Probability Theory

STAT 547C

Winter Term 1 2025/2026

Foreword

These notes are adapted from UBC's STAT 547C Probability Theory (Winter Term 1, 2025/2026). For corrections, please contact me at `nathan.cantafio@stat.ubc.ca`

Contents

1 Measure Theoretic Foundations **2**

1.1 Probability Spaces . . . 2

1.2 Probability Kernels . . . 5

1.3 Radon-Nikodym Derivatives as Densities . . . 10

1.4 Joint Distributions and σ-algebras as Information . . . 14

1.5 Modes of Convergence . . . 16

2 Conditioning and Independence **21**

2.1 Independence . . . 21

2.2 Conditional Expectations . . . 24

2.3 Conditional Distributions . . . 28

2.4 Conditional Independence . . . 30

3 Concentration Inequalities and Divergences **34**

3.1 Concentration Inequalities . . . 34

3.2 Integral Probability Metrics . . . 38

3.3 ϕ–Divergences . . . 40

1 Measure Theoretic Foundations

1.1 Probability Spaces

Definition 1.1. A **probability space** is a measure space $(\Omega, \mathcal{H}, \mathbb{P})$ such that the total mass is unit, i.e., $\mathbb{P}(\Omega) = 1$. Elements $\omega \in \Omega$ are called outcomes, and sets $A \in \mathcal{H}$ are called events.

Definition 1.2. Let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space and $(\mathcal{X}, \mathcal{M})$ be a measurable space. A **random variable** X is a measurable function

$$X : \Omega \to \mathcal{X}.$$

That is, for every $A \in \mathcal{M}$, the pre-image $X^{-1}(A) := \{\omega \in \Omega : X(\omega) \in A\}$ is an element of $\mathcal{H}$. When $(\mathcal{X}, \mathcal{M}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, we simply call X a real-valued random variable.

Definition 1.3. Let $(\mathcal{Y}, \mathcal{N})$ and $(\mathcal{X}, \mathcal{M})$ be measurable spaces. Let $h : \mathcal{Y} \to \mathcal{X}$ be a measurable function and ν be a measure on $(\mathcal{Y}, \mathcal{N})$. The **pushforward measure** of ν by h, denoted by $\nu \circ h^{-1}$ or $h_{\#}\nu$, is the measure on $(\mathcal{X}, \mathcal{M})$ defined by:

$$\left(\nu \circ h^{-1}\right)(A) := \nu(h^{-1}(A)) \quad \text{for all } A \in \mathcal{M},$$

where $h^{-1}(A)$ is the pre-image of A under h.

Proposition 1.4. *(Change of Variables, or Transformation Theorem)* Let $(\mathcal{Y}, \mathcal{N})$ and $(\mathcal{X}, \mathcal{M})$ be measurable spaces. Let $h : \mathcal{Y} \to \mathcal{X}$ be a measurable function and ν be a measure on $(\mathcal{Y}, \mathcal{N})$. Then

$$\int f \,\mathrm{d}\left(\nu \circ h^{-1}\right) = \int f \circ h \,\mathrm{d}\nu$$

for every positive $\mathcal{M}$-measurable function f

Proof. First consider the case where $f = \chi_A$ for some $A \in \mathcal{M}$. Then:

$$\int \chi_A \,\mathrm{d}(\nu \circ h^{-1}) = (\nu \circ h^{-1})(A) = \nu(h^{-1}(A)).$$

On the other hand

$$\int \chi_A \circ h \,\mathrm{d}\nu = \int \chi_{h^{-1}(A)} \,\mathrm{d}\nu = \nu(h^{-1}(A)).$$

Thus, the equality holds for simple functions by linearity. Now let f be any positive $\mathcal{M}$-measurable function. There exists an increasing sequence of non-negative simple functions $s_n \to f$. By the Monotone Convergence Theorem

$$\int f \,\mathrm{d}(\nu \circ h^{-1}) = \lim_{n\to\infty} \int s_n \,\mathrm{d}(\nu \circ h^{-1}) = \lim_{n\to\infty} \int s_n \circ h \,\mathrm{d}\nu.$$

Since $s_n \uparrow f$ implies $s_n \circ h \uparrow f \circ h$, applying MCT again to the right side yields:

$$\lim_{n\to\infty} \int s_n \circ h \,\mathrm{d}\nu = \int f \circ h \,\mathrm{d}\nu.$$

□

Definition 1.5. Let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space. If $X : \Omega \to \mathcal{X}$ is a random variable, the pushforward measure $\mathbb{P} \circ X^{-1}$ is called the **distribution** (or **law**) of X. We typically write:

$$\mathbb{P}(X \in A) := (\mathbb{P} \circ X^{-1})(A).$$

We use the notation $X \sim \mu$ to indicate that the distribution of X is the measure μ on $(\mathcal{X}, \mathcal{M})$.

Proposition 1.6. To specify the distribution of a random variable on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ it suffices to specify the *cumulative distribution function* (or CDF) which is a function $F : \mathbb{R} \to [0, 1]$ such that

$$F(x) = \mathbb{P}(X \in (-\infty, x]) = \mathbb{P}(X \leq x).$$

Proof. We first show that F is non-decreasing and right-continuous.

- Non-decreasing: Let $x < y$. Then $(-\infty, x] \subset (-\infty, y]$. By the monotonicity of measures:

$$F(x) = \mathbb{P}(X \in (-\infty, x]) \leq \mathbb{P}(X \in (-\infty, y]) = F(y).$$

- Right-continuous: Let $x \in \mathbb{R}$ and let $(x_n)_{n\in\mathbb{N}}$ be a decreasing sequence such that $x_n \to x$. Let $A_n = (-\infty, x_n]$. These sets are nested, their intersection is $\bigcap_{n=1}^{\infty} A_n = (-\infty, x]$, and $\mathbb{P}(X \in A_1) = \mathbb{P}(X \leq x_1) = F(x_1) \leq 1 < \infty$. So by continuity from above:

$$\lim_{n\to\infty} F(x_n) = \lim_{n\to\infty} \mathbb{P}(X \in A_n) = \mathbb{P}\left(X \in \bigcap_{n=1}^{\infty} A_n\right) = \mathbb{P}(X \in (-\infty, x]) = F(x).$$

Let $\mathcal{A}$ be the algebra of finite unions of half-open intervals $(a, b]$ in $\mathbb{R}$. Namely, let

$$\mathcal{A} = \{\emptyset\} \cup \left\{ \bigcup_{j=1}^{n} (a_j, b_j] : n \in \mathcal{N}, -\infty \leq a_1 < b_1 < a_2 < \cdots < b_n \leq \infty \right\}.$$

And let $\mu_0(\emptyset) = 0$, $\mu_0\left(\bigcup_{j=1}^{n}(a_j, b_j]\right) = \sum_{j=1}^{n} (F(b_j) - F(a_j))$. One can show that $\mathcal{A}$ is an algebra and that, because F is right-continuous, μ_0 is a pre-measure. By the Carathéodory Extension Theorem, this pre-measure extends to a measure on $\mathcal{M}(\mathcal{A}) = \mathcal{B}(\mathbb{R})$. And since $\mu_0(\mathbb{R}) = \mu(\mathbb{R}) = 1$ is σ-finite, this extension is unique. □

Definition 1.7. Let X be a real-valued random variable. The **expectation** of X is

$$\mathbb{E}X = \int X(\omega) \, \mathrm{d}\mathbb{P}(\omega)$$

Proposition 1.8. Let X be a random variable in $(\mathcal{X}, \mathcal{M})$. Then μ is the distribution of X if and only if

$$\mathbb{E}\left[f \circ X\right] = \int f \, \mathrm{d}\mu$$

for all positive $\mathcal{M}$-measurable f.

Proof. Suppose $\mu = \mathbb{P} \circ X^{-1}$. Then for any positive $\mathcal{M}$-measurable f change of variables yields

$$\int f \, \mathrm{d}\mu = \int f \, \mathrm{d}\big(\mathbb{P} \circ X^{-1}\big) = \int f \circ X \, \mathrm{d}\mathbb{P} = \mathbb{E}\left[f \circ X\right].$$

Conversely, suppose that $\mathbb{E}[f \circ X] = \int f \, \mathrm{d}\mu$ for all positive $\mathcal{M}$-measurable f. Take $f = \chi_A$ for any $A \in \mathcal{M}$. Then

$$\mu(A) = \int \chi_A \, \mathrm{d}\mu = \mathbb{E}[\chi_A \circ X] = \int \chi_A \circ X \, \mathrm{d}\mathbb{P} = \int \chi_A \, \mathrm{d}\left(\mathbb{P} \circ X^{-1}\right) = \mathbb{P}(X \in A)$$

as desired. □

Definition 1.9. Certain expectations have special names and are of particular interest. Let X be a real-valued random variable, $n \in \mathbb{N}$, and r a real-valued constant.

- $\mathbb{E}X^n$ is the **n-th moment of X** and $\mathbb{E}[(X - \mathbb{E}X)^n]$ is the **n-th central moment of X**
- $\mathbb{E}\left[(X - \mathbb{E}X)^2\right]$ is the **variance of X**
- When $X, r \geq 0$, $\mathbb{E}[\exp(-rX)]$ is the **Laplace transform of X**
- $\mathbb{E}[\exp(rX)]$ is the **Moment-generating function of X**
- $\mathbb{E}[\exp(irX)]$ is the **Fourier transform** (sometimes **characteristic function**) **of X** .

Proposition 1.10. Let X be a positive real-valued random variable, and let $r \geq 0$. The Laplace transform $h : [0, \infty] \to [0, \infty]$ given by

$$h(r) = \mathbb{E}[\exp(-rX)]$$

is continuous and decreasing on $(0, \infty)$ and is continuous at 0 if and only if $\mathbb{P}(X < \infty) = 1$.

Proof. Let $X \geq 0$ and $h(r) = \mathbb{E}[e^{-rX}]$.

- Decreasing: Let $0 \leq r_1 < r_2$. Since $X \geq 0$, we have $-r_2X \leq -r_1X$. Since the exponential function is strictly increasing, $e^{-r_2X} \leq e^{-r_1X}$ almost surely. By the monotonicity of the integral:
$$h(r_2) = \mathbb{E}[e^{-r_2X}] \leq \mathbb{E}[e^{-r_1X}] = h(r_1).$$
- Continuity on $(0, \infty)$: Let $r \in (0, \infty)$ and let $(r_n)_{n\in\mathbb{N}}$ be a sequence such that $r_n \to r$. Define $f_n(\omega) = \exp(-r_nX(\omega))$. Since the function $r \mapsto e^{-rx}$ is continuous for any x, $f_n \to \exp(-rX)$ pointwise. Since $r_n, X \geq 0$, we have $|f_n| \leq 1$ for all n. Because $\mathbb{P}$ is a probability measure, the constant function 1 is integrable. Thus, by the Dominated Convergence Theorem:
$$\lim_{n\to\infty} h(r_n) = \lim_{n\to\infty} \mathbb{E}[f_n] = \mathbb{E}[\lim_{n\to\infty} f_n] = \mathbb{E}[e^{-rX}] = h(r).$$
- Continuity at 0: Note that $h(0) = \mathbb{E}[e^0] = \mathbb{E}[1] = 1$. Let $r_n \downarrow 0$. We examine the pointwise limit of $f_n = e^{-r_nX}$:
  - If $X(\omega) < \infty$, then $\lim_{n\to\infty} e^{-r_nX(\omega)} = e^0 = 1$.
  - If $X(\omega) = \infty$, then $e^{-r_nX(\omega)} = 0$ for all n (since $r_n > 0$), so the limit is 0.

  Thus, $e^{-r_nX} \to \chi_{\{X<\infty\}}$ pointwise. Again, by DCT:
$$\lim_{r\to 0^+} h(r) = \mathbb{E}[\chi_{\{X<\infty\}}] = \mathbb{P}(X < \infty).$$
  Hence For h to be continuous at 0, we require $\mathbb{P}(X < \infty) = h(0) = 1$.

□

Definition 1.11. A function $\varphi : (a, b) \to \mathbb{R}$ is **convex** if for all $x, y \in (a, b)$ and $\lambda \in [0, 1]$,

$$\varphi(\lambda x + (1 - \lambda)y) \leq \lambda\varphi(x) + (1 - \lambda)\varphi(y).$$

Geometrically, this means the chord between any two points on the graph lies above the graph.

Theorem 1.12. *(Jensen's Inequality)* Let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space and X an integrable real-valued random variable. If $\varphi : \mathbb{R} \to \mathbb{R}$ is convex, then

$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)].$$

Proof. One can show that if φ is convex, then for every x_0, there exists $L(x) = ax + b$ such that $L(x_0) = \varphi(x_0)$ and $L(x) \leq \varphi(x)$ for all x. Let $x_0 = \mathbb{E}[X]$. Then there exist $a, b \in \mathbb{R}$ such that $\varphi(\mathbb{E}[X]) = a\mathbb{E}[X] + b$ and $L(X) \leq \varphi(X)$ almost surely. Then by the linearity and monotonicity of expectation:

$$\mathbb{E}[\varphi(X)] \geq \mathbb{E}[aX + b] = a\mathbb{E}[X] + b = \varphi(\mathbb{E}[X]).$$

□

1.2 Probability Kernels

Definition 1.13. Let $(\mathcal{X}, \mathcal{M})$ and $(\mathcal{Y}, \mathcal{N})$ be measurable spaces. Let $\overline{\mathbb{R}}_+ = \{z \in \mathbb{R} : z \geq 0\} \cup \{+\infty\}$. A **kernel from $\mathcal{X}$ into $\mathcal{Y}$** is a function $K : \mathcal{X} \times \mathcal{Y} \to \overline{\mathbb{R}}_+$ such that

(i) For every $E \in \mathcal{N}$, the map $x \mapsto K(x, E)$ is $\mathcal{M}$-measurable

(ii) For every $x \in \mathcal{X}$, the map $B \mapsto K(x, E)$ is a measure on $(\mathcal{Y}, \mathcal{N})$

Definition 1.14. The finiteness properties of measures naturally extend to kernels, categorized by whether the constraints hold pointwise for each $x \in \mathcal{X}$ or uniformly across the domain.

(i) K is a **probability kernel** if $K(x, \mathcal{Y}) = 1$ for all $x \in \mathcal{X}$.

(ii) **Finiteness and Boundedness:**

- K is **finite** if $K(x, \mathcal{Y}) < \infty$ for all $x \in \mathcal{X}$.
- K is **bounded** if it is uniformly finite, i.e., $\sup_{x \in \mathcal{X}} K(x, \mathcal{Y}) < \infty$.

(iii) **σ-Finiteness and σ-Boundedness:**

- K is **σ-finite** if for every $x \in \mathcal{X}$, there exists a countable partition $\mathcal{Y} = \bigcup_{n \in \mathbb{N}} E_n$ such that $K(x, E_n) < \infty$ for all $n \in \mathbb{N}$.
- K is **σ-bounded** if there exists a countable partition $\mathcal{Y} = \bigcup_{n \in \mathbb{N}} E_n$ that is independent of x, such that $\sup_{x \in \mathcal{X}} K(x, E_n) < \infty$ for all $n \in \mathbb{N}$.

(iv) **Σ-Finiteness and Σ-Boundedness:**

- K is **Σ-finite** if it can be decomposed into a sum of finite kernels, $K = \sum_{n \in \mathbb{N}} K_n$.
- K is **Σ-bounded** if it can be decomposed into a sum of bounded kernels, $K = \sum_{n \in \mathbb{N}} K_n$.

Definition 1.15. A kernel from $(\mathcal{X}, \mathcal{M})$ into itself, is simply called a **kernel on** $(\mathcal{X}, \mathcal{M})$.

- K is a **Markov kernel** if $K(x, \mathcal{X}) = 1$ for all $x \in \mathcal{X}$. In this case, $K(x, \cdot)$ is a probability measure for every x.
- K is a **sub-Markov kernel** if $K(x, \mathcal{X}) \leq 1$ for all $x \in \mathcal{X}$.

Proposition 1.16. Let K be a kernel from $(\mathcal{X}, \mathcal{M})$ into $(\mathcal{Y}, \mathcal{N})$, and let L be a kernel from $(\mathcal{Y}, \mathcal{N})$ into $(\mathcal{Z}, \mathcal{H})$. The **composition** KL, defined for $x \in \mathcal{X}$ and $B \in \mathcal{H}$ by

$$KL(x, B) := \int_{\mathcal{Y}} L(y, B) K(x, \mathrm{d}y),$$

is a kernel from $(\mathcal{X}, \mathcal{M})$ into $(\mathcal{Z}, \mathcal{H})$.

Proof. To show that KL is a kernel, we must verify two conditions: that $KL(x, \cdot)$ is a measure for each fixed x, and that $KL(\cdot, B)$ is a measurable function for each fixed B.

- Fix $x \in \mathcal{X}$.
  - Since $L(y, \emptyset) = 0$ for all $y \in \mathcal{Y}$, we have $KL(x, \emptyset) = \int_{\mathcal{Y}} 0 \, K(x, \mathrm{d}y) = 0$.
  - Let $\{B_n\}_{n \in \mathbb{N}}$ be a sequence of disjoint sets in $\mathcal{H}$, and let $B = \bigcup_{n=1}^{\infty} B_n$. Since $L(y, \cdot)$ is a measure for each y, we have $L(y, B) = \sum_{n=1}^{\infty} L(y, B_n)$. Then:

    $$KL(x, B) = \int_{\mathcal{Y}} \sum_{n=1}^{\infty} L(y, B_n) K(x, \mathrm{d}y).$$

    By the Monotone Convergence Theorem, we can interchange the sum and the integral:

    $$KL(x, B) = \sum_{n=1}^{\infty} \int_{\mathcal{Y}} L(y, B_n) K(x, \mathrm{d}y) = \sum_{n=1}^{\infty} KL(x, B_n).$$

    Thus, $KL(x, \cdot)$ is a measure on $(\mathcal{Z}, \mathcal{H})$.
- Fix $B \in \mathcal{H}$. We must show that $g(x) = KL(x, B)$ is $\mathcal{M}$-measurable. By definition of L being a kernel, the function $f(y) = L(y, B)$ is $\mathcal{N}$-measurable and non-negative.
  - If $f = \chi_A$ for some $A \in \mathcal{N}$, then $g(x) = \int \chi_A(y) K(x, \mathrm{d}y) = K(x, A)$, which is $\mathcal{M}$-measurable by the definition of K.
  - By linearity, the property holds for all non-negative simple functions $s = \sum c_i \chi_{A_i}$.
  - For a general non-negative measurable f, there exists a sequence of simple functions $s_n \uparrow f$. By the Monotone Convergence Theorem:

    $$g(x) = \int f(y) K(x, \mathrm{d}y) = \lim_{n \to \infty} \int s_n(y) K(x, \mathrm{d}y).$$

    Each $g_n(x) = \int s_n(y) K(x, \mathrm{d}y)$ is measurable. Since the pointwise limit of measurable functions is measurable, $g(x)$ is $\mathcal{M}$-measurable.

□

Proposition 1.17. Let K be a kernel from $(\mathcal{X}, \mathcal{M})$ into $(\mathcal{Y}, \mathcal{N})$, μ be a measure on $(\mathcal{X}, \mathcal{M})$, and f be a non-negative $\mathcal{N}$-measurable function.

(i) The mapping $\mu K : \mathcal{N} \to [0, \infty]$ defined by

$$(\mu K)(B) := \int_{\mathcal{X}} K(x, B) \mu(\mathrm{d}x)$$

is a measure on $(\mathcal{Y}, \mathcal{N})$.

(ii) The mapping $Kf : \mathcal{X} \to [0, \infty]$ defined by

$$(Kf)(x) := \int_{\mathcal{Y}} f(y) K(x, \mathrm{d}y)$$

is a non-negative $\mathcal{M}$-measurable function.

(iii) The following identity holds:

$$\int_{\mathcal{Y}} f \,\mathrm{d}(\mu K) = \int_{\mathcal{X}} (Kf) \,\mathrm{d}\mu = \int_{\mathcal{X}} \int_{\mathcal{Y}} f(y) K(x, \mathrm{d}y) \mu(\mathrm{d}x).$$

Note 1.18. In the language of probability, if μ is the distribution of X and K is the conditional distribution of Y given X, then μK is the *marginal distribution* of Y, $Kf(x)$ is the *conditional expectation* of $f(Y)$ given $X = x$, an $\int f \,\mathrm{d}(\mu K)$ is the *marginal expectation* of $f(Y)$. Part (iii) is the general version of the *Law of Total Expectation*: $\mathbb{E}[f(Y)] = \mathbb{E}[\mathbb{E}[f(Y)|X]]$.

Proof.

(i) μK is a measure:

- $(\mu K)(\emptyset) = \int_{\mathcal{X}} K(x, \emptyset) \mu(\mathrm{d}x) = \int_{\mathcal{X}} 0 \,\mu(\mathrm{d}x) = 0.$
- Let $\{B_n\}_{n \in \mathbb{N}}$ be a sequence of disjoint sets in $\mathcal{N}$. Since $K(x, \cdot)$ is a measure for each x:

$$(\mu K)\left(\bigcup_{n=1}^{\infty} B_n\right) = \int_{\mathcal{X}} K\left(x, \bigcup B_n\right) \mu(\mathrm{d}x) = \int_{\mathcal{X}} \sum_{n=1}^{\infty} K(x, B_n) \mu(\mathrm{d}x).$$

By the Monotone Convergence Theorem, we interchange the sum and integral:

$$\sum_{n=1}^{\infty} \int_{\mathcal{X}} K(x, B_n) \mu(\mathrm{d}x) = \sum_{n=1}^{\infty} (\mu K)(B_n).$$

Thus, μK is a measure on $(\mathcal{Y}, \mathcal{N})$.

(ii) Kf is $\mathcal{M}$-measurable: This is a special case of Proposition 1.21 where f does not depend on x.

- If $f = \chi_B$, then $(Kf)(x) = K(x, B)$, which is $\mathcal{M}$-measurable by definition.
- Extension to simple functions follows by linearity, and to positive measurable functions by the Monotone Convergence Theorem.

(iii) We prove $\int f \,\mathrm{d}(\mu K) = \int (Kf) \,\mathrm{d}\mu$ via the standard machinery of simple $\uparrow$ measurable function:

- If $f = \chi_B$, then $\int \chi_B \,\mathrm{d}(\mu K) = (\mu K)(B)$. By definition, $(\mu K)(B) = \int_{\mathcal{X}} K(x, B) \mu(\mathrm{d}x)$. Since $K(x, B) = (K\chi_B)(x)$, the identity holds.
- By linearity, it holds for all non-negative simple functions.
- For positive $\mathcal{N}$-measurable f, let $s_n \uparrow f$. By the Monotone Convergence Theorem applied twice (once for the measure μK and once for the function sequence $Ks_n \uparrow Kf$):

$$\int f \,\mathrm{d}(\mu K) = \lim_{n \to \infty} \int s_n \,\mathrm{d}(\mu K) = \lim_{n \to \infty} \int (Ks_n) \,\mathrm{d}\mu = \int (Kf) \,\mathrm{d}\mu.$$

□

Example 1.19. *(Pushforward as a Kernel)* Let $h : \mathcal{X} \to \mathcal{Y}$ be a measurable function. We can represent the pushforward measure $h_{\#}\mu$ using the kernel $K(x, B) = \delta_{h(x)}(B)$, where δ is the Dirac measure. Then:

$$(\mu K)(B) = \int_{\mathcal{X}} \delta_{h(x)}(B)\mu(\mathrm{d}x) = \int_{\mathcal{X}} \chi_{h^{-1}(B)}(x)\mu(\mathrm{d}x) = \mu(h^{-1}(B)).$$

Thus, $\mu K = h_{\#}\mu$. This shows that kernels generalize the notion of measurable transformations.

Proposition 1.20. Let K be a Σ-finite kernel from $(\mathcal{X}, \mathcal{M})$ into $(\mathcal{Y}, \mathcal{N})$ and let μ be a measure on $(\mathcal{X}, \mathcal{M})$. There exists a unique measure π on the product space $(\mathcal{X} \times \mathcal{Y}, \mathcal{M} \otimes \mathcal{N})$ such that for all $A \in \mathcal{M}$ and $B \in \mathcal{N}$:

$$\pi(A \times B) = \int_A K(x, B)\mu(\mathrm{d}x).$$

For any non-negative $(\mathcal{M} \otimes \mathcal{N})$-measurable function f, the measure π satisfies:

$$\int_{\mathcal{X}\times\mathcal{Y}} f \,\mathrm{d}\pi = \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} f(x, y)K(x, \mathrm{d}y) \right) \mu(\mathrm{d}x).$$

Furthermore, if μ is σ-finite and K is σ-bounded, then π is σ-finite.

Proof. We construct the measure π by defining it on an algebra and extending it.

1. **Definition on the Algebra:** Let $\mathcal{A}$ be the algebra of finite unions of measurable rectangles. For a rectangle $A \times B$, define

   $$\pi_0(A \times B) = \int_A K(x, B)\mu(\mathrm{d}x).$$

   This set function is well-defined and additive on $\mathcal{A}$. To show π_0 is a pre-measure, we check countable additivity. Let $C = \bigcup C_n$ with $C, C_n \in \mathcal{A}$ disjoint. By applying the Monotone Convergence Theorem to the integral of the kernel, one can show $\pi_0(C) = \sum \pi_0(C_n)$.

2. **Existence and Uniqueness:** By the Carathéodory Extension Theorem, π_0 extends to a measure π on $\sigma(\mathcal{A}) = \mathcal{M} \otimes \mathcal{N}$. If μ is σ-finite and K is σ-bounded, there exist partitions $\{A_i\} \subset \mathcal{M}$ and $\{B_j\} \subset \mathcal{N}$ with $\mu(A_i) < \infty$ and $\sup_x K(x, B_j) < \infty$. The rectangles $A_i \times B_j$ form a countable partition of $\mathcal{X} \times \mathcal{Y}$ with $\pi(A_i \times B_j) < \infty$, ensuring π is σ-finite and the extension is unique.

3. **Integral Identity:** We show $\int f \,\mathrm{d}\pi = \int (Tf) \,\mathrm{d}\mu$ (where Tf is the operator from Proposition 1.17).

   - If $f = \chi_{A\times B}$, then $\int \chi_{A\times B} \,\mathrm{d}\pi = \pi(A \times B)$. By definition:

     $$\pi(A \times B) = \int_{\mathcal{X}} \chi_A(x)K(x, B)\mu(\mathrm{d}x) = \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} \chi_{A\times B}(x, y)K(x, \mathrm{d}y) \right) \mu(\mathrm{d}x).$$

   - Let $\mathcal{C} = \{C \in \mathcal{M} \otimes \mathcal{N} : \text{the identity holds for } f = \chi_C\}$. We have shown rectangles are in $\mathcal{C}$. Using the linearity of the integral and the Monotone Class Lemma, $\mathcal{C} = \mathcal{M} \otimes \mathcal{N}$.

- The identity extends to non-negative simple functions by linearity and to all positive $(\mathcal{M}\otimes\mathcal{N})$-measurable f by the Monotone Convergence Theorem, noting that Proposition 1.21 guarantees the measurability of the inner integral.

□

Proposition 1.21. Let K be a Σ-finite kernel from $(\mathcal{X},\mathcal{M})$ into $(\mathcal{Y},\mathcal{N})$. For every non-negative $(\mathcal{M}\otimes\mathcal{N})$-measurable function $f:\mathcal{X}\times\mathcal{Y}\to[0,\infty]$, the function $Tf:\mathcal{X}\to[0,\infty]$ given by

$$Tf(x)=\int_{\mathcal{Y}} f(x,y)K(x,\mathrm{d}y)$$

is $\mathcal{M}$-measurable.

Proof. Assume first that K is a finite kernel. We proceed by showing that the property of "measurability under the integral" extends from a simple algebra to the full product σ-algebra.

- Let $\mathcal{A}$ be the algebra consisting of finite unions of measurable rectangles $A\times B$ (where $A\in\mathcal{M},B\in\mathcal{N}$). The product σ-algebra $\mathcal{M}\otimes\mathcal{N}$ is the σ-algebra generated by this algebra.
- Let $C\in\mathcal{A}$. Then χ_C can be written as a finite sum of indicators of disjoint rectangles $\sum_{i=1}^{n}\chi_{A_i\times B_i}$. For each rectangle:

  $$T\chi_{A_i\times B_i}(x)=\int_{\mathcal{Y}}\chi_{A_i}(x)\chi_{B_i}(y)K(x,\mathrm{d}y)=\chi_{A_i}(x)K(x,B_i).$$

  This is a product of two $\mathcal{M}$-measurable functions, hence $\mathcal{M}$-measurable. By the linearity of the integral, $T\chi_C$ is measurable for all $C\in\mathcal{A}$.
- Let $\mathcal{C}=\{C\in\mathcal{M}\otimes\mathcal{N}:T\chi_C \text{ is } \mathcal{M}\text{-measurable}\}$, then $\mathcal{C}$ is a monotone class.
  - Increasing Unions: Let $\{C_n\}_{n\in\mathbb{N}}$ be a sequence in $\mathcal{C}$ such that $C_n\uparrow C$. Then $\chi_{C_n}\uparrow\chi_C$ pointwise. By the Monotone Convergence Theorem:

    $$T\chi_C(x)=\int_{\mathcal{Y}}\chi_C(x,y)K(x,\mathrm{d}y)=\lim_{n\to\infty}\int_{\mathcal{Y}}\chi_{C_n}(x,y)K(x,\mathrm{d}y)=\lim_{n\to\infty}T\chi_{C_n}(x).$$

    Since each $T\chi_{C_n}$ is $\mathcal{M}$-measurable and the pointwise limit of measurable functions is measurable, $T\chi_C$ is measurable. Thus $C\in\mathcal{C}$.
  - Decreasing Intersections: Let $\{C_n\}_{n\in\mathbb{N}}$ be a sequence in $\mathcal{C}$ such that $C_n\downarrow C$. Then $\chi_{C_n}\downarrow\chi_C$ pointwise. Since K is a *finite* kernel, we have $T\chi_{C_n}\le K(x,\mathcal{Y})<\infty$. By the Dominated Convergence Theorem:

    $$T\chi_C(x)=\lim_{n\to\infty}T\chi_{C_n}(x).$$

    Again, the limit of measurable functions is measurable, so $C\in\mathcal{C}$.
- Since $\mathcal{A}$ is an algebra and $\mathcal{C}$ is a monotone class containing $\mathcal{A}$, by the Monotone Class Lemma,

  $$\mathcal{M}\otimes\mathcal{N}=\sigma(\mathcal{A})=\mathcal{C}(\mathcal{A})\subset\mathcal{C}.$$

  Thus, $T\chi_C$ is $\mathcal{M}$-measurable for all $C\in\mathcal{M}\otimes\mathcal{N}$.

- The result extends to simple functions by linearity and to all non-negative measurable functions f by taking a sequence $s_n \uparrow f$ and applying the Monotone Convergence Theorem, noting that the pointwise limit of measurable functions is measurable.

Finally, if K is Σ-finite, $K = \sum K_n$ for finite kernels K_n. Then $Tf = \sum T_n f$, and the sum of measurable functions remains measurable. □

Proposition 1.22. Let $(\mathcal{X}, \mathcal{M})$ and $(\mathcal{Y}, \mathcal{N})$ be measurable spaces. Let ν be a σ-finite measure on $(\mathcal{Y}, \mathcal{N})$ and let $k : \mathcal{X} \times \mathcal{Y} \to [0, \infty]$ be an $(\mathcal{M} \otimes \mathcal{N})$-measurable function. Then the mapping $K : \mathcal{X} \times \mathcal{N} \to [0, \infty]$ defined by

$$K(x, B) := \int_B k(x, y)\nu(\mathrm{d}y)$$

is a kernel from $(\mathcal{X}, \mathcal{M})$ into $(\mathcal{Y}, \mathcal{N})$.

Proof. We must verify that $K(x, \cdot)$ is a measure for each fixed x, and that $K(\cdot, B)$ is a measurable function for each fixed B.

- Fix $x \in \mathcal{X}$. The mapping $B \mapsto \int_B k(x, y)\nu(\mathrm{d}y)$ is a measure on $(\mathcal{Y}, \mathcal{N})$ by the standard construction of a measure via a density (Radon-Nikodym theorem in reverse). Specifically, $K(x, \emptyset) = 0$ and countable additivity follows from the Monotone Convergence Theorem applied to the integrand $k(x, y) \sum \chi_{B_n}(y)$.
- Fix $B \in \mathcal{N}$. We must show $x \mapsto K(x, B)$ is $\mathcal{M}$-measurable. Define the function

  $$f(x, y) = k(x, y)\chi_B(y).$$

  Since k is $(\mathcal{M} \otimes \mathcal{N})$-measurable and χ_B is trivially so, f is $(\mathcal{M} \otimes \mathcal{N})$-measurable.

  We can view the measure ν as a constant kernel $L(x, \mathrm{d}y) = \nu(\mathrm{d}y)$. Since ν is σ-finite, L is a Σ-finite kernel. Then:

  $$K(x, B) = \int_{\mathcal{Y}} f(x, y)L(x, \mathrm{d}y).$$

  By Proposition 1.21, this integral defines an $\mathcal{M}$-measurable function of x.

□

1.3 Radon-Nikodym Derivatives as Densities

In this section, we establish the conditions under which one measure can be represented as a density relative to another. For a more exhaustive treatment of the Lebesgue decomposition, see the notes from *MATH 507: Measure Theory and Integration*.

Definition 1.23. Let ν be a signed measure and μ a positive measure on $(\mathcal{X}, \mathcal{M})$. Then ν is said to be **absolutely continuous** w.r.t. μ, denoted $\nu \ll \mu$, if $\mu(E) = 0 \implies \nu(E) = 0$ for all $E \in \mathcal{M}$.

Theorem 1.24. *(Lebesgue-Radon-Nikodym)* Let ν be a σ-finite signed measure, and μ a positive σ-finite measure on $(\mathcal{X}, \mathcal{M})$. Then there exist unique σ-finite measures λ and ρ on $(\mathcal{X}, \mathcal{M})$ such that

$$\lambda \perp \mu, \quad \rho \ll \mu, \quad \text{and} \quad \nu = \lambda + \rho \qquad \text{(Lebesgue)}.$$

Moreover there is an extended μ-integrable function $f : \mathcal{X} \to \mathbb{R}$ such that $d\rho = f d\mu$ (i.e. such that $\rho(E) = \int_E f \, \mathrm{d}\mu$ for all $E \in \mathcal{M}$). And any two such functions are equal μ-a.e. (Radon-Nikodym). As notation we write $f = \frac{d\rho}{d\mu}$ and call this the **Radon-Nikodym derivative**.

As noted in Exercise 3.14 of Folland [1], the assumption that ν is σ-finite is not strictly necessary for the existence of a Radon-Nikodym derivative, provided we allow the density to take values in the extended real line (see Proposition 1.25 below for a precise statement).

Proposition 1.25. Let ν be an arbitrary signed measure and μ a σ-finite measure on $(\mathcal{X}, \mathcal{M})$ such that $\nu \ll \mu$. Then there exists an extended μ-integrable $f : \mathcal{X} \to [-\infty, \infty]$ such that $\mathrm{d}\nu = f\,\mathrm{d}\mu$.

The only condition required for the existence of the derivative $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ is that the dominating measure μ be σ-finite. Indeed, if $\mathrm{d}\nu = f\,\mathrm{d}\mu$, then $\mu(A) = 0 \implies \nu(A) = \int_A f\,\mathrm{d}\mu = 0$, which shows that $\nu \ll \mu$ is both necessary and sufficient.

Proposition 1.26. Let ν and μ be measures $(\mathcal{X}, \mathcal{M})$. Suppose $\nu \ll \mu$ and μ is σ-finite. Then

$$\int f \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\,\mathrm{d}\mu = \int f\,\mathrm{d}\nu$$

for all $\mathcal{M}$-measurable $f : \mathcal{X} \to [0, \infty]$.

Proof. First for any measurable E,

$$\int \chi_E \frac{d\nu}{d\mu} d\mu = \int_E \frac{d\nu}{d\mu} d\mu = \nu(E)$$

since $\frac{d\nu}{d\mu}$ is the Radon-Nikodym density (which exists by σ-finiteness of μ). On the other hand $\int \chi_E d\nu = \nu(E)$. Linearity then extends the result to simple functions. For any positive measurable $f : X \to [0, \infty]$, there is a sequence of simple functions $f_n : X \to [0, \infty)$ increasing to f. For example,

$$f_n(x) = \sum_{m=0}^{4^n - 1} \frac{m}{2^n} \chi_{f^{-1}(I_{m,n})} + 2^n \chi_{f^{-1}([2^n, \infty])}, \qquad I_{m,n} = [m/2^n, (m+1)/2^n].$$

By the M.C.T. applied to $f_n \frac{d\nu}{d\mu}$, the result for simple functions, and then the M.C.T. applied to f_n:

$$\int f \frac{d\nu}{d\mu} d\mu = \int \lim_{n\to\infty} f_n \frac{d\nu}{d\mu} d\mu = \lim_{n\to\infty} \int f_n \frac{d\mu}{d\mu} d\lambda = \lim_{n\to\infty} \int f_n d\nu = \int f d\nu.$$

(Note that $\frac{d\nu}{d\mu}$ is non-negative, so $f_n \frac{d\nu}{d\mu}$ increases to $f \frac{d\nu}{d\mu}$). □

Proposition 1.27. Let μ be a σ-finite measure and $\nu \ll \mu$ with density $p = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}$. Then:

(i) ν is σ-finite if and only if $p < \infty$ μ-a.e.

(ii) $\mu \ll \nu$ (and thus $\mu \equiv \nu$) if and only if $p > 0$ μ-a.e.

Proof.

(i) Suppose ν is σ-finite. Let $\bigcup_{n\in\mathbb{N}} E_n = \mathcal{X}$ be a disjoint union such that $\nu(E_n) < \infty$. Since $\nu(E_n) = \int_{E_n} p\,\mathrm{d}\mu < \infty$, the function p must be finite μ-a.e. on each E_n, and thus μ-a.e. on $\mathcal{X}$. Conversely, if $p < \infty$ μ-a.e., we can define $E_{n,m} = \{x \in F_n : m - 1 \le p(x) < m\}$ where $\{F_n\}$ is a σ-finite partition for μ. Then $\nu(E_{n,m}) = \int_{E_{n,m}} p\,\mathrm{d}\mu \le m\mu(F_n) < \infty$, providing a σ-finite partition for ν.

(ii) Suppose $\mu \ll \nu$. Let $A_0 = \{x : p(x) = 0\}$. Then $\nu(A_0) = \int_{A_0} p\,\mathrm{d}\mu = 0$. Since $\mu \ll \nu$, we must have $\mu(A_0) = 0$, so $p > 0$ μ-a.e. Conversely, if $p > 0$ μ-a.e., let $\nu(A) = 0$. Then $\int_A p\,\mathrm{d}\mu = 0$. Since p is strictly positive μ-a.e., the integrand is zero only if $\mu(A) = 0$. Thus $\mu \ll \nu$.

□

Proposition 1.28. Let ν, μ, λ be measures on $(\mathcal{X}, \mathcal{M})$. Assume all required Radon-Nikodym derivatives exist and λ is σ-finite.

(i) Linearity: For $a, b \geq 0$, $\frac{\mathrm{d}(a\nu+b\mu)}{\mathrm{d}\lambda} = a\frac{\mathrm{d}\nu}{\mathrm{d}\lambda} + b\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}$ (λ-a.e.).

(ii) Chain Rule: If $\nu \ll \mu \ll \lambda$, then $\frac{\mathrm{d}\nu}{\mathrm{d}\lambda} = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}$ (λ-a.e.).

(iii) Reciprocal: If $\nu \equiv \mu$, then $\frac{\mathrm{d}\nu}{\mathrm{d}\mu} = \left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)^{-1}$ (μ-a.e. and ν-a.e.).

(iv) Quotient Rule: $\frac{\mathrm{d}\nu}{\mathrm{d}\mu} = \frac{\mathrm{d}\nu}{\mathrm{d}\lambda} / \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}$ (μ-a.e.).

To prove these properties by evaluating them on sets (measures) rather than general functions, we rely on the defining property of the Radon-Nikodym derivative: p is the density $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ if and only if $\nu(A) = \int_A p \,\mathrm{d}\mu$ for all $A \in \mathcal{M}$.

Proof. In each case, we show that the proposed density, when integrated over an arbitrary set $A \in \mathcal{M}$ with respect to the dominating measure, yields the correct measure of A.

(i) Linearity: Let $A \in \mathcal{M}$. By the definition of the measures ν and μ via their densities with respect to λ:

$$(a\nu + b\mu)(A) = a\nu(A) + b\mu(A) = a\int_A \frac{\mathrm{d}\nu}{\mathrm{d}\lambda}\,\mathrm{d}\lambda + b\int_A \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}\,\mathrm{d}\lambda.$$

By the linearity of the integral:

$$= \int_A \left(a\frac{\mathrm{d}\nu}{\mathrm{d}\lambda} + b\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}\right)\mathrm{d}\lambda.$$

Since this holds for all $A \in \mathcal{M}$, the integrand is the Radon-Nikodym derivative of $(a\nu + b\mu)$ with respect to λ.

(ii) Chain Rule: Let $A \in \mathcal{M}$. By the definition of the density $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$:

$$\nu(A) = \int_A \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\,\mathrm{d}\mu = \int_{\mathcal{X}} \chi_A \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\,\mathrm{d}\mu.$$

Since $\mu \ll \lambda$, we can apply Proposition 1.26 to the function $f = \chi_A \frac{\mathrm{d}\nu}{\mathrm{d}\mu}$:

$$\int_{\mathcal{X}} \left(\chi_A \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right)\mathrm{d}\mu = \int_{\mathcal{X}} \left(\chi_A \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right)\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}\,\mathrm{d}\lambda = \int_A \left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}\right)\mathrm{d}\lambda.$$

Thus, the product of the densities is the derivative of ν with respect to λ.

(iii) Reciprocal: Since $\nu \ll \mu$ and $\mu \ll \nu$, we can apply the Chain Rule from part (ii) with $\lambda = \nu$:

$$\frac{\mathrm{d}\nu}{\mathrm{d}\nu} = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\frac{\mathrm{d}\mu}{\mathrm{d}\nu} \quad \nu\text{-a.e.}$$

By the uniqueness of the Radon-Nikodym derivative, and noting that $\nu(A) = \int_A 1\,\mathrm{d}\nu$ implies $\frac{\mathrm{d}\nu}{\mathrm{d}\nu} = 1$, we have $1 = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ ν-a.e. (and by absolute continuity, μ-a.e.). Solving for $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ yields the reciprocal.

(iv) Quotient Rule: Applying the Chain Rule to $\nu \ll \mu \ll \lambda$ gives:

$$\frac{\mathrm{d}\nu}{\mathrm{d}\lambda} = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}.$$

Dividing by $\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}$ (which is strictly positive μ-a.e. by Proposition 1.27) yields the result.

□

Proposition 1.29. Suppose $\nu \ll \mu$ on $(\mathcal{X}, \mathcal{M})$. For any measurable injection $h : \mathcal{X} \to \mathcal{Y}$, the density of the pushforward measures satisfies:

$$\frac{\mathrm{d}(\nu \circ h^{-1})}{\mathrm{d}(\mu \circ h^{-1})} = \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \circ h^{-1} \quad (\mu \circ h^{-1})\text{-a.e.}$$

Proof. Let f be a positive measurable function on $\mathcal{Y}$. By the definition of pushforward and Change of Variables:

$$\int_{\mathcal{Y}} f \,\mathrm{d}(\nu \circ h^{-1}) = \int_{\mathcal{X}} (f \circ h)\,\mathrm{d}\nu = \int_{\mathcal{X}} (f \circ h)\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\,\mathrm{d}\mu.$$

Since h is an injection, $(f \circ h) \cdot \frac{\mathrm{d}\nu}{\mathrm{d}\mu} = (f \cdot (\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \circ h^{-1})) \circ h$. Applying Change of Variables again:

$$\int_{\mathcal{X}} \left(f \cdot \left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \circ h^{-1} \right) \right) \circ h \,\mathrm{d}\mu = \int_{\mathcal{Y}} f(y) \frac{\mathrm{d}\nu}{\mathrm{d}\mu}(h^{-1}(y))(\mu \circ h^{-1})(\mathrm{d}y).$$

The result follows from the uniqueness of the Radon-Nikodym derivative. □

Proposition 1.30. Let $(\mathcal{X}, \mathcal{M})$ and $(\mathcal{Y}, \mathcal{N})$ be measurable spaces. Let ν, μ be measures on $\mathcal{X}$ and L, K be kernels from $\mathcal{X}$ to $\mathcal{Y}$. If $p = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ exists and $k(x, y) = \frac{\mathrm{d}L(x,\cdot)}{\mathrm{d}K(x,\cdot)}(y)$ exists as an $(\mathcal{M} \otimes \mathcal{N})$-measurable function, then the density of the marginal measure νL with respect to μK is:

$$\frac{\mathrm{d}(\nu L)}{\mathrm{d}(\mu K)}(y) = \int_{\mathcal{X}} p(x)k(x, y)M(y, \mathrm{d}x)$$

where M is the kernel obtained from the disintegration of the joint measure $\pi(\mathrm{d}x, \mathrm{d}y) = \mu(\mathrm{d}x)K(x, \mathrm{d}y)$ with respect to its marginal μK.

Proof. Let f be a positive measurable function on $\mathcal{Y}$. Using Proposition 1.17:

$$(\nu L)f = \int_{\mathcal{X}} (Lf)(x)\nu(\mathrm{d}x) = \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} f(y)L(x, \mathrm{d}y) \right) p(x)\mu(\mathrm{d}x).$$

Using the density $k(x, y)$, we write $L(x, \mathrm{d}y) = k(x, y)K(x, \mathrm{d}y)$:

$$= \int_{\mathcal{X}} \int_{\mathcal{Y}} f(y)k(x, y)p(x)K(x, \mathrm{d}y)\mu(\mathrm{d}x) = \int_{\mathcal{X}\times\mathcal{Y}} f(y)k(x, y)p(x)\pi(\mathrm{d}x, \mathrm{d}y).$$

By the disintegration theorem, $\pi(\mathrm{d}x, \mathrm{d}y) = (\mu K)(\mathrm{d}y)M(y, \mathrm{d}x)$. Substituting this:

$$= \int_{\mathcal{Y}} f(y) \left(\int_{\mathcal{X}} p(x)k(x, y)M(y, \mathrm{d}x) \right) (\mu K)(\mathrm{d}y).$$

Thus, the inner integral is the Radon-Nikodym derivative of νL with respect to μK. □

1.4 Joint Distributions and σ-algebras as Information

Note 1.31. Let $n \in \mathbb{N}$. For each $i \in \{1, \dots, n\}$, let $(\mathcal{X}_i, \mathcal{M}_i)$ be a measurable space. The product space is $\mathcal{X} = \prod_{i=1}^n \mathcal{X}_i$. The **product σ-algebra** $\mathcal{M} = \bigotimes_{i=1}^n \mathcal{M}_i$ is the σ-algebra on $\mathcal{X}$ generated by the algebra $\mathcal{A}$ of **measurable rectangles**:

$$\mathcal{A} = \{A_1 \times \cdots \times A_n : A_i \in \mathcal{M}_i\}.$$

The product measure on $(\mathcal{X}, \mathcal{M})$ is defined via the Carathéodory construction: it is the outer measure induced by the pre-measure on $\mathcal{A}$ (where the pre-measure of a rectangle is the product of its components' measures), restricted to $\mathcal{M}$.

Definition 1.32. Let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space and $X_1, \dots, X_n$ be random variables where $X_i : \Omega \to \mathcal{X}_i$ is $\mathcal{M}_i$-measurable. The **joint map** $X : \Omega \to \prod \mathcal{X}_i$ is defined by:

$$X(\omega) = (X_1(\omega), \dots, X_n(\omega)).$$

The **joint distribution** of $X_1, \dots, X_n$ is the pushforward measure $\mu_X = \mathbb{P} \circ X^{-1}$ on $(\mathcal{X}, \mathcal{M})$.

Proposition 1.33. The joint map X is $\mathcal{M}$-measurable if and only if each X_i is $\mathcal{M}_i$-measurable. Furthermore, μ_X is uniquely determined by its values on measurable rectangles:

$$\mu_X(A_1 \times \cdots \times A_n) = \mathbb{P}(X_1 \in A_1, \dots, X_n \in A_n).$$

Proof. ($\implies$) Let $\pi_i : \mathcal{X} \to \mathcal{X}_i$ be the projection $\pi_i(x) = x_i$, which is measurable by the definition of $\mathcal{M}$. Since $X_i = \pi_i \circ X$, the composition of measurable functions is measurable.
($\impliedby$) For any rectangle A, $X^{-1}(A) = \bigcap_{i=1}^n X_i^{-1}(A_i)$. Since each $X_i^{-1}(A_i) \in \mathcal{H}$ and $\mathcal{H}$ is closed under finite intersection, $X^{-1}(A) \in \mathcal{H}$. Because rectangles generate $\mathcal{M}$, X is measurable. Uniqueness follows from the Carathéodory Extension Theorem, as rectangles form an algebra (and a π-system) that generates $\mathcal{M}$. □

Infinite Products and Stochastic Processes

We now generalize to an arbitrary index set T (e.g., $\mathbb{N}, \mathbb{R}, [0, \infty)$) with potentially heterogeneous state spaces $(\mathcal{X}_t, \mathcal{M}_t)$ for each $t \in T$.

Definition 1.34. Let $\mathcal{X}^T = \prod_{t \in T} \mathcal{X}_t$. The **product σ-algebra** $\mathcal{M}^T = \bigotimes_{t \in T} \mathcal{M}_t$ is the σ-algebra on $\mathcal{X}^T$ generated by the collection of **cylinder sets** of the form:

$$C = \{x \in \mathcal{X}^T : x_{t_1} \in A_{t_1}, \dots, x_{t_k} \in A_{t_k}\}$$

for any finite subset $\{t_1, \dots, t_k\} \subseteq T$ and $A_{t_j} \in \mathcal{M}_{t_j}$.

Note 1.35. A **stochastic process** is the specific case where $(\mathcal{X}_t, \mathcal{M}_t) = (\mathcal{X}, \mathcal{M})$ for all $t \in T$. We call $\mathcal{X}^T$ the **path space**. The joint map $X(\omega) = (X_t(\omega))_{t \in T}$ is $\mathcal{M}^T$-measurable if and only if each coordinate X_t is $\mathcal{M}_t$-measurable. Its probability law μ_X is uniquely determined by its values on cylinder sets (finite-dimensional distributions) via Carathéodory.

Information and Generated σ-algebras

In probability theory, a sub-σ-algebra $\mathcal{G} \subset \mathcal{H}$ represents "information" or "knowledge." To know $\mathcal{G}$ is to know, for every $A \in \mathcal{G}$, whether the event A occurred.

Definition 1.36. *(Generated σ-algebras)*

1. **Single Variable:** For $X : \Omega \to \mathcal{X}$, the σ-algebra **generated by** X is the smallest sub-σ-algebra of $\mathcal{H}$ making X measurable, consisting of all pre-images:
$$\sigma(X) = X^{-1}(\mathcal{M}) = \{X^{-1}(A) : A \in \mathcal{M}\}.$$
2. **Finite Collection:** For $X_1, \dots, X_n$, the generated σ-algebra is $\sigma(X_1, \dots, X_n) = \sigma(X)$, where X is the joint map. This is the smallest σ-algebra containing all sets $\{X_i \in A_i\}$ for $i = 1, \dots, n$.
3. **Arbitrary Collection:** For a collection $\{X_t\}_{t\in T}$, the **σ-algebra generated by the collection**, denoted $\sigma(X_t : t \in T)$, is the smallest σ-algebra making every X_t measurable:
$$\sigma(X_t : t \in T) = \bigvee_{t\in T} \sigma(X_t) = \sigma\left(\bigcup_{t\in T} \sigma(X_t)\right).$$

Proposition 1.37. If $X : \Omega \to \mathcal{X}^T$ is the joint map of the collection $\{X_t\}_{t\in T}$, then:
$$\sigma(X) = X^{-1}(\mathcal{M}^T) = \sigma(X_t : t \in T).$$

Proof. By Proposition 1.33 (extended to the infinite case), X is measurable with respect to a σ-algebra $\mathcal{G}$ if and only if each X_t is measurable with respect to $\mathcal{G}$. Thus, the smallest σ-algebra satisfying one condition must satisfy the other. □

Note 1.38. *(Heuristics of Information)*

- **Indicator Functions:** If $A \in \mathcal{G}$, we "know" A because its indicator $\chi_A(\omega)$ is $\mathcal{G}$-measurable.
- **The Scale of Information:** Larger σ-algebras represent finer knowledge.
    1. **Trivial Information:** $\mathcal{G} = \{\emptyset, \Omega\}$. Only constant functions are measurable; we know nothing of the outcome.
    2. **Generated Information:** $\mathcal{G} = \sigma(X)$. We know the value $X(\omega)$, allowing us to determine $f(X(\omega))$, but we do not know ω itself.
    3. **Full Information:** $\mathcal{G} = \mathcal{H}$. We know the outcome of every possible measurable event.
- **Information Loss:** If $\sigma(Y) \subset \sigma(X)$, then X provides at least as much information as Y. For $X \sim \mathcal{N}(0, 1)$ and $Y = |X|$, $\sigma(Y) \subsetneq \sigma(X)$. Knowing X determines Y, but knowing Y leaves the sign of X uncertain.

Proposition 1.39. *(Doob-Dynkin Lemma)* Let $V : \Omega \to \mathbb{R}$ be a random variable.

(i) **Finite Case:** V is $\sigma(X_1, \dots, X_n)$-measurable if and only if there exists a measurable function $f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathbb{R}$ such that $V = f(X_1, \dots, X_n)$.

(ii) **Infinite Case:** Let T be an arbitrary index set. V is $\sigma(X_t : t \in T)$-measurable if and only if there exists a **countable** sequence of indices $\{t_1, t_2, \dots\} \subseteq T$ and a measurable function $f : \prod_{j=1}^{\infty} \mathcal{X}_{t_j} \to \mathbb{R}$ such that:
$$V = f(X_{t_1}, X_{t_2}, \dots).$$

Proof. Omitted. (Note: The infinite case highlights that even in uncountable processes, any single $\sigma(X_t : t \in T)$-measurable random variable only depends on at most countably many coordinates). □

1.5 Modes of Convergence

We now study the various ways in which a sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ can converge to a limit X. Throughout, let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space and $X, \{X_n\}_{n\in\mathbb{N}}$ be real-valued random variables. Many proofs appeal to concentration inequalities which I have irresponsibly postponed.

Definition 1.40. We define the following two primary modes of convergence:

(i) X_n **converges almost surely** to X, denoted $X_n \xrightarrow{a.s.} X$, if

$$\mathbb{P}\left(\left\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right) = 1.$$

(ii) X_n **converges in probability** to X, denoted $X_n \xrightarrow{P} X$, if for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0.$$

Proposition 1.41. *(Continuous Mapping Theorem for a.s. Convergence)* Let $f : \mathbb{R}^K \to \mathbb{R}$ be a function. If $X_{k,n} \xrightarrow{a.s.} X_k$ for each $k = 1, \ldots, K$, and f is continuous on a set $A \subset \mathbb{R}^K$ such that $\mathbb{P}((X_1, \ldots, X_K) \in A) = 1$, then

$$f(X_{1,n}, \ldots, X_{K,n}) \xrightarrow{a.s.} f(X_1, \ldots, X_K).$$

Proof. Let $\Omega_k = \{\omega \in \Omega : \lim X_{k,n}(\omega) = X_k(\omega)\}$ and $\Omega_A = \{\omega \in \Omega : (X_1(\omega), \ldots, X_K(\omega)) \in A\}$. By assumption, $\mathbb{P}(\Omega_k) = 1$ and $\mathbb{P}(\Omega_A) = 1$. Let $\Omega_0 = \Omega_A \cap \left(\bigcap_{k=1}^K \Omega_k\right)$. Then $\mathbb{P}(\Omega_0) = 1$. For any $\omega \in \Omega_0$, the sequence of vectors $(X_{1,n}(\omega), \ldots, X_{K,n}(\omega))$ converges to $(X_1(\omega), \ldots, X_K(\omega)) \in A$. And by the continuity of f on A, it follows that $f(X_{1,n}(\omega), \ldots, X_{K,n}(\omega)) \to f(X_1(\omega), \ldots, X_K(\omega))$. □

Proposition 1.42. $X_n \xrightarrow{a.s.} X$ if and only if

$$\sum_{n\in\mathbb{N}} \chi_{(\varepsilon,\infty)} \circ |X_n - X| < \infty \qquad a.s.$$

for all $\varepsilon > 0$. Moreover, the following are equivalent:

(i) X_n converges almost surely

(ii) $\sup_{i,j\geq n} |X_i - X_j| \xrightarrow{a.s.} 0$

(iii) $\sup_k |X_{n+k} - X_n| \xrightarrow{a.s.} 0$

Proof. Omitted. □

Lemma 1.43. *(Borel-Cantelli)* Let $\{H_n\}_{n\in\mathbb{N}}$ be a sequence of events. If $\sum_{n=1}^\infty \mathbb{P}(H_n) < \infty$, then

$$\mathbb{P}(\limsup H_n) = 0, \quad \text{or equivalently, } \sum_{n=1}^\infty \chi_{H_n} < \infty \text{ a.s.}$$

Proof. By the Monotone Convergence Theorem, $\mathbb{E}[\sum \chi_{H_n}] = \sum \mathbb{E}[\chi_{H_n}] = \sum \mathbb{P}(H_n)$. Since the sum of probabilities is finite, the expectation of the random sum is finite. A non-negative random variable with finite expectation must be finite almost surely. □

Proposition 1.44. A sequence of random variables X_n converges almost surely to X if either of the following conditions holds:

(i) For every $\varepsilon > 0$, $\sum_{n=1}^{\infty} \mathbb{P}(|X_n - X| > \varepsilon) < \infty$.

(ii) There exists a sequence $\varepsilon_n > 0$ such that $\sum_{n=1}^{\infty} \varepsilon_n < \infty$ and $\sum_{n=1}^{\infty} \mathbb{P}(|X_{n+1} - X_n| > \varepsilon_n) < \infty$.

Proof.

(i) For a fixed $\varepsilon > 0$, let $H_n = \{|X_n - X| > \varepsilon\}$. By the Borel-Cantelli Lemma, $\sum \mathbb{P}(H_n) < \infty$ implies that $\mathbb{P}(\limsup H_n) = 0$. Thus, for almost all ω, $|X_n(\omega) - X(\omega)| > \varepsilon$ for only finitely many n. Since this holds for all $\varepsilon > 0$, $X_n \xrightarrow{a.s.} X$ follows by Proposition 1.44.

(ii) Let $H_n = \{|X_{n+1} - X_n| > \varepsilon_n\}$. By Borel-Cantelli, $|X_{n+1} - X_n| \leq \varepsilon_n$ for all $n \geq N(\omega)$ almost surely. Since $\sum \varepsilon_n < \infty$, the sequence is almost surely Cauchy:

$$|X_{n+k} - X_n| \leq \sum_{j=n}^{n+k-1} |X_{j+1} - X_j| \leq \sum_{j=n}^{\infty} \varepsilon_j \xrightarrow{n\to\infty} 0.$$

By the completeness of $\mathbb{R}$ and Proposition 1.44, X_n converges almost surely.

□

Proposition 1.45. $X_n \xrightarrow{P} X$ if and only if $\mathbb{E}[\min\{1, |X_n - X|\}] \to 0$.

Proof. ($\Longrightarrow$) Suppose $X_n \xrightarrow{P} X$. Fix $\varepsilon > 0$.

$$\mathbb{E}[\min\{1, |X_n - X|\}] = \mathbb{E}[(\min\{1, |X_n - X|\})\chi_{\{|X_n - X| > \varepsilon\}}] + \mathbb{E}[(\min\{1, |X_n - X|\})\chi_{\{|X_n - X| \leq \varepsilon\}}]$$
$$\leq 1 \cdot \mathbb{P}(|X_n - X| > \varepsilon) + \varepsilon \cdot \mathbb{P}(|X_n - X| \leq \varepsilon).$$

Taking $n \to \infty$, the first term goes to 0 by convergence in probability. Since ε is arbitrary, the expectation goes to 0.
($\Longleftarrow$) By Markov's Inequality, for any $\varepsilon \in (0, 1)$:

$$\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(\min\{1, |X_n - X|\} > \varepsilon) \leq \frac{\mathbb{E}[\min\{1, |X_n - X|\}]}{\varepsilon}.$$

If the expectation tends to 0, then $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$ for all ε. □

Theorem 1.46. The following relations hold between the modes of convergence:

(i) If $X_n \xrightarrow{a.s.} X$, then $X_n \xrightarrow{P} X$.

(ii) If $X_n \xrightarrow{P} X$, then there exists a subsequence $\{X_{n_k}\}$ such that $X_{n_k} \xrightarrow{a.s.} X$.

(iii) If every subsequence of $\{X_n\}$ has a further subsequence converging to X a.s., then $X_n \xrightarrow{P} X$.

Proof. Omitted. □

Proposition 1.47. If $X_n \xrightarrow{P} X$ and f is continuous, then $f(X_n) \xrightarrow{P} f(X)$.

Proof. Let $\{n_i\}$ be an arbitrary subsequence of indices. Since $X_n \xrightarrow{P} X$, the subsequence $\{X_{n_i}\}$ also converges in probability to X. By Theorem 1.46 (ii), there exists a further sub-subsequence $\{n_{i_j}\}$ such that $X_{n_{i_j}} \xrightarrow{a.s.} X$. Because f is continuous, the Continuous Mapping Theorem for almost sure convergence implies that $f(X_{n_{i_j}}) \xrightarrow{a.s.} f(X)$. Thus, every subsequence of $\{f(X_n)\}$ contains a further sub-subsequence that converges almost surely to $f(X)$. By the characterization in Theorem 1.46 (iii), it follows that $f(X_n) \xrightarrow{P} f(X)$. □

While almost sure convergence and convergence in probability describe the behaviour of the random variables as functions, L^p convergence describes the behaviour of their moments.

Definition 1.48. For $p \in [1, \infty)$, we say that a sequence of random variables $\{X_n\}$ **converges in** L^p to X, denoted $X_n \xrightarrow{L^p} X$, if $X_n, X \in L^p$ (i.e., $\mathbb{E}|X_n|^p < \infty$) and $\lim_{n\to\infty} \mathbb{E}[|X_n - X|^p] = 0$.

Proposition 1.49. $X_n \xrightarrow{L^p} X \implies X_n \xrightarrow{P} X$.

Proof. This follows immediately from Markov's inequality: $\mathbb{P}(|X_n - X| > \varepsilon) \leq \varepsilon^{-p}\mathbb{E}|X_n - X|^p$. □

Definition 1.50. A collection of random variables $\{X_t : t \in T\}$ is **uniformly integrable** (UI) if

$$\lim_{b\to\infty} \sup_{t\in T} \mathbb{E}\left[\chi_{\{|X_t|>b\}}|X_t|\right] = 0.$$

Theorem 1.51. *(Vitali Convergence Theorem)* Let $X_n, X \in L^1$. Then $X_n \xrightarrow{L^1} X$ if and only if $X_n \xrightarrow{P} X$ and $\{X_n\}$ is uniformly integrable.

Proof. Omitted. □

Definition 1.52. Let μ, $\{\mu_n\}_{n\in\mathbb{N}}$ be probability measures on $(\mathcal{X}, \mathcal{M})$. We say that μ_n **converges weakly** to μ, denoted $\mu_n \xrightarrow{w} \mu$, if for every bounded, continuous function $f : \mathcal{X} \to \mathbb{R}$:

$$\int f \, \mathrm{d}\mu_n \to \int f \, \mathrm{d}\mu.$$

If $X_n \sim \mu_n$ and $X \sim \mu$, we say X_n **converges in distribution** to X, denoted $X_n \xrightarrow{d} X$.

Note 1.53. The definition of weak convergence bridges probability theory and functional analysis.

1. **The Dual Pairing:** Let $V = C_0(\mathcal{X})$ be the Banach space of continuous functions on a normed space $\mathcal{X}$ vanishing at infinity equipped with the supremum norm $\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)|$. We define, for each measure μ, an associated functional $T : V \to \mathbb{R}$ via the integral:

   $$T_\mu f := \int_{\mathcal{X}} f(x)\mu(\mathrm{d}x) = \mathbb{E}[f(X)].$$

   In this context, the measure μ acts as a continuous linear functional on the space of functions.

2. **Radon Measures:** A Borel measure μ is called a **Radon measure** if it is *inner regular*, meaning the measure of any Borel set can be approximated from within by compact sets: $\mu(B) = \sup\{\mu(K) : K \subset B, K \text{ compact}\}$. On most spaces of interest, every probability measure is a Radon measure.

3. **Weak vs. Weak* Convergence:** The *Riesz-Markov-Kakutani Representation Theorem* identifies the dual space V^* of $C_0(\mathcal{X})$ as the space of finite signed Radon measures. In functional analysis we have the following definitions:

   - **Weak convergence** of $\mu_n \to \mu$ would require $\Lambda(\mu_n) \to \Lambda(\mu)$ for all $\Lambda \in V^{**}$.
   - **Weak* convergence** of $\mu_n \to \mu$ requires $T_{\mu_n} f \to T_\mu f$ for all $f \in V$.

   Thus, what probabilists call "weak convergence" is technically weak* convergence on the space of measures. The term "weak" is used in probability I think because convergence in distribution is the least restrictive mode in the hierarchy of convergence.

Proposition 1.54. $X_n \xrightarrow{P} X \implies X_n \xrightarrow{d} X$.

Proof. Let f be bounded and continuous. Since $X_n \xrightarrow{P} X$, by the Continuous Mapping Theorem, $f(X_n) \xrightarrow{P} f(X)$. Since f is bounded, the Dominated Convergence Theorem implies $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$. □

Theorem 1.55. *(Portmanteau)* The following are equivalent for probability measures μ_n, μ:

(i) $\mu_n \xrightarrow{w} \mu$.

(ii) $\limsup \mu_n(F) \leq \mu(F)$ for every closed set $F \subset \mathcal{X}$.

(iii) $\liminf \mu_n(G) \geq \mu(G)$ for every open set $G \subset \mathcal{X}$.

(iv) $\lim \mu_n(A) = \mu(A)$ for every Borel set A such that $\mu(\partial A) = 0$ (the μ-continuity sets).

Proof. Omitted. □

Proposition 1.56. *(Uniqueness of Weak Limits)* If $\mu_n \xrightarrow{w} \mu$ and $\mu_n \xrightarrow{w} \nu$, then $\mu = \nu$.

Proof. Firstly, $\int f \, \mathrm{d}\mu = \int f \, \mathrm{d}\nu$ for all bounded continuous f. Taking $\mu_n = \mu$ as a constant sequence and applying Portmanteau, we have $\mu(G) \leq \nu(G)$ for all open G. By symmetry, $\mu(G) = \nu(G)$ for all open sets. Thus, by outer-regularity $\mu(A) = \nu(A)$ for all $A \in \mathcal{B}(\mathbb{R})$. □

Theorem 1.57. *(Skorokhod Representation Theorem)* Let $X_n \xrightarrow{d} X$. There exists a probability space $(\Omega', \mathcal{H}', \mathbb{P}')$ and random variables Y_n, Y on it such that $Y_n \overset{d}{=} X_n$, $Y \overset{d}{=} X$, and $Y_n \xrightarrow{a.s.} Y$.

Proof. Omitted. □

Proposition 1.58. *(Continuous Mapping Theorem)* Let $h : \mathcal{X} \to \mathcal{Y}$ be an a.s. continuous measurable function. If $X_n \xrightarrow{d} X$, then $h(X_n) \xrightarrow{d} h(X)$.

Proof. There exists Skorokhod representations $\{Y_n\}_{n\in\mathbb{N}}$ and Y such that $Y_n \overset{d}{=} X_n$, $Y \overset{d}{=} X$, and $Y_n \xrightarrow{a.s.} Y$. Then by the Continuous Mapping Theorem, $f(Y_n) \xrightarrow{a.s.} f(Y)$. Finally, one can show that $f(Y_n) \overset{d}{=} f(X_n)$ and $f(Y) \overset{d}{=} f(X)$. So another use of the Skorokhod representation theorem yields $f(X_n) \xrightarrow{d} f(X)$. □

Theorem 1.59. *(Lévy Continuity Theorem)* Let $\phi_n(r) = \mathbb{E}[e^{irX_n}]$ be the characteristic functions. Then $X_n \xrightarrow{d} X$ if and only if $\phi_n(r) \to \phi(r)$ pointwise, where ϕ is continuous at 0.

Proof. Omitted. □

When dealing with sequences of random vectors $\{(X_{1,n}, \dots, X_{K,n})\}_{n\in\mathbb{N}}$ in $\mathbb{R}^K$, we distinguish between the convergence of individual coordinates (**marginal convergence**) and the convergence of the vector (**joint convergence**).

Definition 1.60. We say $X_n = (X_{1,n}, \dots, X_{K,n})$ **converges jointly** to $X = (X_1, \dots, X_K)$ in mode $\alpha \in \{a.s., P, L^p, d\}$ if the sequence converges in the product space $\mathbb{R}^K$ with respect to the Borel σ-algebra $\mathcal{B}(\mathbb{R}^K)$.

Proposition 1.61. Let $\alpha \in \{a.s., P, L^p\}$. Then $X_{k,n} \xrightarrow{\alpha} X_k$ for each $k = 1, \dots, K$ if and only if $(X_{1,n}, \dots, X_{K,n}) \xrightarrow{\alpha} (X_1, \dots, X_K)$.

Proof. ($\Longleftarrow$) follows from the Continuous Mapping Theorem using the coordinate projection

$$f_k(x_1, \dots, x_k) = x_k.$$

($\Longrightarrow$) follows from the fact that the product metric $d(x, y) = (\sum |x_k - y_k|^2)^{1/2}$ is bounded by the sum of the coordinate distances. For instance, for $X_n \xrightarrow{P} X$, we use:

$$\mathbb{P}\left(\sqrt{\sum |X_{k,n} - X_k|^2} > \varepsilon\right) \leq \mathbb{P}\left(\sum |X_{k,n} - X_k| > \varepsilon\right) \leq \sum_{k=1}^{K} \mathbb{P}\left(|X_{k,n} - X_k| > \frac{\varepsilon}{K}\right).$$

Since each term in the finite sum vanishes as $n \to \infty$, the joint probability vanishes. □

Note 1.62. Marginal convergence in distribution does **not** imply joint convergence in distribution:

$$X_{k,n} \xrightarrow{d} X_k \quad \not\Longrightarrow \quad (X_{1,n}, \dots, X_{K,n}) \xrightarrow{d} (X_1, \dots, X_K).$$

Theorem 1.63. *(Cramér-Wold Device)* A sequence of random vectors $X_n \in \mathbb{R}^K$ converges in distribution to X if and only if every linear combination of their components converges in distribution:

$$X_n \xrightarrow{d} X \iff t^\top X_n \xrightarrow{d} t^\top X \quad \forall t \in \mathbb{R}^K.$$

Proof. Omitted. □

Lemma 1.64. *(Slutsky's Theorem)* Let X_n, X and Y_n be random variables. If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{P} c$ for some constant $c \in \mathbb{R}$, then the joint convergence is "rescued": $(X_n, Y_n) \xrightarrow{d} (X, c)$. Consequently:

(i) $X_n + Y_n \xrightarrow{d} X + c$

(ii) $X_n Y_n \xrightarrow{d} cX$

(iii) $X_n / Y_n \xrightarrow{d} X/c$, provided $c \neq 0$.

Proof. First, one shows $(X_n, Y_n) \xrightarrow{d} (X, c)$. This is usually done by showing the joint characteristic function $\mathbb{E}[e^{i(r_1 X_n + r_2 Y_n)}]$ converges to $\mathbb{E}[e^{ir_1 X}]e^{ir_2 c}$. Once joint convergence is established, (i)-(iii) follow by applying the Continuous Mapping Theorem to the functions $g(x, y) = x + y$, xy, and x/y respectively. □

2 Conditioning and Independence

2.1 Independence

Despite my stubborn avoidance up to this point; in order to efficiently extend properties from simple sets to σ-algebras without verifying additivity on algebras at every step, it is useful to introduce the π-λ (or Dynkin system) framework.

Definition 2.1. Let Ω be a set.

1. A collection $\mathcal{P}$ of subsets of Ω is a **π-system** if it is closed under finite intersections:
$$A, B \in \mathcal{P} \implies A \cap B \in \mathcal{P}.$$
2. A collection $\mathcal{L}$ of subsets of Ω is a **λ-system** if:
   (i) $\Omega \in \mathcal{L}$.
   (ii) If $A \in \mathcal{L}$, then $A^c \in \mathcal{L}$.
   (iii) If $\{A_n\}_{n=1}^{\infty}$ is a sequence of disjoint sets in $\mathcal{L}$, then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{L}$.

Note 2.2. A σ-algebra is simply a collection that is both a π-system and a λ-system. The λ-system axioms are essentially the "bare minimum" requirements for a collection to behave consistently with a probability measure (which is defined via disjoint unions).

Theorem 2.3. If $\mathcal{P}$ is a π-system and $\mathcal{L}$ is a λ-system such that $\mathcal{P} \subset \mathcal{L}$, then $\sigma(\mathcal{P}) \subset \mathcal{L}$.

Proof. Let $\mathcal{L}(\mathcal{P})$ be the intersection of all λ-systems containing $\mathcal{P}$. This is the smallest λ-system containing $\mathcal{P}$. Since every σ-algebra is a λ-system, $\mathcal{L}(\mathcal{P}) \subset \sigma(\mathcal{P})$. To prove the theorem, it suffices to show that $\mathcal{L}(\mathcal{P})$ is a π-system; if it is, it must be a σ-algebra, implying $\sigma(\mathcal{P}) \subset \mathcal{L}(\mathcal{P}) \subset \mathcal{L}$.
For any $C \in \mathcal{L}(\mathcal{P})$, define the collection $\mathcal{L}_C = \{B \subset \Omega : B \cap C \in \mathcal{L}(\mathcal{P})\}$. We first verify that $\mathcal{L}_C$ is a λ-system:

1. $\Omega \in \mathcal{L}_C$ because $\Omega \cap C = C \in \mathcal{L}(\mathcal{P})$.
2. If $B \in \mathcal{L}_C$, then $B^c \cap C = C \setminus (B \cap C)$. Since $\mathcal{L}(\mathcal{P})$ is a λ-system and $(B \cap C) \subset C$ (both in $\mathcal{L}(\mathcal{P})$), their difference is in $\mathcal{L}(\mathcal{P})$. Thus $B^c \in \mathcal{L}_C$.
3. Closure under disjoint unions follows from: $(\bigcup B_n) \cap C = \bigcup (B_n \cap C)$.

Now, fix $A \in \mathcal{P}$. For any $B \in \mathcal{P}$, $A \cap B \in \mathcal{P} \subset \mathcal{L}(\mathcal{P})$ (since $\mathcal{P}$ is a π-system). Thus, $\mathcal{P} \subset \mathcal{L}_A$. Since $\mathcal{L}_A$ is a λ-system, the minimality of $\mathcal{L}(\mathcal{P})$ implies $\mathcal{L}(\mathcal{P}) \subset \mathcal{L}_A$.
This means for any $A \in \mathcal{P}$ and $C \in \mathcal{L}(\mathcal{P})$, we have $A \cap C \in \mathcal{L}(\mathcal{P})$. This is equivalent to saying $A \in \mathcal{L}_C$ for all $A \in \mathcal{P}$, which implies $\mathcal{P} \subset \mathcal{L}_C$.
Again, by minimality, $\mathcal{L}(\mathcal{P}) \subset \mathcal{L}_C$ for any $C \in \mathcal{L}(\mathcal{P})$. Thus, for any $B, C \in \mathcal{L}(\mathcal{P})$, we have $B \cap C \in \mathcal{L}(\mathcal{P})$, proving that $\mathcal{L}(\mathcal{P})$ is a π-system. □

Independence is a statement about the factorization of information. We define it for σ-algebras first, as they represent the "available information."

Definition 2.4. A finite collection of sub-σ-algebras $\{\mathcal{F}_1, \ldots, \mathcal{F}_n\}$ is **independent** if for all bounded, $\mathcal{F}_i$-measurable random variables V_i:
$$\mathbb{E}\left[\prod_{i=1}^{n} V_i\right] = \prod_{i=1}^{n} \mathbb{E}[V_i].$$
An arbitrarily indexed collection $\{\mathcal{F}_t\}_{t \in T}$ is independent if every finite sub-collection is independent.

Using the π-λ theorem, we can reduce the requirement of checking all measurable functions to simply checking the "building blocks" of the σ-algebras.

Proposition 2.5. Let $\mathcal{P}_1, \dots, \mathcal{P}_n$ be π-systems. If for all $A_i \in \mathcal{P}_i$:

$$\mathbb{P}\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} \mathbb{P}(A_i),$$

then the σ-algebras $\sigma(\mathcal{P}_1), \dots, \sigma(\mathcal{P}_n)$ are independent.

Proof. We use the π-λ theorem to "lift" the property from sets to σ-algebras one by one. Fix $A_2 \in \mathcal{P}_2, \dots, A_n \in \mathcal{P}_n$. Define the collection:

$$\mathcal{L} = \{A \in \mathcal{H} : \mathbb{P}(A \cap A_2 \cap \cdots \cap A_n) = \mathbb{P}(A)\mathbb{P}(A_2)\dots\mathbb{P}(A_n)\}.$$

One can easily check that $\mathcal{L}$ is a λ-system. Since $\mathcal{P}_1 \subset \mathcal{L}$ and $\mathcal{P}_1$ is a π-system, $\sigma(\mathcal{P}_1) \subset \mathcal{L}$. Now, we can fix $A_1 \in \sigma(\mathcal{P}_1), A_3 \in \mathcal{P}_3, \dots$ and repeat the argument for $\mathcal{P}_2$. After n iterations, the product rule holds for all $A_i \in \sigma(\mathcal{P}_i)$. The extension to expectations of measurable functions follows via the standard machinery (simple functions $\to$ Monotone Convergence). $\square$

Proposition 2.6. Let $\{\mathcal{F}_t : t \in T\}$ be independent sub-σ-algebras.

(i) **Monotonicity:** If $\mathcal{G}_t \subset \mathcal{F}_t$ is a sub-σ-algebra for each t, then $\{\mathcal{G}_t : t \in T\}$ is independent.

(ii) **Grouping:** Let $\{T_s : s \in S\}$ be a partition of T. Define $\mathcal{G}_s = \sigma\left(\bigcup_{t \in T_s} \mathcal{F}_t\right)$. Then $\{\mathcal{G}_s : s \in S\}$ is independent.

(iii) **Sequential Independence:** The sequence $\mathcal{F}_1, \mathcal{F}_2, \dots$ is independent if and only if $\{\mathcal{F}_1, \dots, \mathcal{F}_n\}$ and $\mathcal{F}_{n+1}$ are independent for all $n \geq 1$.

Proof.

(i) Trivial: Any $\mathcal{G}_t$-measurable function is $\mathcal{F}_t$-measurable.

(ii) For each $s \in S$, define the collection of finite intersections:

$$\mathcal{P}_s = \left\{ \bigcap_{j=1}^{k} B_{t_j} : t_j \in T_s, B_{t_j} \in \mathcal{F}_{t_j} \right\}.$$

Each $\mathcal{P}_s$ is a π-system that generates $\mathcal{G}_s$. If we pick $A_{s_1} \in \mathcal{P}_{s_1}, \dots, A_{s_k} \in \mathcal{P}_{s_k}$, the intersection $\bigcap A_{s_i}$ consists of sets from disjoint index sets T_{s_i}. By the original independence of $\{\mathcal{F}_t\}$, the probability of this intersection factors. By Proposition 2.5, the σ-algebras $\{\mathcal{G}_s\}$ are independent.

(iii) ($\implies$) This follows from (ii) by partitioning T into $T_1 = \{1, \dots, n\}$ and $T_2 = \{n+1, n+2, \dots\}$.

($\impliedby$) We proceed by induction. The base case $n = 1$ is trivial. Assume $\{\mathcal{F}_1, \dots, \mathcal{F}_n\}$ is independent. For any positive measurable $V_i \in \mathcal{F}_i$:

$$\mathbb{E}\left[\left(\prod_{i=1}^{n} V_i\right) V_{n+1}\right] = \mathbb{E}\left[\prod_{i=1}^{n} V_i\right] \mathbb{E}[V_{n+1}],$$

where we used the assumption that the "past" $\sigma(\mathcal{F}_1, \dots, \mathcal{F}_n)$ and the "current" $\mathcal{F}_{n+1}$ are independent. Applying the inductive hypothesis to the first term yields the full factorization.

□

Note 2.7. The Grouping Property (ii) matches our intuition: if you have several groups of people, and everyone in the room is independent of everyone else, then any "summary" of Group A (like the average height) must be independent of any "summary" of Group B.

Definition 2.8. Random objects $\{X_t : t \in T\}$ are **independent** if the generated σ-algebras $\{\sigma(X_t) : t \in T\}$ are independent.

Proposition 2.9. Real-valued random variables $X_1, \dots, X_n$ are independent if and only if any of the following hold:

(i) For all $x_i \in \mathbb{R}$, $\mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n) = \prod \mathbb{P}(X_i \leq x_i)$.

(ii) The joint distribution μ_X is the product of the marginals: $\mu_X = \mu_{X_1} \otimes \cdots \otimes \mu_{X_n}$.

(iii) $\phi_{\mathbf{X}}(\mathbf{t}) = \prod_{j=1}^n \phi_{X_j}(t_j)$ for all $\mathbf{t} \in \mathbb{R}^n$. Where, recall, the characteristic function $\phi_X : \mathbb{R} \to \mathbb{C}$ for a real-valued random variable is $\phi_X(t) = \mathbb{E}[e^{itX}]$ and is $\phi_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}[e^{i\langle \mathbf{t}, \mathbf{X}\rangle}]$ where $\mathbf{t} \in \mathbb{R}^n$ for a real-valued random vector $\mathbf{X} = (X_1, \dots, X_n)$.

Proof. We will proceed by showing Def $\iff$ (ii), then (ii) $\iff$ (i), and finally (ii) $\iff$ (iii). Let μ_X be the joint distribution (law) of $\mathbf{X}$ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, and let $\nu = \mu_{X_1} \otimes \cdots \otimes \mu_{X_n}$ be the product measure of the marginals.

Def $\iff$ (ii). By definition, $X_1, \dots, X_n$ are independent if $\sigma(X_1), \dots, \sigma(X_n)$ are independent σ-algebras. ($\implies$) If they are independent, then for any Borel sets $B_1, \dots, B_n \in \mathcal{B}(\mathbb{R})$:

$$\mu_X(B_1 \times \cdots \times B_n) = \mathbb{P}(X_1 \in B_1, \dots, X_n \in B_n) = \prod_{i=1}^n \mathbb{P}(X_i \in B_i) = \prod_{i=1}^n \mu_{X_i}(B_i).$$

This shows that μ_X agrees with the product measure ν on all measurable rectangles. Since measurable rectangles form a π-system that generates $\mathcal{B}(\mathbb{R}^n)$, the uniqueness of measure extension (via the π-λ Theorem) implies $\mu_X = \nu$. ($\impliedby$) If $\mu_X = \nu$, then the probability of any rectangle factors. By the π-system criterion (Prop 2.5), the generated σ-algebras are independent.

(ii) $\iff$ (i). ($\implies$) This is trivial. If $\mu_X = \nu$, we simply evaluate the measures on the specific rectangles $B_i = (-\infty, x_i]$. ($\impliedby$) Let $\mathcal{P}$ be the collection of semi-infinite rectangles of the form $(-\infty, x_1] \times \cdots \times (-\infty, x_n]$. This is a π-system that generates $\mathcal{B}(\mathbb{R}^n)$. If (i) holds, then μ_X and ν agree on $\mathcal{P}$. By the π-λ Theorem, $\mu_X = \nu$.

(ii) $\iff$ (iii). ($\implies$) Suppose $\mu_X = \nu$. By the change of variables formula and Fubini's Theorem:

$$\begin{aligned}\phi_{\mathbf{X}}(\mathbf{t}) = \int_{\mathbb{R}^n} e^{i\langle \mathbf{t}, \mathbf{x}\rangle}\, d\mu_X(\mathbf{x}) &= \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \prod_{j=1}^n e^{it_j x_j}\, d\mu_{X_1}(x_1) \dots d\mu_{X_n}(x_n) \\ &= \prod_{j=1}^n \int_{\mathbb{R}} e^{it_j x_j}\, d\mu_{X_j}(x_j) = \prod_{j=1}^n \phi_{X_j}(t_j).\end{aligned}$$

($\impliedby$) The characteristic function $\phi_{\mathbf{X}}(\mathbf{t})$ is the Fourier-Stieltjes transform of the measure μ_X. The product $\prod \phi_{X_j}(t_j)$ is the Fourier-Stieltjes transform of the product measure ν (as shown in the forward direction). By the Uniqueness Theorem for Fourier Transforms, two finite Borel measures with identical transforms are equal. Thus $\mu_X = \nu$. □

Definition 2.10. Given independent σ-algebras $\mathcal{G}_1, \mathcal{G}_2, \dots$, the **tail σ-algebra** is:

$$\mathcal{T} = \bigcap_{n=1}^{\infty} \sigma(\mathcal{G}_{n+1}, \mathcal{G}_{n+2}, \dots).$$

Theorem 2.11 (Kolmogorov's 0-1 Law)**.** If $A \in \mathcal{T}$, then $\mathbb{P}(A) \in \{0, 1\}$.

Proof. Let $\mathcal{F}_n = \sigma(\mathcal{G}_1, \dots, \mathcal{G}_n)$ and $\mathcal{T}_n = \sigma(\mathcal{G}_{n+1}, \dots)$.

1. By Grouping (Prop 2.6 ii), $\mathcal{F}_n$ and $\mathcal{T}_n$ are independent.
2. Since $A \in \mathcal{T} \subset \mathcal{T}_n$, A is independent of $\mathcal{F}_n$ for every n.
3. Let $\mathcal{P} = \bigcup_n \mathcal{F}_n$. This is a π-system (an algebra) that generates $\mathcal{F}_\infty = \sigma(\mathcal{G}_1, \mathcal{G}_2, \dots)$. Since A is independent of every $\mathcal{F}_n$, it is independent of $\mathcal{P}$.
4. By the π-system criterion, A is independent of $\sigma(\mathcal{P}) = \mathcal{F}_\infty$.
5. However, $A \in \mathcal{T} \subset \mathcal{F}_\infty$. This means A is independent of itself:

$$\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)\mathbb{P}(A) = \mathbb{P}(A)^2.$$

The only solutions to $x = x^2$ are 0 and 1. □

Example 2.12. Let $X_1, X_2, \dots$ be independent random variables and $S_n = \sum_{i=1}^n X_i$.

- The event $\{\lim_{n\to\infty} \frac{S_n}{n} \text{ exists and is } 0\}$ is a tail event. Thus, the Strong Law of Large Numbers can only conclude that this probability is either 0 or 1.
- The event $\{\limsup_{n\to\infty} S_n = \infty\}$ is a tail event.
- The event $\{\sup_n S_n > 0\}$ is **not** a tail event, as it depends on the values of the first few X_i.

2.2 Conditional Expectations

Example 2.13. Let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space and let X be $\overline{\mathbb{R}}_+$-valued random variable. Recall that $\mathbb{E}[X] = \int X(\omega)\, \mathrm{d}\mathbb{P}(\omega)$ is the expected value of X, and is a reasonable "best guess" for the value of $X(\omega)$ without any information on ω.
Let's assume slightly more information; namely that ω lies in H for some $H \in \mathcal{H}$ with $\mathbb{P}(H) \neq 0$. In this case our "best guess" for $X(\omega)$ is now

$$\frac{1}{\mathbb{P}(H)} \int \chi_H(\omega) X(\omega)\, \mathrm{d}\mathbb{P}(\omega) = \frac{\mathbb{E}\left[X\chi_H\right]}{\mathbb{P}(H)}.$$

And in the case $\mathbb{P}(H) = 0$, we can just define this quotient to be 0.
Let's further increase the information; now we allow ourselves to know whether ω lies in H or not. But we do not know which in advance. In that case, our "best guess" is a function of ω that depends on whether $\omega \in H$ or $\omega \in H^c$:

$$\bar{X}(\omega) := \left(\frac{\mathbb{E}\left[X\chi_H\right]}{\mathbb{P}(H)}\right) \chi_H(\omega) + \left(\frac{\mathbb{E}\left[X\chi_{H^c}\right]}{\mathbb{P}(H^c)}\right) \chi_{H^c}(\omega).$$

Now we make three observations:

(i) $\bar{X}$ is measurable with respect to $\mathcal{F} = \{\emptyset, H, H^c, \Omega\}$; the σ-algebra generated $\{H\}$.

(ii) Any $\mathcal{F}$-measurable random variable Y takes the form

$$Y(\omega) = a\chi_H(\omega) + b\chi_{H^c}(\omega).$$

And so for any $\mathcal{F}$-measurable random variable Y:

$$\begin{aligned}\mathbb{E}[XY] &= a\mathbb{E}[X\chi_H] + b\mathbb{E}[X\chi_{H^c}] = a\mathbb{P}(H)\frac{\mathbb{E}[X\chi_H]}{\mathbb{P}(H)} + b\mathbb{P}(H^c)\frac{\mathbb{E}[X\chi_{H^c}]}{\mathbb{P}(H^c)} \\ &= \mathbb{E}\left[a\chi_H \frac{\mathbb{E}[X\chi_H]}{\mathbb{P}(H)} + b\chi_{H^c}\frac{\mathbb{E}[X\chi_{H^c}]}{\mathbb{P}(H^c)}\right] = \mathbb{E}\left[\bar{X}Y\right].\end{aligned}$$

These properties uniquely determine the "conditional expectation" up to almost sure equivalence.

Definition 2.14. Let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space and let $\mathcal{F}$ be a sub σ-algebra of $\mathcal{H}$. Let X be a non-negative real-valued random variable. The **conditional expectation** of X given $\mathcal{F}$, denoted $\mathbb{E}[X|\mathcal{F}]$ is any random variable $\bar{X}$ that satisfies

(i) $\bar{X}$ is $\mathcal{F}$-measurable.

(ii) For every non-negative $\mathcal{F}$-measurable V, $\mathbb{E}[VX] = \mathbb{E}[V\bar{X}]$.

We call $\bar{X}$ a **version** of $\mathbb{E}[X|\mathcal{F}]$. For an arbitrary random variable X, if $\mathbb{E}X$ exists and at least one of the below terms is finite, we define the conditional expectation

$$\mathbb{E}[X|\mathcal{F}] = \mathbb{E}[X^+|\mathcal{F}] - \mathbb{E}[X^-|\mathcal{F}],$$

where $X^+(\omega) = \max\{X(\omega), 0\}$ and $X^-(\omega) = \max\{-X(\omega), 0\}$.

Proposition 2.15. Condition (ii) in definition 2.14 is equivalent to $\mathbb{E}\chi_H X = \mathbb{E}\chi_H \bar{X}$ for all $H \in \mathcal{F}$

Proof. Suppose $\bar{X}$ satisfies condition (ii). Then since $\mathbb{E}[VX] = \mathbb{E}[V\bar{X}]$ holds for any non-negative $\mathcal{F}$-measurable V, it holds in particular for $V = \chi_H$ for any $H \in \mathcal{F}$. For the converse, suppose $\mathbb{E}[\chi_H X] = \mathbb{E}[\chi_H \bar{X}]$ for every $H \in \mathcal{F}$. By linearity, the same equality holds when V is any non-negative $\mathcal{F}$-measurable simple function. The monotone convergence theorem then extends this to any non-negative $\mathcal{F}$-measurable V. □

Proposition 2.16. If X is integrable, so is $\mathbb{E}[X|\mathcal{F}]$.

Proof. We must show $\mathbb{E}[|\mathbb{E}[X|\mathcal{F}]|] < \infty$. By the triangle inequality it suffices to show

$$\mathbb{E}[\mathbb{E}[X^+|\mathcal{F}]] < \infty \quad \text{and} \quad \mathbb{E}[\mathbb{E}[X^-|\mathcal{F}]] < \infty.$$

Let $\bar{X}^+$ and $\bar{X}^-$ be versions of $\mathbb{E}[X^+|\mathcal{F}]$ and $\mathbb{E}[X^-|\mathcal{F}]$ respectively. Applying condition (ii) of Definition 2.14 with $V = 1$ identically gives

$$\mathbb{E}[\bar{X}^+] = \mathbb{E}[X^+] < \infty, \qquad \mathbb{E}[\bar{X}^-] = \mathbb{E}[X^-] < \infty,$$

where finiteness follows from integrability of X. □

Theorem 2.17. Let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space, let $\mathcal{F}$ be a sub-σ-algebra of $\mathcal{H}$, and let X be a random variable s.t. $\mathbb{E}[X]$ exists. Then $\mathbb{E}[X|\mathcal{F}]$ exists and is unique up to almost sure equivalence.

Proof. We first prove existence and uniqueness for non-negative X, then extend to the general case.
Uniqueness. Suppose $X \geq 0$ and $\bar{X}$, $\bar{X}'$ are both versions of $\mathbb{E}[X|\mathcal{F}]$. Then $\bar{X}, \bar{X}', X \geq 0$ are all $\mathcal{F}$-measurable (or $\mathcal{H}$-measurable), so by repeated application of condition (ii),

$$\mathbb{E}[\bar{X}^2] = \mathbb{E}[\bar{X} \cdot \bar{X}] = \mathbb{E}[\bar{X}' \cdot X] = \mathbb{E}[\bar{X}' \cdot \bar{X}'] = \mathbb{E}[(\bar{X}')^2].$$

Therefore

$$\mathbb{E}\big[(\bar{X} - \bar{X}')^2\big] = \mathbb{E}[\bar{X}^2] - 2\mathbb{E}[\bar{X}\bar{X}'] + \mathbb{E}[(\bar{X}')^2] = 0,$$

which implies $\bar{X} = \bar{X}'$ almost surely.
Existence Suppose $X \geq 0$. Define two measures on $(\Omega, \mathcal{F})$ by

$$R(H) = \mathbb{P}(H), \qquad Q(H) = \int_H X(\omega)\ \mathrm{d}\mathbb{P}(\omega), \qquad H \in \mathcal{F}.$$

Then R is a finite (hence σ-finite) measure on $\mathcal{F}$, and $Q \ll R$: if $R(H) = 0$ then $Q(H) = \int_H X\ \mathrm{d}\mathbb{P} = 0$. By the Radon–Nikodym theorem, there exists a non-negative $\mathcal{F}$-measurable function $\bar{X}$ such that for all $H \in \mathcal{F}$,

$$Q(H) = \int_H \bar{X}(\omega)\ \mathrm{d}R(\omega) = \int_H \bar{X}(\omega)\ \mathrm{d}\mathbb{P}(\omega).$$

That is, $\mathbb{E}[\chi_H X] = \mathbb{E}[\chi_H \bar{X}]$ for all $H \in \mathcal{F}$. By the preceding proposition, this is equivalent to condition (ii) of Definition 2.14, so $\bar{X}$ is a version of $\mathbb{E}[X|\mathcal{F}]$.
General case. For arbitrary X with $\mathbb{E}[X]$ existing, write $X = X^+ - X^-$. Since X^+ and X^- are both non-negative, the above gives the existence of versions $\bar{X}^+$ and $\bar{X}^-$ of $\mathbb{E}[X^+|\mathcal{F}]$ and $\mathbb{E}[X^-|\mathcal{F}]$ respectively. Setting $\bar{X} = \bar{X}^+ - \bar{X}^-$ gives a version of $\mathbb{E}[X|\mathcal{F}]$ by definition. Uniqueness follows since if $\bar{X}, \bar{X}'$ are two versions of $\mathbb{E}[X|\mathcal{F}]$, then there exist versions $\bar{X}^+, (\bar{X}')^+$ of $\mathbb{E}[X^+|\mathcal{F}]$ and $\bar{X}^-, (\bar{X}')^-$ of $\mathbb{E}[X^-|\mathcal{F}]$ such that $\bar{X} = \bar{X}^+ - \bar{X}^-$ and $\bar{X}' = (\bar{X}')^+ - (\bar{X}')^-$. By uniqueness in the non-negative case, $\bar{X}^+ \overset{a.s.}{=} (\bar{X}')^+$ and $\bar{X}^- \overset{a.s.}{=} (\bar{X}')^-$, and hence $\bar{X} \overset{a.s.}{=} \bar{X}'$. □

Definition 2.18. Let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space, let X be a random variable, and let Y be a random variable taking values in a measurable space $(\mathcal{Y}, \mathcal{N})$. The **conditional expectation of X given Y** is defined to be $\mathbb{E}[X \mid \sigma(Y)]$, the conditional expectation of X given the σ-algebra generated by Y.

Proposition 2.19. Let X be a non-negative random variable and let Y be a random variable taking values in $(\mathcal{Y}, \mathcal{N})$. Then every version of $\mathbb{E}[X \mid Y]$ has the form $f \circ Y$ for some non-negative $\mathcal{N}$-measurable function f. Conversely, $f \circ Y$ is a version of $\mathbb{E}[X \mid Y]$ if and only if for every non-negative $\mathcal{N}$-measurable h,

$$\mathbb{E}[(f \circ Y)(h \circ Y)] = \mathbb{E}[X(h \circ Y)].$$

Proof. Since every version of $\mathbb{E}[X \mid Y]$ is $\sigma(Y)$-measurable, and since a random variable is $\sigma(Y)$-measurable if and only if it equals $f \circ Y$ for some $\mathcal{N}$-measurable $f \geq 0$ (by Doob–Dynkin), every version has the stated form. The characterization of when $f \circ Y$ is a version then follows directly from condition (ii) of Definition 2.14, since the $\sigma(Y)$-measurable non-negative random variables are precisely those of the form $h \circ Y$ for $\mathcal{N}$-measurable $h \geq 0$. □

Proposition 2.20 (Properties of Conditional Expectation)**.** Let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space, $\mathcal{F}$ a sub-σ-algebra of $\mathcal{H}$, and $a, b, c \in \mathbb{R}$. Assume all conditional expectations below exist. Then the following hold almost surely:

(i) **Monotonicity.** If $X \leq Y$, then $\mathbb{E}[X|\mathcal{F}] \leq \mathbb{E}[Y|\mathcal{F}]$.

(ii) **Linearity.** $\mathbb{E}[aX + bY + c \mid \mathcal{F}] = a\mathbb{E}[X|\mathcal{F}] + b\mathbb{E}[Y|\mathcal{F}] + c$.

(iii) **Insensitivity.** If $X \overset{a.s.}{=} Y$, then $\mathbb{E}[X|\mathcal{F}] \overset{a.s.}{=} \mathbb{E}[Y|\mathcal{F}]$.

(iv) **Monotone Convergence.** If $X_n \geq 0$ and $X_n \nearrow Y$, then $\mathbb{E}[X_n|\mathcal{F}] \nearrow \mathbb{E}[Y|\mathcal{F}]$.

(v) **Fatou's Lemma.** If $X_n \geq 0$, then $\mathbb{E}[\liminf_n X_n \mid \mathcal{F}] \leq \liminf_n \mathbb{E}[X_n|\mathcal{F}]$.

(vi) **D.C.T.** If $|X_n| \leq Y$ with $\mathbb{E}[Y] < \infty$ and $X_n \to X$ a.s., then $\mathbb{E}[X_n|\mathcal{F}] \to \mathbb{E}[X|\mathcal{F}]$.

(vii) **Jensen's Inequality.** If $\varphi : \mathbb{R} \to \mathbb{R}$ is convex, then $\varphi(\mathbb{E}[X|\mathcal{F}]) \leq \mathbb{E}[\varphi(X)|\mathcal{F}]$.

Theorem 2.21. Let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space and let $\mathcal{F}, \mathcal{G}$ be sub-σ-algebras of $\mathcal{H}$. Let W and X be random variables such that $\mathbb{E}[X]$ and $\mathbb{E}[WX]$ exist. Then:

(i) **Conditional determinism.** If W is $\mathcal{F}$-measurable, then

$$\mathbb{E}[WX \mid \mathcal{F}] = W\,\mathbb{E}[X|\mathcal{F}] \quad \text{a.s.}$$

(ii) **Tower property.** If $\mathcal{F} \subseteq \mathcal{G}$, then

$$\mathbb{E}\big[\mathbb{E}[X|\mathcal{G}] \,\big|\, \mathcal{F}\big] = \mathbb{E}\big[\mathbb{E}[X|\mathcal{F}] \,\big|\, \mathcal{G}\big] = \mathbb{E}[X|\mathcal{F}] \quad \text{a.s.}$$

Proof. We prove both parts for $W, X \geq 0$.
(i). Let $\bar{X}$ be a version of $\mathbb{E}[X|\mathcal{F}]$. We claim $W\bar{X}$ is a version of $\mathbb{E}[WX|\mathcal{F}]$. Since W and $\bar{X}$ are both $\mathcal{F}$-measurable, so is $W\bar{X}$. For any non-negative $\mathcal{F}$-measurable V, the product VW is also non-negative and $\mathcal{F}$-measurable, so by condition (ii) applied to $\bar{X}$,

$$\mathbb{E}[V \cdot WX] = \mathbb{E}[(VW)X] = \mathbb{E}[(VW)\bar{X}] = \mathbb{E}[V \cdot W\bar{X}].$$

Hence $W\bar{X} = W\mathbb{E}[X|\mathcal{F}]$ is a version of $\mathbb{E}[WX|\mathcal{F}]$.
(ii). Let $\bar{X} = \mathbb{E}[X|\mathcal{F}]$. Since $\bar{X}$ is $\mathcal{F}$-measurable and $\mathcal{F} \subseteq \mathcal{G}$, $\bar{X}$ is also $\mathcal{G}$-measurable. By conditional determinism applied with the constant $W = 1$ and $\mathcal{G}$ in place of $\mathcal{F}$,

$$\mathbb{E}\big[\mathbb{E}[X|\mathcal{F}] \,\big|\, \mathcal{G}\big] = \mathbb{E}[\bar{X} \mid \mathcal{G}] = \bar{X} \cdot \mathbb{E}[1|\mathcal{G}] = \bar{X} = \mathbb{E}[X|\mathcal{F}].$$

For the other equality, let $\tilde{X} = \mathbb{E}[X|\mathcal{G}]$. We verify that $\bar{X}$ is a version of $\mathbb{E}[\tilde{X}|\mathcal{F}]$ by checking the two conditions of Definition 2.14. First, $\bar{X}$ is $\mathcal{F}$-measurable by definition. Second, for any non-negative $\mathcal{F}$-measurable V, since $\mathcal{F} \subseteq \mathcal{G}$, V is also $\mathcal{G}$-measurable, so by condition (ii) applied to $\tilde{X}$,

$$\mathbb{E}[V\tilde{X}] = \mathbb{E}[VX],$$

and by condition (ii) applied to $\bar{X}$,

$$\mathbb{E}[V\bar{X}] = \mathbb{E}[VX].$$

Hence $\mathbb{E}[V\tilde{X}] = \mathbb{E}[V\bar{X}]$ for all non-negative $\mathcal{F}$-measurable V, so $\bar{X}$ is a version of $\mathbb{E}[\tilde{X}|\mathcal{F}]$, i.e.,

$$\mathbb{E}\big[\mathbb{E}[X|\mathcal{G}] \,\big|\, \mathcal{F}\big] = \mathbb{E}[X|\mathcal{F}].$$

□

2.3 Conditional Distributions

Recall that for any event $H \in \mathcal{H}$, we can write $\mathbb{P}(H) = \mathbb{E}[\chi_H]$. This suggests defining the conditional probability of H given $\mathcal{F}$ as the conditional expectation of χ_H. Collecting these conditional probabilities together for all $H \in \mathcal{H}$ would give us a conditional distribution, analogous to how $\mathbb{P}$ is determined by its values on events. However, there is a subtlety: conditional expectations are only defined up to almost sure equivalence, and the almost sure null set where countable additivity fails may depend on the sequence of events in question. So it is not automatic that we can find versions of $\mathbb{P}(H|\mathcal{F})$ for each $H \in \mathcal{H}$ that simultaneously form a valid measure for every $\omega \in \Omega$. When this is possible, we call the result a regular version of the conditional probability.

Definition 2.22. Let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space and let $\mathcal{F}$ be a sub-σ-algebra of $\mathcal{H}$. For each $H \in \mathcal{H}$, the **conditional probability** of H given $\mathcal{F}$ is the random variable

$$\mathbb{P}(H|\mathcal{F}) = \mathbb{E}[\chi_H|\mathcal{F}].$$

Definition 2.23. Let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space and let $\mathcal{F}$ be a sub-σ-algebra of $\mathcal{H}$. A **regular version** of the conditional probability $\mathbb{P}(\cdot|\mathcal{F})$ is a kernel $Q : \Omega \times \mathcal{H} \to [0,1]$ such that:

(i) For each $H \in \mathcal{H}$, $Q(\cdot, H)$ is a version of $\mathbb{P}(H|\mathcal{F})$.

(ii) For each $\omega \in \Omega$, $Q(\omega, \cdot)$ is a probability measure on $(\Omega, \mathcal{H})$.

Theorem 2.24. If $(\Omega, \mathcal{H})$ is a standard measurable space, then $\mathbb{P}(\cdot|\mathcal{F})$ has a regular version.

Regular versions give us a constructive way to compute conditional expectations as integrals, resolving the nonconstructive nature of Definition 2.14.

Proposition 2.25. If Q is a regular version of $\mathbb{P}(\cdot|\mathcal{F})$, then for every random variable X whose expectation exists,

$$\int Q(\omega, \mathrm{d}\omega')\, X(\omega')$$

is a version of $\mathbb{E}[X|\mathcal{F}]$.

Proof. We prove the result for $X \geq 0$; the general case follows by writing $X = X^+ - X^-$. Let $\bar{X}(\omega) = \int Q(\omega, \mathrm{d}\omega')X(\omega')$. Since Q is a kernel, $\bar{X}$ is $\mathcal{F}$-measurable. It remains to verify condition (ii) of Definition 2.14. For any $H \in \mathcal{H}$, since $Q(\cdot, H)$ is a version of $\mathbb{P}(H|\mathcal{F}) = \mathbb{E}[\chi_H|\mathcal{F}]$, and for any non-negative $\mathcal{F}$-measurable V, applying conditional determinism and the tower property gives

$$\mathbb{E}[V \cdot Q(\cdot, H)] = \mathbb{E}[V\, \mathbb{E}[\chi_H|\mathcal{F}]] = \mathbb{E}[\mathbb{E}[V\chi_H|\mathcal{F}]] = \mathbb{E}[V\chi_H].$$

By linearity and monotone convergence, this extends from indicators to all $X \geq 0$, giving $\mathbb{E}[V\bar{X}] = \mathbb{E}[VX]$ for all non-negative $\mathcal{F}$-measurable V. □

Definition 2.26. Let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space, $\mathcal{F}$ a sub-σ-algebra of $\mathcal{H}$, and Y a random variable taking values in $(\mathcal{Y}, \mathcal{N})$. The **conditional distribution** of Y given $\mathcal{F}$ is a transition kernel L from $(\Omega, \mathcal{F})$ to $(\mathcal{Y}, \mathcal{N})$ such that for all $\omega \in \Omega$ and $B \in \mathcal{N}$,

$$L(\omega, B) = \mathbb{P}(Y \in B|\mathcal{F})(\omega).$$

Theorem 2.27. If $(\mathcal{Y}, \mathcal{N})$ is a standard measurable space, then the conditional distribution of Y given $\mathcal{F}$ exists.

We now turn to the relationship between joint and conditional distributions. The following disintegration theorem says that any joint distribution on a product space can be factored into a marginal and a conditional, provided the second component lives in a standard measurable space.

Theorem 2.28. *(Disintegration)* Let π be a probability measure on $(\mathcal{X}\times\mathcal{Y}, \mathcal{M}\otimes\mathcal{N})$. If $(\mathcal{Y},\mathcal{N})$ is a standard measurable space, then there exists a probability measure μ on $(\mathcal{X},\mathcal{M})$ and a probability kernel K from $(\mathcal{X},\mathcal{M})$ to $(\mathcal{Y},\mathcal{N})$ such that

$$\pi(\mathrm{d}x, \mathrm{d}y) = \mu(\mathrm{d}x)\, K(x, \mathrm{d}y).$$

Proof. Let $\Omega = \mathcal{X}\times\mathcal{Y}$ with $\mathcal{H} = \mathcal{M}\otimes\mathcal{N}$, and define random variables $X(\omega) = x$ and $Y(\omega) = y$ for $\omega = (x, y) \in \Omega$. Since $(\mathcal{Y},\mathcal{N})$ is standard, by Theorem 2.27 there exists a version $L(\omega, B)$ of the conditional distribution of Y given $\sigma(X)$. Since $L(\omega, B)$ is $\sigma(X)$-measurable in ω, by the Doob–Dynkin lemma we can write $L(\omega, B) = K(X(\omega), B)$ for some $\mathcal{M}$-measurable kernel K. Let $\mu(A) = \pi(A\times\mathcal{Y})$ be the marginal distribution of X. Then for any $A \in \mathcal{M}$, $B \in \mathcal{N}$, applying conditional determinism and the tower property gives

$$\begin{aligned}
\pi(A\times B) &= \mathbb{E}[\chi_A(X)\chi_B(Y)] \\
&= \mathbb{E}\big[\mathbb{E}[\chi_A(X)\chi_B(Y)|\sigma(X)]\big] \\
&= \mathbb{E}\big[\chi_A(X)\,\mathbb{E}[\chi_B(Y)|\sigma(X)]\big] \\
&= \mathbb{E}[\chi_A(X)\, K(X, B)] \\
&= \int_A \mu(\mathrm{d}x)\, K(x, B).
\end{aligned}$$

Since the measurable rectangles form a π-system generating $\mathcal{M}\otimes\mathcal{N}$, and both sides are probability measures agreeing on this π-system, they agree on all of $\mathcal{M}\otimes\mathcal{N}$. □

Theorem 2.29. Let X and Y be random variables taking values in $(\mathcal{X},\mathcal{M})$ and $(\mathcal{Y},\mathcal{N})$ respectively, with joint distribution π. Suppose π can be written $\mu(\mathrm{d}x)K(x,\mathrm{d}y)$, i.e., for any $A \in \mathcal{M}$, $B \in \mathcal{N}$,

$$\pi(A\times B) = \int_A \mu(\mathrm{d}x)\, K(x, B).$$

Then μ is the marginal distribution of X, the kernel L defined by $L(\omega, B) = K(X(\omega), B)$ is a version of the conditional distribution of Y given $\sigma(X)$, and for all non-negative $(\mathcal{M}\otimes\mathcal{N})$-measurable f,

$$\mathbb{E}[f(X,Y)|\sigma(X)] = \int K(X, \mathrm{d}y)\, f(X, y) \quad \text{a.s.}$$

Proof. That μ is the marginal distribution of X follows immediately since $\pi(A\times\mathcal{Y}) = \int_A \mu(\mathrm{d}x)K(x,\mathcal{Y}) = \mu(A)$. To see that $L(\omega, B) = K(X(\omega), B)$ is a version of the conditional distribution of Y given $\sigma(X)$, we verify that for any $A \in \mathcal{M}$ and $B \in \mathcal{N}$,

$$\mathbb{E}[\chi_A(X)\, K(X, B)] = \int_A \mu(\mathrm{d}x)K(x, B) = \pi(A\times B) = \mathbb{E}[\chi_A(X)\chi_B(Y)],$$

which by the characterization in Proposition 2.25 confirms $K(X(\omega), B)$ is a version of $\mathbb{P}(Y \in B|\sigma(X))$. The formula for conditional expectations follows from Proposition 2.25 applied with the regular version L. □

Corollary 2.30. Suppose the joint distribution π of X and Y has the representation $\pi(\mathrm{d}x, \mathrm{d}y) = \mu_0(\mathrm{d}x)\,\nu_0(\mathrm{d}y)\,p(x,y)$, where μ_0 and ν_0 are σ-finite measures on $(\mathcal{X}, \mathcal{M})$ and $(\mathcal{Y}, \mathcal{N})$ respectively. Define

$$m(x) = \int \nu_0(\mathrm{d}y)\,p(x,y),$$

and

$$k(x,y) = \begin{cases} p(x,y)/m(x) & m(x) > 0, \\ \int \mu_0(\mathrm{d}x')\,p(x',y) & m(x) = 0. \end{cases}$$

Then $\mu_0(\mathrm{d}x)\,m(x)$ is the marginal distribution of X, and $\nu_0(\mathrm{d}y)\,k(x,y)$ is the conditional distribution of Y given X.

Proof. Set $\mu(\mathrm{d}x) = \mu_0(\mathrm{d}x)\,m(x)$ and $K(x, \mathrm{d}y) = \nu_0(\mathrm{d}y)\,k(x,y)$. We verify that μ is a probability measure, K is a probability kernel, and $\pi(\mathrm{d}x, \mathrm{d}y) = \mu(\mathrm{d}x)K(x, \mathrm{d}y)$, so that Theorem 2.29 applies. That μ is a probability measure follows from Fubini's theorem:

$$\mu(\mathcal{X}) = \int \mu_0(\mathrm{d}x)\,m(x) = \int \mu_0(\mathrm{d}x) \int \nu_0(\mathrm{d}y)\,p(x,y) = \pi(\mathcal{X} \times \mathcal{Y}) = 1.$$

That $K(x, \cdot)$ is a probability measure for each x follows since $k \geq 0$ and

$$K(x, \mathcal{Y}) = \int \nu_0(\mathrm{d}y)\,k(x,y) = 1,$$

which holds in both cases of the definition of k: when $m(x) > 0$, $\int \nu_0(\mathrm{d}y)\,p(x,y)/m(x) = m(x)/m(x) = 1$, and when $m(x) = 0$, $\int \nu_0(\mathrm{d}y) \int \mu_0(\mathrm{d}x')p(x',y) = \pi(\mathcal{X}\times\mathcal{Y}) = 1$ by Fubini. Finally, for any $A \in \mathcal{M}$ and $B \in \mathcal{N}$, another application of Fubini gives

$$\int_A \mu(\mathrm{d}x)\,K(x,B) = \int_A \mu_0(\mathrm{d}x)\,m(x) \int_B \nu_0(\mathrm{d}y)\,k(x,y).$$

When $m(x) > 0$, $m(x) \cdot k(x,y) = p(x,y)$, and when $m(x) = 0$, both $m(x)k(x,y) = 0$ and $\int_B \nu_0(\mathrm{d}y)\,p(x,y) \leq m(x) = 0$, so in both cases $m(x)k(x,y) = p(x,y)$. Therefore

$$\int_A \mu(\mathrm{d}x)\,K(x,B) = \int_A \mu_0(\mathrm{d}x) \int_B \nu_0(\mathrm{d}y)\,p(x,y) = \pi(A \times B). \qquad \square$$

2.4 Conditional Independence

Recall that finitely many sub-σ-algebras $\mathcal{F}_1, \ldots, \mathcal{F}_n \subset \mathcal{H}$ are independent if for all bounded $\mathcal{F}_i$-measurable random variables V_i, $\mathbb{E}[\prod_i V_i] = \prod_i \mathbb{E}[V_i]$. We generalize this by replacing expectations with conditional expectations given some sub-σ-algebra $\mathcal{F}$. Note that taking $\mathcal{F} = \{\Omega, \emptyset\}$ recovers the original definition, since conditioning on the trivial σ-algebra carries no information.

Definition 2.31. Let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space and let $\mathcal{F}$ be a sub-σ-algebra of $\mathcal{H}$. A finite collection of sub-σ-algebras $\mathcal{F}_1, \ldots, \mathcal{F}_n \subset \mathcal{H}$ is **conditionally independent given** $\mathcal{F}$ if for all non-negative $\mathcal{F}_i$-measurable random variables V_i,

$$\mathbb{E}\left[\prod_{i=1}^n V_i \,\middle|\, \mathcal{F}\right] = \prod_{i=1}^n \mathbb{E}[V_i|\mathcal{F}] \quad \text{a.s.}$$

An arbitrarily indexed collection $\{\mathcal{F}_t : t \in T\}$ is conditionally independent given $\mathcal{F}$ if every finite subcollection is. Random variables $\{X_t : t \in T\}$ are conditionally independent given $\mathcal{F}$ if the generated σ-algebras $\{\sigma(X_t) : t \in T\}$ are.

Just as in the unconditional case, it suffices to check the factorization condition on π-systems generating the relevant σ-algebras.

Proposition 2.32. Let $\mathcal{F}, \mathcal{F}_1, \ldots, \mathcal{F}_n$ be sub-σ-algebras of $\mathcal{H}$ with $n \geq 2$. For each $i \leq n$, let $\mathcal{C}_i$ be a π-system containing Ω that generates $\mathcal{F}_i$. Then $\mathcal{F}_1, \ldots, \mathcal{F}_n$ are conditionally independent given $\mathcal{F}$ if and only if for each $H_1 \in \mathcal{C}_1, \ldots, H_n \in \mathcal{C}_n$,

$$\mathbb{P}(H_1 \cap \cdots \cap H_n | \mathcal{F}) = \mathbb{P}(H_1|\mathcal{F}) \cdots \mathbb{P}(H_n|\mathcal{F}) \quad \text{a.s.}$$

Proposition 2.33. $\{\mathcal{F}_t : t \in T\}$ is conditionally independent given $\mathcal{F}$ if and only if $\{\mathcal{G}_t : t \in T\}$ is conditionally independent given $\mathcal{F}$ for every collection satisfying $\mathcal{G}_t \subset \mathcal{F}_t$ for each $t \in T$.

Proof. For necessity, take $\mathcal{G}_t = \mathcal{F}_t$. For sufficiency, note that if $\mathcal{G}_t \subset \mathcal{F}_t$, then for any finite subcollection $t_1, \ldots, t_n$ and any non-negative $\mathcal{G}_{t_i}$-measurable V_i, each V_i is also $\mathcal{F}_{t_i}$-measurable, so conditional independence of $\{\mathcal{F}_t\}$ gives

$$\mathbb{E}\left[\prod_{i=1}^{n} V_i \,\middle|\, \mathcal{F}\right] = \prod_{i=1}^{n} \mathbb{E}[V_i|\mathcal{F}] \quad \text{a.s.} \qquad \square$$

Proposition 2.34. $\{\mathcal{F}_t : t \in T\}$ is conditionally independent given $\mathcal{F}$ if and only if for every subpartition $\{T_s : s \in S\}$ of T, the collection $\{\mathcal{F}_s : s \in S\}$ is conditionally independent given $\mathcal{F}$, where $\mathcal{F}_s = \bigvee_{t \in T_s} \mathcal{F}_t$.

Proof. For necessity, take the subpartition $T_s = \{s\}$, $S = T$. For sufficiency, by definition $\{\mathcal{F}_s : s \in S\}$ is conditionally independent given $\mathcal{F}$ if any finite subcollection $\mathcal{F}_{s_1}, \ldots, \mathcal{F}_{s_n}$ is conditionally independent given $\mathcal{F}$. By Proposition 2.32, it suffices to check that

$$\mathbb{P}(H_1 \cap \cdots \cap H_n | \mathcal{F}) = \mathbb{P}(H_1|\mathcal{F}) \cdots \mathbb{P}(H_n|\mathcal{F})$$

for each $H_i \in \mathcal{C}_i$, where $\mathcal{C}_i$ is a π-system containing Ω and generating $\mathcal{F}_{s_i}$. Take $\mathcal{C}_i$ to be the collection of finite intersections of sets from $\bigcup_{t \in T_{s_i}} \mathcal{F}_t$; this is a π-system containing Ω that generates $\mathcal{F}_{s_i}$. Each $H_1 \cap \cdots \cap H_n$ is then a finite intersection of sets with distinct indices from $\{\mathcal{F}_t : t \in T\}$, and so the conditional independence of this collection yields the required factorization. $\square$

Proposition 2.35. The sequence $\mathcal{F}_1, \mathcal{F}_2, \ldots$ is conditionally independent given $\mathcal{F}$ if and only if for each $k \geq 2$, $\mathcal{F}_k$ is conditionally independent of $\mathcal{F}_1, \ldots, \mathcal{F}_{k-1}$ given $\mathcal{F}$.

Proof. For necessity, subgrouping gives that $\bigvee_{i=1}^{k-1} \mathcal{F}_i$ is conditionally independent of $\mathcal{F}_k$ given $\mathcal{F}$ for each k. For sufficiency, take any non-negative $V_i \in \mathcal{F}_i$. By the tower property and conditional determinism,

$$\mathbb{E}\left[\prod_{i=1}^{n} V_i \,\middle|\, \mathcal{F}\right] = \mathbb{E}\left[\mathbb{E}\left[\prod_{i=1}^{n} V_i \,\middle|\, \mathcal{F} \vee \bigvee_{i=1}^{n-1} \mathcal{F}_i\right] \,\middle|\, \mathcal{F}\right] = \mathbb{E}\left[\prod_{i=1}^{n-1} V_i \cdot \mathbb{E}[V_n | \mathcal{F} \vee \bigvee_{i=1}^{n-1} \mathcal{F}_i] \,\middle|\, \mathcal{F}\right].$$

By Proposition 2.38 and the sequential conditional independence assumption, $\mathbb{E}[V_n|\mathcal{F} \vee \bigvee_{i=1}^{n-1} \mathcal{F}_i] = \mathbb{E}[V_n|\mathcal{F}]$, so

$$= \mathbb{E}\left[\prod_{i=1}^{n-1} V_i \,\middle|\, \mathcal{F}\right] \mathbb{E}[V_n|\mathcal{F}].$$

Repeating this argument on the remaining product yields the full factorization. $\square$

Proposition 2.36. $\{\mathcal{F}_t : t \in T\}$ is conditionally independent given $\mathcal{F}$ if and only if $\{\mathcal{F}_t \vee \mathcal{F} : t \in T\}$ is conditionally independent given $\mathcal{F}$.

Proof. For necessity, apply coarsening since $\mathcal{F}_t \subset \mathcal{F}_t \vee \mathcal{F}$. For sufficiency, by definition we must show that for each finite subcollection $t_1, \ldots, t_n$, the σ-algebras $\mathcal{F}_{t_1} \vee \mathcal{F}, \ldots, \mathcal{F}_{t_n} \vee \mathcal{F}$ are conditionally independent given $\mathcal{F}$. For each i, let $\mathcal{C}_i$ be the π-system of sets of the form $A_i \cap B_i$ with $A_i \in \mathcal{F}_{t_i}$ and $B_i \in \mathcal{F}$; then $\Omega \in \mathcal{C}_i$ and $\sigma(\mathcal{C}_i) = \mathcal{F}_{t_i} \vee \mathcal{F}$. For $H_i = A_i \cap B_i \in \mathcal{C}_i$, by conditional determinism and conditional independence of $\{\mathcal{F}_t\}$,

$$\begin{aligned}
\mathbb{P}(H_1 \cap \cdots \cap H_n | \mathcal{F}) &= \mathbb{E}[\chi_{A_1}\chi_{B_1} \cdots \chi_{A_n}\chi_{B_n} | \mathcal{F}] \\
&= \chi_{B_1} \cdots \chi_{B_n} \, \mathbb{E}[\chi_{A_1} \cdots \chi_{A_n} | \mathcal{F}] \\
&= \chi_{B_1} \cdots \chi_{B_n} \, \mathbb{P}(A_1|\mathcal{F}) \cdots \mathbb{P}(A_n|\mathcal{F}) \\
&= \mathbb{P}(H_1|\mathcal{F}) \cdots \mathbb{P}(H_n|\mathcal{F}).
\end{aligned}$$

By Proposition 2.32, this gives the result. □

Proposition 2.37. $\{\mathcal{F}_t : t \in T\}$ is conditionally independent given $\mathcal{F}$ if and only if for all subsets $S \subset T$, $\{\mathcal{F}_t : t \in T\}$ is conditionally independent given $\mathcal{F} \vee \left(\bigvee_{s \in S} \mathcal{F}_s\right)$.

Proof. For necessity, take $S = \emptyset$. For sufficiency, let $\mathcal{F}' = \mathcal{F} \vee \left(\bigvee_{s \in S} \mathcal{F}_s\right)$ and consider a finite subcollection $t_1, \ldots, t_n \in T$. Suppose without loss of generality that $t_1, \ldots, t_m \in S$ and $t_{m+1}, \ldots, t_n \notin S$ for some $m \leq n$. For non-negative $V_i \in \mathcal{F}_{t_i}$, by conditional determinism,

$$\mathbb{E}\left[\prod_{i=1}^{n} V_i \,\middle|\, \mathcal{F}'\right] = V_1 \cdots V_m \, \mathbb{E}\left[\prod_{i=m+1}^{n} V_i \,\middle|\, \mathcal{F}'\right].$$

Since $\bigvee_{i=m+1}^{n} \mathcal{F}_{t_i}$ is conditionally independent of $\bigvee_{s \in S} \mathcal{F}_s$ given $\mathcal{F}$ by assumption and subgrouping, Proposition 2.38 gives

$$\mathbb{E}\left[\prod_{i=m+1}^{n} V_i \,\middle|\, \mathcal{F}'\right] = \mathbb{E}\left[\prod_{i=m+1}^{n} V_i \,\middle|\, \mathcal{F}\right] = \prod_{i=m+1}^{n} \mathbb{E}[V_i|\mathcal{F}],$$

where the last equality uses the original conditional independence assumption and subgrouping. By conditional determinism and Proposition 2.38,

$$V_1 \cdots V_m \prod_{i=m+1}^{n} \mathbb{E}[V_i|\mathcal{F}] = \prod_{i=1}^{m} \mathbb{E}[V_i|\mathcal{F}'] \cdot \prod_{i=m+1}^{n} \mathbb{E}[V_i|\mathcal{F}'] = \prod_{i=1}^{n} \mathbb{E}[V_i|\mathcal{F}'].$$ □

Proposition 2.38. Let $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}$ be sub-σ-algebras of $\mathcal{H}$. The following are equivalent:

(a) $\mathcal{F}_1$ and $\mathcal{F}_2$ are conditionally independent given $\mathcal{F}$.

(b) $\mathbb{E}[V_2|\mathcal{F} \vee \mathcal{F}_1] = \mathbb{E}[V_2|\mathcal{F}]$ a.s. for every non-negative $\mathcal{F}_2$-measurable V_2.

(c) $\mathbb{E}[V_2|\mathcal{F} \vee \mathcal{F}_1]$ is $\mathcal{F}$-measurable for every non-negative $\mathcal{F}_2$-measurable V_2.

Proof. **(a) $\implies$ (b).** By Propositions 2.36 and 2.33, $\mathcal{F}_1 \vee \mathcal{F}$ and $\mathcal{F}_2$ are conditionally independent given $\mathcal{F}$. Therefore for any non-negative $V_1 \in \mathcal{F}_1 \vee \mathcal{F}$ and $V_2 \in \mathcal{F}_2$,

$$\mathbb{E}[V_1 V_2|\mathcal{F}] = \mathbb{E}[V_1|\mathcal{F}] \, \mathbb{E}[V_2|\mathcal{F}] = \mathbb{E}\big[V_1 \, \mathbb{E}[V_2|\mathcal{F}] \,\big|\, \mathcal{F}\big].$$

Taking expectations and applying the tower property,

$$\mathbb{E}[V_1 V_2] = \mathbb{E}\big[V_1 \, \mathbb{E}[V_2|\mathcal{F}]\big].$$

This is precisely the characterization of $\mathbb{E}[V_2|\mathcal{F} \vee \mathcal{F}_1]$, so $\mathbb{E}[V_2|\mathcal{F} \vee \mathcal{F}_1] = \mathbb{E}[V_2|\mathcal{F}]$ a.s.

(b) $\implies$ (c). If $\mathbb{E}[V_2|\mathcal{F} \vee \mathcal{F}_1] = \mathbb{E}[V_2|\mathcal{F}]$ a.s., then since $\mathbb{E}[V_2|\mathcal{F}]$ is $\mathcal{F}$-measurable by definition, so is $\mathbb{E}[V_2|\mathcal{F} \vee \mathcal{F}_1]$.

(c) $\implies$ (a). Suppose $\mathbb{E}[V_2|\mathcal{F} \vee \mathcal{F}_1]$ is $\mathcal{F}$-measurable for all non-negative $V_2 \in \mathcal{F}_2$. By conditional determinism and the tower property,

$$\mathbb{E}[V_2|\mathcal{F} \vee \mathcal{F}_1] = \mathbb{E}\big[\mathbb{E}[V_2|\mathcal{F} \vee \mathcal{F}_1] \;\big|\; \mathcal{F}\big] = \mathbb{E}[V_2|\mathcal{F}].$$

Therefore for any non-negative $V_1 \in \mathcal{F}_1$,

$$\begin{aligned}
\mathbb{E}[V_1 V_2|\mathcal{F}] &= \mathbb{E}\big[\mathbb{E}[V_1 V_2|\mathcal{F} \vee \mathcal{F}_1] \;\big|\; \mathcal{F}\big] \\
&= \mathbb{E}\big[V_1 \, \mathbb{E}[V_2|\mathcal{F} \vee \mathcal{F}_1] \;\big|\; \mathcal{F}\big] \\
&= \mathbb{E}\big[V_1 \, \mathbb{E}[V_2|\mathcal{F}] \;\big|\; \mathcal{F}\big] \\
&= \mathbb{E}[V_2|\mathcal{F}] \, \mathbb{E}[V_1|\mathcal{F}],
\end{aligned}$$

where the second equality is conditional determinism, the third uses the above, and the fourth uses conditional determinism again since $\mathbb{E}[V_2|\mathcal{F}]$ is $\mathcal{F}$-measurable. This is precisely conditional independence of $\mathcal{F}_1$ and $\mathcal{F}_2$ given $\mathcal{F}$. □

3 Concentration Inequalities and Divergences

3.1 Concentration Inequalities

Proposition 3.1. *(Markov's Inequality)* Let X be as positive real-valued random variable, then

$$\mathbb{P}(X \geq t) \leq t^{-1}\mathbb{E}X$$

for any $t > 0$.

Proof. Notice $t^{-1}X(\omega) \geq \chi_{[t,\infty)}(X(\omega))$ for any positive random variable X and $t > 0$. Hence

$$t^{-1}\mathbb{E}X \geq \mathbb{E}\left[\chi_{[t,\infty)}(X(\omega))\right] = \int \chi_{[t,\infty)}\,\mathrm{d}\big(\mathbb{P}\circ X^{-1}\big) = \big(\mathbb{P}\circ X^{-1}\big)([t,\infty)) = \mathbb{P}(X \geq t).$$

□

Proposition 3.2. *(Generalized Markov's Inequality)* Let $f : \mathbb{R} \to \mathbb{R}$ be a monotone non-decreasing, positive function, X be a real-valued random variable, and $t \in \mathbb{R}$. Then

$$\mathbb{P}(X \geq t) \leq f(t)^{-1}\mathbb{E}\left[f(X)\right].$$

Proof. Since f is monotone non-decreasing, $\{\omega \in \Omega : X(\omega) \geq t\} \subset \{\omega \in \Omega : f(X(\omega)) \geq f(t)\}$. So

$$\mathbb{P}(X \geq t) \leq \mathbb{P}(f(X) \geq f(t)).$$

Since the range of f is positive and since $f(X)$ is itself a random variable, we apply Markov's Inequality to obtain

$$\mathbb{P}(X \geq t) \leq \mathbb{P}(f(X) \geq f(t)) \leq f(t)^{-1}\mathbb{E}\left[f(X)\right].$$

□

Markov's inequality is a powerful starting point, but vanilla Markov by itself does not always yield useful bounds. For instance, applying Markov directly to the empirical average $\bar{X}_N = \frac{1}{N}\sum_{n=1}^{N} X_n$ of i.i.d. positive variables gives

$$\mathbb{P}(\bar{X}_N \geq t) \leq \frac{\mathbb{E}X}{t + \mathbb{E}X},$$

which has no dependence on N and is therefore useless for studying concentration as we collect more data. The key idea is that by choosing different monotone functions f in the generalized Markov inequality, we can obtain much stronger bounds. A rough heuristic: the more quickly f grows, the tighter the bound, but the stronger the assumptions required on X. Two canonical choices are $f(x) = x^2$, leading to Chebyshev's inequality (requiring $\operatorname{Var} X < \infty$), and $f(x) = e^{sx}$, leading to the Chernoff bound (requiring $\mathbb{E}[e^{sX}] < \infty$ for some $s > 0$).

Proposition 3.3. *(Chebyshev's Inequality)* For a random variable X with $\operatorname{Var} X < \infty$,

$$\mathbb{P}(|X - \mathbb{E}X| \geq k\sqrt{\operatorname{Var} X}) \leq \frac{1}{k^2}.$$

Proof. Apply the generalized Markov inequality with $f(x) = x^2$ to the random variable $|X - \mathbb{E}X|$:

$$\mathbb{P}(|X - \mathbb{E}X| \geq k\sqrt{\operatorname{Var} X}) = \mathbb{P}(|X - \mathbb{E}X|^2 \geq k^2 \operatorname{Var} X) \leq \frac{\mathbb{E}(X - \mathbb{E}X)^2}{k^2 \operatorname{Var} X} = \frac{1}{k^2}.$$ □

Proposition 3.4. *(Chernoff Bound)* For a random variable $X \in \mathbb{R}$,

$$\mathbb{P}(X \geq t) \leq \inf_{s \geq 0} e^{-st}\, \mathbb{E}\left[e^{sX}\right].$$

Proof. Apply the generalized Markov inequality with $f(x) = e^{sx}$ for any fixed $s \geq 0$:

$$\mathbb{P}(X \geq t) = \mathbb{P}(e^{sX} \geq e^{st}) \leq e^{-st}\, \mathbb{E}[e^{sX}].$$

Since this holds for every $s \geq 0$, taking the infimum over $s \geq 0$ yields the result. □

To make use of the Chernoff bound we need to control the moment generating function $\mathbb{E}[e^{sX}]$. A natural and widely applicable assumption is subGaussianity.

Definition 3.5. A random variable X is **σ^2-subGaussian** if for all $s \geq 0$,

$$\mathbb{E}[e^{sX}] \leq e^{\sigma^2 s^2/2}.$$

SubGaussian variables necessarily satisfy $\mathbb{E}X = 0$ and $\operatorname{Var} X \leq \sigma^2$, which follows from a Taylor expansion of e^{sX} around $s = 0$ and a careful application of the dominated convergence theorem. Centered Gaussian variables with variance σ^2 are σ^2-subGaussian. Bounded centered variables are also subGaussian, per Hoeffding's lemma below.

Lemma 3.6. *(Hoeffding's Lemma)* If $\mathbb{E}X = 0$ and $X \in [a, b]$ almost surely, then

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

In particular, X is $\frac{(b-a)^2}{4}$-subGaussian.

Proof. Since $X \in [a, b]$ almost surely, convexity of $x \mapsto e^{sx}$ gives

$$e^{sX} \leq \frac{b - X}{b - a} e^{sa} + \frac{X - a}{b - a} e^{sb}.$$

Taking expectations and using $\mathbb{E}X = 0$,

$$\mathbb{E}[e^{sX}] \leq \frac{b}{b-a} e^{sa} - \frac{a}{b-a} e^{sb} =: e^{\varphi(u)},$$

where $\theta = -a/(b-a) \in [0, 1]$, $u = s(b-a)$, and $\varphi(u) = \log\big((1-\theta)e^{-\theta u} + \theta e^{(1-\theta)u}\big)$. It suffices to show $\varphi(u) \leq u^2/8$. One checks that $\varphi(0) = 0$ and $\varphi'(0) = 0$, and a direct calculation gives

$$\varphi''(u) = \theta(1-\theta) \frac{e^{-\theta u} \cdot e^{(1-\theta)u}}{\big((1-\theta)e^{-\theta u} + \theta e^{(1-\theta)u}\big)^2} \leq \frac{1}{4},$$

where the inequality uses $\theta(1-\theta) \leq 1/4$ and AM-GM on the remaining factor. Taylor's theorem then gives $\varphi(u) \leq u^2/8$, and substituting $u = s(b-a)$ yields the result. □

Proposition 3.7. *(Hoeffding's Inequality)* Suppose $X_1, \ldots, X_n$ are independent, with X_i being σ_i^2-subGaussian. Then

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \geq t\right) \leq e^{-\frac{n^2 t^2}{2\sum_i \sigma_i^2}}, \qquad \mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^n X_i\right| \geq t\right) \leq 2e^{-\frac{n^2 t^2}{2\sum_i \sigma_i^2}}.$$

Proof. By the Chernoff bound,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq t\right) \leq \inf_{s\geq 0} e^{-st}\,\mathbb{E}\left[e^{\frac{s}{n}\sum_i X_i}\right].$$

Since $X_1, \ldots, X_n$ are independent, the moment generating function factorizes:

$$\mathbb{E}\left[e^{\frac{s}{n}\sum_i X_i}\right] = \prod_{i=1}^{n} \mathbb{E}\left[e^{\frac{s}{n}X_i}\right].$$

Applying the σ_i^2-subGaussian bound to each factor,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq t\right) \leq \inf_{s\geq 0} e^{-st} \prod_{i=1}^{n} e^{\frac{\sigma_i^2 s^2}{2n^2}} = \inf_{s\geq 0} \exp\left(-st + \frac{s^2 \sum_i \sigma_i^2}{2n^2}\right).$$

The exponent is a convex quadratic in s, minimized at $s^\star = \frac{n^2 t}{\sum_i \sigma_i^2}$, giving

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq t\right) \leq e^{-\frac{n^2 t^2}{2\sum_i \sigma_i^2}}.$$

The two-sided bound follows since $\mathbb{P}(|\bar{X}| \geq t) \leq \mathbb{P}(\bar{X} \geq t) + \mathbb{P}(-\bar{X} \geq t)$, and $-X_i$ is also σ_i^2-subGaussian whenever X_i is. □

While subGaussian variables have moment generating functions bounded for all $s \geq 0$, in many applications this is too strong an assumption. SubExponential variables relax this by only requiring the MGF bound to hold in a neighbourhood of zero, at the cost of a two-regime tail bound.

Definition 3.8. A random variable X is (σ^2, b)**-subExponential** if for all $|s| < 1/b$,

$$\mathbb{E}[e^{sX}] \leq e^{\sigma^2 s^2/2}.$$

As with subGaussian variables, a Taylor expansion around $s = 0$ and the dominated convergence theorem show that $\mathbb{E}X = 0$.

Proposition 3.9. For a (σ^2, b)-subExponential random variable X,

$$\mathbb{P}(X \geq t) \leq \exp\left(-\frac{t\min(t,\, \sigma^2/b)}{2\sigma^2}\right).$$

Proof. By the Chernoff bound, restricting the infimum to $s \in [0, 1/b)$ where the subExponential bound applies,

$$\mathbb{P}(X \geq t) \leq \inf_{0\leq s<1/b} \exp\left(-st + \frac{\sigma^2 s^2}{2}\right).$$

The unconstrained minimizer of the exponent is $s^\star = t/\sigma^2$. We consider two regimes.
Case 1: $t \leq \sigma^2/b$**.** Then $s^\star = t/\sigma^2 \leq 1/b$, so the unconstrained minimizer is feasible, giving

$$\mathbb{P}(X \geq t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right) = \exp\left(-\frac{t\cdot t}{2\sigma^2}\right) = \exp\left(-\frac{t\min(t,\sigma^2/b)}{2\sigma^2}\right).$$

Case 2: $t > \sigma^2/b$**.** The optimum is at the boundary $s = 1/b$, giving

$$\mathbb{P}(X \geq t) \leq \exp\left(-\frac{t}{b} + \frac{\sigma^2}{2b^2}\right) \leq \exp\left(-\frac{t}{b} + \frac{t}{2b}\right) = \exp\left(-\frac{t}{2b}\right) = \exp\left(-\frac{t\min(t,\sigma^2/b)}{2\sigma^2}\right),$$

where the second inequality uses $\sigma^2/b \leq t$. Combining both cases yields the result. □

Most concentration inequalities, including Hoeffding's inequality, rely crucially on independence of the random variables involved. In practice however, one often encounters dependent collections of variables that are known to concentrate marginally, but whose joint distribution is difficult to analyze directly. The following lemma reduces joint concentration to a collection of marginal concentration problems, provided the function of interest is coordinate-wise monotone.

Lemma 3.10. Let $X = (X_n)$ be a countable sequence of real-valued random variables, let $s \in \mathbb{R}$, and let $f : \mathbb{R}^\infty \to \mathbb{R}$ be measurable and coordinate-wise monotone non-decreasing, i.e., $x_n \le y_n$ for each n implies $f(x) \le f(y)$. Define $A_s = \{a \in \mathbb{R}^\infty : f(a) < s\}$. Then

$$\mathbb{P}(f(X) \ge s) \le \inf_{a \in A_s} \sum_n \mathbb{P}(X_n \ge a_n).$$

Proof. Fix any $a \in A_s$, so that $f(a) < s$. Decompose the event $\{f(X) \ge s\}$ according to whether all coordinates satisfy $X_n < a_n$:

$$\mathbb{P}(f(X) \ge s) = \mathbb{P}(f(X) \ge s,\, \forall n :\, X_n < a_n) + \mathbb{P}(f(X) \ge s,\, \exists n :\, X_n \ge a_n).$$

On the event $\{\forall n : X_n < a_n\}$, coordinate-wise monotonicity gives $f(X) \le f(a) < s$, so the first term vanishes. Therefore

$$\mathbb{P}(f(X) \ge s) \le \mathbb{P}(\exists n :\, X_n \ge a_n) \le \sum_n \mathbb{P}(X_n \ge a_n),$$

where the last step is a union bound. Taking the infimum over all $a \in A_s$ yields the result. □

Corollary 3.11. Under the setup of Lemma 3.10, the following three inequalities hold:

(i) $\mathbb{P}\Big(\sup_n X_n \ge s\Big) \le \sum_n \mathbb{P}(X_n \ge s).$

(ii) $\mathbb{P}\Big(\inf_n X_n \ge s\Big) \le \inf_n \mathbb{P}(X_n \ge s).$

(iii) $\mathbb{P}\left(\sum_n X_n \ge s\right) \le \inf_{\sum_n a_n < s} \sum_n \mathbb{P}(X_n \ge a_n).$

Proof. Each result is an application of Lemma 3.10 to a specific choice of f.
(i). Take $f(x) = \sup_n x_n$, which is coordinate-wise monotone. Here $A_s = \{a : \sup_n a_n < s\}$, and each individual probability $\mathbb{P}(X_n \ge a_n)$ is minimized by taking a_n as large as possible, so we take $a_n \nearrow s$ for all n, giving

$$\mathbb{P}(\sup_n X_n \ge s) \le \lim_{b \nearrow s} \sum_n \mathbb{P}(X_n \ge b) = \sum_n \mathbb{P}(X_n \ge s),$$

where the interchange of limit and sum is justified by monotonicity.
(ii). Take $f(x) = \inf_n x_n$, which is also coordinate-wise monotone. Here $A_s = \{a : \inf_n a_n < s\}$, so $a \in A_s$ as long as at least one coordinate is less than s. Setting all but one coordinate equal to ∞ and optimizing,

$$\mathbb{P}(\inf_n X_n \ge s) \le \inf_n \lim_{b \nearrow s} \mathbb{P}(X_n \ge b) = \inf_n \mathbb{P}(X_n \ge s).$$

(iii). Take $f(x) = \sum_n x_n$, which is coordinate-wise monotone. The result is an immediate application of Lemma 3.10 since $A_s = \{a : \sum_n a_n < s\}$. □

3.2 Integral Probability Metrics

A very important topic in both computational and theoretical probability is how to compare the discrepancy between probability distributions. In practice, this comes up in methods such as variational inference and Markov chain Monte Carlo, as well as in statistical estimation problems. In theory, it provides a much richer vocabulary for talking about convergence of distributions than weak convergence alone. Many natural notions of discrepancy between probability distributions fall into the class of **integral probability metrics**, which we now define.

Definition 3.12. Let $(\Omega, \mathcal{H})$ be a measurable space and let $\mathscr{P}$ be the set of all probability measures on $(\Omega, \mathcal{H})$. For a class $\mathcal{F}$ of measurable functions $\Omega \to \mathbb{R}$, the **integral probability metric** (IPM) generated by $\mathcal{F}$ is

$$d_{\mathcal{F}}(\mu, \nu) = \sup_{f \in \mathcal{F}} \left| \int f \, \mathrm{d}\mu - \int f \, \mathrm{d}\nu \right|.$$

We say $\mathcal{F}$ **separates** $\mathscr{P}$ if for all $\mu, \nu \in \mathscr{P}$ with $\mu \neq \nu$, there exists $f \in \mathcal{F}$ such that $\int f \, \mathrm{d}\mu \neq \int f \, \mathrm{d}\nu$.

Proposition 3.13. $d_{\mathcal{F}}$ is a metric on $\mathscr{P}$ if and only if $\mathcal{F}$ separates $\mathscr{P}$.

Proof. For any $\mathcal{F}$, $d_{\mathcal{F}}$ is symmetric and non-negative by inspection, satisfies $d_{\mathcal{F}}(\mu, \mu) = 0$, and satisfies the triangle inequality since

$$\left| \int f \, \mathrm{d}\mu - \int f \, \mathrm{d}\nu \right| \leq \left| \int f \, \mathrm{d}\mu - \int f \, \mathrm{d}\rho \right| + \left| \int f \, \mathrm{d}\rho - \int f \, \mathrm{d}\nu \right|.$$

It remains to check positive definiteness. We have $d_{\mathcal{F}}(\mu, \nu) > 0$ for $\mu \neq \nu$ if and only if there exists $f \in \mathcal{F}$ with $\int f \, \mathrm{d}\mu \neq \int f \, \mathrm{d}\nu$, which is precisely the condition that $\mathcal{F}$ separates $\mathscr{P}$. □

The behaviour of $d_{\mathcal{F}}$ is determined entirely by the choice of $\mathcal{F}$; larger classes correspond to more stringent metrics, since $\mathcal{G} \subset \mathcal{F}$ implies $d_{\mathcal{G}} \leq d_{\mathcal{F}}$. Three canonical choices are:

- **Total variation:** $\mathcal{F}_{\mathrm{TV}} = \{f \text{ measurable} : 0 \leq f(x) \leq 1 \ \forall x\}$.
- **1-Wasserstein:** $\mathcal{F}_{\mathrm{W}} = \left\{ f : \sup_{x \neq y} \frac{|f(x) - f(y)|}{d(x,y)} \leq 1 \right\}$, where d is a metric on Ω.
- **Bounded-Lipschitz:** $\mathcal{F}_{\mathrm{BL}} = \left\{ f : \sup_x |f(x)| \leq 1, \ \sup_{x \neq y} \frac{|f(x) - f(y)|}{d(x,y)} \leq 1 \right\}$, where d is a metric on Ω.

Both the 1-Wasserstein and bounded-Lipschitz distances require (Ω, d) to be a metric space, and are sensitive to how far apart mass is placed by the two distributions. The total variation distance requires no metric structure on Ω, and depends only on how much mass the two distributions place on the same sets, not on how far apart those sets are.

Proposition 3.14. For any probability measures μ, ν on a metric space (Ω, d),

$$d_{\mathrm{BL}}(\mu, \nu) \leq \min\{d_{\mathrm{W}}(\mu, \nu), \, 2d_{\mathrm{TV}}(\mu, \nu)\}.$$

In particular, convergence in total variation or 1-Wasserstein distance implies convergence in bounded-Lipschitz distance. The total variation and 1-Wasserstein distances do not bound each other in general.

Proof. The bound $d_{\mathrm{BL}} \le d_{\mathrm{W}}$ follows immediately from $\mathcal{F}_{\mathrm{BL}} \subset \mathcal{F}_{\mathrm{W}}$. For the bound $d_{\mathrm{BL}} \le 2d_{\mathrm{TV}}$, note that $\mathcal{F}_{\mathrm{BL}} \subset \{f : \sup_x |f(x)| \le 1\}$, so

$$d_{\mathrm{BL}}(\mu,\nu) \le \sup_{|f|\le 1} \left| \int f \, \mathrm{d}\mu - \int f \, \mathrm{d}\nu \right| = 2 \sup_{|f|\le 1} \left| \int \tfrac{f+1}{2} \, \mathrm{d}\mu - \int \tfrac{f+1}{2} \, \mathrm{d}\nu \right|.$$

Since $|f| \le 1$ implies $(f+1)/2 \in [0,1]$, the supremum on the right is bounded by $d_{\mathrm{TV}}(\mu,\nu)$, giving $d_{\mathrm{BL}} \le 2d_{\mathrm{TV}}$. That d_{W} and d_{TV} do not bound each other follows from considering $\mu = \delta_0$, $\nu = \delta_x$: for $x < 1$, $d_{\mathrm{TV}} = 1 > x = d_{\mathrm{W}}$, while for $x > 1$, $d_{\mathrm{W}} = x > 1 = d_{\mathrm{TV}}$. □

Proposition 3.15. Suppose (Ω, d) is bounded with diameter $M = \sup_{x,y\in\Omega} d(x,y) < \infty$. Then

$$d_{\mathrm{W}}(\mu,\nu) \le \frac{M+2}{2} \, d_{\mathrm{BL}}(\mu,\nu).$$

Proof. For any $f \in \mathcal{F}_{\mathrm{W}}$, boundedness of Ω gives $\sup_{x,y} |f(x) - f(y)| \le M$. Let $c_f = \inf_{x\in\Omega} f(x)$ and define

$$g_f(x) = \frac{2}{M+2} \left(f(x) - c_f - \frac{M}{2} \right).$$

Since $|f(x) - c_f| \le M$, we have $|g_f(x)| \le M/(M+2) \le 1$. Moreover,

$$\sup_{x\neq y} \frac{|g_f(x) - g_f(y)|}{d(x,y)} = \frac{2}{M+2} \sup_{x\neq y} \frac{|f(x) - f(y)|}{d(x,y)} \le \frac{2}{M+2} \le 1,$$

so $g_f \in \mathcal{F}_{\mathrm{BL}}$. Since c_f and $M/2$ are constants that integrate to the same value under μ and ν,

$$\int f \, \mathrm{d}\mu - \int f \, \mathrm{d}\nu = \frac{M+2}{2} \left(\int g_f \, \mathrm{d}\mu - \int g_f \, \mathrm{d}\nu \right).$$

Taking the supremum over $f \in \mathcal{F}_{\mathrm{W}}$ and using $g_f \in \mathcal{F}_{\mathrm{BL}}$ yields $d_{\mathrm{W}} \le \frac{M+2}{2} d_{\mathrm{BL}}$. □

The key reason to study d_{BL} is that it metrizes weak convergence. Combined with Proposition 3.14, this means convergence in total variation or 1-Wasserstein distance implies weak convergence, but the converse need not hold in either case.

Proposition 3.16. On a standard Borel space, $\mu_n \xrightarrow{w} \mu$ if and only if $d_{\mathrm{BL}}(\mu_n, \mu) \to 0$.

Proof. Omitted. □

Proposition 3.17. For probability measures μ, ν and any σ-finite measure λ with $\lambda \gg \mu$ and $\lambda \gg \nu$,

$$d_{\mathrm{TV}}(\mu,\nu) = \sup_{A\in\mathcal{H}} |\mu(A) - \nu(A)| = \frac{1}{2} \int \left| \frac{\mathrm{d}\mu}{\mathrm{d}\lambda} - \frac{\mathrm{d}\nu}{\mathrm{d}\lambda} \right| \mathrm{d}\lambda = \inf_{(X,Y)\in\Gamma} \mathbb{P}(X \neq Y),$$

where Γ is the set of joint distributions of (X, Y) with $X \sim \mu$ and $Y \sim \nu$.

Proof. We prove the first two equalities; the coupling representation is stated without proof. Such a λ always exists; for instance take $\lambda = \mu + \nu$. By the Radon–Nikodym theorem,

$$d_{\mathrm{TV}}(\mu,\nu) = \sup_{0\le f\le 1} \int f \left(\frac{\mathrm{d}\mu}{\mathrm{d}\lambda} - \frac{\mathrm{d}\nu}{\mathrm{d}\lambda} \right) \mathrm{d}\lambda.$$

Define $A = \left\{x : \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}(x) - \frac{\mathrm{d}\nu}{\mathrm{d}\lambda}(x) \geq 0\right\}$, which is measurable. To maximize the integral, we should set $f = 1$ where the integrand is positive and $f = 0$ where it is negative, giving $f = \chi_A$; or we can set $f = \chi_{A^c}$ by the symmetric argument. The supremum is therefore achieved at one of these two choices. Now observe that

$$\begin{aligned}\int_A \left(\frac{\mathrm{d}\mu}{\mathrm{d}\lambda} - \frac{\mathrm{d}\nu}{\mathrm{d}\lambda}\right) \mathrm{d}\lambda = \mu(A) - \nu(A) = (1 - \mu(A^c)) - (1 - \nu(A^c)) = \nu(A^c) - \mu(A^c) \\ = \int_{A^c} \left(\frac{\mathrm{d}\nu}{\mathrm{d}\lambda} - \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}\right) \mathrm{d}\lambda,\end{aligned}$$

so the two candidate values of the supremum are equal. Hence

$$d_{\mathrm{TV}}(\mu, \nu) = \mu(A) - \nu(A) = |\mu(A) - \nu(A)|.$$

For any measurable B, choosing $f = \chi_B$ gives $d_{\mathrm{TV}}(\mu, \nu) \geq |\mu(B) - \nu(B)|$, so $d_{\mathrm{TV}}(\mu, \nu) = \sup_{A \in \mathcal{H}} |\mu(A) - \nu(A)|$. For the second equality, since the two terms in the maximum are equal, d_{TV} equals their average, and reintroducing absolute values using the sign of the integrand on A and A^c gives

$$d_{\mathrm{TV}}(\mu, \nu) = \frac{1}{2}\int_A \left|\frac{\mathrm{d}\mu}{\mathrm{d}\lambda} - \frac{\mathrm{d}\nu}{\mathrm{d}\lambda}\right| \mathrm{d}\lambda + \frac{1}{2}\int_{A^c} \left|\frac{\mathrm{d}\nu}{\mathrm{d}\lambda} - \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}\right| \mathrm{d}\lambda = \frac{1}{2}\int \left|\frac{\mathrm{d}\mu}{\mathrm{d}\lambda} - \frac{\mathrm{d}\nu}{\mathrm{d}\lambda}\right| \mathrm{d}\lambda. \qquad \square$$

Proposition 3.18. For probability measures μ, ν on (Ω, d),

$$d_{\mathrm{W}}(\mu, \nu) = \inf_{(X,Y) \in \Gamma} \mathbb{E}[d(X, Y)],$$

where Γ is the set of joint distributions of (X, Y) with $X \sim \mu$ and $Y \sim \nu$. If μ, ν are probability measures on $\mathbb{R}$ with CDFs F_μ and F_ν, then

$$d_{\mathrm{W}}(\mu, \nu) = \int_{\mathbb{R}} |F_\mu(x) - F_\nu(x)| \, \mathrm{d}x.$$

Proof. Omitted. $\square$

In practice, working with IPMs has the advantage of not requiring density functions for μ or ν, since the definition only involves integrals of test functions. However, computing $d_{\mathcal{F}}(\mu, \nu)$ can be difficult due to the need to evaluate a supremum over $\mathcal{F}$. The situation is further complicated in practice, since μ and ν are typically unknown and must be estimated from n draws via empirical distributions μ_n, ν_n. It is not always the case that $d_{\mathcal{F}}(\mu_n, \nu_n) \to d_{\mathcal{F}}(\mu, \nu)$ as $n \to \infty$; guaranteeing this requires that $\mathcal{F}$ consist of sufficiently smooth functions. Notably, the total variation distance fails this requirement: if μ is nonatomic and $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ is the empirical distribution of n i.i.d. draws, then $d_{\mathrm{TV}}(\mu_n, \mu) = 1$ almost surely for all n, since μ_n is purely atomic while μ is not.

3.3 ϕ–Divergences

Recall that integral probability metrics do not require density functions for μ or ν, but can be difficult to compute and estimate from data. When densities are available, ϕ-divergences offer a complementary approach. For a convex function $\phi : \mathbb{R} \to \mathbb{R}$ with $\phi(1) = 0$, the ϕ-divergence between ν and μ is defined as an integral of ϕ applied to the density ratio $\mathrm{d}\nu / \mathrm{d}\mu$. We use the notation "$\|$" to emphasize that ϕ-divergences are not symmetric in general, and so do not define a metric on $\mathscr{P}$. Nevertheless, they are non-negative and, under mild conditions, uniquely identify probability measures.

Definition 3.19. Let $\phi : \mathbb{R} \to \mathbb{R}$ be convex with $\phi(1) = 0$. The **ϕ-divergence** between probability measures ν and μ is

$$d_\phi(\nu\|\mu) = \begin{cases} \displaystyle\int \phi\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x)\right) \mathrm{d}\mu(x) & \mu \gg \nu, \\ \infty & \text{otherwise.} \end{cases}$$

Proposition 3.20. $d_\phi(\nu\|\mu) \geq 0$, and if ϕ is strictly convex then $\nu = \mu$ if and only if $d_\phi(\nu\|\mu) = 0$.

Proof. If $\mu \not\gg \nu$ the result holds by definition, so assume $\mu \gg \nu$. Letting $X \sim \mu$ and applying Jensen's inequality,

$$d_\phi(\nu\|\mu) = \mathbb{E}\left[\phi\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(X)\right)\right] \geq \phi\left(\mathbb{E}\left[\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(X)\right]\right) = \phi(1) = 0.$$

When ϕ is strictly convex, equality holds if and only if $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(X) = c$ almost surely for some constant $c \geq 0$. Since $\int \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\, \mathrm{d}\mu = 1$, we must have $c = 1$, and hence $\mu = \nu$. □

The two key structural properties of ϕ-divergences are convexity and the data processing inequality. Both follow from a general lemma bounding the ϕ-divergence between marginal measures.

Lemma 3.21. Let $(E, \mathcal{M})$ and $(F, \mathcal{N})$ be standard measurable spaces. Let μ, ν be probability measures on $(E, \mathcal{M})$ with $\mu \gg \nu$, and let L, K be probability kernels from $(E, \mathcal{M})$ to $(F, \mathcal{N})$ with $\frac{\mathrm{d}L}{\mathrm{d}K}(x, y)$ existing and measurable with respect to $\mathcal{M} \otimes \mathcal{N}$. Then

$$d_\phi(\nu L\|\mu K) \leq \int \mu(\mathrm{d}x) \int K(x, \mathrm{d}y)\, \phi\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x)\frac{\mathrm{d}L}{\mathrm{d}K}(x, y)\right).$$

Proof. By the earlier result on marginal measure densities, $\frac{\mathrm{d}(\nu L)}{\mathrm{d}(\mu K)}$ exists and satisfies

$$\frac{\mathrm{d}(\nu L)}{\mathrm{d}(\mu K)}(y) = \int M(y, \mathrm{d}x)\, \frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x)\frac{\mathrm{d}L}{\mathrm{d}K}(x, y),$$

where $M(y, \mathrm{d}x)$ is the disintegration of μK over its second component. Therefore $\mu K \gg \nu L$, and

$$d_\phi(\nu L\|\mu K) = \int \mu K(\mathrm{d}y)\, \phi\left(\int M(y, \mathrm{d}x)\, \frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x)\frac{\mathrm{d}L}{\mathrm{d}K}(x, y)\right).$$

Applying Jensen's inequality to ϕ with respect to the probability measure $M(y, \mathrm{d}x)$, then using $\mu K(\mathrm{d}y)M(y, \mathrm{d}x) = \mu(\mathrm{d}x)K(x, \mathrm{d}y)$,

$$d_\phi(\nu L\|\mu K) \leq \int \mu K(\mathrm{d}y) \int M(y, \mathrm{d}x)\, \phi\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x)\frac{\mathrm{d}L}{\mathrm{d}K}(x, y)\right) = \int \mu(\mathrm{d}x) \int K(x, \mathrm{d}y)\, \phi\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x)\frac{\mathrm{d}L}{\mathrm{d}K}(x, y)\right).$$

□

Theorem 3.22. In the setting of Lemma 3.21,

$$d_\phi(\mu L\|\mu K) \leq \int \mu(\mathrm{d}x)\, d_\phi(L(x, \cdot)\|K(x, \cdot)).$$

In particular, for $\alpha \in [0, 1]$ and probability measures p_1, p_2, q_1, q_2,

$$d_\phi(\alpha q_1 + (1 - \alpha)q_2\|\alpha p_1 + (1 - \alpha)p_2) \leq \alpha\, d_\phi(q_1\|p_1) + (1 - \alpha)\, d_\phi(q_2\|p_2).$$

Proof. Setting $\nu = \mu$ in Lemma 3.21 gives $\frac{\mathrm{d}\nu}{\mathrm{d}\mu} = 1$, so

$$d_\phi(\mu L \| \mu K) \le \int \mu(\mathrm{d}x) \int K(x, \mathrm{d}y)\, \phi\!\left(\frac{\mathrm{d}L}{\mathrm{d}K}(x,y)\right) = \int \mu(\mathrm{d}x)\, d_\phi(L(x,\cdot)\|K(x,\cdot)).$$

For the joint convexity statement, let $E = \{1,2\}$ with the discrete σ-algebra, let $\mu\{1\} = \alpha$, $\mu\{2\} = 1-\alpha$, and set $L(1,\cdot) = q_1$, $L(2,\cdot) = q_2$, $K(1,\cdot) = p_1$, $K(2,\cdot) = p_2$. Then $\mu L = \alpha q_1 + (1-\alpha) q_2$ and $\mu K = \alpha p_1 + (1-\alpha) p_2$, and the general result gives the stated inequality. □

Theorem 3.23. *(Data Processing Inequality)* In the setting of Lemma 3.21,

$$d_\phi(\nu K \| \mu K) \le d_\phi(\nu \| \mu).$$

Proof. Set $L = K$ in Lemma 3.21. Then $\frac{\mathrm{d}L}{\mathrm{d}K} = 1$, so

$$d_\phi(\nu K \| \mu K) \le \int \mu(\mathrm{d}x) \int K(x, \mathrm{d}y)\, \phi\!\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x)\right) = \int \mu(\mathrm{d}x)\, \phi\!\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x)\right) = d_\phi(\nu\|\mu). \qquad \square$$

Corollary 3.24. For any measurable $h : E \to F$,

$$d_\phi(\nu \circ h^{-1} \| \mu \circ h^{-1}) \le d_\phi(\nu \| \mu).$$

Proof. Apply Theorem 3.23 with the kernel $K(x, \mathrm{d}y) = \delta_{h(x)}(\mathrm{d}y)$. □

Proposition 3.25. Let h be a measurable bijection with measurable inverse. Then

$$d_\phi(\nu \circ h^{-1} \| \mu \circ h^{-1}) = d_\phi(\nu \| \mu).$$

Proof. If $\mu \gg \nu$, the change of variables formula and injectivity of h give

$$d_\phi(\nu \circ h^{-1} \| \mu \circ h^{-1}) = \int (\mu \circ h^{-1})(\mathrm{d}x)\, \phi\!\left(\frac{\mathrm{d}(\nu \circ h^{-1})}{\mathrm{d}(\mu \circ h^{-1})}(x)\right) = \int \mu(\mathrm{d}x)\, \phi\!\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x)\right) = d_\phi(\nu\|\mu).$$

If $\mu \not\gg \nu$, there exists A with $\mu(A) = 0$ but $\nu(A) > 0$. Setting $B = h(A)$, injectivity gives $h^{-1}(B) = A$, and measurability of h^{-1} gives $B \in \mathcal{N}$. Therefore $(\mu \circ h^{-1})(B) = 0$ but $(\nu \circ h^{-1})(B) > 0$, so $d_\phi(\nu \circ h^{-1} \| \mu \circ h^{-1}) = \infty = d_\phi(\nu\|\mu)$. □

The behaviour of a ϕ-divergence is determined entirely by the choice of ϕ; functions that grow more quickly for large or small values of t yield more stringent divergences. The most commonly used ϕ-divergences are listed below. We assume $\nu \ll \mu$ throughout for simplicity.

Definition 3.26.

(i) **KL divergence** ($\phi(t) = t \log t$):

$$d_{\mathrm{KL}}(\nu\|\mu) = \int \frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x) \log \frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x)\ \mathrm{d}\mu(x) = \int \log \frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x)\ \mathrm{d}\nu(x).$$

(ii) **Reverse KL divergence** ($\phi(t) = -\log t$): when additionally $\mu \ll \nu$,

$$d_{\mathrm{RKL}}(\nu\|\mu) = -\int \log \frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x)\ \mathrm{d}\mu(x) = \int \log \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x)\ \mathrm{d}\mu(x) = d_{\mathrm{KL}}(\mu\|\nu).$$

(iii) **Squared Hellinger distance** ($\phi(t) = \frac{1}{2}(\sqrt{t}-1)^2$): when additionally $\mu \ll \nu$,

$$d_{H^2}(\nu\|\mu) = 1 - \int \sqrt{\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x)}\ \mathrm{d}\mu(x) = d_{H^2}(\mu\|\nu),$$

so the squared Hellinger distance is symmetric.

(iv) **Total variation** ($\phi(t) = \frac{1}{2}|t-1|$): for any σ-finite $\lambda \gg \mu, \nu$,

$$d_{\mathrm{TV}}(\mu,\nu) = \frac{1}{2}\int \left|\frac{\mathrm{d}\nu}{\mathrm{d}\lambda}(x) - \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}(x)\right| \mathrm{d}\lambda(x).$$

Total variation is the only discrepancy that is both a ϕ-divergence and an IPM.

(v) χ^2 **divergence** ($\phi(t) = (t-1)^2$):

$$d_{\chi^2}(\nu\|\mu) = \int \left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x)\right)^2 \mathrm{d}\mu(x) - 1 = \int \frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x)\ \mathrm{d}\nu(x) - 1.$$

(vi) **Reverse χ^2 divergence** ($\phi(t) = (t-1)^2/t$): when additionally $\mu \ll \nu$,

$$d_{R\chi^2}(\nu\|\mu) = \int \left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x)\right)^2 \mathrm{d}\mu(x) - 1 = d_{\chi^2}(\mu\|\nu).$$

The following results compare these divergences to each other. Together they establish an ordering: the χ^2 divergence is the most stringent, followed by KL, then total variation and squared Hellinger (which are comparable to each other), with all being bounded below by zero.

Proposition 3.27. The following inequalities hold whenever the relevant divergences are finite:

(i) Pinsker's Inequality. $d_{\mathrm{TV}}(\mu,\nu) \le \sqrt{\frac{1}{2} d_{\mathrm{KL}}(\nu\|\mu)}$.

(ii) Gilardoni's Inequality. $-\log(1-d_{\mathrm{TV}}) - (1-d_{\mathrm{TV}})\log(1+d_{\mathrm{TV}}) \le d_{\mathrm{KL}}(\nu\|\mu)$.

(iii) Hellinger–Total Variation. $d_{H^2}(\nu\|\mu) \le d_{\mathrm{TV}}(\mu,\nu) \le \sqrt{d_{H^2}(\nu\|\mu)(2 - d_{H^2}(\nu\|\mu))}$.

(iv) KL–χ^2. $d_{\mathrm{KL}}(\nu\|\mu) \le \log(1 + d_{\chi^2}(\nu\|\mu))$.

Items (i) and (ii) are stated without proof.

Proof. (iii). For the lower bound, note that $(\sqrt{t}-1)^2 \le |t-1|$ for all $t \ge 0$, so $d_{H^2} \le d_{\mathrm{TV}}$ follows directly from the definitions. For the upper bound, write

$$d_{\mathrm{TV}}(\mu,\nu) = \frac{1}{2}\int \mu(\mathrm{d}x)\left|\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x) - 1\right| = \frac{1}{2}\int \mu(\mathrm{d}x)\left|\sqrt{\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x)} - 1\right|\left|\sqrt{\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x)} + 1\right|.$$

Applying the Cauchy–Schwarz inequality,

$$d_{\mathrm{TV}}(\mu,\nu) \le \frac{1}{2}\left(\int \mu(\mathrm{d}x)\left(\sqrt{\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x)} - 1\right)^2\right)^{1/2}\left(\int \mu(\mathrm{d}x)\left(\sqrt{\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x)} + 1\right)^2\right)^{1/2}.$$

The first factor equals $\sqrt{2d_{H^2}(\nu\|\mu)}$. For the second, expanding the square and using $\int \sqrt{\frac{\mathrm{d}\nu}{\mathrm{d}\mu}}\,\mathrm{d}\mu = 1 - d_{H^2}$,

$$\int \mu(\mathrm{d}x)\left(\sqrt{\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x)}+1\right)^2 = \int \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\,\mathrm{d}\mu + 2\int \sqrt{\frac{\mathrm{d}\nu}{\mathrm{d}\mu}}\,\mathrm{d}\mu + 1 = 2(2-d_{H^2}).$$

Combining,

$$d_{\mathrm{TV}}(\mu,\nu) \le \frac{1}{2}\sqrt{2d_{H^2}}\cdot\sqrt{2(2-d_{H^2})} = \sqrt{d_{H^2}(2-d_{H^2})}.$$

(iv). We show the equivalent inequality $e^{d_{\mathrm{KL}}(\nu\|\mu)} \le 1 + d_{\chi^2}(\nu\|\mu)$. Since log is concave, Jensen's inequality gives

$$d_{\mathrm{KL}}(\nu\|\mu) = \int \log\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x)\,\mathrm{d}\nu(x) \le \log\int \frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x)\,\mathrm{d}\nu(x) = \log(1+d_{\chi^2}(\nu\|\mu)). \qquad \square$$

A major practical advantage of ϕ-divergences over integral probability metrics is that they can be estimated from draws from μ without requiring a supremum: given $X_1,\dots,X_N \overset{\text{iid}}{\sim} \mu$,

$$d_\phi(\nu\|\mu) \approx \frac{1}{N}\sum_{n=1}^{N}\phi\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(X_n)\right).$$

The cost is that one must be able to evaluate the density ratio $\mathrm{d}\nu/\,\mathrm{d}\mu$. For the (reverse) KL divergence, it suffices to know the density ratio up to a multiplicative constant, which is useful in Bayesian statistics where posterior densities are typically known only up to proportionality.

References

[1] Gerald B Folland. *Real Analysis: Modern Techniques and Their Applications*, volume 168. John Wiley & Sons, New York, 2nd edition, 1999.