Statistical Machine Learning

Christoph Lampert

Institute of Science and Technology

Spring Semester 2013/2014 // Lecture 6
Measure Concentration Inequalities

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

Lemma (Law of Large Numbers)

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

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

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

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

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

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

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

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

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

**Example**

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

Lemma (Chebyshev’s inequality)

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

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

For any \(a \geq 0\):

\[
P[|Z - \mathbb{E}[Z]| \geq a] = P[(Z - \mathbb{E}[Z])^2 \geq a^2] \leq \frac{\mathbb{E}[(Z - \mathbb{E}[Z])^2]}{a^2} = \frac{\text{Var}[Z]}{a^2}.
\]
Chebyshev’s Inequality

Lemma (Chebyshev’s inequality)

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

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

For any \(a \geq 0\):

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

Remark: Chebyshev ineq. has similar role as "3σ-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.
Example (Match Statistics)

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

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.
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 $\text{Var}[Z_i] \leq C$. 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 $\text{Var} \left[ \frac{1}{m} \sum_{i=1}^{m} Z_i \right] = \frac{1}{m} \sum_{i=1}^{m} \text{Var}[Z_i] \leq C$. Chebyshev’s inequality gives us for any $a \geq 0$:

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

Setting $\delta = \frac{C}{ma^2}$ and solving for $a$ yields $a = \sqrt{\frac{C}{\delta m}}$. 
How large should my test set be?

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

Setup: fixed classifier \( g : \mathcal{X} \to \mathcal{Y} \)
- test set \( \mathcal{D} = \{(x_1, y_1) \ldots, (x_n, y_n)\}\) i.i.d. \( p(x, y) \),
- random variables \( Z_i = [g(x_i) \neq y_i] \in \{0, 1\} \),
- \( \mathbb{E}[Z_i] = \mathbb{E}\{[g(x_i) \neq y_i]\} = \mu \) (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. \( \delta = 0.01 \), \( \sqrt{\frac{C}{\delta m}} = \sqrt{\frac{0.25}{0.01m}} = 5\sqrt{\frac{1}{m}} \)

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

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...
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

\[
P[\bar{X} \geq \epsilon] = P[e^{\lambda \bar{X}} \geq e^{\lambda \epsilon}] \leq e^{-\lambda \epsilon} \mathbb{E}[e^{\lambda \bar{X}}]
\]

From 3) the independent of the \( X_i \) we have

\[
\mathbb{E}[e^{\lambda \bar{X}}] = \mathbb{E}\left[\prod_{i=1}^{n} e^{\frac{\lambda X_i}{m}}\right] = \prod_{i=1}^{n} \mathbb{E}[e^{\frac{\lambda X_i}{m}}]
\]

Use 4) Hoeffding’s Lemma for every \( i \):

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

In combination:

\[
P[\bar{X} \geq \epsilon] \leq e^{-\lambda \epsilon} e^{\frac{\lambda^2 (b-a)^2}{8m}}.
\]
Hoeffding’s Inequality – Proof cont.

Previous step:

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

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

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

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

\[ \Pr[\bar{X} \leq \epsilon] \leq e^{\frac{-2m \epsilon^2}{(b-a)^2}} \]

To combine both directions we use the union bound:

\[ P[A \cup B] \leq P[A] + P[B], \]

\[ \Pr[|\bar{X}| \geq \epsilon] = \Pr[(\bar{X} \geq \epsilon) \lor (\bar{X} \leq \epsilon)] \leq 2e^{\frac{-2m \epsilon^2}{(b-a)^2}}. \]
How large should my test set be?

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

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

- test set \( \mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\} \sim i.i.d. p(x, y) \),
- random variables \( Z_i = [g(x_i) \neq y_i] \in \{0, 1\}, \rightarrow b - a = 1 \)
- \( \mathbb{E}[Z_i] = \mathbb{E}\{[g(x_i) \neq y_i]\} = \mu \) (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: $\text{Var}[Z_i] \leq C$

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

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

- Hoeffding’s: $Z_i$ takes values in $[a, b]$:

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

- 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 \( f : \mathcal{X} \rightarrow \mathcal{Y} \)
- \( S = \{(x^1, y^1), \ldots, (x^m, y^m)\} \): training set
- \( \ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R} \): loss function
- \( \mathcal{H} \subseteq \{h : \mathcal{X} \rightarrow \mathcal{Y}\} \): 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) \( f : \mathcal{X} \rightarrow \mathcal{Y} \) that has \( \mathcal{R}(f) = 0 \).
- Can we find such \( f \) from \( S_m \)? If yes, how large must \( m \) be?
A hypothesis class $\mathcal{H}$ is called **PAC learnable** by an algorithm $A$, if

- for every $\epsilon > 0$ (accuracy $\rightarrow$ "approximate correct")
- and every $\delta > 0$ (confidence $\rightarrow$ "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 $f : \mathcal{X} \rightarrow \mathcal{Y}$, 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**

<table>
<thead>
<tr>
<th>Definition (Empirical Risk Minimization (ERM) Algorithm)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>input</strong> hypothesis set ( \mathcal{H} \subseteq { h : \mathcal{X} \rightarrow \mathcal{Y} } ) (not necessarily finite)</td>
</tr>
<tr>
<td><strong>input</strong> training set ( S = { (x^1, y^1), \ldots, (x^m, y^m) } )</td>
</tr>
<tr>
<td><strong>output</strong> ( h = \arg\min_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \left[ h(x^i) \neq y^i \right] ) (best on training set)</td>
</tr>
</tbody>
</table>

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

- $\mathcal{X} = [0, 1]$, $\mathcal{Y} = \{0, 1\}$,
- $\mathcal{H} = \{ h_a(x) = \mathbb{I}[x \geq a], \text{ for } 0 \leq a \leq 1 \}$,
- $f(x) = h_{a_f}(x)$ for some $0 \leq a_f \leq 1$.
- for simplicity: $d(x) \equiv 1$ (uniform distribution in $\mathcal{X}$)
- training set $\mathcal{S} = \{(x^1, y^1), \ldots, (x^m, y^m)\}$
- ERM rule: $h = \arg\min_{h_a \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \mathbb{I}[h_a(x^i) \neq y^i]$, 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

- \( \mathcal{X} = [0, 1] \), \( \mathcal{Y} = \{0, 1\} \),
- \( \mathcal{H} = \{ h_\mathcal{I}(x) \mid \mathcal{I} = \{I_1, \ldots, I_K\} \text{ for some } K \in \mathbb{N} \} \),
- for \( h_\mathcal{I}(x) = \left\lfloor x \in \bigcup_{k=1}^{K} I_k \right\rfloor \) 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 \( S = \{(x^1, y^1), \ldots, (x^m, y^m)\} \)
- ERM rule: \( h = \arg\min_\mathcal{I} \frac{1}{m} \sum_{i=1}^{m} [h_\mathcal{I}(x^i) \neq y^i] \),

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 $\mathcal{H} = \{ f(x) = \text{sign}\langle w, x \rangle \}$ is infinite.
- But: on a computer, $w$ is restricted to 64-bit doubles: $|\mathcal{H}_c| = 2^{64d}$. $m_0(\epsilon, \delta) = \frac{1}{\epsilon} \left( \log(|\mathcal{H}| + \log(1/\delta)) \right) \approx \frac{1}{\epsilon} (44d + \log(1/\delta))$

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

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