Approximate Hypothesis Testing

Nicolas Le Gouic and Robert Graczyk
Information Processing and Communications Laboratory
Télécom Paris, Institut Polytechnique de Paris
Palaiseau, France
nicolas.legouic@ip-paris.fr, robert.graczyk@telecom-paris.fr

Stefan Moser
Signal and Information Processing Laboratory
ETH Zurich
Zurich, Switzerland
moser@isi.ee.ethz.ch

***Abstract*—We establish the sample complexity of Approximate Hypothesis Testing (AHT) where—unlike in classical hypothesis testing—we need only approximate the hypothesis governing the observed samples rather than recover it exactly.**

We show that the AHT sample complexity scales inversely with the multivariate Bhatthacharyya distance evaluated on a "maximally confusable" subset of hypotheses that is characterized by the chosen distance measure and approximation accuracy.

***Index terms*—hypothesis testing, sample complexity, learning, Bhattacharyya distance, Hellinger distance.**

I. Introduction

In Approximate Hypothesis Testing, we seek to estimate an unknown distribution (hypothesis) based on observed samples. Unlike in classical hypothesis testing, the estimate need not be exact, but merely "close" to the distribution that generates the observed samples.

Concretely, we are given n samples $\boldsymbol{X} = (X_1, ..., X_n)$ that are independent and identically distributed (IID) according to some P residing in a predefined finite hypothesis class $\mathcal{H} \subseteq \mathcal{P}(\mathcal{X})$, where $\mathcal{P}(\mathcal{X})$ denotes the set of probability distributions on $\mathcal{X}$. Based on the observed samples $\boldsymbol{X}$, we seek an estimate $\hat{P}$ that is ε-close to P as measured by a distance[1]

$$d : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}_{\geq 0}. \tag{1}$$

We consider a minimax setting where, for any sample-generating distribution P, some $\hat{P}$ that is ε-close must be found with probability at least $1 - \delta$. Our goal is to find the smallest n—i.e., the smallest number of samples—permitting us to do so. We refer to this n as the AHT sample complexity and denote it $n_{\varepsilon,\delta}^{\mathrm{AHT}}(\mathcal{H}, d)$. To determine the AHT sample complexity, we proceed in two steps:

Step 1 We state AHT in terms of yet another hypothesis testing problem: Clustered Hypothesis Testing (CHT). Here, each hypothesis $P \in \mathcal{H}$ is assigned to at least one cluster $\mathcal{C} \in \mathfrak{C}$, with $\mathfrak{C} \subseteq 2^{\mathcal{H}}$. Based on the observed samples $\boldsymbol{X}$, we seek to find a cluster $\mathcal{C}$ containing the data-generating distribution P. We refer to the smallest number of samples permitting us to succeed with probability at least $1 - \delta$ as the CHT sample complexity, denoted $n_{\delta}^{\mathrm{CHT}}(\mathcal{H}, \mathfrak{C})$. By a judicious construction of the cluster family $\mathfrak{C}$—based on the distance $d(\cdot, \cdot)$ and desired accuracy ε—we show that

$$n_{\varepsilon,\delta}^{\mathrm{AHT}}(\mathcal{H}, d) = n_{\delta}^{\mathrm{CHT}}(\mathcal{H}, \mathfrak{C}). \tag{2}$$

Step 2 We solve the CHT problem by meticulously rejecting distributions that *did not* generate the observed samples. We show that it is a worst-case instance in the Rejection Hypothesis Testing (RHT) problem—corresponding to a "maximally confusable" subset of hypotheses $\mathcal{D} \subseteq \mathcal{H}$—that characterizes $n_{\delta}^{\mathrm{CHT}}(\mathcal{H}, \mathfrak{C})$ (and so $n_{\varepsilon,\delta}^{\mathrm{AHT}}(\mathcal{H}, d)$).

Step 1 and Step 2 are presented in Section III and Section IV, respectively. Together, they yield our main result.

Theorem 1. When the hypothesis classes $\mathcal{H}$ is finite, the AHT sample complexity $n_{\varepsilon,\delta}^{\mathrm{AHT}}(\mathcal{H}, d)$ can be bounded as follows:

$$n_{\varepsilon,\delta}^{\mathrm{AHT}}(\mathcal{H}, d) \geq \frac{\ln(1/\delta) - \ln|\mathfrak{C}| - 1}{|\mathfrak{C}| \cdot \min\limits_{\mathcal{D} \in \mathfrak{D}} D_{\mathrm{B}}(\mathcal{D})}, \tag{3.a}$$

$$n_{\varepsilon,\delta}^{\mathrm{AHT}}(\mathcal{H}, d) \leq \frac{\ln(1/\delta) + \ln|\mathcal{H}|}{\min\limits_{\mathcal{D} \in \mathfrak{D}} D_{\mathrm{B}}(\mathcal{D})}. \tag{3.b}$$

Here, $\mathfrak{C}$ denotes the family of clusters pertaining to the CHT problem (Definition 1); $\mathfrak{D} \subseteq 2^{\mathcal{H}}$ is the family of "confusable" hypotheses pertaining to the RHT problem (Definition 2); and

$$\begin{aligned} D_{\mathrm{B}} : 2^{\mathcal{P}(\mathcal{X})} &\to \mathbb{R}_{\geq 0} \cup \{\infty\} \\ \mathcal{P} &\mapsto -\ln\left(\int_{x \in \mathcal{X}} \sqrt[|\mathcal{P}|]{\prod_{P \in \mathcal{P}} \mathrm{d}P(x)}\right) \end{aligned} \tag{4}$$

denotes the multivariate Bhatthacharyya distance of which we list some properties following (11) ahead.

Our result highlights the principal dependancy of $n_{\varepsilon,\delta}^{\mathrm{AHT}}(\mathcal{H}, d)$ on $\min_{\mathcal{D} \in \mathfrak{D}} D_{\mathrm{B}}(\mathcal{D})$, i.e., on a subset of hypotheses that can be easily confused with one another, and so require many samples to be told apart reliably.

Due to their intricate dependence on the hypothesis class $\mathcal{H}$, the distance $d(\cdot, \cdot)$, and the approximation accuracy ε, further assumptions are needed to simplify the bounds in (3). We refer to Section V for an edifying example.

[1] A distance $d(\cdot, \cdot)$ is a pseudometric [1, p. 119] that need not satisfy the triangle inequality, i.e., $d(P, Q) = d(Q, P) \geq 0$ with equality if $P = Q$.

II. Problem Statement

Let $\boldsymbol{X} = (X_1, ..., X_n)$ be a sequence of samples that are IID according to a distribution P only known to reside in some finite hypothesis class $\mathcal{H} \subseteq \mathcal{P}(\mathcal{X})$. Our aim is to determine the smallest number of samples n required to produce, given the samples $\boldsymbol{X}$, an approximation $\hat{P}$ of P satisfying

$$d\big(P, \hat{P}\big) \leq \varepsilon \tag{5}$$

for some distance $d(\cdot, \cdot)$ as specified in (1). Since we consider a minimax problem setting, we need enough samples such that an approximation $\hat{P}$ satisfying (5) exist for every $P \in \mathcal{H}$. The (minimax) AHT sample complexity $n^{\mathrm{AHT}}_{\varepsilon,\delta}(\mathcal{H}, d)$ is therefore formally defined as

$$\begin{aligned} &n^{\mathrm{AHT}}_{\varepsilon,\delta}(\mathcal{H}, d) \\ &\quad := \inf_n \left\{ \inf_{\hat{P}} \max_{P \in \mathcal{H}} \mathop{\mathbb{P}}_{\boldsymbol{X} \sim \mathrm{IID}\, P} \Big[d\big(\hat{P}(\boldsymbol{X}), P\big) > \varepsilon \Big] \leq \delta \right\}, \end{aligned} \tag{6}$$

where the inner infimum is over estimators $\hat{P} : \mathcal{X}^n \to \mathcal{P}(\mathcal{X})$, and where we assume that the problem parameters $\mathcal{H}$, $d(\cdot, \cdot)$, and ε guarantee the existence of $n^{\mathrm{AHT}}_{\varepsilon,\delta}(\mathcal{H}, d)$.

Rather than studying AHT directly (6), we shall consider the following surrogate problem.

III. Clustered Hypothesis Testing

In the Clustered Hypothesis Testing (CHT) problem, each $P \in \mathcal{H}$ is assigned to *at least* one cluster $\mathcal{C} \in \mathfrak{C} \subseteq 2^{\mathcal{H}}$, with $2^{\mathcal{H}}$ denoting the power set of $\mathcal{H}$. The aim of the problem is to find, based on the observed samples $\boldsymbol{X}$, any one cluster $\mathcal{C}$ containing the sample-generating distribution P. The smallest number of samples n whereby we succeed with probability at least $1 - \delta$ is referred to as the CHT sample complexity and denoted $n^{\mathrm{CHT}}_{\delta}(\mathcal{H}, \mathfrak{C})$,

$$\begin{aligned} &n^{\mathrm{CHT}}_{\delta}(\mathcal{H}, \mathfrak{C}) \\ &\quad := \inf_n \left\{ \inf_{\hat{C}} \max_{P \in \mathcal{H}} \mathop{\mathbb{P}}_{\boldsymbol{X} \sim \mathrm{IID}\, P} \Big[P \notin \hat{C}(\boldsymbol{X}) \Big] \leq \delta \right\}, \end{aligned} \tag{7}$$

where the inner infimum is over estimators $\hat{\mathcal{C}} : \mathcal{X}^n \to \mathfrak{C}$, and where the problem parameters are yet again assumed to guarantee the existence of $n^{\mathrm{CHT}}_{\delta}(\mathcal{H}, \mathfrak{C})$.

Example 1. Depicted in Fig. 1 are a hypothesis class $\mathcal{H}$ and cluster family $\mathfrak{C}$. Note that $\mathfrak{C}$ does not partition $\mathcal{H}$, as P_1, P_2, and P_3 are each assigned to two clusters.

Remark 1. *Composite Hypothesis Testing* [2, Chapter 16.4] is a well-known instance of CHT where $\mathcal{H}$ is partitioned into two disjoint clusters; we derive its sample complexity as a straightforward collorary of Theorem 1 in Section V.

The role of CHT as a surrogate for AHT can be motivated by a cluster family $\mathfrak{C}_\varepsilon$ comprising all ε-balls $\mathcal{B}_\varepsilon(Q)$ on $\mathcal{H}$,

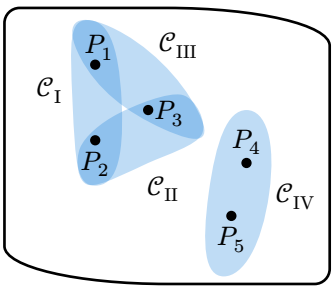

Fig. 1: Hypothesis class $\mathcal{H} = \{P_1, P_2, P_3, P_4, P_5\}$ and cluster family $\mathfrak{C} = \{\mathcal{C}_{\mathrm{I}}, \mathcal{C}_{\mathrm{II}}, \mathcal{C}_{\mathrm{III}}, \mathcal{C}_{\mathrm{IV}}\}$.

$$\mathcal{B}_\varepsilon(Q) := \{P \in \mathcal{H} : d(P, Q) \leq \varepsilon\}, \tag{8.a}$$

$$\mathfrak{C}_\varepsilon := \{\mathcal{B}_\varepsilon(Q) : Q \in \mathcal{P}(\mathcal{X})\}. \tag{8.b}$$

Note that $\mathfrak{C}_\varepsilon$ may contain clusters $\mathcal{C}$ and $\mathcal{C}'$ such that $\mathcal{C} \subset \mathcal{C}'$. Therefore, rather than considering $\mathfrak{C}_\varepsilon$ directly, we shall find it more convenient to consider only its $\subseteq$-maximal elements.

Definition 1 *((d, ε)-Canonical Clustering)*. A cluster family $\mathfrak{C}$ on $\mathcal{H}$ is (d, ε)-canonical if it contains all $\mathcal{C} \in \mathfrak{C}_\varepsilon$ (8.b) that are maximal w.r.t. to the $\subseteq$-preorder[2] and no other $\mathcal{C} \in \mathfrak{C}_\varepsilon$.

With Definition 1 at hand, we may now state the raison d'être of the CHT problem.

Theorem 2. Let $\mathfrak{C}$ be a (d, ε)-canonical cluster family. Then,

$$n^{\mathrm{AHT}}_{\varepsilon,\delta}(\mathcal{H}, d) = n^{\mathrm{CHT}}_{\delta}(\mathcal{H}, \mathfrak{C}). \tag{9}$$

Proof *(Sketch)*. We establish Theorem 2 by reducing AHT to CHT and vice-versa. For the former direction, consider the ball $\mathcal{B}_\varepsilon\big(\hat{P}\big)$, where $\hat{P}$ is the output of an algorithm solving the AHT problem. By (8) and Definition 1, $\hat{P}$ identifies a cluster $\hat{\mathcal{C}} \in \mathfrak{C}$ guaranteed to contain the sample-generating distribution P if (5) is satisfied. For the latter direction (reducing AHT to CHT) consider a cluster $\hat{\mathcal{C}}$ as the output of an algorithm solving the CHT problem. Reversing the preceding argument, $\hat{\mathcal{C}}$ identifies some $\hat{P}$ satisfying (5) if the sample-generating distribution P lies in $\hat{\mathcal{C}}$. ∎

To solve the CHT problem and characterize $n^{\mathrm{CHT}}_{\delta}(\mathcal{H}, \mathfrak{C})$ (and, by Theorem 2, $n^{\mathrm{AHT}}_{\varepsilon,\delta}(\mathcal{H}, d)$) we shall take one last detour.

IV. Rejection Hypothesis Testing

The aim in Rejection Hypothesis Testing (RHT) is to find, based on the observed samples $\boldsymbol{X}$, some $P \in \mathcal{H}$ that *did not* generate $\boldsymbol{X}$.

To motivate RHT in the context of CHT, we shall focus for a moment on hypotheses $\{P_1, P_2, P_3\}$ of Example 1. Suppose that $P = P_1$: by rejecting P_2, we assert that $P \in \mathcal{C}_{\mathrm{III}}$, whereas, by rejecting P_3, we assert that $P \in \mathcal{C}_{\mathrm{I}}$. In either case, we have identified a cluster containing P.

[2] $\omega \in \Omega$ is maximal w.r.t. to a preorder $\leq$ if for no other $\omega' \in \Omega$, $\omega \leq \omega'$; cf. [3, p. 121].

We generalize this observation by introducing the notion of delta sets, which capture the equivalence between rejecting a hypothesis and identifying a cluster.

Definition 2 *(Delta Sets)*. The delta sets $\mathfrak{D}$ (defined for a cluster family $\mathfrak{C}$ on $\mathcal{H}$) consist of all $\mathcal{D} \in 2^{\mathcal{H}} \setminus \mathfrak{C}_\varepsilon$ that are minimal w.r.t. the $\subseteq$-preorder.

Continuing Example 1, we list the delta sets $\mathfrak{D}$ corresponding to the cluster family $\mathfrak{C}$ in Table I below; $D_1, D_2, D_3 \in \mathfrak{D}$ are additionally highlighted in Fig. 2. Note that removing any P from $\mathcal{D} \in \mathfrak{D}$ identifies a cluster $\mathcal{C} \in \mathfrak{C}$, as desired.

TABLE I: DELTA SETS $\mathfrak{D}$ CORRESPONDING TO THE CLUSTER FAMILY $\mathfrak{C}$ OF EXAMPLE 1.

i	I	II	III	IV	V	VI	VII
$\mathcal{D}_i$	P_1 P_2, P_3	P_2 P_5	P_3 P_4	P_2 P_4	P_3 P_5	P_1 P_4	P_1 P_5

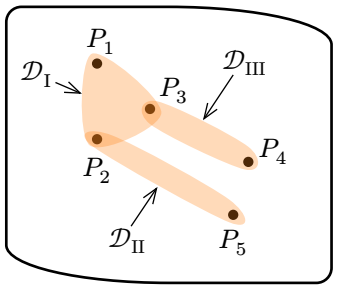

Fig. 2: Highlighted delta sets $\{\mathcal{D}_{\mathrm{I}}, \mathcal{D}_{\mathrm{II}}, \mathcal{D}_{\mathrm{III}}\} \subset \mathfrak{D}$.

We formally link RHT and CHT in the theorem below.

Theorem 3. Assume that the sample-generating distribution P lies in a delta set $D \in \mathfrak{D}$. Then, CHT and RHT are reducible to one another.

Proof *(Sketch)*. Reducing RHT to CHT: Consider $\hat{\mathcal{C}}$, the output of an algorithm for the CHT problem. By Definition 2, $\mathcal{D} \setminus \hat{\mathcal{C}}$ contains a $\hat{P}$ that did not generate $\boldsymbol{X}$ if $P \in \hat{\mathcal{C}}$.

Reducing CHT to RHT: Consider $\hat{P} \in \mathcal{D}$, a distribution rejected by an algorithm for the RHT problem. By Definition 2, $\mathcal{D} \setminus \left\{\hat{P}\right\} \subseteq \hat{\mathcal{C}}$ for some cluster $\hat{\mathcal{C}} \in \mathfrak{C}$ which is guaranteed to contain the distribution P that generated $\boldsymbol{X}$ if $\hat{P}$ did not. ■

The main consequence of Theorem 3, which we state without proof for brevity, is a characterization of $n_\delta^{\mathrm{CHT}}(\mathcal{H}, \mathfrak{C})$ via the multivariate Bhatthacharyya distance when $\mathcal{H}$ is restricted to a delta set $\mathcal{D} \in \mathfrak{D}$,

Lemma 1. When the samples $\boldsymbol{X}$ are generated according to a distribution P from a delta set $\mathcal{D}$ (defined w.r.t. $(\mathcal{H}, \mathfrak{C})$),

$$n_\delta^{\mathrm{CHT}}(\mathcal{H}, \mathfrak{C}) \propto \frac{\ln(1/\delta)}{D_B(\mathcal{D})}. \tag{10}$$

Theorem 1 follows from Lemma 1 by invoking Theorem 2 and showing that it is the sample complexity on a worst-case delta set that characterizes $n_\delta^{\mathrm{CHT}}(\mathcal{H}, \mathfrak{C})$ and, in turn, $n_{\varepsilon,\delta}^{\mathrm{AHT}}(\mathcal{H}, d)$.

To understand the principal dependence of $n_{\varepsilon,\delta}^{\mathrm{AHT}}(\mathcal{H}, d)$ on the (inverse) multivariate Bhatthacharyya distance $D_{\mathrm{B}}(\mathcal{P})$, we shall briefly expand on the latter. To that end, we first express $D_{\mathrm{B}}(\mathcal{P})$ as $-\ln(B(\mathcal{P}))$, with

$$\begin{aligned} B : 2^{\mathcal{P}(\mathcal{X})} &\to \mathbb{R}_{\geq 0} \\ \mathcal{P} &\mapsto \int_{x \in \Omega} \sqrt[|\mathcal{P}|]{\prod_{P \in \mathcal{P}} \mathrm{d}P(x)} \end{aligned} \tag{11}$$

denoting the multivariate Bhatthacharyya coefficient. We note that $B(\mathcal{P})$ satisfies various properties [4] that make it a good similarity measure for distributions: 1. lies between zero and one; 2. equals zero iff some $P, P' \in \mathcal{P}$ have disjoint support; 3. equals one iff all $P \in \mathcal{P}$ are equal; 4. anti-monotonic in $\mathcal{P}$, $\mathcal{P} \subseteq \mathcal{P}' \Rightarrow B(\mathcal{P}) \geq B(\mathcal{P}')$.

To the best of our knowledge and search efforts, this work presents the first operational interpretation of the multivariate Bhatthacharyya coefficient.

V. COMPOSITE HYPOTHESIS TESTING

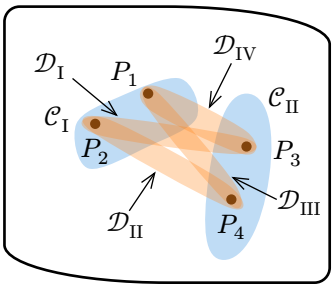

Fig. 3: Disjoint clusters $\mathcal{C}_{\mathrm{I}} = \{P_1, P_2\}$ and $\mathcal{C}_{\mathrm{II}} = \{P_3, P_4\}$ along with their delta sets $\mathcal{D}_1$, $\mathcal{D}_2$, $\mathcal{D}_3$, and $\mathcal{D}_4$.

When $\mathfrak{C}$ consists of two disjoint clusters $\mathcal{C}_1$ and $\mathcal{C}_2$, the CHT problem becomes the Composite Hypothesis Testing problem as introduced in Remark 1. The setup is exemplified in Fig. 3 above. We show how to derive the sample complexity of this problem by means of Theorem 1.

With $d(P, P') = 0$ when $\{P, P'\} = \{P_1, P_2\}$ or $\{P_3, P_4\}$, and ∞ otherwise, we see that irrespective of $\varepsilon > 0$, Theorem 1 asserts that $n_{\varepsilon,\delta}^{\mathrm{AHT}}(\mathcal{H}, d)$ scales inversely with

$$\min_{P \in \{P_1, P_2\}, P' \in \{P_3, P_4\}} D_{\mathrm{B}}(\{P, P'\}), \tag{12}$$

where the minimization encompasses all delta sets. Observing that the squared Hellinger distance $h^2(P, P')$ can be expressed as $1 - B(P, P')$, we have for $h^2(P, P') \leq \frac{1}{2}$ (cf. [2] and [5]),

$$\frac{1}{2} h^2(P, P') \leq D_{\mathrm{B}}(\{P, P'\}) \leq h^2(P, P'). \tag{13}$$

Theorem 1 thus recovers the sample complexity of Composite Hypothesis Testing [2, Chapter 32.2.1]:

$$n^{\text{Composite HT}} \propto \frac{\ln(1/\delta)}{\min\limits_{P \in \mathcal{C}_1, P' \in \mathcal{C}_2} h^2(P, P')}. \tag{14}$$

References

[1] J. L. Kelley, *General Topology*. Mineola, NY: Dover Publications, 2017.

[2] Y. Polyanskiy and Y. Wu, *Information Theory: From Learning to Coding*. Cambridge University Press, 2024.

[3] B. Richmond and T. Richmond, *A Discrete Transition to Advanced Mathematics*. AMS, American Mathematical Society, 2023.

[4] S. M. Kang and R. P. Wildes, "The n-distribution Bhattacharyya Coefficient", 2015.

[5] Z. Bar-Yossef, "The Complexity of Massive Dataset Computations," 2002.