# Probability Theory

Kexing Ying

January 18, 2022

## Contents

# 1 Review of Measure Theory

Modern probability theory is based on measure theory and we will in this section recall some notions from measure theory.

**Definition 1.1** (Algebra)**.** Given a set $\Omega$, a set of subsets $\mathcal{A}$ of $\Omega$ is an algebra if $\Omega \in \mathcal{A}$ and $\mathcal{A}$ is closed under finite union and complements.

It follows straight away that an algebra is also closed under finite intersections.

**Definition 1.2** (Finitely Additive Measure)**.** A function $\mu : \mathcal{A} \to [0, \infty]$ where $\mathcal{A}$ is an algebra, is a finitely additive measure if for any disjoint sets $A, B \in \mathcal{A}$,

$$\mu(A \cup B) = \mu(A) + \mu(B).$$

**Definition 1.3** ($\sigma$-Algebra)**.** A $\sigma$-algebra $\mathcal{F}$ is an algebra that is closed under countable unions.

Similarly, it follows that $\mathcal{F}$ is closed under countable intersections.

**Definition 1.4** (Measure)**.** A function $\mu : \mathcal{F} \to [0, \infty]$ where $\mathcal{F}$ is a $\sigma$-algebra, is a $\sigma$-additive measure (or simply measure) if given a sequence of pairwise disjoint sets $A_1, A_2, ...$ of $\mathcal{F}$, we have

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_n).$$

We call a measure a probability measure if $\mu(\Omega) = 1$.

**Definition 1.5** ($\sigma$-Finite Measure)**.** A measure $\mu$ is said to be $\sigma$-finite if there exists a sequence of pairwise disjoint sets $A_1, A_2, ...$ of $\mathcal{F}$, such that $\bigcup_{i=1}^{\infty} A_i = \Omega$ and for all $i$, $\mu(A_i) < \infty$.

**Definition 1.6** (Probability Space)**.** A probability space is the triple $(\Omega, \mathcal{F}, \mathbb{P})$ consisting of a set $\Omega$, a $\sigma$-algebra $\mathcal{F}$ on $\Omega$ and $\mathbb{P}$ a probability measure on $\mathcal{F}$.

We call elements of $\mathcal{F}$ (i.e. a $\mathcal{F}$-measurable set) an event.

**Proposition 1.1** (Continuity of Measures)**.** Let $(A_n)_{n \in \mathbb{N}} \subseteq \mathcal{F}$, then

- (continuity from below) if $(A_n)$ is increasing, then

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n \to \infty} \mathbb{P}(A_n).$$

- (continuity from above) if $(A_n)$ is decreasing, then

$$\mathbb{P}\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n \to \infty} \mathbb{P}(A_n).$$

We recall the the finiteness of the measure is vital for continuity from below while continuity from above is also valid for general measures.

*Proof.* Exercise. □

**Proposition 1.2.** A finitely additive probability measure on the $\sigma$-algebra $\mathcal{F}$ is a probability measure if and only if it is continuous at 0.

*Proof.* The forward direction follows from above so we will prove the reverse. Suppose $\mu$ is finitely additive and for any decreasing $(A_n) \subseteq \mathcal{F}$ with $\bigcap A_n = \emptyset$, we have $\lim_{n\to\infty} \mu(A_n) = 0$. Then, $\mu$ is continuous from below, and so for any sequence of disjoint sets $(B_n)$, we have $(C_n) := (\bigcup_{i=1}^n B_i)$ is a sequence of increasing sets and thus,

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} B_i\right) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} C_i\right) = \lim_{n\to\infty} \mathbb{P}(C_n) = \lim_{n\to\infty} \mathbb{P}\left(\bigcup_{i=1}^{n} B_i\right) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{P}(B_i)$$

implying $\mu$ is $\sigma$-additive and so, $\mu$ is a measure. □

**Proposition 1.3.** Given a collection $\{\mathcal{F}_i\}_{i\in I}$ $\sigma$-algebras of $\Omega$, $\bigcap_{i\in I} \mathcal{F}_i$ is also a $\sigma$-algebra on $\Omega$.

**Definition 1.7** ($\sigma$-Algebra Generated By Sets)**.** Given a collection of subsets $S$ of $\Omega$, the $\sigma$-algebra generated by $S$ is

$$\sigma(S) := \bigcap \{\mathcal{F} \text{ a } \sigma\text{-algebra} \mid S \subseteq \mathcal{F}\}.$$

**Definition 1.8** (Borel $\sigma$-Algebra)**.** Given a topological space $(X, \mathcal{T})$, the Borel $\sigma$-algebra on $X$ is $\mathcal{B}(X) := \sigma(\mathcal{T})$.

**Definition 1.9** (Product $\sigma$-Algebra)**.** Given measurable spaces $(\Omega_1, \mathcal{F}_1), (\Omega_2, \mathcal{F}_2)$, the product $\sigma$-algebra on $\Omega_1 \times \Omega_2$ is

$$\mathcal{F}_1 \otimes \mathcal{F}_2 := \sigma(\mathcal{F}_1 \times \mathcal{F}_2) = \sigma(\{A_1 \times A_2 \mid A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2\}).$$

**Definition 1.10** (Cylindrical $\sigma$-Algebra)**.** A set $C \subseteq \mathbb{R}^\infty$ is said to be cylindrical if is of the form

$$C = \{x \in \mathbb{R}^\infty \mid (x_1, \cdots, x_n) \in C_n\}$$

where $C_n \in \mathcal{B}(\mathbb{R}^n)$. The set of cylindrical sets $\mathcal{B}(\mathbb{R}^\infty)$ form a $\sigma$-algebra on $\mathbb{R}^\infty$ and is called the cylindrical $\sigma$-algebra.

**Definition 1.11** (Consistent)**.** The sequence of measures $\mathbb{P}_n$ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ is said to be consistent if for all $n \in \mathbb{N}$, $\mathbb{P}_{n+1}(B_n \times \mathbb{R}) = \mathbb{P}_n(B_n)$ for all $B_n \in \mathcal{B}(\mathbb{R}^n)$.

**Theorem 1** (Kolmogorov)**.** Given any consistent sequence of measures $\mathbb{P}_n$, there exists a unique probability measure $\mathbb{P}$ on $(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty))$ such that,

$$\mathbb{P}(\{x \in \mathbb{R}^\infty \mid (x_1, \cdots, x_n) \in \mathcal{B}_n\}) = \mathbb{P}_n(B_n)$$

for all $n \geq 1$, $B_n \in \mathcal{B}(\mathbb{R}^n)$.

*Proof.* Simply define the inner measure on the generating sets as claimed and use the Caratheodory extension (which provides both existence and uniqueness). □

Recall that a nondecreasing function $g$ on $\mathbb{R}$ is continuous up to possibly countably many discontinuities of the first kind. Furthermore, the derivative $g'$ exists $\lambda$-a.e. (where $\lambda$ is the Lebesgue measure on $\mathbb{R}$.

**Proposition 1.4.** Let $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P})$ be a probability space. Defining $F(x) := \mathbb{P}(-\infty, x]$, we have

- $F$ is nondecreasing;
- $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$;
- $F$ is continuous on the right.

*Proof.* Clear by the monotonicity, continuity of measures (from above). □

**Definition 1.12** (Distribution Function)**.** Any function $F : \mathbb{R} \to [0, 1]$ satisfying the above three properties is said to be a distribution function on $\mathbb{R}$.

It is clear that any probability measure induces a distribution. On the other hand the converse is also true.

**Proposition 1.5.** Given a distribution function $F$, there exists a unique probability measure $\mathbb{P}$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $F(x) = \mathbb{P}(-\infty, x]$ for all $x \in \mathbb{R}$.

*Proof.* Use Caratheodory extension theorem on the algebra $\{(-\infty, x] \mid x \in \mathbb{R}\}$ mapping $(-\infty, x] \mapsto F(x)$. The uniqueness of the probability measure follows by the uniqueness of the Caratheodory extension. □

**Definition 1.13** (Null-set)**.** Given a measure $\mu$, a set $S \subseteq \Omega$ is a null-set if there exists some measurable set $N \subseteq \Omega$ with measure 0 such that $S \subseteq N$.

**Definition 1.14** (Complete Measure)**.** A measure $\mu$ is complete if every $\mu$-null set is measurable.

If a measure on the $\sigma$-algebra $\Sigma$ is not complete, we may complete the $\sigma$-algebra by extending $\Sigma$ to

$$\overline{\Sigma} := \sigma(\Sigma \cup \{N \mid N \text{ is a null-set}\}).$$

Clearly, the null-sets will have measure 0.

We note that the probability space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P})$ is not complete as there exists subsets of a Borel null-set which are not Borel. With this in mind, we denote the completion of $\mathcal{B}(\mathbb{R})$ by $\mathcal{M}(\mathbb{R})$ and we say $\mathbb{P}$ is the Lebesgue-Stieltjes measure.

Recall, that in elementary probability theory, we considered three types of distributions, namely discrete, absolutely continuous and singular continuous. Let us now consider them again in a formal measure theoretic setting.

- (Discrete) A random variable $X : \Omega \to \mathbb{R}$ is said to have discrete distribution if there exists some countable (including finite) set $A \subseteq \mathbb{R}$, such that for all $E \in \mathcal{B}(\mathbb{R})$, the push-forward measure satisfies

$$X_*\mathbb{P}(E) = \sum_{x\in A} p(x)\delta_x(E)$$

where $p(x) = X_*\mathbb{P}(\{x\})$ and $\delta_x$ is the Dirac measure at $x$.

We note that a distribution function $F$ corresponds to a discrete random variable if and only if for all $x_0 \in \mathbb{R}$,

$$F(x_0) = \sum_{x \in A \cap \{\leq x_0\}} p(x).$$

It is clear that $\sum_A p(x) = 1$ since $\sum_A p(x) = X_*\mathbb{P}(\mathbb{R}) = 1$.

- (Absolutely continuous) A random variable $X : \Omega \to \mathbb{R}$ is said to be absolutely continuous if $X_*\mathbb{P} = f\lambda$ for some Lebesgue integrable function $f$ and $\lambda$ denotes Lebesgue measure. Thus, the distribution function corresponding to $X$ satisfies

$$F(x) = \int_{(-\infty, x]} f \mathrm{d}\lambda.$$

In particular, recalling the Radon-Nikodym theorem, we have $X$ is absolutely continuous if and only if $X_*\mathbb{P} \ll \lambda$ (hence the name "absolutely continuous").

Before introducing the last type of distribution, let us consider the following definition.

**Definition 1.15** (Concentrated)**.** A measure $\mu$ on the measurable space $X$ is said to be concentrated on a measurable set $A$ if $\mu(E) = 0$ for all $E \subseteq X \setminus A$.

- (Singular continuous) A random variable $X : \Omega \to \mathbb{R}$ is singular continuous if its distribution function $F$ is continuous and $X_*\mathbb{P}$ is concentrated on a set $A$ of Lebesgue measure 0 for which $F'(x) = 0$ for all $x \in A$ almost everywhere.

We note that, since $\lambda(A) = 0$, $X_*\mathbb{P} \perp \lambda$ by the set $A$ (hence the name "singular"). Moreover, by continuity, $X_*\mathbb{P}(\{x\}) = 0$ for all $x \in \mathbb{R}$ in contrast to the discrete measure.

Analogous to the Lebesgue decomposition of measures, we may decompose any distribution function in to a discrete, absolutely continuous and singular continuous distribution.

**Theorem 2** (Hahn Decomposition for Distributions)**.** Given a distribution function $F$, there exists $a_1 + a_2 + a_3 = 1$ and $F_{\text{disc}}, F_{\text{ac}}, F_{\text{sc}}$ discrete, absolutely continuous and singular continuous distribution functions respectively, such that

$$F = a_1 F_{\text{disc}} + a_2 F_{\text{ac}} + a_3 F_{\text{sc}}.$$

*Proof.* Recalling the refinement of the Lebesgue decomposition where we may decompose a measure $\mu$ with

$$\mu = \mu_d + \mu_a + \mu_s$$

where $\mu_d$ is a discrete measure, $\mu_a \ll \lambda$ and $\mu_s$ is singular continuous (i.e. mutually singular with respect to the Lebesgue measure and $\mu_s\{x\}$ for all $x$). Thus, by simply taking the the decomposition of the measure corresponding to $F$ (i.e. $X_*\mathbb{P}$), we obtain the required decomposition after normalization. $\square$

## 1.1 Random Variables

We will continue to let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.

**Definition 1.16** (Random Variable)**.** A function $\xi : \Omega \to \mathbb{R}$ is said to be a random variable if it is $\mathcal{F}$-measurable (i.e. for any $B \in \mathcal{B}(\mathbb{R})$, we have $\xi^{-1}(B) \in \mathcal{F}$).

While we have already introduced the notion of a distribution within the previous section, we will present it here again for organization.

**Definition 1.17** (Distribution of a Random Variable)**.** Given a random variable $\xi$, the distribution of $\xi$ is the push-forward measure of $\mathbb{P}$ along $\xi$. Furthermore, the distribution function corresponding to $\xi$ is

$$F(x) := X_* \mathbb{P}(-\infty, x].$$

**Definition 1.18.** Given a random variable $\xi$, we define $\mathcal{F}_\xi \subseteq \mathcal{F}$ to be the $\sigma$-algebra

$$\mathcal{F}_\xi := \{\xi^{-1}(B) \mid B \in \mathcal{B}(\mathbb{R})\}.$$

This is the least $\sigma$-algebra for which $\xi$ is measurable.

We will recall some standard results about measurable functions. All proofs are left as exercises and can be found in the second year measure theory notes.

**Lemma 1.1.** If $\mathcal{B}(\mathbb{R}) = \sigma(\mathcal{D})$ for a collection of sets $\mathcal{D}$, $\xi$ is a random variable if $\xi^{-1}(D) \in \mathcal{F}$ for all $D \in \mathcal{D}$.

**Lemma 1.2.** Given random variables $f, g$ and $c \in \mathbb{R}$, $f + g, f - g, c \cdot f, |f|, fg$, $\max(f, g)$, and $\min(f, g)$ are all random variables. Furthermore, if $g(x) \neq 0$ for all $x$, then $f/g$ is also a random variable.

**Lemma 1.3.** If $(f_n)$ is a sequence of random variables, then

$$\sup_n f_n, \inf_n f_n, \lim_n f_n$$

are random variables if they exist.

**Lemma 1.4.** If $\xi$ is a random variable and $f : \mathbb{R} \to \mathbb{R}$ is continuous, then $f(\xi)$ is a random variable.

**Definition 1.19** (Simple Function)**.** A random variable $\xi$ is simple if there exists a partition of $\Omega$, $D_1, \cdots, D_n$ such that

$$\xi(\omega) = \sum_{i=1}^{n} x_i 1_{D_i}(\omega)$$

for some $x_1, \cdots, x_n$ for all $\omega \in \Omega$.

**Lemma 1.5.** For any non-negative random variable $\xi$, there exists a sequence of nondecreasing simple random variables $(\xi_n)$ such that for all $\omega \in \Omega$,

$$\xi_n(\omega) \uparrow \xi(\omega).$$

**Definition 1.20** (Random Vector)**.** A function $\xi : \Omega \to \mathbb{R}^n$ is a random vector if it is measurable. Again, we define its distribution to be its push-forward measure.

**Lemma 1.6.** $\xi : \Omega \to \mathbb{R}^n$ is a random vector if and only if $\xi_i := \mathrm{pr}_i \circ \xi$ is a random variable for all $i = 1, \cdots, n$ (where $\mathrm{pr}_i : \mathbb{R}^n \to \mathbb{R}$ is the $i$-th projection function).

**Definition 1.21** (Independent Random Variables)**.** Two random variables $\xi, \eta : \Omega \to \mathbb{R}$ are said to be independent if

$$(\xi, \eta)_* \mathbb{P} = \xi_* \mathbb{P} \otimes \eta_* \mathbb{P}.$$

Since, to check that two measures are equal, it suffices to check equality on generating sets, $\xi, \eta$ are independent if for all $A, B \in \mathcal{B}(\mathbb{R})$,

$$\mathbb{P}(\xi \in A, \eta \in B) = \mathbb{P}(\xi \in A)\mathbb{P}(\eta \in B).$$

Let us quickly recall the construction of the Lebesgue integral.

1. Define the Lebesgue integral for simple functions.
2. Define the Lebesgue integral for non-negative functions by taking the limit of the Lebesgue integral of the monotone sequence of simple functions which converge to the said function.
3. Define the Lebesgue integral for general real-valued functions $f$ by taking $\int f = \int f^+ - \int f^-$ if $\int |f| < \infty$.

**Definition 1.22** (Expectation)**.** Given a random variable $\xi : \Omega \to \mathbb{R}$, the expectation of $\xi$ is simply

$$\mathbb{E}(\xi) := \int \xi \mathrm{d}\mathbb{P}$$

if it exists. Furthermore, we say $\xi$ is integrable if $\mathbb{E}(|\xi|) < \infty$.

**Proposition 1.6.** Let $\xi, \eta$ be integrable random variables and let $c \in \mathbb{R}$, then

- $\mathbb{E}(c) = c$;
- $\mathbb{E}(\xi + \eta) = \mathbb{E}(\xi) + \mathbb{E}(\eta)$;
- $\xi \leq \eta$ a.e. implies $\mathbb{E}(\xi) \leq \mathbb{E}(\eta)$;
- $\xi = \eta$ a.e. implies $\mathbb{E}(\xi) = \mathbb{E}(\eta)$;
- $\xi \geq 0$ a.e. and $\mathbb{E}(\xi) = 0$ implies $\xi = 0$ a.e.

*Proof.* Follows directly from the properties of the Lebesgue integral. □

Let us recall some convergence theorems for the Lebesgue integral.

**Theorem 3** (Dominated Convergence Theorem)**.** Let $(\xi_n)$ be a sequence of random variables such that $\xi_n \to \xi$ almost everywhere. If there exists some integrable $\eta$ such that $|\xi_n| \leq \eta$ for all $n$, then, $\xi$ is integrable and

$$\lim_{n \to \infty} \mathbb{E}(\xi_n) = \mathbb{E}(\xi).$$

**Theorem 4** (Monotone Convergence Theorem)**.** Let $(\xi_n)$ be a sequence of non-negative increasing random variables. Then,

$$\lim_{n \to \infty} \mathbb{E}(\xi_n) = \mathbb{E} \lim_{n \to \infty} \xi_n.$$

We note that the right hand side limit always exists since for all $\omega \in \Omega$, $\xi_n(\omega)$ is increasing any bounded by $\infty$.

We remark that the monotone convergence theorem applies if there exists some random variable $\eta$ such that $\mathbb{E}(\eta) > -\infty$ such that $\eta \leq \xi_n$ for all $n$ by considering $\xi_n - \eta$.

**Corollary 4.1.** If $(\eta_n)$ is a sequence of non-negative random variables, then

$$\sum_{i=1}^{\infty} \mathbb{E}(\eta_i) = \mathbb{E}\left(\sum_{i=1}^{\infty} \eta_i\right).$$

**Corollary 4.2** (Fatou's lemma)**.** Let $\xi_n$ be a sequence of non-negative random variables. Then,

$$\mathbb{E}(\liminf_n \xi_n) \leq \liminf_n \mathbb{E}\xi_n.$$

*Proof.* Apply the monotone convergence theorem to $\lambda_n := \inf_{k>n} \xi_k$. □

Again, the non-negative condition can be replaced by the existence of some random variable $\eta$ such that $\mathbb{E}(\eta) > -\infty$ and $\eta \leq \xi_n$ for all $n$. On the other hand, if $\mathbb{E}(\eta) < \infty$ and $\xi_n \leq \eta$, the theorem holds with limit supremum instead.

We note that in all above theorems, the statement still holds by replacing $\Omega$ with any measurable set by restricting the measure onto that set.

**Theorem 5** (Change of Variables)**.** Given a random variable $\xi$, a measurable function $g : \mathbb{R} \to \mathbb{R}$ and a measurable set $A$, we have

$$\int_A g \mathrm{d}\xi_* \mathbb{P} = \int_{\xi^{-1}(A)} g \circ \xi \mathrm{d}\mathbb{P},$$

where both integrals either exist or not exist simultaneously.

*Proof.* Apply usual method where one first prove the statement for indicator functions. Then, it follows that it holds for simple functions by the linearity of the integral. Finally, for any non-negative measurable function, we take a sequence of monotonically increasing simple functions, and apply the monotone convergence theorem. For arbitrary functions, the result follows by taking $f = f^+ - f^-$. □

**Corollary 5.1** (Law of the Unconscious Statistician)**.** Given a random variable $\xi$ and a measurable function $g : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}(g(\xi)) = \int_{\mathbb{R}} g \mathrm{d}\xi_* \mathbb{P}.$$

**Corollary 5.2.** Let $g : \mathbb{R} \to \mathbb{R}$ be measurable, then, if $\xi$ be a discrete random variable,

$$\mathbb{E}(g(\xi)) = \sum_{x \in A} g(x)p(x).$$

On the other hand, if $\xi$ is absolutely continuous, i.e. there exists some $f$ such that $f\lambda = \xi_* \mathbb{P}$, then

$$\mathbb{E}(g(\xi)) = \int_{\mathbb{R}} g(x)f(x)\lambda(\mathrm{d}x).$$

*Proof.* In the discrete case, we have

$$\mathbb{E}(g(\xi)) = \int g \mathrm{d} \left( \sum_{x \in A} p(x) \delta_x \right) = \sum_{x \in A} p(x) \int g \mathrm{d} \delta_x.$$

By considering $\int g \mathrm{d}\delta_x = \int_{\{x\}} g \mathrm{d}\delta_x + \int_{\mathbb{R} \setminus \{x\}} g \mathrm{d}\delta_x = \delta_x(\{x\}) g(x) + 0 = g(x)$. We have

$$\mathbb{E}(g(\xi)) = \sum_{x \in A} g(x) p(x),$$

as required.

On the other hand, if $f\lambda = \xi_* \mathbb{P}$, we have

$$\mathbb{E}(g(\xi)) = \int_{\mathbb{R}} g \, \mathrm{d}(f\lambda) = \int_{\mathbb{R}} g(x) f(x) \lambda(\mathrm{d}x)$$

as required. □

**Theorem 6** (Fubini's Theorem)**.** Let $(E_1, \Sigma_1, \mu_1), (E_2, \Sigma_2, \mu_2)$ be $\sigma$-finite measure spaces. Then, for any $\Sigma_1 \otimes \Sigma_2$-measurable functions $g : E_1 \times E_2 \to \mathbb{R}$, $g(\cdot, y_0)$ is $\Sigma_1$-measurable for all $y_0 \in E_2$, and $g(x_0, \cdot)$ is $\Sigma_2$-measurable for all $x_0 \in E_1$. Furthermore, $\int_{E_1} g \mathrm{d}\mu_1, \int_{E_2} g \mathrm{d}\mu_2$ are $\Sigma_2$ and $\Sigma_1$-measurable respectively. Finally, if $\int |g| \mathrm{d}\mu_1 \otimes \mu_2 < \infty$, then,

$$\int g \mathrm{d}\mu_1 \otimes \mu_2 = \int \left( \int g(x,y) \mu_2(\mathrm{d}y) \right) \mu_1(\mathrm{d}x) = \int \left( \int g(x,y) \mu_1(\mathrm{d}x) \right) \mu_2(\mathrm{d}y).$$

**Lemma 1.7** (Jensen's Inequality)**.** Let $\xi$ be an integrable random variable and let $g : \mathbb{R} \to \mathbb{R}$ be a measurable, convex function, then,

$$g(\mathbb{E}\xi) \leq \mathbb{E} g(\xi).$$

*Proof.* Recall that the function $g$ is convex if for all $x_0 \in \mathbb{R}$, there exists some $\lambda$ such that $g(x) \geq g(x_0) + (x - x_0)\lambda$ (graphically, $\lambda$ is the slope (more accurately, a subderivative) of $g$ at $x_0$ and so, the inequality is saying that the graph lies above the tangent line).

Setting $x = \xi$ and $x_0 = \mathbb{E}\xi$. Then, the above inequality becomes

$$g(\xi) \geq g(\mathbb{E}\xi) - (\xi - \mathbb{E}\xi)\lambda.$$

Thus, applying the expectation to both sides results in the required inequality by the linearity of the integral. □

**Corollary 6.1** (Lyapunov's Inequality)**.** Let $\xi$ be a random variable and let $0 < s < t$ be real numbers, then

$$\mathbb{E}(|\xi|^s)^{1/s} \leq \mathbb{E}(|\xi|^t)^{1/t}.$$

*Proof.* Use Jensen's inequality with $g(x) = |x|^{t/s}$.

Alternatively, setting $\eta = |\xi|^s$, by Hölder's inequality, we have

$$\|\xi^s\|_1 = \|\eta\|_1 \leq \|\eta\|_{t/s} = \|\xi\|_t^s.$$

Thus, taking both sides to the power of $1/s$, we obtain $\|\xi\|_s = (\|\xi^s\|_1)^{1/s} \leq (\|\xi\|_t^s)^{1/s} = \|\xi\|_t$ as required. □