Statistical learning

— Proof textbook —

Gérard Biau

2022-2023

Lemma 1. *For any classifier g, $L(g^\star) \leq L(g)$. Thus, $g^\star$ is the optimal decision.*

Proof. Let $g : \mathbb{R}^d \to \{0,1\}$ be an arbitrary Borel measurable function. Then

$$\mathbb{P}(g(X) \neq Y) = 1 - \mathbb{P}(g(X) = Y).$$

Thus,

$$\begin{aligned}\mathbb{P}(g(X) \neq Y) - \mathbb{P}(g^\star(X) \neq Y) &= \mathbb{P}(g^\star(X) = Y) - \mathbb{P}(g(X) = Y) \\ &= \mathbb{E}\big(\mathbb{P}(g^\star(X) = Y|X) - \mathbb{P}(g(X) = Y|X)\big) \\ &\geq 0.\end{aligned}$$

To prove this inequality, just note that

$$\begin{aligned}\mathbb{P}(g(X) = Y|X) &= \mathbb{P}(g(X) = 1, Y = 1|X) + \mathbb{P}(g(X) = 0, Y = 0|X) \\ &= \mathbb{1}_{[g(X)=1]}\mathbb{P}(Y = 1|X) + \mathbb{1}_{[g(X)=0]}\mathbb{P}(Y = 0|X).\end{aligned}$$

Similarly,

$$\begin{aligned}\mathbb{P}(g^\star(X) = Y|X) &= \mathbb{P}(g^\star(X) = 1, Y = 1|X) + \mathbb{P}(g^\star(X) = 0, Y = 0|X) \\ &= \mathbb{1}_{[g^\star(X)=1]}\mathbb{P}(Y = 1|X) + \mathbb{1}_{[g^\star(X)=0]}\mathbb{P}(Y = 0|X) \\ &= \max(\mathbb{P}(Y = 0|X), \mathbb{P}(Y = 1|X)),\end{aligned}$$

by definition of $g^\star$. □

Lemma 2. *One has*

(*i*) $L(g_n^\star) - \inf_{g\in\mathscr{C}} L(g) \leq 2\sup_{g\in\mathscr{C}} |L_n(g) - L(g)|$,

and

(*ii*) $|L_n(g_n^\star) - L(g_n^\star)| \leq \sup_{g\in\mathscr{C}} |L_n(g) - L(g)|$.

Proof. We have

$$L(g_n^\star) - \inf_{g\in\mathscr{C}} L(g) \leq |L(g_n^\star) - L_n(g_n^\star)| + |L_n(g_n^\star) - \inf_{g\in\mathscr{C}} L(g)|.$$

Clearly,

$$|L(g_n^\star) - L_n(g_n^\star)| \leq \sup_{g\in\mathscr{C}} |L_n(g) - L(g)|,$$

and

$$|L_n(g_n^\star) - \inf_{g\in\mathscr{C}} L(g)| = |\inf_{g\in\mathscr{C}} L_n(g) - \inf_{g\in\mathscr{C}} L(g)| \leq \sup_{g\in\mathscr{C}} |L_n(g) - L(g)|.$$

This shows the first assertion. The proof of (ii) is straightforward. □

Lemma 3. *Let X be a real-valued random variable with $\mathbb{E}X = 0$ and $X \in [a,b]$ $(a < b)$ with probability one. Then, for all $s \geq 0$,*

$$\mathbb{E}e^{sX} \leq e^{s^2(b-a)^2/8}.$$

Proof. Note that, by the convexity of the exponential function,

$$e^{sx} \leq \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa}, \quad a \leq x \leq b.$$

Exploiting $\mathbb{E}X = 0$, and introducing the notation $p = -\frac{a}{b-a}$, we obtain

$$\begin{aligned}\mathbb{E}e^{sX} &\leq \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb} \\ &= \left(1-p+pe^{s(b-a)}\right)e^{-ps(b-a)} \\ &\stackrel{\text{def}}{=} e^{\phi(u)},\end{aligned}$$

where $u = s(b-a)$ and $\phi(t) = -pt + \log(1-p+pe^t)$. But by straightforward calculation, it is easy to see that the derivative of ϕ is

$$\phi'(t) = -p + \frac{p}{p+(1-p)e^{-t}},$$

and therefore $\phi(0) = \phi'(0) = 0$. Moreover,

$$\phi''(t) = \frac{p(1-p)e^{-t}}{(p+(1-p)e^{-t})^2} \leq 1/4.$$

Thus, by Taylor's theorem, for some $\theta \in [0,u]$,

$$\phi(u) = \phi(0) + u\phi'(0) + \frac{u^2}{2}\phi''(\theta) \leq \frac{u^2}{8} = \frac{s^2(b-a)^2}{8}.$$

□

Lemma 4. *Let X be a random variable taking values in $\mathbb{R}_+$. Assume that there exists a constant $C \geq 1$ such that, for all $t > 0$,*

$$\mathbb{P}(X \geq t) \leq Ce^{-2nt^2}.$$

Then

$$\mathbb{E}X \leq \sqrt{\frac{\log(Ce)}{2n}}.$$

Proof. Observe that

$$\mathbb{E}X^2 = \int_0^\infty \mathbb{P}(X^2 > t)\mathrm{d}t.$$

Therefore, for all $u \geq 0$,

$$\begin{aligned}\mathbb{E}X^2 &= \int_0^u \mathbb{P}(X^2 > t)\mathrm{d}t + \int_u^\infty \mathbb{P}(X^2 > t)\mathrm{d}t \\ &\leq u + C\int_u^\infty e^{-2nt}\mathrm{d}t \\ &= u + \frac{C}{2n}e^{-2nu}.\end{aligned}$$

By choosing $u^\star = \frac{\log C}{2n}$ (which minimizes the right-hand side), we deduce that $\mathbb{E}X^2 \leq \frac{\log(Ce)}{2n}$. The result follows by the Cauchy-Schwarz inequality. □

Theorem 1. *One has*

$$\mathbb{E}\sup_{f\in\mathscr{F}}|P_n f - Pf| \leq 2\mathbb{E}R_n(\mathscr{F}(X_1^n)).$$

In addition, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\sup_{f\in\mathscr{F}}|P_n f - Pf| \leq 2\mathbb{E}R_n(\mathscr{F}(X_1^n)) + \sqrt{\frac{\log(1/\delta)}{2n}},$$

and, with probability at least $1-\delta$,

$$\sup_{f\in\mathscr{F}}|P_n f - Pf| \leq 2R_n(\mathscr{F}(X_1^n)) + 3\sqrt{\frac{\log(2/\delta)}{2n}}.$$

Proof. Introduce a "ghost sample" $X_1', \ldots, X_n'$, independent of the X_i and distributed identically. If $P_n' f = \frac{1}{n}\sum_{i=1}^n f(X_i')$ denotes the empirical averages measured on the ghost sample, then

$$\begin{aligned}\mathbb{E}\sup_{f\in\mathscr{F}}|P_n f - Pf| &= \mathbb{E}\sup_{f\in\mathscr{F}}\left|\mathbb{E}(P_n f - P_n' f \,|\, X_1, \ldots, X_n)\right| \\ &\leq \mathbb{E}\sup_{f\in\mathscr{F}}\mathbb{E}(|P_n f - P_n' f| \,\big|\, X_1, \ldots, X_n) \\ &\quad \text{(by Jensen's inequality)} \\ &\leq \mathbb{E}\sup_{f\in\mathscr{F}}|P_n f - P_n' f| \\ &\quad \text{(since } \sup\mathbb{E}(\cdot) \leq \mathbb{E}\sup(\cdot)\text{)}.\end{aligned}$$

Now, let $\sigma_1, \ldots, \sigma_n$ be independent Rademacher random variables, independent of the X_i and X_i'. Then

$$\begin{aligned}
\mathbb{E} \sup_{f \in \mathscr{F}} |P_n f - P_n' f| &= \mathbb{E}\Big(\sup_{f \in \mathscr{F}} \frac{1}{n} \Big| \sum_{i=1}^n (f(X_i) - f(X_i')) \Big| \Big) \\
&= \mathbb{E}\Big(\sup_{f \in \mathscr{F}} \frac{1}{n} \Big| \sum_{i=1}^n \sigma_i (f(X_i) - f(X_i')) \Big| \Big) \\
&\leq 2\mathbb{E} \sup_{f \in \mathscr{F}} \frac{1}{n} \Big| \sum_{i=1}^n \sigma_i f(X_i) \Big|.
\end{aligned}$$

Thus,

$$\mathbb{E} \sup_{f \in \mathscr{F}} |P_n f - P f| \leq 2\mathbb{E} R_n(\mathscr{F}(X_1^n)),$$

which shows the first assertion of the theorem. Obviously, $\sup_{f \in \mathscr{F}} |P_n f - P f|$ satisfies the bounded difference assumption with $c_i = 1/n$. Therefore, for any $\delta \in (0,1)$, with probability at least $1 - \delta$,

$$\sup_{f \in \mathscr{F}} |P_n f - P f| \leq \mathbb{E} \sup_{f \in \mathscr{F}} |P_n f - P f| + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

The third inequality follows simply by noticing that the random variable $R_n(\mathscr{F}(X_1^n))$ satisfies the conditions of the bounded difference inequality. □

Theorem 2. *Let A, B be bounded subsets of $\mathbb{R}^n$, and let $c \in \mathbb{R}$ be a constant. Then*

$$R_n(A \cup B) \leq R_n(A) + R_n(B), \quad R_n(c \cdot A) = |c| R_n(A),$$

and

$$R_n(A \oplus B) \leq R_n(A) + R_n(B),$$

where $c \cdot A = \{ca : a \in A\}$ and $A \oplus B = \{a + b : a \in A, b \in B\}$. In addition, if $A = \{a^{(1)}, \ldots, a^{(N)}\} \subseteq \mathbb{R}^n$ is a finite set, then

$$R_n(A) \leq \max_{1 \leq j \leq N} \|a^{(j)}\| \frac{\sqrt{2\log(2N)}}{n},$$

where $\|\cdot\|$ denotes Euclidean norm.

Proof. The first three properties are immediate from the definition. To show the last statement, we need the following result:

Lemma 5. *Let $\alpha > 0$, and let $X_1, \dots, X_n$ be real-valued random variables such that, for all $s > 0$ and all $1 \le i \le n$, $\mathbb{E}e^{sX_i} \le e^{s^2\alpha^2/2}$. Then, if $n \ge 2$,*

$$\mathbb{E} \max_{1 \le i \le n} X_i \le \alpha\sqrt{2\log n}.$$

If, in addition, $\mathbb{E}e^{-sX_i} \le e^{s^2\alpha^2/2}$ for all $s > 0$ and $1 \le i \le n$, then, for any $n \ge 1$,

$$\mathbb{E} \max_{1 \le i \le n} |X_i| \le \alpha\sqrt{2\log(2n)}.$$

Proof of Lemma 5. By Jensen's inequality, for all $s > 0$,

$$\begin{aligned} e^{s\mathbb{E}\max_{1\le i\le n} X_i} &\le \mathbb{E}e^{s\max_{1\le i\le n} X_i} = \mathbb{E} \max_{1 \le i \le n} e^{sX_i} \\ &\le \sum_{i=1}^n \mathbb{E}e^{sX_i} \le ne^{s^2\alpha^2/2}. \end{aligned}$$

Thus,

$$\mathbb{E} \max_{1 \le i \le n} X_i \le \frac{\log n}{s} + \frac{s\alpha^2}{2},$$

and taking $s = \sqrt{2\log n}/\alpha$ yields the first inequality. Finally, note that $\max_{1\le i\le n} |X_i| = \max(X_1, -X_1, \dots, X_n, -X_n)$ and apply the first inequality to prove the second one. □

We are now ready to prove Theorem 2. The result is clear if $\max_{1\le j\le N} \|a^{(j)}\| = 0$. Thus, in the sequel, we assume that $\max_{1\le j\le N} \|a^{(j)}\| > 0$. Observe that, for all $s > 0$, by independence, for $a = (a_1, \dots, a_n) \in A$,

$$\begin{aligned} \mathbb{E}\exp\Big(\frac{s}{n}\sum_{i=1}^n \sigma_i a_i\Big) = \prod_{i=1}^n \mathbb{E}\exp\Big(\frac{s}{n}\sigma_i a_i\Big) &\le \prod_{i=1}^n \exp\Big(\frac{s^2 a_i^2}{2n^2}\Big) \\ &\qquad \text{(by Lemma 3)} \\ &= \exp\Big(\frac{s^2\|a\|^2}{2n^2}\Big) \\ &\le \exp\Big(\frac{s^2 \max_{1\le j\le N}\|a^{(j)}\|^2}{2n^2}\Big). \end{aligned}$$

Similarly,

$$\mathbb{E}\exp\Big(-\frac{s}{n}\sum_{i=1}^n \sigma_i a_i\Big) \le \exp\Big(\frac{s^2 \max_{1\le j\le N}\|a^{(j)}\|^2}{2n^2}\Big).$$

Therefore, using Lemma 5 with $\alpha = \max_{1\le j\le N} \|a^{(j)}\|/n$, we conclude that

$$R_n(A) = \mathbb{E} \max_{1\le j\le N} \frac{1}{n}\Big|\sum_{i=1}^n \sigma_i a_i^{(j)}\Big| \le \max_{1\le j\le N} \|a^{(j)}\| \frac{\sqrt{2\log(2N)}}{n}.$$

□

Proposition 1. *For all $n \geq 1$, $\mathbf{S}_{\bar{\mathscr{A}}}(n) = \mathbf{S}_{\mathscr{A}}(n)$, where*

$$\bar{\mathscr{A}} = \big\{\{x \in \mathbb{R}^d : g(x) = 1\} : g \in \mathscr{C}\big\}.$$

In particular, $V_{\bar{\mathscr{A}}} = V_{\mathscr{A}}$.

Proof. Observe that

$$\mathscr{A} = \big\{\bar{A} \times \{0\} \cup \bar{A}^c \times \{1\} : \bar{A} \in \bar{\mathscr{A}}\big\},$$

where the sets $\bar{A}$ are of the form $\{x \in \mathbb{R}^d : g(x) = 1\}$, and the sets in $\mathscr{A}$ are sets of pairs (x, y) for which $g(x) \neq y$.

Let N be a positive integer. We show that for any n pairs from $\mathbb{R}^d \times \{0,1\}$, if N sets from $\mathscr{A}$ pick N different subsets of the n pairs, then there are N corresponding sets in $\bar{\mathscr{A}}$ that pick N different subsets of n points in $\mathbb{R}^d$, and vice versa. Fix n pairs

$$(x_1, 0), \ldots, (x_m, 0), (x_{m+1}, 1), \ldots, (x_n, 1).$$

Note that since ordering does not matter, we may arrange any n pairs in this manner. Assume that for a certain set $\bar{A} \in \bar{\mathscr{A}}$, the corresponding set $A = \bar{A} \times \{0\} \bigcup \bar{A}^c \times \{1\} \in \mathscr{A}$ picks out the pairs

$$(x_1, 0), \ldots, (x_k, 0), (x_{m+1}, 1), \ldots, (x_{m+\ell}, 1),$$

that is, the set of these pairs is the intersection of A and the n pairs. Again, we can assume without loss of generality that the pairs are ordered in this way. This means that $\bar{A}$ picks from the set $\{x_1, \ldots, x_n\}$ the subset $\{x_1, \ldots, x_k, x_{m+\ell+1}, \ldots, x_n\}$, and the two subsets uniquely determine each other. This proves $\mathbf{S}_{\mathscr{A}}(n) \leq \mathbf{S}_{\bar{\mathscr{A}}}(n)$. The other direction is proved in exactly the same way. Equality of the VC dimensions follows from the equality of the shatter coefficients. □

Theorem 3. *Let $f_n^\star \in \arg\min_{f \in \mathscr{F}} A_n(f)$, let B denote a uniform upper bound on $\phi(yf(x))$, and let $\delta \in (0,1)$. Then, with probability at least $1 - \delta$,*

$$L(f_n^\star) \leq A_n(f_n^\star) + 4L_\phi \mathbb{E}R_n(\mathscr{F}(X_1^n)) + B\sqrt{\frac{\log(1/\delta)}{2n}}.$$

Proof. Let $Z_i = (X_i, Y_i)$, $1 \leq i \leq n$. Using the bounded difference inequality and arguments similar to those of the proof of Theorem 1, we have

$$\begin{aligned} L(f_n^\star) \leq A(f_n^\star) &\leq A_n(f_n^\star) + \sup_{f \in \mathscr{F}} |A_n(f) - A(f)| \\ &\leq A_n(f_n^\star) + 2\mathbb{E}R_n(\psi \circ \mathscr{H}(Z_1^n)) + B\sqrt{\frac{\log(1/\delta)}{2n}}, \end{aligned}$$

where $\mathscr{H}$ is the class of functions on $\mathbb{R}^d \times \{-1,1\}$ of the form $yf(x)$, $f \in \mathscr{F}$, and $\psi = \phi - 1$. Thus, by the contraction principle,

$$\begin{aligned} L(f_n^\star) &\le A_n(f_n^\star) + 4L_\phi \mathbb{E}R_n(\mathscr{H}(Z_1^n)) + B\sqrt{\frac{\log(1/\delta)}{2n}} \\ &= A_n(f_n^\star) + 4L_\phi \mathbb{E}R_n(\mathscr{F}(X_1^n)) + B\sqrt{\frac{\log(1/\delta)}{2n}}. \end{aligned}$$

In the last step, we used the fact that $\sigma_i Y_i$ is a symmetric sign variable, independent of the X_i, and therefore $\sigma_i Y_i f(X_i)$ has the same distribution as that of $\sigma_i f(X_i)$. □

Theorem 4.

$$\frac{\lambda}{n\sqrt{2}}\sqrt{\sum_{i=1}^n k(X_i,X_i)} \le R_n(\mathscr{F}_\lambda(X_1^n)) \le \frac{\lambda}{n}\sqrt{\sum_{i=1}^n k(X_i,X_i)}.$$

Proof. Denoting by $\mathbb{E}_\sigma$ expectation with respect to the Rademacher variables $\sigma_1,\dots,\sigma_n$, we have

$$\begin{aligned} R_n(\mathscr{F}_\lambda(X_1^n)) &= \frac{1}{n}\mathbb{E}_\sigma \sup_{\|f\|\le\lambda} \Big| \sum_{i=1}^n \sigma_i f(X_i) \Big| \\ &= \frac{1}{n}\mathbb{E}_\sigma \sup_{\|f\|\le\lambda} \Big| \sum_{i=1}^n \sigma_i \langle f, k(\cdot,X_i)\rangle \Big| \\ &= \frac{\lambda}{n}\mathbb{E}_\sigma \Big\| \sum_{i=1}^n \sigma_i k(\cdot,X_i) \Big\| \end{aligned}$$

by the Cauchy-Schwarz inequality, where $\|\cdot\|$ denotes the norm in the reproducing kernel Hilbert space. The Kahane-Khintchine inequality states that for any vectors $a_1,\dots,a_n$ in a Hilbert space,

$$\frac{1}{2}\mathbb{E}\Big\| \sum_{i=1}^n \sigma_i a_i \Big\|^2 \le \Big(\mathbb{E}\Big\| \sum_{i=1}^n \sigma_i a_i \Big\| \Big)^2 \le \mathbb{E}\Big\| \sum_{i=1}^n \sigma_i a_i \Big\|^2.$$

It is also easy to see that

$$\mathbb{E}\Big\| \sum_{i=1}^n \sigma_i a_i \Big\|^2 = \mathbb{E} \sum_{i,j=1}^n \sigma_i \sigma_j \langle a_i, a_j \rangle = \sum_{i=1}^n \|a_i\|^2.$$

Thus, we obtain

$$\frac{\lambda}{n\sqrt{2}}\sqrt{\sum_{i=1}^n k(X_i,X_i)} \le R_n(\mathscr{F}_\lambda(X_1^n)) \le \frac{\lambda}{n}\sqrt{\sum_{i=1}^n k(X_i,X_i)}.$$

□

Theorem 5. *Let $f_n^\star \in \arg\min_{f\in\mathscr{F}_\lambda} A_n(f)$, using either the exponential or the logit loss function, and let $\delta \in (0,1)$. Then, with probability at least $1-\delta$,*

$$L(f_n^\star) - L^\star \le 2\left(8L_\phi\lambda\sqrt{\frac{V_{\mathscr{C}}\log(n+1)}{n}} + B\sqrt{\frac{\log(1/\delta)}{2n}}\right)^{1/2} + \sqrt{2}\big(\inf_{f\in\mathscr{F}_\lambda} A(f) - A^\star\big)^{1/2}.$$

Proof. We have

$$\begin{aligned}
L(f_n^\star) - L^\star &\le \sqrt{2}(A(f_n^\star) - A^\star)^{1/2} \\
&\le \sqrt{2}\big(A(f_n^\star) - \inf_{f\in\mathscr{F}_\lambda} A(f)\big)^{1/2} + \sqrt{2}\big(\inf_{f\in\mathscr{F}_\lambda} A(f) - A^\star\big)^{1/2} \\
&\le 2\big(\sup_{f\in\mathscr{F}_\lambda} |A_n(f) - A(f)|\big)^{1/2} + \sqrt{2}\big(\inf_{f\in\mathscr{F}_\lambda} A(f) - A^\star\big)^{1/2} \\
&\le 2\left(8L_\phi\lambda\sqrt{\frac{V_{\mathscr{C}}\log(n+1)}{n}} + B\sqrt{\frac{\log(1/\delta)}{2n}}\right)^{1/2} + \sqrt{2}\big(\inf_{f\in\mathscr{F}_\lambda} A(f) - A^\star\big)^{1/2}
\end{aligned}$$

with probability at least $1-\delta$. In the last inequality, we utilized the same bound for $\sup_{f\in\mathscr{F}_\lambda} |A_n(f) - A(f)|$ as in the proof of Theorem 3. □

Theorem 6. *Let* $\mathrm{pen}(n,k)$ *be defined by*

$$\mathrm{pen}(n,k) = 2R_n(\mathscr{F}_k(X_1^n)) + 4\sqrt{\frac{\log k}{n}}.$$

Then

$$\mathbb{E}L(g_{n,\hat{k}}^\star) - L^\star \le \inf_k \left(4\mathbb{E}R_n(\mathscr{F}_k(X_1^n)) + L_k^\star - L^\star + 4\sqrt{\frac{\log k}{n}}\right) + \frac{\sqrt{2}\pi^{5/2}}{3\sqrt{n}}.$$

Proof. By the definition of the selection criterion, we have, for all k,

$$L(g_{n,\hat{k}}^\star) - L^\star \le L_n(g_{n,k}^\star) - L^\star + \mathrm{pen}(n,k) + L(g_{n,\hat{k}}^\star) - L_n(g_{n,\hat{k}}^\star) - \mathrm{pen}(n,\hat{k}).$$

Taking expectations, we obtain

$$\begin{aligned}
\mathbb{E}L(g_{n,\hat{k}}^\star) - L^\star &\le \mathbb{E}\big(L_n(g_{n,k}^\star) - L^\star + \mathrm{pen}(n,k)\big) + \mathbb{E}\big(L(g_{n,\hat{k}}^\star) - L_n(g_{n,\hat{k}}^\star) - \mathrm{pen}(n,\hat{k})\big)_+ \\
&\qquad \text{(where } u_+ = \max(u,0)\text{)} \\
&\le \mathbb{E}\big(L_n(g_{n,k}^\star) - L^\star + \mathrm{pen}(n,k)\big) + \mathbb{E}\big[\sup_k \big(L(g_{n,k}^\star) - L_n(g_{n,k}^\star) - \mathrm{pen}(n,k)\big)_+\big].
\end{aligned}$$

Therefore,

$$
\begin{aligned}
&\mathbb{E}L(g^\star_{n,\hat{k}}) - L^\star \\
&\quad \le \mathbb{E}\big(L_n(g^\star_{n,k}) - L^\star + \mathrm{pen}(n,k)\big) + \mathbb{E}\big[\sup_k \big(\sup_{g\in\mathscr{C}_k} |L_n(g) - L(g)| - \mathrm{pen}(n,k)\big)_+\big] \\
&\quad \le \mathbb{E}\big(L_n(g^\star_{n,k}) - L^\star + \mathrm{pen}(n,k)\big) + \sum_{k\ge 1} \mathbb{E}\big(\sup_{g\in\mathscr{C}_k} |L_n(g) - L(g)| - \mathrm{pen}(n,k)\big)_+ \\
&\quad \le L^\star_k - L^\star + 2\mathbb{E}R_n(\mathscr{F}_k(X_1^n)) + \mathbb{E}\mathrm{pen}(n,k) + \sum_{k\ge 1} \mathbb{E}\big(\sup_{g\in\mathscr{C}_k} |L_n(g) - L(g)| - \mathrm{pen}(n,k)\big)_+ \\
&\qquad \text{(by Theorem 1)} \\
&\quad = L^\star_k - L^\star + 4\mathbb{E}R_n(\mathscr{F}_k(X_1^n)) + 4\sqrt{\frac{\log k}{n}} + \sum_{k\ge 1} \mathbb{E}\big(\sup_{g\in\mathscr{C}_k} |L_n(g) - L(g)| - \mathrm{pen}(n,k)\big)_+ \\
&\qquad \text{(by our choice of the penalty).}
\end{aligned}
$$

Now, observe that, for all $t > 0$,

$$
\begin{aligned}
&\mathbb{P}\big(\big(\sup_{g\in\mathscr{C}_k} |L_n(g) - L(g)| - \mathrm{pen}(n,k)\big)_+ \ge t\big) \\
&\quad = \mathbb{P}\big(\sup_{g\in\mathscr{C}_k} |L_n(g) - L(g)| - \mathrm{pen}(n,k) \ge t\big) \\
&\quad = \mathbb{P}\Big(\sup_{g\in\mathscr{C}_k} |L_n(g) - L(g)| \ge 2R_n(\mathscr{F}_k(X_1^n)) + 4\sqrt{\frac{\log k}{n}} + t\Big) \\
&\quad \le \mathbb{P}\Big(\sup_{g\in\mathscr{C}_k} |L_n(g) - L(g)| \ge \mathbb{E}\sup_{g\in\mathscr{C}_k} |L_n(g) - L(g)| + 2\sqrt{\frac{\log k}{n}} + \frac{t}{2}\Big) \\
&\qquad + \mathbb{P}\Big(R_n(\mathscr{F}_k(X_1^n)) \le \mathbb{E}R_n(\mathscr{F}_k(X_1^n)) - \sqrt{\frac{\log k}{n}} - \frac{t}{4}\Big) \\
&\qquad \text{(by Theorem 1).}
\end{aligned}
$$

Thus, by the bounded difference inequality,

$$
\mathbb{P}\big(\big(\sup_{g\in\mathscr{C}_k} |L_n(g) - L(g)| - \mathrm{pen}(n,k)\big)_+ \ge t\big) \le \frac{2}{k^2} e^{-nt^2/8}.
$$

Consequently,

$$
\begin{aligned}
\mathbb{E}\big(\sup_{g\in\mathscr{C}_k} |L_n(g) - L(g)| - \mathrm{pen}(n,k)\big)_+ &= \int_0^\infty \mathbb{P}\big(\big(\sup_{g\in\mathscr{C}_k} |L_n(g) - L(g)| - \mathrm{pen}(n,k)\big)_+ \ge t\big)\mathrm{d}t \\
&\le \frac{2}{k^2}\int_0^\infty e^{-nt^2/8}\mathrm{d}t \\
&= \frac{2}{k^2}\sqrt{\frac{2\pi}{n}},
\end{aligned}
$$

and so, recalling that $\sum_{k\geq 1}\frac{1}{k^2}=\frac{\pi^2}{6}$,

$$\sum_{k\geq 1}\mathbb{E}\big(\sup_{g\in\mathscr{C}_k}|L_n(g)-L(g)|-\mathrm{pen}(n,k)\big)\leq\frac{\sqrt{2}\pi^{5/2}}{3\sqrt{n}}.$$

Collecting bounds leads to the oracle inequality of the theorem. □

Theorem 7 (Classification and regression)**.** *Let r_n be a regression function estimate of r, and let g_n be the corresponding plug-in classifier. Then*

$$0\leq L(g_n)-L^\star\leq 2\int_{\mathbb{R}^d}|r_n(x)-r(x)|\mu(\mathrm{d}x).$$

In particular, for all $p\geq 1$,

$$0\leq L(g_n)-L^\star\leq 2\Big(\int_{\mathbb{R}^d}|r_n(x)-r(x)|^p\mu(\mathrm{d}x)\Big)^{1/p},$$

and

$$0\leq \mathbb{E}L(g_n)-L^\star\leq 2\,\mathbb{E}^{1/p}|r_n(X)-r(X)|^p.$$

Proof. Proceeding as in the proof of Lemma 1, we may write

$$\begin{aligned}
&\mathbb{P}(g_n(X)\neq Y|X,\mathscr{D}_n)\\
&\quad=1-\mathbb{P}(g_n(X)=Y|X,\mathscr{D}_n)\\
&\quad=1-\big(\mathbb{P}(g_n(X)=1,Y=1|X,\mathscr{D}_n)+\mathbb{P}(g_n(X)=0,Y=0|X,\mathscr{D}_n)\big)\\
&\quad=1-\big(\mathbb{1}_{[g_n(X)=1]}\mathbb{P}(Y=1|X,\mathscr{D}_n)+\mathbb{1}_{[g_n(X)=0]}\mathbb{P}(Y=0|X,\mathscr{D}_n)\big)\\
&\quad=1-\big(\mathbb{1}_{[g_n(X)=1]}r(X)+\mathbb{1}_{[g_n(X)=0]}(1-r(X))\big),
\end{aligned}$$

where, in the last equality, we used the independence of (X,Y) and $\mathscr{D}_n$. Similarly,

$$\mathbb{P}(g^\star(X)\neq Y|X)=1-\big(\mathbb{1}_{[g^\star(X)=1]}r(X)+\mathbb{1}_{[g^\star(X)=0]}(1-r(X))\big).$$

Therefore,

$$\begin{aligned}
&\mathbb{P}(g_n(X)\neq Y|X,\mathscr{D}_n)-\mathbb{P}(g^\star(X)\neq Y|X)\\
&\quad=r(X)(\mathbb{1}_{[g^\star(X)=1]}-\mathbb{1}_{[g_n(X)=1]})+(1-r(X))(\mathbb{1}_{[g^\star(X)=0]}-\mathbb{1}_{[g_n(X)=0]})\\
&\quad=(2r(X)-1)(\mathbb{1}_{[g^\star(X)=1]}-\mathbb{1}_{[g_n(X)=1]})\\
&\quad=|2r(X)-1|\mathbb{1}_{[g_n(X)\neq g^\star(X)]}.
\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{P}(g_n(X) \neq Y | \mathscr{D}_n) - L^\star &= 2\int_{\mathbb{R}^d} |r(x) - 1/2| \mathbb{1}_{[g_n(x) \neq g^\star(x)]} \mu(\mathrm{d}x) \\ &\leq 2\int_{\mathbb{R}^d} |r_n(x) - r(x)| \mu(\mathrm{d}x),\end{aligned}$$

since $g_n(x) \neq g^\star(x)$ implies $|r_n(x) - r(x)| \geq |r(x) - 1/2|$. The other assertions follow from Hölder's and Jensen's inequality, respectively. □

Theorem 8 (Stone's theorem)**.** *Assume that for any distribution of X, the weights satisfy the following conditions:*

(*i*) *There is a constant C such that, for every Borel measurable function $f : \mathbb{R}^d \to \mathbb{R}$ with $\mathbb{E}|f(X)| < \infty$,*

$$\mathbb{E}\Big(\sum_{i=1}^n W_{ni}(X)|f(X_i)|\Big) \leq C\mathbb{E}|f(X)| \quad \textit{for all } n \geq 1.$$

(*ii*) *For all $a > 0$,*

$$\mathbb{E}\Big(\sum_{i=1}^n W_{ni}(X)\mathbb{1}_{[\|X_i - X\| > a]}\Big) \to 0.$$

(*iii*) *One has*

$$\mathbb{E}\max_{1 \leq i \leq n} W_{ni}(X) \to 0.$$

Then the corresponding plug-in classifier g_n is universally consistent, i.e., $\mathbb{E}L(g_n) \to L^\star$ for all distributions of (X,Y).

Proof. According to Theorem 7, it suffices to prove that for every distribution of (X,Y)

$$\mathbb{E}|r_n(X) - r(X)|^2 = \mathbb{E}\int_{\mathbb{R}^d} |r_n(x) - r(x)|^2 \mu(\mathrm{d}x) \to 0.$$

Introduce the notation

$$\hat{r}_n(x) = \sum_{i=1}^n W_{ni}(x) r(X_i).$$

Then, by the simple inequality $(a+b)^2 \leq 2(a^2+b^2)$, we have

$$\begin{aligned}\mathbb{E}|r_n(X) - r(X)|^2 &= \mathbb{E}|r_n(X) - \hat{r}_n(X) + \hat{r}_n(X) - r(X)|^2 \\ &\leq 2\big(\mathbb{E}|r_n(X) - \hat{r}_n(X)|^2 + \mathbb{E}|\hat{r}_n(X) - r(X)|^2\big). \qquad (1)\end{aligned}$$

Therefore, it is enough to show that both terms on the right-hand side tend to zero as n tends to infinity. Since the W_{ni} are nonnegative and sum to one, by Jensen's inequality, the second term is

$$\begin{aligned}\mathbb{E}|\hat{r}_n(X) - r(X)|^2 &= \mathbb{E}\Big|\sum_{i=1}^n W_{ni}(X)(r(X_i) - r(X))\Big|^2 \\ &\le \mathbb{E}\Big(\sum_{i=1}^n W_{ni}(X)|r(X_i) - r(X)|^2\Big).\end{aligned}$$

If the function r, which satisfies $0 \le r \le 1$, is continuous with compact support, then it is uniformly continuous as well: for every $\varepsilon > 0$, there is an $a > 0$ such that for $\|x - x'\| \le a$, $|r(x) - r(x')|^2 \le \varepsilon$. Thus, since $|r(x) - r(x')| \le 1$,

$$\begin{aligned}\mathbb{E}\Big(\sum_{i=1}^n W_{ni}(X)|r(X_i) - r(X)|^2\Big) &\le \mathbb{E}\Big(\sum_{i=1}^n W_{ni}(X)\mathbb{1}_{[\|X_i - X\| > a]}\Big) + \mathbb{E}\Big(\sum_{i=1}^n W_{ni}(X)\varepsilon\Big) \\ &= \mathbb{E}\Big(\sum_{i=1}^n W_{ni}(X)\mathbb{1}_{[\|X_i - X\| > a]}\Big) + \varepsilon.\end{aligned}$$

Therefore, by (ii), since ε is arbitrary,

$$\mathbb{E}\Big(\sum_{i=1}^n W_{ni}(X)|r(X_i) - r(X)|^2\Big) \to 0.$$

In the general case, since the set of continuous functions with compact support is dense in $L^2(\mu)$, for every $\varepsilon > 0$ we can choose r_ε taking values in $[0,1]$ and such that

$$\mathbb{E}|r(X) - r_\varepsilon(X)|^2 \le \varepsilon.$$

By this choice, using the inequality $(a+b+c)^2 \le 3(a^2+b^2+c^2)$ (which follows from the Cauchy-Schwarz inequality),

$$\begin{aligned}&\mathbb{E}|\hat{r}_n(X) - r(X)|^2 \\ &\quad\le \mathbb{E}\Big(\sum_{i=1}^n W_{ni}(X)|r(X_i) - r(X)|^2\Big) \\ &\quad\le 3\mathbb{E}\Big(\sum_{i=1}^n W_{ni}(X)\big(|r(X_i) - r_\varepsilon(X_i)|^2 + |r_\varepsilon(X_i) - r_\varepsilon(X)|^2 + |r_\varepsilon(X) - r(X)|^2\big)\Big).\end{aligned}$$

Thus, using (i),

$$\begin{aligned}&\mathbb{E}|\hat{r}_n(X) - r(X)|^2 \\ &\quad\le 3C\mathbb{E}|r(X) - r_\varepsilon(X)|^2 + 3\mathbb{E}\Big(\sum_{i=1}^n W_{ni}(X)|r_\varepsilon(X_i) - r_\varepsilon(X)|^2\Big) + 3\mathbb{E}|r_\varepsilon(X) - r(X)|^2 \\ &\quad\le 3C\varepsilon + 3\mathbb{E}\Big(\sum_{i=1}^n W_{ni}(X)|r_\varepsilon(X_i) - r_\varepsilon(X)|^2\Big) + 3\varepsilon.\end{aligned}$$

Therefore, $\mathbb{E}|\hat{r}_n(X) - r(X)|^2 \to 0$.

To handle the first term of the right-hand side of (1), observe that, for all $i \neq j$,

$$\begin{aligned}
&\mathbb{E}\big(W_{ni}(X)(Y_i - r(X_i))W_{nj}(X)(Y_j - r(X_j))\big) \\
&\quad = \mathbb{E}\Big[\mathbb{E}\Big(W_{ni}(X)(Y_i - r(X_i))W_{nj}(X)(Y_j - r(X_j)) \,|\, X, X_1, \ldots, X_n, Y_i\Big)\Big] \\
&\quad = \mathbb{E}\Big[W_{ni}(X)(Y_i - r(X_i))W_{nj}(X)\mathbb{E}\big(Y_j - r(X_j) \,|\, X, X_1, \ldots, X_n, Y_i\big)\Big] \\
&\quad = \mathbb{E}\Big[W_{ni}(X)(Y_i - r(X_i))W_{nj}(X)\mathbb{E}(Y_j - r(X_j) \,|\, X_j)\Big] \\
&\qquad \text{(by independence of } (X_j, Y_j) \text{ and } X, X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_n, Y_i) \\
&\quad = \mathbb{E}\Big[W_{ni}(X)(Y_i - r(X_i))W_{nj}(X)(r(X_j) - r(X_j))\Big] \\
&\quad = 0.
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathbb{E}|r_n(X) - \hat{r}_n(X)|^2 &= \mathbb{E}\Big|\sum_{i=1}^n W_{ni}(X)(Y_i - r(X_i))\Big|^2 \\
&= \sum_{i,j=1}^n \mathbb{E}\big(W_{ni}(X)(Y_i - r(X_i))W_{nj}(X)(Y_j - r(X_j))\big) \\
&= \sum_{i=1}^n \mathbb{E}\big(W_{ni}^2(X)(Y_i - r(X_i))^2\big).
\end{aligned}$$

We conclude that

$$\begin{aligned}
\mathbb{E}|r_n(X) - \hat{r}_n(X)|^2 &\leq \mathbb{E}\sum_{i=1}^n W_{ni}^2(X) \leq \mathbb{E}\Big(\max_{1\leq i\leq n} W_{ni}(X) \sum_{j=1}^n W_{nj}(X)\Big) \\
&= \mathbb{E}\max_{1\leq i\leq n} W_{ni}(X) \to 0
\end{aligned}$$

by (iii), and the theorem is proved. □

Lemma 6. *If $x \in supp(\mu)$ and $k/n \to 0$, then*

$$\|X_{(k)}(x) - x\| \to 0 \quad \textit{almost surely}.$$

Proof. Take $\varepsilon > 0$ and note, since x belongs to the support of μ, that $\mu(B(x,\varepsilon)) > 0$. Observe that

$$\Big[\|X_{(k)}(x) - x\| > \varepsilon\Big] = \Big[\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{[X_i \in B(x,\varepsilon)]} < \frac{k}{n}\Big].$$

By the strong law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[X_i \in B(x,\varepsilon)]} \to \mu(B(x,\varepsilon)) \quad \text{almost surely.}$$

Since $k/n \to 0$, we conclude that $\|X_{(k)}(x) - x\| \to 0$ almost surely. □

Lemma 7. *Let ν be a probability measure on $\mathbb{R}^d$. Fix $x' \in \mathbb{R}^d$ and let, for $a \geq 0$,*

$$B_a(x') = \Big\{ x \in \mathbb{R}^d : \nu\big(B(x, \|x' - x\|)\big) \leq a \Big\}.$$

Then

$$\nu(B_a(x')) \leq \gamma_d a,$$

where γ_d is a positive constant depending only upon d.

Proof. Fix $x' \in \mathbb{R}^d$ and let $\mathscr{C}_1, \ldots, \mathscr{C}_{\gamma_d}$ be a collection of cones of angle $0 < \theta \leq \pi/6$ covering $\mathbb{R}^d$, all centered at x' but with different central directions (such a covering is always possible). In other words,

$$\bigcup_{j=1}^{\gamma_d} \mathscr{C}_j = \mathbb{R}^d.$$

We leave it as an easy exercise to show that if $u \in \mathscr{C}_j$, $u' \in \mathscr{C}_j$, and $\|u - x'\| \leq \|u' - x'\|$, then $\|u - u'\| \leq \|u' - x'\|$ (see Figure 1).

In addition,

$$\nu(B_a(x')) \leq \sum_{j=1}^{\gamma_d} \nu(\mathscr{C}_j \cap B_a(x')).$$

Let $x^\star \in \mathscr{C}_j \cap B_a(x')$. Then, by the geometrical property of cones mentioned above, we have

$$\nu\big(\mathscr{C}_j \cap B(x', \|x^\star - x'\|) \cap B_a(x')\big) \leq \nu\big(B(x^\star, \|x' - x^\star\|)\big) \leq a.$$

Since $x^\star$ was arbitrary, we conclude that

$$\nu(\mathscr{C}_j \cap B_a(x')) \leq a.$$

□

Corollary 1. *If distance ties occur with zero probability, then*

$$\sum_{i=1}^{n} \mathbb{1}_{[X \text{ is among the k-NN of } X_i \text{ in } \{X_1, \ldots, X_{i-1}, X, X_{i+1}, \ldots, X_n\}]} \leq k\gamma_d,$$

with probability one.

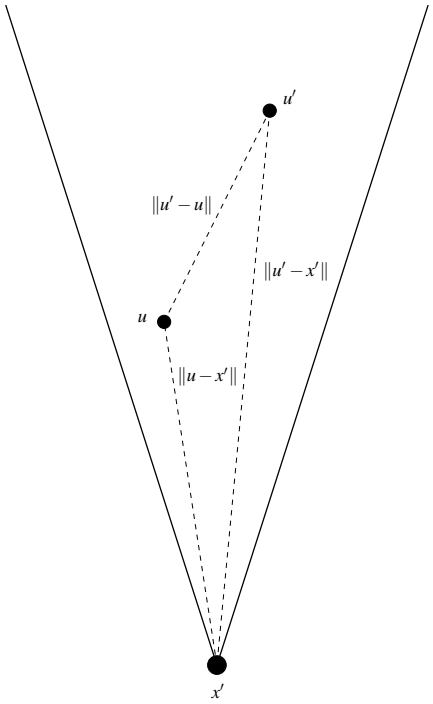

Figure 1: The geometrical property of a cone of angle $0 < \theta \leq \pi/6$ (in dimension 2).

Proof. We apply Lemma 7 with $a = k/n$ and ν the empirical measure μ_n associated with $X_1, \ldots, X_n$. With these choices,

$$B_{k/n}(X) = \Big\{ x \in \mathbb{R}^d : \mu_n\big(B(x, \|X - x\|)\big) \leq k/n \Big\}$$

and, with probability one,

$$\begin{aligned}
&X_i \in B_{k/n}(X) \\
&\Leftrightarrow \mu_n\big(B(X_i, \|X - X_i\|)\big) \leq k/n \\
&\Leftrightarrow X \text{ is among the } k\text{-NN of } X_i \text{ in } \{X_1, \ldots, X_{i-1}, X, X_{i+1}, \ldots, X_n\}.
\end{aligned}$$

(Note that the second equivalence uses the fact that distance ties occur with zero probability.) Thus, by Lemma 7, we conclude that, with probability one,

$$\begin{aligned}
&\sum_{i=1}^{n} \mathbb{1}_{[X \text{ is among the } k\text{-NN of } X_i \text{ in } \{X_1, \ldots, X_{i-1}, X, X_{i+1}, \ldots, X_n\}]} \\
&\quad = \sum_{i=1}^{n} \mathbb{1}_{[X_i \in B_{k/n}(X)]} = n \times \mu_n(B_{k/n}(X)) \leq k\gamma_d.
\end{aligned}$$

□

Lemma 8 (Stone's lemma)**.** *Assume that distance ties occur with zero probability. Then, for every Borel measurable function $f : \mathbb{R}^d \to \mathbb{R}$ such that $\mathbb{E}|f(X)| < \infty$, we have*

$$\sum_{i=1}^{k} \mathbb{E}\big|f(X_{(i)}(X))\big| \le k\gamma_d \mathbb{E}|f(X)|,$$

where γ_d is a positive constant depending only upon d.

Proof. Take f as in the lemma. Then

$$\begin{aligned}
&\sum_{i=1}^{k} \mathbb{E}\big|f(X_{(i)}(X))\big| \\
&= \mathbb{E}\Big(\sum_{i=1}^{n} |f(X_i)| \mathbb{1}_{[X_i \text{ is among the } k\text{-NN of } X \text{ in } \{X_1,\ldots,X_n\}]}\Big) \\
&= \mathbb{E}\Big(|f(X)| \sum_{i=1}^{n} \mathbb{1}_{[X \text{ is among the } k\text{-NN of } X_i \text{ in } \{X_1,\ldots,X_{i-1},X,X_{i+1},\ldots,X_n\}]}\Big) \\
&\quad \text{(by exchanging } X \text{ and } X_i\text{)} \\
&\le \mathbb{E}(|f(X)| k\gamma_d),
\end{aligned}$$

by Corollary 1. □

Theorem 9. *Let g_n be a partitioning classifier with the X-property. If*

(*i*) *$diam(A(X)) \to 0$ in probability,*

and

(*ii*) *$N(X) \to \infty$ in probability,*

then $\mathbb{E}L(g_n) \to L^\star$.

Proof. Let $r(x) = \mathbb{E}(Y|X = x)$. From Theorem 7, we recall that we need only show that $\mathbb{E}|r_n(X) - r(X)| \to 0$, where

$$r_n(x) = \frac{1}{N(x)} \sum_{i=1}^{n} \mathbb{1}_{[X_i \in A(x)]} Y_i.$$

Introduce $\bar{r}(x) = \mathbb{E}(r(X)\,|\,X \in A(x))$. By the triangle inequality,

$$\mathbb{E}|r_n(X) - r(X)| \le \mathbb{E}|r_n(X) - \bar{r}(X)| + \mathbb{E}|\bar{r}(X) - r(X)|.$$

By conditioning on the random variable $N(x)$, and upon noticing that $\mathbb{P}(Y=1|X\in A(x)) = \bar{r}(x)$, it is easy to see that $N(x)r_n(x)$ is distributed as $\mathrm{Bin}(N(x),\bar{r}(x))$, a binomial random variable with parameters $N(x)$ and $\bar{r}(x)$. Thus,

$$\begin{aligned}
&\mathbb{E}\big(|r_n(X)-\bar{r}(X)|\,\big|\,X,\mathbb{1}_{[X_1\in A(X)]},\ldots,\mathbb{1}_{[X_n\in A(X)]}\big)\\
&\quad\le \mathbb{E}\Big(\Big|\frac{\mathrm{Bin}(N(X),\bar{r}(X))}{N(X)}-\bar{r}(X)\Big|\mathbb{1}_{[N(X)>0]}\,\Big|\,X,\mathbb{1}_{[X_1\in A(X)]},\ldots,\mathbb{1}_{[X_n\in A(X)]}\Big)+\mathbb{1}_{[N(X)=0]}\\
&\quad\le \sqrt{\frac{\bar{r}(X)(1-\bar{r}(X))}{N(X)}}\mathbb{1}_{[N(X)>0]}+\mathbb{1}_{[N(X)=0]},
\end{aligned}$$

by the Cauchy-Schwarz inequality. Taking expectations, we see that

$$\mathbb{E}|r_n(X)-\bar{r}(X)| \le \mathbb{E}\Big(\frac{1}{2\sqrt{N(X)}}\mathbb{1}_{[N(X)>0]}\Big)+\mathbb{P}(N(X)=0).$$

Both terms on the right-hand side tend to zero as n tends to infinity by condition (ii).

Next, for $\varepsilon>0$, find a uniformly continuous $[0,1]$-valued function r_ε with compact support so that $\mathbb{E}|r(X)-r_\varepsilon(X)|\le\varepsilon$. By the triangle inequality,

$$\begin{aligned}
\mathbb{E}|\bar{r}(X)-r(X)| &\le \mathbb{E}|\bar{r}(X)-\bar{r}_\varepsilon(X)|+\mathbb{E}|\bar{r}_\varepsilon(X)-r_\varepsilon(X)|+\mathbb{E}|r_\varepsilon(X)-r(X)|\\
&\stackrel{\mathrm{def}}{=} \mathbf{I}+\mathbf{II}+\mathbf{III},
\end{aligned}$$

where $\bar{r}_\varepsilon(x)=\mathbb{E}(r_\varepsilon(X)\,|\,X\in A(x))$. Clearly, $\mathbf{III}\le\varepsilon$ by choice of r_ε. Observe that, for all x,

$$\Big|\frac{1}{\mu(A(x))}\int_{A(x)} r_\varepsilon(z)\mu(\mathrm{d}z)-r_\varepsilon(x)\Big| \le \frac{1}{\mu(A(x))}\int_{A(x)}|r_\varepsilon(z)-r_\varepsilon(x)|\mu(\mathrm{d}z).$$

Thus, since r_ε is uniformly continuous, we can find a $\theta=\theta(\varepsilon)>0$ such that

$$\mathbf{II}\le\varepsilon+\mathbb{P}(\mathrm{diam}(A(X))>\theta).$$

Therefore, $\mathbf{II}\le 2\varepsilon$ for all n large enough, by condition (i). Finally,

$$\mathbf{I}\le\int_{\mathbb{R}^d}\mathbb{E}\big(|r(X)-r_\varepsilon(X)|\,\big|\,X\in A(x)\big)\mu(\mathrm{d}x)=\mathbf{III}\le\varepsilon.$$

Taken together these steps prove the theorem. □

Theorem 10. *Assume that $h\to 0$ and $nh^d\to\infty$. Then the cubic histogram classifier is universally consistent, i.e., $\mathbb{E}L(g_n)\to L^\star$ for all distributions of (X,Y).*

Proof. We check the two simple conditions of Theorem 9. Clearly, the diameter of each cell is $\sqrt{d}h$. Therefore condition (i) follows trivially. To show condition (ii), we need to prove that for any $M < \infty$, $\mathbb{P}(N(X) \le M) \to 0$. Let S be an arbitrary ball centered at the origin. Then the number of cells intersecting S is not more than $c_1 + c_2/h^d$ for some positive constants c_1, c_2. Let μ_n be the empirical measure associated with $X_1, \ldots, X_n$. Then

$$
\begin{aligned}
&\mathbb{P}(N(X) \le M) \\
&\quad \le \sum_{j: A_{nj} \cap S \neq \emptyset} \mathbb{P}(X \in A_{nj}, N(X) \le M) + \mathbb{P}(X \in S^c) \\
&\quad \le \sum_{\substack{j: A_{nj} \cap S \neq \emptyset \\ \mu(A_{nj}) \le 2M/n}} \mu(A_{nj}) + \sum_{\substack{j: A_{nj} \cap S \neq \emptyset \\ \mu(A_{nj}) > 2M/n}} \mu(A_{nj}) \mathbb{P}(n\mu_n(A_{nj}) \le M) + \mu(S^c) \\
&\quad \le \frac{2M}{n}\left(c_1 + \frac{c_2}{h^d}\right) + \sum_{\substack{j: A_{nj} \cap S \neq \emptyset \\ \mu(A_{nj}) > 2M/n}} \mu(A_{nj}) \mathbb{P}\left(\mu_n(A_{nj}) - \mu(A_{nj}) \le M/n - \mu(A_{nj})\right) \\
&\qquad + \mu(S^c) \\
&\quad \le \frac{2M}{n}\left(c_1 + \frac{c_2}{h^d}\right) + \sum_{\substack{j: A_{nj} \cap S \neq \emptyset \\ \mu(A_{nj}) > 2M/n}} \mu(A_{nj}) \mathbb{P}\left(\mu_n(A_{nj}) - \mu(A_{nj}) \le -\mu(A_{nj})/2\right) + \mu(S^c) \\
&\quad \le \frac{2M}{n}\left(c_1 + \frac{c_2}{h^d}\right) + \sum_{\substack{j: A_{nj} \cap S \neq \emptyset \\ \mu(A_{nj}) > 2M/n}} 4\mu(A_{nj}) \frac{\mathbf{Var}(\mu_n(A_{nj}))}{(\mu(A_{nj}))^2} + \mu(S^c) \\
&\qquad \text{(by Chebyshev's inequality)} \\
&\quad \le \frac{2M}{n}\left(c_1 + \frac{c_2}{h^d}\right) + \sum_{\substack{j: A_{nj} \cap S \neq \emptyset \\ \mu(A_{nj}) > 2M/n}} 4\mu(A_{nj}) \frac{1}{n\mu(A_{nj})} + \mu(S^c) \\
&\quad \le \frac{2M+4}{n}\left(c_1 + \frac{c_2}{h^d}\right) + \mu(S^c) \\
&\quad \to \mu(S^c),
\end{aligned}
$$

because $nh^d \to \infty$. Since S is arbitrary, the proof of the theorem is complete. □

Theorem 11. *Assume that X has a density. If $k \to \infty$ and $\frac{n}{k2^k} \to \infty$, then the median tree classifier is consistent, i.e., $\mathbb{E}L(g_n) \to L^\star$. (Note: the conditions on k are fulfilled if $k \le \log_2 n - 2\log_2 \log_2 n$, $k \to \infty$.)*

Proof. We may prove the theorem by checking the conditions of Theorem 9. Condition (ii) follows trivially by the fact that each leaf region contains at least $n/2^k - 2$ points (exercise) and the requirement $\frac{n}{k2^k} \to \infty$. Thus, we need only verify the first condition of Theorem

9. To make the proof more transparent, we first analyze a closely related hypothetical tree, the theoretical median tree. Also, we restrict the analysis to $d = 2$. The multidimensional extension is straightforward. The theoretical median tree rotates through the coordinates and cuts each hyperrectangle precisely so that the two new hyperrectangles have equal μ-measure (see Figure 2 for an example).

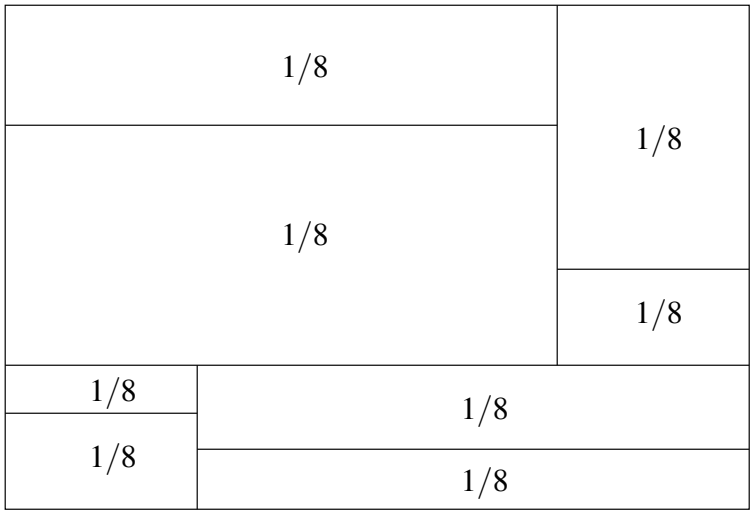

Figure 2: Theoretical median tree with three levels of cuts.

Observe that the rule is invariant under monotone transformations of the coordinate axes. Recall that in such cases there is no harm in assuming that the marginal distributions are all uniform on $[0,1]$. We let $\{H_i, V_i\}$ denote the horizontal and vertical sizes of the rectangles after k levels of cuts. Of course, we begin with $H_1 = V_1 = 1$ when $k = 0$. We now show that, for the theoretical median tree, $\mathrm{diam}(A(X)) \to 0$ in probability, as $k \to \infty$. Note that $\mathrm{diam}(A(X)) \le H(X) + V(X)$, where $H(X)$ and $V(X)$ are the horizontal and vertical sizes of the rectangle $A(X)$. We show that if k is even,

$$\mathbb{E}(H(X) + V(X)) = \frac{2}{2^{k/2}},$$

from which the claim follows. After the k-th round of splits, since all 2^k rectangles have equal probability measure, we have

$$\mathbb{E}(H(X) + V(X)) = \sum_{i=1}^{2^k} \frac{1}{2^k}(H_i + V_i).$$

Apply another round of splits, all vertical. Then each term $\frac{1}{2^k}(H_i + V_i)$ spawns, so to speak, two new rectangles with horizontal and vertical sizes $(H_i', \tilde{V}_i)$ and (H_i''', V_i) with $H_i = H_i' + H_i'''$ that contribute

$$\frac{1}{2^{k+1}}(H_i' + V_i) + \frac{1}{2^{k+1}}(H_i''' + V_i) = \frac{1}{2^{k+1}}H_i + \frac{1}{2^k}V_i.$$

The next round yields horizontal splits, with total contribution now (see Figure 3)

$$\frac{1}{2^{k+2}}(H_i' + V_i' + H_i' + V_i'' + H_i''' + V_i''' + H_i''' + V_i'''') = \frac{1}{2^{k+2}}(2H_i + 2V_i) = \frac{1}{2^{k+1}}(H_i + V_i).$$

Thus, over two iterations of splits, we see that $\mathbb{E}(H(X) + V(X))$ is halved, and the claim follows by simple induction.

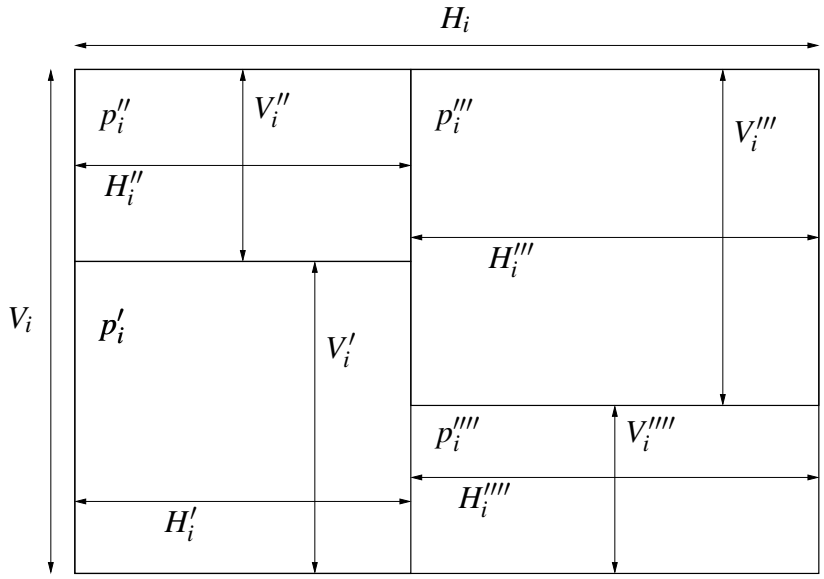

Figure 3: A rectangle after two rounds of splits.

We show now what happens in the real median tree when cuts are based upon a random sample. We deviate of course from the theoretical median tree, but consistency is preserved. The reason, seen intuitively, is that if the number of points in a cell is large, then the sample median will be close to the theoretical median, so that the shrinking-diameter property is preserved. The methodology followed here shows how one may approach the analysis in general by separating the theoretical model from the sample-based model.

We recall the following inequality, due to Massart (1990): let $Z_1, \ldots, Z_n$ be i.i.d. real-valued random variables with distribution function $F(z) = \mathbb{P}(Z_1 \leq z)$, and let $F_n(z) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{[Z_i \leq z]}$ be the empirical distribution function—then, for all n and all $t > 0$,

$$\mathbb{P}\left(\sup_{z \in \mathbb{R}} |F_n(z) - F(z)| > t\right) \leq 2e^{-2nt^2}. \tag{2}$$

As we noted before, all we have to show is that $\mathrm{diam}(A(X)) \to 0$ in probability. Again, we assume without loss of generality that the marginals of X are uniform $[0,1]$, and that $d = 2$.

We show that $\mathbb{E}(H(X) + V(X)) \to 0$. If a rectangle of probability mass p_i and sizes H_i, V_i is split into four rectangles as in Figure 3, with probability masses p_i', p_i'', p_i''', p_i'''', then the

contribution $p_i(H_i+V_i)$ to $\mathbb{E}(H(X)+V(X))$ becomes

$$p_i'(H_i'+V_i')+p_i''(H_i''+V_i'')+p_i'''(H_i'''+V_i''')+p_i''''(H_i''''+V_i'''')$$

after two levels of cuts. This does not exceed

$$\frac{p_i}{2}(1+\varepsilon)(H_i+V_i),$$

if

$$\max(p_i'+p_i'',p_i'''+p_i'''')\le\frac{1}{2}\sqrt{1+\varepsilon}p_i,$$

$$\max(p_i',p_i'')\le\frac{1}{2}\sqrt{1+\varepsilon}(p_i'+p_i''),$$

and

$$\max(p_i''',p_i'''')\le\frac{1}{2}\sqrt{1+\varepsilon}(p_i'''+p_i''''),$$

that is, when all three cuts are within $(1/2)\sqrt{1+\varepsilon}$ of the true median. We call such "$(1/2)\sqrt{1+\varepsilon}$" cuts good. If all cuts are good, we thus note that in two levels of cuts, $\mathbb{E}(H(X)+V(X))$ is reduced by $(1+\varepsilon)/2$. Also, all p_i decrease at a controlled rate. Let G be the event that all $1+2+\cdots+2^{k-1}$ cuts in a median tree with k levels are good. Then, at level k (even), all p_i are at most $(\sqrt{1+\varepsilon}/2)^k$. Thus,

$$\begin{aligned}\sum_{i=1}^{2^k}p_i(H_i+V_i)&\le\Big(\frac{\sqrt{1+\varepsilon}}{2}\Big)^k\sum_{i=1}^{2^k}(H_i+V_i)\\&\le 2^{k/2+1}\Big(\frac{\sqrt{1+\varepsilon}}{2}\Big)^k\end{aligned}$$

since $\sum_{i=1}^{2^k}(H_i+V_i)=2+2+4+8+\cdots+2^{k/2}=2^{k/2+1}$, if k is even. Hence, after k levels of cuts,

$$\mathbb{E}(H(X)+V(X))\le 2^{k/2+1}\mathbb{P}(G^c)+2^{k/2+1}\Big(\frac{\sqrt{1+\varepsilon}}{2}\Big)^k.$$

The last term tends to zero if ε is small enough. We bound $\mathbb{P}(G^c)$ by 2^k times the probability that one cut is bad. Let us cut a cell with N points and probability content p in a given direction. A quick check of the median tree shows that given the position and size of the cell, the N points inside the cell are distributed in an i.i.d. manner according to the restriction of μ to the cell. After the cut, we have $\lfloor(N-1)/2\rfloor$ and $\lceil(N-1)/2\rceil$ points in the new cells, and probability contents p' and p''. It is clear that we may assume without loss of generality that $p=1$. Thus, if all points are projected down in the direction of the cut, and F and F_N denote the distribution function and empirical distribution function of

the obtained one-dimensional data, then

$$\begin{aligned}
\mathbb{P}(\text{cut is not good}\,|\,N) &\le \mathbb{P}\Big(p' > \frac{\sqrt{1+\varepsilon}}{2} \text{ or } p'' > \frac{\sqrt{1+\varepsilon}}{2}\,\Big|\,N\Big) \\
&\le \mathbb{P}\Big(\text{for some } x,\, F(x) > \frac{\sqrt{1+\varepsilon}}{2},\, \text{and } F_N(x) \le \frac{1}{2}\,\Big|\,N\Big) \\
&\le \mathbb{P}\Big(\sup_x (F(x) - F_N(x)) > \frac{1}{2}(\sqrt{1+\varepsilon}-1)\,\Big|\,N\Big) \\
&\le 2\exp\Big(-\frac{1}{2}N\big(\sqrt{1+\varepsilon}-1\big)^2\Big) \\
&\quad \text{(by inequality (2))} \\
&\le 2\exp\Big(-\Big(\frac{n}{2^{k+1}}-1\Big)\big(\sqrt{1+\varepsilon}-1\big)^2\Big) \\
&\quad \text{(as } N \ge n/(2^k) - 2\text{).}
\end{aligned}$$

Hence, for all n large enough,

$$\mathbb{P}(G^c) \le 2^{k+1} e^{-(n/2^{k+2})(\sqrt{1+\varepsilon}-1)^2}$$

and $2^{k/2}\mathbb{P}(G^c) \to 0$ if $n/(k2^k) \to \infty$. □

Lemma 9. *Let $(\mathscr{F}_k)_k$ be a sequence of classes of functions $f : \mathbb{R}^d \to \mathbb{R}$, and let $(\mathscr{C}_k)_k$ be the companion sequence of classes of classifiers. Assume that for every $a, b \in \mathbb{R}^d$ and every continuous function h on $[a,b]^d$,*

$$\lim_{k\to\infty} \inf_{f\in\mathscr{F}_k} \sup_{x\in[a,b]^d} |h(x) - f(x)| = 0.$$

Then, for any distribution of (X,Y),

$$\lim_{k\to\infty} \inf_{g\in\mathscr{C}_k} L(g) - L^\star = 0.$$

Proof. For fixed $\varepsilon > 0$, find a, b such that $\mu([a,b]^d) \ge 1 - \varepsilon/3$, where μ is the probability measure of X. Choose a continuous function r_ε vanishing off $[a,b]^d$ such that

$$\mathbb{E}|r(X) - r_\varepsilon(X)| \le \frac{\varepsilon}{6}.$$

For all k large enough, find $f \in \mathscr{F}_k$ such that

$$\sup_{x\in[a,b]^d} |r_\varepsilon(x) - f(x)| \le \frac{\varepsilon}{6}.$$

For $g(x) = \mathbb{1}_{[f(x)>1/2]}$, we have, by an easy adaptation of Theorem 7,

$$\begin{aligned} L(g) - L^\star &\le 2\mathbb{E}|f(X) - r(X)|\mathbb{1}_{[X\in[a,b]^d]} + \frac{\varepsilon}{3} \\ &\le 2\mathbb{E}|f(X) - r_\varepsilon(X)|\mathbb{1}_{[X\in[a,b]^d]} + 2\mathbb{E}|r_\varepsilon(X) - r(X)| + \frac{\varepsilon}{3} \\ &\le 2 \sup_{x\in[a,b]^d} |f(x) - r_\varepsilon(x)| + 2\mathbb{E}|r_\varepsilon(X) - r(X)| + \frac{\varepsilon}{3} \\ &\le \varepsilon. \end{aligned}$$

□

Theorem 12. *For every continuous function $h : [a,b]^d \to \mathbb{R}$ and for every $\varepsilon > 0$, there exists a neural network with one hidden layer, of the form*

$$\psi(x) = \sum_{i=1}^{k} c_i \sigma(\psi_i(x)) + c_0,$$

such that

$$\sup_{x\in[a,b]^d} |h(x) - \psi(x)| \le \varepsilon.$$

Proof. We prove the theorem for the threshold activation function $\sigma(x) = 2\mathbb{1}_{[x\ge 0]} - 1$. The extension to general nondecreasing activations is left as an exercise. Fix $\varepsilon > 0$. We take the Fourier series approximation of $h(x)$. By the Stone-Weierstrass theorem, there exists a large positive integer M, nonzero real coefficients $\alpha_1, \ldots, \alpha_M, \beta_1, \ldots, \beta_M$, and real numbers $m_{i,j}$ for $i = 1, \ldots, M$, $j = 1, \ldots, d$, such that

$$\sup_{x\in[a,b]^d} \left| h(x) - \sum_{i=1}^{M} (\alpha_i \cos(m_i^\top x) + \beta_i \sin(m_i^\top x)) \right| \le \frac{\varepsilon}{2},$$

where $m_i = (m_{i,1}, \ldots, m_{i,d})$, $1 \le i \le M$. It is clear that every continuous function on the real line can be arbitrarily closely approximated uniformly on compact intervals by one-dimensional neural networks, i.e., by functions of the form

$$\sum_{i=1}^{k} c_i \sigma(a_i x + b_i) + c_0.$$

Just observe that the indicator function of an interval $[b,c]$ may be written as $1/2(\sigma(x - b) + \sigma(-x + c))$. This implies that the sin and cos functions can be approximated arbitrarily

closely by neural networks on compact intervals. In particular, there exist neural networks $u_i(x)$, $v_i(x)$ with $i = 1, \ldots, M$, (i.e., mappings from $\mathbb{R}^d$ to $\mathbb{R}$) such that

$$\sup_{x \in [a,b]^d} |\cos(m_i^\top x) - u_i(x)| \leq \frac{\varepsilon}{4M|\alpha_i|}$$

and

$$\sup_{x \in [a,b]^d} |\sin(m_i^\top x) - v_i(x)| \leq \frac{\varepsilon}{4M|\beta_i|}.$$

Therefore, applying the triangle inequality we get

$$\sup_{x \in [a,b]^d} \left| \sum_{i=1}^M (\alpha_i \cos(m_i^\top x) + \beta_i \sin(m_i^\top x)) - \sum_{i=1}^M (\alpha_i u_i(x) + \beta_i v_i(x)) \right| \leq \frac{\varepsilon}{2}.$$

Since the u_i and v_i are neural networks, their linear combination

$$\psi(x) = \sum_{i=1}^M (\alpha_i u_i(x) + \beta_i v_i(x))$$

is a neural network too and, in fact,

$$\begin{aligned} \sup_{x \in [a,b]^d} |h(x) - \psi(x)| \leq & \sup_{x \in [a,b]^d} \left| h(x) - \sum_{i=1}^M (\alpha_i \cos(m_i^\top x) + \beta_i \sin(m_i^\top x)) \right| \\ & + \sup_{x \in [a,b]^d} \left| \sum_{i=1}^M (\alpha_i \cos(m_i^\top x) + \beta_i \sin(m_i^\top x)) - \psi(x) \right| \\ & \leq \frac{2\varepsilon}{2} = \varepsilon. \end{aligned}$$

□

Lemma 10. *One has $D_k^\star(\mu) \downarrow 0$ as $k \to \infty$.*

Proof. Clearly, the minimal distortion is a nonincreasing function of the order k. Since $\mathbb{R}^d$ is a Polish space, the bounded measure ν defined for every Borel subset A of $\mathbb{R}^d$ by

$$\nu(A) = \int_A \|x\|^2 \mu(\mathrm{d}x)$$

is tight, i.e., for all $\varepsilon \in (0,1]$ there exists a compact K with $\nu(K) \geq 1 - \varepsilon$. Let $\{c_1, c_2, \ldots\}$ be a countable and dense subset of $\mathbb{R}^d$. Since K is compact, one has for all k large enough

$$K \subseteq B \stackrel{\text{def}}{=} \bigcup_{j=1}^k B(c_j, \sqrt{\varepsilon}).$$

Thus, $\nu(B) \geq 1-\varepsilon$. Define now q_{k+1} as the quantizer of order $k+1$ with codebook $\{c_1,\ldots,c_k,0\}$ (assuming, without loss of generality, that $0 \notin \{c_1,c_2,\ldots\}$) and partition $\{A_1,\ldots,A_k,B^c\}$, with $A_1 = B(c_1,\sqrt{\varepsilon})$, and, for $j=2,\ldots,k$, $A_j = B(c_j,\sqrt{\varepsilon}) \setminus A_{j-1}$. Since $\|x-c_j\| \leq \sqrt{\varepsilon}$ when $x \in A_j$, we have

$$\begin{aligned} D^\star_{k+1}(\mu) \leq D_{k+1}(\mu,q_{k+1}) &= \int_{\mathbb{R}^d} \|x-q_{k+1}(x)\|^2 \mu(\mathrm{d}x) \\ &= \sum_{j=1}^k \int_{A_j} \|x-c_j\|^2 \mu(\mathrm{d}x) + \int_{B^c} \|x\|^2 \mu(\mathrm{d}x) \\ &\leq \varepsilon\mu\Big(\bigcup_{j=1}^k A_j\Big) + \nu(B^c) \leq 2\varepsilon, \end{aligned}$$

which concludes the proof. □

Proposition 2. *Let q_{NN} be a NN quantizer with codebook $\mathscr{C} = \{c_1,\ldots,c_k\}$. Then*

$$D(\mu,q_{\mathrm{NN}}) = \mathbb{E} \min_{1\leq j\leq k} \|X-c_j\|^2 = \int_{\mathbb{R}^d} \min_{1\leq j\leq k} \|x-c_j\|^2 \mu(\mathrm{d}x).$$

In addition, for any quantizer $q = (\mathscr{C},\mathscr{P})$, $D(\mu,q_{\mathrm{NN}}) \leq D(\mu,q)$.

NN

Proof. Let $\mathscr{P}_{\mathrm{V}}(\mathscr{C}) = \{A_{\mathrm{V},1},\ldots,A_{\mathrm{V},k}\}$ be the Voronoi partition associated with $\mathscr{C}$. Then

$$\begin{aligned} D(\mu,q_{\mathrm{NN}}) &= \int_{\mathbb{R}^d} \|x-q_{\mathrm{NN}}(x)\|^2 \mu(\mathrm{d}x) = \sum_{j=1}^k \int_{A_{\mathrm{V},j}} \|x-c_j\|^2 \mu(\mathrm{d}x) \\ &= \int_{\mathbb{R}^d} \min_{1\leq j\leq k} \|x-c_j\|^2 \mu(\mathrm{d}x). \end{aligned}$$

This shows the first statement. Next, for $\mathscr{P} = \{A_1,\ldots,A_k\}$,

$$\begin{aligned} D(\mu,q_{\mathrm{NN}}) &= \sum_{j=1}^k \int_{A_j} \min_{1\leq j\leq k} \|x-c_j\|^2 \mu(\mathrm{d}x) \\ &\leq \sum_{j=1}^k \int_{A_j} \|x-c_j\|^2 \mu(\mathrm{d}x) \\ &= \int_{\mathbb{R}^d} \|x-q(x)\|^2 \mu(\mathrm{d}x) = D(\mu,q), \end{aligned}$$

by definition of the distortion. □

Theorem 13. *There exists a quantizer with minimal distortion.*

Sketch of proof. According to Proposition 2, we have to prove that there exists $\mathbf{c}^\star \in \mathbb{R}^{dk}$ such that

$$W(\mu, \mathbf{c}^\star) = \inf_{\mathbf{c} \in \mathbb{R}^{dk}} W(\mu, \mathbf{c}).$$

One first shows (omitted) that there exists an $R > 0$ such that

$$\inf_{\mathbf{c} \in \mathbb{R}^{dk}} W(\mu, \mathbf{c}) = \inf_{\|\mathbf{c}\| \leq R} W(\mu, \mathbf{c}).$$

Then we prove that the function $\mathbb{R}^{dk} \ni \mathbf{c} \mapsto W(\mu, \mathbf{c})$ is continuous. To this aim, observe that the function $x \mapsto \min_{1 \leq j \leq k} \|x - c_j\|$ is continuous. Therefore, for $\mathbf{c}_0 = (c_{1,0}, \ldots, c_{k,0}) \in \mathbb{R}^{dk}$, one has

$$\begin{aligned}
\lim_{\mathbf{c} \to \mathbf{c}_0} W(\mu, \mathbf{c}) &= \int_{\mathbb{R}^d} \lim_{\mathbf{c} \to \mathbf{c}_0} \min_{1 \leq j \leq k} \|x - c_j\|^2 \mu(\mathrm{d}x) \\
&\quad \text{(by the Lebesgue dominated convergence theorem)} \\
&= \int_{\mathbb{R}^d} \min_{1 \leq j \leq k} \|x - c_{j,0}\|^2 \mu(\mathrm{d}x) \\
&\quad \text{(by continuity)} \\
&= W(\mu, \mathbf{c}_0),
\end{aligned}$$

which shows that $W(\mu, \cdot)$ is continuous.

It follows from the continuity of $W(\mu, \cdot)$ and the compactness of the ball $B(0, R)$ of $\mathbb{R}^{dk}$ that the infimum of $W(\mu, \cdot)$ is achieved at some $\mathbf{c}^\star \in \mathbb{R}^{dk}$. But then the quantizer $q^\star = (\mathbf{c}^\star, \mathscr{P}_V(\mathbf{c}^\star))$ has minimal distortion since

$$W(\mu, \mathbf{c}^\star) = \inf_{\mathbf{c} \in \mathbb{R}^{dk}} W(\mu, c) = \inf_q D(\mu, q) = D^\star(\mu).$$

□

Proposition 3. *Let ν_1 and ν_2 be probability measures on $\mathbb{R}^d$ with finite second moment. If q is a NN quantizer, then*

$$\left| D(\nu_1, q)^{1/2} - D(\nu_2, q)^{1/2} \right| \leq \rho_W(\nu_1, \nu_2).$$

Proof. Let (X_0, Y_0) be such that $X_0 \overset{\mathscr{D}}{=} \nu_1$, $Y_0 \overset{\mathscr{D}}{=} \nu_2$, and

$$\rho_W(\nu_1, \nu_2) = \sqrt{\mathbb{E}\|X_0 - Y_0\|^2}.$$

For $q = (\mathbf{c}, \mathscr{P}_{\mathrm{V}}(\mathbf{c}))$, one has

$$\begin{aligned}
D(\nu_1, q)^{1/2} = W(\nu_1, \mathbf{c})^{1/2} &= \sqrt{\mathbb{E} \min_{1 \le j \le k} \|X_0 - c_j\|^2} \\
&= \sqrt{\mathbb{E}(\min_{1 \le j \le k} \|X_0 - c_j\|)^2} \\
&\le \sqrt{\mathbb{E}\big(\min_{1 \le j \le k} (\|X_0 - Y_0\| + \|Y_0 - c_j\|)\big)^2} \\
&= \sqrt{\mathbb{E}\big(\|X_0 - Y_0\| + \min_{1 \le j \le k} \|Y_0 - c_j\|\big)^2} \\
&\le \sqrt{\mathbb{E}\|X_0 - Y_0\|^2} + \sqrt{\mathbb{E} \min_{1 \le j \le k} \|Y_0 - c_j\|^2} \\
&\quad \text{(by the Cauchy-Schwarz inequality)} \\
&= \rho_W(\nu_1, \nu_2) + D(\nu_2, q)^{1/2}.
\end{aligned}$$

One shows with similar arguments that $D(\nu_2, q)^{1/2} \le \rho_W(\nu_1, \nu_2) + D(\nu_1, q)^{1/2}$, and the result follows. □

Theorem 14. *One has $D(\mu, q_n^\star) \to D^\star(\mu)$ almost surely, and $\mathbb{E}D(\mu, q_n^\star) \to D^\star(\mu)$.*

Proof. If $q^\star$ is a NN quantizer optimal for μ, then, by Proposition 3,

$$\begin{aligned}
0 &\le D(\mu, q_n^\star)^{1/2} - D^\star(\mu)^{1/2} \\
&= \big[D(\mu, q_n^\star)^{1/2} - D(\mu_n, q_n^\star)^{1/2}\big] + \big[D(\mu_n, q_n^\star)^{1/2} - D(\mu, q^\star)^{1/2}\big] \\
&\le \big[D(\mu, q_n^\star)^{1/2} - D(\mu_n, q_n^\star)^{1/2}\big] + \big[D(\mu_n, q^\star)^{1/2} - D(\mu, q^\star)^{1/2}\big] \\
&\le 2\rho_W(\mu, \mu_n). \qquad (3)
\end{aligned}$$

But $\rho_W(\mu_n, \mu) \to 0$ almost surely, since $\mathbb{P}(\mu_n \Rightarrow \mu) = 1$ (by Varadarajan's theorem) and, almost surely,

$$\int_{\mathbb{R}^d} \|x\|^2 \mu_n(\mathrm{d}x) \to \int_{\mathbb{R}^d} \|x\|^2 \mu(\mathrm{d}x)$$

(by the strong law of large numbers). We conclude that $D(\mu, q_n^\star) \to D^\star(\mu)$ almost surely.

To prove the second assertion, we introduce $\mathscr{M}(\mu, \mu_n)$, the (random) set of probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and μ_n, respectively. By definition,

$$\rho_W^2(\mu, \mu_n) = \inf_{\nu \in \mathscr{M}(\mu, \mu_n)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \nu(\mathrm{d}x, \mathrm{d}y).$$

Let $C > 0$ be an arbitrary constant, and let $\mathscr{A}$ be the subset of $\mathbb{R}^d \times \mathbb{R}^d$ defined by

$$\mathscr{A} = \big\{(x, y) \in \mathbb{R}^d \times \mathbb{R}^d : \max(\|x\|, \|y\|) \le C\big\}.$$

One has, for all $\nu \in \mathscr{M}(\mu, \mu_n)$,

$$\begin{aligned}
&\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|^2 \nu(\mathrm{d}x, \mathrm{d}y) \\
&\quad = \int_{\mathscr{A}} \|x-y\|^2 \nu(\mathrm{d}x, \mathrm{d}y) + \int_{\mathscr{A}^c} \|x-y\|^2 \nu(\mathrm{d}x, \mathrm{d}y) \\
&\quad \leq \int_{\mathscr{A}} \|x-y\|^2 \nu(\mathrm{d}x, \mathrm{d}y) + 2\int_{\mathscr{A}^c} \|x\|^2 \nu(\mathrm{d}x, \mathrm{d}y) + 2\int_{\mathscr{A}^c} \|y\|^2 \nu(\mathrm{d}x, \mathrm{d}y) \\
&\qquad (\text{since } \|x-y\|^2 \leq 2\|x\|^2 + 2\|y\|^2) \\
&\quad \leq \int_{\mathscr{A}} \|x-y\|^2 \nu(\mathrm{d}x, \mathrm{d}y) + 2\int_{\mathbb{R}^d} \|x\|^2 \mathbb{1}_{[\|x\|>C]} \mu(\mathrm{d}x) + 2\int_{\mathbb{R}^d} \|x\|^2 \mathbb{1}_{[\|x\|\leq C, \|y\|>C]} \nu(\mathrm{d}x, \mathrm{d}y) \\
&\qquad + 2\int_{\mathbb{R}^d} \|y\|^2 \mathbb{1}_{[\|y\|>C]} \mu_n(\mathrm{d}y) + 2\int_{\mathbb{R}^d} \|y\|^2 \mathbb{1}_{[\|x\|>C, \|y\|\leq C]} \nu(\mathrm{d}x, \mathrm{d}y) \\
&\quad \leq \int_{\mathscr{A}} \|x-y\|^2 \nu(\mathrm{d}x, \mathrm{d}y) + 2\int_{\mathbb{R}^d} \|x\|^2 \mathbb{1}_{[\|x\|>C]} \mu(\mathrm{d}x) + 2C^2 \mu_n(\|y\| > C) \\
&\qquad + 2\int_{\mathbb{R}^d} \|y\|^2 \mathbb{1}_{[\|y\|>C]} \mu_n(\mathrm{d}y) + 2C^2 \mu(\|x\| > C).
\end{aligned}$$

Therefore, by Markov's inequality,

$$\begin{aligned}
\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|^2 \nu(\mathrm{d}x, \mathrm{d}y) \leq \int_{\mathscr{A}} \|x-y\|^2 \nu(\mathrm{d}x, \mathrm{d}y) \\
+ 2\int_{\mathbb{R}^d} \|x\|^2 \mathbb{1}_{[\|x\|>C]} \mu(\mathrm{d}x) + 2\int_{\mathbb{R}^d} \|y\|^2 \mathbb{1}_{[\|y\|>C]} \mu_n(\mathrm{d}y) \\
+ 2\int_{\mathbb{R}^d} \|y\|^2 \mathbb{1}_{[\|y\|>C]} \mu_n(\mathrm{d}y) + 2\int_{\mathbb{R}^d} \|x\|^2 \mathbb{1}_{[\|x\|>C]} \mu(\mathrm{d}x).
\end{aligned}$$

Taking the infimum over $\mathscr{M}(\mu, \mu_n)$ on the right-hand side, and then expectation on both sides, we conclude that

$$\mathbb{E}\rho_W^2(\mu, \mu_n) \leq \mathbb{E} \inf_{\nu \in \mathscr{M}(\mu, \mu_n)} \int_{\mathscr{A}} \|x-y\|^2 \nu(\mathrm{d}x, \mathrm{d}y) + 8\int_{\mathbb{R}^d} \|x\|^2 \mathbb{1}_{[\|x\|>C]} \mu(\mathrm{d}x).$$

For fixed $C > 0$, the first term on the right-hand side tends to zero as n tends to infinity by the first statement and the Lebesgue dominated convergence theorem. Since $\int_{\mathbb{R}^d} \|x\|^2 \mu(\mathrm{d}x) < \infty$, the second term can be made arbitrarily small by taking C sufficiently large. Putting all the pieces together, we see that $\mathbb{E}\rho_W^2(\mu, \mu_n)$ tends to zero, and the result easily follows from inequality (3). □

Theorem 15. *If $\|X\| \leq R$ with probability one, then*

$$\mathbb{E}D(\mu, q_n^\star) - D^\star(\mu) \leq \frac{12kR^2}{\sqrt{n}}.$$

Proof. Let us start with some preliminary remarks.

1. Let $\sigma_1, \dots, \sigma_n$ be i.i.d. Rademacher random variables, independent of $X_1, \dots, X_n$, and let $\mathscr{F}$ be a bounded collection of real-valued functions on $\mathbb{R}^d$. Then, by the contraction principle,
$$\mathbb{E} \sup_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i |f(X_i)| \leq \mathbb{E} \sup_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(X_i).$$
(Note that this definition of a Rademacher average does not involve absolute values. This allows us to save a factor of 2 in the contraction principle.)

2. If $\|X\| \leq R$ with probability one, then the optimal codevectors are in $B_R \stackrel{\text{def}}{=} B(0,R)$. To see this, just note that if $\|c\| > R$ and p is the projection onto B_R, then, for all $x \in B_R$,
$$\begin{aligned} \|x-c\|^2 &= \|x-p(c)\|^2 + \|p(c)-c\|^2 - 2\langle x-p(c), c-p(c)\rangle \\ &\geq \|x-p(c)\|^2. \end{aligned}$$
Thus, the distortion is smaller for codevectors in B_R.

3. If $X \stackrel{\mathscr{D}}{=} \mu$, then
$$W(\mu, \mathbf{c}) = \mathbb{E} \min_{1 \leq j \leq k} \|X - c_j\|^2 = \mathbb{E}\|X\|^2 + \mathbb{E} \min_{1 \leq j \leq k} \big(-2\langle X, c_j \rangle + \|c_j\|^2 \big).$$

The last two remarks show that minimizing $W(\mu, \cdot)$ over $\mathbb{R}^{dk}$ is identical to minimizing $\bar{W}(\mu, \cdot)$ over B_R^k, where
$$\bar{W}(\mu, \mathbf{c}) = \mathbb{E} \min_{1 \leq j \leq k} f_{c_j}(X), \quad f_c(x) = -2\langle x, c \rangle + \|c\|^2.$$

The same principle holds with μ_n in place of μ.

We are now ready to prove the theorem. Observe that
$$\begin{aligned} D(\mu, q_n^\star) - D^\star(\mu) &= W(\mu, \mathbf{c}_n^\star) - \inf_{\mathbf{c} \in B_R^k} W(\mu, \mathbf{c}) \\ &= \bar{W}(\mu, \mathbf{c}_n^\star) - \inf_{\mathbf{c} \in B_R^k} \bar{W}(\mu, \mathbf{c}) \\ &= \big[\bar{W}(\mu, \mathbf{c}_n^\star) - \bar{W}(\mu_n, \mathbf{c}_n^\star)\big] + \big[\inf_{\mathbf{c} \in B_R^k} \bar{W}(\mu_n, \mathbf{c}) - \inf_{\mathbf{c} \in B_R^k} \bar{W}(\mu, \mathbf{c})\big] \\ &\leq \sup_{\mathbf{c} \in B_R^k} (\bar{W}(\mu, \mathbf{c}) - \bar{W}(\mu_n, \mathbf{c})) + \sup_{\mathbf{c} \in B_R^k} (\bar{W}(\mu_n, \mathbf{c}) - \bar{W}(\mu, \mathbf{c})). \end{aligned}$$

We are thus interested in upper bounds for the maximal deviation
$$\mathbb{E} \sup_{\mathbf{c} \in B_R^k} (\bar{W}(\mu_n, \mathbf{c}) - \bar{W}(\mu, \mathbf{c})),$$

and note that the other term can be similarly bounded. Let $X_1', \ldots, X_n'$ be a ghost sample, independent of $X_1, \ldots, X_n$ and $\sigma_1, \ldots, \sigma_n$. Then

$$\begin{aligned}
&\mathbb{E} \sup_{\mathbf{c} \in B_R^k} (\bar{W}(\mu_n, \mathbf{c}) - \bar{W}(\mu, \mathbf{c})) \\
&\quad = \mathbb{E} \sup_{\mathbf{c} \in B_R^k} \frac{1}{n} \sum_{i=1}^n \big(\min_{1 \le j \le k} f_{c_j}(X_i) - \mathbb{E} \min_{1 \le j \le k} f_{c_j}(X) \big) \\
&\quad = \mathbb{E} \sup_{\mathbf{c} \in B_R^k} \frac{1}{n} \mathbb{E} \Big[\sum_{i=1}^n \big(\min_{1 \le j \le k} f_{c_j}(X_i) - \min_{1 \le j \le k} f_{c_j}(X_i') \big) \,\Big|\, X_1, \ldots, X_n \Big].
\end{aligned}$$

Thus, upon noting that $\sup \mathbb{E}(\cdot) \le \mathbb{E} \sup(\cdot)$,

$$\begin{aligned}
\mathbb{E} \sup_{\mathbf{c} \in B_R^k} (\bar{W}(\mu_n, \mathbf{c}) - \bar{W}(\mu, \mathbf{c})) &\le \mathbb{E} \sup_{\mathbf{c} \in B_R^k} \frac{1}{n} \sum_{i=1}^n \big(\min_{1 \le j \le k} f_{c_j}(X_i) - \min_{1 \le j \le k} f_{c_j}(X_i') \big) \\
&\le 2 \mathbb{E} \sup_{\mathbf{c} \in B_R^k} \frac{1}{n} \sum_{i=1}^n \sigma_i \min_{1 \le j \le k} f_{c_j}(X_i).
\end{aligned}$$

The proof proceeds now by induction on k, using the contraction principle. Let

$$S_k = \mathbb{E} \sup_{(c_1, \ldots, c_k) \in B_R^k} \frac{1}{n} \sum_{i=1}^n \sigma_i \min_{1 \le j \le k} f_{c_j}(X_i).$$

Case $k = 1$**.** Since $\|X\| \le R$,

$$\begin{aligned}
S_1 &= \mathbb{E} \sup_{c \in B_R} \frac{1}{n} \sum_{i=1}^n \sigma_i \big(-2\langle X_i, c \rangle + \|c\|^2 \big) \\
&\le 2 \mathbb{E} \sup_{c \in B_R} \frac{1}{n} \sum_{i=1}^n \sigma_i \langle X_i, c \rangle + \mathbb{E} \sup_{c \in B_R} \frac{\|c\|^2}{n} \sum_{i=1}^n \sigma_i \\
&\le 2 \mathbb{E} \sup_{c \in B_R} \frac{1}{n} \sum_{i=1}^n \sigma_i \langle X_i, c \rangle + \frac{R^2}{n} \mathbb{E} \Big| \sum_{i=1}^n \sigma_i \Big| \\
&\le 2 \mathbb{E} \sup_{c \in B_R} \frac{1}{n} \sum_{i=1}^n \sigma_i \langle X_i, c \rangle + \frac{R^2}{\sqrt{n}} \\
&\qquad \text{(by the Cauchy-Schwarz inequality).}
\end{aligned}$$

Thus,

$$\begin{aligned}
S_1 &\le 2\mathbb{E}\sup_{c\in B_R}\frac{1}{n}\Big\langle\sum_{i=1}^n \sigma_i X_i, c\Big\rangle + \frac{R^2}{\sqrt{n}} \\
&= \frac{2R}{n}\mathbb{E}\Big\|\sum_{i=1}^n \sigma_i X_i\Big\| + \frac{R^2}{\sqrt{n}} \\
&\le 2R\sqrt{\frac{\mathbb{E}\|X\|^2}{n}} + \frac{R^2}{\sqrt{n}} \\
&\quad \text{(by the Cauchy-Schwarz inequality)} \\
&\le \frac{3R^2}{\sqrt{n}}.
\end{aligned}$$

Case $k = 2$**.** Using the equality $\min(a,b) = \frac{a+b}{2} - \frac{|a-b|}{2}$, we may write

$$\begin{aligned}
S_2 &= \mathbb{E}\sup_{(c_1,c_2)\in B_R^2}\frac{1}{2n}\sum_{i=1}^n \sigma_i\big(f_{c_1}(X_i) + f_{c_2}(X_i) - |f_{c_1}(X_i) - f_{c_2}(X_i)|\big) \\
&\le S_1 + \mathbb{E}\sup_{(c_1,c_2)\in B_R^2}\frac{1}{2n}\sum_{i=1}^n \sigma_i|f_{c_1}(X_i) - f_{c_2}(X_i)|.
\end{aligned}$$

Applying the contraction principle, we obtain

$$S_2 \le S_1 + \mathbb{E}\sup_{(c_1,c_2)\in B_R^2}\frac{1}{2n}\sum_{i=1}^n \sigma_i(f_{c_1}(X_i) - f_{c_2}(X_i)) \le 2S_1.$$

Case $k = 3$**.** Since $S_2 \le 2S_1$,

$$S_3 \le \frac{S_1+S_2}{2} + \frac{S_1+S_2}{2} \le 3S_1.$$

Repeating this process, we find

$$S_k \le kS_1 \le \frac{3kR^2}{\sqrt{n}}.$$

Finally,

$$\mathbb{E}D(\mu, q_n^\star) - D^\star(\mu) \le 4S_k \le \frac{12kR^2}{\sqrt{n}},$$

and the proof is complete. □