September 13 **September 2019**

Tutorial 5: Hypothesis Testing, TV Distance and KL Divergence

Course Instructor: Himanshu Tyagi *Prepared by: Karthik*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Contents

5.1 Preliminaries: Binary Hypothesis Testing

We begin this tutorial with a review of some basic ideas from binary hypothesis testing. In particular, we shall review the frameworks of Bayesian hypothesis testing and Neyman-Pearson hypothesis testing. Let X be a discrete random variable taking values in the discrete set $\mathcal{X}$. Suppose P_0 and P_1 are two distributions on $\mathcal{X}$. It is known that the distribution of X follows one of P_1 and P_2, and the goal is to figure out the distribution of X by observing the value of X. Specifically, it is of interest to resolve the following problem:

$$\begin{aligned} \mathcal{H}_0 : &\quad X \sim P_0 \\ &\quad \text{v.s.} \\ \mathcal{H}_1 : &\quad X \sim P_1. \end{aligned}$$

The above problem is known as a binary hypothesis testing problem. Hypothesis $\mathcal{H}_0$ is known as the *null* hypothesis, and $\mathcal{H}_1$ is known as the *alternative* hypothesis.

A *test* for the above binary hypothesis testing problem is a function $g : \mathcal{X} \to \{0, 1\}$ that upon observing the value of the random variable X declares whether hypothesis $\mathcal{H}_1$ is true or hypothesis $\mathcal{H}_0$ is true. Thus, there may be some values of X for which a test g may declare $\mathcal{H}_0$ to be true, while there may be some other values of X for which g may declare $\mathcal{H}_1$ to be true. Let

$$A \coloneqq \{x \in \mathcal{X} : g(x) = 0\}$$

denote the set of values for which the test g declares $\mathcal{H}_0$ to be true. The set A is commonly referred to as the *rejection region* (region where $\mathcal{H}_1$ is rejected) of the test g. Its complementary set A^c is referred to as the *acceptance* region (region where $\mathcal{H}_1$ is accepted) of the test g. Specifying a test g is equivalent to specifying its rejection region A.

Any test g may make errors in declaring $\mathcal{H}_0$ to be true when actually $\mathcal{H}_1$ is true, or vice-versa. Thus, given any test g, there is a possibility for the occurrence of the following types of errors:

1. Type-I error (false alarm): This is the error that occurs when $\mathcal{H}_1$ is declared to be the true hypothesis when actually $\mathcal{H}_0$ is the true hypothesis. Given any test g with rejection region A, its type-I error probability is given by $P_1(A)$.

2. Type-II error (missed detection): This is the error that occurs when $\mathcal{H}_0$ is declared to be the true hypothesis when actually $\mathcal{H}_1$ is the true hypothesis. Given any test g with rejection region A, its type-II error probability is given by $P_0(A^c)$.

5.2 Bayes Hypothesis Testing Framework

One of the most commonly considered frameworks in binary hypothesis testing is that of Bayes hypothesis testing. In this framework, each of the two hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ is known to be true with prior probabilities p and $1 - p$ respectively, where $p \in [0, 1]$ is a known constant. Given these prior probabilities, the average probability of error of a test g whose rejection region is A, denoted by $P_e(A)$, under these prior probabilities is given by

$$P_e(A) = p \cdot P_0(A^c) + (1 - p) \cdot P_1(A).$$

In the framework of Bayes hypothesis testing, the goal is to arrive at tests g whose average probability of error (for fixed prior probabilities p and $(1 - p)$) is the least possible value. Fixing the prior probabilities of the two hypotheses to be $p = 0.5 = 1-p$, let P_e^{unif} denote the minimum value of average probability of error among all tests. We note that when $p = 0.5 = 1-p$, for any test g with rejection region A, we have

$$\begin{aligned} P_e(A) &= \frac{1}{2}(1 - (P_0(A) - P_1(A))) \\ &= \frac{1}{2}\left(1 - \sum_{x \in A : P_0(x) \geq P_1(x)} (P_0(x) - P_1(x)) - \sum_{x \in A : P_0(x) < P_1(x)} (P_0(x) - P_1(x))\right) \\ &\geq \frac{1}{2}\left(1 - \sum_{x \in A : P_0(x) \geq P_1(x)} (P_0(x) - P_1(x))\right), \end{aligned}$$

and therefore we see that equality above is attained for the set

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

Thus, we get

$$\begin{aligned} P_e^{\text{unif}} &= \frac{1}{2}(1 - (P_0(A^*) - P_1(A^*))) \\ &= \frac{1}{2}(1 - d_{TV}(P_0, P_1)), \end{aligned}$$

where $d_{TV}(P_0, P_1) = P_0(A^*) - P_1(A^*)$ is known as the total variation distance between the probability distributions P_0 and P_1. Therefore, in summary, we note the following:

When the two hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ of a binary hypothesis testing problem are known to be true with equal prior probabilities, the minimum average probability of error is achieved by the test g^* whose rejection region A^* is given by $A^* = \{x \in \mathcal{X} : P_0(x) \geq P_1(x)\}$. The test g^* is known as "Bayes optimal test", and its minimum average probability of error is given by $P_e^{\text{unif}} = 0.5(1 - d_{TV}(P_0, P_1))$. Thus, the total variation distance between two probability distributions arises as a fundamental quantity in Bayes binary hypothesis testing with uniform priors.

Example 5.2.1. Show that the total variation distance $d_{TV}(P_0, P_1)$ is given by the formula

$$d_{TV}(P_0, P_1) = \frac{1}{2} \max_{\text{partitions } A_1, \ldots, A_k,\ k \geq 2} \sum_{i=1}^{k} |P_0(A_k) - P_1(A_k)|,$$

where the maximisation above is over all partitions $A_1, \ldots, A_k$ of the set $\mathcal{X}$ for various values of $k \geq 2$.

To show the above result, we make use of the relation

$$d_{TV}(P_0, P_1) = \frac{1}{2} \sum_{x \in \mathcal{X}} |P_0(x) - P_1(x)|,$$

a fact that can easily be checked and is left as an exercise.

Let $A_1, \ldots, A_k$, $k \geq 2$, be any partition of $\mathcal{X}$. Then, we have

$$\begin{aligned} \frac{1}{2} \sum_{i=1}^{k} |P_0(A_k) - P_1(A_k)| &= \frac{1}{2} \sum_{i=1}^{k} \left| \sum_{x \in A_k} (P_0(x) - P_1(x)) \right| \\ &\overset{(a)}{\leq} \frac{1}{2} \sum_{i=1}^{k} \sum_{x \in A_k} |P_0(x) - P_1(x)| \\ &= \frac{1}{2} \sum_{x \in \mathcal{X}} |P_0(x) - P_1(x)| \\ &= d_{TV}(P_0, P_1), \end{aligned}$$

where (a) above follows from triangle inequality. Since the above series of inequalities is true for any partition of $\mathcal{X}$, we have

$$\frac{1}{2} \max_{\text{partitions } A_1,\ldots,A_k,\ k\geq 2} \sum_{i=1}^{k} |P_0(A_k) - P_1(A_k)| \leq d_{TV}(P_0, P_1). \tag{5.1}$$

Also, we note that

$$\begin{aligned} d_{TV}(P_0, P_1) &= P_0(A^*) - P_1(A^*) \\ &= \frac{1}{2}\Big(P_1((A^*)^c) - P_0((A^*)^c) + P_0(A^*) - P_1(A^*)\Big) \\ &\overset{(a)}{=} \frac{1}{2}\sum_{i=1}^{2} |P_0(A_i) - P_1(A_i)| \\ &\leq \frac{1}{2} \max_{\text{partitions } A_1,\ldots,A_k,\ k\geq 2} \sum_{i=1}^{k} |P_0(A_k) - P_1(A_k)|, \end{aligned} \tag{5.2}$$

where in (a) above, we assign $A_1 = A^*$ and $A_2 = (A^*)^c$, and the last line above follows by taking maximisation over all partitions of $\mathcal{X}$. Combining (5.1) and (5.2), we get the desired result.

5.3 The Neyman-Pearson Framework

Often, the knowledge of the prior probabilities of the two hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ of a binary hypothesis testing problem is not available. In such situations, one often considers the problem of minimising the type-II error, subject to the requirement that the type-I error is below a specified tolerance level. This is the setting of the Neyman-Pearson hypothesis testing framework. Formally, given a binary hypothesis testing problem with $\mathcal{H}_0 : X \sim P_0$ and $\mathcal{H}_1 : X \sim P_1$, and an error tolerance parameter $\epsilon > 0$, we would like to solve the following optimisation problem:

$$\begin{gathered} \min_{A\subseteq\mathcal{X}} P_1(A) \\ \text{subject to } P_0(A^c) \leq \epsilon. \end{gathered}$$

We shall denote by $\beta_\epsilon(P_0, P_1)$ the minimum value of the above optimisation problem, i.e.,

$$\beta_\epsilon(P_0, P_1) \coloneqq \min_{A\subseteq\mathcal{X}} P_1(A), \qquad \text{where } A \text{ is such that } P_0(A^c) \leq \epsilon.$$

We shall soon see that $\beta_\epsilon(P_0, P_1)$ is a fundamental quantity in information theory, and behaves very similar to the quantity $L_\epsilon(P)$ we saw earlier. It is worthwhile to note that an exact characterisation of $\beta_\epsilon(P_0, P_1)$ may not be feasible practically. Therefore, we seek upper and lower bounds for $\beta_\epsilon(P_0, P_1)$.

Suppose that P_1 is the uniform distribution on $\mathcal{X}$. Then, we have

$$\beta_\epsilon(P_0, P_{\text{unif}}) = \min_{A\subseteq\mathcal{X}} \frac{|A|}{|\mathcal{X}|}, \qquad \text{where } A \text{ satisfies } P_0(A) \geq 1-\epsilon.$$

From an earlier tutorial session, we know that if $X \sim P$, then the minimum average number of bits required to represent the values of X belonging to a set A satisfying $P(A) \geq 1-\epsilon$ is given by $L_\epsilon(P)$. Therefore, we get

$$\beta_\epsilon(P_0, P_{\text{unif}}) = \frac{2^{L_\epsilon(P_0)}}{|\mathcal{X}|},$$

or equivalently, we have

$$-\log \beta_\epsilon(P_0, P_{\text{unif}}) = \log |\mathcal{X}| - L_\epsilon(P_0).$$

Note that the right hand side of the above equation represents the difference in compressibility of uniform distribution and that of the distribution P_0, a measure of "distance" between the the uniform distribution and P_0. Thus, the quantity $-\log \beta_\epsilon(P_0, P_{\text{unif}})$ represents a sort of "distance" measure between the uniform distribution and the distribution P_0. We shall soon see that this is true for any two distributions P_0 and P_1.

5.3.1 High Probability Lower and Upper Bounds for $\log \frac{P_0(X)}{P_1(X)}$

Suppose $X \sim P$ is a discrete random variable with pmf p. We recall from one of the earlier tutorial sessions that any high probability (probability at least $1-\epsilon$) upper bound for the random variable $Z = -\log p(X)$ is also an upper bound for the quantity $L_\epsilon(P)$. Similarly, any high probability (probability at least $1-\epsilon$) lower bound for the random variable $Z = -\log p(X)$ is also a lower bound for the quantity $L_\epsilon(P)$ (up to additive factors which may be neglected).

In this subsection, we show that for a binary hypothesis testing problem with $\mathcal{H}_0 : X \sim P_0$ and $\mathcal{H}_1 : X \sim P_1$:

- Any high probability (probability at least $1-\epsilon$) lower bound **under the distribution** P_0 for the random variable $X = \log \frac{P_0(X)}{P_1(X)}$ serves as a lower bound for the quantity $-\log \beta_\epsilon(P_0, P_1)$.
- Any high probability (probability at least $1-\epsilon$) upper bound **under the distribution** P_0 for the random variable $X = \log \frac{P_0(X)}{P_1(X)}$ serves as an upper bound for the quantity $-\log \beta_\epsilon(P_0, P_1)$ (up to additive factors which may be neglected).

We now have the following Lemmas in order.

Lemma 5.3.1. *Consider the binary hypothesis testing problem $\mathcal{H}_0 : X \sim P_0$ v.s. $\mathcal{H}_1 : X \sim P_1$. Fix $\epsilon > 0$. Suppose that there exists a constant $\lambda > 0$ such that the set*

$$A_\lambda := \left\{ x \in \mathcal{X} : \log \frac{P_0(x)}{P_1(x)} \geq \lambda \right\}$$

satisfies $P_0(A_\lambda) \geq 1 - \epsilon$. Then,

$$-\log \beta_\epsilon(P_0, P_1) \geq \lambda.$$

Proof. We note that

$$\begin{aligned} 1 &\geq P_0(A_\lambda) \\ &= \sum_{x \in A_\lambda} P_0(x) \\ &= \sum_{x \in A_\lambda} \frac{P_0(x)}{P_1(x)} P_1(x) \\ &\geq \sum_{x \in A_\lambda} 2^\lambda P_1(x) \\ &= 2^\lambda \cdot P_1(A_\lambda), \end{aligned}$$

from which it follows that $P_1(A_\lambda) \leq 2^{-\lambda}$. We then have the following set of inequalities:

$$\begin{aligned}\beta_\epsilon(P_0, P_1) &= \min_{A \subseteq X : P_0(A) \geq 1-\epsilon} P_1(A) \\ &\leq P_1(A_\lambda) \\ &\leq 2^{-\lambda}.\end{aligned}$$

Taking $-\log$ on both sides of the above inequality, we get the desired result. □

Lemma 5.3.1 says that if there is a constant $\lambda > 0$ that serves as a high probability (probability at least $1-\epsilon$) lower bound on the random variable $Z = \log \frac{P_0(X)}{P_1(X)}$, then λ also serves as a lower bound on the quantity $-\log \beta_\epsilon(P_0, P_1)$. We know from Chebyshev's inequality that for any $\epsilon > 0$,

$$P_0\left(\left|\log \frac{P_0(X)}{P_1(X)} - E_{P_0}\left[\log \frac{P_0(X)}{P_1(X)}\right]\right| > \sqrt{\frac{\mathrm{Var}\left(\log \frac{P_0(X)}{P_1(X)}\right)}{\epsilon}}\right) \leq \epsilon.$$

In particular, we have that for any $\epsilon > 0$,

$$P_0\left(\log \frac{P_0(X)}{P_1(X)} \geq E_{P_0}\left[\log \frac{P_0(X)}{P_1(X)}\right] - \sqrt{\frac{\mathrm{Var}\left(\log \frac{P_0(X)}{P_1(X)}\right)}{\epsilon}}\right) \geq 1 - \epsilon.$$

In conjunction with Lemma 5.3.1, we get

$$-\log \beta_\epsilon(P_0, P_1) \geq E_{P_0}\left[\log \frac{P_0(X)}{P_1(X)}\right] - \sqrt{\frac{\mathrm{Var}\left(\log \frac{P_0(X)}{P_1(X)}\right)}{\epsilon}}. \tag{5.3}$$

Lemma 5.3.2. *Consider the binary hypothesis testing problem $\mathcal{H}_0 : X \sim P_0$ v.s. $\mathcal{H}_1 : X \sim P_1$. Fix $\epsilon > 0$. Suppose that there exists a constant $\lambda > 0$ such that the set*

$$B_\lambda \coloneqq \left\{x \in \mathcal{X} : \log \frac{P_0(x)}{P_1(x)} \leq \lambda\right\}$$

satisfies $P_0(B_\lambda) \geq 1 - \frac{\epsilon}{2}$. Then,

$$-\log \beta_\epsilon(P_0, P_1) \leq \lambda + \log \frac{1}{1-\epsilon}.$$

Proof. Consider any set $A \subseteq \mathcal{X}$ such that $P_0(A) \geq 1 - \frac{\epsilon}{2}$. Then, it follows that $A \cap B_\lambda$ has high probability. In particular,

$$\begin{aligned}P_0(A \cap B_\lambda) &= P_0(A) + P_0(B_\lambda) - P_0(A \cup B_\lambda) \\ &\geq P_0(A) + P_0(B_\lambda) - 1 \\ &= 1 - \epsilon.\end{aligned}$$

Furthermore, we have

$$\begin{aligned}1-\epsilon &\leq P_0(A\cap B_\lambda)\\ &= \sum_{x\in A\cap B_\lambda} P_0(x)\\ &= \sum_{x\in A\cap B_\lambda} \frac{P_0(x)}{P_1(x)}P_1(x)\\ &\leq \sum_{x\in A\cap B_\lambda} 2^\lambda P_1(x)\\ &= 2^\lambda\cdot P_1(A\cap B_\lambda),\end{aligned}$$

from which it follows that $P_1(A\cap B_\lambda)\geq(1-\epsilon)2^{-\lambda}$. Finally, we have

$$P_1(A)\geq P_1(A\cap B_\lambda)\geq(1-\epsilon)2^{-\lambda}.$$

Since the above equation is true for any set A for which $P_0(A)\geq 1-\frac{\epsilon}{2}\geq 1-\epsilon$, it follows that

$$\beta_\epsilon(P_0,P_1)\geq(1-\epsilon)2^{-\lambda}.$$

Taking $-\log$ on both sides of the above equation yields the desired result. □

Lemma 5.3.2 says that if there is a constant $\lambda>0$ that serves as a high probability (probability at least $1-\frac{\epsilon}{2}$) upper bound on the random variable $Z=\log\frac{P_0(X)}{P_1(X)}$, then λ also serves as an upper bound on the quantity $-\log\beta_\epsilon(P_0,P_1)$. We know from Chebyshev's inequality that for any $\epsilon>0$,

$$P_0\left(\left|\log\frac{P_0(X)}{P_1(X)}-E_{P_0}\left[\log\frac{P_0(X)}{P_1(X)}\right]\right|>\sqrt{\frac{\mathrm{Var}\left(\log\frac{P_0(X)}{P_1(X)}\right)}{\epsilon/2}}\right)\leq\frac{\epsilon}{2}.$$

In particular, we have that for any $\epsilon>0$,

$$P_0\left(\log\frac{P_0(X)}{P_1(X)}\leq E_{P_0}\left[\log\frac{P_0(X)}{P_1(X)}\right]+\sqrt{\frac{\mathrm{Var}\left(\log\frac{P_0(X)}{P_1(X)}\right)}{\epsilon/2}}\right)\geq 1-\frac{\epsilon}{2}.$$

In conjunction with Lemma 5.3.2, we get

$$-\log\beta_\epsilon(P_0,P_1)\leq E_{P_0}\left[\log\frac{P_0(X)}{P_1(X)}\right]+\sqrt{\frac{\mathrm{Var}\left(\log\frac{P_0(X)}{P_1(X)}\right)}{\epsilon/2}}+\log\frac{1}{1-\epsilon}. \tag{5.4}$$

From (5.3) and (5.4), we note that when the variance term $\mathrm{Var}\left(\log\frac{P_0(X)}{P_1(X)}\right)$ is negligible, the dominant term in both the upper bound and the lower bound for the quantity $-\log\beta_\epsilon(P_0,P_1)$ is the term

$$D(P_0||P_1):=E_{P_0}\left[\log\frac{P_0(X)}{P_1(X)}\right].$$

This term is known as the Kullback-Leibler divergence (KL divergence) between distributions P_0 and P_1, and represents a measure of "distance" between distributions P_0 and P_1.

5.3.2 From a Single Random Variable to $n > 1$ IID Copies: Stein's Lemma

Consider the binary hypothesis testing problem

$$\begin{aligned} \mathcal{H}_0 : &\quad X_1, \ldots, X_n \overset{iid}{\sim} P_0 \\ &\text{v.s.} \\ \mathcal{H}_1 : &\quad X_1, \ldots, X_n \overset{iid}{\sim} P_1. \end{aligned}$$

It is of interest to determine whether the samples $X_1, \ldots, X_n$ are drawn iid from P_0 or from P_1. A binary hypothesis "test" for the above problem is a function $g : \mathcal{X}^n \to \{0, 1\}$ that, upon observing the samples $X_1, \ldots, X_n$, declares which among the two hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ is true. Equivalently, a test is also characterised by its rejection region

$$A \coloneqq \{(x_1, \ldots, x_n) \in \mathcal{X}^n : g(x_1, \ldots, x_n) = 0\},$$

the set of n-tuples $(x_1, \ldots, x_n)$ where the test g rejects the alternative hypothesis $\mathcal{H}_1$.

As before, the type-I and type-II error probabilities of a test g with rejection region $A \subseteq \mathcal{X}^n$ are given by $P_0^n(A^c)$ and $P_1^n(A)$ respectively, where P_0^n represents the n-fold product (iid) distribution induced by the n samples $X_1, \ldots, X_n$; P_1^n is defined similarly. Given an error tolerance parameter $\epsilon > 0$, we shall denote by $\beta_\epsilon(P_0^n, P_1^n)$ the minimum type-II error among all tests whose type-I error is at most ϵ, i.e.,

$$\beta_\epsilon(P_0^n, P_1^n) \coloneqq \min_{A \subseteq \mathcal{X}^n} P_1^n(A), \qquad \text{where } A \text{ is such that } P_0^n(A^c) \leq \epsilon.$$

Our interest is in quantifying

$$\lim_{n \to \infty} -\frac{1}{n} \log \beta_\epsilon(P_0^n, P_1^n).$$

Towards this, we note that

$$\log \frac{P_0^n(X_1, \ldots, X_n)}{P_1^n(X_1, \ldots, X_n)} = \sum_{i=1}^{n} \log \frac{P_0(X_i)}{P_1(X_i)}.$$

Furthermore, applying (5.3) and (5.4) to the distributions P_0^n and P_1^n, and using

$$E_{P_0}\left[\log \frac{P_0^n(X_1, \ldots, X_n)}{P_1^n(X_1, \ldots, X_n)}\right] = nD(P_0||P_1), \quad \mathrm{Var}\left(\log \frac{P_0^n(X_1, \ldots, X_n)}{P_1^n(X_1, \ldots, X_n)}\right) = n\mathrm{Var}\left(\log \frac{P_0(X_1)}{P_1(X_1)}\right),$$

we get

$$-\frac{1}{n} \log \beta_\epsilon(P_0^n, P_1^n) \leq D(P_0||P_1) + \sqrt{\frac{\mathrm{Var}\left(\log \frac{P_0(X_1)}{P_1(X_1)}\right)}{n\epsilon/2}} + \frac{1}{n} \log \frac{1}{1-\epsilon},$$

$$-\frac{1}{n} \log \beta_\epsilon(P_0^n, P_1^n) \geq D(P_0||P_1) - \sqrt{\frac{\mathrm{Var}\left(\log \frac{P_0(X_1)}{P_1(X_1)}\right)}{n\epsilon}}.$$

From the above inequalities, we get that for every choice of $\epsilon > 0$,

$$\lim_{n \to \infty} -\frac{1}{n} \log \beta_\epsilon(P_0^n, P_1^n) = D(P_0||P_1).$$

The above relation is known as **Stein's lemma**, and is a fundamental result in information theory that characterises the KL divergence $D(P_0||P_1)$ as the largest exponent in the type-II error probability associated with a binary hypothesis test to decide whether the underlying distribution is P_0 or P_1.

Exercise: Which of the following two cases is harder to distinguish? Below, $\delta > 0$ is a fixed and known constant.

1. An unbiased coin versus a coin which shows heads with probability $0.5 + \delta$.
2. A Sholay coin which always shows heads versus a coin which shows heads with probability $1 - \delta$.

In each of the above cases, compute the number of coin tosses required to resolve the problem at hand.