Assessing and Testing Fine-Tuning by Means of Active Information

Ola Hössjer and Daniel Andrés Díaz-Pachón

***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. We introduce a function f that quantifies tuning for each possible outcome of a search. It is possible to use f to exponentially tilt the distribution of the outcome of the search algorithm under the null distribution of no tuning, so that a class of distributions is obtained with a parameter θ that quantifies how tuned an algorithm is. We demonstrate that such algorithms can be obtained by iterating a Metropolis-Hastings type of Markov chain. This makes it possible to compute the active information of these algorithms under equilibrium and non-equilibrium of the Markov chain, with or without stopping when the targeted set of fine-tuned states has been reached. 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 by means of a population genetic example with molecular machines, which is related to a Moran model.**

***Index Terms*—Active information, exponential tilting, fine-tuning, functional information, large deviations, Markov chains, Metropolis-Hastings, Moran model, statistical estimation and testing.**

I. Introduction

WHEN Gödel published his incompleteness theorems [1], there was a commotion in the mathematical world from which it has neither yet recovered nor fully assimilated the consequences [2]. Hilbert's program and the mammoth *Principia Mathematica* of Bertrand Russell and Alfred North Whitehead were shattered to pieces by the implication that no finite set of axioms in a formal system can prove all its true statements, including its own consistency. In similar but lesser scale, when David Wolpert and William MacReady published their No Free Lunch Theorems (NFLTs, [3], [4]), 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 [5], throwing away the dream of a "theory of everything" in machine learning. One of the original conclusions of Wolpert and MacReady was that it was necessary to incorporate "problem-specific knowledge into the behavior of the algorithm" [4]. Thus active information (actinfo) was introduced in order to measure the amount of information carried by such problem-specific knowledge [6], [7]. 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 was originally defined as

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

where $A \subset \Omega$ is a non-empty target, 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. 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

$$\begin{aligned} H_0 &: X \sim P_0, \\ H_1 &: X \sim P, \end{aligned} \tag{2}$$

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

When the sample space Ω is finite or a bounded subset of a Euclidean space, it is well known that the uniform distribution maximizes Shannon entropy. However, the uniform distribution is not a feasible choice of P_0 for unbounded samples spaces. For this reason actinfo has been generalized to deal with unbounded spaces [8], by choosing P_0 to maximize Shannon entropy under side constraints, such as existence of various moments. Actinfo has also been used for mode detection [9], [10] in unsupervised learning, among other applications. In one recent development, based on previous work by Montañez [11], Díaz-Pachón, Sáenz, and Rao made actinfo a tool for hypothesis testing [12]. More specifically, they regarded P as a random measure, so that the actinfo quantity I^+ is random as well, and found expressions for the tail probability of I^+.

In this article we will use actinfo to test the presence of and estimate the degree of fine-tuning (FT). FT was introduced by Carter [13] in physics and cosmology. 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., [14]–[17]) and scientific articles (e.g., [18]–[21]). For a given constant of nature, the FT problem can be divided into two stages:

O. Hössjer is with the Department of Mathematics, Stockholm University, 106 91 Stockholm, Sweden (e-mail: ola@math.su.se)

D. A. Díaz-Pachón is with the Division of Biostatistics, University of Miami, Don Soffer Clinical Research Center, 1120 NW 14th St, Miami Fl, 33136, USA (e-mail: Ddiaz3@miami.edu)

(i) Establishing the life-permitting interval (LPI) that allows the existence of life for the constant.
(ii) Determining the probability of such LPI.

As Adams noticed though, there have been great advances in step (i) but step (ii) is not as well-developed [21]. One of the problems with step (ii) is that most attempts have placed the LPI's in finite subspaces of the Euclidean space, where Bernoulli's Principle of Insufficient Reason (PoIR) does not operate [22], [23]. It was therefore necessary to find a different approach, similar to the one used when actinfo was generalized to unbounded spaces, removing the burden from the PoIR and placing it on the maximum entropy principle [8]. Using also Bayes theory, such was the strategy adopted in [24].

In [24] it was also observed that step (i) belongs to physics, while step (ii) is mathematical. By this is meant that it is physical theories that are going to determine what is the length of the LPI, while finding the probability of the LPI is mainly a mathematical task. This observation has at least two implications. First, it allows to see the connection between FT and search problems. In fact, the LPI for a particular constant of nature is a particular case of a target A in (1), whereas X is the value of this constant of nature for a randomly generated universe. Therefore, techniques used to analyze search problems can also be used for FT. We exploit such a connection in this article, providing a general framework for using actinfo in order to assess the degree of FT. The second implication is that FT can be applied to other realms of science. For instance in biology, Dingjan and Futerman have already explored the FT of cell membranes [25], [26]. Following [27], we continue here this trend by illustrating our theoretical developments with biological applications.

Our article is organized as follows. In Section II we introduce a function f such that $f(x)$ tells how specified each state $x \in \Omega$ is, or how tuned the output x of an algorithm is. This is used in order to interpret I^+ in (1) as a test statistic of fine-tuning that quantifies how much a search algorithm reduces functional information. Then in Section III we introduce a class of probability distributions that involves a parameter θ that quantifies how much problem-specific knowledge has been infused into the search algorithm. We demonstrate that it is possible to obtain such a search algorithm by means of a Metropolis-Hastings type of Markov chain. In Sections IV-V we evaluate the corresponding actinfo under equilibrium and non-equilibrium of this Markov chain, with our without stopping when the targeted set of fine-tuned states has been reached. Nonparametric and parametric estimators of actinfo and tests of FT are proposed in Section VI, when n repeated and independent outputs of the search algorithm are available. In particular, we use large deviations theory to prove that the significance levels of these tests, i.e. the probability to detect FT under H_0, goes to zero at an exponential rate when the sample size n increases. In Section VII we present a population genetics example of molecular machines that is related to the Moran model. A discussion in Section VIII concludes. Proofs of results are gathered in Section IX.

II. Active information and fine-tuning

Consider a function $f : \Omega \to \mathbb{R}$, and assume that the objective of the search algorithm 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 $f(x)$. We will also say that the larger $f(x)$ is, the more *tuned* an algorithm with outcome x is in terms of finding pre-specified states of large f. For this reason $f(x)$ will also be referred to as the degree of tuning associated with x. It is further assumed that the target set in (1) has the form

$$A = \{x \in \Omega;\ f(x) \geq f(x_0)\}. \tag{3}$$

This implies that the purpose of the search algorithm is to find highly specified states. For instance, in cosmology x corresponds to the value of a particular constant of nature, whereas f is a binary function such that $f(x) = 1$ or 0 depending on whether x permits a universe with life or not. From this it follows that, if $f(x_0) = 1$, 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 a second example from biology, we take x to be an amino acid sequence, $f(x)$ is the functionality of the protein that the amino acid corresponds to, and X is the outcome of a random evolutionary process, the goal of which is to generate a functioning protein. This process either makes use of external information (H_1) or not (H_0). In Section VII we give a more refined biological example, where x corresponds to a protein complex or a molecular machine.

There are at least two ways of interpreting x_0. 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 X is another random variable that represents a second (possibly future) search, independent of X'. Then, if we condition on the outcome x_0 of the first search,

$$\begin{aligned} I^+ &= \log P(A) - \log P_0(A) \\ &= I_{f0} - I_f \end{aligned} \tag{4}$$

is the log likelihood ratio for the event that the second search variable X is *at least as tuned* as the observed value $f(x_0)$ of the first search. The corresponding p-value $P_0(A) = P_0(f(X) \geq f(x_0))$ was used in [27] as a measure of FT. It is closely related to the *functional information* $I_{f0} = -\log P_0(A)$ that appears in the last line of (4), and it represents the number of nats of information, under H_0, for X to function at least as well as x_0 (i.e. $f(X) \geq f(x_0)$). This quantity was proposed by Szostak and collaborators in the context of finding amino acid sequences x that correspond to functional proteins [28], [29]. The corresponding functional information $I_f = -\log P(A)$ under H_1 is analogously interpreted as the number of nats of information that the event $f(X) \geq f(x_0)$ mediates under H_1. If $I^+ > 0$, so that $X \in A$ is more likely to occur under H_1 compared to H_0, then actinfo quantifies how much more functional information is attained by observing $X \in A$ under H_0 than under H_1. In particular, when $X \in A$ is observed with certainty under H_1, so that

$I_f = 0$, then actinfo coincides with the functional information under H_0.

It is not necessary though to associate x_0 in (3) with a first search variable X'. Instead we may use some *apriori information* in order to define which values of f represent a high amount of tuning. This gives rise to the second interpretation of x_0, according to which x_0 is used for defining outcomes that are finely and coarsely tuned respectively, with $f_0 = f(x_0)$ a lower bound of FT. According to this interpretation, the two sets A in (3) and its complement

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

represent a dichotomization of tuning, so that A and A^c consist of all states that are finely or coarsely tuned, 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.

$$f_{\max} = \max_{x\in\Omega} f(x) < \infty. \tag{5}$$

Then the most stringent definition of FT

$$f_0 = f_{\max}, \tag{6}$$

only regards outcomes with a maximal degree of tuning as fine-tuned.

To be strict, $x \in A$ is only a necessary condition, but not sufficient, for x to be fine-tuned. Following [24] and [27], FT also requires that $P_0(A)$ is small, or equivalently, that the functional information of A is large. For simplicity of presentation we will however simply speak of A as the set of fine-tuned states, with an implicit understanding that $P_0(A)$ is small.

III. Active information for systems in equilibrium

In order to calculate I^+ we need to specify P_0 and P, the distributions of the random search algorithm under H_0 and H_1, respectively. 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, such was the strategy implemented in [24]. 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, for instance an evolutionary process with no external input.

Although the choice of P is problem specific, one option is to define it as an exponentially tilted version $P = P_\theta$ of P_0 for some $\theta > 0$. Exponential tilting is often used for rare events simulation [30], [31]. Here we use f to define the tilted version of P_0 as

$$P_\theta(x) = \frac{e^{\theta f(x)}}{M(\theta)} P_0(x), \tag{7}$$

with

$$M(\theta) = \sum_{x\in\Omega} e^{\theta f(x)} P_0(x) \tag{8}$$

a normalizing constant assuring that P_θ is a probability measure. For finite sample spaces Ω, we interpret $P_0(x)$ and $P_\theta(x)$ as probabilities, whereas for continuous sample spaces they are probability densities, and then the sum in (8) is replaced by an integral. The larger the tilting parameter $\theta > 0$ is, the more the probability mass of P_θ concentrates on regions of large f. In particular, P_∞ is supported on the set

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

whenever (5) holds.

The parametric family

$$\mathcal{P} = \{P_\theta;\, \theta \geq 0\} \tag{9}$$

of distributions is an exponential family [32, 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:

Proposition 1. *Suppose the target set A is defined as in* (3) *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*

$$I^+(\theta) = \log \frac{P_\theta(A)}{P_0(A)} \tag{10}$$

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, 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.

Inspired by Markov Chain Monte Carlo methods [33], it is possible to define an irreducible 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 (2) when $\theta > 0$), we may interpret $X = X_t$ as an outcome after t iterations of the Markov chain, provided t is so large that equilibrium has been reached. If the Markov chain has an equilibrium distribution (7), it will favor jumps towards regions of large f when $\theta > 0$, more so the higher the value of θ is. In more detail, the transition kernel of the chain is an instance of the well-known Metropolis-Hastings (MH) algorithm [34], [35], which is closely related to simulated annealing [36]. This kernel has a probability or density

$$\pi_\theta(x, y) = r_\theta(x)\delta(x, y) + \alpha_\theta(x, y)q(x, y) \tag{11}$$

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,

$$\alpha_\theta(x, y) = \min\left[1, \frac{e^{\theta f(y)} P_0(y) q(y, x)}{e^{\theta f(x)} P_0(x) q(x, y)}\right] \tag{12}$$

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

$$r_\theta(x) = 1 - \sum_{y\in\Omega} \alpha_\theta(x, y) q(x, y) \tag{13}$$

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 (13) 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 probability $\alpha_\theta(x,Y)$, 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 it is possible to move between any pair of states in Ω in a finite number of steps [37, pp. 243-245].

Notice in particular that 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 probability $P_\theta(Y)/P_\theta(x)$. 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 $x \in \Omega$. It is easy to see that all proposed jumps of this chain are accepted ($\alpha(x,y)=1$), and that $\{X_t\}_{t=1}^\infty$ is a sequence of independent and identically distributed (i.i.d.) random variables with $X_t \sim P_\theta$.

IV. Active information for 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

$$\mathbf{P}_0 = (P_0(x);\, x \in \Omega) \tag{14}$$

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

$$\mathbf{\Pi}_\theta = (\pi_\theta(x,y); x,y \in \Omega) \tag{15}$$

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

$$(P_{\theta t}(x); x \in \Omega) = \mathbf{P}_{\theta t} = \mathbf{P}_0 \mathbf{\Pi}_\theta^t. \tag{16}$$

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 (2). Some basic properties of the corresponding actinfo are summarized in the following proposition:

Proposition 2. *Suppose $X = X_t$ is obtained by iterating t times a Markov chain with initial distribution* (14) *and transition kernel* (15)*. The actinfo then equals*

$$I^+(\theta,t) = \log \frac{P_{\theta t}(A)}{P_0(A)} = \log \frac{\mathbf{P}_0 \mathbf{\Pi}_\theta^t \mathbf{v}}{\mathbf{P}_0 \mathbf{v}}, \tag{17}$$

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

$$\lim_{t\to\infty} I^+(\theta,t) = I^+(\theta). \tag{18}$$

Notice that $I^+(\theta,t) > 0$ corresponds to knowledge of f being used to generate t jumps of the Markov chain, under the alternative hypothesis H_1 in (2).

V. Active information for systems with stopping

In Section IV we assumed that $P \sim P_{\theta t}$ was obtained by starting a random search with null distribution P_0, and then iterating the Markov chain of Section III t times. It is possible though to utilize knowledge of f even more and to stop the Markov chain if the target A in (3) is reached before time t. This can be formalized by introducing the stopping time

$$T = \min\{t \geq 0;\, X_t \in A\} \tag{19}$$

and letting

$$P_{\theta ts}(x) = P(X_{t\wedge T} = x) \tag{20}$$

be the probability distribution of the stopped Markov chain $X_{t\wedge T}$, with the last index s in (20) being an acronym for stopping. In particular,

$$P_{\theta ts}(A) = \sum_{x\in A} P_{\theta ts}(x) = P(T \leq t) \tag{21}$$

is the probability of reaching the target A for the first time after t iterations or earlier. It is possible to use the theory of phase-type distributions in order to compute the target probability $P_{\theta ts}(A)$ in (20) [38], [39]. To this end, we clump all states $x \in A$ into one absorbing state, and decompose the transition kernel in (15) according to

$$\mathbf{\Pi}_\theta = \begin{pmatrix} \mathbf{\Pi}_\theta^{\mathrm{na}} & \mathbf{\Pi}_\theta^{\mathrm{na,a}} \\ \mathbf{0} & 1 \end{pmatrix}, \tag{22}$$

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 (14) to all non-absorbing states. It then follows that

$$P_{\theta ts}(A) = 1 - \mathbf{P}_0^{\mathrm{na}} (\mathbf{\Pi}_\theta^{\mathrm{na}})^t \mathbf{1}, \tag{23}$$

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

After these preliminaries, we are ready to define the actinfo I_s^+ of a search procedure with stopping:

Proposition 3. *Suppose $X = X_t$ is obtained by iterating a Markov chain with initial distribution* (14) *and transition kernel* (15) *(for some $\theta \geq 0$) at most t times, and stopping whenever the set A is reached. Then the actinfo is given by*

$$I_s^+(\theta,t) = \log \frac{P_{\theta ts}(A)}{P_0(A)} = \log \frac{1 - \mathbf{P}_0^{na} (\mathbf{\Pi}_\theta^{na})^t \mathbf{1}}{\mathbf{P}_0 \mathbf{v}}, \tag{24}$$

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 (22) *and* (23)*. This actinfo satisfies*

$$I_s^+(\theta,t) \geq I^+(\theta,t) \tag{25}$$

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

$$\lim_{t\to\infty} I_s^+(\theta,t) = I_{f0} \tag{26}$$

and

$$\sum_{t=0}^{\infty} \left(1 - P_0(A) e^{I_s^+(\theta,t)}\right) = E(T). \tag{27}$$

Inequality (25) states that, for a search procedure with t iterations, knowledge about f that is used for *stopping* the Markov chain in (15) will increase the actinfo, regardless of whether knowledge about f was used ($\theta > 0$) or not ($\theta = 0$) when *iterating* the Markov chain. Equation (26) is a consequence of the fact that the target A is reached eventually 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 (27) 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.

We conclude from Proposition 3 that 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 [40], in the context of gene expression of a number of genes, with x the collection of regulatory regions of all these genes.

VI. Estimating active information and testing fine-tuning

Suppose it is possible to repeat the random search algorithm independently but under the same conditions n times. This corresponds to a sequence $X_1, \ldots, X_n$ of i.i.d random variables $X_i \sim Q$. With repeated experiments, the analogue of the hypothesis testing problem (2) is

$$\begin{array}{ll} H_0: & Q = P_0, \\ H_1: & Q = P. \end{array} \tag{28}$$

Whereas the null distribution P_0 is known, we will assume that P is unknown, apart from the fact that $P(A) > P_0(A)$, so that $I^+ > 0$ in (1). For this reason an estimate $\hat{Q}(A)$ of the target probability $P(A)$ is computed from data, with an associated empirical actinfo

$$\hat{I}^+ = \hat{I}_n^+ = \log \frac{\hat{Q}(A)}{P_0(A)}. \tag{29}$$

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

$$I_Q^+ = \log \frac{Q(A)}{P_0(A)}, \tag{30}$$

which equals 0 under H_0. Let I be the indicator function. The simplest and nonparametric version of the empirical actinfo makes use of the fraction

$$\hat{Q}(A) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \in A) \tag{31}$$

of random searches that fall into A as an estimate of $Q(A)$. In order to test H_0 against H_1 we

$$\text{Reject } H_0 \text{ when } \hat{I}^+ \geq I_{\min}, \tag{32}$$

with a threshold

$$0 < I_{\min} = \log \frac{p_{\min}}{P_0(A)} < I^+ \tag{33}$$

that is the minimal amount of actinfo that the test detects, and with $P_0(A) < p_{\min} < P(A)$ the corresponding lower bound of the target probability that the test detects.

The following result establishes asymptotic normality of the estimator $\hat{I}^+$ and, moreover, the theory of large deviations is used to show that the significance level of the nonparametric test of actinfo goes to zero exponentially fast with n [41]:

Proposition 4. *Suppose the empirical actinfo $\hat{I}^+$ in* (29) *is computed non-parametrically using* (31) *as an estimate of the target probability $Q(A)$. Then $\hat{I}^+$ is an asymptotically normal estimator of I_Q^+ in* (30)*, in the sense that*

$$\sqrt{n}(\hat{I}_n^+ - I_Q^+) \xrightarrow{\mathcal{L}} N(0, V) \text{ as } n \to \infty, \tag{34}$$

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

$$V = \frac{1 - Q(A)}{Q(A)} \tag{35}$$

is the variance of the limiting normal distribution. The significance level of the test (32) *for actinfo satisfies*

$$\lim_{n\to\infty} -\frac{\log\left(P_{H_0}(\hat{I}^+ \geq I_{min})\right)}{n} = C, \tag{36}$$

where I_{min} is the threshold of the test, defined in (33)*, and*

$$C = p_{min}\log\frac{p_{min}}{P_0(A)} + (1 - p_{min})\log\frac{1 - p_{min}}{1 - P_0(A)} \tag{37}$$

is the Kullback-Leibler divergence (41) *between Bernoulli distributions with success probabilities p_{min} and $P_0(A)$ respectively.*

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

Suppose we have a priori knowledge that P is close to the parametric exponential family $\mathcal{P}$ of distributions in (9) for some $\theta > 0$. It is natural then to define a parametric test of actinfo. For this we first need to compute the maximum likelihood estimate

$$\hat{\theta} = \hat{\theta}_n = \arg\max_{\theta \geq 0} \sum_{i=1}^{n} \log P_\theta(X_i) \tag{38}$$

of the tilting parameter θ. This makes it possible to define a parametric estimate

$$\hat{Q}(A) = P_{\hat{\theta}}(A) \tag{39}$$

of the target probability $Q(A)$ that is inserted into (29) in order to define a parametric version of the empirical actinfo $\hat{I}^+$.

To analyze the properties of the estimator (29) and test (32), we first need to introduce

$$\theta^* = \arg\min_{\theta \geq 0} D_{KL}(Q \parallel P_\theta), \tag{40}$$

where

$$D_{KL}(Q \parallel P_\theta) = \sum_{x\in\Omega} Q(x)\log\frac{Q(x)}{P_\theta(x)} \tag{41}$$

is the Kullback-Leibler divergence between Q and P_θ. It follows from (40) that P_{θ^*} is the distribution in $\mathcal{P}$ that best approximates Q. In particular, $\theta^* = \theta$ if $Q = P_\theta$ for some $\theta \geq 0$.

The following proposition shows that $\hat{I}^+$ is an asymptotically normal estimator of $I^+(\theta^*)$ in (10), which differs from I_Q^+ in (30) whenever $Q \notin \mathcal{P}$. Moreover, the proposition also provides large sample properties of the significance level of the test for actinfo:

Proposition 5. *Suppose the empirical actinfo $\hat{I}^+$ in* (29) *is computed parametrically, using an estimate* (39) *of the target probability $Q(A)$. Then $\hat{I}^+$ is an asymptotically normal estimator of $I^+(\theta^*)$, in the sense that*

$$\sqrt{n}\left(\hat{I}_n^+ - I^+(\theta^*)\right) \xrightarrow{\mathcal{L}} N(0, V) \text{ as } n \to \infty, \tag{42}$$

where the variance of the limiting normal distribution is given by

$$V = \frac{Cov^2_{P_{\theta^*}}\left[f(X)I(f(X) \geq f_0)\right] Var_Q\left[f(X)\right]}{P^2_{\theta^*}(A) Var^2_{P_{\theta^*}}\left[f(X)\right]}. \tag{43}$$

Moreover, the significance level of the parametric test for actinfo, based on (29)*,* (33)*, and* (39)*, satisfies*

$$\lim_{n\to\infty} -\frac{\log\left[P_{H_0}\left(\hat{I}^+ \geq I_{min}\right)\right]}{n} = C, \tag{44}$$

where

$$C = \sup_{\phi>0}\left\{\phi E_{P_{min}}[f(X)] - \log M(\phi)\right\}, \tag{45}$$

with $P_{min} = P_{\theta_{min}}$, $\theta_{min} < \theta^*$ is the solution of $P_{\theta_{min}}(A) = p_{min}$, and $M(\phi)$ is defined in (8)*.*

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 (30). 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 $Q \notin \mathcal{P}$. Note that the asymptotic variances in (35) and (43), as well as the rates of exponential significance level decrease in (37) an (45), agree when $Q = P_{\theta^*}$ and $f(x) = f_0 I(x \in A)$, which is a special case of (6).

VII. Example

Assume 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

$$f(x) = \begin{cases} a|x|, & x \neq (1, \ldots, 1), \\ 1, & x = (1, \ldots, 1), \end{cases} \tag{46}$$

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. We assume that $f(x)$ is determined by the number $|x|$ of functioning parts, with a maximal value $f_{\max} = f(1, \ldots, 1) = 1$. Using (6), the most stringent definition of FT, 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 of 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 of the parts that 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.

It is possible to interpret each state x as a *population* of N subjects, all having the same variant x of the molecular machine, and X being the outcome of a random evolutionary process, the purpose of which is to 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 dies out with probability $1 - \alpha_\theta(x, y)$ or it spreads to the whole population (gets fixed) with probability

$$\alpha_\theta(x, y) = C \cdot \left(\frac{e^{\theta f(y)} P_0(y) q(y, x)}{e^{\theta f(x)} P_0(x) q(x, y)}\right)^{1/2}, \tag{47}$$

where

$$C = \left(\max_{x,y} \frac{e^{\theta f(y)} P_0(y) q(y, x)}{e^{\theta f(x)} P_0(x) q(x, y)}\right)^{-1/2} \tag{48}$$

is a constant assuring that (47) 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 (11) and acceptance probability (47) represent the dynamics of the evolutionary process.

We show in Section IX that the equilibrium distribution of this Markov chain is given by P_θ in (7). In particular, Propositions 2–3 remain valid when the Markov chain (11) with acceptance probabilities (47) are used, rather than (12). We will interpret

$$s(x) = e^{\theta f(x)/N} \tag{49}$$

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.

In order to further motivate that the MH-type Markov chain with acceptance probability (47)–(48) represents an evolutionary process, we will show that it closely resembles a Moran model with selection [42]–[44], which is frequently used for describing evolutionary processes. 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 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, a so called fixed state assumption.

Even though the Moran model is specified in continuous time, it is possible to discretize time as $t = 0, 1, 2, \ldots$ by only recording the population when individuals die. If individuals die at rate 1, this means that the next individual dies at rate 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 (get 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

$$q^{\text{Moran}}(x, y) = \begin{cases} \mu(x)q(x, y), & x \neq y, \\ 1 - \mu(x), & x = y. \end{cases} \tag{50}$$

It is shown in Section IX that the acceptance (or fixation) probability of the Moran model is

$$\alpha^{\text{Moran}}_{\theta N}(x, y) \approx \frac{1}{N}\left(1 + \frac{\theta[f(y) - f(x)]}{2}\right) \approx \frac{1}{N}\left(\frac{e^{\theta f(y)}}{e^{\theta f(x)}}\right)^{1/2} \tag{51}$$

when $\theta[f(y) - f(x)]$ is small. It follows from (50)-(51) that the Moran model approximates the Metropolis-Hastings kernel with acceptance probabilities (47)-(48) 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 (47)-(48), 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.

We will assume that the proposal kernel $q(x, y)$ is local and satisfies

$$q(x, y) = \begin{cases} b/[|x| + b(d - |x|)], & y = x + e_j, x_j = 0, \\ 1/[|x| + b(d - |x|)], & y = x + e_j, x_j = 1, \\ 0, & \text{otherwise}, \end{cases} \tag{52}$$

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.

Notice that the kernel q in (52) 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 (47) 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 (51) of the Moran model. In order for the MH-type algorithm to avoid that these states y are reached too often, 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 models needs no such correction). We will therefore choose P_0 in (47) as the stationary distribution of a transition kernel (11) 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, we choose the initial distribution $\mathbf{P}_0$ in (14) as the solution of

$$\begin{cases} \mathbf{P}_0 = \mathbf{P}_0\tilde{\mathbf{\Pi}}_0, \\ \sum_{x\in\Omega} P_0(x) = 1. \end{cases} \tag{53}$$

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 $P_0(x)$.

Here we will study the case $d = 5$, as 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 $I^+(\theta)$ for the equilibrium distribution of the Markov chain as well as the active informations $I^+(\theta, t)$ 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 in order to increase the probability of reaching the fine-tuned state of a working molecular machine $(1, \ldots, 1)$. When a is small it is more difficult to leave this state once it is reached, and consequently $I^+_s(\theta, t)$ is only marginally larger than $I(\theta, t)$.

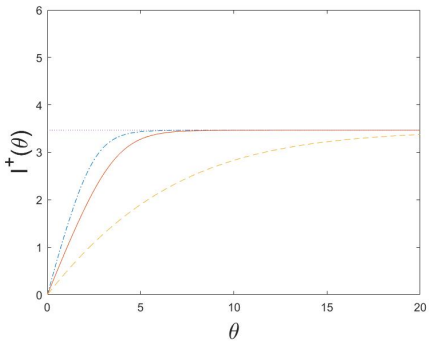

Fig. 1. Plot of $I^+(\theta)$ as a function of θ for a system of molecular machines with $d = 5$ components, $b = 1.0$, and $a = -0.2$ (dash-dotted), $a = 0$ (solid) and $a = 0.2$ (dashed). The horizontal dotted line corresponds to the functional information $I_{f0} = 3.47$.

VIII. Discussion

In this article we provide a general statistical framework for using active information in order to quantify the amount

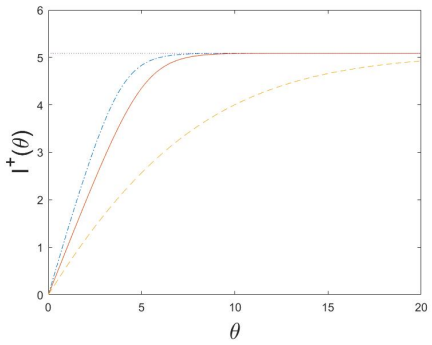

Fig. 2. Plot of $I^+(\theta)$ as a function of θ for a system of molecular machines with $d = 5$ components, $b = 0.5$, and $a = -0.2$ (dash-dotted), $a = 0$ (solid) and $a = 0.2$ (dashed). The horizontal dotted line corresponds to the functional information $I_{f0} = 5.09$.

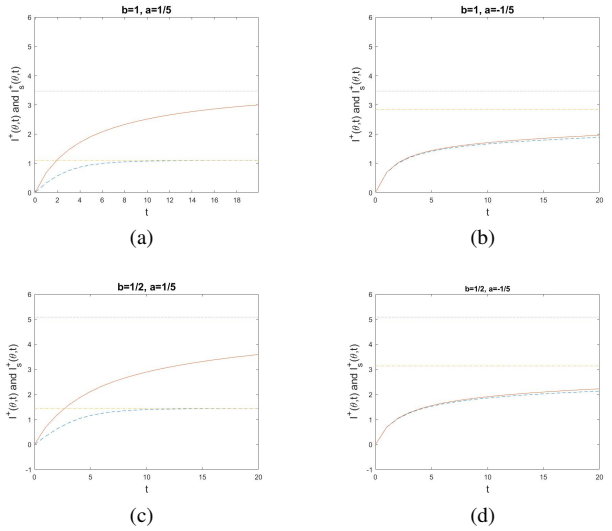

Fig. 3. Plot of $I^+(\theta, t)$ (dashed) and $I_s^+(\theta, t)$ (solid) as a function of t for a system of molecular machines with $d = 5$ components and $\theta = 2.5$. The upper (lower) row corresponds to $b = 1$ ($b = 0.5$), whereas the left (right) column corresponds to $a = 0.2$ and $a = -0.2$.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

of pre-specified external knowledge an algorithm makes use of, or equivalently, how tuned the algorithm is. Our theory is based on quantifying for each state x how specified it is by means of a real-valued function $f(x)$. This makes it possible to introduce an exponential family of distributions for the random outcome of an algorithm, and a corresponding Metropolis-Hastings Markov chain for how the outcome of the algorithm was generated, with or without stopping, when the targeted set of fine-tuned states is reached. We also developed nonparametric and parametric estimators of the actinfo of the algorithm, when independent outcomes of it are available, as well as nonparametric and parametric tests of FT.

As for the example, this is the first time that, to our knowledge, actinfo is applied to the Moran model. In the past though, 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 [45].

It is possible to extend our work in different ways. A first extension would be 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 $f(x)$, so that not only the functionality but also the rarity of the outcome x is taken into account. A class of such specificity functions is

$$g_\theta(x) = \theta f(x) - \log P_0(x), \tag{54}$$

where $\theta > 0$ is a parameter that controls the tradeoff between scenarios where either function or rarity is the most important determinant of specificity. The case $\theta = 0$ corresponds to 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 [11], whereas $g_\theta(x)$ is solely determined by $f(x)$ in the limit when θ gets large.

A third extension is to generalize the notion of actinfo to include not only the probability of reaching a targeted set of fine-tuned 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 functional information [46]–[49]. 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

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

is the reduction in entropy, under the null hypothesis H_0 of the fine-tuned states in A, compared to the entropy under H_0 of all states in Ω. It can be seen that FSC_0 reduces to the functional information I_{f0} when P_0 is uniform over Ω. In a similar vein, we introduce the *active uncertainty reduction*

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

One notices that $\text{UR}^+ = I^+$ 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 (6) holds. It would be of interest to analyze the properties of UR^+ in more detail, for instance investigate how it differs from the actinfo I^+.

A fourth extension is to consider vector-valued objective functions $f : \Omega \to \mathbb{R}^N$, so that

$$f(x) = (f_1(x), \ldots, f_N(x)).$$

This is frequently used in genetic programming [50], as well as for other types of evolutionary programming algorithms

[51], in order to mimic the evolution of N individuals over time. These algorithms typically have a sample space $\Omega = \Omega_{\text{ind}}^N$, where $x = (x_1, \ldots, x_N)$ represents variants of some genomic region for N individuals, and with $x_i \in \Omega_{\text{ind}}$ the variant of this genomic region for individual i. The components $f_i(x) = g(x_i)$ of the objective function are interpreted as the biological fitness $g(x_i)$ for each individual i. Typically, the output $X = X_t$ of the evolutionary algorithm is the last step of a simulation $X_0, \ldots, X_t$ of the population over t generations. If $B \subset \mathbb{R}^N$ corresponds to a targeted set of fitness profiles of the population, the corresponding targeted subset of the sample space is

$$A = \{x \in \Omega;\ f(x) \in B\}. \tag{55}$$

Once the distributions P_0 and P of X are found under the null and alternative hypotheses, it is possible to compute the actinfo I^+ in (1). A typical target set A consists of all populations x for which the average fitness is at least as large as a lower threshold f_0, i.e.

$$B = \{f \in \mathbb{R}^N;\ \bar{f} = \frac{1}{N}\sum_{i=1}^{N} f_i \geq f_0\}. \tag{56}$$

Notice also that the fixed state assumption of Section VII, according to which all individuals have the same genetic variant, corresponds to a scenario where P_0 and P put all their probability masses along the diagonal

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

of Ω. In particular, the target set (55)-(56) reduces to (3) (with $g(x)$ in place of $f(x)$) under such a fixed state assumption.

As a fifth extension, there is sometimes ambiguity in choosing the null distribution of X. For instance, in [24], P_0 was chosen as a maximum entropy distribution with an unknown constraint on its first one or two moments. More generally, if θ_0 is used to parametrize all possible distributions P_{θ_0} under H_0, then

$$\hat{P}_0(A) = \max_{\theta_0} P_{\theta_0}(A)$$

gives a conservative upper bound on the target probability $P_0(A)$ under the null hypothesis. Replacing $P_0(A)$ by $\hat{P}_0(A)$, we thus get a conservative lower bound on the functional information I_{f0} as well as the active information I^+.

IX. Proofs

Proof of Proposition 1:

Introduce

$$\begin{aligned} J(\theta) &= \sum_{x \in A^c} \exp\{\theta[f(x) - f(x_0)]\} P_0(x), \\ K(\theta) &= \sum_{x \in A} \exp\{\theta[f(x) - f(x_0)]\} P_0(x), \end{aligned} \tag{57}$$

when Ω is finite, and replace the sums in (57) by integrals when Ω is continuous. Then

$$\begin{aligned} P_\theta(A) &= \exp[\theta f(x_0)] K(\theta) / \{\exp(\theta f(x_0))[J(\theta) + K(\theta)]\} \\ &= K(\theta)/[J(\theta) + K(\theta)] \\ &= 1/[J(\theta)/K(\theta) + 1]. \end{aligned} \tag{58}$$

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 (58) this implies $P_\theta(A) \to 1$ and $I^+(\theta) \to I_{f0}$ as $\theta \to \infty$. ■

Proof of Proposition 2:

Equation (17) follows from (14), (16) and the fact that

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

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 (18) is equivalent to proving that

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

But this follows from the fact that P_θ is the equilibrium distribution of the Markov chain with transition kernel (15). That is, letting $t \to \infty$ in (16) we find that

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

and therefore

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

■

Proof of Proposition 3:

Equation (25) follows from the definitions of $I^+(\theta, t)$ and $I_s^+(\theta, t)$ in (17) and (24), 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 (19). Since

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

we have proved that $I_s^+(\theta, t)$ is non-decreasing in t. Equation (26) follows from the definition of $I_s^+(\theta, t)$ and the fact that

$$\lim_{t \to \infty} P_{\theta t s}(A) = P(T < \infty) = 1. \tag{59}$$

The last equality of (59) 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 probability 1. In particular, the targeted set A will be reached with probability 1. In order to verify (27), we first deduce

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

from (21), 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

$$\sqrt{n}(\hat{Q}(A) - Q(A)) \xrightarrow{\mathcal{L}} N(0, Q(A)[1-Q(A)]), \tag{60}$$

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 (34) follows from the Delta Method (see, e.g., Theorem 8.12 of [32]) and the fact that

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

In order to establish (36), to begin with, it follows from (29) and (33) 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 Large Deviations theory that (36) holds, with

$$C = \sup_{\phi > 0}[\phi p_{\min} - \lambda(\phi)] \tag{61}$$

the Legendre-Fenchel transformation, and

$$\lambda(\phi) = \log E[\exp(\phi Y)] = \log[1 + p_0(e^{\phi} - 1)] \tag{62}$$

the cumulant generating function of Y [52, pp. 529-533]. Inserting (62) into (61) it can be seen that the maximum in (61) is given by (37). ∎

Proof of Proposition 5: In order to verify (42), we will first show that the estimator (38) of the tilting parameter θ is asymptotically normal

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{\mathcal{L}} N(0, U) \text{ as } n \to \infty, \tag{63}$$

with asymptotic variance

$$U = \frac{\text{Var}_Q[f(X)]}{\text{Var}^2_{P_{\theta^*}}[f(X)]}. \tag{64}$$

To this end, let $'$ refer to 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 [32]) that (63) holds with asymptotic variance

$$U = \frac{\text{Var}_Q[\psi_{\theta^*}(X)]}{E^2_Q[\psi'_{\theta^*}(X)]}. \tag{65}$$

To simplify (65), notice that the score function can be written as

$$\psi_\theta(x) = f(x) - \frac{M'(\theta)}{M(\theta)} = f(x) - E_{P_\theta}[f(X)] \tag{66}$$

for the exponential family of tilted distributions (7)-(8). From this it follows that

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

is a constant, not depending on x. Inserting the last two displayed equations into (65), the formula in (64) for the asymptotic variance of $\hat{\theta}$ is obtained. As a next step we notice that

$$\hat{I}^+ = g(\hat{\theta}), \tag{67}$$

where

$$g(\theta) = \log \frac{P_\theta(A)}{P_0(A)} = \log h(\theta) - \log P_0(A), \tag{68}$$

and

$$h(\theta) = P_\theta(A) = \frac{\sum_{x \in A} e^{\theta f(x)} P_0(x) dx}{M(\theta)} \tag{69}$$

follows from the definition of $P_\theta(x)$ in (7).

Differentiating (69) with respect to θ, we find that

$$\begin{aligned} h'(\theta) = &\sum_{x \in A} f(x) e^{\theta f(x)} P_0(x) dx / M(\theta) \\ &- M'(\theta) \sum_{x \in A} e^{\theta f(x)} P_0(x) dx / M^2(\theta). \end{aligned} \tag{70}$$

And it follows from the RHS of (70) that

$$\begin{aligned} h'(\theta) &= E_{P_\theta}[f(X) I(f(X) \geq f_0)] - P_\theta(A) E_{P_\theta}[f(X)] \\ &= \text{Cov}_{P_\theta}[f(X), I(f(X) \geq f_0)]. \end{aligned} \tag{71}$$

Then we combine (68) and (70), and obtain

$$g'(\theta) = \frac{h'(\theta)}{h(\theta)} = \frac{\text{Cov}_{P_\theta}[f(X), I(f(X) \geq f_0)]}{P_\theta(A)}. \tag{72}$$

Finally we use the Delta Method to conclude that $\hat{I}^+$ is an asymptotic normal estimator (34) of $I^+(\theta^*)$, with asymptotic variance $V = g'(\theta^*)^2 U$, which, in view of (64) and (72), agrees with (43).

In order to prove the large deviation result (44) for the parametric test of FT, let $\theta_{\min}$ be the value of the tilting parameter that satisfies $P_{\theta_{\min}}(A) = p_{\min}$. 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 (66) and introduced $p_{\min} = P_{\theta_{\min}}$. But this last line is a large deviations probability. It then follows from large deviations theory that (44) holds, with C the Legendre-Fenchel transformation in (45). ∎

Details from Section VII:

In order to prove that the Metropolis-Hastings type Markov chain (11) with acceptance probabilities (47) has equilibrium distribution P_θ, we first notice that for any pair of states $x \neq y$, the flow of probability mass

$$\begin{aligned} & P_\theta(x)\pi_\theta(x,y) \\ & = P_\theta(x)q(x,y)\alpha_\theta(x,y) \\ & = \frac{P_0(x)e^{\theta f(x)}}{M(\theta)} q(x,y) \cdot C \left[\frac{e^{\theta f(y)} P_0(y) q(y,x)}{e^{\theta f(x)} P_0(x) q(x,y)} \right]^{1/2} \\ & = C \frac{\left(e^{\theta f(x)} P_0(x) q(x,y) e^{\theta f(y)} P_0(y) q(y,x) \right)^{1/2}}{M(\theta)} \end{aligned} \tag{73}$$

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 (73). A Markov chain with this property is called *reversible* [53, pp. 11-12]. But it is well known that P_θ is a stationary distribution if the Markov chain is reversible with reversible measure P_θ [54, 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 [54, p. 232].

Next we will motivate formula (51) 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 (49), the relative fitness between the $N-1$ individuals of type x and the newly mutated individual of type y is

$$s = \frac{e^{\theta f(y)/N}}{e^{\theta f(x)/N}} = e^{\theta[f(y)-f(x)]/N}. \tag{74}$$

From the theory of Moran models (e.g., [40], [55]), it is well known that the fixation probability for the newly mutated individual is

$$\beta_N(s) = \begin{cases} (1-s^{-1})/(1-s^{-N}), & s \neq 1, \\ 1/N, & s = 1. \end{cases} \tag{75}$$

Inserting (74) into (75) 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 (51). ∎

REFERENCES

[1] K. Gödel, "Über Formal Unentscheidbare Sätze der Principia Mathematica und Verwandter Systeme, I." *Monatshefte für Mathematik und Physik*, vol. 38, pp. 173–198, 1931.

[2] D. R. Hofstadter, *Gödel, Escher, Bach: an Ethernal Golden Braid*. Basic Books, 1999.

[3] D. H. Wolpert and W. G. MacReady, "No Free Lunch Theorems for Search," Santa Fe Institute, Tech. Rep. SFI-TR-95-02-010, 1995.

[4] ——, "No Free Lunch Theorems for Optimization," *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 67–82, 1997. [Online]. Available: https://doi.org/10.1109/4235.585893

[5] D. H. Wolpert, "What is important about the No Free Lunch theorems?" in *Black Box Optimization, Machine Learning and No-Free Lunch Theorems*, P. M. Pardalos, V. Rasskazova, and M. N. Vrahatis, Eds. Springer, 2021.

[6] W. A. Dembski and R. J. Marks II, "Bernoulli's Principle of Insufficient Reason and Conservation of Information in Computer Search," in *Proc. of the 2009 IEEE International Conference on Systems, Man, and Cybernetics. San Antonio, TX*, October 2009, pp. 2647–2652. [Online]. Available: https://doi.org/10.1109/ICSMC.2009.5346119

[7] ——, "Conservation of Information in Search: Measuring the Cost of Success," *IEEE Trans Syst Man Cybern A, Syst Humans*, vol. 5, no. 5, pp. 1051–1061, September 2009. [Online]. Available: https://doi.org/10.1109/TSMCA.2009.2025027

[8] D. A. Díaz-Pachón and R. J. Marks II, "Generalized active information: Extensions to unbounded domains," *BIO-Complexity*, vol. 2020, no. 3, pp. 1–6, 2020. [Online]. Available: https://doi.org/10.5048/BIO-C.2020.3

[9] D. A. Díaz-Pachón, J. P. Sáenz, J. S. Rao, and J.-E. Dazard, "Mode hunting through active information," *Appl Stochastic Models Bus Ind*, vol. 35, no. 2, pp. 376–393, 2019. [Online]. Available: https://doi.org/10.1002/asmb.2430

[10] T. Liu, D. A. Díaz-Pachón, J. S. Rao, and J.-E. Dazard, "High dimensional mode hunting using pettiest component analysis," *Under review*, 2021.

[11] G. D. Montañez, "A Unified Model of Complex Specified Information," *BIO-Complexity*, vol. 4, pp. 1–26, 2018.

[12] D. A. Díaz-Pachón, J. P. Sáenz, and J. S. Rao, "Hypothesis testing with active information," *Stat & Probab Letters*, vol. 161, p. 108742, 2020. [Online]. Available: https://doi.org/10.1016/j.spl.2020.108742

[13] B. Carter, "Large Number Coincidences and the Anthropic Principle in Cosmology," in *Confrontation of Cosmological Theories with Observational Data*, M. S. Longhair, Ed. D. Reidel, 1974, pp. 291–298. [Online]. Available: https://www.doi.org/10.1017/S0074180900235638

[14] P. Davies, *The Accidental Universe*. Cambridge University Press, 1982.

[15] J. D. Barrow and F. J. Tipler, *The Anthropic Cosmological Principle*. Oxford University Press, 1988.

[16] M. J. Rees, *Just Six Numbers: The Deep Forces That Shape The Universe*. Basic Books, 2000.

[17] G. F. Lewis and L. A. Barnes, *A Fortunate Universe: Life In a Finely Tuned Cosmos*. Cambridge University Press, 2016.

[18] M. Tegmark and M. J. Rees, "Why is the cosmic microwave background fluctuation level 10^{-5}," *The Astrophysical Journal*, vol. 499, no. 2, pp. 526–532, 1998. [Online]. Available: https://www.doi.org/10.1086/305673

[19] M. Tegmark, A. Aguirre, M. Rees, and F. Wilczek, "Dimensionless constants, cosmology, and other dark matters," *Phys. Rev. D*, vol. 73, no. 2, p. 023505, 2006. [Online]. Available: https://www.doi.org/10.1103/PhysRevD.73.023505

[20] L. A. Barnes, "The Fine Tuning of the Universe for Intelligent Life," *Publications of the Astronomical Society of Australia*, vol. 29, no. 4, pp. 529–564, 2011. [Online]. Available: https://doi.org/10.1071/AS12015

[21] F. C. Adams, "The degree of fine-tuning in our universe —and others," *Physics Reports*, vol. 807, no. 15, pp. 1–111, May 2019. [Online]. Available: http://www.doi.org/10.1016/j.physrep.2019.02.001

[22] T. McGrew, L. McGrew, and E. Vestrup, "Probabilities and the Fine-Tuning Argument: A Sceptical View," *Mind, New Series*, vol. 110, no. 440, pp. 1027–1037, October 2001. [Online]. Available: https://doi.org/10.1093/mind/110.440.1027

[23] M. Colyvan, J. L. Garfield, and G. Priest, "Problems With the Argument From Fine Tuning," *Synthese*, vol. 145, no. 3, pp. 325–338, 2005. [Online]. Available: https://doi.org/10.1007/s11229-005-6195-0

[24] D. A. Díaz-Pachón, O. Hössjer, and R. J. Marks II, "Is cosmological tuning fine or coarse?" *J Cosmol Astropart Phys*, 2021.

[25] T. Dingjan and A. H. Futerman, "The fine-tuning of cell membrane lipid bilayers accentuates their compositional complexity," *BioEssays*, vol. 43, no. 5, p. e2100021, 2021. [Online]. Available: https://www.doi.org/10.1002/bies.202100021

[26] ——, "The role of the 'sphingoid motif' in shaping the molecular interactions of sphingolipids in biomembranes," *Biochimica et Biophysica Acta (BBA) - Biomembranes*, vol. 1863, no. 11, p. 183701, 2021. [Online]. Available: https://doi.org/10.1016/j.bbamem.2021.183701

[27] S. Thorvaldsen and O. Hössjer, "Using statistical methods to model the fine-tuning of molecular machines and systems," *J Theor Biol*, vol. 501, p. 110352, 2020. [Online]. Available: https://doi.org/10.1016/j.jtbi.2020.110352

[28] J. W. Szostak, "Functional information: Molecular messages functional information: Molecular messages functional information: Molecular messages," *Nature*, vol. 423, p. 689, 2003. [Online]. Available: https://doi.org/10.1038/423689a

[29] R. M. Hazen, P. L. Griffin, J. M. Carothers, and J. W. Szostak, "Functional information and the emergence of biocomplexity," *Proc Natl Acad Sci USA*, vol. 104, no. Suppl 1, pp. 8574–8581, 2007. [Online]. Available: https://doi.org/10.1073/pnas.0701744104

[30] S. Asmussen and P. W. Glynn, *Stochastic Simulation: Algorithms and Analysis*. Springer, 2007.

[31] D. Siegmund, "Importance Sampling in the Monte Carlo Study of Sequential Tests," *Ann Stat*, vol. 4, no. 4, pp. 673–684, 1976. [Online]. Available: https://doi.org/10.1214/aos/1176343541

[32] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, 2nd ed. Springer, 1998.

[33] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*. Springer, 2010.

[34] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970. [Online]. Available: https://doi.org/10.2307/2334940

[35] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, and A. H. Teller, "Equation of state calculations by fast computing machines," *J Chem Phys*, vol. 21, no. 6, pp. 1087–1092, 1953. [Online]. Available: https://doi.org/10.1063/1.1699114

[36] S. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi, "Optimization by Simulated Annealing," *Science*, vol. 220, no. 4598, pp. 671–680, 1983. [Online]. Available: https://doi.org/10.1126/science.220.4598.671

[37] S. Ross, *Introduction to Probability Models*, 8th ed. Academic Press, 2003.

[38] R. Asmussen, O. Nerman, and M. Olsson, "Fitting Phase-type Distributions via the EM Algorithm," *Scand J Stat*, vol. 23, pp. 419–441, 1996. [Online]. Available: https://www.jstor.org/stable/4616418

[39] M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. Johns Hopkins University Press, 1981.

[40] O. Hössjer, G. Bechly, and A. Gauger, "On the waiting time until coordinated mutations get fixed in regulatory sequences," *J Theor Biol*, vol. 524, no. 2021, p. 110657, 2021. [Online]. Available: https://doi.org/10.1016/j.jtbi.2021.110657

[41] S. R. S. Varadhan, *Large Deviations and Applications*. SIAM, 1984.

[42] R. Durrett, *Probability Models for DNA Sequence Evolution*. Springer, 2008.

[43] P. A. P. Moran, "Random processes in genetics," *Math Proc Camb Philos Soc*, vol. 54, no. 1, pp. 60–71, 1958.

[44] ——, "A general theory of the distribution of gene frequencies - I. Overlapping generations," *Proc Roy Soc Lond B*, vol. 149, no. 934, pp. 102–112, 1958. [Online]. Available: https://doi.org/10.1098/rspb.1958.0054

[45] D. A. Díaz-Pachón and R. J. Marks II, "Active Information Requirements for Fixation on the Wright-Fisher Model of Population Genetics," *BIO-Complexity*, vol. 2020, no. 4, pp. 1–6, 2020. [Online]. Available: https://doi.org/10.5048/BIO-C.2020.4.

[46] D. L. Abel and J. T. Trevors, "Three subsets of sequence complexity and their relevance to biopolymeric information," *Theor Biol Med Model*, vol. 2, p. 29, 2005. [Online]. Available: https://doi.org/10.1186/1742-4682-2-29

[47] K. K. Durston and D. K. Y. Chiu, "A functional entropy model for biological sequences," *Dynamics of Continuous, Discrete & Impulsive Systems, Series B Supplement*, 2005.

[48] ——, "Functional Sequence Complexity in Biopolymers," in *The First Gene: The Birth of Programming, Messaging and Formal Control*, D. L. Abel, Ed., 2011, pp. 147–169.

[49] K. K. Durston, D. K. Y. Chiu, D. L. Abel, and J. T. Trevors, "Measuring the functional sequence complexity of proteins," *Theor Biol Med Model*, vol. 4, p. 47, 2007.

[50] M. Mitchell, *An Introduction to Genetic Algorithms*. MIT Press, 1996.

[51] P. A. Vikhar, "Evolutionary algorithms: A critical review and its future prospects," *2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC)*, pp. 261–265, 2016. [Online]. Available: https://doi.org/10.1109/ICGTSPICC.2016.7955308

[52] O. Kallenberg, *Foundations of Modern Probability*, 3rd ed. Springer, 2021, vol. 2.

[53] S. Popov, *Two-Dimensional Random Walk: From Path Counting to Random Interlacements*. Cambridge University Press, 2021.

[54] G. Grimmett and D. Stirzaker, *Probability and Random Processes*, 3rd ed. Oxford Univeristy Press, 2001.

[55] N. L. Komarova, A. Sengupta, and M. A. Nowak, "Mutation-selection networks of cancer initiation: tumor suppressor genes and chromosomal instability," *J Theor Biol*, vol. 223, no. 4, pp. 433–450, 2003. [Online]. Available: https://doi.org/10.1016/s0022-5193(03)00120-6

Ola Hössjer has been Professor of Mathematical Statistics at Stockholm University, Sweden, since 2002. He has done research in statistics and probability theory with applications in population genetics, epidemiology, and insurance mathematics. Hössjer is the author of more than 100 publications, he has supervised 13 PhD students, and in 2009 he received the Gustafsson prize in Mathematics.

Daniel Andrés Díaz-Pachón received his B.S. in Mathematical Statistics at *Universidad Nacional de Colombia*, Colombia (2005); and his PhD in probability theory at *Universidade de São Paulo*, Brazil (2009). In 2011 he moved to the University of Miami, Florida, where he was first a Postdoctoral Associate in Biostatistics (2011–2015), and then became Research Assistant Professor. His research is focused on the intersection of probability theory, statistics, machine learning, and information theory.