Formalizations for the Telephone Theorem and generalized KPD

Leon Lang, Lawrence Chan, Erik Jenner

March 15, 2023

Abstract

In these notes, we clarify the exact setup and proofs of both the telephone theorem and the generalized Koopman-Pitman-Darmois theorem by John Wentworth. For motivation, see our main post, the high-level review of the natural abstractions agenda.

Contents

1 The Telephone Theorem and its Proof

1.1 Setup

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $X_0, X_1, ...$ be a (discrete time) Markov chain defined on Ω, that is, each X_i is a random variable $X_i : \Omega \to \mathbb{R}$ such that

$$P\left(X_i = x_i \mid X_{0:i-1} = x_{0:i-1}\right) = P(X_i = x_i \mid X_{i-1} = x_{i-1})$$

for all $x_{1:i-1}$ with positive support: $P(X_{1:i-1} = x_{1:i-1}) > 0$.

We use $I(X;Y)$ to denote the mutual information between random variables X, Y:

$$I(X;Y) = D_{KL}(P_{(X,Y)} || P_X \otimes P_Y)$$

First, note that the mutual information between X_t and X_0 must converge as $t \to \infty$.

Lemma 1. *As $t \to \infty$, the mutual information $I(X_0; X_t) \to I_\infty$ for some $I_\infty \geq 0$.*

Proof. By the data processing inequality, we have $I(X_0; X_t) \geq I(X_0; X_{t+1})$. Since $I(X_0; X_t) \geq 0$ for all t, the lemma follows by the monotone convergence theorem.

□

1.2 The telephone theorem for finite probability spaces

First, we consider the case where the sample space Ω is finite.

In this case, note that $|X_0(\Omega)| \leq |\Omega| < \infty$, and there are only finitely many values of $I(X_0; X_t)$ (since there are only finitely many partitions of Ω[1]). Combined with Lemma 1, this implies the

[1] Formally, every random variable $X : \Omega \to \mathbb{R}$ induces the partition $\{X^{-1}(x) | x \in X(\Omega)\}$ on Ω. It can be proved that the mutual information of two random variables only depends on the induced partitions on the sample space.

existence of T such that for all $t \geq T$, we have $I(X_t; X_0) = I(X_{t+1}; X_0)$.

We now show that this implies $P(X_0|X_t) = P(X_0|X_{t+1})$ for $t \geq T$ – that is, the conditional probability distributions are equal from then on:

Lemma 2. *For a Markov chain $X_1, X_2, ...$, the following two conditions are equivalent:*

1. $I(X_0; X_t) = I(X_0; X_{t+1})$.
2. $X_0 \perp X_t \mid X_{t+1}$*; that is,* $P(X_0|X_t = x_t) = P(X_0|X_{t+1} = x_{t+1})$ *with probability 1.*[2]

Proof. First, by properties of Markov chains, X_{t+1} and X_0 are independent given X_t and thus $I(X_0; X_t, X_{t+1}) = I(X_0; X_t)$.

The lemma follows from properties of mutual information:

- $(1 \to 2)$ If $I(X_t; X_0) = I(X_{t+1}; X_0)$, then applying the chain rule for mutual information we have

$$\begin{aligned} I(X_0; X_t, X_{t+1}) &= I(X_0; X_{t+1}) + I(X_0; X_t \mid X_{t+1}) \\ &= I(X_0; X_t) + I(X_0; X_t \mid X_{t+1}). \end{aligned}$$

  Yet since $I(X_0; X_t, X_{t+1}) = I(X_t; X_0)$, this implies that $I(X_0; X_t|X_{t+1}) = 0$ and thus $X_0 \perp X_t | X_{t+1}$.

- $(2 \to 1)$ On the other hand, if $X_0 \perp X_t | X_{t+1}$, then together with the original Markov chain we obtain $I(X_0; X_t) = I(X_0; X_t, X_{t+1}) = I(X_0; X_{t+1})$.

□

In order to show the general case, we apply the fact that the likelihood ratio is a minimally sufficient statistic, as shown here:

Lemma 3 (Likelihood ratio is a minimally sufficient statistic)**.** *Let X, Y be two random variables on Ω with codomains $\mathcal{X}$ and $\mathcal{Y}$, where either $\mathcal{X}, \mathcal{Y}$ are countable, or X, Y admit a joint density function. Define $f : \mathcal{Y} \to \Delta(\mathcal{X})$ by the formula*

$$f(y) \coloneqq P(X \mid Y = y).$$

Define $f(Y) \coloneqq f \circ Y : \Omega \to \mathcal{X}$. Then $f(Y)$ is a minimal sufficient statistic of Y for predicting X, that is:

1. $P(X \mid Y) = P(X \mid f(Y)) = P(X \mid P(X \mid Y))$
2. *For any other sufficient statistic $Z : \mathcal{Y} \to \mathcal{Z}$ of Y for predicting X from Y, there exists a function $G : \mathcal{Z} \to \Delta(\mathcal{X})$ with $G(Z(Y)) = f(Y)$.*

Proof. We do the proof in the discrete case. We believe that the proof can be done in the same way for the continuous case, with integrals replacing sums, but have not thought about this in detail.

Let $y \in \mathcal{Y}$ and define $p = P(X \mid Y = y)$. We have the following:

$$\begin{aligned} P(X \mid f(Y) = f(y)) &= P(X \mid f(Y) = p) \\ &= \frac{P(X,\ f(Y) = p)}{P(f(Y) = p)} \\ &= \frac{\sum_{y' : f(y') = p} P(X, y')}{\sum_{y' : f(y') = p} P(y')} \\ &= \frac{\sum_{y' : f(y') = p} P(X \mid y') \cdot P(y')}{\sum_{y' : f(y') = p} P(y')} \\ &= \frac{\sum_{y' : f(y') = p} p \cdot P(y')}{\sum_{y' : f(y') = p} P(y')} \\ &= p \\ &= f(y) \\ &= P(X \mid Y = y). \end{aligned}$$

[2] The second statement follows from the first by applying the Markov chain / independence in both directions: $P(X_0 \mid X_t = x_t) = P(X_0 \mid X_t = x_t, X_{t+1} = x_{t+1}) = P(X_0 \mid X_{t+1} = x_{t+1})$.

For the second statement, notice that

$$P(X \mid Z(Y)) = P(X \mid Y) = f(Y),$$

proving the claim with $G(z) \coloneqq P(X \mid Z(Y) = z)$.

□

1.3 Formal statement and proof

The telephone theorem follows from Lemmas 2 and 3:

Theorem 4 (Telephone theorem for finite probability spaces)**.** *There exist a sequence of functions $f_1, f_2, \ldots$, where $f_i : \mathbb{R} \to \mathbb{R}^{X_0(\Omega)}$, such that:*

1. *there exists a $T \in \mathbb{N}$ such that for all $t \geq T$, $f_t(X_t) = f_{t+1}(X_{t+1})$ with probability 1, and*
2. *for all t, $P(X_0|X_t) = P(X_0|f_t(X_t))$ pointwise on Ω.*

Proof. Let f_t be $x_t \mapsto P(X_0|X_t = x_t)$. That is, f_t the function that maps x_t to a vector in $\mathbb{R}^{X_0(\Omega)}$, where each element is $P(X_0 = x_0^i|X_t = x_t)$.

By Lemma 3 (and by construction), we have $P(X_0|X_t) = P(X_0|f_t(X_t))$ pointwise on Ω.

By Lemmas 1 and 2, there exists $T \in \mathbb{N}$ such that for all $t \geq T$, we have $P(X_0|X_t) = P(X_0|X_{t+1})$ with probability 1. Therefore, for $t \geq T$, we have $f_t(X_t) = f_{t+1}(X_{t+1})$ with probability 1. □

1.4 The telephone theorem for discrete or absolutely continuous Markov chains

Next, we prove the *general* form of the Telephone theorem. A remark on notation: Often, we will write $P(x_0 \mid x_t)$ when we mean, e.g., $P(X_0 = x_0 \mid X_t = x_t)$. This is meant to make the notation less cluttered. When we write $P(X_0 \mid X_t)$, we mean the whole conditional distribution instead of individual values.

Theorem 5 (The telephone theorem)**.** *For any Markov chain $X_0, X_1, \ldots$ that is either discrete or absolutely continuous, there exist a sequence of functions $f_1, f_2, \ldots$, where $f_i : \mathbb{R} \to \mathbb{R}^{X_0(\Omega)}$, such that:*

1. *$f_t(X_t)$ converges in probability to some random variable f_∞, and*
2. *for all t, $P(X_0|X_t) = P(X_0|f_t(X_t))$ pointwise on Ω.*

Proof. Let f_t be $x \mapsto P(X_0|X_t = x)$.

By Lemma 3 and by construction, we have that $P(X_0|X_t) = P(X_0|f_t(X_t))$ pointwise.

By Lemma 1, we have that $I(X_0; X_t)$ converges as $t \to \infty$. This implies that the sequence $I(X_0; X_t)$ is Cauchy: for every $\varepsilon > 0$ there exists a $T \geq 0$ such that for all $t \geq T$ and $k \geq 0$, we have:

$$|I(X_0; X_t) - I(X_0; X_{t+k})| < \varepsilon^5$$

Using the data processing inequality, the Markov property, and properties of mutual information, we have that:

$$\begin{aligned}
|I(X_0; X_t) - I(X_0; X_{t+k})| &= I(X_0; X_t) - I(X_0; X_{t+k}) \\
&= I(X_0; X_t, X_{t+k}) - I(X_0; X_{t+k}) \\
&= I(X_0; X_t|X_{t+k}) \\
&= \int_{\mathcal{X}_{t+k}} D_{\mathrm{KL}}\Big(P(X_0, X_t \mid X_{t+k}) \parallel P(X_0 \mid X_{t+k}) \otimes P(X_t \mid X_{t+k})\Big) dP(X_{t+k}).
\end{aligned}$$

We'll complete the proof assuming that the Markov chain is discrete; the proof in the case where the chain is continuous is similar, but with densities instead of mass functions and integrals instead of sums.

$$\begin{aligned}
&|I(X_0;X_t) - I(X_0; X_{t+k})| \\
&\qquad = \sum_{x \in \mathcal{X}_{t+k}} P(X_{t+k} = x) \cdot D_{\mathrm{KL}}\Big(P(X_0, X_t \mid X_{t+k} = x) \parallel P(X_0 \mid X_{t+k} = x) \otimes P(X_t \mid X_{t+k} = x)\Big) \\
&\qquad < \epsilon^5
\end{aligned}$$

We can split the sum into two cases:

$$S_{bad} = \Big\{ x \in \mathcal{X}_{t+k} \;\Big|\; D_{\mathrm{KL}}\Big(P(X_0, X_t \mid X_{t+k} = x) \;\|\; P(X_0 \mid X_{t+k} = x) \otimes P(X_t \mid X_{t+k} = x) \Big) \geq \varepsilon^4 \Big\}$$

$$S_{good} = \Big\{ x \in \mathcal{X}_{t+k} \;\Big|\; D_{\mathrm{KL}}\Big(P(X_0, X_t | X_{t+k} = x) \;\|\; P(X_0 \mid X_{t+k} = x) \otimes P(X_t \mid X_{t+k} = x) \Big) < \varepsilon^4 \Big\}$$

It follows that $P(S_{bad}) < \varepsilon$. By construction, for $x \in S_{good}$, we have

$$\begin{aligned}
&\delta\Big(P(X_0, X_t \mid X_{t+k} = x),\; P(X_0 \mid X_{t+k} = x) \otimes P(X_t \mid X_{t+k} = x) \Big) \\
&\qquad\qquad < \sqrt{D_{\mathrm{KL}}\Big(P(X_0, X_t \mid X_{t+k} = x) \;\|\; P(X_0 \mid X_{t+k} = x) \otimes P(X_t \mid X_{t+k} = x) \Big)} \\
&\qquad\qquad < \varepsilon^2
\end{aligned}$$

where $\delta(P, Q)$ is the total variation distance between P and Q.

This implies by the correspondence between total variation and L^1 norm that

$$\begin{aligned}
\epsilon^2 &> \delta\Big((P(X_0, X_t \mid X_{t+k} = x),\;\; P(X_0 \mid X_{t+k} = x) \otimes P(X_t \mid X_{t+k} = x) \Big) \\
&= \frac{1}{2} \sum_{x_0, x_t} \Big| P(x_0, x_t \mid X_{t+k} = x) - P(x_0 \mid X_{t+k} = x) \cdot P(x_t \mid X_{t+k} = x) \Big| \\
&= \frac{1}{2} \sum_{x_0, x_t} P(x_t \mid X_{t+k} = x) \cdot \Big| P(x_0 \mid x_t, X_{t+k} = x) - P(x_0 \mid X_{t+k} = x) \Big| \\
&= \frac{1}{2} \sum_{x_t} P(x_t \mid X_{t+k} = x) \sum_{x_0} \Big| P(x_0 \mid x_t) - P(x_0 \mid X_{t+k} = x) \Big|
\end{aligned}$$

since $P(X_0 \mid x_t, X_{t+k} = x) = P(X_0 \mid x_t)$ by the Markov property.

Finally, consider the sets

$$Q_{bad} = \Big\{ (x, x_t) \in S_{good} \times \mathcal{X}_t \;\Big|\; \sum_{x_0} \big| P(x_0 \mid x_t) - P(x_0 \mid X_{t+k} = x) \big| \geq \varepsilon \Big\}$$

$$Q_{good} = \Big\{ (x, x_t) \in S_{good} \times \mathcal{X}_t \;\Big|\; \sum_{x_0} \big| P(x_0 \mid x_t) - P(x_0 \mid X_{t+k} = x) \big| < \varepsilon \Big\}$$

By construction, $P(Q_{bad}) < 2 \cdot \varepsilon$. This gives us, by union bound:

$$\begin{aligned}
P\Big(\Big\| P(X_0 \mid X_t) - P(X_0 \mid X_{t+k}) \Big\|_2 < \varepsilon \Big) &\geq P\Bigg(\sum_{x_0} \big| P(x_0 \mid X_t) - P(x_0 \mid X_{t+k}) \big| < \varepsilon \Bigg) \\
&\geq 1 - P(Q_{bad}) - P(S_{bad}) \\
&> 1 - 3\varepsilon
\end{aligned}$$

As ε is arbitrary this shows that the sequence $f_t(X_t) = P(X_0 \mid X_t)$ is Cauchy in probability. That is already known to prove the claim, but we finish the reasoning for completeness: By applying the Borel-Cantelli lemma to a subsequence t_k such that

$$\sum_{k=1}^{\infty} P\left(\big| f_{t_k}(X_{t_k}) - f_{t_{k+1}}(X_{t_{k+1}}) \big| > \frac{1}{2^k} \right) < \sum_{k=1}^{\infty} \frac{1}{2^k} = 1 < \infty,$$

it follows that $f_{t_k}(X_{t_k})$ is almost surely a Cauchy sequence in $\mathbb{R}$. Let f_∞ be the almost surely limit of $f_{t_k}(X_{t_k})$. Applying the fact that $f_t(X_t)$ is Cauchy in probability, it follows that it converges in probability to f_∞, as desired.

□

2 Generalized Koopman-Pitman-Darmois Theorem

2.1 Notation

In the following, we will often have the following notational situation: there is a set $\{1, \dots, n\}$ of "variable indices" and a finite set I of "function indices" together with attached variables index sets $N_i \subseteq \{1, \dots, n\}$ for $i \in I$. In this situation, define $N(j) \coloneqq \bigcup_{i : j \in N_i} N_i$. For $B \subseteq \{1, \dots, n\}$, define

$$N(B) \coloneqq \bigcup_{j \in B} N(j) = \bigcup_{i : N_i \cap B \neq \emptyset} N_i.$$

Another notational note: for a function $f(X)$, we denote one of its values by $f(x)$. Similarly, given a density $p(\Theta)$, we denote one of its values by $p(\theta)$.

Finally, θ can in the following either live in an open subset $O \subseteq \mathbb{R}^M$ or only take countably many values. Denote by L^2 either the space $L^2(O)$ or $L^2(\{\theta_l\})$ depending on the situation. In any case, it is a Hilbert space.

2.2 The Theorem

We will ignore some regularity conditions and slight generalizations and defer a discussion of them to Remark 8.

Theorem 6 (Generalized Koopman-Pitman-Darmois)**.** *Let $p(\Theta)$ be either a probability mass function over countably many values θ, or a density over $\theta \in O$ for some open set $O \subseteq \mathbb{R}^M$. Furthermore, let*

$$p(X \mid \Theta) = p(X_1, \dots, X_n \mid \Theta)$$

be a conditional probability density over $\mathbb{R}^n$. Together, these can be used to define the posterior mass function or density $p(\Theta \mid X)$ using Bayes rule.

Let a finite set I, subsets $N_i \subseteq \{1, \dots, n\}$ for $i \in I$, and conditional potential functions $\psi_i(\cdot \mid \Theta = \theta) : \mathbb{R}^{N_i} \to \mathbb{R}_{\geq 0}$ be given such that

$$p(X \mid \Theta) = \prod_{i \in I} \psi_i(X_{N_i} \mid \Theta).^{3}$$

We also assume that this distribution has a (low-dimensional) sufficient statistic: there exists a function $G : \mathbb{R}^n \to \mathbb{R}^D$ (for some reasonably small D) such that

$$p(\Theta \mid X) = p\big(\Theta \mid G(X)\big).$$

We assume three differentiability conditions:

(i) G is assumed to be differentiable.

(ii) Let $F : \mathbb{R}^D \to L^2$ be the function from Lemma 3 with the following property:

$$F(G(x)) = p(\Theta \mid X = x).$$

F is assumed to be Fréchet differentiable.

(iii) Each ψ_i, when considered as a function

$$\psi_i : \mathbb{R}^{N_i} \to L^2, \quad x_{N_i} \mapsto \Big[\psi(x_{N_i} \mid \Theta) : \theta \mapsto \psi(x_{N_i} \mid \theta)\Big],$$

is assumed to be continuously Fréchet differentiable, i.e., a C^1 function.

We also assume the existence of a reference parameter θ^0 such that $p(\theta^0) > 0$ and $p(x \mid \theta^0) > 0$ for all $x \in \mathbb{R}^n$. Then there are:

1. *a dimension $K \leq D$;*
2. *a subset $B \subseteq \{1, \dots, n\}$ of size $|B| = K$;*
3. *a set $E \subseteq I$ of "exceptions" given as all i with $N_i \cap N(B) \neq \emptyset$;*
4. *differentiable functions $g_i : \mathbb{R}^{N_i} \to \mathbb{R}^K$ for $i \in \overline{E} = I \setminus E$;*
5. *a function $U(\Theta)$ with values in $\mathbb{R}^K$; and*

[3] This always trivially exists: simply choose $I = \{1\}$, $N_1 = \{1, \dots, n\}$ and $\psi_1 = P$. Thus, there is no strong assumption embedded in this condition. However, *some* probability distributions factor in more interesting ways, e.g. Markov random fields and Bayesian networks with sparse graphs. For those distributions, the conclusion of the theorem becomes interesting.

6. a C^1 function $h : \mathbb{R}^{\overline{N(B)}} \to \mathbb{R}$, where $\overline{N(B)} = \{1, \dots, n\} \setminus N(B)$;

such that the distribution $P(X \mid \Theta)$ factorizes as follows:

$$P(X \mid \Theta) = \frac{1}{Z(\Theta)} \cdot e^{\left[U(\Theta)^T \sum_{i \notin E} g_i(X_{N_i})\right]} \cdot h\big(X_{\overline{N(B)}}\big) \cdot \prod_{i \in E} \psi_i(X_{N_i} \mid \Theta).$$

The specific definition of these functions is revealed in the proof.

Proof. Define $F' : \mathbb{R}^D \to L^2$ by

$$F'(g) \coloneqq \ln \frac{F(g)}{\big[F(g)\big](\theta^0)} - ln \frac{p(\Theta)}{p(\theta^0)}$$

for g for which $\big[F(g)\big](\theta^0) > 0$, and $F'(g) \coloneqq 0$ otherwise. Then, for $x \in \mathbb{R}^n$, we have $\big[F(G(x))\big](\theta^0) = p(\theta^0 \mid x) > 0$ and thus

$$\begin{aligned} F'(G(x)) &= \ln \frac{F(G(x))}{\big[F(G(x))\big](\theta^0)} - \ln \frac{p(\Theta)}{p(\theta^0)} \\ &= \ln \frac{p(\Theta \mid x)}{p(\theta^0 \mid x)} - \ln \frac{p(\Theta)}{p(\theta^0)} \\ &= \ln \frac{p(x \mid \Theta)}{p(x \mid \theta^0)} \\ &= \sum_{i \in I} \ln \frac{\psi_i(x_{N_i} \mid \Theta)}{\psi_i(x_{N_i} \mid \theta^0)} \\ &= \sum_{i \in I} f_i(x_{N_i}) \end{aligned}$$

with $f_i : \mathbb{R}^{N_i} \to L^2$ defined in the obvious way. By the following Lemma 7, we obtain an equality

$$\ln \frac{p(x \mid \Theta)}{p(x \mid \theta^0)} = \sum_{i : N_i \cap N(B) \neq \emptyset} f_i(x_{N_i}) + U \sum_{i : N_i \cap N(B) = \emptyset} g_i(x_{N_i}) + C,$$

where $B \subseteq \{1, \dots, n\}$ is of size $|B| = K \leq D$, $g_i : \mathbb{R}^{N_i} \to \mathbb{R}^K$, linear $U : \mathbb{R}^K \to L^2$ and a constant $C \in L^2$. Exponentiating, we obtain:

$$p(x \mid \Theta) = \frac{1}{e^{-C}} \cdot e^{\left[U \sum_{i : N_i \cap N(B) = \emptyset} g_i(x_{N_i})\right]} \cdot p(x \mid \theta^0) \cdot \prod_{i : N_i \cap N(B) \neq \emptyset} \frac{\psi_i(x_{N_i} \mid \Theta)}{\psi_i(x_{N_i} \mid \theta^0)}.$$

Now, note that $C \in L^2$ and thus $e^{-C} = Z(\Theta)$ is a function of θ. Similarly, $U : \mathbb{R}^K \to L^2$ can be considered as a function with linear outputs $U(\Theta = \theta) : \mathbb{R}^K \to \mathbb{R}$. Defining

$$h\big(x_{\overline{N(B)}}\big) \coloneqq \prod_{i : N_i \cap N(B) = \emptyset} \psi_i(x_{N_i} \mid \theta^0)$$

gives the result. □

2.3 Lemma: Additive Summary Equations

The following is taken from The Additive Summary Equation:

Lemma 7 (Additive Summary Equation)**.** *Let $\mathcal{H}$ be a Hilbert space, $f : \mathbb{R}^n \to \mathcal{H}$ a* continuously *Fréchet differentiable function (i.e., C^1 function), $G : \mathbb{R}^n \to \mathbb{R}^D$ a differentiable function, and $F : \mathbb{R}^D \to \mathcal{H}$ a Fréchet differentiable function. Additionally, assume a finite index set I, subsets $N_i \subseteq \{1, \dots, n\}$, and Fréchet differentiable functions $f_i : \mathbb{R}^{N_i} \to \mathcal{H}$ such that, for all $x \in \mathbb{R}^n$:*

$$F(G(x)) = f(x) = \sum_{i \in I} f_i(x_{N_i}).$$

Then there exists $K \leq D$, a set $B \subseteq \{1, \dots, n\}$ of size $|B| = K$, differentiable functions $g_i : \mathbb{R}^{N_i} \to \mathbb{R}^K$, a linear operator $U : \mathbb{R}^K \to \mathcal{H}$, and a constant $C \in \mathcal{H}$ such that

$$f(x) = \sum_{i : N_i \cap N(B) \neq \emptyset} f_i(x_{N_i}) + U \sum_{i : N_i \cap N(B) = \emptyset} g_i(x_{N_i}) + C.$$

Proof. From the chain rule (which is also valid for Fréchet derivatives), we get

$$Df(x) = D(F \circ G)(x) = DF\big(G(x)\big) \circ DG(x),$$

which is a composition of the two linear operators

$$DG(x) : \mathbb{R}^n \to \mathbb{R}^D, \quad DF\big(G(x)\big) : \mathbb{R}^D \to \mathcal{H}.$$

For a linear operator, define its *rank* to be the dimension of its image. Since $\mathbb{R}^D$ creates a bottleneck of dimension D, we obtain $\operatorname{rank}\big(Df(x)\big) \leq D$ for all $x \in \mathbb{R}^n$.

Let $K := \max_{x \in \mathbb{R}^n} \operatorname{rank}\big(Df(x)\big) \leq D$ and let $x^0 \in \mathbb{R}^n$ be an argmax that achieves that maximum. Now, let $B \subseteq \{1, \dots, n\}$ a subset of size $|B| = K$ such that the vectors in $\mathcal{H}$ contained in the "matrix"

$$\big[Df(x^0)\big]_B$$

generates the image of $Df(x^0)$.[4] Let $U : \mathbb{R}^B \to \mathcal{H}$ consist of an orthonormal basis for the image of $Df(x^0)$; this exists since the image is a finite-dimensional $\mathbb{R}$-vector space. Let $U^* : \mathcal{H} \to \mathbb{R}^B$ be its Hermitian adjoint (i.e., the transpose in the finite-dimensional case). Then $UU^* : \mathcal{H} \to \mathcal{H}$ is a linear operator that projects on the image of $Df(x^0)$, as is well known. In particular, it leaves elements in the image of $Df(x^0)$ invariant.

Now, let $x \in \mathbb{R}^n$ be any element with $x_{N(B)} = x^0_{N(B)}$. Then it follows that

$$\begin{aligned}
\big[Df(x)\big]_B &= \sum_{i : N_i \cap B \neq \emptyset} \big[Df_i(\underbrace{x_{N_i}}_{=x^0_{N_i}})\big]_B + \sum_{i : N_i \cap B = \emptyset} \underbrace{\big[Df_i(x_{N_i})\big]_B}_{=0} \\
&= \sum_{i : N_i \cap B \neq \emptyset} \big[Df_i(x^0_{N_i})\big]_B + \sum_{i : N_i \cap B = \emptyset} \underbrace{\big[Df_i(x^0_{N_i})\big]_B}_{=0} \\
&= \big[Df(x^0)\big]_B.
\end{aligned}$$

Since the image of $Df(x)$ is maximally K-dimensional and it already contains the linearly independent set $\big[Df(x^0)\big]_B$ of size K, we must have that the image of $Df(x)$ coincides with that of $Df(x^0)$. Consequently, remembering that UU^* projects on this image, we have

$$Df(x) = UU^* \circ Df(x) \tag{1}$$

whenever $x_{N(B)} = x^0_{N(B)}$.

Now, let $x \in \mathbb{R}^n$ be *any* point. Define $x' := (x^0_{N(B)}, x_{\overline{N(B)}})$, where $\overline{N(B)} = \{1, \dots, n\} \setminus N(B)$. I.e., x' contains the entries of x in the dimensions in $\overline{N(B)}$ and the entries of x^0 in the dimensions in $N(B)$. Define the differentiable path $\gamma : [0, 1] \to \mathbb{R}^n$ as the path connecting x^0 to x':

$$\gamma(t) = tx' + (1 - t)x^0.$$

Then $f \circ \gamma : [0, 1] \to \mathcal{H}$ is Fréchet differentiable. Thus, it satisfies the fundamental theorem of calculus for C^1 functions, meaning that

$$\begin{aligned}
f(x') - f(x^0) &= (f \circ \gamma)(1) - (f \circ \gamma)(0) \\
&= \int_0^1 (f \circ \gamma)'(t) dt \\
&= \int_0^1 Df\big(\gamma(t)\big) \circ \gamma'(t) dt \\
&= \dots
\end{aligned}$$

Now, note that $\gamma'(t) = x' - x^0 = (0_{N(B)}, (x - x^0)_{\overline{N(B)}})$, and thus:

$$\begin{aligned}
\dots &= \int_0^1 Df\big(\gamma(t)\big)_{\overline{N(B)}} \cdot (x - x^0)_{\overline{N(B)}} dt \\
&= \dots
\end{aligned}$$

[4] To be precise if you've never thought about infinite-dimensional spaces: $Df(x^0)$ may not be a matrix if $\mathcal{H}$ is infinite-dimensional. However, $Df(x^0)$ is fully determined by the n "column" vectors $\big(Df(x^0)(e_i)\big)_{i=1}^n$, and the span of those contains a basis $(Df(x^0)(e_i))_{i \in B}$ of size $|B| = K$. We can view this basis then as a "matrix" or "operator" $\mathbb{R}^B \to \mathcal{H}$ in the obvious sense and denote it by $\big[Df(x^0)\big]_B$.

We have $\gamma(t)_{N(B)} = x^0_{N(B)}$. Thus, $Df\big(\gamma(t)\big) = UU^* \circ Df\big(\gamma(t)\big)$, see Equation (1). Furthermore, if i is such that $N_i \cap N(B) \neq \emptyset$, then $N_i \subseteq N(B)$ and thus $Df_i\big(\gamma(t)\big)_{\overline{N(B)}} = Df_i\big(\gamma(t)_{N_i}\big)_{\overline{N(B)}} = 0$. We obtain:

$$\begin{aligned}
\ldots &= U \sum_{i : N_i \cap N(B) = \emptyset} U^* \int_0^1 Df_i(\gamma(t))_{\overline{N(B)}} \cdot (x - x^0)_{\overline{N(B)}} dt \\
&=: U \sum_{i : N_i \cap N(B) = \emptyset} g_i(x).
\end{aligned}$$

To show that g_i is of the type signature $g_i(x) = g_i(x_{N_i})$, note that reversing the fundamental theorem clearly shows $g_i(x) = U^*\big(f_i(x_{N_i}) - f_i(x^0)\big)$. This also shows the differentiability of g_i.

We have

$$\begin{aligned}
f(x) - f(x') &= \sum_i f_i(x_{N_i}) - f_i(x'_{N_i}) \\
&= \sum_{i : N_i \cap N(B) \neq \emptyset} f_i(x_{N_i}) - f_i(\underbrace{x'_{N_i}}_{= x^0_{N_i}}) + \sum_{i : N_i \cap N(B) = \emptyset} f_i(x_{N_i}) - f_i(\underbrace{x'_{N_i}}_{= x_{N_i}}) \\
&= \sum_{i : N_i \cap N(B) \neq \emptyset} f_i(x_{N_i}) - f_i(x^0_{N_i}).
\end{aligned}$$

We obtain:

$$\begin{aligned}
f(x) &= f(x^0) + f(x') - f(x^0) + f(x) - f(x') \\
&= f(x^0) + U \sum_{i : N_i \cap N(B) = \emptyset} g_i(x_{N_i}) + \sum_{i : N_i \cap N(B) \neq \emptyset} f_i(x_{N_i}) - f_i(x^0_{N_i}) \\
&= \sum_{i : N_i \cap N(B) \neq \emptyset} f_i(x_{N_i}) + U \sum_{i : N_i \cap N(B) = \emptyset} g_i(x_{N_i}) + \sum_{i : N_i \cap N(B) = 0} f_i(x^0_{N_i}).
\end{aligned}$$

Now, setting $C := \sum_{i : N_i \cap N(B) = 0} f_i(x^0_{N_i})$, the result follows. □

Remark 8. *We make several remarks on regularity conditions in the theorem and a slight generalization to open sets in $\mathbb{R}^n$ instead of $\mathbb{R}^n$ itself.*

1. *One crucial condition in the proof was that the function F given by $F(g) = p(\Theta \mid G(X) = g)$ is Fréchet differentiable, at least in points g such that $g = G(x)$ for some x. Ideally, we would formulate that differentiability condition directly in terms of conditions for $p(\Theta)$ and $p(X \mid \Theta)$. We now sketch how that could be done, but leave details to the future:*

   Due to Lemma 3, we have a commutative diagram

$$\begin{tikzcd}
\mathbb{R}^n \arrow[r, "G"'] \arrow[rr, bend left, "p(\Theta \mid \bullet)"] & \mathbb{R}^D \arrow[r, "F"'] & L^2
\end{tikzcd}$$

   Thus, if we can locally invert the differentiable function G around points $g = G(x)$ in $\mathbb{R}^D$, then we can reduce the differentiability of F to that of the minimal sufficient statistic $p(\Theta \mid \bullet)$. There are fairly regular conditions under which this local inversion is possible, see the section on selections in this Wikipedia article.

   However, this leaves the question open of when the minimal sufficient statistic $p(\Theta \mid \bullet)$ is differentiable. Its component functions are given by

$$x \mapsto p(\Theta = \theta \mid x) = \frac{p(x \mid \theta) \cdot p(\theta)}{\int_{\theta'} p(x \mid \theta') \cdot p(\theta')}.$$

   Thus, if $p(x \mid \theta)$ is differentiable in x for all θ, then under some further regularity conditions needed to swap integration and differentiation, the component function will be differentiable. Will this already result in differentiability of $p(\Theta \mid \bullet)$? If θ can take infinitely many values and we stack the derivatives of all component functions together to a derivative of $p(\Theta \mid \bullet)$, then we still need to check when the stack will be a ***bounded*** *linear function. We leave the question of when boundedness can be ensured to the reader.*

2. *In the theorem, we can also replace $\mathbb{R}^n$ by an open set $\mathcal{X} \subseteq \mathbb{R}^n$. Due to the construction in Lemma 7, we need a certain "projected convexity" condition to make this work: there needs to be an x_0 achieving the argmax of the differential of f such that for all $x \in \mathcal{X}$, the point $x' = (x^0_{N(B)}, x_{\overline{N(B)}})$ and the whole line segment between x^0 and x' lie within $\mathcal{X}$. This then allows to integrate along that line segment.*

3. *Finally, note that the theorem says nothing about the regularity of $U(\Theta)$ with respect to θ. This is since in the theorem, we consider it as a function $\theta \mapsto U(\theta) \in \mathbb{R}^K \cong Lin(\mathbb{R}^K, \mathbb{R})$ even though it can more appropriately be described as a function $U : \mathbb{R}^K \to L^2$. In that setting, $U(x)$ is in L^2 for all $x \in \mathbb{R}^K$.*