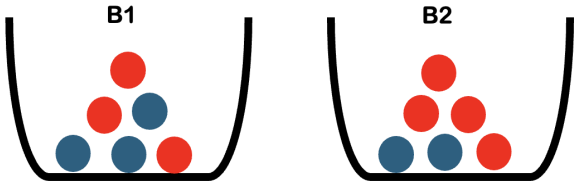
Probability Cheat Sheet

🟠 Bayes Rule

$p(A|B) = \frac{p(A,B)}{p(B)} = \frac{p(B|A)p(A)}{p(B)}$

$p(\ Red\_ball \mid BinB1\ ) =\ ?$

$p(\ BinB1 \mid Red\_Ball\ ) = ?$



Which one above is likelihood and which one is posterior?

Bayes rule can be written as: $Posterior\ = \frac{Likelihood \times Prior}{Evidence}$

🟠 Marginalization

$p(X) = \sum_y p(X,Y)$ ... where $p(X,Y)$ is the joint probability of both X and Y happening.

- Complete the following expression in plain English:

  $p(Messi\ scoring\ a\ goal) =$
  $p(Messi\ scoring\ a\ goal\quad and\quad Messi's\ team\ wins) +$ (what else?)

🟠 Law of Total Probability (LoTP)

- $p(A,B,C) = p(A|B,C)p(B|C)p(C)$
- Can we write the RHS also as: $p(C|B,A)p(A|B)p(C)$?

- Can you apply Marginalization first ... and then LoTP ... to the denominator of Bayes Rule:

  $p(A|B) = \frac{p(B|A)p(A)}{p(B)} = \frac{p(B|A)p(A)}{?}$

🟠 Expectation (a.k.a Averaging)

- Expectation of a random variable:

  $E[X] = \int xf(x)dx = \sum_i x_i p(x_i)$

- When we want to compute expectation of a function of a random variable, it's super easy:

  $E[g(X)] = \int g(x)f(x)dx = \sum_i g(x_i)p(x_i)$

- The above can be written as:

$$E_{x_i \sim p(x)} g(x_i)$$

which means that if someone does not give you the $p(x)$ distribution and only gives you samples from $p(x)$, you can still compute the expectation by averaging over those samples.

🟠 Importance Sampling

$$E[g(X)] = \int g(x)f(x)dx = \int g(x)f(x)\frac{\omega(x)}{\omega(x)}dx = \int \left[\frac{g(x)f(x)}{\omega(x)}\right]\omega(x)dx$$

- Now suppose you still want to calculate the expectation of $g(X)$ but I can no longer give you samples from $f(x)$. Can I give you any other samples that can

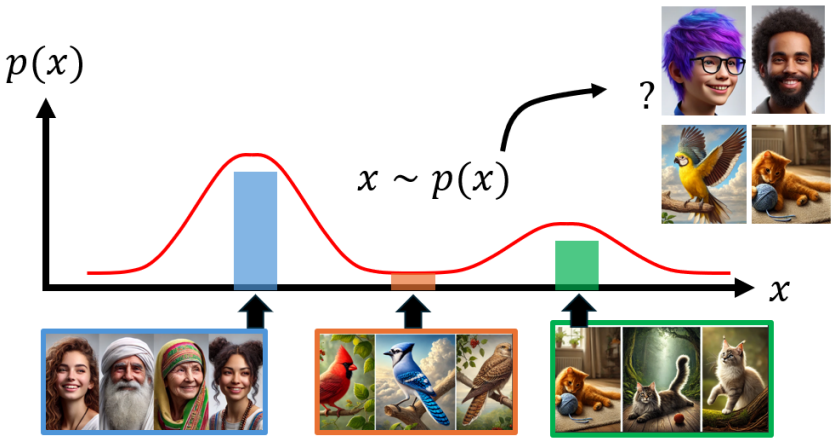
help you calculate $E[g(X)]$?

In other words, can you fill in the blanks for the following equation?

$$E_{\_\_\_\_\_}\left[\frac{g(x)f(x)}{\omega(x)}\right]$$

🟠 Sample From a Distribution

- Imagine the space of all images (i.e., each point x in this space is some image).
  Imagine a distribution
  $p(x)$ over this space, meaning that the volume under this distribution should $= 1$.
  When is it possible to sample from this distribution?
  Can we sample when
  $p(x)$ is Gaussian?
  Can we sample when
  $p(x)$ is a mixture of many Gaussians (like the example shown below)?
  Can we sample if
  $p(x)$ takes any arbitrary shape as long as the volume under the curve is still $= 1$?



🟠 Independence

Check the right column:

	Dependent	Independent
Raining in Shanghai and raining in New York		
A person's age and that person's vocabulary		
I wear a red shirt and my wife wears a red dress		
Butterfly flaps its wings in Brazil and it rains in Miami		

- X and Y are independent if $p(X,Y) = p(X)p(Y) \;\; \forall x, y$

- This also means that ... if I tell you X happened, it won't help you to guess Y any better.

  Mathematically this can be written as: $p(Y|X) = p(Y)$ ... when X and Y are independent.

- For example, say Adam drinking water and Eve drinking water are independent events.

  Now, if I ask you this: Eve is drinking water now, what is the probability that Adam is drinking water?

  Would your guess change in any way if I did not tell you Eve was drinking water now?

Conditional Independence

- X and Y are conditional independent on A if: $p(X, Y|A) = p(X|A)p(Y|A)$

  This means that X and Y are dependent in general ... but if I give you the information A, then they become independent.

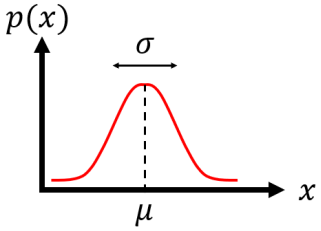
- For example, $p(X = \text{Height of a person})$ and $p(Y = \text{Vocabulary of a person})$ are dependent because knowing one gives you information about the other.

  But,
  $p(X, Y|A = \text{all persons in tenth grade})$ are conditionally independent. Why? Because once you only take students from tenth grade, among them height and vocabulary may have no dependence. A short tenth grader may or may not have a smaller vocabulary than a tall tenth grader.

- Can you think of another example where X and Y are dependent in general, but conditoinally independent on A?

Gaussian and Multivariate Gaussian

$$f(x) := \mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



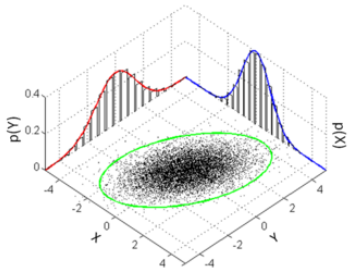
Sum of independent Gaussian is a Gaussian

$\sum_k \mathcal{N}(\mu_k, \sigma_k) = \mathcal{N}(\bar{\mu}, \sum_k \sigma_k)$ where $\bar{\mu}$ is the average mean.

Multivariate Gaussian is K-dimensional Gaussian

$$f(\mathbf{x}) := \mathcal{N}(\mu, \Sigma)$$

where Σ is the covariance matrix $\in R^{K \times K}$
When
$\Sigma = \text{identity matrix } I,$ it's called "Isotropic Gaussian"



Can you imagine what an isotropic matrix looks like when $K = 3$.

🟠 Maximum Likelihood Estimation (MLE)

- Suppose you know that salaries of computer engineers is a Gaussian distribution.
  You got a database of salaries of all computer engineers. Can you calculate θ, the average salary?

- Do you think the answer is: "sum up all the salaries and divide by the number of salary values?"
  If so, why is this the answer?
  If not, why not? Pause here and think before you read on.

- MLE says here is what you should do:
  1. Pretend you know θ and calculate how likely are all the salary values. This means you start with a Gaussian centered at some arbitrary value θ_0, and for that Gaussian, calculate the probability of each salary value and sum up those probabilities. Let's call this sum $\mathcal{L}(\theta)$. Note it's a function of θ because if you change θ, the summed up number will change.

2. Keep moving the θ_0 in some direction that will increase $\mathcal{L}(\theta)$. In other words, do gradient ascent on θ. You will end up at some optimal θ^* and you cannot increase $\mathcal{L}(\theta)$ any more. That θ^* is called your maximum likelihood estimate.

- Formally, we write this as:
  Given parameter
  θ, and data x, define $\mathcal{L}(\theta) = \ p_\theta(x) \ = \ p(x|\theta)$.
  This is called the likelihood function.
  For our Gaussian example above,
  $p(x|\theta)$ is a Gaussian distribution $\mathcal{N}(\theta, \sigma^2)$.
- The MLE solution says:

$$\theta^* = \arg\max_\theta \ \ p(X|\theta)$$

- Now, can you think of modeling the following problem using MLE:
  An opaque bag has many colored balls and you want to know the fraction of red balls inside that bag.
  You draw balls
  5 times from the bag, putting the ball back after every draw, and you observe the colors as: $\{R, B, B, B, R\}$, where R is red and b is blue.
  - What is x in this case?
  - What is θ in this case?
  - What is the probability distribution $p(x|\theta)$? Is it Gaussian? Is it Uniform? Is it exponential?
  - Can you model this problem in the same format as the $\arg\max$ equation above?

🟠 Markov Chain

$$p(X_t|X_0, X_1 \cdots X_{t-1}) = p(X_t|X_{t-1})$$

- The probability of the state only depends on the previous state (one time step ago).
  - E.g., If salary of a new job only depends on the previous job's salary, then it satisfies Markov.
- Example: Which of the following do you think satisfies the Markov property?
  - Say each day's weather is a new random variable. Do you need yesterday's and day before yesterday's weather to predict today's weather?
  - You are predicting how much a student will get in a practice exam. Do you need his score from the previous practice exam and the one before that? Or is only the previous day's practice exam score enough?
  - You are predicting the current location of a taxi in Manhattan. Do you need the last two passenger drop-off locations or is the last passenger drop-off location enough to predict the taxi's current location?