ETH zürich

Seminar for Statistics

Mathematical Tools in Machine Learning

taught by

Prof. Dr. Fadoua Balabdaoui

Lecture in fall 2019

Based on

Shalev-Shwartz and Ben-David, *Understanding machine learning : from theory to algorithms*

Notes by

Sascha Gaudlitz

Special thank goes to Max von Gierke for proof reading

Contents

1 PAC Learning

1.1 Learning Models

Definition 1.1. A *statistical learning problem* is a tuple $(\mathcal{H}, \mathcal{Z}, \mathscr{D}, l)$, where

- $\mathcal{H}$ is a class of functions $h: \mathcal{X} \to \mathcal{Y}$ called *hypothesis class* (and h is called a *prediction rule*, *hypothesis* or *classifier*);
- $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ is the *domain*, where
  - $\mathcal{X}$ is the *state space of the observations*;
  - $\mathcal{Y}$ is the *label space of the observations*;
- $\mathscr{D}$ is a probability distribution on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and
- l is a (measurable) *loss function* $l: \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$.

The *learner* has to find a predictor $h \in \mathcal{H}$ which minimizes the *true loss (risk)*

$$L_{\mathscr{D}}(h) := \mathrm{E}_{(x,y)\sim\mathscr{D}}\big[l(h,(x,y))\big] = \mathrm{E}_{z\sim\mathscr{D}}\big[l(h,z)\big].$$

As $\mathscr{D}$ is not known to the learner, he cannot compute the true risk $L_{\mathscr{D}}(h)$ and hence not its minimizer. He therefore has to find an algorithm $\mathscr{A}: \bigcup_{n\in\mathbb{N}} (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{H}$, which, given a sample $\mathcal{S} := \big\{(x_1, y_1), \ldots, (x_n, y_n)\big\} \in (\mathcal{X} \times \mathcal{Y})^n$ of i.i.d. (according to $\mathscr{D}$) observations, returns an *estimator* $h_{\mathcal{S}} := \mathscr{A}(\mathcal{S})$. One frequently used method to approximate the true risk is to use the *empirical loss (risk)*

$$L_{\mathcal{S}}(h) := \frac{1}{n} \sum_{i=1}^{n} l(h, (x_i, y_i)).$$

A straightforward learning algorithm is the *empirical risk minimization (ERM)* paradigm which sets

$$h_{\mathcal{S}} = \mathscr{A}(\mathcal{S}) : \in \operatorname*{argmin}_{h\in\mathcal{H}} L_{\mathcal{S}}(h).$$

Remark 1.2. Usually, $\mathcal{X} = \mathbb{R}^d$ for some $d > 0$. If the label space is finite, i.e. $\mathcal{Y} = \{1, \ldots, k\}$, $k \in \mathbb{N}$, then the task is (*binary* for $k = 2$) *classification*, for $\mathcal{Y} = \mathbb{R}^q$, $q \in \mathbb{N}$, the task is *regression*. It is common for binary classification to use the labels $-1, 1$ instead of $0, 1$. In some cases, there exists a *labelling function* $f: \mathcal{X} \to \mathcal{Y}$ such that $\mathscr{D} = \mathscr{D}^x \otimes \mathscr{D}^{y|x}$, where $\mathscr{D}^{y|x} = \delta_{f(x)}$ is the Dirac measure in $f(x)$. Common loss functions are

- *0-1-loss*, used for classification:

  $$l_{0-1}(h, z) := l_{0-1}(h, (x, y)) := \mathbf{1}_{\{h(x)\neq y\}}.$$

  Then, for $h \in \mathcal{H}$, holds

  $$L_{\mathscr{D}}(h) = \mathrm{P}_{(x,y)\sim\mathscr{D}}\big(h(x) \neq y\big) = \mathrm{E}_{(x,y)\sim\mathscr{D}}\Big[\mathbf{1}_{\{h(x)\neq y\}}\Big] \text{ and } L_{\mathcal{S}}(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{h(x_i)\neq y_i\}};$$

- *squared loss*, used for regression:

$$l_{\text{sq}}(h, z) := l_{\text{sq}}(h, (x, y)) := (h(x) - y)^2.$$

Here, for $h \in \mathcal{H}$, holds

$$L_{\mathscr{D}}(h) = \mathrm{E}_{(x,y)\sim\mathscr{D}}\Big[(h(x) - y)^2\Big] \text{ and } L_{\mathcal{S}}(h) = \frac{1}{n}\sum_{i=1}^{n}\big(h(x_i) - y_i\big)^2.$$

Example 1.3. Consider the following examples:

- *Multiclass classification*: Consider the problem of classifying some document into one of the following categories: "sports", "health", "mathematics". The *domain* would change to $\mathcal{X} = \mathbb{N}_0^p$ with $p \in \mathbb{N}$ if $x \in \mathcal{X}$ is a vector storing counts of some specific key words. The *label space* becomes $\mathcal{Y} = \{1, \dots, k\}$ with $k \geq 2$ some integer ($k = 3$ in this example). A *training set* is given as $\mathcal{S} = \big\{(x_1, y_1), \dots, (x_n, y_n)\big\}$. A prediction rule $h_{\mathcal{S}}$ is the output of the learning algorithm. For a new document with associated feature vector $x \in \mathcal{X}$, it yields the predicted class $h_{\mathcal{S}}(x) =: y \in \{1, \dots, k\}$.

- *Regression*: In learning problems with regression, the goal is to find a relationship between some response $y \in \mathcal{Y} = \mathbb{R}^d$ and covariates $x \in \mathcal{X} = \mathbb{R}^p$, $d, p \in \mathbb{N}$. A *linear* regression model assumes that $\mathrm{E}[y \mid x] = \beta_0^\top x$, where β_0 is some unknown regression vector in $\mathbb{R}^p$. In this case, the hypothesis space is given by

$$\mathcal{H} = \Big\{h\colon \mathcal{X} \to \mathcal{Y} \;\Big|\; \exists\, \beta \in \mathbb{R}^p \quad h(x) = \beta^\top x\Big\}.$$

The true error (risk) in regression for a predictor $h \in \mathcal{H}$ takes the form

$$L_{\mathscr{D}}(h) = \mathrm{E}_{(x,y)\sim\mathscr{D}}\Big[(h(x) - y)^2\Big]$$

and is also referred to as *mean squared error (MSE)*. Note that in this case we assume a fixed (i.e. non-random) $h \in \mathcal{H}$.

Remark 1.4. Why do we choose to restrict ourselves to the hypothesis class $\mathcal{H}$? One reason for that is that the ERM paradigm might produce an estimator that overfits on the training data. This will be reflected by a small ratio of $L_{\mathcal{S}}(h)/L_{\mathscr{D}}(h)$. Another reason is that large hypothesis classes are hard or even impossible to learn, see Theorem 1.23. Restricting the set of possible classifiers introduces an *inductive bias*.

1.2 Finite Classes

1.2.1 Realizability Assumption

For this subsection, we assume that there exists a labelling function $f\colon \mathcal{X} \to \mathcal{Y}$ such that $\mathscr{D} = \mathscr{D}^x \otimes \delta_{f(x)}$, otherwise the following assumption can never be true.

Definition 1.5. The *realizability assumption* assumes that

$$\exists\, h^* \in \mathcal{H} \quad L_{\mathscr{D}}(h^*) = 0. \tag{1}$$

Remark 1.6. The realizability assumption (1) implies that $L_S(h^*) = 0$. We can see this by noting that (1) implies that $\mathscr{D}\left(\{h^*(x) = y\}\right) = 1$ and as the sample S is i.i.d. from $\mathscr{D}$, this leads to $L_S(h^*) = 0$. Trivially, any ERM rule h^{ERM} then satisfies with probability 1

$$0 = L_S(h^*) \geq L_S\left(h^{\mathrm{ERM}}\right) \geq 0,$$

hence $L_S(h^{\mathrm{ERM}}) = 0$ with probability 1.

Definition 1.7. A hypothesis class $\mathcal{H}$ is *probably approximately correct (PAC) learnable* with respect to some domain $\mathcal{Z}$ and a loss function l if there exists a function $n_{\mathcal{H}}\colon (0,1)^2 \to \mathbb{N}$ and a learning algorithm $\mathscr{A}\colon \bigcup_{n\in\mathbb{N}} \mathcal{Z}^n \to \mathcal{H}$ such that for all $(\varepsilon, \delta) \in (0,1)^2$, all distributions $\mathscr{D}$ on $\mathcal{Z}$ and all samples S of size $n \geq n_{\mathcal{H}}(\varepsilon, \delta)$, $h_S := \mathscr{A}(S) \in \mathcal{H}$ satisfies

$$\mathrm{P}_{S\sim\mathscr{D}^n}\left(L_{\mathscr{D}}(h_S) \leq \varepsilon\right) \geq 1 - \delta. \tag{2}$$

Remark 1.8. Note that $n_{\mathcal{H}}(\varepsilon, \delta)$ from Definition 1.7 is viewed as the smallest sample size for which the learning guarantee (2) is satisfied and is called *sample complexity*.

Theorem 1.9. *Given that the realizability assumption (1) holds, any finite hypothesis class is PAC learnable using the ERM paradigm with sample complexity*

$$n_{\mathcal{H}}(\varepsilon, \delta) \leq \left\lceil \frac{\log\left(|\mathcal{H}|/\delta\right)}{\varepsilon} \right\rceil.$$

Proof. Fix $n \in \mathbb{N}$ and let $\mathcal{H}_B$ be the set of "bad" hypotheses

$$\mathcal{H}_B := \left\{h \in \mathcal{H} \,\middle|\, L_{\mathscr{D}}(h) > \varepsilon\right\}$$

and $\mathcal{M}$ be the set of "misleading" samples

$$\mathcal{M} := \left\{S \in \mathcal{Z}^n \,\middle|\, \exists\, h \in \mathcal{H}_B \quad L_S(h) = 0\right\} = \bigcup_{h\in\mathcal{H}_B} \left\{S \in \mathcal{Z}^n \,\middle|\, L_S(h) = 0\right\}.$$

Since the realizability assumption (1) holds, we know that $L_S(h_S) = 0$ with probability 1. Consequently,

$$\left\{S \in \mathcal{Z}^n \,\middle|\, L_{\mathscr{D}}(h_S) > \varepsilon\right\} \subset \mathcal{M}$$

and by the union bound we find

$$\begin{aligned}
\mathscr{D}^n\left(\left\{S \in \mathcal{Z}^n \,\middle|\, L_{\mathscr{D}}(h_S) > \varepsilon\right\}\right) &\leq \sum_{h\in\mathcal{H}_B} \mathscr{D}^n\left(\left\{S \in \mathcal{Z}^n \,\middle|\, L_S(h) = 0\right\}\right) \\
&= \sum_{h\in\mathcal{H}_B} \mathscr{D}^n\left(\left\{S \in \mathcal{Z}^n \,\middle|\, \forall\, i = 1, \ldots, n \quad h(x_i) = f(x_i)\right\}\right) \\
&= \sum_{h\in\mathcal{H}_B} \prod_{i=1}^{n} \mathscr{D}\left(\left\{(x, y) \in \mathcal{Z} \,\middle|\, h(x) = y\right\}\right).
\end{aligned}$$

Furthermore, for all $h \in \mathcal{H}_B$ holds

$$\mathscr{D}\left(\left\{(x,y) \in \mathcal{Z} \,\middle|\, h(x) = y\right\}\right) = 1 - \mathscr{D}\left(\left\{(x,y) \in \mathcal{Z} \,\middle|\, h(x) \neq y\right\}\right) = 1 - L_{\mathscr{D}}(h) \leq 1 - \varepsilon.$$

Consequently,

$$\mathscr{D}^n\left(\left\{S \in \mathcal{Z}^n \,\middle|\, L_{\mathscr{D}}(h_S) > \varepsilon\right\}\right) \leq |\mathcal{H}_B|(1-\varepsilon)^n \leq |\mathcal{H}_B|\exp\{-n\varepsilon\} \leq |\mathcal{H}|\exp\{-n\varepsilon\}.$$

Then

$$\mathscr{D}^n\left(\left\{S \in \mathcal{Z}^n \,\middle|\, L_{\mathscr{D}}(h_S) \leq \varepsilon\right\}\right) \geq 1 - |\mathcal{H}|\exp\{-n\varepsilon\}.$$

Choosing $n \geq \log\left(|\mathcal{H}|/\delta\right)/\varepsilon$ then yields $\mathscr{D}^n\left(\left\{S \in \mathcal{Z}^n \,\middle|\, L_{\mathscr{D}}(h_S) \leq \varepsilon\right\}\right) \geq 1 - \delta$. □

1.2.2 Bayes Classifier

The following result proves a lower bound for the performance of any learning algorithm on binary classification. Note that it is also valid if there does not exist a labelling function.

Theorem 1.10. *Consider the binary classification task, in this case we use labels* 0 *and* 1, *i.e.* $\mathcal{Y} = \{0,1\}$. *Let* $\mathcal{H}$ *be the set of all measurable functions* $\mathcal{X} \to \mathcal{Y}$ *and consider the* Bayes classifier

$$\varphi\colon \mathcal{X} \to \mathcal{Y}, \quad x \mapsto \mathbf{1}\left(\mathscr{D}^{y|x}\left(\{y = 1 \mid x\}\right) > \frac{1}{2}\right),$$

where we decompose the joint distribution into its $\mathcal{X}$*-marginal and conditional probability according to* $\mathscr{D} = \mathscr{D}^x \otimes \mathscr{D}^{y|x}$. *Then*

$$\varphi = \operatorname*{argmin}_{h \in \mathcal{H}} L_{\mathscr{D}}(h).$$

Proof. Consider any other estimator h and let $\eta(x) \coloneqq \mathscr{D}^{y|x}\left(\{y = 1 \mid x\}\right)$. Then

$$\begin{aligned} L_{\mathscr{D}}(h) &= \mathscr{D}(h(x) \neq y) = \mathrm{E}_{x\sim\mathscr{D}^x}\left[\mathrm{E}_{y\sim\mathscr{D}^{y|x}}\left[\mathbf{1}\left(h(x) \neq y\right) \,\middle|\, x\right]\right] \\ &= \mathrm{E}_{x\sim\mathscr{D}^x}\left[\eta(x)\mathbf{1}\left(h(x) = 0\right) + (1-\eta(x))\mathbf{1}\left(h(x) = 1\right)\right] \\ &= \mathrm{E}_{x\sim\mathscr{D}^x}\left[\mathbf{1}\left(h(x) = 0\right)\left(2\eta(x) - 1\right) + 1 - \eta(x)\right]. \end{aligned}$$

Note that

$$\mathbf{1}\left(h(X) = 0\right)\left(2\eta(X) - 1\right) + 1 - \eta(X) = \begin{cases} \eta(X) & \text{if } \mathbf{1}\left(h(X) = 0\right) = 1, \\ 1 - \eta(X) & \text{if } \mathbf{1}\left(h(X) = 0\right) = 0. \end{cases}$$

Hence $L_{\mathscr{D}}(h)$ is minimal if and only if $h \equiv \varphi$. Note that the proof allows for arbitrary assignment if $\eta(x) = 1/2$. □

Remark 1.11. Note that the Bayes classifier is a likelihood-ratio classifier.

1.2.3 Uniform convergence

Motivation. We now aim to waive the realizability assumption (1). Given a hypothesis class, we do not assume anymore that

$$\exists\, h^* \in \mathcal{H} \quad L_{\mathscr{D}}(h^*) = \mathrm{E}_{(x,y)\sim\mathscr{D}}\big[l(h^*,(x,y))\big] = 0.$$

This leads to another type of learning, the so-called *agnostic PAC learning*.

Definition 1.12. A hypothesis class $\mathcal{H}$ is *agnostic PAC learnable* with respect to some domain $\mathcal{Z}$ and a loss function l if there exists a function $n_{\mathcal{H}}\colon (0,1)^2 \to \mathbb{N}$ and a learning algorithm $\mathscr{A}\colon \bigcup_{n\in\mathbb{N}} \mathcal{Z}^n \to \mathcal{H}$ such that for all $(\varepsilon,\delta) \in (0,1)^2$, all distributions $\mathscr{D}$ on $\mathcal{Z}$ and all samples $\mathcal{S}$ of size $n \geq n_{\mathcal{H}}(\varepsilon,\delta)$, $h_{\mathcal{S}} := \mathscr{A}(\mathcal{S}) \in \mathcal{H}$ satisfies

$$\mathrm{P}_{\mathcal{S}\sim\mathscr{D}^n}\left(L_{\mathscr{D}}(h_{\mathcal{S}}) \leq \min_{h\in\mathcal{H}} L_{\mathscr{D}}(h) + \varepsilon\right) \geq 1-\delta. \tag{3}$$

Remark 1.13. Note that, as in Remark 1.8, $n_{\mathcal{H}}(\varepsilon,\delta)$ from Definition 1.12 is viewed as the smallest sample size for which the agnostic learning guarantee (3) is satisfied and is called *sample complexity*. Furthermore, if the realizability assumption (1) holds, then $\min_{h\in\mathcal{H}} L_{\mathscr{D}}(h) = 0$ and hence agnostic PAC learning gives the same guarantee as PAC learning.

Motivation. We know that under the realizability assumption (1), finite classes are PAC learnable. How can we show that finite classes are agnostic PAC learnable? And what conditions on the loss functions do we need? The tool to answer these question is the *uniform convergence* property.

Definition 1.14. A training set $\mathcal{S}$ is called *ε-representative* with respect to a domain $\mathcal{Z}$, a hypothesis class $\mathcal{H}$, a loss function l and a distribution $\mathscr{D}$ if

$$\forall\, h \in \mathcal{H} \quad \big|L_{\mathcal{S}}(h) - L_{\mathscr{D}}(h)\big| \leq \varepsilon.$$

Lemma 1.15. *Assume that a training set $\mathcal{S}$ is $\varepsilon/2$-representative with respect to $(\mathcal{Z},\mathcal{H},l,\mathscr{D})$. Then, for any ERM predictor $h_{\mathcal{S}}$, i.e. $h_{\mathcal{S}} \in \operatorname{argmin}_{h\in\mathcal{H}} L_{\mathcal{S}}(h)$, we have that*

$$L_{\mathscr{D}}(h_{\mathcal{S}}) \leq \min_{h\in\mathcal{H}} L_{\mathscr{D}}(h) + \varepsilon.$$

Proof. Since $\mathcal{S}$ is $\varepsilon/2$-representative, we have for all $h \in \mathcal{H}$ by the ERM rule

$$L_{\mathscr{D}}(h_{\mathcal{S}}) \leq L_{\mathcal{S}}(h_{\mathcal{S}}) + \frac{\varepsilon}{2} \leq L_{\mathcal{S}}(h) + \frac{\varepsilon}{2} \leq L_{\mathscr{D}}(h) + \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = L_{\mathscr{D}}(h) + \varepsilon. \qquad \square$$

Definition 1.16. We say that a hypothesis class $\mathcal{H}$ has the *uniform convergence property* with respect to $(\mathcal{Z},l)$, if we can find a function $n^{\mathrm{uc}}_{\mathcal{H}}\colon (0,1)^2 \to \mathbb{N}$ such that for all $(\varepsilon,\delta) \in (0,1)^2$, all distributions $\mathscr{D}$ and all samples $\mathcal{S}$ of size $n \geq n^{\mathrm{uc}}_{\mathcal{H}}(\varepsilon,\delta)$ drawn i.i.d. from $\mathscr{D}$ holds

$$\mathrm{P}_{\mathcal{S}\sim\mathscr{D}^n}(\mathcal{S} \text{ is } \varepsilon\text{-representative}) \overset{\mathrm{def}}{=} \mathrm{P}_{\mathcal{S}\sim\mathscr{D}^n}\Big(\forall\, h \in \mathcal{H} \quad \big|L_{\mathcal{S}}(h) - L_{\mathscr{D}}(h)\big| \leq \varepsilon\Big) \geq 1-\delta.$$

Remark 1.17. Recall that we aim to find a learning algorithm $\mathscr{A}$ that, given a sample $\mathcal{S}$, minimizes the error $L_{\mathscr{D}}(\mathscr{A}(\mathcal{S}))$. However, as $\mathscr{D}$ is unknown, our only proxy for $L_{\mathscr{D}}(\mathscr{A}(\mathcal{S}))$ is the empirical risk $L_{\mathcal{S}}(\mathscr{A}(\mathcal{S}))$. The uniform convergence property gives a guarantee that (probably) empirical and true risk are close, i.e. "bad" samples with misleading empirical risk are unlikely.

Theorem 1.18. *If a hypothesis class $\mathcal{H}$ has the uniform convergence property with respect to $(\mathcal{Z}, l)$ with a function $n^{uc}_{\mathcal{H}}$, then this class is agnostic PAC learnable with sample complexity $n_{\mathcal{H}}(\varepsilon,\delta) \le n^{uc}_{\mathcal{H}}(\varepsilon/2,\delta)$. Furthermore, the ERM paradigm is an agnostic PAC learner for $\mathcal{H}$.*

Proof. We know from Lemma 1.15 that for all samples $\mathcal{S}$ of size $n \ge n^{\mathrm{uc}}_{\mathcal{H}}(\varepsilon/2,\delta)$

$$\left\{\forall\, h \in \mathcal{H} \quad \left|L_{\mathcal{S}}(h) - L_{\mathscr{D}}(h)\right| \le \frac{\varepsilon}{2}\right\} \subset \left\{L_{\mathscr{D}}(h_{\mathcal{S}}) \le \min_{h\in\mathcal{H}} L_{\mathscr{D}}(h) + \varepsilon\right\},$$

where $h_{\mathcal{S}} \in \operatorname{argmin}_{h\in\mathcal{H}} L_{\mathcal{S}}(h)$. Hence,

$$\mathrm{P}_{\mathcal{S}\sim\mathscr{D}^n}\left(L_{\mathscr{D}}(h_{\mathcal{S}}) \le \min_{h\in\mathcal{H}} L_{\mathscr{D}}(h) + \varepsilon\right) \ge \mathrm{P}_{\mathcal{S}\sim\mathscr{D}^n}\left(\forall\, h \in \mathcal{H} \quad \left|L_{\mathcal{S}}(h) - L_{\mathscr{D}}(h)\right| \le \frac{\varepsilon}{2}\right) \ge 1-\delta,$$

where the last inequality holds for all $n \ge n^{\mathrm{uc}}_{\mathcal{H}}(\varepsilon/2,\delta)$, implying that $n_{\mathcal{H}}(\varepsilon,\delta) \le n^{\mathrm{uc}}_{\mathcal{H}}(\varepsilon/2,\delta)$. □

We now proceed to prove that finite classes are agnostic PAC learnable. For this we need the following inequality:

Lemma 1.19 (Hoeffding's inequality)**.** *Let $\theta_1,\dots,\theta_n$ be i.i.d. random variables such that $\mathrm{E}[\theta_1] = \mu \in \mathbb{R}$ and $\mathrm{P}(a \le \theta_1 \le b) = 1$ for some $a < b \in \mathbb{R}$. Then, for all $\varepsilon > 0$, it holds with $\bar{S}_n := \sum_{i=1}^n \theta_i/n$ that*

$$\mathrm{P}\left(\left|\bar{S}_n - \mu\right| > \varepsilon\right) \le 2\,exp\left\{-\frac{2n\varepsilon^2}{(b-a)^2}\right\}.$$

In order to prove Hoeffing's inequality, we need the following auxiliary result:

Lemma 1.20. *Let X be a centered random variable with $\mathrm{P}\big(X \in [a,b]\big) = 1$ for $a < b \in \mathbb{R}$. Then, for any $\lambda > 0$, we have*

$$\mathrm{E}\big[exp\{\lambda x\}\big] \le exp\left\{\frac{\lambda^2(b-a)^2}{8}\right\}.$$

Proof. Set $f(x) := \exp\{\lambda x\}$ for $x \in \mathbb{R}$ and $\lambda > 0$ fixed. f is convex on $\mathbb{R}$, hence

$$f(x) \le \frac{b-x}{b-a} f(a) + \frac{x-a}{b-a} f(b)$$

for $x \in [a,b]$. Consequently, as X is supported in $[a,b]$ and has mean 0,

$$\mathrm{E}\big[\exp\{\lambda x\}\big] \le \frac{b-\mathrm{E}[X]}{b-a}\exp\{\lambda a\} + \frac{\mathrm{E}[X]-a}{b-a}\exp\{\lambda b\} = \frac{b}{b-a}\exp\{\lambda a\} - \frac{a}{b-a}\exp\{\lambda b\}.$$

Set $h := \lambda(b-a)$ and $p := -a/(b-a)$. Also, define the function

$$L\colon [0,\infty) \to \mathbb{R}, \quad h \mapsto -hp + \log\big(1-p+p\exp\{h\}\big).$$

We then have that

$$\begin{aligned}
\mathrm{E}\big[\exp\{\lambda x\}\big] &= \frac{b}{b-a}\exp\{\lambda a\} - \frac{a}{b-a}\exp\{\lambda b\} \\
&= \exp\{\lambda a\}\Big(1-p+p\exp\big\{\lambda(b-a)\big\}\Big) \\
&= \exp\{-hp\}\big(1-p+p\exp\{h\}\big) \\
&= \exp\big\{L(h)\big\}.
\end{aligned}$$

If we can now show that $L(h) \le h^2/8$, then we would have

$$\mathrm{E}\big[\exp\{\lambda x\}\big] \le \exp\left\{\frac{\lambda^2(b-a)^2}{8}\right\},$$

which would conclude the proof. In order to show that $L(h) \le h^2/8$ note that $L(0) = 0$. Also, $L'(h) = -p + p\exp\{h\}/(1-p+p\exp\{h\})$ and hence $L'(0) = -p+p = 0$. Furthermore,

$$\begin{aligned}
L''(h) - \frac{1}{4} &= \frac{p(1-p)\exp\{h\}}{(1-p+p\exp\{h\})^2} - \frac{1}{4} = \frac{4p(1-p)\exp\{h\} - (1-p+p\exp\{h\})^2}{4(1-p+p\exp\{h\})^2} \\
&= \frac{4p(1-p)\exp\{h\} - (1-p)^2 - 2p(1-p)\exp\{h\} - p^2\exp\{2h\}}{4(1-p+p\exp\{h\})^2} \\
&= \frac{-(1-p)^2 + 2p(1-p)\exp\{h\} - p^2\exp\{2h\}}{4(1-p+p\exp\{h\})^2} \\
&= \frac{-\big(1-p-p\exp\{h\}\big)^2}{4(1-p+p\exp\{h\})^2} \le 0.
\end{aligned}$$

Hence, for all $h \in (0,\infty)$ holds $L''(h) \le 1/4$.
Using a Taylor expansion of L up to the second order, we can write for $h^* \in [0,h]$

$$L(h) = L(0) + hL'(0) + \frac{h^2}{2}L''(h^*) \le \frac{h^2}{8}.$$ □

Proof of Lemma 1.19. Let $X_i := \theta_i - \mu$, where $\mu = \mathrm{E}[\theta_1]$ and $\bar{X} = \sum_{i=1}^n X_i/n$. Then $X_1,\dots,X_n$ are also i.i.d. and $X_i \in [a-\mu, b-\mu]$. Using monotonicity of $x \mapsto \exp\{\lambda x\}$ for $\lambda > 0$ fixed and the Markov inequality, it follows that

$$\mathrm{P}\big(\bar{X} > \varepsilon\big) = \mathrm{P}\Big(\exp\big\{\lambda\bar{X}\big\} > \exp\{\lambda\varepsilon\}\Big) \le \mathrm{E}\Big[\exp\big\{\lambda\bar{X}\big\}\Big]\exp\{-\lambda\varepsilon\}.$$

Furthermore,

$$\mathrm{E}\Big[\exp\big\{\lambda\bar{X}\big\}\Big] = \mathrm{E}\left[\exp\left\{\frac{\lambda}{n}\sum_{i=1}^{n} X_i\right\}\right] = \mathrm{E}\left[\prod_{i=1}^{n}\exp\left\{\frac{\lambda}{n}X_i\right\}\right] = \prod_{i=1}^{n}\mathrm{E}\left[\exp\left\{\frac{\lambda}{n}X_1\right\}\right]$$
$$= \mathrm{E}\left[\exp\left\{\frac{\lambda}{n}X_1\right\}\right]^n.$$

By Lemma 1.20, we have

$$\mathrm{E}\left[\exp\left\{\frac{\lambda}{n}X_1\right\}\right] \leq \exp\left\{\left(\left(\frac{\lambda}{n}\right)^2\frac{(b-a)^2}{8}\right)\right\}.$$

Hence, for all $\lambda > 0$ holds $\mathrm{P}\big(\bar{X} > \varepsilon\big) \leq \exp\big\{\lambda^2(b-a)^2/(8n) - \lambda\varepsilon\big\}$, which implies

$$\mathrm{P}\big(\bar{X} > \varepsilon\big) \leq \inf_{\lambda>0}\exp\left\{\frac{\lambda^2(b-a)^2}{8n} - \lambda\varepsilon\right\} = \inf_{\lambda>0}\exp\big\{\Psi(\lambda)\big\}$$

for $\Psi(\lambda) = \lambda^2(b-a)^2/(8n) - \lambda\varepsilon$. Note that

$$\Psi'(\lambda) = \frac{\lambda(b-a)^2}{4n} - \varepsilon = 0 \iff \lambda = \lambda^* := \frac{4\varepsilon n}{(b-a)^2} > 0.$$

Since Ψ is strictly convex on $\mathbb{R}$, $(\lambda^*, \Psi(\lambda^*))$ is the global minimum of Ψ on $\mathbb{R}$. Since $\lambda^* > 0$, $\inf_{\lambda>0}\Psi(\lambda) = \Psi(\lambda^*)$. Hence

$$\mathrm{P}\big(\bar{X} > \varepsilon\big) \leq \exp\big\{\Psi(\lambda^*)\big\} = \exp\left\{\frac{(\lambda^*)^2(b-a)^2}{8n} - \lambda^*\varepsilon\right\} = \exp\left\{-\frac{2n\varepsilon^2}{(b-a)^2}\right\}$$

for all $\varepsilon > 0$. The same arguments can be applied to $-X_i$ to show that

$$\mathrm{P}\big(-\bar{X} > \varepsilon\big) \leq \exp\left\{-\frac{2n\varepsilon^2}{(b-a)^2}\right\}.$$

Hence

$$\mathrm{P}\Big(\big|\bar{X}\big| > \varepsilon\Big) \leq 2\exp\left\{-\frac{2n\varepsilon^2}{(b-a)^2}\right\}. \qquad \square$$

We are now ready to state the Theorem about agnostic PAC learnability of finite hypothesis classes:

Theorem 1.21. *Let $\mathcal{H}$ be a finite hypothesis class, $\mathcal{Z}$ a domain and $l\colon \mathcal{H}\times\mathcal{Z}\to[0,1]$ be a loss function. Then $\mathcal{H}$ satisfies the uniform convergence property with*

$$n^{uc}_{\mathcal{H}}(\varepsilon,\delta) \leq \left\lceil \frac{1}{2\varepsilon^2}\log\left(\frac{2|\mathcal{H}|}{\delta}\right)\right\rceil. \tag{4}$$

Furthermore, $\mathcal{H}$ is agnostic PAC learnable with the ERM paradigm. The sample complexity satisfies

$$n_{\mathcal{H}}(\varepsilon,\delta) \le n_{\mathcal{H}}^{uc}\left(\frac{\varepsilon}{2},\delta\right) \le \left\lceil \frac{2}{\varepsilon^2}\log\left(\frac{2|\mathcal{H}|}{\delta}\right)\right\rceil.$$

Proof. Fix $n \in \mathbb{N}$ and $h \in \mathcal{H}$ and set $\theta_i := l(h,(x_i,y_i))$, $i = 1,\dots,n$. Then θ_i, $i = 1,\dots,n$, are i.i.d. and $L_{\mathcal{S}}(h) = \sum_{i=1}^n \theta_i/n$ has expected value $L_{\mathscr{D}}(h)$. By Hoeffding's inequality (Lemma 1.19), it holds that

$$\mathscr{D}^n\left(\left\{\left|L_{\mathcal{S}}(h) - L_{\mathscr{D}}(h)\right| > \varepsilon\right\}\right) \le 2\exp\left\{-2n\varepsilon^2\right\}.$$

Applying the union bound, we can conclude that

$$\begin{aligned}\mathscr{D}^n\left(\left\{\exists\, h \in \mathcal{H} \quad \left|L_{\mathcal{S}}(h) - L_{\mathscr{D}}(h)\right| > \varepsilon\right\}\right) &\le \sum_{h\in\mathcal{H}} \mathscr{D}^n\left(\left\{\left|L_{\mathcal{S}}(h) - L_{\mathscr{D}}(h)\right| > \varepsilon\right\}\right)\\ &\le |\mathcal{H}|2\exp\left\{-2n\varepsilon^2\right\}.\end{aligned}$$

Hence we have shown that $\mathcal{H}$ has the uniform convergence property by choosing $n \ge n_{\mathcal{H}}^{\mathrm{uc}}(\varepsilon,\delta)$ as in (4). Theorem 1.18 yields the second claim. □

Remark 1.22. Theorem 1.21 shows that finite hypothesis classes with bounded loss functons are agnostic PAC learnable. What happens if the hypothesis class $\mathcal{H}$ is infinite? Sections 1.4 and 1.5 give a comprehensive answer to this question using the concept of *Vapnik–Chervonenkis (VC) dimension*. Here, we will give another possible answer using the *discretization trick*:
As an example, consider $\mathcal{X} = \mathbb{R}$ and $\mathcal{Y} = \{-1,1\}$. Let $\mathcal{H} = \{h_\theta \mid \theta \in \mathbb{R}\}$, where $h_\theta(x) = \operatorname{sign}(x-\theta)$ for $x \in \mathbb{R}$. Then $|\mathcal{H}| = \infty$. However, using floating point arithmetic with 64 bits, we can "approximate" $\mathcal{H}$ by $\tilde{\mathcal{H}}$ with $|\tilde{\mathcal{H}}| = 2^{64}$. Since we are in the setting of binary classification, the loss function is given by the 0-1-loss

$$l(h,z) = l(h,(x,y)) := \mathbf{1}_{\{h_\theta(x)\neq y\}} \in [0,1].$$

By Theorem 1.21, $\tilde{\mathcal{H}}$ is agnostic PAC learnable and the sample complexity satisfies

$$n_{\tilde{\mathcal{H}}}(\varepsilon,\delta) \le \left\lceil \frac{2}{\varepsilon^2}\log\left(\frac{2\cdot 2^{64}}{\delta}\right)\right\rceil = \left\lceil \frac{2}{\varepsilon^2}\left(\log\left(\frac{2}{\delta}\right) + 64\log(2)\right)\right\rceil \le \left\lceil \frac{2\log\left(\frac{2}{\delta}\right)+128}{\varepsilon^2}\right\rceil,$$

where we used that $\log(2) < 1$.

1.3 No-free-lunch Theorem and Error decomposition

We proceed to analyze the question whether there can exist some "super learner" $\mathscr{A}: \mathcal{S} \mapsto h_{\mathcal{S}} = \mathscr{A}(\mathcal{S})$ that always minimizes the loss $L_{\mathscr{D}}(h_{\mathcal{S}})$. The answer is no, as the following Theorem shows.

Theorem 1.23 (No free lunch)**.** *Let $\mathscr{A}$ be any learning algorithm for the task of binary classification with respect to some domain $\mathcal{X}$ and the 0-1-loss function l. Let $n < |\mathcal{X}|/2$ be the training set size.*
Then there exists a distribution $\mathscr{D}$ on $\mathcal{X} \times \{0,1\}$ such that

1) $\exists\, f\colon \mathcal{X} \to \{0,1\} \quad L_{\mathscr{D}}(f) = 0$, *but*

2) $\mathrm{P}_{\mathcal{S}\sim\mathscr{D}^n}\big(L_{\mathscr{D}}(\mathscr{A}(\mathcal{S})) \geq 1/8\big) \geq 1/7$.

Proof. Consider a subset $\mathcal{C} \subset \mathcal{X}$ with $|\mathcal{C}| = 2n$. Then there are $T := 2^{|\mathcal{C}|} = 2^{2n}$ possible functions $\mathcal{C} \to \{-1,1\}$. Let us denote these functions by $f_1, \dots, f_T$ and fix some $i \in \{1,\dots,T\}$. Let $\mathscr{D}_i$ be the distribution defined on $\mathcal{C} \times \{-1,1\}$ by

$$\mathscr{D}_i\Big(\big\{(x,y)\big\}\Big) := \begin{cases} 1/|\mathcal{C}| & \text{if } y = f_i(x), \\ 0 & \text{otherwise.} \end{cases}$$

Hence $\mathscr{D}_i$ draws from $\mathcal{C}$ uniformly (with the same probability $1/(2n)$) and conditionally on x assigns $f_i(x)$ to y with probability 1. Clearly,

$$L_{\mathscr{D}_i}(f_i) \overset{\text{def}}{=} \mathrm{P}_{(x,y)\sim\mathscr{D}_i}\big(f_i(x) \neq y\big) = 0.$$

We are going to show for any learning algorithm $\mathcal{A}$ which receives a training set $\mathcal{S}$ of n samples and outputs a classifier $\mathscr{A}(\mathcal{S})\colon \mathcal{C} \to \{-1,1\}$ it holds that

$$\max_{i=1,\dots,T} \mathrm{E}_{\mathcal{S}\sim\mathscr{D}_i^n}\big[L_{\mathscr{D}_i}(\mathscr{A}(\mathcal{S}))\big] \geq \frac{1}{4}. \tag{5}$$

If (5) holds, then there exists a distribution $\mathscr{D}$ on $\mathcal{C} \times \{-1,1\}$ such that there exists a classifier $f\colon \mathcal{C} \to \{-1,1\}$ such that $L_{\mathscr{D}}(f) = 0$ and

$$\mathrm{E}_{\mathcal{S}\sim\mathscr{D}^n}\big[L_{\mathscr{D}}(\mathscr{A}(\mathcal{S}))\big] \geq \frac{1}{4}.$$

It then follows that

$$\mathrm{P}_{\mathcal{S}\sim\mathscr{D}^n}\left(L_{\mathscr{D}}(\mathscr{A}(\mathcal{S})) \geq \frac{1}{8}\right) \geq \frac{1}{7}.$$

Let us now prove (5). For $i \in \{1,\dots,T\}$, consider a training set $\mathcal{S}$ of n i.i.d. samples distributed according to $\mathscr{D}_i$. Such a training set looks like

$$\mathcal{S} = \big\{(x_1, f(x_1)), \dots, (x_n, f(x_n))\big\}.$$

$(x_1,\dots,x_n)$ is a random draw with $|\mathcal{C}|^n = (2n)^n$ possible outcomes, which are all occurring with the same probability $1/|\mathcal{C}|^n$. Set $k := |\mathcal{C}|^n$ and let us list all possible training sets of n examples $S \sim \mathscr{D}_i$ as $\mathcal{S}_1^i, \dots, \mathcal{S}_k^i$. Note that

$$\max_{i=1,\dots,T} \mathrm{E}_{\mathcal{S}\sim\mathscr{D}_i^n}\big[L_{\mathscr{D}_i}(\mathscr{A}(\mathcal{S}))\big] \geq \frac{1}{T}\sum_{i=1}^{T} \mathrm{E}_{\mathcal{S}\sim\mathscr{D}_i^n}\big[L_{\mathscr{D}_i}(\mathscr{A}(\mathcal{S}))\big]$$

$$= \frac{1}{k}\sum_{j=1}^{k}\frac{1}{T}\sum_{i=1}^{T} L_{\mathscr{D}_i}(\mathscr{A}(\mathcal{S}_j^i))$$

$$\geq \min_{j=1,\dots,k}\frac{1}{T}\sum_{i=1}^{T} L_{\mathscr{D}_i}(\mathscr{A}(\mathcal{S}_j^i)). \tag{6}$$

Now, fix $j \in \{1,\dots,k\}$, then a training set $\mathcal{S}_i^j$ is of the form

$$\mathcal{S}_j^i = \left\{(x_1^j, f_i(x_1^j)),\dots,(x_n^j, f_i(x_n^j))\right\}.$$

Let $\left\{v_1,\dots,v_p\right\} := \mathcal{C} \setminus \left\{x_1^j,\dots,x_n^j\right\}$ and note that by $|\mathcal{C}| = 2n$, $p \geq n$. Consider some $h\colon \mathcal{C} \to \{-1,1\}$ and denote by $\mathscr{D}_i^x$ the x-marginal of $\mathscr{D}_i$, then

$$\begin{aligned} L_{\mathscr{D}_i}(h) &= \mathrm{P}_{(x,y)\sim\mathscr{D}_i}\big(h(x) \neq y\big) = \mathrm{P}_{x\sim\mathscr{D}_i^x}\big(h(x) \neq f_i(x)\big) \\ &= \frac{1}{2n}\sum_{c\in\mathcal{C}} \mathbf{1}\left(h(c) \neq f_i(c)\right) \geq \frac{1}{2n}\sum_{r=1}^{p} \mathbf{1}\left(h(v_r) \neq f_i(v_r)\right) \\ &\geq \frac{1}{2p}\sum_{r=1}^{p} \mathbf{1}\left(h(v_r) \neq f_i(v_r)\right). \end{aligned}$$

Replace h by $\mathscr{A}(\mathcal{S}_i^j)$, then we obtain

$$\begin{aligned} \frac{1}{T}\sum_{i=1}^{T} L_{\mathscr{D}_i}(\mathscr{A}(\mathcal{S}_i^j)) &\geq \frac{1}{T}\sum_{i=1}^{T}\frac{1}{2p}\sum_{r=1}^{p} \mathbf{1}\left(\mathscr{A}(\mathcal{S}_i^j)(v_r) \neq f_i(v_r)\right) \\ &\geq \frac{1}{2}\min_{r=1,\dots,p}\frac{1}{T}\sum_{i=1}^{T} \mathbf{1}\left(\mathscr{A}(\mathcal{S}_i^j)(v_r) \neq f_i(v_r)\right). \end{aligned} \tag{7}$$

For $r \in \{1,\dots,p\}$, the functions $f_i, i = 1,\dots,T$, can be partitioned into $T/2$ disjoint pairs where for a pair $(\tilde{f}_i^{(0)}, \tilde{f}_i^{(1)})$ we have that:

1) $\tilde{f}_i^{(0)}(c) = \tilde{f}_i^{(1)}(c) \quad \forall c \in \mathcal{C} \setminus \{v_r\}$,

2) $\tilde{f}_i^{(0)}(v_r) = 0, \tilde{f}_i^{(1)}(v_r) = 1$.

Then, using the fact that $v_r \notin \{x_1^j,\dots,x_n^j\}$

$$\left\{\left(x_1^j, \tilde{f}_i^{(0)}(x_1^j)\right),\dots,\left(x_n^j, \tilde{f}_i^{(0)}(x_n^j)\right)\right\} = \left\{\left(x_1^j, \tilde{f}_i^{(1)}(x_1^j)\right),\dots,\left(x_n^j, \tilde{f}_i^{(1)}(x_n^j)\right)\right\} = \mathcal{S}_j^i.$$

Consequently,

$$\begin{aligned} \sum_{i=1}^{T} \mathbf{1}\left(\mathscr{A}(S_j^i)(v_r) \neq f_i(v_r)\right) = \sum_{l=1}^{T/2}\Bigg[&\mathbf{1}\left(\mathscr{A}(S_j^l)(v_r) \neq \tilde{f}_l^{(0)}(v_r) = 0\right) + \\ &+\mathbf{1}\left(\mathscr{A}(S_j^l)(v_r) \neq \tilde{f}_l^{(1)}(v_r) = 1\right)\Bigg] \end{aligned}$$

$$= T/2.$$

Together with (6) and (7) it now follows that

$$\max_{i=1,\dots,T} \mathrm{E}_{\mathcal{S}\sim\mathscr{D}^n}\left[L_{\mathscr{D}_i}(\mathscr{A}(\mathcal{S}))\right] \geq \frac{1}{2} \min_{j=1,\dots,k} \min_{r=1,\dots,p} \frac{1}{T}\frac{T}{2} = \frac{1}{4},$$

which completes the proof. □

Remark 1.24. Theorem 1.23 shows that if we allow $\mathcal{H} = \{0,1\}^{\mathcal{X}}$ in the binary classification setup, then we are bound to fail PAC learning. This indicates that we need to restrict the "complexity" of the hypothesis class. We will make this precise in subsections 1.4 and 1.5.

Remark 1.25 (Error decomposition and bias-complexity trade-off)**.** Given a sample $\mathcal{S}$, assume that the predictor $h_{\mathcal{S}}$ was obtained by the ERM rule, i.e. $h_{\mathcal{S}} \in \operatorname{argmin}_{h\in\mathcal{H}} L_{\mathcal{S}}(h)$. Then we can decompose the true error

$$L_{\mathscr{D}}(h_{\mathcal{S}}) = \underbrace{\min_{h\in\mathcal{H}} L_{\mathscr{D}}(h)}_{\text{approximation error}} + \underbrace{L_{\mathscr{D}}(h_{\mathcal{S}}) - \min_{h\in\mathcal{H}} L_{\mathscr{D}}(h)}_{\text{estimation error}}.$$

The *approximation error* is not connected to any probabilistic arguments. If $\mathcal{H}$ is large, we expect that this error will be small (if the realizability assumption holds, it will be 0). The approximation error corresponds to the *bias* we introduce with our hypothesis class. The *estimation error* is the result from replacing the unknown distribution by a training set (empirical measure). If $\mathcal{H}$ is big, then we need many examples in the training set, e.g. if $\mathcal{H}$ is finite, we know that the sample complexity $n_{\mathcal{H}}(\varepsilon,\delta)$ is of order $2\log\left(2|\mathcal{H}|/\delta\right)/\varepsilon^2$ by the Theorems 1.9 and 1.21.
The interplay between approximation and estimation error is called *bias-complexity trade-off*.

1.4 Vapnik–Chervonenkis (VC) dimension

In this subsection we will develop a theory for PAC learnability of infinite hypothesis classes.

Example 1.26. Consider the class of all thresholds over $\mathbb{R}$:

$$\mathcal{H} := \{h_a \mid a \in \mathbb{R}\}, \quad h_a(\,\cdot\,) := \mathbf{1}(\,\cdot\, < a).$$

Note that $\mathcal{H}$ is an infinite class. However we claim that $\mathcal{H}$ is PAC learnable with *any* ERM rule given the realizability assumption (1). Furthermore, its sample complexity is

$$n_{\mathcal{H}}(\varepsilon,\delta) \leq \left\lceil \frac{2}{\varepsilon}\log\left(\frac{2}{\delta}\right)\right\rceil$$

for $(\varepsilon,\delta) \in (0,1)^2$.

To see this, let a^* be such that h_{a^*} is a perfect classifier, that is

$$L_{\mathscr{D}}(h_{a^*}) = \mathrm{P}_{(x,y)\sim\mathscr{D}}\big(h_{a^*}(x) \neq y\big) = 0.$$

Let $\mathscr{D}^x$ be the x-marginal, which is assumed to be continuous. Take a_0 and a_1 such that

$$\mathrm{P}_{x\sim\mathscr{D}^x}\big(x \in (a_0, a^*]\big) = \mathrm{P}_{x\sim\mathscr{D}^x}\big(x \in (a^*, a_1]\big) = \frac{\varepsilon}{2}.$$

Let $\mathcal{S} = \big\{(x_1, y_1), \ldots, (x_n, y_n)\big\}$ be some training set with i.i.d. examples and write

$$b_0 := \max\big\{x_i \,\big|\, (x_i, 1) \in \mathcal{S}\big\} \text{ and } b_1 := \min\big\{x_i \,\big|\, (x_i, 0) \in \mathcal{S}\big\}$$

(we assume that both b_0 and b_1 are well-defined). Note that b_0 and b_1 depend on the sample $\mathcal{S}$. Let $b_{\mathcal{S}}$ be the threshold of the ERM rule $h_{b_{\mathcal{S}}} \in \operatorname{argmin}_{a\in\mathbb{R}} L_{\mathcal{S}}(h_a)$. Recall that $L_{\mathcal{S}}(h_{\mathcal{S}}) = 0$ with probability 1 (Remark 1.6). By construction, it follows that $b_0 < b_{\mathcal{S}} \leq b_1$. We will now show that

$$\{\mathcal{S} \mid b_0 \geq a_0 \wedge b_1 \leq a_1\} \subset \big\{\mathcal{S} \,\big|\, L_{\mathscr{D}}(h_{\mathcal{S}}) \leq \varepsilon\big\}. \tag{8}$$

Assume the event on the left of (8). Then

$$\begin{aligned}
L_{\mathscr{D}}(h_{\mathcal{S}}) &= \mathrm{P}_{x\sim\mathscr{D}^x}\Big(\mathbf{1}\left(x < b_{\mathcal{S}}\right) \neq \mathbf{1}\left(x < a^*\right)\Big) \\
&= \mathrm{P}_{x\sim\mathscr{D}^x}\left(x < b_{\mathcal{S}} \wedge x \geq a^*\right) + \mathrm{P}_{x\sim\mathscr{D}^x}\left(x \geq b_{\mathcal{S}} \wedge x < a^*\right) \\
&\leq \mathrm{P}_{x\sim\mathscr{D}^x}\left(a^* \leq x < b_1\right) + \mathrm{P}_{x\sim\mathscr{D}^x}\left(b_0 < x < a^*\right) \\
&\leq \mathrm{P}_{x\sim\mathscr{D}^x}\left(a^* \leq x < a_1\right) + \mathrm{P}_{x\sim\mathscr{D}^x}\left(a_0 < x < a^*\right) \\
&\leq \varepsilon.
\end{aligned}$$

This shows (8). Consequently,

$$\mathrm{P}_{\mathcal{S}\sim\mathscr{D}^n}\big(L_{\mathcal{S}}(h_{\mathcal{S}}) > \varepsilon\big) \leq \mathrm{P}_{\mathcal{S}\sim\mathscr{D}^n}(b_0 < a_0 \vee b_1 > a_1) \leq \mathrm{P}_{\mathcal{S}\sim\mathscr{D}^n}(b_0 < a_0) + \mathrm{P}_{\mathcal{S}\sim\mathscr{D}^n}(b_1 > a_1).$$

Note that

$$\begin{aligned}
\mathrm{P}_{\mathcal{S}\sim\mathscr{D}^n}(b_0 < a_0) &= 1 - \mathrm{P}_{\mathcal{S}\sim\mathscr{D}^n}(b_0 \geq a_0) = 1 - \mathrm{P}_{\mathcal{S}\sim\mathscr{D}^n}\big(\exists\, i \in \{1, \ldots, n\} \quad x_i \in [a_0, a^*)\big) \\
&= 1 - \Big(1 - \mathrm{P}_{\mathcal{S}\sim\mathscr{D}^n}\big(\forall\, i = 1, \ldots, n \quad x_i \notin [a_0, a^*)\big)\Big) \\
&= \Big(\mathrm{P}_{x\sim\mathscr{D}^x}\big(x \notin [a_0, a^*)\big)\Big)^n = \Big(1 - \mathrm{P}_{x\sim\mathscr{D}^x}\big(x \in [a_0, a^*)\big)\Big)^n \\
&= \left(1 - \frac{\varepsilon}{2}\right)^n.
\end{aligned}$$

Similarly, we can show that

$$\begin{aligned}
\mathrm{P}_{\mathcal{S}\sim\mathscr{D}^n}(b_1 > a_1) &= \mathrm{P}_{\mathcal{S}\sim\mathscr{D}^n}\big(\forall\, i = 1, \ldots, n \quad x_i \notin [a^*, a_1)\big) = \Big(\mathrm{P}_{x\sim\mathscr{D}^x}\big(x \notin [a^*, a_1)\big)\Big)^n \\
&= \left(1 - \frac{\varepsilon}{2}\right)^n.
\end{aligned}$$

Therefore,

$$P_{S\sim\mathscr{D}^n}\left(L_{\mathscr{D}}(h_S) > \varepsilon\right) \leq 2\left(1-\frac{\varepsilon}{2}\right)^n \leq 2\exp\left\{-m\frac{\varepsilon}{2}\right\} \overset{!}{\leq} \delta,$$

which yields $n_{\mathcal{H}}(\varepsilon,\delta) \leq \lceil 2\log\left(2/\delta\right)/\varepsilon \rceil$, as claimed.

Hence we have seen that there are infinite hypothesis classes that are PAC learnable, but Theorem 1.23 also shows that too "complex" infinite classes are not PAC learnable. The measure of complexity of a hypothesis class that is most convenient is the *Vapnik–Chervonenkis (VC) dimension.*

Definition 1.27. Let $\mathcal{H}$ be a hypothesis class of functions $\mathcal{X} \to \{-1,1\}$. Let $\mathcal{C} = \{c_1,\dots,c_s\} \subset \mathcal{X}$, $s \in \mathbb{N}$. The *restriction of* $\mathcal{H}$ *to* $\mathcal{C}$ is the set of functions $\mathcal{C} \to \{-1,1\}$ that can be derived from $\mathcal{H}$, i.e.

$$\mathcal{H}_{\mathcal{C}} := \left\{ \left(h(c_1),\dots,h(c_s)\right) \,\middle|\, h \in \mathcal{H} \right\}.$$

Definition 1.28. A hypothesis class $\mathcal{H}$ *shatters* a finite set $\mathcal{C} \subset \mathcal{X}$ if

$$|\mathcal{H}_{\mathcal{C}}| = 2^{|\mathcal{C}|}.$$

Remark 1.29. In other words, $\mathcal{H}$ shatters $\mathcal{C}$ if $\mathcal{H}_{\mathcal{C}}$ is the set of all possible classifiers that can be defined on $\mathcal{C}$.

Definition 1.30. The *Vapnik–Chervonenkis (VC) dimension* of a hypothesis class $\mathcal{H}$, denoted by $\mathrm{VCdim}(\mathcal{H})$, is the maximal size of a set $\mathcal{C} \subset \mathcal{X}$ that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter subsets of any size, then we say that $\mathcal{H}$ has infinite VC dimension.

Remark 1.31. Note that if $\mathcal{H}$ shatters a subset $\mathcal{C}$, then it shatters any subset $\mathcal{C}' \subset \mathcal{C}$. Hence, to show that $\mathrm{VCdim}(\mathcal{H}) \leq d$, we need to show that $\mathcal{H}$ does not shatter any subset $\mathcal{C}$ of size $d+1$. To show that $\mathrm{VCdim}(\mathcal{H}) \geq d$, it is enough to find a subset $\mathcal{C}$ of size d that is shattered by $\mathcal{H}$. Consequently, in order to prove that $\mathrm{VCdim}(\mathcal{H}) = d$, we need to show that

1) $\exists\, \mathcal{C} \subset \mathcal{X}$ of size d that is shattered by $\mathcal{H}$;

2) $\forall\, \mathcal{C} \subset \mathcal{X}$ of size $d+1$ holds that $\mathcal{H}$ does not shatter $\mathcal{C}$.

Remark 1.32. If $\mathcal{H}$ is finite the following estimate is non-trivial: For any $\mathcal{C} \subset \mathcal{X}$ we have $|\mathcal{H}_{\mathcal{C}}| \leq |\mathcal{H}|$. If $\mathcal{C}$ is such that $|\mathcal{H}| < 2^{|\mathcal{C}|}$, then $|\mathcal{H}_{\mathcal{C}}| < 2^{|\mathcal{C}|}$, which implies that $\mathcal{H}$ cannot shatter $\mathcal{C}$. Since this is true for any $\mathcal{C} \subset \mathcal{X}$ with $|\mathcal{C}| > \log_2(|\mathcal{H}|)$, this implies that

$$\mathrm{VCdim}(\mathcal{H}) \leq \log_2\left(|\mathcal{H}|\right).$$

What if $\mathcal{H}$ is very large or even infinite but $\mathcal{X}$ is finite? Then we can reduce $\mathcal{H}$ to have size $2^{|\mathcal{X}|}$ and consequently

$$\mathrm{VCdim}(\mathcal{H}) \leq \log_2\left(2^{|\mathcal{X}|}\right) = |\mathcal{X}|.$$

Consequently, it holds that

$$\mathrm{VCdim}(\mathcal{H}) \leq \min\Big\{|\mathcal{X}|, \log_2\big(|\mathcal{H}|\big)\Big\}.$$

However, the difference $\min(|\mathcal{X}|, \log_2(|\mathcal{H}|)) - \mathrm{VCdim}(\mathcal{H})$ can be quite large, for example for the class of thresholds on $\mathcal{X} = \{1, \dots, k\}$. There, $\mathrm{VCdim}(\mathcal{H}) = 1$, but $\log_2(|\mathcal{H}|) = |\mathcal{X}| = k \to \infty$.

Example 1.33. Consider again the class of thresholds

$$\mathcal{H} := \Big\{x \mapsto \mathbf{1}\,(x < a) \,\Big|\, a \in \mathbb{R}\Big\}.$$

Let $\mathcal{C} = \{c\}$ for some fixed $c \in \mathbb{R}$. There are 2 classifiers on $\mathcal{C}$: $f_0(c) = 0$ and $f_1(c) = 1$. Furthermore, $\mathcal{H}_\mathcal{C} = \{0, 1\}$, as 0 is obtained by $a = c - 1$ and 1 by $a = c + 1$. Hence $\mathcal{H}$ shatters $\mathcal{C}$.
Consider now $\mathcal{C} = \{c_1, c_2\}$ for $c_1 < c_2 \in \mathbb{R}$. There are 4 possible classifiers, but $\mathcal{H}_\mathcal{C} = \big\{(0,0), (1,0), (1,1)\big\}$. Hence $\mathcal{H}$ does not shatters any subset $\mathcal{C} \subset \mathcal{X}$ of size 2.
Together, this implies that $\mathcal{H}$ has VC dimension 1.

Example 1.34. Consider $\mathcal{X} = \mathbb{R}$ and the class of intervals

$$\mathcal{H} = \Big\{x \mapsto \mathbf{1}\big(x \in (a, b)\big) \,\Big|\, a < b \in \mathbb{R}\Big\}.$$

Then $\mathcal{H}$ shatters any subset of size 2, but not $\mathcal{C} = \{c_1, c_2, c_3\}$ with $c_1 < c_2 < c_3$, as $h_{a,b}(c_1) = h_{a,b}(c_3) = 1$ implies $h_{a,b}(c_2)$. Hence $\mathrm{VCdim}(\mathcal{H}) = 2$.

Corollary 1.35. *Let $\mathcal{H}$ be some hypothesis class and n be the size of a training set $\mathcal{S}$. Assume that there exists a subset $\mathcal{C} \subset \mathcal{X}$ such that $|\mathcal{C}| = 2n$ and $\mathcal{C}$ is shattered by $\mathcal{H}$. Then, for any learning algorithm $\mathscr{A}$, we can find a distribution $\mathscr{D}$ on $\mathcal{X} \times \{-1, 1\}$ such that*

1) There exists a classifier $f : \mathcal{X} \to \mathcal{Y}$ with $L_\mathscr{D}(f) = 0$ and

2) $\mathrm{P}_{\mathcal{S} \sim \mathscr{D}^n}\big(L_\mathscr{D}(\mathscr{A}(\mathcal{S})) \geq 1/8\big) \geq 1/7$.

Proof. This immediately follows from the proof of Theorem 1.23, as the shattering allows us to construct any possible classifier f. □

Theorem 1.36. *Let $\mathcal{H}$ be some hypothesis class with infinite VC dimension. Then $\mathcal{H}$ is not PAC-learnable.*

Proof. If $\mathcal{H}$ has infinite VC dimension, then for any $n \in \mathbb{N}$ we can find a set $\mathcal{C} \subset \mathcal{X}$ such that $|\mathcal{C}| = n$ and that $\mathcal{C}$ is shattered by $\mathcal{H}$. Then the claim follows by Corollary 1.35. □

1.5 The Fundamental Theorem of PAC Learning

Motivation. We have seen that infinite VC dimension implies non-PAC-learnability (Theorem 1.36). Theorem 1.41 will show us that finite VC dimension is indeed also sufficient for PAC learnability. For its proof, we need the following concept.

Definition 1.37. Let $\mathcal{H}$ be a hypothesis class. The *growth function* of $\mathcal{H}$, denoted $\tau_{\mathcal{H}}: \mathbb{N} \to \mathbb{N}$, is defined as

$$\tau_{\mathcal{H}}(n) := \max_{C \subset X : |C| = n} |\mathcal{H}_C|.$$

Remark 1.38. In words, $\tau_{\mathcal{H}}(n)$ is the number of different functions from a subset C of size n to $\{-1, 1\}$, that can be obtained by restricting $\mathcal{H}$ to C. Obviously, if $\mathrm{VCdim}(\mathcal{H}) = d$ then for any $n \leq d$ we have $\tau_{\mathcal{H}}(n) = 2^n$. In such cases, $\mathcal{H}$ induces all possible functions from C to $\{-1, 1\}$.

Lemma 1.39 (Sauer's Lemma)**.** *Let $\mathcal{H}$ be a hypothesis class with $\mathrm{VCdim}(\mathcal{H}) \leq d < \infty$. Then, for all $n \in \mathbb{N}$, holds*

$$\tau_{\mathcal{H}}(n) \leq \sum_{i=0}^{d} \binom{n}{i}.$$

In particular, if $n \geq d$, then $\tau_{\mathcal{H}}(n) \leq (en/d)^d$.

Proof. We will prove the lemma by showing the stronger claim: for $C := \{c_1, \ldots, c_n\}$ we have

$$\forall\, \mathcal{H} \quad |\mathcal{H}_C| \leq \left|\{B \subset C \mid \mathcal{H} \text{ shatters } B\}\right|. \tag{9}$$

(9) is sufficient to prove the lemma because if $\mathrm{VCdim}(\mathcal{H}) \leq d$, then no set with size $> d$ is shattered by $\mathcal{H}$ and therefore

$$\left|\{B \subset C \mid \mathcal{H} \text{ shatters } B\}\right| \leq \sum_{i=0}^{d} \binom{n}{i}.$$

If $n \geq d$, then we can further bound

$$\sum_{i=0}^{d} \binom{n}{i} \leq \left(\frac{en}{d}\right)^d,$$

as

$$\left(\frac{d}{n}\right)^d \sum_{i=0}^{d} \binom{n}{i} \leq \sum_{i=0}^{d} \binom{n}{i} \left(\frac{d}{n}\right)^i \leq \sum_{i=0}^{n} \binom{n}{i} \left(\frac{d}{n}\right)^i = \left(1 + \frac{d}{n}\right)^n \leq \exp\{d\}.$$

Hence we only need to show (9), which will be done by induction over n. For n = 1, (9) is trivially fulfilled, as for any $\mathcal{H}$, both sides are either equal to one or two (as the empty set is always shattered by $\mathcal{H}$).

Assume (9) holds for sets of size $k < n$ and let us prove it for sets of size n. To this end, fix $\mathcal{H}$ and $\mathcal{C} \coloneqq \{c_1, \dots, c_m\}$. Denote $\mathcal{C}' \coloneqq \{c_2, \dots, c_m\}$ and define

$$\mathcal{Y}_0 \coloneqq \big\{(y_2, \dots, y_n) \,\big|\, (0, y_2, \dots, y_n) \in \mathcal{H}_\mathcal{C} \vee (1, y_2, \dots, y_n) \in \mathcal{H}_\mathcal{C}\big\},$$
$$\mathcal{Y}_1 \coloneqq \big\{(y_2, \dots, y_n) \,\big|\, (0, y_2, \dots, y_n) \in \mathcal{H}_\mathcal{C} \wedge (1, y_2, \dots, y_n) \in \mathcal{H}_\mathcal{C}\big\}.$$

Clearly, $|\mathcal{H}_\mathcal{C}| = |\mathcal{Y}_0| + |Y_1|$ and, since $\mathcal{Y}_0 = \mathcal{H}_{\mathcal{C}'}$, using the induction assumption (applied on $\mathcal{H}$ and $\mathcal{C}'$) implies

$$|\mathcal{Y}_0| = |\mathcal{H}_{\mathcal{C}'}| \le \big|\{\mathcal{B} \subset \mathcal{C} \,|\, \mathcal{H} \text{ shatters } \mathcal{B}\}\big| = \Big|\big\{\mathcal{B} \subset \mathcal{C} \,\big|\, c_1 \notin \mathcal{B} \wedge \mathcal{H} \text{ shatters } \mathcal{B}\big\}\Big|.$$

Next, let

$$\mathcal{H}' \coloneqq \big\{h \in \mathcal{H} \,\big|\, \exists\, h' \in \mathcal{H} \quad (1 - h'(c_1), h'(c_2), \dots, h'(c_n) = (h(c_1), \dots, h(c_n)\big\} \subset \mathcal{H}.$$

$\mathcal{H}'$ contains pairs of hypotheses that agree in $\mathcal{C}'$ and differ in c_1. Consequently, if $\mathcal{H}'$ shatters a subset $\mathcal{B} \subset \mathcal{C}$, then it also shatters $\mathcal{B} \cup \{c_1\}$ and vice versa. Using $\mathcal{Y}_1 = \mathcal{H}'_{\mathcal{C}'}$ and applying the inductive assumption on $\mathcal{H}'$ and $\mathcal{C}'$, we obtain

$$\begin{aligned}|\mathcal{Y}_1| = \big|\mathcal{H}'_{\mathcal{C}'}\big| &\le \Big|\big\{\mathcal{B} \subset \mathcal{C}' \,\big|\, \mathcal{H}' \text{ shatters } \mathcal{B}\big\}\Big| = \Big|\big\{\mathcal{B} \subset \mathcal{C}' \,\big|\, \mathcal{H}' \text{ shatters } \mathcal{B} \cup \{c_1\}\big\}\Big| \\ &= \Big|\big\{\mathcal{B} \subset \mathcal{C} \,\big|\, c_1 \in \mathcal{B} \wedge \mathcal{H}' \text{ shatters } \mathcal{B}\big\}\Big| \le \big|\{\mathcal{B} \subset \mathcal{C} \,|\, c_1 \in \mathcal{B} \wedge \mathcal{H} \text{ shatters } \mathcal{B}\}\big|.\end{aligned}$$

Hence we have shown that

$$\begin{aligned}|\mathcal{H}_\mathcal{C}| &= |\mathcal{Y}_0| + |\mathcal{Y}_1| \\ &\le \big|\{\mathcal{B} \subset \mathcal{C} \,|\, c_1 \notin \mathcal{B} \wedge \mathcal{H} \text{ shatters } \mathcal{B}\}\big| + \big|\{\mathcal{B} \subset \mathcal{C} \,|\, c_1 \in \mathcal{B} \wedge \mathcal{H} \text{ shatters } \mathcal{B}\}\big| \\ &= \big|\{\mathcal{B} \subset \mathcal{C} \,|\, \mathcal{H} \text{ shatters } \mathcal{B}\}\big|,\end{aligned}$$

which concludes the proof. □

Theorem 1.40. *Let $\mathcal{H}$ be some hypothesis class with growth function $\tau_\mathcal{H}$. Also, suppose that l maps to $[0,1]$. Then for all $\delta > 0$ holds*

$$\mathrm{P}_{\mathcal{S} \sim \mathscr{D}^n}\left(\sup_{h \in \mathcal{H}} \big|L_\mathscr{D}(h) - L_\mathcal{S}(h)\big| \le \frac{4 + \sqrt{\log\big(2\tau_\mathcal{H}(2n)\big)}}{\delta\sqrt{2n}}\right) \ge 1 - \delta.$$

Theorem 1.41 (Fundamental Theorem of PAC learning)**.** *Let $\mathcal{H}$ be some hypothesis class of functions $\mathcal{X} \to \{-1, 1\}$ and let the loss function be the 0-1-loss. Then, the following assertions are equivalent:*

1) $\mathcal{H}$ has the uniform convergence property;

2) any ERM rule is a successful agnostic PAC learner for $\mathcal{H}$;

3) $\mathcal{H}$ is agnostic PAC learnable;

4) $\mathrm{VCdim}(\mathcal{H}) < \infty$.

If, additionally, the realizability assumption (1) *holds, then all the previous assertions are equivalent to*

5) $\mathcal{H}$ *is PAC learnable;*

6) any ERM rule is a successful PAC learner for $\mathcal{H}$*;*

Proof. Let us first summarize the results that we have already proven in Figure 1. We can see that we are only left to prove that finite VC dimension implies the uniform convergence property.

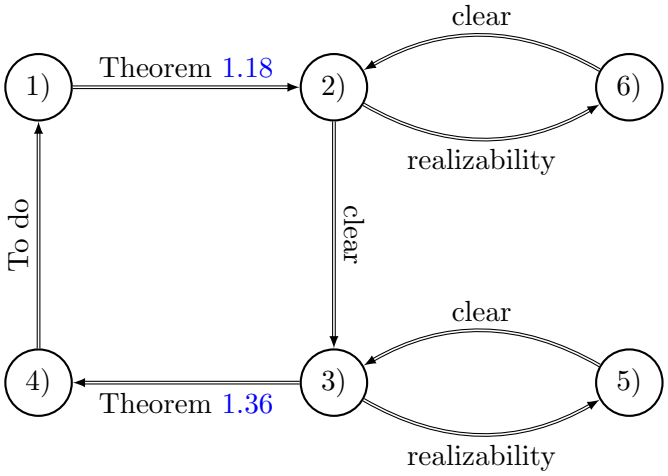


Figure 1: Logical structure for the proof the Theorem 1.41

To see this, note that from Sauer's lemma (Lemma 1.39) it follows that for all $n \geq d/2$ holds $\tau_{\mathcal{H}} \leq (en/d)^d$. Combining this with Theorem 1.40 yields

$$\forall\, \delta > 0 \quad \mathrm{P}_{S\sim\mathscr{D}^n}\left(\sup_{h\in\mathcal{H}}\left|L_{\mathscr{D}}(h) - L_S(h)\right| \leq \frac{4 + \sqrt{d\log\left(2en/d\right)}}{\delta\sqrt{2n}}\right) \geq 1 - \delta.$$

If n is large enough such that $\sqrt{d\log\left(2en/d\right)} \geq 4$, then

$$\forall\, \delta > 0 \quad \mathrm{P}_{S\sim\mathscr{D}^n}\left(\sup_{h\in\mathcal{H}}\left|L_{\mathscr{D}}(h) - L_S(h)\right| \leq \frac{2\sqrt{d\log\left(2en/d\right)}}{\delta\sqrt{2n}}\right) \geq 1 - \delta.$$

Note that

$$\frac{2\sqrt{d\log\left(2en/d\right)}}{\delta\sqrt{2n}} \leq \varepsilon \iff n \geq \frac{2d}{(\varepsilon\delta)^2}\log(n) + \frac{2d\log\left(2e/d\right)}{(\varepsilon\delta)^2}.$$

Recall that for all $\alpha, x > 0$ holds $\log(x) \le \alpha x - \log(\alpha) - 1$ (by convexity, $x \mapsto \alpha x - \log(\alpha) - 1$ is the tangent to $x \mapsto \log(x)$ at $x = 1/\alpha$). Then

$$\frac{2d}{(\varepsilon\delta)^2}\log(n) \le \frac{2d}{(\varepsilon\delta)^2}\left(\frac{(\varepsilon\delta)^2}{4d}n - \log\left(\frac{(\varepsilon\delta)^2}{4d}\right) - 1\right) = \frac{n}{2} - \frac{2d}{(\varepsilon\delta)^2}\log\left(\frac{(\varepsilon\delta)^2}{4d}\right) - \frac{2d}{(\varepsilon\delta)^2}.$$

Consequently, it suffices to take n large enough such that

$$\frac{n}{2} \ge -\frac{2d}{(\varepsilon\delta)^2}\log\left(\frac{(\varepsilon\delta)^2}{4d}\right) - \frac{2d}{(\varepsilon\delta)^2} + \frac{2d\log\big(2e/d\big)}{(\varepsilon\delta)^2}$$

and thus choose

$$n \ge \begin{cases} \frac{4}{(\varepsilon\delta)^2}\log\left(\frac{4}{(\varepsilon\delta)^2}\right) + \frac{4\big(\log(2e)-1\big)}{(\varepsilon\delta)^2} & \text{if } d = 1, \\ \frac{4}{(\varepsilon\delta)^2}\log\left(\frac{4}{(\varepsilon\delta)^2}\right) & \text{if } d \ge 2. \end{cases}$$

Hence $\mathcal{H}$ has the uniform property. □

Remark 1.42. The proof of Theorem 1.41 implies that the sample complexity satisfies

$$n_{\mathcal{H}}^{\mathrm{uc}} \asymp \log\big(1/(\varepsilon\delta)\big)/(\varepsilon\delta)^2.$$

However, it is possible to obtain sharper bounds (Theorem 6.8 on p.48):

Theorem 1.43 (Fundamental Theorem of PAC learning – quantitative version)**.** *Let $\mathcal{H}$ be a hypothesis for binary classification and let the loss function be the 0-1-loss. Assume that $\mathrm{VCdim}(\mathcal{H}) \le d < \infty$. Then, there exists constants $C_1, C_2 > 0$ such that*

1. *$\mathcal{H}$ has the uniform convergence property with sample complexity*

$$C_1\frac{d + \log\big(1/\delta\big)}{\varepsilon^2} \le n_{\mathcal{H}}^{uc} \le C_2\frac{d + \log\big(1/\delta\big)}{\varepsilon^2}; \tag{10}$$

2. *$\mathcal{H}$ is agnostic PAC learnable with sample complexity*

$$C_1\frac{d + \log\big(1/\delta\big)}{\varepsilon^2} \le n_{\mathcal{H}} \le C_2\frac{d + \log\big(1/\delta\big)}{\varepsilon^2}; \tag{11}$$

3. *If the realizability assumption (1) holds, then $\mathcal{H}$ is PAC learnable with sample complexity*

$$C_1\frac{d + \log\big(1/\delta\big)}{\varepsilon} \le n_{\mathcal{H}}^{uc} \le C_2\frac{d\log\big(1/\varepsilon\big) + \log\big(1/\delta\big)}{\varepsilon}. \tag{12}$$

Remark 1.44. It is important to note the following:

1) The Theorems characterising PAC-learnability of infinite classes (Theorems 1.41 and 1.43) only hold for the 0-1-loss function. They can be easily extended to arbitrary *bounded* loss functions, however no statement was made for unbounded loss functions. In contrast, the Theorems for finite classes (Theorems 1.9 and 1.21) hold for arbitrary loss functions.

2) Additionally, we can see that the sample complexity under the realizability assumption (1) scales as ε^{-1} (Theorem 1.9 for finite classes and Theorem 1.43 statement (12) for infinite classes), whereas without the realizability assumption, the sample complexity scales as ε^{-2} (Theorem 1.21 for finite classes and Theorem 1.43 statement (11) for infinite classes). This means that in order the half the error ε, we need to double the sample size if the realizability assumption holds true, and to quadruple it otherwise.

Both points origin in the use of Hoeffding's inequality (Lemma 1.19) for infinite hypothesis classes.

2 Linear Predictors

Motivation. Consider the following class

$$L_d := \left\{ h_{w,b} \,\middle|\, w \in \mathbb{R}^d, b \in \mathbb{R} \right\}, \quad h_{w,b} := \begin{cases} \mathbb{R}^d \to \mathbb{R}, \\ x \mapsto \langle w, x \rangle + b. \end{cases}$$

We call w the vector of *weights* and b the *bias*. Given a function $\varphi \colon \mathbb{R} \to \mathcal{Y}$, we can consider the class

$$\varphi \circ L_d := \left\{ \varphi \circ h_{w,b} \,\middle|\, h_{w,b} \in L_d \right\}.$$

Example 2.1. Let $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{-1, 1\}$ and $\varphi(t) = \operatorname{sign}(t)$. The corresponding class is

$$\mathcal{HS}_d := \operatorname{sign} \circ L_d := \left\{ \operatorname{sign}\left(\langle w, x \rangle + b \right) \,\middle|\, w \in \mathbb{R}^d, b \in \mathbb{R} \right\}$$

and is often used in binary classification. It can be shown that $\mathcal{HS}_d$ has VC dimension $d+1$ and hence is agnostic PAC learnable using the ERM paradigm with sample complexity of order $(d + \log(1/\delta))/\varepsilon^2$ by (11).

Remark 2.2. The bias b can be incorporated into the weight vector by setting

$$w' := (b, w_1, \ldots, w_d)^\top \text{ and } x' := (1, x_1, \ldots, x_d)^\top$$

to obtain the *homogeneous representation*.

2.1 Classification

Let us first consider the task of binary classification, i.e. $\mathcal{Y} = \{-1, 1\}$. In the first part of this subsection we will assume that there exists a labelling function $f: \mathbb{R}^d = \mathcal{X} \to \mathcal{Y} = \{-1, 1\}$ with $f(x) = \operatorname{sign}\big(\langle w^*, x\rangle\big)$ for some $w^* \in \mathbb{R}^d$. This is called the *separable case* and implies that the realizability assumption (1) is satisfied. Recall that this implies that any ERM predictor h_S satisfies $L_S(h_S) = 0$ (Remark 1.6). In the following, we will discuss three independent methods for finding the ERM predictor h_S given a sample S of size n. The first two (Linear Programming and the Perceptron, 2.1.1 and 2.1.2) apply only to the separable case, whereas the third (Logistic Regression, 2.1.3) is more general. In the case of non-separability also other methods like support vector machines should be considered.

2.1.1 Linear Programming

Here, we will be using the concept of *linear programming* in order to find an ERM given linearly separable data $S = \big\{(x_1, y_1), \dots, (x_n, y_n)\big\} \in (\mathbb{R}^d \times \{-1, 1\})^n$.

Definition 2.3. A *linear program (LP)* aims at solving the optimization problem

$$\max_{w \in \mathbb{R}^d} \langle u, w\rangle \quad \text{s.t. } Aw \geq v,$$

where $u \in \mathbb{R}^d, v \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m \times d}$ are given.

Lemma 2.4. *Given linearly separable data* $S = \big\{(x_1, y_1), \dots, (x_n, y_n)\big\} \in (\mathbb{R}^d \times \{-1, 1\})^n$, *the linear program*

$$\max_{w \in \mathbb{R}^d} \langle 0, w\rangle \quad \textit{s.t. } Aw \geq v \tag{13}$$

with $v = (1, \dots, 1)^\top \in \mathbb{R}^n$ *and*

$$A := y \odot \begin{bmatrix} x_1 & \dots & x_n \end{bmatrix}^\top = \begin{pmatrix} y_1(x_1)_1 & \dots & y_1(x_1)_d \\ \vdots & \ddots & \vdots \\ y_n(x_n)_1 & \dots & y_n(x_n)_d \end{pmatrix} \in \mathbb{R}^{n \times d},$$

where $\odot$ *denotes the* componentwise *or* Hadamard product *and* $y := (y_1, \dots, y_n)^\top$ *has a solution* $\bar{w}$ *which separates the data according to*

$$\forall\, i = 1, \dots, n \quad y_i \langle \bar{w}, x_i\rangle \geq 1. \tag{14}$$

Proof. As any solution to (13) satisfies (14) and vice versa, it suffices to show the existence of a solution to (14).
We proceed as follows: Set $\gamma := \min_{i=1,\dots,n} y_i \langle w^*, x_i\rangle$ and $\bar{w} := w^*/\gamma$, where w^* denotes the true data separator which exists as the data is separable but is unknown to us. Then, for all $i = 1, \dots, n$ holds

$$y_i \langle \bar{w}, x_i\rangle = \frac{1}{\gamma} y_i \langle w, x_i\rangle \geq 1. \qquad \square$$

2.1.2 Perceptron

The *Perceptron algorithm* is an iterative method that produces a finite sequence of vectors $w^{(1)}, \dots, w^{(T)}$, whose last element yields a perfect (recall that we are in the separable case) separation of the training data.
The algorithm runs as follows:

Algorithm 1 Perceptron Algorithm

1: Set $t := 1$ and $w^{(1)} := 0 \in \mathbb{R}^d$;
2: **while** $w^{(t)}$ is not a perfect classifier **do**
3: Find $i \in \{1, \dots, n\}$ such that (x_i, y_i) is mislabeled, i.e. $y_i \langle w^{(t)}, x_i \rangle \leq 0$;
4: Update $w^{(t+1)} = w^{(t)} + y_i x_i$;
5: Update $t = t + 1$.
6: **end while**
7: **return** $w^{(t)}$.

Note that the update of the perceptron algorithm guides the sequence towards a more correct labelling:

$$y_i \left\langle w^{(t+1)}, x_i \right\rangle = y_i \left\langle w^{(t)}, x_i \right\rangle + \|x_i\|^2 \geq y_i \left\langle w^{(t)}, x_i \right\rangle.$$

Theorem 2.5. *Assume separability with weight vector $w^\star$ and let*

$$B := \min\left\{ \|w\| \;\middle|\; \forall\, i = 1, \dots, n \quad y_i \langle w, x_i \rangle \geq 1 \right\} \text{ and } R := \max_{i=1,\dots,n} \|x_i\|.$$

Then, the perceptron algorithm stops after $\lceil (RB)^2 \rceil$ iterations at a perfect classifier.

Proof. For $t \in \mathbb{N}$, we will show that if at iteration t, $w^{(t)}$ mislabels some example (x_i, y_i), then we must have

$$\frac{\left\langle w^\star, w^{(t+1)} \right\rangle}{\|w^\star\| \left\| w^{(t+1)} \right\|} \geq \frac{\sqrt{t}}{RB}. \tag{15}$$

First assume that (15) holds. Then, by Cauchy-Schwarz,

$$1 \geq \frac{\left\langle w^\star, w^{(t+1)} \right\rangle}{\|w^\star\| \left\| w^{(t+1)} \right\|} \geq \frac{\sqrt{t}}{RB}$$

and consequently $t \leq (RB)^2$, which would conclude the proof.
It therefore only remains to show (15). To this end, recall that $w^{(1)} = 0 \in \mathbb{R}^d$ and $w^{(2)} = w^{(1)} + y_i x_i$ for some mislabelled pair (x_i, y_i). Suppose that for $t \geq 3$ we have $\langle w^\star, w^{(t)} \rangle \geq t - 1$ and let $i \in \{1, \dots, n\}$ be such that (x_i, y_i) is mislabelled, then

$$\left\langle w^\star, w^{(t+1)} \right\rangle = \left\langle w^\star, w^{(t)} + y_i x_i \right\rangle \geq t - 1 + 1 = t.$$

Therefore, it follows by induction that if $w^{(t)}$ is not a perfect classifier, then $\langle w^*, w^{(t)} \rangle \geq t$. Furthermore,

$$\left\| w^{(t+1)} \right\|^2 = \left\| w^{(t)} + y_i x_i \right\|^2 = \left\| w^{(t)} \right\|^2 + \underbrace{2\left\langle w^{(t)}, y_i x_i \right\rangle}_{\leq 0} + \|x_i\|^2 \leq \left\| w^{(t)} \right\|^2 + R^2$$

and hence for all $t < T$ holds

$$\|w^{(t+1)}\|^2 = \sum_{j=1}^{t} \left(\|w^{(j+1)}\|^2 - \|w^{(j)}\|^2 \right) \leq R^2 t.$$

Therefore,

$$\frac{\left\langle w^*, w^{(t+1)} \right\rangle}{\|w^*\| \|w^{(t+1)}\|} \geq \frac{t}{BR\sqrt{t}} = \frac{\sqrt{t}}{BR}.$$ □

Remark 2.6. The constants B and R from Theorem 2.5 have the following intuitive interpretations:

- $R = \max_{i=1,\dots,n} \|x_i\|$ is a measure of the compactness of the sample data. The idea is that spread out data tend to be harder to separate than compact data. R is known to the observer.
- $B = \min\left\{ \|w\| \;\middle|\; \forall\, i = 1, \dots, n \quad y_i \langle w, x_i \rangle \geq 1 \right\}$ describes the margin that can be achieved by a perfect classifier. To see this, note that for $w \in \mathbb{R}^d$ and $i = 1, \dots, n$ the distance between x_i and the hyperplane defined by $w^\top x = 0$ is given by

  $$\operatorname{dist}(w, x_i) = \frac{\langle w, x_i \rangle}{\|w\|}.$$

  Hence $\langle w, x_i \rangle \geq 1$ for all i is equivalent to $\operatorname{dist}(w, x_i) \geq 1/\|w\|$ and consequently the margin $M(w, \{x_1, \dots, x_n\})$ of the perfect separator given by $w^\top x = 0$ and the data satisfies

  $$M(w, \{x_1, \dots, x_n\}) := \min_{i=1,\dots,n} \operatorname{dist}(w, x_i) \geq \frac{1}{\|w\|}.$$

  Hence, we can conclude that

  $$B = \frac{1}{M(w, \{x_1, \dots, x_n\})},$$

  i.e. B is the inverse of the margin and hence a measure of how difficult the separation of the data is. However, note that B is unknown to the observer.

Remark 2.7. Note that introducing a learning rate $\eta > 0$ to the perceptron only changes the length of the solution vector w and hence neither has an effect on the convergence speed, nor on the final predictions.

2.1.3 Logistic regression

A natural modification in order to deal with predictions that are not -1 or 1 using a linear model is to use the *logistic function*

$$\varphi\colon \mathbb{R} \to [0,1], \quad x \mapsto \frac{1}{1+\exp\{-x\}}$$

and interpret the result $\varphi(\langle w, x\rangle)$ as probability of class 1 to be true given x. Hence the hypothesis class is given by

$$\mathcal{H}_{\text{sig}} := \left\{ h_w\colon\ x \mapsto \frac{1}{1+\exp\left\{-\langle w, x\rangle\right\}} \;\middle|\; w \in \mathbb{R}^d \right\}.$$

Our loss function should penalize misclassifications, i.e. small values of $y_i\langle w, x_i\rangle$ and should therefore be increasing in $y_i\langle w, x_i\rangle$. A possible choice is the *logistic loss*

$$l(h_w, z) := l(h_w, (x,y)) := \log\left(1+\exp\left\{-y\langle w, x\rangle\right\}\right).$$

The ERM rule then requires us, given a sample $\mathcal{S}$ of size n, to find an estimator $h_{\mathcal{S}} \in \mathcal{H}_{\text{sig}}$ satisfying

$$h_{\mathcal{S}} \in \operatorname*{argmin}_{h\in\mathcal{H}_{\text{sig}}} L_{\mathcal{S}}(h_{\mathcal{S}}),$$

which can be done by taking $h_{\mathcal{S}} = h_{w^*}$ for

$$w^* \in \operatorname*{argmin}_{w\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \log\left(1+\exp\left\{-y_i\langle w, x_i\rangle\right\}\right).$$

Remark 2.8. Note that for $w \in \mathbb{R}^d$

$$\exp\left\{-\langle w, x\rangle\right\} = \frac{1}{h_w(x)} - 1 = \frac{1-h_w(x)}{h_w(x)}$$

are the *odds* of class 1. Equivalently,

$$-\langle w, x\rangle = \log\left(\frac{1-h_w(x)}{h_w(x)}\right) =: \operatorname{logit}\left(h_w(x)\right)$$

are the *log-odds*. Hence

$$l(h_w, (x,y)) = \begin{cases} \log\left(1/h_w(x)\right) & \text{if } y = 1, \\ \log\left(1/(1-h_w(x))\right) & \text{if } y = -1. \end{cases}$$

2.2 Linear Regression

In the case of linear regression we have $\mathcal{X} = \mathbb{R}^r, \mathcal{Y} = \mathbb{R}^q$ and we assume

$$Y_i = X_i\beta + \varepsilon_i \in \mathbb{R}^q, \quad i = 1, \dots, n,$$

for $\beta \in \mathbb{R}^d, X_i \in \mathbb{R}^{q\times d}$ of maximal rank $\min\{q, d\}$, $r, q, d \in \mathbb{N}$ and $\varepsilon \sim N(0, \Sigma)$ i.i.d. and $\Sigma \in \mathbb{R}^{q\times q}$ positive definite. Often $X_i := \Psi(x_i)$ for some fixed, known *feature map* $\Psi\colon \mathbb{R}^r \to \mathbb{R}^{q\times d}$. The hypothesis space is given by

$$\mathcal{H} := \left\{ x \mapsto \Psi(x)\beta \;\middle|\; \Psi\colon \mathcal{X} \to \mathbb{R}^{q\times d}, \beta \in \mathbb{R}^d \right\}.$$

It is very convenient to directly work in the feature space $\Psi(\mathcal{X})$. Note that the log-likelihood given an i.i.d. sample $\mathcal{S} = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ is given by

$$\mathcal{L}(\beta|\mathcal{S}) = \sum_{i=1}^n \underbrace{\log\left((2\pi)^{q/2}|\Sigma|^{-1/2}\right)}_{=\text{const}} - \frac{1}{2}\left\|\Sigma^{-1/2}\left(Y_i - X_i\beta\right)\right\|^2.$$

Hence, maximizing the log-likelihood is equivalent to minimizing the squared error:

$$\beta^{\text{MLE}} = \operatorname*{argmin}_{\beta\in\mathbb{R}^d} \sum_{i=1}^n \left\|\Sigma^{-1/2}\left(Y_i - X_i\beta\right)\right\|^2. \tag{16}$$

Note that changing the distribution of the residuals ε_i would yield a different loss function. For example, for exponentially distributed ε_i we would minimize the mean absolute error. We now solve (16) by means of standard vector calculus:

$$\nabla_\beta \left\{ \sum_{i=1}^n \left\|\Sigma^{-1/2}\left(Y_i - X_i\beta\right)\right\|^2 \right\} = \sum_{i=1}^n \left(-2X_i^\top\Sigma^{-1}Y_i + 2X_i^\top\Sigma^{-1}X_i\beta\right);$$
$$\nabla_\beta^2 \left\{ \sum_{i=1}^n \left\|\Sigma^{-1/2}\left(Y_i - X_i\beta\right)\right\|^2 \right\} = 2\sum_{i=1}^n X_i^\top\Sigma^{-1}X_i.$$

Note that as X_i has full rank and Σ^{-1} is positive definite, $X_i^\top\Sigma^{-1}X_i$ is positive definite and consequently

$$\beta^{\text{MSE}} = \beta^{\text{MLE}} := \left(\sum_{i=1}^n X_i^\top\Sigma^{-1}X_i\right)^{-1}\left(\sum_{i=1}^n X_i^\top\Sigma^{-1}Y_i\right) \tag{17}$$

is the unique solution to (16).

Remark 2.9. If $q = 1$, then it is convenient to use the model

$$Y = X\beta + \varepsilon \in \mathbb{R}^n, \text{ where } X := \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \in \mathbb{R}^{n\times d} \text{ and } Y := (Y_1, \dots, Y_n)^\top.$$

This approach allows to replace the independence assumption on the samples by a joint normal distribution by assuming $\varepsilon \sim N(0, \Sigma)$, $\Sigma \in \mathbb{R}^{n\times n}$ positive definite. The least squares / maximum likelihood solution to optimal weights β is then given by

$$\beta^{\mathrm{MLE}} = \left(X^\top \Sigma^{-1} X\right)^{-1} X^\top \Sigma^{-1} Y. \tag{18}$$

Example 2.10. If $q = 1$ and $\Sigma = \sigma^2 \mathbf{I}$, then (17) reduces to

$$\beta^{\mathrm{MLE}} = \left(\frac{1}{\sigma^2}\sum_{i=1}^n X_i^\top X_i\right)^{-1}\left(\frac{1}{\sigma^2}\sum_{i=1}^n X_i^\top Y_i\right) = \left(\sum_{i=1}^n X_i^\top X_i\right)^{-1}\left(\sum_{i=1}^n X_i^\top Y_i\right). \tag{19}$$

How does this compare to (18)? First of all, this comparison is only meaningful if $\Sigma = \sigma^2\mathbf{I}$ in (18). Furthermore, note that X and Y from (18) satisfy

$$X^\top X = \begin{bmatrix} X_1^\top & \dots & X_n^\top \end{bmatrix}\begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} = \sum_{i=1}^n X_i^\top X_i \text{ and}$$
$$X^\top Y = \begin{bmatrix} X_1^\top & \dots & X_n^\top \end{bmatrix} Y = \sum_{i=1}^n X_i^\top Y_i$$

and hence we arrive at the same equations as in (19).
If also $d = 1$, then we arrive at

$$\beta^{\mathrm{MLE}} = \frac{\sigma^2 \sum_{i=1}^n X_i Y_i}{\sigma^2 \sum_{i=1}^n X_i^2} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}.$$

The following example shows the power of using a non-trivial feature mapping Ψ:

Example 2.11 (Polynomial regression)**.** Consider the case of $\mathcal{Y} = \mathcal{X} = \mathbb{R}$ and a given sample $\mathcal{S} = \{(x_1, y_1), \dots, (x_n, y_n)\}$. Define the feature mapping

$$\Psi\colon \mathbb{R} \to \mathbb{R}^d, \quad x \mapsto \left(1, x, x^2, \dots, x^d\right)^\top,\ d \in \mathbb{N}.$$

We can then build a feature matrix

$$X := \begin{bmatrix} \Psi(x_1)^\top \\ \vdots \\ \Psi(x_n)^\top \end{bmatrix} \in \mathbb{R}^{n\times d}.$$

The hypothesis class is consequently

$$\mathcal{H}_{\mathrm{pol}} = \left\{ x \mapsto \sum_{k=0}^d \beta_k x^k \;\middle|\; \beta_0, \dots, \beta_d \in \mathbb{R} \right\}.$$

(19) or (18) then yield the ERM weight vector.

3 Model selection and validation

Motivation. In this section, we will explore methods on how to obtain a good estimate of the true risk $L_{\mathscr{D}}(h)$ for an estimator $h \in \mathcal{H}$, $\mathcal{H}$ some hypothesis class. Recall from Theorem 1.43 with sharp bound (10), any hypothesis class with $\mathrm{VCdim}(\mathcal{H}) = d < \infty$ has the uniform convergence property with sample complexity

$$n_{\mathcal{H}}^{\mathrm{uc}} \leq C_2 \frac{d + \log\left(1/\delta\right)}{\varepsilon^2}$$

for some $C_2 > 0$. Hence, for any $n \geq C_2(d + \log\left(1/\delta\right))/\varepsilon^2$ we have

$$\mathrm{P}_{\mathcal{S}\sim\mathscr{D}^n}\left(\sup_{h\in\mathcal{H}}\left|L_{\mathscr{D}}(h) - L_{\mathcal{S}}(h)\right| \leq \varepsilon\right) \geq 1 - \delta.$$

W.l.o.g. assume that $C_2(d + \log\left(1/\delta\right))/\varepsilon^2 \geq 1$. If

$$n = \left\lfloor 2C_2 \frac{d + \log\left(1/\delta\right)}{\varepsilon^2} \right\rfloor,$$

then

$$C_2 \frac{d + \log\left(1/\delta\right)}{\varepsilon^2} \leq n \leq 2C_2 \frac{d + \log\left(1/\delta\right)}{\varepsilon^2}.$$

With $C := 2C_2$, it follows that

$$\mathrm{P}_{\mathcal{S}\sim\mathscr{D}^n}\left(\sup_{h\in\mathcal{H}}\left|L_{\mathscr{D}}(h) - L_{\mathcal{S}}(h)\right| \leq \sqrt{C\frac{d + \log\left(1/\delta\right)}{n}}\right) \geq 1 - \delta,$$

which gives a concentration result for using the empirical risk $L_{\mathcal{S}}(h)$ as an estimator for the true risk $L_{\mathscr{D}}(h)$.
A very popular method of estimating the true error $L_{\mathscr{D}}(h)$ is to use another set of data $\mathcal{V} = \left\{(x_1^v, y_1^v), \ldots, (x_{n_v}^v, y_{n_v}^v)\right\}$, the *validation set*, which is independent of $\mathcal{S}$. The idea is to estimate $L_{\mathscr{D}}(h)$ via $L_{\mathcal{V}}(h)$, the empirical risk on the validation set.

Theorem 3.1. *Let $h = h_{\mathcal{S}}$ be a data-driven estimator and assume that the loss function is bounded in $[0, 1]$. Then, for all $\delta \in (0, 1)$, we have*

$$\mathrm{P}_{\mathcal{V}\sim\mathscr{D}^{n_v}}\left(\left|L_{\mathcal{V}}(h_{\mathcal{S}}) - L_{\mathscr{D}}(h_{\mathcal{S}})\right| \leq \sqrt{\frac{\log\left(2/\delta\right)}{2n_v}} \,\middle|\, \mathcal{S}\right) \geq 1 - \delta \quad \mathscr{D}^n\text{-a.s.}$$

Furthermore,

$$\mathrm{P}_{(\mathcal{V},\mathcal{S})\sim\mathscr{D}^{n_v+n}}\left(\left|L_{\mathcal{V}}(h_{\mathcal{S}}) - L_{\mathscr{D}}(h_{\mathcal{S}})\right| \leq \sqrt{\frac{\log\left(2/\delta\right)}{2n_v}}\right) \geq 1 - \delta.$$

Proof. Recall that for two independent random variables X, Y with distributions $\mathscr{D}^x$ and $\mathscr{D}^y$ and $\psi \in \mathrm{L}(X,Y)$ holds

$$\mathrm{E}_{(X,Y)\sim\mathscr{D}^x\otimes\mathscr{D}^y}\big[\psi(X,Y)\,\big|\,Y\big] = \mathrm{E}_{X\sim\mathscr{D}^x}\big[\psi(X,y)\big]\Big|_{y=Y} =: g(Y).$$

Furthermore, by the tower property of conditional expectations holds

$$\mathrm{E}_{(X,Y)\sim\mathscr{D}^x\otimes\mathscr{D}^y}\big[\psi(X,Y)\big] = \mathrm{E}_{Y\sim\mathscr{D}^y}\big[g(Y)\big].$$

Now replace X and Y by $\mathcal{S}$ and $\mathcal{V}$ and set $\psi(\mathcal{S},\mathcal{V}) := \mathbf{1}\left(\left|L_{\mathscr{D}}(h_{\mathcal{S}}) - L_{\mathcal{V}}(h_{\mathcal{S}})\right| \geq t\right)$ for some $t > 0$, then

$$\mathscr{D}^n\text{-a.s.}\quad \mathrm{P}_{\mathcal{V}\sim\mathscr{D}^{n_v}}\left(\left|L_{\mathcal{V}}(h_{\mathcal{S}}) - L_{\mathscr{D}}(h_{\mathcal{S}})\right| \geq t \,\middle|\, \mathcal{S} = s\right) = \mathrm{P}_{\mathcal{V}\sim\mathscr{D}^{n_v}}\left(\left|L_{\mathcal{V}}(h_s) - L_{\mathscr{D}}(h_s)\right| \geq t\right)$$

and

$$\begin{aligned}&\mathrm{P}_{(\mathcal{S},\mathcal{V})\sim\mathscr{D}^n\otimes\mathscr{D}^{n_v}}\left(\left|L_{\mathcal{V}}(h_{\mathcal{S}}) - L_{\mathscr{D}}(h_{\mathcal{S}})\right| \geq t \,\middle|\, \mathcal{S} = s\right)\\ &\quad = \int \mathrm{P}_{\mathcal{V}\sim\mathscr{D}^{n_v}}\left(\left|L_{\mathcal{V}}(h_s) - L_{\mathscr{D}}(h_s)\right| \geq t\right)\mathrm{d}\mathscr{D}^n(s).\end{aligned}$$

Since

$$L_{\mathscr{D}}(h_{\mathcal{S}}) = \mathrm{E}_{(x,y)\sim\mathscr{D}}\big[l(h_{\mathcal{S}},(x,y))\big] \text{ and } L_{\mathcal{V}}(h_{\mathcal{S}}) = \frac{1}{n_v}\sum_{i=1}^{n_v} l\left(h_{\mathcal{S}},(x_i,y_i)\right),$$

we can apply Hoeffing's inequality (Lemma 1.19) to obtain

$$\mathrm{P}_{\mathcal{V}\sim\mathscr{D}^{n_v}}\left(\left|L_{\mathcal{V}}(h_{\mathcal{S}}) - L_{\mathscr{D}}(h_{\mathcal{S}})\right| \geq \sqrt{\frac{\log\left(2/\delta\right)}{2n_v}} \,\middle|\, \mathcal{S}\right) \leq 2\exp\left\{-2n_v\frac{\log\left(2/\delta\right)}{2n_v}\right\} = \delta. \qquad \square$$

Remark 3.2. Note that we would get a sharper bound, i.e. a smaller error, than in Theorem 3.1 by choosing n_v such that

$$\frac{\log\left(2/\delta\right)}{2n_v} < C\frac{d+\log\left(1/\delta\right)}{n} \iff n_v \geq n\frac{\log\left(2/\delta\right)}{2C\left(d+\log\left(1/\delta\right)\right)}.$$

For example, it is enough to take $n_v \geq n/(2C)$, since we have $\log(2) + \log\left(1/\delta\right) < d + \log\left(1/\delta\right)$.

Remark 3.3. Note that in order to apply Theorem 3.1, we need an extra sample $\mathcal{V}$, which is independent of the sample $\mathcal{S}$ that was used for training the model. In order to obtain those two sets, it is common to divide the original data into training ($\mathcal{S}$) and validation ($\mathcal{V}$) set. We then call $\mathcal{V}$ *hold-out* set.

Remark 3.4. To solve the model selection task of choosing the "best" predictor $h^* \in \{h_{\mathcal{S},1}, \dots, h_{\mathcal{S},k}\}$, $k \in \mathbb{N}$, out of a class of predictors trained on $\mathcal{S}$, we simply choose

$$h^* := \operatorname*{argmin}_{h \in \{h_{\mathcal{S},1}, \dots, h_{\mathcal{S},k}\}} L_{\mathcal{V}}(h).$$

Theorem 3.5. *Let $\mathcal{H} = \{h_1, \dots, h_k\}$ be an arbitrary set of predictors (based on a training set $\mathcal{S}$) and assume that the loss function l is bounded in $[0,1]$. Assume that a validation set $\mathcal{V}$ of size n_v is sampled independently of $\mathcal{S}$. Then,*

$$\mathscr{D}^n\text{-a.s.} \quad \mathrm{P}_{\mathcal{V} \sim \mathscr{D}^{n_v}} \left(\sup_{h \in \mathcal{H}} \left| L_{\mathscr{D}}(h) - L_{\mathcal{V}}(h) \right| \leq \sqrt{\frac{\log\left(2k/\delta\right)}{2n_v}} \,\middle|\, \mathcal{S} \right) \geq 1 - \delta$$

and

$$\mathrm{P}_{(\mathcal{S},\mathcal{V}) \sim \mathscr{D}^n \otimes \mathscr{D}^{n_v}} \left(\sup_{h \in \mathcal{H}} \left| L_{\mathscr{D}}(h) - L_{\mathcal{V}}(h) \right| \leq \sqrt{\frac{\log\left(2k/\delta\right)}{2n_v}} \right) \geq 1 - \delta.$$

Remark 3.6. Note that if $k = |\mathcal{H}|$ in Theorem 3.5 is not too large, then the bound on the maximal deviation between the true risk and the validation error of any h is sharp.

Remark 3.7. Another idea to generate validation data is to use the same training set for both training and validation. Let $\mathscr{A} : \mathcal{S} \mapsto h_{\mathcal{S}}$ be our learning algorithm. The *k-cross validation* method works the following way:

1) Partition $\mathcal{S}$ into k different subsets (*folds*) of size $\approx n/k$: $\mathcal{S} = \bigcup_{i=1}^{k} \mathcal{S}_i$;

2) For each $i \in \{1, \dots, k\}$, train $\mathscr{A}$ on $\mathcal{S} \setminus \mathcal{S}_i$, this gives an estimator $h_{\mathcal{S}}^{(-i)}$;

3) Evaluate the error by computing $L_{\mathcal{S}_i}(h_{\mathcal{S}}^{(-i)})$;

4) Use as predictor for $L_{\mathscr{D}}(h_{\mathcal{S}})$ the average

$$L_{\mathcal{S}}^{(k)}(h_{\mathcal{S}}) := \frac{1}{k} \sum_{i=1}^{k} L_{\mathcal{S}_i}\left(h_{\mathcal{S}}^{(-i)}\right).$$

If $k = n$, this method is also called *leave-one-out (LOO)* procedure.
Note that $L_{\mathcal{S}}^{(k)}(h_{\mathcal{S}})$ is in general a biased estimator of $L_{\mathscr{D}}(h_{\mathcal{S}})$: Assume, for simplicity, that all folds $\mathcal{S}_i$, $i = 1, \dots, k$, have size $m = n/k \in \mathbb{N}$. Then, by i.i.d. nature of $\mathcal{S}$,

$$\begin{aligned} \mathrm{E}_{\mathscr{D}^n}\left[L_{\mathcal{S}}^{(k)}(h_{\mathcal{S}})\right] &= \frac{1}{k} \sum_{i=1}^{k} \mathrm{E}_{\mathscr{D}^n}\left[L_{\mathcal{S}_i}\left(h_{\mathcal{S}}^{(-i)}\right)\right] = \frac{1}{k} \sum_{i=1}^{n} \mathrm{E}_{\mathscr{D}^n}\left[\mathrm{E}_{\mathscr{D}^n}\left[\frac{1}{m} \sum_{z \in \mathcal{S}_i} l(h_{\mathcal{S}}^{(-i)}, z) \,\middle|\, \mathcal{S}_i\right]\right] \\ &= \frac{1}{k} \sum_{i=1}^{k} \mathrm{E}_{\mathscr{D}^n}\left[L_{\mathscr{D}}\left(h_{\mathcal{S}}^{(-i)}\right)\right] = \mathrm{E}_{\mathscr{D}^{m-n}}\left[L_{\mathscr{D}}(\mathscr{A}(n-m))\right], \end{aligned}$$

where $\mathscr{A}(n-m)$ denotes the output of the learning algorithm when trained on an i.i.d. sample of size $n-m$ according to $\mathscr{D}^{n-m}$. We can see that we have a systematic tendency to underfit. Note that the bias becomes smaller when k increases, as this leads to decreasing m.
In order to apply this for model selection, assume that we are given $r \in \mathbb{N}$ hypothesis classes $\{\mathcal{H}_1, \dots, \mathcal{H}_r\}$. Let $\{h_{\mathcal{S}}^{(1)}, \dots, h_{\mathcal{S}}^{(r)}\}$ be the output of our learning algorithm given data $\mathcal{S}$ in the respective hypothesis classes. Then we would choose

$$h_{\mathcal{S}}^{\text{minCV}} : \in \operatorname*{argmin}_{j=1,\dots,r} L_{\mathcal{S}}^{(k)}\left(h_{\mathcal{S}}^{(j)}\right).$$

There are settings under which the k-cross validation works, but a general theory is hard to establish, as we can see in the following counterexample:

Example 3.8. Consider the setup of binary classification, in this case $\mathcal{Y} = \{0,1\}$ with 0-1-loss l. Assume that the distribution $\mathscr{D}$ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ satisfies $\mathscr{D}^{y|x} = \text{Ber}(1/2)$, i.e. the labels are independent of the features and both labels have the same probability. Given a sample $\mathcal{S}$ of size n, take the estimator

$$h_{\mathcal{S}}(x) := \begin{cases} 0, & \text{if } \sum_{i=1}^n y_i \text{ is odd,} \\ 1, & \text{if } \sum_{i=1}^n y_i \text{ is even.} \end{cases}$$

We consider the case where n is even only, the case n being odd follows analogously. Then

$$L_{\mathscr{D}}(h_{\mathcal{S}}) = \mathrm{P}_{(x,y)\sim\mathscr{D}}\big(h_{\mathcal{S}}(x) \neq y\big) = \mathrm{P}_{(x,y)\sim\mathscr{D}}(y \neq 0) = \frac{1}{2}.$$

However, the leave-one-out cross validation estimate $L_{\mathcal{S}}^{(n)}(h_{\mathcal{S}})$ produces 1 on each fold $\mathcal{S}_i$:

- If $\sum_{j\neq i} y_j$ is odd, then $y_i = 1$ and hence $h_{\mathcal{S}}^{(-i)} = 0 \neq 1 = y$;
- If $\sum_{j\neq i} y_j$ is even, then $y_i = 0$ and hence $h_{\mathcal{S}}^{(-i)} = 1 \neq 0 = y$.

Hence $L_{\mathcal{S}}^{(n)}(h_{\mathcal{S}}) = 1$ but $L_{\mathscr{D}}(h_{\mathcal{S}}) = 1/2$.

Remark 3.9. Let us decompose the true error of a predictor $h_{\mathcal{S}}$ using the validation error:

$$L_{\mathscr{D}}(h_{\mathcal{S}}) = \underbrace{L_{\mathscr{D}}(h_{\mathcal{S}}) - L_{\mathcal{V}}(h_{\mathcal{S}})}_{\text{bounded by Theorem 3.1}} + L_{\mathcal{V}}(h_{\mathcal{S}}) - L_{\mathcal{S}}(h_{\mathcal{S}}) + L_{\mathcal{S}}(h_{\mathcal{S}}).$$

Then we have the following cases:

1) $L_{\mathcal{S}}(h_{\mathcal{S}})$ small but $L_{\mathcal{V}}(h_{\mathcal{S}}) - L_{\mathcal{S}}(h_{\mathcal{S}})$ large: This corresponds to overfitting, i.e. the hypothesis class is too rich;

2) $L_S(h_S)$ big: This corresponds to underfitting, i.e. the hypothesis class is not rich enough.

In order to see 2), decompose $L_S(h_S)$ using $h^* \in \operatorname{argmin}_{h\in\mathcal{H}} L_{\mathscr{D}}(h)$:

$$L_S(h_S) = \underbrace{L_S(h_S) - L_S(h^*)}_{\leq 0} + \underbrace{L_S(h^*) - L_{\mathscr{D}}(h^*)}_{\text{bounded by Theorem 1.41}} + \underbrace{L_{\mathscr{D}}(h^*)}_{\text{approximation error}}.$$

If now $L_S(h_S)$ is big, then so is $L_{\mathscr{D}}(h^*)$, which means that the hypothesis class is too small.

Remark 3.10. We have seen in Remark 3.9 that small $L_S(h_S)$ can either imply a very good fit of our model, or indicates overfitting. A way of determining which of these is the case is to use *learning curve*:

- Compute the training errors $L_S(h_S^{(k)})$ occurring when we train the algorithm on $k\eta|S|$ samples from S, $\eta \in (0,1), k = 1,\dots,\lfloor 1/\eta \rfloor$;
- compute the validation errors $L_V(h_S^{(k)})$ with some validation set V independent of S for $k = 1,\dots,\lfloor 1/\eta \rfloor$.

If the validation error does not drop with the increasing training sizes, we found an indication that the approximation error is not 0. Hence we need to enlarge the hypothesis class.
If the validation error declines with the increasing training sizes but stays nevertheless large, then it is an indication that the size $|S|$ is not enough and we need to obtain more examples.
If the training and validation error are spread apart a lot at the beginning and have different asymptotes, then this is evidence for overfitting.
An example for learning curves can be seen in Figure 2.

4 Optimization methods

4.1 Convex Learning Problems

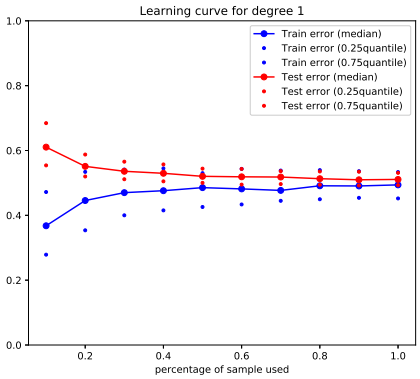
Motivation. We will see in Corollaries 4.45, 4.47 and 4.58, that so-called *convex learning problems* are particularly efficient to solve.

Definition 4.1. A subset C in a vector space X is *convex* if for any two vectors $u, v \in C$, the line segment between u and v is contained in C, i.e.

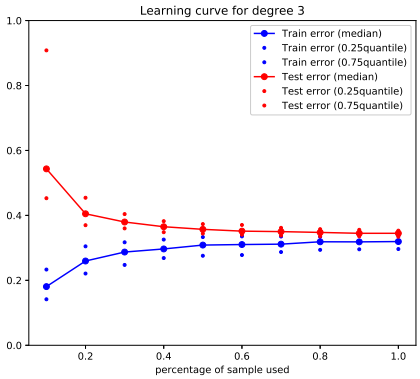
$$\forall\, \alpha \in [0,1], u, v \in C \quad \alpha u + (1-\alpha)v \in C.$$

Definition 4.2. Let C be a convex set. A function $f: C \to \mathbb{R}$ is *convex* if

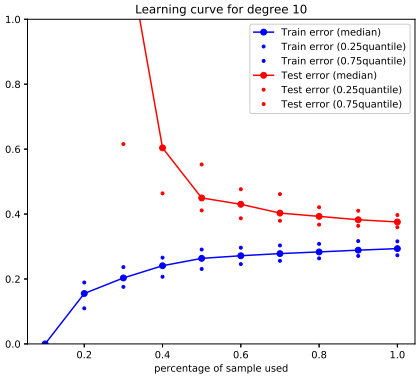
$$\forall\, \alpha \in [0,1], u, v \in C \quad f(\alpha u + (1-\alpha)v) \leq \alpha f(u) + (1-\alpha) f(v).$$



(a) Too small hypothesis class: Underfitting



(b) Correct hypothesis class



(c) Too large hypothesis class: Overfitting

Figure 2: Learning curves for $\mathscr{D}^x = \mathcal{U}([0,1])$ and $\mathscr{D}^{y|x} = (\delta_{5+2x-x^2-3x^3} + \mathcal{U}([-1,1]))/2$, $\mathcal{H}$ the set of polynomials of given maximal degree, $n = 100$, $n_v = 500$.

Remark 4.3. Note that a function $f: \mathcal{C} \to \mathbb{R}$ is convex if and only if

$$\text{epigraph}(f) := \left\{(x,y) \subset \mathcal{X} \times \mathbb{R} \,\middle|\, y \geq f(x)\right\}$$

is a convex set in $\mathcal{X} \times \mathbb{R}$.

Example 4.4. Linear regression with $q = 1$ (cf. 2.2) and quadratic loss $l(h_w, z) = (h_w(x) - y)^2$ with $h_w(x) = \langle w, x \rangle$ is a convex learning problem, as well as Logistic regression (cf. 2.1.3) with the loss exponential loss

$$l(h_w, (x,y)) = \log\left(1 + \exp\left\{-y h_w(x_i)\right\}\right).$$

However, classification with respect to the 0-1-loss is non-convex.

Definition 4.5. We denote with $\mathcal{B}_r(y)$ for $r > 0$ and $y \in \mathbb{R}^d$ the closed Euclidean ball around y with radius r:

$$\mathcal{B}_r(x) := \left\{ x \in \mathbb{R}^d \,\middle|\, \|x - y\| \le r \right\} \subset \mathbb{R}^d.$$

Lemma 4.6. *If $f: \mathcal{C} \to \mathbb{R}$ on a convex set $\mathcal{C}$ is convex, then every local minimum is also a global minimum.*

Proof. Let u be a local minimum of $\mathcal{C}$. Then, there exists $r > 0$ such that for all $v \in \mathcal{B}_r(u) \cap \mathcal{C}$, we have $f(v) \ge f(u)$. Let $w \in \mathcal{C}$, then we can find $\alpha \in (0, 1]$ such that $u + \alpha(w - u) \in \mathcal{B}_r(u) \cap \mathcal{C}$. Therefore, by convexity

$$f(u) \le f(u + \alpha(w - u)) = f((1 - \alpha)u + \alpha w) \le (1 - \alpha) f(u) + \alpha f(w).$$

Hence $\alpha > 0$ implies $f(u) \le f(w)$ for all $w \in \mathcal{C}$. □

Lemma 4.7. *Assume $f: \mathcal{C} \subset \mathbb{R}^d \to \mathbb{R}$ is differentiable with gradient $\nabla f(\,\cdot\,)$. Then f is convex if and only if*

$$\forall\, u, w \in \mathcal{C} \quad f(u) \ge f(w) + \langle \nabla f(w), u - w \rangle, \tag{20}$$

i.e. f stays above the tangent at w.

Proof. Assume f is convex and let $u, w \in \mathcal{C}$. Then for all $\lambda > 0$ holds

$$\begin{aligned}
& f(\lambda u + (1 - \lambda w)) \le \lambda f(u) + (1 - \lambda) f(w) \\
\iff\ & f(\lambda u + (1 - \lambda w)) - f(w) \le \lambda (f(u) - f(w)) \\
\iff\ & \underbrace{\frac{1}{\lambda} (f(\lambda u + (1 - \lambda w)) - f(w))}_{\to \langle \nabla f(w), u - w \rangle \text{ as } \lambda \searrow 0} \le f(u) - f(w).
\end{aligned}$$

Assume now that (20) holds and let $u, w \in \mathcal{C}$ and $v := \lambda u + (1 - \lambda) w$. Then

$$f(u) \ge f(v) + \langle \nabla f(v), u - v \rangle \text{ and } f(w) \ge f(v) + \langle \nabla f(v), w - v \rangle.$$

Hence

$$\lambda f(u) + (1 - \lambda) f(w) \ge f(v) + \langle \nabla f(v), \lambda u + (1 - \lambda) w - v \rangle = f(v) = f(\lambda u + (1 - \lambda w)). \square$$

Lemma 4.8. *Let $f: \mathbb{R} \to \mathbb{R}$ be twice differentiable. Then the following assertions are equivalent:*

1) f is convex;

2) f' is non-decreasing;

3) $f'' > 0$.

Lemma 4.9. *Let $g\colon \mathbb{R} \to \mathbb{R}$ be convex. Then also*

$$f\colon \begin{cases} \mathbb{R}^d \to \mathbb{R}, \\ w \mapsto g(\langle w, x\rangle + y) \end{cases}$$

for $x \in \mathbb{R}^d$, $y \in \mathbb{R}$, is convex.

Proof. For $\alpha \in [0,1]$ and $v, w \in \mathbb{R}^d$ we have

$$\begin{aligned} f(\alpha v + (1-\alpha)w) &= g\left(\alpha\langle v, x\rangle + (1-\alpha)\langle w, x\rangle + y\right) \\ &= g\left(\alpha\left(\langle v, x\rangle + y\right) + (1-\alpha)\left(\langle w, x\rangle + y\right)\right) \\ &\le \alpha g\left(\langle v, x\rangle + y\right) + (1-\alpha) g\left(\langle w, x\rangle + y\right) \\ &= \alpha f(v) + (1-\alpha) f(w). \end{aligned}$$

□

Example 4.10. Lemmata 4.8 and 4.9 show that the functions

$$f\colon \begin{cases} \mathbb{R}^d \to \mathbb{R}, \\ w \mapsto (\langle w, x\rangle - y)^2 \end{cases} \quad \text{and} \quad g\colon \begin{cases} \mathbb{R}^d \to \mathbb{R}, \\ \log\left(1 + \exp\{-y\langle w, x\rangle\}\right) \end{cases}$$

are convex, as

$$\begin{aligned} &\frac{d^2}{da^2}(a-y)^2 = \frac{d}{da} 2(a-y) = 2 > 0 \text{ and} \\ &\frac{d^2}{da^2} \log\left(1 + \exp\{a\}\right) = \frac{d}{da} \frac{\exp\{a\}}{1+\exp\{a\}} = \frac{\exp\{a\}}{(1+\exp\{a\})^2} > 0 \end{aligned}$$

for all $a \in \mathbb{R}$.

Lemma 4.11. *For $i = 1, \dots, r$, $r \in \mathbb{N}$, let $g_i\colon \mathbb{R}^d \to \mathbb{R}$ be a convex function. Then also*

- $\max_{i=1,\dots,r} g_i(\,\cdot\,)$ *and*
- $\sum_{i=1}^r w_i g_i(\,\cdot\,)$ *for $w_i \ge 0$, $i = 1, \dots, r$,*

are convex.

Definition 4.12. A function $f\colon \mathbb{R}^d \to \mathbb{R}^q$, $p, q \in \mathbb{N}$ is called *(ρ-)Lipschitz*, if

$$\exists\, \rho > 0\ \forall\, x, y \in \mathbb{R}^d \quad \left\|f(x) - f(y)\right\| \le \rho \|x - y\|.$$

Theorem 4.13. *A differentiable function $f\colon \mathbb{R}^d \to \mathbb{R}$ is ρ-Lipschitz if and only if $\left\|\nabla f(x)\right\| \le \rho$ for all $x \in \mathbb{R}^d$.*

Proof. Assume that f is ρ-Lipschitz, then for all $x, w \in \mathbb{R}^d$ holds

$$\left|\left\langle \nabla f(x), w\right\rangle\right| = \lim_{h \to 0} \frac{\left|f(x + hw) - f(x)\right|}{|h|} \le \lim_{h \to 0} \frac{\rho |h| \|w\|}{|h|} = \rho \|w\|.$$

Choosing $w = \nabla f(x)$ yields $\|\nabla f(x)\|^2 \le \rho\|\nabla f(x)\|$ and hence $\|\nabla f(x)\| \le \rho$.
Assume now that $\|\nabla f(x)\| \le \rho$ for all $x \in \mathbb{R}^d$. Then by Taylor's Theorem, for all $x, w \in \mathbb{R}^d$ holds for some ξ between x and w that

$$\big\|f(x) - f(w)\big\| = \Big\|\big\langle \nabla f(\xi), x - w\big\rangle\Big\| \le \big\|\nabla f(\xi)\big\|\|x - w\| \le \rho\|x - w\|,$$

which implies that f is ρ-Lipschitz. □

Example 4.14. The following functions are 1-Lipschitz:

- $\mathbb{R}^d \ni x \mapsto \|x\|$;
- $\mathbb{R} \ni x \mapsto \log\big(1 + \exp\{x\}\big)$.

$\mathbb{R} \ni x \mapsto x^2$ is not Lipschitz, it can be made Lipschitz by restricting to domain to a bounded subset $\mathcal{C} \subset \mathbb{R}$.
Note also that $f\colon \mathbb{R}^d \to \mathbb{R}$, $w \mapsto \langle w, x\rangle + b$, where $x \in \mathbb{R}^d, b \in \mathbb{R}$, is $\|x\|$-Lipschitz because

$$\forall\, w_1, w_2 \in \mathbb{R}^d \quad \big|f(w_1) - f(w_2)\big| = \big|\langle w_1 - w_2, x\rangle\big| \le \|x\|\|w_1 - w_2\|.$$

Lemma 4.15. *Let $f(\,\cdot\,) := g_1(g_2(\,\cdot\,))$, where $g_1\colon \mathbb{R}^q \to \mathbb{R}^r$ is ρ_1-Lipschitz and $g_2 := \mathbb{R}^d \to \mathbb{R}^q$ is ρ_2-Lipschitz. Then, f is $(\rho_1\rho_2)$-Lipschitz. In particular, if $g_2(w) = \langle w, v\rangle + b$ for $v \in \mathbb{R}^d$, $b \in \mathbb{R}$, then f is $(\rho_1\|v\|)$-Lipschitz.*

Proof. Note that for all $v, w \in \mathbb{R}^d$ holds

$$\begin{aligned}\big\|f(v) - f(w)\big\| &= \big\|g_1(g_2(v)) - g_1(g_2(w))\big\| \le \rho_1\big\|g_2(v) - g_2(w)\big\| \\ &\le \rho_1\rho_2\|v - w\|.\end{aligned}$$

□

Definition 4.16. A differentiable function $f\colon \mathbb{R}^d \to \mathbb{R}$ is called *β-smooth* for $\beta > 0$ if $\nabla f(\,\cdot\,)$ is β-Lipschitz, i.e.

$$\forall\, v, w \in \mathbb{R}^d \quad \big\|\nabla f(v) - \nabla f(w)\big\| \le \beta\|v - w\|.$$

Lemma 4.17. *If $f\colon \mathbb{R}^d \to \mathbb{R}$ is β-smooth, then*

$$\forall\, v, w \in \mathbb{R}^d \quad f(v) \le f(w) + \big\langle \nabla f(w), v - w\big\rangle + \frac{\beta}{2}\|v - w\|^2.$$

Proof. Let $h(t)\colon [0,1] \ni t \mapsto f(tv + (1-t)w) \in \mathbb{R}$, then h is differentiable and

$$h'(t) = \big\langle \nabla f(w + t(v - w)), v - w\big\rangle, \quad t \in (0,1).$$

Then

$$f(v) - f(w) = h(1) - h(0) = \int_0^1 h'(t)\,\mathrm{d}t = \int_0^1 \big\langle \nabla f(w + t(v - w)), v - w\big\rangle\,\mathrm{d}t.$$

It follows with the Cauchy-Schwarz inequality that

$$\begin{aligned} f(v) - f(w) - \langle \nabla f(w), v - w \rangle &= \int_0^1 \langle \nabla f(w + t(v-w)) - f(w), v - w \rangle \, \mathrm{d}t \\ &\leq \int_0^1 \|\nabla f(w + t(v-w)) - \nabla f(w)\| \|v - w\| \, \mathrm{d}t \\ &\leq \|v - w\| \int_0^1 \beta t \|v - w\| \, \mathrm{d}t \\ &= \frac{\beta}{2} \|v - w\|^2, \end{aligned}$$

which completes the proof. □

Remark 4.18. Note that if $f: \mathbb{R}^d \to \mathbb{R}$ is convex and β-smooth, then by Lemmata 4.7 and 4.17, we have the bounds

$$\forall\, v, w \in \mathbb{R}^d \quad f(w) + \langle \nabla f(w), w - v \rangle \leq f(v) \leq f(w) + \langle \nabla f(w), w - v \rangle + \frac{\beta}{2} \|v - w\|^2.$$

If $v = w - \nabla f(w)/\beta$, then $v - w = -\nabla f(w)/\beta$ and hence

$$\frac{1}{2\beta} \|\nabla f(w)\|^2 \leq f(w) - f(v).$$

If, additionally, $f \geq 0$, then

$$\left\| \nabla f(w)^2 \right\| \leq 2\beta f(w)$$

and we call f *self-bounded*.

Lemma 4.19. *Let $g: \mathbb{R} \to \mathbb{R}$ be β-smooth. Then for some fixed $x \in \mathbb{R}^d$, $b \in \mathbb{R}$,*

$$f: \begin{cases} \mathbb{R}^d \to \mathbb{R}, \\ w \mapsto g(\langle w, x \rangle + b) \end{cases}$$

is $\beta \|x\|^2$-smooth.

Proof. This follows from

$$\begin{aligned} \|\nabla f(v) - \nabla f(w)\| = \left\| \left(g'(\langle w, x \rangle + b) - g'(\langle v, x \rangle + b) \right) x \right\| &\leq \beta \left\| \langle w - v, x \rangle \right\| \|x\| \\ &\leq \beta \|x\|^2 \|w - v\| \end{aligned}$$

for all $v, w \in \mathbb{R}^d$, where we used the Cauchy-Schwarz inequality. □

Example 4.20. Consider the function $g: \mathbb{R} \ni x \mapsto \log\left(1 + \exp\{-yx\}\right) \in \mathbb{R}$, for some fixed $y \in \{-1, 1\}$. Then

$$|g''(x)| = \frac{\exp\{-xy\}}{(1 + \exp\{-xy\})^2} \leq \frac{1}{4}$$

and hence $g'(\,\cdot\,)$ is 1/4-Lipschitz. Thus, the function

$$f\colon \mathbb{R}^d \to \mathbb{R}, \quad w \mapsto \log\Big(\big(1+\exp\big\{-y\langle w,x\rangle\big\}\big)\Big)$$

is $\|x\|^2/4$-smooth.

Motivation. Recall that a learning problem needs a hypothesis class $\mathcal{H}$, a domain $\mathcal{Z} = \mathcal{X}\times\mathcal{Y}$, and a loss function $l\colon \mathcal{H}\times\mathcal{Z}\to[0,\infty)$. Up to now, the elements in $\mathcal{H}$ were functions $h\colon \mathcal{X}\to\mathcal{Y}$. Here, we will assume that each hypothesis function h can be identified with a real d-dimensional vector $w\in\mathbb{R}^d$.

Definition 4.21. A learning problem $(\mathcal{H},\mathcal{Z},l)$ is called *convex* if $\mathcal{H}$ is a convex set and the function $w\mapsto f(w) := l(w,z)$ is convex for any fixed $z\in\mathcal{Z}$.

Example 4.22. Consider a regression problem with $q=1$, where the hypothesis class $\mathcal{H}$ can be identified with $\mathbb{R}^d$ (since $h(x)=\langle h_w,x\rangle$ for some $w\in\mathbb{R}^d$) and the quadratic loss function $l(w,(x,y)) = (\langle h_w,x\rangle - y)^2$. By Lemma 4.9, this yields a convex learning problem.

Lemma 4.23. *If l is a convex loss function and $\mathcal{H}$ is convex, then the* $\mathrm{ERM}_{\mathcal{H}}$ *problem (of minimizing the empirical loss over $\mathcal{H}$) is a convex optimization problem, as it corresponds to the problem of minimizing a convex function over a convex set.*

Proof. Let $S = \big\{(x_1,y_1),\dots,(x_n,y_n)\big\}$ be some training set. Then, when applying the ERM paradigm, we aim at minimizing

$$w\mapsto L_S(w) := \frac{1}{n}\sum_{i=1}^{n} l(w,(x_i,y_i)),$$

which is a convex function. □

Remark 4.24. Note that convexity is not sufficient for a problem to be PAC learnable. It can be shown that for linear regression with $d=q=1$ and $\mathcal{H}=\mathbb{R}$ with quadratic loss l, for any sample S of size $n\in\mathbb{N}$ and any learning algorithm $\mathscr{A}$ we can find $(\varepsilon_0,\delta_0)\in(0,1)$ and a distribution $\mathscr{D}$ on $\mathbb{R}^2$ such that

$$\mathrm{P}_{S\sim\mathscr{D}^n}\Big(L_{\mathscr{D}}(\mathscr{A}(S)) \geq \min_{w\in\mathbb{R}} L_{\mathscr{D}}(w)+\varepsilon_0\Big)\geq\delta_0.$$

Definition 4.25. A learning problem $(\mathcal{H},\mathcal{Z},l)$ is called *convex-Lipschitz-bounded* with parameters $\rho,B>0$, if the class $\mathcal{H}$ is a convex set $\mathcal{H}\subset\mathcal{B}_B(0)$ (that is $\forall w\in\mathcal{H}:\|w\|\leq B$) and for all $z\in\mathcal{Z}$, $w\mapsto l(w,z)$ is convex and ρ-Lipschitz.

Example 4.26. Consider the setting of $\mathcal{X}=\mathcal{B}_\rho(0)\subset\mathbb{R}^d$, $\mathcal{Y}=\mathbb{R}$, $\mathcal{H}=\mathcal{B}_B(0)$ and $l(w,(x,y)) = |\langle w,x\rangle - y|$. By Lemma 4.15, l is $\|x\|$-Lipschitz.

Definition 4.27. A learning problem $(\mathcal{H},\mathcal{Z},l)$ is called *convex-smooth-bounded* with parameters $\beta,B>0$, if $\mathcal{H}$ is a convex set with $\mathcal{H}\subset\mathcal{B}_B(0)$ and for all $z\in\mathcal{Z}$, $w\mapsto l(w,z)$ is convex and β-smooth.

Example 4.28. Consider the setting of $\mathcal{X} = \mathcal{B}_{\sqrt{\beta/2}}(0)$, $\mathcal{Y} = \mathbb{R}$, $\mathcal{H} = \mathcal{B}_B(0)$ and

$$l(w,(x,y)) = (\langle w,x\rangle - y)^2.$$

By Lemma 4.19, $w \mapsto (\langle h_w,x\rangle - y)^2$ is $2\|x\|^2$-smooth, and $2\|x\|^2 \le \beta$. Hence the loss function is β-smooth.

Remark 4.29 (Surrogate loss function)**.** Consider the classification problem with half-spaces with domain $\mathcal{Z} = \mathbb{R}^d \times \{-1,1\}$ and loss function $l_{0-1}(w,(x,y)) = \mathbf{1}(y = \operatorname{sign}(\langle w,x\rangle))$ for $w \in \mathbb{R}^d$. The function $w \mapsto l_{0-1}(w,(x,y))$ is not convex and it can be shown that finding the ERM rule in the non-separable case is NP-hard. In order to make the minimization problem easier, one solution is to upper bound the non-convex function (to be minimized) by a convex surrogate function. For example, consider

$$l^{\text{hinge}}(w,(x,y)) = \max(0, 1 - y\langle w,x\rangle).$$

For all $w \in \mathbb{R}$ and $(x,y) \in \mathcal{Z}$ we have that $l_{0-1}(w,(x,y)) \le l^{\text{hinge}}(w,(x,y))$: Indeed, $y \ne \operatorname{sign}(\langle w,x\rangle)$ holds if and only if $y\langle w,x\rangle \le 0$, and hence

$$l_{0-1}(w,(x,y)) = 1 \implies l^{\text{hinge}}(w,(x,y)) = 1 - y\langle w,x\rangle \ge 1 = l_{0-1}(w,(x,y)).$$

Also, $w \mapsto l^{\text{hinge}}(w,(x,y))$ is convex by Lemma 4.11.
Let $\mathscr{A}$ be a learning algorithm which can learn w using the hinge loss. We aim to achieve

$$L_{\mathscr{D}}^{\text{hinge}}(\mathscr{A}(S)) \le \min_{w\in\mathcal{H}} L_{\mathscr{D}}^{\text{hinge}}(w) + \varepsilon$$

for some small estimation error $\varepsilon > 0$, where for $L_{\mathscr{D}}^{\text{hinge}}$ we replace l in $L_{\mathscr{D}}$ with l^{hinge}. We thus have

$$\begin{aligned} L_{\mathscr{D}}(\mathcal{A}(S)) &\le L_{\mathscr{D}}^{\text{hinge}}(\mathcal{A}(S)) \le \min_{w\in\mathcal{H}} L_{\mathscr{D}}^{\text{hinge}}(w) + \varepsilon \\ &\le \underbrace{\min_{w\in\mathcal{H}} L_{\mathscr{D}}(w)}_{\text{approximation error}} + \underbrace{\min_{w\in\mathcal{H}} L_{\mathscr{D}}^{\text{hinge}}(w) - \min_{w\in\mathcal{H}} L_{\mathscr{D}}(w)}_{\text{optimization error}} + \underbrace{\varepsilon.}_{\text{estimation error}} \end{aligned}$$

The optimization error depends on the unknown distribution $\mathscr{D}$ (and also on our choice for the surrogate function).

4.2 Stochastic Gradient Descend (SGD)

Motivation. We consider again the setting where $\mathcal{H}$ can be identified with some convex subset of $\mathbb{R}^d$ and the loss function $w \mapsto l(w,z)$ is convex for any $z \in \mathcal{Z}$. In this subsection, we will study the properties of a new learning method, namely the *Stochastic Gradient Descent (SGD)*. We start with the simpler version called *Gradient Descent (GD)* and analyze its convergence.

Definition 4.30. The *Gradient Descent (GD)* algorithm with $T \in \mathbb{N}$ steps and step size (*learning rate*) $\eta > 0$ for minimizing a differentiable function $f: \mathbb{R}^d \to \mathbb{R}$ works as follows:

Algorithm 2 Gradient Descent (GD)

1: Set $w^{(1)} := 0 \in \mathbb{R}^d$;
2: **for** $t = 1, \ldots, T-1$ **do**
3: Update $w^{(t+1)} := w^{(t)} - \eta \nabla f(w^{(t)})$;
4: **end for**
5: **return** $\bar{w} := \sum_{t=1}^{T} w^{(t)}/T$.

Remark 4.31. Gradient descent implements the following idea: If f is a differentiable function on $\mathbb{R}^d$ with gradient $\nabla f(w)$, then $\nabla f(w)$ points in the direction of the greatest rate of increase of f around w. If f admits a minimum at w^*, then we "hunt" for this minimizer by iteratively updating

$$w^{(t+1)} := w^{(t)} - \eta \nabla f\left(w^{(t)}\right).$$

Starting from $w^{(1)} = 0$, it can be shown that under some conditions, the output $\bar{w}$ converges to the minimum w^*:

$$\bar{w} := \frac{1}{T} \sum_{t=1}^{T} w^{(t)} \xrightarrow{T \to \infty} w^*.$$

Suppose that f is convex. Then, the starting point of the analysis is to note that

$$f(\bar{w}) - f(w^*) = f\left(\frac{1}{T} \sum_{t=1}^{T} w^{(t)}\right) - f(w^*) \leq \frac{1}{T} \sum_{t=1}^{T} f\left(w^{(t)}\right) - f(w^*) = \frac{1}{T} \sum_{t=1}^{T} \left[f\left(w^{(t)}\right) - f(w^*) \right].$$

Using convexity, we have that $f(w^{(t)}) \leq f(w^*) + \langle w^{(t)} - w^*, \nabla f(w^{(t)}) \rangle$ and hence

$$f(\bar{w}) - f(w^*) \leq \frac{1}{T} \sum_{t=1}^{T} \left\langle w^{(t)} - w^*, \nabla f(w^{(t)}) \right\rangle$$

and the goal now is to upper bound the term on the right side.

Lemma 4.32. *Let $v_1, \ldots, v_T$ be an arbitrary sequence of vectors in $\mathbb{R}^d$. Any algorithm with an initialization $w^{(1)} = 0$ and an update rule of the form*

$$w^{(t+1)} = w^{(t)} - \eta v_t$$

for some $\eta > 0$ satisfies for any $w^* \in \mathbb{R}^d$

$$\sum_{t=1}^{T} \left\langle w^{(t)} - w^*, v_t \right\rangle \leq \frac{1}{2\eta} \left\| w^* \right\|^2 + \frac{\eta}{2} \sum_{t=1}^{T} \| v_t \|^2.$$

In particular, for all $B > 0, \rho > 0$, if we have $\|v_t\| \le \rho$ and if $\eta = \sqrt{B^2/(\rho^2 T)}$, then for any w^* with $\|w^*\| \le B$ we have

$$\frac{1}{T}\sum_{t=1}^{T}\left\langle w^{(t)} - w^*, v_t\right\rangle \le \frac{B\rho}{\sqrt{T}}$$

Proof. Note that

$$\begin{aligned}\left\langle w^{(t)} - w^*, v_t\right\rangle &= \frac{1}{\eta}\left\langle w^{(t)} - w^*, \eta v_t\right\rangle \\ &= \frac{1}{2\eta}\left(-\left\|w^{(t)} - w^* - \eta v_t\right\|^2 + \left\|w^{(t)} - w^*\right\|^2 + \eta^2\|v_t\|^2\right) \\ &= \frac{1}{2\eta}\left(-\left\|w^{(t+1)} - w^*\right\|^2 + \left\|w^{(t)} - w^*\right\|^2\right) + \frac{\eta}{2}\|v_t\|^2\end{aligned}$$

by definition of $w^{(t+1)}$. Summing over $t = 1, \dots, T$ yields

$$\begin{aligned}\sum_{t=1}^{T}\left\langle w^{(t)} - w^*, v_t\right\rangle &= \frac{1}{2\eta}\left(-\left\|w^{(T+1)} - w^*\right\|^2 + \left\|w^{(1)} - w^*\right\|^2\right) + \frac{\eta}{2}\sum_{t=1}^{T}\|v_t\|^2 \\ &\le \frac{1}{2\eta}\left\|w^{(1)} - w^*\right\|^2 + \frac{\eta}{2}\sum_{t=1}^{T}\|v_t\|^2 \\ &= \frac{1}{2\eta}\left\|w^*\right\|^2 + \frac{\eta}{2}\sum_{t=1}^{T}\|v_t\|^2,\end{aligned}$$

since $w^{(1)} = 0$. □

Corollary 4.33. *Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be convex, ρ-Lipschitz and differentiable and let $w^* \in \operatorname{argmin}_{w\in\mathcal{B}_B(0)} f(w)$ for some $B > 0$. If the gradient descent algorithm is run for T steps with $\eta = \sqrt{B^2/(\rho^2 T)}$, then*

$$f(\bar{w}) - f(w^*) \le \frac{B\rho}{\sqrt{T}}.$$

Thus, in order to have $f(\bar{w}) - f(w^*) \le \varepsilon$ for some $\varepsilon > 0$, it suffices to take $T \ge B^2\rho^2/\varepsilon^2$.

Proof. Since f is ρ-Lipschitz and differentiable, we have that $\|\nabla f(w^{(t)}\| \le \rho$ by Theorem 4.13. Apply Lemma 4.32 with $v_t := \nabla f(w^{(t)})$. □

Motivation. We can generalize the Gradient Descent algorithm to convex non-differentiable functions, using the concept of *subgradients*. Recall that if $f\colon \mathbb{R}^d \to \mathbb{R}$ is a convex differentiable function, then for all $u \in \mathbb{R}^d$ holds

$$f(u) \ge f(w) + \left\langle u - w, \nabla f(w)\right\rangle$$

by 4.7. This property can be strengthened by the following result:

Lemma 4.34. *Let $\mathcal{C}$ be an open convex subset of $\mathbb{R}^d$. A function $f\colon \mathcal{C} \to \mathbb{R}$ is convex if and only if*

$$\forall\, w \in \mathcal{C}\ \exists\, v \in \mathbb{R}^d\ \forall\, u \in \mathcal{C} \quad f(u) \geq f(w) + \langle u - w, v\rangle. \tag{21}$$

Definition 4.35. A vector $v \in \mathbb{R}^d$ that satisfies (21) is called a *subgradient of f at $w \in \mathcal{C}$*. The set of all subgradients of f at w is called the *differential set of f at $w \in \mathcal{C}$* and is denoted by $\partial f(w)$.

Remark 4.36. If $f\colon \mathbb{R}^d \to \mathbb{R}$ is differentiable, then $\partial f(w) = \{\nabla f(w)\}$.

Example 4.37. Consider $f\colon \mathbb{R} \ni x \mapsto |x| \in \mathbb{R}$. Then f is differentiable for all $x \neq 0$ and consequently $\partial f(x) = \{-1\}$ for $x < 0$ and $\partial f(x) = \{1\}$ for $x > 0$. For $x = 0$ note that $f(t) \geq f(0) + a(t-0)$, $t \in \mathbb{R}$, holds if and only if $|t| \geq at$, i.e. if $a \leq 1$ or $a \geq 1$. Hence

$$\partial f(x) = \begin{cases} \{-1\} & \text{if } x < 0, \\ [-1, 1] & \text{if } x = 0, \\ \{1\} & \text{if } x > 0. \end{cases}$$

Lemma 4.38. *Let $g_1, \ldots, g_r$ be $r \in \mathbb{N}$ convex differentiable functions $\mathbb{R}^d \to \mathbb{R}$ and $g := \max_{i=1,\ldots,r} g_i$. For a given $w \in \mathbb{R}^d$, let $j \in \{1, \ldots, r\}$ be such that $g(w) = g_j(w)$. Then, $\nabla g_j(w) \in \partial g(w)$.*

Proof. Convexity of g_j implies that for all $u \in \mathbb{R}^d$ holds

$$g_j(u) \geq g_j(w) + \big\langle u - w, \nabla g_j(w)\big\rangle$$

by Lemma 4.7. Since $g(w) = g_j(w)$ and $g(u) \geq g_j(u)$, it follows that

$$g(u) \geq g(w) + \big\langle u - w, \nabla g_j(w)\big\rangle. \qquad \square$$

Example 4.39. Consider the hinge loss function $f\colon \mathbb{R}^d \ni w \mapsto \max(0, 1 - y\langle w, x\rangle) \in \mathbb{R}$ for some vector $x \in \mathbb{R}^d$ and $y \in \{-1, 1\}$. Then, for a given $w \in \mathbb{R}^d$, the vector

$$v := \begin{cases} 0 & \text{if } 1 - y\langle w, x\rangle \leq 0, \\ -yx & \text{otherwise,} \end{cases}$$

is a subgradient of f at w.

Lemma 4.40. *Let $\mathcal{C} \subset \mathbb{R}^d$ be a convex open set and let $f\colon \mathcal{C} \to \mathbb{R}$ be a convex function. Then f is ρ-Lipschitz on $\mathcal{C}$ if and only if $\forall\, w \in \mathcal{C}, v \in \partial f(w)$, we have that $\|v\| \leq \rho$.*

Proof. Suppose that any $v \in \partial f(w)$ satisfies $\|v\| \leq \rho$. By definition of $\partial f(w)$, we have that $f(w) - f(u) \leq \langle v, w - u\rangle$. Applying the Cauchy-Schwarz inequality, we arrive at

$$f(w) - f(u) \leq \|v\|\|w - u\| \leq \rho\|w - u\|.$$

A similar argument can be applied to show that $f(u) - f(w) \leq \rho\|u - w\|$. Hence, f is ρ-Lipschitz.
Suppose now that f is ρ-Lipschitz, and let $w \in \mathcal{C}$ and $v \in \partial f(w)$. If $v = 0$, then nothing is left to show. Suppose now that $v \neq 0$. Since $\mathcal{C}$ is open, we can find a small $\varepsilon > 0$ such that $u = w + \varepsilon v/\|v\| \in \mathcal{C}$. Then, $\langle u - w, v\rangle = \varepsilon\|v\|$, and $\|u - w\| = \varepsilon$. From the definition of the subgradient (21) and ρ-Lipschitzness, we have that

$$\rho\varepsilon = \rho\|u - w\| \geq f(u) - f(w) \geq \langle u - w, v\rangle = \varepsilon\|v\|,$$

implying that $\|v\| \leq \rho$. □

Definition 4.41. The *Subgradient Descent (SubGD)* algorithm with $T \in \mathbb{N}$ steps and step size (*learning rate*) $\eta > 0$ for minimizing a convex and ρ-Lipschitz function $f\colon \mathbb{R}^d \to \mathbb{R}$ works as follows:

Algorithm 3 Subgradient Descent (SubGD)

1: Set $w^{(1)} := 0 \in \mathbb{R}^d$;
2: **for** $t = 1, \ldots, T-1$ **do**
3: compute $v_t \in \partial f(w^{(t)})$;
4: update $w^{(t+1)} := w^{(t)} - \eta v_t$;
5: **end for**
6: **return** $\bar{w} := \sum_{t=1}^{T} w^{(t)}/T$.

Motivation. Note that the function f we want to minimize is the loss function $w \mapsto L_{\mathscr{D}}(w)$ and is unknown to us. Thus, the gradient or sub-gradient at any vector w is also unknown. What should we do?
At iteration t, we can replace the unknown gradient or subgradient by a random vector v_t such that

$$\mathrm{E}\Big[v_t \,\Big|\, w^{(t)}\Big] \in \partial f(w^{(t)}).$$

This idea leads to the *stochastic gradient descent.*

Definition 4.42 (Stochastic Gradient Descend Version 1)**.** The *Stochastic Gradient Descent (SGD)* algorithm with $T \in \mathbb{N}$ steps and step size (*learning rate*) $\eta > 0$ for minimizing a convex and ρ-Lipschitz function $f\colon \mathbb{R}^d \to \mathbb{R}$ works as follows:

Algorithm 4 Stochastic Gradient Descent (SGD) for general minimization

1: Set $w^{(1)} := 0 \in \mathbb{R}^d$;
2: **for** $t = 1, \ldots, T-1$ **do**
3: Generate v_t from a distribution such that $\mathrm{E}\Big[v_t \,\Big|\, w^{(t)}\Big] \in \partial f(w^{(t)})$;
4: Update $w^{(t+1)} := w^{(t)} - \eta v_t$;
5: **end for**
6: **return** $\bar{w} := \sum_{t=1}^{T} w^{(t)}/T$.

Theorem 4.43 (Convergence of the SGD)**.** *Let $B > 0$ and $\rho > 0$ and let f be a convex function with $w^* := \operatorname{argmin}_{w:\|w\| \le B} f(w)$. Assume that the SGD algorithm (as described in Definition 4.42) is run for T iterations with learning rate $\eta = \sqrt{B^2/(\rho^2 T)}$. Assume further that v_t satisfies $\|v_t\| \le \rho$ with probability 1 for all $t \in \{1, \dots, T\}$. Then,*

$$\mathrm{E}\big[f(\bar{w})\big] - f\left(w^*\right) \le \frac{B\rho}{\sqrt{T}}.$$

Proof. Omitted. □

Motivation. We now want to study learning with SGD for risk minimization. Recall that the main goal is to minimize the (true) risk function $L_{\mathscr{D}}(\,\cdot\,)$, where

$$L_{\mathscr{D}}(w) := \mathrm{E}_{z\sim\mathscr{D}}\big[l(w,z)\big].$$

We will analyze the convergence of the SGD for convex and ρ-Lipschitz loss functions: Assume that $w \mapsto l(w,z)$ is convex, ρ-Lipschitz and differentiable for all $z \in \mathcal{Z}$.
Then, by the Dominated Convergence Theorem, the function $w \mapsto L_{\mathscr{D}}(w)$ is differentiable if there exists a function $z \mapsto k(z)$ such that $k \ge 0$ and $\mathrm{E}_{z\sim\mathscr{D}}\big[k(z)\big] < \infty$ and $\big\|\nabla l(w,z)\big\| \le k(z)$ for all w, z. Furthermore, we have that

$$\nabla L_{\mathscr{D}}(w) = \mathrm{E}_{z\sim\mathscr{D}}\big[\nabla l(w,z)\big].$$

Note that if $w \mapsto l(w,z)$ is ρ-Lipschitz and differentiable (our working assumptions), then

$$\forall\, w \in \mathbb{R}^d, z \in \mathcal{Z} \quad \big\|\nabla l(w,z)\big\| \le \rho.$$

Hence we can choose $k(z) = \rho$ for the Dominated Convergence Theorem. Define $v_t := \nabla l(w^{(t)}, z)$, where $z \sim \mathscr{D}$ independent of $w^{(t)}$. We know that $\|v_t\| \le \rho$. Let us compute $\mathrm{E}_{z\sim\mathscr{D}}\Big[v_t \,\Big|\, w^{(t)}\Big]$:

$$\begin{aligned}\mathrm{E}_{z\sim\mathscr{D}}\Big[v_t \,\Big|\, w^{(t)}\Big] &= \mathrm{E}_{z\sim\mathscr{D}}\Big[\nabla l\left(w^{(t)}, z\right) \,\Big|\, w^{(t)}\Big] = \mathrm{E}_{z\sim\mathscr{D}}\Big[\nabla l\left(w^{(t)}, z\right)\Big] \\ &= \nabla \mathrm{E}_{z\sim\mathscr{D}}\Big[l\left(w^{(t)}, z\right)\Big] = \nabla L_{\mathscr{D}}\left(w^{(t)}\right),\end{aligned}$$

which is the only element in $\partial L_{\mathscr{D}}(w^{(t)})$.
Now, assume that $w \mapsto l(w,z)$ is still convex and ρ-Lipschitz, but not necessarily differentiable anymore. Define $v_t \in \partial l(w^{(t)}, z)$ with $z \sim \mathscr{D}$ independent of $w^{(t)}$. We have again that $\|v_t\| \le \rho$, as we are still applying the properties of a sub-gradient vector for a ρ-Lipschitz function. By the properties of sub-gradients, we have that

$$\forall\, u \in \mathbb{R}^d \quad l(u,z) \ge l\left(w^{(t)}, z\right) + \left\langle u - w^{(t)}, v_t\right\rangle.$$

Consequently,

$$\mathrm{E}_{z\sim\mathscr{D}}\Big[l(u,z) \,\Big|\, w^{(t)}\Big] \ge \mathrm{E}_{z\sim\mathscr{D}}\Big[l\left(w^{(t)}, z\right) \,\Big|\, w^{(t)}\Big] + \mathrm{E}_{z\sim\mathscr{D}}\Big[\left\langle u - w^{(t)}, v_t\right\rangle \,\Big|\, w^{(t)}\Big].$$

But this is equivalent to

$$\mathrm{E}_{z\sim\mathscr{D}}\big[l(u,z)\big] \geq \mathrm{E}_{z\sim\mathscr{D}}\Big[l\left(w^{(t)},z\right)\Big] + \Big\langle u - w^{(t)}, \mathrm{E}\Big[v_t \;\Big|\; w^{(t)}\Big]\Big\rangle$$

and to

$$L_{\mathscr{D}}(w) \geq L_{\mathscr{D}}\left(w^{(t)}\right) + \Big\langle u - w^{(t)}, \mathrm{E}\Big[v_t \;\Big|\; w^{(t)}\Big]\Big\rangle.$$

The previous inequality holds for all u and hence $\mathrm{E}\Big[v_t \;\Big|\; w^{(t)}\Big] \in \partial L_{\mathscr{D}}(w^{(t)})$.

Definition 4.44 (Stochastic Gradient Descent, Version 2)**.** Consider the learning problem $(\mathcal{H},\mathcal{Z},l)$ with $\mathcal{Z} = \mathbb{R}^d \times \mathbb{R}$ and $l\colon \mathcal{H}\times\mathcal{Z}\to\mathbb{R}$ convex and ρ-Lipschitz. The *Stochastic Gradient Descent (SGD)* algorithm with $T\in\mathbb{N}$ steps and step size (*learning rate*) $\eta>0$ for minimizing a differentiable loss function $l\colon \mathcal{H}\times\mathcal{Z}\to\mathbb{R}$ works as follows:

Algorithm 5 Stochastic Subgradient Descent (SGD) for loss minimization

1: Set $w^{(1)} := 0 \in \mathbb{R}^d$;
2: **for** $t = 1,\dots,T-1$ **do**
3: Generate $z\sim\mathscr{D}$ independently from $w^{(t)}$;
4: Choose $v_t \in \partial l(w^{(t)},z)$;
5: Update $w^{(t+1)} := w^{(t)} - \eta v_t$;
6: **end for**
7: **return** $\bar{w} := \sum_{t=1}^{T} w^{(t)}/T$.

Corollary 4.45. *Consider a convex, Lipschitz and bounded learning problem with parameters $\rho>0$ and $B>0$, i.e. $\mathcal{H}\subset\{w\in\mathbb{R}^d \mid \|w\|\leq B\}$ and $w\mapsto l(w,z)$ is convex and ρ-Lipschitz. For any $\varepsilon>0$, if we run the SGD algorithm for minimizing $w\mapsto L_{\mathscr{D}}(w)$ over $\mathcal{H}$ with*

$$T \geq \frac{B^2\rho^2}{\varepsilon^2} \text{ and } \eta = \sqrt{\frac{B^2}{\rho^2 T}},$$

then

$$\mathrm{E}\big[L_{\mathscr{D}}(\bar{w})\big] \leq \min_{w\in\mathcal{H}} L_{\mathscr{D}}(w) + \varepsilon.$$

Proof. Follows from Theorem 4.43. □

Motivation. We now proceed to analyze the SGD for convex and smooth loss functions.

Theorem 4.46. *Assume that for all z, $w\mapsto l(w,z)$ is convex and β-smooth for some $\beta>0$. Then, if we run the SGD algorithm with $\eta<1/\beta$, we have for all $w^*\in\mathbb{R}^d$ that*

$$\mathrm{E}\big[L_{\mathscr{D}}\left(\bar{w}\right)\big] \leq \frac{1}{1-\eta\beta}\left(L_{\mathscr{D}}\left(w^*\right) + \frac{\|w^*\|^2}{2\eta T}\right).$$

Proof. Recall that if f is a β-smooth and non-negative function, then $\|\nabla f(w)\|^2 \le 2\beta f(w)$. For $t \in \{1, \dots, T\}$, let us denote $z_t := z \sim \mathscr{D}$ at iteration t such that z_t is independent of $w^{(t)}$ and write $f_t(w) := l(w, z_t)$. As we are in the differentiable case, $v_t = \nabla f_t(w^{(t)})$. By convexity of f_t, Lemma 4.7 implies

$$f_t\left(w^{(t)}\right) \le f_t\left(w^*\right) + \left\langle w^{(t)} - w^*, v_t \right\rangle.$$

Summing over $t = 1, \dots, T$ yields

$$\sum_{t=1}^{T} f_t\left(w^{(t)}\right) \le \sum_{t=1}^{T} f_t\left(w^*\right) + \sum_{t=1}^{T}\left\langle w^{(t)} - w^*, v_t \right\rangle \le \sum_{t=1}^{T} f_t\left(w^*\right) + \frac{\|w^*\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|v_t\|^2.$$

by Lemma 4.32. Using self-boundedness of f_t, it follows that

$$\sum_{t=1}^{T}\|v_t\|^2 \le 2\beta \sum_{t=1}^{T} f_t\left(w^{(t)}\right).$$

Hence,

$$\frac{1}{T}\sum_{t=1}^{T} f_t\left(w^{(t)}\right) \le \frac{1}{T}\sum_{t=1}^{T} f_t\left(w^*\right) + \frac{\|w^*\|^2}{2\eta T} + \frac{\eta\beta}{T}\sum_{t=1}^{T} f_t\left(w^{(t)}\right),$$

which is equivalent to

$$\begin{aligned}
&\frac{1-\eta\beta}{T}\sum_{t=1}^{T} f_t\left(w^{(t)}\right) \le \frac{1}{T}\sum_{t=1}^{T} f_t\left(w^*\right) + \frac{\|w^*\|^2}{2\eta T} \\
&\iff \frac{1}{T}\sum_{t=1}^{T} f_t\left(w^{(t)}\right) \le \frac{1}{1-\eta\beta}\left(\frac{1}{T}\sum_{t=1}^{T} f_t\left(w^*\right) + \frac{\|w^*\|^2}{2\eta T}\right) \\
&\implies \frac{1}{T}\sum_{t=1}^{T} \mathrm{E}\left[f_t\left(w^{(t)}\right)\right] \le \frac{1}{1-\eta\beta}\left(\frac{1}{T}\sum_{t=1}^{T} \mathrm{E}\left[f_t\left(w^*\right)\right] + \frac{\|w^*\|^2}{2\eta T}\right)
\end{aligned}$$

Now, for all $t = 1, \dots, T$ holds

$$\mathrm{E}\left[f_t\left(w^*\right)\right] = \mathrm{E}_{z_t \sim \mathscr{D}}\left[l\left(w^*, z_t\right)\right] = L_{\mathscr{D}}\left(w^*\right)$$

independently of t. Consequently,

$$L_{\mathscr{D}}\left(w^*\right) = \frac{1}{T}\sum_{t=1}^{T} \mathrm{E}\left[f_t\left(w^*\right)\right].$$

Furthermore,

$$\mathrm{E}\left[f_t\left(w^{(t)}\right)\right] = \mathrm{E}\left[l\left(w^{(t)}, z_t\right)\right] = \mathrm{E}\left[L_{\mathscr{D}}\left(w^{(t)}\right)\right].$$

Hence,

$$\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\Big[f_t\left(w^{(t)}\right)\Big] = \frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\Big[L_{\mathscr{D}}\left(w^{(t)}\right)\Big] = \mathrm{E}\left[\frac{1}{T}\sum_{t=1}^{T}L_{\mathscr{D}}\left(w^{(t)}\right)\right].$$

The function $w \mapsto L_{\mathscr{D}}(w) = \mathrm{E}_{z\sim\mathscr{D}}\big[l(w,z)\big]$ is convex (by convexity of $w \mapsto l(w,z)$ uniformly over z). Hence,

$$\frac{1}{T}\sum_{t=1}^{T}L_{\mathscr{D}}\left(w^{(t)}\right) \geq L_{\mathscr{D}}\left(\frac{1}{T}\sum_{t=1}^{T}w^{(t)}\right)$$

by Jensen's inequality. We conclude that

$$\mathrm{E}\big[L_{\mathscr{D}}\left(\bar{w}\right)\big] \leq \frac{1}{1-\eta\beta}\left(L_{\mathscr{D}}\left(w^*\right) + \frac{\|w^*\|^2}{2\eta T}\right),$$

as claimed. □

Corollary 4.47. *Consider a convex, smooth, bounded learning problem with parameters $\beta > 0$ and $B > 0$. Assume in addition that $l(0,z) \leq 1$ for all $z \in \mathcal{Z}$. For all $\varepsilon \in (0,1]$, set $\eta := 1/(\beta(1+3/\varepsilon))$. Then, running the SGD algorithm with $T \geq 12B^2\beta/\varepsilon^2$ yields*

$$\mathrm{E}\big[L_{\mathscr{D}}\left(\bar{w}\right)\big] \leq \min_{w\in\mathcal{H}} L_{\mathscr{D}}(w) + \varepsilon$$

for $\mathcal{H} = \Big\{w \in \mathbb{R}^d \;\Big|\; \|w\| \leq B\Big\}$.

Proof. Note that

$$1-\eta\beta = \frac{3}{\varepsilon+3} \iff \frac{1}{1-\eta\beta} = \frac{\varepsilon+3}{3} = 1+\frac{\varepsilon}{3}.$$

By Theorem 4.46, we know that with $w^* = \operatorname{argmin}_{w\in\mathcal{H}} L_{\mathscr{D}}(w)$

$$\begin{aligned}
\mathrm{E}\big[L_{\mathscr{D}}\left(\bar{w}\right)\big] &\leq \left(1+\frac{\varepsilon}{3}\right)\left(L_{\mathscr{D}}\left(w^*\right) + \frac{B^2}{2\eta T}\right) \\
&\leq \left(1+\frac{\varepsilon}{3}\right)L_{\mathscr{D}}\left(w^*\right) + \left(1+\frac{\varepsilon}{3}\right)\frac{\varepsilon^2+3\varepsilon}{24} \\
&= L_{\mathscr{D}}\left(w^*\right) + \frac{\varepsilon}{3}L_{\mathscr{D}}\left(w^*\right) + \left(1+\frac{\varepsilon}{3}\right)\frac{\varepsilon+3}{24}\varepsilon \\
&\leq L_{\mathscr{D}}\left(w^*\right) + \frac{\varepsilon}{3} + \left(1+\frac{1}{3}\right)\frac{4}{24}\varepsilon \\
&= L_{\mathscr{D}}\left(w^*\right) + \varepsilon\frac{5}{9} \\
&\leq L_{\mathscr{D}}\left(w^*\right) + \varepsilon,
\end{aligned}$$

where we also used that

$$L_{\mathscr{D}}\left(w^*\right) \leq L_{\mathscr{D}}(0) = \mathrm{E}_{z\sim\mathscr{D}}\big[l(0,z)\big] \leq 1.$$

□

4.3 Regularization and stability

4.3.1 Regularized loss minimization (RLM)

Definition 4.48. *Regularized loss minimization (RLM)* is a learning rule which minimizes the criterion

$$w \mapsto L_S(w) + \mathcal{R}(w)$$

for a given *regularization function* $\mathcal{R}$. If

$$\mathcal{R}(w) = \lambda \|w\|^2 = \lambda \sum_{i=1}^{d} w_i^2$$

for some $\lambda > 0$, then we talk about *Tikhonov regularization.*

Example 4.49 (Ridge Regression)**.** When applying Tikhonov regularization to linear regression with $l_{\text{sq}}(w) = (\langle w, x\rangle - y)^2$, we obtain the learning rule

$$\hat{w} \in \operatorname*{argmin}_{w\in\mathbb{R}^d} \left(\lambda \|w\|^2 + \frac{1}{2n} \sum_{i=1}^{n} \left(\langle w, x_i\rangle - y_i \right)^2 \right) = \operatorname*{argmin}_{w\in\mathbb{R}^d} g(w)$$

for

$$g(w) := \lambda \|w\|^2 + \frac{1}{2n} \sum_{i=1}^{n} \left(\langle w, x_i\rangle - y_i \right)^2 = \lambda w^\top w + \frac{1}{2n} \sum_{i=1}^{n} \left(x_i^\top w - y_i \right)^2 .$$

Note that

$$\nabla g(w) = 2\lambda w + \frac{1}{n} \sum_{i=1}^{n} x_i \left(x_i^\top w - y_i \right)$$

and setting $\nabla g(w) = 0$ yields

$$\hat{w} = (A + 2n\lambda \mathbf{I})^{-1} \left(\sum_{i=1}^{n} y_i x_i \right) \text{ for } A := \sum_{i=1}^{n} x_i x_i^\top . \tag{22}$$

Note that A is positive semidefinite, as for all $a \in \mathbb{R}^d$ holds

$$a^\top A a = a^\top \sum_{i=1}^{n} x_i x_i^\top a = \sum_{i=1}^{n} a^\top x_i x_i^\top a = \sum_{i=1}^{n} \left(x_i^\top a \right)^2 \geq 0.$$

As A is also symmetric, there exists $\lambda_1, \ldots, \lambda_d \geq 0$ and $\Omega \in \mathbb{R}^{d\times d}$ orthogonal such that

$$A = \Omega^\top \operatorname{diag}(\lambda_1, \ldots, \lambda_d)\Omega.$$

Consequently,

$$A + 2n\lambda \mathbf{I} = \Omega^\top \operatorname{diag}(\lambda_1 + 2n\lambda, \ldots, \lambda_d + 2n\lambda)\Omega.$$

Since $\lambda_i + 2n\lambda > 0$ for all $i = 1, \ldots, d$, the matrix $A + 2n\lambda\mathbf{I}$ is invertible and $\hat{w}$ is well-defined by (22).

Motivation (Stable rules do not overfit)**.** Let $\mathscr{A}$ be a learning algorithm and $\mathcal{S} = \{z_1, \dots, z_n\}$ be some training set with n i.i.d. examples from the unknown distribution $\mathscr{D}$. In order to assess stability of $\mathscr{A}$, we look at the influence of replacing one example z_i by some $z' \sim \mathscr{D}$. Given $\mathcal{S}$ and an additional example $z' \sim \mathscr{D}$ independent of $\mathcal{S}$, let

$$\mathcal{S}^{(i)} := \{z_1, \dots, z_{i-1}, z', z_{i+1}, \dots, z_n\}.$$

We investigate the difference

$$l(\mathscr{A}(\mathcal{S}^{(i)}), z_i) - l(\mathscr{A}(\mathcal{S}), z_i)$$

of the loss with respect to z_i between the algorithm trained with versus without z_i.

Theorem 4.50. *Let $\mathscr{D}$ be some distribution on $\mathcal{Z}$ and $\mathcal{S} = \{z_1, \dots, z_n\}$ be a training set where $z_1, \dots, z_n$ are i.i.d. according to $\mathscr{D}$ and $z' \sim \mathscr{D}$ independent of $\mathcal{S}$. Then, for any learning algorithm $\mathscr{A}$, we have that*

$$\mathrm{E}_{\mathcal{S}\sim\mathscr{D}^n}\Big[L_{\mathscr{D}}(\mathscr{A}(\mathcal{S})) - L_{\mathcal{S}}(\mathscr{A}(\mathcal{S}))\Big] = \mathrm{E}_{(\mathcal{S},z',i)\sim\mathscr{D}^{n+1}\otimes\mathcal{U}[n]}\Big[l(\mathscr{A}(\mathcal{S}^{(i)}), z_i) - l(\mathscr{A}(\mathcal{S}), z_i)\Big], \quad (23)$$

where $[n] = \{1, \dots, n\}$.

Proof. Note that by independence,

$$\begin{aligned}
\mathrm{E}_{\mathcal{S}\sim\mathscr{D}^n}\Big[L_{\mathscr{D}}(\mathscr{A}(\mathcal{S}))\Big] &= \mathrm{E}_{\mathcal{S}\sim\mathscr{D}^n}\Big[\mathrm{E}_{z'\sim\mathscr{D}}\Big[l\left(\mathscr{A}(\mathcal{S}), z'\right)\Big]\Big] \\
&= \mathrm{E}_{(\mathcal{S},z')\sim\mathscr{D}^{n+1}}\Big[l\left(\mathscr{A}(\mathcal{S}), z'\right)\Big] \\
&= \mathrm{E}_{(\mathcal{S},z')\sim\mathscr{D}^{n+1}}\Big[l\left(\mathscr{A}\left(z_1, \dots, z_{i-1}, z', z_{i+1}, \dots, z_n\right), z_i\right)\Big] \\
&= \mathrm{E}_{(\mathcal{S},z')\sim\mathscr{D}^{n+1}}\Big[l(\mathscr{A}(\mathcal{S}^{(i)}), z_i)\Big],
\end{aligned}$$

since z_i and z' are exchangable. As the last expectation does not depend on the index i, it follows that

$$\mathrm{E}_{(\mathcal{S},z')\sim\mathscr{D}^{n+1}}\Big[l(\mathscr{A}(\mathcal{S}^{(i)}), z_i)\Big] = \mathrm{E}_{(\mathcal{S},z',i)\sim\mathscr{D}^{n+1}\otimes\mathcal{U}[n]}\Big[l(\mathscr{A}(\mathcal{S}^{(i)}), z_i)\Big]. \quad (24)$$

On the other hand, we have that

$$\mathrm{E}_{\mathcal{S}\sim\mathscr{D}^n}\Big[L_{\mathcal{S}}(\mathscr{A}(\mathcal{S}))\Big] = \mathrm{E}_{\mathcal{S}\sim\mathscr{D}^n}\left[\frac{1}{n}\sum_{i=1}^{n} l(\mathscr{A}(\mathcal{S}), z_i)\right] = \mathrm{E}_{(\mathcal{S},i)\sim\mathscr{D}^n\otimes\mathcal{U}[n]}\Big[l(\mathscr{A}(\mathcal{S}), z_i)\Big]$$

by the definition of the training error / risk. Since the last expectation does not depend on z', it follows that

$$\mathrm{E}_{(\mathcal{S},i)\sim\mathscr{D}^n\otimes\mathcal{U}[n]}\Big[l(\mathscr{A}(\mathcal{S}), z_i)\Big] = \mathrm{E}_{(\mathcal{S},z',i)\sim\mathscr{D}^{n+1}\otimes\mathcal{U}[n]}\Big[l(\mathscr{A}(\mathcal{S}), z_i)\Big]. \quad (25)$$

Taking the difference of (24) and (25) completes the proof. □

Remark 4.51. When the right term of (23) is small, this is an indication that replacing z_i by "something else" does not influence the algorithm "too much". In other words, we say that the learning algorithm is *stable* (a change in a single example does not result in a big change on average).

Definition 4.52. Let $\varepsilon\colon \mathbb{N} \to \mathbb{R}$ be a decreasing function. We say that a learning algorithm $\mathscr{A}$ is *on-average-replace-one-stable with rate* $\varepsilon(n)$, if

$$\forall\, \mathscr{D} \text{ distribution } \forall\, \mathcal{S} \sim \mathscr{D}^n\colon \quad \mathrm{E}_{(\mathcal{S},z',i)\sim \mathscr{D}^{n+1}\otimes \mathcal{U}[n]}\Big[l(\mathscr{A}(\mathcal{S}^{(i)}), z_i) - l(\mathscr{A}(\mathcal{S}), z_i)\Big] \leq \varepsilon(n). \tag{26}$$

Remark 4.53. In view of Theorem 4.50, when (26) holds, then for all distribution $\mathscr{D}$, we have that

$$\mathrm{E}_{\mathcal{S}\sim\mathscr{D}^n}\big[L_{\mathscr{D}}(\mathscr{A}(\mathcal{S})) - L_{\mathcal{S}}(\mathscr{A}(\mathcal{S}))\big] \leq \varepsilon(n).$$

This means that $\mathscr{A}$ does not overfit. Hence a "good" learning algorithm should balance between fitting and staying stable.

4.3.2 Tikhonov regularisation as a stabiliser

In the following, we are going to apply Tikhonov regularisation to a convex and ρ-Lipschitz loss function l.

Definition 4.54. A function $f\colon \mathbb{R}^d \to \mathbb{R}$ is said to be λ-*strongly convex* for $\lambda > 0$ if

$$\forall\, u, w \in \mathbb{R}^d, \alpha \in [0,1] \quad f(\alpha w + (1-\alpha)u) \leq \alpha f(w) + (1-\alpha) f(u) - \frac{\lambda}{2}\alpha(1-\alpha)\|w-u\|^2.$$

Lemma 4.55. *The following assertions hold:*

1) The function $f\colon \mathbb{R}^d \to \mathbb{R}$, $w \mapsto \lambda\|w\|^2$ is (2λ)-strongly convex;

2) If $f\colon \mathbb{R}^d \to \mathbb{R}$ is λ-strongly convex and $g\colon \mathbb{R}^d \to \mathbb{R}$ is convex, then $f+g$ is λ-strongly convex ;

3) If $f\colon \mathbb{R}^d \to \mathbb{R}$ is λ-strongly convex and $u \in \mathbb{R}^d$ is a minimizer of f, then

$$\forall\, w \in \mathbb{R}^d \quad f(w) - f(u) \geq \frac{\lambda}{2}\|w-u\|^2.$$

Proof.

1) Note that for all $u, w \in \mathbb{R}^d$ holds $2\langle u, w\rangle = \|u\|^2 + \|w\|^2 - \|u-w\|^2$. Then for $\alpha \in [0,1]$ holds

$$\begin{aligned} f(\alpha w + (1-\alpha)u) &= \lambda\big\|\alpha w + (1-\alpha)u\big\|^2 \\ &= \lambda\alpha^2\|w\|^2 + \lambda(1-\alpha)^2\|u\|^2 + 2\lambda\alpha(1-\alpha)\langle w, u\rangle \end{aligned}$$

$$
\begin{aligned}
&= \lambda\alpha^2\|w\|^2 + \lambda(1-\alpha)^2\|u\|^2 + \lambda\alpha(1-\alpha)\|w\| + \\
&\qquad + \lambda\alpha(1-\alpha)\|u\| - \lambda\alpha(1-\alpha)\|w-u\|^2 \\
&= \lambda\alpha\|w\|^2 + \lambda(1-\alpha)\|u\|^2 - \lambda\alpha(1-\alpha)\|w-u\|^2 \\
&= \alpha f(w) + (1-\alpha) f(u) - \frac{2\lambda}{2}\alpha(1-\alpha)\|w-u\|^2.
\end{aligned}
$$

2) Note that for all $u, w \in \mathbb{R}^d$ and $\alpha \in [0,1]$ holds

$$
f(\alpha w + (1-\alpha)u) + g(\alpha w + (1-\alpha)u) \le \alpha(f(w)+g(w)) + (1-\alpha)(f(u)+g(u)) + \frac{\lambda}{2}\|w-u\|^2.
$$

3) Note that for all $\alpha \in (0,1]$ holds

$$
f(u + \alpha(w-u)) = f(\alpha w + (1-\alpha)u) \le \alpha f(w) + (1-\alpha) f(u) - \frac{\lambda}{2}\alpha(1-\alpha)\|w-u\|^2,
$$

which is equivalent to

$$
\frac{1}{\alpha}\big(f(u+\alpha(w-u)) - f(u)\big) \le f(w) - f(u) - \frac{\lambda}{2}(1-\alpha)\|w-u\|^2.
$$

Since $f(u) = \min_{v\in\mathbb{R}^d} f(v)$, it holds $f(u+\alpha(w-u)) \ge f(u)$ and consequently

$$
0 \le f(w) - f(u) - \frac{\lambda}{2}(1-\alpha)\|w-u\|^2,
$$

which is equivalent to

$$
f(w) \ge f(u) + \frac{\lambda}{2}(1-\alpha)\|w-u\|^2
$$

for all $\alpha \in (0,1]$. Taking $\alpha \searrow 0$, we obtain the claim. □

Corollary 4.56. *If the loss function is convex and ρ-Lipschitz, then the Tikhonov RLM rule is on-average-replace-one-stable with rate $\varepsilon(n) = 2\rho^2/(\lambda n)$. We have*

$$
\mathrm{E}_{S\sim\mathscr{D}^n}\big[L_{\mathscr{D}}(\mathscr{A}(S)) - L_S(\mathscr{A}(S))\big] \le \frac{2\rho^2}{n\lambda}.
$$

Proof. For a training set $S = \{z_1, \dots, z_n\}$, consider the Tikhonov RLM

$$
\mathscr{A}(S) :\in \operatorname*{argmin}_{w\in\mathbb{R}^d}\Big(L_S(w) + \lambda\|w\|^2\Big).
$$

Let us write $f_S(w) := L_S(w) + \lambda\|w\|^2$. Under convexity of $w \mapsto l(w,z)$ uniformly over $\mathcal{Z}$ and using Lemma 4.55, we see that f_S is (2λ)-strongly convex. Also by Lemma 4.55, we have for all $v \in \mathbb{R}^d$ that

$$
f_S(v) - f_S(\mathscr{A}(S)) \ge \lambda\big\|\mathscr{A}(S) - v\big\|^2.
$$

For any $u, v \in \mathbb{R}^d$ and index $i \in \{1, \dots, n\}$, it holds that

$$\begin{aligned} f_S(v) - f_S(u) &= L_S(v) + \lambda\|v\|^2 - L_S(u) - \lambda\|u\|^2 \\ &= L_{S^{(i)}}(v) - \frac{1}{n}l(v, z') + \frac{1}{n}l(v, z_i) + \lambda\|v\|^2 + \\ &\quad - \left(L_{S^{(i)}}(u) - \frac{1}{n}l(u, z') + \frac{1}{n}l(u, z_i) + \lambda\|u\|^2\right) \\ &= L_{S^{(i)}}(v) + \lambda\|v\|^2 - \left(L_{S^{(i)}}(u) + \lambda\|u\|^2\right) + \\ &\quad + \frac{l(v, z_i) - l(v, z')}{n} - \frac{l(u, z_i) - l(u, z')}{n}. \end{aligned}$$

Now take $u = \mathscr{A}(S)$ and $v = \mathscr{A}(S^{(i)})$. Since v minimizes $w \mapsto L_{S^{(i)}}(w) + \lambda\|w\|^2$, it follows that

$$\begin{aligned} \lambda\left\|\mathscr{A}(S^{(i)}) - \mathscr{A}(S)\right\|^2 &\le f_S(\mathscr{A}(S^{(i)})) - f_S(\mathscr{A}(S)) \\ &\le \frac{l(\mathscr{A}(S^{(i)}), z_i) - l(\mathscr{A}(S), z_i)}{n} + \frac{l(\mathscr{A}(S), z') - \mathscr{A}(S^{(i)}), z'))}{n}. \end{aligned}$$

If the loss function is ρ-Lipschitz (in addition to convexity), then

$$\lambda\left\|\mathscr{A}(S^{(i)}) - \mathscr{A}(S)\right\|^2 \le \frac{2\rho}{n}\left\|\mathscr{A}(S^{(i)}) - \mathscr{A}(S)\right\|.$$

Therefore, we obtain

$$\left\|\mathscr{A}(S^{(i)}) - \mathscr{A}(S)\right\| \le \frac{2\rho}{n\lambda}.$$

This also implies by ρ-Lipschitzness that for all S, i, z' with z' independent of S,

$$l(\mathscr{A}(S^{(i)}), z_i) - l(\mathscr{A}(S), z_i) \le \left|l(\mathscr{A}(S^{(i)}), z_i) - l(\mathscr{A}(S), z_i)\right| \le \frac{2\rho^2}{n\lambda}.$$

Theorem 4.50 then yields the claim. □

4.3.3 Controlling the fitting-stability trade-off

Corollary 4.57. *If the loss function is convex and ρ-Lipschitz, then the Tikhonov RLM rule $\mathscr{A}$ satisfies*

$$\forall\, w^* \in \mathbb{R}^d \quad \mathrm{E}_{S\sim\mathscr{D}^n}\left[L_{\mathscr{D}}(\mathscr{A}(S))\right] \le L_{\mathscr{D}}(w^*) + \lambda\left\|w^*\right\|^2 + \frac{2\rho^2}{n\lambda}.$$

Proof. For any learning algorithm $\mathscr{A}$ we have

$$\mathrm{E}_{S\sim\mathscr{D}^n}\left[L_{\mathscr{D}}(\mathscr{A}(S))\right] = \underbrace{\mathrm{E}_{S\sim\mathscr{D}^n}\left[L_S(\mathscr{A}(S))\right]}_{\text{fit}} + \underbrace{\mathrm{E}_{S\sim\mathscr{D}^n}\left[L_{\mathscr{D}}(\mathscr{A}(S)) - L_S(\mathscr{A}(S))\right]}_{\text{stability}}.$$

Let

$$\mathscr{A}(\mathcal{S}) :\in \operatorname*{argmin}_{w\in\mathbb{R}^d} \left(L_{\mathcal{S}}(w) + \lambda \|w\|^2 \right)$$

be the Tikhonov RLM. Then we have for all $w^* \in \mathbb{R}^d$ that

$$L_{\mathcal{S}}(\mathscr{A}(\mathcal{S})) \leq L_{\mathcal{S}}(\mathscr{A}(\mathcal{S})) + \lambda \left\|\mathscr{A}(\mathcal{S})\right\|^2 \leq L_{\mathcal{S}}(w^*) + \lambda \left\|w^*\right\|^2.$$

Note that $\mathrm{E}_{\mathcal{S}\sim\mathscr{D}^n}\left[L_{\mathcal{S}}(w^*)\right] = L_{\mathscr{D}}(w^*)$ and hence

$$\mathrm{E}_{\mathcal{S}\sim\mathscr{D}^n}\left[L_{\mathcal{S}}(\mathscr{A}(\mathcal{S}))\right] \leq L_{\mathscr{D}}(w^*) + \lambda \left\|w^*\right\|^2.$$

This implies that

$$\mathrm{E}_{\mathcal{S}\sim\mathscr{D}^n}\left[L_{\mathscr{D}}(\mathscr{A}(\mathcal{S}))\right] \leq L_{\mathscr{D}}(w^*) + \lambda \left\|w^*\right\|^2 + \mathrm{E}_{\mathcal{S}\sim\mathscr{D}^n}\left[L_{\mathscr{D}}(\mathscr{A}(\mathcal{S})) - L_{\mathcal{S}}(\mathscr{A}(\mathcal{S}))\right].$$

Corollary 4.56 then yields the claim. □

We will take $w^* :\in \operatorname{argmin}_{w\in\mathcal{B}_B(0)\subset\mathbb{R}^d} L_{\mathscr{D}}(w)$ to arrive at:

Corollary 4.58. *Let $(\mathcal{H}, \mathcal{Z}, l)$ for $\mathcal{H} = \mathcal{B}_B(0) \subset \mathbb{R}^d$ be a convex, Lipschitz and bounded learning problem with parameters $\rho, B > 0$. Also, set $\lambda := \sqrt{2\rho^2/(nB^2)}$. Then, the Tikhonov RLM rule satisfies*

$$\mathrm{E}_{\mathcal{S}\sim\mathscr{D}^n}\left[L_{\mathscr{D}}(\mathscr{A}(\mathcal{S}))\right] \leq \min_{w\in\mathcal{H}} L_{\mathscr{D}}(w) + \rho B\sqrt{\frac{8}{n}}.$$

Proof. Follows from Corollary 4.57. □

Works Cited

Shalev-Shwartz, Shai and Shai Ben-David. *Understanding machine learning : from theory to algorithms*. New York, NY: Cambridge University Press, 2014. Print.