Characterization of Computation-Activation Networks by sufficient statistics

Research Notes in the ENEXA and QROM Projects

November 14, 2025

Contents

1 Foundations 1

1.1 Information Theory ([Cover, Thomas - Section 2.10]) . . . 1

1.2 Mathematical Statistic (see [Hogg - Chapter 2]) . . . 2

1.3 Equivalent definitions of sufficient statistics . . . 2

2 Sufficient Statistic for Parametrized Families 3

3 Sufficient Statistic for the Probability 4

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

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

Then we have

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

since

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

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

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*

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

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

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

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 Θ of a parametrized family. The joint distribution of Θ and X is constructed by drawing the parameter variable Θ first with outcome θ and then drawing X from $\mathbb{P}^\theta$.

1.3 Equivalent definitions of sufficient statistics

Theorem 1 (Factorization Theorem of Fisher and Neyman)**.** *Let $\mathbb{P}$ be a joint distribution of variables Z, X with values $\mathrm{val}(Z)$, $\mathrm{val}(X)$ 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) $Z \to Y_T \to X$ is a Markov Chain, i.e.

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

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

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

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

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

This is equivalent to the conditional independence statement $ii)$.

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

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

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

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

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

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

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

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

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

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

and with the above at thus at almost all such pairs

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

This is equivalent to $ii)$. □

2 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 Θ, which takes values $\theta \in \Gamma$ with probability

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

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

$$\mathcal{S} \,:\, \bigtimes_{k\in[d]} [m_k] \to \bigtimes_{l\in[p]} [p_l]$$

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

$$\left(\Theta \perp X_{[d]}\right)\big|\,\mathcal{S}\left(X_{[d]}\right)\,.$$

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

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

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

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

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

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

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

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 $\left(\bigtimes_{l\in[p]}[p_l]\right)\times\Gamma$ and h on $\bigtimes_{k\in[d]}[m_k]$ such that

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

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

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

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

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

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

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

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

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

We therefore find for any $\mathbb{P}^{\theta}\left[X_{[d]}\right]$ a representation as a Computation-Activation Network in $\Lambda^{\mathcal{S},\mathrm{MAX},\nu}$ 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 $\Lambda^{\mathcal{S},\mathrm{MAX},\nu}$ has $\mathcal{S}$ as a sufficient statistic. Let us thus consider a parametric family

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

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

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

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

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

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

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 $\mathbb{P}\left[X_{[d]}\right]$, then*

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

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

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

By basis calculus it follows that

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

and thus

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

□

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