From Targets to Rewards: Continuous Target Sets in the Algorithmic Search Framework

Milo Knell [1,†][a], Sahil Rane [1,†][b], Forrest Bicker [1,†][c], Tiger Che [1,†][d], Alan Wu[2,†][e], George D. Montañez [1][f]

[1] *AMISTAD Lab, Department of Computer Science, Harvey Mudd College, Claremont, CA 91711, USA*

[2] *Department of Computer Science, California Institute of Technology, Pasadena, CA 91125, USA*

{*mknell, srane, fbicker, tche, gmontanez*}*@hmc.edu, alawu@caltech.edu*

† *denotes equal authorship*

Keywords: Algorithmic Search Framework, satisfaction, fuzzy membership

Abstract: Many machine learning tasks have a measure of success that is naturally continuous, such as error under a loss function. We generalize the Algorithmic Search Framework (ASF), used for modeling machine learning domains as discrete search problems, to the continuous space. Moving from discrete target sets to a continuous measure of success extends the applicability of the ASF by allowing us to model fundamentally continuous notions like fuzzy membership. We generalize many results from the discrete ASF to the continuous space and prove novel results for a continuous measure of success. Additionally, we derive an upper bound for the expected performance of a search algorithm under arbitrary levels of quantization in the success measure, demonstrating a negative relationship between quantization and the performance upper bound. These results improve the fidelity of the ASF as a framework for modeling a range of machine learning and artificial intelligence tasks.

1 Introduction

The Algorithmic Search Framework (ASF) is a theoretical model that has been used to rigorously study properties of machine learning (ML), artificial intelligence (AI), and search problems (Montañez, 2017; Montañez et al., 2019; Montañez et al., 2021). This framework has been used to bound the performance of learning models, prove trade-offs between bias and expressivity (Lauw et al., 2020), derive generalization bounds for supervised classification (Ramalingam et al., 2022), and quantify performance bounds on transfer learning (Williams et al., 2020). However, one fundamental limitation of the ASF is that it measures the performance of a machine learning algorithm with a binary target set to which elements in the search space (also referred to as *hypotheses*) either belong or do not belong, rendering them indistinguishable from one another. This limitation makes it impossible to account for the fuzzy membership of hypotheses over a search space where each hypothesis may have a varying degree of fidelity. As a result, the strongest and weakest satisfactory hypotheses are treated equally. By using a continuous metric instead we can incorporate meaningful information about the relative certainty of our hypothesises, allowing us to both strengthen existing results and prove novel theorems.

Many modern applications of machine learning could benefit from a continuous success measure, which we term the **satisfaction** of a hypothesis. Hence, we propose a degree of satisfaction as a continuous-scale measure of the quality of a hypothesis function, instead of having the notion of binary membership in a target set. Examples of continuous measures of success include cross-entropy loss, mean squared error, hinge loss, accuracy, and F_1 score. To accurately model such problems in the ASF, we must account for continuous membership measures. Prior work avoided this limitation by implicitly defining some threshold of acceptability, where the target set was defined as the set of all elements with acceptable satisfaction values (Montañez, 2017). By defining such a threshold we lose information about the underlying satisfaction

[a] https://orcid.org/0009-0002-2951-8324
[b] https://orcid.org/0009-0001-3986-1129
[c] https://orcid.org/0009-0000-9872-7619
[d] https://orcid.org/0009-0000-3586-4288
[e] https://orcid.org/0009-0006-2454-4354
[f] https://orcid.org/0000-0002-1333-4611

structure between different hypotheses in the target set. However, employing a framework that directly interfaces with the underlying satisfaction structure enables us to generalize the results of the ASF, and measure the performance of machine learning algorithms more effectively, accounting for the fuzzy membership of hypotheses functions.

We examine related work to machine learning as search and fuzzy membership, rigorously and mathematically define the continuous ASF, present novel results, and show a real-world example applying these novel bounds.

2 Related Work

Machine learning can be modeled as search (Mitchell, 1982; Montañez, 2017). The conversion of machine learning problems to search problems enables us to perform a variety of analyses on their performance, using a mathematical and information-theoretic perspective. This approach helps us prove bounds on the performance of machine learning algorithms and gain an improved understanding of 'big picture' concepts in machine learning such as the bias-expressivity trade-off (Lauw et al., 2020).

Since its introduction, many results have been proven within the context of the ASF. For instance, researchers have demonstrated that a well-performing machine learning algorithm cannot exist without a predisposition to a certain group of outcomes (*bias*) (Montañez et al., 2019). Defining *expressivity* as the variability of outputs a machine learning algorithm can generate with different training data (Lauw et al., 2020), the ASF has been employed to prove the existence of fundamental trade-offs between bias and expressivity in machine learning (Lauw et al., 2020). Therefore, the ASF has proven effective in establishing fundamental properties of machine learning and machine learning algorithm performance. The framework has been used to prove 'famine' results, such as the fact that favorable algorithms for a specific task are scarce (Montañez, 2017).

In the ASF, researchers make several simplifying assumptions, including the use of binary target sets to measure the satisfaction of hypotheses (Montañez, 2017). Our work extends the ASF and generalizes the previously mentioned results. Moreover, we relax some simplifying assumptions within the ASF and generalize existing results proven within the framework to a continuous measure of satisfaction of hypotheses (Montañez, 2017; Montañez et al., 2019; Lauw et al., 2020). By extending the framework to account for continuous measures of satisfaction, we pave the way for future progress within the ASF, offering a more general framework applicable to a larger set of machine learning problems.

This generalization is especially valuable in the context of recent machine learning advances that incorporate fuzzy membership functions in a variety of capacities. This includes within models to enhance their accuracy, and in performing tasks such as image classification and various engineering applications (Hüllermeier, 2005; Gottwald, 2005; Resti et al., 2022; Ghofrani et al., 2014). Moreover, expanding the framework to encompass continuous target sets broadens its applicability, making it relevant to a broader range of machine learning challenges, thereby increasing its practicality.

3 The Algorithmic Search Framework (ASF)

3.1 The Search Problem

The ASF recasts machine learning problems as search problems, simplifying proofs for results on their performance. Following Montañez (Montañez, 2017), we model the search problem as a modular system of three parts, (Ω, T, F), where Ω represents the discrete, finite search space comprising hypotheses. We search within this space to find an element in the non-empty subset T, known as the target set. The external information resource F guides this search, providing initialization information, and offering evaluations on sampled elements from the search space to further steer our search. For instance, in a machine learning context, the external information resource F could be a training dataset with an accompanying loss function. Therefore, evaluating the external information resource on a particular element of the search space yields the loss function value for a specific hypothesis.

The target set T corresponds to the set of hypotheses that attain sufficiently high levels of satisfaction on a dataset for some desired threshold value of satisfaction. In the context of machine learning, we can interpret the satisfaction level of hypothesis as a notion of accuracy or performance on a test dataset. The loss function included in F directs the algorithm in searching through Ω for a hypothesis in T. This implies that we use our training data in our external information resource F to find a hypothesis in the target set T.

3.2 The Search Algorithm

The search algorithm $\mathcal{A}$ iteratively assigns a probability distribution over the search space, drawing from

its search history and the evaluation of the external information resource on each element, as shown in Figure 1. The search history comprises a query trace and a resource evaluation trace. The query trace holds the history of the elements that have been sampled by the search algorithm, and the resource evaluation trace records the history of the evaluations of the external information resource on these elements. A search algorithm within this framework is considered successful if it samples an element of the target set during its search. Importantly, only the external information resource, not the target set, guides the algorithm during the search process.

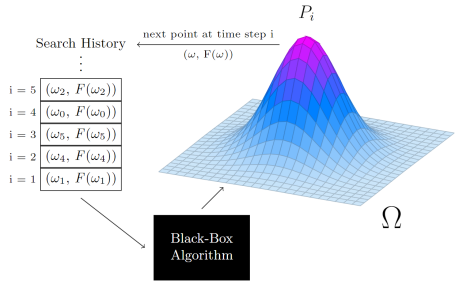

Figure 1: A black-box search algorithm. Reproduced from (Montañez et al., 2019)

4 Continuous ASF

4.1 Definitions

The satisfaction measure serves as the continuous case analog of the target set. It provides an indication of the quality of a hypothesis in the search space, signifying how good or bad a particular hypothesis is.

Definition 4.1 (Satisfaction)**.** The satisfaction function $s(\omega) : \Omega \to [0, c]$ maps from the search space Ω to real-valued quantities denoted *satisfactions*. We assume that these satisfactions exist in the range $[0, c]$ where c is a finite, positive real number. It is possible to assume that total satisfaction sums to 1 over the search space, which is achievable without loss of generality since we can linearly transform any satisfaction space to satisfy this property[1].

Definition 4.2 (Continuous Search Problem)**.** Let the tuple $(\Omega, \mathbf{s}, F)$ define a search problem. The search space Ω contains the elements (hypotheses) to be queried/explored. For each $\omega \in \Omega$, $s(\omega)$ denotes the level of satisfaction corresponding to the hypothesis ω. The function $s(\omega)$ can be represented by a vector $\mathbf{s} \in \mathbb{R}^{|\Omega|}$ where $\mathbf{s}_\omega = s(\omega)$. This deviation from binary membership target sets allows us to account for a continuous measure of satisfaction for hypotheses. F denotes the external information resource available to the learning algorithm, and for each element $\omega \in \Omega$, let $F(\omega)$ be the evaluation of the external information resource corresponding to the element of the search space ω. Thus, the only departure from the classic ASF lies in replacing binary T with continuous satisfaction measure $\mathbf{s}$.

Definition 4.3 (Expected Per-Query Satisfaction)**.** In Montañez's ASF (Montañez, 2017), success is measured using an expected per-query probability of success metric. In the continuous case, this generalizes to an expected per-query satisfaction metric. We do so by weighting the probability that each element is sampled by a search algorithm with its corresponding satisfaction level. Let H be the history of the search algorithm, F the external information resource, $\tilde{P}$ a sequence of probability distributions over the search space assigned by the search algorithm, and $\mathbf{P_i}$ be the probability distribution assigned by the search algorithm over the search space at a time step i in the search history H. Formally, we define the expected per-query satisfaction as

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

Definition 4.4 (Decomposability)**.** We observe that each $q(\mathbf{s}, F)$ can be decomposed into the inner product of $\mathbf{s}$ and $\overline{\mathbf{P}}_F$:

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

where we have defined $\overline{\mathbf{P}}_F := \mathbb{E}_{\tilde{P},H}\left[\frac{1}{|\tilde{P}|}\sum_{i=1}^{|\tilde{P}|}\mathbf{P}_i \mid F\right]$ as the expected average conditional distribution on the search space given F. Intuitively, this is equivalent to weighting each satisfaction value with its corresponding probability mass in $\overline{\mathbf{P}}_F$.

Our notation is summarized in Table 1. We include a real-world example in Section 6, anchoring the ASF to a practical machine-learning problem. Note that there are many learning processes, and some may differ from the examples below (for example, an unsupervised learning problem must have a different measure of success from a supervised classification problem). The ASF is general enough to encompass any algorithmic search problem.

[1] One such transformation is $s'(\omega) = \frac{s(\omega) - \min_\Omega s(\omega)}{\sum_\Omega s(\omega) - \min_\Omega s(\omega)}$.

Table 1: Notation

SYMBOL	DEFINITION
Ω, ω	Search space, element of search space, e.g., an element could be a set of parameters, like weights.
$\mathbf{s}, \mathbf{s}(\omega)$	Satisfaction vector, satisfaction of element ω, e.g., how good a particular hypothesis is, such as performance on the test set.
$T, \mathbf{t}$	Target set in the binary ASF (the equivalent of s in the continuous ASF), e.g., a set of hypothesis that are sufficiently satisfactory, possibly performing above some threshold on the test set. The binary vector representation of this set is given by $\mathbf{t}$.
$F, F(\omega)$	External information resource, evaluation of external information resource on an element ω, e.g., our training measure of how good a hypothesis is, such as training data and some loss function.
$q(\mathbf{s}, F)$	Expected per-query satisfaction.
$q(\mathbf{t}, F)$	Expected per-query probability of success in ASF.
ϕ	Decomposable satisfaction metric, e.g, q.
τ_k	A closed-under-permutation set of $\mathbf{s}$ such that all $\mathbf{s}^\top \mathbf{1} = k$.
$\mathcal{D}_{\tau_k}$	Distribution over a set of satisfaction vectors in τ_k.
$\mathcal{A}$	An abstract search algorithm which iteratively explores the search space.
$\mathbf{P_i}$	Distribution assigned by $\mathcal{A}$ over Ω at step i during the iterative search process. This could be conditioned on F.
$\tilde{P}$	Sequence of probability distributions assigned by algorithm $\mathcal{A}$, each element is a $\mathbf{P_i}$. This could be conditioned on F.
$\overline{\mathbf{P}}_F$	Expected averaged conditional distribution assigned by the search algorithm given the external information resource F.

5 Results

5.1 Famine of Favorable Satisfactions

Theorem 5.1. *For fixed $k \in \mathbb{R}_{\geq 0}$, fixed information resource F, decomposable, non-negative satisfaction metric ϕ such as q, and minimum acceptable per-query satisfaction q_{min}, we define*

$$\tau_k = \{\mathbf{s} \in \mathbb{R}^{|\Omega|} \mid \sum_{i=1}^{|\Omega|} \mathbf{s}_i = k\}, \textit{and}$$

$$\tau_{q_{\min}} = \{\mathbf{s} \in \tau_k \mid \phi(\mathbf{s}, F) \geq q_{min}\}.$$

Then

$$\frac{\mu(\tau_{q_{\min}})}{\mu(\tau_k)} \leq \frac{p}{q_{\min}}$$

where p is the expected per-query satisfaction under uniform random sampling and μ is Lebesgue measure.

This theorem shows that the proportion of satisfaction functions for which our algorithm performs extremely well (with more than q_{min} expected satisfaction) is small. In most practical applications, p is extremely small, as k will typically be extremely small in comparison to the size of the search space. The upper bound for the probability of successful search decreases linearly with the increase of threshold value q_{min}.

5.2 Success Under Dependence

Theorem 5.2. *Let c be a finite positive constant, and restrict $\mathbf{s}$ to an arbitrary quantization $Q = \{i \cdot c \mid i \in \{1, ..., m\}\}$. Let τ_k be the set of satisfaction vectors such that $\tau_k = \{\mathbf{s} \mid \mathbf{s} \in Q^{|\Omega|}, \mathbf{s}^\top \mathbf{1} = k\}$, and let $H(\mathcal{U}_{\tau_k;|\Omega|})$ denote the information-theoretic entropy of the uniform distribution over τ_k for a search space of cardinality $|\Omega|$.*

Let the satisfaction vector $S \sim \mathcal{D}_{\tau_k}$ be a vector-valued random variable over the set τ_k. Let X be the random variable such that $X \sim \overline{\mathbf{P}}_F$ over the elements of Ω. $S(X)$ is similar to $\mathbf{s}(\omega)$, except we are dealing with random variables S and X rather than specific realizations $\mathbf{s}$ and ω. Then for any non-negative constant u,

$$\Pr(S(X) \geq u) \leq \frac{I(S;F) + D_{KL}(\mathcal{D}_{\tau_k} \parallel \mathcal{U}_{\tau_k}) + H(S(X) \mid X)}{H(\mathcal{U}_{\tau_k;|\Omega|-1}) - H(\mathcal{U}_{\tau_{k-u};|\Omega|-1})}.$$

Theorem 5.2 provides a bound on the probability of sampling an element with a sufficiently large satisfaction defined by threshold u. This expression tells us that the upper bound of the probability of success monotonically improves as dependence between the satisfaction vector values and information resource values increases. The term $D_{\mathrm{KL}}(\mathcal{D}_{\tau_k} \parallel \mathcal{U}_{\tau_k})$ represents the Kullback-Leibler (KL) divergence between the actual distribution over the set τ_k, $\mathcal{D}_{\tau_k}$ and the uniform distribution over the same set τ_k, $\mathcal{U}_{\tau_k}$. This can be interpreted as the predictability of the distribution of satisfaction vectors, where large values of KL-divergence

represent more probability mass concentrated on a few elements. The $H(S(X) \mid X)$ term indicates the conditional entropy or surprisal associated with the possible satisfaction values for an element X sampled from the search space. This term is large when there are large variations in the satisfaction values thus resulting in an increase in the upper bound for the probability of sampling an element with a sufficiently large satisfaction value. The denominator in the bound essentially serves as a normalizing factor appropriately scaling the value of the upper bound.

We see that our upper bound increases with an increase in the predictability of the satisfaction vector, the dependence between the satisfaction vector and the external information resource, and the conditional entropy in the satisfaction values associated with an element in the search space. Thus, this theorem gives us an interpretable upper bound on the probability of sampling an element with a sufficiently large satisfaction value. Moreover, this theorem is particularly useful to allow us to determine situations where we cannot expect to perform well.

5.3 Expected Satisfaction Under Dependence

Theorem 5.3. *We will continue using all the definitions from Theorem 5.2. Let*

$$q = \mathbb{E}[S(X)] = \sum_{c \in Q} \Pr(S(X) = c) \cdot c.$$

Then,

$$q \leq \frac{I(S;F) + D_{KL}(\mathcal{D}_{\tau_k} \parallel \mathcal{U}_{\tau_k}) + H(S(X) \mid X)}{\frac{1}{c_m}(H(\mathcal{U}_{\tau_k;|\Omega|-1}) - H(\mathcal{U}_{\tau_{k-c_0};|\Omega|-1}))},$$

where $c_0 = \inf Q$ and $c_m = \sup Q$.

Extending from the bound on $\Pr(S(X) \geq u)$ presented in Theorem 5.2, we present a similar bound for the expected satisfaction, i.e., $q = \mathbb{E}[S(X)]$, without the need to specify a target satisfaction value defined by a constant threshold u. Compared to Theorem 5.2, this bound gives more context of the search problem, and can serve as a more robust metric since it's not susceptible to the skewness and kurtosis of the distribution of satisfaction values over the search space.

The interpretation of this bound is similar to Theorem 5.2 with a small change in the scaling factor in the denominator. Here, the bounded quantity is the expected satisfaction instead of the probability of exceeding a certain satisfaction. Comparing the two bounds in 5.2 and 5.3, we see that the bound in 5.3 is useful when sub-optimal elements contribute to the success of the search algorithm, whereas the bound in 5.2 is useful when sub-optimal elements do not contribute to the success of the search algorithm.

5.4 Difference in Satisfaction

We next quantify and bound the difference between expected per-query satisfaction (i.e., for continuous targets) and expected per-query probability of success (i.e., for binary targets), beginning with a helpful lemma.

Lemma 5.4. *Let g be the threshold value for converting a continuous target set into a discrete (binary) target set where all elements with satisfaction greater than or equal to the threshold g are included in the target set and the rest are excluded. Given a probability vector $\mathbf{w}$, satisfaction vector $\mathbf{s}$, target vector $\mathbf{t}$, and vector $\mathbf{v} = \mathbf{s} - \mathbf{t}$,*

$$|\mathbf{v}^\top \mathbf{w}| \leq \max(1-g, g).$$

Theorem 5.5. *Let $\overline{\mathbf{P}}_{F,\mathbf{s}}$ be the average conditional distribution assigned by the search algorithm under a continuous satisfaction measure, and let $\overline{\mathbf{P}}_{F,\mathbf{t}}$ be the averaged conditional distribution assigned by the search algorithm under a discrete target set. Let r be the maximum rounding amount defined as* $\max(1-g, g)$. *Then,*

$$|q(\mathbf{s},F) - q(\mathbf{t},F)| \leq |T| \sqrt{\frac{1}{2} D_{KL}\left(\overline{\mathbf{P}}_{F,\mathbf{s}} \parallel \overline{\mathbf{P}}_{F,\mathbf{t}}\right)} + r.$$

Theorem 5.5 bounds the difference in the success measure in the discrete and continuous case using the KL-divergence between the distributions learned in the continuous and discrete cases. It indicates that the potential for improved performance obtained by transitioning from a discrete target set to continuous target sets relies on the chosen threshold value g and the size of the target set. The degree of divergence between the outputs of the search algorithm in the two scenarios is measured by KL-divergence. This relationship is logical since the potential for performance improvement between the case with discrete and continuous target sets should be proportional to the amount of rounding required, which is related to both the threshold and the size of the target set. By transitioning from using discrete to continuous target sets, we also have the potential to gain from the divergence between the average conditional probability mass functions produced by the search algorithm in both cases. This is because the external information resource has the potential to be more useful in the case of a continuous satisfaction measure.

6 Example

6.1 Setup

To anchor this theoretical framework we provide an example of how it can be applied to a simple machine learning regression problem. We first create an independent variable X then use some stochastic data generating process to obtain our dependent variable Y. For the purposes of this example we use $Y = 2X + 5 + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, 100)$.

Suppose we try to model the data from this learning problem with a linear regression of the form $Y = aX + b$. We would then be able to model the training process within the ASF as a search over the space of possible learned parameters. We consider a finite set of values from which to take a and b. Let A and B be sets of evenly spaced numbers in the finite interval $[a_{min}, a_{max}]$ and $[b_{min}, b_{max}]$ with a step size x. That is, $A = \{a_{min}, a_{min} + x, ..., a_{max}\}$ and likewise for B. Then, $\Omega = A \times B$ or the Cartesian product of A and B. For this example, we selected $A = [0, 4]$ with a step size of 0.01 and $B = [1, 7]$, also with a step size of 0.01. In general, this is done by inspecting the distribution of X and Y.

We must also determine an external information resource F that will guide our search through the parameter space. For this regression problem, we use the mean squared error (MSE) of our hypothesized model calculated on the training set. That is, for a particular hypothesis ω corresponding to a pair of parameters (a, b) where $a \in A$ and $b \in B$, the evaluation of the external information resource is the mean squared error on the training set, or, $F(\omega) = \frac{1}{n}\sum_i^n (Y_i - (aX_i + b))^2$.

The overarching goal of the search algorithm is to find elements in the search space that have large satisfaction values associated with them. Our search algorithm determines the distribution $\overline{\mathbf{P}}_F$ to attempt to maximize $\mathbf{s}^\top \overline{\mathbf{P}}_F$, but it only has the information provided by F. In this example, the satisfaction values can be interpreted as the mean squared error on the test set. While we perform our search we don't use these values of satisfaction to guide our search, only the evaluation of F. In machine learning terms, the algorithm does not have access to the test data during the training step.

6.2 Example Result

Let us evaluate the bound presented in Theorem 5.2 both on differing levels of quantization and with different evaluation scenarios to gain more insight into the theorem. If our training data was generated by the same process as our testing data and success measured similarly, we would expect our mean squared error on the training data to be reflective of the mean squared error of the test data, thus giving us a high satisfaction for a trained model. However, if the test data was generated via a different process or the measure of success between training and testing data were different, we would not expect our trained model to have high satisfaction.

Consider a case where mean squared error is used to evaluate a hypothesized model on the training data in the information resource, but the satisfaction measure is the mean absolute error. In this case, the information a learning algorithm is guided by during search is systematically different from the information it is supposed to learn. This would be reflected by a lower value in the mutual information term $I(S;F)$.

By making reasonable assumptions about the structure of our example problem, we can compute the value of $I(S;F)$ and the bound for Theorem 5.2. First, we set $k = 1$, and $|\Omega| = |A| \times |B| = 400 \times 600$ (the bounds assigned in the previous section). We compute the bounds for levels of quantization $m = 2$ (binary) and $m = 3$ (ternary). For binary we set our $c = 0.5$ so $s_i \in \{0, 0.5\}$. For ternary, we set $c = \frac{1}{3}$ so $s_i \in \{0, \frac{1}{3}, \frac{2}{3}\}$. We assume that $\mathcal{D}_{\tau_k} = \mathcal{U}_{\tau_k;|\Omega|}$, that is $S \sim \mathcal{U}_{\tau_k;|\Omega|}$. While this assumption is not necessary it simplifies our calculations.

We produce these results in Table 2. The column *Match* means that MSE is used for evaluating in both the train and test phase (i.e F and $\mathbf{s}$), while the column *Not Match* means that MSE was used in the train phase and MAE was used in the test phase. The rows $m = 2$ and $m = 3$ mean a binary level and ternary level of quantization, respectively.

Table 2: Example Results

	Match	**Not Match**
m=2	$I(S;F) = 0.60$ $\Pr(S(X) \geq \frac{1}{2}) \leq 0.31$	$I(S;F) = 0.27$ $\Pr(S(X) \geq \frac{1}{2}) \leq 0.14$
m=3	$I(S;F) = 0.97$ $\Pr(S(X) \geq \frac{2}{3}) \leq 0.34$	$I(S;F) = 0.55$ $\Pr(S(X) \geq \frac{2}{3}) \leq 0.19$

From the table, we can observe that having the same evaluation metric for our train and test set raises our potential for performing well. This change is largely driven by the decreased mutual information term $I(S;F)$ displayed for each. Comparing across levels of quantization does not necessarily make sense, especially since the selection of u (i.e., $\frac{1}{2}$ and $\frac{2}{3}$) differ in the two cases based on the level of quantization.

This demonstrates how these bounds can be applied to real-world problems, and show how changes in $I(S;F)$ influence our ability to perform well. Our

ability to do well on a problem is limited by the quality of our information with respect to what we are trying to learn. In other words: garbage in, garbage out.

7 Conclusion

We extend the Algorithmic Search Framework from discrete target sets to a continuous measure of success, addressing one the framework's core limitations and increasing its versatility. We generalize theorems previously proven using the discrete ASF to the continuous and quantized cases, and derive novel results. Specifically, we prove an upper bound on performance under an arbitrary level of quantization, demonstrating that increasing the granularity of our success metric reduces our maximum theoretical performance. We bound the absolute difference in performance between the binary and continuous cases. We provide an example how the ASF can be applied to a regression problem and show how different processes for generating data or measuring success change key terms, like $I(S;F)$, thus varying our bound on performance.

These results improve the ability of the ASF to model machine learning problems that naturally have continuous measures of success, unlocking the potential to further the body of existing ASF research. There remain many opportunities for extension. One possible application of this framework is that it can be used for an information theory-based analysis of auto-ML algorithms by giving us a framework to better understand the performance of this domain of machine learning algorithms. Strengthening this theoretical framework will give researchers the tools to analyze learning algorithms with a naturally continuous measure of success.

REFERENCES

Ghofrani, F., Helfroush, M. S., Danyali, H., and Kazemi, K. (2014). Improving the performance of machine learning algorithms using fuzzy-based features for medical x-ray image classification. In *Journal of Intelligent & Fuzzy Systems*, volume 6, pages 3169–3180.

Gottwald, S. (2005). Mathematical aspects of fuzzy sets and fuzzy logic: Some reflections after 40 years. In *Mathematical aspects of fuzzy sets and fuzzy logic: Some reflections after 40 years*, volume 156, pages 357–364. 40th Anniversary of Fuzzy Sets.

Hüllermeier, E. (2005). Fuzzy methods in machine learning and data mining: Status and prospects. In *Fuzzy methods in machine learning and data mining: Status and prospects*, volume 156, pages 387–406. 40th Anniversary of Fuzzy Sets.

Lauw, J., Macias, D., Trikha, A., Vendemiatti, J., and Montañez, G. D. (2020). The Bias-Expressivity Tradeoff. In Rocha, A. P., Steels, L., and van den Herik, H. J., editors, *Proceedings of the 12th International Conference on Agents and Artificial Intelligence, ICAART 2020, Volume 2, Valletta, Malta, February 22-24, 2020*, pages 141–150. SCITEPRESS.

Mitchell, T. M. (1982). Generalization as search. *Artificial Intelligence*, 18(2):203–226.

Montañez, G. D. (2017). 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., Bashir, D., and Lauw, J. (2021). Trading Bias for Expressivity in Artificial Learning. In Rocha, A. P., Steels, L., and van den Herik, J., editors, *Agents and Artificial Intelligence*, pages 332–353, Cham. Springer International Publishing.

Montañez, G. D., Hayase, J., Lauw, J., Macias, D., Trikha, A., and Vendemiatti, J. (2019). The Futility of Bias-Free Learning and Search. In *32nd Australasian Joint Conference on Artificial Intelligence*, pages 277–288. Springer.

Ramalingam, R., Dice, N. E., Kaye, M. L., and Montañez, G. D. (2022). Bounding Generalization Error Through Bias and Capacity. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Resti, Y., Irsan, C., Amini, M., Yani, I., Passarella, R., and Zayantii, D. A. (2022). Performance improvement of decision tree model using fuzzy membership function for classification of corn plant diseases and pests. In *Performance Improvement of Decision Tree Model using Fuzzy Membership Function for Classification of Corn Plant Diseases and Pests*, volume 7, page 284–290.

Williams, J., Tadesse, A., Sam, T., Sun, H., and Montañez, G. D. (2020). Limits of Transfer Learning. *The Sixth International Conference on Machine Learning, Optimization, and Data Science (LOD 2020)*.

APPENDIX

Theorem 5.1. For fixed $k \in \mathbb{R}_{\geq 0}$, fixed information resource f, decomposable, non-negative satisfaction metric ϕ, and minimum acceptable per-query satisfaction q_{min}, we define

$$\tau_k = \{\mathbf{s} \in \mathbb{R}^{|\Omega|} \mid \sum_{i=1}^{|\Omega|} \mathbf{s}_i = k\}, \text{and}$$

$$\tau_{q_{\min}} = \{\mathbf{s} \in \tau_k \mid \phi(\mathbf{s}, F) \geq q_{min}\}.$$

Then $\frac{\mu(\tau_{q_{\min}})}{\mu(\tau_k)} \leq \frac{p}{q_{\min}}$ where p is the per-query expected satisfaction under uniform random sampling and μ is Lebesgue measure.

Proof. Under uniform sampling on τ_k, we have

$$\frac{\mu(\tau_{q_{\min}})}{\mu(\tau_k)} = \Pr(\phi(\mathbf{s},F) \geq q_{\min}) \leq \frac{\mathbb{E}_{\mathcal{U}[\tau_k]}[\phi(\mathbf{s},F)]}{q_{\min}}$$

where $\mathcal{U}[\tau_k]$ is the uniform distribution over τ_k, $\mathbf{s} \sim \mathcal{U}[\tau_k]$, and the second step follows from Markov's inequality. By the decomposability of ϕ and linearity of expectation, we have:

$$\frac{\mu(\tau_{q_{\min}})}{\mu(\tau_k)} \leq \frac{\mathbb{E}_{\mathcal{U}[\tau_k]}\left[\mathbf{s}^\top \overline{\mathbf{P}}_{\phi,F}\right]}{q_{\min}} = \frac{\mathbb{E}_{\mathcal{U}[\tau_k]}\left[\mathbf{s}^\top\right] \overline{\mathbf{P}}_{\phi,F}}{q_{\min}}.$$

As $\mathcal{U}[\tau_k]$ is uniform, $\mathbb{E}_{\mathcal{U}[\tau_k]}\left[\mathbf{s}^\top\right] = p \cdot \mathbf{1}^\top$. It follows that

$$\frac{\mu(\tau_{q_{\min}})}{\mu(\tau_k)} \leq \frac{p \cdot \mathbf{1}^\top \overline{\mathbf{P}}_{\phi,F}}{q_{\min}}.$$

Furthermore, as $\overline{\mathbf{P}}_{\phi,F}$ is a probability distribution, $\sum_\omega \overline{\mathbf{P}}_{\phi,F}(\omega) = 1$ and $\mathbf{1}^\top \overline{\mathbf{P}}_{\phi,F} = 1$. Hence, we conclude

$$\frac{\mu(\tau_{q_{\min}})}{\mu(\tau_k)} \leq \frac{p}{q_{\min}}.$$

□

Theorem 5.2. Let X be the random variable such that $X \sim \overline{\mathbf{P}}_F$. For any non-negative $u \in Q$:

$$\Pr(S(X) \geq u) \leq \frac{I(S;F) + D_{\mathrm{KL}}(\mathcal{D}_{\tau_k} \parallel \mathcal{U}_{\tau_k}) + H(S(X) \mid X)}{H(\mathcal{U}_{\tau_k;|\Omega|-1}) - H(\mathcal{U}_{\tau_{k-u};|\Omega|-1})}.$$

Proof. Q is a set of integer multiples of a constant spacing c (an arbitrary quantization) and can be expressed as $Q = \{0, c_0, c_1, ..., c_m\}$, where c_0 corresponds to the minimum positive value and c_m corresponds to the maximum value. Now, observe that

$$H(\mathcal{U}_{\tau_k;|\Omega|}) = \log(|\tau_k|) = \log\left(\binom{|\Omega| + \frac{k}{c} - 1}{\frac{k}{c}}\right).$$

Note that $H(\mathcal{U}_{\tau_k;|\Omega|})$ is monotonically increasing on $|\Omega|$ and k. For notational simplicity, let $P_g = \Pr(S(X) \geq u)$. We see that:

$$\begin{aligned} H(S \mid S(X),X) &= (1-P_g)H(S \mid S(X) < u, X) \\ &\quad + P_g H(S \mid S(X) \geq u, X) \\ &\leq (1-P_g)H(\mathcal{U}_{\tau_k;|\Omega|-1}) \\ &\quad + P_g H(\mathcal{U}_{\tau_{k-u};|\Omega|-1}) \\ &= H(\mathcal{U}_{\tau_k;|\Omega|-1}) \\ &\quad - P_g(H(\mathcal{U}_{\tau_k;|\Omega|-1}) - H(\mathcal{U}_{\tau_{k-u};|\Omega|-1})). \end{aligned}$$

The inequality follows from the fact that the entropy of a distribution of $\mathbf{s}$ is not larger than the entropy of uniform distribution of $\mathbf{s}$.

Also, by the chain rule of conditional entropy,

$$H(S,S(X) \mid X) = H(S \mid S(X),X) + H(S(X) \mid X)$$

$$H(S,S(X) \mid X) = H(S(X) \mid S,X) + H(S \mid X) = H(S \mid X) + 0$$

By the data processing inequality (with the Markov chain $S \to F \to X$),

$$\begin{aligned} H(S \mid F) &\leq H(S \mid X) \\ &= H(S,S(X)|X) \\ &= H(S \mid S(X),X) + H(S(X) \mid X) \end{aligned}$$

Then, we have:

$$\begin{aligned} H(S \mid F) &= H(S \mid S(X),X) + H(S(X) \mid X) \\ &\leq \underbrace{H(\mathcal{U}_{\tau_k;|\Omega|-1})}_{\leq H(\mathcal{U}_{\mathbf{s}})} \\ &\quad - P_g(H(\mathcal{U}_{\tau_k;|\Omega|-1}) - H(\mathcal{U}_{\tau_{k-u};|\Omega|-1})) \\ &\quad + H(S(X) \mid X) \end{aligned}$$

and by the definition of conditional entropy,

$$\begin{aligned} H(S) - I(S;F) \leq H(\mathcal{U}_{\mathbf{s}}) &- P_g(H(\mathcal{U}_{\tau_k;|\Omega|-1}) \\ &- H(\mathcal{U}_{\tau_{k-u};|\Omega|-1})) \\ &+ H(S(X) \mid X). \end{aligned}$$

Thus,

$$\Pr(S(X) \geq u) \leq \frac{I(S;F) + D_{\mathrm{KL}}(\mathcal{D}_{\tau_k} \parallel \mathcal{U}_{\tau_k}) + H(S(X) \mid X)}{H(\mathcal{U}_{\tau_k;|\Omega|-1}) - H(\mathcal{U}_{\tau_{k-u};|\Omega|-1})}.$$

□

Theorem 5.3. Let Q be a set of integer multiples of a constant spacing c and can be expressed as $Q = \{0, c_0, c_1, ..., c_m\}$, where c_0 corresponds to the minimum positive value and c_m corresponds to the maximum value. Let $P_c = \Pr(S(X) = c)$, $q = \sum_{c \in Q, c \geq 0} P_c c$, then we have:

$$q \leq \frac{I(S;F) + D_{\mathrm{KL}}(\mathcal{D}_s \parallel \mathcal{U}_s) + H(S(X) \mid X)}{\frac{1}{c_m}(H(\mathcal{U}_{\tau_k;|\Omega|-1}) - H(\mathcal{U}_{\tau_{k-c_0};|\Omega|-1}))}.$$

Proof.

$$\begin{aligned} H(S \mid S(X),X) &= (1 - \sum P_c) H(S \mid S(X) = 0, X) \\ &\quad + \sum P_c H(S \mid S(X) = c, X) \\ &\leq (1 - \sum P_c) H(\mathcal{U}_{\tau_k;|\Omega|-1}) \\ &\quad + \sum P_c \cdot c \cdot \frac{H(\mathcal{U}_{\tau_{k-c};|\Omega|-1})}{c} \\ &= H(\mathcal{U}_{\tau_k;|\Omega|-1}) \\ &\quad - \sum P_c c \cdot \frac{1}{c}(H(\mathcal{U}_{\tau_k;|\Omega|-1}) \\ &\quad - H(\mathcal{U}_{\tau_{k-c};|\Omega|-1})) \\ &\leq H(\mathcal{U}_{\tau_k;|\Omega|-1}) \\ &\quad - (\sum P_c c) \cdot \frac{1}{c_m}(H(\mathcal{U}_{\tau_k;|\Omega|-1}) \\ &\quad - H(\mathcal{U}_{\tau_{k-c_0};|\Omega|-1})). \end{aligned}$$

The last inequality is due to the monotonicity of $H(\mathcal{U}_{\tau_k;|\Omega|})$. Then, following the same steps as used in Theorem 5.2,

$$\begin{aligned} H(S) - I(S;F) \leq & \underbrace{H(\mathcal{U}_{\tau_k;|\Omega|-1})}_{\leq H(\mathcal{U}_{\mathbf{s}})} \\ & - (\underbrace{\sum P_c c}_{=q}) \frac{1}{c_m} (H(\mathcal{U}_{\tau_k;|\Omega|-1}) \\ & - H(\mathcal{U}_{\tau_{k-c_0};|\Omega|-1})). \end{aligned}$$

After simplification,

$$q \leq \frac{I(S;F) + D_{\mathrm{KL}}(\mathcal{D}_s \parallel \mathcal{U}_s) + H(S(X) \mid X)}{\frac{1}{c_m}(H(\mathcal{U}_{\tau_k;|\Omega|-1}) - H(\mathcal{U}_{\tau_{k-c_0};|\Omega|-1}))}.$$

□

Lemma 5.4. Given a probability vector $\mathbf{w}$ and a vector $\mathbf{v} = \mathbf{s} - \mathbf{t}$

$$|\mathbf{v}^\top \mathbf{w}| \leq \max(1-g, g)$$

Proof. First, note that $g - 1 \leq v_i < g$. This is because v_i is the value that must be subtracted from the i^{th} element of the search space to attain the value t_i, since $\mathbf{t} = \mathbf{s} - \mathbf{v}$ and g is the cutoff threshold for either rounding s_i up to 1 or down to 0. Therefore, in the case that we are rounding up, the most extreme value that can be subtracted is $g - 1$ (which is equivalent to adding $1 - g$ to arrive at 1). In the case that we are rounding down, the largest value we could subtract is strictly less than g. Now, we see that $|\mathbf{v}^\top \mathbf{w}| \leq \|\mathbf{v}\| \|\mathbf{w}\|$, and since $\mathbf{w}$ is a probability vector, $\|\mathbf{w}\| \leq 1$. Therefore, $|\mathbf{v}^\top \mathbf{w}| \leq \|\mathbf{v}\|$ and $\|\mathbf{v}\| \leq \max(g, 1-g)$, given that $g - 1$ is negative. Thus, $|\mathbf{v}^\top \mathbf{w}| \leq \max(g, 1-g)$. □

Theorem 5.5. Given that $\overline{\mathbf{P}}_{F,\mathbf{s}}$ is the averaged conditional distribution assigned by the search algorithm where our target set has a continuous satisfaction measure, $\overline{\mathbf{P}}_{F,\mathbf{t}}$ refers to the averaged conditional distribution assigned by the search algorithm when we use a discrete target set, and g is the threshold value for converting a continuous target set into a discrete target set (all elements with satisfaction greater than or equal to the threshold g are included in the target set and the rest are excluded):

$$\begin{aligned} & |q(\mathbf{s},F) - q(\mathbf{t},F)| \\ & \quad \leq |T| \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\overline{\mathbf{P}}_{F,\mathbf{s}} \parallel \overline{\mathbf{P}}_{F,\mathbf{t}})} + \max(1-g, g). \end{aligned}$$

This theorem bounds the difference in the success measure in the discrete and continuous case using the KL-divergence between the distributions learned in the continuous and discrete cases.

Proof. Consider $|q(\mathbf{s},F) - q(\mathbf{t},F)|$. Using the decomposable probability of success metrics we get:

$$|q(\mathbf{s},F) - q(\mathbf{t},F)| = |\mathbf{s}^\top \overline{\mathbf{P}}_{F,\mathbf{s}} - \mathbf{t}^\top \overline{\mathbf{P}}_{F,\mathbf{t}}|.$$

Now, we define a vector $\mathbf{v}$ such that $\mathbf{v} = \mathbf{s} - \mathbf{t}$. Therefore,

$$\begin{aligned} |\mathbf{s}^\top \overline{\mathbf{P}}_{F,\mathbf{s}} - \mathbf{t}^\top \overline{\mathbf{P}}_{F,\mathbf{t}}| &= |(\mathbf{t}+\mathbf{v})^\top \overline{\mathbf{P}}_{F,\mathbf{s}} - \mathbf{t}^\top \overline{\mathbf{P}}_{F,\mathbf{t}}| \\ &= |\mathbf{t}^\top \overline{\mathbf{P}}_{F,\mathbf{s}} - \mathbf{t}^\top \overline{\mathbf{P}}_{F,\mathbf{t}} + \mathbf{v}^\top \overline{\mathbf{P}}_{F,\mathbf{s}}| \\ &\leq |\mathbf{t}^\top \overline{\mathbf{P}}_{F,\mathbf{s}} - \mathbf{t}^\top \overline{\mathbf{P}}_{F,\mathbf{t}}| + |\mathbf{v}^\top \overline{\mathbf{P}}_{F,\mathbf{s}}|. \end{aligned}$$

Using Lemma 5.4, $|\mathbf{v}^\top \overline{\mathbf{P}}_{F,\mathbf{s}}| \leq \max(1-g, g)$. Therefore,

$$\begin{aligned} & |\mathbf{t}^\top \overline{\mathbf{P}}_{F,\mathbf{s}} - \mathbf{t}^\top \overline{\mathbf{P}}_{F,\mathbf{t}}| + |\mathbf{v}^\top \overline{\mathbf{P}}_{F,\mathbf{s}}| \\ & \qquad \leq |\mathbf{t}^\top (\overline{\mathbf{P}}_{F,\mathbf{s}} - \overline{\mathbf{P}}_{F,\mathbf{t}})| + \max(g, 1-g). \end{aligned}$$

Defining $r := \max(g, 1-g)$, we note that

$$\begin{aligned} & |\mathbf{t}^\top (\overline{\mathbf{P}}_{F,\mathbf{t}} - \overline{\mathbf{P}}_{F,\mathbf{s}})| + \max(g, 1-g) \\ & \qquad \leq |T| \sup_{\omega} (\overline{\mathbf{P}}_{F,\mathbf{s}}(\omega) - \overline{\mathbf{P}}_{F,\mathbf{t}}(\omega)) + r \\ & \qquad \leq |T| \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\overline{\mathbf{P}}_{F,\mathbf{s}} \parallel \overline{\mathbf{P}}_{F,\mathbf{t}})} + r, \end{aligned}$$

where the last step follows from Pinsker's inequality. Hence,

$$|q(\mathbf{s},F) - q(\mathbf{t},F)| \leq |T| \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\overline{\mathbf{P}}_{F,\mathbf{s}} \parallel \overline{\mathbf{P}}_{F,\mathbf{t}})} + r.$$

□

Example. We set $k = 1$. For all levels of quantization, note that $I(S;F)$ can be found computationally by directly computing $s(\omega)$ and $F(\omega)$ for all $\omega \in \Omega$. We know that the D_{KL} term will equal 0, since the KL divergence between two identical distributions is 0. Since $S(X)$ is independent from X, $H(S(X)|X) = H(S(X))$. We also know that

$$H(\mathcal{U}_{\tau_k;|\Omega|}) = -\sum_{s \in \tau_k} \frac{1}{|\tau_k|} \log_2 \left(\frac{1}{|\tau_k|} \right)$$

So then we simply need to find $|\tau_k|$ for a given $|\Omega|$. This can be done by simple combinatorics.

First, we set $m = 2$ and $u = 0.5$ with $c = 0.5$ so $s_i \in \{0, 0.5\}$. We know that $P(S(X) = 0) = \frac{n-2}{n}$ since there are exactly two non-zero elements (both 0.5 to sum up to $k = 1$), and then $P(S(X) = 0.5) = \frac{2}{n}$ so the distribution is known. To find $|\tau_k|$, we know that it is formed from all sets s that sum up to 1, so that is any set with exactly two elements with value 0.5, so then there are $\binom{n}{2}$ such sets. We also need to compute $|\tau_{k-u}|$ for the subtracted term in the bound, this is $\tau_{0.5}$. We know this happens when exactly one element is

0.5, so there are $\binom{n}{1} = n$ such sets.

Second, we set $m = 3$ and $u = \frac{2}{3}$ with $c = \frac{1}{3}$ so $s_i \in \{0, \frac{1}{3}, \frac{2}{3}\}$. There are two possible cases for satisfactory vectors in τ_k: ones that have 3 elements with $\frac{1}{3}$ and those with one $\frac{1}{3}$ and one $\frac{2}{3}$. There are $\binom{n}{3}$ sets satisfying the first case, and $\binom{n}{2} \cdot 2!$ sets satisfying the second so $|\tau_k| = \binom{n}{3} + \binom{n}{2} \cdot 2!$. We must also find $|\tau_{k-u}| = |\tau_{\frac{1}{3}}|$, this happens when exactly one element is a $\frac{1}{3}$ so there are $\binom{n}{1} = n$ such elements. Next, we must find the distribution over $S(X)$. Let a be the probability that we have the case of three $\frac{1}{3}$ (there are $\binom{n}{3}$ such sets) and q be the complementary probability of having one $\frac{1}{3}$ and one $\frac{2}{3}$ (there are $\binom{n}{2} \cdot 2!$ such sets). Then, $P(S(X) = 0) = a \cdot \frac{n-3}{n} + b \cdot \frac{n-2}{n}$ since there are 3 non-zero elements in a and 2 in b. $P(S(X) = \frac{1}{3}) = a \cdot \frac{3}{n} + b \cdot \frac{1}{n}$ since there are 3 $\frac{1}{3}$ in the a case and 1 in the b case. Finally, $P(S(X) = \frac{2}{3}) = b \cdot \frac{1}{n}$ since there is only a $\frac{2}{3}$ in the b case and only 1.

Now that every term has been determined either mathematically or computationally, we can combine them to compute the bounds as given in Section 6.