Review of Probability

1 Definitions

- **Probability Space** : a triple (Ω, F, P)
- Ω : Sample space : set of all possible outcomes
- F : Event space : $\{E_k \subseteq \Omega | k \in K\}$ each even E_i is a "subset of outcomes". the event space defines the events of interest.
- P : Probability Measure : is a function that assigns each event E_k a value $\in [0, 1]$ called the probability of that event.

  conditions on the probability measure :

  1. $P(\emptyset) = 0$, $P(\omega) = 1$
  2. **subadditivity :** if E_k is a set of **countable, DISJOINT** events, then

$$P(\bigcup_{k=1}^{\infty} E_k) = \sum_{k=1}^{\infty} P(E_k)$$

- side note : an event E_k is a range of possible outcomes (continuous in some notion?).

- This general framework allows you to talk about probability **regardless of the nature of the space**, make **general proofs** hold for all spaces.
- for example the difference between using discrete and continuous random variables is simply 2 differently defined problems in terms of their sample space, and probability measure.
- for the continuous case the probability is defined by the integral

$$P(E) = \int_E p(\omega) dw$$

  - $\omega \in E$
  - capital P $P(E)$ defines a measure, over an event space. How you would calculate probabilities of sets of events, collectively.
  - small p $p(\omega)$ defines the density associated with a particular event.
  - is small p a tool defined by the probability measure big P?

- More generally, **Probability measures, define a way to calculate the "size" (assigning a probability) of a set (set of events)**

- **LAW OF TOTAL PROBABILITY** : if B_i is a countable partition of Ω , then

$$P(A) = \sum_k P(A \cap B_k)$$

  Recall **Marginalization** : you use law of total probability over undersirable variables (B_k) to discard them from known joint distributions and find the marginal distribution / "sub-joint (if more than 1 remain)" of the desired variables $P(A)$.
  Note: In the lecture (oct 1) from 22:48, prof says, only the intersections $A \cap B_k$ are "measurable" (probabilities can be assigned to). B_k are simply mutually exclusive partitions, which need not be measurable ((need not abide by unknown conditions that validate measurability of a partition) (all possible events are not measurable, but are all partitions events?)

- **PushForward Measures** : given a sample space X, and its measure P and another sample space Y, with a function that maps **events** from x to y $f : X \to Y$. if you want to map probabilities defined on X to probabilities on Y you can define a new measure, called a **pushforward measure** that is as follows:

$$(f_* P)(X) = P(f^{-1}(E))$$

- "$\square_*$" is the push-forward operation
  - $P(f^{-1}(E))$: Using known measure P defined on X, you assign probabilities to pre-images of subsets of Y.
  - f should be onto to have an inverse
  - From a generative standpoint : if you have function $f : X \to Y$ that is takes elements ($\in X$) drawn from gaussian and maps into elements on some wacky space (Y). then the events in Y are straightforward to *measure* (assign probability to) by the above method if $A \subseteq X, B \subseteq Y$:

    $$f(A) = B$$

    then probability of event B happening is the same as the probability of event A happening as defined by P_X

    This is interesting because f can be arbitrarily complicated (but should be measurable: maps measurable sets in one sample space to measurable sets in another space)
  - The function f can take collections of outcomes / outcomes separately (if measurable). Generally, f is a map from a sample space X to sample space Y. it can map individual outcomes, or you can look at a group of mappings and call it a mapping between events

- **STUDENT QUESTION : Why is the sample space defined as a separate entity and not as a union of all event spaces** (motivation : in practice the sample space is sometimes intractable/unknowable): **prof says, in such cases different constructions can be made, where you start by "taking collections of random values" and see how they stitch together,to define a consistent object"**
  — " **Kolomogorov Theorems**"
  — arise when studying gaussian processes, ∞ dim random processes, **eg: function estimation (like neural networks??!)**
  — the objective in these problems is to prove that there is this consistent space within which the problem is defined.
  "you don't start with a thing (like the sample space Ω), you have to show that there is a thing that corresponds to the operation that you are trying to do"

- SIDE QUESTION 1

- Such general machinery can help when you think about random processes beyond "simply number or vectors" eg: growing connections over graphs. (building a graph by coin flips)
  - there is a mapping between the result of a coin flip to a random configuration of a graph. You associate a graph configuration with a sequence of coin flips

- Generally if you want to find the probability density over a new variable using a push forward measure from an old variable, you $\sum_{x \inf^{-1}(y)}$ over the probability densities of pre-images

Special case: Change of variables: If V is a random variable on $\Omega \subseteq \mathbb{R}^n$ with density p_V and $f : \Omega \to Y \subseteq \mathbb{R}^m$, then $W \triangleq f(V)$ is a random variable on Y with probability density:

$$p_W(y) = \sum_{x \in f^{-1}(y)} p_V(x) \left| \frac{df}{dx}(x) \right|^{-1}$$

Figure 1: change of variables using pushforward measure

2 Joint and Marginal Distributions

- given a sample space X, and a sample space Y, the joint probability distribution is defined on the product space $X \times Y$ which is also a valid sample space.

- when you define a joint probability distribution more specifically a measure on the product space $X \times Y$, you "automatically get" projection maps π_x,, and π_y, which takes an element of the product space as input, and outputs a corresponding element from a particular sub-space eg : $pi_x : X \times Y \to X$.
- usage : $P_x \triangleq (\pi_x)_* P_{XY}$: which asks you to sum over the densities of pre-images of some x in $X \times Y$ defined by π_x (**marginalizes** out all variables other than X).

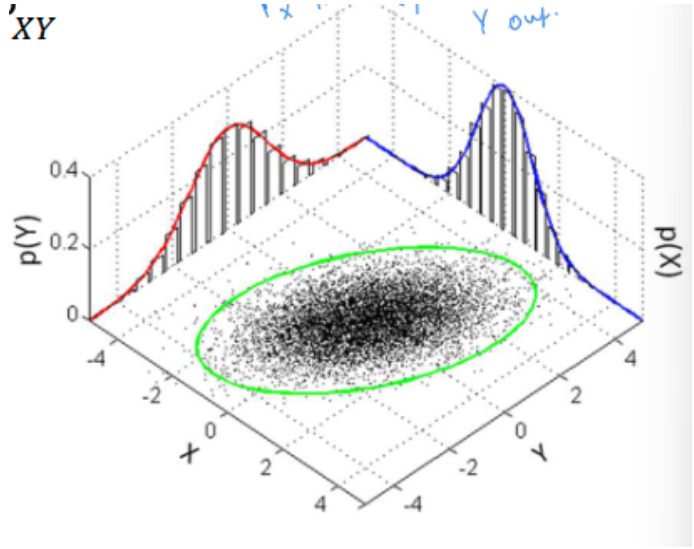

Figure 2: Marginalization via pushing forward projection maps

Marginalization : if $X \in R^m$, and $Y \in R^n$ are both sample spaces, and you define the product space $X \times Y$

$$P_x(E) = (\pi_x)_* P_{XY}(E)$$

$$\Rightarrow p_x(x) = \int_{R^n} p_{x,y}(x,y)dy$$

3 Conditional probability and independence

- $P(A|B) = \frac{P(A \cap B)}{P(B)}$, $p_{x|y}(x|Y=y) = \frac{p_{XY}(x,y)}{p_y(y)}$
- Know that an even B has occurred, what is the probability that A has occurred.
- How the distribution of A is affected given that B has occurred.
- If the distribution of A hasn't changed at all when and after B occurred, that means A is independent of B. $P(A|B) = P(A)$. It follows that if A and B are independent

$$P(A \cap B) = P(A)P(B)$$

- punchline: if A, and B are independent, then knowing that B has happened, doesn't impact our estimate of A's distribution in any way. (doesn't inform anything about A)
- What would independence look like : : Independence is not a property of the sets, by a property of the measures, making it different to visualize. If you had to , you could plot the probability densities over sets, above (coming out of the screen) a plot like Figure 3.
  if you look these from a view angle parallel to the screen and along the axes of a variable, seeing only the density, the distribution over all of A, would be **Homogenous, even when regions corresponding to other variables intersect** these variables, that don't cause any changes in the probability density of A are independent to A.

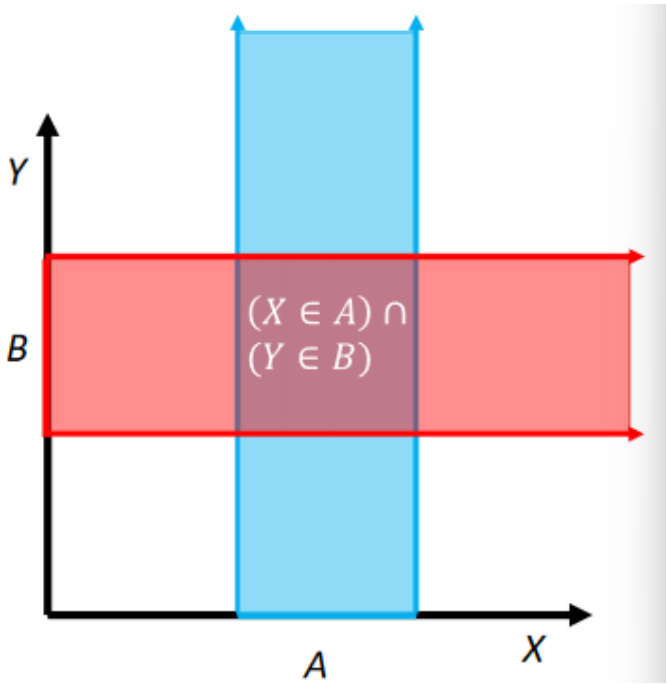

Figure 3: Intersection of sets A, B

4 Bayes Rule

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

Scenarios

1. **Disease test**

   - X = {sick, not sick}, Y = {positive, negative} (test result for disease present?)
   - $P(Y|X)$ **: Forward model** : probability that test says sick (Y = positive), and person is sick (X = sick). Which is a feature of the test by design. If you are using a test, you should know how well it works. you probably know if a person is sick , how probably is it for the test to say the person is sick as well.
   - $P(X)$ **: prior** : information that y ou have already that led you to an estimate about the probability that the person is sick.
   - $P(Y)$ **: "normalization constant"** : some characteristic of the test, here "how often the test says a person is sick"
   - $P(X|Y)$ **: Posterior** : if a test says a person is sick, and you already have an estimate of how probable it is for the person to be sick, How does the result of the test, your estimate?

2. State estimation using sensor data

   - X = differnet possible states, Y = different possible measurements
   - $P(Y|X)$ **: Sensor model : Forward process / likelihood**: what is the probability that you sensor makes some measurement y, when observing some state x. this is direct consequence of the behaviour of the sensor which can be estimated by subjecting the sensor to different states in a lab. **completely known before hand**
   - **P(X) : prior** : the estimate you have of how you likely it is for you to be in some state x, based on whatever previous information. This is a running estimate that you usually update each iteration. what you compute for the current time step will be called the belief, and the current believe, will become the prior in the next iteration.

- **P(Y) : "Normalization constant / evidence"** : In practice it is something you use just to make sure the numerator integrates to 1. It can be interpreted as the probability that your sensor might output this particular measurement, as opposed to any other measurement regardless of what state is observed.
- $P(X|Y)$ **: posterior :** What your updated belief will be given that your sensor made some measurement y, and you had some prior information that led you previously believe that you were in state x with probability p(x+)

3. **PUNCHLINE :** Bayes rule tells you how to **update your prior** ($P(X)$)/existing beliefs in the face of **new information** ($P(Y|X)$) (eg: sensor measurements/test results) to make a **new belief** $P(X|Y)$(**posterior**)

5 mulivariate gaussians

- gaussian are the maximum entropy distributions that has their's moments
- they are solid general models because **they are closed under marginalization, conditioning, affine transformations**.

uncertainity on lie groups can be modeled by using the exponential map as a pushforward measure on a distribution defined on the tangent space of the identity.

6 SIDE QUESTIONS

1. Is a **Random variable** completely defined by a triple (Ω, F, P)
2. When would you want to push forward an uncertainty distribution from the group to the tangent space at the identity using a log map.