Detection

Binary Hypothesis Training:

- Two hypotheses H_0 and H_1
- If H_0 occurs, X has PMF $P_{X|H0}(x)$
- If H_1 occurs, X has PMF $P_{X|H1}(x)$
- U_0 = {decide H_0 based on X}
- U_1 = {decide H_1 based on X}
- Error occurs if we decide the wrong hypothesis based on X

- Goal is to minimize the probability of error:

$$\mathsf{P}[\text{error}] = \mathsf{P}\big[U_1\big|H_0\big]\mathsf{P}[H_0] + \mathsf{P}\big[U_0\big|H_1\big]\mathsf{P}[H_1]$$

- Likelihood Ratio: $L(x) = \dfrac{P_{X|H_1}(x)}{P_{X|H_0}(x)}$.
- Log-Likelihood Ratio: $\ln\big(L(x)\big) = \ln\left(\dfrac{P_{X|H_1}(x)}{P_{X|H_0}(x)}\right)$.

Maximum Likelihood (ML) Rule):

- Intuition: Choose the hypothesis that best explains the observation.
- In terms of the conditional PMFs,

  Decide H_0 if $P_{X|H_0}(x) > P_{X|H_1}(x)$.
  Decide H_1 if $P_{X|H_0}(x) < P_{X|H_1}(x)$.
  Decide either if $P_{X|H_0}(x) = P_{X|H_1}(x)$.

  if diff

- In terms of the likelihood ratio,

  Decide H_0 if $L(x) < 1$.
  Decide H_1 if $L(x) > 1$.
  Decide either if $L(x) = 1$.

  if the same

- In terms of the log-likelihood ratio,

  Decide H_0 if $\ln\big(L(x)\big) < 0$.
  Decide H_1 if $\ln\big(L(x)\big) > 0$.
  Decide either if $\ln\big(L(x)\big) = 0$.

  with e

Maximum a Posteriori (MAP) Rule:

- Intuition: Choose the most likely hypothesis given the observation.
- Attains the minimum probability of error.
- In terms of the conditional PMFs,

  Decide H_0 if $P_{X|H_0}(x)\mathsf{P}[H_0] > P_{X|H_1}(x)\mathsf{P}[H_1]$.
  Decide H_1 if $P_{X|H_0}(x)\mathsf{P}[H_0] < P_{X|H_1}(x)\mathsf{P}[H_1]$.
  Decide either if $P_{X|H_0}(x)\mathsf{P}[H_0] = P_{X|H_1}(x)\mathsf{P}[H_1]$.

- In terms of the likelihood ratio,

  Decide H_0 if $L(x) < \frac{\mathsf{P}[H_0]}{\mathsf{P}[H_1]}$.
  Decide H_1 if $L(x) > \frac{\mathsf{P}[H_0]}{\mathsf{P}[H_1]}$.
  Decide either if $L(x) = \frac{\mathsf{P}[H_0]}{\mathsf{P}[H_1]}$.

- In terms of the log-likelihood ratio,

  Decide H_0 if $\ln\big(L(x)\big) < \ln\big(\frac{\mathsf{P}[H_0]}{\mathsf{P}[H_1]}\big)$.
  Decide H_1 if $\ln\big(L(x)\big) > \ln\big(\frac{\mathsf{P}[H_0]}{\mathsf{P}[H_1]}\big)$.
  Decide either if $\ln\big(L(x)\big) = \ln\big(\frac{\mathsf{P}[H_0]}{\mathsf{P}[H_1]}\big)$.

Estimation

- Two continuous random variables X and Y. Observe one and want to estimate other
- Here, we assume that we observe Y and estimate X using a function of Y, written as x”hat”(Y)
- Need to measure the error between the trus X and our estimate x”hat”(Y).

- In this class, we focused on the mean-squared error (MSE):

$$\mathsf{MSE} = \mathsf{E}\Big[\big(X - \hat{x}(Y)\big)^2\Big]$$

- **ML Estimator:** $\hat{x}_{\mathsf{ML}}(y) = \arg\max_x f_{Y|X}(y|x)$
- **MAP Estimator:** $\hat{x}_{\mathsf{MAP}}(y) = \arg\max_x f_{Y|X}(y|x)f_X(x)$

Minimum Mean-Squared Error (MMSE) Estimator:

- Attains the minimum MSE amongst all possible estimators
- Given by the conditional expectation:

  $\hat{x}_{\mathsf{MMSE}}(y) = \mathsf{E}\big[X|Y=y\big]$
- Unbiased: $\mathsf{E}\big[\hat{x}_{\mathsf{MMSE}}(Y)\big] = \mathsf{E}[X]$
- Orthogonality Principle: $\mathsf{E}\big[\big(X - \hat{x}_{\mathsf{MMSE}}(Y)\big)g(y)\big] = 0$ for any function $g(y)$ (including $\hat{x}_{\mathsf{MMSE}}(Y)$).

Linear Least-Squares Error (LLSE) Estimator:

- Attains the minimum MSE amongst all linear estimators.
- Two formulas: $\hat{x}_{\mathsf{LLSE}}(y) = \mathsf{E}[X] + \dfrac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}\big(y - \mathsf{E}[Y]\big)$

  $\hat{x}_{\mathsf{LLSE}}(y) = \mathsf{E}[X] + \rho_{X,Y}\dfrac{\sigma_X}{\sigma_Y}\big(y - \mathsf{E}[Y]\big)$
- Unbiased: $\mathsf{E}\big[\hat{x}_{\mathsf{LLSE}}(Y)\big] = \mathsf{E}[X]$.
- Orthogonality Principle: $\mathsf{E}\big[\big(X - \hat{x}_{\mathsf{LLSE}}(Y)\big)(aY+b)\big] = 0$ for any linear function $ay+b$ (including $\hat{x}_{\mathsf{LLSE}}(y)$).

Sums of Random Variables

More than Two Random Variables:

- Joint PMF for n discrete random variables $P_{X_1,\ldots,X_n}(x_1,\ldots,x_n)$
- Joint PDF for n continuous random variables $f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)$
- Expected Value of a Sum: $\mathsf{E}\big[X_1 + \cdots + X_n\big] = \sum_{i=1}^{n}\mathsf{E}[X_i]$
- Variance of a Sum: $\mathsf{Var}\big[X_1 + \cdots + X_n\big] = \sum_{i=1}^{n}\sum_{j=1}^{n}\mathsf{Cov}[X_i, X_j]$

Independent and Identically Distributed Random Var:

- Joint PMF $P_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = P_X(x_1)\cdots P_X(x_n)$
- Joint PDF $f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = f_X(x_1)\cdots f_X(x_n)$
- Expected Value of a Sum: $\mathsf{E}\big[X_1 + \cdots + X_n\big] = n\mathsf{E}[X]$
- Variance of a Sum: $\mathsf{Var}\big[X_1 + \cdots + X_n\big] = n\mathsf{Var}[X]$

Limit Theorems

- **Markov's Inequality:** For any non-negative random variable X and constant $c > 0$,

$$\mathsf{P}[X \geq c] \leq \frac{E[X]}{c}\ .$$

- **Chebyshev's Inequality:** For any random variable X and constant $c > 0$,

$$P\big[|X - \mathsf{E}[X]| \geq c\big] \leq \frac{\mathsf{Var}[X]}{c^2}\ .$$

- **Weak Law of Large Numbers:** Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables. For any constant $c > 0$,

$$\mathsf{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}X_i - \mathsf{E}[X]\right| \geq c\right) \leq \frac{\mathsf{Var}[X]}{nc^2}\ .$$

- **Hoeffding's Inequality:** Let $X_1, X_2, \ldots, X_n$ be i.i.d. Bernoulli(p) random varibles. For any constant $c > 0$,

$$\mathsf{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}X_i - \mathsf{E}[X]\right| \geq c\right) \leq 2e^{-2nc^2}\ .$$

Intro to Machine Learning

- The goal is to decide between two hypotheses, but we do not have access to the underlying probability model
- Instead we have a dataset D consisting of n sample

$$\mathcal{D} = \{(\vec{X}_1, Y_1), (\vec{X}_2, Y_2), \ldots, (\vec{X}_n, Y_n)\}$$

The i^{th} sample (X_i, Y_i) has an observation vector X_i and a label Y_i which we assume is -1 or +1

- We use this dataset to come up with a classifier h(X), which is a function that maps any possible observation vector into a guess of its label
- Ideally, given enough samples, we would approach the probability of error of the optimal MAP rule that knows the probability distribution

Training and test Error:

- To make sure we are not overfitting, we split our dataset into non-overlapping training and test datasets:

$$\mathcal{D}_{\text{train}} = \{(\vec{X}_{\text{train},1}, Y_{\text{train},1}), (\vec{X}_{\text{train},2}, Y_{\text{train},2}), \ldots, (\vec{X}_{\text{train},n_{\text{train}}}, Y_{\text{train},n_{\text{train}}})\}$$
$$\mathcal{D}_{\text{test}} = \{(\vec{X}_{\text{test},1}, Y_{\text{test},1}), (\vec{X}_{\text{test},2}, Y_{\text{test},2}), \ldots, (\vec{X}_{\text{test},n_{\text{test}}}, Y_{\text{test},n_{\text{test}}})\}.$$

- The training set is used to construct our classifier h(X) and the test set can only be used to evaluate its performance.

$$\text{Training Error} = \frac{1}{n_{\text{train}}} \sum_{i=1}^{n_{\text{train}}} g_{\text{not\_equal}}(h(\vec{X}_{\text{train},i}), Y_{\text{train},i}),$$

$$\text{Test Error} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} g_{\text{not\_equal}}(h(\vec{X}_{\text{test},i}), Y_{\text{test},i})$$

$$\text{where} \quad g_{\text{not\_equal}}(y_{\text{guess}}, y_{\text{true}}) = \begin{cases} 1 & \text{if } y_{\text{guess}} \neq y_{\text{true}} \\ 0 & \text{if } y_{\text{guess}} = y_{\text{true}} \end{cases}$$

Basic Classifiers:

- The **closest average** classifier first computes the average vector for each label, based on the training set. Given a new observation vector, it computes the distance to each average and choose the label with the smallest distance
- The **nearest neighbor** classifier, when given a new observation vector, computes the distance to every sample in the training set to find the closest point It then outputs the label of this point as its guess
- The **linear regression** classifier first computes a weighted sum of a new observation vector, and outputs the sign of this weighted sum and its guess. The training set is used to select the weights

Dimensionality Reduction :

- **Principal component analysis** allows us to reduce the dimensionality of our observations, by only keeping the (orthogonal) directions corresponding to the largest variance

Markov Chain

- Sequence of (discrete) random variables X_0, X_1, X_2... such that, given the history $X_0, \ldots, X_n$, the next state X_{n+1} only depends on the current state X_n This is sometimes called the Markov Property:

$$\mathsf{P}\big[X_{n+1} = j \big| X_n = i,\, X_{n-1} = i_{n-1}, \ldots,\, X_0 = i_0\big] = P_{ij} \quad \text{for all } n$$

- The *transition probabilities* P_{ij} are the probabilities of moving from state *i* to state *j* in one time step
- The n-step transition probabilities $P_{ij}(N)$ are the probabilities of moving from state *i* to state *j* in exactly *n* time steps. They can be determined via the **Chapman-Kolmogorov** equations:

$$P_{ij}(n+m) = \sum_{k=0}^{K} P_{ik}(n) P_{kj}(m)$$

State Transition Matrix:

- The state transition matrix is:

$$; \mathbf{P} = \begin{bmatrix} P_{00} & P_{01} & \cdots & P_{0K} \\ P_{10} & P_{11} & \cdots & P_{1K} \\ \vdots & \vdots & \ddots & \vdots \\ P_{K0} & P_{K1} & \cdots & P_{KK} \end{bmatrix}$$

- Row index is for the current state, column index is for the next state
- All rows must sum to 1

Probability State Vector:

- The probability state vector $\vec{\mathbf{p}}(n) = [p_0(n)\ p_1(n)\ \cdots\ p_K(n)]$ captures the PMF across the states at time n.
- $p_i(n) = \mathsf{P}[X_n = i]$
- The entries of $\vec{\mathbf{p}}(n)$ must sum to 1.
- Moving forward one time step: $\vec{\mathbf{p}}(n+1) = \vec{\mathbf{p}}(n)\mathbf{P}$.
- Moving forward n time steps: $\vec{\mathbf{p}}(n) = \vec{\mathbf{p}}(0)\mathbf{P}^n$.

State Classification:

- State j is accessible from state i if it is possible to reach state j starting from state i in one or more time steps. Notation: $i \to j$
- States i and j communicate if $i \to j$ and $j \to i$. Notation: $i \leftrightarrow j$
- A communicating class C is a subset of states such that if $i \in C$, then $j \in C$ if and only if $i \leftrightarrow j$.
- A Markov chain is irreducible if all of its states belong to a single communicating class.
- A state j is transient if there is a state k such that $j \to k$ but $k \nrightarrow j$.
- Any state that is not transient is recurrent.
- The period d of a state i is the greatest common divisor of the length of all cycles from i back to itself.
- If the period is 1, then the state is called aperiodic. The entire Markov chain is aperiodic if all states are aperiodic.

Limiting State Probability Vector:

- For an irreducible, aperiodic Markov chain, there is a unique limiting state probability vector
$$\vec{\boldsymbol{\pi}} = \lim_{n\to\infty} \vec{\mathbf{p}}(n)$$

Properties of the Limiting State Probability Vector:

- Normalization: $\sum_{i=0}^{K} \pi_i = 1$
- Any initial state probability vector $\vec{\mathbf{p}}(0)$ will converge to $\vec{\boldsymbol{\pi}}$.
- Steady-State Distribution: $\vec{\boldsymbol{\pi}} = \vec{\boldsymbol{\pi}}\mathbf{P}$.

Handling Transient States:

- If there is only one recurrent communicating class, then there is still a unique limiting state probability vector
$$\vec{\boldsymbol{\pi}} = \lim_{n\to\infty} \vec{\mathbf{p}}(n).$$
- If state *j* is transient, then $\pi_j = 0$

LAST EXAM STUFF

Important Families of Discrete Random Variables:

Bernoulli Random Variables:

- X is a Bernoulli(p) random variable if it has PMF
$$P_X(x) = \begin{cases} 1-p & x = 0, \\ p & x = 1\ . \end{cases}$$
- Range: $S_X = \{0, 1\}$.
- Expected Value: $\mathsf{E}[X] = p$.
- Variance: $\mathsf{Var}[X] = p(1-p)$.
- Interpretation: Single trial with success probability p.

Geometric Random Variables:

- X is a Geometric(p) random variable if it has PMF
$$P_X(x) = \begin{cases} p(1-p)^{x-1} & x = 1, 2, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$
- Range: $S_X = \{1, 2, \ldots\}$.
- Expected Value: $\mathsf{E}[X] = \frac{1}{p}$.
- Variance: $\mathsf{Var}[X] = \frac{1-p}{p^2}$.
- Interpretation: $\#$ of independent Bernoulli(p) trials until first success.

Binomial Random Variables:

- X is a Binomial(n, p) random variable if it has PMF
$$P_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0, 1, \ldots, n, \\ 0 & \text{otherwise.} \end{cases}$$
- Range: $S_X = \{0, 1, \ldots, n\}$.
- Expected Value: $\mathsf{E}[X] = np$.
- Variance: $\mathsf{Var}[X] = np(1-p)$.
- Interpretation: $\#$ of successes in n independent Bernoulli(p) trials.

Discrete Uniform Random Variables:

- X is a Discrete Uniform(k, ℓ) random variable if it has PMF
$$P_X(x) = \begin{cases} \frac{1}{\ell - k + 1} & x = k, k+1, \ldots, \ell, \\ 0 & \text{otherwise.} \end{cases}$$
- Range: $S_X = \{k, k+1, \ldots, \ell\}$.
- Expected Value: $\mathsf{E}[X] = \frac{k+\ell}{2}$.
- Variance: $\mathsf{Var}[X] = \frac{(\ell-k)(\ell-k+2)}{12} = \frac{(\ell-k+1)^2-1}{12}$.
- Interpretation: equally likely to take any integer value from k to ℓ.

Poisson Random Variables:

- X is a Poisson(α) random variable if it has PMF
$$P_X(x) = \begin{cases} \frac{\alpha^x}{x!} e^{-\alpha} & x = 0, 1, \ldots \\ 0 & \text{otherwise.} \end{cases}$$
- Range: $S_X = \{0, 1, \ldots\}$.
- Expected Value: $\mathsf{E}[X] = \alpha$.
- Variance: $\mathsf{Var}[X] = \alpha$.
- Interpretation: $\#$ of arrivals in a fixed time window.

Important Families of Continuous Random Variables:

Uniform Random Variables:

- X is a Uniform(a, b) random variable if it has PDF
$$f_X(x) = \begin{cases} \frac{1}{b-a} & a \le x < b \\ 0 & \text{otherwise.} \end{cases}$$
- CDF:
$$F_X(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \le x < b \\ 1 & b \le x \end{cases}$$
- Expected Value: E[X] = (a+b) / 2
- Variance: Var[X] = $(b-a)^2$ / 12

Exponential Random Variables:

- X is an Exponential(λ) random variable if it has PDF
$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \ge 0 \\ 0 & x < 0 \end{cases}$$
- CDF:
$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & x \ge 0 \\ 0 & x < 0 \end{cases}$$
- Expected Value: E[X] = 1/λ
- Variance: Var[X] = $1/\lambda^2$

Gaussian Random Variables:

- X is an Gaussian(μ,σ) random variable if it has PDF

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- CDF:

$$F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) \text{ where } \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{w^2}{2}}\, dw$$

  - Φ(z) is the standard normal CDF and Q(z) = 1-Φ(z) is the standard normal complementary CDF
  - Φ(-z) = 1-Φ(z) = Q(z)
  - Expected Value: E[X] = μ
  - Variance: Var[X] = σ^2

- $\text{Var}[X] = E[X^2] - E[X]^2$
- $\text{Var}[aX+b] = a^2\,\text{Var}[X]$
- $\sigma_X = \sqrt{\text{Var}[X]}$
- $E[X] = \int \mathscr{X} f_X(\mathscr{X})\, d\mathscr{X}$

- The covariance of random variables X and Y is

$$\text{Cov}[X,Y] = \text{E}\big[(X - \text{E}[X])(Y - \text{E}[Y])\big]$$

- Captures how X and Y vary together (linearly).
- Another useful formula is $\text{Cov}[X,Y] = \text{E}[XY] - \text{E}[X]\text{E}[Y]$
- Sometimes, $\text{E}[XY]$ is called the correlation. Don't confuse this with the correlation coefficient (next slide).

<u>Covariance of Linear Functions:</u> For any constants a,b,c, and d,

$$\text{Cov}[aX+b, cY+d] = a\,c\,\text{Cov}[X,Y] \ .$$

<u>Variance of Sums:</u>

$$\text{Var}[X+Y] = \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}[X,Y] \ .$$

<u>Variance of Weighted Sums:</u>

$$\text{Var}[aX+bY+c] = a^2\text{Var}[X] + b^2\text{Var}[Y] + 2ab\text{Cov}[X,Y] \ .$$

- The correlation coefficient is $\rho_{X,Y} = \dfrac{\text{Cov}[X,Y]}{\sqrt{\text{Var}[X]\text{Var}[Y]}}$

- $\text{P}[A \cap B] = \text{P}[A]\,\text{P}[B|A] = \text{P}[B]\,\text{P}[A|B].$

- The events A and B are conditionally independent given C if

$$\text{P}[A \cap B|C] = \text{P}[A|C]\,\text{P}[B|C] \ .$$

<u>Pairs of Continuous Random Variables:</u>

- Joint PDF: $f_{X,Y}(x,y) = \dfrac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y}$
- Range $S_{X,Y} = \{(x,y) : f_{X,Y}(x,y) > 0\}$.
- Marginal PDFs $f_X(x)$ and $f_Y(y)$ are just the PMFs of the individual random variables X and Y, respectively.

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy \qquad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx$$

- Conditional PDFs give the probability density of one random variable when the other is fixed to a certain value:

$$f_{X|Y}(x|y) = \begin{cases} \frac{f_{X,Y}(x,y)}{f_Y(y)} & f_Y(y) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

$$f_{Y|X}(y|x) = \begin{cases} \frac{f_{X,Y}(x,y)}{f_X(x)} & f_X(x) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

<u>EXAMPLE PROBLEMS</u>

T/F

(a) Let A and B be events. Then, $P[B|A] = P[A|B]\dfrac{1-P[B^c]}{1-P[A^c]}$.

True. By Bayes' Theorem and the complement property,

$$P[B|A] = \frac{P[A|B]P[B]}{P[A]} = P[A|B]\frac{1-P[B^c]}{1-P[A^c]}$$

.

(b) Let X and Y be a pair of uncorrelated random variables and let $Z = 2X - Y$. Then, $\text{Var}[Z] = 4\text{Var}[X] - \text{Var}[Y]$.

False. $\text{Var}[2X - Y] = 4\text{Var}[X] + \text{Var}[Y]$.

(c) Let X and Y be jointly Gaussian random variables with $E[X] = E[Y] = 1$, $\text{Var}[X] = \text{Var}[Y] = 5$, and $\rho_{X,Y} = 0$. Also, let $Z = X + Y$. Then, $P[Z \geq 2] = 1/2$.

True. Since Z is a linear function of jointly Gaussian random variables, it is also Gaussian. It has mean $E[Z] = E[X+Y] = E[X] + E[Y] = 2$. Since the Gaussian PDF is symmetric about its mean, $P[Z \geq 2] = 1/2$.

(d) Suppose we have a Markov chain with two communicating classes. Then, there is no unique limiting state probability vector.

False. If one of the classes is transient, there will still be a unique limiting state probability vector.

(e) Let X be a zero-mean random variable with variance 1. Then, $P[X^2 \geq 2] \leq 1/2$.

True. Define $Y = X^2$ and note that $E[Y] = E[X^2] = \text{Var}[X] + (E[X])^2 = 1$. Using Markov's inequality, $P[Y \geq 2] \leq \frac{E[Y]}{2} = 1/2$.

(f) Let X be a non-negative random variable and $Y = \ln(X)$. Then, $F_Y(y) = F_X(e^y)$.

True. $F_Y(y) = P[Y \leq y] = P[\ln(X) \leq y] = P[X \leq e^y] = F_X(e^y)$.

(a) Let X and Y be independent random variables, each with mean 1 and variance 1. Then, $E[X^2Y^2] = 4$.

True. $\text{E}[X^2] = \text{Var}[X] + (\text{E}[X])^2 = 1+1 = 2$, and $\text{E}[Y^2] = \text{Var}[Y] + (\text{E}[Y])^2 = 1+1 = 2$. Using independence, $\text{E}[X^2Y^2] = \text{E}[X^2]\text{E}[Y^2]$, so the result follows.

(b) Let X and Y be continuous random variables. Then, for any constants $a > 0$ and $b > 0$, $P[\{X > a\} \cap \{Y > b\}] = \left(\int_a^\infty f_X(x)\,dx\right) \cdot \left(\int_b^\infty f_Y(y)\,dy\right)$

False. The probability of interest can be written as $\int_a^\infty \int_b^\infty f_{X,Y}(x,y)\,dy\,dx$, but this cannot be simplified to the given expression unless X and Y are independent.

(c) Let X and Y be jointly Gaussian random variables with $E[X] = E[Y] = 0$, $\text{Var}[X] = \text{Var}[Y] = 1$, and $\rho_{X,Y} = 0$. Also, let $U = X + 2Y$ and $V = 2Y - X$. Then, U and V are independent.

False. Note that $\text{E}[U] = 0$, $\text{E}[V] = 0$, and

$$\text{E}[UV] = \text{E}[(X+2Y)(2Y-X)] = \text{E}[4Y^2 - X^2] = 4 - 1 = 3,$$

so $\text{Cov}[U,V] = \text{E}[UV] - \text{E}[U]\text{E}[V] = 3 \neq 0$. Correlated Gaussian random variables are not independent.

(d) Let X and Y be a pair of uncorrelated random variables ($\text{Cov}[X,Y] = 0$). Then, $E[(X - \hat{x}_{\text{LLSE}}(Y))^2] = \text{Var}[X]$.

True. Since $\text{Cov}[X,Y] = 0$, $\hat{x}_{\text{LLSE}}(Y) = \text{E}[X]$. Substituting this, we are left with the definition of variance.

(e) For a Markov chain and any states i and j, $P[X_2 = j|X_0 = i] = P[X_2 = j|X_1 = i]$.

False. A two-step transition probability is not generally the same as a one-step transition probability.

(f) For the MAP decision rule, $P[\text{error}] \leq \min(P[H_0], P[H_1])$.

True. $\text{P[error]} = \text{P}[H_0]$ is achieved by always deciding H_1, and $\text{P[error]} = \text{P}[H_1]$ is achieved by always deciding H_0, so the MAP decision rule (which minimizes the probability of error) will do at least as well as the minimum.

(g) If X and Y are independent, $\hat{x}_{\text{MMSE}}(y) = 0$.

False. X and Y being independent makes $\hat{x}_{\text{MMSE}}(y) = \text{E}[X]$, which is not necessarily zero.

(h) If $E[X] = E[Y] = 0$, then $E[XY] \leq \sqrt{\text{Var}[X]\text{Var}[Y]}$.

True. This follows from $\rho_{X,Y} \leq 1$.

6.1 Binary Hypothesis Testing

$P_e = \mathbb{P}[\{\text{error}\}] = \mathbb{P}[\{\text{error}\} \mid H_0]\,\mathbb{P}[H_0] + \mathbb{P}[\{\text{error}\} \mid H_1]\,\mathbb{P}[H_1]$

$\quad = \mathbb{P}[\{y \in A_1\} \mid H_0]\,\mathbb{P}[H_0] + \mathbb{P}[\{y \in A_0\} \mid H_1]\,\mathbb{P}[H_1]$

<u>Error</u>:

Discrete: $\mathbb{P}[\{y \in A_1\} \mid H_0] = \sum_{y \in A_1} P_{Y|H_0}(y)$

Continuous: $\mathbb{P}[\{y \in A_1\} \mid H_0] = \int_{A_1} f_{Y|H_0}(y)\,dy$

Probability of false alarm: $P_{FA} = \mathbb{P}[\{y \in A_1\} \mid H_0]$

Probability of missed detection: $P_{MD} = \mathbb{P}[\{y \in A_0\} \mid H_1]$

so: $P_e = P_{FA}\,\mathbb{P}[H_0] + P_{MD}\,\mathbb{P}[H_1]$

Maximum Likelihood (ML) Rule: selects H w/ highest likelihood value

Discrete:

- $D^{ML} = \begin{cases} 1, & P_{Y|H_1}(y) \geq P_{Y|H_0}(y) \\ 0, & P_{Y|H_1}(y) < P_{Y|H_0}(y) \end{cases}$

Continuous:

- $D^{ML} = \begin{cases} 1, & f_{Y|H_1}(y) \geq f_{Y|H_0}(y) \\ 0, & f_{Y|H_1}(y) < f_{Y|H_0}(y) \end{cases}$

Maximum a posteriori (MAP) rule: selects H that is most likely given observation

Discrete:

- $D^{MAP} = \begin{cases} 1, & P_{Y|H_1}(y)\,\mathbb{P}[H_1] \geq P_{Y|H_0}(y)\,\mathbb{P}[H_0] \\ 0, & P_{Y|H_1}(y)\,\mathbb{P}[H_1] < P_{Y|H_0}(y)\,\mathbb{P}[H_0] \end{cases}$

Continuous:

- $D^{map}(y) = \begin{cases} 1, & f_{Y|H_1}(y)\,\mathbb{P}[H_1] \geq f_{Y|H_0}(y)\,\mathbb{P}[H_0] \\ 0, & f_{Y|H_1}(y)\,\mathbb{P}[H_1] < f_{Y|H_0}(y)\,\mathbb{P}[H_0] \end{cases}$

* MAP Rule is optimal

6.2: Likelihood Ratio

Likelihood Ratio $L(y)$:

- $L(y) = \begin{cases} \dfrac{P_{Y|H_1}(y)}{P_{Y|H_0}(y)} & y \text{ is discrete} \\ \dfrac{f_{Y|H_1}(y)}{f_{Y|H_0}(y)} & y \text{ is continuous} \end{cases}$

$D^{ML}: \begin{cases} 1, & L(y) \geq 1 \\ 0, & L(y) < 1 \end{cases} = \begin{cases} 1, & \ln(L(y)) \geq 0 \\ 0, & \ln(L(y)) < 0 \end{cases}$

$D^{map}: \begin{cases} 1, & L(y) \geq \frac{\mathbb{P}[H_0]}{\mathbb{P}[H_1]} \\ 0, & L(y) < \frac{\mathbb{P}[H_0]}{\mathbb{P}[H_1]} \end{cases} = \begin{cases} 1, & \ln(L(y)) \geq \ln\left(\frac{\mathbb{P}[H_0]}{\mathbb{P}[H_1]}\right) \\ 0, & \ln(L(y)) < \ln\left(\frac{\mathbb{P}[H_0]}{\mathbb{P}[H_1]}\right) \end{cases}$

* Remember for Gaussian:

$F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$ *standard normal CDF

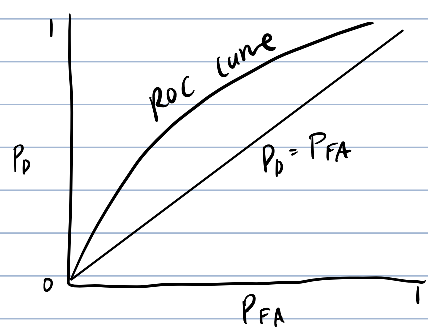
$Q(z) = 1 - \Phi(z) = \Phi(-z)$

Probability of Detection

- $P_D = \mathbb{P}[\{y \in A_1\} \mid H_1] = 1 - \mathbb{P}[\{y \in A_0\} \mid H_1] = 1 - P_{MD}$

ROC Curve:

1. Find $L(y)$
2. Create a likelihood ratio test w/ Threshold T: $D_T(y) = \begin{cases} 1, & L(y) \geq T \\ 0, & L(y) < T \end{cases}$
3. Starting from $T=0$ to $T=\infty$, calculate $A_1(T)$, $P_{FA}(T)$, $P_D(T)$
4. Plot resulting $(P_{FA}(T), P_D(T))$ pairs



properties:

- concave
- Lies on or above $P_D = P_{FA}$ line

6.3 Detection w/ unequal costs

Minimum Bayes Risk Detection:

→ cost C_{ij}

- C_{10} = cost of false alarm
- C_{01} = cost of missed detection

Expected Risk:

- Discrete: $\mathbb{E}[R] = \sum_{y \in A_0} C_{01} P_{Y|H_1}(y) \mathbb{P}[H_1] + \sum_{y \in A_1} C_{10} P_{Y|H_0}(y) \mathbb{P}[H_0]$

- Continuous: $\mathbb{E}[R] = \int_{A_0} C_{01} f_{Y|H_1}(y) \mathbb{P}[H_1]\, dy + \int_{A_1} C_{10} f_{Y|H_0}(y) \mathbb{P}[H_0]\, dy$

Minimum Baye's Risk decision Rule:

$$D^{MBR}(y) = \begin{cases} 1, & L(y) \geq \dfrac{C_{10} \mathbb{P}[H_0]}{C_{01} \mathbb{P}[H_1]} \\ 0, & L(y) < \dfrac{C_{10} \mathbb{P}[H_0]}{C_{01} \mathbb{P}[H_1]} \end{cases}$$

M-ary Hypotheses: $H_0, H_1, \dots, H_{M-1}$

* $\arg\max(-g(y)) = \arg\min(g(y))$

$P_e = \sum_{i=0}^{M-1} \mathbb{P}[\{y \notin A_i\} | H_i] \mathbb{P}[H_i]$

- $D^{ML} = \arg\max_{i \in \{0,1,\dots M-1\}} P_{Y|H_i}(y)$ (discrete)
- $D^{ML} = \arg\max_{i \in \{0,1,\dots M-1\}} f_{Y|H_i}(y)$ (continuous)

$D^{MAP} = \arg\max_{i \in \{0,1,\dots,M-1\}} P_{Y|H_i}(y) \mathbb{P}[H_i]$

$D^{MAP} = \arg\max_{i \in \{0,1,\dots M-1\}} f_{Y|H_i}(y) \mathbb{P}[H_i]$

Requires knowing $\mathbb{P}[H_i]$

7.1: ML, MAP, and MMSE Estimation

Mean-Squared Error: X is unobserved, y is observed

- $MSE = \mathbb{E}[(X - \hat{x}(y))^2]$ ← quality of prediction
  - $\hat{x}(y)$ = estimation rule
- Estimator $\hat{x}(y)$ is unbiased if the error $x - \hat{x}(y)$ has zero mean: $\mathbb{E}[X - \hat{x}(y)] = 0$

ML Estimator: $\hat{x}_{ML}(y)$

y discrete: $\hat{x}_{ML} = \arg\max_{x \in R_x} P_{Y|X}(y|x)$ y continuous: $\hat{X}_{ML}(y) = \arg\max_{x \in R_x} f_{Y|X}(y|x)$

MAP estimator: $\hat{x}_{MAP}(y)$

X, y discrete: $\arg\max_{x \in R_x} P_{Y|X}(y|x) P_X(x) = \arg\max_{x \in R_x} P_{X|Y}(x|y)$

X, y Jointly Continuous: $\arg\max_{x \in R_x} f_{Y|X}(y|x) f_X(x) = \arg\max_{x \in R_x} f_{X|Y}(x|y)$

X Discrete, y Continuous: $\arg\max_{x \in R_x} f_{Y|X}(y|x) P_X(x)$

X Continuous, y Discrete: $\arg\max_{x \in R_x} P_{Y|X}(y|x) f_X(x)$

Unbiased? $\mathbb{E}[\hat{X}_{ML}(y)] = \mathbb{E}[\mathbb{E}[\hat{x}_{ML}(y) | X]] = \mathbb{E}[X]$ ✓ yes

* In general, ML & MAP estimates are biased :(

MMSE estimator: OPTIMAL

* Atains smallest possible MSE

$$\hat{X}_{MMSE}(y) = \mathbb{E}[X | Y=y] \to \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\, dx$$

- unbiased
- Error of MMSE estimator is orthogonal to any function $g(y)$ of the observation

$$\mathbb{E}[(X - \hat{x}_{MMSE}(y)) g(y)] = 0$$

Jointly Gaussian Random Vars

If X and y are jointly Gaussian, MMSE estimator is a linear function of y

$$\hat{x}_{MMSE}(y) = \mathbb{E}[X|Y=y] = \mu_x + \rho_{x,y}\frac{\sigma_x}{\sigma_y}(y-\mu_y) = \mathbb{E}[X] + \frac{Cov[X,Y]}{Var[Y]}(y-\mathbb{E}[Y])$$

$$MSE_{MMSE} = \mathbb{E}[(X-\hat{x}_{MMSE}(y))^2] = (1-\rho_{x,y}^2)\sigma_x^2 = Var[X] - \frac{(Cov[X,Y])^2}{Var[Y]}$$

7.2: LLSE and Vector Estimation

The linear least squares error $\hat{x}_{LLSE}(y)$ attains the smallest possible mean-squared error among all linear estimators

$$\hat{x}_{LLSE}(y) = \mu_x + \rho_{x,y}\frac{\sigma_x}{\sigma_y}(y-\mu_y) = \mathbb{E}[X] + \frac{Cov[X,Y]}{Var[Y]}(y-\mathbb{E}[Y])$$

$$MSE_{LLSE} = \sigma_x^2(1-\rho_{x,y}^2) = Var[X] - \frac{(Cov[X,Y])^2}{Var[Y]}$$

* unbiased: $\mathbb{E}[\hat{x}_{LLSE}(y)] = \mathbb{E}[X]$
* Error is orthogonal to any linear function $ay+b$: $\mathbb{E}[(X-\hat{x}_{LLSE}(Y))(aY+b)] = 0$

$$\hat{\underline{x}}_{LLSE}(\underline{Y}) = \mathbb{E}[\underline{X}] + \Sigma_{\underline{x},\underline{y}}\Sigma_{\underline{y}}^{-1}(\underline{y}-\mathbb{E}[\underline{Y}])$$

Cross Covariance matrix $\underline{X}, \underline{Y}$: $\Sigma_{\underline{x},\underline{y}} = \mathbb{E}[(\underline{X}-\mathbb{E}[\underline{X}])(\underline{Y}-\mathbb{E}[\underline{Y}])^T] = \begin{bmatrix} Cov[X_1,Y_1] & \cdots & Cov[X_1,Y_m] \\ \vdots & \ddots & \vdots \\ Cov[X_n,Y_1] & \cdots & Cov[X_n,Y_m] \end{bmatrix}$

Recall LLSE estimator Properties:

* unbiased
* $\mathbb{E}[(\underline{X}-\hat{\underline{x}}_{LLSE}(\underline{Y}))(A\underline{Y}+b)^T] = 0$ ← all zeros matrix

8.1 Sums of Random Variables: Bounds

· Random variables $X_1, X_2, \dots X_n$ are independent and identically distributed (i.i.d.) if they are independent and have the same PMF and PDF

Discrete PMF: $\prod_{i=1}^{n} P_X(x_i)$ Continuous PDF: $\prod_{i=1}^{n} f_X(x_i)$

· $\mathbb{E}[S_n] = \mathbb{E}[\sum_{i=1}^{n} X_i] = n\mathbb{E}[X]$

· $Var[S_n] = Var[\sum_{i=1}^{n} X_i] = nVar[X]$ * compute using marginal PMF $P_X(x)$ or PDF $f_X(x)$

$$M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$\mathbb{E}[M_n] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \mathbb{E}[X]$ $Var[M_n] = Var[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}Var[X]$

Markov's Inequality:

$\mathbb{P}[\{X \geq c\}] \leq \mathbb{E}[X]/c$

Chebyshev's Inequality:

$\mathbb{P}[|X-\mathbb{E}[X]| \geq c] \leq Var[X]/c^2$

Hoeffding Inequality

$\mathbb{P}[|\sum_{i=1}^{n} X_i - \mathbb{E}[\sum_{i=1}^{n} X_i]| \geq c] \leq 2\exp(\frac{-2c^2}{n(b-a)^2})$

8.2: Sums of Random Variables: Limits

· Sample Mean: $M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

· Weak Law of Large Numbers: $\lim_{n\to\infty}\mathbb{P}[|M_n-\mu| > \varepsilon] = 0$

· Strong Law of Large Numbers: $\mathbb{P}[\lim_{n\to\infty} M_n = \mu] = 1$

· Central Limit Theorem: $\lim_{n\to\infty} F_{Y_n}(y) = \Phi(y)$ where $Y_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$

CDF for Gaussian(0,1)

* sum of many small independent effect look gaussian eventually

9.1 Confidence Intervals

Sample Variance V_n = $V_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_n)^2$ *unbiased estimator of variance

· Confidence level : $1-\alpha$ satisfies

$\mathbb{P}[A \le \mu \le B] = 1-\alpha$ (Confidence Interval : $[A, B]$)

Assumptions:

① Known Variance :

Confidence Interval : $[M_n - \varepsilon, M_n + \varepsilon]$

Confidence Level : $1 - \sigma^2/n\varepsilon^2$

② CLT and known Variance :

$CI = [M_n + \gamma\sigma/\sqrt{n}, \; M_n + \gamma\sigma/\sqrt{n}]$

$CL = 1-\alpha$

*find γ so that $Q(\gamma) = \alpha/2$

* $\varepsilon = \gamma\sigma/\sqrt{n}$

* always round up

Chi-square RV w/ n degrees-of-freedom =

$Y \sim \chi^2(n)$: $Y = \sum_{i=1}^{n} X_i^2$ where $X_i \to X_n$ i.i.d. $N(0,1)$

(sum of squares of n independent standard Gaussian RVs)

→ Mean: $\mathbb{E}[Y] = n$

→ Var: $Var[X] = 2n$ CDF: lookup table

Student's t-distribution with n-degrees of freedom :

$W \sim T(n)$: where $Z \sim N(0,1)$, $Y \sim \chi^2(n)$ and Y, Z independent

$W = Z/\sqrt{Y/n}$

→ Mean: $\mathbb{E}[W] = 0$

→ $Var[W] = n/n-2$ for $n>2$ (∞ for $n=1,2$)

Assumptions:

③ Gaussian w/ unknown variance

$CI = [M_n - \beta\sqrt{V_n}/\sqrt{n}, \; M_n + \beta\sqrt{V_n}/\sqrt{n}]$ } Student's t-distribution

$CL: 1-\alpha$

*Find β so that $F_W(\beta) = 1 - \alpha/2 = 0.95$

④ Find Gaussian w/ unknown mean and variance

$CI: \left[\frac{(n-1)V_n}{t_2}, \frac{(n-1)V_n}{t_1}\right]$ } Chi-square distribution

$CL: 1-\alpha$

* set t_1 so $F_Y(t_1) = \alpha/2$ and t_2 so $F_Y(t_2) = 1 - \alpha/2$

9.2 Significance Testing

Rejection Region : R_0: $\mathbb{P}[Y \in R_0 | H_0] = \int_{R_0} f_{Y|H_0}(y)\,dy = \alpha$

· If $Y \in R_0$, reject the null hypothesis

· If $Y \notin R_0$, fail to reject

Threshold Tests:

① 1-sided Test, Right Tail : $R_0 = \{y > t_u\}$ $\mathbb{P}[Y > t_u | H_0] = \alpha$

② 1-sided Test, Left Tail : $R_0 = \{y < t_e\}$ $\mathbb{P}[Y < t_e | H_0] = \alpha$

③ 2-sided Test : $R_0 = \{y > t_u\} \cup \{y < t_e\}$

$\mathbb{P}[Y < t_e | H_0] = \mathbb{P}[Y > t_u | H_0] = \alpha/2$

* Reject null hypothesis if p-value < α

① 1-sided test, Right Tail : p-value = $1 - F_{Y|H_0}(y)$

② 1-sided test, Left Tail : p-value = $F_{Y|H_0}(y)$

③ Two-sided Test : p-value = $2\min(F_{Y|H_0}(y), 1 - F_{Y|H_0}(y))$

Four different Tests:

① **One-sample Z test:** Observations are i.i.d. Gaussian (μ, θ) under H_0.
μ & θ^2 are known

Test: if mean has changed.

Two Sided Test: ① Compute sample mean $M_n = \frac{1}{n}\sum_{i=1}^{n} x_i$

② Compute Z-stat: $Z = \frac{\sqrt{n}(M_n - \mu)}{\theta}$

③ Compute p-value: $2\Phi(-|Z|)$

④ Pval $< \alpha$: Reject

Pval $\geq \alpha$: Fail to reject

② **One-sample T-test:** Observation i.i.d. Gaussian w/ known μ and unknown θ^2

Test: if mean has changed

Two-sided test: ① Compute sample mean

② Compute sample variance $V_n = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - M_n)^2$

③ Compute the t-statistic $T = \frac{\sqrt{n}(M_n - \mu)}{\sqrt{V_n}} \sim T(n-1)$

④ Compute p-val: $2F_T(-|T|)$

⑤ Reject: pval $< \alpha$

Fail to reject: pval $\geq \alpha$.

③ **Two-sample Z test:** Collect two datasets: $X_1, \dots X_n$ is i.i.d Gaussian(μ, θ^2) and $Y_1 \dots Y_n$ is i.i.d. Gaussian w/ the same mean $\mu_1 = \mu_2$

Means unknown variances known

Test: if means are different $\mu_1 \neq \mu_2$

Two-sided Test: ① Compute sample means

② Compute Z-stat: $\frac{M_n^{(1)} - M_n^{(2)}}{\sqrt{\theta_1^2/n_1 + \theta_2^2/n_2}} \sim N(0,1)$ under H_0

③ Compute Pval: $2\Phi(-|Z|)$

④ Reject: Pval $< \alpha$

Fail to Reject: Pval $\geq \alpha$

④ **Two-sample T-test:** Same as sample Z, except $\mu_1 = \mu_2$ and vars $\theta_1^2 = \theta_2^2$ (known μ and θ^2)

Test: if means are different $\mu_1 \neq \mu_2$

Two sided Test: ① Compute sample means

② Compute sample vars

③ Compute: $\hat{\theta} = \sqrt{\frac{(n_1-1)V_{n_1}^{(1)} + (n_2-1)V_{n_2}^{(2)}}{n_1 + n_2 - 2}}$

④ Compute T stat: $T = \frac{M_{n_1} - M_{n_2}}{\hat{\theta}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

⑤ Compute pval: $2F_T(-|T|)$

⑥ Reject pval $< \alpha$, Fail pval $\geq \alpha$