REPRESENTATION OF DISCRETE MAXIMUM ENTROPY DISTRIBUTIONS AS TENSOR NETWORKS

RESEARCH NOTES IN THE ENEXA AND QROM PROJECTS

November 21, 2025

ABSTRACT

We here summarize results of the main report on maximum entropy distributions. The principle of maximum entropy serves as motivation for Computation Activation Networks. We then restrict to these distributions and study relative entropy minimization problems.

In the report, first discussion on the mean polytope are in Chapter 3, and the general maximum entropy problem in Chapter 4. Chapter 8 contains the discussion on maximum entropy distributions in case of boolean statistics.

Contents

1 Contents

We in this paper provide tensor network representations

- Representation of any distribution with a sufficient statistics: Generic activation tensors.
- 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)

2 Motivation

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.

3 The Maximum Entropy Problem

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

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

We express the computation of the mean parameter in the contraction of the selection encoding $\sigma^{\mathcal{S}}\left[X_{[d]}, L\right]$ of $\mathcal{S}$

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

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

$$\max_{\mathbb{P}\left[X_{[d]}\right]\in\Lambda^{\delta,\mathrm{MAX},\nu}} \mathbb{H}\left[\mathbb{P}\left[X_{[d]}\right]\right] \quad \text{subject to} \quad \left\langle \mathbb{P}\left[X_{[d]}\right], \sigma^{\mathcal{S}}\left[X_{[d]}, L\right]\right\rangle \left[L\right] = \mu^*[L] \tag{$\mathrm{P}_{\mathcal{S},\mu,\nu}$}$$

where $\Lambda^{\delta,\mathrm{MAX},\nu}$ is the set of distributions, which are representable with respect to the base measure ν.

A quick argument shows, that maximum entropy distributions always have $\mathcal{S}$ as a sufficient statistics.

Theorem 1. *Any maximum entropy distribution with respect to a moment constraint on $\mathcal{S}$ and a base measure ν has the sufficient statistic $\mathcal{S}$.*

Proof. Let $\mathbb{P}\left[X_{[d]}\right]$ be a feasible distribution for the maximum entropy problem, which does not have a sufficient statistic $\mathcal{S}$. Then we find $x_{[d]}, \tilde{x}_{[d]} \in \bigtimes_{k\in[d]}[m_k]$ with $x_{[d]} \neq \tilde{x}_{[d]}$, $\mathcal{S}\left(x_{[d]}\right) = \mathcal{S}\left(\tilde{x}_{[d]}\right)$, $\nu\left[X_{[d]} = x_{[d]}\right] \neq 0$, $\nu\left[X_{[d]} = \tilde{x}_{[d]}\right]$ and $\mathbb{P}\left[X_{[d]} = x_{[d]}\right] \neq \mathbb{P}\left[X_{[d]} = \tilde{x}_{[d]}\right]$. We then define a distribution $\tilde{\mathbb{P}}[X_{[d]}]$ coinciding with $\mathbb{P}\left[X_{[d]}\right]$ except for the coordinates $x_{[d]}, \tilde{x}_{[d]}$, where we set

$$\tilde{\mathbb{P}}[X_{[d]} = x_{[d]}] = \tilde{\mathbb{P}}[X_{[d]} = \tilde{x}_{[d]}] = \frac{\mathbb{P}\left[X_{[d]} = x_{[d]}\right] + \tilde{\mathbb{P}}[X_{[d]} = \tilde{x}_{[d]}]}{2}$$

We notice that $\tilde{\mathbb{P}}[X_{[d]}]$ is also a feasible distribution with an larger entropy than $\mathbb{P}\left[X_{[d]}\right]$. Therefore, a distribution which does not have the sufficient statistic $\mathcal{S}$ cannot be a maximum entropy distribution. □

This shows that any maximum entropy distribution is in $\Lambda^{\mathcal{S},\mathrm{MAX}}$, where MAX is the maximal hypergraph $\mathrm{MAX} = ([p], \{[p]\})$. We search for sparse representations of the corresponding activation tensors and investigate in which cases the maximum entropy distribution is also in $\Lambda^{\mathcal{S},\mathcal{G}}$ for sparser hypergraphs $\mathcal{G}$.

3.1 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.

4 Tensor Notation

4.1 Computation-Activation Networks

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

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

A computation network is any representation of $\beta^{\mathcal{S}}\left[Y_{[p]}, X_{[d]}\right]$ 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 $\mathcal{S}$ and τ the tensor

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

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 $\mathcal{G}$. The family of by $\mathcal{S}$ and a $\mathcal{G}$ computable distributions are

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

4.2 CP decompositions

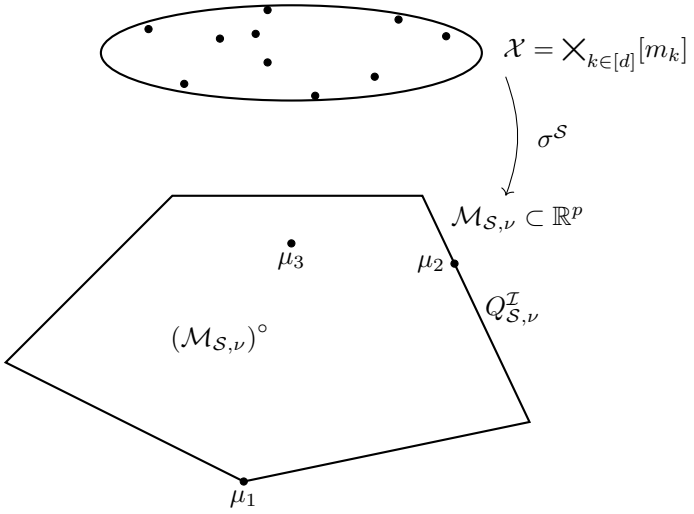
We here introduce the CP decomposition of tensors and the restriction to bas+. This will be used to represent face measures as computation activation networks.

5 The mean polytope

The mean polytope is the set of mean parameters to any distribution. We define it

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

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



5.1 Convex hull

The mean polytope is the convex hull

$$\mathcal{M}_{\mathcal{S},\nu} = \mathrm{conv}\left(\sigma^{\mathcal{S}}\left[X_{[d]} = x_{[d]}, L\right] \,:\, x_{[d]} \in \bigtimes_{k\in[d]} [m_k],\, \nu\left[X_{[d]} = x_{[d]}\right] = 1 \right) .$$

It is thus a convex polytope, inherited by the convex polytope of distributions (the standard simplex). We can characterize the maximum entropy distribution based on the position of the mean parameter in the mean polytope. To be more precise, any polytope decomposes into effective interiors of its faces and we characterize the maximum entropy distribution depending on the face to the mean parameter.

5.2 Faces

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

Definition 1. *Given a mean parameter polytope $\mathcal{M}_{\mathcal{S},\nu}$ in the half space representation of Thm.* **??***, and any subset $\mathcal{I} \subset [n]$ we say that the set*

$$Q^{\mathcal{I}}_{\mathcal{S},\nu} = \{\mu[L] \in \mathcal{M}_{\mathcal{S},\nu} \,:\, \forall_{i\in\mathcal{I}}\ \langle \mu[L], a_i[L] \rangle [\varnothing] = b_i\}$$

is the face to the constraints $\mathcal{I}$.

While all inequalities in a half-space representation are satisfied for any element of the polytope, we defined faces by the additional sharp satisfaction of a subset of the half-space inequalities. In this way, the faces build the boundary of $\mathcal{M}_{\mathcal{S},\nu}$. This can be easily verified, since for any vector $\mu[L] \in \mathcal{M}_{\mathcal{S},\nu}$, for which no halfspace inequalities hold sharply, also a neighborhood satisfies the halfspace inequalities. If any halfspace inequality holds sharply, in the other case, the vector is a member of the corresponding face.

If $\mathcal{S}$ is not minimal with respect to ν, we find a non-vanishing vector $V[L]$ and a scalar $\lambda \in \mathbb{R}$ such that

$$\left\langle \sigma^{\mathcal{S}}\left[X_{[d]}, L\right], V[L], \nu\left[X_{[d]}\right] \right\rangle \left[X_{[d]}\right] = \lambda \cdot \nu\left[X_{[d]}\right] .$$

This implies, that any probability distribution $\mathbb{P}\left[X_{[d]}\right]$ representable with ν satisfies

$$\left\langle \mathbb{P}\left[X_{[d]}\right], \sigma^{\mathcal{S}}\left[X_{[d]}, L\right], V[L], \nu\left[X_{[d]}\right] \right\rangle [\varnothing] = \lambda \cdot \left\langle \mathbb{P}\left[X_{[d]}\right], \nu\left[X_{[d]}\right] \right\rangle [\varnothing] = \lambda .$$

Any $\mu[L] \in \mathcal{M}_{\mathcal{S},\nu}$ then satisfies

$$\langle \mu[L], V[L] \rangle [\varnothing] = \lambda .$$

Thus, the polytope $\mathcal{M}_{\mathcal{S},\nu}$ is contained in an affine linear subspace and has vanishing interior. We can further understand this equation as two half-space inequalities

$$\langle \mu[L], V[L] \rangle [\varnothing] \leq \lambda \quad \text{and} \quad \langle \mu[L], V[L] \rangle [\varnothing] \geq \lambda ,$$

which can be integrated into any half-space representation. We conclude, that in the case of non-minimal statistics, the whole polytope $\mathcal{M}_{\mathcal{S},\nu}$ is a face itself, since it satisfies these half-space inequalities sharply.

Lemma 1. *For each face $Q^{\mathcal{I}}_{\mathcal{S},\nu}$ we have*

$$Q^{\mathcal{I}}_{\mathcal{S},\nu} = \operatorname{conv}\left(\sigma^{\mathcal{S}}\left[X_{[d]} = x_{[d]}, L\right] \,:\, x_{[d]} \in (\sigma^{\mathcal{S}})^{-1}(Q^{\mathcal{I}}_{\mathcal{S},\nu}),\, \nu\left[X_{[d]} = x_{[d]}\right] = 1\right) .$$

Proof. This holds, since each face is the convex hull of the contained vertices (see Proposition 2.2 and 2.3 in Ziegler). Since the vertices are contained in the image of the statistic encoding $\sigma^{\mathcal{S}}$, the vertices contained in $Q^{\mathcal{I}}_{\mathcal{S},\nu}$ are contained in the set

$$\sigma^{\mathcal{S}}\left[X_{[d]} = x_{[d]}, L\right] \,:\, x_{[d]} \in (\sigma^{\mathcal{S}})^{-1}(Q^{\mathcal{I}}_{\mathcal{S},\nu}) . \qquad \square$$

Lem. 1 implies in particular, that faces are mean parameter polytopes with respect to refined base measures. For reference in later chapters, we define these refined base measures next as face measures.

Definition 2. *The face measure to the face $Q^{\mathcal{I}}_{\mathcal{S},\nu}$ of $\mathcal{M}_{\mathcal{S},\nu}$ is the boolean tensor $\nu^{\mathcal{S},\mathcal{I}}\left[X_{[d]}\right]$ with coordinates to $x_{[d]} \in \bigtimes_{k\in[d]}[m_k]$ by*

$$\nu^{\mathcal{S},\mathcal{I}}\left[X_{[d]} = x_{[d]}\right] = \begin{cases} 1 & \text{if} \quad \gamma^{x_{[d]}} \in Q^{\mathcal{I}}_{\mathcal{S},\nu} \\ 0 & \text{else} \end{cases} .$$

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

Lemma 2. *For any face $Q^{\mathcal{I}}_{\mathcal{S},\nu}$ of $\mathcal{M}_{\mathcal{S},\nu}$, we have with the refined base measure*

$$\tilde{\nu}\left[X_{[d]}\right] = \left\langle \nu\left[X_{[d]}\right], \nu^{\mathcal{S},\mathcal{I}}\left[X_{[d]}\right] \right\rangle \left[X_{[d]}\right]$$

that

$$Q^{\mathcal{I}}_{\mathcal{S},\nu} = \mathcal{M}_{\mathcal{S},\tilde{\nu}} .$$

Proof. We notice that for any $x_{[d]} \in \bigtimes_{k\in[d]}[m_k]$, $x_{[d]} \in (\sigma^{\mathcal{S}})^{-1}(Q^{\mathcal{I}}_{\mathcal{S},\nu})$ is equal to $\nu^{\mathcal{S},\mathcal{I}}\left[X_{[d]} = x_{[d]}\right] = 1$ and thus

$$\left\{x_{[d]} \,:\, x_{[d]} \in (\sigma^{\mathcal{S}})^{-1}(Q^{\mathcal{I}}_{\mathcal{S},\nu}),\, \nu\left[X_{[d]} = x_{[d]}\right] = 1\right\} = \left\{x_{[d]} \,:\, \tilde{\nu}\left[X_{[d]} = x_{[d]}\right] = 1\right\} .$$

In combination with Lem. 1 we then get

$$\begin{aligned} Q^{\mathcal{I}}_{\mathcal{S},\nu} &= \operatorname{conv}\left(\sigma^{\mathcal{S}}\left[X_{[d]} = x_{[d]}, L\right] \,:\, x_{[d]} \in (\sigma^{\mathcal{S}})^{-1}(Q^{\mathcal{I}}_{\mathcal{S},\nu}),\, \nu\left[X_{[d]} = x_{[d]}\right] = 1\right) \\ &= \operatorname{conv}\left(\sigma^{\mathcal{S}}\left[X_{[d]} = x_{[d]}, L\right] \,:\, x_{[d]} \,:\, \tilde{\nu}\left[X_{[d]} = x_{[d]}\right] = 1\right) \\ &= \mathcal{M}_{\mathcal{S},\tilde{\nu}} . \end{aligned} \qquad \square$$

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 3. *If and only if for a distribution $\mathbb{P}\left[X_{[d]}\right]$ and a face $\mathcal{I}$ we have*

$$\left\langle \mathbb{P}\left[X_{[d]}\right], \sigma^{\mathcal{S}}\left[X_{[d]}, L\right] \right\rangle [L] \in Q^{\mathcal{I}}_{\mathcal{S},\nu} ,$$

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

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

Proof. We have

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

Now, the $x_{[d]}$ with $\nu^{\mathcal{S},\mathcal{I}}\left[X_{[d]} = x_{[d]}\right] = 1$ are exactly those, for which the conditions $\mathcal{I}$ hold straight. If and only if for a $x_{[d]}$ with $\nu^{\mathcal{S},\mathcal{I}}\left[X_{[d]} = x_{[d]}\right] = 0$ we have $\mathbb{P}\left[X_{[d]} = x_{[d]}\right] > 0$, one of the conditions $\mathcal{I}$ would not hold straight. Thus, if and only if $\mathbb{P}\left[X_{[d]}\right]$ is representable with respect to $\nu^{\mathcal{S},\mathcal{I}}\left[X_{[d]}\right]$, we have $\mu\left[L\right] \in Q^{\mathcal{I}}_{\mathcal{S},\nu}$. □

For members of exponential families, we can make a stronger statement than Lem. 3. If for any $\mathbb{P}\left[X_{[d]}\right] \in \Gamma^{\mathcal{S},\nu}$ and a face $\mathcal{I}$ we have $\left\langle \mathbb{P}\left[X_{[d]}\right], \sigma^{\mathcal{S}}\left[X_{[d]}, L\right]\right\rangle\left[L\right] \in \left(Q^{\mathcal{I}}_{\mathcal{S},\nu}\right)^{\circ}$ then $\mathbb{P}\left[X_{[d]}\right]$ is positive with respect to the base measure

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

Let us now investigate tensor network representations of face measures, based on the basis encoding $\beta^{\mathcal{S}}$ of a statistic. Vertices of $\mathcal{M}_{\mathcal{S},\nu}$ are faces with single elements, that is $\{\mu\left[L\right]\}$. By Lem. 1 there must be μ must lie in the image of $\sigma^{\mathcal{S}}$, since otherwise $\mathcal{M}_{\mathcal{S},\nu}$ would be empty. The vertex measure is then

$$\nu^{\mathcal{S},\mathcal{I}}\left[X_{[d]}\right] = \left\langle \beta^{\mathcal{S}}\left[Y_{[p]}, X_{[d]}\right], \epsilon_{\mu}\left[Y_{[p]}\right]\right\rangle\left[X_{[d]}\right]$$

Here we use that each $\mu \in Q^{\mathcal{I}}_{\mathcal{S},\nu} \cap \operatorname{im}\left(\sigma^{\mathcal{S}}\right)$ has integer-valued coordinates and denote

$$\epsilon_{\mu}\left[Y_{[p]}\right] = \bigotimes_{l \in [p]} \epsilon_{\mu[L=l]}\left[Y_l\right].$$

Theorem 2 (Face measure representation)**.** *For any face $Q^{\mathcal{I}}_{\mathcal{S},\nu}$ of $\mathcal{M}$ we have*

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

where

$$\kappa^{\mathcal{I}}\left[Y_{[p]}\right] = \sum_{\mu \in Q^{\mathcal{I}}_{\mathcal{S},\nu} \cap \operatorname{im}(\sigma^{\mathcal{S}})} \epsilon_{\mu}\left[Y_{[p]}\right].$$

Proof. For any $\mu \in Q^{\mathcal{I}}_{\mathcal{S},\nu} \cap \operatorname{im}\left(\sigma^{\mathcal{S}}\right)$ the tensor

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

is the indicator of the preimage of μ under $\sigma^{\mathcal{S}}$. Since preimages the elements in $Q^{\mathcal{I}}_{\mathcal{S},\nu} \cap \operatorname{im}\left(\sigma^{\mathcal{S}}\right)$ are disjoint, the support of $\tau^{\mu}\left[X_{[d]}\right]$ is disjoint and their sum

$$\sum_{\mu \in Q^{\mathcal{I}}_{\mathcal{S},\nu} \cap \operatorname{im}(\sigma^{\mathcal{S}})} \tau^{\mu}\left[X_{[d]}\right]$$

is the indicator of the preimage of $Q^{\mathcal{I}}_{\mathcal{S},\nu}$ under $\sigma^{\mathcal{S}}$, which is the face measure $\nu^{\mathcal{S},\mathcal{I}}\left[X_{[d]}\right]$. Exploiting linearity of contraction we have

$$\begin{aligned}
\nu^{\mathcal{S},\mathcal{I}}\left[X_{[d]}\right] &= \sum_{\mu \in Q^{\mathcal{I}}_{\mathcal{S},\nu} \cap \operatorname{im}(\sigma^{\mathcal{S}})} \tau^{\mu}\left[X_{[d]}\right] \\
&= \left\langle \beta^{\mathcal{S}}\left[Y_{[p]}, X_{[d]}\right], \sum_{\mu \in Q^{\mathcal{I}}_{\mathcal{S},\nu} \cap \operatorname{im}(\sigma^{\mathcal{S}})} \epsilon_{\mu}\left[Y_{[p]}\right]\right\rangle\left[X_{[d]}\right] \\
&= \left\langle \beta^{\mathcal{S}}\left[Y_{[p]}, X_{[d]}\right], \kappa^{\mathcal{I}}\left[Y_{[p]}\right]\right\rangle\left[X_{[d]}\right].
\end{aligned}$$

□

Motivated from the face measure representation, we define a CP rank for faces and show that normalized face measures are computable with respect to a corresponding CP format.

Definition 3. *The bas+* CP *rank of a face is*

$$\operatorname{rank}^{\text{bas+}}\left(Q^{\mathcal{I}}_{\mathcal{S},\nu}\right) = \min_{\{y_{[p]} : \epsilon_{y_{[p]}} \in Q^{\mathcal{I}}_{\mathcal{S},\nu}\} \subset \mathcal{U} \subset \epsilon(\times_{l \in [p]}[n_l]), \mathcal{U} \cup \{y_{[p]} : \epsilon_{y_{[p]}} \in \mathcal{M}_{\mathcal{S},\nu}/Q^{\mathcal{I}}_{\mathcal{S},\nu}\} = \varnothing} \operatorname{rank}^{\text{bas+}}\left(\sum_{v \in \mathcal{U}} \epsilon_v\left[Y_{[p]}\right]\right).$$

The face measures are contraction of the vertex subset encodings with the computation. They are Computation-Activation Networks, when choosing the CP graph with rank at least $\mathrm{rank}^{\mathrm{bas+}}\left(Q^{\mathcal{I}}_{\mathcal{S},\nu}\right)$.

Lemma 4. *For each face of the mean polytope we have*

$$\nu^{\mathcal{S},\mathcal{I}}\left[X_{[d]}|\varnothing\right] \in \Lambda^{\mathcal{S},\mathrm{CP}^{\mathrm{rank}^{\mathrm{bas+}}\left(Q^{\mathcal{I}}_{\mathcal{S},\nu}\right)}},$$

where $\mathrm{CP}^{\mathrm{rank}^{\mathrm{bas+}}\left(Q^{\mathcal{I}}_{\mathcal{S},\nu}\right)}$ *is the CP graph with a hidden variable of dimension* $\mathrm{rank}^{\mathrm{bas+}}\left(Q^{\mathcal{I}}_{\mathcal{S},\nu}\right)$.

Proof. We find by definition a set $\mathcal{U}$ of basis vectors containing the vertices of the face $\nu^{\mathcal{S},\mathcal{I}}$ but no further vertices, which has a bas+ CP rank of $\mathrm{rank}^{\mathrm{bas+}}\left(Q^{\mathcal{I}}_{\mathcal{S},\nu}\right)$. We have therefore, that $\left\langle \beta^{\mathcal{S}}\left[Y_{[p]}, X_{[d]}\right], \epsilon_{\mathcal{U}}\left[Y_{[p]}\right]\right\rangle \left[X_{[d]}|\varnothing\right]$ is in $\Lambda^{\mathcal{S},\mathrm{CP}^{\mathrm{rank}^{\mathrm{bas+}}\left(Q^{\mathcal{I}}_{\mathcal{S},\nu}\right)}}$. Further it holds that

$$\begin{aligned}
\left\langle \beta^{\mathcal{S}}\left[Y_{[p]}, X_{[d]}\right], \epsilon_{\mathcal{U}}\left[Y_{[p]}\right]\right\rangle \left[X_{[d]}\right] &= \sum_{y_{[p]}\in\mathcal{U}\,:\,\epsilon_{y_{[p]}}\in Q^{\mathcal{I}}_{\mathcal{S},\nu}} \left\langle \beta^{\mathcal{S}}\left[Y_{[p]}, X_{[d]}\right], \epsilon_{y_{[p]}}\left[Y_{[p]}\right]\right\rangle \left[X_{[d]}\right] \\
&\quad + \sum_{y_{[p]}\in\mathcal{U}\,:\,\epsilon_{y_{[p]}}\notin Q^{\mathcal{I}}_{\mathcal{S},\nu}} \left\langle \beta^{\mathcal{S}}\left[Y_{[p]}, X_{[d]}\right], \epsilon_{v}\left[Y_{[p]}\right]\right\rangle \left[X_{[d]}\right] \\
&= \sum_{y_{[p]}\in\mathcal{U}\,:\,\epsilon_{y_{[p]}}\in Q^{\mathcal{I}}_{\mathcal{S},\nu}} \left\langle \beta^{\mathcal{S}}\left[Y_{[p]}, X_{[d]}\right], \epsilon_{y_{[p]}}\left[Y_{[p]}\right]\right\rangle \left[X_{[d]}\right] \\
&= \nu^{\mathcal{S},\mathcal{I}}\left[X_{[d]}\right] .
\end{aligned}$$

Here we used, that for $y_{[p]} \in \mathcal{U}$ with $\epsilon_{y_{[p]}} \notin Q^{\mathcal{I}}_{\mathcal{S},\nu}$ is not in the image of $\mathcal{S}$ and therefore the contraction of its one-hot encoding with the computation cores vanishes. Thus, $\left\langle \beta^{\mathcal{S}}\left[Y_{[p]}, X_{[d]}\right], \epsilon_{\mathcal{U}}\left[Y_{[p]}\right]\right\rangle \left[X_{[d]}|\varnothing\right]$ coincides with the normalized face measure, which is therefore in $\Lambda^{\mathcal{S},\mathrm{CP}^{\mathrm{rank}^{\mathrm{bas+}}\left(Q^{\mathcal{I}}_{\mathcal{S},\nu}\right)}}$. □

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

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

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

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

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

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

Extending Example 1, we can provide a coarse estimation of the hypergraph $\mathcal{G}$ required to decompose $\kappa^{\mathcal{I}}$ for generic faces $Q^{\mathcal{I}}_{\mathcal{S},\nu}$. We notice that $\kappa^{\mathcal{I}}\left[Y_{[p]}\right]$ in Thm. 2 is a sparse tensor with basis CP rank $\left|Q^{\mathcal{I}}_{\mathcal{S},\nu} \cap \mathrm{im}\left(\sigma^{\mathcal{S}}\right)\right|$ (see Chapter **??**).

$$\ell_0\left(\kappa^{\mathcal{I}}\left[Y_{[p]}\right]\right) = \left|\sigma^{\mathcal{S}}\left(\nu^{\mathcal{S},\mathcal{I}}\right)\right|$$

where $\left|\sigma^{\mathcal{S}}\left(\nu^{\mathcal{S},\mathcal{I}}\right)\right|$ is the number of different statistic encoding vectors to the support of the face measure $\nu^{\mathcal{S},\mathcal{I}}$. Using the formalism of sparse CP decompositions Chapter **??**, this characterize the basis CP rank of $\kappa^{\mathcal{I}}\left[Y_{[p]}\right]$. However, the basis CP rank is only an upper bound to generic CP rank, which can be loose. By Example 2 we provide with the maximal face an example, where the basis CP rank is given by the tensor space dimension, whereas the generic CP rank is one and the normalized face measure is thus still in $\Lambda^{\mathcal{S},\mathrm{EL}}$.

Example 2 (Maximal face)**.** *The maximal face $Q^{\varnothing}_{\mathcal{S},\nu} = \mathcal{M}_{\mathcal{S},\nu}$ coincides with the mean parameter itself. 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 $\nu^{\mathcal{S},\varnothing}$ to the maximal face is in $\Lambda^{\mathcal{S},\mathrm{EL}}$.

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

Definition 4 (Effective Interior)**.** *Let $\mathcal{U} \subset \mathbb{R}^p$ be an arbitrary set and $\mathcal{L}$ the minimal affine subspace of $\mathbb{R}^p$ containing $\mathcal{U}$. Then the effective interior, denoted $(\mathcal{U})^\circ$ is the the interior of $\mathcal{U}$ in the space $\mathcal{L}$.*

Lemma 5. *Any polytope is a disjoint union of the effective interiors of its faces, that is*

$$\mathcal{M}_{\mathcal{S},\nu} = \dot{\bigcup_{\mathcal{I}\subset[n]}} \left(Q^{\mathcal{I}}_{\mathcal{S},\nu}\right)^\circ .$$

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

In this way, we find to each $\mu \in \mathcal{M}_{\mathcal{S},\nu}$ a unique exponential family with statistics $\mathcal{S}$ and base measure by a face measure, such that μ is reproduced by an element of that exponential family. We will show in Chapter **??**, that these reproducing distributions maximize the entropy among any other reproducing distribution.

6 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.

6.1 Main result

Theorem 3 (Generic characterization of Maximum Entropy Solutions)**.** *Let $\mathcal{S}$ be a statistic and ν a base measure. For any $\mu\,[L]$ the maximum entropy problem has a feasible distribution, if and only if $\mu\,[L] \in \mathcal{M}_{\mathcal{S},\nu}$. In case $\mu\,[L] \in \mathcal{M}_{\mathcal{S},\nu}$ there is a unique face $Q^{\mathcal{I}}_{\mathcal{S},\nu}$ such that μ is in the effective interior of $Q^{\mathcal{I}}_{\mathcal{S},\nu}$. Then the solution of the maximum entropy problem is the member*

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

of the exponential family $\Gamma^{\mathcal{S},\nu^{\mathcal{S},\mathcal{I}}}$, where $\nu^{\mathcal{S},\mathcal{I}}$ is the . If ν is an elementary Computation-Activation Network, then $\mathbb{P}^{(\mathcal{S},B^{\mathcal{S},\nu^{\mathcal{S},\mathcal{I}}}(\mu),\nu^{\mathcal{S},\mathcal{I}})}$ is a Computation-Activation Network with respect to the CP graph of rank $\operatorname{rank}\left(Q^{\mathcal{I}}_{\mathcal{S},\nu}\right)$.

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

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

- By Lem. 5 for any μ we find a unique face $Q^{\mathcal{I}}_{\mathcal{S},\nu}$.
- By Lem. 3 all feasible distributions are representable by the with the face measure refined base measure. The maximum entropy solution is thus the same as for the $(\mathcal{S}, \mu, \left\langle \nu, \nu^{\mathcal{S},\mathcal{I}} \right\rangle \left[X_{[d]}\right])$ instance, which we characterize in the following.
- By Lem. 2 the face $Q^{\mathcal{I}}_{\mathcal{S},\nu}$ coincides with the polytope $\mathcal{M}_{\mathcal{S},\langle\nu,\nu^{\mathcal{S},\mathcal{I}}\rangle\left[X_{[d]}\right]}$ and in particular μ is in the effective interior of that polytope.
- We can now apply Thm. 4 and get a characterization of the maximum entropy solution as a member of the exponential family.

Representation Claim: By Thm. 2 we can represent the face measure as a CP Computation-Activation Network. Since both the ν (by assumption) and the soft activation (always) is elementary, they do not change the CP rank when contracting to the face activating tensor of minimal rank. □

6.2 Maximum entropy on the interior

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

Theorem 4. *If and only if μ^* is in the effective interior of $\mathcal{M}_{\mathcal{S},\nu}$, then the unique solution of the maximum entropy problem is the distribution*

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

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

Proof. By The 3.3 in Wainwright and Jordan, since by assumption

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

there is a canonical parameter θ with

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

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

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

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

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

Therefore, if $\tilde{\mathbb{P}}$ does not coincide with$\mathbb{P}^{(\mathcal{S},\hat{\theta},\nu)}$, it is not a maximum entropy distribution. □

Exponential families are in $\Lambda^{\mathcal{S},\mathrm{EL}}$, if and only if $\langle\nu\rangle\left[X_{[d]}|\varnothing\right] \in \Lambda^{\mathcal{S},\mathrm{EL}}$. If $\langle\nu\rangle\left[X_{[d]}|\varnothing\right] \in \Lambda^{\mathcal{S},\mathrm{EL}}$ and $\mu\left[L\right] \in \left(\mathcal{M}_{\mathcal{S},\nu}\right)^{\circ}$ we therefore have a sparse representation of the maximum entropy distribution with elementary activation tensors.

6.3 Mean parameter on faces

We always find a unique face of the polytope with the mean parameter being in the interior (see Lem. 5). Any distribution reproducing the mean parameter is realizable with respect to the face measure of that face (see Lem. **??**). We conclude that the maximum entropy distribution of $\mu^*[L]$ with respect to $\mathcal{S}, \nu$ is also the maximum entropy distribution

Theorem 5. *Given $\mathcal{S}$ and $\mu\left[L\right] \in \mathcal{M}_{\mathcal{S},\nu}$, let $\mathcal{I}$ be the smallest face of $\mathcal{M}_{\mathcal{S},\nu}$ such that*

$$\mu\left[L\right] \in Q^{\mathcal{I}}_{\mathcal{S},\nu}\,.$$

Then the corresponding maximum entropy distribution is in $\Lambda^{\mathcal{S},\mathcal{G},\nu}$ if and only if the face measure (see Def. 2)

$$\kappa^{\mathcal{I}}\left[Y_{[p]}\right] = \sum_{\mu \in Q^{\mathcal{I}}_{\mathcal{S},\nu} \cap \mathrm{im}(\sigma^{\mathcal{S}})} \epsilon_{\mu}\left[Y_{[p]}\right]$$

is in $\mathcal{T}^{\mathcal{G}}$.

Proof. By Thm. **??** the maximum entropy distribution is an element of the exponential family with by the face measure refined base measure $\tilde{\nu}$. Let $\theta\left[L\right]$ be a canonical parameter such that

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

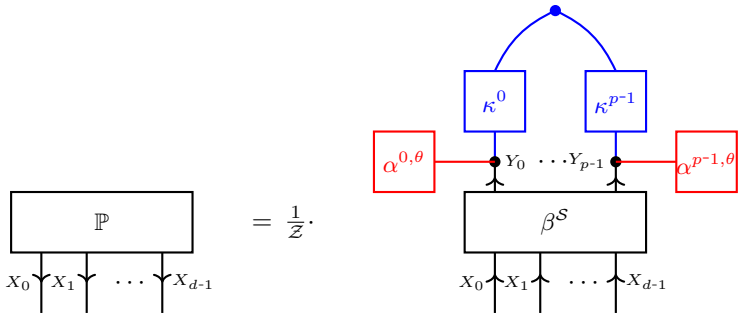


Figure 1: Tensor network decomposition of maximum entropy distributions to the constraint $\mu\left[L\right] = \left\langle \mathbb{P}, \sigma^{\mathcal{S}} \right\rangle \left[L\right]$. Blue: Constraint activation cores κ^l in a CP decomposition, representing the face measure to the minimal face, such that $\mu \in Q^{\mathcal{I}}_{\mathcal{S},\nu}$. Red: Probabilistic activation cores $\alpha^{l,\theta}\left[Y_l\right]$ in an elementary decomposition, where each leg core is a scaled exponentials evaluated on the enumerated image $\mathrm{im}\left(s_l\right)$.

that is $\mathbb{P}^{\mathcal{S},\theta,\tilde{\nu}}\left[X_{[d]}\right]$ is the maximum entropy distribution. We apply Thm. 2 to represent the face measure by

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

Then for the tensor

$$\tau\left[Y_{[p]}\right] = \left\langle \{\alpha^{l,\theta}\left[Y_l\right] : l \in [p]\} \cup \{\kappa^{\mathcal{I}}\left[Y_{[p]}\right]\} \right\rangle \left[Y_{[p]}\right]$$

we have

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

Thus, the maximum entropy distribution is in $\Lambda^{\mathcal{S},\mathcal{G},\nu}$, if τ admits a tensor network decomposition with respect to $\mathcal{G}$. Since the hard activation cores are elementary, this is the case when $\kappa^{\mathcal{I}}$ admits a tensor network decomposition with respect to $\mathcal{G}$. □

7 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 $\mathcal{F} : \bigtimes_{k\in[d]}[m_k] \to \bigtimes_{l\in[p]}[2]$ the mean polytope is a subset of the cube $[0,1]^p$. In this case, any boolean vector in $\mathcal{M}_{\mathcal{F},\nu}$ is a vertex. It follows, that any distribution reproducing a mean parameter $\mu\left[L\right]$ on the effective interior of $\mathcal{M}_{\mathcal{F},\nu}$ 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 $\Lambda^{\mathcal{S},\mathrm{EL}}$, if and only if the face measure is in $\Lambda^{\mathcal{S},\mathrm{EL}}$. This is exactly the case, when the face is an intersection of the mean polytope with a face of the cupe $[0,1]^p$.

7.1 Set of maximum entropy distributions

Theorem 6. *Any distribution in $\Lambda^{\mathcal{F},\mathrm{EL}}$ is a maximum entropy distribution. Any maximum entropy distribution is realized by $\Lambda^{\mathcal{F},\mathrm{EL}}$ if and only if the mean parameter is on an effective 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. □

The mean parameters, which can be realized by a distribution in $\Lambda^{\mathcal{F},\mathrm{EL}}$ are those, which are on the effective interior of the intersection of the mean polytope with a face of the cube.

7.2 Example

Add examples: Partition Statistics - Simplices, (Logically) Independent formulas - Hypercubes

Example 3 (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

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

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

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

This polytope has a non cube-like face (sketched blue in Figure 2), 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

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

and has the face measure

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

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

$$\mu_\lambda\left[L\right] = \begin{bmatrix} \lambda & \lambda \end{bmatrix}^T .$$

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

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

we get the correspondence by the sigmoid

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

Note, that the hard activation core $\kappa^{(0,0),(1,1)}\left[Y_0, Y_1\right]$ 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

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

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.

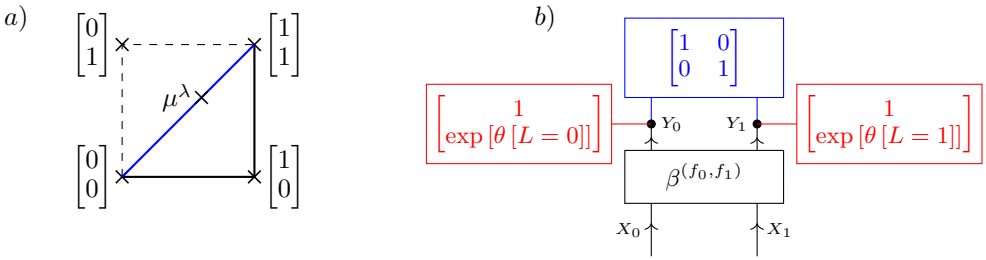


Figure 2: a) Mean polytope of the statistic $\mathcal{F} = (X_0 \wedge X_1, X_0 \Rightarrow X_1)$ (thick), as a subset of the cube $[0,1]^2$ (dashed). The blue line is the face of the polytope, which is not cube like, that is not an intersection of the polytope with the faces of the polytope. We further 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 $\theta [L = 0] + \theta [L = 1]$.

8 Generic Base Measures

The definition of the Shannon entropy is dependent on the chosen base measure. We now allow for generic non-Boolean base measures $\nu \left[X_{[d]}\right]$, which are non-negative and non-vanishing tensors. We define

- The set of by ν representable distributions

$$\Gamma^\nu = \{\mathbb{P}\left[X_{[d]}\right] : \forall x_{[d]} \in \bigtimes_{k\in[d]} [m_k] : \mathbb{P}\left[X_{[d]} = x_{[d]}\right] > 0, \left\langle \mathbb{P}\left[X_{[d]}\right], \nu\left[X_{[d]}\right]\right\rangle [\varnothing] = 1, \left\langle \mathbb{P}\left[X_{[d]}\right], \mathbb{I}\left[X_{[d]}\right] - \nu\left[X_{[d]}\right]\right\rangle [\varnothing]$$

- The entropy of $\mathbb{P}\left[X_{[d]}\right] \in \Gamma^\nu$ by

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

- The mean parameter of $\mathbb{P}\left[X_{[d]}\right] \in \Gamma^\nu$ with respect to the statistic $\mathcal{S}$ by

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

- Polytope of mean parameters by

$$\mathcal{M}_{\mathcal{S},\nu} = \text{conv}\left(\sigma^{\mathcal{S}} X_{[d]} = x_{[d]}, L : \nu\left[X_{[d]} = x_{[d]}\right] \neq 0\right)$$

- The maximum entropy problem by

$$\text{argmax}_{\mathbb{P}\in\Gamma^\nu} \mathbb{H}^\nu[\mathbb{P}] \quad \text{subject to} \quad \mu[L] = \left\langle \mathbb{P}\left[X_{[d]}\right], \sigma^{\mathcal{S}}\left[X_{[d]}, L\right], \nu\left[X_{[d]}\right]\right\rangle [L]$$

The main theorem of this work generalizes to this situation, with minor changes in the proofs.

9 Mean as a Statistic

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

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

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

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

Theorem 7. *The mean statistic is sufficient for the family of maximum entropy distributions* $(\mathcal{S}, \nu)$.

Proof. It sufficies to show that the likelihood is a function of μ_D. Let us choose a face $\mathcal{I}$ of $\mathcal{M}_{\mathcal{S},\nu}$, 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

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

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

References

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 Polytopes*. Springer, 1995th edition edition. ISBN 978-0-387-94365-7.