ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Chair for Mathematical Information Science

Prof. Dr. H. Bölcskei
Sternwartstrasse 7
CH-8092 Zürich

Mathematics of Information

Spring semester 2022

Solutions to Problem Set 12

Problem 1 Sup-norm Continuity.

a) The mean functional is not continuous w.r.t. the sup-norm. As a counterexample, take F to be the CDF of a random variable $X = 0$ a.s., and F_n the CDF of a random variable

$$X_n = \begin{cases} 0, & \text{with probability } 1 - 1/n, \\ n, & \text{with probability } 1/n. \end{cases}$$

We first observe that $\gamma_1(F_n) = 1$ and $\gamma_1(F) = 0$, so that $\gamma_1(F_n)$ does not converge to $\gamma_1(F)$. On the other hand, $\|F_n - F\|_\infty = 1/n$, which implies that, given $\delta > 0$ and for n large enough, $\|F_n - F\|_\infty \leq \delta$. These observations allow us to conclude that the mean functional is not continuous w.r.t. the sup-norm.

b) The Cramér-von Mises functional is continuous w.r.t. the sup-norm. To see that, fix $\epsilon > 0$ and take $\delta \coloneqq \epsilon/4$. Consider F and G two CDFs such that $\|F_1 - F_2\|_\infty \leq \delta$, we have

$$\begin{aligned} |\gamma_2(F_1) - \gamma_2(F_2)| &= \left| \int (F_1(x) - F_0(x))^2 - (F_2(x) - F_0(x))^2 dF_0(x) \right| \\ &= \left| \int (F_1(x) - F_2(x))(F_1(x) + F_2(x) - 2F_0(x)) dF_0(x) \right| \\ &\leq \int |F_1(x) - F_2(x)| \, (F_1(x) + F_2(x) + 2F_0(x)) dF_0(x) \\ &\leq 4\delta = \epsilon. \end{aligned}$$

This proves the continuity of Cramér-von Mises functional w.r.t. the sup-norm.

c) The α-quantile functional is not continuous w.r.t. the sup-norm. As a counterexample, we take $\alpha = 1/2$, we let F be the CDF of the random variable

$$X = \begin{cases} 0, & \text{with probability } 1/2, \\ 1, & \text{with probability } 1/2, \end{cases}$$

and, for $n \geq 2$, we let F_n be the CDF of the random variable

$$X_n = \begin{cases} 0, & \text{with probability } 1/2 - 1/n, \\ n, & \text{with probability } 1/2 + 1/n. \end{cases}$$

We observe that $\gamma_3(F) = 0$ and $\gamma_3(F_n) = 1$, so that $\gamma_3(F_n)$ does not converge to $\gamma_3(F)$. On the other hand, $\|F_n - F\|_\infty = 1/n$, which implies that, given $\delta > 0$ and for n large enough, $\|F_n - F\|_\infty \leq \delta$.

Problem 2 VC dimension.

a) We write $\nu(\mathcal{F})$ for the VC dimension of the class $\mathcal{F}$.

i) Given one point $x_1 \in [0,1]$, choosing $t < x_1$, respectively $t > x_1$, the function $\mathbb{1}_{\cdot<t}$ labels x_1 as 0, respectively 1. Therefore, we have $\nu(\mathcal{F}_1) \geq 1$.

Given two distinct points $x_1 < x_2$ in $[0,1]$, there is no $t \in [0,1]$ such that $\mathbb{1}_{\cdot<t}$ produces the labeling $(0,1)$. Therefore, we have $\nu(\mathcal{F}_1) < 2$.

We conclude that $\nu(\mathcal{F}_1) = 1$.

ii) Given two points $x_1 < x_2$ in $[0,1]$, $\mathbb{1}_{\cdot<t}$ produces the labelings $(0,0)$, $(1,0)$ and $(1,1)$ depending on whether we choose t such that $t < x_1 < x_2$, $x_1 < t < x_2$ or $x_1 < x_2 < t$. Moreover, choosing $x_1 < t < x_2$, the function $1 - \mathbb{1}_{\cdot<t}$ yields the labeling $(0,1)$. Therefore, we have $\nu(\mathcal{F}_2) \geq 2$.

Given three distinct points $x_1 < x_2 < x_3$ in $[0,1]$, there is no function in $\mathcal{F}_2$ producing the labeling $(0,1,0)$. Therefore, we have $\nu(\mathcal{F}_2) < 3$.

We conclude that $\nu(\mathcal{F}_2) = 2$.

iii) Given two points $x_1 < x_2$ in $[0,1]$, $\mathbb{1}_{t_1<\cdot<t_2}$ produces the labelings $(0,0)$, $(1,0)$, $(0,1)$ and $(1,1)$ depending on whether we choose t_1 and t_2 such that $t_2 < x_1 < x_2$, $t_1 < x_1 < t_2 < x_2$, $x_1 < t_1 < x_2 < t_2$ or $x_1 < x_2 < t_1$. Therefore, we have $\nu(\mathcal{F}_3) \geq 2$.

Given three distinct points $x_1 < x_2 < x_3$ in $[0,1]$, there is no function in $\mathcal{F}_3$ producing the labeling $(1,0,1)$. Therefore, we have $\nu(\mathcal{F}_3) < 3$.

We conclude that $\nu(\mathcal{F}_3) = 2$.

iv) Given three distinct points $x_1 < x_2 < x_3$ in $[0,1]$, all the labelings where all the 1 are consecutive (namely all the labelings except $(1,0,1)$) can be obtained by a function of the form $\mathbb{1}_{t_1<\cdot<t_2}$ by choosing appropriate t_1 and t_2. Moreover, the labeling $(1,0,1)$ is obtained by taking $x_1 < t_1 < x_2 < t_2 < x_3$ and considering the function $1 - \mathbb{1}_{t_1<\cdot<t_2}$. Therefore, we have $\nu(\mathcal{F}_4) \geq 3$.

Given four distinct points $x_1 < x_2 < x_3 < x_4$ in $[0,1]$, there is no function in $\mathcal{F}_4$ producing the labeling $(1,0,1,0)$. Therefore, we have $\nu(\mathcal{F}_4) < 4$.

We conclude that $\nu(\mathcal{F}_4) = 3$.

v) Given $2N$ distinct points $x_1 < x_2 < \cdots < x_{2N}$ in $[0,1]$, a labeling contains at most N groups of consecutive 1. Associating to each such group an indicator function $\mathbb{1}_{t_{2n-1}<\cdot<t_{2n}}$, the sum $\sum_{n=1}^{N} \mathbb{1}_{t_{2n-1}<\cdot<t_{2n}}$ produces the desired labeling. Therefore, we have $\nu(\mathcal{F}_5) \geq 2N$.

Given $2N+1$ distinct points $x_1 < x_2 < \cdots < x_{N+1}$ in $[0,1]$, there is no function in $\mathcal{F}_5$ producing the following alternating labeling $(1,0,1,\ldots,0,1)$. Therefore, we have $\nu(\mathcal{F}_5) < 2N+1$.

We conclude that $\nu(\mathcal{F}_5) = 2N$.

vi) Let $n \geq 1$ be an integer.

We consider the set of points $(x_1,\ldots,x_n)$, defined as $x_i \coloneqq 2^{-i}$, with arbitrary labels $(y_1,\ldots,y_n) \in \{0,1\}^n$. Now, choose the parameter $t = \pi\left(1 + \sum_{i=1}^{n} 2^i y_i'\right)$, where $y_i' \coloneqq 1 - y_i$. We show that this single parameter will always correctly classify the entire

sample. For any $j \in [n]$, we have

$$\begin{aligned} tx_j &= t2^{-j} \\ &= \pi \left(2^{-j} + \sum_{i=1}^{n} 2^{i-j} y'_i \right) \\ &= \pi \left(2^{-j} + \sum_{k=1}^{j-1} 2^{-k} y'_{j-k} + y'_j + \sum_{\ell=1}^{n-j} 2^{\ell} y'_{\ell} \right). \end{aligned}$$

The last term can be dropped since it only contributes to multiples of 2π. The remaining term can be bounded as

$$\pi y'_j < \pi \left(2^{-j} + \sum_{k=1}^{j-1} 2^{-k} y'_{j-k} + y'_j \right) \leq \pi \left(\sum_{k=1}^{j} 2^{-k} + y'_j \right) < \pi \left(1 + y'_j \right).$$

Thus, if $y_j = 1$, i.e. $y'_j = 0$, we have $0 < tx_j < \pi$, which implies $\text{sign}(\sin(tx_j)) = 1$. Similarly, for $y_j = 0$, we have $\text{sign}(\sin(tx_j)) = -1$.

We have proven that $\mathcal{F}_6$ shatters a set of n points for any integer $n \geq 1$.

We conclude that $\nu(\mathcal{F}_6) = \infty$.

b) Given a set of n points $\{x_1, \ldots, x_n\}$ and a set S, we say that S yields the labeling $\{y_1, \ldots, y_n\}$ if $y_i = \mathbb{1}_S(x_i)$, for all $i = 1, \ldots, n$. If $y_i = 1$, we say that x_i is labelled positively.

i) The set S^c yields the labeling obtained by flipping all the labels of the labeling generated by the set S. Therefore, all the labelings of a given set of points are realizable by $\mathcal{S}_1^c$ if and only if they are realizable by $\mathcal{S}_1$. This means that $\nu(\mathcal{S}_1^c) = \nu(\mathcal{S}_1) < \infty$.

ii) Consider a set of points P of size n. Note that a point $x \in P$ is labelled positively by $S_1 \cap S_2$ if and only if it is labelled positively by S_1 and by S_2. Therefore, fixing S_1 and letting S_2 take all possible values in $\mathcal{S}_2$ yields at most $|\mathcal{S}_2(P)|$ different labelings. Likewise, fixing S_2 and letting S_1 take all possible values in $\mathcal{S}_1$ yields at most $|\mathcal{S}_1(P)|$ different labelings. Therefore, we have that $|(\mathcal{S}_1 \sqcap \mathcal{S}_2)(P)| \leq |\mathcal{S}_1(P)||\mathcal{S}_2(P)|$.

For sufficiently large n, Sauer-Shelah's lemma yields

$$|(\mathcal{S}_1 \sqcap \mathcal{S}_2)(P)| \leq |\mathcal{S}_1(P)||\mathcal{S}_2(P)| \leq (n+1)^{\nu(\mathcal{S}_1)+\nu(\mathcal{S}_2)} < 2^n,$$

which proves that $\mathcal{S}_1 \sqcap \mathcal{S}_2$ is of finite VC-dimension.

iii) Note that $\mathcal{S}_1 \sqcup \mathcal{S}_2 = (\mathcal{S}_1^c \sqcap \mathcal{S}_2^c)^c$. Then, by *i)* and *ii)*, the set $\mathcal{S}_1 \sqcup \mathcal{S}_2$ has finite VC-dimension.

Problem 3 Failure of uniform law.

a) Since $S \in \mathcal{S}$ is by definition finite and $\mathbb{P}$ has no atom, the probability of the set S is necessarily zero:

$$\mathbb{P}[S] = 0, \quad \forall S \in \mathcal{S},$$

which can be equivalently stated as

$$\mathbb{E}[f(X)] = 0, \quad \forall f \in \mathcal{F}_{\mathcal{S}},$$

where the expectation is taken according to $\mathbb{P}$.

On the other hand, considering the independent random variables $\{X_i\}_{i=1}^n$, we have

$$\sup_{S \in \mathcal{S}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_S(X_i) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_1,\dots,X_n\}}(X_i) = 1.$$

Combining the previous observations, we obtain

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}_{\mathcal{S}}} = \sup_{S \in \mathcal{S}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_S(X_i) - \mathbb{E}[f(X)] \right| = 1,$$

which does not converge to zero as the sample size grows.

b) Recall the definition of the Rademacher complexity

$$\mathcal{R}_n(\mathcal{F}_{\mathcal{S}}) = \mathbb{E}_{X,\varepsilon} \left[\sup_{S \in \mathcal{S}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \mathbb{1}_S(X_i) \right| \right].$$

Note that for any realization of X and of the Rademacher random variables, there is a set $S \in \mathcal{S}$ such that

$$\mathbb{1}_S(X_i) \coloneqq \begin{cases} 1 & \text{if } \varepsilon_i = +1, \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, there is a set $S' \in \mathcal{S}$ such that $\mathbb{1}_{S'}(X_i) = 1 - \mathbb{1}_S(X_i)$. Letting $N_+(\varepsilon)$ denote the number of $+1$ in the Rademacher sequence $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$, we have

$$\mathcal{R}_n(\mathcal{F}_{\mathcal{S}}) \geq \mathbb{E}_{\varepsilon} \left[\max \left\{ \frac{N_+(\varepsilon)}{n}, 1 - \frac{N_+(\varepsilon)}{n} \right\} \right] \geq \frac{1}{2}.$$

Problem 4 VC dim. of rectangles (Winter Exam 2020, Problem 4).

a) Consider the points $x_1 = 0$ and $x_2 = 1$ in $\mathbb{R}$. The four possible labelings

$$\begin{cases} h_{(-1,-1)}(x_1) = 0 & \text{and} \quad h_{(-1,-1)}(x_2) = 0, \\ h_{(0,0)}(x_1) = 1 & \text{and} \quad h_{(0,0)}(x_2) = 0, \\ h_{(1,1)}(x_1) = 0 & \text{and} \quad h_{(1,1)}(x_2) = 1, \\ h_{(0,1)}(x_1) = 1 & \text{and} \quad h_{(0,1)}(x_2) = 1, \end{cases}$$

are produced by $\mathcal{H}_1$. Therefore, there is a set of 2 points shattered by $\mathcal{H}_1$, which implies $\dim_{VC}(\mathcal{H}_1) \geq 2$.

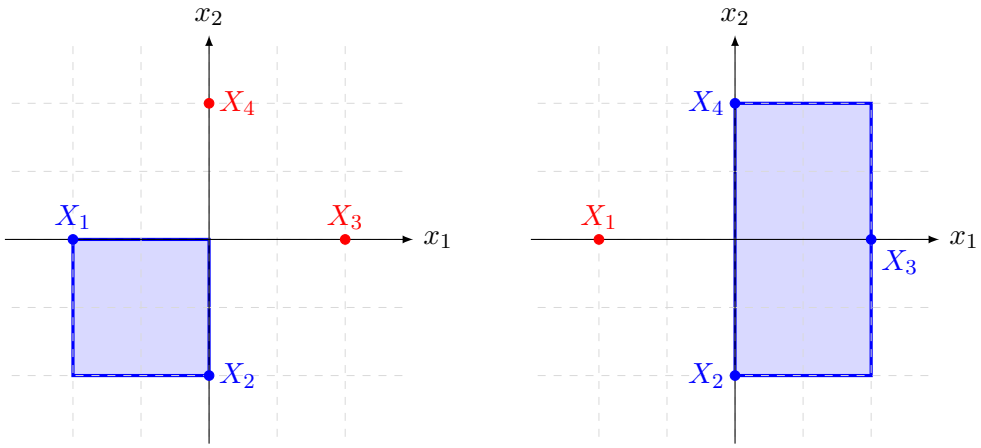
On the other hand, for any set of three distinct points x_1, x_2, and x_3 that we choose without loss of generality such that $x_1 < x_2 < x_3$, there is no closed interval containing x_1 and x_3 but not x_2. Therefore, there is no set of 3 points shattered by $\mathcal{H}_1$, which implies $\dim_{VC}(\mathcal{H}_1) < 3$.

We have therefore proven that $\dim_{VC}(\mathcal{H}_1) = 2$.

b) i) The 4 points $X_1 = (-1, 0)$, $X_2 = (0, -1)$, $X_3 = (1, 0)$, and $X_4 = (0, 1)$ are shattered by $\mathcal{H}_2$. To see this, we fix a labeling $(y_1, y_2, y_3, y_4) \in \{0, 1\}^4$. The rectangle $h_{(-y_1,-y_2,y_3,y_4)} \in \mathcal{H}_2$ labels correctly all four points:

$$\begin{cases} h_{(-y_1,-y_2,y_3,y_4)}(X_1) = y_1, \\ h_{(-y_1,-y_2,y_3,y_4)}(X_2) = y_2, \\ h_{(-y_1,-y_2,y_3,y_4)}(X_3) = y_3, \\ h_{(-y_1,-y_2,y_3,y_4)}(X_4) = y_4. \end{cases}$$

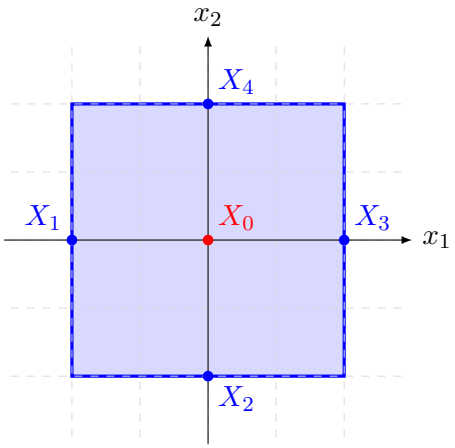
The figures below picture the rectangle $h_{(-1,-1,0,0)}$ (left) and the rectangle $h_{(0,-1,1,1)}$ (right).



ii) Assume that we can find a_1, b_1, a_2, and b_2 such that

$$\begin{cases} h_{(a_1,a_2,b_1,b_2)}(0,0) = 0, \\ h_{(a_1,a_2,b_1,b_2)}(-1,0) = h_{(a_1,a_2,b_1,b_2)}(1,0) = 1, \\ h_{(a_1,a_2,b_1,b_2)}(0,-1) = h_{(a_1,a_2,b_1,b_2)}(0,1) = 1. \end{cases}$$

Since $h_{(a_1,a_2,b_1,b_2)}(-1,0) = h_{(a_1,a_2,b_1,b_2)}(1,0) = 1$, we must have $a_1 \leq -1$ and $1 \leq b_1$, which implies $a_1 \leq 0 \leq b_1$. Likewise, since $h_{(a_1,a_2,b_1,b_2)}(0,-1) = h_{(a_1,a_2,b_1,b_2)}(0,1) = 1$, we must have $a_2 \leq -1$ and $1 \leq b_2$, which implies $a_2 \leq 0 \leq b_2$. The inequalities $a_1 \leq 0 \leq b_1$ and $a_2 \leq 0 \leq b_2$ together imply $h_{(a_1,a_2,b_1,b_2)}(0,0) = 1$, which yields a contradiction, thereby concluding the proof.



iii) Take any set of five distinct points in $\mathbb{R}^2$. We call X_1 the "leftmost" point (the point of smallest first coordinate x_1^-), X_2 the "lowest" point (the point of smallest second coordinate x_2^-), X_3 the "rightmost" point (the point of largest first coordinate x_1^+) and X_4 the "highest" point (the point of largest second coordinate x_2^+). Note that these extremal points X_1, X_2, X_3, and X_4 are not necessarily distinct (e.g., there could be a point of both largest first and largest second coordinates, i.e., $X_3 = X_4$). We consider the labeling y that assigns 1 to X_1, X_2, X_3 and X_4 and 0 to a point X_0 distinct from X_1, X_2, X_3, and X_4 (which exists since we consider 5 points in total). If it exists, a rectangle $h_{(a_1,a_2,b_1,b_2)}$ realizing the desired labeling has to be such that $a_1 \leq x_1^-$, $x_1^+ \leq b_1$, $a_2 \leq x_2^-$, and $x_2^+ \leq b_2$. Since by construction of X_1, X_2, X_3 and X_4, we must have $x_1^- \leq x_1^0 \leq x_1^+$ and $x_2^- \leq x_2^0 \leq x_2^+$, we necessarily get $a_1 \leq x_1^0 \leq b_1$ and $a_2 \leq x_2^0 \leq b_2$, which does not produce the correct labeling for the point X_0. This contradiction proves that there is no rectangle $h_{(a_1,a_2,b_1,b_2)}$ yielding the desired labelling y. Therefore, no set of 5 points can be shattered by $\mathcal{H}_2$.

iv) The class $\mathcal{H}_2$ shatters *a* set of 4 points but does not shatter *any* set of 5 points. By definition of VC dimension, we therefore have $\dim_{VC}(\mathcal{H}_2) = 4$.

c) We will prove that $\dim_{VC}(\mathcal{H}_d) = 2d$. We apply the same procedure as in (b). We first prove that the set of $2d$ points $\{X_i^+, X_i^-\}_{i=1}^d$ is shattered by $\mathcal{H}_d$, where X_i^σ has all coordinates equal to zero, except its i-th coordinate which is equal to the sign σ:

$$X_i^+ = (0,\dots,0,\underbrace{+1}_{i\text{-th coordinate}},0,\dots,0) \quad \text{and} \quad X_i^- = (0,\dots,0,\underbrace{-1}_{i\text{-th coordinate}},0,\dots,0).$$

Fix a labelling $(y_{X_1^+}, y_{X_1^-}, \dots, y_{X_d^+}, y_{X_d^-})$. The rectangle $h_{(-y_{X_1^-},\dots,-y_{X_d^-},y_{X_1^+},\dots,y_{X_d^+})}$ labels

correctly all the $2d$ points.

We now prove that no set S of $2d+1$ distinct points can be shattered by $\mathcal{H}_d$. Such a set S contains a subset $S_X = \{X_i^+, X_i^-\}_{i=1}^d$ of $2d$ (non-necessarily distinct) points where X_i^+ is the point of largest coordinate i in S, call it x_i^+, and X_i^- is the point of lowest coordinate i in S, call it x_i^-. Consider the labelling y assigning 1 to elements of S_X and 0 to other elements. If it exists, a rectangle $h_{(a_1,\dots,a_d,b_1,\dots,b_d)}$ yielding the desired labelling should be such that $a_i \leq x_i^-$ and $x_i^+ \leq b_i$. We now take a point $X_0 = (x_0^1, \dots, x_0^d)$ in S but not in S_X (which exists since S has $2d+1$ elements and S_X has at most $2d$ elements). Since, by definition of S_X, we have $x_i^- \leq x_0^i \leq x_i^+$, we necessarily have $a_i \leq x_0^i \leq b_i$, for $i = 1, \dots, d$, which does not yield the correct labelling for the point X_0. This proves that there is no rectangle $h_{(a_1,\dots,a_d,b_1,\dots,b_d)}$ yielding the labelling y. Therefore, no set of $2d+1$ points can be shattered by $\mathcal{H}_d$.

The class $\mathcal{H}_d$ shatters *a* set of $2d$ points but does not shatter *any* set of $2d+1$ points. By definition of the VC dimension, we have $\dim_{VC}(\mathcal{H}_d) = 2d$.

Problem 5 Rademacher complexity (Exam 2020, Problem 3).

a) The solution to this problem follows the same structure as the first step of the proof of Theorem 12.10 in the lecture notes. We first prove that the empirical Rademacher complexity satisfies the bounded difference property. To this end, we start by noting that the Rademacher complexity is invariant under permutation of its inputs so that it is sufficient to establish that

$$\left|\mathcal{R}_n\left(\mathcal{F}\left(x_1^n\right)/n\right) - \mathcal{R}_n\left(\mathcal{F}\left(y_1^n\right)/n\right)\right| \leq L, \tag{1}$$

where L is some positive constant and where we defined $x_1^n \coloneqq \{x_1, \dots, x_n\}$ with $x_i \in \mathbb{R}^d$, $i = 1, \dots, n$, and $y_1^n \coloneqq \{y_1, x_2, \dots, x_n\}$ with $y_1 \in \mathbb{R}^d$. Developing the left-hand side of (1), we get

$$\begin{aligned}
&\left|\mathcal{R}_n\left(\mathcal{F}\left(x_1^n\right)/n\right) - \mathcal{R}_n\left(\mathcal{F}\left(y_1^n\right)/n\right)\right| \\
&= \left|\frac{1}{n}\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \varepsilon_i f(x_i)\right|\right] - \frac{1}{n}\mathbb{E}_\varepsilon\left[\sup_{g\in\mathcal{F}}\left|\varepsilon_1 g(y_1) + \sum_{i=2}^n \varepsilon_i g(x_i)\right|\right]\right| \\
&= \left|\frac{1}{n}\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \varepsilon_i f(x_i)\right| - \sup_{g\in\mathcal{F}}\left|\varepsilon_1 g(y_1) + \sum_{i=2}^n \varepsilon_i g(x_i)\right|\right]\right| \\
&\leq \left|\frac{1}{n}\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left(\left|\sum_{i=1}^n \varepsilon_i f(x_i)\right| - \left|\varepsilon_1 f(y_1) + \sum_{i=2}^n \varepsilon_i f(x_i)\right|\right)\right]\right| \\
&\leq \frac{1}{n}\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left(\left|\sum_{i=1}^n \varepsilon_i f(x_i) - \varepsilon_1 f(y_1) - \sum_{i=2}^n \varepsilon_i f(x_i)\right|\right)\right] \\
&= \frac{1}{n}\mathbb{E}_{\varepsilon_1}\left[\sup_{f\in\mathcal{F}}\left|\varepsilon_1\left(f(x_1) - f(y_1)\right)\right|\right] \\
&\leq \frac{2b}{n},
\end{aligned}$$

which proves that the empirical Rademacher complexity, seen as a function of fixed data points satisfies the bounded difference property with $L \coloneqq \frac{2b}{n}$. The conditions are now satisfied for the application of the bounded difference inequality. For $\delta > 0$, we set $\epsilon \coloneqq \sqrt{\frac{2b^2 \log(1/\delta)}{n}}$ and get

$$\begin{aligned}
&\mathbb{P}\left[\mathbb{E}[\mathcal{R}_n\left(\mathcal{F}\left(X_1^n\right)/n\right)] - \mathcal{R}_n\left(\mathcal{F}\left(X_1^n\right)/n\right) > \sqrt{\frac{2b^2\log(1/\delta)}{n}}\right] \\
&\leq e^{-\frac{2n}{4b^2}\frac{2b^2\log(1/\delta)}{n}} \\
&= \delta.
\end{aligned}$$

For the complementary event, we therefore have with probability at least $1-\delta$ that

$$\mathcal{R}_n\left(\mathcal{F}\right) = \mathbb{E}[\mathcal{R}_n\left(\mathcal{F}\left(X_1^n\right)/n\right)] \leq \mathcal{R}_n\left(\mathcal{F}\left(X_1^n\right)/n\right) + \sqrt{\frac{2b^2\log(1/\delta)}{n}}.$$

b) We rewrite the function class $\mathcal{F}_{|\cdot|}$ as

$$\mathcal{F}_{|\cdot|} = \phi \circ \{\mathcal{F}_1 - \mathcal{F}_2\},$$

where $\mathcal{F}_1 - \mathcal{F}_2 = \{f_1 - f_2 \mid f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}$ and $\phi\colon x \mapsto |x|$ is a 1-Lipschitz function from $\mathbb{R}$ to $\mathbb{R}$ such that $\phi(0) = 0$. Applying the Ledoux-Talagrand contraction lemma now yields

$$\begin{aligned}
\mathcal{R}_n\left(\mathcal{F}_{|\cdot|}\left(x_1^n\right)/n\right) &= \mathcal{R}_n\left(\phi \circ \{\mathcal{F}_1 - \mathcal{F}_2\}\left(x_1^n\right)/n\right) \\
&\leq 2\,\mathcal{R}_n\left(\{\mathcal{F}_1 - \mathcal{F}_2\}\left(x_1^n\right)/n\right) \\
&= \frac{2}{n}\,\mathbb{E}_\varepsilon\left[\sup_{f_1,f_2\in\mathcal{F}_1\times\mathcal{F}_2}\left|\sum_{i=1}^n \varepsilon_i\left(f_1(x_i) - f_2(x_i)\right)\right|\right] \\
&\leq \frac{2}{n}\,\mathbb{E}_\varepsilon\left[\sup_{f_1\in\mathcal{F}_1}\left|\sum_{i=1}^n \varepsilon_i f_1(x_i)\right|\right] + \frac{2}{n}\,\mathbb{E}_\varepsilon\left[\sup_{f_2\in\mathcal{F}_2}\left|\sum_{i=1}^n \varepsilon_i f_2(x_i)\right|\right] \\
&= 2\,\mathcal{R}_n\left(\mathcal{F}_1\left(x_1^n\right)/n\right) + 2\,\mathcal{R}_n\left(\mathcal{F}_2\left(x_1^n\right)/n\right).
\end{aligned}$$

c) We first prove that the maximum of two real numbers a and b can be expressed as

$$\max\{a,b\} = \frac{|a-b| + a + b}{2}. \tag{2}$$

If $a \geq b$, then

$$\max\{a,b\} = a = \frac{a-b+a+b}{2} = \frac{|a-b|+a+b}{2},$$

and if $b \geq a$, then

$$\max\{a,b\} = b = \frac{b-a+a+b}{2} = \frac{|a-b|+a+b}{2},$$

which establishes the hint. Inserting (2) into the definition of the empirical Rademacher

complexity of $\mathcal{F}_{\max}$, we obtain

$$\begin{aligned}
\mathcal{R}_n\left(\mathcal{F}_{\max}\left(x_1^n\right)/n\right) &= \frac{1}{n}\mathbb{E}_\varepsilon\left[\sup_{f_1,f_2\in\mathcal{F}_1\times\mathcal{F}_2}\left|\sum_{i=1}^n \varepsilon_i \max\{f_1(x_i), f_2(x_i)\}\right|\right] \\
&= \frac{1}{n}\mathbb{E}_\varepsilon\left[\sup_{f_1,f_2\in\mathcal{F}_1\times\mathcal{F}_2}\left|\sum_{i=1}^n \varepsilon_i \frac{|f_1(x_i)-f_2(x_i)| + f_1(x_i) + f_2(x_i)}{2}\right|\right] \\
&\leq \frac{\mathcal{R}_n\left(\mathcal{F}_{|\cdot|}\left(x_1^n\right)/n\right) + \mathcal{R}_n\left(\mathcal{F}_1\left(x_1^n\right)/n\right) + \mathcal{R}_n\left(\mathcal{F}_2\left(x_1^n\right)/n\right)}{2} \\
&\leq \frac{3}{2}\Big(\mathcal{R}_n\left(\mathcal{F}_1\left(x_1^n\right)/n\right) + \mathcal{R}_n\left(\mathcal{F}_2\left(x_1^n\right)/n\right)\Big),
\end{aligned}$$

where the last inequality follows from the bound established in subproblem (b).

Problem 6 Metric entropy and VC dim. (Exam 2020, Problem).

a) By construction M is the size of a maximal δ-packing of $\mathcal{F}_{\mathcal{S}}$ in the $L_1(\mathbb{Q})$-norm, i.e.,

$$M \coloneqq M(\delta; \mathcal{F}_{\mathcal{S}}, L_1(\mathbb{Q}))\,.$$

We know from the lecture that, for any metric space $(\mathbb{T}, \rho)$, the following relation between the δ-packing number and the δ-covering number holds

$$N(\delta; \mathbb{T}, \rho) \leq M(\delta; \mathbb{T}, \rho)\,.$$

Direct application of this result therefore yields

$$N(\delta; \mathcal{F}_{\mathcal{S}}, L_1(\mathbb{Q})) \leq M.$$

b) Using that $\{\mathbb{1}_{S_1}, \ldots, \mathbb{1}_{S_M}\}$ is a δ-packing, we derive

$$\begin{aligned}
\mathbb{P}\left[X \notin (S_i \triangle S_j)\right] &= 1 - \mathbb{P}\left[X \in (S_i \triangle S_j)\right] \\
&= 1 - \mathbb{E}\left[\mathbb{1}_{S_i \triangle S_j}(X)\right] \\
&= 1 - \mathbb{E}\left[\left|\mathbb{1}_{S_i}(X) - \mathbb{1}_{S_j}(X)\right|\right] \\
&< 1 - \delta.
\end{aligned}$$

c) $S_i \cap \{X_1, \ldots, X_n\}$ and $S_j \cap \{X_1, \ldots, X_n\}$ are distinct if and only if there exists k such that either $X_k \in S_i$ and $X_k \notin S_j$ or $X_k \in S_j$ and $X_k \notin S_i$, which is equivalent to $X_k \in (S_i \triangle S_j)$.

d) From subproblem (b) we know that

$$\mathbb{P}\left[X_k \notin (S_i \triangle S_j)\right] < 1 - \delta, \quad \forall k = 1, \ldots, n,$$

and by the independence assumption on the X_k, we get

$$\mathbb{P}[X_k \notin (S_i \triangle S_j),\ \forall k = 1, \ldots, n] < (1-\delta)^n.$$

Taking the union bound over all $\binom{M}{2}$ pairs of subsets, the probability that there exists at least one pair of subsets $\{S_i, S_j\}$ such that $S_i \cap \{X_1, \ldots, X_n\}$ and $S_j \cap \{X_1, \ldots, X_n\}$ are identical is upper bounded by $\binom{M}{2}(1-\delta)^n$. We are interested in the complementary event, which hence has probability lower-bounded by $1 - \binom{M}{2}(1-\delta)^n$.

e) Set $n = \frac{3\log(M)}{\delta} - 1$. Using the result in subproblem (d), the probability that every set S_i picks out a different subset of $\{X_1, \ldots, X_n\}$ is lower-bounded according to

$$\begin{aligned} 1 - \binom{M}{2}(1-\delta)^n &= 1 - \frac{M(M-1)}{2}(1-\delta)^n \\ &\geq 1 - M^2 e^{-n\delta} \\ &= 1 - e^{-n\delta + 2\log(M)} \\ &= 1 - e^{-\log(M) + \delta} \\ &> 0, \end{aligned}$$

where the last inequality is a consequence of

$$\log(M) > \frac{\delta(\nu+1)}{3} \geq \delta,$$

which, in turn, follows from $\nu \geq 2$. As $1 - \binom{M}{2}(1-\delta)^n > 0$, we can conclude that there must exist a set of n points $\{x_1, \ldots, x_n\}$ such that

$$M \leq |\mathcal{S}(\{x_1, \ldots, x_n\})|, \tag{3}$$

as desired.

f) As $3\log(M) > \delta(\nu+1)$, we have $n > \nu$. One can therefore apply the Vapnik-Chervonenkis-Sauer-Shelah lemma, which yields

$$|\mathcal{S}(\{x_1, \ldots, x_n\})| \leq (n+1)^\nu = \left(\frac{3\log(M)}{\delta}\right)^\nu. \tag{4}$$

Combining (3) and (4) yields

$$M \leq \left(\frac{3\log(M)}{\delta}\right)^\nu. \tag{5}$$

g) We first prove the relation given in the hint, namely

$$\sup_{t \geq 0} \left(t^{2\nu} e^{-t}\right) \leq (2\nu)^{2\nu-1}.$$

By differentiating the function $t \mapsto t^{2\nu} e^{-t}$, we find that it attains its maximum at $t^* = 2\nu$. Therefore,

$$\sup_{t \geq 0} \left(t^{2\nu} e^{-t}\right) \leq (2\nu)^{2\nu} e^{-2\nu} \leq (2\nu)^{2\nu-1},$$

where in the last inequality we used that $te^{-t} \leq 1$, for $t \geq 0$. Now, we rewrite inequality (5) in the previous subproblem as

$$M^2 \leq \left(\frac{3\log(M)}{\delta}\right)^{2\nu},$$

which, upon application of $t^{2\nu}e^{-t} \leq (2\nu)^{2\nu-1}$, for $t \geq 0$, with $t = \log(M)$, yields

$$M \leq \frac{(\log(M))^{2\nu}}{M} \left(\frac{3}{\delta}\right)^{2\nu} \leq (2\nu)^{2\nu-1} \left(\frac{3}{\delta}\right)^{2\nu}.$$

Using the bound

$$N(\delta; \mathcal{F}_{\mathcal{S}}, L_1(\mathbb{Q})) \leq M$$

established in the first subproblem, we obtain the desired result:

$$N(\delta; \mathcal{F}_{\mathcal{S}}, L_1(\mathbb{Q})) \leq (2\nu)^{2\nu-1} \left(\frac{3}{\delta}\right)^{2\nu}.$$