UNIVERSIDAD CARLOS III DE MADRID
Departamento de Teoría de la Señal y Comunicaciones

DOCTORAL THESIS

INFORMATION–ESTIMATION RELATIONSHIPS OVER BINOMIAL, NEGATIVE BINOMIAL AND POISSON MODELS

Author: CAMILO GIL TABORDA
Supervised by: FERNANDO PÉREZ CRUZ
September 2014

Tesis Doctoral: INFORMATION–ESTIMATION RELATIONSHIPS OVER BINOMIAL, NEGATIVE BINOMIAL AND POISSON MODELS

Autor: Camilo Gil Taborda

Director: D. Fernando Pérez Cruz

Fecha:

Tribunal

Presidente:

Vocal:

Secretario:

Acknowledgements

There is not enough space and time to write this section. I hope to have the rest of my life to thank all the people that have made me happy!

Abstract

This thesis presents several relationships between information theory and estimation theory over random transformations that are governed through probability mass functions of the type binomial, negative binomial and Poisson.

The pioneer expressions that arose relating these fields date back to the 60's when Duncan [14, 15] proved that the input–output mutual information of a channel affected by Gaussian noise can be expressed as a time integral of the causal minimum mean square error. With the time, additional works due to Zakai[1], Kadota [30], Mayer-Wolf [40], Lipster [38] and Guo *et al.* [23] –among others– suggested the fact that there could be a hidden structure relating concepts such as the mutual information with some estimation quantities over a wide range of scenarios. The most prominent work in this field states that, over a real-valued Gaussian channel, the derivative of the input–output mutual information with respect to the signal to noise ratio is proportional to the mean square error achieved when measuring the loss between the input X and its conditional mean estimate based on an observation on the output. The minimum value of the mean square error is achieved precisely by the conditional mean estimate of the input, which gives rise to the well known "I-MMSE"[2] relationship. Similar expressions can be derived by studying the derivative of the relative entropy between two distributions obtained at the output of a Gaussian channel.

The expressions proved for the Gaussian channel translate verbatim to the Poisson channel [3] where the main difference lies in the loss function used to state the connection between information and estimation. In this framework, regarding the derivative of the input–output mutual information, it is further known that the considered loss function achieves its minimum value when is measured the difference between the input and its conditional mean estimate. This behavior has two main implications: in the context of the information–estimation relationships, it gives rise to the "I-MMLE"[3]

[1] In [23] it is pointed out that results obtained initially by Duncan regarding relationships between information and estimation where also obtained independently by Zakai.

[2] In "I-MMSE" the "I" stands for "Information" and the "MMSE" stands for "Minimum Mean Square Error".

[3] In "I-MMLE" the "I" stands for "Information" and the "MMLE" stands for "Minimum Mean Loss Error". Notice that the "I-MMSE" relationship could also be denominated as an "I-MMLE" relationship.

relationship over the Poisson model; second, it converts the loss function to a Bregman divergence, a property that is shared with the square distance used to state information–estimation relations in the Gaussian channel.

Based on the previous results we explore similar relationships in the context of the binomial and negative binomial models. In each model, using a deterministic input preprocessing, we develop several information–estimation relationships, depending solely on input statistics and its respective conditional estimates, that in some scenarios are given through Bregman divergences as was done formerly for the Gaussian and Poisson models. Working over models whose mean is given by a linear scaling of the input X through a parameter θ, we show for the binomial and negative binomial models, that the derivative of the input–output mutual information is given through a Bregman divergence where the arguments are the mean of the model and its conditional estimate. This condition gives rise to relationships that are of the same kind as the "I-MMSE" and the "I-MMLE" found initially for the Gaussian and Poisson models. Similar expressions are developed for the relative entropy, where the arguments of the Bregman divergence are the conditional mean estimate of the model θX assuming that $X \sim P_X$ and its correspondent mismatched version when $X \sim Q_X$. Making the input scaling factor tends to zero, we show that the derivative of the input–output mutual information is proportional to the expectation of a Bregman divergence between the input X and its mean $\mathsf{E}[X]$. This behavior is similar to that proved for the case of the Gaussian channel where, when the signal to noise ratio goes to zero, the derivative of the mutual information tends to the variance of the input.

Furthermore, using an arbitrary input preprocessing function that is not necessarily linear, we prove that several scenarios lead to information–estimation expressions that are given through Bregman divergences even though this is not always the case. In those cases where the information–estimation relationship is given through the minimum of a Bregman divergence, an information–estimation relationship similar to the "I-MMSE" and "I-MMLE" relationships can be stated. Finally, we provide conditions for which the results obtained for the binomial and negative binomial models converge asymptotically to information–estimation relationships over the Poisson model. This technique let us present connections between information and estimation over the Poisson model that cover wider scenarios than those studied so far.

Resumen

En esta tesis se estudian diversas relaciones entre la teoría de la estimación y la teoría de la información sobre transformaciones aleatorias donde la relación entre la entrada y la salida está dada a través de distribuciones de probabilidad del tipo binomial, binomial negativo y Poisson.

Las primeras expresiones encontradas que relacionan estos dos campos datan de la década de lo 60's cuando Duncan [14, 15] provó que la información mutua entre la entrada y la salida de un canal del tipo Gaussiano equivale a la integral en el tiempo del mínimo error cuadrático medio entre la entrada y su estimación condicionada a la observación de la salida.

Estudios posteriores, hechos por Zakai[4], Kadota [30], Mayer-Wolf [40] y Lipster [38], -entre otros- sugirieron la existencia de relaciones más fuertes entre la teoría de la estimación y la teoría de la información que tenían validez sobre un amplio espectro de transformaciones aleatorias. A la fecha, el resultado más destacado concerniente a las relaciones entre estas dos teorías establece que, sobre un canal del tipo Gaussiano, la derivada de la información mutua con respecto a la relación señal a ruido es proporcional al error cuadrático medio obtenido entre la entrada y su correspondiente estimación a través de la media condicionada al valor de la salida. En este caso, una propiedad fundamental de la conexión entre estimación e información se basa en que el valor del error cuadrático medio es mínimo cuando la estimación de la entrada se hace a través de la media condicional, lo que da lugar a lo que es conocido en la literatura como la relación "I-MMSE"[5]. Expresiones similares entre información y estimación son obtenidas para el caso de la entropía relativa entre dos distribuciones obtenidas a la salida del canal Gaussiano, donde de nuevo, el nexo, estimación–información está dado a través del error cuadrático.

En el contexto del canal del tipo Poisson –usado frecuentemente en el modelado de canales ópticos– las relaciones entre información y estimación encontradas hasta el momento tienen forma similar a las encontradas en el caso del canal Gaussiano donde la única diferencia radica en la función de pérdida utilizada. En otras palabras, mientras que en caso del canal

[4]En [23] se indica que algunos de los resultados obtenidos inicialmente por Duncan, relacionados con las relaciones entre la teoría de la información y la teoría de la estimación fueron también obtenidos por Zakai de manera independiente.

[5]En esta notación, la "I" representa "Información" y "MMSE" representa "Mínimo error cuadrático medio" por sus siglas en inglés.

Gaussiano la relación información–estimación está dada a través del error cuadrático, en el caso del canal Poisson, la relación está dada a través de la divergencia de Bregman que aparece en la representación exponencial de la distribución Poisson. Como consecuencia de esto, debido a las propiedades de las divergencias de Bregman se concluye que la función de pérdida es mínima cuando se utiliza para comparar la diferencia entre la entrada del canal Poisson y su estimación a través de la media condicional. Este comportamiento da lugar a lo que es conocido en el ámbito del canal Poisson a la relación "I-MMLE"[6], de manera análoga a la relación "I-MMSE" en el caso del canal Gaussiano.

Basados en los resultados anteriores en esta tesis son presentadas relaciones similares en el contexto de los modelos binomial y binomial negativo. En cada modelo, asumiendo un pre-procesado determinista de la entrada, son demostradas diversas relaciones entre información y estimación que están dadas en términos de estadísticos de la entrada y sus correspondientes estimas condicionales. En algunos casos, dichas relaciones son a través de divergencias de Bregman aunque ése no es siempre el caso. Cuando el pre–procesado de la entrada es lineal se muestra que para los modelos binomial y binomial negativo, la derivada de la información mutua está dada a través de divergencias de Bregman donde los argumentos de la función de pérdida son la media del modelo (que depende de la entrada) y su media condicional. Estas características dan lugar a relaciones de la misma naturaleza que las denominadas "I-MMSE" en el caso del canal Gaussiano y la "I-MMLE" en el caso del canal Poisson. Expresiones similares son obtenidas en el caso de la entropía relativa. Posteriormente, cuando el parámetro que afecta linealmente a la entrada se hace tender a cero, es demostrado que la derivada de la información mutua es proporcional al valor esperado de la divergencia de Bregman (usada para expresar la derivada de la información mutua en el modelo Poisson) entre la entrada y su media. Este comportamiento es similar al obtenido en el caso del canal Gaussiano donde el valor de la derivada de la información mutua cuando la relación señal a ruido tiende a cero está dado por el valor esperado de la distancia Euclídea entre la entrada y su media (varianza).

Utilizando un pre-procesado arbitrario en la entrada que no es lineal necesariamente, es mostrado que diversos escenarios dan lugar a relaciones entre información y estimación a través de divergencias de Bregman. Cuando dichas divergencias de Bregman son minimizadas se

[6]En este caso, la "I" se refiere a "Información" y "MMLE" se refiere a "Mínimo error de pérdida medio" por sus siglas en inglés.

puede establecer la existencia de una relación entre información y estimación del mismo tipo que las denominadas "I-MMSE" y "I-MMLE" estudiadas anteriormente. Para concluir, se presentan diversos escenarios en los modelos binomial y binomial negativo sobre los que las relaciones entre información y estimación encontradas convergen asintóticamente a las relaciones encontradas en el caso del modelo Poisson. Esta técnica permite la obtención de resultados, desconocidos hasta ahora para el modelo Poisson, a partir de los resultados obtenidos para los modelos binomial y binomial negativo.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The problem of communication defined by Shannon on his seminal papers [50, 51, 52] as "that of reproducing at one point either exactly or approximately a message selected at another point" found a mathematical translation through the concept of mutual information via the coding theorem. Roughly speaking, such theorem states that, for a random transformation between two random variables X and Y, the maximum rate at which reliable communication is possible, known as capacity, is given by the maximum of the mutual information between X and Y, where the maximization space is with respect to the set of all distributions of X while the random transformation $P_{Y|X}$ is kept fix. Denoted by $I(X;Y)$, the input–output mutual information between X and Y is defined as follows.

Definition 1. *[13] The mutual information $I(X;Y)$ between two random variables X and Y is defined as,*

$$I(X;Y) = D(P_{XY}||P_X \times P_Y) \tag{1.1}$$

where, for two probability measures P and Q, defined on the same measurable space, their relative entropy $D(P||Q)$ is defined as

$$D(P||Q) = \begin{cases} \int \left[\log \frac{dP}{dQ}\right] dP & \text{if} \quad P \ll Q \\ \infty & \text{otherwise.} \end{cases} \tag{1.2}$$

Remark 1. *Formally, for a continuous output Y i.e., if for every x $P_{Y|X}(\cdot|x)$ is absolutely continuous with respect to the Lebesgue measure, the mutual information can be expressed as:*

$$I(X;Y) = \mathsf{E}\left[\log \frac{f_{Y|X}(Y)}{f_Y(Y)}\right], \tag{1.3}$$

where $f_{Y|X}$ stands for the probability density function of a random transformation between an input X and a continuos output Y. Similarly, in the case of a discrete output Y, (1.1) *is equivalently expressed as:*

$$I(X;Y) = \mathsf{E}\left[\log \frac{P_{Y|X}(Y|X)}{P_Y(Y)}\right]. \tag{1.4}$$

Estimation theory deals mainly with ways of estimating information based on observations of random variables. Towards this end, there are several known estimation rules, *i.e.*, functions of the observed variable that give an estimation of the transmitted information (input random variable) according to some restrictions that account for constraints in complexity, error probability, optimality, etc. In this setting, suppose that our optimality criterion is the mean square error, *i.e.*, for a given estimate of the transmitted signal $\hat{x} = h(y)$, our goal is to minimize the function,

$$\mathsf{mse} = \mathsf{E}[(X - \hat{X})^2]. \tag{1.5}$$

It is well known (see for example [31, p. 313]) that the function $h(y)$ that minimizes the mean square error is given by the mean of the posterior probability density function $P_{X|Y}$, *i.e.*, $h(y) = \mathsf{E}[X|Y = y]$ which is referred to as the conditional mean estimate. Mathematically speaking, the previous idea is expressed as follows,

$$\mathsf{E}[X|Y = y] = \arg\min_{\hat{X}(y)} \mathsf{E}[(X - \hat{X}(y))^2]. \tag{1.6}$$

In recent years, multiple relationships between information and estimation have played a fundamental role in the development of theoretical and practical advances in the study of communication systems. Initially in 2004, Guo, Shamai and Verdú, stated a key information-estimation relationship over the Gaussian channel [23]. Its simplicity and massive usage of the Gaussian distribution to model several scenarios in the problem of communication highlight the importance of this expression. Based on the previous key quantities used in the estimation and information fields, in the following section we proceed to state a link between them for the Gaussian channel.

1.1 Notation conventions

In this chapter and throughout the thesis, we use the following notation. Let X and Y be the input and output, respectively, of a random transformation

$P_{Y|X}$ that depends on a parameter θ. Each domain set is denoted as $\mathcal{X}$, $\mathcal{Y}$ and Θ. Let P_X and Q_X denote two distributions over the input X that are independent of θ. Additionally, let P_Y and Q_Y be two probability distributions induced at the output Y by each input distribution P_X and Q_X respectively, through the random transformation $P_{Y|X}$. Unless otherwise stated, the output distributions P_Y and Q_Y depend implicitly on θ.

1.2 Gaussian Channel

The Gaussian channel in its simplest form is defined as follows. For a real-valued input X, the output Y is given by,

$$Y = \sqrt{\mathsf{snr}}X + N, \tag{1.7}$$

where, snr stands for the signal to noise ratio of the channel when $\mathsf{E}[X^2] = 1$ and N is Gaussian distributed with zero mean and unit variance. In this scenario, regardless of the input distribution, the link between information and estimation is given by [23],

$$\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}} I(X;Y) = \frac{1}{2}\mathsf{E}\left[(X - \mathsf{E}[X|Y])^2\right], \tag{1.8}$$

$$\triangleq \frac{1}{2}\mathsf{mmse}(\mathsf{snr}), \tag{1.9}$$

where $\mathsf{mmse}(\cdot)$ stands for the Minimum Mean Square Error in concordance with (1.5) and (1.6). The left hand side (LHS) of (1.8) represents the derivative of the input-output mutual information and the right hand side (RHS) represents the square distance between the input X and its estimate through the conditional mean estimate.

Several additional expressions can be derived upon the landmark expression stated in (1.8). As is shown in (1.1), an information theory concept closely related with the mutual information is the relative entropy between two distributions P and Q defined in (1.2). Although the relative entropy is not symmetric, in some scenarios it is considered as a measure of distance between the distributions P and Q, given that it is always positive [13, Theorem 9.6.1]. Hence, a second expression linking information measures with the estimation theory field through the conditional mean estimate is stated as follows in the context of the relative entropy. Suppose that P_X and Q_X are two distributions of X such that $\mathsf{E}[X^2]$ and $\mathsf{E}_Q[X^2]$ are finite, where we use the subscript Q to indicate that $X \sim Q_X$ and analogously omit it when $X \sim P_X$. Then, for an input-output relationship

given by (1.7) the derivative of the relative entropy between the output distributions P_Y and Q_Y with respect to changes in the snr parameter is given by [62],

$$\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}} D(P_Y||Q_Y) = \frac{1}{2}\mathsf{E}\left[\left(\mathsf{E}[X|Y] - \mathsf{E}_Q[X|Y]\right)^2\right], \tag{1.10}$$

where $\mathsf{E}_Q[X|Y]$ represents the conditional mean estimate of X when $X \sim Q_X$. Notice that (1.10) has a similar structure to that found in (1.8), in the sense that it is given in terms of the square distance between the conditional mean estimates $\mathsf{E}[X|Y]$ and $\mathsf{E}_Q[X|Y]$.

Additionally, notice that, based on the expression given for the mutual information and the relative entropy, we get that

$$\frac{1}{2}\mathsf{E}\left[\left(X - \mathsf{E}_Q[X|Y]\right)^2\right] = \frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}}\left(I(X;Y) + D(P_Y||Q_Y)\right), \tag{1.11}$$

i.e., the expectation of the difference between the input X and the conditional mean estimate $\mathsf{E}_Q[X|Y]$ (when assumed a mismatched prior Q_X), is proportional to the derivative with respect to the snr of the sum of the mutual information and relative entropy.

Several applications arise from the aforementioned information-estimation relationships, such as power allocation over parallel Gaussian channels [39, 44], multiuser detection [26] and nonlinear filtering [23, 66]. Furthermore, the expression given in (1.8) can be used to state a proof of Shannon's entropy power inequality [63, 24, 59] and to study the capacity region of several multiuser channels [58, 73, 10]. The above expressions can also be generalized to multiple input multiple output (MIMO) scenarios over vector valued channels [23] and to continuous time channel models [23, 66].

1.3 Poisson Channel

In the search of multiple expressions relating the information field with the estimation field, one natural model to consider is the Poisson channel. This channel is frequently used to model optical communication systems where the transmitter sends information by modulating the intensity of an optical signal, while the receiver tries to make an informed guess of the transmitted message by using the arrival moments of the individual photons [70, 71].

In the following, we briefly state the Poisson model considered and then present analogous information-estimation relationship to those stated for the Gaussian channel.

Let X and Y be the input and output, respectively, of a random transformation where, for a given $X = x$, the conditional distribution[1] of the channel is Poisson with mean $\theta X > 0$, *i.e.*,

$$P_{Y|X}(y|x) = \frac{(\theta x)^y}{y!} e^{-\theta x}, \quad y = 0, 1, 2, \ldots \tag{1.12}$$

where θ stands for an input scaling factor playing a similar role as the snr in the context of the Gaussian channel.

Guo, Shamai and Verdú, after presenting their seminal work on the information-estimation relationships for the Gaussian channel, stated further expressions for the Poisson channel over discrete and continuous models. In this case, the counterpart relationship to (1.8) relating the mutual information with the conditional mean estimate can be stated as follows. Let X be a positive random variable such that $\mathsf{E}[X \log X] < \infty$. Then, considering a Poisson channel (1.12) with mean $\theta X > 0$ we have that [25],

$$\frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) = \mathsf{E}\left[X \log \frac{X}{\mathsf{E}[X|Y]}\right], \tag{1.13}$$

for all $\theta > 0$. Thinking in terms of loss functions, one question that naturally arises is whether the function inside the expectation in (1.13) corresponds to a function with similar properties to those fulfilled by the square distance which is the loss function used in the definition of the mmse (1.5). To tackle this issue, lets define the function $\ell_P : (0,\infty) \times (0,\infty) \to (0,\infty)$ as follows,[2]

$$\ell_P(a, \hat{a}) = a \log \frac{a}{\hat{a}} - (a - \hat{a}), \quad a, \hat{a} \in (0,\infty). \tag{1.14}$$

In the context of the Poisson channel with input X and output Y, for a given realization of the input, let $a = x$, and for a given realization of the output let $\hat{a}(y) = \mathsf{E}[X|Y = y]$. Based on the function ℓ_P, calculating its expected value we get that,

$$\mathsf{E}\left[\ell_P(X, \mathsf{E}[X|Y])\right] = \mathsf{E}\left[X \log \frac{X}{\mathsf{E}[X|Y]} - (X - \mathsf{E}[X|Y])\right] \tag{1.15}$$

$$= \mathsf{E}\left[X \log \frac{X}{\mathsf{E}[X|Y]}\right] \tag{1.16}$$

$$= \frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y). \tag{1.17}$$

[1] Throughout this monograph and unless other condition is stated, we assume that all the random transformations $P_{Y|X}$ considered depend on a parameter θ over which we take derivatives. Dependency of $P_{Y|X}$ on θ is implicit.

[2] Here, the subscript P stands for "Poisson".

In order to state several properties of the function ℓ_P, lets define the function $\phi(a) = a \log a$ with $\theta \in (0, \infty)$. For all $a > 0$, notice that $\frac{\mathrm{d}^2\phi(a)}{\mathrm{d}a^2} > 0$, proving by this way the strictly convex behavior of the function $\phi(\cdot)$. The first order Taylor approximation of the function $\phi(a)$ around a point $\hat{a}$ is given by the function,

$$\phi(\hat{a}) + \frac{\mathrm{d}}{\mathrm{d}a}\phi(a)\bigg|_{a=\hat{a}}(a - \hat{a}). \tag{1.18}$$

The difference between the function $\phi(a)$ and its first order Taylor approximation around the point $\hat{a}$, called error, is given by,

$$\begin{aligned}
\mathsf{error}(a, \hat{a}) &= \phi(a) - \phi(\hat{a}) - (a - \hat{a})\frac{\mathrm{d}}{\mathrm{d}a}\phi(a)\bigg|_{a=\hat{a}} && (1.19)\\
&= \phi(a) - \phi(\hat{a}) - (a - \hat{a})\left(\log \hat{a} + 1\right) && (1.20)\\
&= a \log a - a \log \hat{a} - (a - \hat{a}) && (1.21)\\
&= \ell_P(a, \hat{a}). && (1.22)
\end{aligned}$$

Such a function is called a Bregman divergence. Bregman divergences are functions that quantify the difference between the value of a strictly convex function and its first order Taylor approximation. Deeper mathematical treatment of this kind of functions is given in Chapter 3. Comparing (1.22) with (1.13) we see that the derivative of the relative entropy corresponds to the Bregman divergence associated with the function $\phi(a) = a \log a$. Notice that the strictly convexity of the function $\phi(\cdot)$ ensures that the value of the function $\ell_P(a, \hat{a})$ is always positive.

Along the same lines, the square distance considered for the Gaussian channel belongs to the set of Bregman divergences. Indeed, for $\phi(a) = a^2$, the difference between this function and its first order Taylor approximation based on the point $\hat{a}$ is given by,

$$\begin{aligned}
\ell_G(a, \hat{a}) &= \phi(a) - \phi(\hat{a}) - (a - \hat{a})\frac{\mathrm{d}}{\mathrm{d}\theta}\phi(a)\bigg|_{a=\hat{a}} && (1.23)\\
&= a^2 - \hat{a}^2 - (a - \hat{a})2\hat{a} && (1.24)\\
&= a^2 - 2\hat{a}a + \hat{a}^2 && (1.25)\\
&= (a - \hat{a})^2, && (1.26)
\end{aligned}$$

where we use the subscript G to make reference that ℓ_G corresponds to the loss function used over the Gaussian channel to describe information

measures. One consequence of the functions ℓ_G and ℓ_P being Bregman divergences, explored in more detail in Chapter 3, is that,

$$\arg\min_{z\in\sigma(Y)} \mathsf{E}\left[\ell(X,z)\right] = \mathsf{E}[X|Y=y], \tag{1.27}$$

where $\sigma(Y)$ stands for the σ-algebra generated by the output of the channel Y and ℓ is any Bregman divergence. In other words, the conditional mean estimate of X based on an observation of Y minimizes the expected loss of the functions ℓ_G and ℓ_P. This behavior implies that the derivative of the mutual information over the Gaussian and Poisson channels corresponds to the minimum value of a given loss function. This leads to the "I-MMSE" relationship pertaining to the Gaussian channel defined in [23] and its analogous for the Poisson channel defined by Atar *et al.* in [3] known as the "I-MMLE" relationship.

Additional information measures over the Poisson channel gave rise to an alternative representation in terms of estimation quantities depending on conditional mean estimates and its mismatched versions. Let P_X and Q_X be two distributions over an input random variable X that is bounded. Each distribution obtained at the output of the Poisson channel with mean θX is denoted by P_Y and Q_Y, respectively. The Poisson counterpart to the expression given in (1.10) for the Gaussian channel, regarding the derivative of the relative entropy is given by [3],

$$\frac{\mathrm{d}}{\mathrm{d}\theta} D(P_Y||Q_Y) = \mathsf{E}[\ell_P(\mathsf{E}[X|Y], \mathsf{E}_Q[X|Y])]. \tag{1.28}$$

To complete the analogy with the Gaussian channel, notice finally that

$$\mathsf{E}\left[\ell_P(X, \mathsf{E}_Q[X|Y])\right] = \frac{\mathrm{d}}{\mathrm{d}\theta}\left(I(X;Y) + D(P_Y||Q_Y)\right). \tag{1.29}$$

The Poisson channel has motivated research in several directions in the past. A comprehensive history of these results can be found in [25]. Additionally, the information–estimation relationship found in [3] was applied in [34] to determine the secrecy capacity of the degraded Poisson wiretap channel, which essentially consists of determining the maximum achievable rate between a transmitter and an intended receiver while ensuring that only a negligible amount of information is leaked to an eavesdropper. Moreover [65], proposes a generalization of the classical Bregman divergence in order to produce a vectorial version of (1.10) and (1.28) given in [62] and [3], respectively. It is worth to point out here that the vector Poisson channel has several applications in fields such as medical imaging [17] or document classification [72].

1.4 A Common Framework

As a consequence of the functions ℓ_P and ℓ_G being Bregman divergences, we highlight the following properties:

- ▷ Several information measures find an alternative representation through the expectation of certain functions that depend on input statistics and conditional mean estimates. Those representations can be associated with loss functions due to the fact that the connection between information and estimation is given in terms of Bregman divergences.

- ▷ The derivative of the mutual information with respect to the input scaling factor is equal to the minimum of the expected loss of a Bregman divergence.

- ▷ Due to the strict positiveness of Bregman divergences, the derivative of the mutual information and the derivative of the relative entropy are increasing in the parameters considered over each model. Conditions to establish the monotonicity of the mutual information when expressed in terms of conditional quantities are studied in [45]. They give rise to applications in the field of network information theory [19]. Specifically, in a broadcast channel setting, the increasing nature of the mutual information gives rise to the "More Capable" Broadcast Channels for which the capacity region is known.

- ▷ Assuming a mismatched distribution Q_X, the expectation of the loss function $\mathsf{E}\left[\ell(X, \mathsf{E}_Q[X|Y])\right]$ is proportional to the sum of the derivative of the relative entropy and the derivative of the mutual information. Based on the previous assertion, the sum of the relative entropy and the mutual information gets the following alternative representation,

$$I(X;Y) + D(P_Y||Q_Y) = \int_0^\theta \mathsf{E}\left[\ell\left(X, \mathsf{E}_Q[X|Y]\right)\right] d\gamma \tag{1.30}$$

  which is increasing in θ.

- ▷ All the expressions obtained over a discrete time framework can be translated akin to the continuous time case.

1.5 Scope of this thesis

Based on the information–estimation expressions found for the Gaussian and Poisson channels, we observe that expressions given in (1.8) and (1.13) to describe the derivative of the mutual information over each channel can be represented through a unique relation;

$$\frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) \propto \frac{1}{\theta} \mathsf{E}[\ell(\theta X, \mathsf{E}[\theta X|Y])]. \tag{1.31}$$

Note that the arguments of the loss function are given by the mean of the channel θX and its conditional mean estimate $\mathsf{E}[\theta X|Y]$. In the case of the Gaussian channel, the loss function ℓ corresponds to the square distance denoted by ℓ_G and in the case of the Poisson channel the loss function is the one defined in (1.14). Similarly, based in (1.10) and in (1.28), for the relative entropy we have that,

$$\frac{\mathrm{d}}{\mathrm{d}\theta} D(P_Y||Q_Y) \propto \frac{1}{\theta} \mathsf{E}[\ell(\mathsf{E}[\theta X|Y], \mathsf{E}_Q[\theta X|Y])], \tag{1.32}$$

where the unique difference with the expression given for the mutual information lies in the arguments of the loss function. In this thesis, we explore whether the expressions given in (1.31) and (1.32) continue to hold in other scenarios different to the ones previously treated. This would allow us to translate to other models all the functional properties proven for the behavior of the mutual information and relative entropy for the Gaussian and Poisson channels.

Specifically, the core of this thesis is the study of several information-estimation relationships over different models, predominantly, Gaussian, Poisson, binomial and negative binomial.

Initially in Chapter 2 we present the state of the art in the study of the representation of information measures in terms of estimation quantities over a wide range of scenarios. In addition, we show advantages and disadvantages of the results known so far.

Based on the importance of the expressions found in (1.8), (1.10) for the Gaussian channel and in (1.13), (1.28) for the Poisson channel and its representation in terms of loss functions, in Chapter 3 we provide a self-contained treatment for the set of Bregman divergences jointly with their main properties. One characteristic that arises naturally is their close relationship with exponential family distributions. This connection has lead to several applications in fields such as Machine Learning [5], [68], [27],

Statistical Learning [18] and Optimization [9]. Even though the presentation of the binomial and negative binomial models is relegated to Chapters 4 and 5, in Chapter 3 we also show the exponential representation of such distributions in order to make comparisons in future sections.

In Chapter 4 we formally present the structure of the binomial model considered in this thesis. Initially we state the information-estimation relationship that constitutes the binomial counterpart to the relationship stated in (1.8) and in (1.13) for the Gaussian and Poisson channels, respectively. The relationship is given in terms of a Bregman divergence, which for the mutual information attains its minimum value at the conditional mean estimate, like the "I-MMSE" relationship for the Gaussian channel and the "I-MMLE" relationship for the Poisson channel. Furthermore, an information-estimation relationship through a Bregman divergence is also derived for relative entropy. Finally, we show that there exists a Bregman divergence for which its expectation in the mismatched case, corresponds to the derivative of the sum between the relative entropy and the mutual information.

In Chapter 5 we state for the negative binomial model similar results to those found for the binomial model. Specifically, we show that all the results found in Chapter 4 translate verbatim to the negative binomial model using an alternative Bregman divergence.

Assuming a specific configuration in the constitution of the binomial model, we show in Chapter 6 that asymptotically,[3] results found for the binomial model converge to those obtained for the Poisson model over different scenarios. We conclude this thesis in Chapter 7 with some remarks and conclusions build upon the results found for the binomial, negative binomial and Poisson models. Finally, additional results are presented, where the interplay between the information theory and the estimation theory leads to striking questions.

[3]Specifically this statement holds when the number of trials n used to generate the binomial model goes to infinite.

Chapter 2

State of the Art

Posterior to the landmark expression given in (1.8), several attempts have been done in order to explore deeper relationships linking the information field with the estimation field. Most of the work done relies on the analysis of the information–estimation expressions over models where the noise is additive. Additional works have been done in order to find an information–estimation expression that works over all kind of channels. In this case, the main issue with the expressions found lies on the difficulty to extract useful conclusions from the results obtained.

A striking property that is worth pointing out is the fact that the Gaussian and Poisson distributions belong to the set of infinitely divisible distributions. Specifically, a random variable X is infinitely divisible if, for all $n \in \mathbb{N}$, there exists a sequence of independent identically distributed (i.i.d) random variables $Y_1, \ldots, Y_n$ such that,

$$X \stackrel{d}{=} Y_1 + \cdots + Y_n, \tag{2.1}$$

where $\stackrel{d}{=}$ denotes equality in distribution. To see further details about this type of random variables and its properties see [1, 32, 33]. Recently, it has been brought to our attention the publication of a paper that studies relations between information and estimation over Lévy channels, which rely heavily on infinitely divisible distributions. We refer the interested reader to [28, 29].

This chapter is mainly dedicated to the presentation of different information estimation relationships studied previously. In Section 2.1 we show an expression for the derivative of the mutual information in terms of conditional estimates for the case of additive noise channels. Section 2.2 deals with a similar scenario where it is assumed that the distribution of the

noise belongs to the set of exponential family distributions. The last scenario considered in this chapter, in Section 2.3, states an information estimation expression that works over a wide range of random transformations, and is due to Palomar *et al.* [42]. However, the gain in generality comes at the cost of increased complexity of the result. At the end, we provide a section with a set of conclusions derived upon those results illustrated along this chapter.

2.1 Additive Noise Channels

In this section, we show different information-estimation results regarding additive noise channels due to Guo *et al.* [22]. Assume that the output of a random transformation $P_{Y|X}$ is given by the following expression,

$$Y = f(\theta, X) + W, \tag{2.2}$$

where W is an arbitrary continuous random variable that is independent of X and $f(\theta, X)$ is a measurable deterministic preprocessing of the input X that is differentiable with respect to the parameter θ. In this framework, several regularity conditions are assumed in order to guarantee the validity of the results obtained. Suppose that the following conditions are satisfied:

A1. The input P_X and the noise probability density function P_W are fixed and independent of θ.

A2. The expression $\frac{\mathrm{d}}{\mathrm{d}w} P_W(w)$ is uniformly continuous over the support of P_W.

A3. The function $\mathsf{E}\left[\frac{\mathrm{d}}{\mathrm{d}W} \log P_W(W)|Y = y\right]$ exists for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

A4. For every θ and $\omega(y)$ integrable, we have

$$\left|\log P_Y(y) \frac{d}{d\theta} P_Y(y)\right| \leq \omega(y) \tag{2.3}$$

in a neighborhood of θ_0.

Notice that these technical conditions are satisfied by the most common distributions used in communications.

Theorem 1 ([22]). *Consider a general additive noise channel given by* (2.2)*. Then, for every input distribution and noise distributions that satisfy*

conditions A1-A4,

$$\frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) = -\mathsf{E}\left[\mathsf{E}\left[f'(\theta,X)|Y\right]\mathsf{E}\left[\frac{\mathrm{d}}{\mathrm{d}W}\log P_W(W)\middle|Y\right]\right] \tag{2.4}$$

where $f'(\theta,X)$ stands for the derivative of the preprocessing function with respect to θ, i.e.*,*

$$f'(\theta,X) \triangleq \frac{\partial}{\partial\theta} f(\theta,X). \tag{2.5}$$

The previous expression has several advantages. First, it covers a wide set of scenarios studied in the communication systems. Second, the expression found holds regardless of the input distribution; basically it is composed of two terms involving conditional estimates, one depending on the estimate of the derivative of the input preprocessing and the other depending on the distribution of the noise. Additionally, as was pointed out before, the regularity conditions *A1-A4* are sufficiently mild for Theorem 1 to be useful. The mayor drawback in the expression given by Theorem 1 is the fact that in most of the cases, the term $\frac{\mathrm{d}}{\mathrm{d}w}\log P_W(w)$ depends on the output Y which means that the expression at the RHS of (2.4) not only depends on conditional mean estimates of the input, in contrast to the initial expressions given in (1.8) and (1.13).

Example 1 (Scalar Gaussian Channel with variance σ^2).
Consider a Gaussian channel where the input-output relationship is given by,

$$Y = \sqrt{\theta}X + W \tag{2.6}$$

where W is a Gaussian distribution with zero mean and variance σ^2. In this case notice that,

$$\frac{\mathrm{d}}{\mathrm{d}w}\log P_W(w) = -\frac{w}{\sigma^2}. \tag{2.7}$$

Therefore, applying Theorem 1 we get that,

$$\frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) = \frac{1}{2\sigma^2\theta^{1/2}} \mathsf{E}\left[\mathsf{E}[X|Y]\mathsf{E}\left[W|Y\right]\right] \tag{2.8}$$

$$= \frac{1}{2\sigma^2\theta^{1/2}} \mathsf{E}\left[\mathsf{E}[X|Y]\mathsf{E}\left[Y - \theta^{1/2}X|Y\right]\right] \tag{2.9}$$

$$= \frac{1}{2\sigma^2\theta^{1/2}} \mathsf{E}\left[\mathsf{E}[XY] - \theta^{1/2}\mathsf{E}\left[X|Y\right]^2\right] \tag{2.10}$$

$$= \frac{1}{2\sigma^2} \mathsf{E}\left[X^2 - \mathsf{E}\left[X|Y\right]^2\right] \tag{2.11}$$

$$= \frac{1}{2\sigma^2} \mathsf{mmse}(\theta). \tag{2.12}$$

Notice that, in order to obtain an information–estimation expression that is determined entirely by the input and its conditional mean estimates, the steps from (2.9) to (2.12) are required. When dealing with other models different to the Gaussian, these steps may not hold.

2.2 Additive Channels with Noise from the Exponential Family

In the following section we state some results given by Raginsky and Coleman [45], pertaining to additive noise channels where the distribution of the noise is a member of the exponential family. Consider the following input-output relationship,

$$Y = X + W, \tag{2.13}$$

where W is a random variable that belongs to the exponential family, *i.e.*, for a given function $\rho : \mathcal{Y} \to \mathbb{R}$,

$$P_W(y) = e^{-\theta\rho(y) - A(\theta)} \tag{2.14}$$

which depends implicitly on θ and ρ. The quantity $A(\theta)$ known as the log partition function or cumulant function is defined as,

$$A(\theta) = \log Z(\theta) \tag{2.15}$$

where,

$$Z(\theta) = \int e^{-\theta\rho(y)}\mu(\mathrm{d}y) < \infty, \quad \theta \in (0, \infty), \tag{2.16}$$

with μ a $\sigma-$finite fixed reference measure defined over a measurable space $(\mathcal{Y}, \mathcal{F}_{\mathcal{Y}}, \mu)$ that is translation invariant, *i.e.*, for any $A \in \mathcal{F}_{\mathcal{Y}}$ and $y \in \mathcal{Y}$, $\mu(A - y) = \mu(A)$. Based on the input-output structure of the model, the conditional distribution of the channel is given by,

$$P_{Y|X}(y|x) = \frac{e^{-\theta\rho(y-x)}}{Z(\theta)}. \tag{2.17}$$

Even though it is not explicit in the notation, the conditional distribution $P_{Y|X}$ depends on the parameter θ, denominated natural parameter. In what follows we study the behavior of the mutual information with respect to changes in the value of the natural parameter. Several known channels treated in the communications field match with the scenario described by (2.13). The Gaussian channel is one of those models, where $\rho(y) = y^2$ and $Z(\theta) = \sqrt{\pi/\theta}$. Additionally, using a module 2 arithmetic, the Binary Symmetric Channel (BSC) matches with the model given in (2.13) with $\rho(y) = y$ and $Z(\theta) = 1 + e^{-\theta}$. To state the main result, *i.e.*, the derivative of the input–output mutual information, assume that the input distribution P_X is such that, for all θ in some neighborhood of every $\theta_0 > 0$,

$$\left| \frac{\mathrm{d}}{\mathrm{d}\theta} A(\theta|y) P_Y(y) \right| \leq \omega(y) \tag{2.18}$$

for some integrable function $\omega(y)$, where

$$A(\theta|y) \triangleq \log \int e^{-\theta\rho(y-x)} \mathrm{d}P_X(x). \tag{2.19}$$

In [45] it is shown that the technical condition (2.18) holds for a wide variety of cases, and is used for the sake of formality. In the following theorem we state an information-estimation relationship for additive noise channels where the distribution of the noise belongs to the exponential family.

Theorem 2 ([45]). *Let the input distribution P_X be such that condition* (2.18) *holds. Then,*

$$\frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) = \theta \frac{\mathrm{d}^2}{\mathrm{d}\theta^2} A(\theta) + cov\{\mathsf{E}[\rho(Y - X)|Y], A(\theta|Y)\}, \tag{2.20}$$

where the expectation is with respect to the distribution P_{XY}.

In light of the previous result there are several comments to make: even though the expression obtained for the derivative of the mutual information

holds for a wide range of channels, the main drawback is that (2.20) depends on the output Y not only through conditional mean estimates of the input. In other words, comparing (2.20) with the expressions obtained for the Gaussian channel in (1.8) and in (1.13) for the Poisson channel, the main difference relies on the fact that (1.8) and (1.13) only depend on input statistics and conditional mean estimates of the input, meanwhile, (2.20) depends directly on the expected value of the function $\rho(Y - X)$. This explicit dependency on Y, for example in the case of the Gaussian channel, masks the fact that the derivative of the mutual information corresponds to the minimum square error. Additionally, in those cases where the output Y takes values on positive and negative sets, it is harder to state the monotonicity of the mutual information based on the expression found for its derivatives.

2.3 Representation of Mutual Information via Input Estimates

In this section we show a general information-estimation relationship that works over any kind of channels without assuming any special structure. After presenting the main result stated initially in [42] we discuss several benefits and drawbacks from the obtained expressions.

Consider again a random transformation $P_{Y|X}$ that depends on a parameter θ (not necessarily from the exponential family or additive). In this section, unless otherwise is stated, all distributions over the output alphabet $\mathcal{Y}$ depend on the parameter θ through the conditional $P_{Y|X}$. Before proceeding formally with the results, assume the following regularity conditions:

- For a given $x \in \mathcal{X}$ and a distribution Q_Y independent of θ,

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathsf{E}_Q\left[P_{Y|X}(Y|x)\right] = \mathsf{E}_Q\left[\frac{\mathrm{d}}{\mathrm{d}\theta}P_{Y|X}(Y|x)\right]. \tag{2.21}$$

- For a given $y \in \mathcal{Y}$ and P_X independent of θ,

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathsf{E}\left[P_{Y|X}(y|X)\right] = \mathsf{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}P_{Y|X}(y|X)\right]. \tag{2.22}$$

These assumptions are required in order to state the proof of Theorem 3. They are satisfied by most of the distributions used in communications

as is illustrated in [42] through examples. Theorem 3, due to Palomar and Verdú, characterizes the derivative of the input-output mutual information for an arbitrarily random transformation over arbitrary alphabets.

Theorem 3 ([42]). *Let $P_{Y|X}$ be a random transformation which is differentiable with respect to θ and let X be a random input with distribution P_X independent of θ. Then,*

$$\frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) = \mathsf{E}\left[\log P_{X|Y}(X|Y) \frac{\mathrm{d}}{\mathrm{d}\theta} \log P_{Y|X}(Y|X)\right], \tag{2.23}$$

where the expectation is with respect to the joint distribution P_{XY}.

In [42], Theorem 3 was specialized to the following class of channels.

Theorem 4 ([42]). *Let $P_{Y|X}$ be a differentiable random transformation that depends on θ, where the output alphabet is continuous. If*

$$\frac{\mathrm{d}}{\mathrm{d}\theta} P_{Y|X}(y|x) = -\Phi^{\theta}(x) \frac{\mathrm{d}}{\mathrm{d}y} P_{Y|X}(y|x), \tag{2.24}$$

where $\Phi^{\theta}(x)$ is some function of θ and x, then,

$$\frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) = \mathsf{E}\left[\Phi^{\theta}(x) \frac{\mathrm{d}}{\mathrm{d}y} \log P_{X|Y}(X|Y)\right]. \tag{2.25}$$

Several conclusions can be drawn from Theorems 3 and 4. To better illustrate them, consider the following example.

Example 2. *Let $P_{Y|X}$ be a Poisson channel* (1.12) *with mean θX. Based on Theorem 3 we have that,*

$$\frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) = \mathsf{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\left(\log P_{Y|X}(Y|X)\right) \log P_{X|Y}(X|Y)\right] \tag{2.26}$$

$$= \mathsf{E}\left[\frac{1}{P_{Y|X}(Y|X)} \frac{\mathrm{d}}{\mathrm{d}\theta}\left(P_{Y|X}(Y|X)\right) \log P_{X|Y}(X|Y)\right] \tag{2.27}$$

$$= \mathsf{E}\left[\left(\frac{Y}{\theta} - X\right) \log P_{X|Y}(X|Y)\right]. \tag{2.28}$$

Even though results stated in Theorems 3 and 4 work over a wide range of scenarios, in several cases the expression found involves the calculation of the posterior $P_{X|Y}$ which in general is difficult to obtain. Additionally, notice that the expression found in (2.28) differs from that given in (1.13) in

the sense that it still depends directly on the output Y. Further calculations are needed in order to obtain the expression given in (1.13).

In general, the results presented by Palomar and Verdú prove their advantages when they can be applied to any random transformation without assuming any structure in the input-output relationship, *i.e.*, it is not assumed an additive nature for the noise or even that it belongs to any special set of distributions, like the exponential family. For instance, Theorem 3 allows for a characterization of the behavior of the mutual information over scenarios such as the discrete memoryless channel, additive noise models, Poisson models, etc. The major difficulty in this case is the fact that the information estimation expression found, uses the calculation of the posterior $P_{X|Y}$ which is not easy to obtain in most of the cases.

2.4 Concluding Remarks

Throughout this chapter, we presented several attempts made in the past in order to find representations of information measures such as the mutual information and relative entropy in terms of conditional estimates. Initially, assuming certain structures in the input-output relationship, information-estimation relationships similar to those shown in the Introduction were presented. However, most of the expressions found depend on the output of the channel Y not only through conditional estimates of functions of the input. This is in contrast to the behavior of (1.8) and (1.13) which only depend on conditional mean estimates of the input. In Section 2.3 we presented an information-estimation relationship that works over any kind of random transformation and, under mild regularity conditions, let us express the derivative of the mutual information in terms of expectations of functions that depend on the calculation of the posterior $P_{X|Y}$, which generally are not easy to compute. Notice that the resulting expressions depend again on expectations of the output Y as is illustrated in Example 2. It is worth pointing out that expressions that depend only on input statistics and conditional mean estimates through the expectation of loss functions give rise to the "I-MMSE" and "I-MMLE" relationships. A key point in those relationships is that the expectation of such loss functions achieves its minimum value when they are evaluated at the conditional mean estimate of the input. This conclusion can not be stated from the expressions studied in this section when the information–estimation expression depends on the output Y not only through conditional estimates.

Chapter 3

Bregman Divergences

Bregman divergences were proposed by Lev M. Bregman in 1967 to develop solutions to convex optimization problems [9]. These class of functions have found several applications that lie in fields such as machine learning [5, 48, 27], estimation theory [4, 18] and computational geometry [41], among others. The structure of this chapter is as follows. In Section 3.1 we state the definition of a Bregman divergence jointly with a set of properties. Subsequently, in Section 3.2 we present a one–to–one relationship between the Bregman divergences and the exponential family distributions. This connection opens up several questions regarding the information estimation relationships over those models where the random transformation $P_{Y|X}$ belongs to the exponential family distributions.

3.1 Definition and Properties

Bregman divergences, posteriorly to their application to the solution of optimization problems, became popular due to their behavior when dealing with the expectation of such functions. Before stating this behavior mathematically together with other useful properties, we first provide some background.

Let $(\Omega, \mathcal{F}, P)$ be a probability space and let X be a $\mathcal{F}-$measurable random variable that we want to estimate. Let y be an observation of the random variable Y. The available information about X that can be obtained by observing y is represented by $\sigma(Y)$. Mathematically, $\sigma(Y)$ is the $\sigma-$algebra generated by Y and contains all Borel-measurable functions of Y.

Definition 2. *Let $\phi : \Omega \to \mathbb{R}$ be a continuously differentiable strictly convex*

function defined on a convex set $\Omega \subseteq \mathbb{R}^d$. *Consider two points* $a, \hat{a} \in \Omega$. *Then, the Bregman divergence between* a *and* $\hat{a}$, *associated with the function* $\phi(\cdot)$ *is defined by,*

$$\ell_\phi(a, \hat{a}) = \phi(a) - \phi(\hat{a}) - \langle a - \hat{a}, \nabla\phi(\hat{a})\rangle, \tag{3.1}$$

where, $\langle \cdot, \cdot \rangle$ *denotes the inner product and* $\nabla\phi(\hat{a})$ *represents the gradient of the function* $\phi(\cdot)$ *evaluated at* $\hat{a}$.

Note that $\ell_\phi(a, \hat{a})$ is equal to the difference between $\phi(a)$ and its first order Taylor approximation based on the behavior of the function $\phi(\cdot)$ at the point $\hat{a}$. This definition is graphically illustrated in Figure 3.1.

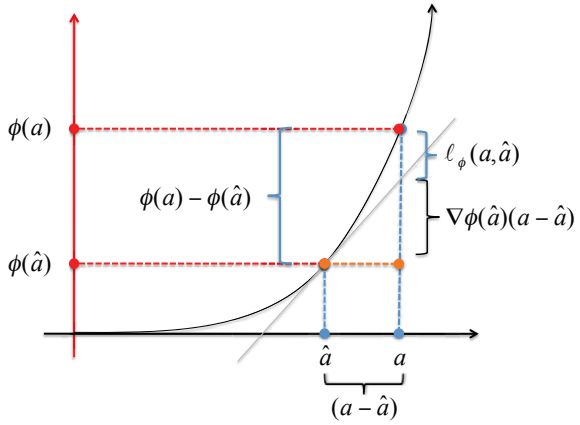

Figure 3.1: Bregman divergence definition.

Bregman divergences are pseudo-metrics that neither need to satisfy the triangle inequality nor need to be symmetric. In Theorem 5 we state the basic properties of the Bregman divergences family. We give posteriorly the proof to these properties for the sake of completeness.

Theorem 5 (Bregman divergences: basic properties). *Let* $\phi : \Omega \to \mathbb{R}$ *be a strictly convex function and let* ℓ_ϕ *be the Bregman divergence associated with* $\phi(\cdot)$. *Then,*

(i) $\ell_\phi(a, \hat{a}) \geq 0$ *for all* $a, \hat{a} \in \Omega$ *with equality if and only if* $a = \hat{a}$.

(ii) $\ell_\phi(a, \hat{a})$ *is strictly convex on its first argument.*

(iii) For $a, \hat{a}, \tilde{a} \in \Omega$,

$$\ell_\phi(a, \hat{a}) = \ell_\phi(a, \tilde{a}) + \ell_\phi(\tilde{a}, \hat{a}) - \langle a - \tilde{a}, \nabla\phi(\hat{a}) - \nabla\phi(\tilde{a})\rangle. \tag{3.2}$$

Proof. To prove Property *(i)*, let $a, \hat{a} \in \mathbb{R}^d$ with $a \neq \hat{a}$ and consider ϕ restricted to the line passing through them, *i.e.*, the function defined by $g(t) = \phi(t\hat{a} + (1-t)a)$, for $t \in [0,1]$. Additionally, notice that,

$$\frac{\mathrm{d}}{\mathrm{d}t} g(t) = \nabla\phi(t\hat{a} + (1-t)a)^T(\hat{a} - a). \tag{3.3}$$

Due to the fact that ϕ is strictly convex, g is strictly convex. Therefore, we get that,

$$g(1) > g(0) + \frac{\mathrm{d}}{\mathrm{d}t} g(t)\big|_{t=0}, \tag{3.4}$$

which implies that,

$$g(1) = \phi(\hat{a}) > g(0) + \frac{\mathrm{d}}{\mathrm{d}t} g(t)\big|_{t=0} = \phi(a) + \nabla\phi(a)^T(\hat{a} - a) \tag{3.5}$$

and therefore,

$$\begin{aligned} 0 &< \phi(\hat{a}) - \phi(a) - \nabla\phi(a)^T(\hat{a} - a) && (3.6)\\ &= \ell_\phi(\hat{a}, a). && (3.7) \end{aligned}$$

Property *(ii)* is a direct consequence of the definition of Bregman divergences jointly with the strictly convexity nature of the function $\phi(a)$.

To prove property *(iii)*, let $a, \hat{a}, \tilde{a} \in \Omega$, then

$$\begin{aligned} &\ell_\phi(a, \tilde{a}) + \ell_\phi(\tilde{a}, \hat{a}) - \langle a - \tilde{a}, \nabla\phi(\hat{a}) - \nabla\phi(\tilde{a})\rangle \\ &= \phi(a) - \phi(\tilde{a}) - (a - \tilde{a})\nabla\phi(\tilde{a}) \\ &\quad + \phi(\tilde{a}) - \phi(\hat{a}) - (\tilde{a} - \hat{a})\phi(\hat{a}) - (a - \tilde{a})\nabla\phi(\hat{a}) + \nabla\phi(\tilde{a})(a - \tilde{a}) && (3.8)\\ &= \phi(a) - \phi(\hat{a}) - \nabla\phi(\hat{a})(a - \hat{a}) && (3.9)\\ &= \ell_\phi(a, \hat{a}). && (3.10) \end{aligned}$$

□

We refer to [8] for further properties satisfied by the set of convex functions.

Theorem 6 ([4]). *Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be a strictly convex differentiable function and let ℓ_ϕ be the corresponding Bregman divergence. Let X be an arbitrary random variable taking values in $\mathbb{R}^d$ such that $\mathsf{E}[X]$ and $\mathsf{E}[\phi(X)]$ are finite. Then, among all the functions of Y, the conditional expectation is the unique minimizer of the Bregman divergence, i.e.,*

$$\arg\min_{z \in \sigma(Y)} \mathsf{E}\left[\ell_\phi(X, z)\right] = \mathsf{E}[X|Y]. \tag{3.11}$$

Proof. Using Property *(iii)* of Theorem 5, we have that,

$$\begin{aligned}\mathsf{E}\left[\ell_\phi(X,z)\right] &= \mathsf{E}\left[\ell_\phi(X,\mathsf{E}[X|Y]) + \ell_\phi(\mathsf{E}[X|Y],z)\right] \\ &\quad - \langle X - \mathsf{E}[X|Y], \nabla\phi(z) - \nabla\phi(\mathsf{E}[X|Y])\rangle \qquad (3.12)\\ &= \mathsf{E}\left[\ell_\phi(X,\mathsf{E}[X|Y])\right] + \mathsf{E}\left[\ell_\phi(\mathsf{E}[X|Y],z)\right], \qquad (3.13)\end{aligned}$$

which implies that,

$$\mathsf{E}\left[\ell_\phi(\mathsf{E}[X|Y],z)\right] = \mathsf{E}\left[\ell_\phi(X,z)\right] - \mathsf{E}\left[\ell_\phi(X,\mathsf{E}[X|Y])\right], \qquad (3.14)$$

which due to Property *(i)* of Theorem 5 is minimum when $z = \mathsf{E}[X|Y]$. □

In Table 3.1 we provide several Bregman divergences used throughout this thesis. With a slight abuse of notation, we use ℓ_G to denote the square distance, ℓ_I stands for the Itakura–Saito distance, ℓ_P stands for the Bregman divergence build upon the convex function $\phi(x) = x\log x$, and ℓ_b, ℓ_{nb} stand for the Bregman divergences used to state information estimation relationships in the case of the binomial and negative binomial models, respectively.

Notation	$\ell_\phi(a,\hat{a})$	$\phi(a)$	domain of $\ell_\phi(\cdot,\cdot)$
$\ell_G(a,\hat{a})$	$(a-\hat{a})^2$	a^2	$\mathbb{R}^2$
$\ell_I(a,\hat{a})$	$\frac{a}{\hat{a}} - \log\frac{a}{\hat{a}} - 1$	$-\log a$	$(0,\infty)^2$
$\ell_P(a,\hat{a})$	$a\log\frac{a}{\hat{a}} - (a-\hat{a})$	$a\log a$	$[0,\infty)^2$
$\ell_b(a,\hat{a})$	$a\log\frac{a(1-\hat{a})}{\hat{a}(1-a)} - \frac{a-\hat{a}}{1-\hat{a}}$	$a\log\frac{a}{1-a}$	$(0,1)^2$
$\ell_{nb}(a,\hat{a})$	$a\log\frac{a(1+\hat{a})}{\hat{a}(1+a)} - \frac{a-\hat{a}}{1+\hat{a}}$	$a\log\frac{a}{1+a}$	$[0,\infty)^2$

Table 3.1: Example Bregman divergences.

3.2 Exponential Families and Bregman Divergences

Given a random variable $X \in \mathbb{R}^d$, let ξ be a collection of functions $\xi_i : \mathbb{R}^d \to \mathbb{R}$, known as sufficient statistics. For a given vector of sufficient statistics ξ, let $\theta = \{\theta_1, \cdots, \theta_{|\xi|}\}$ be the associated vector of natural or exponential parameters where $|\xi|$ is the number of functions in the collection ξ. For each $x \in \mathbb{R}^d$ we use $\langle\theta, \xi(x)\rangle$ to denote the Euclidean inner product in $\mathbb{R}^{|\xi|}$ of the two vectors θ and $\xi(x)$.

Definition 3. *The exponential family associated with ξ consists of the following parametrized collection of density functions*

$$P_\theta(x) = \exp\{\langle \theta, \xi(x)\rangle - A(\theta)\}\nu(x) \tag{3.15}$$

In Section 2.2 we restricted ourselves to a scalar version through the function $\rho(y)$(see (2.14)). The log-partition function $A(\theta)$, corresponding to the distributions indicated in (3.15) is given by,

$$A(\theta) = \log \int \exp \langle \theta, \xi(x)\rangle \nu(\mathrm{d}x). \tag{3.16}$$

With the set of sufficient statistics ξ fixed, each parameter vector θ indexes a particular member P_θ of the family. The canonical parameters θ of interest belong to the set

$$\Theta \triangleq \{\theta \in \mathbb{R}^{|\xi|} | A(\theta) < \infty\}. \tag{3.17}$$

A distribution of the exponential family is said to be minimal if there does not exist a non-zero vector $\lambda \in \mathbb{R}^{|\xi|}$ such that the linear combination,

$$\langle \lambda, \xi(x)\rangle = \sum_{i \in \{1,\cdots,|\xi|\}} \lambda_i \xi_i(x) \tag{3.18}$$

is equal to a constant.

Definition 4. *Consider a d-dimensional real-valued random vector X distributed according to an exponential family density P_θ specified by the natural parameter $\theta \in \mathrm{int}(\Theta)$. The expectation of a natural statistics $\xi_i(X)$ with respect to P_θ is called the expectation parameter, given by,*

$$\mu_i = \mu_i(\theta) = \mathsf{E}[\xi_i(X)] = \int \xi_i(x) P_\theta(x) \mathrm{d}x. \tag{3.19}$$

Remark 2. *It can be shown that the Hessian $\nabla^2 A(\theta)$ is positive semidefinite [64]. Therefore, $A(\theta)$ is strictly convex in θ when P_θ is minimal.*

Remark 3. *The derivative of the log-partition function $A(\theta)$ with respect to the i-th component of the natural parameter is given by*

$$\frac{\partial}{\partial \theta_i} A(\theta) = \mu_i(\theta) = \mathsf{E}\left[\xi_i(X)\right] = \int \xi_i P_\theta(x) \mathrm{d}x. \tag{3.20}$$

The proof to (3.20) *can be obtained by differentiating inside the integral in the definition of $A(\theta)$. See [64] for further details.*

Definition 5 ([46]). *Let $A(\theta)$ be a real-valued function on $\mathbb{R}^d$. Then, its conjugate function $\phi(\mu)$ is given by,*

$$\phi(\mu) = \sup_{\theta \in \mathrm{dom}(A)} \{\langle \mu, \theta \rangle - A(\theta)\}. \tag{3.21}$$

The following theorem, due to Banerjee *et al.*, demostrates the existence of a one-to-one relationship between the set of Bregman divergences and the exponential family distributions.

Theorem 7 ([5]). *Let P_θ be a probability density function of a exponential family distribution where $\phi(\mu)$ is the conjugate function of $A(\theta)$. Let $\theta \in \Theta$ be the natural parameter and $\mu \in \mathrm{int}(\mathrm{dom}(\phi))$ be the corresponding expectation parameter* (3.19)*. Let d_ϕ be the Bregman divergence associated with the function ϕ. Then P_θ can be uniquely expressed as,*

$$P_\theta(x) = \exp\{-d_\phi(x, \mu)\} b_\phi(x), \tag{3.22}$$

where $b_\phi(x) \triangleq \exp\{\phi(x)\}\nu(x)$.

Now we proceed to illustrate this theorem with some distributions from the exponential family set. In what follows we give an answer to the question whether the function d_ϕ used in Theorem 7 corresponds to the functions ℓ_G and ℓ_P used to express information measures over the Gaussian and Poisson channels, illustrated previously in (1.8) and (1.13).

Example 3 (Exponential form: Gaussian distribution).
Let $P_\theta(x)$ be a standard Gaussian distribution with mean μ and variance σ^2. Then,

$$P_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \tag{3.23}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}} \tag{3.24}$$

$$= e^{x\cdot\theta - \frac{\sigma^2\theta^2}{2}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \tag{3.25}$$

where, to obtain (3.25) *we replace $\theta = \mu/\sigma^2$. Consequently by* (3.15) *and*

(3.21),

$$\phi(\mu) = \sup_{\theta\in\mathbb{R}}\{\mu\cdot\theta - A(\theta)\} \tag{3.26}$$

$$= \sup_{\theta\in\mathbb{R}}\{\mu\cdot\theta - \frac{\sigma^2\theta^2}{2}\} \tag{3.27}$$

$$= \mu\cdot\frac{\mu}{\sigma^2} - \frac{\sigma^2}{2}\frac{\mu^2}{\sigma^4} \tag{3.28}$$

$$= \frac{1}{2}\frac{\mu^2}{\sigma^2}, \tag{3.29}$$

which means that, by the definition of the Bregman divergence,

$$d_\phi(x,\mu) = \frac{1}{2\sigma^2}\ell_G(x,\mu). \tag{3.30}$$

According to Theorem 7 we have that,

$$P_\theta(x) = e^{-\frac{1}{2\sigma^2}\ell_G(x,\mu)}b_\phi(x) \tag{3.31}$$

$$= e^{-\frac{1}{2\sigma^2}\ell_G(x,\mu)}e^{\phi(x)}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{x^2}{2\sigma^2}} \tag{3.32}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}\ell_G(x,\mu)}, \tag{3.33}$$

where, based on (3.20) *with* $A(\theta) = \sigma^2\theta^2/2$,

$$\mu(\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta}A(\theta) = \mathsf{E}[X] = \sigma^2\theta = \mu. \tag{3.34}$$

Example 4 (Exponential form: Poisson distribution). *Let* $P_\theta(x)$ *be a Poisson distribution with mean* λ*. For* $\theta = \log\lambda$*, we have that*

$$P_\theta(x) = \frac{\lambda^x}{x!}e^{-\lambda} \tag{3.35}$$

$$= \frac{1}{x!}e^{-\lambda + x\log\lambda} \tag{3.36}$$

$$= \frac{1}{x!}e^{x\cdot\theta-\lambda} \tag{3.37}$$

$$= \frac{1}{x!}e^{x\cdot\theta-e^\theta}. \tag{3.38}$$

For the Poisson distribution with mean λ*, using* (3.15) *and* (3.21)*, we obtain*

$$\phi(\mu) = \sup_{\theta\in\mathbb{R}}\{\mu\cdot\theta - A(\theta)\} \tag{3.39}$$

$$= \sup_{\theta\in\mathbb{R}}\{\mu\cdot\theta - e^\theta\} \tag{3.40}$$

$$= \mu\log\mu - \mu, \tag{3.41}$$

which implies that, by the definition of the Bregman divergence,

$$d_\phi(x,\mu) = \phi(x) - \phi(\mu) - \langle x-\mu, \nabla\phi(\mu)\rangle \tag{3.42}$$

$$= x\log\frac{x}{\mu} - (x-\mu) \tag{3.43}$$

$$= \ell_P(x,\mu). \tag{3.44}$$

Hence, we get that

$$P_\theta(x) = \frac{\lambda^x}{x!}e^{-\lambda} \tag{3.45}$$

$$= \exp\{-\ell_P(x,\mu)\}\frac{\exp\{x\log x - x\}}{x!} \tag{3.46}$$

where, with $A(\theta) = e^\theta$,

$$\mu(\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta}A(\theta) = \mathsf{E}[X] = e^\theta = \lambda. \tag{3.47}$$

The remainder of Section 3.2 is devoted to the analysis of the exponential representation form of the binomial and negative binomial distributions. Notice that each exponential representation is given by a unique function d_ϕ (see Theorem 7). For the sake of clarity we postpone to Chapters 4 and 5 respectively, the comparisons between the functions d_ϕ, given by the exponential representation, and the correspondent functions ℓ_ϕ that give rise to several information estimation relationships in the context of the binomial and negative binomial models.

Example 5 (Exponential form: binomial distribution). *Let* $P_\theta(x)$ *be a binomial distribution with parameters* (n,p)*, i.e.,*

$$P_\theta(y) = \binom{n}{y}p^y(1-p)^{n-y}, \quad y\in\{0,1,\cdots,n\}, \tag{3.48}$$

with $p\in(0,1)$ *and* $n\in\mathbb{Z}_0^+$*. Following a similar procedure to that used in Examples 3 and 4, we get that,*

$$\phi(\mu) = \mu\log\frac{\mu}{n-\mu} - n\log\frac{n}{n-\mu}. \tag{3.49}$$

Then, the Bregman divergence associated with (3.48) *is given by,*

$$d_\phi(x,\mu) = x\log\frac{x(n-\mu)}{(n-x)\mu} - n\log\frac{n-\mu}{n-x}. \tag{3.50}$$

which implies that,

$$P_\theta(y) = e^{-d_\phi(y,\mu)} e^{\phi(y)} \binom{n}{y} \tag{3.51}$$

where $\phi(y)$ is given in (3.49) *and*

$$\mu(\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta} A(\theta) = \mathsf{E}[Y] = n \frac{e^\theta}{1+e^\theta} = np. \tag{3.52}$$

with $A(\theta) = n \log(e^\theta + 1)$.

Example 6 (Exponential form: negative binomial distribution). *Let $P_\theta(x)$ be a negative binomial distribution with parameters (r, q), i.e.,*

$$P_\theta(y) = \binom{y+r-1}{y} q^y (1-q)^r, \quad y = 0, 1, \cdots \tag{3.53}$$

Then,

$$P_\theta(y) = e^{-d_\phi(y,\mu)} e^{\phi(y)} \binom{r+y-1}{y} \tag{3.54}$$

where,

$$d_\phi(y,\mu) \triangleq d_{nb} = y \log \frac{y(r+\mu)}{(r+y)\mu} + r \log \frac{r+\mu}{r+y} \tag{3.55}$$

with

$$\phi(y) = y \log \frac{y}{r+y} - r \log \frac{r+y}{r}. \tag{3.56}$$

3.3 Concluding Remarks

In this chapter we have defined and proved several properties satisfied by Bregman divergences. Their positiveness implies that the derivative of the mutual information is positive for all values of θ where the expressions given in (1.8) and (1.13) are valid. This means also that the mutual information is an increasing function of θ, a fact that, for instance, has applications for the class of "More Capable" Broadcast channels, studied previously in [19, 20, 12]. Similar conclusions can be derived for relative entropies. Furthermore, property stated in Theorem 6, regarding the minimization of the expected value of Bregman divergences, give rise to the "I-MMSE" and

"I-MMLE" relationships in the context of Gaussian and Poisson models. This property plays a fundamental role in the characterization of similar relationships over the binomial and negative binomial models, which are studied in Chapters 4 and 5, respectively.

Additionally we present a result obtained previously in [5] that shows a one-to-one relationship between the exponential family distributions and the Bregman divergences functions. Taking into account this relationship we prove that the Bregman divergence that appears in the exponential form of the Gaussian and Poisson distributions is related with the Bregman divergence used in (1.8) and (1.13) to represent the derivative of the mutual information. Later on, in Chapters 4 and 5 we explore in the context of the binomial and negative binomial models whether this connection stills been valid. Furthermore, facing different structures in the constitution of each model, we show that the connection between information and estimation is not always given through a unique Bregman divergence.

Chapter 4

Binomial Model

In this chapter[1] we present new information estimation expressions for a random transformation based on the binomial distribution. The binomial model is useful in the treatment of the deletion channel, which was introduced by Levenshtein [37] and is widely used to model packet switching networks and systems with synchronization errors.

Initially, adopting a linear scaling of the input θX where θ is the input scaling factor, we show that the derivative of the relative entropy between two distributions obtained at the output of the model can be represented through the expectation of a Bregman divergence ℓ_b, similar to (1.10) and (1.28) showed previously for the Gaussian and Poisson models respectively. When we analyze the behavior of the input–output mutual information, we show that its derivative with respect to the input scaling factor can also be represented through the expectation of the function ℓ_b. The context over which this behavior is analyzed, leads to an information–estimation relationship similar to the "I-MMSE" and "I-MMLE" relationships showed previously for the Gaussian and Poisson models.

For an arbitrary θ-dependent preprocessing of the input X_θ, we show different expressions for the mutual information and the relative entropy in terms of conditional mean estimates. Among other applications, these expressions let us state a relationship between the Bregman divergence used in the exponential form of the binomial distribution (3.50) and the derivative of some information measures expressed through the Bregman divergence ℓ_b.

Along the process we highlight several connections between the Poisson and binomial models.

[1] Some results presented through this chapter were published jointly with Professors F. Pérez-Cruz and D. Guo in [54, 57, 56, 21, 55].

4.1 Model definition

The binomial model is based on the binomial distribution that describes the probability of having y successful trials in n independent Bernoulli trials, each with probability p to succeed:

$$P(Y=y)=\binom{n}{y}p^y(1-p)^{n-y},\quad y=0,1,\ldots,n. \tag{4.1}$$

We define the binomial model of order n as a random transformation that maps an input random variable X to an integer output random variable Y, where, conditioned on $X=x$, Y has distribution Binomial $(n, f(\theta,x))$. In this model $f(\theta,x)$ represents a deterministic θ-dependent preprocessing of the input. The conditional probability mass function (pmf) of the model is given by,

$$P^n_{Y|X}(y|x)=\binom{n}{y}(f(\theta,x))^y(1-f(\theta,x))^{n-y},\quad y=0,1,\ldots,n. \tag{4.2}$$

Dependency of Y on θ is implicit for notational convenience. Variables X and Y are viewed as the input and output of the binomial model, respectively, where $f(\theta,x)$ controls the probability of success of each of the Bernoulli trials that contributes to Y. Also for notational convenience, we use the shorthand

$$X_\theta = f(\theta, X). \tag{4.3}$$

Furthermore, it is assumed that $f(\theta,X)$ is differentiable with respect to θ, and we denote the derivative of the preprocessed input by,

$$X'_\theta = \frac{\partial f(\theta,X)}{\partial\theta}. \tag{4.4}$$

To present different relations between estimation and information, let P_X and Q_X be two input distributions. The two input distributions induce two output distributions through the same n-th order binomial model (4.2), denoted by P^n_Y and Q^n_Y, respectively. We shall use $\mathsf{E}^n[\cdot]$ to represent the expectation with respect to the probability measure $P^n_{XY}=P_XP^n_{Y|X}$. We use the subscript Q, *i.e.*, $\mathsf{E}^n_Q[\cdot]$, when the expectation is with respect to the probability measure $Q^n_{XY}=Q_XP^n_{Y|X}$. Similarly, we use $\mathsf{E}^n[X_\theta|Y]$ (resp. $\mathsf{E}^n_Q[X_\theta|Y]$) to denote the conditional estimate of X_θ given the corresponding output Y of the n-th order binomial model, when $X\sim P_X$ (resp. $X\sim Q_X$). In particular, in view of (4.2),

$$\mathsf{E}^n[Y|X_\theta=z]=nz, \tag{4.5}$$

where we use the superscript n to specify that the expectation pertains to the n-th order binomial model.

4.2 Information-Estimation Relationships

Connections between information and estimation have proved their own benefits in the past, predominantly in the Gaussian and Poisson models. Starting from their simplicity and ending up in proofs for the entropy power inequality [63] and applications in the field of secrecy capacity [34], these kind of relationships have important implications in those problems tackled by the information theory community. Along this section we present several information-estimation expressions assuming different scenarios in the constitution of the binomial model.

4.2.1 Linear Scaling ($X_\theta = \theta X$)

We begin with a set of results concerning linear scaling of the input where $X_\theta = \theta X$ such that $\theta X \in (0, 1)$.

Theorem 8. *Let X be a random variable taking its values in $(0, x_{\max})$, following distribution P_X or Q_X. Let Y be the output of the n-th order binomial model described by (4.2) with $X_\theta = \theta X$. Then,*

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\theta} D(P_Y^n \| Q_Y^n) \\
&\quad = \frac{n}{\theta} \mathsf{E}^{n-1}\left[\ell_b(\theta X, \mathsf{E}_Q^{n-1}[\theta X|Y])\right] - \frac{n}{\theta}\mathsf{E}^{n-1}\left[\ell_b(\theta X, \mathsf{E}^{n-1}[\theta X|Y])\right] \quad (4.6)\\
&\quad = \frac{n}{\theta} \mathsf{E}^{n-1}\left[\ell_b(\mathsf{E}^{n-1}\left[\theta X|\, Y\right], \mathsf{E}_Q^{n-1}\left[\theta X|\, Y\right])\right] \quad (4.7)
\end{aligned}$$

and

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\theta} D(P_Y^n \| Q_Y^n) &= \mathsf{E}^n\left[\frac{Y}{\theta}\ell_I\left(X^{-1} - \theta, \mathsf{E}_Q^n[X^{-1}|Y] - \theta\right)\right] \\
&\quad - \mathsf{E}^n\left[\frac{Y}{\theta}\ell_I\left(X^{-1} - \theta, \mathsf{E}^n[X^{-1}|Y] - \theta\right)\right] \quad (4.8)\\
&= \mathsf{E}^n\left[\frac{Y}{\theta}\ell_I\left(\mathsf{E}^n[X^{-1}|Y] - \theta, \mathsf{E}_Q^n[X^{-1}|Y] - \theta\right)\right] \quad (4.9)
\end{aligned}$$

hold for all $\theta \in (0, x_{\max}^{-1})$.

Proof. See Section 4.4.5 for the proof of (4.7) and Section 4.4.1 for the proof of (4.9). □

Theorem 8 provides two sets of expressions, both with their merits, for the rate of change of the relative entropy as a function of the scaling α. The RHS of both (4.7) and (4.9) involves some loss (Bregman divergence) due to the mismatched estimation by assuming a distribution Q_X while the true distribution is P_X. In (4.7) the loss function considers the conditional mean of X and in (4.9) the loss function considers the conditional mean of X^{-1}. In (4.7) the RHS is proportional to the expected distance between two posterior mean estimates with respect to the binomial model of order $n-1$, measured through the Bregman divergence ℓ_b, while the expectation in the LHS is with respect to the binomial model of order n. Moreover, in (4.9), all the expectations are with respect to the underlying n-th order binomial model. However, the RHS becomes the inner product of Y/θ and the loss function does not rely on the posterior mean estimate.

The following theorem shows that the derivative of the input–output mutual information across the binomial model can also be related to a Bregman divergence.

Theorem 9. *Let $X \sim P_X$ be a random variable taking its value in $(0, x_{\max})$. Let Y be the output of the n-th order binomial model described by (4.2) with $X_\theta = \theta X$. Then,*

$$\frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) = \frac{n}{\theta} \mathsf{E}^{n-1}\left[\ell_b(\theta X, \mathsf{E}^{n-1}[\theta X|Y])\right] \tag{4.10}$$

$$= \mathsf{E}^n\left[\frac{Y}{\theta}\ell_I\left(X^{-1} - \theta, \mathsf{E}^n[X^{-1}|Y] - \theta\right)\right] \tag{4.11}$$

hold for all $\theta \in (0, x_{\max}^{-1})$.

Proof. See Section 4.4.6 for the proof of (4.10) and Section 4.4.2 for the proof of (4.11). □

Based on the definition of the function ℓ_b and subsequently rearranging algebraically the terms of the expression given in (4.10), we obtain an alternative representation for the mutual information expression written in terms of the function ℓ_P.

Corollary 1. *Assume the same set of conditions used in Theorem 9. Then,*

$$\frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) = n\mathsf{E}^{n-1}\left[\ell_P\left(X, \frac{\mathsf{E}^{n-1}[X|Y](1-\theta X)}{1-\theta\mathsf{E}^{n-1}[X|Y]}\right)\right]. \tag{4.12}$$

Mathematically speaking, (4.10) and (4.12) are equal, but based on the arguments used in (4.12), the function ℓ_P does not achieve its minimum

value, given that the second argument is not the conditional mean of the input X. In the case of (4.10) notice that the function ℓ_b do achieve its minimum value through the use of the arguments X and its conditional mean estimate.

Adding (4.7) and (4.10) and using property (iii) of the Bregman divergence functions yield the following result.

Theorem 10. *Let X and Y be defined as in Theorem 8. Then,*

$$\mathsf{E}^{n-1}\left[\ell_b(\theta X, \mathsf{E}_Q^{n-1}\left[\theta X|\, Y\right])\right] = \frac{\theta}{n}\frac{\mathrm{d}}{\mathrm{d}\theta}\left[I(X;Y) + D(P_Y^n \| Q_Y^n)\right] \tag{4.13}$$

holds for all $\theta \in (0, x_{\max}^{-1})$.

Theorem 10 shows that the mismatched estimation penalty incurred when we estimate the random variable X through the mismatched prior distribution Q_X is proportional to the derivative of sum of the relative entropy between P_X and Q_X and the input-output mutual information. In this case, the mismatched penalty measured through the loss function ℓ_b is minimum when $Q_X = P_X$ for all $x \in \mathcal{X}$. For $Q_X = P_X$, the relative entropy between P_Y^n and Q_Y^n is zero and the minimum of the loss function is proportional to the derivative of the mutual information:

$$\min_{Q_X} \mathsf{E}^{n-1}\left[\ell_b(\theta X, \mathsf{E}_Q^{n-1}[\theta X|Y])\right] = \frac{\theta}{n}\frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y). \tag{4.14}$$

The counterpart results to (4.13) regarding the Gaussian and Poisson channels are, respectively, shown in [66] and [3].

4.2.2 Arbitrary Scaling

We now study the more general case in which $f(\theta, X) = X_\theta \in (0,1)$ depends on the parameter θ in an arbitrary manner, which is not necessarily linear. In particular, the first part of Theorems 8 and 9 can be regarded as corollary of these general results. We assume that the set of feasible values for the parameter θ, denoted by Θ is an open real number set.

Theorem 11. *Under both distributions P_X and Q_X, let $X_\theta \in (0,1)$ and X'_θ be integrable and bounded. Let Y be the output of the n-th order binomial defined in* (4.2) *with X_θ as the input. Then,*

$$\frac{\mathrm{d}}{\mathrm{d}\theta} D(P_Y^n \| Q_Y^n) = \mathsf{F}_Q^{n-1}(X_\theta) - \mathsf{F}_P^{n-1}(X_\theta) \tag{4.15}$$

holds for all $\theta \in \Theta$*, where*

$$\begin{aligned}&\mathsf{F}_Q^{n-1}(X_\theta)\\&= n\mathsf{E}^{n-1}\left[X'_\theta \log \frac{(1-\mathsf{E}_Q^{n-1}[X_\theta|Y])X_\theta}{(1-X_\theta)\mathsf{E}_Q^{n-1}[X_\theta|Y]} - \frac{\mathsf{E}_Q^{n-1}[X'_\theta|Y](X_\theta - \mathsf{E}_Q^{n-1}[X_\theta|Y])}{(1-\mathsf{E}_Q^{n-1}[X_\theta|Y])\mathsf{E}_Q^{n-1}[X_\theta|Y]}\right].\end{aligned} \tag{4.16}$$

Proof. See Section 4.4.3. □

For $X_\theta = \theta X$, (4.16) simplifies to

$$\begin{aligned}&\mathsf{F}_Q^{n-1}(\theta X)\\&= n\mathsf{E}^{n-1}\left[X \log \frac{(1-\mathsf{E}_Q^{n-1}[\theta X|Y])\theta X}{(1-\theta X)\mathsf{E}_Q^{n-1}[\theta X|Y]} - \frac{\mathsf{E}_Q^{n-1}[X|Y](\theta X - \mathsf{E}_Q^{n-1}[\theta X|Y])}{(1-\mathsf{E}_Q^{n-1}[\theta X|Y])\mathsf{E}_Q^{n-1}[\theta X|Y]}\right]\end{aligned} \tag{4.17}$$

$$= \frac{n}{\theta}\mathsf{E}^{n-1}\left[\ell_b(\theta X, \mathsf{E}_Q^{n-1}[\theta X|Y])\right]. \tag{4.18}$$

in which we can see that, in the general case, X'_θ might preclude us from using the Bregman divergence for the binomial model.

Theorem 12. *Let* X_θ*,* X'_θ *and* Y *be defined as in Theorem 11. Then,*

$$\frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) = \mathsf{F}_P^{n-1}(X_\theta) \tag{4.19}$$

$$= n\ \mathsf{E}^{n-1}\left[X'_\theta \log \frac{(1-\mathsf{E}^{n-1}[X_\theta|Y])X_\theta}{(1-X_\theta)\mathsf{E}^{n-1}[X_\theta|Y]}\right], \tag{4.20}$$

holds for all $\theta \in \Theta$*.*

Proof. See Section 4.4.4. □

We illustrate with a simple example how Theorem 12 can be applied and show that the achieved result cannot be represented by a Bregman divergence. We set $X_\theta = X + \theta$ such that $0 < x + \theta < 1$ for all $x \in \mathcal{X}$. A similar scenario was studied in [25] for the Poisson model. Applying (4.20), we get that the derivative of the mutual information can be expressed as:

$$\frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) = n\mathsf{E}^{n-1}\left[\log \frac{(X+\theta)(1-\mathsf{E}^{n-1}[X+\theta|Y])}{(1-(X+\theta))\mathsf{E}^{n-1}[X+\theta|Y]}\right] \tag{4.21}$$

$$= n\mathsf{E}^{n-1}\left[g\left((X+\theta), \mathsf{E}^{n-1}[X+\theta|Y]\right)\right], \tag{4.22}$$

where,

$$g(a,\hat{a}) = \log \frac{a(1-\hat{a})}{(1-a)\hat{a}} - \frac{a-\hat{a}}{(1-\hat{a})\hat{a}}. \tag{4.23}$$

Although $g(a,\hat{a})$ can be obtained from (3.1) with $\phi(a) = \log(a) - \log(1-a)$, it is not a Bregman divergence because $\log(a) - \log(1-a)$ is non-convex on $(0,1)$.

Recall from the analysis given in (3.30) and (3.44) that results obtained for the Gaussian and Poisson models suggest a straight relationship between the Bregman divergence used in the exponential form of the conditional distribution of the model and the derivatives of the information measures treated. Hence a natural extension to these results consists on finding out the relationship between the Bregman divergence used to obtain the exponential form of the binomial distribution (3.50), and the Bregman divegence ℓ_b used in Theorems 8 and 9 to express the derivative of the relative entropy and mutual information.

Theorem 13. *Let X be a positive bounded random variable that can be distributed as either P_X or Q_X. Let Y be the output of a n-th binomial model with parameters $(n, \theta X/n)$. Then,*

$$\frac{1}{\theta}\mathsf{E}^{n-1}\left[d_b\left(\theta X, \mathsf{E}^{n-1}[\theta X|Y]\right)\right] < \frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y), \tag{4.24}$$

and

$$\frac{1}{\theta}\mathsf{E}^{n-1}\left[d_b\left(\mathsf{E}^{n-1}[\theta X|Y], \mathsf{E}_Q^{n-1}[\theta X|Y]\right)\right] < \frac{\mathrm{d}}{\mathrm{d}\theta} D(P_Y^n||Q_Y^n), \tag{4.25}$$

where we use d_b, to denote the Bregman divergence associated with the exponential representation of the binomial distribution, given in (3.50).

Proof. See Section 4.4.7. □

Expression (4.24) shows that in the case of the binomial model the derivative of the mutual information, up to a scaling factor, constitutes an upper bound for the expectation of the Bregman divergence d_b associated with the binomial distribution on its exponential form (see (3.50) and (3.51)). The generality of the bound relies on the fact that it holds regardless of the input distribution. A similar analysis applies in the case of the derivative of the relative entropy.

4.2.3 Low Input Scaling

Previously, in works due to Guo *et al.* [69] and to Lapidoth*et al.* [35] was analyzed the low input scaling behavior for the Gaussian and Poisson channels. In the former, the expression found has a fundamental meaning in the wideband communications regime; it determines the minimum energy required per bit to achieve reliable communication [61]; in the latter, it was established that, at low input scaling regimes, the capacity of the Poisson channel scales like $\mathcal{E} \log 1/\mathcal{E}$ where $\mathcal{E}$ represents the average input power.

The expressions given for the low input scaling regime play a fundamental roll in terms of efficiency. This claim is based on the fact that, an ideal scenario to work over is that that achieves the maximum rate of change of the mutual information when small increases in the amplifying factor are allowed[2]. Additionally, based on the concavity of the mutual information over the input scaling space, the derivative of the mutual information is maximum when the input scaling tends to zero.

Based on the expressions found for the derivative of the mutual information and relative entropy, in this section, we study the behavior of such derivatives when the input scaling θ goes to zero. These expressions show that, over certain scenarios, the low input scaling regime for the binomial models has the same behavior that the one found for the Poisson channel and is independent of the number trials n made to constitute the binomial model.

Theorem 14. *Let $X \sim P_X$ be a positive bounded random variable taking its values in $(0, \theta_{\max}^{-1})$. Let Y be the output of the n-order binomial model described by (4.2) with $X_\theta = \theta X$. Then,*

$$\lim_{\theta \to 0} \frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) = n\mathsf{E}\left[\ell_P(X, \mathsf{E}[X])\right] \tag{4.26}$$

$$= n\mathsf{E}\left[X \log \frac{X}{\mathsf{E}[X]}\right]. \tag{4.27}$$

Proof. See Section 4.4.8. □

Even though the most prominent information-estimation relationship is given in terms of the function ℓ_b, Theorem 14 shows the close relationship between the Bregman divergence ℓ_P and the binomial model. In the context

[2]This claim works over such channels where the mutual information is increasing in θ. An analogous statement can be done when the mutual information has a decreasing nature.

of the Poisson model, Atar *et al.* [3] describe the function shown in (4.27) as the input-dependent expression that is the analogous to the variance in the Gaussian case. This claim appears as consequence of the behavior of the function ℓ_P.

Corollary 2. *Let $X \sim P_X$ be a positive bounded random variable taking its values in $(0, n/\theta_{\max})$. Let Y be the output of the n-th order binomial model described (4.2) with $X_\theta = \theta X/n$. Then,*

$$\lim_{\theta \to 0} \frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) = \mathsf{E}\left[\ell_P(X, \mathsf{E}[X])\right] \tag{4.28}$$

$$= \mathsf{E}\left[X \log \frac{X}{\mathsf{E}[X]}\right]. \tag{4.29}$$

Proof. See Section 4.4.9. □

The analogous expression for the relative entropy to that given in Theorem 14 at low input scaling factor is given as follows.

Theorem 15. *Assume the same set of conditions used in Theorem 14. Then,*

$$\lim_{\theta \to 0} \frac{\mathrm{d}}{\mathrm{d}\theta} D(P_Y^n || Q_Y^n) = n\ell_P(\mathsf{E}[X], \mathsf{E}_Q[X]). \tag{4.30}$$

Proof. Starting from the expression given for the derivative of the relative entropy in Theorem 8, the proof to Theorem 15 is similar to the proof of Theorem 14. □

Based on expression obtained for Theorem 15 notice that the low input scaling behavior for the relative entropy only depends on the mean of each input distribution, feature shared with those results given previously for the Gaussian model.[3]

Corollary 3. *Assume the same set of conditions used in Corollary 2. Then,*

$$\lim_{\theta \to 0} \frac{\mathrm{d}}{\mathrm{d}\theta} D(P_Y^n || Q_Y^n) = \ell_P(\mathsf{E}[X], \mathsf{E}_Q[X]). \tag{4.32}$$

Proof. See Section 4.4.10. □

[3]Let Y be the output of the Gaussian model shown in (1.7). Then, it can be shown that,

$$\lim_{\mathsf{snr} \to 0} \mathsf{E}[X|Y] = \mathsf{E}[X] \tag{4.31}$$

Therefore, $\lim_{\mathsf{snr} \to 0} D(P_Y||Q_Y)$ only depends on $\mathsf{E}[X]$ and $\mathsf{E}_Q[X]$.

4.3 Concluding Remarks

Throughout this chapter we study several information-estimation relationships that arise in the context of the binomial models. In this case we have that, for a linear input scaling $X_\theta = \theta X$, the derivative of the input-output mutual information is related to the conditional mean estimate through a Bregman divergence, as was shown previously for the Gaussian and Poisson models. Based on this fact, we can state for the binomial model a relationship that plays a similar role to the "I-MMSE" relationship in the case of the Gaussian channel and to the "I-MMLE" relationship in the case of the Poisson channel.

Over a second scenario, we characterize the derivative of the relative entropy between two distributions P_Y^n and Q_Y^n obtained at the output of a binomial model. In the linear case, the information-estimation relationship is given in terms of the same Bregman divergence used to describe the mutual information.

A striking property that appears in the process is the fact that those expressions given in terms of the Bregman divergence ℓ_b can also be given in terms of the Bregman divergence ℓ_P used in the context of the Poisson models. The difference in this case arises in the arguments that are used by each function. When dealing with the mutual information expression, the arguments of the function ℓ_b are θX and its conditional mean estimate, which guarantees that the expected loss achieves its minimum value. In the case of the function ℓ_P, the arguments used are X and $\frac{(1-\theta X)\hat{X}}{1-\theta\hat{X}}$ where $\hat{X} = \mathsf{E}^{n-1}[X|Y]$ at which the function ℓ_P does not achieves its minimum value, given that the second argument is not exactly the conditional mean estimate of the first.

When dealing with an arbitrary function X_θ we show that there is a connection between information and estimation through the function $\mathsf{F}(X_\theta)$, which let us express the derivative of the relative entropy and mutual information in terms of conditional estimates. This relationship has several advantages:

- It is useful to prove that not in all the cases the information–estimation relationship is given through a Bregman divergence.

- It is demonstrated that, up to a scaling factor, the expectation of the Bregman divergence d_b between the mean of the model θX and its conditional estimate $\mathsf{E}^{n-1}[\theta X|Y]$ is an upper bound for the derivative of the input–output mutual information which is expressed through

the function ℓ_b. A similar upper bound can be found for the derivative of the relative entropy. The generality of these expressions lie on the fact that they hold regardless of the input distribution.

- Broadcast channels context. In several circumstances it is useful to establish scenarios over which the mutual information is monotone over the parameter θ. This property leads to the notion of "more capable Broadcast channels" for which the capacity region is completely characterized [19, 36, 20].

In the low input scaling regime, we show that, as long as the parameter θ goes to zero, the limit of the derivative of the mutual information and its similar, in the context of the relative entropy, is governed by the Bregman divergence ℓ_P. Specifically, in the case of the mutual information, the low input scaling behavior only depends on the function $\mathsf{E}[\ell_P(X, \mathsf{E}[X]) = \mathsf{E}\left[X \log \frac{X}{\mathsf{E}[X]}\right]$ which in [3] is cited as the input-dependent function over the Poisson model, that plays a roll similar to the variance in the Gaussian model. In the case of the derivative of the relative entropy, we prove that its low input scaling regime only depends on the mean of each input distribution.

4.4 Proofs

In this appendix, we collect the proofs for the binomial model. Organization of this section is given mainly because of the dependency between each result. Initially we prove (4.9) in Theorem 8 and (4.11) in Theorem 9. Later we prove Theorems 11 and 12 which let us conclude, as particular cases, expressions given in (4.7) in Theorem 8 and (4.10) in Theorem 9. We conclude this section with the proofs to results pertaining the low input scaling behavior of the binomial model.

4.4.1 Proof of (4.9) in Theorem 8.

Proof of (4.9) hinges on the following lemma, which translates the derivative of the output pmf to a certain difference function.

Lemma 1. *Let P_Y^n be the pmf of the output of the binomial model described by (4.2) with θX as the input. For every $y = 0, \ldots, n$,*

$$\frac{\mathrm{d}}{\mathrm{d}\theta} P_Y^n(y) = \frac{1}{\theta} \left(y P_Y^n(y) - (y+1) P_Y^n(y+1) \right), \tag{4.33}$$

where we use the convention that $P_Y^n(n+1) = 0$.

Lemma 1 resembles a result for Gaussian models in [23], where the derivative with respect to the scaling parameter translates to the derivative with respect to the output variable. For the binomial model, the output is discrete and the result consists of the difference of the output distribution (modulated by the variable y) in lieu of derivative.

Proof. We start with

$$P_Y^n(y) = \mathsf{E}\left[\binom{n}{y}(\theta X)^y (1-\theta X)^{n-y}\right]. \tag{4.34}$$

Evidently,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\theta}P_Y^n(y) &= \mathsf{E}\left[\binom{n}{y}\frac{\mathrm{d}}{\mathrm{d}\theta}\left((\theta X)^y (1-\theta X)^{n-y}\right)\right] && (4.35)\\
&= \frac{y}{\theta}\mathsf{E}\left[\binom{n}{y}(\theta X)^y (1-\theta X)^{n-y}\right] \\
&\quad - \frac{(n-y)}{\theta}\mathsf{E}\left[\binom{n}{y}(\theta X)^{y+1}(1-\theta X)^{n-y-1}\right] && (4.36)\\
&= \frac{y}{\theta}P_Y^n(y) - \frac{y+1}{\theta}\mathsf{E}\left[\binom{n}{y+1}(\theta X)^{y+1}(1-\theta X)^{n-y-1}\right] && (4.37)
\end{aligned}$$

where in (4.35) we use the interchangeability property (see Lemma 2 in Appendix 4.4.3). We note that (4.35)–(4.36) hold for $y = 0, \ldots, n$. In arriving at (4.37), we use (4.34) and the convention that $\binom{n}{n+1} = 0$. In fact, the second term in (4.36) and the second term in (4.37) are both equal to 0 for $y = n$. Using (4.34) again, we arrive at (4.33) from (4.37). □

Proof of (4.9) *in Theorem 8.* From the definition of relative entropy,

$$D(P_Y^n \| Q_Y^n) = \sum_{y=0}^{n} P_Y^n(y) \log \frac{P_Y^n(y)}{Q_Y^n(y)}, \tag{4.38}$$

it is not difficult to show that

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\theta}D(P_Y^n \| Q_Y^n) \\
&= \sum_{y=0}^{n}\left(\log\frac{P_Y^n(y)}{Q_Y^n(y)}\right)\frac{\mathrm{d}P_Y^n(y)}{\mathrm{d}\theta} - \frac{P_Y^n(y)}{Q_Y^n(y)}\frac{\mathrm{d}Q_Y^n(y)}{\mathrm{d}\theta} && (4.39)\\
&= \theta^{-1}(A - B), && (4.40)
\end{aligned}$$

where

$$A = \theta \sum_{y=0}^{n} \left(\log \frac{P_Y^n(y)}{Q_Y^n(y)} \right) \frac{\mathrm{d}P_Y^n(y)}{\mathrm{d}\theta} \tag{4.41}$$

$$= \sum_{y=0}^{n} \left(\log \frac{P_Y^n(y)}{Q_Y^n(y)} \right) (yP_Y^n(y) - (y+1)P_Y^n(y+1)) \tag{4.42}$$

$$= \sum_{y=1}^{n} \left(\log \frac{P_Y^n(y)}{Q_Y^n(y)} \right) yP_Y^n(y) - \sum_{y=0}^{n-1} \left(\log \frac{P_Y^n(y)}{Q_Y^n(y)} \right) (y+1)P_Y^n(y+1) \tag{4.43}$$

$$= \sum_{y=1}^{n} \left(\log \frac{P_Y^n(y)}{Q_Y^n(y)} \right) yP_Y^n(y) - yP_Y^n(y) \log \frac{P_Y^n(y-1)}{Q_Y^n(y-1)} \tag{4.44}$$

$$= \sum_{y=1}^{n} yP_Y^n(y) \log \frac{P_Y^n(y)Q_Y^n(y-1)}{P_Y^n(y-1)Q_Y^n(y)} \tag{4.45}$$

and

$$B = \theta \sum_{y=0}^{n} \frac{P_Y^n(y)}{Q_Y^n(y)} \frac{\mathrm{d}Q_Y^n(y)}{\mathrm{d}\theta} \tag{4.46}$$

$$= \sum_{y=0}^{n} \frac{P_Y^n(y)}{Q_Y^n(y)} (yQ_Y^n(y) - (y+1)Q_Y^n(y+1)) \tag{4.47}$$

$$= \sum_{y=1}^{n} yP_Y^n(y) - \sum_{y=0}^{n-1} \frac{P_Y^n(y)}{Q_Y^n(y)} (y+1)Q_Y^n(y+1)) \tag{4.48}$$

$$= \sum_{y=1}^{n} yP_Y^n(y) - \sum_{y=1}^{n} yQ_Y^n(y) \frac{P_Y^n(y-1)}{Q_Y^n(y-1)} \tag{4.49}$$

$$= \sum_{y=1}^{n} yP_Y^n(y) \left(1 - \frac{P_Y^n(y-1)Q_Y^n(y)}{P_Y^n(y)Q_Y^n(y-1)} \right), \tag{4.50}$$

where (4.39) relies on $\frac{\mathrm{d}}{\mathrm{d}\theta} \sum_{y=0}^{n} P_Y^n(y) = 0$ and, in (4.42) and (4.47) we apply the result obtained in Lemma 1. Since (4.33) holds for any input distribution P_X, it remains true if P_X is replaced by another distribution Q_X, as long as the input θX belongs to the interval $(0, 1)$. Moreover,

$$P_Y^n(y-1) = \mathsf{E} \left[\binom{n}{y-1} (\theta X)^{y-1} (1 - \theta X)^{n-y+1} \right] \tag{4.51}$$

$$= \frac{y}{n-y+1} \mathsf{E}^n \left[\frac{1-\theta X}{\theta X} \middle| Y = y \right] P_Y^n(y), \tag{4.52}$$

where (4.52) is because

$$\mathsf{E}\left[h(X)P_{Y|X}^n(y|X)\right] = \mathsf{E}^n\left[h(X)|Y=y\right]P_Y^n(y) \tag{4.53}$$

holds for every measurable function $h(x)$.

Similarly,

$$Q_Y^n(y-1) = \frac{y}{n-y+1}\mathsf{E}_Q^n\left[\frac{1-\theta X}{\theta X}\middle| Y=y\right]Q_Y^n(y) \tag{4.54}$$

so that

$$\frac{P_Y^n(y-1)Q_Y^n(y)}{P_Y^n(y)Q_Y^n(y-1)} = \frac{\mathsf{E}^n[X^{-1}-\theta|Y=y]}{\mathsf{E}_Q^n[X^{-1}-\theta|Y=y]}. \tag{4.55}$$

Therefore,

$$\frac{\mathrm{d}}{\mathrm{d}\theta}D(P_Y^n\|Q_Y^n) = \frac{1}{\theta}\sum_{y=1}^n yP_Y^n(y)\ell_I\left(\mathsf{E}^n[X^{-1}-\theta|Y=y], \mathsf{E}_Q^n[X^{-1}-\theta|Y=y]\right) \tag{4.56}$$

$$= \mathsf{E}^n\left[\frac{Y}{\theta}\ell_I\left(\mathsf{E}^n[X^{-1}-\theta|Y], \mathsf{E}_Q^n[X^{-1}|Y]-\theta\right)\right] \tag{4.57}$$

by plugging (4.55) into (4.45) and (4.50) and subsequently (4.40). □

4.4.2 Proof of (4.11) in Theorem 9.

Using Theorem 8,

$$\frac{\mathrm{d}}{\mathrm{d}\theta}I(X;Y) = \int \frac{\mathrm{d}}{\mathrm{d}\theta}D(P_{Y|X=x}^n\|P_Y^n)\mathrm{d}P_X(x) \tag{4.58}$$

$$= \int \mathsf{E}_{P_{Y|X=x}^n}\left[\frac{Y}{\theta}\ell_I\left(x^{-1}-\theta, \mathsf{E}^n[X^{-1}|Y]-\theta\right)\right]\mathrm{d}P_X(x) \tag{4.59}$$

$$= \mathsf{E}^n\left[\frac{Y}{\theta}\ell_I\left(X^{-1}-\theta, \mathsf{E}^n[X^{-1}|Y]-\theta\right)\right], \tag{4.60}$$

where we have fix $x \in \mathcal{X}$ and let $P_Y^n = P_{Y|X=x}^n$, which can be regarded as the output distribution of a binomial model with deterministic input θx. We have also relabeled Q_Y^n by P_Y^n in the second argument of the relative entropy.

4.4.3 Proof of Theorem 11.

Before proceeding with a formal proof for Theorem 11, we present several results regarding the binomial model with parameters (n, X_θ). Recall that $X_\theta = f(\theta, X)$ and $X'_\theta = \partial f(\theta, X)/\partial\theta$.

Lemma 2. *[Exchangeability Property]. Let the function $X_\theta = f(\theta, X) : \Theta \times \mathcal{X} \to (0,1)$ be such that $|\frac{\partial f(\theta,x)}{\partial\theta}| < M$ for all $(\theta, x) \in \Theta \times \mathcal{X}$ where $M \in \mathbb{R}^+$. Then,*

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathsf{E}\left[P^n_{Y|X}(y|X)\right] = \mathsf{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}P^n_{Y|X}(y|X)\right] \tag{4.61}$$

$$= yP^n_Y(y)\mathsf{E}^n\left[\frac{X'_\theta}{X_\theta}\middle| Y = y\right] - (n-y)P^n_Y(y)\mathsf{E}^n\left[\frac{X'_\theta}{1-X_\theta}\middle| Y = y\right], \tag{4.62}$$

where $P^n_{Y|X}$ is given by (4.2).

Proof. By [6, Theorem 12.13] the derivative and the integral operators can be exchanged in order if the following conditions hold:

(i) The derivative $\frac{\mathrm{d}}{\mathrm{d}\theta}P^n_{Y|X}(y|x)$ exists for all values of $(\theta, x) \in \Theta \times \mathcal{X}$.

(ii) There exists a function $\omega(x)$ such that for all $(\theta, x) \in \Theta \times \mathcal{X}$,

$$\left|\frac{\mathrm{d}}{\mathrm{d}\theta}P^n_{Y|X}(y|x)\right| \leq \omega(x) \tag{4.63}$$

and $\mathsf{E}[\omega(X)] < \infty$.

The first condition is verified as follows,

$$\frac{\mathrm{d}}{\mathrm{d}\theta}P^n_{Y|X}(y|x) = \underbrace{P^n_{Y|X}(y|x)\frac{\partial f(\theta,x)}{\partial\theta}\frac{y}{f(\theta,x)}}_{(\mathbb{I})} - \underbrace{P^n_{Y|X}(y|x)\frac{\partial f(\theta,x)}{\partial\theta}\frac{n-y}{1-f(\theta,x)}}_{(\mathbb{III})}, \tag{4.64}$$

in which we have used the definition of $P^n_{Y|X}(y|x)$ in (4.2). Due to the fact that $0 < f(\theta, x) < 1$ and $|\frac{\partial f(\theta,x)}{\partial\theta}| < M$ for all $(\theta, x) \in \Theta \times \mathcal{X}$ we can conclude that (4.64) is well defined over the domain considered.

For *(ii)*, we proceed to prove the existence of integrable upper and lower bounds for (𝕀) and (𝕀𝕀𝕀) in (4.64) independently, and we then combine them. The term (𝕀) is zero for $y = 0$. For $y \neq 0$ we have that,

$$\begin{aligned}
&\left| P^n_{Y|X}(y|x) \frac{\partial f(\theta,x)}{\partial \theta} \frac{y}{f(\theta,x)} \right| \\
&\quad = n \binom{n-1}{y-1} (f(\theta,x))^{y-1} (1 - f(\theta,x))^{n-y} \left| \frac{\partial f(\theta,x)}{\partial \theta} \right| && (4.65) \\
&\quad = n\, P^{n-1}_{Y|X}(y-1|x) \left| \frac{\partial f(\theta,x)}{\partial \theta} \right| < Mn, && (4.66)
\end{aligned}$$

because $|\frac{\partial f(\theta,x)}{\partial \theta}| < M$, $0 < P^{n-1}_{Y|X}(y|x) < 1$ for all $(\theta, x) \in \Theta \times \mathcal{X}$ and $y \in \{1, 2, \ldots, n\}$.

For the second term, (𝕀𝕀𝕀), observe that it is zero for $y = n$. For $y \in \{0, 1, \ldots, n-1\}$ we have

$$\begin{aligned}
&\left| P^n_{Y|X}(y|x) \frac{n-y}{1 - f(\theta,x)} \frac{\partial f(\theta,x)}{\partial \theta} \right| \\
&\quad = n \binom{n-1}{y} (f(\theta,x))^{y} (1 - f(\theta,x))^{n-y-1} \left| \frac{\partial f(\theta,x)}{\partial \theta} \right| && (4.67) \\
&\quad = n\, P^{n-1}_{Y|X}(y|x) \left| \frac{\partial f(\theta,x)}{\partial \theta} \right| < Mn. && (4.68)
\end{aligned}$$

Combining (4.66) and (4.68), we get Condition *(ii)*. □

Next, we provide a set of lemmas that let us write different expressions regarding the binomial channel of order n in terms of conditional estimates carried over a binomial model of order $n - 1$.

Lemma 3.

$$\begin{aligned}
&\mathsf{E}^n \left[Y \mathsf{E}^n \left[\frac{X'_\theta}{X_\theta} \middle| Y \right] \log \frac{P^n_Y(Y)}{Q^n_Y(Y)} \right] \\
&\qquad = n \mathsf{E}^{n-1} \left[\mathsf{E}^{n-1}[X'_\theta|Y] \log \frac{\mathsf{E}^{n-1}[X_\theta|Y] P^{n-1}_Y(Y)}{\mathsf{E}^{n-1}_Q[X_\theta|Y] Q^{n-1}_Y(Y)} \right]. \quad (4.69)
\end{aligned}$$

Proof. Before proceeding with the proof, we highlight several facts that are

useful later. First, for a given $y \neq 0$,

$$
\begin{aligned}
& y\mathsf{E}^n\left[\frac{X'_\theta}{X_\theta}\middle| Y=y\right] P_Y^n(y) \\
&\quad = \mathsf{E}\left[\frac{\partial f(\theta,X)}{\partial\theta}\frac{n(n-1)!}{(y-1)!(n-y)!}(f(\theta,X))^{y-1}(1-f(\theta,X))^{n-y}\right] \qquad (4.70)\\
&\quad = n\mathsf{E}\left[\frac{\partial f(\theta,X)}{\partial\theta}P_{Y|X}^{n-1}(y-1|X)\right]. \qquad (4.71)
\end{aligned}
$$

Second, for an arbitrary function $\theta(y)$ we have,

$$
\begin{aligned}
\sum_{y=1}^{n}\theta(y)\log\frac{P_Y^n(y)}{Q_Y^n(y)} &= \sum_{y=0}^{n-1}\theta(y+1)\log\frac{P_Y^n(y+1)}{Q_Y^n(y+1)} \qquad (4.72)\\
&= \sum_{y=0}^{n-1}\theta(y+1)\log\frac{\mathsf{E}\left[(f(\theta,X))^{y+1}(1-f(\theta,X))^{n-y-1}\right]}{\mathsf{E}_Q\left[(f(\theta,X))^{y+1}(1-f(\theta,X))^{n-y-1}\right]} \qquad (4.73)\\
&= \sum_{y=0}^{n-1}\theta(y+1)\log\frac{\mathsf{E}\left[f(\theta,X)P_{Y|X}^{n-1}(y|X)\right]}{\mathsf{E}_Q\left[f(\theta,X)P_{Y|X}^{n-1}(y|X)\right]}. \qquad (4.74)
\end{aligned}
$$

In order to prove the lemma, we transform the LHS of (4.69) as follows,

$$
\begin{aligned}
& \sum_{y=0}^{n} y\mathsf{E}\left[\frac{X'_\theta}{X_\theta}\middle| Y=y\right] P_Y^n(y)\log\frac{P_Y^n(y)}{Q_Y^n(y)} \\
&= n\sum_{y=1}^{n}\mathsf{E}\left[\frac{\partial f(\theta,X)}{\partial\theta}P_{Y|X}^{n-1}(y-1|X)\right]\log\frac{P_Y^n(y)}{Q_Y^n(y)} \qquad (4.75)\\
&= n\sum_{y=0}^{n-1}\mathsf{E}\left[\frac{\partial f(\theta,X)}{\partial\theta}P_{Y|X}^{n-1}(y|X)\right]\log\frac{\mathsf{E}\left[f(\theta,X)P_{Y|X}^{n-1}(y|X)\right]}{\mathsf{E}_Q\left[f(\theta,X)P_{Y|X}^{n-1}(y|X)\right]} \qquad (4.76)\\
&= n\sum_{y=0}^{n-1}\mathsf{E}^{n-1}[X'_\theta|Y=y]P_Y^{n-1}(y)\log\frac{\mathsf{E}^{n-1}[X_\theta|Y=y]P_Y^{n-1}(y)}{\mathsf{E}_Q^{n-1}[X_\theta|Y=y]Q_Y^{n-1}(y)}, \qquad (4.77)
\end{aligned}
$$

where we eliminate the first term of the sum ($y = 0$) and apply (4.71) to yield (4.75); we have a change of variable from y to $(y-1)$ and apply (4.74) to yield (4.76); and, to obtain (4.77), we have relied on (4.53).

Finally, we apply the definition of conditional mean estimate and replace $f(\theta, X)$ by X_θ and $\partial f(\theta, X)/\partial\theta$ by X'_θ to establish (4.69). □

Lemma 4.

$$\mathsf{E}^n\left[(n-Y)\mathsf{E}^n\left[\frac{X'_\theta}{1-X_\theta}\middle|Y\right]\log\frac{P_Y^n(Y)}{Q_Y^n(Y)}\right]$$
$$=n\mathsf{E}^{n-1}\left[\mathsf{E}^{n-1}[X'_\theta|Y]\log\frac{\mathsf{E}^{n-1}[1-X_\theta|Y]P_Y^{n-1}(Y)}{\mathsf{E}_Q^{n-1}[1-X_\theta|Y]Q_Y^{n-1}(Y)}\right]. \tag{4.78}$$

Proof. In this proof, initially we highlight two facts that are used later in the proof. First, for every $y\neq n$ we have,

$$(n-y)\mathsf{E}^n\left[\frac{X'_\theta}{1-X_\theta}\middle|Y=y\right]P_Y^n(y)$$
$$=\mathsf{E}\left[\frac{\partial f(\theta,X)}{\partial\theta}\frac{n(n-1)!}{y!(n-y-1)!}(f(\theta,X))^y(1-f(\theta,X))^{n-y-1}\right] \tag{4.79}$$
$$=n\mathsf{E}\left[\frac{\partial f(\theta,X)}{\partial\theta}P_{Y|X}^{n-1}(y|X)\right]. \tag{4.80}$$

Second, for every $y\neq n$,

$$\log\frac{P_Y^n(y)}{Q_Y^n(y)}=\log\frac{\mathsf{E}\left[(f(\theta,X))^y(1-f(\theta,X))^{n-y}\right]}{\mathsf{E}_Q\left[(f(\theta,X))^y(1-f(\theta,X))^{n-y}\right]} \tag{4.81}$$
$$=\log\frac{\mathsf{E}\left[(1-f(\theta,X))P_{Y|X}^{n-1}(y|X)\right]}{\mathsf{E}_Q\left[(1-f(\theta,X))P_{Y|X}^{n-1}(y|X)\right]}. \tag{4.82}$$

To prove Lemma 4, we transform the LHS of (4.78) as follows,

$$\sum_{y=0}^{n}(n-y)\mathsf{E}^n\left[\frac{X'_\theta}{1-X_\theta}\middle|Y=y\right]P_Y^n(y)\log\frac{P_Y^n(y)}{Q_Y^n(y)}$$
$$=n\sum_{y=0}^{n-1}\mathsf{E}\left[\frac{\partial f(\theta,X)}{\partial\theta}P_{Y|X}^{n-1}(y|X)\right]\log\frac{P_Y^n(y)}{Q_Y^n(y)} \tag{4.83}$$
$$=n\sum_{y=0}^{n-1}\mathsf{E}\left[\frac{\partial f(\theta,X)}{\partial\theta}P_{Y|X}^{n-1}(y|X)\right]\log\frac{\mathsf{E}\left[(1-f(\theta,X))\,P_{Y|X}^{n-1}(y|X)\right]}{\mathsf{E}_Q\left[(1-f(\theta,X))\,P_{Y|X}^{n-1}(y|X)\right]} \tag{4.84}$$
$$=n\sum_{y=0}^{n-1}\mathsf{E}^{n-1}[X'_\theta|Y=y]\log\frac{\mathsf{E}^{n-1}[1-X_\theta|Y=y]P_Y^{n-1}(y)}{\mathsf{E}_Q^{n-1}[1-X_\theta|Y=y]Q_Y^{n-1}(y)}P_Y^{n-1}(y), \tag{4.85}$$

where in (4.83) we have used (4.80) and eliminated the last term of the sum because it is zero; in (4.84) we write the expression in (4.83), using (4.82),

in terms of binomial distributions with parameters $(n-1, f(\theta, x))$; finally we apply (4.53) to obtain (4.85). □

Lemma 5.

$$\mathsf{E}^n\left[Y\mathsf{E}_Q^n\left[\frac{X'_\theta}{X_\theta}\middle|Y\right]\right] = n\mathsf{E}^{n-1}\left[\mathsf{E}^{n-1}[X_\theta|Y]\frac{\mathsf{E}_Q^{n-1}[X'_\theta|Y]}{\mathsf{E}_Q^{n-1}[X_\theta|Y]}\right]. \tag{4.86}$$

Lemma 6.

$$\mathsf{E}^n\left[(n-Y)\mathsf{E}_Q^n\left[\frac{X'_\theta}{1-X_\theta}\middle|Y\right]\right] = n\mathsf{E}^{n-1}\left[\mathsf{E}_Q^{n-1}[X'_\theta|Y]\frac{\mathsf{E}^{n-1}[1-X_\theta|Y]}{\mathsf{E}_Q^{n-1}[1-X_\theta|Y]}\right]. \tag{4.87}$$

We omit the proof of Lemmas 5 and 6 as they are easily obtained using proofs of Lemmas 3 and 4, respectively.

Proof of (4.15) *in Theorem 11.* This proof is based on writing the derivatives of P_Y^n and Q_Y^n in terms of conditional estimations. By definition, the derivative of the relative entropy with respect to θ is given by,

$$\frac{\mathrm{d}}{\mathrm{d}\theta}D(P_Y^n\|Q_Y^n) = \underbrace{\sum_{y=0}^{n}\frac{\mathrm{d}P_Y^n(y)}{\mathrm{d}\theta}\log\frac{P_Y^n(y)}{Q_Y^n(y)}}_{(\mathbb{I})} - \underbrace{\sum_{y=0}^{n}\frac{P_Y^n(y)}{Q_Y^n(y)}\frac{\mathrm{d}Q_Y^n(y)}{\mathrm{d}\theta}}_{\mathbb{III}}, \tag{4.88}$$

where the derivative penetrates the sum by the finiteness of the output alphabet $\mathcal{Y}$. Calculation of the first term $(\mathbb{I})$ is obtained as follows,

$$\begin{aligned}
&\sum_{y=0}^{n}\frac{\mathrm{d}P_Y^n(y)}{\mathrm{d}\theta}\log\frac{P_Y^n(y)}{Q_Y^n(y)} \\
&= \sum_{y=0}^{n} y\,\mathsf{E}^n\left[\frac{X'_\theta}{X_\theta}\middle|Y=y\right]\log\frac{P_Y^n(y)}{Q_Y^n(y)}P_Y^n(y) \\
&\quad - \sum_{y=0}^{n}(n-y)\mathsf{E}^n\left[\frac{X'_\theta}{1-X_\theta}\middle|Y=y\right]P_Y^n(y)\log\frac{P_Y^n(y)}{Q_Y^n(y)} \qquad (4.89)\\
&= n\sum_{y=0}^{n-1}\mathsf{E}^{n-1}[X'_\theta|Y=y]\log\frac{\mathsf{E}^{n-1}[X_\theta|Y=y](1-\mathsf{E}_Q^{n-1}[X_\theta|Y=y])}{\mathsf{E}_Q^{n-1}[X_\theta|Y=y](1-\mathsf{E}^{n-1}[X_\theta|Y=y])}P_Y^{n-1}(y), \qquad (4.90)
\end{aligned}$$

where (4.89) is a consequence of the exchangeability property illustrated in Lemma 2, (4.90) is the expression obtained in (4.89) written in terms of the conditional mean estimates over a binomial channel with $n-1$ trials, which is proven explicitly in Lemmas 3 and 4.

The second term of (4.88), ($\mathbb{III}$), simplifies to,

$$\sum_{y=0}^{n} \frac{P_Y^n(y)}{Q_Y^n(y)} \frac{\mathrm{d}Q_Y^n(y)}{\mathrm{d}\theta}$$

$$= \sum_{y=0}^{n} y \, \mathsf{E}_Q^n \left[\frac{X'_\theta}{X_\theta} \middle| Y = y \right] P_Y^n(y) - \sum_{y=0}^{n} (n-y) \, \mathsf{E}_Q^n \left[\frac{X'_\theta}{1-X_\theta} \middle| Y = y \right] P_Y^n(y) \tag{4.91}$$

$$= n \sum_{y=0}^{n-1} \mathsf{E}_Q^{n-1}[X'_\theta | Y = y] \frac{\mathsf{E}^{n-1}[X_\theta | Y = y] - \mathsf{E}_Q^{n-1}[X_\theta | Y = y]}{\mathsf{E}_Q^{n-1}[1 - X_\theta | Y = y] \mathsf{E}_Q^{n-1}[X_\theta | Y = y]} P_Y^{n-1}(y), \tag{4.92}$$

where (4.91) appears as consequence of the exchangeability property and (4.92) is due to Lemmas 5 and 6.

Putting together expressions obtained in (4.90) and (4.92) yields the desired result. □

4.4.4 Proof of Theorem 12.

We prove Theorem 12 by applying Theorem 11 to the relation between the mutual information and relative entropy,

$$I(X;Y) = \int D(P_{Y|X=x}^n \| P_Y^n) \mathrm{d}P_X(x), \tag{4.93}$$

in which $P_{Y|X=x}^n$ represents the output distribution of a binomial deterministic model with input distribution being a point mass at x, denoted as $\delta(x)$. The derivative of the mutual information with respect to θ is given

by,

$$\frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) = \int \frac{\mathrm{d}}{\mathrm{d}\theta} D(P^n_{Y|X=x} \| P^n_Y) \mathrm{d}P_X(x) \tag{4.94}$$

$$= \int \left(\mathsf{F}^{n-1}_{\delta(x),P}(f(\theta,X)) - \mathsf{F}^{n-1}_{\delta(x),\delta(x)}(f(\theta,X)) \right) \mathrm{d}P_X(x) \tag{4.95}$$

$$= \int \mathsf{F}^{n-1}_{\delta(x),P}(f(\theta,X)) \mathrm{d}P_X(x) \tag{4.96}$$

$$= \mathsf{F}^{n-1}_{P}(X_\theta) \tag{4.97}$$

$$= n\, \mathsf{E}^{n-1} \left[X'_\theta \log \frac{X_\theta (1 - \mathsf{E}^{n-1}[X_\theta|Y])}{(1-X_\theta)\mathsf{E}^{n-1}[X_\theta|Y]} \right], \tag{4.98}$$

where[4],

$$\mathsf{F}^{n-1}_{\delta(x),P}(f(\theta,X)) = n\mathsf{E}_{P^{n-1}_{Y|X=x}} \left[\frac{\partial f(\theta,x)}{\partial \theta} \log \frac{f(\theta,x)(1-\mathsf{E}^{n-1}[f(\theta,X)|Y])}{(1-f(\theta,x))\mathsf{E}^{n-1}[f(\theta,X)|Y]} - \frac{\mathsf{E}^{n-1}\left[\frac{\partial f(\theta,X)}{\partial \theta} \middle| Y \right] (f(\theta,x) - \mathsf{E}^{n-1}[f(\theta,X)|Y])}{(1-\mathsf{E}^{n-1}[f(\theta,X)|Y])\mathsf{E}^{n-1}[f(\theta,X)|Y]} \right]. \tag{4.99}$$

This definition is different from the one in (4.16), because the true input distribution is a delta function and we only take the expectation with respect to $P^{n-1}_{Y|X=x}$ and not P^{n-1}_{XY}. In (4.96), we use the fact that $\mathsf{F}^{n-1}_{\delta(x),\delta(x)}(f(\theta,X)) = 0$, because $\mathsf{E}^{n-1}_{\delta(x)}[f(\theta,X)|Y] = f(\theta,x)$. In (4.97), we make use of the definition in (4.16), when we integrate over P_X. Finally (4.98) is due to

$$\mathsf{E}^{n-1} \left[\frac{\mathsf{E}^{n-1}[X'_\theta|Y](X_\theta - \mathsf{E}^{n-1}[X_\theta|Y])}{(1-\mathsf{E}^{n-1}[X_\theta|Y])\mathsf{E}^{n-1}[X_\theta|Y]} \right]$$

$$= \mathsf{E}^{n-1} \left[\mathsf{E}^{n-1} \left[\frac{\mathsf{E}^{n-1}[X'_\theta|Y](X_\theta - \mathsf{E}^{n-1}[X_\theta|Y])}{(1-\mathsf{E}^{n-1}[X_\theta|Y])\mathsf{E}^{n-1}[X_\theta|Y]} \middle| Y \right] \right] \tag{4.100}$$

$$= \mathsf{E}^{n-1} \left[\frac{\mathsf{E}^{n-1}[X'_\theta|Y](\mathsf{E}^{n-1}[X_\theta|Y] - \mathsf{E}^{n-1}[X_\theta|Y])}{(1-\mathsf{E}^{n-1}[X_\theta|Y])\mathsf{E}^{n-1}[X_\theta|Y]} \right] \tag{4.101}$$

$$= 0. \tag{4.102}$$

[4]In (4.99), the subscript P in $\mathsf{F}^{n-1}_{\delta(x),P}(X_\theta)$ makes reference to the fact that the conditionals $\mathsf{E}^{n-1}[X_\theta|Y]$ at the RHS of the equation are with respect to the distribution induced by $P_X \times P^{n-1}_{Y|X}$.

4.4.5 Proof of (4.7) in Theorem 8.

Replacing X_θ by θX in (4.15), we get,

$$\frac{\mathrm{d}}{\mathrm{d}\theta} D(P_Y^n \| Q_Y^n) = \mathsf{F}_Q^{n-1}(\theta X) - \mathsf{F}_P^{n-1}(\theta X) \tag{4.103}$$

$$= n\mathsf{E}^{n-1}\left[X \log \frac{1 - \mathsf{E}_Q^{n-1}[\theta X|Y]}{\mathsf{E}_Q^{n-1}[\theta X|Y]} - \frac{(X - \mathsf{E}_Q^{n-1}[X|Y])}{1 - \mathsf{E}_Q^{n-1}[\theta X|Y]}\right] - n\mathsf{E}^{n-1}\left[X \log \frac{1 - \mathsf{E}^{n-1}[\theta X|Y]}{\mathsf{E}^{n-1}[\theta X|Y]}\right] \tag{4.104}$$

$$= \frac{n}{\theta}\mathsf{E}^{n-1}\left[\theta X \log \frac{(1 - \mathsf{E}_Q^{n-1}[\theta X|Y])\mathsf{E}^{n-1}[\theta X|Y]}{\mathsf{E}_Q^{n-1}[\theta X|Y](1 - \mathsf{E}^{n-1}[\theta X|Y])}\right] - \frac{n}{\theta}\mathsf{E}^{n-1}\left[\frac{(\mathsf{E}^{n-1}[\theta X|Y] - \mathsf{E}_Q^{n-1}[\theta X|Y])}{1 - \mathsf{E}_Q^{n-1}[\theta X|Y]}\right] \tag{4.105}$$

$$= \frac{n}{\theta}\mathsf{E}^{n-1}\left[\ell_b(\mathsf{E}^{n-1}[\theta X|Y], \mathsf{E}_Q^{n-1}[\theta X|Y])\right]. \tag{4.106}$$

4.4.6 Proof of (4.10) in Theorem 9.

Replacing X_θ by θX in (4.19), we get,

$$\frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) = \mathsf{F}_P^{n-1}(\theta X) \tag{4.107}$$

$$= n\mathsf{E}^{n-1}\left[X \log \frac{\theta X(1 - \mathsf{E}^{n-1}[\theta X|Y])}{(1-\theta X)\mathsf{E}^{n-1}[\theta X|Y]} - \frac{(X - \mathsf{E}^{n-1}[X|Y])}{1 - \mathsf{E}^{n-1}[\theta X|Y]}\right] \tag{4.108}$$

$$= \frac{n}{\theta}\mathsf{E}^{n-1}\left[\ell_b(\theta X, \mathsf{E}^{n-1}[\theta X|Y])\right]. \tag{4.109}$$

4.4.7 Proof of Theorem 13

Let d_b denote the Bregman divergence associated with the exponential form of the binomial distribution given in (3.50). Then, for $a, \hat{a} \in (0, n)$,

$$d_b(a, \hat{a}) = n\, d_{b^*}\left(\frac{a}{n}, \frac{\hat{a}}{n}\right) \tag{4.110}$$

where d_{b^*} is the Bregman divergence associated with the convex function

$$b^*(a) = a \log \frac{a}{1-a} - \log \frac{1}{1-a}. \tag{4.111}$$

Using the linearity of the Bregman divergences we obtain that,

$$d_{b^*}\left(\frac{a}{n},\frac{\hat{a}}{n}\right) = \ell_b\left(\frac{a}{n},\frac{\hat{a}}{n}\right) - d_{\overline{\phi}}\left(\frac{a}{n},\frac{\hat{a}}{n}\right), \tag{4.112}$$

$$> 0, \tag{4.113}$$

where $d_{\overline{\phi}}(a,\hat{a})$ constitutes the Bregman divergence build upon the convex function $\overline{\phi}(a) = -\log(1-a)$. Hence, for a given $Y = y \in \mathbb{Z}_0^+$, for $a = \theta X$ and $\hat{a}(y) = \mathsf{E}^{n-1}[\theta X|Y=y]$ we get,

$$\frac{1}{\theta}\mathsf{E}^{n-1}\left[d_b\left(\theta X, \mathsf{E}^{n-1}[\theta X|Y]\right)\right] < \frac{n}{\theta}\mathsf{E}^{n-1}\left[\ell_b\left(\frac{\theta X}{n}, \mathsf{E}^{n-1}\left[\frac{\theta X}{n}\middle|Y\right]\right)\right] \tag{4.114}$$

$$= \frac{\mathrm{d}}{\mathrm{d}\theta}I(X;Y), \tag{4.115}$$

where to obtain (4.115) we use the expression found in Theorem 12 with $X_\theta = \theta X/n$. The equivalent expression for the derivative of the relative entropy is consequence of the inequalities given by (4.112) and (4.113).

4.4.8 Proof of Theorem 14

Carrying out the expectation over (4.10) let us get that,

$$\frac{\mathrm{d}}{\mathrm{d}\theta}I(X;Y) = n\mathsf{E}^{n-1}\left[X\log\frac{X(1-\theta\mathsf{E}^{n-1}[X|Y])}{(1-\theta X)\mathsf{E}^{n-1}[X|Y]}\right].$$

Therefore, calculating the limit, we obtain,

$$\lim_{\theta\to 0}\frac{\mathrm{d}}{\mathrm{d}\theta}I(X;Y) = n\mathsf{E}[X\log X] - n\lim_{\theta\to 0}\sum_{y=0}^{n-1}A(y)\log\frac{A(y)}{B(y)-\theta A(y)}, \tag{4.116}$$

where we use the shorthands $A(y) \triangleq \mathsf{E}[XP_{Y|X}^{n-1}(y|X)]$ and $B(y) \triangleq \mathsf{E}[P_{Y|X}^{n-1}(y|X)]$. First notice that,

$$\lim_{\theta\to 0}A(y) = \lim_{\theta\to 0}\binom{n-1}{y}\mathsf{E}[X(\theta X)^y(1-\theta X)^{n-y-1}]$$

$$= \lim_{\theta\to 0}\binom{n-1}{y}\sum_{k=0}^{n-y-1}\binom{n-y-1}{k}(-1)^k\theta^{k+y}\mathsf{E}[X^{k+y+1}] \tag{4.117}$$

$$= \mathsf{E}[X]\mathbb{1}_{\{y=0\}}, \tag{4.118}$$

where $\mathbb{1}_{\{U\}}$ is the indicator function of U, which is equal to 1 if condition U is satisfied and equal to 0 otherwise. Following a similar procedure, we obtain that,

$$\lim_{\theta \to 0} B(y) = \mathbb{1}_{\{y=0\}}. \tag{4.119}$$

Based on (4.118), we get that

$$\lim_{\theta \to 0} \sum_{y=0}^{n-1} A(y) \log A(y) = \mathsf{E}[X] \log \mathsf{E}[X]. \tag{4.120}$$

Additionally, as consequence of (4.117) notice that the expression $A(y)$ is a polynomial in θ with exponents between y and $(n-1)$. Similarly the function $B(y) - \theta A(y)$ is a polynomial in θ with exponents between y and n. Therefore, for $y \neq 0$ we get

$$\lim_{\theta \to 0} A(y) \log(B(y) - \theta A(y)) = - \lim_{\theta \to 0} \frac{(A(y))^2 \frac{\mathrm{d}}{\mathrm{d}\theta}(B(y) - \theta A(y))}{(B(y) - \theta A(y)) \frac{\mathrm{d}}{\mathrm{d}\theta} A(y)} \tag{4.121}$$

$$= 0, \tag{4.122}$$

where in (4.121) we use L'Hospital's rule [47], and (4.122) can be stated once we show that the polynomial obtained in the numerator in (4.121) contains terms with higher degree than its counterparts in the denominator. In effect, notice that $(A(y))^2 \frac{\mathrm{d}}{\mathrm{d}\theta} (B(y) - \theta A(y))$ is a polynomial where the least exponent of θ is $2y + (y-1) = 3y - 1$, meanwhile, for $(B(y) - \theta A(y)) \frac{\mathrm{d}}{\mathrm{d}\theta} A(y)$ the least exponent on θ is $y + (y-1) = 2y - 1$. Then, when we multiply and divide by θ^{2y-1}, the grade of the polynomial in the numerator is greater than the grade of the polynomial in the denominator, fact that let us conclude that the entire expression, in the limit, tends to zero. For $y = 0$ we have

$$\lim_{\theta \to 0} A(y) \log \left(B(y) - \theta A(y)\right) = 0. \tag{4.123}$$

Replacing (4.120), (4.122) and (4.123) in (4.116) let us get the desired result.

4.4.9 Proof of Corollary 2

In this case, the derivative of the mutual information with respect to θ is given by Theorem 12,

$$\frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) = \mathsf{E}^{n-1}\left[X \log \frac{X(1 - \frac{1}{n}\mathsf{E}^{n-1}[\theta X|Y])}{(1 - \frac{\theta X}{n})\mathsf{E}^{n-1}[X|Y]} \right].$$

Then, carrying out a similar procedure to that shown in the proof of Theorem 14, let us obtain the desired result.

4.4.10 Proof of Corollary 3

Let Y be the output of a binomial model with parameters $(n, \theta X/n)$. Then, by Theorem 11, the derivative of the relative entropy between the marginals P_Y^n and Q_Y^n with $X_\theta = \theta X/n$, is given by,

$$\frac{\mathrm{d}}{\mathrm{d}\theta} D(P_Y^n || Q_Y^n) = \mathsf{F}_Q\left(\frac{\theta X}{n}\right) - \mathsf{F}_P\left(\frac{\theta X}{n}\right) \tag{4.124}$$

$$= \mathsf{E}^{n-1}\left[X \log \frac{\left(1 - \mathsf{E}_Q^{r+1}\left[\,\frac{\theta X}{n}\middle|\, Y\right]\right)\mathsf{E}^{n-1}[X|Y]}{\left(1 - \mathsf{E}^{n-1}\left[\frac{\theta X}{n}|Y\right]\right)\mathsf{E}_Q^{n-1}\left[X|\,Y\right]}\right]$$

$$- \mathsf{E}^{n-1}\left[\frac{\left(\mathsf{E}^{n-1}[X|Y] - \mathsf{E}_Q^{n-1}\left[X|\,Y\right]\right)}{\left(1 - \mathsf{E}_Q^{n-1}\left[\,\frac{\theta X}{n}\middle|\, Y\right]\right)}\right]. \tag{4.125}$$

We get the desired result once we make θ tend to zero inside the expectation in (4.125). This procedure is justified in the proof of Theorem 14 in Section 4.4.8.

Chapter 5

Negative Binomial Model

In this chapter[1] we explore several expressions linking the information field with the estimation field for the negative binomial model. The negative binomial model can be interpreted as an over-dispersed Poisson model, as the variance of the negative binomial is always larger than its mean. From a Bayesian perspective, if the mean of a Poisson random variable is gamma distributed, the unconditional distribution (integrate out the mean) is a negative binomial. From this viewpoint, the negative binomial can be used to model lasers or LEDs whose mean number of photons vary in each firing.

The methodology employed throughout this chapter is similar to that used in Chapter 4 for the binomial model. Initially assuming a linear preprocessing of the input $X_\theta = \theta X$ we show that information measures such as the relative entropy and mutual information, find an alternative representation in terms of conditional estimates through the expectation of a Bregman divergence ℓ_{nb}. This fact leads to different consequences in the behavior of the mutual information, similar to those developed for models such as the Gaussian, Poisson and binomial. Later, assuming an arbitrary deterministic preprocessing of the input X_θ we present a general information-estimation expression that in some cases is given through a Bregman divergence. As consequence of these expressions we state a relationship between the Bregman divergence used in the exponential representation of the negative binomial distribution (3.55) and the Bregman divergence ℓ_{nb} used to express the derivative of some information measures. Finally, based on the results obtained over the linear preprocessing of the input $X_\theta = \theta X$ we show that, over certain scenarios, the low input scaling

[1] Some results presented through this chapter were published jointly with Professors F. Pérez-Cruz and D. Guo in [54, 57, 56, 21, 55].

behavior of the relative entropy and mutual information is similar to that found for the binomial model.

5.1 Model definition

The negative binomial distribution of order r is defined by the following pmf:

$$P(Y = y) = \binom{y + r - 1}{y}(1 - q)^r q^y, \quad y \in \{0, 1, \dots\} \triangleq \mathbb{Z}_0^+, \tag{5.1}$$

which is the probability that y successful trials are seen before the r-th failure occurs, where the trials are independent Bernoulli random variables, each with probability q to succeed. We refer to (5.1) as a negative binomial distribution with parameters (r, q).

We define a negative binomial model as a random transformation from a random variable X to an integer random variable Y, where, conditioned on $X = x$, Y has negative binomial distribution with parameters $(r, f(\theta, x)/(1+ f(\theta, x))$. In this case the mapping $f(\theta, X)$ represents a preprocessing of the input which depends on a parameter θ. Therefore, the random transformation that governs the model is given by the following conditional pmf,

$$P_{Y|X}^r(y|x) = \binom{y + r - 1}{y}\left(\frac{f(\theta, x)}{1 + f(\theta, x)}\right)^y \left(\frac{1}{1 + f(\theta, x)}\right)^r, \quad y \in \mathbb{Z}_0^+. \tag{5.2}$$

For notational convenience we use the shorthand

$$X_\theta = f(\theta, X). \tag{5.3}$$

We further assume that the function X_θ is differentiable and denote its derivative as,

$$X'_\theta = \frac{\partial f(\theta, X)}{\partial \theta}. \tag{5.4}$$

Throughout this section we employ similar conventions to those used in the case of the binomial model. In particular, we use P_X and Q_X to denote two input probability laws and P_Y^r and Q_Y^r to denote the corresponding output distributions, where the underlying negative binomial model is of order r.

5.2 Information-Estimation Relationships

In this section we prove that, based on alternative expressions found for the negative binomial model, all of those consequences explored previously for models such as the Gaussian, Poisson and binomial, translate akin to the context of the negative binomial model.

5.2.1 Linear Scaling ($X_\theta = \theta X$)

We start with a simple linear scaling, where $X_\theta = \theta X$ is positive and bounded for some $\theta > 0$.

Theorem 16. *Let X be a positive bounded random variable with distribution P_X or Q_X. Let Y be the output of an r-th order negative binomial model described by* (5.2) *with $X_\theta = \theta X$. Then,*

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\theta} D(P_Y^r \| Q_Y^r) \\
&= \frac{r}{\theta}\left(\mathsf{E}^{r+1}\left[\ell_{nb}\left(\theta X, \mathsf{E}_Q^{r+1}[\theta X|Y]\right)\right] - \mathsf{E}^{r+1}\left[\ell_{nb}\left(\theta X, \mathsf{E}^{r+1}[\theta X|Y]\right)\right]\right) \quad (5.5)\\
&= \frac{r}{\theta}\mathsf{E}^{r+1}\left[\ell_{nb}(\mathsf{E}^{r+1}\left[\theta X|\, Y\right], \mathsf{E}_Q^{r+1}\left[\theta X|\, Y\right])\right] \quad (5.6)
\end{aligned}$$

and

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\theta} D(P_Y^r \| Q_Y^r) &= \mathsf{E}^r\left[\frac{Y}{\theta}\ell_I\left(X^{-1}+\theta, \mathsf{E}_Q^r[X^{-1}|Y]+\theta\right)\right] \\
&\quad - \mathsf{E}^r\left[\frac{Y}{\theta}\ell_I\left(X^{-1}+\theta, \mathsf{E}^r[X^{-1}|Y]+\theta\right)\right] \quad (5.7)\\
&= \mathsf{E}^r\left[\frac{Y}{\theta}\ell_I\left(\mathsf{E}^r[X^{-1}|Y]+\theta, \mathsf{E}_Q^r[X^{-1}|Y]+\theta\right)\right] \quad (5.8)
\end{aligned}$$

hold for all $\theta > 0$.

Proof. See Section 5.4.5 for the proof of (5.6) and Section 5.4.1 for the proof (5.8). □

As in the binomial model, the derivative of the relative entropy admits two different representations; one considers conditional estimates over negative binomial models of order $(r+1)$, and the other involves conditional estimates over models of order r.

A related result connects the derivative of the input–output mutual information to the expected estimation error under two different Bregman divergences.

Theorem 17. *Let X be a positive bounded random variable with distribution P_X. Let Y be the output of a r-th order negative binomial model described by (5.2) with $X_\theta = \theta X$. Then,*

$$\frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) = \frac{r}{\theta} \mathsf{E}^{r+1}\left[\ell_{nb}(\theta X, \mathsf{E}^{r+1}\left[\theta X \middle| Y\right])\right] \tag{5.9}$$

$$= \mathsf{E}^{r}\left[\frac{Y}{\theta}\ell_I\left(X^{-1}+\theta, \mathsf{E}^{r}[X^{-1}|Y]+\theta\right)\right] \tag{5.10}$$

hold for all $\theta > 0$.

Proof. See Section 5.4.6 for the proof of (5.9) and Section 5.4.2 for the proof of (5.10). □

One striking property that appears again, but now in the context of the negative binomial model is the fact that the mutual information can alternatively be represented through the expectation of the function ℓ_P; function that gives rise to the "I-MMLE" relationship for the Poisson model, proved initially in [25].

Corollary 4. *Assume the same set of conditions used in Theorem 17. Then,*

$$\frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) = r\mathsf{E}^{r+1}\left[\ell_P\left(X, \frac{\mathsf{E}^{r+1}[X|Y](1+\theta X)}{1+\theta\mathsf{E}^{r+1}[X|Y]}\right)\right]. \tag{5.11}$$

Similarly to the case studied for the binomial model, notice that even though (5.9) and (5.11) are mathematically identical, the expectation in (5.9) achieves its minimum value meanwhile the expectation in (5.11) does not. This is a direct consequence of the arguments used over each Bregman divergence. As consequence of Property *(iii)* stated in Theorem 5 we can state the following result.

Theorem 18. *Let X and Y be defined as in Theorem 16. Then,*

$$\mathsf{E}^{r+1}\left[\ell_{nb}(\theta X, \mathsf{E}_Q^{r+1}\left[\theta X \middle| Y\right])\right] = \frac{\theta}{r}\frac{\mathrm{d}}{\mathrm{d}\theta}\left[D(P_Y^r \| Q_Y^r) + I(X;Y)\right] \tag{5.12}$$

holds for all $\theta > 0$.

Theorem 18 shows that the mismatched estimation penalty incurred when we estimate the random variable X through the mismatched prior distribution Q_X is proportional to the sum of derivative of the relative entropy between P_X and Q_X and the input-output mutual information.

In addition, the minimum value achieved by the LHS of (5.12) is proportional to the derivative of the mutual information, as was pointed out in Theorem 10 for the binomial model.

5.2.2 Arbitrary Scaling

We now generalize Theorem 16 and 17 to the case where the input X_θ depends on θ in an arbitrary manner. Throughout this section, the set of feasible values for the parameter θ, denoted as Θ is an open real number set. We also assume that the function X_θ is always positive and bounded.

Theorem 19. *Under both distributions, P_X and Q_X, let X_θ and X'_θ be integrable and bounded. Let Y be the output of the r-th order negative binomial model described by (5.2) with X_θ as input. Then,*

$$\frac{\mathrm{d}}{\mathrm{d}\theta} D(P_Y^r \| Q_Y^r) = \mathsf{G}_Q^{r+1}(X_\theta) - \mathsf{G}_P^{r+1}(X_\theta) \tag{5.13}$$

holds for all $\theta \in \Theta$, where $\mathsf{G}_Q^{r+1}(X_\theta)$ is given by,

$$\begin{aligned}\mathsf{G}_Q^{r+1}(X_\theta) = r\mathsf{E}^{r+1}&\left[X'_\theta \log \frac{\left(1+\mathsf{E}_Q^{r+1}\left[X_\theta|\,Y\right]\right)X_\theta}{(1+X_\theta)\,\mathsf{E}_Q^{r+1}\left[X_\theta|\,Y\right]}\right]\\ &- r\mathsf{E}^{r+1}\left[\frac{\mathsf{E}_Q^{r+1}[X'_\theta|Y]\left(X_\theta - \mathsf{E}_Q^{r+1}\left[X_\theta|\,Y\right]\right)}{\left(1+\mathsf{E}_Q^{r+1}\left[X_\theta|\,Y\right]\right)\mathsf{E}_Q^{r+1}\left[X_\theta|\,Y\right]}\right].\end{aligned} \tag{5.14}$$

Proof. See Section 5.4.3. □

The generalized version of Theorem 17 is stated as follows.

Theorem 20. *Let X_θ, X'_θ and Y be defined as in Theorem 19. Then,*

$$\frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) = \mathsf{G}_P^{r+1}(X_\theta) \tag{5.15}$$

$$= r\mathsf{E}^{r+1}\left[X'_\theta \log \frac{\left(1+\mathsf{E}^{r+1}\left[X_\theta|\,Y\right]\right)X_\theta}{(1+X_\theta)\mathsf{E}^{r+1}\left[X_\theta|\,Y\right]}\right] \tag{5.16}$$

holds for all $\theta \in \Theta$.

Proof. See Section 5.4.4. □

As we did in Section 4.2.2, we let $X_\theta = X + \theta$ where $\theta \in \mathbb{R}^+$ and compute the derivative of the mutual information using Theorem 20,

$$\frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) = -r\mathsf{E}^{r+1}\left[\ell_\psi\left(X+\theta, \mathsf{E}^{r+1}[X+\theta|Y]\right)\right], \tag{5.17}$$

where $\ell_\psi(\cdot,\cdot)$ is the Bregman divergence build with $\psi(a) = -\log\frac{a}{1+a}$ on $\mathbb{R}^+$. Expression given in (5.17) suggests that the mutual information decays when the parameter θ increases. This behavior is similar to that found over a Poisson model with mean $X+\theta$ given in [25]. In Section 4.2.2 we showed a case where the result could not be expressed as a Bregman divergence, whereas here the reverse is true.

A second application of the results stated in Theorems 19 and 20, where the input of the model is preprocessed through the function X_θ is stated as follows. In the search of a relationship between the Bregman divergence used in the exponential form of the negative binomial distribution (3.55) and the function ℓ_{nb} we arrive to the following theorem, which in words, states that the expectation of the function used in the exponential form of the negative binomial distribution in (3.55), up to a scaling factor, constitutes an upper bound for the derivative of the mutual information. A similar upper bound is stated in the case of the derivative of the relative entropy.

Theorem 21. *Let X be a positive bounded random variable that can be distributed as either P_X or Q_X. Let Y be the output of a r-th negative binomial model with parameters $\left(r, \frac{\theta X}{r+\theta X}\right)$. Then,*

$$\frac{1}{\theta}\mathsf{E}^{r+1}\left[d_{nb}\left(\theta X, \mathsf{E}^{r+1}[\theta X|Y]\right)\right] > \frac{\mathrm{d}}{\mathrm{d}\theta}I(X;Y) \tag{5.18}$$

and

$$\frac{1}{\theta}\mathsf{E}^{r+1}\left[d_{nb}\left(\mathsf{E}^{r+1}[\theta X|Y], \mathsf{E}_Q^{r+1}[\theta X|Y]\right)\right] > \frac{\mathrm{d}}{\mathrm{d}\theta}D(P_Y^r||Q_Y^r), \tag{5.19}$$

where we denote d_{nb} as the Bregman divergence used in (3.55) *to build the exponential form of the negative binomial distribution.*

Proof. See Section 5.4.7. □

5.2.3 Low input scaling

As was pointed out in Section 4.2.3, the low input scaling behavior of several models has been studied previously, showing these results their own merits in terms of efficiency for the Gaussian, Poisson and binomial models. In this section, we explore the low input scaling behavior pertaining the negative binomial model.

Based on the expression given for the derivative of the input-output mutual information, in the following theorem we show that the low input

scaling behavior in the negative binomial model can be expressed in terms of the Bregman divergence ℓ_P used to describe the mutual information when dealing with Poisson models.

Theorem 22. *Let $X \sim P_X$ be a positive bounded random variable. Let Y be the output of the r-order negative binomial model described by* (5.2) *with $X_\theta = \theta X$. Then,*

$$\lim_{\theta \to 0} \frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) = r\mathsf{E}\left[\ell_P(X, \mathsf{E}[X])\right] \tag{5.20}$$

$$= r\mathsf{E}\left[X \log \frac{X}{\mathsf{E}[X]}\right]. \tag{5.21}$$

Proof. See Section 5.4.8. □

Similarly as was stated in the case of the binomial model, Theorem 22 shows the importance of the function $\mathsf{E}[\ell_P(X, \mathsf{E}[X])]$ at low input scaling factors for the negative binomial model. At the end, we can conclude that up to a scaling factor, the low input scaling behavior of the binomial and negative binomial models is the same. Equality on their behavior is stated formally as follows.

Corollary 5. *Let $X \sim P_X$ be a positive bounded random variable. Let Y be the output of the r-order negative binomial model described by* (5.2) *with $X_\theta = \theta X/r$. Then,*

$$\lim_{\theta \to 0} \frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) = \mathsf{E}\left[\ell_P(X, \mathsf{E}[X])\right] \tag{5.22}$$

$$= \mathsf{E}\left[X \log \frac{X}{\mathsf{E}[X]}\right]. \tag{5.23}$$

Proof. See Section 5.4.9. □

The expression for the relative entropy at low input scaling regime is stated as follows.

Theorem 23. *Let X be a positive bounded random variable that can be distributed as either P_X or Q_X. Let Y be the output of the r-order negative binomial model described by* (5.2) *with $X_\theta = \theta X$. Then,*

$$\lim_{\theta \to 0} \frac{\mathrm{d}}{\mathrm{d}\theta} D(P_Y^r || Q_Y^r) = r\ell_P(\mathsf{E}[X], \mathsf{E}_Q[X]). \tag{5.24}$$

Proof. The proof to this theorem is similar to that developed to state Theorem 22, shown in Section 5.4.8. □

Corollary 6. *Let X be a positive bounded random variable that can be distributed as either P_X or Q_X. Let Y be the output of the r-order negative binomial model described by (5.2) with $X_\theta = \theta X/r$. Then,*

$$\lim_{\theta \to 0} \frac{\mathrm{d}}{\mathrm{d}\theta} D(P_Y^r || Q_Y^r) = \ell_P(\mathsf{E}[X], \mathsf{E}_Q[X]). \tag{5.25}$$

Proof. See Section 5.4.10. □

5.3 Concluding Remarks

In this chapter, we show several connections between information and estimation for the negative binomial model, which can be used over specific circumstances as a statistical model for photon-emission based systems.

The most prominent expression found for the negative binomial model, states a relationship between information measures such as the relative entropy and mutual information and the conditional mean estimate through a Bregman divergence ℓ_{nb}. This fact reveals that the information measures that we are treating behave similarly over all the statistical models studied before. Specifically, we can state that both, mutual information and relative entropy are increasing in the value of θ; the expression for the derivative of the mutual information corresponds to the minimum value attained by the expectation of a Bregman divergence; the expectation of the Bregman divergence ℓ_{nb} over a mismatched prior Q_X is directly related with the derivative of the sum between the relative entropy and the mutual information, etc. Additionally, we highlight that, at first sight, expressions given by (5.9) and (5.11) are equal but they hide one property in the background. While the function ℓ_{nb} attains its minimum value when is evaluated with arguments θX and $\theta \mathsf{E}^{r+1}[X|Y]$ the function ℓ_P used in (5.11) does not attain its minimum value at those values at which it is evaluated (Theorem 6 in Chapter 2). This difference shows the importance of the function ℓ_{nb} in the case of the negative binomial model.

Assuming an arbitrarily deterministic preprocessing of the input X_θ we state a general expression linking information measures with conditional estimates. This approach has several advantages:

- Over several scenarios in the negative binomial model we show that the information–estimation relationship is still characterized by a Bregman

divergence, even though this is not always the case and is not always through the function ℓ_{nb}. In other words, depending on the shape of X_θ in some cases the information–estimation relationship is through a Bregman divergence different to ℓ_{nb}.

- In the search of a relationship between the functions d_{nb} and ℓ_{nb} we show that, up to a scaling factor, the derivative of the input–output mutual information, expressed through the function ℓ_{nb} is a lower bound for the expectation of the Bregman divergence d_{nb} where the arguments are the mean of the model θX and its conditional estimate $\mathsf{E}^{r+1}[\theta X|Y]$. Recall from the analysis given in Section (3.2) that, d_{nb} is the Bregman divergence used in the exponential form of the negative binomial distribution. A similar argument can be applied in the case of the derivative of the relative entropy.

- As was stated in the binomial case, identifying those scenarios over which the mutual information has a monotone behavior with respect to changes in the parameter θ is useful in the context of the broadcast channels given that the monotonicity gives rise to the "More Capable" Broadcast channels. This kind of channels have a completely characterized capacity region [19, 36, 20].

At low input scaling regimes, the function ℓ_P used to state information-estimation expressions for the Poisson model, appears to describe the behavior of the negative binomial model. The expressions found, which only depend on input statistics strengthen the importance of the function $\mathsf{E}[X \log \frac{X}{\mathsf{E}[X]}]$ in the context of the mutual information and $\ell_P(\mathsf{E}[X], \mathsf{E}_Q[X])$ in the context of the relative entropy. We note that meanwhile in the relative entropy side, the behavior is basically given by the first moment statistics of the input, in the case of the mutual information the behavior of the model depends also on the statistics $\mathsf{E}[X \log X]$.

5.4 Proofs

In this section, we develop the proofs for the negative binomial model. We first present a proof of (5.8) in Theorem 16 and (5.10) in Theorem 17. Subsequently we present proofs for Theorem 19 and 20 which let us conclude statements proposed in (5.6) in Theorem 16 and (5.9) in Theorem 17. We close this section with the proof to results related with the low input scaling regime of the negative binomial model.

5.4.1 Proof of (5.8) in Theorem 16

First we state a Lemma that let us calculate the derivative of the marginal P_Y^r with respect to the parameter θ.

Lemma 7. *[Exchangeability Property]. Let the function $X_\theta = f(\theta, X) : \Theta \times \mathcal{X} \to \mathbb{R}^+$ be such that X_θ and X'_θ are bounded, with $|X'_b| < M$ for some $M \in \mathbb{R}$. Then,*

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathsf{E}\left[P^r_{Y|X}(y|X)\right] = \mathsf{E}\left[\left(\frac{\mathrm{d}}{\mathrm{d}\theta}P^r_{Y|X}(y|X)\right)\right] \tag{5.26}$$

$$= yP^r_Y(y)\mathsf{E}^n\left[\frac{X'_\theta}{X_\theta}\middle| Y = y\right] - (y+r)P^r_Y(y)\mathsf{E}^n\left[\frac{X'_\theta}{1+X_\theta}\middle| Y = y\right], \tag{5.27}$$

where $P^r_{Y|X}(y|x)$ is given by (5.2).

Proof. By [6, Theorem 12.13] the derivative and the integral operators can be exchanged in order if the following conditions hold:

(i) The derivative $\frac{\mathrm{d}}{\mathrm{d}\theta}P^r_{Y|X}(y|x)$ is well defined for all $(\theta, x) \in \Theta \times \mathcal{X}$.

(ii) There exists a function $\omega(x)$ such that for all $(\theta, x) \in \Theta \times \mathcal{X}$,

$$\left|\frac{\mathrm{d}}{\mathrm{d}\theta}P^r_{Y|X}(y|x)\right| < \omega(x), \tag{5.28}$$

and $\mathsf{E}[\omega(X)] < \infty$.

For *(i)*, the derivative of the conditional $P^r_{Y|X}$ is given by

$$\frac{\mathrm{d}}{\mathrm{d}\theta}P^r_{Y|X}(y|x) = \underbrace{P^r_{Y|X}(y|x)\frac{\partial f(\theta,x)}{\partial\theta}\frac{y}{f(\theta,x)}}_{(\mathbb{I})} - \underbrace{P^r_{Y|X}(y|x)\frac{\partial f(\theta,x)}{\partial\theta}\frac{(y+r)}{f(\theta,x)+1}}_{(\mathbb{III})}. \tag{5.29}$$

Therefore, due to the fact that $f(\theta, x) > 0$ and $|\frac{\partial f(\theta,x)}{\partial\theta}| < M$ for all $(\theta, x) \in \Theta \times \mathcal{X}$, we conclude that (5.29) is well defined over the domain $\Theta \times \mathcal{X}$.

For *(ii)*, we proceed to prove the existence of integrable upper and lower bounds for (𝕀) and (𝕀𝕀𝕀) in (5.29). Notice that the term (𝕀) is zero for $y = 0$. For $y \neq 0$ we have,

$$\begin{aligned}
&\left| P^r_{Y|X}(y|x) \frac{\partial f(\theta, x)}{\partial \theta} \frac{y}{f(\theta, x)} \right| \\
&\quad = \frac{r(r+y-1)!}{(y-1)!r!} \left(\frac{f(\theta, x)}{f(\theta, x)+1} \right)^{y-1} \left(\frac{1}{f(\theta, x)+1} \right)^{r+1} \left| \frac{\partial f(\theta, x)}{\partial \theta} \right| \qquad (5.30) \\
&\quad = r P^{r+1}_{Y|X}(y-1|x) \left| \frac{\partial f(\theta, x)}{\partial \theta} \right| < Mr, \qquad (5.31)
\end{aligned}$$

where $P^{r+1}_{Y|X}$ represents a negative binomial distribution with parameters $(r+1, f(\theta, x)/(f(\theta, x)+1))$.

We can apply a similar procedure to (𝕀𝕀𝕀), which leads to,

$$\begin{aligned}
\left| P^r_{Y|X}(y|x) \frac{\partial f(\theta, x)}{\partial \theta} \frac{(y+r)}{f(\theta, x)+1} \right| &= \frac{r(y+r)!}{y!r!} \frac{(f(\theta, x))^y}{(f(\theta, x)+1)^{y+r+1}} \left| \frac{\partial f(\theta, x)}{\partial \theta} \right| \qquad (5.32) \\
&= r P^{r+1}_{Y|X}(y|x) \left| \frac{\partial f(\theta, x)}{\partial \theta} \right| < Mr. \qquad (5.33)
\end{aligned}$$

Combining (5.29), (5.31) and (5.33) yields Condition *(ii)*. □

Proof of (5.8) hinges on the following result that reduces the derivative of the output pmf to a certain difference.

Lemma 8. *Let P^r_Y be the pmf of the output of the negative binomial model described by (5.2), with $X_\theta = \theta X$, where the input X is always positive and follows an arbitrary distribution P_X. For every $y \in \mathbb{Z}^+$,*

$$\frac{\mathrm{d}}{\mathrm{d}\theta} P^r_Y(y) = \frac{1}{\theta} \left(y P^r_Y(y) - (y+1) P^r_Y(y+1) \right). \qquad (5.34)$$

Proof. We start with

$$P^r_Y(y) = \mathsf{E}\left[\binom{y+r-1}{y} \left(\frac{1}{\theta X + 1} \right)^r \left(\frac{\theta X}{\theta X + 1} \right)^y \right]. \qquad (5.35)$$

Evidently,

$$\frac{\mathrm{d}}{\mathrm{d}\theta}P_Y^r(y) = \mathsf{E}\left[\binom{y+r-1}{y}\frac{\mathrm{d}}{\mathrm{d}\theta}\left(\left(\frac{1}{\theta X+1}\right)^r\left(\frac{\theta X}{\theta X+1}\right)^y\right)\right] \tag{5.36}$$

$$= \mathsf{E}\left[\binom{y+r-1}{y}\left(\frac{1}{\theta X+1}\right)^r\left(\frac{\theta X}{\theta X+1}\right)^y\left(\frac{y}{\theta}-\frac{(y+r)X}{\theta X+1}\right)\right] \tag{5.37}$$

$$= \frac{y}{\theta}P_Y^r(y) - \frac{y+1}{\theta}\mathsf{E}\left[\binom{y+r}{y+1}\left(\frac{1}{\theta X+1}\right)^r\left(\frac{\theta X}{\theta X+1}\right)^{y+1}\right] \tag{5.38}$$

$$= \frac{1}{\theta}\left(yP_Y^r(y) - (y+1)P_Y^r(y+1)\right), \tag{5.39}$$

where in (5.36) we have used the exchangeability property shown in Lemma 7. Since (5.34) holds for any input distribution P_X, it remains true if P_X is replaced by another distribution Q_X, as long as the input is always nonnegative. It is interesting to see that (5.34) is literally identical to (4.33). □

Proof of (5.8) in Theorem 16, based on Lemma 8, resembles that of (4.9) in Theorem 8. Using similar techniques to those shown in Section 4.4.1 we arrive to,

$$\frac{\mathrm{d}}{\mathrm{d}\theta}D(P_Y^r\|Q_Y^r) = \frac{1}{\theta}\sum_{y=1}^{\infty} yP_Y^r(y)\left(T(y)-\log T(y)-1\right), \tag{5.40}$$

where

$$T(y) = \frac{P_Y^r(y-1)Q_Y^r(y)}{P_Y^r(y)Q_Y^r(y-1)}. \tag{5.41}$$

Moreover,

$$P_Y^r(y-1) = \mathsf{E}\left[\binom{y+r-2}{y-1}\left(\frac{1}{\theta X+1}\right)^r\left(\frac{\theta X}{\theta X+1}\right)^{y-1}\right] \tag{5.42}$$

$$= \mathsf{E}\left[\frac{y}{y+r-1}\binom{y+r-1}{y}\left(\frac{1}{\theta X+1}\right)^r\left(\frac{\theta X}{\theta X+1}\right)^{y-1}\right] \tag{5.43}$$

$$= \frac{y}{y+r-1}\mathsf{E}^r\left[1+\frac{1}{\theta X}\middle|Y=y\right]P_Y^r(y). \tag{5.44}$$

Similarly,

$$Q_Y^r(y-1) = \frac{y}{y+r-1}\mathsf{E}_Q^r\left[1+\frac{1}{\theta X}\middle|Y=y\right]Q_Y^r(y). \tag{5.45}$$

Therefore,

$$T(y) = \frac{\mathsf{E}^r\left[X^{-1} + \theta | Y = y\right]}{\mathsf{E}_Q^r\left[X^{-1} + \theta | Y = y\right]}. \tag{5.46}$$

Equation (5.8) is thus established using the definition of ℓ_I, given in Table 3.1, Chapter 3, jointly with (5.40) and (5.46).

5.4.2 Proof of (5.10) in Theorem 17

Fix $x \in \mathcal{X}$ and let $Q_Y^r = P_{Y|X=x}^r$. Taking into account the expression obtained for the derivative of the relative entropy we have,

$$\frac{\mathrm{d}}{\mathrm{d}\theta} D(Q_Y^r \| P_Y^r) = \frac{\mathrm{d}}{\mathrm{d}\theta} D(P_{Y|X=x}^r \| P_Y^r) \tag{5.47}$$

$$= \mathsf{E}_{P_{Y|X=x}^r}\left[\frac{Y}{\theta}\ell_I\left(x^{-1} + \theta, \mathsf{E}^r[X^{-1}|Y] + \theta\right)\right]. \tag{5.48}$$

Therefore,

$$\frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) = \int \frac{\mathrm{d}}{\mathrm{d}\theta} D(P_{Y|X=x}^r \| P_Y^r) \mathrm{d}P_X(x) \tag{5.49}$$

$$= \int \mathsf{E}_{P_{Y|X=x}^r}\left[\frac{Y}{\theta}\ell_I\left(x^{-1} + \theta, \mathsf{E}^r[X^{-1}|Y] + \theta\right)\right] \mathrm{d}P_X(x) \tag{5.50}$$

$$= \mathsf{E}^r\left[\frac{Y}{\theta}\ell_I\left(X^{-1} + \theta, \mathsf{E}^r[X^{-1}|Y] + \theta\right)\right]. \tag{5.51}$$

5.4.3 Proof of Theorem 19

Next, we provide a set of lemmas to express certain quantities over the negative binomial model of order r in terms of conditional estimates over negative binomial models of order $r+1$. We use $X_\theta = f(\theta, X)$ and $X'_\theta = \partial f(\theta, X)/\partial\theta$.

Lemma 9.

$$\mathsf{E}^r\left[Y\ \mathsf{E}^r\left[\frac{X'_\theta}{X_\theta}\middle| Y\right] \log \frac{P_Y^r(Y)}{Q_Y^r(Y)}\right]$$

$$= r\mathsf{E}^{r+1}\left[\mathsf{E}^{r+1}[X'_\theta|Y] \log \frac{\mathsf{E}^{r+1}[X_\theta|Y] P_Y^{r+1}(Y)}{\mathsf{E}_Q^{r+1}[X_\theta|Y] Q_Y^{r+1}(Y)}\right]. \tag{5.52}$$

Proof. We shall use two facts in this proof. First, for every positive integer y,

$$
\begin{aligned}
& y\mathsf{E}^r\left[\frac{X'_\theta}{X_\theta}\middle| Y=y\right] P_Y^r(y) \\
&= r\mathsf{E}\left[\frac{\partial f(\theta,X)}{\partial\theta}\frac{(r+y-1)!}{(y-1)!r!}\left(\frac{f(\theta,X)}{1+f(\theta,X)}\right)^{y-1}\left(\frac{1}{1+f(\theta,X)}\right)^{r+1}\right] \qquad (5.53)\\
&= r\mathsf{E}\left[\frac{\partial f(\theta,X)}{\partial\theta}P_{Y|X}^{r+1}(y-1|X)\right]. \qquad (5.54)
\end{aligned}
$$

Second, for an arbitrary function $\theta(y)$,

$$
\begin{aligned}
& \sum_{y=1}^{\infty}\theta(y)\log\frac{\mathsf{E}\left[P_{Y|X}^r(y|X)\right]}{\mathsf{E}_Q\left[P_{Y|X}^r(y|X)\right]} \\
&= \sum_{y=0}^{\infty}\theta(y+1)\log\frac{\mathsf{E}\left[P_{Y|X}^r(y+1|X)\right]}{\mathsf{E}_Q\left[P_{Y|X}^r(y+1|X)\right]} \qquad (5.55)\\
&= \sum_{y=0}^{\infty}\theta(y+1)\log\frac{\mathsf{E}\left[\left(\frac{f(\theta,X)}{1+f(\theta,X)}\right)^{y+1}\left(\frac{1}{1+f(\theta,X)}\right)^{r}\right]}{\mathsf{E}_Q\left[\left(\frac{f(\theta,X)}{1+f(\theta,X)}\right)^{y+1}\left(\frac{1}{1+f(\theta,X)}\right)^{r}\right]} \qquad (5.56)\\
&= \sum_{y=0}^{\infty}\theta(y+1)\log\frac{\mathsf{E}\left[f(\theta,X)P_{Y|X}^{r+1}(y|X)\right]}{\mathsf{E}_Q\left[f(\theta,X)P_{Y|X}^{r+1}(y|X)\right]}. \qquad (5.57)
\end{aligned}
$$

In order to prove Lemma 9, we manipulate the LHS of (5.52) as follows,

$$\sum_{y=0}^{\infty} y\, \mathsf{E}^r \left[\frac{X'_\theta}{X_\theta} \middle| Y = y \right] \log \frac{P_Y^r(y)}{Q_Y^r(y)} P_Y^r(y)$$

$$= r \sum_{y=1}^{\infty} \mathsf{E} \left[\frac{\partial f(\theta, X)}{\partial \theta} P_{Y|X}^{r+1}(y-1|X) \right] \log \frac{\mathsf{E}\left[P_{Y|X}^r(y|X) \right]}{\mathsf{E}_Q\left[P_{Y|X}^r(y|X) \right]} \tag{5.58}$$

$$= r \sum_{y=0}^{\infty} \mathsf{E} \left[\frac{\partial f(\theta, X)}{\partial \theta} P_{Y|X}^{r+1}(y|X) \right] \log \frac{\mathsf{E}\left[f(\theta, X) P_{Y|X}^{r+1}(y|X) \right]}{\mathsf{E}_Q\left[f(\theta, X) P_{Y|X}^{r+1}(y|X) \right]} \tag{5.59}$$

$$= r \sum_{y=0}^{\infty} \mathsf{E}^{r+1}[X'_\theta | Y = y] \log \frac{\mathsf{E}^{r+1}[X_\theta | Y = y] P_Y^{r+1}(y)}{\mathsf{E}_Q^{r+1}[X_\theta | Y = y] Q_Y^{r+1}(y)} P_Y^{r+1}(y), \tag{5.60}$$

where (5.58) is consequence of (5.54) jointly with the fact that the first term of the sum ($y = 0$) disappears; in (5.59) we changed y by $(y + 1)$ and apply (5.57); and (5.60) is the result of applying, in (5.59), the definition of conditional estimate over a negative binomial model and replacing $f(\theta, X)$ by X_θ and $f'(\theta, X)$ by X'_θ. □

Lemma 10.

$$\mathsf{E}^r \left[(Y + r)\, \mathsf{E}^r \left[\frac{X'_\theta}{1 + X_\theta} \middle| Y \right] \log \frac{P_Y^r(Y)}{Q_Y^r(Y)} \right]$$
$$= r \mathsf{E}^{r+1} \left[\mathsf{E}^{r+1}[X'_\theta | Y] \log \frac{\mathsf{E}^{r+1}[1 + X_\theta | Y] P_Y^{r+1}(Y)}{\mathsf{E}_Q^{r+1}[1 + X_\theta | Y] Q_Y^{r+1}(Y)} \right].$$

Proof. We use the following two facts: First,

$$(y + r) \mathsf{E}^r \left[\frac{X'_\theta}{1 + X_\theta} \middle| Y = y \right] P_Y^r(y)$$
$$= r \mathsf{E} \left[\frac{\partial f(\theta, X)}{\partial \theta} \frac{(r + y)!}{y!\, r!} \frac{(f(\theta, X))^y}{(1 + f(\theta, X))^{y+r+1}} \right] \tag{5.61}$$
$$= r \mathsf{E} \left[\frac{\partial f(\theta, X)}{\partial \theta} P_{Y|X}^{r+1}(y|X) \right]. \tag{5.62}$$

Second,

$$\log \frac{\mathsf{E}\left[P_{Y|X}^r(y|X)\right]}{\mathsf{E}_Q\left[P_{Y|X}^r(y|X)\right]} = \log \frac{\mathsf{E}\left[(1+f(\theta,X))\left(\frac{f(\theta,X)}{1+f(\theta,X)}\right)^y\left(\frac{1}{1+f(\theta,X)}\right)^{r+1}\right]}{\mathsf{E}_Q\left[(1+f(\theta,X))\left(\frac{f(\theta,X)}{1+f(\theta,X)}\right)^y\left(\frac{1}{1+f(\theta,X)}\right)^{r+1}\right]} \tag{5.63}$$

$$= \log \frac{\mathsf{E}\left[(1+f(\theta,X))P_{Y|X}^{r+1}(y|X)\right]}{\mathsf{E}_Q\left[(1+f(\theta,X))P_{Y|X}^{r+1}(y|X)\right]}. \tag{5.64}$$

Therefore, taking the LHS of (5.61) we have,

$$\sum_{y=0}^{\infty}(y+r)\,\mathsf{E}^r\left[\frac{X'_\theta}{1+X_\theta}\middle|Y=y\right]\log\frac{P_Y^r(y)}{Q_Y^r(y)}P_Y^r(y)$$

$$= r\sum_{y=0}^{\infty}\mathsf{E}\left[\frac{\partial f(\theta,X)}{\partial\theta}P_{Y|X}^{r+1}(y|X)\right]\log\frac{\mathsf{E}\left[(1+f(\theta,X))P_{Y|X}^{r+1}(y|X)\right]}{\mathsf{E}_Q\left[(1+f(\theta,X))P_{Y|X}^{r+1}(y|X)\right]} \tag{5.65}$$

$$= r\sum_{y=0}^{\infty}\mathsf{E}^{r+1}[X'_\theta|Y=y]\log\frac{\mathsf{E}^{r+1}[1+X_\theta|Y=y]P_Y^{r+1}(y)}{\mathsf{E}_Q^{r+1}[1+X_\theta|Y=y]Q_Y^{r+1}(y)}P_Y^{r+1}(y), \tag{5.66}$$

where (5.65) is consequence of (5.62) and (5.64). □

Lemma 11.

$$\mathsf{E}^r\left[Y\,\mathsf{E}_Q^r\left[\frac{X'_\theta}{X_\theta}\middle|Y\right]\right] = r\mathsf{E}^{r+1}\left[\mathsf{E}_Q^{r+1}[X'_\theta|Y]\frac{\mathsf{E}^{r+1}[X_\theta|Y]}{\mathsf{E}_Q^{r+1}[X_\theta|Y]}\right]. \tag{5.67}$$

Lemma 12.

$$\mathsf{E}^r\left[(Y+r)\,\mathsf{E}_Q^r\left[\frac{X'_\theta}{X_\theta+1}\middle|Y\right]\right] = r\mathsf{E}^{r+1}\left[\mathsf{E}_Q^{r+1}[X'_\theta|Y]\frac{\mathsf{E}^{r+1}[X_\theta+1|Y]}{\mathsf{E}_Q^{r+1}[X_\theta+1|Y]}\right]. \tag{5.68}$$

We omit the proof of Lemmas 11 and 12 as the results can be obtained using the same steps used to prove Lemmas 9 and 10, respectively.

Proof of (5.13) *in Theorem 19.* The derivative of the relative entropy with respect to θ for the negative binomial model is given by,

$$\frac{\mathrm{d}}{\mathrm{d}\theta}D(P_Y^r\|Q_Y^r) = \underbrace{\sum_{y=0}^{\infty}\frac{\mathrm{d}P_Y^r(y)}{\mathrm{d}\theta}\log\frac{P_Y^r(y)}{Q_Y^r(y)}}_{(\mathbb{I})} - \underbrace{\sum_{y=0}^{\infty}\frac{P_Y^r(y)}{Q_Y^r(y)}\frac{\mathrm{d}Q_Y^r(y)}{\mathrm{d}\theta}}_{(\mathbb{III})}, \tag{5.69}$$

where the derivative penetrates the sum by the Lebesgue convergence theorem [47]. Calculation of the first term ($\mathbb{I}$) is obtained as follows,

$$\begin{aligned}
&\sum_{y=0}^{\infty} \frac{\mathrm{d}P_Y^r(y)}{\mathrm{d}\theta} \log \frac{P_Y^r(y)}{Q_Y^r(y)} \\
&= \sum_{y=0}^{\infty} y \, \mathsf{E}^r \left[\frac{X'_\theta}{X_\theta} \middle| Y = y \right] P_Y^r(y) \log \frac{P_Y^r(y)}{Q_Y^r(y)} \\
&\quad - \sum_{y=0}^{\infty} (y+r) \mathsf{E}^r \left[\frac{X'_\theta}{X_\theta + 1} \middle| Y = y \right] P_Y^r(y) \log \frac{P_Y^r(y)}{Q_Y^r(y)} \qquad (5.70) \\
&= r \sum_{y=0}^{\infty} \mathsf{E}^{r+1}[X'_\theta | Y = y] \log \frac{\mathsf{E}^{r+1}[X_\theta | Y = y](1 + \mathsf{E}_Q^{r+1}[X_\theta | Y = y])}{\mathsf{E}_Q^{r+1}[X_\theta | Y = y]\,(1 + \mathsf{E}^{r+1}[X_\theta | Y = y])} P_Y^{r+1}(y), \qquad (5.71)
\end{aligned}$$

where (5.70) is a consequence of the exchangeability property illustrated in Lemma 7, (5.71) is the expression obtained in (5.70) written in terms of the conditional mean estimation over a negative binomial model with $r+1$ failures, which is proven in Lemmas 9 and 10.

The second term of (5.69), ($\mathbb{III}$), simplifies to,

$$\begin{aligned}
&\sum_{y=0}^{\infty} \frac{P_Y^r(y)}{Q_Y^r(y)} \frac{\mathrm{d}Q_Y^r(y)}{\mathrm{d}\theta} \\
&= \sum_{y=0}^{\infty} y \, \mathsf{E}_Q^r \left[\frac{X'_\theta}{X_\theta} \middle| Y = y \right] P_Y^r(y) - \sum_{y=0}^{\infty} (y+r) \, \mathsf{E}_Q^r \left[\frac{X'_\theta}{X_\theta + 1} \middle| Y = y \right] P_Y^r(y) \qquad (5.72) \\
&= r \sum_{y=0}^{\infty} P_Y^{r+1}(y) \mathsf{E}_Q^{r+1}[X'_\theta | Y = y] \\
&\quad \times \left(\frac{\mathsf{E}^{r+1}[X_\theta | Y = y]}{\mathsf{E}_Q^{r+1}[X_\theta | Y = y]} - \frac{\left(1 + \mathsf{E}^{r+1}[X_\theta | Y = y]\right)}{\left(1 + \mathsf{E}_Q^{r+1}[X_\theta | Y = y]\right)} \right), \qquad (5.73)
\end{aligned}$$

where (5.72) appears as consequence of the exchangeability property (Lemma 7) and (5.73) is a result obtained of applying Lemmas 11 and 12 to the expression in (5.72).

Plugging (5.71) and (5.73) in (5.69) establishes (5.13). □

5.4.4 Proof of Theorem 20

This proof is similar to the proof for Theorem 12 in Section 4.4.4. Hence,

$$\frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) = \int \frac{\mathrm{d}}{\mathrm{d}\theta} D(P^r_{Y|X=x} \| P^r_Y) \mathrm{d}P_X(x) \tag{5.74}$$

$$= \int \left(\mathsf{G}^{r+1}_{\delta(x),P}(f(\theta,X) - \mathsf{G}^{r+1}_{\delta(x),\delta(x)}(f(\theta,X)) \right) \mathrm{d}P_X(x) \tag{5.75}$$

$$= \mathsf{G}^{r+1}_{P}(X_\theta) \tag{5.76}$$

$$= \mathsf{E}^{r+1} \left[X'_\theta \log \frac{(1+\mathsf{E}^{r+1}[X_\theta|Y])X_\theta}{(1+X_\theta)\mathsf{E}^{r+1}[X_\theta|Y]} \right], \tag{5.77}$$

where $\mathsf{G}^{r+1}_{\delta(x),P}(f(\theta,X))$ is given by[2],

$$\mathsf{G}^{r+1}_{\delta(x),P}(f(\theta,X)) = r\mathsf{E}_{P^{r+1}_{Y|X=x}} \Bigg[\frac{\partial f(\theta,x)}{\partial \theta} \log \frac{(1+\mathsf{E}^{r+1}[f(\theta,X)|Y])f(\theta,x)}{(1+f(\theta,x))\mathsf{E}^{r+1}[f(\theta,X)|Y]} - \frac{\mathsf{E}^{r+1}\left[\frac{\partial f(\theta,X)}{\partial \theta}\middle|Y\right](f(\theta,x) - \mathsf{E}^{r+1}[f(\theta,X)|Y])}{(1+\mathsf{E}^{r+1}[f(\theta,X)|Y])\mathsf{E}^{r+1}[f(\theta,X)|Y]} \Bigg]. \tag{5.78}$$

In this case, we note that (5.75) is consequence of (5.13) in Theorem 19 and (5.76) is due to the definition given in (5.14).

[2]In (5.78), the subscript P in $\mathsf{G}^{r+1}_{\delta(x),P}(X_\theta)$ makes reference to the fact that the conditionals $\mathsf{E}^{r+1}[X_\theta|Y]$ at the RHS of the equation are with respect to the distribution induced by $P_X \times P^{r+1}_{Y|X}$.

5.4.5 Proof of (5.6) in Theorem 16

Replacing in (5.13) X_θ by θX,

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\theta} D(P_Y^r \| Q_Y^r) \\
&= \mathsf{G}_Q^{r+1}(\theta X) - \mathsf{G}_P^{r+1}(\theta X) && (5.79)\\
&= r\mathsf{E}^{r+1}\left[X \log \frac{1+\mathsf{E}_Q^{r+1}[\theta X|Y]}{\mathsf{E}_Q^{r+1}[\theta X|Y]} - \frac{(X-\mathsf{E}_Q^{r+1}[X|Y])}{(1+\mathsf{E}_Q^{r+1}[\theta X|Y])}\right] \\
&\quad - r\mathsf{E}^{r+1}\left[X \log \frac{1+\mathsf{E}^{r+1}[\theta X|Y]}{\mathsf{E}^{r+1}[\theta X|Y]}\right] && (5.80)\\
&= \frac{r}{\theta}\mathsf{E}^{r+1}\left[\theta X \log \frac{(1+\mathsf{E}_Q^{r+1}[\theta X|Y])\mathsf{E}^{r+1}[\theta X|Y]}{\mathsf{E}_Q^{r+1}[\theta X|Y](1+\mathsf{E}^{r+1}[\theta X|Y])}\right] \\
&\quad - \frac{r}{\theta}\mathsf{E}^{r+1}\left[\frac{(\mathsf{E}^{r+1}[\theta X|Y]-\mathsf{E}_Q^{r+1}[\theta X|Y])}{1+\mathsf{E}_Q^{r+1}[\theta X|Y]}\right] && (5.81)\\
&= \frac{r}{\theta}\mathsf{E}^{r+1}\left[\ell_{nb}(\mathsf{E}^{r+1}[\theta X|Y], \mathsf{E}_Q^{r+1}[\theta X|Y])\right]. && (5.82)
\end{aligned}$$

5.4.6 Proof of (5.9) in Theorem 17

Replacing in (5.15) X_θ by θX,

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) \\
&= \mathsf{G}_P^{r+1}(\theta X) && (5.83)\\
&= r\mathsf{E}^{r+1}\left[X \log \frac{\theta X(1+\mathsf{E}^{r+1}[\theta X|Y])}{(1+\theta X)\mathsf{E}^{r+1}[\theta X|Y]} - \frac{(X-\mathsf{E}^{r+1}[X|Y])}{(1+\mathsf{E}^{r+1}[\theta X|Y])}\right] && (5.84)\\
&= \frac{r}{\theta}\mathsf{E}^{r+1}\left[\ell_{nb}(\theta X, \mathsf{E}^{r+1}[\theta X|Y])\right]. && (5.85)
\end{aligned}$$

5.4.7 Proof of Theorem 21

Let Y be the output of a negative binomial model $\left(r, \frac{\theta X}{r+\theta X}\right)$ and $a, \hat{a} \in \mathbb{R}^+$. Just for simplicity in the notation we denote as d_{nb} the Bregman divergence used in (3.55) to build the exponential representation form of the negative binomial distribution, *i.e.*,

$$d_{nb}(a,\hat{a}) = a \log \frac{a(r+\hat{a})}{(r+a)\hat{a}} + r \log \frac{r+\hat{a}}{r+a}. \tag{5.86}$$

Then,

$$d_{nb}(a,\hat{a}) = r d_{nb^*}\left(\frac{a}{r},\frac{\hat{a}}{r}\right) \tag{5.87}$$

where d_{nb^*} is the Bregman divergence build upon the convex function

$$nb^*(a) = a\log\frac{a}{1+a} + \log\frac{1}{1+a}. \tag{5.88}$$

Using the linearity of the Bregman divergences we get that,

$$d_{nb^*}\left(\frac{a}{r},\frac{\hat{a}}{r}\right) = \ell_{nb}\left(\frac{a}{r},\frac{\hat{a}}{r}\right) + d_{\overline{\phi}}\left(\frac{a}{r},\frac{\hat{a}}{r}\right) \tag{5.89}$$

$$> \ell_{nb}\left(\frac{a}{r},\frac{\hat{a}}{r}\right) \tag{5.90}$$

where $d_{\overline{\phi}}$ is the Bregman divergence build using the function

$$\overline{\phi}(a) = -\log(1+a). \tag{5.91}$$

Therefore, for a given $Y = y \in \mathbb{Z}_0^+$, with $a = \theta X$ and $\hat{a}(y) = \mathsf{E}^{r+1}[\theta X|Y=y]$ we get,

$$\frac{1}{\theta}\mathsf{E}^{r+1}\left[d_{nb}\left(\theta X, \mathsf{E}^{r+1}[\theta X|Y]\right)\right] > \frac{r}{\theta}\mathsf{E}^{r+1}\left[\ell_{nb}\left(\frac{\theta X}{r}, \mathsf{E}^{r+1}\left[\frac{\theta X}{r}\middle|Y\right]\right)\right] \tag{5.92}$$

$$= \frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y), \tag{5.93}$$

where to obtain (5.93) we use the expression found in Theorems 20 with $X_\theta = \frac{\theta X}{r}$. The upper bound for the derivative of the relative entropy can be easily found once we change θX by $\mathsf{E}^{r+1}[X|Y]$ and $\mathsf{E}^{r+1}[X|Y]$ by $\mathsf{E}_Q^{r+1}[\theta X|Y]$ in (5.92).

5.4.8 Proof of Theorem 22

Before proceed formally with the proof of Theorem 22 we state two lemmas.

Lemma 13. *Let Y be the output of a negative binomial model with parameters $(r+1, \theta X/(1+\theta X))$ where X is a positive bounded random variable. Then,*

$$\lim_{\theta\to 0}\mathsf{E}[X P_{Y|X}^{r+1}(y|X)] = \mathsf{E}[X]\mathbb{1}_{\{y=0\}} \tag{5.94}$$

and

$$\lim_{\theta\to 0}\mathsf{E}[P_{Y|X}^{r+1}(y|X)] = \mathbb{1}_{\{y=0\}}. \tag{5.95}$$

Proof. By hypothesis we have that $\mathsf{E}[X] < \infty$. Additionally, notice that,

$$\lim_{\theta \to 0} x P_{Y|X}^{r+1}(y|x) = x P_{Y|X}^{r+1}(y|x)\Big|_{\theta=0} = x \mathbb{1}_{\{y=0\}} \tag{5.96}$$

and

$$\lim_{\theta \to 0} P_{Y|X}^{r+1}(y|x) = P_{Y|X}^{r+1}(y|x)\Big|_{\theta=0} = \mathbb{1}_{\{y=0\}}. \tag{5.97}$$

Therefore, taking into account (5.96) and (5.97) and based on the Dominated Convergence Theorem [47] jointly with [6, Theorem 12.11] we can introduce the limit inside the expectation in (5.94) and (5.95) and state the desired result. □

Lemma 14. *Let Y be the output of a binomial model with parameters $(r+1, \theta X/(1+\theta X))$. Then, for a given $Y = y \in \mathbb{Z}_0^+$,*

$$\begin{aligned} &\lim_{\theta \to 0} P_Y^{r+1}(y) \mathsf{E}^{r+1}\left[X|Y=y\right] \log \frac{\mathsf{E}^{r+1}[X|Y=y]}{1+\theta \mathsf{E}^{r+1}[X|Y=y]} \\ &\quad = P_Y^{r+1}(y) \mathsf{E}^{r+1}\left[X|Y=y\right] \log \frac{\mathsf{E}^{r+1}[X|Y=y]}{1+\theta \mathsf{E}^{r+1}[X|Y=y]}\Bigg|_{\theta=0}. \end{aligned} \tag{5.98}$$

Proof. First notice that,

$$\begin{aligned} &P_Y^{r+1}(y) \mathsf{E}^{r+1}\left[X|Y=y\right] \log \frac{\mathsf{E}^{r+1}[X|Y=y]}{1+\theta \mathsf{E}^{r+1}[X|Y=y]} \\ &= \mathsf{E}[X P_{Y|X}^{r+1}(y|X)] \\ &\quad \times \left(\log \mathsf{E}[X P_{Y|X}^{r+1}(y|X)] - \log \left(\mathsf{E}[P_{Y|X}^{r+1}(y|X)] + \theta \mathsf{E}[X P_{Y|X}^{r+1}(y|X)] \right) \right). \end{aligned} \tag{5.99}$$

Based on the following inequality,

$$\begin{aligned} &0 < \mathsf{E}[X P_{Y|X}^{r+1}(y|X)] \\ &< \binom{r+y}{y} \left(\frac{\theta x_{\max}}{1+\theta x_{\max}} \right)^y \left(\frac{1}{1+\theta x_{\max}} \right)^{r+1} (1+\theta x_{\max})^{r+1} \mathsf{E}[X], \end{aligned} \tag{5.100}$$

we get that,

$$\mathsf{E}[X P_{Y|X}^{r+1}(y|X)]\Big|_{\theta=0} = \mathsf{E}[X] \mathbb{1}_{\{y=0\}}, \tag{5.101}$$

and

$$\mathsf{E}[P_{Y|X}^{r+1}(y|X)]\Big|_{\theta=0} = \mathbb{1}_{\{y=0\}}. \tag{5.102}$$

Expressions given in (5.101) and (5.102) together with (5.99) let us state that,

$$P_Y^{r+1}(y)\mathsf{E}^{r+1}\left[X|Y=y\right]\log\frac{\mathsf{E}^{r+1}[X|Y=y]}{1+\theta\mathsf{E}^{r+1}[X|Y=y]}\Big|_{\theta=0} = \mathsf{E}[X]\log\mathsf{E}[X]\mathbb{1}_{\{y=0\}}, \tag{5.103}$$

where we use the convention $0\log 0 = 0$. On the other hand, the limit in (5.99) is calculated as follows. For $y=0$, based on Lemma 13 we get,

$$\lim_{\theta\to 0} P_Y^{r+1}(y)\mathsf{E}^{r+1}\left[X|Y=y\right]\log\frac{\mathsf{E}^{r+1}[X|Y=y]}{1+\theta\mathsf{E}^{r+1}[X|Y=y]} = \mathsf{E}[X]\log\mathsf{E}[X]\mathbb{1}_{\{y=0\}}. \tag{5.104}$$

For $y\neq 0$ the limit over the first term at the RHS of (5.99), based on Lemma 13 is given by,

$$\lim_{\theta\to 0}\mathsf{E}[XP_{Y|X}^{r+1}(y|X)]\log\mathsf{E}[XP_{Y|X}^{r+1}(y|X)] = 0. \tag{5.105}$$

Applying L'Hospital rule [47] over the second term at the RHS of (5.99) let us get that,

$$\begin{aligned}&\lim_{\theta\to 0}\mathsf{E}[XP_{Y|X}^{r+1}(y|X)]\log\Big(\mathsf{E}[P_{Y|X}^{r+1}(y|X)]+\theta\mathsf{E}[XP_{Y|X}^{r+1}(y|X)]\Big)\\ &\quad= -\lim_{\theta\to 0}\frac{A_y^2(\theta)\frac{\mathrm{d}}{\mathrm{d}\theta}\left\{B_y(\theta)+A_y(\theta)\right\}}{(B_y(\theta)+\theta A_y(\theta))\frac{\mathrm{d}}{\mathrm{d}\theta}A_y(\theta)},\end{aligned} \tag{5.106}$$

where, for simplicity in the notation we define,

$$A_y(\theta)\triangleq\mathsf{E}[XP_{Y|X}^{r+1}(y|X)] = \theta^y\mathsf{E}\left[\binom{r+y}{y}X\left(\frac{X}{1+\theta X}\right)^y\left(\frac{1}{1+\theta X}\right)^{r+1}\right] \tag{5.107}$$

and

$$B_y(\theta)\triangleq\mathsf{E}[P_{Y|X}^{r+1}(y|X)] = \theta^y\mathsf{E}\left[\binom{r+y}{y}\left(\frac{X}{1+\theta X}\right)^y\left(\frac{1}{1+\theta X}\right)^{r+1}\right]. \tag{5.108}$$

Differentiating inside the expectation (see Lemma 7) in (5.107) and (5.108) let us get that,

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}\theta}A_y(\theta)\\&=\theta^{y-1}\mathsf{E}\left[\binom{r+y}{y}\left(\frac{X}{1+\theta X}\right)^y\left(\frac{1}{1+\theta X}\right)^{r+1}\left(\frac{yX}{1+\theta X}-\frac{\theta X^2(r+1)}{1+\theta X}\right)\right]\end{aligned} \tag{5.109}$$

and

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}\theta}B_y(\theta)\\&=\theta^{y-1}\mathsf{E}\left[\binom{r+y}{y}\left(\frac{X}{1+\theta X}\right)^y\left(\frac{1}{1+\theta X}\right)^{r+1}\left(\frac{y}{1+\theta X}-\frac{\theta X(r+1)}{1+\theta X}\right)\right].\end{aligned} \tag{5.110}$$

Expressions given in (5.107) and (5.108) imply that,

$$\begin{aligned}&\lim_{\theta\to 0}\frac{A_y(\theta)}{B_y(\theta)+\theta A_y(\theta)}\\&=\lim_{\theta\to 0}\frac{\mathsf{E}\left[X\left(\frac{X}{1+\theta X}\right)^y\left(\frac{1}{1+\theta X}\right)^{r+1}\right]}{\mathsf{E}\left[\left(\frac{X}{1+\theta X}\right)^y\left(\frac{1}{1+\theta X}\right)^{r+1}\right]+\theta\mathsf{E}\left[X\left(\frac{X}{1+\theta X}\right)^y\left(\frac{1}{1+\theta X}\right)^{r+1}\right]} &(5.111)\\&=\frac{\mathsf{E}[X^{y+1}]}{\mathsf{E}[X^y]}, &(5.112)\end{aligned}$$

where in order to obtain (5.112) we use the fact that,[3]

$$\lim_{\theta\to 0}\mathsf{E}\left[X^i\left(\frac{X}{1+\theta X}\right)^y\left(\frac{1}{1+\theta X}\right)^{r+1}\right]=\mathsf{E}[X^{y+i}],\quad \text{for } i=0,1. \tag{5.113}$$

Additionally, from Lemma 13 we get that,

$$\lim_{\theta\to 0}A_y(\theta)=0. \tag{5.114}$$

[3]To obtain (5.113) we use [6, Theorem 12.11] which let us calculate the limit inside the expectation. (Similarly as was done in Lemma 13)

Finally, calculating,

$$\lim_{\theta\to 0}\frac{\frac{\mathrm{d}}{\mathrm{d}\theta}\left(B_y(\theta)+\theta A_y(\theta)\right)}{\frac{\mathrm{d}}{\mathrm{d}\theta}A_y(\theta)}=\lim_{\theta\to 0}\frac{\frac{\mathrm{d}}{\mathrm{d}\theta}B_y(\theta)+A_y(\theta)+\theta\frac{\mathrm{d}}{\mathrm{d}\theta}A_y(\theta)}{\frac{\mathrm{d}}{\mathrm{d}\theta}A_y(\theta)} \tag{5.115}$$

$$=\frac{\mathsf{E}[X^y]}{\mathsf{E}[X^{y+1}]}, \tag{5.116}$$

which is consequence of the fact that,

$$\lim_{\theta\to 0}\mathsf{E}\left[X^i\left(\frac{X}{1+\theta X}\right)^y\left(\frac{1}{1+\theta X}\right)^{r+1}\left(\frac{y}{1+\theta X}-\frac{\theta X(r+1)}{1+\theta X}\right)\right]$$
$$=y\mathsf{E}[X^{y+i}],\quad \text{for } i=0,1. \tag{5.117}$$

Putting together (5.112), (5.114) and (5.116), and replacing them in (5.106) yields,

$$\lim_{\theta\to 0}\mathsf{E}[XP_{Y|X}^{r+1}(y|X)]\log\left(\mathsf{E}[P_{Y|X}^{r+1}(y|X)]+\theta\mathsf{E}[XP_{Y|X}^{r+1}(y|X)]\right)=0, \tag{5.118}$$

for $y\neq 0$. Consequently, taking into account (5.105), (5.118) and (5.99) we have,

$$\lim_{\theta\to 0}P_Y^{r+1}(y)\mathsf{E}^{r+1}[X|Y=y]\log\frac{\mathsf{E}^{r+1}[X|Y=y]}{1+\theta\mathsf{E}^{r+1}[X|Y=y]}=0. \tag{5.119}$$

for $y\neq 0$. Putting together (5.119) and (5.104) and comparing them with (5.103) let us obtain the desired result. □

Proof of Theorem 22. Let Y be the output of a negative binomial model with parameters $(r,\theta X/(1+\theta X))$. Based on the expression given for the derivative of the input output mutual information given in Theorem 17 we

prove that,

$$\lim_{\theta\to 0}\frac{\mathrm{d}}{\mathrm{d}\theta}I(X;Y)$$

$$= \lim_{\theta\to 0}\frac{r}{\theta}\mathsf{E}^{r+1}\left[\ell_{nb}(\theta X, \mathsf{E}^{r+1}[X|Y])\right] \tag{5.120}$$

$$= r\lim_{\theta\to 0}\mathsf{E}^{r+1}\left[X\log\frac{\left(1+\mathsf{E}^{r+1}[\theta X|Y]\right)X}{(1+\theta X)\mathsf{E}^{r+1}[X|Y]}\right] \tag{5.121}$$

$$= \lim_{\theta\to 0} r\left[X\log\frac{X}{1+\theta X}\right]$$

$$- r\lim_{\theta\to 0}\sum_{y=0}^{\infty}P_Y^{r+1}(y)\mathsf{E}^{r+1}\left[X|Y=y\right]\log\frac{\mathsf{E}^{r+1}[X|Y=y]}{1+\theta\mathsf{E}^{r+1}[X|Y=y]} \tag{5.122}$$

$$= r\mathsf{E}\left[X\log X\right] - r\sum_{y=0}^{\infty}\lim_{\theta\to 0}P_Y^{r+1}(y)\mathsf{E}^{r+1}\left[X|Y=y\right]\log\frac{\mathsf{E}^{r+1}[X|Y=y]}{1+\theta\mathsf{E}^{r+1}[X|Y=y]}. \tag{5.123}$$

The first term of (5.123) is justified by the Dominated Convergence Theorem [6, Theorem 8.8]. In the second term, the exchange between the limit and the sum is justified once we prove the following conditions,

(i) The function,

$$P_Y^{r+1}(y)\mathsf{E}^{r+1}\left[X|Y=y\right]\log\frac{\mathsf{E}^{r+1}[X|Y=y]}{1+\theta\mathsf{E}^{r+1}[X|Y=y]} \tag{5.124}$$

is summable over $\mathbb{Z}_0^+$ for each $\theta\in[0,\delta), \delta>0$.

(ii) There exists summable functions $\xi(y)$ and $\omega(y)$ such that

$$\xi(y) \le P_Y^{r+1}(y)\mathsf{E}^{r+1}\left[X|Y=y\right]\log\frac{\mathsf{E}^{r+1}[X|Y=y]}{1+\theta\mathsf{E}^{r+1}[X|Y=y]} \le \omega(y). \tag{5.125}$$

(iii) For a given $Y=y\in\mathbb{Z}_0^+$,

$$\lim_{\theta\to 0}P_Y^{r+1}(y)\mathsf{E}^{r+1}\left[X|Y=y\right]\log\frac{\mathsf{E}^{r+1}[X|Y=y]}{1+\theta\mathsf{E}^{r+1}[X|Y=y]}, \tag{5.126}$$

exists and is equal to,

$$P_Y^{r+1}(y)\mathsf{E}^{r+1}\left[X|Y=y\right]\log\frac{\mathsf{E}^{r+1}[X|Y=y]}{1+\theta\mathsf{E}^{r+1}[X|Y=y]}\bigg|_{\theta=0}. \tag{5.127}$$

In effect, notice that (5.124) can be upper bounded as follows,

$$\begin{aligned}
&\left|\mathsf{E}^{r+1}[X|Y=y]\log\frac{\mathsf{E}^{r+1}[X|Y=y]}{1+\theta\mathsf{E}^{r+1}[X|Y=y]}\right| \\
&\quad\leq \left|\mathsf{E}^{r+1}[X|Y=y]\log\mathsf{E}^{r+1}[X|y=y]\right| \\
&\qquad+\left|\mathsf{E}^{r+1}[X|Y=y]\log 1+\theta\mathsf{E}^{r+1}[X|y=y]\right| &(5.128)\\
&\quad\leq M^*+x_{\max}\log 1+\theta x_{\max}., &(5.129)
\end{aligned}$$

where $M^*=\sup\{x_{\max}\log x_{\max},e^{-1}\}$. To prove Condition *(i)*, it remains to show that (5.124) is summable for $\theta=0$. Effectively, from the analysis given from (5.100) to (5.103) we have that,

$$\begin{aligned}
&P_Y^{r+1}(y)\mathsf{E}^{r+1}\left[X|Y=y\right]\log\frac{\mathsf{E}^{r+1}[X|Y=y]}{1+\theta\mathsf{E}^{r+1}[X|Y=y]}\bigg|_{\theta=0}\\
&=\mathsf{E}[X]\log\mathsf{E}[X]\mathbb{1}_{\{y=0\}}. &(5.130)
\end{aligned}$$

To state Condition *(ii)*, taking into account (5.129) we have that,

$$\begin{aligned}
&\left|\mathsf{E}^{r+1}[X|Y=y]\log\frac{\mathsf{E}^{r+1}[X|Y=y]}{1+\theta\mathsf{E}^{r+1}[X|Y=y]}\right|P_Y^{r+1}(y)\\
&\quad\leq\left(M^*+x_{\max}\log 1+\delta x_{\max}\right)\overline{P}_Y^{r+1}(y)(1+\delta x_{\max})^{r+1} &(5.131)\\
&\quad\triangleq\omega(y) &(5.132)
\end{aligned}$$

where

$$\overline{P}_Y^{r+1}(y)=\binom{r+y}{y}\left(1-\frac{1}{1+\delta x_{\max}}\right)^y\left(\frac{1}{1+\delta x_{\max}}\right)^{r+1}. \tag{5.133}$$

Additionally, we note that,

$$\sum_{y=0}^{\infty}\omega(y)=\left(M^*+x_{\max}\log 1+\delta x_{\max}\right)(1+\delta x_{\max})^{r+1}<\infty. \tag{5.134}$$

Finally, Condition *(iii)* is proven in Lemma 14 which let us state that,

$$\lim_{\theta\to 0}P_Y^{r+1}(y)\mathsf{E}^{r+1}\left[X|Y=y\right]\log\frac{\mathsf{E}^{r+1}[X|Y=y]}{1+\theta\mathsf{E}^{r+1}[X|Y=y]}=\mathsf{E}[X]\log\mathsf{E}[X]\mathbb{1}_{\{y=0\}}. \tag{5.135}$$

Replacing (5.135) in (5.123) yields,

$$\lim_{\theta\to 0}\frac{\mathrm{d}}{\mathrm{d}\theta}I(X;Y)=r\mathsf{E}\left[X\log\frac{X}{\mathsf{E}[X]}\right]. \tag{5.136}$$

□

5.4.9 Proof of Corollary 5

For a negative binomial model with parameters $(r, \theta X/(r+\theta X))$, the derivative of the mutual information with respect to θ, shown in Theorem 20 is given by,

$$\frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) = \mathsf{E}^{r+1}\left[X \log \frac{\left(1+\mathsf{E}^{r+1}\left[\frac{\theta X}{r}\middle|Y\right]\right)X}{\left(1+\frac{\theta X}{r}\right)\mathsf{E}^{r+1}[X|Y]}\right]. \tag{5.137}$$

Taking the limit over θ let us get that,

$$\lim_{\theta\to 0}\frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) = \mathsf{E}\left[X \log \frac{X}{\mathsf{E}[X]}\right]. \tag{5.138}$$

The exchange between the limit and the sum carried out in order to obtain (5.138) can be justified through the same procedure used to state Theorem 22, shown in Section 5.4.8.

5.4.10 Proof of Corollary 6

Let Y be the output of a negative binomial model with parameters $(r, \theta X/(\theta X + r))$. Then, based on Theorem 19, replacing X_θ by $\theta X/r$ in (5.14) and consequently in (5.13) let us get that,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\theta} D(P_Y^r||Q_Y^r) &= \mathsf{G}_Q^{r+1}\left(\frac{\theta X}{r}\right) - \mathsf{G}_P^{r+1}\left(\frac{\theta X}{r}\right)\\ &= \mathsf{E}^{r+1}\left[X\log\frac{\left(1+\mathsf{E}_Q^{r+1}\left[\frac{\theta X}{r}\middle|Y\right]\right)\mathsf{E}^{r+1}[X|Y]}{\left(1+\mathsf{E}^{r+1}\left[\frac{\theta X}{r}|Y\right]\right)\mathsf{E}_Q^{r+1}[X|Y]}\right]\\ &\quad - \mathsf{E}^{r+1}\left[\frac{\left(\mathsf{E}^{r+1}[X|Y]-\mathsf{E}_Q^{r+1}[X|Y]\right)}{\left(1+\mathsf{E}_Q^{r+1}\left[\frac{\theta X}{r}\middle|Y\right]\right)}\right].\end{aligned} \tag{5.139}$$

Making θ tend to zero inside the sum of (5.139) let us get the desired result. Formally this procedure is justified by the procedure carried out in Section 5.4.8 in order to obtain Theorem 22.

Chapter 6

Connection with the Poisson model

In this chapter, we treat asymptotically the behavior of the binomial model and its connection with the Poisson model. Expressions given in Chapter 3 relating the information field with the estimation field, for an arbitrarily input pre-processing X_θ let us study the behavior of the Poisson model as long as the number of trials over the binomial model goes to infinite. The Poisson model has been widely treated in the past by the information theory community due to its application in the area of optic communications; it is often used to model pulse-amplitude modulated optical communication with a direct-detection receiver [49]. The methodology used in this section is as follows. At first instance we work over two special cases of the Poisson model studied previously in [25]; the linear amplifying factor and the presence of an additive dark current. In these two cases, based on a binomial model, we show that under mild conditions in the model, as long as the number of trials n goes to infinite, the input output mutual information and the relative entropy converge to their counterpart results found previously for the Poisson model.

Beyond these results, we characterize for the relative entropy and mutual information the information–estimation relationship that arises over a binomial model with an arbitrary input preprocessing X_θ when the parameter n goes to infinite. Subsequently, departing from a Poisson model where the mean is given by a linear scaling of the input, we show that the low input scaling regime coincides with that found for the binomial and negative binomial models. Furthermore, we show that in general, for a Poisson model with an arbitrary mean X_θ the derivative of the mutual

information and the derivative of the relative entropy are equal to the derivatives calculated asymptotically over a binomial model with parameters $(n, X_\theta/n)$. We finish this section with a set of conclusions derived upon those information–estimation expressions obtained. Additionally we show that there is a relationship between the negative binomial models and the Poisson models that let us obtain for the Poisson model, the same results that were obtained using the binomial model results.

6.1 Poisson model: Definition

The Poisson model is based on the Poisson distribution which can be used to express the probability of a given number of events occurring in a fixed interval of time if these events occur with a known average rate and independently of the time since the last event. Mathematically, the pmf is given by the following expression,

$$P(Y = y) = \frac{\lambda^y}{y!} e^{-\lambda}, \quad y \in \mathbb{Z}_0^+, \tag{6.1}$$

where $\lambda > 0$ is the mean of the distribution. We define the Poisson model as a random transformation that maps an input random variable X to an integer output random variable Y, where conditioned on $X = x$ the output Y is Poisson distributed with mean $f(\theta, x)$. Similarly to the previous models studied, the function $f(\theta, X)$ represents a deterministic $\theta-$preprocessing of the input. Hence, the conditional pmf of the model is given by,

$$P_{Y|X}(y|x) = \frac{f(\theta, x)^y}{y!} e^{-f(\theta,x)}, \quad y \in \mathbb{Z}_0^+. \tag{6.2}$$

Similarly to the previous models, for notational convenience, dependency of Y on θ is implicit. Also we use the shorthand,

$$X_\theta = f(\theta, X). \tag{6.3}$$

Furthermore, it is assumed that $f(\theta, X)$ is differentiable with respect to θ, where we denote its derivative as,

$$X'_\theta = \frac{\partial}{\partial} f(\theta, X). \tag{6.4}$$

Suppose that the input X can be distributed according to either P_X or Q_X. Over a Poisson model with mean X_θ we denote each output distribution as

P_Y and Q_Y where the former assumes that $X \sim P_X$ and the latter that $X \sim Q_X$. When dealing with conditional estimates, $\mathsf{E}[X_\theta|Y]$ stands for the estimate of the function X_θ, based on the observation of the output Y when $X \sim P_X$; similarly, $\mathsf{E}_Q[X_\theta|Y]$ represents the conditional estimate of X_θ when $X \sim Q_X$.

6.2 Information-Estimation expressions based on the Binomial model

Throughout this section we gather all the scenarios over the Poisson model that can be analyzed through the results obtained for the binomial model. Hence, the main objective in this case lies in showing the conditions over the binomial model that asymptotically end up over a Poisson model and, by that way let us translate several information-estimation expressions obtained in the binomial context by taking the limit as the number of trials n goes to infinite. Therefore, a key point in the discussion is the structure used to generate a binomial model that asymptotically converges to a Poisson model.

6.2.1 Linear Scaling

The binomial model converges to a Poisson model with rate θX as long as the parameter n tends to infinite and the value of p is given by the function $\theta X/n$ [53, pag. 140]. As $n \to \infty$, the parameter $p = \theta X/n$ goes to zero while the mean of the conditional distribution of the model θX remains fixed. Hence, for a binomial model $(n-1, \theta X/n)$,

$$\lim_{n\to\infty} \binom{n-1}{y} \left(\frac{\theta x}{n}\right)^y \left(1 - \frac{\theta x}{n}\right)^{n-y-1} = \frac{(\theta x)^y}{y!} e^{-\theta x}, \quad y \in \mathbb{Z}_0^+. \tag{6.5}$$

We now explore how the previous results are related to the Poisson model. First, for the Poisson model, the relation between the mutual information can be also expressed using the Itakuro-Saito divergence and the X^{-1} moment:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) &= -\mathsf{E}\left[\frac{Y}{\theta} \log \frac{X^{-1}}{\mathsf{E}[X^{-1}|Y]}\right] \\ &= \mathsf{E}\left[\frac{Y}{\theta} \ell_I\left(X^{-1}, \mathsf{E}\left[X^{-1}|Y\right]\right)\right]. \end{aligned} \tag{6.6}$$

And similarly for the relative entropy,

$$\frac{\mathrm{d}}{\mathrm{d}\theta} D(P_Y \| Q_Y) = \mathsf{E}\left[\frac{Y}{\theta}\ell_I\left(\mathsf{E}\left[X^{-1}|Y\right], \mathsf{E}_Q\left[X^{-1}|Y\right]\right)\right]. \tag{6.7}$$

These results can be obtained by direct substitution of the results in [3, 25].

We now show that expressions given in Theorems 8 and 9 converge to their Poisson counterpart when we make n tend to infinite over a binomial model with parameters $(n, \theta X/n)$.

Theorem 24. *Let $X \sim P_X$ be a random variable taking its value in $(0, x_{\max})$. Let Y be the output of the n-th order binomial model described by (4.2) with $X_\theta = \theta X/n$. Then,*

$$\lim_{n\to\infty} \frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) = \frac{1}{\theta}\mathsf{E}\left[\ell_P(\theta X, \mathsf{E}[\theta X|Y])\right] \tag{6.8}$$

$$= \mathsf{E}\left[\frac{Y}{\theta}\ell_I\left(X^{-1}, \mathsf{E}\left[X^{-1}|Y\right]\right)\right] \tag{6.9}$$

hold for all finite θ. At the RHS of (6.8) and (6.9), the expectation is with respect to $P_{Y|X}P_X$ where $P_{Y|X}$ is the conditional Poisson distribution with mean θX.

Proof. See Section 6.5.1. □

Theorem 25. *Let X be a random variable taking its values in $(0, x_{\max})$, following distribution P_X or Q_X. Let Y be the output of the n-th order binomial model described by (4.2) with $X_\theta = \theta X/n$. Then,*

$$\lim_{n\to\infty} \frac{\mathrm{d}}{\mathrm{d}\theta} D(P_Y^n \| Q_Y^n) = \mathsf{E}\left[\ell_P(\mathsf{E}[X|Y], \mathsf{E}_Q[X|Y])\right] \tag{6.10}$$

$$= \mathsf{E}\left[\frac{Y}{\theta}\ell_I\left(\mathsf{E}[X^{-1}|Y], \mathsf{E}_Q\left[X^{-1}|Y\right]\right)\right] \tag{6.11}$$

hold for all finite θ. At the RHS of (6.10) and (6.11), the expectations $\mathsf{E}[\cdot]$ and $\mathsf{E}_Q[\cdot]$ are, respectively, with respect to $P_{Y|X}P_X$ and $P_{Y|X}Q_X$ where $P_{Y|X}$ is a Poisson model with rate θX.

Proof. The proof of this Theorem follows the same steps carried out in Section 6.5.1 to develop the proof of Theorem 24. □

6.2.2 Additive Dark current

A second scenario frequently covered in all the research done around the Poisson model, considers in the generation of the model, the presence of an additive factor θ, known as dark current. Following the analysis given in [53, pag. 140], we have that, for a binomial model with parameters $(n-1,(X+\theta)/n)$,

$$\lim_{\boldsymbol{n}\to\infty}\binom{n-1}{y}\left(\frac{x+\theta}{n}\right)^{y}\left(1-\frac{x+\theta}{n}\right)^{n-y-1}=\frac{(x+\theta)^{y}}{y!}e^{-(x+\theta)},\quad y\in\mathbb{Z}_{0}^{+}. \tag{6.12}$$

In this case, considering a Poisson model with mean $X+\theta$, the derivative of the input-output mutual information can be expressed in terms of the Bregman divergence ℓ_P through the function $(X+\theta)^{-1}$:

$$\frac{\mathrm{d}}{\mathrm{d}\theta}I(X;Y)=-\mathsf{E}\left[\ell_I(X+\theta,\mathsf{E}[X+\theta|Y])\right] \tag{6.13}$$

$$=-\mathsf{E}\left[\frac{Y}{\theta}\ell_P\left((X+\theta)^{-1},\mathsf{E}[(X+\theta)^{-1}|Y]\right)\right]. \tag{6.14}$$

Furthermore when dealing with the relative entropy concept, we have the following expression,

$$\frac{\mathrm{d}}{\mathrm{d}\theta}D(P_Y||Q_Y)=-\mathsf{E}\left[\ell_I(\mathsf{E}[X+\theta|Y],\mathsf{E}_Q[X+\theta|Y])\right] \tag{6.15}$$

$$=-\mathsf{E}\left[\frac{Y}{\theta}\ell_P\left(\mathsf{E}[(X+\theta)^{-1}|Y],\mathsf{E}_Q[(X+\theta)^{-1}|Y]\right)\right]. \tag{6.16}$$

In the following Theorems we show that, over a Binomial model with parameters $(n,(X+\theta)/n)$, as long as the number of trials n goes to infinite, the derivative of the input-output mutual information and the derivative of the relative entropy tend to (6.13) and (6.15) respectively.

Theorem 26. *Let $X\sim P_X$ be a random variable taking its values in $(0,x_{\max})$. Let Y be the output of the n-th order binomial model described by (4.2) with $X_\theta=(X+\theta)/n$. Then,*

$$\lim_{n\to\infty}\frac{\mathrm{d}}{\mathrm{d}\theta}I(X;Y)=-\mathsf{E}\left[\ell_I(X+\theta,\mathsf{E}[X+\theta|Y])\right] \tag{6.17}$$

$$=-\mathsf{E}\left[\frac{Y}{\theta}\ell_P((X+\theta)^{-1},\mathsf{E}[(X+\theta)^{-1}|Y])\right]. \tag{6.18}$$

Proof. The proof to this Theorem follows the same steps carried out in Section 6.5.1 simply by changing θX by $X + \theta$. □

Theorem 27. *Let X be a random variable taking its values in $(0, x_{\max})$, following distribution P_X or Q_X. Let Y be the output of the n-th order binomial model described by (4.2) with $X_\theta = (X + \theta)/n$. Then*

$$\lim_{n\to\infty} \frac{\mathrm{d}}{\mathrm{d}\theta} D(P_Y^n || Q_Y^n) = -\mathsf{E}\left[\ell_I(\mathsf{E}[X+\theta|Y], \mathsf{E}_Q[X+\theta|Y])\right] \tag{6.19}$$

$$= -\mathsf{E}\left[\frac{Y}{\theta}\ell_P(\mathsf{E}_Q[(X+\theta)^{-1}|Y], \mathsf{E}[(X+\theta)^{-1}|Y])\right]. \tag{6.20}$$

Proof. The proof to this Theorem is just a combination of the proofs of Theorems 25 and 26. □

6.2.3 General case

In this section we briefly present information-estimation relationships when the parameter n goes to infinite over a binomial model with parameters $(n-1, X_\theta/n)$. Similarly to (6.5) and (6.12), we get that [53, pag. 140],

$$\lim_{n\to\infty} \binom{n-1}{y} \left(\frac{f(\theta,x)}{n}\right)^y \left(1 - \frac{f(\theta,x)}{n}\right)^{n-y-1} = \frac{(f(\theta,x))^y}{y!} e^{-f(\theta,x)}, \quad y \in \mathbb{Z}_0^+. \tag{6.21}$$

Therefore, for a generic binomial model where the Bernoulli trials are governed by an arbitrarily function X_θ/n we get that, as long as the number of trials increases, the binomial model converges to a Poisson model with mean X_θ. Based on this behavior in the following theorems we show the corresponding information-estimation expressions for a binomial model when the parameter n goes to infinite. Later, we show that these expressions coincide with those who describe the mutual information and relative entropy over the Poisson model.

Theorem 28. *Let $X \sim P_X$ be a random variable such that X_θ and X'_θ are measurable bounded functions with $X_\theta > 0$. Let Y be the output of the n-th order binomial model (4.2) with parameters $(n, X_\theta/n)$. Then,*

$$\lim_{n\to\infty} \frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) = \mathsf{E}\left[X'_\theta \log \frac{X_\theta}{\mathsf{E}[X_\theta|Y]}\right]. \tag{6.22}$$

Proof. See Section 6.5.2. □

In this case, we note that the result given in Theorems 28 covers information estimation expressions found in the cases of linear input scaling and additive dark current noise. Notice additionally that the expression found in (6.22) let us conclude that in all the cases where X_θ is of the form $h(\theta)X$ or $X + h(\theta)$, for differentiable measurable functions $h(\theta)$, the derivative of the input output mutual information is proportional to the expectation of a Bregman divergence.

Theorem 29. *Let X be a random variable distributed as either P_X or Q_X such that X_θ and X'_θ are measurable bounded functions with $X_\theta > 0$. Let Y be the output of the n-th order binomial model (4.2) with parameters $(n, X_\theta/n)$. Then,*

$$\lim_{n\to\infty} D(P_Y^n||Q_Y^n)$$
$$= \mathsf{E}\left[\mathsf{E}[X'_\theta|Y]\log\frac{\mathsf{E}[X_\theta|Y]}{\mathsf{E}_Q[X_\theta|Y]} - \frac{\mathsf{E}_Q[X'_\theta|Y]\left(\mathsf{E}[X_\theta|Y] - \mathsf{E}_Q[X_\theta|Y]\right)}{\mathsf{E}_Q[X_\theta|Y]}\right]. \quad (6.23)$$

Proof. See Section 6.5.3. □

Notice that, even thought the conditional $P^n_{Y|X}$ tends to a conditional $P_{Y|X}$ that is Poisson, it is not proven that the derivative of the mutual information and the derivative of the relative entropy found in (6.22) and (6.23) correspond to the respective derivatives over a Poisson model.

6.3 Information-Estimation expressions based on the Poisson model

As has been shown before, the low input scaling regime in some scenarios leads to several optimum conditions in the communication problem. In this section we show that the binomial, negative binomial and Poisson models have the same behavior at low input scaling regimes. Specifically, we study the behavior of several information measures over Poisson models where the mean is determined by a linear scaling of the input and the input scaling goes to zero. As was pointed out for the binomial and negative binomial models in Sections 4.2.3 and 5.2.3 respectively, the Bregman divergence ℓ_P between the input X and the mean $\mathsf{E}[X]$, also seems to play a fundamental role over the Poisson model under these circumstances.

Theorem 30. *Let $X \sim P_X$ be a positive bounded random variable. Let Y be the output of a Poisson model* (6.2) *with mean $X_\theta = \theta X$. Then,*

$$\lim_{\theta \to 0} \frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) = \mathsf{E}\left[X \log \frac{X}{\mathsf{E}[X]}\right]. \tag{6.24}$$

Proof. See Section 6.5.4. □

Previously, Atar *et al.* [3], proved the concavity on θ of the mutual information over a Poisson model with mean θX. This implies that the value of the derivative of the mutual information with respect to θ achieves its maximum value when θ goes to zero. In terms of efficiency, a direct implication of the expression given in Theorem 30 is that the bigger the value of $\mathsf{E}[\ell_P(X, \mathsf{E}[X])]$, the bigger the gain in terms of mutual information for a small increase in θ.

The following theorem describes the relative entropy between two output distributions P_Y and Q_Y when the parameter θ goes to zero in the constitution of the model.

Theorem 31. *Let X be a positive bounded random variable that can be distributed accordingly to P_X or Q_X. Let Y be the output of a Poisson model* (6.2) *with mean θX. Then,*

$$\lim_{\theta \to 0} \frac{\mathrm{d}}{\mathrm{d}\theta} D(P_Y || Q_Y) = \ell_P(\mathsf{E}[X], \mathsf{E}_Q[X]). \tag{6.25}$$

Proof. The proof to this Theorem follows the same steps carried out in the proof of Theorem 30 given in Section 6.5.4. □

A remarkable feature to notice in the behavior of the relative entropy is that its behavior only depends on the mean of each input distribution, meanwhile in the case of the mutual information, its rate of change depends on the mean of the input and also on the input statistics $\mathsf{E}[X \log X]$.

The rest of this section is dedicated to the development of information-estimation expressions based on the assumption that the mean of the Poisson model (6.2) is an arbitrary function X_θ. The main result at the end let us state that the expressions given in Theorems 28 and 29, based on the binomial model, correspond to the derivatives of the mutual information and relative entropy of the Poisson model treated.

Theorem 32. *Let $X \sim P_X$ be a random variable such that X_θ and X'_θ are measurable bounded functions with $X_\theta > 0$. Let Y be the output of the*

Poisson model (6.2) *with mean* X_θ*. Then,*

$$\frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) = \mathsf{E}\left[X'_\theta \log \frac{X_\theta}{\mathsf{E}[X_\theta|Y]}\right]. \tag{6.26}$$

Proof. See Section 6.5.5. □

Analogously, for the derivative of the relative entropy we get the following result.

Theorem 33. *Let* X *be a random variable distributed as either* P_X *or* Q_X *such that* X_θ *and* X'_θ *are measurable bounded functions with* $X_\theta > 0$*. Let* Y *be the output of a Poisson model* (6.2) *with mean* X_θ*. Then,*

$$D(P_Y||Q_Y) = \mathsf{E}\left[\mathsf{E}[X'_\theta|Y] \log \frac{\mathsf{E}[X_\theta|Y]}{\mathsf{E}_Q[X_\theta|Y]} - \frac{\mathsf{E}_Q[X'_\theta|Y]\left(\mathsf{E}[X_\theta|Y] - \mathsf{E}_Q[X_\theta|Y]\right)}{\mathsf{E}_Q[X_\theta|Y]}\right]. \tag{6.27}$$

Proof. The proof to this theorem follows is similar to the proof of Theorem 32 shown in Section 6.5.5. □

6.4 Concluding remarks

Throughout this section we have presented several information estimation expressions for the Poisson model. Initially, based on a set of results obtained for the Binomial model, it is proven that they converge to their counterpart results over the Poisson model when the number of trials of the binomial model tends to infinite. It is remarkable that those binomial models that asymptotically converge to the Poisson model are constituted with n Bernoulli trials with probability of success $p = X_\theta/n$. This leads to a binomial model where the mean $np = X_\theta$ is independent of n.

In the context of the negative binomial model it is worth pointing out the following scenario. Let $P^r_{Y|X}$ be the conditional distribution of a negative binomial model with parameters $\left(r, \frac{X^*_\theta}{1+X^*_\theta}\right)$ with $X^*_\theta = X_\theta/r$, *i.e.*,

$$P^r_{Y|X}(y|x) = \binom{r+y-1}{y} \left(\frac{f(\theta,x)}{r+f(\theta,x)}^{\,y}\right)^y \left(\frac{r}{r+f(\theta,x)}\right)^r, \quad y \in \mathbb{Z}^+_0. \tag{6.28}$$

The random transformation modeled through (6.28) corresponds to a negative binomial model where the probability of each Bernoulli trial is given

by $\frac{X_\theta}{r+X_\theta}$, which ends up in a model with mean equal to X_θ, independent of r. Asymptotically, as the parameter r goes to infinite we have that [53, pag. 287],

$$\lim_{r\to\infty} P^r_{Y|X}(y|x) = \frac{(f(\theta,x))^y}{y!}e^{-f(\theta,x)}, \quad y \in \mathbb{Z}_0^+. \tag{6.29}$$

In words, (6.29) means that, for a negative binomial model with parameters $(r, X_\theta/(r+X_\theta))$ where the mean is given by X_θ, the conditional $P^r_{Y|X}$ converges to a Poisson model of mean X_θ, as long as r goes to infinite. This translates, in the case of the mutual information to expressions as the following. Let Y be the output of a negative binomial model where the conditional is given by (6.28). Then, by Theorem 20 in Section 5.2.2, we get that,

$$\frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) = \mathsf{G}_P^{r+1}\left(\frac{X_\theta}{r}\right) \tag{6.30}$$

$$= \mathsf{E}^{r+1}\left[X'_\theta \log \frac{\left(1+\mathsf{E}^{r+1}\left[\frac{X_\theta}{r}\middle| Y\right]\right)X_\theta}{\left(1+\frac{X_\theta}{r}\right)\mathsf{E}^{r+1}[X_\theta|Y]}\right], \tag{6.31}$$

which asymptotically tends to,

$$\lim_{r\to\infty}\frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) = \mathsf{E}\left[X'_\theta \log \frac{X_\theta}{\mathsf{E}[X_\theta|Y]}\right]. \tag{6.32}$$

Therefore, similarly to the binomial model, those negative binomial models where the mean X_θ is independent of r, asymptotically produce an input-output mutual information that is equal to the input-output mutual information over a Poisson model with mean X_θ.

Finally, in the context of the Poisson model with a mean given by an input scaling, it is shown that the low input scaling regime over the Poisson model is equal to that found for the binomial and negative binomial models in Sections 4.2.3 and 5.2.3, respectively.

6.5 Proofs

6.5.1 Proof of Theorem 24

In this section we prove that, over a binomial model with parameters $(n, \theta X/n)$, the derivative of the mutual information converge to their Poisson counterparts as n tends to infinite. First, we prove some auxiliary lemmas and later we proceed to show the proof of Theorem 24.

Lemma 15. *Consider a Binomial model with parameters* $(n-1, \theta X/n)$ *where* $0 < \frac{\theta X}{n} < 1$. *Then,*

$$\lim_{n\to\infty} \mathsf{E}^{n-1}[X|Y=y] = \mathsf{E}[X|Y=y], \tag{6.33}$$

where at the LHS of (6.33)*, the input output relationship is given by a binomial model with parameters* $(n-1, \theta X/n)$ *and at the RHS of* (6.33) *the input output relationship is given by a Poisson model with rate* θX.

Proof. As $n \to \infty$, the binomial model converges to the Poisson model,

$$\lim_{n\to\infty} P_{Y|X}^{n-1}(y|x) = \lim_{n\to\infty} \binom{n-1}{y} \left(\frac{\theta x}{n}\right)^y \left(1-\frac{\theta x}{n}\right)^{n-1-y} \tag{6.34}$$

$$= \frac{(\theta x)^y}{y!} \lim_{n\to\infty} \frac{n(n-1)\cdots(n-y)}{(n-\theta x)^{y+1}} \left(1-\frac{\theta x}{n}\right)^n \tag{6.35}$$

$$= \frac{(\theta x)^y}{y!} e^{-\theta x} \tag{6.36}$$

$$= P_{Y|X}(y|x). \tag{6.37}$$

By the Dominated convergence theorem [6, Theorem 8.8],

$$\lim_{n\to\infty} \mathsf{E}^{n-1}[X|Y=y] = \frac{\lim_{n\to\infty} \mathsf{E}\left[X P_{Y|X}^{n-1}(y|X)\right]}{\lim_{n\to\infty} \mathsf{E}\left[P_{Y|X}^{n-1}(y|X)\right]}$$

$$= \frac{\mathsf{E}[X P_{Y|X}(y|X)]}{\mathsf{E}[P_{Y|X}(y|X)]} \tag{6.38}$$

where $P_{Y|X}$ is Poisson with mean θX. □

In the next Lemma, we show that the tail of a binomial distribution with parameters $(n-1, \theta X/n)$ is upper bounded by a Poisson distribution with rate θX for all y sufficiently large.

Lemma 16. *Let* $P_{Y|X}^{n-1}$ *be a binomial distribution with parameters* $(n-1, \theta X/n)$ *where* X *is a positive bounded random variable. Define*

$$\mu(y,x) = \begin{cases} \frac{(\theta x)^y}{y!} e^{-\theta x} & \text{if } y \geq 2\lceil \theta x_{\max} \rceil, \\ 1 & \text{if } 0 \leq y < 2\lceil \theta x_{\max} \rceil. \end{cases} \tag{6.39}$$

Then, for all $n > \theta x_{\max}$ *and* $y \in \{0, 1, \dots\}$,

$$\mu(y,x) > P_{Y|X}^{n-1}(y|x). \tag{6.40}$$

Proof. The case of $y < 2\lceil \theta x_{\max} \rceil$ is because $P_{Y|X}^{n-1}(y|x)$ is a pmf. For $y \geq 2\lceil \theta x_{\max} \rceil$, we rearrange $P_{Y|X}^{n-1}(y|x)$:

$$\begin{aligned} P_{Y|X}^{n-1}(y|x) &= \binom{n-1}{y} \left(\frac{\theta x}{n}\right)^y \left(1 - \frac{\theta x}{n}\right)^{n-y-1} \\ &= \frac{(\theta x)^y}{y!} \left(1 - \frac{\theta x}{n}\right)^n \underbrace{\frac{n}{n-\theta x} \cdots \frac{n-y}{n-\theta x}}_{\Gamma}. \end{aligned} \tag{6.41}$$

It then suffices to show that $\Gamma < 1$ for $y \geq 2\lceil \theta x_{\max} \rceil$ and that

$$\left(1 - \frac{\theta x}{n}\right)^n < e^{-\theta x}, \tag{6.42}$$

which is due to

$$n \log\left(1 - \frac{\theta x}{n}\right) = -n \sum_{i=1}^{\infty} \frac{1}{i} \left(\frac{\theta x}{n}\right)^i \tag{6.43}$$

$$= -\theta x - n \sum_{i=2}^{\infty} \frac{1}{i} \left(\frac{\theta x}{n}\right)^i \tag{6.44}$$

$$< -\theta x. \tag{6.45}$$

To complete the proof, we can notice that,

$$\Gamma = \prod_{k=0}^{y} \frac{n-k}{n-\theta x} \tag{6.46}$$

$$\leq \prod_{k=0}^{y} \frac{n-k}{n - \lceil \theta x_{\max} \rceil} \tag{6.47}$$

$$= \prod_{k=0}^{2\lceil \theta x_{\max} \rceil} \frac{n-k}{n - \lceil \theta x_{\max} \rceil} \prod_{k=2\lceil \theta x_{\max} \rceil + 1}^{y} \frac{n-k}{n - \lceil \theta x_{\max} \rceil} \tag{6.48}$$

$$\leq \prod_{k=0}^{2\lceil \theta x_{\max} \rceil} \frac{n-k}{n - \lceil \theta x_{\max} \rceil} \tag{6.49}$$

$$= \prod_{k=0}^{\lceil \theta x_{\max} \rceil - 1} \frac{(n-k)(n - (2\lceil \theta x_{\max} \rceil - k))}{(n - \lceil \theta x_{\max} \rceil)^2} \tag{6.50}$$

$$\leq 1, \tag{6.51}$$

where (6.47) is because $x_{\max}$ is the largest value that x can take, (6.49) is because each term in the inner product in (6.48) is less than 1, and (6.51) inequality is because the area of a square is greater than the area of a rectangle with the same perimeter. □

Proof of (6.8) *in Theorem 24.* We first define

$$m = \lceil \theta x_{\max} \rceil \tag{6.52}$$

where $m < \infty$ for any finite θ. Applying Theorem 12 with $X_\theta = \theta X/n$ we have that,

$$\lim_{n\to\infty} \frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) = \lim_{n\to\infty} \frac{n}{\theta} \mathsf{E}^{n-1}\left[\ell_b\left(\frac{\theta X}{n}, \mathsf{E}^{n-1}\left[\frac{\theta X}{n}\middle| Y\right]\right)\right] \tag{6.53}$$

$$= \lim_{n\to\infty} \mathsf{E}^{n-1}\left[X \log \frac{X\left(1 - \frac{\theta}{n}\mathsf{E}^{n-1}[X|Y]\right)}{\left(1-\frac{\theta X}{n}\right)\mathsf{E}^{n-1}[X|Y]}\right] \tag{6.54}$$

$$= \lim_{n\to\infty} \mathsf{E}^{n-1}\left[X \log \frac{X}{1-\frac{\theta X}{n}}\right] - \lim_{n\to\infty} \mathsf{E}^{n-1}\left[\mathsf{E}^{n-1}[X|Y] \log \frac{\mathsf{E}^{n-1}[X|Y]}{1-\frac{\theta}{n}\mathsf{E}^{n-1}[X|Y]}\right] \tag{6.55}$$

for $n \geq m$. To complete the proof, we need to interchange the limit and the expectation in both terms in (6.55). To do so, we rely on the domineted convergence theorem [6, Theorem 8.8]. For the first term at the RHS of (6.55) we have the following analysis. We first prove that, for all $n \geq m$, the expectations are finite, *i.e.*,

$$\mathsf{E}^{n-1}\left[X \log \frac{X}{1-\frac{\theta X}{n}}\right] = \mathsf{E}^{n-1}\left[X \log X\right] + \mathsf{E}^{n-1}\left[X \log \frac{1}{1-\frac{\theta X}{n}}\right] \tag{6.56}$$

$$\leq x_{\max} \log x_{\max} + x_{\max} \log \frac{1}{1-\frac{\theta x_{\max}}{m}} \tag{6.57}$$

$$< \infty. \tag{6.58}$$

The maximum is achieved when P_X is a delta function at $x_{\max}$ and $n = m$. As n increases this value decreases and this expectation is always finite.

Second, we prove that there exists a function $\omega(x) \geq \left| x \log \frac{x}{1-\frac{\theta x}{n}} \right|$ such that $\mathsf{E}[\omega(X)] < \infty$. In effect,

$$\left| x \log \frac{x}{1-\frac{\theta x}{n}} \right| \leq |x \log x| + \left| x \log \left(1 - \frac{\theta x}{n}\right) \right| \tag{6.59}$$

$$\leq M^* + x_{\max} \left| \log \left(1 - \frac{\theta x_{\max}}{m}\right) \right| \tag{6.60}$$

$$\triangleq \omega(x), \tag{6.61}$$

where,

$$M^* \triangleq \sup \left\{ e^{-1}, x_{\max} \log x_{\max} \right\}. \tag{6.62}$$

Now we can exchange the limit and the expectation. Hence,

$$\lim_{n\to\infty} \mathsf{E}^{n-1} \left[X \log \frac{X}{1-\frac{\theta X}{n}} \right] = \mathsf{E} \left[\lim_{n\to\infty} X \log \frac{X}{1-\frac{\theta X}{n}} \right] \tag{6.63}$$

$$= \mathsf{E} \left[X \log X \right]. \tag{6.64}$$

To calculate the second term at the RHS of (6.55) we define

$$\vartheta^{n-1}(y) = \mathsf{E}^{n-1}[X|Y=y] \log \frac{\mathsf{E}^{n-1}[X|Y=y]}{1-\frac{\theta}{n}\mathsf{E}^{n-1}[X|Y=y]} \tag{6.65}$$

for $y \in \{0, 1, \ldots, n-1\}$. To show that

$$\lim_{n\to\infty} \sum_{y=0}^{\infty} P_Y^{n-1}(y) \vartheta^{n-1}(y) = \sum_{y=0}^{\infty} \lim_{n\to\infty} P_Y^{n-1}(y) \vartheta^{n-1}(y), \tag{6.66}$$

where we define $P_Y^{n-1}(y) = 0$ for all $y \geq n$, we verify the following conditions [6, Theorem 8.8]:

(i) $\sum_{y=0}^{\infty} P_Y^{n-1}(y) \vartheta^{n-1}(y) < \infty$, for all $n \geq m$.

(ii) There exists a summable function $\omega(y)$ such that, for all $n \geq m$,

$$\left| P_Y^{n-1}(y) \vartheta^{n-1}(y) \right| \leq \omega(y). \tag{6.67}$$

To prove Condition *(i)*, note that $0 \leq \mathsf{E}^{n-1}[X|Y=y] \leq x_{\max}$ leads to

$$\left| \vartheta^{n-1}(y) \right| \leq M^* + x_{\max} \left| \log \left(1 - \frac{\theta x_{\max}}{m}\right) \right|, \tag{6.68}$$

where M^* is given in (6.62), which yields,

$$\left|\sum_{y=0}^{\infty}\vartheta^{n-1}(y)P_Y^{n-1}(y)\right| \leq \sum_{y=0}^{\infty}|\vartheta^{n-1}(y)|P_Y^{n-1}(y) \tag{6.69}$$

$$= M^* + x_{\max}\left|\log\left(1-\frac{\theta x_{\max}}{m}\right)\right| \tag{6.70}$$

$$< \infty \tag{6.71}$$

for all $n \geq m$. To prove Condition *(ii)*, note that

$$P_Y^{n-1}(y) = \mathsf{E}\left[P_{Y|X}^{n-1}(y|X)\right] < \mathsf{E}\left[\mu(y,X)\right] \triangleq \nu(y). \tag{6.72}$$

due to Lemma 16. At this point the summability of the function $\nu(y)$ is guaranteed:

$$\sum_{y=0}^{\infty}\nu(y) = \sum_{y=0}^{2\lceil\theta x_{\max}\rceil} 1 + \sum_{y=2\lceil\theta x_{\max}+1\rceil}^{\infty} \mathsf{E}\left[\mu(y,X)\right] \tag{6.73}$$

$$< 2\lceil\theta x_{\max}\rceil + 2 \tag{6.74}$$

$$< \infty. \tag{6.75}$$

Based on (6.68) and (6.72), we have

$$P_Y^{n-1}(y)\left|\vartheta^{n-1}(y)\right|$$

$$\leq \left(M^* + x_{\max}\left|\log\left(1-\frac{\theta x_{\max}}{m}\right)\right|\right)\nu(y) \tag{6.76}$$

$$\triangleq \omega(y), \tag{6.77}$$

and consequently,

$$\sum_{y=0}^{\infty}\omega(y)$$

$$= \left(M^* + x_{\max}\left|\log\left(1-\frac{\theta x_{\max}}{m}\right)\right|\right)(2\lceil\theta x_{\max}\rceil + 2) \tag{6.78}$$

$$< \infty \tag{6.79}$$

for all $n \geq m$, which let us verify Condition *(ii)*. We can now exchange the

limit and the expectation, leading to,

$$\begin{aligned}
&\lim_{n\to\infty} P_Y^{n-1}(y)\vartheta^{n-1}(y) \\
&= \lim_{n\to\infty} P_Y^{n-1}(y)\mathsf{E}^{n-1}[X|Y=y]\log\frac{\mathsf{E}^{n-1}[X|Y=y]}{1-\frac{\theta}{n}\mathsf{E}^{n-1}[X|Y=y]} \qquad (6.80)\\
&= P_Y(y)\mathsf{E}[X|Y=y]\log\mathsf{E}[X|Y=y], \qquad (6.81)
\end{aligned}$$

where, to obtain (6.81) we have applied results in Lemma 15:

$$\lim_{n\to\infty}\mathsf{E}^{n-1}[X|Y=y] = \mathsf{E}[X|Y], \qquad (6.82)$$

and

$$\lim_{n\to\infty}\mathsf{E}[P_{Y|X}^{n-1}(y|X)] = P_Y(y). \qquad (6.83)$$

Combining (6.63) and (6.81) yields the desired result. □

Proof of (6.9) *in Theorem 24.* We now show that (6.8) and (6.9) are identical, *i.e.*,

$$\begin{aligned}
\mathsf{E}\left[X\log\frac{X}{\mathsf{E}[X|Y]}\right] &= \mathsf{E}\left[\sum_{y=0}^{\infty} X\log\frac{X}{\mathsf{E}[X|Y=y]}P_{Y|X}(y|X)\right] \qquad (6.84)\\
&= \mathsf{E}\left[\sum_{y=0}^{\infty} X\log\left(X\mathsf{E}[X^{-1}|Y=y+1]\right)P_{Y|X}(y|X)\right] \qquad (6.85)\\
&= -\mathsf{E}\left[\frac{1}{\theta}\sum_{y=0}^{\infty}\frac{(\theta X)^{y+1}}{y!}e^{-\theta X}\log\frac{X^{-1}}{\mathsf{E}[X^{-1}|Y=y+1]}\right] \qquad (6.86)\\
&= -\mathsf{E}\left[\frac{1}{\theta}\sum_{y=1}^{\infty}\frac{(\theta X)^{y}}{(y-1)!}e^{-\theta X}\log\frac{X^{-1}}{\mathsf{E}[X^{-1}|Y=y]}\right] \qquad (6.87)\\
&= -\mathsf{E}\left[\sum_{y=0}^{\infty}\frac{y}{\theta}\log\frac{X^{-1}}{\mathsf{E}[X^{-1}|Y=y]}P_{Y|X}(y|X)\right] \qquad (6.88)\\
&= -\mathsf{E}\left[\frac{Y}{\theta}\log\frac{X^{-1}}{\mathsf{E}[X^{-1}|Y]}\right] \qquad (6.89)\\
&= \mathsf{E}\left[\frac{Y}{\theta}\ell_I\left(X^{-1},\mathsf{E}\left[X^{-1}|Y\right]\right)\right], \qquad (6.90)
\end{aligned}$$

where in (6.85) we use the following fact:

$$\frac{1}{\mathsf{E}[X|Y=y]} = \frac{\mathsf{E}\left[\frac{(\theta X)^y}{y!}e^{-\theta X}\right]}{\mathsf{E}\left[X\frac{(\theta X)^y}{y!}e^{-\theta X}\right]} \tag{6.91}$$

$$= \frac{\mathsf{E}\left[X^{-1}(\theta X)^{y+1}e^{-\theta X}\right]}{\mathsf{E}\left[(\theta X)^{y+1}e^{-\theta X}\right]} \tag{6.92}$$

$$= \mathsf{E}[X^{-1}|Y=y+1]. \tag{6.93}$$

□

6.5.2 Proof of Theorem 28

Based on Theorem 12, the derivative of the input output mutual information over a binomial model with parameters $(n, X_\theta/n)$ is given by,

$$\frac{\mathrm{d}}{\mathrm{d}\theta}I(X;Y) = \mathsf{E}^{n-1}\left[X'_\theta \log \frac{(1-\mathsf{E}^{n-1}[X_\theta/n|Y])X_\theta}{(1-X_\theta/n)\mathsf{E}^{n-1}[X_\theta|Y]}\right]. \tag{6.94}$$

Hence, taking the limit, we get that,

$$\lim_{n\to\infty}\frac{\mathrm{d}}{\mathrm{d}\theta}I(X;Y) = \lim_{n\to\infty}\mathsf{E}^{n-1}\left[X'_\theta \log \frac{X_\theta}{(1-X_\theta/n)}\right] - \lim_{n\to\infty}\mathsf{E}^{n-1}\left[X'_\theta \log \frac{\mathsf{E}^{n-1}[X_\theta|Y]}{(1-\mathsf{E}^{n-1}[X_\theta/n|Y])}\right] \tag{6.95}$$

$$= \mathsf{E}\left[X'_\theta \log \frac{X_\theta}{\mathsf{E}[X_\theta|Y]}\right], \tag{6.96}$$

where in order to obtain (6.96) rigorously speaking, we use the same procedure used in the proof of Theorem 24 jointly with the dominated convergence theorem [6, Theorem 8.8] and with Lemmas 15 and 16 generalized to the case when $\theta X = X_\theta$ where X_θ is a bounded measurable function of the input X.

6.5.3 Proof of Theorem 29

The derivative of the relative entropy between two output distributions P_Y^n and Q_Y^n for a binomial model with parameters $(n, X_\theta/n)$, given by Theorem

11 in Section 4.2.2 is expressed as follows,

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\theta} D(P_Y^n || Q_Y^n) \\
&= \mathsf{F}_Q^{n-1}(X_\theta/n) - \mathsf{F}_P^{n-1}(X_\theta/n) \qquad (6.97) \\
&= \mathsf{E}^{n-1}\left[X'_\theta \log \frac{(1-\mathsf{E}_Q^{n-1}[X_\theta/n|Y])X_\theta}{(1-X_\theta/n)\mathsf{E}^{n-1}[X_\theta|Y]} \right] \\
&\quad - \mathsf{E}^{n-1}\left[\frac{\mathsf{E}^{n-1}[X'_\theta|Y](X_\theta - \mathsf{E}_Q^{n-1}[X_\theta/n|Y])}{(1-\mathsf{E}_Q^{n-1}[X_\theta/n|Y])\mathsf{E}_Q^{n-1}[X_\theta]} \right] \\
&\quad - \mathsf{E}^{n-1}\left[X'_\theta \log \frac{(1-\mathsf{E}^{n-1}[X_\theta/n|Y])}{(1-X_\theta/n)\mathsf{E}^{n-1}[X_\theta|Y]} \right] \qquad (6.98) \\
&= \mathsf{E}^{n-1}\left[X'_\theta \log \frac{(1-\mathsf{E}_Q^{n-1}[X_\theta/n|Y])\mathsf{E}^{n-1}[X_\theta|Y]}{(1-\mathsf{E}^{n-1}[X_\theta/n|Y])\mathsf{E}_Q^{n-1}[X_\theta|Y]} \right] \\
&\quad - \mathsf{E}^{n-1}\left[\frac{\mathsf{E}^{n-1}[X'_\theta|Y](X_\theta - \mathsf{E}_Q^{n-1}[X_\theta/n|Y])}{(1-\mathsf{E}_Q^{n-1}[X_\theta/n|Y])\mathsf{E}_Q^{n-1}[X_\theta]} \right]. \qquad (6.99)
\end{aligned}$$

Carrying out the limit, let us get,

$$\begin{aligned}
&\lim_{n\to\infty} \frac{\mathrm{d}}{\mathrm{d}\theta} D(P_Y^n || Q_Y^n) \\
&= \mathsf{E}\left[\mathsf{E}[X'_\theta|Y] \log \frac{\mathsf{E}[X_\theta|Y]}{\mathsf{E}_Q[X_\theta|Y]} - \frac{\mathsf{E}_Q[X'_\theta|Y](\mathsf{E}[X_\theta|Y] - \mathsf{E}_Q[X_\theta|Y])}{\mathsf{E}_Q[X_\theta|Y]} \right] \qquad (6.100)
\end{aligned}$$

where in order to obtain (6.100) we introduce the limit inside the expectation operator in (6.99). This procedure can be justified using the same steps used in the proof of Theorem 24 changing θX by X_θ.

6.5.4 Proof of Theorem 30

Before proceed with a formal proof to Theorem 30 we state some lemmas required along the proof.

Lemma 17. *Let $X \sim P_X$ be a positive bounded random variable. Then, for a Poisson model* (6.2) *with mean $\theta X > 0$,*

$$\lim_{\theta\to 0} \mathsf{E}[X P_{Y|X}(y|X)] = \mathsf{E}[X] \mathbb{1}_{\{y=0\}} \qquad (6.101)$$

and

$$\lim_{\theta\to 0} \mathsf{E}[P_{Y|X}(y|X)] = \mathbb{1}_{\{y=0\}}. \qquad (6.102)$$

Proof. By hypothesis we have that $\mathsf{E}[X] < \infty$. Therefore, based on the fact that,

$$\lim_{\theta \to 0} x P_{Y|X}(y|x) = x \mathbb{1}_{\{y=0\}} \tag{6.103}$$

and

$$\lim_{\theta \to 0} P_{Y|X}(y|x) = \mathbb{1}_{\{y=0\}} \tag{6.104}$$

together with the Dominated Convergence Theorem [6, Theorem 8.8] let us state the desired result. □

Lemma 18. *For a bounded input X, let Y be the output of a Poisson model (6.2) with mean θX. Then, for a given $Y = y \in Z_o^+$,*

$$\begin{aligned} &\lim_{\theta \to 0} P_Y(y) \mathsf{E}[X|Y=y] \log \mathsf{E}[X|Y=y] \\ &\qquad = P_Y(y) \mathsf{E}[X|Y=y] \log \mathsf{E}[X|Y=y]\big|_{\theta=0}. \end{aligned} \tag{6.105}$$

Proof. For a given $Y = y \in Z_o^+$ notice that,

$$0 < \mathsf{E}[X P_{Y|X}(y|X)] \leq \frac{(\theta x_{\max})^y}{y!} \mathsf{E}[X], \tag{6.106}$$

which implies that,

$$\mathsf{E}[X P_{Y|X}(y|X)]\big|_{\theta=0} = \mathsf{E}[X] \mathbb{1}_{\{y=0\}} \tag{6.107}$$

and analogously,

$$\mathsf{E}[P_{Y|X}(y|X)]\big|_{\theta=0} = \mathbb{1}_{\{y=0\}}. \tag{6.108}$$

Based on (6.107) and (6.108) we get that,

$$\begin{aligned} &P_Y(y) \mathsf{E}[X|Y=y] \log \mathsf{E}[X|Y=y]\big|_{\theta=0} \\ &\qquad = \mathsf{E}[X P_{Y|X}(y|X)] \left(\log \mathsf{E}[X P_{Y|X}(y|X)] - \log \mathsf{E}[P_{Y|X}(y|X)] \right) && (6.109) \\ &\qquad = \mathsf{E}[X] \log \mathsf{E}[X] \mathbb{1}_{\{y=0\}} && (6.110) \end{aligned}$$

where we use the convention $0 \log 0 = 0$. On the other hand, based on Lemma 17 we have that,

$$\lim_{\theta \to 0} \mathsf{E}[X P_{Y|X}(y|X)] = \mathsf{E}[X] \mathbb{1}_{\{y=0\}} \tag{6.111}$$

and

$$\lim_{\theta\to 0} \mathsf{E}[P_{Y|X}(y|X)] = \mathbb{1}_{\{y=0\}}. \tag{6.112}$$

Therefore, for $y = 0$,

$$\begin{aligned}
&\lim_{\theta\to 0} P_Y(0)\mathsf{E}[X|Y=0]\log \mathsf{E}[X|Y=0] \\
&\quad = \lim_{\theta\to 0} \mathsf{E}[XP_{Y|X}(0|X)]\left(\log \mathsf{E}[XP_{Y|X}(0|X)] - \log \mathsf{E}[P_{Y|X}(0|X)]\right) \qquad (6.113)\\
&\quad = \mathsf{E}[X]\log \mathsf{E}[X]. \qquad (6.114)
\end{aligned}$$

Indeed, notice that our main objective now is to calculate, for $y \neq 0$ the following expression,

$$\begin{aligned}
&\lim_{\theta\to 0} P_Y(y)\mathsf{E}[X|Y=y]\log \mathsf{E}[X|Y=y] \\
&\quad = \lim_{\theta\to 0} \mathsf{E}[XP_{Y|X}(y|X)]\left(\log \mathsf{E}[XP_{Y|X}(y|X)] - \log \mathsf{E}[P_{Y|X}(y|X)]\right). \qquad (6.115)
\end{aligned}$$

Taking into account (6.111), what is left to calculate (6.115) is to analyze the following expression,

$$\lim_{\theta\to 0} \mathsf{E}[XP_{Y|X}(y|X)]\log \mathsf{E}[P_{Y|X}(y|X)] \tag{6.116}$$

when $y \neq 0$. To do so, notice that,

$$\begin{aligned}
\mathsf{E}[XP_{Y|X}(y|X)] &= \mathsf{E}\left[X\frac{(\theta X)^y}{y!}e^{-\theta X}\right] &(6.117)\\
&= \mathsf{E}\left[\frac{1}{y!}\sum_{k=0}^{\infty}\frac{(-1)^k}{k!}\theta^{k+y}X^{k+y+1}\right] &(6.118)\\
&= \frac{1}{y!}\sum_{k=0}^{\infty}\frac{(-1)^k}{k!}\theta^{k+y}\mathsf{E}[X^{k+y+1}] &(6.119)
\end{aligned}$$

where, to obtain (6.119) we use the uniform convergence [47, Theorem 7.10] of the function

$$\sum_{k=0}^{\infty}\frac{(-1)^k}{y!}\frac{\theta^{k+y}}{k!}x^{k+y+1}, \tag{6.120}$$

jointly with [47, Theorem 7.16]. To calculate (6.116), applying the L'Hospital rule [47] let us get,

$$\lim_{\theta\to 0}\mathsf{E}[XP_{Y|X}(y|X)]\log\mathsf{E}[P_{Y|X}(y|X)] = \\ -\lim_{\theta\to 0}\frac{\frac{\mathrm{d}}{\mathrm{d}\theta}\mathsf{E}[P_{Y|X}(y|X)]\mathsf{E}[XP_{Y|X}(y|X)]^2}{\frac{\mathrm{d}}{\mathrm{d}\theta}\mathsf{E}[XP_{Y|X}(y|X)]\mathsf{E}[P_{Y|X}(y|X)]}. \tag{6.121}$$

In order to resolve (6.121), let us define,

$$g_k(\theta) \triangleq \frac{(-1)^k}{y!}\frac{\theta^{k+y}}{k!}\mathsf{E}[X^{k+y+1}]. \tag{6.122}$$

The derivative of each term $g_\theta(\theta)$ with respect to θ, denoted as $g'_k(\theta)$ is given by,

$$g'_k(\theta) = \frac{(-1)^k}{y!}\frac{\theta^{k+y-1}}{k!}(k+y)\mathsf{E}[X^{k+y+1}] \tag{6.123}$$

which constitutes a sequence of continuous functions for $\theta\in[0,\delta)$, $\delta>\mathbb{R}^+$. Based on [47, Theorem 7.10] we have that,

$$\sum_{k=0}^{\infty} g'_k(\theta) \tag{6.124}$$

converges uniformly for $\theta\in[0,\delta)$. Hence we get that,

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathsf{E}[XP_{Y|X}(y|X)] = \sum_{k=0}^{\infty}\frac{(-1)^k}{y!}\frac{\theta^{k+y-1}}{k!}(k+y)\mathsf{E}[X^{k+y+1}] \tag{6.125}$$

constitutes a polynomial in θ where the least exponent is of grade $(k-1)$. Similarly we have that,

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathsf{E}[P_{Y|X}(y|X)] = \sum_{k=0}^{\infty}\frac{(-1)^k}{y!}\frac{\theta^{k+y-1}}{k!}(k+y)\mathsf{E}[X^{k+y}] \tag{6.126}$$

which also represents a polynomial where the least exponent is of grade $(k-1)$. Based on these properties notice that due to (6.119) and (6.126) the product,

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathsf{E}[P_{Y|X}(y|X)]\mathsf{E}[XP_{Y|X}(y|X)]^2 \\ =\left(\sum_{k=0}^{\infty}\frac{(-1)^k}{y!}\frac{\theta^{k+y-1}}{k!}(k+y)\mathsf{E}[X^{k+y}]\right)\left(\sum_{k=0}^{\infty}\frac{(-1)^k}{y!}\frac{\theta^{k+y}}{k!}\mathsf{E}[X^{k+y+1}]\right)^2 \tag{6.127}$$

is a polynomial on θ with exponents belonging to the set $\{3y-1, 3y, \cdots\}$. Similarly,

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\theta}\mathsf{E}[XP_{Y|X}(y|X)]\mathsf{E}[P_{Y|X}(y|X)] \\
&= \left(\sum_{k=0}^{\infty} \frac{(-1)^k}{y!}\frac{\theta^{k+y-1}}{k!}(k+y)\mathsf{E}[X^{k+y+1}]\right)\left(\sum_{k=0}^{\infty} \frac{(-1)^k}{y!}\frac{\theta^{k+y}}{k!}\mathsf{E}[X^{k+y}]\right)
\end{aligned} \tag{6.128}$$

constitutes a polynomial on θ with exponents within the set $\{2y-1, 2y, \cdots\}$. Therefore, for $y \neq 0$ in (6.121), we get,

$$\begin{aligned}
&\lim_{\theta\to 0}\mathsf{E}[XP_{Y|X}(y|X)]\log\mathsf{E}[P_{Y|X}(y|X)] \\
&\quad= -\lim_{\theta\to 0}\frac{\frac{\mathrm{d}}{\mathrm{d}\theta}\mathsf{E}[P_{Y|X}(y|X)]\mathsf{E}[XP_{Y|X}(y|X)]^2}{\frac{\mathrm{d}}{\mathrm{d}\theta}\mathsf{E}[XP_{Y|X}(y|X)]\mathsf{E}[P_{Y|X}(y|X)]} && (6.129)\\
&\quad= -\lim_{\theta\to 0}\frac{\left(\frac{y}{y!}\theta^{y-1}\mathsf{E}[X^y] + \cdots\right)\left(\frac{1}{y!}\theta^{y}\mathsf{E}[X^{y+1}] + \cdots\right)^2}{\left(\frac{y}{y!}\theta^{y-1}\mathsf{E}[X^{y+1}] + \cdots\right)\left(\frac{1}{y!}\theta^{y}\mathsf{E}[X^{y}] + \cdots\right)} && (6.130)\\
&\quad= -\lim_{\theta\to 0}\frac{\left(\frac{y}{(y!)^3}\theta^{3y-1}\mathsf{E}[X^y]\mathsf{E}[X^{y+1}]^2 + \cdots\right)}{\left(\frac{y}{(y!)^2}\theta^{2y-1}\mathsf{E}[X^{y+1}]\mathsf{E}[X^{y}] + \cdots\right)} && (6.131)\\
&\quad= -\lim_{\theta\to 0}\frac{\left(\frac{1}{y!}\theta^{y}\mathsf{E}[X^y]\mathsf{E}[X^{y+1}]^2 + \cdots\right)}{(\mathsf{E}[X^{y+1}]\mathsf{E}[X^{y}] + \cdots)} && (6.132)\\
&\quad= 0 && (6.133)
\end{aligned}$$

where (6.130) appears as consequence of (6.128) and (6.127), (6.131) appears when we multiply the first terms of each element in (6.130) and (6.132) appears when we multiply and divide by θ^{2y-1}. Consequently for $y \neq 0$,

$$\begin{aligned}
&\lim_{\theta\to 0} P_Y(y)\mathsf{E}[X|Y=y]\log\mathsf{E}[X|Y=y] \\
&\quad= \lim_{\theta\to 0}\mathsf{E}[XP_{Y|X}(y|X)]\left(\log\mathsf{E}[XP_{Y|X}(y|X)] - \log\mathsf{E}[P_{Y|X}(y|X)]\right) && (6.134)\\
&\quad= 0 && (6.135)
\end{aligned}$$

where (6.135) appears as consequence of (6.133) and (6.111). We obtain the desired result once we compare (6.135) and (6.114) with (6.110). □

Proof of Theorem 30. Departing from the expression for the derivative of the input output mutual information over a Poisson model with mean θX given in (1.13), we prove that,

$$\lim_{\theta\to 0}\frac{\mathrm{d}}{\mathrm{d}\theta}I(X;Y)=\lim_{\theta\to 0}\mathsf{E}\left[X\log\frac{X}{\mathsf{E}[X|Y]}\right] \tag{6.136}$$

$$=\mathsf{E}\left[X\log X\right]-\sum_{y=0}^{\infty}\lim_{\theta\to 0}P_Y(y)\mathsf{E}[X|Y=y]\log\mathsf{E}[X|Y=y]. \tag{6.137}$$

To exchange the sum with the limit in (6.137) we verify the following conditions,

(i) The function $P_Y(y)\mathsf{E}[X|Y=y]\log\mathsf{E}[X|Y=y]$ is summable over $\mathbb{Z}_0^+$ for each $\theta\in[0,\delta),\delta\in\mathbb{R}^+$.

(ii) There exists summable functions $\xi(y)$ and $\omega(y)$ such that, for all $y\in\mathbb{Z}_0^+$,

$$\xi(y)\leq P_Y(y)\mathsf{E}[X|Y=y]\log\mathsf{E}[X|Y=y]\leq\omega(y) \tag{6.138}$$

(iii) For a given $Y=y$, the limit,

$$\lim_{\theta\to 0}P_Y(y)\mathsf{E}[X|Y=y]\log\mathsf{E}[X|Y=y] \tag{6.139}$$

exists and is equal to,

$$P_Y(y)\mathsf{E}[X|Y=y]\log\mathsf{E}[X|Y=y]\big|_{\theta=0}. \tag{6.140}$$

In effect, notice that the function $|\mathsf{E}[X|Y=y]\log\mathsf{E}[X|Y=y]|$ is upper bounded by $M^*=\sup\{e^{-1},x_{\max}\log x_{\max}\}$ for all $\theta\in(0,\delta)$. This condition, jointly with (6.110) let us state Conditions *(i)* and *(ii)*. Finally, Condition *(iii)* is verified in Lemma 18. Consequently, we get that,

$$\lim_{\theta\to 0}P_Y(y)\mathsf{E}[X|Y=y]\log\mathsf{E}[X|Y=y]=\mathsf{E}[X]\log\mathsf{E}[X]\mathbb{1}_{\{y=0\}}. \tag{6.141}$$

Replacing (6.141) in (6.137) yields,

$$\lim_{\theta\to 0}\frac{\mathrm{d}}{\mathrm{d}\theta}I(X;Y)=\mathsf{E}[X\log X]-\mathsf{E}[X]\log\mathsf{E}[X]. \tag{6.142}$$

□

6.5.5 Proof of Theorem 32

Let P_Y be the marginal distribution obtained at the output of a Poisson model with mean X_θ. In the following Lemma, for a given $Y = y$, we show an expression for the derivative of the marginal distribution P_Y with respect to changes in the input scaling θ. Throughout this section we assume that the set of feasible values is an open set, denoted by Θ.

Lemma 19. *Let X_θ be a measurable differentiable bounded function such that $|X'_\theta| < M, M \in \mathbb{R}^+$ for all $x \in \mathcal{X}$ and and $\theta \in \Theta$. Then, for a Poisson model with mean $X_\theta > 0$,*

$$\frac{\mathrm{d}}{\mathrm{d}\theta} P_Y(y) = \mathsf{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta} P_{Y|X}(y|X)\right] \tag{6.143}$$

$$= y\mathsf{E}\left[\frac{X'_\theta}{X_\theta}\middle| Y\right] P_Y(y) - \mathsf{E}\left[X'_\theta|Y\right]. \tag{6.144}$$

Proof. Based on [6, Theorem 12.13] we proceed to verify the following conditions in order to exchange the derivative with the expectation,

(i) The derivative $\frac{\mathrm{d}}{\mathrm{d}\theta} P_{Y|X}$ exists for all values of $\theta \in \Theta$.

(ii) There exists a function $\omega(x)$ such that

$$\left|\frac{\mathrm{d}}{\mathrm{d}\theta} P_{Y|X}(y|x)\right| \leq \omega(x) \tag{6.145}$$

and $\mathsf{E}[\omega(X)] < \infty$.

In effect,

$$\frac{\mathrm{d}}{\mathrm{d}\theta} P_{Y|X}(y|x) = \frac{y}{f(\theta,x)} f'(\theta,x) P_{Y|X}(y|x) - f'(\theta,x) P_{Y|X}(y|x) \tag{6.146}$$

which is well defined for all values of $\theta \in \Theta$, letting us verify Condition *(i)*. Condition *(ii)* is verified as follows;

$$\left|\frac{\mathrm{d}}{\mathrm{d}\theta} P_{Y|X}(y|x)\right| \leq \left|\frac{y}{f(\theta,x)} f'(\theta,x)\right| P_{Y|X}(y|x) + \left|f'(\theta,x)\right| P_{Y|X}(y|x) \tag{6.147}$$

$$= \left|f'(\theta,x)\right| \frac{f(\theta,x)^{y-1}}{(y-1)!} e^{-f(\theta,x)} + \left|f'(\theta,x)\right| P_{Y|X}(y|x) \tag{6.148}$$

$$\leq M\left(P_{Y|X}(y-1|x)\mathbb{1}_{\{y>0\}} + P_{Y|X}(y|x)\right) \tag{6.149}$$

$$< 2M \tag{6.150}$$

$$\triangleq \omega(x). \tag{6.151}$$

Verification of Conditions *(i)* and *(ii)* let us conclude that

$$\frac{\mathrm{d}}{\mathrm{d}\theta}P_Y(y) = \mathsf{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}P_{Y|X}(y|X)\right] \tag{6.152}$$

$$= y\,\mathsf{E}\left[\frac{X'_\theta}{X_\theta}\middle| Y = y\right] P_Y(y) - \mathsf{E}\left[X'_\theta|Y=y\right] P_Y(y) \tag{6.153}$$

where (6.153) can be seen as consequence of (6.146). □

Proof of Theorem 32. Let Y be the output of a Poisson model with mean X_θ. By definition we have that,

$$I(X;Y) = \mathsf{E}\left[\log\frac{P_{Y|X}(Y|X)}{P_Y(Y)}\right]. \tag{6.154}$$

Taking the derivative with respect to the parameter θ let us get,

$$\frac{\mathrm{d}}{\mathrm{d}\theta}I(X;Y) = \mathsf{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\sum_{y=0}^{\infty} P_{Y|X}(y|X)\log\frac{P_{Y|X}(y|X)}{P_Y(y)}\right] \tag{6.155}$$

$$= \mathsf{E}\left[\sum_{y=0}^{\infty}\frac{\mathrm{d}}{\mathrm{d}\theta}P_{Y|X}(y|X)\log\frac{P_{Y|X}(y|X)}{P_Y(y)} - \frac{P_{Y|X}(y|X)}{P_Y(y)}\frac{\mathrm{d}}{\mathrm{d}\theta}P_Y(y)\right], \tag{6.156}$$

where in (6.155) the derivative penetrates the expectation operator based on the Dominated Convergence Theorem [47] together with Theorem [6, 12.13].

The first term at the RHS of (6.156) is calculated as follows,

$$\mathsf{E}\left[\sum_{y=0}^{\infty}\frac{\mathrm{d}}{\mathrm{d}\theta}P_{Y|X}(y|X)\log\frac{P_{Y|X}(y|X)}{P_Y(y)}\right]$$

$$=\mathsf{E}\left[\sum_{y=0}^{\infty}\left(y\frac{X'_\theta}{X_\theta}P_{Y|X}(y|X)-X'_\theta P_{Y|X}(y|X)\right)\log\frac{P_{Y|X}(y|X)}{P_Y(y)}\right] \tag{6.157}$$

$$=\mathsf{E}\left[\sum_{y=1}^{\infty}\frac{X'_\theta}{X_\theta}\frac{(X_\theta)^y}{(y-1)!}e^{-X_\theta}\log\frac{P_{Y|X}(y|X)}{P_Y(y)}-\sum_{y=0}^{\infty}X'_\theta P_{Y|X}(y|X)\log\frac{P_{Y|X}(y|X)}{P_Y(y)}\right] \tag{6.158}$$

$$=\mathsf{E}\left[\sum_{y=0}^{\infty}X'_\theta P_{Y|X}(y|X)\log\frac{X_\theta P_{Y|X}(y|X)}{\mathsf{E}[X_\theta P_{Y|X}(y|X)]}-X'_\theta P_{Y|X}(y|X)\log\frac{P_{Y|X}(y|X)}{P_Y(y)}\right] \tag{6.159}$$

$$=\mathsf{E}\left[\sum_{y=0}^{\infty}X'_\theta P_{Y|X}(y|X)\log\frac{X_\theta P_Y(y)}{\mathsf{E}[X_\theta P_{Y|X}(y|X)]}\right] \tag{6.160}$$

$$=\mathsf{E}\left[X'_\theta\log\frac{X_\theta}{\mathsf{E}[X_\theta|Y]}\right] \tag{6.161}$$

where; in (6.157) we calculate the derivative of the conditional $P_{Y|X}$ with respect to θ, (6.158) appears when we eliminate the first term of the sum given that it is zero, (6.159) appears when we change y by $y-1$ in the index of the sum, and (6.161) is consequence of the fact that $\mathsf{E}[X_\theta|Y=y]=\frac{\mathsf{E}[X_\theta P_{Y|X}(y|X)]}{P_Y(y)}$. Subsequently, the second term at the RHS of (6.156) is given

by,

$$\mathsf{E}\left[\sum_{y=0}^{\infty}\frac{P_{Y|X}(y|X)}{P_Y(y)}\frac{\mathrm{d}}{\mathrm{d}\theta}P_Y(y)\right]$$

$$=\mathsf{E}\left[\sum_{y=0}^{\infty}P_{Y|X}(y|X)\left(y\mathsf{E}\left[\frac{X'_\theta}{X_\theta}\middle|Y=y\right]-\mathsf{E}\left[X'_\theta|Y=y\right]\right)\right] \tag{6.162}$$

$$=\mathsf{E}\left[\sum_{y=1}^{\infty}\frac{(X_\theta)^y}{(y-1)!}e^{-X_\theta}\frac{\mathsf{E}\left[X'_\theta/X_\theta P_{Y|X}(y|X)\right]}{\mathsf{E}\left[P_{Y|X}(y|X)\right]}-\sum_{y=0}^{\infty}P_{Y|X}(y|X)\mathsf{E}[X'_\theta|Y=y]\right] \tag{6.163}$$

$$=\mathsf{E}\left[\sum_{y=0}^{\infty}X_\theta P_{Y|X}(y|X)\frac{\mathsf{E}[X'_\theta(X_\theta)^y e^{-X_\theta}]}{\mathsf{E}[X_\theta(X_\theta)^y e^{-X_\theta}]}-P_{Y|X}(y|X)\mathsf{E}[X'_\theta|Y=y]\right] \tag{6.164}$$

$$=\sum_{y=0}^{\infty}\mathsf{E}[X'_\theta P_{Y|X}(y|X)]-\sum_{y=0}^{\infty}P_Y(y)\mathsf{E}[X'_\theta|Y=y] \tag{6.165}$$

$$=0, \tag{6.166}$$

where (6.162) is consequence of Lemma 19, (6.163) appears when we eliminate the first term of the sum because it is zero, (6.164) appears when we change y by $y-1$ in the index of the sum, (6.165) is consequence of Fubinni's Theorem [16, pag. 467] and finally, (6.166) appears as consequence of the fact that $\mathsf{E}[X'_\theta|Y=y]=\frac{\mathsf{E}[X'_\theta P_{Y|X}(y|X)]}{P_Y(y)}$. □

Chapter 7

Conclusions and Ongoing Work

7.1 Conclusions

This thesis explores information–estimation relationships over binomial, negative binomial and Poisson models. As a starting point, we show that those results obtained initially for the Gaussian and Poisson models share the property that they can be represented entirely through the use of Bregman divergences with the characteristic that those results only depend on input statistics and their respective conditional estimates. This property, as is shown in Chapter 2 is not fulfilled by the vast majority of results developed recently, where the expressions given depend on the output of the model Y not only through conditional estimates of the input. This fact plays a fundamental role, given that the connections between information and estimation only depending on conditional estimates through the Bregman divergence give rise to the denominated "I-MMSE" and "I-MMLE" relationships over the Gaussian and Poisson models, respectively.

Based on the previous results we show similar relationships in the context of the binomial and negative binomial models. Over each model, using a deterministic input preprocessing function X_θ, we develop several information–estimation relationships, depending solely on input statistics and its respective conditional estimates, that in some scenarios are given through Bregman divergences as was done formerly for the Gaussian and Poisson models. We highlight the following features over the results obtained:

- Linear input scaling in the mean of the model: This scenario is similar

to those studied previously for the Gaussian and Poisson models. We show for the binomial and negative binomial models that the derivative of the input–output mutual information is given through a Bregman divergence where the arguments are the mean of the model θX and its conditional estimate. Mathematically speaking, we prove for the mutual information that,

$$\frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) \propto \mathsf{E}\left[\ell(\theta X, \mathsf{E}[\theta X|Y])\right] \tag{7.1}$$

where the loss function ℓ is a Bregman divergence. This condition gives rise to relationships that are of the same kind that the "I-MMSE" and the "I-MMLE" found initially for the Gaussian and Poisson models. Similar expressions are developed for the relative entropy concept, where the arguments of the Bregman divergence used are the conditional estimate of the mean θX under the distribution P_X and its correspondent mismatched version when $X \sim Q_X$, *i.e.*,

$$\frac{\mathrm{d}}{\mathrm{d}\theta} D(P_Y||Q_Y) \propto \mathsf{E}\left[\ell(\mathsf{E}[\theta X|Y], \mathsf{E}_Q[\theta X|Y])\right] \tag{7.2}$$

where again ℓ is a Bregman divergence.

- Low input scaling regime: Assuming that the input scaling factor θ goes to zero, we show that the derivative of the input–output mutual information over the binomial and negative binomial models is proportional to the input statistics $\mathsf{E}[X \log X] - \mathsf{E}[X] \log \mathsf{E}[X]$ which corresponds to the Bregman divergence ℓ_P between X and its expectation $\mathsf{E}[X]$. When we make the mean of each model equal to θX[1], we prove that the low input scaling regime is equal for the binomial and negative binomial models. Similar results are obtained for the relative entropy, where the low input scaling regime is characterized by the expression $\ell_P(\mathsf{E}[X], \mathsf{E}_Q[X])$.

- Arbitrary input preprocessing: Using an arbitrary input preprocessing we prove that several scenarios lead to information–estimation expressions that are given through Bregman divergences even though this is not always the case. In those cases where the information–estimation relationship is given through a Bregman divergence, we

[1]To obtain a binomial model with mean θX the required input preprocessing function is $\frac{\theta X}{n}$. In the case of the negative binomial model, to obtain a model with mean θX the input preprocessing function is $\frac{\theta X}{r}$.

can translate to the considered scenario those properties shown in the "I-MMSE" and "I-MMLE" relationships. For those models and scenarios over which the mutual information has a monotone behavior in θ, in the broadcast channel setting, lead to the class of broadcast channels known as "More Capable" channels which have a complete characterized capacity region.

One question that naturally arises along the content presented in Chapters 1 to 3 is whether there is a relationship between the Bregman divergence used in the exponential form of the conditional of the model and the Bregman divergence used to express the information–estimation relationships. Based on the expressions found in the arbitrary input preprocessing case we show that, up to a scaling factor, the input–output mutual information is upper (resp. lower) bounded by the expectation of the Bregman divergence used in the exponential representation of the binomial (resp. negative binomial) model. Similar bounds are obtained for the relative entropy.

- Extension to the Poisson models: It is well known that, over a binomial model where the mean of the model is independent of the number of trials n and the probability of each Bernoulli trial tends to zero, as long as n goes to infinity, the binomial model converges to the Poisson model. This property lets us extend the information–estimation relationships found for the Binomial model to the Poisson model.

  On the other hand, we provide conditions over the negative binomial model that let us obtain asymptotically the same results obtained for the Poisson model. This claim let us see the duality between the binomial and negative binomial models.

  At low input scaling regime, we prove that, under mild conditions, the Poisson model behavior is equal to that found for the binomial and negative binomial models.

7.2 Ongoing Work

This section is a composition, mainly, of a set of results that constitute an ongoing work. In the first case we show an alternative and as we believe novel approach to the "I-MMSE" relationship that depends heavily on the fact that the Gaussian distribution is a solution to the Heat equation. Then, we show several relationships between lautum information [43] and estimation

that can be derived using the same techniques employed in Chapters 4, 5 and 6.

7.2.1 Information–estimation relationships through partial differential equations

Let X and Y be the input and output of a random transformation where the input output relationship of the model considered is given by,

$$Y = X + N \tag{7.3}$$

where N stands for a random variable that is Gaussian distributed with zero mean and variance γ. Based on (7.3) the conditional distribution of Y given X is a Gaussian distribution with mean X and variance γ, *i.e.*,

$$P_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi\gamma}} e^{-\frac{1}{2\gamma}(y-x)^2}. \tag{7.4}$$

In the following theorem we repeat the information-estimation relationship between the mutual information and the minimum mean square error (MMSE) due to Guo *et al.* in [23].

Theorem 34. *Let X be a random variable such that $\mathsf{E}[X]^2 < \infty$. Then, for the Gaussian model given in* (7.3)*,*

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} I(X;Y) = -\frac{1}{2\gamma^2} \mathsf{E}\left[(X - \mathsf{E}[X|Y])^2\right]. \tag{7.5}$$

Proof. The mutual information can be written as

$$I(X;Y) = D(P_{Y|X}||P_N|P_X) - D(P_Y||P_N) \tag{7.6}$$

where P_N stands for a standard Gaussian distribution with zero mean and variance γ. The first term on the RHS of (7.6) is

$$\begin{aligned} D(P_{Y|X}||P_N|P_X) &= \mathsf{E}\left[\mathsf{E}\left[\log e^{\frac{YX}{\gamma} - \frac{1}{2}\frac{X^2}{\gamma}} \middle| X\right]\right] \\ &= \mathsf{E}\left[\frac{YX}{\gamma} - \frac{1}{2}\frac{X^2}{\gamma}\right] \\ &= \frac{1}{2\gamma}\mathsf{E}\left[X^2\right]. \end{aligned} \tag{7.7}$$

The derivative of this term with respect to γ is given by,

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} D(P_{Y|X}||P_N|P_X) = -\frac{1}{2\gamma^2}\mathsf{E}\left[X^2\right]. \tag{7.8}$$

Now we proceed to calculate the derivative with respect to γ of the second term at the RHS of (7.6). Taking into account that the Gaussian distribution is a solution of a particular case of the Heat equation we have that [67],

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} P_{Y|X}(y|x) = \frac{1}{2}\frac{\partial^2}{\partial y^2} P_{Y|X}(y|x) \tag{7.9}$$

which implies that [11, eq. (12)]

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} P_Y(y) = \frac{1}{2}\frac{\partial^2}{\partial y^2} P_Y(y). \tag{7.10}$$

Therefore,

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} D(P_Y||P_N) = \int \frac{\mathrm{d}}{\mathrm{d}\gamma} P_Y(y) \log \frac{P_Y(y)}{P_N(y)} \mathrm{d}y - \int \frac{\mathrm{d}}{\mathrm{d}\gamma} P_N(y) \frac{P_Y(y)}{P_N(y)} \mathrm{d}y \tag{7.11}$$

$$= \underbrace{\frac{1}{2}\int \frac{\partial^2 P_Y(y)}{\partial y^2} \log \frac{P_Y(y)}{P_N(y)} \mathrm{d}y}_{(\mathbb{I})} - \underbrace{\frac{1}{2}\int \frac{\partial^2 P_N(y)}{\partial y^2} \frac{P_Y(y)}{P_N(y)} \mathrm{d}y}_{(\mathbb{III})}. \tag{7.12}$$

where, to obtain (7.12) we use (7.10). To solve ($\mathbb{I}$), we integrate by parts over y, to obtain,

$$\frac{1}{2}\int \frac{\partial^2 P_Y(y)}{\partial y^2} \log \frac{P_Y(y)}{P_N(y)} \mathrm{d}y = \frac{1}{2}\frac{\partial P_Y(y)}{\partial y} \log \frac{P_Y(y)}{P_N(y)}\Big|_{-\infty}^{\infty} \tag{7.13}$$

$$= 0 - \frac{1}{2}\int \frac{\partial P_Y(y)}{\partial y} \frac{\partial}{\partial y}\left(\log \frac{P_Y(y)}{P_N(y)}\right) \mathrm{d}y \tag{7.14}$$

where the last step is justified by [7, eq. (9)]. Furthermore, we have that,

$$
\begin{aligned}
&-\frac{1}{2}\int \frac{\partial P_Y(y)}{\partial y}\frac{\partial}{\partial y}\left(\log\frac{P_Y(y)}{P_N(y)}\right)\mathrm{d}y \\
&= -\frac{1}{2}\int \left(\frac{\partial P_Y(y)}{\partial y}\right)^2 \frac{1}{P_Y(y)}\mathrm{d}y + \frac{1}{2}\int \frac{\partial P_Y(y)}{\partial y}\frac{\partial P_N(y)}{\partial y}\frac{1}{P_N(y)}\mathrm{d}y && (7.15)\\
&= \frac{1}{2}\int \frac{\partial P_Y(y)}{\partial y}\left(\frac{y}{\gamma} - \frac{\mathsf{E}[X|Y=y]}{\gamma}\right)\mathrm{d}y - \frac{1}{2}\int \frac{y}{\gamma}\frac{\partial P_Y(y)}{\partial y}\mathrm{d}y && (7.16)\\
&= \frac{1}{2}\int \left(\frac{y}{\gamma} - \frac{\mathsf{E}[X|Y=y]}{\gamma}\right)\frac{\mathsf{E}[X|Y=y]}{\gamma}P_Y(y)\mathrm{d}y && (7.17)\\
&= \frac{1}{2\gamma^2}\int x\int y P_{Y|X}(y|x)\mathrm{d}y P_X(x)\mathrm{d}x - \frac{1}{2\gamma^2}\mathsf{E}\left[\mathsf{E}[X|Y=y]^2\right] && (7.18)\\
&= \frac{1}{2\gamma^2}\mathsf{E}\left[(X-\mathsf{E}[X|Y])^2\right], && (7.19)
\end{aligned}
$$

where in (7.16) and (7.17) we use Lemma 20, proven at the end of this section.

Term labeled as ($\mathbb{III}$) is calculated similarly,

$$
\begin{aligned}
\frac{1}{2}\int \frac{\partial^2 P_N(y)}{\partial y^2}\frac{P_Y(y)}{P_N(y)}\mathrm{d}y &= -\frac{1}{2\gamma}\int \frac{\partial(yP_N(y))}{\partial y}\frac{P_Y(y)}{P_N(y)}\mathrm{d}y && (7.20)\\
&= -\frac{1}{2\gamma} + \frac{1}{2\gamma^2}\int y^2 P_Y(y)\mathrm{d}y && (7.21)\\
&= -\frac{1}{2\gamma} + \frac{1}{2\gamma^2}(\gamma + \mathsf{E}[X^2]) && (7.22)\\
&= \frac{\mathsf{E}[X^2]}{2\gamma^2} && (7.23)
\end{aligned}
$$

where to obtain (7.20) and (7.21) we use Lemma 20. Putting together (7.19) and (7.23) yields

$$
\frac{\mathrm{d}}{\mathrm{d}\gamma}D(P_Y||P_N) = -\frac{1}{2\gamma^2}\mathsf{E}\left[\mathsf{E}[X|Y]^2\right] \tag{7.24}
$$

Finally, based on the expressions given in (7.8) and (7.24) we get that

$$
\frac{\mathrm{d}}{\mathrm{d}\gamma}I(X;Y) = -\frac{1}{2\gamma^2}\mathsf{E}\left[(X-\mathsf{E}[X|Y])^2\right]. \tag{7.25}
$$

□

Notice that the two main ingredients in the alternative proof of Theorem 34 are the Heat Equation and Lemma 20, which characterizes the derivative of the marginal P_Y with respect to y. A striking question to ask is whether there is a similar information–estimation relationship (in terms of the square distance) for other channel laws $P_{Y|X}$ that satisfy the Heat Equation. Clearly, any deviation from (7.5) would be due to the fact that P_Y depends on the particular choice of $P_{Y|X}$ which in turn enters the picture via Lemma 20.

Lemma 20. *Let Y be a random variable obtained at the output of the Gaussian channel described by* (7.3)*. Then,*

$$\frac{\partial P_Y(y)}{\partial y} = -\frac{y}{\gamma} P_Y(y) + \frac{1}{\gamma} \mathsf{E}[X|Y=y] P_Y(y) \tag{7.26}$$

Proof. Based on [7, p. 269] we have that

$$\frac{\partial P_Y(y)}{\partial y} = -\frac{1}{\gamma} \mathsf{E}[(y-X) P_{Y|X}(y|X)] \tag{7.27}$$

$$= -\frac{y}{\gamma} P_Y(y) + \mathsf{E}[X|Y=y] P_Y(y) \tag{7.28}$$

□

7.2.2 Lautum Information and Estimation Theory

In this section we show expressions relating the lautum information [43] with some estimation quantities over the Poisson and binomial models.

Introduced in [43] by Palomar and Verdú as an alternative measure of dependency, the lautum information between two random variables X and Y is defined as follows,

$$L(X;Y) \triangleq D(P_X P_Y || P_{XY}) \tag{7.29}$$

$$= \mathsf{E}_{P_X P_Y}\left[\log \frac{P_X(\bar{X}) P_Y(\bar{Y})}{P_{XY}(\bar{X},\bar{Y})}\right] \tag{7.30}$$

where the roles of the joint P_{XY} and the product of marginals distributions $P_X P_Y$ are swapped in comparison with the definition of the mutual information. Note that in the definition given in (7.29), variables $(\bar{X},\bar{Y})$ are independent with the same marginals as (X,Y). In [43], the authors also present several properties and characterizations satisfied by the lautum information. There, it is shown that the lautum information appears in

problems such as test of independence, capacity per unit cost over certain channels [60], gambling strategies, etc. Moreover [2] gives an information-estimation for the lautum information in the context of the Gaussian channel.

Lets consider a Poisson model where, conditioned on X, Y is distributed as a Poisson with parameter θX. Recall from (1.13) that the derivative of the input–output mutual information is given by,

$$\frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) = \mathsf{E}\left[X \log \frac{X}{\mathsf{E}[X|Y]}\right]. \tag{7.31}$$

Employing a similar procedure to that used in Chapter 6 (Theorem 32), we find that the derivative of the lautum information for the Poisson model is characterized as follows.

Theorem 35. *Let $X \sim P_X$ be a positive bounded random variable. Let Y be the output of a Poisson model with mean $\theta X > 0$. Then,*

$$\frac{\mathrm{d}}{\mathrm{d}\theta} L(X;Y) = \mathsf{E}\left[X \log \frac{\mathsf{E}[X|Y]}{\mathsf{E}[\log X]}\right]. \tag{7.32}$$

Proof. Omitted. See Section 6.5.5 for further details. □

Algebraic manipulations over (7.31) and (7.32) let us obtain the following expression for the sum of the derivative of the mutual information and the derivative of the lautum information. This quantity was previously studied in the context of the Gaussian channel in [43, 2]. We present the analogous version for the Poisson model for the sake of completeness.

Corollary 7. *Assume the same conditions used in Theorem 35. Then,*

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\left(I(X;Y) + L(X;Y)\right) = \mathsf{E}\left[X \log \frac{X}{\mathsf{E}[\log X]}\right]. \tag{7.33}$$

Proof. Adding up (7.31) and (7.32) let us obtain the desired result. □

Let us next consider the binomial model given in (4.2). Specifically, for a given X, let Y be binomial distributed with parameters $(n, \theta X)$. Then by Theorem 9,

$$\frac{\mathrm{d}}{\mathrm{d}\theta} I(X;Y) = n\mathsf{E}^{n-1}\left[X \log \frac{X(1-\theta\mathsf{E}^{n-1}[X|Y])}{(1-\theta X)\mathsf{E}^{n-1}[X|Y]}\right]. \tag{7.34}$$

Following the procedure illustrated in the proof of Theorem 8 we obtain the following expression for the derivative of the lautum information with respect to θ.

Theorem 36. *Let $X \sim P_X$ be a random variable taking its value in $(0, x_{\max})$. Let Y be the output of the n-th order binomial model described by (4.2) with $X_\theta = \theta X$. Then,*

$$\frac{\mathrm{d}}{\mathrm{d}\theta} L(X;Y) = n\mathsf{E}^{n-1}\left[X \log \frac{\mathsf{E}^{n-1}[X|Y]}{(1-\theta\mathsf{E}^{n-1}[X|Y])}\right] - n\mathsf{E}[X]\mathsf{E}\left[\log \frac{X}{1-\theta X} + 1\right] + n\mathsf{E}\left[\frac{X}{1-\theta X}\right]\mathsf{E}[1-\theta X]. \tag{7.35}$$

Proof. Omitted. See Section 4.4.5 for further details. □

The derivative of the sum between the mutual information and the lautum information is expressed as follows.

Corollary 8. *Assume the same set of conditions used in Theorem 36. Then,*

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\left(I(X;Y) + L(X;Y)\right) = n\left[X \log \frac{X}{1-\theta X}\right] - n\mathsf{E}[X]\mathsf{E}\left[\log \frac{X}{1-\theta X} + 1\right] + n\mathsf{E}\left[\frac{X}{1-\theta X}\right]\mathsf{E}[1-\theta X]. \tag{7.36}$$

Proof. Adding up (7.35) and (7.36) let us obtain the desired result. □

Additionally, based on the procedures illustrated in the proof of Theorem 17, similar extensions can be carried out in the context of the negative binomial model.

So far is unknown whether the expressions given in (7.33) and (7.36) have an operational or conceptual meaning in the context of the Poisson and binomial models as was given in [2] for the Gaussian channel.

Bibliography

[1] D. Applebaum. *Lévy processes and stochastic calculus*. Cambridge university press, 2009.

[2] H. Asnani, K. Venkat, and T. Weissman. Relations between information and estimation in the presence of feedback. In Giacomo Como, Bo Bernhardsson, and Anders Rantzer, editors, *Information and Control in Networks*, volume 450 of *Lecture Notes in Control and Information Sciences*, pages 157–175. Springer International Publishing, 2014.

[3] R. Atar and T. Weissman. Mutual information, relative entropy, and estimation in the Poisson channel. *IEEE Trans. on Inf. Theory*, 58(3):1302–1319, March 2012.

[4] A. Banerjee, X. Guo, and H. Wang. On the optimality of conditional expectation as a Bregman predictor. *IEEE Trans. on Inf. Theory*, 51(7):2664 –2669, july 2005.

[5] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *J. Mach. Learn. Res.*, 6:1705–1749, December 2005.

[6] Robert Gardner Bartle. *A Modern Theory of Integration*, volume 32. American Mathematical Society, 1 edition, 2001.

[7] N. M. Blachman. The convolution inequality for entropy powers. *IEEE Trans. on Inf. Theory*, 11(2):267–271, Apr 1965.

[8] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2009.

[9] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex

programming. *Ussr Computational Mathematics and Mathematical Physics*, 7:200–217, 1967.

[10] R. Bustin, L. Ruoheng, and H. V. Poor. An MMSE approach to the secrecy capacity of the MIMO Gaussian wiretap channel. *in Proc. IEEE Int. Symp. Inf. Theory*, pages 2602–2606, June 2009.

[11] M. H. M. Costa. A new entropy power inequality. *IEEE Trans. on Inf. Theory*, 31(6):751–760, Nov 1985.

[12] T. Cover. Broadcast channels. *IEEE Trans. on Inf. Theory*, 18(1):2–14, Jan 1972.

[13] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications, 1st. edition, 1991.

[14] T. E. Duncan. Evaluation of likelihood functions. *Information and Control*, 13(1):62 – 74, 1968.

[15] T. E. Duncan. On the calculation of mutual information. *SIAM Journal on Applied Mathematics*, 19(1):pp. 215–220, 1970.

[16] R. Durrett. *Probability: Theory and examples*. Cengage Learning, 3 edition, 2011.

[17] I. A. Elbakri and J. A. Fessler. Statistical image reconstruction for polyenergetic x-ray computed tomography. *IEEE Trans on Medical Imaging*, 21(2):89–99, Feb 2002.

[18] B. A. Frigyik, S. Srivastava, and M. R. Gupta. Functional Bregman divergence and Bayesian estimation of distributions. *IEEE Trans. on Inf. Theory*, 54(11):5130–5139, Nov 2008.

[19] A. E. Gamal. The capacity of a class of broadcast channels. *IEEE Trans. on Inf. Theory*, 25(2):166–169, Mar 1979.

[20] A. E. Gamal and Young-Han Kim. *Network Information Theory*. Cambridge University Press, New York, NY, USA, 2012.

[21] D. Guo. On information-estimation relationships over binomial and negative binomial models. In *Proc. IEEE Int. Symp. on Inf. Theory*, pages 459–463, July 2013.

[22] D. Guo, S. Shamai, and S. Verdú. Additive non-Gaussian noise channels: mutual information and conditional mean estimation. In *Proc. IEEE Int. Symp. on Inf. Theory*, pages 719–723, Sept 2005.

[23] D. Guo, S. Shamai, and S. Verdú. Mutual information and minimum mean-square error in Gaussian channels. *IEEE Trans. on Inf. Theory*, 51(4):1261–1282, Apr. 2005.

[24] D. Guo, S. Shamai, and S. Verdú. Proof of entropy power inequalities via MMSE. In *Proc. IEEE Int. Symp. Inf. Theory*, pages 1011–1015, July 2006.

[25] D. Guo, S. Shamai, and S. Verdú. Mutual information and conditional mean estimation in Poisson channels. *IEEE Trans. on Inf. Theory*, 54(5):1837–1849, May. 2008.

[26] D. Guo and S. Verdú. Randomly spread CDMA: asymptotics via statistical physics. *IEEE Trans. on Inf. Theory*, 51(6):1983–2010, June 2005.

[27] R. Iyer and J. A. Bilmes. Submodular-Bregman and the Lovász-Bregman divergences with applications. In *NIPS*, pages 2942–2950, 2012.

[28] J. Jiao, K. Venkat, and T. Weissman. Relations between information and estimation in scalar Lévy channels. *CoRR*, abs/1404.6812, 2014.

[29] J. Jiao, K. Venkat, and T. Weissman. Relations between information and estimation in scalar Lévy channels. In *Proc. IEEE Int. Symp. Inf. Theory*, pages 2212–2216, July 2014.

[30] T. Kadota, M. Zakai, and J. Ziv. Mutual information of the white Gaussian channel with and without feedback. *IEEE Trans. on Inf. Theory*, 17(4):368–371, Jul 1971.

[31] S. M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall PTR, Fundamentals of Statistical Signal Processing, 1993.

[32] S. Ken-Iti. *Lévy Processes and Infinitely Divisible Distributions*. Cambridge University Press, 1999.

[33] A. E. Kyprianou. *Introductory lectures on fluctuations of Lévy processes with applications*. Springer, 2006.

[34] A. Laourine and A. B. Wagner. The degraded Poisson wiretap channel. *IEEE Trans. on Inf. Theory*, 58(12):7073–7085, Dec 2012.

[35] A. Lapidoth, J. H. Shapiro, V. Venkatesan, and L. Wang. The discrete-time Poisson channel at low input powers. *IEEE Trans. on Inf. Theory*, 57(6):3260–3272, June 2011.

[36] A. Lapidoth, I. E. Telatar, and R. Urbanke. On wide-band broadcast channels. *IEEE Trans. on Inf. Theory*, 49(12):3250–3258, Dec 2003.

[37] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet Physics-Doklady*, volume 10, 1966.

[38] R. S. Liptser. Optimal encoding and decoding for transmission of a Gaussian Markov signal in a noiseless-feedback channel. *Problemy Peredachi Informatsii*, 10(4):3–15, 1974.

[39] A. Lozano, A. M. Tulino, and S. Verdú. Optimum power allocation for parallel Gaussian channels with arbitrary input distributions. *IEEE Trans. on Inf. Theory*, 52(7):3033–3051, July 2006.

[40] E. Mayer-Wolf and M. Zakai. On a formula relating the Shannon information to the Fisher information for the filtering problem. 61:164–171, 1984.

[41] F. Nielsen, J. Boissonnat, and R. Nock. Bregman Voronoi diagrams: Properties, algorithms and applications. *CoRR*, abs/0709.2196, 2007.

[42] D. P. Palomar and S. Verdú. Representation of mutual information via input estimates. *IEEE Trans. on Inf. Theory*, 53(2):453–470, Feb 2007.

[43] D. P. Palomar and S. Verdú. Lautum information. *IEEE Trans. on Inf. Theory*, 54(3):964–975, March 2008.

[44] F. Pérez-Cruz, M. R. D. Rodrigues, and S. Verdú. MIMO Gaussian channels with arbitrary inputs: Optimal precoding and power allocation. *IEEE Trans. on Inf. Theory*, 56(3):1070–1084, March 2010.

[45] M. Raginsky and T. P. Coleman. Mutual information and posterior estimates in channels of exponential family type. In *IEEE Information Theory Workshop, 2009*, pages 399–403, Oct 2009.

[46] R. T. Rockafellar. *Convex analysis*, volume 28. Princeton university press, 1997.

[47] W. Rudin. *Principles of Mathematical Analysis*. International series in pure and applied mathematics. McGraw-Hill, 1964.

[48] R. Santos-Rodríguez, A. Guerrero-Curieses, R. Alaiz-Rodríguez, and J. Cid-Sueiro. Cost-sensitive learning based on Bregman divergences. *Machine Learning*, 76(2-3):271–285, 2009.

[49] S. Shamai. Capacity of a pulse amplitude modulated direct detection photon channel. *IEEE Proceedings in Communications, Speech and Vision*, 137(6):424–430, Dec 1990.

[50] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, July, October 1948.

[51] C. E. Shannon. Communication in the presence of noise. *Proc. IRE*, 37:10–21, 1949.

[52] C. E. Shannon. Two-way communication channels. In *Proc. 4th Berkeley Symp. Math. Stat. Prob*, volume 1, pages 611–644. USA, 1961.

[53] D. Stirzaker. *Elementary Probability*. Cambridge University Press, 2003.

[54] C. G. Taborda, D. Guo, and F. Pérez-Cruz. Information-estimation relationships over binomial and negative binomial models. *IEEE Trans. on Inf. Theory*, 60(5):2630–2646, May 2014.

[55] C. G. Taborda and F. Pérez-Cruz. Derivative of the relative entropy over the Poisson and binomial model. In *Proc. IEEE Inf. Theory Workshop*, pages 386–390, September 2012.

[56] C. G. Taborda and F. Pérez-Cruz. Mutual information and relative entropy over binomial and negative binomial models. In *Proc. IEEE Int. Symp. on Inf. Theory*, pages 696–700, July 2012.

[57] C. G. Taborda, F. Pérez-Cruz, and D. Guo. New information–estimation relationships over binomial, negative binomial and Poisson models. In *Proc. IEEE Int. Symp. on Inf. Theory*, pages 2207–2211, July 2014.

[58] D. N. C. Tse and R. D. Yates. Fading broadcast channels with state information at the receivers. *IEEE Trans. on Inf. Theory*, 58(6):3453–3471, June 2012.

[59] A. M. Tulino and S. Verdú. Monotonic decrease of the non-Gaussianness of the sum of independent random variables: A simple proof. *IEEE Trans. on Inf. Theory*, 52(9):4295–4297, Sept 2006.

[60] S. Verdú. On channel capacity per unit cost. *IEEE Trans. on Inf. Theory*, 36(5):1019–1030, Sep 1990.

[61] S. Verdú. Spectral efficiency in the wideband regime. *IEEE Trans. on Inf. Theory*, 48(6):1319–1343, Jun 2002.

[62] S. Verdú. Mismatched estimation and relative entropy. *IEEE Trans. on Inf. Theory*, 56(8):3712–3720, Aug. 2010.

[63] S. Verdú and D. Guo. A simple proof of the entropy-power inequality. *IEEE Trans. on Inf. Theory*, 52(5):2165–2166, May 2006.

[64] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

[65] L. Wang, D. E. Carlson, M. R. D. Rodrigues, R. Calderbank, and L. Carin. A Bregman matrix and the gradient of mutual information for vector Poisson and Gaussian channels. *IEEE Trans. on Inf. Theory*, 60(5):2611–2629, May 2014.

[66] T. Weissman. The relationship between causal and noncausal mismatched estimation in continuous-time AWGN channels. *IEEE Trans. on Inf. Theory*, 56(9):4256 –4273, Sept. 2010.

[67] D.V. Widder. *The Heat Equation*. Pure and Applied Mathematics, a Series of Monographs and Tex. Elsevier Science, 1976.

[68] L. Wu, S. C. H. Hoi, J. Rong, J. Zhu, and Y. Nenghai. Learning Bregman distance functions for semi-supervised clustering. *IEEE Transactions on Knowledge and Data Engineering*, 24(3):478–491, March 2012.

[69] Y. Wu, D. Guo, and S. Verdú. Derivative of mutual information at zero SNR: The Gaussian-noise case. *IEEE Trans. on Inf. Theory*, 57(11):7307–7312, Nov 2011.

[70] A. D. Wyner. Capacity and error exponent for the direct detection photon channel. I. *IEEE Trans. on Inf. Theory*, 34(6):1449–1461, Nov 1988.

[71] A. D. Wyner. Capacity and error exponent for the direct detection photon channel. II. *IEEE Trans. on Inf. Theory*, 34(6):1462–1471, Nov 1988.

[72] M. Zhou, L. A. Hannah, D. B. Dunson, and L. Carin. Beta-negative binomial process and Poisson factor analysis. In *Proc. Int. Conf. Artificial Intelligence Statistics, AISTATS*, 2012.

[73] Y. Zhu and D. Guo. Ergodic fading one-sided interference channels without state information at transmitters. *CoRR*, abs/0911.1082, 2009.