NOTES ON THE VAPNIK–CHERVONENKIS THEOREM: BACKGROUND AND PROOF

ROLAND WALKER

1. Introduction

Vladimir Vapnik and Alexey Chervonenkis proved their eponymous theorem in 1968. The original Russian proof was published in 1971 and then translated to English by B. Seckler later that year. The English translation was most recently reprinted in 2015 [4].

These notes, which provide a relatively self-contained proof of the VC Theorem, assume the reader has some comfort with the basics of real analysis (e.g., Chapters 1 and 2 of [2]) but little or no background in probability theory. In addition to the original paper, we used Chapter 6 and Appendix B of [3] as a reference for the proof of the VC theorem and Appendix A of [1] as a reference for the proof of Chernoff's theorem.

2. Products of σ-algebras

Let I be a nonempty set, and let $(X_i, \mathcal{A}_i)_{i\in I}$ be a family of measurable spaces (i.e., each X_i is a nonempty set and each $\mathcal{A}_i$ is a σ-algebra on X_i).

Definition 2.1. The product $\bigotimes_{i\in I} \mathcal{A}_i$ is the σ-algebra on $\prod_{i\in I} X_i$ given by

$$\bigotimes_{i\in I} \mathcal{A}_i = \sigma\left(\left\{\pi_i^{-1}(A_i) : i \in I,\ A_i \in \mathcal{A}_i\right\}\right).$$

Moreover, if $I = \{0, \dots, n-1\}$ for some $n \geq 2$, we often write $\mathcal{A}_0 \otimes \cdots \otimes \mathcal{A}_{n-1}$ for $\bigotimes_{i\in I} \mathcal{A}_i$ just as we often write $X_0 \times \cdots \times X_{n-1}$ for $\prod_{i\in I} X_i$.

Lemma 2.2. *If I is countable, then*

$$\bigotimes_{i\in I} \mathcal{A}_i = \sigma\left(\left\{\prod_{i\in I} A_i : A_i \in \mathcal{A}_i\right\}\right).$$

Proof. A σ-algebra is closed under taking countable intersections. □

Lemma 2.3. *If $(\mathcal{E}_i)_{i\in I}$ is such that each $\mathcal{A}_i = \sigma(\mathcal{E}_i)$, then*

$$\bigotimes_{i\in I} \mathcal{A}_i = \sigma\left(\left\{\pi_i^{-1}(E_i) : i \in I,\ E_i \in \mathcal{E}_i\right\}\right).$$

If, in addition, I is countable, then

$$\bigotimes_{i\in I} \mathcal{A}_i = \sigma\left(\left\{\prod_{i\in I} E_i : E_i \in \mathcal{E}_i\right\}\right).$$

Lemma 2.4. *If $I = J \sqcup K$, with both J and K nonempty, then*

$$\bigotimes_{i\in I} \mathcal{A}_i = \left(\bigotimes_{j\in J} \mathcal{A}_j\right) \otimes \left(\bigotimes_{k\in K} \mathcal{A}_k\right). \tag{2.1}$$

Proof. By Lemma 2.3, the right-hand side of (2.1) is the σ-algebra generated by sets of the form $\pi_j^{-1}(A_j) \cap \pi_k^{-1}(A_k)$ where $j \in J$, $k \in K$, $A_j \in \mathcal{A}_j$, and $A_k \in \mathcal{A}_k$. □

Corollary 2.5. *The operator $\otimes$ is associative.*

3. Product Measures

Let $n \geq 2$, and let $(X_i, \mathcal{A}_i, \mu_i)_{i<n}$ be a family of measure spaces; i.e., each $\mu_i : \mathcal{A}_i \to [0, \infty]$ is a measure (see [2, p. 24]) on the measurable space $(X_i, \mathcal{A}_i)$. Let $\mathcal{R}$ denote the collection of rectangular sets in $\mathcal{A}_0 \otimes \cdots \otimes \mathcal{A}_{n-1}$; i.e.,

$$\mathcal{R} = \{A_0 \times \cdots \times A_{n-1} : A_i \in \mathcal{A}_i\}.$$

It follows that $\mathcal{R}$ is an elementary family (see [2, p. 23]), so the set

$$\mathcal{F} = \left\{ \bigsqcup_{j<m} R_j : 1 \leq m < \omega,\ R_j \in \mathcal{R} \right\}.$$

consisting of all finite disjoint unions of rectangles is an algebra [2, Proposition 1.7].

Let $\rho : \mathcal{R} \to [0, \infty]$ be defined by

$$A_0 \times \cdots \times A_{n-1} \mapsto \mu_0(A_0) \cdots \mu_{n-1}(A_{n-1}).$$

Claim 3.1. *Suppose $(S_j)_{j<\omega} \subseteq \mathcal{R}$ is a family of pairwise disjoint rectangles and $R = \bigsqcup_{j<\omega} S_j$. If $R \in \mathcal{R}$, then $\rho(R) = \sum_{j<\omega} \rho(S_j)$.*

Proof. Suppose $R = A_0 \times \cdots \times A_{n-1}$ and each $S_j = B_0^j \times \cdots \times B_{n-1}^j$ with each A_i and B_i^j in $\mathcal{A}_i$. Since

$$\begin{aligned} 1_{A_0}(x_0) \cdots 1_{A_{n-1}}(x_{n-1}) &= 1_{A_0 \times \cdots \times A_{n-1}}(x_0, \ldots, x_{n-1}) \\ &= \sum_{j<\omega} 1_{B_0^j \times \cdots \times B_{n-1}^j}(x_0, \ldots, x_{n-1}) \\ &= \sum_{j<\omega} 1_{B_0^j}(x_0) \cdots 1_{B_{n-1}^j}(x_{n-1}) \end{aligned}$$

for all $(x_0, \ldots, x_{n-1}) \in X_0 \times \cdots \times X_{n-1}$, [2, Theorem 2.15] asserts that

$$\begin{aligned} &\mu_0(A_0) \cdots \mu_{n-1}(A_{n-1}) \\ &\quad = \int_{X_{n-1}} \cdots \int_{X_0} 1_{A_0}(x_0) \cdots 1_{A_{n-1}}(x_{n-1})\ d\mu_0(x_0) \cdots d\mu_{n-1}(x_{n-1}) \\ &\quad = \sum_{j<\omega} \int_{X_{n-1}} \cdots \int_{X_0} 1_{B_0^j}(x_0) \cdots 1_{B_{n-1}^j}(x_{n-1})\ d\mu_0(x_0) \cdots d\mu_{n-1}(x_{n-1}) \\ &\quad = \sum_{j<\omega} \mu_0(B_0^j) \cdots \mu_{n-1}(B_{n-1}^j). \end{aligned}$$

□

Let $\nu : \mathcal{F} \to [0, \infty]$ be defined by

$$\nu\left(\bigsqcup_{j<m} R_j\right) = \sum_{j<m} \rho(R_j).$$

In order to show that ν is well-defined, suppose that $\bigsqcup_{j<m} R_j$ and $\bigsqcup_{k<m} S_k$ describe the same set in $\mathcal{F}$. For each $j < m$, suppose $R_j = A_0^j \times \cdots \times A_{n-1}^j$ and $S_k = B_0^k \times \cdots \times B_{n-1}^k$ with each A_i^j and B_i^k in $\mathcal{A}_i$. By Claim 3.1, we have

$$\begin{aligned}
\nu\left(\bigsqcup_{j<m} R_j\right) &= \sum_{j<m} \mu_0\left(A_0^j\right) \cdots \mu_{n-1}\left(A_{n-1}^j\right) \\
&= \sum_{j,k<m} \mu_0\left(A_0^j \cap B_0^k\right) \cdots \mu_{n-1}\left(A_{n-1}^j \cap B_{n-1}^k\right) \\
&= \sum_{k<m} \mu_0\left(B_0^k\right) \cdots \mu_{n-1}\left(B_{n-1}^k\right) \\
&= \nu\left(\bigsqcup_{k<m} S_k\right).
\end{aligned}$$

Next, we show that ν is a premeasure on $\mathcal{F}$ (see [2, p.30]). Let $\bigsqcup_{j<m} R_j \in \mathcal{F}$, and let $\left(\bigsqcup_{k<m_\ell} S_k^\ell\right)_{\ell<\omega} \subseteq \mathcal{F}$ be pairwise disjoint. Suppose $\bigsqcup_{j<m} R_j = \bigsqcup_{\ell<\omega}\left(\bigsqcup_{k<m_\ell} S_k^\ell\right)$. By Claim 3.1, it follows that

$$\begin{aligned}
\nu\left(\bigsqcup_{j<m} R_j\right) &= \sum_{j<m} \rho(R_j) \\
&= \sum_{j<m} \sum_{\ell<\omega} \sum_{k<m_\ell} \rho\left(R_j \cap S_k^\ell\right) \\
&= \sum_{\ell<\omega} \sum_{k<m_\ell} \sum_{j<m} \rho\left(R_j \cap S_k^\ell\right) \\
&= \sum_{\ell<\omega} \sum_{k<m_\ell} \rho\left(S_k^\ell\right) \\
&= \sum_{\ell<\omega} \nu\left(\bigsqcup_{k<m_\ell} S_k^\ell\right).
\end{aligned}$$

Let ν^* be the outer measure associated with ν; i.e.,

$$\nu^* : \mathcal{P}(X_0 \times \cdots \times X_{n-1}) \to [0, \infty]$$

where

$$\nu^*(A) = \inf\left\{\sum_{j<\omega} \nu(F_j) : F_j \in \mathcal{F},\ A \subseteq \bigcup_{j<\omega} F_j\right\}.$$

Definition 3.2. The product measure $\mu_0 \times \cdots \times \mu_{n-1}$ is the restriction of ν^* to $\mathcal{A}_0 \otimes \cdots \otimes \mathcal{A}_{n-1}$.

By [2, Proposition 1.13], this product is indeed a measure which extends ρ. If, in addition, each μ_i is σ-finite, then [2, Proposition 1.14] implies that the product is the unique measure extending ρ to $\mathcal{A}_0 \otimes \cdots \otimes \mathcal{A}_{n-1}$.

Lemma 3.3. *If each μ_i is σ-finite, then the product $\mu_0 \times \cdots \times \mu_{n-1}$ is associative.*

Proof. Suppose $I \sqcup J = \{0, \ldots, n-1\}$ where both I and J are nonempty. Let $\mu_I = \prod_{i \in I} \mu_i$ and $\mu_J = \prod_{j \in J} \mu_j$. It follows that $(\mu_I \times \mu_J)\restriction_{\mathcal{R}} = \rho$. □

4. Pushforwards

Suppose $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ are measurable spaces and $f : X \to Y$ is an $(\mathcal{A}, \mathcal{B})$-measurable function.

Definition 4.1. If $\mu : \mathcal{A} \to [0, \infty]$ is a measure, then we call $\mu \circ f^{-1} : \mathcal{B} \to [0, \infty]$ its *pushforward by f*.

Claim 4.2. *The pushforward $\mu \circ f^{-1}$ is a measure.*

Proof. Notice that $\mu \circ f^{-1}(\varnothing) = \mu(\varnothing) = 0$. Suppose $(B_i : i < \omega) \subseteq \mathcal{B}$ is pairwise disjoint. It follows that $(f^{-1}(B_i) : i < \omega) \subseteq \mathcal{A}$ is also pairwise disjoint, so

$$\mu \circ f^{-1}\left(\bigcup B_i\right) = \mu\left(\bigcup f^{-1}(B_i)\right) = \sum \mu \circ f^{-1}(B_i).$$

□

5. Probability Spaces

Definition 5.1. A *probability space* is a measure space $(\Omega, \mathcal{A}, P)$ with $P(\Omega) = 1$.

Definition 5.2. If $(\Omega, \mathcal{A}, P)$ is a probability space, then the P-measurable sets (i.e., the elements of $\mathcal{A}$) are called *events*.

6. Random Elements and Variables

Let $(\Omega, \mathcal{A}, P)$ be a probability space.

Definition 6.1. A *random element* of a measurable space $(\Psi, \mathcal{B})$ is an $(\mathcal{A}, \mathcal{B})$-measurable function $X : \Omega \to \Psi$. Furthermore, if $\Psi = \mathbb{R}$ and $\mathcal{B} = \mathcal{B}(\mathbb{R})$, then we call X a *random variable*.

When describing events using preimages of random elements, we often use

$$\begin{aligned} &[X \in B] \text{ for } \{\omega \in \Omega : X(\omega) \in B\}, \\ &[X > r] \text{ for } \{\omega \in \Omega : X(\omega) > r\}, \\ &\text{etc.} \end{aligned}$$

This abbreviation practice is common in the literature of probability theory. As an aid to the reader, we set off such abbreviations with square brackets rather than braces.

Definition 6.2. We say that a collection of random elements $X_0, \ldots, X_{n-1}$ of measurable spaces $(\Psi_0, \mathcal{B}_0), \ldots, (\Psi_{n-1}, \mathcal{B}_{n-1})$, respectively, are *mutually independent* iff: for all $(B_0, \ldots, B_{n-1}) \in \mathcal{B}_0 \times \cdots \times \mathcal{B}_{n-1}$, we have

$$P[X_0 \in B_0, \ldots, X_{n-1} \in B_{n-1}] = P[X_0 \in B_0] \cdots P[X_{n-1} \in B_{n-1}].$$

Definition 6.3. If X is a random element of $(\Psi, \mathcal{B})$, then the *probability distribution* of X is the pushforward $P \circ X^{-1} : \mathcal{B} \to [0, 1]$.

Lemma 6.4. *A collection of random elements $X_0, \dots, X_{n-1}$ of measurable spaces $(\Psi_0, \mathcal{B}_0), \dots, (\Psi_{n-1}, \mathcal{B}_{n-1})$, respectively, is mutually independent if and only if the probability distribution of the random element $\bar{X}$ of*

$$(\Psi_0 \times \cdots \times \Psi_{n-1},\ \mathcal{B}_0 \otimes \cdots \otimes \mathcal{B}_{n-1})$$

given by

$$\bar{X}(\omega) = (X_0(\omega), \dots, X_{n-1}(\omega))$$

is the product $\mu_0 \times \cdots \times \mu_{n-1}$ where each $\mu_i = P \circ X_i^{-1}$ is the probability distribution of X_i.

Proof. Let $(B_0, \dots, B_{n-1}) \in \mathcal{B}_0 \otimes \cdots \otimes \mathcal{B}_{n-1}$. Since

$$\bar{X}^{-1}(B_0 \times \cdots \times B_{n-1}) = X_0^{-1}(B_0) \cap \cdots \cap X_{n-1}^{-1}(B_{n-1}) \in \mathcal{A}$$

and since preimages preserve complements and arbitrary unions, it follows that $\bar{X}$ is $(\mathcal{A},\ \mathcal{B}_0 \otimes \cdots \otimes \mathcal{B}_{n-1})$-measurable.

($\Rightarrow$) Notice that

$$\begin{aligned} P \circ \bar{X}^{-1}(B_0 \times \cdots \times B_{n-1}) &= P[X_0 \in B_0, \dots, X_{n-1} \in B_{n-1}] \\ &= P[X_0 \in B_0] \cdots P[X_{n-1} \in B_{n-1}] \\ &= \mu_0(B_0) \cdots \mu_{n-1}(B_{n-1}). \end{aligned}$$

Since each μ_i is finite, the product $\mu_0 \times \cdots \times \mu_{n-1}$ is the unique measure on $\mathcal{B}_0 \otimes \cdots \otimes \mathcal{B}_{n-1}$ with this property for all rectangles.

($\Leftarrow$) Notice that

$$\begin{aligned} P[X_0 \in B_0, \dots, X_{n-1} \in B_{n-1}] &= P \circ \bar{X}^{-1}(B_0 \times \cdots \times B_{n-1}) \\ &= \mu_0 \times \cdots \times \mu_{n-1}(B_0 \times \cdots \times B_{n-1}) \\ &= \mu_0(B_0) \cdots \mu_{n-1}(B_{n-1}) \\ &= P[x_0 \in B_0] \cdots P[x_{n-1} \in B_{n-1}]. \end{aligned}$$

□

Definition 6.5. If X is a random variable, its *expected value* is given by

$$E(X) = \int_\Omega X\ dP$$

provided the integral is well-defined (i.e., either $\int_\Omega X^+\ dP$ or $\int_\Omega X^-\ dP$ is finite).

For the remainder, we tacitly assume all random variables have well-defined expectations.

Definition 6.6. Given a random variable $X : \Omega \to [0, \infty)$, for each $n < \omega$, let ϕ_n^X denote the simple function

$$\sum_{i<n^2} \frac{i}{n} 1_{X^{-1}(B_i)},$$

where each

$$B_i = \left[\frac{i}{n}, \frac{i+1}{n}\right).$$

Lemma 6.7. *If X and Y are mutually independent random variables, then for all $m, n < \omega$, we have*

$$E\left(\phi_m^X \phi_n^Y\right) = E\left(\phi_m^X\right) E\left(\phi_n^Y\right).$$

Proof. The result follows since for all $r, s \in \mathbb{R}$ and all $A, B \in \mathcal{B}(\mathbb{R})$, we have

$$\begin{aligned} E\left(r1_{X^{-1}(A)} \cdot s1_{Y^{-1}(B)}\right) &= rsE\left(1_{X^{-1}(A)\cap Y^{-1}(B)}\right) \\ &= rsP\left(X^{-1}(A) \cap Y^{-1}(B)\right) \\ &= rP\left(X^{-1}(A)\right) \cdot sP\left(Y^{-1}(B)\right) \\ &= E\left(r1_{X^{-1}(A)}\right) \cdot E\left(s1_{Y^{-1}(B)}\right). \end{aligned}$$

□

Lemma 6.8. *If $X_0, \ldots, X_{n-1}$ are mutually independent random variables, then $E(X_0 \cdots X_{n-1}) = E(X_0) \cdots E(X_{n-1})$.*

Proof. We proceed by induction on n. Suppose the lemma holds for $n \geq 1$. Given mutually independent random variables $X_0, \ldots, X_{n-1}, Y$, let $X = X_0 \cdots X_{n-1}$. Lemma 6.4 implies that X and Y are mutually independent.

Suppose that X and Y are non-negative. The Monotone Convergence Theorem [2, Theorem 2.14] asserts that

$$E\left(\phi_i^X\right) \to E(X), \quad E\left(\phi_i^Y\right) \to E(Y), \quad \text{and} \quad E\left(\phi_i^X \phi_i^Y\right) \to E(XY),$$

so by Lemma 6.7, we have $E(XY) = E(X)E(Y)$.

The general case follows since

$$\begin{aligned} E(XY) &= E((X^+ - X^-)(Y^+ - Y^-)) \\ &= E(X^+Y^+) - E(X^+Y^-) - E(X^-Y^+) + E(X^-Y^-) \\ &= E(X^+)E(Y^+) - E(X^+)E(Y^-) - E(X^-)E(Y^+) + E(X^-)E(Y^-) \\ &= (E(X^+) - E(X^-))(E(Y^+) - E(Y^-)) \\ &= E(X)E(Y). \end{aligned}$$

□

Definition 6.9. If X is a random variable, its *variance* is given by

$$V(X) = E((X - E(X))^2).$$

Lemma 6.10. *If $X_0, \ldots, X_{n-1}$ are mutually independent random variables, then*

$$V(X_0 + \cdots + X_{n-1}) = V(X_0) + \cdots + V(X_{n-1}).$$

Proof. We proceed by induction on n. Suppose the lemma holds for $n \geq 1$. Given mutually independent random variables $X_0, \ldots, X_{n-1}, Y$, let $X = X_0 \cdots X_{n-1}$. Lemma 6.4 implies that X and Y are mutually independent, so we have

$$\begin{aligned} V(X+Y) &= E\left((X + Y - E(X+Y))^2\right) \\ &= E\left(X^2 + 2XY + Y^2 - 2(X+Y)E(X+Y) + E(X+Y)^2\right) \\ &= E(X^2 + 2XY + Y^2 - 2XE(X) - 2XE(Y) - 2YE(X) \\ &\qquad - 2YE(Y) + E(X)^2 + 2E(X)E(Y) + E(Y)^2) \\ &= E(X^2 - 2XE(X) + E(X)^2) + E(Y^2 - 2YE(Y) + E(Y)^2) \\ &\qquad + 2E(XY - XE(Y) - YE(X) + E(X)E(Y)) \\ &= V(X) + V(Y) + 2(E(X)E(Y) - E(X)E(Y) - E(Y)E(X) \\ &\qquad + E(X)E(Y)) \\ &= V(X) + V(Y). \end{aligned}$$

□

7. Average Measures

Definition 7.1. Given a measurable space $(X, \mathcal{A})$ and $b_0, \ldots, b_{n-1} \in X$, let $\mathrm{Av}_{\bar{b}}$ denote the *average measure* given by

$$\mathrm{Av}_{\bar{b}}(A) = \frac{1}{n} \sum_{i<n} 1_{\{b_i\}}(A)$$

for all $A \in \mathcal{A}$.

8. Chernoff's Bound

Let $(\Omega, \mathcal{A}, P)$ be a probability space, and let $X : \Omega \to \mathbb{R}^{\geq 0}$ be a random variable.

Lemma 8.1. *Given $r \geq 0$ and $s > 0$, if $P[X > r] \geq s$, then $E(X) > rs$.*

Proof. Since

$$[X > r] = \bigcup_{\delta > 0} [X > r + \delta],$$

there are $\delta, \varepsilon > 0$ such that $P[X > r + \delta] > \varepsilon$, so

$$E(X) \geq rP[X > r] + \delta P[X > r + \delta] \geq rs + \delta\varepsilon.$$

□

Lemma 8.2. *(Markov's Inequality) For $r > 0$, we have*

$$P[X > rE(X)] < \frac{1}{r}.$$

Proof. Assume that $P[X > rE(X)] \geq 1/r$ for some $r > 0$. The previous lemma implies that $E(X) > rE(X) \cdot 1/r = E(X)$, a contradiction. □

Lemma 8.3. *If $x > 0$, then $\cosh x \leq e^{x^2/2}$.*

Proof. Let

$$f(x) = \cosh x = \frac{e^x + e^{-x}}{2}.$$

It follows that

$$f'(x) = \sinh x = \frac{e^x - e^{-x}}{2}.$$

Furthermore, we have

$$f^{(k)}(x) = \begin{cases} \cosh x & \text{if } k \text{ even}, \\ \sinh x & \text{if } k \text{ odd}, \end{cases}$$

so Taylor's theorem asserts that

$$f(x) = \sum_{k=0}^{\infty} \frac{x^k}{k!} f^{(k)}(0) = \sum_{k=0}^{\infty} \frac{x^k}{k!} \begin{cases} \frac{1}{2} & \text{if } k \text{ even}, \\ 0 & \text{if } k \text{ odd} \end{cases}$$

since for all $y \in (0, x)$, the remainder vanishes, i.e.,

$$R_f(k) = \frac{x^k}{k!} f^{(k)}(y) \leq \frac{x^k}{k!} f^{(k)}(x) \to 0.$$

Let $g(x) = e^{x^2/2}$. By induction, $g^{(k)}(x) = p_k(x)g(x)$ for some

$$p_k(x) = a_{n_k}x^{n_k} + \cdots + a_0$$

with nonnegative integer coefficients such that

- $a_0 > 0$ if k is even,
- $a_1 > 0$ if k is odd, and
- $a_i = 0$ if $i \not\equiv k \pmod 2$.

It follows that

$$g^{(k)}(0) \geq \begin{cases} 1 & \text{if } k \text{ even,} \\ 0 & \text{if } k \text{ odd.} \end{cases}$$

For all $n \geq 1$, Taylor's Theorem asserts that

$$g(x) = \sum_{k=0}^{n-1} \frac{x^k}{k!} g^{(k)}(0) + R_g(n)$$

with remainder

$$R_g(n) = \frac{x^n}{n!} f^{(n)}(y)$$

for some $y \in (0, x)$. Since each remainder is positive, we have shown that $f(x) \leq g(x)$ for all $x > 0$. □

Theorem 8.4. *(Chernoff's Bound) Given $\varepsilon > 0$, if $\sigma_o, \ldots, \sigma_{n-1}$ are mutually independent random variables, each with probability distribution* $\mathrm{Av}_{-1,1}$*, then*

$$P\left[\sum_{i<n} \sigma_i > \varepsilon\right] < e^{-\varepsilon^2/2n}.$$

Proof. By Lemma 8.3, we have

$$E\left(e^{\delta\sigma_i}\right) = \frac{e^\delta + e^{-\delta}}{2} = \cosh(\delta) \leq e^{\delta^2/2}$$

for each $i < n$, and since expectations multiply (Lemma 6.8), it follows that

$$E\left(e^{\delta\sigma_0 + \cdots + \delta\sigma_{n-1}}\right) \leq e^{n\delta^2/2}.$$

Now we can apply Markov's inequality (Lemma 8.2) to obtain

$$P\left[\sum_{i<n} \sigma_i > \varepsilon\right] = P\left[e^{\delta\sigma_0 + \cdots + \delta\sigma_{n-1}} > e^{\delta\varepsilon}\right] < \frac{E\left(e^{\delta\sigma_0 + \cdots + \delta\sigma_{n-1}}\right)}{e^{\delta\varepsilon}} \leq \frac{e^{n\delta^2/2}}{e^{\delta\varepsilon}}.$$

This bound becomes minimal when $\delta = \varepsilon/n$.

□

9. The Weak Law of Large Numbers

Let $(\Omega, \mathcal{A}, P)$ be a probability space.

Lemma 9.1 (Chevyshev's Inequality)**.** *Given $\varepsilon > 0$, if X is a random variable, then*

$$P(|X - E(X)| \geq \varepsilon) \leq \frac{V(X)}{\varepsilon^2}.$$

Proof. See [3, Proposition B.3]. □

Proposition 9.2 (The Weak Law of Large Numbers)**.** *Given $\varepsilon > 0$, if $A \in \mathcal{A}$, then for all $n \geq 1$, we have*

$$P^n\left(\left\{\bar{b} \in \Omega^n : |\mathrm{Av}_{\bar{b}}(A) - P(A)| \geq \varepsilon\right\}\right) \leq \frac{1}{4n\varepsilon^2}.$$

Proof. See [3, Proposition B.4]. □

10. The Vapnik–Chervonenkis Theorem

Let X be a nonempty set, $\mathcal{A} \subseteq \mathcal{P}(X)$ a σ-algebra, and $\mu : \mathcal{A} \to [0,1]$ a probability measure. Fix $n < \omega$, and let $x_0, \ldots, x_{n-1}$ be mutually independent random elements of $(X, \mathcal{A})$ each with probability distribution μ.

Theorem 10.1 (The Vapnik–Chervonenkis Theorem)**.** *If $\varepsilon > 0$ and $\mathcal{S}$ is a nonempty countable collection of subsets from $\mathcal{A}$, then*

$$\mu^n\left[\sup_{S \in \mathcal{S}} |\mathrm{Av}_{\bar{x}}(S) - \mu(S)| > \varepsilon\right] \leq 8\pi_{\mathcal{S}}(n) e^{-n\varepsilon^2/32}. \tag{10.1}$$

Proof. For each $S \in \mathcal{S}$, the function

$$\bar{x} \mapsto \mathrm{Av}_{\bar{x}}(S) - \mu(S) = \frac{1}{n}\sum_{i<n} 1_S(x_i) - \mu(S)$$

is measurable since $S \in \mathcal{A}$. Furthermore, since $\mathcal{S}$ is countable, the function

$$\bar{x} \mapsto \sup_{S \in \mathcal{S}} |\mathrm{Av}_{\bar{x}}(S) - \mu(S)|$$

is also measurable [2, Proposition 2.7], so the inequality (10.1) is well-defined.

Let $y_0, \ldots, y_{n-1}$ be random elements of $(X, \mathcal{A})$ each with probability distribution μ, and let $\sigma_0, \ldots, \sigma_{n-1}$ be random variables each with probability distribution $\nu = \mathrm{Av}_{-1,1}$. Suppose all the random elements and variables named above are mutually independent.

For an explicit construction, consider the set $\Omega = X^{2n} \times \{-1,1\}^n$. Let each x_i be $\pi_i : \Omega \to X$, each y_i be $\pi_{n+i} : \Omega \to X$, and each σ_i be $\pi_{2n+i} : \Omega \to \mathbb{R}$. Let

$$\mathcal{F} = \bigotimes_{i<2n} \mathcal{A} \otimes \bigotimes_{i<n} \mathcal{P}(\{-1,1\})$$

and $P : \mathcal{F} \to [0,1]$ be the probability measure determined by

$$P\left(\prod_{i<2n} A_i \times \prod_{i<n} B_i\right) = \prod_{i<2n} \mu(A_i) \cdot \prod_{i<n} \frac{|B_i|}{2}$$

for rectangular sets where each $A_i \in \mathcal{A}$ and each $B_i \subseteq \{-1,1\}$. This yields a probability space $(\Omega, \mathcal{F}, P)$ where all the previously named random variables/elements are mutually independent and possess the desired distributions.

For each $i < n$ and $S \in \mathcal{S}$, let

$$f_i(S) = 1_S(x_i) - 1_S(y_i)$$

and

$$g_i(S) = \sigma_i \cdot f_i(S).$$

Notice that

$$P[f_i(S) = 1] = P[x_i \in S,\ y_i \notin S] = \mu(S) \cdot (1 - \mu(S))$$

and

$$\begin{aligned} P[g_i(S) = 1] &= P[\sigma_i = 1,\ x_i \in S,\ y_i \notin S] + P[\sigma_i = -1,\ x_i \notin S,\ y_i \in S] \\ &= \frac{1}{2} \cdot \mu(S) \cdot (1 - \mu(S)) + \frac{1}{2} \cdot (1 - \mu(S)) \cdot \mu(S) \\ &= \mu(S) \cdot (1 - \mu(S)). \end{aligned}$$

Similarly, we have

$$P[f_i(S) = -1] = P[g_i(S) = -1] = \mu(S) \cdot (1 - \mu(S))$$

and

$$P[f_i(S) = 0] = P[g_i(S) = 0] = 1 - 2\mu(S) \cdot (1 - \mu(S)).$$

Notice that for fixed $S \in \mathcal{S}$, if we let each h_i be either f_i or g_i, then the variables $h_0(S), \ldots, h_{n-1}(S)$ are mutually independent. However, it is not true in general that $f_i(S)$ and $g_i(S)$ are mutually independent since both depend on x_i and y_i. Explicitly, we have

$$\begin{aligned} P[f_i(S) = 1,\ g_i(S) = 1] &= P[\sigma_i = 1,\ x_i \in S,\ y_i \notin S] \\ &= \frac{1}{2}\mu(S) \cdot (1 - \mu(S)) \end{aligned}$$

and

$$P[f_i(S) = 1] \cdot P[g_i(S) = 1] = \mu(S)^2 \cdot (1 - \mu(S))^2,$$

so $f_i(S)$ and $g_i(S)$ are mutually independent if and only if $\mu(S) = 0$ or 1.

Consider the map $F : \Omega \to \Omega$ defined by

$$\begin{aligned} &F(a_0, \ldots, a_{n-1}, b_0, \ldots, b_{n-1}, e_0, \ldots, e_{n-1}) \\ &\qquad = (c_0, \ldots, c_{n-1}, d_0, \ldots, d_{n-1}, e_0, \ldots, e_{n-1}) \end{aligned}$$

where each

$$(c_i, d_i) = \begin{cases} (a_i, b_i) & e_i = 1 \\ (b_i, a_i) & e_i = -1. \end{cases}$$

Notice that F is its own inverse and, therefore, a bijection. Given a rectangular set

$$R = \prod_{i<n} A_i \times \prod_{i<n} B_i \times \prod_{i<n} E_i \in \mathcal{F},$$

we have

$$\begin{aligned} P(R) &= \sum_{\bar{e} \in E_0 \times \cdots \times E_{n-1}} P\left[\bigwedge_{i<n} x_i \in A_i,\ \bigwedge_{i<n} y_i \in B_i,\ \bigwedge_{i<n} \sigma_i = e_i\right] \\ &= \sum_{\bar{e} \in E_0 \times \cdots \times E_{n-1}} P\left[\bigwedge_{i<n} (e_i = 1 \to x_i \in A_i \wedge y_i \in B_i),\right. \\ &\qquad\qquad \left. \bigwedge_{i<n} (e_i = -1 \to x_i \in B_i \wedge y_i \in A_i),\ \bigwedge_{i<n} \sigma_i = e_i\right] \\ &= P(F(R)). \end{aligned}$$

Since P is a product measure, which is the restriction of an outer measure (see Definition 3.2), it follows that F is measure preserving. Furthermore, since an elementary event $(\bar{a}, \bar{b}, \bar{e})$ is contained in

$$\left[\sup_{S\in\mathcal{S}}\left|\frac{1}{n}\sum_{i<n} f_i(S)\right| > \frac{\varepsilon}{2}\right] \tag{10.2}$$

if and only if $F(\bar{a}, \bar{b}, \bar{e})$ is contained in

$$\left[\sup_{S\in\mathcal{S}}\left|\frac{1}{n}\sum_{i<n} g_i(S)\right| > \frac{\varepsilon}{2}\right], \tag{10.3}$$

events (10.2) and (10.3) have the same probability.

Let $D = [\sup_{S\in\mathcal{S}} |\mathrm{Av}_{\bar{x}}(S) - \mathrm{Av}_{\bar{y}}(S)| > \varepsilon/2]$. We can use the result of the previous paragraph to conclude that

$$\begin{aligned}
P(D) &= P\left[\sup_{S\in\mathcal{S}}\left|\frac{1}{n}\sum_{i<n} f_i(S)\right| > \frac{\varepsilon}{2}\right] \\
&= P\left[\sup_{S\in\mathcal{S}}\left|\frac{1}{n}\sum_{i<n} g_i(S)\right| > \frac{\varepsilon}{2}\right] \\
&\le P\left(\left[\sup_{S\in\mathcal{S}}\left|\frac{1}{n}\sum_{i<n} \sigma_i\cdot 1_S(x_i)\right| > \frac{\varepsilon}{4}\right] \cup \left[\sup_{S\in\mathcal{S}}\left|\frac{1}{n}\sum_{i<n} \sigma_i\cdot 1_S(y_i)\right| > \frac{\varepsilon}{4}\right]\right) \qquad (10.4)\\
&\le 2P\left[\sup_{S\in\mathcal{S}}\left|\frac{1}{n}\sum_{i<n} \sigma_i\cdot 1_S(x_i)\right| > \frac{\varepsilon}{4}\right] \qquad (10.5)
\end{aligned}$$

where we obtain (10.4) since

$$\begin{aligned}
\left|\sum_{i<n} \sigma_i\cdot(1_S(x_i) - 1_S(y_i))\right| &= \left|\sum_{i<n} \sigma_i\cdot 1_S(x_i) - \sum_{i<n} \sigma_i\cdot 1_S(y_i))\right| \\
&\le \left|\sum_{i<n} \sigma_i\cdot 1_S(x_i)\right| + \left|\sum_{i<n} \sigma_i\cdot 1_S(y_i))\right|
\end{aligned}$$

and we obtain (10.5) by subadditivity.

Let

$$h(S) = \left|\frac{1}{n}\sum_{i<n} \sigma_i\cdot 1_S(x_i)\right|.$$

For each $\bar{a}\in X^n$, there is a subset $\mathcal{S}_{\bar{a}}\subseteq\mathcal{S}$ of size at most $\pi_{\mathcal{S}}(n)$ such that

$$\left[\sup_{S\in\mathcal{S}} h(S) > \frac{\varepsilon}{4},\ \bar{x}=\bar{a}\right] = \bigcup_{S\in\mathcal{S}}\left[h(S) > \frac{\varepsilon}{4},\ \bar{x}=\bar{a}\right] = \bigcup_{S\in\mathcal{S}_{\bar{a}}}\left[h(S) > \frac{\varepsilon}{4},\ \bar{x}=\bar{a}\right],$$

and for each $S\in\mathcal{S}_{\bar{a}}$, Chernoff's Bound (Theorem 8.4) asserts that

$$\nu^n\left[h(S) > \frac{\varepsilon}{4},\ \bar{x}=\bar{a}\right] < 2e^{-n\varepsilon^2/32}.$$

It follows that

$$\begin{aligned}\nu^n\left[\sup_{S\in\mathcal{S}} h(S) > \frac{\varepsilon}{4},\ \bar{x}=\bar{a}\right] &\leq \pi_{\mathcal{S}}(n)\cdot\nu^n\left[h(S)>\frac{\varepsilon}{4},\ \bar{x}=\bar{a}\right]\\ &< 2\pi_{\mathcal{S}}(n)e^{-n\varepsilon^2/32}.\end{aligned}$$

Let $C = [\sup_{S\in\mathcal{S}} h(S) > \varepsilon/4]$. Continuing from (10.5), we have

$$\begin{aligned}P(D) &\leq 2P(C)\\ &= 2\int_\Omega 1_C\ dP\\ &= 2\int_{X^n}\int_{X^n}\int_{\{-1,1\}^n} 1_C\ d\bar{\sigma}\,d\bar{y}\,d\bar{x}\\ &\leq 2\int_{X^n}\int_{X^n} 2\pi_{\mathcal{S}}(n)e^{-n\varepsilon^2/32}\ d\bar{y}\,d\bar{x}\\ &= 4\pi_{\mathcal{S}}(n)e^{-n\varepsilon^2/32}.\end{aligned}$$

For every $\bar{a}\in X^n$, let

$$B_{\bar{a}} = \left\{\bar{b}\in X^n : \sup_{S\in\mathcal{S}} |\mathrm{Av}_{\bar{a}}(S) - \mathrm{Av}_{\bar{b}}(S)| > \frac{\varepsilon}{2}\right\}.$$

Let

$$A = \left\{\bar{a}\in X^n : \mu^n(B_{\bar{a}}) \geq \frac{1}{2}\right\}.$$

Looking ahead to (10.6), we see that the function $\bar{x}\mapsto\mu^n(B_{\bar{x}})$ is measurable by Tonelli [2, Theorem 2.37], so A is μ^n-measurable. We now have

$$\begin{aligned}P(D) &= \int_\Omega 1_D\ dP\\ &= \int_{X^n}\int_{X^n}\int_{\{-1,1\}^n} 1_D\ d\bar{\sigma}\,d\bar{y}\,d\bar{x}\\ &\geq \int_A\int_{X^n} 1_D\ d\bar{y}\,d\bar{x}\\ &= \int_A \mu^n(B_{\bar{x}})\ d\bar{x} \qquad (10.6)\\ &\geq \frac{1}{2}\mu^n(A),\end{aligned}$$

so

$$\mu^n(A) \leq 2P(D) = 8\pi_{\mathcal{S}}(n)e^{-n\varepsilon^2/32}. \qquad (10.7)$$

Notice that the right-hand side of (10.7) is the same as the right-hand side of (10.1), so our proof will be complete if we can show

$$\left\{\bar{a}\in X^n : \sup_{S\in\mathcal{S}} |\mathrm{Av}_{\bar{a}}(S) - \mu(S)| > \varepsilon\right\} \subseteq A.$$

Given $\bar{a}\in A^c$, it follows that $\mu^n(B_{\bar{a}}^c) > 1/2$. Let $S\in\mathcal{S}$ and

$$B = \left\{\bar{b}\in X^n : |\mathrm{Av}_{\bar{b}}(S) - \mu(S)| > \frac{\varepsilon}{2}\right\}.$$

The Weak Law of Large Numbers (Proposition 9.2) implies that

$$\mu^n(B) \leq \frac{1}{n\varepsilon^2}.$$

Our theorem is vacuously true if the right-hand side of (10.1) is at least 1, so we may assume $n \geq 2/\varepsilon^2$. It follows that $B^c \cap B_{\bar{a}}^c$ is nonempty. Furthermore, for any $\bar{b} \in B^c \cap B_{\bar{a}}^c$, we have

$$|\mathrm{Av}_{\bar{a}}(S) - \mathrm{Av}_{\bar{b}}(S)| + |\mathrm{Av}_{\bar{b}}(S) - \mu(S)| \leq \varepsilon,$$

so we conclude that

$$|\mathrm{Av}_{\bar{a}}(S) - \mu(S)| \leq \varepsilon.$$

□

References

[1] Noga Alon and Joel H. Spencer, *The probabilistic method*, third ed., Wiley-Interscience Series in Discrete Mathematics and Optimization, John Wiley & Sons, Inc., Hoboken, NJ, 2008, With an appendix on the life and work of Paul Erdős. MR 2437651

[2] Gerald B. Folland, *Real analysis*, second ed., Pure and Applied Mathematics (New York), John Wiley & Sons, Inc., New York, 1999, Modern techniques and their applications, A Wiley-Interscience Publication. MR 1681462

[3] Pierre Simon, *A guide to NIP theories*, Lecture Notes in Logic, vol. 44, Association for Symbolic Logic, Chicago, IL; Cambridge Scientific Publishers, Cambridge, 2015. MR 3560428

[4] V. N. Vapnik and A. Ya. Chervonenkis, *On the uniform convergence of relative frequencies of events to their probabilities*, Measures of complexity, Springer, Cham, 2015, Reprint of Theor. Probability Appl. **16** (1971), 264–280, pp. 11–30. MR 3408730