Mathematical aspects in machine learning

Lorenz Richter
BTU Cottbus-Senftenberg

March 6, 2018

Contents

1 Statistical learning theory **1**
1.1 Concentration inequalities 4
1.2 Error bounds for finite classes 7
1.3 Error bounds for infinite classes 9

2 Kernels **15**

3 Algorithms **18**
3.1 Neural networks 18
3.2 Adaptive boosting 23
3.3 Support vector machines 25

4 Variational inference **28**
4.1 A unifying framework for implicit generative models 31

5 Bibliography **36**

1 Statistical learning theory

We consider data from an input space $\mathcal{X}$ and an output space $\mathcal{Y}$, specifically the sample $S_n = ((X_i, Y_i))_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$. The goal is to learn a (prediction) function $f : \mathcal{X} \to \mathcal{Y}$ that maps input to output data.

Example 1.1. $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{-1, 1\}$ *(binary classification),* $\mathcal{Y} = \mathbb{N}$, $\mathcal{Y} = \mathbb{R}$ *(regression).*

In order to develop a proper theory we need to make some assumptions:

- There exists an unknown probability distribution $\mathbb{P}$ over $\mathcal{X} \times \mathcal{Y}$.
- The data S_n are i.i.d. from $\mathbb{P}$, i.e. $(X_i, Y_i) \sim \mathbb{P}$ for every $i = 1, \dots, n$.
- The future data (sometimes called test data) also come from $\mathbb{P}$.

In order to measure how good we learn the prediction function f we consider a loss function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ that indicates deviations between predictions and true values. Being stochastically more meaningful, we consider for any function f the expected loss

$$L(f) := \mathbb{E}_{(X,Y)\sim\mathbb{P}}[l(f(X), Y)],$$

which in statistics is sometimes called "(Bayes-)Risk". The goal is to come up with a learning algorithm $\mathcal{A} : S_n \mapsto \hat{f}_n$ (i.e. $\hat{f}_n := \mathcal{A}(S_n)$, indicating that the function depends on the n training data) s.t. $L(\hat{f}_n)$ is small. We therefore consider $\mathbb{E}[l(\hat{f}_n(X), Y)|S_n]$, i.e. the expectation conditioned on the samples, which is still a random quantity (as it depends on the sample data)[1].

[1] We will continuously omit the measure $\mathbb{P}$ in the expected value.

Definition 1.2. *A predictor f^B is called Bayes-optimal if it minimizes the expected loss, i.e.*

$$L(f^B) = \inf_f L(f) =: L^B.$$

Remark 1.3. *Bayes-optimality depends on $\mathbb{P}, \mathcal{X} \times \mathcal{Y}$ and l.*

Example 1.4 (Zero noise or function learning)**.** *One could consider the case that our targets Y are deterministically prescribed by a function g, i.e. $\mathbb{P}(Y = g(X)|X = x) = 1$. However, this is a rather unrealistic case.*

Example 1.5 (Binary classification)**.** *Consider $\mathcal{Y} = \{0, 1\}$ and the loss $l(y', y) = \mathbb{1}\{y' \neq y\}$. One can show that the Bayes-optimal predictor (in this case classifier) is $f^B(x) = \mathbb{1}\{\eta(x) > \frac{1}{2}\}$ with $\eta(x) = \mathbb{P}(Y = 1|X = x)$. We have*

$$\begin{aligned}\mathbb{P}(f(X) \neq Y|X = x) &= 1 - \mathbb{P}(f(X) = Y|X = x) \\ &= 1 - \left(\mathbb{1}_{\{f(x)=1\}}\eta(x) - \mathbb{1}_{\{f(x)=-1\}}(1 - \eta(x))\right).\end{aligned}$$

This yields

$$\begin{aligned}\mathbb{P}(f(X) \neq Y|X = x) - \mathbb{P}(f^B(X) \neq Y|X = x) &= \eta(x)\left(\mathbb{1}_{\{f^B(x)=1\}} - \mathbb{1}_{\{f(x)=1\}}\right) + (1 - \eta(x))\left(\mathbb{1}_{\{f^B(x)=-1\}} - \mathbb{1}_{\{f(x)=-1\}}\right) \\ &= (2\eta(x) - 1)\left(\mathbb{1}_{\{f^B(x)=1\}} - \mathbb{1}_{\{f(x)=1\}}\right) \geq 0\end{aligned}$$

and by integration w.r.t. x the statement

Example 1.6 (Regression)**.** *We consider $\mathcal{Y} = \mathbb{R}$ and $l(y', y) = (y' - y)^2$. Then the Bayes-optimal predictor is $f^B(x) = \mathbb{E}[Y|X = x]$, which we can see by noting that*

$$\begin{aligned}L(f) &= \mathbb{E}\left[(f(X) - Y)^2\right] \\ &= \mathbb{E}\left[(f(X) - f^B(X) + f^B(X) - Y)^2\right] \\ &= \mathbb{E}\left[(f(X) - f^B(X))^2\right] + \mathbb{E}\left[(f^B(X) - Y)^2\right] - 2\mathbb{E}\left[(f(X) - f^B(X))(f^B(X) - Y)\right]\end{aligned}$$

is minimized by $f = f^B$ since the last term is equal to

$$\mathbb{E}_X\left[\mathbb{E}_{Y|X}\left[(f(X) - f^B(X))(f^B(X) - Y)\right]\right] = \mathbb{E}_X\left[(f(X) - f^B(X))\mathbb{E}_{Y|X}\left[(f^B(X) - Y)\right]\right] = 0.$$

Let us come back to finding a good function f. Somehow we need to work with our sample S_n. We therefore define the empirical risk to be

$$L_n(f) = \frac{1}{n}\sum_{i=1}^{n} l(f(X_i), Y_i).$$

We see that $\mathbb{E}[L_n(f)] = L(f)$ and can therefore consider empirical risk minimization (ERM), namely

$$\hat{f}_n := \arg\inf_f L_n(f)$$

as a reasonable strategy to learn the function f. However, the following example shows that this is not always a good idea.

Example 1.7. *Consider $\mathcal{X} = [0, 1], \mathcal{Y} = \{0, 1\}, l(y', y) = \mathbb{1}\{y' \neq y\}$ and $\mathbb{P}(Y = 1|x) = \frac{1}{2}$ for all $x \in \mathcal{X}$. Define the predictor as*

$$\hat{f}_n(x) = \begin{cases} y & \text{if } (x, y) \in S_n \text{ for some value } y \\ 0 & \text{otherwise}\end{cases}.$$

Note again that this predictor is not fixed, but depends on the data. We can see that it performs very badly: $L_n(\hat{f}_n) = 0$, but $L(\hat{f}_n) = \frac{1}{2}$. This is a stereotypical example of what is described as overfitting.

Theorem 1.8 (No-free-lunch)**.** *Let $\mathcal{A}$ be any learning algorithm for binary classification and let $n < \frac{|\mathcal{X}|}{2}$ be a sample size. Then there exists a probability distribution $\mathbb{P}$ s.t.*

(i) there exists a function $f : \mathcal{X} \to \mathcal{Y}$ with $L(f) = 0$,

(ii) with probability $\geq \frac{1}{7}$ over the sample $S_n \sim \mathbb{P}^n$ we have $L(\mathcal{A}(S_n)) > \frac{1}{8}$.

Proof. See chapter 5 in [30][2]. □

Remark 1.9. *This means that there is no perfect algorithm that works for all distributions. Bayes optimality cannot be obtained independent of $\mathbb{P}$.*

In order to address the problem of overfitting[3], we consider the function class $\mathcal{F}$ (usually a class chosen with prior knowledge regarding the prediction problem) and restrict our predictors to come from this class, i.e. $\hat{f}_n \in \mathcal{F}$. We can then decompose our expected error as

$$L(\hat{f}_n) - L(f^B) = \underbrace{L(\hat{f}_n) - \inf_{f\in\mathcal{F}} L(f)}_{\text{estimation error (variance)}} + \underbrace{\inf_{f\in\mathcal{F}} L(f) - L(f^B)}_{\text{approximation error (bias)}}.$$

Here, $\hat{f}_n$ shall be the predictor that minimizes the empirical error in our hypothesis class, i.e. $L_n(\hat{f}_n) = \inf_{f\in\mathcal{F}} L_n(f)$. In the sequel, we will only study the estimation error, i.e. we aim at statements of the form

$$\mathbb{P}\{L(\hat{f}_n) - L(f^*) > \epsilon\} \leq B(\epsilon, n, \mathcal{F}),$$

where B decreases to 0 as $n \to \infty$ for any ϵ and $f^* = \arg\inf_{f\in\mathcal{F}} L(f)$, i.e. the best predictor in the class $\mathcal{F}$ (we will not study the approximation error here, details can for instance be found in [9]). Equivalently, we will consider the following kind of deviation inequality: With probability $1-\delta$ for $\delta \in (0,1)$ we want

$$L(\hat{f}_n) - L(f^*) \leq D(\delta, n, \mathcal{F}).$$

We will particularly be interested in statements for all $f \in \mathcal{F}$, i.e. we want uniform and not only pointwise bounds. Note that we have $L(\hat{f}_n) - L_n(\hat{f}_n) \leq \sup_{f\in\mathcal{F}}\{L(f) - L_n(f)\}$, so we will then also investigate the difference between expected and empirical loss, $L(\hat{f}_n) - L_n(\hat{f}_n)$, which we call generalization error[4].

More abstractly, we consider the following learnability definition.

Definition 1.10 (PAC learnability [32])**.** *A hypothesis class $\mathcal{F}$ is "probably approximately correctly" (PAC) learnable if there is an algorithm $\mathcal{A}$ such that for all $\epsilon > 0$ and $\delta > 0$ there is a sample size $n(\epsilon, \delta) \in \mathbb{N}$ such that for all distributions $\mathbb{P}$ over $\mathcal{X} \times \mathcal{Y}$*

$$\mathbb{P}(L(\mathcal{A}(S_n)) - L(f^*) < \epsilon) \geq 1 - \delta.$$

Remark 1.11. *One interesting aspect in introducing hypothesis classes $\mathcal{F}$ is model selection. Consider a family of hypothesis classes $\{\mathcal{F}_\alpha\}_{\alpha\in A}$ with $\mathcal{F} = \cup_\alpha \mathcal{F}_\alpha$ (e.g. $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots$). Let $\hat{f}_n^\alpha$ be the predictor corresponding to the ERM over $\mathcal{F}_\alpha$. One considers $\alpha_n := \alpha(S_n)$, i.e. a data-dependent selection of the possible classes. We are then interested in $L(\hat{f}_n^{\alpha_n}) - \inf_{\alpha\in A} \inf_{f\in\mathcal{F}_\alpha} L(f)$.*

Remark 1.12. *If we consider only one possible function, i.e. $\mathcal{F} = \{f_1\}$, then nothing will be learnt and we know from the law of large numbers that*

$$L_n(\hat{f}_n) - L(\hat{f}_n) = \frac{1}{n}\sum_{i=1}^{n} l(f_1(X_i), Y_i) - \mathbb{E}[l(f_1(X), Y)] \to 0$$

for $n \to \infty$ at rate $\frac{1}{\sqrt{n}}$.

So in the rather pathological case of $\mathcal{F} = \{f_1\}$ we have convergence of $L(\hat{f}_n) - L(f^*)$. But does it also hold with N predictors at hand, and if yes, how fast does it happen? To answer these questions the following so-called concentration inequalities will be of help.

[2] The proof is similar to the concept of VC dimension, which we will discuss later.

[3] Another way to address overfitting is to modify the criterion to be minimized by for instance adding a penalty term for too "complicated" functions – see structural risk minimization and (normalized) regularization.

[4] Compare to the uniform law of large numbers.

1.1 Concentration inequalities

Concentration inequalities are statements of the form[5]

$$\mathbb{P}(|\bar{Z}_n - \mathbb{E}[Z]| < \epsilon) \geq 1 - \delta.$$

Theorem 1.13 (Markov inequality)**.** *Consider a random variable Z with $Z \geq 0$. Then for all $\epsilon > 0$ we have*

$$\mathbb{P}(Z > \epsilon) \leq \frac{\mathbb{E}[Z]}{\epsilon}.$$

Proof. See [16]. □

Theorem 1.14 (Chebysheff inequality)**.** *Consider a random variable Z. Then for all $\epsilon > 0$ we have*

$$\mathbb{P}(|Z - \mathbb{E}[Z]| > \epsilon) \leq \frac{\mathrm{Var}(Z)}{\epsilon^2}.$$

Proof. See [16]. □

We motivate the next bound with

Example 1.15. *Consider $X \sim \mathcal{N}(\mu, \sigma^2)$, then we have*

$$\mathbb{P}(|X - \mu| > \epsilon) = \int_{\epsilon}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \mathrm{d}x \leq \int_{\epsilon}^{\infty} \frac{x}{\epsilon\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \mathrm{d}x$$

and therefore $\mathbb{P}(|X - \mu| \geq \epsilon) \leq \frac{\sigma}{\epsilon}\sqrt{\frac{2}{\pi}} e^{-\frac{\epsilon^2}{2\sigma^2}}$. We will see that this is a special case of the following theorem, when replacing σ^2 with the maximal possible variance of a random variable in $[0,1]$ (i.e. $\frac{1}{4}$). It is also intuitive to look at $\bar{X}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$, i.e. $\mathbb{P}(|\bar{X}_n - \mu| > \epsilon) = \mathbb{P}(|Z| > \sqrt{n}\epsilon) \leq \frac{\sigma}{\sqrt{n}\epsilon}\sqrt{\frac{2}{\pi}} e^{-\frac{n\epsilon^2}{2\sigma^2}}$, which decreases exponentially for large n.

Lemma 1.16 (Hoeffding's lemma)**.** *For a random variable $Z \in [a, b]$ with $\mathbb{E}[Z] = 0$ we can bound the moment generating function for all $\lambda \in \mathbb{R}$ as*

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

Proof. Since $e^{\lambda Z}$ is convex we have for all $a \leq Z \leq b$

$$e^{\lambda Z} \leq \frac{b - Z}{b - a} e^{\lambda a} + \frac{Z - a}{b - a} e^{\lambda b}$$

and therefore

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda Z}\right] &\leq \frac{b}{b-a} e^{\lambda a} - \frac{a}{b-a} e^{\lambda b} \\
&= \left(-\frac{a}{b-a}\right) e^{\lambda a} \left(-\frac{b}{a} + e^{\lambda b - \lambda a}\right) \\
&= \left(-\frac{a}{b-a}\right) e^{\lambda a} \left(-\frac{b-a}{a} - 1 + e^{\lambda(b-a)}\right) \\
&= \left(1 - c + ce^{d}\right) e^{-cd} \\
&=: e^{L(d)}
\end{aligned}$$

with $c = -\frac{a}{b-a} > 0$, $d = \lambda(b - a)$ and

$$L(d) = -cd + \log(1 - c + ce^d).$$

[5] Sometimes the notation $Pf = \mathbb{E}[f(X,Y)]$ and $P_n f = \frac{1}{n}\sum_{i=1}^{n} f(X_i, Y_i)$ for the empirical measure is used. Compare this to the theory of empirical processes, where uniform deviations of averages from their expectations are investigated, i.e. $\sup_{f\in\mathcal{F}} \{Pf - P_n f\}$.

Taking derivatives w.r.t. d brings

$$L(0) = L'(0) = 0$$

and

$$L''(d) = \frac{ce^d\left(1 - c + ce^d\right) - c^2e^{2d}}{\left(1 - c + ce^d\right)^2} = \frac{ce^d}{1 - c + ce^d}\left(1 - \frac{ce^d}{1 - c + ce^d}\right) = t(1 - t) \leq \frac{1}{4}$$

with $t = \frac{ce^d}{1-c+ce^d} > 0$. By Taylor's theorem there exists an $e \in [0, d]$ s.t.

$$L(d) = L(0) + dL'(0) + \frac{1}{2}d^2L''(e) \leq \frac{1}{8}d^2 = \frac{\lambda^2}{8}(b - a)^2.$$

This implies the bound (by convexity of the logarithm). □

Theorem 1.17 (Hoeffding's inequality). *Let $Z_1, \ldots, Z_n$ be independent (but not necessarily identically distributed). Assume w.l.o.g. that $Z_i \in [0, 1]$ for $i = 1, \ldots, n$. Then with $\bar{Z}_n = \frac{1}{n}\sum_{i=1}^n Z_i$*

$$\mathbb{P}\left(\left|\bar{Z}_n - \mathbb{E}\left[\bar{Z}_n\right]\right| > \epsilon\right) \leq 2e^{-2n\epsilon^2}.$$

Proof. Consider $Z_i \in [a_i, b_i]$. With Markov's inequality we have for $\lambda > 0$

$$\begin{aligned}
\mathbb{P}\left(\bar{Z}_n - \mathbb{E}\left[\bar{Z}_n\right] > \epsilon\right) &\leq \frac{\mathbb{E}\left[e^{\lambda(\bar{Z}_n - \mathbb{E}[\bar{Z}_n])}\right]}{e^{\lambda\epsilon}} \\
&= \frac{\mathbb{E}\left[e^{\frac{\lambda}{n}\sum_{i=1}^n (Z_i - \mathbb{E}[Z_i])}\right]}{e^{\lambda\epsilon}} \\
&= \frac{\prod_{i=1}^n \mathbb{E}\left[e^{\frac{\lambda}{n}(Z_i - \mathbb{E}[Z_i])}\right]}{e^{\lambda\epsilon}} \\
&\leq \frac{\prod_{i=1}^n e^{\frac{\lambda^2}{8n^2}(b_i - a_i)^2}}{e^{\lambda\epsilon}} \\
&= \exp\left(\frac{\sum_{i=1}^n (b_i - a_i)^2}{8n^2}\lambda^2 - \lambda\epsilon\right),
\end{aligned}$$

where in the last inequality we used Hoeffding's lemma. This general proof technique is called Chernoff bound. We now minimize w.r.t λ and get $\lambda^* = \frac{4n^2\epsilon}{\sum_{i=1}^n (b_i - a_i)^2}$ and therefore

$$\mathbb{P}\left(\bar{Z}_n - \mathbb{E}\left[\bar{Z}_n\right] > \epsilon\right) \leq \exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right),$$

which yields the desired result with $a_i = 0$, $b_i = 1$ and by noting that $\mathbb{P}(|Z| \leq \epsilon) \leq \mathbb{P}(Z \leq \epsilon) + \mathbb{P}(-Z \leq \epsilon)$. □

Remark 1.18. *Note that this theorem also holds for functions of random variables. One can for instance consider $Z = g(X, Y)$ with $g(X, Y) = l(f(X), Y)$. However, also note that it only holds for a fixed function and not uniformly for all $f \in \mathcal{F}$. For a fixed sample one finds an $f \in \mathcal{F}$ that yields a very large error, as for instance seen in example 1.7.*

Remark 1.19. *We can formulate Hoeffding's bound as a deviation inequality by realizing the equivalence of the statements $\mathbb{P}(|Z| > \epsilon) \leq \delta$ and $\mathbb{P}(|Z| < \epsilon) \geq 1 - \delta$. Set $\delta := 2e^{-2n\epsilon^2}$, then $\epsilon = \sqrt{\frac{\log\frac{2}{\delta}}{2n}}$, and we have with probability at least $1 - \delta$ that*

$$\left|\bar{Z}_n - \mathbb{E}\left[\bar{Z}_n\right]\right| \leq \sqrt{\frac{\log\frac{2}{\delta}}{2n}}.$$

Hoeffding's inequality does not use any knowledge about the distribution of variables. Therefore one can consider Bernstein's inequality, which uses the variance of the distribution to get a tighter bound.

Theorem 1.20 (Bernstein's inequality)**.** *Let $Z_1, \dots, Z_n$ be independent (but not necessarily identically distributed). Assume w.l.o.g. $|Z_i| < 1$. Then with probability $1-\delta$ we have*

$$\left|\bar{Z}_n - \mathbb{E}\left[\bar{Z}_n\right]\right| \leq \sqrt{2 \operatorname{Var}\left(\bar{Z}_n\right) \log \frac{2}{\delta}} + \frac{2 \log \frac{2}{\delta}}{3n}.$$

Proof. See [5]. □

Remark 1.21. *If the variance is very small then the first term on the right hand side becomes negligible.*

Theorem 1.22 (McDiarmid's inequality)**.** *Consider a function $g : \mathcal{Z}^n \to \mathbb{R}$. Assume that there exist constants $c_1, \dots, c_n$ s.t.*

$$\sup_{Z_1,\dots,Z_n,Z'} |g(Z_1,\dots,Z_n) - g(Z_1,\dots,Z_{i-1},Z',Z_{i+1},\dots,Z_n)| \leq c_i$$

for all $i = 1, \dots, n$. Let $Z_1, \dots, Z_n \sim \mathbb{P}$ i.i.d. with values in $\mathcal{Z}$ and let $g := g(Z_1, \dots, Z_n)$. Then with probability $\geq 1-\delta$ we have

$$|g - \mathbb{E}\left[g\right]| \leq \sqrt{\frac{1}{2}\left(\sum_{i=1}^n c_i^2\right) \log \frac{2}{\delta}}.$$

Proof. First note that

$$\mathbb{P}\left(|g - \mathbb{E}[g]| \geq \epsilon\right) = \mathbb{P}\left(g - \mathbb{E}[g] \geq \epsilon\right) + \mathbb{P}\left(g - \mathbb{E}[g)] \leq -\epsilon\right),$$

from which we will show the first inequality – the second one works analogously. Define for $i = 2, \dots, n$ the martingale difference sequence

$$V_i := \mathbb{E}[g|Z_1,\dots,Z_i] - \mathbb{E}[g|Z_1,\dots,Z_{i-1}] \qquad \text{and} \qquad V_1 := \mathbb{E}[g|Z_1] - \mathbb{E}[g],$$

then we have

$$g(Z_1,\dots,Z_n) - \mathbb{E}[g(Z_1,\dots,Z_n)] = \sum_{i=1}^n V_i.$$

With a similar argument as in Lemma 1.16 (see [5], section 6.1) we get

$$\mathbb{E}\left[e^{\lambda V_i}|Z_1,\dots,Z_{i-1}\right] \leq \exp\left(\frac{\lambda^2 c_i^2}{8}\right).$$

Then we have for all $\lambda > 0$

$$\begin{aligned}
\mathbb{P}\left(g(Z_1,\dots,Z_n) - \mathbb{E}[g(Z_1,\dots,Z_n)] \geq \epsilon\right) &= \mathbb{P}\left(\sum_{i=1}^n V_i \geq \epsilon\right) \\
&= \mathbb{P}\left(e^{\lambda \sum_{i=1}^n V_i} \geq e^{\lambda\epsilon}\right) \\
&\leq e^{-\lambda\epsilon} \mathbb{E}\left[e^{\lambda \sum_{i=1}^n V_i}\right] \\
&= e^{-\lambda\epsilon} \mathbb{E}\left[e^{\lambda \sum_{i=1}^{n-1} V_i} \mathbb{E}\left[e^{\lambda V_n}|Z_1,\dots,Z_{n-1}\right]\right] \\
&\leq e^{-\lambda\epsilon} e^{\frac{\lambda^2 c_n^2}{8}} \mathbb{E}\left[e^{\lambda \sum_{i=1}^{n-1} V_i}\right] \\
&\leq e^{-\lambda\epsilon} e^{\frac{\lambda^2}{8} \sum_{i=1}^n c_i^2}.
\end{aligned}$$

Taking $\lambda = \frac{4\epsilon}{\sum_{i=1}^n c_i^2}$ brings the statement. □

Remark 1.23. *Note that with $g(Z_1, \dots, Z_n) = \frac{1}{n}\sum_{i=1}^n Z_i$ we rediscover Hoeffding's inequality.*

1.2 Error bounds for finite classes

Let us first consider the finite hypothesis space $\mathcal{F} = \{f_1, \dots, f_N\}$.

Theorem 1.24 (Realizable case)**.** *Consider $\mathcal{Y} = \{0,1\}$ in the zero-noise case, i.e. $L(f^B) = 0$ with $f^B \in \mathcal{F}$ (as in example 1.4). As before we call such an f^B, since it is in $\mathcal{F}$, just f^*. Then with probability $\geq 1-\delta$ we have*

$$L(\hat{f}_n) - L(f^*) = L(\hat{f}_n) \leq \frac{\log N + \log\left(\frac{1}{\delta}\right)}{n}.$$

Proof. Let us try to bound $\mathbb{P}(\text{ERM selects } f \text{ with } L(f) > \epsilon)$. By definition we know that $L_n(\hat{f}_n) = 0$, since ERM selects only these predictors f for which $L_n(f) = 0$. Now take any $f \in \mathcal{F}$ s.t. $L(f) = \mathbb{P}(f(X) \neq Y) > \epsilon$, i.e. consider the functions that are expected to make some error. That implies that $\mathbb{P}(L(f) = 0) < 1 - \epsilon$ and therefore for n samples we get

$$\mathbb{P}(L_n(f) = 0) < (1-\epsilon)^n \leq e^{-n\epsilon},$$

where the last step follows from $1 + x \leq e^x$. Now we consider

$$\begin{aligned}
\mathbb{P}(\text{ERM selects } \hat{f}_n \text{ with } L(\hat{f}_n) > \epsilon) &\leq \mathbb{P}\left(\exists f \in \mathcal{F} \text{ s.t. } L(f) > \epsilon, L_n(f) = 0\right)\\
&= \mathbb{P}\left(\cup_{\substack{f\in\mathcal{F}\\ L(f)>\epsilon}} L_n(f) = 0\right)\\
&\leq \sum_{\substack{f\in\mathcal{F}\\ L(f)>\epsilon}} \mathbb{P}(L_n(f) = 0)\\
&\leq \sum_{\substack{f\in\mathcal{F}\\ L(f)>\epsilon}} e^{-n\epsilon}\\
&\leq N e^{-n\epsilon} =: \delta.
\end{aligned}$$

This gives $\epsilon = \frac{\log N + \log\left(\frac{1}{\delta}\right)}{n}$. So with probability at least $1-\delta$ we have the complement event, i.e.

$$\mathbb{P}(\text{ERM selects } f \text{ with } L(f) \leq \epsilon) \geq 1 - \delta.$$

The event $\{\text{ERM selects } f \text{ with } L(f) \leq \epsilon\}$ is just $\{L(\hat{f}_n) \leq \epsilon\}$ and therefore we get the desired bound. □

Remark 1.25. *One can show that given our assumptions this bound is actually the best possible one.*

Let us now drop the assumption $L(f^*) = 0$.

Theorem 1.26 (Agnostic case)**.** *We consider again $\mathcal{F} = \{f_1, \dots, f_N\}$. Let l be bounded. Then with probability $\geq 1-\delta$ we have*

$$L(\hat{f}_n) - L(f^*) \leq \sqrt{2\frac{\log(2N) + \log\left(\frac{1}{\delta}\right)}{n}}.$$

Proof.

$$\begin{aligned}
L(\hat{f}_n) - L(f^*) &= L(\hat{f}_n) - L_n(\hat{f}_n) + \underbrace{L_n(\hat{f}_n) - L_n(f^*)}_{\leq 0} + L_n(f^*) - L(f^*)\\
&\leq |L(\hat{f}_n) - L_n(\hat{f}_n)| + |L_n(f^*) - L(f^*)|\\
&\leq 2\sup_{f\in\mathcal{F}} |L(f) - L_n(f)|.
\end{aligned}$$

Since our $\mathcal{F}$ is finite, we can actually consider the maximum and compute

$$\begin{aligned}
\mathbb{P}\left(\max_{i=1,\dots,N} |L(f_i) - L_n(f_i)| > \epsilon\right) &= \mathbb{P}\left(\cup_{i=1}^N \{|L(f_i) - L_n(f_i)| > \epsilon\}\right)\\
&\leq \sum_{i=1}^N \mathbb{P}(|L(f_i) - L_n(f_i)| > \epsilon)\\
&\leq 2N e^{-2n\epsilon^2} =: \delta,
\end{aligned}$$

where the last step follows from Hoeffding's inequality, which also holds when taking functions of the random variables. If we solve for ϵ we get $\epsilon = \sqrt{\frac{\log \frac{2N}{\delta}}{2n}}$. Considering again the complement event (as in the proof of theorem 1.24) yields our desired statement. □

We have now seen the two extreme cases: The rather unrealistic realizible setting ($L(f^*) = 0$) and the totally agnostic setting ($L(f^*) = \frac{1}{2}$). The first one has convergence rate $\mathcal{O}(\frac{1}{n})$, the latter one $\mathcal{O}(\frac{1}{\sqrt{n}})$. Let us now look at cases in between those extremes ("from slow to fast rates").

Theorem 1.27. *We again consider $\mathcal{F} = \{f_1, \dots, f_N\}$ and our loss being bounded. With probability $\geq 1 - \delta$ we have*

$$L(\hat{f}_n) - L(f^*) \leq C \left(\sqrt{L(f^*) \frac{\log 4N + \log\left(\frac{1}{\delta}\right)}{n}} + \frac{\log\left(\frac{4N}{\delta}\right)}{n} \right),$$

where C is a constant.

Proof. First note that

$$\mathrm{Var}(\mathbb{1}_{\{f(X) \neq Y\}}) = \mathbb{E}[\mathbb{1}^2] - \mathbb{E}[\mathbb{1}]^2 \leq \mathbb{E}[\mathbb{1}] = L(f).$$

Now take Bernstein's inequality with $\delta \to \frac{\delta}{N}$:

$$\mathbb{P}\left(\exists f \in \mathcal{F} : \underbrace{|L_n(f) - L(f)| \geq \sqrt{2L(f)\frac{\log \frac{2N}{\delta}}{n}} + \frac{2\log\frac{2N}{\delta}}{3n}}_{=:E(f)} \right) = \mathcal{P}(\cup_{f \in \mathcal{F}} E(f)) \leq \sum_{f \in \mathcal{F}} \mathbb{P}(E) \leq \sum_{f \in \mathcal{F}} \frac{\delta}{N} = \delta.$$

Forming the complement yields

$$\mathbb{P}\left(\forall f \in \mathcal{F} : |L_n(f) - L(f)| \leq \sqrt{2L(f)\frac{\log \frac{2N}{\delta}}{n}} + \frac{2\log\frac{2N}{\delta}}{3n} \right) \geq 1 - \delta.$$

We can therefore pick $f = \hat{f}_n$ and get that with probability at least $1 - \delta$

$$|L_n(\hat{f}_n) - L(\hat{f}_n)| \leq \sqrt{2L(\hat{f}_n)\frac{\log \frac{2N}{\delta}}{n}} + \frac{2\log\frac{2N}{\delta}}{3n}.$$

Using another time Bernstein's inequality for f^* yields

$$L_n(\hat{f}_n) \leq L_n(f^*) \leq L(f^*) + \sqrt{2L(f^*)\frac{\log \frac{2}{\delta}}{n}} + \frac{2\log\frac{2}{\delta}}{3n}$$

We can now combine the preceding two equations to get

$$\frac{L(\hat{f}_n) - L(f^*)}{\sqrt{L(\hat{f}_n)}} \leq \sqrt{2\frac{\log\frac{4N}{\delta}}{n}}$$

and use the fact $\frac{A-B}{\sqrt{A}} \leq C \Rightarrow A \leq B + C^2 + \sqrt{B}C$ for $A, B, C \geq 0$ to get the result.

□

Remark 1.28. *This bound is again tight. Note that if we assume $n < \frac{1}{L(f^*)} =: \gamma$ we are in a fast-rates-regime, i.e. $\mathcal{O}(\frac{1}{n})$.*

Remark 1.29. *Note that we often do not use ERM in practice, as the minimization can be NP-hard.*

1.3 Error bounds for infinite classes

We now want to consider function classes $\mathcal{F}$ that have infinitely many elements. The former bounds will then be meaningless (as we have $N = \infty$). When looking into the proofs, we realize that our union bounds as in theorem 1.26 will yield an infinite sum that has to be bounded in order to be usefull. If the function class is even uncountable we need a completely different approach. For that, we make the following definition.

Definition 1.30 (Rademacher complexity). *Consider a function class $\mathcal{F}$ and i.i.d. random variables $Z_1, \dots, Z_n \sim \mathbb{P}$. Let $\sigma = \{\sigma_1, \dots, \sigma_n\}$ be a set of i.i.d. random variables s.t. $\mathbb{P}(\sigma_i = \pm 1) = \frac{1}{2}$. We then define the empirical Rademacher complexity*[6] *as*

$$\hat{R}_n(\mathcal{F}, S_n) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f(Z_i) \right| \middle| S_n \right].$$

For the Rademacher complexity we average over all samples $S_n = (Z_1, \dots, Z_n)$, i.e.

$$R_n(\mathcal{F}) = \mathbb{E}_{S_n} \left[\hat{R}_n(\mathcal{F}) \right].$$

Remark 1.31. *The supremum can be interpreted as finding the function f out of $\mathcal{F}$ that makes the samples correlate the most with some random values σ. Loosely speaking, the idea is to find a function that looks like random noise the most. The Rademacher complexity then gives a measure of how well we can find a function out of the function class $\mathcal{F}$ that behaves like random noise. Note that $\hat{R}_n(\mathcal{F}) = 0$ if $\mathcal{F}$ consists of only one function and $\hat{R}_n(\mathcal{F}) = 1$ if $\mathcal{F}$ consists of all functions. In fact, we have $\hat{R}_n(\mathcal{F}) \in [0, 1]$. Note that in general computing Rademacher complexities can be difficult (actually as difficult as computing empirical risk minimizers).*

Remark 1.32. *For $\mathcal{F} \subset \mathcal{G}$ we have $R_n(\mathcal{F}) \leq R_n(\mathcal{G})$.*

Theorem 1.33. *Consider any prediction task $(\mathbb{P}, \mathcal{X}, \mathcal{Y}, l, \mathcal{F})$ s.t. $l(f(X), Y) \in [0, 1]$. Then with probability $\geq 1 - \delta$*

$$\sup_{f \in \mathcal{F}} |L(f) - L_n(f)| \leq 2 R_n(l \circ \mathcal{F}) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}},$$

where $l \circ \mathcal{F} = \{l(f(X), Y) : f \in \mathcal{F}, (X, Y) \in \mathcal{X} \times \mathcal{Y}\}$.

Proof. Comparing to theorem 1.22, let $g(Z_1, \dots, Z_n) = \sup_{f \in \mathcal{F}} |L_n(f) - L(f)|$. If we change one datum Z_i to Z_i' we get

$$\begin{aligned} \left| \sup_{f \in \mathcal{F}} |L_n(f) - L(f)| - \sup_{f \in \mathcal{F}} \left| L_n^i(f) - L(f) \right| \right| &\leq \left| \frac{1}{n} \sup_{f \in \mathcal{F}} |l(f(X_i), Y_i) - l(f(X_i'), Y_i')| \right| \\ &\leq \frac{1}{n}, \end{aligned}$$

where L_n^i is the empirical risk with the changed datum. We can therefore use McDiarmid's inequality

$$\sup_{f \in \mathcal{F}} |L_n(f) - L(f)| \leq \mathbb{E} \left[\sup_{f \in \mathcal{F}} |L_n(f) - L(f)| \right] + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}.$$

The idea now is to introduce a ghost sample S_n' that is identically distributed as S_n. We then have

$$\begin{aligned} \mathbb{E}_{S_n} \left[\sup_{f \in \mathcal{F}} |L_n(f) - L(f)| \right] &= \mathbb{E}_{S_n} \left[\sup_{f \in \mathcal{F}} \left| \mathbb{E}_{S_n'} \left[L_n(f) - L_n'(f) \right] \right| \right] \\ &\leq \mathbb{E}_{S_n, S_n'} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \left(l(f(X_i), Y_i) - l(f(X_i'), Y_i') \right) \right| \right] \\ &= \mathbb{E}_{S_n, S_n', \sigma} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i \left(l(f(X_i), Y_i) - l(f(X_i'), Y_i') \right) \right| \right] \\ &\leq \mathbb{E}_{S_n, \sigma} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i l(f(X_i), Y_i) \right| \right] + \mathbb{E}_{S_n', \sigma} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i l(f(X_i'), Y_i') \right| \right] \\ &= 2 R_n(l \circ \mathcal{F}), \end{aligned}$$

[6] In the context of empirical processes $R_n(f) = \frac{1}{n} \sum_{i=1}^n \sigma_i f(Z_i)$ is sometimes called Rademacher process.

where we used Jensen's inequality in the second step. □

Remark 1.34. *Note by looking at the proof of theorem 1.26 that we actually have*

$$L(\hat{f}_n) - L(f^*) \leq \sup_{f\in\mathcal{F}}(L(f) - L_n(f)) + \sup_{f\in\mathcal{F}}(L_n(f) - L(f))$$

and can therefore bound the estimation error.

Let us now investigate how to control the Rademacher complexity. First, we rediscover the finite class case.

Lemma 1.35. *Assume $\mathcal{F} = \{f_1, \dots, f_N\}$. Then*

$$\hat{R}_n(\mathcal{F}) \leq \frac{\sqrt{2\log N}}{n} \max_{j=1,\dots,N} \sqrt{\sum_{i=1}^n f_j^2(Z_i)}$$

and therefore

$$L(\hat{f}_n) - L(f^*) \leq C\left(\sqrt{\frac{\log N}{n}} + \sqrt{\frac{\log\frac{1}{\delta}}{n}}\right).$$

Proof. We have for any $\lambda > 0$

$$\begin{aligned}
\exp\left(\lambda\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \sigma_i f_j(Z_i)\right]\right) &\leq \mathbb{E}\left[\exp\left(\lambda \max_{j=1,\dots,N}\frac{1}{n}\sum_{i=1}^n \sigma_i f_j(Z_i)\right)\right] \\
&= \mathbb{E}\left[\max_{j=1,\dots,N}\exp\left(\frac{\lambda}{n}\sum_{i=1}^n \sigma_i f_j(Z_i)\right)\right] \\
&\leq \sum_{j=1}^N \prod_{i=1}^n \mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sigma_i f_j(Z_i)\right)\right] \\
&\leq \sum_{j=1}^N \prod_{i=1}^n \exp\left(\frac{\lambda^2}{2n^2} f_j^2(Z_i)\right) \\
&= \sum_{j=1}^N \exp\left(\frac{\lambda^2}{2n^2}\sum_{i=1}^n f_j^2(Z_i)\right) \\
&\leq N \max_{j=1,\dots,N}\exp\left(\frac{\lambda^2}{2n^2}\sum_{i=1}^n f_j^2(Z_i)\right),
\end{aligned}$$

where we used Hoeffding's lemma. We get

$$\hat{R}_n(\mathcal{F}) \leq \frac{1}{\lambda}\left(\log N + \max_{j=1,\dots,N}\frac{\lambda^2}{2n^2}\sum_{i=1}^n f_j^2(Z_i)\right).$$

Minimizing w.r.t. λ brings $\lambda^* = \sqrt{\frac{\log N 2n^2}{\max_{j=1,\dots,N}\sum_{i=1}^n f_j^2(Z_i)}}$ and therefore the result. □

Remark 1.36. *Sometimes the Rademacher complexity is defined without the absolute values. Note, however, that we have $\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i f(Z_i)\right|\right] = \mathbb{E}\left[\sup_{f\in\mathcal{F}\cup-\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \sigma_i f(Z_i)\right]$.*

Let us now consider the infinite class case, i.e. a general $\mathcal{F}$ for binary classification with $l(y', y) = \mathbb{1}\{y' \neq y\}$. From theorem 1.33 we know that

$$L(\hat{f}_n) - L(f^*) \leq 4R_n(l\circ\mathcal{F}) + \sqrt{\frac{2\log\frac{1}{\delta}}{n}}.$$

Still, in R_n we take the supremum over all $f \in \mathcal{F}$, thus we need to find a strategy how to deal with that in the infinite class case. The idea is to argue that we can divide $\mathcal{F}$ into finitely many equivalence classes. Note that in the computation of R_n we consider the sum over n summands, where each of them has the value 0 or 1. We can say that each function f induces an n-dimensional vector with the losses of the sample,

$$f \mapsto (l(f(X_1), Y_1), \dots, l(f(X_n), Y_n)) \in \{0,1\}^n,$$

and we realize that we can have at most 2^n such vectors, i.e. no matter how big the function class is, we have to consider at most 2^n cases. Therefore, we can divide $\mathcal{F}$ into 2^n equivalence classes and identify functions that lead to the same loss vector with one another. Unfortunately, however, this is not enough, as plugging in 2^n instead of N in lemma 1.35 just yields a constant bound that is not decreasing with n. The next idea is to consider only function classes $\mathcal{F}$ that yield (much) less than 2^n possible permutations. Let us therefore define the following set of vectors, the projections of the predictors on the sample,

$$\mathcal{F}_{S_n} := \{ (l(f(X_1), Y_1), \dots, l(f(X_n), Y_n)), f \in \mathcal{F} \} \subset \{0,1\}^n,$$

and ask what the cardinatlity of the vector over all possible data sets S_n is. For this we define the so called growth function (or shattering number)

$$S_{\mathcal{F}}(n) = \sup_{S_n} |\mathcal{F}_{S_n}| \leq 2^n,$$

which is the worst case cardinality of our equivalence classes. The hope is that the growth function will be indeed much smaller than 2^n. Note that $S_{\mathcal{F}}(n)$ does not depend on the sample distribution $\mathbb{P}$, but only on $\mathcal{F}$.

Now we introduce a crucial trick.

Lemma 1.37 (Symmetrization). *For all $\epsilon \geq \sqrt{\frac{2}{n}}$*

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} |L_n(f) - L(f)| > \epsilon \right) \leq 2\mathbb{P}\left(\sup_{f \in \mathcal{F}} |L_n(f) - L_n'(f)| > \frac{\epsilon}{2} \right),$$

where L_n' is the empirical error of a ghost sample.

Proof. We again consider a ghost sample $S_n' = (Z_1', \dots, Z_n')$ that is distributed as the original sample S_n and denote its empirical error as L_n'. Let $f_n \in \mathcal{F}$ maximize $|L_n(f) - L(f)|$. Note that if $|L_n(f_n) - L(f_n)| > \epsilon$ and $|L(f_n) - L_n'(f_n)| \leq \frac{\epsilon}{2}$, then

$$\epsilon < |L_n(f_n) - L(f_n)| = |L_n(f_n) - L_n'(f_n) + L_n'(f_n) - L(f_n)| \leq |L_n(f_n) - L_n'(f_n)| + |L_n'(f_n) - L(f_n)| \leq \frac{\epsilon}{2} + |L_n'(f_n) - L(f_n)|$$

and therefore $|L_n'(f_n) - L(f_n)| > \frac{\epsilon}{2}$. This then yields

$$\begin{aligned} \mathbb{1}\Big\{|L_n(f_n) - L(f_n)| > \epsilon\Big\} \mathbb{1}\left\{|L(f_n) - L'(f_n)| \leq \frac{\epsilon}{2}\right\} &= \mathbb{1}\left\{|L_n(f_n) - L(f_n)| > \epsilon, |L(f_n) - L'(f_n)| \leq \frac{\epsilon}{2}\right\} \\ &\leq \mathbb{1}\left\{|L(f_n) - L_n'(f_n)| > \frac{\epsilon}{2}\right\}. \end{aligned}$$

Taking the expectation w.r.t. S_n' brings

$$\mathbb{1}\Big\{|L_n(f_n) - L(f_n)| > \epsilon\Big\} \mathbb{P}'\left(|L(f_n) - L'(f_n)| \leq \frac{\epsilon}{2}\right) \leq \mathbb{P}'\left(|L_n(f_n) - L_n'(f_n)| > \frac{\epsilon}{2}\right).$$

Note that by Chebyshev's inequality and with $X \in [0,1] \Rightarrow \operatorname{Var}(X) \leq \frac{1}{4}$ we get

$$\mathbb{P}'\left(|L(f_n) - L'(f_n)| \leq \frac{\epsilon}{2}\right) \leq \frac{4\operatorname{Var}'(f_n)}{n\epsilon^2} \leq \frac{1}{n\epsilon^2} \leq \frac{1}{2},$$

which then yields

$$\mathbb{1}\Big\{\sup_{f \in \mathcal{F}} |L_n(f) - L(f)| > \epsilon\Big\} \leq 2\mathbb{P}'\left(\sup_{f \in \mathcal{F}} |L_n(f) - L_n'(f)| > \frac{\epsilon}{2} \right).$$

Taking expectation w.r.t. S_n brings the desired result. □

Remark 1.38. *Note that $L_n(f) - L(f)$ can take any real value, whereas $L_n(f) - L_n'(f)$ can only take finitely many values, i.e. we can hope to apply the union bound in the uncountable case.*

Theorem 1.39 (Vapnik-Chervonenkis bound)**.** *We again consider $\mathcal{Y} = \{0,1\}$. Then with probability $\geq 1-\delta$*

$$L(\hat{f}_n) - L(f^*) \leq 2\sqrt{\frac{4}{n}\log\left(\frac{4S_{\mathcal{F}}(2n)}{\delta}\right)}.$$

Proof. With the preceding lemma we have

$$\begin{aligned}
\mathbb{P}\left(\sup_{f\in\mathcal{F}} |L_n(f) - L(f)| \geq \epsilon\right) &\leq 2\mathbb{P}\left(\sup_{f\in\mathcal{F}} |L_n(f) - L'_n(f)| > \frac{\epsilon}{2}\right) \\
&= 2\mathbb{P}\left(\max_{l\in\mathcal{F}_{S_n,S'_n}} |\tilde{L}_n(l) - \tilde{L}'_n(l)| > \frac{\epsilon}{2}\right) \\
&\leq 2\sum_{l\in\mathcal{F}_{S_n,S'_n}} \mathbb{P}\left(|\tilde{L}_n(l) - \tilde{L}'_n(f)| > \frac{\epsilon}{2}\right) \\
&\leq 2\sum_{l\in\mathcal{F}_{S_n,S'_n}} 2e^{-\frac{n}{4}\epsilon^2} \\
&\leq 4S_{\mathcal{F}}(2n)e^{-\frac{n}{4}\epsilon^2},
\end{aligned}$$

where with some abuse of notation $\tilde{L}_n(l) - \tilde{L}'_n(l) = \frac{1}{n}\left(\sum_{i=1}^n l_i - \sum_{i=n+1}^{2n} l_i\right)$ with $l_i = l(f(X_i), Y_i)$, for which we used Hoeffding's inequality. □

Still we have to answer the question when this bound goes to zero for $n \to \infty$, or in other words, what the geometry of $\mathcal{F}$ has to be s.t. $S_{\mathcal{F}}(n)$ does not grow too fast. The crucial tool that shall answer these questions is the following.

Definition 1.40 (VC dimension)**.** *The VC dimension of a function class $\mathcal{F}$ is defined as*

$$VC(\mathcal{F}) = \sup\{n : S_{\mathcal{F}}(n) = 2^n\}.$$

If such an n does not exist we set $\mathrm{VC}(\mathcal{F}) = \infty$.

Remark 1.41. $\mathrm{VC}(\mathcal{F}) = d$ *implies that there exists a sample $X_1, \dots, X_d \in \mathcal{X}$ s.t. for any $Y_1, \dots, Y_d \in \mathcal{Y}$ there exists an $f \in \mathcal{F}$ with $f(X_i) = Y_i$. We say that the sample is "shattered".*

Remark 1.42. *The VC dimension can be understood as measuring the effective size of a function class. Note that it is distribution independent. One can consider a concept called VC entropy to include properties of $\mathbb{P}$ (see for instance chapter 4.3 in [6]).*

Class $\mathcal{F}$	$\mathrm{VC}(\mathcal{F})$
$\{f_1, \dots, f_N\}$	$\leq \log_2 N$
Intervals on the real line	2
Discs in $\mathbb{R}^2$	3
Rectangles in $\mathbb{R}^d$	$2d$
Half-spaces in $\mathbb{R}^d$	$d+1$
Convex polygons in $\mathbb{R}^2$	∞
$\{\mathrm{sgn}(\sin(\alpha x)) : \alpha \in \mathbb{R}\}$	∞

Figure 1: Examples of VC dimensions.

Example 1.43 (VC dimensions)**.** *The last example in figure 1 is interesting as it shows that the VC dimension is not necessarily linked to the number of free parameters. To show this take for instance $X_i = 10^{-i}$ and $\alpha =$*

$\pi\left(1+\sum_{i=1}^{n}\frac{1}{2}(1-Y_i)10^i\right)$. *Then*

$$\begin{aligned}f(X_j) &= \operatorname{sgn}\left(\sin\left(10^{-j}\pi+\sum_{i=1}^{n}\frac{\pi}{2}(1-Y_i)10^{i-j}\right)\right)\\ &= \operatorname{sgn}\left(\sin\left(10^{-j}\pi+\sum_{i\le j, Y_i=-1}\frac{\pi}{2}(1-Y_i)10^{i-j}\right)\right)\\ &= \operatorname{sgn}\left(\sin\left(\frac{\pi}{2}(1-Y_j)+\pi\left(\underbrace{10^{-j}+\sum_{i<j,Y_i=-1}10^{i-j}}_{<1}\right)\right)\right),\end{aligned}$$

after the properties of the geometric series. Now, if $Y_j=1$ the sine function takes values between 0 and π and if $Y_j=-1$ it takes values between π and 2π – therefore the classification is correct.

The VC dimension brings the following theorem.

Theorem 1.44 (Sauer, Shelah, Perles, Vapnik, Chervonenkis)**.** *Assume* $\mathrm{VC}(\mathcal{F})=d<\infty$. *Then for all* n

$$S_{\mathcal{F}}(n)\le\sum_{i=0}^{d}\binom{n}{i}$$

and for all $n\ge d$ we get

$$S_{\mathcal{F}}(n)\le\left(\frac{en}{d}\right)^d.$$

Proof. For $n=d=1$ we have

$$S_{\mathcal{F}}(1)\le\binom{1}{0}+\binom{1}{1}=2,$$

which clearly holds. By induction assume that the statement holds for $n-1$ and $d-1$ as well as for $n-1$ and d. We want to show that it also holds for n and d. Let $S_1=(Z_1,\dots,Z_n)$ and $S_2=(Z_2,\dots,Z_n)$, let $\mathcal{F}_{S_1}$ and $\mathcal{F}_{S_2}$ be the corresponding projections onto $\mathcal{F}$. For $f,g\in\mathcal{F}$ write

$$f\sim g \qquad \text{if} \qquad f(Z_1)=1-g(Z_1) \quad \text{and} \quad f(Z_j)=g(Z_j) \qquad \text{for } j=2,\dots,n.$$

and let

$$\mathcal{G}=\{f\in\mathcal{F}:\exists g\in\mathcal{F} \text{ s.t. } f\sim g\}.$$

Now consider $\mathcal{G}_{S_2}$, then

$$|\mathcal{F}_{S_1}|=|\mathcal{F}_{S_2}|+|\mathcal{G}_{S_2}|.$$

Note that $\mathrm{VC}(\mathcal{F}_{S_2})\le d$ and $\mathrm{VC}(\mathcal{G}_{S_2})\le d-1$. The latter follows since, if $\mathcal{G}_{S_2}$ shatters a set, we can add Z_1 to create a set that is shattered by $\mathcal{F}_{S_1}$. Therefore we have by assumption that

$$\begin{aligned}|\mathcal{F}_{S_1}| &\le \sum_{i=0}^{d}\binom{n-1}{i}+\sum_{i=0}^{d-1}\binom{n-1}{i}\\ &= \sum_{i=0}^{d}\left(\binom{n-1}{i}+\binom{n-1}{i-1}\right)\\ &= \sum_{i=0}^{d}\binom{n}{i}.\end{aligned}$$

To prove the second statement we note that for $n \geq d$

$$\begin{aligned}\sum_{i=0}^{d}\binom{n}{i} &\leq \left(\frac{n}{d}\right)^d \sum_{i=0}^{d}\binom{n}{i}\left(\frac{d}{n}\right)^i \\ &\leq \left(\frac{n}{d}\right)^d \sum_{i=0}^{n}\binom{n}{i}\left(\frac{d}{n}\right)^i \\ &= \left(\frac{n}{d}\right)^d \left(1+\frac{d}{n}\right)^n \\ &\leq \left(\frac{n}{d}\right)^d e^d.\end{aligned}$$

□

In analogy to the Vapnik-Chervonenkis bound one can also bound the Rademacher complexity. For this we need

Lemma 1.45. *Let $X_1, \dots, X_n$ be random variables with $\mathbb{E}\left[e^{\lambda X_i}\right] \leq e^{\frac{\lambda^2 \xi^2}{2}}$ for all $\lambda > 0$ and $\xi > 0$. Then*

$$\mathbb{E}\left[\max_{1\leq i\leq n} X_i\right] \leq \xi\sqrt{2\log n}.$$

Proof. With Jensen's inequality we have

$$\begin{aligned}\exp\left(\lambda\mathbb{E}\left[\max_{1\leq i\leq n} X_i\right]\right) &\leq \mathbb{E}\left[\exp\left(\lambda\max_{1\leq i\leq n} X_i\right)\right] \\ &= \mathbb{E}\left[\max_{1\leq i\leq n}\exp\left(\lambda X_i\right)\right] \\ &\leq \sum_{i=1}^{n}\mathbb{E}\left[e^{\lambda X_i}\right] \\ &\leq ne^{\frac{\lambda^2\xi^2}{2}}.\end{aligned}$$

Therefore,

$$\mathbb{E}\left[\max_{1\leq i\leq n} X_i\right] \leq \frac{\log n}{\lambda} + \frac{\lambda\xi^2}{2}$$

and minimizing w.r.t. λ gives the desired bound. □

Theorem 1.46. *Let $\mathcal{F}$ be a set of binary functions. Then*

$$R_n(\mathcal{F}) \leq \sqrt{\frac{2\log S_{\mathcal{F}}(n)}{n}}.$$

Proof. We have

$$\begin{aligned}R_n(\mathcal{F}) &= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(Z_i)\right|\right] \\ &= \mathbb{E}_{S_n}\left[\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(Z_i)\right|\middle| S_n\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\max_{v_i\in\mathcal{F}_{S_n}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i v_i\right|\middle| S_n\right]\right].\end{aligned}$$

Note that $\frac{\sigma_i v_i}{n}$ has mean zero and $-\frac{1}{n} \leq \frac{\sigma_i v_i}{n} \leq \frac{1}{n}$. Therefore, with lemma 1.16 we have $\mathbb{E}\left[e^{\lambda\sigma_i v_i}\right] \leq e^{\frac{\lambda^2}{2n^2}}$ and $\mathbb{E}\left[e^{\lambda\sum_{i=1}^{n}\frac{\sigma_i v_i}{n}}\right] \leq e^{\frac{\lambda^2}{2n}}$ and thus with the preceding lemma 1.45 we get

$$\mathbb{E}\left[\max_{v\in\mathcal{F}_{S_n}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i v_i\right|\middle| S_n\right] \leq \sqrt{\frac{2\log|\mathcal{F}_{S_n}|}{n}} \leq \sqrt{\frac{2\log S_{\mathcal{F}}(n)}{n}}.$$

□

Remark 1.47. *For $n \geq d$ the growth function $S_{\mathcal{F}}(n)$ only grows polynomially and therefore gives a helpful bound when plugging it in to theorem 1.39, namely $\leq C\sqrt{\frac{d\log n}{n}}$. In other words, a finite VC dimension brings learnability. In fact, using techniques called "chaining" and "covering numbers" one can even get a bound $\leq C\sqrt{\frac{d}{n}}$ – details can for instance be found in [12], chapter 4.3.*

In analogy to theorem 1.27 one can show

Theorem 1.48. *With probability at least $1-\delta$ we have*

$$L(\hat{f}_n) - L(f^*) \leq 2\sqrt{L(f^*)\frac{\log S_{\mathcal{F}}(2n) + \log\frac{4}{\delta}}{n}} + \frac{\log S_{\mathcal{F}}(2n) + \log\frac{4}{\delta}}{n}.$$

Proof. See chapter 6.2 in [6]. □

Lemma 1.49 (Contraction principle). *Assume $\varphi : \mathbb{R} \to \mathbb{R}, \varphi(0) = 0$ and φ is L_φ-Lipschitz. Then for all function classes $\mathcal{F}$*

$$R_n(\varphi \circ \mathcal{F}) \leq L_\varphi R_n(\mathcal{F}).$$

Proof. Let $g_i(z)$ and $h_i(z)$ be functions s.t. for any z, z'

$$|g_i(z) - g_i(z')| \leq |h_i(z) - h_i(z')|.$$

Let us show

$$\mathbb{E}_{\sigma_1,\ldots,\sigma_n,x}\left[\sup_z \left\{c(x,z) + \sum_{i=1}^n \sigma_i g_i(z)\right\}\right] \leq \mathbb{E}_{\sigma_1,\ldots,\sigma_n,x}\left[\sup_z \left\{c(x,z) + \sum_{i=1}^n \sigma_i h_i(z)\right\}\right]$$

by induction. Obviously it is true for $n = 0$. Let us assume it holds for n, then

$$\begin{aligned}
\mathbb{E}_{\sigma_1,\ldots,\sigma_{n+1},x}\left[\sup_z \left\{c(x,z) + \sum_{i=1}^{n+1} \sigma_i g_i(z)\right\}\right] &= \mathbb{E}_{\sigma_1,\ldots,\sigma_n,x}\left[\sup_{z,z'} \left\{\frac{c(x,z)+c(x,z')}{2} + \sum_{i=1}^n \sigma_i \frac{g_i(z)-g_i(z')}{2} + \frac{g_{n+1}(z)-g_{n+1}(z')}{2}\right\}\right] \\
&\leq \mathbb{E}_{\sigma_1,\ldots,\sigma_n,x}\left[\sup_{z,z'} \left\{\frac{c(x,z)+c(x,z')}{2} + \sum_{i=1}^n \sigma_i \frac{g_i(z)-g_i(z')}{2} + \frac{h_{n+1}(z)-h_{n+1}(z')}{2}\right\}\right] \\
&= \mathbb{E}_{\sigma_1,\ldots,\sigma_{n+1},x}\left[\sup_y \left\{c(x,z) + \sum_{i=1}^n \sigma_i g_i(y) + \sigma_{n+1} h_{n+1}(z)\right\}\right] \\
&\leq \mathbb{E}_{\sigma_1,\ldots,\sigma_{n+1},x}\left[\sup_z \left\{c(x,z) + \sum_{i=1}^{n+1} \sigma_i h_i(z)\right\}\right].
\end{aligned}$$

One can show the same also with absolute values. Now take $c(x,z) = 0$, $f(z) = \varphi(f(x)y)$ and $h(z) = L_\varphi f(x)y$ to get the result. □

Remark 1.50. *This lemma brings that Rademacher complexity is bounded whenever the loss is Lipschitz and therefore also brings learning statements for the regression setting. To give some examples, hinge loss given by $l(y,y') = \max\{1-yy', 0\}$ is 1-Lipschitz, logistic loss given by $l(y,y') = \log(1+e^{-yy'})$ is 1-Lipschitz, squared loss $l(y,y') = (y'-y)^2$ is $4B$-Lipschtiz when $|y|, |y'| \leq B$, absolute loss $l(y,y') = |y-y'|$ is 1-Lipschitz.*

2 Kernels

Sometimes working with the data $X_1, \ldots, X_n$ in the space $\mathcal{X}$ does not yield reasonable results. Therefore we consider mapping them into a higher dimensional space, do computations there and map the results back to the original space. This can be understood as preprocessing the data via a function $\Phi : \mathcal{X} \to \mathcal{H}$, a so called feature mapping, where usually $\mathcal{H}$ is a Hilbert space and $\dim \mathcal{X} \ll \dim \mathcal{H}$. One way of seeing this is to incorporate nonlinearities with a nonlinear Φ. In fact, these methods will provide an elegant approach to achieve nonlinear algorithms from linear ones. One can also look at it from a different point of view, relating to the previous chapter: We want to find a function class $\mathcal{H}$ that allows statistical generalization and computational efficiency. Note that in practice there are many ways to choose Φ. The general credo is: In high dimensional feature spaces almost everything looks linear and is therefore hopefully linearly separable.

Although we map to the usually infinite dimensional space $\mathcal{H}$ we will see that we can still do computations in the original space $\mathcal{X}$; so called "kernels" will help us with that.

Definition 2.1. *A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called symmetric positive semi-definite (s.p.s.d.) kernel if*

(i) k is symmetric: $k(X, X') = k(X', X)$ for all $X, X' \in \mathcal{X}$,

(ii) k is positive semi-definite: $\sum_{i,j=1}^{n} a_i a_j k(X_i, X_j) \geq 0$ for any $X_1, \dots, X_n \in \mathcal{X}, a_1, \dots, a_n \in \mathbb{R}$.

Remark 2.2. *Note that k is positive semi-definite if the Gram matrix $K \in \mathbb{R}^{n \times n}$ defined by $K_{ij} = k(X_i, X_j)$ is positive semi-definite, which then implies the eigenvalues being positive. Note also that $n = 1$ implies $k(X, X') \geq 0$ and $n = 2$ implies $k(X, X')^2 \leq k(X, X)k(X', X')$ in analogy to the Cauchy-Schwarz inequality (i.e. two properties that we also have with the usual scalar product).*

The trick is to never compute scalar products in $\mathcal{H}$ directly, but rather compute them via a kernel k, i.e. we never evaluate Φ directly as long as we can evaluate the corresponding scalar products.

Lemma 2.3. *Let $\mathcal{H}$ be a Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. For any map $\Phi : \mathcal{X} \to \mathcal{H}$ there is a s.p.s.d. kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ s.t. k is reproducing the inner product in $\mathcal{H}$, namely*

$$k(X, X') = \langle \Phi(X), \Phi(X') \rangle_{\mathcal{H}} \qquad \forall\, X, X' \in \mathcal{X}.$$

Proof. k is symmetric since $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is symmetric. Since $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is bilinear and positive definite we have

$$\sum_{i,j=1}^{n} a_i a_j k(X_i, X_j) = \sum_{i,j=1}^{n} a_i a_j \langle \Phi(X_i), \Phi(X_j) \rangle_{\mathcal{H}} = \left\langle \sum_{i=1}^{n} a_i \Phi(X_i), \sum_{j=1}^{n} a_j \Phi(X_j) \right\rangle_{\mathcal{H}} \geq 0.$$

□

Remark 2.4. *A complete characterization for the positive definiteness of a function is provided by Bochner's theorem. A continuous function $f : \mathbb{R}^d \to \mathbb{R}$ is positive definite if and only if it is the Fourier transform of a finite non-negative Borel measure $\mathbb{P}$, i.e. $f(X) = \int_{\mathbb{R}^d} e^{i\langle X, \omega \rangle} \mathrm{d}\mathbb{P}(\omega)$.*

Interestingly, the converse of the previous lemma 2.3 is also true:

Theorem 2.5 (Aronszajn, Moore)**.** *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a s.p.s.d. kernel. Then there exists a unique Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ of functions on $\mathcal{X}$ and a map $\Phi : \mathcal{X} \to \mathcal{H}$ s.t.*

$$k(X, X') = \langle \Phi(X), \Phi(X') \rangle_{\mathcal{H}} \qquad \forall\, X, X' \in \mathcal{X}.$$

Proof. Define the linear map $\Phi : \mathcal{X} \to \mathcal{H}$ by $\Phi(X) = k(X, \cdot)$. Let $\mathcal{H}_0 = \operatorname{span}\{\Phi(X) : X \in \mathcal{X}\}$ and define an inner product on $\mathcal{H}_0$ by $\left\langle \sum_{i=1}^{m} a_i \Phi(X_i), \sum_{j=1}^{l} a_j \Phi(X'_j) \right\rangle = \sum_{i=1}^{m} \sum_{j=1}^{l} a_i b_j k(X_i, X'_j)$. Now let $\mathcal{H}$ be the completion of $\mathcal{H}_0$ w.r.t. the inner product, so $\mathcal{H}$ contains all Cauchy sequences of the type $f(X) = \sum_{i=1}^{\infty} a_i k(X_i, X)$. We can check that

$$\langle f, \Phi(x) \rangle = \left\langle \sum_{i=1}^{\infty} a_i \Phi(X_i), \Phi(X) \right\rangle_{\mathcal{H}} = \sum_{i=1}^{\infty} a_i k(X_i, X) = f(x).$$

In particular, for $f = \Phi(X')$ we get $\langle \Phi(X'), \Phi(X) \rangle_{\mathcal{H}} = k(X, X')$. To show uniqueness, let $\mathcal{G}$ be another Hilbert space for which k is reproducing, that is

$$\langle \Phi(X'), \Phi(X) \rangle_{\mathcal{H}} = k(X, X') = \langle \Phi(X'), \Phi(X) \rangle_{\mathcal{G}}.$$

By linearity we have $\langle \cdot, \cdot \rangle_{\mathcal{H}} = \langle \cdot, \cdot \rangle_{\mathcal{G}}$ on $\mathcal{H}_0$, i.e. $\mathcal{H}_0 \subset \mathcal{G}$ and since $\mathcal{G}$ is complete, $\mathcal{H} \subset \mathcal{G}$. Let now $f \in \mathcal{G}$ with $f = f_{\mathcal{H}} + f_{\mathcal{H}^{\perp}}$, where $f_{\mathcal{H}} \in \mathcal{H}$ and $f_{\mathcal{H}^{\perp}} \in \mathcal{H}^{\perp}$. We then have

$$f(X) = \langle f, \Phi(X) \rangle_{\mathcal{H}} = \langle f_{\mathcal{H}}, \Phi(X) \rangle_{\mathcal{H}} = f_{\mathcal{H}}(X)$$

and thus $f_{\mathcal{H}^{\perp}} = 0$ and $\mathcal{H} = \mathcal{G}$. □

Remark 2.6. *The feature map Φ is a mapping from $\mathcal{X}$ to the functions $f : \mathcal{X} \to \mathbb{R}$ defined by $X \mapsto k(\cdot, X)$, where one can think of the kernel as a nonlinear similarity measure; it can be interpreted as a similarity of X to all other points in $\mathcal{X}$. Note the "reproducing" property $\langle k(\cdot, X), f(\cdot)\rangle_{\mathcal{H}} = f(X)$, i.e. evaluation at a point X, and that f is linear in the Hilbert space. We can see the Hilbert space as the linear span of kernel functions: $\mathcal{H} = \operatorname{span}\{k(\cdot, X) : X \in \mathcal{X}\}$.*

Definition 2.7. *A reproducing kernel Hilbert space (RKHS) is a Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$ with a reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, i.e. the span of $\{k(\cdot, X) : X \in \mathcal{X}\}$ is dense in $\mathcal{H}$, and $k(X, \cdot) \in \mathcal{H}$ is the point evaluation function for $\mathcal{H}$, i.e. $f(X) = \langle k(X, \cdot), f\rangle_{\mathcal{H}}$.*

Remark 2.8. *Note that not all Hilbert spaces have a reproducing kernel. For instance $L^2 = \{f(x) : \int_0^1 f(x)^2 \mathrm{d}x < \infty\}$ with $\langle f, g\rangle_{L^2} = \int_0^1 f(x)g(x)\mathrm{d}x$ does not contain its point evaluation function δ, since*

$$f(x) = \int\limits_0^1 f(x)\delta(x)\mathrm{d}x, \qquad \int\limits_0^1 \delta^2(x)dx = \infty.$$

Example 2.9. *Some common kernels are*

- *linear kernel: $k(X, X') = X^\top X'$,*
- *polynomial kernel: $k(X, X') = (X^\top X' + c)^m$,*
- *Gaussian kernel: $k(X, X') = \exp\left(-\frac{1}{\sigma}\|X - X'\|^2\right)$.*

Example 2.10. *Consider the polynomial kernel of degree 2. Then we have*

$$k(X, X')_2 = (X^\top X')^2 = (X_1X_1' + X_2X_2')^2 = (X_1^2, \sqrt{2}X_1X_2, X_2^2)^\top (X'^2_1, \sqrt{2}X'_1X'_2, X'^2_2) = \Phi(X)^\top \Phi(X')$$

with $\Phi : \mathbb{R}^2 \to \mathbb{R}^3$. This works similarly for $k_m(X, X') = (X^\top X)^m$ with a feature map $\Phi : \mathbb{R}^d \to \mathbb{R}^D$, where $D = \binom{d+m-1}{m}$, which grows exponentially with m. Luckily we never have to compute this feature map explicitly.

Theorem 2.11 (Mercer[7])**.** *Define the operator $T_k f(x) = \int_{\mathcal{X}} k(x, x')f(x')\mathrm{d}x'$. If the kernel k is continuous and $\int_{\mathcal{X}^2} k(x, x')f(x)f(x')\mathrm{d}x\mathrm{d}x' \geq 0$ for all $f \in L^2(\mathcal{X})$, then T_k has eigenfunctions $\psi_i \in L^2(\mathcal{X})$ with eigenvalues $\lambda_i \geq 0$ and we can write*

$$k(X, X') = \sum_{i=1}^{\infty} \lambda_i \psi_i(X)\psi_i(X')$$

for all $X, X' \in \mathcal{X}$. Furthermore, this series converges uniformly.

Proof. See [21], theorem 3.a.1. □

Remark 2.12. *For our feature mapping we could then consider*

$$\Phi(X) = (\sqrt{\lambda_1}\psi_1(X), \sqrt{\lambda_2}\psi_2(X), \dots)^\top.$$

As we also have $\Phi(X) = k(\cdot, X)$ we see that kernels are not unique.

Remark 2.13. *The Cauchy-Schwarz inequality gives*

$$|f(x) - f(x')| = |\langle f, \Phi(x) - \Phi(x')\rangle_{\mathcal{H}}| \leq \|f\|_{\mathcal{H}} \|\Phi(x) - \Phi(x')\|_{\mathcal{H}},$$

which separates the model on the left-hand side in a term independent of the data and a distance w.r.t. the geometry of the kernel (independent of f).

[7] Mercer's theorem is also used to analyze stochastic processes: k is the covariance function, and representing it as an inner product between sequences of orthogonal eigenfunctions allows a stochastic process to be represented as a sum of these eigenfunctions with random coefficients (see Karhunen-Loéve theorem).

Theorem 2.14 (Representer theorem)**.** *Given a positiv definite kernel $k\colon \mathcal{X}\times\mathcal{X} \to \mathbb{R}$, a sample $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y}$, a strictly monotonic increasing function $h : \mathbb{R} \to \mathbb{R}$ and an arbitrary loss function $l : (\mathcal{X} \times \mathcal{Y})^n \to \mathbb{R}$, then any $f \in \mathcal{H}$ minimizing the regularized loss functional*

$$l((f(X_1), Y_1), \ldots, (f(X_n), Y_n)) + h(\|f\|_{\mathcal{H}})$$

admits a representation of the form

$$f(X) = \sum_{i=1}^{n} \alpha_i k(X_i, X).$$

If l is convex and non-negativ, then for all $\lambda > 0$ there is a unique solution to the minimization of

$$l((f(X_1), Y_1), \ldots, (f(X_n), Y_n)) + \lambda\|f\|^2_{\mathcal{H}}.$$

Proof. Let $\mathcal{H}_{||} = \operatorname{span}\{k(X_1, \cdot), \ldots, k(X_n, \cdot)\}$. Then we have for every $f = f_{||} + f_\perp \in \mathcal{H}$ with $f_{||} \in \mathcal{H}_{||}, f_\perp \in \mathcal{H}^\perp_{||}$

$$\forall i = 1, \ldots, n : f(X_i) = f_{||}(X_i), \qquad \|f\|^2_{\mathcal{H}} = \|f_{||}\|^2_{\mathcal{H}} + \|f_\perp\|^2_{\mathcal{H}}.$$

If f is a minimizer we must have $f_\perp = 0$. Therefore $f \in \mathcal{H}_{||}$ and the first claim follows.

If l is convex and non-negativ, then also $f \mapsto K(f) := l((f(X_1), Y_1), \ldots, (f(X_n), Y_n)) + \lambda\|f\|^2_{\mathcal{H}}$ is. For $f \in \mathcal{H}$ with $\lambda\|f\|^2_{\mathcal{H}} > l(0, \ldots, 0)$ we have $K(f) > K(0)$. Then there exist $f_n \in \mathcal{H}$ with $\|f_n\|_{\mathcal{H}} \leq \lambda^{-\frac{1}{2}} l(0, \ldots, 0)^{\frac{1}{2}}$ and $K(f_n) \mapsto \min_{f\in\mathcal{H}} K(f)$ for $n \to \infty$. We again consider $f_n = (f_{||})_n + (f_\perp)_n$ and as before have $(f_{||})_n \mapsto \min_{f\in\mathcal{H}} K(f)$. Now $(f_{||})_n$ lies on the compact finite-dimensional ball $\{f_{||} \in \mathcal{H}_{||} : \|f_{||}\| \leq \lambda^{-\frac{1}{2}} l(0, \ldots, 0)^{\frac{1}{2}}\}$ and every limit point of $(f_{||})_n$ solves the minimization problem. For two different solutions f_1 and f_2 the parallelogram identity would yield for $\tilde{f} = \frac{1}{2}(f_1 + f_2)$

$$\|\tilde{f}\|_{\mathcal{H}} = \frac{1}{4}\left(2\|f_1\|^2_{\mathcal{H}} + 2\|f_2\|^2_{\mathcal{H}} - \|f_1 - f_2\|^2_{\mathcal{H}}\right) < \frac{1}{2}\left(\|f_1\|^2_{\mathcal{H}} + \|f_2\|^2_{\mathcal{H}}\right).$$

Due to the convexity of l this yields $K(\tilde{f}) < \frac{1}{2}(K(f_1) + K(f_2))$, i.e. strong convexity, which contradicts the minimality of f_1, f_2. Therefore we have uniqueness. □

Remark 2.15. *With the preceding theorem we can minimize in $\alpha \in \mathbb{R}^n$ and the infinite dimensional optimization problem reduces to a finite dimensional one, i.e. the solution of the minimization problem over $\mathcal{H}$ lies on a finite-dimensional subspace.*

Definition 2.16 (Universal kernel)**.** *A kernel k on a compact metric space $\mathcal{X}$ is said to be universal if the reproducing kernel Hilbert space $\mathcal{H}$ is dense (w.r.t. the uniform norm) in the space of continuous functions on $\mathcal{X}$.*

Remark 2.17. *For certain conditions on the loss, if k is universal, then* $\inf_{f\in\mathcal{H}} L(f) = L(f^B)$*, i.e. approximation error is zero. No finite dimensional RKHS is universal.*

Remark 2.18. *Many of the traditional algorithms can be transformed to a version with kernels instead of scalar products, leading to kernel ridge regression, kernel PCA, kernel K-means, support vector machines etc.*

Theorem 2.19. *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a s.p.s.d. kernel and $\Phi : \mathcal{X} \to \mathcal{H}$ the corresponding feature map. Let $S_n \subset \mathcal{X} \times \mathcal{Y}$ be a sample s.t. $k(X, X) \leq R^2$ for all $(X, Y) \in S_n$ and $\mathcal{F}_\Lambda = \{X \mapsto \langle f, \Phi(X)\rangle : \|f\|_{\mathcal{H}} \leq \Lambda\}$ for a $\lambda \geq 0$. Then*

$$R_n(\mathcal{F}) \leq \frac{\Lambda\sqrt{\operatorname{tr}(K)}}{n} \leq \sqrt{\frac{R^2\Lambda^2}{n}}.$$

Proof. The proof is analog to the one of theorem 3.27. □

3 Algorithms

3.1 Neural networks

We will start with the simplest neural network, the preceptron. It learns the function $f(X) = \operatorname{sgn}(\theta^\top X)$ via the following algorithm.

```
Input: $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathbb{R}^d \times \{-1, 1\}$
$\theta_0 = (0, \dots, 0)^\top \in \mathbb{R}^d, k = 0$
while some $(X_i, Y_i)$ is misclassified, i.e. $Y_i \neq \mathrm{sgn}(\theta_k^\top X_i)$ do
    pick some misclassified $(X_i, Y_i)$ at random
    $\theta_{k+1} \leftarrow \theta_k + Y_i X_i$
    $k \leftarrow k + 1$
end while
```

The corresponding class of functions that can be learnt, $\mathcal{F} = \{x \mapsto \mathrm{sgn}(\theta^\top x), \theta \in \mathbb{R}^d, x \in \mathbb{R}^d\}$, is called linear threshold functions.

Theorem 3.1. *Given that the data is linearly separable, i.e. there is a θ s.t. $Y_i \theta^\top X_i > 0$ for all i, the algorithm terminates (with zero empirical risk) after at most $\frac{R^2}{\gamma^2}$ iterations, where $R = \max_i \|X_i\|$ and $\gamma = \min_i \frac{\theta^\top X_i Y_i}{\|\theta\|}$.*

Proof. We have

$$\theta_{k+1}^\top \theta = (\theta_k + Y_i X_i)^\top \theta \geq \theta_t^\top \theta + \gamma \|\theta\|$$

and since $\theta_0 = 0$ this brings $\theta_k^\top \theta \geq k\gamma\|\theta\|$. On the other hand, since X_i was incorrectly classified,

$$\|\theta_{k+1}\|^2 = \|\theta_k + Y_i X_i\|^2 = \|\theta_k\|^2 + 2Y_i \theta_k^\top X_i + \|X_i\|^2 \leq \|\theta_k\|^2 + R^2,$$

so $\|\theta_k\|^2 \leq kR^2$. Combining those we get with Cauchy-Schwarz

$$t\gamma\|\theta\| \leq \theta_k^\top \theta \leq \|\theta_k\|\|\theta\| \leq \sqrt{k}R\|\theta\|$$

and therefore the statement. □

Remark 3.2. *As seen before the VC dimension of the perceptron in $\mathbb{R}^d$ is $d + 1$.*

Theorem 3.3. *For the class $\mathcal{F} = \{\theta^\top x \in [-1, 1], x \in \mathbb{R}^n, \theta \in \mathbb{R}^n, \|\theta\|_2 \leq A, \|x\|_2 \leq B\}$ we have*

$$\hat{R}_n(\mathcal{F}, S_n) \leq \frac{AB}{\sqrt{n}}.$$

Proof.

$$\begin{aligned}
\hat{R}_n(\mathcal{F}, S_n) &= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f \in \mathcal{F}}\left|\sum_{i=1}^n \sigma_i f(X_i)\right|\right] = \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{\|\theta\|_2 \leq A}\left|\sum_{i=1}^n \sigma_i \theta^\top X_i\right|\right] \\
&\leq \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{\|\theta\|_2 \leq A}\left|\theta^\top \sum_{i=1}^n \sigma_i X_i\right|\right] \leq \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{\|\theta\|_2 \leq A}\|\theta\|\left\|\sum_{i=1}^n \sigma_i X_i\right\|\right] \\
&= \frac{A}{n}\mathbb{E}_\sigma\left[\sqrt{\sum_{i,j=1}^n \sigma_i \sigma_j X_i^\top X_j}\right] \leq \frac{A}{n}\sqrt{\sum_{i,j=1}^n \mathbb{E}_\sigma\left[\sigma_i \sigma_j\right] X_i^\top X_j} \\
&\leq \frac{A}{n}\sqrt{\sum_{i=1}^n \|X_i\|^2} \leq \frac{AB}{\sqrt{n}}.
\end{aligned}$$

□

Theorem 3.4. *For the perceptron with n input data, i.e. $\mathcal{F} = \{x \mapsto \mathrm{sgn}(\theta^\top x - \theta_0) : \theta \in \mathbb{R}^d, x \in \mathbb{R}^d, \theta_0 \in \mathbb{R}\}$, the growth function is*

$$S_{\mathcal{F}}(n) = 2\sum_{k=0}^d \binom{n-1}{k}.$$

Proof. See [1], theorem 3.1. □

Definition 3.5 (Neural network). *A (feed-forward) neural network is a function $\Phi : \mathbb{R}^d \to \mathbb{R}^{N_L}$ given by*

$$\Phi(x) = W_L(\sigma(W_{L-1}(\sigma(\dots\sigma(W_1(x)))))),$$

where L is the number of layers, N_l the number of neurons in layer l for $1 \leq l \leq L$, $\sigma : \mathbb{R} \to \mathbb{R}$ is a (non-linear) function (applied component-wise) and $W_l : \mathbb{R}^{N_{l-1}} \to \mathbb{R}^{N_l}$ are affine linear maps, i.e. $W_l(x) = A_l x + b_l$, where $A_l \in \mathbb{R}^{N_l \times N_{l-1}}$ and $b_l \in \mathbb{R}^{N_l}$. Let $A = (A_1, \dots, A_L), b = (b_1, \dots, b_L)$, we then call (A, b) the architecture of the network. We write $\Phi \in \mathcal{NN}_{d,M,N,L,\sigma}$, where $N = \sum_{l=0}^{L} N_l$ and $M = \sum_{l=0}^{L} nz(A_l)$, $nz(\cdot)$ denoting the number of non-zeros entries.

Remark 3.6. *There are three major areas of research in the field of neural networks: the architecture of a network and therefore its function approximation properties, its optimization procedure, i.e. how empirical risk minimization is possible given the complex structure of the loss function, as well as the network's statistical generalization abilities – of course these questions interact.*

Remark 3.7 (Backpropagation). *In practice the training of a network is done with a variant of stochastic gradient decent, which is called backpropagation.*

We now study the approximation properties of neural networks. Let us start with an example that demonstrates why we should not take polynomials as activation functions.

Example 3.8. *If σ is a polynomial of degree q then $\sigma(Ax + b)$ is also a polynomial of degree q, hence $\Phi(x)$ is a polynomial of degree at most Lq, i.e. in this case $C(\mathbb{R}^d)$ cannot be approximated well.*

Theorem 3.9 (Weierstraß). *Let $f : [a, b] \to \mathbb{R}$ be a continuous function. For every $\epsilon > 0$ there exists a polynomial $p(x)$ such that for all $x \in [a, b]$*

$$|p(x) - f(x)| < \epsilon.$$

Proof. See 15.5 in [22]. See also (5.10) in [14]. □

Remark 3.10. *A generalization of this theorem is known as the Stone-Weierstraß approximation theorem, see for instance chapter 6 in [8].*

Theorem 3.11 (Universal approximation theorem). *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be continuous, but not a polynomial. Let $d \geq 1, L = 2$ and $K \subset \mathbb{R}^d$ be compact. Then for any continuous function $f : \mathbb{R}^d \to \mathbb{R}^{N_L}$ and every $\epsilon > 0$ there exist $M, N \in \mathbb{N}$ and $\Phi \in \mathcal{NN}_{d,M,N,L,\sigma}$ s.t.*

$$\sup_{x \in K} |\Phi(x) - f(x)| < \epsilon.$$

Proof. We consider the case $N_2 = 1$ w.l.o.g. Note that for an arbitrary N_1 the neural network can be written as

$$\Phi(x) = \sum_{i=1}^{N_1} a_i^{(2)} \sigma\left(\langle a_i^{(1)}, x\rangle - b_i^{(1)}\right) - b^{(2)},$$

where the i-th row of A_1 is ${a_i^{(1)}}^\top \in \mathbb{R}^d$ and $A_2 = {a^{(2)}}^\top \in \mathbb{R}^{N_1}$. Therefore we want to show that $\operatorname{span}\{\sigma(\langle a, x\rangle - b), a \in \mathbb{R}^d, b \in \mathbb{R}\}$ is dense in the space of continuous functions from K to $\mathbb{R}$. We consider first the case $d = 1, \sigma$ smooth. Since σ is not a polynomial there is an $x_0 \in \mathbb{R}$ with $\sigma^{(k)}(-x_0) \neq 0$ for $k \in \mathbb{Z}$. We have

$$\sigma(hx - x_0) \to \sigma(-x_0) \quad \text{as} \quad h \to 0,$$

i.e. every constant can be approximated arbitrarily well. Since

$$\frac{\sigma((\lambda + h)x - x_0) - \sigma(\lambda x - x_0)}{h} \to x\sigma'(\lambda x - x_0) \quad \text{as} \quad h \to 0$$

(consider $\tilde{h} = hx$) we have

$$\frac{\sigma((\lambda + h)x - x_0) - \sigma(\lambda x - x_0)}{h} \to x\sigma'(-x_0) \quad \text{as} \quad h, \lambda \to 0,$$

which implies that any polynomial can be approximated arbitrarily well. Theorem 3.9 concludes the proof. Now we consider the case of an arbitrary d and a smooth σ. We take $\operatorname{span}\{g(\langle a, x\rangle - b), a \in \mathbb{R}^d, b \in \mathbb{R}, g \in C(\mathbb{R})\}$. This is dense in $C(K)$, since we can choose $g = \sin, g = \cos$ and then use approximation by Fourier series. So we can approximate any $f \in C(K)$ by

$$\sum_{i=1}^{N} d_i g_i(\langle v_i, x\rangle - e_i), \quad v_i \in \mathbb{R}^d, d_i, e_i \in \mathbb{R}, g_i \in C(\mathbb{R}).$$

Now we can apply the previous result for $d = 1$ to approximate functions $t \mapsto g_i(t - e_i)$ for $1 \leq i \leq N$ using neural networks. For the case of a non-smooth σ we just pick a family $(g_\epsilon)_{\epsilon>0}$ with $\sigma * g_\epsilon$ smooth and $\sigma * g_\epsilon \to \sigma$ as $\epsilon \to 0$ uniformly on K. Then we apply the previous result to $\sigma * g_\epsilon$ and let $\epsilon \to 0$. □

Remark 3.12. *First versions of this proof can be found in [10, 18]; there one can also find more functional analytic arguments using for instance the Hahn-Banach theorem and the measure formulation of the Riesz representation theorem. We see that even "shallow" neural networks have the universal approximation property – however, note that the theorem gives no information on M and N. A more quantitative analysis can for instance be found in [24, 27].*

For better understanding the restrictions on M and N let us borrow from information theory.

Definition 3.13 (Rate distortion theory)**.** *Let $d, l \in \mathbb{N}$. We define the set of binary encoders*

$$\mathcal{E}^l := \left\{E : L^2(\mathbb{R}) \to \{0,1\}^l\right\}$$

and the set of binary decoders

$$\mathcal{D}^l := \left\{D : \{0,1\}^l \to L^2(\mathbb{R})\right\}.$$

For $\epsilon > 0$ and $\mathcal{C} \subset L^2(\mathbb{R}^d)$ the minimax code length is given by

$$L(\epsilon, \mathcal{C}) := \min\{l \in \mathbb{N} : \exists (E, D) \in \mathcal{E}^l \times \mathcal{D}^l : \sup_{f \in \mathcal{C}} \|D(E(f)) - f\|_{L^2(\mathbb{R}^d)} < \epsilon\}$$

and the optimal exponent $\gamma^*(\mathcal{C})$ is defined by

$$\gamma^*(\mathcal{C}) := \inf\{\gamma \in \mathbb{R} : L(\epsilon, \mathcal{C}) = \mathcal{O}(\epsilon^{-\gamma})\}.$$

Remark 3.14. *Certainly $l \to \infty$ if $\epsilon \to 0$. γ then determines the rate, i.e. how fast this happens and $\gamma^*(\mathcal{C})$ can be understood as a measure of the complexity of the function class $\mathcal{C}$.*

Theorem 3.15. *Let $\Phi \in \mathcal{NN}_{d,\infty,\infty,\infty,\sigma}$ be an approach to learn the function f with at most error ϵ s.t. all weights can be encoded with $-c\log_2 \epsilon$ bits, i.e. $\sup_{f \in \mathcal{C}} \|f - \Phi\|_{L^2(\mathbb{R}^d)} < \epsilon$. Then for all $\gamma < \gamma^*(\epsilon)$*

$$\epsilon^\gamma \sup_{f \in \mathcal{C}} M(\Phi) = \infty.$$

as $\epsilon \to 0$.

Proof. See theorem 2.7 in [4]. □

Remark 3.16. *The theorem shows that there exists a fundamental lower bound on the number of weights. Loosely speaking, M converges faster to infinity than ϵ to zero. A next question is: What happens for $\gamma = \gamma^*(\mathcal{C})$? Indeed one can show that in this case, given some assumptions, there exists a neural network with only $\mathcal{O}(\epsilon^{-\gamma^*})$ weights*[8].

Definition 3.17 (Affine system)**.** *Let $d \in \mathbb{N}, (A_j)_{j \in J} \subset \mathrm{GL}(\mathbb{R}^d), \psi_1, \ldots, \psi_s \in L^2(\mathbb{R}^d)$ compactly supported. We define affine systems as*

$$\{\det(A_j)^{\frac{1}{2}} \psi_k(A_j x - b) : k = 1, \ldots, s, b \in \mathbb{Z}, j \in J\}.$$

[8] The idea is to use results from approximation theory, e.g. approximate functions with wavelets, shearlets, curvelets etc., which in some sense yield an optimal approximation, and show that those basis elements can again be approximated by neural networks (see for instance [4]; here the M-term approximation concept is mimicd by neural networks; furthermore there is numerical evidence that meaningful basis functions get learnt).

Theorem 3.18. *Let $(\varphi_i)_{i\in I} \subset L^2(\Omega)$, Ω bounded, be an affine system. Suppose for the activation function $\sigma : \mathbb{R} \to \mathbb{R}$ there exists a constant $c > 0$ such that for all $l, \epsilon > 0$ there exists a $\Phi \in \mathcal{NN}_{L,c,d,\sigma}$ with*

$$\|\psi_k - \Phi\|_{L^2([-l,l]^d)} < \epsilon.$$

Let $\mathcal{C} \subset L^2(\mathbb{R}^d)$. Then if $\epsilon > 0, M \in \mathbb{N}, f \in L^2(\Omega) \subset \mathcal{C}$ such that there exists $(d_i)_{i=1}^M$ with

$$\|f - \sum_{i=1}^{M} d_i\varphi_i\| < \epsilon$$

then there exists a neural network Φ with $\mathcal{O}(M)$ edges such that

$$\|f - \Phi\| < 2\epsilon.$$

Proof. See theorem 3.4 in [4]. □

Let us now state some results regarding statistical generalization and note that it is still not completely understood why deep networks generalize well. To give an example, it is not clear why neural networks with sigmoidal activation functions should have a bounded VC dimension (recall also from example 1.43 that the VC dimension is not necessarily linked to the number of free parameters and note that certain activation functions can make the VC dimension of neural networks unbounded, see also theorem 7.1 in [1]).

Theorem 3.19. *Let σ be a piecewise polynomial of degree at most l with p pieces. Let further $h = \mathbb{1}_{\mathbb{R}_+}$ and for $d, M, N \in \mathbb{N}$ let $H_{d,M,N,\infty,\sigma} := \{h \circ \Phi : \Phi \in \mathcal{NN}_{d,M,N,\infty,\sigma}\}$. Then*

$$\mathrm{VC}(H_{d,M,N,\infty,\sigma}) = \mathcal{O}(M(M + lN\log_2 p)).$$

Proof. See [1], theorem 8.7. Also compare to theorems 6.1 - 6.4. □

Theorem 3.20. *Let σ be a piecewise polynomial of degree at most l with p pieces. Let further $h = \mathbb{1}_{\mathbb{R}_+}$ and for $d, M, N, L \in \mathbb{N}$ let $H_{d,M,N,L,\sigma} := \{h \circ \Phi : \Phi \in \mathcal{NN}_{d,M,N,L,\sigma}\}$. Then*

$$\mathrm{VC}(H_{d,M,N,L,\sigma}) \leq 2ML\log_2\left(\frac{4MLpN}{\log 2}\right) + 2ML^2\log_2(l+1) + 2L.$$

For fixed $p, l, N \leq M$ this yields

$$\mathrm{VC}(H_{d,M,N,L,\sigma}) \leq \mathcal{O}(ML\log_2 M + ML^2).$$

Proof. See [1], theorem 8.8. □

Remark 3.21. *When taking the ReLU $\sigma(x) = \max(0, x)$ as activation function we get a VC dimension that is $\mathcal{O}(ML\log M)$.*

Interestingly, the VC dimension can be used to proof the following approximation theorem.

Theorem 3.22. *Let $d, N, L \in \mathbb{N}$ and $H = \{h \in C^n([0,1]^d) : \|h\|_{C^n} \leq 1\}$. Further let σ be a ReLU. If for every $\epsilon > 0$ and every $h \in H$ there exists a neural network $\Phi \in \mathcal{NN}_{d,M(\epsilon),N(\epsilon),L,\sigma}$ s.t.*

$$\|h - \Phi\|_\infty < \epsilon$$

then $M(\epsilon) = \Omega(\epsilon^{-\frac{d}{n(1+\delta)}})$ for $\delta > 0$.

Proof. See [34]. □

3.2 Adaptive boosting

The idea of adaptive boosting is to have multiple maybe not so well performing classifiers, and combine them to a single classifier that performs well [3, 20, 29]. We again consider binary classification, i.e. $\mathcal{Y} = \{-1, 1\}$ and choose K multiple base classifiers $\{f_k\}_{k=1}^K$ with $f_k \in \mathcal{F}$. We then consider the composition classifier $\tilde{f}_K(X) = \operatorname{sgn}\left\{\sum_{k=1}^K \alpha_k f_k(X)\right\}$, where $\alpha_k \in \mathbb{R}$ are weights – so the overall classification can be interpreted as a "weighted majority rule": each datum X gets mapped to a vector $(f_1(X), \ldots, f_K(X)) \in \{-1, 1\}^T$ and the weights tell us how to combine these multiple predictions. We now want to study the generalization and therefore estimation error. We have

$$\begin{aligned}
L_n(\tilde{f}_K) &= \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{\tilde{f}_K(X_i) \neq Y_i\} \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{Y_i \sum_{k=1}^K \alpha_k f_k(X_i) \leq 0\} \\
&\leq \frac{1}{n}\sum_{i=1}^n \exp\left(-Y_i \sum_{k=1}^K \alpha_k f_k(X_i)\right) =: \mathcal{L}_n(\tilde{f}_K),
\end{aligned}$$

where the last line follows from $\mathbb{1}\{z \leq 0\} \leq e^{-z}$. Instead of minimizing $L_n(\tilde{f}_K)$ directly, which can be difficult as it is a non-continuous function, we can minimize the upper bound $\mathcal{L}_n(\tilde{f}_K)$ and hope that we will find a "similar" minimium. Note that we still consider a minimization over $f_1, \ldots, f_K$ as well as $\alpha_1, \ldots, \alpha_K$. Since this is still computationally heavy we consider an iterative approach, a greedy optimitazion: Imagine we already tuned $f_1, \ldots, f_K$ and $\alpha_1, \ldots, \alpha_K$, how do we then choose f_{K+1} and α_{K+1}? Let us change the writing a bit:

$$\begin{aligned}
\mathcal{L}_n(\tilde{f}_{K+1}) &= \sum_{i=1}^n \underbrace{\frac{1}{n}\exp\left(-Y_i \sum_{k=1}^K \alpha_k f_k(X_i)\right)}_{=: w_i^{(K)}} \exp\left(-Y_i \alpha_{K+1} f_{K+1}(X_i)\right) \\
&= \sum_{i=1}^n w_i^{(K)} \exp\left(-Y_i \alpha_{K+1} f_{K+1}(X_i)\right).
\end{aligned}$$

Let $\tilde{w}_i^{(K)} := \frac{w_i^{(K)}}{\sum_{i=1}^n w_i^{(K)}}$, then

$$\begin{aligned}
\frac{\mathcal{L}_n(\tilde{f}_{K+1})}{\mathcal{L}_n(\tilde{f}_K)} &= \sum_{i=1}^n \tilde{w}_i^{(K)} \exp\left(-Y_i \alpha_{K+1} f_{K+1}(X_i)\right) \\
&= e^{-\alpha_{K+1}} \sum_{i: f_{K+1}(X_i) = Y_i} \tilde{w}_i^{(K)} + e^{\alpha_{K+1}} \sum_{i: f_{K+1}(X_i) \neq Y_i} \tilde{w}_i^{(K)} \\
&= \left(e^{\alpha_{K+1}} - e^{-\alpha_{K+1}}\right) \sum_{i=1}^n \tilde{w}_i^{(K)} \mathbb{1}\{f_{K+1}(X_i) \neq Y_i\} + e^{-\alpha_{K+1}}.
\end{aligned}$$

Note that $f_{K+1} \in \mathcal{F}$ only appears in $\epsilon_K := \sum_{i=1}^n \tilde{w}_i^{(K)} \mathbb{1}\{f_{K+1}(X_i) \neq Y_i\}$. Minimization of the above expression gives

$$\alpha_{K+1}^* = \frac{1}{2}\log\frac{1 - \epsilon_K}{\epsilon_K},$$

which is positive if $\epsilon_K < \frac{1}{2}$, which we expect since $\epsilon_K = \frac{1}{2}$ would just correspond to random guessing. Note that

$$w_i^{(K+1)} = w_i^{(K)} e^{-Y_i f_{K+1}(X_i) \alpha_{K+1}},$$

which explains the name of "adaptive boosting". While learning we concentrate on those predictors that have made mistakes in the previous training rounds. We now make the assumption:

$$\epsilon_k < \frac{1}{2} - \gamma, \quad \gamma > 0$$

for every $k = 1, \dots, K$. Note that this assumption depends on choosing a good function class $\mathcal{F}$. We then have

$$\begin{aligned}
\mathcal{L}_n(\tilde{f}_{K+1}) &= \mathcal{L}_n(\tilde{f}_K)\left(\left(e^{\alpha^*_{K+1}} - e^{-\alpha^*_{K+1}}\right)\epsilon_K + e^{-\alpha^*_{K+1}}\right) \\
&= \mathcal{L}_n(\tilde{f}_K)2\sqrt{\epsilon_K(1-\epsilon_K)} \\
&\leq \mathcal{L}_n(\tilde{f}_K)2\sqrt{\left(\frac{1}{2}-\gamma\right)\left(\frac{1}{2}+\gamma\right)} \\
&= \mathcal{L}_n(\tilde{f}_K)\sqrt{1-4\gamma^2} \\
&\leq \mathcal{L}_n(\tilde{f}_K)e^{-2\gamma^2}
\end{aligned}$$

and therefore

$$L_n(\tilde{f}_K) \leq \mathcal{L}_n(\tilde{f}_K) \leq (e^{-2\gamma^2})^K \mathcal{L}(\tilde{f}_0) = e^{-2K\gamma^2}.$$

Since $L_n(\tilde{f}_K) \in \{0, \frac{1}{n}, \frac{2}{n}, \dots, 1\}$ we now know that AdaBoost achieves zero empirical loss in finite steps. For investigating the generalization error let us say $\tilde{f}_K \in C(\mathcal{F}, K)$ and note that we have $C(\mathcal{F}, K) \subset C(\mathcal{F}, K+1)$. This means that the approximation error gets smaller, however at the same time the estimation error could get worse. Theorems 1.39 and 1.44 give us

$$L(\tilde{f}_K) \leq L_n(\tilde{f}_K) + \mathcal{O}\left(\sqrt{\frac{\mathrm{VC}(C(\mathcal{F}, K))\log n + \log\frac{1}{\delta}}{n}}\right),$$

where the first summand goes to zero, however, one can show that $\mathrm{VC}(C(\mathcal{F}, K)) \approx K\,\mathrm{VC}(\mathcal{F})$ (see lemma 10.3 in [30]). So the error should grow sublinearly in K. Still, in practice, one can observe that the loss on a test set keeps decreasing even after the empirical risk has gone to zero. In order to understand why this is the case another approach is necessary that shall be summarized with the following theorem.

Lemma 3.23. *For the Rademacher complexity of convex hulls of function classes we have*

$$\hat{R}_n(\mathrm{conv}(\mathcal{F})) = \hat{R}_n(\mathcal{F}).$$

Proof.

$$\begin{aligned}
\hat{R}_n(\mathrm{conv}(\mathcal{F})) &= \mathbb{E}_\sigma\left[\sup_{f_k\in\mathcal{F}, \alpha\geq 0, \|\alpha\|_1=1}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i \sum_{k=1}^K \alpha_k f_k(X_i)\right|\right] \\
&= \mathbb{E}_\sigma\left[\sup_{f_k\in\mathcal{F}}\sup_{\alpha\geq 0, \|\alpha\|_1=1}\left|\frac{1}{n}\sum_{k=1}^K \alpha_k \sum_{i=1}^n \sigma_i f_k(X_i)\right|\right] \\
&= \mathbb{E}_\sigma\left[\sup_{f_k\in\mathcal{F}}\max_{1\leq k\leq K}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i f_k(X_i)\right|\right] \\
&= \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i)\right|\right] \\
&= \hat{R}_n(\mathcal{F}).
\end{aligned}$$

□

Remark 3.24. *We have noted before that this is not valid for the VC dimension.*

Theorem 3.25. *Let $\mathcal{Y} = \{-1, 1\}$, $l(y', y) = \mathbb{1}\{y' \neq y\}$. We consider the base class $\mathcal{F}$ with $\mathrm{VC}(\mathcal{F}) < \infty$ and classifiers of the form $\tilde{f}(x) = \mathrm{sgn}\{g(x)\}$, where $g \in \mathrm{conv}(\mathcal{F}) := \{g(x) = \sum_{k=1}^K \alpha_k f_k(x), f_1, \dots, f_K \in \mathcal{F}, \alpha_k > 0, \sum_{k=1}^K \alpha_k = 1\}$. Then for all $\eta > 0$ with probability $\geq 1 - \delta$*

$$L(\tilde{f}) \leq \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{g(X_i)Y_i \leq \eta\} + \frac{2}{\eta}\sqrt{\frac{2\,\mathrm{VC}(\mathcal{F})\log n}{n}} + \sqrt{\frac{\log\frac{1}{\delta}}{2n}}.$$

Proof. Let

$$L(\tilde{f}) = \mathbb{E}[\mathbb{1}\{\tilde{f}(X) \neq Y\}] = \mathbb{E}[\mathbb{1}\{g(X)Y \leq 0\}].$$

Take any function $\varphi : \mathbb{R} \to [0,1]$ that is L_φ-Lipschitz and for that $\varphi(x) \geq \mathbb{1}\{x \leq 0\}$. Then with theorem 1.33

$$\begin{aligned} L(\tilde{f}) &\leq \mathbb{E}[\varphi(g(X)Y)] \\ &\leq \frac{1}{n}\sum_{i=1}^{n} \varphi(g(X_i)Y_i) + 2R_n(\varphi \circ \operatorname{conv}(\mathcal{F})) + \sqrt{\frac{\log\frac{1}{\delta}}{2n}} \end{aligned}$$

with

$$\begin{aligned} R_n(\varphi \circ \operatorname{conv}(\mathcal{F})) &= \mathbb{E}_{S_n,\sigma}\left[\sup_{g \in \operatorname{conv}(\mathcal{F})} \frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i \varphi(g(X_i)Y_i)\right|\right] \\ &= \mathbb{E}_{S_n,\sigma}\left[\sup_{g \in \operatorname{conv}(\mathcal{F})} \frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i \left(\varphi(g(X_i)Y_i) - \varphi(0)\right)\right|\right] \\ &\leq L_\varphi \mathbb{E}_{S_n,\sigma}\left[\sup_{g \in \operatorname{conv}(\mathcal{F})} \frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i g(X_i)Y_i\right|\right] \\ &= L_\varphi \mathbb{E}_{S_n,\sigma}\left[\sup_{g \in \mathcal{F}} \frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i g(X_i)\right|\right] \\ &\leq L_\varphi \sqrt{\frac{2\,\mathrm{VC}(\mathcal{F})\log n}{n}}, \end{aligned}$$

where we used lemmas 1.49 and 3.23, theorem 1.46 and the fact that $\sigma_i \overset{d}{=} \sigma_i Y_i$. Setting

$$\varphi(x) = \begin{cases} 1 & \text{if } x \leq 0 \\ 0 & \text{if } x \geq \eta \\ 1 - \frac{x}{\eta} & \text{in between} \end{cases}$$

yields $L_\varphi = \frac{1}{\eta}$ and completes the proof. □

Remark 3.26. *$|g(X_i)Y_i|$ can be interpreted as a margin or a confidence of a prediction $g(X_i)$. Then, the indicator function not only counts the number of wrong classifications, but also the correct classifications that have however not been confident enough (depending on the parameter η).*

3.3 Support vector machines

Let us w.l.o.g. consider homogeneous hyperplanes, i.e. our classifiers are $f(X_i) = \operatorname{sgn}(\langle X_i, w\rangle)$. In the linear separable case we have $Y_i\langle X_i, w\rangle > 0$ for all $i = 1, \ldots, n$ (i.e. a hard margin), but there are infinitely many hyperplanes separating our data. The idea is to choose the hyperplane that maximizes the distance to closest data points. The distance to a point X_0 is $\frac{|\langle X_0, w\rangle|}{\|w\|}$ so that we consider the optimization problem

$$\max_{w \in \mathbb{R}^d} \min_{i=1,\ldots,n} \frac{|\langle X_i, w\rangle|}{\|w\|} \qquad \text{s.t.} \qquad Y_i\langle X_i, w\rangle > 0.$$

Note that scaling $\tilde{w} = cw, c \in \mathbb{R}$ does not change anything. Therefore we will only consider vectors w s.t. $\min_{i=1,\ldots,n} |\langle X_i, w\rangle| = 1$ (then no data point is in the margin region). This yields

$$\max_{w \in \mathbb{R}^d} \frac{1}{\|w\|} \qquad \text{s.t.} \qquad Y_i\langle X_i, w\rangle > 1.$$

For notational convenience we consider equivalently

$$\min_{w \in \mathbb{R}^d} \frac{1}{2}\|w\|^2 \qquad \text{s.t.} \qquad Y_i\langle X_i, w\rangle > 1.$$

This form of the hard margin SVM is called primal form. The corresponding dual form is

$$\min_{w\in\mathbb{R}^d} \frac{1}{2}\|w\|^2 + \frac{\lambda}{n}\sum_{i=1}^{n} \mathbb{1}\{Y_i\langle X_i, w\rangle < 1\},$$

where $\lambda > 0$ and for $\lambda \to \infty$ it becomes equivalent to the primal form. Note that we can also write

$$\min_{f\in\mathcal{H}} \frac{1}{2}\|f\|_{\mathcal{H}}^2 + \frac{\lambda}{n}\sum_{i=1}^{n} \mathbb{1}\{Y_i\langle X_i, w\rangle < 1\},$$

with a linear kernel on $\mathbb{R}^d$ and the corresponding RKHS $\mathcal{H}$. Then the representer theorem 2.14 gives

$$f_\lambda^*(X) = \sum_{i=1}^{n} \alpha_{i,\lambda} k(X_i, X) = \sum_{i=1}^{n} \alpha_{i,\lambda}\langle X_i, X\rangle,$$

i.e. we have a solution of the form $w_\lambda^* = \sum_{i=1}^n \alpha_{i,\lambda} X_i$ and can therefore consider the minimization problem

$$\min_{\alpha\in\mathbb{R}^n} \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i\alpha_j\langle X_i, X_j\rangle \qquad s.t. \qquad Y_i\sum_{j=1}^{n}\alpha_j\langle X_i, X_j\rangle \geq 1.$$

Let us now drop the separability assumption and consider the <u>soft margin</u> SVM, meaning that we let the constraint be partially violated, namely

$$\min_{\substack{w\in\mathbb{R}^d\\ \xi\in\mathbb{R}^n}} \left\{\frac{1}{2}\|w\|^2 + \lambda\sum_{i=1}^{n}\xi_i\right\} \qquad \text{s.t.} \qquad Y_i\langle X_i, w\rangle \geq 1-\xi_i, \qquad \xi_i \geq 0.$$

Note that if $\xi_i > 1$ we allow the classifier to make a mistake on X_i and that $\lambda \to \infty$ leads to the hard margin SVM, given the separability assumption. If w is fixed the optimal slack variables are

$$\xi_i^* = (1 - Y_i\langle X_i, w\rangle)_+$$

and our unconstrained optimization problem becomes

$$\min_{w\in\mathbb{R}^d} \left\{\frac{1}{2}\|w\|^2 + \lambda\sum_{i=1}^{n}(1 - Y_i\langle X_i, w\rangle)_+\right\}.$$

If a point X_i is inside or at the boundary of the margin region $\{X \in \mathbb{R}^d : |\langle X, w\rangle| \leq 1\}$ it is called support vector. The corresponding (convex) loss is called hinge loss and we have $(1 - Yf(X))_+ \geq \mathbb{1}\{f(X)Y \leq 0\}$. Now, the representer theorem gives

$$\min_{\alpha\in\mathbb{R}^n} \left\{\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j\langle X_i, X_j\rangle + \lambda\sum_{i=1}^{n}\left(1 - Y_i\sum_{j=1}^{n}\alpha_j\langle X_i, X_j\rangle\right)_+\right\},$$

which is a convex problem.

Note that one can also approach the minimization problem by exploiting Lagrange duality, i.e.

$$\max_{\alpha,\beta\geq 0}\min_{w,\xi} \left\{\frac{1}{2}\|w\|^2 + \lambda\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i(1 - Y_i\langle X_i, w\rangle - \xi_i) - \sum_{i=1}^{n}\beta_i\xi_i\right\}.$$

One finds $\beta_i = \lambda - \alpha_i$ and $w = \sum_{i=1}^n \alpha_i Y_i X_i$ and finally the optimization problem

$$\max_{0\leq\alpha\leq c} \left\{\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j Y_i Y_j\langle X_i, X_j\rangle\right\}.$$

To get nonlinear decision boundaries one can use the kernel trick from section 2 and replace $\langle\cdot,\cdot\rangle$ with $k(\cdot,\cdot)$. In the primal version this can lead to an optimization problem over an infinite dimensional space, however in the dual version we only need to optimize over n variables.

Kernel SVMs use hyperplanes in the feature space $\mathcal{H}$, therefore the VC dimension of them is $\dim(\mathcal{H})+1$. We get:

Kernel	VC dimension
linear	$d+1$
polynomial of degree m	$\binom{d+m}{m}$
Gaussian	∞

Figure 2: VC dimensions of SVMs for different kernels.

Let us write $f_\alpha = \sum_{i=1}^n \alpha_i k(X_i,\cdot)$, then the optimization problem for the soft margin SVM can be written as

$$\min_{\Lambda\in\mathbb{R}} \min_{\substack{\alpha\in\mathbb{R}^n \\ \|f_\alpha\|_\mathcal{H}=\Lambda}} \left\{ \frac{1}{2}\Lambda^2 + \frac{\lambda}{n}\sum_{i=1}^n (1-Y_i f_\alpha(X_i))_+ \right\} = \min_{\Lambda\in\mathbb{R}} \min_{\substack{\alpha\in\mathbb{R}^n \\ \|f_\alpha\|_\mathcal{H}\leq\Lambda}} \left\{ \frac{1}{2}\Lambda^2 + \frac{\lambda}{n}\sum_{i=1}^n (1-Y_i f_\alpha(X_i))_+ \right\}.$$

To see this fix Λ' to be the solution of the inner optimization problem with a minimal value $f_\alpha^{\Lambda'}$. Now assume $\|f_\alpha^{\Lambda'}\|_\mathcal{H} = \Lambda'' < \Lambda'$, i.e. we consider functions that do not lie at the boundary. This yields

$$\begin{aligned}
\min_{\substack{\alpha\in\mathbb{R}^n \\ \|f_\alpha\|_\mathcal{H}\leq\Lambda''}} \left\{ \frac{1}{2}\Lambda''^2 + \frac{\lambda}{n}\sum_{i=1}^n (1-Y_i f_\alpha(X_i))_+ \right\} &\leq \frac{1}{2}\Lambda''^2 + \frac{\lambda}{n}\sum_{i=1}^n (1-Y_i f_\alpha^{\Lambda'}(X_i))_+ \\
&< \frac{1}{2}\Lambda'^2 + \frac{\lambda}{n}\sum_{i=1}^n (1-Y_i f_\alpha^{\Lambda'}(X_i))_+ \\
&= \min_{\substack{\alpha\in\mathbb{R}^n \\ \|f_\alpha\|_\mathcal{H}\leq\Lambda'}} \left\{ \frac{1}{2}\Lambda'^2 + \frac{\lambda}{n}\sum_{i=1}^n (1-Y_i f_\alpha(X_i))_+ \right\}.
\end{aligned}$$

This shows that Λ' is not a solution to the outer optimization problem. Therefore we have $\|f_\alpha^{\Lambda'}\|_\mathcal{H} = \Lambda'$. To conclude, there exists a number Λ_λ s.t. the unconstrained optimization problem is equivalent to

$$\min_{\alpha\in\mathbb{R}^n} \frac{1}{n}\sum_{i=1}^n (1-Y_i f_\alpha(X_i))_+ \qquad \text{s.t.} \qquad \|f_\alpha\|_\mathcal{H} \leq \Lambda_\lambda.$$

Thus the soft-margin kernel SVM with regularizer λ results in minimizing the empirical risk over the ball in the RKHS $\mathcal{H}$ with the radius Λ_λ. Let

$$\mathcal{F}_\Lambda := \left\{ f_\alpha \in \mathcal{H} : \alpha\in\mathbb{R}^n, \|f_\alpha\|_\mathcal{H} \leq \Lambda \right\}.$$

Note that a larger λ corresponds to balls of larger radius Λ_λ.

Theorem 3.27. *Consider $\mathcal{Y}=\{-1,1\}$ and let k be a kernel satisfying $k(X,X)\leq\Gamma$ for any $X\in\mathcal{X}$. Consider for $\eta>0$*

$$\varphi(x) = \begin{cases} 1 & \textit{if } x\leq 0 \\ 0 & \textit{if } x\geq\eta \\ 1-\frac{x}{\eta} & \textit{in between} \end{cases}.$$

Then with probability $\geq 1-\delta$ we have for every $f\in\mathcal{F}_\Lambda, \Lambda>0$

$$L(f) = \mathbb{P}\left(Y\neq \operatorname{sgn}(f(X))\right) \leq \frac{1}{n}\sum_{i=1}^n \varphi(Y_i f(X_i)) + \frac{2\Lambda}{\eta}\sqrt{\frac{\Gamma}{n}} + \sqrt{\frac{\log\frac{1}{\delta}}{n}}$$

Remark 3.28. *Note that $\eta\to 0$ leads to approaching the binary loss. If we use a Gaussian kernel, the RKHS is infinite dimensional, and VC theory tells us that the problem is not learnable (compare to figure 2). This is in accordance with the observation that the upper bound explodes when $\eta\to 0$. Note that the empirical φ-loss is small if f makes a lot of correct confident answers with large margin.*

Remark 3.29. *Note that theorem 3.27 is not directly applicable to soft-margin SVMs since the hinge loss is unbounded. However, looking into the proof gives that we can replace φ with any L-Lipschitz function bounded in $[0, M]$ s.t. $\varphi(x) \geq \mathbb{1}\{x \leq 0\}$, resulting in $\frac{2}{\eta}$ being replaced by $2L$.*

Proof. Recall that theorems 1.33 and 3.25 bring with probability $\geq 1 - \delta$ that for all $f \in \mathcal{F}_\Lambda$

$$\mathbb{P}\left(Y \neq \operatorname{sgn}(f(X))\right) \leq \frac{1}{n}\sum_{i=1}^{n} \varphi(Y_i f(X_i)) + \frac{2}{\eta} R_n(\mathcal{F}_\Lambda) + \sqrt{\frac{\log \frac{1}{\delta}}{n}}.$$

It remains to bound the Rademacher complexity. We have $\mathcal{F}_\Lambda \subset \{f \in \mathcal{H} : \|f\|_\mathcal{H} \leq \Lambda\}$ and therefore

$$\begin{aligned}
\mathbb{E}_\sigma\left[\sup_{f \in \mathcal{F}_\Lambda}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(X_i)\right|\right] &\leq \mathbb{E}_\sigma\left[\sup_{\substack{f\in\mathcal{H}\\ \|f\|_\mathcal{H}\leq\Lambda}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(X_i)\right|\right] \\
&= \mathbb{E}_\sigma\left[\sup_{\substack{f\in\mathcal{H}\\ \|f\|_\mathcal{H}\leq\Lambda}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i \langle f, k(X_i,\cdot)\rangle_\mathcal{H}\right|\right] \\
&= \mathbb{E}_\sigma\left[\sup_{\substack{f\in\mathcal{H}\\ \|f\|_\mathcal{H}\leq\Lambda}}\left|\left\langle f, \frac{1}{n}\sum_{i=1}^{n}\sigma_i k(X_i,\cdot)\right\rangle_\mathcal{H}\right|\right] \\
&= \frac{\Lambda}{n}\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^{n}\sigma_i k(X_i,\cdot)\right\|_\mathcal{H}\right] \\
&\leq \frac{\Lambda}{n}\sqrt{\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^{n}\sigma_i k(X_i,\cdot)\right\|_\mathcal{H}^2\right]},
\end{aligned}$$

where we used the Cauchy-Schwarz and Jensen's inequality. Noting that

$$\left\|\sum_{i=1}^{n}\sigma_i k(X_i,\cdot)\right\|_\mathcal{H}^2 = \sum_{i=1}^{n}\sigma_i^2 k(X_i,X_i) + \sum_{i\neq j}^{n}\sigma_i\sigma_j k(X_i,X_j)$$

and $\sigma_i^2 = 1, \mathbb{E}[\sigma_i\sigma_j] = 0$ concludes the proof. □

4 Variational inference

One motivation for variational inference comes from Bayes' theorem.

Theorem 4.1 (Bayes' theorem)**.** *We consider random variables X, Z and their probability densities p_X, p_Z. Then*

$$p_Z(z|X = x) = \frac{p_X(x|Z = z)p_Z(z)}{p_X(x)},$$

where usually X is data and Z are parameters. $p_Z(z|X = x)$ is called posterior distribution, $p_X(x|Z = z)$ likelihood, $p_Z(z)$ prior and $p_X(x)$ evidence[9]*.*

Proof. See [16]. □

Remark 4.2. *Unlike in the frequentist approach, in the Bayesian attempt we usually do not consider the parameters Z explicitly, but only consider probability distributions over them – this is also why we do not use greek letters and sometimes call them "latent variables" instead of parameters.*

Remark 4.3. *Note Bayes' theorem for multiple variables, namely*

$$p(z, w|x) = \frac{p(x, w, z)}{p(x)} = \frac{p(z|x, w)p(x, w)}{p(x)} = p(z|x, w)p(w|x).$$

[9] We will continuously omit the capital letters of the random variables and just write $p(z|x)$ for $p_Z(z|X = x)$ etc.

Remark 4.4. *Note the notational subtleties: Here, the value x, meaning the random variable $X = x$, can refer to data from $\mathcal{X} \times \mathcal{Y}$ as well as to data from the input space $\mathcal{X}$ only (when for instance considering "unsupervised learning"). In the former case, one can for instance exploit the posterior distribution to make predictions via*

$$p(y^*|x, y, x^*) = \int_{\mathcal{Z}} p(y^*|x^*, z)p(z|x, y)\mathrm{d}z = \mathbb{E}_{p(z|x,y)}[p(y^*|x^*, z)],$$

where we average over all point estimates and weight them according to their probabilities – $\mathcal{Z}$ is the latent variable space, which is often just $\mathbb{R}^p$. Here $x^* \in \mathcal{X}$ and $y^* \in \mathcal{Y}$ are the values of future data and the values x and y of the random variables $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$ are the observed data.

Often, computing the posterior is hard – the difficulty mostly lies in the term $p(x) = \int_{\mathcal{Z}} p(z, x)\mathrm{d}z$. There are two main approaches to still handle this integral: sampling and variational inference. We want to focus on the latter.

Definition 4.5 (Kullback-Leibler (KL) divergence)**.** *Given two probability measures $\mathbb{P}$ and $\mathbb{Q}$ with respective densities p and q the Kullback-Leibler divergence is defined as*

$$\mathrm{KL}(p\|q) = \int_{\mathcal{X}} \log \frac{p(x)}{q(x)} p(x)\mathrm{d}x.$$

Remark 4.6. *With an abuse of notation we sometimes write $\mathrm{KL}(\mathbb{P}\|\mathbb{Q})$ and sometimes $\mathrm{KL}(p\|q)$, but mean the same thing. We have $\mathrm{KL}(p\|q) \geq 0$ as well as the property that $\mathrm{KL}(p\|q) = 0$ if and only if $p = q$ almost everywhere. Note, however, that the KL divergence is not symmetric and does not fulfill the triangle inequality, i.e. it is not a metric. Still, it can be related to other metrics, e.g. to the total variation distance*[10] *as in Pinsker's inequality: $\|p - q\|_{\mathrm{TV}} \leq \sqrt{2\,\mathrm{KL}(p\|q)}$ (see e.g. lemma 2.5 in [31]).*

Remark 4.7. *A generalization of the KL divergence is called f-divergence, which we will define later in definition 4.18.*

The idea of variational inference is to find a density $q(z)$ out of a family of (tractable) densities such that it is close to the desired (intractable) quantity $p(z|x)$. For this we consider $\mathrm{KL}(q(z)\|p(z|x)) = \mathbb{E}_{q(z)}\left[\log \frac{q(z)}{p(z|x)}\right]$ and note that minimizing this quantity w.r.t. to $q(z)$ is equivalent to maximizing the so called "evidence lower bound" (ELBO)[11]

$$\mathcal{L}(q(z)) = \mathbb{E}_{q(z)}[\log p(z, x) - \log q(z)].$$

Formally, we want to find the closest fit in a set of feasible distributions $\mathcal{Q}$, i.e.

$$q^*(z) = \operatorname*{arg\,min}_{q \in \mathcal{Q}} \mathrm{KL}(q(z)\|p(z|x)),$$

where usually the set $\mathcal{Q}$ is parametrized by some parameter $\lambda \in \mathbb{R}^l$. Therefore, our prediction problem becomes an optimization problem. For the prediction of new data as in remark 4.4 we can then write $p(y^*|x, y, x^*) = \mathbb{E}_{q^*(z)}[p(y^*|x^*, z)]$.

Remark 4.8. *Note the identity*

$$\mathcal{L}(q(z)) = \mathbb{E}_{q(z)}[\log p(x|z)] - \mathrm{KL}(q(z)\|p(z)),$$

where the first term can be interpreted as a goodness of reconstruction and last term as a regularizer in the optimization problem. Yet another way of writing the ELBO is

$$\mathcal{L}(q(z)) = \mathbb{E}_{q(z)}[\log p(z, x)] - \mathbb{E}_{q(z)}[\log q(z)],$$

i.e. as a sum of the negative energy and the entropy – which, when maximizing over $q \in \mathcal{Q}$, is sometimes termed "free energy" (see e.g. [17]).

A common approach to make computations easier is the following.

[10] The total variation distance is defined as $\mathrm{TV}(\mathbb{P}, \mathbb{Q}) = \sup_{A \in \Sigma} |\mathbb{P}(A) - \mathbb{Q}(A)|$, where Σ is the set of all Borel subsets of $\mathcal{X}$.

[11] The name comes from the fact that $\mathcal{L}(q(z)) \leq \log p(x)$ and that $p(x)$ is sometimes called "evidence".

Definition 4.9 (Mean-field variational distribution)**.** *We call q a mean-field variational distribution if it factorizes in the latent variables, namely*

$$q(z) = \prod_{i=1}^{m} q_i(z_i).$$

This then leads to the practical approach where one can iteratively optimize each factor of the mean-field variational density separately, while holding the others fixed.

Theorem 4.10 (Coordinate ascent variational inference)**.** *For the optimal update of a $q_j(z_j)$ with all other $q_{-j}(z_{-j}) := \prod_{i\neq j} q_i(z_i)$ being fixed we have $q_j^*(z_j) \propto \exp\left(\mathbb{E}_{\prod_{i\neq j} q_i(z_i)}[\log(p(z_j|z_{-j}, x))]\right)$.*

Proof.

$$\begin{aligned}\mathcal{L}(q(z)) &= \mathbb{E}_{q(z)}[\log(p(z_j|z_{-j}, x)p(z_{-j}, x)) - \log(q_j(z_j)q_{-j}(z_{-j}))] \\ &= \mathbb{E}_{q(z)}[\log p(z_j|z_{-j}, x) - \log q_j(z_j)] + \mathbb{E}_{\prod_{i\neq j} q_i(z_i)}[\log p(z_{-j}, x) - \log q_{-j}(z_{-j})].\end{aligned}$$

The idea is to take z_{-j} fixed and consider the ELBO as a function of z_j, i.e.

$$\begin{aligned}\mathcal{L}(q_j(z_j)) &= \mathbb{E}_{q(z)}[\log(p(z_j|z_{-j}, x)) - \log q_j(z_j)] + C \\ &= \mathbb{E}_{q_j(z_j)}\left[\mathbb{E}_{\prod_{i\neq j} q_i(z_i)}[\log(p(z_j|z_{-j}, x))] - \log q_j(z_j)\right] + C \\ &= -\operatorname{KL}\left(q_j(z_j)\middle\| \exp\left(\mathbb{E}_{\prod_{i\neq j} q_i(z_i)}[\log(p(z_j|z_{-j}, x))]\right)\right) + C,\end{aligned}$$

where the term C does not depend on z_j. Maximization of this expression yields $q_j^*(z_j) \propto \exp\left(\mathbb{E}_{\prod_{i\neq j} q_i(z_i)}[\log(p(z_j|z_{-j}, x))]\right)$, where "$\propto$" comes from the fact that in the end q^* shall be normalized. □

Remark 4.11. *Note that using the above suggestion as iterative updates only guarantees convergence to a local maximum.*

Remark 4.12. *In practice one often assumes $p(z_j|z_{-j}, x)$ to be in the exponential family, i.e. $p(z_j|z_{-j}, x) = h(z_j)\exp\left(\eta(z_{-j}, x)^\top T(z_j) - A(\eta(z_{-j}, x))\right)$. Then the optimal update is $q_j^*(z_j) \propto h(z_j)\exp\left(\mathbb{E}_{\prod_{i\neq j} q_i(z_i)}[\eta(z_{-j}, x)]^\top T(z_j)\right)$, which is also in the exponential family.*

An approach to deal with more general densities $q(z)$ (which for instance do not follow the mean-field assumption) is called "black box variational inference", where we usually have the situation that we cannot compute expectations in a closed form anymore.

$$\begin{aligned}\mathcal{L}(q(z|\lambda)) &= \mathbb{E}_{q(z|\lambda)}[\underbrace{\log p(z, x) - \log q(z|\lambda)}_{=:f(x,z|\lambda)}] \\ &\approx \frac{1}{K}\sum_{k=1}^{K} f(z_k, x|\lambda)\end{aligned}$$

with $z_k \sim q(z|\lambda)$, where $\lambda \in \mathbb{R}^l$ is a parameter specifying the distribution q. For the optimization one usually computes the gradient w.r.t. λ, i.e.

$$\begin{aligned}\nabla_\lambda \mathcal{L}(q(z|\lambda)) &= \nabla_\lambda \int_{\mathcal{X}} f(x, z|\lambda)q(z|\lambda)\mathrm{d}z \\ &= \int_{\mathcal{X}} (\nabla_\lambda f(x, z|\lambda) + \nabla_\lambda \log q(z|\lambda))q(z|\lambda)\mathrm{d}z \\ &= \mathbb{E}_{q(z|\lambda)}[\nabla_\lambda f(x, z|\lambda) + \nabla_\lambda \log q(z|\lambda)].\end{aligned}$$

Here, a practical problem is the high variance of a corresponding gradient estimator. One approach to tackle this, is the reparametrization trick. Instead of $z \sim p_z = q(z|\lambda)$ we consider $z = h(\epsilon, \lambda)$ with $\epsilon \sim p_\epsilon$, which does not depend on λ. We then have

$$F_\epsilon(x) = \mathbb{P}(\epsilon \leq x) = \mathbb{P}(h(\epsilon) \leq h(x)) = \int\limits_{-\infty}^{h(x)} p_z(z)\mathrm{d}z$$

and with the chain rule

$$p_\epsilon(x) = F'_\epsilon(x) = p_z(h(x))h'(x).$$

We therefore get

$$\mathbb{E}_{p_z}[f(z)] = \int\limits_{\mathcal{X}} f(x)p_z(x)\mathrm{d}x = \int\limits_{h^{-1}(\mathcal{X})} f(h(x))p_z(h(x))\,|h'(x)|\,\mathrm{d}x = \int\limits_{h^{-1}(\mathcal{X})} f(h(x))p_\epsilon(x)\mathrm{d}x = \mathbb{E}_{p_\epsilon}[f(h(\epsilon))]$$

and we can do

$$\begin{aligned}\nabla_\lambda \mathcal{L}(q(z|\lambda)) &= \nabla_\lambda \mathbb{E}_{q(z|\lambda)}[f(x, z|\lambda)] \\ &= \mathbb{E}_{p_\epsilon}[\nabla_\lambda f(x, h(\epsilon)|\lambda)],\end{aligned}$$

i.e. we can sample from p_ϵ, which does not depend on λ.

Remark 4.13. *We can now combine Bayesian inference with complicated models, such as neural networks. Consider for instance so called variational autoencoders with an encoder $p_{W_1}(z|x)$ (in analogy to $q(z|\lambda)$) and a decoder $p_{W_2}(x|z)$, where W_1 and W_2 are the weights of a neural network and z is the "representation" of the data*[12] *(see e.g. [19]).*

4.1 A unifying framework for implicit generative models

Assume the data $X_1, \ldots, X_n \in \mathcal{X}$ are generated from a distribution $\mathbb{P}$. The goal is to find a model distribution $\mathbb{Q}$ that is "similar" to $\mathbb{P}$. The procedure is the following: We sample Z from a latent space $\mathcal{Z}$ according to $\mathbb{Q}_Z$ (or its density q_Z, which can be seen in analogy to the prior in theorem 4.1) and map each $Z \in \mathcal{Z}$ to the space $\mathcal{X}$ by a (possibly random) mapping from $\mathcal{Z}$ to $\mathcal{X}$. This results in latent variable models $\mathbb{Q}$ with a density of the form

$$q(x) := \int\limits_{\mathcal{Z}} q(x|z)q_Z(z)\mathrm{d}z, \tag{1}$$

where the conditional density $q(x|z)$ incorporates the randomness[13].

Remark 4.14. *We can alternatively consider a deterministic mapping $T : \mathcal{Z} \to \mathcal{X}$, where we sample $z \in \mathcal{Z}$ from $\mathbb{Q}_Z$, and get $x = T(z)$ s.t. $x \sim \mathbb{Q}$.*

Example 4.15. *Consider for instance the case $q(x|z) = \mathcal{N}(x|\mu(z), \Sigma(z))$ and a discrete space $\mathcal{Z}$. This corresponds to a mixture of Gaussians and we can understand the random mapping T as a map from $\mathcal{Z}$ to probability distributions over $\mathcal{X}$, from which we can again sample. In applications $T(z)$, $\mu(z)$ and $\Sigma(z)$ are often neural networks (i.e. flexible functions) and the task it to tune their weights s.t. the distance between $\mathbb{P}$ and $\mathbb{Q}$ gets minimized. Note that usually $q(x|z)$ has no analytical expression, but it is easy to sample from, and as before we usually cannot compute the integral analytically, we but need to sample.*

Theorem 4.16 (Variational auto-encoder)**.** *Take any conditional distribution $\tilde{\mathbb{Q}}(Z|X)$ with corresponding density $\tilde{q}(z|x)$. Then the minimization*

$$\inf_{\mathbb{Q}} \inf_{\tilde{\mathbb{Q}}} \left\{ \mathbb{E}_{\mathbb{P}}[\mathrm{KL}(\tilde{\mathbb{Q}}(Z|X) \| \mathbb{Q}_Z) - \mathbb{E}_{\tilde{\mathbb{Q}}(Z|X)}[\log q(x|z)]] \right\}$$

is comparable to $\inf_{\mathbb{Q}} \mathrm{KL}(\mathbb{P}\|\mathbb{Q})$.

[12] From a coding theory perspective, the unobserved variables z have an interpretation as a latent representation or code.

[13] One can also define the quantity $q(z) = \int_{\mathcal{X}} q(z|x)p(x)\mathrm{d}x$, which is sometimes called "aggregated posterior".

Proof. We have

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}}\left[\mathrm{KL}(\tilde{q}(z|x)\|q_Z(z)) - \mathbb{E}_{\tilde{\mathbb{Q}}}[\log q(x|z)]\right] &\geq \mathbb{E}_{\mathbb{P}}\left[\mathrm{KL}(\tilde{q}(z|x)\|q_Z(z)) - \mathbb{E}_{\tilde{\mathbb{Q}}}[\log q(x|z)]\right] - \mathbb{E}_{\mathbb{P}}\left[\mathrm{KL}(\tilde{q}(z|x)\|q(z|x))\right] \\
&= \mathbb{E}_{x\sim\mathbb{P}}\left[\mathbb{E}_{z\sim\tilde{\mathbb{Q}}}\left[\log\frac{\tilde{q}(z|x)}{q_Z(z)}\right] - \mathbb{E}_{z\sim\tilde{\mathbb{Q}}}\left[\log q(x|z)\right]\right] - \mathbb{E}_{x\sim\mathbb{P}}\left[\mathbb{E}_{z\sim\tilde{\mathbb{Q}}}\left[\log\frac{\tilde{q}(z|x)}{q(z|x)}\right]\right] \\
&= \mathbb{E}_{x\sim\mathbb{P},z\sim\tilde{\mathbb{Q}}}\left[\log\frac{\tilde{q}(z|x)}{q_Z(z)} - \log q(x|z) - \log\frac{\tilde{q}(z|x)}{q(z|x)}\right] \\
&= -\mathbb{E}_{x\sim\mathbb{P},z\sim\tilde{\mathbb{Q}}}\left[\log\frac{q_Z(z)q(x|z)}{q(z|x)}\right] \\
&= -\mathbb{E}_{x\sim\mathbb{P},z\sim\tilde{\mathbb{Q}}}[\log q(x)] \\
&= -\mathbb{E}_{x\sim\mathbb{P}}[\log q(x)],
\end{aligned}$$

where we used Bayes' theorem and where the last expression does not depend on $\tilde{\mathbb{Q}}$ anymore (this is just the expectation of the ELBO). Note that minimizing $-\mathbb{E}_{x\sim\mathbb{P}}[\log q(x)]$ w.r.t q is the same as minimizing $\mathrm{KL}(\mathbb{P}\|\mathbb{Q})$ w.r.t. $\mathbb{Q}$. By "comparable" we mean that we actually minimize an upper bound. If $\tilde{\mathbb{Q}}$ is not restricted we can just choose $\tilde{\mathbb{Q}} = \mathbb{Q}$ and have

$$-\mathbb{E}_{\mathbb{P}}[\log q(x)] = \mathbb{E}_{\mathbb{P}}\left[\mathrm{KL}(\tilde{q}(z|x)\|q_Z(z)) - \mathbb{E}_{\tilde{\mathbb{Q}}}[\log q(x|z)]\right],$$

so that both minimization problems become equivalent. $q(z|x)$ can be understood is the intractable true posterior. □

Remark 4.17. *Note that often variational autoencoders just minimize the upper bound (as for instance in [19] – read more in [7]). $\tilde{q}(z|x)$ can be considered as encoder, $q(x|z)$ as decoder. Choices of the models can be $q(x|z) = \mathcal{N}(x; T(z), \sigma^2\mathbb{1}), q_Z(z) = \mathcal{N}(z; 0, \mathbb{1})$ and $q(z|x) = \mathcal{N}(z; \mu(x), \Sigma(x))$, where T, μ and Σ are neural networks. The minimization can then be done with gradient descent in the weights of the networks. Note that this min min problem is expected to be more stable than the min max problems that will appear later.*

In general, when considering distances between $\mathbb{P}$ and $\mathbb{Q}$ we differentiate between different distance measures as well as between primal and corresponding dual formulations. Those different approaches naturally lead to different numerical methods. Let us start by looking at one broad class of distance measures.

Definition 4.18 (f-divergence)**.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability measures with corresponding densities p and q. We define the f-divergence as*

$$D_f(\mathbb{P}\|\mathbb{Q}) = \int_{\mathcal{X}} f\left(\frac{p(x)}{q(x)}\right) q(x)\mathrm{d}x,$$

where $f : \mathbb{R}_+ \to \mathbb{R}$ is a convex, lower-semicontinuous function satisfying $f(1) = 0$.

Remark 4.19. *Taking $f(x) = x\log x$ yields the KL divergence and $f(x) = \frac{1}{2}|x-1|$ the total variation distance [23].*

Remark 4.20. *One can show $D_f(\mathbb{P}\|\mathbb{Q}) = D_g(\mathbb{P}\|\mathbb{Q})$ for any $g(x) = f(x) + \lambda(x-1), \lambda \in \mathbb{R}$.*

Theorem 4.21 (Variational dual representation of f-divergence)**.** *It holds*

$$D_f(\mathbb{P}\|\mathbb{Q}) = \sup_{U\in\mathcal{F}}\left\{\mathbb{E}_{\mathbb{P}}[U(X)] - \mathbb{E}_{\mathbb{Q}}[f^*(U(X))]\right\},$$

where $f^*(x) = \sup_{u\in\mathcal{X}}\{xu - f(u)\}$ and $U : \mathcal{X} \to \mathbb{R}$ is any function.

Proof.

$$\begin{aligned}
D_f(\mathbb{P}\|\mathbb{Q}) &= \int_{\mathcal{X}} f\left(\frac{p(x)}{q(x)}\right) q(x)\mathrm{d}x \\
&= \int_{\mathcal{X}} \sup_{u\in\mathcal{X}}\left\{\frac{p(x)}{q(x)}u - f^*(u)\right\} q(x)\mathrm{d}x \\
&= \sup_{U:\mathcal{X}\to\mathbb{R}} \int_{\mathcal{X}} \left(U(x)p(x) - f^*(U(x))q(x)\right)\mathrm{d}x,
\end{aligned}$$

since we have $f(x) = \sup_{u\in\mathbb{R}}\{xu - f^*(u)\}$, as f is convex and closed (see [25] for further details). The supremum in the second line is done for every x independently, which is why we consider the supremum over functions of x when taking it out of the integral. □

Remark 4.22. *If the function class out of which we take $U : \mathcal{X} \to \mathbb{R}$ is not rich enough we only have a lower bound and not equality.*

Remark 4.23. *Note the following variational representations of the KL divergence: For any $\phi \in L^\infty(\mathbb{Q})$ we have*

$$\mathrm{KL}(\mathbb{P} \,\|\, \mathbb{Q}) \geq \mathbb{E}_{\mathbb{P}}[\phi] - \log \mathbb{E}_{\mathbb{Q}}[e^{\phi}] \geq \mathbb{E}_{\mathbb{P}}[\phi] - \mathbb{E}_{\mathbb{Q}}[e^{\phi} + 1].$$

We have equality when taking the supremum over ϕ [11, 28]. This shows there are multiple variational representations and one should ask which one is the tightest. For the optimal variation one finds

$$D_f(\mathbb{P}\|\mathbb{Q}) = \sup_{U:\mathcal{X}\to\mathbb{R},\lambda\in\mathbb{R}} \left\{ \int_{\mathcal{X}} U(x)p(x)\mathrm{d}x - \int_{\mathcal{X}} (f^*)^{\dagger}(U(x)+\lambda)q(x)\mathrm{d}x + \lambda \right\}$$

with $f^{\dagger}(x) := \inf_{t\geq 0} f(x+1)$.

We realize that taking particular choices of f-divergences and considering different optimization approaches leads to different so called "generative adversarial networks". The minimization of the variational dual representation of f-divergences reads, when again considering the codes Z,

$$\inf_{\mathbb{Q}} D_f(\mathbb{P}\|\mathbb{Q}) = \inf_{T\in\mathcal{T}} \sup_{U\in\mathcal{U}} \left\{ \mathbb{E}_{\mathbb{P}}[U(X)] - \mathbb{E}_{\mathbb{Q}_Z}[f^*(U(T(Z)))] \right\},$$

i.e. we minimize over the (random) mappings $T \in \mathcal{T}$ as defined before. In practice we estimate the expectations with sample means

$$\inf_{T\in\mathcal{T}} \sup_{U\in\mathcal{U}} \left\{ \frac{1}{N}\sum_{i=1}^{N} U(X_i) - \frac{1}{M}\sum_{j=1}^{M} f^*(U(T(Z_j))) \right\}$$

and parametrize $U = U_\omega$ and $T = T_\theta$ with flexible functions (e.g. neural networks) and optimize in the parameters ω and θ with stochastic gradient descent. Note that this optimization problem becomes difficult numerically and that (relating to statistical learning theory) it is not yet clear why choosing certain classes $\mathcal{T}$ and $\mathcal{U}$ works well in terms of generalization.

When using a particular f, namely $f(x) = -(x+1)\log\frac{x+1}{2} + x\log x$, we recover the Jensen-Shannon divergence.

Definition 4.24 (Jensen-Shannon divergence)**.** *For two probability measures $\mathbb{P}$ and $\mathbb{Q}$ the Jensen-Shannon divergence is defined as*

$$\mathrm{JS}(\mathbb{P}\|\mathbb{Q}) = \frac{1}{2}\mathrm{KL}\left(\mathbb{P}\,\middle\|\,\frac{\mathbb{P}+\mathbb{Q}}{2}\right) + \frac{1}{2}\mathrm{KL}\left(\mathbb{Q}\,\middle\|\,\frac{\mathbb{P}+\mathbb{Q}}{2}\right).$$

Remark 4.25. *The Jensen-Shannon divergence is symmetric and $\sqrt{\mathrm{JS}(\mathbb{P}\|\mathbb{Q})}$ satisfies the triangle inequality. We have $0 \leq \mathrm{JS}(\mathbb{P}\|\mathbb{Q}) \leq \log 2$ and $\mathrm{JS}(\mathbb{P}\|\mathbb{Q}) \leq \frac{1}{2}\sqrt{\mathrm{KL}(\mathbb{P}\|\mathbb{Q})}$.*

Taking the above specified f we get $f^*(x) = -\log(2 - e^x)$. Let us take $U = g_f \circ V$, with $g_f(v) = \log 2 - \log(1 - e^v)$, then $\inf_{\mathbb{Q}} D_f(\mathbb{P}\|\mathbb{Q})$ is equivalent to

$$\inf_T \sup_V \left\{ \mathbb{E}_{\mathbb{P}}\left[\log\frac{1}{1+e^{-V(X)}}\right] + \mathbb{E}_{\mathbb{Q}_Z}\left[\log\left(1 - \frac{1}{1+e^{-V(T(Z))}}\right)\right] \right\} + 2\log 2,$$

which can be compared to the original attempt of so called "generative adversarial networks" (GANs), which considers the optimization problem

$$\inf_T \sup_V \left\{ \mathbb{E}\left[\log V(X)\right] + \mathbb{E}\left[\log\left(1 - V(T(Z))\right)\right] \right\},$$

where V is called the "discriminator" and T the "generator" and we have an interpretation of a minimax two-player game [15]. In the original paper it is already argued that $D_{\mathrm{GAN}}(\mathbb{P},\mathbb{Q}) \leq 2\,\mathrm{JS}(\mathbb{P}\,\|\,\mathbb{Q}) - \log(4)$. More details can be found in [26].

Remark 4.26. *Practical issues are that $\mathbb{P}$ and $\mathbb{Q}$ could be not supported on the same manifold and that $f-$divergences seem to be too strong. Therefore it makes sense to also consider weaker notions of distance measures*[14].

One other class of distance measures is the following.

Definition 4.27 (Optimal transport)**.** *We consider a cost $c(x,y) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$. We define the Wasserstein distance between two probability measures $\mathbb{P}$ and $\mathbb{Q}$ as*

$$W_c(\mathbb{P},\mathbb{Q}) = \inf_{\Gamma \in \mathcal{P}(X\sim\mathbb{P}, Y\sim\mathbb{Q})} \mathbb{E}_{(X,Y)\sim\Gamma}[c(X,Y)],$$

where $\mathcal{P}(X \sim \mathbb{P}, Y \sim \mathbb{Q})$ is a set of all joint distributions of (X,Y) with marginals $\mathbb{P}$ and $\mathbb{Q}$ respectively.

Remark 4.28. *Taking a metric $d(x,y)$ (usually $\|x-y\|$) for $c(x,y)$ leads to the Wasserstein-1 distance, which we call W_1.*

Example 4.29. *Consider $Z \sim \mathcal{U}([0,1])$ and look at the random variables $(0,Z)$ prescribed by $\mathbb{P}_0$ and (θ,Z) prescribed by $\mathbb{P}_\theta$. Then*

$$\begin{aligned}
&W_1(\mathbb{P}_0,\mathbb{P}_\theta) = |\theta| \\
&\mathrm{JS}(\mathbb{P}_0 \,\|\, \mathbb{P}_\theta) = \begin{cases} \log 2, & \theta \neq 0 \\ 0, & \theta = 0 \end{cases} \\
&\mathrm{KL}(\mathbb{P}_0 \,\|\, \mathbb{P}_\theta) = \mathrm{KL}(\mathbb{P}_\theta \,\|\, \mathbb{P}_0) = \begin{cases} \infty, & \theta \neq 0 \\ 0, & \theta = 0 \end{cases} \\
&\mathrm{TV}(\mathbb{P}_0,\mathbb{P}_\theta) = \begin{cases} 1, & \theta \neq 0 \\ 0, & \theta = 0 \end{cases},
\end{aligned}$$

i.e. only W_1 is continuous in θ and provides reasonable gradients w.r.t. θ. This example can be generalized (see theorem 1 in [2]).

Theorem 4.30. *Let $\mathbb{P}$ be a distribution on a compact space $\mathcal{X}$ and $(\mathbb{P}_n)_{(n\in\mathbb{N})}$ be a sequence of distributions on $\mathcal{X}$. Then the following are equivalent:*

(i) $W_1(\mathbb{P}_n,\mathbb{P}) \to 0$ *as* $n \to \infty$.

(ii) $\mathbb{P}_n \to \mathbb{P}$ *in distribution for* $n \to \infty$.

Proof. See theorem 2 in [2]. □

Theorem 4.31 (Kantorovich-Rubinstein duality)**.** *We have*

$$W_1(\mathbb{P},\mathbb{Q}) = \sup_{U\in\mathcal{F}_L} \left|\mathbb{E}_{\mathbb{P}}[U(X)] - \mathbb{E}_{\mathbb{Q}}[U(X)]\right|,$$

where $\mathcal{F}_L$ are all bounded 1-Lipschitz functions on $(\mathcal{X},d)$.

Proof. See theorem 5.10 in [33]. □

We consider the randomness of $\mathbb{Q}$ just as described in the beginning of section 4.1 and can write $\inf_{\mathbb{Q}} W_1(\mathbb{P},\mathbb{Q})$ as

$$\inf_T \sup_U \left\{\mathbb{E}_{\mathbb{P}}[U(X)] - \mathbb{E}_{\mathbb{Q}_Z}[U(T(Z))]\right\},$$

which can be compared to the objective we had before. A corresponding algorithm, which essentially uses $\nabla_\theta W_1(\mathbb{P},\mathbb{Q}_\theta) = -\mathbb{E}_{\mathbb{Q}_Z}[\nabla_\theta U_\omega(T_\theta(Z))]$, is proposed in [2].

Remark 4.32. *There is a dual version for other Wasserstein distances as well, but they do not fit to the GAN framework.*

Another approach is to try to minimize the primal Wasserstein distance directly.

[14] Informally, a distance induces a weaker topology when it makes it easier for a sequence of distributions to converge.

Theorem 4.33. *For any function $T : \mathcal{Z} \to \mathcal{X}$ we have*

$$W_c(\mathbb{P}, \mathbb{Q}) = \inf_{\mathbb{Q}(Z|X) : \tilde{\mathbb{Q}} = \mathbb{Q}_Z} \mathbb{E}_{\mathbb{P}} \left[\mathbb{E}_{\mathbb{Q}(Z|X)} \left[c(X, T(Z)) \right] \right],$$

where $\tilde{\mathbb{Q}} = \mathbb{E}_{\mathbb{P}}[q(z|x)]$ is the marginal distribution of Z if $X \sim \mathbb{P}$ and $Z \sim \mathbb{Q}(Z|X)$.

Proof. See [7]. □

Remark 4.34. *Numerically solving this constrained minimization problem is hard. One can for instance consider the relaxed problem*

$$\inf_{\mathbb{Q}(Z|X) \in \mathcal{Q}} \mathbb{E}_{\mathbb{P}} \left[\mathbb{E}_{\mathbb{Q}(Z|X)} \left[c(X, T(Z)) \right] \right] + \lambda D(\tilde{\mathbb{Q}}, \mathbb{Q}_Z),$$

with $\lambda > 0$ and a distance D.

Finally, we consider the third class of probability distances.

Definition 4.35 (Integral probability metrics)**.** *Let $\mathcal{F}$ be a class of bounded real-valued functions. We define the metric*

$$\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) := \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{\mathbb{P}}[f(X)] - \mathbb{E}_{\mathbb{Q}}[f(X)] \right|.$$

Remark 4.36. *The described distances are not exclusive. The total variation distance for instance is both an f-divergence and an integral probability metric. By the Kantorovich-Rubinstein duality we know that $W_1(\mathbb{P}, \mathbb{Q}) = \gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})$ if $\mathcal{F}$ is the set of* 1*-Lipschitz functions. When $\mathcal{F}$ is the set of all measurable functions bounded between* -1 *and* 1*, we have $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = \mathrm{TV}(\mathbb{P}, \mathbb{Q})$.*

Definition 4.37 (Maximum mean discrepancy)**.** *Consider a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and its RKHS $\mathcal{H}$. Let $\mathcal{F} = \{f \in \mathcal{H} : \|f\|_\infty \leq 1\}$. The maximum mean discrepancy is defined as*

$$\gamma_k(\mathbb{P}, \mathbb{Q}) := \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{\mathbb{P}}[f(X)] - \mathbb{E}_{\mathbb{Q}}[f(X)] \right|.$$

Remark 4.38. *The maximum mean discrepancy is a special case of an integral probability metric. See for instance proposition 11.3.2 and the following in [13] for further details.*

5 Bibliography

[1] Anthony, M. and Bartlett, P. L. (1999). *Neural network learning: Theoretical foundations*. Cambridge university press.

[2] Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein gan. *arXiv preprint arXiv:1701.07875*.

[3] Bartlett, P. L. and Traskin, M. (2007). Adaboost is consistent. *Journal of Machine Learning Research*, 8(Oct):2347–2368.

[4] Bölcskei, H., Grohs, P., Kutyniok, G., and Petersen, P. (2017). Optimal approximation with sparsely connected deep neural networks. *arXiv preprint arXiv:1705.01714*.

[5] Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

[6] Bousquet, O., Boucheron, S., and Lugosi, G. (2004). Introduction to statistical learning theory. pages 169–207.

[7] Bousquet, O., Gelly, S., Tolstikhin, I., Simon-Gabriel, C.-J., and Schoelkopf, B. (2017). From optimal transport to generative modeling: the vegan cookbook. *arXiv preprint arXiv:1705.07642*.

[8] Cheney, E. W. (1966). *Introduction to approximation theory*. McGraw-Hill.

[9] Cucker, F. and Zhou, D. X. (2007). *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge University Press.

[10] Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314.

[11] Dai Pra, P., Meneghini, L., and Runggaldier, W. J. (1996). Connections between stochastic control and dynamic games. *Mathematics of Control, Signals, and Systems (MCSS)*, 9(4):303–326.

[12] Devroye, L. and Lugosi, G. (2001). *Combinatorial methods in density estimation*. Springer, New-York.

[13] Dudley, R. M. (2002). *Real analysis and probability*, volume 74. Cambridge University Press.

[14] Georgii, H.-O. (2015). *Stochastik: Einführung in die Wahrscheinlichkeitstheorie und Statistik*. Walter de Gruyter GmbH & Co KG.

[15] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, pages 2672–2680.

[16] Hartmann, C. (2017). *Wahrscheinlichkeitstheorie*. Lecture script, BTU Cottbus-Senftenberg.

[17] Hartmann, C., Richter, L., Schütte, C., and Zhang, W. (2017). Variational characterization of free energy: theory and algorithms. *Entropy*, 19(11):626.

[18] Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.

[19] Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

[20] Koltchinskii, V. and Panchenko, D. (2002). Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, pages 1–50.

[21] König, H. (2013). *Eigenvalue distribution of compact operators*, volume 16. Birkhäuser.

[22] Königsberger, K. (1999). *Analysis 1. Vierte Auflage*. Springer-Lehrbuch. Springer-Verlag, Berlin.

[23] Liese, F. and Vajda, I. (2006). On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412.

[24] Maiorov, V. and Pinkus, A. (1999). Lower bounds for approximation by mlp neural networks. *Neurocomputing*, 25(1-3):81–91.

[25] Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861.

[26] Nowozin, S., Cseke, B., and Tomioka, R. (2016). f-gan: Training generative neural samplers using variational divergence minimization. *Advances in Neural Information Processing Systems*, pages 271–279.

[27] Pinkus, A. (1999). Approximation theory of the mlp model in neural networks. *Acta numerica*, 8:143–195.

[28] Ruderman, A., Reid, M., García-García, D., and Petterson, J. (2012). Tighter variational representations of f-divergences via restriction to probability measures. *arXiv preprint arXiv:1206.4664*.

[29] Schapire, R. E., Freund, Y., Bartlett, P., Lee, W. S., et al. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *The annals of statistics*, 26(5):1651–1686.

[30] Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms.* Cambridge university press.

[31] Tsybakov, A. B. (2009). *Introduction to nonparametric estimation.* Springer Series in Statistics. Springer, New York.

[32] Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142.

[33] Villani, C. (2008). *Optimal transport: old and new*, volume 338. Springer Science & Business Media.

[34] Yarotsky, D. (2017). Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114.