# Statistical Machine Learning

Christoph Lampert

IST AUSTRIA

*Institute of Science and Technology*

Spring Semester 2013/2014 // Lecture 6

## Measure Concentration Inequalities

- $Z$ 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 $Z_1, Z_2, \ldots,$ be i.i.d. random variables with mean $\mu$, then*

$$\frac{1}{m}\sum_{i=1}^{m} Z_i \quad \xrightarrow{m\to\infty} \quad \mu \qquad \textit{with probability } 1.$$

**Measure concentration inequalities** quantify the deviation between the two values for finite $m$.

# Markov's Inequality

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

**Lemma (Markov's inequality)**

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

**Proof.** Step 1) We can write

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

Step 2) Since $\mathbb{P}[Z \geq x]$ is non-increasing in $x$, we have

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

# Markov's Inequality

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

**Lemma (Markov's inequality)**

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

**Proof.** Step 1) We can write

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

Step 2) Since $\mathbb{P}[Z \geq x]$ is non-increasing in $x$, we have

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

**Example**

Is it possible that more than half of the population have a salary more than twice the average? No, by $a = 2\mu$.

## Chebyshev's Inequality

### Lemma (Chebyshev's inequality)

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

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

For any $a \geq 0$:

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

## Chebyshev's Inequality

### Lemma (Chebyshev's inequality)

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

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

For any $a \geq 0$:

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

**Remark**: Chebyshev ineq. has similar role as "$3\sigma$-rule" for Gaussians:

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

but Chebyshev holds for arbitrary probability distributions.

# Chebyshev's Inequality

## Example (Match Statistics)

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

What if we pretended $Z$ is Gaussian?

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

With Chebyshev:

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

# Applying Chebyshev's Inequality

## 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 $1-\delta$:*

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

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

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

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

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

# How large should my test set be?

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

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

- test set $\mathcal{D} = \{(x_1, y_1) \ldots, (x_n, y_n)\} \overset{i.i.d.}{\sim} p(x, y)$,
- random variables $Z_i = [\![ g(x_i) = y_i ]\!] \in \{0, 1\}$,
- $\mathbb{E}[Z_i] = \mathbb{E}\{[\![ g(x_i) \neq y_i ]\!]\} = \mu$ (test error of $g$)
- $\text{Var}[Z_i] = \mathbb{E}\{(Z_i - \mu)^2\} = \mu(1-\mu)^2 + (1-\mu)\mu^2 = \mu(1-\mu) \Rightarrow C = \frac{1}{4}$

Setup: fixed confidence, e.g. $\delta = 0.01$, $\sqrt{\frac{C}{\delta m}} = \sqrt{\frac{0.25}{0.01m}} = 5\sqrt{\frac{1}{m}}$

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\right| \leq 5\sqrt{\frac{1}{m}}\right] \geq 0.99$$

To be 99%-certain that the error is within $\pm 5\%$, use $m \geq 10,000$.

# Hoeffding's Lemma and Inequality

## 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* $\epsilon > 0$,

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

Proof: Blackboard...

# Hoeffding's Inequality – Proof

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

Note: $\mathbb{E}[X_i] = 0$; $\mathbb{E}[\bar{X}] = 0$; 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 independent 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:

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

# Hoeffding's Inequality – Proof cont.

Previous step:

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

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

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

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

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

To combine both directions we use the *union bound*:
$P[A \cup B] \leq P[A] + P[B]$,

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

$\square$

# How large should my test set be?

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

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

- test set $\mathcal{D} = \{(x_1, y_1) \ldots, (x_n, y_n)\} \overset{i.i.d.}{\sim} p(x, y)$,
- random variables $Z_i = [\![g(x_i) = y_i]\!] \in \{0, 1\}$, $\to \quad b - a = 1$
- $\mathbb{E}[Z_i] = \mathbb{E}\{[\![g(x_i) \neq y_i]\!]\} = \mu$ (test error of $g$)

Setup: fixed confidence $\delta = 0.01$: $m = \log(\frac{2}{\delta})/\epsilon^2 \Rightarrow \epsilon = \sqrt{\log(200)/m}$

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\right| \leq 5.3\sqrt{\frac{1}{m}}\right] \geq 0.99$$

To be 99%-certain that the error is within $\pm 5\%$, use $m \geq 11300$.

# Difference: Chebyshev's vs. Hoeffding's Inequality

With $\Delta = \frac{1}{m}\sum_{i=1}^{m} Z_i$ and $\mu = \mathbb{E}[\frac{1}{m}\sum_{i=1}^{m} Z_i]$:

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

$$\mathbb{P}\left[\,|\Delta - \mu| > \sqrt{\frac{C}{\delta m}}\,\right] \leq \delta, \qquad \mathbb{P}\left[\,|\Delta - \mu| > \epsilon\,\right] \leq \frac{C}{\epsilon^2 m}$$

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

$$\mathbb{P}\left[\,|\Delta - \mu| > \sqrt{\frac{(b-a)^2 \log\frac{2}{\delta}}{m}}\,\right] \leq \delta, \qquad \mathbb{P}\left[\,|\Delta - \mu| > \epsilon\,\right] \leq 2e^{-\frac{m\epsilon^2}{(b-a)^2}}.$$

- interval decreases like $\frac{1}{\sqrt{m}}$, prob. decreases like $e^{-m}$

Both are typical **PAC (probably approximately correct)** statements:
"With **prob.** $1-\delta$, the estimated $\Delta$ is an $\epsilon$-**close approximation** of $\mu$."

# PAC Learning

Learning scenario:

- $\mathcal{X}$: input set
- $\mathcal{Y}$: output/label set, for now: $\mathcal{Y} = \{-1, 1\}$ or $\mathcal{Y} = \{0, 1\}$
- $p(x, y)$: data distribution (unknown to us)
- *new:* assume **deterministic** labels, $y = f(x)$ for unknown $f : \mathcal{X} \to \mathcal{Y}$
- $S = \{(x^1, y^1), \dots, (x^m, y^m)\} \overset{i.i.d.}{\sim} p(x, y)$: training set
- $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$: loss function
- $\mathcal{H} \subseteq \{h : \mathcal{X} \to \mathcal{Y}\}$: hypothesis set (the lerner's choice)

Quantity of interest:

- $\mathcal{R}_p(h) = \mathbb{P}_{(x,y)\sim p(x,y)}\{h(x) \neq y\} = \mathbb{P}_{x\sim p(x)}\{h(x) \neq f(x)\}$

What can we learn?

- We know: there is (at least one) $f : \mathcal{X} \to \mathcal{Y}$ that has $\mathcal{R}(f) = 0$.
- Can we find such $f$ from $S_m$? If yes, how large must $m$ be?

## 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 sample size)

such that

- for every probability distribution $d$ over $\mathcal{X}$,
- and for every labeling function $f : \mathcal{X} \to \mathcal{Y}$, with $\mathcal{R}_d(f) = 0$,

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

$$\forall m \geq m_0(\epsilon, \delta) \quad \mathbb{P}_{S \sim d^m}[\, \mathcal{R}_d(A[S]) > \epsilon \,] \leq \delta.$$

# Empirical Risk Minimization

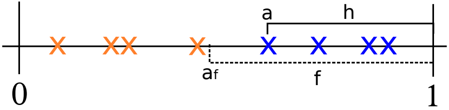
**Definition (Empirical Risk Minimization (ERM) Algorithm)**

**input** hypothesis set $\mathcal{H} \subseteq \{h : \mathcal{X} \to \mathcal{Y}\}$ (not necessarily finite)
**input** training set $S = \{(x^1, y^1), \dots, (x^m, y^m)\}$
**output** $h = \operatorname*{argmin}_{h \in H} \frac{1}{m} \sum_{i=1}^{m} [\![ h(x^i) = y^i ]\!]$ (best on training set)

ERM learns the classifier of minimal training error.

- We saw already: this might or might not work well.
- Can we characterize when ERM works and when it fails?

# Example: Learning Thresholding Functions

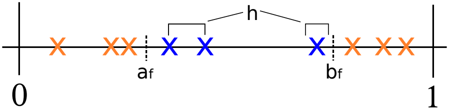



- $\mathcal{X} = [0, 1]$, $\mathcal{Y} = \{0, 1\}$,
- $\mathcal{H} = \{h_a(x) = [\![\, x \geq a \,]\!],$ for $0 \leq a \leq 1\}$,
- $f(x) = h_{a_f}(x)$ for some $0 \leq a_f \leq 1$.
- for simplicity: $d(x) \equiv 1$ (uniform distribution in $\mathcal{X}$)
- training set $S = \{(x^1, y^1), \dots, (x^m, y^m)\}$
- ERM rule: $\quad h = \operatorname*{argmin}_{h_a \in H} \frac{1}{m} \sum_{i=1}^{m} [\![ h_a(x^i) = y^i ]\!]$,

  pick *smallest possible "$+1$" region (largest $a$)* when not unique (to make algorithm deterministic)

Claim: ERM learns $f$ (in the PAC sense) Proof: blackboard...

# 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 some $K \in \mathbb{N}\}$,
  for $h_{\mathcal{I}}(x) = [\![\, x \in I_1 \vee I_2 \vee \cdots \vee I_K ]\!]$ with $I_i = [a_k, b_k]$
- $f(x) = h_{[a_f, b_f]}(x)$ for some $0 \leq a_f \leq b_f \leq 1$.
- for simplicity: $d(x) \equiv 1$ (uniform distribution in $\mathcal{X}$)
- training set $S = \{(x^1, y^1), \dots, (x^m, y^m)\}$
- ERM rule: $h = \operatorname*{argmin}_{\mathcal{I}} \frac{1}{m} \sum_{i=1}^{m} [\![ h_{\mathcal{I}}(x^i) = y^i ]\!]$,

  pick *smallest possible "$+1$" region* when not unique

Claim: ERM fails to learn $f$ in the PAC sense Proof: blackboard...

Can we prove more general statements?

**Theorem (Learnability of finite hypothesis classes (realizable case))**

*Let* $\mathcal{H} = \{h_1, \ldots, h_K\}$ *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 empirical risk minimization algorithm with* $m_0(\epsilon, \delta) = \frac{1}{\epsilon}(\,\log(|\mathcal{H}| + \log(1/\delta)\,)$

Proof: blackboard.

## Examples: Finite hypothesis classes

Model selection:

- Clients offer me trained classifiers: *decision tree*, *LogReg* or an *SVM*? Which one should I buy?

Finite precision:

- For $x \in \mathbb{R}^d$, the hypothesis set $\mathcal{H} = \{f(x) = \operatorname{sign}\langle w, x\rangle\}$ is infinite.
- But: on a computer, $w$ is restricted to 64-bit `doubles`: $|\mathcal{H}_c| = 2^{64d}$.
  $m_0(\epsilon, \delta) = \frac{1}{\epsilon}(\,\log(|\mathcal{H}| + \log(1/\delta)\,) \approx \frac{1}{\epsilon}(44d + \log(1/\delta))$

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) feasible.

# Agnostic PAC Learning

More realistic scenario: labeling isn't a deterministic function

- $\mathcal{X}$: input set
- $\mathcal{Y}$: output/label set, for now: $\mathcal{Y} = \{-1, 1\}$ or $\mathcal{Y} = \{0, 1\}$
- $p(x, y)$: data distribution (unknown to us)
- ~~assume **deterministic** labels, $y = f(x)$ for unknown $f : \mathcal{X} \to \mathcal{Y}$~~
- $S = \{(x^1, y^1), \dots, (x^m, y^m)\} \overset{i.i.d.}{\sim} p(x, y)$: training set
- $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$: loss function
- $\mathcal{H} \subseteq \{h : \mathcal{X} \to \mathcal{Y}\}$: hypothesis set (the lerner's choice)

Quantity of interest:

- $\mathcal{R}_p(h) = \mathbb{P}_{(x,y)\sim p(x,y)}\{h(x) \neq y\}$

What can we learn?

- there might not be an $f : \mathcal{X} \to \mathcal{Y}$ 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 sample size)

such that

- for every probability distribution $d(x, y)$ 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

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

$$\forall m \geq m_0(\epsilon, \delta) \quad \mathbb{P}_{S \sim d^m}[\, \mathcal{R}_d(A[S]) - \min_{\bar{h} \in \mathcal{H}} \mathcal{R}_d(\bar{h}) \; > \; \epsilon \,] \; \leq \; \delta.$$

## Uniform Convergence is Sufficient for Learnability

There's three main quantities:

- training error, $\hat{\mathcal{R}}_S(h) = \frac{1}{m}\sum_{i=1}^{m}[\![h(x^i) = y^i]\!]$ for any $h \in \mathcal{H}$,
- generalizaton error, $\mathcal{R}_d(h) = \mathbb{E}_{(x,y)\sim d}[\![h(x) = y]\!]$ for any $h \in \mathcal{H}$,
- best achievable generalization error, $\mathbf{min}_{\bar{h}\in\mathcal{H}}\,\mathcal{R}(\bar{h})$.

# Uniform Convergence is Sufficient for Learnability

There's three main quantities:

- training error, $\hat{\mathcal{R}}_S(h) = \frac{1}{m}\sum_{i=1}^{m}[\![h(x^i) = y^i]\!]$ for any $h \in \mathcal{H}$,
- generalizaton error, $\mathcal{R}_d(h) = \mathbb{E}_{(x,y)\sim d}[\![h(x) = y]\!]$ for any $h \in \mathcal{H}$,
- best achievable generalization error, $\mathbf{min}_{\bar{h}\in\mathcal{H}} \mathcal{R}(\bar{h})$.

## Definition ($\epsilon$-representative sample)

A training set $S$ is called $\epsilon$-**representative** (for the current situation), if

$$\forall h \in \mathcal{H} \quad |\hat{\mathcal{R}}_S(h) - \mathcal{R}_d(h)| \leq \epsilon.$$

## Uniform Convergence is Sufficient for Learnability

There's three main quantities:

- training error, $\hat{\mathcal{R}}_S(h) = \frac{1}{m}\sum_{i=1}^{m} [\![h(x^i) = y^i]\!]$ for any $h \in \mathcal{H}$,
- generalizaton error, $\mathcal{R}_d(h) = \mathbb{E}_{(x,y)\sim d}[\![h(x) = y]\!]$ for any $h \in \mathcal{H}$,
- best achievable generalization error, $\mathbf{min}_{\bar{h}\in\mathcal{H}} \mathcal{R}(\bar{h})$.

### Definition ($\epsilon$-representative sample)

A training set $S$ is called $\epsilon$-**representative** (for the current situation), if

$$\forall h \in \mathcal{H} \quad |\hat{\mathcal{R}}_S(h) - \mathcal{R}_d(h)| \leq \epsilon.$$

### Lemma ("ERM works well for $(\epsilon/2)$-representative training sets")

*Let $S$ be $(\epsilon/2)$-representative. Then any $h_{ERM}$ with $\hat{\mathcal{R}}_S(h_{ERM}) = \mathbf{min}_{\bar{h}\in\mathcal{H}} \hat{\mathcal{R}}_S(\bar{h})$ (i.e. a possible output of ERM) fulfills*

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

**Lemma ("ERM works well for $(\epsilon/2)$-representative training sets")**

*Let $S$ be $(\epsilon/2)$-representative. Then any $h_{ERM}$ with $\hat{\mathcal{R}}_S(h_{ERM}) = \mathbf{min}_{\bar{h}\in\mathcal{H}} \hat{\mathcal{R}}_S(\bar{h})$ (i.e. a possible output of ERM) fulfills*

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

**Proof.** For any $h \in \mathcal{H}$:

$$\mathcal{R}_d(h_{ERM}) \leq \hat{\mathcal{R}}_S(h_{ERM}) + \frac{\epsilon}{2} \leq \hat{\mathcal{R}}_S(h) + \frac{\epsilon}{2} \leq \mathcal{R}(h) + \frac{\epsilon}{2} + \frac{\epsilon}{2} = \mathcal{R}(h) + \epsilon$$

Taking the minimum over $h \in \mathcal{H}$ on both sides, we obtain

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

## Uniform Conference

### Definition

A hypothesis class $\mathcal{H}$ is said to have the **uniform convergence property**, if

- for every $\epsilon > 0$, and every $\delta > 0$

there exists an

- $m^{UC} = m^{UC}(\epsilon, \delta) \in \mathbb{N}$

such that

- for every probability distribution $d$ over $\mathcal{X}$,

if

- $S$ is a training set of size $m \geq m^{UC}$, sampled i.i.d. from $d$,

the probability that $S$ is $\epsilon$-representable is at least $1 - \delta$.

$$\forall m \geq m_0(\epsilon, \delta) \quad \mathbb{P}_{S \sim d^m}[\forall h \in \mathcal{H} : |\hat{\mathcal{R}}_S(h) - \mathcal{R}_d(h)| \leq \epsilon] \geq 1 - \delta,$$

or $\quad \forall m \geq m_0(\epsilon, \delta) \quad \mathbb{P}_{S \sim d^m}[\exists h \in \mathcal{H} : |\hat{\mathcal{R}}_S(h) - \mathcal{R}_d(h)| > \epsilon] \leq \delta.$

## Uniform Conference is sufficient for PAC Learnability

### Lemma

*Let $\mathcal{H}$ have the uniform convergence property with sample complexity $m_{UC}(\epsilon, \delta)$, then $\mathcal{H}$ is agnostically PAC-learnable with sample complexity $m(\epsilon, \delta) \leq m_{UC}(\delta/2, \delta)$, and the ERM rule is a successful PAC learning algorithm.*

**Proof.** combine two previous lemma:

- uniform convergence (for $\delta, \epsilon/2$)

implies

- $\epsilon/2$-representativeness,

implies

- $(\delta, \epsilon)$ learnability (by ERM rule).

# Finite Hypothesis Classes are PAC Learnable

**Theorem**

*Every finite hypothesis set, $\mathcal{H}$, is agnostic PAC-learnable by the ERM algorithm.*

# Finite Hypothesis Classes are PAC Learnable

**Theorem**

*Every finite hypothesis set, $\mathcal{H}$, is agnostic PAC-learnable by the ERM algorithm.*

**Proof.** it's enough to show that $\mathcal{H}$ has the uniform convergence property.

**Part 1)** let $\epsilon > 0$, $\delta > 0$ be fixed. We want to find $m$

$$\mathrm{Pr}_{S\sim d}\left[\forall h \in \mathcal{H} : |\hat{\mathcal{R}}_S(h) - \mathcal{R}_d(h)| \leq \epsilon\right] \geq 1 - \delta,$$

for $|S| = m$, or equivalently, the $m$ such that

$$\mathrm{Pr}_{S\sim d}\left[\exists h \in \mathcal{H} : |\hat{\mathcal{R}}_S(h) - \mathcal{R}_d(h)| > \epsilon\right] \leq \delta,$$

Note:

$$\left\{S : \ \exists h \in \mathcal{H} : |\hat{\mathcal{R}}_S(h) - \mathcal{R}_d(h)| > \epsilon\right\} = \bigcup_{h\in\mathcal{H}} \{S : |\hat{\mathcal{R}}_S(h) - \mathcal{R}_d(h)| > \epsilon\}$$

Union bound:

$$\mathrm{Pr}_{S\sim d}\left[\exists h \in \mathcal{H} : |\hat{\mathcal{R}}_S(h) - \mathcal{R}_d(h)| > \epsilon\right] \leq \sum_{h\in\mathcal{H}} \mathrm{Pr}_{S\sim d}\left[|\hat{\mathcal{R}}_S(h) - \mathcal{R}_d(h)| > \epsilon\right]$$

## Finite Hypothesis Classes are PAC Learnable – Proof cont.

**Part 2)** We show that for any fixed $h$ (chosen before $S$ is sampled), $\Pr_{S\sim d}\left[|\hat{\mathcal{R}}_S(h) - \mathcal{R}_d(h)| > \epsilon\right]$ is small for large enough $m = |S|$.

Reminder:

- $\hat{\mathcal{R}}_S(h) = \frac{1}{m}\sum_{i=1}^{m}[\![h(x^i) = y^i]\!]$, and $\mathcal{R}_d(h) = \mathbb{E}_{(x,y)\sim d}[\![h(x) = y]\!]$.

$h$ is fixed, independent of $S$. Therefore, $\mathbb{E}_{S\sim d}\{\hat{\mathcal{R}}_S(h)\} = \mathcal{R}_d(h)$.

Apply **Hoeffding's inequality**: $\mathbb{P}_S[|\hat{\mathcal{R}}_S(h) - \mathcal{R}_d(h)| > \epsilon] \leq 2e^{-2m\epsilon^2}$

**Part 3)** We insert Part 2) into Part 1):

$$\Pr_{S\sim d}\left[\exists h \in \mathcal{H} : |\hat{\mathcal{R}}_S(h) - \mathcal{R}_d(h)| > \epsilon\right] \leq \sum_{h\in\mathcal{H}} 2e^{-2m\epsilon^2} = 2|\mathcal{H}|e^{-2m\epsilon^2}$$

If we choose $m \geq \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2}$, then $2|\mathcal{H}|e^{-2m\epsilon^2} \leq 2|\mathcal{H}|e^{-\log(2|\mathcal{H}|/\delta)} = \delta$.

$\mathcal{H}$ has the uniform convergence property with $m_{UC}(\epsilon,\delta) = \left\lceil \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2} \right\rceil$.