Probabilistic Error Guarantees for Abductive Inference

Kerria Pang-Naylor
AMISTAD Lab
Harvey Mudd College
Claremont, CA, USA
kpangnaylor@g.hmc.edu

Ian Li
AMISTAD Lab
Harvey Mudd College
Claremont, CA, USA
ili@g.hmc.edu

Kishore Rajesh
AMISTAD Lab
Harvey Mudd College
Claremont, CA, USA
kirajesh@g.hmc.edu

George D. Montañez
AMISTAD Lab
Harvey Mudd College
Claremont, CA, USA
gmontanez@g.hmc.edu

***Abstract*—Abductive reasoning is ubiquitous in artificial intelligence and everyday thinking. However, formal theories that provide probabilistic guarantees for abductive inference are lacking. We present a quantitative formalization of abductive logic that combines Bayesian probability with the interpretation of abduction as a search process within the Algorithmic Search Framework (ASF). By incorporating uncertainty in background knowledge, we establish two novel sets of probabilistic bounds on the success of abduction when (1) selecting the *single* most likely cause while assuming noiseless observations, and (2) selecting *any* cause above some probability threshold while accounting for noisy observations. To our knowledge, no existing abductive or general inference bounds account for noisy observations. Furthermore, while most existing abductive frameworks assume exact underlying prior and likelihood distributions, we assume only percentile-based confidence intervals for such values. These milder assumptions result in greater flexibility and applicability of our framework. We also explore additional information-theoretic results from the ASF and provide mathematical justifications for everyday abductive intuitions.**

***Index Terms*—Abductive Reasoning, Probabilistic Inference, Bayesian Decision Theory, Algorithmic Search Framework**

I. Introduction

Imagine a patient visits a doctor because of a persistent cough, fever, and shortness of breath. As the doctor considers these symptoms and the prevalence of certain illnesses in the area, the doctor may hypothesize that the patient has pneumonia. This is an example of abductive reasoning, or *abduction*.

Abduction is the process of finding the best causal explanation given some observed effects. Abductive reasoning can be categorized into strategies that can generate new hypotheses, known as *creative abduction*, and those that select the best candidate given a set of possible explanations, known as *selective abduction* [1]. We focus on selective abduction, which can be formalized with Bayesian Decision Theory [2]. Given observation(s) O, we select a hypothesis C_i from a finite set of hypotheses C. Per Bayesian probability, we denote $\Pr(C_i|O)$ as the *posterior*, where the most probable cause is that with the highest posterior. By Bayes' theorem,

$$\Pr(C_i|O) = \frac{\Pr(O|C_i)\Pr(C_i)}{\Pr(O)}.$$

However, during the hypothesis selection process, the relevant observations $\Pr(O)$ remain constant. Thus, the relevant form of Bayes' theorem becomes

$$\Pr(C_i|O) \propto \Pr(O|C_i)\Pr(C_i).$$

To perform selective abduction, one simply chooses the hypothesis whose likelihood and prior have the greatest product.

Abduction accompanies induction and deduction as one of three forms of logical reasoning [3], [4]. In supervised machine learning, inductive and abductive processes serve as the underlying logic behind model training and application (see Figure 1) [5]. While both inductive and abductive reasoning are applied ubiquitously in the field, inductive reasoning is currently the more well-understood process; we have already gained a theoretical understanding of inductive accuracy [6]–[8]. However, to our knowledge, there currently exist no formal frameworks with accuracy bounds for abductive reasoning.

In a broader context, artificial intelligence researchers such as Erik Larson argue that obtaining a theory of abduction is a necessary step towards bridging machine and human intelligence. Abduction, more specifically creative abduction, encapsulates human intuition or "guessing" capability lacking in current models. Larson describes machine understanding of abductive reasoning as the central "blind spot" of artificial intelligence:

> *"Abductive inference is required for general intelligence, purely inductively inspired techniques like machine learning remain inadequate...The field requires a fundamental theory of abduction." [9]*

Our work primarily aims to (1) provide currently lacking accuracy bounds for abductive reasoning and (2) serve as a preliminary version of this "fundamental theory of abduction" needed for abductive machine understanding. We propose a general probabilistic framework for *selective* abduction built from Bayesian Decision Theory [10] (detailed in Section III), serving as a jumping off point for future work on creative abduction. Through this Bayesian framework, we first derive upper and lower probabilistic bounds of abductive accuracy when assuming underlying q-percentile uncertainty bounds of prior and likelihood probabilities for each cause (Section IV). This first set of accuracy bounds treats successful

This work is supported by NSF grant no. 2243941 and the Harvey Mudd College Experiential Learning Fund.

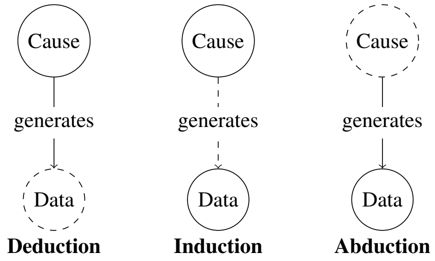

Fig. 1: Three methods of inference. The dotted lines show which part of each process is being inferred.

abduction as choosing the *single* true hypothesis assuming the selection of the single highest posterior. We then extend this by reframing abduction as a search process within the Algorithmic Search Framework (ASF) [11], which lets us describe and bound the probability of selecting *any* hypothesis with a posterior probability above a certain threshold while accounting for noisy observations (Section V-A). Lastly, in addition to deriving bounds on abductive accuracy, we apply the framework to quantitatively justify common-sense heuristic abduction (Section V-B, V-C).

II. Related work

We review applications of abductive logic in machine learning and artificial intelligence, and survey existing abductive frameworks and current literature on Bayesian inference.

A. Logic in Machine Learning

Peirce introduced abduction alongside induction and deduction as the three pillars of logical inference [12]. Induction, inferring causal relationships from data, is central to machine learning [5]. Inductive logic is core to the training process, where labeled examples are used to develop generalized relationships within a model. Deductive and abductive logic are employed within machine learning's underlying inductive framework by applying the relationships derived through inductive training [13]. Deduction facilitates data generation by selecting a class (cause) to produce feature data (observations). Conversely, abduction involves assigning class labels (causes) to unlabeled data (observations) using a trained model that embeds established causal relationships (see Figure 1) [13].

Induction corresponds to the training phase, where input-output relationships are learned, while abduction relates to classification, using known relationships to infer likely causes. Table I outlines the connections between logical inference [13] and machine learning.

From this perspective, machine learning applies abductive logic in model inference. For example, machine learning emulates the abductive reasoning used in spam detection and medical diagnosis by applying trained algorithms to unlabeled data (i.e., text from emails or radiology scans). However, model inference is just one of many applications of abduction in machine learning. Our work addresses the theoretical limits of the success of abductive reasoning generalizable to applications such as these.

TABLE I: Schematic outline of the processes of inference in supervised machine learning [13].

Logical Inference	**Machine Learning**
Induction: $P(a)$ $\therefore \forall w P(w)$	Training: $(x^1, y_1), ..., (x^n, y_n)$ $\therefore f : \mathcal{X} \to \mathcal{Y}$
Abduction: $Q(a)$ $P(w) \to Q(w)$ $\therefore P(a)$	Classification: $\mathbf{x}^m$ $y_m = f(\mathbf{x}^m)$ $\therefore y_m$

B. Applying Abduction in Machine Learning

In addition to its synonymy with the higher-level logic of model inference, abduction is central to several common machine-learning processes. Abduction is the underlying logic of Bayesian networks, which are used for tasks such as clustering, supervised classification, anomaly detection, and temporal modeling [14]. Additionally, *maximum a posteriori* (MAP) applies abductive reasoning through Bayes' theorem to optimize model parameters. In relational learning, where data is represented through relationships with other data, abduction guides search and generates missing input data [13]. In computer vision, integrating abductive reasoning with convolutional neural networks (CNNs) enhances spatial-temporal reasoning and image segmentation tasks [15].

C. Formalizations of Abduction

Various formalizations of abduction have been explored in symbolic AI literature [16]. Set-cover-based approaches involve selecting a subset of hypotheses from a larger set, requiring complete causal relationships [17]. Knowledge-level approaches propose explanations based on beliefs [18]. Abductive Logic Programming (ALP) represents inferences as entailments from a prior knowledge base to the veracity of specific causes [19].

Probabilistic Horn Abduction extends Prolog by combining exact probabilities of hypotheses with Bayes' theorem to generate posterior probabilities built from multiple observations [20], [21]. Unlike our proposed framework, it assumes exact prior and likelihood probabilities and does not incorporate confidence ranges for these distributions [20]. Developments in probabilistic abductive logic programming [22], [23] depart from our work in similar ways, as exact probabilities are assumed and general bounds for abductive success are not provided.

A recently developed framework applying stochastic mathematical systems (SMSs) models abduction by representing reasoning as stochastic systems, with the human reasoner SMS generating hypotheses and an oracle SMS evaluating their validity based on explanatory power and evidence [24]. Like Probabilistic Horn Abduction, it also does not account for the uncertainties in underlying distributions.

These methods lack probabilistic guarantees for the correctness of the abductive inferences and do not quantify associated uncertainties. Our approach addresses this gap by integrating formal machine learning frameworks, which allows for more precise quantification of the uncertainties involved in abductive inferences.

D. Bayesian Inference

Bayesian inference forms the basis of our framework, deriving accuracy bounds using q_p and q_l confidence intervals for prior and likelihood distributions, respectively. These intervals represent confidence in causal relationships (q_l) and general world knowledge (q_p), providing flexibility in representation.

Bayesian inference estimations and bounds are well-explored in the literature, with numerous known methods of deriving accuracy bounds for inference of specific algorithms or tasks [25]–[27] However, general methods for deriving bounds using techniques like multi-valued mapping [28] or prior measure intervals [29] are less common. To our knowledge, no existing method derives Bayesian inference bounds based on specific prior and likelihood confidence intervals with probabilities q_l, as our framework does.

Our work is the first to leverage the ASF [11] to construct a formalization of abduction or abduction by Bayesian inference. Unlike other established frameworks [20], [21], [30], the ASF accounts for noisy observations – observations that may not fully reflect "true" events. The framework makes very few assumptions of given information resources, F, which (in the case of abduction) embeds observation data. Such data is abstracted as binary strings, with no conditions placed on what form the binary strings take, only that we have functions available to extract feedback from the strings for individual search queries. Thus, with no restrictions placed on the information resources, the ASF accommodates both noisy and noiseless observations. To our knowledge, there are no abductive or general inference bounds with this specific property. Existing work has only analyzed the correlation of real dataset noise with the accuracy of Bayesian inference for specific algorithms, assuming specific data qualities [31].

III. Preliminaries

We formalize the fundamental building blocks of abduction, causes and observations, as vectors, and use this basis to establish the likelihood and posterior uncertainty intervals on which the abductive search process relies.

A. Vectorizing Observations

We formalize observations as binary vectors, where each scalar component corresponds to the existence of a specific *observation feature* or certain observed outcome. For example, suppose you swallow an unknown pill and then your headache disappears. A representative observation vector might be $\langle 1, 1\rangle$ with each feature representing (1) "Did you swallow a pill?" and (2) "Did the headache go away?" (respectively).

Definition III.1. ($\mathcal{O}$) Let $\mathcal{O}$ denote the vector space of discrete topology containing all binary-featured observation vectors.

Since any outcome must strictly occur or not occur, the set of possibilities within $\mathcal{O}$ is mutually exclusive and collectively exhaustive. In the case where an observation is a continuous variable, such as temperature, we would convert the variable by adding additional features representing levels of the value, such as ["cold", "lukewarm","hot"].[1]

TABLE II: Example likelihood distribution for cause aspirin.

Pill taken?	Headache relieved?	$\mathbf{x}$	$\Pr(\mathbf{x}\|do(\text{aspirin} = \texttt{True}))$
no	no	$\langle 0, 0\rangle$	0.05
no	yes	$\langle 0, 1\rangle$	0.10
yes	no	$\langle 1, 0\rangle$	0.15
yes	yes	$\langle 1, 1\rangle$	0.70

B. Vectorizing Causes and Likelihood Probability Mass Functions

A cause C_i has some probability of instigating any possible observation vector $\mathbf{x} \in \mathcal{O}$, inducing a conditional probability mass distribution (i.e., likelihood function) $\Pr(\mathbf{x}|C_i)$ over all observations $\mathbf{x} \in \mathcal{O}$. Note that every observation $\mathbf{x} \in \mathcal{O}$ is disjoint, and we assume exactly one observation vector is produced and observed.

Following the earlier example, the likelihood distribution over the observation space for the cause "aspirin" expresses the probability that, *assuming* aspirin was taken, phenomena $\mathbf{x} \in \mathcal{O}$ would follow. Knowing that aspirin typically relieves headaches and is ingested in pill form, the likelihood distribution over $\mathcal{O}$ with dimensions {"Pill taken?", "Headache relieved?"} may be similar to Table 2. Such a likelihood distribution depends only on the cause C_i, and will act over $\mathcal{O}$.

Assuming that exactly one of the observation vectors must occur, we know that the probabilities for each collectively must sum to one. Considering all the possible ways there are to assign probabilities to a collectively exhaustive and mutually exclusive set of options forms a mathematical simplex, $\mathcal{S}$. For k observation features (where $\dim(\mathcal{O}) = k$), simplex $\mathcal{S}$ forms a continuous $2^k - 1$ dimensional hyperplane containing all possible "cause vectors", each corresponding with some likelihood probability mass function over the 2^k observation vectors in $\mathcal{O}$. Each scalar component of a 2^k dimensional "cause" vector $\mathbf{c} \in \mathcal{S}$ denotes how much probability mass is placed on a corresponding observation vector in $\mathcal{O}$. Since we define a "cause" as the event representation of a likelihood distribution over $\mathcal{O}$, a single cause vector in $\mathcal{O}$ can actually represent multiple concurrent events or causes.

Ensuring that every cause $\mathbf{c} \in \mathcal{S}$ corresponds to a valid probability mass function on $\mathcal{O}$ requires the following two properties: (1) the simplex is bounded within $[0, 1]$ on every dimension such that no $\mathbf{c} \in \mathcal{S}$ holds a component that indicates an invalid probability, and (2) the sum of all components of a cause vector equals 1.

C. Defining Posterior Confidence Bounds

During the decision-making process, we compare different posterior probabilities for the same observation $\mathbf{x}$. Since the evidence, $\Pr(\mathbf{x})$, is constant, we will only compare the

[1] Note that any probability mass function over $\mathcal{O}$ would, by default, place zero mass on contradictory observation vectors, such as one that is both "hot" and "cold".

product of the likelihood and prior across causes, namely, $\Pr(\mathbf{x}|\mathbf{c})\Pr(\mathbf{c})$, which we denote as the "posterior".

In applications, we often lack these exact likelihood and prior distributions. Instead, we may estimate such probabilities through numerical techniques, including asymptotic estimations, Monte Carlo methods, numerical integration, and various sampling methods [32]–[34]. Other distribution estimation methods include smoothing and reduction methods, and Markov chain algorithms can be further used to combine estimation methods [32]. To account for uncertainty, we estimate likelihood, prior, and posterior probabilities through confidence intervals.

We define two functions denoting the upper bound likelihood probability, $l_U(\mathbf{c},\mathbf{x})$, and lower bound likelihood, $l_L(\mathbf{c},\mathbf{x})$, of the q_l-percentile likelihood uncertainty interval, where $l_U(\mathbf{c},\mathbf{x}) \geq l_L(\mathbf{c},\mathbf{x})$. The prior q_r-percentile uncertainty interval is similarly represented through an upper and lower bound $r_U(\mathbf{c})$ and $r_L(\mathbf{c})$ (respectively). The upper and lower confidence bounds of the posterior, $p_U(\mathbf{c},\mathbf{x})$ and $p_L(\mathbf{c},\mathbf{x})$, can then be found by simply multiplying the upper or lower bounds of the likelihood and prior probabilities together:

$$p_U(\mathbf{c},\mathbf{x}) = l_U(\mathbf{c},\mathbf{x})r_U(\mathbf{c}),$$

$$p_L(\mathbf{c},\mathbf{x}) = l_L(\mathbf{c},\mathbf{x})r_L(\mathbf{c}).$$

This bound assumes there is a q_l probability that the likelihood lies in its q_l-percentile interval $[l_L(\mathbf{c},\mathbf{x}), l_U(\mathbf{c},\mathbf{x})]$ and, likewise, that there is a q_r probability that the prior lies in its q_r-percentile interval $[r_L(\mathbf{c},\mathbf{x}), r_U(\mathbf{c},\mathbf{x})]$. Thus, the interval $[p_U(\mathbf{c},\mathbf{x}), p_L(\mathbf{c},\mathbf{x})]$ defines the q-percentile confidence interval for posterior $\Pr(\mathbf{c}|\mathbf{x})$ where $q = q_l q_r$.

D. Narrowing the Space of Possible Causes

We have established $\mathcal{S}$ as the *infinite* space containing all possible likelihood distributions over $\mathcal{O}$ and, thus, the space of *all* possible causes. However, this space includes likelihood distributions generated by causes that are implausible. In the real world, we often choose the most likely cause from a smaller set of plausible causes; for example, one would not consider an atomic bomb to be a plausible cause for your headache disappearing. Rather than considering the entirety of $\mathcal{S}$ as the pool of possible causes, we assume that some finite subset $\mathcal{C} \subset \mathcal{S}$ with cardinality $k = |\mathcal{C}|$ has been pre-selected as the finite set of *plausible* causes assumed to contain the true cause. We further assume $\mathcal{C}$ includes a "cause" C_{other}, whose posterior encapsulates the (likely low) combined probability of all other causes in $\mathcal{S}$ occurring. With this, we assume that all causes in $\mathcal{C}$ are disjoint and that $\mathcal{C}$ contains the *one* true explanation for observation $\mathbf{x}$ (namely, what actually caused it).

Definition III.2. ($\mathcal{C}$) Let $\mathcal{C} \subset \mathcal{S}$ denote the relevant finite subset of possible cause vectors in $\mathcal{S}$.

For notational simplicity, we additionally denote each cause as $C_i \in \mathcal{C}$ and its corresponding "true" posterior probability as M_i in posterior set $\mathcal{M}$. We likewise simplify the notation of the upper and lower bounds of q-percentile uncertainty interval posterior M_i as follows: from $p_U(\mathbf{c},\mathbf{x})$ and $p_L(\mathbf{c},\mathbf{x})$ to u_i and l_i, respectively. For future reference, we define the following:

Definition III.3. (M_i) Let $M_i \in \mathcal{M}$ denote the "true" posterior probability of cause $C_i \in \mathcal{C}$, where $M_i = \Pr(C_i|\mathbf{x})\Pr(C_i)$. Then M_i falls into the following uncertainty interval with probability q:

$$M_i \in [l_i, u_i].$$

Since we assume each $C_i \in \mathcal{C}$ is disjoint, and that $\mathcal{C}$ surely contains the true explanation for observation $\mathbf{x}$, each posterior probability $\Pr(C_i|\mathbf{x})$ sums to 1. Thus,

$$\sum_{M_i \in \mathcal{M}} M_i = \Pr(\mathbf{x}).$$

Definition III.4. ($\mathcal{U}$) Let $\mathcal{U}$ denote the set containing the q-percentile uncertainty interval bounds $[l_i, u_i]$ for each posterior $M_i \in \mathcal{M}$.

IV. Abduction by Bayesian Inference

A. Cause Selection with Uncertainty Intervals

Given the set of q-percentile confidence posterior probability uncertainty bounds $[l_i, u_i] \in \mathcal{U}$ for each cause $C_i \in \mathcal{C}$, one selects the cause whose *point estimate posterior probability* is highest. Since the true posterior probabilities of each cause are unknown, this process may incorrectly select a cause whose posterior is not the true maximum. We quantify this rate of incorrect selection in the case where every posterior $M_i \in \mathcal{M}$ is contained in respective confidence bound $[l_i, u_i]$. Let predicate $\text{IsMax}(M_i)$ denote whether posterior M_i is truly the highest posterior. We first define the probability range where the maximum posterior must lie, $[l, u]$.

Definition IV.1. Let each posterior $M_i \in \mathcal{M}$ occur within q-percentile confidence interval $[l_i, u_i] \in \mathcal{U}$. Then, we set

$$l = \max(\{l_i | i \in \mathbb{Z}_+, i \leq |M|\}),$$

$$u = \max(\{u_i | i \in \mathbb{Z}_+, i \leq |M|\}).$$

Proposition IV.1. *Assuming that every $M_i \in \mathcal{M}$ lies in respective q-percentile confidence interval $[l_i, u_i] \in \mathcal{U}$, the max posterior is bounded by u and l.*

Thus, in the case that *every* confidence bound fully contains its respective posterior almost surely (instead of just with probability q), any posterior M_i whose uncertainty bounds $[l_i, u_i]$ overlap with $[l, u]$ is potentially the maximum posterior with some probability $\Pr(\text{IsMax}(M_i))$.

Theorem IV.2. *Let $\mathcal{M}' \subseteq \mathcal{M}$ denote the set of posteriors whose confidence intervals intersect with $[l, u]$. Let $p_{M_i}(x)$ be the probability density function of the position of M_i. The probability that $M_i \in \mathcal{M}'$ is the maximum posterior is as follows:*

$\Pr(\textit{IsMax}(M_i)) =$

$$\int_l^u \Pr\left(\bigcap_{j=1, j\neq i}^{|\mathcal{M}|} (M_j < x) \,\middle|\, M_i = x\right) p_{M_i}(x)dx$$

This accounts for any estimated posterior probability distribution within $[l_i, u_i]$, but assumes M_i is contained by $[l_i, u_i]$ with probability 1.

B. Bayes Error Rate

However, even assuming the cause with the true highest posterior is successfully identified, there is the unavoidable error from non-zero posteriors of the "losing" categories. The true cause of a feature may simply not have the highest posterior. This minimum achievable error is expressed by Bayes Error Rate (BER):

Definition IV.3. (ϵ, [35]) Let ϵ denote Bayes multiclass error rate (BER) for every $C_i \in \mathcal{C}$. For $|\mathcal{C}| = k$ possible causes:

$$\epsilon = 1 - \int \Pr(\mathbf{x}) \max_i \Pr(C_i|\mathbf{x}) d\mathbf{x}.$$

However, the formula above is often impractical to compute for $k > 2$ causes. Instead, one can derive bounds for the multi-cause BER with techniques such as the Bhattacharyya bound, estimations using Friedman-Rafsky test statistics, and non-parametric bounds using Henze-Penrose divergence [36]. We adopt a recent method[2] of upper bounding BER through global minimal spanning trees [35] and adopt a pairwise computational lower bounding method for BER [37].

Definition IV.4. (ϵ_{upper}, [35]) Let ϵ_{upper} denote the upper bound of BER such that $\epsilon \leq \epsilon_{\text{upper}}$. Then, for $|\mathcal{C}| = k$,

$$\epsilon_{\text{upper}} = 2 \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \delta_{ij}$$

where $\delta_{ij} := \int \frac{\Pr(C_i) \Pr(C_j) \Pr(\mathbf{x}|C_i) \Pr(\mathbf{x}|C_j)}{\Pr(C_i) \Pr(\mathbf{x}|C_i) + \Pr(C_j) \Pr(\mathbf{x}|C_j)} d\mathbf{x}$.

Definition IV.5. (ϵ_{lower}, [38], [37]) Let ϵ_{lower} denote the lower bound of BER such that $\epsilon \geq \epsilon_{\text{lower}}$. BER may be lower bounded by applying pairwise computations of Bayes error ϵ_{ij} for i and j between every unique cause pair (C_i, C_j) where $C_i \in \mathcal{C}, C_i \in \mathcal{C}, i \neq j$:

$$\epsilon_{lower} = \frac{2}{k} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} (\Pr(C_i) + \Pr(C_j))\epsilon_{ij}.$$

C. Abductive Error Guarantees

Assume an algorithm selects from the set of possible causes $\mathcal{C}$ the cause with the highest estimated posterior. The preceding subsections detail the two possible sources of error:

1) Incomplete or imprecise background information (e.g., not knowing all the potential causes and causal relationships). This uncertainty is represented through q-percentile posterior confidence intervals in $\mathcal{U}$.
2) The true cause is not the cause with the highest true posterior. If the exact likelihood and prior is given, this minimum achievable error is simply expressed through the Bayes Error Rate (Definition IV.3).

We derive bounds of the error rate by combining these two possible sources of error. Let W denote the event of incorrect abduction (not selecting the true cause). Then, the probability of correctly selecting the maximum posterior M_i and incorrect abduction is

$$\begin{aligned} \Pr(\text{W}, \text{IsMax}(M_i)) &= \Pr(\text{W}|\text{IsMax}(M_i)) \Pr(\text{IsMax}(M_i)) \\ &= \epsilon \Pr(\text{IsMax}(M_i)). \end{aligned}$$

The probability of both incorrectly selecting the maximum posterior and incorrect abduction is

$$\begin{aligned} \Pr(\text{W}, \neg\text{IsMax}(M_i)) &= \Pr(\text{W}|\neg\text{IsMax}(M_i)) \Pr(\neg\text{IsMax}(M_i)) \\ &= (1 - \Pr(M_i|\mathbf{x}))(1 - \Pr(\text{IsMax}(M_i))). \end{aligned}$$

Such definitions let us derive upper and lower bounds for the error rate assuming that all posteriors $M_i \in \mathcal{M}$ lie in q-percentile confidence intervals $[l_i, u_i] \in \mathcal{U}$ with probability 1. Let γ_i denote the error rate given this assumption.

Theorem IV.6. *Let γ_i denote the error rate of selected cause C_i when assuming posterior M_i lies in confidence interval $[l_i, u_i]$ almost surely. Then, γ_i is bounded above by*

$$\gamma_i \leq \epsilon_{upper} \Pr(\textit{IsMax}(M_i)) + (1 - l_i)(1 - \Pr(\textit{IsMax}(M_i)))$$

where ϵ_{upper} may be derived by Definition IV.4

Theorem IV.7. *Let γ_i denote the error rate of selected cause C_i when assuming posterior M_i lies in confidence interval $[l_i, u_i]$ almost surely. Then, γ_i is bounded below by*

$$\gamma_i \geq \epsilon_{lower} \Pr(\textit{IsMax}(M_i)) + (1 - u_i)(1 - \Pr(\textit{IsMax}(M_i)))$$

where ϵ_{lower} may be derived by Definition IV.5

We extend this result to the general case where all posteriors $M_i \in \mathcal{M}$ are assumed to each lie in their respective confidence intervals $[l_i, u_i] \in \mathcal{U}$ with probability q.

Theorem IV.8. *Let q^k be the probability that all $M_i \in \mathcal{M}$ lie in their respective confidence bounds $[l_i, u_i] \in \mathcal{U}$. Let $\gamma_{i,\ upper}$ be the upper bound of γ_i defined in Theorem IV.6. Then, the upper bound of the general error rate is given by*

$$\Pr(W) \leq 1 - q^k(1 - \gamma_{i,\ upper}).$$

Theorem IV.9. *Let q^k be the probability that all $M_i \in \mathcal{M}$ lie in their respective confidence bounds $[l_i, u_i] \in \mathcal{U}$. Let $\gamma_{i,\ lower}$ be the lower bound of γ_i defined in Theorem IV.7. Then, the lower bound of the general error rate is given by*

$$\Pr(W) \geq \gamma_{i,\ lower} q^k.$$

We should note that the bounds presented in this section assume noiseless observations. That is, we assume observation **x** is a wholly accurate description of the "true" outcomes of a cause. A noisy observation vector may have entries that deviate from the "true" outcome of a cause, akin to the possibility of a faulty observer or inaccurate data pipeline with which observations is processed (i.e., faulty equipment, random errors in sampling, etc.). The next section explores a

[2] This method provides a tighter bound than aforementioned techniques [35].

different set of bounds describing the selection of *any* cause whose probability is above some threshold. With this broader definition of "success," we can account for noisy observations through applying the Algorithmic Search Framework [11].

V. Search and Heuristic Applications

The Algorithmic Search Framework (ASF) characterizes learning problems as search, allowing one to equate the chance of success of any learning algorithm to that of a search process described by the three-tuple (Ω, T, F) – the *search space*, *target set*, and *external information resource*, respectively [11]. This framework formalizes the seminal work of Mitchell [39] and extends results beyond binary classification problems [40]. The ASF provides formal bounds accounting for noise, and formalizes insights into the frequency of favorable search strategies and problems [40].

We have previously discussed abductive success in terms of finding the one "true" cause for some observation vector (which may or may not have the highest posterior) *assuming* the selection of the single highest posterior. Furthermore, we assumed noiseless observations. By reframing the ASF for abduction, we describe an algorithm's ability to identify the cause(s) with posteriors above some threshold in terms of information-theoretic properties within (Ω, T, F) and generalize to noisy observation vectors.

A. ASF: Success of Abduction through Search

We define each term of (Ω, T, F) as follows.

Search Space (Ω) constitutes the finite set of pre-selected, plausible causes for the given observation vector $\mathbf{x}$; it is synonymous with $\mathcal{C}$ defined in III.2. P_i over search space Ω denotes the probability distribution over the space at step i, and $P_i(T)$ is the probability of success – namely, the amount of probability mass placed on the target set T at time i [11]. In our adaptation, P_i denotes the posterior distribution of $\Pr(C_i|\mathbf{x})$ over all possible causes C_i in Ω. P_i may be derived from aforementioned bounds $[l_i, u_i] \in \mathcal{U}$ of posterior-adjacent value $\Pr(C_i|\mathbf{x})\Pr(\mathbf{x})$ (Definition IV.1) with two modifications: (1) P_i denotes the *point estimate probability* of the posterior within these confidence bounds, and (2) this point estimate of $\Pr(C_i|\mathbf{x})\Pr(\mathbf{x})$ is inversely scaled by $\Pr(\mathbf{x})$ such that P_i is a valid probability mass function that sums to one.

Target Set (T), a subset of the search space Ω, contains the set of the "more plausible" causes with posterior probability P_i above or at *minimum performance value* in $(0, 1]$. Search aims to identify causes in Ω that lie in T, a task whose difficulty increases as the threshold for T rises.

External Information Resource (F) is a finite-length binary string drawn from a distribution with an "API"-like interface, meaning one can extract information from F [11]. In our case, it embeds (1) the observation vector $\mathbf{x}$ whose cause we determine, and (2) the upper and lower bounds of the q-confidence intervals for likelihood and prior probabilities across Ω for every cause $C_i \in \Omega$. More specifically, F contains the likelihood bounds $l_U(\mathbf{c}, \mathbf{x})$ and $l_L(\mathbf{c}, \mathbf{x})$ and prior bounds $r_U(\mathbf{c}, \mathbf{x})$ and $r_L(\mathbf{c}, \mathbf{x})$, which inform the construction of posterior probability distribution P_i over Ω for the search process as defined previously. Since F is a function of random data, it is itself a random variable.

Framing abduction through the ASF, we apply established derivations of the maximal probability of success defined in terms of information-theoretic properties of (Ω, T, F) and the complexity of the search problem [11]. And as explained in Section IV, the ASF places few restrictions on information resources F, and thus allows for both noisy or noiseless observations.

Theorem V.1. *[11] The probability of a successful abduction, q, is bounded above by*

$$q \leq \frac{I(T;F) + D(P_T||\mathcal{U}_T) + 1}{I_\Omega},$$

where $I_\Omega = -\log \frac{|T|}{|\Omega|}$, $D(P_T||\mathcal{U}_T)$ *is the Kullback-Leibler divergence between the marginal distribution on target sets and the uniform distribution on possible target sets, and* $I(T;F)$ *is the mutual information between the target and observation.*

We interpret $I(T;F)$ as the dependence between the target set and the observation, $D(P_T||\mathcal{U}_T)$ as the non-uniformness of the target, and I_Ω as the sparseness of the targets inside the search space. When the true cause is highly correlated with the observations (i.e., less random), the achievable success rate is high. When the search space consists of a large number of causes, the achievable success rate is lower. This gives us an additional information-theoretic upper bound on the probability of successful abduction.

B. ASF: High-Likelihood Causes are Rare

Any high-posterior cause must also confer high-likelihood to observed effects, due to the multiplicative nature of posterior computation. Yet a cause can only make an observation vector more probable at the cost of making others less probable. Such high-likelihood causes must necessarily be rare to the degree they confer high joint-probability on the observations, as shown by the following theorem [11].

Theorem V.2. *(Famine of Favorable Strategies Theorem, [11]) For any fixed search problem* (Ω, T, F), *set of probability mass functions* $\mathcal{P} = \{P : P \in [0,1]^{|\Omega|}, \sum_j P_j = 1\}$, *and a fixed threshold* $q_{\min} \in [0, 1]$,

$$\frac{\mu(\mathcal{G}_{t,q_{\min}})}{\mu(\mathcal{G}_\mathcal{P})} \leq \frac{p}{q_{\min}},$$

where $p = \frac{|T|}{|\Omega|}, \mathcal{G}_{t,q_{\min}} = \{P : P \in \mathcal{P}, t^\top P \geq q_{\min}\}$, *and* μ *is Lebesgue measure.*

In contrast to Section V-A, we consider a different search problem in applying Theorem V.2. The search space Ω no longer consists of posteriors, but is now the space of all possible observation vectors, some of which are "close enough" to the true vector to comprise a noisy target set, T. Causes sample observation vectors by producing effects: a blind, weighted search. F becomes irrelevant. Theorem V.2 then tells us that the proportion of causes which confer at least $q_{\min}$ probability to the observation set is necessarily small whenever $q_{\min}$ is high, if we are only willing to tolerate so much noise in our observations (leading to small $|T|$).

One might argue that although not many causes can confer high *joint* likelihood to the observations, several independent causes might together constitute an abductive explanation for the observed phenomena, if each sufficiently raises the likelihood of a *single* observed feature. Simple arithmetic renders this possibility unpersuasive. Assuming independent causes for each observed feature, the probability of jointly occurring outcomes in an observation vector **x** scales exponentially with $|\mathbf{x}|$ or the number of features. For instance, if two features have a 50/50 chance of occurring coincidentally, then the chance of them occurring together is $1/2 \cdot 1/2 = 1/4$. For four such features, the probability drops to 6.25%. Thus, the coincidental co-occurrence of independent causes that together explain an observation vector is unlikely as the number of observations increases.

C. ASF: Non-coincidental Causes

When we perform abduction, we often must reason about whether a potential cause is simply coincidental. When one defines only high probability causes as "plausible" (i.e., threshold $q_{\min}$ is closer to 1), then intuitively, it seems we are less likely to select unrelated coincidental causes. Theorem V.2 formally confirms this intuition: if we desire a strong probability of success, fewer causes will satisfy the threshold $q_{\min}$. In this case, the target set becomes more sparse, and it will be difficult for an algorithm to successfully find targets from the search space. Therefore, favorable causes are rare, and are thus more likely to be causal than coincidental.

D. Increasing Certainty in Abductive Inference

Inductive inference error guarantees derive their strength from data abundance: increasing the number of observed examples typically increases the tightness of such bounds. In contrast, abductive inference proceeds from a single observation. How do we increase confidence in our abductive judgment? In the real world, our confidence in abductive reasoning typically depends on the amount of evidence supporting or contradicting a potential hypothesis. Though consisting of a single example, there are often many features of that observation, which may or may not be well-explained by a proposed cause. This suggests a "horizontal" mode of confirmation built on many conditionally independent features, rather than the "vertical" mode of confirmation based on many observed examples typical of inductive inference. We note the importance of conditional independence among features, since features that necessarily imply each other even given the cause do not give us additional confidence in our abductive judgment.

Recall that observation vector $\mathbf{x} \in \mathcal{O}$ consists of binary features representing the existence or non-existence of some conditionally independent observed outcome. Letting $x_1, \ldots, x_n$ represent each feature of $\mathbf{x} \in \mathcal{O}$ where $|\mathbf{x}| = \dim(O) = n$, we quantitatively demonstrate this phenomenon with the following.

Theorem V.3. *For each conditionally independent feature $x_1, \ldots, x_n$, define $\beta_i > 0$ such that for all $i = 1...n$,*

$$\Pr(x_i|C) = \beta_i \Pr(x_i|\overline{C}).$$

Let $\beta = \sqrt[n]{\prod_i^n \beta_i}$, the geometric mean of $\{\beta_i\}$. If $\beta > 1$, then

$$\lim_{n\to\infty} \frac{\Pr(x_1, \ldots, x_n|C)}{\Pr(x_1, \ldots, x_n|\overline{C})} = \lim_{n\to\infty} \beta^n = \infty.$$

Each conditionally independent observation feature can either support ($\beta_i > 1$) or contradict ($\beta_i < 1$) the proposed cause. If features support the current cause C on average (i.e., $\beta > 1$), then the confidence of abduction (ratio between likelihood under C over $\overline{C}$) approaches infinity as the number of (on average) supporting features increases.

VI. Discussion

Formalizing abduction as the selection high posterior cause(s) from a finite pool of causes, we establish two novel sets of probabilistic bounds on the success of abduction when (1) selecting the *single* most likely cause while assuming noiseless observations (Theorems IV.8 and IV.9), and (2) selecting *any* cause above some probability threshold while accounting for noisy observations (Theorem V.1).

Regarding the practicality of our results, it has been shown that bounds on the Bayes Error Rate can be empirically estimated by learning from training data instead of density estimation [35]. Unlike traditional methods for estimating BER, such as those based on pairwise HP divergence or generalized Jensen-Shannon (JS) divergence, which becomes computationally infeasible as the number of classes or dimensions increases. The GHP-based method is shown to be computationally more efficient, making it more suitable for large-scale applications like neural networks [35]. Then, in practice, it is possible to model a selective abduction problem using a Bayesian Neural Network and obtain approximate posterior distributions [41], [42], which can be directly used in our bounds for abductive inference.

The presented formalization and bounds are also tools to understand the limits of human-like reasoning abilities, and with it, the limits of decision-making in artificial intelligence. Theorem V.2 demonstrates how high-likelihood causes are rare, and more supporting observations increase confidence in a causal relationship rather than coincidence. V.3 captures the degree of certainty of our everyday abductive inferences. For example, suppose we must decide whether to convict a suspect of a crime. If pieces of evidence collectively support that the suspect is guilty, our confidence to convict grows as the amount of such evidence grows. Conversely, we would be less confident if pieces of evidence were contradictory or refuted a suspect's involvement.

VII. Conclusion

Abductive reasoning is a key component of critical thinking and discovery. State-of-the-art artificial intelligence is currently incapable of performing abductive reasoning at a human level.

To achieve true human-like reasoning, it is important to consider the process of abduction and its innate limitations.

Our work formalizes selective abduction, deriving formal error guarantees for abductive reasoning within a finite space of causes. Future work might explore creative abduction using our framework as a starting point. Creative abduction can be represented through a search space that is potentially infinite. Rather than filtering $\mathcal{S}$ to a finite pool $\mathcal{C}$, we represent hypothesis generation as optimization within a countably infinite subset of $\mathcal{S}$. Statistical bounds within such a framework would hold implications for general scientific reasoning and human creativity.

References

[1] G. Schurz, "Patterns of Abduction," *Synthese*, vol. 164, no. 2, pp. 201–234, Aug. 2007. [Online]. Available: https://doi.org/10.1007/s11229-007-9223-4

[2] J.-W. Romeijn, "Abducted by Bayesians?" *Journal of Applied Logic*, vol. 11, no. 4, pp. 430–439, 2013.

[3] C. T. Rodrigues, "The Method of Scientific Discovery in Peirce's Philosophy: Deduction, Induction, and Abduction," *Logica Universalis*, vol. 5, pp. 127–164, 2011.

[4] C. S. Peirce, M. R. Cohen, and J. Dewey, "Deduction, Induction, and Hypothesis," in *Chance, love, and logic*. Routledge, 2017, pp. 131–153.

[5] R. J. Mooney, "Integrating Abduction and Induction in Machine Learning," *Abduction and Induction: essays on their relation and integration*, pp. 181–191, 2000.

[6] T. G. Dietterich, "Limitations on Inductive Learningg," in *Proceedings of the Sixth International Workshop on Machine Learning*. Elsevier, 1989, pp. 124–128.

[7] R. Cosentino, R. Balestriero, R. Baranuik, and B. Aazhang, "Deep Autoencoders: From Understanding to Generalization Guarantees," in *Mathematical and Scientific Machine Learning*. PMLR, 2022, pp. 197–222.

[8] S. Garg, S. Balakrishnan, Z. Kolter, and Z. Lipton, "Ratt: Leveraging Unlabeled Data to Guarantee Generalization," in *International Conference on Machine Learning*. PMLR, 2021, pp. 3598–3609.

[9] E. J. Larson, *The Myth of Artificial Intelligence*. London, England: Harvard University Press, Apr. 2021, pp. 89–234.

[10] J. O. Berger, *Identification of Correlated Damage Parameters Under Noise and Bias Using Bayesian Inference*. Springer Science & Business Media, 2013, pp. 1–45, 74–117.

[11] G. D. Montañez, "The Famine of Forte: Few Search Problems Greatly Favor Your Algorithm," in *Systems, Man, and Cybernetics (SMC), 2017 IEEE International Conference on*. IEEE, 2017, pp. 477–482.

[12] T. Shanahan, "The First Moment of Scientific Inquiry: CS Peirce on the Logic of Abduction," *Transactions of the Charles S. Peirce Society*, vol. 22, no. 4, pp. 449–466, 1986.

[13] F. Bergadano, V. Cutello, and D. Gunetti, *Abduction in Machine Learning*. Dordrecht: Springer Netherlands, 2000, pp. 197–229. [Online]. Available: https://doi.org/10.1007/978-94-017-1733-5_5

[14] B. Mihaljević, C. Bielza, and P. Larrañaga, "Bayesian Networks for Interpretable Machine Learning and Optimization," *Neurocomputing*, vol. 456, pp. 648–665, 2021.

[15] C. Zhang, B. Jia, S.-C. Zhu, and Y. Zhu, "Abstract Spatial-Temporal Reasoning via Probabilistic Abduction and Execution," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9731–9741, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:232380362

[16] G. Paul, *AI Approaches to Abduction*. Dordrecht: Springer Netherlands, 2000, pp. 35–98. [Online]. Available: https://doi.org/10.1007/978-94-017-1733-5_2

[17] D. Allemang, M. C. Tanner, T. Bylander, and J. R. Josephson, "Computational Complexity of Hypothesis Assembly," in *IJCAI, International Joint Conference on Artificial Intelligence*, vol. 87, 1987, pp. 1112–1117.

[18] H. J. Levesque, "A Knowledge-Level Account of Abduction," in *IJCAI, International Joint Conference on Artificial Intelligence*, vol. 11, 1989, pp. 1061–1067.

[19] A. C. Kakas, R. A. Kowalski, and F. Toni, "Abductive Logic Programming," *Journal of Logic and Computation*, vol. 2, no. 6, pp. 719–770, 12 1992. [Online]. Available: https://doi.org/10.1093/logcom/2.6.719

[20] D. Poole, "Representing Diagnostic Knowledge for Probabilistic Horn Abduction," in *IJCAI*, 1991, pp. 1129–1137.

[21] H. T. Ng and R. J. Mooney, "An Efficient First-Order Horn-Clause Abduction System Based on the ATMS," in *Proceedings of the Ninth National Conference on Artificial Intelligence. Anaheim, CA*, 1991, pp. 494–499.

[22] C.-R. Turliuc, N. Maimari, A. Russo, and K. Broda, "On minimality and integrity constraints in probabilistic abduction," 12 2013.

[23] D. Azzolini, E. Bellodi, S. Ferilli, F. Riguzzi, and R. Zese, "Abduction with probabilistic logic programming under the distribution semantics," *International Journal of Approximate Reasoning*, vol. 142, 11 2021.

[24] D. H. Wolpert and D. B. Kinney, "A Stochastic Model of Mathematics and Science," *Foundations of Physics*, vol. 54, no. 2, Apr. 2024. [Online]. Available: http://dx.doi.org/10.1007/s10701-024-00755-9

[25] D. Yekutieli, "Identification of Correlated Damage Parameters Under Noise and Bias Using Bayesian Inference," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 74, no. 3, pp. 515–541, 2012. [Online]. Available: http://www.jstor.org/stable/41674641

[26] D. Pati, A. Bhattacharya, and Y. Yang, "On statistical optimality of variational Bayes," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2018, pp. 1579–1588.

[27] B.-E. Chérief-Abdellatif, P. Alquier, and M. E. Khan, "A Generalization Bound for Online Variational Inference," in *Asian conference on machine learning*. PMLR, 2019, pp. 662–677.

[28] A. Dempster, "A Generalization of Bayesian Inference," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 30, no. 2, pp. 205–247, 1968. [Online]. Available: http://www.jstor.org/stable/2984504

[29] ——, "Upper and Lower Probabilities Induced by a Multivalued Mapping," *The Annals of Mathematical Statistics*, vol. 38, no. 2, pp. 325–339, 1967. [Online]. Available: http://www.jstor.org/stable/2239146

[30] D. Poole, "Probabilistic Horn abduction and Bayesian networks," *Artificial Intelligence*, vol. 64, no. 1, pp. 81–129, 1993. [Online]. Available: https://www.sciencedirect.com/science/article/pii/000437029390061F

[31] D. An, J.-H. Choi, and N. H. Kim, "Identification of Correlated Damage Parameters Under Noise and Bias Using Bayesian Inference," *Structural Health Monitoring*, vol. 11, no. 3, pp. 293–303, 2012. [Online]. Available: https://doi.org/10.1177/1475921711424520

[32] L. Tierney, "Markov Chains for Exploring Posterior Distributions," *The Annals of Statistics*, vol. 22, no. 4, pp. 1701–1728, 1994. [Online]. Available: http://www.jstor.org/stable/2242477

[33] S. Chib, "Calculating Posterior Distributions and Modal Estimates in Markov Mixture Models," *Journal of Econometrics*, vol. 75, no. 1, pp. 79–97, 1996. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0304407695017704

[34] R. A. Levine and G. Casella, "Implementations of the Monte Carlo EM Algorithm," *Journal of Computational and Graphical Statistics*, vol. 10, no. 3, pp. 422–39, 2001. [Online]. Available: http://www.jstor.org/stable/1391097

[35] S. Y. Sekeh, B. Oselio, and A. O. Hero, "Learning to Bound the Multi-Class Bayes Error," *IEEE Transactions on Signal Processing*, vol. 68, pp. 3793–3807, 2020.

[36] ——, "Multi-Class Bayes Error Estimation With a Global Minimal Spanning Tree," in *2018 56th annual allerton conference on communication, control, and computing (allerton)*. IEEE, 2018, pp. 676–681.

[37] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, 1991.

[38] A. A. Wisler, V. Berisha, D. Wei, K. N. Ramamurthy, and A. Spanias, "Empirically-Estimable Multi-Class Classification Bounds," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2594–2598, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:10821499

[39] T. M. Mitchell, "Generalization as search," *Artificial intelligence*, vol. 18, no. 2, pp. 203–226, 1982.

[40] G. D. Montanez, "Why machine learning works," *URL https://www. cs. cmu. edu/˜ gmontane/montanez_dissertation. pdf*, 2017.

[41] P. Myshkov and S. Julier, "Posterior distribution analysis for bayesian inference in neural networks," in *Workshop on Bayesian deep learning, NIPS*, 2016.

[42] T. Charnock, L. Perreault-Levasseur, and F. Lanusse, "Bayesian neural networks," in *Artificial Intelligence for High Energy Physics*. World Scientific, 2022, pp. 663–713.