Appendice B

B Elementi di Teoria dell'Informazione[1]

B.1 Introduzione

E' noto da tempo che i fenomeni percettivi possono essere formalizzati e studiati mediante la Teoria dell'Informazione (TI). Attneave (1954), asseriva:

> *La principale funzione di una macchina percettiva consiste nell'eliminazione dell'informazione ridondante, procedendo così ad una descrizione o codifica dell'informazione stessa in una forma più economica rispetto a quella presente sui recettori.*

In altri termini, i pesi sinaptici di una rete neurale elaborano gli stimoli esterni presenti sui recettori (ingressi della rete). L'informazione è, di conseguenza, massimizzata quando il segnale passa nei vari stadi (o strati) di processamento della rete e rispettando opportuni vincoli.

E' evidente, quindi, che tra gli strumenti formali per lo studio delle reti neurali, e più in generale dei circuiti, debba essere presente anche la Teoria dell'Informazione.

B.2 Entropia

Sia X una variabile aleatoria (VA) discreta quantizzata, in pratica, con un numero finito di livelli distribuiti uniformemente, tale che:

$$X = \{ x_k \mid k = 0, \pm 1, \ldots, \pm K \} ; \qquad (2K+1) \text{ livelli,}$$

dove x_k, rappresenta il k-esimo valore di X, definiamo la probabilità dell'occorrenza dell'evento $X = x_k$ come:

[1] Tratto da Haykyn: Neural Networks …., 1999.

$$p_k = P(X = x_k)$$

ovviamente, valgono gli assiomi del calcolo della probabilità:

$$0 \le p_k \le 1 \quad \text{e} \qquad \sum_{k=-K}^{K} p_k = 1;$$

Se l'evento $X=x_k$ occorre con probabilità pari a 1, non c'è "sorpresa" e perciò non c'è neanche "informazione". Nel caso in cui $p_k<1$, allora saranno presenti anche altri valori ($p_i>0$ per $i \neq k$) ne segue allora, che l'informazione ricevuta sarà maggiore. In altre parole possiamo dire che l'"informazione" è correlata in qualche modo alla "sorpresa" a alla "incertezza". Possiamo anche affermare che l'informazione è proporzionale all'inverso della probabilità dell'occorrenza.

Definizione: *si definisce quantità* ***guadagno di informazione****, dopo l'osservazione dell'evento $X=x_k$ con probabilità p_k; la quantità:*

$$I(x_k) = \log\left(\frac{1}{p_k}\right) = -\log p_k .$$

$I(x_k)$ è una quantità discreta. L'unita di misura dell'informazione, se il logaritmo è a base 2, è il [*bit*]. Quando il logaritmo è naturale, si usa il [*nat*].

Proprietà:

1. $I(x_k) = 0$ per $p_k=1$.

   Ovvero, non c'è guadagno di informazione se l'occorrenza l'evento è certa.

2. $I(x_k) \ge 0$ per $0 \le p_k \le 1$.

L'evento $X=x_k$ può portare o non portare informazione. Sicuramente, però, non ci sarà perdita d'informazione.

3. $I(x_k) > I(x_i)$ per $p_k < p_i$.

   L'evento meno probabile è quello che porta maggiore informazione.

Definizione: *si definisce **entropia** il valore medio del guadagno d'informazione $I(x_k)$ relativo a tutti i (2K+1) livelli ammissibili.*

$$H(X) = E\left[I(x_k)\right] = \sum_{k=-K}^{K} p_k I(x_k) = -\sum_{k=-K}^{K} p_k \log p_k \,.$$

L'entropia $H(X)$ rappresenta una misura della quantità media dell'informazione portata da un messaggio.

Proprietà:

1. $0 \le H(X) \le (2K+1)$.
2. $H(X) = 0$, se e solo se per qualche k, $p_k=1$.

   Questo limite inferiore coincide con l'evento certo.

3. $H(X) = \log_2(2K+1)$, se e solo se $p_k=1/(2K+1)$ per tutti i k.

   L'entropia e massima se i livelli sono equiprobabile, ovvero, distribuiti uniformemente.

   Tale limite superiore viene definito come *condizione di massima incertezza*.

La dimostrazione della 3. è derivata dal seguente

Lemma: Data due distribuzioni (di quantità di masse) p_k e q_k, per una VA discreta X, allora la quantità:

$$\sum_k p_k \log\left(\frac{p_k}{q_k}\right) \ge 0$$

è pari a zero se e solo se $p_k = q_k$, per ogni k.

B.3 Entropia relativa di Kullback-Leibler

Siano $p_X(x)$ e $q_X(x)$ le probabilità che la VA X sia nello stato x sotto due differenti condizioni operative desritta da p e q; segue la

Definizione: *Si definisce **entropia relativa** (o **divergenza** o distanza o **cross-entropia**), tra funzioni di probabilità $p_X(x)$ e $q_X(x)$; come;*

$$D_{p\|q} = \sum_{x\in\mathcal{X}} p_X(x) \log\left(\frac{p_X(x)}{q_X(x)}\right);$$

dove con $\mathcal{X}$ si è indicato l'alfabeto dei simboli della VA X. La quantità $q_X(x)$ rappresenta il riferimento di misura.

L'entropia relativa di Kullback-Leibler, a volte indicata come $K(p,q)$, può essere presentata anche nel contesto della geometria differenziale Amari(1985), Amari (1992), come metrica di Riemann nello spazio delle distribuzioni.

Osservazione: l'entropia relativa $K(p,q)$, non è una vera distanza, infatti, non vale la proprietà di simmetria: $K(p,q) \neq K(q,p)$; e può essere interpretata come una "*quasi distanza*".

B.4 Entropia differenziale

I concetti base della TI, possono essere facilmente estesi nel contesto delle variabili aleatorie continue.

Definizione: *Sia X una VA continua con funzione di densità di probabilità (pdf) $f_X(x)$. Per analogia con il caso di VA discrete definiamo **entropia differenziale** la quantità:*

$$h(X) = -\int_{-\infty}^{\infty} f_X(x)\log f_X(x)dx = -E\left[\log f_X(x)\right].$$

La $h(X)$, rappresenta una quantità matematica, piuttosto che una misura della aleatorietà di X.

La relazione tra la definizione di entropia per VA discrete e continue può essere studiata come caso limite di una VA discreta che assume il valore $x_k=k\delta_x$, dove k = 0, ±1, ±2, … e δ_x tende a zero. Per definizione la X assume un certo valore costante nell'intervallo $[x_k,\ x_k+k\delta_x]$ con probabilità $f_X(x_k)\ \delta_x$. La entropia ordinaria della VA X può essere scritta considerando il limite per δ_x tendente a zero come:

$$\begin{aligned} H(X) &= -\lim_{\delta x\to 0}\sum_{k=-\infty}^{\infty} f_X(x_k)\delta x\log[f_X(x_k)\delta x] \\ &= \lim_{\delta x\to 0}\left[\sum_{k=-\infty}^{\infty} f_X(x_k)[\log f_X(x_k)]\delta x + \sum_{k=-\infty}^{\infty} f_X(x_k)\delta x\right] \\ &= -\int_{-\infty}^{\infty} f_X(x)\log f_X(x)dx - \lim_{\delta x\to 0}\log\delta x\int_{-\infty}^{\infty} f_X(x)dx \\ &= h(X) - \lim_{\delta x\to 0}\log\delta x \end{aligned}$$

Ricordiamo che $\int_{-\infty}^{\infty} f_X(x)dx = 1$, e che per δ_x tendente a zero la $h(X)$ tende all'infinito. Questo significa che l'entropia di una variabile continua è infinitamente larga potendo assumere infiniti valori con probabilità infinitesima nell'intervallo $(-\infty,+\infty)$.

Il problema relativo al termine $\log\delta_x$ può essere evitato considerando la definizione di entropia differenziale. In questo caso il termine $-\log\delta_x$ è assunto come riferimento.

Generalizzando al caso di VA vettoriali $\mathbf{X} = [X_1, X_2, \ldots, X_N]$ l'entropia differenziale è definita come:

$$h(\mathbf{X}) = -\int_{-\infty}^{\infty} f_{\mathrm{X}}(\mathbf{x})\log f_{\mathrm{X}}(\mathbf{x})d\mathbf{x} = -E\left[\log f_{\mathrm{X}}(\mathbf{x})\right]$$

dove $f_{\mathrm{X}}(\mathbf{x})$ è la pdf congiunta di $\mathbf{X}$.

Lemma: entropia della distribuzione uniforme.

Se una VA X continua è distribuita uniformemente nell'intervallo $[0, a]$ la sua entropia differenziale è $h(X)=\log a$.

Infatti, dalla definizione possiamo ricavare:

$$h(X) = -\int_{-\infty}^{\infty}\frac{1}{a}\log\frac{1}{a}dx = -\int_{0}^{a}dx\frac{1}{a}\log\frac{1}{a} = \log a$$

Proprietà: dell'entropia differenziale

1. $h(X)=h(X+c)$, per c=costante;

   ovvero, la traslazione non cambia il valore dell'entropia.

2. $h(aX)=h(X) + \log|a|$, con a = fattore di scalatura

   nel caso di VA $\mathbf{X}$ vettoriale quest'ultima diventa:

3. $h(\mathbf{AX})=h(\mathbf{X}) + \log|\det(\mathbf{A})|$.

B.5 Principio di Massima Entropia

Supponiamo di avere un sistema stocastico con un insieme noto di variabili di stato ma con distribuzione sconosciuta. Supponiamo, inoltre, che per mezzo di un qualche meccanismo di apprendimento, siano noti alcuni vincoli sulle funzioni di probabilità della VA degli stati del sistema. I vincoli possono essere espressi come medie di insieme o come valori limite sulle VA. Consideriamo, ora, il problema della scelta di un modello di

probabilità che sia ottimo, secondo qualche criterio, in grado di formalizzare le conoscenze a priori sul sistema stocastico stesso.

È possibile dimostrare che esistono infiniti modelli che soddisfano i vincoli noti su cui è possibile effettuare una scelta.

La risposta al problema della scelta della distribuzione può essere affrontato nel contesto del principio della Massima Entropia (*Max Ent*) (Jaynes 1957, 1982):

> *Se una inferenza è fatta sulla base di informazioni incomplete, occorre scegliere quella distribuzione che massimizza l'entropia ed è soggetta ai vincoli imposti dalla distribuzione.*

In effetti, la nozione di entropia fornisce gli strumenti per effettuare un insieme di misure in modo tale da favorire le pdf con alto valore di entropia.

Il principio di Massima Entropia può, quindi, essere visto come un *Problema di Ottimizzazione Vincolato*, che può essere risolto per mezzo delle definizioni introdotte dall'entropia differenziale.

Dato un insieme di pdf della VA X, consideriamo il problema di determinare quella che massimizza la:

$$h(X) = -\int_{-\infty}^{\infty} f_X(x) \log f_X(x) dx;$$

soggetta ai vincoli:

1. $f_X(x) \geq 0$; dove l'uguaglianza vale esternamente al supporto di x.

2. $\int_{-\infty}^{\infty} f_X(x) dx = 1$

3. $\int_{-\infty}^{\infty} f_X(x) g_i(x) dx = \alpha_i$ per i=1,2,…, m

dove $g_i(x)$ è una qualche funzione di x.

I vincoli espressi dalle 1 e dalle 2 descrivono delle proprietà fondamentali delle pdf. Le 3, invece, dice che i momenti di X dipendono dalla formulazione della funzione $g_i(x)$. In effetti la 3, somma le conoscenze a priori disponibili sulla VA X.

Per risolvere tale problema di ottimizzazione occorre definire una *funzione obiettivo* e, perciò, viene usato il metodo dei *moltiplicatori di Lagrange* :

$$J(f) = -\int_{-\infty}^{\infty} \left[-f_X(x) \log f_X(x) dx + \lambda_0 f_X(x) + \sum_{i=1}^{m} \lambda_i g_i(x) f_X(x) \right] dx$$

dove i termini λ_0, λ_1, …, λ_m; sono i *moltiplicatori di Lagrange*. Differenziando e integrando rispetto alla $f_X(x)$, e uguagliando a zero, si ottiene:

$$-1 - \log f_X(x) + \lambda_0 + \sum_{i=1}^{m} \lambda_i g_i(x) = 0$$

Risolvendo rispetto alla $f_X(x)$ si ottiene:

$$f_X(x) = \exp\left[-1 + \lambda_0 + \sum_{i=1}^{m} \lambda_i g_i(x) \right];$$

dove i moltiplicatori di Lagrange sono scelti in accordo con i vincoli 1, 2 e 3. L'ultima espressione definisce la pdf che massimizza l'entropia per il problema in questione.

Esempio: distribuzione Gaussiana mono dimensionale.

Problema: determinate la $f_X(x)$ tale che la $h(X)$ sia massima quando le conoscenze a priori disponibili sono il valore medio μ e la varianza σ^2.

Per definizione abbiamo che.

$$\int_{-\infty}^{\infty}(x-\mu)^2 f_X(x)dx = \sigma^2 = \text{costante}.$$

Per il rispetto del vincolo 3, si ha:

$$g_1(x) = (x-\mu)^2$$
$$\alpha_1 = \sigma^2 \quad .$$

Considerando il risultato ottenuto con i moltiplicatori di Lagrange, abbiamo che:

$$f_X(x) = \exp\left[-1+\lambda_0+\lambda_1(x-\mu)^2\right];$$

per il vincolo 2, che richiede la convergenza dell'integrale della $f_X(x)$, segue che il coefficiente λ_1 deve essere negativo. Sostituendo la precedente espressione nelle disequazioni dei vincoli 1 e 2 e, risolvendo rispetto a λ_0 e λ_1, otteniamo:

$$\lambda_0 = 1 - \log(2\pi\sigma^2)$$

e

$$\lambda_1 = -\frac{1}{2\sigma^2}.$$

La pdf che massimizza l'informazione desiderata, risulta, quindi, essere:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right);$$

che corrisponde proprio ad una pdf Gaussiana mono dimensionale della VA X con valore medio μ e varianza σ^2.

L'entropia differenziale che massimizza la precedente $f_X(x)$ risulta pertanto essere:

$$h(X) = \frac{1}{2}[1+\log(2\pi\sigma^2)];$$

notare che per la proprietà della traslazione, il valore medio non interessa quest'ultima espressione.

Esempio: distribuzione Gaussiana multivariata

Il secondo esempio che consideriamo consiste nella determinazione dell'entropia differenziale di una VA multidimensionale **X** con pdf Gaussiana. Per la proprietà di traslazione consideriamo il vettore *m*-dimensionale **X** a media nulla. La statistica del secondo ordine è descritta dalla matrice di covarianza $\mathbf{Q}_X = \mathbf{X}\mathbf{X}^T$ (prodotto esterno). Ne segue che la pdf multidimensionale è:

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{m/2}(\det(\mathbf{Q}_{\mathbf{X}}))^{1/2}}\exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{Q}_{\mathbf{X}}^{-1}\mathbf{x}\right).$$

Ricordando la definizione di entropia differenziale per VA multi dimensionali:

$$h(\mathbf{X}) = -\int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x})\log f_{\mathbf{X}}(\mathbf{x})d\mathbf{x} = -E\left[\log f_{\mathbf{X}}(\mathbf{x})\right];$$

e sostituendo la precedente in quest'ultima espressione, otteniamo il risultato:

$$h(\mathbf{X}) = \frac{1}{2}\left[m + m\log(2\pi) + \log\left|\det(\mathbf{Q}_{\mathbf{X}})\right|\right].$$

Possiamo osservare, quindi, che per il principio di *Massima Entropia*, in questo caso differenziale, nota una matrice di covarianza $\mathbf{Q}_X$ la pdf che massimizza l'entropia differenziale è Gaussiana.

B.6 Mutua Informazione per VA discrete

Il problema primario dei sistemi auto organizzanti, consiste nella determinazione di un algoritmo di apprendimento che determini una certa relazione ingresso/uscita. Tale relazione va determinata esclusivamente sulla base dei soli dati in ingresso.

Consideriamo un sistema con ingresso e uscita le VA discrete X e Y, i cui valori sono indicati con x e y rispettivamente.

L'entropia $H(X)$, fornisce una misura dell'incertezza a priori sugli ingressi X.

La questione, è quella di determinare l'incertezza sui dati X dopo le osservazioni Y. In questo contesto la notazione della *mutua informazione* fornisce importanti strumenti e proprietà.

Definizione: *si definisce* ***entropia condizionale*** *di X dato Y come segue:*

$$H(X|Y) = H(X,Y) - H(Y);$$

con la proprietà:

$$0 \le H(X|Y) \le H(X).$$

L'entropia condizionale, rappresenta l'ammontare di incertezza rimanente sulla VA d'ingresso X dopo l'osservazione della VA Y in uscita del sistema.

La quantità $H(X,Y)$ che compare nella precedente espressione prende il nome di *entropia congiunta* ed è definita come segue:

Definizione: *si definisce* ***entropia congiunta*** *delle VA discrete X e Y come:*

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y)$$

dove p(x,y) rappresenta la funzione di massa di probabilità congiunta delle X e Y, mentre $\mathcal{X}$ e $\mathcal{Y}$ i rispettivi alfabeti.

Osservazione: l'entropia, rappresenta la misura dell'incertezza sulla VA X. L'entropia condizionale $H(X|Y)$, invece, rappresenta l'ammontare di incertezza rimanente su X dopo l'osservazione della Y in uscita del sistema. È facilmente deducibile, quindi, che la differenza $H(X) - H(X|Y)$, rappresenta l'incertezza sull'ingresso del sistema risolta dell'osservazione dell'uscita. Questa quantità può quindi essere definita come segue:

Definizione: *si definisce* ***mutua informazione*** *I(X;Y) tra le VA discrete X e Y la quantità:*

$$I(X;Y) = H(X) - H(X|Y)$$

$$= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\left(\frac{p(x,y)}{p(x)\,p(y)}\right)$$

L'entropia è un caso particolare di mutua informazione:

$$H(X) = I(X;X)$$

Proprietà:

1. *Simmetria*: la mutua informazione tra X e Y è simmetrica:

   $$I(X;Y) = I(Y;X);$$

   dove $I(Y;X)$ rappresenta la misura dell'incertezza di Y che è risolta della osservazione degli ingressi X.

2. *Non negatività*: la mutua informazione tra X e Y è non negativa:

   $$I(X;Y) \ge 0.$$

Questa proprietà indica che non è possibile perdere mediamente informazione osservando le uscite. Inoltre, la mutua informazione è nulla, se e solo se, gli ingressi e le uscite del sistema sono statisticamente indipendenti.

4. *Reciprocità*: la mutua informazione tra X e Y è esprimibile in termini di entropia di Y, come:

$$I(X;Y) = H(Y) - H(Y|X)$$

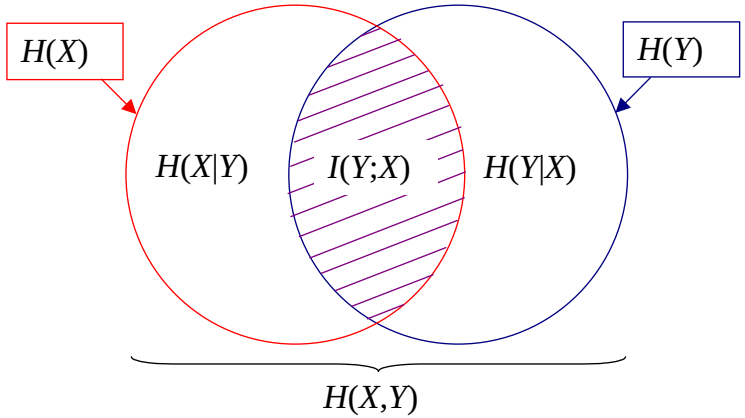

Figura 1 – Illustrazione delle relazioni tra la *mutua informazione* $I(X;Y)$ e le entropie $H(X)$ e $H(Y)$.

L'entropia dell'ingresso X, è rappresentata dal cerchio a sinistra mentre quella dell'uscita Y, con il cerchio a destra. La mutua informazione è rappresentata dalla sovrapposizione dei due cerchi.

B.7 Mutua Informazione per VA continue

Per analogia con la definizione precedente possiamo estendere tale definizione a VA continue.

Definizione: *la mutua informazione tra le VA X e Y continue è definita come:*

$$I(X;Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\log\left(\frac{f_X(x\,|\,y)}{f_X(x)}\right)dxdy$$

dove con $f_{X,Y}(x,y)$ rappresenta la pdf congiunta di X e Y, mentre, $f_X(x|y)$ è la pdf condizionata di X nota Y.

Notare, inoltre, che $f_{X,Y}(x,y) = f_X(x|y)\, f_Y(y)$ e quindi possiamo scrivere:

$$I(X;Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\log\left(\frac{f_{X,Y}(x,y)}{f_X(x)\,f_Y(y)}\right)dxdy$$

Per analogia valgono le proprietà

Proprietà:

1. *Simmetria/Reciprocità*: la mutua informazione tra X e Y è simmetrica e reciproca:

   $$I(X;Y) = h(X) - h(X|Y) = h(Y) - h(Y|X) = I(Y;X);$$

   dove $I(Y;X)$ rappresenta la misura dell'incertezza di Y che è risolta della osservazione degli ingressi X.

2. *Non negatività*: la mutua informazione tra X e Y è non negativa:

   $$I(X;Y) \geq 0.$$

   Questa proprietà indica che non è possibile perdere mediamente informazione osservando le uscite. Inoltre, la mutua informazione è nulla, se e solo se, gli ingressi e le uscite del sistema sono statisticamente indipendenti.

La funzione $h(X|Y)$ rappresenta *l'entropia differenziale condizionata* di X data Y, definita come;

$$h(X \mid Y) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\log f_X(x \mid y)dxdy$$

Se le VA X e Y sono statisticamente indipendenti, segue che la probabilità congiunta può essere fattorizzate come:

$$f_{X,Y}(x,y) = f_X(x)\, f_Y(y);$$

dove con $f_X(x)$ e $f_Y(y)$ abbiamo indicato le pdf marginali di X e Y rispettivamente. In modo equivalente, possiamo scrivere che:

$$f_X(x \mid y) = f_X(x);$$

ovvero, nel caso di VA statisticamente indipendenti, la conoscenza dell'uscita non contribuisce in alcun modo alla determinazione della pdf di X.

La definizione di mutua informazione data per VA X e Y scalari discrete, può essere generalizzata al caso di VA multimodali $\mathbf{X}$ e $\mathbf{Y}$ continue, come segue:

$$I(\mathbf{X};\mathbf{Y}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})\log\left(\frac{f_{\mathbf{X}}(\mathbf{x} \mid \mathbf{y})}{f_{\mathbf{X}}(\mathbf{x})}\right)d\mathbf{x}d\mathbf{y}$$

Per analogia, valgono le proprietà già formulate nel caso di VA discrete.

B.8 Divergenza di Kullback-Leibler (DKL)

Definizione: *siano* $f_{\mathbf{X}}(\mathbf{x})$ *e* $g_{\mathbf{X}}(\mathbf{x})$ *due pdf della VA* (m x 1) *multimodale* $\mathbf{X}$,*si definisce* ***Divergenza di Kullback-Leibler****, tra* $f_{\mathbf{X}}(\mathbf{x})$ *e* $g_{\mathbf{X}}(\mathbf{x})$ *come segue:*

$$D_{f_{\mathbf{X}}\|g_{\mathbf{X}}} = \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x})\log\left(\frac{f_{\mathbf{X}}(\mathbf{x})}{g_{\mathbf{X}}(\mathbf{x})}\right)d\mathbf{x}$$

Proprietà:

1. La $D_{f_{\mathbf{X}}\|g_{\mathbf{X}}} \geq 0$ è sempre positiva o zero quando $f_{\mathbf{X}}(\mathbf{x}) = g_{\mathbf{X}}(\mathbf{x})$ ovvero si ha una sovrapposizione perfetta tra le due pdf.
2. La $D_{f_{\mathbf{X}}\|g_{\mathbf{X}}}$, è invariante rispetto ai seguenti cambiamenti-trasformazioni sul vettore di ingresso $\mathbf{x}$:
   - permutazione dei componenti di $\mathbf{x}$;
   - cambiamenti di scala (amplificazioni);
   - trasformazioni non lineari monotone.

Osservazione: La mutua informazione $I(\mathbf{X};\mathbf{Y})$ tra la coppia di vettori $\mathbf{X}$ e $\mathbf{Y}$ ha una interessante interpretazione in termini di divergenza di DKL. Prima notiamo che:

$$f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y}) = f_{\mathbf{Y}}(\mathbf{y} \mid \mathbf{x})\, f_{\mathbf{X}}(\mathbf{x});$$

e quindi possiamo riscrivere l'espressione della mutua informazione come:

$$I(\mathbf{X};\mathbf{Y}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})\log\left(\frac{f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})}{f_{\mathbf{X}}(\mathbf{x})\, f_{\mathbf{Y}}(\mathbf{y})}\right)d\mathbf{x}d\mathbf{y}.$$

Confrontando quest'ultima con la definizione di DKL, segue il seguente risultato:

$$I(\mathbf{X};\mathbf{Y}) = D_{f_{\mathbf{X},\mathbf{Y}}\| f_{\mathbf{X},\mathbf{Y}}}.$$

La mutua informazione $I(\mathbf{X};\mathbf{Y})$ tra $\mathbf{X}$ e $\mathbf{Y}$ è equivalente alla divergenza di Kullback-Leibler tra le pdf congiunte $f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})$ e il prodotto tra le pdf $f_{\mathbf{X}}(\mathbf{x})$ e $f_{\mathbf{Y}}(\mathbf{y})$.

B.9 DKL tra una pdf fX(x) il prodotto delle sue pdf marginali

Consideriamo, ora, la DKL tra una pdf $f_{\mathbf{X}}(\mathbf{x})$ di una VA **X**, vettore di dimensione $m\times 1$ e il prodotto delle sue m pdf marginali.

Definizione: *la i-esima pdf marginale dell'elemento X_i rispetto alla VA* **X***, è definita come:*

$$\tilde{f}_{X_i}(x_i) = \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}^{(i)}, \qquad i = 1, 2, \ldots, m;$$

dove $\mathbf{x}^{(i)}$ *è il vettore il rimanente, di dimensione* $(m\text{-}1)\times 1$*, che si ottiene dopo aver rimosso il l'i-esimo elemento da* **x**.

La DKL tra la pdf $f_{\mathbf{X}}(\mathbf{x})$ e la distribuzione fattoriale $\prod_{i=1}^{m} \tilde{f}_{X_i}(x_i)$ è data da:

$$D_{f_{\mathbf{X}} \| \tilde{f}_{\mathbf{X}}} = \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x}) \log \left(\frac{f_{\mathbf{X}}(\mathbf{x})}{\prod_{i=1}^{m} \tilde{f}_{X_i}(x_i)} \right) d\mathbf{x};$$

che può essere scritta in forma espansa come:

$$D_{f_{\mathbf{X}} \| \tilde{f}_{\mathbf{X}}} = \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x}) \log f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} - \sum_{i=1}^{m} \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x}) \log \tilde{f}_{\mathbf{X}_i}(x_i) d\mathbf{x}.$$

Osservazione: il primo integrale di quest'ulrima espressione, coincide, per definizione, alla entropia differenziale $-h(\mathbf{X})$ della VA **X**.

Per quanto riguarda il secondo termine possiamo, inoltre, osservare che:

$d\mathbf{x} = d\mathbf{x}^{(i)} dx_i$.

Ne segue quindi:

$$\int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x}) \log \tilde{f}_{\mathbf{X}_i}(x_i) d\mathbf{x} = \int_{-\infty}^{\infty} \log \tilde{f}_{\mathbf{X}_i}(x_i) \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}^{(i)} dx_i;$$

dove l'integrale interno nella parte a destra, è fatto rispetto al vettore $(m\text{-}1)\times 1$, $d\mathbf{x}^{(i)}$, mentre quello esterno, rispetto allo scalare dx_i . Dalla definizione di i-esima pdf marginale di **X**, $\tilde{f}_{X_i}(x_i)$, segue, inoltre:

$$\begin{aligned} \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x}) \log \tilde{f}_{\mathbf{X}_i}(x_i) d\mathbf{x} &= \int_{-\infty}^{\infty} \tilde{f}_{\mathbf{X}_i}(x_i) \log \tilde{f}_{\mathbf{X}_i}(x_i) dx_i \\ &= -\tilde{h}(X_i), \qquad i = 1, 2, \ldots, m \end{aligned}$$

dove $\tilde{h}(X_i)$ è la i-esima entropia differenziale marginale (basata, cioè, sulla pdf marginale $\tilde{f}_{X_i}(x_i)$). Combinando quest'ultima con l'espressione della DKL $D_{f_{\mathbf{X}} \| \tilde{f}_{\mathbf{X}}}$, otteniamo:

$$D_{f_{\mathbf{X}} \| \tilde{f}_{\mathbf{X}}} = -h(\mathbf{X}) + \sum_{i=1}^{m} \tilde{h}(X_i).$$

Questa espressione della DKL marginale è particolarmente importante nello studio della separazione cieca di sorgenti.