Trading Bias for Expressivity in Artificial Learning[⋆]

George D. Montañez, Daniel Bashir, Julius Lauw

AMISTAD Lab
Department of Computer Science, Harvey Mudd College, Claremont, CA 91711, USA
{gmontanez, dbashir, julauw}@hmc.edu

Abstract. Bias, arising from inductive assumptions, is necessary for successful artificial learning, allowing algorithms to generalize beyond training data and outperform random guessing. We explore how bias relates to algorithm flexibility (expressivity). Expressive algorithms alter their outputs as training data changes, allowing them to adapt to changing situations. Using a measure of algorithm flexibility rooted in the information-theoretic concept of entropy, we examine the trade-off between bias and expressivity, showing that while highly biased algorithms may outperform uniform random sampling, they cannot also be highly expressive. Conversely, maximally expressive algorithms necessarily have performance no better than uniform random guessing. We establish that necessary trade-offs exist in trying to design flexible yet strongly performing learning systems.

Keywords: Machine Learning, Search, Algorithmic Bias, Inductive Bias, Entropic Expressivity

1 INTRODUCTION

Assumptions are essential for learning [7, 9, 10]. Unless a learning algorithm is biased towards certain outcomes, it cannot outperform random guessing [10]. However, biased algorithms are less flexible; increasing performance comes at a price. Being predisposed towards some outcomes means being predisposed away from others. The degree to which an algorithm can respond to data and output a variety of different responses is its *expressivity*. We investigate the inherent tension between bias and expressivity in learning algorithms, presenting a number of theorems which show that the two are at odds. Flexible and expressive algorithms can change their outcomes in response to changes in data, but highly flexible algorithms cannot widely deviate in performance from uniform random sampling. Conversely, highly biased algorithms can be successful on only a narrow set of problems, limiting their expressivity. Biased algorithms are specialized and our work explores the costs of this specialization in terms of reduced flexibility.

We build on existing research in theoretical machine learning viewing machine learning, AI, and optimization as black-box search processes [8], allowing us to prove

[⋆] This work was supported in part by a generous grant from the Walter Bradley Center for Natural and Artificial Intelligence.

theorems simultaneously applying to many different types of learning, including classification, regression, unsupervised clustering, and density estimation. Within the algorithmic search framework, we define a form of expressivity, *entropic expressivity*, which measures the information-theoretic entropy of an algorithm's induced probability distribution over its search space. An algorithm with high entropic expressivity will spread its probability mass more uniformly on the search space, allowing it to sample widely and without strong preference within that space, displaying flexibility. Conversely, an algorithm with low entropic expressivity concentrates its mass on few regions of the search space, displaying bias towards those outcomes. No algorithm can be both highly expressive and highly biased.

2 BIAS IN MACHINE LEARNING

The word "bias" used in this paper may call a number of associations to the reader's mind. In this section, we seek to disambiguate between different definitions of bias as they arise in machine learning, clarify how we use the term bias, and discuss how our notion interacts with other definitions.

Definition 1. *(Prejudicial Bias (Tim Jones, IBM)) Prejudicial Bias is a prejudice in favor of or against a person, group, or thing that is considered to be unfair. Such bias can result in disparate outcomes when machine learning algorithms, such as facial recognition systems, behave differently when applied to different groups of people.*

Definition 2. *(Inductive Bias (Tom Mitchell)) In the case of a machine leaning algorithm, Inductive Bias is any basis for choosing one generalization over another, other than strict consistency with the observed training instances.*

Informally, the definition of bias that we will employ in this paper quantifies the predisposition an algorithm has towards certain outcomes over others. This controls how an algorithm interprets data and influences how "well-suited" an algorithm is to a particular task, resulting in performance deviations from uniform random sampling. For example, the naïve Bayes classifier is predisposed towards hypotheses that interpret the training data as a set of conditionally independent inputs, where each feature depends only on its output label.

Our mathematical definition of bias is inspired by Definition 2, and is similar to the notion of bias used in statistical parameter estimation, being the difference between an expected random outcome and a baseline value. Bias, as used in this paper, is quite different from the notion of bias in Definition 1. The machine learning community has paid a great deal of attention to confronting the problematic and unethical consequences of prejudicial bias in our field. We believe any applications that use insights from our work and similar work on bias should reflect the responsible consideration of the potential for prejudicial bias and seek to eliminate, or at least minimize, its impact.

As an illustrative example, prejudicial bias can arise when either the researchers designing a machine learning system or the data used to train that system themselves exhibit problematic bias. For example, in the facial recognition case, a vision model being used to recognize people might be trained on images of those that are mostly of

European descent. As a result, the vision model might have trouble recognizing people of non-European racial backgrounds and skin tones.

We would also like to briefly comment on how our notion of bias interacts with prejudicial bias. As we have discussed, a facial recognition system may not work as well for some groups of people as it does for others either because it chose a hypothesis based on training data that reflects such a bias, because it was predisposed towards prejudicially biased hypotheses in the first place (inductive bias), or some combination of both. Under our definition of bias, the algorithm that is inductively biased towards prejudicially biased hypotheses would indeed be considered more biased than a "baseline" algorithm that is not more likely to select one hypothesis over another. At the same time, and perhaps less intuitively, we would also consider an algorithm strongly predisposed *not* to select a racially biased hypothesis more biased than the baseline, precisely because this reflects the algorithm's predisposition towards certain outcomes (and away from others).

We strongly oppose the use of prejudicial biases in learning systems, and support continued efforts to expose and eliminate them. Knowledge that all nontrivial systems must be biased is some way can help us identify and critically evaluate the biases inherent in our own systems, removing prejudicial biases wherever we find them.

3 RELATED WORK

Our notion of algorithmic expressivity stands among many other measures of expressivity found in the statistical learning literature. Among the most well-established are the Vapnik-Chernovekis (VC) dimension [16], a loose upper bound based on the number of points that can be perfectly classified by a learning algorithm for any possible labeling of the points; the Fat-shattering VC dimension, an extension to the VC dimension developed by Kearns and Schapire that solves the issue of dependence on dimensionality when the algorithm operates within a restricted space [4]; and Rademacher complexity, a measure of algorithmic expressivity developed by Barlett and Mendelson that eliminates the need for assuming restrictions on the distribution space of an algorithm [1]. More recent work has attempted to capture the expressivity of deep neural networks in particular, by using structural properties of neural networks to consider their representative power [11]. While this more recent work is of interest, we seek a much more general notion of algorithmic expressivity under which other such notions might be captured.

This paper delves into the relationships between algorithmic (inductive) bias, a concept explored by Mitchell due to its importance for generalization [7], and algorithmic expressivity. Towards this end, we build on the search and bias frameworks developed in [10], where Montañez et al. prove the necessity of bias for better-than-random performance of learning algorithms and that no algorithm may be simultaneously biased towards many distinct target sets. In addition to giving explicit bounds on the trade-offs between bias and algorithmic expressivity, we establish a general measure of algorithmic flexibility that applies to clustering and optimization [9] in addition to the problems considered in Vapnik's learning framework [15], such as classification, regression, and density estimation. Our framework's generality is conducive to applying theoretical derivations of algorithmic expressivity to many different types of learning algorithms.

3.1 Relation to Lauw et al.'s "The Bias-Expressivity Trade-off"

This manuscript is most closely related to Lauw et al.'s "The Bias-Expressivity Trade-off" [6], being an extended presentation of that work. Our discussion of the various uses of the term "bias" in machine learning acts as a helpful supplement to the space-constrained introduction made there. We also provide an improved presentation and motivation for the concepts of algorithmic bias and entropic expressivity found in that paper. The theorems from that work, along with their proofs and key figures, are reproduced here.

4 ALGORITHMIC SEARCH FRAMEWORK

4.1 The Search Problem

The framework used in this paper views machine learning problems as instances of algorithmic search problems [8]. In the algorithmic search framework, a search problem is defined as a 3-tuple, (Ω, T, F). An algorithm samples elements from a finite **search space** Ω in order to find a particular nonempty subset T of Ω, called the **target set**. The elements of Ω and T are encoded in a **target function**, a $|\Omega|$-length binary vector where an entry has value 1 if it belongs to the target set T and 0 otherwise. The **external information resource** F provides initialization information for the search and evaluates points in Ω to guide the search process. In a traditional machine learning scenario, the search space Ω corresponds to a hypothesis space an algorithm may have available (such as the space of linear functions). The external information resource F would be a dataset with an accompanying loss function. The target set T would correspond to those hypotheses which achieve sufficiently low empirical risk on a dataset given some desired threshold. The loss function included in F guides the algorithm in searching through Ω for a hypothesis in T.

4.2 The Search Algorithm

Black-box search algorithms can be viewed as processes that induce probability distributions over a search space and subsequently sample according to those distributions, in order to locate target elements within the space. Black-box algorithms use their search history to produce a sequence of distributions. The search history contains information gained during the course of the search by sampling elements of Ω and evaluating them according to the external information resource F, along with any information given as initialization information. As the search proceeds, a sequence of probability distributions gets generated, one distribution per iteration, and a point is sampled from each distribution at every iteration. The sampled point and its evaluation under F are added back to the search history, and the algorithm updates its sampling distribution over Ω. A search algorithm is *successful* if at any point of the search it samples an element $\omega \in T$

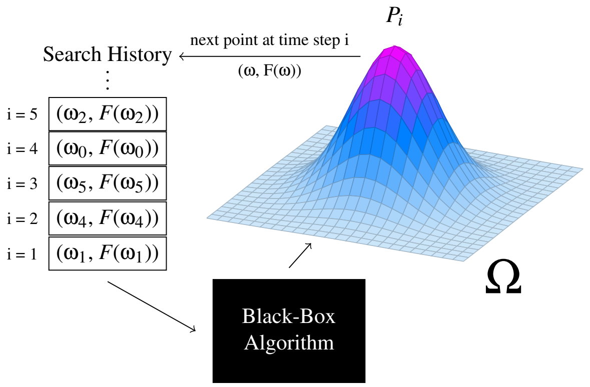
P_i
Search History
next point at time step i
(ω, F(ω))
i = 5 $(\omega_2, F(\omega_2))$
i = 4 $(\omega_0, F(\omega_0))$
i = 3 $(\omega_5, F(\omega_5))$
i = 2 $(\omega_4, F(\omega_4))$
i = 1 $(\omega_1, F(\omega_1))$
Black-Box Algorithm
Ω

Fig. 1: Graphical representation of the search process. As a black-box algorithm searches a space by sampling from Ω, it induces a probability distribution P_i at each time step, based on its current search history. A sampled point $\omega \in \Omega$ is evaluated according to the external information resource F, and the tuple $(\omega, F(\omega))$ is added to the search history. The process repeats until a termination criterion is met. Figure reproduced from [6].

contained in the target set. Success is determined retrospectively, since the algorithm has no knowledge of the target T during its search apart from that information given by F. Figure 1 gives a graphical representation of the search process.

4.3 Measuring Performance

A more fine-grained measure of search performance is given by the expected per-query probability of success [8], which normalizes the expected total cumulative probability of success by the number of queries taken. Different algorithms may choose to terminate their searches using different criteria, taking a different number of sampling steps. Using the expected total probability of success without normalization would unfairly reward algorithms making a larger number of queries. As an additional benefit, the expected per-query probability of success gracefully handles algorithms which may repeatedly resample the same points, such as genetic algorithms [3, 12].

Following Montañez [8], **the expected per-query probability of success** is defined as

$$q(T,F) = \mathbb{E}_{\tilde{P},H}\left[\frac{1}{|\tilde{P}|}\sum_{i=1}^{|\tilde{P}|} P_i(\omega \in T)\middle| F\right] \tag{1}$$

where $\tilde{P} = [P_1, P_2, \ldots, P_N]$ is a sequence of induced probability distributions on search space Ω (with P_i denoting the distribution at the ith iteration), H is the search history, T is the target set, and F is the external information resource. Thus, the expected per-query probability of success measures the expected amount of probability mass placed on the target set, averaged over the entire search. As Lauw et al. explain [6], *“The* *outer expectation accounts for stochastic differences in multiple runs of the algorithm, whereas the inner quantity is equivalent to the expected probability of success for a uniformly sampled time step of a given run."*

5 INDUCTIVE ORIENTATION AND ALGORITHMIC BIAS

We begin by introducing a geometric concept of algorithm behavior, the *inductive orientation*, which will allow us to define algorithmic bias in a way that is simultaneously quantitative and geometric. We then review the definition of algorithmic bias introduced in Lauw et al. [6], and define it in terms of inductive orientation.

5.1 Inductive Orientation

The definition of expected per-query probability of success given in Equation 1 naturally suggests a way of characterizing the behavior of an algorithm based on its expected average probability distribution on the search space. For a fixed target set t, define a corresponding $|\Omega|$-length binary target vector $\mathbf{t}$, which has a 1 at its ith index if the ith element of Ω is in the target set t, and has a zero otherwise. For a random information resource $F \sim \mathcal{D}$, we can see that

$$\begin{aligned} q(t,F) &= \mathbb{E}_{\tilde{P},H}\left[\frac{1}{|\tilde{P}|}\sum_{i=1}^{|\tilde{P}|} P_i(\omega \in t)\middle| F\right] \\ &= \mathbb{E}_{\tilde{P},H}\left[\frac{1}{|\tilde{P}|}\sum_{i=1}^{|\tilde{P}|} \mathbf{t}^\top \mathbf{P}_i \middle| F\right] \\ &= \mathbf{t}^\top \mathbb{E}_{\tilde{P},H}\left[\frac{1}{|\tilde{P}|}\sum_{i=1}^{|\tilde{P}|} \mathbf{P}_i \middle| F\right] \\ &= \mathbf{t}^\top \overline{\mathbf{P}}_F \end{aligned}$$

where we have defined $\overline{\mathbf{P}}_F := \mathbb{E}_{\tilde{P},H}[\frac{1}{|\tilde{P}|}\sum_{i=1}^{|\tilde{P}|} \mathbf{P}_i \mid F]$ as the expected average conditional distribution on the search space given F. We notice two things. First, the expected per-query probability of success is equivalent to an inner product between two vectors, one representing the target and the other representing where the algorithm tends to place probability mass in expectation. As Montanez et al. note [10], the degree to which the expected distribution aligns geometrically to the target vector is the degree to which a search algorithm will be successful. Second, as Sam et al. have shown, this implies that the expected per-query probability of success is a decomposable probability-of-success metric [13].

The **inductive orientation** of an algorithm is then defined as

$$\overline{\mathbf{P}}_{\mathcal{D}} = \mathbb{E}_{F\sim\mathcal{D}}[\overline{\mathbf{P}}_F] \tag{2}$$

given a marginal distribution $\mathcal{D}$ on the information resource F. From this definition, we see that

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}[q(t,F)] &= \mathbb{E}_{\mathcal{D}}[\mathbf{t}^{\top}\overline{\mathbf{P}}_F] && (3)\\ &= \mathbf{t}^{\top}\,\mathbb{E}_{\mathcal{D}}[\overline{\mathbf{P}}_F] && (4)\\ &= \mathbf{t}^{\top}\overline{\mathbf{P}}_{\mathcal{D}}. && (5)\end{aligned}$$

Thus, the expected per-query probability of success relative to a randomized F can be computed simply and geometrically, by taking an inner product of the inductive orientation with the target vector.

We can further extend the concept of inductive orientation with regards to any decomposable probability-of-success metric ϕ, since we can take any weighted average of the probability distributions in a search, not just the uniform average. We define the **ϕ-inductive orientation** for decomposable metric $\phi(t,F) = \mathbf{t}^{\top}\mathbf{P}_{\phi,F}$ as

$$\mathbf{P}_{\phi,\mathcal{D}} = \mathbb{E}_{F\sim\mathcal{D}}[\mathbf{P}_{\phi,F}], \tag{6}$$

of which $\overline{\mathbf{P}}_{\mathcal{D}}$ is simply a special case for uniform weighting [13].

5.2 Algorithmic Bias

We now review the definition of bias introduced in [10] and show how it can be defined as a linear function of the inductive orientation. We then restate some existing results for bias, which show the need for bias in learning systems.

Definition 3. *(Algorithmic Bias) Given a fixed target function* $\mathbf{t}$ *(corresponding to target set t), let $p = \|\mathbf{t}\|^2/|\Omega|$ denote the expected per-query probability of success under uniform random sampling, let $\mathbf{P}_{\mathcal{U}} = \mathbf{1}\cdot|\Omega|^{-1}$ be the inductive orientation vector for a uniform random sampler, and let $F \sim \mathcal{D}$, where $\mathcal{D}$ is a distribution over a collection of information resources $\mathcal{F}$. Then,*

$$\begin{aligned}\mathrm{Bias}(\mathcal{D},\mathbf{t}) &= \mathbb{E}_{\mathcal{D}}[q(t,F)-p]\\ &= \mathbf{t}^{\top}(\overline{\mathbf{P}}_{\mathcal{D}} - \mathbf{P}_{\mathcal{U}})\\ &= \mathbf{t}^{\top}\mathbb{E}_{\mathcal{D}}\left[\overline{\mathbf{P}}_F\right] - \mathbf{t}^{\top}(\mathbf{1}\cdot|\Omega|^{-1})\\ &= \mathbf{t}^{\top}\int_{\mathcal{F}}\overline{\mathbf{P}}_f\,\mathcal{D}(f)\,\mathrm{d}f - \frac{\|\mathbf{t}\|^2}{|\Omega|}.\end{aligned}$$

The above definition is in complete agreement with that given by Lauw et al. [6], but makes clearer the relation of bias to inductive orientation, the bias being a linear function of the orientation vector. The first equality in the definition highlights the semantic meaning of bias, being a deviation in performance from uniform random sampling. The second equality highlights the cause of this deviation, namely the algorithm's inductive orientation encoding assumptions concerning where target elements are likely to reside, distributing its probability mass unevenly within the search space. The larger the deviation from uniform mass placement, the greater the opportunity for improved (or degraded) performance.

As a special case, we can define bias with respect to a finite set of information resources, as follows.

Definition 4. *(Bias for a finite set of information resources) Let $\mathcal{U}[\mathcal{B}]$ denote a uniform distribution over a finite set of information resources $\mathcal{B}$. Then,*

$$\begin{aligned}\mathrm{Bias}(\mathcal{B},\mathbf{t}) &= \mathrm{Bias}(\mathcal{U}[\mathcal{B}],\mathbf{t})\\ &= \mathbf{t}^{\top}\left(\frac{1}{|\mathcal{B}|}\sum_{f\in\mathcal{B}}\overline{\mathbf{P}}_f\right) - \frac{\|\mathbf{t}\|^2}{|\Omega|}.\end{aligned}$$

6 EXISTING BIAS RESULTS

We restate a number of theorems given in Montañez et al. [10] which are useful for understanding the results in the present paper.

Theorem 1 (Improbability of Favorable Information Resources). *Let $\mathcal{D}$ be a distribution over a set of information resources $\mathcal{F}$, let F be a random variable such that $F \sim \mathcal{D}$, let $t \subseteq \Omega$ be an arbitrary fixed k-sized target set with corresponding target function* $\mathbf{t}$*, and let $q(t,F)$ be the expected per-query probability of success for algorithm $\mathcal{A}$ on search problem (Ω,t,F). Then, for any $q_{\min} \in [0,1]$,*

$$\Pr(q(t,F) \geq q_{\min}) \leq \frac{p+\mathrm{Bias}(\mathcal{D},\mathbf{t})}{q_{\min}}$$

where $p = \frac{k}{|\Omega|}$.

This theorem tells us that sampling highly favorable information resources remains unlikely for any distribution without high algorithmic bias for the given target. Given that we typically search for very small targets in very large spaces (implying a tiny p), the bound is restrictive for values of $q_{\min}$ approaching 1, unless the bias is strong. The upper bound is controlled linearly by the bias of the sampling distribution with respect to the fixed target and algorithm. Furthermore, bias is a conserved quantity: to be highly biased towards one target means to be equally biased against other targets. Thus, choosing an inductive orientation and bias represents a zero-sum game, as the next result shows.

Theorem 2 (Conservation of Bias). *Let $\mathcal{D}$ be a distribution over a set of information resources and let $\tau_k = \{\mathbf{t}|\mathbf{t} \in \{0,1\}^{|\Omega|}, \|\mathbf{t}\| = \sqrt{k}\}$ be the set of all $|\Omega|$-length k-hot vectors[1]. Then for any fixed algorithm $\mathcal{A}$,*

$$\sum_{\mathbf{t}\in\tau_k}\mathrm{Bias}(\mathcal{D},\mathbf{t}) = 0.$$

[1] k-hot vectors are binary vectors containing exactly k ones.

This result can be viewed as a special case of No Free Lunch [14, 17] behavior, since bias is a relative performance measure between two algorithm strategies, and the set of all k-sized targets is closed under permutation [14], a necessary and sufficient condition for the original No Free Lunch theorems.

Theorem 3 (Famine of Favorable Information Resources). *Let $\mathcal{B}$ be a finite set of information resources and let $t \subseteq \Omega$ be an arbitrary fixed k-size target set with corresponding target function* $\mathbf{t}$*. Define*

$$\mathcal{B}_{q_{\min}} = \{f \mid f \in \mathcal{B}, q(t,f) \geq q_{\min}\},$$

where $q(t,f)$ is the expected per-query probability of success for algorithm $\mathcal{A}$ on search problem (Ω, t, f) and $q_{\min} \in [0,1]$ represents the minimum acceptable per-query probability of success. Then,

$$\frac{|\mathcal{B}_{q_{\min}}|}{|\mathcal{B}|} \leq \frac{p + \mathrm{Bias}(\mathcal{B}, \mathbf{t})}{q_{\min}}$$

where $p = \frac{k}{|\Omega|}$.

By the above theorem, the proportion of $q_{\min}$-favorable information resources is bounded by the problem difficulty and the average bias of the set as a whole. For any fixed value of bias, fixed target, and fixed algorithm, the proportion of highly favorable information resources remains strictly bound.

Lastly, we see that without bias, the single-query probability of success for any algorithm is equivalent to uniform random sampling: it is the same as flipping coins. Algorithms must have nonuniform inductive orientations to perform well, and any choice of inductive orientation is a choice against some targets sets, thus encoding trade-offs among the various possible targets.

Theorem 4 (Futility of Bias-Free Search). *For any fixed algorithm $\mathcal{A}$, fixed target $t \subseteq \Omega$ with corresponding target function* $\mathbf{t}$*, and distribution over information resources $\mathcal{D}$, if* $\mathrm{Bias}(\mathcal{D}, \mathbf{t}) = 0$*, then*

$$\Pr(\omega \in t; \mathcal{A}) = p$$

where $\Pr(\omega \in t; \mathcal{A})$ *represents the single-query probability of successfully sampling an element of t using $\mathcal{A}$, marginalized over information resources $F \sim \mathcal{D}$, and p is the single-query probability of success under uniform random sampling.*

At this point, we remind the reader that although the above theorems are stated with reference to search and sampling, they apply far more widely to most forms of artificial learning, such as AI methods and other types of machine learning [9], being formalized within the algorithmic search framework for that purpose [8].

7 MAIN RESULTS

In this section, we present results building on the definitions of algorithmic bias and inductive orientation given in Section 5. We reproduce the main results of Lauw et al. [6] and add new discussion concerning the relevance of each result. These results include an upper bound on the bias of a learning algorithm in relation to its minimum value over the set of possible targets, a concentration bound on the difference between estimated and actual bias, and bounds relating algorithmic bias to entropic expressivity. These bounds capture an inherent trade-off between the expressivity and bias for artificial learning systems.

7.1 Bias Bounds

Theorem 5 (Bias Upper Bound). *Let $\tau_k = \{\mathbf{t} | \mathbf{t} \in \{0,1\}^{|\Omega|}, ||\mathbf{t}|| = \sqrt{k}\}$ be the set of all $|\Omega|$-length k-hot vectors and let $\mathcal{B}$ be a finite set of information resources. Then,*

$$\sup_{\mathbf{t}\in\tau_k} \operatorname{Bias}(\mathcal{B},\mathbf{t}) \leq \left(\frac{p-1}{p}\right) \inf_{\mathbf{t}\in\tau_k} \operatorname{Bias}(\mathcal{B},\mathbf{t})$$

where $p = \frac{k}{|\Omega|}$.

This result presents limitations on the amount of bias that can be induced within a learning algorithm from all possible target sets of a fixed size. From Theorem 5, we see that the maximum amount of bias that can be induced in a learning algorithm is related to the minimum amount that can be induced. The two are related by, at most, a constant factor $\frac{p-1}{p}$, where p is the proportion of elements in the $|\Omega|$-sized search space that are in the target set.

Figure 2 demonstrates the relationship between the value of p and the upper bound on bias in Theorem 5. We see that as p increases, the upper bound on bias tightens considerably. This is due to the fact that the target set k increases in size relative to the size of Ω, which substantially increases the probability that the algorithm will do well on a greater number of target sets because of target element density in the search space. This indicates that the algorithm is not predisposed towards any particular target set, giving evidence against the presence of strong bias (Theorem 2).

Theorem 6 (Difference Between Estimated and Actual Bias). *Let $\mathbf{t}$ be a fixed target function, let $\mathcal{D}$ be a distribution over a set of information resources $\mathcal{B}$, and let $X = \{X_1, \ldots, X_n\}$ be a finite sample independently drawn from $\mathcal{D}$. Then,*

$$\mathbb{P}(|\operatorname{Bias}(X,\mathbf{t}) - \operatorname{Bias}(\mathcal{D},\mathbf{t})| \geq \varepsilon) \leq 2e^{-2n\varepsilon^2}.$$

While the bias with respect to an underlying distribution over information resources may not be accessible, it may be possible to estimate it by drawing, independently at random, a sample from that distribution. Theorem 6 quantifies how well the empirical bias estimates the true bias with high probability.

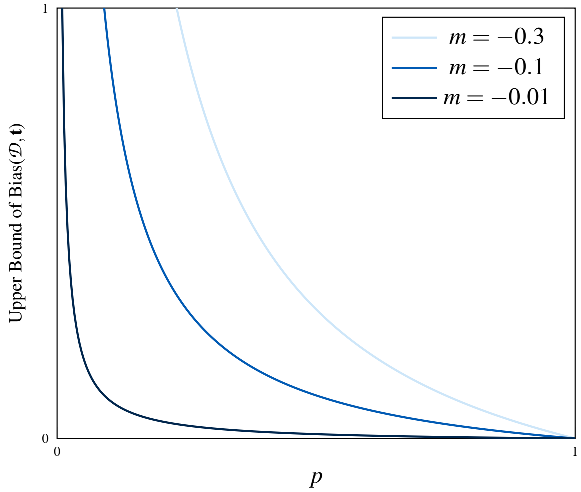

Fig. 2: As plotted for different values of $m = \frac{p-1}{p}$, the upper bound of the supremum of bias changes with different values of p, where the supremum is over all possible target sets of some fixed size k. Figure reproduced from [6].

7.2 Entropic Expressivity

Just as the bias of an algorithm can be defined as a function of inductive orientation, so can the expressivity. We now formalize such a definition.

Definition 5 (Entropic Expressivity). *The entropic expressivity of a search algorithm is the information-theoretic entropy of its inductive orientation, namely,*

$$H(\overline{\mathbf{P}}_{\mathcal{D}}) = H(\mathcal{U}) - D_{\mathrm{KL}}(\overline{\mathbf{P}}_{\mathcal{D}} \,||\, \mathcal{U})$$

where $D_{\mathrm{KL}}(\overline{\mathbf{P}}_{\mathcal{D}} \,||\, \mathcal{U})$ is the Kullback-Leibler divergence between distribution $\overline{\mathbf{P}}_{\mathcal{D}}$ and the uniform distribution $\mathcal{U}$, both being distributions over search space Ω.

Informally, the expressivity of an algorithm is how well it responds to changes in data: different datasets produce different probability distributions over the search space. More formally, expressivity is the degree to which an algorithm's inductive orientation vector blurs towards uniformity, meaning that in expectation it places probability mass on many different regions of the search space, in reaction to changing information resources. Again, expressive algorithms will shift mass around different regions of the space as the dataset changes. This stands in contrast to a highly biased algorithm, which places substantial mass in a limited number of regions of the search space. Consequently, the inductive orientation of the more flexible algorithm will be closer to uniform than that of the highly biased algorithm. Using the information-theoretic entropy for discrete probability mass functions, our notion of entropic expressivity characterizes this aspect of algorithmic flexibility.

7.3 Expressivity and Bias Trade-Off

We now present results relating entropic expressivity to algorithmic bias, bounding expressivity in terms of bias and bias in terms of expressivity, demonstrating a trade-off between the two quantities.

Theorem 7 (Expressivity Bounded by Bias). *Let* $\varepsilon := \mathrm{Bias}(\mathcal{D}, \mathbf{t})$. *Given a fixed k-hot target vector* $\mathbf{t}$ *and a distribution over information resources* $\mathcal{D}$*, the entropic expressivity,* $H(\overline{\mathbf{P}}_{\mathcal{D}})$*, of a search algorithm can be bounded in terms of bias,* ε*, by*

$$H(\overline{\mathbf{P}}_{\mathcal{D}}) \in \left[H(p+\varepsilon), \left((p+\varepsilon)\log_2\left(\frac{k}{p+\varepsilon}\right) \right.\right.$$
$$\left.\left. + (1-(p+\varepsilon))\log_2\left(\frac{|\Omega|-k}{1-(p+\varepsilon)}\right)\right)\right].$$

Theorem 7 gives the minimum and maximum amount of entropic expressivity for a given amount of bias and fixed target set size. We see that bias limits the entropic expressivity of a learning algorithm, by reducing the uniformity of its inductive orientation. To get a better sense of this result, we show how the range of entropic expressivity varies in response to plugging in different values of algorithmic bias. In Table 1, we consider some exemplar cases, where a learning algorithm has minimum bias, zero bias, as well as maximum bias, and present the entropic expressivity range computed based on Theorem 7.

Table 1: As computed for cases where there is minimum bias, zero bias, and maximum bias, the bound for the range of entropic expressivity changes with different levels of bias relative to target function $\mathbf{t}$. Table reproduced from [6].

$\mathrm{Bias}(\mathcal{D}, \mathbf{t})$	$\mathbf{t}^\top \overline{\mathbf{P}}_{\mathcal{D}}$	**Expressivity Range**
$-p$ (Minimum bias)	0	$[0, \log_2(\lvert\Omega\rvert - k)]$
0 (No bias)	p	$[H(p), \log_2 \lvert\Omega\rvert]$
$1-p$ (Maximum bias)	1	$[0, \log_2 k]$

As an algorithm's bias increases, its entropic expressivity becomes more tightly bounded. In the majority of cases, the size k of the target set is substantially smaller than the size $|\Omega|$ of the search space. Therefore, in the case of minimum bias, entropic expressivity has only a loosely bounded range of $[0, \log_2(|\Omega| - k)]$. Similarly, when there is no bias, the expressivity can take on values up to $\log_2 |\Omega|$. In the case of maximum bias, however, the algorithm becomes extremely predisposed towards a particular outcome. Given that $k \ll |\Omega|$, $\log_2 k \ll \log_2(|\Omega - k|)$, indicating a considerable tightening of the bound. With these observations in hand, we now present our main result, which demonstrates a quantifiable trade-off between the algorithmic bias and entropic expressivity of artificial learning systems.

Theorem 8 (Bias-Expressivity Trade-off). *Given a distribution over information resources $\mathcal{D}$ and a fixed target $t \subseteq \Omega$, entropic expressivity is bounded above in terms of bias,*

$$H(\overline{\mathbf{P}}_{\mathcal{D}}) \leq \log_2 |\Omega| - 2\,\mathrm{Bias}(\mathcal{D},\mathbf{t})^2.$$

Additionally, bias is bounded above in terms of entropic expressivity,

$$\begin{aligned}\mathrm{Bias}(\mathcal{D},\mathbf{t}) &\leq \sqrt{\frac{1}{2}(\log_2 |\Omega| - H(\overline{\mathbf{P}}_{\mathcal{D}}))} \\ &= \sqrt{\frac{1}{2} D_{KL}(\overline{\mathbf{P}}_{\mathcal{D}} \mid\mid \mathcal{U})}.\end{aligned}$$

This theorem extends the results of Theorem 7 to demonstrate a mathematical relationship between algorithmic bias and entropic expressivity. As in Theorem 7, entropic expressivity is bound from above in terms of algorithmic bias. As the level of algorithmic bias on a specified target set increases, the level of entropic expressivity in the underlying inductive orientation decreases. We see that there is an opposing relationship between entropic expressivity and bias, such that higher values of algorithmic bias result in smaller values of entropic expressivity, and vice versa. We upper-bound algorithmic bias in terms of entropic expressivity, which again demonstrates this trade-off. The higher the entropic expressivity of a learning algorithm, the lower the bias. This result establishes that if a learning algorithm is strongly oriented towards any specific outcome, the algorithm becomes less flexible and less expressive over all elements, and the more flexible an algorithm, the less it can specialize towards specific outcomes.

Finally, we present a bound for algorithmic bias in terms of the expected entropy of induced strategy distributions. Similar to the trade-off relationship between algorithmic bias and entropic expressivity, the following corollary further establishes a trade-off between algorithmic bias and the expected entropy of induced strategy distributions.

Corollary 1 (Bias Bound Under Expected Expressivity).

$$\begin{aligned}\mathrm{Bias}(\mathcal{D},\mathbf{t}) &\leq \sqrt{\frac{1}{2}(\log_2 |\Omega| - \mathbb{E}_{\mathcal{D}}[H(\overline{\mathbf{P}}_F)])} \\ &= \sqrt{\mathbb{E}_{\mathcal{D}}\left[\frac{1}{2} D_{KL}(\overline{\mathbf{P}}_F \mid\mid \mathcal{U})\right]}.\end{aligned}$$

8 CONCLUSION

We extend the algorithmic search framework to consider a new notion of bias, being the difference in performance from uniform random sampling caused by the inductive assumptions encoded within an algorithm. We also define the entropic expressivity of a learning algorithm and characterize its relation to bias. Given an underlying distribution on information resources, entropic expressivity quantifies the expected degree of uniformity for strategy distributions, namely, the uniformity of the resulting inductive orientation. In addition to upper-bounding the bias on an arbitrary target set and the probability of a large difference between the estimated and true biases, we upper- and lower-bound the entropic expressivity with respect to the bias on a given target. These bounds concretely demonstrate the trade-off between bias and expressivity.

The bias-variance trade-off [2, 5] is well-known in machine learning. Our results present a similar trade-off, providing bounds for bias and expressivity in terms of one another. Our notion of bias corresponds to the expected deviation from uniform random sampling that results from an algorithm's inductive assumptions, while expressivity, similar to variance, captures how an algorithm's output distribution over its search space changes in expectation with regards to the underlying distribution on information resources (e.g., training data).

As shown by Mitchell [7] and later Montañez et al. [10], bias is necessary for learning algorithms to perform better than random chance. However, this comes at the cost of reducing the algorithm's ability to respond to varied training data. A maximally biased algorithm will have very little flexibility, while a maximally flexible algorithm, making no assumptions about its input, cannot perform better than uniform random sampling (Theorem 4). Fundamentally, bias and expressivity are both functions of an algorithm's inductive orientation: the more strongly pronounced its inductive orientation, the better an algorithm can generalize, but the less flexible it will be. Understanding the nature of this trade-off can help us design the type of behavior we want, according to the situation at hand.

References

1. Bartlett, P.L., Mendelson, S.: Rademacher and gaussian complexities: Risk bounds and structural results. J. Mach. Learn. Res. **3**, 463–482 (Mar 2003), ISSN 1532-4435, URL `http://dl.acm.org/citation.cfm?id=944919.944944`
2. Geman, S., Bienenstock, E., Doursat, R.: Neural networks and the bias/variance dilemma. Neural computation **4**(1), 1–58 (1992)
3. Goldberg, D.: Genetic algorithms in search optimization and machine learning. Addison-Wesley Longman Publishing Company (1999)
4. Kearns, M.J., Schapire, R.E.: Efficient distribution-free learning of probabilistic concepts. In: Proceedings [1990] 31st Annual Symposium on Foundations of Computer Science, pp. 382–391 vol.1 (Oct 1990), https://doi.org/10.1109/FSCS.1990.89557
5. Kohavi, R., Wolpert, D.H., et al.: Bias plus variance decomposition for zero-one loss functions. In: ICML, vol. 96, pp. 275–83 (1996)
6. Lauw, J., Macias, D., Trikha, A., Vendemiatti, J., Montañez, G.D.: The bias-expressivity trade-off. In: Rocha, A.P., Steels, L., van den Herik, H.J. (eds.) Proceedings of the 12th International Conference on Agents and Artificial Intelligence, ICAART 2020, Volume 2, Valletta, Malta, February 22-24, 2020, pp. 141–150, SCITEPRESS (2020), https://doi.org/10.5220/0008959201410150, URL `https://doi.org/10.5220/0008959201410150`

7. Mitchell, T.D.: The need for biases in learning generalizations. In: Rutgers University: CBM-TR-117 (1980)
8. Montañez, G.D.: The Famine of Forte: Few Search Problems Greatly Favor Your Algorithm. In: 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 477–482, IEEE (2017)
9. Montanẽz, G.D.: Why Machine Learning Works. In: Dissertation, Carnegie Mellon University (2017)
10. Montañez, G.D., Hayase, J., Lauw, J., Macias, D., Trikha, A., Vendemiatti, J.: The futility of bias-free learning and search. In: 32nd Australasian Joint Conference on Artificial Intelligence, pp. 277–288, Springer (2019)
11. Raghu, M., Poole, B., Kleinberg, J., Ganguli, S., Sohl-Dickstein, J.: On the expressive power of deep neural networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol. 70, pp. 2847–2854, PMLR, International Convention Centre, Sydney, Australia (Aug 2017), URL http://proceedings.mlr.press/v70/raghu17a.html
12. Reeves, C., Rowe, J.E.: Genetic algorithms: principles and perspectives: a guide to GA theory, vol. 20. Springer Science & Business Media (2002)
13. Sam, T., Williams, J., Abel, T., Huey, S., Montañez, G.D.: Decomposable probability-of-success metrics in algorithmic search. In: Rocha, A.P., Steels, L., van den Herik, H.J. (eds.) Proceedings of the 12th International Conference on Agents and Artificial Intelligence, ICAART 2020, Volume 2, Valletta, Malta, February 22-24, 2020, pp. 785–792, SCITEPRESS (2020), https://doi.org/10.5220/0009098807850792, URL https://doi.org/10.5220/0009098807850792
14. Schumacher, C., Vose, M.D., Whitley, L.D.: The no free lunch and problem description length. In: Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2001), pp. 565–570 (2001)
15. Vapnik, V.N.: An overview of statistical learning theory. IEEE transactions on neural networks **10**(5), 988–999 (1999)
16. Vapnik, V.N., Chervonenkis, A.Y.: On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probability and its Applications **16**(2), 264–280 (1971)
17. Wolpert, D.H., Macready, W.G.: No free lunch theorems for optimization. IEEE transactions on evolutionary computation **1**(1), 67–82 (1997)

APPENDIX

For completeness, in this appendix we reproduce all proofs from Lauw et al. [6] in their entirety, without modification.

Lemma 1 (Existence of subset with at most uniform mass). *Given an n-sized subset S of the sample space of an arbitrary probability distribution with total probability mass M_S, there exists a k-sized proper subset $R \subset S$ with total probability mass M_R such that*

$$M_R \leq \frac{k}{n} M_S.$$

Proof. We proceed by induction on the size k.

Base Case: When $k = 1$, there exists an element with total probability mass at most $\frac{M_S}{n}$, since for any element in S that has probability mass greater than the uniform mass $\frac{M_S}{n}$, there exists an element with mass strictly less than $\frac{M_S}{n}$ by the law of total probability. This establishes our base case.

Inductive Hypothesis: Suppose that a k-sized subset $R_k \subset S$ exists with total probability mass M_{R_k} such that $M_{R_k} \leq \frac{k}{n} M_S$.

Induction Step: We show that there exists a subset $R_{k+1} \subset S$ of size $k+1$ with total probability mass $M_{R_{k+1}}$ such that $M_{R_{k+1}} \leq \frac{k+1}{n} M_S$.

First, let $M_{R_k} = \frac{k}{n} M_S - s$, where $s \geq 0$ represents the slack between M_{R_k} and $\frac{k}{n} M_S$. Then, the total probability mass on ${R_k}^{\mathrm{c}} := S \setminus R_k$ is

$$M_{{R_k}^{\mathrm{c}}} = M_S - M_{R_k} = M_S - \frac{k}{n} M_S + s.$$

Given that $M_{{R_k}^{\mathrm{c}}}$ is the total probability mass on set ${R_k}^{\mathrm{c}}$, either each of the $n-k$ elements in ${R_k}^{\mathrm{c}}$ has a uniform mass of $M_{{R_k}^{\mathrm{c}}}/(n-k)$, or they do not. If the probability mass is uniformly distributed, let e be an element with mass exactly $M_{{R_k}^{\mathrm{c}}}/(n-k)$. Otherwise, for any element e' with mass greater than $M_{{R_k}^{\mathrm{c}}}/(n-k)$, by the law of total probability there exists an element $e \in {R_k}^{\mathrm{c}}$ with mass less than $M_{{R_k}^{\mathrm{c}}}/(n-k)$. Thus, in either case there exists an element $e \in {R_k}^{\mathrm{c}}$ with mass at most $M_{{R_k}^{\mathrm{c}}}/(n-k)$.

Then, the set $R_{k+1} = R_k \cup \{e\}$ has total probability mass

$$\begin{aligned}
M_{R_{k+1}} &\leq M_{R_k} + \frac{M_{{R_k}^{\mathrm{c}}}}{n-k} \\
&= \frac{k}{n} M_S - s + \frac{M_S - \frac{k}{n} M_S + s}{n-k} \\
&= \frac{k M_S (n-k) + n(M_S - \frac{k}{n} M_S + s)}{n(n-k)} - s \\
&= \frac{knM_S - k^2 M_S + nM_S - kM_S + ns}{n(n-k)} - s \\
&= \frac{(n-k)(kM_S + M_S) + ns}{n(n-k)} - s \\
&= \frac{k+1}{n} M_S + \frac{s}{n-k} - s \\
&= \frac{k+1}{n} M_S + \frac{s(1+k-n)}{n-k} \\
&\leq \frac{k+1}{n} M_S
\end{aligned}$$

where the final inequality comes from the fact that $k < n$. Thus, if a k-sized subset $R_k \in S$ exists such that $M_{R_k} \leq \frac{k}{n} M_S$, a $k+1$-sized subset $R_{k+1} \in S$ exists such that $M_{R_{k+1}} \leq \frac{k+1}{n} M_S$.

Since the base case holds true for $k = 1$ and the inductive hypothesis implies that this rule holds for $k+1$, we can always find a k-sized subset $R_k \in S$ such that

$$M_{R_k} \leq \frac{k}{n} M_S.$$

Lemma 2 (Maximum probability mass over a target set). *Let $\tau_k = \{\boldsymbol{t} | \boldsymbol{t} \in \{0,1\}^{|\Omega|}, ||\boldsymbol{t}|| = \sqrt{k}\}$ be the set of all $|\Omega|$-length k-hot vectors. Given an arbitrary probability distribution* $\mathbf{P}$*,*

$$\sup_{\mathbf{t} \in \tau_k} \mathbf{t}^\top \mathbf{P} \leq 1 - \left(\frac{1-p}{p}\right) \inf_{\mathbf{t} \in \tau_k} \mathbf{t}^\top \mathbf{P}$$

where $p = \frac{k}{|\Omega|}$.

Proof. We proceed by contradiction. Suppose that

$$\sup_{\mathbf{t} \in \tau_k} \mathbf{t}^\top \mathbf{P} > 1 - \left(\frac{1-p}{p}\right) \inf_{\mathbf{t} \in \tau_k} \mathbf{t}^\top \mathbf{P}.$$

Then, there exists some target function $\mathbf{t} \in \tau_k$ such that

$$\mathbf{t}^\top \mathbf{P} > 1 - \left(\frac{1-p}{p}\right) \inf_{\mathbf{t} \in \tau_k} \mathbf{t}^\top \mathbf{P}.$$

Let $\mathbf{s}$ be the complementary target function to $\mathbf{t}$ such that $\mathbf{s}$ is an $|\Omega|$-length, $(|\Omega| - k)$-hot vector that takes value 1 where $\mathbf{t}$ takes value 0 and takes value 0 elsewhere. Then, by the law of total probability,

$$\mathbf{s}^\top \mathbf{P} < \left(\frac{1-p}{p}\right) \inf_{\mathbf{t} \in \tau_k} \mathbf{t}^\top \mathbf{P}.$$

By Lemma 1, there exists a k-sized subset of the complementary target set with total probability mass q such that

$$\begin{aligned} q &\leq \frac{k}{|\Omega| - k} (\mathbf{s}^\top \mathbf{P}) \\ &< \frac{k}{|\Omega| - k} \left(\left(\frac{1-p}{p}\right) \inf_{\mathbf{t} \in \tau_k} \mathbf{t}^\top \mathbf{P} \right) \\ &= \frac{k}{|\Omega| - k} \left(\left(\frac{|\Omega| - k}{k}\right) \inf_{\mathbf{t} \in \tau_k} \mathbf{t}^\top \mathbf{P} \right) \\ &= \inf_{\mathbf{t} \in \tau_k} \mathbf{t}^\top \mathbf{P}. \end{aligned}$$

Thus, we can always find a target set with total probability mass strictly less than $\inf_{\mathbf{t} \in \tau_k} \mathbf{t}^\top \mathbf{P}$, which is a contradiction.

Therefore, we have proven that

$$\sup_{\mathbf{t} \in \tau_k} \mathbf{t}^\top \mathbf{P} \leq 1 - \left(\frac{1-p}{p}\right) \inf_{\mathbf{t} \in \tau_k} \mathbf{t}^\top \mathbf{P}.$$

Theorem 5 (Bias Upper Bound). *Let $\tau_k = \{\mathbf{t} | \mathbf{t} \in \{0,1\}^{|\Omega|}, ||\mathbf{t}|| = \sqrt{k}\}$ be the set of all $|\Omega|$-length k-hot vectors and let $\mathcal{B}$ be a finite set of information resources. Then,*

$$\sup_{\mathbf{t}\in\tau_k} \operatorname{Bias}(\mathcal{B},\mathbf{t}) \leq \left(\frac{p-1}{p}\right) \inf_{\mathbf{t}\in\tau_k} \operatorname{Bias}(\mathcal{B},\mathbf{t})$$

where $p = \frac{k}{|\Omega|}$.

Proof. First, define

$$m := \inf_{\mathbf{t}\in\tau_k} \mathbb{E}_{\mathcal{U}[\mathcal{B}]}[\mathbf{t}^\top \overline{\mathbf{P}}_F] = \inf_{\mathbf{t}\in\tau_k} \operatorname{Bias}(\mathcal{B},\mathbf{t}) + p$$

and

$$M := \sup_{\mathbf{t}\in\tau_k} \mathbb{E}_{\mathcal{U}[\mathcal{B}]}[\mathbf{t}^\top \overline{\mathbf{P}}_F] = \sup_{\mathbf{t}\in\tau_k} \operatorname{Bias}(\mathcal{B},\mathbf{t}) + p.$$

By Lemma 2,

$$M \leq 1 - \left(\frac{1-p}{p}\right) m.$$

Substituting the values of m and M,

$$\begin{aligned}\sup_{\mathbf{t}\in\tau_k} \operatorname{Bias}(\mathcal{B},\mathbf{t}) &\leq 1 - p - \left(\frac{1-p}{p}\right) \\ &\quad \left(\inf_{\mathbf{t}\in\tau_k} \operatorname{Bias}(\mathcal{B},\mathbf{t}) + p\right) \\ &= \left(\frac{p-1}{p}\right) \inf_{\mathbf{t}\in\tau_k} \operatorname{Bias}(\mathcal{B},\mathbf{t}).\end{aligned}$$

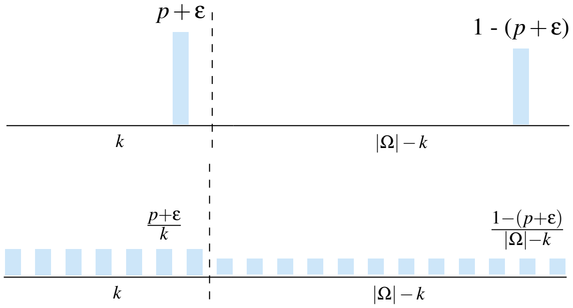

Fig. 3: Assuming positive bias, this figure shows two discrete probability distributions over Ω. The top is of an algorithm with high KL divergence while the bottom is of an algorithm with low KL divergence. Figure reproduced from Lauw et al. [6].

Theorem 6 (Difference Between Estimated and Actual Bias). *Let* $\mathbf{t}$ *be a fixed target function, let* $\mathcal{D}$ *be a distribution over a set of information resources* $\mathcal{B}$*, and let* $X = \{X_1, \dots, X_n\}$ *be a finite sample independently drawn from* $\mathcal{D}$*. Then,*

$$\mathbb{P}(|\operatorname{Bias}(X,\mathbf{t}) - \operatorname{Bias}(\mathcal{D},\mathbf{t})| \geq \varepsilon) \leq 2e^{-2n\varepsilon^2}.$$

Proof. Define

$$\begin{aligned}\overline{B}_X &:= \frac{1}{n}\sum_{i=1}^{n} \mathbf{t}^\top \overline{\mathbf{P}}_{X_i} \\ &= \operatorname{Bias}(X,\mathbf{t}) + p.\end{aligned}$$

Given that X is an iid sample from $\mathcal{D}$, we have

$$\begin{aligned}\mathbb{E}[\overline{B}_X] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{t}^\top \overline{\mathbf{P}}_{X_i}\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\mathbf{t}^\top \overline{\mathbf{P}}_{X_i}\right] \\ &= \operatorname{Bias}(\mathcal{D},\mathbf{t}) + p.\end{aligned}$$

By Hoeffding's inequality and the fact that

$$0 \leq \overline{B}_X \leq 1$$

we obtain

$$\begin{aligned}\mathbb{P}(|\operatorname{Bias}(X,\mathbf{t}) - \operatorname{Bias}(\mathcal{D},\mathbf{t})| \geq \varepsilon) &= \mathbb{P}(|\overline{B}_X - \mathbb{E}[\overline{B}_X]| \geq \varepsilon) \\ &\leq 2e^{-2n\varepsilon^2}.\end{aligned}$$

Theorem 7 (Expressivity Bounded by Bias). *Let* $\varepsilon := \operatorname{Bias}(\mathcal{D},\mathbf{t})$. *Given a fixed k-hot target vector* $\mathbf{t}$ *and a distribution over information resources* $\mathcal{D}$*, the entropic expressivity,* $H(\overline{\mathbf{P}}_{\mathcal{D}})$*, of a search algorithm can be bounded in terms of bias,* ε*, by*

$$\begin{aligned}H(\overline{\mathbf{P}}_{\mathcal{D}}) \in \Bigg[H(p+\varepsilon), \Bigg((p+\varepsilon)\log_2\left(\frac{k}{p+\varepsilon}\right) \\ + (1-(p+\varepsilon))\log_2\left(\frac{|\Omega|-k}{1-(p+\varepsilon)}\right)\Bigg)\Bigg].\end{aligned}$$

Proof. Following definition 5, the expressivity of a search algorithm varies solely with respect to $D_{\mathrm{KL}}(\overline{\mathbf{P}}_{\mathcal{D}} \;||\; \mathcal{U})$ since we always consider the same search space and thus $H(\mathcal{U})$ is a constant value. We obtain a lower bound of the expressivity by maximizing the value of $D_{\mathrm{KL}}(\overline{\mathbf{P}}_{\mathcal{D}} \;||\; \mathcal{U})$ and an upper bound by minimizing this term.

First, we show that $H(p+\varepsilon)$ is a lower bound of expressivity by constructing a distribution that deviates the most from a uniform distribution over Ω. By the definition of $\operatorname{Bias}(\mathcal{D},\boldsymbol{t})$, we place $(p+\varepsilon)$ probability mass on the target set t and $1-(p+\varepsilon)$

probability mass on the remaining $(n-k)$ elements of Ω. We distribute the probability mass such that all of the $(p+\varepsilon)$ probability mass of the target set is concentrated on a single element and all of the $1-(p+\varepsilon)$ probability mass of the complement of the target set is concentrated on a single element. In this constructed distribution where $D_{\mathrm{KL}}(\overline{\mathbf{P}}_{\mathcal{D}} \,||\, \mathcal{U})$ is maximized, the value of expressivity is

$$\begin{aligned}
H(\overline{\mathbf{P}}_{\mathcal{D}}) &= -\sum_{\omega\in\Omega} \overline{P}_{\mathcal{D}}(\omega)\log_2 \overline{P}_{\mathcal{D}}(\omega) \\
&= -(p+\varepsilon)\log_2(p+\varepsilon) \\
&\quad -(1-(p+\varepsilon))\log_2(1-(p+\varepsilon)) \\
&= H(p+\varepsilon)
\end{aligned}$$

where the $H(p+\varepsilon)$ is the entropy of a Bernoulli distribution with parameter $(p+\varepsilon)$. The entropy of this constructed distribution gives a lower bound on expressivity,

$$H(\overline{\mathbf{P}}_{\mathcal{D}}) \geq H(p+\varepsilon).$$

Now, we show that

$$(p+\varepsilon)\log_2\left(\frac{k}{p+\varepsilon}\right)+(1-(p+\varepsilon))\log_2\left(\frac{|\Omega|-k}{1-(p+\varepsilon)}\right)$$

is an upper bound of expressivity by constructing a distribution that deviates the least from a uniform distribution over Ω. In this case, we uniformly distribute $\frac{1}{|\Omega|}$ probability mass over the entire search space, Ω. Then, to account for the ε level of bias, we add $\frac{\varepsilon}{k}$ probability mass to elements of the target set and we remove $\frac{\varepsilon}{n-k}$ probability mass to elements of the complement of the target set. In this constructed distribution where

$D_{\mathrm{KL}}(\overline{\mathbf{P}}_{\mathcal{D}} \,||\, \mathcal{U})$ is minimized, the value of expressivity is

$$\begin{aligned}
H(\overline{\mathbf{P}}_{\mathcal{D}}) &= -\sum_{\omega\in\Omega} \overline{P}_{\mathcal{D}}(\omega)\log_2 \overline{P}_{\mathcal{D}}(\omega) \\
&= -\sum_{\omega\in t}\left(\frac{1}{|\Omega|}+\frac{\varepsilon}{k}\right)\log_2\left(\frac{1}{|\Omega|}+\frac{\varepsilon}{k}\right) \\
&\quad -\sum_{\omega\in t^{\mathrm{c}}}\left(\frac{1}{|\Omega|}-\frac{\varepsilon}{|\Omega|-k}\right)\log_2\left(\frac{1}{|\Omega|}-\frac{\varepsilon}{|\Omega|-k}\right) \\
&= -\sum_{\omega\in t}\left(\frac{p+\varepsilon}{k}\right)\log_2\left(\frac{p+\varepsilon}{k}\right) \\
&\quad -\sum_{\omega\in t^{\mathrm{c}}}\left(\frac{1-(p+\varepsilon)}{|\Omega|-k}\right)\log_2\left(\frac{1-(p+\varepsilon)}{|\Omega|-k}\right) \\
&= -k\left(\frac{p+\varepsilon}{k}\right)\log_2\left(\frac{p+\varepsilon}{k}\right) \\
&\quad -(|\Omega|-k)\left(\frac{1-(p+\varepsilon)}{|\Omega|-k}\right)\log_2\left(\frac{1-(p+\varepsilon)}{|\Omega|-k}\right) \\
&= (p+\varepsilon)\log_2\left(\frac{k}{p+\varepsilon}\right) \\
&\quad +(1-(p+\varepsilon))\log_2\left(\frac{|\Omega|-k}{1-(p+\varepsilon)}\right).
\end{aligned}$$

The entropy on this constructed distribution gives an upper bound on expressivity,

$$\begin{aligned}
H(\overline{\mathbf{P}}_{\mathcal{D}}) \le\, &(p+\varepsilon)\log_2\left(\frac{k}{p+\varepsilon}\right) \\
&+(1-(p+\varepsilon))\log_2\left(\frac{|\Omega|-k}{1-(p+\varepsilon)}\right).
\end{aligned}$$

These two bounds give us a range of possible values of expressivity given a fixed level of bias, namely

$$\begin{aligned}
H(\overline{\mathbf{P}}_{\mathcal{D}}) \in \Bigg[H(p+\varepsilon), \Bigg(&(p+\varepsilon)\log_2\left(\frac{k}{p+\varepsilon}\right) \\
&+(1-(p+\varepsilon))\log_2\left(\frac{|\Omega|-k}{1-(p+\varepsilon)}\right)\Bigg)\Bigg].
\end{aligned}$$

Theorem 8 (Bias-Expressivity Trade-off). *Given a distribution over information resources $\mathcal{D}$ and a fixed target $t \subseteq \Omega$, entropic expressivity is bounded above in terms of bias,*

$$H(\overline{\mathbf{P}}_{\mathcal{D}}) \le \log_2|\Omega| - 2\,\mathrm{Bias}(\mathcal{D},\mathbf{t})^2.$$

Additionally, bias is bounded above in terms of entropic expressivity,

$$\begin{aligned}\mathrm{Bias}(\mathcal{D},\mathbf{t}) &\leq \sqrt{\frac{1}{2}(\log_2|\Omega| - H(\overline{\mathbf{P}}_{\mathcal{D}}))} \\ &= \sqrt{\frac{1}{2}D_{KL}(\overline{\mathbf{P}}_{\mathcal{D}} \mid\mid \mathcal{U})}.\end{aligned}$$

Proof. Let $\omega \in t$ denote the measurable event that ω is an element of target set $t \subseteq \Omega$, and let Σ be the sigma algebra of measurable events. First, note that

$$\begin{aligned}\mathrm{Bias}(\mathcal{D},t)^2 &= |\,\mathrm{Bias}(\mathcal{D},t)|^2 \\ &= |\mathbf{t}^\top \mathbb{E}_{\mathcal{D}}[\overline{\mathbf{P}}_F] - p|^2 \\ &= |\mathbf{t}^\top \overline{\mathbf{P}}_{\mathcal{D}} - p|^2 \\ &= |\overline{P}_{\mathcal{D}}(\omega \in t) - p|^2 \\ &\leq \frac{1}{2}D_{\mathrm{KL}}(\overline{\mathbf{P}}_{\mathcal{D}} \mid\mid \mathcal{U}) \\ &= \frac{1}{2}(H(\mathcal{U}) - H(\overline{\mathbf{P}}_{\mathcal{D}})) \\ &= \frac{1}{2}(\log_2|\Omega| - H(\mathbb{E}_{\mathcal{D}}[\overline{\mathbf{P}}_F]))\end{aligned}$$

where the inequality is an application of Pinsker's Inequality. The quantity $D_{\mathrm{KL}}(\overline{\mathbf{P}}_{\mathcal{D}} \mid\mid \mathcal{U})$ is the Kullback-Leibler divergence between distributions $\overline{\mathbf{P}}_{\mathcal{D}}$ and $\mathcal{U}$, which are distributions on search space Ω.

Thus,

$$H(\mathbb{E}_{\mathcal{D}}[\overline{\mathbf{P}}_F]) \leq \log_2|\Omega| - 2\,\mathrm{Bias}(\mathcal{D},\mathbf{t})^2$$

and

$$\begin{aligned}\mathrm{Bias}(\mathcal{D},t) &\leq \sqrt{\frac{1}{2}(\log_2|\Omega| - H(\overline{\mathbf{P}}_{\mathcal{D}}))} \\ &= \sqrt{\frac{1}{2}D_{\mathrm{KL}}(\overline{\mathbf{P}}_{\mathcal{D}} \mid\mid \mathcal{U})} \\ &= \sqrt{\frac{1}{2}(\log_2|\Omega| - H(\mathbb{E}_{\mathcal{D}}[\overline{\mathbf{P}}_F]))}.\end{aligned}$$

Corollary 1 (Bias Bound Under Expected Expressivity).

$$\begin{aligned}\mathrm{Bias}(\mathcal{D},\mathbf{t}) &\leq \sqrt{\frac{1}{2}(\log_2|\Omega| - \mathbb{E}_{\mathcal{D}}[H(\overline{\mathbf{P}}_F)])} \\ &= \sqrt{\mathbb{E}_{\mathcal{D}}\left[\frac{1}{2}D_{KL}(\overline{\mathbf{P}}_F \mid\mid \mathcal{U})\right]}.\end{aligned}$$

Proof. By the concavity of the entropy function and Jensen's Inequality, we obtain

$$\mathbb{E}_{\mathcal{D}}[H(\overline{\mathbf{P}}_F)] \leq H(\mathbb{E}_{\mathcal{D}}[\overline{\mathbf{P}}_F]) \leq \log_2 |\Omega| - 2\,\mathrm{Bias}(\mathcal{D}, \boldsymbol{t})^2.$$

Thus, an upper bound of bias is

$$\begin{aligned}
\mathrm{Bias}(\mathcal{D}, t) &\leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\overline{\mathbf{P}}_{\mathcal{D}} \mid\mid \mathcal{U})} \\
&= \sqrt{\frac{1}{2}(\log_2 |\Omega| - H(\mathbb{E}_{\mathcal{D}}[\overline{\mathbf{P}}_F]))} \\
&\leq \sqrt{\frac{1}{2}(\log_2 |\Omega| - \mathbb{E}_{\mathcal{D}}[H([\overline{\mathbf{P}}_F])])} \\
&= \sqrt{\mathbb{E}_{\mathcal{D}}\left[\frac{1}{2} D_{\mathrm{KL}}(\overline{\mathbf{P}}_F \mid\mid \mathcal{U})\right]},
\end{aligned}$$

where the final equality follows from the linearity of expectation and the definition of KL-divergence.