entropy

MDPI

Article

Assessing, Testing and Estimating the Amount of Fine-Tuning by Means of Active Information

Daniel Andrés Díaz-Pachón [1] and Ola Hössjer [2,*]

[1] Division of Biostatistics, University of Miami, Miami, FL 33136, USA
[2] Department of Mathematics, Stockholm University, 114 19 Stockholm, Sweden
* Correspondence: ola@math.su.se; Tel.: +46-706721218

Abstract: A general framework is introduced to estimate how much external information has been infused into a search algorithm, the so-called active information. This is rephrased as a test of fine-tuning, where tuning corresponds to the amount of pre-specified knowledge that the algorithm makes use of in order to reach a certain target. A function f quantifies specificity for each possible outcome x of a search, so that the target of the algorithm is a set of highly specified states, whereas fine-tuning occurs if it is much more likely for the algorithm to reach the target as intended than by chance. The distribution of a random outcome X of the algorithm involves a parameter θ that quantifies how much background information has been infused. A simple choice of this parameter is to use θf in order to exponentially tilt the distribution of the outcome of the search algorithm under the null distribution of no tuning, so that an exponential family of distributions is obtained. Such algorithms are obtained by iterating a Metropolis–Hastings type of Markov chain, which makes it possible to compute their active information under the equilibrium and non-equilibrium of the Markov chain, with or without stopping when the targeted set of fine-tuned states has been reached. Other choices of tuning parameters θ are discussed as well. Nonparametric and parametric estimators of active information and tests of fine-tuning are developed when repeated and independent outcomes of the algorithm are available. The theory is illustrated with examples from cosmology, student learning, reinforcement learning, a Moran type model of population genetics, and evolutionary programming.

Keywords: active information; exponential tilting; fine-tuning; functional information; large deviations; Markov chains; Metropolis–Hastings; Moran model; statistical estimation and testing

Citation: Díaz-Pachón, D.A.; Hössjer, O. Assessing, Testing and Estimating the Amount of Fine-Tuning by Means of Active Information. *Entropy* **2022**, *24*, 1323. https://doi.org/10.3390/e24101323

Academic Editors: Augustine Wong and Xiaoping Shi

Received: 23 August 2022
Accepted: 19 September 2022
Published: 21 September 2022

1. Introduction

When Gödel published their incompleteness theorems [1], the mathematical world was shaken such that to date it has neither recovered nor fully assimilated the consequences [2]. Hilbert's program to base mathematics on a finite set of axioms had previously been pursued by Alfred North Whitehead and Bertrand Russell [3]. However, this approach turned out to be wrong when Gödel proved that no finite set of axioms in a formal system can prove all its true statements, including its own consistency. At a similar but lesser scale, when David Wolpert and William MacReady published their No Free Lunch Theorems (NFLTs, [4,5]), there was disquiet in the community because these results imply that there is no one-size-fit-all algorithm that can do well in all searches [6], and thus that a "theory of everything" is not possible in machine learning. Wolpert and MacReady concluded that it was necessary to incorporate "problem-specific knowledge into the behavior of the algorithm" [5]. Thus, active information (actinfo) was introduced in order to measure the amount of information carried out by such problem-specific knowledge [7,8]. More specifically, the NFLTs say that no search works better on average than a blind search, i.e., a search according to a uniform distribution. Accordingly, actinfo is defined as

$$I^+ = \log \frac{P(A)}{P_0(A)}, \tag{1}$$

where $A \subset \Omega$ is the non-empty target of the search algorithm, a subset of the finite sample space Ω, and P_0 is a uniform probability measure ($P_0(A) = |A|/|\Omega|$). P must be seen here as the probability measure induced by the problem-specific knowledge of the researcher, whereas P_0 is the underlying distribution assumed in the NFLTs. This corresponds to absence of problem-specific knowledge, in accordance with Bernoulli's Principle of Insufficient Reason (PoIR). An equivalent characterization of actinfo is the reduction in functional information

$$I^+ = I_{f0} - I_f = -\log P_0(A) - (-\log P(A)) \tag{2}$$

between algorithms that do not and do make use background knowledge. The name functional information was introduced by Szostak and collaborators [9,10]. It refers to applications wherein A corresponds to all outcomes of an algorithm that are functional according to some criterion. Then, I_{f0} and I_f are the self-information (measured in nats) of the event that an algorithm X produces a functional outcome, given that it was generated under P_0 and P, respectively.

Suppose we do not know whether problem-specific knowledge has been used or not when the random search $X \in \Omega$ was generated. This corresponds to a hypothesis testing problem

$$\begin{aligned} H_0 &: X \sim P_0, \\ H_1 &: X \sim P, \end{aligned} \tag{3}$$

where data are generated from distributions P_0 and P under the null and alternative hypotheses H_0 and H_1, respectively. It follows from (1) that I^+ is the log likelihood ratio when testing H_0 against H_1, if data are censored so that only $X \in A$ is known.

When the sample space Ω is finite or a continuous, bounded subset of a Euclidean space, the PoIR can be motivated by the fact that the uniform distribution maximizes the Shannon entropy, since it thereby maximizes ignorance about the outcome of X. However, the uniform distribution is neither a feasible choice of P_0 for infinite, countable sample spaces nor for continuous, unbounded samples spaces. For this reason, actinfo was generalized to deal with unbounded spaces [11], by choosing P_0 to maximize Shannon entropy under side constraints ξ, such as the existence of various moments. This gives rise to a family of null distributions $P_0 = P_{0\xi}$, with a ξ a nuisance parameter that has to be estimated or controlled for in order to estimate or give bounds for the active information.

Actinfo has also been used for mode detection in unsupervised learning, among other applications [12,13]. Based on previous works by Montañez [14,15], actinfo has been used in the past for testing hypotheses [16]. More specifically, P is regarded as a random measure in [16], so that I^+ is random as well and expressions for the tail probability of I^+ can be found.

1.1. Fine-Tuning

Fine-tuning (FT) was introduced by Carter in physics and cosmology [17]. According to FT, the constants in the laws of nature and/or the boundary conditions in the standard models of physics must belong to intervals of low probability in order for life to exist. Since its inception, FT has generated a great deal of fascination, seen in multiple divulgation books (e.g., [18–21]) and scientific articles (e.g., [22–25]). For a given constant of nature X, the connection between FT and active information can be described in three steps:

(i) Establishing the life-permitting interval (LPI) A that allows the existence of life for the constant, with $\Omega = (0, \infty) = \mathbb{R}^+$ or $\Omega = \mathbb{R}$ denoting the range of values that this constant could possibly take, including those that do not permit life.

(ii) Determining the probability $P_0(A)$ of such a LPI. If $P_0 = P_{0\xi}$ contains unknown parameters ξ, find an upper bound

$$P_{0\max}(A) = \max_{\xi} P_{0\xi}(A) \tag{4}$$

of $P_0(A)$.

(iii) Suppose that H_1 corresponds to an agent who uses background knowledge of what is required for life to exist in order to bring about a constant of nature X that definitely permits life ($P(A) = 1$). The active information $I^+ = I_{f0} = -\log P_0(A)$ is then a measure of how much background knowledge this agent has infused. Following [26,27], we conclude that X is finely tuned when the lower bound $-\log P_{0\max}(A)$ of $I^+ = I_{f0}$ is large enough. That is, FT corresponds to infusing a high degree of background knowledge into a problem.

Fine-tuning has also been used in biology. Dingjan and Futerman explored it for cell membranes [28,29], whereas Thorvaldsen and Hössjer [30] formalized it for a large class of biological models. According to [30], a system is fine-tuned if it satisfies the two following requirements:

(a) It has an independent specification;
(b) It is very unlikely to occur by chance.

1.2. The Present Article

In this article, actinfo will not only be used in the algorithmic sense. It will also be employed for testing the presence of and estimating the degree of fine-tuning (FT) of a search algorithm or agent who brings about X. Our definition of FT relies on (a) and (b), and in order to formalize these two concepts, we introduce a specificity function f, which quantifies, in terms of $f(x)$, how specified an outcome $x \in \Omega$ is. The target A, on the other hand, is a set of highly specified states, that is, all states with a degree of specificity that exceed a given threshold f_0. Then, I^+ in (1) is a test statistic for testing whether an algorithm has a much larger probability of reaching the set of highly specified states compared to a random search. This is a test of FT, since reaching the target corresponds to specificity (a), whereas reaching it with a much higher probability than expected by chance corresponds to (b).

To calculate I^+, the distributions P_0 and P of the random search algorithm under H_0 and H_1, respectively, need to be defined. As mentioned above, the null distribution P_0 is typically chosen according to some criterion, such as a maximizer of entropy, possibly with some extra constraints on moments for unbounded Ω, which was the strategy implemented in [26,27]. Another possibility is to choose P_0 as the equilibrium distribution of a Markov chain that models the dynamics of the system under the null hypothesis of no external input. In general, $P_0 = P_{0\xi t}$ involves a number of nuisance parameters ξ, and sometimes, also the time point t when an algorithm that does not make use of external information stops. The choice of $P = P_{\theta\xi t}$ is problem specific, and it possibly involves the nuisance parameters ξ of the null distribution, the time point t when the algorithm stops, as well as the tuning parameters θ that correspond to infusing the background knowledge into the search problem. Therefore, in its most general form, the actinfo (1) is a function $I^+ = I^+(\theta, \xi, t)$ of the tuning parameters θ, the nuisance parameters ξ, and the time point t.

This general framework has many applications based on different choices of f, A, P_0, and P. For some models, f is a binary function that quantifies functionality, so that A is the set of objects of a certain type (e.g., universes, proteins, protein complexes, or cellular networks) that are functional or permit life, among the set Ω of all such objects.

Another possibility is to choose A as the set of populations x whose (expected) fitness $f(x)$ exceeds a given threshold. In this setting, $P_{0\xi t}(A)$ corresponds to the probability that a randomly chosen population would evolve and reach target A of high fitness at time t, given that no background knowledge of the specificity function f is used to generate X, so that natural selection does not occur. The functional information $I_{f0} = -\log P_{0\xi t}(A)$ corresponds to the amount of external information that an evolutionary algorithm infuses under H_1, given that it brings about X so that A happens with certainty ($P(A) = 1$) within time t. In this case, the population is finely tuned when I_{f0} is large enough. More generally, we say that an evolutionary algorithm that generates $X \sim P = P_{\theta\xi t}$ after t time steps is finely

tuned when $I^+(\theta, \xi, t)$ is large enough. Typically, θ involves the selection parameters that determine to which extent a population evolves towards higher fitness.

A third possibility is to choose $f(X) = X$ as the test score of a randomly chosen student, whereas $A = [f_0, \infty)$ is the set of results of those students who pass the test with a score of at least f_0. Assume that $f(X) \sim N(\xi, 1)$ for a randomly chosen student who did not prepare for the test (H_0), whereas $f(X) \sim N(\xi + \theta t, 1)$ for a randomly chosen student who prepared for the test for a period of length t (H_1). Then, $P_0(A) = P_{0\xi}(A) = 1 - \Phi(f_0 - \xi)$, whereas $P(A) = P_{\theta\xi t}(A) = 1 - \Phi(f_0 - \xi - \theta t)$, where Φ is the cumulative distribution function of a standard normal distribution. In particular, the tuning parameter $\theta > 0$ corresponds to the amount of knowledge that a student is expected to generate per unit time of study.

The unified treatment of search problems and FT of this paper, is organized as follows: Section 2 introduces the specification function f and the set A of highly specified states. Section 3 introduces a class of probability distributions $P = P_\theta$ for which the specificity function f is used to exponentially tilt the null distribution P_0, so that outcomes with high specificity are more likely to occur, and with a scalar tuning parameter θ of P_θ that corresponds to the amount of exponential tilting. A proof is presented that it is possible to obtain a Metropolis–Hastings type Markov chain in discrete time $t = 0, 1, 2, \ldots$, whose outcome $X = X_t$ at time t has the aforementioned exponentially tilted distribution under equilibrium, that is, when t is large. The corresponding actinfo $I^+(\theta, t)$ is shown to increase monotonically with t towards an equilibrium limit. The actinfo of a search algorithm $X = X_{t\wedge T}$ that stops at time T, when the targeted set A of highly specified states has been reached, is also shown to increase more rapidly. Section 4 introduces various nonparametric and parametric estimators of actinfo, and corresponding tests of FT, when n repeated and independent outputs of the search algorithm are available. In particular, large deviations theory is used to prove that the significance levels of these tests, i.e., the probability of detecting FT under H_0, goes to zero at an exponential rate when the sample size n increases. Section 5 presents a number of examples from cosmology, student learning, reinforcement learning, and population genetics, that illustrate our approach. A discussion in Section 6 follows, whereas the proof and further details about the models are presented in Section 7.

2. Specificity and Target

Consider a function $f : \Omega \to \mathbb{R}$ and assume that the objective of the search algorithm, or the agent that brings about X, is to find regions in Ω where f is large. The rationale for this is an independent *specification*, where a more specified state $x \in \Omega$ corresponds to a larger $f(x)$. It is further assumed that the target set in (1) is given by

$$A = \{x \in \Omega;\, f(x) \geq f(x_0)\} \tag{5}$$

for some $x_0 \in \Omega$. This implies that the purpose of the search algorithm or the agent is to bring about an X that is at least as specified as x_0. We will refer to f as a specificity function of the agent or an objective function of the search algorithm.

Several examples of specificity functions are provided in Section 5. For instance, Example 2 deals with student learning. For a special case of this model, $f(x) = x$ represents the test score of a student, whereas x_0 is a reference value that corresponds to the minimum score needed to pass the test.

For cosmological FT (Example 1), x is the value of a particular constant of nature and the specificity function equals

$$f(x) = 1_{\{x \in A\}}, \tag{6}$$

where $1_{\{\cdot\}}$ is the indicator function. That is, f has a binary range, with $f(x) = 1$ and 0 corresponding to whether x permits a universe with life, and in particular, x_0 is a universe that permits life. From this, A is the LPI of this constant. Moreover, X is the value of this constant of nature for a randomly generated universe, with a distribution that either incorporates external information (H_1) or not (H_0).

In the context of proteins, x is taken to be an amino acid sequence, whereas $f(x)$ in (6) quantifies whether the protein that the amino acid corresponds to is functional (1) or not (0). For instance, X could be the outcome of a random evolutionary process, the goal of which is to generate a functioning protein, and this process either makes use of external information (H_1) or not (H_0). In a more refined example (Example 4), x is a molecular machine that consists of a possibly large number of proteins (or parts), and $f(x)$ is (a monotone function of) the fitness of x.

Interpretation of Target

There are at least two ways of interpreting x_0, and hence also the target set A. According to the first interpretation, x_0 is the outcome of random variable $X' \in \Omega$; that is, the outcome of a first search. Suppose that X is another random variable that represents a second (possibly future) search, independent of X'. Then, if we condition the outcome x_0 of the first search, the actinfo I^+ in (1) is the log likelihood ratio for the event that the second search variable X is *at least as specified* as the observed value $f(x_0)$ of the first search.

There is, however, no need to associate x_0 in (5) with a first search variable X'. Instead, some a priori information may be used to define which values of f represent a high amount of specificity. This gives rise to the second interpretation of x_0, according to which x_0 is used for defining outcomes with a high and low degree of specificity, using $f_0 = f(x_0)$ as a cutoff. According to this interpretation, the two sets A in (5) and its complement

$$A^c = \Omega \setminus A = \{x;\, f(x) < f(x_0)\}$$

represent a dichotomization of specificity, so that A and A^c consist of all states with high and low specificity, respectively. With this interpretation of x, I^+ is the log likelihood ratio for testing FT based on the search variable X. In particular, suppose that the specificity function f is bounded, i.e.,

$$f_{\max} = \max_{x \in \Omega} f(x) < \infty. \tag{7}$$

Then, the most stringent definition of high specificity,

$$f_0 = f_{\max}, \tag{8}$$

only regards outcomes with a maximal value of f as highly specified, so that

$$A = \Omega_{\max} = \{x \in \Omega;\, f(x) = f_{\max}\}. \tag{9}$$

Note that (6) is a special case of (9).

3. Active Information for Exponentially Tilted Systems

Throughout Section 3, ξ is assumed to be known and the null distribution does not involve any time index t. Therefore, P_0 is known, whereas $P = P_{\theta t}$ involves the tuning parameters θ and the time index t. It will be further assumed in Sections 3.1 and 3.2 that the system is in equilibrium, or that the time index t is fixed, so that t can also be dropped under H_1 ($P = P_\theta$).

3.1. Exponential Tilting

Let P_θ be an exponentially tilted version of P_0 for some scalar tuning parameter $\theta > 0$, which will also be called a tilting parameter. Exponential tilting is often used for rare events simulation [31,32]. Here, f is used to define the tilted version of P_0 as

$$P_\theta(x) = \frac{e^{\theta f(x)}}{M(\theta)} P_0(x), \tag{10}$$

with

$$M(\theta) = \sum_{x \in \Omega} e^{\theta f(x)} P_0(x) \tag{11}$$

a normalizing constant assuring that P_θ is a probability measure. For countable sample spaces Ω, we interpret $P_0(x)$ and $P_\theta(x)$ as the probability masses, whereas for continuous sample spaces, they are probability densities and the sum in (11) is replaced by an integral. The larger the tilting parameter $\theta > 0$, the more the probability mass of P_θ concentrates on regions of large f. In particular, P_∞, the weak limit of P_θ as $\theta \to \infty$, is supported on (9) whenever (7) holds.

The parametric family

$$\mathcal{P} = \{P_\theta;\, \theta \geq 0\} \tag{12}$$

of distributions is an exponential family [33] (Section 1.5), and each $P_\theta \in \mathcal{P}$ gives rise to a separate version of actinfo. This is summarized in the following proposition (cf. Section 7 for a proof):

Proposition 1. *Suppose the target set A is defined as in (5) for some $x_0 \in \Omega$ such that $P_0(A) > 0$. Then, $P_\theta(A)$ is a strictly increasing function of $\theta \geq 0$ with $P_\infty(A) = 1$. Consequently, the actinfo*

$$I^+(\theta) = \log \frac{P_\theta(A)}{P_0(A)} \tag{13}$$

is a strictly increasing function of $\theta \geq 0$, with $I^+(0) = 0$ and $I^+(\infty) = I_{f0} = -\log P_0(A)$.

The intuitive interpretation of Proposition 1 is that the larger θ is, the more problem-specific knowledge is infused into P_θ in terms of shifting probability mass towards regions in Ω where f, the specificity function, is large.

A simple instance of exponential tilting is the student learning example of Section 1.2. Recall that $f(x) = x$ is the test score of a student, with $X \sim N(\xi, 1)$ for a randomly chosen student who did not prepare for the test (H_0), whereas $X \sim N(\xi + \theta, 1)$ is the test score of a randomly chosen student who prepared for the test during $t = 1$ units of time (H_1). It is clear that

$$\begin{aligned} P_0(x) &= e^{-(x-\xi)^2/2}/\sqrt{2\pi}, \\ P_\theta(x) &= e^{-(x-\xi-\theta)^2/2}/\sqrt{2\pi} = P_0(x)e^{\theta x}/M(\theta). \end{aligned}$$

3.2. Metropolis–Hastings Systems with Exponential Tilting Equilibrium

Inspired by Markov Chain Monte Carlo methods [34], consider a Markov chain $X_0, X_1, \ldots \in \Omega$ for which P_θ is the equilibrium distribution. Consequently, if $P = P_\theta$ (that is, under the alternative hypothesis H_1 in (3) when $\theta > 0$), $X = X_t$ may be interpreted as the outcome of an algorithm after t iterations, provided that t is so large that the equilibrium has been reached. The assumption is made that this algorithm knows f and tries to explore the whole state space Ω. If the Markov chain has an equilibrium distribution (10), this corresponds to an algorithm that favors jumps towards the regions of large f when $\theta > 0$, an effect which is accentuated the higher the value of θ is. In further detail, the transition kernel of the chain is an instance of the well-known Metropolis–Hastings (MH) algorithm [35,36], which is closely related to simulated annealing [37]. This kernel has a probability or density

$$\pi_\theta(x, y) = r_\theta(x)\delta(x, y) + \alpha_\theta(x, y)q(x, y) \tag{14}$$

for jumps from x to y, where $\delta(x,\cdot)$ is a point mass at $x \in \Omega$, $q(x,\cdot)$ is a proposal distribution of jumps from a current position x of the Markov chain,

$$\alpha_\theta(x,y) = \min\left[1, \frac{e^{\theta f(y)} P_0(y) q(y,x)}{e^{\theta f(x)} P_0(x) q(x,y)}\right] \tag{15}$$

is the probability of accepting a proposed move from x to y, whereas

$$r_\theta(x) = 1 - \sum_{y\in\Omega} \alpha_\theta(x,y) q(x,y) \tag{16}$$

is the probability that the Markov chain rejects a proposed move away from x (for continuous sample spaces $q(x,\cdot)$ is a probability density and then the sum in (16) is replaced by an integral). The transition of the Markov chain from $X_t = x$ to the next state X_{t+1} is described in two steps as follows. First, a candidate $Y \sim q(x,\cdot)$ is proposed. Then, in the second step, this candidate is either accepted with a probability of $\alpha_\theta(x,Y)$, so that $X_{t+1} = Y$, or it is rejected with probability $1 - \alpha_\theta(x,Y)$, so that $X_{t+1} = X_t$. It is well known that P_θ is the equilibrium distribution of this Markov chain whenever it is irreducible; that is, provided the proposal distribution q is defined in such a way that moving between any pair of states in Ω in a finite number of steps is possible [38], pp. 243–245.

In particular, if q is symmetric and P_0 is uniform, then a proposed upward move with $f(Y) > f(x)$ and $P_\theta(Y) > P_\theta(x)$ is always accepted, whereas a proposed downward move with $f(Y) < f(x)$ is accepted with a probability of $P_\theta(Y)/P_\theta(x)$. The Markov chain only makes local jumps if $q(x,\cdot)$ puts all its probability mass in a small neighborhood of x, for any $x \in \Omega$. At the other extreme is a chain with the global proposal distribution $q(x,\cdot) \sim P_\theta$ for any $x \in \Omega$; all proposed jumps of this chain are then accepted ($\alpha(x,y) = 1$), and $\{X_t\}_{t=1}^{\infty}$ is a sequence of independent and identically distributed (i.i.d.) random variables with $X_t \sim P_\theta$.

The choice of proposal distribution q is problem specific. In this section, we defined q for the Metropolis–Hastings type algorithms that require knowledge of the specificity function f, since the acceptance probability (15) is a function of f. Proposed moves also occur for evolutionary algorithms (Examples 4 and 5 of Section 5). These algorithms are typically the result of many small changes, with specificity corresponding to functionality or fitness. The proposed moves are local mutations that either survive (are accepted) or do not. Other algorithms (such as reinforcement learning in Example 3 of Section 5) only make use of estimates of the specificity function. However, it is still meaningful for these algorithms to talk about proposed moves that are initially large (exploration phase) followed by a subsequent period of small or no moves (exploitation phase). In the context of Metropolis–Hastings algorithms, this is the strategy of simulated annealing, where large moves are initially proposed (corresponding to high temperatures), followed by subsequent small proposed moves (corresponding to low temperatures).

3.3. Active Information for Metropolis–Hastings Systems in Non-Equilibrium

Suppose, for simplicity, that the sample space Ω is finite, and that the states in Ω are listed in some order. Let

$$\mathbf{P}_0 = (P_0(x);\, x \in \Omega) \tag{17}$$

be a row vector of length $|\Omega|$ with all the null distribution probabilities, and let

$$\mathbf{\Pi}_\theta = (\pi_\theta(x,y); x, y \in \Omega) \tag{18}$$

be a square matrix of order $|\Omega|$ that defines the transition kernel of the Markov chain $\{X_t\}_{t=0}^{\infty}$ of Section 3.2. If $X_0 \sim P_0$, then by the Kolmogorov–Chapman equation $X_t \sim P_{\theta t}$, where

$$(P_{\theta t}(x); x \in \Omega) = \mathbf{P}_{\theta t} = \mathbf{P}_0 \mathbf{\Pi}_\theta^t. \tag{19}$$

Hence, if $P = P_{\theta t}$, then $X = X_t$ corresponds to observing the Markov chain at time t, under the alternative hypothesis H_1 in (3). Some basic properties of the corresponding actinfo are summarized in the following proposition, which is proved in Section 7:

Proposition 2. *Suppose that $X = X_t$ is obtained by iterating t times a Markov chain with initial distribution (17) and transition kernel (18). The actinfo then equals*

$$I^+(\theta, t) = \log \frac{P_{\theta t}(A)}{P_0(A)} = \log \frac{\mathbf{P}_0 \mathbf{\Pi}_\theta^t \mathbf{v}}{\mathbf{P}_0 \mathbf{v}}, \tag{20}$$

where $\mathbf{v}$ is a column vector of length $|\Omega|$ with ones in positions $x \in A$ and zeros in positions $x \in A^c$. In particular, $I^+(\theta, 0) = 0$ and

$$\lim_{t \to \infty} I^+(\theta, t) = I^+(\theta). \tag{21}$$

Therefore, $I^+(\theta, t) > 0$ corresponds to knowledge of f being used to generate t jumps of the Markov chain, under the alternative hypothesis H_1 in (3).

3.4. Active Information for Metropolis–Hastings Systems with Stopping

In Section 3.3, $P \sim P_{\theta t}$ was obtained by starting a random search with null distribution P_0, and then iterating the Markov chain of Section 3.2 t times. However, knowledge of f can be utilized even more and stop the Markov chain if the target A in (5) is reached before time t. This can be formalized by introducing the stopping time

$$T = \min\{t \geq 0;\, X_t \in A\} \tag{22}$$

and letting

$$P_{\theta ts}(x) = P(X_{t \wedge T} = x) \tag{23}$$

be the probability distribution of the stopped Markov chain $X_{t \wedge T}$, with the last index s in (23) being an acronym for stopping. In particular,

$$P_{\theta ts}(A) = \sum_{x \in A} P_{\theta ts}(x) = P(T \leq t) \tag{24}$$

is the probability of reaching the target A for the first time after t iterations or earlier. The theory of phase-type distributions can then be used to compute the target probability $P_{\theta ts}(A)$ in (23) [39,40]. To this end, clump all states $x \in A$ into one absorbing state, and decompose the transition kernel in (18) according to

$$\mathbf{\Pi}_\theta = \begin{pmatrix} \mathbf{\Pi}_\theta^{\mathrm{na}} & \mathbf{\Pi}_\theta^{\mathrm{na,a}} \\ \mathbf{0} & 1 \end{pmatrix}, \tag{25}$$

where $\mathbf{\Pi}_\theta^{\mathrm{na}}$ is a square matrix of order $|A^c|$ containing the transition probabilities between all non-absorbing states in A^c, whereas $\mathbf{\Pi}_\theta^{\mathrm{na,a}}$ is a column vector of length $|A^c|$ with transition probabilities $\pi(x, A)$ from all the non-absorbing states $x \in A^c$ into the absorbing state A. Moreover, $\mathbf{P}_0^{\mathrm{na}} = (P_0(x);\, x \in A^c)$ is a row vector of length $|A^c|$ that is the restriction of the start-distribution $\mathbf{P}_0$ in (17) to all non-absorbing states. Then

$$P_{\theta ts}(A) = 1 - \mathbf{P}_0^{\mathrm{na}}(\mathbf{\Pi}_\theta^{\mathrm{na}})^t\mathbf{1}, \tag{26}$$

where $\mathbf{1}$ is a column vector of $|A^c|$ ones.

The actinfo I_s^+ of a search procedure with stopping is thus defined:

Proposition 3. *Suppose that $X = X_t$ is obtained by iterating a Markov chain with an initial distribution (17) and a transition kernel (18) (for some $\theta \geq 0$) at most t times, and stopping whenever the set A is reached. Then, the actinfo is given by*

$$I_s^+(\theta, t) = \log \frac{P_{\theta ts}(A)}{P_0(A)} = \log \frac{1 - \mathbf{P}_0^{na}(\mathbf{\Pi}_\theta^{na})^t\mathbf{1}}{\mathbf{P}_0\mathbf{v}}, \tag{27}$$

with $\mathbf{P}_0$ and $\mathbf{v}$ as in Proposition 2, whereas $\mathbf{P}_0^{na}$, $\mathbf{\Pi}_\theta^{na}$, and $\mathbf{1}$ are defined below (25) and (26). This actinfo satisfies

$$I_s^+(\theta, t) \geq I^+(\theta, t) \tag{28}$$

and $I_s^+(\theta, t)$ is a non-decreasing function of t such that

$$\lim_{t\to\infty} I_s^+(\theta, t) = I_{f0} \tag{29}$$

and

$$\sum_{t=0}^{\infty}\left(1 - P_0(A)e^{I_s^+(\theta,t)}\right) = E(T). \tag{30}$$

Proposition 3 is proven in Section 7. Inequality (28) states that, for a search procedure with t iterations, knowledge about f that is used for *stopping* the Markov chain in (18) will increase the actinfo, regardless of whether knowledge about f was used ($\theta > 0$) or not ($\theta = 0$) when *iterating* the Markov chain. Equation (29) is a consequence of the fact that target A is eventually reached with probability 1, so that the actinfo of a search procedure with stopping equals the functional information $I_{f0} = -\log P_0(A)$ after many iterations of the Markov chain. Moreover, Equation (30) tells that the rate at which $P_0(A)e^{I_s^+(\theta,t)}$ approaches 1 is determined by the expected waiting time $E(T)$ of reaching the target.

From Proposition 3, actinfo for a system with stopping is closely related to the phase-type distribution of the waiting time T until the target is reached. This has been studied in [41], in the context of the expression of a number of genes, with x being the collection of the regulatory regions of all these genes.

4. Estimating Active Information and Testing Fine-Tuning

In Section 3, we gave explicit expressions of the actinfo, for Metropolis–Hastings algorithms with a scalar tuning parameter θ. In general, however, it might be infeasible to calculate I^+, either because the sample space is very large, or the nuisance parameters ξ and/or the tuning parameters θ are unknown. If is of interest then to consider ways of estimating I^+ from data, for instance through Monte Carlo-based methods. To this end, we will assume that the random search algorithm is repeated independently, under the same conditions, n times. For instance, suppose that $\{X_{it}\}_{t=0}^{\infty}$ corresponds to independent realizations $i = 1, \ldots, n$ of a search algorithm. If these independent realizations are recorded or stopped at one single time point, the outcome is either $X_i = X_{it}$ for $i = 1, \ldots, n$, or $X_i = X_{i,t\wedge T_i}$, for $i = 1, \ldots, n$, depending on whether the search algorithm is stopped at a fixed time point t or at random time points $\{T_i\}_{i=1}^n$. In either case, an output of i.i.d. random variables

$$X_1, \ldots, X_n \sim Q \tag{31}$$

is obtained. These repeated outcomes of the search algorithm will be used to test for and estimate the degree of fine-tuning. The methodology depends on whether the null distribution P_0 is known or involves unknown nuisance parameters.

4.1. Null Distribution Known

Suppose the null distribution P_0 is known. The sample in (31) is then used for testing between the two hypotheses

$$\begin{array}{ll} H_0: & Q = P_0, \\ H_1: & Q \in \mathcal{P}_1, \end{array} \tag{32}$$

with

$$\mathcal{P}_1 = \{P;\, P(A) \geq p_{\min}\} \tag{33}$$

the set of distributions that correspond to fine-tuning. Suppose an estimate $\hat{Q}(A)$ of the probability that $X \in A$ is computed from data (31), with an associated empirical actinfo

$$\hat{I}^+ = \hat{I}_n^+ = \log \frac{\hat{Q}(A)}{P_0(A)}. \tag{34}$$

If $\hat{Q}(A)$ is a consistent estimator of $Q(A)$, then for large sample sizes, $\hat{I}^+$ will be close to

$$I_Q^+ = \log \frac{Q(A)}{P_0(A)}, \tag{35}$$

which equals 0 under H_0 and $I^+ = I_P^+$ under H_1, for some particular $P \in \mathcal{P}_1$. To test H_0 against H_1,

$$\text{reject } H_0 \text{ when } \hat{I}^+ \geq I_{\min}, \tag{36}$$

where $I_{\min}$ is a pre-specified lower bound on the range of values of the actinfo that corresponds to FT.

4.1.1. Nonparametric Estimator and Test

In Section 3, $P = P_\theta$, $P = P_{\theta t}$, or $P = P_{\theta ts}$ were used for distributions that make use of pre-specified knowledge. These distributions involve the tilting parameter θ, and possibly also the number of iterations t of the algorithm and a stopping time T. In this section, however, no other assumption than $P \in \mathcal{P}_1$ is made on P, and a nonparametric version of the empirical actinfo is used. The fraction

$$\hat{Q}(A) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_i \in A\}} \tag{37}$$

of random searches that fall into A is used as an estimate of $Q(A)$. Therefore, (37) only requires the knowledge of the set A, not of the function f.

The following result establishes the asymptotic normality of the nonparametric version of the estimator $\hat{I}^+$ in (34). Moreover, large deviations [42] are used to show that the significance level of the nonparametric version of the FT test (36) goes to zero exponentially fast with n (see Section 7 for more details of the proof).

Proposition 4. *Suppose the empirical actinfo $\hat{I}_n^+$ in (34) is computed nonparametrically, using (37) as an estimate of the target probability $Q(A)$. Then, $\hat{I}_n^+$ is an asymptotically normal estimator of I_Q^+ in (35), in the sense that*

$$\sqrt{n}(\hat{I}_n^+ - I_Q^+) \xrightarrow{\mathcal{L}} N(0, V) \text{ as } n \to \infty, \tag{38}$$

where $\xrightarrow{\mathcal{L}}$ *refers to convergence in distribution, and*

$$V = \frac{1 - Q(A)}{Q(A)} \tag{39}$$

is the variance of the limiting normal distribution. The significance level of the test (36) *for fine-tuning, with threshold* I_{min}*, satisfies*

$$\lim_{n\to\infty} -\frac{\log\left(P_{H_0}(\hat{I}_n^+ \geq I_{min})\right)}{n} = C, \tag{40}$$

where

$$C = p_{min} \log \frac{p_{min}}{P_0(A)} + (1 - p_{min}) \log \frac{1 - p_{min}}{1 - P_0(A)} \tag{41}$$

is the Kullback–Leibler divergence between Bernoulli distributions with success probabilities $p_{min} = P_0(A)\exp(I_{min})$ *and* $P_0(A)$*, respectively.*

Remark 1. *The conclusion of Proposition 4 is that the probability of observing actinfo that corresponds to fine-tuning by chance decays at rate* e^{-Cn} *when the sample size n becomes large.*

4.1.2. Parametric Estimator and Test

Suppose that there is a priori knowledge that P is close to the parametric exponential family $\mathcal{P}$ of distributions in (10)–(12) for some value $\theta > 0$ of the tilting parameter. A parametric test of actinfo is naturally defined. For this, first compute the maximum likelihood estimate

$$\hat{\theta} = \hat{\theta}_n = \arg\max_{\theta\geq 0} \sum_{i=1}^{n} \log P_\theta(X_i) \tag{42}$$

of θ, and use it to define a parametric estimate

$$\hat{Q}(A) = P_{\hat{\theta}}(A) \tag{43}$$

of the target probability $Q(A)$ that is inserted into (34) to define a parametric version of the empirical actinfo $\hat{I}^+$. As opposed to (37), the estimate (43) requires the full knowledge of f.

To analyze the properties of the estimator (34) and test (36), introduce

$$\theta^* = \arg\min_{\theta\geq 0} D_{KL}(Q \parallel P_\theta), \tag{44}$$

where

$$D_{KL}(Q \parallel P_\theta) = \sum_{x\in\Omega} Q(x) \log \frac{Q(x)}{P_\theta(x)} \tag{45}$$

is the Kullback–Leibler divergence between Q and P_θ. From (44), P_{θ^*} is the distribution in $\mathcal{P}$ that best approximates Q. In particular, $\theta^* = \theta$ if $Q \in \mathcal{P}$ and $Q = P_\theta$ for some $\theta \geq 0$.

The following proposition shows that $\hat{I}^+$ is an asymptotically normal estimator of $I^+(\theta^*)$ in (13), which differs from I_Q^+ in (35) whenever $Q \notin \mathcal{P}$. Moreover, the proposition also provides large sample properties of the significance level of the test for actinfo (cf. Section 7 for details of the proof):

Proposition 5. *Suppose the empirical actinfo* $\hat{I}_n^+$ *in* (34) *is computed parametrically, using an estimate* (43) *of the target probability* $Q(A)$*. Then,* $\hat{I}_n^+$ *is an asymptotically normal estimator of* $I^+(\theta^*)$*, in the sense that*

$$\sqrt{n}\big(\hat{I}_n^+ - I^+(\theta^*)\big) \xrightarrow{\mathcal{L}} N(0, V) \text{ as } n \to \infty, \tag{46}$$

where the variance of the limiting normal distribution is given by

$$V = \frac{Cov^2_{P_{\theta^*}}[f(X)I(f(X) \geq f_0)]Var_Q[f(X)]}{P^2_{\theta^*}(A)Var^2_{P_{\theta^*}}[f(X)]}. \tag{47}$$

Moreover, the significance level of the parametric test for fine-tuning, based on (36) *and* (43)*, satisfies*

$$\lim_{n\to\infty} -\frac{\log\big[P_{H_0}\big(\hat{I}_n^+ \geq I_{min}\big)\big]}{n} = C, \tag{48}$$

for

$$C = \sup_{\phi>0}\big\{\phi E_{P_{min}}[f(X)] - \log M(\phi)\big\}, \tag{49}$$

where $P_{min} = P_{\theta_{min}}$*,* $\theta_{min} < \theta^*$ *is the solution of* $P_{\theta_{min}}(A) = p_{min} = P_0(A)\exp(I_{min})$*,* $M(\phi)$ *is given by* (11)*, whereas* p_{min} *is defined in* (33)*.*

4.1.3. Comparison between Nonparametric and Parametric Estimates of Actinfo

The two versions of empirical actinfo are complementary. The nonparametric version is preferable in the sense that it makes less assumptions about the distribution P of the random algorithm under H_1, and in particular, it is a consistent estimator of I_Q^+ in (35). The parametric version of $\hat{I}^+$, on the other hand, is preferable when $nQ(A)$ is small, since it makes use of all data in order to estimate $Q(A)$, although it is not a consistent estimator of I_Q^+ when $Q \notin \mathcal{P}$. The asymptotic variances in (39) and (47), as well as the rates of exponential significance level decrease in (41) and (49), agree when $Q = P_{\theta^*}$ and $f(x) = f_0 1_{\{x\in A\}}$, which is a special case of (8).

4.2. Null Distribution Unknown

Suppose that the null distribution $P_0 = P_{0\xi}$ involves an unknown nuisance parameter $\xi \in \Xi$. The objective is then to test the two hypotheses

$$\begin{aligned} H_0 &: \quad Q \in \mathcal{P}_0, \\ H_1 &: \quad Q \in \mathcal{P}_1, \end{aligned} \tag{50}$$

where the set of distribution under the null and alternative hypotheses equals

$$\mathcal{P}_0 = \{P_{0\xi};\, \xi \in \Xi\} \tag{51}$$

and (33), respectively.

4.2.1. One Sample Available

The actinfo

$$I_Q^+ = I_Q^+(\xi) = \log\frac{Q(A)}{P_{0\xi}(A)} \tag{52}$$

cannot be consistently estimated if only one sample (31) is available. The best course of action is thus to estimate a lower bound

$$\hat{I}^+ = \hat{I}_n^+ = \log\frac{\hat{Q}(A)}{P_{0\max}(A)} \tag{53}$$

of I^+, with $P_{0\max}(A)$ defined in (4) and $\hat{Q}(A)$ an estimate of $Q(A)$. This estimator will have an asymptotic bias

$$B = I_Q^+(\xi^*) - I_Q^+ = \log \frac{P_{0\xi}(A)}{P_{0\max}(A)} \leq 0, \tag{54}$$

where ξ^* is the nuisance parameter that maximizes $P_{0\xi}(A)$ [43]. For the numerator of (53), either the nonparametric estimate of $Q(A)$ in (37) can be used, or a parametric class

$$\mathcal{P} = \{P_{\theta\xi}; \theta \in \Theta, \xi \in \Xi\}$$

of distributions can be used that involves a tuning parameter vector θ and a vector of nuisance parameters ξ. If Q is thought to be close to $\mathcal{P}$, the parametric estimate

$$\hat{Q}(A) = P_{\hat{\theta}\hat{\xi}}(A) \tag{55}$$

of $Q(A)$ is used, which generalizes (43), with

$$(\hat{\theta}, \hat{\xi}) = \arg\max_{\theta,\xi} \sum_{i=1}^{n} \log P_{\theta\xi}(X_i). \tag{56}$$

When the sample size n tends towards infinity, the estimator (56) will converge to

$$(\theta^*, \xi^*) = \arg\min_{\theta,\xi} D_{KL}(Q \parallel P_{\theta\xi}). \tag{57}$$

The following result is an extension of Propositions 4 and 5, when nuisance parameters ξ are added and a general type of tuning parameter θ (not necessarily a scalar tilting parameter) is used. A short proof of the proposition is offered in Section 7.

Proposition 6. *Suppose that the null distribution $P_0 = P_{0\xi}$ involves an unknown parameter ξ and the actinfo I_Q^+ in (52) is estimated by $\hat{I}_n^+$ in (53), using an estimator $\hat{Q}(A)$ of the target probability $Q(A)$ that is either nonparametric (37) or parametric (55). Given these assumptions, $\hat{I}_n^+$ is an asymptotically normal estimator, in the sense that*

$$\sqrt{n}(\hat{I}_n^+ - I_Q^+ - B) \xrightarrow{\mathcal{L}} N(0, V) \text{ as } n \to \infty. \tag{58}$$

The asymptotic bias B in (58) is defined in (54) whereas the asymptotic variance V is defined in (39) for the nonparametric estimator of I_Q^+, whereas

$$\begin{aligned} V &= E[\psi_{\theta^*\xi^*}(X)|X \in A]E[\psi'_{\theta^*\xi^*}(X)]^{-1}E[\psi^T_{\theta^*\xi^*}(X)\psi_{\theta^*\xi^*}(X)] \\ &\cdot \quad E[(\psi'_{\theta^*\xi^*})^T(X)]^{-1}E[\psi_{\theta^*\xi^*}(X)|X \in A]^T \end{aligned} \tag{59}$$

for the parametric estimator of I_Q^+, with $\psi_{\theta\xi}(x) = d \log P_{\theta\xi}(x)/d(\theta, \xi)$, (θ^*, ξ^*) defined as in (57), and T refering to matrix transposition. Moreover, the significance level of the test (36) of FT, with threshold I_{min}, satisfies

$$\lim_{n\to\infty} -\frac{\log\left[P_{0\xi}\left(\hat{I}_n^+ \geq I_{min}\right)\right]}{n} = C, \tag{60}$$

with

$$C = p_{min}e^{-B} \log \frac{p_{min}e^{-B}}{P_{0\xi}(A)} + (1 - p_{min}e^{-B}) \log \frac{1 - p_{min}e^{-B}}{1 - P_{0\xi}(A)} \tag{61}$$

for the nonparametric version of the test, with $p_{min} = P_{0\xi}(A)\exp(I_{min})$. For the parametric versions of the FT-test, and in the special case when θ is a scalar exponential tilting parameter, C is given by (49), with $P_{min} = P_{\theta_{min}\xi}$, and θ_{min} the solution of $P_{\theta_{min}\xi}(A) = p_{min}e^{-B}$.

Remark 2. *The negative bias term B makes the test of FT in Proposition 6 more conservative than the tests in Propositions 4 and 5. This can be seen, for instance, by comparing the two large deviation rates C in (41) and (61). The rate in (61) is larger, since p_{min} is multiplied by a term e^{-B}. This corresponds to the fact that to falsely reject H_0 in Proposition 6 is more difficult.*

4.2.2. Two Samples Available

In addition to the first sample (31), suppose a second sample

$$X_{01}, \ldots, X_{0n_0} \sim P_{0\xi} \tag{62}$$

of n_0 i.i.d. observations under the null distribution is available. A consistent estimator

$$\hat{I}^+ = \hat{I}^+_{nn_0} = \log \frac{\hat{Q}(A)}{P_{0\hat{\xi}}(A)} \tag{63}$$

of I^+_Q in (52) is then available, with

$$\hat{\xi} = \arg\max_{\xi} \sum_{i=1}^{n_0} \log P_{0\xi}(X_{0i}). \tag{64}$$

The following result provides asymptotic properties of the estimator (63) of actinfo, and the corresponding test (36) of FT with threshold $I_{\min}$ (cf. Section 7 for a proof):

Proposition 7. *Suppose that the null distribution $P_0 = P_{0\xi}$ involves an unknown nuisance parameter ξ, and that the active information I^+_Q in (52) is estimated by $\hat{I}^+_{nn_0}$ in (63), making use of two samples (31) and (62), of sizes n and n_0, from Q and $P_{0\xi}$, respectively. Further assume that the estimator $\hat{Q}(A)$ of $Q(A)$ is either nonparametric (37) or parametric (55). If $n, n_0 \to \infty$ in such a way that*

$$\frac{n}{n_0} \to \lambda > 0, \tag{65}$$

then

$$\sqrt{n}(\hat{I}^+_{nn_0} - I^+_Q) \xrightarrow{\mathcal{L}} N(0, V_1 + \lambda V_2), \tag{66}$$

where

$$V_2 = E[\psi_\xi(X)|X \in A]E[\psi^T_\xi(X)\psi_\xi(X)]^{-1}E[\psi_\xi(X)|X \in A]^T, \tag{67}$$

and $\psi_\xi(x) = d \log P_{0\xi}(x)/d\xi$. If the nonparametric estimator of $Q(A)$ is used, then V_1 equals V in (39), whereas if the parametric estimator $Q(A)$ is used, then V_1 equals V in (59). The significance level of the test (36) of FT, with threshold I_{min}, satisfies the same type of large deviation result (60) as in Proposition 6, for the nonparametric and parametric versions of the test (in the latter case assuming that θ is a scalar tilting parameter), but in the definitions of the nonparametric and parametric large deviation rates C, the bias term $B = 0$.

5. Examples

In this section, we provide five examples. The first cosmology example is a continuation of Section 1.1, with specificity corresponding to a universe that permits life. The second example of student learning was introduced in Section 1.2, with specificity being the test score of a student who prepares for a test. The third example concerns reinforcement learning, with specificity the cumulative reward of a certain trajectory of actions and environments. The last two examples concern evolutionary algorithms for generating molecular machines, with specificity corresponding to the functionality or fitness of these machines. These evolutionary algorithms can be viewed as extensions or variants of the Metropolis–Hastings algorithms of Section 3.2, where proposed moves correspond to mutations, whereas accepted moves correspond to mutations that survive and then possibly spread to a whole population.

Example 1 (Cosmology [26,27])**.** Suppose that there is a positive constant of nature $X \in \Omega = \mathbb{R}^+$, a life-permitting interval $A \subset \Omega$, and a specificity function (6) that equals 1 inside $A = (a, b)$ and zero elsewhere. The maximum entropy distribution under a first moment constraint $\xi = E(X)$ is exponential with expected value. Consequently,

$$P_{0\xi}(A) = \frac{1}{\xi} \int_a^b e^{-x/\xi} dx.$$

The null and alternative hypotheses for the fine-tuning test are given in (50), where under H_1, the agent brings about a life-permitting value of X with probability 1 ($P(A) = 1$). Only one universe is observed, with a value $X = X_1$ of the constant. Therefore, there is a sample (31) of size $n = 1$, whereas no null sample (62) is available. Since $X_1 \in A$ is life-permitting, $\hat{Q}(A) = 1$. The estimate (53) of actinfo then simplifies to

$$\hat{I}^+ = \log \frac{1}{P_{0\max}(A)} = -\log P_{0\max}(A). \tag{68}$$

Let $x = (a + b)/2$ be the midpoint of the LPI and suppose that half of its relative size $\epsilon = (b - a)/(2x)$ is small. The probability in (68) is then approximated by

$$P_{0\max}(A) \approx (b - a) \max_{\xi > 0} \frac{e^{-x/\xi}}{\xi} \approx 2\epsilon e^{-1}.$$

From (68), the estimated actinfo

$$\hat{I}^+ \approx 1 - \log(\epsilon) - \log(2)$$

is a monotone decreasing function of ϵ.

Example 2 (Evaluation of student test scores [44])**.** As a generalization of the example given in Section 1.2, suppose that a number of students perform a test. Let $x = (z, y) = (z_1, \ldots, z_{d-1}, y) \in \mathbb{R}^d$ summarize the chararcteristics of a student with covariates z that are used to predict the outcome y of the test. The specificity function $f(x) = x_d = y$ equals the student's test score, and (5) corresponds to the set of students that pass the test, with a minimally allowed score of f_0. The population of students follows a $(d-1)$-dimensional multivariate normal distribution $Z \sim N(\boldsymbol{m}, \boldsymbol{\Sigma})$, where $\boldsymbol{m} = (m_1, \ldots, m_{d-1})$ and $\boldsymbol{\Sigma} = (\sigma_{jk})_{j,k=1}^{d-1}$ are known. The conditional distribution of the response follows a multiple linear regression model

$$Y|Z = z \sim N\left(\xi_0 + \sum_{j=1}^{d-1} \xi_j z_j + t(\theta_0 + \sum_{j=1}^{d-1} \theta_j z_j), \sigma^2\right),$$

for a student with a covariate vector z who prepared for the test for a period of length t. The nuisance parameter vector $\xi = (\xi_0, \ldots, \xi_{d-1}, \sigma^2)$ involves the error variance and the regression parameters for students who did not train for the test, whereas the tuning parameter vector $\theta = (\theta_0, \ldots, \theta_{d-1})$ involves the regression parameters that correspond to the effect of preparing for the test. The unconditional distribution of the response is normal, $Y \sim N(\mu, V)$, with

$$\begin{aligned} \mu &= \mu(\theta, \xi, t) = (\xi_0 + t\theta_0) + \textstyle\sum_{j=1}^{d-1} (\xi_j + t\theta_j) m_j, \\ V &= V(\theta, \xi, t) = \sigma^2 + \textstyle\sum_{j,k=1}^{d-1} (\xi_j + t\theta_j)(\xi_k + t\theta_k)\sigma_{jk}. \end{aligned}$$

Therefore, the probability that a randomly chosen student that studied for the test for a period of length t passes is

$$P(A) = P_{\theta\xi t}(A) = P(Y \geq f_0) = 1 - \Phi\left(\frac{f_0 - \mu}{\sqrt{V}}\right), \tag{69}$$

where Φ is the cumulative distribution function of a standard normal distribution. The null distribution $P_0 = P_{0\xi}$ corresponds to putting $t = 0$ in (69). Thus, the actinfo

$$I^+ = I^+(\theta, \xi, t) = \log \frac{1 - \Phi\left((f_0 - \mu(\theta, \xi, t)/\sqrt{V(\theta, \xi, t)}\right)}{1 - \Phi\left((f_0 - \mu(0, \xi, 0)/\sqrt{V(0, \xi, 0)}\right)} \tag{70}$$

quantifies how much learning, during a period of length t, increases the probability of passing the test. To compute an estimate $\hat{I}^+$ of I^+ in (70), estimates $\hat{\xi}$ and $\hat{\theta}$ of ξ and θ are needed. This can be achieved by collecting two training samples, as in (63). Another option is to compute the least squares estimates $(\hat{\xi}, \hat{\theta})$ of the nuisance and the tuning parameters jointly, without bias, from one single dataset $\{(t_i, z_i, y_i)\}_{i=1}^n$, provided that the time periods t_i vary, so that all parameters are identifiable.

Example 3 (Reinforcement learning (RI) [45])**.** Consider an agent whose purpose is to maximize the reward $f(x)$ of a trajectory x that they to some extent will be able to control, for a time period of length t. At each time point u, there are m possible environments $\mathcal{S} = \{s_1, \ldots, s_m\}$ and q possible actions $\mathcal{A} = \{a_1, \ldots, a_q\}$ to take. The state space $\mathcal{X} = \mathcal{A}^t \times \mathcal{S}^{t+1}$ consists of all possible trajectories

$$x = (a_0, \ldots, a_{t-1}, s_0, \ldots, s_t)$$

of environments and actions, where s_u is the environment and a_u the action taken at time u. A corresponding random trajectory is denoted with capital letters

$$X = (A_0, \ldots, A_{t-1}, S_0, \ldots, S_t).$$

If the environment of the system is $S_u = s$ at time u, and action $A_u = a$ is taken, the probability of moving to environment s' is $P_a(s, s') = P(S_{u+1} = s' | S_u = s, A_u = u)$, with an instantaneous reward of $R_a(s, s')$. If future rewards are discounted by a factor γ, the total reward, over a time horizon of length t, is

$$f(x) = \sum_{u=0}^{t} R_{a_u}(s_u, s_{u+1})\gamma^u.$$

Let f_0 be a lower bound for a trajectory's total discounted reward to be acceptable, so that A in (5) is the set of all acceptable trajectories. The agent takes action according to some *policy* to make the expected total reward of a trajectory as large as possible. To this end, consider stationary policies, where the action A_u taken by the agent at each time point u is only determined by the current environment s_u, according to some matrix $\mathbf{\Pi} = (\pi(s, a); s \in \mathcal{S}, a \in \mathcal{A})$ of transition probabilities $\pi(s, a) = P(A_u = u | S_u = s)$. For a completely random policy

$$\pi(s, a) = \xi_a; \quad a = 1, \ldots, q,$$

the action is not influenced by the current environment, and it is completely specified by the vector $\xi = (\xi_1, \ldots, \xi_q)$ of nuisance parameters. Thus, $P_0(A) = P_{0\xi t}(f(X) \geq f_0)$ is the probability that an ignorant agent with policy determined by ξ, will have an acceptable trajectory. An agent who knows the reward function R_a and the dynamics P_a of the environment will try to take this knowledge into account to formulate a policy that makes the reward as large as possible. A deterministic policy $\theta : \mathcal{S} \to \mathcal{A}$ is a function that takes a unique action for each environment, so that

$$\pi(s, a) = 1_{\{a = \theta(s)\}}.$$

Thus, $P(A) = P_{\theta t}(f(X) \geq f_0)$ is the probability that an agent with deterministic policy θ obtains an acceptable trajectory. The active information

$$I^+ = I^+(\theta, \xi, t) = \log \frac{P_\theta(\sum_{u=0}^t R_{A_u}(S_u, S_{u+1})\gamma^u \geq f_0)}{P_{0\xi}(\sum_{u=0}^t R_{A_u}(S_u, S_{u+1})\gamma^u \geq f_0)} \tag{71}$$

quantifies, on a logarithmic scale, how much more likely it is for an agent with policy θ to obtain an acceptable trajectory, compared to an ignorant agent with policy ξ. The values ξ and θ are varied during the exploration phase of RI, but they are assumed to be known during the exploitation phase of RI. Suppose that we want to compute the actinfo (71) during the exploitation phase. Since $P_0(A)$ and $P(A)$ are typically unknown, they have to be estimated by Monte Carlo. To this end, assume we have two samples (31) and (62) of n and n_0 trajectories available, from $Q = P_{\theta t}$ and $Q = P_{0\xi t}$, respectively. Then, $\hat{I}^+$ in (63) can be used to estimate the actinfo (71).

Example 4 (Molecular machines and Moran models [15,30,41])**.** Suppose that Ω consists of all 2^d binary sequences $x = (x_1, \ldots, x_d)$ of length d, with a null distribution $P_0(x)$ that will be chosen below. The specificity function f is defined as

$$f(x) = \begin{cases} a|x|, & x \neq (1, \ldots, 1), \\ 1, & x = (1, \ldots, 1), \end{cases} \tag{72}$$

where $|x| = \sum_{i=1}^d x_i$ and $a \leq 1/d$ is a fixed parameter. We regard x as a molecular machine with d parts, with $x_i = 1$ or 0 depending on whether part i functions or not. The specificity $f(x)$ quantifies how well the machine works, for instance, its ability to regulate activity *in vitro* or *in vivo* in a living cell. It is assumed that $f(x)$ is determined by the number $|x|$ of functioning parts, with a maximal value $f_{\max} = f(1, \ldots, 1) = 1$. Using (8), the most stringent definition of high specificity, it follows that $A = \{(1, \ldots, 1)\}$ only contains one element, a molecular machine for which all parts are in shape. The parameter a is crucial. If $0 < a \leq 1/d$, it follows that a molecular machine works better the more the parts that are in shape. On the other hand, if $a < 0$, then a molecular machine with some parts in shape, but not all, functions worse the more parts are in shape, since all units must work in order for the whole machine to function, and there is a cost $-a$ associated with carrying each part that is in shape, as long as the whole system does not function.

Each state x is interpreted as a *population* of N subjects, all having the same variant x of the molecular machine. With this interpretation, $X = X_t$ is the outcome of a random evolutionary process where all subjects of the population, at any time point t, have the same state. However, this state may vary over time when all subjects of population simultaneously experience the same change. The question of interest is whether this process can modify the population so that all its members have a functioning molecular machine. A transition of this process from x is caused by a mutation with distribution $q(x, \cdot)$, where $q(x, x) = 0$. Suppose a mutation from x to y is possible, i.e., $q(x, y) > 0$. A mutation from x to y first occurs in one individual and then it either (momentarily) dies out with probability $1 - \alpha_\theta(x, y)$ or it (momentarily) spreads to the whole population (becomes fixed) with probability

$$\alpha_\theta(x, y) = C \cdot \left(\frac{e^{\theta f(y)} P_0(y) q(y, x)}{e^{\theta f(x)} P_0(x) q(x, y)} \right)^{1/2}, \tag{73}$$

where

$$C = \left(\max_{x,y} \frac{e^{\theta f(y)} P_0(y) q(y, x)}{e^{\theta f(x)} P_0(x) q(x, y)} \right)^{-1/2} \tag{74}$$

is a constant assuring that (73) never exceeds 1, and the maximum is taken over all x, y such that $x \neq y$ and both of $q(x,y)$ and $q(y,x)$ are positive. The Markov chain with transition probabilities (14) and acceptance probability (73) represent the dynamics of the evolutionary process.

As shown in Section 7, the equilibrium distribution of this Markov chain is given by P_θ in (10). In particular, Propositions 2 and 3 remain valid when the Markov chain (14) with acceptance probabilities (73) are used, rather than (15). We will interpret

$$s(x) = e^{\theta f(x)/N} \tag{75}$$

as the selection coefficient or fitness of individuals with a molecular machine of type x, that is, $s(x)$ is proportional to the fertility rate of individuals of type x.

The MH-type Markov chain with acceptance probability (73) and (74) represents an evolutionary process that closely resembles a Moran model with the selection [46–48], which is frequently used for describing evolutionary processes (as can be seen in Section 7). The Moran model is a continuous time Markov chain for a population with overlapping generations where individuals die at the same rate, and are replaced by the offspring of individuals in the population proportionally to their selection coefficients $s(x)$. New types arise when an offspring of parents of type x mutate with probability $\mu(x)$. If the mutation rate is small ($\mu(x) \ll N^{-1}$ for all $x \in \Omega$), then to a good approximation the whole population will have the same type at any point in time, which is a so-called fixed state assumption.

Even though the Moran model is specified in continuous time, time can be discretized as $t = 0, 1, 2, \ldots$ by only recording the population when individuals die. If individuals die at a rate of 1, then the next individual dies at a rate of N, so that time is counted in units of N^{-1} generations. The fixed state assumption is motivated by assuming that newborn offspring with a new mutation either dies out or spreads to the whole population (becoming fixed in the population) right after birth. In this context, q corresponds to the way in which mutations change the type of the individual, whereas $\alpha_\theta = \alpha_{\theta N}$ is the probability of fixation. If $q(x,y)$ is the conditional probability that an offspring of a type x parent mutates to y, given that a mutation occurs, then the proposal kernel of the Moran model is

$$q^{\text{Moran}}(x,y) = \begin{cases} \mu(x)q(x,y), & x \neq y, \\ 1 - \mu(x), & x = y. \end{cases} \tag{76}$$

As shown in Section 7, the acceptance (or fixation) probability of the Moran model is

$$\alpha_{\theta N}^{\text{Moran}}(x,y) \approx \frac{1}{N}\left(1 + \frac{\theta[f(y) - f(x)]}{2}\right) \approx \frac{1}{N}\left(\frac{e^{\theta f(y)}}{e^{\theta f(x)}}\right)^{1/2} \tag{77}$$

when $\theta[f(y) - f(x)]$ is small. From (76) and (77), the Moran model approximates the Metropolis–Hastings kernel with acceptance probabilities (73) and (74) with good accuracy when (i) $\mu(x) \equiv \mu$; (ii) P_0 is uniform; and (iii) the proposal kernel q is symmetric (i.e., $q(x,y) = q(y,x)$), although the time scales of the two processes are different. More specifically, if (i)–(iii) hold, a time-shifted version of the Moran model approximates the MH-type model with acceptance probabilities (73) and (74), so that each time step of the MH-type Markov chain corresponds to C/μ generations of a Moran model. However, even under assumptions (i)–(iii), the stationary distribution of the Moran model differs slightly from P_θ.

The proposal kernel $q(x,y)$ is assumed to be local and satisfying

$$q(x,y) = \begin{cases} b/[|x| + b(d - |x|)], & y = x + e_j, x_j = 0, \\ 1/[|x| + b(d - |x|)], & y = x + e_j, x_j = 1, \\ 0, & \text{otherwise}, \end{cases} \tag{78}$$

where $e_j = (0, \ldots, 0, 1, 0, \ldots, 0)$ is a row vector of length d with a 1 in position $j \in \{1, \ldots, d\}$ and zeros elsewhere, whereas $x + e_j$ refers to component-wise addition modulo 2, corresponding to a switch of component j of x. A change of component j from 0 to 1 is caused by a beneficial mutation, whereas a change from 1 to 0 corresponds to a deleterious mutation. Consequently, $b > 0$ is the ratio between the rates at which beneficial and deleterious mutations occur.

The kernel q in (78) is symmetric only when beneficial and deleterious mutations have the same rate ($b = 1$). The more general case of asymmetric q is handled differently by the MH-type algorithm and the Moran model. Whereas the MH-type algorithm elevates the acceptance probability (73) of seldom-proposed states y (those y for which $q(x, y)$ is small for many x), this is not the case for the acceptance probability (77) of the Moran model. To avoid that these states y are reached too often by the MH-type algorithm, the null distribution P_0 of no selection has to be chosen so that $P_0(y)$ is small for rarely proposed states (whereas the Moran model needs no such correction). Therefore P_0 in (73) will be chosen as the stationary distribution of a transition kernel (14) for which $\theta = 0$ and all candidates are accepted ($\alpha_0(x, y) = 1$). That is, if $\tilde{\mathbf{\Pi}}_0$ refers to the transition matrix of such a Markov chain, the initial distribution $\mathbf{P}_0$ in (17) is chosen as the solution of

$$\begin{cases} \mathbf{P}_0 = \mathbf{P}_0 \tilde{\mathbf{\Pi}}_0, \\ \sum_{x \in \Omega} P_0(x) = 1. \end{cases} \tag{79}$$

The null distribution $P_0 = P_{0b}$ in (79) involves one single nuisance parameter $\xi = b$. In the special case, when beneficial and deleterious mutations have the same rate ($b = 1$), this procedure generates a uniform distribution $P_0(x) \equiv 2^{-d}$. On the other hand, states x with many functioning parts will be harder to reach by the Markov process $\tilde{\mathbf{\Pi}}_0$ when beneficial mutations occur less frequently than deleterious ones ($0 < b < 1$), resulting in smaller values of $P_0(x)$. The distribution under the alternative hypothesis, $P = P_{\tilde{\theta} b t}$, involves the nuisance parameter b, the time point t at which the state of the population is recorded, and $\tilde{\theta} = (a, \theta)$, the two parameters that determine how much background information the MH-type evolutionary algorithm makes use of. For simplicity, a and b are here regarded as constants and we only include θ and t in the notation. This gives rise to an active information

$$I^+(\theta, t) = \log \frac{P_\theta(X_t = (1, \ldots, 1))}{P_0(X_t = (1, \ldots, 1))}. \tag{80}$$

The MH-type algorithm is studied for $d = 5$, and illustrated in Figures 1–3. Note that the functional information I_{f0} is a decreasing function of b, since it is more surprising to find a working molecular machine by chance when the rate of beneficial mutations b is small. Moreover, the active information $I^+(\theta) = \lim_{t \to \infty} I^+(\theta, t)$ for the equilibrium distribution of the Markov chain as well as the active information $I^+(\theta, t)$ and $I_s^+(\theta, t)$ for a system in non-equilibrium, without and with stopping, are increasing functions of θ, and decreasing functions of a and b. The smaller a or b is, the more external information can be infused to increase the probability of reaching the fine-tuned state of a working molecular machine $(1, \ldots, 1)$. When a is small, to leave this state once it is reached becomes more difficult, and consequently $I_s^+(\theta, t)$, is only marginally larger than $I(\theta, t)$.

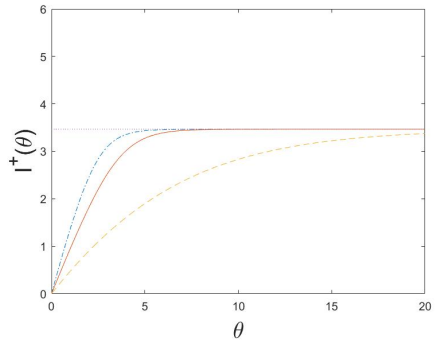

Figure 1. Plot of $I^+(\theta) = \lim_{t\to\infty} I^+(\theta, t)$ in (80) as a function of θ for a system of molecular machines with transition kernel (73), proposal distribution (78), and null distribution (79). The system has $d = 5$ components, $b = 1.0$, and $a = -0.2$ (dash−dotted), $a = 0$ (solid) and $a = 0.2$ (dashed). The horizontal dotted line corresponds to the functional information $I_{f0} = 3.47$.

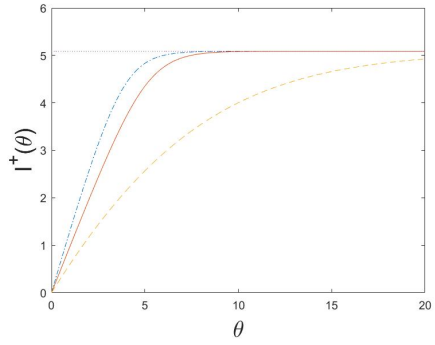

Figure 2. Plot of $I^+(\theta) = \lim_{t\to\infty} I^+(\theta, t)$ in (80) as a function of θ for a system of molecular machines with transition kernel (73), proposal distribution (78), and null distribution (79). The system has $d = 5$ components, $b = 0.5$, and $a = -0.2$ (dash−dotted), $a = 0$ (solid), and $a = 0.2$ (dashed). The horizontal dotted line corresponds to the functional information $I_{f0} = 5.09$.

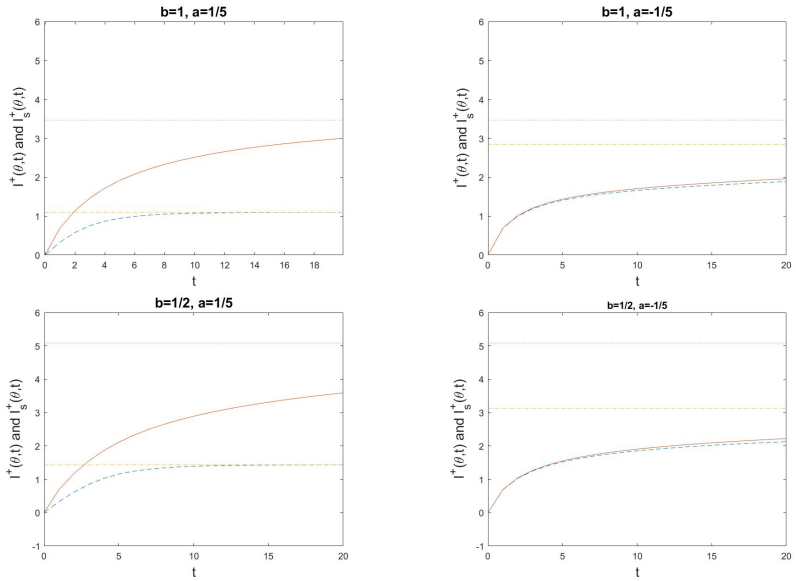

Figure 3. Plot of $I^+(\theta, t)$ in (80) (dashed) and $I_s^+(\theta, t)$ (solid) as a function of t for a system of molecular machines with transition kernel (73), proposal distribution (78), and null distribution (79). The system has $d = 5$ components and $\theta = 2.5$. The upper (lower) row corresponds to $b = 1$ ($b = 0.5$), whereas the left (right) column corresponds to $a = 0.2$ ($a = -0.2$). The horizontal lines in each figure illustrate $I^+(\theta)$ (dash−dotted) and the functional information I_{f0} (dotted).

Example 5 (Evolutionary programming algorithms)**.** Suppose that $\Omega = \Omega_{\text{ind}}^N$ is a set of genetic variants from some genomic region, $x = (x_1, \ldots, x_N)$, for the members of a population of size N. That is, $x_k \in \Omega_{\text{ind}}$ is the variant of this genomic region for individual k. If, for instance, the region codes for the molecular machine of Example 4, we let $x_k = (x_{k1}, \ldots, x_{kd}) \in \{0,1\}^d = \Omega_{\text{ind}}$, with $x_{kj} = 1$ or 0 depending on whether component j of this machine works for individual k. Let $g(x_k)$ be the biological fitness, or the expected number of offspring, of k. In the context of molecular machines, the logarithm of $g(x_k)$ could be a function of the number of functioning parts of a machine of type x_k. The specificity function of a population in state x is the average fitness

$$f(x) = \frac{1}{N}\sum_{k=1}^{N} g(x_k)$$

of its individuals. The targeted set A in (5) corresponds to all genetic profiles with an average fitness at least f_0. This type of model is frequently used in genetic programming as well as in other types of evolutionary programming algorithms to mimic the evolution of N individuals over time [49,50]. Typically, the output $X = X_t$ of the evolutionary algorithm is the last step of a simulation $\{X_s = (X_{s1}, \ldots, X_{sN})\}_{s=0}^t$ of the population over t generations. Once the distributions $P_0 = P_{0\xi t}$ and $P = P_{\theta\xi t}$ of X are found under the null hypothesis H_0 and the alternative hypothesis H_1, the actinfo I^+ can be computed, according to (1). This actinfo quantifies, on a logarithmic scale, how much more likely it is for the average fitness of the population to exceed f_0 at time t, for a population with externally infused information (H_1) compared to an evolutionary process where no such external information is used (H_0). For instance, if a molecular machine needs all its parts in order to function ($g(x_k) = 1(|x_k| = d)$), then the actinfo at time t equals

$$I^+ = I^+(\theta, \xi, t) = \log \frac{P_{\theta\xi t}(|\{k;\, 1 \le k \le N, X_k = (1, \ldots, 1)\}| \ge f_0 N)}{P_{0\xi t}(|\{k;\, 1 \le k \le N, X_k = (1, \ldots, 1)\}| \ge f_0 N)}, \tag{81}$$

with $X = (X_1, \ldots, X_N)$. Since the state space Ω is very large, it is often complicated to find explicit, analytical expressions for the actinfo I^+ in (81). Suppose that the nuisance parameters ξ of the null distribution $P_0 = P_{0\xi}$ are known. This makes the framework of Section 4.1 applicable, running the evolutionary algorithm n times. That is, n i.i.d. copies $\{X_{is}\}_{s=0}^t$ of the population trajectory are generated up to time t for $i = 1, \ldots, n$. Then, $X_i = X_{it} = (X_{it1}, \ldots, X_{itN})$, $i = 1, \ldots, n$, are used for computing an estimate $\hat{I}_n^+$ of the actinfo, and test for fine-tuning, according to Section 4.1.

Recall the fixed state assumption of Example 4, whereby all individuals of the population, at any time point, have the same state. Such an assumption is only realistic when $N\mu \ll 1$, that is, when either the mutation rate μ and/or the population size N is small. This corresponds to a scenario where P_0 and P put all their probability masses along the diagonal

$$\Omega_{\text{diag}} = \{x \in \Omega;\, x_1 = \ldots = x_N\} \tag{82}$$

of Ω. Since (82) is equivalent to the reduced state space Ω_{ind}, the fixed state assumption greatly simplifies the analysis. For instance, it often makes it possible to find analytical expressions for the actinfo I^+, rather than having to estimate it.

6. Discussion

In this article, a general statistical framework is provided for using active information to quantify the amount of pre-specified external knowledge an algorithm makes use of, or equivalently, how tuned the algorithm is. The theory is based on quantifying, for each state x, how specified it is by means of a real-valued function $f(x)$. An algorithm with external information either directly makes use of knowledge of f, or at least it incorporates knowledge that tends to move the output of the algorithm towards more specified regions. The Metropolis–Hastings Markov chain directly incorporates knowledge of f in terms of

the acceptance probability of proposed moves. The learning ability of this algorithm was analyzed by studying its active information, with or without stopping, when the targeted set of highly specified states is reached. When the independent outcomes of an algorithm are available, nonparametric and parametric estimators of the actinfo of the algorithm were also developed, as well as nonparametric and parametric tests of FT.

This work can be extended in different ways. A first extension is to find conditions under which the actinfo $I^+(\theta, t)$ of a stochastic algorithm based on a random start (according to the null distribution of a non-guided algorithm) followed by t iterations of the Metropolis–Hastings Markov chain (without stopping) is a non-decreasing function of t. We conjecture that this is typically the case but have not obtained any general conditions on the distribution q of proposed candidates for this result to hold.

A second extension is to widen the notion of specificity, so that not only the functionality $f(x)$ but also the rarity $P_0(x)$ of the outcome x under the null distribution is taken into account. A class of such specificity functions is

$$g_\theta(x) = \theta f(x) - \log P_0(x), \tag{83}$$

where $\theta > 0$ is a parameter that controls the tradeoff between scenarios where either functionality or rarity under the null is the most important determinant of specificity. The case $\theta = 0$ in (83) corresponds to the function having no impact, so that $g_0(x)$ reduces to Shannon's self information of x. The case $g_1(x)$ was proposed in [15], whereas $g_\theta(x)$ is solely determined by $f(x)$ in the limit when θ becomes large.

A third extension is to generalize the notion of actinfo to include not only the probability of reaching a targeted set of highly specified states A under H_0 and H_1, but also account for the conditional distribution of the states within A, given that A has been reached. This is related to the way in which *functional sequence complexity* generalizes the functional information [51–54]. Let $H(Q) = -\sum_x Q(x) \log[Q(x)]$ refer to the Shannon entropy of a distribution Q, whereas $H(Q_A)$ is the Shannon entropy of the corresponding conditional distribution $Q_A(x) = Q(x|A)$, given that A has been reached. The functional sequence complexity

$$\begin{aligned} \mathrm{FSC}_0 &= H(P_0) - H(P_{0A}) \\ &= E_{P_0}\{\log[P_0(X \mid A)] \mid X \in A\} - E_{P_0}\{\log[P_0(X)]\} \end{aligned}$$

is the reduction in entropy, under the null hypothesis H_0 of the highly specified states in A, compared to the entropy under H_0 of all states in Ω. FSC_0 then reduces to the functional information I_{f0} when P_0 is uniform over Ω. In a similar vein, the *active uncertainty reduction* is introduced:

$$\begin{aligned} \mathrm{UR}^+ &= \sum_{x \in A} P_A(x) \log P(x) - \sum_{x \in A} P_{0A}(x) \log P_0(x) \\ &= E_P[\log P(X) | X \in A] - E_{P_0}[\log P_0(X) | X \in A]. \end{aligned}$$

Then, $\mathrm{UR}^+ = I^+$ when P_{0A} and P_A are uniformly distributed on A. This happens, for instance, when P_0 has a uniform distribution on Ω and $P = P_\theta$ for some $\theta > 0$, and if (8) holds. The properties of UR^+ deserve to be analyzed in more detail, for instance, by investigating how it differs from the actinfo I^+.

A fourth extension would be to apply the concept of actinfo to other genetic models. For instance, Example 4 is the first time that, to our knowledge, actinfo is applied to the Moran model. In the past, however, actinfo was used in population genetics to study fixation times for the Wright–Fisher model of population genetics, a model for which time is discrete and generations do not overlap [55].

7. Proofs

Proof of Proposition 1. Introduce

$$\begin{aligned} J(\theta) &= \sum_{x \in A^c} \exp\{\theta[f(x) - f(x_0)]\} P_0(x), \\ K(\theta) &= \sum_{x \in A} \exp\{\theta[f(x) - f(x_0)]\} P_0(x), \end{aligned} \tag{84}$$

when Ω is countable, and replace the sums in (84) by integrals when Ω is continuous. Then

$$\begin{aligned} P_\theta(A) &= \exp[\theta f(x_0)] K(\theta) / \{\exp(\theta f(x_0))[J(\theta) + K(\theta)]\} \\ &= K(\theta)/[J(\theta) + K(\theta)] \\ &= 1/[J(\theta)/K(\theta) + 1]. \end{aligned} \tag{85}$$

Since $P_0(A) < 1$, it follows that $J(\theta)$ is a strictly decreasing function of $\theta \geq 0$, whereas $K(\theta)$ is a non-decreasing function of θ. From this, it follows that $P_\theta(A)$ is a strictly increasing function of θ, and consequently $I^+(\theta) = \log[P_\theta(A)/P_0(A)]$ is a strictly increasing function of θ as well.

Moreover, $K(\theta) \geq P_0(A) > 0$ for all $\theta \geq 0$, and $J(\theta) \to 0$ as $\theta \to \infty$ follows by dominated convergence. In conjunction with (85), this implies that $P_\theta(A) \to 1$ and $I^+(\theta) \to I_{f0}$ as $\theta \to \infty$. □

Proof of Proposition 2. Equation (20) follows from (17), (19) and the fact that

$$\begin{aligned} P_0(A) &= \sum_{x \in A} P_0(x) = \mathbf{P}_0 \mathbf{v}, \\ P_{\theta t}(A) &= \sum_{x \in A} P_{\theta t}(x) = \mathbf{P}_{\theta t} \mathbf{v} s. = \mathbf{P}_0 \mathbf{\Pi}_\theta^t \mathbf{v}, \end{aligned}$$

since $\mathbf{v}$ is a column vector of length $|\Omega|$ with ones in positions $x \in A$ and zeros in positions $x \in A^c$.

Equation (21) is equivalent to proving that

$$P_{\theta t}(A) \to P_\theta(A) \text{ as } t \to \infty.$$

However, this follows from the fact that P_θ is the equilibrium distribution of the Markov chain with transition kernel (18). That is, letting $t \to \infty$ in (19), we find that

$$\mathbf{P}_{\theta t} = \mathbf{P}_0 \mathbf{\Pi}_\theta^t \to \mathbf{P}_\theta,$$

and therefore

$$P_{\theta t}(A) = \mathbf{P}_{\theta t} \mathbf{v} s. \to \mathbf{P}_\theta \mathbf{v} s. = P_\theta(A), \text{ as } t \to \infty.$$

□

Proof of Proposition 3. Equation (28) follows from the definitions of $I^+(\theta, t)$ and $I_s^+(\theta, t)$ in (20) and (27), and the fact that

$$P_{\theta t}(A) = P(X_t \in A) \leq P(X_{t \wedge T} \in A) = P_{\theta t s}(A),$$

where the inequality is a consequence of the definition of T in (22). Since

$$P_{\theta t s}(A) = P(T \leq t) \leq P(T \leq t + 1) = P_{\theta, t+1, s}(A),$$

we proved that $I_s^+(\theta, t)$ is non-decreasing in t. Equation (29) follows from the definition of $I_s^+(\theta, t)$ and the fact that

$$\lim_{t \to \infty} P_{\theta t s}(A) = P(T < \infty) = 1. \tag{86}$$

The last equality of (86) is a consequence of the fact that the Markov chain with transition kernel $\mathbf{\Pi}_\theta$ is irreducible, so that any state $x \in \Omega$ will be reached with a probability of 1. In particular, the targeted set A will be reached with a probability of 1. In order to verify (30), we first deduce

$$P(T > t) = 1 - P_0(A)e^{I_s^+(\theta,t)}$$

from (24), and then we make use of the equality

$$E(T) = \sum_{t=0}^{\infty} P(T > t).$$

□

Proof of Proposition 4. Since $n\hat{Q}(A) \sim \text{Bin}(n, Q(A))$ has a binomial distribution, it follows from the central limit theorem that

$$\sqrt{n}(\hat{Q}(A) - Q(A)) \xrightarrow{\mathcal{L}} N(0, Q(A)[1 - Q(A)]), \tag{87}$$

as $n \to \infty$. Notice that $\hat{I}^+ = g(\hat{Q}(A))$, where $g(Q) = \log[Q/P_0(A)]$ and $g'(Q) = 1/Q$. Equation (38) follows from the Delta method (see, e.g., Theorem 8.12 of [33]) and the fact that

$$V = g'(Q(A))^2 \cdot Q(A)[1 - Q(A)].$$

In order to establish (40), to begin with, it follows from (34) and the definition of $p_{\min}$ that

$$\begin{aligned} P_{H_0}(\hat{I}^+ \geq I_{\min}) &= P_{H_0}(\hat{Q}(A) \geq p_{\min}) \\ &= P_{H_0}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i \geq p_{\min}\right), \end{aligned}$$

where $Y_i = I(X_i \in A) \sim \text{Be}(p_0)$ are independent Bernoulli variables under H_0 with success probability $p_0 = P_0(A)$. It follows from the large deviations theory that (40) holds, with

$$C = \sup_{\phi > 0}[\phi p_{\min} - \lambda(\phi)] \tag{88}$$

the Legendre–Fenchel transformation, and

$$\lambda(\phi) = \log E[\exp(\phi Y)] = \log[1 + p_0(e^{\phi} - 1)] \tag{89}$$

the cumulant generating function of Y [56], pp. 529–533. Inserting (89) into (88), it can be seen that the maximum in (88) is given by (41). □

Proof of Proposition 5. In order to verify (46), we will first show that the estimator (42) of the tilting parameter θ is asymptotically normal

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{\mathcal{L}} N(0, U) \text{ as } n \to \infty, \tag{90}$$

with asymptotic variance

$$U = \frac{\text{Var}_Q[f(X)]}{\text{Var}^2_{P_{\theta^*}}[f(X)]}. \tag{91}$$

To this end, let $'$ refer to the derivatives with respect to the tilting parameter θ. Define the score function

$$\psi_\theta(x) = \frac{d \log P_\theta(x)}{d\theta} = \frac{P'_\theta(x)}{P_\theta(x)}$$

and its derivative

$$\psi'_\theta(x) = \frac{d\psi_\theta(x)}{d\theta}.$$

It is a standard result from the asymptotic theory of maximum likelihood estimation and M-estimation (see, e.g., Chapter 6 of [33]) that (90) holds with asymptotic variance

$$U = \frac{\mathrm{Var}_Q[\psi_{\theta^*}(X)]}{E_Q^2[\psi'_{\theta^*}(X)]}. \tag{92}$$

To simplify (92), notice that the score function can be written as

$$\psi_\theta(x) = f(x) - \frac{M'(\theta)}{M(\theta)} = f(x) - E_{P_\theta}[f(X)] \tag{93}$$

for the exponential family of tilted distributions (10) and (11). From this, it follows that

$$\psi'_\theta(x) = \frac{M''(\theta)}{M(\theta)} - \left(\frac{M'(\theta)}{M(\theta)}\right)^2 = \mathrm{Var}_{P_\theta}[f(X)]$$

is a constant, not depending on x. Inserting the last two displayed equations into (92), the formula in (91) for the asymptotic variance of $\hat{\theta}$ is obtained. As a next step, we notice that

$$\hat{I}^+ = g(\hat{\theta}), \tag{94}$$

where

$$g(\theta) = \log \frac{P_\theta(A)}{P_0(A)} = \log h(\theta) - \log P_0(A), \tag{95}$$

and

$$h(\theta) = P_\theta(A) = \frac{\sum_{x\in A} e^{\theta f(x)} P_0(x)dx}{M(\theta)} \tag{96}$$

follows from the definition of $P_\theta(x)$ in (10).

Differentiating (96) with respect to θ, we find that

$$\begin{aligned} h'(\theta) = &\sum_{x\in A} f(x) e^{\theta f(x)} P_0(x)dx / M(\theta) \\ &- M'(\theta) \sum_{x\in A} e^{\theta f(x)} P_0(x)dx / M^2(\theta). \end{aligned} \tag{97}$$

Furthermore, it follows from the RHS of (97) that

$$\begin{aligned} h'(\theta) &= E_{P_\theta}[f(X) I(f(X) \geq f_0)] - P_\theta(A) E_{P_\theta}[f(X)] \\ &= \mathrm{Cov}_{P_\theta}[f(X), I(f(X) \geq f_0)]. \end{aligned} \tag{98}$$

Then, we combine (95) and (97), and obtain

$$g'(\theta) = \frac{h'(\theta)}{h(\theta)} = \frac{\mathrm{Cov}_{P_\theta}[f(X), I(f(X) \geq f_0)]}{P_\theta(A)}. \tag{99}$$

Finally, we use the Delta method to conclude that $\hat{I}^+$ is an asymptotic normal estimator (38) of $I^+(\theta^*)$, with asymptotic variance $V = g'(\theta^*)^2 U$, which, in view of (91) and (99), agrees with (47).

In order to prove the large deviation result (48) for the parametric test of FT, let $\theta_{\min}$ be the value of the tilting parameter that satisfies $P_{\theta_{\min}}(A) = p_{\min} = P_0(A)\exp(I_{\min})$. Then, notice that

$$\begin{aligned} P_{H_0}(\hat{I}^+ \geq I_{\min}) &= P_{H_0}(\hat{Q}(A) \geq p_{\min}) \\ &= P_{H_0}(\hat{\theta} \geq \theta_{\min}) \\ &= P_{H_0}(\sum_{i=1}^{n} \psi_{\theta_{\min}}(X_i)/n \geq 0) \\ &= P_{H_0}\left(\sum_{i=1}^{n} f(X_i)/n \geq E_{p_{\min}}[f(X)]\right), \end{aligned}$$

where, in the third step, we utilized that $\hat{\theta} \geq \theta_{\min}$ is equivalent to the derivative of the log likelihood of data being non-negative at $\theta_{\min}$, and in the fourth step, we made use of (93) and introduced $p_{\min} = P_{\theta_{\min}}$. However, this last line is a large deviations probability. It then follows from a large deviations theory that (48) holds, with C the Legendre–Fenchel transformation in (49). □

Proof of Proposition 6. Since the bias corrected empirical actinfo

$$\hat{I}_n^+ - B = \log \frac{\hat{Q}(A)}{P_{0\tilde{\xi}}(A)} \tag{100}$$

behaves like (34), with $P_0 = P_{0\tilde{\xi}}$, the asymptotic normality result for the nonparametric version of the estimator of I_Q^+ follows from Proposition 4.

For the parametric version of the estimator of I_Q^+, we will (briefly) generalize the asymptotic normality proof of Proposition 5. It follows from (53) and (55) that

$$\hat{I}_n^+ = g(\hat{\theta}, \hat{\xi}),$$

where

$$g(\theta, \xi) = \log \frac{P_{\theta\xi}(A)}{P_{0\max}(A)}. \tag{101}$$

Making use of the delta method, it follows that the asymptotic variance of the parametric version of $\hat{I}_n^+$ equals

$$V = g'(\theta^*, \xi^*)\mathrm{AsVar}(\hat{\theta}, \hat{\xi})g'(\theta^*, \xi^*)^T, \tag{102}$$

with the asymptotic variance of $(\hat{\theta}, \hat{\xi})$ defined through

$$\sqrt{n}\left((\hat{\theta}, \hat{\xi}) - (\theta^*, \xi^*)\right) \xrightarrow{\mathcal{L}} N(0, \mathrm{AsVar}(\hat{\theta}, \hat{\xi}))$$

as $n \to \infty$. Since $(\hat{\theta}, \hat{\xi})$ in (56) is an M-estimator, it follows that its asymptotic variance equals

$$\mathrm{AsVar}(\hat{\theta}, \hat{\xi}) = E[\psi'_{\theta^*\xi^*}(X)]^{-1} E[\psi^T_{\theta^*\xi^*}(X)\psi_{\theta^*\xi^*}(X)] E[(\psi'_{\theta^*\xi^*})^T(X)]^{-1}. \tag{103}$$

The gradient of (101) is

$$g'(\theta, \xi) = \frac{P'_{\theta\tilde{\xi}}(A)}{P_{\theta\tilde{\xi}}(A)} = E[\psi_{\theta\tilde{\xi}}(X) | X \in A], \tag{104}$$

where $\psi_{\theta\tilde{\xi}} = P'_{\theta\tilde{\xi}}(x)/P_{\theta\tilde{\xi}}(x)$ is the likelihood score function for the combined parameter vector (θ, ξ). Putting things together, the asympotic variance formula (59) for the parametric version of $\hat{I}_n^+$ follows from (102)–(104).

The significance level of the FT test can be written as

$$P_{0\xi}(\hat{I}_n^+ \geq I_{\min}) = P_{0\xi}(\hat{I}_n^+ - B \geq I_{\min} - B).$$

Since $p_{\min} = P_{0\xi}(A)\exp(I_{\min})$, we have that

$$I_{\min} - B = \log \frac{p_{\min} e^{-B}}{P_{0\xi}(A)}. \tag{105}$$

From this and (100), it follows that the nonparametric test of FT behaves as the corresponding nonparametric test of Proposition 4, with the null probability $P_0(A)$ replaced by $P_{0\xi}(A)$, and $p_{\min}$ replaced by $p_{\min}e^{-B}$. Therefore, the large deviation result (61) follows from (41). In a similar way, the large deviation result for the parametric version of the FT-test (in the special case when θ is a scalar exponential tilting parameter) follows from (100), (105) and Proposition 5. □

Proof of Proposition 7. Because of (52) and (63), we have that

$$\sqrt{n}(\hat{I}^+_{nn_0} - I^+_Q) = \sqrt{n}\log\frac{\hat{Q}(A)}{Q(A)} - \sqrt{\frac{n}{n_0}}\sqrt{n_0}\log\frac{P_{0\hat{\xi}}(A)}{P_{0\xi}(A)}, \tag{106}$$

where

$$\sqrt{n}\log\frac{\hat{Q}(A)}{Q(A)} \xrightarrow{\mathcal{L}} N(0, V_1) \text{ as } n \to \infty \tag{107}$$

and

$$\sqrt{n_0}\log\frac{P_{0\hat{\xi}}(A)}{P_{0\xi}(A)} \xrightarrow{\mathcal{L}} N(0, V_2) \text{ as } n_0 \to \infty \tag{108}$$

respectively. It follows from the proofs of Propositions 4 and 5 that the asymptotic variance for V_1 in (107) is the same as V in (39) and (59), for the nonparametric and parametric versions of $\hat{Q}(A)$, respectively. The asymptotic variance V_2 in (108) is given by (67). This is proven using the delta method (similarly as for Proposition 6), making use of the fact that $\hat{\xi}$ is the maximum likelihood estimator of ξ with asymptotic variance that is the inverse $E[\psi_\xi^T(X)\psi_\xi(X)]^{-1}$ of the Fisher information matrix. The asymptotic normality result (66) then follows from (106)–(108), the fact that $n/n_0 \to \lambda$, and the independence of the two samples.

The large deviations results are proven in a similar way as in Proposition 6, replacing $P_{0\max}(A)$ by $P_{0\hat{\xi}}(A)$. Using statistical consistency $\hat{\xi} \xrightarrow{p} \xi$ as $n_0 \to \infty$, it follows that the large deviation rates C of Proposition 7, for the nonparametric and parametric versions of the FT tests, are the same as in Proposition 6, with bias term $B = 0$. □

Details from Example 4. In order to prove that the Metropolis–Hastings-type Markov chain (14) with acceptance probabilities (73) has an equilibrium distribution of P_θ, we first notice that for any pair of states $x \neq y$, the flow of probability mass

$$\begin{aligned}
&P_\theta(x)\pi_\theta(x,y)\\
&= P_\theta(x)q(x,y)\alpha_\theta(x,y)\\
&= \frac{P_0(x)e^{\theta f(x)}}{M(\theta)}q(x,y)\cdot C\left[\frac{e^{\theta f(y)}P_0(y)q(y,x)}{e^{\theta f(x)}P_0(x)q(x,y)}\right]^{1/2}\\
&= C\frac{\left(e^{\theta f(x)}P_0(x)q(x,y)e^{\theta f(y)}P_0(y)q(y,x)\right)^{1/2}}{M(\theta)}
\end{aligned} \tag{109}$$

from x to y is symmetric with respect to x and y. Therefore, the flow $P_\theta(y)\pi_\theta(y,x)$ of probability mass in the opposite direction, from y to x, is the same as in (109). A Markov

chain with this property is called *reversible* [57], pp. 11–12. However, it is well known that P_θ is a stationary distribution if the Markov chain is reversible with reversible measure P_θ [58], p. 238. If, additionally, the proposal distribution q is such that it is possible to move between any pair of states in a finite number of steps, it follows that the Markov chain is irreducible and hence that P_θ is its unique stationary distribution, which is also the equilibrium distribution of the Markov chain [58], p. 232.

We will then motivate formula (77) for the acceptance probability of a Moran model. Assume that the population evolves over time as a Moran model, and that all individuals have type x. If one individual mutates from x to y, because of (75), the relative fitness between the $N-1$ individuals of type x and the newly mutated individual of type y is

$$s = \frac{e^{\theta f(y)/N}}{e^{\theta f(x)/N}} = e^{\theta[f(y)-f(x)]/N}. \tag{110}$$

From the theory of Moran models (e.g., [41,59]), it is well known that the fixation probability for the newly mutated individual is

$$\beta_N(s) = \begin{cases} (1-s^{-1})/(1-s^{-N}), & s \neq 1, \\ 1/N, & s = 1. \end{cases} \tag{111}$$

Inserting (110) into (111), we find (when $s \neq 1$, or equivalently when $\Delta = \theta[f(y) - f(x)] \neq 0$) that

$$\beta_N(s) = \frac{1-e^{-\Delta/N}}{1-e^{-\Delta}} \approx \frac{1}{N} \cdot \frac{\Delta}{1-e^{-\Delta}} \approx \frac{1}{N} \cdot (1+\frac{\Delta}{2}),$$

which is equivalent to (77).

Author Contributions: D.A.D.-P. and O.H. contributed equally to all parts of the manuscript, including conceptualization, methodology, writing, review, and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors want to thank two anonymous reviewers for valuable comments that considerably improved the quality of the paper. SDG.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gödel, K. Über Formal Unentscheidbare Sätze der Principia Mathematica und Verwandter Systeme, I. *Monatshefte Math. Phys.* **1931**, *38*, 173–198. [CrossRef]
2. Hofstadter, D.R. *Gödel, Escher, Bach: An Ethernal Golden Braid*; Basic Books: New York, NY, USA, 1999.
3. Whitehad, A.N.; Russell, B. *Principia Mathematica*; Cambridge University Press: Cambridge, UK, 1927.
4. Wolpert, D.H.; MacReady, W.G. *No Free Lunch Theorems for Search*; Technical Report SFI-TR-95-02-010; Santa Fe Institute: Santa Fe, NM, USA, 1995.
5. Wolpert, D.H.; MacReady, W.G. No Free Lunch Theorems for Optimization. *IEEE Trans. Evol. Comput.* **1997**, *1*, 67–82. [CrossRef]
6. Wolpert, D.H. What is important about the No Free Lunch theorems? In *Black Box Optimization, Machine Learning and No-Free Lunch Theorems*; Pardalos, P.M., Rasskazova, V., Vrahatis, M.N., Eds.; Springer: Berlin/Heidelberg, Germany, 2021.
7. Dembski, W.A.; Marks, R.J., II. Bernoulli's Principle of Insufficient Reason and Conservation of Information in Computer Search. In Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics. San Antonio, TX, USA, 11–14 October 2009; pp. 2647–2652. . [CrossRef]
8. Dembski, W.A.; Marks, R.J., II. Conservation of Information in Search: Measuring the Cost of Success. *IEEE Trans. Syst. Man, Cybern. Part Syst. Hum.* **2009**, *5*, 1051–1061. [CrossRef]

9. Hazen, R.M.; Griffin, P.L.; Carothers, J.M.; Szostak, J.W. Functional information and the emergence of biocomplexity. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 8574–8581. [CrossRef] [PubMed]
10. Szostak, J.W. Functional information: Molecular messages. *Nature* **2003**, *423*, 689. [CrossRef] [PubMed]
11. Díaz-Pachón, D.A.; Marks, R.J., II. Generalized active information: Extensions to unbounded domains. *BIO-Complexity* **2020**, *2020*, 1–6. [CrossRef]
12. Díaz-Pachón, D.A.; Sáenz, J.P.; Rao, J.S.; Dazard, J.E. Mode hunting through active information. *Appl. Stoch. Model. Bus. Ind.* **2019**, *35*, 376–393. [CrossRef]
13. Liu, T.; Díaz-Pachón, D.A.; Rao, J.S.; Dazard, J.E. High Dimensional Mode Hunting Using Pettiest Component Analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* 2022, *accepted*. [CrossRef]
14. Montañez, G.D. The famine of forte: Few search problems greatly favor your algorithm. In Proceedings of the 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Banff, AB, Canada, 5–8 October 2017; pp. 477–482. [CrossRef]
15. Montañez, G.D. A Unified Model of Complex Specified Information. *BIO-Complexity* **2018**, *2018*, 1–26. [CrossRef]
16. Díaz-Pachón, D.A.; Sáenz, J.P.; Rao, J.S. Hypothesis testing with active information. *Stati. Probab. Lett.* **2020**, *161*, 108742. [CrossRef]
17. Carter, B. Large Number Coincidences and the Anthropic Principle in Cosmology. In *Confrontation of Cosmological Theories with Observational Data*; Longhair, M.S., Ed.; D. Reidel: Dordrecht, The Netherlands, 1974; pp. 291–298.
18. Barrow, J.D.; Tipler, F.J. *The Anthropic Cosmological Principle*; Oxford University Press: Oxford, UK, 1988.
19. Davies, P. *The Accidental Universe*; Cambridge University Press: Cambridge, UK, 1982.
20. Lewis, G.F.; Barnes, L.A. *A Fortunate Universe: Life In a Finely Tuned Cosmos*; Cambridge University Press: Cambridge, UK, 2016. [CrossRef]
21. Rees, M.J. *Just Six Numbers: The Deep Forces That Shape The Universe*; Basic Books: New York, NY, USA, 2000.
22. Adams, F.C. The degree of fine-tuning in our universe—Furthermore, others. *Phys. Rep.* **2019**, *807*, 1–111. [CrossRef]
23. Barnes, L.A. The Fine Tuning of the Universe for Intelligent Life. *Publ. Astron. Soc. Aust.* **2012**, *29*, 529–564. [CrossRef]
24. Tegmark, M.; Rees, M.J. Why is the cosmic microwave background fluctuation level 10^{-5}. *Astrophys. J.* **1998**, *499*, 526–532. [CrossRef]
25. Tegmark, M.; Aguirre, A.; Rees, M.; Wilczek, F. Dimensionless constants, cosmology, and other dark matters. *Phys. Rev. D* **2006**, *73*, 023505. [CrossRef]
26. Díaz-Pachón, D.A.; Hössjer, O.; Marks, R.J., II. Is Cosmological Tuning Fine or Coarse? *J. Cosmol. Astropart. Phys.* **2021**, *2021*, 020. [CrossRef]
27. Díaz-Pachón, D.A.; Hössjer, O.; Marks, R.J., II. Sometimes size does not matter. *Found. Phys.* 2022, *under revision*.
28. Dingjan, T.; Futerman, A.H. The fine-tuning of cell membrane lipid bilayers accentuates their compositional complexity. *BioEssays* **2021**, *43*, e2100021. [CrossRef]
29. Dingjan, T.; Futerman, A.H. The role of the 'sphingoid motif' in shaping the molecular interactions of sphingolipids in biomembranes. *Biochim. Biophys. Acta BBA Biomembr.* **2021**, *1863*, 183701. [CrossRef]
30. Thorvaldsen, S.; Hössjer, O. Using statistical methods to model the fine-tuning of molecular machines and systems. *J. Theor. Biol.* **2020**, *501*, 110352. [CrossRef]
31. Asmussen, S.; Glynn, P.W. *Stochastic Simulation: Algorithms and Analysis*; Springer: Berlin/Heidelberg, Germany, 2007.
32. Siegmund, D. Importance Sampling in the Monte Carlo Study of Sequential Tests. *Ann. Stat.* **1976**, *4*, 673–684. [CrossRef]
33. Lehmann, E.L.; Casella, G. *Theory of Point Estimation*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 1998.
34. Robert, C.P.; Casella, G. *Monte Carlo Statistical Methods*; Springer: Berlin/Heidelberg, Germany, 2010.
35. Hastings, W.K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **1970**, *57*, 97–109. [CrossRef]
36. Metropolis, N.; Rosenbluth, A.W.; Rosenbluth, M.N.; Teller, A.H. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953**, *21*, 1087–1092. [CrossRef]
37. Kirkpatrick, S.; Gelatt, C.D., Jr..; Vecchi, M.P. Optimization by Simulated Annealing. *Science* **1983**, *220*, 671–680. [CrossRef]
38. Ross, S. *Introduction to Probability Models*, 8th ed.; Academic Press: Cambridge, MA, USA, 2003.
39. Asmussen, R.; Nerman, O.; Olsson, M. Fitting Phase-type Distributions via the EM Algorithm. *Scand. J. Stat.* **1996**, *23*, 419–441.
40. Neuts, M.F. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*; Johns Hopkins University Press Hoboken, NJ, USA, 1981.
41. Hössjer, O.; Bechly, G.; Gauger, A. On the waiting time until coordinated mutations get fixed in regulatory sequences. *J. Theor. Biol.* **2021**, *524*, 110657. [CrossRef] [PubMed]
42. Varadhan, S.R.S. *Large Deviations and Applications*; SIAM: Philadelphia, PA, USA, 1984.
43. Hössjer, O.; Díaz-Pachón, D.A.; Chen, Z.; Rao, J.S. Active information, missing data, and prevalence estimation. *arXiv* **2022**, arXiv:2206.05120. https://doi.org/10.48550/arXiv.2206.05120.
44. Hössjer, O.; Díaz-Pachón, D.A.; Rao, J.S. Active Information, Learning, and Knowledge Acquisition. *PsyArXiv* **2022**. [CrossRef]
45. Kaelbling, L.P.; Littman, M.L.; Moore, A.W. Reinforcement Learning: A Survey. *J. Artif. Intell. Res.* **1996**, *4*, 237–285. [CrossRef]
46. Durrett, R. *Probability Models for DNA Sequence Evolution*; Springer Berlin/Heidelberg, Germany, 2008.
47. Moran, P.A.P. Random processes in genetics. *Math. Proc. Camb. Philos. Soc.* **1958**, *54*, 60–71. [CrossRef]

48. Moran, P.A.P. A general theory of the distribution of gene frequencies—I. Overlapping generations. *Proc. Roy. Soc. Lond. B* **1958**, *149*, 102–112.
49. Mitchell, M. *An Introduction to Genetic Algorithms*; MIT Press: Cambridge, MA, USA, 1996.
50. Vikhar, P.A. Evolutionary algorithms: A critical review and its future prospects. In Proceedings of the 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC), Jalgaon, India, 22–24 Decembe 2016; pp. 261–265.
51. Abel, D.L.; Trevors, J.T. Three subsets of sequence complexity and their relevance to biopolymeric information. *Theor. Biol. Med. Model* **2005**, *2*, 29. [CrossRef] [PubMed]
52. Durston, K.K.; Chiu, D.K.Y. A functional entropy model for biological sequences. Dynamics of Continuous, Discrete & Impulsive Systems, Series B: Applications & Algorithms, Supplement. In Proceedings of the International Conference on Engineering Applications and Compuational Algorithms, Guelph, ON, Canada, 27–29 July 2005; Liu, X., Ed.; pp. 722-725.
53. Durston, K.K.; Chiu, D.K.Y. Functional Sequence Complexity in Biopolymers. In *The First Gene: The Birth of Programming, Messaging and Formal Control*; Abel, D.L., Ed.; LongView Press: New York, NY, USA, 2011; pp. 147–169.
54. Durston, K.K.; Chiu, D.K.Y.; Abel, D.L.; Trevors, J.T. Measuring the functional sequence complexity of proteins. *Theor. Biol. Med. Model* **2007**, *4*, 47. [CrossRef] [PubMed]
55. Díaz-Pachón, D.A.; Marks, R.J., II. Active Information Requirements for Fixation on the Wright-Fisher Model of Population Genetics. *BIO-Complexity* **2020**, *2020*, 1–6. [CrossRef]
56. Kallenberg, O. *Foundations of Modern Probability*, 3rd ed.; Springer Berlin/Heidelberg, Germany, 2021; Volume 2.
57. Popov, S. *Two-Dimensional Random Walk: From Path Counting to Random Interlacements*; Cambridge University Press: Cambridge, UK, 2021. [CrossRef]
58. Grimmett, G.; Stirzaker, D. *Probability and Random Processes*, 3rd ed.; Oxford Univeristy Press: Oxford, UK, 2001.
59. Komarova, N.L.; Sengupta, A.; Nowak, M.A. Mutation-selection networks of cancer initiation: Tumor suppressor genes and chromosomal instability. *J. Theor. Biol.* **2003**, *223*, 433–450. [CrossRef]