1. To prove the equivalent forms of the total variation distance for discrete distributions $P_0$ and $P_1$ where total variation distance is defined as

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

(a)

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

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

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

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

(b) Follows from the above two definitions.

(c) Consider the following subset of $\mathcal{X}$.

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

For any subset $B \subset \mathcal{X}$, $P_1(B) - P_0(B) \leq P_1(A^*) - P_0(A^*)$. Therefore,

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

i.e.,

$$d(P_0, P_1) = \sum_{x \in \mathcal{X} : P_1(x) \geq P_0(x)} P_1(x) - P_0(x).$$

(d) Similar arguments as the previous one.

(e) From the equivalent forms in (1c) and (1d), 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 the total variation distance.

(i) From (1c), it follows that \( d(P_0, P_1) \) is a sum of non negative quantities and hence \( d(P_0, P_1) \geq 0 \). From the same expression, it follows that

\[
\begin{align*}
    d(P_0, P_1) &= \frac{1}{2} \sum_{x \in X'} |P_0(x) - P_1(x)| \\
    &\leq \frac{1}{2} \left[ \sum_{x \in X'} |P_0(x)| + \sum_{x \in X'} |P_1(x)| \right] \\
    &\leq \frac{1}{2} [1 + 1]
\end{align*}
\]

(by triangle inequality)

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

(a) From (1c), 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 sum of non negative quantities \( |P_0(x) - P_1(x)| \) have to be all zero. i.e., \( P_0 = P_1 \).

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

\[
\begin{align*}
P_1(\{x \in X : P_1(x) \geq P_0(x)\}) &= 1 \text{ and,} \\
P_0(\{x \in X : P_1(x) \geq P_0(x)\}) &= 0.
\end{align*}
\]

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

To show the converse, we appeal to the characterization in (1a).

Let \( d(P_0, P_1) = 1 \). For any subset \( A \subset X \), \( P_1(A) - P_0(A) \leq 1 \) 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. Hypotheses Testing

(i) \( p = 0.5, q = 0.51 \)

\[
\begin{align*}
    D(P||Q) &= \frac{1}{2} \log \frac{0.5}{0.51} + \frac{1}{2} \log \frac{0.5}{0.49} \\
    &= 2.886 \times 10^{-4}
\end{align*}
\]

\[
\begin{align*}
    \text{Var} \left( \log \frac{p}{q} \right) &= E \left[ \left( \log \frac{p}{q} \right)^2 \right] - D(P||Q)^2 \\
    &= \frac{1}{2} \left( \log \frac{0.5}{0.51} \right)^2 + \frac{1}{2} \left( \log \frac{0.5}{0.49} \right)^2 - (2.886 \times 10^{-4})^2 \\
    &= 8.3277 \times 10^{-4}
\end{align*}
\]

From Chebyshev’s inequality we have,

\[
\mathbb{P} \left( \log \frac{p(x^n)}{q(x^n)} \geq nD(P||Q) - \sqrt{\frac{n\text{Var} \left( \log \frac{p}{q} \right)}{\epsilon}} \right) \geq 1 - \epsilon.
\]

Thus, from Little-Big lemmas we have

\[
\lambda = nD(P||Q) - \sqrt{\frac{n\text{Var} \left( \log \frac{p}{q} \right)}{\epsilon}}.
\]
For \( n = 10,000, \epsilon = 0.5 \) and substituting for divergence and variance from above

\[
\lambda = -25.978.
\]

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

\[
\mathbb{P}\left( \log \frac{p(x^n)}{q(x^n)} \leq nD(P||Q) + \sqrt{\frac{n\text{Var}\left( \log \frac{p}{q} \right)}{\delta}} \right) \geq 1 - \delta.
\]

Thus,

\[
\Rightarrow \lambda' = nD(P||Q) + \sqrt{\frac{n\text{Var}\left( \log \frac{p}{q} \right)}{\delta}}.
\]

Choosing \( \delta = \epsilon = 0.01 \), and using values for the other parameters as above, we get \( \lambda' = 31.744 \). From the two lemmas, \( \lambda \) and \( \lambda' \),

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

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

As we see, the bounds derived using Chebyshev’s inequality is bad and gives trivial bound on one side.

(ii) From Stein’s lemma, for large values of \( n \), \( \beta_{0.01}(P,Q) \approx 2^{-nD(P||Q)} \). For \( n = 10000, \ 2^{-nD(P||Q)} = 2^{-2.886} \). 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) Now, for the new pair of coins we have \( p = 1 \) and \( q = 0.01 \). Repeating the calculations as in 4.(i) and 4.(ii), we have,

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

\[
\text{Var}\left( \log \frac{p}{q} \right) = 0
\]

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

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

\[
\beta_{0.01}(P,Q) = 2^{-6.6439 \times 10^4},
\]

which is same as what is given by Stein’s lemma.

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

4. Calculating the entropy.

(i) $H(X) = H(X_1, X_2, ..., X_n)$ where $X_i \sim$ i.i.d. $Ber(p)$. Therefore, $H(X) = nH(X_1) = n[-p \log p - (1-p) \log(1-p)]$.

(ii) $X \sim Geo(p)$. Consider $h(X) = -\log P(X)$ where $P(X = k) = (1-p)^{k-1}p$. We also know that for $X \sim Geo(p)$, $E[X] = \frac{1}{p}$. Now,

$$H(X) = \mathbb{E}[h(X)] = \mathbb{E}[\log \{ (1-p)^{X-1}p \}] = \mathbb{E}[(1-X) \log(1-p) - \log p] = (1 - \frac{1}{p}) \log(1-p) - \log p = -(1-p) \log(1-p) - \frac{p \log p}{p}$$

(iii) $X \sim Exp(\lambda)$. For every $x$, $Y \sim Poi(x)$.

$$P_Y(y) = \int_0^\infty P_X(x)P_{Y|x}(y|x)$$

$$= \int_0^\infty \lambda e^{-\lambda x} x^y e^{-x} \frac{x^y e^{-x}}{y!} \cdot dx$$

$$= \frac{\lambda}{y!} \int_0^\infty x^y e^{-(\lambda+1)x} \cdot dx$$

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

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

$$= \frac{\lambda}{(\lambda + 1)^{y+1}}$$

$$= \left( 1 - \frac{\lambda}{\lambda + 1} \right)^y \left( \frac{\lambda}{\lambda + 1} \right)$$

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. KL divergence, $D(P||Q)$ where
(i) $P \equiv N(\mu_1, \sigma^2)$, $Q \equiv N(\mu_2, \sigma^2)$.

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

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

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

$$= \frac{(\mu_1 - \mu_2)^2}{2\sigma^2}.$$  

(ii) $P \equiv Pois(\lambda_1)$, $Q \equiv Pois(\lambda_2)$.

$$D(P||Q) = \sum_{k=0}^{\infty} \frac{\lambda_2^k e^{-\lambda_1}}{k!} \cdot \log \left( \frac{\lambda_1^k e^{-\lambda_1}}{\lambda_2^k e^{-\lambda_2}} \right)$$

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

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

$$= \lambda_1 \log \left( \frac{\lambda_1}{\lambda_2} \right) + (\lambda_2 - \lambda_1).$$

(iii) $P \equiv Geo(p)$, $Q \equiv Geo(q)$.

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

$$= \mathbb{E}_P \left[ (X-1) \log \left( \frac{1-p}{1-q} \right) + \log \left( \frac{p}{q} \right) \right]$$

$$= \left( \frac{1}{p} - 1 \right) \log \left( \frac{1-p}{1-q} \right) + \log \left( \frac{p}{q} \right).$$  

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

(i) Using the characterization of total variation distance in (1e), for $P \equiv Ber(p)$, $Q \equiv Ber(q)$,

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

$$= |p - q|.$$  

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

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

\(^1\ln \text{ stands for natural logarithm, i.e., } \log_e(\cdot)\)
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,
\[
D(P||Q) \geq \frac{2}{\ln 2} (p - q)^2.
\]