STAT 425: Introduction to Nonparametric Statistics **Winter 2026**

Lecture 9: Empirical Risk Minimization and Concentration Inequalities

Instructor: Yen-Chi Chen

9.1 Introduction

Recall that in (binary) classification problem, we observe

$$(X_1, Y_1), \cdots, (X_n, Y_n)$$

that are IID from some distribution P and each Y_i is a binary class label, i.e. $Y = 0$ or 1 and each $X_i \in \mathcal{X} \subset \mathbb{R}^d$ contains d variables/features/covariates.

A classifier $c : \mathcal{X} \mapsto \{0, 1\}$ is a function such that the input is the feature of a new observation (without label) and then the output is a predicted class label.

To measure the quality of our classifier, we use a loss function $L(c(x), y)$ and a common loss function is the $0-1$ loss, which is $L(c(x), y) = I(c(x) \neq y)$. Note that the MLE method for the logistic regression uses another loss function. The expected loss is the risk function

$$R(c) = \mathbb{E}(L(c(X), Y)),$$

where $(X, Y) \sim P$. $R(c)$ is the expected loss for the future prediction.

Let $\mathcal{C}$ be a collection of classifiers, we want to find the one $c^* \in \mathcal{C}$ such that the risk function is minimized, i.e.,

$$c^* = \mathsf{argmin}_{c \in \mathcal{C}} R(c).$$

This classifier is called the Bayes classifier and it it is the one with the best predictive performance for the future data.

Because $R(c)$ is an unknown quantity, so a sample analogue is

$$\widehat{R}_n(c) = \frac{1}{n} \sum_{i=1}^{n} L(c(X_i), Y_i).$$

$\widehat{R}_n(c)$ is called the **empirical risk**. By the law of large number, $\widehat{R}_n(c) - R(c) = o_P(1)$ for any given c.

The **empirical risk minimization (ERM)** is to find the classifier c^* by minimizing the empirical risk

$$\widehat{c} = \mathsf{argmin}_{c \in \mathcal{C}} \widehat{R}_n(c). \tag{9.1}$$

As we have seen in the past few lectures, the ERM may not work because we not only need $\widehat{R}_n(c) - R(c) \approx 0$ for a given c but also $\widehat{R}_n(c) - R(c) \approx 0$ *uniformly for all* $c \in \mathcal{C}$. Namely, we need

$$\sup_{c \in \mathcal{C}} |\widehat{R}_n(c) - R(c)| = o_P(1). \tag{9.2}$$

When $\mathcal{C}$ contains only finite number of classifiers, say N classifiers, you can show that (using Hoeffding's inequality in Section 9.3)

$$\sup_{c\in\mathcal{C}} |\widehat{R}_n(c) - R(c)| = O_P\left(\sqrt{\frac{\log N}{n}}\right),$$

so as long as N is not too large compared to the sample size n, we eventually have (9.2).

Even when $\mathcal{C}$ contains infinite number of classifiers, as long as its VC dimension is not too large, the VC theory (Section 9.4) tells us that

$$\sup_{c\in\mathcal{C}} |\widehat{R}_n(c) - R(c)| = O_P\left(\sqrt{\nu\frac{\log n}{n}}\right),$$

where ν is the VC dimension of $\mathcal{C}$ (which is often a fixed number).

9.2 Excess Risk

Recall that the **excess risk** of a classifier c is defined as

$$\mathcal{E}(c) = R(c) - R(c^*) = R(c) - \min_{c\in\mathcal{C}} R(c).$$

Namely, the excess risk describes the amount of decreased performance of a classifier c compared to the Bayes classifier c^*.

If a classifier $\widehat{c}_n$ is from the ERM (equation (9.1)), how will its excess risk be like? It turns out that we can control the excess risk using the uniform error bound

$$\epsilon_n = \sup_{c\in\mathcal{C}} |\widehat{R}_n(c) - R(c)|.$$

Because $\widehat{c}$ is the minimizer of $\widehat{R}_n(c)$,

$$\widehat{R}_n(\widehat{c}) \le \widehat{R}_n(c)$$

for any classifier $c \in \mathcal{C}$, including c^*. Thus,

$$\widehat{R}_n(\widehat{c}) \le \widehat{R}_n(c^*).$$

Because ϵ_n is the uniform bound, which implies

$$|\widehat{R}_n(\widehat{c}) - R(\widehat{c})| \le \epsilon_n, \quad |\widehat{R}_n(c^*) - R(c^*)| \le \epsilon_n.$$

This further implies

$$\begin{aligned} R(\widehat{c}) &\le \widehat{R}_n(\widehat{c}) + \epsilon_n \\ &\le \underbrace{\widehat{R}_n(c^*)}_{\le R(c^*)+\epsilon_n} + \epsilon_n \\ &\le R(c^*) + 2\epsilon_n \end{aligned}$$

and

$$\mathcal{E}(\widehat{c}) = R(\widehat{c}) - R(c^*) \le 2\epsilon_n = 2 \times \sup_{c\in\mathcal{C}} |\widehat{R}_n(c) - R(c)|.$$

Namely, the excess risk of a classifier from ERM is no more than 2 times the uniform error.

So when the uniform error ϵ_n converges to 0, the excess risk converges to 0, implying that the classifier $\widehat{c}$ is as good as the Bayes classifier.

Moreover, the distribution of ϵ_n can be used to construct a confidence interval for the Bayes risk $R(c^*) = \min_{c\in\mathcal{C}} R(c)$. Let $t_{1-\alpha}$ be the $1-\alpha$ upper quantile of ϵ_n, i.e,

$$P(\epsilon_n \le t_{1-\alpha}) = 1-\alpha.$$

Recall that to construct a CI, we often need a lower bound and an upper bound.

Because

$$\widehat{R}_n(\widehat{c}) \le \widehat{R}_n(c^*) \le R(c^*) + \epsilon_n,$$

a lower bound of $R(c^*)$ can be obtained by

$$\widehat{R}_n(\widehat{c}) - \epsilon_n \le R(c^*).$$

For the upper bound,

$$R(c^*) \le R(\widehat{c}) \le \widehat{R}_n(\widehat{c}) + \epsilon_n.$$

Thus, we conclude that

$$\widehat{R}_n(\widehat{c}) - \epsilon_n \le R(c^*) \le \widehat{R}_n(\widehat{c}) + \epsilon_n.$$

Namely,

$$|R(c^*) - \widehat{R}_n(\widehat{c})| \le \epsilon_n.$$

Thus, a $1-\alpha$ CI of the Bayes risk $R(c^*)$ is

$$\left[\widehat{R}_n(\widehat{c}) - t_{1-\alpha}, \widehat{R}_n(\widehat{c}) + t_{1-\alpha}\right].$$

9.3 Hoeffding's Inequality

For bounded IID random variables, the Hoeffding's inequality is a powerful to bound the behavior of their average.

We first start with a property of a bounded random variable.

Lemma 9.1 *Let X be a random variable with $\mathbb{E}(X) = 0$ and $a \le X \le b$. Then*

$$\mathbb{E}(e^{tX}) \le e^{t^2(b-a)^2/8}$$

for any positive number t.

Proof: We will use the fact that $x \mapsto e^{tx}$ is a convex function for all positive t. Recall that a function $g(x)$ is a convex function if for any two point $a < b$ and $\alpha \in [0,1]$,

$$g(\alpha a + (1-\alpha)b) \le \alpha g(a) + (1-\alpha)g(b).$$

Because $X \in [a,b]$, we define α_X to

$$X = \alpha_X b + (1-\alpha_X)a.$$

This implies

$$\alpha_X = \frac{X-a}{b-a}$$

Using the fact that $x \mapsto e^{tx}$ is convex,

$$e^{tX} \leq \alpha_X e^{tb} + (1-\alpha_X)e^{ta} = \frac{X-a}{b-a}e^{tb} + \frac{b-X}{b-a}e^{ta}.$$

Now taking the expectation in both sides,

$$\mathbb{E}(e^{tX}) \leq \frac{\mathbb{E}(X)-a}{b-a}e^{tb} + \frac{b-\mathbb{E}(X)}{b-a}e^{ta} = \frac{b}{b-a}e^{ta} - \frac{a}{b-a}e^{tb} = e^{g(s)}, \tag{9.3}$$

where $s = t(b-a)$ and $g(s) = -\gamma s + \log(1-\gamma+\gamma e^s)$ and $\gamma = -a/(b-a)$. Note that $g(0) = g'(0) = 0$ and $g''(s) \leq 1/4$ for all positive s. Using Taylor's theorem,

$$g(s) = g(0) + sg'(0) + \frac{1}{2}s^2 g''(s^*)$$

for some $s^* \in [0, s]$. Thus, we conclude $g(s) \leq \frac{1}{2} \times s^2 \times \frac{1}{4} = \frac{1}{8}s^2$.

Then equation (9.3) implies

$$\mathbb{E}(e^{tX}) \leq e^{g(s)} \leq e^{\frac{s^2}{8}} = e^{\frac{t^2(b-a)^2}{8}}.$$

■

With this lemma, we can prove the Hoeffding's inequality.

Theorem 9.2 *Let $X_1, \cdots, X_n$ be IID with $\mathbb{E}(X_1) = \mu$ and $a \leq X_1 \leq b$. Then*

$$P\left(|\bar{X}_n - \mu| \geq \epsilon\right) \leq 2e^{-2n\epsilon^2/(b-a)^2}.$$

Proof:

We first prove that $P\left(\bar{X}_n - \mu \geq \epsilon\right) \leq e^{-2n\epsilon^2/(b-a)^2}$.

Let $Y_i = X_i - \mu$. Because the exponential function is monotonic, for any positive r,

$$\begin{aligned}
P\left(\bar{X}_n - \mu \geq \epsilon\right) &= P\left(\bar{Y}_n \geq \epsilon\right) \\
&= P\left(\sum_{i=1}^n Y_i \geq n\epsilon\right) \\
&= P\left(e^{\sum_{i=1}^n Y_i} \geq e^{n\epsilon}\right) \\
&= P\left(e^{t\sum_{i=1}^n Y_i} \geq e^{tn\epsilon}\right) \\
&\leq \frac{\mathbb{E}(e^{t\sum_{i=1}^n Y_i})}{e^{tn\epsilon}} \quad \text{by Markov's inequality} \\
&= e^{-tn\epsilon}\mathbb{E}(e^{tY_1} \cdot e^{tY_2} \cdots e^{tY_n}) \\
&= e^{-tn\epsilon}\mathbb{E}(e^{tY_1}) \cdot \mathbb{E}(e^{tY_2}) \cdots \mathbb{E}(e^{tY_n}) \\
&= e^{-tn\epsilon}\mathbb{E}(e^{tY_1})^n \\
&\leq e^{-tn\epsilon}e^{nt^2(b-a)^2/8} \text{ by Lemma 9.1.}
\end{aligned}$$

Because the above inequality holds for all positive t, we can choose t to optimize the bound. To get the bound as sharp as possible, we would like to make it as small as possible. Thus, we need to find t such that

$$-tn\epsilon + nt^2(b-a)^2/8$$

is minimized. Taking derivatives with respect to t and set it to be 0, we obtain

$$t_* = \frac{4\epsilon}{(b-a)^2}$$

and

$$-t_* n\epsilon + nt_*^2(b-a)^2/8 = -2n\epsilon^2/(b-a)^2.$$

Thus, the inequality becomes

$$P\left(\bar{X}_n - \mu \geq \epsilon\right) \leq e^{-t_* n\epsilon} e^{nt_*^2(b-a)^2/8} = e^{-2n\epsilon^2/(b-a)^2}.$$

The same proof also applies to the case $P\left(\bar{X}_n - \mu \leq \epsilon\right)$ and we will obtain the same bound. Therefore, we conclude that

$$P\left(|\bar{X}_n - \mu| \geq \epsilon\right) \leq 2e^{-2n\epsilon^2/(b-a)^2}.$$

■

An inequality like the one in Theorem 9.2 is called a concentration inequality and this phenomena caused by this type of inequalities is called concentration of measures – indeed the probability measure of $\bar{X}_n$ is concentrating around μ, its expectation. Concentration inequalities are common theoretical tools in analyzing the property of a statistic. And they are particularly powerful when applies to estimators and classifiers.

9.3.1 Application in Classification

The Hoeffding's inequality has a direct application to classification problem Remember that in binary classification case with $0-1$ loss, the loss at each observation is

$$L(c(X_i), Y_i) = I(c(X_i) \neq Y_i),$$

a bounded random variable with $a = 0$ and $b = 1$! The empirical risk

$$\widehat{R}_n(c) = \frac{1}{n}\sum_{i=1}^{n} L(c(X_i), Y_i)$$

is an unbiased estimator of $\mathbb{E}(\widehat{R}_n(c)) = R(c)$. Thus, the Hoeffding's inequality implies that for a given classifier c,

$$P(|\widehat{R}_n(c) - R(c)| \geq \epsilon) \leq 2e^{-2n\epsilon^2}. \tag{9.4}$$

A powerful feature of the Hoeffding's inequality is that it holds *regardless of the classifier*. Namely, even if we are considering many different types of classifiers, some are decision trees, some are kNN, some are logistic regression, they all satisfy equation (9.4).

Assume that the collection of classifier $\mathcal{C}$ contains only N classifiers $c_1, \cdots, c_N$. Then we have

$$\begin{aligned} P\left(\sup_{c\in\mathcal{C}} |\widehat{R}_n(c) - R(c)| \geq \epsilon\right) &= P\left(\max_{j=1,\cdots,N} |\widehat{R}_n(c_j) - R(c_j)| \geq \epsilon\right) \\ &\leq \sum_{j=1}^{N} P\left(|\widehat{R}_n(c_j) - R(c_j)| \geq \epsilon\right) \\ &\leq N2e^{-2n\epsilon^2}. \end{aligned}$$

Thus, as long as $N \cdot 2e^{-2n\epsilon^2} \to 0$, we have

$$P\left(\sup_{c\in\mathcal{C}} |\widehat{R}_n(c) - R(c)| \geq \epsilon\right) \to 0 \Leftrightarrow \sup_{c\in\mathcal{C}} |\widehat{R}_n(c) - R(c)| \xrightarrow{P} 0,$$

the desired property we want in equation (9.2).

9.3.2 Convergence Rate

The Hoeffding's inequality directly implies that

$$|\bar{X}_n - \mu| = O_P\left(\frac{1}{\sqrt{n}}\right).$$

It actually implies a bound on the expectation $\mathbb{E}|\bar{X}_n - \mu|$ as well.

To see this, recall that for a positive random variable T,

$$\mathbb{E}(T) = \int_0^\infty P(T > t)dt.$$

Thus, for any positive s,

$$\begin{aligned}
\mathbb{E}\left(|\bar{X}_n - \mu|^2\right) &= \int_0^\infty P(|\bar{X}_n - \mu|^2 > t)dt \\
&= \int_0^\infty P(|\bar{X}_n - \mu| > \sqrt{t})dt \\
&= \int_0^s \underbrace{P(|\bar{X}_n - \mu| > \sqrt{t})}_{\leq 1} dt + \int_c^\infty \underbrace{P(|\bar{X}_n - \mu| > \sqrt{t})}_{\text{Hoeffding}} dt \\
&\leq \int_0^s 1dt + \int_c^\infty 2e^{-2nt/(b-a)^2} dt \\
&= s + \frac{(b-a)^2}{n} e^{-ns/(b-a)^2}.
\end{aligned}$$

It is not easy to optimize this but there are some simple choice that gives a good result. We will choose $s = \frac{2(b-a)^2 \log \frac{1}{b-a}}{n}$, which leads to

$$\mathbb{E}\left(|\bar{X}_n - \mu|^2\right) \leq \frac{2(b-a)^2 \log \frac{1}{b-a}}{n} + \frac{1}{n} = O\left(\frac{1}{n}\right).$$

Finally, using the Cauchy-Schwarz inequality,

$$\mathbb{E}|\bar{X}_n - \mu| \leq \sqrt{\mathbb{E}\left(|\bar{X}_n - \mu|^2\right)} = O\left(\frac{1}{\sqrt{n}}\right), \tag{9.5}$$

which is the desired bound on the expectation.

In the case of classification, when that the collection of classifier $\mathcal{C}$ contains N classifiers, we have

$$P\left(\sup_{c\in\mathcal{C}} |\widehat{R}_n(c) - R(c)| \geq \epsilon\right) \leq N \cdot 2e^{-2n\epsilon^2}.$$

This implies

$$\sup_{c\in\mathcal{C}} |\widehat{R}_n(c) - R(c)| = O_P\left(\sqrt{\frac{\log N}{n}}\right).$$

Moreover, using a similar calculation as we have done for deriving equation (9.5), one can show that

$$\mathbb{E}\left(\sup_{c\in\mathcal{C}} |\widehat{R}_n(c) - R(c)|\right) = O\left(\sqrt{\frac{\log N}{n}}\right).$$

9.3.3 Application in Histogram

We have derived the convergence rate for a histogram density estimator at a given point x. However, we have not yet analyzed the *uniform convergence rate* of the histogram. Using the Hoeffding's inequality, we are able to obtain such a uniform convergence rate.

Assume that $X_1, \cdots, X_n \sim F$ with a PDF p that has a non-zero density over $[0, 1]$. If we use a histogram with M bins:

$$B_1 = \left[0, \frac{1}{M}\right), B_2 = \left[\frac{1}{M}, \frac{2}{M}\right), \cdots, B_M = \left[\frac{M-1}{M}, 1\right].$$

Let $\widehat{p}_M(x)$ be the histogram density estimator:

$$\widehat{p}_M(x) = \frac{M}{n}\sum_{i=1}^{n} I(X_i \in B(x)),$$

where $B(x)$ is the bin that x belongs to.

The goal is to bound

$$\sup_x |\widehat{p}_M(x) - p(x)|.$$

We know that the difference $\widehat{p}_M(x) - p(x)$ can be written as

$$\widehat{p}_M(x) - p(x) = \underbrace{\widehat{p}_M(x) - \mathbb{E}(\widehat{p}_M(x))}_{\text{stochastic variation}} + \underbrace{\mathbb{E}(\widehat{p}_M(x)) - p(x)}_{\text{bias}}.$$

The bias analysis we have done in Lecture 6 can be generalized to every point x, so we have

$$\sup_x |\mathbb{E}(\widehat{p}_M(x)) - p(x)| = O\left(\frac{1}{M}\right).$$

So we only need to bound the stochastic variation part $\sup_x |\widehat{p}_M(x) - \mathbb{E}(\widehat{p}_M(x))|$. Although we are taking supremum over every x, there are only M bins $B_1, \cdots, B_M$ so we can rewrite the stochastic part as

$$\begin{aligned}\sup_x |\widehat{p}_M(x) - \mathbb{E}(\widehat{p}_M(x))| &= \max_{j=1,\cdots,M}\left|\frac{M}{n}\sum_{i=1}^{n} I(X_i \in B_j) - MP(X_i \in B_j)\right| \\ &= M \cdot \max_{j=1,\cdots,M}\left|\frac{1}{n}\sum_{i=1}^{n} I(X_i \in B_j) - P(X_i \in B_j)\right|.\end{aligned}$$

Because the indicator function takes only two values: 0 and 1, we have

$$\begin{aligned} P\left(\sup_x |\widehat{p}_M(x) - \mathbb{E}(\widehat{p}_N(x))| > \epsilon\right) &= P\left(M \cdot \max_{j=1,\cdots,M} \left|\frac{1}{n}\sum_{i=1}^n I(X_i \in B_j) - P(X_i \in B_j)\right| > \epsilon\right) \\ &= P\left(\max_{j=1,\cdots,M} \left|\frac{1}{n}\sum_{i=1}^n I(X_i \in B_j) - P(X_i \in B_j)\right| > \underbrace{\frac{\epsilon}{M}}_{=\epsilon'}\right) \\ &\leq M \cdot 2e^{-2n\epsilon'^2} \\ &= M \cdot 2e^{-2n\epsilon^2/M^2}. \end{aligned}$$

Thus,

$$\sup_x |\widehat{p}_M(x) - \mathbb{E}(\widehat{p}_M(x))| = O_P\left(\sqrt{\frac{M^2 \log M}{n}}\right)$$

This, together with the uniform bias, implies

$$\begin{aligned} \sup_x |\widehat{p}_M(x) - p(x)| &\leq \sup_x |\mathbb{E}(\widehat{p}_N(x)) - p(x)| + \sup_x |\widehat{p}_M(x) - \mathbb{E}(\widehat{p}_N(x))| \\ &= O\left(\frac{1}{M}\right) + O_P\left(\sqrt{\frac{M^2 \log M}{n}}\right). \end{aligned}$$

Note that this bound is not the tightest bound we can obtain. Using the Bernstein's inequality[1], you can obtain a faster convergence rate:

$$\sup_x |\widehat{p}_M(x) - p(x)| = O\left(\frac{1}{M}\right) + O_P\left(\sqrt{\frac{M \log M}{n}}\right).$$

9.4 VC Theory

Hoeffding's inequality has helped us to partially solve the problem of proving equation (9.2) when the number of classifiers are fixed. However, in many cases we are working on infinite number of classifiers so that the Hoeffding's inequality does not work.

For instance, in the logistic regression, the collection of all possible classifiers is

$$\mathcal{C}_{\text{logistic}} = \{\widetilde{c}_{\beta_0,\beta} : \beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^d\},$$

where

$$\widetilde{c}_{\beta_0,\beta}(x) = I\left(\frac{e^{\beta_0+\beta^T x}}{1+e^{\beta_0+\beta^T x}} > \frac{1}{2}\right).$$

The parameters (indices) are $\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^d$, both contain infinite number of points (the real line contain infinite number of points). Thus, the Hoeffding's inequality will not work in this case.

To overcome this problem, we first think about the performance of the classifier $\widetilde{c}_{\beta_0,\beta}(x)$ for different values of (β_0, β). For simplicity, we assume that the classifiers c_θ are indexed by a quantity $\theta \in \mathbb{R}^p$, i.e., there are p parameters controlling each classifier and let

$$\mathcal{C}_\Theta = \{c_\theta : \theta \in \mathbb{R}^p\}.$$

[1] https://en.wikipedia.org/wiki/Bernstein_inequalities_(probability_theory)

In the case of the logistic regression, $p = d + 1$ and $\theta = (\beta_0, \beta)$.

Consider two classifiers c_{θ_1} and c_{θ_2}. When θ_1, θ_2 are close, we will expect their performance to be similar. Just like in the logistic regression, when two classifiers with similar parameters, their decision boundary (the boundary between the two class labels) should be similar. Namely, the performance of many classifiers may be similar to each other so when using the concentration inequality, we do not need to treat all of them as independent classifiers.

Thus, although there are infinite number of classifiers in $\mathcal{C}_\Theta$, there might only be a few *effetive independent* classifiers. Instead of using the total number of classifiers in constructing a uniform bound, we will use

$$P\left(\sup_{c\in\mathcal{C}} |\widehat{R}_n(c) - R(c)| \geq \epsilon\right) \leq N_{\text{eff}} \cdot 2e^{-2n\epsilon^2}, \tag{9.6}$$

where N_{eff} is some measure of effective independent classifiers. In a sense, N_{eff} measures the complexity of $\mathcal{C}_\Theta$.

There are many ways of measuring the number of effective independent classifiers such as the covering number and bracketing number, we will introduce a simple and famous quantity call the VC dimension.

9.4.1 Shattering Number and VC Dimension

In the $0-1$ loss case, given a classifier c, the empirical risk

$$\widehat{R}_n(c) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \neq c(X_i))$$

and the actual risk

$$R(c) = \mathbb{E}(I(Y_1 \neq c(X_1))) = P(Y_1 \neq c(X_1)).$$

If we define a new random variable $Z_i = (X_i, Y_i)$, then the event

$$Y_i \neq c(X_i) \Longleftrightarrow Z_i \in A_c,$$

for some region A_c that is determined by the classifier c. Thus, we can then rewrite the difference between the empirical risk and the actual risk as

$$\widehat{R}_n(c) - R(c) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \neq c(X_i)) - P(Y_1 \neq c(X_1)) = \frac{1}{n}\sum_{i=1}^{n} I(Z_i \in A_c) - P(Z_1 \in A_c)).$$

The uniform bound in equation (9.2) can then be rewritten as

$$\sup_{c\in\mathcal{C}} |\widehat{R}_n(c) - R(c)| = \sup_{c\in\mathcal{C}} \left|\frac{1}{n}\sum_{i=1}^{n} I(Z_i \in A_c) - P(Z_1 \in A_c))\right| = \sup_{A\in\mathcal{A}} \left|\frac{1}{n}\sum_{i=1}^{n} I(Z_i \in A) - P(Z_1 \in A))\right|,$$

where $\mathcal{A} = \{A_c : c \in \mathcal{C}\}$ is a collection of regions. Thus, the effective number N_{eff} should be related to how complex the set $\mathcal{A}$ is. If $\mathcal{A}$ contains many very different regions, then N_{eff} is expected to be large.

Note that the quantity $\frac{1}{n}\sum_{i=1}^{n} I(Z_i \in A)$ is just the ratio of our data that falls within the area A_c whereas the quantity $P(Z_1 \in A)$ is the probability of observing an random quantity Z_1 within the area A_c. Thus, eventually we are trying to bound the difference between using the ratio of a sample as an estimator of the probability of falling into a given region, which is the classical problem in STAT 101. However, the challenge here is that we are considering many many many regions so the problem is much more complicated.

The VC theory shows the following bound:

$$P\left(\sup_{A\in\mathcal{A}}\left|\frac{1}{n}\sum_{i=1}^{n}I(Z_i\in A)-P(Z_1\in A))\right|>\epsilon\right)\leq 8(n+1)^{\nu(\mathcal{A})}e^{-n\epsilon^2/32}. \tag{9.7}$$

Namely, $8(n+1)^{\nu(\mathcal{A})}=N_{\mathsf{eff}}$ is the effective number we want and the quantity $\nu(\mathcal{A})$ is called the **VC dimension** of $\mathcal{A}$[2] .

To define the VC dimension, we first introduce the concept of *shattering*. Shattering is a concept about using sets in a collection of regions $\mathcal{R}$ to partition a set of points. Here are some simple examples of $\mathcal{R}$:

1. $\mathcal{R}_1=\{(\infty,t]:t\in\mathbb{R}\}$. This is related to the uniform bound of the EDF.
2. $\mathcal{R}_2=\{(a,b):a\leq b\}$.
3. $\mathcal{R}_3=\{[x_0-r,x_0+r]:r\geq 0\}$ for some fixed x_0.
4. $\mathcal{R}_4=\{[x_0-r,x_0+r]:r\geq 0,x_0\in\mathbb{R}\}$. Note that the difference between $\mathcal{R}_3$ and $\mathcal{R}_4$ is that this collection is much larger than the previous one because the x_0 in the previous example is fixed whereas here x_0 is allowed to change.

The above examples are where $\mathcal{R}$ contains subregions in 1D. But it can be generalized to other dimensions. Here are some examples about d-dimension regions:

1. $\mathcal{R}_5=\{(\infty,t]\times(\infty,s]:t,s\in\mathbb{R}\}$. In this case, $d=2$ and this is related to the performance of a $2D$ EDF.
2. $\mathcal{R}_6=\{(a_1,b_1)\times(a_2,b_2)\times\cdots\times(a_d,b_d):a_j\leq b_j,\ j=1,\cdots,d\}$. In this case the dimension is d.
3. $\mathcal{R}_7=\{B(x_0,r):r\geq 0\}$, where $B(x_0,r)=\{x:\|x-x_0\|\leq r\}$ is a d-dimensional ball with radius r centered at x_0.
4. $\mathcal{R}_8=\{\mathcal{B}_{\beta,c}:\beta\in\mathbb{R}^d,c\in\mathbb{R}\}$, where $\mathcal{B}_{\beta,c}=\{x:\beta^Tx+c\leq 0\}$ is the half space in d-dimension. This is related to the linear classifier or the logistic regression.

Let $F=\{x_1,\cdots,x_n\}\subset\mathbb{R}^d$ be a set of points. For a subset $G\subset F$, we say $\mathcal{R}$ **picks out** G if there exists $R\in\mathcal{R}$ such that

$$G=R\cap F.$$

Namely, there exists a subregions in $\mathcal{R}$ that we can find the subset G by the intersection of F and this subregion.

Example 1. For example, consider the case $d=1$ and $\mathcal{R}=\mathcal{R}_1$ and let $F=\{1,5,8\}$. For the case $G=\{1,5\}$, it can be picked out by $\mathcal{R}_1$ because we can choose $(-\infty,6]\in\mathcal{R}_1$ such that $(-\infty,6]\cap F=G$. Note that the choice may not be unique; $(-\infty,7]\in\mathcal{R}_1$ also picks up G. However, the subset $G'=\{5,8\}$ cannot be picked out by any element of $\mathcal{R}_1$.

Let $\mathcal{S}(\mathcal{R},F)$ be the number of subsets of F that can be picked out by $\mathcal{R}$. By definition, if F contains n elements, then $\mathcal{S}(\mathcal{R},F)\leq 2^n$ (because there is at most 2^n unique subsets of n elements, including the empty set).

[2]Note that the quantity in the exponential part becomes $-n\epsilon^2/32$ rather than $-2n\epsilon^2$. The rates $n\epsilon^2$ are the same but just the constants are different. This change is due to the fact that we are not just using Hoeffding's inequality but also another technique called symmetrization in the proof so the constant changes.

Example 2. In the case of Example 1, $\mathcal{S}(\mathcal{R}_1, F)$ is 4 because $\mathcal{R}_1$ can only picks out the following subsets:

$$\emptyset, \{1\}, \{1,5\}, \{1,5,8\}.$$

However, if we consider $\mathcal{R}_2$, the quantity becomes $\mathcal{S}(\mathcal{R}_2, F) = 7$. The only case that $\mathcal{R}_2$ cannot pick out is $\{1,8\}$.

Example 3. Now assume that we are working on $F' = \{1,5\}$, then $\mathcal{S}(\mathcal{R}_1, F') = 3$ because we can picks out

$$\emptyset, \{1\}, \{1,5\}$$

and $\mathcal{S}(\mathcal{R}_2, F') = 4 = 2^2$ because every subset of F' can be picked out.

F is **shattered** by $\mathcal{R}$ if $\mathcal{S}(\mathcal{R}, F) = 2^n$, where n is the number of elements in F. Namely, F is shattered by $\mathcal{R}$ if every subset of F can be picked out by $\mathcal{R}$. In Example 2 and 3, we see that $F' = \{1,5\}$ is shattered by $\mathcal{R}_2$ but $F = \{1,5,8\}$ is not.

The **VC dimension** of $\mathcal{R}$, denoted as $\nu(\mathcal{R})$, is the maximal number of distinct points that can shattered by $\mathcal{R}$. Namely, if $\mathcal{R}$ has a VC dimension ν, then we will be able to find a set of ν points such that $\mathcal{R}$ can pick out every subsets of this set. With the VC dimension, we can then use the equation (9.7) to obtain equation (9.2).

Here are the VC dimension corresponds to the regions we have mentioned

1. $\mathcal{R}_1 = \{(\infty, t] : t \in \mathbb{R}\} \Longrightarrow \nu(\mathcal{R}_1) = 1.$
2. $\mathcal{R}_2 = \{(a,b) : a \leq b\} \Longrightarrow \nu(\mathcal{R}_2) = 2.$
3. $\mathcal{R}_3 = \{[x_0 - r, x_0 + r] : r \geq 0\}$ for some fixed $x_0 \implies \nu(\mathcal{R}_3) = 1$.
4. $\mathcal{R}_4 = \{[x_0 - r, x_0 + r] : r \geq 0, x_0 \in \mathbb{R}\} \Longrightarrow \nu(\mathcal{R}_4) = 2.$
5. $\mathcal{R}_5 = \{(\infty, t] \times (\infty, s] : t, s \in \mathbb{R}\} \Longrightarrow \nu(\mathcal{R}_5) = 1.$
6. $\mathcal{R}_6 = \{(a_1, b_1) \times (a_2, b_2) \times \cdots \times (a_d, b_d) : a_j \leq b_j,\ j = 1, \cdots, d\} \Longrightarrow \nu(\mathcal{R}_6) = 2d.$
7. $\mathcal{R}_7 = \{B(x_0, r) : r \geq 0\}$, where $B(x_0, r) = \{x : \|x - x_0\| \leq r\}$. $\Longrightarrow \nu(\mathcal{R}_7) = 1.$
8. $\mathcal{R}_8 = \{\mathcal{B}_{\beta,c} : \beta \in \mathbb{R}^d, c \in \mathbb{R}\} \Longrightarrow \nu(\mathcal{R}_8) = d + 1.$

Example: uniform convergence of the EDF. Let $\widehat{F}_n(x)$ be the EDF of a CDF $F(x)$ based on an IID random sample from F. Using VC theory, we can prove that

$$\begin{aligned} P(\sup_x |\widehat{F}_n(x) - F(x)| > \epsilon) &= P\left(\sup_{A \in \mathcal{R}_1} \left| \frac{1}{n} \sum_{i=1}^{n} I(X_i \in A) - P(X_1 \in A)) \right| > \epsilon \right) \\ &\leq 8(n+1)^{\nu(\mathcal{R}_1)} e^{-n\epsilon^2/32} \\ &= 8(n+1) e^{-n\epsilon^2/32} \end{aligned}$$

converges to 0 as $n \to \infty$[3]. Moreover, using the method in Section 9.3.2, we have

$$\sup_x |\widehat{F}_n(x) - F(x)| = O_P\left(\sqrt{\frac{\log n}{n}}\right)$$

[3]Note that there is a better bound for the EDF called the DKW (Dvoretsky-Kiefer-Wolfowitz) inequality:

$$P\left(\sup_x |\widehat{F}_n(x) - F(x)| > \epsilon\right) \leq 2e^{-2n\epsilon^2}.$$

and

$$\mathbb{E}\left(\sup_x |\widehat{F}_n(x) - F(x)|\right) = O\left(\sqrt{\frac{\log n}{n}}\right).$$