Probability Theory

CMPUT 261: Introduction to Artificial Intelligence

P&M §8.1

Logistics & Assignment #1

Assignment #1 was released last week

- Available on eClass
- Due: Thursday **September 28 at 11:59pm**

Recap: Search

- Agent searches **internal representation** to find solution
- **Fully-observable, deterministic, offline, single-agent** problems
- **Graph search** finds a **sequence of actions** to a goal node
    - Efficiency gains from using **heuristic functions** to encode **domain knowledge**
- **A*** makes optimal** use of the heuristic
- **Branch-and-bound** uses heuristic close to optimal use of heuristic and has exponentially better space complexity
    - But only when the upper bound is set *just so*
- **IDA*** uses iterative deepening to set the upper bound

Lecture Outline

1. Recap
2. Uncertainty
3. Probability Semantics
4. Conditional Probability
5. Expected Value

After this lecture, you should be able to:

- Compute joint, marginal, and conditional probabilities
- Compute expected values
- Apply Bayes' Rule to compute posterior probabilities
- Apply the Chain Rule (of Probabilities) to compute joint probabilities

Uncertainty

- In search problems, agent has **perfect knowledge** of the world and its dynamics
- In most applications, an agent cannot just **make assumptions** and then act according to those assumptions
- Knowledge is **uncertain**:
  - Must consider **multiple** hypotheses
  - Must **update** beliefs about which hypotheses are likely given **observations**

Example: Wearing a Seatbelt

- An agent has to decide between **three actions**:
    1. Drive without wearing a seatbelt
    2. Drive while wearing a seatbelt
    3. Stay home
- If the agent *knows* that an accident **will** happen, it will just stay home
- If the agent *knows* that an accident **will not** happen, it will not bother to wear a seatbelt!
- Wearing a seatbelt only makes sense because the agent is **uncertain** about whether driving will lead to an accident.

Measuring Uncertainty

- **Probability** is a way of **measuring** uncertainty
- We assign a number between 0 and 1 to **events** (hypotheses):
  - **0** means absolutely certain that statement is **false**
  - **1** means absolutely certain that statement is **true**
  - **Intermediate** values mean more or less certain
- Probability is a measurement of **uncertainty**, **not truth**
  - A statement with probability .75 is not "mostly true"
  - Rather, we **believe** it is more **likely** to be true than not

Subjective vs. Objective: The Frequentist Perspective

- Probabilities can be interpreted
  as **objective** statements about the **world**, or
  as **subjective** statements about an agent's **beliefs**.

- Objective view is called **frequentist:**
  - The probability of an event is the proportion of times it would happen **in the long run** of **repeated experiments**
  - Every event has a single, **true** probability
  - Events that can only happen **once** don't have a well-defined probability

Subjective vs. Objective: The Bayesian Perspective

- Probabilities can be interpreted
  as **objective** statements about the **world**, or
  as **subjective** statements about an agent's **beliefs**.
- Subjective view is called **Bayesian:**
  - The probability of an event is a measure of an agent's **belief** about its likelihood
  - Different agents can legitimately have **different beliefs**, so they can legitimately assign **different probabilities** to the same event
  - There is **only one way** to **update** those beliefs in response to new data
- In this course, we will primarily take the **Bayesian** view

Example: Dice

- Diane rolls a **fair, six-sided die**, and gets the number X
  - **Question:** What is $P(X = 5)$? (the probability that Diane rolled a 5)
- Diane truthfully tells Oliver that she rolled an **odd** number.
  - **Question:** What should **Oliver** believe $P(X = 5)$ is?
- Diane truthfully tells Greta that she rolled a number ≥ 5.
  - **Question:** What should **Greta** believe $P(X = 5)$ is?
- **Question:** What is $P(X = 5)$?

Example: Dice

- Diane rolls a **fair, six-sided die**, and gets the number X
  - **Question:** What is $P(X = 5)$? (the probability that Diane rolled a 5) $P(X = 5)$
- Diane truthfully tells Oliver that she rolled an **odd** number.
  - **Question:** What should **Oliver** believe $P(X = 5)$ is? → $P(X = 5 \mid X \text{ is odd})$
- Diane truthfully tells Greta that she rolled a number ≥ 5.
  - **Question:** What should **Greta** believe $P(X = 5)$ is? → $P(X = 5 \mid X \geq 5)$
- **Question:** What is $P(X = 5)$? → $P(X = 5 \mid X \text{ is odd}, X \geq 5)$

Semantics: Possible Worlds

- **Random variables** take values from a **domain**.
  We will write them as uppercase letters (e.g., X, Y, D, etc.)

- A **possible world** is a **complete assignment** of values to variables
  We will usually write a single "world" as ω and the set of all possible worlds as Ω
  In this lecture: worlds are **discrete** (i.e., we can take sums)

- A **probability measure** is a function $P : \Omega \rightarrow \mathbb{R}$ over **possible worlds** ω satisfying:

1. $$\sum_{\omega \in \Omega} P(\omega) = 1$$
2. $$P(\omega) \geq 0 \;\; \forall \omega \in \Omega$$

Propositions

- A **primitive proposition** is an equality or inequality expression
  E.g., $X = 5$ or $X \geq 4$
- A **proposition** is built up from other propositions using **logical connectives**.
  E.g., $(X = 1 \vee X = 3 \vee X = 5)$
- The **probability** of a proposition is the sum of the probabilities of the possible worlds in which that proposition is true:

$$P(\alpha) = \sum_{\omega : \omega \models \alpha} P(\omega)$$

$\omega \models \alpha$ means "α is true in ω"

- Therefore:

$$P(\alpha \vee \beta) \geq P(\alpha)$$

$\alpha \vee \beta$ means "α OR β"

$$P(\alpha \wedge \beta) \leq P(\alpha)$$

$\alpha \wedge \beta$ means "α AND β"

$$P(\neg\alpha) = 1 - P(\alpha)$$

$\neg\alpha$ means "NOT α"

Joint Distributions

- In our dice example, there was a **single** random variable
- We typically want to think about the interactions of **multiple** random variables
- A **joint distribution** assigns a probability to each full assignment of values to variables
  - e.g., $P(X = 1, Y = 5)$. Equivalent to $P(X = 1 \wedge Y = 5)$
  - Can view this as another way of specifying a single **possible world**

Joint Distribution Example

- What might a day be like in Edmonton? Random variables:
  - **Weather**, with domain {clear, snowing}
  - **Temperature**, with domain {mild, cold, very_cold}
- **Joint distribution** P(Weather, Temperature):

Weather	Temperature	P
clear	mild	0.20
clear	cold	0.30
clear	very cold	0.25
snowing	mild	0.05
snowing	cold	0.10
snowing	very cold	0.10

Marginalization

Question:

What is the **marginal distribution** of Weather?

- **Marginalization** is using a joint distribution $P(X_1, \ldots, X_m, \ldots X_n)$ to compute a distribution over a smaller number of variables $P(X_1, \ldots, X_m)$
  - Smaller distribution is called the **marginal distribution** of its variables (e.g., marginal distribution of $X_1, \ldots, X_m$)
- We compute the marginal distribution by summing out the other variables:

$$P(X, Y) = \sum_{w \in \mathrm{dom}(W)} \sum_{z \in \mathrm{dom}(Z)} P(W = w, X, Y, Z = z)$$

Weather	Temperature	P
clear	mild	0.20
clear	cold	0.30
clear	very cold	0.25
snowing	mild	0.05
snowing	cold	0.10
snowing	very cold	0.10

Conditional Probability

- Agents need to be able to **update** their beliefs based on new **observations**
- This process is called **conditioning**
- We write $P(h \mid e)$ to denote "probability of **hypothesis** h given that we have observed **evidence** e"
    - $P(h \mid e)$ is the **probability of h conditional on e**

Semantics of Conditional Probability

- Evidence e lets us **rule out** all of the worlds that are incompatible with e
  - E.g., if I observe that the weather is clear, I should no longer assign **any** probability to the worlds in which it is snowing
  - We need to **normalize** the probabilities of the remaining worlds to ensure that the probabilities of possible worlds sum to 1

$$P(\omega \mid e) = \begin{cases} c \times P(\omega) & \text{if } \omega \models e, \\ 0 & \text{otherwise.} \end{cases}$$

Semantics of Conditional Probability

- Evidence e lets us **rule out** all of the worlds that are incompatible with e
  - E.g., if I observe that the weather is clear, I should no longer assign **any** probability to the worlds in which it is snowing
  - We need to **normalize** the probabilities of the remaining worlds to ensure that the probabilities of possible worlds sum to 1

$$P(\omega \mid e) = \begin{cases} \frac{1}{P(e)} \times P(\omega) & \text{if } \omega \models e, \\ 0 & \text{otherwise.} \end{cases}$$

Conditional Probability Example

- My initial marginal belief about the weather was: $P(Weather = snow) = 0.25$
- Suppose I observe that the temperature is **mild**.
  - **Question:** What probability should I **now** assign to $Weather = snow$?

1. **Rule out** incompatible worlds
2. **Normalize** remaining probabilities
3. Result: $P(Weather = snow \mid Temperature = mild) = 0.20$

Weather	P
clear	.20 / (.20 + .05) = **0.8**
snowing	.05 / (.20 + .05) = **0.2**

~~clear~~	~~very cold~~	~~0.25~~
snowing	mild	0.05
~~snowing~~	~~cold~~	~~0.10~~
~~snowing~~	~~very cold~~	~~0.10~~

Chain Rule

Definition: conditional probability

$$P(h \mid e) = \frac{P(h, e)}{P(e)}$$

- We can run this **in reverse** to get

$$P(h, e) = P(h \mid e) \times P(e)$$

Definition: chain rule

$$P(\alpha_1, \ldots, \alpha_n) = P(\alpha_1) \times P(\alpha_2 \mid \alpha_1) \times \cdots \times P(\alpha_n \mid \alpha_1, \ldots, \alpha_{n-1})$$
$$= \Pi_{i=1}^{n} P(\alpha_i \mid \alpha_1, \ldots, \alpha_{i-1})$$

Bayes' Rule

- From the **chain rule**, we have

$$P(h, e) = P(h \mid e)P(e)$$
$$= P(e \mid h)P(h)$$

- **Often**, $P(e \mid h)$ is easier to compute than $P(h \mid e)$.

Bayes' Rule:

$$P(h \mid e) = \frac{P(e \mid h)\,P(h)}{P(e)}$$

Posterior: $P(h \mid e)$

Likelihood: $P(e \mid h)$

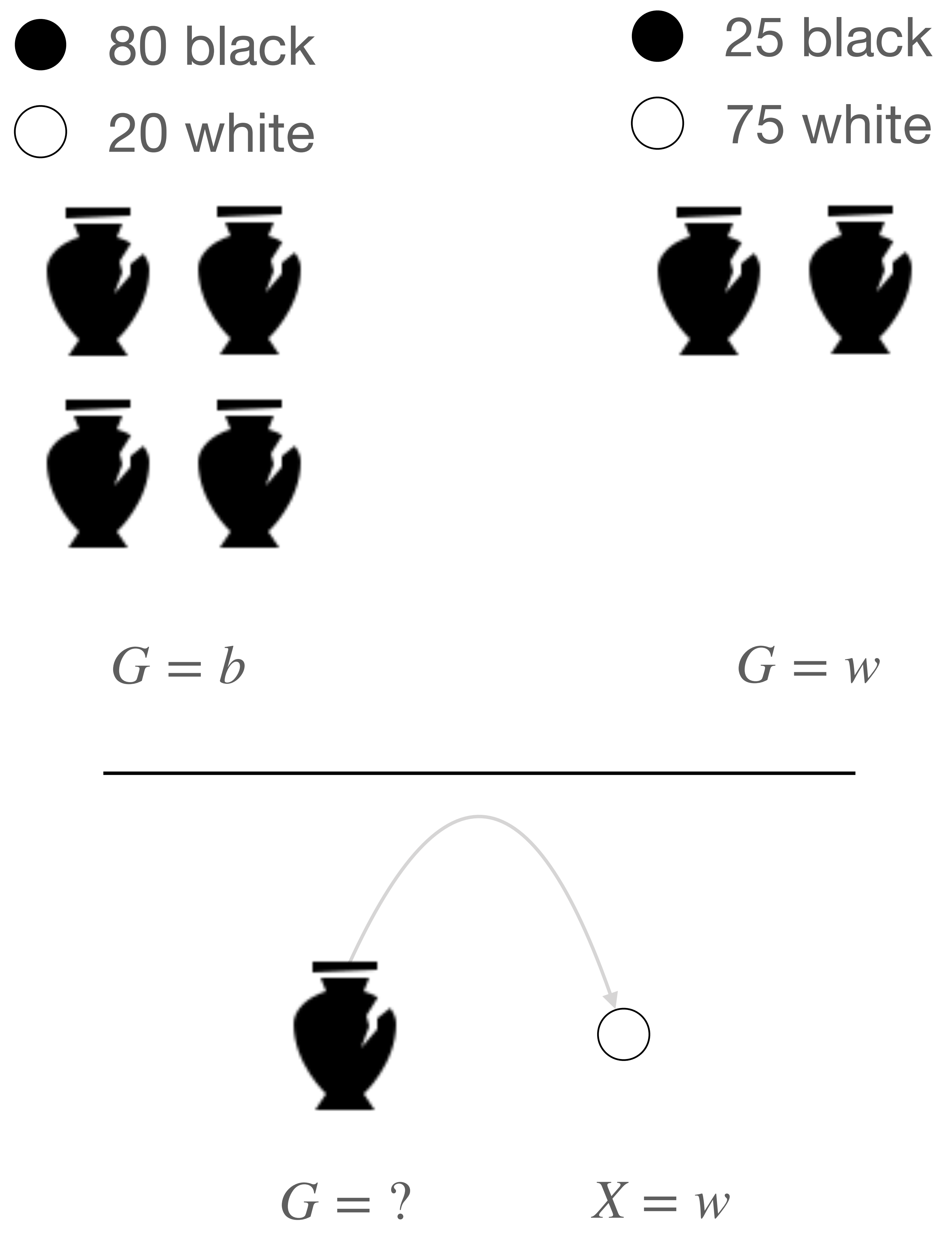
Prior: $P(h)$

Evidence: $P(e)$

Bayes' Rule Example: Urns

- 6 urns with 100 balls each
- Four have 80 black balls, 20 white; the other 2 have 25 black balls, 75 white
- I roll a fair die and choose the urn with the corresponding number
  - **Q:** With what probability are the majority of the balls in the chosen urn white? i.e., $\Pr(G = w)$
- I draw a ball from the urn; it's white! i.e., $X = w$
- **Conditional on that observation**, with what probability are most of the balls in the urn white?

i.e., $\Pr(G = w \mid X = w)$



icons by Icons8

Bayes' Rule Example: Urns

$$\Pr(G = w) = \frac{2}{6}$$

$$\Pr(X = w \mid G = w) = 0.75$$

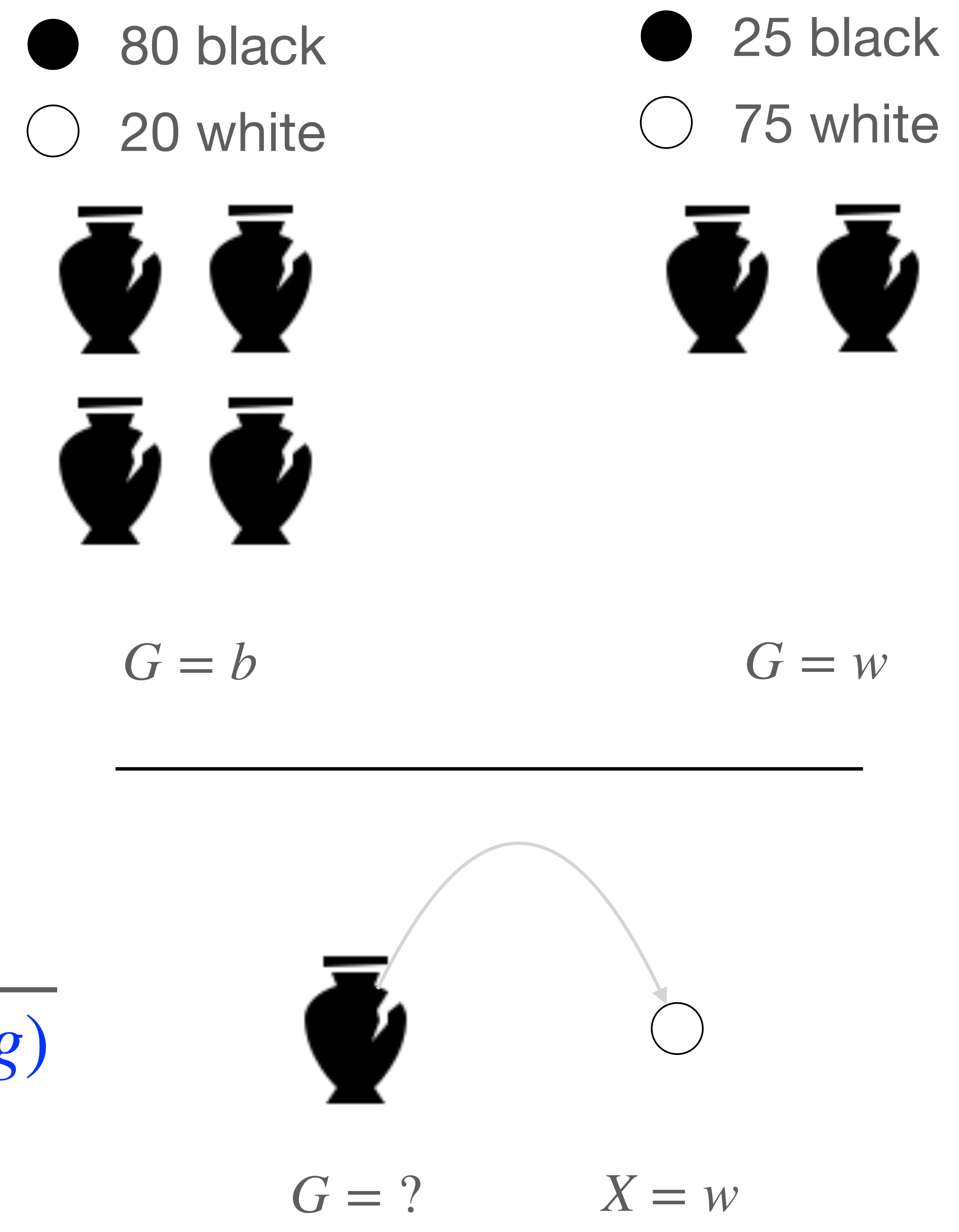
$$\Pr(G = w \mid X = w) = ?$$

$$\Pr(G = w \mid X = w) = \frac{\Pr(X = w \mid G = w)\Pr(G = w)}{\Pr(X = w)}$$

$$= \frac{\Pr(X = w \mid G = w)\Pr(G = w)}{\sum_{g \in \mathrm{dom}(G)} \Pr(X = w, G = g)}$$

$$= \frac{\Pr(X = w \mid G = w)\Pr(G = w)}{\sum_{g \in \mathrm{dom}(G)} \Pr(X = w \mid G = g)\Pr(G = g)}$$

$$= \frac{0.75 \times 0.33}{0.75 \times 0.33 + 0.20 \times 0.67}$$



icons by Icons8

Expected Value

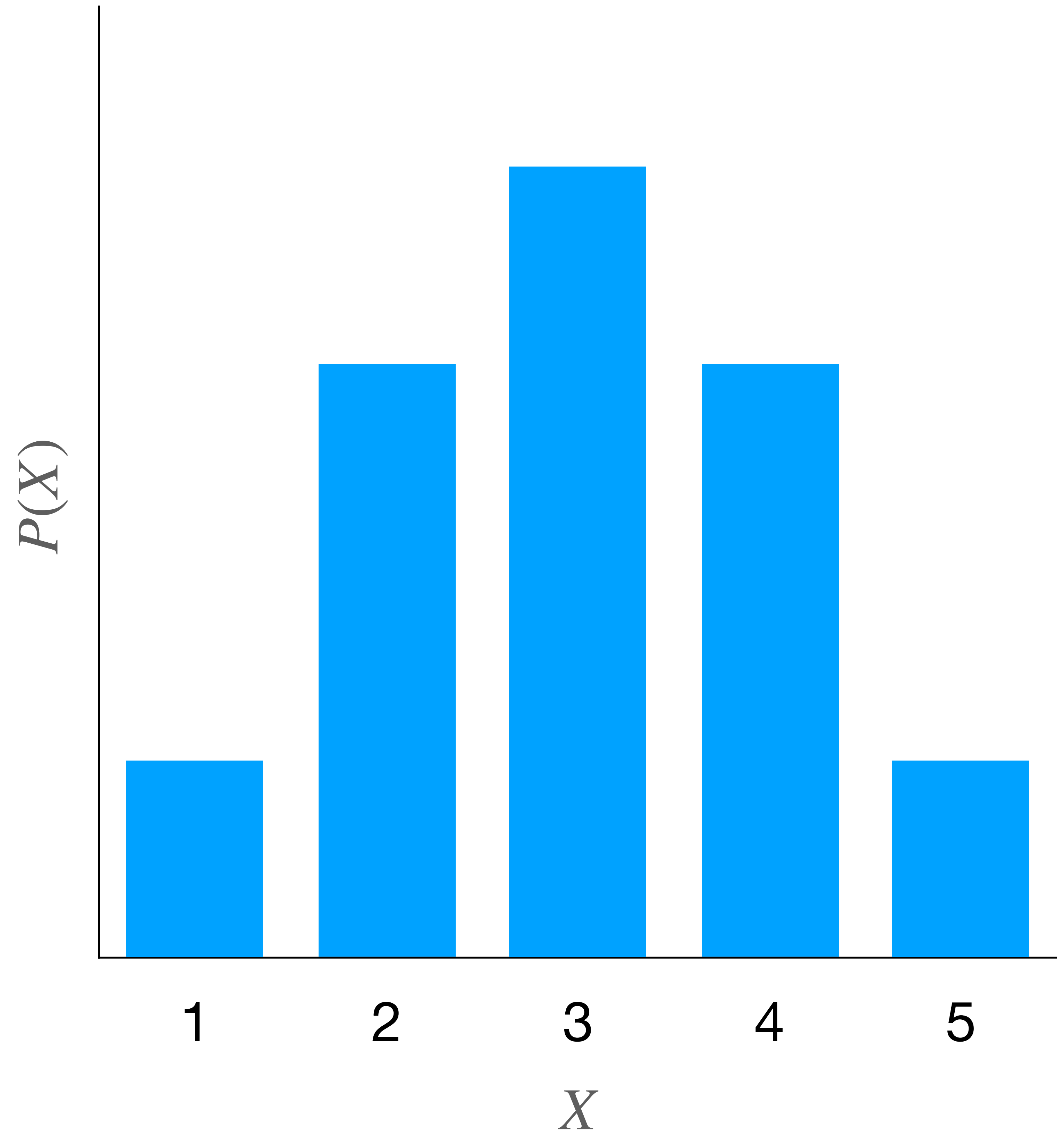
- The **expected value** of a **function** f on a random variable is the weighted **average** of that function over the domain of the random variable, **weighted** by the **probability** of each value:

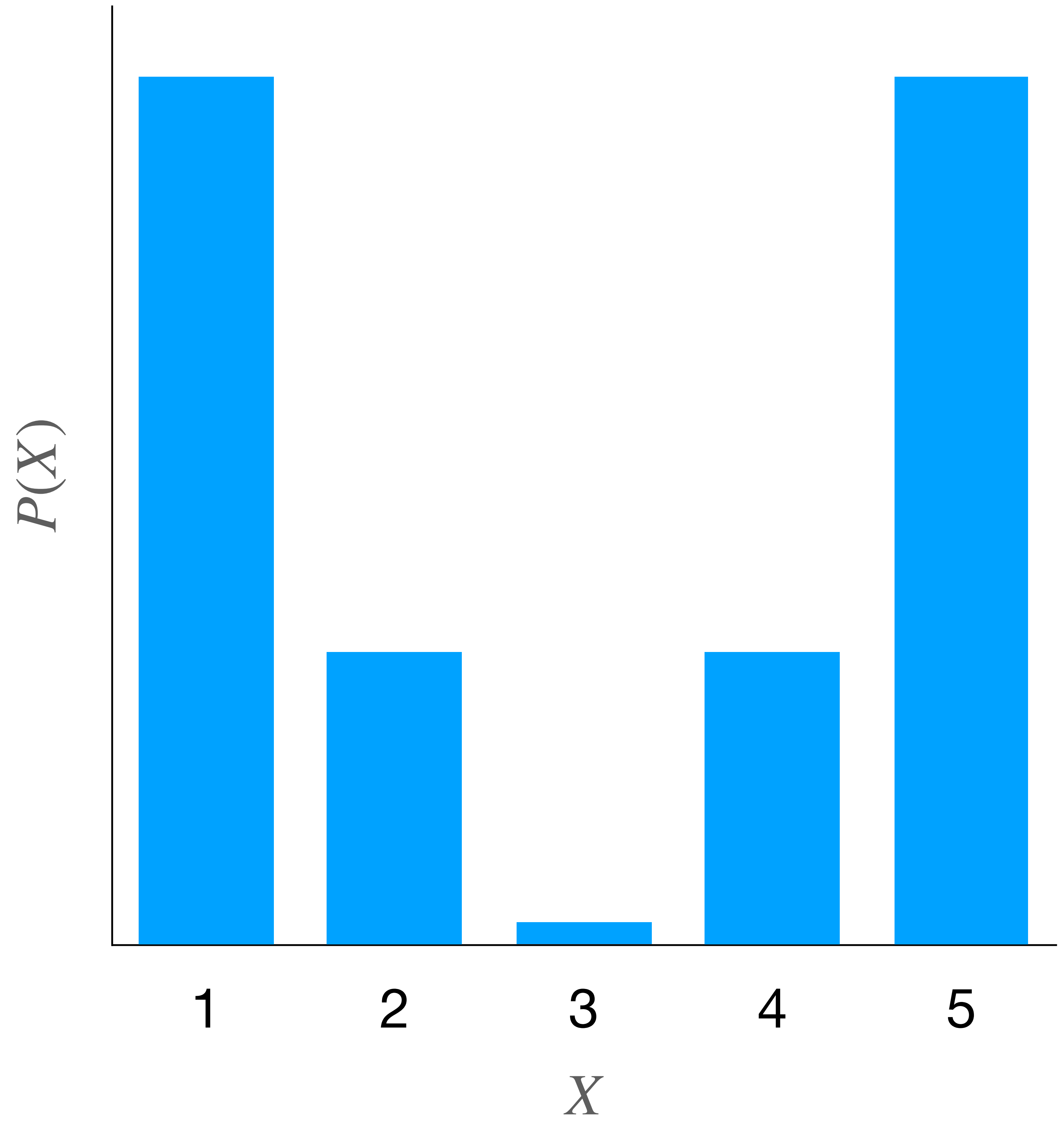
$$\mathbb{E}\left[f(X)\right] = \sum_{x \in dom(X)} P(X = x) f(x)$$

- The **conditional expected value** of a **function** f is the average value of the function over the domain, weighted by the **conditional probability** of each value:

$$\mathbb{E}\left[f(X) \mid Y = y\right] = \sum_{x \in dom(X)} P(X = x \mid Y = y) f(x)$$

Expected Value Examples





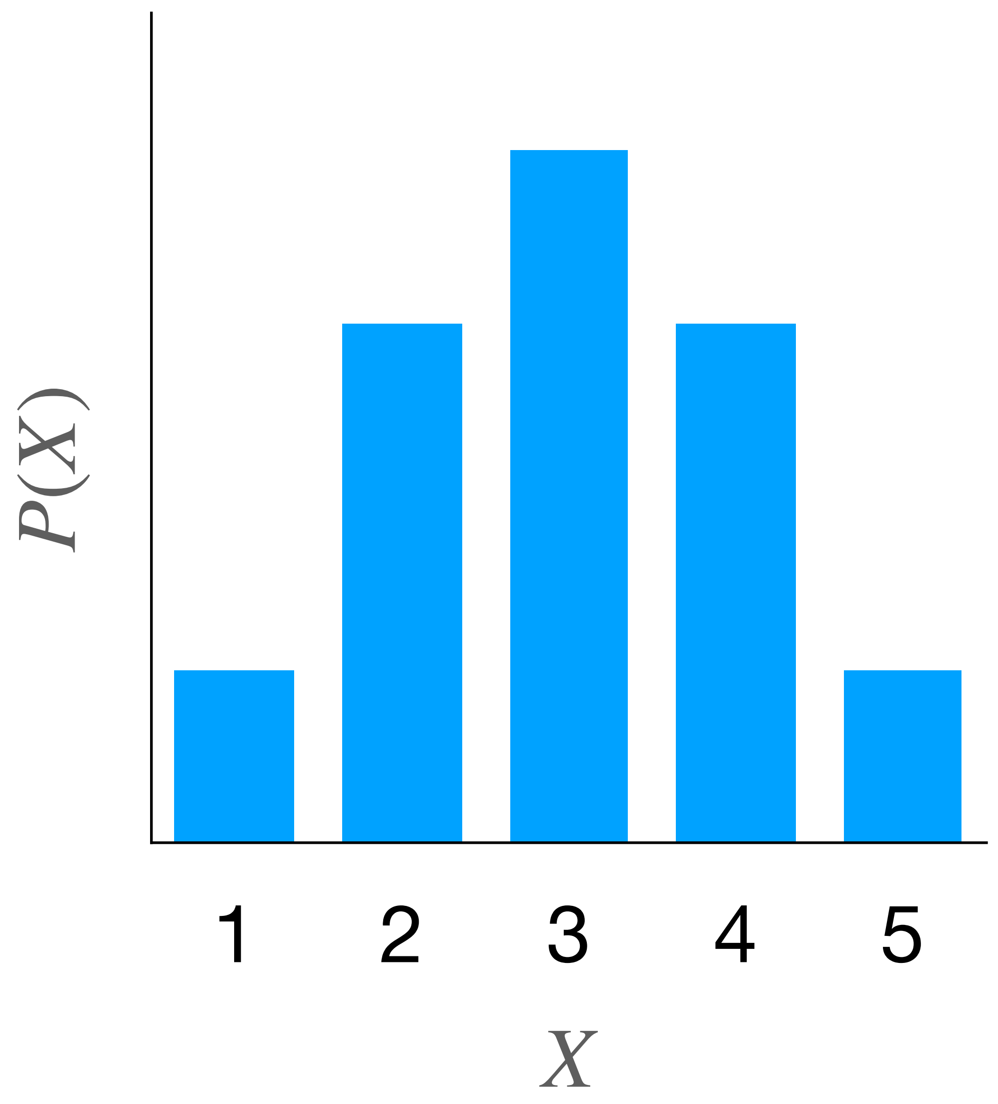
$$\mathbb{E}[X] = 3$$

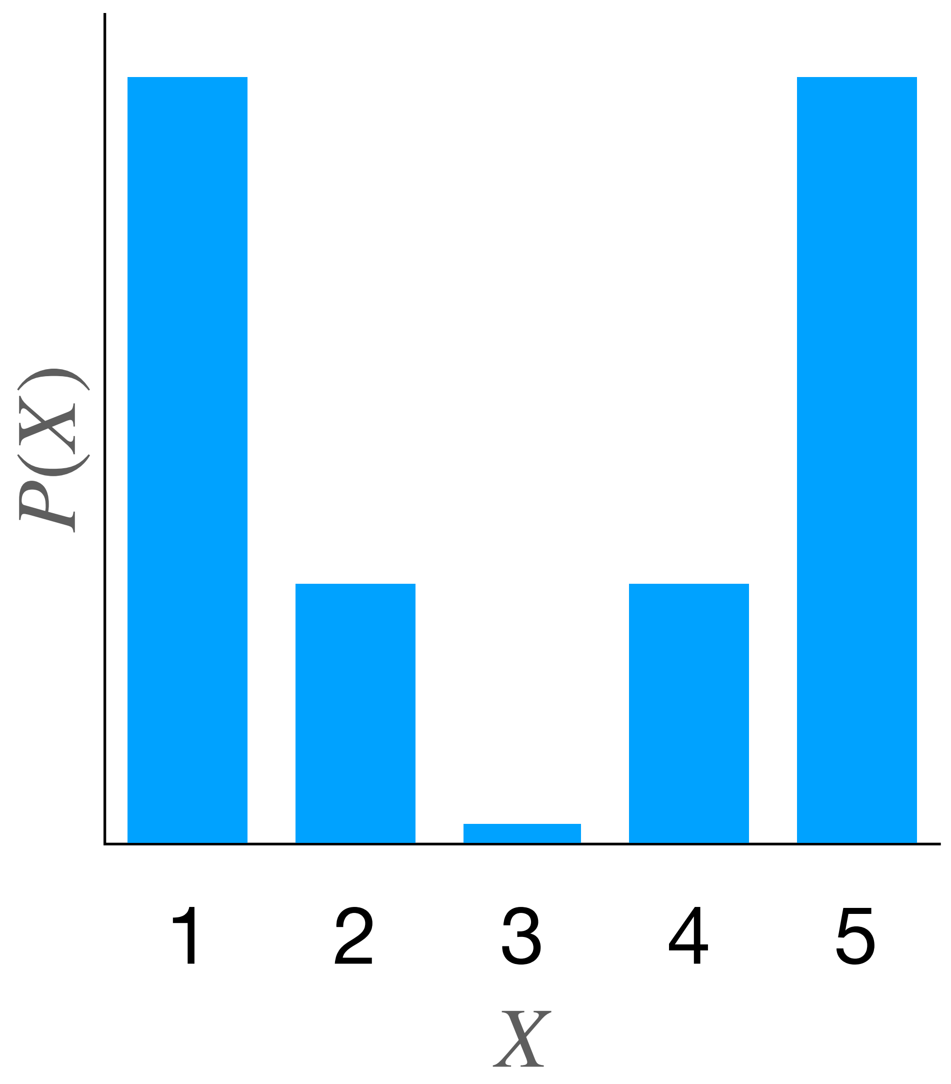
$$\mathbb{E}[X^2] \simeq 10$$

$$\mathbb{E}[X] = 3$$

$$\mathbb{E}[X^2] \simeq 12$$

Expected Value Examples





$$\begin{aligned}\mathbb{E}[X] &= \sum_{x \in \mathrm{dom}(X)} x \Pr[X = x] \\ &= \frac{1}{12}1 + \frac{3}{12}2 + \frac{4}{12}3 + \frac{3}{12}4 + \frac{1}{12}5 \\ &= 3\end{aligned}$$

$$\begin{aligned}\mathbb{E}[X] &= \sum_{x \in \mathrm{dom}(X)} x \Pr[X = x] \\ &= (0.37)1 + (0.125)2 + (0.01)3 + (0.125)4 + (0.37)5 \\ &= 3\end{aligned}$$

Summary

- **Probability** is a **numerical** measure of **uncertainty**
- Formal semantics:
  - Weights over **possible worlds** sum to 1
  - Probability of a proposition is **total weight** of **possible worlds** in which that proposition is **true**
- **Conditional probability** updates beliefs based on **evidence**
- **Expected value** of a function is its **probability-weighted average** over possible worlds