Machine Learning 2

DS 4420 / ML 2

Math review

Byron C Wallace

Probability

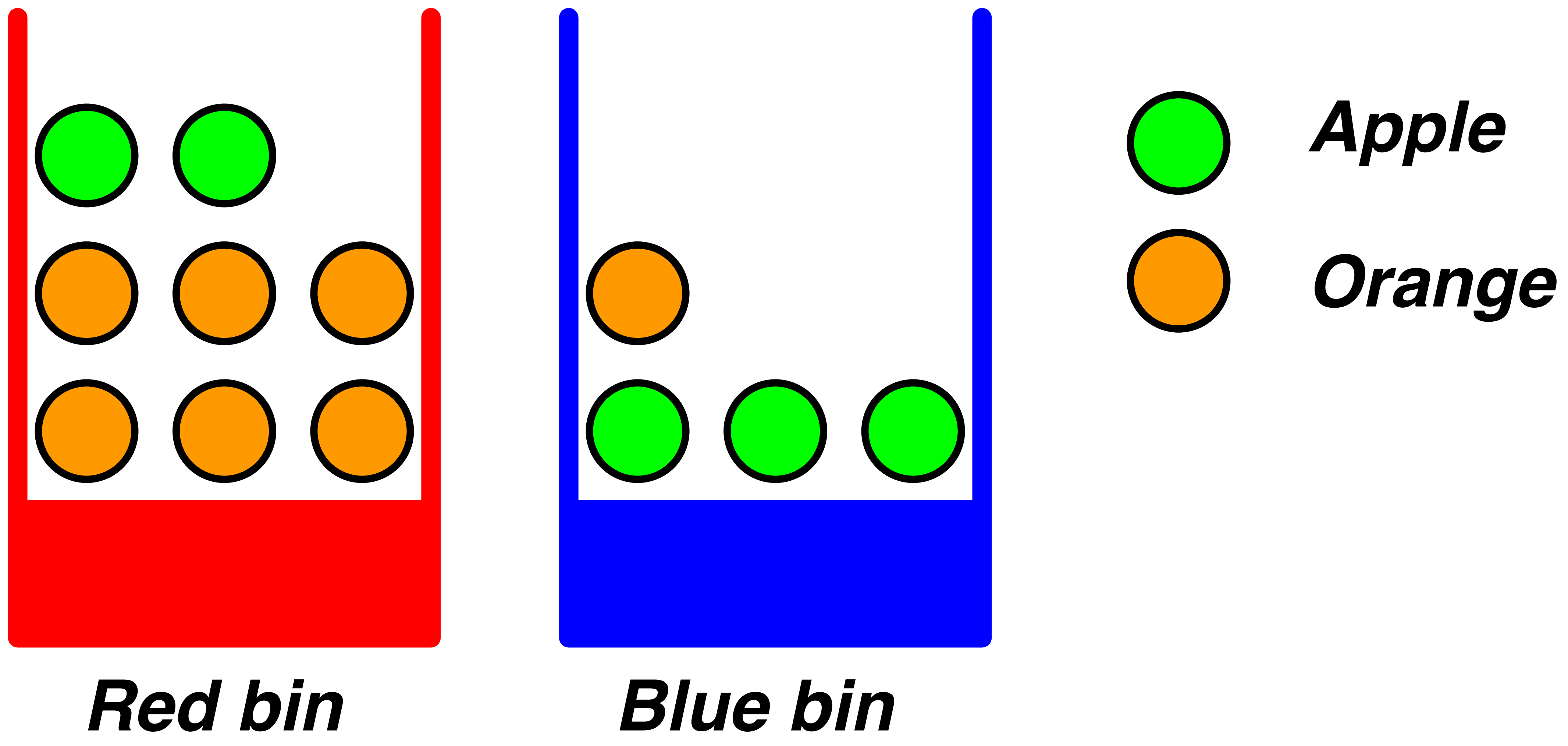
Examples: Independent Events

What's the probability of getting a sequence of 1,2,3,4,5,6 if we roll a dice six times?

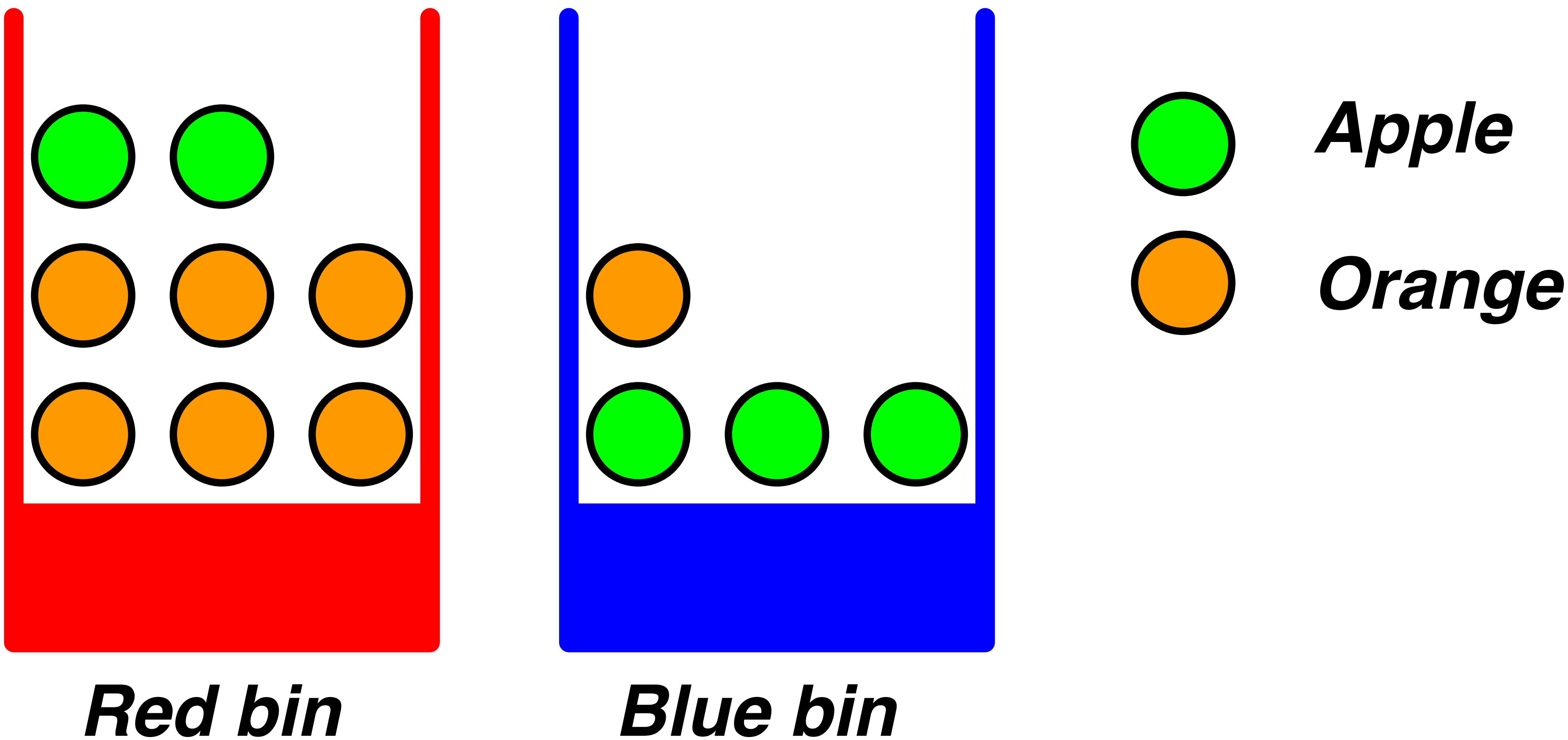
Examples: Independent Events

A school survey found that 9 out of 10 students like pizza. If three students are chosen at random with replacement, what is the probability that all three students like pizza?

Urns!



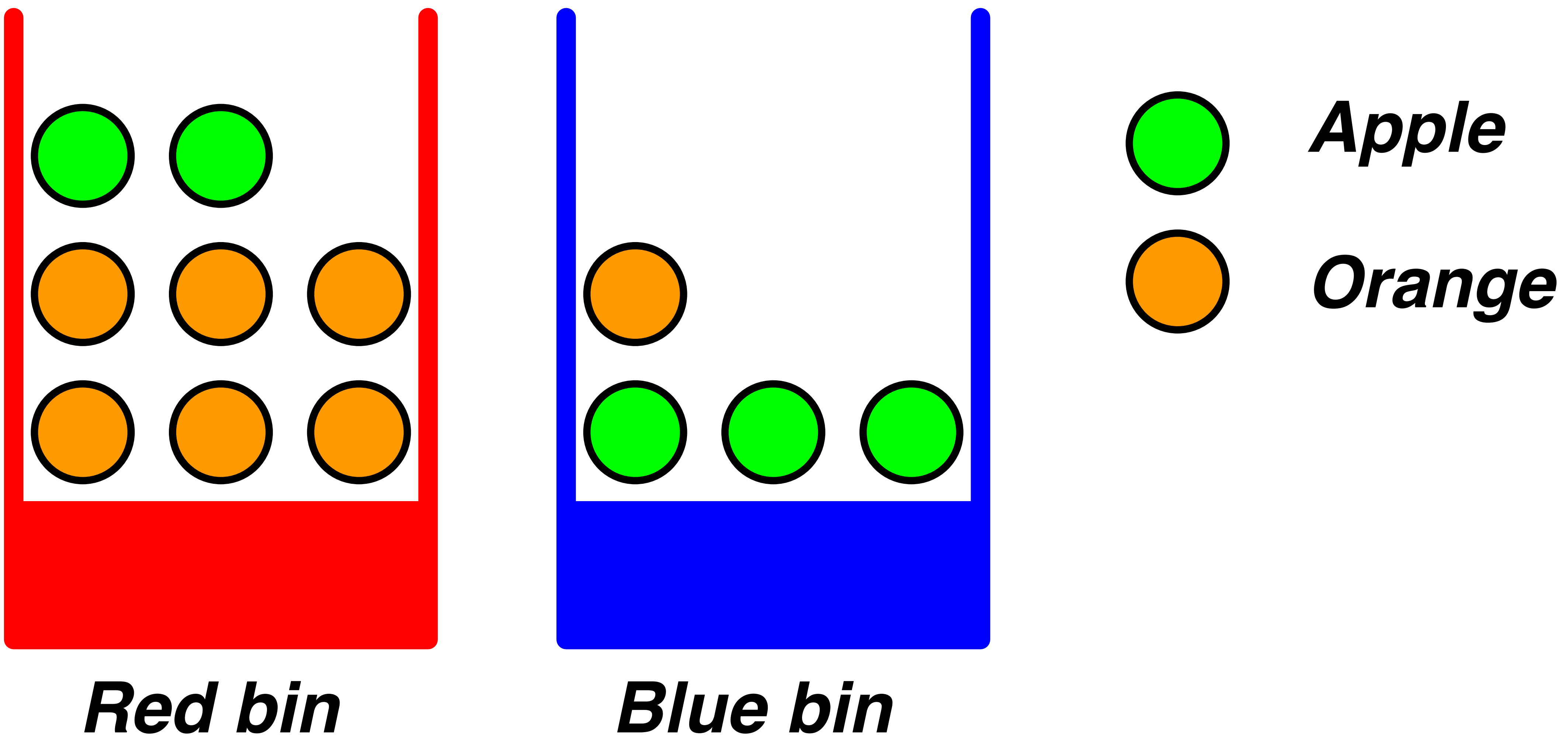
Dependent Events



*If I randomly pick a fruit from the **red** bin,*

*what is the probability that I get an **apple**?*

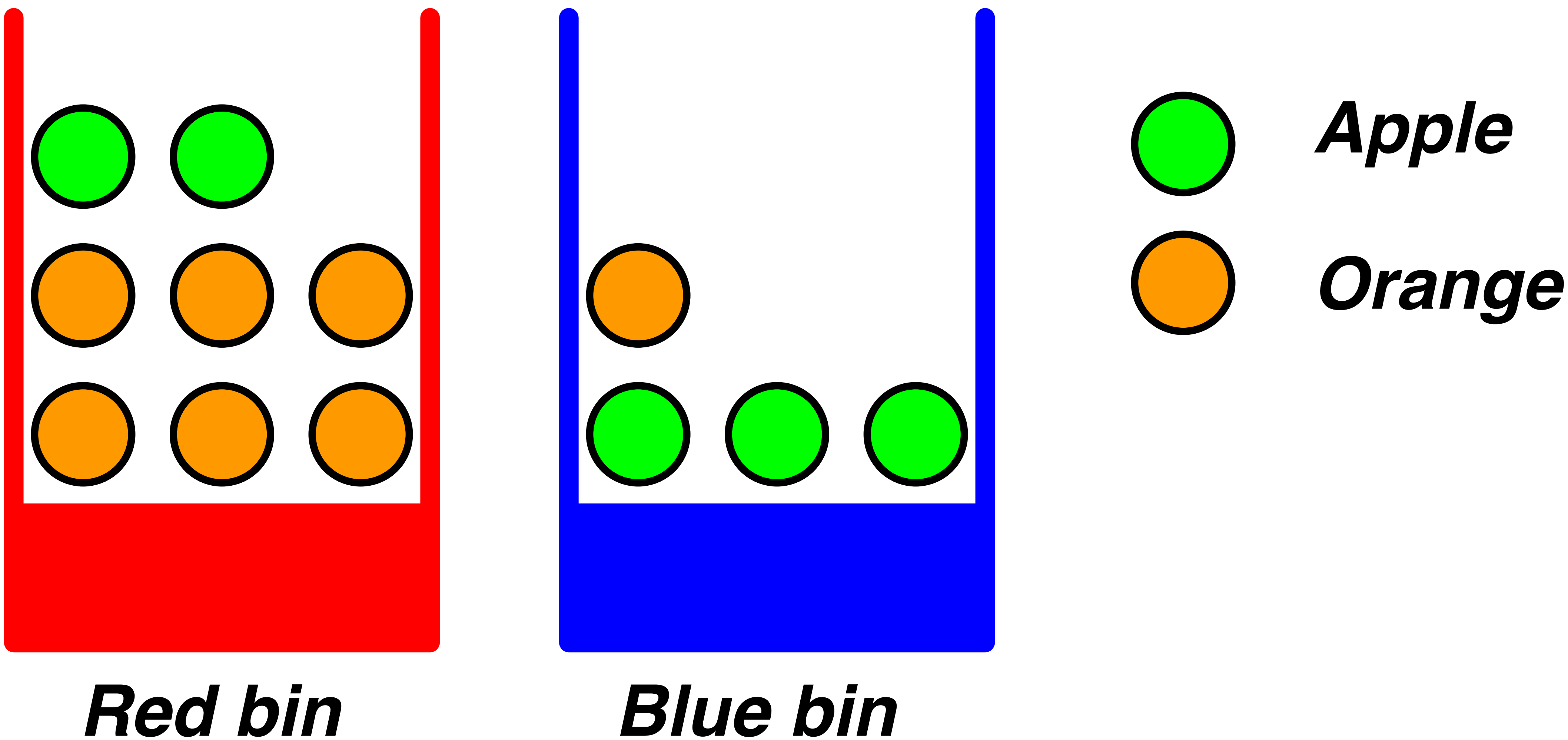
Dependent Events



Conditional Probability

P(fruit = apple | bin = red) = 2 / 8

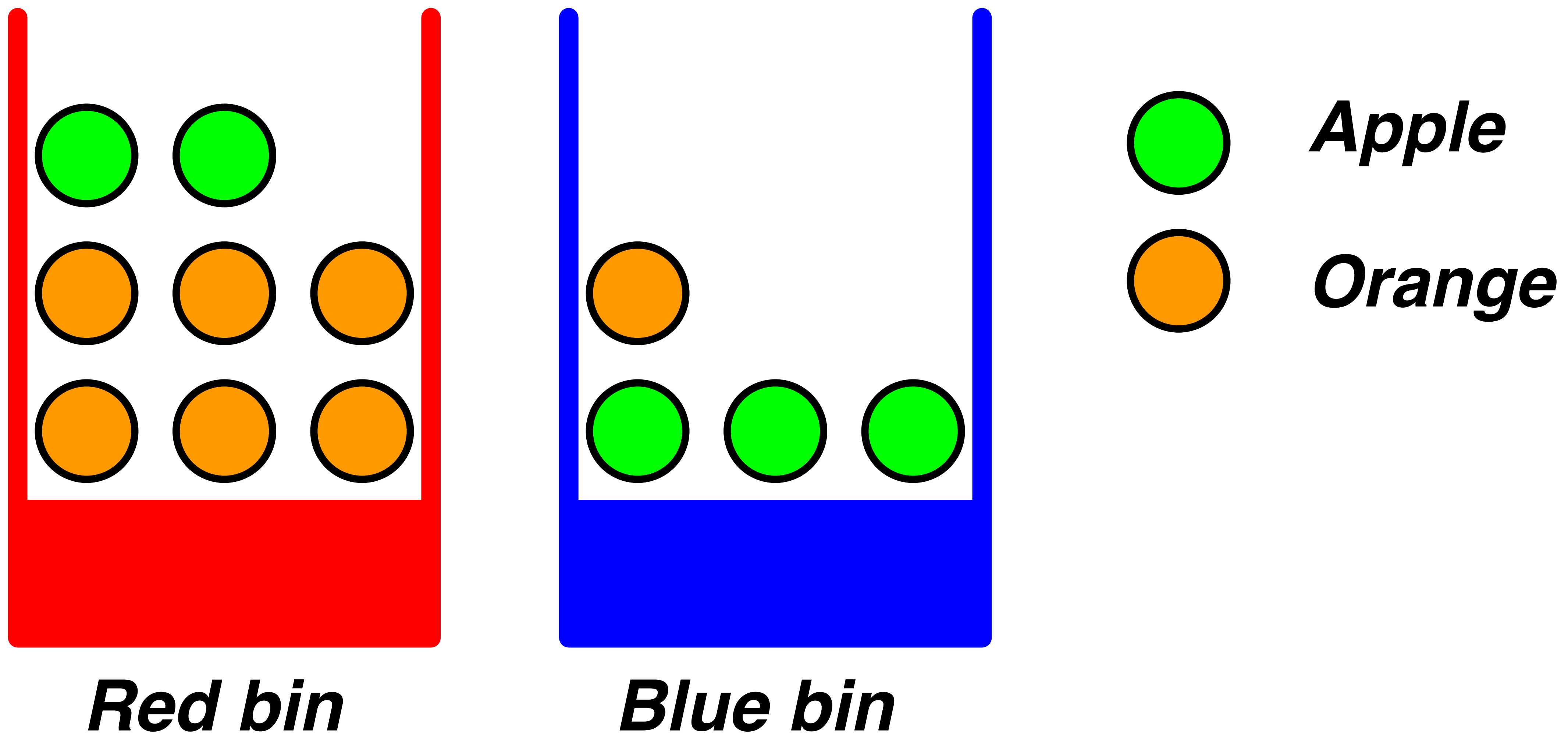
Dependent Events



Joint Probability

P(fruit = apple , bin = red) = 2 / 12

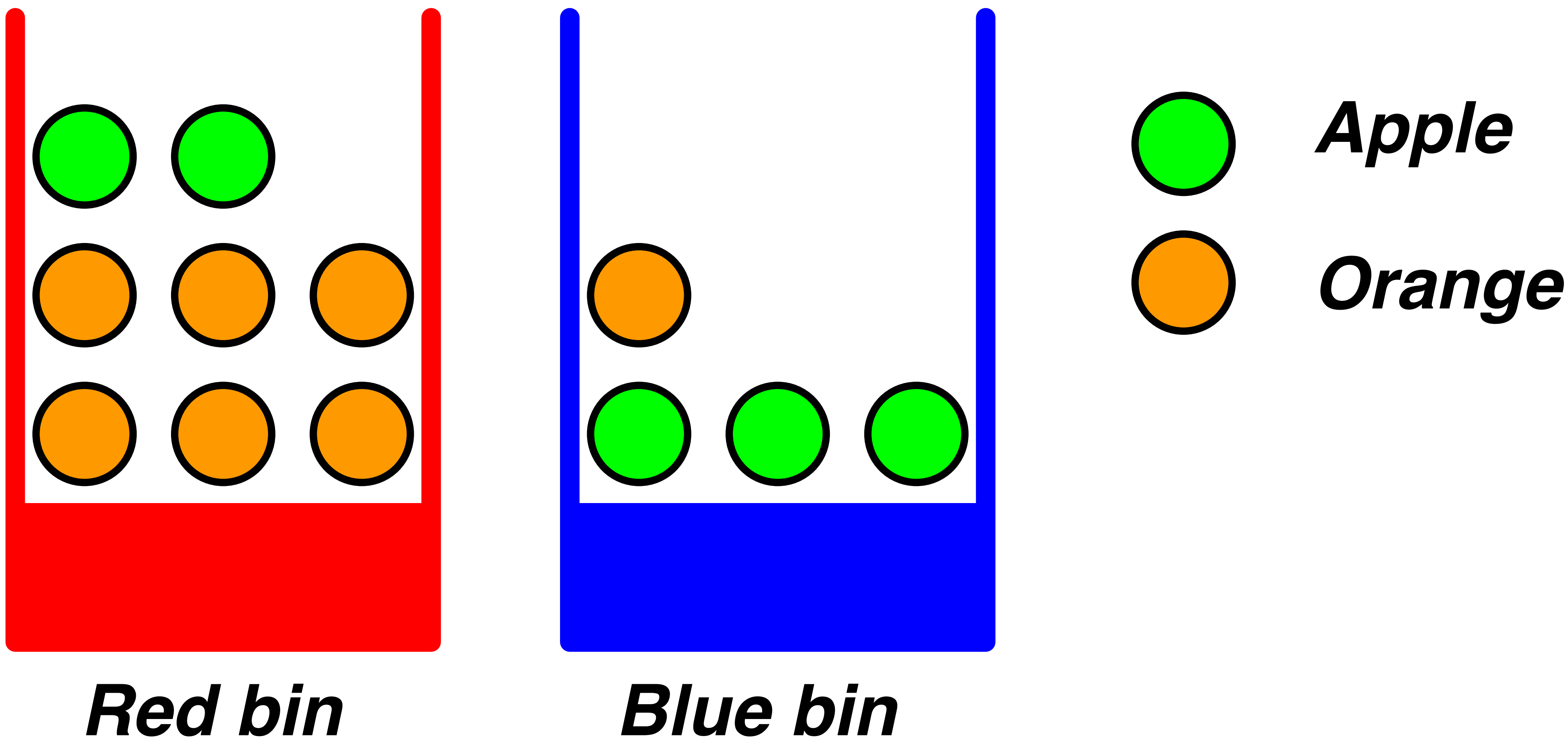
Dependent Events



Joint Probability

P(fruit = apple , bin = blue) = ?

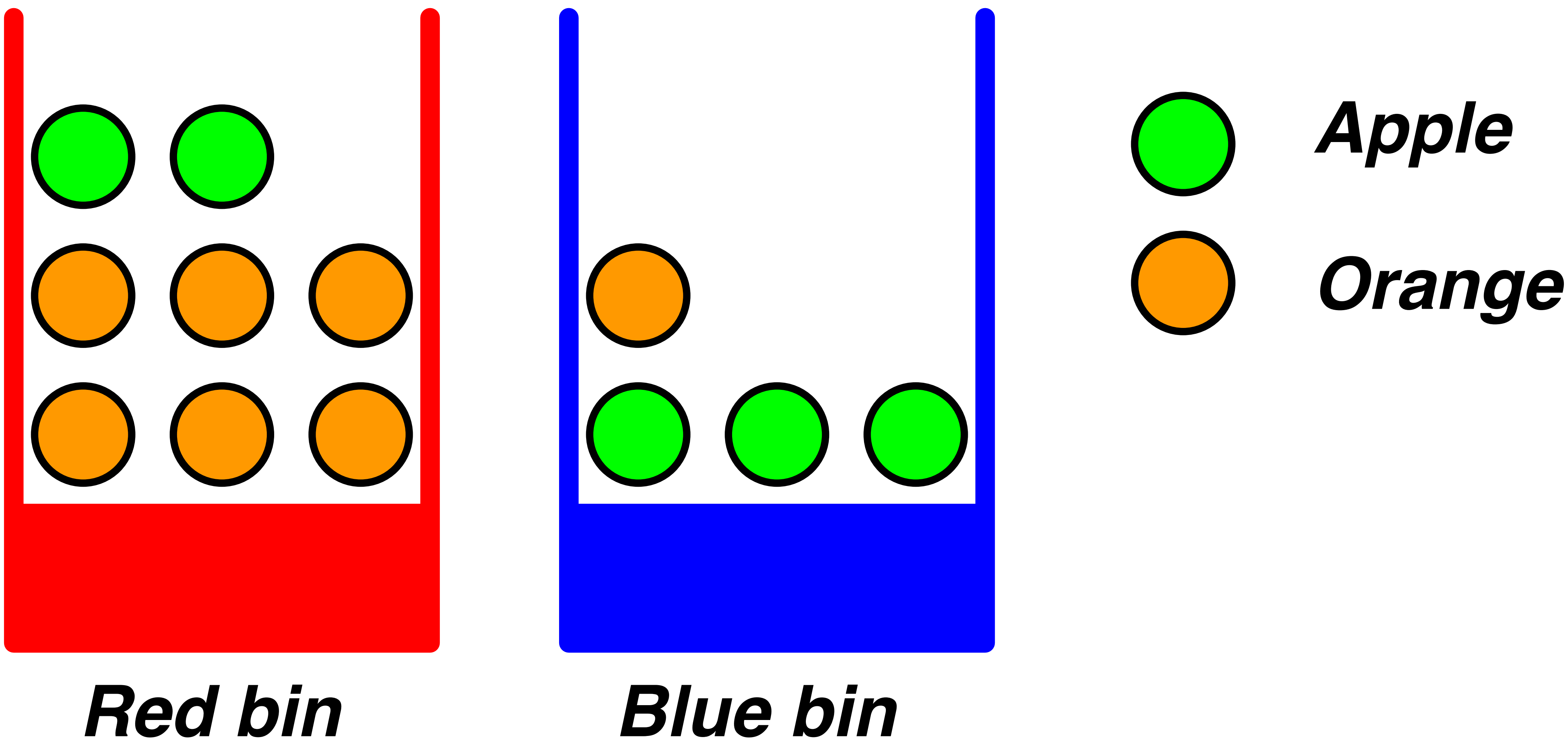
Dependent Events



Joint Probability

P(fruit = apple , bin = blue) = 3 / 12

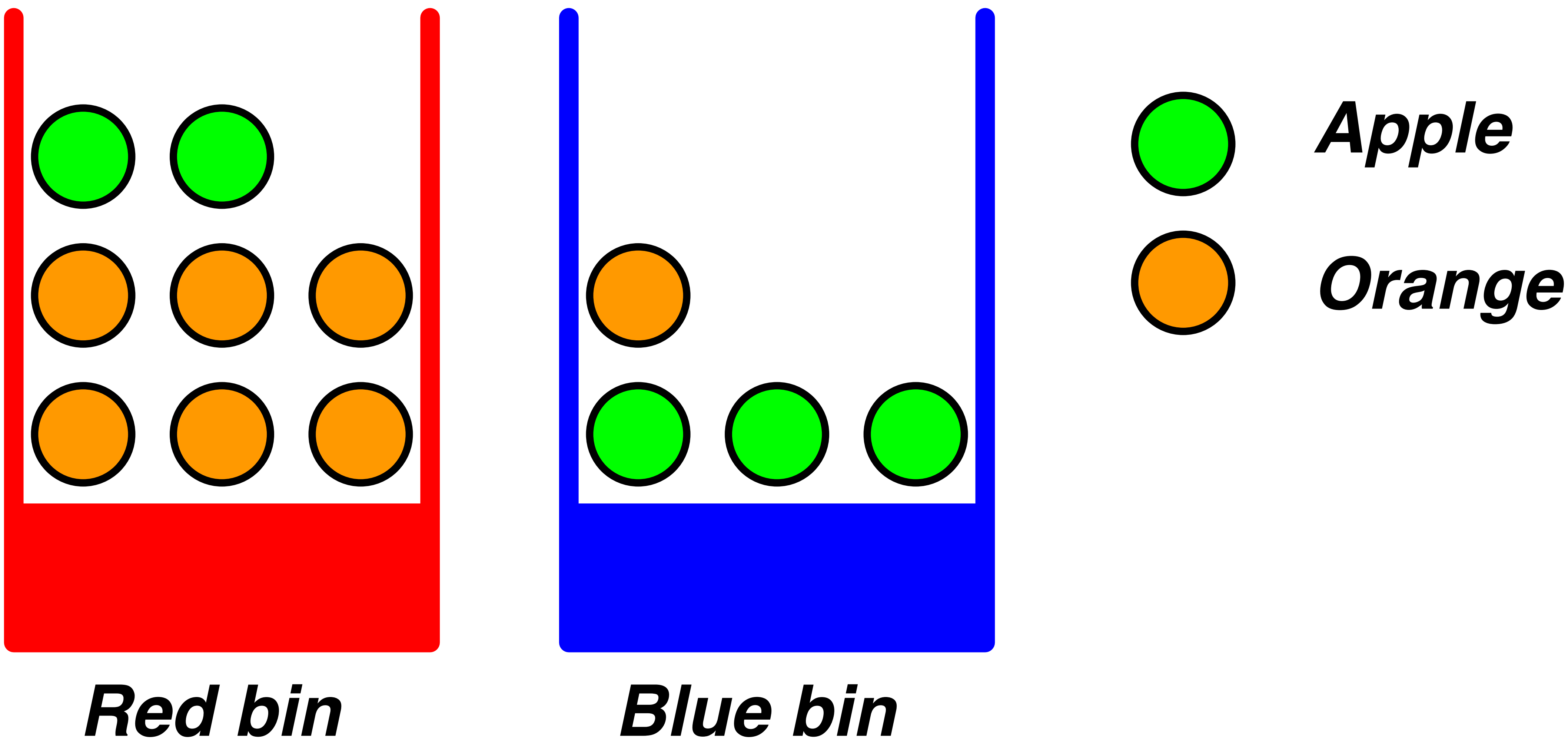
Dependent Events



Joint Probability

P(fruit = orange , bin = blue) = ?

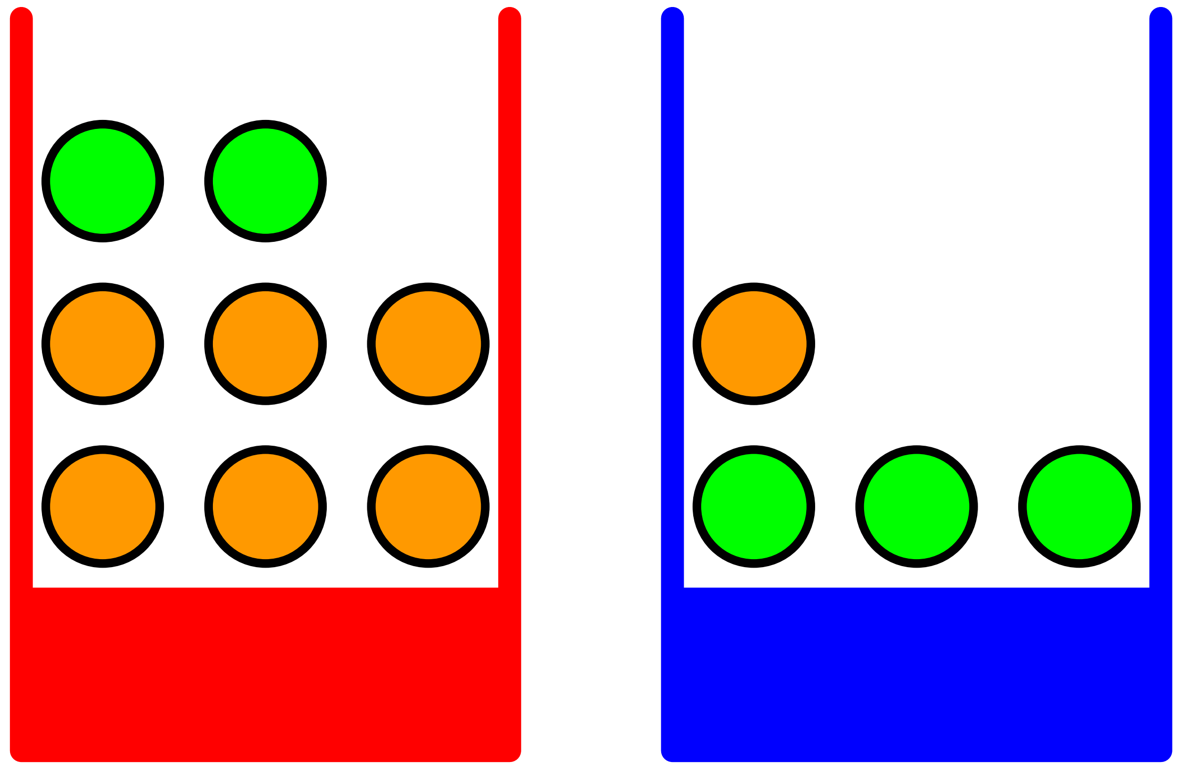
Dependent Events



Joint Probability

P(fruit = orange , bin = blue) = 1 / 12

Two rules of Probability



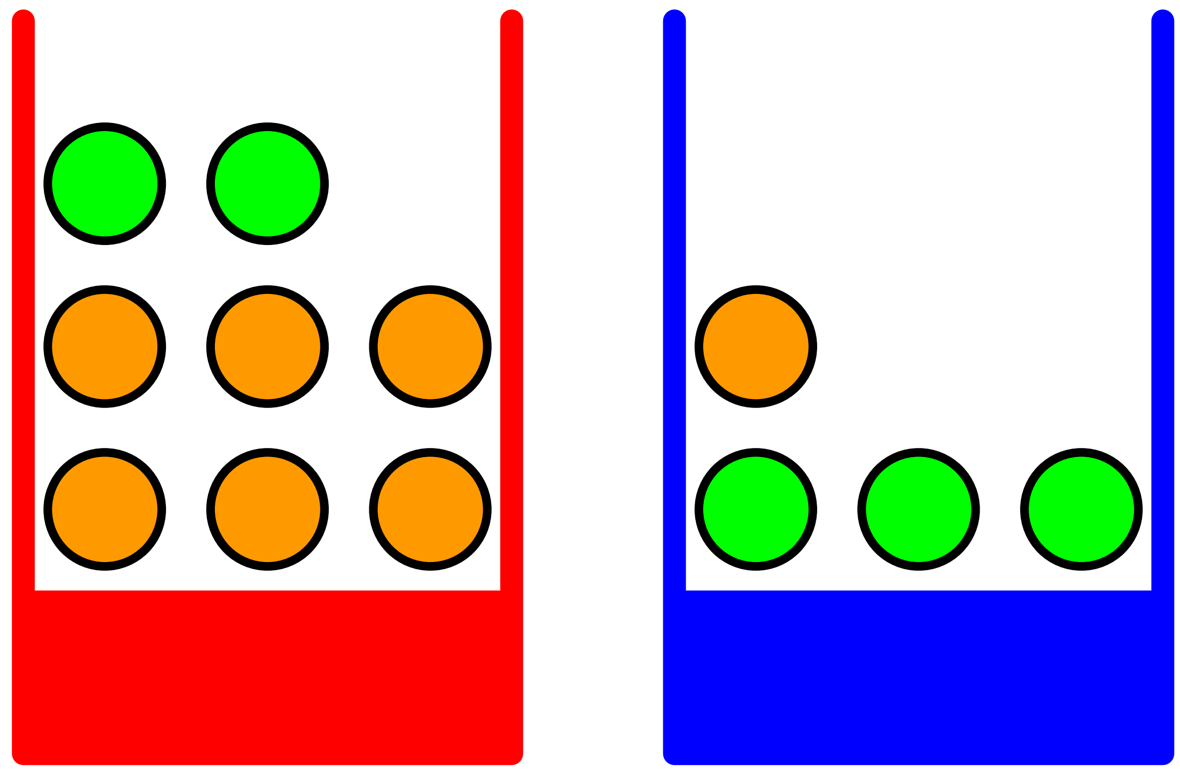
1. Sum Rule (Marginal Probabilities)

P(fruit = apple) = P(fruit = apple , bin = blue)

+ P(fruit = apple , bin = red)

= ?

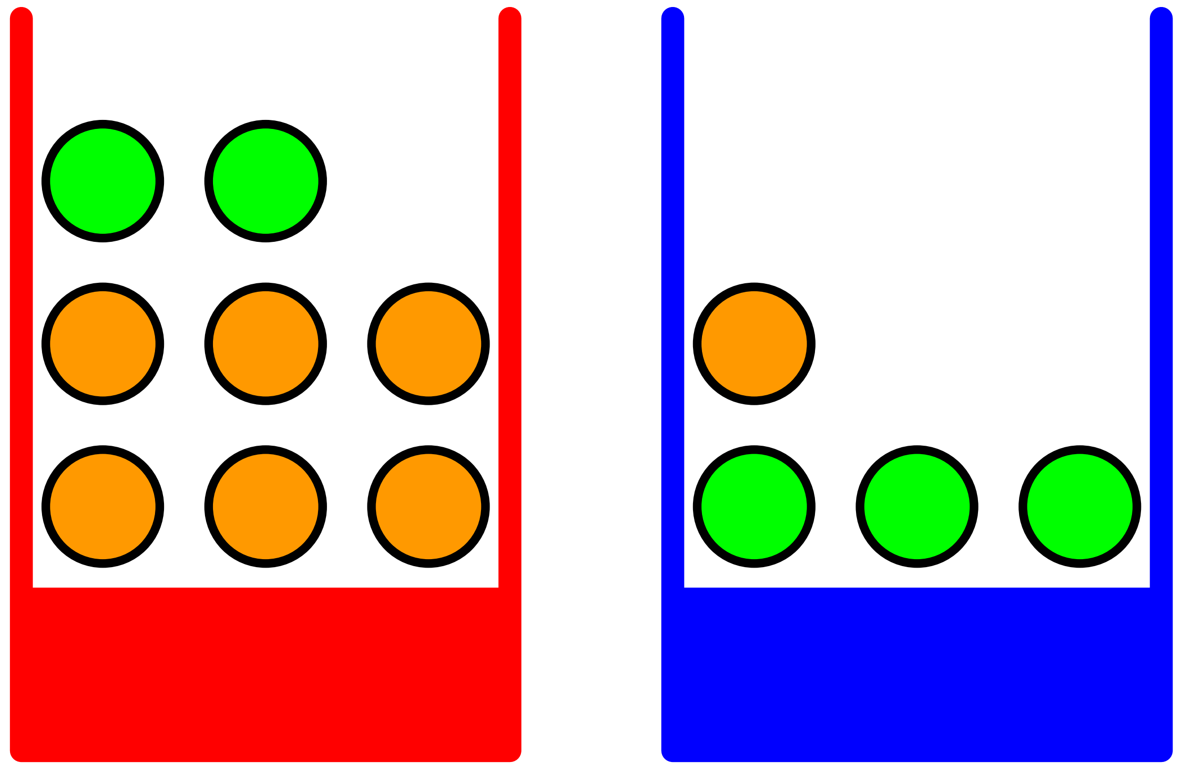
Two rules of Probability



1. Sum Rule (Marginal Probabilities)

P(fruit = apple) = ?

Two rules of Probability



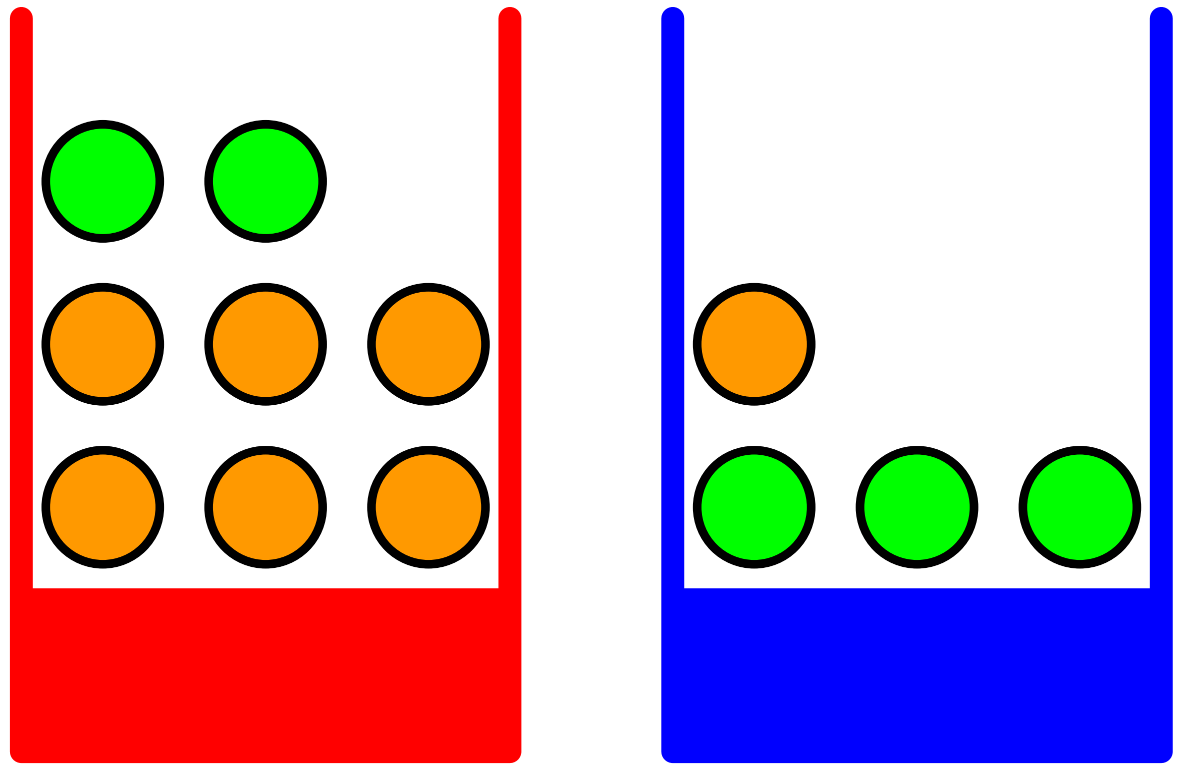
1. Sum Rule (Marginal Probabilities)

P(fruit = apple) = P(fruit = apple , bin = blue)

+ P(fruit = apple , bin = red)

= 3 / 12 + 2 / 12 = **5 / 12**

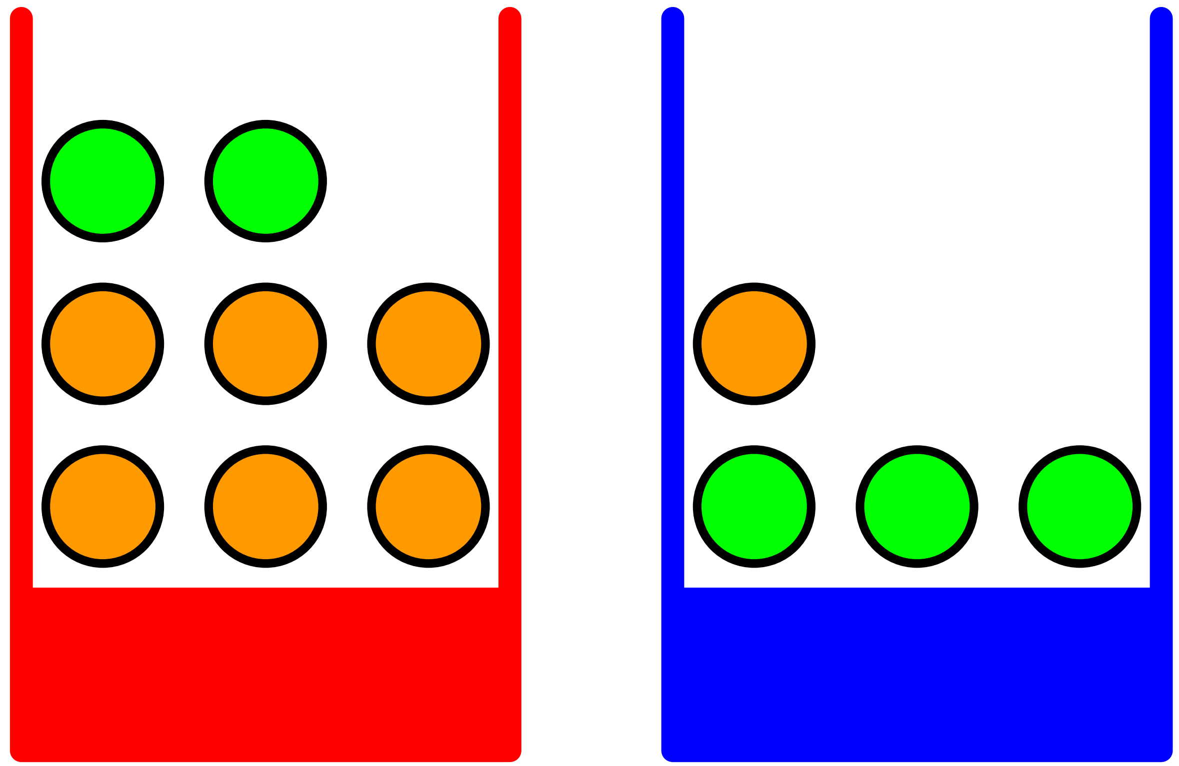
Two rules of Probability



2. Product Rule

P(fruit = apple , bin = red) = ?

Two rules of Probability



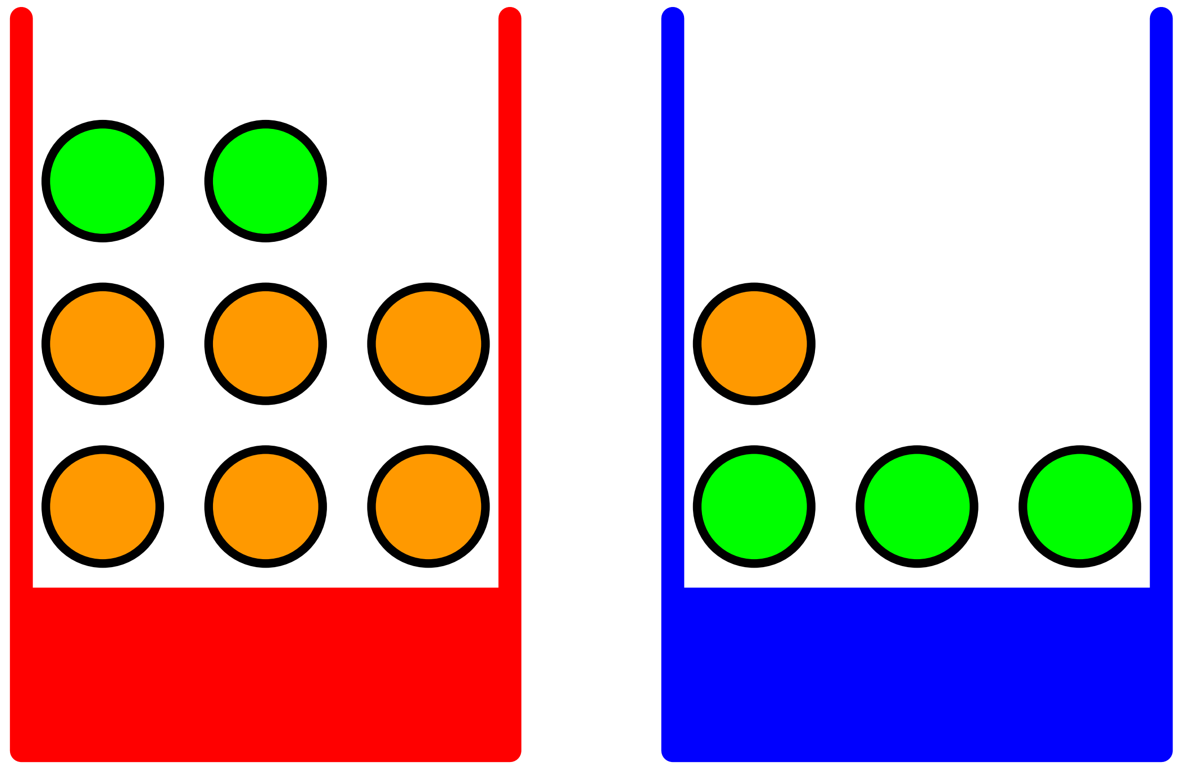
2. Product Rule

P(fruit = apple , bin = red) =

P(fruit = apple | bin = red) p(bin = red)

= ?

Two rules of Probability



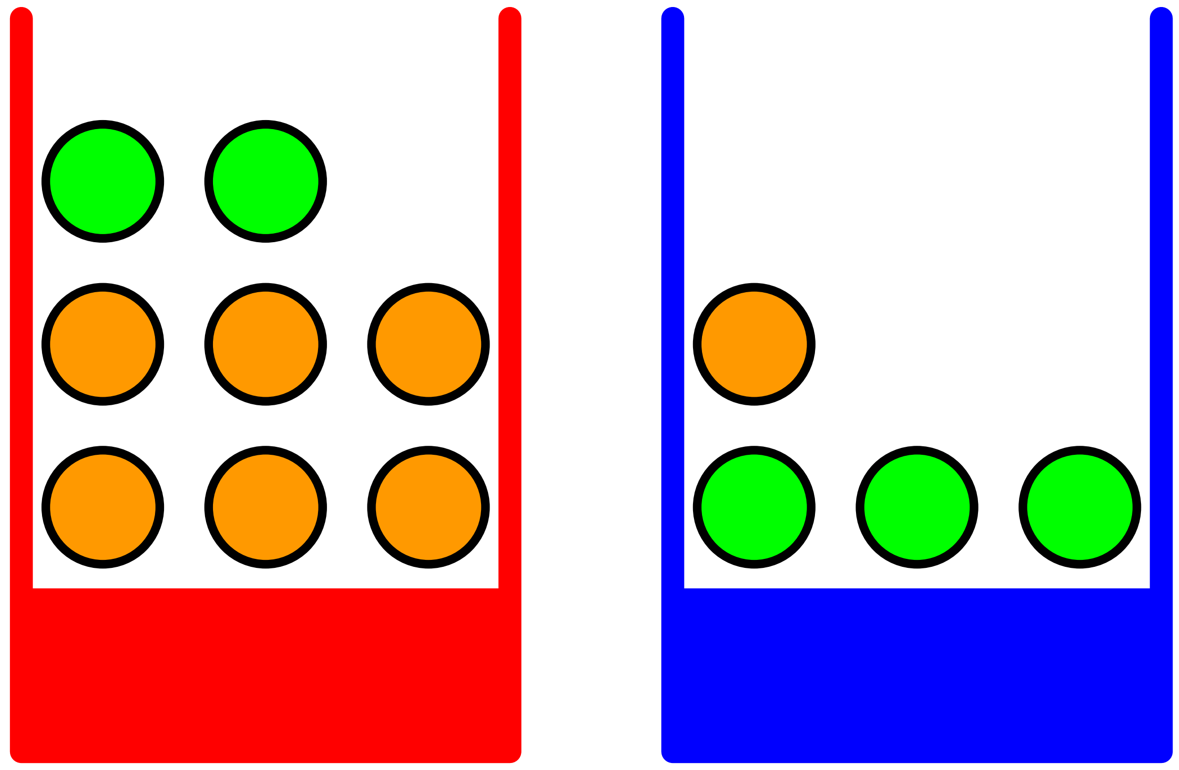
2. Product Rule

P(fruit = apple , bin = red) =

P(fruit = apple | bin = red) p(bin = red)

= 2 / 8 * 8 / 12 = **2 / 12**

Two rules of Probability



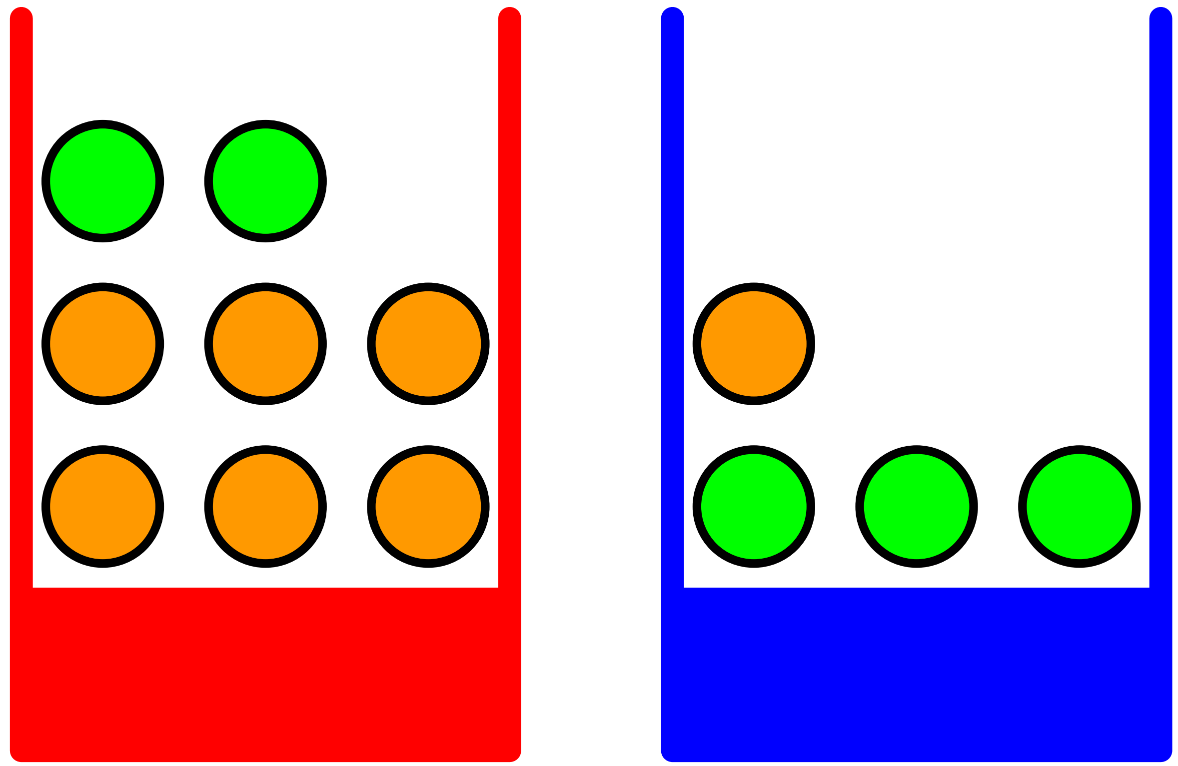
2. Product Rule (reversed)

P(fruit = apple , bin = red) =

P(bin = red | fruit = apple) p(fruit = apple)

= ?

Two rules of Probability



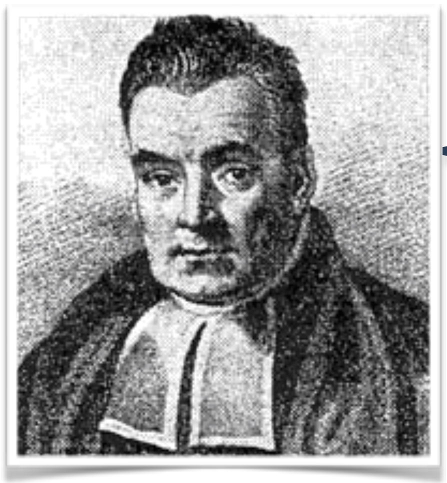
2. Product Rule (reversed)

P(fruit = apple , bin = red) =

P(bin = red | fruit = apple) p(fruit = apple)

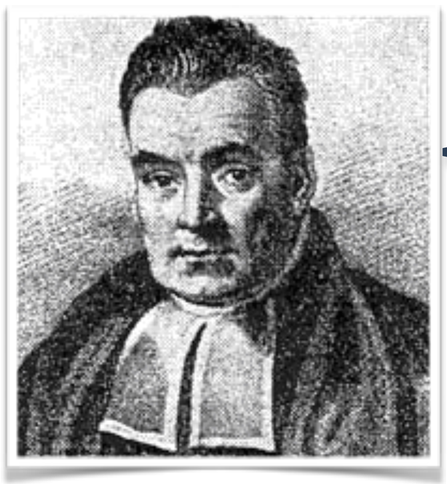
= 2 / 5 * 5 / 12 = **2 / 12**

Bayes' Rule



$$\underbrace{p(x \mid y)}_{\text{Posterior}} = \underbrace{p(y \mid x)}_{\text{Likelihood}}\underbrace{p(x)}_{\text{Prior}}/p(y)$$

Bayes' Rule



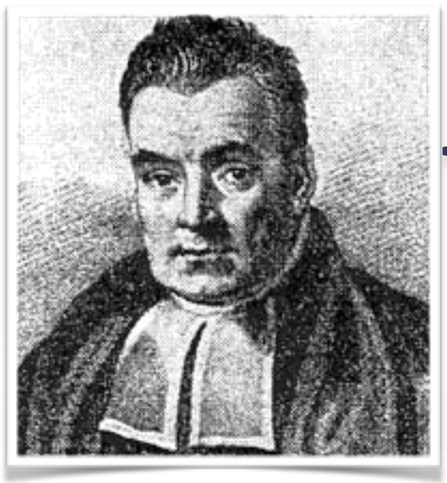
$$p(x \mid y) = p(y \mid x)p(x)/p(y)$$

Posterior Likelihood Prior

Sum Rule: $p(y) = \sum_{x} p(y, x) \quad p(x) = \sum_{y} p(y, x)$

Product Rule: $p(y, x) = p(y \mid x)p(x) = p(x \mid y)p(y)$

Bayes' Rule



$$\underbrace{p(x \mid y)}_{\text{Posterior}} = \underbrace{p(y \mid x)}_{\text{Likelihood}}\,\underbrace{p(x)}_{\text{Prior}}/p(y)$$

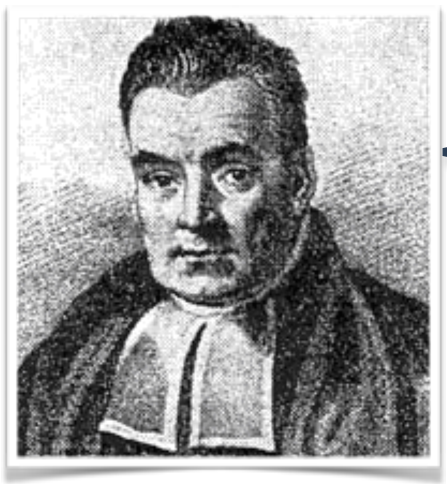
$p(x)$ *Probability of rare disease:* 0.005

$p(y \mid x)$ *Probability of detection:* 0.98

Probability of false positive: 0.05

$p(x \mid y)$ *Probability of disease when test positive?*

Bayes' Rule



$$p(x \mid y) = p(y \mid x)p(x)/p(y)$$

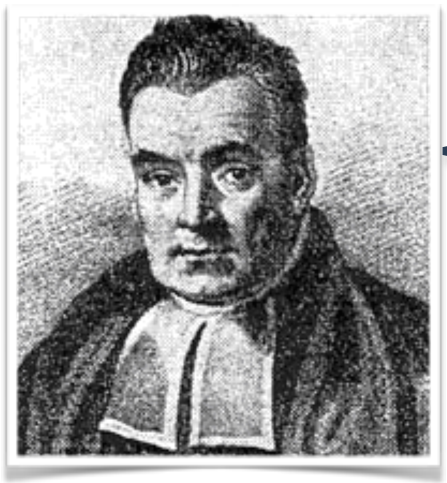
Posterior   Likelihood   Prior

$$p(y,x) = p(y \mid x)p(x)$$

$$p(y) = \sum_x p(y,x)$$

$$p(x \mid y)$$

Bayes' Rule



$$p(x \mid y) = p(y \mid x)p(x)/p(y)$$

Posterior Likelihood Prior

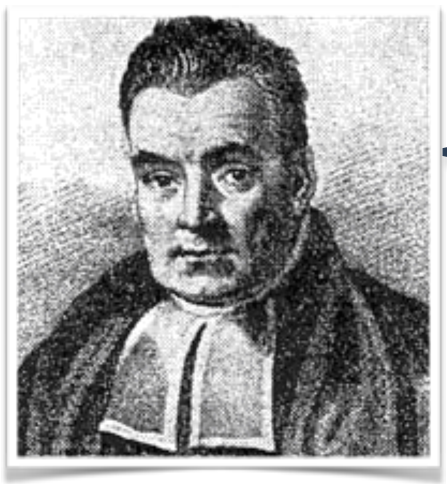
$$p(y, x) = p(y \mid x)p(x)$$

*0.98 * 0.005 = 0.0049*

$$p(y) = \sum_x p(y, x)$$

$$p(x \mid y)$$

Bayes' Rule



$$p(x \mid y) = p(y \mid x)p(x)/p(y)$$

Posterior Likelihood Prior

$$p(y, x) = p(y \mid x)p(x)$$

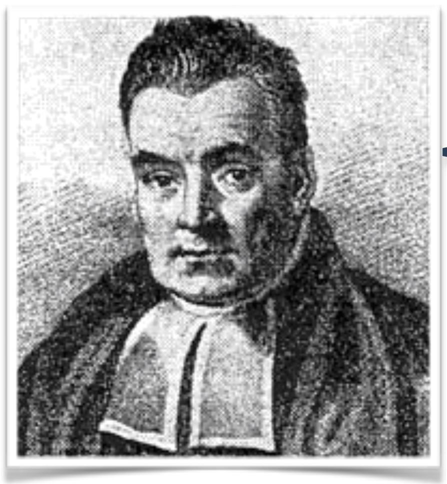
*0.98 * 0.005 = 0.0049*

$$p(y) = \sum_x p(y, x)$$

*0.98 * 0.005 + 0.05 * 0.995 = 0.0547*

$$p(x \mid y)$$

Bayes' Rule



$$p(x \mid y) = p(y \mid x)p(x)/p(y)$$

Posterior Likelihood Prior

$$p(y, x) = p(y \mid x)p(x)$$

*0.98 * 0.005 = 0.0049*

$$p(y) = \sum_x p(y, x)$$

*0.98 * 0.005 + 0.05 * 0.995 = 0.0547*

$$p(x \mid y)$$

0.0049 / 0.0547 = 0.089

Random Variables

- ***Random Variable:*** A variable with a stochastic ***outcome***

$$X = x \qquad x \in \{1, 2, 3, 4, 5, 6\}$$

Random Variables

- ***Random Variable:*** A variable with a stochastic *outcome*

  $X = x \qquad x \in \{1, 2, 3, 4, 5, 6\}$

- ***Event:*** A set of *outcomes*

  $X >= 3 \qquad \{3, 4, 5, 6\}$

Random Variables

- ***Random Variable:*** A variable with a stochastic *outcome*

  $X = x \qquad x \in \{1, 2, 3, 4, 5, 6\}$

- ***Event:*** A set of *outcomes*

  $X >= 3 \qquad \{3, 4, 5, 6\}$

- ***Probability:*** The chance that a randomly selected *outcome* is part of an *event*

  $P(X >= 3) = 4 / 6$

Distribution

- A ***distribution*** maps ***outcomes*** to ***probabilities***

$$P(X = x) = 1 / 6$$

- Commonly used (or abused) shorthand:

$P(x)$ *is equivalent to* $P(X = x)$

Probability Spaces

Definition: A probability space (Ω, F, P) consists of

- A sample space Ω (i.e. the set of *outcomes*)
- A set of events F (i.e. the set possible sets)
- A probability measure P (maps events to probabilities)

Probability Spaces

Definition: A probability space (Ω, F, P) consists of

- A sample space **Ω** (i.e. the set of *outcomes*)
- A set of events **F** (i.e. the set possible sets)
- A probability measure **P** (maps events to probabilities)

Axioms of Probability

$$P : \mathcal{F} \to \mathbb{R} \qquad P(E) \geq 0 \;\; \forall\, E \in \mathcal{F} \qquad P(\Omega) = 1$$

$$P(E_1, E_2) = P(E_1) + P(E_2) \;\; \text{when} \;\; E_1 \cap E_2 = \emptyset$$

Conditional Probabilities

- **Definition:** Joint Probability

Outcomes in both *A* and *B*

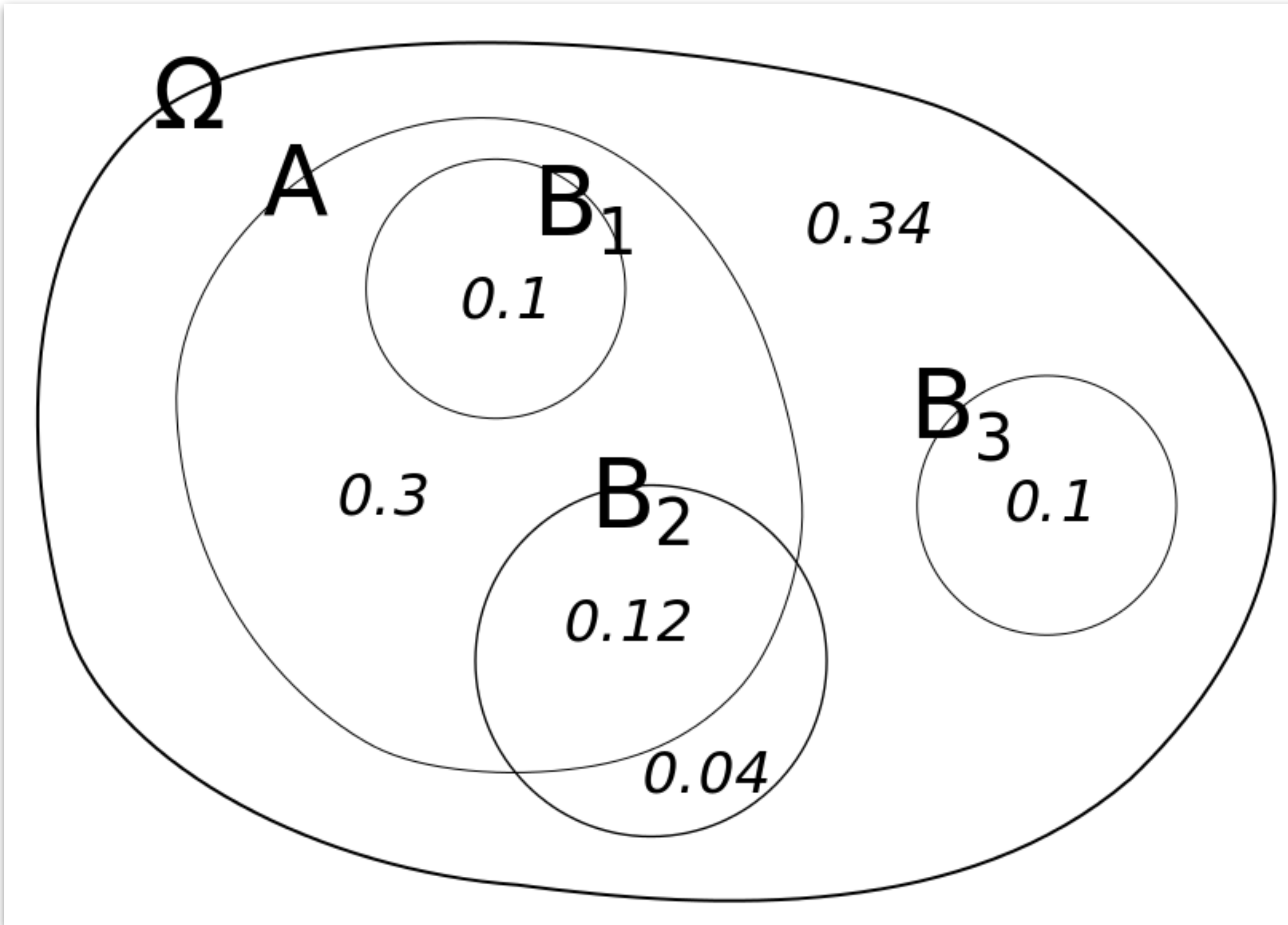
$$P(A, B) = P(A \cap B)$$

Events (i.e. sets of outcomes)

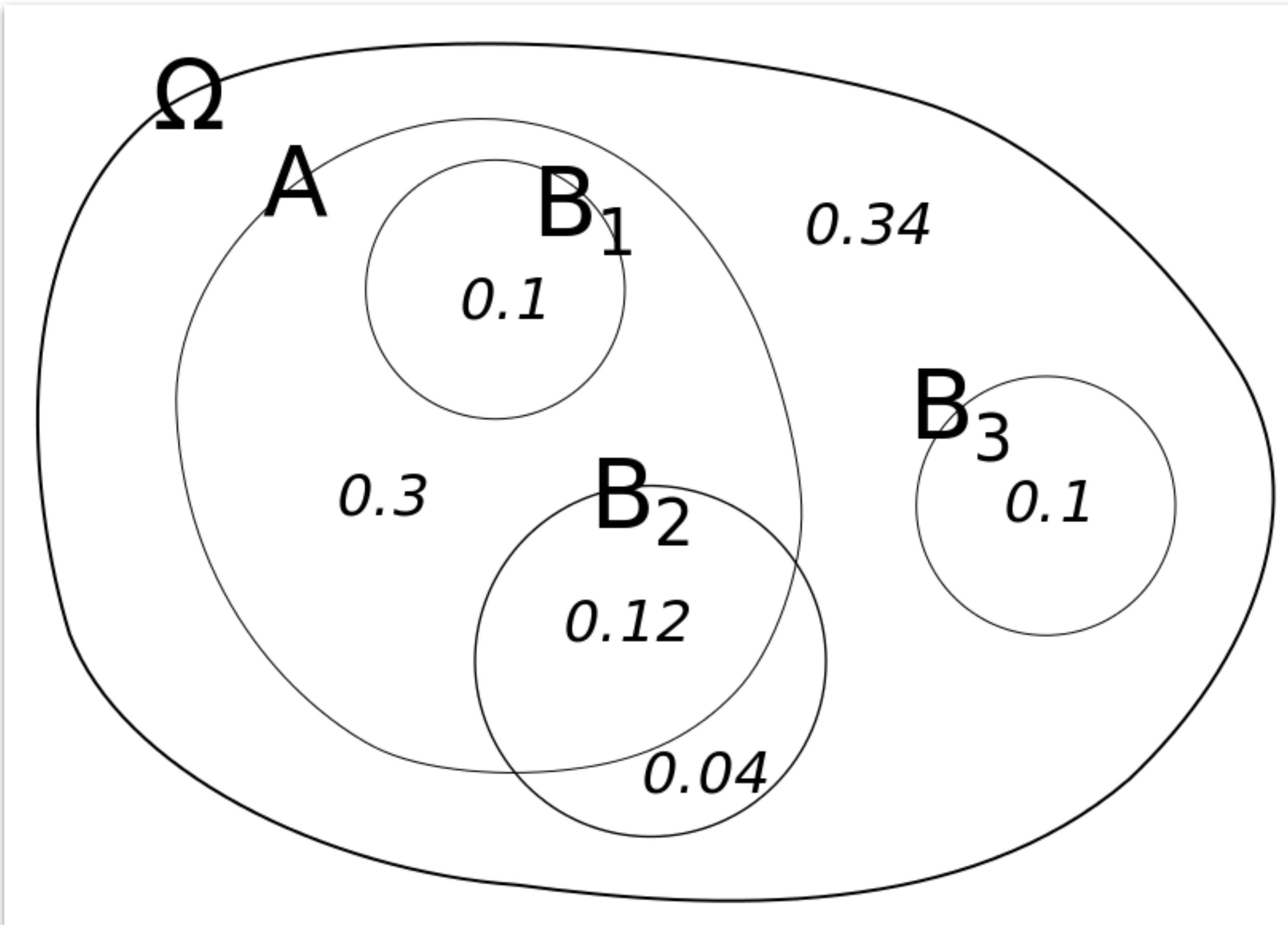
- **Definition:** Conditional Probability

$$P(A \mid B) = \frac{P(A, B)}{P(B)}$$

Conditional Probability

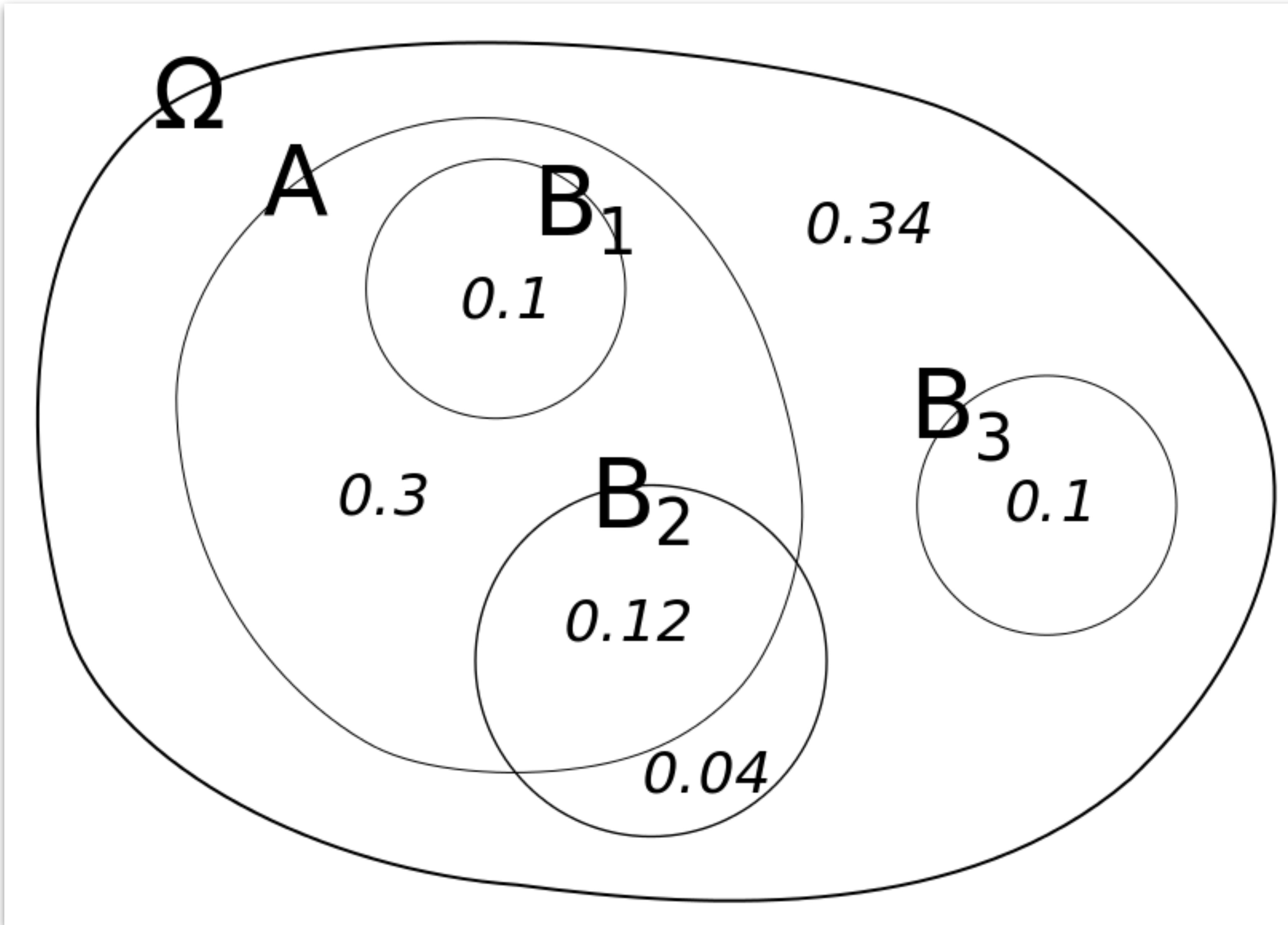


Conditional Probability



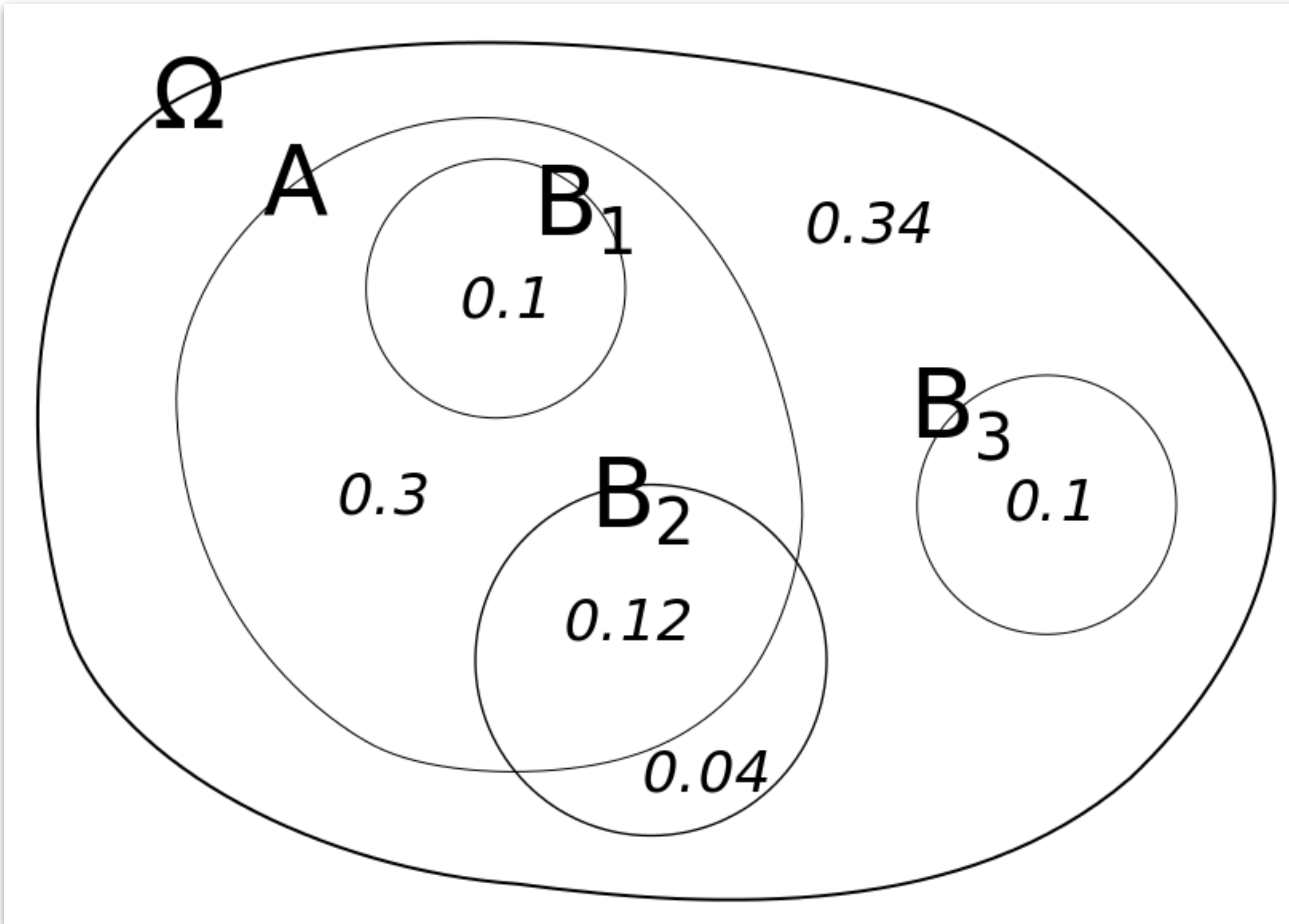
What is the probability $P(B_3)$? 0.1 / 0.34

Conditional Probability



What is the probability $P(B_2 \mid A)$? 0.12 / 0.3

Conditional Probability



What is the probability $P(B_1 \mid B_3)$? 0.0 / 0.1

Examples: Conditional Probability

1. A math teacher gave her class two tests.
    - *25% of the class passed both tests*
    - *42% of the class passed the first test.*

What percent of those who passed the first test also passed the second test?

Examples: Conditional Probability

2. Suppose that for houses in New England
   - *84% of the houses have a garage*
   - *65% of the houses have a garage and a back yard.*

What is the probability that a house has a backyard given that it has a garage?

To Jupyter…

Probability Density Functions

- **Problem:** If *X* is a *continuous* variable, then *P*(*X*=*x*) is **0** for any outcome ***x***

Single Outcome

$$X \sim \mathrm{Normal}(0,1)$$

$$P(X = \pi) = 0$$

$$P(3.1 \leq X \leq 3.2) \neq 0$$

Event

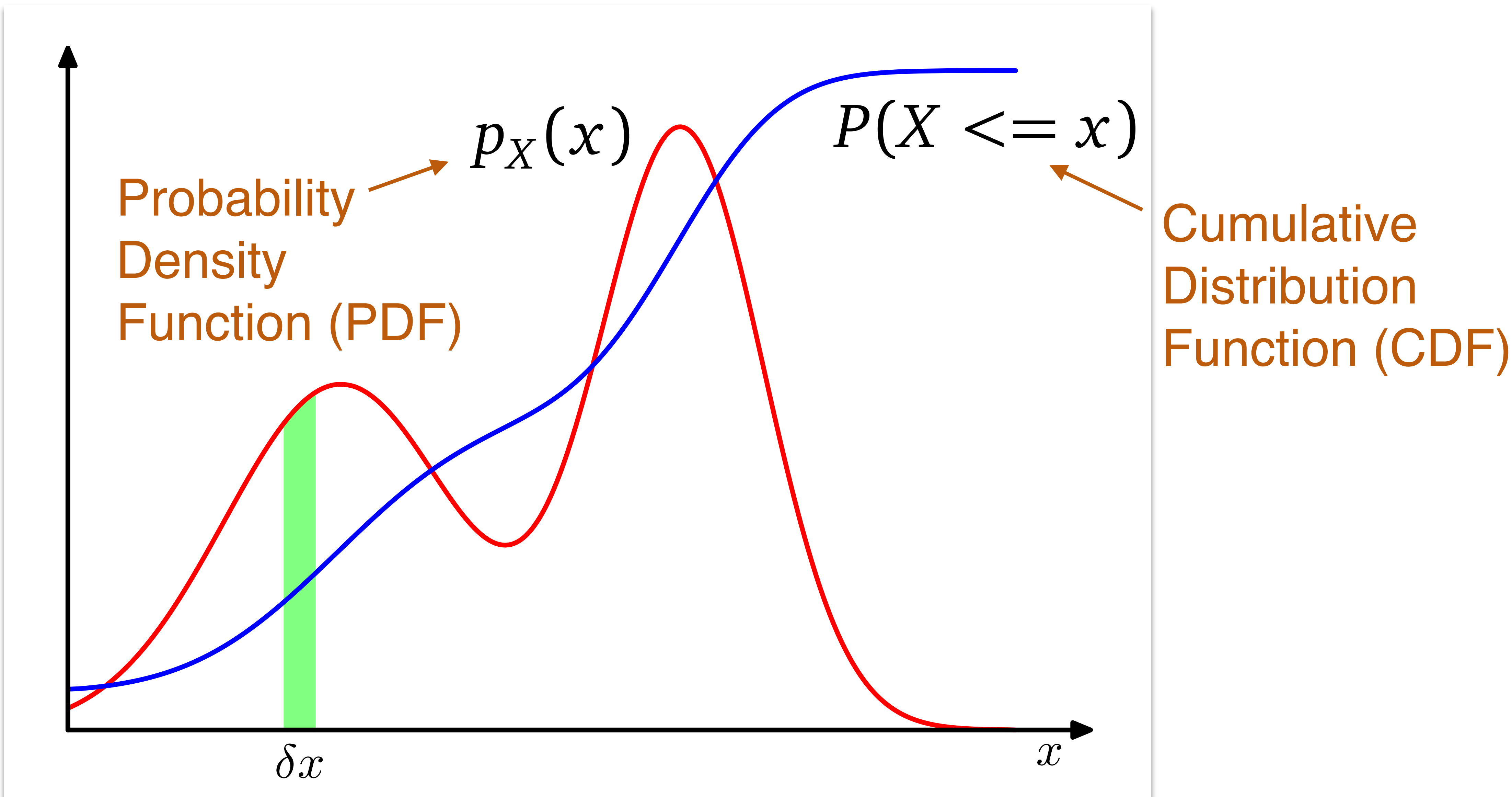
- **Solution:** Define a density function as a derivative

Capital P for probability

$$p_X(x) = \lim_{\delta \to 0} \frac{P(x - \delta < X < x + \delta)}{2\delta}$$

Small p for density

Probability Density Functions



$$p_X(x) = \lim_{\delta \to 0} \frac{P(x - \delta < X < x + \delta)}{2\delta}$$

Expected Values

$$X \sim p(x)$$ ← X is a random variable with density p(x)

Statistics

$$\mathbb{E}[X] = \sum_x p(x)\, x$$

$$\mathbb{E}[X] = \int dx\, p(x)\, x$$

Machine Learning

$$\mathbb{E}_{p(x\,|\,y)}[f(x)] = \sum_x p(x\,|\,y)\, f(x)$$

$$\mathbb{E}_{p(x\,|\,y)}[f(x)] = \int dx\, p(x\,|\,y)\, f(x)$$

Mean, Variance, Covariance

Mean

$$\mu_X = \mathbb{E}[X]$$

Variance

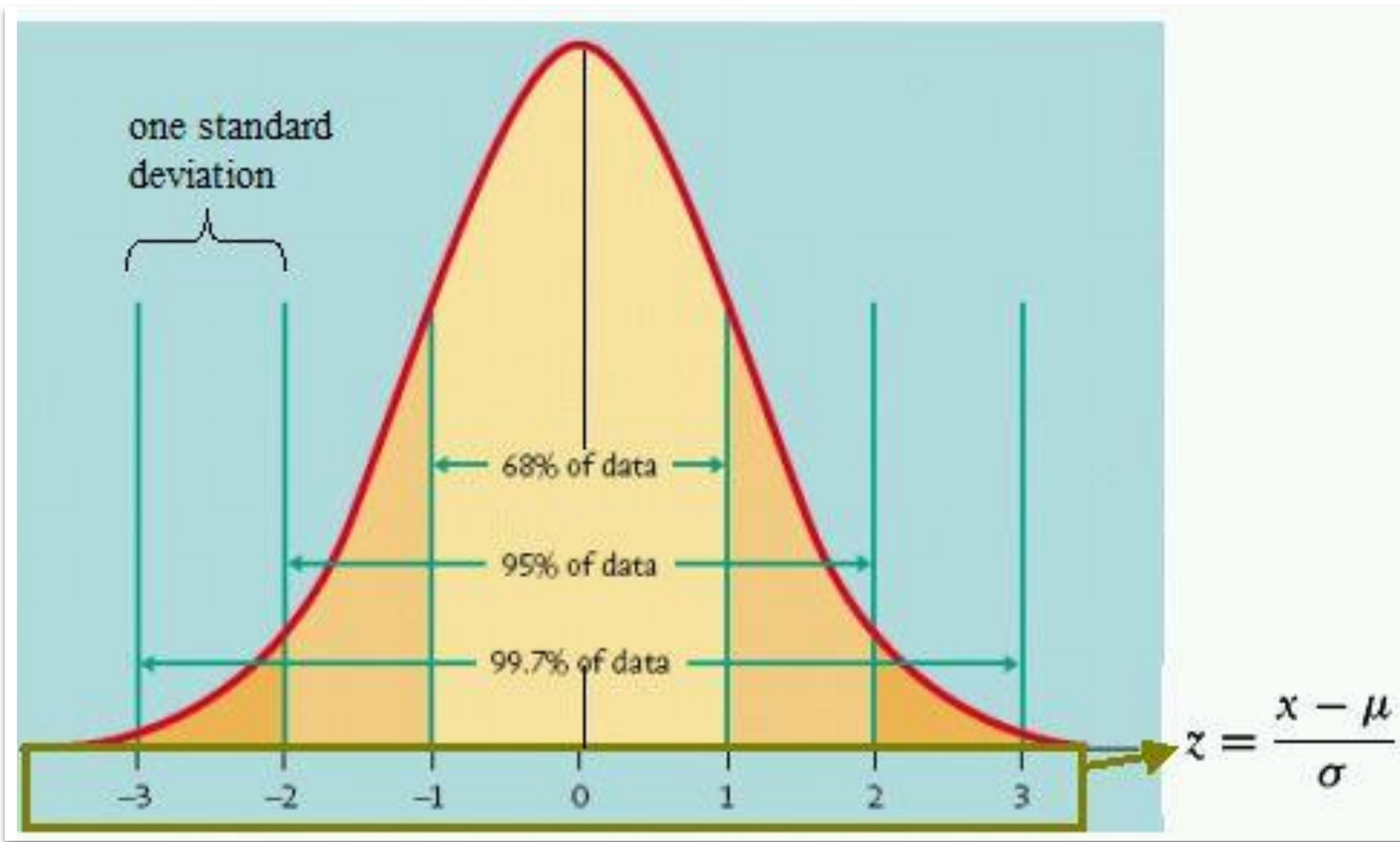
$$\sigma_X^2 = \mathrm{Var}[X] = \mathbb{E}[(X - \mu_X)^2]$$

Covariance

$$\Sigma_{X,Y} = \mathrm{Cov}[X, Y] = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$$

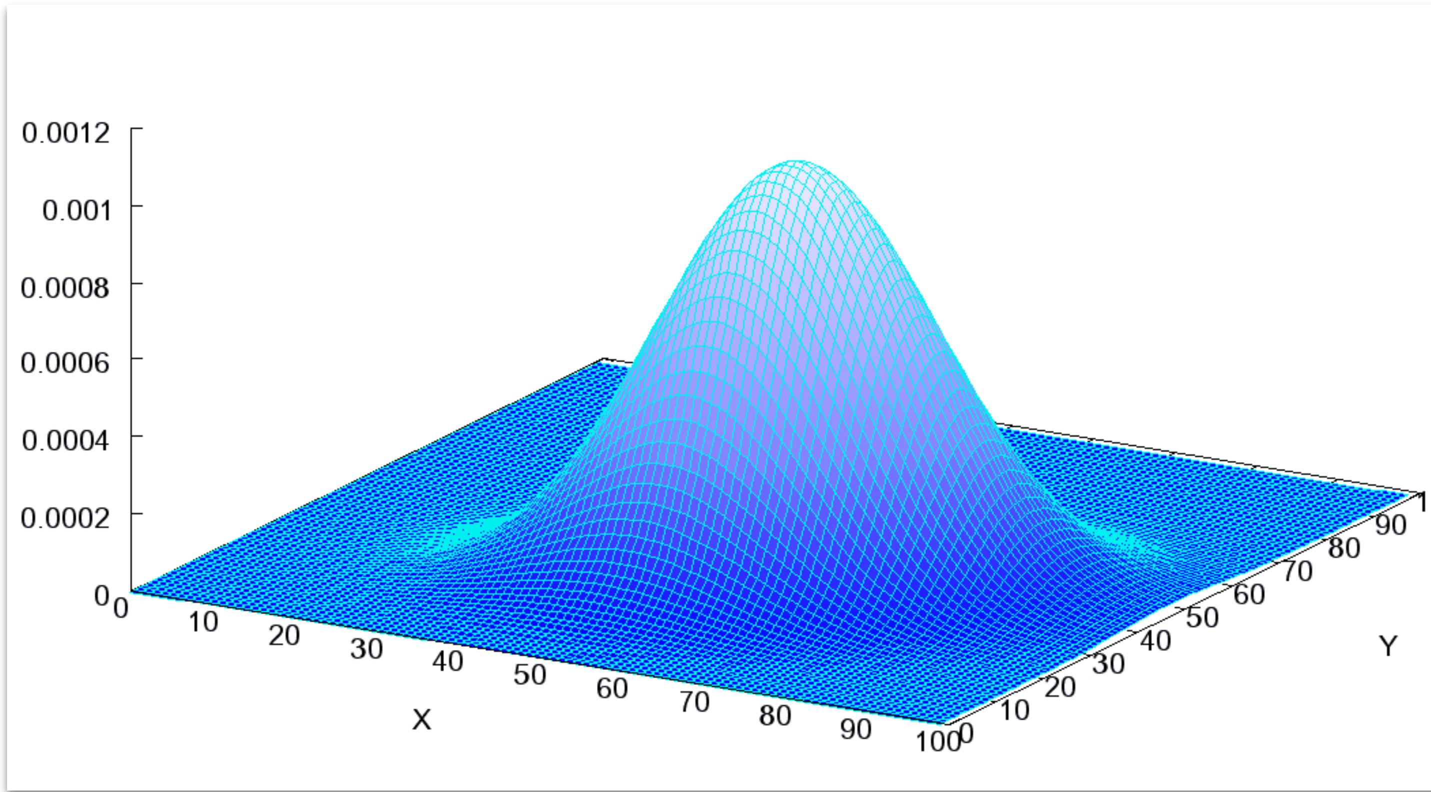
Properties of Gaussians

Normal Distribution



Density: $$p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{1}{2}(x-\mu)^2/\sigma^2}$$

Multivariate Normal



Vectors
$(x_1, \ldots, x_D)$
$(\mu_1, \ldots, \mu_D)$

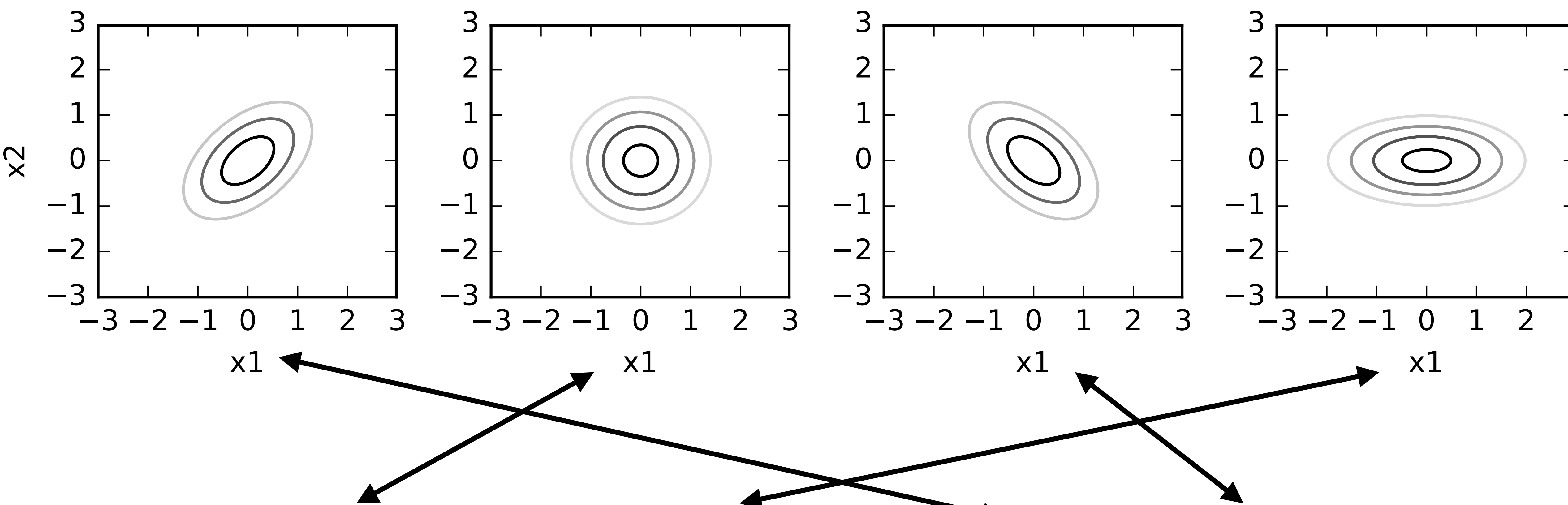
Density: $p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} \exp^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$

Covariance Matrix

Mean: $\mathbb{E}[X_d] = \mu_d$ **Covariance:** $\mathrm{Cov}[X_d, X_e] = \Sigma_{de}$

Covariance Matrices

Density: $f(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \exp^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$



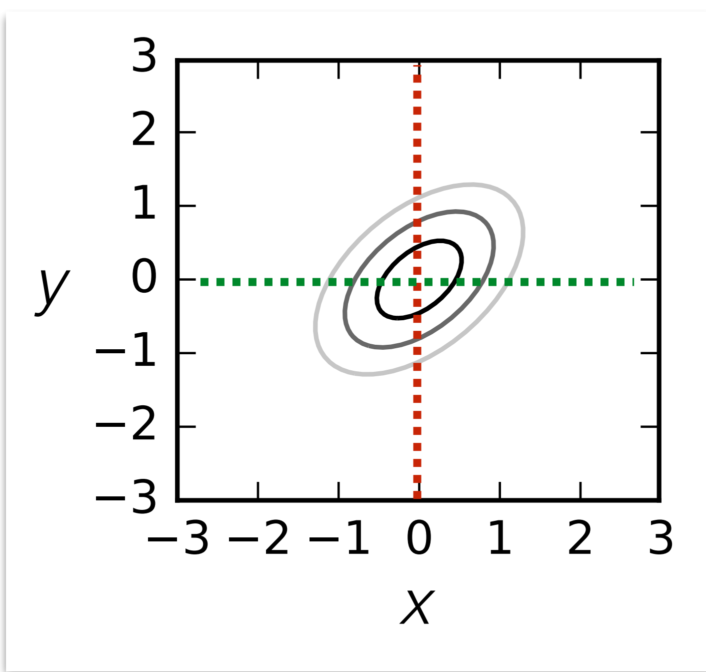
$\begin{bmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{bmatrix}$ $\begin{bmatrix} 2.0 & 0.0 \\ 0.0 & 0.5 \end{bmatrix}$ $\begin{bmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{bmatrix}$ $\begin{bmatrix} 1.0 & -0.5 \\ -0.5 & 1.0 \end{bmatrix}$

Question: Which covariance matrix Σ corresponds to which plot?

Marginals and Conditionals

Suppose that ***x*** and ***y*** are jointly Gaussian:

$$\mathbf{z} = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^T & \mathbf{B} \end{bmatrix} \right)$$



Question: What are the marginal distributions *p*(***x***) and *p*(***y***)?

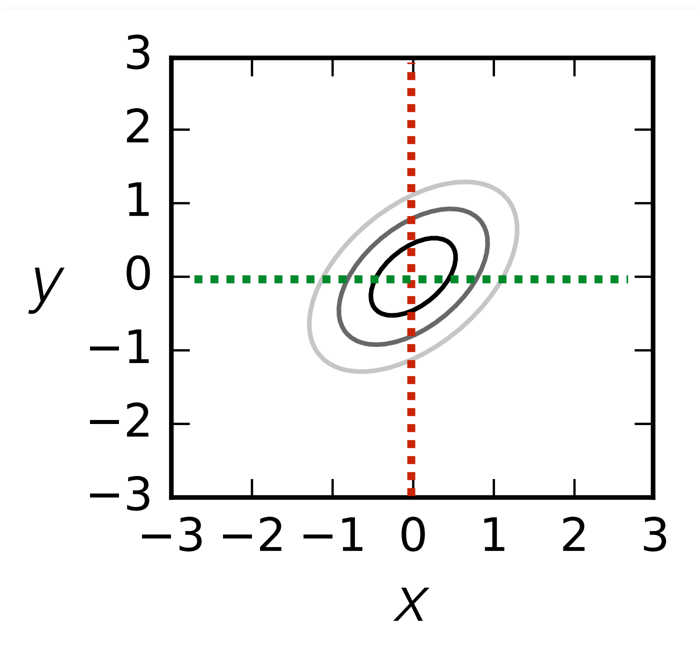
$$\mathbf{x} \sim \mathcal{N}(\mathbf{a}, \mathbf{A})$$

$$\mathbf{y} \sim \mathcal{N}(\mathbf{b}, \mathbf{B})$$

Marginals and Conditionals

Suppose that ***x*** and ***y*** are jointly Gaussian:

$$\mathbf{z} = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^T & \mathbf{B} \end{bmatrix} \right)$$



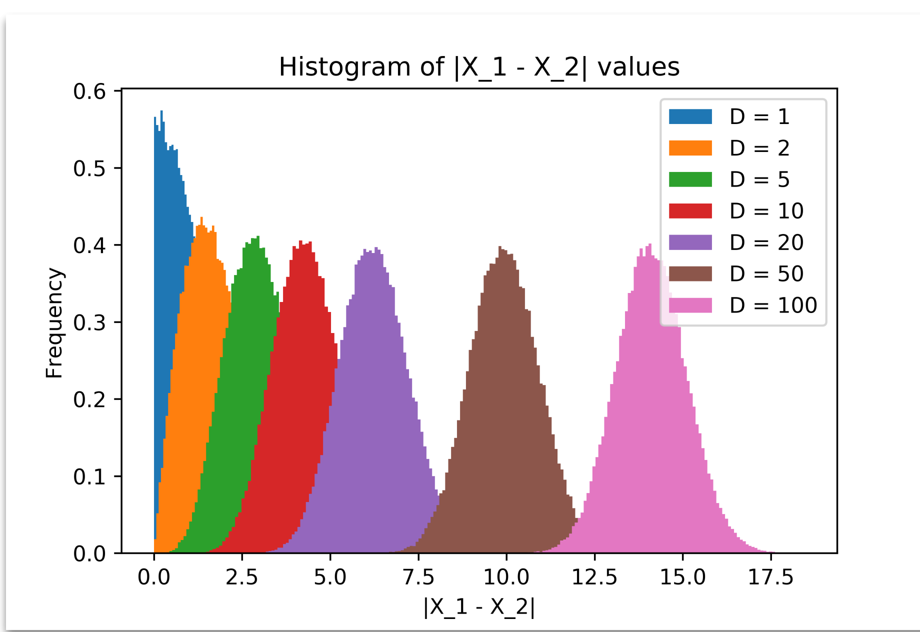
Once can derive the conditional distributions $p(\mathbf{x} \mid \mathbf{y})$ and $p(\mathbf{y} \mid \mathbf{x})$ in closed form as well; they are also Normals!

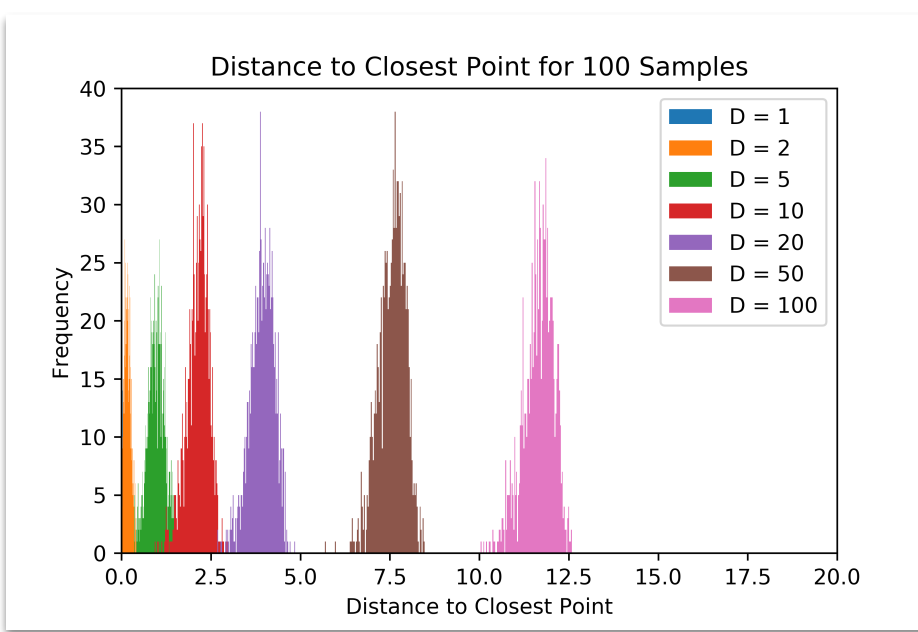
Curse of Dimensionality

Question: Suppose that X_1 and X_2 are independent Gaussian variables with <u>diagonal covariance</u>

$$X_1 \sim \mathrm{Normal}(0, \sigma^2 I_D) \qquad X_2 \sim \mathrm{Normal}(0, \sigma^2 I_D)$$

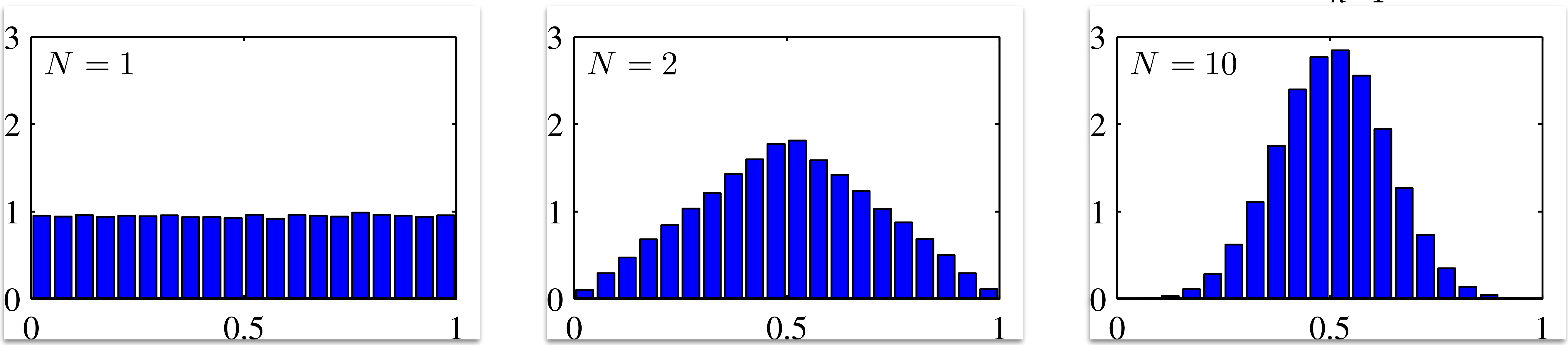
How does the distribution on the distance $|X_1 - X_2|$ change as we increase the dimension D?





Central Limit Theorem

$\bar{X} = X_1$ $\qquad$ $\bar{X} = \frac{1}{2}(X_1 + X_2)$ $\qquad$ $\bar{X} = \frac{1}{N}\sum_{n=1}^{N} X_n$



If X_1, …, X_N are

1. Independent identically distributed (i.i.d.)
2. Have finite variance $0 < \sigma_X{}^2 < \infty$

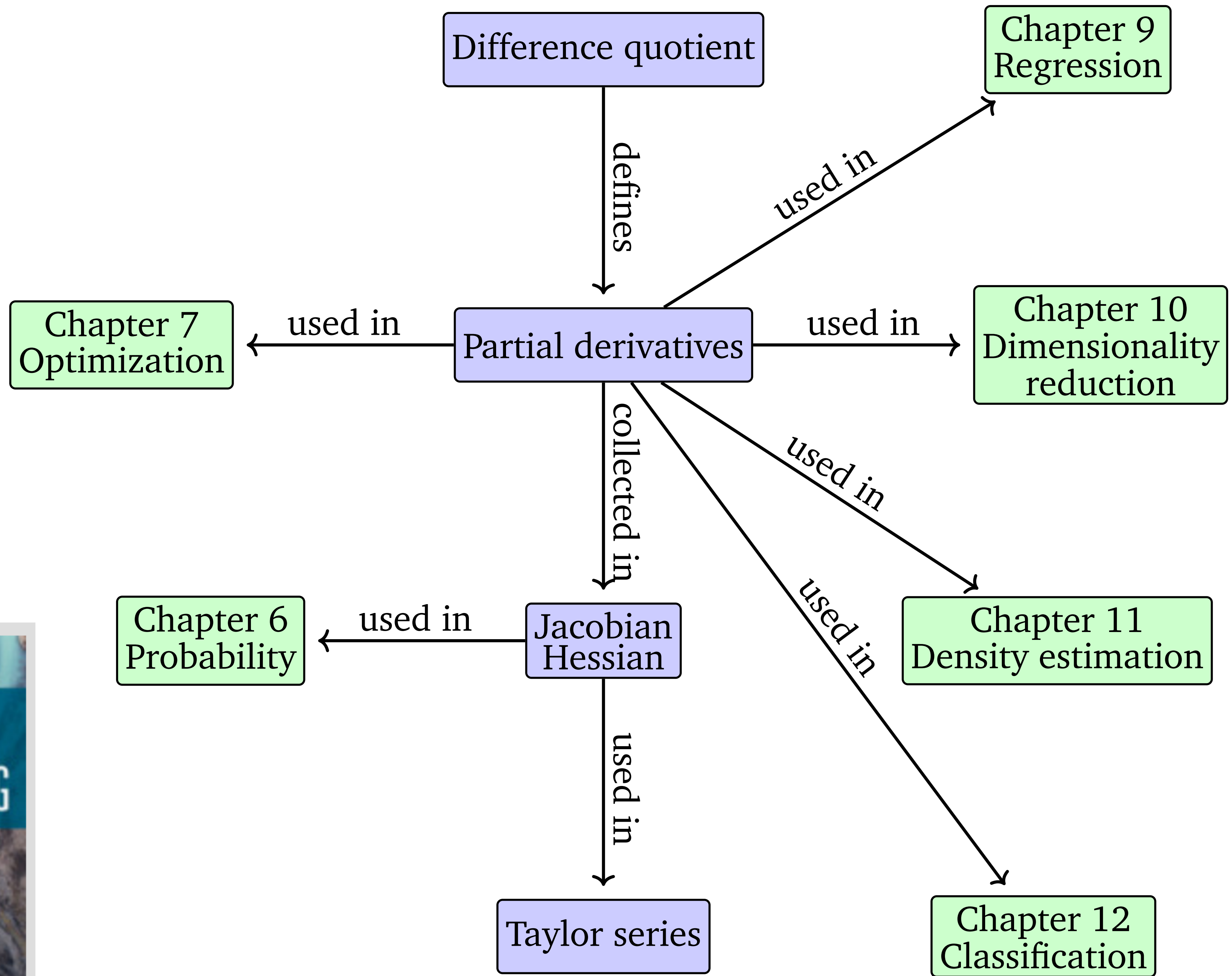
Then, as N approaches **∞**, the mean is distributed as

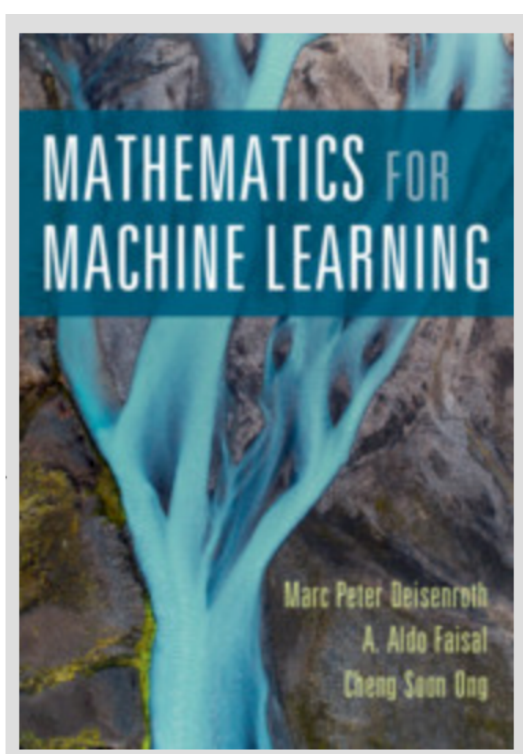
$$p(\bar{x}) = \mathrm{Normal}(\bar{x}; \mu_X, \sigma_X^2/N) \qquad \bar{X} = \frac{1}{N}\sum_{n=1}^{N} X_n$$

Summary of Gaussians

In sum, the Normal (or Gaussian) distribution pops up everywhere and is easy to work with; familiarize yourself with it!

Some Calc Review


Difference quotient
defines
Partial derivatives
used in
Chapter 9
Regression
used in
Chapter 7
Optimization
used in
Chapter 10
Dimensionality
reduction
collected in
used in
Chapter 11
Density estimation
used in
Jacobian
Hessian
used in
Chapter 6
Probability
used in
Taylor series
Chapter 12
Classification


MATHEMATICS FOR
MACHINE LEARNING
Marc Peter Deisenroth
A. Aldo Faisal
Cheng Soon Ong

Univariate Functions

$$y = f(x),\ x, y \in \mathbb{R}$$

Difference Quotient

$$\frac{\delta y}{\delta x} := \frac{f(x + \delta x) - f(x)}{\delta x}$$

Univariate Functions

Derivative (formally)

$$\frac{\mathrm{d}f}{\mathrm{d}x} := \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

Univariate Functions

Derivative (formally)

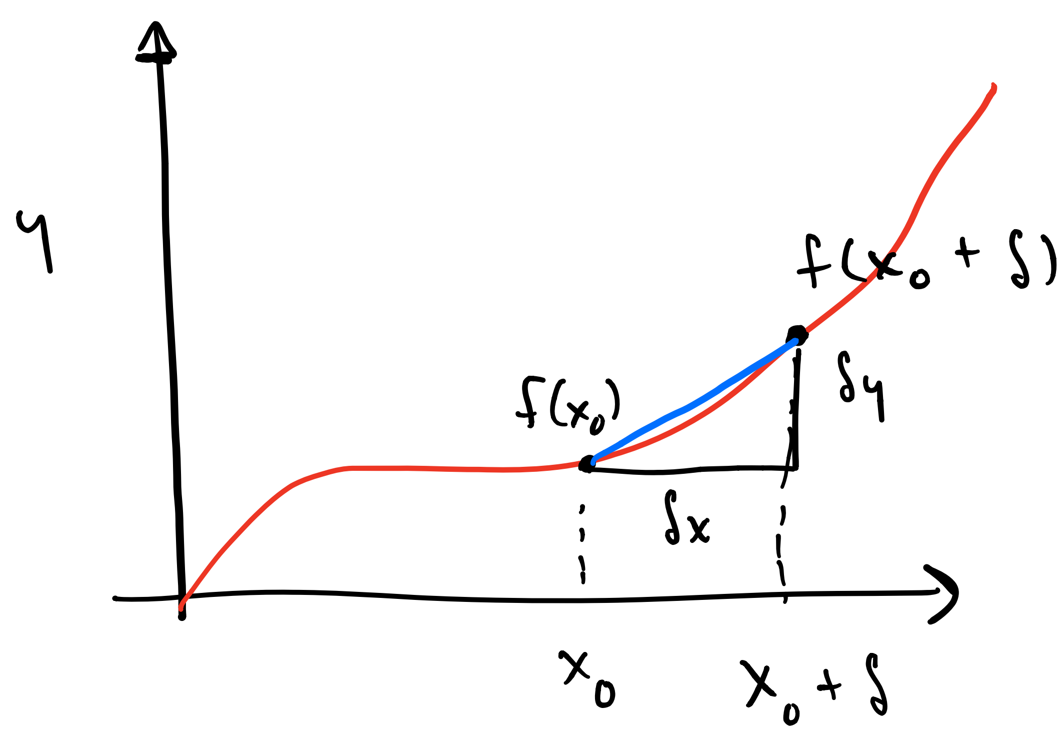
$$\frac{\mathrm{d}f}{\mathrm{d}x} := \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

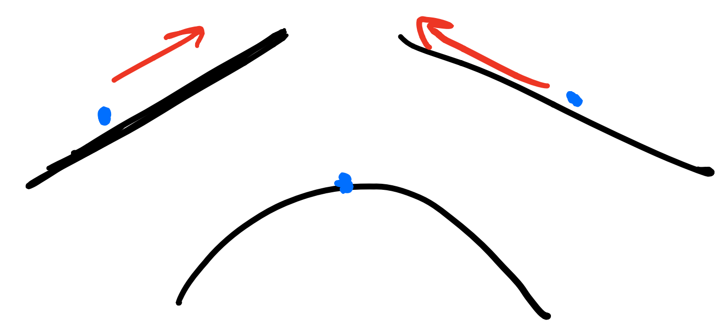
The derivative points in the direction of steepest ascent

Consider univariate case.

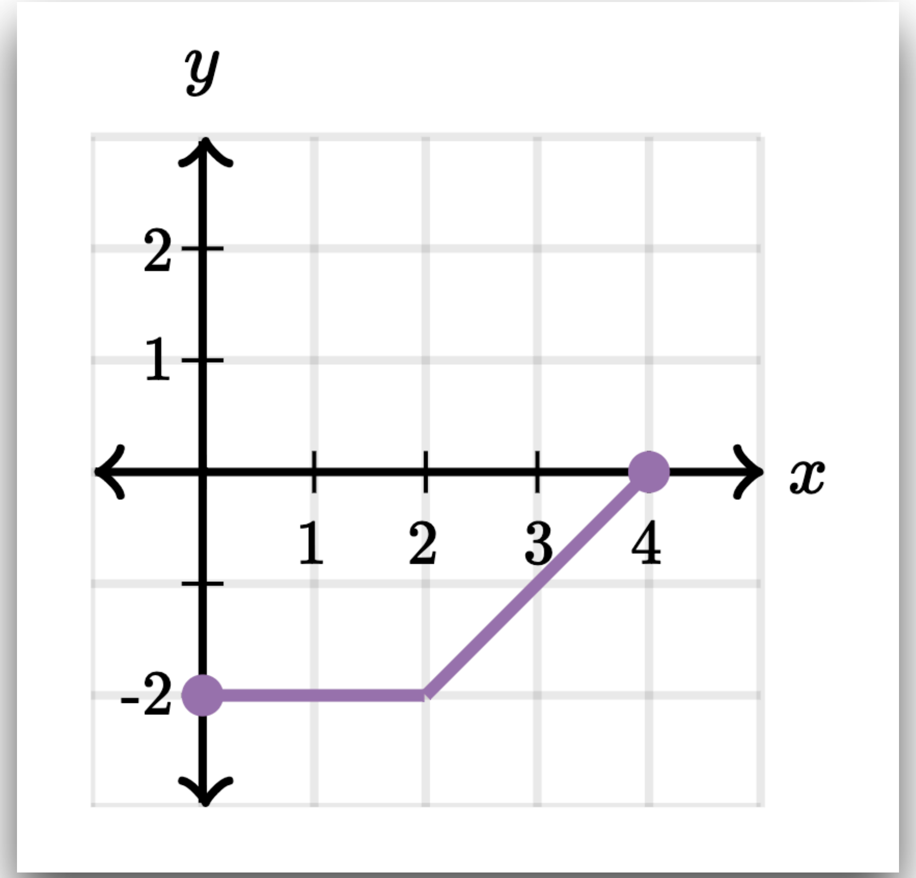
$$y = f(x)$$

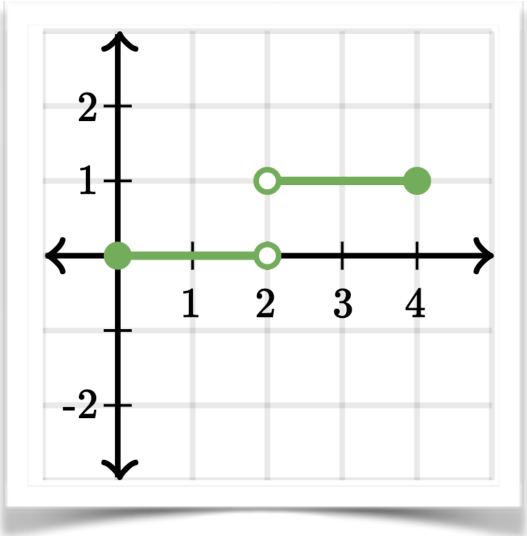
$$\frac{\delta y}{\delta x} \stackrel{\text{def}}{=} \frac{f(x+\delta x) - f(x)}{\delta x}$$



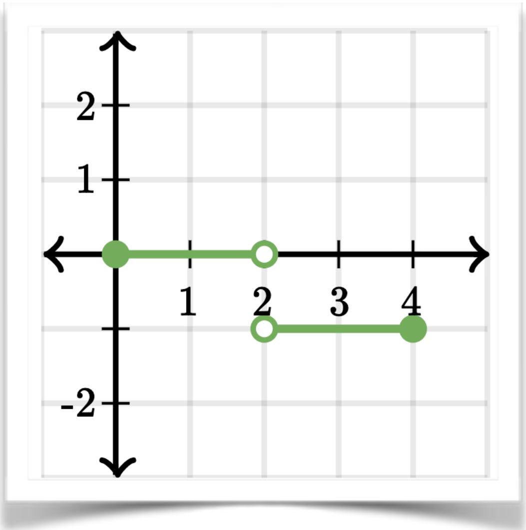


Spot the derivative

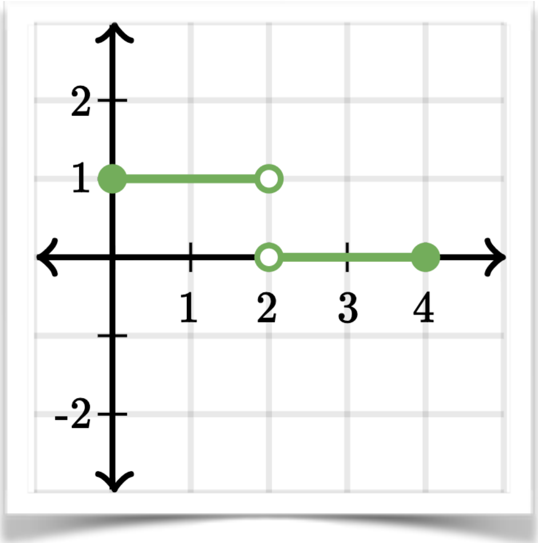




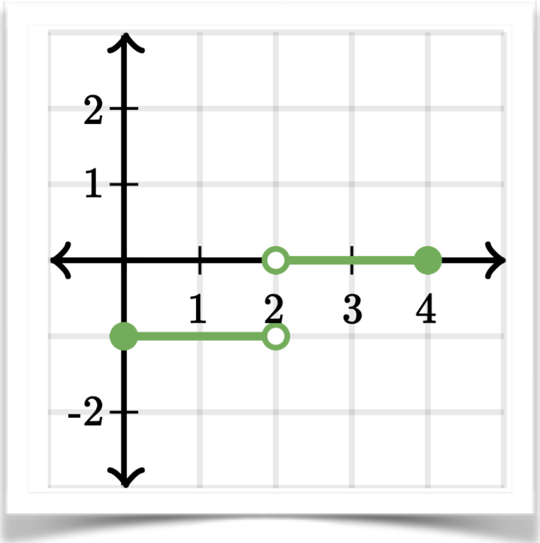
A



B



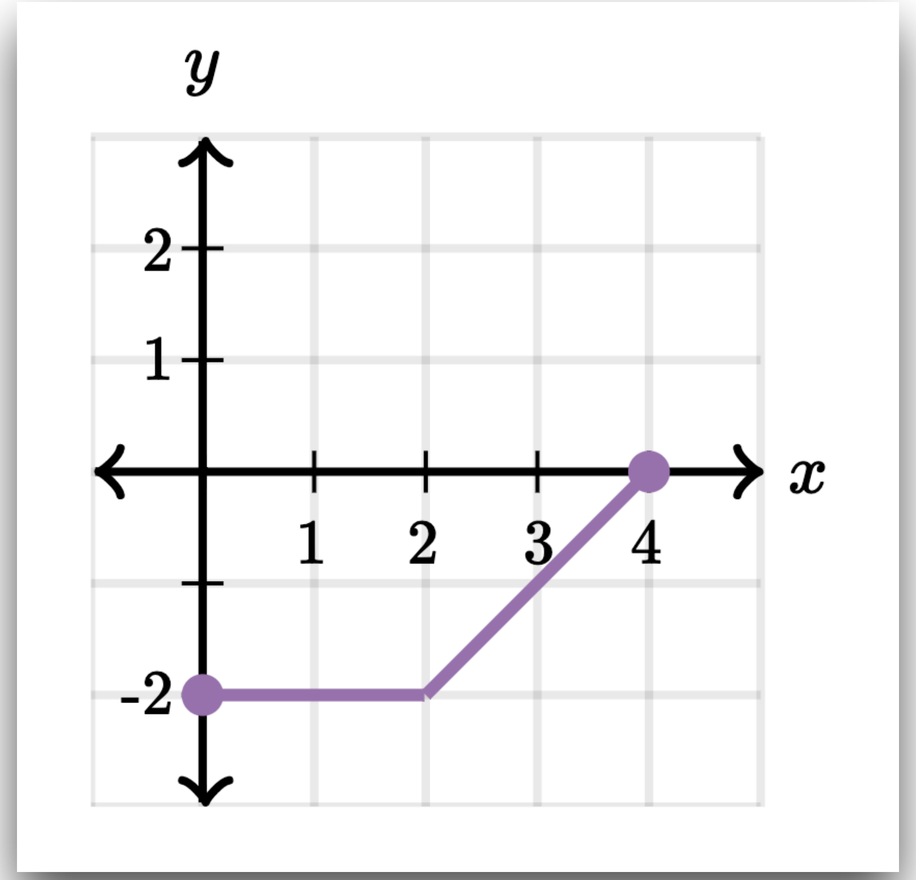
C

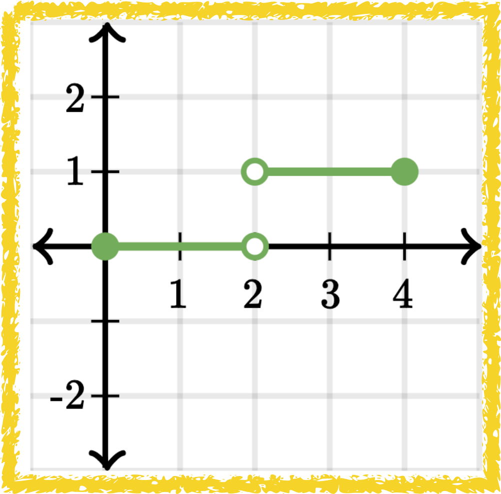


D

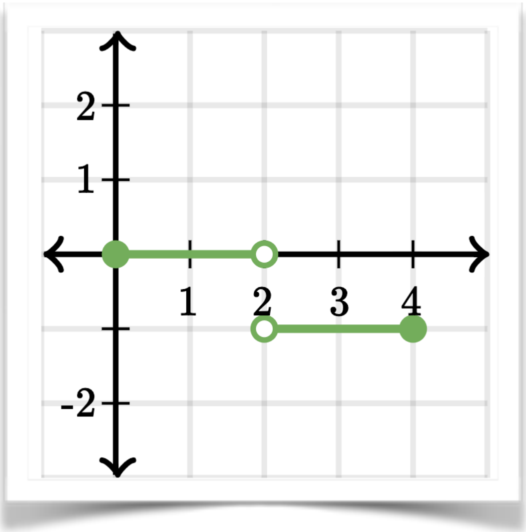
Example from Khan academy

Spot the derivative

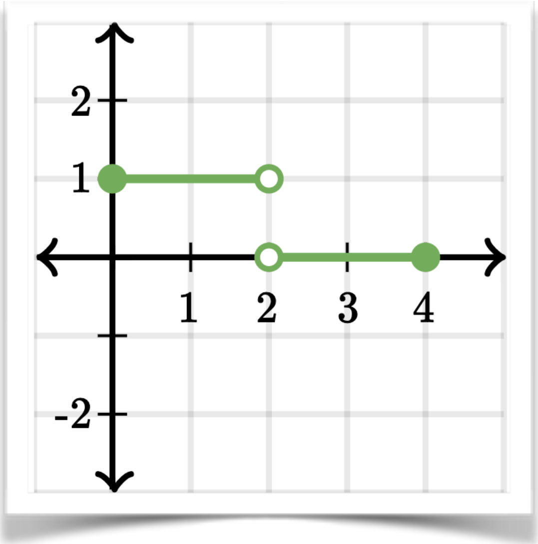




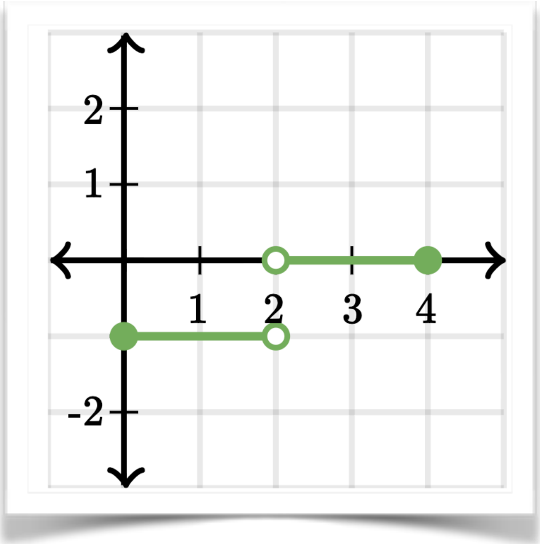
A



B



C



D

Example from Khan academy

Sum Rule

$$(f(x) + g(x))' = f'(x) + g'(x)$$

Product Rule

$$(f(x)g(x))' = f'(x)g(x) + f(x)g'(x)$$

Chain Rule

$$\big(g(f(x))\big)' = (g \circ f)'(x) = g'(f(x))f'(x)$$

Gradients

Usually in ML we care about multivariate functions

$\boldsymbol{x} \in \mathbb{R}^n$ of n variables $x_1, \ldots, x_n$

$f \ : \ \mathbb{R}^n \ \to \ \mathbb{R}$

Gradients

Partial derivatives are taken w.r.t. one dimension at a time:

$$\frac{\partial f}{\partial x_1} = \lim_{h \to 0} \frac{f(x_1 + h, x_2, \ldots, x_n) - f(\boldsymbol{x})}{h}$$

$$\vdots$$

$$\frac{\partial f}{\partial x_n} = \lim_{h \to 0} \frac{f(x_1, \ldots, x_{n-1}, x_n + h) - f(\boldsymbol{x})}{h}$$

Gradients

Group the gradients into a vector (the *gradient*)

$$\nabla_{\boldsymbol{x}} f = \mathrm{grad} f = \frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{x}} = \begin{bmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} & \frac{\partial f(\boldsymbol{x})}{\partial x_2} & \cdots & \frac{\partial f(\boldsymbol{x})}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{1 \times n}$$

Example

$$f(x_1, x_2) = x_1^2 x_2 + x_1 x_2^3$$

$$\frac{\partial f(x_1, x_2)}{\partial x_1} = 2x_1 x_2 + x_2^3$$
$$\frac{\partial f(x_1, x_2)}{\partial x_2} = x_1^2 + 3x_1 x_2^2$$

$$\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{x}} = \begin{bmatrix} \frac{\partial f(x_1, x_2)}{\partial x_1} & \frac{\partial f(x_1, x_2)}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2x_1 x_2 + x_2^3 & x_1^2 + 3x_1 x_2^2 \end{bmatrix} \in \mathbb{R}^{1\times 2}$$

Rules still hold!

Sum rule: $$\frac{\partial}{\partial \boldsymbol{x}}\big(f(\boldsymbol{x}) + g(\boldsymbol{x})\big) = \frac{\partial f}{\partial \boldsymbol{x}} + \frac{\partial g}{\partial \boldsymbol{x}}$$

Product rule: $$\frac{\partial}{\partial \boldsymbol{x}}\big(f(\boldsymbol{x})g(\boldsymbol{x})\big) = \frac{\partial f}{\partial \boldsymbol{x}}g(\boldsymbol{x}) + f(\boldsymbol{x})\frac{\partial g}{\partial \boldsymbol{x}}$$

Chain rule: $$\frac{\partial}{\partial \boldsymbol{x}}(g \circ f)(\boldsymbol{x}) = \frac{\partial}{\partial \boldsymbol{x}}\big(g(f(\boldsymbol{x}))\big) = \frac{\partial g}{\partial f}\frac{\partial f}{\partial \boldsymbol{x}}$$

Rules still hold!

Sum rule: $$\frac{\partial}{\partial \boldsymbol{x}}\big(f(\boldsymbol{x}) + g(\boldsymbol{x})\big) = \frac{\partial f}{\partial \boldsymbol{x}} + \frac{\partial g}{\partial \boldsymbol{x}}$$

Product rule: $$\frac{\partial}{\partial \boldsymbol{x}}\big(f(\boldsymbol{x})g(\boldsymbol{x})\big) = \frac{\partial f}{\partial \boldsymbol{x}}g(\boldsymbol{x}) + f(\boldsymbol{x})\frac{\partial g}{\partial \boldsymbol{x}}$$

Chain rule: $$\frac{\partial}{\partial \boldsymbol{x}}(g \circ f)(\boldsymbol{x}) = \frac{\partial}{\partial \boldsymbol{x}}\big(g(f(\boldsymbol{x}))\big) = \frac{\partial g}{\partial f}\frac{\partial f}{\partial \boldsymbol{x}}$$

… but be mindful of dims!

For review: Problem 5.7 in MML

5.7 Compute the derivatives $\mathrm{d}f/\mathrm{d}\boldsymbol{x}$ of the following functions by using the chain rule. Provide the dimensions of every single partial derivative. Describe your steps in detail.

a.

$$f(z) = \log(1 + z)\,, \quad z = \boldsymbol{x}^\top \boldsymbol{x}\,, \quad \boldsymbol{x} \in \mathbb{R}^D$$

b.

$$f(\boldsymbol{z}) = \sin(\boldsymbol{z})\,, \quad \boldsymbol{z} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}\,, \quad \boldsymbol{A} \in \mathbb{R}^{E \times D}, \boldsymbol{x} \in \mathbb{R}^D, \boldsymbol{b} \in \mathbb{R}^E$$

where $\sin(\cdot)$ is applied to every element of $\boldsymbol{z}$.

(a) $\quad f(z) = \log(\overbrace{1+z}^{u}) \qquad z = x^T x \qquad x \in \mathbb{R}^D$

$$\frac{df}{dx} = \frac{d}{du}\log u \,\frac{d}{dx}(1 + x^T x)$$

$$= \frac{1}{(1+x^T x)}\, 2x = \frac{2x}{(1+x^T x)}$$

(b) $f(z) = \sin(z) \qquad z = Ax + b \qquad A \in \mathbb{R}^{E \times D} \qquad x \in \mathbb{R}^{D}$

$b \in \mathbb{R}^{E}$

$$\frac{\partial f}{\partial x} = \frac{d \sin(u)}{du} \frac{du}{dx}$$

$$= \cos(u) \frac{\partial}{\partial x} Ax + b$$

$$= \cos(\underbrace{Ax + b}_{E \times 1}) \cdot \underbrace{A}_{E \times D}$$

intermezzo: the joys of auto-diff…

… or, first steps in pytorch