Hypothesis testing with active information

Daniel Andrés Díaz–Pachón[a,*], Juan Pablo Sáenz[b], J. Sunil Rao[a]

[a]*Division of Biostatistics - University of Miami, Don Soffer Clinical Research Center, 1120 NW 14th St, Miami FL, 33136*
[b]*Department of Industrial Engineering - University of Miami, McArthur Engineering Building, 1251 Memorial Drive, Coral Gables FL, 33146*

Abstract

We develop hypothesis testing for active information —the averaged quantity in the Kullback-Liebler divergence. To our knowledge, this is the first paper to derive *exact* probabilities of type-I errors for hypothesis testing in the area.

1. Introduction

The No Free Lunch Theorems (NFLT) [10, 11], introduced by Wolpert and Macready, proved that no search behaves better on average than a blind search. For instance, selecting a point according to a standard normal r.v. restricted to the interval $[-x, x]$ might be good at finding a target close to 0, but it won't work as well towards the limits of the interval; other r.v.'s will do better in the latter case.

However, in applications, searches like evolutionary algorithms usually do better than blind chance. Wolpert and Macready attributed it to the incorporation of "problem-specific knowledge into the behavior of the [optimization or search] algorithm." Active information (actinfo) was thus introduced by Dembski and Marks in order to measure this amount of knowledge infused by the programmer to reach a given target [2, 3]. Actinfo is obtained by measuring how much information is infused in an algorithm in order to reach a given target and then subtracting from it the information of reaching that same target by a uniformly-at-random search.

Recently, Montañez proposed a model unifying different versions of complex specified information, setting in turn the stage to use actinfo in hypothesis testing [9]. In this paper we build on Montañez's idea of hypothesis testing. Thus, the main goal of this article is to set a general framework for hypothesis testing. Accordingly, we begin by doing a basic review of active information and the framework of hypothesis testing developed by Montañez.

1.1. Active information

Let's consider a search space Ω and a target $T \subset \Omega$ such that $|T| \ll |\Omega|$, where $|\cdot|$ stands for cardinality. The NFLT show that no search behaves better on average than

---

*Corresponding author
Email addresses: `Ddiaz3@miami.edu` (Daniel Andrés Díaz–Pachón), `j.saenz4@umiami.edu` (Juan Pablo Sáenz), `JRao@biostat.med.miami.edu` (J. Sunil Rao)

$\mathbf{U}(\cdot)$, where $\mathbf{U}$ is uniform over Ω. (This implies that $|\Omega| < \infty$, since in any other case the uniform distribution does not exist.) Let's define the endogenous information as $-\log \mathbf{U}(T)$; we will denote it by I_0. Defined this way, I_0 measures the inherent difficulty of reaching T.

Of course, alternative searches can —and must— be developed once relevant additional knowledge is acquired. Partial or total knowledge on the position of the target or the space structure might alter the chances of reaching T. Such knowledge assigns a new probability p of reaching T. We now call $-\log p$ the exogenous information, and we will denote it by I_1.

The difference $I_+ := I_0 - I_1 = \log(p/\mathbf{U})$ is called active information. It measures the amount of information added to the search by the programmer with respect to the one provided by the blind search.

1.2. Hypothesis testing with active information

In order to explain Montañez's ideas in [9], we begin with a null hypothesis that is rejected whenever its p-value, p_{val}, satisfies that $p_{\mathrm{val}} < \alpha$ for $\alpha \in (0, 1)$. Or, equivalently, we reject the null hypothesis when

$$\log \frac{p_{\mathrm{val}}}{\alpha} < 0. \tag{1}$$

Since, under the null hypothesis, p-values are uniformly distributed in $(0, 1)$ when the test statistic defining the p-value is continuous,

$$\mathbf{P}[p_{\mathrm{val}} < \alpha] \leq \alpha. \tag{2}$$

From this, we obtain directly that

$$\mathbf{P}\left[\log \frac{p_{\mathrm{val}}}{\alpha} < x\right] \leq \alpha \exp(x). \tag{3}$$

This last equation corresponds to Theorem 1 of [9]. The following Theorem can be easily understood from the previous discussion:

Theorem 1 (Theorem 2 of [9]: Conservation of canonical specified complexity)**.** *Let $p(x)$ be any probability measure on space $\mathcal{X}$, and let $v : \mathcal{X} \to \mathbb{R}^+$ be an integrable function, such that $v(\mathcal{X}) \leq r$ for a constant $r \in \mathbb{R}^+$. For $X \sim p$,*

$$\boldsymbol{P}\left[-\log r \frac{p(X)}{v(X)} \geq x\right] \leq \exp(-x). \tag{4}$$

Notice that up to this point we have not specified the base of the logarithm, since in principle it could be given in different units —bits, trits, dits, nats, etc. Throughout this article, unless we explicitly specify the base, we are not assuming any particular unit.

2. Coin tossing

Imagine that we are tossing a coin. Against the backdrop of a search, p can be thought as the probability of reaching the target "heads" under a new search strategy, and $1/2$ as the default probability of reaching "heads" in the absence of further knowledge. For these

reasons, appealing to the definitions in the previous section, we call them respectively exogenous probability and endogenous probability. When these probabilities are close, $\log p/(1/2) \approx 0$. Of course, we know that if p is too far removed from $1/2$, say $p = 1$, then the distribution is not uniform. But how close is close enough so that it is not justified to reject the hypothesis that the probability p is effectively the same as that of a uniform? Active information allows us to link information to rejection regions specifying how many information units we need in an event in order to reject the null hypothesis to a given α-level:

$$\mathbf{H}_0 : 0 \approx I_+. \tag{5}$$

2.1. One-sided test

In order to be able to use I_+ as a test statistic, it has to be random. So let's suppose p is distributed as a continuous uniform random variable in $(0, 1]$. (This is equivalent to set ourselves in a Bayesian framework in which we have a Bernoulli distribution with random parameter p such that the parameter is uniformly distributed in $(0, 1]$.) In bits, this distribution is given by:

$$\mathbf{P}\left[I_+ \leq b\right] = \mathbf{P}\left[\log_2 \frac{p}{1/2} \leq b\right] = \begin{cases} 2^{b-1}, & \text{if } b \in (-\infty, 1]; \\ 1, & \text{if } b > 1. \end{cases} \tag{6}$$

This enables us to find the probability of type-one errors exactly:

$$\mathbf{P}\left[I_+ > b\right] < \alpha \Leftrightarrow b > 1 + \log_2(1 - \alpha), \tag{7}$$

for $0 < \alpha < 1$. Table 1 in the supplementary material gives α-levels from (7).

2.2. Two-sided test

Since I_+ can also be negative (when $p < 1/2$), it is needed to consider a deviation from the two sides, which in nats becomes:

$$\mathbf{P}\left[|I_+| \leq n\right] = \begin{cases} \sinh n, & n \leq \ln 2; \\ 1 - \frac{\cosh n - \sinh n}{2}, & n > \ln 2. \end{cases} \tag{8}$$

Remark 1. The interesting result obtained in (8) in terms of the hyperbolic functions explains why we took the base e. Notice then that the density function of the r.v. $Y = |I_+|$ is given by

$$f_Y(n) = \begin{cases} \cosh n & n \leq \ln 2; \\ \frac{\cosh n - \sinh n}{2} & n > \ln 2. \end{cases} \tag{9}$$

In other words, the density of $|I_+|$ can be seen as a catenary function $y = \cosh n$, and its distribution as the arc length of the catenary from 0 to n, provided $n \leq \ln 2$.

Going to nats provides an easy way to find an α-value, since the inverses of hyperbolic functions are well-known. We obtain from (8) that

$$\mathbf{P}\left[|I_+| > n\right] = \begin{cases} 1 - \sinh n, & n \leq \ln 2; \\ \frac{e^{-n}}{2}, & n > \ln 2. \end{cases} \tag{10}$$

3

Thus, for $\mathbf{P}\left[|I_+| > n\right] < \alpha$, we obtain that

$$
n > \begin{cases} \ln\left(1 - \alpha + \sqrt{(1-\alpha)^2 + 1}\right), & \alpha \le 1/4; \\ -\ln(2\alpha), & \alpha > 1/4. \end{cases} \tag{11}
$$

Table 2 in the supplementary material gives some α-levels from (11).

3. General discrete uniform $U(N)$

We have detailed the analysis for a discrete uniform r.v. with two points. But this can be easily generalized to any $N \in \mathbb{N}$. That is, consider $\Omega = \{1, \ldots, N\}$. Let $X \sim U(N)$, and let Y be an arbitrary r.v. in the same space.

If for a given singleton the probability under Y is p, the active information becomes

$$
I_+ = \log_N \frac{p}{1/N} = 1 + \log_N(p). \tag{12}
$$

Here we consider the base of the logarithm to be N, so that we measure our information in "N-its". This simplifies the notation and can be easily converted to bits by means of the equation $N^x = 2^b$. Assuming $p \sim \mathcal{U}((0,1])$, the distribution of I_+ becomes

$$
\mathbf{P}\left[I_+ \le x\right] = \begin{cases} N^{x-1}, & x \in (-\infty, 1]; \\ 1, & x > 1; \end{cases} \tag{13}
$$

From (13), when $x < 1$,

$$
\mathbf{P}\left[I_+ > x\right] < \alpha \Leftrightarrow 1 + \log_N(1 - \alpha) < x, \tag{14}
$$

and $\mathbf{P}\left[I_+ > x\right] = 0$ whenever $x > 1$.

On the other hand, in the two-sided case, considered again in nats:

$$
\mathbf{P}\left[|I+| \le n\right] = \begin{cases} \frac{2}{N}\sinh n, & n \le \ln N; \\ 1 - \frac{\cosh n - \sinh n}{N}, & n > \ln N. \end{cases} \tag{15}
$$

So that for $\mathbf{P}\left[|I+| > n\right] < \alpha$, we obtain

$$
n > \begin{cases} \ln\left(\frac{(1-\alpha)N}{2} + \sqrt{\left(\frac{(1-\alpha)N}{2}\right)^2 + 1}\right), & \alpha \le \frac{1}{N}; \\ -\ln(\alpha N), & \alpha > \frac{1}{N}. \end{cases}
$$

4. General prior F

Our guiding principle has been to set p as a uniform r.v. But what if we want to use a different prior for p? For instance, taking Jeffrey's prior, we could consider $p \sim \beta(\frac{1}{2}, \frac{1}{2})$. Then, provided that we know the prior distribution for p, the actual distribution of I_+

4

can be derived directly. For instance, for a discrete space with N elements, the test statistic has distribution

$$\mathbf{P}[I_+ \le b] = F\left(\frac{2^b}{N}\right),$$

$$\mathbf{P}[I_+ \le n] = F\left(\frac{e^n}{N}\right),$$

where the first distribution is given in bits and the second in nats, and F is the prior, the distribution of p.

Finally, up to this point we have referenced actinfo to a discrete uniform rv with parameter N, since it has maxent over all finite spaces of size N, making it a natural choice to compare against [4]. In Section 1 of the supplementary material, we present a general version that removes this condition.

5. Discussion

One of the main strengths of actinfo in comparison with other strategies is that it provides an objective way to measure differences: although a quotient of probabilities can be somewhat obscure to interpret, the logarithm of that quotient has a more straightforward interpretation in terms of bits (or any other information unit) added. Another advantage is that the quotient of the probabilities of a given event of two dimensioned r.v.'s is dimensionless. This seems to support the idea that actinfo is a more basic unit than others in statistics and information theory. In fact, this is highlighted by the fact that the Kullback-Liebler distance —the average of the actinfo— is invariant under parameter transformation (see [8], p. 18-22), a fact of relevance both in Bayesian and frequentist statistics.

We have constructed test statistics building from the most simple case of a space with two singletons to the more general situations. Two differences are worth mentioning with respect to Montanez's results in [9]. First, Montanez developed some tables similar to ours. However, since he considered $-\log(p/v)$ instead of $\log(p/v)$, the values he obtained differ from ours. It seems to us more intuitive to consider the specification function v in the denominator. In fact, when the specification is the endogenous distribution and the alternate distribution (search) is given by p, it is natural to consider this order, since it coincides with what was written in Section 1.1. Second, knowing specifically the distribution of p enables us to find the exact probabilities of the false positives, which is more potent than the inequality in Theorem 1.

Actinfo is at the core of the algorithm called AIMH (active information mode hunting). This algorithm is more efficient to find modes in large dimensions than its competitors, as illustrated in [6] (for other models of bump hunting see e.g., [1, 5, 7]). However, other applications are possible; for instance, actinfo is able to compare two different learning strategies. One of these situations can be envisioned when we want to compare an unsupervised learner to a supervised one. Indeed, in this case the unsupervised learner can be taken as the one to which the actinfo is referenced, and the supervised case might be the one for which we are measuring how much information is added in order to reach the target. We can expect big additions of information of supervised learners that perform better than unsupervised ones. In fact, Section 2 of the Supplementary material ensures that we can compare any two strategies.

Bibliography

[1] Jean-Eudes Dazard and J. Sunil Rao. Local Sparse Bump Hunting. *J. Comput. Graph. Stat.*, 19 (4):900–929, 2010. URL https://doi.org/10.1198/jcgs.2010.09029.

[2] William A. Dembski and Robert J. Marks II. Bernoulli's Principle of Insufficient Reason and Conservation of Information in Computer Search. In *Proc. of the 2009 IEEE International Conference on Systems, Man, and Cybernetics. San Antonio, TX*, pages 2647–2652, October 2009. URL https://doi.org/10.1109/ICSMC.2009.5346119.

[3] William A. Dembski and Robert J. Marks II. Conservation of Information in Search: Measuring the Cost of Success. *IEEE Transactions on Systems, Man and Cybernetics A, Systems & Humans*, 5(5):1051–1061, September 2009. URL https://doi.org/10.1109/TSMCA.2009.2025027.

[4] Daniel Andrés Díaz-Pachón and Robert J. Marks II. Generalized active information: Extensions to unbounded domains. *BIO-Complexity*, 2020(3):1–6, 2020. URL https://doi.org/10.5048/BIO-C.2020.3.

[5] Daniel Andrés Díaz-Pachón, Jean-Eudes Dazard, and J. Sunil Rao. Unsupervised Bump Hunting Using Principal Components. In S. Ejaz Ahmed, editor, *Big and Complex Data Analysis: Methodologies and Applications*, pages 325–345. Springer International Publishing, 2017. URL https://doi.org/10.1007/978-3-319-41573-4_16.

[6] Daniel Andrés Díaz-Pachón, Juan P. Sáenz, J. Sunil Rao, and Jean-Eudes Dazard. Mode hunting through active information. *Applied Stochastic Models in Business and Industry*, 35(2):376–393, 2019. URL https://doi.org/10.1002/asmb.2430.

[7] Jerome H. Friedman and Nicholas I. Fisher. Bump hunting in high-dimensional data. *Stat. Comput.*, 9:123–143, 1999. URL https://doi.org/10.1023/A:1008894516817.

[8] S. Kullback. *Information Theory and Statistics*. Wiley, 1959.

[9] G. D. Montañez. A unified model of complex specified information. *BIO-Complexity*, 4:1–26, 2018.

[10] David H. Wolpert and William G. MacReady. No free lunch theorems for search. Technical Report SFI-TR-95-02-010, Santa Fe Institute, 1995.

[11] David H. Wolpert and William G. MacReady. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997.