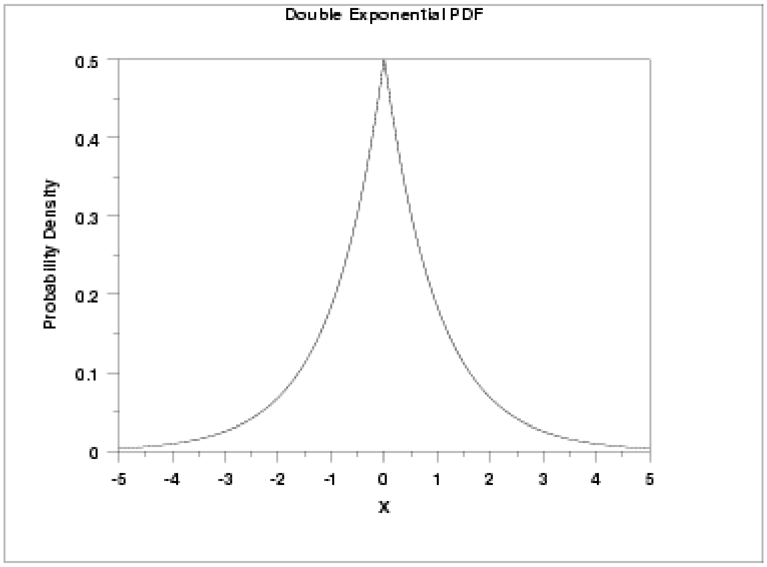
probability and information theory

- probability is a way to formally deal with uncertainty.
- information theory enables us to quantify the amount of uncertainty in a probability distribution.
- stochastic means non-deterministic, with some amount of uncertainty.
- three possible sources of uncertainty in ML:
  - inherent stochasticity in system being modeled.
  - incomplete observability of some parts of the system
  - incomplete modeling → when we use a model that discards some of the information we have observed.
- in many cases, it is more practical and effective to use a simple but uncertain rule rather than a complex but certain one.
- In the **frequentist** approach, it is asserted that the only sense in which probabilities have meaning is as the limiting value of the number of successes in a sequence of trials, i.e. as

  $p = lim_{n->\infty}\frac{k}{n}$

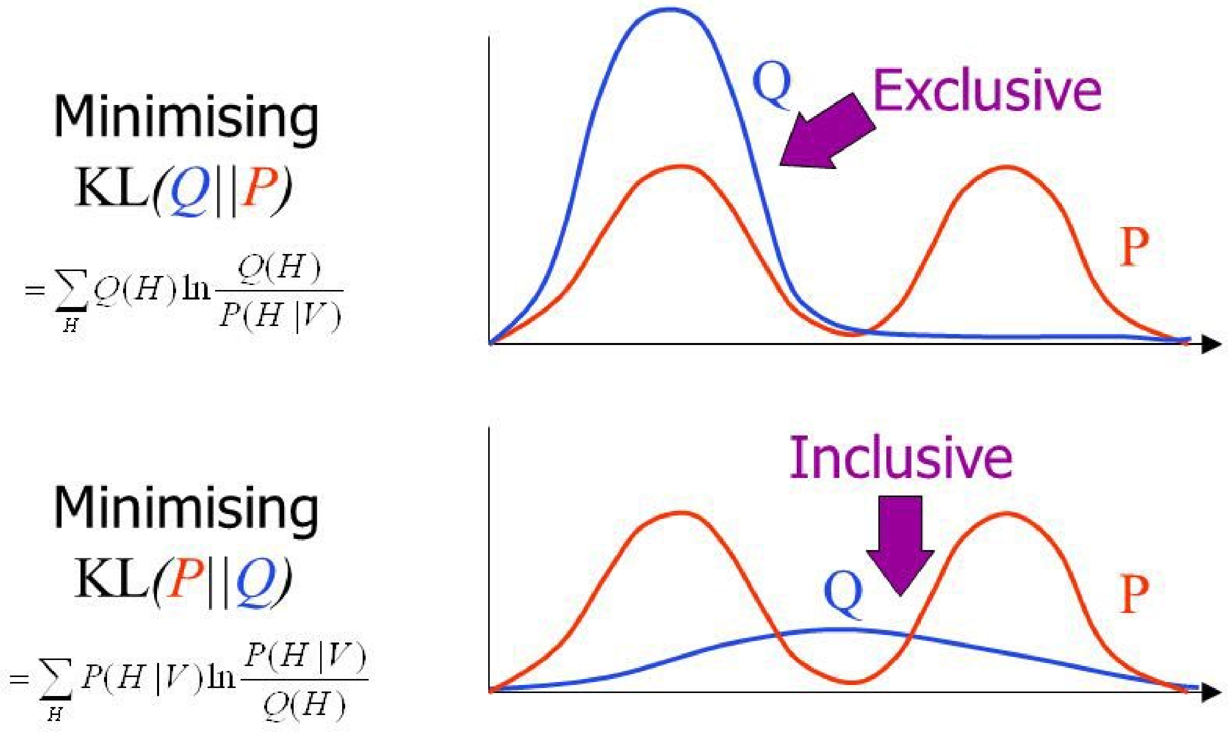
  where k is the number of successes and nn is the number of trials. In particular, it doesn't make any sense to associate a probability distribution with a *parameter*.
- In the **Bayesian** approach, we interpret probability distributions as quantifying our uncertainty about the world. In particular, this means that we can now meaningfully talk about probability distributions of parameters, since even though the parameter is fixed, our knowledge of its true value may be limited.
- we can think of a random variable as a set of states with a likelihood of being in each state. or it can be thought of, as we discussed in our notes on the probability class, as a function from sample space to real numbers.
- a probability distribution over many variables is called a joint probability distribution.
- when we have $P(x, y)$ as a grid, we can get $P(x)$ and $P(y)$ by summing the rows/columns and writing the answer in the margins. Hence these are called marginal probabilities.
- conditional probability $P(Y = y|X = x) = \frac{P(Y=y, X=x)}{P(X=x)}$.
- random variables are independent if their joint distribution can be expressed as a product of individual factors.
- conditional independence implies they are independent when conditional probability is considered.
- Expectations are linear: $E[af(x) + bg(x)] = aE[f(x)] + bE[g(x)]$
- variance: $Var(x) = E[(f(x) - E[f(x)])^2]$. The square root of this is standard deviation.
- $Cov(f(x), g(x)) = E[(f(x) - E[f(x)])(g(x) - E[g(x)])]$.
- Correlation normalizes covariance in order to measure only how much the variables are related, rather than also being affected by the scale of the separate variables.
- independence implies 0 covariance. Reverse is not true.
- normal distribution is a good default distribution in many cases because:
  - central limit theorem shows that the sum of many independent random variables is approximately normally distributed.
  - it can be proved that, out of all possible probability distributions over the real number line with the same variance, the normal distribution encodes the maximum amount of uncertainty → least amount of prior knowledge. proof not given in this chapter.

- If we have a vector of random variables, $Cov(\mathbf{x})_{i,j} = Cov(x_i, x_j)$. If a multivariate normal has covariance of diagonal matrix, that means that the elements of the vector of random variables are independent.
- Laplace distribution: $P(x; \mu, \gamma) = \frac{1}{2\gamma} e^{-\frac{|x-\mu|}{\gamma}}$



- Dirac delta is such that it is zero valued everywhere except at 0 but integrates to 1. it is defined in terms of its result when integrated, this is not a normal function where you can find the output for each point in the domain.
- A mixture distribution is made up of several component distributions. On each trial, we choose which component to sample from and then sample from it. if we don't know which component it comes from, we model it with a latent variable, and we can estimate it with MLE, for example.
- the word "prior" indicates that it expresses the model's belief about something it is about to estimate. Posterior is the belief after observations have been accounted for.
- softplus function is $f(x) = log(1 + e^x)$.
- $\sigma(x)$ is the sigmoid function. σ^{-1} is called logit in statistics.
- bayes' rule $P(x|y) = \frac{P(y|x)P(x)}{P(y)}$
- suppose we have two random variables two random variables x and y such that $y = g(x)$. When we are expressing probabilities interachangably with x and y, we need to account for the distortion of space introduced by g. What needs to stay true is $|p_y(g(x))dy| = |p_x(x)dx|$
- information theory:
  - originally invented to study how to design codes to send messages with as little expected message length as possible
  - we can use it to quantify similarity between probability distributions.
  - basic intuition is that learning that an unlikely event has occurred is more informative than learning that a likely one has occurred.
  - likely events should have low information content so as to use as little bits (message length) as possible.
  - We define self-information of an event $X = x$ to be $I(x) = -log\ P(x)$
  - One nat is the amount of information gained by observing an event of probability $1/e$.

    - Think of it like this: if an event is more unlikely it takes more information to represent it. the expected self information is called the entropy of the distribution (also called Shannon entropy in discrete case or differential entropy in continuous case): $H(x) = -\int p(x) log\, p(x) dx$
    - KL divergence between two probability distributions: $D_{KL}[P||Q] = E_{x from P} log\frac{P(x)}{Q(x)}$. We can show that this is always non-negative. it is kind of a measure of distance between two probability distributions.
    - Observe that the above is not symmetric.
    - In a learning model, note how which divergence is minimized changes the result totally.



- Cross entropy $H(P, Q) = -E_{x from P} log\, Q(x)$
- Structured probabilistic models or probabilistic graphical models can be either directed or undirected depending on how we want to represent independence relationships. we will visit this in the notes on Structured Probabilistic Models for DL