Probability Theory

A measure space such that the measure of the entire space equals 1 is a **probability space**. Hereafter it is understood that a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is fixed once and for all. In the context of probability measures, some mathematical objects are given different names since the subject was developed independently of Lebesgue measure theory. For instance

- A *measurable set* is called an **event**.
- A *measurable function* X is called a **random variable**. If X takes values in $\mathbb{R}$, this is called a **real random variable** or simply a random variable. In general, if $X\colon \Omega \to E$, X is an E**-valued random variable**. It is understood that E is a measurable space.
- The pushforward $\mu := \mathbb{P} \circ X^{-1}$ is called the **law of** X.
- The *almost everywhere* or *a.e.* notation is replaced by **almost sure** or **a.s.** notation.
- The probabilistic notation is in use: $\mathbb{P}\left(\{\omega \in \Omega \,:\, \text{property } p(\omega) \text{ holds}\}\right)$ is denoted $\mathbb{P}\left(p\right)$. For instance $\mathbb{P}(X \geq 5)$ means $\mathbb{P}\left(\{\omega \in \Omega \,:\, X(\omega) \geq 5\}\right)$. This is due to the fact that the probability space is non-canonical, and one usually is interested in properties that hold regardless of the actual choice of the probability space. The measure-theoretic setting is mostly used to ensure the existence of these spaces.
- The integral is called **expected value** or **expectation** and is denoted $\mathbb{E}[X] := \int X(\omega) d\mathbb{P}(\omega)$. For instance, for an E-valued random variable X and $f\colon E \to \mathbb{R}$ such that $f(X) \in L^1(\mathbb{P})$

$$\mathbb{E}[f(X)] = \int_E f d\mu$$

  where μ is the law of X.

Discrete Probabilities

Let $(\Omega, \mathcal{F})$ be a measurable space. A probability μ on Ω is called *discrete* if there exists a countable subset $S \subset \Omega$ such that $\mu(S) = 1$. For a discrete measure it is quite intuitive that it is enough to know the probability of each point $x \in S$ to characterize a probability μ.

Indeed the following result can be easily established.

Proposition 0.1. *There is a one-to-one correspondence between*

a. The set of discrete probabilities $\mathcal{P}_{\text{discrete}}(\Omega)$ on Ω.
b. The maps $\Omega \ni \omega \mapsto \mu_\omega \in [0,1]$, such that $\sum_{\omega\in\Omega} \mu_\omega = 1$ (the convergence of the sum implies that the set $\{\omega \,:\, \mu_\omega > 0\}$ is at most countable).

Such a correspondence is given as follows:

- *For $\mu \in \mathcal{P}_{\text{discrete}}(\Omega)$, set $\mu_\omega := \mu(\{\omega\})$. Then $\omega \mapsto \mu_\omega$ satisfies the hypothesis as in b.*
- *For a map $\mu_\cdot$ as in b., define a probability on Ω setting $\mu(A) := \sum_{\omega\in A} \mu_\omega$.*

It is clear that for countable spaces every probability is discrete. In this case, one usually calls $(\mu_\omega)_\omega$ a probability.

This approach easily extends to atomic probabilities.

Definition 0.1 (atomic probabilities)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A measurable $A \in \mathcal{F}$ is called an **atom** for $\mathbb{P}$ if $\mathbb{P}(A) > 0$ and there exists no measurable $B \subset A$ with $0 < \mathbb{P}(B) < \mathbb{P}(A)$.

$\mathbb{P}$ is called **atomic** if every set of strictly positive measure contains an atom.

Conditional Expectation

For discrete random variables, the meaning of conditional expectations and probability is quite intuitive. If A is an event with $\mathbb{P}(A) \in (0,1)$, and X a real-valued random variable, then

$$\mathbb{E}[X|A] = \mathbb{E}[X\mathbf{1}_A]/\mathbb{P}(A), \qquad \mathbb{E}[X|A^c] = \mathbb{E}[X\mathbf{1}_{A^c}]/\mathbb{P}(A^c) \tag{1}$$

is an intuitive definition.

There is however an important notion that extends this elementary approach. In Equation 1, we can identify the two values given in the formula with a function on Ω that takes one constant value on A and another constant value on A^c. In measure-theoretic terms, this is a function that is measurable w.r.t. the σ-algebra $\mathcal{G}_A := \{\emptyset, A, A^c, \Omega\}$. As we will see, it is convenient to denote this function as follows

$$\mathbb{E}[X|\mathcal{G}_A](\omega) = \begin{cases} \mathbb{E}[X\mathbf{1}_A]/\mathbb{P}(A) & \text{if } \omega \in A \\ \mathbb{E}[X\mathbf{1}_{A^c}]/\mathbb{P}(A^c) & \text{if } \omega \in A^c \end{cases}$$

Then the random variable $Z := \mathbb{E}[X|\mathcal{G}_A]$ has two properties

a. As remarked, Z is $\mathcal{G}_A$-measurable (as a function from Ω to $\mathbb{R}$).
b. For any random variable Y that is also $\mathcal{G}_A$-measurable, say $Y(\omega) = \alpha$ on A and $Y(\omega) = \beta$ on A^c, it holds

$$\mathbb{E}[ZY] = \mathbb{E}[XY] \tag{2}$$

Taking $\alpha = 1$ and $\beta = 0$ and vice versa, it is not hard to check that Z is the only random variable (up to a.s. equivalence) having such properties.

Exercise 0.1. Verify that Equation 2 holds and that Z is uniquely defined up to a.s. equivalence.

These aforementioned properties a. and b. are what allow us to define conditional expectation for any sub-σ-algebra.

Remark. Let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$ and let $X \in L^1(\mathbb{P})$ be a random variable. Then there exists a unique (up to a.s. equivalence) random variable Z such that

a. Z is $\mathcal{G}$-measurable.
b. For all bounded $\mathcal{G}$-measurable random variables Y it holds

$$\mathbb{E}[ZY] = \mathbb{E}[XY] \tag{3}$$

Indeed, consider the (signed) finite measure ν on $(\Omega, \mathcal{G})$ given by $\nu(A) := \mathbb{E}[X\mathbf{1}_A]$. Denote $\mathbb{P}_\mathcal{G}$ the restriction of $\mathbb{P}$ to $\mathcal{G}$.

ν is absolutely continuous w.r.t. $\mathbb{P}_\mathcal{G}$ (since $\nu(A) = 0$ whenever $\mathbb{P}(A) = 0$). By the Radon-Nikodym theorem (applied to the positive and negative parts of X), there exists a unique function $Z\colon \Omega \to \mathbb{R}$, which is $\mathcal{G}$-measurable, such that b. holds.

Definition 0.2. The unique (up to a.s. equivalence) random variable Z defined in the above remark is called the **conditional expectation of** X **given** $\mathcal{G}$ or the **expectation of** X **knowing** $\mathcal{G}$.

Proposition 0.2 (Properties of the conditional expectation)**.** *It holds*

1. *If $\mathcal{H}$ is a sub-σ-algebra of $\mathcal{G}$, then a.s.*
$$\mathbb{E}[\mathbb{E}[X|\mathcal{G}]|\mathcal{H}] = \mathbb{E}[X|\mathcal{H}]$$
In particular taking $\mathcal{H}$ the trivial σ-algebra $\mathbb{E}[\mathbb{E}[X|\mathcal{G}]] = \mathbb{E}[X]$.
2. *If X is independent of $\mathcal{G}$, then $\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X]$.*
3. *If Y is $\mathcal{G}$-measurable, and $XY \in L^1(\Omega)$, then $\mathbb{E}[XY|\mathcal{G}] = \mathbb{E}[X|\mathcal{G}]Y$ a.s..*
4. *Convergence theorems for integrals extend to their conditional version. E.g. if $X_n \uparrow X$ a.s. (monotone convergence), then $\mathbb{E}[X_n|\mathcal{G}] \uparrow \mathbb{E}[X|\mathcal{G}]$.*

Exercise 0.2. Prove Proposition 0.2.

Exercise 0.3. Assume that $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$, and let $\mathcal{G} \subset \mathcal{F}$ be a sub-σ-algebra of $\mathcal{F}$. Prove that $\mathbb{E}[X|\mathcal{G}]$ is the orthogonal projection (in the Hilbert space $L^2(\Omega, \mathcal{F}, \mathbb{P})$) of X on the (closed) subspace $L^2(\Omega, \mathcal{G}, \mathbb{P})$.

Use this fact to give an immediate interpretation of the property 1 in Proposition 0.2.

Topologies on the space of Probability Measures

Let $(E, \mathcal{E})$ be a measurable space. Let's review some common distances on the space $\mathcal{P}(E)$ of probability measures on E

Definition 0.3. For $\mu, \nu \in \mathcal{P}(E)$ define the **total variation** distance as

$$\|\mu - \nu\|_{TV} := \sup_A |\mu(A) - \nu(A)| = \frac{1}{2} \sup_{|f| \leq 1} \int f \, d\mu - \int f \, d\nu$$

where the suprema are taken over measurable events $A \subset E$ and measurable functions $f \colon E \to \mathbb{R}$ with $|f| \leq 1$.

Remark. It holds $\|\mu - \nu\|_{TV} \leq 1$. Moreover if $E = \mathbb{R}$ and $\mu = \varrho dx$ and $\nu = \varrho' dx$, then $\|\mu - \nu\|_{TV} = \frac{1}{2}\|\varrho - \varrho'\|_{L^1}$.

Definition 0.4. Assume that E is a metric space with distance d, and that $\mathcal{E}$ is the associated Borel σ-algebra. For $A \subset E$ and $\varepsilon > 0$, define $A^\varepsilon := \{x \in E \, : \, d(x, A) < \varepsilon\}$. For $\lambda > 0$, the **Lévy-Prokhorov distance** $d_\lambda \colon \mathcal{P}(E) \times \mathcal{P}(E) \to [0, \lambda]$ is defined as

$$d_\lambda(\mu, \nu) := \inf \left\{ \varepsilon > 0 \, : \, \mu(A) \leq \nu(A^\varepsilon) + \frac{\varepsilon}{\lambda}, \nu(A) \leq \mu(A^\varepsilon) + \frac{\varepsilon}{\lambda}, \qquad \text{for all closed } A \subset E \right\}$$

Remark. It holds

1. $d_\lambda(\delta_x, \delta_y) = \min(d(x, y), \lambda)$. In other words if λ is larger than the diameter of E, d_λ is indeed a lift of d to $\mathcal{P}(E)$.
2. If d and d' generate the same topology on E, then d_λ and $d'_{\lambda'}$ generate the same topology on $\mathcal{P}(E)$.
3. If E is a Polish space, $\mathcal{P}(E)$ equipped with the Lévy-Prokhorov distance is a Polish space.

4. If E is a Polish space, a sequence μ_n converges to μ in $\mathcal{P}(E)$ (with the Lévy-Prokhorov distance) iff

$$\lim_n \int f d\mu_n = \int f d\mu$$

for all $f \in C_{\mathrm{b}}(E)$.

5. A subset $\mathcal{K} \subset \mathcal{P}(E)$ is precompact in the Lévy-Prokhorov topology iff[1]

$$\inf_{K \text{compact}} \sup_{\mu \in \mathcal{K}} \mu(K^c) = 0 \tag{4}$$

6. We can rephrase the last point as follows: for (μ_n) a sequence in $\mathcal{P}(E)$, there exists a subsequence n_k and a non-negative Borel measure μ with $\mu(E) \leq 1$ such that $\mu_{n_k}(f) \to \mu(f)$ for every $f \in C_{\mathrm{b}}(E)$. μ is a probability iff for each $\varepsilon > 0$ there exists a compact $K^\varepsilon \subset E$ such that

$$\varlimsup_k \mu_{n_k}(K^\varepsilon) \geq 1 - \varepsilon$$

Entropies

The notion of entropy can be introduced in several different contexts and with slightly different meanings.

Definition 0.5 (Relative entropy)**.** Let $(E, \mathcal{E}, m)$ be a probability space with reference measure m. The **relative entropy** (in Mathematics) or **Kullback–Leibler divergence** (in Computer Science) between μ and m is

$$H(\mu|m) := \sup_f \int f d\mu - \log \int e^f dm$$

where the supremum is taken over all bounded measurable functions $f \colon E \to \mathbb{R}$.

Proposition 0.3. *It holds:*

1. $H(\mu|m) \geq 0$ *and* $H(\mu|m) = 0$ *iff* $\mu = m$.
2. $H(\cdot|\cdot)$ *is jointly convex, namely* $H(\alpha\mu + (1-\alpha)\mu'|\alpha m + (1-\alpha)m') \leq \alpha H(\mu|m) + (1-\alpha)H(\mu'|m')$.
3. *If* E *is a Polish space* $H(\mu|m)$ *is jointly lower semicontinuous in the Lévy-Prokhorov metric.*
4. *For* $h(v) := v \log v$ *(or equivalently* $h(v) = v \log v - v + 1$*)*

$$H(\mu|m) = \begin{cases} \int h(\varrho) dm & \text{if } \mu = \varrho m \text{ (in the sense of Radon-Nikodym)} \\ +\infty & \text{otherwise} \end{cases} \tag{5}$$

5. *If* $m, m' \in \mathcal{P}(E)$*, with* $m = \frac{1}{Z} e^{-V} m'$ *for some measurable* $V \in L^1(\mu)$ *and* $Z > 0$*, then*

$$H(\mu|m) = H(\mu|m') + \int V d\mu + \log Z \tag{6}$$

6. *For each event* $A \subset E$

$$\mu(A) \leq \frac{H(\mu|m) + \log 2}{1 + \log(1/m(A))}$$

[1] The property Equation 4 is usually called **tightness** of the family $\mathcal{K}$ of probabilities.

7. *It holds*

$$\begin{aligned}\|\mu - m\|_{TV}^2 &\leq \tfrac{1}{2} H(\mu|m) \\ \|\mu - m\|_{TV}^2 &\leq 1 - \exp(-H(\mu|m))\end{aligned} \tag{7}$$

Proof. We give a sketch of the proofs. While some of the arguments may sound abstract, they become quite elementary on finite or countable spaces.

1. Take f constant in the definition of H (also follows from the next point).
2. The function $\mathcal{P}(E) \times \mathcal{P}(E) \ni (\mu, m) \mapsto \mu(f) - \log m(e^f) \in \mathbb{R}$ is convex for each f, as the supremum of convex functions is convex.
3. For the sake of simplicity, let's consider the case where E is compact. Then since $C(E)$ is dense in $L^1(\mu)$, it is the same to take the supremum over $f \in C(E)$ in the definition of H. But then the map $\mathcal{P}(E) \times \mathcal{P}(E) \ni (\mu, m) \mapsto \mu(f) - \log m(e^f) \in \mathbb{R}$ is continuous in the Lévy-Prokhorov metric, and the supremum of continuous functions is lower semicontinuous. If E is locally compact, we may replace $C(E)$ with functions with compact support. In the general case, we may replace $C(E)$ with functions that are uniformly continuous w.r.t. to a fixed totally bounded metric on E (such a metric always exists on Polish spaces).
4. If there exists a set A such that $\mu(A) > 0$ but $m(A) = 0$, then for $c > 0$ take $f = c\mathbf{1}_A$. We get

$$H(\mu|m) \geq c\mu(A) - \log(e^c m(A) + e^0 m(A^c)) = c\mu(A)$$

   Since this holds for any $c > 0$, $H(\mu|m) = +\infty$. If such a set A does not exist, by Radon-Nikodym, we can assume that $\mu = \varrho m$ for some $\varrho \in L^1(m)$. First assume that ϱ is bounded and bounded away from 0. Then take $f = \log \varrho + g$ for some arbitrary g bounded measurable, to get

$$\begin{aligned}H(\mu|m) &= \sup_g \int \log \varrho d\mu + \int g d\mu - \log \int e^{\log \varrho + g} dm \\ &= \int h(\varrho) dm + \sup_g \int g d\mu - \log \int e^g d\mu\end{aligned}$$

   The last supremum is non-positive by Jensen inequality, so the $\sup_g$ equals 0 (achieved for g constant). It is then easy to adapt the argument when $\log \varrho$ is unbounded.
5. Straightforward from the properties of the logarithm and the chain rule.
6. Take $f = c\mathbf{1}_A$ and optimize over $c > 0$.
7. This is the Pinsker inequality, which can be proved by elementary methods but is beyond the scope of this note.

□

Relation to the classical entropy

Remark. On a finite space E, one may define $\mathrm{Ent}(\mu) := \sum_{x \in E} \mu_x \log \mu_x$. Notice that, in the relative entropy notation, this is nothing but

$$\mathrm{Ent}(\mu) = H(\mu|m') - \log(|E|)$$

where m' is the uniform probability on E, namely $m'_x = 1/|E|$ for all $x \in E$. From Equation 6, if $m_x = e^{-V(x)}/Z$ for some $V : E \to \mathbb{R} \cup \{+\infty\}$ and $Z = \sum_x e^{-V(x)}$, it follows

$$H(\mu|m) = \mathrm{Ent}(\mu) + \int V d\mu + \log(Z/|E|)$$

In the physical literature, $S(\mu) := -\mathrm{Ent}(\mu)$ is called the entropy, $V(x)$ is interpreted as $V(x) = \beta h(x)$ where $h(x)$ is the energy of the configuration of the state $x \in E$, $\beta = 1/(\kappa T)$ where κ is a universal constant (the *Boltzmann*

constant), and T is the *temperature*. $\int h\mu$ is interpreted as the *energy* of the "state" μ. So that one defines the *free energy*

$$F(\mu) = \text{energy - } \kappa T \text{ entropy } = \int h d\mu - \kappa T S(\mu) = \tfrac{1}{\beta} H(\mu|m) - \tfrac{1}{\beta}\log(Z/|E|)$$

In particular the statement "$\mu \mapsto H(\mu|m)$ is minimized at $\mu = m$" is equivalently rephrased as "the measure $m = e^{-h/(\kappa T)}/Z$ minimizes the free energy". Unfortunately the different notations and nomenclature persist to this day, and as a general rule:

- $H(\mu|m)$ is used in the Probability literature, regardless of the space.
- $D_{KL}(\mu||m)$ is used in the Computer Science literature. This coincides with $H(\mu|m)$.
- $F(\mu)$ is used in the Physical literature. This differs by some constants from $H(\mu|m)$, which are not relevant for fixed β or h (e.g. if we look at these as functions of μ), but are relevant if we consider $H(\mu|m)$ as a function of both μ and m.

Large Deviations

In this section we review some basic concentration properties of sequences of probability measures. As above, E is a Polish space, equipped with its Borel σ-algebra.

Definition 0.6. A function $I\colon E \to (-\infty, \infty]$ is **lower semicontinuous** (or **lsc**) if for all $c \in \mathbb{R}$ the set $\{I \leq c\}$ is closed.

I is called **coercive** if $\{I \leq c\}$ is either empty or precompact.

In particular, if I is lsc and coercive, then it admits a minimum on E[2].

Proposition 0.4. *Let (μ_n) be a sequence of probability measures on E, and $\mathbf{a} = (a_n)$ a sequence of reals with $\lim_{n\to\infty} a_n = +\infty$. Let $B_\varepsilon(x)$ be the ball of radius ε centered at x. Define*

$$\underline{I}(x) \equiv \underline{I}^{\mathbf{a}}(x) := -\lim_{\varepsilon\to 0} \varliminf_n \frac{1}{a_n}\log \mu_n(B_\varepsilon(x)) \in [0,\infty]$$
$$\overline{I}(x) \equiv \overline{I}^{\mathbf{a}}(x) := -\lim_{\varepsilon\to 0} \varlimsup_n \frac{1}{a_n}\log \mu_n(B_\varepsilon(x)) \in [0,\infty]$$

Then:

1. *$\underline{I}$ and $\overline{I}$ are lsc.*
2. *$\underline{I}$ is the optimal (i.e. the largest) lsc function, and $\overline{I}$ is the optimal (i.e. smallest) function, such that the following inequalities hold*
$$\mu_n(K) \leq \exp\left(-a_n \inf_{x\in K} \underline{I}(x) + o(a_n)\right), \qquad \text{for all } K \subset E$$
$$\mu_n(O) \geq \exp\left(-a_n \inf_{x\in O} \underline{I}(x) + o(a_n)\right), \qquad \text{for all } O \subset E$$
3. *Equivalently, they are the optimal (lsc) functions such that for each $f \in C_b(E)$*
$$\mu_n(e^{a_n f}) \leq \exp\left(a_n \sup_x (f(x) - \underline{I}(x)) + o(a_n)\right)$$
$$\mu_n(e^{a_n f}) \geq \exp\left(a_n \sup_x (f(x) - \overline{I}(x)) + o(a_n)\right)$$

[2]This easy-to-prove fact, known as Bolzano-Weierstrass theorem, is usually discussed in calculus classes, and the rigor of the proof has played an important role to inspire modern Mathematics.

Remark. Let $x \in E$. For each sequence $\nu_n \to \delta_x$ it holds

$$\underline{\lim}_n \frac{1}{a_n} H(\nu_n|\mu_n) \geq \underline{I}^{\mathbf{a}}(x)$$

Moreover there exists a sequence $\nu_n \to \delta_x$ such that

$$\underline{\lim}_n \frac{1}{a_n} H(\nu_n|\mu_n) \leq \overline{I}^{\mathbf{a}}(x)$$

$\underline{I}^{\mathbf{a}}(x), \overline{I}^{\mathbf{a}}(x)$ are the optimal functions for which these two statements hold.

Large Deviations should be compared with the narrow convergence Definition 0.4, in which convergence is substantially equivalent to $\mu_n(f) \to \mu(f)$ for $f \in C_{\mathrm{b}}(E)$ and similar inequalities hold on open and closed sets. Informally speaking, convergence of probability measures corresponds to the case $a_n = 1$ of Proposition 0.4.

Definition 0.7. Let (μ_n), (a_n) be as in Proposition 0.4. (μ_n) satisfies a **Large Deviations Principle** with speed a_n and rate function $I\colon E \to [0,\infty]$ if $\underline{I} = \overline{I} =: I$. The Large Deviations Principle is **non-trivial** if there exists $x \in E$ such that $I(x) \in (0,\infty)$.

Proposition 0.5. *Let (μ_n), (a_n) be as in Proposition 0.4. There exists a subsequence n_k along which a Large Deviations Principle holds.*

Inequalities

We list some notable inequalities here. The proofs are elementary, except for the Ahlswede-Daykin inequality on product spaces.

Markov inequalities

Proposition 0.6 (Markov inequality)**.** *Let X be a real random variable, then for $c > 0$*

$$\mathbb{P}[X \geq c] \leq \mathbb{E}[|X|]/c$$

While this appears to be a crude inequality, we can apply it to any non-decreasing function $\varphi\colon \mathbb{R} \to \mathbb{R}^+$ to get for any X and $c \in \mathbb{R}$

$$\mathbb{P}[X \geq c] \leq \mathbb{P}[\varphi(X) \geq \varphi(c)] \leq \mathbb{E}[\varphi(X)]/\varphi(c)$$

On the other hand, the latter is trivially an equality for $\varphi(x) = \mathbf{1}_{(\infty,c]}(x)$. This entails the following stronger statement

Proposition 0.7 (Markov equality)**.** *For any real random variable and $c \in \mathbb{R}$ it holds*

$$\mathbb{P}[X \geq c] = \inf_{\varphi} \mathbb{E}[\varphi(X)]/\varphi(c)$$

where the infimum is taken over all non-decreasing $\varphi\colon \mathbb{R} \to [0,\infty)$ with $\varphi(c) > 0$.

In particular

- *taking $\varphi(x) = |x - \mathbb{E}[X]|$ we get the **Chebyshev inequality***

$$\mathbb{P}[|X - \mathbb{E}[X]| \geq c] \leq \mathrm{Var}[X] c^{-2}$$

- *taking $\varphi(x) = e^{\lambda x}$ for $\lambda > 0$, we get the* ***Chernoff inequality***

$$\mathbb{P}[X \geq c] \leq \exp(-\psi(c))$$

where ψ is given by the Legendre duality formula $\psi(c) := \sup_{\lambda \geq 0} \lambda c - \log \mathbb{E}[e^{\lambda X}] \in [0, +\infty]$.

Jensen inequalities

Since a convex function is a supremum of affine functions, exchanging the supremum with the expected value we get the famous Jensen inequality, which holds a rather general linear space.

Proposition 0.8 (Jensen inequality). *If f is a convex function, X a random variable such that $f(X) \in L^1(\mathbb{P})$, it holds*

$$\mathbb{E}[f(X)|\mathcal{G}] \geq f\left(\mathbb{E}[X|\mathcal{G}]\right)$$

Hölder inequalities

Proposition 0.9 (Hölder inequality). *If $p_1, \dots, p_n, q \in [1, \infty]$ are such that $\sum_i 1/p_i \leq 1/q$ then*

$$\mathbb{E}[|X_1 \cdots X_n|^q]^{1/q} \leq \mathbb{E}[|X_1|^{p_1}]^{1/p_1} \dots \mathbb{E}[|X_n|^{p_n}]^{1/p_n}$$

Kochen-Stone lemma

This is the counterpart of the Borel-Cantelli lemma

Proposition 0.10 (Kochen-Stone Lemma). *Let (A_n) be a sequence of events such that*

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$$

Then

$$\mathbb{P}(\limsup_{n\to\infty} A_n) \geq \limsup_{k\to\infty} \frac{\left(\sum_{n=1}^{k} \mathbb{P}(A_n)\right)^2}{\sum_{1\leq m,n\leq k} \mathbb{P}(A_m \cap A_n)} \tag{8}$$

In particular, if the (A_n) are pairwise independent (or A_n is independent of A_m for all but finitely many m), then under Equation 8 $\mathbb{P}(\limsup_{n\to\infty} A_n) = 1$.

Correlation inequalities

In this section we introduce two classes of non-trivial correlation inequalities: those in the GKS/Ginibre class, and those in the FKG class.

Ginibre inequality

Definition 0.8 (Convex cone)**.** A subset C of a real vector space is called a (blunt) **convex cone** if it is closed under linear combinations with non-negative coefficents: if $u, v \in C$, then $\alpha u + \beta v \in C$ for $\alpha, \beta \geq 0$.

The smallest convex cone containing a subset A of the vector space is called the **convex cone generated by** A (this is well-defined as the intersection of all convex cones containing A)

Definition 0.9. Let $A \subset L^1(\mathbb{P})$ be a set of integrable random variables. We say that A satisfies the **Ginibre condition** if for each $N \geq 1$, $X_1, \dots, X_N \in A$, and $\epsilon_1, \dots, \epsilon_N \in \{-1, +1\}$

$$\int \prod_{i=1}^{N} \left(X_i(\omega) + \epsilon_i X_i(\omega')\right) \mathbb{P}(d\omega)\mathbb{P}(d\omega') \geq 0$$

In other words, if $(Y_1, \dots, Y_N)$ is an independent copy of $(X_1, \dots, X_N)$ then $\mathbb{E}[(X_1 \pm Y_1) \cdots (X_N \pm Y_N)] \geq 0$, regardless of the signs $\pm$ in each factor.

Theorem 0.1 (Ginibre inequality)**.** *Let $A \subset L^1(\mathbb{P})$ be a set of integrable random variables satisfying the Ginibre condition. If X, Y, H are random variables in the convex cone generated by A, and $e^{-H}, Xe^{-H}, Ye^{-H} \in L^1(\mathbb{P})$, then*

$$\mathbb{E}\left[XYe^{-H}\right]\mathbb{E}\left[e^{-H}\right] \geq \mathbb{E}\left[Xe^{-H}\right]\mathbb{E}\left[Ye^{-H}\right]$$

General AD and FKG inequalities

To properly define a general version of the AD, Holley and FKG inequalities, we first recall the notion of distributive lattice.

Definition 0.10 (Measurable distributive lattice)**.** A partial order relation $\preccurlyeq$ defined on a measurable space $(\Omega, \mathcal{G})$ is **measurable** if the set $\{(x, y) \,:\, x \preccurlyeq y\}$ is measurable.

A set Ω equipped with a partial order relation $\preccurlyeq$ is a **distributive lattice** if for any elements $x, y, z \in \Omega$:

a. There exist a unique greatest lower bound (meet) $x \wedge y$ and a unique least upper bound (join) $x \vee y$ (**lattice property**).
b. The operations are **distributive**: $x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z)$.

A **measurable distributive lattice** is a triple $(\Omega, \mathcal{F}, \preccurlyeq)$ where $(\Omega, \mathcal{F})$ is a measurable space and $\preccurlyeq$ is a measurable partial order that defines a distributive lattice structure on Ω.

Definition 0.11 (Correlation inequalities)**.** Let μ be a σ-finite measure on a measurable distributive lattice $(\Omega, \mathcal{F}, \preccurlyeq)$.

- We say that μ satisfies the **Ahlswede-Daykin inequality** if, for all measurable $f_1, f_2, f_3, f_4 \colon \Omega \to [0, \infty]$ such that for $x, y \in \Omega$

  $$f_1(x \vee y) f_2(x \wedge y) \geq f_3(x) f_4(y) \qquad x, y \in \Omega$$

  it holds

  $$\mu(f_1)\mu(f_2) \geq \mu(f_3)\mu(f_4) \tag{9}$$

- We say that μ satisfies the **Holley inequality** if, for all measurable functions $h, g_1, g_2 \colon \Omega \to [0, \infty]$ such that h is non-decreasing and

  $$g_1(x \vee y) g_2(x \wedge y) \geq g_1(x) g_2(y) \qquad x, y \in \Omega$$

  it holds

  $$\mu(h\, g_1)\mu(g_2) \geq \mu(g_1)\mu(h g_2) \tag{10}$$

- If μ is a probability, we say that μ satisfies the **FKG inequality** if for all measurable non-decreasing functions $f, g \colon \Omega \to [0, \infty]$

$$\mu(fg) \geq \mu(f)\mu(g)$$

Proposition 0.11. *If a σ-finite measure μ satisfies the Ahlswede-Daykin inequality, then it satisfies the Holley inequality.*

If a probability μ satisfies the Holley inequality, then it satisfies the FKG inequality.

Proof. For the first statement, take in the Equation 9, $f_1 = hg_1$, $f_2 = g_2$, $f_3 = g_1$, $f_4 = hg_2$. It is easy to see that these satisfy the conditions for the Ahlswede-Daykin inequality, since $h(x \vee y) \geq h(y)$. But for such a choice of the four functions, the Ahlswede-Daykin inequality reduces to the Holley's inequality.

For the second statement, we can assume $f \in L^1(\mu)$ up to a simple approximation argument. Then take in Equation 10 $g_1 = f$, $g_2 = 1$, $h = g$. □

A special class of measurable distributive lattices are product lattices. Suppose that, for t in some arbitrary index set T, $(\Omega_t, \mathcal{F}_t, \preccurlyeq_t)$ is a measurable distributive lattice and assume that $\preccurlyeq_t$ is a **total** order relation. Then the product space $\Omega = \prod_{t \in T} \Omega_t$ is naturally equipped with a partial order relation: $\omega \preccurlyeq \omega'$ iff $\omega_t \preccurlyeq_t \omega'_t$ for all $t \in T$. In this case we say that Ω is a measurable **product** distributive lattice.

The main statement of the next theorem is proved in (Batty and Bollmann 1980).

Theorem 0.2 (General Ahlswede-Daykin theorem)**.** *Any product measure on a product distributive lattice satisfies the Ahlswede-Daykin inequality.*

In particular, since any finite distributive lattice can be regarded as a sublattice of a (finite) product distributive lattice, we have that the counting measure over a finite distributive lattice satisfies the Ahlswede-Daykin inequality.

Batty, CJK, and HW Bollmann. 1980. "Generalised Holley-Preston Inequalities on Measure Spaces and Their Products." *Zeitschrift für Wahrscheinlichkeitstheorie Und Verwandte Gebiete* 53 (2): 157–73.