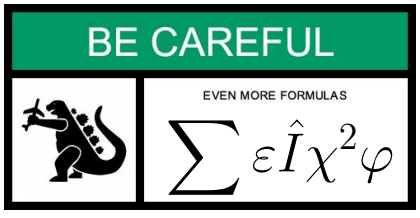
Quantifying and Evaluating Uncertainty in the Internal Representations of Robots

Vladimir Sukhoy and Ryan Kirk
{sukhoy, rakirk}@iastate.edu

Image credit: http://rsss.anu.edu.au/ maier/workshop.html

This document is a final project report for HCI 585X – the course on Developmental Robotics taught by Dr. Alexander Stoytchev in Spring 2011.



I. Research Question

How can a robot guide itself using reduction in the uncertainty of internal representations?

II. Why?

Artificial Intelligence has enjoyed tremendous success over the last twenty five years. Its tools and techniques are now main stream within computer science, and at the core of so many of the systems we use every day. Search algorithms, the backbone of traditional AI, are used throughout operating systems, compilers and networks. More modern machine learning techniques are used to adapt these same systems in real-time. Satisfiability of logic formulas has become a central notion in understanding computability questions and once esoteric notions like semantic anthologies are being used to power the search engines that have become organizers of the world's knowledge, replacing libraries, encyclopedias, and automating business interfaces. And who would have guessed that AI powered robots in people's homes would now be counted in the millions. So much accomplishment to bring pride to us all.

But at the same time Artificial Intelligence has not yet succeeded in its most fundamental ambitions. Our systems are still fragile when outside their carefully circumscribed domains. The best poker playing program can't even understand the notion of a chess move, let alone the conceptual idea of animate versus inanimate. A six year old child can discuss all three domains, but may not be very good at any of them compared to our specialized systems. The challenge for AI, still, is to capture the fundamental nature of generalized perception, intelligence, and action. Worthy challenges for AI that would have tremendous practical impact, are, in my opinion:

- **the generic visual object recognition capabilities of a two year old child**
- **the manual dexterity of a six year old child**
- **the social interaction and language capabilities of a ten year old child**

So much work for all of us to be challenged by.

Rodney Brooks[1]

The field of Artificial Intelligence, much like AIs in John Gibson's science fiction novels, is currently fractured [2]. In the early days, AI and Robotics were about machines with human-like intelligence. The goal of AI was to provide the software, while Robotics sought to provide the hardware.

The situation is now quite different: Computer Vision does not have much to say to Genetic Algorithms, while Decision Theory is quite disjoint from Natural Language Processing. It can be assumed that only the advances in integrated systems and successful applications of first principles can bridge the gaps between many disjoint sub-fields of AI and Robotics. Only these advances can lead to the integrated intelligence with human-like capabilities.

Human development and the human brain are the key areas where the inspiration for a unified approach towards artificial intelligence must come from. If there is agreement on anything in AI, then the fact that humans are intelligent must be it. Therefore, the key integrating principles, if they can be at all isolated, must be present in human development and the human brain.

Human development can really be thought of as series of linear periods of growth interrupted by brief periods of dynamic change. Here, a case will be made for the idea that these dynamic periods of growth can be thought of as phase shifts within the brain. Conceptualizing these periods of growth as phase shifts allows for the inclusion of methods from physics that deal with phase shifts in matter. These methods from Physics center on the conservation of energy, the notion of entropy and the quantification of uncertainty. The field of information theory commonly employs methods to quantify uncertainty within a data set. The goal of this project will be to explore what role, if any, uncertainty has in dynamic periods of growth within humans and within robots. The experiment will focus upon replicating dynamic periods of growth within humans using a robot that learns through by uncertainty driven exploration.

III. Infant Feasibility Test

The goal of this section is to show that the goals of this project are accomplished by two year old human infants.

1) **Self-detection** is a basic capability of animals and human infants. For the purposes of this test, it is sufficient to note that even two month old infants are capable of self-detection in a TV screen [3].
2) **Button-pressing** is a trivial task for a two year old. According to Piaget [4], the tertiary circular relations emerge between 12 and 18 months of age. For instance, if pressing one button on the telephone does not activate it, infants press other buttons until the desired effect is achieved [5, p. 112]. There is also evidence that even 9 months old infants can predict that a bright light will flash or there will be an interesting sound when an experimenter presses a colored button [6].
3) **Phase transitions.** There is significant evidence that phase transitions routinely occur in the human brain [7]. During rhythmic movements, human motor commands exhibits traits that can be explained by a model based on a dynamical

[1]This essay was prepared for the workshop on the Future of AI held at the 25th anniversary of AAAI [1].

systems undergoing a phase transition [7]. Another example is that hypnic jerk – an involuntary movement that often occurs at the onset of sleep – can be explained by a behavior of the dynamical systems near its instability [8]. A popular theory [9] of cognition and action uses an approach based on dynamical systems, which can exhibit phase transitions.

Finally, there is evidence that visual attention can be modelled using the approach based on the change in uncertainty of the internal model [10].

IV. Related Work

A. Categorization emerges from entropy organizing sensory inputs

Inference of concepts comes from the ability to classify perceptions[11]. Here perceptions are distinguished from stimuli since perceptions are stimuli that have already been non-consciously pre-processed. Concept formation requires distinguishing various attributes of a one perception from the attributes of other perceptions[11][12]. What results is the classification of perceptions based upon their perceptual attributes. Classification is an important characteristic of intelligence that is examined, in depth, by many branches of psychology. Behaviorism is interested in the association of one concept with another and the connection that a stimuli can have on the participants behavior/response. Behaviorism is concerned with learning curves and extinction rates which measure the persistence of learning over time [13][14][15]. In both cases, it could be said that a phase transition occurs within the participant when they learn to associate or disassociate.

In cognitive psychology, categorization is often used as dependent variable to measure changes within the internal milieu [11][16][17][18]. Exemplar theory and prototype theory both agree that the human brain creates models of the world that new perceptions are compared against for categorical purposes[16][17]. Piaget mentioned that humans tend to either absorb or assimilate information into fresh categories, or that humans tend to fit or accommodate into existing constructs[15][19].

If this is the case, then distinctive information becomes important since it can be used to simply and efficiently separate prior distributions from conditional distributions. The principle of salience is well established and falls in line with the tendency to use unique information to categorize. Using salient features would reduce the amount of information (as measured by information compression) to categorize new perceptions into existing constructs [18]. This makes sense and can help explain humans tendency to create stereotypical classifications: humans are simply classifying concepts/events based upon the one attribute thought to give the most mutual information about the new perception based upon the individuals previous knowledge of the concept/event. Information gain can also be used to explain the inability of humans to perceive differences that fall below a critical threshold, the Just Noticeable Difference[18][15][20][21]. The familiarity effect relates to humans positive affect bias towards information that is familiar and this may relate to information gain because it takes less effort to process information that is familiar [18][15][20]. A reduction in cognitive effort may be an intrinsic motivator and may act as a natural reward[22]. Signal detection theory relates to the effort needed to detect the difference between two concepts. If two concepts are too similar they will commonly be classified as the same concept or, at the least, will not be distinguished from each other[21]. Cognitive psychology is concerned with the likelihood of committing a Type I or a Type II error and certain social systems (such as the justice system) can be defined by their predisposition towards a particular form of error. The mutual information between two categories would, using information gain theory, determine the tendency to commit Type I and Type II error[23][21][18].

The principle of salience is well established and falls in line with the tendency to use unique information to categorize. Using salient features would reduce the amount of information (as measured by information compression) to categorize new perceptions into existing constructs. This makes sense and can help explain humans tendency to create stereotypical classifications: humans are simply classifying concepts/events based upon the one attribute thought to give the most mutual information about the new perception based upon the individuals previous knowledge of the concept/event. Information gain can also be used to explain the inability of humans to perceive differences that fall below a critical threshold known as the Just Noticeable Difference.

The familiarity effect relates to humans positive affect bias towards information that is familiar and this may relate to information gain because it takes less effort to process information that is familiar. A reduction in cognitive effort may be an intrinsic motivator and may act as a natural reward. Signal detection theory relates to the effort needed to detect the difference between two concepts [21]. If two concepts are too similar they will commonly be classified as the same concept or, at the least, will not be distinguished from each other. Cognitive psychology is concerned with the likelihood of committing a Type I or a Type II error and certain social systems (such as the justice system) can be defined by their predisposition towards a particular form of error. The mutual information between two categories would, using information gain theory, determine the tendency to commit Type I and Type II error.

B. Body-schema emerges from entropy organizing categories

Self-detection and a sense of self are vital to the development of a self-identity. The ability to detect ones self has been shown to arise in humans as young as four months, but this sense of self is not immediately complete [24]. Self-detection and self-identity are not distinct to humans, but are concepts exhibited in other animals considered to be intelligent, such as great apes, dolphins and elephants. Understanding the self will reduce the amount of information required to process ones

environment. Tool use requires some self-concept because the tool user will need to understand what their body is capable of doing in order to know how to use a tool. The ability to transfer skills learned during the performance of one task to the performance of another task requires understanding that the same person, the self, was the actor committing both activities. Thus, generalizing tool use from one situation to another also requires a sense of self. Finally, the development of a sense of self-identity is important for the development of the ability to perceive others as individuals. Perceiving others as individuals is necessary for the development of society since social interactions often require understanding the needs of others.

C. An uncertainty-based approach to the development of A.I.

The minimum information required to classify an event would be that information that distinguishes it from an existing distribution (read: existing concept, category or idea). Thus, stimuli would need to be processed serially through some innate discretization, dimensional-reduction, pre-processing, invariant pattern/algorithm before learning could occur. For simplicity, this function will be referred to as the neuronal function [25]. Processing stimuli in such a way allows for the comparison of information from different modalities to be tied into a contiguous conception of reality. As intelligence arises, it develops through particular stages. Not all capacities will be immediately present. These capacities are probably absent at first and arise as a result of discontinuous development within an intelligent system. Here, two distinct periods of dynamic development will be examined: the development of the sense of self and that development of concept inference. The development of both of these distinctive qualities of intelligence can be tied to differences in memory, learning and behavioral outcomes.

In both categorization and self-detection, the principle of information gain could be applied to explain the development of these processes. An information gain approach would utilize the minimal information necessary to separate a signal from a prior distribution. A characteristic of reduction of entropy within a system is dynamic change. These changes can be measured physically through changes in state or through changes in behavior and/or interaction patterns or through measuring differences in memory and learning outcomes. These dynamic changes can also sometimes be quantified through measuring artifacts and differences in temporal signals. Examples of temporal signals include wave patterns within a physical system, input patterns into a computational system and electroencephalograph (EEG) patterns within human systems. Here, a method will be used that strives to model human developmental processes in a synthetic, robotic system. This method, discussed in depth in the approach section, will utilize a traditional approach to information processing that strives to compress, classify and ultimately categorize information. Such an approach has been used both in robotics and in analyzing EEG data [26].

The human infant is a ball of potential that, at birth, comes pre-wired with the minimal biological structures and the minimal information necessary in the brain's 'innate invariant neuronal patterns' for the developmental life cycle for the normal human to occur. Of course, a host of mitigating social and developmental factors can vastly change the realized/actualized potential of an individual; however, what is clear, is that we all start in very similar forms. At birth, an infant is a bundle of reflexes that lead it to certain 'survival instincts', imprinting potentials and, importantly, exploratory behaviors. This exploration is necessary for the neuronal function to begin to lead the human down the path of development.

D. Robots learning button-pressing through exploratory behavior

The analysis of the related work on pressing buttons in Robotics produced several categories for the proposed approaches. These categories highlight different aspects of the manipulation problem and the proposed methodology for solving it.

1) Detecting buttons is hard, pressing them is straightforward. Work in this category is focused on a single aspect of the problem: detecting buttons using vision. Once a button is detected, it is assumed to be easily pressable. Due to the narrow focus of this line of work, the feedback that a button might generate is often completely ignored. The evaluation is often performed for elevator buttons [27] [28] [29]. The fact that buttons in a typical elevator are arranged in a grid pattern and have numeric labels is often used to boost the performance of the learning algorithm [27] [28].

2) Both detecting and pressing buttons is very hard. Another category of the related work assumes that detecting and manipulating buttons, switches, levers, knobs, and similar widgets designed for humans is intractable for robots. To help robots solve these problems, different types of environmental augmentations are proposed. For example, reflective markers [30] or RFID tags [31] [32] can be attached to the widgets. A tag may inform the robot where the widget is, how to activate it and what happens when it is activated. The main focus of this line of work is on robotic applications enabled by different types of environmental augmentations [32].

3) Understanding social context is crucial for both pressing and detecting buttons. The third category of the related work focuses on social aspects of manipulation. These approaches seek to interpret human-provided social cues associated with robotic actions. For instance, a robot can learn how to detect buttons when humans point at it [33] [34] and learn to press it from human demonstrations [35].

4) Pressing and detecting buttons must be learned together. The last category of the related work differs from other categories in two ways: 1) both pressing and detecting tasks are regarded as challenging, but solvable; 2) the visual model for detecting buttons is trained from multimodal events produced by pressing them. Previously, it was shown that a robot can bootstrap the visual model for detecting buttons by exploring them with pushing behaviors [36]. Our work builds on these ideas and shows how both skills can be developed together in real time.

1) GPU Programming: GPU programming [37] was used in this work to achieve real time performance for the visual pipeline. Image convolutions are extensively used in the pipeline and implementing them on GPU [38] was a key enabling factor for the speedup. Previously, SIFT [39] and SURF [40] visual pipelines were implemented using GPU to improve their performance.

2) Developmental Psychology: E.J. Gibson [41]. Experience obtained while exploring objects stimulates further interest [42]. 7-11 months old infants are not interested in seeing object or people who manipulate objects until the infants have had the chance to play with them. Infants as young as 9 m.o. can predict interesting events when an experimenter presses a colored button [6]. Infants perform repetitive movements when they learn to manipulate [4].

V. Experimental Setup and Methods

A. Approach

This paper outlines mathematical formulations describing the nature of the entropy-based learning algorithm of interest in this experiment. These formulations will be used to guide learning across several tasks characteristic of human development in order to determine whether a system of learning based upon entropy will result in the natural creation of categories and/or decisions. In particular, three scenarios will be examined:

1) **Self-detection through engaging objects with exploratory shaking behaviors.**
2) **Self-detection through engaging objects with exploratory pushing behavior.**
3) **Self-detection through game play.**

These three separate experimental situations are meant to test the devised algorithms' efficacy in both the creation of categories and in the establishment of self-detection.

B. Research question

The of this experiment is to effectively test the hypothesis that entropy-based learning, defined in scope through the mathematical formulations, can account for learning across several developmental periods as such as categorization and the development of the self-identity as measured through self detection.

C. Self-detection Problem

The self-detection problem:

Given a feature, the robot needs to decide if it is self or not self.

Information theory is used to solve this problem. The key idea is to determine whether the stream of information from the feature is influenced by the stream of information that controls movement of the robot. If this is true, the feature is self. If this is not true then the feature is not self.

More specifically, the relationship between movement of the feature, quantified by a variable M, and a the temporal delay since the last motor command, quantified by a variable D, is used. The mutual information $I(M;D) \geq 0$ quantifies the amount of information shared between the two variables. The criterion for self-detection:

$$\text{Feature} = \begin{cases} \text{Self} & \text{if } I(M;D) > 0, \\ \text{Not Self} & \text{if } I(M;D) = 0. \end{cases}$$

D. Mutual Information Estimation

The mutual information $(M;D)$ can be written using the Shannon entropy [43] function H as follows:

$$I(M;D) = H(M) + H(D) - H(M,D), \tag{1}$$

where the entropy $H(X)$ quantifies the uncertainty in X:

$$H(X) = H(p(x_1), \ldots, p(x_m)) = -\sum_{i=1}^{m} p_i \log p_i, \tag{2}$$

where $p_i = p(x_i)$.

The robot does not know true distributions for M, D, or the true joint distribution. It can only estimate these distributions from the data. For instance, the temporal delay can be discretized using a histogram, while movement can be reduced to a binary variable.

E. Bias and Variance of Statistical Estimators

Suppose that X is a random variable. Also suppose that θ is some quantity that is a function of the probability distribution of X (for instance, θ can be $H(X)$ – i.e., the entropy of X). Finally, suppose that the probability distribution of X is not known, but a number of individual observations from X are available. In this case it is only possible to compute an estimate for θ using this sample. The result of this computation is an estimator $\hat{\theta}$.

How good is $\hat{\theta}$? How far is $\hat{\theta}$ from θ? The answer to this question is the mean-squared error $\mathrm{MSE}(\hat{\theta})$, which is defined as the expected value for the squared difference between $\hat{\theta}$ and its true value θ:

$$\mathrm{MSE}(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right].$$

It turns out that $\mathrm{MSE}(\hat{\theta})$ can be expressed in terms of the systematic error that $\hat{\theta}$ makes against the true value θ (this systematic error is quantified by $\mathrm{Bias}(\hat{\theta})$) and the random spread of $\hat{\theta}$ itself (this spread is quantified by the variance $\mathrm{Var}(\hat{\theta})$).

More formally,

$$\begin{aligned} \mathrm{Bias}(\hat{\theta}) &= E[\hat{\theta} - \theta], \\ \mathrm{Var}(\hat{\theta}) &= E\left[(\hat{\theta} - E[\theta])^2\right]. \end{aligned} \tag{3}$$

$\mathrm{Bias}(\theta)$ is a linear operator, which is shown by the following couple of properties.

Property 1. *If* $\hat{\gamma} = \hat{\alpha} + \hat{\beta}$*, then* $\mathrm{Bias}(\hat{\gamma}) = \mathrm{Bias}(\hat{\alpha}) + \mathrm{Bias}(\hat{\beta})$.

Property 2. $\mathrm{Bias}(-\hat{\alpha}) = -\mathrm{Bias}(\hat{\alpha})$.

The relationship between bias, variance, and MSE is formalized by the following property.

Property 3. $\mathrm{MSE}(\hat{\theta}) = (\mathrm{Bias}(\hat{\theta}))^2 + \mathrm{Var}(\hat{\theta})$

It turns out that it is impossible to estimate entropy with zero bias [44, p. 1236]. It is a common mistake in the literature to provide error bars for entropy that only show the sample variance for the values generated by the entropy estimator $\hat{\theta}$. This sample variance only accounts for $\mathrm{Var}(\hat{\theta})$ and ignores $\mathrm{Bias}(\hat{\theta})$. This mistake may result in misleadingly small error bars, which do not reflect the actual quality of the estimates.

F. Histogram Properties

Let $X \in \{x_1, x_2, \ldots, x_m\}$ be a discrete random variable. Also let $p_i = p(x_i)$ be the shorthand notation for $\Pr(X = x_i)$, i.e., the probability of $X = x_i$. Suppose that the true probability distances of X – i.e, the vector $p = (p_1, p_2, \ldots, p_m)$ – is unknown.

Let $S = \{s_1, s_2, \ldots, s_N\}$ be a sample that is drawn i.i.d. from X. In other words, the following three assumptions hold: 1) each s_k is drawn from a random variable X_k, 2) the random variables $\{X_1, X_2, \ldots, X_N\}$ are independent, and 3) the set of values and the probability distribution for each X_k matches the set of values and the probability distribution of X (i.e., $X_k = X$ for $k = 1, \ldots, N$).

How to estimate the probability distribution $p = (p_1, p_2, \ldots, p_m)$ from the sample $S = \{s_1, s_2, \ldots, s_N\}$? It is possible to construct a histogram of S by collecting its N elements into m bins that correspond to the m possible values of X, i.e., $x_1, x_2, \ldots, x_m$. In other words, if $c = (c_1, c_2, \ldots, c_m)$ is a vector of bin counters for the histogram, then each c_i is equal to the number of elements in S that are equal to x_i. Note that $c_1 + c_2 + \ldots + c_m = N$. Also note that

$$\lim_{N \to \infty} \frac{c_i}{N} = p_i,$$

which follows from the law of large numbers.

Note that the vector of bin counters c is a multinomial random variable. The probability of observing a particular value of c is equal to the multinomial probability mass function parametrized by $c = (c_1, c_2, \ldots, c_m)$ and the unknown probability distribution $p = (p_1, p_2, \ldots, p_m)$:

$$\Pr(c) = \frac{N!}{c_1! c_2! \cdots c_m!} p_1^{c_1} p_2^{c_2} \cdots p_m^{c_m}. \tag{4}$$

The following property gives the characteristic function for the multinomial distribution. The characteristic function is a very powerful tool. It allows to compute various moments (e.g., expected value, variance, and covariance) of the elements of c. In addition, it is used to prove Pearson's Theorem (Theorem 6) and Fisher's Theorem (Theorem 7), which are essential for statistical reasoning about entropy and mutual information.

Property 4. *The joint characteristic function* $\varphi_c(t)$ *for the multinomial random variable* c *is*

$$\varphi_c(t) = \left(p_1 e^{\imath t_1} + p_2 e^{\imath t_2} + \ldots + p_m e^{\imath t_m}\right)^N, \tag{5}$$

where $\imath$ denotes the imaginary unit (i.e., $\imath^2 = -1$) and $t = (t_1, t_2, \ldots, t_m)$.

Note that

$$\begin{aligned}
\varphi'_i(t) &= \frac{\partial \varphi_c(t)}{\partial t_i} = N \left(\sum_{k=1}^{m} p_k e^{\imath t_k} \right)^{N-1} p_i \imath e^{\imath t_i}, \\
\varphi''_{ii}(t) &= \frac{\partial^2 \varphi_c(t)}{\partial t_i^2} = N(N-1) \left(\sum_{k=1}^{m} p_k e^{\imath t_k} \right)^{N-2} (p_i \imath e^{\imath t_i})^2 + N \left(\sum_{k=1}^{m} p_k e^{\imath t_k} \right)^{N-1} p_i \imath^2 e^{\imath t_i} \\
&= - \left(N(N-1) \left(\sum_{k=1}^{m} p_k e^{\imath t_k} \right)^{N-2} p_i^2 e^{2\imath t_i} + N \left(\sum_{k=1}^{m} p_k e^{\imath t_k} \right)^{N-1} p_i e^{\imath t_i} \right), \\
\varphi''_{ij}(t) &= \frac{\partial^2 \varphi_c(t)}{\partial t_i \partial t_j} = N(N-1) \left(\sum_{k=1}^{m} p_k e^{\imath t_k} \right)^{N-2} p_i (\imath e^{\imath t_i}) p_j (\imath e^{\imath t_j}) \\
&= -N(N-1) \left(\sum_{k=1}^{m} p_k e^{\imath t_k} \right)^{N-2} p_i p_j e^{\imath t_i} e^{\imath t_j}.
\end{aligned}$$

To illustrate the power of φ_c, consider computing the expected value, the variance, and the covariance of the elements of c:

$$E[c_i] = -\imath \varphi'_i(0) = -\imath N (\underbrace{p_1 + p_2 + \ldots + p_m}_{1})^{N-1} p_i \imath \cancel{e^{\imath 0}} = -\imath \imath N p_i = N p_i, \tag{6}$$

$$\begin{aligned}
E\left[c_i^2\right] &= \imath^2 \varphi''_{ii}(0) = -(-\imath)^2 \left(N(N-1) p_i^2 + N p_i \right) = N(N-1) p_i^2 + N p_i, \\
\mathrm{Var}(c_i) &= E\left[c_i^2\right] - \left(E[c_i]\right)^2 = N(N-1) p_i^2 + N p_i - (N p_i)^2 = \cancel{N^2 p_i^2} - N p_i^2 + N p_i - \cancel{N^2 p_i^2} = N p_i (1 - p_i), \qquad (7) \\
E[c_i c_j] &= \imath^2 \varphi''_{ij}(0) = N(N-1) p_i p_j, \\
\mathrm{Cov}(c_i, c_j) &= E[c_i c_j] - E[c_i] E[c_j] = N(N-1) p_i p_j - N^2 p_i p_j = -N p_i p_j. \qquad (8)
\end{aligned}$$

Also note that

$$E\left[\frac{c_i}{N}\right] = \frac{E[c_i]}{N} = \frac{N p_i}{N} = p_i. \tag{9}$$

G. Maximum Likelihood Estimators

Let θ be an unknown parameter that is determined by the unknown probability distribution $p = (p_1, p_2, \ldots, p_m)$. How to estimate θ using the sample $S = (s_1, s_2, \ldots, s_N)$? One way is to use the maximum likelihood (MLE) estimator $\hat{\theta}_{\text{MLE}}$, which is maximizes the likelihood function $L(\tilde{\theta} \mid S)$. More formally,

$$\hat{\theta}_{\text{MLE}} = \operatorname*{argmax}_{\tilde{\theta}} L(\tilde{\theta} \mid S). \tag{10}$$

The likelihood function $L(\tilde{\theta} \mid S)$ is defined as the conditional probability of observing the sample S given that the true value of the parameter θ is equal to $\tilde{\theta}$. More formally,

$$L(\tilde{\theta} \mid S) = \Pr(S \mid \theta = \tilde{\theta}). \tag{11}$$

If the task is to estimate the multinomial probability distribution (i.e., $\theta = p$), then the likelihood function is simply the multinomial probability mass function (4), where the candidate $\tilde{p} = (\tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_m)$ replaces the true probability distribution $p = (p_1, p_2, \ldots, p_m)$:

$$\begin{aligned}
L(\tilde{p} \mid S) &= \Pr(S \mid p = \tilde{p}) = \Pr(c \mid p = \tilde{p}) \\
&= \frac{N!}{c_1! c_2! \cdots c_m!} (\tilde{p}_1)^{c_1} (\tilde{p}_2)^{c_2} \cdots (\tilde{p}_m)^{c_m}.
\end{aligned} \tag{12}$$

The histogram-based estimator $\hat{p} = {}^{c}/_{N}$ is in fact the maximum likelihood estimator for the unknown probability distribution p. The following proposition states this formally.

Proposition 1. $\hat{p}_{\text{MLE}} = {}^{c}/_{N}$.

Maximum likelihood estimators are functionally invariant, which makes it easy to construct these estimators. In other words, the MLE of a function can be obtained by applying the function to the MLE of its arguments. For instance, the MLE $\hat{H}_{\text{MLE}}$ for the entropy $H(p) = -(p_1 \log p_1 + p_2 \log p_2 + \ldots p_m \log p_m)$ can be obtained by plugging $\hat{p}_{\text{MLE}}$ into the entropy function, i.e., $\hat{H}_{\text{MLE}} = H(\hat{p}_{\text{MLE}})$. The following theorem states the functional invariance property of the MLE more formally.

Theorem 1. ***Functional Invariance for MLE.***
Let θ be a parameter defined on a set Θ. Also let $g : \Theta \to \mathrm{T}$ be a function. If $\hat{\theta}_{MLE}$ is the MLE for θ, then $\hat{\tau}_{MLE} = g(\hat{\theta}_{MLE})$ is the MLE for the parameter $\tau = g(\theta)$.

H. Entropy

In his famous paper [43], Shannon proposed to quantify the uncertainty of a probability distribution $p = (p_1, p_2, \ldots, p_k)$, where $0 \leq p_i \leq 1$ for $i = 1, \ldots, k$ and $p_1 + p_2 + \ldots + p_k = 1$ using entropy $H(p)$, which is defined as follows:

$$H(p_1, \ldots, p_k) = -\sum_{i=1}^{k} p_i \log p_i. \tag{13}$$

Note that if $p_i = 0$, then it is assumed that $p_i \log p_i = 0$. This rule is in agreement with the following limit:

$$\lim_{x \to 0} x \log x = 0.$$

Similarly, it is possible to define entropy for a discrete random variable $X \in \{x_1, x_2, \ldots, x_m\}$ as shown below:

$$H(X) = H(p(x_1), p(x_2), \ldots, p(x_m)) = -\sum_{i=1}^{m} p(x_i) \log p(x_i),$$

where $p(x_i) = \Pr(X = x_i)$. Without loss of generality, it is assumed that $p(x_i) > 0$. If there is a degenerate x_i which can never occur, then it is possible to exclude it from the set of values that X can take. In other words, define another random variable $Y = \{x_1, \ldots, x_m\} / \{x_i : p(x_i) = 0\}$ and consider Y instead of X.

Entropy also quantifies the amount of information in a random variable X. If the logarithm base is 2, the entropy is measured in bits. If the base is e, the entropy is measured in nats. Logarithm base 10 results in measuring entropy in digits. In this report the entropy is measured in bits.

I. Entropy Estimators

The true value for Shannon entropy $H(X)$ cannot be computed when the true probability distribution of X is unknown. In other words, computing $H(X)$ requires knowledge of the exact value for each $p(x_1), p(x_2), \ldots, p(x_m)$.

When these values are not available, but a sample $S = \{s_1, s_2, \ldots, s_N\}$, which was drawn from X is available, then it is possible to construct a histogram of S and estimate $H(X)$ using this histogram.

Theorem 1 (i.e., functional invariance of the maximum likelihood estimates) implies that if $\hat{p}_{\mathrm{MLE}} = {}^{c}/_{N}$ is plugged into the Shannon's formula (13), then the resulting function is the entropy maximum likelihood estimator $\hat{H}_{\mathrm{MLE}}$. In other words,

$$\hat{H}_{\mathrm{MLE}} = -\sum_{i=1}^{m} \frac{c_i}{N} \log_2 \frac{c_i}{N}. \tag{14}$$

Despite its intuitive appeal, $\hat{H}_{\mathrm{MLE}}$ is not the best estimator for the true value of entropy $H(X)$. It is well known that $\hat{H}_{\mathrm{MLE}}$ can have a significant negative bias, unless $N \gg m$ [44]. A better estimator was proposed by Miller [45]. The improvement is achieved by adding a correction term to $\hat{H}_{\mathrm{MLE}}$ in order to improve its quality. The resulting estimator $\hat{H}_{\mathrm{MM}}$ is often called the entropy MLE with Miller-Madow bias correction [44] [45]:

$$\hat{H}_{\mathrm{MM}} = \hat{H}_{\mathrm{MLE}} + \frac{m-1}{2N(\ln 2)} = -\sum_{i=1}^{m} \frac{c_i}{N} \log_2 \frac{c_i}{N} + \frac{m-1}{2N(\ln 2)}.$$

Miller's correction increases the estimate to compensate for the negative bias of $\hat{H}_{\mathrm{MLE}}$. Note that the the correction term is a function of the number of bins m and the number of samples N.

In this project the Miller-Madow entropy estimator was used because it is better than the MLE and because it preserves the similarity to Shannon's formula, which enables more compact derivations. For large samples with $N \gg m$, $\hat{H}_{\mathrm{MM}}$ is a very good estimator [44]. More advanced estimators, which can perform even better than $\hat{H}_{\mathrm{MM}}$ when ${}^{N}/_{m}$ is low, are described in [44].

J. Rigorous Bound for the Bias of Entropy Estimators

Consider two vectors: $p = (p_1, p_2, \ldots, p_m)$ and $\hat{p} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_m)$ such that

$$\sum_{i=1}^{m} p_i = \sum_{i=1}^{m} \tilde{p}_i = 1, \qquad 0 \leq p_i, \tilde{p}_i \leq 1, \quad (i = 1, \ldots, m).$$

Suppose that if $p_i = 0$, then $\tilde{p}_i = 0$. Note that this is true if $\tilde{p} = \hat{p}_{\text{MLE}}$, where $\hat{p}_{\text{MLE}}$ is computed from a sample histogram.

In this section the vector p is the unknown probability distribution and the vector $\tilde{p}$ is its estimate. The following results show that the difference $H(\tilde{p}) - H(p)$ can be expressed in terms that can be bounded. The overall argument follows [44].

An algebraic trick that consists of adding and subtracting a specially constructed term allows us to expres $H(\tilde{p}) - H(p)$ as follows:

$$H(\hat{p}) - H(p) = H(\hat{p}) - H(p) + T - T = R - T. \tag{15}$$

The remainder term R is defined as follows:

$$R{=}H(\hat{p}){-}H(p){+}T{=}H(\hat{p}){-}H(p){+}\sum_{i=1}^{m}(\hat{p}_i{-}p_i)\log_2 p_i. \tag{16}$$

The trick can only be applied if the term T is finite. The following proposition, which is straightforward to prove, is a formal statement of this assumption.

Proposition 2. $T = \sum_{i=1}^{m}(\hat{p}_i - p_i)\log_2 p_i$ *is finite.*

Recall from Section V-H that

$$H(\hat{p}) = H\left(\frac{c_1}{N}, \frac{c_2}{N}, \ldots, \frac{c_m}{N}\right) = \hat{H}_{\text{MLE}}, \tag{17}$$

This allows us to express the difference $\hat{H}_{\text{MLE}} - H(p)$ as follows

$$H(\hat{p}) - H(p) = \hat{H}_{\text{MLE}} - H(p) = R - \sum_{i=1}^{m}\left(\frac{c_i}{N} - p_i\right)\log_2 p_i. \tag{18}$$

Note that the expected value $E\left[H(\hat{p}) - H(p)\right]$ is the $\text{Bias}(\hat{H}_{\text{MLE}})$. The following proposition, which is straightforward to prove using algebra, shows that only the expected value of the remainder term R contributes to the $\text{Bias}(\hat{H}_{\text{MLE}})$.

Proposition 3. $\text{Bias}(\hat{H}_{MLE}) = E[R]$.

Note that the remainder term R is equal to the negative value of the Kullback-Leibler divergence $-D_{\text{KL}}(\hat{p}\|p)$. This follows from a sequence of straightforward algebraic operations, which is shown below:

$$\begin{aligned} R &= H(\hat{p}) - H(p) + \sum_{i=1}^{m}(\hat{p}_i - p_i)\log_2 p_i = -\sum_{i=1}^{m}\hat{p}_i \log_2 \hat{p}_i + \cancel{\sum_{i=1}^{m} p_i \log_2 p_i} + \sum_{i=1}^{m}\hat{p}_i \log_2 p_i - \cancel{\sum_{i=1}^{m} p_i \log_2 p_i} \\ &= -\sum_{i=1}^{m}\hat{p}_i(\log_2 \hat{p}_i - \log_2 p_i) = -\sum_{i=1}^{m}\hat{p}_i \log_2 \frac{\hat{p}_i}{p_i} = -D_{\text{KL}}(\hat{p}\|p). \end{aligned} \tag{19}$$

Now it is possible to use the results from statistics that give bounds for D_{KL} in order to find the bounds for $\text{Bias}(\hat{H}_{\text{MLE}})$. The following theorem, the proof for which can be found in [46], uses Pearson's chi-square statistic $\mathcal{X}^2$, which is related, but not the same as the χ^2 distribution, to give the bounds for D_{KL}.

Theorem 2. *The Kullback-Leibler divergence D_{KL} satisfies*

$$0 \leq D_{KL}(\hat{p}\|p) \leq \log_2(1 + \mathcal{X}^2(\hat{p}, p))$$

where $\mathcal{X}^2(\hat{p}, p)$ is Pearson's chi-square statistic, defined as

$$\mathcal{X}^2(\hat{p}, p) = \sum_{i=1}^{m}\frac{(\hat{p}_i - p_i)^2}{p_i}.$$

It turns out that when $\hat{p} = \hat{p}_{\text{MLE}}$, the expected value of the $\mathcal{X}^2$ statistic is a function of the number of histogram bins m and sample size N.

Theorem 3. *If $\hat{p}_i = \frac{c_i}{N}$, then $E\left[\mathcal{X}^2(\hat{p}, p)\right] = \frac{m-1}{N}$.*

It is now possible to use these intermediate results to derive upper and lower bounds for $\mathrm{Bias}(\hat{H}_{\mathrm{MLE}})$.

Proposition 4. $L_{\hat{H}_{MLE}} \leq \mathrm{Bias}(\hat{H}_{MLE}) \leq U_{\hat{H}_{MLE}}$, *where* $L_{\hat{H}_{MLE}} = -\log_2\left(1+\frac{m-1}{N}\right)$ *and* $U_{\hat{H}_{MLE}} = 0$.

It is straightforward to apply the Miller-Madow correction to the bounds for $\mathrm{Bias}(\hat{H}_{\mathrm{MLE}})$ and obtain the bounds for $\mathrm{Bias}(\hat{H}_{\mathrm{MM}})$.

Proposition 5. $L_{\hat{H}_{MM}} \leq \mathrm{Bias}(\hat{H}_{MM}) \leq U_{\hat{H}_{MM}}$, *where*

$$L_{\hat{H}_{MM}} = L_{\hat{H}_{MLE}} + \frac{m-1}{2N(\ln 2)} = -\log_2\left(1+\frac{m-1}{N}\right) + \frac{m-1}{2N(\ln 2)},$$
$$U_{\hat{H}_{MM}} = U_{\hat{H}_{MLE}} + \frac{m-1}{2N(\ln 2)} = \frac{m-1}{2N(\ln 2)}.$$

K. Mutual Information

Let $X \in \{x_1, x_2, \ldots, x_m\}$ and $Y \in \{y_1, y_2, \ldots y_n\}$ be two random variables. The mutual information $I(X;Y)$, which quantifies the dependencies between X and Y, is defined as follows:

$$I(X;Y) = \sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j)\log\left(\frac{p(x_i, y_j)}{p(x_i)p(y_j)}\right), \tag{20}$$

where $p(x_i, y_j) = \Pr(X = x_i \cap Y = y_j)$, $p(x_i) = \Pr(X = x_i)$, and $p(y_j) = \Pr(Y = y_j)$, which implies the following constraints:

$$0 \leq p(x_i) \leq 1, \quad 0 \leq p(y_j) \leq 1, \quad 0 \leq p(x_i, y_j) \leq 1; \tag{21}$$
$$p(x_i) = \sum_{j=1}^{n} p(x_i, y_j), \quad p(y_j) = \sum_{i=1}^{m} p(x_i, y_j); \tag{22}$$
$$\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j) = \sum_{i=1}^{m} p(x_i) = \sum_{j=1}^{n} p(y_j) = 1. \tag{23}$$

If $p(x_i, y_j) = 0$, then it is assumed that the corresponding entry in the definition of $I(X;Y)$ is zero, i.e.

$$p(x_i, y_j)\log\left(\frac{p(x_i, y_j)}{p(x_i)p(y_j)}\right) = 0, \quad \text{if } p(x_i, y_j) = 0. \tag{24}$$

The following property restricts the range of values that $I(X;Y)$ might attain:

Property 5. $0 \leq I(X;Y) \leq \log(\min(m, n))$.

The following property establishes the relationship between the mutual information of X and Y and their independence.

Theorem 4. $I(X;Y) = 0$ *if and only if* X *and* Y *are independent – i.e.,* $p(x_i, y_j) = p(x_i)p(x_j)$ *for* $i = 1, \ldots, m$ *and* $j = 1, \ldots, n$.

L. Mutual Information Estimation

Note that the mutual information can be expressed in terms of entropy as shown below:

$$I(X;Y) = H(X) + H(Y) - H(X,Y).$$

This fact allows us to estimate $I(X;Y)$ using entropy estimators:

$$\hat{I}(X;Y) = \hat{H}(X) + \hat{H}(Y) - \hat{H}(X,Y).$$

In particular, it is possible to use $\hat{H}_{\mathrm{MLE}}$ to obtain the corresponding $\hat{I}_{\mathrm{MLE}}$ and $\hat{H}_{\mathrm{MM}}$ to obtain the corresponding $\hat{I}_{\mathrm{MM}}$:

$$\hat{I}_{\mathrm{MLE}}(X;Y) = \hat{H}_{\mathrm{MLE}}(X) + \hat{H}_{\mathrm{MLE}}(Y) - \hat{H}_{\mathrm{MLE}}(X,Y), \tag{25}$$
$$\hat{I}_{\mathrm{MM}}(X;Y) = \hat{H}_{\mathrm{MM}}(X) + \hat{H}_{\mathrm{MM}}(Y) - \hat{H}_{\mathrm{MM}}(X,Y). \tag{26}$$

For the joint distribution (X, Y), it possible to use the histogram to estimate it. The following formulas clarify the notation:

$$c_i^X = \sum_{j=1}^{n} c_{ij}^{XY}, \quad c_j^Y = \sum_{i=1}^{m} c_{ij}^{XY}, \quad \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}^{XY} = N. \tag{27}$$

Note that $c_1^X + c_2^X + \ldots + c_m^X = c_1^Y + c_2^Y + \ldots + c_n^Y = N$.

Note that for $\hat{H}_{\text{MM}}(X,Y)$ it is assumed that all possible pairwise combinations (x_i, y_j) are possible. Thus, $|(X,Y)| = |X| \cdot |Y| = mn$. It follows that

$$\hat{H}_{\text{MM}}(X,Y) = \hat{H}_{\text{MLE}}(X,Y) + \frac{mn-1}{2N(\ln 2)}. \tag{28}$$

Therefore, $\hat{I}_{\text{MM}}(X,Y)$ can be computed as follows:

$$\begin{aligned}
\hat{I}_{\text{MM}}(X;Y) &= \hat{H}_{\text{MM}}(X) + \hat{H}_{\text{MM}}(Y) - \hat{H}_{\text{MM}}(X,Y) \\
&= \hat{H}_{\text{MLE}}(X) + \frac{m-1}{2N(\ln 2)} + \hat{H}_{\text{MLE}}(Y) + \frac{n-1}{2N(\ln 2)} - \left(\hat{H}_{\text{MLE}}(X,Y) + \frac{mn-1}{2N(\ln 2)} \right) \\
&= \hat{H}_{\text{MLE}}(X) + \hat{H}_{\text{MLE}}(Y) - \hat{H}_{\text{MLE}}(X,Y) + \frac{m-1}{2N(\ln 2)} + \frac{n-1}{2N(\ln 2)} - \frac{mn-1}{2N(\ln 2)} \\
&= \hat{I}_{\text{MLE}}(X;Y) + \frac{m-1+n-1-(mn-1)}{2N(\ln 2)} = \hat{I}_{\text{MLE}}(X;Y) - \frac{(m-1)(n-1)}{2N(\ln 2)}.
\end{aligned} \tag{29}$$

M. Rigorous Bounds for the Bias of Mutual Information Estimators

By the linearity of bias,

$$\text{Bias}(\hat{I}_{\text{MLE}}(X;Y)) = \text{Bias}(\hat{H}_{\text{MLE}}(X)) + \text{Bias}(\hat{H}_{\text{MLE}}(Y)) - \text{Bias}(\hat{H}_{\text{MLE}}(X,Y)), \tag{30}$$

which allows us to re-use the bounds for $\text{Bias}(\hat{H}_{\text{MLE}})$ and $\text{Bias}(\hat{H}_{\text{MM}})$ to obtain the bounds for $\text{Bias}(\hat{I}_{\text{MLE}})$ and $\text{Bias}(\hat{I}_{\text{MM}})$.

Proposition 6. $L_{\hat{I}_{MLE}} \leq \text{Bias}(\hat{I}_{MLE}) \leq U_{\hat{I}_{MLE}}$, *for*

$$\begin{aligned}
L_{\hat{I}_{MLE}} &= L_{\hat{H}_{MLE}(X)} + L_{\hat{H}_{MLE}(Y)} - U_{\hat{H}_{MLE}(X,Y)}, \\
U_{\hat{I}_{MLE}} &= U_{\hat{H}_{MLE}(X)} + U_{\hat{H}_{MLE}(Y)} - L_{\hat{H}_{MLE}(X,Y)},
\end{aligned}$$

where $L_{\hat{H}_{MLE}}$ *and* $U_{\hat{H}_{MLE}}$ *for* X, Y, *and the joint distribution* (X,Y) *are defined according to Proposition 4.*

Proposition 7. $L_{\hat{I}_{MM}} \leq \text{Bias}(\hat{I}_{MM}) \leq U_{\hat{I}_{MM}}$, *where*

$$\begin{aligned}
L_{\hat{I}_{MM}} &= L_{\hat{I}_{MLE}} - \frac{(m-1)(n-1)}{2N(\ln 2)}, \\
U_{\hat{I}_{MM}} &= U_{\hat{I}_{MLE}} - \frac{(m-1)(n-1)}{2N(\ln 2)}.
\end{aligned}$$

N. Properties of Mutual Information Estimators

The following proposition shows that $\hat{I}_{\text{MLE}}(X;Y)$ is nonnegative. This result is used in the next section to derive the confidence bound for rejecting the null hypothesis that $I(X;Y) = 0$. The proof of this basic fact is provided here for the sake of completeness.

Proposition 8. $\hat{I}_{MLE}(X;Y) \geq 0$.

Recall from (15), (17), and Proposition 2 that $\hat{H}_{\text{MLE}}(X)$, $\hat{H}_{\text{MLE}}(Y)$, and $\hat{H}_{\text{MLE}}(X,Y)$ can be expressed as follows:

$$\begin{aligned}
\hat{H}_{\text{MLE}}(X) &= H(X) + R_{\hat{H}_{\text{MLE}}(X)} - T_{\hat{H}_{\text{MLE}}(X)}, \\
\hat{H}_{\text{MLE}}(Y) &= H(Y) + R_{\hat{H}_{\text{MLE}}(Y)} - T_{\hat{H}_{\text{MLE}}(Y)}, \\
\hat{H}_{\text{MLE}}(X,Y) &= H(X,Y) + R_{\hat{H}_{\text{MLE}}(X,Y)} - T_{\hat{H}_{\text{MLE}}(X,Y)},
\end{aligned}$$

where

$$\begin{aligned}
R_{\hat{H}_{\text{MLE}}(X)} &= -D_{\text{KL}}(\hat{p}^X \| p^X), \\
R_{\hat{H}_{\text{MLE}}(Y)} &= -D_{\text{KL}}(\hat{p}^Y \| p^Y), \\
R_{\hat{H}_{\text{MLE}}(X,Y)} &= -D_{\text{KL}}(\hat{p}^{XY} \| p^{XY}),
\end{aligned}$$

and

$$T_{\hat{H}_{\text{MLE}}(X)} = \sum_{i=1}^{m} (\hat{p}_i^X - p_i^X) \log_2 p_i^X, \tag{31}$$

$$T_{\hat{H}_{\text{MLE}}(Y)} = \sum_{j=1}^{n} (\hat{p}_j^Y - p_j^Y) \log_2 p_j^Y, \tag{32}$$

$$T_{\hat{H}_{\text{MLE}}(X,Y)} = \sum_{i=1}^{m} \sum_{j=1}^{n} (\hat{p}_{ij}^{XY} - p_{ij}^{XY}) \log_2 p_{ij}^{XY}. \tag{33}$$

In the above expressions, p^X, p^Y, and p^{XY} denote the true probability distributions of X, Y, and (X, Y), while $\hat{p}^X$, $\hat{p}^Y$, and $\hat{p}^{XY}$ denote the corresponding estimates, i.e.,

$$p_i^X = p(x_i), \qquad p_j^Y = p(y_j), \qquad p_{ij}^{XY} = p(x_i, y_j), \tag{34}$$

$$\hat{p}_i^X = {}^{c_i^X}\!/_N, \qquad \hat{p}_j^Y = {}^{c_j^Y}\!/_N, \qquad \hat{p}_{ij}^{XY} = {}^{c_{ij}^{XY}}\!/_N, \tag{35}$$

where $i = 1, \ldots, m$ and $j = 1, \ldots, n$.

Note that from (22) and (27) it follows that

$$\hat{p}_i^X = \sum_{j=1}^{n} \hat{p}_{ij}^{XY}, \; \hat{p}_j^Y = \sum_{i=1}^{m} \hat{p}_{ij}^{XY}, \; p_i^X = \sum_{j=1}^{n} p_{ij}^{XY}, \; p_j^Y = \sum_{i=1}^{m} p_{ij}^{XY}. \tag{36}$$

Therefore, $\hat{I}_{\text{MLE}}(X;Y)$, which was defined in (25), can be expressed as follows:

$$\begin{aligned}
\hat{I}_{\text{MLE}}(X;Y) &= \hat{H}_{\text{MLE}}(X) + \hat{H}_{\text{MLE}}(Y) - \hat{H}_{\text{MLE}}(X,Y) \\
&= \left(H(X) + R_{\hat{H}_{\text{MLE}}(X)} - T_{\hat{H}_{\text{MLE}}(X)}\right) + \left(H(Y) + R_{\hat{H}_{\text{MLE}}(Y)} - T_{\hat{H}_{\text{MLE}}(Y)}\right) \\
&\quad - \left(H(X,Y) + R_{\hat{H}_{\text{MLE}}(X,Y)} - T_{\hat{H}_{\text{MLE}}(X,Y)}\right) \\
&= (H(X) + H(Y) - H(X,Y)) + \left(R_{\hat{H}_{\text{MLE}}(X)} + R_{\hat{H}_{\text{MLE}}(Y)} - R_{\hat{H}_{\text{MLE}}(X,Y)}\right), \\
&\quad + \left(T_{\hat{H}_{\text{MLE}}(X)} + T_{\hat{H}_{\text{MLE}}(Y)} - T_{\hat{H}_{\text{MLE}}(X,Y)}\right) \\
&= I(X;Y) + R_{\hat{I}_{\text{MLE}}(X;Y)} - T_{\hat{I}_{\text{MLE}}(X;Y)}.
\end{aligned} \tag{37}$$

where

$$R_{\hat{I}_{\text{MLE}}(X;Y)} = R_{\hat{H}_{\text{MLE}}(X)} + R_{\hat{H}_{\text{MLE}}(Y)} - R_{\hat{H}_{\text{MLE}}(X,Y)}, \tag{38}$$

$$T_{\hat{I}_{\text{MLE}}(X;Y)} = T_{\hat{H}_{\text{MLE}}(X)} + T_{\hat{H}_{\text{MLE}}(Y)} - T_{\hat{H}_{\text{MLE}}(X,Y)}. \tag{39}$$

The following proposition shows that the term $T_{\hat{I}_{\text{MLE}}}$ simplifies to zero if the mutual information between X and Y is zero.

Proposition 9. *If $I(X;Y) = 0$, then $T_{\hat{I}_{MLE}(X;Y)} = 0$.*

The following proposition shows that $R_{\hat{I}_{\text{MLE}}(X;Y)}$ reduces to KL-divergence when $I(X;Y) = 0$. The proof is a sequence of straightforward algebraic transformations, which are provided here for completeness.

Proposition 10. *If $I(X;Y) = 0$, then*

$$R_{\hat{I}_{MLE}(X;Y)} = D_{KL}(\hat{p}^{XY} \| \hat{p}^X \otimes \hat{p}^Y),$$

where $\otimes$ denotes the outer product: $(\hat{p}^X \otimes \hat{p}^Y)_{ij} = \hat{p}_i^X \hat{p}_j^Y$ for $i = 1, \ldots, m$ and $j = 1, \ldots, n$.

The following proposition, which follows from Theorem 2, allows us to use Pearson's $\mathcal{X}^2$ statistic for $\hat{p}^{XY}$ and $\hat{p}^X \otimes \hat{p}^Y$ as an upper bound for $D_{\text{KL}}(\hat{p}^{XY} \| \hat{p}^X \otimes \hat{p}^Y)$ for the purpose of bounding $\hat{I}_{\text{MLE}}(X;Y)$. This is useful because $\mathcal{X}^2(\hat{p}^{XY}, \hat{p}^X \otimes \hat{p}^Y)$ has a limiting distribution with $(m-1)(n-1)$ degrees of freedom as $N \to \infty$. As described in the next subsection, this allows us to use the χ^2 cumulative distribution function to compute the p-value for the null hypothesis $I(X;Y) = 0$ and the confidence bound for the alternative hypothesis $I(X;Y) \neq 0$.

Proposition 11. $D_{KL}(\hat{p}^{XY} \| \hat{p}^X \otimes \hat{p}^Y) \leq \dfrac{\mathcal{X}^2(\hat{p}^{XY}, \hat{p}^X \otimes \hat{p}^Y)}{\ln 2}$.

The following inequality, which is obtained from (37) by applying Proposition 8, Proposition 9, Proposition 10, and Proposition 11 summarizes the results of this section:

$$0 \leq \hat{I}_{\text{MLE}}(X;Y) \leq \frac{\mathcal{X}^2(\hat{p}^{XY}, \hat{p}^X \otimes \hat{p}^Y)}{\ln 2}, \text{ if } I(X;Y) = 0. \tag{40}$$

O. Statistical Reasoning about Mutual Information

The following theorem, which was proved by Lévy [47] [48], allows us to infer the convergence in distribution for the limit of a sequence of random variables from pointwise convergence of characteristic functions of variables from the sequence. This theorem can be used to prove the Central Limit Theorem or to prove Pearson's Theorem (Theorem 6).

Theorem 5. ***Lévy's continuity theorem.***
Let $\{X_i\}$ be a sequence of random variables such that $X_i \in \mathbb{R}^n$ for $i = 1, \ldots, \infty$. Also let $F_i(x) = \Pr(X_i \leq x)$ and $\varphi_i(t) = E[e^{itX_i}]$ be the cumulative distribution function for X_i and the characteristic distribution function for X_i respectively, where $x, t \in \mathbb{R}^n$.

The following equivalence relation holds: $\lim_{i\to\infty} F_i(x) = F(x)$ and $F(x)$ is a cumulative distribution function if and only if 1) $\lim_{i\to\infty} \varphi_i(t) = \varphi(t)$ and 2) $\lim_{t\to 0} \varphi(t) = \varphi(0)$.

In other words, a necessary and sufficient condition for the pointwise convergence of the sequence of cumulative distribution function $\{F_i(x)\}$ to $F(x)$ such that $F(x)$ is also a cumulative distribution function is that 1) $\{\varphi_i(t)\}$ converges pointwise to a limit function $\varphi(t)$ and 2) $\varphi(t)$ is continuous at $t = 0$.

The following two theorems allow to perform statistical hypothesis testing for the hypothesis $I(X;Y) = 0$ using the bound (40). The first theorem is due to Karl Pearson [49] and the second is due to Ronald Fisher [50] [51]. These two results lie in the theoretical foundation of modern statistics. The proof for the Pearson's Theorem, which follows [52, pp. 418–419] is provided here.

Theorem 6. ***Pearson's Theorem.***
If $p = (p_1, p_2, \ldots, p_m)$ is a probability distribution such that $p > 0$, $m > 1$, and $\hat{p} = \hat{p}_{MLE} = {}^{c}/_{N}$ is its maximum likelihood estimator obtained from a vector of bin counters $c = (c_1, c_2, \ldots, c_m)$ of a histogram that was constructed using a sample of size N that was drawn i.i.d. from p, then

$$N\mathcal{X}^2(\hat{p}, p) \xrightarrow{d} \chi^2_{m-1} \quad \text{as } N \to \infty, \tag{41}$$

where

$$\mathcal{X}^2(\hat{p}, p) = \sum_{i=1}^{m} \frac{(\hat{p}_i - p_i)^2}{p_i} = \sum_{i=1}^{m} \frac{({}^{c_i}/_{N} - p_i)^2}{p_i}$$

is the Pearson's $\mathcal{X}^2$ statistic, χ^2_{m-1} is a random variable distributed according to χ^2 distribution with $m-1$ degrees of freedom, and $\xrightarrow{d}$ denotes convergence in distribution.

In other words, if

$$F_{\mathcal{X}^2}(\zeta) = \Pr(\mathcal{X}^2(\hat{p}, p) \leq \zeta),$$
$$F_{\chi^2_{m-1}}(\zeta) = \Pr(\chi^2_{m-1} \leq \zeta),$$

where $\zeta \in \mathbb{R}$, are cumulative distribution functions for $\mathcal{X}^2(\hat{p}, p)$ (i.e., a random variable defined by Pearson's $\mathcal{X}^2$ statistic) and χ^2_{m-1} (i.e., a random variable distributed according to χ^2 distribution with $m-1$ degrees of freedom) respectively, then

$$\lim_{N\to\infty} F_{\mathcal{X}^2}(N\zeta) = F_{\chi^2_{m-1}}(\zeta).$$

Proof: Let $\xi = (\xi_1, \xi_2, \ldots, \xi_m)$, where

$$\xi_i = \frac{c_i - Np_i}{\sqrt{Np_i}}, \quad i = 1, \ldots, m. \tag{42}$$

Note that

$$\xi_i^2 = \frac{(c_i - Np_i)^2}{Np_i} = N\frac{(c_i - Np_i)^2}{N^2 p_i} = N\frac{\frac{(c_i - Np_i)^2}{N^2}}{p_i} = N\frac{\left(\frac{c_i - Np_i}{N}\right)^2}{p_i} = N\frac{({}^{c_i}/_{N} - p_i)^2}{p_i}.$$

It follows that

$$\sum_{i=1}^{m} \xi_i^2 = \sum_{i=1}^{m} N\frac{({}^{c_i}/_{N} - p_i)^2}{p_i} = N\sum_{i=1}^{m} \frac{({}^{c_i}/_{N} - p_i)^2}{p_i} = N\mathcal{X}^2(\hat{p}, p).$$

Before the distribution of $(\xi_1^2 + \xi_2^2 + \ldots + \xi_N^2)$ is discussed, consider first each individual ξ_i for $i = 1, \ldots, m$.
Recall from (6) and (7) that

$$E[c_i] = Np_i,$$
$$\mathrm{Var}(c_i) = E[c_i^2] - (E[c_i])^2 = Np_i(1 - p_i).$$

Therefore,

$$E[c_i^2] = \mathrm{Var}(c_i) + (E[c_i])^2 = Np_i(1-p_i) + (Np_i)^2 = Np_i(1-p_i+Np_i).$$

These results can be used to express $E[\xi_i]$, $E[\xi_i^2]$, and $\mathrm{Var}(\xi_i)$ as follows:

$$E[\xi_i] = E\left[\frac{c_i - Np_i}{\sqrt{Np_i}}\right] = \frac{E[c_i - Np_i]}{\sqrt{Np_i}} = \frac{E[c_i] - Np_i}{\sqrt{Np_i}} = \frac{\cancel{Np_i} - \cancel{Np_i}}{\sqrt{Np_i}} = 0,$$

$$E[\xi_i^2] = E\left[\frac{(c_i - Np_i)^2}{Np_i}\right] = E\left[\frac{c_i^2 - 2Np_ic_i + (Np_i)^2}{Np_i}\right] = \frac{E\left[c_i^2 - 2Np_ic_i + (Np_i)^2\right]}{Np_i} = \frac{E[c_i^2] - E[2Np_ic_i] + E[(Np_i)^2]}{Np_i}$$

$$= \frac{\cancel{Np_i}(1-p_i+Np_i) - 2\cancel{Np_i}E[c_i] + (Np_i)^2}{\cancel{Np_i}} = 1 - p_i + \cancel{Np_i} - \cancel{2Np_i} + \cancel{Np_i} = 1 - p_i,$$

$$\mathrm{Var}(\xi_i) = E[\xi_i^2] - (E[\xi_i])^2 = 1 - p_i.$$

Thus, the expected value of ξ_i is 0 and its variance is $1 - p_i$ for $i = 1, \ldots, m$. Unfortunately, these results do not tell much about the distribution of $(\xi_1^2 + \xi_2^2 + \ldots \xi_m^2)$, because ξ_i are *not independent*. In fact, ξ_i are defined as transformed versions of c_i and $0 \neq \mathrm{Cov}(c_i, c_j) = -Np_ip_j$ for $i, j = 1, \ldots, m$ and $i \neq j$. Nonzero covariance implies that the variables are not dependent.

The proof is accomplished in two steps: 1) showing that the joint characteristic function $\varphi_\xi(t)$ of the random vector ξ converges pointwise to the joint characteristic function $\varphi_{\mathcal{N}^{m-1}}$ of a random vector $\mathcal{N}^{m-1}$ that consists of $m-1$ independent standard (i.e., with zero mean and unit variance) normal variables $\mathcal{N}$ and 2) applying Lévy's continuity theorem (Theorem 5) that allows us to conclude that $\xi \xrightarrow{d} \mathcal{N}^{m-1}$ as $N \to \infty$. Once this is established, the definition of χ^2 distribution and the fact that m is finite imply that

$$\sum_{i=1}^{m} \xi_i^2 \xrightarrow{d} \sum_{i=1}^{m-1} \mathcal{N}^2 \triangleq \chi^2_{m-1}.$$

Recall from Property 4 that the joint characteristic function $\varphi_c(t)$ of a multinomial vector c is equal to

$$\varphi_c(t) = E[e^{\imath t \cdot c}] = \left(p_1 e^{\imath t_1} + p_2 e^{\imath t_2} + \ldots + p_m e^{\imath t_m}\right)^N.$$

Therefore, the joint characteristic function of a random vector ξ, which is defined in (V-O) using c, can be expressed as follows:

$$\varphi_\xi(t) = E[e^{\imath t \cdot \xi}] = E\left[\exp\left(\sum_{k=1}^{m} \imath t_k \xi_k\right)\right] = E\left[\exp\left(\sum_{k=1}^{m} \imath t_k \frac{c_k - Np_k}{\sqrt{Np_k}}\right)\right] = E\left[\exp\left(\sum_{k=1}^{m} \frac{\imath t_k c_k}{\sqrt{Np_k}} - \sum_{k=1}^{m} \imath t_k \sqrt{Np_k}\right)\right]$$

$$= E\Bigg[\underbrace{\exp\left(\sum_{k=1}^{m} \frac{\imath t_k c_k}{\sqrt{Np_k}}\right)}_{\text{random}} \underbrace{\exp\left(-\sum_{k=1}^{m} \imath t_k \sqrt{Np_k}\right)}_{\text{fixed}}\Bigg] = \exp\left(-\sum_{k=1}^{m} \imath t_k \sqrt{Np_k}\right) E\left[\exp\left(\sum_{k=1}^{m} \frac{\imath t_k c_k}{\sqrt{Np_k}}\right)\right]$$

$$= \exp\left(-\imath\sqrt{N}\sum_{k=1}^{m} t_k \sqrt{p_k}\right) \underbrace{E\left[\exp\left(\sum_{k=1}^{m} \frac{\imath t_k c_k}{\sqrt{Np_k}}\right)\right]}_{U}. \tag{43}$$

Let $\tau = (\tau_1, \tau_2, \ldots, \tau_m)$ be a vector where

$$\tau_k = \frac{t_k}{\sqrt{Np_k}}$$

for $k = 1, \ldots, m$. Note that

$$\tau \cdot c = \sum_{k=1}^{m} \tau_k c_k = \sum_{k=1}^{m} \frac{t_k c_k}{\sqrt{Np_k}}.$$

It follows that the term U in (43) can be written as follows:

$$U = E\left[\exp\left(\sum_{k=1}^{m} \frac{\imath t_k c_k}{\sqrt{Np_k}}\right)\right] = E\left[\exp\left(\imath \sum_{k=1}^{m} \frac{t_k c_k}{\sqrt{Np_k}}\right)\right] = E[e^{\imath \tau \cdot c}] = \varphi_c(\tau) = (p_1 e^{\imath\tau_1} + p_2 e^{\imath\tau_2} + \ldots + p_m e^{\imath\tau_m})^N$$

$$= \left(\sum_{k=1}^{m} p_k e^{\imath \tau_k}\right)^N = \left(\sum_{k=1}^{m} p_k \exp\left(\frac{\imath t_k}{\sqrt{Np_k}}\right)\right)^N.$$

Plugging the above expression for U into (43) allows us to write $\varphi_\xi(t)$ as follows:

$$\varphi_\xi(t) = \exp\left(-\imath\sqrt{N}\sum_{k=1}^{m} t_k\sqrt{p_k}\right)\left(\sum_{k=1}^{m} p_k \exp\left(\frac{\imath t_k}{\sqrt{Np_k}}\right)\right)^N.$$

To find the limiting function for $\varphi_\xi(t)$ as $N\to\infty$, consider first $\ln\varphi_\xi(t)$:

$$\begin{aligned}\ln\varphi_\xi(t) &= \ln\left(\exp\left(-\imath\sqrt{N}\sum_{k=1}^{m} t_k\sqrt{p_k}\right)\left(\sum_{k=1}^{m} p_k\exp\left(\frac{\imath t_k}{\sqrt{Np_k}}\right)\right)^N\right) = \ln\left(\exp\left(-\imath\sqrt{N}\sum_{k=1}^{m} t_k\sqrt{p_k}\right)\right) + \ln\left(\left(\sum_{k=1}^{m} p_k \exp\left(\frac{\imath t_k}{\sqrt{Np_k}}\right)\right)^N\right)\\ &= -\imath\sqrt{N}\sum_{k=1}^{m} t_k\sqrt{p_k} + N\ln\Bigg(\underbrace{\sum_{k=1}^{m} p_k\exp\left(\frac{\imath t_k}{\sqrt{Np_k}}\right)}_{V(t)}\Bigg).\end{aligned} \tag{44}$$

Note that $V(t)$ is infinitely differentiable because $\exp\left(\frac{\imath t_k}{\sqrt{Np_k}}\right)$ is infinitely differentiable for each $k=1,\dots,m$, which implies that the sum is also infinitely differentiable.

Let

$$v_t(s) = V(st) = \sum_{k=1}^{m} p_k \exp\left(\frac{\imath s t_k}{\sqrt{Np_k}}\right),$$

where $s\in\mathbb{R}$. Note that $v_t(s)$ is infinitely differentiable, which implies that it can be expressed using its MacLaurin series as shown below:

$$v_t(s) = v_t(0) + \sum_{n=1}^{\infty}\frac{v_t^{(n)}(0)}{n!}s^n. \tag{45}$$

Note that

$$v_t(0) = V(0) = \sum_{k=1}^{m} p_k\exp(0) = \sum_{k=1}^{m} p_k = 1 \tag{46}$$

$$\begin{aligned}v_t'(s) &= \frac{d}{ds}v_t(s) = \frac{d}{ds}\left(\sum_{k=1}^{m} p_k\exp\left(\frac{\imath s t_k}{\sqrt{Np_k}}\right)\right) = \sum_{k=1}^{m}\frac{d}{ds}\left(p_k\exp\left(\frac{\imath s t_k}{\sqrt{Np_k}}\right)\right) = \sum_{k=1}^{m} p_k\frac{d}{ds}\exp\left(\frac{\imath s t_k}{\sqrt{Np_k}}\right)\\ &= \sum_{k=1}^{m} p_k\exp\left(\frac{\imath s t_k}{\sqrt{Np_k}}\right)\frac{d}{ds}\left(\frac{\imath s t_k}{\sqrt{Np_k}}\right) = \sum_{k=1}^{m} p_k\exp\left(\frac{\imath s t_k}{\sqrt{Np_k}}\right)\frac{\imath t_k}{\sqrt{Np_k}} = \frac{\imath}{\sqrt{N}}\sum_{k=1}^{m}\frac{p_k t_k}{\sqrt{p_k}}\exp\left(\frac{\imath s t_k}{\sqrt{Np_k}}\right)\\ &= \frac{\imath}{\sqrt{N}}\sum_{k=1}^{m} t_k\sqrt{p_k}\exp\left(\frac{\imath s t_k}{\sqrt{Np_k}}\right),\end{aligned} \tag{47}$$

$$\begin{aligned}v_t''(s) &= \frac{d}{ds}v_t'(s) = \frac{d}{ds}\left(\frac{1}{\sqrt{N}}\sum_{k=1}^{m}\imath t_k\sqrt{p_k}\exp\left(\frac{\imath s t_k}{\sqrt{Np_k}}\right)\right) = \frac{\imath}{\sqrt{N}}\sum_{k=1}^{m} t_k\sqrt{p_k}\frac{d}{ds}\exp\left(\frac{\imath s t_k}{\sqrt{Np_k}}\right)\\ &= \frac{\imath}{\sqrt{N}}\sum_{k=1}^{m} t_k\cancel{\sqrt{p_k}}\exp\left(\frac{\imath s t_k}{\sqrt{Np_k}}\right)\frac{\imath t_k}{\cancel{\sqrt{Np_k}}} = \frac{\imath^2}{N}\sum_{k=1}^{m} t_k^2\exp\left(\frac{\imath s t_k}{\sqrt{Np_k}}\right) = -\frac{1}{N}\sum_{k=1}^{m} t_k^2\exp\left(\frac{\imath s t_k}{\sqrt{Np_k}}\right),\end{aligned} \tag{48}$$

$$\begin{aligned}v_t'''(s) &= \frac{d}{ds}v_t''(s) = \frac{d}{ds}\left(-\frac{1}{N}\sum_{k=1}^{m} t_k^2\exp\left(\frac{\imath s t_k}{\sqrt{Np_k}}\right)\right) = -\frac{1}{N}\sum_{k=1}^{m} t_k^2\frac{d}{ds}\exp\left(\frac{\imath s t_k}{\sqrt{Np_k}}\right) = -\frac{1}{N}\sum_{k=1}^{m} t_k^2\exp\left(\frac{\imath s t_k}{\sqrt{Np_k}}\right)\frac{\imath t_k}{\sqrt{Np_k}}\\ &= -\frac{\imath}{N\sqrt{N}}\sum_{k=1}^{m}\frac{t_k^3}{\sqrt{p}_k}\exp\left(\frac{\imath s t_k}{\sqrt{Np_k}}\right).\end{aligned} \tag{49}$$

Recall that MacLaurin series (45) can be also written as follows:

$$v_t(s) = v_t(0) + v_t'(0)s + \frac{1}{2}v_t''(0)s^2 + R_2^v(s), \tag{50}$$

where $R_2^v(s)$ is the remainder term of the second order, which can be written in the Lagrange form as follows:

$$R_2^v(s) = \frac{1}{3!}v_t'''(\eta)s^3 = \frac{1}{6}v_t'''(\eta)s^3, \tag{51}$$

for some $\eta\in(0,s)$.

The expression for $V(t)$ can be obtained from (50) and (51) for the case when $s=1$ using (46) (47) (48) (49), as shown below:

$$\begin{aligned}
V(t) = v_t(1) &= v_t(0) + v_t'(0) + \frac{1}{2}v_t''(0) + \frac{1}{6}v_t'''(\eta) \\
&= 1 + \frac{\imath}{\sqrt{N}}\sum_{k=1}^{m} t_k\sqrt{p_k}\exp(0) + \frac{1}{2}\left(-\frac{1}{N}\sum_{k=1}^{m} t_k^2\exp(0)\right) + \frac{1}{6}\left(-\frac{\imath}{N\sqrt{N}}\sum_{k=1}^{m}\frac{t_k^3}{\sqrt{p}_k}\exp\left(\frac{\imath\eta t_k}{\sqrt{Np_k}}\right)\right) \\
&= 1 + \frac{\imath}{\sqrt{N}}\sum_{k=1}^{m} t_k\sqrt{p_k} - \frac{1}{2N}\sum_{k=1}^{m} t_k^2 \underbrace{-\frac{\imath}{6N\sqrt{N}}\sum_{k=1}^{m}\frac{t_k^3}{\sqrt{p}_k}\exp\left(\frac{\imath\eta t_k}{\sqrt{Np_k}}\right)}_{O\left(N^{-\frac{3}{2}}\right)} \\
&= 1 + \frac{\imath}{\sqrt{N}}\sum_{k=1}^{m} t_k\sqrt{p_k} - \frac{1}{2N}\sum_{k=1}^{m} t_k^2 + O\left(N^{-\frac{3}{2}}\right) = 1 + \underbrace{\frac{1}{\sqrt{N}}\left(\imath\sum_{k=1}^{m} t_k\sqrt{p}_k - \frac{1}{2\sqrt{N}}\sum_{k=1}^{m} t_k^2 + O\left(N^{-1}\right)\right)}_{W(t)}.
\end{aligned} \tag{52}$$

Let $w_t(s) = \ln(1+sW(t))$. Note that $w_t(1)$ is equal to the term $\ln V(t)$ in (44). Recall that $w_t(s)$ can be expressed using Mercator series, which is the Taylor series for the natural logarithm around 1, as shown below:

$$w_t(s) = \underbrace{w_t(0)}_{0} + \sum_{n=1}^{\infty}\frac{1}{n!}w_t^{(n)}(0)s^n = \sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n}(W(t))^n s^n = W(t)s - \frac{(W(t)s)^2}{2} + \frac{(W(t)s)^3}{3} - \frac{(W(t)s)^4}{4} + \dots$$

Similarly to (50), the above MacLaurin series can be written using the remainder term as follows:

$$w_t(s) = w_t'(0)s - \frac{w_t''(0)s^2}{2} + R_2^w(s) = W(t)s - \frac{(W(t)s)^2}{2} + R_2^w(s), \tag{53}$$

where $R_2^w(s)$ is the remainder term of the second order that can be expressed in the Lagrange form as shown below:

$$R_2^w(s) = \frac{1}{3!}w_t'''(\eta)s^3 = \frac{1}{3}(W(\eta)s)^3, \tag{54}$$

for some $\eta \in (0, s)$.

Similarly to (52) for $V(t)$, an expression for $\ln(1+W(t))$ can be obtained from the Taylor series for $w_t(s)$ (53) and its remainder term (54) by letting $s=1$, which results in the following expression:

$$\ln(1+W(t)) = w_t(1) = W(t) - \frac{W^2(t)}{2} + R_2^w(1). \tag{55}$$

Note that

$$\begin{aligned}
R_2^w(1) &= \frac{1}{3}W^3(\eta) = \frac{1}{3}\left(\frac{1}{\sqrt{N}}\left(\imath\sum_{k=1}^{m} t_k\sqrt{p_k} - \frac{1}{2\sqrt{N}}\sum_{k=1}^{m} t_k^2 + O(N^{-1})\right)\right)^3 \\
&= \frac{1}{3}\frac{1}{N\sqrt{N}}\left(\imath\sum_{k=1}^{m} t_k\sqrt{p_k} - \frac{1}{2\sqrt{N}}\sum_{k=1}^{m} t_k^2 + O(N^{-1})\right)^3 = O\left(\frac{1}{N\sqrt{N}}\right) = O\left(N^{-\frac{3}{2}}\right), \\
\frac{W^2(t)}{2} &= \frac{1}{2}\left(\frac{1}{\sqrt{N}}\left(\imath\sum_{k=1}^{m} t_k\sqrt{p_k} - \frac{1}{2\sqrt{N}}\sum_{k=1}^{m} t_k^2 O\left(N^{-1}\right)\right)\right)^2 = \frac{1}{2N}\left(\imath\sum_{k=1}^{m} t_k\sqrt{p_k} - \frac{1}{2\sqrt{N}}\sum_{k=1}^{m} t_k^2 + O\left(\frac{1}{N}\right)\right)^2 \\
&= \frac{1}{2N}\left(\imath\sum_{k=1}^{m} t_k\sqrt{p_k} + \underbrace{\frac{1}{\sqrt{N}}\left(-\frac{1}{2}\sum_{k=1}^{m} t_k^2 + O\left(\frac{1}{\sqrt{N}}\right)\right)}_{O\left(\frac{1}{\sqrt{N}}\right)}\right)^2 = \frac{1}{2N}\left(\imath\sum_{k=1}^{m} t_k\sqrt{p_k} + O\left(\frac{1}{\sqrt{N}}\right)\right)^2 \\
&= \frac{1}{2N}\left(\left(\imath\sum_{k=1}^{m} t_k\sqrt{p_k}\right)^2 + \underbrace{2\,O\left(\frac{1}{\sqrt{N}}\right)\imath t_k\sqrt{p_k} + O^2\left(\frac{1}{\sqrt{N}}\right)}_{O\left(\frac{1}{\sqrt{N}}\right)}\right) = \frac{1}{2N}\left(\imath^2\left(\sum_{k=1}^{m} t_k\sqrt{p_k}\right) + O\left(\frac{1}{\sqrt{N}}\right)\right)
\end{aligned} \tag{56}$$

$$= \frac{1}{2N}\left(-\left(\sum_{k=1}^{m} t_k\sqrt{p_k}\right)^2 + O\left(\frac{1}{\sqrt{N}}\right)\right) = -\frac{1}{2N}\left(\sum_{k=1}^{m} t_k\sqrt{p_k}\right)^2 + \frac{1}{2N}O\left(\frac{1}{\sqrt{N}}\right)$$
$$= -\frac{1}{2N}\left(\sum_{k=1}^{m} t_k\sqrt{p_k}\right)^2 + O\left(\frac{1}{N\sqrt{N}}\right). \tag{57}$$

Using the formulas (52), (57), and (56) for the terms $W(t)$, $W^2(t)/2$, and $R_2^w(t)$, respectively, it is possible to find the formulas for the terms $NW(t)$, $NW^2(t)/2$, and $NR_2^w(t)$ that constitute the second term $N\ln V(t) = N\ln(1+W(t))$ within (44), according to (55). The resulting three expressions are shown below:

$$NW(t) = N\left(\frac{1}{\sqrt{N}}\left(\imath\sum_{k=1}^{m} t_k\sqrt{p_k} - \frac{1}{2\sqrt{N}}\sum_{k=1}^{m} t_k^2 + O(N^{-1})\right)\right) = \sqrt{N}\left(\imath\sum_{k=1}^{m} t_k\sqrt{p_k} - \frac{1}{2\sqrt{N}}\sum_{k=1}^{m} t_k^2 + O(N^{-1})\right)$$
$$= \imath\sqrt{N}\sum_{k=1}^{m} t_k\sqrt{p_k} - \cancel{\sqrt{N}}\frac{1}{2\cancel{\sqrt{N}}}\sum_{k=1}^{m} t_k^2 + \cancel{\sqrt{N}}O\left(\frac{1}{\cancel{\sqrt{N}}\sqrt{N}}\right) = \imath\sqrt{N}\sum_{k=1}^{m} t_k\sqrt{p_k} - \frac{1}{2}\sum_{k=1}^{m} t_k^2 + O\left(\frac{1}{\sqrt{N}}\right), \tag{58}$$

$$N\frac{W^2(t)}{2} = \cancel{N}\left(-\frac{1}{2\cancel{N}}\left(\sum_{k=1}^{m} t_k\sqrt{p_k}\right)^2 + O\left(\frac{1}{\cancel{N}\sqrt{N}}\right)\right) = -\frac{1}{2}\left(\sum_{k=1}^{m} t_k\sqrt{p_k}\right)^2 + O\left(\frac{1}{\sqrt{N}}\right), \tag{59}$$

$$NR_2^w(t) = \cancel{N}O\left(\frac{1}{\cancel{N}\sqrt{N}}\right) = O\left(\frac{1}{\sqrt{N}}\right). \tag{60}$$

It is now possible to expres $\ln\phi_\xi(t)$ plugging (58), (59), and (60) into (44) with respect to (55), which results in the following formulas:

$$\ln\varphi_\xi(t) = -\imath\sqrt{N}\sum_{k=1}^{m} t_k\sqrt{p_k} + N\ln V(t) = -\imath\sqrt{N}\sum_{k=1}^{m} t_k\sqrt{p_k} + N\ln(1+W(t))$$
$$= -\imath\sqrt{N}\sum_{k=1}^{m} t_k\sqrt{p_k} + N\left(W(t) - \frac{W^2(t)}{2} + R_2^w(t)\right) = -\imath\sqrt{N}\sum_{k=1}^{m} t_k\sqrt{p_k} + NW(t) - N\frac{W^2(t)}{2} + NR_2^w(1)$$
$$= \cancel{-\imath\sqrt{N}\sum_{k=1}^{m} t_k\sqrt{p_k}} + \cancel{\imath\sqrt{N}\sum_{k=1}^{m} t_k\sqrt{p_k}} - \frac{1}{2}\sum_{k=1}^{m} t_k^2 + O\left(\frac{1}{\sqrt{N}}\right) + \frac{1}{2}\left(\sum_{k=1}^{m} t_k\sqrt{p_k}\right)^2 - O\left(\frac{1}{\sqrt{N}}\right) + O\left(\frac{1}{\sqrt{N}}\right)$$
$$= -\frac{1}{2}\left(\sum_{k=1}^{m} t_k^2 - \left(\sum_{k=1}^{m} t_k\sqrt{p_k}\right)^2\right) + O\left(\frac{1}{\sqrt{N}}\right).$$

The last expression makes it possible to find the limiting function for the joint characteristic function $\varphi_\xi(t)$ for the random vector ξ, which was defined in :

$$\lim_{N\to\infty}\varphi_\xi(t) = \lim_{N\to\infty}\exp\left(\ln\varphi_\xi(t)\right) = \lim_{N\to\infty}\exp\left(-\frac{1}{2}\left(\sum_{k=1}^{m} t_k^2 - \left(\sum_{k=1}^{m} t_k\sqrt{p_k}\right)^2\right) + \underbrace{O\left(\frac{1}{\sqrt{N}}\right)}_{\to 0}\right) = \exp\left(-\frac{1}{2}\left(\sum_{k=1}^{m} t_k^2 - \left(\sum_{k=1}^{m} t_k\sqrt{p_k}\right)^2\right)\right)$$
$$= \exp\left(-\frac{1}{2}Q(t)\right), \tag{61}$$

where $Q(t)$ is the quadratic from that is defined defines as follows:

$$Q(t) = \sum_{k=1}^{m} t_k^2 - \left(\sum_{k=1}^{m} t_k\sqrt{p_k}\right)^2. \tag{62}$$

Note that $Q(t)$ can be also expressed using the matrix form, i.e. $Q(t) = tAt^T$, where $t = (t_1, t_2, \ldots, t_m)$ denotes the $1\times m$ single row vector, t^T denotes its transpose, and A is the $m\times m$ square matrix, which is defined as follows:

$$A = I - \varrho^T\varrho,$$

where I denotes the $m\times m$ identity matrix and $\varrho = \sqrt{p} = (\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_m})$. Indeed, the following sequence of algebraic

transformations shows that the matrix form tAt^T is equivalent to the definition of $Q(t)$ in (62):

$$tAt^T = t(I - \varrho^T\varrho)t^T = tIt^T - t(\varrho^T\varrho)t^T = tt^T - (t\varrho^T)(\varrho t^T) = \sum_{k=1}^{m} t_k^2 - \left(\sum_{k=1}^{m} t_k\sqrt{p_k}\right)\left(\sum_{k=1}^{m} \sqrt{p_k}t_k\right)$$

$$= \sum_{k=1}^{m} t_k^2 - \left(\sum_{k=1}^{m} t_k\sqrt{p_k}\right)^2 = Q(t).$$

Note that

$$\|\varrho\|_2^2 = \sum_{k=1}^{m} \varrho_k^2 = \sum_{k=1}^{m} (\sqrt{p_k})^2 = \sum_{k=1}^{m} p_k = 1.$$

It follows that A is an orthonormal projection matrix and that the rank of A is $m-1$.

Let $G = \left\{g^{(1)}, g^{(2)}, \ldots, g^{(m)}\right\}$ be an orthonormal basis in $\mathbb{R}^m$ such that $g^{(m)} = \varrho$. It is always possible to construct a G by applying the Gram-Schmidt orthonormalization process to a set of $m+1$ vectors

$$\left\{ \begin{array}{c} \varrho, \\ (1,0,0,0,\ldots,0,0), \\ (0,1,0,0,\ldots,0,0), \\ \vdots \\ (0,0,0,0,\ldots,0,1) \end{array} \right\},$$

which results in a set of m orthonormal vectors that span $\mathbb{R}^m$ such that one of these vectors is ϱ.

Let $u = (u_1, u_2, \ldots, u_m)$ be the coordinates of a vector $t = (t_1, t_2, \ldots, t_m)$ in the coordinate system defined by the basis G. In other words,

$$t = u_1 g^{(1)} + u_2 g^{(2)} + \ldots + u_m g^{(m)},$$

where $u_k = t \cdot g^{(k)} = t_1 g_1^{(k)} + t_2 g_2^{(k)} + \ldots + t_m g_m^{(k)}$. Note that

$$\|u\|_2^2 = \sum_{k=1}^{m} u_k^2 = \sum_{k=1}^{m} t_k^2 = \|t\|_2^2,$$

$$u_m^2 = \left(t \cdot g^{(m)}\right)^2 = (t \cdot \varrho)^2 = (t \cdot \sqrt{\varrho})^2 = \left(\sum_{k=1}^{m} t_k\sqrt{p_k}\right)^2.$$

Therefore,

$$Q(t) = \sum_{k=1}^{m} t_k^2 - \left(\sum_{k=1}^{m} t_k\sqrt{p_k}\right) = \sum_{k=1}^{m} u_k^2 - u_m^2 = \sum_{k=1}^{m-1} u_k^2.$$

Recall from (61) that the limit for the joint characteristic function $\varphi_\xi(t)$ as $N \to \infty$ is equal to

$$\lim_{N\to\infty} \varphi_\xi(t) = \exp\left(-\frac{1}{2}Q(t)\right) = \exp\left(-\frac{1}{2}\sum_{k=1}^{m-1} u_k^2\right) = \varphi_{\mathcal{N}^{m-1}}(u_1, u_2, \ldots, u_{m-1}),$$

where $\varphi_{\mathcal{N}^{m-1}}$ is the joint characteristic function for a vector of $m-1$ independent standard normal variables.

Therefore, Levy's continuity theorem (Theorem 5) implies implies that $\xi \xrightarrow{d} \mathcal{N}^{m-1}$ as $N \to \infty$. It follows that

$$N\mathcal{X}^2(\hat{p}, p) = \sum_{k=1}^{m} \xi_k^2 \xrightarrow{d} \sum_{i=1}^{m-1} \mathcal{N}^2 = \chi_{m-1}^2.$$

This completes the proof for Pearson's theorem. ■

Theorem 7. ***Fisher's Theorem.***
Let s and r be integers such that $s < r$ and let $f : \mathbb{R}^s \to \mathbb{R}^r$ be a function that satisfies the following four conditions in each point $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_s)$ within a non-degenerate interval $A \subseteq \mathbb{R}^s$:
a) the sum of r individual entries in $f(\alpha)$ is equal to one:

$$f_1(\alpha) + f_2(\alpha) + \ldots + f_r(\alpha) = 1; \tag{63}$$

b) there exists $\varepsilon > 0$, which does not depend on α, such that f is bounded away from zero by ε^2:

$$f_i(\alpha) > \varepsilon^2, \qquad (i = 1, \ldots, r); \tag{64}$$

c) partial derivatives f' and f'' of the first and second orders are continuous, i.e.,

$$f'_{i,j}(\alpha) = \frac{\partial f_i(\alpha)}{\partial \alpha_j} \in \mathcal{C}, \quad f''_{i,jk}(\alpha) = \frac{\partial^2 f_i(\alpha)}{\partial \alpha_j \partial \alpha_k} \in \mathcal{C}, \tag{65}$$

where $i = 1, \ldots, r$ and $j, k = 1, \ldots, s$;
c) the matrix $D(\alpha) = \left[f'_{i,j}(\alpha)\right]$, where $i = 1, \ldots, r$ and $j = 1, \ldots, s$, has rank s, i.e.,

$$\operatorname{rank}(D(\alpha)) = \operatorname{rank}\begin{bmatrix} f'_{1,1}(\alpha) & \cdots & f'_{1,s}(\alpha) \\ \vdots & \ddots & \vdots \\ f'_{r,1}(\alpha) & \cdots & f'_{r,s}(\alpha) \end{bmatrix} = s. \tag{66}$$

Suppose that $p = (p_1, p_2, \ldots, p_r)$ is a probability distribution such that $p = f(\alpha^*)$ for an α^* that is an inner point of A. Also suppose that $c = (c_1, c_2, \ldots, c_r)$ is a vector of bin counters for a histogram that was constructed using a sample of size N that was drawn i.i.d. from p (i.e., $c_1 + c_2 + \ldots + c_r = N$). Let $\hat{p} = \hat{p}_{MLE} = {}^{c}/_{N} = ({}^{c_1}/_{N}, {}^{c_2}/_{N}, \ldots, {}^{c_r}/_{N})$ be the maximum likelihood estimator for p.

This theorem states that:
1) The system of s equations, which is defined as follows:

$$\sum_{i=1}^{r} \frac{c_i - N f_i(\alpha)}{f_i(\alpha)} f'_{i,j}(\alpha) = 0, \quad (j = 1, \ldots, s), \tag{67}$$

has a unique solution $\hat{\alpha} = (\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_s)$.
2) $\hat{\alpha} \xrightarrow{p} \alpha^*$. In other words, $\hat{\alpha}$, which is the solution for (67)*, converges in probability to α^*, which can be also expressed as follows:*

$$\lim_{N \to \infty} \Pr\left(\max_{i \in 1, \ldots, r} |\hat{\alpha}_i - \alpha_i^*| \geq \delta \right) = 0, \tag{68}$$

for any $\delta > 0$.
3) $N\mathcal{X}^2(\hat{p}_{MLE}, \hat{p}_f) \xrightarrow{d} \chi^2_{r-s-1}$, where $\xrightarrow{d}$ denotes convergence in distribution, $\hat{p}_f = f(\hat{\alpha})$, $\mathcal{X}^2$ is the Pearson's $\mathcal{X}^2$ statistic, i.e.,

$$\mathcal{X}^2(\hat{p}_{MLE}, \hat{p}_f) = \sum_{i=1}^{r} \frac{\left((\hat{p}_{MLE})_i (\hat{p}_f)_i\right)^2}{(\hat{p}_f)_i} = \sum_{i=1}^{r} \frac{\left({}^{c_i}/_{N} - f_i(\hat{\alpha})\right)^2}{f_i(\hat{\alpha})},$$

and χ^2_{r-s-1} is the χ^2 distribution with $r - s - 1$ degrees of freedom. In other words,

$$\lim_{N \to \infty} \Pr(\mathcal{X}^2(\hat{p}_{MLE}, \hat{p}_f) < \eta) = \Pr(\chi^2_{r-s-1} < \eta), \tag{69}$$

for any $\eta \in \mathbb{R}$.

P. Experimental utilization of this approach

The entropy-based self-detection algorithm formulated above was implemented in three separate scenarios:

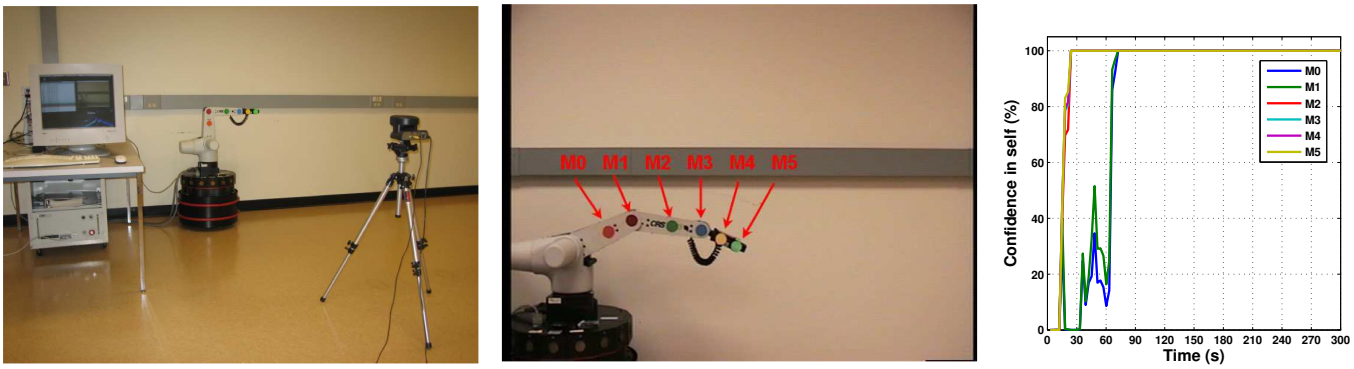
1) *Self-detection through engaging objects with exploratory shaking behaviors:*
   A robotic learning approach will be utilized to categorize objects and their relationships to other objects using mutual information about the objects. The robot will attempt to categorize objects and distinguish these as self or other in identity.
2) *Self-detection through engaging objects with exploratory pushing behavior:*
   A robotic learning approach will be utilized to categorize perceptions as belonging to the self or to other based upon its use of exploratory pushing behaviors and the confidence in mutual information gained about the object given information about the hand.
3) *Self-detection through game play:*
   Finally, a robot will learn to categorize its virtual perceptions of elements as belonging to self in agency or to other and the resultant confidence in self identity over time will be recorded.

VI. Results

A. Detection of Self Using Mutual Information

Evaluation was performed on the dataset from [53]. The dataset consists of the timestamped color marker tracking coordinates, recorded while the robot was performing motor babbling. Fig. 1(a) and 1(b) give a visual summary of the self-detection experiment used to collect this dataset. For a detailed description of the dataset and the experiment, the reader is referred to [53, Chapter 5].

Results for entropy-based self-detection show that the method can be applied for this task. For one of the CRS+ A251 datasets, the results are shown in Fig. 1(c). The algorithm detects all seven marker on its body as self within 90 seconds.



(a) Robot (b) Color Markers (c) Results

Fig. 1: Results for the dataset collected using the CRS+ A251 robot: (a) the robot used to collect the dataset; (b) six color markers used for self-detection; (c) results for the entropy-based approach. Results indicate high confidence in the self hypothesis. The experimental setup is described in more detail in [53, Chapter 5], from which (a) and (b) are reprinted with permission.

B. Understanding Containers Using Mutual Information

One of the goals of Shane Griffith's course project [54] was to use the link between controllability and movement dependencies to infer that a container allows the robot to influence movement of its contents.

In this experiment the robot performed 100 trials, each of which consisted of the following actions: 1) grasping a block, 2) shaking the block, 3) dropping the block onto an object, which was either a container or a non-container, 4) grasping the object, 5) shaking the object, and 6) dropping the object.

If the object was a container, then the robot was shaking both the container and the block inside it. Otherwise, the robot was only shaking the object and not the block.

The video recorded by the robot during these experiments was processed to extract three binary variables for each frame: 1) movement of the robotic hand, 2) movement of the block, and 3) movement of the object. The mutual information was estimated using a sliding temporal window of size 3 seconds for each of the three pairs of these variables. Example results for a single trial are shown in Fig. 2. In particular, Fig. 2(d) shows the number of dependent movement variables (a pair of variables was dependent if the p-value for the null hypothesis $I = 0$ was below the predefined threshold of .01, which corresponds to the confidence level above 99%). As the robot was shaking the block, there was one pair of dependent movement variables – i.e., hand–block. Later, when the robot was shaking the object, which in this case was a container with the block inside it, the mutual information indicated that there are three pairs of dependent movement variables: hand–object, hand–block, and object–block.

This experiment showed that the ability to estimate and reason about mutual information can capture differences between containers and non-containers. This difference was captured as the number of pairs of dependent movement variables. Only for a container the number of these pairs could be equal to three in this experiment.

C. Learning to Play Video Games Using Mutual Information

One of the motivations behind Pavel Kazatsker's course project [55] was to teach robots to play video games. The ability to identify controllable elements of the game and associate them with the robot's actions is the key to this process.

During the experiment, the robot played an analog of Arkanoid. Video recorded by the camera in one of the robot's eyes was segmented into moving components. Next, mutual information between movement variables for these components and the temporal delay after the last motor command was estimated and those components for which a very high level of confidence (above 99.99%) for the hypothesis $I > 0$ was achieved were labeled as "self". A brief summary of these results is shown in Fig. 3. In particular, the robot was able to label component 13, which corresponded to the "paddle" in Arkanoid, as "self" within 20 seconds. This "paddle" is indeed the only element of this game that the robot could control.

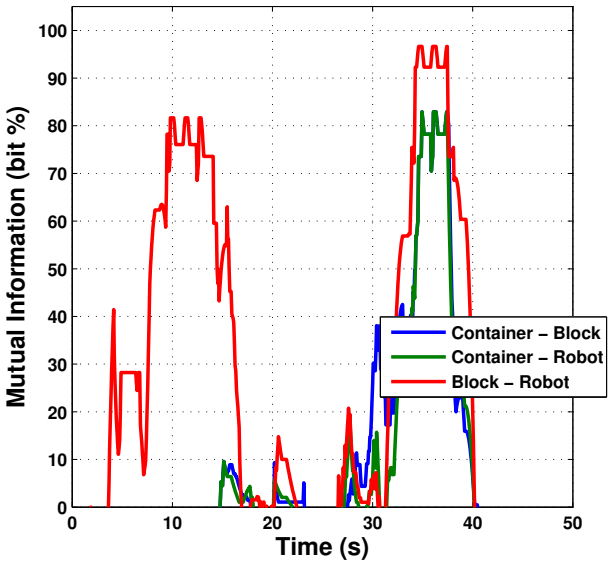
This evaluation showed that mutual information can be useful for teaching robots how to control virtual objects. In particular, it was shown that the robot can detect that it can control a "paddle" in Arkanoid.

D. Support Affordance Detection Using Mutual Information

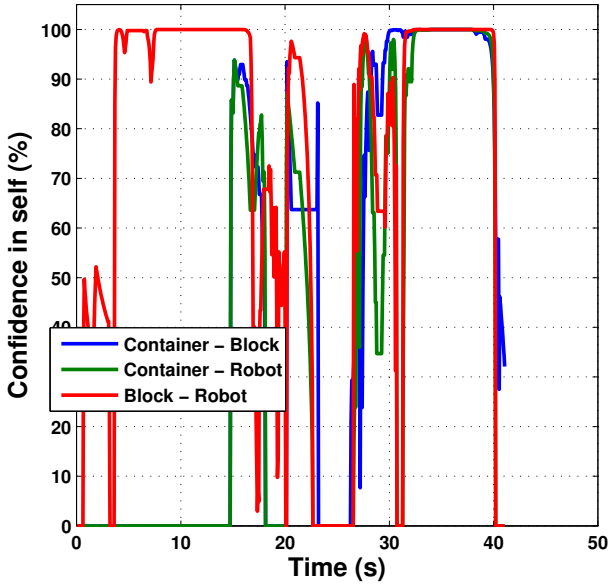
The goal of the project by Karl Deakyne, Yehoshua Meyer, and Brian Russell was to describe how a robot can learn support affordances. One of the aspects of this project was to show how a robot can detect that an object is no longer supported.

In this experiment the robot pushed an object until it slid off a cliff. Mutual information for the visual displacement of the robot's hand and the object from frame to frame in the video was analyzed to see if it is possible to detect the moment when

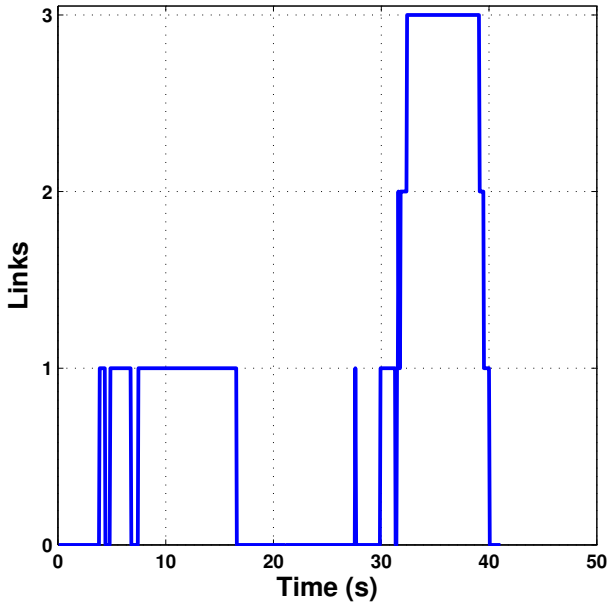
(a) Robot Grasping a Container



(b) Mutual Information



(c) Confidence in Nonnegative Mutual Information



(d) Number of Mutually Dependent Object Pairs

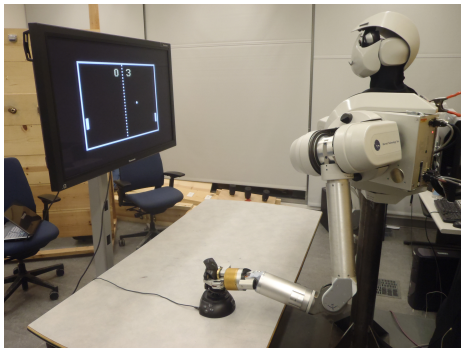
Fig. 2: Summary of the results for understanding containers through mutual information. See text for more details.

an object starts sliding. The example evaluation for one of the trials is shown in Fig. 4. Upon reaching the cliff, the mutual information starts rapidly decreasing. This is also observed for the confidence in $I > 0$. In other words, the cliff traversed by an object translates into a cliff in mutual information and confidence level.

This experiment showed that mutual information can be used to detect when a robot has pushed an object over a cliff.

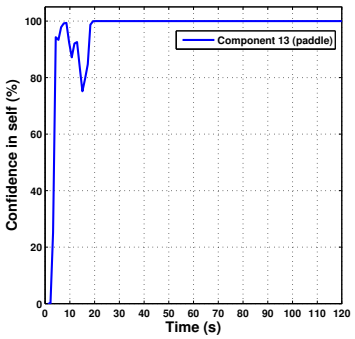
VII. Discussion

Consistent with preliminary research, the adoption of the entropy-based learning approach as outlined through the aforementioned formulations leads to fast, clear and consistent results in both categorization and self-identity tasks. In the first experiment, exploratory behaviors were used to determine the difference between self versus object. In this experiment, the realization of the self directly related to the confidence (p-value) of mutual information gained from the object given the



(a) Robot Playing a Video Game

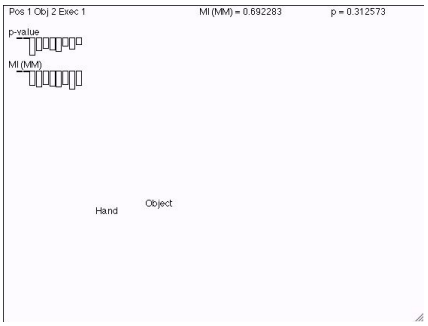
(b) Internal Representation of Arkanoid



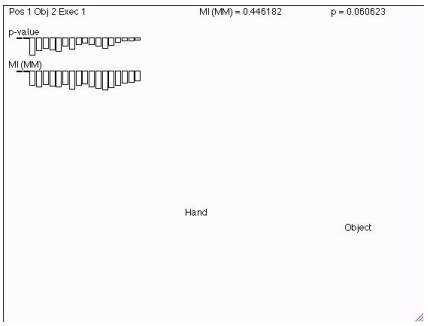
(c) Self-detection for Paddle Component

Fig. 3: Summary of the results for learning to play video games using mutual information.

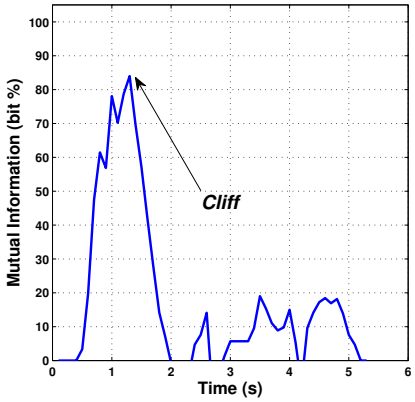
(a) Robot Pushing an Object onto a Ramp



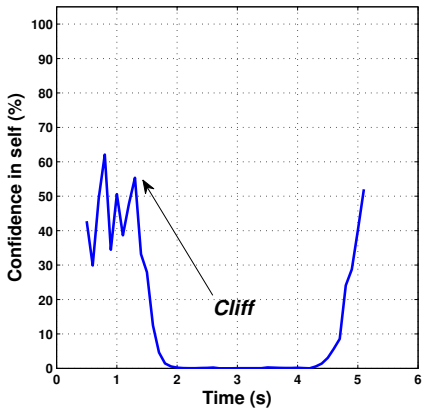
(b) Internal Representation Before a Cliff



(c) Internal Representation After a Cliff



(d) Cliff in Mutual Information



(e) Cliff in Confidence for Hypothesis $I > 0$

Fig. 4: Summary of the results for support affordance detection using mutual information.

information about the robot's hand. In the second experiment, grasping behavior led to confidence in categorization between self and object within 30 seconds. This is remarkable since the robot had no prior knowledge of either the block or of the self. The robot effectively categorized its perceptions using an entropy-based approach to classification. Finally, self-detection within the game environment experiment was consistently achieved within 20 seconds.In conclusion, using a single modality for exploration within a synthetic system was enough to allow that system to learn to categorize and to learn a sense of self/body-schema.

The reason that entropy-based learning works well in all of these situations supports the hypothesis that multiple periods of development could be guided by the same developmental function. Each stage of development is characterized by different levels of agency and intelligence, but is, nonetheless, created through similar processes. This model represents a possible explanation of the multi-faceted nature of human development through the use of a unified lens. This has implications for developmental robotics because similar dynamic periods of growth occur at many points within the human developmental process. This also has implications for philosophy since, it had been proposed long ago, that a single modality is all that is required for a system to self-organize. Finally, this has implications for Psychology because a synthetic system, using entropy as a learning function, expressed dynamic developmental patterns similar to humans developmental patterns. Nonetheless, future

work needs to be done to examine whether more stages of development can be modeled using this approach. Future work also needs to examine the optimality of this approach to other learning approaches.

VIII. Future work

This is an ongoing work in progress with the goal of developing a more unified approach to learning for use within developmental robotics. Presented here is a learning approach that unified categorical learning with self-detection learning. This content is intended circulation in the near future. Future work will focus upon perfecting the current approach and upon adding additional developmental periods into the framework that utilize this same learning approach in order to examine the effectiveness of this approach upon a variety of developmental periods. Of particular interest is the use of this approach within the auditory modality. Processing of sound within the human mind begins at birth. Inside the human mind, sound is first encoded into digital, neuronal signals by the inner ears hardware. Human development begins with the necessary structural-functional elements for auditory perception in place. Since the tools of the auditory system are arguably innate or, at least, change very little, an experiment interested in development will focus upon how a learning system could use such existing hardware to develop an understanding of its world. This understanding is thought to rely upon the categorizing auditory stimuli. The act of categorizing auditory stimuli is the first area of active processing that can be focused upon within the auditory human system. The theory of signal detection relates to the minimum difference between two sensory inputs in order for the correct change detection to occur. These changes can occur within several dimensions of a modality such as seeing, hearing or even within categorization itself. Within the auditory system there are thresholds for frequency, amplitude and duration [21]. There are absolute thresholds for the human auditory system. However, the theory of just noticeable difference suggests that there are various thresholds (limens) within human modalities that need to be overcome before change detection can occur at all [21]. In hearing, these thresholds occur for the frequency, amplitude and duration of sound. The minimum frequency of detection appears to have absolute thresholds based upon the architecture of the human ear with detection becoming poor beyond thresholds of 100 to 10,000 Hz [21]. The ability to detect sound at varying frequencies also relates to the intensity of the sound with the most detectable sounds ranging from 2000 to 5000 Hz [21]. The microphone apparatus used for the robot to detect sound will likely have a predisposition towards certain frequencies just as humans have. The minimum length for the detection of sound to occur depends upon the amplitude of the sound with most sounds (those louder than 40 dB) capable of being heard in durations longer than 5 ms [56]. In terms of relative threshold, the minimum detectable difference in time is about 10% of the signal's duration for those sounds lasting between 50 and 500 msec. Thus, the minimum difference in duration for any sound in this experiment should not be shorter than 5 ms [21][56]. Ultimately, the sensitivity of the algorithm will determine the just noticeable difference between the duration of two sounds; however, it is reasonable to state that no sounds shorter than 5 ms should be considered since comparison to human development is the goal of this experiment. The minimum amplitude of a sound depends upon other factors such as frequency and duration. The human ear detects changes in energy levels and the level of energy. The maximum sound pressure level is thought to be about 120 dB. It is thought that the minimal detectable change in energy will depend upon both the change in intensity and the change in duration. A louder tone will require less change in duration to be detectable than a quieter tone [21]. Up to about 200-300 ms, the relationship between duration and intensity appears to be logarithmic such that a change in intensity of about 10 dB is equal to a change of about 10 times in duration. It is proposed that this relationship breaks down at longer periods of time since the total operable window for energy detection within the ear is about 200 ms [21]. The minimum change in amplitude for signal detection will likely be different for the robot than for the humans. However, it is reasonable to require the minimum window for comparison to be 200 ms in order to create innate similarity between their synthetic architecture to the human architecture. Ultimately, the psychophysics of change detection will be an important principle to compare the similarity of the development of the robot to the development of humans. The very presence of this threshold will be enough to warrant similarities in learning. Absolute thresholds can be hard-coded without affecting the validity of the developmental cycle since these thresholds are thought to be innate and relate to the architecture and only restrict learning to various thresholds without effecting developmental processes.

Statement of Goals

The following statement gives a high-level summary of the overall progress for the three goals of this project:

- Goal I. Develop Mathematical apparatus for self-detection based on Information Theory.
  Achieved in full.
- Goal II. Show that reduction in uncertainty can guide skill acquisition.
  Future work.
- Goal III. Formulate a research question about phase transitions in robotic systems that learn and develop.
  Future work, but significant progress was made.

Statement of Success

No manuscripts were submitted for publication during the active phase of the project. However, the methods were applied in a number of different contexts and showed potential for generalization. Thus, the project is likely to achieve full success retroactively.

ACKNOWLEDGMENTS

The authors would like to thank Shane Griffith, Pavel Kazatsker, and Brian Russell for providing the data from their experiments.

REFERENCES

[1] R. Brooks, *The AI Journey: Future Challenges*, 2004, http://www.ai.rutgers.edu/aaai25/brooks.htm.
[2] P. McCorduck, *Machines Who Think*, 2nd ed. Peters, Ltd., 2004.
[3] J. Watson, "Detection of self: The perfect algorithm," in *Self-awareness in Animals and Humans: Developmental Perspectives*. Cambridge University Press, 1994, pp. 131–148.
[4] J. Piaget, *The origins of intelligence in children*. Intl. Univ. P., 1952.
[5] E. Hutchison, *Dimensions of human behavior. The changing life course*. Sage Publications, 2003.
[6] P. Hauf and G. Aschersleben, "Action-effect anticipation in infant action control," *Psych. Research*, vol. 72, no. 2, pp. 203–210, 2008.
[7] J. Kelso, *Dynamic patterns: the self-organization of brain and behavior*. MIT Press, 1995.
[8] A. Phillips and P. Robinson, "A quantitative model of sleep-wake dynamics based on the physiology of the brainstem ascending arousal system," *J Biol Rhythms*, vol. 22, pp. 167–179, April 2007.
[9] E. Thelen and L. Smith, *A Dynamic Systems Approach to the Development of Cognition and Action (Cognitive Psychology)*. MIT Press, 1996.
[10] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Vision Research*, vol. 49, no. 10, pp. 1295–1306, 2009.
[11] E. Tulving, "Episodic memory: From mind to brain," *Annual Review of Psychology, 53, 1-25.*, 2002.
[12] L. Barrett, "Are emotions natural kinds? perspectives on psychological science 1(1)," 2006.
[13] K. Miller, *Principles of everyday behavioral analysis*, 2006.
[14] B. F. Skinner, "Whatever happened to psychology as the science of behavior," *American Psychologist, 42. p. 780-86.*, 1987.
[15] B. Hergenhahn, *An introduction to the history of psychology (4th)*, 2001.
[16] K. J. Holyoak, *Relations in semantic memory: Still puzzling after all these years*, 2008.
[17] D. L. Medin, "Concepts and conceptual structure," *American Psychologist, 44, 1469-1481.*, 1989.
[18] G. Moskowitz, *Social cognition: Understanding self and others*, 2005.
[19] W. Crain, *Theories of development: Concepts and applications (2nd)*, 1985.
[20] S. Blackmore, *Consciousness: An introduction*, 2004.
[21] S. Gelfand, *Hearing: An introduction to psychological and physiological acoustics (3rd)*, 1998.
[22] Bargh, "The ecology of automaticity: Toward establishing the conditions needed to produce automatic processing effects," *The American Journal of Psychology 105. p. 181-99.*, 1992.
[23] T. Rogers. (2001) Type i and type ii errors - making mistakes in the justice system.
[24] P. Rochat, "Five levels of self-awareness as they unfold early in life," *Consciousness and cognition 12(2003) p. 717-31.*, 2003.
[25] J. Hawkins and S. Blakeslee, *On intelligence: How a new understanding of the brain will lead to the creation of truly intelligent machines*, 2004.
[26] D. Streeter and J. Raviv, "Research on advanced computer methods for biological data processing," 1965.
[27] E. Klingbeil, B. Carpenter, O. Russakovsky, and A. Ng, "Autonomous operation of novel elevators for robot navigation," in *Proc. of ICRA*, 2010, pp. 751–758.
[28] J. Miura, K. Iwase, and Y. Shirai, "Interactive teaching of a mobile robot," in *Proc. of ICRA*, 2005, pp. 3378–3383.
[29] K.-T. Song and T.-Z. Wu, "Visual servo control of a mobile manipulator using one-dimensional windows," in *Proc. of Industrial Electronics Society*, vol. 2, 1999, pp. 686–691.
[30] R. Katsuki, J. Ota, T. Yamura, T. Mizuta, T. Arai, T. Ueyama, and T. Nishiyama, "Handling of objects with marks by a robot," in *Proceedings of IROS 2003*, October 2003, pp. 130–135.
[31] T. Deyle, H. Nguyen, M. Reynolds, and C. Kemp, "RFID-guided robots for pervasive automation," *Pervasive Computing*, vol. 9, no. 2, pp. 37–45, 2010.
[32] H. Nguyen, T. Deyle, M. Reynolds, and C. Kemp, "PPS-tags: Physical, Perceptual and Semantic tags for autonomous mobile manipulation," in *Proc. of the IROS Workshop on Semantic Perception for Mobile Manipulation*, 2009.
[33] A. Thomaz, "Socially guided machine learning," Ph.D. dissertation, Massachusetts Institute of Technology, 2006.
[34] C. Breazeal and A. Thomaz, "Learning from human teachers with socially guided exploration," in *Proceedings of ICRA*, 2008, pp. 3539–3544.
[35] J. Lieberman, "Teaching a robot manipulation skills through demonstration," Master's thesis, Massachusetts Institute of Technology, 2004.
[36] V. Sukhoy and A. Stoytchev, "Learning to detect the functional components of doorbell buttons using active exploration and multimodal correlation," in *In Proceedings of the 2010 IEEE-RSJ Conference on Humanoid Robots*, Nashville, TN, 2010, pp. 572–579.
[37] D. Luebke, M. Harris, N. Govindaraju, A. Lefohn, M. Houston, J. Owens, M. Segal, M. Papakipos, and I. Buck, "GPGPU: general-purpose computation on graphics hardware," in *Proceedings of the 2006 ACM/IEEE conference on Supercomputing*, 2006.
[38] V. Podlozhnyuk, "Image convolution with CUDA," *NVIDIA Corporation white paper*, Jan 2007.
[39] S. Heymann, K. Müller, B. Smolic, A. Fröhlich, and T. Wiegand, "SIFT implementation and optimization for general-purpose GPU," in *WSCG*, 2007, pp. 317–322.
[40] T. Terriberry, L. French, and J. Helmsen, "GPU accelerating speeded-up robust features," 2008.
[41] E. Gibson, "Exploratory behavior in the development of perceiving, acting, and the acquiring of knowledge," *Annual review of psychology*, vol. 39, no. 1, pp. 1–42, 1988.
[42] P. Hauf, G. Aschersleben, and W. Prinz, "Baby do-baby see!: How action production influences action perception in infants," *Cognitive Development*, vol. 22, no. 1, pp. 16 – 32, 2007.
[43] C. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, July,October 1948.
[44] L. Paninski, "Estimation of entropy and mutual information," *Neural Comput.*, vol. 15, pp. 1191–1253, June 2003.
[45] G. Miller, "Note on the bias of information estimates," in *Information Theory in Psychology: Problems and Methods*, 1955, pp. 95–100.
[46] A. Gibbs and F. Su, "On choosing and bounding probability metrics," *International Statistical Review*, vol. 70, no. 3, pp. 419–435, 2002.
[47] P. Lévy, *Calcul des probabilités*. Gauthier-Villars, 1925.
[48] ——, "Théorie de l'Addition des Variables Aléatoires," in *Monographies des Probabilités*, E. Borel, Ed., vol. 1. Paris, France: Gauthier-Villars, pp. 17–345.
[49] K. Pearson, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *Philsophical Magazine Series 5*, vol. 50, pp. 157–175, July 1900.
[50] R. Fisher, "On the interpretation of χ^2 from contingency tables, and the calculation of *p*," *Journal of the Royal Statistical Society*, vol. 85, no. 1, pp. 87–94, 1922.
[51] ——, "The conditions under which χ^2 measures the discrepancy between observation and hypothesis," *Journal of the Royal Statistical Society*, vol. 87, pp. 442–450, 1924.
[52] H. Cramér, *Mathematical methods of statistics*. Princeton University Press, 1946.

[53] A. Stoytchev, "Robot tool behavior: A developmental approach to autonomous tool use," Ph.D. dissertation, College of Computing, Georgia Institute of Technology, August 2007.
[54] S. Griffith, "Learning to predict the controllability of containers and their contents," 2011, a course project proposal for CPR E 585X.
[55] P. Kazatsker, "CPRE 585 X Project Proposal," 2011, a course project proposal for CPR E 585X.
[56] R. Irwin, L. Hinchcliff, and S. Kemp, "Temporal acuity in normal and hearing-imparied listeners," *Audiology*, vol. 20, pp. 234–248, 1981.