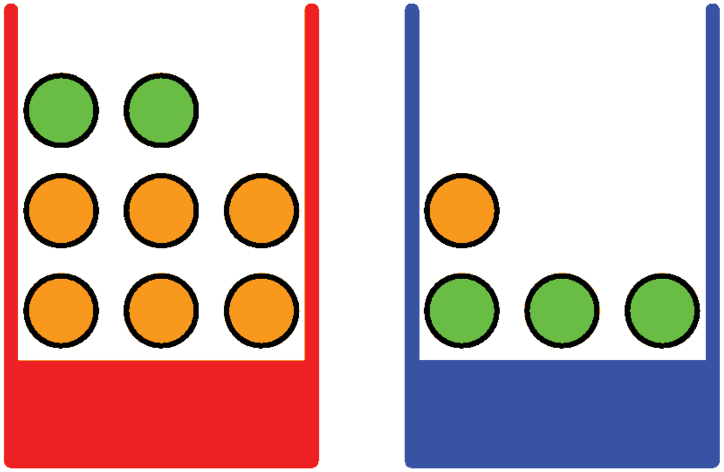
Background: Probabilities, Probability Densities, and Gaussian Distributions

CSE 4309 – Machine Learning

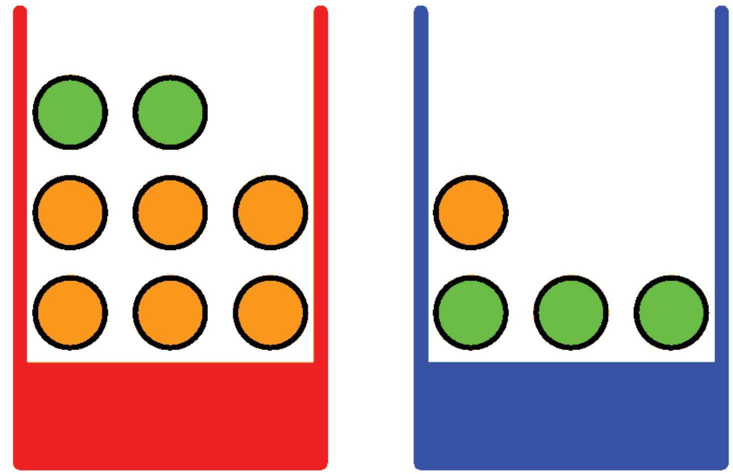
Vassilis Athitsos
Computer Science and Engineering Department
University of Texas at Arlington

A Probability Example



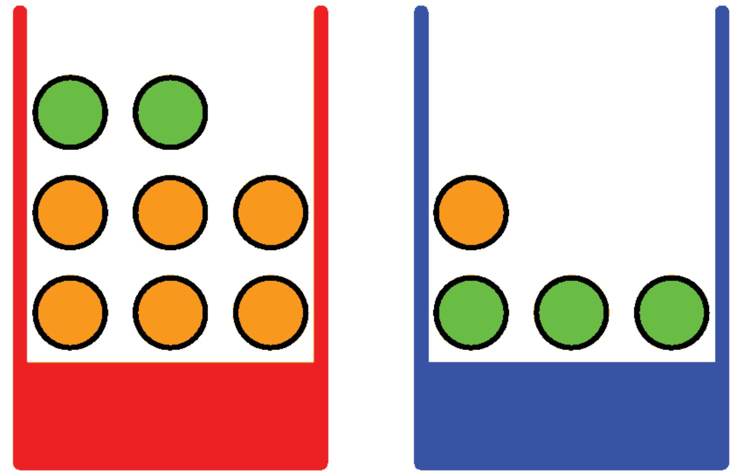
- We have two boxes.
  - A red box, that contains two apples and six oranges.
  - A blue box, that contains three apples and one orange.
- An experiment is conducted as follows:
  - We randomly pick a box.
    - We pick the red box 40% of the time.
    - We pick the blue box 60% of the time.
  - We randomly pick up one item from the box.
    - All items are equally likely to be picked.
  - We put the item back in the box.

Random Variables



- A random variable is a variable whose possible values depend on random events.
- The process we just described generates two random variables:
  - B: The identity of the box that we picked.
    - Possible values: r for the red box, b for the blue box.
  - F: The type of the fruit that we picked.
    - Possible values: a for apple, o for orange.

Probabilities



- We pick the red box 40% of the time, and the blue box 60% of the time. We write those probabilities as:
  - $p(B = r) = 0.4$.
  - $p(B = b) = 0.6$.

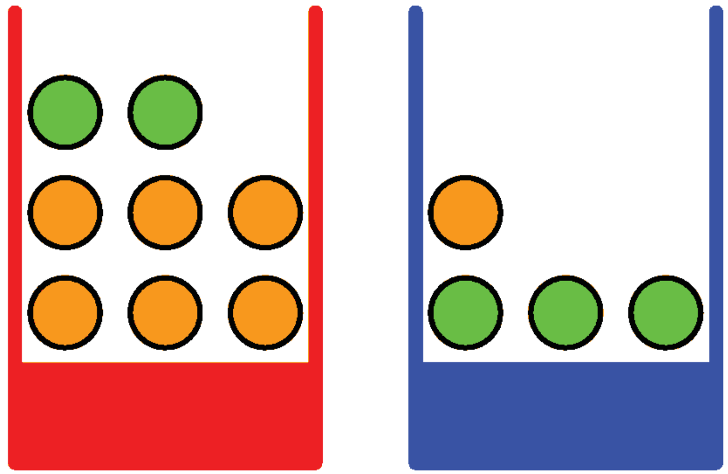
Probabilities

- Let p be some function, taking input values from some space X, and producing real numbers as output.
- Suppose that the space X is a set of **atomic events**, that cannot happen together at the same time.
- Function p is called a **probability function** if and only if it satisfies all the following properties:
- $\forall x \in X, p(x) \in [0, 1]$
  - The probability of any event cannot be less than 0 or greater than 1.
- $\sum_{x \in X} p(x) = 1.$
  - The sum of probabilities of all possible atomic events is 1.

Atomic Events

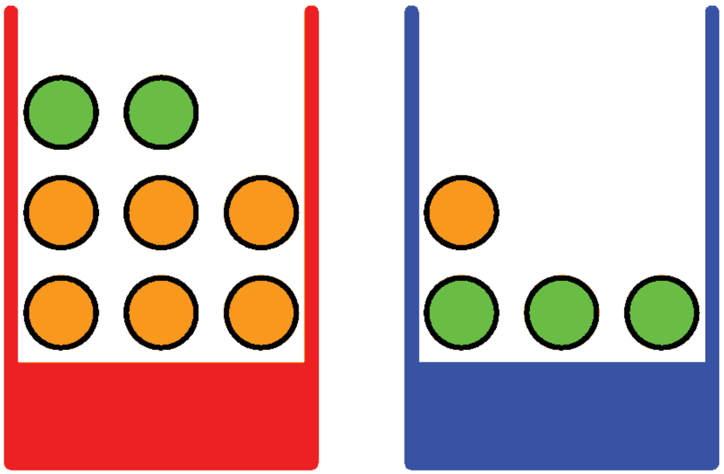
- Consider rolling a regular 6-faced die.
- There are six atomic events: we may roll a 1, a 2, a 3, a 4, a 5, or a 6.
- The sum of probabilities of the atomic events has to be 1.
- The event "the roll is an even number" is not an atomic event. Why?
  - It can happen together with atomic even "the roll is a 2".
- In the boxes and fruits examples:
  - One set of atomic events is the set of box values: {r, b}.
  - Another set of atomic events is the set of fruit types: {a, o}.

Conditional Probabilities



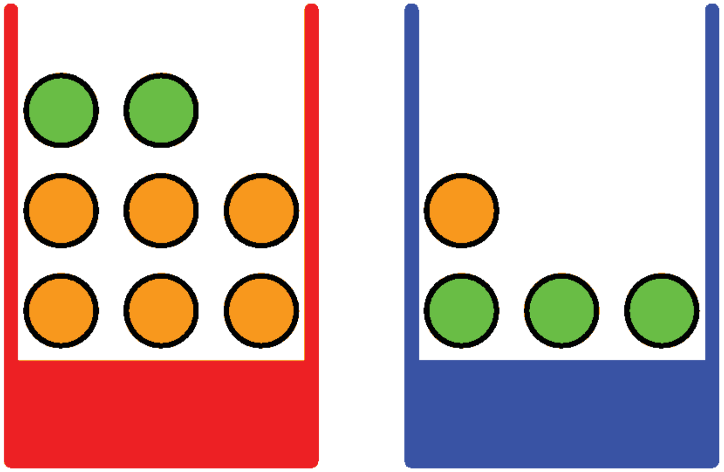
- If we pick an item from the red box, that item will be:
  - an apple two out of eight times.
  - an orange six out of eight times.
- If we pick an item from the blue box, that item will be:
  - an apple three out of four times.
  - an orange one out of four times.
- We write those probabilities as:
  - p(F = a | B = r) = 2/8.
  - p(F = o | B = r) = 6/8.
  - p(F = a | B = b) = 3/4.
  - p(F = o | B = b) = 1/4.
- These are called **conditional probabilities**.

Joint Probabilities



- Consider the probability that we pick the blue box and an apple.
- We write this as p(B = b, F = a).
- This is called a **joint probability**, since it is the probability of two random variables **jointly** taking some specific values.
- How do we compute p(B = b, F = a)?

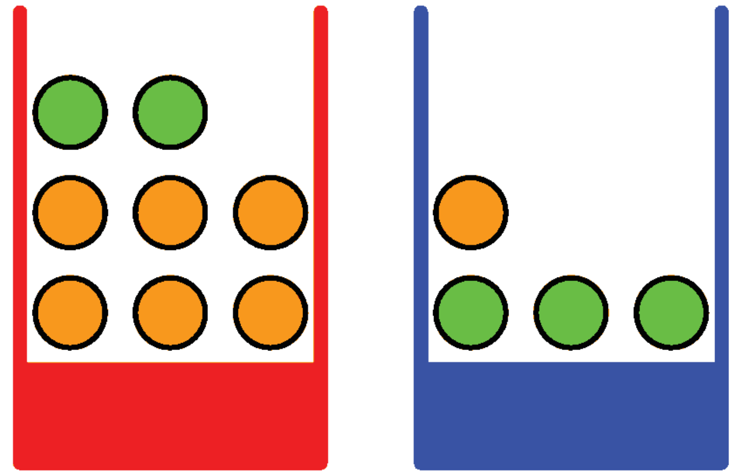
Joint Probabilities



- Consider the probability that we pick the blue box and an apple.
- We write this as p(B = b, F = a).
- This is called a **joint probability**, since it is the probability of two random variables **jointly** taking some specific values.
- How do we compute p(B = b, F = a)?
- p(B = b, F = a) = p(B = b) * p(F = a | B = b)

  = 0.6 * 0.75 = 0.45.

Conditional and Joint Probabilities



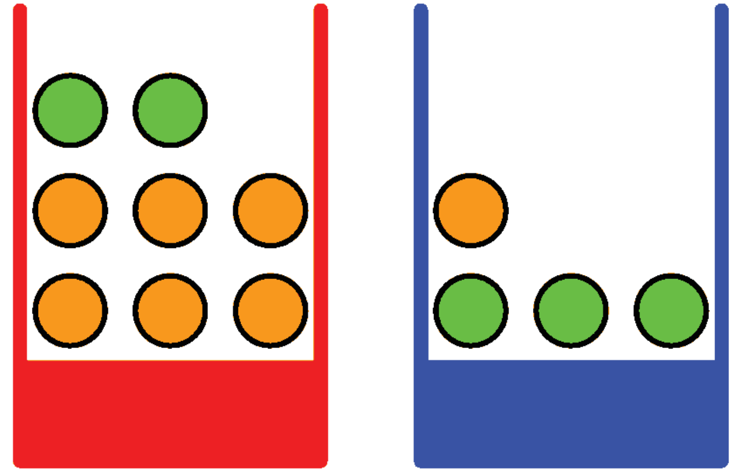
- $p(B = b, F = a) = p(B = b) * p(F = a \mid B = b)$

  $= 0.6 * 0.75 = 0.45.$

- In general, conditional probabilities and joint probabilities are connected with the following formula:

$$p(X, Y) = p(X) * p(Y \mid X) = p(Y) * p(X \mid Y)$$

The Product Rule

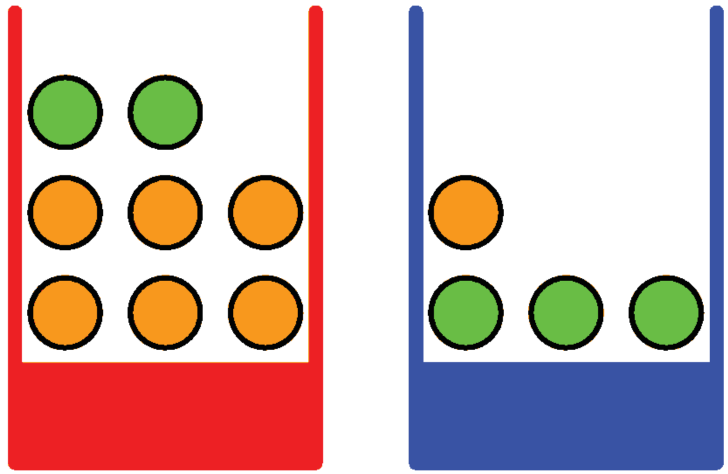


- In general, conditional probabilities and joint probabilities are connected with the following formula:

$$p(X, Y) = p(X) * p(Y \mid X) = p(Y) * p(X \mid Y)$$

- This formula is called the **product rule, and** can be used both ways:
  - You can compute conditional probabilities if you know the corresponding joint probabilities.
  - You can compute joint probabilities if you know the corresponding conditional probabilities.

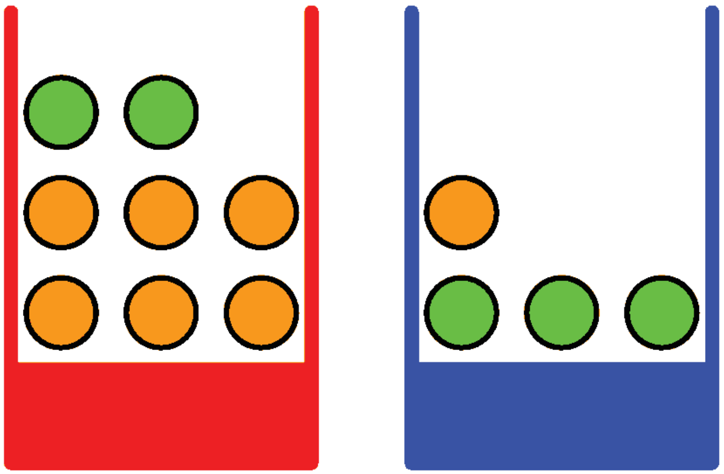
The Product Rule – Chained Version



- If $X_1, \ldots, X_n$ are n random variables:

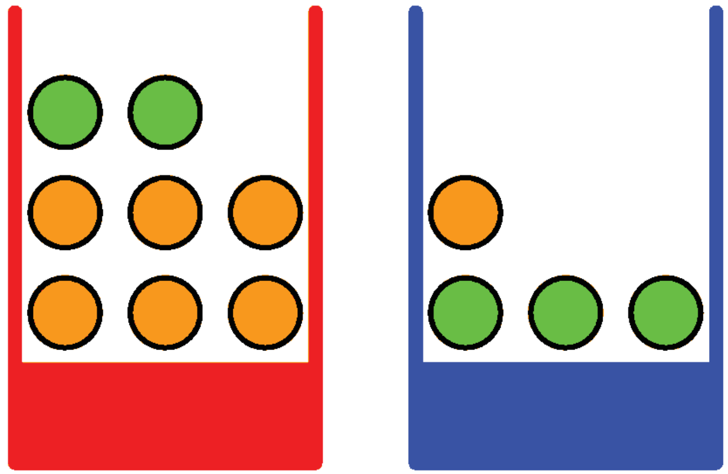
$$
\begin{aligned}
p(X_1, \ldots, X_n) = \; & p(X_1 | X_2, \ldots, X_n) * \\
& p(X_2 | X_3, \ldots, X_n) * \\
& p(X_3 | X_4, \ldots, X_n) * \\
& \ldots * \\
& p(X_{n-1} | X_n) * \\
& p(X_n)
\end{aligned}
$$

More Random Variables



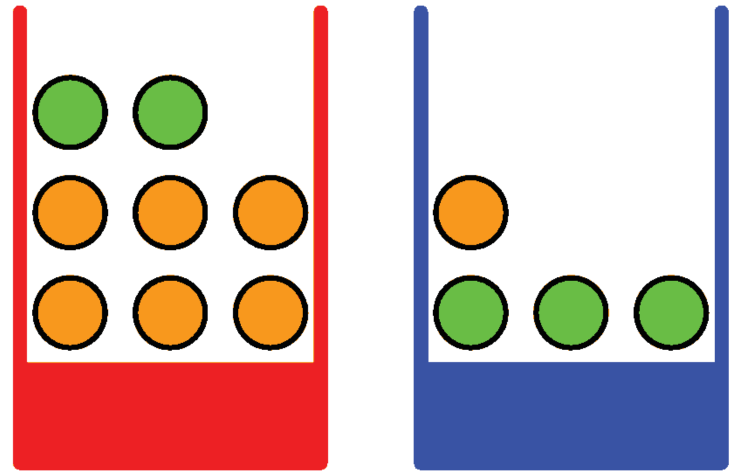
- Suppose we that we conduct two experiments, using the same protocol we described before.
- This way, we obtain four random variables:
  - B_1: the identity of the first box we pick.
  - F_1: the identity of the first fruit we pick.
  - B_2: the identity of the second box we pick.
  - F_2: the identity of the second fruit we pick.

Independence



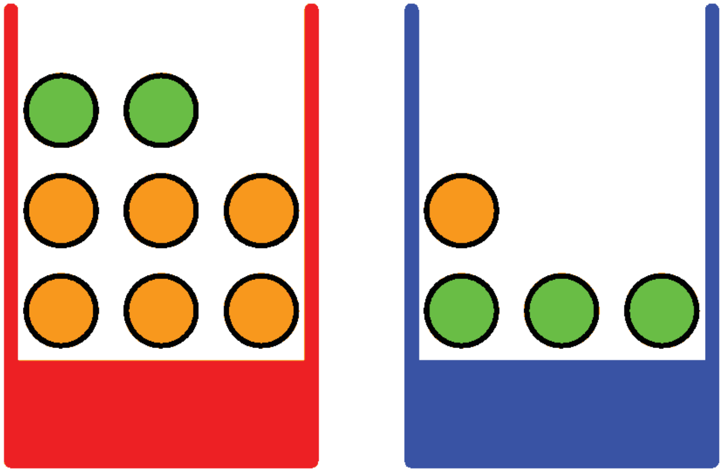
- What is $p(B_1 = r)$?
  What is $p(B_2 = r)$?

Independence



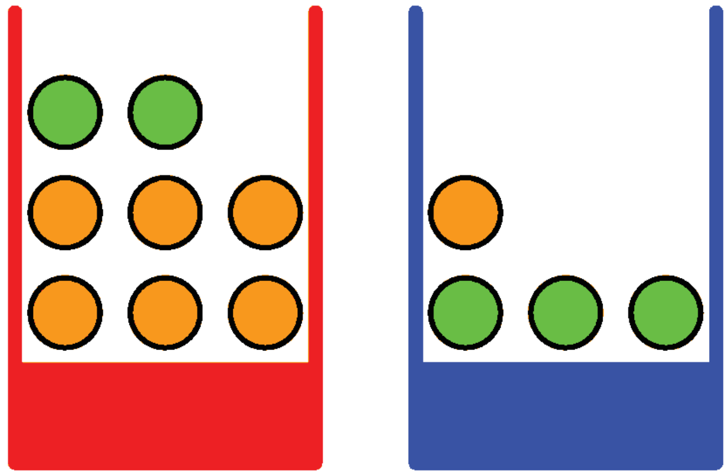
- What is $p(B_1 = r)$?
  What is $p(B_2 = r)$?
  - $p(B_1 = r) = p(B_2 = r) = 0.4$.
  - Why?

Independence



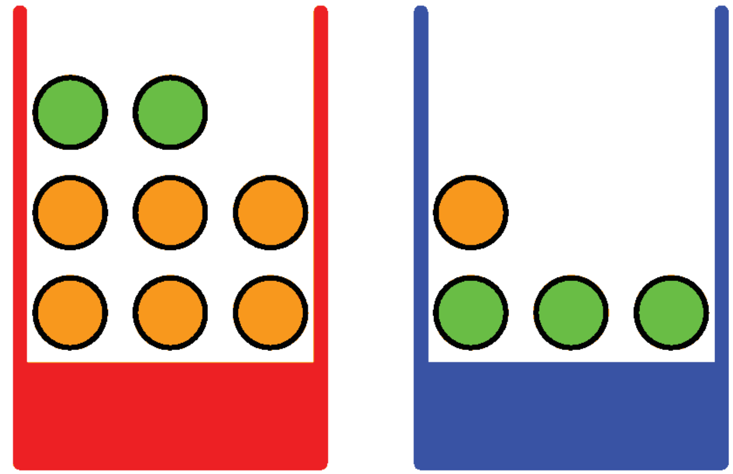
- What is $p(B_1 = r)$?
  What is $p(B_2 = r)$?
  - $p(B_1 = r) = p(B_2 = r) = 0.4$.
  - Why? Because each time we pick a box randomly, with the same odds of picking red or blue as any other time.

Independence



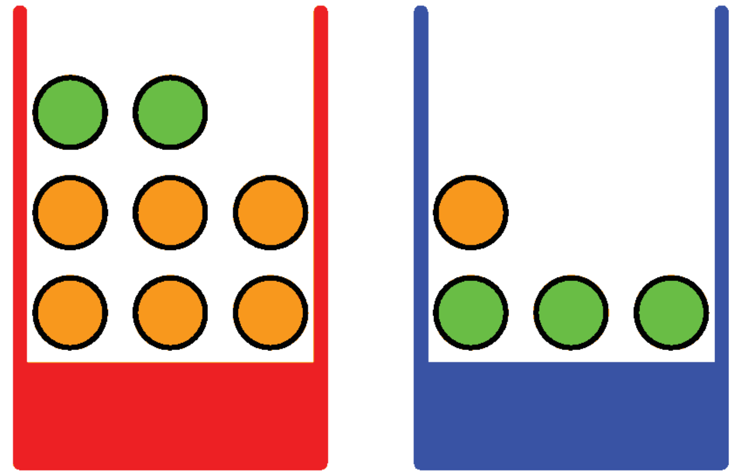
- What is $p(B_1 = r)$?
  What is $p(B_2 = r)$?
  - $p(B_1 = r) = p(B_2 = r) = 0.4$.
  - Why? Because each time we pick a box randomly, with the same odds of picking red or blue as any other time.
- What is $p(B_2 = r \mid B_1 = r)$?

Independence



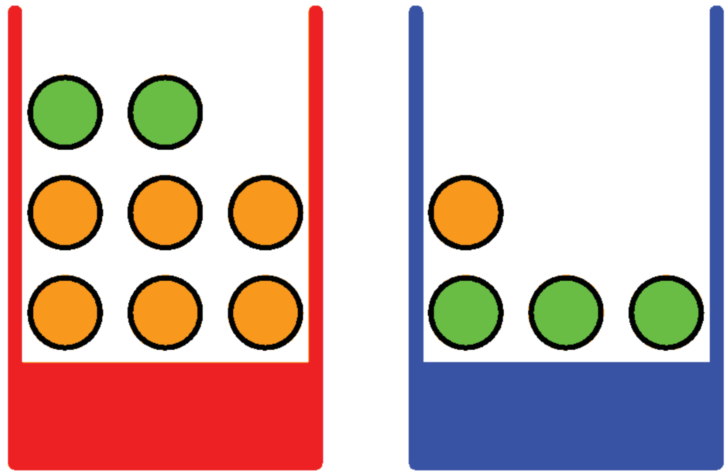
- What is $p(B_1 = r)$?
  What is $p(B_2 = r)$?
  – $p(B_1 = r) = p(B_2 = r) = 0.4$.
  – Why? Because each time we pick a box randomly, with the same odds of picking red or blue as any other time.
- What is $p(B_2 = r \mid B_1 = r)$?
  $p(B_2 = r \mid B_1 = r) = p(B_2 = r) = 0.4$
- Why?

Independence



- What is $p(B_1 = r)$?
  What is $p(B_2 = r)$?
  - $p(B_1 = r) = p(B_2 = r) = 0.4$.
  - Why? Because each time we pick a box randomly, with the same odds of picking red or blue as any other time.
- What is $p(B_2 = r \mid B_1 = r)$?
  $p(B_2 = r \mid B_1 = r) = p(B_2 = r) = 0.4$
- Why? Because the odds of picking red or blue remain the same every time.
- We say that **B_2 is independent of B_1**.

Independence

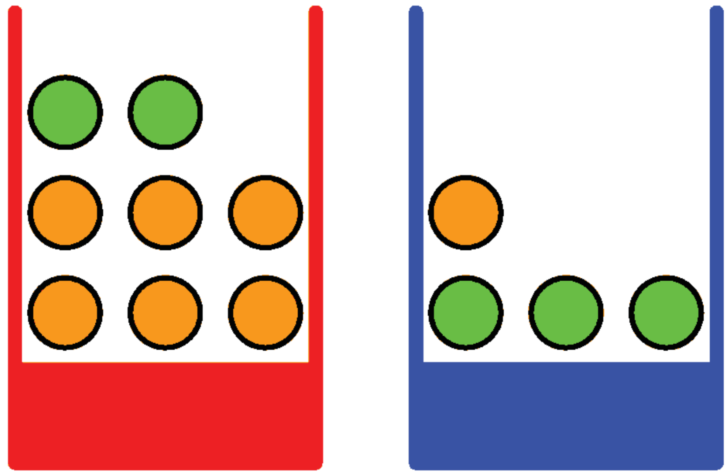


- In general, if we have two random variables X and Y, we say that X is independent of Y if and only if:

$$p(X \mid Y) = p(X)$$

- X is independent of Y if and only if Y is independent of X.

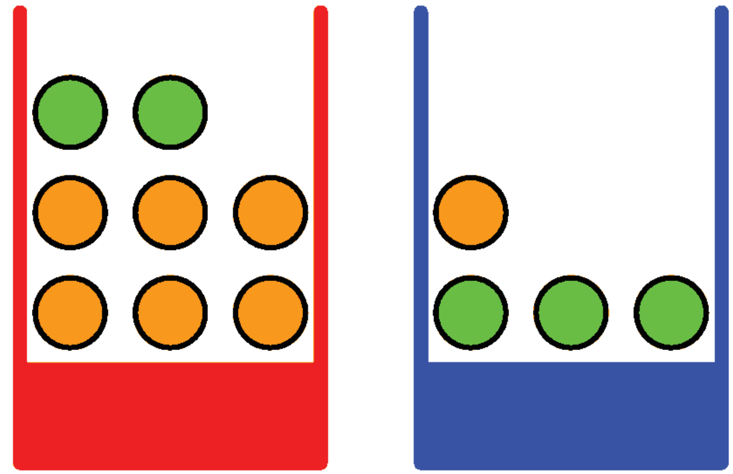
The Product Rule for Independent Variables



- The product rule states that, for any random variables X and Y, p(X, Y) = p(X) * p(Y | X).
- If X and Y are independent, then p(Y | X) = p(Y).
- Therefore, if X and Y are independent:
  p(X, Y) = p(X) * p(Y).
- If X_1, …, X_n are pairwise independent random variables:

$$p(X_1, \dots, X_n) = \prod_{i=1}^{n} p(X_i)$$

The Sum Rule



- What is $p(F_1 = a)$?

$p(F_1 = a) = p(F_1 = a, B_1 = r) + p(F_1 = a, B_1 = b)$

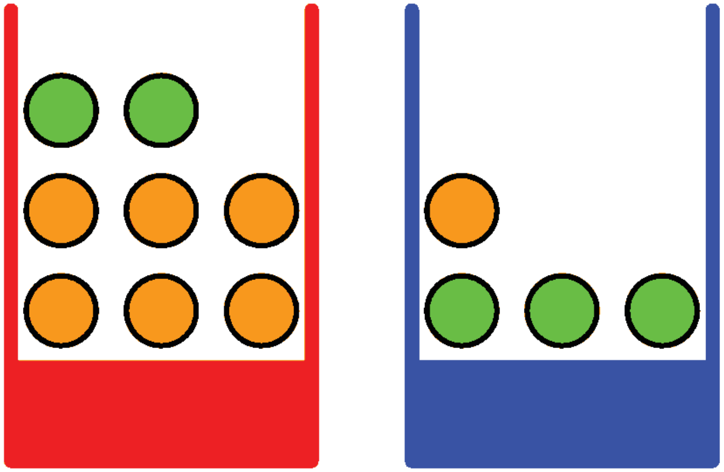
- Equivalently, we can compute $p(F_1 = a)$ as:

$p(F_1 = a) = p(F_1 = a \mid B_1 = r)\, P(B_1 = r) + p(F_1 = a \mid B_1 = b)\, P(B_1 = b)$

- Those formulas are the two versions of the **sum rule**.
- In general, for any two random variables X and Y: suppose that Y takes values from some set $\mathbb{Y}$. Then, the **sum rule** is stated as follows:

$$p(X) = \sum_{y \in \mathbb{Y}} p(X, Y = y) \quad = \sum_{y \in \mathbb{Y}} p(X \mid Y = y)\, p(Y)$$

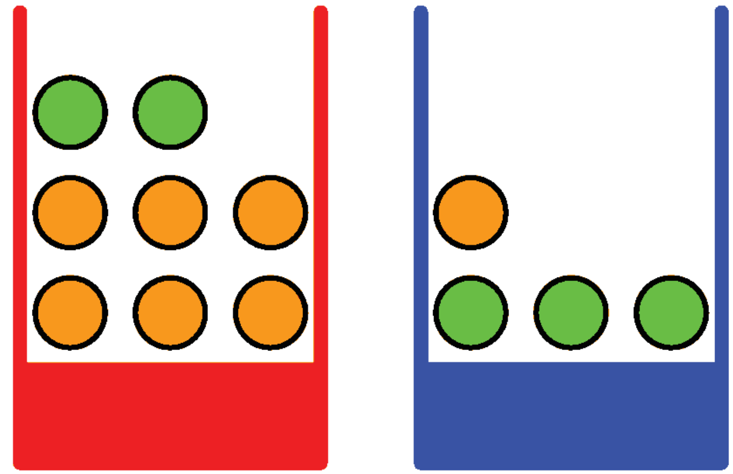
The Sum Rule



- Applying the sum rule:

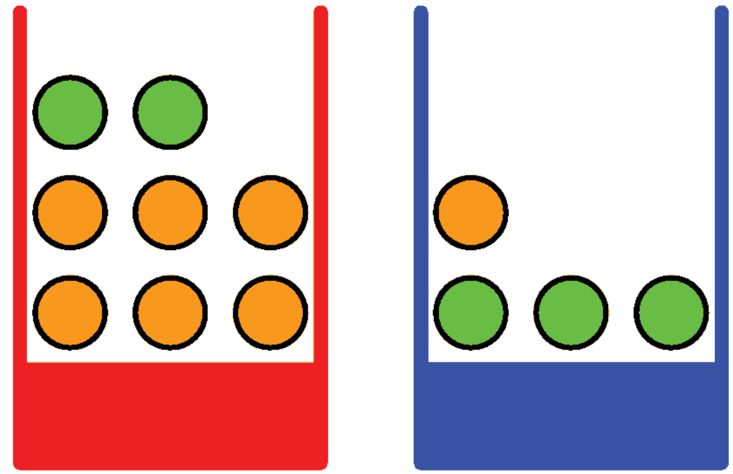
$$p(F_1 = a) = p(F_1 = a \mid B_1 = r)\, P(B_1 = r) + p(F_1 = a \mid B_1 = b)\, P(B_1 = b)$$
$$= 0.25 * 0.4 + 0.75 * 0.6 = 0.55$$

Another Example



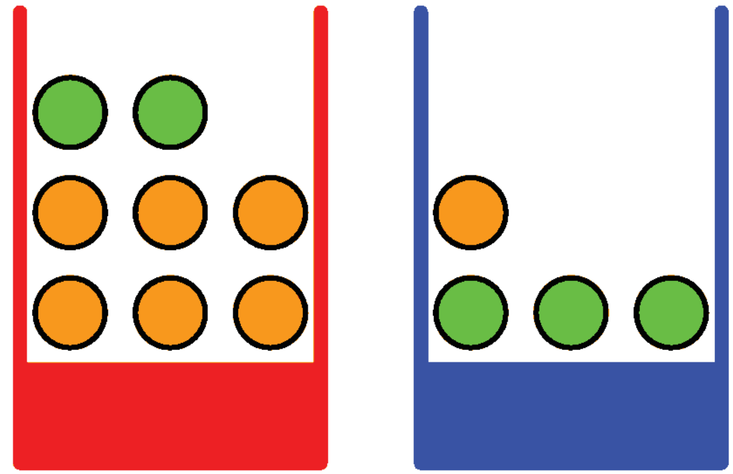
- Is F_1 independent of B_1?
- $p(F_1 = a) = 0.55$ (see previous slides).
- $p(F_1 = a \mid B_1 = r) = 0.25$
- $p(F_1 = a) \neq p(F_1 = a \mid B_1 = r)$, therefore F_1 and B_1 **are not independent**.
  - Note: to prove that F_1 and B_1 are independent, we would need to verify that $(F_1 = x) = p(F_1 = f \mid B_1 = y)$ **for every possible value x of F_1 and y of B_1**.
  - However, finding a single case, such as $(F_1 = a, B_1 = r)$, where $p(F_1 = a) \neq p(F_1 = a \mid B_1 = r)$, is sufficient to prove that F_1 and B_1 are **not independent**.

Bayes Rule



- Suppose that $F_1 = a$.
  - The first fruit we picked is an apple.
- What is $p(B_1 = r \mid F_1 = a)$?

Bayes Rule

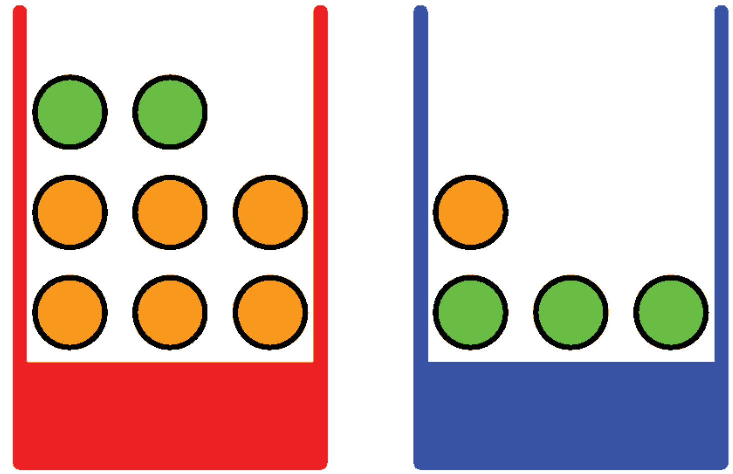


- Suppose that F_1 = a.
  - The first fruit we picked is an apple.
- What is $p(B_1 = r \mid F_1 = a)$?
- This can be computed using **Bayes rule**: if X and Y are any random variables, then:

$$p(X \mid Y) = \frac{p(Y \mid X)\, p(X)}{p(Y)}$$

- Where is this formula coming from?

Bayes Rule



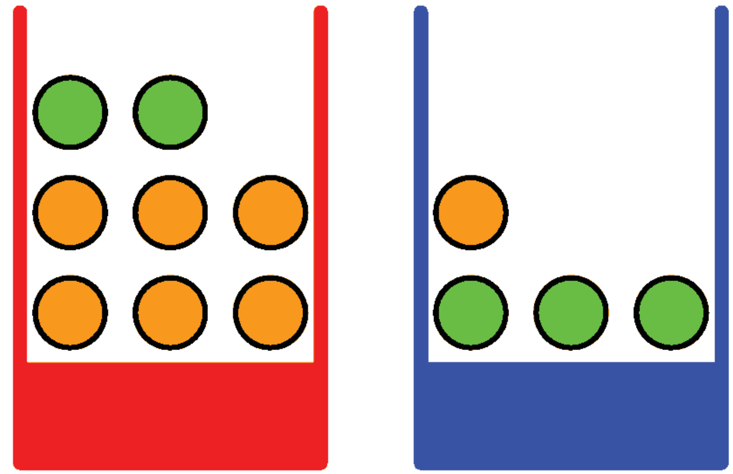
- Suppose that F_1 = a.
  - The first fruit we picked is an apple.
- What is $p(B_1 = r \mid F_1 = a)$?
- This can be computed using **Bayes rule**: if X and Y are any random variables, then:

$$p(X \mid Y) = \frac{p(Y \mid X)\, p(X)}{p(Y)}$$

- This formula comes from the relationship between conditional and joint probabilities:

$$p(X, Y) = p(X \mid Y) P(Y) = p(Y \mid X)\, p(X)$$

Bayes Rule



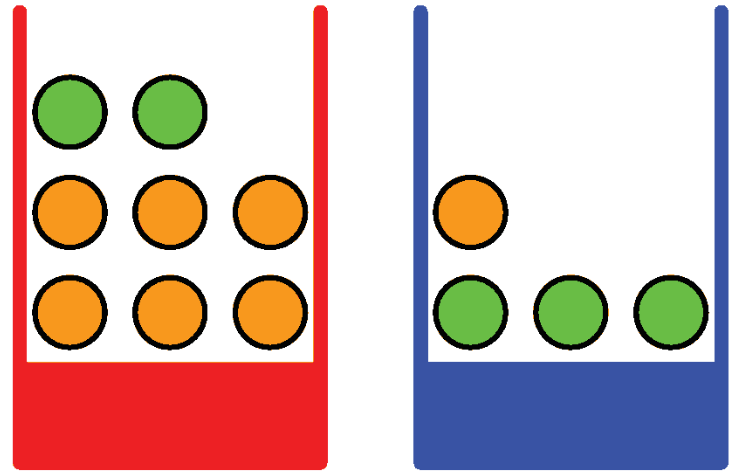
- Suppose that F_1 = a.
  - The first fruit we picked is an apple.
- What is $p(B_1 = r \mid F_1 = a)$?
- In our case, Bayes rule is applied as follows:

$$p(B_1 = r \mid F_1 = a) = \frac{p(F_1 = a \mid B_1 = r)\, p(B_1 = r)}{p(F_1 = a)}$$

$$= \frac{0.25 \ * 0.4}{0.55} = 0.1818$$

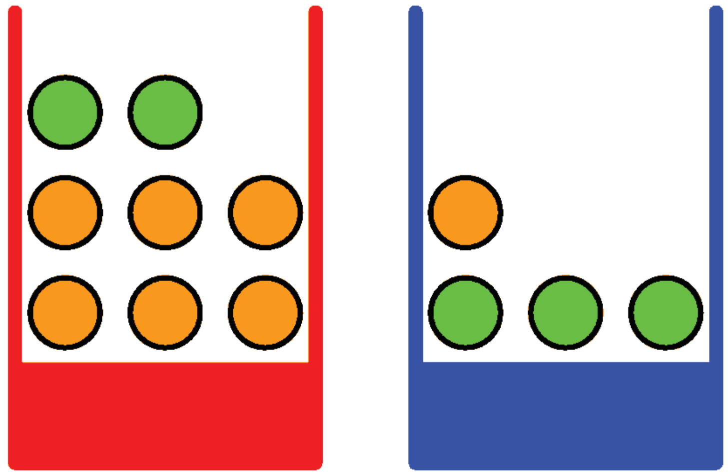
- Reminder: We computed earlier, using the sum rule, that $P(F_1 = a) = 0.55$.

Priors and Posteriors



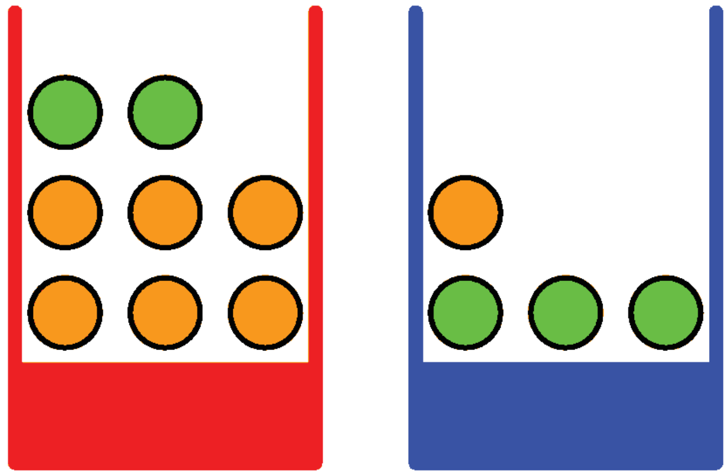
- So: before we knew that the first fruit is an apple, we had $p(B_1 = r) = 0.4$.
- This is called the **prior probability** of $B_1 = r$.
  - It is called **prior**, because it is the default probability, when no other knowledge is available.
- After we saw that the first fruit was an apple, we have $p(B_1 = r \mid F_1 = a) = 0.1818$
- This is called the **posterior probability** of $B_1 = r$, given the knowledge that the first fruit was an apple.

Conditional Independence



- Let's modify our protocol:
- We pick a box B, with odds as before:
  - We pick red 40% of the times, blue 60% of the times.
- We pick a fruit of type F_1.
  - All fruits in the box have equal chances of getting picked.
- We put that first fruit back in the box.
- We pick a second fruit of type F_2 **from the same box B**.
  - Again, all fruits in the box have equal chances of getting picked.
  - Possibly we pick the same fruit as the first time.

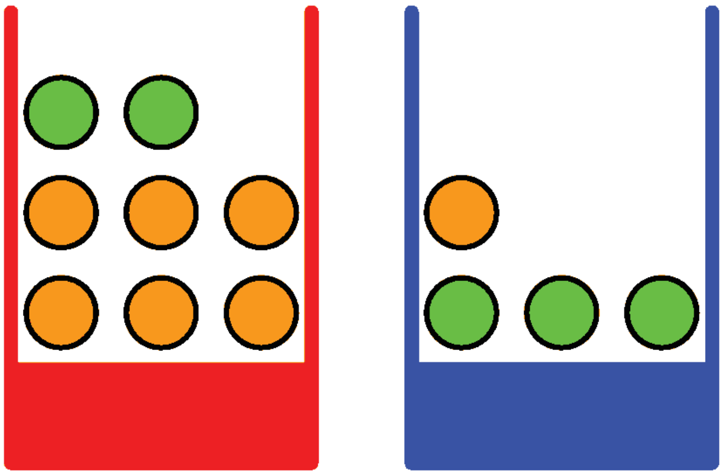
Conditional Independence



- Using this new protocol:
  Are F_1 and F_2 independent?
- F_1 and F_2 are independent iff $p(F_2) = p(F_2 \mid F_1)$.
- So, we must compute and compare $p(F_2)$ and $p(F_2 \mid F_1)$.
- By applying the sum rule, we already computed that:

$$p(F_2 = a) = p(F_2 = a \mid B = r)\, p(B = r) + p(F_2 = a \mid B = b)\, p(B = b) = 0.55.$$

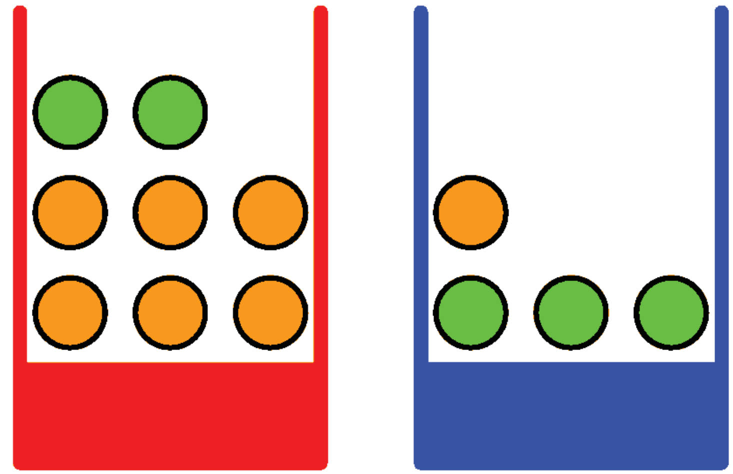
Conditional Independence



$p(F_2 = a \mid F_1 = a)$
$= p(F_2 = a \mid F_1 = a, B = r)\, p(B = r \mid F_1 = a) +$
$p(F_2 = a \mid F_1 = a , B = b)\, p(B = b \mid F_1 = a)$

- Here, note that $p(F_2 = a \mid F_1 = a, B = r) = p(F_2 = a \mid B = r)$.
- Why is that true?

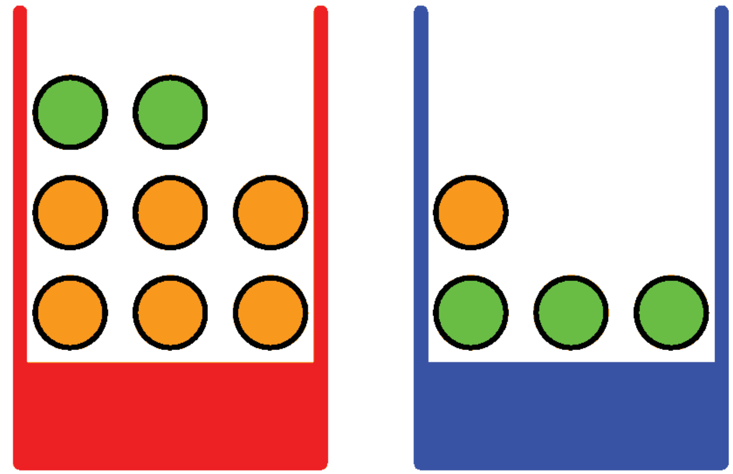
Conditional Independence



$p(F_2 = a | F_1 = a)$
$= p(F_2 = a \mid F_1 = a, B = r)\, p(B = r \mid F_1 = a) +$
$p(F_2 = a \mid F_1 = a , B = b)\, p(B = b \mid F_1 = a)$

- Here, note that $p(F_2 = a \mid F_1 = a, B = r) = p(F_2 = a \mid B = r)$.
- Why is that true?
  - If we know that B = r, the first fruit does not provide any additional information about the second fruit.
- If X, Y, Z are random variables, we say that X and Y are **conditionally independent given Z** when: $p(X \mid Y, Z) = p(X \mid Z)$.
- Thus, F_2 and F_1 are conditionally independent given B.

Conditional Independence



- Continuing with our computation:

$p(F_2 = a | F_1 = a)$
$= p(F_2 = a \mid F_1 = a, B = r)\ p(B = r \mid F_1 = a) +$
$\quad p(F_2 = a \mid F_1 = a, B = b)\ p(B = b \mid F_1 = a)$

- F_2 and F_1 are conditionally independent given B, so we get:

$p(F_2 = a \mid B = r)\ p(B = r \mid F_1 = a) + p(F_2 = a \mid B = b)\ p(B = b \mid F_1 = a)$

- Using values computed in earlier slides, we get:

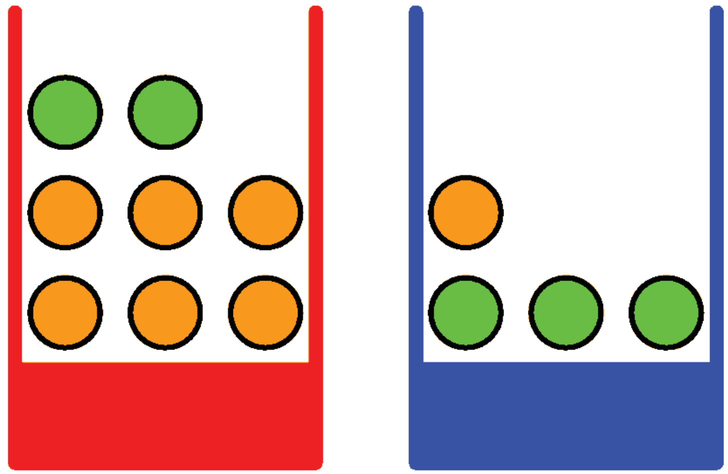
$0.25 * 0.1818 + 0.75 * p(B = b \mid F_1 = a)$

$= 0.25 * 0.1818 + 0.75 * p(F_1 = a \mid B = b) * p(B = b) / p(F_1 = a)$

- We use Bayes rule to compute $p(B = b \mid F_1 = a)$, so we get:

$0.25 * 0.1818 + 0.75 * 0.75 * 0.6 / 0.55 = 0.6591.$

Conditional Independence



- Putting the previous results together:
- $p(F_2 = a) = 0.55$.
- $p(F_2 = a \mid F_1 = a) = 0.6591$.
- So, $P(F_2) \neq P(F_2 \mid F_1)$. Therefore, F_1 and F_2 are NOT independent.
- On the other hand: $p(F_2 \mid F_1, B) = p(F_2 \mid B)$.
- Therefore, F_1 and F_2 are **conditionally independent** given B.

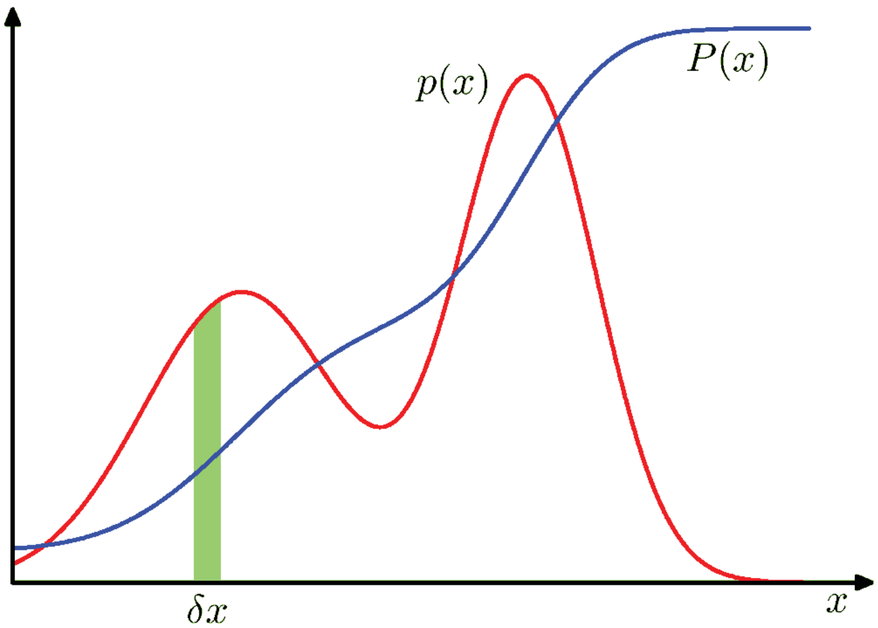
Regarding Notation

- Suppose that X and Y are random variables, and suppose that c and d are some values that X and Y can take.
- If $p(X = c) = p(X = c \mid Y = d)$, does this mean that X and Y are independent?

Regarding Notation

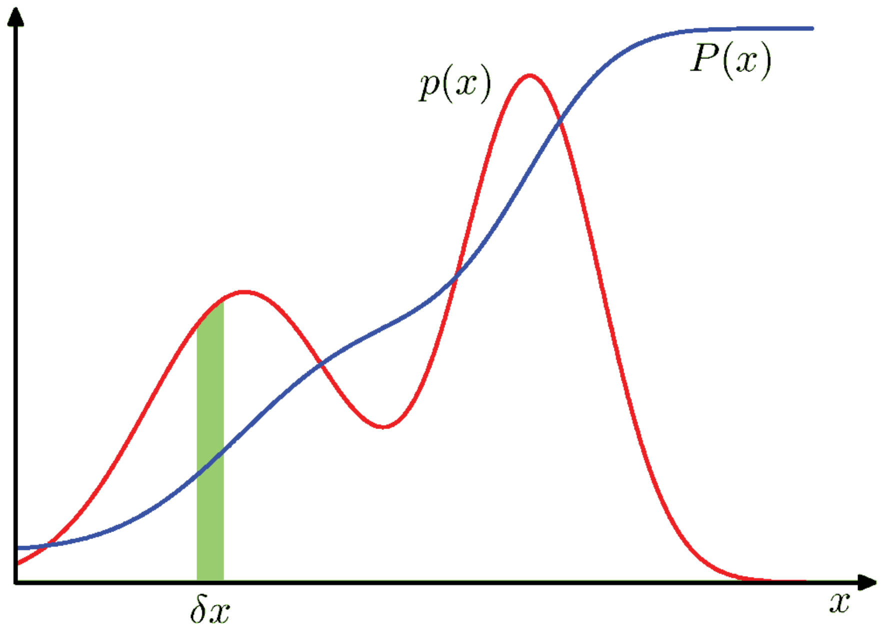
- Suppose that X and Y are random variables, and suppose that c and d are some values that X and Y can take.
- If $p(X = c) = p(X = c \mid Y = d)$, does this mean that X and Y are independent?
- NO. The requirement for independence is that:
  $p(X) = p(X \mid Y)$.
- $p(X) = p(X \mid Y)$ means that, **for any possible value x of X, any possible value y of y, $p(X = x) = p(X = x \mid Y = y)$**.
- If $p(X = c) = p(X = c \mid Y = d)$, that information regards only some specific values of X and Y, not all possible values.

Probability Densities



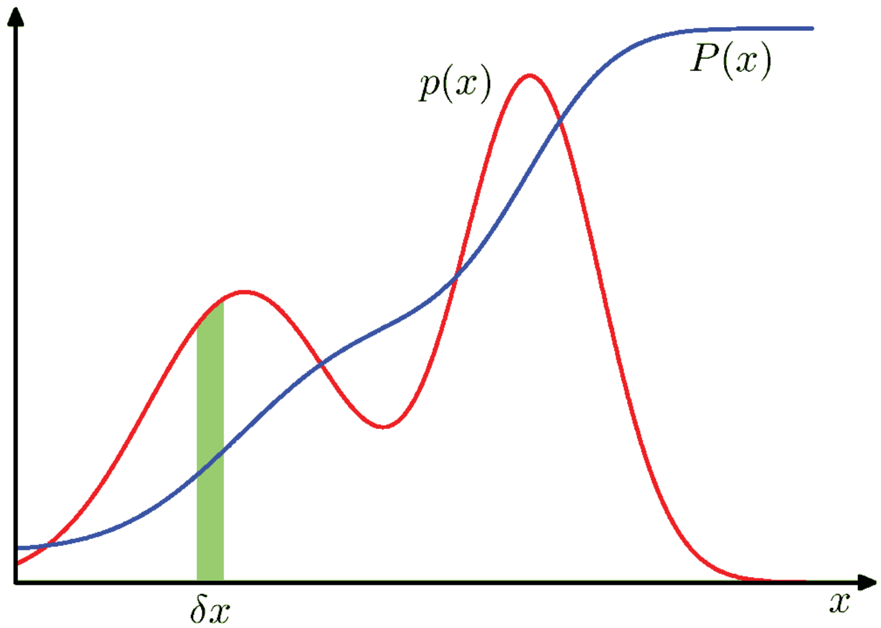
- Some times, random variables take values from a continuous space.
  - For example: temperature, time, length.
- In those cases, typically (but not always) the probability of any specific value is 0. What we care about is the probability of values belonging to a certain range.

Probability Densities



- Suppose X is a real-valued random variable X.
- Consider a very small number δx.
- Intuitively, the probability density P(X = x) expresses the probability that the value of X falls in the interval (x, x + δx), **<u>divided by</u>** δx.

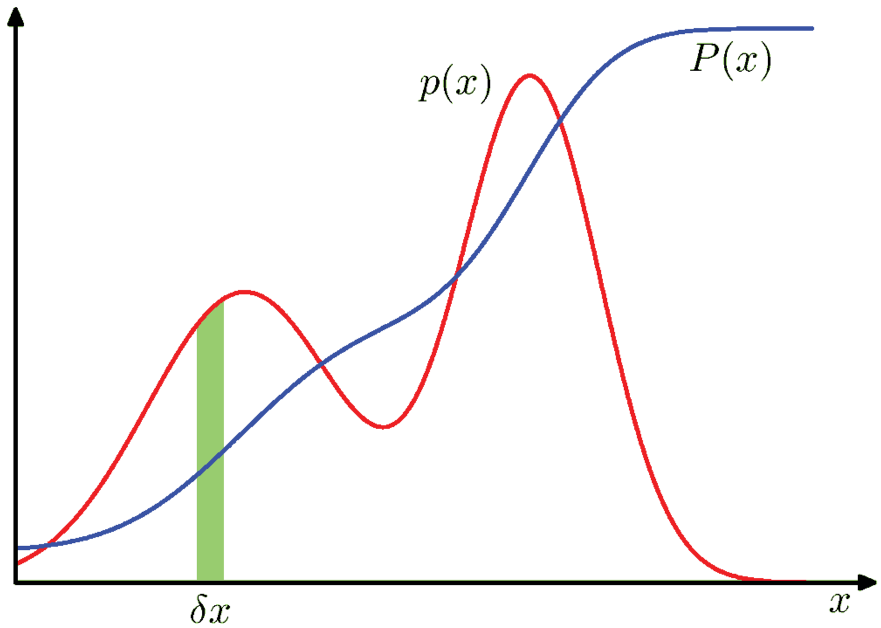
Probability Densities



- Mathematically: the **probability density** P(x) of a random variable X is defined as:

$$P(x) = \lim_{\delta x \to 0} \frac{p(X \in (x, x + \delta x))}{\delta x}$$

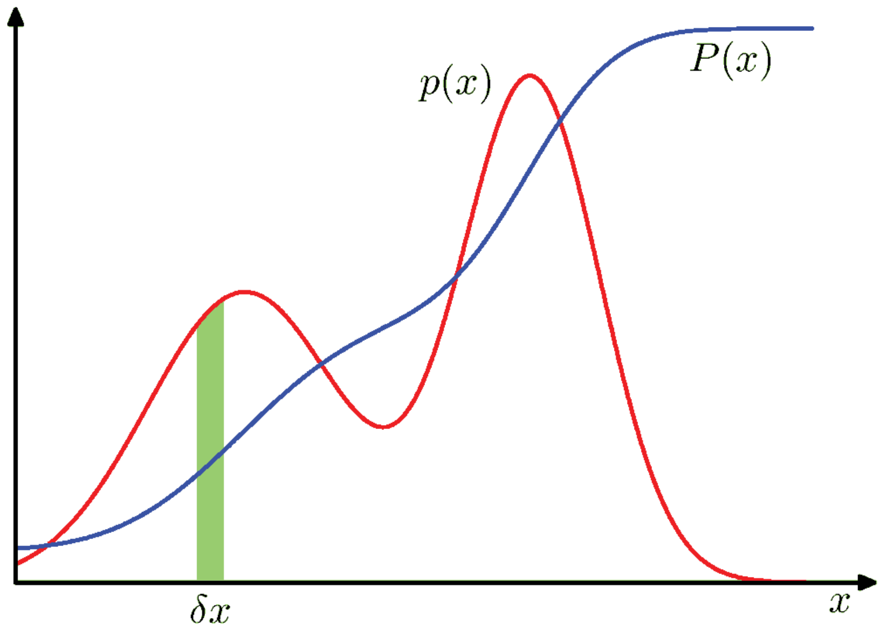
Integrating over Probability Densities



- To compute the probability that X belongs to an interval (a, b), we integrate over the density P(x):

$$p(X \in (a,b)) = \int_a^b P(x)dx$$

Constraints on Density Functions



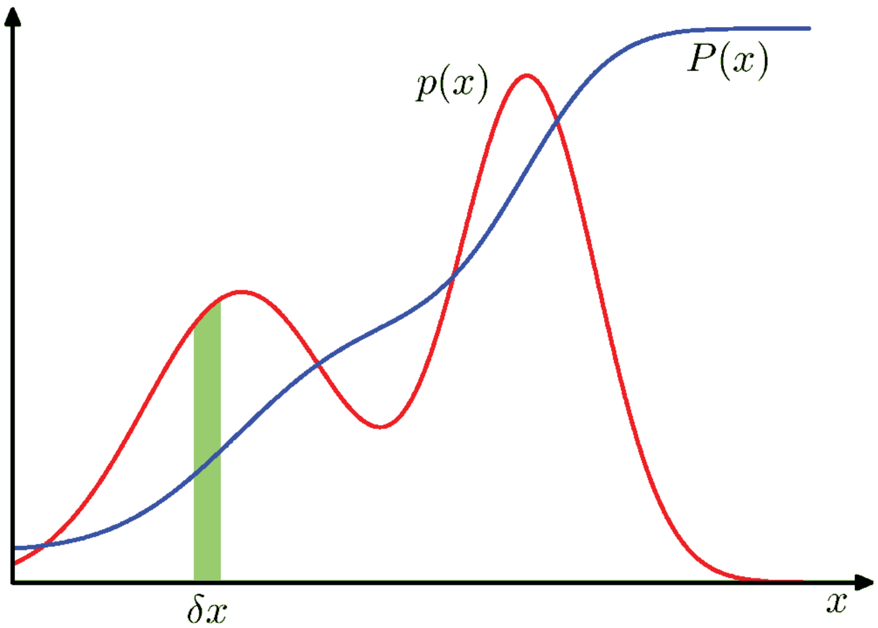
- Note that P(x) can be larger than 1, because P(x) is a **density**, not a probability.
- However, $p(X \in (a,b)) \leq 1$, always.
- P(x) >= 0, always. We cannot have negative probabilities or negative densities.
- $\int_{-\infty}^{\infty} P(x)dx = 1$. A real-valued random variable x always has a value between $-\infty$ and ∞.

Example of Densities > 1

Here is a density function: $P(\mathrm{x}) = \begin{cases} 0, & \text{if x} < 5.3 \\ 10, & \text{if x} \epsilon [5.3, 5.4] \\ 0, & \text{if x} > 5.4 \end{cases}$

- **A density is not a probability**.
- A density is converted to a probability by integrating over an interval.
- A density can have values > 1, at some small range, as long as integrals over any interval are <= 1.
- In the example above:
  - $\forall a, b \int_a^b P(x)dx \leq 1$
  - $\int_{-\infty}^{\infty} P(x)dx = \int_{5.3}^{5.4} P(x)dx = 1$

Cumulative Distributions



- The probability that x belongs to the interval $(-\infty, z)$ is called the **cumulative distribution P(z).**

- P(z) can be computed by integrating the density over the interval $(-\infty, z)$:

$$P(z) = \int_{-\infty}^{z} p(x)dx$$

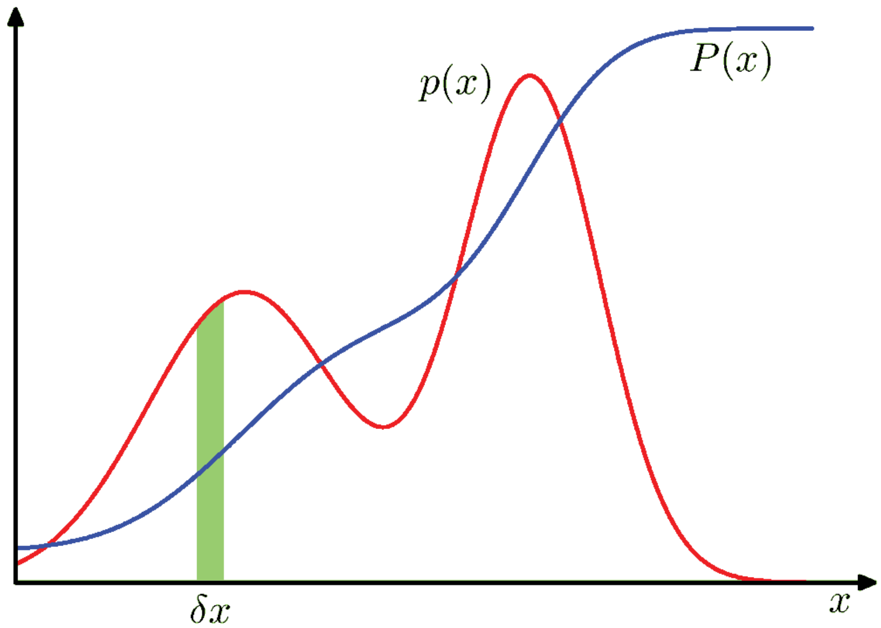
Higher Dimensions

- If we have several continuous random variables $x_1, \ldots, x_D$, we can define a **joint probability density** $P(x) = P(x_1, \ldots, x_D)$.
- It still must be the case that:

$$P(x) \geq 0$$

$$\int p(x)dx = 1$$

Sum and Product Rules for Densities



- Suppose that x and y are continuous random variables.
- The sum rule is written as:

$$P(x) = \int_{-\infty}^{\infty} P(x, y) dy$$

- The product rule remains the same:

$$P(x, y) = P(y \,|x)\, P(x)$$

Expectation

- The average value of some function f(x) under a probability distribution, or probability density, is called the **expectation** of f(x).
- The expectation of f(x) is denoted as $\mathbb{E}|f|$.
- If p(x) is a probability function:

$$\mathbb{E}|f| = \sum_{x} (p(x)f(x))$$

- If P(x) is a density function:

$$\mathbb{E}|f| = \int_{-\infty}^{\infty} P(x)f(x)dx$$

Mean Value

- The mean of a probability distribution is defined as:

$$\mathbb{E}|x| = \sum_{x} (p(x)x)$$

- The mean of a probability density function is defined as:

$$\mathbb{E}|x| = \int_{-\infty}^{\infty} P(x)\, x\, dx$$

Variance and Standard Deviation

- The variance of a probability distribution, or a probability density function, is defined in several equivalent ways, as:

$$var[x] = \mathbb{E}|x^2| - \mathbb{E}|x|^2$$
$$= \mathbb{E}|(x - \mathbb{E}|x|)^2|$$

- For probability functions, this becomes:

$$var[x] = \sum_x (p(x)(x - \mathbb{E}|x|)^2)$$

- For probability density functions, it becomes:

$$\mathbb{E}|x| = \int_{-\infty}^{\infty} P(x)(x - \mathbb{E}|x|)^2 dx$$

- The **standard deviation** of a probability distribution, or a probability density function, is the square root of its variance.

Gaussians

- A popular way to estimate **probability density functions** is to model them as Gaussians.
  - These Gaussian densities are also called **normal distributions**.
- In one dimension, a normal distribution is defined as:

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

- To define a Gaussian, what parameters do we need to specify?

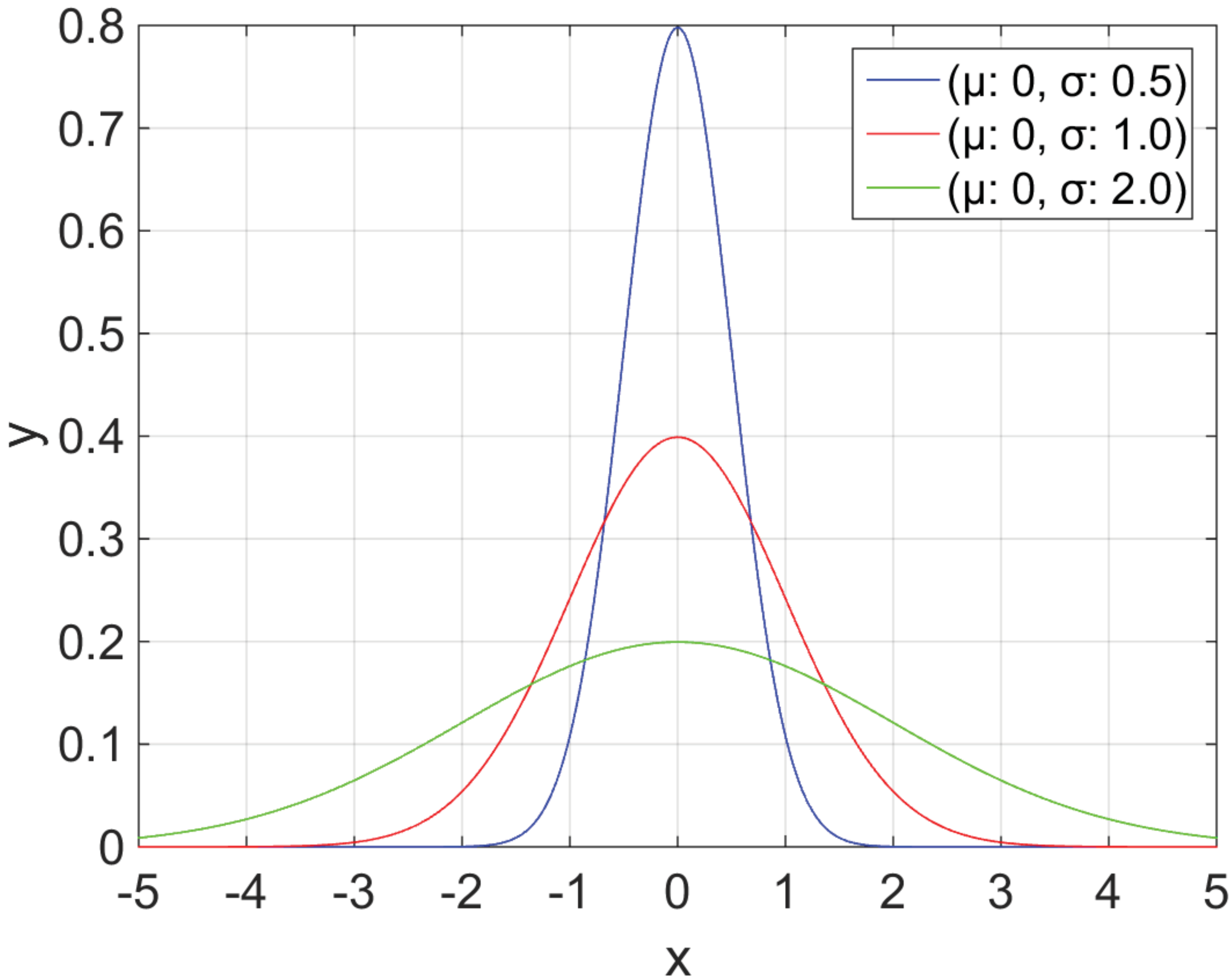
Gaussians

- A popular way to estimate **probability density functions** is to model them as Gaussians.
  - These Gaussian densities are also called **normal distributions**.
- In one dimension, a normal distribution is defined as:

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

- To define a Gaussian, what parameters do we need to specify? Just two parameters:
  - μ, which is the **mean** (average) of the distribution.
  - σ, which is the **standard deviation** of the distribution.
  - Note: σ^2 is obviously the **variance** of the distribution.

Examples of Gaussians

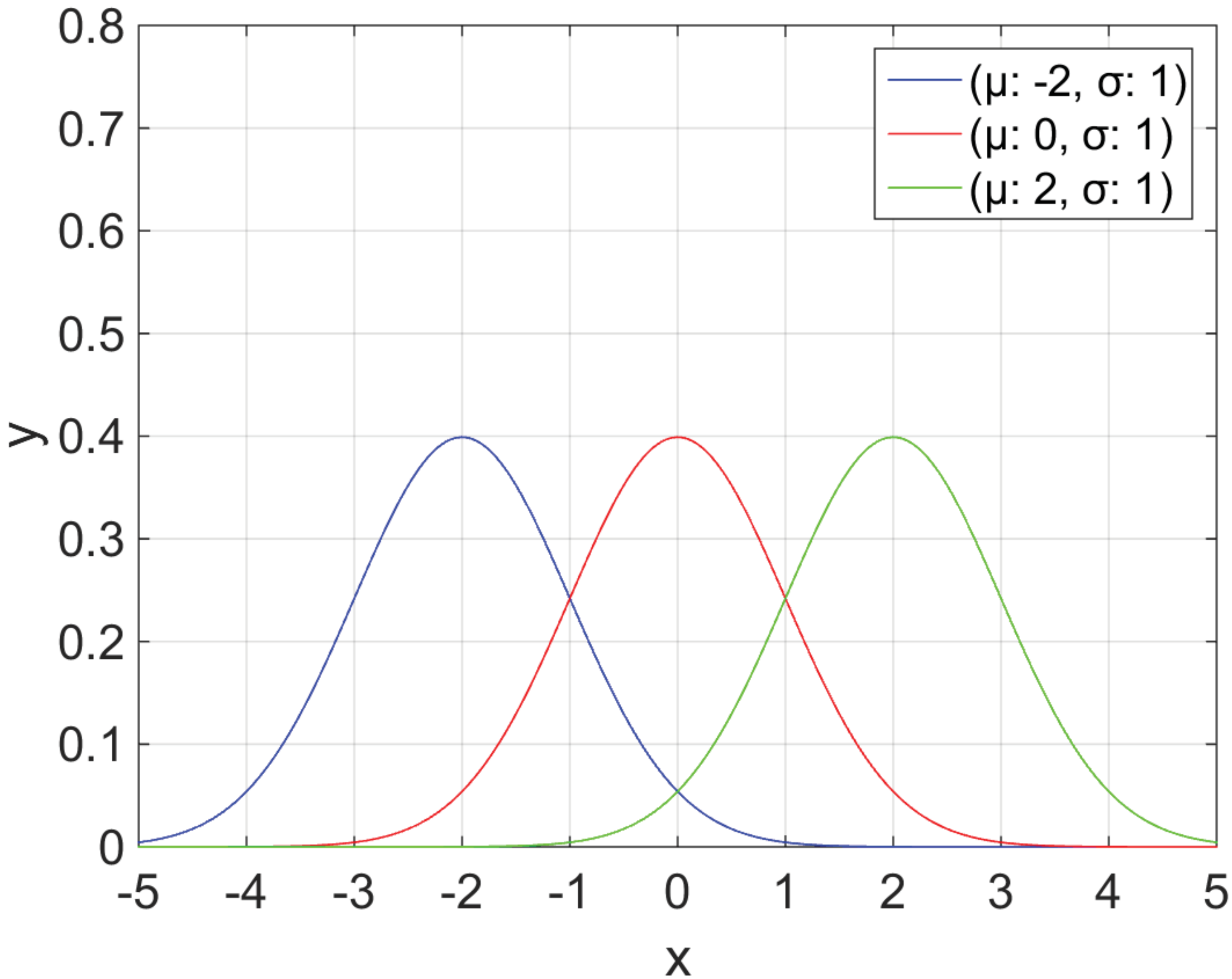


Increasing the standard deviation makes the values more spread out.

Decreasing the std makes the distribution more peaky.

The integral is always equal to 1.

Examples of Gaussians



Changing the mean moves the distribution to the left or to the right.

Estimating a Gaussian

- In one dimension, a Gaussian is defined like this:

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

- Given a set of n real numbers $x_1, \ldots, x_n$, we can easily find the best-fitting Gaussian for that data.
- The mean μ is simply the average of those numbers:

$$\mu = \frac{1}{n}\sum_{1}^{n} x_i$$

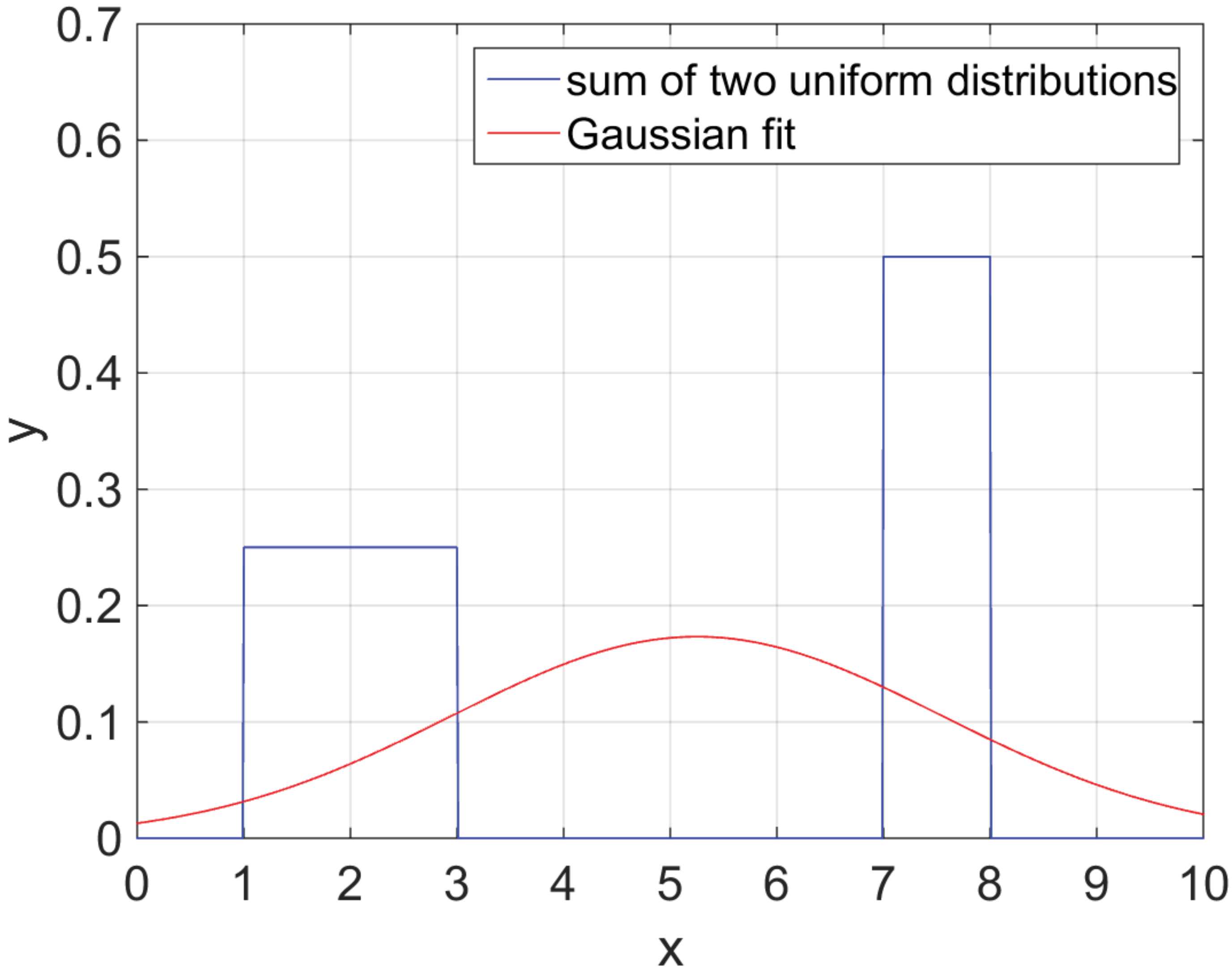
- The standard deviation σ is computed as:

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{1}^{n}(x_i - \mu)^2}$$

Estimating a Gaussian

- Fitting a Gaussian to data does not guarantee that the resulting Gaussian will be an accurate distribution for the data.
- The data may have a distribution that is very different from a Gaussian.
- This also happens when fitting a line to data.
  - We can estimate the parameters for the best-fitting line.
  - Still, the data itself may not look at all like a line.

Example of Fitting a Gaussian

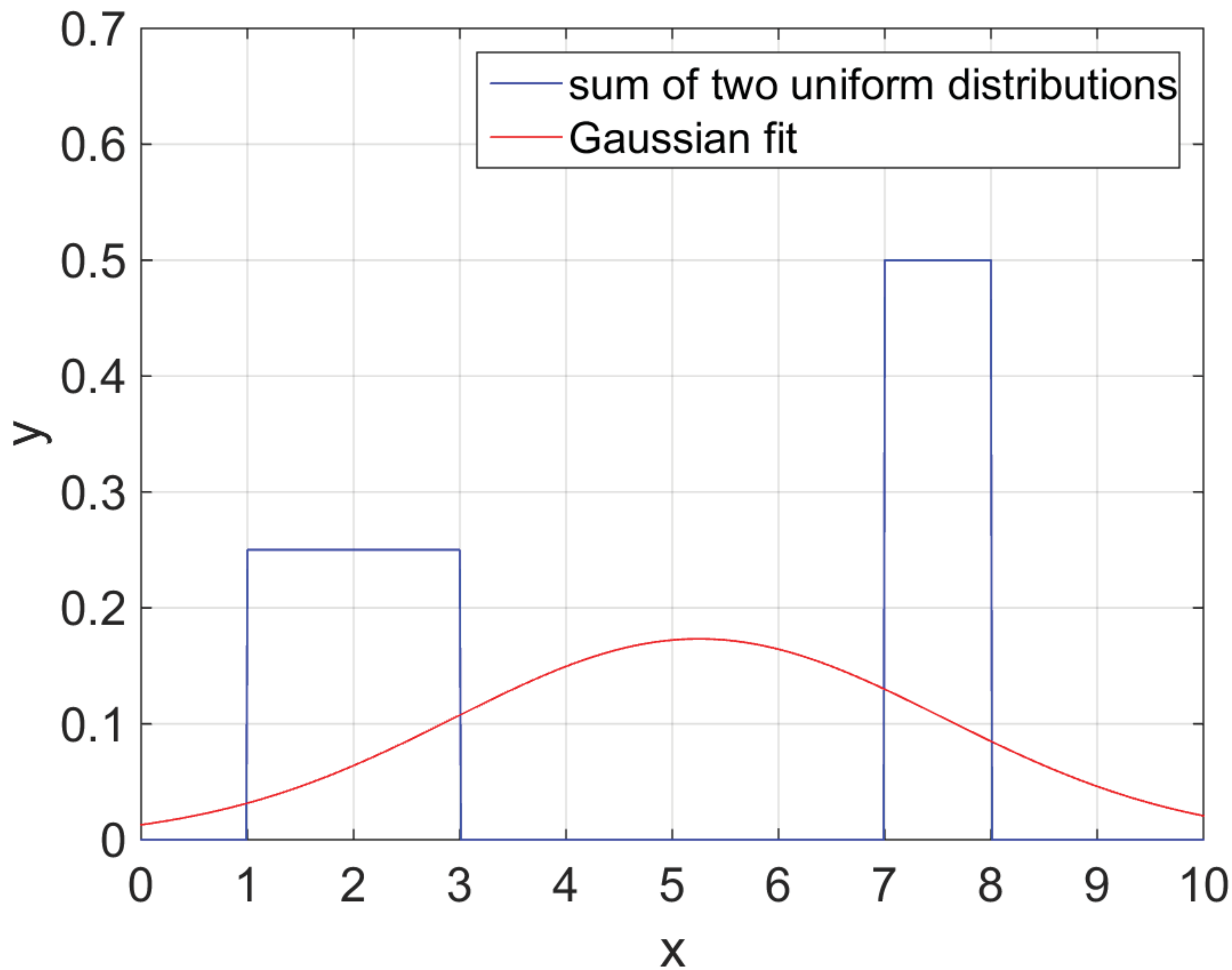


The blue curve is a density function F such that:

- $F(x) = 0.25$ for $1 \leq x \leq 3$.
- $F(x) = 0.5$ for $7 \leq x \leq 8$.

The red curve is the Gaussian fit G to data generated using F.

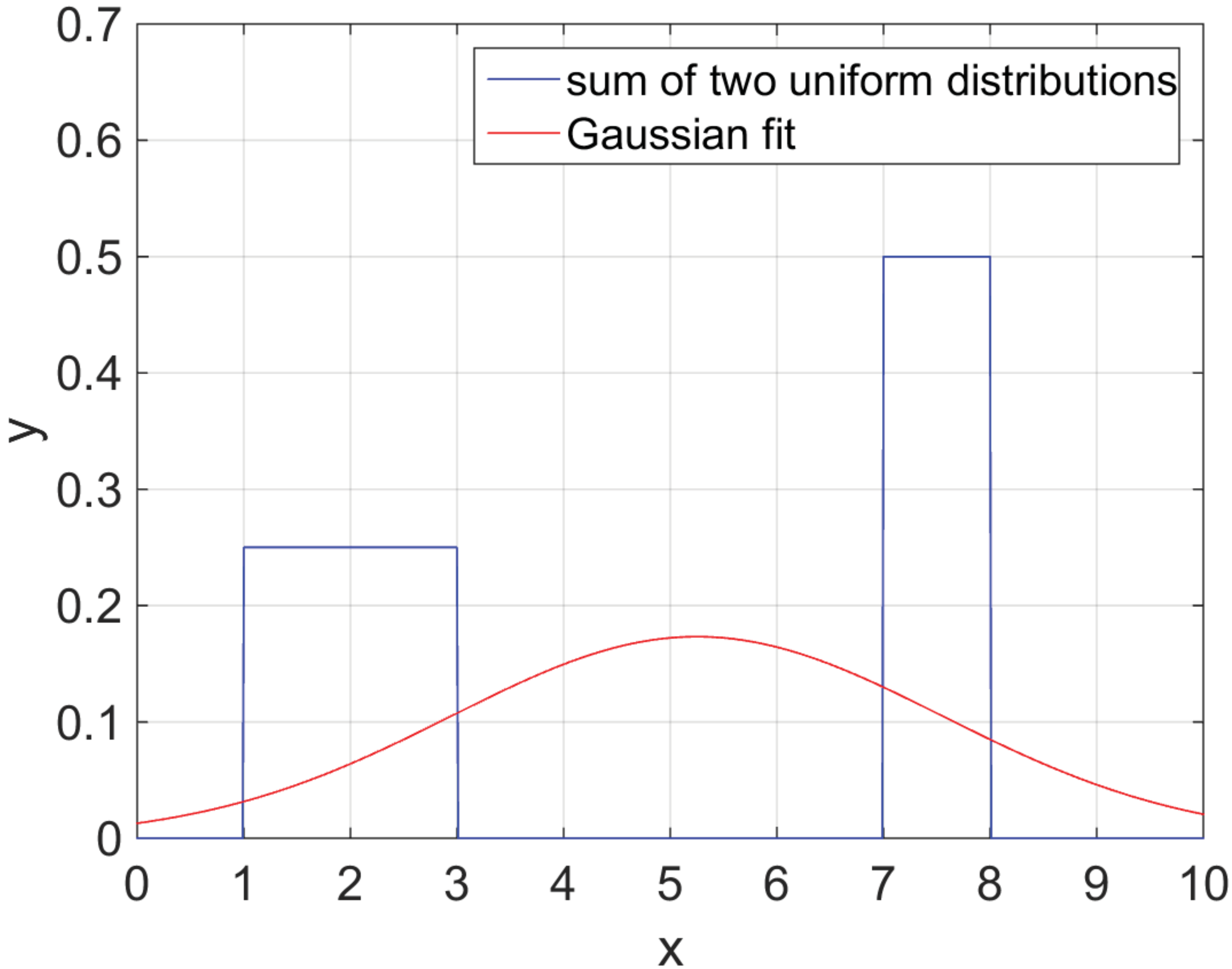
Example of Fitting a Gaussian



Note that the Gaussian does not fit the data well.

X	F(x)	G(x)
1	0.25	0.031
2	0.25	0.064
3	0.25	0.107
4	0	0.149
5	0	0.172
6	0	0.164
7	0.5	0.130
8	0.5	0.085

Example of Fitting a Gaussian

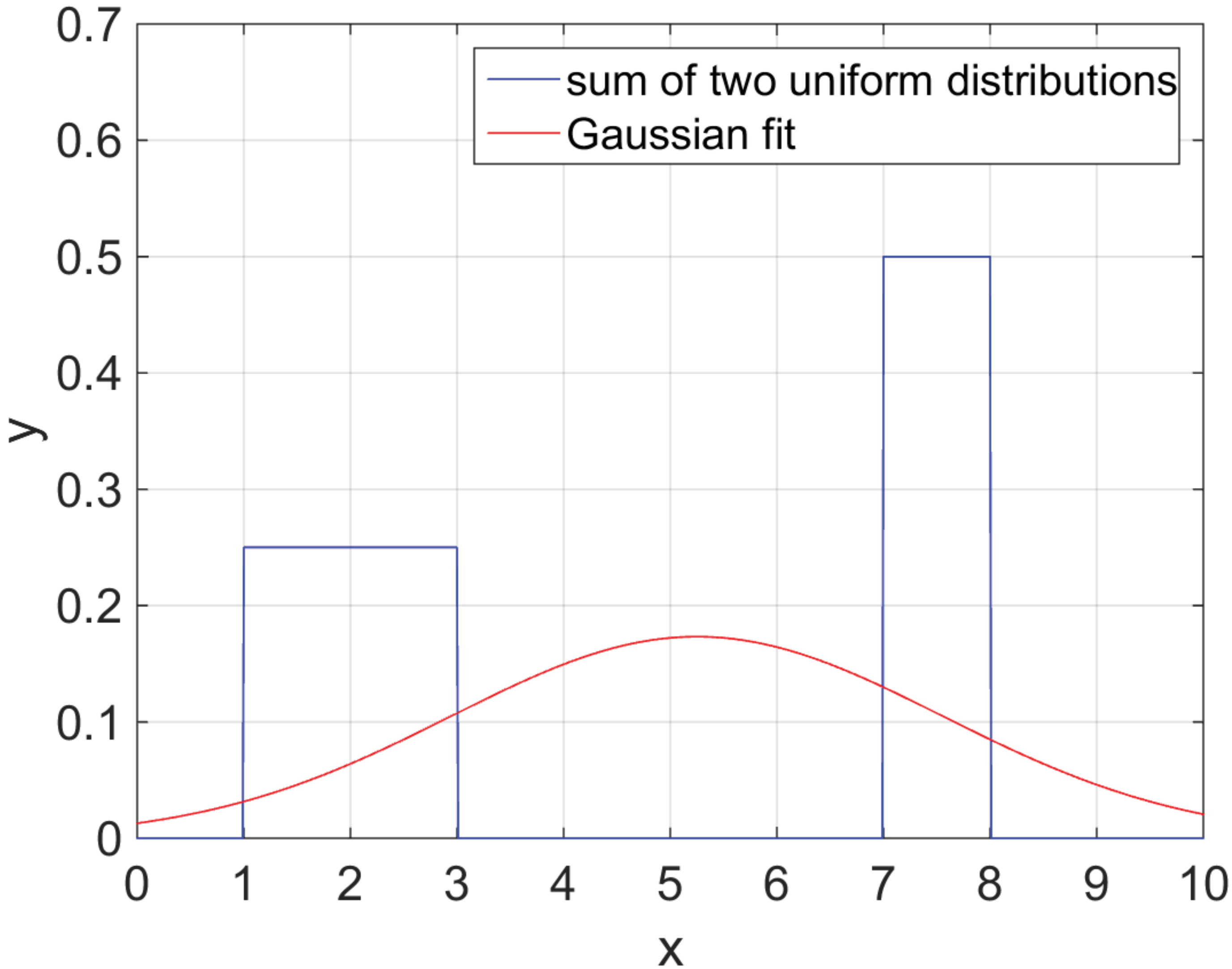


The peak value of G is 0.173, for x=5.25.

F(5.25) = 0!!!

X	F(x)	G(x)
1	0.25	0.031
2	0.25	0.064
3	0.25	0.107
4	0	0.149
5	0	0.172
6	0	0.164
7	0.5	0.130
8	0.5	0.085

Example of Fitting a Gaussian



The peak value of F is 0.5, for $7 \le x \le 8$. In that range, $G(x) \le 0.13$.

X	F(x)	G(x)
1	0.25	0.031
2	0.25	0.064
3	0.25	0.107
4	0	0.149
5	0	0.172
6	0	0.164
7	0.5	0.130
8	0.5	0.085

Multidimensional Gaussians

- So far we have discussed Gaussians for the case where our training examples x_1, x_2, …, x_n are real numbers.
- What if each x_j is a vector?
  - Let D be the dimensionality of the vector.
  - Then, we can write x_j as ($x_{j,1}$, $x_{j,2}$, …, $x_{j,D}$), where each $x_{j,d}$ is a real number.
- We can define Gaussians for vector spaces as well.
- To fit a Gaussian to vectors, we must compute two things:
  - The mean (which is also a D-dimensional vector).
  - The **covariance matrix** (which is a DxD matrix).

Multidimensional Gaussians - Mean

- Let x_1, x_2, ..., x_n be D-dimensional vectors.
- $x_j = (x_{j,1}, x_{j,2}, \ldots, x_{j,D})$, where each $x_{j,d}$ is a real number.
- Then, the mean $\mu = (\mu_1, \ldots, \mu_D)$ is computed as:

$$\boxed{\mu = \frac{1}{n} \sum_{1}^{n} x_j}$$

- Therefore, $\mu_d = \frac{1}{n} \sum_{1}^{n} x_{j,d}$

Multidimensional Gaussians – Covariance Matrix

- Let $x_1, x_2, \ldots, x_n$ be D-dimensional vectors.
- $x_j = (x_{j,1}, x_{j,2}, \ldots, x_{j,D})$, where each $x_{j,d}$ is a real number.
- Let Σ be the covariance matrix. Its size is DxD.
- Let $\sigma_{r,c}$ be the value of Σ at row r, column c.

$$\boxed{\sigma_{r,c} = \frac{1}{n-1} \sum_{j=1}^{n} (x_{j,r} - \mu_r)(x_{j,c} - \mu_c)}$$

Multidimensional Gaussians – Evaluation

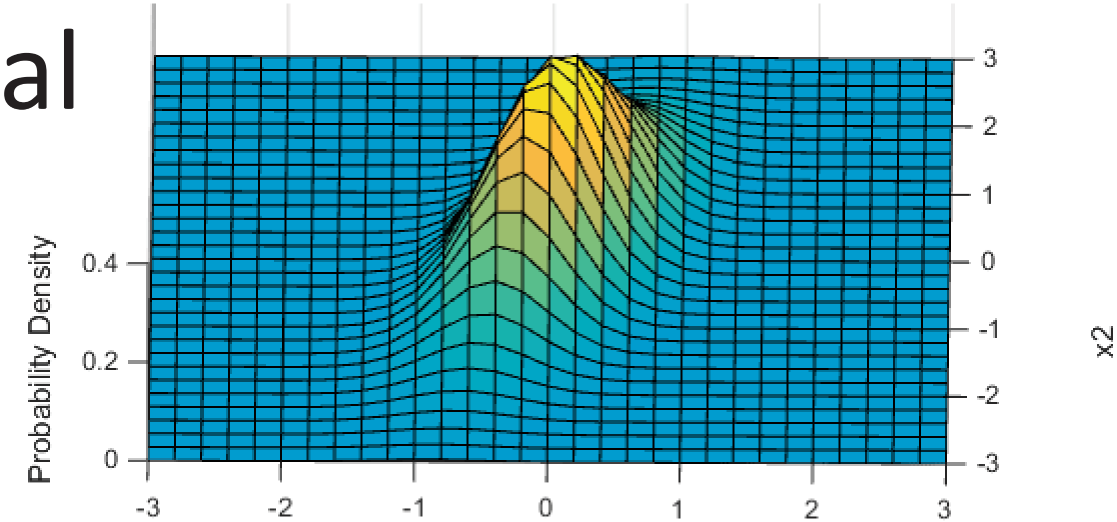
- Let x = (x_1, x_2, …, x_D) be a D-dimensional vector.
- Let N be a D-dimensional Gaussian with mean μ and covariance matrix Σ.
- Let $\sigma_{r,c}$ be the value of Σ at row r, column c.
- Then, the density N(x) of the Gaussian at point x is:

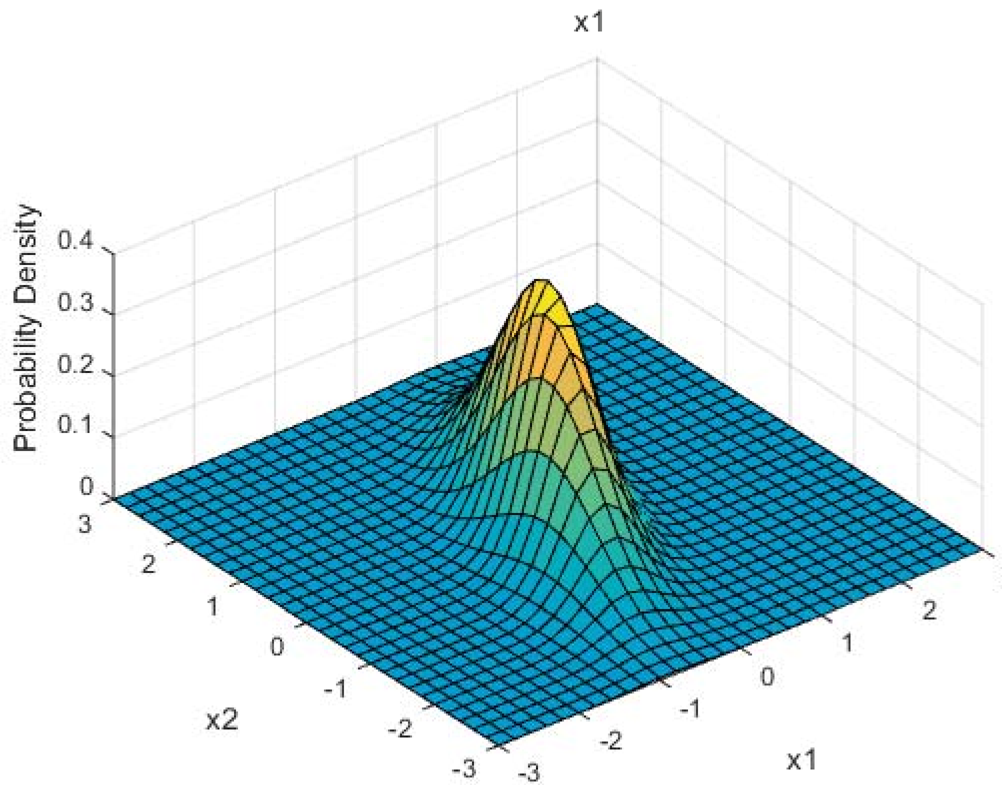
$$\boxed{N(x) = \frac{1}{\sqrt{(2\pi)^D |\Sigma|}} \exp\left(-\frac{1}{2}\,(x - \mu)^{\mathrm{T}} \Sigma^{-1} (x - \mu)\right)}$$

- |Σ| is the determinant of Σ.
- Σ^{-1} is the matrix inverse of Σ.
- $(x - \mu)^{\mathrm{T}}$ is a 1xD row vector, $(x - \mu)$ is a Dx1 column vector.

A 2-Dimensional Example

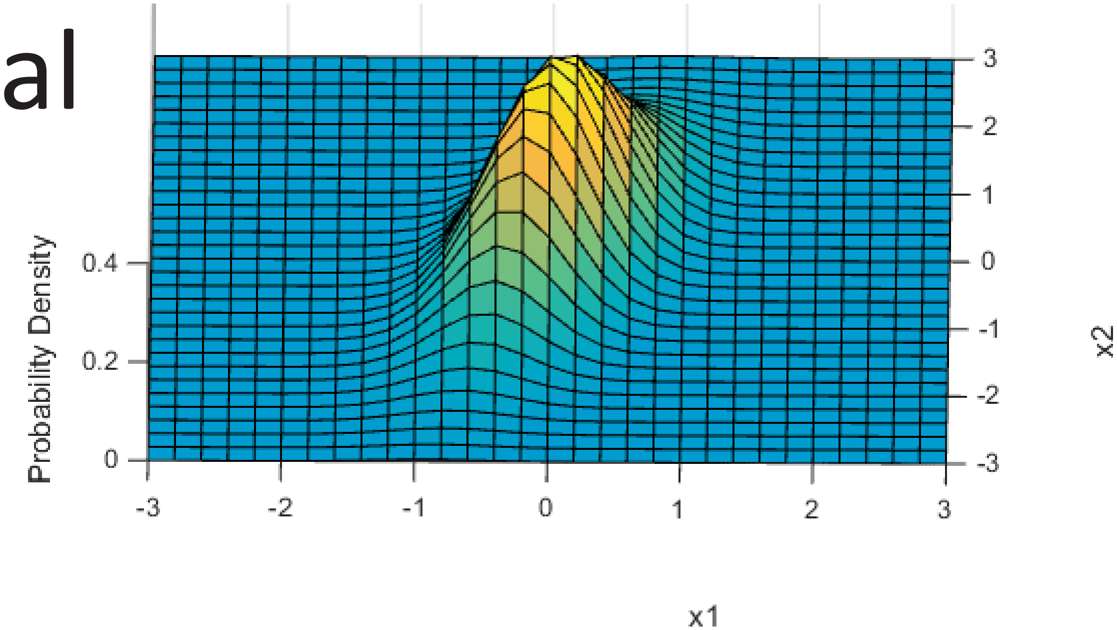
- Here you see (from different points of view) a visualization of a two dimensional Gaussian.
  - Axes: x_1, x_2, value.
- Its peak value is on the mean, which is (0,0).
- It has a ridge directed (in the top figure) from the bottom left to the top right.

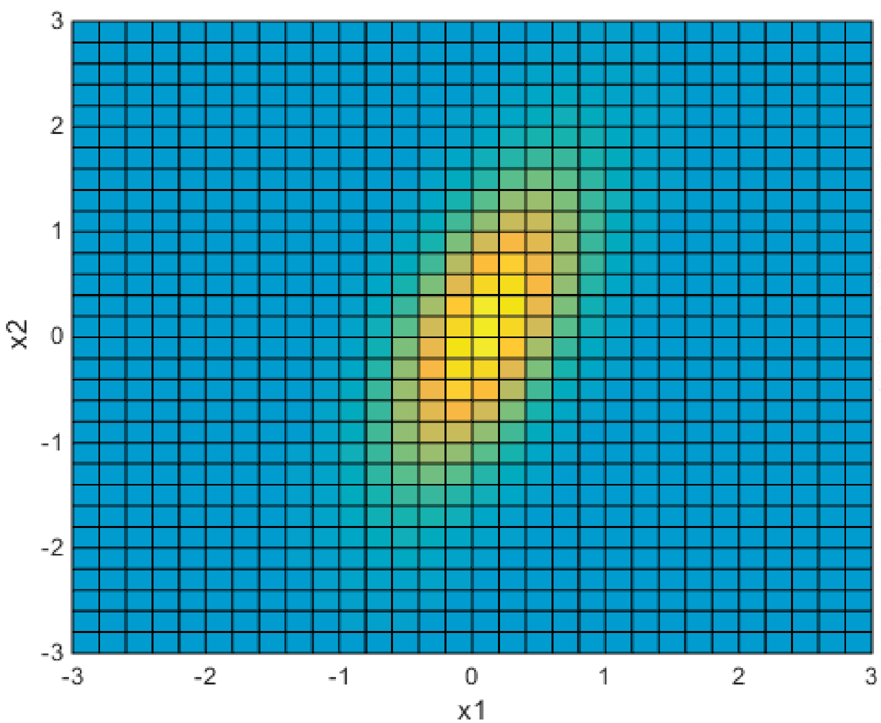


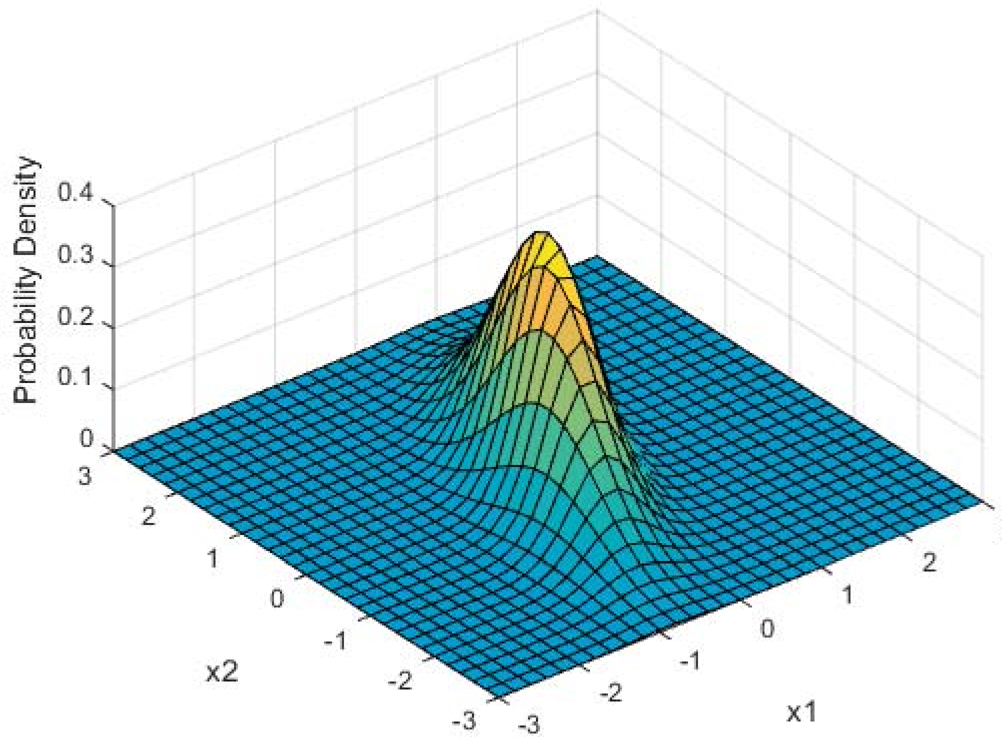


A 2-Dimensional Example

- The view from the top shows that, for any value A, the set of points (x, y) such that N(x, y) = A form an ellipse.
  - Each value corresponds to a color.







Multidimensional Gaussians – Training

- Let N be a D-dimensional Gaussian with mean μ and covariance matrix Σ.
- How many parameters do we need to specify N?
  - The mean μ is defined by D numbers.
  - The covariance matrix Σ requires D^2 numbers $\sigma_{r,c}$.
  - Strictly speaking, Σ is symmetric, $\sigma_{r,c} = \sigma_{c,r}$.
  - So, we need roughly $D^2/2$ parameters.
- The number of parameters is quadratic to D.
- The number of training data we need for reliable estimation is also quadratic to D.

Gaussians: Recap

- 1-dimensional Gaussians are easy to estimate from relatively few examples.
  – They are specified using only two parameters, μ and σ.
- D-dimensional Gaussians are specified using $O(D^2)$ parameters.
- Gaussians take a specific shape, which may not fit well the actual distribution of the data.