1. Proving equivalent forms of total variation distance $d(P_0, P_1)$ for discrete distributions $P_0$ and $P_1$ where

$$d(P_0, P_1) \overset{\text{def}}{=} \sup_{A \subset \mathcal{X}} P_0(A) - P_1(A)$$

(i) 

$$d(P_0, P_1) = \sup_{A \subset \mathcal{X}} P_0(A) - P_1(A)$$

$$= \sup_{A \subset \mathcal{X}} [1 - P_0(A^c)] - [1 - P_1(A^c)]$$

$$= \sup_{A \subset \mathcal{X}} P_1(A^c) - P_0(A^c)$$

$$= \sup_{A^c \subset \mathcal{X}} P_1(A^c) - P_0(A^c)$$

$$= \sup_{A \subset \mathcal{X}} P_1(A) - P_0(A).$$

(ii) Follows from the above two definitions.

(iii) Consider the following subset of $\mathcal{X}$:

$$A^* = \{x \in \mathcal{X} : P_0(x) \geq P_1(x)\}.$$ 

Claim: $\sup_{A \subset \mathcal{X}} P_0(A) - P_1(A) = P_0(A^*) - P_1(A^*)$.

The proof is by contradiction. Assume there exists $B \subset \mathcal{X}$, $B \neq A^*$, that attains the maximum. Let $B \subset A^*$. This gives

$$P_0(A^*) - P_1(A^*) = P_0(B) - P_1(B) + \sum_{x \in A^* \setminus B} P_0(x) - P_1(x)$$

$$\geq P_0(B) - P_1(B),$$

contradicting the assumption that $B$ is the maximizer. Likewise, if $B \supset A^*$, then

$$P_0(B) - P_1(B) = P_0(A^*) - P_1(A^*) + \sum_{x \in B \setminus A^*} P_0(x) - P_1(x)$$

$$\leq P_0(A^*) - P_1(A^*),$$

again leading to a contradiction. Therefore,

$$d(P_0, P_1) = \sup_{A \subset \mathcal{X}} P_0(A) - P_1(A) = P_0(A^*) - P_1(A^*),$$

i.e.,

$$d(P_0, P_1) = \sum_{x \in \mathcal{X} : P_0(x) \geq P_1(x)} P_0(x) - P_1(x).$$
(iv) Similar arguments as in 1(iii).

(v) From the equivalent forms in 1(iii) and 1(iv), we have

\[
2 \cdot d(P_0, P_1) = \sum_{x \in \mathcal{X}: P_1(x) \geq P_0(x)} |P_0(x) - P_1(x)| + \sum_{x \in \mathcal{X}: P_0(x) \geq P_1(x)} |P_0(x) - P_1(x)|
\]

\[
= \sum_{x \in \mathcal{X}} |P_0(x) - P_1(x)|.
\]

2. Properties of total variation distance.

(i) From part 1(iii), it follows that \(d(P_0, P_1)\) is a sum of non negative quantities and hence \(d(P_0, P_1) \geq 0\). Also, from 1(v), it follows that

\[
d(P_0, P_1) = \frac{1}{2} \sum_{x \in \mathcal{X}} |P_0(x) - P_1(x)|
\]

\[
\leq \frac{1}{2} \left[ \sum_{x \in \mathcal{X}} |P_0(x)| + \sum_{x \in \mathcal{X}} |P_1(x)| \right] \quad \text{(by triangle inequality)}
\]

\[
\leq \frac{1}{2} [1 + 1]
\]

Thus, \(0 \leq d(P_0, P_1) \leq 1\).

(ii) From 1(v), it is clear that if \(P_0 = P_1\), then \(d(P_0, P_1) = 0\). Conversely, if \(d(P_0, P_1) = 0\), the summands \(|P_0(x) - P_1(x)|\) have to be all zero, implying that \(P_0(x) = P_1(x)\) for all \(x \in \mathcal{X}\).

(iii) When \(P_0\) and \(P_1\) have disjoint supports,

\[
P_0(\{x \in \mathcal{X} : P_0(x) \geq P_1(x)\}) = 1 \quad \text{and,}
\]

\[
P_1(\{x \in \mathcal{X} : P_0(x) \geq P_1(x)\}) = 0.
\]

Then, using the characterization in 1(iii), \(d(P_0, P_1) = 1\).

To show the converse, we appeal to the characterization in 1(i). Let \(d(P_0, P_1) = 1\). For any subset \(A \subset \mathcal{X}\), \(P_1(A) - P_0(A) \leq 1\) and strict inequality holds if \(P_0(A) > 0\). Therefore, to attain the supremum, the set \(A\) must be such that \(P_1(A) = 1\) and \(P_0(A) = 0\). That is to say that \(P_1\) is supported on \(A\) and \(P_0\) on \(A^c\) (disjoint supports).

3. Hypothesis Testing

(i) \(p = 0.5, q = 0.51; P \equiv \text{Ber}(p), Q \equiv \text{Ber}(q)\)

\[
D(P||Q) = \frac{1}{2} \log \frac{0.5}{0.51} + \frac{1}{2} \log \frac{0.5}{0.49}
\]

\[
= 2.886 \times 10^{-4}
\]

\[
\text{Var}_P \left( \log \frac{P(X)}{Q(X)} \right) = \mathbb{E}_P \left[ \left( \log \frac{P(X)}{Q(X)} \right)^2 \right] - D(P||Q)^2
\]

\[
= \frac{1}{2} \left( \log \frac{0.5}{0.51} \right)^2 + \frac{1}{2} \left( \log \frac{0.5}{0.49} \right)^2 - (2.886 \times 10^{-4})^2
\]

\[
= 8.3277 \times 10^{-4}
\]
From Chebyshev’s inequality we have,

\[ P \left( \log \frac{P^n(X^n)}{Q^n(X^n)} \geq n D(P||Q) - \frac{n \text{Var}_P \left( \frac{\log P(X)}{Q(X)} \right)}{\epsilon} \right) \geq 1 - \epsilon. \]

Thus, from Little-Big lemmas we have

\[ \lambda = n D(P||Q) - \sqrt{\frac{n \text{Var}_P \left( \frac{\log P(X)}{Q(X)} \right)}{\epsilon}}. \]

For \( n = 10,000, \epsilon = 0.5 \) and substituting for divergence and variance from above

\[ \lambda = -25.978. \]

Similarly, from the other side of Chebyshev’s inequality we have,

\[ P \left( \log \frac{P^n(X^n)}{Q^n(X^n)} \leq n D(P||Q) + \frac{n \text{Var}_P \left( \frac{\log P(X)}{Q(X)} \right)}{\delta} \right) \geq 1 - \delta, \]

that is,

\[ \lambda' = n D(P||Q) + \sqrt{\frac{n \text{Var}_P \left( \frac{\log P(X)}{Q(X)} \right)}{\delta}}. \]

Choosing \( \delta = \epsilon = 0.01 \), and using values for the other parameters as above, we get \( \lambda' = 31.744 \). Using these values in the Little-Big lemmas, we get

\[ 2^{-31.744} \leq \beta_{0.01}(P^n, Q^n) \leq 2^{25.978}, \]

that is,

\[ 2^{-31.744} \leq \beta_{0.01}(P^n, Q^n) \leq 1. \]

We see that the bounds derived using Chebyshev’s inequality are weak and gives trivial bound on one side.

(ii) From Stein’s lemma, for large values of \( n \), \( \beta_{0.01}(P^n, Q^n) \approx 2^{-n D(P||Q)} \). For \( n = 10000, \ 2^{-n D(P||Q)} = 2^{-2.880} \). This suggests that the bounds derived in 4.(i) are very weak and don’t compare well with the asymptotic behaviour of \( \beta_{0.01}(P, Q) \).

(iii) For this part we have \( p = 1 \) and \( q = 0.01 \).
Repeating the calculations as in 4.(i) and 4.(ii), we have,

\[ D(P||Q) = \log \frac{1}{0.01} = 6.6439 \]

\[ \text{Var}_P \left( \log \frac{P(X)}{Q(X)} \right) = 0 \]

\[ \lambda = nD(P||Q) - \sqrt{\frac{n\text{Var}_P \left( \log \frac{P(X)}{Q(X)} \right)}{\epsilon}} \]
\[ = 6.6439 \times 10^4 \]

\[ \lambda' = nD(P||Q) + \sqrt{\frac{n\text{Var}_P \left( \log \frac{P(X)}{Q(X)} \right)}{\delta}} \]
\[ = 6.6439 \times 10^4 \]

\[ \beta_{0.01}(P^n, Q^n) = 2^{-6.6439 \times 10^4}, \]
which is the same as what is given by Stein’s lemma.

Alternatively, observe that the acceptance region contains only one element which is the all-ones sequence (all the other sequences have zero probability under the null hypotheses). Also, since the probability of the rejection region under null hypotheses is zero, false alarm probability is zero for all \( n \). Thus the missed detection probability \( \beta_\epsilon(P, Q) \), is the probability of getting an all-ones sequence under \( Q \), which is equal to \( q^n = 2^{\log q^n} = 2^{-n \log \frac{1}{q}} = 2^{-nD(P||Q)} \), and is independent of \( \epsilon \). This is exactly what we obtained above by Stein’s lemma and also from the bounds.

4. Calculating the entropy.

(i) \( X_i \overset{iid}{\sim} \text{Ber}(p), \ i \in [n] \).

Therefore, \( H(X) = nH(X_1) = n[-p \log p - (1 - p) \log(1 - p)] \).

(ii) \( X \sim \text{Geo}(p) \).

We have \( P(X = k) = (1 - p)^{k-1}p \) and \( \mathbb{E}[X] = \frac{1}{p} \). Now,

\[ H(X) = \mathbb{E}[-\log P(X)] \]
\[ = \mathbb{E}[-\log \{(1 - p)^{X-1}\}] \]
\[ = \mathbb{E}[(1 - X) \log(1 - p) - \log p] \]
\[ = (1 - \frac{1}{p}) \log(1 - p) - \log p \]
\[ = \frac{-(1 - p) \log(1 - p) - p \log p}{p} \]
(iii) \( X \sim \text{Exp}(\lambda) \). For every \( x \), we have that \( Y \sim \text{Poi}(x) \). Then,

\[
P_Y(y) = \int_0^\infty P_X(x)P_{Y|X}(y|x)\,dx
\]

\[
= \int_0^\infty \lambda e^{-\lambda x} x^y e^{-x} \frac{y^y}{y!} \, dx
\]

\[
= \frac{\lambda}{y!} \int_0^\infty x^y e^{-(\lambda+1)x} \, dx
\]

\[
= \frac{\lambda \Gamma(y+1)}{y!(\lambda+1)^{y+1}}
\]

\[
= \frac{\lambda y!}{(\lambda+1)^{y+1}}
\]

\[
= \left( 1 - \frac{\lambda}{\lambda+1} \right)^y \left( \frac{\lambda}{\lambda+1} \right).
\]

Thus the random variable \( Y \) is a translated Geo\( \left( \frac{\lambda}{\lambda+1} \right) \) random variable translated by 1. Since entropy is translation invariant, from 4.(ii),

\[
H(Y) = \frac{(\lambda+1)h(\frac{\lambda}{\lambda+1})}{\lambda},
\]

where \( h(p) = -(1-p) \log(1-p) - p \log p \)

5. Calculating the KL divergence.

(i) \( P \equiv \text{N}(\mu_1, \sigma^2), Q \equiv \text{N}(\mu_2, \sigma^2) \).

\[
D(P||Q) = \int_\mathbb{R} \frac{e^{-(x-\mu_1)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} \left[ \frac{(x-\mu_2)^2 - (x-\mu_1)^2}{2\sigma^2} \right] dx
\]

\[
= \int_\mathbb{R} \frac{e^{-(x-\mu_1)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} \left[ 2x(\mu_1 - \mu_2) + \mu_2^2 - \mu_1^2 \right] dx
\]

\[
= \frac{2(\mu_1 - \mu_2)}{2\sigma^2} \int_\mathbb{R} \frac{e^{-(x-\mu_1)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} \cdot x \cdot dx + \int_\mathbb{R} \frac{e^{-(x-\mu_1)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} \cdot (\frac{\mu_2^2 - \mu_1^2}{2\sigma^2}) dx
\]

\[
= \frac{(\mu_1 - \mu_2)^2}{2\sigma^2}
\]

(ii) \( P \equiv \text{Poi}(\lambda_1), Q \equiv \text{Poi}(\lambda_2) \).

\[
D(P||Q) = \sum_{k=0}^{\infty} \frac{\lambda_1^k e^{-\lambda_1}}{k!} \log \frac{\lambda_1^k e^{-\lambda_1}}{\lambda_2^k e^{-\lambda_2}}
\]

\[
= \sum_{k=0}^{\infty} \frac{\lambda_1^k e^{-\lambda_1}}{k!} \log \left( \frac{\lambda_1}{\lambda_2} \right)^k + \sum_{k=0}^{\infty} \frac{\lambda_1^k e^{-\lambda_1}}{k!} (\lambda_2 - \lambda_1)
\]

\[
= \log \left( \frac{\lambda_1}{\lambda_2} \right) \left[ \sum_{k=0}^{\infty} \frac{\lambda_1^k e^{-\lambda_1}}{k!} \cdot k \right] + (\lambda_2 - \lambda_1) \left[ \sum_{k=0}^{\infty} \frac{\lambda_1^k e^{-\lambda_1}}{k!} \right]
\]

\[
= \lambda_1 \log \left( \frac{\lambda_1}{\lambda_2} \right) + (\lambda_2 - \lambda_1).
\]
(iii) \( P \equiv \text{Geo}(p), Q \equiv \text{Geo}(q) \).

\[
D(P||Q) = \mathbb{E}_P \left[ \log \frac{(1-p)^{X-1}p}{(1-q)^{X-1}q} \right] \\
= \mathbb{E}_P \left[ (X-1) \log \left( \frac{1-p}{1-q} \right) + \log \left( \frac{p}{q} \right) \right] \\
= \left( \frac{1}{p} - 1 \right) \log \left( \frac{1-p}{1-q} \right) + \log \left( \frac{p}{q} \right)
\]

6. Total variation distance and KL divergence between two Bernoulli distributions.

(i) Using the characterization of total variation distance in 1(v), for \( P \equiv \text{Ber}(p), Q \equiv \text{Ber}(q) \),

\[
d(P, Q) = \frac{1}{2} \left( |p - q| + |(1-p) - (1-q)| \right) \\
= |p - q|.
\]

(ii) * (Pinsker’s inequality for Bernoulli) Consider \( D(P||Q) \) in nats\(^1\), i.e.,

\[
D(P||Q) = p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q}.
\]

Fix a \( q \). Now, let

\[
g(p) = D(P||Q) - 2(p-q)^2 \\
= p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q} - 2p^2 + 4pq - 2q^2 \\
\frac{dg(p)}{dp} = \ln \frac{p}{q} - \ln \frac{1-p}{1-q} - 4(p-q).
\]

Setting \( \frac{dg(p)}{dp} \) equal to 0, we get \( p^* = q \). Observe that \( \frac{d^2g(p)}{dp^2} = \frac{1}{p} + \frac{1}{1-p} - 4 \geq 0 \). Therefore, \( p^* \) is indeed the minimizer for \( g(p) \) and \( g(p^*) = 0 \). Thus, it follows that,

\[
D(P||Q) \geq 2(p-q)^2
\]

Equivalently, in bits,

\[
D(P||Q) \geq \frac{2}{\ln 2} (p-q)^2.
\]

\(^1\)expressed in base \( e \); \( \ln \) stands for natural logarithm, i.e., \( \log_e(\cdot) \)