Measure-Theoretic Probability

A Supplementary Reference for Students

This document covers what measure theory formalizes in probability, what it actively generates beyond formalization, its connection to entropy, the Radon-Nikodym derivative, characteristic functions, and a comparison with maximum likelihood estimation.

Contents

What Measure Theory Formalizes in Probability

Conditional Probability

The elementary definition

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

breaks down when $P(B) = 0$, which occurs constantly with continuous random variables. Measure theory redefines conditional expectation $E[X \mid \mathcal{G}]$ as a **projection** onto a sub-sigma-algebra $\mathcal{G} \subseteq \mathcal{F}$, via the Radon-Nikodym theorem. Conditional probability becomes a special case:

$$P(A \mid \mathcal{G}) = E[\mathbf{1}_A \mid \mathcal{G}]$$

This works even when conditioning on zero-probability events.

Expectation

Elementary expectation is either a sum or a Riemann integral — two separate definitions. Measure theory unifies both as a single **Lebesgue integral**:

$$E[X] = \int_\Omega X \, dP$$

Discrete and continuous cases are the same operation on different measures.

Convergence of Random Variables

Elementary probability has no precise language for limits of random variables. Measure theory gives four distinct notions:

Type	Definition
Almost sure	$P(\omega : X_n(\omega) \to X(\omega)) = 1$
In probability	$P(\|X_n - X\| > \varepsilon) \to 0$
In L^p	$E[\|X_n - X\|^p] \to 0$
In distribution	$F_n(x) \to F(x)$ at continuity points

These are genuinely different and the relationships between them are non-trivial.

Law of Large Numbers

Measure theory states it precisely as **almost sure convergence**:

$$P\left(\omega : \frac{1}{n}\sum_{i=1}^{n} X_i(\omega) \to \mu\right) = 1$$

The "almost sure" qualifier — convergence except on a set of probability zero — requires measure theory to even express.

Central Limit Theorem

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0,1)$$

The arrow $\xrightarrow{d}$ means convergence in distribution — a sequence of measures converging weakly to the normal measure. This is a statement about measures, not numbers.

Stochastic Processes

A stochastic process is an uncountably infinite collection of random variables $\{X_t\}_{t\geq 0}$. The **Kolmogorov extension theorem** guarantees existence — given consistent finite-dimensional distributions, a probability space exists that supports the entire process. Without this, Brownian motion cannot be rigorously constructed.

Zero Probability Events

Measure theory introduces **almost sure** (P-a.s.) reasoning — properties that hold except on sets of measure zero. This is essential because

$$P(X = x) = 0 \quad \text{for all } x \in \mathbb{R}$$

for any continuous random variable, yet X still takes a value.

The pattern. In every case the structure is the same:

$$\text{elementary intuition} \xrightarrow{\text{measure theory}} \text{precise statement that works in all cases}$$

What Measure Theory Generates Beyond Formalization

Conditional Expectation as a Random Variable

Elementary conditioning gives a number. Measure theory reveals that $E[X \mid \mathcal{G}]$ is itself a **random variable** — a function on Ω. This enables:

- Tower property: $E[E[X \mid \mathcal{G}]] = E[X]$
- Martingale theory — built entirely on iterated conditional expectations

Martingale Theory

A martingale is a process where:

$$E[X_{n+1} \mid \mathcal{F}_n] = X_n$$

This requires a **filtration** — a growing sequence of sigma-algebras $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots$ formalizing information accumulating over time. This gives optimal stopping theory, derivative pricing, and the martingale convergence theorem.

Radon-Nikodym as an Active Tool

The Radon-Nikodym derivative $\frac{dQ}{dP}$ is the **likelihood ratio** between two measures. It appears in:

- Hypothesis testing — Neyman-Pearson lemma
- Bayesian inference — posterior as a tilted prior measure
- Importance sampling — changing the sampling measure
- Girsanov theorem — changing drift in stochastic differential equations

Construction of New Probability Spaces

Measure theory provides tools to *build* probability spaces:

- **Product measures** $P_1 \otimes P_2$ — joint distributions of independent processes
- **Kolmogorov extension theorem** — infinite-dimensional spaces enabling stochastic processes
- **Disintegration of measures** — rigorous foundation of Bayesian hierarchical models

Ergodic Theory

A measure-preserving transformation $T : \Omega \to \Omega$ satisfies:

$$P(T^{-1}A) = P(A) \quad \forall A \in \mathcal{F}$$

The **ergodic theorem** states:

$$\frac{1}{n}\sum_{k=0}^{n-1} f(T^k\omega) \to \int f\, dP \quad \text{a.s.}$$

Time averages equal space averages. This connects to MCMC convergence, statistical mechanics, and systems biology.

Weak Convergence and Empirical Processes

Weak convergence of measures:

$$P_n \xrightarrow{w} P \iff \int f\, dP_n \to \int f\, dP$$

for all bounded continuous f. This enables empirical process theory, bootstrap theory, and Donsker's theorem — the empirical CDF converges to a Brownian bridge.

Entropy and Information Theory

The Kullback-Leibler divergence is a Radon-Nikodym derivative:

$$D_{KL}(P\|Q) = \int \log \frac{dP}{dQ}\, dP$$

This unifies information theory, statistical divergences, and variational inference.

The overall picture. Measure theory reveals structure that was invisible before:

sigma-algebras → information → filtrations → martingales → optimal stopping

Radon-Nikodym → change of measure → Girsanov → stochastic calculus

weak convergence → empirical processes → statistical learning theory

Measure Theory and Entropy

The Problem With Discrete Entropy

Shannon entropy

$$H(X) = -\sum_i p_i \log p_i$$

works for discrete distributions but breaks down for continuous ones — replacing the sum with an integral does not yield something consistent.

The Measure-Theoretic Fix

Given two measures P and Q on $(\Omega, \mathcal{F})$ with $P \ll Q$, the KL divergence is:

$$D_{KL}(P\|Q) = \int_\Omega \log \frac{dP}{dQ}\, dP$$

This is the fundamental object. Everything else is a special case.

Entropy as a Special Case

Shannon entropy is KL divergence from the counting measure μ:

$$H(P) = -D_{KL}(P\|\mu) = -\sum_i p_i \log p_i$$

Differential entropy is KL divergence from Lebesgue measure λ:

$$h(P) = -\int f(x) \log f(x)\, dx$$

Both are the **same operation** with different reference measures.

Entropy Is Relative

The measure-theoretic view reveals: **entropy is not absolute — it is always relative to a reference measure.** This is why differential entropy can be negative — it is not a true entropy but a relative entropy against Lebesgue measure.

The Full Chain

$$D_{KL}(P\|Q) = \int \log \frac{dP}{dQ}\, dP = E_P\left[\log \frac{dP}{dQ}\right]$$

Concept	How it arises
Shannon entropy	Q = counting measure
Differential entropy	Q = Lebesgue measure
Cross entropy	$-\int \log q\, dP$
Fisher information	Second-order expansion of D_{KL}
Mutual information	$D_{KL}(P_{XY}\|P_X \otimes P_Y)$

Mutual information measures exactly how far the joint distribution is from independence:

$$I(X;Y) = D_{KL}(P_{XY}\|P_X \otimes P_Y)$$

Connection to Statistical Mechanics

Boltzmann entropy $S = -k_B \sum_i p_i \log p_i$ has the same structure. The **maximum entropy principle** is a variational problem on the space of probability measures:

$$\max_P H(P) \quad \text{subject to} \quad \int f_i \, dP = c_i$$

The solution is always an **exponential family distribution**, connecting thermodynamics, statistical mechanics, and Bayesian statistics.

The Radon-Nikodym Derivative

The Core Idea

Given two measures P and Q on $(\Omega, \mathcal{F})$ with $P \ll Q$, the Radon-Nikodym theorem guarantees a unique function f such that:

$$P(A) = \int_A f \, dQ \quad \forall A \in \mathcal{F}$$

This f is written $\frac{dP}{dQ}$ and answers: *how much more mass does P assign relative to Q, at each point?*

It Behaves Like a Derivative

Chain rule. If $P \ll Q \ll R$:

$$\frac{dP}{dR} = \frac{dP}{dQ} \cdot \frac{dQ}{dR} \quad R\text{-a.s.}$$

Inversion. If $P \ll Q$ and $Q \ll P$:

$$\frac{dP}{dQ} = \left(\frac{dQ}{dP}\right)^{-1}$$

Change of Variables for Measures

When $P_X \ll \lambda$ (Lebesgue), the Radon-Nikodym derivative $\frac{dP_X}{d\lambda} = f_X$ is the PDF:

$$E_P[g(X)] = \int g \, dP_X = \int g(x) f_X(x) \, dx$$

Likelihood Ratios and Hypothesis Testing

The likelihood ratio between hypotheses P_0 and P_1 is:

$$\Lambda = \frac{dP_1}{dP_0}$$

The **Neyman-Pearson lemma** says the most powerful test rejects when $\Lambda > k$. Hypothesis testing theory is built on this object.

Bayesian Inference

Bayes' theorem in measure-theoretic form:

$$\frac{dP(\theta \mid x)}{d\pi(\theta)} = \frac{\mathcal{L}(x \mid \theta)}{\int \mathcal{L}(x \mid \theta) \, d\pi(\theta)}$$

The posterior measure is absolutely continuous with respect to the prior π. Bayesian updating is **reweighting a measure**.

Girsanov Theorem

Given Brownian motion W_t under measure P, define a new measure Q via:

$$\left.\frac{dQ}{dP}\right|_{\mathcal{F}_t} = \exp\left(\int_0^t \theta_s \, dW_s - \frac{1}{2}\int_0^t \theta_s^2 \, ds\right)$$

Under Q, the process $W_t - \int_0^t \theta_s \, ds$ is a Brownian motion. You have **changed the drift** of the process by changing the measure. This is the engine of Black-Scholes pricing and stochastic optimal control.

Importance Sampling

$$E_P[f] = \int f\, dP = \int f \frac{dP}{dQ}\, dQ = E_Q\left[f \cdot \frac{dP}{dQ}\right]$$

The Radon-Nikodym derivative $\frac{dP}{dQ}$ is the **importance weight**. Every importance sampling algorithm is an application of this identity.

Conditional Expectation

$E[X \mid \mathcal{G}]$ is itself defined via Radon-Nikodym. Define $\nu(A) = \int_A X\, dP$ for $A \in \mathcal{G}$. Then $\nu \ll P|_{\mathcal{G}}$ and:

$$E[X \mid \mathcal{G}] = \frac{d\nu}{dP|_{\mathcal{G}}}$$

Conditional expectation *is* a Radon-Nikodym derivative.

Summary of Appearances

Situation	**Radon-Nikodym role**
Defining PDFs	P_X relative to Lebesgue
Hypothesis testing	P_1 relative to P_0
Bayesian updating	Posterior relative to prior
Importance sampling	Target relative to proposal
Girsanov / finance	Risk-neutral relative to real-world
Conditional expectation	Sub-measure relative to restriction
KL divergence	P relative to Q inside a log

Characteristic Functions and Distribution Identification

The Characteristic Function

The **characteristic function** of a random variable X is:

$$\varphi_X(t) = E[e^{itX}] = \int_{-\infty}^{\infty} e^{itx}\, dP_X(x)$$

This is the **Fourier transform of the distribution measure** P_X. It always exists because $|e^{itx}| = 1$, so the integral is always finite.

Uniqueness

$$\varphi_X = \varphi_Y \iff P_X = P_Y$$

The characteristic function **uniquely determines the distribution**. The Fourier inversion theorem gives:

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \varphi_X(t)\, dt$$

whenever P_X has a density f.

Examples

Distribution	$\varphi_X(t)$
$\mathcal{N}(\mu, \sigma^2)$	$\exp\left(i\mu t - \dfrac{\sigma^2 t^2}{2}\right)$
Poisson(λ)	$\exp\big(\lambda(e^{it} - 1)\big)$
Cauchy$(0, 1)$	$\exp(-\lvert t\rvert)$
Uniform(a, b)	$\dfrac{e^{itb} - e^{ita}}{it(b - a)}$
Exponential(λ)	$\dfrac{\lambda}{\lambda - it}$

Moments From Characteristic Functions

If $E[|X|^n] < \infty$:

$$E[X^n] = \frac{1}{i^n} \varphi_X^{(n)}(0)$$

The Taylor expansion encodes all moments:

$$\varphi_X(t) = \sum_{n=0}^{\infty} \frac{(it)^n}{n!} E[X^n]$$

Cumulant Generating Function

$$\log \varphi_X(t) = \sum_{n=1}^{\infty} \frac{(it)^n}{n!} \kappa_n$$

where κ_n are **cumulants**: $\kappa_1 = \mu$, $\kappa_2 = \sigma^2$, $\kappa_3 =$ skewness, $\kappa_4 =$ excess kurtosis. For the normal distribution all cumulants beyond the second are exactly zero.

Sums of Independent Random Variables

If X and Y are independent:

$$\varphi_{X+Y}(t) = \varphi_X(t) \cdot \varphi_Y(t)$$

Convolution of distributions becomes **pointwise multiplication** in Fourier space.

Proof of the Central Limit Theorem

Let $X_1, X_2, \ldots$ be i.i.d. with mean 0, variance σ^2. The characteristic function of $\frac{1}{\sqrt{n}} \sum_{k=1}^{n} X_k$ is:

$$\left[\varphi_X\left(\frac{t}{\sqrt{n}}\right)\right]^n$$

Taylor expanding around 0:

$$\varphi_X\left(\frac{t}{\sqrt{n}}\right) = 1 - \frac{\sigma^2 t^2}{2n} + o\left(\frac{1}{n}\right)$$

Taking the limit:

$$\left[1 - \frac{\sigma^2 t^2}{2n} + o\left(\frac{1}{n}\right)\right]^n \to \exp\left(-\frac{\sigma^2 t^2}{2}\right)$$

This is the characteristic function of $\mathcal{N}(0, \sigma^2)$. The CLT follows from the uniqueness theorem.

Lévy Continuity Theorem

$$P_n \xrightarrow{d} P \iff \varphi_n(t) \to \varphi(t) \quad \forall t$$

Convergence in distribution is equivalent to **pointwise convergence of characteristic functions**.

Empirical Characteristic Function

$$\hat{\varphi}(t) = \frac{1}{n} \sum_{k=1}^{n} e^{itX_k}$$

is a consistent estimator of $\varphi_X(t)$. Parameters are estimated by minimizing:

$$\int |\hat{\varphi}(t) - \varphi(t;\theta)|^2 \, w(t) \, dt$$

over θ, where $w(t)$ is a weight function.

Characteristic Functions vs. Maximum Likelihood

Where MLE Breaks Down

MLE requires the PDF in closed form:

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} \log f(x_i; \theta)$$

Stable distributions are the canonical failure case. Their characteristic function is:

$$\varphi(t) = \exp\Big(i\mu t - |ct|^{\alpha}\Big(1 - i\beta\,\mathrm{sgn}(t)\tan\frac{\pi\alpha}{2}\Big)\Big)$$

The PDF has no closed form except for three special cases:

α	**Distribution**
2	Normal
1	Cauchy
1/2	Lévy

For all other $\alpha \in (0, 2)$, MLE is unavailable. The characteristic function remains explicit.

Other Failure Cases

Situation	**Problem for MLE**	**CF approach**
Stable distributions	No closed form PDF	CF is explicit
Contaminated data	Sensitive to outliers	Downweight via $w(t)$
Mixture models	Multimodal likelihood	CF of mixture is mixture of CFs
Convolutions	Requires convolution integral	CF of sum is a product

Robustness

MLE achieves the Cramér-Rao bound under correct specification but fails under misspecification. CF estimation with $w(t)$ concentrated near $t \approx 0$ is **robust** because $|e^{itx}| = 1$ regardless of how large x is — outlier phase contributions average out.

Honest Comparison

Property	**MLE**	**CF Estimation**
Efficiency (correct model)	Achieves Cramér-Rao bound	Generally less efficient
Robustness	Sensitive to misspecification	Tunable via $w(t)$
Requires closed form PDF	Yes	No
Handles convolutions	Hard	Natural
Software availability	Everywhere	Limited

The Deep Reason

MLE operates in **data space** — it needs $\frac{dP_X}{d\lambda}$, the density with respect to Lebesgue measure. CF estimation operates in **Fourier space** — it needs $\int e^{itx}\, dP_X(x)$, an expectation that always exists.

$$\text{MLE needs: } \frac{dP_X}{d\lambda} \qquad \text{CF needs: } \int e^{itx}\, dP_X(x)$$

The second is a strictly weaker requirement than the first.