

---

# CHARACTERIZATION OF COMPUTATION-ACTIVATION NETWORKS BY SUFFICIENT STATISTICS

---

RESEARCH NOTES IN THE ENEXA AND QROM PROJECTS

November 15, 2025

## Contents

<b>1</b>	<b>Foundations</b>	<b>1</b>
1.1	Information Theory [Cover, Thomas - Section 2.10] . . . . .	1
1.2	Mathematical Statistic [Hogg - Chapter 2] . . . . .	2
<b>2</b>	<b>The Computation Mechanism of Tensor Network Decompositions</b>	<b>2</b>
<b>3</b>	<b>Sufficient Statistic for Parametrized Families</b>	<b>3</b>
<b>4</b>	<b>Sufficient Statistic for the Probability</b>	<b>5</b>
<b>5</b>	<b>Minimal sufficient statistics</b>	<b>5</b>

## 1 Foundations

### 1.1 Information Theory [Cover, Thomas - Section 2.10]

Consider two variables  $Z$  and  $X$  with a joint distribution  $\mathbb{P}[Z, X]$ , and a function  $T$  on the states of  $X$ . We augment this joint distribution by a variable  $Y_T$ , which is the head variable to the function  $T$

$$\mathbb{P}[Z, X, Y_T] = \langle \mathbb{P}[Z, X], \beta^T[Y_T, X] \rangle [Z, X, Y_T]$$

Then we have

$$(Y_T \perp Z) | X$$

since

$$\mathbb{P}[Y_T | Z, X] = \beta^T[Y_T, X] \otimes \mathbb{I}[Z] .$$

Thus, the variables are a Markov Chain  $Z \rightarrow X \rightarrow Y$ .

**Definition 1.** We call  $T$  sufficient statistic of  $Z$ , if and only if

$$I(Z; X) = I(Z; T(X)) .$$

**Lemma 1.** If there is a function  $Q$  such that

$$\mathbb{P}[Z, X] = \langle \mathbb{P}[X], \beta^Q[Z, X] \rangle [Z, X] ,$$

and  $T$  is sufficient for  $Z$ , then there is a function  $R$  such that

$$Q = R \circ T .$$

*Proof.* Since  $Z$  has a deterministic dependence on  $X$  we have  $\mathbb{H}[Z|X] = 0$  and by the sufficient statistic assumption (using that  $I(X; Y_T) = H(Y_T) - H(X|Y_T)$ ) we have

$$\mathbb{H}[Z|Y_T] = \mathbb{H}[Z|X] = 0.$$

Now,  $\mathbb{H}[Z|Y_T]$  is equal to the existence of a function  $R$  mapping the states of  $Y$  to  $Z$ , such that for any state  $y$

$$\mathbb{P}[Z|Y_T = y] = \epsilon_{R(y)}[Z].$$

Since  $Y$  itself is computable by  $X$  with the function  $T$ , and  $Z$  with  $Q$ , we have

$$Q = R \circ T.$$

□

This lemma is applied when characterizing sufficient statistics for  $Z = \mathbb{P}[X]$ .

## 1.2 Mathematical Statistic [Hogg - Chapter 2]

In mathematical statistic, sufficient statistics are used to characterize parameter estimation problems, i.e. where  $Z$  is a parameter variable  $\Theta$  of a parametrized family. The joint distribution of  $\Theta$  and  $X$  is constructed by drawing the parameter variable  $\Theta$  first with outcome  $\theta$  and then drawing  $X$  from  $\mathbb{P}^\theta$ .

## 2 The Computation Mechanism of Tensor Network Decompositions

Sufficient statistics imply tensor network decompositions of joint distributions using basis encodings of them. The basis encoding of the sufficient statistics computes the sufficient statistic in the basis calculus scheme. We thus call this decomposition mechanism the computation mechanism.

**Theorem 1** (Factorization Theorem of Fisher and Neyman). *Let  $\mathbb{P}$  be a joint distribution of variables  $Z, X$  with values  $\text{val}(Z)$ ,  $\text{val}(X)$  and let  $T(X)$  be a statistic. The following are equivalent:*

i) *The Data Processing Inequality holds straight, i.e.*

$$I(Z; X) = I(Z; Y_T).$$

ii)  *$Z \rightarrow Y_T \rightarrow X$  is a Markov Chain, i.e.*

$$(Z \perp X) | Y_T$$

iii) *There are functions  $g : \text{im}(T) \times \text{val}(Z) \rightarrow \mathbb{R}$  and  $h : \text{val}(X) \rightarrow \mathbb{R}$  such that for any  $(x, z) \in \text{val}(Z) \times \text{val}(X)$*

$$\mathbb{P}[Z = z, X = x] = g(T(x), z) \cdot h(x).$$

*Proof.* i)  $\Leftrightarrow$  ii): We have always

$$I(Z; X) = I(Z; X, Y_T) = I(Z; Y_T) + I(Z; X|Y_T)$$

and thus if and only if i) holds

$$I(Z; X|Y_T) = 0.$$

Using the KL-divergence characterization of the mutual information, this is equal to

$$\mathbb{P}[Z, X|Y_T] = \langle \mathbb{P}[Z|Y_T], \mathbb{P}[X|Y_T] \rangle [Z, X, Y_T].$$

This is equivalent to the conditional independence statement ii).

ii)  $\Rightarrow$  iii): For all  $z \in \text{val}(Z)$  and  $x \in \text{val}(X)$  we have

$$\begin{aligned} \mathbb{P}[Z = z|X = x] &= \mathbb{P}[Z = z|X = x, Y_T = T(x)] \\ &= \mathbb{P}[Z = z|Y_T = T(x)] \end{aligned}$$

Here we used that  $Y_T$  has a deterministic dependence on  $X$  and ii). There is thus a function  $g$  such that for all  $z \in \text{val}(Z)$  and  $x \in \text{val}(X)$

$$g(T(x), z) = \mathbb{P}[Z = z|X = x].$$

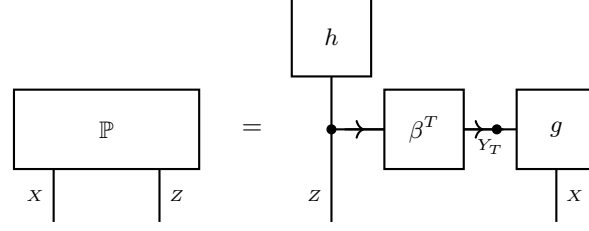


Figure 1: Sketch of the computation decomposition of a joint distribution of  $X, Z$  given a sufficient statistic  $T$ . This decomposition follows from the Fisher-Neyman factorization Thm. 1.

We further define a function  $h(x) = \mathbb{P}[X = x]$  and get

$$\begin{aligned} \mathbb{P}[Z = z, X = x] &= \mathbb{P}[X = x] \cdot \mathbb{P}[Z = z | X = x] \\ &= g(T(x), z) \cdot h(x). \end{aligned}$$

*iii)  $\Rightarrow$  ii):* Using *iii)* we have for all supported  $(x, z) \in \text{val}(Z) \times \text{val}(X)$

$$\begin{aligned} \mathbb{P}[Z = z | X = x] &= \frac{\mathbb{P}[Z = z, X = x]}{\mathbb{P}[X = x]} \\ &= \frac{g(T(x), z) \cdot h(x)}{\int g(T(x), z) \cdot h(x) dz} \\ &= \frac{g(T(x), z)}{\int g(T(x), z) dz} \\ &= \frac{\left( \int_{\tilde{x}: T(\tilde{x})=T(\tilde{x})} h(x) dx \right) \cdot g(T(x), z)}{\left( \int_{\{\tilde{x}: T(\tilde{x})=T(\tilde{x})\}} h(x) dx \right) \cdot \int g(T(x), z) dz} \\ &= \frac{\mathbb{P}[Z = z, Y_T = T(x)]}{\mathbb{P}[Y_T = T(x)]} \\ &= \mathbb{P}[Z = z | Y_T = T(x)] \end{aligned}$$

We have at almost all  $y \in \text{val}(Y_T)$ ,  $z \in \text{val}(Z)$  and  $x \in \text{val}(X)$  that  $y = T(x)$  and

$$\mathbb{P}[Z = z | X = x, Y_T = y] = \mathbb{P}[Z = z | X = x]$$

and with the above at thus at almost all such pairs

$$\mathbb{P}[Z = z | X = x, Y_T = y] = \mathbb{P}[Z = z | Y_T = y].$$

This is equivalent to *ii)*. □

Thm. 1 thus states, that whenever a sufficient statistic  $T$  of  $X$  exists for a variable  $Z$ , then the joint distribution of  $X$  and  $Z$  decomposes as sketched in Figure 1.

### 3 Sufficient Statistic for Parametrized Families

Sufficient statistics are treated in mathematical statistics and in information theory. We here choose a definition of information theory and apply a factorization theorem of mathematical statistics to relate with Computation-Activation Networks. The distribution of a canonical parameter is now drawn from a (possibly continuous) random variable  $\Theta$ , which takes values  $\theta \in \Gamma$  with probability

$$\tilde{\mathbb{P}}[\Theta = \theta].$$

**Definition 2** (Sufficient statistics for Parameters). *Let  $\{\mathbb{P}^\theta[X_{[d]}] : \theta \in \Gamma\}$  be a family of probability distributions and*

$$\mathcal{S} : \prod_{k \in [d]} [m_k] \rightarrow \prod_{l \in [p]} [p_l]$$

be a function. We say that  $\mathcal{S}$  is sufficient for  $\Theta$ , if for any distribution  $\tilde{\mathbb{P}}[\Theta]$  of  $\Theta$ , when drawing  $X_{[d]}$  from  $\mathbb{P}^\theta[X_{[d]}]$  with probability  $\tilde{\mathbb{P}}[\Theta = \theta]$ , we have that

$$(\Theta \perp X_{[d]}) \mid \mathcal{S}(X_{[d]}).$$

We can characterize Computation-Activation Networks with arbitrary base measures based on sufficient statistics.

**Theorem 2** (Characterization of Computation-Activation Networks). *Let  $\{\mathbb{P}^\theta[X_{[d]}] : \theta \in \Gamma\}$  be a family of probability distributions with a sufficient statistic  $\mathcal{S}$ . Then there is a non-negative (possibly non-Boolean) base measure  $\nu[X_{[d]}]$  and a map*

$$h : \Gamma \rightarrow \bigotimes_{l \in [p]} \mathbb{R}^{p_l}$$

such that for all  $\theta \in \Gamma$

$$\mathbb{P}^\theta[X_{[d]}] = \langle h(\Gamma)[Y_{[p]}], \beta^{\mathcal{S}}[Y_{[p]}, X_{[d]}], \nu[X_{[d]}] \rangle [X_{[d]} | \emptyset].$$

We further have that for a set  $\{\mathbb{P}^\theta[X_{[d]}] : \theta \in \Gamma\}$   $\mathcal{S}$  is a sufficient statistic, if and only if there is a non-negative (possibly non-Boolean) base measure  $\nu[X_{[d]}]$  with

$$\{\mathbb{P}^\theta[X_{[d]}] : \theta \in \Gamma\} \subset \Lambda^{\mathcal{S}, \text{MAX}, \nu}.$$

*Proof.* By the Fisher-Neyman Factorization Thm. 1 we have that  $\mathcal{S}$  is a sufficient statistic if and only if there are real-valued functions  $g$  on  $(\times_{l \in [p]} [p_l]) \times \Gamma$  and  $h$  on  $\times_{k \in [d]} [m_k]$  such that

$$\mathbb{P}^\theta[X_{[d]} = x_{[d]}] = g(\mathcal{S}(x_{[d]}), \Gamma) \cdot h(x_{[d]}). \quad (1)$$

We define a base measure by the coordinate encoding of  $h$  by

$$\nu[X_{[d]}] = \sum_{x_{[d]} \in \times_{k \in [d]} [m_k]} h(x_{[d]}) \epsilon_{x_{[d]}}[X_{[d]}]$$

and for each  $\theta \in \Gamma$  an activation tensor

$$\xi^\theta[Y_{[p]}] = \sum_{y_{[p]}} g(y_{[p]}, \theta) \epsilon_{y_{[p]}}[Y_{[p]}].$$

With this we have for any  $\theta \in \Gamma$

$$\langle h(\Gamma)[Y_{[p]}], \beta^{\mathcal{S}}[Y_{[p]}, X_{[d]}], \nu[X_{[d]}] \rangle [\emptyset] = 1$$

and thus for any  $x_{[d]} \in \times_{k \in [d]} [m_k]$  applying basis calculus

$$\begin{aligned} \langle h(\Gamma)[Y_{[p]}], \beta^{\mathcal{S}}[Y_{[p]}, X_{[d]}], \nu[X_{[d]}] \rangle [X_{[d]} = x_{[d]} | \emptyset] &= h(\Gamma)[Y_{[p]} = \mathcal{S}(x_{[d]})] \cdot \nu[X_{[d]} = x_{[d]}] \\ &= g(\mathcal{S}(x_{[d]}), \Gamma) \cdot h(x_{[d]}) \\ &= \mathbb{P}^\theta[X_{[d]} = x_{[d]}]. \end{aligned}$$

We therefore find for any  $\mathbb{P}^\theta[X_{[d]}]$  a representation as a Computation-Activation Network in  $\Lambda^{\mathcal{S}, \text{MAX}, \nu}$  with the activation tensor  $h(\Gamma)[Y_{[p]}]$ .

To show the second claim, we are left to show that any set of Computation-Activation Networks in  $\Lambda^{\mathcal{S}, \text{MAX}, \nu}$  has  $\mathcal{S}$  as a sufficient statistic. Let us thus consider a parametric family

$$\{\mathbb{P}^\theta[X_{[d]}] : \theta \in \Gamma\} \subset \Lambda^{\mathcal{S}, \text{MAX}, \nu}.$$

By this inclusion we find for any  $\theta \in \Gamma$  an activation core  $\alpha^\theta[Y_{[p]}]$ . We then construct functions  $g$  and  $h$  by

$$g(y_{[p]}, \Gamma) = \alpha^\theta[Y_{[p]} = y_{[p]}] \quad \text{and} \quad h(x_{[d]}) = \nu[X_{[d]} = x_{[d]}]$$

and notice that the equivalent condition (1) to  $\mathcal{S}$  being a sufficient statistic is satisfied.  $\square$

## 4 Sufficient Statistic for the Probability

We here consider sufficient statistics for the parameter of a parametrized family, while in the report we considered sufficient statistics for the probability mass as a random variable. In both cases this results from the information theoretic viewpoint, that a function  $T$  of  $X$  is a sufficient statistic for a variable  $Z$ , if

$$(Z \perp X) | T(X).$$

While we choose for  $Z$   $Y_\theta$  above, we now choose for  $Z$  the variable  $Y_{\mathbb{P}}$ . This variable can be computed by contraction with

$$\beta^{\mathbb{P}} [Y_{\mathbb{P}}, X_{[d]}].$$

If  $T$  is a sufficient statistic for  $Y_{\mathbb{P}}$ , we call it probability sufficient for  $\mathbb{P}$ .

**Theorem 3** (Theorem 2.19 in the report). *If and only if a statistic  $\mathcal{S}$  is probability sufficient for  $\mathbb{P} [X_{[d]}]$ , then*

$$\mathbb{P} [X_{[d]}] \in \Lambda^{\mathcal{S}, \text{MAX}, \mathbb{I}}.$$

*Proof.* By Lem. 1 we have a function  $R$  such that for all  $x_{[d]} \in \times_{k \in [d]} [m_k]$

$$\mathbb{P} [X_{[d]} = x_{[d]}] = (R \circ \mathcal{S})(x_{[d]}).$$

By basis calculus it follows that

$$\mathbb{P} [X_{[d]}] = \langle R(I_{\mathcal{S}}[Y_{[p]}]), \beta^{\mathcal{S}} [Y_{[p]}, X_{[d]}] \rangle [X_{[d]}]$$

and thus

$$\mathbb{P} [X_{[d]}] \in \Lambda^{\mathcal{S}, \text{MAX}, \mathbb{I}}. \quad \square$$

Note that by this theorem we can restrict ourselves to the Computation-Activation Networks with trivial base measure for the characterization of distributions with a probability sufficient statistic.

## 5 Minimal sufficient statistics

Minimal sufficient statistics are defined by existences of functions from any sufficient statistics.

**Definition 3.** *A sufficient statistic  $T$  of  $Z$  is minimal, if and only if for any sufficient statistic  $U$  of  $Z$  there is a function  $R$  such that  $T = R \circ U$ .*

Note that by construction, we can choose the same base measure  $h$  when factorizing with respect to different sufficient statistics. The activation cores  $g^{(U)}$  to an arbitrary sufficient statistic  $U$  can thus be further decomposed by the basis encoding of  $R$  and an activation core  $g^{(T)}$  to a minimal sufficient statistic as

$$g^{(U)}[Y_U, Z] = \langle \beta^R [Y_T, Y_U], g^{(T)}[Y_T, Z] \rangle [Y_U, Z].$$

Minimal sufficient statistics thus provide the best embedding into a Computation-Activation Networks, by decomposing the activation tensor into refining Computation-Activation Network.