Empirical Risk Minimization

Vincent Divol

The problem of **supervised learning** can be expressed in the following way: we are given a family of inputs $(x_1, \dots, x_n)$ on a set $\mathcal{X}$, and associated outputs $(y_1, \dots, y_n)$ on a set $\mathcal{Y}$. Given a new input $x \in \mathcal{X}$, can we predict what the associated output $y \in \mathcal{Y}$ will be? There are two large families of learning problems, the case where $\mathcal{Y}$ is a finite set (**classification task**), and the case where the ouputs y_is take continuous values, typically $\mathcal{Y} = \mathbb{R}$ (**regression task**).

Example 0.1.

1. Each input x_i is a picture that represents an animal y_i. In this case, a picture x_i is represented by the RGB value (Red, Green, Blue) of each of its pixels, so that $\mathcal{X} = \mathbb{R}^{3K}$, where K is a number of pixels. The set $\mathcal{Y}$ is the set of animals that are depicted. This is a classification task.

2. Each input x_i is a review of a movie, and y_i is the rating associated with the review. Given a new review x, the goal is to guess if the user who wrote the review liked the movie (y is high) or disliked it (y is low). The set of inputs $\mathcal{X}$ is the set of texts, whereas $y \in \mathcal{Y} = [0, 1]$ represents a grade between 0 and 1. This is a regression task.

3. Each input x_i is a patient that is described by different physiological parameters (including e.g. sex, age, blood pressure, etc.) and a treatment that was given to them, whereas y_i is equal to 1 (the patient is cured) or -1 (the patient is not cured) ($\mathcal{Y} = \{-1, 1\}$). The goal is then to understand what is the efficiency of different treatments for different profiles of patients.

Before going further, there is an important question to address, concerning the model assumptions made on the inputs $(x_1, \dots, x_n)$ and the outputs $(y_1, \dots, y_n)$. As it is most often the case in machine learning approaches, we will assume that both the inputs and the ouputs are **random**. In particular, it may be possible that for two inputs $x_i = x_j$ that are equal, the corresponding outputs y_i and y_j are different (this for instance makes sense in Example (3), where different patients having the same profile, given the same treatment, may experience different outcomes). We will moreover always assume that the pairs $((x_1, y_1), \dots, (x_n, y_n))$ are **independent and identically distributed** (i.i.d.). This is the simplest assumption, that is reasonable in a large number of cases, although there exist relevant problems where such an assumption is too strong (e.g. if the observations (x_i, y_i) arrive in a sequential manner, it may then be the case that the law of the variable (x_i, y_i) depends on the time i at which it was observed).

To distinguish between deterministic variables and random ones, we will use a **bold font** to refer to the latter, that is $(\mathbf{x}, \mathbf{y})$ is random, whereas (x, y) is deterministic. The law of the observations $(\mathbf{x_i}, \mathbf{y_i})$ will be denoted by P, whereas $P_{\mathbf{x}}$ is the law of the first marginal $\mathbf{x_i}$ and $P_{\mathbf{y}}$ is the law of the second marginal $\mathbf{y_i}$. Remark that P is a probability measure on the space $\mathcal{X} \times \mathcal{Y}$, whereas $P_{\mathbf{x}}$ is a probability measure on $\mathcal{X}$ and $P_{\mathbf{y}}$ is a probability measure on $\mathcal{Y}$. We will write $\mathbb{E}_P[f(\mathbf{x}, \mathbf{y})]$ for the expectation of some function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ with respect to P, whereas the conditional expectation of $f(\mathbf{x}, \mathbf{y})$ given that $\mathbf{x} = x$ is written as $\mathbb{E}_P[f(\mathbf{x}, \mathbf{y}) | \mathbf{x} = x]$. We will sometimes only write $\mathbb{E}$ instead of $\mathbb{E}_P$ when it is clear what the underlying law is.

1 Risks and losses

How do we measure the quality of a prediction?

The term ”predicting” is here rather vague, and the data scientist may want to give it different meanings depending on the context. For instance, in binary classification ($\mathcal{Y} = \{-1, 1\}$), a possible goal is to minimize the number of misclassifications on average. However, the two ouputs -1 and 1 may not play a symmetrical role, and we may want to favorize predictions that make very few mistakes when choosing -1, at the price of making more mistakes when choosing 1. For example, in medical settings, the output $y = -1$ can represent the fact that the patient x is sick and deserves further treatment, while $y = 1$ means that the patient x is healthy. Then, it is a more serious mistake to predict that $y = 1$ when in fact $y = -1$ than the opposite, and we want to take this into account when assessing the quality of a predictor.

The problem is even more striking when the ouput space $\mathcal{Y}$ is multidimensional, say $\mathcal{Y} = \mathbb{R}^d$. In that case, possible ways of measuring how a prediction $y' = (y'_1, \ldots, y'_d)$ is close to the output $y = (y_1, \ldots, y_d)$ include:

- the L_∞-norm: $\|y' - y\|_\infty := \max_{j=1,\ldots,d} |y'_j - y_j|$,
- the L_p-norm $\|y' - y\|_p := (\sum_{j=1}^d |y'_j - y_j|^p)^{1/p}$,
- the weighted L_p-norm $\|y' - y\|_{p,w} := (\sum_{j=1}^d w_j |y'_j - y_j|^p)^{1/p}$, where $w = (w_1, \ldots, w_d)$ is a vector of positive weights,
- the dot product $y' \cdot y = \sum_{j=1}^d y'_j y_j$.

There are no "better" choices of distances among the one listed above, they each represent a different way of measuring how two points in the output space $\mathcal{Y}$ are similar. For instance, choosing the ℓ_1-norm instead of the ℓ_2-norm indicates that we want to penalize less the fact that a huge error was made on one of the entries y'_j of the prediction, which might be a desirable feature in some problems.

More generally, we will work with a general loss function ℓ on the set of outputs.

Definition 1.1 (Loss function). *A **loss function** is a nonnegative function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, +\infty)$.*

The goal is then to find a function $f : \mathcal{X} \to \mathcal{Y}$ such that $\ell(f(\mathbf{x}), \mathbf{y})$ is small on new samples $((\mathbf{x'_1}, \mathbf{y'_1}), \ldots, (\mathbf{x'_{n'}}, \mathbf{y'_{n'}}))$, that we call the testing sample. We will here always assume that the testing sample is also i.i.d. of law P[1]. The goal is then to minimize the average loss on the testing sample, that we call the **expected risk** or the **test error**.

Definition 1.2 (Expected risk). *Let P be a probability measure on $\mathcal{X} \times \mathcal{Y}$ and ℓ be a loss function. Given a function $f : \mathcal{X} \to \mathcal{Y}$, the P-**risk of** f is given by*

$$\mathcal{R}_P(f) := \mathbb{E}_P[\ell(\mathbf{y}, f(\mathbf{x}))]. \tag{1}$$

[1] In many practical situations, the law of the testing sample is actually different from the law P on which the predictor was trained. This situation is referred to as **covariate shift** in the literature and requires the development of specific techniques. This issue will never be addressed in these notes.

The best prediction $f_P^\star$ is the one that minimizes $\mathcal{R}_P(f)$, and is called the **Bayes predictor**. It turns out that we can give an expression of the Bayes predictor.

Theorem 1.3 (Optimality of Bayes predictor)**.** *Let P be a probability measure on $\mathcal{X} \times \mathcal{Y}$. The function $f \mapsto \mathcal{R}_P(f)$ is minimized at $f_P^\star$ that is defined by*

$$f_P^\star(x) \in \operatorname*{arg\,min}_{z \in \mathcal{Y}} \mathbb{E}_P[\ell(\mathbf{y}, z)|\mathbf{x} = x]. \tag{2}$$

Proof. Let $f : \mathcal{X} \to \mathcal{Y}$ be a function. Define the function $\Psi : (x, z) \in \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{E}_P[\ell(\mathbf{y}, z)|\mathbf{x} = x]$. By definition, we have the equality $\Psi(x, f_P^\star(x)) = \min_{z \in \mathcal{Y}} \Psi(x, z)$. By the law of total expectation, we obtain

$$\mathcal{R}_P(f) = \mathbb{E}_P[\ell(\mathbf{y}, f(\mathbf{x}))] = \mathbb{E}_P[\Psi(\mathbf{x}, f(\mathbf{x}))] \geq \mathbb{E}_P[\Psi(\mathbf{x}, f_P^\star(\mathbf{x}))] = \mathcal{R}_P(f_P^\star). \tag{3}$$

As this hold for every function f, this implies the conclusion. □

Let us consider concrete examples of losses ℓ and associated Bayes predictors.

Example 1.4.

- Consider the problem of binary classification ($\mathcal{Y} = \{-1, 1\}$) with the loss $\ell(y, y') = \mathbf{1}\{y \neq y'\}$. Then, $\mathbb{E}_P[\ell(\mathbf{y}, 1)|\mathbf{x} = x] = P(\mathbf{y} = 1|\mathbf{x} = x)$. We call this quantity the **regression function** $\eta(x)$. We have $\mathbb{E}_P[\ell(\mathbf{y}, -1)|\mathbf{x} = x] = 1 - \eta(x)$. Therefore,

$$f_P^\star(x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2, \\ -1 & \text{otherwise.} \end{cases} \tag{4}$$

  In other words, the Bayes predictor follows the following intuitive rule: if the probability of observing the output $\mathbf{y} = 1$ given that $\mathbf{x} = x$ is larger than $1/2$, then we predict 1. Otherwise, we predict -1.

- Let $\mathcal{Y} = \mathbb{R}$ and let $\ell(y, y') = |y - y'|^2$. Remark that the function $z \mapsto \mathbb{E}[(\mathbf{a} - z)^2]$ is minimized at $z = \mathbb{E}[\mathbf{a}]$. This implies that the function $z \mapsto \mathbb{E}_P[\ell(\mathbf{y}, z)|\mathbf{x} = x] = \mathbb{E}_P[|\mathbf{y} - z|^2|\mathbf{x} = x]$ is minimized for $z = \mathbb{E}_P[\mathbf{y}|\mathbf{x} = x]$. Therefore, the Bayes predictor is in this case given by the conditional expectation

$$f_P^\star(x) = \mathbb{E}_P[\mathbf{y}|\mathbf{x} = x]. \tag{5}$$

2 Empirical risk

If the Bayes predictor is indeed the optimal one, it has a major drawback: computing it requires to know what the law of the sample P is! In practice, we only have access to the training sample $((\mathbf{x_1}, \mathbf{y_1}), \ldots, (\mathbf{x_n}, \mathbf{y_n}))$, and P is unknown. We cannot therefore use the Bayes predictor, and our goal will be to design predictors f that can be computed based on the observations, and that will (hopefully) behave almost as well as the Bayes predictor.

A powerful method to do so consists in minimizing the **empirical risk**.

Definition 2.1 (Empirical risk)**.** *Let $(\mathbf{x_1}, \mathbf{y_1}), \ldots, (\mathbf{x_n}, \mathbf{y_n})$ be a training sample from law P and ℓ be a loss function. The **empirical risk** of the sample is given by*

$$f \mapsto \mathcal{R}_n(f) := \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{y_i}, f(\mathbf{x_i})). \tag{6}$$

The law of large number indicates that $\mathcal{R}_n(f) \simeq \mathcal{R}_P(f)$ when n is very large. Therefore, one may expect that minimizing $\mathcal{R}_n$ is a good strategy to build a predictor with small P-risk. There is however a caveat: for most losses ℓ, one can always find **many** functions f such that $\mathcal{R}_n(f) = 0$, some of them being **very irregular**. For instance, in a regression setting with $\mathcal{X} = \mathcal{Y} = \mathbb{R}$ and $\ell(y, y') = |y - y'|$, there are infinitely many functions (continuous or discontinuous) such that $f(x_i) = y_i$, and $\mathcal{R}_n(f) = 0$ for all such functions. Such functions f will then behave badly on new observations $(\mathbf{x}, \mathbf{y})$ sampled according to P: the risk $\mathcal{R}_P(f)$ will be large although $\mathcal{R}_n(f) = 0$.

This minimizing strategy must therefore be improved. An idea consists in minimizing $\mathcal{R}_n$ over a restricted class of functions $\mathcal{F}$, that will encode some regularity that we expect the Bayes predictor to have.

Definition 2.2 (Empirical risk minimizer)**.** *Let $(\mathbf{x_1}, \mathbf{y_1}), \ldots, (\mathbf{x_n}, \mathbf{y_n})$ be a training sample from law P and ℓ be a loss function. Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $\mathcal{Y}$, that we also call a **model**. An **empirical risk minimizer** $\hat{f}_{\mathcal{F}}$ is any function in $\mathcal{F}$ that attains the minimum*

$$\min_{f \in \mathcal{F}} \mathcal{R}_n(f). \tag{7}$$

There is a crucial element that needs to be directly mentioned: the computation of the empirical risk minimizer requires the minimization of a potentially complicated functional on an arbitrary set $\mathcal{F}$. This problem is in

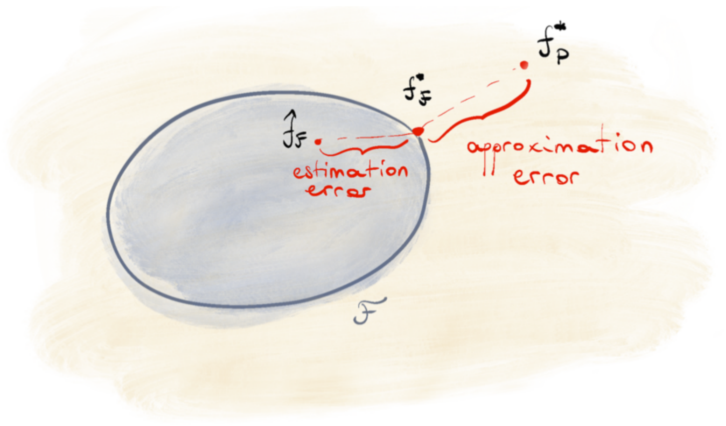

Figure 1: Decomposition of the excess risk $\mathcal{R}_P(\hat{f}_P) - \mathcal{R}(f_P^\star)$ into the approximation error $\mathcal{R}_P(f_{\mathcal{F}}^\star) - \mathcal{R}(f_P^\star)$ and the estimation error $\mathcal{R}_P(\hat{f}_{\mathcal{F}}) - \mathcal{R}(f_{\mathcal{F}}^\star)$.

general untractable, and we will address in the next chapters how to solve it in the case where the loss ℓ is convex. There are however some specific examples where no optimization procedures are required, and an explicit form of the solution exists, as in the following example.

Example 2.3 (Linear regression). Consider the regression problem on $\mathbb{R}^d$ with the quadratic loss (that is $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$, and $\ell(y, y') = |y - y'|^2$). In this setting, a popular choice is to consider the class $\mathcal{F}_{\text{lin}}$ of linear predictors of the form $f_\theta : x \mapsto \theta^T x$. The empirical risk is then given by

$$\begin{aligned}\mathcal{R}_n(f_\theta) &= \frac{1}{n}\sum_{i=1}^n |\mathbf{y_i} - f_\theta(\mathbf{x_i})|^2 = \frac{1}{n}\sum_{i=1}^n |\mathbf{y_i} - \theta^T \mathbf{x_i}|^2 \\ &= \frac{1}{n}\|\mathbf{Y} - \mathbf{X}\theta\|^2,\end{aligned}$$

where

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y_1} \\ \vdots \\ \mathbf{y_n} \end{pmatrix} \text{ and } \mathbf{X} = \begin{pmatrix} \mathbf{x_1} \\ \vdots \\ \mathbf{x_n} \end{pmatrix}.$$

In this case, the empirical risk minimizer $\hat{f}_{\mathcal{F}_{\text{lin}}}$ is given by a linear regression. We may also further restrict the model by considering only vectors θ having a small ℓ_2-norm (ridge regression) or a small ℓ_1-norm (lasso regression).

Decomposition of the empirical risk: underfitting and overfitting

Let us analyze the performance of an empirical risk minimizer. Let $\mathcal{R}_P^\star := \min_f \mathcal{R}_P(f) = \mathcal{R}_P(f_P^\star)$. We want to understand when the risk of $\hat{f}_{\mathcal{F}}$ is not much larger than $\mathcal{R}_P^\star$, that is we want to bound the **excess risk**

$$\mathcal{R}_P(\hat{f}_{\mathcal{F}}) - \mathcal{R}_P^\star. \tag{8}$$

The **excess risk** can be decomposed into

$$\mathcal{R}_P(\hat{f}_{\mathcal{F}}) - \mathcal{R}_P^\star = \underbrace{(\mathcal{R}_P(\hat{f}_{\mathcal{F}}) - \inf_{f\in\mathcal{F}} \mathcal{R}_P(f))}_{\text{estimation error}} + \underbrace{(\inf_{f\in\mathcal{F}} \mathcal{R}_P(f) - \mathcal{R}_P^*)}_{\text{approximation error}}. \tag{9}$$

Remark first that the two error terms are nonnegative.

- The **approximation error** $\inf_{f\in\mathcal{F}} \mathcal{R}_P(f) - \mathcal{R}_P^*$ is a deterministic quantity (it does not depend on the observations) that measures how far the best predictor on $\mathcal{F}$ is from the best predictor (the Bayes predictor). The larger $\mathcal{F}$ is, the smaller this error becomes.

- It is less immediate to understand how the **estimation error** $\mathcal{R}_P(\hat{f}_{\mathcal{F}}) - \inf_{f\in\mathcal{F}} \mathcal{R}_P(f)$ behaves. Assume that the infimum $\inf_{f\in\mathcal{F}} \mathcal{R}_P(f)$ is attained at some function $f_{\mathcal{F}}^\star$. We can then write

$$\begin{aligned}
&\mathcal{R}_P(\hat{f}_{\mathcal{F}}) - \inf_{f\in\mathcal{F}} \mathcal{R}_P(f) = \mathcal{R}_P(\hat{f}_{\mathcal{F}}) - \mathcal{R}_P(f_{\mathcal{F}}^\star) \\
&= (\mathcal{R}_P(\hat{f}_{\mathcal{F}}) - \mathcal{R}_n(\hat{f}_{\mathcal{F}})) + (\mathcal{R}_n(\hat{f}_{\mathcal{F}}) - \mathcal{R}_n(f_{\mathcal{F}}^\star)) + (\mathcal{R}_n(f_{\mathcal{F}}^\star) - \mathcal{R}_P(f_{\mathcal{F}}^\star)) \\
&\le (\mathcal{R}_P(\hat{f}_{\mathcal{F}}) - \mathcal{R}_n(\hat{f}_{\mathcal{F}})) + 0 + (\mathcal{R}_n(f_{\mathcal{F}}^\star) - \mathcal{R}_P(f_{\mathcal{F}}^\star)) \\
&\le \sup_{f\in\mathcal{F}} (\mathcal{R}_P(f) - \mathcal{R}_n(f)) + (\mathcal{R}_n(f_{\mathcal{F}}^\star) - \mathcal{R}_P(f_{\mathcal{F}}^\star)),
\end{aligned} \tag{10}$$

  where the first inequality follows from the fact that $\hat{f}_{\mathcal{F}}$ minimizes $\mathcal{R}_n$ on $\mathcal{F}$ by definition, so that $\mathcal{R}_n(\hat{f}_{\mathcal{F}}) \le \mathcal{R}_n(f_{\mathcal{F}}^\star)$. If we believe that this inequality is tight (and it is in many cases), then the estimation error is linked to the quantity $\sup_{f\in\mathcal{F}}(\mathcal{R}_P(f) - \mathcal{R}_n(f))$, that is the uniform deviation between the empirical risk $\mathcal{R}_n$ and its expectation $\mathcal{R}_P$ on the class $\mathcal{F}$. This quantity increases with the size of $\mathcal{F}$ and decreases with the number of observations n.

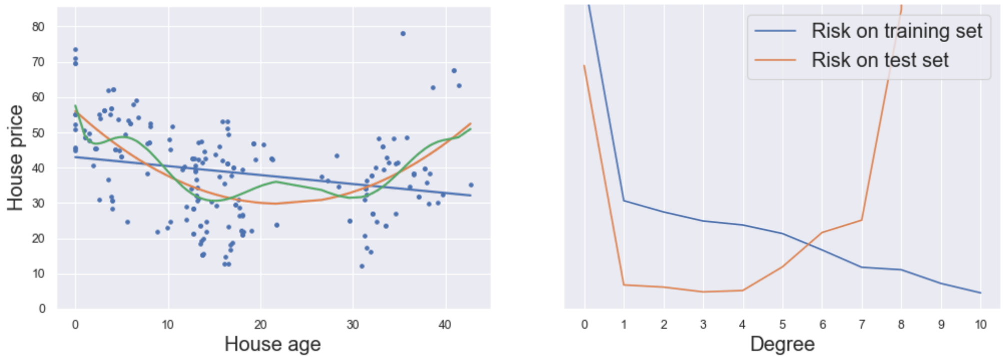

Figure 2: Left: linear predictor $\hat{f}_1$ (blue), quadratic predictor $\hat{f}_2$ (orange) and predictor of degree 10 (green). Right: Empirical risk $\mathcal{R}_n(\hat{f}_d)$ as a function of d (in blue) and average risk of $\hat{f}_d$ on the testing sample (in orange). As expected, the empirical risk $\mathcal{R}_n(\hat{f}_d)$ is a nonincreasing function of d.

Example 2.4 (Polynomial regression). Let $\mathcal{X} = [0, 1]$, $\mathcal{Y} = \mathbb{R}$ and ℓ be the square loss. Let $\mathcal{F}_d$ be the set of polynomials of degree d and let $\hat{f}_d := \hat{f}_{\mathcal{F}_d}$. We test the performance of the estimator $\hat{f}_d$ on the *Real estate valuation data set* [Yeh and Hsu, 2018] (taken from the UCI Machine Learning Repository). On this dataset, the goal is to predict the price y of a house based on several features (coordinates, house age, number of nearby convenience stores, etc.). For visualization sake, we consider a single feature x representing the house age and compute the predictors $\hat{f}_d$ for d ranging from 1 to 10 (see Figure 2). We observe the two predicted regimes. For $d = 1$, the model is too simple and both the empirical risk $\mathcal{R}_n(\hat{f}_1)$ and the risk $\mathcal{R}_P(\hat{f}_1)$ (that is approximated by the empirical risk on the testing sample) are large: the model is underfitting. For $d = 10$, the empirical risk becomes small, but the risk $\mathcal{R}_P(\hat{f}_{10})$ is really large: our model is too complicated and we are overfitting.

We have discovered a fundamental phenomenon: the excess risk of an empirical risk minimizer is driven by two contrary forces. The first one is the approximation error, that measures how far the model $\mathcal{F}$ is close from ”the truth”, and will be large if our model is overly simplistic, a regime that we call **underfitting**. The second one is the estimation error, that measures how the set of observations $(\mathbf{y_1}, f(\mathbf{x_1})), \ldots, (\mathbf{y_n}, f(\mathbf{x_n}))$ is able to capture the behavior of the expectation $\mathbb{E}_P[\ell(\mathbf{y}, f(\mathbf{x}))]$ over all functions $\mathcal{F}$. If $\mathcal{F}$ is very large, then there will typically be many different functions with very small empirical risk (as in Example 2.4), so that the minimizer $\hat{f}_P$ might be very different from $f_P^\star$. We call this regime **overfitting**.

3 Bound on the estimation error in binary classification

We focus in this section on the classification task $\mathcal{Y} = \{-1, 1\}$ with the $0-1$ loss $\ell(y, y') = \mathbf{1}\{y \neq y'\}$. Our aim is to understand how the estimation error $\mathcal{R}_P(\hat{f}_{\mathcal{F}}) - \inf_{f \in \mathcal{F}} \mathcal{R}_P(f)$ scales with $\mathcal{F}$. We first consider the case where the set $\mathcal{F}$ of predictors is finite, and then consider the more delicate case of infinite classes of predictors $\mathcal{F}$ by introducing the concept of VC dimension.

3.1 Finite number of predictors

Assume first that the set $\mathcal{F} = \{f_1, \ldots, f_k\}$ is finite. The estimation error is bounded using this general inequality.

Theorem 3.1 (Maximal inequality)**.** *Let $\mathbf{z_1}, \ldots, \mathbf{z_k}$ be real valued random variables such that there exists a constant $\sigma > 0$ with $\mathbb{E}[e^{\lambda \mathbf{z_j}}] \leq e^{\lambda^2 \sigma^2/2}$ for every $\lambda > 0$. Then,*

$$\mathbb{E}[\max_{j=1,\ldots,k} \mathbf{z_j}] \leq \sigma \sqrt{2 \log k}. \tag{11}$$

Proof. A first (naive) idea to bound the maximum of a collection of numbers (a_j) consists in using that

$$\max_{j=1,\ldots,k} a_j \leq \sum_{j=1}^{k} a_j. \tag{12}$$

Of course, this bound is often very bad. However, using (12) makes sense if the maximum of the a_js (say a_{j_0}) is much larger than the other ones: in that case, the sum is roughly equal to the max.

We will enforce this situation by replacing each a_j by $\exp(\lambda a_j)$ for some parameter $\lambda > 0$. If λ is very large, then indeed $\exp(\lambda a_{j_0})$ is much larger than the other values $\exp(\lambda a_j)$, and therefore the bound

$$\max_{j=1,\ldots,k} e^{\lambda a_j} \leq \sum_{j=1}^{k} e^{\lambda a_j} \tag{13}$$

becomes a much more reasonable one. Another way of writing this equation is the following:

$$\max_{j=1\ldots k} a_j \leq \frac{1}{\lambda} \log \left(\sum_{j=1}^{k} e^{\lambda a_j} \right). \tag{14}$$

Let us now fix $a_j = \mathbf{z_j}$. We obtain

$$\mathbb{E}[\max_{j=1\ldots k} \mathbf{z_j}] \leq \mathbb{E}\left[\frac{1}{\lambda} \log \left(\sum_{j=1}^{k} e^{\lambda \mathbf{z_j}} \right) \right]. \tag{15}$$

We are now in position to use Jensen's inequality.

Lemma 3.2 (Jensen's inequality)**.** *Let* $\mathbf{x}$ *be a real valued random variable and* $\varphi : \mathbb{R} \to \mathbb{R}$ *be a convex function. Then,*

$$\varphi(\mathbb{E}[\mathbf{x}]) \leq \mathbb{E}[\varphi(\mathbf{x})]. \tag{16}$$

Applying Jensen's inequality to $\varphi = \exp$, we obtain

$$\mathbb{E}[\max_{j=1\ldots k} \mathbf{z_j}] \leq \frac{1}{\lambda} \log \left(\mathbb{E}\left[\sum_{j=1}^{k} e^{\lambda \mathbf{z_j}} \right] \right). \tag{17}$$

The assumption $\mathbb{E}[e^{\lambda \mathbf{z_j}}] \leq e^{\lambda^2 \sigma^2 / 2}$ yields

$$\mathbb{E}[\max_{j=1\ldots k} \mathbf{z_j}] \leq \frac{\log k}{\lambda} + \frac{\lambda \sigma^2}{2}. \tag{18}$$

We choose $\lambda > 0$ to minimize this expression: the optimal value is $\lambda = \sqrt{2 \log k}/\sigma$ and we obtain the final bound. □

Let us now turn to the quantity $\mathbb{E}[\mathcal{R}_P(\hat{f}_{\mathcal{F}}) - \inf_{f \in \mathcal{F}} \mathcal{R}_P(f)]$. According to the inequality (10), it is enough to bound

$$\begin{aligned}
&\mathbb{E}\left[\sup_{f \in \mathcal{F}} (\mathcal{R}_P(f) - \mathcal{R}_n(f)) + (\mathcal{R}_n(f^\star_{\mathcal{F}}) - \mathcal{R}_P(f^\star_{\mathcal{F}})) \right] \\
&= \mathbb{E}[\max_{j=1,\ldots,k} (\mathcal{R}_P(f_j) - \mathcal{R}_n(f_j))] + \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{\mathbf{y_i} \neq f^\star_P(\mathbf{x_i})\} \right] - \mathbb{P}(\mathbf{y} \neq f^\star_P(\mathbf{x})) \\
&= \mathbb{E}[\max_{j=1,\ldots,k} (\mathcal{R}_P(f_j) - \mathcal{R}_n(f_j))].
\end{aligned}$$

Let us apply the maximal inequality to the random variables $\mathbf{z_j} = \mathcal{R}_P(f_j) - \mathcal{R}_n(f_j)$ for $j = 1, \dots, k$. We have

$$\mathcal{R}_P(f_j) - \mathcal{R}_n(f_j) = \mathbb{P}(f_j(\mathbf{x}) = \mathbf{y}) - \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{f_j(\mathbf{x_i}) = \mathbf{y_i}\}. \tag{19}$$

The independence of the observations $(\mathbf{x_i}, \mathbf{y_i})$ yields

$$\mathbb{E}[e^{\lambda(\mathcal{R}_P(f_j) - \mathcal{R}_n(f_j))}] = \prod_{i=1}^{n} \mathbb{E}[e^{\frac{\lambda}{n}(\mathbb{P}(f_j(\mathbf{x})=\mathbf{y}) - \mathbf{1}\{f_j(\mathbf{x_i})=\mathbf{y_i}\})}]. \tag{20}$$

Let $p_j = \mathbb{P}(f_j(\mathbf{x_i}) = \mathbf{y_i})$. Then,

$$\mathbb{E}[e^{\frac{\lambda}{n}(p_j - \mathbf{1}\{f_j(\mathbf{x_i})=\mathbf{y_i}\})}] = p_j e^{-\frac{\lambda}{n}(1-p_j)} + (1-p_j)e^{\frac{\lambda}{n}p_j}. \tag{21}$$

The maximum of this quantity is obtained at $p_j = 1/2$, so that

$$\mathbb{E}[e^{\frac{\lambda}{n}(p_j - \mathbf{1}\{f_j(\mathbf{x_i})=\mathbf{y_i}\})}] \leq \frac{e^{\frac{\lambda}{n}} + e^{-\frac{\lambda}{n}}}{2} \leq e^{\lambda^2/(2n^2)}, \tag{22}$$

where we use the standard inequality $e^{\lambda} + e^{-\lambda} \leq 2e^{\lambda^2/2}$. Therefore, the random variable $\mathbf{z_j}$s satisfy the condition of Theorem 3.1 with $\sigma = 1/\sqrt{n}$. We thus obtain the following result.

Theorem 3.3 (Error bound in expectation on the estimation error in binary classification: finite case)**.** *Assume that $\mathcal{F}$ contains k elements and that ℓ is the $0-1$ loss. Then*

$$\mathbb{E}[\mathcal{R}_P(\hat{f}_{\mathcal{F}}) - \inf_{f\in\mathcal{F}} \mathcal{R}_P(f)] \leq \mathbb{E}[\sup_{f\in\mathcal{F}}(\mathcal{R}_P(f) - \mathcal{R}_n(f))] \leq \sqrt{\frac{2\log k}{n}}. \tag{23}$$

By the central limit theorem, we expect the fluctuations of $\mathcal{R}_n(f)$ around $\mathcal{R}_P(f)$ to be of order $1/\sqrt{n}$. Theorem 3.3 asserts that the uniform deviations of $\mathcal{R}_n(f)$ over a family of k functions f are also of order $1/\sqrt{n}$.

3.2 Vapnik-Chervonenkis dimension

The left hand side of Theorem 3.3 diverges as the size k of the class $\mathcal{F}$ of predictors grow (although at a slow $\sqrt{\log k}$ rate). Does this mean that

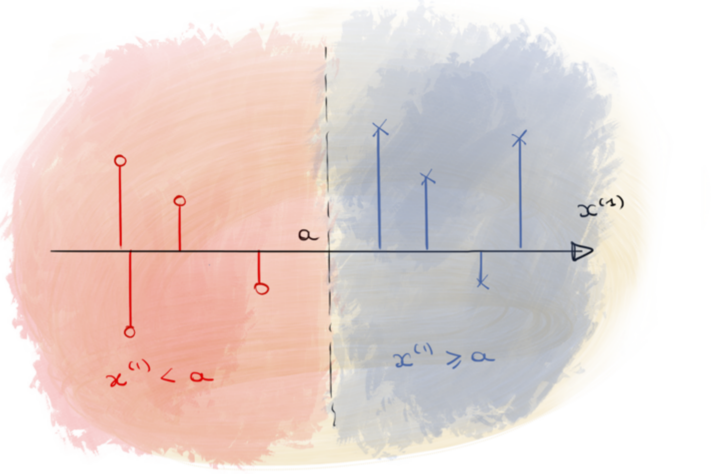

Figure 3: Given n points $x_1, \ldots, x_n$ in $\mathbb{R}^d$, the number $\mathcal{N}_{\mathcal{F}_0}(x_1, \ldots, x_n)$ of classifications using a classifier $f \in \mathcal{F}_0$ is at most $n+1$. Indeed, if we order the points such that their first coordinates are in increasing order $x_1^{(1)} \leq x_2^{(1)} \leq \cdots \leq x_n^{(1)}$, then choosing a classifier in $\mathcal{F}_0$ amounts to choosing the largest index which will be classified as -1, and there are $n+1$ such indices.

everything is hopeless when $\mathcal{F}$ is infinite and that we should stick with a finite set $\mathcal{F}$ in practice? Hopefully not! Indeed, in many situations, one chooses $\mathcal{F}$ to be some infinite "well-behaved" family. This is for instance the case for linear regression, where the set of predictors is the (infinite) set $\mathcal{F}_{\text{lin}} = \{x \mapsto x^T\theta,\ \theta \in \mathbb{R}^d\}$. For the classification problem, it turns out the size of the set $\mathcal{F}$ is not a good measure of its complexity. Rather, the "effective" size of a set $\mathcal{F}$ is measured by the number of classifications $(f(\mathbf{x_1}), \ldots, f(\mathbf{x_n}))$ over all $f \in \mathcal{F}$.

Let us first remark that, even if $\mathcal{F}$ is infinite, then the set of possible classifications $\mathcal{C}_{\mathcal{F}}(\mathbf{x_1}, \ldots, \mathbf{x_n}) := \{(f(\mathbf{x_1}), \ldots, f(\mathbf{x_n})),\ f \in \mathcal{F}\}$ is always finite, and of size at most 2^n (each $f(\mathbf{x_i})$ is equal to ± 1). Using a technical tool called *symmetrization*, one can show that one can indeed replace the size of $\mathcal{F}$ in Theorem 3.3 by the size of the classification set $\mathcal{C}_{\mathcal{F}}(\mathbf{x_1}, \ldots, \mathbf{x_n})$.

Lemma 3.4. *Let $\mathcal{N}_{\mathcal{F}}(\mathbf{x_1}, \ldots, \mathbf{x_n})$ be the size of the set $\mathcal{C}_{\mathcal{F}}(\mathbf{x_1}, \ldots, \mathbf{x_n})$. Then,*

$$\mathbb{E}[\sup_{f \in \mathcal{F}}(\mathcal{R}_P(f) - \mathcal{R}_n(f))] \leq 2\mathbb{E}\left[\sqrt{\frac{2\log \mathcal{N}_{\mathcal{F}}(\mathbf{x_1}, \ldots, \mathbf{x_n})}{n}}\right]. \qquad (24)$$

For many different examples, the quantity $\mathcal{N}_{\mathcal{F}}(\mathbf{x_1}, \ldots, \mathbf{x_n})$ is much smaller than the maximum possible value 2^n, making inequality (24) non trivial. Consider for instance the set $\mathcal{F}_0 = \{x \mapsto \mathbf{1}\{x^{(1)} \geq a\},\ a \in \mathbb{R}\}$. Then, the set $\mathcal{C}_{\mathcal{F}_0}(\mathbf{x_1}, \ldots, \mathbf{x_n})$ contains at most $n+1$ elements (see Figure 3). We therefore directly obtain a bound of order $\sqrt{\log n/n}$ in this case, which is comparable to the bound that we obtained in the previous section (Theorem 3.3), although $\mathcal{F}_0$ is infinite.

For general sets $\mathcal{F}$, bounding directly $\mathcal{N}_{\mathcal{F}}(\mathbf{x_1}, \ldots, \mathbf{x_n})$ is delicate, and there are even certain values of n for which finding a good bound is hopeless. Indeed, assume that for every $x_1, \ldots, x_n \in \mathcal{X}$ and every $y_1, \ldots, y_n \in \{0, 1\}$, one can find a function $f \in \mathcal{F}$ with $f(x_i) = y_i$ for $i = 1, \ldots, n$. Then, the empirical risk $\mathcal{R}_n(\hat{f}_{\mathcal{F}}) = \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{f(\mathbf{x_i}) \neq \mathbf{y_i}\}$ is always equal to 0. We are exactly in the overfitting regime where we expect the estimation error to be large. In that case, we have $\mathcal{N}_{\mathcal{F}}(\mathbf{x_1}, \ldots, \mathbf{x_n}) = 2^n$, making inequality (24) vacuous.

Definition 3.5 (Vapnik-Chervonenkis dimension)**.** *Let $\mathcal{F}$ be a set of functions from $\mathcal{X}$ to $\{-1, 1\}$. The **Vapnik-Chervonenkis dimension** $\mathrm{VC}(\mathcal{F})$ of $\mathcal{F}$ is defined as the largest number n such that there exists a configuration $x_1, \ldots, x_n \in \mathcal{X}$ such that for every possible classifications $y_1, \ldots, y_n \in \{-1, 1\}$, there exists $f \in \mathcal{F}$ with $f(x_i) = y_i$ for $i = 1, \ldots, n$. We set $\mathrm{VC}(\mathcal{F}) = +\infty$ if this condition holds for every $n \in \mathbb{N}$.*

According to the previous discussion, for $n \leq \mathrm{VC}(\mathcal{F})$, the set $\mathcal{F}$ is overfitting and there is no hope in bounding the estimation error. However, should $n \gg \mathrm{VC}(\mathcal{F})$, then the next lemma asserts that $\mathcal{N}_{\mathcal{F}}(x_1, \ldots, x_n)$ scales at most polynomially with n. In particular, it is *much* smaller than 2^n!

Lemma 3.6 (Sauer's lemma)**.** *Let $\mathcal{F}$ be a set with finite VC dimension. Let $n > 2\mathrm{VC}(\mathcal{F})$. Then, for every $x_1, \ldots, x_n \in \mathcal{X}$, we have*

$$\log \mathcal{N}_{\mathcal{F}}(x_1, \ldots, x_n) \leq \mathrm{VC}(\mathcal{F}) \log\left(\frac{en}{\mathrm{VC}(\mathcal{F})}\right). \tag{25}$$

Putting Lemma 3.4 and Lemma 3.6 together, we obtain the following result.

Theorem 3.7 (Error bound in expectation on the estimation error in binary classification: with VC dimension)**.** *Assume that $\mathcal{F}$ has a finite VC dimension*

VC$(\mathcal{F})$ *and that* ℓ *is the* $0-1$ *loss. Then, for* $n \geq 2\mathrm{VC}(\mathcal{F})$,

$$\begin{aligned}\mathbb{E}[\mathcal{R}_P(\hat{f}_{\mathcal{F}}) - \inf_{f\in\mathcal{F}} \mathcal{R}_P(f)] &\leq \mathbb{E}[\sup_{f\in\mathcal{F}}(\mathcal{R}_P(f) - \mathcal{R}_n(f))] \\ &\leq 2\sqrt{\frac{2\mathrm{VC}(\mathcal{F})}{n}\log\left(\frac{en}{\mathrm{VC}(\mathcal{F})}\right)}.\end{aligned} \tag{26}$$

We conclude by giving some properties of the VC dimension.

Proposition 3.8. *Let* $\mathcal{F}$ *be a set of functions from* $\mathcal{X}$ *to* $\{-1,1\}$.

1. *If* $\mathcal{F}$ *is of size* k, *then* $\mathrm{VC}(\mathcal{F}) \leq \log_2(k)$.

2. *It* $\mathcal{X} = \mathbb{R}^d$ *and* $\mathcal{F}$ *is the set of linear classifiers (that is* $f \in \mathcal{F}$ *is of the form* $f(x) = 1$ *if* x *belongs to some halfspace* H *and* -1 *otherwise), then* $\mathrm{VC}(\mathcal{F}) = d+1$.

3. *Let* $s \geq 1$ *be an integer and let* $\mathcal{F}_s$ *be the set of classifiers of the form* $\max_{j=1,\dots,s} f_j$ *for some functions* $f_j s$ *in* $\mathcal{F}$. *Then,*

$$\mathrm{VC}(\mathcal{F}_s) \leq \mathrm{VC}(\mathcal{F})(2s\log_2(3s)). \tag{27}$$

Remark 3.9. So far, we have only given tools to bound the estimation error $\mathcal{R}_P(\hat{f}_{\mathcal{F}}) - \inf_{f\in\mathcal{F}} \mathcal{R}_P(f)$. What about the approximation error $\inf_{f\in\mathcal{F}} \mathcal{R}_P(f) - \mathcal{R}_P^*$? This quantity will depend on the regularity of the Bayes predictor $f_P^\star$. Assume for the sake of simplicity that $\mathbf{x}$ is uniform on the cube $\mathcal{X} = [0,1]^d$ and $\mathbf{y} = f_0(\mathbf{x})$ for some function $f_0 : \mathcal{X} \to \{-1,1\}$ (that is there is no noise) for $d \geq 2$. Then, the Bayes risk $\mathcal{R}_P^\star$ is equal to 0 and the approximation error is equal to

$$\inf_{f\in\mathcal{F}} \mathbb{P}(f(\mathbf{x}) \neq f_0(\mathbf{x})). \tag{28}$$

This probability represents the area of the cube where f and f_0 differ. Let us consider a toy example to get some intuition. Assume that f_0 is equal to 1 on some smooth convex set of volume 1 and -1 outside. Consider the model $\mathcal{F}_s$ consisting of intersections of s halfplanes, that is every $f \in \mathcal{F}_s$ is of the form

$$f(x) = \begin{cases} 1 & \text{if } x \in \bigcap_{j=1}^s H_j \\ -1 & \text{otherwise,} \end{cases} \tag{29}$$

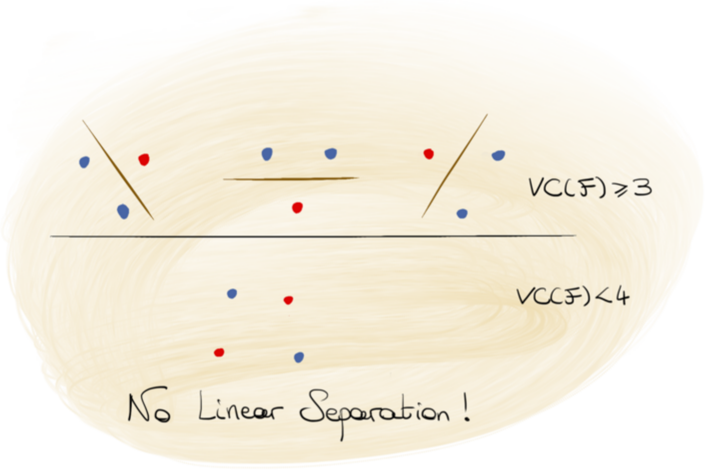

Figure 4: Let $\mathcal{F}$ be the set of linear classifier in $\mathbb{R}^2$. There exists a set of three points, such that linear classifiers can output all possible classifications (top). However, for all configurations of four points, there exists a classification that cannot be realized by a linear classifier (bottom). Therefore, $\mathrm{VC}(\mathcal{F}) = 3$.

where $H_1, \dots, H_s$ are s halfplanes. It is then known [Bronstein, 2008] that the approximation error is bounded by $c_d s^{-2/(d-1)}$ for some constant c_d. As expected, this quantity decreases as s gets larger. Using Theorem 3.7 and Proposition 3.8, we obtain the following bound on the excess risk (for n larger than s):

$$\mathbb{E}[\mathcal{R}_P(\hat{f}_{\mathcal{F}}) - \mathcal{R}_P^\star] \leq c_d' \sqrt{\frac{s \log(s) \log(n)}{n}} + c_d s^{-2/(d-1)}, \tag{30}$$

where c_d' is some positive constant. Letting $s = n^{(d-1)/(d+3)}$, we obtain a bound of order

$$\mathbb{E}[\mathcal{R}_P(\hat{f}_{\mathcal{F}}) - \mathcal{R}_P^\star] \leq c_d'' \log(n) n^{-2/(d+3)} \tag{31}$$

for some other constant c_d''. Note that this rate of convergence is extremely slow for large d: we refer to this phenomenon as the **curse of dimensionality**.

Appendix

Symmetrization

We provide here a proof of Lemma 3.4. This is a delicate proof, that uses a key technical tool used symmetrization. We call a random sign $\mathbf{e}$ that is equal to $+1$ with probability $1/2$ and -1 with probability $1/2$ as a **Rademacher random variable**.

Lemma 3.10. *Let T be a set and let $\mathbf{a_1}, \dots, \mathbf{a_n}$ be i.i.d. random variables in $\mathbb{R}^T$: each $\mathbf{a_i}$ is a function from T to $\mathbb{R}$. We assume that for every $t \in T$, $\mathbb{E}[\mathbf{a_i}(t)]$ is finite. Let $\mathbf{e_1}, \dots, \mathbf{e_n}$ be n i.i.d. Rademacher random variables, independent from the $\mathbf{a_i}$s. We have*

$$\mathbb{E}[\sup_{t \in T} \frac{1}{n} \sum_{i=1}^{n} (\mathbf{a_i}(t) - \mathbb{E}[\mathbf{a_i}(t)])] \leq 2 \cdot \mathbb{E}[\sup_{t \in T} \sum_{i=1}^{n} \mathbf{e_i} \mathbf{a_i}(t)]. \tag{32}$$

Proof. We introduce $\mathbf{a'_1}, \dots, \mathbf{a'_n}$ an independent copy from $\mathbf{a_1}, \dots, \mathbf{a_n}$. The random vectors $\mathbf{a_i} - \mathbf{a'_i}$ are independent and symmetric, such that $\mathbf{a'_i} - \mathbf{a_i}$ has the same law as $\mathbf{a_i} - \mathbf{a'_i}$. One can check that $\mathbf{a_i} - \mathbf{a'_i}$ has the same law as $\mathbf{e_i}(\mathbf{a_i} - \mathbf{a'_i})$. Therefore,

$$\begin{aligned}
\mathbb{E}[\sup_{t \in T} \frac{1}{n} \sum_{i=1}^{n} (\mathbf{a_i}(t) - \mathbb{E}[\mathbf{a_i}(t)])] &= \mathbb{E}[\sup_{t \in T} \frac{1}{n} \sum_{i=1}^{n} (\mathbf{a_i}(t) - \mathbb{E}[\mathbf{a'_i}(t)])] \\
&= \mathbb{E}[\sup_{t \in T} \mathbb{E}[\frac{1}{n} \sum_{i=1}^{n} (\mathbf{a_i}(t) - \mathbf{a'_i}(t)) | \mathbf{a_1}, \dots, \mathbf{a_n}])].
\end{aligned}$$

The function $z \mapsto \sup_{t \in t} z(t)$ is convex. Therefore, by Jensen's inequality,

$$\begin{aligned}
\mathbb{E}[\sup_{t\in T} \frac{1}{n}\sum_{i=1}^{n}(\mathbf{a_i}(t) - \mathbb{E}[\mathbf{a_i}(t)])] &\leq \mathbb{E}[\mathbb{E}[\sup_{t\in T} \frac{1}{n}\sum_{i=1}^{n}(\mathbf{a_i}(t) - \mathbf{a_i'}(t))]] \\
&= \mathbb{E}[\sup_{t\in T} \frac{1}{n}\sum_{i=1}^{n}(\mathbf{a_i}(t) - \mathbf{a_i'}(t))] \\
&= \mathbb{E}[\sup_{t\in T} \frac{1}{n}\sum_{i=1}^{n}\mathbf{e_i}(\mathbf{a_i}(t) - \mathbf{a_i'}(t))] \\
&= \mathbb{E}[\sup_{t\in T} \frac{1}{n}\sum_{i=1}^{n}\mathbf{e_i}\mathbf{a_i}(t)] + \mathbb{E}[\sup_{t\in T} \frac{1}{n}\sum_{i=1}^{n}-\mathbf{e_i}\mathbf{a_i'}(t))] \\
&= 2\cdot\mathbb{E}[\sup_{t\in T} \frac{1}{n}\sum_{i=1}^{n}\mathbf{e_i}\mathbf{a_i}(t)].
\end{aligned}$$

□

We apply this general inequality with $T = \mathcal{F}$ to the random variables $\mathbf{a_i}(f) = \ell_{01}(f(\mathbf{x_i}), \mathbf{y_i}) = \mathbf{1}\{f(\mathbf{x_i}) \neq \mathbf{y_i}\}$ to obtain

$$\begin{aligned}
\mathbb{E}[\sup_{f\in\mathcal{F}}(\mathcal{R}_P(f) - \mathcal{R}_n(f))] &\leq 2\cdot\mathbb{E}[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\mathbf{e_i}\mathbf{1}\{f(\mathbf{x_i}) \neq \mathbf{y_i}\}] \\
&= 2\cdot\mathbb{E}[\sup_{u\in\mathcal{C}_{\mathcal{F}}(\mathbf{x_1},\ldots,\mathbf{x_n})} \frac{1}{n}\sum_{i=1}^{n}\mathbf{e_i}\mathbf{1}\{u_i \neq \mathbf{y_i}\}],
\end{aligned} \tag{33}$$

where $u = (u_1, \ldots, u_n)$ is any element of $\mathcal{C}_{\mathcal{F}}(\mathbf{x_1}, \ldots, \mathbf{x_n})$, the set of classifications of $\mathbf{x_1}, \ldots, \mathbf{x_n}$ using a classifier $f \in \mathcal{F}$. Conditionally on the observations $(\mathbf{x_1}, \mathbf{y_1}), \ldots, (\mathbf{x_n}, \mathbf{y_n})$, the random variables $\frac{1}{n}\sum_{i=1}^{n}\mathbf{e_i}\mathbf{1}\{u_i \neq \mathbf{y_i}\}$ satisfy the assumptions of Theorem 3.1 with $\sigma = 1/\sqrt{n}$. Also, there are $\mathcal{N}_{\mathcal{F}}(\mathbf{x_1}, \ldots, \mathbf{x_n})$ such random variables. Then, by applying Theorem 3.1 conditionally on the observations $(\mathbf{x_1}, \mathbf{y_1}), \ldots, (\mathbf{x_n}, \mathbf{y_n})$, we obtain that

$$\mathbb{E}[\sup_{f\in\mathcal{F}}(\mathcal{R}_P(f) - \mathcal{R}_n(f))] \leq 2\cdot\mathbb{E}\left[\sqrt{\frac{2\log\mathcal{N}_{\mathcal{F}}(\mathbf{x_1}, \ldots, \mathbf{x_n})}{n}}\right], \tag{34}$$

that is exactly Lemma 3.4.

To summarize, we have used symmetrization to replace a supremum over an infinite number of random variables (each random variable $\mathcal{R}_P(f) - \mathcal{R}_n(f)$ are different because $\mathcal{R}_P(f)$ is a priori different for every function $f \in \mathcal{F}$) to a supremum over only a finite number of $\mathcal{C}_{\mathcal{F}}(\mathbf{x_1}, \ldots, \mathbf{x_n})$ random variables.

References

[Bronstein, 2008] Bronstein, E. M. (2008). Approximation of convex sets by polytopes. *Journal of Mathematical Sciences*, 153(6):727–762.

[Yeh and Hsu, 2018] Yeh, I.-C. and Hsu, T.-K. (2018). Building real estate valuation models with comparative approach through case-based reasoning. *Applied Soft Computing*, 65:260–271.