Lecture 1

Quantum Mechanics: Motivation and Axioms

Goal: motivate QM through comparison with earlier mathematical models in physics (for those who know them). A crash course on probability theory. QM as a generalization of probability theory to include interference of amplitudes (“wave mechanics”) and noncommuting measurements (“matrix mechanics”).

Mathematical Paradigms in Classical Physics

Classical Mechanics(18th century): describes the motion of rigid bodies

States: position, momentum, angular position/momentum. For N particles phase space is $\mathcal{O}(N)$ dim. E.g. $\mathbb{R}^{\mathcal{O}(N)}, M^{\mathcal{O}(N)}$.

Dynamics: systems of $\mathcal{O}(N)$ coupled nonlinear ordinary differential equations.

Newton's 2nd Law,
Euler, Laplace,
Lagrange,Hamilton

Observables and Measurements: all components of the state are simultaneously measurable and do not update the state.

Classical Field Theory (19th century): describes the motion of fields e.g. fluctuations in a medium, electromagnetism, gravitational fields.

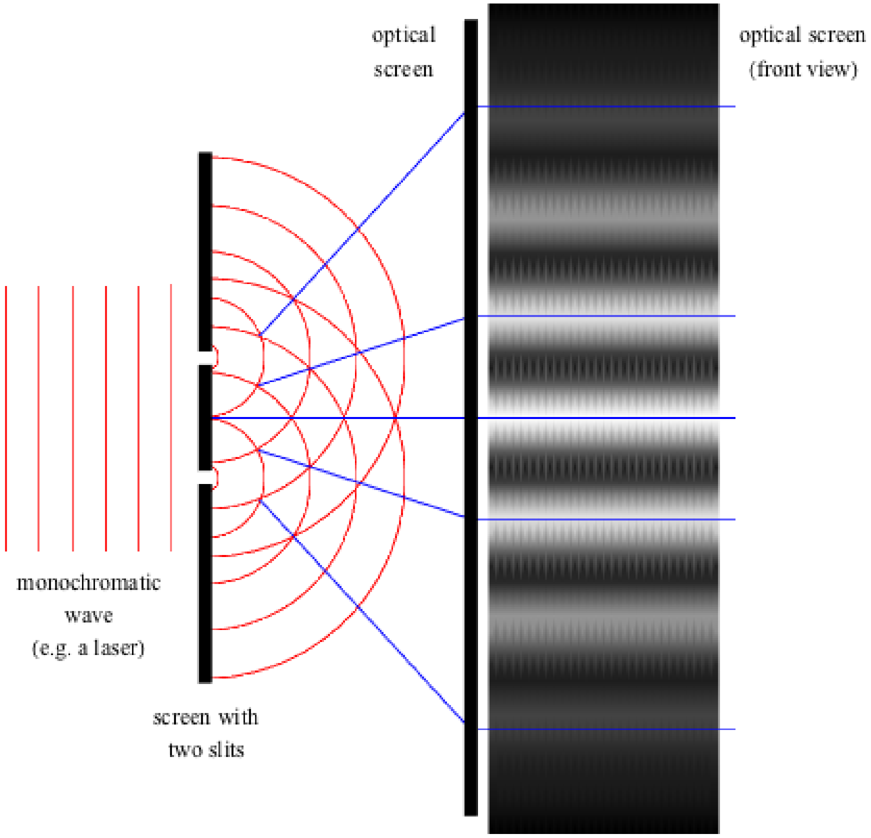
States: scalar/vector/tensor valued functions on manifolds describing uncountably many interacting degrees of freedom

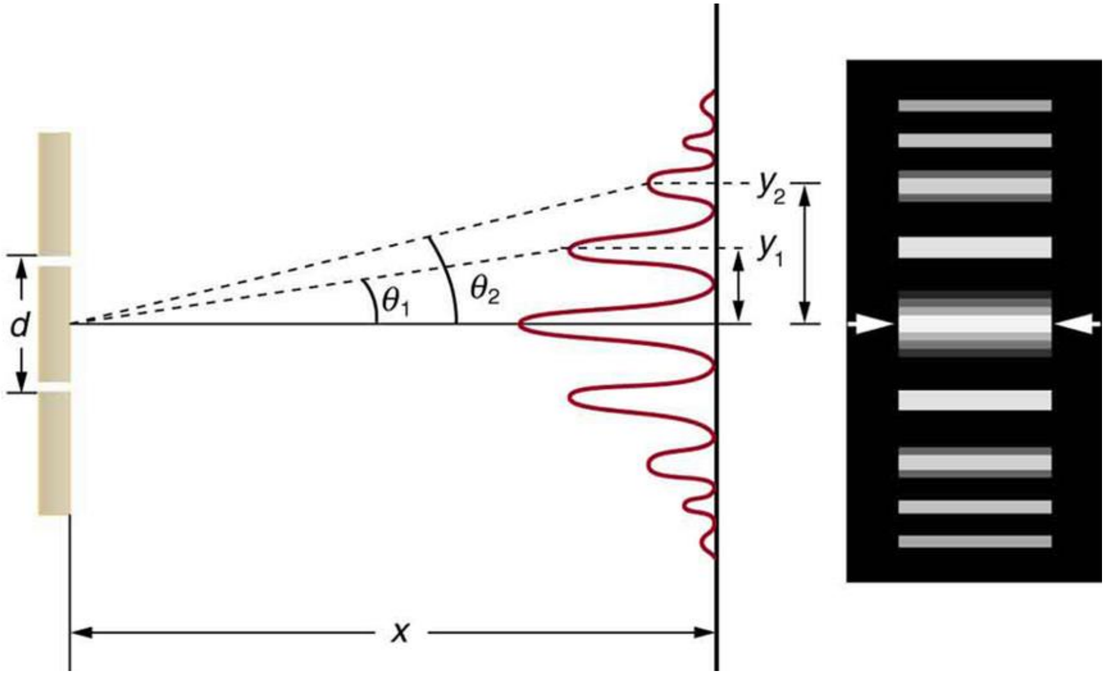
Dynamics: systems of $\mathcal{O}(1)$ coupled nonlinear partial differential equations. Discretization yields $\mathcal{O}(N)$ dim state space, finite element / difference dynamics yield *N* coupled nonlinear ODEs with

$$\boxed{N \to \infty}$$

Observables and Measurements: notion of wave amplitudes interfering to yield visible magnitudes (double slit experiment)

Young’s double slit experiment (light)


optical screen
optical screen (front view)
monochromatic wave (e.g. a laser)
screen with two slits



Mathematical Paradigms in Classical Physics

Statistical Mechanics (19th century): describes behavior of large numbers of microscopic particles, including stochastic effects as simplifying assumption.

(e.g. kinetic theory of gases, Brownian motion, contrast with thermo)

States: probability distributions over phase spaces. For N particles, state is prob distribution over $\mathcal{O}(N)$ dimensional vec space.

Dynamics: main focus on infinite-time limit (equilibrium). Noneq effects described by stochastic ordinary differential equations.

Observables and Measurements: probabilistic theory => observables are expectation values. Measurements can update the state (Bayes)

Summary of Mathematical Physics

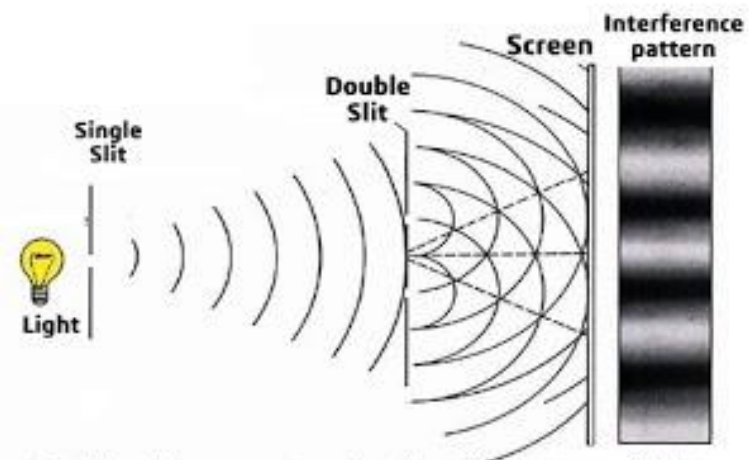
Classical Mechanics: simple correspondence of states and observables, focus on fine-grained behavior of few-body systems.

Classical Field theory: many interacting degrees of freedom described by PDEs / infinite systems of ODEs, amplitudes combine to yield observable magnitudes

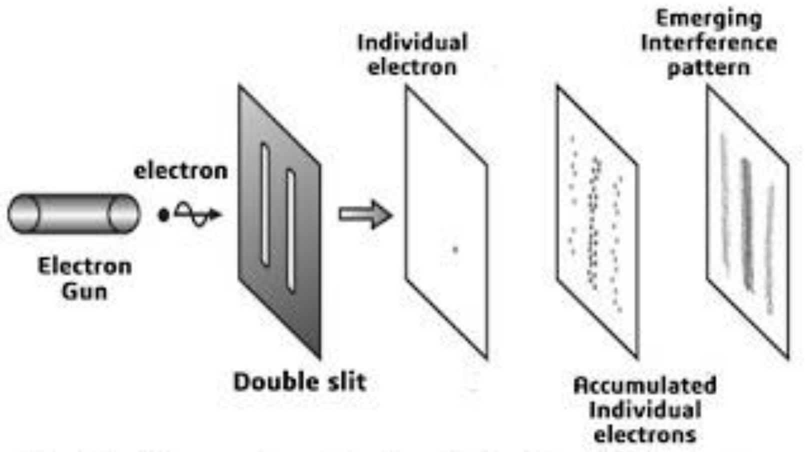
Statistical Mechanics: states are prob distributions over phase space of many interacting particles. Observables are expectations. Stochastic "incompleteness" may be ontological ("of nature") or epistemic ("of knowledge").

Quantum Mechanics: states are amplitude distributions over space of many interacting particles. Observables are expectations. Stochastic behavior cannot be explained as a lack of knowledge about "hidden variables."

Electron double slit experiment



Double-slit apparatus showing the pattern of light


Emerging Interference pattern
Individual electron
electron
Electron Gun
Double slit
Accumulated Individual electrons

Double-slit apparatus showing the pattern of electron

Motivation for Wave Mechanics: interference pattern for electrons passing through the double slit apparatus cannot be explained by mixtures of probability waves, requires amplitude interference.

Stochastic theory of classical statistical mechanics preceded QM in attempting to describe the miscroscopic world, therefore it is mathematical naturally (and somewhat historically accurate) to view QM as a mininimal generalization of stochastic theory to include amplitude interference.

A bit about bits

Physics has traditionally approximated nature as a continuum. This was and is based on inductive reasoning and our limited scale of observation.

A quantitative theory of information that is grounded in the physical world needs to avoid non-discrete systems that store unbounded amounts of information.

Example: if matter were a continuum and you could use the precision of the real numbers, then you could encode an unlimited quantity of data by making a sufficiently precise mark on a rod. The standard model treats space as a continuum and fundamental particles as points.

Dense and uncountable sets are taken as useful approximations in physics, we have no good reason to believe in physical infinity (alternative: unbounded).

Discrete: all elements of the set are separated by finite distance. (natural numbers, integers)

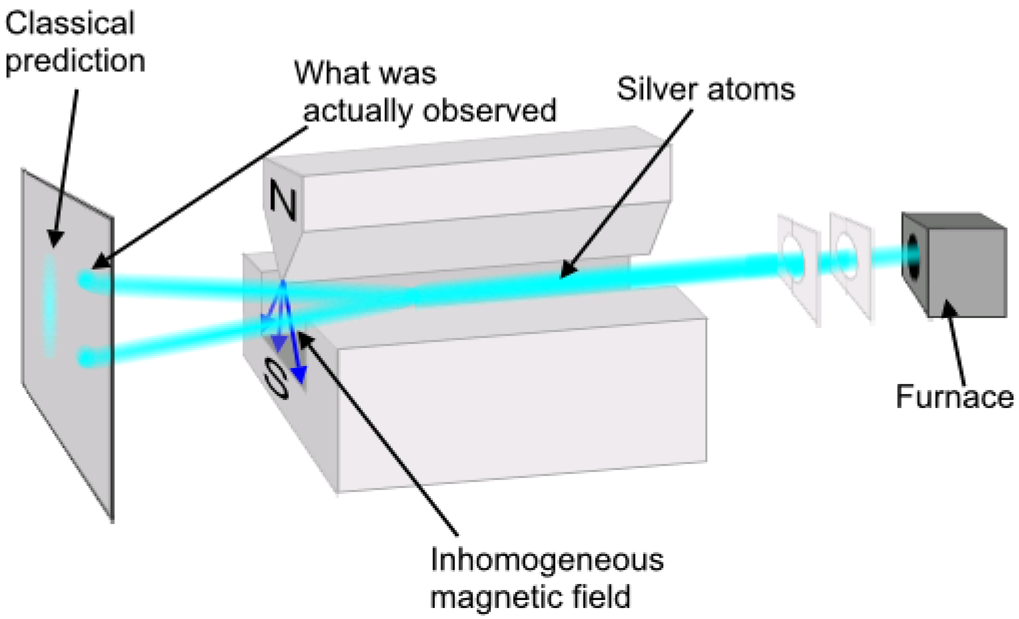
Dense: contains points arbitrarily close to any given point (e.g. the rationals are dense in $\mathbb{R}$)

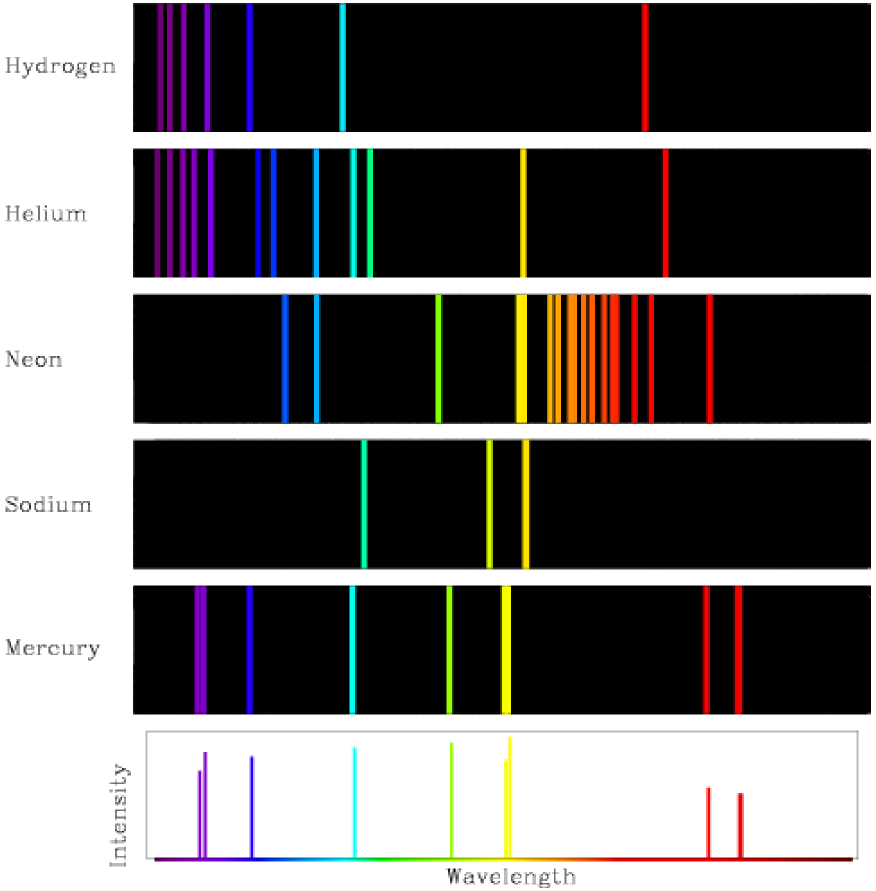
Uncountable: being in correspondence with the real numbers, the cardinality of the continuum

Complete: every sequence for which the distance between consecutive pairs of points goes to zero has a limit in the set (the rationals are not complete. The real numbers are defined as the unique complete totally ordered field)

Fundamental Discreteness: Particle spin, frequencies of atomic transitions

Stern-Gerlach Experiment with spin ½





A bit about bits

Once we have accepted discreteness, we may enumerate all possible states and events with strings of digits.

Bits are the simplest choice, with an alphabet of just two symbols “0” and “1”

The number of digits used to express *N* in base *b* is $\lfloor \log_b(N) \rfloor + 1$, so using any larger constant as the base only compresses the representation by a constant factor

Important notation: $\{0,1\}^n = \{0,1\} \times ... \times \{0,1\}$ is the set of *n*-bit strings.

Work this out if it is unfamiliar, e.g. $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R} = \{(x,y) : x \in \mathbb{R}, y \in \mathbb{R}\}$

A bit about bits

Consider a coin that is heads with probability ½ + ε , and tails with probability ½ - ε.   If ε can be any real number, then is this probabilistic bit (pbit) discrete or not?

The important point is that we cannot learn ε by flipping the coin, even if we are allowed to flip identical coins any finite number of times *N*.

Put another way, if we used *N* pbits to store information, then the most amount of information that could be pulled back out of them by measurement is N bits.

Exactly similar statements hold in quantum mechanics.   Just like ε in the pbit, we sometimes use continuous real parameters to describe states in QM, but this doesn't give them infinite information content because we must look carefully at what we can observe and measure about the states.

A closer look at probability theory

A set of events Ω, and a probability distribution $\mu : \Omega \to [0,1]$

Let $\Omega = \{0,1\}^n$. We can think μ as a normalized vector in 2^n dimensions

$$\boldsymbol{\mu} = (\mu_0, \mu_1, \ldots, \mu_{2^n-1}) \quad , \quad \sum_{i=0}^{2^n-1} \mu_i = 1$$

(Get used to "zero-indexing" and converting freely between bit strings and integers)

Mathematical terminology: the 1-norm of a vector is the sum (integral) of the absolute value of its components. The mathematical spaces ℓ_1, L_1 are the set of real-valued sequences (functions on domain $\mathbb{R}$) with a finite 1-norm.

("$\boldsymbol{\mu}$ is a unit vector in the 1-norm", "$\boldsymbol{\mu}$ is little ell one normalized")

A closer look at probability theory

Probability distributions are vectors, but real-valued linear combinations of probability distributions will not always be probability distributions.

$\frac{1}{2}(\boldsymbol{\mu_1}+\boldsymbol{\mu_2})$ ✓

$\frac{1}{2}(\boldsymbol{\mu_1}-\boldsymbol{\mu_2})$ ✗

Instead use convex combinations: $\alpha_1\mathbf{x}_1 + \ldots + \alpha_k\mathbf{x}_k \ , \ \alpha_i \geq 0 \ , \ \sum_i \alpha_i = 1$

Convex combinations (aka "mixtures") preserve nonnegativity and normalization

The nonnegative orthant of a vector space is called a "convex cone" because it is closed under convex combinations.

A closer look at probability theory

Let $\Omega_{\text{coin}} = \{0, 1\}$ be the state space for a coin (“0” = heads, “1” = tails)

If we have two coins, the set of possible states is $\Omega_{\text{coin}} \times \Omega_{\text{coin}} = \{00, 01, 10, 11\}$

A biased coin could be described by a vector $\boldsymbol{\mu} = (1-\epsilon, \epsilon) \quad , \quad \epsilon \in [0, 1]$

For two **independent** copies of this biased coin, the joint distribution is

00	$(1-\epsilon)^2$
01	$(1-\epsilon)\epsilon$
10	$\epsilon(1-\epsilon)$
11	ϵ^2

$$\boldsymbol{\mu}_{12} = \begin{bmatrix} (1-\epsilon)^2 \\ \epsilon(1-\epsilon) \\ \epsilon(1-\epsilon) \\ \epsilon^2 \end{bmatrix} = \begin{bmatrix} (1-\epsilon) \\ \epsilon \end{bmatrix} \otimes \begin{bmatrix} (1-\epsilon) \\ \epsilon \end{bmatrix} = \boldsymbol{\mu}_1 \otimes \boldsymbol{\mu}_2$$

A closer look at probability theory

Probability theory for composite systems: previous example shows that composite state space is the cartesian product of component state spaces.

If the individual components are statistically independent, then the product rule for combining independent probabilities means joint distribution is a tensor product.

$$\mathbf{a} \otimes \mathbf{b} = \begin{bmatrix} a_1 \\ a_2 \\ \dots \\ a_N \end{bmatrix} \otimes \begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_N \end{bmatrix} = \begin{bmatrix} a_1 b_1 \\ \dots \\ a_1 b_N \\ a_2 b_1 \\ \dots \\ a_2 b_N \\ a_N b_1 \\ \dots \\ a_N b_N \end{bmatrix}$$

Product distributions describe variables that are uncorrelated.

Although the number of possible events scales **exponentially** with the number of subsystems, independence lets us describe them using a number of variables that is **linear** in the number of subsystems.

A closer look at probability theory

In general the component subsystems of a composite system are not independent, but rather they are correlated. For example, weather and clothes:

$\Omega_{\text{weather}} = \{0, 1\}$ ("0" = warm, "1" = cold)

$\Omega_{\text{clothes}} = \{0, 1\}$ ("0" = no jacket, "1" = jacket)

Suppose the weather qualifies as warm ¾ of the time, $\mu_{\text{weather}} = (3/4, 1/4)$

But I always check the forecast and choose my clothes accordingly, then

$$\mu_{\text{weather,clothes}} = \begin{bmatrix} 3/4 \\ 0 \\ 0 \\ 1/4 \end{bmatrix}$$

Due to the correlations, it is impossible to factorize this distribution into a product of independent distributions.

A closer look at probability theory

Although correlated distributions do not factorize into products, we can still ask about the distribution we would see by looking at just one of the subsystems.

If $\boldsymbol{\mu}$ is a joint distribution on $\Omega_1 \times \Omega_2$ then the marginal on the first subsystem is

$$\mu_1(X) = \sum_{Y \in \Omega_2} \mu(X, Y)$$

The marginal represents the reduced state of the system. It may be that $\boldsymbol{\mu}$ records intricate correlations on $\Omega_1 \times \Omega_2$, even while the marginal $\boldsymbol{\mu}_1$ is uniform.

The joint distribution on a composite systems is the tensor product of the marginals on the component subsystems iff they are statistically independent.

A closer look at probability theory

Along with the marginal distribution, another important notion for subsystems is the **conditional distribution**. If $\boldsymbol{\mu}$ is a joint distribution on $\Omega_1 \times \Omega_2$ then conditional probability of $X \in \Omega_1$ given $Y \in \Omega_2$ is

$$\mu(X|Y) = \frac{\mu(X,Y)}{\sum_{X' \in \Omega_1} \mu(X',Y)}$$

The expression in the denominator means that the distribution is “renormalized” to account for $Y \in \Omega_2$ being held fixed.

Bayes’ Theorem: $\mu(X|Y) = \dfrac{\mu(Y|X)\mu(X)}{\mu(Y)}$

A closer look at probability theory

What are the **observables** associated with states described by probability vectors?

One could imagine observing a sequence of events $X_1, X_2, ..., X_m \in \Omega$ drawn according to a fixed distribution $\boldsymbol{\mu}$. These are called **samples** from $\boldsymbol{\mu}$.

A single sample $X \in \Omega$ tells us almost nothing about $\boldsymbol{\mu}$. Strictly speaking, it only tells us $\mu(X) > 0$. (In some contexts observing X makes us believe X is likely).

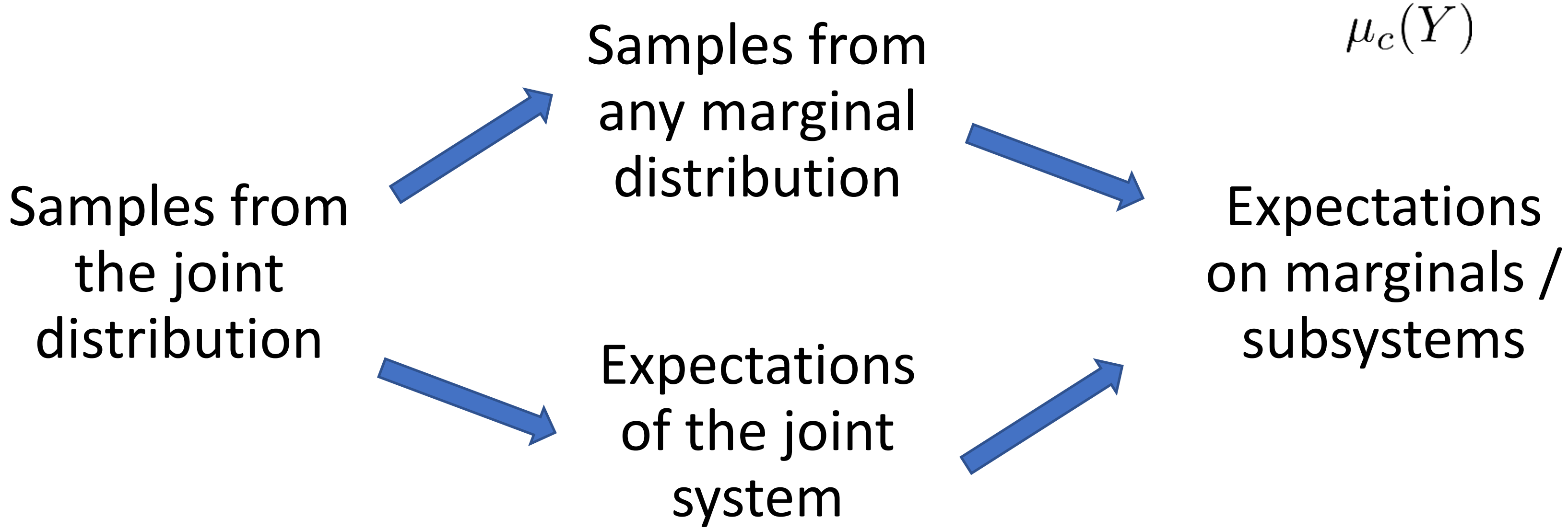
To draw more meaningful conclusions, we need to combine many samples to see the average, or expected behavior of the system.

$$\langle f \rangle_\mu = \sum_{X \in \Omega} f(X)\mu(X)$$

A closer look at probability theory

The formula for expectation values already includes expectations of the marginals as a special case. In the joint system $\Omega_{\text{weather}} \times \Omega_{\text{clothes}}$, we might ask about the expected warmth-level of the clothes by defining $f(X,Y) = Y$, so that

$$\langle f \rangle_\mu = \sum_{(X,Y)\in\Omega_{\mathrm{w}}\times\Omega_{\mathrm{c}}} f(X,Y)\mu(X,Y) = \sum_{Y\in\Omega_c} Y \left(\underbrace{\sum_{X\in\Omega_w} \mu(X,Y)}_{\mu_c(Y)} \right)$$


Samples from the joint distribution

Samples from any marginal distribution

Expectations of the joint system

Expectations on marginals / subsystems

A closer look at probability theory

What does it mean for probability distributions to be close? One natural condition is to demand closeness of all expectations. For this it is necessary and sufficient to be close in **total variation distance**:

$$d_{\mathrm{TVD}}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = \frac{1}{2}\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_1 = \frac{1}{2}\sum_{X\in\Omega} |\mu_1(X) - \mu_2(X)|$$

(“close in the 1-norm”, “close in trace distance”, “additive error”)

Operational meaning is closeness of expectations:

$$d_{\mathrm{TVD}}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = \max_{\text{observables } O\in[-1,1]} |\langle O\rangle_{\mu_1} - \langle O\rangle_{\mu_2}|$$

Approximate sampling: sampling from a distribution that is close in TVD suffices for practical purposes, and can be much easier than exact sampling.

A closer look at probability theory

Definition of TVD means that distributions can be close, even if the probability assigned to some events differs by a large factor. Example:

$$\boldsymbol{\mu}_1 = (2^{-n}, ..., 2^{-n}) \quad , \quad \boldsymbol{\mu}_2 = (0, 2^{-n+1}, 2^{-n}, ..., 2^{-n})$$

$$d_{\mathrm{TVD}}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = 2^{-n} + 2^{-n+1} = \mathcal{O}(2^{-n})$$

The two distributions are exponentially close, even if there is one event (“00…0”) that is infinitely many times more likely to occur in $\boldsymbol{\mu}_2$. Example shows that measuring expectations in physics won’t let us see individual probabilities.

Stronger notion of closeness is a multiplicative bound on the probability of events:

$$(1-\alpha)\mu_1(X) \leq \mu_2(X) \leq (1+\alpha)\mu_1(X)$$

(“multiplicative error”). **Exercise**: relate multiplicative and additive error in terms of the size of the event space $|\Omega|$ (“a dimension factor”).

A closer look at probability theory

We have already seen that verifying closeness of expectations does not verify individual probabilities of events. But how can a finite set of samples ever be used to determine expectation values with high accuracy? Don't we need to flip a coin infinitely many times for it to truly be heads 50% of the time?

Suppose we flip 100 coins and get 55 heads, and 45 tails. This has a finite probability to occur, if the coin is fair this probability is

$$\binom{100}{45} 2^{-100} = 0.04847...$$

Hmm 5%, should we conclude from this experiment that the coin is fair?

To form a more meaningful statement, lets assume $\boldsymbol{\mu}_{\text{coin}} = \left(\frac{1}{2} + \epsilon, \frac{1}{2} - \epsilon\right)$ and say “based on my experiment, I am [some percentage]% confident that the bias on this coin is less than [some value]”

A closer look at probability theory

Many questions of this nature can be answered by **Hoeffding's inequality**.

Let $X_1, X_2, \ldots, X_m \in \Omega$ be independent samples from a distribution $\boldsymbol{\mu}$, and let the observable *f* on Ω take values in [0,1]. Then the estimator defined by

$$\bar{f} = \frac{1}{m}\left(f(X_1) + f(X_2) + \ldots + f(X_m)\right)$$

satisfies $\Pr_{\mu}\left[\left|\bar{f} - \langle f \rangle_{\mu}\right| > t\right] \le e^{-2mt^2}$ for all $t \ge 0$.

Monte Carlo: learn expectations from relatively few independent samples. Named after famous casino, and the idea that one can learn win rates from playing relatively few rounds of a game, instead of counting all possible events.

A closer look at probability theory

Let $X_1, X_2, ..., X_m \in \Omega$ be independent samples from a distribution $\boldsymbol{\mu}$, and let the observable *f* on Ω take values in [0,1]. Then the estimator defined by

$$\bar{f} = \frac{1}{m}\left(f(X_1) + f(X_2) + ... + f(X_m)\right)$$

satisfies $\mathrm{Pr}_\mu\left[\left|\bar{f} - \langle f\rangle_\mu\right| > t\right] \leq e^{-2mt^2}$ for all $t \geq 0$.

Analyze example of 55 heads, and 45 tails. $\bar{f} = 0.55$, $\langle f\rangle_\mu = 0.5 + \epsilon$

$$\mathrm{Pr}_\mu\left[|\epsilon - 0.05| > t\right] \leq e^{-200t^2}$$

Suppose we want 95% confidence, so we choose t to make the RHS less than 0.05,

$$\mathrm{Pr}_\mu\left[|\epsilon - 0.05| > 0.122\right] \leq 0.05$$