MAXIMUM ENTROPY DISTRIBUTIONS AS COMPUTATION ACTIVATION NETWORKS

RESEARCH NOTES IN THE ENEXA PROJECT

July 31, 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.

1 Motivation

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.

2 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

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

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 $\Lambda^{\mathcal{S},\mathcal{G}^{\max}}$, where $\mathcal{G}^{\max}$ is the maximal hypergraph $\mathcal{G}^{\max} = ([p], \{[p]\})$. 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 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 The mean polytope

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

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

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

4.1 Convex hull and faces

The mean polytope is the convex hull

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

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.

Faces of the mean polytope are itself mean polytopes with respect to refined base measures.

4.2 Maximum entropy on the interior

A classical result states, that the maximum entropy distribution is in the exponential family $\Gamma^{\mathcal{S},\nu}$.

Theorem 2. *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 [Wainright and Jordan, 2008], since by assumption

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

there is a canonical parameter θ with

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

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

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

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 $\mu\,[L] \in (\mathcal{M}_{\mathcal{S},\nu})^{\circ}$ we therefore have a sparse representation of the maximum entropy distribution with elementary activation tensors.

4.3 Mean parameter on faces

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

5 Characterization for boolean statistics

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 $\mu\,[L]$ 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 $[0,1]^p$.

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.