





# The Bias-Expressivity Trade-off

Julius Lauw<sup>†</sup><sup>a</sup>, Dominique Macias<sup>†</sup><sup>b</sup>, Akshay Trikha<sup>†</sup><sup>c</sup>, Julia Vendemiatti<sup>†</sup><sup>d</sup>  
and George D. Montañez<sup>†</sup><sup>e</sup>

AMISTAD Lab, Department of Computer Science, Harvey Mudd College, Claremont, CA 91711, U.S.A.

<sup>†</sup>denotes equal authorship.

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

**Abstract:** Learning algorithms need bias to generalize and perform better than random guessing. We examine the flexibility (expressivity) of biased algorithms. An expressive algorithm can adapt to changing training data, altering its outcome based on changes in its input. We measure expressivity by using an information-theoretic notion of entropy on algorithm outcome distributions, demonstrating a trade-off between bias and expressivity. To the degree an algorithm is biased is the degree to which it can outperform uniform random sampling, but is also the degree to which it becomes inflexible. We derive bounds relating bias to expressivity, proving the necessary trade-offs inherent in trying to create strongly performing yet flexible algorithms.


## 1 INTRODUCTION


Biased algorithms, namely those which are more heavily predisposed to certain outcomes than others, have difficulty changing their behavior in response to new information or new training data. Yet bias is needed for learning (Montañez et al., 2019). Given a set of information resources (or a distribution over them), an algorithm that can output many different responses is said to be more *expressive* than one that cannot. We explore the inverse relationship between algorithmic bias and expressivity for learning algorithms. This work builds on recent results in theoretical machine learning, which highlight the necessity of incorporating biases tailored to specific learning problems in order to achieve learning performance that is better than uniform random sampling of the hypothesis space (Montañez et al., 2019). A trade-off exists between specialization and flexibility of learning algorithms. While algorithmic bias can be viewed as an algorithm’s ability to ‘specialize’, expressivity characterizes the ‘flexibility’ of a learning algorithm. Using the algorithmic search framework for learning (Montañez, 2017b), we define a specific form of


expressivity, called *entropic expressivity*, which is a function of the information-theoretic entropy of an algorithm’s induced probability distribution over its search space. Under this notion of expressivity, the degree to which a search algorithm is able to spread its probability mass on many distinct target sets captures the extent to which the same algorithm is said to be capable of ‘expressing’ a preference towards different search outcomes. No algorithm can be both highly biased and highly expressive.


## 2 RELATED WORK


Inspired by Mitchell’s work highlighting the importance of incorporating biases in classification algorithms to generalize beyond training data (Mitchell, 1980), we propose a method to measure algorithmic expressivity in terms of the amount of bias induced by a learning algorithm. This paper delves further into the relationships between algorithmic bias and expressivity by building on the search and bias theoretical frameworks defined in (Montañez et al., 2019). Montañez et al. proved that bias is necessary for a learning algorithm to perform better than uniform random sampling, and algorithmic bias was shown to encode trade-offs, such that no algorithm can be concurrently biased towards many distinct target sets. In this paper, we apply these properties of algorithmic bias to derive an upper bound on the level of bias encoded

<sup>a</sup> <https://orcid.org/0000-0003-4201-0664>

<sup>b</sup> <https://orcid.org/0000-0002-6506-4094>

<sup>c</sup> <https://orcid.org/0000-0001-8207-6399>

<sup>d</sup> <https://orcid.org/0000-0002-6547-9601>

<sup>e</sup> <https://orcid.org/0000-0002-1333-4611>

in a learning algorithm, in order to gain insights on the expressivity of learning algorithms.

Within the statistical learning literature, there exist various measures characterizing algorithmic expressivity. For instance, the Vapnik-Chervonekis (VC) dimension (Vapnik and Chervonekis, 1971) provides a loose upper bound on algorithmic expressivity in general by characterizing the number of data points that can be exactly classified by the learning algorithm, for any possible labeling of the points. However, the disadvantages of the VC dimension include its inherent dependence on the dimensionality of the space on which the learning algorithm operates on (V'yugin, 2015), as well as the fact that it is only restricted to classification problems. Building on the original VC dimension idea, Kearns and Schapire developed a generalization of the VC dimension with the Fat-shattering VC dimension by deriving dimension-free bounds with the assumption that the learning algorithm operates within a restricted space (Kearns and Schapire, 1990). Further, Bartlett and Mendelson created Rademacher complexity as a more general measure of algorithmic expressivity by eliminating the assumption that learning algorithms are restricted within a particular distribution space (Bartlett and Mendelson, 2003).

In this paper, we establish an alternative general measure of algorithmic expressivity based on the algorithmic search framework (Montañez, 2017a). Because this search framework applies to clustering and optimization (Montañez, 2017b) as well as to the general machine learning problems considered in Vapnik's learning framework (Vapnik, 1999), such as classification, regression, and density estimation, theoretical derivations of the expressivity of search algorithms using this framework directly apply to the expressivity of many types of learning algorithms.

### 3 SEARCH FRAMEWORK

#### 3.1 The Search Problem

We formulate machine learning problems as search problems using the algorithmic search framework (Montañez, 2017a). Within the framework, a search problem is represented as a 3-tuple  $(\Omega, T, F)$ . The finite **search space** from which we can sample is  $\Omega$ . The subset of elements in the search space that we are searching for is the **target set**  $T$ . A **target function** that represents  $T$  is an  $|\Omega|$ -length vector with entries having value 1 when the corresponding elements of  $\Omega$  are in the target set and 0 otherwise. The **external information resource**  $F$  is a finite binary string that

provides initialization information for the search and evaluates points in  $\Omega$ , acting as an oracle that guides the search process. In learning scenarios this is typically a dataset with accompanying loss function.

#### 3.2 The Search Algorithm

Given a search problem, a history of elements already examined, and information resource evaluations, an algorithmic search is a process that decides how to next query elements of  $\Omega$ . As the search algorithm samples, it adds the record of points queried and information resource evaluations, indexed by time, to the search history. The algorithm uses the history to update its sampling distribution on  $\Omega$ . An algorithm is successful if it queries an element  $\omega \in T$  during the course of its search. Figure 1 visualizes the search process.

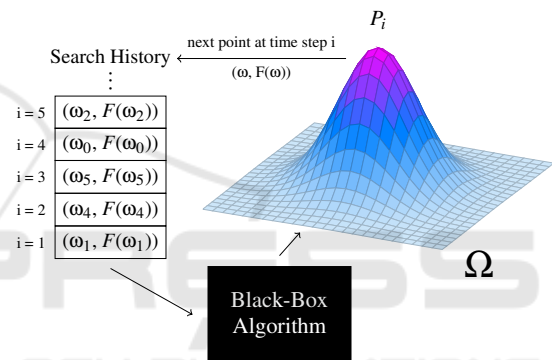


Figure 1: As a black-box optimization algorithm samples from  $\Omega$ , it produces an associated probability distribution  $P_i$  based on the search history. When a sample  $\omega_k$  corresponding to location  $k$  in  $\Omega$  is evaluated using the external information resource  $F$ , the tuple  $(\omega_k, F(\omega_k))$  is added to the search history.

#### 3.3 Measuring Performance

Following Montañez, we measure a learning algorithm's performance using the expected per-query probability of success (Montañez, 2017a). This quantity gives a normalized measure of performance compared to an algorithm's total probability of success, since the number of sampling steps may vary depending on the algorithm used and the particular run of the algorithm, which in turn effects the total probability of success. Furthermore, the per-query probability of success naturally accounts for sampling procedures that may involve repeatedly sampling the same points in the search space, as is the case with genetic algorithms (Goldberg, 1999; Reeves and Rowe, 2002), allowing this measure to deftly handle search algorithms that manage trade-offs between exploration and exploitation.

The expected per-query probability of success is defined as

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

where  $\tilde{P}$  is a sequence of probability distributions over the search space (where each timestep  $i$  produces a distribution  $P_i$ ),  $T$  is the target,  $F$  is the information resource, and  $H$  is the search history. The number of queries during a search is equal to the length of the probability distribution sequence,  $|\tilde{P}|$ . 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.

## 4 BIAS

In this section, we review the definition of bias introduced in (Montañez et al., 2019) and restate some results related to that concept, showing the need for bias in learning algorithms.

**Definition 4.1.** (Bias between a Distribution over Information Resources and a Fixed Target) Let  $\mathcal{D}$  be a distribution over a space of information resources  $\mathcal{F}$  and let  $F \sim \mathcal{D}$ . For a given  $\mathcal{D}$  and a fixed  $k$ -hot<sup>1</sup> target function  $\mathbf{t}$  (corresponding to target set  $t$ ),

$$\begin{aligned} \text{Bias}(\mathcal{D}, \mathbf{t}) &= \mathbb{E}_{\mathcal{D}}[q(t, F)] - \frac{k}{|\Omega|} \\ &= \mathbb{E}_{\mathcal{D}} \left[ \mathbf{t}^\top \bar{P}_F \right] - \frac{\|\mathbf{t}\|^2}{|\Omega|} \\ &= \mathbf{t}^\top \mathbb{E}_{\mathcal{D}} [\bar{P}_F] - \frac{\|\mathbf{t}\|^2}{|\Omega|} \\ &= \mathbf{t}^\top \int_{\mathcal{F}} \bar{P}_f \mathcal{D}(f) df - \frac{\|\mathbf{t}\|^2}{|\Omega|} \end{aligned}$$

where  $\bar{P}_f$  is the vector representation of the averaged probability distribution (conditioned on  $f$ ) induced on  $\Omega$  during the course of the search, which implies  $q(t, f) = \mathbf{t}^\top \bar{P}_f$ .

**Definition 4.2.** (Bias between a Finite Set of Information Resources and a Fixed Target) Let  $\mathcal{U}[\mathcal{B}]$  denote a uniform distribution over a finite set of information resources  $\mathcal{B}$ . For a random quantity  $F \sim \mathcal{U}[\mathcal{B}]$ , the averaged  $|\Omega|$ -length simplex vector  $\bar{P}_F$ , and a fixed

$k$ -hot target function  $\mathbf{t}$ ,

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

**Theorem 4.1** (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  $(\Omega, t, F)$ . Then, for any  $q_{\min} \in [0, 1]$ ,*

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

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

**Theorem 4.2** (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. Then for any fixed algorithm  $\mathcal{A}$ ,*

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

**Theorem 4.3** (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  $(\Omega, t, f)$  and  $q_{\min} \in [0, 1]$  represents the minimum acceptable per-query probability of success. Then,

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

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

**Theorem 4.4** (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  $\text{Bias}(\mathcal{D}, \mathbf{t}) = 0$ , 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.

<sup>1</sup> $k$ -hot vectors are binary and have exactly  $k$  ones.

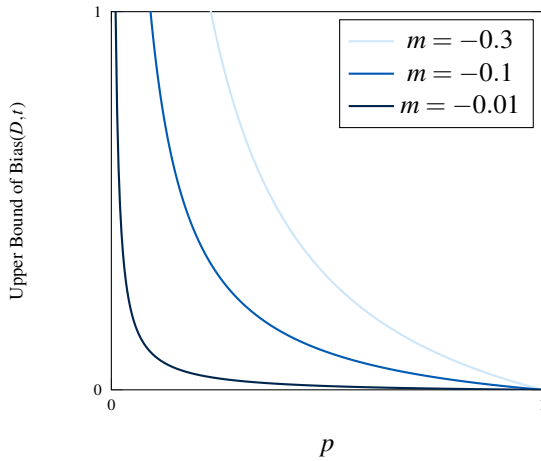


Figure 2: This graph shows how the upper bound of the supremum of the bias over all possible target sets of size  $k$  varies with different values of  $p$ , for different values of  $m = \frac{p-1}{p}$ .

## 5 MAIN RESULTS

Having reviewed the definitions of bias and prior results related to it, we now present our own results, with full proofs given in the Appendix. We proceed by presenting new results regarding bias and defining entropic expressivity. We explore expressivity in relation to bias, demonstrating a trade-off between them.

**Theorem 5.1** (Bias Upper Bound). *Let  $\tau_k = \{t | t \in \{0, 1\}^{|\Omega|}, |t| = \sqrt{k}\}$  be the set of all  $|\Omega|$ -length  $k$ -hot vectors and let  $\mathcal{B}$  be a finite set of information resources. Then,*

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

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

Theorem 5.1 confirms the intuition that the bounds on the maximum and minimum values the bias can take over all possible target sets are related by at most a constant factor. Note that from this theorem we can also derive a lower bound on the infimum of the bias by simply dividing by the constant factor.

We also consider the bound's behavior as  $p$  varies in Figure 2. As  $p$  increases, which can only happen as the size of the target set  $k$  increases relative to the size of  $\Omega$ , the upper bound on bias tightens. This is because if the target set size is a great proportion of the search space, it is more likely that the algorithm will do well on a greater number of target sets. Thus, it will be less biased towards any given one of them, by conservation of bias (Theorem 4.2).

**Theorem 5.2** (Difference between Estimated and Actual Bias). *Let  $t$  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,*

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

This theorem bounds the difference in the bias defined with respect to a distribution over information resources,  $\text{Bias}(\mathcal{D}, t)$ , and the bias defined on a finite set of information resources sampled from  $\mathcal{D}$ . In practice, we may not have access to the underlying distribution of information resources but we may be able to sample from such an unknown distribution. This theorem tells us how close empirically computed values of bias will be to the true value of bias, with high probability.

**Definition 5.1** (Entropic Expressivity). Given a distribution over information resources  $\mathcal{D}$ , we define the *entropic expressivity* of a search algorithm as the information-theoretic entropy of the averaged strategy distributions over  $\mathcal{D}$ , namely,

$$\begin{aligned} H(\bar{P}_{\mathcal{D}}) &= H(\mathbb{E}_{\mathcal{D}}[\bar{P}_F]) \\ &= H(\mathcal{U}) - D_{\text{KL}}(\bar{P}_{\mathcal{D}} \| \mathcal{U}) \end{aligned}$$

where  $F \sim \mathcal{D}$  and the quantity  $D_{\text{KL}}(\bar{P}_{\mathcal{D}} \| \mathcal{U})$  is the Kullback-Leibler divergence between distribution  $\bar{P}_{\mathcal{D}}$  and the uniform distribution  $\mathcal{U}$ , both being distributions over search space  $\Omega$ .

Definition 5.1 uses the standard information-theoretic entropy for discrete probability mass functions,  $H(\cdot)$ . Our notion of expressivity characterizes the flexibility of an algorithm by measuring the entropy of its induced probability vectors (strategies) averaged over the distribution on information resources. Algorithms that place probability mass on many different regions of the search space will tend to have a more uniform averaged probability vector. Entropic expressivity captures this key aspect of the flexibility of an algorithm.

We now present results relating this notion of expressivity to algorithmic bias.

**Theorem 5.3** (Expressivity Bounded by Bias). *Given a fixed  $k$ -hot target function  $t$  and a distribution over information resources  $\mathcal{D}$ , the entropic expressivity of a search algorithm can be bounded in terms of  $\epsilon := \text{Bias}(\mathcal{D}, t)$ , by*

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

This theorem shows that entropic expressivity is bounded above and below with respect to the level of bias on a fixed target. Table 1 demonstrates the different expressivity ranges for varying levels of bias. While these ranges may be quite large, maximizing the level of bias significantly reduces the range of possible values of entropic expressivity.

Table 1: Varying ranges of entropic expressivity for different levels of bias on target  $t$ .

Bias( $\mathcal{D}, t$ )	$\mathbb{E}[t^\top \bar{P}_F]$	Expressivity Range
$-p$ (Minimum bias)	0	$[0, \log_2( \Omega  - k)]$
0 (No bias)	$p$	$[H(p), \log_2  \Omega ]$
$1 - p$ (Maximum bias)	1	$[0, \log_2 k]$

**Theorem 5.4** (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,*

$$H(\bar{P}_{\mathcal{D}}) \leq \log_2 |\Omega| - 2 \text{Bias}(\mathcal{D}, t)^2$$

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

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

Theorem 5.4 demonstrates a trade-off between bias and entropic expressivity. We bound entropic expressivity above in terms of bias and bias above in terms of entropic expressivity such that higher values of bias decrease the range of possible values of expressivity and higher values of expressivity decrease the range of possible values of bias. Thus, a higher level of bias on a specified target restricts the expressivity of the underlying strategy distribution and a higher level of expressivity on the underlying strategy distribution restricts the amount of bias on any arbitrary target. Intuitively, this trade-off means that preferences towards specific targets reduces the potential flexibility of our algorithm over all elements and vice versa.

Lastly, we give a corollary bound allowing us to bound bias as a function of the expected entropy of induced strategy distributions, rather than the entropic expressivity.

**Corollary 5.4.1** (Bias Bound Under Expected Expressivity).

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

## 6 CONCLUSION

Expanding results on the algorithmic search framework, we supplement the notion of bias and define entropic expressivity, as well as its relation to bias. We upper bound the bias on an arbitrary target set with respect to the minimum bias toward a target set over all possible target sets of a fixed size. Moreover, we upper bound the probability of the difference between the estimated bias and the true bias exceeding some threshold, showing an exponential rate of measure concentration in the number of samples. Entropic expressivity characterizes the degree of uniformity for strategy distributions in expectation for an underlying distribution of information resources. We provide upper and lower bounds of the entropic expressivity with respect to the bias on a specified target and we demonstrate a trade-off between bias and expressivity.

While bias is needed for better-than-chance performance of learning algorithms, bias also hinders the flexibility of an algorithm by reducing the different ways it can respond to varied training data. Although algorithms predisposed to certain outcomes will not adapt as well as algorithms without strong predispositions, maximally flexible algorithms (those without any bias) can only perform as well as uniform random sampling (Theorem 4.4). This paper explores the trade-off, giving bounds for bias in terms of expressivity, and bounds for expressivity in terms of bias, demonstrating that such a trade-off exists. Although the notions of bias are different, the bias-expressivity trade-off can be viewed as a type of bias-variance trade-off (Geman et al., 1992; Kohavi et al., 1996), where bias here is not an expected error but an expected deviation from uniform random sampling performance caused by an algorithm’s inductive assumptions, and variance is not a fluctuation in observed error caused by changing data but is instead a “fluctuation” in algorithm outcome distributions caused by the same. Therefore, our results may provide new insights for that well-studied phenomenon.

## ACKNOWLEDGEMENTS

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

## REFERENCES

- Bartlett, P. L. and Mendelson, S. (2003). Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482.
- Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58.
- Goldberg, D. (1999). *Genetic algorithms in search optimization and machine learning*. Addison-Wesley Longman Publishing Company.
- Kearns, M. J. and Schapire, R. E. (1990). Efficient distribution-free learning of probabilistic concepts. In *Proceedings [1990] 31st Annual Symposium on Foundations of Computer Science*, pages 382–391 vol.1.
- Kohavi, R., Wolpert, D. H., et al. (1996). Bias plus variance decomposition for zero-one loss functions. In *ICML*, volume 96, pages 275–83.
- Mitchell, T. D. (1980). The need for biases in learning generalizations. In *Rutgers University: CBM-TR-117*.
- Montañez, G. D. (2017a). The Famine of Forte: Few Search Problems Greatly Favor Your Algorithm. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 477–482. IEEE.
- Montañez, G. D. (2017b). Why Machine Learning Works. In *Dissertation*. Carnegie Mellon University.
- Montañez, G. D., Hayase, J., Lauw, J., Macias, D., Trikha, A., and Vendemiatti, J. (2019). The futility of bias-free learning and search. In *AI 2019: Advances in Artificial Intelligence: 32nd Australasian Joint Conference, Adelaide, SA, Australia, December 2–5, 2019, Proceedings*, pages 277–288. Springer Nature.
- Reeves, C. and Rowe, J. E. (2002). *Genetic algorithms: principles and perspectives: a guide to GA theory*, volume 20. Springer Science & Business Media.
- Vapnik, V. and Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities.
- Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999.
- V’yugin, V. (2015). *VC Dimension, Fat-Shattering Dimension, Rademacher Averages, and Their Applications*, pages 57–74.

## APPENDIX

**Lemma 6.1** (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  $R_k^c := S \setminus R_k$  is

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

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

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

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

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

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

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

*Proof.* We proceed by contradiction. Suppose that

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

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

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

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,

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

By Lemma 6.1, there exists a  $k$ -sized subset of the complementary target set with total probability mass  $q$  such that

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

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

Therefore, we have proven that

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

□

**Theorem 5.1** (Bias Upper Bound). *Let  $\tau_k = \{\mathbf{t} \mid \mathbf{t} \in \{0,1\}^{|\Omega|}, \|\mathbf{t}\| = \sqrt{k}\}$  be the set of all  $|\Omega|$ -length  $k$ -hot vectors and let  $\mathcal{B}$  be a finite set of information resources. Then,*

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

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

*Proof.* First, define

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

and

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

By Lemma 6.2,

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

Substituting the values of  $m$  and  $M$ ,

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

□

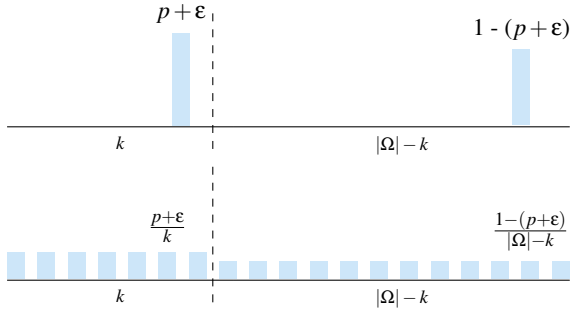


Figure 3: Assuming positive bias, this figure shows two discrete probability distributions over  $\Omega$ . The top is of an algorithm with high KL divergence while the bottom is of an algorithm with low KL divergence.

**Theorem 5.2** (Difference between Estimated and Actual Bias). *Let  $t$  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,*

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

*Proof.* Define

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

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

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

By Hoeffding's inequality and the fact that

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

we obtain

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

□

**Theorem 5.3** (Expressivity Bounded by Bias). *Given a fixed  $k$ -hot target function  $t$  and a distribution over information resources  $\mathcal{D}$ , the entropic expressivity of a search algorithm can be bounded in terms of  $\epsilon := \text{Bias}(\mathcal{D}, t)$ , by*

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

*Proof.* Following definition 5.1, the expressivity of a search algorithm varies solely with respect to  $D_{\text{KL}}(\bar{P}_{\mathcal{D}} \parallel \mathcal{U})$  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  $D_{\text{KL}}(\bar{P}_{\mathcal{D}} \parallel \mathcal{U})$  and an upper bound by minimizing this term.

First, we show that  $H(p + \epsilon)$  is a lower bound of expressivity by constructing a distribution that deviates the most from a uniform distribution over  $\Omega$ . By the definition of  $\text{Bias}(\mathcal{D}, t)$ , we place  $(p + \epsilon)$  probability mass on the target set  $t$  and  $1 - (p + \epsilon)$  probability mass on the remaining  $(n - k)$  elements of  $\Omega$ . We distribute the probability mass such that all of the  $(p + \epsilon)$  probability mass of the target set is concentrated on a single element and all of the  $1 - (p + \epsilon)$  probability mass of the complement of the target set is concentrated on a single element. In this constructed distribution where  $D_{\text{KL}}(\bar{P}_{\mathcal{D}} \parallel \mathcal{U})$  is maximized, the value of expressivity is

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

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

$$H(\bar{P}_{\mathcal{D}}) \geq H(p + \epsilon).$$

Now, we show that

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

is an upper bound of expressivity by constructing a distribution that deviates the least from a uniform distribution over  $\Omega$ . In this case, we uniformly distribute  $\frac{1}{|\Omega|}$  probability mass over the entire search space,  $\Omega$ . Then, to account for the  $\epsilon$  level of bias, we add  $\frac{\epsilon}{k}$  probability mass to elements of the target set and we remove  $\frac{\epsilon}{n-k}$  probability mass to elements of the complement of the target set. In this constructed distribution where  $D_{\text{KL}}(\bar{P}_{\mathcal{D}} \parallel \mathcal{U})$  is minimized, the value of expressivity is



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

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

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

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

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

□

**Theorem 5.4** (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,*

$$H(\bar{P}_{\mathcal{D}}) \leq \log_2 |\Omega| - 2 \text{Bias}(\mathcal{D}, t)^2$$

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

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

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

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

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

Thus,

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

and

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

□

**Corollary 5.4.1** (Bias Bound Under Expected Expressivity).

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

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

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

Thus, an upper bound of bias is

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

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

