COM-621: Advanced Topics in Information Theory
Project, Spring 2023

From Simple to Composite Hypothesis Testing Through Universal Distributions

Adway Girish

Last Updated: July 7, 2023

Abstract

For a simple binary hypothesis testing setup, Hoeffding's bound gives the optimal error exponent of one type of error when the other decays exponentially fast, with an exponent smaller than the KL divergence between the null and alternative hypothesis distributions. We consider an extension of this result to a composite hypothesis testing setup. In particular, we study an axiomatic approach based on a "universal distribution", introduced by Tomamichel and Hayashi. By looking at the proof restricted to a specific setting, we try to understand the use of the axiomatic approach, and in particular, the universal distributions.

1 Preliminaries

For completeness, we first summarize the notation used and make some relevant definitions.

1.1 Probabilities and types

Let $\mathcal{X}$ and $\mathcal{Y}$ be two finite sets. For any pair of random variables (X, Y) on $\mathcal{X} \times \mathcal{Y}$, with the joint probability mass function (pmf) P_{XY}, which we denote as $(X, Y) \sim P_{XY}$, we let P_X and P_Y be the marginal pmfs of X and Y respectively. We refer to the n-fold product distribution $P \times \cdots \times P$ on $\mathcal{X}^n$ by P^n. We use $\mathcal{P}(\mathcal{X})$ to denote the set of probability distributions on $\mathcal{X}$.

The *type* [1] of a sequence $x^n \in \mathcal{X}^n$ is the distribution $P \in \mathcal{P}(\mathcal{X})$, given by $P(a) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x_i = a\}$ for all $a \in \mathcal{X}$, also called the empirical distribution. The set of all types of sequences in $\mathcal{X}^n$ is denoted by $\mathcal{T}_n(\mathcal{X})$. Even though the number of sequences in $\mathcal{X}^n$ is $|\mathcal{X}|^n$, which grows exponentially in n, there are at most $(n+1)^{|\mathcal{X}|}$ types, which is polynomial in n. If a distribution only depends on the type of the sequence, it is called a *permutation-invariant* distribution, since the type remains unchanged under permutations. The set of all permutation-invariant distributions on $\mathcal{X}$ is denoted by $\mathcal{P}^{\text{sym}}(\mathcal{X})$. For any set of distributions $\mathcal{Q}$, we use $\mathcal{Q}^{\text{sym}}$ to refer to the set of permutation-invariant distributions in $\mathcal{Q}$.

1.2 Rényi information measures

Given two distributions P and Q on $\mathcal{X}$, we define the Rényi divergence of order α, $\alpha \in (0,1) \cup (1,\infty)$ as

$$D_\alpha(P \,||\, Q) \triangleq \frac{1}{\alpha - 1} \log \sum_{x \in \mathcal{X}: P(x) > 0} P(x)^\alpha Q(x)^{1-\alpha}.$$

When $\alpha = 1$, we define $D_1(P \,||\, Q)$ to be the limit, which happens to be the KL divergence,

$$D_1(P \,||\, Q) \triangleq \lim_{\alpha \to 1} D_\alpha(P \,||\, Q) = \sum_{x \in \mathcal{X}: P(x) > 0} P(x) \log \frac{P(x)}{Q(x)} = D_{\mathsf{KL}}(P \,||\, Q).$$

The classical mutual information between random variables X and Y is related to the KL divergence through the following equivalent formulations:

$$\begin{aligned} I(X;Y) &= D_{\mathsf{KL}}(P_{XY} \,||\, P_X P_Y) \\ &= \min_{Q_Y \in \mathcal{P}(\mathcal{Y})} D_{\mathsf{KL}}(P_{XY} \,||\, P_X Q_Y) \\ &= \min_{Q_X \in \mathcal{P}(\mathcal{X}), Q_Y \in \mathcal{P}(\mathcal{Y})} D_{\mathsf{KL}}(P_{XY} \,||\, Q_X Q_Y). \end{aligned}$$

It is also possible to similarly define a Rényi mutual information in terms of the Rényi divergence, but the above expressions are, in general, not equal for $\alpha \neq 1$. This leads to different defintions of Rényi mutual information, such as that by Arimoto [2] and Sibson [3]. We are interested in Sibson's definition, given by

$$I_\alpha^S(X;Y) = \min_{Q_Y \in \mathcal{P}(\mathcal{Y})} D_\alpha(P_{XY} \,||\, P_X \times Q_Y). \tag{1}$$

2 Simple and Composite Hypothesis Testing, Error Exponents

Consider a simple binary hypothesis testing setup. We have two probability distributions P and Q on a finite set $\mathcal{X}$, and a random variable X is drawn from either P or Q. Given n independently and identically distributed (i.i.d.) samples of X as $X^n = (X_1, \ldots, X_n)$, the task is to identify whether X^n is drawn from P^n or Q^n. This can be stated as follows:

$$\begin{aligned} \text{null hypothesis} &: X^n \sim P^n, \\ \text{alternative hypothesis} &: X^n \sim Q^n. \end{aligned}$$

We make the decision through a test function $T : \mathcal{X}^n \to \{0,1\}$, where $T(X^n) = 1$ (0) means that we accept (reject) the null hypothesis, or equivalently, that we decide that X^n is drawn according to P^n (Q^n). Two kinds of errors are possible: the *type-I error*, given by $\mathsf{p}_n = P^n\{T(X^n) = 0\}$ and the *type-II error*, given by $\mathsf{q}_n = Q^n\{T(X^n) = 1\}$. These are the probabilities of deciding incorrectly when X^n is drawn from P^n and Q^n respectively. Clearly, both cannot be made arbitrarily small simultaneously.

This trade-off is captured through the Chernoff-Stein lemma [4]: If we require that $\mathsf{p}_n \leq \epsilon$ for some constant $\epsilon \in (0,1)$, then the optimal test has a type-II error that decays exponentially with n as $\mathsf{q}_n = \exp\big(-nD_{\mathsf{KL}}(P \,||\, Q) + o(n)\big)$, Conversely, this also means that if $\mathsf{q}_n \leq \exp(-nR)$ for some $R > D_{\mathsf{KL}}(P \,||\, Q)$, then p_n cannot be upper bounded by any constant strictly smaller than 1. On the other hand, if $\mathsf{q}_n \leq \exp(-nR)$ for some $R \in (0, D_{\mathsf{KL}}(P \,||\, Q))$, then we have Hoeffding's bound

[5], which says that the optimal type-I error is not only upper bounded by a constant, but also decays to zero exponentially with n, as

$$\mathsf{p}_n = \exp\Big(-n \sup_{0<\alpha<1} \frac{1-\alpha}{\alpha}\big(D_\alpha(P\,||\,Q) - R\big) + o(n)\Big).$$

These results can also be extended to the case where the alternative hypothesis is *composite*, i.e., $X^n \sim Q_n$, where Q_n belongs to some set of distributions $\mathcal{Q}_n$ (note that we no longer require Q_n to have a product structure). This composite hypothesis problem is then stated as follows:

$$\begin{aligned} \text{null hypothesis} &: X^n \sim P^n, \\ \text{alternative hypothesis} &: X^n \sim Q_n, \text{ for some } Q_n \in \mathcal{Q}_n. \end{aligned}$$

If $\mathcal{Q}_1 = \{Q\}$ and $\mathcal{Q}_n = \mathcal{Q}_1 \times \cdots \times \mathcal{Q}_1$ n times, then this reduces to the simple hypothesis testing problem above. For composite hypothesis testing, the type-I error remains unchanged, but the type-II error is the maximum over all distributions in $\mathcal{Q}_n$, i.e., $\mathsf{q}_n = \max_{Q_n \in \mathcal{Q}_n} Q_n\{T(X^n) = 1\}$.

Tomamichel and Hayashi [6] show that, under some conditions on $\mathcal{Q}_n$, the trade-off between the optimal type-I and type-II errors is given by the same expression as Hoeffding's bound, except with $D_\alpha(P\,||\,Q)$ replaced with

$$D_\alpha(P\,||\,\mathcal{Q}_1) \triangleq \min_{Q \in \mathcal{Q}_1} D_\alpha(P\,||\,Q). \tag{2}$$

The required conditions on $\mathcal{Q}_n$ are:

(Axiom 1) The set $\mathcal{Q}_1$ is compact convex, and the minimizer in (2) is unique and lies in the relative interior of $\mathcal{Q}_1$.

(Axiom 2) The set $\mathcal{Q}_n$ contains the element Q^n for every $Q \in \mathcal{Q}_1$.

(Axiom 3) For all $\alpha > 0$ and $n \in \mathbb{N}$, we have $D_\alpha(P^n\,||\,\mathcal{Q}_n) \geq n D_\alpha(P\,||\,\mathcal{Q})$.

(Axiom 4) There exists a sequence of pmfs $\{U_n\}_{n\in\mathbb{N}}$ with $U_n \in \mathcal{P}^{\text{sym}}(\mathcal{X}^n)$ and a polynomial $v(n)$ such that, for all $n \in \mathbb{N}$ and $Q_n \in \mathcal{Q}_n^{\text{sym}}$, we have

$$\begin{aligned} Q_n(x^n) \leq v(n)\, U_n(x^n), \ \forall\ x^n \in \mathcal{X}^n, \text{ and} \\ D_\alpha(P^n\,||\,U_n) \geq D_\alpha(P^n\,||\,\mathcal{Q}_n). \end{aligned}$$

Further, $\mathcal{Q}_n$ is closed under symmetrization.

Axioms 2 and 3 together imply that the inequality in Axiom 3 actually holds with equality. Axiom 1 can be relaxed to require only a convex re-parametrization on an interval (a, b) containing 1, and both Axiom 1 and its relaxation are only required to show the optimality of the exponent, which we do not consider here. Axiom 4 introduces the notion of "universal distributions", inspired by a similar idea in the quantum setting, which is central to proving the above extension. In the next section, we look at the proof specialized to a particular composite hypothesis testing setup, to illustrate the utility of these universal distributions. The setup we consider also provides an operational meaning to Sibson's mutual information (1), which appears in the error exponent.

3 Sibson's Mutual Information as an Error Exponent

As motivated above, consider the following composite hypothesis testing setup:

$$\begin{aligned} &\text{null hypothesis} : (X^n, Y^n) \sim P_{XY}^n, \\ &\text{alternative hypothesis} : X^n \sim P_X^n, \text{ independent of } Y^n. \end{aligned} \tag{3}$$

Then, Hoeffding's bound can be extended to this setup as follows.

Theorem 1. *For the composite hypothesis setup in (3), there exists a sequence of tests such that the type-II error $\mathsf{q}_n \le \exp(-nR)$ for some $R > 0$, and the type-I error decays exponentially fast as*

$$\mathsf{p}_n = \exp\left(-n \sup_{\alpha \in (0,1)} \frac{1-\alpha}{\alpha}\big(I_\alpha^S(X;Y) - R\big) + o(n)\right).$$

Moreover, this exponent is optimal, i.e., for any sequence of tests such that $\mathsf{q}_n \le \exp(-nR)$, p_n cannot decay any faster than the right-hand side above.

Proof. (achievability only) The alternative hypothesis in (3) is equivalent to saying that $(X^n, Y^n) \in \mathcal{Q}_n$, where $\mathcal{Q}_n = \{P_X^n \times Q_{Y^n} : Q_{Y^n} \in \mathcal{P}(\mathcal{Y}^n)\}$. Let $\{U_n\}_{n\in\mathbb{N}}$ be a sequence of distributions on (X^n, Y^n), given by

$$U_n(x^n, y^n) = P_X^n(x^n) \sum_{T \in \mathcal{T}_n(\mathcal{Y})} \frac{1}{|\mathcal{T}_n(\mathcal{Y})|} \frac{1}{|T|} \mathbb{1}\{y^n \text{ is of type } T\},$$

i.e., the product distribution whose marginal on Y^n is a uniform distribution over all sequences y^n of a given type, with each type picked uniformly as well. It is easy to see that this choice of U_n satisfies Axiom 4. First, for any $Q_n \in \mathcal{Q}_n^{\text{sym}}$, we have $Q_n(x^n, y^n) \le |\mathcal{T}_n(\mathcal{Y})| U_n(x^n, y^n)$. Further, since $U_n \in \mathcal{Q}_n$, we also have $D_\alpha(P^n \,||\, U_n) \ge D_\alpha(P^n \,||\, \mathcal{Q}_n)$.

Now fix an $\alpha \in (0,1)$. Define the sequence of tests

$$T_n(x^n, y^n) = \begin{cases} 1 & \text{if } P_{XY}^n(x^n, y^n) \ge \exp(\lambda_n) U_n(x^n, y^n), \\ 0 & \text{else,} \end{cases}$$

where λ_n is as specified later. We can then upper bound the type-II error q_n as

$$\begin{aligned} &\max_{Q_n \in \mathcal{Q}_n} Q_n\big[P_{XY}^n(X^n, Y^n) \ge \exp(\lambda_n)\, U_n(X^n, Y^n)\big] \\ &= \max_{Q_n \in \mathcal{Q}_n^{\text{sym}}} Q_n\big[P_{XY}^n(X^n, Y^n) \ge \exp(\lambda_n)\, U_n(X^n, Y^n)\big] \\ &\le |\mathcal{T}_n(\mathcal{Y})| \sum_{x^n, y^n} U_n\big[P_{XY}^n(X^n, Y^n) \ge \exp(\lambda_n)\, U_n(X^n, Y^n)\big] \\ &= |\mathcal{T}_n(\mathcal{Y})| \sum_{x^n, y^n} U_n(x^n, y^n) \mathbb{1}\big\{P_{XY}^n(x^n, y^n) \ge \exp(\lambda_n)\, U_n(x^n, y^n)\big\} \\ &\le |\mathcal{T}_n(\mathcal{Y})| \exp(-\alpha\lambda_n) \sum_{x^n, y^n} U_n(x^n, y^n)^{1-\alpha} P_{XY}^n(x^n, y^n)^\alpha \mathbb{1}\big\{P_{XY}^n(x^n, y^n) \ge \exp(\lambda_n)\, U_n(x^n, y^n)\big\} \\ &\le |\mathcal{T}_n(\mathcal{Y})| \exp(-\alpha\lambda_n) \sum_{x^n, y^n} U_n(x^n, y^n)^{1-\alpha} P_{XY}^n(x^n, y^n)^\alpha \\ &= |\mathcal{T}_n(\mathcal{Y})| \exp(-\alpha\lambda_n) \exp\big((\alpha-1) D_\alpha(P_{XY}^n \,||\, U_n)\big), \end{aligned}$$

where the first equality follows since $\mathcal{Q}_n$ is closed under symmetrization and the test T_n is permutation-invariant (thanks to U_n). Choosing $\lambda_n = \frac{1}{\alpha}\big(\log|\mathcal{T}_n(\mathcal{Y})| + nR + (\alpha-1)D_\alpha(P_{XY}^n \,||\, U_n)\big)$, we have that $\mathsf{q}_n \leq \exp(-nR)$, as required. By a similar calculation, we can also upper bound the type-I error p_n as

$$\begin{aligned}\mathsf{p}_n &\leq \exp\left(\frac{1-\alpha}{\alpha}\big(\log|\mathcal{T}_n(\mathcal{Y})| + nR - D_\alpha(P_{XY}^n \,||\, U_n)\big)\right) \\ &\leq \exp\left(\frac{1-\alpha}{\alpha}\big(\log|\mathcal{T}_n(\mathcal{Y})| + nR - D_\alpha(P_{XY}^n \,||\, \mathcal{Q}_n)\big)\right),\end{aligned}$$

where the second step follows from $D_\alpha(P_{XY}^n \,||\, U_n) \geq D_\alpha(P_{XY}^n \,||\, \mathcal{Q}_n)$. All that is left to do is to show that Sibson's mutual information $I_\alpha^S(X;Y) \leq D_\alpha(P_{XY}^n \,||\, \mathcal{Q}_n)$, and then we have the exponent as stated in the theorem. This can be seen from

$$D_\alpha(P_{XY}^n \,||\, \mathcal{Q}_n) = \min_{Q_n \in \mathcal{P}(\mathcal{Y}^n)} D_\alpha(P_{XY}^n \,||\, P_X^n \times Q_n) = D_\alpha(P_{XY}^n \,||\, P_X^n \times Q_n^*),$$

where Q_n^* is the minimum-achieving distribution, given by (up to a normalizing constant factor)

$$Q_n^*(y^n) \propto \left(\sum_{x^n} P_X^n(x^n) P_{Y^n|X^n}(y^n \mid x^n)^\alpha\right)^{\frac{1}{\alpha}} = \prod_{i=1}^n \left(\sum_x P_X(x) P_{Y|X}(y_i \mid x)^\alpha\right)^{\frac{1}{\alpha}} = \prod_{i=1}^n Q_1^*(y_i),$$

i.e., the minimum-achieving distributions have a product structure. Hence, we also have that $D_\alpha(P_{XY}^n \,||\, \mathcal{Q}_n) = nD_\alpha(P_{XY} \,||\, \mathcal{Q}_1) = n \min_{Q \in \mathcal{P}(\mathcal{Y})} D_\alpha(P_{XY} \,||\, P_X \times Q) = nI_\alpha^S(X;Y)$, and we are done. □

The key role that the universal distribution plays is in bounding the type-I and type-II errors simultaneously. They "dominate" (permutation-invariant) Q_n up to a polynomial factor which vanishes in the exponent, while still having a larger divergence with P^n. Attempts to replace U_n with other functions of Q_n such as the maximum or minimum prove futile; for these choices, when one of the errors is bounded satisfactorily, the other cannot be.

4 Conclusion

We have looked at an extension of results from simple hypothesis testing to composite hypothesis testing via an axiomatic framework. The most powerful of these axioms is that involving the universal distributions. They provide a convenient way to deal with maxima over arbitrary sets by allowing us to restrict the maximization to just permutation-invariant sets. It would be interesting to study what these universal distributions actually represent, for more general classes of distributions, and whether requiring their existence is a particularly strong condition. Another direction that looks promising is to see if these universal distribution techniques can be applied to other settings not limited to hypothesis testing, to show, for example, achievability results in channel coding setups.

References

[1] I. Csiszar, "The method of types," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2505–2523, 1998. DOI: 10.1109/18.720546.

[2] S. Arimoto, "Information measures and capacity of order α for discrete memoryless channels," *Topics in information theory*, 1977.

[3] R. Sibson, "Information radius," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 14, no. 2, pp. 149–160, 1969.

[4] H. Chernoff, "Large-sample theory: Parametric case," *The Annals of Mathematical Statistics*, vol. 27, no. 1, pp. 1–22, 1956.

[5] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *The Annals of Mathematical Statistics*, pp. 369–401, 1965.

[6] M. Tomamichel and M. Hayashi, "Operational interpretation of Rényi information measures via composite hypothesis testing against product and Markov distributions," *IEEE Transactions on Information Theory*, vol. 64, no. 2, pp. 1064–1082, 2018. DOI: 10.1109/TIT.2017.2776900.