Introduction to Statistical Learning

Exercise set #3

Exercise 1 - (Properties of Rademacher averages) Let $\mathcal{T}, \mathcal{T}_1, \mathcal{T}_2$ be classes of real-valued functions. Prove the following properties:

1. If $c \in \mathbb{R}$, then $\hat{\mathcal{R}}_n(c\mathcal{T}) = |c|\hat{\mathcal{R}}_n(\mathcal{T})$.
2. If $\mathcal{T}_1 \subseteq \mathcal{T}_2$, then $\hat{\mathcal{R}}_n(\mathcal{T}_1) \leq \hat{\mathcal{R}}_n(\mathcal{T}_2)$.
3. $\hat{\mathcal{R}}_n(\mathcal{T}_1 + \mathcal{T}_2) = \hat{\mathcal{R}}_n(\mathcal{T}_1) + \hat{\mathcal{R}}_n(\mathcal{T}_2)$.
4. Let $\text{conv}(\mathcal{T})$ be the convex hull of $\mathcal{T}$. Prove that : $\hat{\mathcal{R}}_n(\text{conv}(\mathcal{T})) = \hat{\mathcal{R}}_n(\mathcal{T})$.
5. If $\psi : \mathbb{R} \to \mathbb{R}$ is k-Lipschitz, then $\hat{\mathcal{R}}_n(\psi \circ \mathcal{T}) \leq k\hat{\mathcal{R}}_n(\mathcal{T})$.

Correction 1 - Throughout, $D_n \coloneqq (X_1, \dots, X_n)$ denotes the sample, and the empirical Rademacher average of a class $\mathcal{T}$ is the random variable

$$\hat{\mathcal{R}}_n(\mathcal{T}) \coloneqq \mathbb{E}_\varepsilon \left(\sup_{f \in \mathcal{T}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \mid D_n \right),$$

where the expectation is taken over ε only, the sample being fixed. Its expectation $\mathcal{R}_n(\mathcal{T}) \coloneqq \mathbb{E}(\hat{\mathcal{R}}_n(\mathcal{T}))$ is the (deterministic) Rademacher average. All the following computations hold for any fixed sample.

1.

$$\begin{aligned}
\hat{\mathcal{R}}_n(c\mathcal{T}) &= \mathbb{E}_\varepsilon \left(\sup_{f \in c\mathcal{T}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \mid D_n \right) \\
&= \mathbb{E}_\varepsilon \left(\sup_{f \in \mathcal{T}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i c f(X_i) \mid D_n \right) \\
&= \mathbb{E}_\varepsilon \left(\sup_{f \in \mathcal{T}} \frac{1}{n} \sum_{i=1}^n \underbrace{\varepsilon_i \text{sgn}(c)}_{\mathcal{L}(\varepsilon_i) = \mathcal{L}(\text{sgn}(c)\varepsilon_i)} |c| f(X_i) \mid D_n \right) \\
&= |c| \mathbb{E}_\varepsilon \left(\sup_{f \in \mathcal{T}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \mid D_n \right) \\
&= |c| \hat{\mathcal{R}}_n(\mathcal{T})
\end{aligned}$$

> **Conclusion.** For any $c \in \mathbb{R}$, $\hat{\mathcal{R}}_n(c\mathcal{T}) = |c|\hat{\mathcal{R}}_n(\mathcal{T})$.

2. Because $\mathcal{T}_1 \subseteq \mathcal{T}_2$, for any $X_1, \dots, X_n, \varepsilon$, we have

$$\sup_{f \in \mathcal{T}_1} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \leq \sup_{f \in \mathcal{T}_2} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i).$$

Hence, because expectancy is increasing,

$$\hat{\mathcal{R}}_n(\mathcal{T}_1) = \mathbb{E}_\varepsilon \left(\sup_{f \in \mathcal{T}_1} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \mid D_n \right)$$

$$\leq \mathbb{E}_\varepsilon \left(\sup_{f \in \mathcal{T}_2} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \mid D_n \right) = \hat{\mathcal{R}}_n(\mathcal{T}_2)$$

Conclusion. If $\mathcal{T}_1 \subseteq \mathcal{T}_2$, then $\hat{\mathcal{R}}_n(\mathcal{T}_1) \leq \hat{\mathcal{R}}_n(\mathcal{T}_2)$: the empirical Rademacher average is nondecreasing with respect to inclusion.

3.

$$\begin{aligned}
\hat{\mathcal{R}}_n(\mathcal{T}_1 + \mathcal{T}_2) &= \mathbb{E}_\varepsilon \left(\sup_{f \in \mathcal{T}_1 + \mathcal{T}_2} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \mid D_n \right) \\
&= \mathbb{E}_\varepsilon \left(\sup_{\substack{f_1 \in \mathcal{T}_1 \\ f_2 \in \mathcal{T}_2}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i (f_1(X_i) + f_2(X_i)) \mid D_n \right) \\
&= \mathbb{E}_\varepsilon \left(\sup_{\substack{f_1 \in \mathcal{T}_1 \\ f_2 \in \mathcal{T}_2}} \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i f_1(X_i) \right) + \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i f_2(X_i) \right) \mid D_n \right) \\
&= \mathbb{E}_\varepsilon \left(\sup_{f_1 \in \mathcal{T}_1} \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i f_1(X_i) \right) + \sup_{f_2 \in \mathcal{T}_2} \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i f_2(X_i) \right) \mid D_n \right) \\
&= \mathbb{E}_\varepsilon \left(\sup_{f_1 \in \mathcal{T}_1} \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i f_1(X_i) \right) \mid D_n \right) + \mathbb{E}_\varepsilon \left(\sup_{f_2 \in \mathcal{T}_2} \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i f_2(X_i) \right) \mid D_n \right) \\
&= \hat{\mathcal{R}}_n(\mathcal{T}_1) + \hat{\mathcal{R}}_n(\mathcal{T}_2)
\end{aligned}$$

Conclusion. $\hat{\mathcal{R}}_n(\mathcal{T}_1 + \mathcal{T}_2) = \hat{\mathcal{R}}_n(\mathcal{T}_1) + \hat{\mathcal{R}}_n(\mathcal{T}_2)$.

4.

$$\begin{aligned}
\hat{\mathcal{R}}_n(\mathrm{conv}(\mathcal{T})) &= \mathbb{E}_\varepsilon \left(\sup_{f \in \mathrm{conv}(\mathcal{T})} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \mid D_n \right) \\
&= \mathbb{E}_\varepsilon \left(\sup_{\substack{K \in \mathbb{N}^* \\ \alpha_1, \dots, \alpha_K \in [0,1] \\ \sum_{k=1}^K \alpha_k = 1 \\ f_1, \dots, f_K \in \mathcal{T}}} \sum_{k=1}^K \alpha_k \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i f_k(X_i) \right) \mid D_n \right) \\
&= \mathbb{E}_\varepsilon \left(\sup_{\substack{K \in \mathbb{N}^* \\ \alpha_1, \dots, \alpha_K \in [0,1] \\ \sum_{k=1}^K \alpha_k = 1}} \sum_{k=1}^K \alpha_k \sup_{f \in \mathcal{T}} \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right) \mid D_n \right) \\
&= \mathbb{E}_\varepsilon \left(\sup_{f \in \mathcal{T}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \mid D_n \right) \\
&= \hat{\mathcal{R}}_n(\mathcal{T}).
\end{aligned}$$

Conclusion. $\hat{\mathcal{R}}_n(\mathrm{conv}(\mathcal{T})) = \hat{\mathcal{R}}_n(\mathcal{T})$: taking convex combinations of functions of $\mathcal{T}$ does not increase the empirical Rademacher average.

5. Let $\psi : \mathbb{R} \to \mathbb{R}$ such that for any $x, y \in \mathbb{R}$, $|\psi(x) - \psi(y)| \leq k|x - y|$, and $\psi \circ \mathcal{T} \coloneqq \{\psi \circ f,\ f \in \mathcal{T}\}$.

The idea is to replace the $\psi(f(X_i))$ by $kf(X_i)$ one coordinate at a time. For $j \in \{0, \dots, n\}$, let

$$\Phi_j \coloneqq \mathbb{E}_\varepsilon \left(\sup_{f \in \mathcal{T}} \left[k \sum_{i \le j} \varepsilon_i f(X_i) + \sum_{i > j} \varepsilon_i \psi(f(X_i)) \right] \mid D_n \right),$$

so that $\Phi_0 = n\hat{\mathcal{R}}_n(\psi \circ \mathcal{T})$ and $\Phi_n = kn\hat{\mathcal{R}}_n(\mathcal{T})$. Let us show that $\Phi_{j-1} \le \Phi_j$ for any $j \in \{1, \dots, n\}$. Fix ε_i for $i \neq j$, and let $u(f) \coloneqq k \sum_{i<j} \varepsilon_i f(X_i) + \sum_{i>j} \varepsilon_i \psi(f(X_i))$, which does not depend on ε_j. Since ε_j is uniform on $\{-1, +1\}$,

$$\mathbb{E}_{\varepsilon_j} \left(\sup_{f \in \mathcal{T}} \left[u(f) + \varepsilon_j \psi(f(X_j)) \right] \right) = \frac{1}{2} \sup_{f \in \mathcal{T}} \left[u(f) + \psi(f(X_j)) \right] + \frac{1}{2} \sup_{f \in \mathcal{T}} \left[u(f) - \psi(f(X_j)) \right].$$

Let $\eta > 0$, and $f_1, f_2 \in \mathcal{T}$ that are η-close to the supremum. Let $s \coloneqq \operatorname{sgn}(f_1(X_j) - f_2(X_j))$. Because ψ is k-Lipschitz, $\psi(f_1(X_j)) - \psi(f_2(X_j)) \le k|f_1(X_j) - f_2(X_j)| = ks(f_1(X_j) - f_2(X_j))$, hence

$$\begin{aligned}
\mathbb{E}_{\varepsilon_j} \left(\sup_{f \in \mathcal{T}} \left[u(f) + \varepsilon_j \psi(f(X_j)) \right] \right) &\le \frac{1}{2} \left[u(f_1) + u(f_2) + \psi(f_1(X_j)) - \psi(f_2(X_j)) \right] + \eta \\
&\le \frac{1}{2} \left[u(f_1) + ksf_1(X_j) \right] + \frac{1}{2} \left[u(f_2) - ksf_2(X_j) \right] + \eta \\
&\le \frac{1}{2} \sup_{f \in \mathcal{T}} \left[u(f) + ksf(X_j) \right] + \frac{1}{2} \sup_{f \in \mathcal{T}} \left[u(f) - ksf(X_j) \right] + \eta \\
&= \mathbb{E}_{\varepsilon_j} \left(\sup_{f \in \mathcal{T}} \left[u(f) + k\varepsilon_j f(X_j) \right] \right) + \eta,
\end{aligned}$$

where the last equality holds because $\{s, -s\} = \{-1, +1\}$. Letting $\eta \to 0$ and integrating over the ε_i, $i \neq j$ (independent of ε_j), we get $\Phi_{j-1} \le \Phi_j$.

Conclusion. $n\hat{\mathcal{R}}_n(\psi \circ \mathcal{T}) = \Phi_0 \le \Phi_1 \le \cdots \le \Phi_n = kn\hat{\mathcal{R}}_n(\mathcal{T})$, hence $\hat{\mathcal{R}}_n(\psi \circ \mathcal{T}) \le k\hat{\mathcal{R}}_n(\mathcal{T})$.

Exercise 2 - (Rademacher average for linear and kernel classes) Let $\mathcal{T}, \mathcal{T}_1, \mathcal{T}_2$ be classes of real-valued functions. Prove the following properties:

1. Consider $\mathcal{B}_\infty(C) = \{x \in \mathbb{R}^d \mid \|x\|_\infty \le C\}$ and $\mathcal{G} = \{x \in \mathcal{B}_\infty(C) \mapsto w^\top x : \|w\|_1 \le B\}$. Show that the following bound holds :

$$\hat{\mathcal{R}}_n(\mathcal{G}) \le \frac{BC\sqrt{2\ln(2d)}}{\sqrt{n}}.$$

2. Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ a positive definite and symmetric kernel function with feature mapping Φ that is : for any (x, x'), we have $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$ where $\langle , \rangle$ is the product of the Hilbert space induced by k. Given a sample $X_1, \dots, X_n$, define its Gram matrix as $K = (k(X_i, X_j))_{1 \le i,j \le n}$. Consider the class of functions $\mathcal{H} = \{x \mapsto \langle w, \Phi(x) \rangle : \|w\|_k \le M\}$, and prove that :

$$\hat{\mathcal{R}}_n(\mathcal{H}) \le \frac{M\sqrt{\operatorname{Trace}(K)}}{n}.$$

Correction 2 -

1.

$$\hat{\mathcal{R}}_n(\mathcal{G}) = \mathbb{E}_\varepsilon \left(\sup_{f \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \mid D_n \right)$$

$$
\begin{aligned}
&= \mathbb{E}_\varepsilon\left(\sup_{\|w\|_1\le B}\frac{1}{n}\sum_{i=1}^n \varepsilon_i w^\top X_i \mid D_n\right)\\
&= \mathbb{E}_\varepsilon\left(\sup_{\|w\|_1\le B}\frac{1}{n}\sum_{i=1}^n \varepsilon_i \sum_{j=1}^d w^{(j)} X_i^{(j)} \mid D_n\right)\\
&= \mathbb{E}_\varepsilon\left(\sup_{\|w\|_1\le B}\frac{1}{n}\sum_{j=1}^d w^{(j)} \sum_{i=1}^n \varepsilon_i X_i^{(j)} \mid D_n\right)\\
&= \mathbb{E}_\varepsilon\left(\sup_{\|w\|_1\le B}\frac{1}{n} w^\top \left(\sum_{i=1}^n \varepsilon_i X_i\right) \mid D_n\right)\\
&\le \mathbb{E}_\varepsilon\left(\sup_{\|w\|_1\le B}\frac{1}{n}\|w\|_1 \left\|\sum_{i=1}^n \varepsilon_i X_i\right\|_\infty \mid D_n\right) \quad \text{using Hölder inequality}\\
&\le \mathbb{E}_\varepsilon\left(\frac{B}{n}\left\|\sum_{i=1}^n \varepsilon_i X_i\right\|_\infty \mid D_n\right)\\
&= \mathbb{E}_\varepsilon\left(\frac{B}{n}\max_{v\in A_X}\sum_{i=1}^n \varepsilon_i v_i \mid D_n\right)\\
&\le \frac{B}{n}\max_{v\in A_X}\|v\|_2\sqrt{2\ln|A_X|} \quad \text{using Massart's Lemma}\\
&\le \frac{B}{n} C\sqrt{n}\sqrt{2\ln(2d)} = \frac{BC\sqrt{2\ln(2d)}}{\sqrt{n}},
\end{aligned}
$$

where $A_X \subset \mathbb{R}^n$ is the finite set gathering, for each coordinate j, the j-th coordinates of the sample, with both signs:

$$A_X \coloneqq \left\{\pm\left(X_1^{(j)},\dots,X_n^{(j)}\right),\ j\in\{1,\dots,d\}\right\},$$

so that

$$\left\|\sum_{i=1}^n \varepsilon_i X_i\right\|_\infty = \max_{1\le j\le d}\left|\sum_{i=1}^n \varepsilon_i X_i^{(j)}\right| = \max_{v\in A_X}\sum_{i=1}^n \varepsilon_i v_i,$$

because $|a| = \max(a,-a)$. We have $|A_X| \le 2d$ and, since $|X_i^{(j)}| \le \|X_i\|_\infty \le C$, any $v \in A_X$ satisfies $\|v\|_2 = \sqrt{\sum_{i=1}^n (X_i^{(j)})^2} \le C\sqrt{n}$.

Conclusion. For any sample,

$$\hat{\mathcal{R}}_n(\mathcal{G}) \le \frac{BC\sqrt{2\ln(2d)}}{\sqrt{n}}.$$

2.

$$
\begin{aligned}
\hat{\mathcal{R}}_n(\mathcal{H}) &= \mathbb{E}_\varepsilon\left(\sup_{f\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i) \mid D_n\right)\\
&= \mathbb{E}_\varepsilon\left(\sup_{\|w\|_k\le M}\frac{1}{n}\sum_{i=1}^n \varepsilon_i \langle w, \Phi(X_i)\rangle \mid D_n\right)\\
&= \frac{1}{n}\mathbb{E}_\varepsilon\left(\sup_{\|w\|_k\le M}\left\langle w, \sum_{i=1}^n \varepsilon_i \Phi(X_i)\right\rangle \mid D_n\right)\\
&\le \frac{1}{n}\mathbb{E}_\varepsilon\left(\sup_{\|w\|_k\le M}\|w\|_k \left\|\sum_{i=1}^n \varepsilon_i \Phi(X_i)\right\|_k \mid D_n\right)\\
&\le \frac{M}{n}\mathbb{E}_\varepsilon\left(\left\|\sum_{i=1}^n \varepsilon_i \Phi(X_i)\right\|_k \mid D_n\right)
\end{aligned}
$$

$$
\begin{aligned}
&\leq \frac{M}{n}\mathbb{E}_\varepsilon\left(\left(\left\langle\left(\sum_{i=1}^n \varepsilon_i\Phi(X_i)\right),\left(\sum_{i=1}^n \varepsilon_i\Phi(X_i)\right)\right\rangle_k\right)^{1/2} \mid D_n\right)\\
&\leq \frac{M}{n}\mathbb{E}_\varepsilon\left(\left(\left\langle\left(\sum_{i=1}^n \varepsilon_i\Phi(X_i)\right),\left(\sum_{i=1}^n \varepsilon_i\Phi(X_i)\right)\right\rangle_k\right) \mid D_n\right)^{1/2} \qquad \text{(Jensen)}\\
&\leq \frac{M}{n}\mathbb{E}_\varepsilon\left(\sum_{1\leq i,j\leq n}\langle\varepsilon_i\Phi(X_i),\varepsilon_j\Phi(X_j)\rangle_k \mid D_n\right)^{1/2}\\
&\leq \frac{M}{n}\mathbb{E}_\varepsilon\left(\sum_{1\leq i,j\leq n}\varepsilon_i\varepsilon_j\langle\Phi(X_i),\Phi(X_j)\rangle_k \mid D_n\right)^{1/2}\\
&\leq \frac{M}{n}\mathbb{E}_\varepsilon\left(\sum_{1\leq i\neq j\leq n}\varepsilon_i\varepsilon_j\langle\Phi(X_i),\Phi(X_j)\rangle_k + \sum_{i=1}^n\|\Phi(X_i)\|_k^2 \mid D_n\right)^{1/2}\\
&\leq \frac{M}{n}\left(\sum_{1\leq i\neq j\leq n}\mathbb{E}_\varepsilon(\varepsilon_i\varepsilon_j)\langle\Phi(X_i),\Phi(X_j)\rangle_k + \sum_{i=1}^n\|\Phi(X_i)\|_k^2\right)^{1/2} \qquad (\varepsilon_i^2=1)\\
&\leq \frac{M}{n}\left(\sum_{1\leq i\neq j\leq n}\underbrace{\mathbb{E}_\varepsilon(\varepsilon_i)}_{=0}\underbrace{\mathbb{E}_\varepsilon(\varepsilon_j)}_{=0}\langle\Phi(X_i),\Phi(X_j)\rangle_k + \sum_{i=1}^n\|\Phi(X_i)\|_k^2\right)^{1/2} \qquad (\varepsilon_i \text{ i.i.d.})\\
&\leq \frac{M}{n}\left(\underbrace{\sum_{i=1}^n\|\Phi(X_i)\|_k^2}_{=\mathrm{Trace}(K)}\right)^{1/2}\\
&\leq \frac{M\sqrt{\mathrm{Trace}(K)}}{n}
\end{aligned}
$$

Conclusion. For any sample,

$$\hat{\mathcal{R}}_n(\mathcal{H}) \leq \frac{M\sqrt{\mathrm{Trace}(K)}}{n}.$$

In particular, if $k(x,x)\leq B^2$ for any x, then $\mathrm{Trace}(K)\leq nB^2$, so that $\hat{\mathcal{R}}_n(\mathcal{H})\leq MB/\sqrt{n}$.

Exercise 3 - (Relation between Rademacher average and combinatorial complexity) Consider a class $\mathcal{G}$ of binary valued functions with shattering coefficient function denoted by $\gamma(\mathcal{G},n)$. Show that :

$$\mathbb{E}(\hat{\mathcal{R}}_n(\mathcal{G})) \leq \sqrt{\frac{2\ln(\gamma(\mathcal{G},n))}{n}}$$

and if $V(\mathcal{G})<\infty$ show that, for some constant C, we have :

$$\mathbb{E}(\hat{\mathcal{R}}_n(\mathcal{G})) \leq C\sqrt{\frac{V(\mathcal{G})\log(n)}{n}}.$$

Correction 3 - Let $D_n \coloneqq (Z_1,\ldots,Z_n)$ be the sample. By definition, the empirical Rademacher average is a random variable depending on D_n:

$$\hat{\mathcal{R}}_n(\mathcal{G}) = \mathbb{E}_\varepsilon\left(\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^n\varepsilon_i g(Z_i) \mid D_n\right),$$

where the expectation is taken over ε only, the sample being fixed. Let $\mathcal{G}_{D_n}$ be the ensemble of $(g(z_1), \dots, g(z_n))$ for any $g \in \mathcal{G}$:

$$\mathcal{G}_{D_n} \coloneqq \{(g(z_i))_{1 \le i \le n},\ g \in \mathcal{G}\},$$

so that a change a variables yield

$$\hat{\mathcal{R}}_n(\mathcal{G}) = \mathbb{E}_\varepsilon \left(\sup_{z \in \mathcal{G}_{D_n}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i z_i \mid D_n \right).$$

But because g is $\{-1, +1\}$-valued, we have $\|z\| \le \sqrt{n}$. Applying Massart's lemma, this gives

$$\hat{\mathcal{R}}_n(\mathcal{G}) \le \frac{\sqrt{2n \ln |\mathcal{G}_{D_n}|}}{n}.$$

But we also have the bound

$$|\mathcal{G}_{D_n}| \le \gamma(\mathcal{G}, n) \coloneqq \max_{(z_i)_{1 \le i \le n}} |\{(g(z_i))_{1 \le i \le n},\ g \in \mathcal{G}\}|.$$

Hence, for any sample D_n,

$$\hat{\mathcal{R}}_n(\mathcal{G}) \le \sqrt{\frac{2 \ln \gamma(\mathcal{G}, n)}{n}},$$

and taking the expectation over D_n gives $\mathbb{E}(\hat{\mathcal{R}}_n(\mathcal{G})) \le \sqrt{\frac{2 \ln \gamma(\mathcal{G}, n)}{n}}$. Now, assume that $V(\mathcal{G}) < +\infty$. If $n \le V(\mathcal{G})$, it means that $\gamma(\mathcal{G}, n) = 2^n$, so that

$$\hat{\mathcal{R}}_n(\mathcal{G}) \le \sqrt{2}.$$

If $n > V(\mathcal{G}) =: V$, $\gamma(\mathcal{G}, n) \le \left(\frac{en}{V}\right)^V$, and

$$\begin{aligned}
\hat{\mathcal{R}}_n(\mathcal{G}) &\le \sqrt{\frac{2 \ln \left(\frac{en}{V}\right)^V}{n}} \\
&\le \sqrt{\frac{2V \left(\ln \left(\frac{n}{V} + 1\right)\right)}{n}} \\
&\le \sqrt{2(\ln V - 1) \frac{V \ln n}{n}} \\
&\le C \sqrt{\frac{V \ln n}{n}}.
\end{aligned}$$

Conclusion. In both cases, for any sample D_n, $\hat{\mathcal{R}}_n(\mathcal{G}) \le C\sqrt{\frac{V \ln n}{n}}$. Taking the expectation,

$$\mathbb{E}(\hat{\mathcal{R}}_n(\mathcal{G})) \le C \sqrt{\frac{V \ln n}{n}}.$$

Exercise 4 - (Contraction principle)

1. Consider $\psi : \mathbb{R} \to \mathbb{R}$ a Lipschitz function with constant κ. Show that, for any class $\mathcal{F}$ of real-valued functions, we have:
$$\hat{\mathcal{R}}_n(\psi \circ \mathcal{F}) \le \kappa \hat{\mathcal{R}}_n(\mathcal{F}).$$
Hint: Use a recurrence argument and the definition of the supremum applied to the following functions of f: $u_{j-1}(f) + (\psi \circ f)(X_j)$ and $u_{j-1}(f) - (\psi \circ f)(X_j)$.

2. Consider the following function: for any $\rho > 0$,

$$\psi_\rho(t) = (1 - t/\rho)\,\mathbb{I}\{0 \le t \le \rho\} + \mathbb{I}\{t \le 0\}.$$

Let $\mathcal{F}$ be a class of real-valued functions and a fixed value of ρ. We assume (X, Y), (X_1, Y_1), $\ldots$, (X_n, Y_n) are i.i.d. binary classification data with labels in $\{-1, +1\}$. Denote by $L(f) = \mathbb{P}(Y \cdot f(X) < 0)$ and $\widehat{L}_{n,\rho}(f) = \frac{1}{n}\sum_{i=1}^n \psi_\rho(Y_i \cdot f(X_i))$. For any $\delta > 0$, show that with probability at least $1 - \delta$, the two following inequalities hold:

$$\sup_{f\in\mathcal{F}} \left(L(f) - \widehat{L}_{n,\rho}(f)\right) \le \frac{2}{\rho}\mathbb{E}(\hat{\mathcal{R}}_n(\mathcal{F})) + \sqrt{\frac{\log(1/\delta)}{2n}}$$

and

$$\sup_{f\in\mathcal{F}} \left(L(f) - \widehat{L}_{n,\rho}(f)\right) \le \frac{2}{\rho}\hat{\mathcal{R}}_n(\mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}.$$

Correction 4 -

1. This is exactly Exercise 1, question 5, with κ in place of k and $\mathcal{F}$ in place of $\mathcal{T}$: see the proof there.

> **Remark**
>
> As written in the statement, $\psi_\rho(0) = 2$, since both indicators are equal to 1 at $t = 0$. We use instead the intended definition $\psi_\rho(t) = (1 - t/\rho)\,\mathbb{I}\{0 < t \le \rho\} + \mathbb{I}\{t \le 0\} = \min(1, \max(0, 1 - t/\rho))$, which is $[0,1]$-valued and $1/\rho$-Lipschitz.

2. Let $\mathcal{F}' \coloneqq \{(x, y) \mapsto yf(x), f \in \mathcal{F}\}$, and $\tilde{\mathcal{F}} \coloneqq \{\psi_\rho \circ f, f \in \mathcal{F}'\}$. $\tilde{\mathcal{F}}$ is a set of $[0,1]$-valued functions. Hence, we can apply the results from Exercise 3, TD 2, (second part, with $\delta/2 \leftarrow \delta$) so that with probability $1 - \delta$

$$\begin{aligned}
&\forall f \in \tilde{\mathcal{F}}, \quad \mathbb{E}(f(Z_1)) - \frac{1}{n}\sum_{i=1}^n f(Z_i) \le 2\mathcal{R}_n(\tilde{\mathcal{F}}) + \sqrt{\frac{\log(1/\delta)}{2n}},\\
\iff &\forall f \in \mathcal{F}, \quad \mathbb{E}(\psi_\rho(Y_1 f(X_1))) - \hat{L}_{n,\rho}(f) \le 2\mathcal{R}_n(\psi_\rho \circ \mathcal{F}') + \sqrt{\frac{\log(1/\delta)}{2n}},\\
\implies &\forall f \in \mathcal{F}, \quad \mathbb{E}(\psi_\rho(Y_1 f(X_1))) - \hat{L}_{n,\rho}(f) \le \frac{2}{\rho}\mathcal{R}_n(\mathcal{F}') + \sqrt{\frac{\log(1/\delta)}{2n}} \quad (\psi_\rho \text{ is } 1/\rho\text{-Lipschitz.}).
\end{aligned}$$

But because $\mathbb{1}_{t\le 0} \le \psi_\rho(t)$, we can write

$$L(f) = \mathbb{E}(\mathbb{1}_{Yf(X)<0}) \le \mathbb{E}(\psi_\rho(Yf(X))) \le \hat{L}_{n,\rho}(f) + \frac{2}{\rho}\mathcal{R}_n(\mathcal{F}') + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

Finally, we have

$$\begin{aligned}
\mathcal{R}_n(\mathcal{F}') &= \mathbb{E}_{X,Y}\left(\mathbb{E}_\varepsilon\left(\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i Y_i f(X_i)\right)\right)\\
&= \mathbb{E}_X\left(\mathbb{E}_\varepsilon\left(\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right)\right) \quad \text{because } \mathcal{L}(\varepsilon_i) = \mathcal{L}(\varepsilon_i Y_i),\\
&= \mathcal{R}_n(\mathcal{F}).
\end{aligned}$$

With probability $1-\delta$,

$$\forall f \in \mathcal{F}, \quad L(f) - \hat{L}_{n,\rho}(f) \leq \frac{2}{\rho}\mathbb{E}(\hat{\mathcal{R}}_n(\mathcal{F})) + \sqrt{\frac{\log(1/\delta)}{2n}},$$

so that the inequality is verified for the supremum.

For the second inequality, we apply the full result of Exercise 3, TD 2 (second part) to the $[0,1]$-valued class $\tilde{\mathcal{F}}$: with probability $1-\delta$,

$$\forall f \in \mathcal{F}, \quad \mathbb{E}(\psi_\rho(Y_1 f(X_1))) - \hat{L}_{n,\rho}(f) \leq 2\hat{\mathcal{R}}_n(\tilde{\mathcal{F}}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}.$$

This time, we need to bound the empirical Rademacher average $\hat{\mathcal{R}}_n(\tilde{\mathcal{F}})$ itself, and not its expectation. The argument used above actually hold for any fixed sample D_n, where $D_n := ((X_1,Y_1),\dots,(X_n,Y_n))$. Conditionally on D_n, the $Y_i \in \{-1,+1\}$ are fixed and the ε_i are i.i.d. and symmetric, so that $(\varepsilon_1 Y_1,\dots,\varepsilon_n Y_n)$ has the same law as $(\varepsilon_1,\dots,\varepsilon_n)$. Hence

$$\hat{\mathcal{R}}_n(\mathcal{F}') = \mathbb{E}_\varepsilon\left(\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i Y_i f(X_i) \mid D_n\right) = \mathbb{E}_\varepsilon\left(\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i) \mid D_n\right) = \hat{\mathcal{R}}_n(\mathcal{F}).$$

Therefore $\hat{\mathcal{R}}_n(\tilde{\mathcal{F}}) \leq \frac{1}{\rho}\hat{\mathcal{R}}_n(\mathcal{F})$ for any sample. Combining this with $L(f) \leq \mathbb{E}(\psi_\rho(Yf(X)))$ as above, we get the following.

Conclusion. With probability $1-\delta$,

$$\sup_{f\in\mathcal{F}}\left(L(f) - \hat{L}_{n,\rho}(f)\right) \leq \frac{2}{\rho}\hat{\mathcal{R}}_n(\mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}.$$

Exercise 5 - (SVM consistency) Consider a kernel k, uniformly bounded by B^2, with associated RKHS $\mathcal{F}$ and norm $\|\cdot\|_{\mathcal{F}}$, and assume that $\inf_{f\in\mathcal{F}} A(f) = \inf_f A(f)$ holds, where

$$A(f) = \mathbb{E}\left(\max\{0, 1 - Y\cdot f(X)\}\right).$$

Consider also the following estimator for fixed $\lambda > 0$:

$$\widehat{f}_n^\lambda = \operatorname*{arg\,min}_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^n \max\{0, 1 - Y_i\cdot f(X_i)\} + \lambda\|f\|_{\mathcal{F}}^2\right\}.$$

1. Show that $\|\widehat{f}_n^\lambda\|^2 \leq 1/\lambda$.
2. Show that, with probability at least $1-\delta$:

$$A(\widehat{f}_n^\lambda) \leq \widehat{A}_n(\widehat{f}_n^\lambda) + \frac{2B}{\sqrt{n\lambda}} + \left(1 + B/\sqrt{\lambda}\right)\sqrt{\frac{\log(2/\delta)}{2n}}.$$

3. Now set $\lambda = \lambda_n$ such that $\lambda_n \to 0$ and $n\lambda_n \to \infty$ when $n\to\infty$, and prove that, for any $\epsilon > 0$, we have:

$$\sum_{n\geq 0}\mathbb{P}\left(A(\widehat{f}_n^{\lambda_n}) - \inf_f A(f) \geq \epsilon\right) < \infty.$$

Deduce that $A(\widehat{f}_n^{\lambda_n})$ tends to $\inf_f A(f)$ almost surely. What can be said about $L(\operatorname{sgn}(\widehat{f}_n^{\lambda_n}))$?

4. Present an alternative proof of SVM consistency based on the property of stability.

Correction 5 -

1. Let
$$\hat{A}_n^\lambda(f) \coloneqq \frac{1}{n}\sum_{i=1}^n \max\{0, 1 - Y_i \cdot f(X_i)\} + \lambda\|f\|_{\mathcal{F}}^2.$$
First, we can notice that $\hat{A}_n^\lambda(f \equiv 0) = 1$, so that $\hat{A}_n^\lambda(\hat{f}_n^\lambda) \leq \hat{A}_n^\lambda(0) = 1$ because $\hat{f}_n^\lambda$ minimizes $\hat{A}_n^\lambda$. Moreover, $\frac{1}{n}\sum_{i=1}^n \max\{0, 1 - Y_i \cdot f(X_i)\} \geq 0$ so that
$$\lambda\|\hat{f}_n^\lambda\|_{\mathcal{F}}^2 \leq \hat{A}_n^\lambda(\hat{f}_n^\lambda) \leq 1.$$

> **Conclusion.** Hence, $\|\hat{f}_n^\lambda\|_{\mathcal{F}}^2 \leq \frac{1}{\lambda}$.

2. (i) We have $A(f) = \mathbb{E}(\varphi(Yf(X)))$, with $\varphi(t) \coloneqq \max(0, 1-t)$. φ is 1-Lipschitz.
   (ii) $\hat{A}_n(f) \coloneqq \frac{1}{n}\sum_{i=1}^n \max\{0, 1 - Y_i \cdot f(X_i)\}$ so that $\hat{A}_n^\lambda(f) = \hat{A}_n(f) + \lambda\|f\|_{\mathcal{F}}^2$.
   (iii) Let $\mathcal{F}_\lambda \coloneqq \{f \in \mathcal{F} : \|f\|_{\mathcal{F}} \leq 1/\sqrt{\lambda}\}$, which contains $\hat{f}_n^\lambda$ by question 1. For any $f \in \mathcal{F}_\lambda$ and any x, the RKHS property and the Cauchy-Schwarz inequality give
   $$|f(x)| = |\langle f, k(x,\cdot)\rangle_{\mathcal{F}}| \leq \|f\|_{\mathcal{F}}\,\|k(x,\cdot)\|_{\mathcal{F}} = \|f\|_{\mathcal{F}}\sqrt{k(x,x)} \leq \frac{B}{\sqrt{\lambda}}.$$
   Hence, for $y \in \{-1, +1\}$, $yf(x) \in [-B/\sqrt{\lambda}, B/\sqrt{\lambda}]$. Since $\varphi \geq 0$ and nonincreasing,
   $$0 \leq \varphi(yf(x)) \leq \varphi\left(-\frac{B}{\sqrt{\lambda}}\right) = 1 + \frac{B}{\sqrt{\lambda}}.$$
   Consequently, for any $f \in \mathcal{F}_\lambda$ and any x, x', y, y',
   $$|\varphi(yf(x)) - \varphi(y'f(x'))| \leq 1 + \frac{B}{\sqrt{\lambda}}.$$
   (iv) Let $Z \coloneqq (X, Y)$, $Z_i \coloneqq (X_i, Y_i)$, $c_\lambda \coloneqq 1 + B/\sqrt{\lambda}$, and consider the loss class
   $$\mathcal{G}_\lambda \coloneqq \{g_f,\ f \in \mathcal{F}_\lambda\}, \quad \text{where} \quad g_f(x,y) \coloneqq \varphi(yf(x)),$$
   so that $\mathbb{E}(g_f(Z)) = A(f)$ and $\frac{1}{n}\sum_{i=1}^n g_f(Z_i) = \hat{A}_n(f)$. By (iii), $\mathcal{G}_\lambda$ is $[0, c_\lambda]$-valued, so that $\mathcal{G}_\lambda/c_\lambda \coloneqq \{g/c_\lambda,\ g \in \mathcal{G}_\lambda\}$ is $[0,1]$-valued. We can thus apply to $\mathcal{G}_\lambda/c_\lambda$ the first result of Exercise 3, TD 2 (second part, with $\delta/2 \leftarrow \delta$). With probability $1-\delta$,
   $$\forall f \in \mathcal{F}_\lambda, \quad \frac{A(f)}{c_\lambda} - \frac{\hat{A}_n(f)}{c_\lambda} \leq 2\mathcal{R}_n\left(\mathcal{G}_\lambda/c_\lambda\right) + \sqrt{\frac{\log(1/\delta)}{2n}}.$$
   Since $\mathcal{R}_n(\mathcal{G}_\lambda/c_\lambda) = \mathcal{R}_n(\mathcal{G}_\lambda)/c_\lambda$ (Exercise 1, question 1), multiplying by c_λ gives, with probability $1-\delta$,
   $$\forall f \in \mathcal{F}_\lambda, \quad A(f) \leq \hat{A}_n(f) + 2\mathcal{R}_n(\mathcal{G}_\lambda) + c_\lambda\sqrt{\frac{\log(1/\delta)}{2n}}.$$
   (v) We now bound $\mathcal{R}_n(\mathcal{G}_\lambda)$. Let $\tilde{\mathcal{F}}_\lambda \coloneqq \{(x,y) \mapsto yf(x),\ f \in \mathcal{F}_\lambda\}$, so that $\mathcal{G}_\lambda = \varphi \circ \tilde{\mathcal{F}}_\lambda$. Let $K \coloneqq (k(X_i, X_j))_{1\leq i,j\leq n}$ be the Gram matrix of the sample. For any fixed sample D_n,
   $$\begin{aligned}
   \hat{\mathcal{R}}_n(\mathcal{G}_\lambda) = \hat{\mathcal{R}}_n(\varphi \circ \tilde{\mathcal{F}}_\lambda) &\leq \hat{\mathcal{R}}_n(\tilde{\mathcal{F}}_\lambda) && (\varphi \text{ is 1-Lipschitz}) \\
   &= \hat{\mathcal{R}}_n(\mathcal{F}_\lambda) && (\mathcal{L}(\varepsilon_i Y_i) = \mathcal{L}(\varepsilon_i)) \\
   &\leq \frac{1}{\sqrt{\lambda}}\frac{\sqrt{\mathrm{Trace}(K)}}{n} && (\text{Exercise 2, question 2 with } M = 1/\sqrt{\lambda}) \\
   &\leq \frac{1}{\sqrt{\lambda}}\frac{\sqrt{nB^2}}{n} = \frac{B}{\sqrt{n\lambda}}.
   \end{aligned}$$

Taking the expectation over D_n, we get $\mathcal{R}_n(\mathcal{G}_\lambda) \leq \frac{B}{\sqrt{n\lambda}}$.

Conclusion. Combining (iv) and (v), with probability $1-\delta$,

$$\forall f \in \mathcal{F}_\lambda, \quad A(f) \leq \hat{A}_n(f) + \frac{2B}{\sqrt{n\lambda}} + \left(1 + \frac{B}{\sqrt{\lambda}}\right)\sqrt{\frac{\log(1/\delta)}{2n}}.$$

Because $\hat{f}_n^\lambda \in \mathcal{F}_\lambda$,

$$A(\hat{f}_n^\lambda) \leq \hat{A}_n(\hat{f}_n^\lambda) + \frac{2B}{\sqrt{n\lambda}} + \left(1 + \frac{B}{\sqrt{\lambda}}\right)\sqrt{\frac{\log(1/\delta)}{2n}}.$$

Remark

The bound obtained is with $\log(1/\delta)$ instead of $\log(2/\delta)$.

3. Let $A^* \coloneqq \inf_f A(f)$, and recall that $A^* = \inf_{f\in\mathcal{F}} A(f)$ by assumption.

Remark

The assumption $n\lambda_n \to \infty$ is not quite enough to get a summable series from question 2. Indeed, its deviation term contains $\frac{B}{\sqrt{\lambda_n}}\sqrt{\frac{\log(2/\delta)}{2n}} = B\sqrt{\frac{\log(2/\delta)}{2n\lambda_n}}$, which must go to 0 for a summable choice of $\delta = \delta_n$, such as $\delta_n = 1/n^2$. This requires $\frac{n\lambda_n}{\log n} \to \infty$, which we assume from now on (e.g. $\lambda_n = n^{-1/2}$). Under $n\lambda_n \to \infty$ only, the same argument with a fixed δ shows that $\mathbb{P}(A(\hat{f}_n^{\lambda_n}) - A^* \geq \epsilon) \to 0$, i.e. convergence in probability.

(i) *Approximation.* Fix $\epsilon > 0$. Since $A^* = \inf_{f\in\mathcal{F}} A(f)$, there exists $f_\epsilon \in \mathcal{F}$ such that $A(f_\epsilon) \leq A^* + \epsilon/3$. Note that f_ϵ is deterministic: it depends neither on n nor on the sample. It replaces $f^* = \arg\min_f A(f)$, which may not exist, or may not belong to $\mathcal{F}$.

(ii) *Decomposition.* We write

$$A(\hat{f}_n^{\lambda_n}) - A^* \leq \underbrace{A(\hat{f}_n^{\lambda_n}) - \hat{A}_n(\hat{f}_n^{\lambda_n})}_{(a)} + \underbrace{\hat{A}_n(\hat{f}_n^{\lambda_n}) - \hat{A}_n(f_\epsilon)}_{(b)} + \underbrace{\hat{A}_n(f_\epsilon) - A(f_\epsilon)}_{(c)} + \frac{\epsilon}{3}.$$

Let $\delta_n \in (0,1)$, to be chosen later.

(a) By question 2 with $\lambda = \lambda_n$ and $\delta = \delta_n$, with probability $1-\delta_n$,

$$(a) \leq C_n \coloneqq \frac{2B}{\sqrt{n\lambda_n}} + \left(1 + \frac{B}{\sqrt{\lambda_n}}\right)\sqrt{\frac{\log(2/\delta_n)}{2n}}.$$

(b) This term is bounded deterministically. $\hat{f}_n^{\lambda_n}$ minimizes $\hat{A}_n + \lambda_n\|\cdot\|_{\mathcal{F}}^2$ over $\mathcal{F}$, so that,

$$\hat{A}_n(\hat{f}_n^{\lambda_n}) + \lambda_n\|\hat{f}_n^{\lambda_n}\|_{\mathcal{F}}^2 \leq \hat{A}_n(f_\epsilon) + \lambda_n\|f_\epsilon\|_{\mathcal{F}}^2.$$

Hence

$$(b) \leq \lambda_n\|f_\epsilon\|_{\mathcal{F}}^2 - \lambda_n\|\hat{f}_n^{\lambda_n}\|_{\mathcal{F}}^2 \leq \lambda_n\|f_\epsilon\|_{\mathcal{F}}^2.$$

(c) Since f_ϵ is fixed, $\hat{A}_n(f_\epsilon) = \frac{1}{n}\sum_{i=1}^n \varphi(Y_i f_\epsilon(X_i))$ is an average of i.i.d. variables with expectation $A(f_\epsilon)$. As in 2.(iii), these variables lie in $[0, c_\epsilon]$ with $c_\epsilon \coloneqq 1 + B\|f_\epsilon\|_{\mathcal{F}}$. Hoeffding's inequality (TD 2, Exercise 1, question 2, applied to the $\varphi(Y_i f_\epsilon(X_i))/c_\epsilon$) gives, with probability $1-\delta_n$,

$$(c) \leq c_\epsilon\sqrt{\frac{\log(1/\delta_n)}{2n}}.$$

(iii) *Summability.* Summing the three inequalities, with probability $1 - 2\delta_n$,

$$A(\hat{f}_n^{\lambda_n}) - A^* \leq R_n + \frac{\epsilon}{3}, \quad \text{where} \quad R_n \coloneqq C_n + \lambda_n \|f_\epsilon\|_{\mathcal{F}}^2 + c_\epsilon \sqrt{\frac{\log(1/\delta_n)}{2n}}.$$

Let $\delta_n \coloneqq 1/n^2$. Then $R_n \to 0$ because:

* $\frac{2B}{\sqrt{n\lambda_n}} \to 0$ since $n\lambda_n \to \infty$;
* $\left(1 + \frac{B}{\sqrt{\lambda_n}}\right) \sqrt{\frac{\log(2n^2)}{2n}} = \sqrt{\frac{\log(2n^2)}{2n}} + B\sqrt{\frac{\log(2n^2)}{2n\lambda_n}} \to 0$ since we assumed $\frac{n\lambda_n}{\log n} \to \infty$;
* $\lambda_n \|f_\epsilon\|_{\mathcal{F}}^2 \to 0$ since $\lambda_n \to 0$ and f_ϵ does not depend on n;
* $c_\epsilon \sqrt{\frac{\log(n^2)}{2n}} \to 0$.

Hence there exists $N_\epsilon \geq 1$ such that $R_n \leq \epsilon/3$ for any $n \geq N_\epsilon$. For such n, with probability $1 - 2/n^2$, $A(\hat{f}_n^{\lambda_n}) - A^* \leq 2\epsilon/3 < \epsilon$, that is,

$$\mathbb{P}\left(A(\hat{f}_n^{\lambda_n}) - A^* \geq \epsilon\right) \leq \frac{2}{n^2}.$$

Conclusion. Bounding the first N_ϵ terms by 1,

$$\sum_{n \geq 0} \mathbb{P}\left(A(\hat{f}_n^{\lambda_n}) - \inf_f A(f) \geq \epsilon\right) \leq N_\epsilon + \sum_{n \geq N_\epsilon} \frac{2}{n^2} < \infty.$$

(iv) *Almost sure convergence.* For $\epsilon > 0$, let $E_n(\epsilon) \coloneqq \{A(\hat{f}_n^{\lambda_n}) - A^* \geq \epsilon\}$. Since $\sum_n \mathbb{P}(E_n(\epsilon)) < \infty$, the Borel–Cantelli lemma gives $\mathbb{P}(\limsup_n E_n(\epsilon)) = 0$, i.e., almost surely, $A(\hat{f}_n^{\lambda_n}) - A^* < \epsilon$ for some n large enough.

Conclusion. Because this is true for $\varepsilon > 0$ arbitrarily small, $A(\hat{f}_n^{\lambda_n}) \to \inf_f A(f)$ almost surely.

(v) *Consequence for the classification error.* Let $\eta(x) \coloneqq \mathbb{P}(Y = 1 \mid X = x)$, and let $L^* \coloneqq \inf_g L(g)$ be the Bayes error. Zhang's inequality for the hinge loss gives for any measurable f,

$$L(\operatorname{sgn}(f)) - L^* \leq A(f) - A^*.$$

Conclusion. $0 \leq L(\operatorname{sgn}(\hat{f}_n^{\lambda_n})) - L^* \leq A(\hat{f}_n^{\lambda_n}) - A^* \to 0$ almost surely: the SVM classifier $\operatorname{sgn}(\hat{f}_n^{\lambda_n})$ is strongly Bayes-consistent.

4. See here, Chapter 14, Corollary 14.8 (Stability-based learning bound for SVMs).