REPRESENTATION OF DISCRETE MAXIMUM ENTROPY DISTRIBUTIONS AS TENSOR NETWORKS

RESEARCH NOTES IN THE ENEXA AND QROM PROJECTS

November 26, 2025

ABSTRACT

This work investigates discrete distributions, which maximize the entropy with respect to linear constraints by the expectation of statistics. We introduce Computation-Activation Networks, which are tensor network formats representing generic maximum entropy distribution. The characterization exploits the face lattice of corresponding mean parameter polytopes to statistics, depending on which we provide rank bounds of the tensor network format. For boolean statistics we provide an interpretation based on propositional logics, which we utilize to construct statistics of generic lattice. We furthermore show that the empirical means are minimal sufficient statistics for the estimation of maximum entropy distributions.

Contents

1 Introduction

1.1 Motivation

Entropy as Information Quantifier: Based on Shannons source code theorem, one can interpret entropy as an quantifier of the information content in a random variable. This has fundamental equivalences in physics and machine learning.

Maximum Entropy in Physics: E.g. Maxwell-Boltzmann distributions.

Maximum Entropy in Learning: Consider a learning problem where we want to estimate a model based on observed data. The maximum entropy problem principle approaches this problem by designing statistics of the data, which means shall be reproduced in the model, and choosing the model reproducing the means of the statistic with least structure. The entropy of a distribution quantifies the degree of structureless in a distribution and is therefore maximized to solve the learning task.

Hybrid Reasoning: Combining logical and probabilistic AI is a common aim of statistical relational AI, neuro-symbolic AI and explainable AI. Logical and probabilistic models can be treated in the tensor network formalism, which thus serves as a unifying representation language.

The maximum support problem: Factorization using the exponential represent always distributions with the maximal support (i.e. that of the base measure). Thus they do not discuss situations appearing in hybrid reasoning. We here investigate the more generic case, where distributions have different support.

1.2 Tensor Notation

Introduce here tensors, variables, contractions.

1.3 The Maximum Entropy Problem

Given a non-negative and non-vanishing base measure $\nu\left[X_{[d]}\right]$, a probability distribution is a non-negative tensor $\mathbb{P}\left[X_{[d]}\right]$ such that

$$\left\langle \mathbb{P}\left[X_{[d]}\right], \nu\left[X_{[d]}\right]\right\rangle [\varnothing] = 1\,.$$

We denote the set of such distributions by

$$\Lambda^{\delta,\mathrm{MAX},\nu} = \left\{\mathbb{P}\left[X_{[d]}\right] \,:\, \mathbb{P}\left[X_{[d]}\right] \geq 0\left[X_{[d]}\right], \left\langle \mathbb{P}\left[X_{[d]}\right], \nu\left[X_{[d]}\right]\right\rangle [\varnothing] = 1\right\}\,.$$

The mean parameter of a distribution $\mathbb{P}\left[X_{[d]}\right]$ to a statistic $T : \bigtimes_{k\in[d]}[m_k] \to \bigtimes_{s\in[n]}[p_s]$ is the vector $\mu[L] \in \mathbb{R}^p$ with the coordinates

$$\mu[L = l] = \mathbb{E}[f_l] = \left\langle f_l\left[X_{[d]}\right], \mathbb{P}\left[X_{[d]}\right], \nu\left[X_{[d]}\right]\right\rangle [\varnothing]\,.$$

The entropy of a distribution is

$$\mathbb{H}_\nu\left[\mathbb{P}\left[X_{[d]}\right]\right] = \left\langle \mathbb{P}\left[X_{[d]}\right], \ln\left[\mathbb{P}\left[X_{[d]}\right]\right], \nu\left[X_{[d]}\right]\right\rangle [\varnothing]\,.$$

The maximum entropy problem given a mean parameter $\mu^*[L]$ is stated by

$$\max_{\mathbb{P}\left[X_{[d]}\right]\in\Lambda^{\delta,\mathrm{MAX},\nu}} \mathbb{H}_\nu\left[\mathbb{P}\left[X_{[d]}\right]\right] \quad \text{subject to} \quad \forall_{l\in[p]} \,:\, \left\langle \mathbb{P}\left[X_{[d]}\right], T_l\left[X_{[d]}\right], \nu\left[X_{[d]}\right]\right\rangle [L] = \mu^*[L = l] \qquad (\mathrm{P}_{T,\mu,\nu})$$

1.4 Contributions

We in this paper provide tensor network representations

- Representation of distributions with maximum entropy, in case of positive realizability: Elementary activation tensors
- Representation of generic maximum entropy distributions: CP activation tensors.

Now, we want to characterize the CP rank of the activation tensors

- Depends on the face of the mean polytope, which contains the mean parameter
- We have thus a well-defined "CP rank" of faces
- Largest faces and vertices have always CP rank of 1, intermediate faces can have larger CP rank

For boolean statistics we further provide insights for boolean statistics (see Chapter 8.5):

- Example of independent statistics (see Exa. 8.28): Always elementary activation tensors (hypercubes)
- Example of partition statistics (see Exa. 8.30):
- Generic criterion for elementary activation: "Cube-like" polytopes (see Def. 8.29)

1.5 Outlook

To prepare for the presentation of our main results we introduce

- Computation-Activation Networks: A tensor network architecture, which will be used to represent maximum entropy distributions
- Mean polytopes: Polytopes, which contain all realizable mean parameter vectors.

We will then show, that dependent on the position of the mean parameter in the mean polytope, we can characterize the corresponding maximum entropy distribution by a Computation-Activation Network.

2 Tensor Network Representation of Exponential Families

We now introduce an architecture of tensor networks, namely Computation-Activation Networks, and show that exponential families can be represented by them.

2.1 Computation-Activation Networks

Given a statistic $T : \bigtimes_{k\in[d]}[m_k] \to \bigtimes_{s\in[n]}[p_s]$ we build its basis encoding tensor

$$\beta^T\left[Y_{[p]}, X_{[d]}\right] = \sum_{x_{[d]}\in\bigtimes_{k\in[d]}[m_k]} \epsilon_{T(x_{[d]})}\left[Y_{[p]}\right] \otimes \epsilon_{x_{[d]}}\left[X_{[d]}\right] .$$

A computation network is any representation of $\beta^T\left[Y_{[p]}, X_{[d]}\right]$ as a tensor network. These can be constructed in the case statistics being a composition of connective functions.

An activation tensor is $\tau\left[Y_{[p]}\right]$ and the Computation-Activation Network of T and τ the tensor

$$\mathbb{P}\left[X_{[d]}\right] = \left\langle \beta^T\left[Y_{[p]}, X_{[d]}\right], \tau\left[Y_{[p]}\right]\right\rangle\left[X_{[d]}|\varnothing\right] .$$

We are interested in decomposition formats of $\tau\left[Y_{[p]}\right]$, where we use sets of tensor networks $\mathcal{T}^{\mathcal{G}}$ on a hypergraph $\mathcal{G}$.

Definition 1. *The family of by T and a $\mathcal{G}$ computable distributions are*

$$\Lambda^{T,\mathcal{G},\nu} = \left\{ \frac{\left\langle \tau\left[Y_{[p]}\right], \beta^T\left[Y_{[p]}, X_{[d]}\right]\right\rangle\left[X_{[d]}\right]}{\left\langle \tau\left[Y_{[p]}\right], \beta^T\left[Y_{[p]}, X_{[d]}\right], \nu\left[X_{[d]}\right]\right\rangle[\varnothing]} \;:\; \tau\left[Y_{[p]}\right] \in \mathcal{T}^{\mathcal{G}} \right\} .$$

2.2 Exponential Families

Definition 2 (Exponential Family)**.** *Given a statistic function*

$$T \,:\, \bigtimes_{k\in[d]} [m_k] \to \mathbb{R}^p$$

and a base measure

$$\nu \,:\, \bigtimes_{k\in[d]} [m_k] \to \mathbb{R}^+$$

with $\langle \nu \rangle [\varnothing] \neq 0$*, the set* $\Gamma^{T,\nu} = \{\mathbb{P}^{(T,\theta,\nu)} \,:\, \theta\,[L] \in \mathbb{R}^p\} \subset \Lambda^{\delta,\mathrm{MAX},\nu}$ *of probability distributions*

$$\mathbb{P}^{(T,\theta,\nu)}\left[X_{[d]}\right] = \frac{\exp\left[\left\langle \sigma^T\left[X_{[d]}, L\right], \theta\,[L]\right\rangle \left[X_{[d]}\right]\right]}{\left\langle \exp\left[\left\langle \sigma^T\left[X_{[d]}, L\right], \theta\,[L]\right\rangle \left[X_{[d]}\right]\right], \nu\left[X_{[d]}\right]\right\rangle [\varnothing]}$$

is called the exponential family to T*.*

To present a tensor network representation, we introduce image interpretation maps

$$I_l \,:\, [|\mathrm{im}\,(s_l)|] \to \mathrm{im}\,(s_l)\ ,$$

which enumerate the possible values of each feature. We treat these maps as tensors with in a variable Y_l with values in $[|\mathrm{im}\,(s_l)|]$.

Theorem 1 (Exponential Families are in Computation-Activation Networks)**.** *Given any base measure* ν *and a sufficient statistic* T *we enumerate for each coordinate* $l \in [p]$ *the image* $\mathrm{im}\,(s_l)$ *by a variable* Y_l *taking values in* $[|\mathrm{im}\,(s_l)|]$*, given an interpretation map*

$$I_l \,:\, [|\mathrm{im}\,(s_l)|] \to \mathrm{im}\,(s_l)\ .$$

For any canonical parameter vector $\theta\,[L] \in \mathbb{R}^p$ *we build the activation cores* $\alpha^{l,\theta}\,[Y_l]$ *for each coordinate* $y_l \in [|\mathrm{im}\,(s_l)|]$ *by*

$$\alpha^{l,\theta}\,[Y_l = y_l] = \exp\left[\theta\,[L = l] \cdot I_l(y_l)\right]$$

and have (see Figure 1)

$$\mathbb{P}^{(T,\theta,\nu)}\left[X_{[d]}\right] = \frac{\left\langle \{\beta^{s_l}\left[Y_l, X_{[d]}\right] \,:\, l \in [p]\} \cup \{\alpha^{l,\theta}\,[Y_l] \,:\, l \in [p]\}\right\rangle \left[X_{[d]}\right]}{\left\langle \{\beta^{s_l}\left[Y_l, X_{[d]}\right] \,:\, l \in [p]\} \cup \{\alpha^{l,\theta}\,[Y_l] \,:\, l \in [p]\} \cup \{\nu\left[X_{[d]}\right]\}\right\rangle [\varnothing]}\ .$$

Proof. For each $x_{[d]} \in \bigtimes_{k\in[d]}[m_k]$ we have

$$\begin{aligned}
&\left\langle \{\beta^{s_l}\left[Y_l, X_{[d]}\right] \,:\, l \in [p]\} \cup \{\alpha^{l,\theta}\,[Y_l] \,:\, l \in [p]\}\right\rangle \left[X_{[d]} = x_{[d]}\right] \\
&\quad = \prod_{l\in[p]} \left\langle \beta^{s_l}\left[Y_l, X_{[d]} = x_{[d]}\right], \alpha^{l,\theta}\,[Y_l]\right\rangle [\varnothing] \\
&\quad = \prod_{l\in[p]} \exp\left[\theta\,[L = l] \cdot I_l(T_l\left[x_{[d]}\right])\right] \\
&\quad = \exp\left[\sum_{l\in[p]} \theta\,[L = l] \cdot I_l(T_l\left[x_{[d]}\right])\right] \\
&\quad = \exp\left[\left\langle \sigma^T\left[X_{[d]}, L\right], \theta\,[L]\right\rangle \left[X_{[d]} = x_{[d]}\right]\right]\ .
\end{aligned}$$

Therefore we have

$$\left\langle \{\beta^{s_l}\left[Y_l, X_{[d]}\right] \,:\, l \in [p]\} \cup \{\alpha^{l,\theta}\,[Y_l] \,:\, l \in [p]\}\right\rangle \left[X_{[d]}\right] = \exp\left[\left\langle \sigma^T\left[X_{[d]}, L\right], \theta\,[L]\right\rangle \left[X_{[d]}\right]\right]\ .$$

The claim follows, since this implies that also the contraction of both sides with $\nu\left[X_{[d]}\right]$ is equivalent. □

We will use the following well known property, that there is a one-to-one map between the canonical parameters and the mean parameters.

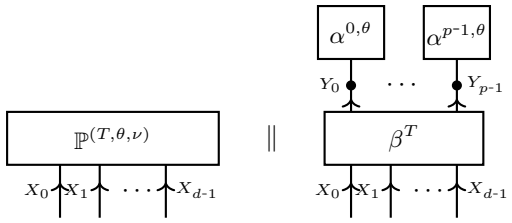


Figure 1: Representation of a member in an exponential family by a Computation-Activation Network with elementary activation. Since the right hand side is not normalized both sides are equal up to a constant.

Lemma 1. *The set of mean parameters of the members of an exponential family is the relative interior of the mean polytope. For $\theta, \tilde{\theta} \in \mathbb{R}^p$ with*

$$\forall l \in [p] : \left\langle \mathbb{P}^{T,\theta,\nu}[X_{[d]}], T_l\left[X_{[d]}\right], \nu\left[X_{[d]}\right]\right\rangle[\varnothing] = \left\langle \mathbb{P}^{T,\tilde{\theta},\nu}[X_{[d]}], T_l\left[X_{[d]}\right], \nu\left[X_{[d]}\right]\right\rangle[\varnothing]$$

we furthermore have $\mathbb{P}^{T,\theta,\nu}[X_{[d]}] = \mathbb{P}^{T,\tilde{\theta},\nu}[X_{[d]}]$.

Proof. See The 3.3 in Wainwright and Jordan. □

Based on this property we define the forward and backward mappings to an exponential family.

Definition 3. *The forward map of an exponential family is the map*

$$F^{(T,\nu)} : \mathbb{R}^p \to (\mathcal{M}_{T,\nu})^\circ$$

defined as

$$\forall l \in [p] : F^{(T,\nu)}(\theta)[L = l] = \left\langle \mathbb{P}^{T,\theta,\nu}[X_{[d]}], T_l\left[X_{[d]}\right], \nu\left[X_{[d]}\right]\right\rangle[\varnothing]$$

Any map $B^{(T,\nu)} : (\mathcal{M}_{T,\nu})^\circ \to \mathbb{R}^p$ with $B^{(T,\nu)} \circ F^{(T,\nu)} = \mathrm{Id}$ is called a backward map.

3 The mean polytope

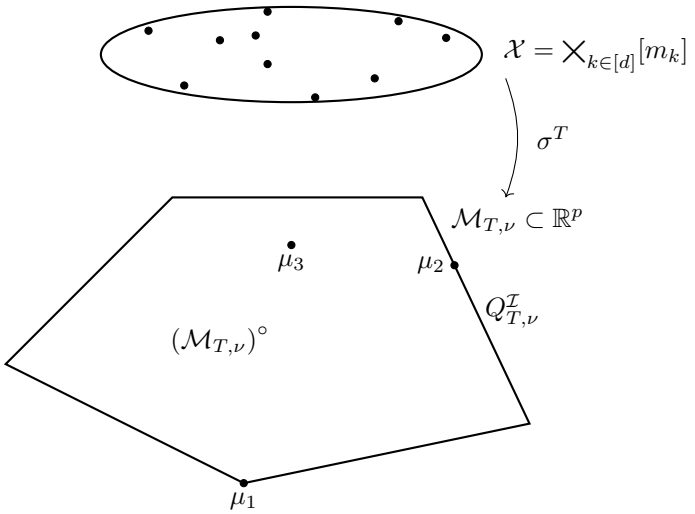
The mean polytope is the set of possible mean parameters given the by a base measure representable distributions. To ease the notation, we use the selection encoding $\sigma^T\left[X_{[d]}, L\right]$ of a statistic

$$\sigma^T\left[X_{[d]}, L = l\right] = T_l\left[X_{[d]}\right] .$$

We define it as

$$\mathcal{M}_{T,\nu} = \left\{\left\langle \mathbb{P}, \sigma^T, \nu\right\rangle[L] : \mathbb{P}\left[X_{[d]}\right] \in \Lambda^{\delta,\mathrm{MAX},\nu}\right\} ,$$

where we denote by $\Lambda^{\delta,\mathrm{MAX},\nu}$ the set of all probability distributions representable with respect to ν.



3.1 Convex hull

Lemma 2. *For any statistic T and base measure ν, the set of mean parameters is the convex hull of the set*

$$\mathcal{N}_{T,\nu} = \{\sigma^T\left[X_{[d]} = x_{[d]}, L\right] \,:\, x_{[d]} \in \bigtimes_{k\in[d]} [m_k],\, \nu\left[X_{[d]} = x_{[d]}\right] \neq 0\}$$

that is $\mathcal{M}_{T,\nu} = \mathrm{conv}\left(\mathcal{N}_{T,\nu}\right)$.

Proof. This follows from

$$\Lambda^{\delta,\mathrm{MAX},\nu} = \mathrm{conv}\left(\frac{1}{\nu\left[X_{[d]} = x_{[d]}\right]} \cdot \epsilon_{x_{[d]}}\left[X_{[d]}\right] \,:\, x_{[d]} \in \bigtimes_{k\in[d]} [m_k],\, \nu\left[X_{[d]} = x_{[d]}\right] \neq 0\right).$$

and for any vertex of this simplex we have

$$\left\langle \frac{1}{\nu\left[X_{[d]} = x_{[d]}\right]} \cdot \epsilon_{x_{[d]}}\left[X_{[d]}\right], \sigma^T\left[X_{[d]}, L\right], \nu\left[X_{[d]}\right] \right\rangle [L] = \sigma^T\left[X_{[d]} = x_{[d]}, L\right].$$

□

Thus the polytope of mean parameters depends on the base measure only through its support.

3.2 Faces

Let us now continue with the investigation of the faces of the mean parameter polytope.

Definition 4. *We say that the convex hull of a subset $\mathcal{N}^{\mathcal{F}}_{T,\nu} \subset \mathcal{N}_{T,\nu}$ is a face of a mean polytope $\mathcal{M}_{T,\nu}$, if and only if there is a face normal vector $\theta_{\mathcal{F}}[L] \in \mathbb{R}^p$ such that*

$$\mathcal{N}^{\mathcal{F}}_{T,\nu} = \mathrm{argmax}_{\mu[L]\in\mathcal{N}_{T,\nu}} \left\langle \mu[L], \theta_{\mathcal{F}}[L] \right\rangle [\varnothing].$$

We denote the face as $\mathcal{F} = \mathrm{conv}\left(\mathcal{N}^{\mathcal{F}}_{T,\nu}\right)$. The set of all faces of the mean polytope $\mathcal{M}_{T,\nu}$ is denoted by $L\left(\mathcal{M}_{T,\nu}\right)$.

$L\left(\mathcal{M}_{T,\nu}\right)$ is called a lattice **?**.

We notice, that each face itself is a convex polytope. What is more, we can characterize these as mean parameter polytopes with respect to refined base measures to be defined next.

Definition 5. *The face measure to the face $\mathcal{F}$ of $\mathcal{M}_{T,\nu}$ is the boolean tensor $\nu^{T,\mathcal{F}}\left[X_{[d]}\right]$ with coordinates to $x_{[d]} \in \bigtimes_{k\in[d]}[m_k]$ by*

$$\nu^{T,\mathcal{F}}\left[X_{[d]} = x_{[d]}\right] = \begin{cases} \nu\left[X_{[d]} = x_{[d]}\right] & \text{if} \quad T\left(x_{[d]}\right) \in \mathcal{N}^{\mathcal{F}}_{T,\nu} \\ 0 & \text{else} \end{cases}.$$

We now specify the mean parameter polytope to any face using the face measure as a refinement of the base measure.

Lemma 3. *For any face $\mathcal{F}$ of $\mathcal{M}_{T,\nu}$ we have*

$$\mathcal{F} = \mathcal{M}_{T,\nu^{T,\mathcal{F}}}.$$

Proof. For any $x_{[d]} \in \bigtimes_{k\in[d]}[m_k]$ we have $\nu^{T,\mathcal{F}}\left[X_{[d]} = x_{[d]}\right] \neq 0$ if and only if $T\left(x_{[d]}\right) \in \mathcal{N}^{\mathcal{F}}_{T,\nu}$ and $\nu\left[X_{[d]} = x_{[d]}\right] \neq 0$. Thus we have

$$\begin{aligned} \mathcal{F} &= \mathrm{conv}\left(\mathcal{N}^{\mathcal{F}}_{T,\nu}\right) \\ &= \mathrm{conv}\left(\sigma^T\left[X_{[d]} = x_{[d]}, L\right] \,:\, T\left(x_{[d]}\right) \in \mathcal{N}^{\mathcal{F}}_{T,\nu},\, \nu\left[X_{[d]} = x_{[d]}\right] \neq 0\right) \\ &= \mathrm{conv}\left(\sigma^T\left[X_{[d]} = x_{[d]}, L\right] \,:\, \nu^{T,\mathcal{F}}\left[X_{[d]} = x_{[d]}\right] \neq 0\right) \\ &= \mathcal{M}_{T,\nu^{T,\mathcal{F}}}. \end{aligned}$$

□

Representability of a distribution with respect to face measures is an equivalent condition for the mean parameter of a distribution to be on a face, as we show next.

Lemma 4. *If and only if for a distribution $\mathbb{P}\left[X_{[d]}\right] \in \Lambda^{\delta,\mathrm{MAX},\nu}$ and a face $\mathcal{F}$ we have*

$$\left\langle \sigma^T\left[X_{[d]}, L\right], \mathbb{P}\left[X_{[d]}\right], \nu\left[X_{[d]}\right]\right\rangle [L] \in \mathcal{F},$$

then $\mathbb{P}\left[X_{[d]}\right]$ is representable with respect to the base measure

$$\nu^{T,\mathcal{I}}\left[X_{[d]}\right].$$

Proof. We have

$$\mu[L] = \sum_{x_{[d]}} \mathbb{P}\left[X_{[d]} = x_{[d]}\right] \cdot \nu\left[X_{[d]} = x_{[d]}\right] \cdot \gamma^T\left[X_{[d]} = x_{[d]}, L\right].$$

Let now $\theta_{\mathcal{F}}[L]$ be a face normal to the face $Q^{\mathcal{I}}_{T,\nu}$. We then have

$$\langle \mu[L], \theta_{\mathcal{F}}[L]\rangle [\varnothing] = \sum_{x_{[d]} \in \bigtimes_{k\in[d]}[m_k]} \mathbb{P}\left[X_{[d]} = x_{[d]}\right] \cdot \nu\left[X_{[d]} = x_{[d]}\right] \cdot \left\langle \gamma^T\left[X_{[d]} = x_{[d]}, L\right] \theta_{\mathcal{F}}[L]\right\rangle [\varnothing].$$

Now if and only if $\mathbb{P}\left[X_{[d]} = x_{[d]}\right] \cdot \nu\left[X_{[d]} = x_{[d]}\right]$ is supported only for $x_{[d]}$ with $T\left(x_{[d]}\right) \in \mathcal{N}^{\mathcal{F}}_{T,\nu}$ we have that

$$\langle \mu[L], \theta_{\mathcal{F}}[L]\rangle [\varnothing] = \max_{\mu[L] \in \mathcal{M}_{T,\nu}} \langle \mu[L], \theta_{\mathcal{F}}[L]\rangle [\varnothing]$$

which is equal to $\mu[L] \in Q^{\mathcal{I}}_{T,\nu}$. Thus, if and only if $\mu[L] \in Q^{\mathcal{I}}_{T,\nu}$ then $\mathbb{P}\left[X_{[d]}\right]$ is only supported at $x_{[d]}$ in the support of $\nu^{T,\mathcal{I}}\left[X_{[d]}\right]$.

□

Let us now investigate tensor network representations of face measures, based on the basis encoding β^T of a statistic.

Theorem 2 (Face measure representation)**.** *For any face $\mathcal{F}$ of $\mathcal{M}$ we have*

$$\nu^{T,\mathcal{I}}\left[X_{[d]}\right] = \left\langle \beta^T\left[Y_{[p]}, X_{[d]}\right], \kappa^{\mathcal{F}}\left[Y_{[p]}\right], \nu\left[X_{[d]}\right]\right\rangle \left[X_{[d]}\right]$$

where

$$\kappa^{\mathcal{F}}\left[Y_{[p]}\right] = \sum_{\mu \in \mathcal{N}^{\mathcal{F}}_{T,\nu}} \epsilon_\mu\left[Y_{[p]}\right].$$

Proof. For any $\mu \in \mathcal{N}^{\mathcal{F}}_{T,\nu}$ the tensor

$$\tau^\mu\left[X_{[d]}\right] = \left\langle \beta^T\left[Y_{[p]}, X_{[d]}\right], \epsilon_\mu\left[Y_{[p]}\right]\right\rangle \left[X_{[d]}\right]$$

is the indicator of the preimage of μ under σ^T. Since preimages the elements in $\mathcal{N}^{\mathcal{F}}_{T,\nu}$ are disjoint, the support of $\tau^\mu\left[X_{[d]}\right]$ is disjoint and their sum

$$\sum_{\mu \in \mathcal{N}^{\mathcal{F}}_{T,\nu}} \tau^\mu\left[X_{[d]}\right]$$

is the indicator of the preimage of $\mathcal{F}$ under σ^T. The face measure obeys thus

$$\begin{aligned}
\nu^{T,\mathcal{F}}\left[X_{[d]}\right] &= \left\langle \left(\sum_{\mu \in \mathcal{N}^{\mathcal{F}}_{T,\nu}} \tau^\mu\left[X_{[d]}\right] \right), \nu\left[X_{[d]}\right] \right\rangle \left[X_{[d]}\right] \\
&= \sum_{\mu \in \mathcal{N}^{\mathcal{F}}_{T,\nu}} \left\langle \beta^T\left[Y_{[p]}, X_{[d]}\right], \epsilon_\mu\left[Y_{[p]}\right], \nu\left[X_{[d]}\right]\right\rangle \left[X_{[d]}\right] \\
&= \left\langle \beta^T\left[Y_{[p]}, X_{[d]}\right], \kappa^{\mathcal{F}}\left[Y_{[p]}\right], \nu\left[X_{[d]}\right]\right\rangle \left[X_{[d]}\right]
\end{aligned}$$

□

We now investigate the representation of face measures by Computation-Activation Networks.

Definition 6. *Let $\mathcal{G}$ be a hypergraph which nodes include $[p]$. We say that a face $Q^{\mathcal{I}}_{T,\nu}$ is representable by $\mathcal{G}$ if and only if there is a set $\mathcal{U}$ of basis vectors with*

$$\mathcal{U} \,:\, \mathcal{U} \cap \mathcal{N}_{T,\nu} = \mathcal{N}^{\mathcal{F}}_{T,\nu}$$

and there is a tensor network $\tau^{\mathcal{G}}$ with respect to the hypergraph $\mathcal{G}$ such that

$$\left\langle \tau^{\mathcal{G}} \right\rangle \left[Y_{[p]}\right] = \sum_{v \in \mathcal{U}} \epsilon_v \left[Y_{[p]}\right] .$$

We call any such tensor network a face activating tensor network.

Lemma 5. *If any only if a face is representable by a hypergraph $\mathcal{G}$ we have*

$$\nu^{T,\mathcal{I}} \left[X_{[d]} | \varnothing\right] \in \Lambda^{T,\mathcal{G},\nu} .$$

Proof. For any $\mathcal{U}$ with $\mathcal{U} \,:\, \mathcal{U} \cap \mathcal{N}_{T,\nu} = \mathcal{N}^{\mathcal{F}}_{T,\nu}$ and tensor network $\tau^{\mathcal{G}}$ respecting the assumptions of Def. 6 we have

$$\begin{aligned}
\left\langle \left\langle \tau^{\mathcal{G}} \right\rangle \left[Y_{[p]}\right], \beta^T \left[Y_{[p]}, X_{[d]}\right] \right\rangle \left[X_{[d]}\right] &= \left\langle \sum_{v \in \mathcal{U}} \epsilon_v \left[Y_{[p]}\right], \beta^T \left[Y_{[p]}, X_{[d]}\right] \right\rangle \left[X_{[d]}\right] \\
&= \left\langle \sum_{v \in \mathcal{N}^{\mathcal{F}}_{T,\nu}} \epsilon_v \left[Y_{[p]}\right], \beta^T \left[Y_{[p]}, X_{[d]}\right] \right\rangle \left[X_{[d]}\right] .
\end{aligned}$$

Here we used in the second equation that only the vertices in $\mathcal{N}_{T,\nu}$ are in the image of T. It follows, that

$$\nu^{T,\mathcal{I}} \left[X_{[d]} | \varnothing\right] = \left\langle \left\langle \tau^{\mathcal{G}} \right\rangle \left[Y_{[p]}\right], \beta^T \left[Y_{[p]}, X_{[d]}\right], \nu \left[X_{[d]}\right] \right\rangle \left[X_{[d]} | \varnothing\right]$$

and thus $\nu^{T,\mathcal{I}} \left[X_{[d]} | \varnothing\right] \in \Lambda^{T,\mathcal{G},\nu}$. □

Let us now investigate, which normalized face measures can be computed using T and a hypergraph $\mathcal{G}$.

Example 1 (Vertices)**.** *Vertices $Q^{\mathcal{I}}_{T,\nu}$ are proper faces of affine dimension 0, that is they consist in single vectors. Since all vertices are in the image $\sigma^T(\mathcal{X})$, there exists an index tuple $x_{[d]} \in \mathcal{X}$ such that $\nu\left[X_{[d]} = x_{[d]}\right] = 1$ and*

$$Q^{\mathcal{I}}_{T,\nu} = \left\{\sigma^T \left[X_{[d]} = x_{[d]}, L\right]\right\} .$$

Then $\kappa^{\mathcal{F}}\left[Y_{[p]}\right]$ is the one-hot encoding of the by an interpretation map I assigned index to $\sigma^T\left[X_{[d]} = x_{[d]}, L\right]$, that is

$$\kappa^{\mathcal{F}}\left[Y_{[p]}\right] = \epsilon_{I^{-1}(\sigma^T\left[X_{[d]} = x_{[d]}, L\right])}\left[Y_{[p]}\right] .$$

In particular, the activation core is elementary and the face measure to any vertex is in $\Lambda^{T,\mathrm{EL},\nu}$.

While vertices are the minimal non-vanishing faces in the face-lattice (see **?**), we now show that also the maximal face, namely the polytope itself, is representable with respect to the elementary hypergraph EL.

Example 2 (Maximal face)**.** *The maximal face $Q^{\varnothing}_{T,\nu} = \mathcal{M}_{T,\nu}$ coincides with the mean polytope itself and is given by the choice $\theta_{\varnothing}[L] = 0[L]$. In this case the corresponding activation tensor to the face measure is trivial, that is*

$$\kappa^{\varnothing}\left[Y_{[p]}\right] = \mathbb{I}\left[Y_{[p]}\right] .$$

$\kappa^{\varnothing}$ is elementary and the normalized face measure $\nu^{T,\varnothing}$ to the maximal face is in $\Lambda^{T,\mathrm{EL}}$.

Extending Example 1, we can provide a coarse estimation of the hypergraph $\mathcal{G}$ required to decompose $\kappa^{\mathcal{F}}$ for generic faces $Q^{\mathcal{I}}_{T,\nu}$.

Lemma 6. *Any face $Q^{\mathcal{I}}_{T,\nu}$ is representable by a* CP *graph with hidden rank*

$$r = \min\left(\left|\mathcal{N}^{\mathcal{F}}_{T,\nu}\right|, \left|\mathcal{N}_{T,\nu}\right| - \left|\mathcal{N}^{\mathcal{F}}_{T,\nu}\right| + 1\right) .$$

Proof. We show the claim by constructing two face activating tensor networks to $Q^{\mathcal{I}}_{T,\nu}$ in a CP hypergraph with hidden rank $\left|\mathcal{N}^{\mathcal{F}}_{T,\nu}\right|$ and in a CP hypergraph with hidden rank $\left|\mathcal{N}_{T,\nu}\right| - \left|\mathcal{N}^{\mathcal{F}}_{T,\nu}\right| + 1$. To show the first representation we enumerate the vertices by a variable I with dimension $r = \left|\mathcal{N}^{\mathcal{F}}_{T,\nu}\right|$, i.e. $\mathcal{N}^{\mathcal{F}}_{T,\nu} = \{v^i[L] \,:\, i \in [n]\}$. Then we define for $l \in [p]$ core tensors $\tau^l\left[Y_l, I\right]$

$$\tau^l\left[Y_l, I = i\right] = \epsilon_{v^i[L=l]}\left[Y_l\right] .$$

Then we have

$$\left\langle \{\tau^l\left[Y_l, I\right] \,:\, l \in [p]\} \right\rangle \left[Y_{[p]}\right] = \sum_{v \in \mathcal{N}^{\mathcal{F}}_{T,\nu}} \epsilon_v\left[Y_{[p]}\right]$$

and thus have found an activation tensor network in a CP graph with hidden rank $\left|\mathcal{N}^{\mathcal{F}}_{T,\nu}\right|$ representing the face $Q^{\mathcal{I}}_{T,\nu}$.

We continue with the second representation, for which we enumerate the set $\mathcal{N}_{T,\nu}/\mathcal{N}^{\mathcal{F}}_{T,\nu}$ by $v^i[L]$ where $i \in [\left|\mathcal{N}_{T,\nu}\right| - \left|\mathcal{N}^{\mathcal{F}}_{T,\nu}\right|]$. We define variable I with dimension $r = \left|\mathcal{N}_{T,\nu}\right| - \left|\mathcal{N}^{\mathcal{F}}_{T,\nu}\right| + 1$ and define for $l \in [p]$ core tensors

$$\tau^l\left[Y_l, I = i\right] = \begin{cases} -\epsilon_{v^i[L=l]}\left[Y_l\right] & \text{if} \quad i < \left|\mathcal{N}_{T,\nu}\right| - \left|\mathcal{N}^{\mathcal{F}}_{T,\nu}\right| \\ \mathbb{I}\left[Y_l\right] & \text{if} \quad i = \left|\mathcal{N}_{T,\nu}\right| - \left|\mathcal{N}^{\mathcal{F}}_{T,\nu}\right| \end{cases} .$$

We then have

$$\begin{aligned} \left\langle \{\tau^l\left[Y_l, I\right] \,:\, l \in [p]\} \right\rangle \left[Y_{[p]}\right] &= \mathbb{I}\left[Y_{[p]}\right] - \sum_{v \in \mathcal{N}_{T,\nu}/\mathcal{N}^{\mathcal{F}}_{T,\nu}} \epsilon_v\left[Y_{[p]}\right] \\ &= \sum_{v \in \mathcal{N}^{\mathcal{F}}_{T,\nu}} \epsilon_v\left[Y_{[p]}\right] + \sum_{v \in \left(\times_{l \in [p]}[p_l]\right)/\mathcal{N}_{T,\nu}} \epsilon_v\left[Y_{[p]}\right] . \end{aligned}$$

We have thus found a face activating tensor network for $Q^{\mathcal{I}}_{T,\nu}$ in a CP format with hidden rank $r = \left|\mathcal{N}_{T,\nu}\right| - \left|\mathcal{N}^{\mathcal{F}}_{T,\nu}\right| + 1$. □

3.3 Partition into Relative Interiors of Faces

Let us now introduce relative interiors, which enables us to find disjoint partitions of the mean polytope.

Definition 7 (Relative Interior)**.** *Let $\mathcal{U} \subset \mathbb{R}^p$ be an arbitrary set and $\mathcal{L}$ be the affine hull of $\mathcal{U}$. Then the relative interior, denoted $(\mathcal{U})^\circ$ is the interior of $\mathcal{U}$ in the affine subspace $\mathcal{L}$.*

Lemma 7. *Any polytope is a disjoint union of the relative interiors of its faces, that is*

$$\mathcal{M}_{T,\nu} = \dot{\bigcup_{\mathcal{F} \in L(\mathcal{M}_{T,\nu})}} \left(Q^{\mathcal{I}}_{T,\nu}\right)^\circ .$$

Proof. For any $\mu \in \mathcal{M}_{T,\nu}$ we find a face such that $\mu \in Q^{\mathcal{I}}_{T,\nu}$. If $\mu \notin \left(Q^{\mathcal{I}}_{T,\nu}\right)^\circ$, then there is a face $Q^{\tilde{\mathcal{F}}}_{T,\nu} \subset Q^{\mathcal{I}}_{T,\nu}$ of smaller affine dimension such that $\mu \in \mathcal{F}$. When continuing this process we reach a face such that $\mu \notin \left(Q^{\mathcal{I}}_{T,\nu}\right)^\circ$, since the faces with affine dimension 0 are vertices and they coincide with their relative interior because they contain a single vector. □

Definition 8. *To each $\mu \in \mathcal{M}_{T,\nu}$ we denote the unique face $Q^{\mathcal{I}}_{T,\nu}$ with $\mu \in \left(Q^{\mathcal{I}}_{T,\nu}\right)^\circ$ by $Q^{\mathcal{F}(\mu)}_{T,\nu}$.*

4 Main results: Tensor network representation of maximum entropy distributions

Given the mean polytope discussion we now characterize the tensor network representation of maximum entropy distributions.

4.1 Maximum entropy on the interior

A classical result states, that the maximum entropy distribution is in the exponential family $\Gamma^{T,\nu}$ (see e.g. Koller and Friedman).

Theorem 3. *If and only if μ^* is in the relative interior of $\mathcal{M}_{T,\nu}$, then the unique solution of the maximum entropy problem is the distribution*

$$\mathbb{P}^{T,\mu^*,\nu}[X_{[d]}] \in \Gamma^{T,\nu}$$

with $\left\langle \mathbb{P}^{T,\mu^*,\nu}[X_{[d]}], \sigma^T\left[X_{[d]}, L\right]\right\rangle[L] = \mu^*[L]$.

Proof. By Lem. 1

$$\mu[L] \in (\mathcal{M}_{T,\nu})^\circ,$$

there is a canonical parameter θ with

$$\left\langle \mathbb{P}^{T,\theta,\nu}[X_{[d]}], \sigma^T\left[X_{[d]}, L\right], \nu\left[X_{[d]}\right]\right\rangle[L] = \mu[L]\ .$$

For any other feasible distribution $\tilde{\mathbb{P}}[X_{[d]}]$ we also have $\left\langle \tilde{\mathbb{P}}[X_{[d]}], \sigma^T\left[X_{[d]}, L\right], \nu\left[X_{[d]}\right]\right\rangle[L] = \mu[L]$ and thus

$$\begin{aligned}
\mathbb{H}_\nu\left[\tilde{\mathbb{P}}, \mathbb{P}^{(T,\theta,\nu)}\right] &= -\left\langle \tilde{\mathbb{P}}, \ln\left[\mathbb{P}^{(T,\theta,\nu)}[X_{[d]}]\right], \nu\left[X_{[d]}\right]\right\rangle[\varnothing] \\
&= -\left\langle \tilde{\mathbb{P}}, \sigma^T\left[X_{[d]}, L\right], \theta[L], \nu\left[X_{[d]}\right]\right\rangle[\varnothing] + A^{(T,\nu)}(\theta) \\
&= -\langle \theta[L], \mu[L]\rangle[\varnothing] + A^{(T,\nu)}(\theta) \\
&= \mathbb{H}_\nu\left[\mathbb{P}^{(T,\theta,\nu)}\right].
\end{aligned}$$

With the Gibbs inequality we have if $\tilde{\mathbb{P}} \neq \mathbb{P}^{(T,\theta,\nu)}$

$$\mathbb{H}_\nu\left[\mathbb{P}^{(T,\hat{\theta},\nu)}\right] - \mathbb{H}_\nu\left[\tilde{\mathbb{P}}\right] = \mathbb{H}_\nu\left[\tilde{\mathbb{P}}, \mathbb{P}^{(T,\hat{\theta},\nu)}\right] - \mathbb{H}_\nu\left[\tilde{\mathbb{P}}\right] > 0$$

and thus $\mathbb{H}_\nu\left[\tilde{\mathbb{P}}\right] < \mathbb{H}_\nu\left[\tilde{\mathbb{P}}\right]$. Therefore, if $\tilde{\mathbb{P}}$ does not coincide with $\mathbb{P}^{(T,\hat{\theta},\nu)}$, it is not a maximum entropy distribution. □

4.2 Main result

Our main result generalizes the maximum entropy characterization of Thm. 3 to arbitrary mean parameters.

Theorem 4 (Generic characterization of Maximum Entropy Solutions)**.** *Let T be a statistic and ν a base measure. For any $\mu[L]$ the maximum entropy problem has a feasible distribution, if and only if $\mu[L] \in \mathcal{M}_{T,\nu}$. In case $\mu[L] \in \mathcal{M}_{T,\nu}$ denote the unique face of $\mathcal{M}_{T,\nu}$ with μ in its relative interior by $Q^{\mathcal{I}}_{T,\nu}$ (see Def. 7). Then the solution of the maximum entropy problem is the member*

$$\mathbb{P}^{(T, B^{T,\nu^{T,\mathcal{I}}}(\mu), \nu^{T,\mathcal{I}})}$$

of the exponential family $\Gamma^{T,\nu^{T,\mathcal{I}}}$, where $\nu^{T,\mathcal{I}}$ is the face measure (see Def. 5). If for a hypergraph $\mathcal{G}$, which nodes appear all in at least one edge, the face is representable with respect to $\mathcal{G}$ (see Def. 6), then the maximum entropy distribution is in $\Lambda^{T,\mathcal{G},\nu}$.

Note, that while we use refined base measures $\nu^{T,\mathcal{I}}$ to characterize the maximum entropy distribution, Thm. 4 states a representation with respect to the original base measure ν.

To prepare for the proof of this theorem we first show in an auxiliary lemma that we can reduce the set of feasible distributions in Problem ($\mathrm{P}_{T,\mu,\nu}$).

Lemma 8. *For any $\mu[L] \in \mathcal{M}_{T,\nu}$ and a face $Q^{\mathcal{I}}_{T,\nu}$ with $\mu[L] \in \left(Q^{\mathcal{I}}_{T,\nu}\right)^\circ$ we have that the solutions of $\mathrm{P}_{T,\mu,\nu}$ and $\mathrm{P}_{T,\mu,\nu^{T,\mathcal{I}}}$ coincide.*

Proof. By Lem. 4 all feasible distributions are representable by the with the face measure refined base measure. We have that any for $\mathrm{P}_{T,\mu,\nu}$ feasible distribution $\mathbb{P}\left[X_{[d]}\right]$ satisfies

$$\left\langle \mathbb{P}\left[X_{[d]}\right], \nu^{T,\mathcal{I}} \right\rangle [\varnothing] = 1$$

and thus $\mathbb{P}\left[X_{[d]}\right] \in \Lambda^{T,\nu^{T,\mathcal{I}}}$. Conversely, any $\mathbb{P}\left[X_{[d]}\right] \in \Lambda^{T,\nu^{T,\mathcal{I}}}$ satisfies

$$\left\langle \mathbb{P}\left[X_{[d]}\right], \nu^{T,\mathcal{I}} \right\rangle [\varnothing] = \left\langle \mathbb{P}\left[X_{[d]}\right], \nu \right\rangle [\varnothing] = 1$$

and thus $\mathbb{P}\left[X_{[d]}\right] \in \Lambda^{T,\nu}$. Problem $\mathrm{P}_{T,\mu,\nu}$ is thus equal to

$$\mathrm{argmax}_{\mathbb{P}\left[X_{[d]}\right] \in \Lambda^{T,\nu^{T,\mathcal{I}}}} \mathbb{H}_{\nu}\left[\mathbb{P}\left[X_{[d]}\right]\right] \quad \text{subject to} \quad \forall_{l \in [p]} : \left\langle \mathbb{P}\left[X_{[d]}\right], T_l\left[X_{[d]}\right], \nu\left[X_{[d]}\right] \right\rangle [L] = \mu^*[L = l]$$

We further have that any $\mathbb{P}\left[X_{[d]}\right] \in \Lambda^{T,\nu^{T,\mathcal{I}}}$

$$\mathbb{H}_{\nu}\left[\mathbb{P}\left[X_{[d]}\right]\right] = \mathbb{H}_{\nu^{T,\mathcal{I}}}\left[\mathbb{P}\left[X_{[d]}\right]\right]$$

and arrive together with the above equivalence at the claim. □

Proof of Thm. 4. **Feasibility Claim:** If and only if $\mu[L] \in \mathcal{M}_{T,\nu}$ then there is by definition a by ν representable $\mathbb{P}\left[X_{[d]}\right]$ reproducing $\mu[L]$. Thus if and only if $\mu[L] \in \mathcal{M}_{T,\nu}$ there is a feasible distribution for the maximum entropy problem.

Characterization Claim: We use the following argumentation to show the second claim:

- By Lem. 7 for any μ we find a unique face $Q^{\mathcal{I}}_{T,\nu}$.
- By Lem. 8 we can reduce the maximum entropy problem Problem ($\mathrm{P}_{T,\mu,\nu}$) to $\mathrm{P}_{T,\mu,\nu^{T,\mathcal{I}}}$ to the base measure $\nu^{T,\mathcal{I}}$.
- By Lem. 3 the face $Q^{\mathcal{I}}_{T,\nu}$ coincides with the polytope $\mathcal{M}_{T,\nu^{T,\mathcal{I}}}$ and in particular μ is in the relative interior of that polytope.
- We can now apply Thm. 3 and get a characterization of the maximum entropy solution as a member of the exponential family.

Representation Claim: By Def. 6 we find a tensor network $\tau^{\mathcal{G}}$ on $\mathcal{G}$ representing the face measure, that is

$$\nu^{T,\mathcal{I}}\left[X_{[d]}\right] = \left\langle \{\tau^{\mathcal{G}}\} \cup \{\beta^T\left[Y_{[p]}, X_{[d]}\right], \nu\left[X_{[d]}\right]\} \right\rangle \left[X_{[d]}\right] .$$

We now contract the activation vectors of the exponential family on this tensor network (see Figure 2). To this end we choose a hyperedge $e(l)$ to each node $l \in [p]$, which is possible by assumption, and define a tensor network $\tilde{\tau}^{\mathcal{G}}$ by core tensors

$$\tilde{\tau}^{e}[e] = \left\langle \{\tau^{e}[Y_e]\} \cup \{\alpha^{l,\theta}[Y_l] : e = e(l)\} \right\rangle [X_e] .$$

Now we have that

$$\left\langle \{\tilde{\tau}^{\mathcal{G}}\} \cup \{\beta^T\left[Y_{[p]}, X_{[d]}\right], \nu\left[X_{[d]}\right]\} \right\rangle \left[X_{[d]}\right] = \left\langle \alpha^{\theta}\left[Y_{[p]}\right], \beta^T\left[Y_{[p]}, X_{[d]}\right], \nu^{T,\mathcal{I}}\left[X_{[d]}\right] \right\rangle \left[X_{[d]}\right]$$

Thus, the activation tensor network $\tilde{\tau}^{\mathcal{G}}$ represents the maximum entropy distribution $\mathbb{P}^{T,\mu,\nu}$ in the family $\Lambda^{T,\mathcal{G},\nu}$ of Computation-Activation Networks. □

4.3 Family of maximum entropy distributions

The family of maximum entropy distributions is the set

$$\{\mathbb{P}^{T,\mu,\nu} : \mu \in \mathcal{M}_{T,\nu},\ \mathbb{P}^{T,\mu,\nu} \quad \text{is a solution of} \quad Problem\ (\mathrm{P}_{T,\mu,\nu})\} .$$

By Thm. 1 this is the union of exponential families to each face measure. Note that these families are disjoint, since each member of an exponential family has support by the support of the face measure.

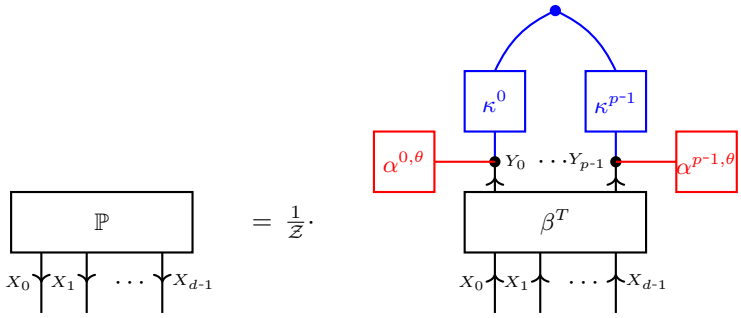


Figure 2: Tensor network decomposition of maximum entropy distributions to the constraint $\mu[L] = \langle \mathbb{P}, \sigma^T \rangle [L]$. Blue: Constraint activation cores κ^l in a CP decomposition, representing the face measure to the minimal face, such that $\mu \in Q^{\mathcal{I}}_{T,\nu}$. Red: Probabilistic activation cores $\alpha^{l,\theta}[Y_l]$ in an elementary decomposition, where each leg core is a scaled exponentials evaluated on the enumerated image $\operatorname{im}(s_l)$.

5 Characterization for boolean statistics

We here study the face CP ranks in case of boolean statistics. We further show that any elementary Computation-Activation Network to boolean statistics is a maximum entropy distribution.

For boolean statistics $F : \bigtimes_{k\in[d]}[m_k] \to \bigtimes_{l\in[p]}[2]$ the mean polytope is a subset of the cube $[0,1]^p$. In this case, any boolean vector in $\mathcal{M}_{F,\nu}$ is a vertex. It follows, that any distribution reproducing a mean parameter $\mu[L]$ on the relative interior of $\mathcal{M}_{F,\nu}$ is positive with respect to ν.

We apply the exponential distribution characterization of the maximum entropy distribution and get that the maximum entropy distribution is in $\Lambda^{T,\mathrm{EL}}$, if and only if the face measure is in $\Lambda^{T,\mathrm{EL}}$. This is exactly the case, when the face is an intersection of the mean polytope with a face of the cupe $[0,1]^p$.

5.1 Set of maximum entropy distributions

Example 3 (Hypercube)**.** *In cases where $X_{[d]}$ are boolean and we have $l \in [p]$ features*

$$f_l\left[X_{[d]} = x_{[d]}\right] = x_l$$

the mean polytope is the hypercube

$$\mathcal{M}_{\{f_l\,:\,l\in[p]\},\mathbb{I}} = [0,1]^p\,.$$

Its face lattice $L([0,1]^p)$ can be enumerated by choosing a subset $A \subset [p]$ and indices $y_A \in \bigtimes_{l\in A}[2]$ and represented by the cartesian products

$$\mathcal{F}^{(A,y_A)} = \bigtimes_{l\in[p]} \mathcal{I}_l$$

where

$$\mathcal{I}_l = \begin{cases} [0,1] & \text{if} \quad l \notin A \\ \{y_l\} & \text{if} \quad l \in A \end{cases}.$$

Each of these faces can be represented with respect to the elementary graph EL*, namely by the tensor product of leg vectors*

$$\alpha^l[Y_l] = \begin{cases} \mathbb{I}[Y_l] & \text{if} \quad l \notin A \\ \epsilon_{y_l}[Y_l] & \text{if} \quad l \in A \end{cases}.$$

This will later be interpreted by propositional logics as the example of atomic formulas.

Definition 9. *We say a polytope $\mathcal{M}_{F,\nu}$ is cube-like, if for any face $Q^{\mathcal{I}}_{F,\mathbb{I}}$ we find a face of the hypercube parametrized by (A, y_A) (see Example 3), such that*

$$\mathcal{F} = \mathcal{M}_{F,\nu} \cap \mathcal{F}^{(A,y_A)}\,.$$

Theorem 5. *Any distribution in $\Lambda^{F,\mathrm{EL},\nu}$ is a maximum entropy distribution with respect to (F,μ,ν) where μ is its mean parameter. Any maximum entropy distribution is realized by $\Lambda^{F,\mathrm{EL},\nu}$ if and only if the mean parameter is in the relative interior of a cube-like face.*

Proof. First claim by decomposing any elementary tensor into exponential and hard activation core. Second claim by characterization of elementary faces by cube-likeness. □

We can now use the same notation as applied for hypercubes to classify the faces of a cube-like polytope.

5.2 Interpretation by propositional formulas

We can understand each feature as a propositonal formula and the variables $X_{[d]}$ as atoms (possibly after a binarization).

Each vertex of the cube, which is not a vertex of the polytope corresponds with the unsatisfiability of a formula

$$\bigwedge_{l\in[p]} \neg^{1-\mu[L=l]} f_l\left[X_{[d]}\right]$$

which is equal with any of the entailment statements for $A \subset [p]$

$$\left(\bigwedge_{l\in A} \neg^{1-\mu[L=l]} f_l\left[X_{[d]}\right]\right) \models \left(\bigwedge_{l\in A} \neg^{\mu[L=l]} f_l\left[X_{[d]}\right]\right) .$$

Along this interpretation we can easily construct examples of statistics, which polytopes are not cube-like.

Example 4 (Maximum entropy distribution with non-elementary activation cores)**.** *Consider two atomic variables X_0 and X_1 and a statistic $\mathcal{F}$ consisting in the formulas*

$$f_0 = (X_0 \wedge X_1) \quad , \quad f_1 = (X_0 \Rightarrow X_1)$$

with the coordinatewise expressions

$$f_0 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \quad , \quad f_1 = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} .$$

We can think of X_0 as a feature on an invoice, and X_1 as a feature on the accounting proposal.

From this we have

$$\beta^{(f_0,f_1)}[Y_0=0,Y_1=0,X_0,X_1] = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \quad , \quad \beta^{(f_0,f_1)}[Y_0=0,Y_1=1,X_0,X_1] = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \quad ,$$
$$\beta^{(f_0,f_1)}[Y_0=1,Y_1=0,X_0,X_1] = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \quad \textit{and} \quad \beta^{(f_0,f_1)}[Y_0=1,Y_1=1,X_0,X_1] = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} .$$

Since the only vanishing slice of $\beta^{\mathcal{F}}$ with respect to the head variables is that to $y_{0,1} = (1,0)$, the vertices of the mean polytope are the vectors to the other head indices. The mean polytope is the convex hull of these vertices

$$\mathcal{M}_{(f_0,f_1)} = \mathrm{conv}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}\right) .$$

*This polytope has a non cube-like face (sketched blue in Figure 3), which is the convex hull of the vertices $[0\,0]^T$, $[1\,1]^T$. This face is parametrized by the (*CP*-rank 2) hard activation core*

$$\kappa^{(0,0),(1,1)}[Y_0,Y_1] = \epsilon_{(0,0)}[Y_0,Y_1] + \epsilon_{(1,1)}[Y_0,Y_1] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

and has the face measure

$$\left\langle \kappa^{(0,0),(1,1)}[Y_0,Y_1], \beta^{\mathcal{F}}[Y_0,Y_1,X_0,X_1] \right\rangle [X_0,X_1] = \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix} .$$

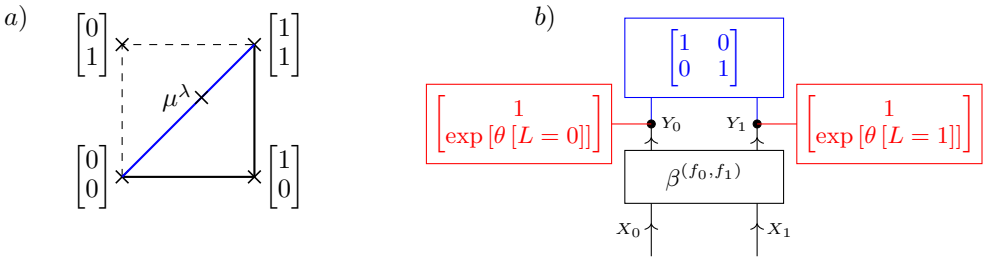


Figure 3: a) Mean polytope of the statistic $\mathcal{F} = (X_0 \wedge X_1, X_0 \Rightarrow X_1)$ (thick), as a subset of the cube $[0,1]^2$ (dashed). The blue line is the face of the polytope, which is not cube like, that is not an intersection of the polytope with the faces of the polytope. We further define for $\lambda \in (0,1)$ a mean parameter $\mu_\lambda [L] = [\lambda\, \lambda]^T$ which is on the interior of the blue face. b) Corresponding Computation-Activation Network being the maximum entropy distribution reproducing $\mu_\lambda [L]$, when λ is the sigmoid of $\theta [L = 0] + \theta [L = 1]$.

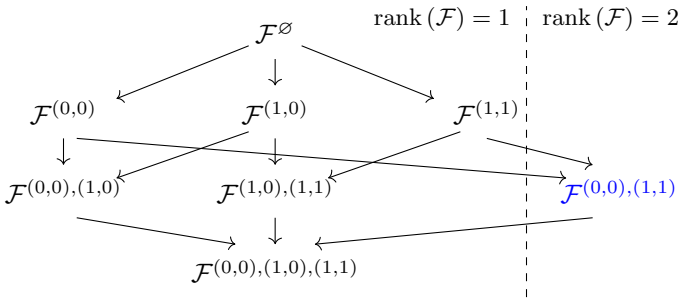


Figure 4: Face lattice $L((X_0 \wedge X_1, X_0 \Rightarrow X_1))$ to Example 4. The directed arrows represent inclusion of the faces, which is a partial order of the faces. The face $\mathcal{F}^{(0,0),(1,1)}$ is the only face, which is not representable by an elementary hypergraph. It is representable in a CP with hidden rank 2

Any mean parameter μ on the interior of that face can be parametrized by a scalar $\lambda \in (0,1)$

$$\mu_\lambda [L] = [\lambda \quad \lambda]^T .$$

With the canonical parameters $\theta [L] \in \mathbb{R}^2$ of the maximum entropy distributions on this face by

$$\mathbb{P} [X_0, X_1] = \frac{1}{1 + \exp [\theta [L = 0] + \theta [L = 1]]} \begin{bmatrix} 0 & 0 \\ 1 & \exp [\theta [L = 0] + \theta [L = 1]] \end{bmatrix}$$

we get the correspondence by the sigmoid

$$\lambda = \frac{1}{1 + \exp [-(\theta [L = 0] + \theta [L = 1])]} .$$

Note, that the hard activation core $\kappa^{(0,0),(1,1)} [Y_0, Y_1]$ to the blue face is the only non-elementary activation core. While the vertices have always elementary cores, the further non-vertex faces have elementary activation cores

$$\kappa^{(0,0),(1,0),(1,1)} [Y_0, Y_1] = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \mathbb{I} [Y_0] \otimes \mathbb{I} [Y_1] \quad , \quad \kappa^{(0,0),(1,0)} [Y_0, Y_1] = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} = \mathbb{I} [Y_0] \otimes \epsilon_0 [Y_1] \quad ,$$

$$\kappa^{(1,0),(1,1)} [Y_0, Y_1] = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} = \epsilon_0 [Y_0] \otimes \mathbb{I} [Y_1] .$$

The maximum entropy distributions to mean parameters on the interior of all other faces than the blue face are represented by Computation-Activation Networks with only elementary activation cores.

Example 5 (Atomic formulas)**.** *Let us consider the case of atomic formulas. The mean polytope in this case is the d-dimensional hypercube*

$$\mathcal{M}_{\mathcal{F}_{[d]}, \mathbb{I}} = [0,1]^p$$

which is called a simple polytope, since each vertex is contained in the minimal number of d facets.

The faces of a hypercube are enumerated in the following way. Each face is characterized by the projections onto each variable, which is either $\{0\}$, $\{1\}$ or $[0,1]$. The projections are represented by the tuple (A, y_A) defined in the following way:

- *We define the set $A \subset [d]$ of variables, such that the projection onto the variable is $\{0\}$ or $\{1\}$*
- *We define to each $l \in A$ an index $y_l = 0$ if the projection is $\{0\}$ and $y_l = 1$ if the projection is $\{1\}$.*

Trivially, each face of the hypercube is a cube face and $\mathcal{M}^{\mathrm{EL}}_{\mathcal{F}_{[d]}} = \mathcal{M}_{\mathcal{F}_{[d]}}$.

More general: If and only if no combination of possibly negated formulas is unsatisfiable, then the mean polytope is a hypercube.

Example 6 (Minterm formulas)**.** *The set of minterm formulas is indexed by $x_{[d]} \in \bigtimes_{k\in[d]}[m_k]$ and given by*

$$f_{x_{[d]}}\left[X_{[d]}\right] = \bigwedge_{l\in[d]} \neg^{1-x_l} f_l\left[X_{[d]}\right] = \epsilon_{x_{[d]}}\left[X_{[d]}\right]$$

where by $f_l\left[X_{[d]}\right]$ we denote the l-th atomic formula (see Example 5). The mean polytope is in the case of the minterm statistic (also referred to as universal statistic) and a boolean base measure ν the standard simplex of dimension

$$\left| x_{[d]} \in \bigtimes_{k\in[d]} [m_k] \,:\, \nu\left[X_{[d]} = x_{[d]}\right] \neq 0 \right| - 1\,,$$

that is the set

$$\mathcal{M}^{\mathrm{EL}}_{\mathcal{F}_\wedge,\nu} = \mathrm{conv}\left(\epsilon_{x_{[d]}}\left[Y_{[p]}\right] \,:\, x_{[d]} \in \bigtimes_{k\in[d]} [m_k],\, \nu\left[X_{[d]} = x_{[d]}\right] \neq 0 \right).$$

In this case, $\Lambda^{\mathcal{F}_\wedge,\mathrm{EL}}$ contains any distribution and therefore trivially realizes any mean parameter in $\mathcal{M}_{\mathcal{F}_\wedge,\mathbb{I}}$.

The faces of the standard simplex are itself standard simplices to base measures $\tilde{\nu}$ with $\tilde{\nu} \prec \nu$. We store them by the tuple (A, y_A), where A is the support of $\tilde{\nu}$ and $y_l = 0$ for $l \in A$. Each face is a cube face, since it is the intersection of $\Lambda^{\mathcal{F}_\wedge,\mathrm{EL}}$ with the cube face (A, y_A). In particular, we have $\mathcal{M}^{\mathrm{EL}}_{\mathcal{F}_\wedge} = \mathcal{M}_{\mathcal{F}_\wedge}$.

More general the mean polytope is a standard simplex, if and only if each formula contradicts all others.

Example 7 (TT representation of diagonal faces)**.** *Consider the vertex set*

$$\mathcal{N} \coloneqq \{0,1\}^p / v\left[L\right]$$

convex polytope. We are interested in the face $\mathcal{F}^{\lhd,1}$ with the normal $\frac{1}{2}(2v\left[L\right] - \mathbb{I}\left[L\right])$. Its vertices are the p elements of $\{0,1\}^p$, which differ from $v\left[L\right]$ in exactly one coordinate (i.e. those with Hamming distance of 1 from $v\left[L\right]$). We denote these vertices by v^l for $l \in [p]$, which coordinates to $\tilde{l} \in [p]$ are

$$v^l\left[L = \tilde{l}\right] = \begin{cases} v\left[L = \tilde{l}\right] & \text{if} \quad \tilde{l} \neq l \\ 1 - v\left[L = \tilde{l}\right] & \text{if} \quad \tilde{l} = l \end{cases}.$$

We now represent their sum in an TT *with hidden ranks $r_0 = \ldots = r_{p-2} = 2$. The boolean hidden variables are denoted be $I_{[p-1]}$ and can be interpreted as indicators, whether the coordinate flip has happened in $[l]$ coordinates. We now construct a* TT *cores for $l = 0$ by*

$$\tau^0\left[Y_0, I_0\right] = \epsilon_{v[L=0]}\left[Y_0\right] \otimes \epsilon_0\left[I_0\right] + \epsilon_{1-v[L=0]}\left[Y_0\right] \otimes \epsilon_1\left[I_0\right]$$

further for $l \notin \{0, p-1\}$ the cores

$$\begin{aligned}\tau^l\left[I_{l-1}, Y_l, I_l\right] = &\epsilon_0\left[I_{l-1}\right] \otimes \epsilon_{1-v[L=l]}\left[Y_l\right] \otimes \epsilon_1\left[I_l\right] + \epsilon_0\left[I_{l-1}\right] \otimes \epsilon_{v[L=l]}\left[Y_l\right] \otimes \epsilon_0\left[I_l\right] \\ &+ \epsilon_1\left[I_{l-1}\right] \otimes \epsilon_{v[L=l]}\left[Y_l\right] \otimes \epsilon_1\left[I_l\right]\end{aligned}$$

and for $l = p-1$

$$\tau^{p-1}\left[I_{p-1}, Y_{p-1}\right] = \epsilon_0\left[I_{p-1}\right] \otimes \epsilon_{1-v[L=p-1]}\left[Y_{p-1}\right] + \epsilon_1\left[I_{p-1}\right] \otimes \epsilon_{v[L=p-1]}\left[Y_{p-1}\right].$$

For this tensor network in the TT *format we have*

$$\sum_{l\in[p]} \epsilon_{v^l}\left[Y_{[p]}\right] = \left\langle \left\{\tau^0\left[Y_0, I_0\right], \tau^{p-1}\left[I_{p-1}, Y_{p-1}\right]\right\} \cup \left\{\tau^l\left[I_{l-1}, Y_l, I_l\right] \,:\, l \notin \{0, p-1\}\right\}\right\rangle \left[Y_{[p]}\right] .$$

The TT *multirank of* 2 *is furthermore minimal, since each matrification based on a partition of* $[p]$ *into non-empty sets has a matrix rank of* 2*. With respect to such partitions also the tensor* $\epsilon_v\left[Y_{[p]}\right] + \left(\sum_{l\in[p]} \epsilon_{v^l}\left[Y_{[p]}\right]\right)$ *has matrix ranks of* 2*. Since these are the only two activation tensors for the face* $\mathcal{F}^{\triangleleft,1}$*.*

Example 8 (Generalization of Example 7 to larger TT ranks)**.** *We now generalize the construction of Example 7 by using the Hadamard distance* $d(\cdot,\cdot)$ *in* $\{0,1\}^p$*, which counts the number of coordinates two vertices differ in. For* $s \in \{1,\ldots,p-1\}$ *we define for a fixed* $v\left[L\right] \in \{0,1\}^p$ *a polytope as the convex hull*

$$\left\{\tilde{v}[L] \,:\, d(\tilde{v}, v) \geq s\right\} .$$

The face to the normal $\frac{1}{2}(2v\left[L\right] - \mathbb{I}\left[L\right])$ *is the convex hull*

$$\mathcal{F}^{\triangleleft,s} \coloneqq \operatorname{conv}\left(\left\{\tilde{v}[L] \,:\, d(\tilde{v}, v) = s\right\}\right)$$

containing $\binom{p}{s}$ *vertices. We label these vertices by subsets* $A \subset [p]$ *of cardinality* s *and define for* $l \in [p]$

$$v^A\left[L = l\right] = \begin{cases} v\left[l \in [p]\right] & \text{if} \quad l \notin A \\ 1 - v\left[l \in [p]\right] & \text{if} \quad l \notin A \end{cases}$$

We now construct a TT *with hidden variables* I_l *and ranks* $r_l = \min(l+1, p-l+1, s+1)$ *to represent the sum of their one-hot encodings. To this end, let there be* TT *cores for* $l = 0$ *by*

$$\tau^0\left[Y_0, I_0\right] = \epsilon_{v[L=0]}\left[Y_0\right] \otimes \epsilon_0\left[I_0\right] + \epsilon_{1-v[L=0]}\left[Y_0\right] \otimes \epsilon_1\left[I_0\right]$$

further for $l \notin \{0, p-1\}$ *the cores*

$$\tau^l\left[I_{l-1}, Y_l, I_l\right] = \sum_{l\in\{\max(s-p+l,0),\ldots,\max(l,s)\}} \epsilon_l\left[I_l\right] \otimes \left(\epsilon_{v[L=l]}\left[Y_l\right] \otimes \epsilon_l\left[I_{l+1}\right] + \epsilon_{1-v[L=l]}\left[Y_l\right] \otimes \epsilon_{l+1}\left[I_{l+1}\right]\right)$$

and for $l = p-1$

$$\tau^{p-1}\left[I_{p-1}, Y_{p-1}\right] = \epsilon_0\left[I_{p-1}\right] \otimes \epsilon_{1-v[L=p-1]}\left[Y_{p-1}\right] + \epsilon_1\left[I_{p-1}\right] \otimes \epsilon_{v[L=p-1]}\left[Y_{p-1}\right] .$$

Based on the interpretation, that the hidden variables I_l *count the Hamming distance of the vectors* $v\left[L\right]|_{\mathbb{R}^l \times 0_{p-l}}$ *and the respective* $v^A\left[L\right]|_{\mathbb{R}^l \times 0_{p-l}}$ *one can show that*

$$\sum_{A\subset[p]\,:\,|A|=s} \epsilon_{v^A}\left[Y_{[p]}\right] = \left\langle \left\{\tau^0\left[Y_0, I_0\right], \tau^{p-1}\left[I_{p-1}, Y_{p-1}\right]\right\} \cup \left\{\tau^l\left[I_{l-1}, Y_l, I_l\right] \,:\, l \notin \{0, p-1\}\right\}\right\rangle \left[Y_{[p]}\right] .$$

Example 9 (Generalization of Example 8 to arbitrary HT formats)**.** *Instead of aligning the Hamming count variables linearly, one can find a representation of the activation tensor*

$$\sum_{A\subset[p]\,:\,|A|=s} \epsilon_{v^A}\left[Y_{[p]}\right]$$

in an arbitrary tree hypergraph format, which hidden ranks are bounded by the number of leafs in the subtree and $s+1$*.*

At any leaf of the tree we define a $2x2$ *matrix*

$$\tau^l\left[Y_l, I_l\right] = \epsilon_{v[L=l]}\left[Y_l\right] \otimes \epsilon_0\left[I_l\right] + \epsilon_{1-v[L=l]}\left[Y_l\right] \otimes \epsilon_1\left[I_l\right] .$$

At each intermediate non-root hyperedge we choose an outgoing counting variable $I_{e^{\text{out}}}$ *with dimension by* $n_{e^{\text{out}}} = \min(\sum_{v\in e^{\text{in}}} n_{e^{\text{in}}}, s+1)$ *and define a tensor with the slices*

$$\tau^e\left[I_e = i_e, I_{e^{\text{in}}}\right] = \beta^+\left[Y_+ = i_{e^{\text{out}}}, I_{e^{\text{in}}}\right] .$$

We further build at the root hyperedge e *a tensor*

$$\tau^e\left[I_e\right] = \sum_{i\,:\,\langle i\rangle[\varnothing]=s} \epsilon_{i_e}\left[I_e\right] .$$

Using the counting variable interpretation one can now show that

$$\sum_{A\subset[p]\,:\,|A|=s} \epsilon_{v^A}\left[Y_{[p]}\right] = \left\langle \left\{\tau^e\left[I_e\right] \,:\, e \in \mathcal{E}\right\} \cup \left\{\tau^l\left[Y_l, I_l\right] \,:\, v \in \mathcal{V}\right\}\right\rangle \left[Y_{[p]}\right]$$

5.3 Construction of Hybrid Logic Networks

We now constructively show, that any convex polytope with boolean vertices in $\mathbb{R}^p$ (a so called 0-1 polytope, see Ziegler) is the mean polytope of a family of Hybrid Logic Networks.

Theorem 6. *Let $\mathcal{M}$ an arbitrary polytope with boolean vertices in $\mathbb{R}^p$. Then we construct propositional formulas on atoms $X_{[p]}$ by*

$$f_0\left[X_0\right] = \begin{cases} \top & \text{if} \quad \mathcal{M}|_{\mathbb{R}^1 \times 0_{p-1}} = \{1\} \\ \bot & \text{if} \quad \mathcal{M}|_{\mathbb{R}^1 \times 0_{p-1}} = \{0\} \\ X_0 & \text{if} \quad \mathcal{M}|_{\mathbb{R}^1 \times 0_{p-1}} = [0,1] \end{cases}$$

and iteratively for $l \in [p]$ with $l \geq 1$ by

$$f_l\left[X_{[l+1]}\right] = \bigwedge_{v[L] \in \mathcal{M}|_{\mathbb{R}^l \times 0_{p-l}} \cap \{0,1\}^p} \left(\left(\bigwedge_{\tilde{l} \in [l+1]} \neg^{1-v[L=l]} f_{\tilde{l}}\left[X_{[\tilde{l}]}\right] \right) \Rightarrow \begin{cases} \top & \text{if} \quad \mathcal{M}|_{v \times \mathbb{R}^1 \times 0_{p-l-1}} = \{1\} \\ \bot & \text{if} \quad \mathcal{M}|_{v \times \mathbb{R}^1 \times 0_{p-l-1}} = \{0\} \\ X_l & \text{if} \quad \mathcal{M}|_{v \times \mathbb{R}^1 \times 0_{p-l-1}} = \{0,1\} \end{cases} \right).$$

Here we denote by $\mathcal{M}|_V$ the projections of the vertices in $\mathcal{M}$ onto the subspaces V, and by 0_p the zero vector in $\mathbb{R}^p$.

Proof. We show per induction, that for any $l \in [p]$ the family of Hybrid Logic Networks with the statistic $f_{[l+1]}$ by the first $l+1$ formulas has the mean polytope

$$\mathcal{M}_{f_{[l+1]},\mathbb{I}} = \mathcal{M}|_{\mathbb{R}^l \times 0_{p-l}}\,. \tag{1}$$

$l = 0$: The polytope $\mathcal{M}|_{\mathbb{R}^l \times 0_{p-l}} = \{1\}$ (respectively $\mathcal{M}|_{\mathbb{R}^l \times 0_{p-l}} = \{0\}$) is reproduced by f_0 being a tautology (respectively a contradiction). In the case $\mathcal{M}|_{\mathbb{R}^l \times 0_{p-l}} = [0,1]$ the polytope is reproduced by the any formula, which is neither a tautology nor a contradiction, and the atomic formula X_0 is an example of such an contingency. Since the projection of a 0-1 polytope onto the first coordinates is itself a 0-1 polytope, these are the only possible cases and we conclude that in all

$$\mathcal{M}_{f_{[l+1]},\mathbb{I}} = \mathcal{M}|_{\mathbb{R}^l \times 0_{p-l}}\,.$$

$l \to l+1$: Let us assume, that (1) holds for a $l \in [p]$. Then for any $x_{[d]} \in \bigtimes_{k \in [d]} [m_k]$ there is exactly one $v[L] \in \mathcal{M}|_{\mathbb{R}^l \times 0_{p-l}}$ such that $x_{[d]}$ is a model of

$$f_{l,v} \coloneqq \bigwedge_{\tilde{l} \in [l+1]} \neg^{1-v[L=l]} f_{\tilde{l}}\left[X_{[\tilde{l}]}\right]\,.$$

This holds, since by the mutual contradiction of these formulas at most one can have $x_{[d]}$ as a model. Further, if none would have $x_{[d]}$ as a model, then the to $x_{[d]}$ corresponding vector of satisfactions $(f_{\tilde{l}}\left[X_{[d]} = x_{[d]}\right])_{\tilde{l} \in [l]}$ is not in $\mathcal{M}|_{\mathbb{R}^l \times 0_{p-l}}$, which can not be the case. Now, the implications to all $\tilde{v}$ except for v are for the models $x_{[d]}$ of $f_{l,v}$ satisfied, and the satisfaction of f_l thus only depends on the head of the implication to v. It follows, that the vertices of $\mathcal{M}_{f_{[l+2]},\mathbb{I}}$ sharing the first l coordinates with v are determined by the head of the implication at v. With the same arguments as in the case $l = 0$ we now notice, that in the three cases we construct vertex sets $v \times \{1\}$, $v \times \{1\}$ or $v \times \{0,1\}$ if and only if they appear in the polytope $\mathcal{M}|_{\mathbb{R}^{l+1} \times 0_{p-l-1}}$. This establishes for each vertex v of $\mathcal{M}_{f_{[l+1]},\mathbb{I}}$ that

$$\left(\mathcal{M}_{f_{[l+2]},\mathbb{I}}\right)|_{v \times \mathbb{R} \times 0_{p-l-1}} = \left(\mathcal{M}|_{\mathbb{R}^l \times 0_{p-l}}\right)|_{v \times \mathbb{R} \times 0_{p-l-1}}\,.$$

Since for any vertex in $\mathcal{M}$ we find a unique vertex in $\mathcal{M}_{f_{[l+1]},\mathbb{I}}$ sharing the first l coordinates, we have that (1) holds for $l+1$.

By induction the equation (1) holds for arbitrary $l \in [p]$. For $l = p-1$ the equation is the claim. □

6 Mean as a Statistic

We in this section take the perspective of estimating a maximum entropy distribution given observed data.

- μ_D is an unbiased estimator of μ, i.e. $\mathbb{E}\left[\mu_D\left[L\right]\right] = \mu$

- μ_D is a consistent estimator of μ, i.e. $\mu_D[L] \to \mu$ by the law of large numbers coordinatewise almost everywhere.
- μ_D is the minimal variance unbiased estimator of μ.

The mean parameter μ_D given a dataset can be understood as a statistic of the dataset. We here show that for the family of maximum entropy distributions this statistic is a minimal sufficient statistic.

The family of maximum entropy distributions is the set

$$\left\{\mathbb{P}^{\mu}\left[X_{[d]}\right] : \mu \in \mathcal{M}_{T,\nu}\right\}$$

which has been characterized above by a union of exponential families with respect to face measures.

Taking a frequentist perspective we now understand datasets by random variables $X_{[d]\times[m]}$, where for $j \in [m]$ the variables $X_{[d],j}$ are drawn i.i.d. from a maximum entropy distribution. The mean statistic is then a tensor

$$\mu_D\left[X_{[d]\times[m]}, L\right]$$

whith coordinates

$$\mu_D\left[X_{[d]\times[m]} = x_{[d]\times[m]}, L\right] = \frac{1}{m}\sum_{j\in[m]} \sigma^T\left[X_{[d],j} = x_{[d],j}, L\right] .$$

Theorem 7. *The mean statistic is a sufficient statistic for the family of maximum entropy distributions (T,ν). If the base measure has maximal support, the mean statistic is in addition minimal.*

Proof. It sufficies to show that the likelihood is a function of μ_D. Let us choose a face $\mathcal{I}$ of $\mathcal{M}_{T,\nu}$, then the likelihood is different from 0 if and only if the empirical distribution is representable with respect to the face measure. This is the case if and only if μ_D is on the face. In case that μ_D is on the face, then the likelihood of any distribution on that face exponential family is

$$\exp\left[m\cdot\left(\left\langle\mu_D\left[X_{[d]\times[m]}, L\right], \theta[L]\right\rangle[\varnothing] - A^{(T,\nu)}(\theta)\right)\right]$$

We have thus shown that the likelihood is always a function of μ_D.

The minimality claim can be shown by the constant quotient criterion (see Thm. 6.2.13 in Casella and Berger). □

7 Outlook

This is relevant for Parameter Estimation: Instabilities when fitting a mean parameter on a non cube-like face.

Border-rank development to resolve the issue. CITE

References

George Casella and Roger Berger. *Statistical Inference*. Cengage Learning. ISBN 978-0-534-24312-8.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 1. edition edition. ISBN 978-0-262-01319-2.

Martin J. Wainwright and Michael Irwin Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers Inc. ISBN 978-1-60198-184-4.

Günter M. Ziegler. Lectures on 0/1-polytopes. In Gil Kalai and Günter M. Ziegler, editors, *Polytopes — Combinatorics and Computation*, pages 1–41. Birkhäuser. ISBN 978-3-0348-8438-9. doi: 10.1007/978-3-0348-8438-9_1. URL https://doi.org/10.1007/978-3-0348-8438-9_1.