Probability Theory

CMPUT 366: Intelligent Systems

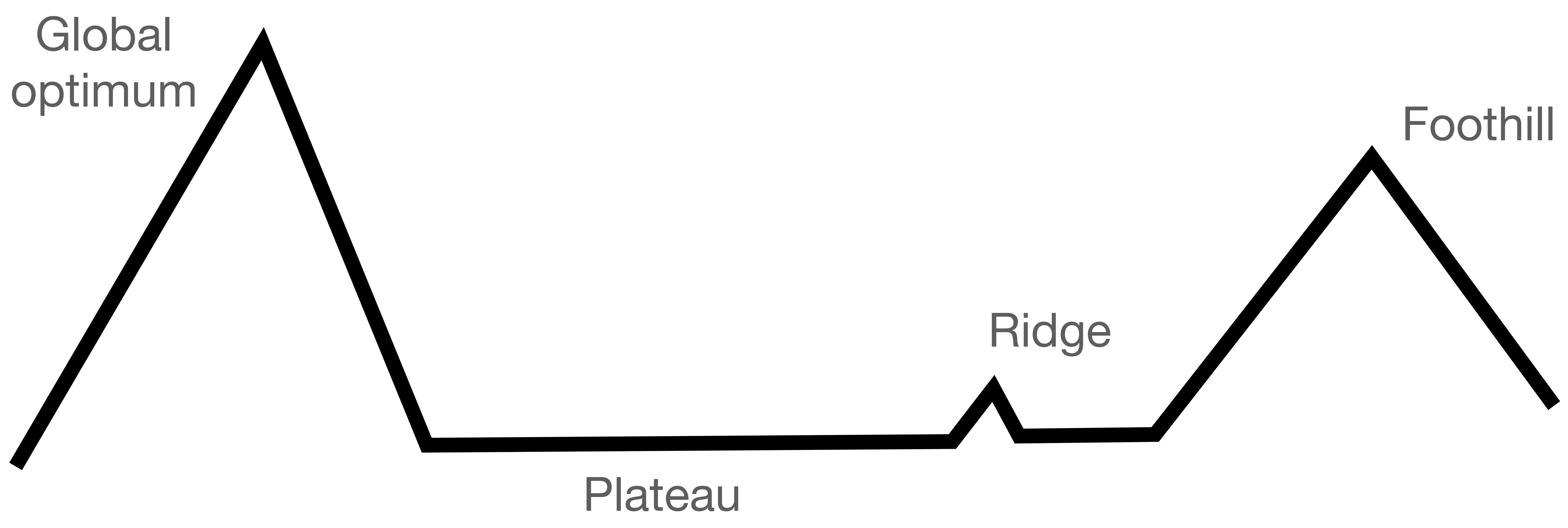
P&M §8.1

Logistics & Assignment #1

- **Midterm** is **March 15** (see eClass for other important dates)
- **Assignment #1 was released on Monday**
  See eClass
  - Due **February 8** at 11:55pm
- Office hours have begun!
  - Not mandatory; for getting help from TAs
  - New Monday office hours: **6:00-7:00pm** Mountain time
  - Python refreshers **TODAY**, Monday

Recap:
Hill Climbing Problems

1. **Foothills:** **Local maxima** that are not global maxima
2. **Plateaus:** Regions of the state space where the **score is uninformative**
3. **Ridges:** Foothills that would not be foothills with a **larger neighbourhood**
4. **Ignorance of the global optimum:** Unless we reach a satisfying assignment, we cannot be sure that an optimum returned by local search is the **global optimum**.



Recap:
Randomized Algorithms

- Adding **random moves** can fix some hill climbing problems
- Two main kinds of random move:
    1. **Random restart:** Start searching from a **completely** random new location
    2. **Random step:** Choose a random **neighbour**
- **Stochastic random search**: Add both kinds of random moves to hill climbing

Recap:
Stochastic Local Search

Input: a constraint satisfaction problem; a *neighbours* function; a *score* function to maximize; a *stop_walk* criterion; a *random_step* criterion

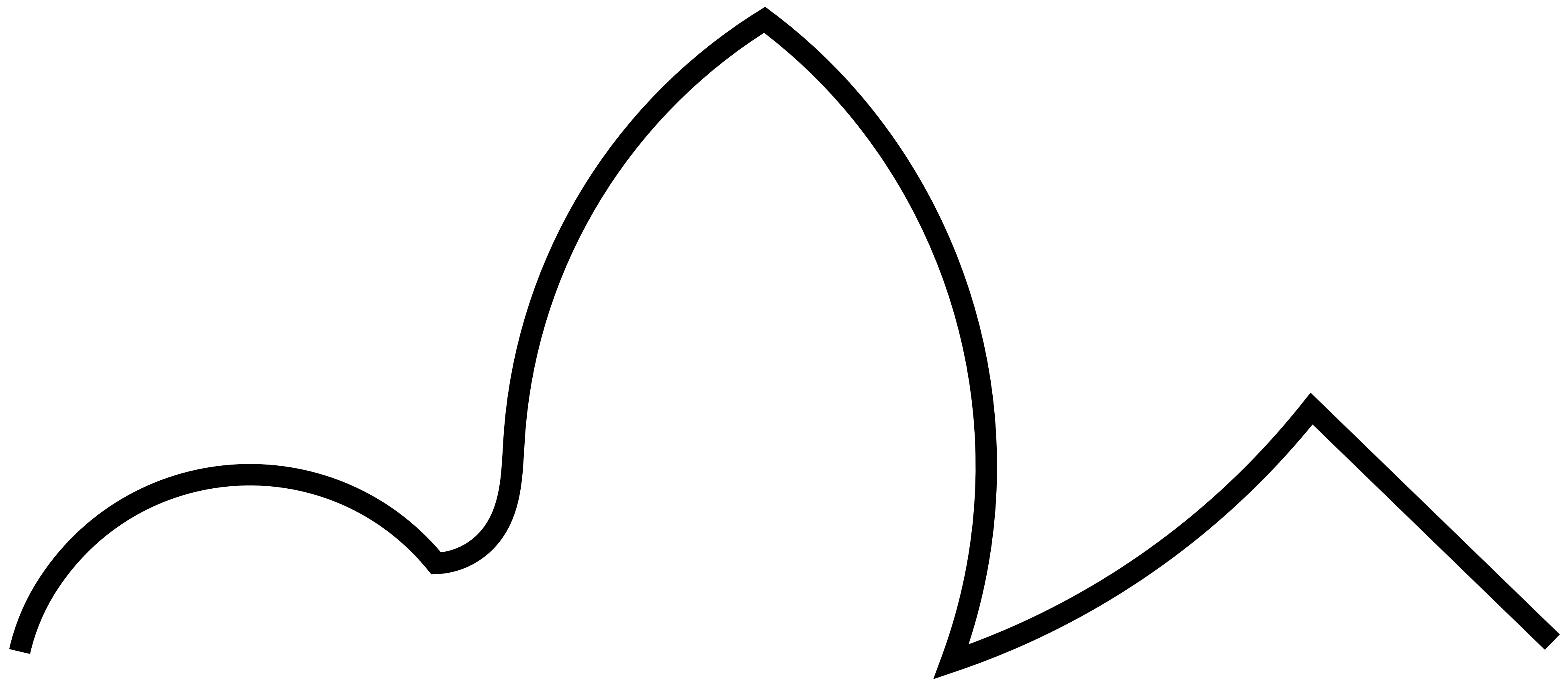
```
current := random assignment of values to variables
incumbent := current
repeat
    if incumbent is a satisfying assignment:
        return incumbent
    if stop_walk():
        current := new random assignment of values to variables
    else if random_step():
        current := a random element from neighbours(current)
    else:
        current := n from neighbours(current) with maximum score(n)
    if score(current) > score(incumbent):
        incumbent := current
```

Two Examples

- Consider two partial algorithms:

  1. Hill climbing plus **random restart**
  2. Hill climbing plus **random steps**

- **Question:** Which finds the maximum most easily on each of these two search spaces? Why?



Simulated Annealing

- **Idea:** **Start out** by searching pretty randomly, but become **more directed**
  - **Intuition:** Move to a good neighbourhood quickly, then search intensively in that neighbourhood
- Maintain a "**temperature**" T
- Choose new nodes more randomly at higher temperatures;
  Gradually decrease the temperature (according to a **cooling schedule**)
- At each step:
  1. Randomly choose a neighbour new
  2. **Always accept** (i.e., assign to $current$) if $score(new) > score(current)$
  3. Else, accept with **probability** $e^{[(score(new) - score(current))/T]}$

Simulated Annealing cont.

$$e^{[(score(new)-score(current))/T]}$$

- Worse $score(new)$ means lower acceptance probability
- Always negative (**why?**)

- Higher T makes negative value smaller
- Higher acceptance probability

- Small neighbourhoods are good, because they are easier to search
- Large neighbourhoods are good, because they are more likely to contain an improvement
- Simulated annealing allows for a large neighbourhood and efficient searching
    - You don't have to generate the whole neighbourhood, just randomly construct a single neighbour

Local Search Summary

- For some problems, we only care about finding a **goal node**, not the actions we took to find it
- **Local search:** Look for goal states by iteratively moving from a **current state** to a **neighbouring state**
  - **Hill climbing:** Always move to the **highest-score** neighbour
  - **Random step:** Sometimes choose a **random** neighbour
  - **Random restart:** Sometimes start again from an **entirely random** state
  - **Simulated annealing:** Random moves start very **random**, become more **greedy** over time

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**
- **Local search** finds a goal node by repeatedly making **small changes** to the current state
  - **Random steps** and **random restarts** help handle **local optima**, **completeness**

Lecture Outline

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

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 thinks that an accident **will** happen, it will just stay home
- If the agent thinks 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 $P(X = 5)$?

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

- A **probability measure** is a function $P : \Omega \to \mathbb{R}$ 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
- We compute the marginal distribution by summing out the other variables:

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

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

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

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 should I **now** believe about the weather?

1. **Rule out** incompatible worlds
2. **Normalize** remaining probabilities

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

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

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: $P(e)$

Expected Value

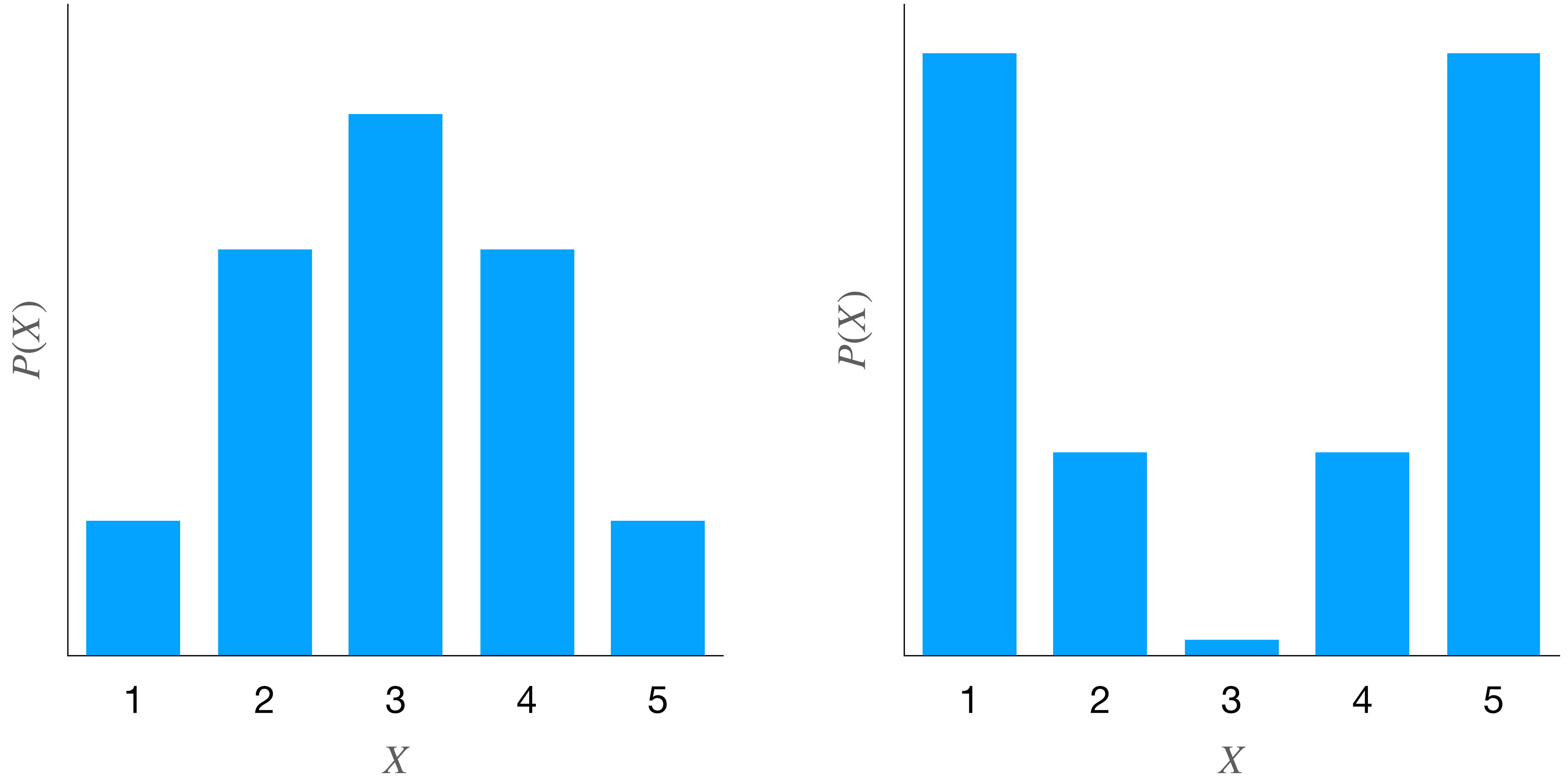
- The **expected value** of a **function** f on a random variable is the weighted **average** of that function over the domain of the random variable, **weighted** by the **probability** of each value:

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

- 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



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

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

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

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

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