# Topics

1. Pytorch basics: https://colab.research.google.com/github/wecacuee/ECE490-Neural-Networks/blob/master//notebooks/06-pytorch/NumpyTutorial-Pytorched.ipynb
2. Autograd Mathematics: https://colab.research.google.com/github/wecacuee/ECE490-Neural-Networks/blob/master/notebooks/03-autograd/AutogradNumpy.ipynb
3. Probability problems ( below)

# Probability definitions

## Q1: Define Sample Space

Sample space is the set all possible of outcomes of an experiment, denoted by $\Omega$.

For example, For 2-coin tosses the sample space is

$$\Omega_{\text{2-coin}} = \{HH, HT, TH, TT\}$$

For roll of a dice with 6-sides

$$\Omega_{\text{dice}} = \{1, 2, 3, 4, 5, 6\}$$

For weight measurements of an individual, the sample space is the set of all positive real numbers

$$\Omega_{\text{weight}} = \mathbb{R}^{+}$$

## Q2: Define Event Space

An event is the set of outcomes that we might be interested in.

Event space is a set of subsets of the sample space.

or example, For 2-coin tosses the set of all subsets of the sample space in cluding the null set $\{\}$ and the full sample $\Omega$

$$\mathcal{F}_{\text{2-coin}} = \{\{\}, \{HH\}\{HT\}, \{TH\}, \{TT\}, \{HH, HT\}, \ldots, \underbrace{\{HH, HT, TH, TT\}}_{\Omega}\}$$

For weight measurements of an individual, the event space is be the set of all unions and intersections of intervals (open and closed) of sample space (positive real numbers).

$$\mathcal{F}_{\text{weight}} = \{\cup_i \cap_j [a_{ij}, b_{ij}] : a_{ij} < b_{ij}, a_{ij} \in \mathbb{R}, b_{ij} \in \mathbb{R}\}$$

## Q3: Define Power set

The set of all possible subsets of a set $\Omega$ is called a power set and is denoted by $2^{\Omega}$.

For roll of a dice with 6-sides

$$2^{\Omega} = \{\{\}, \{HH\}\{HT\}, \{TH\}, \{TT\}, \{HH, HT\}, \ldots, \underbrace{\{HH, HT, TH, TT\}}_{\Omega}\}$$

For discrete sample space, event space is the power set of the sample space.

## Q4: Define Probability measure

Probability measure is a function $P : \mathcal{F} \to [0, 1]$ that maps from event space to real numbers between $[0, 1]$ and satisfy the following Kolmogorov axioms

1. $P(E) \in [0, 1]$ for all $E \in \mathcal{F}$, where $\mathcal{F}$ is event space
2. $P(\Omega) = 1$, where $\Omega$ is sample space
3. For all disjoint set of events $A_1$, $A_2$ ($A_1 \cap A_2 = \phi$), the probability of union of events is the sum of individual event probabilities:

   $$P(A_1) + P(A_2) = P(A_1 \cup A_2)$$

   when $A_1 \cap A_2 = \phi$.

   In general, for a countably infinite set of event $A_1, A_2, \ldots A_n \ldots \infty$,

   $$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$$

   when $A_i \cap A_j = \infty$ for all $i \neq j$.

## Q5: Define Probability space

The triple of sample space $\Omega$, event space $\mathcal{F}$ and a probability measure $P : \mathcal{F} \to [0, 1]$ is called a probability space.

## Q6: Define Random variable

A random variable is a function $X : \Omega \to \mathbb{Q}$ that maps from sample space $\Omega$ to a space of integers $\mathbb{Z}$ or real numbers $\mathbb{R}$ (in general a measurable space), such that a preimage $X^{-1}(B) \in \Omega$ of any set of numbers $B \in \mathbb{Q}$ exists in the sample space.

For example, a 2-coin toss:

$$\Omega = \{HH, HT, TH, TT\}$$

A random variable maps the elements of sample space to a number,

$$X(HH) = 0, X(HT) = 1, X(TH) = 2, X(TT) = 3$$

By slight abuse of notation, the random variable also maps events to a set of numbers $X : \mathcal{F} \to B$,

$$X(\{HT, TH, TT\}) = \{1, 2, 3\}$$

## Q7: What is the difference between discrete and continuous random variable

Discrete random variable: When the random variable maps the sample space to integers, then the random variable is discrete.

Continuous random variable: When the random variable maps the sample space to real numbers then the random variable is continuous.

## Q8: Define Probability mass function (PMF)

For a discrete random variable (RV) the Probability mass function (PMF) is a function that assigns probability value to every discrete value of the random variable, such that

$$\sum_{x \in \Omega} P(X = x) = 1.$$

For example, a die roll

$$\Omega = \{1, \ldots, 6\}$$

$$P(X = 1) = 1/6, P(X = 2) = 1/6, \ldots, P(X = 6) = 1/6$$

PMF is denoted as multiple symbols $P(X = x) = P_X(x) = P(x)$

## Q9: Define probability density function (PDF)

For a continuous random variable $X : \Omega \to \mathbb{R}$, the probability density function (PDF) is a function $f_X : \mathbb{R} \to [0, \infty)$ such that:

1. $f_X(x) \geq 0$ for all $x \in \mathbb{R}$
2. $\int_{\mathbb{R}} f_X(x) dx = 1$
3. $P(a \leq X \leq b) = P(X \in [a, b]) = \int_a^b f_X(x) dx$

## Q10: Define joint probability mass function

$$P(X = x, Y = y) = P((X = x) \cap (Y = y)) = P((X = x) \text{ AND } (Y = y))$$

## Q11: Define joint probability density function

For two continuous random variable $X$ and $Y$, the joint probability density function (PDF) is a function $f_{X,Y} : (\mathbb{R}, \mathbb{R}) \to [0, \infty)$ such that:

1. $f_{X,Y}(x, y) \geq 0$ for all $x, y \in \mathbb{R}$
2. $\int_{\mathbb{R}} \int_{\mathbb{R}} f_{X,Y}(x, y) dx dy = 1$
3. $P(a \leq X \leq b, c \leq Y \leq d) = P(X \in [a, b], Y \in [c, d]) = \int_c^d \int_a^b f_{X,Y}(x, y) dx dy$

## Q12: Define cumulative distribution function

A cumulative distribution function (CDF) is $F_X(x)$ is defined as

$$F_X(x) = P(X \leq x).$$

For a discrete random variable, CDF is the sum of probability mass function

$$F_X(x) = P(X \leq x) = \sum_{a \leq x} P_X(a)$$

For a continuous random variable, CDF is the integral of probability density function

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(z) dz$$

## Q13: Define conditional probability

Conditional probability of event $A$ given event $B$ is defined as

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

when $P(B) \neq 0$.

## Q14: State Bayes theorem

For any two events, $A$ and $B$

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

## Q15: State Bayes theorem in terms of likelihood, prior, evidence and posterior

For an observable event $D$ and a hidden event $\theta$, the posterior $P(\theta|D)$ can be estimated using Bayes theorem in terms of likelihood $P(D|\theta)$, prior $P(\theta)$ and evidence $P(D)$ as

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

## Q16: Define statistical independence

Two random variables $X$ and $Y$ are said to be independent, denoted as $X \perp Y$ if any of the following equivalent condition hold for all $x, y$ :

1. $$P(X = x, Y = y) = P(X = x)P(Y = y)$$
2. $$P(X = x|Y = y) = P(X = x)$$
3. $$P(Y = y|X = x) = P(Y = y)$$

## Q17: Define conditional independence

Two random variables $X$ and $Y$ are said to be conditionally independent given random variable $Z$, denoted as $X \perp Y|Z$ if for all $x, y, z$ :

$$P(X = x, Y = y|Z = z) = P(X = x|Z = z)P(Y = y|Z = z)$$

## Q18: Identically independently distributed (IID)

The random variables (RVs) $X_1, X_2, \ldots, X_n$ are identically independently distributed if they are mutually independent $X_i \perp X_j$ and have the same probability distributions $P_{X_i}(x_i) = P_{X_j}(x_j)$.

## Q19: Expectation of a function of a random variable

The expectation of a function $g(X)$ of a discrete random variable $X$ is defined as:

$$\mathbb{E}_X[g(X)] = \sum_{x \in \mathbb{Z}} P(X = x)g(x)$$

The expectation of a function $g(X)$ of a continuous random variable $X$ is defined as:

$$\mathbb{E}_X[g(X)] = \int_{x \in \mathbb{R}} f_X(x)g(x)dx$$

## Q20: What is the difference between sample mean and expectation

Sample mean of n samples is

$$\mu(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Expectation of a discrete random variable is

$$\mathbb{E}_X[X] = \sum_{x \in \Omega_X} P(X = x)x$$

Sample mean converges to the expectation when $n$ with high probability:

$$\lim_{n \to \infty} \mu(X_1, \ldots, X_n) = E_X[X]$$

## Q21: Define variance of a function of a random variable

The expectation of a function $g(X)$ of a random variable $X$ is given by

$$\mathbb{V}_X[g(X)] = \mathbb{E}_X\left[(g(X) - \mathbb{E}_X[g(X)])^2\right]$$

## Q22: Define a covariance matrix

For random vector $\mathrm{X} = [X_1, X_2, \ldots, X_n]$, the covariance matrix of $X$ is defined as:

$$\mathbb{V}_X[\mathrm{X}] = \mathbb{E}_X\left[(\mathrm{X} - \mathbb{E}_X[\mathrm{X}])\,(\mathrm{X} - \mathbb{E}_X[\mathrm{X}])^\top\right]$$

## Q23:

Given the dataset $\mathcal{D} = \{(\mathrm{x}_1, y_1), \ldots, (\mathrm{x}_n, y_n)\}$, a model $\hat{y}_i = f(\mathrm{x}_i; \theta)$, and a loss function $l(y_i, \hat{y}_i)$, show that the following optimization problem can be interpreted as maximum likelihood estimation. In the process show that for the interpretation, we need the IID (independently, identically distributed) assumption over the dataset. List any other assumptions that you need for the interpretation.

$$\theta^* = \arg\ \min_{\theta} \sum_{i=1}^{n} l(y_i, f(\mathrm{x}_i; \theta))$$

## A23:

Let the $\mathrm{x}_i$ and $y_i$ be random vectors for all $i$. Model the probability distribution as a negative log of the loss function:

$$P((\mathrm{x}_i, y_i)|\theta) = \frac{1}{Z}\exp(-l(y_i, f(\mathrm{x}_i; \theta)).$$

If the samples are IID, then we can write the probability of the entire dataset as products of sample probabilities

$$P(\mathcal{D}|\theta) = \prod_{i=1}^{n} P((\mathrm{x}_i, y_i)|\theta)$$

$$P(\mathcal{D}|\theta) = \prod_{i=1}^{n} \frac{1}{Z} \exp(-l(y_i, f(\mathrm{x}_i; \theta)).$$

A product of exponents is the summation of their powers,

$$P(\mathcal{D}|\theta) = \frac{1}{Z} \exp(-\sum_{i=1}^{n} l(y_i, f(\mathrm{x}_i; \theta)).$$

Denote

$$L(\mathcal{D}; \theta) = \sum_{i=1}^{n} l(y_i, f(\mathrm{x}_i; \theta).$$

The original optimization problem can be written as:

$$\theta^* = \arg \min_{\theta} L(\mathcal{D}; \theta)$$

Taking negative exponent on both sides turns the problem into a maximization problem because $\exp(-y)$ is a monotonically decreasing function.

$$\theta^* = \arg \max_{\theta} \exp(-L(\mathcal{D}; \theta))$$

This problem is the same as maximizing the likelihood $P(\mathcal{D}|\theta)$, hence maximum likelihood estimate.

## Q24:

Given the dataset $\mathcal{D} = \{(\mathrm{x}_1, y_1), \ldots, (\mathrm{x}_n, y_n)\}$, a model $\hat{y}_i = f(\mathrm{x}_i; \theta)$, a regularizer $R(\theta)$ and a loss function $l(y_i, \hat{y}_i)$, show that the following optimization problem can be interpreted as maximum-a-posteriori estimation. In the process show that for the interpretation, we need the IID (independently, identically distributed) assumption over the dataset. List any other assumptions that you need for the interpretation.

$$\theta^* = \arg \min_{\theta} \sum_{i=1}^{n} l(y_i, f(\mathrm{x}_i; \theta)) + \lambda R(\theta),$$

where $\lambda$ is some positive constant that balances between the loss function and the regularizer.

## A24:

Let the $\mathrm{x}_i$ and $y_i$ be random vectors for all $i$. Model the probability distribution as a negative log of the loss function:

$$P((\mathrm{x}_i, y_i)|\theta) = \frac{1}{Z}\exp(-l(y_i, f(\mathrm{x}_i; \theta)).$$

If the samples are IID, then we can write the probability of the entire dataset as products of sample probabilities

$$P(\mathcal{D}|\theta) = \prod_{i=1}^{n} P((\mathrm{x}_i, y_i)|\theta)$$

$$P(\mathcal{D}|\theta) = \prod_{i=1}^{n} \frac{1}{Z}\exp(-l(y_i, f(\mathrm{x}_i; \theta)).$$

A product of exponents is the summation of their powers,

$$P(\mathcal{D}|\theta) = \frac{1}{Z}\exp(-\sum_{i=1}^{n} l(y_i, f(\mathrm{x}_i; \theta)).$$

Denote

$$L(\mathcal{D}; \theta) = \sum_{i=1}^{n} l(y_i, f(\mathrm{x}_i; \theta).$$

The original optimization problem can be written as:

$$\theta^* = \arg \min_{\theta} L(\mathcal{D}; \theta) + \lambda R(\theta)$$

Taking negative exponent on both sides turns the problem into a maximization problem because $\exp(-y)$ is a monotonically decreasing function.

$$\theta^* = \arg \max_{\theta} \exp(-L(\mathcal{D}; \theta)) \exp(-\lambda R(\theta))$$

The first term is the same as maximizing the likelihood $P(\mathcal{D}|\theta)$. If we interpret the second term as a prior:

$$P(\theta) = \frac{1}{Z'}\exp(-\lambda R(\theta)),$$

then we can rewrite the original optimization problem as

$$\theta^* = \arg \max_{\theta} P(\mathcal{D}|\theta)P(\theta)$$

By Bayes theorem $P(\mathcal{D}|\theta)P(\theta) = P(\theta|\mathcal{D})P(\mathcal{D})$, hence we can write the optimization problem as maximizing the posterior

$$\theta^* = \arg \max_{\theta} P(\theta|\mathcal{D})P(\mathcal{D}).$$

We can ignore the evidence term $P(\mathcal{D})$, because it is independent of $\theta$ the optimization variable. The original problem reduces to maximizing the posterior, hence maximum a posteriori:

$$\theta^* = \arg \max_{\theta} P(\theta|\mathcal{D})$$

## Q25: Define L-p norm for $p = \{1, 2, \ldots\}$

$$\|\mathrm{x}\|_p = \left(|x_1|^p + |x_2|^p + \cdots + |x_n|^p\right)^{\frac{1}{p}}$$

## Q26: Find the minimum point for the following regularized least square problem and

$$\mathrm{w}^* = \arg \min_{\mathrm{w}} \|\mathrm{y} - \mathrm{X}\mathrm{w}\|^2 + \lambda \|\mathrm{w}\|^2,$$

where $\mathrm{w} \in \mathbb{R}^n$, $\mathrm{y} \in \mathbb{R}^m$, $\mathrm{X} \in \mathbb{R}^{m \times n}$ and $\lambda \in \mathbb{R}^+$

## A26:

Let $f(\mathrm{w}) = \|\mathrm{y} - \mathrm{X}\mathrm{w}\|^2 + \lambda \|\mathrm{w}\|^2$

Write $f(\mathrm{w})$ in terms of inner product,

$$f(\mathrm{w}) = (\mathrm{y} - \mathrm{X}\mathrm{w})^\top (\mathrm{y} - \mathrm{X}\mathrm{w}) + \lambda \mathrm{w}^\top \mathrm{w}$$

Expand and collect the terms,

$$f(\mathrm{w}) = \mathrm{w}^\top (\mathrm{X}^\top \mathrm{X} + \lambda I_n)\mathrm{w} - 2\mathrm{y}^\top \mathrm{X}\mathrm{w} + \mathrm{y}^\top \mathrm{y}$$

Taking the derivative of $f(\mathrm{w})$ we get,

$$\frac{\partial}{\partial \mathrm{w}} f(\mathrm{w}) = 2\mathrm{w}^\top (\mathrm{X}^\top \mathrm{X} + \lambda I_n) - 2\mathrm{y}^\top \mathrm{X}.$$

At the maximum point $\mathrm{w}^*$ the derivative of $f(\mathrm{w})$ is zero,

$$\frac{\partial}{\partial \mathrm{w}} f(\mathrm{w}) \bigg|_{\mathrm{w}^*} = 0_n^\top,$$

Equating the derivative to zero at $\mathrm{w}^*$, we can solve for $\mathrm{w}^*$,

$$2\mathrm{w}^{*\top} (\mathrm{X}^\top \mathrm{X} + \lambda I_n) - 2\mathrm{y}^\top \mathrm{X} = 0_n^\top.$$

Rearranging we get,

$$\mathrm{w}^* = (\mathrm{X}^\top \mathrm{X} + \lambda I_n)^{-1} \mathrm{X}^\top \mathrm{y}$$