Searching Large Spaces: Displacement and the No Free Lunch Regress

William A. Dembski

Conceptual Foundations of Science
Baylor University, Box 7130
Waco, Texas 76798

William_Dembski@baylor.edu

version 2.1, 4 March 2005

Abstract

Searching for small targets in large spaces is a common problem in the sciences. Because blind search is inadequate for such searches, it needs to be supplemented with additional information, thereby transforming a blind search into an assisted search. This additional information can be quantified and indicates that assisted searches themselves result from searching higher-level search spaces—by conducting, as it were, a search for a search. Thus, the original search gets displaced to a higher-level search. The key result in this paper is a displacement theorem, which shows that successfully resolving such a higher-level search is exponentially more difficult than successfully resolving the original search. Leading up to this result, a measure-theoretic version of the No Free Lunch theorems is formulated and proven. The paper shows that stochastic mechanisms, though able to explain the success of assisted searches in locating targets, cannot, in turn, explain the source of assisted searches.

1 Blind Search

Most searches that come up in scientific investigation occur over spaces that are far too large to be searched exhaustively. Take the search for a very modest protein, one that is, say, 100 amino acids in length (most proteins are at least 250 to 300 amino acids in length). The space of all possible protein sequences that are 100 amino acids in length has size 20^{100}, or approximately 1.27×10^{130}. Exhaustively searching a space this size to find a target this small is utterly

beyond not only present computational capacities but also the computational capacities of the universe as we know it. Seth Lloyd (2002), for instance, has argued that 10^{120} is the maximal number of bit operations that the known, observable universe could have performed throughout its entire multi-billion year history.

If exhaustive searches are infeasible for large spaces, what about random searches? Do random searches fare better at successfully finding small targets in large spaces? The answer depends on what one means by the term *random*? Ordinarily what is meant is this: the internal structure of some large space Ω treats all its points as equivalent in the sense that the internal structure doesn't distinguish certain regions of the space from others. This internal structure is typically captured via a geometry and thus encapsulated in a metric D on Ω. In turn, D induces a uniform probability $\mathbf{U}$ on Ω (see Dembski 1990; spaces that are not uniformizable are the exception rather than the rule for most of the search problems that come up in scientific investigation). In general, random searches of Ω therefore come down to either uniform random sampling or random walks without drift.

When it comes to locating small targets in large spaces, random sampling and random walks are equally ineffective. Uniform random sampling treats Ω like a giant urn from which items are drawn (with replacement) according to the uniform probability $\mathbf{U}$. Each draw constitutes a brand-new attempt to locate a point in the target. Because such draws are probabilistically independent of each other, uniform random sampling cannot build on past successes in attempting to reach the target. For each sampling event, success at reaching the target is all-or-nothing.

The mathematics here is elementary: a uniform random sample from Ω of size m will have a better-than-even chance of containing a point in a small target T ($\subset \Omega$) only if m is close to the reciprocal of the uniform probability of the target. That's because for a small target T, the probability $p = \mathbf{U}(T)$ will be correspondingly small. Thus, hitting the target in at least one of m independent trials has probability $1-(1-p)^m$. This number approaches $1-e^{-1} \approx 0.63 > 1/2$ as m approaches $1/p$, but remains close to 0 as m falls far short of $1/p$. Thus, for instance, to stand a reasonable chance of locating a particular protein 100 amino acids in length, uniform random sampling requires a sample size m on the order of 10^{130} (p here is 20^{-100}, so $1/p = m \approx 10^{130}$). It follows that uniform random sampling is no better than exhaustive search in decreasing the number of points that, on average, need to be examined before the search succeeds.

Nor are random walks more efficient in this regard. Whereas uniform random sampling at each step selects a point with respect to the uniform probability $\mathbf{U}$ over the entire search space Ω, a random walk at each step selects a point with respect to the uniform probability defined over a neighborhood of fixed proximity to the previously selected point. For instance, given that x_k is selected at step k, x_{k+1} may be selected by uniformly sampling $B_\varepsilon(x_k) = \{y \in \Omega : D(y, x_k) \leqslant \varepsilon\}$ (i.e., the ball of radius ε around x_k). In practice, defining uniform probabilities on such neighborhoods is unproblematic because the uniform probability on a neighborhood like $B_\varepsilon(x_k)$ is simply the uniform probability of the whole space

conditioned on the neighborhood, i.e., $\mathbf{U}(\cdot|B_\varepsilon(x_k))$ (see Dembski 1990 for the conditions under which this result holds as well as for pathological spaces that constitute an exception).

Of course, we need to make sure that the random walk has no unfair advantage in finding the target. For instance, the random walk must not have an inherent tendency to drift toward the target T. This would unduly increase the probability of the walk reaching the target. Moreover, the initial starting point of the random walk needs to be selected by sampling with respect to the uniform probability $\mathbf{U}$ over the entire search space Ω. This ensures that the starting point is not deliberately taken so close to T (or, worse yet, inside T itself) that it is likely to reach the target in but few steps.

Given these provisos, the stopping time for the random walk to reach the target T will have expected value whose order of magnitude is $\frac{1}{\mathbf{U}(T)}$ (for the relevant mathematics, see Spitzer 2001). The stopping time here is simply the number of steps required for the random walk to reach the target. Thus, for instance, the average number of steps for a random walk to locate a particular protein 100 amino acids in length will have order of magnitude $\frac{1}{\mathbf{U}(T)} = \frac{1}{1/20^{100}} = 20^{100} \approx 10^{130}$. Thus, we find that random walks, as with uniform random sampling, are no better than exhaustive search in decreasing the number of points that, on average, need to be examined before the search succeeds.

Uniform random sampling and random walks both presuppose a uniform probability $\mathbf{U}$ on the search space Ω. This presupposition places a mathematical restriction on Ω. Nevertheless, it imposes no practical limitation on this space. Many of the spaces that come up in practice are finite and thus have a straightforward uniform probability, namely, one that assigns the same probability to each point in the space (i.e., $1/N$ if N is the number of points in the space). More generally, for a space to be uniformizable, it must be a compact metric space. Granted, this precludes search spaces of infinite diameter. Nonetheless, even with search spaces whose diameter is potentially infinite, in practice we limit searches to bounded subspaces that are compact. For instance, if we are trying to account for the fine-tuning of the gravitational constant as the solution of a search through the space of all possible values that the constant might take (namely, the positive reals, a space that is unbounded and noncompact), it is in practice enough to consider a closed interval around the actual gravitational constant (such intervals are compact) and show that possible gravitational constants compatible with a life-permitting universe have small probability within this interval (see Collins 2003).

Random walks are discrete stochastic processes and thus depend on traversing a finite number of steps. By increasing the number of steps and making them small enough, random walks converge to Brownian motions, which are continuous stochastic processes (see Billingsley 1999: ch. 2). Here again, provided that a Brownian motion is without drift and that its starting point is selected with respect to $\mathbf{U}$ over all of Ω, the stopping time (which now reflects a continuous measure of time) for it to reach the target T will have expected value proportional to $\frac{1}{\mathbf{U}(T)}$ (see Port and Stone 1978: ch. 2; see also Doob 1984). This

means that however time is scaled, the time it takes for the Brownian motion to reach the target will, on average, be bounded below by some fixed factor times $\frac{1}{\mathbf{U}(T)}$. It follows that if the time to search Ω is limited, targets T whose uniform probability is small enough will, for practical purposes, be unsearchable by means of Brownian motions.

Exhaustive search, uniform random sampling, random walks, and Brownian motion all fall under *blind search*. In general, blind search can be characterized as a conversation between two interlocutors, call them Alice and Bob. Alice has access only to the search space Ω. Bob not only has access to Ω, but also to the target T (T is a subset of Ω; both Ω and T are nonempty). We think of T as a problem that Bob has posed to Alice, and we think of the points in T as solutions to the problem. Alice's job is to find at least one such solution. What makes Alice's job difficult is that nothing about Ω provides a clue about T. If we think of Ω as a giant urn filled with balls, it's not as though the balls are color-coded so that those with certain colors are more likely to be solutions. Rather, we should imagine that all the balls have the same color and that the only information Alice has about the balls is their color.

Bob, in contrast, knows a lot more. For any candidate solution x in Ω, Bob is able to tell Alice whether x is in T (i.e., whether x is in fact a solution). In attempting to find a solution, Alice now successively selects candidate solutions $x_1, x_2, \ldots, x_m$ from Ω, at each step querying Bob whether a candidate is in fact a solution. Note that Alice limits herself to a finite search with at most m steps—she does not have infinite resources to continue the search indefinitely. (For simplicity, in the sequel, we limit ourselves to discrete searches with finitely many steps; searches involving continuous time, as with Brownian motion, add no fundamental new insight to the discussion.) Presented by Alice with a candidate solution, Bob truthfully answers whether it is in fact in T. The search is successful if Bob informs Alice that one of her m candidate solutions resides in T.

This is blind search. Two things render it blind. First, the search space Ω provides no clue about the solution space or target T. This can be captured mathematically by placing a uniform probability $\mathbf{U}$ on Ω and by limiting Alice to sampling candidate solutions from Ω based only on the knowledge of $\mathbf{U}$ and the underlying metric structure D (and, perhaps, other structures internal to Ω that provide no clue about the target—genetic algorithms, for instance, allow for candidate solutions to be "mated"). Second, Bob refuses to divulge to Alice anything about a candidate solution except whether it is in the solution space T. The information that Bob gives Alice is all-or-nothing—a candidate solution is either in or out.

In a blind search, Bob provides the minimum amount of information that Alice needs if she is to stand any chance of finding a solution. Moreover, as we've just seen, for locating small targets in large spaces, this information is essentially useless. Provided the targets are small enough, blind search will, in practice, not find them. For a search procedure to be more effective than blind search, it therefore needs to provide additional information.

2 Assisted Search

Let us, therefore, define an *assisted search* as any search procedure that provides more information about candidate solutions than a blind search. The prototypical example of an assisted search is an Easter egg hunt in which instead of saying "yes" or "no" for each possible place where an egg might be hidden, one says "warmer" or "colder" depending on whether the distance to an egg is narrowing or widening. This additional information clearly assists those who are looking for the Easter eggs, especially when the eggs are well hidden and blind search would be unlikely to find them.

To characterize assisted search, let us again employ the services of Alice and Bob. As before, Bob has a full grasp of the target T, so that for any proposed solution x in Ω, he is able to answer whether x is in T. Alice, on the other hand, knows only Ω and whatever information Bob is willing to divulge that might help her to find a solution in T. We may assume that in knowing Ω, Alice knows Ω's geometric structure (as induced by the metric D) as well as the uniform probability $\mathbf{U}$. We may also assume that Alice knows enough about the problem in question to ascertain the probability $p = \mathbf{U}(T)$. Finally, we assume that m is an upper bound on the number of candidate solutions in Ω that Alice can verify with Bob, and that mp, the approximate probability for locating the target by uniform random sampling given a sample size m, is so small that Alice has no hope of attaining T via a blind search.

Alice and Bob are playing a game of "m questions." Initially Bob divulged too little information for Alice to have any hope of winning the game. The game initially played was thus one of blind search. Generous fellow that Bob is, he is now going to divulge additional information to Alice so that she may reasonably hope to win the game. The game now being played is therefore one of assisted search. To make the game interesting, Bob needs to give Alice enough information so that she stands a reasonable chance of locating the target T by proposing m candidate solutions. The challenge for Alice is to make optimal use of whatever information Bob gives her so that her m questions are as effective as possible for locating the target.

We can represent this situation mathematically as follows. Think of Bob as possessing an *information function* $\mathbf{j}$ that maps $cand(\Omega, m) =_{def} \bigcup_{k=1}^{m} \Omega^k$ into a space of responses Λ (here Ω^k is the k-fold Cartesian product of Ω with itself and Λ is some nonempty set). Call $cand(\Omega, m)$ the *candidate solution space for Ω with sample size m*. The idea behind $\mathbf{j}$ is this: As Alice proposes candidate solutions $x_1, x_2, \ldots, x_m$, at each step k ($1 \leqslant k \leqslant m$) Bob responds with an item of information $\mathbf{j}(x_1, \ldots, x_k)$ from the response space Λ. This way of representing the information that Bob gives to Alice ensures that Bob need not merely respond to each candidate solution in isolation but rather to the whole sequence of candidates that Alice has proposed leading up to it. Thus, for each candidate solution x_k, Bob is able to keep in memory previous candidate solutions $x_1, \ldots, x_{k-1}$. This is important because it enables Bob to convey dependencies among previously proposed candidates (as in an Easter egg

hunt, where comparative responses such as "warmer" and "colder" depend on previous locations visited by the Easter egg hunter). These dependencies can provide Alice with crucial information for locating the target T.

It's now clear why blind search provides Alice with so little information for locating T. In the scheme just outlined, blind search is represented (up to isomorphism) as follows: $\Lambda = \{0, 1\}$ and $\mathbf{j}(x_1, ..., x_k) = 1$ if x_k is in the target T, 0 otherwise. Here 0 tells Alice that x_k is not in the target, 1 that it is. Observe that for the k-tuple $(x_1, ..., x_k)$, the value of $\mathbf{j}$ depends only the last element of the k-tuple, namely, x_k, and merely records whether this candidate solution is in T. Let us call such a $\mathbf{j}$ an *indicator function* for the target T. Information functions like this characterize blind search.

Since assisted search is supposed to augment the information inherent in blind search, the information function associated with an assisted search needs to contain strictly more information than is contained in the indicator function of the corresponding blind search. This strict increase in information can be characterized as follows: An information function $\mathbf{\breve{j}}$ strictly augments the information in an indicator function $\mathbf{j}$ associated with a target T provided there is a function φ from Λ to $\{0, 1\}$ such that $\varphi \circ \mathbf{\breve{j}} = \mathbf{j}$ and for any such φ there is no function ψ from $\{0, 1\}$ back to Λ such that $\psi \circ \varphi \circ \mathbf{\breve{j}} = \mathbf{\breve{j}}$. In other words, composing $\mathbf{\breve{j}}$ with some function allows Alice to recover the indicator function for T, but $\mathbf{\breve{j}}$ cannot in turn be recovered from this composition.

The underlying intuition here can be understood from the measure-theoretic idea that the amount of information associated with a random variable depends on the complexity of the σ-algebra induced by it (see Bauer 1981: 309–319). This can be generalized as follows: a function $f : \mathcal{A} \to \mathcal{B}$ conveys information about $\mathcal{A}$ to the degree that the values it assumes in $\mathcal{B}$ discriminate among the elements of $\mathcal{A}$. Think of the information associated with f as the partition of $\mathcal{A}$ resulting from all the inverse images of elements of $\mathcal{B}$, i.e., $Part(f) = \{A \subset \mathcal{A} : A = f^{-1}(b) \text{ for some } b \in \mathcal{B}\}$. Accordingly, the composition of a function g with f (i.e., $g \circ f$) can at best retain all the information that f provides about $\mathcal{A}$, and in fact will strictly diminish it if f cannot be recovered from $g \circ f$ by further composition of functions. This follows because the partition of $\mathcal{A}$ associated with $g \circ f$ cannot be finer than the one associated with f and may in fact be coarser. In other words, $Part(g \circ f) \subset Part(f)$ with inequality always a possibility. This is just elementary set theory.

The information function $\mathbf{j}$ was defined deterministically. It could also be defined stochastically. Thus, instead of $\mathbf{j}$ mapping $cand(\Omega, m)$ into Λ, it could be defined from $cand(\Omega, m) \times \Gamma$ into Λ where Γ is a probability space that supplies a randomizing element to $\mathbf{j}$. It follows that for a stochastic information function $\mathbf{j}$, the item of information associated with a k-tuple $(x_1, ..., x_k)$ need not assume a univocal value, as in the deterministic case, but can vary according to a probability distribution. With regard to Alice and Bob, this means that Bob isn't required every time to provide the same answer to the same k-tuple proposed by Alice.

The information function $\mathbf{j}$ (whether deterministic or stochastic) is what Bob brings to the game of "m questions" that he is playing with Alice. What does

Alice bring to the game? Alice brings a strategy to take advantage of this information. Mathematically, this can be expressed as follows. Think of Alice as possessing a *strategy function* $\mathbf{s}$ that maps $strat(\Omega, \Lambda, m) =_{def} \bigcup\limits_{k=0}^{m-1} [\Omega^k \times \Lambda^k]$ to the search space Ω. Call $strat(\Omega, \Lambda, m)$ the *strategy space for Ω and Λ with sample size m* (note that unlike for the candidate solution space, the strategy space is defined for a grand union whose index begins at 0 rather than 1 and ends at $m-1$ rather than m—the reason will be clear in a moment). The idea behind $\mathbf{s}$ is this: As Alice proposes candidate solutions $x_1, x_2, \ldots, x_m$ and at each step k ($1 \leqslant k \leqslant m$) receives feedback from Bob, transmitted as items of information $\mathbf{j}(x_1, \ldots, x_k)$, Alice responds by selecting a candidate solution $\mathbf{s}(x_1, x_2 \ldots, x_{k-1}, \mathbf{j}(x_1), \mathbf{j}(x_1, x_2) \ldots, \mathbf{j}(x_1, x_2 \ldots, x_{k-1})) = x_k$ (note that for $k = 1$ this needs to be interpreted as $\mathbf{s}(\emptyset, \emptyset) = x_1$ and for $k = 2$ this is just $\mathbf{s}(x_1, \mathbf{j}(x_1)) = x_2$). The function $\mathbf{s}$ is Alice's strategy for proposing candidate solutions in response to both previous candidate solutions (proposed by her) and previous items of information (given by Bob).

As with information functions, strategy functions can be deterministic or stochastic (in which case $\mathbf{s}$ needs to map $strat(\Omega, \Lambda, m) \times \Gamma'$ into Ω where Γ' is a probability space that supplies a randomizing element to $\mathbf{s}$). It follows that for a stochastic strategy function $\mathbf{s}$, the candidate solution associated with $(x_1, x_2 \ldots, x_{k-1}, \mathbf{j}(x_1), \mathbf{j}(x_1, x_2) \ldots, \mathbf{j}(x_1, x_2 \ldots, x_{k-1}))$ need not be uniquely determined, but can vary according to a probability distribution. In that case, Alice isn't required every time to provide the same candidate solution given the same history of prior exchanges between Alice and Bob.

In general, we may therefore characterize an assisted search $\mathfrak{A}$ on the space Ω with target T, metric structure D, uniform probability $\mathbf{U}$, and sample size m as a pairing of strategy function $\mathbf{s}$ and information function $\mathbf{j}$, i.e., $\mathfrak{A} = (\mathbf{s}, \mathbf{j})$. In case of a blind search, $\mathbf{j}$ is just the indicator function for the target T and $\mathbf{s}$ is any sampling scheme with no inherent bias toward or prior knowledge about the target T. This last condition, which attempts to purify the strategy function of any special information from the environment regarding the target, is difficult to formulate with full generality (cf. Culberson 1998). Nonetheless, there are clear instances where this condition is unproblematically fulfilled—exhaustive search, uniform random sampling, and random walks being cases in point. Uniform random sampling, as a form of blind search, is conveniently represented as the ordered pair $\mathfrak{B} = (\mathbf{U}, 1_T)$. Here 1_T is the indicator function with respect to T ($1_T(x) = 1$ for $x \in T$, 0 otherwise). The first element in this ordered pair is interpreted as Alice's strategy of uniformly randomly sampling from the search space whenever she proposes a candidate solution; the second is interpreted as the information function that Bob uses to guide Alice.

This framework for understanding assisted search is entirely general. By choosing the response space Λ as the set of nonnegative reals, Bob can define the information function $\mathbf{j}$ as a hill-climbing comparator, thereby turning Alice's search into a straightforward hill-climbing search. More generally, Bob can define the information function as a stochastic, time covarying fitness landscape,

thereby turning Alice's search into an evolutionary search. Provided there is enough structure on the search space Ω to allow for elements to be mutated and recombined, Alice can select her strategy function **s** so that her search becomes a genetic algorithm (GA). Note that this framework is entirely compatible with searches in which at each step Alice proposes not a single candidate solution but a whole set of them (i.e., a population). In this case, Bob's information function and Alice's strategy function will be invariant under permutation for the indexes corresponding to each such population. Supervised learning using artificial neural networks falls within this framework (Reed and Marks 1999). So do iterative forms of optimization, including those that employ populations whose agents, as it were, communicate by walkie-talkies, as in particle swarm optimization (Shi 2004). So do self-learning or self-play optimization scenarios, in which agents compete in successive rounds of selection (Fogel et al. 2005). Moreover, this assimilation of self-play to the framework outlined here holds despite the claim that such optimizations offer free lunches (Wolpert and Macready 2005).

3   A Simplification

The obvious question that now needs to be addressed is how to assess the efficacy of assisted search over blind search. Clearly, assisted search needs to be more effective at locating small targets in large spaces than blind search. The question is how go gauge this effectiveness. Specifically, given an assisted search $\mathfrak{A}$ and a blind search $\mathfrak{B}$, is there some way to measure the degree to which $\mathfrak{A}$ is better at locating a target than $\mathfrak{B}$?

To answer this question, we need first to establish a baseline for how effective blind search is at locating a target. Given a search space Ω with metric D and uniform probability **U**, there are many ways a blind search might attempt to locate a target T contained in Ω. Suppose, for instance, Ω is finite with N elements enumerated in the following order: $a_1, a_2, \ldots, a_N$. In this case, a blind search might simply be an exhaustive search that runs through these elements in order starting with a_1 and ultimately (if time an resources permit) ending with a_N. But this raises a difficulty: if early in this enumeration of elements, one of the a_is happens to fall in T (in the most extreme case, if a_1 were to fall in T), this blind (qua exhaustive) search for T would in fact be quite effective. Of course, as the target varied (even if the size or probability of the target remained constant), the search would in most cases be less effective. The underlying difficulty here is that the effectiveness of this search depends on idiosyncrasies in the relation between target and enumeration of elements defining the search. Move the target or change the enumeration, and the effectiveness of the search may rocket or plummet. It follows that exhaustive searches like this preclude a usable baseline for the effectiveness of blind search.

Similar difficulties arise for random walks. Even if the starting point of a random walk is taken by uniform random sampling of the entire search space Ω (i.e., by picking a point at random in Ω according to the uniform probability

U), if the search space is disconnected in the sense that it decomposes into geometrically isolated portions or islands, the random walk, by always only taking baby steps (i.e., small steps within neighborhoods of fixed proximity from previous points in the random walk), may get stuck on some island of the search space and never be able to reach a target located on another island. Thus, the effectiveness of such a search can depend on idiosyncrasies in the relation between the connectivity of the search space and the target. Moreover, even if the search space is connected, there can be bottlenecks that unduly hinder the random walk from locating the target. Still another issue is the size of the jumps at each step in the random walk: if these are too large or too small, the random walk may consistently miss certain types of targets.

Because of such difficulties, neither exhaustive searches nor random walks are helpful in setting a baseline for how effective blind search is at locating a target. There is one form of blind search, however, that avoids all such dependencies, namely, uniform random sampling. Because at each step in the search uniform random sampling always samples the entire search space with respect to the uniform probability **U** and because each step in the search is probabilistically independent of the others, the effectiveness with which this form of blind search locates a target T depends solely on two things: the probability of the target, namely, $p = \mathbf{U}(T)$, and the number of points in the search space capable of being sampled, namely, the sample size m. As we saw in section 1, for a sample of size m, uniform random sampling locates a target of probability p with probability $1 - (1-p)^m$. If, as is typical with small targets in large spaces, $p = \mathbf{U}(T)$ is much smaller than $1/m$, then $1 - (1-p)^m$ is approximately mp. In the sequel, we take uniform random sampling, which in section 2 we denoted by $\mathfrak{B} = (\mathbf{U}, 1_T)$, as the baseline for blind search.

Given this baseline, the next question is how to assess the degree to which an assisted search $\mathfrak{A} = (\mathbf{s}, \mathbf{j})$ does a better job locating the target T than $\mathfrak{B} = (\mathbf{U}, 1_T)$. Given the extreme generality with which assisted search is characterized (see section 2), it is not immediately obvious how to compare the assisted search $\mathfrak{A} = (\mathbf{s}, \mathbf{j})$ with the baseline search $\mathfrak{B} = (\mathbf{U}, 1_T)$. $\mathfrak{A}$ may incorporate prior knowledge about the target; then again, it may not. $\mathfrak{A}$ may gradually converge on the target; then again, it may waste most of its initial candidate solutions, after which it suddenly zeros in on the target with a vengeance. Repeated applications of $\mathfrak{A}$ may increase the chances of locating the target; then again, repeated applications of $\mathfrak{A}$ may provide no advantage over a single application.

Given so much free rein with assisted search, simplifying the representation of $\mathfrak{A}$ in relation to $\mathfrak{B}$ is desirable. The following considerations suggest a canonical simplification. Since the sample size m sets an upper bound on the number of points in the search space that may be sampled, both $\mathfrak{A}$ and $\mathfrak{B}$ may be seen as proposing m candidate solutions, call them $x_1, x_2, \ldots, x_m$ for $\mathfrak{A}$ and $y_1, y_2, \ldots, y_m$ for $\mathfrak{B}$. These candidate solutions can be viewed as instantiations of Ω-valued random variables, respectively $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_m$. Since $\mathfrak{B}$ is just uniform random sampling, the Y_is are independent and identically distributed with probability distribution **U**. Of course, no neat characterization like this is possible for the X_is, whose joint probabilistic

structure may be completely intractable. Nonetheless, just as the key to assessing the efficacy of $\mathfrak{B}$ in locating the target T is the probability that at least one of the Y_is successfully locates T, so too it makes sense to identify the efficacy of $\mathfrak{A}$ in locating the target T as the probability that at least one of the X_is successfully locates T.

In the case of $\mathfrak{B}$, this probability is one we've already seen, namely,

$$
\begin{aligned}
\mathbf{P}(Y_1 \;\in\; & T \;\vee Y_2 \in T \;\vee \cdots \vee Y_m \in T) \\
=\;& \sum_{i=1}^{m} \mathbf{P}(Y_1 \notin T \;\wedge Y_2 \notin T \;\wedge \cdots \wedge Y_{i-1} \notin T \wedge Y_i \in T) \\
=\;& \sum_{i=1}^{m} \mathbf{P}(Y_1 \notin T\;)\mathbf{P}(Y_2 \notin T\;) \cdots \mathbf{P}(Y_{i-1} \notin T)\mathbf{P}(Y_i \in T) \quad\;\; \text{[by indep.]} \\
=\;& \sum_{i=1}^{m} (1-p)^{i-1}p \\
=\;& 1 - (1-p)^m.
\end{aligned}
$$

Here $\vee$ denotes disjunction and $\wedge$ conjunction. In the case of $\mathfrak{A}$, however, the corresponding probability does not simplify. Given the extreme generality with which assisted search is characterized, it appears that the best we can do to assess the effectiveness of $\mathfrak{A}$ in locating T is to calculate the probability $\mathbf{P}(X_1 \in T \;\vee X_2 \in T \;\vee \cdots \vee X_m \in T)$.

This, however, suggests an analogy with the case of uniform random sampling. Since $\mathbf{P}(X_1 \in T \;\vee X_2 \in T \;\vee \cdots \vee X_m \in T)$ is a probability, and since for fixed m and for arbitrary α in the unit interval, the expression $1 - (1-\alpha)^m$ varies between 0 and 1, we can find a uniquely determined number q in the unit interval for which $\mathbf{P}(X_1 \in T \vee X_2 \in T \vee \cdots \vee X_m \in T) = 1-(1-q)^m$. In place of the random variables $X_1, X_2, \ldots, X_m$, we may therefore substitute independent and identically distributed random variables $X_1', X_2', \ldots, X_m'$ whose probability of locating T in m steps precisely equals the probability of $X_1, X_2, \ldots, X_m$ locating T in m steps. In that case,

$$
\begin{aligned}
1 - (1-q)^m \;=\;& \mathbf{P}(X_1 \in T \;\vee X_2 \in T \;\vee \cdots \vee X_m \in T) \\
=\;& \mathbf{P}(X_1' \in T \;\vee X_2' \in T \;\vee \cdots \vee X_m' \in T) \\
=\;& \sum_{i=1}^{m} \mathbf{P}(X_1' \notin T \;\wedge X_2' \notin T \;\wedge \cdots \wedge X_{i-1}' \notin T \wedge X_i' \in T) \\
=\;& \sum_{i=1}^{m} \mathbf{P}(X_1' \notin T\;)\mathbf{P}(X_2' \notin T\;) \cdots \mathbf{P}(X_{i-1}' \notin T)\mathbf{P}(X_i' \in T).
\end{aligned}
$$

But from this, given that the X_i's are independent and identically distributed, it follows that for $1 \leqslant i \leqslant m$, $\mathbf{P}(X_i' \in T\;) = q$.

The canonical simplification that gauges $\mathfrak{A}$'s effectiveness in locating T as compared with that of $\mathfrak{B}$ can now be summarized as follows. $\mathfrak{B}$'s effectiveness in locating T is the effectiveness of uniform random sampling locating T in m steps. Thus, to compare $\mathfrak{A}$'s effectiveness in locating T, substitute for the m random variables induced by $\mathfrak{A}$ corresponding random variables that are independent and identically distributed, and whose probability of locating T in m steps is identical with the probability of the original random variables induced by $\mathfrak{A}$ locating T in m steps.

Accordingly, determining the relative effectiveness with which $\mathfrak{A}$ and $\mathfrak{B}$ locate T in m steps is a matter of comparing two probability measures: one induced by the Ω-valued random variable X' defined as any one of the independent and identically distributed random variables $X'_1, X'_2, \ldots, X'_m$ that jointly represent $\mathfrak{A}$; and the other induced by the Ω-valued random variable Y defined as any one of the independent and identically distributed (in this case uniformly distributed) random variables $Y_1, Y_2, \ldots, Y_m$ that jointly represent $\mathfrak{B}$.

Since Y is uniformly distributed on Ω, Y induces the uniform probability $\mathbf{U}$ on Ω. Thus, for an arbitrary Borel set S in Ω, $\mathbf{P}(Y \in S) = \mathbf{U}(S)$. Similarly, X' induces a probability measure μ on Ω. Thus, for an arbitrary Borel set S included in Ω, $\mathbf{P}(X' \in S) = \mu(S)$. Comparing $\mathfrak{A}$ and $\mathfrak{B}$ in their effectiveness to locate T in m steps therefore amounts to substituting the probability measure μ for $\mathfrak{A}$ and the uniform probability $\mathbf{U}$ for $\mathfrak{B}$, and then assessing the difference in magnitude between two probabilities, $\mu(T) = q$ and $\mathbf{U}(T) = p$. The degree to which $\mathfrak{A}$ is more effective than $\mathfrak{B}$ can thus be defined as the degree to which μ concentrates more mass on T than $\mathbf{U}$, and thus, in particular, the degree to which q is greater than p. Note that in this simplification, $\mathfrak{A}$ need not uniquely determine μ (in fact, except in the simplest cases, μ will be vastly underdetermined). At issue is what probability μ assigns to T and how this probability compares to the uniform probability of T.

Does this simplification adequately capture the key features of blind and assisted searches that enable their relative effectiveness to be accurately gauged? To be sure, these simplifications dispense with a lot of the structure of $\mathfrak{A}$ and focus on a particular type of blind search, namely, uniform random sampling. But representing blind search as uniform random sampling is, as argued at the start of this section, well warranted. Moreover, any information lost by substituting μ for $\mathfrak{A}$ is irrelevant for gauging $\mathfrak{A}$'s effectiveness in locating T since its effectiveness coincides with the probability of $\mathfrak{A}$ locating T in m steps, and this probability is captured by μ.

In any case, the simplification does not introduce any bias or distortion. Uniform random sampling is an exact instance of blind search, not a rough estimate of it. Moreover, I've indicated why it should be taken as the canonical form of blind search. At the same time, the probability measure μ induced by $\mathfrak{A}$ and, specifically, the probability that μ assigns to T, namely q, are not artificial additions to $\mathfrak{A}$ but rather fully inherent in $\mathfrak{A}$. After all, $1 - (1-q)^m$ is the exact probability that the assisted search $\mathfrak{A}$ locates the target T in m steps.

4   Accounting for Net Gains in Information

The simplification of the previous section essentially substitutes probability measures for search procedures, $\mathbf{U}$ for blind search and μ for assisted search. Now, in practice, once μ is in hand, the search for T can be considered successfully resolved provided that the probability μ assigns to T, namely q, is sufficiently large so that the probability of the assisted search successfully finding T in m steps, namely $1 - (1 - q)^m$, is reasonably close to 1. For this last condition to hold, a good rule of thumb is that the order of magnitude of q needs to be no smaller than $1/m$. In that case, $1 - (1-q)^m \geqslant 1 - (1-1/m)^m \approx 1 - e^{-1} \approx 0.63 > 1/2$.

Thus, even though $p = \mathbf{U}(T)$ is so small that random sampling with respect to $\mathbf{U}$ is extremely unlikely to locate T in m steps, by representing the assisted search as random sampling with respect μ, and given that $q \ (= \mu(T))$ is reasonably large (i.e., no smaller than $1/m$), the assisted search is highly like to locate the target. And, in practice, successfully concluding a search is all we care about, whether it is a computer algorithm successfully locating an optimum or nature successfully evolving a biological function.

In focusing on the conditions that μ must satisfy for an assisted search to have a high probability of concluding successfully, let us not lose track of a deeper question, namely, what is the source of the probability measure μ that makes a successful search possible. Since we will be referring to $\mathbf{U}$ and μ repeatedly, let us assign names to them. $\mathbf{U}$, as already noted, is the *uniform probability*. μ, as the probability induced by an assisted search, will be called the *exchange probability* (i.e., it exchanges with the uniform probability). In practice, we are concerned with how well an assisted search, as represented by the exchange probability, facilitates finding a target. But whence the assisted search and the exchange probability that represents it? In human engineering contexts, these facilitators of successful searches typically need to be invented (i.e., they are the result of intelligent design). But must they always be invented? Could they not simply be brute givens, gifts that a profligate environment bestows on certain searches without any need of explanation? In particular, might not the measure μ simply be a free lunch?

Let's examine these questions more closely. First off, it should be clear that an assisted search needs to input a substantial amount of novel information to make the search successful. The problem with blind search is that as the search proceeds, no signal is coming back to the search to ensure that it is getting closer to the target. Blind search offers extremely limited feedback and no sense of approximation to a target. Assisted search, by contrast, attempts to provide the feedback necessary for such approximation. Now, the precise form of this feedback, though fully encapsulated in the strategy and information functions that constitute an assisted search, is largely lost from the exchange probability μ (i.e., a lot of the details about an assisted search get lost when substituting an exchange probability for an assisted search). Nonetheless, in the very act of replacing $\mathbf{U}$, μ can be seen to introduce novel information, and this information can be measured.

To see this, recall that the information in an event A, denoted by $\mathbf{I}(A)$, is

by definition $-\log_2 \mathbf{P}(A)$ and measures the number of bits inherent in A's occurrence. This definition extends to two events A and B so that $\mathbf{I}(B|A)$ equals by definition $-\log_2 \mathbf{P}(B|A)$ and measures the number of bits inherent in B's occurrence given that A is known to have occurred. Now, this last definition generalizes to probability measures, so that for probability measures ν and ξ, $\mathbf{I}(\xi|\nu)$ equals by definition $\log_2 \int_\Omega \left(\frac{d\xi}{d\nu}\right)^2 d\nu$ provided that ξ is absolutely continuous with respect to ν (see Dembski 2004; note that $\frac{d\xi}{d\nu}$ is a Radon-Nikodym derivative). $\mathbf{I}(\xi|\nu)$ is a special case of Rényi information (Rényi 1961; Cover and Thomas 1991: 499–501) and can be thought of as measuring the number of bits required for ξ to update ν. Think of it this way: initially we thought that ν characterized some state of affairs but then we learned that ξ characterized that state of affairs; $\mathbf{I}(\xi|\nu)$ then measures the amount of information we've learned in updating ν by ξ. In the case at hand, μ updates $\mathbf{U}$ and thereby introduces $\mathbf{I}(\mu|\mathbf{U})$ bits of novel information.

What is this last number? Before answering this question, let us introduce a simplifying assumption that results in no loss of generality in the ensuing discussion. Specifically, let us assume that the collection of all probability measures ν that are absolutely continuous with respect to $\mathbf{U}$ are dense in the weak topology on $\mathbf{M}(\Omega)$, the latter space being the set of all probability measures defined on the Borel sets of Ω. This assumption holds for most search spaces that come up in practice, certainly for finite sets Ω as well as for the manifolds and topological groups that are the basis for much of physics (for exceptions see Dembski 1990). Given this assumption, μ will be indistinguishable from a probability measure that is absolutely continuous with respect to $\mathbf{U}$. As a consequence, we are justified treating μ as absolutely continuous with respect to $\mathbf{U}$ and therefore taking $\mathbf{I}(\mu|\mathbf{U})$ to be well-defined.

The remaining task, then, is to evaluate $\mathbf{I}(\mu|\mathbf{U})$. Because μ may unduly concentrate probability in portions of Ω to which $\mathbf{U}$ assigns little probability, $\mathbf{I}(\mu|\mathbf{U})$ may blow up enormously even if the assisted search that induces μ provides little help in locating T (as when q is very close to p). Thus, what's needed is a lower bound on $\mathbf{I}(\mu|\mathbf{U})$ for all probability measures μ such that $\mu(T) = q$. In other words, for fixed T and q, we need to evaluate

$$\inf\{\mathbf{I}(\mu|\mathbf{U}) : \mu(T) = q\}.$$

Fortunately, this is easily done. It turns out that the following probability measure, denoted by μ_0, is in this set and achieves the infimum:

$$d\mu_0 =_{def} [\frac{q}{p}1_T + \frac{1-q}{1-p}1_{T^c}]d\mathbf{U}.$$

Here 1_T is the indicator function for T and 1_{T^c} is the indicator function for the set-theoretic complement of T.

Let us call μ_0 the *canonical exchange probability* associated with an assisted search whose probability of successfully locating T is $1 - (1 - q)^m$. The reason this probability measure achieves the infimum is that $\mathbf{I}(\mu|\mathbf{U})$ is, in essence, a

variance (see Dembski 2004) and the probability density with respect to $\mathbf{U}$ that defines μ_0 (i.e., $\frac{q}{p}1_T + \frac{1-q}{1-p}1_{T^c}$) is the least squares solution for all such variances.

To see this, note first that μ_0 is indeed an exchange probability since

$$\mu_0(T) = \int_T (\frac{q}{p}1_T + \frac{1-q}{1-p}1_{T^c})d\mathbf{U} = \frac{q}{p}\mathbf{U}(T) = q.$$

In addition, note that

$$\int_T (\frac{q}{p}1_T + \frac{1-q}{1-p}1_{T^c})^2 d\mathbf{U} = \frac{q^2}{p} + \frac{(1-q)^2}{1-p}.$$

Next, consider an arbitrary exchange probability μ that is absolutely continuous with respect to $\mathbf{U}$ and for which

$$\mu(T) = \int_T \frac{d\mu}{d\mathbf{U}} = q.$$

In that case

$$
\begin{aligned}
\int_\Omega (\frac{q}{p}1_T + \frac{1-q}{1-p}1_{T^c} - \frac{d\mu}{d\mathbf{U}})^2 d\mathbf{U} \;\; &= \;\; \int_T (\frac{q}{p}1_T - \frac{d\mu}{d\mathbf{U}})^2 d\mathbf{U} \\
&\quad + \int_{T^c} (\frac{1-q}{1-p}1_{T^c} - \frac{d\mu}{d\mathbf{U}})^2 d\mathbf{U} \\
&= \;\; -\frac{q^2}{p} + \int_T (\frac{d\mu}{d\mathbf{U}})^2 d\mathbf{U} \\
&\quad - \frac{(1-q)^2}{1-p} + \int_{T^c} (\frac{d\mu}{d\mathbf{U}})^2 d\mathbf{U} \\
&= \;\; -\int_T (\frac{q}{p}1_T + \frac{1-q}{1-p}1_{T^c})^2 d\mathbf{U} + \int_\Omega (\frac{d\mu}{d\mathbf{U}})^2 d\mathbf{U} \\
&\geqslant \;\; 0,
\end{aligned}
$$

which implies that

$$\mathbf{I}(\mu|\mathbf{U}) = \log_2 \int_\Omega (\frac{d\mu}{d\mathbf{U}})^2 d\mathbf{U} \geqslant \log_2 \int_T (\frac{q}{p}1_T + \frac{1-q}{1-p}1_{T^c})^2 d\mathbf{U} = \mathbf{I}(\mu_0|\mathbf{U}).$$

It immediately follows $\{\mathbf{I}(\mu|\mathbf{U}) : \mu(T) = q\}$ attains its infimum at the canonical exchange probability μ_0.

$\mathbf{I}(\mu_0|\mathbf{U})$ evaluates as follows:

$$\mathbf{I}(\mu_0|\mathbf{U}) = \log_2 \left[\frac{q^2}{p} + \frac{(1-q)^2}{1-p} \right].$$

This last expression simplifies if p and q are small and p is much smaller than q (as is so often the case when searching for small targets in large spaces—typically q has order of magnitude $1/m$ and p is much, much smaller than q). Under these circumstances, we can drop the term $\frac{(1-q)^2}{1-p}$ within the logarithm (because it's now negligible) and write:

$$\mathbf{I}(\mu_0|\mathbf{U}) \approx \log_2 \frac{q^2}{p} = \log_2 \frac{1}{p} - 2\log_2 \frac{1}{q}.$$

Note that in the extreme case where the search is guaranteed to find a solution (i.e., $q = 1$), this last equation becomes exact again and is

$$\mathbf{I}(\mu_0|\mathbf{U}) = \log_2 \frac{1}{p},$$

which is just the old-fashioned, event-based information of the target T with respect to the uniform probability $\mathbf{U}$.

As an example to which we can put actual numbers for $\mathbf{I}(\mu_0|\mathbf{U})$, take the search for a given protein 100 amino acids in length (recall section 1). Call this protein the target T. There are roughly 10^{130} amino acid sequences of length 100. This space of sequences is the search space Ω. The uniform probability $\mathbf{U}$ on Ω assigns equal probability to each point in the space. Thus $\mathbf{U}(T) = p$ is roughly equal to 10^{-130}. The fastest supercomputer at the time of this writing maxes out at under 100 teraflops, which is 10^{14} floating point operations per second. Let us therefore imagine that this is the fastest rate at which points in Ω can be sampled and that points can be sampled for three years (which is just under 10^8 seconds). A conservative estimate on the number of possible proteins that can be sampled is therefore $m = 10^{14} \times 10^8 = 10^{22}$. If we now assume an assisted search has probability $1 - (1-q)^m$ of reaching the target and that q has order of magnitude roughly $1/m$ (as is typical in successful assisted searches), then

$$\mathbf{I}(\mu_0|\mathbf{U}) \approx \log_2 \frac{q^2}{p} = \log_2 \frac{1}{p} - 2\log_2 \frac{1}{q} \approx \log_2 10^{130} - 2\log_2 10^{22} \approx 286.$$

Accordingly, the assisted search here adds at least 286 bits of information, which for event-based information measures corresponds to an improbability of 10^{-86}. Note, the more effective the search, the greater the number of bits measured by $\mathbf{I}(\mu_0|\mathbf{U})$. Thus, when the search guarantees a solution (i.e., $q = 1$), $\mathbf{I}(\mu_0|\mathbf{U}) \approx \log_2 10^{130} \approx 432$, which is the maximal number of bits that an assisted search can produce through a canonical exchange probability relative to this target.

It follows that assisted search, even with so modest a problem as finding a specific protein 100 amino acids in length, requires a considerable amount of information if it is to surpass blind search and successfully locate a target. How are we to explain this net increase in information? One way is to explain it away by suggesting that no targets are in fact being searched. Rather, a space

of possibilities is merely being explored, and we, as pattern-seeking animals, are merely imposing patterns, and therefore targets, after the fact (see, for instance, Shermer 2003).

This explanation may work in certain instances where humans make up patterns as they go along. But many patterns—whether in the arts or in engineering or in the natural sciences—are objectively given. For instance, it is an objective fact whether a given polymer has a certain strength and resilience. Thus, searching through a polymer configuration space to find a polymer with at least that level of strength and resilience constitutes a search for an objectively given pattern qua target. If such a polymer is found and if the target within which it resides has small uniform probability, then a considerable amount of information needs to be incorporated in an assisted search for it to be successful, a fact that will be reflected in the information measure $\mathbf{I}$ as applied to the canonical exchange probability (i.e., $\mathbf{I}(\mu_0|\mathbf{U})$).

Besides explaining it away, there are two main options for explaining the net increase in information that an assisted search brings to an otherwise blind search. One is that an assisted search is intelligently designed by a purposive agent (cf. engineering). The other is that it is a fortuitous gift bestowed by an environment under the control of stochastic mechanisms (cf. evolutionary biology). I will argue that this latter option is inadequate and that the increase in information captured by $\mathbf{I}(\mu_0|\mathbf{U})$ is properly viewed as the result of a form of intelligence that cannot be reduced to stochastic mechanisms.

By intelligence, here, I mean something quite definite, namely, the causal factors that change one probability distribution into another and thus, in the present discussion, transform a blind search into an assisted search. A logically equivalent, information-theoretic reformulation of this definition takes intelligence as those causal factors that induce a net increase in information as measured by the information measure $\mathbf{I}$. Note that by a stochastic mechanism, here, I mean any causal process governed exclusively by the interplay between chance and necessity and characterized by unbroken deterministic and nondeterministic laws.

Intelligence acts by changing probabilities. Equivalently, intelligence acts by generating information. For instance, a slab of marble temporarily has a high probability of remaining unchanged. Then, without warning, Michelangelo decides to sculpt David, and the probability of that marble slab taking on a new *form* (i.e., receiving new in*form*ation) now changes dramatically.

This definition of intelligence as the causal factors responsible for changes in probabilities or, equivalently, for net increases in information is noncircular and, on reflection, should seem unproblematic. If there is a problem, it concerns whether intelligence is reducible to stochastic mechanisms. The neo-Darwinian theory of evolution, for instance, purports to account for biological complexity and diversity through an intelligence that is a stochastic mechanism, namely, the joint action of natural selection and random genetic mutations. To be sure, this mechanism operates in nature and is responsible for significant changes in the biological world. Nevertheless, is it the case that this mechanism accounts for biological complexity and diversity without remainder? In other words, is the

intelligence responsible for biological complexity and diversity entirely reducible to this mechanism?

In general, to justify the reduction of intelligence to stochastic mechanisms, these mechanisms need to supply a complete, self-consistent account of how changes in probability or net increases in information arise. As we shall see, the mathematics of blind and assisted searches precludes such an account, whether for neo-Darwinian assisted searches or for assisted searches in general.

5   Displacement

To see why intelligence, as defined in the last section, is not reducible to stochastic mechanisms in accounting for the probability change associated with an assisted search replacing a blind search, let us review where we are in the argument. We started with a large search space Ω that is a compact metric space under the metric D. D, in turn, induces a uniform probability $\mathbf{U}$. Our task is to find a target T in Ω, but $\mathbf{U}(T) = p$ is so small that blind search, as represented by uniform random sampling, is highly unlikely to locate T with any feasible sample size m. In other words, $1-(1-p)^m$, the probability of locating the target by uniform random sampling in m steps, is close to 0. We therefore look to an assisted search that induces a canonical exchange probability μ_0. $\mu_0(T) = q$, and the probability of the assisted search locating the target T in m steps is $1 - (1 - q)^m$. This last probability needs to be reasonably close to 1 (as is the case when q has order of magnitude at least $1/m$). When $1-(1-q)^m$ is close to 1, we have a satisfactory explanation for why the assisted search that induces μ_0 successfully locates T. The same cannot be said for blind search (qua uniform random sampling) since its probability of locating T in m steps is minuscule.

All this is unproblematic. But it leaves unanswered the precise causal factors responsible for the assisted search that induces μ_0. In general terms, μ_0 derives from an ambient environment (or context) in which the search for T is embedded. This environment is intelligent in that, as defined in the last section, it is capable of altering probabilities by replacing a blind search with an assisted search. But in this replacement, is such an environment entirely reducible to stochastic mechanisms? If so, then a stochastic mechanism must account for how μ_0 came to replace $\mathbf{U}$ in the search for T.

Here's how this would work: the original search space Ω has little structure; by itself, it cannot, except as a highly improbable event (improbability being gauged by $\mathbf{U}$), deliver a solution from the target T. Ω must therefore be embedded in a larger environment capable of delivering an assisted search that induces μ_0, which then in turn is capable of delivering a solution from the target T. But if this larger environment is driven by stochastic mechanisms and if a stochastic mechanism within Ω' is responsible for delivering μ_0, then μ_0 is itself the solution of a stochastically driven search.

What is this search for μ_0? To answer this question, let us first ask, What is the target of this new search? The original target was T, and the original search aimed to find some element in T (i.e., a solution in this target). But if μ_0, in

representing an assisted search that effectively locates the original target T, is a solution for some new search, to what new target qua solution space does μ_0 belong? Clearly, the solutions in this new target will need to comprise all other probability measures on Ω that represent assisted searches at least as effective as the search for T represented by μ_0: to omit any of these is to artificially constrict the new target, making unduly difficult the search for assisted searches that make the original search for T feasible; on the other hand, to include any more than these is to include assisted searches that are strictly less effective than the search represented by μ_0, and thus to open the new target to assisted searches that may be unlikely to locate the original target T. Such solutions would be no solutions at all.

Since any probability measure ν on Ω for which $\nu(T) \geqslant \mu_0(T)$ represents an assisted search at least as effective in locating T as the assisted search that induced μ_0, the new target is therefore properly defined as follows:

$$\overline{T} =_{def} \{\nu \in \mathbf{M}(\Omega) : \nu(T) \geqslant \mu_0(T)\}.$$

Here $\mathbf{M}(\Omega)$ denotes the set of all Borel probability measures on Ω. Note that any probability measure within $\overline{T}$ is at least as effective as μ_0 for locating T, whereas any probability measure outside will be strictly less effective than μ_0. Also, it follows immediately from the results of section 4 and from the fact that $\mu_0(T) = q$ that $\mathbf{I}(\mu_0|\mathbf{U}) = \inf\{\mathbf{I}(\nu|\mathbf{U}) : \nu(T) \geqslant q\} = \inf\{\mathbf{I}(\nu|\mathbf{U}) : \nu \in \overline{T}\}$. Thus, each element of $\overline{T}$ contributes at least as much information as μ_0 to an assisted search for T.

Given this characterization of $\overline{T}$, what is the search space in which the search for $\overline{T}$ takes place? $\overline{T}$, the new target being searched, is a set of probability measures. Moreover, since Ω has a topology induced by the metric D, those probability measures need to respect that topology. Probability measures that respect Ω's topology are those defined on the Borel sets of Ω. For this reason, $\overline{T}$ was explicitly defined in reference to $\mathbf{M}(\Omega)$, which comprises the probability measures defined on the Borel sets of Ω. Accordingly, the search for $\overline{T}$ is a search within $\mathbf{M}(\Omega)$. This latter space is the most parsimonious way to cash out the search space in which the search for $\overline{T}$ takes place.

To summarize, the problem we face is how to account for an assisted search that renders the probability of locating a target T highly probable. Initially, the problem was to find T in Ω using only Ω's metric structure D and the uniform probability $\mathbf{U}$ induced by D on Ω. Because this problem proved to be intractable for blind search (specifically, the problem was effectively unsolvable for uniform random sampling), an assisted search was required. The assisted search, represented by the canonical exchange probability μ_0, adequately explained finding a solution within T. Yet, procuring such an assisted search required its own search, namely, the search for $\overline{T} = \{\nu \in \mathbf{M}(\Omega) : \nu(T) \geqslant \mu_0(T)\}$ within $\mathbf{M}(\Omega)$. The initial search explored Ω to find T. The new search explores $\mathbf{M}(\Omega)$ to find $\overline{T}$. The problem of finding T has therefore been *displaced* to the new problem of finding $\overline{T}$. This is the *displacement problem*. As we shall see, the displacement problem precludes stochastic mechanisms from reductively explaining assisted

searches. Stochastic mechanisms may be involved, but they cannot be the whole story.

6   Higher-Level Search

How, then, shall we understand the search for $\overline{T}$ within $\mathbf{M}(\Omega)$? By analogy with the search for T in Ω, let us first determine which structures in $\mathbf{M}(\Omega)$ are relevant to the search for $\overline{T}$. As it is, geometric and measure-theoretic structures on Ω extend straightforwardly and canonically to corresponding structures on $\mathbf{M}(\Omega)$. For instance, take the metric D on Ω. Because D makes Ω a compact metric space, D a fortiori makes Ω a complete separable metric space. Now, most of the interesting mathematical work in probability theory focuses on separable metric spaces and, specifically, on separable topological spaces that can be metrized with a complete metric—these are known as *Polish* spaces (see Cohn 1996: ch. 8).

As it turns out, $\mathbf{M}(\Omega)$ is itself a separable metric space in the Kantorovich-Wasserstein metric $\overline{D}$, which induces the weak topology on $\mathbf{M}(\Omega)$. For Borel probability measures μ and v on Ω, this metric is defined as follows:

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

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

In the first equation here, $\mathbf{P}_2(\mu,\nu)$ 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 is ≤ 1 ($\|f\|_L = \sup\{|f(x) - f(y)|/D(x,y) : x,y \in \Omega, \ x \neq y\}$). Both the infimum and the supremum in these equations define metrics. The first is called the Wasserstein metric, the second the Kantorovich metric. Though the two expressions appear quite different, they are known to be equal (see Dudley 1976).

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

Perhaps the best way to see that $\overline{D}$ scrupulously extends the metric structure of Ω to $\mathbf{M}(\Omega)$ is to consider the following reformulation of this metric. Let $\mathbf{M}_{av}(\Omega) = \{\frac{1}{n}\sum_{1\leq i\leq n}\delta_{x_i} : x_i \in \Omega, \ n \text{ a positive integer}\}$. 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.

Now, for any measures μ and v 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

$$\overline{D}_{perm}(\tfrac{1}{n}\sum_{1\le i \le n}\delta_{x_i}, \tfrac{1}{n}\sum_{1\le i \le n}\delta_{y_i}) =_{def} \min\{\tfrac{1}{n}\sum_{1\le i \le n}D(x_i, y_{\sigma i}) : \sigma \in \mathbf{S}_n\}$$

where $\mathbf{S}_n$ is the symmetric group on the numbers 1 to n. $\overline{D}_{perm}$ 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 $\overline{D}_{perm}$ 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 $\frac{1}{n}\sum_{1\le i \le n}\delta_{x_i}$ and $\frac{1}{n}\sum_{1\le i \le n}\delta_{y_i}$ in $\mathbf{M}_{av}(\Omega)$, and for any measures $\frac{1}{mn}\sum_{1\le i \le mn}\delta_{z_i} = \frac{1}{n}\sum_{1\le i \le n}\delta_{x_i}$ and $\frac{1}{mn}\sum_{1\le i \le mn}\delta_{w_i} = \frac{1}{n}\sum_{1\le i \le n}\delta_{y_i}$,

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

This equality does in fact hold. Crucial in its proof is Philip Hall's well-known "marriage lemma" from combinatorial theory. Given this equality, it follows that $\overline{D}_{perm} = \overline{D}$ on $\mathbf{M}_{av}(\Omega)$ and, because $\mathbf{M}_{av}(\Omega)$ is dense in $\mathbf{M}(\Omega)$, that $\overline{D}_{perm}$ extends uniquely to $\overline{D}$ on all of $\mathbf{M}(\Omega)$ (for the proof see Dembski 2004). $\overline{D}_{perm}$, as a characterization of the Kantorovich-Wasserstein metric, will be important in the sequel.

Thus, given that D metrizes Ω, $\overline{D}$ is the canonical metric that metrizes $\mathbf{M}(\Omega)$. And, just as D induces a compact topology on Ω, so does $\overline{D}$ induce a compact topology on $\mathbf{M}(\Omega)$. This last result is a direct consequence of $\overline{D}$ 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 (Prohorov's theorem is actually somewhat stronger; see Billingsley 1999: 59).

Given that $\overline{D}$ makes $\mathbf{M}(\Omega)$ a compact metric space, the next question is whether this metric induces a uniform probability $\overline{\mathbf{U}}$ on $\mathbf{M}(\Omega)$. Accordingly, given that $\mathbf{M}^0(\Omega) = \Omega$, $\mathbf{M}^1(\Omega) = \mathbf{M}(\Omega)$, $\mathbf{M}^2(\Omega) = \mathbf{M}(\mathbf{M}(\Omega))$, and in general $\mathbf{M}^k(\Omega) = \mathbf{M}(\mathbf{M}^{k-1}(\Omega))$, it would follow that $\overline{\mathbf{U}}$ resides in $\mathbf{M}^2(\Omega)$. As it turns out, $\mathbf{M}(\Omega)$ is uniformizable with respect to $\overline{D}$. To see this, note that if

$$\mathbf{U}_\varepsilon = \tfrac{1}{n}\sum_{1\le i \le n}\delta_{x_i}$$

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., $\overline{D}(\mathbf{U}_\varepsilon, \mathbf{U}) \le \varepsilon$. And from this it follows that for $n^* = \binom{2n-1}{n} = \frac{(2n-1)!}{n!(n-1)!}$, $\{\theta_1, \theta_2, \ldots, \theta_{n^*}\} \subset \mathbf{M}(\Omega)$ is an $\frac{\varepsilon}{n}$-lattice as the θ_ks run through all finitely supported probability measures $\frac{1}{n}\sum_{1\le i \le n}\delta_{w_i}$ where the w_is are drawn from $\{x_1, x_2, \ldots, x_n\}$ allowing repetitions. Note that $\binom{2n-1}{n}$ is the number of ways of filling n slots with n identical items (see "orderings involving two kinds of elements" and "occupancy problems" in Feller 1968: ch. 1).

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

$$\overline{\mathbf{U}}_{\varepsilon} = \frac{1}{n^*} \sum\nolimits_{1 \leq k \leq n^*} \delta_{\theta_k}$$

converges to the uniform probability $\overline{\mathbf{U}}$ in the weak topology on $\mathbf{M}^2(\Omega)$. More-over, for $\overline{\overline{D}}$, which is the iterated Wasserstein-Kantorovich metric on $\mathbf{M}^2(\Omega)$ (note the double bars over D; $\mathbf{M}^2(\Omega)$ is likewise a compact metric space with respect to $\overline{\overline{D}}$), $\overline{\overline{D}}(\overline{\mathbf{U}}_{\varepsilon}, \overline{\mathbf{U}}) \leq \frac{\varepsilon}{n}$. The details here are all elementary combinatorics and unpacking the definition of uniform probability as given in Dembski (1990).

The search for $\overline{T}$ within $\mathbf{M}(\Omega)$ therefore precisely parallels the search for T within Ω. Just as the D induces a compact metric structure on Ω that in turn induces a uniform probability $\mathbf{U}$, so does the Kantorovich-Wasserstein metric $\overline{D}$ induce a compact metric structure on $\mathbf{M}(\Omega)$ that in turn induces a uniform probability $\overline{\mathbf{U}}$. The obvious question, now, is how $\overline{\mathbf{U}}(\overline{T})$ compares to $\mathbf{U}(T)$ and what that the comparison indicates about the search for $\overline{T}$ within $\mathbf{M}(\Omega)$ being either easier or more difficult than the search for T within Ω. We'll return to this question, but first we examine some deeper connections between Ω and $\mathbf{M}(\Omega)$.

7 Liftings, Lowerings, and No Free Lunch

The parallels between Ω and $\mathbf{M}(\Omega)$ don't stop with these spaces being compact metric spaces that support uniform probabilities. In addition, these spaces allow for considerable interaction through what may be called *liftings* and *lowerings*. Think of $\mathbf{M}(\Omega)$ as a higher-order space and Ω as a lower-order space with structures in the lower having corresponding structures in the higher and vice versa. Thus, structures in Ω may be *lifted* to structures in $\mathbf{M}(\Omega)$ and structures in $\mathbf{M}(\Omega)$ may correspondingly be *lowered* to structures in Ω.

Consider, for instance, the following. For a bounded continuous real-valued function f on Ω (actually, all continuous real-valued functions on Ω, $\mathbf{M}(\Omega)$, $\mathbf{M}^2(\Omega)$, etc. are bounded since these spaces are compact), we can define a (bounded) continuous real-valued function $\overline{f}$ on $\mathbf{M}(\Omega)$ as $\theta \mapsto \int_{\Omega} f(x)d\theta(x)$. Note that $\overline{f}$ is indeed continuous on $\mathbf{M}(\Omega)$ because weak convergence is defined in this space as θ_n converges to θ provided that for all bounded continuous real-valued functions h on Ω, $\int_{\Omega} h(x)d\theta_n(x)$ converges to $\int_{\Omega} h(x)d\theta(x)$. Note also that for $\theta = \delta_x$ (i.e., the point mass at x), $\overline{f}(\delta_x) = f(x)$. Call $\overline{f}$ the *lifting* of f from Ω to $\mathbf{M}(\Omega)$. Likewise, for F a continuous real-valued function on $\mathbf{M}(\Omega)$, define $\widetilde{F}$ on Ω as $x \mapsto \widetilde{F}(\delta_x)$. $\widetilde{F}$ is continuous on Ω. Call $\widetilde{F}$ the *lowering* of F from $\mathbf{M}(\Omega)$ to Ω. It then follows that $\widetilde{\overline{f}} = f$ though in general it need not be the case that $\overline{\widetilde{F}} = F$ (lowerings can lose information whereas liftings do not). Note that we can give up on continuity in defining liftings and lowerings. Thus, for instance, if f is a nonnegative measurable function on Ω, we can define $\overline{f}$ on $\mathbf{M}(\Omega)$ as before, namely $\overline{f}(\theta) = \int_{\Omega} f(x)d\theta(x)$. Lowerings can likewise be generalized to arbitrary measurable functions.

We now prove a result that is the key to lifting and lowering probability measures between $\mathbf{M}(\Omega)$ and $\mathbf{M}^2(\Omega)$:

THEOREM (NO FREE LUNCH). $\mathbf{U} = \int_{\mathbf{M}(\Omega)} \theta d\overline{\mathbf{U}}(\theta).$

REMARKS. The integral on the right side of this equation is vector-valued. Vector-valued integration has been well understood since the work of Gelfand (1936) and Pettis (1938) in the 1930s and has been widely applied since then (see Dinculeanu 2000). 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 Ω) equals integrating over the continuous linear functions applied inside the integral. Essentially, linear functionals reduce vector-valued integration to ordinary integration. Thus, the equality in the statement of this theorem means that for all continuous real-valued f on Ω,

$$\int_{\Omega} f(x) d\mathbf{U}(x) = \int_{\mathbf{M}(\Omega)} \left[\int_{\Omega} f(x) d\theta(x) \right] d\overline{\mathbf{U}}(\theta).$$

Because of the compactness of Ω and $\mathbf{M}(\Omega)$, the existence and uniqueness of $\int_{\mathbf{M}(\Omega)} \theta d\overline{\mathbf{U}}(\theta)$ is not a problem.

According to the equation $\mathbf{U} = \int_{\mathbf{M}(\Omega)} \theta d\overline{\mathbf{U}}(\theta)$, averaging all probability measures on $\mathbf{M}(\Omega)$ with respect to the uniform probability $\overline{\mathbf{U}}$ is just the uniform probability $\mathbf{U}$ on Ω. This, in measure-theoretic terms, restates the No Free Lunch theorems of Wolpert and Macready (1997), which say that when averaged over all fitness functions (whether time-dependent or time-independent fitness functions), no evolutionary search procedure outperforms any other. Thus, in particular, these searches, when averaged, do not outperform blind search. If we now think of θ under the integral in $\int_{\mathbf{M}(\Omega)} \theta d\overline{\mathbf{U}}(\theta)$ as an exchange probability for an assisted search, this formulation of the No Free Lunch theorem says that the average performance of all assisted searches is no better than uniform random sampling, which throughout this paper epitomizes blind search.

PROOF. We prove the theorem in the case where Ω is finite (the infinite case is proven by considering finitely supported uniform probabilities on lattices, and then taking the limit as the mesh of these lattices goes to zero—see Dembski 1990). Suppose, therefore, that $\Omega = \{x_1, x_2, \ldots, x_n\}$. For large N, consider all probabilities of the form

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

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 (see Feller 1968: 38). Therefore, define

$$\overline{\mathbf{U}}_N =_{def} \frac{1}{N^*} \sum_{1 \leq k \leq N^*} \delta_{\theta_k}$$

so that the θ_ks run through all such θ. As we saw in the last section, $\overline{\mathbf{U}}_N$ converges in the weak topology to $\overline{\mathbf{U}}$ as $N \uparrow \infty$.

It's enough, therefore, to show that

$$\int_{\mathbf{M}(\Omega)} \theta d\overline{\mathbf{U}}_N(\theta) = \frac{1}{N^*} \sum_{1 \leq k \leq N^*} \theta_k$$

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

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

For definiteness, let us consider x_1. We can think of x_1 as being 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 θ_ks with weight M/N at x_1 is

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

The term $\binom{N-M+n-2}{n-2}$ is the number of ways of occupying $n-1$ slots with $N-M$ identical items (see Feller 1968: 38). Accordingly, the total weight that the θ_ks assign to x_1 when normalized by $1/N^*$ is

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

This last expression is messy and difficult to evaluate directly. But it does not need to be evaluated directly. Because this expression is independent x_1 and is also the probability of x_1, it follows that the probability of all x_is under $\frac{1}{N^*} \sum_{1 \leq k \leq N^*} \theta_k$ is the same. In other words, for each x_i in Ω,

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

This is what needed to be proved. ∎

COROLLARY. Suppose μ is a probability measure on Ω that is absolutely continuous with respect to $\mathbf{U}$. Let $\frac{d\mu}{d\mathbf{U}}$ denote the Radon-Nikodym derivative of μ with respect to $\mathbf{U}$ and let $\overline{\frac{d\mu}{d\mathbf{U}}}$ denote its lifting. If we now define the lifting of μ as $\overline{\mu} = \overline{\frac{d\mu}{d\mathbf{U}}} d\overline{\mathbf{U}}$, then $\overline{\mu}$ is a probability measure on $\mathbf{M}(\Omega)$. Moreover, since $\mathbf{U}$ is absolutely continuous with itself so that $\frac{d\mathbf{U}}{d\mathbf{U}}$ is identically equal to 1 on Ω, it follows that the lifting of $\frac{d\mathbf{U}}{d\mathbf{U}}$, i.e., $\overline{\frac{d\mathbf{U}}{d\mathbf{U}}}$, is identically equal to 1 on $\mathbf{M}(\Omega)$, and thus the lifting of $\mathbf{U}$, as just defined, is in fact the uniform probability on $\mathbf{M}(\Omega)$. Thus, whether we interpret $\overline{\mathbf{U}}$ as the uniform probability on $\mathbf{M}(\Omega)$ in the sense of Dembski (1990) or as the lifting of the uniform probability $\mathbf{U}$ on Ω, both signify the same probability measure on $\mathbf{M}(\Omega)$.

PROOF. It is enough to see that $\overline{\mu}$ is indeed a probability measure on $\mathbf{M}(\Omega)$, and for this it is enough to see that

$$\int_{\mathbf{M}(\Omega)} \overline{\frac{d\mu}{d\mathbf{U}}}(\theta) d\overline{\mathbf{U}}(\theta) = \int_{\mathbf{M}(\Omega)} \left[\int_{\Omega} \frac{d\mu}{d\mathbf{U}}(x) d\theta(x) \right] d\overline{\mathbf{U}}(\theta)$$

$$= \int_{\Omega} \frac{d\mu}{d\mathbf{U}}(x) d \left[\int_{\mathbf{M}(\Omega)} \theta d\overline{\mathbf{U}}(\theta) \right] (x)$$

$$= \int_{\Omega} \frac{d\mu}{d\mathbf{U}}(x) d\mathbf{U}(x) \quad \text{[by the NFL theorem]}$$

$$= \int_{\Omega} d\mu$$

$$= 1. \quad \blacksquare$$

By analogy with the No Free Lunch theorem, it is tempting to think that for μ absolutely continuous with respect to $\mathbf{U}$,

$$\mu = \int_{\mathbf{M}(\Omega)} \theta d\overline{\mu}(\theta).$$

This equality, however, does not hold. To see this, suppose that μ is absolutely continuous with respect to $\mathbf{U}$ and that for some set A in Ω, $\mu(A) = 0$ but $\mathbf{U}(A) > 0$. In that case

$$\left[\int_{\mathbf{M}(\Omega)} \theta d\overline{\mu}(\theta) \right] (A) = \int_{\Omega} 1_A(x) d \left[\int_{\mathbf{M}(\Omega)} \theta d\overline{\mu}(\theta) \right] (x)$$

$$= \int_{\Omega} 1_A(x) d \left[\int_{\mathbf{M}(\Omega)} \theta \overline{\frac{d\mu}{d\mathbf{U}}}(\theta) d\overline{\mathbf{U}}(\theta) \right] (x)$$

$$= \int_{\mathbf{M}(\Omega)} \theta(A) \left[\int_{\Omega} \frac{d\mu}{d\mathbf{U}} d\theta \right] d\overline{\mathbf{U}}(\theta)$$

$$> 0.$$

This last inequality holds because the set of θs in $\mathbf{M}(\Omega)$ for which both $\theta(A)$ and $\theta(A)$ are both strictly positive has nonzero measure under $\overline{\mathbf{U}}$, thus making the integrand $\theta(A) \left[\int_{\Omega} \frac{d\mu}{d\mathbf{U}} d\theta \right]$ strictly greater than zero on a set of positive $\overline{\mathbf{U}}$ probability.

For completeness, given a probability measure Θ on $\mathbf{M}(\Omega)$ that is absolutely continuous with respect to $\overline{\mathbf{U}}$, we define its lowering $\widetilde{\Theta}$ as follows: since Θ is absolutely continuous with respect to $\overline{\mathbf{U}}$, take its Radon-Nikodym derivative $\frac{d\Theta}{d\overline{\mathbf{U}}}$ on $\mathbf{M}(\Omega)$ and lower it to $\widetilde{\frac{d\Theta}{d\mathbf{U}}}$ on Ω as defined at the start of this section. The

lowering of Θ is then defined as the measure $\widetilde{\frac{d\Theta}{d\mathbf{U}}}d\mathbf{U}$ on Ω. Note that this need not be a probability measure since $\overline{\Omega} =_{def} \{\delta_x \in \mathbf{M}(\Omega) : x \in \Omega\}$ defines a set of $\overline{\mathbf{U}}$-probability zero in $\mathbf{M}(\Omega)$, allowing $\frac{d\Theta}{d\mathbf{U}}$ to be arbitrarily defined on that set. Nonetheless, if μ on Ω not only is absolutely continuous with respect to $\mathbf{U}$ but also has a continuous Radon-Nikodym derivative $\frac{d\mu}{d\mathbf{U}}$ on Ω, then $\overline{\frac{d\mu}{d\mathbf{U}}}$ is itself continuous on $\mathbf{M}(\Omega)$, and $\widetilde{\frac{d\mu}{d\mathbf{U}}} = \frac{d\mu}{d\mathbf{U}}$, implying that $\widetilde{\overline{\mu}} = \mu$.

8   The Displacement Theorem

THEOREM (DISPLACEMENT THEOREM). Suppose that Ω has uniform probability $\mathbf{U}$ and that $\mathbf{M}(\Omega)$ has uniform probability $\overline{\mathbf{U}}$. Let T be a target in Ω such that $\mathbf{U}(T) = p \ (> 0)$ and let μ be an exchange probability representing an assisted search for T such that $\mu(T) = q \ (> p)$. Define $\overline{T} = \{\nu \in \mathbf{M}(\Omega) : \nu(T) \geqslant q\}$. Then, if $\Omega = \{x_1, x_2, \ldots, x_K\}$ is finite with K elements such that $T = \{x_1, x_2, \ldots, x_{K_1}\}$ for $K_1 = Kp$ (i.e., $p = K_1/K$),

$$
\begin{aligned}
\overline{\mathbf{U}}(\overline{T}) \;\;&= \;\; \frac{\Gamma(K)}{\Gamma(K(1-p))\Gamma(Kp)} \int\limits_0^{1-q} t^{K(1-p)-1}(1-t)^{Kp-1}dt \\[2mm]
&< \;\; \sqrt{K} \cdot \frac{\sqrt{p}}{q} \cdot \left[\left(\frac{1-q}{1-p}\right)^{1-p} \left(\frac{q}{p}\right)^p \right]^K \quad \text{for } K \geqslant \frac{2q-1}{q-p}. \\[2mm]
&< \;\; \sqrt{K} \cdot \frac{\sqrt{p}}{q} \cdot \left[1 - r^2\right]^K \quad \text{for } r = q - p \text{ and } K \geqslant \frac{2q-1}{q-p}.
\end{aligned}
$$

Because both these last expressions are exponential in K and because both the terms in brackets are strictly less than 1 for $q > p$, it follows that the last two expressions converge to zero as $K \uparrow \infty$. Thus, for infinite Ω and nonatomic $\mathbf{U}$, $\overline{\mathbf{U}}(\overline{T}) = 0$.

REMARKS. Here Γ is the gamma function, i.e., $\Gamma(x) = \int_0^\infty e^{-t}t^{x-1}dt$ for x positive (recall that for positive integers n, $\Gamma(n) = (n-1)!$). For x and y positive, the beta function is defined as follows:

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

It follows that for finite Ω, $\overline{\mathbf{U}}(\overline{T})$ is given by a cumulative beta distribution.

Although there is no closed expression for $\overline{\mathbf{U}}(\overline{T})$ in terms of p, q, and K, in some cases $\overline{\mathbf{U}}(\overline{T})$ can be readily evaluated. Recall finding a protein 100 amino acids in length (section 6). Here we set $p = 10^{-130}$, $q = 10^{-22}$, $K = 10^{130}$, and $K_1 = 1$. In that case

$$\overline{\mathbf{U}}(\overline{T}) \;=\; \frac{\Gamma(K)}{\Gamma(K(1-p))\Gamma(Kp)} \int\limits_{0}^{1-q} t^{K(1-p)-1}(1-t)^{Kp-1}dt$$

$$=\; \frac{\Gamma(10^{130})}{\Gamma(10^{130}-1)\Gamma(1)} \int\limits_{0}^{1-10^{-22}} t^{10^{130}-2}(1-t)^{0}dt$$

$$=\; (10^{130}-1)\cdot \frac{t^{10^{130}-1}}{10^{130}-1}\Big|_{0}^{1-10^{-22}}$$

$$=\; (1-10^{-22})^{10^{130}-1}$$

$$=\; (1-\frac{10^{108}}{10^{130}})^{10^{130}-1}$$

$$\approx\; e^{-10^{108}}.$$

Note that since $K = 10^{130} = 1/p$, plugging numbers into the upper bound estimate $\sqrt{K}\cdot\frac{\sqrt{p}}{q}\cdot\left[\left(\frac{1-q}{1-p}\right)^{1-p}\left(\frac{q}{p}\right)^{p}\right]^{K}$ yields:

$$\sqrt{K}\cdot\frac{\sqrt{p}}{q}\cdot\left(\frac{1-q}{1-p}\right)^{K-1}\cdot\left(\frac{q}{p}\right) \;=\; 10^{65}\cdot 10^{-43}\cdot 10^{108}\left(\frac{1-10^{-22}}{1-10^{-130}}\right)^{10^{130}-1}$$

$$\leqslant\; 10^{130}\cdot\left(\frac{1-10^{-22}}{1-10^{-130}}\right)^{10^{130}}$$

$$\approx\; 10^{130}\frac{e^{-10^{108}}}{e^{-1}}$$

$$=\; e^{-10^{108}+1+130\ln 10}$$

$$<\; e^{-10^{108}+301}$$

It follows not just in this instance that the higher-order probability $\overline{\mathbf{U}}(\overline{T})$ is exponential in the lower-order probability $\mathbf{U}(T)$, but that this result holds quite generally since, as K increases but p and q remain fixed, this bound will decrease all the more. For instance, substituting the multiple nK for K in this bound yields

$$\sqrt{nK}\cdot\frac{\sqrt{p}}{q}\cdot\left[\left(\frac{1-q}{1-p}\right)^{1-p}\left(\frac{q}{p}\right)^{p}\right]^{nK} \;=\; \sqrt{nK}\cdot\frac{\sqrt{p}}{q}\cdot\left(\frac{1-q}{1-p}\right)^{n(K-1)}\cdot\left(\frac{q}{p}\right)^{n}$$

$$=\; \sqrt{n}\cdot 10^{65-43+108n}\cdot\left(\frac{1-10^{-22}}{1-10^{-130}}\right)^{n(10^{130}-1)}$$

$$\leqslant\; \sqrt{n}\cdot 10^{22+108n}\cdot\left(\frac{1-10^{-22}}{1-10^{-130}}\right)^{n10^{130}}$$

$$\approx\; e^{\ln\sqrt{n}}\cdot e^{\ln 10(108n+22)}\cdot e^{-(10^{108}\cdot n)+n}$$

$$\leqslant\; e^{-10^{108}n+250n+\ln\sqrt{n}+51}.$$

26

Not only is this bound exponential in n, but the term $-10^{108}n$ completely dominates the exponent.

Accordingly, this inequality allows maximal tolerance when the cardinality of the underlying search space Ω equals $1/p$ and gets tighter as Ω gets refined. In particular, refining the probability space Ω only makes the lifted probability $\overline{\mathbf{U}}(\overline{T})$ worse (i.e., smaller still). In the limit, as Ω becomes infinite and $\mathbf{U}$ constitutes a nonatomic uniform probability on Ω, $\overline{\mathbf{U}}(\overline{T}) = 0$. By definition, $\mathbf{U}$ is nonatomic iff for every x in Ω, $\mathbf{U}(\{x\}) = 0$ (see Parthasarathy 1967: 53–55). This assumption is necessary for the Displacement Theorem to show that $\overline{\mathbf{U}}(\overline{T}) = 0$ because it assures that as K increases and as finitely supported uniform probabilities converge weakly to $\mathbf{U}$, those probabilities, when applied to T, are converging to $\mathbf{U}(T)$ (cf. conditionalization of uniform probabilities on subspaces in Dembski 1990).

PROOF. We start with the finite case: $\Omega = \{x_1, x_2, \ldots, x_K\}$, $T = \{x_1, x_2, \ldots, x_{K_1}\}$, $1 \leqslant K_1 < K$, and $p = K_1/K$. For a given N, consider all probabilities $\theta = \frac{1}{N} \sum_{1 \leq i \leq N} \delta_{w_i}$ where the w_is are drawn from $\{x_1, x_2, \ldots, x_K\}$ allowing repetitions. By elementary combinatorics (recall the proof of the No Free Lunch Theorem—note the slight change in notation), there are $N^* = \binom{N+K-1}{K-1} = \frac{(N+K-1)!}{N!(K-1)!}$ such probabilities. Accordingly, define $\overline{\mathbf{U}}_N = \frac{1}{N^*} \sum_{1 \leq i \leq N^*} \delta_{\theta_i}$ where the θ_is range over these θs. $\overline{\mathbf{U}}_N$ then converges weakly to $\mathbf{U}$.

Next, consider the probabilities $\theta = \frac{1}{N} \sum_{1 \leq i \leq N} \delta_{w_i}$ where the w_is are drawn from $\{x_1, x_2, \ldots, x_K\}$ allowing repetitions but also for which the number of w_is among $T = \{x_1, x_2, \ldots, x_{K_1}\}$ is at least $\lfloor Nq \rfloor$ (here $\lfloor x \rfloor$ denotes the greatest integer contained in x; thus $\lfloor 5.4 \rfloor = 5$). If we let Q_N denote the set of all such finitely supported probabilities that assign probability at least q to T, and if we let $K_2 = K - K_1$ and $q(N) = N - \lfloor Nq \rfloor$, then by elementary combinatorics Q_N has the following number of elements:

$$|Q_N| = \sum_{j=0}^{q(N)} \binom{N - j + K_1 - 1}{K_1 - 1} \binom{j + K_2 - 1}{K_2 - 1}.$$

Accordingly, it follows that

$$\overline{\mathbf{U}}(\overline{T}) = \lim_{N \to \infty} \frac{|Q_N|}{N^*} = \lim_{N \to \infty} \frac{\sum_{j=0}^{q(N)} \binom{N-j+K_1-1}{K_1-1}\binom{j+K_2-1}{K_2-1}}{\binom{N+K-1}{K-1}} \tag{*}$$
.

This last limit simplifies. Note first that

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

implying that $\binom{N+K-1}{K-1} \sim \frac{N^{K-1}}{(K-1)!}$, i.e.,

$$\lim_{N \to \infty} \frac{\binom{N+K-1}{K-1}}{N^{K-1}/(K-1)!} = 1.$$

Similarly, note that $\binom{N-j+K_1-1}{K_1-1} = \frac{(N-j)^{K_1-1}+(\text{lower order terms in } N-j)}{(K_1-1)!}$ and that $\binom{j+K_2-1}{K_2-1} = \frac{j^{K_2-1}+(\text{lower order terms in } j)}{(K_2-1)!}$ so that

$$\binom{N-j+K_1-1}{K_1-1} \sim \frac{(N-j)^{K_1-1}}{(K_1-1)!} \text{ and } \binom{j+K_2-1}{K_2-1} \sim \frac{j^{K_2-1}}{(K_2-1)!}.$$

Accordingly, there is no problem substituting $\frac{N^{K-1}}{(K-1)!}$ for $\binom{N+K-1}{K-1}$ in the denominator of (*). In the numerator of (*), the corresponding substitution does not follow immediately because the limits of summation depend on N via $q(N)$. Nonetheless, because in general $\sum_{j=0}^{M} j^n = \frac{M^{n+1}}{n+1} + (\text{lower order terms}$ in $M)$, the effect of summing lower order terms up to $q(N)$ is still swamped by dividing out by $\binom{N+K-1}{K-1} \sim \frac{N^{K-1}}{(K-1)!}$. More precisely, the lower-order terms in N and j from the numerator have exponents that sum to no more than $(K_1 - 1) + (K_2 - 1) - 1 = K - 3$. Summing from zero to $q(N)$ adds 1 to this exponent, making it $K - 2$. Therefore, dividing by $\frac{N^{K-1}}{(K-1)!}$, where the exponent is higher still, takes to zero these lower-order terms as $N \longrightarrow \infty$.

We thus have the following simplification:

$$\overline{\mathbf{U}}(\overline{T}) = \lim_{N \to \infty} \frac{|Q_N|}{N^*} = \lim_{N \to \infty} \frac{\sum_{j=0}^{q(N)} \frac{(N-j)^{K_1-1}}{(K_1-1)!} \frac{j^{K_2-1}}{(K_2-1)!}}{\frac{N^{K-1}}{(K-1)!}} \qquad (**)$$

This last limit can now be rewritten as follows:

$$\lim_{N \to \infty} \frac{\sum_{j=0}^{q(N)} \frac{(N-j)^{K_1-1}}{(K_1-1)!} \frac{(j)^{K_2-1}}{(K_2-1)!}}{\frac{N^{K-1}}{(K-1)!}}$$

$$= \lim_{N \to \infty} \frac{(K-1)!}{(K_1-1)!(K_2-1)!} \sum_{j=0}^{q(N)} \frac{1}{N} \frac{(N-j)^{K_1-1}}{N^{K_1-1}} \frac{j^{K_2-1}}{N^{K_2-1}}$$

$$= \frac{\Gamma(K)}{\Gamma(K_1)\Gamma(K_2)} \lim_{N \to \infty} \sum_{j=0}^{q(N)} \frac{1}{N}(1 - \frac{j}{N})^{K_1-1}(\frac{j}{N})^{K_2-1}$$

$$= \frac{\Gamma(K)}{\Gamma(Kp)\Gamma(K(1-p))} \int_0^{1-q} (1-t)^{Kp-1} t^{K(1-p)-1} dt$$

$$= \frac{\Gamma(K)}{\Gamma(K(1-p))\Gamma(Kp)} \int_0^{1-q} t^{K(1-p)-1}(1-t)^{Kp-1} dt.$$

This proves our main result, which is that

$$\overline{\mathbf{U}}(\overline{T}) = \frac{\Gamma(K)}{\Gamma(K(1-p))\Gamma(Kp)} \int_0^{1-q} t^{K(1-p)-1}(1-t)^{Kp-1}dt.$$

In this last expression, let us use Stirling's exact formula to calculate the factor in front of the integral (without loss of generality assume p is rational and K such that Kp is an integer). 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 $n! = \sqrt{2\pi}n^{n+1/2}e^{-n+\varepsilon(n)/(12n)}$ (Spivak 1980: 543). It now follows that

$\vdots$

$$\frac{\Gamma(K)}{\Gamma(K(1-p))\Gamma(Kp)}$$

$$= \frac{K!}{(K(1-p))!(Kp)!} \cdot \frac{K(1-p)Kp}{K}$$

$$= Kp(1-p) \cdot \frac{\sqrt{2\pi}K^{K+1/2}e^{-K+\varepsilon(K)/(12K)}}{\sqrt{2\pi}(Kp)^{Kp+1/2}e^{-Kp+\varepsilon(Kp)/(12Kp)}} \cdot$$

$$\frac{1}{\sqrt{2\pi}(K(1-p))^{K(1-p)+1/2}e^{-K(1-p)+\varepsilon(K(1-p))/(12K(1-p))}}$$

$$= \sqrt{\frac{Kp(1-p)}{2\pi}} \cdot \left((1-p)^{(1-p)}p^p\right)^{-K} \cdot e^{\frac{\varepsilon(K)}{12K} - \frac{\varepsilon(Kp)}{12Kp} - \frac{\varepsilon(K(1-p))}{12K(1-p)}}$$

$$\leqslant \sqrt{\frac{Kp(1-p)}{2\pi}} \cdot e^{\frac{1}{12}} \cdot \left((1-p)^{(1-p)}p^p\right)^{-K}$$

$$< \sqrt{Kp(1-p)} \cdot \left((1-p)^{(1-p)}p^p\right)^{-K} \quad \text{for all } K.$$

Moreover, the integral in this expression can be bounded as follows for large K:

$$\int_0^{1-q} t^{K(1-p)-1}(1-t)^{Kp-1}dt \leqslant (1-q)(1-q)^{K(1-p)-1}q^{Kp-1}$$

$$= (1-q)^{K(1-p)}q^{Kp-1}$$

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 to $t = 1$, 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 $m = K(1-p)-1$ and $n = Kp - 1$, the integrand in the preceding integral increases from 0 to

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

Since $p < q$ and therefore $1-p > 1-q$, elementary manipulations show that this cutoff is at least $1 - q$ whenever $K \geqslant \frac{2q-1}{q-p}$ (which is automatic if $q < \frac{1}{2}$ since then the right side is negative). Thus, if $K \geqslant \frac{2q-1}{q-p}$, when integrated over the interval $[0, 1-q]$, the integrand reaches its maximum at $1-q$. 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,

$$\frac{\Gamma(K)}{\Gamma(K(1-p))\Gamma(Kp)} \int_0^{1-q} t^{K(1-p)-1}(1-t)^{Kp-1}dt$$

$$< \quad \sqrt{Kp(1-p)} \cdot \left((1-p)^{(1-p)}p^p\right)^{-K} \cdot (1-q)^{K(1-p)}q^{Kp-1}$$

$$= \quad \sqrt{\frac{Kp(1-p)}{q^2}} \left((1-p)^{(1-p)}p^p\right)^{-K} \cdot ((1-q)^{(1-p)}q^p)^K$$

$$= \quad \sqrt{\frac{Kp(1-p)}{q^2}} \left[\left(\frac{1-q}{1-p}\right)^{1-p}\left(\frac{q}{p}\right)^p\right]^K$$

$$< \quad \sqrt{K} \cdot \frac{\sqrt{p}}{q} \cdot \left[\left(\frac{1-q}{1-p}\right)^{1-p}\left(\frac{q}{p}\right)^p\right]^K \quad \text{for } K \geqslant \frac{2q-1}{q-p}.$$

This last expression is exponential in K, with its limit going to zero as $K \uparrow \infty$. This is because the term in brackets, to which the power of K is taken, is strictly less than 1 whenever $q > p$. To see this, note first that in general for $0 < a < b$ and for $r > -a$, $(1+\frac{r}{a})^a < (1+\frac{r}{b})^b$ (Hardy et al. 1952: 37). If we now write $r = q - p$, which is greater than zero, the term in brackets can be rewritten as

$$\left(1 + \frac{-r}{1-p}\right)^{1-p}\left(1 + \frac{r}{p}\right)^p,$$

which by a double application of the previous inequality (choosing b in each case to be 1) is therefore strictly less than $(1-r)(1+r) = 1 - r^2 < 1$. Hence the final inequality in the statement of the theorem. ∎

9   Conclusion: The No Free Lunch Regress

What is the significance of the Displacement Theorem? It is this. Blind search for small targets in large spaces is highly unlikely to succeed. For a search to succeed, it therefore needs to be an assisted search. Such a search, however, resides in a target of its own. And a blind search for this new target is even less

likely to succeed than a blind search for the original target (the Displacement Theorem puts precise numbers to this). Of course, this new target can be successfully searched by replacing blind search with a new assisted search. But this new assisted search for this new target resides in a still higher-order search space, which is then subject to another blind search, more difficult than all those that preceded it, and in need of being replaced by still another assisted search. And so on. This regress, which I call the No Free Lunch Regress, is the upshot of this paper. It shows that stochastic mechanisms cannot explain the success of assisted searches.

This last statement contains an intentional ambiguity. In one sense, stochastic mechanisms fully explain the success of assisted searches because these searches themselves constitute stochastic mechanisms that, with high probability, locate small targets in large search spaces. Yet, in another sense, for stochastic mechanisms to explain the success of assisted searches means that such mechanisms have to explain how those assisted searches, which are so effective at locating small targets in large spaces, themselves arose with high probability. It's in this latter sense that the No Free Lunch Regress asserts that stochastic mechanisms cannot explain the success of assisted searches.

To appreciate the significance of the No Free Lunch Regress in this latter sense, consider the case of evolutionary biology. Evolutionary biology holds that various (stochastic) evolutionary mechanisms operating in nature facilitate the formation of biological structures and functions. These include preeminently the Darwinian mechanism of natural selection and random variation, but also others (e.g., genetic drift, lateral gene transfer, and symbiogenesis). There is a growing debate whether the mechanisms currently proposed by evolutionary biology are adequate to account for biological structures and functions (see, for example, Depew and Weber 1995, Behe 1996, and Dembski and Ruse 2004). Suppose they are. Suppose the evolutionary searches taking place in the biological world are highly effective assisted searches qua stochastic mechanisms that successfully locate biological structures and functions. Regardless, that success says nothing about whether stochastic mechanisms are in turn responsible for bringing about those assisted searches.

Evolving biological systems invariably reside in larger environments that subsume the search space in which those systems evolve. Moreover, these larger environments are capable of dramatically changing the probabilities associated with evolution as occurring in those search spaces. Take an evolving protein or an evolving strand of DNA. The search spaces for these are quite simple, comprising sequences that at each position select respectively from either twenty amino acids or four nucleotide bases. But these search spaces embed in incredibly complex cellular contexts. And the cells that supply these contexts themselves reside in still higher-level environments.

As a consequence, the uniform probability on the search space almost never characterizes the system's evolution. Rather, according to evolutionary biology, the larger environment bestows a nonuniform probability that brings the search (i.e., an assisted search) to a successful conclusion. This, in a nutshell, was Richard Dawkins's (1996) argument in *Climbing Mount Improbable*: biological

structures that at first blush seem vastly improbable (i.e., with respect to the uniform probability, blind search, pure randomness, call it what you will) become quite probable once the appropriate evolutionary mechanisms are factored in to reset the probabilities.

Even if we accept the full efficacy of evolutionary mechanisms to evolve biological structures and functions, the challenge that displacement poses to evolutionary biology still stands. A larger environment bestows a nonuniform probability qua assisted search. Fine. Presumably this nonuniform probability, which is defined over the search space in question, splinters off from richer probabilistic structures defined over the larger environment. We can, for instance, imagine the search space being embedded in the larger environment, and such richer probabilistic structures inducing a nonuniform probability (qua assisted search) on this search space, perhaps by conditioning on a subspace or by factorizing a product space. But, if the larger environment is capable of inducing such probabilities, what exactly are the structures of the larger environment that endow it with this capacity? Are any canonical probabilities defined over this larger environment (e.g., a uniform probability)? Do any of these higher-level probabilities induce the nonuniform probability that characterizes effective search of the original search space? What stochastic mechanisms might induce such higher-level probabilities?

For any interesting instances of biological evolution, we don't know the answer to these questions. But suppose we could answer these questions. As soon as we could, the No Free Lunch Regress would kick in, applying to the larger environment once its probabilistic structure becomes evident. And so, this probabilistic structure would itself require explanation in terms of stochastic mechanisms. On the other hand, lacking answers to these questions, we lack a stochastic mechanism to explain the nonuniform probabilities (and corresponding assisted searches) that the larger environment is supposed to induce and that makes effective search of the original space possible. In either case, the No Free Lunch Regress blocks our attempts to account for assisted searches in terms of stochastic mechanisms.

Evolutionary biologists at this point sometimes object that evolutionary mechanisms like Darwinian natural selection are indeed a free lunch because they are so simple, generating, as Richard Dawkins (1987: 316) puts it, biological complexity out of "primeval simplicity." But ascribing simplicity to these mechanisms betrays wishful thinking. The information that assisted searches bring to otherwise blind searches is measurable and substantial, and discloses an underlying complexity (see section 4). Just because it's possible to describe the mechanism that assists a search in simple terms does not mean that the mechanism, as actually operating in nature and subject to countless contingencies (Michael Polanyi called them boundary conditions), is in fact simple.

A final question therefore presents itself, namely, Is it even reasonable, whether in biology or elsewhere, to think that the assisted searches that successfully locate small targets in large spaces should be conceived as purely the result of stochastic mechanisms? What if, additionally, they inevitably result from a form of intelligence that is not reducible to stochastic mechanisms—a form of

intelligence that transcends chance and necessity? The No Free Lunch Regress, by demonstrating the incompleteness of stochastic mechanisms to explain assisted searches, fundamentally challenges the materialist dogma that reduces all intelligence to chance and necessity.

References

[1] Bauer, Heinz, *Probability Theory and Elements of Measure Theory*, trans. R. B. Burckel (London: Academic Press, 1981).

[2] Behe, Michael J., *Darwin's Black Box: The Biochemical Challenge to Evolution* (New York: Free Press, 1996).

[3] Billingsley, Patrick, *Convergence of Probability Measures*, 2nd ed. (New York: Wiley, 1999).

[4] Cohn, Donald L., *Measure Theory* (Boston: Birkhäuser, 1996).

[5] Collins, Robin, "Evidence of Fine-Tuning," in N. A. Manson, ed., *God and Design: The Teleological Argument and Modern Science*, 178–199 (London: Routledge, 2003).

[6] Cover, Thomas M., and Joy A. Thomas, *Elements of Information Theory* (New York: Wiley, 1991).

[7] Culberson, Joseph C., "On the Futility of Blind Search: An Algorithmic View of 'No Free Lunch'," *Evolutionary Computation* 6(2) (1998): 109–127.

[8] Dawkins, Richard, *The Blind Watchmaker: Why the Evidence of Evolution Reveals a University without Design* (New York: Norton, 1987).

[9] Dawkins, Richard, *Climbing Mount Improbable* (New York: Norton, 1996).

[10] Dembski, William A., "Uniform Probability," *Journal of Theoretical Probability* 3(4) (1990): 611–626, available online at www.designinference.com/documents/2004.12.Uniform_Probability.pdf.

[11] Dembski, William A., *No Free Lunch: Why Specified Complexity Cannot Be Purchased without Intelligence* (Lanham, Md.: Rowman and Littlefield, 2002).

[12] Dembski, William A., "Information as a Measure of Variation," (2004): under submission with *Complexity*, available online at www.designinference.com/documents/2004.08.Variational_Information.pdf.

[13] Dembski, William A. and Michael Ruse, eds., *Debating Design: From Darwin to DNA* (Cambridge: Cambridge University Press, 2004).

[14] Depew, David J. and Bruce H. Weber, *Darwinism Evolving: Systems Dynamics and the Genealogy of Natural Selection* (Cambridge, Mass.: MIT Press, 1995).

[15] Dinculeanu, Nicolae, *Vector Integration and Stochastic Integration in Banach Spaces* (New York: Wiley, 2000).

[16] Doob, J. L., *Classical Potential Theory and Its Probabilistic Counterpart* (New York: Springer-Verlag, 1984).

[17] Dudley, R. M., *Probability and Metrics* (Aarhus: Aarhus University Press, 1976).

[18] Feller, William, *An Introduction to Probability Theory and Its Applications*, 3rd ed., vol. 1 (New York: Wiley, 1968).

[19] Fogel, David B., Timothy J. Hays, Sarah L. Hahn, and James Quon, "A Self-Learning Evolutionary Chess Program," *Proceedings of the IEEE* (2005): forthcoming.

[20] Gelfand, I. M., "Sur un Lemme de la Theorie des Espaces Lineaires," *Comm. Inst. Sci. Math. de Kharko.* 13(4) (1936): 35–40.

[21] Hardy, G. H., J. E. Littlewood, and G. Pólya, *Inequalities*, 2nd ed. (Cambridge: Cambridge University Press, 1952).

[22] Lloyd, Seth, "Computational Capacity of the Universe," *Physical Review Letters* 88(23) (2002): 7901–4.

[23] Parthasarathy, K. R., *Probability Measures on Metric Spaces* (New York: Academic Press, 1967).

[24] Pettis, B. J., "On Integration in Vector Spaces," *Transactions of the American Mathematical Society* 44 (1938): 277–304.

[25] Port, Sidney C. and Charles J. Stone, *Brownian Motion and Classical Potential Theory* (New York: Academic Press, 1978).

[26] Reed, Russell D. and Robert J. Marks II, *Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks* (Cambridge, Mass.: MIT Press, 1999).

[27] Rényi, Alfred, "On Measures of Information and Entropy," in J. Neyman, ed., *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1 (Berkeley, Calif.: University of California Press, 1961): 547–561.

[28] Shermer, Michael, *How We Believe: Science, Skepticism, and the Search for God*, 2nd ed. (New York: Owl Books, 2003).

[29] Shi, Yuhui, "Particle Swarm Optimization," *IEEE Connections* 2(1) (2004): 8–13.

[30] Spitzer, Frank, *Principles of Random Walk*, 2nd ed. (New York: Springer-Verlag, 2001).

[31] Spivak, Michael, *Calculus*, 2nd ed. (Berkeley, Calif.: Publish or Perish, 1980).

[32] Wolpert, David H. and William G. Macready, "No Free Lunch theorems for Optimization," *IEEE Transactions on Evolutionary Computation* 1(1) (1997): 67–82.

[33] Wolpert, David H. and William G. Macready, "Coevolutionary Free Lunches" (2005): preprint.