Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Prof. Dr. Helmut Bölcskei
Chair for Mathematical Information Science

Solutions to the **Examination on Mathematics of Information** **August 14, 2020**

Problem 1

(a) (i) We have

$$\sum_{j=1}^{N} \tilde{\mu}_j^2 = \sum_{j=1}^{N} \max_{k\in[N]} |\langle g_k, h_j\rangle|^2 \geq \sum_{j=1}^{N} |\langle g_1, h_j\rangle|^2 = \|g_1\|_2^2 = 1,$$

where we used the fact that $\{h_j\}_{j=1}^N$ is an orthonormal basis for $\mathbb{C}^N$.

(ii) Let $\boldsymbol{t}$ be a random variable taking values in $[N] := \{1, 2, \dots, N\}$ and distributed according to $\widetilde{\mathcal{P}}$. We then have

$$\kappa^{-1} = \min_{j\in[N]} p_j\, \tilde{\mu}_j^{-2} \leq p_{\boldsymbol{t}}\, \tilde{\mu}_{\boldsymbol{t}}^{-2}.$$

Taking the expectation of both sides with respect to $\boldsymbol{t}$ yields $\kappa^{-1} \leq \mathbb{E}_{\boldsymbol{t}\sim\widetilde{\mathcal{P}}}\left[p_{\boldsymbol{t}}\, \tilde{\mu}_{\boldsymbol{t}}^{-2}\right]$.

(iii) We have

$$\begin{aligned}\kappa^{-1} &\leq \mathbb{E}_{\boldsymbol{t}\sim\widetilde{\mathcal{P}}}\left[p_{\boldsymbol{t}}\, \tilde{\mu}_{\boldsymbol{t}}^{-2}\right] = \sum_{j=1}^{N} \mathbb{P}_{\boldsymbol{t}\sim\widetilde{\mathcal{P}}}\left(\boldsymbol{t} = j\right) \cdot p_j \tilde{\mu}_j^{-2} \\ &= \sum_{j=1}^{N} \frac{\tilde{\mu}_j^2}{\tilde{\mu}_1^2 + \cdots + \tilde{\mu}_N^2} p_j \tilde{\mu}_j^{-2} = \frac{1}{\tilde{\mu}_1^2 + \cdots + \tilde{\mu}_N^2} \underbrace{\sum_{j=1}^{N} p_j}_{=1} = \frac{1}{\tilde{\mu}_1^2 + \cdots + \tilde{\mu}_N^2},\end{aligned}$$

and hence $\kappa \geq \sum_{j=1}^N \tilde{\mu}_j^2$. Moreover, equality holds if and only if

$$\min_{j\in[N]} p_j\, \tilde{\mu}_j^{-2} = \mathbb{E}_{\boldsymbol{t}\sim\widetilde{\mathcal{P}}}\left[p_{\boldsymbol{t}} \tilde{\mu}_{\boldsymbol{t}}^{-2}\right].$$

This is the case if and only if the random variable $p_{\boldsymbol{t}}\, \tilde{\mu}_{\boldsymbol{t}}^{-2}$ is constant, i.e., there exists a $c \in \mathbb{R}$ such that $p_j\, \tilde{\mu}_j^{-2} = c$, for $j \in [N]$. If this is the case, then, using that $\mathcal{P}$ is a probability distribution, we have

$$c \sum_{k=1}^{N} \tilde{\mu}_k^2 = \sum_{k=1}^{N} p_k = 1,$$

and thus $p_j = c\tilde{\mu}_j^2 = \left(\sum_{k=1}^N \tilde{\mu}_k^2\right)^{-1} \tilde{\mu}_j^2$, for all $j \in [N]$, i.e., $\mathcal{P} = \widetilde{\mathcal{P}}$.

(b) (i) We have

$$\begin{aligned}
\mathbb{E}\left[|\langle x, Y_\ell\rangle|^2\right] &= \mathbb{E}\left[\langle x, Y_\ell\rangle\overline{\langle x, Y_\ell\rangle}\right] = \mathbb{E}\left[\sum_{k,k'=1}^N A_{\ell k}x_k\overline{A_{\ell k'}x_{k'}}\right] \\
&= \sum_{k,k'=1}^N x_k\overline{x_{k'}}\, m^{-1}\mathbb{E}_{\boldsymbol{t}\sim\mathcal{P}}\left[p_{\boldsymbol{t}}^{-1}\langle g_k, h_{\boldsymbol{t}}\rangle\overline{\langle g_{k'}, h_{\boldsymbol{t}}\rangle}\right] \\
&= \sum_{k,k'=1}^N x_k\overline{x_{k'}}\, m^{-1}\sum_{j=1}^N p_j\cdot p_j^{-1}\langle g_k, h_j\rangle\overline{\langle g_{k'}, h_j\rangle} \\
&= \sum_{k,k'=1}^N x_k\overline{x_{k'}}\, m^{-1}\left\langle\sum_{j=1}^N\langle g_k, h_j\rangle h_j, g_{k'}\right\rangle \\
&= \sum_{k,k'=1}^N x_k\overline{x_{k'}}\, m^{-1}\underbrace{\langle g_k, g_{k'}\rangle}_{=\delta_{kk'}} = m^{-1}\|x\|_2^2 = m^{-1},
\end{aligned}$$

and so $\mathbb{E}[X_\ell] = \mathbb{E}[|\langle x, Y_\ell\rangle|^2] - \frac{1}{m} = 0$. Next, for $\ell\in[m]$, applying the Cauchy-Schwarz inequality and $|\langle g_j, h_{\boldsymbol{t}_\ell}\rangle| \le \tilde{\mu}_{\boldsymbol{t}_\ell}$ yields

$$\begin{aligned}
|\langle x, Y_\ell\rangle|^2 &\le \Big(\sum_{j\in\mathrm{supp}(x)}|A_{\ell j}|^2\Big)\|x\|_2^2 = \sum_{j\in\mathrm{supp}(x)} m^{-1}p_{\boldsymbol{t}_\ell}^{-1}|\langle g_j, h_{\boldsymbol{t}_\ell}\rangle|^2 \\
&\le \sum_{j\in\mathrm{supp}(x)} m^{-1}p_{\boldsymbol{t}_\ell}^{-1}\tilde{\mu}_{\boldsymbol{t}_\ell}^2 = sm^{-1}p_{\boldsymbol{t}_\ell}^{-1}\tilde{\mu}_{\boldsymbol{t}_\ell}^2 \le sm^{-1}\kappa,
\end{aligned}$$

and so

$$|X_\ell| \le \max\left\{|\langle x, Y_\ell\rangle|^2, \tfrac{1}{m}\right\} \le m^{-1}\max\{s\kappa, 1\} = sm^{-1}\kappa,$$

where the second inequality follows from $\kappa \ge \sum_{j=1}^N \tilde{\mu}_j^2 \ge 1$. Finally, we estimate

$$\begin{aligned}
\mathbb{E}\left[|X_\ell|^2\right] &= \mathbb{E}\left[|\langle x, Y_\ell\rangle|^4\right] - \tfrac{2}{m}\,\mathbb{E}\left[|\langle x, Y_\ell\rangle|^2\right] + \tfrac{1}{m^2} \\
&\le sm^{-1}\kappa\,\mathbb{E}\left[|\langle x, Y_\ell\rangle|^2\right] - \tfrac{2}{m}\,\mathbb{E}\left[|\langle x, Y_\ell\rangle|^2\right] + \tfrac{1}{m^2} \\
&= sm^{-1}\kappa\tfrac{1}{m} - \tfrac{2}{m}\tfrac{1}{m} + \tfrac{1}{m^2} < sm^{-2}\kappa,
\end{aligned}$$

for all $\ell\in[m]$, as desired.

(ii) We have

$$\|Ax\|_2^2 - 1 = \sum_{\ell=1}^m |\langle x, Y_\ell\rangle|^2 \; - 1 = \sum_{\ell=1}^m X_\ell,$$

and thus obtain using Bernstein's inequality

$$\begin{aligned}
\mathbb{P}\left(\big|\|Ax\|_2^2 - 1\big| \ge t\right) &= \mathbb{P}\left(\left|\sum_{\ell=1}^m X_\ell\right| \ge t\right) \le 2\exp\left(-\frac{t^2/2}{m\cdot sm^{-2}\kappa + sm^{-1}\kappa\frac{t}{3}}\right) \\
&\le 2\exp\left(-\frac{t^2/2}{sm^{-1}\kappa + \frac{1}{3}sm^{-1}\kappa}\right) = 2\exp\left(-\frac{3t^2m}{8s\kappa}\right),
\end{aligned}$$

for $t \in (0,1)$, as desired.

Problem 2

(a) (i) We have

$$\begin{aligned}\|x\|_1 &= \sum_{k=1}^{N} |x_k| = \sum_{k\in S} |x_k| = \sum_{k\in S} x_k \cdot \frac{\overline{x_k}}{|x_k|} \\ &= \sum_{k\in S} x_k\, \overline{\big(\operatorname{sgn}(x)\big)_k} = \sum_{k=1}^{N} x_k\, \overline{\big(\operatorname{sgn}(x)\big)_k} = \langle x, \operatorname{sgn}(x)\rangle,\end{aligned}$$

as desired.

(ii) Using the triangle inequality, the Hölder inequality, and $\|z\|_1 = \|z_S\|_1 + \|z_{\overline{S}}\|_1$, we obtain

$$\begin{aligned}\|x\|_1 &= \langle x, \operatorname{sgn}(x)\rangle = \langle z+v, \operatorname{sgn}(x)\rangle \\ &\le |\langle z, \operatorname{sgn}(x)\rangle| + |\langle v, \operatorname{sgn}(x)\rangle| = |\langle z_S, \operatorname{sgn}(x_S)\rangle| + |\langle v, \operatorname{sgn}(x)\rangle| \\ &\le \|z_S\|_1 \underbrace{\|\operatorname{sgn}(x_S)\|_\infty}_{=1} + |\langle v, \operatorname{sgn}(x)\rangle| = \|z\|_1 - \|z_{\overline{S}}\|_1 + \ |\langle v, \operatorname{sgn}(x)\rangle| \\ &= \|z\|_1 - \|v_{\overline{S}}\|_1 + \ |\langle v, \operatorname{sgn}(x)\rangle|.\end{aligned}$$

(b) (i) We have

$$0 = Av = \sum_{k=1}^{N} \boldsymbol{a}_k v_k = \sum_{k\in S} \boldsymbol{a}_k v_k + \sum_{k\in\overline{S}} \boldsymbol{a}_k v_k = A_S v_S + A_{\overline{S}}\, v_{\overline{S}}.$$

(ii) First note that

$$A_S v_S = -A_{\overline{S}}\, v_{\overline{S}} = -\sum_{k\in\overline{S}} \boldsymbol{a}_k v_k.$$

Using this, the definition of the matrix operator norm, the triangle inequality, and the definition of α, we can argue as follows

$$\begin{aligned}\|v_S\|_2 &= \|(A_S^{\mathsf{H}} A_S)^{-1} A_S^{\mathsf{H}} A_S v_S\| \le \|(A_S^{\mathsf{H}} A_S)^{-1}\|_2 \|A_S^{\mathsf{H}} A_S v_S\|_2 \\ &= \|(A_S^{\mathsf{H}} A_S)^{-1}\|_2 \Big\| A_S^{\mathsf{H}} \sum_{k\in\overline{S}} \boldsymbol{a}_k v_k \Big\|_2 = \|(A_S^{\mathsf{H}} A_S)^{-1}\|_2 \Big\| \sum_{k\in\overline{S}} A_S^{\mathsf{H}} \boldsymbol{a}_k v_k \Big\|_2 \\ &\le \|(A_S^{\mathsf{H}} A_S)^{-1}\|_2 \sum_{k\in\overline{S}} \|A_S^{\mathsf{H}} \boldsymbol{a}_k\|_2 |v_k| \\ &\le \|(A_S^{\mathsf{H}} A_S)^{-1}\|_2 \cdot \max_{k\in\overline{S}} \|A_S^{\mathsf{H}} \boldsymbol{a}_k\|_2 \cdot \sum_{k\in\overline{S}} |v_k| \\ &= \alpha \|v_{\overline{S}}\|_1,\end{aligned}$$

yielding the desired identity.

(c) As $v \in \ker(A)$ and $u \in \operatorname{im}(A^{\mathsf{H}}) = (\ker(A))^{\perp}$, we have $\langle v, u \rangle = 0$, and so

$$\begin{aligned}
|\langle v, \operatorname{sgn}(x)\rangle| = |\langle v, \operatorname{sgn}(x) - u\rangle| &= |\langle v_S, \operatorname{sgn}(x_S) - u_S\rangle + \langle v_{\overline{S}}, \underbrace{\operatorname{sgn}(x_{\overline{S}})}_{=0} - u_{\overline{S}}\rangle| \\
&\leq |\langle v_S, \operatorname{sgn}(x_S) - u_S\rangle| + |\langle v_{\overline{S}}, -u_{\overline{S}}\rangle| \\
&\leq \|u_S - \operatorname{sgn}(x_S)\|_2 \|v_S\|_2 + \|u_{\overline{S}}\|_\infty \|v_{\overline{S}}\|_1,
\end{aligned}$$

where in the last step we used the Cauchy-Schwarz and the Hölder inequality.

(d) (i) Let $v \in \ker(A) \setminus \{0\}$. We now have

$$\begin{aligned}
|\langle v, \operatorname{sgn}(x)\rangle| &\leq \|\operatorname{sgn}(x_S) - u_S\|_2 \|v_S\|_2 + \|u_{\overline{S}}\|_\infty \|v_{\overline{S}}\|_1 \\
&\leq \|\operatorname{sgn}(x_S) - u_S\|_2 \cdot \alpha \|v_{\overline{S}}\|_1 + \|u_{\overline{S}}\|_\infty \|v_{\overline{S}}\|_1 \\
&= \big(\alpha \, \|\operatorname{sgn}(x_S) - u_S\|_2 + \|u_{\overline{S}}\|_\infty\big) \cdot \|v_{\overline{S}}\|_1,
\end{aligned}$$

and hence the desired inequality follows by the condition

$$\alpha \, \|\operatorname{sgn}(x_S) - u_S\|_2 + \|u_{\overline{S}}\|_\infty < 1,$$

provided we can show that $\|v_{\overline{S}}\|_1 > 0$ (this is necessary as we have to prove a strict inequality between $|\langle v, \operatorname{sgn}(x)\rangle|$ and $\|v_{\overline{S}}\|_1$).

Indeed, if $\|v_{\overline{S}}\|_1 = 0$, we would have $\|v_S\|_2 \leq \alpha \|v_{\overline{S}}\|_1 = 0$, and so $v_{\overline{S}} = 0$ and $v_S = 0$. This, however, stands in contradicton to the assumption $v \neq 0$, and hence we must have $\|v_{\overline{S}}\|_1 > 0$, as desired.

(ii) Let $z \in \mathbb{C}^N \setminus \{x\}$ be such that $Az = Ax$, but otherwise arbitrary, and set $v = x - z$. Then $v \neq 0$ and $Av = Ax - Az = 0$, and thus $v \in \ker(A) \setminus \{0\}$. Hence

$$\|x\|_1 \leq \|z\|_1 - \|v_{\overline{S}}\|_1 + \; |\langle v, \operatorname{sgn}(x)\rangle| < \|z\|_1.$$

As z was arbitrary, this establishes that x is the unique minimizer of

$$\min_{z \in \mathbb{C}^N} \|z\|_1 \quad \text{subject to} \quad Az = Ax.$$

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

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

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

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

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

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

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

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

(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 4

(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.,

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

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:

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