Assignment - 2020101055

Submission by Sukhjinder Kumar (2020101055)

1. Define convex set and convex function.

   A set C is convex if the line segment between any two points in C lies in C, i.e., if for any $x_1, x_2 \in C$ and any θ with 0 ≤ θ ≤ 1, we have $\theta x_1 + (1-\theta)x_2 \in C$. A function $f : R^n \rightarrow R$ is convex if dom f is a convex set and if for all x, y ∈ dom f , and θ with 0 ≤ θ ≤ 1, we have f (θx + (1 − θ)y) ≤ θf (x) + (1 − θ)f (y).

2. Give an example of a function that is not convex.

   f(x) = - x^2

3. Define conjugate of a function.

   Let $f : R^n \rightarrow R$. The function $f^* : R^n \rightarrow R$, defined as $f^*(y) = sup(y^T x - f(x))$. x ∈ dom f is called the conjugate of the function f . The domain of the conjugate function consists of $y \in R^n$ for which the supremum is finite, i.e., for which the difference $y^T x - f(x)$ is bounded above on dom f .

4. Show that the conjugate function is always convex.

   We see immediately that f^* is a convex function, since it is the pointwise supremum of a family of convex (indeed, affine) functions of y. This is true whether or not f is convex.

5. Define strong and weak duality.

   If the equality $d^\star = p^\star$ holds, i.e., the optimal duality gap is zero, then we say that strong duality holds. Else, we say that weak duality holds.

6. What are the optimization solvers used for min-max problems in GAN or WGAN?

   SGD, ADAM, RMSProp, SVRG.

7. What is Jensen's inequality?

If x is a random variable such that x ∈ dom f with probability one, and f is convex, then we have $f(E[x]) \leq E[f(x)]$, provided the expectations exist.

8. State log-sum inequality.

$$\ln\left(\sum_{i=1}^{n} a_i b_i\right) \geq \sum_{i=1}^{n} a_i \ln(b_i)$$

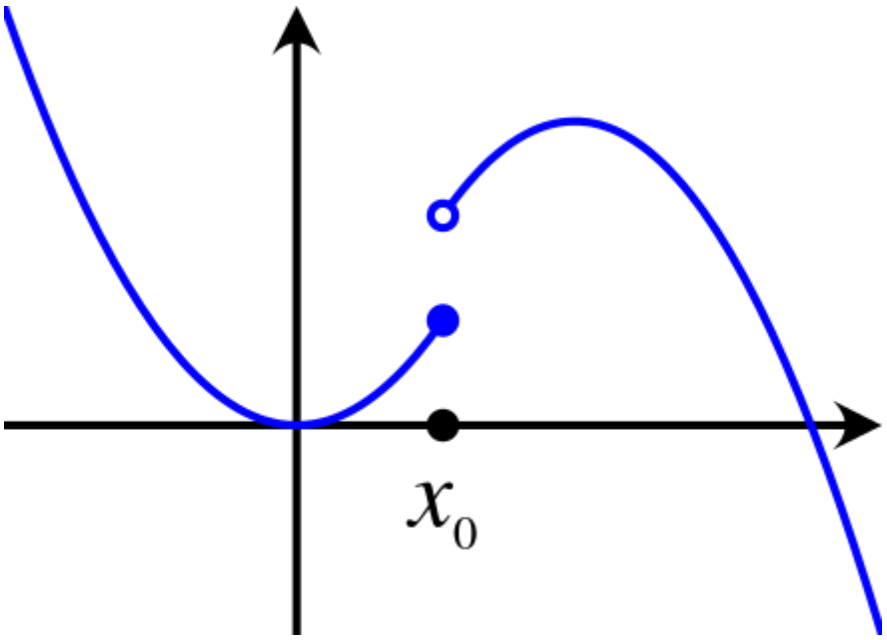
Consider $f : R^n \to R$ such that f(x) = xlogx. By using the second order condition, we can say that f is convex. Applying Jensen's inequality on this function with $\alpha_i = b_i/b$ and $t_i = a_i/b_i$, we get the required result.

9. Define lower semi-continuous function.

Let $f : R^d \to R \cup +\infty$, and $\bar{x} \in R^d$. We say that f is lower semi-continuous at x̄ if $f(\bar{x}) \leq \lim_{x \to \bar{x}} \inf f(x)$. We say that f is lower semi-continuous (l.s.c. for short) if f is lower semi-continuous at every $\bar{x} \in R^d$

10. Is lower semi-continuous function necessarily convex? Give an example.

No. Example →



11. Write an optimization problem with equality and inequality constraints.

Minimise $f_0(x)$ subject to $f_i(x) \leq 0, i = 1, ..., m$ and $h_i(x) = 0, i = 1, ..., p$

12. Write the dual problem.

$$g(\lambda, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) = \inf_{x \in \mathcal{D}} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right).$$

13. Show that the dual function is concave.

    Since the dual function is the pointwise infimum of a family of affine functions of (λ, ν), it is concave always.

14. Derive a min-max characterization for weak and strong duality (for inequality constraints only).

    Weak duality can also be obtained as a consequence of the following minimax inequality, which is valid for any function φ of two vector variables x, y, and any subsets X , Y: $\max_{y \in Y} \min_{x \in X} \varphi(x, y) \leq \min_{x \in X} \max_{y \in Y} \varphi(x, y)$.

    To prove this, start from $\forall$ x, y : $\min_{x' \in X} \varphi(x', y) \leq \max_{y' \in Y} \varphi(x, y')$. And take the minimum over x ∈ X on the right-hand side, then the maximum over y ∈ Y on the left-hand side. Weak duality is indeed a direct consequence of the above. To see this, start from the unconstrained formulation, and apply the above inequality with φ = L the Lagrangian of the original problem, and y = λ the vector of Lagrange multipliers.

15. What is Slater's condition?

    There exists an x ∈ relint D such that $f_i(x) < 0, i = 1, ..., m$, Ax = b. Such a point is sometimes called strictly feasible, since the inequality constraints hold with strict inequalities. Slater's theorem states that strong duality holds, if Slater's condition holds (and the problem is convex).

16. If f^* denote the conjugate of f , then when do we have $f^{**} = f$?

    If f is a convex function and its domain is a convex set, then f ** = f.

17. Define independent and mutually exclusive events.

    Two events A and B are independent if we have P(A,B)=P(A)P(B). Two events A and B are mutually exclusive if we have P(A∩B) = 0.

18. Define conditional expectation.

Conditional expectation is a concept in probability theory that describes the expected value of a random variable given certain information or conditions. More specifically, the conditional expectation of a random variable X given a sigma-algebra F is a function E[X | F] that represents the expected value of X, given the information contained in F.

In simpler terms, the conditional expectation of X given F represents the average value of X over all outcomes that are consistent with the information contained in F. It is denoted by E[X | F] and can be defined as: $E[X|F] = \int x f(x|F) dx$ where f(x | F) is the conditional probability density function of X given F, and the integral is taken over all possible values of X.

19. Write PDF of Bernoulli, Binomial, Standard normal and Normal distribution.

    a. Bernoulli distribution: The Bernoulli distribution is a discrete probability distribution that describes the probability of a single trial that can have only two possible outcomes, typically denoted as success (S) and failure (F). The probability of success is p, and the probability of failure is (1-p). The probability mass function (PMF) of a Bernoulli distribution is: $P(X = x) = p^x \times (1-p)^{(1-x)}$, for x ∈ {0, 1}, where X is the random variable representing the outcome of the trial.

    b. Binomial distribution: The binomial distribution is a discrete probability distribution that describes the probability of having a certain number of successes in a fixed number of independent trials, each with the same probability of success. It is characterized by two parameters: n (the number of trials) and p (the probability of success). The probability mass function (PMF) of a binomial distribution is: $P(X = k) = \binom{n}{k} \times p^k \times (1-p)^{(n-k)}$, for k ∈ {0, 1, ..., n}, where X is the random variable representing the number of successes, (n choose k) is the binomial coefficient, and p is the probability of success.

    c. Standard normal distribution: The standard normal distribution is a continuous probability distribution with mean 0 and standard deviation 1. It is also known as the Gaussian distribution or the bell curve distribution. The probability density function (PDF) of the standard normal distribution is: $f(x) = (1/\sqrt{2\pi}) \times e^{-x^2/2}$, for x ∈ R, where x is the random variable and e is the base of the natural logarithm.

    d. Normal distribution: The normal distribution is a continuous probability distribution that can take any real value, characterized by two parameters: mean μ and standard deviation σ. It is also known as the Gaussian distribution. The probability density

function (PDF) of a normal distribution is: $f(x) = 1/(\sigma\sqrt{2\pi}) \times e^{-(x-\mu)^2/2\sigma^2}$, for x ∈ R, where x is the random variable, μ is the mean, σ is the standard deviation, and e is the base of the natural logarithm.

20. When do we say that two random variables are independent?

    In probability theory, two random variables X and Y are said to be independent if the occurrence or outcome of one variable does not affect the probability distribution of the other variable. In other words, if the knowledge of the value of one variable does not change the probability distribution of the other variable, then the variables are considered independent. Mathematically, two random variables X and Y are independent if and only if the joint probability distribution of X and Y can be expressed as the product of their marginal probability distributions: P(X = x, Y = y) = P(X = x) * P(Y = y) This means that the probability of X taking on a certain value and Y taking on another certain value is simply the product of the probability of X taking on the first value and the probability of Y taking on the second value, for any possible pair of values (x, y). In summary, two random variables are independent if the value of one variable does not affect the probability distribution of the other variable.

21. Define marginal PDF for a function of two random variables.

    The marginal probability density function (PDF) for a function of two random variables is the probability distribution of that function when we integrate over the other variable. More specifically, suppose we have two random variables X and Y with joint probability density function f(x,y). Then, the marginal PDF of X is the function g(x) defined as: $g(x) = \int f(x,y)dy$where the integral is taken over all possible values of y that X can take. In other words, the marginal PDF of X is the probability distribution of X alone, without any consideration of the value of Y. Similarly, the marginal PDF of Y is the function h(y) defined as: $h(y) = \int f(x,y)dx$ where the integral is taken over all possible values of x that Y can take. The marginal PDF of Y is the probability distribution of Y alone, without any consideration of the value of X. The marginal PDFs are important because they allow us to study the properties of a single random variable, without worrying about the other variables. They are often used in applications where one of the variables is of primary interest, such as in regression analysis, where we want to model the relationship between a response variable and a set of predictor variables.

22. Define law of total variance, law of total expectation, law of total probability.

The laws of total variance, total expectation, and total probability are important concepts in probability theory that relate to the decomposition of the variance, expectation, and probability of a random variable into sub-parts based on a partition of the sample space.

1. Law of total variance:

The law of total variance states that the total variance of a random variable can be expressed as the sum of the conditional variances of the random variable given a partition of the sample space, weighted by their corresponding probabilities. Mathematically, if X is a random variable and A1, A2, ..., An is a partition of the sample space, then the law of total variance can be written as: Var(X) = E[Var(X | Ai)] + Var(E[X | Ai]) where E[.] denotes the expectation operator and Var(.) denotes the variance operator.

2. Law of total expectation:

The law of total expectation, also known as the law of iterated expectations, states that the expected value of a random variable can be expressed as the weighted average of its conditional expectations given a partition of the sample space. Mathematically, if X is a random variable and A1, A2, ..., An is a partition of the sample space, then the law of total expectation can be written as: E[X] = E[E[X | Ai] * P(Ai)], where E[.] denotes the expectation operator and P(.) denotes the probability operator.

3. Law of total probability:

The law of total probability states that the probability of an event can be expressed as the sum of the conditional probabilities of the event given a partition of the sample space, weighted by their corresponding probabilities. Mathematically, if A1, A2, ..., An is a partition of the sample space and B is an event, then the law of total probability can be written as: P(B) = ∑ P(B | Ai) * P(Ai), where P(.) denotes the probability operator and the sum is taken over all possible events Ai in the partition.

23. Define correlation and covariance.

Correlation and covariance are measures of the linear relationship between two random variables.

Covariance:

The covariance between two random variables X and Y is a measure of how much their values tend to vary together. It is defined as the expected value of the product of the deviations of X and Y from their respective means. Mathematically, the covariance between X and Y is defined as: cov(X, Y) = E[(X - E[X])(Y - E[Y])], where E[.] denotes the expectation operator. If cov(X, Y) is positive, it indicates that X and Y tend to increase or decrease together, while a negative covariance indicates that they tend to move in opposite directions. A covariance of zero indicates that there is no linear relationship between X and Y.

Correlation:

The correlation between two random variables X and Y is a measure of the strength of their linear relationship. It is defined as the covariance between X and Y, normalized by the product of their standard deviations. Mathematically, the correlation between X and Y is defined as: corr(X, Y) = cov(X, Y) / (std(X) * std(Y)), where std(.) denotes the standard deviation operator. The correlation takes on values between -1 and 1, where a correlation of -1 indicates a perfect negative linear relationship, a correlation of 1 indicates a perfect positive linear relationship, and a correlation of 0 indicates no linear relationship. The correlation is unitless, meaning that it does not depend on the scales or units of measurement of the variables.

24. State method of transform for vectors of random variables.

    The method of transforms is a technique used to find the distribution of a function of one or more random variables. Specifically, the method of transforms for vectors of random variables involves finding the joint distribution of a transformed vector of random variables using the joint distribution of the original vector of random variables.

    Suppose we have a vector of random variables X = (X1, X2, ..., Xn) with joint probability density function fX(x1, x2, ..., xn), and let Y = (Y1, Y2, ..., Ym) be a vector of functions of X, such that Yi = g(Xi) for some function g.

    Then, the joint probability density function of Y, denoted by fY(y1, y2, ..., ym), can be found using the method of transforms:

    1. First, find the joint cumulative distribution function (CDF) of Y as a function of y1, y2, ..., ym:

       FY(y1, y2, ..., ym) = P(Y1 ≤ y1, Y2 ≤ y2, ..., Ym ≤ ym)

2. Next, differentiate the joint CDF with respect to each yi to obtain the joint PDF of Y:

   fY(y1, y2, ..., ym) = (∂^m FY(y1, y2, ..., ym)) / (∂y1 ∂y2 ... ∂ym)

   where ∂^m denotes the mth partial derivative.

The method of transforms can be used to find the distribution of any vector of functions of the original vector of random variables X, as long as the inverse of the transformation exists and is differentiable. It is a powerful tool for solving many problems in probability and statistics, including finding the distribution of sums, products, ratios, and functions of multiple random variables.

25. State Markov's, Chebychev's, Cauchy-Schwarz, Jensen's inequality.

Markov's inequality:

Markov's inequality is a probability inequality that provides an upper bound on the probability that a non-negative random variable X is greater than or equal to some positive value a. Mathematically, it states that for any non-negative random variable X and any positive constant a,

P(X ≥ a) ≤ E[X] / a,

where E[X] denotes the expected value of X. Markov's inequality is useful for bounding tail probabilities of random variables and is widely used in probability theory and statistics.

Chebyshev's inequality:

Chebyshev's inequality is a probability inequality that provides an upper bound on the probability that a random variable X deviates from its mean by more than a certain amount. Specifically, it states that for any random variable X with finite mean and variance, and any positive constant k,

P(|X - E[X]| ≥ kσ) ≤ 1 / k^2,

where σ^2 denotes the variance of X. Chebyshev's inequality is a more general result than Markov's inequality and is often used to establish concentration bounds for random variables.

Cauchy-Schwarz inequality:

The Cauchy-Schwarz inequality is a fundamental inequality in mathematics that relates the inner product of two vectors to their magnitudes. Specifically, it states

that for any two real vectors x and y,

$|(x, y)| \leq ||x|| \, ||y||$,

where (x, y) denotes the inner product of x and y, and ||x|| and ||y|| denote the magnitudes of x and y, respectively. In probability theory, the Cauchy-Schwarz inequality is often used to prove other inequalities, such as Jensen's inequality.

Jensen's inequality:

Jensen's inequality is a fundamental inequality in mathematics that relates the convexity of a function to the expectation of a random variable. Specifically, it states that for any convex function f(x), and any random variable X,

$f(E[X]) \leq E[f(X)]$,

where E[.] denotes the expectation operator. Jensen's inequality is widely used in probability theory and statistics to establish bounds on functions of random variables. For example, it can be used to prove that the logarithm of the expected value of a non-negative random variable is less than or equal to the expected value of the logarithm of the random variable.

26. Define convergence in probability, convergence in mean, convergence in distribution.

    In probability theory and statistics, there are several types of convergence of random variables. Three commonly used types of convergence are:

    1. Convergence in Probability:

    A sequence of random variables {Xn} converges to a random variable X in probability if for any positive number ε, the probability that the absolute difference between Xn and X is greater than ε goes to zero as n goes to infinity. Mathematically, we say that {Xn} converges to X in probability if:

    $\lim P(|Xn - X| > \varepsilon) = 0$, as $n \rightarrow \infty$

    This type of convergence is denoted as $Xn \rightarrow p\ X$, and is also known as convergence in the weak sense.

    2. Convergence in Mean:

    A sequence of random variables {Xn} converges to a random variable X in mean (or converges in the mean squared) if the expected value of the absolute difference between Xn and X converges to zero as n goes to infinity. Mathematically, we say

that {Xn} converges to X in mean if: lim E[|Xn - X|] = 0, as n → ∞. This type of convergence is denoted as Xn →m X, and is also known as convergence in the strong sense or convergence in the mean squared.

3. Convergence in Distribution:

A sequence of random variables {Xn} converges to a random variable X in distribution if the cumulative distribution function (CDF) of Xn converges to the CDF of X at all points where the CDF of X is continuous. Mathematically, we say that {Xn} converges to X in distribution if: lim F_n(x) = F(x), as n → ∞

where F_n(x) and F(x) denote the CDFs of Xn and X, respectively. This type of convergence is denoted as Xn →d X, and is also known as weak convergence.

Convergence in probability and convergence in distribution are weaker forms of convergence than convergence in mean. Convergence in probability implies convergence in distribution, but the converse is not always true. Convergence in mean implies convergence in probability and convergence in distribution, but the converse is not always true either.

27. State law of large numbers.

The law of large numbers is a fundamental theorem in probability theory that states that the sample average of a large number of independent and identically distributed (i.i.d.) random variables converges in probability to the expected value of the random variable. In other words, as the sample size becomes larger and larger, the average of the observations approaches the expected value of the underlying distribution.

Formally, let X1, X2, ..., Xn be a sequence of i.i.d. random variables with finite expected value μ, and let S_n = (X1 + X2 + ... + Xn)/n be their sample mean. Then, according to the law of large numbers, we have:

$$\lim P(|S_n - \mu| > \varepsilon) = 0, as n \to \infty,$$

for any positive ε, where P denotes probability. In other words, the probability that the difference between the sample mean and the expected value exceeds a certain amount decreases to zero as the sample size becomes larger.

The law of large numbers is a fundamental result in probability theory and has many important applications in statistics, finance, and other fields. It provides a theoretical

foundation for the use of sample means to estimate population means and is a key component of many statistical methods.

28. State central limit theorem (CLT).

    The central limit theorem (CLT) is a fundamental theorem in probability theory that describes the distribution of the sample mean of a large number of independent and identically distributed (i.i.d.) random variables. The theorem states that under certain conditions, the sample mean will follow a normal distribution, regardless of the underlying distribution of the individual random variables.

    Formally, let X1, X2, ..., Xn be a sequence of i.i.d. random variables with mean μ and variance σ^2, and let S_n = (X1 + X2 + ... + Xn)/n be their sample mean. Then, according to the central limit theorem, as the sample size n becomes larger, the distribution of S_n approaches a normal distribution with mean μ and variance σ^2/n, i.e.,

    S_n ~ N(μ, σ^2/n)

    This result holds true for a wide variety of underlying distributions of the individual random variables, as long as they have a finite mean and variance. The central limit theorem is a key result in statistics, as it justifies the use of the normal distribution as an approximation for the distribution of the sample mean, even when the underlying distribution is not normal.

    The central limit theorem has many important applications in statistics, including hypothesis testing, confidence intervals, and regression analysis. It is also used in fields such as finance, engineering, and physics, where the normal distribution is commonly used to model many natural phenomena.

29. Explain how CLT is used in practice in estimations.

    The central limit theorem (CLT) is a fundamental result in probability theory that is widely used in practice for making statistical estimations. The CLT states that the distribution of the sample mean of a large number of independent and identically distributed (i.i.d.) random variables will approach a normal distribution, regardless of the underlying distribution of the individual random variables. This property of the sample mean makes it a powerful tool for making estimations about the population mean or other parameters of interest.

In practice, the CLT is often used in conjunction with statistical inference techniques, such as hypothesis testing and confidence interval estimation. For example, suppose we want to estimate the mean μ of a population based on a random sample of n observations. Using the CLT, we can construct a confidence interval for μ based on the sample mean and standard error, which is given by:

CI = ($\bar{x}$ - zα/2 * s/√n, $\bar{x}$ + zα/2 * s/√n)

where $\bar{x}$ is the sample mean, s is the sample standard deviation, n is the sample size, and zα/2 is the critical value from the standard normal distribution corresponding to the desired level of confidence (e.g., α/2 = 0.025 for a 95% confidence interval).

This confidence interval provides a range of plausible values for the population mean based on the sample data, and the width of the interval is determined by the sample size and variability. As the sample size increases, the interval becomes narrower and more precise, due to the decreasing standard error.

Another way CLT is used in practice is to test hypotheses about population parameters. For example, we may want to test whether the population mean is equal to a certain value. Using the CLT, we can construct a test statistic based on the sample mean and standard error, and compare it to a critical value from the standard normal distribution to determine whether to reject or fail to reject the null hypothesis.

Overall, the central limit theorem is a powerful tool for making statistical estimations and testing hypotheses in practice, and its applications are widespread in fields such as finance, economics, medicine, and engineering.

30. Define sample mean.

    The sample mean is a statistical measure that represents the average value of a set of observations or data points drawn from a larger population. It is defined as the sum of all the observations divided by the total number of observations in the sample.

    Suppose we have a sample of n observations, denoted by x1, x2, ..., xn. The sample mean, denoted by $\bar{x}$ (pronounced "x bar"), is given by: $\bar{x}$ = (x1 + x2 + ... + xn) / n

    The sample mean is often used as an estimator of the population mean, which is the true average value of the variable of interest in the entire population. When the

sample is drawn randomly from the population, the sample mean is an unbiased estimator of the population mean, meaning that its expected value equals the true population mean.

The sample mean is a commonly used statistic in many fields, including finance, economics, engineering, and the social sciences. It is also used in statistical inference, such as hypothesis testing and confidence interval estimation, and plays a central role in many statistical models and methods.

31. What is MLE?

    MLE stands for Maximum Likelihood Estimation, which is a method used to estimate the parameters of a statistical model. The basic idea of MLE is to find the values of the model parameters that maximize the likelihood function, which is a function that measures how well the model fits the observed data.

    In other words, MLE is a method for finding the "best-fit" parameters of a statistical model, given a set of observed data. The likelihood function is constructed based on the probability density function (PDF) or probability mass function (PMF) of the model, and the goal of MLE is to find the parameter values that maximize this function.

    The MLE method involves finding the values of the model parameters that maximize the likelihood function, which is typically done using numerical optimization techniques such as gradient descent or Newton's method. Once the MLE estimates of the parameters are obtained, they can be used to make predictions or draw inferences about the population of interest.

    MLE is widely used in statistical inference and machine learning, and is an important tool for fitting models to data and making predictions. It is a versatile and powerful technique that can be applied to a wide range of statistical models, including linear regression, logistic regression, time series models, and many others.

32. What is MAP estimation.

    MAP estimation, or Maximum A Posteriori estimation, is a method used to estimate the parameters of a statistical model based on both prior knowledge and observed data. It is a Bayesian approach to parameter estimation that takes into account both the likelihood of the observed data and the prior probability of the parameters.

In MAP estimation, the goal is to find the parameter values that maximize the posterior probability distribution, which is a combination of the likelihood function and the prior distribution of the parameters. The posterior distribution gives the probability distribution of the parameters given the observed data and prior knowledge.

Mathematically, the MAP estimate is given by: $\theta_{MAP} = \arg\max(P(\theta|D))$

where θ is the parameter vector to be estimated, D is the observed data, and P(θ|D) is the posterior distribution of the parameters given the observed data.

Compared to MLE (Maximum Likelihood Estimation), MAP estimation incorporates prior information or beliefs about the parameters, which can help to reduce the variance of the parameter estimates, especially when the amount of observed data is limited.

MAP estimation is widely used in signal processing, image processing, and machine learning, among other fields. It is often used in cases where there is some prior knowledge or belief about the parameters, and where it is desirable to incorporate this information into the parameter estimation process

33. Sample mean an unbiased estimator for true mean.

    The sample mean is an unbiased estimator for the true mean of a population. An estimator is said to be unbiased if its expected value is equal to the true value of the parameter being estimated.

    Suppose we have a random sample of size n, denoted by X1, X2, ..., Xn, drawn from a population with mean μ and variance σ^2. The sample mean, denoted by $\bar{X}$, is defined as the sum of the sample values divided by the sample size, i.e.,

    $\bar{X}$ = (X1 + X2 + ... + Xn) / n. The expected value of the sample mean is: E($\bar{X}$) = E[(X1 + X2 + ... + Xn) / n] = (1/n)[E(X1) + E(X2) + ... + E(Xn)] = μ,

    where the last equality follows from the linearity of the expected value operator. Therefore, the sample mean is an unbiased estimator for the true population mean.

    Intuitively, the idea behind the unbiasedness of the sample mean is that, on average, the sample mean will be equal to the true population mean. This is true regardless of the distribution of the population, as long as the sample is drawn randomly from the population.

Note that unbiasedness does not guarantee that the sample mean will be close to the true mean for any particular sample. The sample mean can still be affected by random sampling variation, and its precision depends on the sample size and the variability of the population. However, on average, the sample mean will provide an accurate estimate of the true population mean.

34. Define a push-forward distribution.

    A push-forward distribution, also known as image measure or induced measure, is a way to define a probability distribution on one space based on a probability distribution on another space and a measurable function that maps between the two spaces.

    Let (Ω, F, P) be a probability space and let X be a random variable on Ω with probability distribution P_X. Let (E, G) be another measurable space and let f: Ω → E be a measurable function that maps each outcome in Ω to an outcome in E. Then, the push-forward distribution of P_X under f, denoted by f_*P_X or P_X ∘ f^(-1), is defined as the probability distribution of the random variable Y = f(X) on E.

    More formally, the push-forward distribution f_*P_X is defined by: f_*P_X(A) = P_X(f^(-1)(A)) = P(X ∈ f^(-1)(A)), for any measurable subset A of E, where f^(-1)(A) is the preimage of A under f.

    Intuitively, the push-forward distribution assigns probabilities to subsets of the target space E based on how likely they are to be the image of subsets of the source space Ω under the function f.

    Push-forward distributions are used in many areas of probability and statistics, including probability theory, stochastic processes, and machine learning. They provide a way to describe the distribution of a random variable in terms of its relationship with other random variables or measurable functions

35. How can we generate samples from N(μ, σ 2) from N(0, 1)? Here N stands for normal distribution.

    We can generate samples from N(μ, σ^2) using the following transformation:

    Let Z be a standard normal random variable (i.e., Z ~ N(0,1)). Then, we can obtain a sample from N(μ, σ^2) by using the transformation: X = μ + σZ, where X ~ N(μ, σ^2).

To see why this works, note that if we substitute Z = (X - μ) / σ in the standard normal distribution, we obtain: P(Z ≤ z) = P((X - μ) / σ ≤ z) = P(X ≤ μ + σz)

which is the cumulative distribution function (CDF) of the normal distribution with mean μ and variance σ^2. Therefore, by generating samples from a standard normal distribution and applying the above transformation, we can obtain samples from a normal distribution with any mean μ and variance σ^2.

There are many methods for generating samples from a standard normal distribution, such as the Box-Muller transform or the Marsaglia polar method. Once we have obtained a sample from the standard normal distribution, we can apply the transformation above to obtain a sample from N(μ, σ^2).

36. State how inverse CDF can be used for generating samples given the initial samples $x_1, ..., x_m$.

    The inverse CDF (cumulative distribution function) method can be used to generate samples from any probability distribution for which the CDF is known or can be computed. The basic idea of the method is to generate samples from a uniform distribution on the interval [0, 1], and then apply the inverse of the CDF to obtain samples from the desired distribution.

    Given m initial samples x1, ..., xm from a uniform distribution on [0, 1], the inverse CDF method can be used to generate samples from a probability distribution with CDF F(x) as follows:

    1. Sort the initial samples in increasing order, i.e., x(1) ≤ x(2) ≤ ... ≤ x(m).
    2. For each i = 1, ..., n, compute the value y(i) = F^(-1)(x(i)), where F^(-1) is the inverse of the CDF F.
    3. The resulting samples y(1), ..., y(m) are samples from the distribution with CDF F(x).

    Note that the inverse CDF method can be computationally expensive when the inverse of the CDF is difficult to compute or when the CDF has a complex form. In these cases, other methods such as rejection sampling or Markov chain Monte Carlo (MCMC) may be more efficient.

37. More concretely, let us assume that we did an experiment to obtain following samples: S = {1, 1, 3, 1, 5, 2, 4, 2, 7, 1, 2, 3, 5, 6, 6, 3, 9, 9, 1, 3, 3, 4, 8}. We know that these samples are generated from the set {1, . . . , 9}. However, we don’t know

the underlying distribution. Hence we don’t know the real distribution. That is, we don’t know the probability of each numbers. Use inverse CDF method to generate more samples for this underlying distribution. This is exact same as quiz1, got full marks for it. Just generate CDF from data given and use inverse transform method.

38. What are the difficulties for generating samples using inverse CDF approach?

    While the inverse CDF (cumulative distribution function) method is a powerful tool for generating samples from a wide range of probability distributions, there are several difficulties that can arise when using this approach:

    a. Inverse CDF may be difficult to compute: In order to use the inverse CDF method, we need to have an explicit formula or algorithm for computing the inverse CDF function. In some cases, this may not be possible or may be very difficult, especially for complex or high-dimensional distributions.

    b. Limited accuracy of numerical methods: Even if we have a formula or algorithm for computing the inverse CDF, the numerical methods used to evaluate it may be subject to limited accuracy. This can result in errors or inaccuracies in the generated samples.

    c. Inefficient for high-dimensional distributions: The inverse CDF method can be computationally expensive, especially for high-dimensional distributions where evaluating the inverse CDF function for each sample can be very time-consuming.

    d. Not suitable for non-smooth distributions: The inverse CDF method assumes that the distribution has a smooth CDF function, which may not be the case for some distributions, such as discrete distributions or distributions with discontinuous CDFs.

    e. Rejection sampling may be more efficient: In some cases, rejection sampling may be more efficient than the inverse CDF method, especially when the distribution is complex or high-dimensional. Rejection sampling involves generating samples from a simpler distribution and then using a rejection criterion to accept or reject them based on whether they belong to the desired distribution.

39. Prove the relationship between MSE, Bias and Variance.

The Mean Squared Error (MSE) of an estimator is defined as the expected value of the squared difference between the estimator and the true value: MSE = $E[(\theta - \bar{y})^2]$, where θ is the true value of the parameter being estimated, and ȳ is the estimate obtained from the data.

The MSE can be decomposed into two components: the squared bias of the estimator and the variance of the estimator: MSE = $Bias^2$ + Var, where Bias = $E[\bar{y}]$ - θ and Var = $E[(\bar{y} - E[\bar{y}])^2]$

The bias measures the difference between the expected value of the estimator and the true value, while the variance measures the variability of the estimator around its expected value.

To prove this relationship, we start by expanding the MSE:

MSE = $E[(\theta - \bar{y})^2]$ = $E[\theta^2 - 2\theta\bar{y} + \bar{y}^2]$ = $E[\theta^2]$ - 2E[θȳ] + $E[\bar{y}^2]$

Now we can use the following properties of expected values:

E[θ] = θ, E[c] = c (where c is a constant), and E[X + Y] = E[X] + E[Y] (where X and Y are random variables).

Using these properties, we can simplify the expression for the MSE:

MSE = $E[\theta^2]$ - 2θE[ȳ] + $E[\bar{y}^2]$ + 2θE[ȳ] - 2E[θȳ] + E[θȳ] - θ^2 + θ^2

= $E[\theta^2]$ - θ^2 + $E[\bar{y}^2]$ - 2E[θȳ] + $E[\theta^2]$ - θ^2

= $(E[\theta^2] - \theta^2)$ + ($E[\bar{y}^2]$ - 2E[θȳ] + $E[\theta^2] - \theta^2$)

= $Bias^2$ + Var

Therefore, we have shown that the MSE can be decomposed into the squared bias and the variance of the estimator. This decomposition provides insight into the trade-off between bias and variance when designing estimators. In general, reducing bias tends to increase variance, and reducing variance tends to increase bias. The goal is to find an estimator that strikes a balance between these two factors to achieve low MSE.

40. Prove that if MSE(Θ̂) n goes to zero and n tends to ∞, then Θ̂n is an consistent estimator.

    To prove that if MSE(Θ) n goes to zero as n tends to ∞, then Θ n is a consistent estimator, we need to show that Θ n converges in probability to the true value Θ. That is, $\lim_{n\to\infty} P(|\Theta_n - \Theta| > \varepsilon) = 0$, for any ε > 0. Using the definition of MSE,

we can write $MSE(\Theta)_n = E[(\Theta_n - \Theta)^2]$Expanding the square, we get $MSE(\Theta)_n = E[\Theta_n^2 - 2\Theta n\Theta + \Theta^2]$ = $E[\Theta_n^2] - 2\Theta E[\Theta_n] + \Theta^2$ = $Var(\Theta_n) + Bias^2(\Theta_n, \Theta)$, where $Bias(\Theta_n, \Theta) = E[\Theta_n] - \Theta$ is the bias of the estimator. If MSE(Θ) n tends to zero as n tends to ∞, then both the variance and the squared bias must tend to zero: $\lim_{n\to\infty} Var(\Theta_n) + Bias^2(\Theta_n, \Theta) = 0$

Since the variance is non-negative, it follows that the bias must also tend to zero: $\lim_{n\to\infty} Bias^2(\Theta_n, \Theta) = 0$. Therefore, we have shown that if MSE(Θ) n tends to zero as n tends to ∞, then the estimator Θ n is consistent, meaning that it converges in probability to the true value Θ.

41. A bag contains 3 balls. Each ball is either red or blue. Let θ be the number of blue balls. Here θ = 0, 1, 2, 3. Define the samples (X 1 , X 2 , X 3 , X 4) = (1, 1, 0, 1). Estimate number of blue balls.

    Given that the bag contains 3 balls, there are only four possible configurations for the number of blue balls: θ = 0, 1, 2, or 3.

    Let X be the number of blue balls in the sample of size 4. We can model X as a binomial distribution with parameters n = 4 and θ, the probability of drawing a blue ball from the bag. Therefore, the likelihood function is given by: $L(\theta|x) = Pr(X = x|\theta) = \binom{4}{x}\theta^x(1-\theta)^{(4-x)}$.

    We can use maximum likelihood estimation to find the value of θ that maximizes the likelihood function, given the observed sample (X_1, X_2, X_3, X_4) = (1, 1, 0, 1).

    The log-likelihood function is given by: $logL(\theta|x) = log[\binom{4}{x}\theta^x(1-\theta)^{(4-x)}] = log\binom{4}{x} + xlog\theta + (4-x)log(1-\theta)$

    Taking the derivative of the log-likelihood function with respect to θ, we get: d/dθ log L(θ|x) = x/θ - (4-x)/(1-θ). Setting this derivative to zero, we can solve for θ: x/θ = (4-x)/(1-θ), θ = x/2

    Therefore, the maximum likelihood estimate for the number of blue balls in the bag is θ = x/2 = 1/2.

    Thus, based on the observed sample (X 1 , X 2 , X 3 , X 4) = (1, 1, 0, 1), we estimate that there are 1/2 * 3 = 1.5 blue balls in the bag. Since we cannot have a fractional number of balls, we round our estimate to the nearest integer, which is 2. Therefore, we estimate that there are 2 blue balls in the bag.

42. Define KL divergence.

KL divergence (Kullback-Leibler divergence) is a measure of the difference between two probability distributions P and Q. It is defined as:

$$KL(P||Q) = \sum_x P(x)log[P(x)/Q(x)]$$

where x is the possible outcomes of a random variable, P(x) is the probability of x according to distribution P, and Q(x) is the probability of x according to distribution Q. The logarithm is typically taken to base 2 or natural logarithm.

KL divergence is asymmetric, i.e., in general, KL(P || Q) ≠ KL(Q || P). It also does not satisfy the triangle inequality, i.e., in general, KL(P || R) + KL(R || Q) ≥ KL(P || Q).

KL divergence is often used in machine learning and information theory for model selection, optimization, and evaluation. For example, it can be used to compare the output of a probabilistic model with the true distribution of a dataset, or to measure the difference between a prior and posterior distribution in Bayesian inference.

43. Show that KL(p, q) is convex in p and q.

To show that KL(p, q) is convex in p and q, we need to show that the Hessian matrix of KL(p, q) with respect to p and q is positive semidefinite, i.e., all its eigenvalues are nonnegative. First, we write the KL divergence as: $KL(p, q) = \sum_x p(x)log[p(x)/q(x)]$. The gradient of KL(p, q) with respect to p is: ∇p KL(p, q) = (1 + log[p(x)/q(x)]) * δ(x), where δ(x) is the Kronecker delta function. Similarly, the gradient of KL(p, q) with respect to q is: ∇q KL(p, q) = -p(x)/q(x) * δ(x).

The Hessian matrix of KL(p, q) with respect to p and q is:

H(p, q) = [∂2 KL(p, q) / ∂p2 ∂2 KL(p, q) / ∂p∂q]

[∂2 KL(p, q) / ∂q∂p ∂2 KL(p, q) / ∂q2]

The second derivative of KL(p, q) with respect to p is: ∂2 KL(p, q) / ∂p2 = -∑x log[p(x)/q(x)] * δ(x). The second derivative of KL(p, q) with respect to q is: ∂2 KL(p, q) / ∂q2 = ∑x p(x)/q(x)

The mixed derivative of KL(p, q) with respect to p and q is: ∂2 KL(p, q) / ∂p∂q = 0

Now, we can evaluate the eigenvalues of the Hessian matrix H(p, q) by plugging in the above expressions for the second derivatives:

λ1 = -∑x log[p(x)/q(x)] * δ(x)

λ1 = ∑x p(x)/q(x)

Since both λ1 and λ2 are nonnegative (by the properties of logarithm and the fact that p(x) and q(x) are probabilities), we can conclude that the Hessian matrix H(p, q) is positive semidefinite, and hence KL(p, q) is convex in p and q.

44. Show an example to show that KL(p, q) / = KL(q, p).

    Let us consider two probability distributions: p = [0.3, 0.7], q = [0.7, 0.3]. Then, the KL divergence from p to q is: KL(p, q) = 0.3 * log(0.3/0.7) + 0.7 * log(0.7/0.3) = 0.241, And the KL divergence from q to p is: KL(q, p) = 0.7 * log(0.7/0.3) + 0.3 * log(0.3/0.7) = 0.511. Therefore, we can see that KL(p, q) is not equal to KL(q, p).

45. Define $f-$divergence.

    The f-divergence is a family of measures of the difference between two probability distributions. It is a generalization of the Kullback-Leibler (KL) divergence and includes other divergences such as the Total Variation (TV) distance and the Hellinger distance as special cases. The f-divergence is defined as:

    $$D_f(p||q) = \int q(x) f(p(x)/q(x)) dx$$

    where p and q are two probability distributions over a common space, and f is a convex function on (0,∞) such that f(1) = 0. The f-divergence satisfies many properties similar to the KL divergence, such as non-negativity, symmetry (in some cases), and the data processing inequality. The choice of the function f determines the properties of the divergence and the types of differences that it measures between the distributions. Some commonly used f-divergences include the Jensen-Shannon divergence, the Pearson chi-square divergence, and the Neyman chi-square divergence.

46. Derive KL−divergence as a special case of $f-$divergence.

    The KL-divergence (also known as the Kullback-Leibler divergence) is a special case of the f-divergence when f(x) = x log(x) and the probability distributions are discrete.

    Let P = {p_1, p_2, ..., p_n} and Q = {q_1, q_2, ..., q_n} be two discrete probability distributions over the same sample space. The KL-divergence from Q to P is then defined as:

    $$KL(P||Q) = \sum_i p_i log(p_i/q_i)$$

    Now, let's substitute f(x) = x log(x) into the formula for f-divergence:

$$D_f(P||Q) = \int q(x) f(p(x)/q(x)) dx$$
$$D_f(P||Q) = \sum_i q_i f(p_i/q_i)$$
$$D_f(P||Q) = \sum_i q_i (p_i/q_i) log(p_i/q_i)$$
$$D_f(P||Q) = \sum_i p_i log(p_i/q_i)$$

which is exactly the formula for the KL-divergence from Q to P. Therefore, we can conclude that the KL-divergence is a special case of the f-divergence with the specific choice of f(x) = x log(x).

47. Define JS divergence.

    JS (Jensen-Shannon) divergence is a measure of similarity between two probability distributions. It is defined as half the sum of the KL-divergences between the two distributions and their average distribution: JS(P || Q) = 1/2 * (KL(P || M) + KL(Q || M)), where P and Q are the two probability distributions, M is their average distribution, and KL is the KL-divergence between two distributions. The JS divergence is always non-negative and is equal to zero if and only if the two distributions are identical.

    JS divergence can also be interpreted as a smoothed version of the KL divergence, where the divergence between P and Q is reduced if there is a third distribution M that is close to both P and Q. This makes JS divergence more robust than KL divergence when dealing with small differences between distributions or when comparing distributions with different supports.

48. Define Wasserstein divergence.

    Wasserstein divergence, also known as Earth Mover's Distance (EMD), is a measure of distance between two probability distributions. Unlike other divergence measures like KL divergence or JS divergence, which measure the dissimilarity between the shapes of the distributions, Wasserstein divergence focuses on the transportation cost required to transform one distribution into another.

    Formally, the Wasserstein divergence between two distributions P and Q on a metric space (X, d) is defined as:

    $$W(P, Q) = inf \int d(x, y) d\gamma(x, y) : \gamma \in \Pi(P, Q)$$

    where Π(P, Q) is the set of all joint probability distributions over X × X with marginals P and Q, and the infimum is taken over all such distributions. In other words,

Wasserstein divergence measures the minimum "cost" of transforming one distribution into another, where the cost is defined as the distance between the points in the two distributions multiplied by their probability of occurrence.

Wasserstein divergence has several advantages over other divergence measures, including its ability to handle distributions with different supports and its stability under small perturbations of the distributions. It has found applications in a variety of fields, including image processing, computer vision, and machine learning.

49. Define total variation divergence.

    Total variation (TV) divergence, also known as statistical distance or variation distance, is a measure of distance between two probability distributions. It is based on the idea of measuring the total amount of difference between the two distributions, in terms of the magnitude of their differences at each point in their support.

    Formally, the total variation divergence between two distributions P and Q on a discrete space X is defined as:

    $TV(P, Q) = 1/2 \sum_x \in X|P(x) - Q(x)|$

    or for continuous distributions, it is defined as: $TV(P, Q) = 1/2 \int |P(x) - Q(x)|dx$

    In other words, TV divergence measures the "total variation" or "distance" between two probability distributions, by summing up the absolute differences between the probabilities of each possible outcome or point in their support.

    Total variation divergence is a useful measure of distance between probability distributions, particularly in cases where the distributions are discrete or have a finite number of support points. It is often used in information theory, statistical inference, and machine learning applications.

50. Give an example where KL divergence can become infinite but Wasserstein distance may not.

    Consider two probability distributions P and Q defined as follows: P = {0: 1/2, 1: 1/2}, Q = {0: 0, 1: 1}

    The KL divergence between P and Q is given by: KL(P || Q) = ∑ p(x) log[p(x)/q(x)] = (1/2) log(1/0) + (1/2) log(1/1) = ∞

On the other hand, the Wasserstein distance between P and Q is given by: W(P, Q) = inf { ∑ c(i,j) d(i,j) } = inf { c(0,1) + c(1,0) } = 1, where c(i,j) is the cost of transporting mass from i to j and d(i,j) is the distance between i and j. In this case, the optimal transport plan is to move all the mass from 1 to 0, with a cost of 1.

Thus, we see that the KL divergence between P and Q is infinite, while the Wasserstein distance between them is finite. This example illustrates that the KL divergence can become infinite even when the distributions have overlapping support, while the Wasserstein distance is always well-defined and bounded.

51. What is smoothing in the context of KL divergence?

    In the context of KL divergence, smoothing refers to the process of adding a small positive constant to the probabilities of a discrete probability distribution. This is done to avoid the problem of having zero probability values, which can cause the KL divergence to become undefined or infinite.

    For example, suppose we have two discrete probability distributions P and Q, where P has some zero probabilities. If we directly calculate the KL divergence between P and Q, we may get an undefined or infinite value. However, by adding a small positive constant to each probability in P, we can ensure that there are no zero probabilities, and the KL divergence can be properly calculated. This process of adding a small constant is called smoothing.

    Smoothing is commonly used in machine learning and natural language processing applications, where discrete probability distributions are often encountered.

52. What could be negative effects of smoothing?

    While smoothing can be helpful in some cases, there are also potential negative effects that should be considered. Here are a few examples:

    a. Over-smoothing: If the smoothing parameter is set too high, it can result in over-smoothing, which means that the original probability distribution is significantly altered, and some important features may be lost. This can lead to poor performance in downstream tasks that rely on the probability distribution.

    b. Bias: Smoothing can introduce bias in the probability distribution by artificially increasing the probabilities of some events, which can result in a distortion of the true underlying probabilities.

c. Loss of information: Smoothing can also result in a loss of information, as it reduces the differences between the probabilities of different events. This can make it harder to distinguish between events that are actually different.

d. Sensitivity to smoothing parameter: The optimal smoothing parameter can be data-dependent, which means that it can be difficult to choose a good value for the parameter that works well across all datasets.

Overall, smoothing is a trade-off between reducing the noise in the probability distribution and preserving the underlying structure of the data. Careful consideration should be given to the specific problem at hand and the effects of smoothing on the performance of downstream tasks.

53. Show that Minimising KL-divergence is equivalent to MLE. This question is same as quiz1.

54. Name the paper where $f-$divergence was proposed.

    The paper where f-divergence was proposed is "Divergence measures based on the Shannon entropy" by Csiszár, Imre (1963).