Probability Theory Review

Reinforcement Learning (INF8250AE)
Fall 2025

Polytechnique Montréal

Motivation

Uncertainty arises through:

- Noisy measurements
- Variability between samples
- Finite size of data sets

Probability provides a consistent framework for the quantification and manipulation of uncertainty.

Sample Space

Sample space Ω is the set of all possible outcomes of an experiment.

Observations $\omega \in \Omega$ are points in the space also called sample outcomes, realizations, or elements.

Events $E \subset \Omega$ are subsets of the sample space.

In this experiment we flip a coin twice:

Sample space All outcomes $\Omega = \{HH, HT, TH, TT\}$

Observation $\omega = HT$ valid sample since $\omega \in \Omega$

Event Both flips same $E = \{HH, TT\}$ valid event since $E \subset \Omega$

Probability

The probability of an event E, $P(E)$, satisfies three axioms:

1: $P(E) \geq 0$ for every E

2: $P(\Omega) = 1$

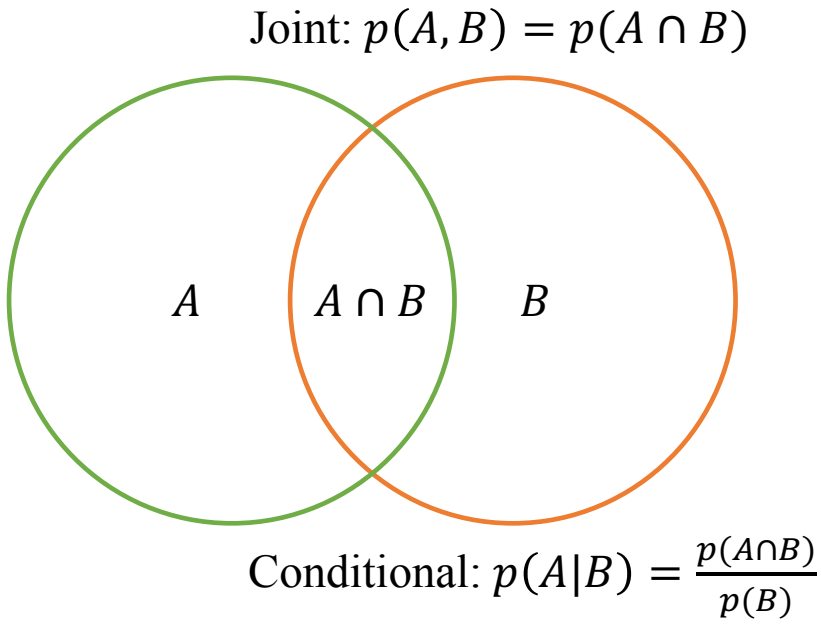
3: If $E_1, E_2, \ldots$ are disjoint then

$$P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$$

Joint and Conditional Probabilities

Joint Probability of A and B is denoted $P(A, B)$.

Conditional Probability of A given B is denoted $P(A|B)$.



$$p(A, B) = p(A|B)p(B) = p(B|A)p(A)$$

Conditional Example

Probability of passing the midterm is 60% and probability of passing both the final and the midterm is 45%.
What is the probability of passing the final given the student passed the midterm?

$$\begin{aligned} P(F|M) &= P(M,F)/P(M) \\ &= 0.45/0.60 \\ &= 0.75 \end{aligned}$$

Independence

Events A and B are independent if $P(A, B) = P(A)P(B)$.

- Indepentent: A: first toss is HEAD; B: second toss is HEAD;

$$P(A, B) = 0.5 * 0.5 = P(A)P(B)$$

- Not Indepentent: A: first toss is HEAD; B: first toss is HEAD;

$$P(A, B) = 0.5 \neq P(A)P(B)$$

Independence

Events A and B are conditionally independent given C if

$$P(A, B|C) = P(B|C)P(A|C)$$

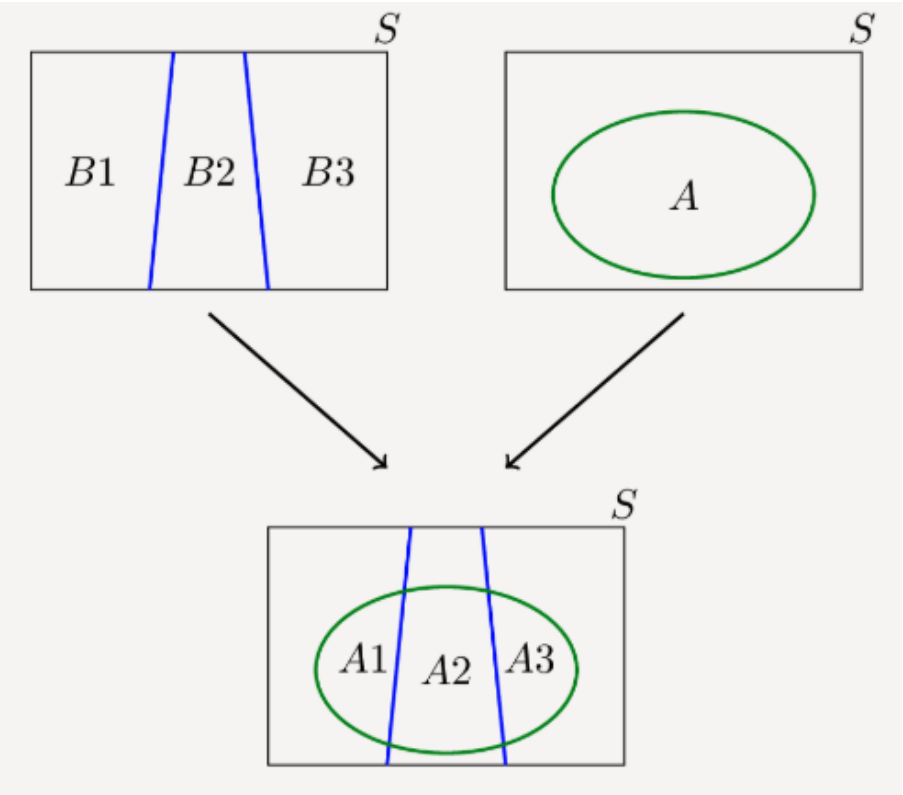
Consider two coins [1]: A regular coin and a coin which always outputs HEAD or always outputs TAIL.
A=The first toss is HEAD; B=The second toss is HEAD;
C=The regular coin is used. D=The other coin is used.
Then A and B are conditionally independent given C, but A and B are NOT conditionally independent given D.

[1] `www.probabilitycourse.com/chapter1/1_4_4_conditional_independence.php`

Marginalization and Law of Total Probability

Law of Total Probability [2]

$$P(A) = \sum_i P(A, B_i) = \sum_i P(A|B_i)P(B_i)$$



[2] www.probabilitycourse.com/chapter1/1_4_2_total_probability.php

Bayes' Rule

Bayes' Rule:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)}$$

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

Bayes' Example

Suppose you have tested positive for a disease. What is the probability you actually have the disease?
This depends on the prior probability of the disease:

- $P(T = 1|D = 1) = 0.95$ (likelihood)
- $P(T = 1|D = 0) = 0.10$ (likelihood)
- $P(D = 1) = 0.1$ (prior)

So $P(D = 1|T = 1) =?$

Bayes' Example

Suppose you have tested positive for a disease. What is the probability you actually have the disease?

$$P(T = 1|D = 1) = 0.95 \text{ (true positive)}$$

$$P(T = 1|D = 0) = 0.10 \text{ (false positive)}$$

$$P(D = 1) = 0.1 \text{ (prior)}$$

So $P(D = 1|T = 1) =?$

Use Bayes' Rule:

$$P(D = 1|T = 1) = \frac{P(T = 1|D = 1)P(D = 1)}{P(T = 1)} = \frac{0.95 \times 0.1}{P(T = 1)} = 0.51$$

$$P(T = 1) = P(T = 1|D = 1)P(D = 1) + P(T = 1|D = 0)P(D = 0)$$

$$= 0.95 \times 0.1 + 0.1 \times 0.90 = 0.185$$

Chain Rule of Probability

For a set of random variables $X_1, X_2, \ldots, X_n$:

$P(X_1, X_2, \ldots, X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)\ldots P(X_n|X_1, \ldots, X_{n-1})$*Implic*

Markov Property: A process satisfies the Markov property if the future depends only on the present:

$$P(X_{t+1}|X_t, X_{t-1}, \ldots, X_0) = P(X_{t+1}|X_t)$$

Example in RL: joint probability of a trajectory $\tau = (x_0, a_0, x_1, a_1, \ldots, x_T)$

$$P(\tau) = P(x_0) \prod_{t=0}^{T-1} \pi(a_t|x_t)P(x_{t+1}|x_t, a_t)$$

Random Variable

How do we connect sample spaces and events to data?
A random variable is a mapping which assigns a real number $X(\omega)$ to each observed outcome $\omega \in \Omega$

For example, let's flip a coin 10 times. $X(\omega)$ counts the number of Heads we observe in our sequence. If $\omega = HHTHTHHTHT$ then $X(\omega) = 6$.

Discrete and Continuous Random Variables

Discrete Random Variables

- Takes countably many values, e.g., number of heads
- Distribution defined by probability mass function (PMF)
- Marginalization: $p(x) = \sum_y p(x, y)$

Continuous Random Variables

- Takes uncountably many values, e.g., time to complete task
- Distribution defined by probability density function (PDF)
- Marginalization: $p(x) = \int_y p(x, y) \mathrm{d}y$

I.I.D.

Random variables are said to be independent and identically distributed (i.i.d.) if they are sampled from the same probability distribution and are mutually independent.
This is a common assumption for observations. For example, coin flips are assumed to be iid.

Probability Distribution Statistics

Mean: First Moment, μ

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} x_i p(x_i) \qquad \text{(univariate discrete r.v.)}$$

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x p(x) dx \qquad \text{(univariate continuous r.v.)}$$

Variance: Second (central) Moment, σ^2

$$\begin{aligned} \text{Var}[X] &= \int_{-\infty}^{\infty} (x-\mu)^2 p(x) \mathrm{d}x \\ &= \mathbb{E}\left[(X-\mu)^2\right] \\ &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \end{aligned}$$

It is common to use capital letters such as X to denote a random variable drawn from a distribution $p(x)$. That is why we wrote $\mathbb{E}[X]$ instead of $\mathbb{E}[x]$, but the latter may also be used sometimes. We may go back and forth between these two.

Conditional Expectation: Intuition

Idea: The conditional expectation is the "best prediction" of a random variable Y given information about another variable X.

$$\mathbb{E}[Y \mid X = x] = \text{expected value of } Y, \text{ knowing that } X \text{ is } x.$$

Example:
- Imagine you want to predict tomorrow's temperature (Y).
- Without info: just take the average temperature.
- With info about the season (X): your prediction should change depending on X.

Note:
$\mathbb{E}[Y \mid X = x]$ is a *function of* x that gives the expected value of Y once random variable X revealed to be x.

Conditional Expectation: Formal Definition

For discrete variables:

$$\mathbb{E}\left[Y \mid X = x\right] = \sum_{y} y \Pr(Y = y \mid X = x)$$

For continuous variables:

$$\mathbb{E}\left[Y \mid X = x\right] = \int_{-\infty}^{\infty} y \, f_{Y|X}(y \mid x) \, dy$$

Conditional Expectation: Examples

Example 1: Dice

- Y = sum of two dice, X = value of the first die.
- Without info: $\mathbb{E}[Y] = 7$.
- If $X = 4$: $\mathbb{E}[Y \mid X = 4] = 4 + 3.5 = 7.5$.

Example 2: Students and Grades

- Y = final grade, X = hours studied.
- $\mathbb{E}[Y \mid X]$ = average grade for students who studied X hours.

Example 3: Value function in RL

Value function is a conditional expectation:

$$V^{\pi}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, s_0 = s\right].$$

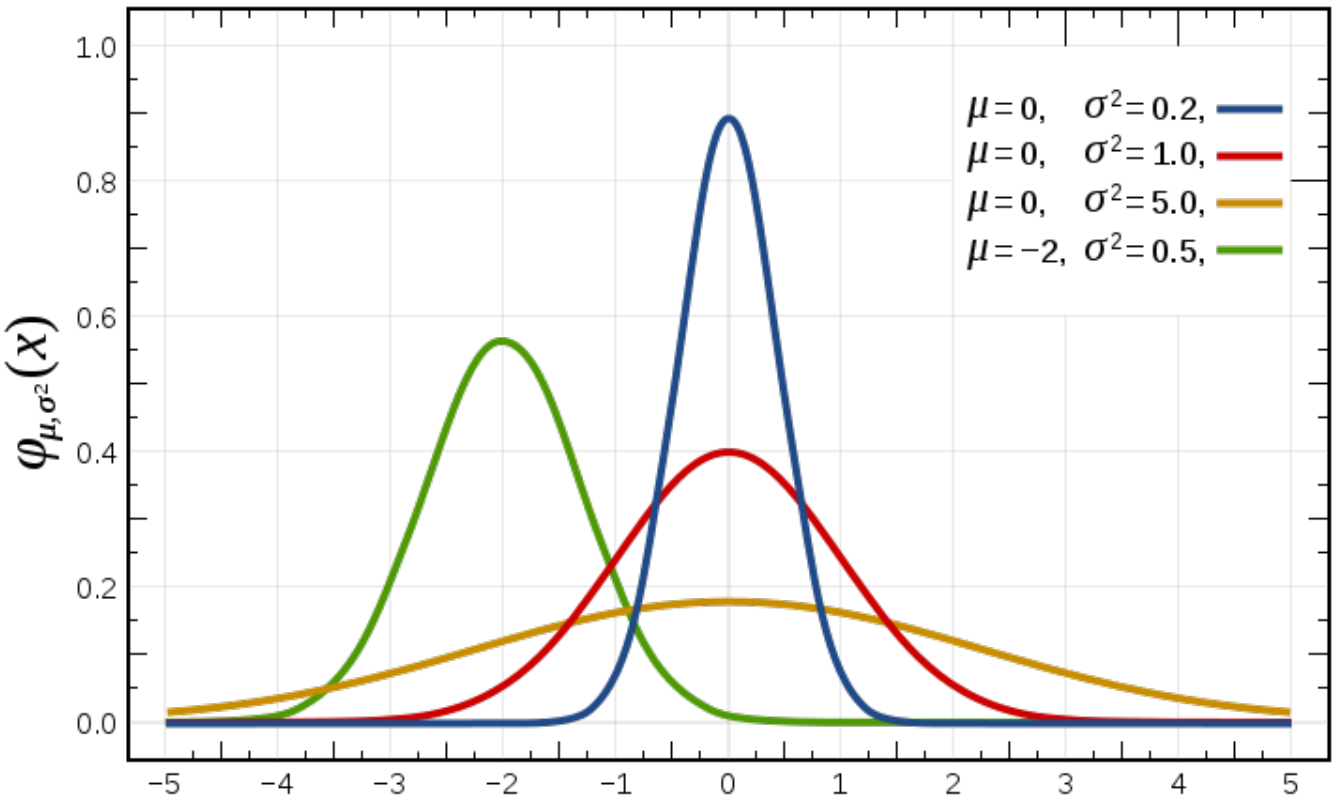
Discrete and Continuous Distributions

- Multinomial/Discrete: $P(X = k) = p_k$, used in categorical action policies.
- Gaussian: $X \sim \mathcal{N}(\mu, \sigma^2)$, used for continuous actions in policy gradient methods.
- Boltzmann / Softmax: $\pi(a|s) = \frac{\exp(Q(s,a)/\tau)}{\sum_b \exp(Q(s,b)/\tau)}$, used in discrete action exploration.

Univariate Gaussian Distribution

Also known as the Normal Distribution, $\mathcal{N}(\mu, \sigma^2)$

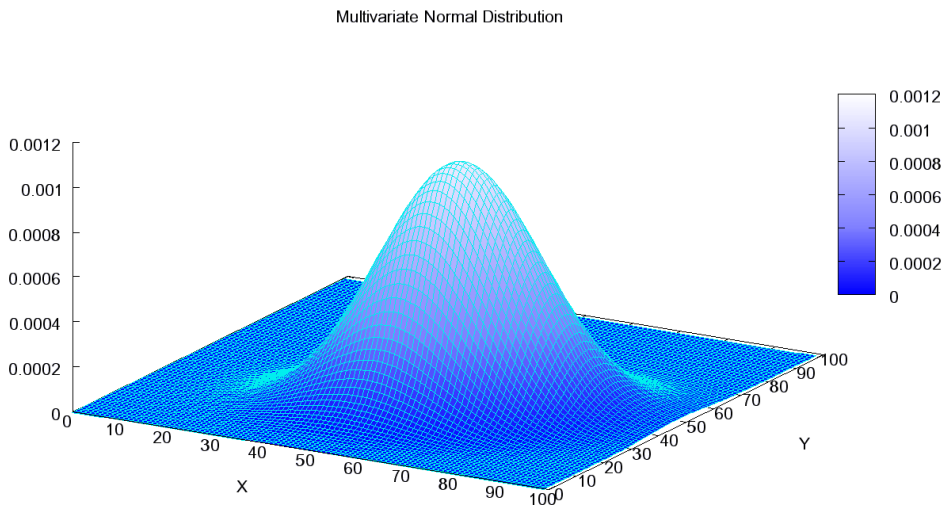
$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



Multivariate Gaussian Distribution

Multidimensional generalization of the Gaussian.
$\mathbf{x}$ is a D-dimensional vector
μ is a D-dimensional mean vector
Σ is a $D \times D$ covariance matrix with determinant $|\Sigma|$

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



Covariance Matrix

Recall that $\mathbf{x}$ and μ are D-dimensional vectors
Covariance matrix Σ is a matrix whose (i, j) entry is the covariance

$$\begin{aligned}\Sigma_{ij} &= \mathbf{Cov}\left(\mathbf{X}_i, \mathbf{X}_j\right) \\ &= \mathbb{E}\left[(\mathbf{X}_i - \mu_i)(\mathbf{X}_j - \mu_j)\right] \\ &= \mathbb{E}\left[\mathbf{X}_i \mathbf{X}_j\right] - \mu_i \mu_j.\end{aligned}$$

Notice that the diagonal entries are the variance of each elements.
The covariant matrix has the property that it is symmetric and positive-semidefinite (this is useful for whitening).

Inferring Parameters

We have data X and we assume it is sampled from some distribution. How do we figure out the parameters that "best" fit that distribution?

Maximum Likelihood Estimation (MLE)

$$\hat{\theta}_{MLE} = \underset{\theta}{\text{argmax}} P(X|\theta)$$

Maximum A posteriori Probability (MAP)

$$\hat{\theta}_{MAP} = \underset{\theta}{\text{argmax}} P(\theta|X)$$

Note:

- **MLE:** No prior needed, objective and data-driven, but can overfit with limited data.
- **MAP:** Incorporates prior knowledge, often more robust with small datasets, but depends on the choice of prior.

MLE for Univariate Gaussian Distribution

We are trying to infer the parameters mean μ and variance σ^2 of a univariate Gaussian Distribution:

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(x-\mu)^2).$$

The likelihood that our observations $X_1, \ldots, X_N$ were generated by a univariate Gaussian with parameters μ and σ^2 is

$$\text{Likelihood} = p(X_1, \ldots, X_N|\mu,\sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(X_i-\mu)^2).$$

MLE for Univariate Gaussian Distribution

For MLE we want to maximize this likelihood, which is difficult because it is represented by a product of terms

$$\text{Likelihood} = p(X_1, \ldots, X_N | \mu, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(X_i - \mu)^2)$$

So we take the log of the likelihood so the product becomes a sum

$$\begin{aligned} \text{Log Likelihood} &= \log p(X_1, \ldots, X_N | \mu, \sigma^2) \\ &= \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(X_i - \mu)^2). \end{aligned}$$

Since log is monotonically increasing, their maximizers are the same, i.e. $\underset{\theta}{\text{argmax}}\{L(\theta)\} = \underset{\theta}{\text{argmax}}\{\log L(\theta)\}$.

MLE for Univariate Gaussian Distribution

The log Likelihood simplifies to

$$\mathcal{L}(\mu, \sigma) = \sum_{i=1}^{N} \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(X_i - \mu)^2) \right]$$

$$= -\frac{1}{2} N \log(2\pi\sigma^2) - \sum_{i=1}^{N} \frac{(X_i - \mu)^2}{2\sigma^2}$$

Which we want to maximize. How?

MLE for Univariate Gaussian Distribution

To maximize we take the derivatives, set equal to 0, and solve:

$$\mathcal{L}(\mu, \sigma) = -\frac{1}{2} N \log(2\pi\sigma^2) - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2}$$

Derivative w.r.t. μ, set equal to 0, and solve for $\hat{\mu}$

$$\frac{\partial \mathcal{L}(\mu, \sigma)}{\partial \mu} = 0 \implies \hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} X_i.$$

Therefore the $\hat{\mu}$ that maximizes the likelihood is the average of the data points, which is called the sample average or empirical expectation too.
Derivative w.r.t. σ^2, set equal to 0, and solve for $\hat{\sigma}^2$

$$\frac{\partial \mathcal{L}(\mu, \sigma)}{\partial \sigma^2} = 0 \implies \hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (X_i - \hat{\mu})^2.$$

Trick: Monte Carlo Estimation of Expectations

- Goal: estimate $\mathbb{E}[f(X)]$ when the distribution $p(x)$ is complex or unknown
- Monte Carlo estimate using N samples $x_1, \ldots, x_N \sim p(x)$:

$$\mathbb{E}[f(X)] \approx \frac{1}{N} \sum_{i=1}^{N} f(x_i)$$

- Useful in RL for:
  - Estimating expected returns
  - Policy evaluation when dynamics are unknown
- Trick: more samples reduce variance, but cost increases

Variance Reduction in Monte Carlo

- Monte Carlo estimates can have high variance.
- Common variance reduction techniques:
  - Baseline / Control Variates: subtract a baseline b from the reward
    $$\nabla_\theta \mathbb{E}[R] = \mathbb{E}[(R - b)\nabla_\theta \log p_\theta(x)]$$
    reduces variance without changing expectation.
  - Antithetic Sampling: use negatively correlated samples to cancel variance.
  - Importance Sampling: reweight samples from another distribution to reduce variance.
- Widely used in policy gradient RL methods.

Trick: Log-Derivative (Score Function) Trick

- Goal: gradient of expectation when reparametrization not possible

$$\nabla_\theta \mathbb{E}\left[f(X)\right] = \mathbb{E}\left[f(X)\nabla_\theta \log p_\theta(X)\right]$$

- Advantage: avoids differentiating $f(X)$ through stochastic X

Trick: Sampling from Multinomial / Categorical Distribution

- Suppose probabilities $p_1, \ldots, p_k$
- Sample $U \sim \text{Uniform}(0, 1)$
- Compute cumulative distribution $c_j = \sum_{i=1}^{j} p_i$
- Return smallest j such that $U < c_j$
- RL Example: sample discrete action from policy

Trick: Sampling from Continuous Distributions via Inverse CDF

- Given CDF $F_X(x)$ of X, sample $U \sim \text{Uniform}(0, 1)$
- Transform: $X = F_X^{-1}(U)$
- Works for arbitrary continuous distributions
- RL Example: sample continuous action from custom policy

Trick: Reparametrization Trick in Gaussian Policy

- Policy: $a \sim \pi_\theta(a|s) = \mathcal{N}(\mu_\theta(s), \sigma_\theta(s)^2)$
- Standard sampling: $a \sim \mathcal{N}(\mu_\theta(s), \sigma_\theta(s)^2)$
- Reparametrization trick:

$$a = \mu_\theta(s) + \sigma_\theta(s)\epsilon, \quad \epsilon \sim \mathcal{N}(0, 1)$$

- Allows gradients to flow through $\mu_\theta(s)$ and $\sigma_\theta(s)$
- Reduces variance in gradient estimates

Trick: Numerically Stable Softmax Trick

- Standard softmax:

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}}$$

- Problem: if z_i are large, e^{z_i} can overflow.
- Trick: subtract the maximum value from all logits:

$$\text{softmax}(z_i) = \frac{e^{z_i - \max_j z_j}}{\sum_j e^{z_j - \max_j z_j}}$$

- Numerically stable and equivalent mathematically.
- Important in RL for computing action probabilities in policies.

Log-Sum-Exp Trick

- Problem: computing $\log \sum_i e^{z_i}$ can overflow or underflow.
- Trick: factor out the max

$$\log \sum_i e^{z_i} = \max_i z_i + \log \sum_i e^{z_i - \max_i z_i}$$

- Numerically stable and widely used in RL (e.g., softmax policy, partition functions).

Running / Exponential Moving Average

- Keep a running estimate of mean / variance for normalization:

$$\mu_t = (1-\alpha)\mu_{t-1} + \alpha x_t, \quad \sigma_t^2 = (1-\alpha)\sigma_{t-1}^2 + \alpha(x_t - \mu_t)^2$$

- Useful for:
  - Normalizing observations or rewards in RL
  - Reducing variance in learning

Quantifying Information

Intuition behind information content:

- **Rare events carry more information:** An unlikely event teaches us more than a common one.
- **Additivity for independent events:** If two independent events happen, total information should sum:

$$I(A \text{ and } B) = I(A) + I(B)$$

- **Mathematical form:** We want a function $I(p)$ such that:

$$I(p_1 \cdot p_2) = I(p_1) + I(p_2)$$

$$\Rightarrow I(p) = -\log(p)$$

- **Why negative?** Information content is proportional to the inverse of probability.

Example: Observing Through a Window

Imagine someone looks out a window and reports what passes by:

- "A person walks by." Common: $p = 0.8$

$$I = -\log_2 0.8 \approx 0.32 \text{ bits}$$

- "A T-Rex walks by!" Rare: $p = 0.00001$

$$I = -\log_2 0.00001 \approx 16.6 \text{ bits}$$

Conclusion: Rare events carry more information; common events convey little new knowledge.

Entropy: Average Information Content

Definition: Entropy measures the average information of a random variable X with distribution $p(x)$:

$$H(X) = \mathbb{E}[I(X)] = -\sum_x p(x) \log p(x) \quad \text{(discrete)}$$

- High entropy: distribution is spread out, more uncertainty, more information on average.
- Low entropy: distribution is concentrated, less uncertainty, less information on average.
- Continuous case:

$$H(X) = -\int p(x) \log p(x)\, dx$$

- Entropy is maximized for the uniform distribution.

Intuition: Think of entropy as the "expected surprise" when observing a random event.

Practical Usage: Entropy Regularization in RL

Idea: Add policy entropy to the reward to encourage exploration:

$$\tilde{R}_t = R_t + \beta H(\pi(\cdot|s_t))$$

- $H(\pi(\cdot|s)) = -\sum_a \pi(a|s) \log \pi(a|s)$
- Prevents premature convergence to deterministic policies
- β controls exploration-exploitation tradeoff

KL Divergence: Measuring Difference Between Distributions

Definition: The Kullback-Leibler (KL) divergence measures how one probability distribution Q differs from a reference distribution P:

$$D_{\mathrm{KL}}(P \,\|\, Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)} \quad \text{(discrete)}$$

$$D_{\mathrm{KL}}(P \,\|\, Q) = \int p(x) \log \frac{p(x)}{q(x)} dx \quad \text{(continuous)}$$

- $D_{\mathrm{KL}}(P \,\|\, Q) \geq 0$ (non-negative)
- $D_{\mathrm{KL}}(P \,\|\, Q) = 0$ iff $P = Q$
- Not symmetric: $D_{\mathrm{KL}}(P \,\|\, Q) \neq D_{\mathrm{KL}}(Q \,\|\, P)$

Intuition: KL divergence tells us the extra number of bits needed to encode samples from P if we use a code optimized for Q instead of P. In other words, it calculates how much more surprised we are on average when the outcomes follow P but we assume they follow Q.
In RL: P can be a reference policy (teacher), Q the agent policy. Minimizing KL encourages the agent to behave similarly to the reference, reducing "surprise" in expected actions.

KL Divergence Example: Looking Through the Window

Suppose someone reports what they see through a window. Consider two distributions:

- P: True probabilities of things passing by
  - Person: 0.9
  - Dog: 0.09
  - T-Rex: 0.01
- Q: Our model guesses
  - Person: 0.95
  - Dog: 0.05
  - T-Rex: 0

$$D_{\mathrm{KL}}(P\|Q) = 0.9 \log \frac{0.9}{0.95} + 0.09 \log \frac{0.09}{0.05} + 0.01 \log \frac{0.01}{0} \approx \infty$$

- The KL divergence is huge because Q assigns zero probability to a rare but possible event (T-Rex)
- KL captures the penalty for being surprised by unlikely events

KL Divergence in Reinforcement Learning

Context: In RL, KL divergence is used to measure how much a new policy π_θ deviates from a reference or old policy π_{old}.

$$D_{\text{KL}}(\pi_{\text{old}} \,\|\, \pi_\theta) = \sum_a \pi_{\text{old}}(a|s) \log \frac{\pi_{\text{old}}(a|s)}{\pi_\theta(a|s)}$$

- **Policy Constraints:** Algorithms like PPO and TRPO limit the KL divergence between successive policies to avoid overly large updates.
- **Surprise Interpretation:**
  - High KL means the new policy behaves very differently from the old one $\Rightarrow$ "surprise" about actions.
  - Low KL means small updates, preserving stability.
- **Expected KL:** Often we compute $\mathbb{E}_{s\sim d_\pi}[D_{\text{KL}}(\pi_{\text{old}}(\cdot|s)\|\pi_\theta(\cdot|s))]$ over states visited by the old policy.
- **Usage:**
  - Regularize policy updates: keep new policy close to old.
  - Guide exploration: encourage moderate deviation when needed.