $p(x,y) = p(x)p(y|x)$ with $p(y|x) = \delta_{y=f^*(x)}$
- hypothesis set $\mathcal{H} \subseteq \{h(x) = \operatorname{sign}(x - \theta) : \theta \in [0,1]\}$, "all threshold functions"
- training set $\mathcal{D}_m = \{(x_1,y_1),\dots,(x_m,y_m)\} \overset{i.i.d.}{\sim} p(x,y)$

How well will be able to determine θ^* from $\mathcal{D}_m$?

Example (Learning a threshold)

- $\mathcal{X} = [0,1]$, $\quad \mathcal{Y} = \{\pm 1\}$, $\quad \ell(y,y') = [\![y \neq y']\!]$
- true labeling function $f^*(x) = \operatorname{sign}(x - \theta^*)$ for some $\theta^* \in [0,1]$
- data distribution $p(x,y) = p(x)p(y|x)$ with $p(y|x) = \delta_{y=f^*(x)}$
- hypothesis set $\mathcal{H} \subseteq \{h(x) = \operatorname{sign}(x-\theta) : \theta \in [0,1]\}$, "all threshold functions"
- training set $\mathcal{D}_m = \{(x_1,y_1),\dots,(x_m,y_m)\} \overset{i.i.d.}{\sim} p(x,y)$

How well will be able to determine θ^* from $\mathcal{D}_m$?

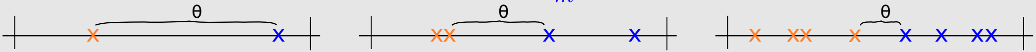


1) for any finite m some uncertainty about θ^* will remain
$\rightarrow$ we cannot hope to find f^* perfectly, only better and better approximations to it

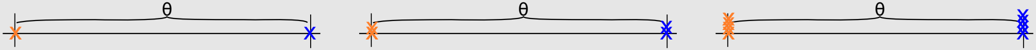
Example (Learning a threshold)

- $\mathcal{X} = [0,1], \quad \mathcal{Y} = \{\pm 1\}, \quad \ell(y,y') = [\![y \neq y']\!]$
- true labeling function $f^*(x) = \operatorname{sign}(x - \theta^*)$ for some $\theta^* \in [0,1]$
- data distribution $p(x,y) = p(x)p(y|x)$ with $p(y|x) = \delta_{y=f^*(x)}$
- hypothesis set $\mathcal{H} \subseteq \{h(x) = \operatorname{sign}(x - \theta) : \theta \in [0,1]\}$, "all threshold functions"
- training set $\mathcal{D}_m = \{(x_1,y_1),\dots,(x_m,y_m)\} \overset{i.i.d.}{\sim} p(x,y)$

How well will be able to determine θ^* from $\mathcal{D}_m$?



1) for any finite m some uncertainty about θ^* will remain
$\rightarrow$ we cannot hope to find f^* perfectly, only better and better approximations to it



2) for any finite m, there is a chance that the training data will be unlucky (and useless)
$\rightarrow$ we cannot be 100% certain that the approximation will behave well

Definition (Probably Approximately Correct (PAC) Learnability)

A hypothesis class $\mathcal{H}$ is called **PAC learnable** by an algorithm A, if

- for every $\epsilon > 0$ (accuracy → "approximate correct")
- and every $\delta > 0$ (confidence → "probably")

there exists an

- $m_0 = m_0(\epsilon, \delta) \in \mathbb{N}$ (minimal training set size)

such that

- for any probability distribution p over $\mathcal{X}$, and
- for any labeling function $f \in \mathcal{H}$, with $\mathcal{R}(f) = 0$,

when we run the learning algorithm A on a training set consisting of $m \geq m_0$ examples sampled i.i.d. from p, the algorithm returns a hypothesis $h \in \mathcal{H}$ that, with probability at least $1 - \delta$, fulfills $\mathcal{R}_p(h) \leq \epsilon$.

$$\forall m \geq m_0(\epsilon, \delta) \quad \Pr_{\mathcal{D}_m \sim p}[\,\mathcal{R}_d(A[\mathcal{D}_m]) > \epsilon\,] \leq \delta.$$

Note: for "efficient learning", A must run in poly$(m, \frac{1}{\epsilon}, \frac{1}{\delta},$ *"size of $\mathcal{D}_m$"*$)$.

Empirical Risk Minimization

What *learning algorithm*?

Definition (Empirical Risk Minimization (ERM) Algorithm)

input hypothesis set $\mathcal{H} \subseteq \{h : \mathcal{X} \to \mathcal{Y}\}$ (not necessarily finite)

input training set $\mathcal{D} = \{(x_1, y_1), \dots, (x_m, y_m)\}$

output $h \in \operatorname*{\mathbf{argmin}}_{h \in H} \frac{1}{m} \sum_{i=1}^{m} \ell(\, y_i, h(x_i)\,)$ (lowest training error)

ERM learns a classifier that has minimal training error.

- There might be multiple, we can't control which one.
- We already saw cases where ERM worked well and some where it didn't.
- Can we characterize when ERM works and when it fails?

Examples

A constant decision is PAC-learnable by ERM

- $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{\pm 1\}$, $\ell(y, y') = [\![y, y']\!]$
- $\mathcal{H} = \{h_+, h_-\}$ with $h_+(x) = +1$ and $h_-(x) = -1$
- p arbitrary

ERM needs only $m_0 = 1$ example, then its solution is unique and perfect.

Examples

A constant decision is PAC-learnable by ERM

- $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{\pm 1\}$, $\ell(y, y') = [\![y, y']\!]$
- $\mathcal{H} = \{h_+, h_-\}$ with $h_+(x) = +1$ and $h_-(x) = -1$
- p arbitrary

ERM needs only $m_0 = 1$ example, then its solution is unique and perfect.

Coordinate classifiers

- $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{\pm 1\}$, $\ell(y, y') = [\![y \neq y']\!]$
- $\mathcal{H} = \{h_1, \ldots, h_d\}$ with $h_i(x) = \operatorname{sign} x[i]$

Lemma

If p is uniform in $[-1, 1]^d$, ERM works for $m_0(\epsilon, \delta) = \lceil \log_2 \frac{d-1}{\delta} \rceil$

Proof: textbook

For general p, we might have to return hypothesis with $\epsilon > 0$, and have m_0 depend on ϵ.

Which $\mathcal{H}$ are PAC-learnable by ERM?

Can we prove general statements?

Theorem (PAC Learnability of finite hypothesis classes)

Let $\mathcal{H} = \{h_1, \dots, h_K\}$ be a finite hypothesis class and $f \in \mathcal{H}$ (i.e. the true labeling function is one of the hypotheses). Then $\mathcal{H}$ is PAC-learnable by the ERM algorithm with

$$m_0(\epsilon, \delta) = \lceil \frac{1}{\epsilon}(\log(|\mathcal{H}| + \log(1/\delta))\rceil$$

Proof: textbook

Which $\mathcal{H}$ are PAC-learnable by ERM?

Can we prove general statements?

Theorem (PAC Learnability of finite hypothesis classes)

Let $\mathcal{H} = \{h_1, \dots, h_K\}$ be a finite hypothesis class and $f \in \mathcal{H}$ (i.e. the true labeling function is one of the hypotheses). Then $\mathcal{H}$ is PAC-learnable by the ERM algorithm with

$$m_0(\epsilon, \delta) = \lceil \frac{1}{\epsilon}(\log(|\mathcal{H}| + \log(1/\delta)) \rceil$$

Proof: textbook

Corollary

Let $\mathcal{D}$ be a training set of size m. Let f_{ERM} be the result of running ERM on $\mathcal{D}$. Then

$$\mathcal{R}(f_{ERM}) \leq \frac{\log |\mathcal{H}| + \log(1/\delta)}{m} \tag{1}$$

Examples: Finite hypothesis classes

Model selection:

- Classifiers trained with K different hyperparameter settings. Can we be sure to pick the right one?

Finite precision:

- For $\mathcal{X} \subset \mathbb{R}^d$, the hypothesis set $\mathcal{H} = \{f(x) = \operatorname{sign}\langle w, x\rangle\}$ is infinite.
- But: on a computer, w is restricted, e.g. to 32-bit `floats`: $|\mathcal{H}_c| = 2^{32d}$.
  $m_0(\epsilon, \delta) = \frac{1}{\epsilon}(\log(|\mathcal{H}| + \log(1/\delta)) \approx \frac{1}{\epsilon}(22d + \log(1/\delta))$

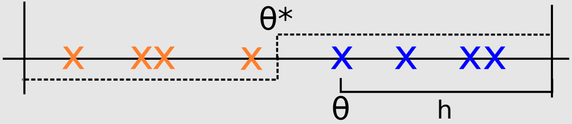
Implementation:

- $\mathcal{H} = \{$ all algorithms implementable in 10 KB C-code $\}$ is finite.

Logarithmic dependence on $|\mathcal{H}|$ makes even large (finite) hypothesis sets (kind of) practical.

What about infinite/continuous hypothesis classes?

Example (PAC-Learning for threshold functions)

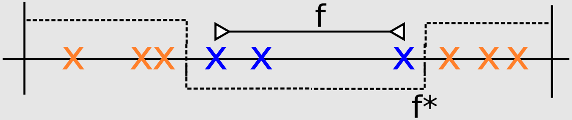


- $\mathcal{X} = [0,1]$, $\quad \mathcal{Y} = \{-1,1\}$, $\quad \mathcal{H} = \{h_\theta(x) = \text{sign}(x-\theta)]\!]$, for $\theta^* \in [0,1]\}$,
- $f^*(x) = h_{\theta^*}(x)$ for some $\theta^* \in [0,1]$
- ERM rule: $\quad \theta = \mathop{\mathbf{argmin}}_{\theta \in [0,1]} \frac{1}{m} \sum_{i=1}^{m} [\![h_\theta(x_i) \neq y_i]\!]$,
  any rule to make unique, e.g. "pick the smallest possible $+1$ region"

Claim: ERM learns f^* (in the PAC sense). Proof: textbook...

Which $\mathcal{H}$ are PAC-learnable by ERM?

Example (Learning Intervals)



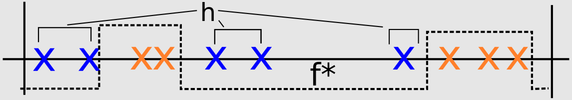
- $\mathcal{X} = [0,1]$, $\mathcal{Y} = \{0,1\}$, $\mathcal{H} = \Big\{ h_{[\theta_L,\theta_R]}(x) = [\![x \geq \theta_L \wedge x \leq \theta_R]\!]$, for $0 \leq \theta_L \leq \theta_R \leq 1 \Big\}$,
- $f(x) = h_{[\theta_L^*,\theta_R^*]}(x)$ for some $0 \leq \theta_L^* \leq \theta_R^* \leq 1$.
- training set $S = \{(x_1,y_1),\dots,(x_m,y_m)\}$
- ERM rule: $\quad h = \operatorname*{\mathbf{argmin}}_{[a,b]} \frac{1}{m}\sum_{i=1}^{m} [\![h_{[a,b]}(x_i) \neq y_i]\!]$,

  to make unique pick *smallest possible "+1" interval*

Claim: ERM learns f^* (in the PAC sense). Proof: textbook...

Which $\mathcal{H}$ are PAC-learnable by ERM?

Example (Learning Unions of Intervals)



- $\mathcal{X} = [0,1]$, $\mathcal{Y} = \{0,1\}$, $\mathcal{H} = \{h_{\mathcal{I}}(x)$ for $\mathcal{I} = \{I_1, \dots, I_K\}$ for any $K \in \mathbb{N}\}$, for $h_{\mathcal{I}}(x) = [\![x \in \bigcup_{k=1}^{K} I_k]\!]$ with $I_i = [\theta_L^i, \theta_R^i]$
- $f(x) = h_{\mathcal{I}^*}(x)$ for some set of intervals $\mathcal{I}^*$
- training set $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$
- ERM rule: $h = \operatorname*{\mathbf{argmin}}_{\mathcal{I}} \frac{1}{m} \sum_{i=1}^{m} [\![h_{\mathcal{I}}(x_i) \neq y_i]\!]$,

  to make unique pick *smallest possible "$+1$" region*

Claim: ERM **does not** learn f^* (in the PAC sense).

Proof: textbook... (though obvious here: $h_{\text{ERM}} \equiv 0$ except in $x_1, \dots, x_m$)

There's No Free Lunch

Observation: ERM can learn all finite classes, but it fails on some infinite ones.

Is there a better algorithm than ERM, one that *always works*?

Observation: ERM can learn all finite classes, but it fails on some infinite ones.

Is there a better algorithm than ERM, one that *always works*?

No-Free-Lunch Theorem

- $\mathcal{X}$ input set, $\mathcal{Y} = \{0, 1\}$ label set, $\ell : \mathcal{Y} \times \mathcal{Y} \to \{0, 1\}$: 0/1-loss,
- A an arbitrary learning algorithm for binary classification,
- m (training size) any number smaller than $|\mathcal{X}|/2$

There exists

- a data distribution p over $\mathcal{X} \times \mathcal{Y}$, and
- a function $f : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$ with $\mathcal{R}(f) = 0$, but

$$\Pr_{\mathcal{D} \sim p^{\otimes m}} [\, \mathcal{R}(A[\mathcal{D}]) \geq 1/8 \,] \geq 1/7.$$

Summary: For every learning algorithm there exists a task on which it fails!

[David Wolpert. *"The Lack of A Priori Distinctions between Learning Algorithms"*, Neural Computation, 1996]

Agnostic PAC Learning

More realistic scenario: labeling isn't a deterministic function

- input set $\mathcal{X}$, label set $\mathcal{Y} = \{\pm 1\}$, data distribution $p(x, y)$
- ~~**deterministic** labels, $y = f(x)$ for unknown $f : \mathcal{X} \to \mathcal{Y}$~~
- loss function $\ell(y, y') = [\![y \neq y']\!]$
- $\mathcal{H} \subseteq \{h : \mathcal{X} \to \mathcal{Y}\}$: hypothesis set
- $\mathcal{D} = \{(x_1, y_1), \dots, (x_m, y_m)\} \overset{i.i.d.}{\sim} p(x, y)$: training set

Quantity of interest:

- $\mathcal{R}(h) = \mathop{\mathbb{E}}_{(x,y)\sim p(x,y)} \ell(y, h(x)) = \Pr_{(x,y)\sim p(x,y)} \{h(x) \neq y\}$

What can we learn?

- there might not be any $f : \mathcal{X} \to \mathcal{Y}$ that has $\mathcal{R}(f) = 0$.
- but: can we at least find the best h from the hypothesis set?

Definition (Agnostic PAC Learning)

A hypothesis class $\mathcal{H}$ is called **agnostic PAC learnable** by A, if

- for every $\epsilon > 0$ (accuracy → "approximate correct")
- and every $\delta > 0$ (confidence → "probably")

there exists an

- $m_0 = m_0(\epsilon, \delta) \in \mathbb{N}$ (minimal training set size)

such that

- for every probability distribution $p(x, y)$ over $\mathcal{X} \times \mathcal{Y}$,

when we run the learning algorithm A on a training set consisting of $m \geq m_0$ examples sampled i.i.d. from d, the algorithm returns a hypothesis $h \in \mathcal{H}$ that, with probability at least $1 - \delta$, fulfills

$$\mathcal{R}(h) \leq \min_{\bar{h} \in \mathcal{H}} \mathcal{R}(\bar{h}) + \epsilon.$$

$$\forall m \geq m_0(\epsilon, \delta) \quad \Pr_{\mathcal{D} \sim p^{\otimes m}} \left[\mathcal{R}(A[\mathcal{D}]) - \min_{h \in \mathcal{H}} \mathcal{R}(h) \ > \ \epsilon \right] \ \leq \ \delta.$$

Theorem (Agnostic PAC Learnability of finite hypothesis classes)

Let $\mathcal{H} = \{h_1, \ldots, h_K\}$ be a finite hypothesis class.

Then $\mathcal{H}$ is agnostic PAC-learnable by ERM with $m_0(\epsilon, \delta) = \lceil \frac{2}{\epsilon^2}(\log(|\mathcal{H}| + \log(2/\delta))\rceil$.

Proof. later

Theorem (Agnostic PAC Learnability of finite hypothesis classes)

Let $\mathcal{H} = \{h_1, \ldots, h_K\}$ *be a finite hypothesis class.*

Then $\mathcal{H}$ *is agnostic PAC-learnable by ERM with* $m_0(\epsilon, \delta) = \lceil \frac{2}{\epsilon^2}(\log(|\mathcal{H}| + \log(2/\delta))\rceil$.

Proof. later

Corollary

Let $\mathcal{D}$ *be a training set of size* m. *Let* f_{ERM} *be the result of running ERM on* $\mathcal{D}$. *Then*

$$\mathcal{R}(f_{\mathit{ERM}}) \leq \hat{\mathcal{R}}(f_{\mathit{ERM}}) + \sqrt{\frac{2(\log(|\mathcal{H}| + \log(2/\delta))}{m}} \tag{2}$$

Excurse: Concentration of Measure

Christoph Lampert

IST AUSTRIA

Institute of Science and Technology

Fall Semester 2020/2021
Lecture 7

- Z random variables, taking values $z \in \mathcal{Z} \subseteq \mathbb{R}$.
- $p(Z = z)$ probability distribution
  - $\mu = \mathbb{E}[\,Z\,]$ mean
  - $\mathrm{Var}[z] = \mathbb{E}[\,(Z - \mu)^2\,]$ variance

Lemma (Law of Large Numbers)

Let $Z_1, Z_2, \dots,$ be i.i.d. random variables with mean $\mathbb{E}[Z] < \infty$, then

$$\frac{1}{m}\sum_{i=1}^{m} Z_i \quad \xrightarrow{m\to\infty} \quad \mathbb{E}[Z] \qquad \textit{with probability } 1.$$

In machine learning, we have finite data, so $m \to \infty$ is less important.
Concentration of measure inequalities quantify the deviation between average and expectation for finite m.

Assumption: $\mathcal{Z} \subseteq \mathbb{R}_+$, i.e. Z takes only non-negative values.

Lemma (Markov's inequality)

$$\forall a > 0 : \quad \Pr[Z \geq a] \leq \frac{\mathbb{E}[Z]}{a}.$$

Proof. Step 1) We can write

$$\mathbb{E}[Z] = \int_{x=0}^{\infty} \Pr[Z \geq x] \, dx$$

Step 2) Since $\Pr[Z \geq x]$ is non-increasing in x, we have for any $a \geq 0$:

$$\mathbb{E}[Z] \geq \int_{x=0}^{a} \Pr[Z \geq x] \, \mathsf{dx} \geq \int_{x=0}^{a} \Pr[Z \geq a] \, \mathsf{dx} = a \Pr[Z \geq a]$$

Proof sketch of Step 1 inequality (ignoring questions of measurability and exchange of limit processes and writing the expression as if Z had a density $p(z)$)

$$\Pr[Z \geq x] = \int_{z=x}^{\infty} p(z)dz = \int_{z=0}^{\infty} [\![z \geq x]\!]\; p(z)\; dz$$

$$\begin{aligned}
\int_{x=0}^{\infty} \Pr[Z \geq x]\; dx &= \int_{x=0}^{\infty} \int_{z=0}^{\infty} [\![z \geq x]\!]\; p(z)dz\; dx \\
&= \int_{z=0}^{\infty} \int_{x=0}^{\infty} [\![z \geq x]\!]\; dx\; p(z)dz \\
&= \int_{z=0}^{\infty} \underbrace{\int_{x=0}^{z} dx}_{=z}\; p(z)dz \\
&= \int_{z=0}^{\infty} z\; p(z)dz \\
&= \mathbb{E}[Z]
\end{aligned}$$

Assumption: $\mathcal{Z} \subseteq \mathbb{R}_+$, i.e. Z takes only non-negative values.

Lemma (Markov's inequality)

$$\forall a \geq 0 : \quad \Pr[Z \geq a] \leq \frac{\mathbb{E}[Z]}{a}.$$

Corollary

$$\forall a \geq 0 : \quad \Pr[Z \geq a\,\mathbb{E}[Z]] \leq \frac{1}{a}.$$

Example

Is it possible that more than half of the population have a salary more than twice the mean salary?

Assumption: $\mathcal{Z} \subseteq \mathbb{R}_+$, i.e. Z takes only non-negative values.

Lemma (Markov's inequality)

$$\forall a \geq 0 : \quad \Pr[Z \geq a] \leq \frac{\mathbb{E}[Z]}{a}.$$

Corollary

$$\forall a \geq 0 : \quad \Pr[Z \geq a\,\mathbb{E}[Z]] \leq \frac{1}{a}.$$

Example

Is it possible that more than half of the population have a salary more than twice the mean salary? No, by corrolary with $a = 2$.

Assumption: $\mathcal{Z} \subseteq \mathbb{R}_+$, i.e. Z takes only non-negative values.

Lemma (Markov's inequality)

$$\forall a \geq 0 : \quad \Pr[Z \geq a] \leq \frac{\mathbb{E}[Z]}{a}.$$

Corollary

$$\forall a \geq 0 : \quad \Pr[Z \geq a\,\mathbb{E}[Z]] \leq \frac{1}{a}.$$

Example

Is it possible that more than half of the population have a salary more than twice the mean salary? No, by corrolary with $a = 2$.

Example

Is it possible that more than 90% of the population have a salary less than one tenth of the mean?

Assumption: $\mathcal{Z} \subseteq \mathbb{R}_+$, i.e. Z takes only non-negative values.

Lemma (Markov's inequality)

$$\forall a \geq 0: \quad \Pr[Z \geq a] \leq \frac{\mathbb{E}[Z]}{a}.$$

Corollary

$$\forall a \geq 0: \quad \Pr[Z \geq a\,\mathbb{E}[Z]] \leq \frac{1}{a}.$$

Example

Is it possible that more than half of the population have a salary more than twice the mean salary? No, by corrolary with $a = 2$.

Example

Is it possible that more than 90% of the population have a salary less than one tenth of the mean? Easily: $p(\$1) = 0.99$, $p(\$100000) = 0.01$.

Lemma (Chebyshev's inequality)

$$\forall a \geq 0: \quad \Pr[|Z - \mathbb{E}[Z]| \geq a] \leq \frac{\mathit{Var}[Z]}{a^2}$$

Proof. Apply Markov's Inequality to the random variable $(Z - \mathbb{E}[Z])^2$.

Lemma (Chebyshev's inequality)

$$\forall a \geq 0 : \quad \Pr[|Z - \mathbb{E}[Z]| \geq a] \leq \frac{\mathit{Var}[Z]}{a^2}$$

Proof. Apply Markov's Inequality to the random variable $(Z - \mathbb{E}[Z])^2$.

For any $a \geq 0$:

$$\Pr[\,|Z - \mathbb{E}[Z]| \geq a] = \Pr[(Z - \mathbb{E}[Z])^2 \geq a^2] \overset{\text{Markov}}{\leq} \frac{\mathbb{E}[\,(Z - \mathbb{E}[Z])^2\,]}{a^2} = \frac{\mathsf{Var}[Z]}{a^2}.$$

Lemma (Chebyshev's inequality)

$$\forall a \geq 0 : \quad \Pr[|Z - \mathbb{E}[Z]| \geq a] \leq \frac{\mathit{Var}[Z]}{a^2}$$

Proof. Apply Markov's Inequality to the random variable $(Z - \mathbb{E}[Z])^2$.

For any $a \geq 0$:

$$\Pr[\,|Z - \mathbb{E}[Z]| \geq a] = \Pr[(Z - \mathbb{E}[Z])^2 \geq a^2] \overset{\text{Markov}}{\leq} \frac{\mathbb{E}[\,(Z - \mathbb{E}[Z])^2\,]}{a^2} = \frac{\mathrm{Var}[Z]}{a^2}.$$

Remark: Chebyshev ineq. has similar role as "σ-rules" for Gaussians:

- 68% of probability mass of a Gaussian lie within $\mu \pm \sigma$,
- 95% of probability mass of a Gaussian lie within $\mu \pm 2\sigma$,
- 99.7% of probability mass of a Gaussian lie within $\mu \pm 3\sigma$,

Chebyshev holds for arbitrary probability distributions, not just Gaussians.

Example (Soccer Match Statistics)

- $z = -1$ for loss, $z = 0$ for draw, $z = 1$ for win.
- $p(-1) = \frac{1}{10}$, $p(1) = \frac{1}{10}$, $p(0) = \frac{4}{5}$.
- $\mathbb{E}[Z] = 0$.
- Var$[Z] = \mathbb{E}[\,(Z)^2] = \frac{1}{10}(-1)^2 + \frac{4}{5}0^2 + \frac{1}{10}(1)^2 = \frac{1}{5}$

What if we pretended Z is Gaussian?

- $\mu = 0$, $\sigma = \sqrt{\frac{1}{5}} \approx 0.45$,
- we expect $\leq 5\%$ prob.mass outside of the 2σ-interval $[-0.9, 0.9]$
- but really, its 20%!

With Chebyshev:

- $\Pr[|Z| \geq 0.9] \leq \frac{1}{5}/(0.9)^2 \approx 0.247$, so bound is correct

Lemma (Quantitative Version of the Law of Large Numbers)

Set $Z_1, \ldots, Z_m$ be i.i.d. random variables with $\mathbb{E}[Z_i] = \mu$ and Var$[Z_i] \leq C$. Then, for any $\delta \in (0,1)$, the following inequality holds with probability at least $1-\delta$:

$$\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\right| < \sqrt{\frac{C}{\delta m}}.$$

Equivalent formulations:

$$\Pr\left[\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\right| < \sqrt{\frac{C}{\delta m}}\right] \geq 1-\delta.$$

$$\Pr\left[\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\right| \geq \sqrt{\frac{C}{\delta m}}\right] \leq \delta.$$

Lemma (Quantitative Version of the Law of Large Numbers)

Set $Z_1, \ldots, Z_m$ be i.i.d. RVs with $\mathbb{E}[Z_i] = \mu$ and Var$[Z_i] \leq C$. Then, for any $\delta \in (0,1)$,

$$\Pr\left[\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\right| \geq \sqrt{\frac{C}{\delta m}}\right] \leq \delta.$$

Lemma (Quantitative Version of the Law of Large Numbers)

Set $Z_1, \ldots, Z_m$ be i.i.d. RVs with $\mathbb{E}[Z_i] = \mu$ and Var$[Z_i] \leq C$. Then, for any $\delta \in (0,1)$,

$$\Pr\left[\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\right| \geq \sqrt{\frac{C}{\delta m}}\right] \leq \delta.$$

Proof. The Z_i are indep., so $\mathrm{Var}\left[\frac{1}{m}\sum_{i=1}^m Z_i\right] = \frac{1}{m^2}\sum_{i=1}^m \mathrm{Var}[Z_i] \leq \frac{C}{m}$.

2) Chebyshev's inequality gives us for any $a \geq 0$:

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\right| \geq a\right] \leq \frac{\mathrm{Var}\left[\frac{1}{m}\sum_{i=1}^m Z_i\right]}{a^2} \leq \frac{C}{ma^2}.$$

Setting $\delta = \frac{C}{ma^2}$ and solving for a yields $a = \sqrt{\frac{C}{\delta m}}$.

Sanity check: How large should my test set be?

Setup: fixed classifier $g : \mathcal{X} \to \mathcal{Y}$, 0/1-loss: $\ell(\bar{y}, y) = [\![\bar{y} \neq y]\!]$

- test set $\mathcal{D} = \{(x^1, y^1) \dots, (x^m, y^m)\} \overset{i.i.d.}{\sim} p(x, y)$,
- random variables $Z_i = [\![g(x^i) \neq y^i]\!] \in \{0, 1\}$,
- $\mathbb{E}[Z^i] = \mathbb{E}\{[\![g(x^i) \neq y^i]\!]\} = \mu$ (generalization error of g)
- $\text{Var}[Z^i] = \mathbb{E}\{(Z^i - \mu)^2\} = \mu(1-\mu)^2 + (1-\mu)\mu^2 = \mu(1-\mu) \leq \frac{1}{4} =: C$

Setup: fixed confidence, e.g. $\delta = 0.1$, $\sqrt{\frac{C}{\delta m}} = \sqrt{\frac{0.25}{0.1 m}} = \sqrt{\frac{2.5}{m}}$

$$\mathbb{P}\left[\left| \frac{1}{m} \sum_{i=1}^{m} Z_i - \mu \right| \leq \sqrt{\frac{2.5}{m}} \right] \geq 0.9$$

To be 90%-certain that the error is within ± 0.05, use $m \geq 1,000$.

Sanity check: How large should my test set be?

Setup: fixed classifier $g : \mathcal{X} \to \mathcal{Y}$, $0/1$-loss: $\ell(\bar{y}, y) = [\![\bar{y} \neq y]\!]$

- test set $\mathcal{D} = \{(x^1, y^1) \dots, (x^m, y^m)\} \overset{i.i.d.}{\sim} p(x, y)$,
- random variables $Z_i = [\![g(x^i) \neq y^i]\!] \in \{0, 1\}$,
- $\mathbb{E}[Z^i] = \mathbb{E}\{[\![g(x^i) \neq y^i]\!]\} = \mu$ (generalization error of g)
- $\text{Var}[Z^i] = \mathbb{E}\{(Z^i - \mu)^2\} = \mu(1-\mu)^2 + (1-\mu)\mu^2 = \mu(1-\mu) \leq \frac{1}{4} =: C$

Setup: fixed confidence, e.g. $\delta = 0.1$, $\sqrt{\frac{C}{\delta m}} = \sqrt{\frac{0.25}{0.1m}} = \sqrt{\frac{2.5}{m}}$

$$\mathbb{P}\left[\left| \frac{1}{m} \sum_{i=1}^{m} Z_i - \mu \right| \leq \sqrt{\frac{2.5}{m}} \right] \geq 0.9$$

To be 90%-certain that the error is within ± 0.05, use $m \geq 1,000$.
10× more certain: to be 99%-certain that the error is within ± 0.05, use $m \geq 10,000$.

Sanity check: How large should my test set be?

Setup: fixed classifier $g : \mathcal{X} \to \mathcal{Y}$, 0/1-loss: $\ell(\bar{y}, y) = [\![\bar{y} \neq y]\!]$

- test set $\mathcal{D} = \{(x^1, y^1) \dots, (x^m, y^m)\} \overset{i.i.d.}{\sim} p(x, y)$,
- random variables $Z_i = [\![g(x^i) \neq y^i]\!] \in \{0, 1\}$,
- $\mathbb{E}[Z^i] = \mathbb{E}\{[\![g(x^i) \neq y^i]\!]\} = \mu$ (generalization error of g)
- $\mathrm{Var}[Z^i] = \mathbb{E}\{(Z^i - \mu)^2\} = \mu(1-\mu)^2 + (1-\mu)\mu^2 = \mu(1-\mu) \leq \frac{1}{4} =: C$

Setup: fixed confidence, e.g. $\delta = 0.1$, $\sqrt{\frac{C}{\delta m}} = \sqrt{\frac{0.25}{0.1m}} = \sqrt{\frac{2.5}{m}}$

$$\mathbb{P}\left[\Big| \frac{1}{m} \sum_{i=1}^{m} Z_i - \mu \Big| \leq \sqrt{\frac{2.5}{m}} \right] \geq 0.9$$

To be 90%-certain that the error is within ± 0.05, use $m \geq 1,000$.
10× more certain: to be 99%-certain that the error is within ± 0.05, use $m \geq 10,000$.
10× more accuracy: to be 90%-certain that the error is within ± 0.005, use $m \geq 100,000$.

Sanity check: How large should my test set be?

Setup: fixed classifier $g : \mathcal{X} \to \mathcal{Y}$, 0/1-loss: $\ell(\bar{y}, y) = [\![\bar{y} \neq y]\!]$

- test set $\mathcal{D} = \{(x^1, y^1) \dots, (x^m, y^m)\} \overset{i.i.d.}{\sim} p(x, y)$,
- random variables $Z_i = [\![g(x^i) \neq y^i]\!] \in \{0, 1\}$,
- $\mathbb{E}[Z^i] = \mathbb{E}\{[\![g(x^i) \neq y^i]\!]\} = \mu$ (generalization error of g)
- $\mathrm{Var}[Z^i] = \mathbb{E}\{(Z^i - \mu)^2\} = \mu(1-\mu)^2 + (1-\mu)\mu^2 = \mu(1-\mu) \leq \frac{1}{4} =: C$

Setup: fixed confidence, e.g. $\delta = 0.1$, $\sqrt{\frac{C}{\delta m}} = \sqrt{\frac{0.25}{0.1 m}} = \sqrt{\frac{2.5}{m}}$

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\right| \leq \sqrt{\frac{2.5}{m}}\right] \geq 0.9$$

To be 90%-certain that the error is within ± 0.05, use $m \geq 1,000$.
10× more certain: to be 99%-certain that the error is within ± 0.05, use $m \geq 10,000$.
10× more accuracy: to be 90%-certain that the error is within ± 0.005, use $m \geq 100,000$.

... admittedly not very impressive. Luckily, a bit tighter bounds are coming up next.

Lemma (Hoeffding's Lemma)

Let Z be a random variable that takes values in $[a, b]$ and $\mathbb{E}[Z] = 0$. Then, for every $\lambda > 0$,

$$\mathbb{E}[e^{\lambda X}] \leq e^{\frac{\lambda^2 (b-a)^2}{8}}.$$

Proof: Exercise...

Lemma (Hoeffding's Inequality)

Let $Z_1, \ldots, Z_m$ be i.i.d. random variables that take values in the interval $[a, b]$. Let $\bar{Z} = \frac{1}{m}\sum_{i=1}^{m} Z_i$ and denote $\mathbb{E}[\bar{Z}] = \mu$. Then, for any $\epsilon > 0$,

$$\mathbb{P}\left[\left(\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\right) > \epsilon\right] \leq e^{-m\frac{\epsilon^2}{(b-a)^2}}.$$

and

$$\mathbb{P}\left[\left(\mu - \frac{1}{m}\sum_{i=1}^{m} Z_i\right) > \epsilon\right] \leq e^{-m\frac{\epsilon^2}{(b-a)^2}}.$$

and

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\right| > \epsilon\right] \leq 2e^{-m\frac{\epsilon^2}{(b-a)^2}}.$$

Hoeffding's Inequality – Proof

Define new RVs: $X_i = Z_i - \mathbb{E}[Z_i]$, $\bar{X} = \frac{1}{m}\sum_i X_i$

- $\mathbb{E}[X_i] = 0$; $\mathbb{E}[\bar{X}] = 0$; each X_i takes values in $[a - \mathbb{E}[Z_i], b - \mathbb{E}[Z_i]]$

Use 1) monotonicity of $\exp$ and 2) Markov's inequality to check

$$\mathbb{P}[\bar{X} \geq \epsilon] \overset{1)}{=} \mathbb{P}[e^{\lambda\bar{X}} \geq e^{\lambda\epsilon}] \overset{2)}{\leq} e^{-\lambda\epsilon}\,\mathbb{E}[e^{\lambda\bar{X}}]$$

From 3) the independence of the X_i we have

$$\mathbb{E}[e^{\lambda\bar{X}}] = \mathbb{E}[\prod_{i=1}^{n} e^{\lambda X_i/m}] \overset{3)}{=} \prod_{i=1}^{n}\mathbb{E}[e^{\lambda X_i/m}]$$

Use 4) Hoeffding's Lemma for every i:

$$\mathbb{E}[e^{\lambda X_i/m}] \overset{4)}{\leq} e^{\frac{\lambda^2(b-a)^2}{8m^2}}.$$

In combination:

$$\mathbb{P}[\bar{X} \geq \epsilon] \leq e^{-\lambda\epsilon}e^{\frac{\lambda^2(b-a)^2}{8m}}$$

Previous step:

$$\mathbb{P}[\bar{X} \geq \epsilon] \;\leq\; e^{-\lambda\epsilon} e^{\frac{\lambda^2 (b-a)^2}{8m}}$$

So far, λ was arbitrary. Now we set $\lambda = \frac{4m\epsilon}{(b-a)^2}$

$$\mathbb{P}[\bar{X} \geq \epsilon] \;\leq\; e^{-\frac{4m\epsilon}{(b-a)^2}\epsilon + \left(\frac{4m\epsilon}{(b-a)^2}\right)^2 \frac{(b-a)^2}{8m}} = e^{-\frac{2m\epsilon^2}{(b-a)^2}}$$

This proves the first statement.

If we repeat the same steps again for $-\bar{X}$ instead of X, we get

$$\mathbb{P}[\bar{X} \leq -\epsilon] \;\leq\; e^{-\frac{2m\epsilon^2}{(b-a)^2}}$$

This proves the second statement.

Use the *union bound*: $\mathbb{P}[A \vee B] \leq \mathbb{P}[A] + \mathbb{P}[B]$, to combine both directions:

$$\mathbb{P}[|\bar{X}| \geq \epsilon] \;=\; \mathbb{P}[\;(\bar{X} \geq \epsilon) \vee (\bar{X} \leq -\epsilon)\;] \;\leq\; 2e^{-\frac{2m\epsilon^2}{(b-a)^2}}.$$

$\square$

How large should my test set be?

$$\mathbb{P}\Big[\Big|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\Big| > \epsilon\Big] \leq 2e^{-\frac{2m\epsilon^2}{(b-a)^2}}.$$

Setup: fixed classifier $g : \mathcal{X} \to \mathcal{Y}$

- test set $\mathcal{D} = \{(x^1, y^1) \dots, (x^m, y^m)\} \overset{i.i.d.}{\sim} p(x, y)$,
- random variables $Z_i = [\![g(x^i) \neq y^i]\!] \in \{0, 1\}$, $\to \quad b - a = 1$
- $\mathbb{E}[Z_i] = \mathbb{E}\{[\![g(x^i) \neq y^i]\!]\} = \mu \quad$ (test error of g)

Setup: $m = \frac{1}{2}\log(\frac{2}{\delta})/\epsilon^2$.

For fixed confidence $\delta = 0.1 \Rightarrow \epsilon = \sqrt{\log(20)/(2m)} \approx 1.22\sqrt{\frac{1}{m}}$

$$\mathbb{P}\left[\Big|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\Big| \leq 1.22\sqrt{\frac{1}{m}}\right] \geq 0.9$$

How large should my test set be?

$$\mathbb{P}\Big[\,\Big|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\Big| > \epsilon\,\Big] \leq 2e^{-\frac{2m\epsilon^2}{(b-a)^2}}.$$

Setup: fixed classifier $g : \mathcal{X} \to \mathcal{Y}$

- test set $\mathcal{D} = \{(x^1, y^1)\dots,(x^m, y^m)\} \overset{i.i.d.}{\sim} p(x, y)$,
- random variables $Z_i = [\![g(x^i) \neq y^i]\!] \in \{0, 1\}$, $\to \quad b - a = 1$
- $\mathbb{E}[Z_i] = \mathbb{E}\{[\![g(x^i) \neq y^i]\!]\} = \mu \qquad$ (test error of g)

Setup: $m = \frac{1}{2}\log(\frac{2}{\delta})/\epsilon^2$.

For fixed confidence $\delta = 0.1 \Rightarrow \epsilon = \sqrt{\log(20)/(2m)} \approx 1.22\sqrt{\frac{1}{m}}$

$$\mathbb{P}\Big[\Big|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\Big| \leq 1.22\sqrt{\frac{1}{m}}\Big] \geq 0.9$$

To be 90%-certain that the error is within ± 0.05, use $m \geq 600$.

How large should my test set be?

$$\mathbb{P}\left[\,\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\right| > \epsilon\,\right] \leq 2e^{-\frac{2m\epsilon^2}{(b-a)^2}}.$$

Setup: fixed classifier $g : \mathcal{X} \to \mathcal{Y}$

- test set $\mathcal{D} = \{(x^1, y^1)\ldots,(x^m, y^m)\} \overset{i.i.d.}{\sim} p(x, y)$,
- random variables $Z_i = [\![g(x^i) \neq y^i]\!] \in \{0, 1\}$, $\to \quad b - a = 1$
- $\mathbb{E}[Z_i] = \mathbb{E}\{[\![g(x^i) \neq y^i]\!]\} = \mu \qquad$ (test error of g)

Setup: $m = \frac{1}{2}\log(\frac{2}{\delta})/\epsilon^2$.

For fixed confidence $\delta = 0.1 \Rightarrow \epsilon = \sqrt{\log(20)/(2m)} \approx 1.22\sqrt{\frac{1}{m}}$

$$\mathbb{P}\left[\,\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\right| \;\leq 1.22\sqrt{\frac{1}{m}}\right] \geq 0.9$$

To be 90%-certain that the error is within ± 0.05, use $m \geq 600$.

$10\times$ more certain: to be 99%-certain that the error is within ± 0.05, use $m \geq 1,060$.

How large should my test set be?

$$\mathbb{P}\Big[\,\Big|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\Big| > \epsilon\,\Big] \leq 2e^{-\frac{2m\epsilon^2}{(b-a)^2}}.$$

Setup: fixed classifier $g : \mathcal{X} \to \mathcal{Y}$

- test set $\mathcal{D} = \{(x^1, y^1) \ldots, (x^m, y^m)\} \overset{i.i.d.}{\sim} p(x, y)$,
- random variables $Z_i = [\![g(x^i) \neq y^i]\!] \in \{0, 1\}$, $\rightarrow \quad b - a = 1$
- $\mathbb{E}[Z_i] = \mathbb{E}\{[\![g(x^i) \neq y^i]\!]\} = \mu \qquad$ (test error of g)

Setup: $m = \frac{1}{2}\log(\frac{2}{\delta})/\epsilon^2$.

For fixed confidence $\delta = 0.1 \Rightarrow \epsilon = \sqrt{\log(20)/(2m)} \approx 1.22\sqrt{\frac{1}{m}}$

$$\mathbb{P}\Big[\,\Big|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\Big| \leq 1.22\sqrt{\frac{1}{m}}\,\Big] \geq 0.9$$

To be 90%-certain that the error is within ± 0.05, use $m \geq 600$.

$10\times$ more certain: to be 99%-certain that the error is within ± 0.05, use $m \geq 1,060$.

$10\times$ more accuracy: to be 90%-certain that the error is within ± 0.005, use $m \geq 59,914$.

Difference: Chebyshev's vs. Hoeffding's Inequality

With $\hat{\mathcal{R}} = \frac{1}{m}\sum_{i=1}^{m} Z_i$ and $\mathcal{R} = \mathbb{E}[\frac{1}{m}\sum_{i=1}^{m} Z_i]$:

- Chebyshev's: $\mathrm{Var}[Z_i] \le C$

$$\mathbb{P}\left[|\hat{\mathcal{R}} - \mathcal{R}| > \sqrt{\frac{C}{\delta m}}\right] \le \delta, \qquad \mathbb{P}\left[|\hat{\mathcal{R}} - \mathcal{R}| > \epsilon\right] \le \frac{C}{\epsilon^2 m}$$

- interval decreases like $\frac{1}{\sqrt{m}}$, confidence grows like $1 - \frac{1}{m}$
- Hoeffding's: Z_i takes values in $[a, b]$:

$$\mathbb{P}\left[|\hat{\mathcal{R}} - \mathcal{R}| > \sqrt{\frac{(b-a)^2 \log\frac{2}{\delta}}{m}}\right] \le \delta, \quad \mathbb{P}\left[|\hat{\mathcal{R}} - \mathcal{R}| > \epsilon\right] \le 2e^{-\frac{2m\epsilon^2}{(b-a)^2}}.$$

- interval decreases like $\frac{1}{\sqrt{m}}$, confidence grows like $1 - e^{-m}$

Both are typical **PAC (probably approximately correct)** statements:
"With **prob.** $1-\delta$, the estimated $\hat{\mathcal{R}}$ is an ϵ-**close approximation** of $\mathcal{R}$."

Back to PAC Learning

Christoph Lampert

Institute of Science and Technology AUSTRIA

Fall Semester 2020/2021
Lecture 7

Theorem (Finite hypothesis classes are agnostic PAC learnable)

Let $\mathcal{H} = \{h_1, \dots, h_K\}$ *be a finite hypothesis class.*

For any $\delta > 0$ *and* $\epsilon > 0$ *let* $m_0(\epsilon, \delta) = \lceil \frac{2}{\epsilon^2}(\log(|\mathcal{H}| + \log(2/\delta))\rceil$*. For any* $m \geq m_0$*, let* $\mathcal{D}_m = \{(x_1, y_1), \dots, (x_m, y_m)\} \overset{i.i.d.}{\sim} p(x, y)$ *be a training set and let* $f_{\textsf{ERM}}$ *be the result of running an ERM algorithm on* $\mathcal{D}$*. Then, it holds with probability at least* $1 - \delta$ *over the sampled* $\mathcal{D}$ *that*

$$\mathcal{R}(f_{\textsf{ERM}}) \leq \min_{h \in \mathcal{H}} \mathcal{R}(h) + \epsilon$$

Theorem (Finite hypothesis classes are agnostic PAC learnable)

Let $\mathcal{H} = \{h_1, \ldots, h_K\}$ *be a finite hypothesis class.*

For any $\delta > 0$ *and* $\epsilon > 0$ *let* $m_0(\epsilon, \delta) = \lceil \frac{2}{\epsilon^2}(\log(|\mathcal{H}| + \log(2/\delta))\rceil$. *For any* $m \geq m_0$, *let* $\mathcal{D}_m = \{(x_1, y_1), \ldots, (x_m, y_m)\} \overset{i.i.d.}{\sim} p(x, y)$ *be a training set and let* f_{ERM} *be the result of running an ERM algorithm on* $\mathcal{D}$. *Then, it holds with probability at least* $1 - \delta$ *over the sampled* $\mathcal{D}$ *that*

$$\mathcal{R}(f_{\mathsf{ERM}}) \leq \min_{h \in \mathcal{H}} \mathcal{R}(h) + \epsilon$$

Proof strategy.

- Step 1: show that $\mathcal{R}(h)$ and $\hat{\mathcal{R}}_m(h)$ close together with high probability uniformly in h:
- Step 1: apply this result specifically to f_{ERM} and $\mathbf{argmin}_{h \in \mathcal{H}} \mathcal{R}(h)$

Lemma

For any $\epsilon > 0$, $\delta > 0$, the following holds with probability at least $1 - \delta$ w.r.t. $\mathcal{D}_m$:

$$\forall h \in \mathcal{H} \qquad |\mathcal{R}(h) - \hat{\mathcal{R}}_m(h)| \leq \sqrt{\frac{\log|\mathcal{H}| + \log\frac{2}{\delta}}{2m}}$$

Proof:

1. For any individual $h \in \mathcal{H}$, we get from Hoeffding's inequality:
$$\mathbb{P}[\underbrace{\left|\mathcal{R}(h) - \hat{\mathcal{R}}_m(h)\right| > \epsilon}_{\text{call this event "}C_h\text{"}}] \leq 2e^{-2m\epsilon^2}.$$

Lemma

For any $\epsilon > 0$, $\delta > 0$, the following holds with probability at least $1-\delta$ w.r.t. $\mathcal{D}_m$:

$$\forall h \in \mathcal{H} \qquad |\mathcal{R}(h) - \hat{\mathcal{R}}_m(h)| \leq \sqrt{\frac{\log|\mathcal{H}| + \log\frac{2}{\delta}}{2m}}$$

Proof:

1. For any individual $h \in \mathcal{H}$, we get from Hoeffding's inequality:
$$\mathbb{P}[\underbrace{\big|\mathcal{R}(h) - \hat{\mathcal{R}}_m(h)\big| > \epsilon}_{\text{call this event "}C_h\text{"}}] \leq 2e^{-2m\epsilon^2}.$$

2. From a union bound, $\Pr\{\bigvee_{h\in\mathcal{H}} C_h\} \leq \sum_{h\in\mathcal{H}} \Pr\{C_h\}$, we obtain
$$\mathbb{P}[\exists h \in \mathcal{H} : |\mathcal{R}(h) - \hat{\mathcal{R}}_m(h)| > \epsilon] \leq |\mathcal{H}| 2e^{-2m\epsilon^2}.$$

Lemma

For any $\epsilon > 0$, $\delta > 0$, the following holds with probability at least $1-\delta$ w.r.t. $\mathcal{D}_m$:

$$\forall h \in \mathcal{H} \qquad |\mathcal{R}(h) - \hat{\mathcal{R}}_m(h)| \le \sqrt{\frac{\log|\mathcal{H}| + \log\frac{2}{\delta}}{2m}}$$

Proof:

1. For any individual $h \in \mathcal{H}$, we get from Hoeffding's inequality:
$$\mathbb{P}[\,\underbrace{\left|\mathcal{R}(h) - \hat{\mathcal{R}}_m(h)\right| > \epsilon}_{\text{call this event "}C_h\text{"}}\,] \le 2e^{-2m\epsilon^2}.$$

2. From a union bound, $\Pr\{\bigvee_{h\in\mathcal{H}} C_h\} \le \sum_{h\in\mathcal{H}} \Pr\{C_h\}$, we obtain
$$\mathbb{P}[\,\exists h \in \mathcal{H} : |\mathcal{R}(h) - \hat{\mathcal{R}}_m(h)| > \epsilon\,] \le |\mathcal{H}|2e^{-2m\epsilon^2}.$$

3. Setting the right hand side to be δ, we solve for ϵ, obtaining $\epsilon = \sqrt{\frac{\log|\mathcal{H}|+\log\frac{2}{\delta}}{2m}}$.

Lemma

For any $\epsilon > 0$, $\delta > 0$, the following holds with probability at least $1-\delta$ w.r.t. $\mathcal{D}_m$:

$$\forall h \in \mathcal{H} \qquad |\mathcal{R}(h) - \hat{\mathcal{R}}_m(h)| \leq \sqrt{\frac{\log|\mathcal{H}| + \log\frac{2}{\delta}}{2m}}$$

Proof:

1. For any individual $h \in \mathcal{H}$, we get from Hoeffding's inequality:
$$\mathbb{P}[\,\underbrace{\big|\mathcal{R}(h) - \hat{\mathcal{R}}_m(h)\big| > \epsilon}_{\text{call this event "}C_h\text{"}}\,] \leq 2e^{-2m\epsilon^2}.$$

2. From a union bound, $\Pr\{\bigvee_{h\in\mathcal{H}} C_h\} \leq \sum_{h\in\mathcal{H}} \Pr\{C_h\}$, we obtain
$$\mathbb{P}[\,\exists h \in \mathcal{H} : |\mathcal{R}(h) - \hat{\mathcal{R}}_m(h)| > \epsilon\,] \leq |\mathcal{H}|2e^{-2m\epsilon^2}.$$

3. Setting the right hand side to be δ, we solve for ϵ, obtaining $\epsilon = \sqrt{\frac{\log|\mathcal{H}|+\log\frac{2}{\delta}}{2m}}$.

4. The statement of the lemma follows, because
$$\Pr\left\{\forall h \in \mathcal{H} : |\mathcal{R}(h) - \hat{\mathcal{R}}(h)| \leq \epsilon\right\} = 1 - \Pr\left\{\exists h \in \mathcal{H} : |\mathcal{R}(h) - \hat{\mathcal{R}}(h)| \leq \epsilon\right\}$$

Lemma

For any $\epsilon > 0$, $\delta > 0$, the following holds with probability at least $1 - \delta$ w.r.t. $\mathcal{D}_m$:

$$\forall h \in \mathcal{H} \qquad |\mathcal{R}(h) - \hat{\mathcal{R}}_m(h)| \leq \sqrt{\frac{\log |\mathcal{H}| + \log \frac{2}{\delta}}{2m}} =: \alpha$$

Step 2: we use the lemma to bound the difference between

- $h_{\text{ERM}} \in \mathbf{argmin}_{\bar{h} \in \mathcal{H}} \hat{\mathcal{R}}_m(\bar{h})$ (result of ERM)
- $h^* \in \mathbf{argmin}_{\bar{h} \in \mathcal{H}} \mathcal{R}(\bar{h})$ (if exists, otherwise argue with arbitrarily close approximation)

$$\mathcal{R}(h_{\text{ERM}}) - \mathcal{R}(h^*) =$$

Lemma

For any $\epsilon > 0$, $\delta > 0$, the following holds with probability at least $1 - \delta$ w.r.t. $\mathcal{D}_m$:

$$\forall h \in \mathcal{H} \qquad |\mathcal{R}(h) - \hat{\mathcal{R}}_m(h)| \leq \sqrt{\frac{\log|\mathcal{H}| + \log\frac{2}{\delta}}{2m}} =: \alpha$$

Step 2: we use the lemma to bound the difference between

- $h_{\text{ERM}} \in \mathbf{argmin}_{\bar{h}\in\mathcal{H}} \hat{\mathcal{R}}_m(\bar{h})$ (result of ERM)
- $h^* \in \mathbf{argmin}_{\bar{h}\in\mathcal{H}} \mathcal{R}(\bar{h})$ (if exists, otherwise argue with arbitrarily close approximation)

$$\mathcal{R}(h_{\text{ERM}}) - \mathcal{R}(h^*) = \mathcal{R}(h_{\text{ERM}}) - \hat{\mathcal{R}}(h_{\text{ERM}}) + \hat{\mathcal{R}}(h_{\text{ERM}}) - \hat{\mathcal{R}}(h^*) + \hat{\mathcal{R}}(h^*) - \mathcal{R}(h^*)$$

Lemma

For any $\epsilon > 0$, $\delta > 0$, the following holds with probability at least $1-\delta$ w.r.t. $\mathcal{D}_m$:

$$\forall h \in \mathcal{H} \qquad |\mathcal{R}(h) - \hat{\mathcal{R}}_m(h)| \leq \sqrt{\frac{\log|\mathcal{H}| + \log\frac{2}{\delta}}{2m}} =: \alpha$$

Step 2: we use the lemma to bound the difference between

- $h_{\textsf{ERM}} \in \mathbf{argmin}_{\bar{h}\in\mathcal{H}}\, \hat{\mathcal{R}}_m(\bar{h})$ (result of ERM)
- $h^* \in \mathbf{argmin}_{\bar{h}\in\mathcal{H}}\, \mathcal{R}(\bar{h})$ (if exists, otherwise argue with arbitrarily close approximation)

$$\mathcal{R}(h_{\textsf{ERM}}) - \mathcal{R}(h^*) = \underbrace{\mathcal{R}(h_{\textsf{ERM}}) - \hat{\mathcal{R}}(h_{\textsf{ERM}})}_{\leq\alpha} + \hat{\mathcal{R}}(h_{\textsf{ERM}}) - \hat{\mathcal{R}}(h^*) + \underbrace{\hat{\mathcal{R}}(h^*) - \mathcal{R}(h^*)}_{\leq\alpha}$$

Lemma

For any $\epsilon > 0$, $\delta > 0$, *the following holds with probability at least* $1 - \delta$ *w.r.t.* $\mathcal{D}_m$*:*

$$\forall h \in \mathcal{H} \qquad |\mathcal{R}(h) - \hat{\mathcal{R}}_m(h)| \leq \sqrt{\frac{\log|\mathcal{H}| + \log\frac{2}{\delta}}{2m}} =: \alpha$$

Step 2: we use the lemma to bound the difference between

- $h_{\mathsf{ERM}} \in \mathbf{argmin}_{\bar{h} \in \mathcal{H}} \hat{\mathcal{R}}_m(\bar{h})$ (result of ERM)
- $h^* \in \mathbf{argmin}_{\bar{h} \in \mathcal{H}} \mathcal{R}(\bar{h})$ (if exists, otherwise argue with arbitrarily close approximation)

$$\mathcal{R}(h_{\mathsf{ERM}}) - \mathcal{R}(h^*) = \underbrace{\mathcal{R}(h_{\mathsf{ERM}}) - \hat{\mathcal{R}}(h_{\mathsf{ERM}})}_{\leq \alpha} + \hat{\mathcal{R}}(h_{\mathsf{ERM}}) - \hat{\mathcal{R}}(h^*) + \underbrace{\hat{\mathcal{R}}(h^*) - \mathcal{R}(h^*)}_{\leq \alpha}$$

$$\leq \hat{\mathcal{R}}(h_{\mathsf{ERM}}) - \hat{\mathcal{R}}(h^*) + 2\alpha$$

Lemma

For any $\epsilon > 0$, $\delta > 0$, the following holds with probability at least $1 - \delta$ w.r.t. $\mathcal{D}_m$:

$$\forall h \in \mathcal{H} \qquad |\mathcal{R}(h) - \hat{\mathcal{R}}_m(h)| \leq \sqrt{\frac{\log|\mathcal{H}| + \log\frac{2}{\delta}}{2m}} =: \alpha$$

Step 2: we use the lemma to bound the difference between

- $h_{\text{ERM}} \in \mathbf{argmin}_{\bar{h} \in \mathcal{H}} \hat{\mathcal{R}}_m(\bar{h})$ (result of ERM)
- $h^* \in \mathbf{argmin}_{\bar{h} \in \mathcal{H}} \mathcal{R}(\bar{h})$ (if exists, otherwise argue with arbitrarily close approximation)

$$\mathcal{R}(h_{\text{ERM}}) - \mathcal{R}(h^*) = \underbrace{\mathcal{R}(h_{\text{ERM}}) - \hat{\mathcal{R}}(h_{\text{ERM}})}_{\leq \alpha} + \hat{\mathcal{R}}(h_{\text{ERM}}) - \hat{\mathcal{R}}(h^*) + \underbrace{\hat{\mathcal{R}}(h^*) - \mathcal{R}(h^*)}_{\leq \alpha}$$

$$\leq \underbrace{\hat{\mathcal{R}}(h_{\text{ERM}}) - \hat{\mathcal{R}}(h^*)}_{\leq 0} + 2\alpha$$

Lemma

For any $\epsilon > 0$, $\delta > 0$, the following holds with probability at least $1-\delta$ w.r.t. $\mathcal{D}_m$:

$$\forall h \in \mathcal{H} \qquad |\mathcal{R}(h) - \hat{\mathcal{R}}_m(h)| \leq \sqrt{\frac{\log|\mathcal{H}| + \log\frac{2}{\delta}}{2m}} =: \alpha$$

Step 2: we use the lemma to bound the difference between

- $h_{\text{ERM}} \in \mathbf{argmin}_{\bar{h}\in\mathcal{H}} \hat{\mathcal{R}}_m(\bar{h})$ (result of ERM)
- $h^* \in \mathbf{argmin}_{\bar{h}\in\mathcal{H}} \mathcal{R}(\bar{h})$ (if exists, otherwise argue with arbitrarily close approximation)

$$\mathcal{R}(h_{\text{ERM}}) - \mathcal{R}(h^*) = \underbrace{\mathcal{R}(h_{\text{ERM}}) - \hat{\mathcal{R}}(h_{\text{ERM}})}_{\leq\alpha} + \hat{\mathcal{R}}(h_{\text{ERM}}) - \hat{\mathcal{R}}(h^*) + \underbrace{\hat{\mathcal{R}}(h^*) - \mathcal{R}(h^*)}_{\leq\alpha}$$

$$\leq \underbrace{\hat{\mathcal{R}}(h_{\text{ERM}}) - \hat{\mathcal{R}}(h^*)}_{\leq 0} + 2\alpha \leq 2\sqrt{\frac{\log|\mathcal{H}| + \log\frac{2}{\delta}}{2m}} \overset{m\geq m_0}{\leq} \epsilon$$