Paper:

The Search for a Search: Measuring the Information Cost of Higher Level Search

William A. Dembski[*] **and Robert J. Marks II**[**]

[*]Center for Science & Culture, Discovery Institute
Seattle, WA 98104, USA
E-mail: Robert_Marks@baylor.edu
[**]Dept. of Electrical & Computer Engineering, Baylor University
Waco, TX 76798, USA

[Received September 19, 2009; accepted April 1, 2010]

Needle-in-the-haystack problems look for small targets in large spaces. In such cases, blind search stands no hope of success. Conservation of information dictates any search technique will work, on average, as well as blind search. Success requires an assisted search. But whence the assistance required for a search to be successful? To pose the question this way suggests that successful searches do not emerge spontaneously but need themselves to be discovered via a search. The question then naturally arises whether such a higher-level "search for a search" is any easier than the original search. We prove two results: (1) The Horizontal No Free Lunch Theorem, which shows that average relative performance of searches never exceeds unassisted or blind searches, and (2) The Vertical No Free Lunch Theorem, which shows that the difficulty of searching for a successful search increases exponentially with respect to the minimum allowable active information being sought.

Keywords: No Free Lunch Theorems, active information, active entropy, assisted search, endogenous information

1. Introduction

Conservation of information theorems [1–3], especially the No Free Lunch Theorems (NFLT's) [4–8], show that without prior information about a search environment or the target sought, one search strategy is, on average, as good as any other [9]. This is the result of the *Horizontal NFLT* presented in Section 3.2.

A search's difficulty can be measured by its *endogenous information* [1, 10–14] defined as

$$I_\Omega = -\log_2 p \quad \dots\dots\dots\dots\dots \quad (1)$$

where p is the probability of a success from a random query [1]. When there is knowledge about the target location or search space structure, the degree to which the search is improved is determined by the resulting *active information* [1, 10–14]. Even moderately sized searches are virtually certain to fail in the absence of knowledge about the target location or the search space structure. Knowledge concerning membership of the search problem in a structured class [15], for example, can constitute search space structure information [16].

Endogenous and active information together allow for a precise characterization of the conservation of information. The average active information, or *active entropy*, of an unassisted search is zero when no assumption is made concerning the search target. If any assumption is made concerning the target, the active entropy becomes negative. This is the Horizontal NFLT presented in Section 3.2. It states that an arbitrary search space structure will, on average, result in a worse search than assuming nothing and simply performing an unassisted search.

The measure of endogenous and active information can also be applied in a meta sense to a *Search for a Search* (S4S). As one might expect, no active information in the S4S translates to zero active information in the lower-level search (which means that, on average, the lower-level search so found cannot do better than an unassisted search). This result holds for still higher level searches such as a "search for a search for a search" and so on. Thus, without active information introduced somewhere in the search hierarchy, none will be available for the original search. If, on the other hand, active information is introduced anywhere in the hierarchy, it projects onto the original search space as active information.

The target of a S4S is a search algorithm that equals or exceeds a minimally acceptable active information threshold. A higher-level S4S will, itself, have a difficulty as measured by the S4S's endogenous information. How much? Previous results have shown the difficulty of a S4S as measured by its endogenous information is lower bounded by the desired active information of the target search [14]. We establish a much more powerful result. According to the *Vertical NFLT* introduced in Section 4.3, the difficulty of a S4S under loose conditions, as measured by the S4S endogenous information, increases exponentially with respect to the active information threshold required in the lower-level search space.

2. Information in Search

All but the most trivial searches require information about the search environment (e.g., smooth landscapes) or target location (e.g., fitness measures) if they are to be successful. Conservation of information theorems [2–5] show that one search algorithm will, on average, perform as well as any other and thus that no search algorithm will, on average, outperform an unassisted, or blind, search. But clearly, many of the searches that arise in practice do outperform blind unassisted search. How, then, do such searches arise and where do they obtain the information that enables them to be successful?

2.1. Blind and Assisted Queries and Searches

Let $T_1 \in \Omega_1$ denote a target in a search space, Ω_1, so that for uniformity the probability of success, p, is ratio of the cardinality of T_1 to that of Ω_1. The probability, p, is the chance of obtaining an element in the target with a single query assuming a uniform distribution. If nothing is known about the search space or target location, the uniformity of the distribution follows from *Bernoulli's Principle of Insufficient Reason* (PrOIR) [14, 17, 18]. Bernoulli's PrOIR states that in the absence of any prior information, "we must assume that the events ... have equal probability" [17, 18]. Uniformity is equivalent to the assumption that the search space is at maximum informational entropy.

Consider Q queries (samples) from Ω_1 without replacement. Such searches can be construed as a single query when the search space is appropriately defined. For a fixed Q, the augmented search space, Ω_Q, consists of all sequences of length Q chosen from Ω_1 where no element appears more than once. Note that, appropriately, the original search space, Ω_1, is Ω_Q for $Q = 1$. In general

$$|\Omega_Q| = \frac{|\Omega_1|!}{(|\Omega_1| - Q)!}. \qquad (2)$$

The requirement of sampling without replacement requires $Q \leq |\Omega_1|$. If nothing is known about the location or the target, then, from Bernoulli's PrOIR, any element of Ω_Q is as likely to contain the target as any other element.

$$p_Q = \frac{|T_Q|}{|\Omega_Q|} \qquad (3)$$

where $T_Q \in \Omega_Q$ consists of all the elements containing the original target, T_1.

When, for example, T_1 consists of a single element in Ω (i.e., $|T_1| = 1$) and there is no knowledge about the search space structure or target location, then $p_1 = 1/|\Omega_1|$ and

$$p_Q = \frac{Q}{|\Omega_1|} = Qp_1. \qquad (4)$$

From Eqs. (2) and (3), we have

$$\begin{aligned} |T_Q| &= p_Q|\Omega_Q| \\ &= \frac{Q}{|\Omega_1|}|\Omega_Q| \\ &= \frac{Q(|\Omega_1|-1)!}{(|\Omega_1|-Q)!}. \end{aligned} \qquad (5)$$

There appears, at first flush, to be knowledge about the search space Ω_Q for $Q > 1$ since a sequence in Ω_Q has the same chance of containing a target as an element in Ω_Q with a perturbation of the same elements. This, however, is knowledge we cannot exploit with a single query to Ω_Q. Also, if the space is exhaustively contracted to eliminate all elements allowing a perturbation rearrangement to another element, the chance of success in a single query remains the same.

Keeping track of Q subscripts and exhaustively contracted perturbation search spaces is distracting. As such, let Ω denote any search space, like Ω_Q, wherein Bernoulli's PrOIR is applicable for a single query. We adopt similar notation for the target subspace, T, and the probability, p, of finding the target using a single query. A single query to a multi-query space can then be considered a search. If a search is chosen randomly from such a multi-query space, we are still preforming a blind (or unassisted) search.

3. Active Information and No Free Lunch

Define an *assisted query* as any choice from Ω_1 that provides more information about the search environment or candidate solutions than a blind search. Gauging the effectiveness of assisted search in relation to blind search is our next task. Random sampling with respect to the uniform probability **U** sets a baseline for the effectiveness of blind search. For a finite search space with $|\Omega|$ elements, the probability of locating the target T has uniform probability given by Eq. (3) without the subscripts:

$$p = \frac{|T|}{|\Omega|}. \qquad (6)$$

Let q denote the probability of success of an assisted search. We assume that we can always do at least as well as uniform random sampling. The question is, how much better can we do? Given a small target T in Ω and probability measures **U** and φ characterizing blind and assisted search respectively, assisted search will be more effective than blind search when $p < q$, as effective if $p = q$, and less effective if $p > q$. In other words, an assisted search can be more likely to locate T than blind search, equally likely, or less likely. If less likely, the assisted search is counterproductive, actually doing worse than blind search. For instance, we might imagine an Easter egg hunt where one is told "warmer" if one is moving away from an egg and "colder" if one is moving toward it. If one acts on this guidance, one will be less likely to find the egg than if one simply did a blind search. Because the assisted search is in this instance misleading, it actually undermines the success of the search.

An instructive way to characterize the effectiveness of assisted search in relation to blind search is with the likelihood ratio $\frac{q}{p}$. This ratio achieves a maximum of $\frac{1}{p}$ when $q = 1$, in which case the assisted search is guaranteed to succeed. The assisted search is better at locating T than blind search provided the ratio is bigger than 1, equivalent to blind search when it is equal to 1, and worse than blind search when it is less than 1. Finally, this ratio achieves a minimum value of 0 when $q = 0$, indicating that the assisted search is guaranteed to fail.

3.1. Active Information

Let $\mathbf{U}$ denote a uniform distribution on Ω characteristic of an unassisted search and φ the (nonuniform) measure on Ω for an assisted search. Let $\mathbf{U}(\mathbf{T})$ and $\varphi(T)$ denote the probability over the target set $T \in \Omega$. Define the *active information* of the assisted search as

$$\begin{aligned} I_+(\varphi|\mathbf{U}) &:= \log_2 \frac{\varphi(T)}{\mathbf{U}(T)} \quad \ldots\ldots\ldots\ldots \quad (7) \\ &= \log_2 \frac{q}{p} \end{aligned}$$

Active information measures the effectiveness of assisted search in relation to blind search using a conventional information measure. It [1] characterizes the amount of information [19] that φ (representing the assisted search) adds with respect to $\mathbf{U}$ (representing the blind search) in the search for T. Active information therefore quantifies the effectiveness of assisted search against a blind-search baseline. The NFLT dictates that any search without active information will, on average, perform no better than blind search.

The maximum value that active information can attain is $(I_+)_{\max} = -\log_2 p = I_\Omega$ indicating an assisted search guaranteed to succeed (i.e., *a perfect search*) [1]); and the minimum it can attain is $(I_+)_{\min} = -\infty$, indicating an assisted search guaranteed to fail.

Equation (8) can be written as the difference between two positive numbers:

$$\begin{aligned} I_+ &= \log_2 \frac{1}{p} - \log_2 \frac{1}{q} \quad \ldots\ldots\ldots\ldots \quad (8) \\ &= I_\Omega - I_S. \end{aligned}$$

We call the first term the *endogenous information*, I_Ω, given in Eq. (1). Endogenous information represents the fundamental difficulty of the search in the absence of any external information. The endogenous information therefore bounds the active information in Eq. (8).

$$-\infty \leq I_+ \leq I_\Omega.$$

The second term in Eq. (9) is the *exogenous information*.

$$I_S := -\log_2 q \quad \ldots\ldots\ldots\ldots\ldots \quad (9)$$

I_S represents the difficulty that remains once the assisted search is brought to bear. Active information, as the difference between endogenous and exogenous information, thus represents the difficulty inherent in blind search that the assisted search overcomes. Thus, for instance, an assisted search that is no better at locating a target than blind search entails zero active information.

Like other log measurements (e.g., dB), the active information in Eq. (8) is measured with respect to a reference point. Here endogenous information based on blind search serves as the reference point. Active information can also be measured with respect to other reference points. We therefore define the active information more generally as

$$I_+(\varphi|\psi) := \log_2 \frac{\varphi(T)}{\psi(T)} = \log_2 \varphi(T) - \log_2 \psi(T) \quad (10)$$

for φ and ψ arbitrary probability measures over the compact metric space Ω (with metric D), and T an arbitrary Borel set of Ω such that $\psi(T) > 0$.

3.2. Horizontal No Free Lunch

The following No Free Lunch theorem (NFLT) underscores the parity of average search performance.

Theorem 1: **Horizontal No Free Lunch**. Let φ and ψ be arbitrary probability measures and let $\widetilde{T} = \{T_i\}_{i=1}^N$ be an exhaustive partition of Ω all of whose partition elements have positive probability with respect to ψ. Define *active entropy* as the average active information that φ contributes to ψ with respect to the partition $\widetilde{T}$ as

$$H_+^{\widetilde{T}}(\varphi|\psi) := \sum_{i=1}^{N} \psi(T_i) I_+^{T_i}(\varphi|\psi) = \sum_{i=1}^{N} \psi(T_i) \log_2 \frac{\varphi(T_i)}{\psi(T_i)} \quad (11)$$

Then $H_{\widetilde{T}}(\varphi|\psi) \leqslant 0$ with equality if and only if $\varphi(T_i) = \psi(T_i)$ for $1 \leqslant i \leqslant N$.

Proof: The expression in Eq. (11) is immediately recognized as the negative of the *Kullback-Leibler distance* [19]. Since the Kullback-Leibler distance is always nonnegative, the expression in Eq. (11) does not exceed zero. Zero is achieved if for each i, $\varphi(T_i) = \psi(T_i)$.

Because the active entropy is strictly negative, any uninformed assisted search (φ) will on average perform worse than the baseline search. Moreover, the degree to which it performs worse will track the degree to which the assisted search singles out and confers disproportionately high probability on only a few targets in the partition. This suggests that success of an assisted search depends on its attending to a few select targets at the expense of neglecting most of the remaining targets.

Corollary: **Horizontal No Free Lunch – No Information Known**. Given any probability measure on Ω, the active entropy for any partition with respect to a uniform probability baseline will be nonpositive.

Remarks: If no information about a search exists, so that the underlying measure is uniform, then, on average, any other assumed measure will result in negative active information, thereby rendering the search performance worse than random search.

Proof: Using $\psi = \mathbf{U}$ and Eq. (6), the expression Eq. (11) in Theorem 3.2 becomes

$$H_{+}^{\widetilde{T}}(\varphi|\mathbf{U}) = \sum_{i=1}^{N} \mathbf{U}(T_i) I_{+}^{T_i}(\varphi|\mathbf{U}). \quad . \ . \ . \ . \ . \ . \ (12)$$

From Theorem 3.2, the expression in Eq. (12) is nonpositive.

4. The Search for a Good Search

What is the source of active information in a search? Typically, programmers with knowledge about the search (e.g., domain expertise) introduce it. But what if they lack such knowledge? Since active information is indispensable for the success of the search, they will then need to S4S. In this case, a good search is one that generates the active information necessary for success. In this section, we prove a *vertical no free lunch theorem*. It establishes that, under general conditions, the difficulty of the S4S as measured against an endogenous information baseline, increases exponentially with respect to the minimum active information needed for the original search.

4.1. The Probabilistic Hierarchy

Our first task is to extend the definition of active information in Eq. (10) to the probabilistic hierarchy that sits on top of Ω. Doing so will allow us to characterize information-theoretically "the search for a good search," "the search for the search for a good search," etc.

Given Ω, define $\mathbf{M}(\Omega)$ as the collection of all probability measures defined on the Borel sets of Ω. We show in Appendix A that $\mathbf{M}(\Omega)$ is itself a compact metric space with Borel sets, and thus admits the set of all probability measures on it, which we denote by $\mathbf{M}^2(\Omega) = \mathbf{M}(\mathbf{M}(\Omega))$ and which is again a compact metric space, and so on.

Accordingly, given that $\mathbf{M}^0(\Omega) := \Omega, \mathbf{M}^1(\Omega) = \mathbf{M}(\Omega)$, in general $\mathbf{M}^k(\Omega) = \mathbf{M}(\mathbf{M}^{k-1}(\Omega))$ is a compact metric space. $\mathbf{M}^k(\Omega)$ is the k^{th} order probability space on Ω. This is the probabilistic hierarchy. For notational simplicity, we will use the abbreviation $\mathbf{M}^k(\Omega) = \mathbf{M}^k$.

To extend the definition of active information in Eq. (10) up the probabilistic hierarchy, we need to show how higher-order probabilities are canonically transformed into lower-order probabilities. Denote an arbitrary probability measure on $\mathbf{M}^k$ by Θ_{k+1} (which is in $\mathbf{M}^{k+1}$). By vector-valued integration [20–22]

$$\Theta_k = \int_{\mathbf{M}^k} \mu d\Theta_{k+1}(\mu). \quad . \ . \ . \ . \ . \ . \ . \ . \ . \ (13)$$

Active information in Ω can thus be interpreted as being propagated from active information down the probabilistic hierarchy. Because of the compactness of Ω and $\mathbf{M}(\Omega)$, the existence and uniqueness of Eq. (13) is assured. Note that such integrals exist provided that all continuous linear functionals applied to them (which, in this case, amounts to integrating with respect to all bounded continuous real-valued functions on $\mathbf{M}^k$) equals integrating over the continuous linear functions applied inside the integral. Linear functionals thereby reduce vector-valued integration to ordinary integration.

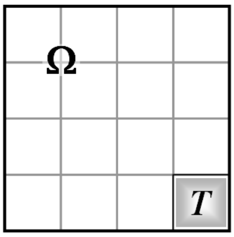


Fig. 1. Search space used in Examples 1 and 2.

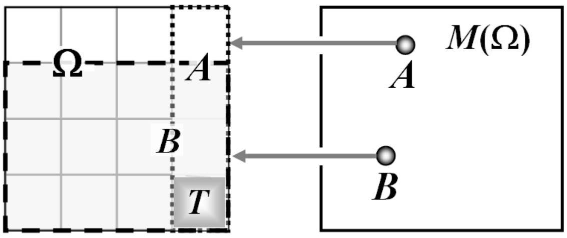


Fig. 2. Propagation of active information. See Example 1.

1. **Propagation of Active Information from $\mathbf{M}(\Omega)$ to Ω. Fig. 1** illustrates the simple problem of searching for one of 16 possible squares in a 4×4 grid ($= \Omega$). The square in question is the target marked T. Here $p = \frac{1}{16}$ and the endogenous information is $I_\Omega = 4$ bits.

   Figure 2 illustrates how active information for searching $\mathbf{M}(\Omega)$ propagates down the probabilistic hierarchy to active information for searching Ω. Consider the following two probability distributions in $\mathbf{M}(\Omega)$:

   - A is the uniform distribution over the rightmost four squares in the search space.
   - B is the uniform distribution over the bottom twelve squares in the search space.

   Suppose we know that one of these distributions characterizes the search for T in Ω, but we know nothing else about the search for T. In that case, Bernoulli's PrOIR would have the search for T characterized by a probability measure Θ_2 ($\in \mathbf{M}^2(\Omega)$) that assigns probability $\frac{1}{2}$ to both A and B. By Eq. (13), Θ_2 induces the following probability of success on Ω:

   $$\begin{aligned} q &= \Pr[T|A]\Pr[A] + \Pr[T|B]\Pr[B] \\ &= \frac{1}{4} \times \frac{1}{2} + \frac{1}{12} \times \frac{1}{2} = \frac{1}{6}. \end{aligned}$$

   The corresponding exogenous information is therefore $I_S = -\log_2 \frac{1}{6} = 2.59$ bits and the active information is $I_+ = 4.00 - 2.59 = 1.41$ bits. Thus, the higher-order probability measure Θ_2 propagated approximately $1\frac{1}{2}$ bits of active information down the probabilistic hierarchy.

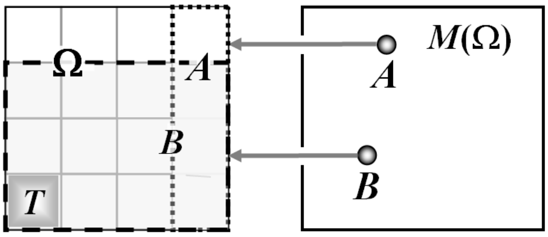


Fig. 3. Propagation of negative active information. See Example 2.

2. **Negative Active Information from $\mathbf{M}(\Omega)$.** If incorrect information assists a search, the active information can be negative. Consider, again, **Fig. 1** and assume, as in the previous example, that one of the distributions A and B in $\mathbf{M}(\Omega)$ characterizes the search for T, though we have no way of preferring one to the other. This time, as shown in **Fig. 3**, assume that the target T, instead of being in the lower right corner, is in the lower left corner. The probability of success is then

$$\begin{aligned} q &= \Pr[T|A]\Pr[A] + \Pr[T|B]\Pr[B] \\ &= 0 + \frac{1}{12} \times \frac{1}{2} = \frac{1}{24} \end{aligned}$$

which corresponds to exogenous information $I_S = 4.59$ and to active information $I_+ = 4.00 - 4.59 = -0.59$ bits.

3. **Propagation of Active Information from $\mathbf{M}^2$.** **Fig. 4** illustrates the propagation of active information from $\mathbf{M}^2(\Omega)$ down the probabilistic hierarchy. One of the distributions in $\mathbf{M}^2(\Omega)$ is the Bernoulli distribution Θ_2 that assigns $\frac{7}{8}$ probability to the distribution A and $\frac{1}{8}$ to B. Over Ω, the distribution A is uniform on the right four squares of Ω and B is uniform on the bottom eight squares. The probability of success is then

$$\begin{aligned} q &= \Pr[T|A]\Pr[A] + \Pr[T|C]\Pr[C] \\ &= \frac{1}{4} \times \frac{7}{8} + \frac{1}{8} \times \frac{1}{8} = \frac{15}{64}. \end{aligned}$$

Therefore, $I_S = -\log \frac{15}{64} = 2.09$ bits corresponding to an active information of $I_+ = 4.00 - 2.09 = 1.91$ bits.

4.2. Conservation of Uniformity

In the absence of active information, uniform probability measures propagate down the probabilistic hierarchy to lower-level uniform probability measures. In other words, higher-order uniform probability measures never induce anything other than uniform probability measures at lower levels in the probabilistic hierarchy.

Theorem 2: **Conservation of Uniformity**. Let $\mathbf{U}_1 := \mathbf{U}$ be the uniform probability on a compact metric space Ω with metric D [23], and let $\mathbf{U}_2$ be the uniform probability on the compact metric space $\mathbf{M}$ with canonical (Kantorovich-Wasserstein) metric [24]. Then

$$\mathbf{U}_1 = \int_{\mathbf{M}^1} \mu d\mathbf{U}_2(\mu), \quad \ldots\ldots\ldots\ldots \quad (14)$$

and, more generally,

$$\mathbf{U}_k = \int_{\mathbf{M}^k} \mu d\mathbf{U}_{k+1}(\mu) \quad \ldots\ldots\ldots\ldots \quad (15)$$

where $\mathbf{U}_{k+1}$ is the uniform distribution on $\mathbf{M}^k$.

The proof is in Appendix B.

Remarks: Because of the compactness of Ω and $\mathbf{M}^1 = \mathbf{M}$, the integral in Eq. (14) exists and is uniquely determined. This theorem shows that averaging the probability measures of $\mathbf{M}$ with respect to the uniform probability $\mathbf{U}_2$ ($\in \mathbf{M}^2$) is just the uniform probability $\mathbf{U}_1$ ($\in \mathbf{M}^1$). Thus, if we think of each μ under the integral in Eq. (14) as a probability measure representing an assisted search, this theorem says that the average performance of all assisted searches is equivalent to uniform random sampling.

4.3. The Displacement Principle: The Vertical No Free Lunch Theorem

The *displacement principle* states that the search for a good search is at least as difficult as a given search. We prove that in the search for a good search, the endogenous information for the higher-level search grows exponentially with respect to the active information needed to successfully search for the original target. Since endogenous information gauges the inherent difficulty of a search, this shows that the difficulty of the search for a good search grows exponentially with the difficulty of the original search.

More precisely,[1] the original search seeks a target $T = T_1$ in $\Omega = \Omega_1 = \mathbf{M}^0$. In the search for the search, we have a target T_2 of searches that equal or exceed a given performance level among the set of measures $\Omega_2 = \mathbf{M}^1$. (We will henceforth interchangeably use the notation Ω_{k+1} and $\mathbf{M}^k$.)

Consider, then, a minimally acceptable active information $\breve{I}_+$. Over the set of all measures $\mathbf{M}^1$, the target set of searches is then

$$T_2 = \left\{ \varphi \in \Omega_2 \;\middle|\; I_+(\varphi|\mathbf{U}) \geq \breve{I}_+ \right\}. \quad \ldots\ldots \quad (16)$$

Theorem 3: **The Strict Vertical No Free Lunch Theorem.** Let $\breve{I}_+ = \log(\hat{q}/p)$ be the minimally acceptable active information[2] of a search such that

$$p \ll \hat{q} \ll 1 \quad \ldots\ldots\ldots\ldots\ldots \quad (17)$$

and

$$p = \frac{1}{K} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (18)$$

where we have adopted the notation $K := |\Omega|$. Given that $M^1 = \Omega_2$ and that information is measured in nats (as op-

[1] The subscripts here refer to the level of the search. In Section 2.1, the subscripts refer to multiple queries.

[2] A breve as is used on $\breve{I}_+$ denotes a lower bound while a cap, as is used on $\hat{q}$, denotes an upper bound.

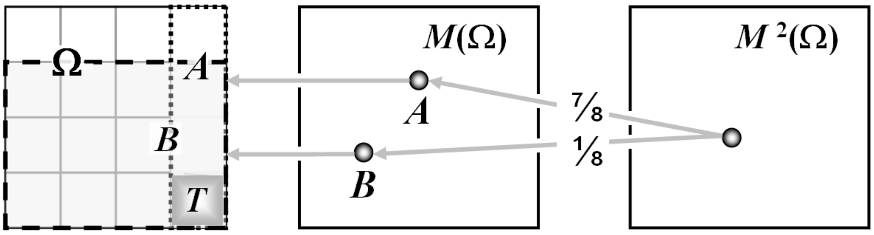


Fig. 4. Propagation of active information through two levels of the probability hierarchy. See Example 3.

posed to bits),[3] the endogenous information of a search for a search that achieves at least an active information $\check{I}_+$ is exponential with respect to $\check{I}_+$:

$$I_{\Omega_2} \simeq e^{\check{I}_+}. \qquad (19)$$

The proof is supplied in Appendix C.

Theorem 4: **The General Vertical No Free Lunch Theorem.** Let

$$p \ll \hat{q} < 1 \qquad (20)$$

and

$$pK \geq 1. \qquad (21)$$

Then the difficulty of the search for the search as measured by the endogenous information, I_{Ω_2}, is bounded by

$$I_{\Omega_2} \geq \check{I}_{\Omega_2} \qquad (22)$$

where

$$\check{I}_{\Omega_2} = \frac{I_\Omega - \log(K)}{2} - I_+ + KD(p\|\hat{q}). \qquad (23)$$

Here $D(p\|\hat{q})$ is the Kullback-Leibler distance:

$$D(p\|\hat{q}) = (1-p)\ln\left(\frac{1-p}{1-\hat{q}}\right) + p\ln\left(\frac{p}{\hat{q}}\right) \qquad (24)$$

The proof of Theorem 4.3 is given in Appendix D. Although the strict vertical NFLT in Eq. (18) and the general vertical NFLT in Eq. (21) make different assumptions, the following result shows that Theorem 4.3 is a special case of Theorem 4.3.

Theorem 5: **Strict Case Subsumed in General Case.** If Eqs. (17) and (18) are in force, then Eq. (23) becomes $\check{I}_{\Omega_2} \to e^{I_+}$ which is consistent with Eq. (19).

The proof is given in Appendix E.

5. Conclusion

The Horizontal NFLT illustrates the law of conservation of information by revealing that unsubstantiated arbitrary assumptions about a search will, on average, result in a search with less than average performance as measured by the search's active information. This results from the average active information, e.g., the active entropy, being always negative.

The Vertical NFLT establishes the troubling property that, under a loose set of conditions, the difficulty of a Search for a Search (S4S) increases exponentially as a function of minimal acceptable active information being sought.

Acknowledgements

Seminars organized by Claus Emmeche and Jakob Wolf at the Niels Bohr Institute in Copenhagen (May 13, 2004) and by Baylor University's Department of Electrical and Computer Engineering (January 27, 2005) were the initial impetus for this paper. Thanks are due to them and also to C. J., who provided crucial help with the combinatorics required to prove the Vertical No Free Lunch Theorem. The authors also express their appreciation to Dietmar Eben for comments made on an earlier draft of the manuscript.

References:

[1] W. A. Dembski and R. J. Marks II, "Conservation of Information in Search: Measuring the Cost of Success," IEEE Trans. on Systems, Man and Cybernetics A, Systems and Humans, Vol.39, No.5, pp. 1051-1061, September 2009.

[2] C. Schaffer, "A conservation law for generalization performance," Proc. Eleventh Int. Conf. on Machine Learning, H. Willian and W. Cohen, San Francisco: Morgan Kaufmann, pp. 295-265, 1994.

[3] T. M. English, "Some information theoretic results on evolutionary optimization," Proc. of the 1999 Congress on Evolutionary Computation, 1999, CEC 99, Vol.1, pp. 6-9, July 1999.

[4] D. Wolpert and W. G. Macready, "No free lunch theorems for optimization," IEEE Trans. Evolutionary Computation, Vol.1, No.1, pp. 67-82, 1997.

[5] M. Koppen, D. H. Wolpert, and W. G. Macready, "Remarks on a recent paper on the 'no free lunch' theorems," IEEE Trans. on Evolutionary Computation, Vol.5 , Issue 3, pp. 295-296, June 2001.

[6] Y.-C. Ho and D. L. Pepyne, "Simple explanantion of the No Free Lunch Theorem," Proc. of the 40th IEEE Conf. on Decision and Control, Orlando, Florida, 2001.

[7] Y.-C. Ho, Q.-C. Zhao, and D. L. Pepyne, "The No Free Lunch Theorems: Complexity and Security," IEEE Trans. on Automatic Control, Vol.48, No.5, pp. 783-793, May 2003.

[8] W. A. Dembski, "No Free Lunch: Why Specified Complexity Cannot Be Purchased without Intelligence," Lanham, Md.: Rowman and Littlefield, 2002.

[9] J. C. Culberson, "On the Futility of Blind Search: An Algorithmic View of 'No Free Lunch'," Evolutionary Computation, Vol.6, No.2, pp. 109-127, 1998.

[10] W. A. Dembski and R. J. Marks II, "Conservation of Information in Search: Measuring the Cost of Success," IEEE Trans. on Systems, Man and Cybernetics A, Systems and Humans, Vol.39, No.5, pp. 1051-1061, September 2009.

[11] W. Ewert, W. A. Dembski, and R. J. Marks II, "Evolutionary Synthesis of Nand Logic: Dissecting a Digital Organism," Proc. of the 2009 IEEE Int. Conf. on Systems, Man, and Cybernetics. San Antonio, TX, USA, pp. 3047-3053, October 2009.

[12] W. Ewert, G. Montaez, W. A. Dembski, and R. J. Marks II, "Efficient Per Query Information Extraction from a Hamming Oracle," Proc. of the the 42nd Meeting of the Southeastern Symposium on System Theory, IEEE, University of Texas at Tyler, pp. 290-297, March 7-9, 2010.

[3] For a probability, π, information in bits is $b = -\log_2 \pi$ and, in nats, is $\eta = -\ln \pi$ where ln denotes the natural logarithm [19]. One nat equals $\ln 2 = 0.639$ bits.

[13] W. A. Dembski and R. J. Marks II, "Life's Conservation Law: Why Darwinian Evolution Cannot Create Biological Information," in Bruce Gordon and William Dembski, editors, The Nature of Nature, Wilmington, Del.: ISI Books, 2010.

[14] W. A. Dembski and R. J. Marks II, "Bernoulli's Principle of Insufficient Reason and Conservation of Information in Computer Search," Proceedings of the 2009 IEEE Int. Conf. on Systems, Man, and Cybernetics, San Antonio, TX, USA, pp. 2647-2652, October 2009.

[15] M. Hutter, "A Complete Theory of Everything (will be subjective)," (in review) Arxiv preprint arXiv:0912.5434, 2009, arxiv.org, Dec. 20, 2009.

[16] B. Weinberg and E. G. Talbi, "NFL theorem is unusable on structured classes of problems," Congress on Evolutionary Computation, CEC2004, Vol.1, 19-23, pp. 220-226, June 2004.

[17] J. Bernoulli, "Ars Conjectandi (The Art of Conjecturing)," Tractatus De Seriebus Infinitis, 1713.

[18] A. Papoulis, "Probability, Random Variables, and Stochastic Processes," 3rd ed., New York: McGraw-Hill, 1991.

[19] T. M. Cover and J. A. Thomas, "Elements of Information Theory," 2nd Edition, Wiley-Interscience, 2006.

[20] N. Dinculeanu, "Vector Integration and Stochastic Integration in Banach Spaces," New York: Wiley, 2000.

[21] I. M. Gelfand, "Sur un Lemme de la Theorie des Espaces Lineaires," Comm. Inst. Sci. Math. de Kharko., Vol.13, No.4, pp. 35-40, 1936.

[22] B. J. Pettis, "On Integration in Vector Spaces," Trans. of the American Mathematical Society, Vol.44, pp. 277-304, 1938.

[23] W. A. Dembski, "Uniform Probability," J. of Theoretical Probability, Vol.3, No.4, pp. 611-626, 1990.

[24] F. Giannessi and A. Maugeri, "Variational Analysis and Applications (Nonconvex Optimization and Its Applications)," Springer, 2005.

[25] D. L. Cohn, "Measure Theory," Boston: Birkhäuser, 1996.

[26] R. M. Dudley, "Probability and Metrics," Aarhus: Aarhus University Press, 1976.

[27] P. de la Harpe, "Topics in Geometric Group Theory," University of Chicago Press, 2000.

[28] P. Billingsley, "Convergence of Probability Measures," 2nd ed., New York: Wiley, 1999.

[29] W. Feller, "An Introduction to Probability Theory and Its Applications," 3rd ed., Vol.1, New York: Wiley, 1968.

[30] R. J. Marks II, "Handbook of Fourier Analysis and Its Applications," Oxford University Press, 2009.

[31] M. Spivak, "Calculus," 2nd ed., Berkeley, Calif.: Publish or Perish, 1980.

Appendix A. The Probabilistic Hierarchy

Geometric and measure-theoretic structures on Ω extend straightforwardly and canonically to corresponding structures on $\mathbf{M}(\Omega)$, the collection of all probability measures on the Borel sets of Ω. And from there they extend to $\mathbf{M}^2(\Omega) = \mathbf{M}(\mathbf{M}(\Omega))$, the collection of all probability measures on the Borel sets of $\mathbf{M}(\Omega)$, and so on up the probabilistic hierarchy, which is defined inductively as $\mathbf{M}^k(\Omega) = \mathbf{M}(\mathbf{M}^{k-1}(\Omega))$.

To see how structures on Ω extend up the probabilistic hierarchy, begin with the metric D on Ω. Because D makes Ω a compact metric space, D *a fortiori* makes Ω a complete separable metric space. Separable topological spaces that can be metrized with a complete metric are known as *Polish* spaces ([25], ch. 8).

$\mathbf{M}(\Omega)$ is itself a separable metric space in the Kantorovich-Wasserstein metric D_1 which induces the weak topology on $\mathbf{M}(\Omega)$. For Borel probability measures μ and ν on Ω, this metric is defined as

$$D_1(\mu,\nu) = \inf\left\{\int D(x,y)\zeta(dx,dy);\zeta \in \mathbf{P}_2(\mu,\nu)\right\}$$

$$= \sup\left\{\left|\int f(x)\mu(dx) - \int f(x)\nu(dx)\right|;\|f\|_L \le 1\right\}$$

In the first equation, $\mathbf{P}_2(\mu,\nu)$ is the collection of all Borel probability measures on $\Omega \times \Omega$ with marginal distributions μ on the first factor and ν on the second. In the second equation here, f ranges over all continuous real-valued functions on Ω for which the Lipschitz seminorm does not exceed one. The Lipschitz seminorm is defined as follows:

$$\|f\|_L = \sup\left\{\frac{|f(x)-f(y)|}{D(x,y)};x,y \in \Omega,\ x \neq y\right\}$$

Both the infimum and the supremum in these equations define metrics. The first is the Wasserstein metric, the second the Kantorovich metric. Though the two expressions appear different, the quantities they represent are known to be identical [26].

The Kantorovich-Wasserstein metric D_1 is the canonical extension to $\mathbf{M}(\Omega)$ of the metric D on Ω. It extends the metric structure of Ω as faithfully as possible to $\mathbf{M}(\Omega)$. For instance, if δ_x and δ_y are point masses in $\mathbf{M}(\Omega)$, then $D_1(\delta_x,\delta_y) = D(x,y)$. It follows that the canonical embedding of Ω into $\mathbf{M}(\Omega)$, i.e., $x \mapsto \delta_x$, is in fact an isometry.

To see that D_1 canonically extends the metric structure of Ω to $\mathbf{M}(\Omega)$, consider the following reformulation of this metric. Let

$$\mathbf{M}_{av}(\Omega) = \left\{\frac{1}{n}\sum_{1\le i\le n}\delta_{x_i} : x_i \in \Omega\right\}$$

where n ranges over all positive integers. It is readily seen that $\mathbf{M}_{av}(\Omega)$ is dense in $\mathbf{M}(\Omega)$ in the weak topology. Note that the x_is are not required to be distinct, implying that $\mathbf{M}_{av}(\Omega)$ consists of all convex linear combinations of point masses with rational weights; note also that such combinations, when restricted to a countable dense subset of Ω, form a countable dense subset of $\mathbf{M}(\Omega)$ in the weak topology, showing that $\mathbf{M}(\Omega)$ is itself separable in the weak topology.

For any measures μ and ν in $\mathbf{M}_{av}(\Omega)$, it is possible to find a positive integer n such that

$$\mu = \frac{1}{n}\sum_{1\le i\le n}\delta_{x_i}$$

and

$$\nu = \frac{1}{n}\sum_{1\le i\le n}\delta_{y_i}.$$

Next, define

$$D_1^{perm}\left(\frac{1}{n}\sum_{1\le i\le n}\delta_{x_i},\frac{1}{n}\sum_{1\le i\le n}\delta_{y_i}\right)$$

$$:= \min\left\{\frac{1}{n}\sum_{1\le i\le n}D(x_i,y_{\sigma i}) : \sigma \in \mathbf{S}_n\right\}$$

where $\mathbf{S}_n$ is the symmetric group on the numbers 1 to n. D_1^{perm} looks for the best way to match up point masses for any pair of measures in $\mathbf{M}_{av}(\Omega)$ *vis-a-vis* the metric D.

It is straightforward to show that D_1^{perm} is well-defined and constitutes a metric on $\mathbf{M}_{av}(\Omega)$. The only point in need of proof here is whether for arbitrary measures $\frac{1}{n}\sum_{1\leq i\leq n}\delta_{x_i}$ and $\frac{1}{n}\sum_{1\leq i\leq n}\delta_{y_i}$ in $\mathbf{M}_{av}(\Omega)$, and for any measures

$$\frac{1}{mn}\sum_{1\leq i\leq mn}\delta_{z_i}=\frac{1}{n}\sum_{1\leq i\leq n}\delta_{x_i}$$

and

$$\frac{1}{mn}\sum_{1\leq i\leq mn}\delta_{w_i}=\frac{1}{n}\sum_{1\leq i\leq n}\delta_{y_i},$$

we have

$$\min\left\{\frac{1}{n}\sum_{1\leq i\leq n}D(x_i,y_{\sigma i}):\sigma\in\mathbf{S}_n\right\}$$
$$=\min\left\{\frac{1}{mn}\sum_{1\leq i\leq mn}D(z_i,w_{\rho i}):\rho\in\mathbf{S}_{mn}\right\}.$$

The proof follows from Hall's *marriage lemma* [27]. Given this equality, it follows that $D_1^{perm}=D_1$ on $\mathbf{M}_{av}(\Omega)$ and, because $\mathbf{M}_{av}(\Omega)$ is dense in $\mathbf{M}(\Omega)$, that D_1^{perm} extends uniquely to D_1 on all of $\mathbf{M}(\Omega)$.

Thus, given that D metrizes Ω, D_1 is the canonical metric that metrizes $\mathbf{M}(\Omega)$. And, just as D induces a compact topology on Ω, so does D_1 induce a compact topology on $\mathbf{M}(\Omega)$. This last result is a direct consequence of D_1 inducing the weak topology on $\mathbf{M}(\Omega)$ and of Prohorov's theorem, which ensures that $\mathbf{M}(\Omega)$ is compact in the weak topology provided that Ω is compact ([28]: 59).

Given that D_1 makes $\mathbf{M}(\Omega)$ a compact metric space, the next question is whether this metric induces a uniform probability $\mathbf{U}_2$ on $\mathbf{M}(\Omega)$ (such a $\mathbf{U}_2$ would reside in $\mathbf{M}^2(\Omega)$; $\mathbf{U}_1=\mathbf{U}$ is the original uniform probability on Ω). As it is, $\mathbf{M}(\Omega)$ is uniformizable with respect to D_1 and therefore has such a uniform probability. To see this, note that if

$$\mathbf{U}^\varepsilon=\frac{1}{n}\sum_{1\leq i\leq n}\delta_{x_i}$$

denotes a finitely supported uniform probability that is based on an ε-lattice $\{x_1,x_2,\ldots,x_n\}\subset\Omega$, then $\mathbf{U}^\varepsilon$ approximates $\mathbf{U}$ to within ε, i.e., $D_1(\mathbf{U}^\varepsilon,\mathbf{U})\leqslant\varepsilon$. Given that $\binom{2n-1}{n}$ is the number of ways of filling n slots with n identical items ([29], ch. 1), it follows that for

$$n^*=\binom{2n-1}{n}=\frac{(2n-1)!}{n!\,(n-1)!},$$
$$\{\mu_1,\mu_2,\ldots,\mu_{n^*}\}\subset\mathbf{M}(\Omega)$$

is an $\frac{\varepsilon}{n}$-lattice as the μ_ks run through all finitely supported probability measures $\frac{1}{n}\sum_{1\leq i\leq n}\delta_{w_i}$ where the w_i's are drawn from $\{x_1,x_2,\ldots,x_n\}$ allowing repetitions.

It follows that as $\mathbf{U}^\varepsilon$ converges to $\mathbf{U}$ in the weak topology on $\mathbf{M}(\Omega)$, the sample distribution

$$\mathbf{U}_2^\varepsilon=\frac{1}{n^*}\sum_{1\leq k\leq n^*}\delta_{\mu_k}$$

converges to the uniform probability $\mathbf{U}_2$ in the weak topology on $\mathbf{M}^2(\Omega)$. Moreover, for D_2, which is the Kantorovich-Wasserstein metric on $\mathbf{M}^2(\Omega)$ ($\mathbf{M}^2(\Omega)$ is likewise a compact metric space with respect to D_2), $D_2(\mathbf{U}_1^\varepsilon,\mathbf{U}_1)\leqslant\frac{\varepsilon}{n}$. The details for proving these claims derive from elementary combinatorics and from unpacking the definition of uniform probability [23].

In summary, given the compact metric space Ω with metric D and uniform probability $\mathbf{U}$ induced by D, both D and $\mathbf{U}$ extend canonically up the probabilistic hierarchy: $D:=D_0$ on $\mathbf{M}^0(\Omega):=\Omega$, D_1 on $\mathbf{M}^1(\Omega)$)), D_2 on $\mathbf{M}^2(\Omega)$, etc. Moreover, $\mathbf{U}_1\in\mathbf{M}^1(\Omega)$, $\mathbf{U}_2\in\mathbf{M}^2(\Omega)$, etc.

Appendix B. Proof of the Conservation of Uniformity

We prove Theorem 4.2 for Ω finite (the infinite case is proved by considering finitely supported uniform probabilities on ε-lattices, and then taking the limit as the ε-mesh of these lattices goes to zero [23]).

Proof: Let $\Omega=\{x_1,x_2,\ldots,x_n\}$. For large N, consider all probabilities of the form

$$\mu=\sum_{1\leq i\leq n}\frac{N_i}{N}\delta_{x_i}$$

such that the N_is are nonnegative integers that sum to N. From elementary combinatorics, there are $N^*=\binom{N+n-1}{n-1}$ distinct probabilities like this [29]. Therefore, define

$$\mathbf{U}_2[N]:=\frac{1}{N^*}\sum_{1\leq k\leq N^*}\delta_{\mu_k}$$

so that the μ_k's run through all such μ. From Appendix A it follows that $\mathbf{U}_2[N]$ converges in the weak topology to $\mathbf{U}_2$ as $N\uparrow\infty$.

It's enough, therefore, to show that

$$\int_{\mathbf{M}(\Omega)}\mu d\mathbf{U}_2[N]=\frac{1}{N^*}\sum_{1\leq k\leq N^*}\mu_k$$

is the uniform probability on Ω. And for this, it is enough to show that for x_i in Ω,

$$\frac{1}{N^*}\sum_{1\leq k\leq N^*}\mu_k(\{x_i\})=\frac{1}{n}.$$

For definiteness, let us consider x_1. We can think of x_1 as occupied with weights that run through all the multiples of $1/N$ ranging from 0 to N. Hence, for fixed integer M ($0\leqslant M\leqslant N$), the contribution of the μ_ks with weight M/N at x_1 is

$$M\cdot\binom{N-M+n-2}{n-2}.$$

Note that the term $\binom{N-M+n-2}{n-2}$ is the number of ways of occupying $n-1$ slots with $N-M$ identical items [29].

Accordingly, the total weight that the μ_ks assign to x_1, when normalized by $1/N^*$, is

$$\frac{1}{N^*}\sum_{1\le k\le N^*}\mu_k(\{x_1\}) = \frac{1}{N^*}\sum_{0\le M\le N} M\cdot\binom{N-M+n-2}{n-2}.$$

Because this expression is independent of x_1 and is also the probability of x_1, it follows that the probability of all x_is under $\frac{1}{N^*}\sum_{1\le k\le N^*}\mu_k$ is the same. Hence for each x_i in Ω,

$$\mathbf{U}(\{x_i\}) = \frac{1}{N^*}\sum_{1\le k\le N^*}\mu_k(\{x_i\}) = \frac{1}{n}.$$

This is what needed to be proved.

Appendix C. Proof of the Strict Vertical No Free Lunch Theorem

We offer two derivations of Theorem 4.3.

C.1. Combinatoric Approach.

We start with the finite case. First, we assume

$$\Omega = \{x_1, x_2, \ldots, x_K\},$$
$$T = \{x_1, x_2, \ldots, x_{|T|}\},$$

where, from Eq. (6), we use $|T| = pK$. (For the proof of Theorem 4.3, we use Eq. (18).) For a given N, consider all probabilities

$$\mu = \frac{1}{N}\sum_{1\le i\le N}\delta_{w_i}$$

where the w_i's are drawn from Ω allowing repetitions. By elementary combinatorics there are

$$N^* = \binom{N+K-1}{K-1} = \frac{\Gamma(N+K)}{\Gamma(N+1)\Gamma(K)}$$

such probabilities.[4]

Accordingly, define

$$\mathbf{U}_2[N] := \frac{1}{N^*}\sum_{1\le i\le N^*}\delta_{\mu_i}$$

where the μ_is range over these μs. $\mathbf{U}_2[N]$ then converges weakly to $\mathbf{U}_2$.

Next, consider the probabilities

$$\mu = \frac{1}{N}\sum_{1\le i\le N}\delta_{w_i}$$

where the w_is are drawn from Ω, allowing repetitions, but also for which the number of w_is among T is at least $\lfloor N q\rfloor$ If we let Q_N denote the set of all such finitely supported probabilities that assign probability at least $\hat{q}$ to T, and if we let

$$q(N) = N - \lfloor N\hat{q}\rfloor \xrightarrow[N\to\infty]{} (1-\hat{q})N,$$

then by elementary combinatorics Q_N has the following cardinality.

$$|Q_N| = \sum_{j=0}^{q(N)}\binom{N-j+pK-1}{pK-1}\binom{j+(1-p)K-1}{(1-p)K-1}. \tag{25}$$

Because we assume Eq. (18), this gives

$$|Q_N| = \sum_{j=0}^{q(N)}\binom{j+K-2}{K-2}.$$

It follows that

$$p_2 := \mathbf{U}_2(T_2) = \lim_{N\to\infty}\frac{|Q_N|}{N^*}$$

$$= \lim_{N\to\infty}\frac{\sum_{j=0}^{q(N)}\binom{j+K-2}{K-2}}{\binom{N+K-1}{K-1}}. \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \tag{26}$$

This last limit simplifies. Note first that

$$\binom{N+K-1}{K-1} = \frac{N^{K-1} + (\text{lower order terms in } N)}{\Gamma(K)}.$$

Thus

$$\binom{N+K-1}{K-1} \xrightarrow[N\to\infty]{} \frac{N^{K-1}}{\Gamma(K)}. \quad . \; . \; . \; . \; . \; . \; . \; . \tag{27}$$

Similarly, note that

$$\binom{j+K-2}{K-2} = \frac{j^{K-2} + (\text{lower order terms in } j)}{\Gamma(K+1)}.$$

Since

$$\sum_{j=0}^{M} j^n = \frac{M^{n+1}}{n+1} + (\text{ lower order terms in } M), \quad . \tag{28}$$

we can write

$$\sum_{j=0}^{q[n]}\binom{j+K-2}{K-2} \xrightarrow[N\to\infty]{} \sum_{j=0}^{(1-\hat{q})N}\frac{j^{K-2}}{\Gamma(K+1)}$$
$$\xrightarrow[N\to\infty]{} \frac{((1-\hat{q})N)^{K-1}}{\Gamma(K)}. \quad . \tag{29}$$

Substituting Eqs. (17) and (18) into Eq. (26) gives

$$p_2 \xrightarrow[N\to\infty]{} (1-\hat{q})^{K-1}. \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \tag{30}$$

The endogenous information for the search for the search

$$I_{\Omega_2} = -\ln p_2 \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \tag{31}$$

is then

$$I_{\Omega_2} = -(K-1)\ln(1-\hat{q}). \quad . \; . \; . \; . \; . \; . \; . \tag{32}$$

[4] $\Gamma(\cdot)$ is the *gamma function*. For M a nonnegative integer, $\Gamma(M+1) = M!$. We use $\Gamma(\cdot)$ rather than the factorial because we do not distinguish between integer and real arguments in our treatment. More rigourously, the *beta function* [30],

$$\beta(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$$

takes on the role of the binomial coefficient for noninteger arguments.

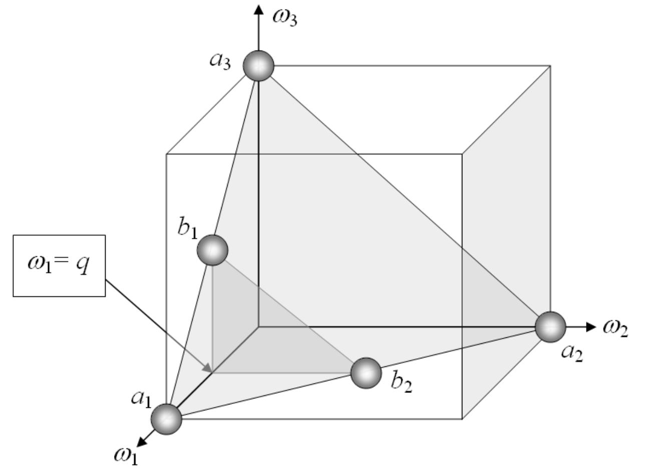


Fig. 5. A three dimensional simplex in $\{\omega_1, \omega_2, \omega_3\}$. The numerical values of a_1, a_2 and a_3 are one.

Assuming Eq. (17) requires that $\ln(1-\hat{q}) \xrightarrow[\hat{q}\to 0]{} -\hat{q}$ and, using Eq. (18)

$$K-1 \xrightarrow[p\to 0]{} \frac{1}{p}.$$

Therefore Eq. (32) becomes

$$I_{\Omega_2} \simeq \frac{\hat{q}}{p} = e^{I_+}.$$

C.2. Geometric Approach

Define the set of all possible probability density functions on Ω as

$$\mathbf{pdf}(\Omega) = \left\{ f(n) \,\middle|\, 1 \le n \le K, \sum_{n=1}^{K} f(n) = 1 \right\}.$$

The set $\mathbf{pdf}(\Omega)$ lies on a simplex. To illustrate, the simplex equilateral triangle (a_1, a_2, a_3) is shown in **Fig. 5** for $K = 3$ in the coordinate system $\{\omega_1, \omega_2, \omega_3\}$. We choose a point on the simplex at random assuming a uniform distribution. Then

$$\Pr[\hat{q} \ge q] = \Pr[f(0) \ge q].$$

This is equivalent in **Fig. 5** to choosing a point on the triangle (a_1, b_1, b_2). The ratio the area of this smaller triangle to the area of the simplex triangle is the probability a successful pdf will be chosen. The two triangles are congruent. In higher dimensions, the success area will be congruent with the simplex. The ratio scales with the dimension, so

$$p_2 = \Pr[q \ge \hat{q}] = (1-\hat{q})^{K-1}. \quad \ldots\ldots\ (33)$$

This is equivalent to Eq. (30), from which the rest of the derivation follows.

Appendix D. Proof of the General Vertical No Free Lunch Theorem

Proof of Theorem 4.3.

The preamble to this proof is identical to the development in Appendix C.1 up to Eq. (25), from which point we continue. For $pK \ge 1$, it follows that

$$p_2 := \mathbf{U}_2(T_2) = \lim_{N\to\infty} \frac{|Q_N|}{N^*}$$

$$= \lim_{N\to\infty} \frac{\sum_{j=0}^{q(N)} \binom{N-j+pK-1}{pK-1}\binom{j+(1-p)K-1}{(1-p)K-1}}{\binom{N+K-1}{K-1}}. \quad \ldots\ (34)$$

Similarly, note that

$$\binom{N-j+pK-1}{pK-1}$$

$$= \frac{(N-j)^{pK-1} + (\text{lower order terms in } N-j)}{\Gamma(pK)}$$

and that

$$\binom{j+(1-p)K-1}{(1-p)K-1}$$

$$= \frac{j^{((1-p)K)-1} + (\text{lower order terms in } j)}{\Gamma((1-p)K)}$$

so that

$$\binom{N-j+pK-1}{pK-1} \xrightarrow[N\to\infty]{} \frac{(N-j)^{pK-1}}{\Gamma(pK)}$$

and

$$\binom{j+(1-p)K-1}{(1-p)K-1} \xrightarrow[N\to\infty]{} \frac{j^{(1-p)K-1}}{\Gamma((1-p)K)}.$$

Using this and the denominator simplification in Eq. (27), we find that Eq. (34) becomes

$$p_2 \xrightarrow[N\to\infty]{} \binom{K}{pK} \sum_{j=0}^{(1-\hat{q})N} \frac{1}{N}\left(1-\frac{j}{N}\right)^{pK-1}\left(\frac{j}{N}\right)^{\Gamma((1-p)K)}$$

This last limit becomes the cumulative beta distribution [30]:

$$p_2 = \binom{K}{pK} \int_0^{1-\hat{q}} t^{(1-p)K-1}(1-t)^{pK-1} dt. \quad .\ (35)$$

We now apply Stirling's exact formula to calculate the factor in front of the integral in Eq. (35). According to Stirling's exact formula, for every positive integer n,

$$\sqrt{2\pi} n^{n+1/2} e^{-n} < n! < \sqrt{2\pi} n^{n+1/2} e^{-n+1/(12n)},$$

which implies that there is a function $\varepsilon(n)$ satisfying $0 < \varepsilon(n) < 1$ such that [31]

$$n! = \sqrt{2\pi} n^{n+1/2} e^{-n+\varepsilon(n)/(12n)}.$$

Since $\Gamma(n+1) = n!$, it now follows for all K that

$$\begin{aligned}
\binom{K}{pK} &= \frac{\Gamma(K)}{\Gamma(pK)\Gamma((1-p)K)} \\
&= \frac{\Gamma(K+1)}{\Gamma((1-p)K)\Gamma(pK+1)} \cdot \frac{(1-p)pK^2}{K} \\
&= p(1-p)K \\
&\quad \times \frac{\sqrt{2\pi}\, K^{K+1/2} e^{-K+\varepsilon(K)/(12K)}}{\sqrt{2\pi}(pK)^{pK+1/2} e^{-pK+\varepsilon(pK)/(12pK)}} \\
&\quad \times \frac{e^{(1-p)K-\varepsilon((1-p)K)/(12(1-p)K)}}{\sqrt{2\pi}\,((1-p)K)^{(1-p)K+1/2}} \\
&= \sqrt{\frac{p(1-p)K}{2\pi}} \cdot \left((1-p)^{(1-p)} p^p\right)^{-K} \\
&\quad \times \exp\left(\frac{\varepsilon(K)}{12K} - \frac{\varepsilon(pK)}{12pK} - \frac{\varepsilon((1-p)K)}{12(1-p)K}\right) \\
&\leq \sqrt{\frac{p(1-p)K}{2\pi}} \cdot e^{\frac{1}{12}} \cdot \left((1-p)^{(1-p)} p^p\right)^{-K} \\
&< \sqrt{p(1-p)K} \cdot \left((1-p)^{(1-p)} p^p\right)^{-K}.
\end{aligned}$$

Moreover, the integral in Eq. (35) can be bounded for large K.

$$\begin{aligned}
&\int_0^{1-\hat{q}} t^{(1-p)K-1}(1-t)^{pK-1}dt \\
&\leq (1-\hat{q})(1-\hat{q})^{(1-p)K-1}\hat{q}^{pK-1} \\
&= (1-\hat{q})^{(1-p)K}\hat{q}^{pK-1}
\end{aligned}$$

How large does K have to be for this last inequality to hold? Consider the function $t^m(1-t)^n$. For $t=0$ as well as for $t=1$, this function is 0. Elsewhere on the unit interval it is strictly positive. From 0 onwards, the function is therefore monotonically increasing to a certain point. Up to what point? To the point where the derivative of $t^m(1-t)^n$, namely $mt^{m-1}(1-t)^n - nt^m(1-t)^{n-1} = t^{m-1}(1-t)^{n-1}[m(1-t)-nt]$, equals 0. And this occurs when the expression in brackets equals 0, which is when $t = m/(m+n)$. Thus, letting $m=(1-p)K-1$ and $n = pK-1$, the integrand in the preceding integral increases from 0 to

$$\frac{(1-p)K-1}{(K-2)} = \frac{K}{K-2}(1-p) - \frac{1}{K-2}.$$

Since $p < \hat{q}$ and therefore $1-p > 1-\hat{q}$, elementary manipulations show that this cutoff is at least $1-\hat{q}$ whenever $K \geq \frac{2\hat{q}-1}{\hat{q}-p}$, which is automatic if $\hat{q} < \frac{1}{2}$ since then the right side is negative. Thus, if

$$K \geq \frac{2\hat{q}-1}{\hat{q}-p}, \quad \ldots\ldots\ldots\ldots\ldots \quad (36)$$

when integrated over the interval $[0, 1-\hat{q}]$, the integrand reaches its maximum at $1-\hat{q}$. That maximum times the length of the interval of integration therefore provides an upper bound for the integral, which justifies the preceding inequality.

It now follows that for large K obeying Eq. (36), we have

$$\begin{aligned}
p_2 &= \binom{K}{pK} \int_0^{1-q} t^{(1-p)K-1}(1-t)^{pK-1}dt \quad . \quad (37) \\
&< \sqrt{p(1-p)K}\left((1-p)^{(1-p)}p^p\right)^{-K} \\
&\quad \times (1-\hat{q})^{(1-p)K}\hat{q}^{pK-1} \\
&= \sqrt{\frac{p(1-p)K}{\hat{q}^2}}\left((1-p)^{(1-p)}p^p\right)^{-K} \\
&\quad \times ((1-\hat{q})^{(1-p)}\hat{q}^p)^K \\
&= \sqrt{\frac{p(1-p)K}{\hat{q}^2}}\left[\left(\frac{1-\hat{q}}{1-p}\right)^{1-p}\left(\frac{\hat{q}}{p}\right)^p\right]^K \\
&< \frac{\sqrt{pK}}{\hat{q}} \cdot \left[\left(\frac{1-\hat{q}}{1-p}\right)^{1-p}\left(\frac{\hat{q}}{p}\right)^p\right]^K
\end{aligned}$$

Applying the endogenous information in Eq. (31) to the inequality in Eq. (38) gives Eq. (23) and the proof is complete.

Appendix E. Proof that the Strict Vertical NFL is a Special Case of the General

Proof of Theorem 4.3.

When Eq. (17) is true, we will first show that

$$\frac{D(p\|q)}{p} \xrightarrow[p\ll q\ll 1]{} e^{I_+}. \quad \ldots\ldots\ldots\ldots \quad (38)$$

Since $\ln(1-t) \xrightarrow[t\to 0]{} -t$, Eq. (24) becomes

$$D(p\|q) \xrightarrow[p\ll q\ll 1]{} (1-p)(q-p) - pI_+.$$

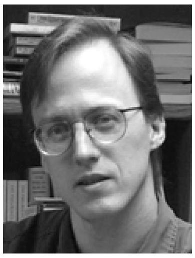
Since $1-p \xrightarrow[p\ll 1]{} 1$ and $q-p \xrightarrow[p\ll q]{} q$, the pI_+ term is dwarfed and we have

$$D(p\|q) \xrightarrow[p\ll q\ll 1]{} q.$$

and

$$\frac{D(p\|q)}{p} \xrightarrow[p\ll q\ll 1]{} \frac{q}{p} = e^{I_+},$$

which is the result of the strict vertical NFLT in Eq. (19).



Name:
William A. Dembski

Affiliation:
Senior Fellow, Center for Science and Culture, Discovery Institute

Address:
208 Columbia Street, Seattle, WA 98104, USA

Brief Biographical History:
1996- Senior Fellow with Discovery Institute
2000- Associate Research Professor in the conceptual foundations of science at Baylor University
2007- Senior Research Scientist with the Evolutionary Informatics Lab

Main Works:
- "Randomness by Design," Nous, Vol.25, No.1, March 1991.
- "The Design Inference: Eliminating Chance through Small Probabilities," Cambridge University Press, 1998.
- "No Free Lunch: Why Specified Complexity Cannot Be Purchased without Intelligence," Rowman & Littlefield, 2002.
- "Conservation of Information in Search: Measuring the Cost of Success," IEEE Trans. on Systems, Man and Cybernetics A, Systems & Humans, Vol.5, No.5, September 2009 (with R. J. Marks II).

Membership in Academic Societies:
- American Mathematical Society (AMS)
- Senior Member, Institute of Electrical and Electronics Engineers (IEEE)

Name:
Robert J. Marks II

Affiliation:
Department of Electrical & Computer Engineering, Baylor University

Address:
One Bear Place, Waco, TX 76798-7356, USA

Brief Biographical History:
1977-2003 Department of Electrical Engineering, University of Washington, Seattle, WA
2003- Distinguished Professor of Electrical & Computer Engineering, Baylor University, Waco, Texas

Main Works:
- R. J. Marks II, "Handbook of Fourier Analysis and Its Applications," Oxford University Press, 2009.
- E. C. Green, B. R. Jean, and R. J. Marks II, "Artificial Neural Network Analysis of Microwave Spectrometry on Pulp Stock: Determination of Consistency and Conductivity," IEEE Trans. on Instrumentation and Measurement, Vol.55, No.6, pp. 2132-2135, December 2006.
- J. Park, D. C. Park, R. J. Marks II, and M. A. El-Sharkawi, "Recovery of Image Blocks Using the Method of Alternating Projections," IEEE Trans. on Image Processing, Vol.14, No.4, pp. 461-471, April 2005.
- S. T. Lam, P. S. Cho, R. J. Marks, and S. Narayanan, "Detection and correction of patient movement in prostate brachytherapy seed reconstruction," Phys. Med. Biol., Vol.50, No.9, pp. 2071-2087, May 7, 2005.
- I. N. Kassabalidis, M. El-Sharkawi, and R. J. Marks II, "Dynamic Security Border Identification Using Enhanced Particle Swarm," IEEE Trans. on Power Systems, Vol.17, Issue 3, pp. 723-729, Aug. 2002.
- R. D. Reed and R. J. Marks II, "Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks," MIT Press, Cambridge, MA, 1999.
- R. J. Marks II, "Introduction to Shannon Sampling and Interpolation Theory," Springer-Verlag, 1991.
- R. J. Marks II, Editor, "Advanced Topics in Shannon Sampling and Interpolation Theory," Springer-Verlag, 1993.
- R. J. Marks II, Editor, "Fuzzy Logic Technology and Applications," IEEE Technical Activities Board, Piscataway, 1994.
- J. Zurada, R. J. Marks II, and C. J. Robinson, Editors, "Computational Intelligence: Imitating Life," IEEE Press, 1994.

Membership in Academic Societies:
- Fellow, Optical Society of America (OSA)
- Fellow, Institute of Electrical and Electronics Engineers (IEEE)

Erratum

The Horizontal No Free Lunch theorem takes $\tilde{T}$, the set of possible targets to be a partition of the search space Ω. However, in the case of multiple queries, they are treated as a single query to a larger search space Ω_Q. Any distribution of targets on the original search space will produce overlapping targets on Ω_Q. This prevents considering the set of possible targets as a partition of the search space Ω_Q. As such the horizontal no free lunch theorem does not properly handle multiple queries.

Theorem 1. *Given a uniform distribution over targets of cardinality k, and baseline uniform distribution, the average active information will be non-positive*

Proof. The average active information formula in its general form is

$$E[I_+] = E\left[-\log\frac{\phi(T)}{\psi(T)}\right] = E\left[\log\frac{\psi(T)}{\phi(T)}\right] \tag{1}$$

where T the random variable taking the domain of the powerset of Ω, $\phi(T)$ is the fixed probability of the baseline uniform search algorithm succeeding given a specific target and $\psi(T)$ is the fixed probability of the search algorithm being considered succeeding given a specific target.

By Jensen's inequality

$$E\left[\log\frac{\psi(T)}{\phi(T)}\right] \leq \log E\left[\frac{\psi(T)}{\phi(T)}\right] \tag{2}$$

Any given target can be represented by a function producing 1 for every point ω in the target, and 0 otherwise. All functions with the same number of targets will be closed under permutation. The algorithms succeed if they pick a point in the target, which is equivalent running single query and getting a 1. The No Free Lunch theorem holds for any distribution which is closed under permutation, and thus all algorithms will have on average, the same probability of success.

$$E[\psi(T)] = E[\phi(T)] = \frac{k}{|\Omega|} \tag{3}$$

We note that $\phi(T)$ is a constant, performing the same regardless of the identity of the target. It is a baseline uniform search, and thus will select every target with the same probability. Therefore,

$$E\left[\frac{\psi(T)}{\phi(T)}\right] = \frac{E[\psi(T)]}{\phi(T)} = \frac{\phi(T)}{\phi(T)} = 1 \tag{4}$$

This effectively restates the No Free Lunch theorem, the expected performance of two search algorithms does not differ. Using equation 2:

$$E[I_+] = E\left[\log\frac{\psi(T)}{\phi(T)}\right] \leq \log E\left[\frac{\psi(T)}{\phi(T)}\right] = \log\frac{E[\psi(T)]}{\phi(T)} = \log 1 = 0 \tag{5}$$

Therefore the expected active information is always non-positive.

Additionally, equality only occours if

$$\frac{\psi(T)}{\phi(T)} \tag{6}$$

is the same for all values of T. This will only occour if both algorithms succeed and fail with the same probability for every target.

□

Corollary 1. *Given a distribution over all targets, such that targets of the same cardinality have the same probability: the average active information will be non-positive*

$$E\left[\log\frac{\psi(T)}{\phi(T)}\right] \leq 0 \tag{7}$$

Proof. We can express the average active information as

$$E\left[\log\frac{\psi(T)}{\phi(T)}\right] = \sum_{k=0}^{k=|\Omega|} \Pr\left[|T| = k\right] E\left[\log\frac{\psi(T)}{\phi(T)}\middle| |T| = k\right] \tag{8}$$

Since the probabilities for sets of the same size are the same, Theorem 1 applies to the inner expected value. The expression then takes the weighted average of non-positive values and thus will be non-positive. □

Theorem 2. *For every distribution of targets there is some baseline search for which the active information will always be non-positive.*

$$E[I_+] = E\left[-\log\frac{\phi(T)}{\psi(T)}\right] \leq 0 \tag{9}$$

for some ϕ.

Proof. We can rewrite the average active information as:

$$E[I_+] = E\left[-\log\frac{\phi(T)}{\psi(T)}\right] = E\left[-\log\phi(T)\right] - E\left[-\log\psi(T)\right] \tag{10}$$

For every possible distribution over targets, there exists an algorithm, ϕ, that maximizes $E\left[-\log\phi(T)\right]$. $E\left[-\log\phi(T)\right] \geq E\left[-\log\psi(T)\right]$ since ϕ maximizes the expectation, and thus, $E\left[-\log\phi(T)\right] - E\left[-\log\psi(T)\right] \leq 0$. □

The Horizontal No Free Lunch Theorem does hold in all cases; however, determining the baseline may be non-trivial.

We thank Dietmar Eben for his attention to our work and pointing out the problem with overlapping targets.