Homework #1 Solutions

1. Exercise 1.21 (a - e) in Gallager

Solution:

(a) 
\[
E[(X - \bar{X})(Y - \bar{Y})] = E[XY] - \bar{X}E[Y] - \bar{Y}E[X] + \bar{X}\bar{Y}
\]
\[X,Y\text{uncorrelated} \Rightarrow E[XY] = \bar{X}\bar{Y}
\]
\[XY - \bar{X}\bar{Y} = 0
\]

(b) 
\[
E[X + Y] = \bar{X} + \bar{Y}
\]
\[
Var(X + Y) = E[(X + Y - \bar{X} - \bar{Y})^2]
\]
\[= E[(X - \bar{X})^2 + (Y - \bar{Y})^2] + 2E[(X - \bar{X})(Y - \bar{Y})]
\]
\[= Var(X) + Var(Y)
\]

(c) Proof by induction:
In (b.) we showed that the assumption is true for \(n = 2\).
We’ll assume that the assumption is right for \(n\):
\[
Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i)
\]
and show that in the assumption holds for \(n+1\): \(Z = \sum_{i=1}^{n} X_i\)
\[
Var\left(\sum_{i=1}^{n+1} X_i\right) = Var(Z + X_{n+1})
\]
\[= Var(Z) + Var(X_{n+1})
\]
\[\text{induction assumption}\]
\[
= \sum_{i=1}^{n} Var(X_i) + Var(X_{n+1}) = \sum_{i=1}^{n+1} Var(X_i)
\]

(d) Discrete case:
\[
E[XY] = \sum_{i,j=1}^{\infty} XY P(X = x_i, Y = y_j)
\]
\[X,Y\text{independent} \Rightarrow \sum_{i,j=1}^{\infty} XY P(X = x_i)P(Y = y_j)
\]
\[= \sum_{i=1}^{\infty} XP(X = x_i) + \sum_{j=1}^{\infty} Y P(Y = y_j) = E[X]E[Y]
\]
2. A criterion for almost sure convergence.

Let \((X_n, n \geq 0)\) be a sequence of random variables. Show that it convergence a.s. to \(X\)

\[ P\{\omega \in \Omega : \lim_{n \to \infty} X_n(\omega) = X(\omega)\} = 1 \]

if and only if

\[ \forall \epsilon > 0 : \lim_{m \to \infty} P\{|X_n - X| < \epsilon, \forall n \geq m\} = 1 \]

**Solution:** In this problem we will view probability in terms of subsets of the sample space. Recall that each random variable can be viewed as some function of elements of the sample space: \(X = X(w), w \in \Omega\). Then, saying that \(P\{X > 3\} = P(B)\), where \(B\) is a subset of \(\Omega : B = \{w \in \Omega : X(w) > 3\}\). Usually it makes no sense to always mention the sample space, but if one is working with quantifiers, this comes handy. Further in this problem, instead of exactly mentioning \(w\), we will state the subsets in terms of conditions on random variables: \(B = \{X > 3\}\) would represent both the condition on \(X\) and the subset of \(w\), for which the condition holds. The two most important examples are the condition that involve \(\forall\) and \(\exists\): assuming \(B_m, \forall m\) are some subsets of the sample space, \(P(\forall m B_m \text{ holds}) = P(\cap_{m=1}^{\infty} B_m)\) and \(P(\exists m : B_m \text{ holds}) = P(\cup_{m=1}^{\infty} B_m)\). You should take some time to make yourself used to this.

a. \(P\{\lim_{n \to \infty} X_n = X\} = 1 \Rightarrow \forall \epsilon > 0 : \lim_{m \to \infty} P\{|X_n - X| < \epsilon, \forall n \geq m\} = 1\)

Assume that the LHS is true. First, let’s define an event \(B_{\epsilon,m} = \{\forall n \geq m : |X_n - X| < \epsilon\}\) and \(B_\epsilon = \cup_{m=1}^{\infty} B_{\epsilon,m}\) then, by the definition of the limit:

\[ \lim_{n \to \infty} X_n = X \Leftrightarrow \forall \epsilon > 0 \exists m : B_{\epsilon,m} \text{ holds} \Leftrightarrow \cap_{\epsilon > 0} B_\epsilon \text{ holds} \]

Let’s fix any \(\delta > 0\), then \(\cap_{\epsilon > 0} B_\epsilon \subset B_\delta\) and

\[ 1 = P\{\lim_{n \to \infty} X_n = X\} = P\{\cap_{\epsilon > 0} B_\epsilon\} \leq P\{B_\delta\} \]

Therefore, \(P\{B_\delta\} = 1\). To find the probability of \(B_\delta\) in terms of \(B_{\delta,m}\), which is exactly the probability stated in the right hand side of the criterion, we will use the additivity of probability.

\[ E[XY] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} XY f_{X,Y}(X = x, Y = y) dx dy \]

\[ = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} XY f_X(X = x) f_Y(Y = y) dx dy \]

\[ = \int_{-\infty}^{\infty} X f_X(X = x) dx + \int_{-\infty}^{\infty} Y f_Y(Y = y) dy = E[X]E[Y] \]

(e) If \(P(X = 1) = P(X = -1) = 1/4, P(X = 0) = 1/2\)
and \(P(Y = -X) = P(Y = X) = 1/2\) than:

\[ P(Y = 1) = P(Y = -1) = 1/4, P(Y = 0) = 1/2 \]

\[ E[XY] = \sum_{x,y=-1}^{1} XY P(X = x, Y = y) = 0 \Rightarrow X,Y \text{uncorrelated} \]

\[ P(X = 0|Y) \neq P(X = 0) \Rightarrow X,Y \text{ are dependent} \]
for events $\hat{B}_{\delta,m} = B_{\delta,m} \setminus B_{\delta,m-1}$, $\hat{B}_{\delta,1} = B_{\delta,1}$. The reason for introducing the events is that they do not intersect: $\hat{B}_{\delta,i} \cap \hat{B}_{\delta,j} = \emptyset$, so by the probability axioms $P\{\bigcup \hat{B}_{\delta,m}\} = \sum P\{\hat{B}_{\delta,m}\}$. Also note that $B_{\delta,m} \subseteq B_{\delta,m-1}$, so $B_{\delta,m} = \bigcup_{i=1}^{m} \hat{B}_{\delta,i}$.

$$P\{B_{\delta}\} = P\{\bigcup_{m=1}^{\infty} \hat{B}_{\delta,m}\} = \sum_{m=1}^{\infty} P\{\hat{B}_{\delta,m}\} = \lim_{m \to \infty} \sum_{i=1}^{m} P\{\hat{B}_{\delta,i}\} = \lim_{m \to \infty} P\{B_{\delta,m}\}$$

which finishes the proof of the first part as we get for $\forall \delta > 0$:

$$1 = P\{B_{\delta}\} = \lim_{m \to \infty} P\{B_{\delta,m}\} = \lim_{m \to \infty} P\{\forall n \geq m : |X_n - X| < \delta\}$$

b. $\forall \epsilon > 0 : \lim_{m \to \infty} P\{|X_n - X| < \epsilon, \forall n \geq m\} = 1 \Rightarrow P\{\lim_{n \to \infty} X_n = X\} = 1$

Assume the LHS is true. Let’s find $P\{\lim_{n \to \infty} X_n \neq X\}$:

$$P\{\lim_{n \to \infty} X_n \neq X\} = P\{\exists \epsilon > 0 : B_{\epsilon} \text{ doesn’t hold}\} = P\{\cup_{\epsilon > 0} B_{\epsilon}^c\}$$

Now let’s again fix some $\delta > 0$, by the LHS $P\{B_{\delta}\} = 1$, so we can upper bound the union by any element:

$$P\{\lim_{n \to \infty} X_n \neq X\} = P\{\cup_{\epsilon > 0} B_{\epsilon}^c\} \leq P\{B_{\delta}^c\} = 1 - P\{B_{\delta}\} = 0,$$

so $P\{\lim_{n \to \infty} X_n \neq X\} = 0$ and $P\{\lim_{n \to \infty} X_n = X\} = 1$, which completes the proof.

3. Whitening.

Let $U_1$, $U_2$, and $U_3$ be independent random variables, where $U_i$ has mean zero and variance $\sigma_{u_i}^2 = i$ for $i = 1, 2, 3$. Define

$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$$

such that

$$X_i = \sum_{j=1}^{i} U_j, \quad i = 1, 2, 3.$$ 

a. Find the covariance matrix of $X : K_X$.

b. Find a decorrelating linear transformation for $X$

**Solution:**

a. The covariance matrix

$$\Sigma_X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 3 & 3 \\ 1 & 3 & 6 \end{bmatrix}$$

b. if $Y = AX \Rightarrow \Sigma_Y = A\Sigma_X A^T$

$\Sigma_X = V \Sigma_A V^T$ where $\Sigma_A$ is diagonal matrix and $V$ is a unitary matrix.

Therefor if we choose $A = V^T$ we’ll get:

$$\Sigma_Y = A\Sigma_A V^T A^T = V^T V \Sigma_A V^T V = \Sigma_A$$
Where:

\[
\mathbf{V} = \begin{bmatrix}
0.8628 & 0.4684 & 0.1902 \\
-0.4927 & 0.6950 & 0.5237 \\
0.1131 & -0.5455 & 0.8304 \\
\end{bmatrix}
\]

\[
\Sigma_\Lambda = \begin{bmatrix}
0.5601 & 0 & 0 \\
0 & 1.3192 & 0 \\
0 & 0 & 8.1207 \\
\end{bmatrix}
\]

4. Law Of Large Numbers. The Law of Large Numbers is said to hold for a sequence of random variables \( S_1, S_2, S_3, S_4, \ldots \) if for every \( \epsilon > 0 \),

\[
\lim_{n \to \infty} \Pr(\left| \frac{1}{n} S_n - E[\frac{1}{n} S_n] \right| > \epsilon) = 0.
\]

In class we have shown that the Law of Large Numbers holds if \( S_n = X_1 + \cdots + X_n \), where the \( X_i \)'s are i.i.d. random variables. This problem explores if the Law of Large Numbers holds under other circumstances.

Packets are sent from a source to a destination node over the Internet. Each packet is sent on a certain route. Each route has a failure probability of \( p \) and different routes fail independently. If a route fails, all packets sent along that route are lost. You can assume that the routing protocol has no knowledge of which route fails.

For each of the following routing protocols, determine whether the Law of Large Numbers holds when \( S_n \) is defined as the total number of received packets out of \( n \) packets sent. Do a calculation to justify each answer. (Whenever convenient, you can assume below that \( n \) is even.)

(a) **YES** or **NO**: Each packet is sent on a completely different route.

(b) **YES** or **NO**: The packets are split into \( n/2 \) pairs of packets. Each pair is sent together on its own route (i.e., different pairs are sent on different routes).

(c) **YES** or **NO**: The packets are split into 2 groups of \( n/2 \) packets. All the packets in each group are sent on the same route, and the two groups are sent on different routes.

(d) **YES** or **NO**: All the packets are sent on one route.

**Solution:**

(a) Yes. Define a set of indicator functions

\[
Z_i := 1_{\{\text{Packet } i \text{ is received}\}}, \quad 1 \leq i \leq n.
\]

Apparently, \( Z_i \)'s are i.i.d. Bernoulli variables with bounded variance. Since \( S_n = \sum_{i=1}^{n} Z_i \), LLN ensures that

\[
\lim_{n \to \infty} \mathbb{P} \left( \frac{1}{n} |S_n - E[S_n]| > \epsilon \right) = 0.
\]

(b) Yes. Suppose that for Packet \( 2i - 1 \) and \( 2i \) are paired. Define a set of indicator functions

\[
Z_i := 1_{\{\text{Packet } 2i \text{ and } 2i-1 \text{ are received}\}}, \quad 1 \leq i \leq \frac{n}{2}.
\]
Apparently, $Z_i$’s are i.i.d. Bernoulli variables with mean $1 - p$ and bounded variance. Since $S_n = \sum_{i=1}^{n/2} 2Z_i$, LLN ensures that
\[
\lim_{n \to \infty} \mathbb{P} \left( \frac{1}{n} \left| S_n - n\mathbb{E}[Z_i] \right| > \epsilon \right) = 0,
\]
\[
\implies \lim_{n \to \infty} \mathbb{P} \left( \frac{1}{n} \left| S_n - n(1-p) \right| > \epsilon \right) = 0.
\]
Observing that $\mathbb{E}[S_n] = n(1 - p)$ completes the proof.

(c) No. Observe that $S_n$ can only take 3 different values ($S_n = n$, $S_n = n/2$, or $S_n = 0$). For instance, $S_n = n$ when all packets are received, which occurs with probability $(1 - p)^2 > 0$, and $S_n = 0$ when none of the packets is received, which occurs with probability $p^2 > 0$. Consequently, $S_n$ cannot converge to a constant value in probability.

(d) No. $S_n$ can only take 2 different values: i) $S_n = n$ when all packets are received, which occurs with probability $p > 0$, and ii) $S_n = 0$ when none of the packets is received, which occurs with probability $p > 0$. Consequently, $S_n$ cannot converge to a constant value in probability.

5. Download the data package at [http://web.stanford.edu/class/ee278/homeworks/hw2-data.zip](http://web.stanford.edu/class/ee278/homeworks/hw2-data.zip), where you’ll find the MNIST handwritten digit dataset. The data package includes two image subsets (concerning digit 0 and digit 2) each containing 4999 instances of $28 \times 28$ pixels.

(a) Each image can be vectorized as a 784-dimensional vector. Suppose that each image in subset $l$ are independently drawn from some distribution $\mathbb{P}_l$ ($l = 1, 2$). Compute an estimate of the covariance matrix for each subset of images (i.e. the set of digit-0 images and the set of digit-2 images).

(b) For each subset of images, compute and plot the first few eigenvectors (as images) of the estimated covariance matrix. These are the principal components of the data.

(c) Compute an estimate of the covariance matrix for the entire set of images (including all digit-0 and digit-2 images). Plot its first few eigenvectors as images. Compare and contrast these plots with the ones you obtain in Part (b).

(d) Plot the eigenvalues for the covariance matrices computed in Part (a) and Part (c). Explain your findings in the plots. For instance, from the eigenvalues, what can you say about the “effective” dimensionality of the data? What does your observation suggest for feature selection for say recognizing digits?

Please include your code. You can use any programming language (e.g. Matlab, Python) for this problem, but you are not allowed to use off-the-shelf functions like “PCA” or “cov”. For Matlab, you might find the functions `imshow` and `imread` useful in plotting and reading image data.

**Solution:**

(a) One estimation of the covariance matrix is the following:
\[
\frac{1}{n} \sum_{i=1}^{n} \left( X_i - \frac{1}{n} \sum_{j=1}^{n} X_j \right) \left( X_i - \frac{1}{n} \sum_{j=1}^{n} X_j \right)^\top.
\]
Do remember to remove the mean component from each instance $X_i$. 

---

*Homework #1 Solutions*
(b) Perform an eigen-decomposition on the estimated covariance matrices to get the principal components. Below we attach plots of the top-5 principal components.

(c) One can use the same procedure as in Part (b) to obtain the principal components for the entire set of all digit-0 and digit-2 images. One can see that the eigen-images (i.e. the images representing these principal components) look like certain mixture / superposition of the eigen-images obtained in Part (b).

The Matlab code is attached as follows.

```matlab
numImage = 4999;
N = 28;
X = zeros(N*N, numImage);

for i = 1: numImage
    filename = strcat( num2str(i,'%05d'), '.pgm' );
    X(:,i) = reshape( imread( filename ), N^2, 1);
end

% remove the mean component
meanX = mean(X')';
centeredX = X - kron(meanX, ones(1,numImage));
CovX = 1/numImage * centeredX * centeredX';

% eigen-decomposition / PCA
[U,S] = eigs(CovX, 100, 'lm');
[S,I] = sort(diag(S),'descend');
U = U(:,I);

for i = 1: 5
    figure
    imshow( reshape( U(:,i), 28,28) / max( abs(U(:,i)) ) );
    title( strcat( num2str(i), 'th component') )
end

figure
plot(S)
ylabel('eigenvalues of the covariance matrix')
```

(d) The eigenvalue distributions are plotted (we only plot partial eigenvalues for better illustration). One can see from the plot that the magnitude of the eigenvalues decays very fast. For instance, for the covariance matrix for the entire set of all digit-0 and digit-2 images, the first 35 principal components (out of a total number of 784 eigenvalues) already account for 80% of the total energy. That said, the “effective” dimensionality of the data is significantly lower than the ambient dimension of the data.

We can also see that the magnitudes of the last several hundred eigenvalues are negligible compared to the total energy of the data. In some sense, most information is contained within the first several principal components since they explain most of the variability of the data. This suggests that for
practical applications, we might want to project the data onto the first several principal components as the most important features of the data.