Overview of The Theory and Practice of Induction by Alignment

CJ McCartney

April 2020

Abstract

Induction is the discovery of models given samples.

This paper demonstrates formally from first principles that there exists an optimally likely model for any sample, given certain general assumptions. Also, there exists a type of encoding, parameterised by the model, that compresses the sample. Further, if the model has certain entropy properties then it is insensitive to small changes. In this case, approximations to the model remain well-fitted to the sample. That is, accurate classification is practicable for some samples.

Then the paper derives directly from theory a practicable unsupervised machine learning algorithm that optimises the likelihood of the model by maximising the alignment of the model variables. Alignment is a statistic which measures the law-likeness or the degree of dependency between variables. It is similar to mutual entropy but is a better measure for small samples. If the sample variables are not independent then the resultant models are well-fitted. Furthermore, the models are structures that can be analysed because they consist of trees of context-contingent sub-models that are built layer by layer upwards from the substrate variables. In the top layers the variables tend to be diagonalised or equational. In this way, the model variables are meaningful in the problem domain.

If there exist causal alignments between the induced model variables and a label variable, then a semi-supervised sub-model can be obtained by minimising the conditional entropy. Similar to a Bayesian network, this sub-model can then make predictions of the label.

The paper shows that this semi-supervised method is related to the supervised method of optimising artificial neural networks by least-squares gradient-descent. That is, some gradient-descent parameterisations satisfy the entropy properties required to obtain likely and well-fitted neural nets.

1 Preface

This paper consists of the 'Overview' section extracted from the paper 'The Theory and Practice of Induction by Alignment'. The 'Overview' section covers the important points of the theory and some interesting parts of the practice. The overview also has a summary of the set-theoretic notation used throughout.

Terms in italics have a mathematical definition to avoid ambiguity. So '*independent*' is a well defined property, whereas 'independent' has its dictionary definition.

For further discussion see `https://greenlake.co.uk/`.

2 Overview

This section provides an overview of the main points of the paper. Detailed explanations are excluded for brevity. The overview is presented as a series of assertions of fact, but only some are proven and many are conjectured, especially statements regarding correlations. In some cases, however, there are multiple strands of evidence that corroborate a conjecture. This is particularly true for the conjectures regarding the general *induction* of *models* given *samples*. Given a set of *induction* assumptions these conjectures relate (i) the maximisation of the *likelihood* of a *sample*, and also the minimisation of the *likelihood's sensitivity* to *model* and *distribution*, to (ii) properties such as *encoding space*, *entropy* and *alignment*. The different sets of *induction* assumptions can be categorised in various complementary ways: (a) *classical induction* versus *aligned induction*, (b) *law-like* conditional *draws* of *samples* from *distributions* versus the *compression* of *encodings* of *samples* by *model*, (c) simple *transform models* versus *layered, contingent models*, and (d) intractable theoretical *induction* assumptions versus tractable and practicable *induction* assumptions. The existence of working implementations of practicable *induction* such as *artificial neural networks* and *alignment inducers* provides concrete support to the theory.

2.1 Notation

The notation is briefly summarised below. The appendices contain further details.

The notation used throughout this discussion is conventional set-theoretic with some additions. Sets are often defined using set-builder notation, for example $Z = \{\mathrm{f}(x) : x \in X,\ \mathrm{p}(x)\}$ where $\mathrm{f}(x)$ is a function, X is another set and $\mathrm{p}(x)$ is a predicate.

Tuples, or lists, can be defined similarly where the order is not important, for example, $\sum(\mathrm{f}(x) : x \in X,\ \mathrm{p}(x))$.

The powerset function is defined as $\mathrm{P}(A) := \{X : X \subseteq A\}$.

The partition function B is the set of all partitions of an argument set. A partition is a set of non-empty disjoint subsets, called components, which union to equal the argument, $\forall P \in \mathrm{B}(A)\ \forall C \in P\ (C \neq \emptyset)$, $\forall P \in \mathrm{B}(A)\ \forall C, D \in P\ (C \neq D \implies C \cap D = \emptyset)$ and $\forall P \in \mathrm{B}(A)\ (\bigcup P = A)$.

A relation $A \in \mathrm{P}(\mathcal{X} \times \mathcal{Y})$ between the set $\mathcal{X}$ and the set $\mathcal{Y}$ is a set of pairs, $\forall (x, y) \in A\ (x \in \mathcal{X}\ \wedge\ y \in \mathcal{Y})$. The domain of a relation is $\mathrm{dom}(A) := \{x : (x, y) \in A\}$ and the range is $\mathrm{ran}(A) := \{y : (x, y) \in A\}$.

Functions are special cases of relations such that each element of the domain appears exactly once. Functions can be finite or infinite. For example, $\{(1, 2), (2, 4)\} \subset \{(x, 2x) : x \in \mathbf{R}\}$. The powerset of functional relations between sets is denoted $\to$. For example, $\{(x, 2x) : x \in \mathbf{R}\} \in \mathbf{R} \to \mathbf{R}$. The application of the function $F \in \mathcal{X} \to \mathcal{Y}$ to an argument $x \in \mathcal{X}$ is denoted by $F(x) \in \mathcal{Y}$ or $F_x \in \mathcal{Y}$. Functions $F \in \mathcal{X} \to \mathcal{Y}$ and $G \in \mathcal{Y} \to \mathcal{Z}$ can be composed $G \circ F \in \mathcal{X} \to \mathcal{Z}$. The inverse of a function, $\mathrm{inverse} \in (\mathcal{X} \to \mathcal{Y}) \to (\mathcal{Y} \to \mathrm{P}(\mathcal{X}))$, is defined $\mathrm{inverse}(F) := \{(y, \{x : (x, z) \in F,\ z = y\}) : y \in \mathrm{ran}(F)\}$, and is sometimes denoted F^{-1}. The range of the inverse is a partition of the domain, $\mathrm{ran}(F^{-1}) \in \mathrm{B}(\mathrm{dom}(F))$.

Functions may be recursive. Algorithms are represented as recursive functions.

The powerset of bi-directional relations, or one-to-one functions, is denoted $\leftrightarrow$. The cardinality of the domain of a bi-directional function equals the range, $F \in \mathrm{dom}(F) \leftrightarrow \mathrm{ran}(F) \implies |\mathrm{dom}(F)| = |\mathrm{ran}(F)|$.

Total functions are denoted with a colon. For example, the left total function $F \in X :\to Y$ requires that $\mathrm{dom}(F) = X$ but only that $\mathrm{ran}(F) \subseteq Y$.

An order D on some set X is a choice of the enumerations, $D \in X :\leftrightarrow: \{1 \dots |X|\}$. Given the order, any subset $Y \subseteq X$ can be enumerated. Define $\mathrm{order}(D, Y) \in Y :\leftrightarrow: \{1 \dots |Y|\}$ such that $\forall a, b \in Y\ (D_a \leq D_b \implies \mathrm{order}(D, Y)(a) \leq \mathrm{order}(D, Y)(b))$.

The set of natural numbers $\mathbf{N}$ is taken to include 0. The set $\mathbf{N}_{>0}$ excludes 0. The *space* of a non-zero natural number is the natural logarithm, $\mathrm{space}(n) := \ln n$. The set of rational numbers is denoted $\mathbf{Q}$. The set of log-rational numbers is denoted $\ln \mathbf{Q}_{>0} = \{\ln q : q \in \mathbf{Q}_{>0}\}$. The set of real numbers is denoted $\mathbf{R}$.

The factorial of a non-zero natural number $n \in \mathbf{N}_{>0}$ is written $n! = \prod\{1 \dots n\}$.

The unit-translated gamma function is the real function that corresponds to the factorial function. It is defined $(\Gamma_!) \in \mathbf{R} \to \mathbf{R}$ as $\Gamma_! x = \Gamma(x+1)$ which is such that $\forall n \in \mathbf{N}_{>0}\ (\Gamma_! n = \Gamma(n+1) = n!)$.

Given a relation $A \subset \mathcal{X} \times \mathcal{Y}$ such that an order operator is defined on the range, $\mathcal{Y}$, the max function returns the maximum subset, $\max \in \mathrm{P}(\mathcal{X} \times \mathcal{Y}) \to (\mathcal{X} \to \mathcal{Y})$

$$\max(A) := \{(x, y) : (x, y) \in A,\ \forall (r, s) \in A\ (s \leq y)\}$$

For convenience define the functions $\mathrm{maxd}(A) := \mathrm{dom}(\max(A))$ and $\mathrm{maxr}(A) := m$, where $\{m\} = \mathrm{ran}(\max(A))$. The corresponding functions for minimum, min, mind and minr, are similarly defined.

Given a relation $A \subset \mathcal{X} \times \mathcal{Y}$ such that the arithmetic operators are defined on the range, $\mathcal{Y}$, the sum function is defined $\mathrm{sum}(A) := \sum(y : (x, y) \in A)$. The relation can be normalised, $\mathrm{normalise}(A) := \{(x, y/\mathrm{sum}(A)) : (x, y) \in A\}$. Define notation $\hat{A} := \mathrm{normalise}(A)$. A normalised relation is such that its sum is one, $\mathrm{sum}(\hat{A}) = 1$.

The set of *probability functions* $\mathcal{P}$ is the set of rational valued functions such that the values are bounded $[0, 1]$ and sum to 1, $\mathcal{P} \subset \mathcal{X} \to \mathbf{Q}_{[0,1]}$ and $\forall P \in \mathcal{P}\ (\mathrm{sum}(P) = 1)$. The normalisation of a positive rational valued function $F \in \mathcal{X} \to \mathbf{Q}_{\geq 0}$ is a *probability function*, $\hat{F} \in \mathcal{P}$.

The *entropy* of positive rational valued functions, $\mathrm{entropy} \in (\mathcal{X} \to \mathbf{Q}_{\geq 0}) \to \mathbf{Q}_{\geq 0} \ln \mathbf{Q}_{>0}$, is defined as $\mathrm{entropy}(N) := -\sum(\hat{N}_x \ln \hat{N}_x : x \in \mathrm{dom}(N),\ N_x >$

0). The *entropy* of a singleton is zero, $\text{entropy}(\{(\cdot,1)\}) = 0$. *Entropy* is maximised in uniform functions as the cardinality tends to infinity, $\text{entropy}(X \times \{1/|X|\}) = \ln |X|$.

Given some finite function $F \in \mathcal{X} \to \mathcal{Y}$, where $0 < |F| < \infty$, a *probability function* may be constructed from its distribution, $\{(y, |X|) : (y, X) \in F^{-1}\}^{\wedge} \in (\mathcal{Y} \to \mathbf{Q}_{\geq 0}) \cap \mathcal{P}$. The *probability function* of an arbitrarily chosen finite function is likely to have high *entropy*.

A *probability function* $P(z) \in (X :\to \mathbf{Q}_{\geq 0}) \cap \mathcal{P}$, parameterised by some parameter $z \in Z = \text{dom}(P)$, has a corresponding *likelihood function* $L(x) \in Z :\to \mathbf{Q}_{\geq 0}$, parameterised by coordinate $x \in X$, such that $L(x)(z) = P(z)(x)$. The *maximum likelihood estimate* $\tilde{z}$ of the parameter, z, at coordinate $x \in X$ is the mode of the *likelihood function*,

$$\begin{aligned}
\{\tilde{z}\} &= \text{maxd}(L(x)) \\
&= \text{maxd}(\{(z, P(z)(x)) : z \in Z\}) \\
&= \{z : z \in Z,\ \forall z' \in Z\ (P(z)(x) \geq P(z')(x))\}
\end{aligned}$$

A list is a object valued function of the natural numbers $\mathcal{L}(\mathcal{X}) \subset \mathbf{N} \to \mathcal{X}$, such that $\forall L \in \mathcal{L}(\mathcal{X})\ (L \neq \emptyset \implies \text{dom}(L) = \{1 \ldots |L|\})$. Two lists $L, M \in \mathcal{L}(\mathcal{X})$ may be concatenated, $\text{concat}(L, M) := L \cup \{(|L| + i, x) : (i, x) \in M\}$.

A tree is recursively defined as a tree valued function of objects, $\text{trees}(\mathcal{X}) = \mathcal{X} \to \text{trees}(\mathcal{X})$. The nodes of the tree $T \in \text{trees}(\mathcal{X})$ are $\text{nodes}(T) := T \cup \bigcup\{\text{nodes}(R) : (x, R) \in T\}$, and the elements are $\text{elements}(T) := \text{dom}(\text{nodes}(T))$. The paths of a tree $\text{paths}(T) \subset \mathcal{L}(\mathcal{X})$ is a set of lists. Given a set of lists $Q \subset \mathcal{L}(\mathcal{X})$ a tree can be constructed $\text{tree}(Q) \in \text{trees}(\mathcal{X})$.

2.2 Maximum Entropy

Let $X \subset \mathcal{X}$ be a finite set of micro-states, $0 < |X| < \infty$. Consider a system of n distinguishable particles, each in a micro-state. The set of states of the system is the set of micro-state functions of particle identifier, $\{1 \ldots n\} :\to X$. The cardinality of the set of states is $|X|^n$.

Each state implies a distribution of particles over micro-states,

$$I = \big\{(R, \{(x, |C|) : (x, C) \in R^{-1}\}) : R \in \{1 \ldots n\} :\to X\big\}$$

That is, a state $R \in \{1 \ldots n\} :\to X$ has a particle distribution $I(R) \in X \to \{1 \ldots n\}$ such that $\text{sum}(I(R)) = n$.

The cardinality of states for each particle distribution, $I(R)$, is the multinomial coefficient,

$$\begin{aligned} W &= \{(N, |D|) : (N, D) \in I^{-1}\} \\ &= \{(N, \frac{n!}{\prod_{(x,\cdot) \in N} N_x!}) : (N, \cdot) \in I^{-1}\} \end{aligned}$$

That is, there are $W(I(R))$ states that have the same particle distribution, $I(R)$, as state R. The normalisation of the state distribution over particle distributions is a *probability function*, $\hat{W} \in ((X \to \{1 \ldots n\}) \to \mathbf{Q}_{>0}) \cap \mathcal{P}$.

In the case where the number of particles is large, $n \gg \ln n$, the logarithm of the multinomial coefficient of a particle distribution $N \in X \to \{1 \ldots n\}$ approximates to the scaled *entropy*,

$$\ln \frac{n!}{\prod_{(x,\cdot) \in N} N_x!} \approx n \times \text{entropy}(N)$$

so the probability of the particle distribution varies with its *entropy*, $\hat{W}(N) \sim \text{entropy}(N)$.

The least probable particle distributions are singletons,

$$\text{mind}(W) = \big\{\{(x, n)\} : x \in X\big\}$$

because they have only one state, $\forall x \in X\ (W(\{(x, n)\}) = 1)$. The *entropy* of a singleton distribution is zero, $\text{entropy}(\{(x, n)\}) = 0$.

In the case where the number of particles per micro-state is integral, $n/|X| \in \mathbf{N}_{>0}$, the modal particle distribution is the uniform distribution,

$$\text{maxd}(W) = \big\{\{(x, n/|X|) : x \in X\}\big\}$$

The *entropy* of the uniform distribution is maximised, $\text{entropy}(\{(x, n/|X|) : x \in X\}) = \ln |X|$.

The normalisation of a particle distribution $N \in X \to \{1 \ldots n\}$ is a micro-state *probability function*, $\hat{N} \in (X \to \mathbf{Q}_{\geq 0}) \cap \mathcal{P}$, which is independent of the number of particles, $\text{sum}(\hat{N}) = 1$.

So in the case where a problem domain is parameterised by an *unknown* micro-state *probability function* otherwise arbitrarily chosen from a *known* subset $Q \subseteq (X \to \mathbf{Q}_{\geq 0}) \cap \mathcal{P}$, where the number of particles is *known* to be large, the *maximum likelihood estimate* $\tilde{P} \in Q$ is the *probability function* with the greatest *entropy*, $\forall P \in Q\ (\mathrm{entropy}(\tilde{P}) \geq \mathrm{entropy}(P))$ or $\tilde{P} \in \mathrm{maxd}(\{(P, \mathrm{entropy}(P)) : P \in Q\})$.

2.3 Histograms

2.3.1 States, histories and histograms

The set of all *variables* is $\mathcal{V}$. The set of all *values* is $\mathcal{W}$. A *system* $U \in \mathcal{V} \to \mathrm{P}(\mathcal{W})$ is a functional mapping between *variables* and non-empty sets of *values*, $\forall (v, W) \in U\ (|W| > 0)$. The *variables* of a *system* is the domain, $\mathrm{vars}(U) := \mathrm{dom}(U)$.

In a *system* of finite *variables*, $\forall v \in \mathrm{vars}(U)\ (|U_v| < \infty)$, each *variable* has a set of discrete *values*. The *values* need not be ordered. The *valency* of a *variable* v is the cardinality of its *values*, $|U_v|$. The *volume* of a set of *variables* in a *system* $V \subseteq \mathrm{vars}(U)$ is the product of the *valencies*, $\prod_{v \in V} |U_v| \geq 1$.

The set of *states* is the set of *value* valued functions of *variable*, $\mathcal{S} = \mathcal{V} \to \mathcal{W}$. The *variables* of a *state* $S \in \mathcal{S}$ is the function domain, $\mathrm{vars}(S) := \mathrm{dom}(S)$.

The *state*, S, is in a *system* U if (i) the *variables* of the *state* are *variables* of the *system*, $\mathrm{vars}(S) \subseteq \mathrm{vars}(U)$, and (ii) the *value* of each *variable* in the *state* is in the *system*, $\forall v \in \mathrm{vars}(S)\ (S_v \in U_v)$.

Given a set of *variables* in a *system* $V \subseteq \mathrm{vars}(U)$, the *cartesian* set of all possible *states* is $\prod_{v \in V}(\{v\} \times U_v)$, which has cardinality equal to the *volume* $\prod_{v \in V} |U_v|$.

The *variables* $V = \mathrm{vars}(S)$ of a *state* S may be *reduced* to a given subset $K \subseteq V$ by taking the subset of the *variable-value* pairs,

$$S \;\%\; K := \{(v, u) : (v, u) \in S,\ v \in K\}$$

A set of *states* $Q \subset \mathcal{S}$ in the same *variables* $\forall S \in Q\ (\mathrm{vars}(S) = V)$ may be *split* into a subset of its *variables* $K \subseteq V$ and the complement $V \setminus K$,

$$\mathrm{split}(K, Q) \quad = \quad \{(S \;\%\; K,\ S \;\%\; (V \setminus K)) : S \in Q\}$$

Two *states* $S, T \in \mathcal{S}$ are said to *join* if their union is also a *state*, $S \cup T \in \mathcal{S}$. That is, a *join* is functional,

$$\begin{aligned} S \cup T \in \mathcal{S} &\iff |\mathrm{vars}(S) \cup \mathrm{vars}(T)| = |S \cup T| \\ &\iff \forall v \in \mathrm{vars}(S) \cap \mathrm{vars}(T)\ (S_v = T_v) \end{aligned}$$

States in disjoint *variables* always *join*, $\forall S, T \in \mathcal{S}\ (\mathrm{vars}(S) \cap \mathrm{vars}(T) = \emptyset \implies S \cup T \in \mathcal{S})$. *States* in the same *variables* only *join* if they are equal, $\forall S, T \in \mathcal{S}\ (\mathrm{vars}(S) = \mathrm{vars}(T) \implies (S \cup T \in \mathcal{S} \iff S = T))$.

The set of *event identifiers* is the universal set $\mathcal{X}$. An *event* (x, S) is a pair of an *event identifier* and a *state*, $(x, S) \in \mathcal{X} \times \mathcal{S}$. A *history* H is a *state* valued function of *event identifiers*, $H \in \mathcal{X} \to \mathcal{S}$, such that all of the *states* of its *events* share the same set of *variables*, $\forall (x, S), (y, T) \in H\ (\mathrm{vars}(S) = \mathrm{vars}(T))$. The set of *histories* is denoted $\mathcal{H} \subset \mathcal{X} \to \mathcal{S}$.

The set of *variables* of a *history* is the set of the *variables* of any of the *events* of the *history*, $\mathrm{vars}(H) = \mathrm{vars}(S)$ where $(x, S) \in H$.

The *event identifiers* of a *history* need not be ordered, so a *history* is not necessarily sequential or chronological.

The inverse of a *history*, H^{-1}, is called the *classification*. So a *classification* is an *event identifier* set valued function of *state*, $H^{-1} \in \mathcal{S} \to \mathrm{P}(\mathcal{X})$. The *event identifier* components are non-empty, $\forall (S, X) \in H^{-1}\ (X \neq \emptyset)$.

The *reduction* of a *history* is the *reduction* of its *events*, $H\%V := \{(x, S\%V) : (x, S) \in H\}$.

The *addition* operation of *histories* is defined as the disjoint union of the *events* if both *histories* have the same *variables*,

$$H_1 + H_2 \ := \ \{((x, \cdot), S) : (x, S) \in H_1\} \ \cup \ \{((\cdot, y), T) : (y, T) \in H_2\}$$

where $\mathrm{vars}(H_1) = \mathrm{vars}(H_2)$. The *size* of the *sum* equals the sum of the *sizes*, $|H_1 + H_2| = |H_1| + |H_2|$.

The *multiplication* operation of *histories* is defined as the product of the *events* where the *states join*,

$$\begin{aligned} H_1 * H_2 \ := \ \{((x, y), S \cup T) : (x, S) \in H_1,\ (y, T) \in H_2, \\ \forall v \in \mathrm{vars}(S) \cap \mathrm{vars}(T)\ (S_v = T_v)\} \end{aligned}$$

The *size* of the *product* equals the product of the *sizes* if the *variables* are disjoint, $\mathrm{vars}(H_1) \cap \mathrm{vars}(H_2) = \emptyset \implies |H_1 * H_2| = |H_1| \times |H_2|$. The *variables* of the *product* is the union of the *variables* if the *size* is non-zero, $H_1 * H_2 \neq \emptyset \implies \mathrm{vars}(H_1 * H_2) = \mathrm{vars}(H_1) \cup \mathrm{vars}(H_2)$.

The set of all *histograms* $\mathcal{A}$ is a subset of the positive rational valued functions of *states*, $\mathcal{A} \subset \mathcal{S} \to \mathbf{Q}_{\geq 0}$, such that each *state* of each *histogram* has the same set of *variables*, $\forall A \in \mathcal{A}\ \forall S, T \in \mathrm{dom}(A)\ (\mathrm{vars}(S) = \mathrm{vars}(T))$.

The set of *variables* of a *histogram* $A \in \mathcal{A}$ is the set of the *variables* of any of the elements of the *histogram*, $\mathrm{vars}(A) = \mathrm{vars}(S)$ where $(S, q) \in A$. The *dimension* of a *histogram* is the cardinality of its *variables*, $|\mathrm{vars}(A)|$. The *counts* of a *histogram* is the range. The *states* of a *histogram* is the domain. Define the shorthand $A^{\mathrm{S}} := \mathrm{dom}(A)$. The *size* of a *histogram* is the sum of the *counts*, $\mathrm{size}(A) := \mathrm{sum}(A)$. The *size* is always positive, $\mathrm{size}(A) \geq 0$. If the *size* is non-zero the normalised *histogram* has a *size* of one, $\mathrm{size}(A) > 0 \implies \mathrm{size}(\hat{A}) = 1$. In this case the normalised *histogram* is a *probability function*, $\mathrm{size}(A) > 0 \implies \hat{A} \in \mathcal{P}$.

The *volume* of a *histogram* A of *variables* V in a *system* U is the *volume* of the *variables*, $\prod_{v \in V} |U_v|$.

A *histogram* with no *variables* is called a *scalar*. The *scalar* of *size* z is $\{(\emptyset, z)\}$. Define $\mathrm{scalar}(z) := \{(\emptyset, z)\}$. A *singleton* is a *histogram* with only one *state*, $\{(S, z)\}$. A *uniform histogram* A has unique non-zero *count*, $|\{c : (S, c) \in A,\ c > 0\}| = 1$.

The set of *integral histograms* is the subset of *histograms* which have integal *counts* $\mathcal{A}_{\mathrm{i}} = \mathcal{A} \cap (\mathcal{S} \to \mathbf{N})$. A *unit histogram* is a special case of an *integral histogram* in which all its *counts* equal one, $\mathrm{ran}(A) = \{1\}$. The *size* of a *unit histogram* equals its cardinality, $\mathrm{size}(A) = |A|$. A set of *states* $Q \subset \mathcal{S}$ in the same *variables* may be promoted to a *unit histogram*, $Q^{\mathrm{U}} := Q \times \{1\} \in \mathcal{A}_{\mathrm{i}}$.

The *unit effective histogram* of a *histogram* is the *unit histogram* of the *states* where the *count* is non-zero. Define the shorthand $A^{\mathrm{F}} := \{(S, 1) : (S, c) \in A,\ c > 0\} \in \mathcal{A}_{\mathrm{i}}$.

Given a *system* U define the *cartesian histogram* of the set of *variables* V as $V^{\mathrm{C}} := \left(\prod_{v \in V} (\{v\} \times U_v)\right) \times \{1\} \in \mathcal{A}_{\mathrm{i}}$. The *size* of the *cartesian histogram* equals its cardinality which is the *volume* of the *variables*, $\mathrm{size}(V^{\mathrm{C}}) = |V^{\mathrm{C}}| =$

$\prod_{v \in V} |U_v|$. The *unit effective histogram* is a subset of the *cartesian histogram* of its *variables*, $A^{\mathrm{F}} \subseteq V^{\mathrm{C}}$, where $V = \mathrm{vars}(A)$. A *complete histogram* has the *cartesian* set of *states*, $A^{\mathrm{S}} = V^{\mathrm{CS}}$.

A *partition* P is a partition of the *cartesian states*, $P \in \mathrm{B}(V^{\mathrm{CS}})$. The *partition* is a set of disjoint *components*, $\forall C, D \in P\ (C \neq D \implies C \cap D = \emptyset)$, that union to equal the *cartesian states*, $\bigcup P = V^{\mathrm{CS}}$. The *unary partition* is $\{V^{\mathrm{CS}}\}$. The *self partition* is $V^{\mathrm{CS}\{\}} = \{\{S\} : S \in V^{\mathrm{CS}}\}$. A *partition variable* $P \in \mathrm{vars}(U)$ in a *system* U is such that its set of *values* equals its set of *components*, $U_P = P$. So the *valency* of a *partition variable* is the cardinality of the *components*, $|U_P| = |P|$.

A *regular histogram* A of *variables* V in *system* U has unique *valency* of its *variables*, $|\{|U_v| : v \in V\}| = 1$. The *volume* of a *regular histogram* is $d^n = |V^{\mathrm{C}}| = \prod_{v \in V} |U_v|$, where *valency* d is such that $\{d\} = \{|U_v| : v \in V\}$ and *dimension* $n = |V|$.

The *counts* of the *integral histogram* $A \in \mathcal{A}_{\mathrm{i}}$ of a *history* $H \in \mathcal{H}$ are the cardinalities of the *event identifier* components of its *classification*, $A = \mathrm{histogram}(H)$ where $\mathrm{histogram}(H) := \{(S, |X|) : (S, X) \in H^{-1}\}$. In this case the *histogram* is a distribution of *events* over *states*. If the *history* is bijective, $H \in \mathcal{X} \leftrightarrow \mathcal{S}$, then its *histogram* is a *unit histogram*, $A = \mathrm{ran}(H) \times \{1\}$.

A *sub-histogram* A of a *histogram* B is such that the *effective states* of A are a subset of the *effective states* of B and the *counts* of A are less than or equal to those of B, $A \leq B := A^{\mathrm{FS}} \subseteq B^{\mathrm{FS}} \wedge \forall S \in A^{\mathrm{FS}}\ (A_S \leq B_S)$. The *histogram* of a *sub-history* $G \subseteq H$ is a *sub-histogram*, $\mathrm{histogram}(G) \leq \mathrm{histogram}(H)$.

The *reduction* of a *histogram* is the *reduction* of its *states*, adding the *counts* where two different *states reduce* to the same *state*,

$$A\%V := \{(R, \sum(c : (T,c) \in A,\ T \supseteq R)) : R \in \{S\%V : S \in A^{\mathrm{S}}\}\}$$

Reduction leaves the *size* of a *histogram* unchanged, $\mathrm{size}(A\%V) = \mathrm{size}(A)$, but the number of *states* may be fewer, $|(A\%V)^{\mathrm{S}}| \leq |A^{\mathrm{S}}|$. The *reduction* to the empty set is a *scalar*, $A\%\emptyset = \{(\emptyset, z)\}$, where $z = \mathrm{size}(A)$. The *histogram* of a *reduction* of a *history* equals the *reduction* of the *histogram* of the *history*,

$$\mathrm{histogram}(H\ \%\ V) = \mathrm{histogram}(H)\ \%\ V$$

The *addition* of *histograms* A and B is defined,

$$\begin{aligned}&A + B := \\ &\quad \{(S,c) : (S,c) \in A,\ S \notin B^{\mathrm{S}}\} \cup \\ &\quad \{(S,c+d) : (S,c) \in A,\ (T,d) \in B,\ S = T\} \cup \\ &\quad \{(T,d) : (T,d) \in B,\ T \notin A^{\mathrm{S}}\}\end{aligned}$$

where $\mathrm{vars}(A) = \mathrm{vars}(B)$. The *sizes* add, $\mathrm{size}(A + B) = \mathrm{size}(A) + \mathrm{size}(B)$. The *histogram* of an *addition* of *histories* equals the *addition* of the *histograms* of the *histories*,

$$\mathrm{histogram}(H_1 + H_2) = \mathrm{histogram}(H_1) + \mathrm{histogram}(H_2)$$

The *multiplication* of *histograms* A and B is the product of the *counts* where the *states join*,

$$A*B := \{(S \cup T, cd) : (S,c) \in A,\ (T,d) \in B,\ \forall v \in \mathrm{vars}(S) \cap \mathrm{vars}(T)\ (S_v = T_v)\}$$

If the *variables* are disjoint, the *sizes* multiply, $\mathrm{vars}(A) \cap \mathrm{vars}(B) = \emptyset \implies \mathrm{size}(A * B) = \mathrm{size}(A) \times \mathrm{size}(B)$. *Multiplication* by a *scalar* scales the *size*, $\mathrm{size}(\mathrm{scalar}(z)*A) = z \times \mathrm{size}(A)$. The *histogram* of a *multiplication* of *histories* equals the *multiplication* of the *histograms* of the *histories*,

$$\mathrm{histogram}(H_1 * H_2) = \mathrm{histogram}(H_1) * \mathrm{histogram}(H_2)$$

The *reciprocal* of a *histogram* is $1/A := \{(S, 1/c) : (S,c) \in A,\ c > 0\}$. Define *histogram division* as $B/A := B * (1/A)$.

A *histogram* A is *causal* in a subset of its *variables* $K \subset V$ if the *reduction* of the *effective states* to the subset, K, is functionally related to the *reduction* to the complement, $V \setminus K$,

$$\{(S \ \%\ K,\ S \ \%\ (V \setminus K)) : S \in A^{\mathrm{FS}}\} \ \in \ K^{\mathrm{CS}} \to (V \setminus K)^{\mathrm{CS}}$$

or

$$\mathrm{split}(K, A^{\mathrm{FS}}) \ \in \ K^{\mathrm{CS}} \to (V \setminus K)^{\mathrm{CS}}$$

A *histogram* A is *diagonalised* if no pair of *effective states* shares any *value*, $\forall S,T \in A^{\mathrm{FS}}\ (S \neq T \implies S \cap T = \emptyset)$. A *diagonalised histogram* A is *fully diagonalised* if its *effective* cardinality equals the minimum *valency* of its *variables*, $|A^{\mathrm{F}}| = \mathrm{minr}(\{(v, |U_v|) : v \in V\})$. The cardinality of the *effective states* of a *fully diagonalised regular histogram* is the *valency*, $|A^{\mathrm{F}}| = d$, where

$\{d\} = \{|U_v| : v \in V\}$. In a *diagonalised histogram* the *causality* is bijective or equational,

$$\forall u, w \in V \ (\{(S\%\{u\}, S\%\{w\}) : S \in A^{\mathrm{FS}}\} \ \in \ \{u\}^{\mathrm{CS}} \leftrightarrow \{w\}^{\mathrm{CS}})$$

Given some *slice state* $R \in K^{\mathrm{CS}}$, where $K \subset V$ and $V = \mathrm{vars}(A)$, the *slice histogram*, $A * \{R\}^{\mathrm{U}} \subset A$, is said to be *contingent* on the *incident slice state*. For example, if the *slice histogram* is *diagonalised*, $\mathrm{diagonal}(A * \{R\}^{\mathrm{U}} \ \% \ (V \setminus K))$, then the *histogram*, A, is said to be *contingently diagonalised*.

The *perimeters* of a *histogram* $A \in \mathcal{A}$ is the set of its *reductions* to each of its *variables*, $\{A\%\{w\} : w \in V\}$, where $V = \mathrm{vars}(A)$. The *independent* of a *histogram* is the product of the normalised *perimeters* scaled to the *size*,

$$A^{\mathrm{X}} := Z * \prod_{w \in V} \hat{A}\%\{w\}$$

where $z = \mathrm{size}(A)$ and $Z = \mathrm{scalar}(z) = A\%\emptyset$. The *independent* of a *histogram* is such that (i) the *states* are a superset, $A^{\mathrm{XS}} \supseteq A^{\mathrm{S}}$, (ii) the *size* is unchanged, $\mathrm{size}(A^{\mathrm{X}}) = \mathrm{size}(A)$, and (iii) the *variables* are unchanged, $\mathrm{vars}(A^{\mathrm{X}}) = \mathrm{vars}(A)$. A *histogram* is said to be *independent* if it equals its *independent*, $A = A^{\mathrm{X}}$. The *independent* of an *independent histogram* is the *independent*, $A^{\mathrm{XX}} = A^{\mathrm{X}}$. The scaled *product* of (i) any *reduction* of a normalised *independent histogram* to any subset of its *variables* $K \subseteq V$, and (ii) the *reduction* to the complement, $V \setminus K$, is the *independent*, $Z * (\hat{A}^{\mathrm{X}} \ \% \ K) * (\hat{A}^{\mathrm{X}} \ \% \ (V \setminus K)) = A^{\mathrm{X}}$.

Scalar histograms are *independent*, $\{(\emptyset, z)\} = \{(\emptyset, z)\}^{\mathrm{X}}$. *Singleton histograms*, $|A^{\mathrm{F}}| = 1$, are *independent*, $\{(S, z)\} = \{(S, z)\}^{\mathrm{X}}$. If the *histogram* is *mono-variate*, $|V| = 1$, then it is *independent* $A = A\%\{w\} = A^{\mathrm{X}}$ where $\{w\} = V$. *Uniform-cartesian histograms*, which are *scalar* multiples of the *cartesian*, $A = V_z^{\mathrm{C}}$ where $V_z^{\mathrm{C}} = \mathrm{scalar}(z/v) * V^{\mathrm{C}}$, $z = \mathrm{size}(A)$ and $v = |V^{\mathrm{C}}|$, are *independent*, $V_z^{\mathrm{C}} = V_z^{\mathrm{CX}}$.

A *completely effective pluri-variate independent histogram*, $A^{\mathrm{XF}} = V^{\mathrm{C}}$ where $|V| > 1$, for which all of the *variables* are *pluri-valent*, $\forall w \in V \ (|U_w| > 1)$, must be *non-causal*,

$$\begin{aligned} &\forall K \subset V \ (K \notin \{\emptyset, V\} \implies \\ &\quad \{(S \ \% \ K, \ S \ \% \ (V \setminus K)) : S \in A^{\mathrm{XFS}}\} \ \notin \ K^{\mathrm{CS}} \to (V \setminus K)^{\mathrm{CS}}) \end{aligned}$$

The set of *substrate histories* $\mathcal{H}_{U,V,z}$ is the set of *histories* having *event identifiers* $\{1 \ldots z\}$, fixed *size* z and fixed *variables* V,

$$\begin{aligned}\mathcal{H}_{U,V,z} &:= \{1 \ldots z\} :\to V^{\mathrm{CS}} \\ &= \{H : H \subseteq \{1 \ldots z\} \times V^{\mathrm{CS}},\ \mathrm{dom}(H) = \{1 \ldots z\},\ |H| = z\}\end{aligned}$$

The cardinality of the *substrate histories* is $|\mathcal{H}_{U,V,z}| = v^z$ where $v = |V^{\mathrm{C}}|$. If the *volume*, v, is finite, the set of *substrate histories* is finite, $|\mathcal{H}_{U,V,z}| < \infty$.

The corresponding set of *integral substrate histograms* $\mathcal{A}_{U,\mathrm{i},V,z}$ is the set of *complete integral histograms* in *variables* V with *size* z,

$$\begin{aligned}\mathcal{A}_{U,\mathrm{i},V,z} &:= \{\mathrm{histogram}(H) : H \in \mathcal{H}_{U,V,z}\} \\ &= \{A : A \in V^{\mathrm{CS}} :\to \{0 \ldots z\},\ \mathrm{size}(A) = z\}\end{aligned}$$

Note that the *histogram* function is redefined here to return *complete histograms*, $\mathrm{histogram}(H) := \{(S, |X|) : (S, X) \in H^{-1}\}\ +\ V^{\mathrm{CS}} \times \{0\}$.

The cardinality of *integral substrate histograms* is the cardinality of weak compositions,

$$|\mathcal{A}_{U,\mathrm{i},V,z}| = \frac{(z + v - 1)!}{z!\ (v - 1)!}$$

If the *volume*, v, is finite, the set of *integral substrate histograms* is finite, $|\mathcal{A}_{U,\mathrm{i},V,z}| < \infty$.

2.3.2 Entropy and alignment

The *entropy* of a *non-zero histogram* $A \in \mathcal{A}$ is defined as the expected negative logarithm of the normalised *counts*,

$$\mathrm{entropy}(A) := -\sum_{S \in A^{\mathrm{FS}}} \hat{A}_S \ln \hat{A}_S$$

(Note that in conventional terminology the *entropy* would be written $H[V]$.) The *sized entropy* is $z \times \mathrm{entropy}(A)$ where $z = \mathrm{size}(A)$. The *entropy* of a *singleton* is zero, $z \times \mathrm{entropy}(\{(\cdot, z)\}) = 0$. *Entropy* is highest in *cartesian histograms*, which are *uniform* and have maximum *effective volume*. The maximum *sized entropy* is $z \times \mathrm{entropy}(V^{\mathrm{C}}_z) = z \ln v$ where $v = |V^{\mathrm{C}}|$.

Given a *histogram* A and a set of query *variables* $K \subset V$, the *scaled* label *entropy* is the degree to which the *histogram* is ambiguous or *non-causal* in the query *variables*, K. It is the sum of the *sized entropies* of the *contingent slices reduced* to the label *variables*, $V \setminus K$,

$$\sum_{R \in (A\%K)^{\mathrm{FS}}} (A\%K)_R \times \mathrm{entropy}(A * \{R\}^{\mathrm{U}} \ \% \ (V \setminus K))$$

The *scaled* label *entropy* is also known as the *scaled* query *conditional entropy*,

$$\begin{aligned}
& \sum_{R \in (A\%K)^{\mathrm{FS}}} (A\%K)_R \times \mathrm{entropy}(A * \{R\}^{\mathrm{U}} \ \% \ (V \setminus K)) \\
= & \ -\sum_{S \in A^{\mathrm{FS}}} A_S \ln \frac{A_S}{(A\%K * V^{\mathrm{C}})_S} \\
= & \ -\sum_{S \in A^{\mathrm{FS}}} A_S \ln(A/(A\%K))_S \\
= & \ z \times \mathrm{entropy}(A) - z \times \mathrm{entropy}(A\%K)
\end{aligned}$$

The query *conditional entropy* is a special case of negative relative entropy, $\mathrm{entropy}(A) - \mathrm{entropy}(A\%K) = -\ \mathrm{entropyRelative}(A, A\%K)$. See appendix 'Entropy and Gibbs' inequality'. (Note that in conventional terminology the query *conditional entropy* would be written $H[V \setminus K \mid K] = H[V] - H[K]$. See the discussion of Bayes' theorem in section 'Transforms and probability', below.)

When the *histogram*, A, is *causal* in the query *variables*, $\mathrm{split}(K, A^{\mathrm{FS}}) \in K^{\mathrm{CS}} \to (V \setminus K)^{\mathrm{CS}}$, the label *entropy* is zero because each *slice* is an *effective singleton*, $\forall R \in (A\%K)^{\mathrm{FS}}\ (|A^{\mathrm{F}} * \{R\}^{\mathrm{U}}| = 1)$. In this case the label *state* is unique for every *effective* query *state*. By contrast, when the label *variables* are *independent* of the query *variables*, $A = Z * \hat{A}\%K * \hat{A}\%(V \setminus K)$, the label *entropy* is maximised.

The *multinomial coefficient* of a *non-zero integral histogram* $A \in \mathcal{A}_{\mathrm{i}}$ is

$$\frac{z!}{\prod_{S \in A^{\mathrm{S}}} A_S!} \ \in \ \mathbf{N}_{>0}$$

where $z = \mathrm{size}(A) > 0$. In the case where the *histogram* is *non-integral* the *multinomial coefficient* is defined by the unit-translated gamma function,

$$\frac{\Gamma_! z}{\prod_{S \in A^{\mathrm{S}}} \Gamma_! A_S}$$

Given an arbitrary *substrate history* $H \in \mathcal{H}_{U,V,z}$ and its *histogram* $A = \text{histogram}(H)$, the cardinality of *histories* having the same *histogram*, A, is the *multinomial coefficient*,

$$|\{G : G \in \mathcal{H}_{U,V,z},\ \text{histogram}(G) = A\}| \quad = \quad \frac{z!}{\prod_{S \in A^{\mathrm{S}}} A_S!}$$

In the case where the *counts* are not small, $z \gg \ln z$, the logarithm of the *multinomial coefficient* approximates to the *sized entropy*,

$$\ln \frac{z!}{\prod_{S \in A^{\mathrm{S}}} A_S!} \quad \approx \quad z \times \text{entropy}(A)$$

so the *entropy*, $\text{entropy}(A)$, is a measure of the probability of the *histogram* of a randomly chosen *history*. *Singleton histograms* are least probable and *uniform histograms* are most probable.

The *sized relative entropy* between a *histogram* and its *independent* is the *sized mutual entropy*,

$$\sum_{S \in A^{\mathrm{FS}}} A_S \ln \frac{A_S}{A^{\mathrm{X}}_S}$$

It can be shown that the *size* scaled expected logarithm of the *independent* with respect to the *histogram* equals the *size* scaled expected logarithm of the *independent* with respect to the *independent*,

$$\sum_{S \in A^{\mathrm{FS}}} A_S \ln A^{\mathrm{X}}_S = \sum_{S \in A^{\mathrm{XFS}}} A^{\mathrm{X}}_S \ln A^{\mathrm{X}}_S$$

so the *sized mutual entropy* is the difference between the *sized independent entropy* and the *sized histogram entropy*,

$$\sum_{S \in A^{\mathrm{FS}}} A_S \ln \frac{A_S}{A^{\mathrm{X}}_S} \quad = \quad z \times \text{entropy}(A^{\mathrm{X}}) - z \times \text{entropy}(A)$$

The *sized mutual entropy* can be viewed as a measure of the probability of the *independent*, A^{X}, relative to the *histogram*, A, given arbitrary *substrate history*. Equivalently, *sized mutual entropy* can be viewed as a measure of the surprisal of the *histogram*, A, relative to the *independent*, A^{X}. That is, *sized mutual entropy* is a measure of the dependency between the *variables* in the *histogram*, A.

The *sized mutual entropy* is the *sized relative entropy* so it is always positive,

$$z \times \text{entropy}(A^{\mathrm{X}}) - z \times \text{entropy}(A) \;\; \geq \;\; 0$$

and so the *independent entropy* is always greater than or equal to the *histogram entropy*

$$\text{entropy}(A^{\mathrm{X}}) \;\; \geq \;\; \text{entropy}(A)$$

That is, *histograms* of *substrate histories* arbitrarily chosen from a uniform distribution are probably *independent* or nearly *independent*. The expected *sized mutual entropy* is low.

An example of a dependency between *variables* is where a *histogram* A is *causal* in a subset of its *variables* $K \subset V$. In this case the *histogram* cannot be *independent*, $A \neq A^{\mathrm{X}}$, and so the *sized mutual entropy* must be greater than zero,

$$\{(S \,\%\, K,\ S \,\%\, (V \setminus K)) : S \in A^{\mathrm{FS}}\} \;\in\; K^{\mathrm{CS}} \to (V \setminus K)^{\mathrm{CS}} \implies z \times \text{entropy}(A^{\mathrm{X}}) - z \times \text{entropy}(A) \; > \; 0$$

The *alignment* of a *histogram* $A \in \mathcal{A}$ is defined

$$\text{algn}(A) \;\; := \;\; \sum_{S \in A^{\mathrm{S}}} \ln \Gamma_! A_S - \sum_{S \in A^{\mathrm{XS}}} \ln \Gamma_! A^{\mathrm{X}}_S$$

where $\Gamma_!$ is the unit-translated gamma function.

In the case where both the *histogram* and its *independent* are *integral*, $A, A^{\mathrm{X}} \in \mathcal{A}_{\mathrm{i}}$, then the *alignment* is the difference between the sum log-factorial *counts* of the *histogram* and its *independent*,

$$\text{algn}(A) \;\; = \;\; \sum_{S \in A^{\mathrm{S}}} \ln A_S! - \sum_{S \in A^{\mathrm{XS}}} \ln A^{\mathrm{X}}_S!$$

Alignment is the logarithm of the ratio of the *independent multinomial coefficient* to the *multinomial coefficient*,

$$\text{algn}(A) \;\; = \;\; \ln \left(\frac{z!}{\prod_{S \in A^{\mathrm{XS}}} A^{\mathrm{X}}_S!} \; / \; \frac{z!}{\prod_{S \in A^{\mathrm{S}}} A_S!} \right)$$

so *alignment* is the logarithm of the probability of the *independent*, A^{X}, relative to the *histogram*, A. Equivalently, *alignment* is the logarithm of the surprisal of the *histogram*, A, relative to the *independent*, A^{X}. *Alignment* is a measure of the dependency between the *variables* in the *histogram*, A.

Alignment is approximately equal to the *sized mutual entropy*,

$$\begin{aligned}\mathrm{algn}(A) &\approx z \times \mathrm{entropy}(A^{\mathrm{X}}) - z \times \mathrm{entropy}(A) \\ &= \sum_{S \in A^{\mathrm{FS}}} A_S \ln \frac{A_S}{A^{\mathrm{X}}_S}\end{aligned}$$

so the *histogram* of an arbitrary *history* usually has low *alignment*. Note that, because *sized entropy* is only an approximation to the logarithm of the *multinomial coefficient*, especially at low *sizes*, *alignment* is the better measure of the surprisal of the *histogram*, A, relative to the *independent*, A^{X}, than *sized mutual entropy*.

The *alignment* of an *independent histogram*, $A = A^{\mathrm{X}}$, is zero. In particular, *scalar histograms*, $V = \emptyset$, *mono-variate histograms*, $|V| = 1$, *uniform cartesian histograms*, $A = V^{\mathrm{C}}_z$, and *effective singleton histograms*, $|A^{\mathrm{F}}| = 1$, all have zero *alignment*.

The maximum *alignment* of a *histogram* A occurs when the *histogram* is both *uniform* and *fully diagonalised*. No pair of *effective states* shares any *value*, $\forall S,T \in A^{\mathrm{FS}}\ (S \neq T \implies S \cap T = \emptyset)$, and all *counts* are equal along the *diagonal*, $\forall S,T \in A^{\mathrm{FS}}\ (A_S = A_T)$. The maximum *alignment* of a *regular histogram* with *dimension* $n = |V|$ and *valency* d is

$$d \ln \Gamma_! \frac{z}{d} \ - \ d^n \ln \Gamma_! \frac{z}{d^n}$$

The maximum *alignment* is approximately $z \ln d^{n-1} = z \ln v/d$, where $v = d^n$. It can be compared to the maximum *sized entropy* of the *'co-histogram' reduced* by one *variable* along the *diagonal*.

Although *alignment* varies against *sized entropy*, $\mathrm{algn}(A) \sim -z \times \mathrm{entropy}(A)$, the maximum *alignment* does not occur when the *entropy* is minimised. At minimum *entropy* the *histogram* is a *singleton*, but the *alignment* is zero because *singletons* are *independent*.

An example of an *aligned histogram* A is where the *histogram* is *causal* in a subset of its *variables* $K \subset V$. In this case the *histogram* cannot be *independent*, $A \neq A^{\mathrm{X}}$, and so the *alignment* must be greater than zero,

$$\{(S\%K,\ S\%(V \setminus K)) : S \in A^{\mathrm{FS}}\} \in K^{\mathrm{CS}} \to (V \setminus K)^{\mathrm{CS}} \implies \mathrm{algn}(A) > 0$$

At maximum *alignment* the *histogram* is *fully diagonalised*, so all pairs of *variables* are necessarily bijectively *causal* or equational,

$$\forall u, w \in V\ (\{(S\%\{u\}, S\%\{w\}) : S \in A^{\mathrm{FS}}\}\ \in\ \{u\}^{\mathrm{CS}} \leftrightarrow \{w\}^{\mathrm{CS}})$$

The *alignment* is approximately equal to the *scaled mutual entropy*, so the *alignment* varies against the *scaled* label *entropy* or *scaled* query *conditional entropy*,

$$\begin{aligned}
\mathrm{algn}(A) &\approx z \times \mathrm{entropy}(A^{\mathrm{X}}) - z \times \mathrm{entropy}(A) \\
&\sim z \times \mathrm{entropy}(A\%K) + z \times \mathrm{entropy}(A\%(V \setminus K)) - z \times \mathrm{entropy}(A) \\
&\sim -(z \times \mathrm{entropy}(A) - z \times \mathrm{entropy}(A\%K)) \\
&= -\sum_{R \in (A\%K)^{\mathrm{FS}}} (A\%K)_R \times \mathrm{entropy}(A * \{R\}^{\mathrm{U}}\ \%\ (V \setminus K))
\end{aligned}$$

The *conditional entropy* is directed from the query *variables* to the label *variables*, whereas the *alignment* is symmetrical with respect to the *variables*.

2.3.3 Encoding and compression

A *substrate history probability function* $P \in (\mathcal{H}_{U,V,z} :\to \mathbf{Q}_{\geq 0}) \cap \mathcal{P}$ is a normalised distribution over *substrate histories*, $\sum(P_H : H \in \mathcal{H}_{U,V,z}) = 1$. The entropy of the *probability function* is $\mathrm{entropy}(P)$. Note that *history probability function* entropy is not to be confused with *histogram entropy*. A *history probability function* is a distribution over *histories*, $\mathcal{H}_{U,V,z} \to \mathbf{Q}_{\geq 0}$, whereas a *histogram* is a distribution of *events* over *states*, $V^{\mathrm{CS}} \to \mathbf{Q}_{\geq 0}$.

History coders define the conversion of lists of *histories*, $\mathcal{L}(\mathcal{H})$, to and from the natural numbers, $\mathbf{N}$. A *substrate history coder* $C \in \mathrm{coders}(\mathcal{H}_{U,V,z})$ defines an *encode* function of any list of *substrate histories* into a positive integer, $\mathrm{encode}(C) \in \mathcal{L}(\mathcal{H}_{U,V,z}) :\to \mathbf{N}$, and the corresponding *decode* function of the integer back into the list of *histories*, $\mathrm{decode}(C) \in \mathbf{N} \times \mathbf{N} \to: \mathcal{L}(\mathcal{H}_{U,V,z})$, given the length of the list.

A third function is the *space* function, $\mathrm{space}(C) \in \mathcal{H}_{U,V,z} :\to \ln \mathbf{N}_{>0}$, which defines the logarithm of the cardinality of the encoding states of a *substrate history*. The encoding integer of a single *history* is always less than this cardinality, $\forall H \in \mathcal{H}_{U,V,z}\ (\mathrm{encode}(C)(\{(1, H)\}) < \exp(\mathrm{space}(C)(H)))$. The *space* of an encoded list of *histories* is the sum of the *spaces* of the *histories*. The *space* function is also denoted $C^{\mathrm{s}} = \mathrm{space}(C)$.

Given a *substrate history probability function* $P \in (\mathcal{H}_{U,V,z} :\to \mathbf{Q}_{\geq 0}) \cap \mathcal{P}$, the *expected substrate history space* is $\sum(P_H C^{\mathrm{s}}(H) : H \in \mathcal{H}_{U,V,z})$. The *expected space* is always greater than or equal to the *probability function* entropy (or Shannon entropy in nats), $\sum(P_H C^{\mathrm{s}}(H) : H \in \mathcal{H}_{U,V,z}) \geq \mathrm{entropy}(P)$.

A *minimal history coder* $C_{\mathrm{m},U,V,z} \in \mathrm{coders}(\mathcal{H}_{U,V,z})$ encodes the *history* by encoding the index of an enumeration of the entire set of *substrate histories*, $\mathrm{encode}(C_{\mathrm{m},U,V,z})(\{(1,H)\}) \in \{0 \ldots v^z - 1\}$. The *space* is fixed, $C^{\mathrm{s}}_{\mathrm{m},U,V,z}(H) = \ln |\mathcal{H}_{U,V,z}| = z \ln v$. In the case where the *probability function* is uniform, $P = \mathcal{H}_{U,V,z} \times \{1/v^z\}$, the *expected space* equals the *probability function* entropy, $\sum(P_H C^{\mathrm{s}}_{\mathrm{m},U,V,z}(H) : H \in \mathcal{H}_{U,V,z}) = \mathrm{entropy}(P) = z \ln v$. In other words, when the *history* is arbitrary then the *minimal history coder* has the least *expected space*.

There are two *canonical history coders*, the *index history coder* C_{H} and the *classification coder* C_{G}. The *index substrate history coder* $C_{\mathrm{H},U,V,z} \in \mathrm{coders}(\mathcal{H}_{U,V,z})$ is the simpler of the two. It encodes each *history* by indexing the *volume* for each *event*. The *space* of an index into a *volume* $v = |V^{\mathrm{CS}}|$ is $\ln v$. So the total *space* of any *substrate history* $H \in \mathcal{H}_{U,V,z}$ is

$$C^{\mathrm{s}}_{\mathrm{H},U,V,z}(H) \quad = \quad z \ln v$$

The *space* is fixed because it does not depend on the *histogram*, A. The *index history space* equals the *minimal history space*, $C^{\mathrm{s}}_{\mathrm{H},U,V,z}(H) = C^{\mathrm{s}}_{\mathrm{m},U,V,z}(H) = z \ln v$, but the *encode* functions are different. In the case of an arbitrary *history*, or uniform *history probability function*, the *index history coder* also has least *expected space*.

The *classification substrate history coder* $C_{\mathrm{G},U,V,z} \in \mathrm{coders}(\mathcal{H}_{U,V,z})$ encodes each *history* in two steps. First the *histogram* is encoded by choosing one of the *integral substrate histograms*, $\mathcal{A}_{U,\mathrm{i},V,z}$. The choice has fixed *space*

$$\ln |\mathcal{A}_{U,\mathrm{i},V,z}| \quad = \quad \ln \frac{(z+v-1)!}{z!\ (v-1)!}$$

Given the *histogram*, A, the cardinality of *classifications* equals the *multinomial coefficient*. Now the choice of the *classification*, H^{-1}, is encoded in a *space* equal to the logarithm of the *multinomial coefficient*,

$$\ln \frac{z!}{\prod_{S \in A^{\mathrm{S}}} A_S!}$$

The total *space* to encode the *history* in the *classification substrate history coder* is

$$C^{\mathrm{s}}_{\mathrm{G},U,V,z}(H) \quad = \quad \ln \frac{(z+v-1)!}{z!\ (v-1)!} \ + \ \ln \frac{z!}{\prod_{S \in A^{\mathrm{s}}} A_S!}$$

The *space* is not fixed because it depends on the *histogram*, A.

The *classification space* may be approximated in terms of *sized entropy*,

$$C^{\mathrm{s}}_{\mathrm{G},U,V,z}(H) \ \approx \ (z+v)\ln(z+v) \ - \ z \ln z \ - \ v \ln v \ + \ z \times \mathrm{entropy}(A)$$

The maximum *sized entropy*, $z \times \mathrm{entropy}(A)$, is $z \ln v$, so when the *entropy* is high the *classification space* is greater than the *index space*, $C^{\mathrm{s}}_{\mathrm{G},U,V,z}(H) > C^{\mathrm{s}}_{\mathrm{H},U,V,z}(H)$, but when the *entropy* is low the *classification space* is less than the *index space*, $C^{\mathrm{s}}_{\mathrm{G},U,V,z}(H) < C^{\mathrm{s}}_{\mathrm{H},U,V,z}(H)$. The break-even *sized entropy* is approximately

$$z \times \mathrm{entropy}(A) \ \approx \ z \ln v \ - \ ((z+v)\ln(z+v) \ - \ z \ln z \ - \ v \ln v)$$

In the case where the *size* is much less than the *volume*, $z \ll v$, the break-even *sized entropy* is approximately $z \times \mathrm{entropy}(A) \approx z \ln z$.

2.4 Induction without model

Induction may be defined as the determination of the *likely* properties of *unknown history probability functions*.

Let P be a *substrate history probability function*, $P \in (\mathcal{H}_{U,V,z} :\to \mathbf{Q}_{\geq 0}) \cap \mathcal{P}$. Let the domain of the *probability function*, $\mathrm{dom}(P) = \mathcal{H}_{U,V,z}$, be *known*. The simplest case of *induction* is that nothing else is *known* about the *probability function*, P. If the *probability function* is assumed to be the normalisation of the distribution of a finite *history* valued function of undefined particle, $\mathcal{X} \to \mathcal{H}$, and this particle function is assumed to be chosen arbitrarily, then the *maximum likelihood estimate* $\tilde{P}$ for the *probability function*, P, maximises the entropy, $\mathrm{entropy}(\tilde{P})$, at the mode. So the *likely history probability function*, $\tilde{P}$, is the uniform distribution,

$$\tilde{P} \ = \ \mathcal{H}_{U,V,z} \times \{1/v^z\}$$

That is, the *likely substrate histories* are arbitrary or random.

The next case is where a *history* $H \in \mathcal{H}_{U,V,z}$ is *known* to be *necessary*, $P(H) = 1$. In this case the *probability function*, P, is,

$$\begin{aligned} P &= \{(H,1)\} \cup \\ &\quad\ \{(G,0) : G \in \mathcal{H}_{U,V,z},\ G \neq H\} \end{aligned}$$

If the *history*, H, is *known*, then the *probability function*, P, is *known*. The *maximum likelihood estimate* equals the *probability function*, $\tilde{P} = P$. The entropy is zero, $\mathrm{entropy}(\tilde{P}) = 0$.

2.4.1 Classical induction

In *classical induction* the *history probability functions* are constrained by *histogram*.

Let his = histogram. Now consider the case where the *histogram* $A \in \mathcal{A}_{U,\mathrm{i},V,z}$ is *known* to be *necessary*, $\sum(P(H) : H \in \mathcal{H}_{U,V,z},\ \mathrm{his}(H) = A) = 1$. The *maximum likelihood estimate* which maximises the entropy, $\mathrm{entropy}(\tilde{P})$, is

$$\begin{aligned} \tilde{P} &= \{(H,1) : H \in \mathcal{H}_{U,V,z},\ \mathrm{his}(H) = A\}^{\wedge} \cup \\ &\quad\ \{(G,0) : G \in \mathcal{H}_{U,V,z},\ \mathrm{his}(G) \neq A\} \\ &= \{(H, 1/\frac{z!}{\prod_{S \in A^{\mathrm{S}}} A_S!}) : H \in \mathcal{H}_{U,V,z},\ \mathrm{his}(H) = A\} \cup \\ &\quad\ \{(G,0) : G \in \mathcal{H}_{U,V,z},\ \mathrm{his}(G) \neq A\} \end{aligned}$$

where $()^{\wedge}$ = normalise. That is, the *maximum likelihood estimate*, $\tilde{P}$, is such that all *histories* with the *histogram*, $\mathrm{his}(H) = A$, are uniformly probable and all other *histories*, $\mathrm{his}(G) \neq A$, are impossible, $\tilde{P}(G) = 0$. If the *histogram*, A, is *known*, then the *likely probability function*, $\tilde{P}$, is *known*. Note that the *likely history probability function* entropy varies with the *histogram entropy*, $\mathrm{entropy}(\tilde{P}) \sim \mathrm{entropy}(A)$.

Next consider the case where either *histogram* A or *histogram* B are *known* to be *necessary*, $\sum(P(H) : H \in \mathcal{H}_{U,V,z},\ (\mathrm{his}(H) = A\ \vee\ \mathrm{his}(H) = B)) = 1$. The *maximum likelihood estimate* which maximises the entropy, $\mathrm{entropy}(\tilde{P})$, is

$$\begin{aligned}
\tilde{P} &= \{(H,1) : H \in \mathcal{H}_{U,V,z},\ (\text{his}(H) = A \ \vee\ \text{his}(H) = B)\}^{\wedge}\ \cup \\
&\quad\ \{(G,0) : G \in \mathcal{H}_{U,V,z},\ \text{his}(G) \neq A,\ \text{his}(G) \neq B\} \\
&= \{(H, 1/\left(\frac{z!}{\prod_{S \in A^{\mathrm{S}}} A_S!} + \frac{z!}{\prod_{S \in B^{\mathrm{S}}} B_S!}\right)) : \\
&\qquad\qquad H \in \mathcal{H}_{U,V,z},\ (\text{his}(H) = A\ \vee\ \text{his}(H) = B)\}\ \cup \\
&\quad\ \{(G,0) : G \in \mathcal{H}_{U,V,z},\ \text{his}(G) \neq A,\ \text{his}(G) \neq B\}
\end{aligned}$$

That is, the *maximum likelihood estimate*, $\tilde{P}$, is such that all *histories* with either *histogram*, A or B, are uniformly probable and all other *histories*, $\text{his}(G) \neq A$ and $\text{his}(G) \neq B$, are impossible, $\tilde{P}(G) = 0$. If the *histograms*, A and B, are *known*, then the *likely probability function*, $\tilde{P}$, is *known*.

Given a *history* $H_E \in \mathcal{H}_{U,V,z_E}$, of *size* $z_E = |H_E|$, consider the case where its subsets of *size* z are *known* to be necessary, $\sum(P(H) : H \subseteq H_E,\ |H| = z) = 1$. The given *history*, H_E, is called the *distribution history*. A subset $H \subseteq H_E$ is a *sample history drawn* from the *distribution history*, H_E. The *maximum likelihood estimate* which maximises the entropy, $\text{entropy}(\tilde{P})$, is

$$\begin{aligned}
\tilde{P} &= \{(H,1) : H \subseteq H_E,\ |H| = z\}^{\wedge}\ \cup \\
&\quad\ \{(G,0) : G \in \mathcal{H}_{U,V,z},\ G \not\subseteq H_E\}\ \cup \\
&\quad\ \{(G,0) : G \in \mathcal{H}_{U,V,z},\ |G| \neq z\} \\
&= \{(H, 1/\binom{z_E}{z}) : H \subseteq H_E,\ |H| = z\}\ \cup \\
&\quad\ \{(G,0) : G \in \mathcal{H}_{U,V,z},\ G \not\subseteq H_E\}\ \cup \\
&\quad\ \{(G,0) : G \in \mathcal{H}_{U,V,z},\ |G| \neq z\}
\end{aligned}$$

That is, the *maximum likelihood estimate*, $\tilde{P}$, is such that all *drawn histories* $H \subseteq H_E$ of *size* $|H| = z$ are uniformly probable and all other *histories*, $G \not\subseteq H_E$ or $|G| \neq z$, are impossible, $\tilde{P}(G) = 0$. If the *distribution histogram*, H_E, is *known*, then the *likely probability function*, $\tilde{P}$, is *known*.

Now consider the case where the *drawn histogram* A is *known* to be *necessary*, $\sum(P(H) : H \subseteq H_E,\ \text{his}(H) = A) = 1$. The *maximum likelihood* *estimate* which maximises the entropy, $\text{entropy}(\tilde{P})$, is

$$\begin{aligned}\tilde{P} &= \{(H,1) : H \subseteq H_E,\ \text{his}(H) = A\}^{\wedge} \ \cup \\ &\quad \{(G,0) : G \in \mathcal{H}_{U,V,z},\ G \nsubseteq H_E\} \ \cup \\ &\quad \{(G,0) : G \in \mathcal{H}_{U,V,z},\ \text{his}(G) \neq A\} \\ &= \{(H, 1/\prod_{S \in A^{\mathrm{S}}} \binom{E_S}{A_S}) : H \subseteq H_E,\ \text{his}(H) = A\} \ \cup \\ &\quad \{(G,0) : G \in \mathcal{H}_{U,V,z},\ G \nsubseteq H_E\} \ \cup \\ &\quad \{(G,0) : G \in \mathcal{H}_{U,V,z},\ \text{his}(G) \neq A\}\end{aligned}$$

where the *distribution histogram* $E = \text{his}(H_E)$.

That is, the *maximum likelihood estimate*, $\tilde{P}$, is such that all *drawn histories* $H \subseteq H_E$ with the *histogram*, $\text{his}(H) = A$, are uniformly probable and all other *histories*, $G \nsubseteq H_E$ or $\text{his}(G) \neq A$, are impossible, $\tilde{P}(G) = 0$. If the *histogram*, A, is *known* and the *distribution histogram*, H_E, is *known*, then the *likely probability function*, $\tilde{P}$, is *known*.

The *historical distribution* $Q_{\mathrm{h},U}$ is defined

$$Q_{\mathrm{h},U}(E,z)(A) \ := \ \prod_{S \in A^{\mathrm{S}}} \binom{E_S}{A_S} = \prod_{S \in A^{\mathrm{S}}} \frac{E_S!}{A_S!\ (E_S - A_S)!}$$

where $A \leq E$. The *frequency* of *histogram* A in the *historical distribution*, $Q_{\mathrm{h},U}$, parameterised by *draw* (E,z), is the cardinality of *histories drawn without replacement* having *histogram* A,

$$Q_{\mathrm{h},U}(E,z)(A) \ = \ |\{H : H \subseteq H_E,\ \text{his}(H) = A\}|$$

The *historical probability distribution* is normalised,

$$\hat{Q}_{\mathrm{h},U}(E,z)(A) \ := \ 1/\binom{z_E}{z} \times Q_{\mathrm{h},U}(E,z)(A)$$

The *likely history probability function*, $\tilde{P}$, can be re-written in terms of the *historical distribution*,

$$\begin{aligned}\tilde{P} &= \{(H, 1/Q_{\mathrm{h},U}(E,z)(A)) : H \subseteq H_E,\ \text{his}(H) = A\} \ \cup \\ &\quad \{(G,0) : G \in \mathcal{H}_{U,V,z},\ G \nsubseteq H_E\} \ \cup \\ &\quad \{(G,0) : G \in \mathcal{H}_{U,V,z},\ \text{his}(G) \neq A\}\end{aligned}$$

So the *likely history probability function* entropy, $\mathrm{entropy}(\tilde{P})$, is maximised when the *historical distribution frequency*, $Q_{\mathrm{h},U}(E,z)(A)$, is maximised.

Consider the case where the *histogram*, A, is *known*, but the *distribution histogram*, E, is *unknown* and hence the *likely history probability function*, $\tilde{P}$, is *unknown*. The *historical probability distribution* is a *probability function*, $\hat{Q}_{\mathrm{h},U}(E,z) \in \mathcal{P}$, parameterised by the *distribution histogram*, E, so there is a corresponding *likelihood function* $L_{\mathrm{h},U}(A) \in \mathcal{A}_{U,\mathrm{i},V,z_E} \to \mathbf{Q}_{\geq 0}$ such that $L_{\mathrm{h},U}(A)(E) = \hat{Q}_{\mathrm{h},U}(E,z)(A)$. The *maximum likelihood estimate* $\tilde{E}$ for the *distribution histogram*, E, is a modal value of this *likelihood function*,

$$\begin{aligned}
\tilde{E} &\in \mathrm{maxd}(L_{\mathrm{h},U}(A)) \\
&= \mathrm{maxd}(\{(D, Q_{\mathrm{h},U}(D,z)(A)) : D \in \mathcal{A}_{U,\mathrm{i},V,z_E}\})
\end{aligned}$$

The *likely distribution histogram*, $\tilde{E}$, is *known* if the *distribution histogram size*, z_E, is *known* and the *histogram*, A, is *known*. If it is assumed that the *distribution histogram* equals the *likely distribution histogram*, $E = \tilde{E}$, then the *likely history probability* is *known*, $\tilde{P}(H) = 1/Q_{\mathrm{h},U}(\tilde{E},z)(A)$ where $\mathrm{his}(H) = A$.

The *multinomial distribution* $Q_{\mathrm{m},U}$ is defined

$$Q_{\mathrm{m},U}(E,z)(A) \;:=\; \frac{z!}{\prod_{S \in A^{\mathrm{S}}} A_S!} \prod_{S \in A^{\mathrm{S}}} E_S^{A_S}$$

where $A^{\mathrm{F}} \leq E^{\mathrm{F}}$. The *frequency* of *histogram* A in the *multinomial distribution*, $Q_{\mathrm{m},U}$, parameterised by *draw* (E,z), is the cardinality of *histories drawn with replacement* having *histogram* A,

$$Q_{\mathrm{m},U}(E,z)(A) \;=\; |\{L : L \in H_E^z,\ \mathrm{his}(\{((i,x),S) : (i,(x,S)) \in L\}) = A\}|$$

where $H_E^z \in \mathcal{L}(H_E)$ is the set of lists of the *distribution history events* of length z.

The *multinomial probability distribution* is normalised,

$$\begin{aligned}
\hat{Q}_{\mathrm{m},U}(E,z)(A) &:= \frac{1}{z_E^z} \times Q_{\mathrm{m},U}(E,z)(A) \\
&= \frac{z!}{\prod_{S \in A^{\mathrm{S}}} A_S!} \prod_{S \in A^{\mathrm{S}}} \hat{E}_S^{A_S}
\end{aligned}$$

so the *multinomial probability*, $\hat{Q}_{\mathrm{m},U}(E,z)(A) = \hat{Q}_{\mathrm{m},U}(\hat{E},z)(A)$, does not depend on the *distribution histogram size*, z_E.

As the *distribution histogram size*, z_E, tends to infinity, the *historical probability* tends to the *multinomial probability*. That is, for large *distribution histogram size*, $z_E \gg z$, the *historical probability* may be approximated by the *multinomial probability*, $\hat{Q}_{\mathrm{h},U}(E,z)(A) \approx \hat{Q}_{\mathrm{m},U}(E,z)(A)$.

In the case where the *distribution histogram* is *known* to be *cartesian*, $E = V_{z_E}^{\mathrm{C}}$, but the *distribution histogram size*, z_E, is *unknown*, except that it is *known* to be large, $z_E \gg z$, then the case where the *drawn histogram*, A, is *known* to be *necessary*, $\sum(P(H) : H \subseteq H_E,\ \mathrm{his}(H) = A) = 1$, approximates to the case where the *substrate histogram*, A, is *known* to be *necessary*, $\sum(P(H) : H \in \mathcal{H}_{U,V,z},\ \mathrm{his}(H) = A) = 1$. That is,

$$\begin{aligned}
\tilde{P} &= \{(H, 1/\prod_{S \in A^{\mathrm{S}}} \binom{V_{z_E}^{\mathrm{C}}(S)}{A(S)}) : H \subseteq H_E,\ \mathrm{his}(H) = A\} \ \cup \\
&\quad \{(G,0) : G \in \mathcal{H}_{U,V,z},\ G \nsubseteq H_E\} \ \cup \\
&\quad \{(G,0) : G \in \mathcal{H}_{U,V,z},\ \mathrm{his}(G) \neq A\} \\
&\approx \{(H, 1/\frac{z!}{\prod_{S \in A^{\mathrm{S}}} A_S!}) : H \in \mathcal{H}_{U,V,z},\ \mathrm{his}(H) = A\} \ \cup \\
&\quad \{(G,0) : G \in \mathcal{H}_{U,V,z},\ \mathrm{his}(G) \neq A\}
\end{aligned}$$

In this case, the *likely history probability function* entropy varies with the *histogram entropy*, $\mathrm{entropy}(\tilde{P}) \sim \mathrm{entropy}(A)$.

In the case where the *distribution histogram*, E, is *unknown*, and the *distribution histogram size*, z_E, is also *unknown*, except that it is *known* to be large, $z_E \gg z$, then the *maximum likelihood estimate* $\tilde{E}$ for the *distribution probability histogram*, $\hat{E}$, may be approximated by a modal value of a *likelihood function* which depends on the *multinomial distribution* instead,

$$\tilde{E} \in \mathrm{maxd}(\{(D, Q_{\mathrm{m},U}(D,z)(A)) : D \in \mathcal{A}_{U,V,1}\})$$

The *mean* of the *multinomial probability distribution* is the *sized distribution histogram*,

$$\mathrm{mean}(\hat{Q}_{\mathrm{m},U}(E,z)) = \mathrm{scalar}(z) * \hat{E}$$

so the *maximum likelihood estimate*, $\tilde{E}$, for the *distribution probability histogram*, $\hat{E}$, is the *sample probability histogram*, $\hat{A}$,

$$\tilde{E} = \hat{A}$$

If it is assumed that the *distribution probability histogram* equals the *likely distribution probability histogram*, $\hat{E} = \tilde{E} = \hat{A}$, then the *likely history probability* varies against the *sample-distributed multinomial probability*, $\tilde{P}(H) \sim 1/\hat{Q}_{\mathrm{m},U}(\hat{A},z)(A)$.

The *sample-distributed multinomial log-likelihood* is

$$\ln \hat{Q}_{\mathrm{m},U}(A,z)(A) = \ln z! - z \ln z - \sum_{S \in A^{\mathrm{S}}} \ln A_S! + \sum_{S \in A^{\mathrm{FS}}} A_S \ln A_S$$

which varies against the sum of the logarithms of the *counts*

$$\ln \hat{Q}_{\mathrm{m},U}(A,z)(A) \quad \sim \quad - \sum_{S \in A^{\mathrm{FS}}} \ln A_S$$

So the *log-likelihood* varies weakly against the *histogram entropy*,

$$\ln \hat{Q}_{\mathrm{m},U}(A,z)(A) \quad \sim \quad - \,\mathrm{entropy}(A)$$

If it is assumed that the *distribution probability histogram* equals the *likely distribution probability histogram*, $\hat{E} = \tilde{E} = \hat{A}$, then the *likely history probability function* entropy varies against the *histogram entropy*, $\mathrm{entropy}(\tilde{P}) \sim -\,\mathrm{entropy}(A)$, in contrast to the case where the *distribution histogram* is *cartesian*.

The *Fisher information* of a *probability function* varies with the negative curvature of the *likelihood function* near the *maximum likelihood estimate* of the parameter. So the *Fisher information* is a measure of the sensitivity of the *likelihood function* with respect to the *maximum likelihood estimate*. The *Fisher information* of the *multinomial probability distribution*, $\hat{Q}_{\mathrm{m},U}(E,z)$, is the *sum sensitivity*

$$\mathrm{sum}(\mathrm{sensitivity}(U)(\hat{Q}_{\mathrm{m},U}(E,z))) \quad = \quad \sum_{S \in V^{\mathrm{CS}}} \frac{z}{\hat{E}_S(1-\hat{E}_S)}$$

The *sum sensitivity* varies against the *sized entropy*,

$$\mathrm{sum}(\mathrm{sensitivity}(U)(\hat{Q}_{\mathrm{m},U}(E,z))) \quad \sim \quad -z \times \mathrm{entropy}(E)$$

So, in the case of *sample-distributed multinomial probability distribution*, $\hat{Q}_{\mathrm{m},U}(A,z)$, the *sum sensitivity* varies weakly with the *log-likelihood*,

$$\begin{aligned} \mathrm{sum}(\mathrm{sensitivity}(U)(\hat{Q}_{\mathrm{m},U}(A,z))) \quad &\sim \quad -z \times \mathrm{entropy}(A) \\ &\sim \quad \ln \hat{Q}_{\mathrm{m},U}(A,z)(A) \end{aligned}$$

If it is assumed that the *distribution probability histogram* equals the *likely distribution probability histogram*, $\hat{E} = \tilde{E} = \hat{A}$, then, as the *likely history probability function* entropy, $\mathrm{entropy}(\tilde{P})$, increases, the *sensitivity* to the *distribution histogram*, $\tilde{E}$, increases.

The lower the *entropy* of the *sample* the more *likely* the *normalised sample histogram*, $\hat{A}$, equals the *normalised distribution histogram*, $\hat{E}$, but the larger the *likely* difference between them if they are not equal.

Now consider the case where either the *drawn histogram* A or the *drawn histogram* B are *known* to be *necessary*, $\sum(P(H) : H \subseteq H_E,\ (\mathrm{his}(H) = A \vee \mathrm{his}(H) = B)) = 1$. The *maximum likelihood estimate* which maximises the entropy, $\mathrm{entropy}(\tilde{P})$, is

$$\begin{aligned}
\tilde{P} &= \{(H,1) : H \subseteq H_E,\ (\mathrm{his}(H) = A\ \vee\ \mathrm{his}(H) = B)\}^{\wedge}\ \cup \\
&\quad \{(G,0) : G \in \mathcal{H}_{U,V,z},\ G \nsubseteq H_E\}\ \cup \\
&\quad \{(G,0) : G \in \mathcal{H}_{U,V,z},\ \mathrm{his}(G) \neq A,\ \mathrm{his}(G) \neq B\} \\
&= \{(H, 1/(Q_{\mathrm{h},U}(E,z)(A) + Q_{\mathrm{h},U}(E,z)(B))) : \\
&\qquad\qquad H \subseteq H_E,\ (\mathrm{his}(H) = A\ \vee\ \mathrm{his}(H) = B)\}\ \cup \\
&\quad \{(G,0) : G \in \mathcal{H}_{U,V,z},\ G \nsubseteq H_E\}\ \cup \\
&\quad \{(G,0) : G \in \mathcal{H}_{U,V,z},\ \mathrm{his}(G) \neq A,\ \mathrm{his}(G) \neq B\}
\end{aligned}$$

That is, the *maximum likelihood estimate*, $\tilde{P}$, is such that all *drawn histories* $H \subseteq H_E$ with either *histogram*, A or B, are uniformly probable and all other *histories*, $G \nsubseteq H_E$ or $\mathrm{his}(G) \neq A$ and $\mathrm{his}(G) \neq B$, are impossible, $\tilde{P}(G) = 0$. If the *histograms*, A and B, are *known* and the *distribution histogram*, H_E, is *known*, then the *likely probability function*, $\tilde{P}$, is *known*.

The *likely probability* of *drawing histogram* A from *necessary drawn histograms* A or B is

$$\sum(\tilde{P}(H) : H \in \mathcal{H}_{U,V,z},\ \mathrm{his}(H) = A) = \frac{Q_{\mathrm{h},U}(E,z)(A)}{Q_{\mathrm{h},U}(E,z)(A) + Q_{\mathrm{h},U}(E,z)(B)}$$

The *likely history probability function* entropy, $\mathrm{entropy}(\tilde{P})$, is maximised when the sum of the *historical frequencies*, $Q_{\mathrm{h},U}(E,z)(A) + Q_{\mathrm{h},U}(E,z)(B)$, is maximised.

Consider the case where the *drawn histograms*, A and B, are *known*, but the *distribution histogram*, E, is *unknown* and hence the *likely history probability function*, $\tilde{P}$, is *unknown*. The *maximum likelihood estimate* $\tilde{E}$ for the *distribution histogram*, E, is a modal value of the *likelihood function*,

$$\tilde{E} \in \text{maxd}(\{(D, Q_{\text{h},U}(D,z)(A) + Q_{\text{h},U}(D,z)(B)) : D \in \mathcal{A}_{U,\text{i},V,z_E}\})$$

The *likely distribution histogram*, $\tilde{E}$, is *known* if the *distribution histogram size*, z_E, is *known* and the *drawn histograms*, A and B, are *known*. If it is assumed that the *distribution histogram* equals the *likely distribution histogram*, $E = \tilde{E}$, then the *likely history probability* is *known*, $\tilde{P}(H) = 1/(Q_{\text{h},U}(\tilde{E},z)(A) + Q_{\text{h},U}(\tilde{E},z)(B))$ where $\text{his}(H) = A$ or $\text{his}(H) = B$.

In the case where the *distribution histogram*, E, is *unknown*, and the *distribution histogram size*, z_E, is also *unknown*, except that it is *known* to be large, $z_E \gg z$, then the *maximum likelihood estimate* $\tilde{E}$ for the *distribution probability histogram*, $\hat{E}$, may be approximated by a modal value of a *likelihood function* which depends on the *multinomial distribution* instead,

$$\tilde{E} \in \text{maxd}(\{(D, Q_{\text{m},U}(D,z)(A) + Q_{\text{m},U}(D,z)(B)) : D \in \mathcal{A}_{U,V,1}\})$$

Now the *likely distribution histogram*, $\tilde{E}$, is *known* if there is a computable solution and the *drawn histograms*, A and B, are *known*.

Consider the case where the *histogram* is *uniformly possible*. Instead of assuming the *substrate history probability function* $P \in (\mathcal{H}_{U,V,z} :\to \mathbf{Q}_{\geq 0}) \cap \mathcal{P}$ to be the distribution of an arbitrary *history* valued function of undefined particle, $\mathcal{X} \to \mathcal{H}$, assume that it is the distribution of an arbitrary *history* valued function, $\mathcal{X} \to \mathcal{H}$, given an arbitrary *histogram* valued function, $\mathcal{X} \to \mathcal{A}$. In this case, the *history* valued function is chosen arbitrarily from the constrained subset

$$\begin{aligned}&\{G : F \in \mathcal{X} \to (\mathcal{A} \times (\mathcal{X} \to \mathcal{H})),\\ &\qquad (\cdot,(A,G)) \in F,\ \forall (\cdot,H) \in G\ (\text{his}(H) = A)\} \subset \mathcal{X} \to \mathcal{H}\end{aligned}$$

In the case where there is no *distribution history*, the *maximum likelihood estimate* which maximises the entropy, $\text{entropy}(\tilde{P})$, is

$$\begin{aligned}\tilde{P} &= \left(\bigcup\{\{(H,1) : H \in \mathcal{H}_{U,V,z},\ \text{his}(H) = A\}^{\wedge} : A \in \mathcal{A}_{U,\text{i},V,z}\}\right)^{\wedge}\\ &= \{(H, 1/|\mathcal{A}_{U,\text{i},V,z}| \times 1/\frac{z!}{\prod_{S \in A^{\text{S}}} A_S!}) : H \in \mathcal{H}_{U,V,z},\ A = \text{his}(H)\}\end{aligned}$$

That is, the *maximum likelihood estimate*, $\tilde{P}$, is such that all *histograms* are uniformly probable, $\forall A \in \mathcal{A}_{U,\text{i},V,z}\ (\sum(\tilde{P}(H) : H \in \mathcal{H}_{U,V,z},\ \text{his}(H) = A) =$

$1/|\mathcal{A}_{U,\mathrm{i},V,z}|$), and then all *histories* with the same *histogram*, $\mathrm{his}(H) = A$, are uniformly probable. The *likely probability function*, $\tilde{P}$, is *known.*

In the case where there is a *distribution history* H_E, the *maximum likelihood estimate* which maximises the entropy, $\mathrm{entropy}(\tilde{P})$, is

$$\begin{aligned}
\tilde{P} &= \Big(\bigcup\{\{(H,1) : H \subseteq H_E,\ \mathrm{his}(H) = A\}^{\wedge} : A \in \mathcal{A}_{U,\mathrm{i},V,z}\}\Big)^{\wedge} \cup \\
&\quad \{(G,0) : G \in \mathcal{H}_{U,V,z},\ G \not\subseteq H_E\} \\
&= \Big(\bigcup\{\{(H, 1/Q_{\mathrm{h},U}(E,z)(A)) : H \subseteq H_E,\ \mathrm{his}(H) = A\} : \\
&\qquad\qquad A \in \mathcal{A}_{U,\mathrm{i},V,z}\}\Big)^{\wedge} \cup \\
&\quad \{(G,0) : G \in \mathcal{H}_{U,V,z},\ G \not\subseteq H_E\}
\end{aligned}$$

That is, the *maximum likelihood estimate*, $\tilde{P}$, is such that all *drawn histograms*, $A \leq E$, are uniformly probable, and then all *drawn histories* $H \subseteq H_E$ with the same *histogram*, $\mathrm{his}(H) = A$, are uniformly probable. If the *distribution histogram*, H_E, is *known*, then the *likely probability function*, $\tilde{P}$, is *known.*

Consider the case where a *drawn sample* A is *known*, but the *distribution histogram*, E, is *unknown* and hence the *likely history probability function*, $\tilde{P}$, is *unknown.* The *maximum likelihood estimate* $\tilde{E}$ for the *distribution histogram*, E, is the same as for *necessary histogram*,

$$\tilde{E} \in \mathrm{maxd}(\{(D, Q_{\mathrm{h},U}(D,z)(A)) : D \in \mathcal{A}_{U,\mathrm{i},V,z_E}\})$$

The *likely distribution histogram*, $\tilde{E}$, is *known* if the *distribution histogram size*, z_E, is *known* and the *histogram*, A, is *known.* If it is assumed that the *distribution histogram* equals the *likely distribution histogram*, $E = \tilde{E}$, then the *likely history probability* is *known*, $\tilde{P}(H) = 1/|\{B : B \in \mathcal{A}_{U,\mathrm{i},V,z},\ B \leq \tilde{E}\}| \times 1/Q_{\mathrm{h},U}(\tilde{E},z)(A)$ where $\mathrm{his}(H) = A$.

In the case where the *distribution histogram*, E, is *unknown*, and the *distribution histogram size*, z_E, is also *unknown*, except that it is *known* to be large, $z_E \gg z$, then the *maximum likelihood estimate* $\tilde{E}$ for the *distribution probability histogram*, $\hat{E}$, may be approximated by a modal value of a *likelihood function* which depends on the *multinomial distribution* instead,

$$\tilde{E} \in \mathrm{maxd}(\{(D, Q_{\mathrm{m},U}(D,z)(A)) : D \in \mathcal{A}_{U,V,1}\})$$

Again, the *maximum likelihood estimate*, $\tilde{E}$, for the *distribution probability histogram*, $\hat{E}$, is the *sample probability histogram*, $\hat{A}$,

$$\tilde{E} \ = \ \hat{A}$$

If it is assumed that the *distribution probability histogram* equals the *likely distribution probability histogram*, $\hat{E} = \tilde{E} = \hat{A}$, then the *likely history probability* varies against the *sample-distributed multinomial probability*, $\tilde{P}(H) \sim 1/|\mathcal{A}_{U,\mathrm{i},V,z}| \times 1/\hat{Q}_{\mathrm{m},U}(\hat{A},z)(A)$.

So the properties of *uniform possible histogram* are similar to *necessary histogram* except that more *histories* are possible but less probable.

2.4.2 Aligned induction

In *aligned induction* the *history probability functions* are constrained by *independent histogram*.

The *independent histogram* valued function of *integral substrate histograms* $Y_{U,\mathrm{i},V,z}$ is defined

$$Y_{U,\mathrm{i},V,z} \ := \ \{(A, A^{\mathrm{X}}) : A \in \mathcal{A}_{U,\mathrm{i},V,z}\}$$

The finite set of *iso-independents* of *independent histogram* A^{X} is

$$Y^{-1}_{U,\mathrm{i},V,z}(A^{\mathrm{X}}) \ = \ \{B : B \in \mathcal{A}_{U,\mathrm{i},V,z},\ B^{\mathrm{X}} = A^{\mathrm{X}}\}$$

Given any subset of the *integral substrate histograms* $I \subseteq \mathcal{A}_{U,\mathrm{i},V,z}$ that contains the *histogram*, $A \in I$, the degree to which the subset is said to be *aligned-like* is called the *iso-independence*. The *iso-independence* is defined as the ratio of (i) the cardinality of the intersection between the *integral substrate histograms* subset and the set of *integral iso-independents*, and (ii) the cardinality of the union,

$$\frac{1}{|\mathcal{A}_{U,\mathrm{i},V,z}|} \ \leq \ \frac{|I \ \cap \ Y^{-1}_{U,\mathrm{i},V,z}(A^{\mathrm{X}})|}{|I \ \cup \ Y^{-1}_{U,\mathrm{i},V,z}(A^{\mathrm{X}})|} \ \leq \ 1$$

Consider the case where the *independent* A^{X} of *drawn histories* is *known* to be *necessary*, $\sum(P(H) : H \subseteq H_E,\ \mathrm{his}(H)^{\mathrm{X}} = A^{\mathrm{X}}) = 1$. The *maximum* *likelihood estimate* which maximises the entropy, $\text{entropy}(\tilde{P})$, is

$$\begin{aligned}
\tilde{P} &= \{(H,1) : H \subseteq H_E,\ \text{his}(H)^{\mathrm{X}} = A^{\mathrm{X}}\}^{\wedge}\ \cup \\
&\quad \{(G,0) : G \in \mathcal{H}_{U,V,z},\ G \not\subseteq H_E\}\ \cup \\
&\quad \{(G,0) : G \in \mathcal{H}_{U,V,z},\ \text{his}(G)^{\mathrm{X}} \neq A^{\mathrm{X}}\} \\
&= \{(H, 1/\textstyle\sum(Q_{\mathrm{h},U}(E,z)(B) : B \in Y^{-1}_{U,\mathrm{i},V,z}(A^{\mathrm{X}}))) : \\
&\qquad\qquad\qquad\qquad H \subseteq H_E,\ \text{his}(H)^{\mathrm{X}} = A^{\mathrm{X}}\}\ \cup \\
&\quad \{(G,0) : G \in \mathcal{H}_{U,V,z},\ G \not\subseteq H_E\}\ \cup \\
&\quad \{(G,0) : G \in \mathcal{H}_{U,V,z},\ \text{his}(G)^{\mathrm{X}} \neq A^{\mathrm{X}}\}
\end{aligned}$$

That is, the *maximum likelihood estimate*, $\tilde{P}$, is such that all *drawn histories* $H \subseteq H_E$ with the *independent*, $\text{his}(H)^{\mathrm{X}} = A^{\mathrm{X}}$, are uniformly probable and all other *histories*, $G \not\subseteq H_E$ or $\text{his}(G)^{\mathrm{X}} \neq A^{\mathrm{X}}$, are impossible, $\tilde{P}(G) = 0$. If the *independent*, A^{X}, is *known* and the *distribution histogram*, H_E, is *known*, then the *likely probability function*, $\tilde{P}$, is *known*.

The *likely probability* of *drawing histogram* A from *necessary drawn independent* A^{X} is

$$\sum(\tilde{P}(H) : H \in \mathcal{H}_{U,V,z},\ \text{his}(H) = A) = \frac{Q_{\mathrm{h},U}(E,z)(A)}{\sum Q_{\mathrm{h},U}(E,z)(B) : B \in Y^{-1}_{U,\mathrm{i},V,z}(A^{\mathrm{X}})}$$

The *likely history probability function* entropy, $\text{entropy}(\tilde{P})$, is maximised when the sum of the *iso-independent historical frequencies*, $\sum Q_{\mathrm{h},U}(E,z)(B) : B \in Y^{-1}_{U,\mathrm{i},V,z}(A^{\mathrm{X}})$, is maximised.

Consider the case where the *independent*, A^{X}, is *known*, but the *distribution histogram*, E, is *unknown* and hence the *likely history probability function*, $\tilde{P}$, is *unknown*. The *maximum likelihood estimate* $\tilde{E}$ for the *distribution histogram*, E, is a modal value of the *likelihood function*,

$$\tilde{E} \in \text{maxd}(\{(D, \textstyle\sum(Q_{\mathrm{h},U}(D,z)(B) : B \in Y^{-1}_{U,\mathrm{i},V,z}(A^{\mathrm{X}}))) : D \in \mathcal{A}_{U,\mathrm{i},V,z_E}\})$$

The *likely distribution histogram*, $\tilde{E}$, is *known* if the *distribution histogram size*, z_E, is *known* and the *independent*, A^{X}, is *known*. If it is assumed that the *distribution histogram* equals the *likely distribution histogram*, $E = \tilde{E}$, then the *likely history probability* is *known*, $\tilde{P}(H) = 1/\sum(Q_{\mathrm{h},U}(\tilde{E},z)(B) : B \in Y^{-1}_{U,\mathrm{i},V,z}(A^{\mathrm{X}}))$ where $\text{his}(H)^{\mathrm{X}} = A^{\mathrm{X}}$.

In the case where the *distribution histogram*, E, is *unknown*, and the *distribution histogram size*, z_E, is also *unknown*, except that it is *known* to be large, $z_E \gg z$, then the *maximum likelihood estimate* $\tilde{E}$ for the *distribution probability histogram*, $\hat{E}$, may be approximated by a modal value of a *likelihood function* which depends on the *multinomial distribution* instead,

$$\tilde{E} \;\in\; \mathrm{maxd}(\{(D, \sum(Q_{\mathrm{m},U}(D,z)(B) : B \in Y^{-1}_{U,\mathrm{i},V,z}(A^{\mathrm{X}}))) : D \in \mathcal{A}_{U,V,1}\})$$

which has a solution $\tilde{E} = \hat{A}^{\mathrm{X}}$. So the *maximum likelihood estimate*, $\tilde{E}$, for the *distribution probability histogram*, $\hat{E}$, is the *independent probability histogram*, $\hat{A}^{\mathrm{X}}$,

$$\tilde{E} \;=\; \hat{A}^{\mathrm{X}}$$

In the case where the *independent* is *integral*, $A^{\mathrm{X}} \in \mathcal{A}_{\mathrm{i}}$, the sum of the *iso-independent independent-distributed multinomial probabilities* varies with the *independent independent-distributed multinomial probability*,

$$\sum(Q_{\mathrm{m},U}(A^{\mathrm{X}},z)(B) : B \in Y^{-1}_{U,\mathrm{i},V,z}(A^{\mathrm{X}})) \;\sim\; Q_{\mathrm{m},U}(A^{\mathrm{X}},z)(A^{\mathrm{X}})$$

So, if it is assumed that the *distribution probability histogram* equals the *likely distribution probability histogram*, $\hat{E} = \tilde{E} = \hat{A}^{\mathrm{X}}$, then the *likely history probability* varies against the *independent-distributed multinomial probability* of the *independent*, $\tilde{P}(H) \sim 1/\hat{Q}_{\mathrm{m},U}(A^{\mathrm{X}},z)(A^{\mathrm{X}})$.

In this case, the *likely probability* of *drawing histogram* A from *necessary drawn independent* A^{X} is approximately

$$\begin{aligned}
&\sum(\tilde{P}(H) : H \in \mathcal{H}_{U,V,z},\ \mathrm{his}(H) = A) \\
&\qquad\approx\; \frac{Q_{\mathrm{m},U}(A^{\mathrm{X}},z)(A)}{\sum Q_{\mathrm{m},U}(A^{\mathrm{X}},z)(B) : B \in Y^{-1}_{U,\mathrm{i},V,z}(A^{\mathrm{X}})} \\
&\qquad\sim\; \frac{Q_{\mathrm{m},U}(A^{\mathrm{X}},z)(A)}{Q_{\mathrm{m},U}(A^{\mathrm{X}},z)(A^{\mathrm{X}})}
\end{aligned}$$

The negative logarithm of the ratio of the *histogram independent-distributed multinomial probability* to the *independent independent-distributed multinomial probability* equals the *alignment*,

$$-\ln \frac{Q_{\mathrm{m},U}(A^{\mathrm{X}},z)(A)}{Q_{\mathrm{m},U}(A^{\mathrm{X}},z)(A^{\mathrm{X}})} \;=\; \mathrm{algn}(A)$$

So the logarithm of the *likely probability* of *drawing histogram* A from *necessary drawn independent* A^{X} varies against the *alignment*,

$$\ln \sum (\tilde{P}(H) : H \in \mathcal{H}_{U,V,z},\ \mathrm{his}(H) = A) \quad \sim \quad -\ \mathrm{algn}(A)$$

The *independent*, A^{X}, which has zero *alignment*, $\mathrm{algn}(A^{\mathrm{X}}) = 0$, is the most *probable histogram*, $\forall B \in Y^{-1}_{U,\mathrm{i},V,z}(A^{\mathrm{X}})\ (Q_{\mathrm{m},U}(A^{\mathrm{X}}, z)(A^{\mathrm{X}}) \geq Q_{\mathrm{m},U}(A^{\mathrm{X}}, z)(B))$. As the *alignment* increases, $\mathrm{algn}(A) > 0$, the *likely histogram probability*, $Q_{\mathrm{m},U}(A^{\mathrm{X}}, z)(A)/\sum(Q_{\mathrm{m},U}(A^{\mathrm{X}}, z)(B) : B \in Y^{-1}_{U,\mathrm{i},V,z}(A^{\mathrm{X}}))$, decreases.

The *likely history probability function* entropy varies with the *independent entropy*, $\mathrm{entropy}(\tilde{P}) \sim \mathrm{entropy}(A^{\mathrm{X}})$.

Define the *dependent histogram* $A^{\mathrm{Y}} \in \mathcal{A}_{U,V,z}$ as the *maximum likelihood estimate* of the *distribution histogram* of the *multinomial probability* of the *histogram* A conditional that it is an *iso-independent*,

$$\{A^{\mathrm{Y}}\} \quad = \quad \mathrm{maxd}(\{(D, \frac{Q_{\mathrm{m},U}(D, z)(A)}{\sum Q_{\mathrm{m},U}(D, z)(B) : B \in Y^{-1}_{U,\mathrm{i},V,z}(A^{\mathrm{X}})}) : D \in \mathcal{A}_{U,V,z}\})$$

Note that the *dependent*, A^{Y}, is not always computable, but an approximation to any accuracy can be made to it. In the case where the *histogram* is *independent*, the *dependent* equals the *independent*, $A = A^{\mathrm{X}} \implies A^{\mathrm{Y}} = A = A^{\mathrm{X}}$. The *dependent alignment* is greater than or equal to the *histogram alignment*, $\mathrm{algn}(A^{\mathrm{Y}}) \geq \mathrm{algn}(A) \geq \mathrm{algn}(A^{\mathrm{X}}) = 0$. In the case where the *histogram* is *uniformly diagonalised*, the *histogram alignment*, $\mathrm{algn}(A)$, is at the maximum, and the *dependent* equals the *histogram*, $A^{\mathrm{Y}} = A$.

Now consider the case where, given *necessary drawn independent* A^{X}, it is *known*, in addition, that the *sample histogram* A is the most *probable histogram*, regardless of its *alignment*. That is, the *likely probability* of *drawing histogram* A from *necessary drawn independent* A^{X},

$$\sum (\tilde{P}(H) : H \in \mathcal{H}_{U,V,z},\ \mathrm{his}(H) = A) \ = \ \frac{Q_{\mathrm{h},U}(E, z)(A)}{\sum Q_{\mathrm{h},U}(E, z)(B) : B \in Y^{-1}_{U,\mathrm{i},V,z}(A^{\mathrm{X}})}$$

is maximised.

In the case where the *sample*, A, is *known*, but the *distribution histogram*,

E, is *unknown*, the *maximum likelihood estimate* $\tilde{E}$ for the *distribution histogram*, E, is a modal value of the *likelihood function*,

$$\tilde{E} \in \text{maxd}(\{(D, \frac{Q_{\text{h},U}(D,z)(A)}{\sum Q_{\text{h},U}(D,z)(B) : B \in Y^{-1}_{U,\text{i},V,z}(A^{\text{X}})}) : D \in \mathcal{A}_{U,\text{i},V,z_E}\})$$

The *likely distribution histogram*, $\tilde{E}$, is *known* if the *distribution histogram size*, z_E, is *known* and the *sample*, A, is *known*. If it is assumed that the *distribution histogram* equals the *likely distribution histogram*, $E = \tilde{E}$, then the *likely history probability* is *known*, $\tilde{P}(H) = 1/\sum(Q_{\text{h},U}(\tilde{E},z)(B) : B \in Y^{-1}_{U,\text{i},V,z}(A^{\text{X}}))$ where $\text{his}(H)^{\text{X}} = A^{\text{X}}$.

If the *histogram* is *independent*, $A = A^{\text{X}}$, then the additional constraint of *probable sample* makes no change to the *maximum likelihood estimate*, $\tilde{E}$,

$$\begin{aligned} A = A^{\text{X}} \implies & \\ & \text{maxd}(\{(D, \frac{Q_{\text{h},U}(D,z)(A)}{\sum Q_{\text{h},U}(D,z)(B) : B \in Y^{-1}_{U,\text{i},V,z}(A^{\text{X}})}) : D \in \mathcal{A}_{U,\text{i},V,z_E}\}) \\ = & \text{maxd}(\{(D, \sum(Q_{\text{h},U}(D,z)(B) : B \in Y^{-1}_{U,\text{i},V,z}(A^{\text{X}}))) : D \in \mathcal{A}_{U,\text{i},V,z_E}\}) \end{aligned}$$

If the *histogram* is not *independent*, $\text{algn}(A) > 0$, however, then the *likely history probability function* entropy, $\text{entropy}(\tilde{P})$, is lower than it is in the case of *necessary independent* unconstrained by *probable sample*.

In the case where the *distribution histogram*, E, is *unknown*, and the *distribution histogram size*, z_E, is also *unknown*, except that it is *known* to be large, $z_E \gg z$, then the *maximum likelihood estimate* $\tilde{E}$ for the *distribution probability histogram*, $\hat{E}$, is now approximated by a modal value of the conditional *likelihood function*,

$$\tilde{E} \in \text{maxd}(\{(D, \frac{Q_{\text{m},U}(D,z)(A)}{\sum Q_{\text{m},U}(D,z)(B) : B \in Y^{-1}_{U,\text{i},V,z}(A^{\text{X}})}) : D \in \mathcal{A}_{U,V,1}\})$$

The solution to this is the *normalised dependent*, $\tilde{E} = \hat{A}^{\text{Y}}$. The *maximum likelihood estimate* is near the *sample*, $\tilde{E} \sim \hat{A}$, only in as much as it is far from the *independent*, $\tilde{E} \nsim \hat{A}^{\text{X}}$. This may be compared to the case unconstrained by *probable sample* where the *maximum likelihood estimate* equals the *independent*, $\tilde{E} = \hat{A}^{\text{X}}$. In the *probable sample* case the *sized maximum likelihood estimate* is *aligned*, $\text{algn}(A^{\text{Y}}) > 0$, so there are fewer ways to *draw* the *iso-independents* and the *likely history probability function* entropy, $\text{entropy}(\tilde{P})$, is lower. At maximum *alignment*, where the *histogram* is *uniformly diagonalised*, the *dependent* equals the *histogram*, $A^{\text{Y}} = A$, and the *likely history* *probability function* entropy, $\mathrm{entropy}(\tilde{P})$, is least.

The *iso-independent conditional multinomial probability distribution* is defined,

$$\hat{Q}_{\mathrm{m,y},U}(E,z)(A) \ := \ \frac{1}{|\mathrm{ran}(Y_{U,\mathrm{i},V,z})|} \frac{Q_{\mathrm{m},U}(E,z)(A)}{\sum Q_{\mathrm{m},U}(E,z)(B) : B \in Y^{-1}_{U,\mathrm{i},V,z}(A^{\mathrm{X}})}$$

So the optimisation can be rewritten,

$$\tilde{E} \ \in \ \mathrm{maxd}(\{(D, \hat{Q}_{\mathrm{m,y},U}(D,z)(A)) : D \in \mathcal{A}_{U,V,1}\})$$

The logarithm of the *independent-distributed iso-independent conditional multinomial probability* varies against the *alignment*,

$$\ln \frac{Q_{\mathrm{m},U}(A^{\mathrm{X}},z)(A)}{\sum Q_{\mathrm{m},U}(A^{\mathrm{X}},z)(B) : B \in Y^{-1}_{U,\mathrm{i},V,z}(A^{\mathrm{X}})} \ \sim \ -\ \mathrm{algn}(A)$$

Conversely, the logarithm of the *dependent-distributed iso-independent conditional multinomial probability* varies with the *alignment*,

$$\ln \frac{Q_{\mathrm{m},U}(A^{\mathrm{Y}},z)(A)}{\sum Q_{\mathrm{m},U}(A^{\mathrm{Y}},z)(B) : B \in Y^{-1}_{U,\mathrm{i},V,z}(A^{\mathrm{X}})} \ \sim \ \mathrm{algn}(A)$$

That is, the *log-likelihood* varies with the *sample alignment*,

$$\ln \hat{Q}_{\mathrm{m,y},U}(A^{\mathrm{Y}},z)(A) \ \sim \ \mathrm{algn}(A)$$

In the case where the *alignment* is low the *sum sensitivity* varies with the *alignment*

$$\mathrm{sum}(\mathrm{sensitivity}(U)(\hat{Q}_{\mathrm{m,y},U}(A^{\mathrm{Y}},z))) \ \sim \ \mathrm{algn}(A)$$

and in the case where the *alignment* is high the *sum sensitivity* varies against the *alignment*

$$\mathrm{sum}(\mathrm{sensitivity}(U)(\hat{Q}_{\mathrm{m,y},U}(A^{\mathrm{Y}},z))) \ \sim \ -\ \mathrm{algn}(A)$$

At intermediate *alignments* the *sum sensitivity* is independent of the *alignment.*

So, in the *probable sample* case, if it is assumed that the *distribution probability histogram* equals the *likely distribution probability histogram*, $\hat{E} = \tilde{E} = \hat{A}^{\mathrm{Y}}$, then the *likely history probability function* entropy varies against the *alignment*, $\text{entropy}(\tilde{P}) \sim -\text{algn}(A)$.

As the *alignment*, $\text{algn}(A)$, increases towards its maximum, the *likely distribution probability histogram* tends to the *histogram*, $\tilde{E} = \hat{A}^{\mathrm{Y}} \sim \hat{A}$, and the *log-likelihood*, $\ln \hat{Q}_{\mathrm{m,y},U}(A^{\mathrm{Y}}, z)(A)$, increases, but the *sensitivity* to *distribution histogram*, E, decreases. In other words, the more *aligned* the *sample* the more *likely* the *normalised sample histogram*, $\hat{A}$, equals the *normalised distribution histogram*, $\hat{E}$, and the smaller the *likely* difference between them if they are not equal.

Consider the case where the *independent* is *uniformly possible*. Assume that the *substrate history probability function* $P \in (\mathcal{H}_{U,V,z} :\to \mathbf{Q}_{\geq 0}) \cap \mathcal{P}$ is the distribution of an arbitrary *history* valued function, $\mathcal{X} \to \mathcal{H}$, given an arbitrary *independent* valued function, $\mathcal{X} \to \mathcal{A}$. In this case, the *history* valued function is chosen arbitrarily from the constrained subset

$$\begin{aligned}&\{G : F \in \mathcal{X} \to (\mathcal{A} \times (\mathcal{X} \to \mathcal{H})),\\ &\qquad (\cdot, (A, G)) \in F,\ \forall (\cdot, H) \in G\ (\text{his}(H)^{\mathrm{X}} = A)\} \ \subset\ \mathcal{X} \to \mathcal{H}\end{aligned}$$

Uniformly possible independent is a weaker constraint than *uniformly possible histogram*, so the subset of *history* valued functions is larger.

In the case where there is a *distribution history* H_E, the *maximum likelihood estimate* which maximises the entropy, $\text{entropy}(\tilde{P})$, is

$$\begin{aligned}\tilde{P} &= \Big(\bigcup\{\{(H,1) : H \subseteq H_E,\ \text{his}(H)^{\mathrm{X}} = A\}^{\wedge} : A \in \text{ran}(Y_{U,\mathrm{i},V,z})\}\Big)^{\wedge}\ \cup\\ &\quad\ \{(G,0) : G \in \mathcal{H}_{U,V,z},\ G \nsubseteq H_E\}\\ &= \Big(\bigcup\{\{(H, 1/\textstyle\sum(Q_{\mathrm{h},U}(E,z)(B) : B \in Y^{-1}_{U,\mathrm{i},V,z}(A^{\mathrm{X}}))) :\\ &\qquad\qquad H \subseteq H_E,\ \text{his}(H)^{\mathrm{X}} = A\}\ :\ A \in \text{ran}(Y_{U,\mathrm{i},V,z})\}\Big)^{\wedge}\ \cup\\ &\quad\ \{(G,0) : G \in \mathcal{H}_{U,V,z},\ G \nsubseteq H_E\}\end{aligned}$$

That is, the *maximum likelihood estimate*, $\tilde{P}$, is such that all *drawn independents* are uniformly probable, and then all *drawn histories* $H \subseteq H_E$ with the same *independent*, $\text{his}(H)^{\mathrm{X}} = A$, are uniformly probable. If the *distribution histogram*, H_E, is *known*, then the *likely probability function*, $\tilde{P}$, is *known*.

The properties of *uniformly possible independent* are the same as for *necessary independent*, except that the probabilities are scaled. So, in the case where the *distribution histogram*, E, is *unknown*, and the *distribution histogram size*, z_E, is also *unknown*, except that it is *known* to be large, $z_E \gg z$, then the *likely history probability* varies against the *independent-distributed multinomial probability* of the *independent*,

$$\tilde{P}(H) \quad \sim \quad 1/|\mathrm{ran}(Y_{U,\mathrm{i},V,z})| \times 1/\hat{Q}_{\mathrm{m},U}(A^{\mathrm{X}}, z)(A^{\mathrm{X}})$$

That is, more *histories* are possible but less probable.

2.5 Models

2.5.1 Transforms

Transforms are the simplest *models*. All *models* can be converted to *transforms*.

Given a *histogram* $X \in \mathcal{A}$ and a subset of its *variables* $W \subseteq \mathrm{vars}(X)$, the pair $T = (X, W)$ forms a *transform*. The *variables*, W, are the *derived variables*. The complement $V = \mathrm{vars}(X) \setminus W$ are the *underlying variables*. The set of all *transforms* is

$$\mathcal{T} \quad := \quad \{(X, W) : X \in \mathcal{A},\ W \subseteq \mathrm{vars}(X)\}$$

The *transform histogram* is $X = \mathrm{his}(T)$. The *transform derived* is $W = \mathrm{der}(T)$. The *transform underlying* is $V = \mathrm{und}(T)$. The set of *underlying variables* of a *transform* is also called the *substrate*.

The *null transform* is $(X, \emptyset)$. The *full transform* is $(X, \mathrm{vars}(X))$.

Given a *histogram* $A \in \mathcal{A}$, the *multiplication* of the *histogram*, A, by the *transform* $T \in \mathcal{T}$ equals the *multiplication* of the *histogram*, A, by the *transform histogram* $X = \mathrm{his}(T)$ followed by the *reduction* to the *derived variables* $W = \mathrm{der}(T)$,

$$A * T \; = \; A * (X, W) \quad := \quad A * X \;\%\; W$$

If the *histogram variables* are a superset of the *underlying variables*, $\mathrm{vars}(A) \supseteq \mathrm{und}(T)$, then the *histogram*, A, is called the *underlying histogram* and the *multiplication*, $A * T$, is called the *derived histogram*. The *derived histogram variables* equals the *derived variables*, $\mathrm{vars}(A * T) = \mathrm{der}(T)$.

The application of the *null transform* of the *cartesian* is the *scalar*, $A * (V^{\mathrm{C}}, \emptyset) = A\%\emptyset = \mathrm{scalar}(\mathrm{size}(A))$, where $V = \mathrm{vars}(A)$. The application of the *full transform* of the *cartesian* is the *histogram*, $A * (V^{\mathrm{C}}, V) = A\%V = A$.

Given a *histogram* $A \in \mathcal{A}$ and a *transform* $T \in \mathcal{T}$, the *formal histogram* is defined as the *independent derived*, $A^{\mathrm{X}} * T$. The *abstract histogram* is defined as the *derived independent*, $(A * T)^{\mathrm{X}}$.

In the case where the *formal* and *abstract* are equal, $A^{\mathrm{X}} * T = (A * T)^{\mathrm{X}}$, the *abstract* equals the *independent abstract*, $(A*T)^{\mathrm{X}} = A^{\mathrm{X}}*T = (A^{\mathrm{X}}*T)^{\mathrm{X}}$, and so only depends on the *independent*, A^{X}, not on the *histogram*, A. The *formal* equals the *formal independent*, $A^{\mathrm{X}} * T = (A * T)^{\mathrm{X}} = (A^{\mathrm{X}} * T)^{\mathrm{X}}$, and so is itself *independent*.

A *transform* $T \in \mathcal{T}$ is *functional* if there is a *causal* relation between the *underlying variables* $V = \mathrm{und}(T)$ and the *derived variables* $W = \mathrm{der}(T)$,

$$\mathrm{split}(V, X^{\mathrm{FS}}) \in V^{\mathrm{CS}} \to W^{\mathrm{CS}}$$

where $X = \mathrm{his}(T)$. The set of *functional transforms* $\mathcal{T}_{\mathrm{f}} \subset \mathcal{T}$ is the subset of all *transforms* that are *causal*.

A *functional transform* $T \in \mathcal{T}_{\mathrm{f}}$ has an *inverse*,

$$T^{-1} \ := \ \{((S\%V, c), S\%W) : (S, c) \in X\}^{-1}$$

A *transform* T is *one functional* in *system* U if the *reduction* of the *transform histogram* to the *underlying variables* equals the *cartesian histogram*, $X\%V = V^{\mathrm{C}}$. So the *causal* relation is a *derived state* valued left total function of *underlying state*, $\mathrm{split}(V, X^{\mathrm{S}}) \in V^{\mathrm{CS}} :\to W^{\mathrm{CS}}$. The set of *one functional transforms* $\mathcal{T}_{U,\mathrm{f},1} \subset \mathcal{T}_{\mathrm{f}}$ is

$$\begin{aligned}\mathcal{T}_{U,\mathrm{f},1} \ &= \ \{(\{(S \cup R, 1) : (S, R) \in Q\}, W) : \\ &\qquad V, W \subseteq \mathrm{vars}(U),\ V \cap W = \emptyset,\ Q \in V^{\mathrm{CS}} :\to W^{\mathrm{CS}}\}\end{aligned}$$

The application of a *one functional transform* to an *underlying histogram* preserves the *size*, $\mathrm{size}(A * T) = \mathrm{size}(A)$.

The *one functional transform inverse* is a *unit component* valued function of *derived state*, $T^{-1} \in W^{\mathrm{CS}} \to \mathrm{P}(V^{\mathrm{C}})$. That is, the range of the *inverse* corresponds to a *partition* of the *cartesian states* into *components*, $\mathrm{ran}(T^{-1}) \in \mathrm{B}(V^{\mathrm{C}})$.

The application of a *one functional transform* T to its *underlying cartesian* V^{C} is the *component cardinality histogram*, $V^{\mathrm{C}} * T = \{(R, |C|) : (R, C) \in T^{-1}\}$. The *effective cartesian derived volume* is less than or equal to the *derived volume*, $|(V^{\mathrm{C}} * T)^{\mathrm{F}}| = |T^{-1}| \leq |W^{\mathrm{C}}|$.

A *one functional transform* $T \in \mathcal{T}_{U,\mathrm{f},1}$ may be applied to a *history* $H \in \mathcal{H}$ in the *underlying variables* of the *transform*, $\mathrm{vars}(H) = \mathrm{und}(T)$, to construct a *derived history*,

$$H * T \;\; := \;\; \{(x,R) : (x,S) \in H,\; \{R\} = (\{S\}^{\mathrm{U}} * T)^{\mathrm{FS}}\}$$

The *size* is unchanged, $|H*T| = |H|$, and the *event identifiers* are conserved, $\mathrm{dom}(H * T) = \mathrm{dom}(H)$.

Given a *partition* $P \in \mathrm{B}(V^{\mathrm{CS}})$ of the *cartesian states* of *variables* V, a *one functional transform* can be constructed. The *partition transform* is

$$P^{\mathrm{T}} \;\; := \;\; (\{(S \cup \{(P,C)\},1) : C \in P,\; S \in C\},\{P\})$$

The set of *derived variables* of the *partition transform* is a singleton of the *partition variable*, $\mathrm{der}(P^{\mathrm{T}}) = \{P\}$. The *derived volume* is the *component* cardinality, $|\{P\}^{\mathrm{C}}| = |P|$. The *underlying variables* are the given *variables*, $\mathrm{und}(P^{\mathrm{T}}) = V$.

The *unary partition transform* is $T_{\mathrm{u}} = \{V^{\mathrm{CS}}\}^{\mathrm{T}}$. The *self partition transform* is $T_{\mathrm{s}} = V^{\mathrm{CS}\{\}\mathrm{T}}$.

Given a *one functional transform* $T \in \mathcal{T}_{U,\mathrm{f},1}$, the *natural converse* is

$$T^{\dagger} \;\; := \;\; (X/(X\%W),V)$$

where $(X,W) = T$ and $V = \mathrm{und}(T)$. The *natural converse* may be expressed in terms of *components*,

$$T^{\dagger} \; := \; (\sum_{(R,C) \in T^{-1}} \{R\}^{\mathrm{U}} * \hat{C},\; V)$$

Given a *histogram* $A \in \mathcal{A}$ in the *underlying variables*, $\mathrm{vars}(A) = V$, the *naturalisation* is the application of the *natural converse transform* to the *derived histogram*, $A * T * T^{\dagger}$. The *naturalisation* can be rewritten $A * X \;\%\; W * X \;/\; (X\%W) \;\%\; V$. The *naturalisation* is in the *underlying variables*, $\mathrm{vars}(A * T * T^{\dagger}) = V$. The *size* is conserved, $\mathrm{size}(A * T * T^{\dagger}) = \mathrm{size}(A)$. The *naturalisation derived* equals the *derived*, $A * T * T^{\dagger} * T = A * T$.

The *naturalisation* equals the *sum* of the *scaled components*, $A * T * T^{\dagger} = \sum \mathrm{scalar}((A * T)_R) * \hat{C} \;:\; (R,C) \in T^{-1}$. So each *component* is *uniform*, $\forall (R,C) \in T^{-1}\; (|\mathrm{ran}(A * T * T^{\dagger} * C)| = 1)$.

The *naturalisation* of the *unary partition transform*, $T_{\mathrm{u}} = \{V^{\mathrm{CS}}\}^{\mathrm{T}}$, is the *sized cartesian*, $A * T_{\mathrm{u}} * T_{\mathrm{u}}^{\dagger} = V_z^{\mathrm{C}}$, where $z = \mathrm{size}(A)$. The *naturalisation* of the *self partition transform*, $T_{\mathrm{s}} = V^{\mathrm{CS}\{\}\mathrm{T}}$, is the *histogram*, $A * T_{\mathrm{s}} * T_{\mathrm{s}}^{\dagger} = A$.

A *histogram* is *natural* when it equals its *naturalisation*, $A = A * T * T^{\dagger}$. The *cartesian* is *natural*, $V^{\mathrm{C}} = V^{\mathrm{C}} * T * T^{\dagger}$.

Given a *one functional transform* $T \in \mathcal{T}_{U,\mathrm{f},1}$ with *underlying variables* $V = \mathrm{und}(T)$, and a *histogram* $A \in \mathcal{A}$ in the same *variables*, $\mathrm{vars}(A) = V$, the *sample converse* is

$$(\hat{A} * X, V)$$

where $X = \mathrm{his}(T)$.

Related to the *sample converse*, the *actual converse* is defined as the *summed normalised* application of the *components* to the *sample histogram*,

$$T^{\odot A} \ := \ (\sum_{(R,C) \in T^{-1}} \{R\}^{\mathrm{U}} * (A * C)^{\wedge},\ V)$$

The application of the *actual converse transform* to the *derived histogram* equals the *histogram*, $A * T * T^{\odot A} = A$.

Given a *one functional transform* $T \in \mathcal{T}_{U,\mathrm{f},1}$ with *underlying variables* $V = \mathrm{und}(T)$, and a *histogram* $A \in \mathcal{A}$ in the same *variables*, $\mathrm{vars}(A) = V$, the *independent converse* is defined as the *summed normalised independent* application of the *components* to the *sample histogram*,

$$T^{\dagger A} \ := \ (\sum_{(R,C) \in T^{-1}} \{R\}^{\mathrm{U}} * (A * C)^{\wedge\mathrm{X}},\ V)$$

The *idealisation* is the application of the *independent converse transform* to the *derived histogram*, $A * T * T^{\dagger A}$. The *idealisation* is in the *underlying variables*, $\mathrm{vars}(A * T * T^{\dagger A}) = V$. The *size* is conserved, $\mathrm{size}(A * T * T^{\dagger A}) = \mathrm{size}(A)$. The *idealisation derived* equals the *derived*, $A * T * T^{\dagger A} * T = A * T$.

The *idealisation* equals the *sum* of the *independent components*, $A * T * T^{\dagger A} = \sum (A * C)^{\mathrm{X}} : (R, C) \in T^{-1}$. So each *component* is *independent*, $\forall (R, C) \in T^{-1}\ (A * T * T^{\dagger A} * C = (A * T * T^{\dagger A} * C)^{\mathrm{X}} = (A * C)^{\mathrm{X}})$.

The *idealisation* of the *unary partition transform*, $T_{\mathrm{u}} = \{V^{\mathrm{CS}}\}^{\mathrm{T}}$, is the *sized cartesian*, $A * T_{\mathrm{u}} * T_{\mathrm{u}}^{\dagger A} = V_z^{\mathrm{C}}$. The *idealisation* of the *self partition transform*, $T_{\mathrm{s}} = V^{\mathrm{CS}\{\}\mathrm{T}}$, is the *histogram*, $A * T_{\mathrm{s}} * T_{\mathrm{s}}^{\dagger A} = A$.

The *idealisation independent* equals the *independent*, $(A * T * T^{\dagger A})^{\mathrm{X}} = A^{\mathrm{X}}$. The *idealisation formal* equals the *formal*, $(A * T * T^{\dagger A})^{\mathrm{X}} * T = A^{\mathrm{X}} * T$. The *idealisation abstract* equals the *abstract*, $(A * T * T^{\dagger A} * T)^{\mathrm{X}} = (A * T)^{\mathrm{X}}$.

A *histogram* is *ideal* when it equals its *idealisation*, $A = A * T * T^{\dagger A}$.

The sense in which a *transform* is a simple *model* can be seen by considering queries on a *sample histogram*. Let *histogram* A have a set of *variables* $V = \mathrm{vars}(A)$ which is partitioned into query *variables* $K \subset V$ and label *variables* $V \setminus K$. Let $T = (X, W)$ be a *one functional transform* having *underlying variables* equal to the query *variables*, $\mathrm{und}(T) = K$. Given a query *state* $Q \in K^{\mathrm{CS}}$ that is *ineffective* in the *sample*, $Q \notin (A\%K)^{\mathrm{FS}}$, but is *effective* in the *sample derived*, $R \in (A * T)^{\mathrm{FS}}$ where $\{R\} = (\{Q\}^{\mathrm{U}} * T)^{\mathrm{FS}}$, the *probability histogram* for the label is

$$(\{Q\}^{\mathrm{U}} * T * (\hat{A} * X, V))^{\wedge} \ \% \ (V \setminus K) \ \in \ \mathcal{A} \cap \mathcal{P}$$

where the *sample converse transform* is $(\hat{A} * X, V)$. This can be expressed more simply in terms of the *actual converse*,

$$\{Q\}^{\mathrm{U}} * T * T^{\odot A} \ \% \ (V \setminus K) \ \in \ \mathcal{A} \cap \mathcal{P}$$

The query of the *sample* via *model* can also be written without the *transforms*, $(\{Q\}^{\mathrm{U}} * X \ \% \ W * X * A)^{\wedge} \ \% \ (V \setminus K)$. The query *state*, Q, in the query *variables*, K, is raised to the query *derived state*, R, in the *derived variables*, W, then lowered to *effective sample states*, in the *sample variables*, V, and finally *reduced* to label *states*, in the label *variables*, $V \setminus K$. Even though the *sample* itself does not contain the query, $\{Q\}^{\mathrm{U}} * \hat{A} = \emptyset$, the *sample derived* does contain the query *derived*, $\{R\}^{\mathrm{U}} * (\hat{A} * T) \neq \emptyset$, and so the resultant labels are those of the corresponding *effective component*, $(A * C)^{\wedge} \ \% \ (V \setminus K)$, where $(R, C) \in T^{-1}$.

The set of *substrate histories* $\mathcal{H}_{U,V,z}$ is defined above as the set of *histories* having *event identifiers* $\{1 \ldots z\}$, fixed *size* z and fixed *variables* V,

$$\mathcal{H}_{U,V,z} \ := \ \{1 \ldots z\} :\to V^{\mathrm{CS}}$$

The corresponding set of *integral substrate histograms* $\mathcal{A}_{U,\mathrm{i},V,z}$ is the set of *complete integral histograms* in *variables* V with *size* z,

$$\mathcal{A}_{U,\mathrm{i},V,z} \quad := \quad \{A : A \in V^{\mathrm{CS}} :\to \{0 \ldots z\},\ \mathrm{size}(A) = z\}$$

The set of *substrate transforms* $\mathcal{T}_{U,V}$ is the subset of *one functional transforms*, $\mathcal{T}_{U,V} \subset \mathcal{T}_{U,\mathrm{f},1}$, that have *underlying variables* V and *derived variables* which are *partitions*,

$$\mathcal{T}_{U,V} \quad = \quad \{(\prod_{(X,\cdot) \in F} X, \bigcup_{(\cdot,W) \in F} W) : F \subseteq \{P^{\mathrm{T}} : P \in \mathrm{B}(V^{\mathrm{CS}})\}\}$$

Let v be the *volume* of the *substrate*, $v = |V^{\mathrm{C}}|$. The cardinality of the *substrate transforms set* is $|\mathcal{T}_{U,V}| = 2^{\mathrm{bell}(v)}$, where $\mathrm{bell}(n)$ is Bell's number, which has factorial computation complexity. If the *volume*, v, is finite, the set of *substrate transforms* is finite, $|\mathcal{T}_{U,V}| < \infty$.

2.5.2 Transform entropy

Let T be a *one functional transform*, $T \in \mathcal{T}_{U,\mathrm{f},1}$, having *underlying variables* $V = \mathrm{und}(T)$. Let A be a *histogram*, $A \in \mathcal{A}$, in the *underlying variables*, $\mathrm{vars}(A) = V$, having *size* $z = \mathrm{size}(A) > 0$. The *underlying volume* is $v = |V^{\mathrm{C}}|$. The *derived volume* is $w = |T^{-1}|$.

The *derived entropy* or *component size entropy* is

$$\mathrm{entropy}(A * T) \quad := \quad -\sum_{(R,\cdot) \in T^{-1}} (\hat{A} * T)_R \times \ln\ (\hat{A} * T)_R$$

The *derived entropy* is positive and less than or equal to the logarithm of the *derived volume*, $0 \leq \mathrm{entropy}(A * T) \leq \ln w$.

Complementary to the *derived entropy* is the *expected component entropy*,

$$\begin{aligned}
\mathrm{entropyComponent}(A, T) \quad &:= \quad \sum_{(R,C) \in T^{-1}} (\hat{A} * T)_R \times \mathrm{entropy}(A * C) \\
&= \quad \sum_{(R,\cdot) \in T^{-1}} (\hat{A} * T)_R \times \mathrm{entropy}(\{R\}^{\mathrm{U}} * T^{\odot A})
\end{aligned}$$

The *cartesian derived entropy* or *component cardinality entropy* is

$$\mathrm{entropy}(V^{\mathrm{C}} * T) \quad := \quad -\sum_{(R,\cdot) \in T^{-1}} (\hat{V}^{\mathrm{C}} * T)_R \times \ln\ (\hat{V}^{\mathrm{C}} * T)_R$$

The *cartesian derived entropy* is positive and less than or equal to the logarithm of the *derived volume*, $0 \leq \text{entropy}(V^{\mathrm{C}} * T) \leq \ln w$.

The *cartesian derived derived sum entropy* or *component size cardinality sum entropy* is

$$\text{entropy}(A * T) + \text{entropy}(V^{\mathrm{C}} * T)$$

The *component size cardinality cross entropy* is the negative *derived histogram expected normalised cartesian derived count logarithm*,

$$\text{entropyCross}(A * T, V^{\mathrm{C}} * T) \quad := \quad - \sum_{(R,\cdot) \in T^{-1}} (\hat{A} * T)_R \times \ln \ (\hat{V}^{\mathrm{C}} * T)_R$$

The *component size cardinality cross entropy* is greater than or equal to the *derived entropy*, $\text{entropyCross}(A * T, V^{\mathrm{C}} * T) \geq \text{entropy}(A * T)$.

The *component cardinality size cross entropy* is the negative *cartesian derived expected normalised derived histogram count logarithm*,

$$\text{entropyCross}(V^{\mathrm{C}} * T, A * T) \quad := \quad - \sum_{(R,\cdot) \in T^{-1}} (\hat{V}^{\mathrm{C}} * T)_R \times \ln \ (\hat{A} * T)_R$$

The *component cardinality size cross entropy* is greater than or equal to the *cartesian derived entropy*, $\text{entropyCross}(V^{\mathrm{C}} * T, A * T) \geq \text{entropy}(V^{\mathrm{C}} * T)$.

The *component size cardinality sum cross entropy* is,

$$\text{entropy}(A * T + V^{\mathrm{C}} * T)$$

The *component size cardinality sum cross entropy* is positive and less than or equal to the logarithm of the *derived volume*, $0 \leq \text{entropy}(A * T + V^{\mathrm{C}} * T) \leq \ln w$.

In all cases the *cross entropy* is maximised when high *size components* are low *cardinality components*, $(\hat{A} * T)_R \gg (\hat{V}^{\mathrm{C}} * T)_R$ or $\text{size}(A * C)/z \gg |C|/v$, and low *size components* are high *cardinality components*, $(\hat{A} * T)_R \ll (\hat{V}^{\mathrm{C}} * T)_R$ or $\text{size}(A * C)/z \ll |C|/v$, where $(R, C) \in T^{-1}$.

The *cross entropy* is minimised when the *normalised derived histogram* equals the *normalised cartesian derived*, $\hat{A} * T = \hat{V}^{\mathrm{C}} * T$ or $\forall (R, C) \in T^{-1}$ $(\text{size}(A * C)/z = |C|/v)$. In this case the *cross entropy* equals the corresponding *component entropy*.

The *component size cardinality relative entropy* is the *component size cardinality cross entropy* minus the *component size entropy*,

$$\begin{aligned}
&\text{entropyRelative}(A * T, V^{\mathrm{C}} * T) \\
&\qquad := \sum_{(R,\cdot) \in T^{-1}} (\hat{A} * T)_R \times \ln \frac{(\hat{A} * T)_R}{(\hat{V}^{\mathrm{C}} * T)_R} \\
&\qquad = \text{entropyCross}(A * T, V^{\mathrm{C}} * T) \ - \ \text{entropy}(A * T)
\end{aligned}$$

The *component size cardinality relative entropy* is positive, $\text{entropyRelative}(A * T, V^{\mathrm{C}} * T) \geq 0$.

The *component cardinality size relative entropy* is the *component cardinality size cross entropy* minus the *component cardinality entropy*,

$$\begin{aligned}
&\text{entropyRelative}(V^{\mathrm{C}} * T, A * T) \\
&\qquad := \sum_{(R,\cdot) \in T^{-1}} (\hat{V}^{\mathrm{C}} * T)_R \times \ln \frac{(\hat{V}^{\mathrm{C}} * T)_R}{(\hat{A} * T)_R} \\
&\qquad = \text{entropyCross}(V^{\mathrm{C}} * T, A * T) \ - \ \text{entropy}(V^{\mathrm{C}} * T)
\end{aligned}$$

The *component cardinality size relative entropy* is positive, $\text{entropyRelative}(V^{\mathrm{C}} * T, A * T) \geq 0$.

The *size-volume* scaled *component size cardinality sum relative entropy* is the *size-volume* scaled *component size cardinality sum cross entropy* minus the *size-volume* scaled *component size cardinality sum entropy*,

$$\begin{aligned}
&(z + v) \times \text{entropy}(A * T + V^{\mathrm{C}} * T) \\
&\qquad\qquad - z \times \text{entropy}(A * T) \ - v \times \text{entropy}(V^{\mathrm{C}} * T)
\end{aligned}$$

The *size-volume* scaled *component size cardinality sum relative entropy* is positive, $(z+v) \times \text{entropy}(A * T + V^{\mathrm{C}} * T) - z \times \text{entropy}(A * T) - v \times \text{entropy}(V^{\mathrm{C}} * T) \geq 0$. The *size-volume* scaled *component size cardinality sum relative entropy* is less than the *size-volume* scaled logarithm of the *derived volume*, $(z+v) \times \text{entropy}(A * T + V^{\mathrm{C}} * T) - z \times \text{entropy}(A * T) - v \times \text{entropy}(V^{\mathrm{C}} * T) < (z + v) \ln w$.

In all cases the *relative entropy* is maximised when (a) the *cross entropy* is maximised and (b) the *component entropy* is minimised. That is, the *relative entropy* is maximised when both (i) the *component size entropy*,

entropy$(A * T)$, and (ii) the *component cardinality entropy*, entropy$(V^{\mathrm{C}} * T)$, are low, but low in different ways so that the *component size cardinality sum cross entropy*, entropy$(A * T + V^{\mathrm{C}} * T)$, is high.

Let *histogram* A have a set of *variables* $V = \mathrm{vars}(A)$ which is partitioned into query *variables* $K \subset V$ and label *variables* $V \setminus K$. Let $T \in \mathcal{T}_{U,\mathrm{f},1}$ be a *one functional transform* having *underlying variables* equal to the query *variables*, $\mathrm{und}(T) = K$. As shown above, given a query *state* $Q \in K^{\mathrm{CS}}$ that is *effective* in the *sample derived*, $R \in (A * T)^{\mathrm{FS}}$ where $\{R\} = (\{Q\}^{\mathrm{U}} * T)^{\mathrm{FS}}$, the *probability histogram* for the label is

$$\{Q\}^{\mathrm{U}} * T * T^{\odot A} \ \% \ (V \setminus K) \quad \in \quad \mathcal{A} \cap \mathcal{P}$$

If the normalised *histogram*, $\hat{A} \in \mathcal{A} \cap \mathcal{P}$, is treated as a *probability function* of a single-*state* query, the *scaled expected entropy* of the *modelled transformed conditional product*, or *scaled* label *entropy*, is

$$\begin{aligned}
&\sum_{(R,C) \in T^{-1}} (A * T)_R \times \mathrm{entropy}(A * C \ \% \ (V \setminus K)) \\
= &\sum_{(R,\cdot) \in T^{-1}} (A * T)_R \times \mathrm{entropy}(\{R\}^{\mathrm{U}} * T^{\odot A} \ \% \ (V \setminus K))
\end{aligned}$$

This is similar to the definition of the *scaled expected component entropy*, above,

$$\begin{aligned}
z \times \mathrm{entropyComponent}(A, T) \quad &:= \quad \sum_{(R,C) \in T^{-1}} (A * T)_R \times \mathrm{entropy}(A * C) \\
&= \quad \sum_{(R,\cdot) \in T^{-1}} (A * T)_R \times \mathrm{entropy}(\{R\}^{\mathrm{U}} * T^{\odot A})
\end{aligned}$$

but now the *component* is *reduced* to the label *variables*, $V \setminus K$.

The label *entropy*, may be contrasted with the *alignment* between the *derived variables*, W, and the label *variables*, $V \setminus K$,

$$\mathrm{algn}(A * \mathrm{his}(T) \ \% \ (W \cup V \setminus K))$$

The *alignment* varies against the *scaled* label *entropy* or *scaled* query *conditional entropy*. Let $B = A * \text{his}(T)\ \%\ (W \cup V \setminus K)$,

$$\begin{aligned}
&\text{algn}(A * \text{his}(T)\ \%\ (W \cup V \setminus K)) \\
&\quad = \text{algn}(B) \\
&\quad \approx z \times \text{entropy}(B^{\text{X}}) - z \times \text{entropy}(B) \\
&\quad \sim z \times \text{entropy}(B\%W) + z \times \text{entropy}(B\%(V \setminus K)) - z \times \text{entropy}(B) \\
&\quad \sim -(z \times \text{entropy}(B) - z \times \text{entropy}(B\%W)) \\
&\quad = -\sum_{R \in (B\%W)^{\text{FS}}} (B\%W)_R \times \text{entropy}(B * \{R\}^{\text{U}}\ \%\ (V \setminus K)) \\
&\quad = -\sum_{(R,C) \in T^{-1}} (A * T)_R \times \text{entropy}(A * C\ \%\ (V \setminus K))
\end{aligned}$$

The label *entropy*, may also be compared to the *slice entropy*, which is the sum of the *sized entropies* of the *contingent slices reduced* to the label *variables*, $V \setminus K$,

$$\sum_{R \in (A\%K)^{\text{FS}}} (A\%K)_R \times \text{entropy}(A * \{R\}^{\text{U}}\ \%\ (V \setminus K))$$

In the case where the relation between the *derived variables* and the label *variables* is functional or *causal*,

$$\text{split}(W, (A * \text{his}(T)\ \%\ (W \cup V \setminus K))^{\text{FS}}) \ \in\ W^{\text{CS}} \to (V \setminus K)^{\text{CS}}$$

the label *entropy* is zero,

$$\sum_{(R,C) \in T^{-1}} (A * T)_R \times \text{entropy}(A * C\ \%\ (V \setminus K)) \ =\ 0$$

So label *entropy* is a measure of the ambiguity in the relation between the *derived variables* and the label *variables*. Negative label *entropy* may be viewed as the degree to which the *derived variables* of the *model* predict the label *variables*. In the cases of low label *entropy*, or high *causality*, the *derived variables* and the label *variables* are correlated and therefore *aligned*, $\text{algn}(A * \text{his}(T)\ \%\ (W \cup V \setminus K)) > 0$. In these cases the *derived histogram* tends to the *diagonal*, $\text{algn}(A * T) > 0$.

2.5.3 Functional definition sets

This section may be skipped until section 'Artificial neural networks'.

A *functional definition set* $F \in \mathcal{F}$ is a set of *unit functional transforms*, $\forall T \in F~(T \in \mathcal{T}_{\mathrm{f}})$. *Functional definition sets* are also called *fuds*. *Fuds* are constrained such that *derived variables* can appear in only one *transform*. That is, the sets of *derived variables* are disjoint,

$$\forall F \in \mathcal{F}~\forall T_1, T_2 \in F~(T_1 \neq T_2 \implies \mathrm{der}(T_1) \cap \mathrm{der}(T_2) = \emptyset)$$

The set of *fud histograms* is $\mathrm{his}(F) := \{\mathrm{his}(T) : T \in F\}$. The set of *fud variables* is $\mathrm{vars}(F) := \bigcup\{\mathrm{vars}(X) : X \in \mathrm{his}(F)\}$. The *fud derived* is $\mathrm{der}(F) := \bigcup_{T \in F} \mathrm{der}(T) \setminus \bigcup_{T \in F} \mathrm{und}(T)$. The *fud underlying* is $\mathrm{und}(F) := \bigcup_{T \in F} \mathrm{und}(T) \setminus \bigcup_{T \in F} \mathrm{der}(T)$. The set of *underlying variables* of a *fud* is also called the *substrate*.

A *functional definition set* is a *model*, so it can be converted to a *functional transform*,

$$F^{\mathrm{T}} \quad := \quad (\prod \mathrm{his}(F) ~\%~ (\mathrm{der}(F) \cup \mathrm{und}(F)),~ \mathrm{der}(F))$$

The resultant *transform* has the same *derived* and *underlying variables* as the *fud*, $\mathrm{der}(F^{\mathrm{T}}) = \mathrm{der}(F)$ and $\mathrm{und}(F^{\mathrm{T}}) = \mathrm{und}(F)$.

The set of *one functional definition sets* $\mathcal{F}_{U,1}$ in *system* U is the subset of the *functional definition sets*, $\mathcal{F}_{U,1} \subset \mathcal{F}$, such that all *transforms* are *one functional* and the *fuds* are not *circular*. The *transform* of a *one functional definition set* is a *one functional transform*, $\forall F \in \mathcal{F}_{U,1}~(F^{\mathrm{T}} \in \mathcal{T}_{U,\mathrm{f},1})$.

A *dependent variable* of a *one functional definition set* $F \in \mathcal{F}_{U,1}$ is any *variable* that is not a *fud underlying variable*, $\mathrm{vars}(F) \setminus \mathrm{und}(F)$. Each *dependent variable depends* on an *underlying* subset of the *fud*, $\mathrm{depends} \in \mathcal{F} \times \mathrm{P}(\mathcal{V}) \to \mathcal{F}$ where $\forall w \in \mathrm{vars}(F) \setminus \mathrm{und}(F)~(\mathrm{depends}(F, \{w\}) \subseteq F)$.

Each *dependent variable* is in a *layer*. The *layer* is the length of the longest path of *underlying transforms* to the *dependent variable*. Given *fud* $F \in \mathcal{F}_{U,1}$, let l be the highest *layer*, $l = \mathrm{layer}(F, \mathrm{der}(F))$, where $\mathrm{layer} \in \mathcal{F} \times \mathrm{P}(\mathcal{V}) \to \mathbf{N}$ is defined in terms of $\mathrm{depends} \in \mathcal{F} \times \mathrm{P}(\mathcal{V}) \to \mathcal{F}$. Let F_i be the subset of the *fud* in a particular *layer*, $F_i = \{T : T \in F,~\mathrm{layer}(F, \mathrm{der}(T)) = i\}$. Then $F = \bigcup_{i \in \{1 \ldots l\}} F_i$.

A *one functional definition set* $F \in \mathcal{F}_{U,1}$ is *non-overlapping* if the sets of *variables* of the *underlying transforms* of each of the *fud derived variables* are disjoint, $\forall v, w \in \mathrm{der}(F)~(v \neq w~\wedge~\mathrm{vars}(\mathrm{depends}(F, \{v\})) \cap$

vars(depends($F, \{w\}$)) $= \emptyset$). A *one functional transform* $T \in \mathcal{T}_{U,\mathrm{f},1}$ is *non-overlapping* if it is equal to the *transform* of a *non-overlapping fud*, $T = F^{\mathrm{T}}$. If the *transform*, T, is *non-overlapping*, then its *formal* is always *independent*, $A^{\mathrm{X}} * T = (A^{\mathrm{X}} * T)^{\mathrm{X}}$, where A is any *underlying histogram*, $\mathrm{vars}(A) \supseteq \mathrm{und}(T)$.

Given a set of *substrate variables* V, the set of *substrate functional definition sets* $\mathcal{F}_{U,V}$ is the subset of *one functional definition sets*, $\mathcal{F}_{U,V} \subset \mathcal{F}_{U,1}$, that (i) have *underlying variables* which are subsets of the *substrate*, $\forall F \in \mathcal{F}_{U,V}\ (\mathrm{und}(F) \subseteq V)$, and (ii) consist of *partition transforms*, $\forall F \in \mathcal{F}_{U,V}\ \forall T \in F\ \exists P \in \mathrm{B}(\mathrm{und}(T)^{\mathrm{CS}})\ (T = P^{\mathrm{T}})$. In addition, *partition circularities* are excluded by ensuring that the *partitions* are unique in the *fud* when *flattened* to *substrate*.

Let v be the *volume* of the *substrate*, $v = |V^{\mathrm{C}}|$. If the *volume*, v, is finite, the set of *substrate fuds* is finite, $|\mathcal{F}_{U,V}| < \infty$.

Avoiding *partition circularities* is computationally expensive. The *infinite-layer substrate functional definition sets* $\mathcal{F}_{\infty,U,V}$ is the superset of the *substrate functional definition sets*, $\mathcal{F}_{\infty,U,V} \supset \mathcal{F}_{U,V}$, that drop the exclusion of *partition circularities*. The *infinite-layer substrate fud set* is defined recursively,

$$\mathcal{F}_{\infty,U,V} \quad = \quad \{F : F \subseteq \mathrm{powinf}(U,V)(\emptyset),\ \mathrm{und}(F) \subseteq V\}$$

where

$$\begin{aligned}\mathrm{powinf}(U,V)(F) := F \cup G \cup \mathrm{powinf}(U,V)(F \cup G) :\\ G = \{P^{\mathrm{T}} : K \subseteq \mathrm{vars}(F) \cup V,\ P \in \mathrm{B}(K^{\mathrm{CS}})\}\end{aligned}$$

The cardinality of the *infinite-layer substrate fud set* is infinite, $|\mathcal{F}_{\infty,U,V}| = \infty$.

2.5.4 Decompositions

This section may be skipped until section 'Tractable and practicable aligned induction'.

A *functional definition set decomposition* is a *model* that consists of a tree of *fuds* that are *contingent* on *components*.

The set of *functional definition set decompositions* $\mathcal{D}_{\mathrm{F}}$ is a subset of the trees of pairs of (i) *states*, $\mathcal{S}$, and (ii) *functional definition sets*, $\mathcal{F}$

$$\mathcal{D}_{\mathrm{F}} \quad \subset \quad \mathrm{trees}(\mathcal{S} \times \mathcal{F})$$

Let D be a *fud decomposition*, $D \in \mathcal{D}_{\mathrm{F}}$. The set of *fuds* is $\mathrm{fuds}(D) := \{F : ((\cdot, F), \cdot) \in \mathrm{nodes}(D)\}$. The *underlying* is $\mathrm{und}(D) := \bigcup\{\mathrm{und}(F) : F \in \mathrm{fuds}(D)\}$. The set of *underlying variables* of a *decomposition* is also called the *substrate*.

Fud decompositions are constrained such that each of the *states* in child pairs are *states* in the *derived variables* of the parent *fud*,

$$\forall D \in \mathcal{D}_{\mathrm{F}}\ \forall((\cdot, F), E) \in \mathrm{nodes}(D)\ \forall((S, \cdot), \cdot) \in E\ (S \in \mathrm{dom}((F^{\mathrm{T}})^{-1}))$$

The root nodes have no parent and so their *states* are constrained to be null, $\forall D \in \mathcal{D}_{\mathrm{F}}\ \ \forall((S, \cdot), \cdot) \in D\ (S = \emptyset)$. Given a *fud decomposition* $D \in \mathcal{D}_{\mathrm{F}}$ having *underlying variables* $V = \mathrm{und}(D)$, each *fud* $F \in \mathrm{fuds}(D)$ is *contingent* on the *component* $C \in \mathrm{B}(V^{\mathrm{C}})$ implied by the union of the ancestor *derived states* in the *derived variables* of the union of the ancestor *fuds*. Let L be a path in the *fud decomposition*, $L \in \mathrm{paths}(D)$. Then for each child *fud* $(\cdot, F) = L_i$, where $i \in \{2 \ldots |L|\}$, the union of the ancestor *derived states* is $R = \bigcup\{S : j \in \{2 \ldots i\},\ (S, \cdot) = L_j\}$, the union of the ancestor *fuds* is $G = \bigcup\{H : j \in \{1 \ldots i-1\},\ (\cdot, H) = L_j\}$, and so the *contingent component* is $(G^{\mathrm{T}})^{-1}(R)$. In the case where the *underlying* of the ancestor *fud*, G, is the whole *substrate*, $\mathrm{und}(G) = V$, then the *component* is $C = (G^{\mathrm{T}})^{-1}(R) \subseteq V^{\mathrm{C}}$.

The function $\mathrm{cont} \in \mathcal{D}_{\mathrm{F}} \to \mathrm{P}(\mathcal{A} \times \mathcal{F})$ returns the set of *component-fud* pairs of the *fud decomposition*. When the *fud decomposition*, D, is applied to a *histogram* $A \in \mathcal{A}$ in *variables* $\mathrm{vars}(A) = V$, each *fud transform* is applied to the *contingent slice*, $A * C * F^{\mathrm{T}}$ where $(C, F) \in \mathrm{cont}(D)$. Two *fuds* on the same path $(\cdot, F_1) \in L_j$ and $(\cdot, F_2) \in L_i$ where $L \in \mathrm{paths}(D)$ and $j < i$, are such that the *fud* $(C_1, F_1) \in \mathrm{cont}(D)$ nearer the root has a *component* which is a superset of the *component* of the *fud* $(C_2, F_2) \in \mathrm{cont}(D)$ nearer the leaves, $C_1 \supset C_2$. So the *slice* nearer the root is greater than or equal to the *slice* nearer the leaves, $A * C_1 \geq A * C_2$. That is, the *fuds* are more and more selectively *contingent* along the *fud decomposition's* paths, and so are applied to smaller and smaller *slices*.

In the case where each of the *slice derived* are *diagonalised*, $\forall (C, F) \in \mathrm{cont}(D)\ (\mathrm{diagonal}(A * C * F^{\mathrm{T}}))$, the *fud decomposition*, D, is a *contingent, layered, redundant model* of the *sample histogram*, A.

A *fud decomposition* is a *model*, so it can be converted to a *functional transform*, $D^{\mathrm{T}} \in \mathcal{T}_{\mathrm{f}}$. The *partition* of the *fud decomposition transform* is equal to the set of *components* corresponding to those *fud derived states* that are not parent *derived states* in the *decomposition* tree, $\bigcup\{\mathrm{dom}((F^{\mathrm{T}})^{-1})\setminus\{S : ((S,\cdot),\cdot) \in E\} : ((\cdot,F),E) \in \mathrm{nodes}(D)\}$. The resultant *transform* has the same *underlying variables* as the *fud decomposition*, $\mathrm{und}(D^{\mathrm{T}}) = \mathrm{und}(D)$.

The tree of a *fud decomposition* is sometimes unwieldy, so consider the *fud decomposition fud*, $D^{\mathrm{F}} \in \mathcal{F}$, which is the intermediate *fud* used in the construction of the *fud decomposition transform*, D^{T}. The *decomposition fud* is defined as the union of the *decomposition fuds* and the *nullable fud*, $D^{\mathrm{F}} := \bigcup \mathrm{fuds}(D) \cup \mathrm{nullable}(U)(D^{\mathrm{D}})$. The *nullable fud*, $\mathrm{nullable}(U)(D^{\mathrm{D}})$, is defined in section 'Decompositions', below. It consists of a *layer* of *transforms* which is added on top of the union of the *decomposition fuds*, $\bigcup \mathrm{fuds}(D)$. Each *derived variable* in the *fud* union, $w \in \mathrm{der}(F)$ where $F \in \mathrm{fuds}(D)$, is in the *underlying* of a corresponding *transform*, $w \in \mathrm{und}(T_w)$, in the *nullable layer*. The *transform derived* consists of a *nullable variable* $\{w'\} = \mathrm{der}(T_w)$. This *nullable variable*, w', has the same *values* as its *underlying variable*, w, but with an additional null *value*, $U_{w'} = U_w \cup \{\mathrm{null}\}$. If the *fud*, F, is not the *root fud*, there is also a *contingent variable* c with *values* corresponding to the *fud's in-slice* and *out-slice states*, $U_c = \{\mathrm{in}, \mathrm{out}\}$. That is, given *contingent state* $S \in C^{\mathrm{S}}$, where $(C,F) \in \mathrm{cont}(D)$, the *derived state*, R, is such that $(c,\mathrm{in}) \in R$. Similarly, if $S \in V^{\mathrm{CS}} \setminus C^{\mathrm{S}}$, then $(c,\mathrm{out}) \in R$. The *underlying* of *nullable variable's transform* will also contain the *contingent variable*, $\{c, w\} = \mathrm{und}(T_w)$. The *nullable variable*, w', is constrained by the *transform*, T_w, to be in the null *value* whenever the *contingent variable*, c, is in the out *value*, and to be in the *value* of the *underlying variable*, w, otherwise. That is, $(c,\mathrm{out}) \in R \implies (w',\mathrm{null}) \in R$, and $(c,\mathrm{in}) \in R \implies (w', R_w) \in R$. In this way, there is no need to navigate the *slices* of the *decomposition*. The *fud decomposition fud*, D^{F}, can be analysed by examining the *effective states* of *reductions* to its *nullable derived variables*, $\mathrm{der}(D^{\mathrm{F}})$.

Given a set of *substrate variables* V, the set of *substrate fud decompositions* $\mathcal{D}_{\mathrm{F},U,V}$ is a subset of *fud decompositions*, $\mathcal{D}_{\mathrm{F},U,V} \subset \mathcal{D}_{\mathrm{F}}$, that contain only *substrate fuds*, $\forall D \in \mathcal{D}_{\mathrm{F},U,V}\ \forall F \in \mathrm{fuds}(D)\ (F \in \mathcal{F}_{U,V})$. In addition, each *fud* is unique in a path, $\forall D \in \mathcal{D}_{\mathrm{F},U,V}\ \forall L \in \mathrm{paths}(D)\ (|\{F : (\cdot,(\cdot,F)) \in L\}| = |L|)$.

Let v be the *volume* of the *substrate*, $v = |V^{\mathrm{C}}|$. If the *volume*, v, is finite, the set of *substrate fud decompositions* is finite, $|\mathcal{D}_{\mathrm{F},U,V}| < \infty$.

Similarly, the *infinite-layer substrate fud decompositions* $\mathcal{D}_{\mathrm{F},\infty,U,V}$ is the superset of the *substrate fud decompositions*, $\mathcal{D}_{\mathrm{F},\infty,U,V} \supset \mathcal{D}_{\mathrm{F},U,V}$, that contain only *infinite-layer substrate fuds*, $\forall D \in \mathcal{D}_{\mathrm{F},\infty,U,V}\ \forall F \in \mathrm{fuds}(D)\ (F \in \mathcal{F}_{\infty,U,V})$.

The cardinality of the *infinite-layer substrate fud decomposition set* is infinite, $|\mathcal{D}_{\mathrm{F},\infty,U,V}| = \infty$.

2.6 Induction with model

2.6.1 Classical induction

Given *substrate transform* $T \in \mathcal{T}_{U,V}$, the *derived histogram* valued *integral substrate histograms* function $D_{U,\mathrm{i},T,z}$ is defined

$$D_{U,\mathrm{i},T,z} \ := \ \{(A, A * T) : A \in \mathcal{A}_{U,\mathrm{i},V,z}\}$$

The finite set of *iso-deriveds* of *derived histogram* $A * T$ is

$$D_{U,\mathrm{i},T,z}^{-1}(A * T) \ = \ \{B : B \in \mathcal{A}_{U,\mathrm{i},V,z},\ B * T = A * T\}$$

The degree to which an *integral iso-set* $I \subseteq \mathcal{A}_{U,\mathrm{i},V,z}$ that contains the *histogram*, $A \in I$, is said to be *law-like* is called the *iso-derivedence*. The *iso-derivedence* is defined as the ratio of (i) the cardinality of the intersection between the *integral iso-set* and the set of *integral iso-deriveds*, and (ii) the cardinality of the union,

$$\frac{1}{|\mathcal{A}_{U,\mathrm{i},V,z}|} \ \leq \ \frac{|I \ \cap \ D_{U,\mathrm{i},T,z}^{-1}(A * T)|}{|I \ \cup \ D_{U,\mathrm{i},T,z}^{-1}(A * T)|} \ \leq \ 1$$

In *classical modelled induction* the *history probability functions* are constrained by *derived histogram*.

Let P be a *substrate history probability function*, $P \in (\mathcal{H}_{U,V,z} :\to \mathbf{Q}_{\geq 0}) \cap \mathcal{P}$. Given a *history* $H_E \in \mathcal{H}_{U,V,z_E}$, of *size* $z_E = |H_E|$, consider the case where the *derived histogram* $A * T$ of *drawn histories* is *known* to be *necessary*, $\sum(P(H) : H \subseteq H_E,\ \mathrm{his}(H) * T = A * T) = 1$. The *maximum likelihood estimate* which maximises the entropy, $\mathrm{entropy}(\tilde{P})$, is

$$\begin{aligned}
\tilde{P} \ &= \ \{(H, 1) : H \subseteq H_E,\ \mathrm{his}(H) * T = A * T\}^{\wedge} \ \cup \\
&\quad\ \{(G, 0) : G \in \mathcal{H}_{U,V,z},\ G \nsubseteq H_E\} \ \cup \\
&\quad\ \{(G, 0) : G \in \mathcal{H}_{U,V,z},\ \mathrm{his}(G) * T \neq A * T\} \\
&= \ \{(H, 1/\textstyle\sum(Q_{\mathrm{h},U}(E, z)(B) : B \in D_{U,\mathrm{i},T,z}^{-1}(A * T))) : \\
&\qquad\qquad\qquad H \subseteq H_E,\ \mathrm{his}(H) * T = A * T\} \ \cup \\
&\quad\ \{(G, 0) : G \in \mathcal{H}_{U,V,z},\ G \nsubseteq H_E\} \ \cup \\
&\quad\ \{(G, 0) : G \in \mathcal{H}_{U,V,z},\ \mathrm{his}(G) * T \neq A * T\}
\end{aligned}$$

That is, the *maximum likelihood estimate*, $\tilde{P}$, is such that all *drawn histories* $H \subseteq H_E$ with the *derived*, $\mathrm{his}(H) * T = A * T$, are uniformly probable and all other *histories*, $G \not\subseteq H_E$ or $\mathrm{his}(G) * T \neq A * T$, are impossible, $\tilde{P}(G) = 0$. If (i) the *transform*, T, is *known*, (ii) the *derived*, $A * T$, is *known* and (iii) the *distribution histogram*, H_E, is *known*, then the *likely probability function*, $\tilde{P}$, is *known*.

The *likely probability* of *drawing histogram* A from *necessary drawn derived* $A * T$ is

$$\sum(\tilde{P}(H) : H \in \mathcal{H}_{U,V,z},\ \mathrm{his}(H) = A) \ = \ \frac{Q_{\mathrm{h},U}(E, z)(A)}{\sum Q_{\mathrm{h},U}(E, z)(B) : B \in D^{-1}_{U,\mathrm{i},T,z}(A * T)}$$

The *likely history probability function* entropy, $\mathrm{entropy}(\tilde{P})$, is maximised when the sum of the *iso-derived historical frequencies*, $\sum Q_{\mathrm{h},U}(E, z)(B) : B \in D^{-1}_{U,\mathrm{i},T,z}(A * T)$, is maximised.

Consider the case where the *transform*, T, is *known* and the *derived*, $A * T$, is *known*, but the *distribution histogram*, E, is *unknown* and hence the *likely history probability function*, $\tilde{P}$, is *unknown*. The *maximum likelihood estimate* $\tilde{E}$ for the *distribution histogram*, E, is a modal value of the *likelihood function*,

$$\tilde{E} \ \in \ \mathrm{maxd}(\{(D, \sum(Q_{\mathrm{h},U}(D, z)(B) : B \in D^{-1}_{U,\mathrm{i},T,z}(A * T))) : D \in \mathcal{A}_{U,\mathrm{i},V,z_E}\})$$

The *likely distribution histogram*, $\tilde{E}$, is *known* if the *distribution histogram size*, z_E, is *known*, the *transform*, T, is *known* and the *derived*, $A * T$, is *known*. If it is assumed that the *distribution histogram* equals the *likely distribution histogram*, $E = \tilde{E}$, then the *likely history probability* is *known*, $\tilde{P}(H) = 1/\sum(Q_{\mathrm{h},U}(\tilde{E}, z)(B) : B \in D^{-1}_{U,\mathrm{i},T,z}(A*T))$ where $\mathrm{his}(H)*T = A*T$.

In the case where the *distribution histogram*, E, is *unknown*, and the *distribution histogram size*, z_E, is also *unknown*, except that it is *known* to be large, $z_E \gg z$, then the *maximum likelihood estimate* $\tilde{E}$ for the *distribution probability histogram*, $\hat{E}$, may be approximated by a modal value of a *likelihood function* which depends on the *multinomial distribution* instead,

$$\tilde{E} \ \in \ \mathrm{maxd}(\{(D, \sum(Q_{\mathrm{m},U}(D, z)(B) : B \in D^{-1}_{U,\mathrm{i},T,z}(A * T))) : D \in \mathcal{A}_{U,V,1}\})$$

The *normalised naturalisation*, $\hat{A} * T * T^{\dagger}$, is a solution. The *naturalisation*, $A * T * T^{\dagger}$, is the *independent analogue* of the *iso-deriveds*. So the *maximum* *likelihood estimate*, $\tilde{E}$, for the *distribution probability histogram*, $\hat{E}$, is the *naturalisation probability histogram*, $\hat{A} * T * T^{\dagger}$,

$$\tilde{E} \;=\; \hat{A} * T * T^{\dagger}$$

In the case where the *naturalisation* is *integral*, $A * T * T^{\dagger} \in \mathcal{A}_{\mathrm{i}}$, the sum of the *iso-derived naturalisation-distributed multinomial probabilities* varies with the *naturalisation naturalisation-distributed multinomial probability*,

$$\begin{aligned}\sum Q_{\mathrm{m},U}(A * T * T^{\dagger}, z)(B) : B \in D^{-1}_{U,\mathrm{i},T,z}(A * T) \;\sim& \\ & Q_{\mathrm{m},U}(A * T * T^{\dagger}, z)(A * T * T^{\dagger})\end{aligned}$$

So, if it is assumed that the *distribution probability histogram* equals the *likely distribution probability histogram*, $\hat{E} = \tilde{E} = \hat{A} * T * T^{\dagger}$, then the *likely history probability* varies against the *naturalisation-distributed multinomial probability* of the *naturalisation*, $\tilde{P}(H) \sim 1/\hat{Q}_{\mathrm{m},U}(A * T * T^{\dagger}, z)(A * T * T^{\dagger})$.

The cardinality of the set of *integral iso-deriveds* may be stated explicitly as the product of the weak compositions of the *components*,

$$|D^{-1}_{U,\mathrm{i},T,z}(A * T)| \;=\; \prod_{(R,C) \in T^{-1}} \frac{((A * T)_R + |C| - 1)!}{(A * T)_R! \; (|C| - 1)!}$$

So the *integral iso-deriveds log-cardinality* varies against the *size-volume* scaled *component size cardinality sum relative entropy*,

$$\begin{aligned}\ln |D^{-1}_{U,\mathrm{i},T,z}(A * T)| \;\sim& \\ -\big((z + v) \times \mathrm{entropy}&(A * T + V^{\mathrm{C}} * T) \\ &-z \times \mathrm{entropy}(A * T) \; - v \times \mathrm{entropy}(V^{\mathrm{C}} * T)\big)\end{aligned}$$

where *size* $z = \mathrm{size}(A) = \mathrm{size}(A * T)$ and *volume* $v = |V^{\mathrm{C}}|$. In the domain where the *size* is less than or equal to the *volume*, $z \leq v$, the *integral iso-deriveds log-cardinality* varies against the *size* scaled *component size cardinality relative entropy*,

$$\ln |D^{-1}_{U,\mathrm{i},T,z}(A * T)| \;\sim\; -z \times \mathrm{entropyRelative}(A * T, V^{\mathrm{C}} * T)$$

So the logarithm of the *likely probability* of *drawing histogram* A from *necessary drawn derived* $A * T$ varies with the *relative entropy*,

$$\begin{aligned}\ln \sum(\tilde{P}(H) : H \in \mathcal{H}_{U,V,z},\; \mathrm{his}(H) = A) \;\sim& \\ & z \times \mathrm{entropyRelative}(A * T, V^{\mathrm{C}} * T)\end{aligned}$$

The *naturalisation*, $A*T*T^{\dagger}$, is the most *probable histogram*, $\forall B \in D^{-1}_{U,\mathrm{i},T,z}(A*T)$ $(Q_{\mathrm{m},U}(A*T*T^{\dagger},z)(A*T*T^{\dagger}) \geq Q_{\mathrm{m},U}(A*T*T^{\dagger},z)(B))$. In the case where the *histogram* is *natural*, $A = A*T*T^{\dagger}$, then, as the *relative entropy*, $\mathrm{entropyRelative}(A*T, V^{\mathrm{C}}*T)$, increases, the *likely histogram probability*, $Q_{\mathrm{m},U}(A,z)(A)/\sum(Q_{\mathrm{m},U}(A,z)(B) : B \in D^{-1}_{U,\mathrm{i},T,z}(A*T))$, increases.

The *likely history probability function* entropy varies with the *naturalisation entropy*, $\mathrm{entropy}(\tilde{P}) \sim \mathrm{entropy}(A*T*T^{\dagger})$, and against the *relative entropy*, $\mathrm{entropy}(\tilde{P}) \sim -\ \mathrm{entropyRelative}(A*T, V^{\mathrm{C}}*T)$.

Consider the case where a *drawn histogram* A is *known*, but neither the *distribution histogram*, E, is *known* nor the *transform*, T, is *known*, and hence the *likely history probability function*, $\tilde{P}$, is *unknown*. The *maximum likelihood estimate* $(\tilde{E},\tilde{T})$ for the pair of the *distribution histogram*, E, and the *transform*, T, is a modal value of the *likelihood function*,

$$\begin{aligned}(\tilde{E},\tilde{T}) \in\ & \\ & \mathrm{maxd}(\{((D,M), \sum(Q_{\mathrm{h},U}(D,z)(B) : B \in D^{-1}_{U,\mathrm{i},M,z}(A*M))) : \\ & \qquad D \in \mathcal{A}_{U,\mathrm{i},V,z_E},\ M \in \mathcal{T}_{U,V}\})\end{aligned}$$

All solutions are such that the *transform maximum likelihood estimate* is *unary*, $\tilde{T} = T_{\mathrm{u}}$ where $T_{\mathrm{u}} = \{V^{\mathrm{CS}}\}^{\mathrm{T}}$. This is the trivial case where the set of *iso-derived histograms* is the entire set of *substrate histograms*, $D^{-1}_{U,\mathrm{i},T_{\mathrm{u}},z}(A*T_{\mathrm{u}}) = \mathcal{A}_{U,\mathrm{i},V,z}$. In this case *necessary derived*, $H \subseteq H_E$ and $\mathrm{his}(H)*T_{\mathrm{u}} = A*T_{\mathrm{u}}$, reduces to *drawn history*, $H \subseteq H_E$. If it is assumed that the *transform* equals the *likely transform*, $T = \tilde{T} = T_{\mathrm{u}}$, then the *likely history probability function* which maximises the entropy, $\mathrm{entropy}(\tilde{P})$, is

$$\begin{aligned}\tilde{P} &= \{(H, 1/\binom{z_E}{z}) : H \subseteq H_E,\ |H| = z\} \cup \\ & \quad \{(G,0) : G \in \mathcal{H}_{U,V,z},\ G \nsubseteq H_E\}\end{aligned}$$

That is, in the case of *unknown transform*, the *maximum likelihood estimate*, $\tilde{P}$, is such that all *drawn histories* $H \subseteq H_E$ of *size* $|H| = z$ are uniformly probable and all other *histories*, $G \nsubseteq H_E$, are impossible, $\tilde{P}(G) = 0$.

Define the *derived-dependent* $A^{\mathrm{D}(T)} \in \mathcal{A}_{U,V,z}$ as the *maximum likelihood estimate* of the *distribution histogram* of the *multinomial probability* of the *histogram*, A, conditional that it is an *iso-derived*,

$$\{A^{\mathrm{D}(T)}\} = \mathrm{maxd}(\{(D, \frac{Q_{\mathrm{m},U}(D,z)(A)}{\sum Q_{\mathrm{m},U}(D,z)(B) : B \in D^{-1}_{U,\mathrm{i},T,z}(A * T)}) : D \in \mathcal{A}_{U,V,z}\})$$

The *derived-dependent*, $A^{\mathrm{D}(T)}$, is the *dependent analogue* of the *iso-deriveds*. Note that the *derived-dependent*, $A^{\mathrm{D}(T)}$, is not always computable, but an approximation to any accuracy can be made to it. In the case where the *histogram* is *natural*, the *derived-dependent* equals the *naturalisation*, $A = A * T * T^{\dagger} \implies A^{\mathrm{D}(T)} = A = A * T * T^{\dagger}$.

Now consider the case where, given *necessary drawn derived* $A * T$, it is *known*, in addition, that the *sample histogram* A is the most *probable histogram* of the *iso-derived*. That is, the *likely probability* of *drawing histogram* A from *necessary drawn derived* $A * T$,

$$\sum(\tilde{P}(H) : H \in \mathcal{H}_{U,V,z},\ \mathrm{his}(H) = A) = \frac{Q_{\mathrm{h},U}(E,z)(A)}{\sum Q_{\mathrm{h},U}(E,z)(B) : B \in D^{-1}_{U,\mathrm{i},T,z}(A * T)}$$

is maximised.

In the case where the *transform*, T, is *known* and the *sample*, A, is *known*, but the *distribution histogram*, E, is *unknown*, the *maximum likelihood estimate* $\tilde{E}$ for the *distribution histogram*, E, is a modal value of the *likelihood function*,

$$\tilde{E} \in \mathrm{maxd}(\{(D, \frac{Q_{\mathrm{h},U}(D,z)(A)}{\sum Q_{\mathrm{h},U}(D,z)(B) : B \in D^{-1}_{U,\mathrm{i},T,z}(A * T)}) : D \in \mathcal{A}_{U,\mathrm{i},V,z_E}\})$$

The *likely distribution histogram*, $\tilde{E}$, is *known* if the *distribution histogram size*, z_E, is *known*, the *transform*, T, is *known* and the *sample*, A, is *known*. If it is assumed that the *distribution histogram* equals the *likely distribution histogram*, $E = \tilde{E}$, then the *likely history probability* is *known*, $\tilde{P}(H) = 1/\sum(Q_{\mathrm{h},U}(\tilde{E},z)(B) : B \in D^{-1}_{U,\mathrm{i},T,z}(A * T))$ where $\mathrm{his}(H) * T = A * T$.

If the *histogram* is *natural*, $A = A * T * T^{\dagger}$, then the additional constraint of *probable sample* makes no change to the *maximum likelihood estimate*, $\tilde{E}$,

$$A = A * T * T^{\dagger} \implies$$
$$\begin{aligned} &\text{maxd}(\{(D, \frac{Q_{\text{h},U}(D,z)(A)}{\sum Q_{\text{h},U}(D,z)(B) : B \in D^{-1}_{U,\text{i},T,z}(A*T)}) : D \in \mathcal{A}_{U,\text{i},V,z_E}\}) \\ = \;\; &\text{maxd}(\{(D, \sum(Q_{\text{h},U}(D,z)(B) : B \in D^{-1}_{U,\text{i},T,z}(A*T))) : D \in \mathcal{A}_{U,\text{i},V,z_E}\}) \end{aligned}$$

If the *histogram* is not *natural*, $A \neq A * T * T^{\dagger}$, however, then the *likely history probability function* entropy, entropy($\tilde{P}$), is lower than it is in the case of *necessary derived* unconstrained by *probable sample*.

In the case where the *distribution histogram*, E, is *unknown*, and the *distribution histogram size*, z_E, is also *unknown*, except that it is *known* to be large, $z_E \gg z$, then the *maximum likelihood estimate* $\tilde{E}$ for the *distribution probability histogram*, $\hat{E}$, is now approximated by a modal value of the conditional *likelihood function*,

$$\tilde{E} \;\; \in \;\; \text{maxd}(\{(D, \frac{Q_{\text{m},U}(D,z)(A)}{\sum Q_{\text{m},U}(D,z)(B) : B \in D^{-1}_{U,\text{i},T,z}(A*T)}) : D \in \mathcal{A}_{U,V,1}\})$$

The solution to this is the *normalised derived-dependent*, $\tilde{E} = \hat{A}^{\text{D}(T)}$. The *maximum likelihood estimate* is near the *sample*, $\tilde{E} \sim \hat{A}$, only in as much as it is far from the *naturalisation*, $\tilde{E} \nsim \hat{A} * T * T^{\dagger}$.

The *iso-derived conditional multinomial probability distribution* is defined

$$\hat{Q}_{\text{m,d},T,U}(E,z)(A) \;\; := \;\; \frac{1}{|\text{ran}(D_{U,\text{i},T,z})|} \frac{Q_{\text{m},U}(E,z)(A)}{\sum Q_{\text{m},U}(E,z)(B) : B \in D^{-1}_{U,\text{i},T,z}(A*T)}$$

So the optimisation can be rewritten,

$$\tilde{E} \;\; \in \;\; \text{maxd}(\{(D, \hat{Q}_{\text{m,d},T,U}(D,z)(A)) : D \in \mathcal{A}_{U,V,1}\})$$

In the case where the *histogram* is *natural*, $A = A * T * T^{\dagger}$, the *log likelihood* varies against the *iso-derived log-cardinality*,

$$\begin{aligned} \ln \hat{Q}_{\text{m,d},T,U}(A,z)(A) \;\; &\propto \;\; \ln \frac{Q_{\text{m},U}(A,z)(A)}{\sum Q_{\text{m},U}(A,z)(B) : B \in D^{-1}_{U,\text{i},T,z}(A*T)} \\ &\sim \;\; -\ln |D^{-1}_{U,\text{i},T,z}(A*T)| \end{aligned}$$

So the *log likelihood* varies with the *size-volume* scaled *component size cardinality sum relative entropy*,

$$\begin{aligned} &\ln \hat{Q}_{\text{m,d},T,U}(A,z)(A) \;\; \sim \\ &\quad (z+v) \times \text{entropy}(A * T + V^{\text{C}} * T) \\ &\quad\quad -z \times \text{entropy}(A * T) \;\; - v \times \text{entropy}(V^{\text{C}} * T) \end{aligned}$$

In the domain where the *size* is less than or equal to the *volume*, $z \leq v$, the *log likelihood* varies with the *size* scaled *component size cardinality relative entropy*,

$$\ln \hat{Q}_{\mathrm{m,d},T,U}(A,z)(A) \quad \sim \quad z \times \mathrm{entropyRelative}(A * T, V^{\mathrm{C}} * T)$$

In other words, the *log likelihood* is maximised where (i) the *derived entropy*, $\mathrm{entropy}(A * T)$, is minimised, and (ii) the *cross entropy*, $\mathrm{entropyCross}(A * T, V^{\mathrm{C}} * T)$, is maximised, so that high *counts* are in low cardinality *components* and high cardinality *components* have low *counts*.

If the *histogram* is *natural*, $A = A * T * T^{\dagger}$, and the *component size cardinality relative entropy* is high, $\mathrm{entropyCross}(A * T, V^{\mathrm{C}} * T) > \ln |T^{-1}|$, it can also be shown that the *log likelihood* varies against the *derived multinomial probability*,

$$\ln \hat{Q}_{\mathrm{m,d},T,U}(A,z)(A) \quad \sim \quad - \ \ln \hat{Q}_{\mathrm{m},U}(A * T,z)(A * T)$$

In this case the *sum sensitivity* of the *iso-derived conditional multinomial probability distribution* varies with the *underlying-derived multinomial probability distribution sum sensitivity difference*,

$$\begin{aligned}&\mathrm{sum}(\mathrm{sensitivity}(U)(\hat{Q}_{\mathrm{m,d},T,U}(A,z))) \sim \\ &\quad \mathrm{sum}(\mathrm{sensitivity}(U)(\hat{Q}_{\mathrm{m},U}(A,z))) - \mathrm{sum}(\mathrm{sensitivity}(U)(\hat{Q}_{\mathrm{m},U}(A * T,z)))\end{aligned}$$

and so is less than or equal to the *sum sensitivity* of the *multinomial probability distribution*,

$$\mathrm{sum}(\mathrm{sensitivity}(U)(\hat{Q}_{\mathrm{m,d},T,U}(A,z))) \quad \leq \quad \mathrm{sum}(\mathrm{sensitivity}(U)(\hat{Q}_{\mathrm{m},U}(A,z)))$$

Furthermore, the *sum sensitivity* varies against the *log-likelihood*,

$$\mathrm{sum}(\mathrm{sensitivity}(U)(\hat{Q}_{\mathrm{m,d},T,U}(A,z))) \quad \sim \quad - \ln \hat{Q}_{\mathrm{m,d},T,U}(A,z)(A)$$

That is, in the high *relative entropy natural* case, the maximisation of the *log-likelihood* also tends to minimise the *sum sensitivity* to the *maximum likelihood estimate*. This is opposite to the relationship between the *sum sensitivity* and the *log-likelihood* in *classical non-modelled induction*, which was found to be weakly positively correlated.

As the *relative entropy*, $\mathrm{entropyRelative}(A * T, V^{\mathrm{C}} * T)$, increases, the *log-likelihood*, $\ln \hat{Q}_{\mathrm{m,d},T,U}(A,z)(A)$, increases, but the *sensitivity* to *distribution histogram*, E, decreases. In other words, the higher the *sample relative entropy* the more *likely* the *normalised sample histogram*, $\hat{A}$, equals the *normalised distribution histogram*, $\hat{E}$, and the smaller the *likely* difference between them if they are not equal.

Given *necessary derived* and *probable sample*, consider the case where a *drawn histogram* A is *known*, but neither the *distribution histogram*, E, is *known* nor the *transform*, T, is *known*, and hence the *likely history probability function*, $\tilde{P}$, is *unknown*. The *maximum likelihood estimate* $(\tilde{E},\tilde{T})$ for the pair of the *distribution histogram*, E, and the *transform*, T, is a modal value of the *likelihood function*,

$$(\tilde{E},\tilde{T}) \in \text{maxd}(\{((D,M), \frac{Q_{\text{h},U}(D,z)(A)}{\sum Q_{\text{h},U}(D,z)(B) : B \in D^{-1}_{U,\text{i},M,z}(A * M)}) : D \in \mathcal{A}_{U,\text{i},V,z_E},\ M \in \mathcal{T}_{U,V}\})$$

All solutions are such that the *transform maximum likelihood estimate* is *self*, $\tilde{T} = T_{\text{s}}$ where $T_{\text{s}} = V^{\text{CS}\{\}\text{T}}$. This is the trivial case where the set of *iso-derived histograms* is just the *sample*, $D^{-1}_{U,\text{i},T_{\text{s}},z}(A * T_{\text{s}}) = \{A\}$. In this case *necessary derived*, $\text{his}(H) * T_{\text{s}} = A * T_{\text{s}}$, reduces to *necessary histogram*, $\text{his}(H) = A$. If it is assumed that the *transform* equals the *likely transform*, $T = \tilde{T} = T_{\text{s}}$, then the *likely history probability function* which maximises the entropy, $\text{entropy}(\tilde{P})$, is

$$\begin{aligned}\tilde{P} &= \{(H, 1/Q_{\text{h},U}(E,z)(A)) : H \subseteq H_E,\ \text{his}(H) = A\} \cup \\ &\quad \{(G,0) : G \in \mathcal{H}_{U,V,z},\ G \not\subseteq H_E\} \cup \\ &\quad \{(G,0) : G \in \mathcal{H}_{U,V,z},\ \text{his}(G) \neq A\}\end{aligned}$$

That is, in the case of *unknown transform*, the *maximum likelihood estimate*, $\tilde{P}$, is such that all *drawn histories* $H \subseteq H_E$ with the *histogram*, $\text{his}(H) = A$, are uniformly probable and all other *histories*, $G \not\subseteq H_E$ or $\text{his}(G) \neq A$, are impossible, $\tilde{P}(G) = 0$.

In this case the *maximum likelihood estimate*, $\tilde{E}$, for the *distribution probability histogram*, $\hat{E}$, is the *sample probability histogram*, $\hat{A}$,

$$\tilde{E} = \hat{A} = \hat{A} * T_{\text{s}} * T_{\text{s}}^{\dagger}$$

Consider the case where the *derived* is *uniformly possible*. Given *substrate transform* $T \in \mathcal{T}_{U,V}$, assume that the *substrate history probability function* $P \in (\mathcal{H}_{U,V,z} :\to \mathbf{Q}_{\geq 0}) \cap \mathcal{P}$ is the distribution of an arbitrary *history* valued function, $\mathcal{X} \to \mathcal{H}$, given an arbitrary *derived* valued function, $\mathcal{X} \to \mathcal{A}$. In this case, the *history* valued function is chosen arbitrarily from the constrained subset

$$\{G : F \in \mathcal{X} \to (\mathcal{A} \times (\mathcal{X} \to \mathcal{H})),\ (\cdot,(A',G)) \in F,\ \forall (\cdot,H) \in G\ (\text{his}(H) * T = A')\} \subset \mathcal{X} \to \mathcal{H}$$

Uniformly possible derived is a weaker constraint than *uniformly possible histogram*, so the subset of *history* valued functions is larger.

This subset of the *substrate history probability functions* can be generalised for all *substrate transforms* as the subset derived from

$$\bigcup_{T \in \mathcal{T}_{\mathrm{f}}} (\mathcal{X} \to (\mathcal{A} \times_T (\mathcal{X} \to \mathcal{H})))$$

where $\mathcal{T}_{\mathrm{f}}$ is the set of all *functional transforms*, and the fibre product $\times_T$ is defined

$$\begin{aligned} \mathcal{A} \times_T (\mathcal{X} \to \mathcal{H}) \ &:= \\ &\{(A', G) : (A', G) \in \mathcal{A} \times (\mathcal{X} \to \mathcal{H}),\ \forall (\cdot, H) \in G\ (\mathrm{his}(H) * T = A')\} \end{aligned}$$

In the case where there is a *distribution history* H_E and a *substrate transform* $T \in \mathcal{T}_{U,V}$, the *maximum likelihood estimate* which maximises the entropy, $\mathrm{entropy}(\tilde{P})$, is

$$\begin{aligned} \tilde{P} \ &= \ \Big(\bigcup \{\{(H,1) : H \subseteq H_E,\ \mathrm{his}(H) * T = A'\}^{\wedge} : A' \in \mathrm{ran}(D_{U,\mathrm{i},T,z})\}\Big)^{\wedge} \ \cup \\ &\quad\ \{(G,0) : G \in \mathcal{H}_{U,V,z},\ G \not\subseteq H_E\} \\ &= \ \Big(\bigcup \{\{(H, 1/\textstyle\sum(Q_{\mathrm{h},U}(E,z)(B) : B \in D^{-1}_{U,\mathrm{i},T,z}(A'))) : \\ &\qquad\qquad H \subseteq H_E,\ \mathrm{his}(H) * T = A'\} \ : \ A' \in \mathrm{ran}(D_{U,\mathrm{i},T,z})\}\Big)^{\wedge} \ \cup \\ &\quad\ \{(G,0) : G \in \mathcal{H}_{U,V,z},\ G \not\subseteq H_E\} \end{aligned}$$

That is, the *maximum likelihood estimate*, $\tilde{P}$, is such that all *drawn deriveds* are uniformly probable, and then all *drawn histories* $H \subseteq H_E$ with the same *derived*, $\mathrm{his}(H) * T = A'$, are uniformly probable. If the *distribution histogram*, H_E, is *known* and the *substrate transform*, T, is *known*, then the *likely probability function*, $\tilde{P}$, is *known*.

In the case where the *distribution histogram* is *uniform*, $\hat{E} = \hat{V}^{\mathrm{C}}$, so that all *histories* are *substrate histories*, $\{H : H \in \mathcal{H}_{U,V,z},\ \mathrm{his}(H) * T = A'\}$, the more probable *histograms*, $A \in \mathrm{maxd}(\{(B, \sum(\tilde{P}_H : H \in \mathcal{H}_{U,V,z},\ \mathrm{his}(H) = B)) : B \in \mathcal{A}_{U,\mathrm{i},V,z}\})$, tend to be such that they are *uniform* within the *component*, $\forall C \in T^{\mathrm{P}}\ \forall R, S \in C\ (A_R \approx A_S)$, or *naturalised*, $A \approx A * T * T^{\dagger}$.

The properties of *uniformly possible derived* are the same as for *necessary derived*, except that the probabilities are scaled. So, in the case where the *distribution histogram*, E, is *unknown*, and the *distribution histogram size*, z_E, is also *unknown*, except that it is *known* to be large, $z_E \gg z$, then the *likely history probability* varies against the *naturalisation-distributed multinomial probability* of the *naturalisation*,

$$\tilde{P}(H) \quad \sim \quad 1/|\mathrm{ran}(D_{U,\mathrm{i},T,z})| \times 1/\hat{Q}_{\mathrm{m},U}(A * T * T^{\dagger}, z)(A * T * T^{\dagger})$$

That is, more *histories* are possible but less probable.

Now consider the case where, given *uniform possible derived*, it is *known*, in addition, that the *sample histogram* A is the most *probable histogram* of its *iso-derived*.

The *iso-derived conditional multinomial probability distribution*, is defined above as

$$\hat{Q}_{\mathrm{m,d},T,U}(E, z)(A) \quad := \quad \frac{1}{|\mathrm{ran}(D_{U,\mathrm{i},T,z})|} \frac{Q_{\mathrm{m},U}(E, z)(A)}{\sum Q_{\mathrm{m},U}(E, z)(B) : B \in D_{U,\mathrm{i},T,z}^{-1}(A * T)}$$

The *iso-derived conditional multinomial probability* already includes the *uniform possible* scaling factor of $1/|\mathrm{ran}(D_{U,\mathrm{i},T,z})|$.

The cardinality of the *derived*, $|\mathrm{ran}(D_{U,\mathrm{i},T,z})|$, is equal to the cardinality of the *derived substrate histograms*,

$$|\mathrm{ran}(D_{U,\mathrm{i},T,z})| \quad = \quad \frac{(z + w' - 1)!}{z!\ (w' - 1)!}$$

where $w' = |T^{-1}|$. So the additional term, $-\ln |\mathrm{ran}(D_{U,\mathrm{i},T,z})|$, in the *uniform possible log likelihood*, $\ln \hat{Q}_{\mathrm{m,d},T,U}(E, z)(A)$, varies against the *derived volume*, w', where the *derived volume* is less than the *size*, $w' < z$, otherwise against the *size* scaled *log derived volume*, $z \ln w'$,

$$-\ln |\mathrm{ran}(D_{U,\mathrm{i},T,z})| \quad \sim \quad -\ ((w'\ :\ w' < z)\ +\ (z \ln w'\ :\ w' \geq z))$$

In the case where the *sample* is *natural*, $A = A * T * T^{\dagger}$, the *uniform possible log likelihood* varies (i) against the *derived volume*, w', where the *derived volume* is less than the *size*, $w' < z$, otherwise against the *size* scaled *log derived volume*, $z \ln w'$, and (ii) with the *size-volume* scaled *component size cardinality sum relative entropy*,

$$\begin{aligned}
&\ln \hat{Q}_{\mathrm{m,d},T,U}(A, z)(A) \ \sim \\
&\quad -\ ((w'\ :\ w' < z)\ +\ (z \ln w'\ :\ w' \geq z)) \\
&\quad +\ (z + v) \times \mathrm{entropy}(A * T + V^{\mathrm{C}} * T) \\
&\qquad\qquad -z \times \mathrm{entropy}(A * T)\ - v \times \mathrm{entropy}(V^{\mathrm{C}} * T)
\end{aligned}$$

In other words, the *log likelihood* is maximised where (i) the *derived volume*, w', is minimised, (ii) the *derived entropy*, $\mathrm{entropy}(A * T)$, is minimised, and (iii) the *cross entropy*, $\mathrm{entropyCross}(A * T, V^{\mathrm{C}} * T)$, is maximised, so that high *counts* are in low cardinality *components* and high cardinality *components* have low *counts*.

As in the case of *necesary derived* and *probable sample*, above, if the *histogram* is *natural*, $A = A * T * T^{\dagger}$, and the *component size cardinality relative entropy* is high, $\mathrm{entropyCross}(A * T, V^{\mathrm{C}} * T) > \ln w'$, the *sum sensitivity* of the *iso-derived conditional multinomial probability distribution* is less than or equal to the *sum sensitivity* of the *multinomial probability distribution*,

$$\mathrm{sum}(\mathrm{sensitivity}(U)(\hat{Q}_{\mathrm{m,d},T,U}(A, z))) \quad \leq \quad \mathrm{sum}(\mathrm{sensitivity}(U)(\hat{Q}_{\mathrm{m},U}(A, z)))$$

and varies against the *log-likelihood*,

$$\mathrm{sum}(\mathrm{sensitivity}(U)(\hat{Q}_{\mathrm{m,d},T,U}(A, z))) \quad \sim \quad -\ln \hat{Q}_{\mathrm{m,d},T,U}(A, z)(A)$$

Given *uniform possible derived* and *probable sample*, consider the case where a *drawn histogram* A is *known*, but neither the *distribution histogram*, E, is *known* nor the *transform*, T, is *known*, and hence the *likely history probability function*, $\tilde{P}$, is *unknown*. In the case where the *distribution histogram size*, z_E, is also *unknown*, except that it is *known* to be large, $z_E \gg z$, then the *maximum likelihood estimate* $(\tilde{E}, \tilde{T})$ for the pair of the *distribution histogram*, E, and the *transform*, T, is approximated by a modal value of the conditional *likelihood function*,

$$(\tilde{E}, \tilde{T}) \quad \in \quad \mathrm{maxd}(\{((D, M), \hat{Q}_{\mathrm{m,d},M,U}(D, z)(A)) : D \in \mathcal{A}_{U,V,1},\ M \in \mathcal{T}_{U,V}\})$$

If there is a unique maximum for the *distribution probability histogram*, $\tilde{E}$, this can be rewritten in terms of the *derived-dependent*,

$$\tilde{T} \quad \in \quad \mathrm{maxd}(\{(M, \hat{Q}_{\mathrm{m,d},M,U}(A^{\mathrm{D}(M)}, z)(A)) : M \in \mathcal{T}_{U,V}\})$$

The *derived-dependent*, $A^{\mathrm{D}(T)}$, is not always computable, but an approximation to any accuracy can be made to it, so a computable approximation of the *maximum likelihood estimate*, $\tilde{T}$, can be made for the *unknown transform*, T. In some cases the *likely transform*, $\tilde{T}$, is not trivial, $\tilde{T} \neq T_{\mathrm{u}}$ and $\tilde{T} \neq T_{\mathrm{s}}$.

If it is also *known* that the *sample* is *natural*, the optimisation can be restricted to *natural transforms*, $A = A * T * T^{\dagger} \implies A^{\mathrm{D}(T)} = A$. In this case the optimisation is

$$\tilde{T} \quad \in \quad \mathrm{maxd}(\{(M, \hat{Q}_{\mathrm{m,d},M,U}(A, z)(A)) : M \in \mathcal{T}_{U,V},\ A = A * M * M^{\dagger}\})$$

or

$$\begin{aligned}\tilde{T} &\in \text{maxd}(\{(M, \frac{1}{|\text{ran}(D_{U,\text{i},M,z})|} \frac{Q_{\text{m},U}(A,z)(A)}{\sum Q_{\text{m},U}(A,z)(B) : B \in D_{U,\text{i},M,z}^{-1}(A * M)}) : \\ &\qquad M \in \mathcal{T}_{U,V},\ A = A * M * M^{\dagger}\})\end{aligned}$$

The numerator is constant, so the optimisation can be simplified,

$$\begin{aligned}\tilde{T} &\in \text{mind}(\{(M, |\text{ran}(D_{U,\text{i},M,z})| \sum Q_{\text{m},U}(A,z)(B) : B \in D_{U,\text{i},M,z}^{-1}(A * M)) : \\ &\qquad M \in \mathcal{T}_{U,V},\ A = A * M * M^{\dagger}\})\end{aligned}$$

In this case the *maximum likelihood estimate*, $\tilde{E}$, for the *distribution probability histogram*, $\hat{E}$, is the *sample probability histogram*, $\hat{A}$,

$$\tilde{E} = \hat{A} = \hat{A} * \tilde{T} * \tilde{T}^{\dagger}$$

Note that, although computable, this optimisation is intractable because the cardinality of the *substrate transforms*, $|\mathcal{T}_{U,V}|$, is factorial in the *volume*, v. Tractable optimisations require the computation to be at most polynomial.

Note, also, that, although the *sensitivity* to *distribution*, E, is defined above for *uniform possible derived*, the *sensitivity* to *model*, T, is not yet defined.

2.6.2 Specialising coder induction

It is shown above that there are two *canonical history coders*, the *index history coder* C_{H} and the *classification coder* C_{G}. Given *variables* V and *size* z, the *index substrate history coder*, $C_{\text{H},U,V,z}$, encodes each *substrate history* $H \in \mathcal{H}_{U,V,z}$ in a fixed *space* of $C^{\text{s}}_{\text{H},U,V,z}(H) = z \ln v$, where *volume* $v = |V^{\text{C}}|$. By contrast, the *classification substrate history coder*, $C_{\text{G},U,V,z}$, encodes each *history* in a *space* which depends on the *histogram* $A = \text{his}(H)$,

$$C^{\text{s}}_{\text{G},U,V,z}(H) = \ln \frac{(z+v-1)!}{z!\ (v-1)!} + \ln \frac{z!}{\prod_{S \in A^{\text{S}}} A_S!}$$

When the *histogram entropy*, $\text{entropy}(A)$, is high the *classification space* is greater than the *index space*, $C^{\text{s}}_{\text{G},U,V,z}(H) > C^{\text{s}}_{\text{H},U,V,z}(H)$, but when the *entropy* is low the *classification space* is less than the *index space*, $C^{\text{s}}_{\text{G},U,V,z}(H) < C^{\text{s}}_{\text{H},U,V,z}(H)$. In the case where the *size* is much less than the *volume*, $z \ll v$, the break-even *sized entropy* is approximately $z \times \text{entropy}(A) \approx z \ln z$.

Given *substrate transform* $T \in \mathcal{T}_{U,V}$, the *specialising derived substrate history coder*, $C_{\mathrm{G,H},U,T,z}$, is intermediate between the *classification coder*, $C_{\mathrm{G},U,V,z}$, and the *index coder*, $C_{\mathrm{H},U,V,z}$. Given a *substrate history* $H \in \mathcal{H}_{U,V,z}$, the *derived history*, $H * T$, is encoded in a *classification coder*, $C_{\mathrm{G},U,W,z}$, where *derived variables* $W = \mathrm{der}(T)$. Then each *sub-history* H_C, corresponding to a *component* of the *partition*, $H_C \subseteq H$, where $(R, C) \in T^{-1}$, is encoded in a *index coder*, C_{H,U,C,z_C}, where $z_C = (A * T)_R$. The *specialising space* is

$$\begin{aligned} C^{\mathrm{s}}_{\mathrm{G,H},U,T,z}(H) \;=\; & \ln \frac{(z + w' - 1)!}{z!\,(w' - 1)!} + \ln \frac{z!}{\prod_{(R,\cdot)\in T^{-1}} (A * T)_R!} + \\ & \sum_{(R,C)\in T^{-1}} (A * T)_R \ln |C| \end{aligned}$$

where $w' = |T^{-1}|$.

In the case where the *transform* is *self*, $T = T_{\mathrm{s}}$ where $T_{\mathrm{s}} = V^{\mathrm{CS\{\}T}}$, then the *specialising space* equals the *classification space*, $C^{\mathrm{s}}_{\mathrm{G,H},U,T_{\mathrm{s}},z}(H) = C^{\mathrm{s}}_{\mathrm{G},U,V,z}(H)$. In the case where the *transform* is *unary*, $T = T_{\mathrm{u}}$ where $T_{\mathrm{u}} = \{V^{\mathrm{CS}}\}^{\mathrm{T}}$, then the *specialising space* equals the *index space*, $C^{\mathrm{s}}_{\mathrm{G,H},U,T_{\mathrm{u}},z}(H) = C^{\mathrm{s}}_{\mathrm{H},U,V,z}(H)$.

The *specialising space* depends only on the *transform*, T, and the *derived*, $A * T$. Define the *specialising space* function $\mathrm{sp}(T)(A * T) := C^{\mathrm{s}}_{\mathrm{G,H},U,T,z}(H)$.

The *specialising space* varies (i) with the *derived volume*, w', where the *derived volume* is less than the *size*, $w' < z$, otherwise with the *size* scaled *log derived volume*, $z \ln w'$, and (ii) against the *size* scaled *component size cardinality relative entropy*,

$$\begin{aligned} C^{\mathrm{s}}_{\mathrm{G,H},U,T,z}(H) \;\sim\; & (w' \;:\; w' < z) + (z \ln w' \;:\; w' \geq z) \\ & - z \times \mathrm{entropyRelative}(A * T, V^{\mathrm{C}} * T) \end{aligned}$$

In general, the *specialising space* is less than either of the two *canonical spaces* where the *derived entropy*, $\mathrm{entropy}(A * T)$, is low, but the *expected component entropy*, $\mathrm{entropyComponent}(A, T)$, is high. So the *specialising space* is minimised when (a) the *derived volume*, w', is minimised, (b) the *derived entropy*, $\mathrm{entropy}(A * T)$, is minimised, (c) high *size components* are low *cardinality components* and low *size components* are high *cardinality components*, and (d) the *expected component entropy* is maximised.

In *specialising induction* the *history probability functions* are constrained by *specialising space* which in turn depends on *derived histogram*.

In the discussion of 'Maximum Entropy', above, it was shown that, of a subset of the micro-state valued functions of distinguishable particle, the *maximum likelihood estimate* of the implied *probability function* is the *probability function* with the greatest entropy.

Consider a system of r undefined particles where the micro-state is a *substrate history*, $H \in \mathcal{H}_{U,V,z}$. The set of *substrate history* valued functions having exactly r particles with integer identifier is $\{1 \ldots r\} :\to \mathcal{H}_{U,V,z} \subset \mathcal{X} \to \mathcal{H}$. Given *substrate transform* $T \in \mathcal{T}_{U,V}$, let the subset $S \subset \{1 \ldots r\} :\to \mathcal{H}_{U,V,z}$ be such that the expected *specialising space* is a constant, $\forall R \in S\ (\sum(C^{\mathrm{s}}_{\mathrm{G,H},U,T,z}(H)/r : (\cdot, H) \in R) = \epsilon)$. Of this subset, S, the implied *probability function* with the greatest entropy, $\tilde{P} \in \mathrm{maxd}(\{(N, \mathrm{entropy}(N)) : R \in S,\ N = \{(H, |C|/r) : (H, C) \in R^{-1}\}\})$, approximates to a Boltzmann distribution.

Given *substrate transform* $T \in \mathcal{T}_{U,V}$, the *maximum likelihood estimate* $\tilde{P}$ of the *substrate history probability function* $P \in (\mathcal{H}_{U,V,z} :\to \mathbf{Q}_{\geq 0}) \cap \mathcal{P}$, which maximises the entropy, $\mathrm{entropy}(\tilde{P})$, is

$$\begin{aligned}
\tilde{P} &= \{(H, \exp(-C^{\mathrm{s}}_{\mathrm{G,H},U,T,z}(H))) : H \in \mathcal{H}_{U,V,z}\}^{\wedge} \\
&= \{(H, \exp(-\mathrm{sp}(T)(\mathrm{his}(H) * T))) : H \in \mathcal{H}_{U,V,z}\}^{\wedge} \\
&= \{(H, \frac{\exp(-\mathrm{sp}(T)(\mathrm{his}(H) * T))}{\sum \exp(-\mathrm{sp}(T)(\mathrm{his}(G) * T)) : G \in \mathcal{H}_{U,V,z}}) : H \in \mathcal{H}_{U,V,z}\}
\end{aligned}$$

where exp is the exponential function. The *likely* probability of a *history*, $\tilde{P}(H)$, is inversely proportional to the bounding integer, for which the *space* is the logarithm, of the integer encoding of the *history* in the *specialising coder*. The *maximum likelihood estimate*, $\tilde{P}$, is such that all *substrate histories* $H \in \mathcal{H}_{U,V,z}$ with the same *specialising space*, $C^{\mathrm{s}}_{\mathrm{G,H},U,T,z}(H)$, are equally probable and all *histories* are possible, $\tilde{P}(H) > 0$. If the *transform*, T, is *known*, then the *likely probability function*, $\tilde{P}$, is *known* and an approximation to the *expected specialising space*, ϵ, is *known*.

The *specialising space*, $\mathrm{sp}(T)(\mathrm{his}(H) * T) = C^{\mathrm{s}}_{\mathrm{G,H},U,T,z}(H)$, depends only on the *transform*, T, and the *derived*, $\mathrm{his}(H) * T$, so all *substrate histories* with the same *derived*, $\mathrm{his}(H) * T = A * T$, are equally probable. All *histories* are possible, $\tilde{P}(H) > 0$, so *specialising coder induction* is similar to *uniformly possible derived induction*, above, except that the *deriveds* are not necesssarily equally probable.

The *likely history probability function* entropy, $\text{entropy}(\tilde{P})$, is maximised when the expected numerator, $\exp(-\text{sp}(T)(\text{his}(H) * T))$, is minimised. The *expected specialising space* is $\sum(\tilde{P}(H) \times \text{sp}(T)(\text{his}(H)*T) : H \in \mathcal{H}_{U,V,z}) \approx \epsilon$, so the *likely history probability function* entropy varies with the *expected specialising space*, $\text{entropy}(\tilde{P}) \sim \epsilon$.

Now consider the case where, given *specialising*, it is *known*, in addition, that the *sample histogram* A is the most *probable histogram*. That is, the *likely probability* of *histogram* A,

$$\sum(\tilde{P}(H) : H \in \mathcal{H}_{U,V,z},\ \text{his}(H) = A) \ = \ \frac{z!}{\prod_{S \in A^{\text{S}}} A_S!} \times \frac{\exp(-\text{sp}(T)(A * T))}{\sum \exp(-\text{sp}(T)(\text{his}(G) * T)) : G \in \mathcal{H}_{U,V,z}}$$

is maximised.

The *specialising probability distribution* is defined

$$\hat{Q}_{\text{G,H},T,U}(z) \ := \ \{(A, \frac{z!}{\prod_{S \in A^{\text{S}}} A_S!} \times \exp(-\text{sp}(T)(A * T))) : A \in \mathcal{A}_{U,\text{i},V,z}\}^{\wedge}$$

The *specialising log likelihood* varies (i) with the *size* scaled *underlying entropy* (ii) against the *derived volume*, w', where the *derived volume* is less than the *size*, $w' < z$, otherwise against the *size* scaled *log derived volume*, $z \ln w'$, and (iii) with the *size* scaled *component size cardinality relative entropy*,

$$\begin{aligned} \ln \hat{Q}_{\text{G,H},T,U}(z)(A) \ \sim \ & z \times \text{entropy}(A) \\ & - ((w' \ : \ w' < z) \ + \ (z \ln w' \ : \ w' \geq z)) \\ & + z \times \text{entropyRelative}(A * T, V^{\text{C}} * T) \end{aligned}$$

In other words, the *log likelihood* is maximised where (i) the *derived volume*, w', is minimised, (ii) the *derived entropy*, $\text{entropy}(A * T)$, is minimised, (iii) the *cross entropy*, $\text{entropyCross}(A * T, V^{\text{C}} * T)$, is maximised, so that high *counts* are in low cardinality *components* and high cardinality *components* have low *counts*, and (iv) the *expected component entropy*, $\text{entropyComponent}(A, T)$, is maximised.

In the case of *probable sample*, the *likely history probability function* entropy varies against the *relative entropy*, $\text{entropy}(\tilde{P}) \sim -\,\text{entropyRelative}(A * T, V^{\text{C}} * T)$. Similarly, the *expected specialising space* varies against the *relative entropy*, $\epsilon \sim -\,\text{entropyRelative}(A * T, V^{\text{C}} * T)$.

Given *specialising* and *probable sample*, consider the case where the *histogram* A is *known*, but the *transform*, T, is *unknown*, and hence the *likely history probability function*, $\tilde{P}$, is *unknown*. The *maximum likelihood estimate* $\tilde{T}$ for the *transform*, T, is approximated by a modal value of the *specialising likelihood*,

$$\tilde{T} \;\in\; \text{maxd}(\{(M, \hat{Q}_{\text{G},\text{H},M,U}(z)(A)) : M \in \mathcal{T}_{U,V}\})$$

Note that, as in the case of *uniform possible derived induction*, although computable, this optimisation is intractable because the cardinality of the *substrate transforms*, $|\mathcal{T}_{U,V}|$, is factorial in the *volume*, v.

Unlike *uniform possible derived induction*, in *specialising induction* there is no *distribution history*, H_E, and so no *sensitivity* to *distribution*, E. A *sensitivity* to *model*, T, can be defined, however, as the negative logarithm of the cardinality of the *maximum likelihood estimate models*,

$$-\;\ln |\text{max}(\{(M, \hat{Q}_{\text{G},\text{H},M,U}(z)(A)) : M \in \mathcal{T}_{U,V}\})|$$

That is, as the cardinality of the modal *models* of the *log likelihood* function increases, the *sensitivity* to *model* decreases. It can be shown that the *sensitivity* to *model* varies against the *size-volume* scaled *component size cardinality sum relative entropy*,

$$\begin{aligned} -\;\ln |\text{max}(\{(M, \hat{Q}_{\text{G},\text{H},M,U}(z)(A)) : M \in \mathcal{T}_{U,V}\})| \;\sim& \\ -\big((z+v) \times \text{entropy}(A * T + V^{\text{C}} * T)& \\ -z \times \text{entropy}(A * T) \;-v \times \text{entropy}(V^{\text{C}} * T)\big)& \end{aligned}$$

So the *sensitivity* to *model* varies against the *log likelihood*,

$$-\;\ln |\text{max}(\{(M, \hat{Q}_{\text{G},\text{H},M,U}(z)(A)) : M \in \mathcal{T}_{U,V}\})| \quad\sim\quad -\;\ln \hat{Q}_{\text{G},\text{H},T,U}(z)(A)$$

As the *relative entropy*, $\text{entropyRelative}(A * T, V^{\text{C}} * T)$, increases, the *log-likelihood*, $\ln \hat{Q}_{\text{G},\text{H},T,U}(z)(A)$, increases, but the *sensitivity* to *model*, T, decreases. In other words, the higher the *sample relative entropy* the more *likely* the *maximum likelihood estimate*, $\tilde{T}$, equals the *model*, T, and the smaller the *likely* difference between them if they are not equal.

It is shown above, in the case of *uniform possible derived* and *natural sample*, $A = A * T * T^{\dagger}$, that the *log likelihood* varies against the *derived volume* and with the *size-volume* scaled *component size cardinality sum relative* *entropy*,

$$
\begin{aligned}
&\ln \hat{Q}_{\mathrm{m,d},T,U}(A,z)(A) \ \sim \\
&\quad -\ ((w' \ :\ w' < z) \ +\ (z \ln w' \ :\ w' \geq z)) \\
&\quad +\ (z+v) \times \mathrm{entropy}(A * T + V^{\mathrm{C}} * T) \\
&\qquad\qquad\qquad -z \times \mathrm{entropy}(A * T) \ - v \times \mathrm{entropy}(V^{\mathrm{C}} * T)
\end{aligned}
$$

so the *iso-derived conditional log likelihood* varies with the *specialising log likelihood*,

$$\ln \hat{Q}_{\mathrm{m,d},T,U}(A,z)(A) \ \ \sim \ \ \ln \hat{Q}_{\mathrm{G,H},T,U}(z)(A)$$

and the *iso-derived conditional model sensitivity* varies against the *iso-derived conditional log likelihood*,

$$
\begin{aligned}
&-\ \ln |\max(\{(M, \hat{Q}_{\mathrm{m,d},M,U}(A,z)(A)) : M \in \mathcal{T}_{U,V},\ A = A * M * M^{\dagger}\})| \ \sim \\
&\qquad\qquad\qquad\qquad\qquad\qquad -\ \ln \hat{Q}_{\mathrm{m,d},T,U}(A,z)(A)
\end{aligned}
$$

The *iso-derived conditional model sensitivity* may be compared to the *iso-derived conditional distribution sensitivity* which also varies against the *iso-derived conditional log likelihood*,

$$\mathrm{sum}(\mathrm{sensitivity}(U)(\hat{Q}_{\mathrm{m,d},T,U}(A,z))) \ \ \sim \ \ -\ln \hat{Q}_{\mathrm{m,d},T,U}(A,z)(A)$$

That is, in *classical modelled induction*, the *log likelihood* is maximised and the *sensitivities* to both *distribution* and *model* are minimised where (i) the *derived volume* is minimised, (ii) the *derived entropy* is minimised, (iii) the *cross entropy* is maximised, so that high *counts* are in low cardinality *components* and high cardinality *components* have low *counts*, and (iv) the *expected component entropy* is maximised.

2.6.3 Artificial neural networks

In the discussion of *classical modelled induction*, above, it is shown that, given *uniform possible derived* and *probable sample* $A \in \mathcal{A}_{U,V,z}$, where the *sample* is *natural*, $A = A * T * T^{\dagger}$, the *maximum likelihood estimate* $\tilde{T}$ for *unknown transform* $T \in \mathcal{T}_{U,V}$, is

$$\tilde{T} \ \ \in \ \ \mathrm{maxd}(\{(M, \hat{Q}_{\mathrm{m,d},M,U}(A,z)(A)) : M \in \mathcal{T}_{U,V},\ A = A * M * M^{\dagger}\})$$

Similarly, given *specialising* and *probable sample*, the *maximum likelihood estimate*, $\tilde{T}$, for the *transform*, T, is approximated by a modal value of the *specialising likelihood*,

$$\tilde{T} \ \ \in \ \ \mathrm{maxd}(\{(M, \hat{Q}_{\mathrm{G,H},M,U}(z)(A)) : M \in \mathcal{T}_{U,V}\})$$

In both cases, although computable, the optimisations are intractable because the cardinality of the *substrate transforms*, $|\mathcal{T}_{U,V}|$, is factorial in the *volume*, v. In order to make the optimisation tractable and then practicable, the search must be restricted to a subset of the *models*.

Artificial neural network induction is an example of practicable *classical modelled induction*. Here the *models* are artificial neural networks which correspond to *functional definition sets* of *transforms* representing the neurons. The optimisation consists of a sequence of these networks. The graph of the network remains constant, but the weights between neurons of successive networks are altered to decrease a loss function step by step. The weights of the initial network are chosen at random. The optimisation proceeds until the loss falls below a threshold. The *fud* of the terminating network is then the practicable *model*. The network graph is chosen depending on the given *sample*. In some cases of configuration the *entropy* properties of the resultant *model* are those of *classical induction*.

The *one functional transforms*, $\mathcal{T}_{U,\mathrm{f},1}$, are *derived state* valued left total functions of *underlying state*,

$$\forall T \in \mathcal{T}_{U,\mathrm{f},1}\ (\mathrm{split}(V, X^{\mathrm{S}}) \in V^{\mathrm{CS}} :\to W^{\mathrm{CS}})$$

where $(X, W) = T$ and $V = \mathrm{und}(T)$. In order to construct a coordinate from a *state* define $()^{\square} \in \mathcal{S} \to \mathcal{L}(\mathcal{W})$ as

$$S^{\square} := \{(i, u) : ((v, u), i) \in \mathrm{order}(D_{\mathcal{V}\times\mathcal{W}}, S)\}$$

where $D_{\mathcal{V}\times\mathcal{W}}$ is an *order* on the *variables* and *values*. The converse function to construct a *state* from a coordinate $()^{V} \in \mathcal{L}(\mathcal{W}) \to \mathcal{S}$ is

$$S^{V} := \{(v, S_i) : (v, i) \in \mathrm{order}(D_{\mathcal{V}}, V)\}$$

Now *one functional transforms* may be represented as *derived value* coordinate valued left total functions of *underlying value* coordinate,

$$\begin{aligned}\{(S^{\square}, R^{\square}) : (S, R) \in \mathrm{split}(V, X^{\mathrm{S}})\} &\in \{S^{\square} : S \in V^{\mathrm{CS}}\} :\to \{R^{\square} : R \in W^{\mathrm{CS}}\} \\ &\subset \mathcal{W}^{n} \to \mathcal{W}^{m}\end{aligned}$$

where $n = |V|$ and $m = |W|$.

So an alternative definition for a *one functional transform* is a tuple of (i) the *underlying variables*, V, (ii) the *derived variables*, W, and (iii) a *derived value* coordinate valued left total function of *underlying value* coordinate, f,

$$\begin{aligned}
&\mathcal{T}_{U,\mathrm{f},1} = \\
&\quad \{(V, W, f) : V, W \in \mathrm{P}(\mathrm{vars}(U)),\ V \cap W = \emptyset, \\
&\qquad\qquad f \in \{S^{\square} : S \in V^{\mathrm{CS}}\} :\to \{R^{\square} : R \in W^{\mathrm{CS}}\}\}
\end{aligned}$$

The *histogram* of a *function-defined one functional transform* $T = (V, W, f) \in \mathcal{T}_{U,\mathrm{f},1}$ is

$$\mathrm{histogram}(T) \ := \ \{S \cup f(S^{\square})^W : S \in V^{\mathrm{CS}}\} \times \{1\}$$

In the special case where the *transform* is *mono-derived-variate*, $T = (V, \{w\}, f)$, the function may be simplified to $f \in \{S^{\square} : S \in V^{\mathrm{CS}}\} :\to U_w$, and the *histogram* is

$$\mathrm{histogram}(T) := \{S \cup \{(w, f(S^{\square}))\} : S \in V^{\mathrm{CS}}\} \times \{1\}$$

In the further special case of *mono-derived-variate transform* where its *variables* are real, $\forall v \in V\ (U_v = \mathbf{R})$ and $U_w = \mathbf{R}$, then the function is a real valued left total function of a real coordinate, $f \in \mathbf{R}^n :\to \mathbf{R}$. Here the *cartesian states* are $V^{\mathrm{CS}} = \prod_{v \in V}(\{v\} \times \mathbf{R})$, so the *histogram* is

$$\begin{aligned}
\mathrm{histogram}(T) &= \{S \cup \{(w, f(S^{\square}))\} : S \in \prod_{v \in V}(\{v\} \times \mathbf{R})\} \times \{1\} \\
&= \{S^V \cup \{(w, f(S))\} : S \in \mathbf{R}^n\} \times \{1\}
\end{aligned}$$

The *cartesian volume* is infinite, $|V^{\mathrm{C}}| = |\mathbf{R}^n|$, so the cardinality of the *histogram* is infinite, $|\mathrm{histogram}(T)| = |\mathbf{R}^n|$.

The reals form a metric space so a real valued function of real coordinates may be discretised given a finite subset of the reals $D \subset \mathbf{R} : |D| < \infty$. The discretised function is

$$\mathrm{discrete}(D, n)(f) := \{(X, \mathrm{nearest}(D, f(X))) : X \in D^n\} \in D^n :\to D$$

where $\mathrm{nearest} \in \mathrm{P}(\mathbf{R}) \times \mathbf{R} \to \mathbf{R}$ is defined

$$\mathrm{nearest}(D, r) := t : \{t\} \in \mathrm{mind}(\{(s, (|r - s|, s)) : s \in D\})$$

The cardinality of the discretised *transform's histogram* is finite,

$$|\mathrm{histogram}((V, \{w\}, \mathrm{discrete}(D, n)(f)))| = |D^n| = |D|^n$$

An example of a *transform* defined by a real valued function occurs in the function composition of artificial neural networks. Here a *transform* represents a model of a neuron called a perceptron, $T = (V, w, f_\sigma(Q))$, where the *dimension* is $n = |V|$ and the function $f_\sigma(Q) \in \mathbf{R}^n :\to \mathbf{R}$ is parameterised by (i) some differentiable function $\sigma \in \mathbf{R} :\to \mathbf{R}$, called the activation function, and (ii) a vector of weights, $Q \in \mathbf{R}^{n+1}$, and is defined

$$f_\sigma(Q)(S) := \sigma(\sum_{i \in \{1 \ldots n\}} Q_i S_i \ + \ Q_{n+1})$$

The function composition of artificial neural networks may be represented by *fuds* of these *transforms*. Define nets as a subset of the set of lists of tuples of the graph and real weights,

$$\text{nets} \ := \ \{G : G \in \mathcal{L}(\mathrm{P}(\mathcal{V}) \times \mathcal{V} \times \mathcal{L}(\mathbf{R})), \ \forall (\cdot, (V, \cdot, Q)) \in G \ (|Q| = |V| + 1)\}$$

Define the set of *transforms*, $\mathrm{fud}(\sigma) \in \text{nets} \to \mathrm{P}(\mathcal{T}_{\mathrm{f}})$ as

$$\begin{aligned}\mathrm{fud}(\sigma)(G) \ &:= \\ &\{(\{S^V \cup \{(w, f_\sigma(Q)(S))\} : S \in \mathbf{R}^n\} \times \{1\}, \{w\}) : \\ &\qquad (\cdot, (V, w, Q)) \in G, \ n = |V|\}\end{aligned}$$

The *fud* search is restricted to the *neural net substrate fud set*, $\mathcal{F}_{\infty,U,V,\sigma} = \mathcal{F}_{\infty,U,V} \cap (\mathrm{fud}(\sigma) \circ \text{nets})$.

An example of a *neural net substrate fud* $F \in \mathcal{F}_{\infty,U,V,\sigma}$ has $l = \mathrm{layer}(F, \mathrm{der}(F))$ *layers* of fixed *breadth* equal to the *underlying dimension*, $\forall i \in \{1 \ldots l\} \ (|F_i| = n)$ where $n = |V|$ and $F_i = \{T : T \in F, \ \mathrm{layer}(F, \mathrm{der}(T)) = i\}$, such that the *underlying* of each *transform* is the *derived* of the *layer* below, $\forall T \in F_1 \ (\mathrm{und}(T) = V)$ and $\forall i \in \{2 \ldots l\} \ \forall T \in F_i \ (\mathrm{und}(T) = \mathrm{der}(F_{i-1}))$.

The optimisation of artificial neural networks can be divided into unsupervised and supervised types. In the supervised case there is additional *knowledge*. First, there exists an *unknown distribution histogram* E from which the *known sample histogram*, A, is *drawn*, $A < E$. Secondly, the *substrate* can be partitioned into query *variables* $K \subset V$ and label *variables*, $V \setminus K$, such that the *distribution histogram*, E, is *causal* between the query *variables* and the label *variables*,

$$\mathrm{split}(K, E^{\mathrm{FS}}) \ \in \ K^{\mathrm{CS}} \to (V \setminus K)^{\mathrm{CS}}$$

and so the *sample histogram*, A, is also *causal*,

$$\mathrm{split}(K, A^{\mathrm{FS}}) \ \in \ K^{\mathrm{CS}} \to (V \setminus K)^{\mathrm{CS}}$$

That is, in the supervised case, there is a functional relation such that there is exactly one label *state* for every *effective* query *state*. In an optimisation, a *fud* $F \in \mathcal{F}_{\infty,U,K,\sigma}$ has its *underlying variables* restricted to the query *variables*, $\mathrm{und}(F) \subseteq K$. The optimisation maximises the *causality* between the *derived variables* and the label *variables* by minimising the loss function. At the optimum there is no error and the relation is functional,

$$\mathrm{split}(W_F, (A * X_F \ \% \ (W_F \cup V \setminus K))^{\mathrm{FS}}) \ \in \ W_F^{\mathrm{CS}} \to (V \setminus K)^{\mathrm{CS}}$$

where $X_F = \mathrm{histogram}(F^{\mathrm{T}})$ and $W_F = \mathrm{der}(F)$. At zero loss the label *state* is implied for all query *states* that are *effective* in the *sample derived*,

$$\mathrm{split}(K, (K^{\mathrm{C}} * F^{\mathrm{T}} * (A * X_F) \ \% \ (V \setminus K))^{\mathrm{FS}}) \ \in \ K^{\mathrm{CS}} \to (V \setminus K)^{\mathrm{CS}}$$

That is, a query *state* $Q \in K^{\mathrm{CS}}$ that is *effective* in the *sample derived* $R \in (A * F^{\mathrm{T}})^{\mathrm{FS}}$, where $\{R\} = (\{Q\}^{\mathrm{U}} * F^{\mathrm{T}})^{\mathrm{FS}}$, but that is not necessarily *effective* in the *sample* itself, $Q \notin (A\%K)^{\mathrm{FS}}$, still has an implied label *state*, $\{L\} = (A * X_F * \{R\}^{\mathrm{U}} \ \% \ (V \setminus K))^{\mathrm{FS}}$ where $L \in (V \setminus K)^{\mathrm{CS}}$.

In the case where the *derived variables* of the *fud* is a *literal frame* of the label *variables*, $W_F :\leftrightarrow: (V \setminus K)$ and $\forall v \in (V \setminus K)\ (U_v \subseteq \mathbf{R})$, the least squares loss function $\mathrm{lsq} \in \mathcal{A} \times \mathcal{F} \times \mathrm{P}(\mathcal{V}) \to \mathbf{R}$ is

$$\mathrm{lsq}(A, F, K) \ := \ \sum_{(S,c) \in A * X_F} \Big(c \times \sum_{i \in \{1 \ldots m\}} ((S\%W_F)^{[]}_i - (S\%(V \setminus K))^{[]}_i)^2\Big)$$

where $m = |W_F| = |(V \setminus K)|$. The loss function is a continuous real valued function and so its derivative with respect to each weight can be defined. In this case the optimisation is least squares gradient descent.

If the optimisation of artificial neural networks is of the unsupervised type, there is no *knowledge* of a *causal* label. Here the method of least squares gradient descent is still used but the label is simply a copy of the *substrate*, V, itself. Usually the network graph is constrained so that a middle *layer* $a \in \{2 \ldots l-1\}$ has narrower *breadth* than the *substrate*, $|F_a| < n$.

In the computations of *alignment* and *entropy* that follow, the *derived variables* are *discretised* to the *values* of the label *variables*, $D = \cup\{U_v : v \in (V \setminus K)\}$.

In some cases of *sample* and network optimisation configuration, the negative least squares loss (a) varies against the *effective derived volume*

$$-\ \mathrm{lsq}(A, F_D, K) \ \sim \ -\ |(A * F_D^{\mathrm{T}})^{\mathrm{F}}|$$

(b) varies against the *derived entropy* of the *fud transform*,

$$-\ \text{lsq}(A, F_D, K) \quad \sim \quad -\ \text{entropy}(A * F_D^{\mathrm{T}})$$

(c) varies with the *component size cardinality relative entropy*,

$$-\ \text{lsq}(A, F_D, K) \quad \sim \quad \text{entropyRelative}(A * F_D^{\mathrm{T}}, V^{\mathrm{C}} * F_D^{\mathrm{T}})$$

and (d) varies with the *expected component entropy*,

$$-\ \text{lsq}(A, F_D, K) \quad \sim \quad \text{entropyComponent}(A, F_D^{\mathrm{T}})$$

The initial *fud* F_R has arbitrary weights, so is likely to have a high least squares loss. That is, far from the *derived variables* and the label *variables* being *causally* related, $W_D^{\mathrm{CS}} \to (V \setminus K)^{\mathrm{CS}}$, they are likely to be *independent*,

$$\text{algn}(A * X_{F_R} * \{W_D^{\mathrm{CS}\{\}\mathrm{T}}, (V \setminus K)^{\mathrm{CS}\{\}\mathrm{T}}\}^{\mathrm{T}}) \quad \approx \quad 0$$

where $\{W_D^{\mathrm{CS}\{\}\mathrm{T}}, (V \setminus K)^{\mathrm{CS}\{\}\mathrm{T}}\}$ is the *fud* of the *self transforms* of the (i) *discretised derived variables* and (ii) label *variables*.

As the optimisation proceeds from the initial *fud*, F_R, to the optimal *fud* F, the loss decreases and the relation between the top *layer* and the label becomes more *causal*,

$$\text{algn}(A * X_F * \{W_D^{\mathrm{CS}\{\}\mathrm{T}}, (V \setminus K)^{\mathrm{CS}\{\}\mathrm{T}}\}^{\mathrm{T}}) \quad > \quad 0$$

The negative least squares loss varies with the *alignment* of the *self partition transforms*, so varies against the *derived entropy* of the *fud transform*,

$$\begin{aligned} -\ \text{lsq}(A, F_D, K) \quad &\sim \quad \text{algn}(A * X_F * \{W_D^{\mathrm{CS}\{\}\mathrm{T}}, (V \setminus K)^{\mathrm{CS}\{\}\mathrm{T}}\}^{\mathrm{T}}) \\ &\sim \quad -z \times \text{entropy}(A * F_D^{\mathrm{T}}) \end{aligned}$$

That is, as the loss, $\text{lsq}(A, F_D, K)$, is minimised, the *derived entropy*, $\text{entropy}(A * F_D^{\mathrm{T}})$, tends to be minimised. The minimisation of *derived entropy* is a property of *classical induction*.

The negative least squares loss only varies with the *component size cardinality relative entropy*, $\text{entropyRelative}(A * F_D^{\mathrm{T}}, V^{\mathrm{C}} * F_D^{\mathrm{T}})$, in the case where the *histogram*, A, is clustered by the label *variables*. This requires *alignment* within the query *variables*, $\text{algn}(A\%K) > 0$. Clustering may be described as follows.

Consider the case of a *multi-variate* set of real *valued* query *variables* K, where $k = |K| \geq 2$ and $\forall x \in K\ (U_x \subseteq \mathbf{R})$, and a *neural net fud* $F \in \mathcal{F}_{\infty,U,K,\sigma}$ consisting of two *transforms*, $F = \{T_1, T_2\}$, each having the query *variables* as the *underlying*, $\mathrm{und}(T_1) = \mathrm{und}(T_2) = K$. For a coordinate $S \in \mathbf{R}^k$ the weights of the *transforms* form a pair of hyperplanes,

$$\sum_{i \in \{1 \ldots k\}} Q_{1,i} S_i \ + \ Q_{1,k+1} \ = \ 0$$

and

$$\sum_{i \in \{1 \ldots k\}} Q_{2,i} S_i \ + \ Q_{2,k+1} \ = \ 0$$

where $Q_1, Q_2 \in \mathbf{R}^{k+1}$ are the weights correspondng to T_1, T_2. If the hyperplanes of the arbitrarily weighted initial *fud*, F_R, intersect, the acute angle between them is expected to be 45°. That is, given an activation function, σ, which is a step function, or a binary set of *discrete values*, $D = \{0, 1\}$, the probability distribution of the *component cardinalities* of the initial *fud* is bi-modal. If $(\cdot, C_1), (\cdot, C_2) \in (F^{\mathrm{T}}_{R,\{0,1\}})^{-1}$ are such that $|C_1| < |C_2|$, then it is expected that $3|C_1| = |C_2|$. So the *component cardinality entropy* of the initial *fud* is expected to be less than maximal,

$$\mathrm{entropy}(K^{\mathrm{C}} * F^{\mathrm{T}}_{R,D}) \ < \ \mathrm{entropy}(W^{\mathrm{C}}_D)$$

The *derived entropy* of the initial *fud* is expected to be approximately equal to the *component cardinality entropy*,

$$\mathrm{entropy}(A * F^{\mathrm{T}}_{R,D}) \ \approx \ \mathrm{entropy}(K^{\mathrm{C}} * F^{\mathrm{T}}_{R,D})$$

and so the *component size cardinality relative entropy* of the initial *fud* is expected to be small,

$$\mathrm{entropyRelative}(A * F^{\mathrm{T}}_{R,D}, K^{\mathrm{C}} * F^{\mathrm{T}}_{R,D}) \ \approx \ 0$$

If the *histogram*, A, is approximately uniformly distributed over the *volume*, then the *component size cardinality relative entropy* remains small during the optimisation,

$$\mathrm{entropyRelative}(A * F^{\mathrm{T}}_{D}, K^{\mathrm{C}} * F^{\mathrm{T}}_{D}) \ \approx \ 0$$

In contrast, consider the case where the *histogram*, A, is not uniformly distributed, but clustered by label *state*. Let $Y_L \subset K^{\mathrm{CS}}$ be the set of the centres of the clusters for *effective* label *state* $L \in (A\%(V \setminus K))^{\mathrm{FS}}$. The maximum radius $r_L \in \mathbf{R}_{>0}$ is such that

$$\forall S \in A^{\mathrm{FS}} \ \Diamond L = S\%(V \setminus K) \ \exists Q \in Y_L \ (\sum_{i \in \{1 \ldots k\}} (Q^{[]}_i - S^{[]}_i)^2 \ \leq \ r_L^2)$$

Let r_C be the radius of *component* C. In the case where the *histogram* is clustered such that the cluster radius of a label *state* is much smaller than the least initial *component* radius, $\forall (\cdot, C) \in (F^{\mathrm{T}}_{R,\{0,1\}})^{-1}$ $(r_L \ll r_C)$, then optimised rotations of the hyperplanes, that sweep up nearby clusters in the same label *state*, tend to be such that the magnitude of the change in the fractional *component size*, $|(A * F^{\mathrm{T}}_{2,D})(R) - (A * F^{\mathrm{T}}_{1,D})(R)|/z$, is greater than magnitude of the change in the fractional *component cardinality*, $|(K^{\mathrm{C}} * F^{\mathrm{T}}_{2,D})(R) - (K^{\mathrm{C}} * F^{\mathrm{T}}_{1,D})(R)|/|K^{\mathrm{C}}|$. So, in the clustered case, as the optimisation decreases the *derived entropy*, $\mathrm{entropy}(A * F^{\mathrm{T}}_D)$, the *component sizes* and *component cardinalities* become less synchronised and the *component size cardinality relative entropy* increases,

$$\begin{aligned} -\ \mathrm{lsq}(A, F_D, K) &\sim -z \times \mathrm{entropy}(A * F^{\mathrm{T}}_D) \\ &\sim z \times \mathrm{entropyRelative}(A * F^{\mathrm{T}}_D, K^{\mathrm{C}} * F^{\mathrm{T}}_D) \\ &= z \times \mathrm{entropyRelative}(A * F^{\mathrm{T}}_D, V^{\mathrm{C}} * F^{\mathrm{T}}_D) \end{aligned}$$

The same reasoning applies to *fuds* consisting of more than two *transforms*, $|F| > 2$, but note that at higher *fud* cardinalities the initial *component cardinality entropy*, $\mathrm{entropy}(K^{\mathrm{C}} * F^{\mathrm{T}}_{R,D})$, tends to be multi-modal and so approximates more closely to the *uniform cartesian derived entropy*, $\mathrm{entropy}(W^{\mathrm{C}}_D)$. So there is less freedom for the *relative entropy* of the *fud* to increase during optimisation. In the case of *multi-layer fuds*, however, the *breadth* can be constrained and so the *relative entropy* of deeper, narrrower *fuds* may be higher than in shallower, wider *fuds* of the same cardinality.

In general, in the clustered case, the optimised *fud* is such that high *counts* are in low cardinality *components* and high cardinality *components* have low *counts*. The maximisation of *relative entropy* is a property of *classical induction.*

The *accuracy* of the approximation of *artificial neural network induction* to *classical induction* can be defined as the ratio of the practicable *model sample-distributed iso-derived conditional log likelihood* to the maximum *model sample-distributed iso-derived conditional log likelihood*,

$$0 \ < \ \frac{\hat{Q}_{\mathrm{m,d},F^{\mathrm{T}},U}(A,z)(A)}{\hat{Q}_{\mathrm{m,d},\tilde{T},U}(A,z)(A)} \ \leq \ 1$$

The *accuracy* varies against the *sensitivity* to *model*,

$$\frac{\hat{Q}_{\mathrm{m,d},F^{\mathrm{T}},U}(A,z)(A)}{\hat{Q}_{\mathrm{m,d},\tilde{T},U}(A,z)(A)} \quad \sim \quad -(-\ln|\max(\{(M,\hat{Q}_{\mathrm{m,d},M,U}(A,z)(A)) : M \in \mathcal{T}_{U,V}\})|)$$

and so varies with the *log-likelihood*,

$$\frac{\hat{Q}_{\mathrm{m,d},F^{\mathrm{T}},U}(A,z)(A)}{\hat{Q}_{\mathrm{m,d},\tilde{T},U}(A,z)(A)} \quad \sim \quad \ln \hat{Q}_{\mathrm{m,d},T,U}(A,z)(A)$$

That is, although the *model* obtained from *least squares gradient descent* is merely an approximation, in the cases where the *log-likelihood* is high, and so the *sensitivity* to *model* is low, the approximation may be reasonably close nonetheless.

2.6.4 Aligned induction

Given *substrate transform* $T \in \mathcal{T}_{U,V}$, the *abstract histogram* valued *integral substrate histograms* function $Y_{U,\mathrm{i},T,\mathrm{W},z}$ is defined

$$Y_{U,\mathrm{i},T,\mathrm{W},z} \quad := \quad \{(A,(A*T)^{\mathrm{X}}) : A \in \mathcal{A}_{U,\mathrm{i},V,z}\}$$

The finite set of *iso-abstracts* of *abstract histogram* $(A*T)^{\mathrm{X}}$ is

$$Y^{-1}_{U,\mathrm{i},T,\mathrm{W},z}((A*T)^{\mathrm{X}}) \quad = \quad \{B : B \in \mathcal{A}_{U,\mathrm{i},V,z},\ (B*T)^{\mathrm{X}} = (A*T)^{\mathrm{X}}\}$$

The degree to which an *integral iso-set* $I \subseteq \mathcal{A}_{U,\mathrm{i},V,z}$ that contains the *histogram*, $A \in I$, is said to be *entity-like* is called the *iso-abstractence*. The *iso-abstractence* is defined as the ratio of (i) the cardinality of the intersection between the *integral iso-set* and the set of *integral iso-abstracts*, and (ii) the cardinality of the union,

$$\frac{1}{|\mathcal{A}_{U,\mathrm{i},V,z}|} \ \leq \ \frac{|I \ \cap \ Y^{-1}_{U,\mathrm{i},T,\mathrm{W},z}((A*T)^{\mathrm{X}})|}{|I \ \cup \ Y^{-1}_{U,\mathrm{i},T,\mathrm{W},z}((A*T)^{\mathrm{X}})|} \ \leq \ 1$$

Law-like iso-sets are subsets of the set of *iso-abstracts*,

$$D^{-1}_{U,\mathrm{i},T,z}(A*T) \quad \subseteq \quad Y^{-1}_{U,\mathrm{i},T,\mathrm{W},z}((A*T)^{\mathrm{X}})$$

and so are also *entity-like*.

The *formal histogram* valued *integral substrate histograms* function $Y_{U,\mathrm{i},T,\mathrm{V},z}$ is defined

$$Y_{U,\mathrm{i},T,\mathrm{V},z} \ := \ \{(A, A^{\mathrm{X}} * T) : A \in \mathcal{A}_{U,\mathrm{i},V,z}\}$$

The finite set of *iso-formals* of *formal histogram* $A^{\mathrm{X}} * T$ is

$$Y^{-1}_{U,\mathrm{i},T,\mathrm{V},z}(A^{\mathrm{X}} * T) \ = \ \{B : B \in \mathcal{A}_{U,\mathrm{i},V,z},\ B^{\mathrm{X}} * T = A^{\mathrm{X}} * T\}$$

Aligned-like iso-sets are subsets of the set of *iso-formals*,

$$Y^{-1}_{U,\mathrm{i},V,z}(A^{\mathrm{X}}) \ \subseteq \ Y^{-1}_{U,\mathrm{i},T,\mathrm{V},z}(A^{\mathrm{X}} * T)$$

The *formal-abstract* pair valued *integral substrate histograms* function $Y_{U,\mathrm{i},T,z}$ is defined

$$Y_{U,\mathrm{i},T,z} \ := \ \{(A, (A^{\mathrm{X}} * T, (A * T)^{\mathrm{X}})) : A \in \mathcal{A}_{U,\mathrm{i},V,z}\}$$

The finite set of *iso-transform-independents* of $(A^{\mathrm{X}} * T, (A * T)^{\mathrm{X}})$ is

$$Y^{-1}_{U,\mathrm{i},T,z}((A^{\mathrm{X}} * T, (A * T)^{\mathrm{X}})) \ = \ \{B : B \in \mathcal{A}_{U,\mathrm{i},V,z},\ B^{\mathrm{X}} * T = A^{\mathrm{X}} * T,\ (B * T)^{\mathrm{X}} = (A * T)^{\mathrm{X}}\}$$

The *iso-transform-independents* is the intersection of the *iso-formals* and the *iso-abstracts*,

$$Y^{-1}_{U,\mathrm{i},T,z}((A^{\mathrm{X}} * T, (A * T)^{\mathrm{X}})) \ = \ Y^{-1}_{U,\mathrm{i},T,\mathrm{V},z}(A^{\mathrm{X}} * T) \ \cap \ Y^{-1}_{U,\mathrm{i},T,\mathrm{W},z}((A * T)^{\mathrm{X}})$$

In *aligned modelled induction* the *history probability functions* are constrained by *formal* and *abstract histograms*.

Let P be a *substrate history probability function*, $P \in (\mathcal{H}_{U,V,z} :\to \mathbf{Q}_{\geq 0}) \cap \mathcal{P}$. Given a *history* $H_E \in \mathcal{H}_{U,V,z_E}$, of *size* $z_E = |H_E|$, consider the case where both the *formal histogram* $A^{\mathrm{X}} * T$ of *drawn histories* is *known* to be *necessary* and the *abstract histogram* $(A * T)^{\mathrm{X}}$ of *drawn histories* is *known* to be *necessary*, $\sum(P(H) : H \subseteq H_E,\ \mathrm{his}(H)^{\mathrm{X}} * T = A^{\mathrm{X}} * T,\ (\mathrm{his}(H) * T)^{\mathrm{X}} = (A * T)^{\mathrm{X}}) = 1$. The *maximum likelihood estimate* which maximises the entropy, $\mathrm{entropy}(\tilde{P})$, is

$$
\begin{aligned}
\tilde{P} \;=\; & \{(H,1) : \\
& \quad H \subseteq H_E,\ \mathrm{his}(H)^{\mathrm{X}} * T = A^{\mathrm{X}} * T,\ (\mathrm{his}(H) * T)^{\mathrm{X}} = (A * T)^{\mathrm{X}}\}^{\wedge}\ \cup \\
& \{(G,0) : G \in \mathcal{H}_{U,V,z},\ G \not\subseteq H_E\}\ \cup \\
& \{(G,0) : G \in \mathcal{H}_{U,V,z},\ \mathrm{his}(G)^{\mathrm{X}} * T \neq A^{\mathrm{X}} * T\}\ \cup \\
& \{(G,0) : G \in \mathcal{H}_{U,V,z},\ (\mathrm{his}(G) * T)^{\mathrm{X}} \neq (A * T)^{\mathrm{X}}\} \\
=\; & \{(H, 1/\textstyle\sum(Q_{\mathrm{h},U}(E,z)(B) : B \in Y^{-1}_{U,\mathrm{i},T,z}((A^{\mathrm{X}} * T, (A * T)^{\mathrm{X}})))) : \\
& \quad H \subseteq H_E,\ \mathrm{his}(H)^{\mathrm{X}} * T = A^{\mathrm{X}} * T,\ (\mathrm{his}(H) * T)^{\mathrm{X}} = (A * T)^{\mathrm{X}}\}\ \cup \\
& \{(G,0) : G \in \mathcal{H}_{U,V,z},\ G \not\subseteq H_E\}\ \cup \\
& \{(G,0) : G \in \mathcal{H}_{U,V,z},\ \mathrm{his}(G)^{\mathrm{X}} * T \neq A^{\mathrm{X}} * T\}\ \cup \\
& \{(G,0) : G \in \mathcal{H}_{U,V,z},\ (\mathrm{his}(G) * T)^{\mathrm{X}} \neq (A * T)^{\mathrm{X}}\}
\end{aligned}
$$

That is, the *maximum likelihood estimate*, $\tilde{P}$, is such that all *drawn histories* $H \subseteq H_E$ with both the *formal*, $\mathrm{his}(H)^{\mathrm{X}} * T = A^{\mathrm{X}} * T$ and the *abstract*, $(\mathrm{his}(H) * T)^{\mathrm{X}} = (A * T)^{\mathrm{X}}$, are uniformly probable and all other *histories*, $G \not\subseteq H_E$ or $\mathrm{his}(G)^{\mathrm{X}} * T \neq A^{\mathrm{X}} * T$ or $(\mathrm{his}(G) * T)^{\mathrm{X}} \neq (A * T)^{\mathrm{X}}$, are impossible, $\tilde{P}(G) = 0$. If (i) the *transform*, T, is *known*, (ii) the *formal*, $A^{\mathrm{X}} * T$, is *known*, (iii) the *abstract*, $(A * T)^{\mathrm{X}}$, is *known* and (iv) the *distribution histogram*, H_E, is *known*, then the *likely probability function*, $\tilde{P}$, is *known*.

The *likely probability* of *drawing histogram* A from *necessary drawn formal* $A^{\mathrm{X}} * T$ and *necessary drawn abstract* $(A * T)^{\mathrm{X}}$ is

$$
\sum(\tilde{P}(H) : H \in \mathcal{H}_{U,V,z},\ \mathrm{his}(H) = A) \;=\; \frac{Q_{\mathrm{h},U}(E,z)(A)}{\sum Q_{\mathrm{h},U}(E,z)(B) : B \in Y^{-1}_{U,\mathrm{i},T,z}((A^{\mathrm{X}} * T, (A * T)^{\mathrm{X}}))}
$$

The *likely history probability function* entropy, $\mathrm{entropy}(\tilde{P})$, is maximised when the sum of the *iso-transform-independent historical frequencies*,

$$
\sum Q_{\mathrm{h},U}(E,z)(B) : B \in Y^{-1}_{U,\mathrm{i},T,z}((A^{\mathrm{X}} * T, (A * T)^{\mathrm{X}}))
$$

is maximised.

Consider the case where the *transform*, T, is *known*, the *formal*, $A^{\mathrm{X}} * T$, is *known*, and the *abstract*, $(A * T)^{\mathrm{X}}$, is *known*, but the *distribution histogram*, E, is *unknown* and hence the *likely history probability function*, $\tilde{P}$, is *unknown*. The *maximum likelihood estimate* $\tilde{E}$ for the *distribution histogram*, E, is a modal value of the *likelihood function*,

$$\begin{aligned}\tilde{E} \ \in \ \ & \mathrm{maxd}(\{(D, \sum(Q_{\mathrm{h},U}(D,z)(B) : B \in Y^{-1}_{U,\mathrm{i},T,z}((A^{\mathrm{X}} * T, (A * T)^{\mathrm{X}})))) : \\ & \qquad D \in \mathcal{A}_{U,\mathrm{i},V,z_E}\})\end{aligned}$$

The *likely distribution histogram*, $\tilde{E}$, is *known* if the *distribution histogram size*, z_E, is *known*, the *transform*, T, is *known*, the *formal*, $A^{\mathrm{X}} * T$, is *known*, and the *abstract*, $(A * T)^{\mathrm{X}}$, is *known*. If it is assumed that the *distribution histogram* equals the *likely distribution histogram*, $E = \tilde{E}$, then the *likely history probability* is *known*, $\tilde{P}(H) = 1/\sum(Q_{\mathrm{h},U}(\tilde{E}, z)(B) : B \in Y^{-1}_{U,\mathrm{i},T,z}((A^{\mathrm{X}} * T, (A * T)^{\mathrm{X}})))$ where $\mathrm{his}(H)^{\mathrm{X}} * T = A^{\mathrm{X}} * T$ and $(\mathrm{his}(H) * T)^{\mathrm{X}} = (A * T)^{\mathrm{X}}$.

In the case where the *distribution histogram*, E, is *unknown*, and the *distribution histogram size*, z_E, is also *unknown*, except that it is *known* to be large, $z_E \gg z$, then the *maximum likelihood estimate* $\tilde{E}$ for the *distribution probability histogram*, $\hat{E}$, may be approximated by a modal value of a *likelihood function* which depends on the *multinomial distribution* instead,

$$\begin{aligned}\tilde{E} \ \in \ \ & \mathrm{maxd}(\{(D, \sum(Q_{\mathrm{m},U}(D,z)(B) : B \in Y^{-1}_{U,\mathrm{i},T,z}((A^{\mathrm{X}} * T, (A * T)^{\mathrm{X}})))) : \\ & \qquad D \in \mathcal{A}_{U,V,1}\})\end{aligned}$$

If it is *known*, in addition, that the *formal* equals the *abstract*, $A^{\mathrm{X}} * T = (A*T)^{\mathrm{X}}$, then the *normalised naturalised abstract*, $(\hat{A}*T)^{\mathrm{X}}*T^{\dagger}$, is a solution. In this case the *naturalised abstract*, $(A * T)^{\mathrm{X}} * T^{\dagger}$, or *naturalised formal*, $A^{\mathrm{X}}*T*T^{\dagger} = (A*T)^{\mathrm{X}}*T^{\dagger}$, is the *independent analogue* of the *iso-transform-independents*. So the *maximum likelihood estimate*, $\tilde{E}$, for the *distribution probability histogram*, $\hat{E}$, is the *naturalised abstract probability histogram*, $(\hat{A}* T)^{\mathrm{X}} * T^{\dagger}$,

$$\tilde{E} \ = \ (\hat{A} * T)^{\mathrm{X}} * T^{\dagger}$$

Formal-abstract equivalence, $A^{\mathrm{X}}*T = (A*T)^{\mathrm{X}}$, is also called *mid transform*. In this case the *abstract* equals the *independent abstract*, $(A*T)^{\mathrm{X}} = A^{\mathrm{X}}*T = (A^{\mathrm{X}}*T)^{\mathrm{X}}$, and so does not depend on the *histogram alignment*, $\mathrm{algn}(A)$. The *formal* equals the *formal independent*, $A^{\mathrm{X}} * T = (A * T)^{\mathrm{X}} = (A^{\mathrm{X}} * T)^{\mathrm{X}}$, and so does not depend on its own *alignment*, $\mathrm{algn}(A^{\mathrm{X}} * T) = 0$.

The *naturalised abstract* is the *independent analogue* of the *iso-transform-independents*, so, in the case where the *naturalised abstract* is *integral*, $(A *$

$T)^{\mathrm{X}} * T^{\dagger} \in \mathcal{A}_{\mathrm{i}}$, the sum of the *iso-transform-independent naturalised-abstract-distributed multinomial probabilities* varies with the *naturalised-abstract naturalised abstract-distributed multinomial probability*,

$$\sum Q_{\mathrm{m},U}((A * T)^{\mathrm{X}} * T^{\dagger}, z)(B) : B \in Y^{-1}_{U,\mathrm{i},T,z}((A^{\mathrm{X}} * T, (A * T)^{\mathrm{X}})) \sim Q_{\mathrm{m},U}((A * T)^{\mathrm{X}} * T^{\dagger}, z)((A * T)^{\mathrm{X}} * T^{\dagger})$$

So, if it is assumed that the *distribution probability histogram* equals the *likely distribution probability histogram*, $\hat{E} = \tilde{E} = (\hat{A} * T)^{\mathrm{X}} * T^{\dagger}$, then the *likely history probability* varies against the *naturalised-abstract-distributed multinomial probability* of the *naturalised abstract*, $\tilde{P}(H) \sim 1/\hat{Q}_{\mathrm{m},U}((A * T)^{\mathrm{X}} * T^{\dagger}, z)((A * T)^{\mathrm{X}} * T^{\dagger})$. The *likely history probability function* entropy varies with the *naturalised abstract entropy*, $\mathrm{entropy}(\tilde{P}) \sim \mathrm{entropy}((A * T)^{\mathrm{X}} * T^{\dagger})$.

Given *necessary formal*, *necessary abstract* and *mid transform*, consider the case where a *drawn histogram* A is *known*, but neither the *distribution histogram*, E, is *known* nor the *transform*, T, is *known*, and hence the *likely history probability function*, $\tilde{P}$, is *unknown*. The *maximum likelihood estimate* $(\tilde{E}, \tilde{T})$ for the pair of the *distribution histogram*, E, and the *transform*, T, is a modal value of the *likelihood function*,

$$(\tilde{E}, \tilde{T}) \in \mathrm{maxd}(\{((D, M), \sum(Q_{\mathrm{h},U}(D, z)(B) : B \in Y^{-1}_{U,\mathrm{i},M,z}((A^{\mathrm{X}} * M, (A * M)^{\mathrm{X}})))) : D \in \mathcal{A}_{U,\mathrm{i},V,z_E},\ M \in \mathcal{T}_{U,V},\ A^{\mathrm{X}} * M = (A * M)^{\mathrm{X}}\})$$

In some cases of *drawn sample*, A, the *transform maximum likelihood estimate*, $\tilde{T}$, is not trivial. That is, the *transform maximum likelihood estimate* is not necessarily *unary*, $T_{\mathrm{u}} = \{V^{\mathrm{CS}}\}^{\mathrm{T}}$, nor *self*, $T_{\mathrm{s}} = V^{\mathrm{CS}\{\}\mathrm{T}}$. In the cases where the *transform maximum likelihood estimate* is trivial, $\tilde{T} \in \{T_{\mathrm{u}}, T_{\mathrm{s}}\}$, *aligned modelled induction* reduces to *aligned non-modelled induction*,

$$\begin{aligned}\tilde{P} \;=\; & \{(H,1) : H \subseteq H_E,\ \mathrm{his}(H)^{\mathrm{X}} = A^{\mathrm{X}}\}^{\wedge} \cup \\ & \{(G,0) : G \in \mathcal{H}_{U,V,z},\ G \nsubseteq H_E\} \cup \\ & \{(G,0) : G \in \mathcal{H}_{U,V,z},\ \mathrm{his}(G)^{\mathrm{X}} \neq A^{\mathrm{X}}\}\end{aligned}$$

Define the *transform-dependent* $A^{\mathrm{Y}(T)} \in \mathcal{A}_{U,V,z}$ as the *maximum likelihood estimate* of the *distribution histogram* of the *multinomial probability* of the *histogram*, A, conditional that it is an *iso-transform-independent*,

$$\{A^{\mathrm{Y}(T)}\} = \mathrm{maxd}(\{(D, \frac{Q_{\mathrm{m},U}(D,z)(A)}{\sum Q_{\mathrm{m},U}(D,z)(B) : B \in Y^{-1}_{U,\mathrm{i},T,z}((A^{\mathrm{X}} * T, (A * T)^{\mathrm{X}}))}) : D \in \mathcal{A}_{U,V,z}\})$$

The *transform-dependent*, $A^{\mathrm{Y}(T)}$, is the *dependent analogue* of the *iso transform independents*. Note that the *transform-dependent*, $A^{\mathrm{Y}(T)}$, is not always computable, but an approximation to any accuracy can be made to it. In the case where the *formal* equals the *abstract*, $A^{\mathrm{X}} * T = (A * T)^{\mathrm{X}}$, and the *histogram* equals the *naturalised abstract*, the *transform-dependent* equals the *naturalised abstract*, $A = (A * T)^{\mathrm{X}} * T^{\dagger} \implies A^{\mathrm{Y}(T)} = A = (A * T)^{\mathrm{X}} * T^{\dagger}$.

Now consider the case where, given *necessary formal*, *necessary abstract* and *mid transform*, it is *known*, in addition, that the *sample histogram* A is the most *probable histogram* of the *iso-transform-independents*. That is, the *likely probability* of *drawing histogram* A from *necessary formal-abstract* $(A^{\mathrm{X}} * T, (A * T)^{\mathrm{X}})$,

$$\sum(\tilde{P}(H) : H \in \mathcal{H}_{U,V,z},\ \mathrm{his}(H) = A) = \frac{Q_{\mathrm{h},U}(E,z)(A)}{\sum Q_{\mathrm{h},U}(E,z)(B) : B \in Y^{-1}_{U,\mathrm{i},T,z}((A^{\mathrm{X}} * T, (A * T)^{\mathrm{X}}))}$$

is maximised.

In the case where the *transform*, T, is *known* and the *sample*, A, is *known*, but the *distribution histogram*, E, is *unknown*, the *maximum likelihood estimate* $\tilde{E}$ for the *distribution histogram*, E, is a modal value of the *likelihood function*,

$$\tilde{E} \in \mathrm{maxd}(\{(D, \frac{Q_{\mathrm{h},U}(D,z)(A)}{\sum Q_{\mathrm{h},U}(D,z)(B) : B \in Y^{-1}_{U,\mathrm{i},T,z}((A^{\mathrm{X}} * T, (A * T)^{\mathrm{X}}))}) : D \in \mathcal{A}_{U,\mathrm{i},V,z_E}\})$$

The *likely distribution histogram*, $\tilde{E}$, is *known* if the *distribution histogram size*, z_E, is *known*, the *transform*, T, is *known* and the *sample*, A, is *known*. If it is assumed that the *distribution histogram* equals the *likely distribution histogram*, $E = \tilde{E}$, then the *likely history probability* is *known*, $\tilde{P}(H) = 1/\sum(Q_{\mathrm{h},U}(\tilde{E},z)(B) : B \in Y^{-1}_{U,\mathrm{i},T,z}((A^{\mathrm{X}} * T, (A * T)^{\mathrm{X}})))$ where $\mathrm{his}(H)^{\mathrm{X}} * T = A^{\mathrm{X}} * T$ and $(\mathrm{his}(H) * T)^{\mathrm{X}} = (A * T)^{\mathrm{X}}$.

If the *histogram* is *naturalised abstract*, $A = (A * T)^{\mathrm{X}} * T^{\dagger}$, then the additional constraint of *probable sample* makes no change to the *maximum likelihood estimate*, $\tilde{E}$,

$$\begin{aligned}
&A = (A * T)^{\mathrm{X}} * T^{\dagger} \implies \\
&\quad \mathrm{maxd}(\{(D, \frac{Q_{\mathrm{h},U}(D,z)(A)}{\sum Q_{\mathrm{h},U}(D,z)(B) : B \in Y^{-1}_{U,\mathrm{i},T,z}((A^{\mathrm{X}} * T, (A * T)^{\mathrm{X}}))}) : \\
&\qquad\qquad D \in \mathcal{A}_{U,\mathrm{i},V,z_E}\}) \\
&= \mathrm{maxd}(\{(D, \sum (Q_{\mathrm{h},U}(D,z)(B) : B \in Y^{-1}_{U,\mathrm{i},T,z}((A^{\mathrm{X}} * T, (A * T)^{\mathrm{X}})))) : \\
&\qquad\qquad D \in \mathcal{A}_{U,\mathrm{i},V,z_E}\})
\end{aligned}$$

If the *histogram* is not *naturalised abstract*, $A \neq (A * T)^{\mathrm{X}} * T^{\dagger}$, however, then the *likely history probability function* entropy, $\mathrm{entropy}(\tilde{P})$, is lower than it is in the case of *necessary formal-abstract* unconstrained by *probable sample*.

In the case where the *distribution histogram*, E, is *unknown*, and the *distribution histogram size*, z_E, is also *unknown*, except that it is *known* to be large, $z_E \gg z$, then the *maximum likelihood estimate* $\tilde{E}$ for the *distribution probability histogram*, $\hat{E}$, is now approximated by a modal value of the conditional *likelihood function*,

$$\begin{aligned}
\tilde{E} \;\in\; &\mathrm{maxd}(\{(D, \frac{Q_{\mathrm{m},U}(D,z)(A)}{\sum Q_{\mathrm{m},U}(D,z)(B) : B \in Y^{-1}_{U,\mathrm{i},T,z}((A^{\mathrm{X}} * T, (A * T)^{\mathrm{X}}))}) : \\
&\qquad\qquad D \in \mathcal{A}_{U,V,1}\})
\end{aligned}$$

The solution to this is the *normalised transform-dependent*, $\tilde{E} = \hat{A}^{\mathrm{Y}(T)}$. The *maximum likelihood estimate* is near the *sample*, $\tilde{E} \sim \hat{A}$, only in as much as it is far from the *naturalised abstract*, $\tilde{E} \nsim (\hat{A} * T)^{\mathrm{X}} * T^{\dagger}$.

The *iso-transform-independent conditional multinomial probability distribution* is defined

$$\begin{aligned}
&\hat{Q}_{\mathrm{m,y},T,U}(E,z)(A) := \\
&\quad \frac{1}{|\mathrm{ran}(Y_{U,\mathrm{i},T,z})|} \frac{Q_{\mathrm{m},U}(E,z)(A)}{\sum Q_{\mathrm{m},U}(E,z)(B) : B \in Y^{-1}_{U,\mathrm{i},T,z}((A^{\mathrm{X}} * T, (A * T)^{\mathrm{X}}))}
\end{aligned}$$

So the optimisation can be rewritten,

$$\tilde{E} \;\in\; \mathrm{maxd}(\{(D, \hat{Q}_{\mathrm{m,y},T,U}(D,z)(A)) : D \in \mathcal{A}_{U,V,1}\})$$

Consider the case where the *distribution* equals the *transform-dependent*, $\hat{E} = \hat{A}^{\mathrm{Y}(T)}$. First, the logarithm of the *iso-transform-independent conditional multinomial probability* of the *histogram*, A, with respect to the *dependent analogue* or *transform-dependent*, $A^{\mathrm{Y}(T)}$, varies against the logarithm of the *iso-transform-independent conditional multinomial probability* with respect to the *independent analogue* or *naturalised abstract*, $(A * T)^{\mathrm{X}} * T^{\dagger}$,

$$\begin{aligned}
&\ln \frac{Q_{\mathrm{m},U}(A^{\mathrm{Y}(T)}, z)(A)}{\sum Q_{\mathrm{m},U}(A^{\mathrm{Y}(T)}, z)(B) : B \in Y^{-1}_{U,\mathrm{i},T,z}((A^{\mathrm{X}} * T, (A * T)^{\mathrm{X}}))} \\
\sim\quad &-\ln \frac{Q_{\mathrm{m},U}((A * T)^{\mathrm{X}} * T^{\dagger}, z)(A)}{\sum Q_{\mathrm{m},U}((A * T)^{\mathrm{X}} * T^{\dagger}, z)(B) : B \in Y^{-1}_{U,\mathrm{i},T,z}((A^{\mathrm{X}} * T, (A * T)^{\mathrm{X}}))}
\end{aligned}$$

Second, the negative logarithm of the *iso-transform-independent conditional multinomial probability* of the *histogram*, A, with respect to the *naturalised abstract*, $(A * T)^{\mathrm{X}} * T^{\dagger}$, varies with the negative logarithm of the *lifted iso-transform-independent conditional multinomial probability* of the *derived*, $A * T$, with respect to the *abstract*, $(A * T)^{\mathrm{X}}$,

$$\begin{aligned}
&-\ln \frac{Q_{\mathrm{m},U}((A * T)^{\mathrm{X}} * T^{\dagger}, z)(A)}{\sum Q_{\mathrm{m},U}((A * T)^{\mathrm{X}} * T^{\dagger}, z)(B) : B \in Y^{-1}_{U,\mathrm{i},T,z}((A^{\mathrm{X}} * T, (A * T)^{\mathrm{X}}))} \\
\sim\quad &-\ln \frac{Q_{\mathrm{m},U}((A * T)^{\mathrm{X}}, z)(A * T)}{\sum Q_{\mathrm{m},U}((A * T)^{\mathrm{X}}, z)(B') : B' \in Y'^{-1}_{U,\mathrm{i},T,z}((A^{\mathrm{X}} * T, (A * T)^{\mathrm{X}}))}
\end{aligned}$$

where $Y'^{-1}_{U,\mathrm{i},T,z}((A^{\mathrm{X}} * T, (A * T)^{\mathrm{X}})) = \{B * T : B \in Y^{-1}_{U,\mathrm{i},T,z}((A^{\mathrm{X}} * T, (A * T)^{\mathrm{X}}))\}$.

Third, the negative logarithm of the *lifted iso-transform-independent conditional multinomial probability* with respect to the *abstract*, $(A * T)^{\mathrm{X}}$, varies with the negative logarithm of the *relative multinomial probability* with respect to the *abstract*, $(A * T)^{\mathrm{X}}$, which is the *derived alignment*,

$$\begin{aligned}
&-\ln \frac{Q_{\mathrm{m},U}((A * T)^{\mathrm{X}}, z)(A * T)}{\sum Q_{\mathrm{m},U}((A * T)^{\mathrm{X}}, z)(B') : B' \in Y'^{-1}_{U,\mathrm{i},T,z}((A^{\mathrm{X}} * T, (A * T)^{\mathrm{X}}))} \\
\sim\quad &-\ln \frac{Q_{\mathrm{m},U}((A * T)^{\mathrm{X}}, z)(A * T)}{Q_{\mathrm{m},U}((A * T)^{\mathrm{X}}, z)((A * T)^{\mathrm{X}})} \\
=\quad &\mathrm{algn}(A * T)
\end{aligned}$$

So the *log-likelihood* varies with the *derived alignment*,

$$\ln \hat{Q}_{\mathrm{m},\mathrm{y},T,U}(A^{\mathrm{Y}(T)}, z)(A) \quad \sim \quad \mathrm{algn}(A * T)$$

The *mid transform* constraint allows the *log-likelihood*, which is a function of the *histogram*, A, to be *lifted* to the *derived alignment*, which is a function of the *derived*, $A * T$. So a *model* optimisation need only search in the *derived volume*, $|T^{-1}|$, which is typically much smaller than the *underlying volume*, $|T^{-1}| \ll |V^{\mathrm{C}}|$. It is this relation between the *log-likelihood* and the *derived alignment* that makes *aligned induction* practicable.

The case of *classical modelled induction*, where the *derived* is *necessary*, may be termed *law-like* because the set of *iso-derived*, $D^{-1}_{U,\mathrm{i},T,z}(A * T)$, is *law-like*. All *drawn histories* $H \subseteq H_E$, are such that their *derived histograms* are fixed, $\mathrm{his}(H) * T = A * T$.

By contrast, the case of *aligned modelled induction*, where the *abstract* is *necessary*, may be termed *entity-like* because the set of *iso-abstracts*, $Y^{-1}_{U,\mathrm{i},T,\mathrm{W},z}((A*T)^{\mathrm{X}})$, is *entity-like*. All *drawn histories* are such that their *abstract histograms* are fixed, $(\mathrm{his}(H) * T)^{\mathrm{X}} = (A * T)^{\mathrm{X}}$. That is, the *derived variables* are separately *necessary*, $\forall u \in W\ (\mathrm{his}(H) * T\ \%\ \{u\} = A * T\ \%\ \{u\})$. *Necessary abstract* is a weaker constraint than *necessary derived* because the *iso-abstracts* are a superset of the *iso-derived*, $D^{-1}_{U,\mathrm{i},T,z}(A * T) \subseteq Y^{-1}_{U,\mathrm{i},T,\mathrm{W},z}((A * T)^{\mathrm{X}})$. In fact, *aligned induction* is stricter than pure *entity-like* because the *formal* is *necessary* too, $\mathrm{his}(H)^{\mathrm{X}} * T = A^{\mathrm{X}} * T$, and so *aligned induction* is also *aligned-like*, $Y^{-1}_{U,\mathrm{i},V,z}(A^{\mathrm{X}}) \subseteq Y^{-1}_{U,\mathrm{i},T,\mathrm{V},z}(A^{\mathrm{X}} * T)$. *Aligned induction*, however, is not necessarily *law-like*, $\mathrm{his}(H) * T \neq A * T$, and so does not always approximate to *classical induction*. *Mid transform* is stricter still, but this constraint does not necessarily increase *law-likeness*, but merely allows *lifting*.

Consider the case where, given *necessary formal*, *necessary abstract*, *mid transform* and *probable sample*, it is *known*, in addition, that the *sample histogram* is *ideal*, $A = A * T * T^{\dagger A}$. The *idealisation independent* equals the *independent*, $(A*T*T^{\dagger A})^{\mathrm{X}} = A^{\mathrm{X}}$, so the *idealisation* is *aligned-like*. The *ideal sample* approximates to the *independent analogue* of the *iso-derived*, which is the *naturalisation*, $A \approx A * T * T^{\dagger}$, and so, if it is also the case that *derived alignment* is high, $\mathrm{algn}(A * T) \gg 0$, the *iso-transform-independent conditional multinomial log-likelihood* with respect to the *dependent analogue* or *transform-dependent*, $A^{\mathrm{Y}(T)}$, varies with the *iso-derived conditional multinomial log-likelihood* with respect to the *independent analogue* or *naturalisation*, $A * T * T^{\dagger}$,

$$\begin{aligned}\ln \hat{Q}_{\mathrm{m,y},T,U}(A^{\mathrm{Y}(T)}, z)(A) &\sim \ln \hat{Q}_{\mathrm{m,d},T,U}(A * T * T^{\dagger}, z)(A) \\ &\sim \ln \hat{Q}_{\mathrm{m,d},T,U}(A, z)(A)\end{aligned}$$

So the *log likelihood* varies with the *size-volume* scaled *component size cardinality sum relative entropy*,

$$\begin{aligned}\ln \hat{Q}_{\mathrm{m,y},T,U}(A^{\mathrm{Y}(T)}, z)(A) \ &\sim \\ &(z+v) \times \mathrm{entropy}(A * T + V^{\mathrm{C}} * T) \\ &\qquad -z \times \mathrm{entropy}(A * T) \ - v \times \mathrm{entropy}(V^{\mathrm{C}} * T)\end{aligned}$$

and the *maximum likelihood estimate derived* approximates to the *normalised sample derived*,

$$\begin{aligned}\tilde{E} * T &= \hat{A}^{\mathrm{Y}(T)} * T \\ &\approx \hat{A} * T\end{aligned}$$

In the case where the *underlying alignment* is intermediate, $\mathrm{algn}(A) \gg 0$, and the *component size cardinality relative entropy* is high, $\mathrm{entropyCross}(A * T, V^{\mathrm{C}} * T) > \ln |T^{-1}|$, the *sum sensitivity* varies against the *log likelihood*,

$$\mathrm{sum}(\mathrm{sensitivity}(U)(\hat{Q}_{\mathrm{m,y},T,U}(A^{\mathrm{Y}(T)}, z))) \ \sim \ - \ \ln \hat{Q}_{\mathrm{m,y},T,U}(A^{\mathrm{Y}(T)}, z)(A)$$

and the *model sensitivity* varies against the *log likelihood*,

$$\begin{aligned}- \ \ln |\max(\{(M, \hat{Q}_{\mathrm{m,y},M,U}(A^{\mathrm{Y}(M)}, z)(A)) : M \in \mathcal{T}_{U,V}, \\ A^{\mathrm{X}} * M = (A * M)^{\mathrm{X}}, \ A = A * M * M^{\dagger A}\})| \\ \sim \ - \ \ln \hat{Q}_{\mathrm{m,y},T,U}(A^{\mathrm{Y}(T)}, z)(A)\end{aligned}$$

That is, given *mid-ideal transform*, the maximisation of the *derived alignment* tends to make the properties of *aligned modelled induction* similar to those of *classical modelled induction*.

Given *necessary formal-abstract*, *mid-ideal transform* and *probable sample*, consider the case where a *drawn histogram* A is *known*, but neither the *distribution histogram*, E, is *known* nor the *transform*, T, is *known*, and hence the *likely history probability function*, $\tilde{P}$, is *unknown*. In the case where the *distribution histogram size*, z_E, is also *unknown*, except that it is *known* to be large, $z_E \gg z$, then the *maximum likelihood estimate* $(\tilde{E}, \tilde{T})$ for the pair of the *distribution histogram*, E, and the *transform*, T, is approximated by a modal value of the conditional *likelihood function*,

$$\begin{aligned}(\tilde{E}, \tilde{T}) \ &\in \\ &\mathrm{maxd}(\{((D, M), \frac{Q_{\mathrm{m},U}(D, z)(A)}{\sum Q_{\mathrm{m},U}(D, z)(B) : B \in Y^{-1}_{U,\mathrm{i},M,z}((A^{\mathrm{X}} * M, (A * M)^{\mathrm{X}}))}) : \\ &\qquad D \in \mathcal{A}_{U,V,1}, \ M \in \mathcal{T}_{U,V}, \ A^{\mathrm{X}} * M = (A * M)^{\mathrm{X}}, \ A = A * M * M^{\dagger A}\})\end{aligned}$$

So the *likely distribution* equals the *likely transform-dependent*, $\tilde{E} = \hat{A}^{\mathrm{Y}(\tilde{T})}$, and the *likely model* is such that

$$\begin{aligned}\tilde{T} \in &\\ &\mathrm{maxd}(\{(M, \frac{Q_{\mathrm{m},U}(A^{\mathrm{Y}(M)}, z)(A)}{\sum Q_{\mathrm{m},U}(A^{\mathrm{Y}(M)}, z)(B) : B \in Y^{-1}_{U,\mathrm{i},M,z}((A^{\mathrm{X}} * M, (A * M)^{\mathrm{X}}))}) :\\ &\qquad M \in \mathcal{T}_{U,V},\ A^{\mathrm{X}} * M = (A * M)^{\mathrm{X}},\ A = A * M * M^{\dagger A}\})\end{aligned}$$

The *log-likelihood* varies with the *derived alignment*, so an approximation to the *likely model* is

$$\begin{aligned}\tilde{T} \ \in \ &\mathrm{maxd}(\{(M, \mathrm{algn}(A * M)) :\\ &\qquad M \in \mathcal{T}_{U,V},\ A^{\mathrm{X}} * M = (A * M)^{\mathrm{X}},\ A = A * M * M^{\dagger A}\})\end{aligned}$$

This optimisation is still intractable, because the cardinality of the *substrate transforms*, $|\mathcal{T}_{U,V}|$, is factorial in the *volume*, v. The computation of the *derived alignment*, $\mathrm{algn}(A*M)$, is tractable, however, and so limited searches can be made tractable and then practicable.

In *classical modelled induction* the constraint must be weakened from *necessary derived* to *uniform possible derived* if the *likely model* is to be non-trivial, $\tilde{T} \notin \{T_{\mathrm{u}}, T_{\mathrm{s}}\}$. *Uniform possible* is not required for *aligned modelled induction* because the *likely model* is sometimes non-trivial when constrained by *necessary formal-abstract*, which is already weaker than *necessary derived*.

Consider, however, the case where the *formal-abstract* pair is *uniformly possible*. Given *substrate transform* $T \in \mathcal{T}_{U,V}$, assume that the *substrate history probability function* $P \in (\mathcal{H}_{U,V,z} :\to \mathbf{Q}_{\geq 0}) \cap \mathcal{P}$ is the distribution of an arbitrary *history* valued function, $\mathcal{X} \to \mathcal{H}$, given an arbitrary *formal-abstract* valued function, $\mathcal{X} \to \mathcal{A}^2$. In this case, the *history* valued function is chosen arbitrarily from the constrained subset

$$\begin{aligned}&\{G : F \in \mathcal{X} \to (\mathcal{A}^2 \times (\mathcal{X} \to \mathcal{H})),\ (\cdot, ((A', B'), G)) \in F,\\ &\qquad \forall (\cdot, H) \in G\ (\mathrm{his}(H)^{\mathrm{X}} * T = A' \wedge (\mathrm{his}(H) * T)^{\mathrm{X}} = B')\} \ \subset \ \mathcal{X} \to \mathcal{H}\end{aligned}$$

In the case of *mid transform*, $A^{\mathrm{X}} * T = (A * T)^{\mathrm{X}}$, the constrained subset is simpler,

$$\begin{aligned}&\{G : F \in \mathcal{X} \to (\mathcal{A} \times (\mathcal{X} \to \mathcal{H})),\ (\cdot, (A', G)) \in F,\\ &\qquad \forall (\cdot, H) \in G\ (\mathrm{his}(H)^{\mathrm{X}} * T = (\mathrm{his}(H) * T)^{\mathrm{X}} = A')\} \ \subset \ \mathcal{X} \to \mathcal{H}\end{aligned}$$

This subset of the *substrate history probability functions* can be generalised for all *substrate transforms* as the subset derived from

$$\bigcup_{T \in \mathcal{T}_{\mathrm{f}}} (\mathcal{X} \to (\mathcal{A} \times_T (\mathcal{X} \to \mathcal{H})))$$

where $\mathcal{T}_{\mathrm{f}}$ is the set of all *functional transforms*, and the fibre product $\times_T$ is defined

$$\begin{aligned} \mathcal{A} \times_T (\mathcal{X} \to \mathcal{H}) \ :=& \\ & \{(A', G) : (A', G) \in \mathcal{A} \times (\mathcal{X} \to \mathcal{H}), \\ & \quad \forall (\cdot, H) \in G \ (\mathrm{his}(H)^{\mathrm{X}} * T = (\mathrm{his}(H) * T)^{\mathrm{X}} = A')\} \end{aligned}$$

In the case of *uniform possible formal-abstract*, where there is a *distribution history* H_E and a *substrate transform* $T \in \mathcal{T}_{U,V}$, the *maximum likelihood estimate* which maximises the entropy, $\mathrm{entropy}(\tilde{P})$, is

$$\begin{aligned} \tilde{P} \ =& \ \Big(\bigcup \{\{(H, 1) : H \subseteq H_E, \ \mathrm{his}(H)^{\mathrm{X}} * T = A', \ (\mathrm{his}(H) * T)^{\mathrm{X}} = B'\}^{\wedge} \ : \\ & \qquad (A', B') \in \mathrm{ran}(Y_{U,\mathrm{i},T,z})\}\Big)^{\wedge} \ \cup \\ & \{(G, 0) : G \in \mathcal{H}_{U,V,z}, \ G \not\subseteq H_E\} \\ =& \ \Big(\bigcup \{\{(H, 1/\sum(Q_{\mathrm{h},U}(E, z)(B) : B \in Y^{-1}_{U,\mathrm{i},T,z}((A', B'))) : \\ & \qquad H \subseteq H_E, \ \mathrm{his}(H)^{\mathrm{X}} * T = A', \ (\mathrm{his}(H) * T)^{\mathrm{X}} = B'\} \ : \\ & \qquad (A', B') \in \mathrm{ran}(Y_{U,\mathrm{i},T,z})\}\Big)^{\wedge} \ \cup \\ & \{(G, 0) : G \in \mathcal{H}_{U,V,z}, \ G \not\subseteq H_E\} \end{aligned}$$

That is, the *maximum likelihood estimate*, $\tilde{P}$, is such that all *drawn formal-abstracts* are uniformly probable, and then all *drawn histories* $H \subseteq H_E$ with the same *formal-abstract*, $\mathrm{his}(H)^{\mathrm{X}} * T = A'$ and $(\mathrm{his}(H) * T)^{\mathrm{X}} = B'$, are uniformly probable. If the *distribution histogram*, H_E, is *known* and the *substrate transform*, T, is *known*, then the *likely probability function*, $\tilde{P}$, is *known*.

The properties of *uniformly possible formal-abstract* are the same as for *necessary formal-abstract*, except that the probabilities are scaled by the fraction $1/|\mathrm{ran}(Y_{U,\mathrm{i},T,z})|$.

Given *uniform possible formal-abstract*, *mid-ideal transform* and *probable sample*, consider the case where a *drawn histogram* A is *known*, but neither the *distribution histogram*, E, is *known* nor the *transform*, T, is *known*, and hence the *likely history probability function*, $\tilde{P}$, is *unknown.* In the case where the *distribution histogram size*, z_E, is also *unknown*, except that it is *known* to be large, $z_E \gg z$, then the *maximum likelihood estimate* $(\tilde{E},\tilde{T})$ for the pair of the *distribution histogram*, E, and the *transform*, T, is approximated by a modal value of the conditional *likelihood function*,

$$\begin{aligned}
&(\tilde{E},\tilde{T}) \in \\
&\quad \mathrm{maxd}(\{((D,M),\hat{Q}_{\mathrm{m,y},M,U}(D,z)(A)) : \\
&\qquad D \in \mathcal{A}_{U,V,1},\ M \in \mathcal{T}_{U,V},\ A^{\mathrm{X}} * M = (A * M)^{\mathrm{X}},\ A = A * M * M^{\dagger A}\})
\end{aligned}$$

So the *likely distribution* equals the *likely transform-dependent*, $\tilde{E} = \hat{A}^{\mathrm{Y}(\tilde{T})}$, and the *likely model* is such that

$$\begin{aligned}
\tilde{T} \in\ & \mathrm{maxd}(\{(M,\hat{Q}_{\mathrm{m,y},M,U}(A^{\mathrm{Y}(M)},z)(A)) : \\
& \qquad M \in \mathcal{T}_{U,V},\ A^{\mathrm{X}} * M = (A * M)^{\mathrm{X}},\ A = A * M * M^{\dagger A}\})
\end{aligned}$$

The *log-likelihood* varies with the *derived alignment*, so an approximation to the *likely model* is

$$\begin{aligned}
\tilde{T} \in\ & \mathrm{maxd}(\{(M,\mathrm{algn}(A * M)) : \\
& \qquad M \in \mathcal{T}_{U,V},\ A^{\mathrm{X}} * M = (A * M)^{\mathrm{X}},\ A = A * M * M^{\dagger A}\})
\end{aligned}$$

Note, however, that this approximation is looser than in the *necessary formal-abstract* case because the scaling fraction, $1/|\mathrm{ran}(Y_{U,\mathrm{i},\tilde{T},z})|$, is ignored.

2.6.5 Tractable and practicable aligned induction

In the discussion of *aligned induction* above it is shown that, given *necessary formal-abstract*, *mid-ideal transform* and *probable sample*, the *maximum likelihood estimate* $\tilde{T}$ for the *transform*, T, is approximated by a maximisation of the *derived alignment*,

$$\begin{aligned}
\tilde{T} \in\ & \mathrm{maxd}(\{(M,\mathrm{algn}(A * M)) : \\
& \qquad M \in \mathcal{T}_{U,V},\ A^{\mathrm{X}} * M = (A * M)^{\mathrm{X}},\ A = A * M * M^{\dagger A}\})
\end{aligned}$$

This optimisation is intractable because the cardinality of the *substrate transforms*, $|\mathcal{T}_{U,V}|$, is factorial in the *volume*, v. Consider how limited searches can be made tractable and then practicable.

Given *sample histogram* $A \in \mathcal{A}_{U,\mathrm{i},V,z}$, the tractable *limited-models summed alignment valency-density substrate aligned non-overlapping infinite-layer fud* *decomposition inducer* is defined

$$\begin{aligned}
&I'^{*}_{z,\mathrm{Sd},\mathrm{D},\mathrm{F},\infty,\mathrm{n},\mathrm{q}}(A) = \\
&\quad \{(M, I^{*}_{\approx\mathbf{R}}(\sum \mathrm{algn}(A * C * F^{\mathrm{T}})/w_F^{1/m_F} : (C,F) \in \mathrm{cont}(M))) : \\
&\qquad\qquad M \in \mathcal{D}_{\mathrm{F},\infty,U,V} \cap \mathrm{trees}(\mathcal{S} \times (\mathcal{F}_{\mathrm{n}} \cap \mathcal{F}_{\mathrm{q}})), \\
&\qquad\qquad\qquad \forall (C,F) \in \mathrm{cont}(M)\ (\mathrm{algn}(A * C * F^{\mathrm{T}}) > 0)\}
\end{aligned}$$

where *derived variables* $W_F = \mathrm{der}(F)$, *derived volume* $w_F = |W_F^{\mathrm{C}}|$, *derived dimension* $m_F = |W_F|$ and $I^{*}_{\approx\mathbf{R}}$ computes an approximation to a real number. The geometric average of the *fud derived valencies* is w_F^{1/m_F}.

Here the *model* has been extended from *transforms*, $M \in \mathcal{T}_{U,V}$, to *functional definition set decompositions*, $M \in \mathcal{D}_{\mathrm{F},\infty,U,V}$. At the same time the set of *fud decompositions* has been restricted to those having (a) *fuds* that are *non-overlapping*, $\mathcal{F}_{\mathrm{n}}$, (b) *fuds* with a *limited-underlying*, *limited-derived*, *limited-layer* and *limited-breadth* structure, $\mathcal{F}_{\mathrm{q}} = \mathcal{F}_{\mathrm{u}} \cap \mathcal{F}_{\mathrm{d}} \cap \mathcal{F}_{\mathrm{h}} \cap \mathcal{F}_{\mathrm{b}}$, and (c) *fuds* with *derived alignment*, $\mathrm{algn}(A * C * F^{\mathrm{T}}) > 0$. The tractable optimal *model* is

$$D_{\mathrm{Sd}} \ \in \ \mathrm{maxd}(I'^{*}_{z,\mathrm{Sd},\mathrm{D},\mathrm{F},\infty,\mathrm{n},\mathrm{q}}(A))$$

The maximisation of the *contingent fud derived alignment valency-density*, $\mathrm{algn}(A*C*F^{\mathrm{T}})/w_F^{1/m_F}$, of the *non-overlapping fud* $(C,F) \in \mathrm{cont}(D_{\mathrm{Sd}})$ for the *sample slice* $A*C$, tends to *mid fud transform*, $(A*C)^{\mathrm{X}}*F^{\mathrm{T}} \approx (A*C*F^{\mathrm{T}})^{\mathrm{X}}$. Then the maximisation of the *summed alignment valency-density*, $\sum \mathrm{algn}(A * C * F^{\mathrm{T}})/w_F^{1/m_F} : (C,F) \in \mathrm{cont}(D_{\mathrm{Sd}})$, for all of the *contingent slices*, tends to *mid-ideal fud decomposition transform*, $A \approx A * D_{\mathrm{Sd}}^{\mathrm{T}} * D_{\mathrm{Sd}}^{\mathrm{T}\dagger A}$. The *summed alignment valency-density* varies with the *derived alignment*, $\mathrm{algn}(A * D_{\mathrm{Sd}}^{\mathrm{T}})$, so the tractable *model* approximates to the *likely model*, $D_{\mathrm{Sd}}^{\mathrm{T}} \approx \tilde{T}$, depending on the limits chosen.

The *derived alignment accuracy* of the approximation can be defined as the exponential of the difference in *derived alignments*,

$$0 \ < \ \frac{\exp(\mathrm{algn}(A * D_{\mathrm{Sd}}^{\mathrm{T}}))}{\exp(\mathrm{algn}(A * \tilde{T}))} \ \leq \ 1$$

This definition of *accuracy* is consistent with the gradient of the likelihood function at the mode, so the *derived alignment accuracy* varies against the *sensitivity* to *model*,

$$\frac{\exp(\text{algn}(A * D_{\text{Sd}}^{\text{T}}))}{\exp(\text{algn}(A * \tilde{T}))} \sim -(-\ln|\max(\{(M, \text{algn}(A * M)) : M \in \mathcal{T}_{U,V}, A^{\text{X}} * M = (A * M)^{\text{X}},\ A = A * M * M^{\dagger A}\})|)$$

The *log-likelihood* varies against the *sensitivity* to *model*, so the *derived alignment accuracy* varies with the *derived alignment*,

$$\frac{\exp(\text{algn}(A * D_{\text{Sd}}^{\text{T}}))}{\exp(\text{algn}(A * \tilde{T}))} \sim \text{algn}(A * T)$$

That is, although the *model* obtained from the tractable *summed alignment valency-density inducer* is merely an approximation, in the cases where the *log-likelihood* or *derived alignment* is high, and so the *sensitivity* to *model/distribution* is low, the approximation may be reasonably close nonetheless.

The maximisation of *derived alignment* tends to make the properties of *mid-ideal aligned induction* similar to those of *natural classical induction*. This is also the case for the tractable optimisation, so the tractable *model* approximates to the *likely classical model*, $D_{\text{Sd}}^{\text{T}} \approx \tilde{T}$, where

$$\tilde{T} \in \text{maxd}(\{(M, \hat{Q}_{\text{m,d},M,U}(A, z)(A)) : M \in \mathcal{T}_{U,V},\ A = A * M * M^{\dagger}\})$$

That this is true may be seen by considering the *entropy* properties. The correlations for *summed alignment valency-density* are similar to those for *iso-derived log-likelihood*. The *summed alignment valency-density* (a) varies against the *derived volume* $w' = |(D_{\text{Sd}}^{\text{T}})^{-1}|$,

$$\text{algnValDensSum}(U)(A, D_{\text{Sd}}) \sim 1/w'$$

(b) varies against the *derived entropy*,

$$\text{algnValDensSum}(U)(A, D_{\text{Sd}}) \sim -z \times \text{entropy}(A * D_{\text{Sd}}^{\text{T}})$$

(c) varies with the *component size cardinality relative entropy*,

$$\text{algnValDensSum}(U)(A, D_{\text{Sd}}) \sim z \times \text{entropyRelative}(A * D_{\text{Sd}}^{\text{T}}, V^{\text{C}} * D_{\text{Sd}}^{\text{T}})$$

and (d) varies with the *expected component entropy*,

$$\text{algnValDensSum}(U)(A, D_{\text{Sd}}) \sim z \times \text{entropyComponent}(A, D_{\text{Sd}}^{\text{T}})$$

where

$$\begin{aligned}\text{algnValDensSum}(U)(A,D) \;&:=\\ &\sum \text{algn}(A * C * F^{\mathrm{T}})/w_F^{1/m_F} : (C,F) \in \text{cont}(D)\end{aligned}$$

The maximisation of the *derived alignment valency-density*, $\text{algn}(A * C * F^{\mathrm{T}})/w_F^{1/m_F}$, of the *contingent fud* $(C,F) \in \text{cont}(D_{\mathrm{Sd}})$, tends to *diagonalise* the *mid fud transform*, $\text{diagonal}(A * C * F^{\mathrm{T}})$, so minimising the *fud derived entropy*, $\text{entropy}(A * C * F^{\mathrm{T}})$, and hence minimising the overall *decomposition transform derived entropy*, $\text{entropy}(A * D_{\mathrm{Sd}}^{\mathrm{T}})$. The *component cardinality entropy*, $\text{entropy}(C * F^{\mathrm{T}})$, also decreases but is synchronised with the *derived entropy*, $\text{entropy}(A * C * F^{\mathrm{T}})$, so the *mid component size cardinality relative entropy* tends to remain small, $\text{entropyRelative}(A * C * F^{\mathrm{T}}, C * F^{\mathrm{T}}) \approx 0$. The maximisation of the *valency-density*, however, shortens the *diagonal* and so the *off-diagonal derived states* tend to be *ineffective*. The recursive *slicing* during the *decomposition* then removes the *ineffective components*, concentrating the *effective derived states* in smaller *components*, and so maximising the overall *decomposition transform component size cardinality relative entropy*, $\text{entropyRelative}(A * D_{\mathrm{Sd}}^{\mathrm{T}}, V^{\mathrm{C}} * D_{\mathrm{Sd}}^{\mathrm{T}})$, when fully *idealised*.

The *limited-models summed alignment valency-density substrate aligned non-overlapping infinite-layer fud decomposition inducer*, $I'_{z,\mathrm{Sd,D,F},\infty,\mathrm{n,q}}$, limits the optimisation to make *aligned induction* tractable. By additionally imposing a sequence on the search and other constraints, tractable *induction* is made practicable in the *highest-layer summed shuffle content alignment valency-density fud decomposition inducer*, $I'_{z,\mathrm{Scsd,D,F},\infty,\mathrm{q},P,\mathrm{d}}$. (The details of the implementation are defined later.) Now, given a set of search parameters P, the *fud decomposition* is

$$D_{\mathrm{Scsd},P} \in \text{maxd}(I'^{*}_{z,\mathrm{Scsd,D,F},\infty,\mathrm{q},P,\mathrm{d}}(A))$$

The set of practicable searched *models* is approximately a subset of the tractable searched *models*, so the practicable *derived alignment* is less than or equal to the tractable *derived alignment*,

$$\text{algn}(A * D_{\mathrm{Scsd},P}^{\mathrm{T}}) \;\;\leq\;\; \text{algn}(A * D_{\mathrm{Sd}}^{\mathrm{T}})$$

Even so, in the cases where the *log-likelihood* or *derived alignment* is high, and so both the *sensitivity* to *model* and the *sensitivity* to *distribution* are low, the approximation to the *maximum likelihood estimate*, $D_{\mathrm{Scsd},P}^{\mathrm{T}} \approx \tilde{T}$, may be reasonably close nonetheless.

The *highest-layer summed shuffle content alignment valency-density fud decomposition inducer*, $I'_{z,\mathrm{Scsd,D,F},\infty,\mathrm{q},P,\mathrm{d}}$, is an example of practicable *aligned induction*. *Artificial neural network induction* is an example of practicable *classical induction*. Let the *ANN classical model* $F^{\mathrm{T}}_{\mathrm{gr,lsq},P} \approx \tilde{T}$ be obtained by *least squares gradient descent* given a *sample* A subject to the constraints of (i) real *valued variables*, (ii) *causal histogram*, (iii) a *literal frame*, and (iv) clustered *histogram*. The *ANN classical induction* is supervised, requiring that there is a *causal* relation between query *variables* $K \subset V$ and label *variables*, $V \setminus K$,

$$\mathrm{split}(K, A^{\mathrm{FS}}) \ \in \ K^{\mathrm{CS}} \to (V \setminus K)^{\mathrm{CS}}$$

At the optimum there is no error and the relation between the *classical derived variables* and the label *variables* is functional,

$$\mathrm{split}(W, (A * X \ \% \ (W \cup V \setminus K))^{\mathrm{FS}}) \ \in \ W^{\mathrm{CS}} \to (V \setminus K)^{\mathrm{CS}}$$

where $(X, W) = F^{\mathrm{T}}_{\mathrm{gr,lsq},P}$.

By contrast, *aligned induction* is unsupervised, so no label is required. *Aligned induction*, however, must have *alignments* between the *underlying variables*,

$$\mathrm{algn}(A) \ > \ 0$$

If there is a label, the *aligned induction model* does not necessarily have a *causal* relation between the *derived variables* and the label *variables*, so the label *entropy* may be non-zero,

$$\sum_{(R,C) \in T^{-1}} (A * T)_R \times \mathrm{entropy}(A * C \ \% \ (V \setminus K)) \ \geq \ 0$$

or

$$\sum_{(R,\cdot) \in T^{-1}} (A * T)_R \times \mathrm{entropy}(\{R\}^{\mathrm{U}} * T^{\odot A} \ \% \ (V \setminus K)) \ \geq \ 0$$

where $T = D^{\mathrm{T}}_{\mathrm{Scsd},P}$.

The *ANN classical induction* also requires that the *sample*, A, is clustered. This implies that the query *variables*, K, are real-*valued*, so that there is a metric. The practicable *aligned inducer* requires that the *underlying variables* be discrete, so they must be bucketed if they are in fact continuous.

The *ANN fud*, $F_{\mathrm{gr,lsq},P}$, has a fixed graph so that the *derived variables* have a *literal frame* mapping to the label *variables* in the loss function. This graph is defined a priori in the parameter set, P, and depends on the query *variables*, K, and the label *variables*, $V \setminus K$. The *aligned inducer model*, $D_{\mathrm{Scsd},P}$, is a *fud decomposition* in which the *fuds* are built upwards from the *substrate*, and the only parameters are limits to gross *fud* structure. In addition, a *decomposition* allows *fuds* to be built on *contingent slices*, $A*C$ where $(C,F) \in \mathrm{cont}(D_{\mathrm{Scsd},P})$, which depend on the *components* corresponding to *effective derived states* of ancestor *fuds*. In this way, the *derived variables* near the root of the *decomposition* are most general, applying to the largest *slices*, while the *derived variables* near the leaves of the *decomposition* are most specific, applying to the smallest *slices* as the *alignments* are removed in the *idealisation*. So in the *decomposition*, $D_{\mathrm{Scsd},P}$, each *contingent fud derived*, $A * C * F^{\mathrm{T}}$, may be meaningful in the problem domain. By contrast, the *ANN fud derived variables* apply to the entire query *volume*, K^{C}, and so the *derived*, $A * F^{\mathrm{T}}_{\mathrm{gr,lsq},P}$, is less context specific.

202004041015.PAP4