ON SIGNS OF DERIVATIVES OF ENTROPY ALONG MARKOV SEMI-GROUPS: A COLLECTION OF SOME RECENT RESULTS

CHANDRA NAIR

Abstract. Some results on the signs of the derivative of entropy along Markov semi-groups is presented. In part, the results are motivated by McKean's conjecture about Gaussian optimality of the derivatives and consequent alternating signs of them along the heat semi-group. We establish log-convexity of Fisher information along the heat semi-group. We consider a particular (and natural) family of discrete semi-groups and show that the signs of the derivative of entropy do alternate for the first nine-derivatives but fail to do so for the tenth derivative. These results are obtained as a consequence of different collaborations by the author.

1. Introduction

1.1. **Background.** Let X be a random variable with a finite variance. Let μ_{X_t} denote the density function of $X_t := X + \sqrt{t}Z$, where $Z \sim \mathcal{N}(0,1)$ is the standard Gaussian and is assumed to be independent of X. One can view this operation as a family of Markov operators on densities defined by

$$W_t(\mu) = \nu_t * \mu,$$

where ν_t denote the density function of $\sqrt{t}Z$. Then this family of operators satisfy a semi-group property, i.e.

$$W_{t_2}(W_{t_1}(\mu)) = W_{t_2+t_1}(\mu).$$

This is called the heat semi-group in literature since it arises as the solution to the heat equation.

Let $g_X^{(k)}(t) := \frac{d^k}{dt^k} h(\mu_{X_t})$, where

$$h(\mu_{X_t}) = -\int_{\mathbb{R}} \mu_{X_t} \ln \mu_{X_t} dx,$$

refers to the differential entropy of X_t. Let G be a Gaussian random variable with the same variance as X. In Section 12 of [5], McKean observed that $g_G^{(0)}(t) \geq g_X^{(0)}(t) \geq 0$, $g_G^{(1)}(t) \leq g_X^{(1)}(t) \leq 0$, and $g_G^{(2)}(t) \geq g_X^{(2)}(t) \geq 0$.

Conjecture 1.1 (McKean [5])**.** *The following inequality*

$$(-1)^k g_G^{(k)}(t) \geq (-1)^k g_X^{(k)}(t) \geq 0$$

holds for every $k \geq 3$.

Definition 1.2. *A function $f : \mathbb{R} \to \mathbb{R}$ continuous on $[0,\infty)$ and infinitely differentiable on $(0,\infty)$ is said to be completely monotone if $(-1)^k \frac{d^k}{dt^k} f(t) \geq 0$ for any $t > 0$ and $k \in \mathbb{Z}^+$.*

Thus a weaker version of McKean's conjecture is that Fisher Information, $I_X(t) = \frac{1}{2}\frac{d}{dt} h(\mu_{X_t})$, is completely monotone along the heat semi-group.

Theorem 1.3 (Theorem A in [3])**.** *Let f be a completely monotone function on $[0,\infty)$. Then for each $t \in (0,\infty)$ and $0 < k < n$,*

$$(-1)^{nk} |f^{(k)}(t)|^n \leq (-1)^{nk} |f^{(n)}(t)|^k |f(t)|^{n-k}.$$

CUHK
E-mail address: chandra@ie.cuhk.edu.hk.
Date: 07/02/2020. *Last Update*: February 23, 2023.

In particular, taking $k = 1$ and $n = 2$ we obtain that f is log-convex with respect to t. In [1] the authors showed that $g_X^{(3)}(t) \geq 0$, $g_X^{(4)}(t) \leq 0$ using techniques in [7], which was in turn motivated by the calculations of Bakry. However, their proof techniques did not extend to higher derivatives. They also made a weaker conjecture (Conjecture 2 in [1]) that Fisher information, $I(\mu_t^X)$, is log-convex in t. We extend the ideas developed in [1] and [9], and in Theorem 3.3 show that Fisher-information is log-convex with respect to t. In Section 3 we study a similar problem along a discrete semi-group and establish that signs of the derivatives of entropy do alternate for the first nine derivatives but fail to do so for the tenth derivative (in general).

1.1.1. *Alterate Motivation.* The author's primary motivation for this series of work comes from the study of optimization problems of the following type, that occur often times in evaluation of achievable rate regions or outer bounds to the capacity regions in network information theory settings. Let $T_{Y|X}$ denote a channel that maps input distributions μ_X into output distributions $\mu_Y = T\mu_X$. If X and Y takes values in a finite alphabet space, then consider the problem[1] of computing the maximum, over μ_X, of

$$F_\lambda(\mu_X) := \lambda H(\mu_X) - H(T\mu_X),$$

where $H(\mu_X) = -\sum_{x\in\mathcal{X}} \mu_X(x) \log \mu_X(x)$ denotes the Shannon entropy of X; and $\lambda \geq 0$ is a fixed constant. When $\lambda \geq 1$, it is immediate from the data-processing inequality that the functional $F_\lambda(\mu_X)$ is concave in μ_X. However for $\lambda \in [0, 1)$, this is not necessarily true. In particular for $\lambda = 0$, $F_0(\mu_X)$ is convex in μ_X. Therefore, from a optimization perspective, computing the optimizers of $F_\lambda(\mu_X)$ becomes a non-convex optimization problem at least for some values of λ in the range $[0, 1)$.

When the channel $T_{Y|X}$ is the binary-symmetric-channel (BSC), say with crossover probability p, consider the following reparameterization of μ_X, defined by $\mu_X(0) = H_2^{-1}(u)$, where $H_2^{-1} : [0, 1] \mapsto [0, \frac{1}{2}]$ denotes the inverse binary entropy function. Under this reparameterization, for $BSC(p)$, observe that

$$F_\lambda(\mu_X) = \lambda u - H_2(p * H_2^{-1}(u)).$$

It was shown in [8] that $H_2(p * H_2^{-1}(u))$ is convex in u and hence $\lambda u - H_2(p * H_2^{-1}(u))$ is a concave function in u for any λ. Therefore this non-linear parameterization converted the non-convex optimization problem to a convex-optimization problem. It is also worth remarking that the convexity of $H_2(p * H_2^{-1}(u))$ was developed by Wyner and Ziv in the context of evaluating the superposition-coding region for a degraded binary symmetric broadcast channel.

Additive White Gaussian Noise channels are in many ways the continuous analogue of Binary Symmetric Channels. Therefore it is natural to see if there is an analogous result in the additive Gaussian noise setting, where under a suitable parameterization of μ_X, $h(\mu_X)$ - the differential entropy - becomes linear in the parameter and $h(T_G\mu_X)$ becomes convex in the parameter, where T_G refers to the Markov operator corresponding to the channel with additive Gaussian noise W.

For distributions on binary alphabets, there is only one degree of freedom and hence the parameterization of $\mu_X(0) = H_2^{-1}(u)$ is forced on us, if we wish to make $H_2(\mu_X)$ linear. In the continuous world we assume that μ_X evolves along the heat flow, i.e. $X_t := X + \sqrt{t}Z, t > 0$, where Z is the standard Gaussian and independent of X. Therefore we seek a parameterization $t = \phi(u)$ such that $h(X + \sqrt{\phi(u)}Z)$ is linear in u and investigate whether, the output entropy, $h(\mu_Y) = h(X + \sqrt{\phi(u)}Z + W)$ is convex in u, where W is some Gaussian independent of X and Z. Let μ_t^X denote the distribution of $X_t = X + \sqrt{t}Z$. A bit of algebra immediately shows that this question is equivalent to asking whether the Fisher information $I(\mu_t^X)$ is log-convex in t, for all random variables X (see Remark 2.2).

1.2. **Preliminaries.** Given a random variable X on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with values in $\mathbb{R}$, let the cumulative distribution function of X be $\tilde{F}(x) := Pr(X \leq x)$, $x \in \mathbb{R}$. For Z some independent standard Gaussian random variable with mean zero and variance one, consider $X_t := X + \sqrt{t}Z$, $t > 0$, with probability density function $f_t(x)$ with respect to the Lebesgue measure on $\mathbb{R}$. The density $f_t(x)$, $x \in \mathbb{R}$, can be written as

$$f_t(x) = \int_{\mathbb{R}} \frac{-z}{\sqrt{2\pi t}} e^{-\frac{z^2}{2t}} \tilde{F}(x - z)dz.$$

[1]This particular problem does arise in the context of computing certain strong data-processing constants as well as in the Ahlswede-Korner source coding setting.

It is well-known in literature that the probability density function $f_t(x)$ of X_t is always upper bounded by $1+t$, strictly positive and infinitely differentiable with respect to $x \in (-\infty, \infty)$ and $t \in (0, \infty)$, and satisfy that

$$\lim_{|x|\to\infty} \frac{\partial^n f_t(x)}{\partial x^n} = 0, \forall n \in \mathbb{Z}_+.$$

Besides, $f_t(x)$ also satisfies the heat equation, see, e.g., [6].

$$\frac{\partial}{\partial t} f_t(x) = \frac{1}{2} \frac{\partial^2}{\partial x^2} f_t(x). \tag{1}$$

The differential entropy of X_t, $h(X_t)$, $t > 0$, is defined as

$$h(X_t) = -\int_{\mathbb{R}} f_t(x) \ln f_t(x) dx.$$

When X has a finite variance P, $h(X_t)$ exists and is maximized by X following a Gaussian distribution with variance P. The Fisher information of X_t is defined as

$$I(\mu_t^X) := \int_{\mathbb{R}} \left(\frac{\partial}{\partial x} \ln f_t(x) \right)^2 f_t(x) dx.$$

One can verify that the Fisher information $I(\mu_t^X), t > 0$, always exists and is infinitively differentiable with respect to $t \in (0, \infty)$, see, e.g., [1]. The Fisher information $I(\mu_t^X)$ is closely related to the differential entropy of X_t via the de Bruijin's identity when X has a finite variance, see, e.g., [2]

$$\frac{\partial}{\partial t} h(X_t) = \frac{1}{2} I(\mu_t^X). \tag{2}$$

Notation: For convenience of writing, we will suppress the dependence on t and write $v(x) := \ln f_t(x), t > 0$, and $v_k(x) := \frac{\partial^k \ln f_t(x)}{\partial x^k}, k \in \mathbb{Z}_+$, i.e., $v_k(x)$ is the k-th derivative of v as a function of $x \in \mathbb{R}$. Well-definedness of $v_k(x)$ for any $k \in \mathbb{Z}_+$ follows from the known properties of $f_t(x)$.

Proposition 1.4 (Proposition 2 in [1]). *For any $r, m_i, k_i \in \mathbb{Z}_+$,*

$$\int_{\mathbb{R}} \left| \prod_{i=1}^{r} v_{k_i}^{m_i}(x) \right| f_t(x) dx < \infty,$$

and

$$\lim_{|x|\to\infty} \left| \prod_{i=1}^{r} v_{k_i}^{m_i}(x) \right| f_t(x) = 0.$$

We define $\langle \varphi \rangle := \int_{\mathbb{R}} \varphi f_t(x) dx$ to denote the integration with respect to the probability measure $f_t(x)$. Under this notation

$$I(\mu_t^X) = \langle v_1^2 \rangle. \tag{3}$$

The following, integration by part lemma, turns out to be useful in our proof.

Lemma 1.5 (Lemma 3 in [9]). *For $k \geq 2$, let $\varphi(x)$ be some function continuously differentiable with respect to x satisfying that $\lim_{|x|\to\infty} \varphi v_{k-1} f_t = 0$, then*

$$\langle \varphi v_k + \varphi v_1 v_{k-1} + \frac{\partial \varphi}{\partial x} v_{k-1} \rangle = 0.$$

One can see that this lemma follows from the basic integration by parts property.

Proof.

$$\begin{aligned}
&\langle \varphi v_k + \varphi v_1 v_{k-1} + \frac{\partial \varphi}{\partial x} v_{k-1} \rangle \\
=&\int_{\mathbb{R}} \left(\varphi v_k f_t + \varphi v_{k-1} \frac{\partial f_t}{\partial x} + \frac{\partial \varphi}{\partial x} v_{k-1} f_t \right) dx \\
\overset{(a)}{=}&\int_{\mathbb{R}} \left(\frac{\partial}{\partial x} \varphi v_{k-1} f_t \right) dx \\
=&\varphi v_{k-1} f_t |_{-\infty}^{\infty} \\
\overset{(b)}{=}&0.
\end{aligned}$$

Equality (a) follows from the integration by parts property, and equality (b) follows from the condition that $\lim_{|x|\to\infty} \varphi v_{k-1} f_t = 0$. □

Notice that by Proposition 1.4 we could choose φ in Lemma 1.5 to be in the form of $\prod_{i=1}^{r} v_{k_i}^{m_i}(x)$, where $r, m_i, k_i \in \mathbb{Z}_+$.

Lemma 1.6 ([1], [9])**.** *Let $\varphi(x)$ be some function continuously differentiable with respect to x satisfying that $\lim_{|x|\to\infty} \varphi v_1 f_t = 0$. For $k \geq 0$, the following hold:*

$$\begin{aligned}
\frac{\partial}{\partial t} v_k &= \frac{1}{2} \left(v_{k+2} + \sum_{i=0}^{k} \binom{k}{i} v_{i+1} v_{k-i+1} \right), \\
\frac{\partial}{\partial t} \langle \varphi \rangle &= \langle \frac{\partial}{\partial t} \varphi - \frac{1}{2} \frac{\partial \varphi}{\partial x} v_1 \rangle.
\end{aligned}$$

Proof. The proof idea is to interchange integral and derivatives by Proposition 1.4 and the Dominated Convergence Theorem, and the calculations follow from the following observations (for details, see Appendix A in [9]). We again present the outline here.

$$\begin{aligned}
2\frac{\partial}{\partial t} v_k &= 2\frac{\partial}{\partial t} \left(\frac{\partial^k}{\partial x^k} \ln f_t(x) \right) \\
&= 2\frac{\partial^k}{\partial x^k} \left(\frac{\partial}{\partial t} \ln f_t(x) \right) \\
&\overset{(a)}{=} \frac{\partial^k}{\partial x^k} \left(\frac{\frac{\partial^2}{\partial x^2} f_t(x)}{f_t(x)} \right) \\
&= \frac{\partial^k}{\partial x^k} \left(v_2 + v_1^2 \right) \\
&\overset{(b)}{=} v_{k+2} + \sum_{i=0}^{k} \binom{k}{i} v_{i+1} v_{k-i+1}.
\end{aligned}$$

Equality (a) is due to the heat equation (1) and (b) can be established by mathematical induction.

For the second part, observe that

$$\begin{aligned}
\frac{\partial}{\partial t} \langle \varphi \rangle &= \langle \frac{\partial}{\partial t} \varphi \rangle + \int_{\mathbb{R}} \varphi \frac{\partial f_t}{\partial t} dx \\
&\overset{(a)}{=} \langle \frac{\partial}{\partial t} \varphi \rangle + \frac{1}{2} \int_{\mathbb{R}} \varphi \frac{\partial^2 f_t}{\partial x^2} dx \\
&\overset{(b)}{=} \langle \frac{\partial}{\partial t} \varphi \rangle - \frac{1}{2} \int_{\mathbb{R}} \frac{\partial \varphi}{\partial x} \frac{\partial f_t}{\partial x} dx \\
&= \langle \frac{\partial}{\partial t} \varphi \rangle - \frac{1}{2} \langle \frac{\partial \varphi}{\partial x} v_1 \rangle.
\end{aligned}$$

Equality (a) is again due to the heat equation (1) and (b) follows from integration by parts and the assumption that $\lim_{|x|\to\infty} \varphi v_1 f_t = 0$. □

One can compute the derivatives of the Fisher information $I(\mu_t^X)$ with respect to t as following: see [4] and [9].

Lemma 1.7 ([1], [9])**.** *For $t > 0$, Fisher information $I(\mu_t^X)$ and its derivatives up to second order can be expressed as:*

$$\frac{d}{dt} I(\mu_t^X) = -\langle v_2^2 \rangle,$$
$$\frac{d^2}{dt^2} I(\mu_t^X) = \langle v_3^2 + 2v_1^2 v_2^2 + 4v_1 v_2 v_3 \rangle.$$

Proof. We outline the proof via applications of Lemmas 1.6 and 1.5. Observe that

$$\begin{aligned}\frac{d}{dt} I(\mu_t^X) &= \frac{d}{dt}\langle v_1^2 \rangle \\ &\overset{(a)}{=} \langle 2v_1 \frac{\partial v_1}{\partial t} - v_2 v_1^2 \rangle \\ &\overset{(b)}{=} \langle v_1(v_3 + 2v_1 v_2) - v_2 v_1^2 \rangle \\ &\overset{(c)}{=} -\langle v_2^2 \rangle.\end{aligned}$$

Here $(a), (b)$ follow from Lemma 1.6, and (c) follows from Lemma 1.5 by setting $\varphi = v_1$ and $k = 3$. Similarly, note that

$$\begin{aligned}\frac{d^2}{dt^2} I(\mu_t^X) &= -\frac{d}{dt}\langle v_2^2 \rangle \\ &\overset{(a)}{=} \langle -2v_2 \frac{\partial v_2}{\partial t} + v_2 v_3 v_1 \rangle \\ &\overset{(b)}{=} \langle -v_2(v_4 + 2v_1 v_3 + 2v_2^2) + v_2 v_3 v_1 \rangle \\ &\overset{(c)}{=} \langle v_3^2 - 2v_2^3 \rangle \\ &\overset{(c)}{=} \langle v_3^2 + 2v_1^2 v_2^2 + 4v_1 v_2 v_3 \rangle.\end{aligned}$$

Here $(a), (b)$ follow from Lemma 1.6, (c) follows from Lemma 1.5 by setting $\varphi = v_2$ and $k = 4$, and (c) follows from Lemma 1.5 by setting $\varphi = v_2^2$ and $k = 2$. □

Remark 1.8. *There are several equivalent ways of expressing $\frac{d^2}{dt^2} I(\mu_t^X)$ using Lemma 1.6. For instance, [9] expressed it as $\langle v_3^2 - 2v_2^3 \rangle$. We choose this particular representation, $\langle v_3^2 + 2v_1^2 v_2^2 + 4v_1 v_2 v_3 \rangle$, as it turns out to be useful to prove the log-convexity of Fisher information.*

2. Main

2.1. Log-convexity of Fisher Information.

Theorem 2.1. *[with Michel Ledoux and Yannan Dustin Wang] Let X be a random variable on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with values in $\mathbb{R}$, and Z some independent standard Gaussian random variable. Consider $X_t := X + \sqrt{t}Z, t > 0$, with probability density function $f_t(x)$ with respect to the Lebesgue measure on $\mathbb{R}$. The Fisher information of X_t is log-convex in t, i.e.*

$$\ln I(\mu_t^X) = \ln \int_{\mathbb{R}} \left(\frac{\partial}{\partial t} \ln f_t(x) \right)^2 f_t(x) dx$$

is convex in t.

Proof. Log-convexity of Fisher information is equivalent to showing

$$\left(\frac{d}{dt} I(\mu_t^X) \right)^2 \leq I(\mu_t^X) \frac{d^2}{dt^2} I(\mu_t^X).$$

Using Lemma 1.7, this is equivalent to showing

$$\langle v_2^2 \rangle^2 \leq \langle v_1^2 \rangle \langle v_3^2 + 2v_1^2 v_2^2 + 4v_1 v_2 v_3 \rangle. \tag{4}$$

In Lemma 1.5, the choices that $k = 2, \varphi = v_2$ and that $k = 2, \varphi = v_1^2$ will lead to the following two equalities respectively

$$\langle v_2^2 + v_1^2 v_2 + v_1 v_3 \rangle = 0 \tag{5}$$

$$\langle v_1^4 + 3v_1^2 v_2 \rangle = 0. \tag{6}$$

Consequently, for any $\alpha \in \mathbb{R}$ we have

$$\langle v_2^2 \rangle = -\langle v_1 (v_3 + \alpha v_1 v_2 - \frac{1-\alpha}{3} v_1^3) \rangle.$$

The Cauchy-Schwarz inequality yields,

$$\langle v_2^2 \rangle^2 \leq \langle v_1^2 \rangle \langle (v_3 + \alpha v_1 v_2 - \frac{1-\alpha}{3} v_1^3)^2 \rangle.$$

Thus to show inequality (4), it suffices to show that

$$\langle (v_3 + \alpha v_1 v_2 - \frac{1-\alpha}{3} v_1^3)^2 \rangle \leq \langle v_3^2 + 2v_1^2 v_2^2 + 4 v_1 v_2 v_3 \rangle \tag{7}$$

holds for some $\alpha \in \mathbb{R}$. Expanding, (7) is equivalent to

$$\begin{aligned} \langle (2-\alpha^2) v_1^2 v_2^2 + (4 - 2\alpha) v_1 v_2 v_3 - \frac{1}{9}(1-\alpha)^2 v_1^6 \\ + \frac{2}{3}(1-\alpha) v_1^3 v_3 + \frac{2}{3}\alpha(1-\alpha) v_1^4 v_2 \rangle \geq 0. \end{aligned}$$

In Lemma 1.5, the choices that $k = 3, \varphi = v_1^3$ and that $k = 2, \varphi = v_1^4$ will lead to the following two equalities respectively.

$$\begin{aligned} \langle v_1^3 v_3 + v_2 v_1^4 + 3 v_1^2 v_2^2 \rangle = 0 \\ \langle v_1^6 + 5 v_1^4 v_2 \rangle = 0. \end{aligned}$$

Thus proving inequality (7) for some $\alpha \in \mathbb{R}$ is equivalent to proving the following inequality

$$\begin{aligned} \langle (2-\alpha^2) v_1^2 v_2^2 + (4 - 2\alpha) v_1 v_2 v_3 - \frac{1}{9}(1-\alpha)^2 v_1^6 \\ + \frac{2}{3}(1-\alpha) v_1^3 v_3 + \frac{2}{3}\alpha(1-\alpha) v_1^4 v_2 \rangle \\ + \beta \langle v_1^3 v_3 + v_2 v_1^4 + 3 v_1^2 v_2^2 \rangle + \gamma \langle v_1^6 + 5 v_1^4 v_2 \rangle \geq 0 \end{aligned} \tag{8}$$

for some $\alpha, \beta, \gamma \in \mathbb{R}$.

We successively choose the values α, β, γ to eliminate the terms whose signs are not clear: first set $\alpha = 2$ to get rid of $\langle v_1 v_2 v_3 \rangle$, then $\beta = \frac{2}{3}$ to eliminate $\langle v_1^3 v_3 \rangle$, and finally $\gamma = \frac{2}{15}$ to handle $\langle v_1^4 v_2 \rangle$. With these choices, the above inequality (8) reduces to $\frac{1}{45}\langle v_1^6 \rangle \geq 0$, which holds trivially. □

Remark 2.2. *Let $\phi(u)$, with $\phi(0) = 0$ and $\phi(1) = 1$, be the uniquely defined increasing function of u such that $h(X + \sqrt{\phi(u)} Z)$ is linear in u. Then we have*

$$\begin{aligned} 0 &= \frac{d^2}{du^2} h(X + \sqrt{\phi(u)} Z) \\ &= \frac{1}{2} \left(\frac{d^2 \phi(u)}{du^2} I(\mu^X_{\phi(u)}) + \left(\frac{d\phi(u)}{du} \right)^2 \frac{d}{d\phi(u)} I(\mu^X_{\phi(u)}) \right). \end{aligned}$$

Now, showing that $\frac{d^2}{du^2} h(X + \sqrt{\phi(u)} Z + W) \geq 0$, for $W \sim \mathcal{N}(0, \sigma^2)$ independent of (X, Z), is equivalent to showing that

$$0 \leq \frac{1}{2} \left(\frac{d^2 \phi(u)}{du^2} I(\mu^{X+W}_{\phi(u)}) + \left(\frac{d\phi(u)}{du} \right)^2 \frac{d}{d\phi(u)} I(\mu^{X+W}_{\phi(u)}) \right).$$

This can be rewritten using the equality above as requiring

$$\frac{\frac{d}{d\phi(u)} I(\mu^{X+W}_{\phi(u)})}{I(\mu^{X+W}_{\phi(u)})} \geq \frac{\frac{d}{d\phi(u)} I(\mu^{X}_{\phi(u)})}{I(\mu^{X}_{\phi(u)})}.$$

Since $I(\mu^{X+W}_{\phi(u)}) = I(\mu^{X}_{\phi(u_1)})$ for some $u_1 \geq u$, the above inequality is equivalent to showing that

$$\frac{\frac{d}{dt}I(\mu_t^X)}{I(\mu_t^X)}$$

is increasing in t or equivalently, that $\log I(\mu_t^X)$ is convex in t. Thus, the result we showed can be considered as a continuous analogue of the convexity result for BSC established by Wyner and Ziv.

2.1.1. *A conditional version of log-convexity.* Given two jointly distributed random variables (X, W), such that the conditional distributions (and densities) are well defined, one can define

$$I_w(\mu_t^{X|W}) := \int_{\mathbb{R}} \left(\frac{\partial}{\partial x} \ln f_t(x|w)\right)^2 f_t(x|w)dx.$$

Further, conditional Fisher information, is traditionally defined as

$$I(\mu_t^{X|W}) := \mathrm{E}(I_W(\mu_t^{X|W})) = \int_{\mathbb{R}^2} \left(\frac{\partial}{\partial x} \ln f_t(x|w)\right)^2 f_t(x|w) f_W(w)dxdw.$$

Note that Theorem 3.3 implies that

$$\left(\frac{d}{dt}I_w(\mu_t^{X|W})\right)^2 \leq \left(\frac{d^2}{dt^2}I_w(\mu_t^{X|W})\right)\left(I_w(\mu_t^{X|W})\right).$$

Since $\frac{d}{dt}I_w(\mu_t^{X|W}) \leq 0$, we obtain

$$-\frac{d}{dt}I_w(\mu_t^{X|W}) \leq \sqrt{\left(\frac{d^2}{dt^2}I_w(\mu_t^{X|W})\right)}\sqrt{\left(I_w(\mu_t^{X|W})\right)}.$$

Taking expectation with respect to W, we obtain

$$\begin{aligned}-\frac{d}{dt}I(\mu_t^{X|W}) = -\mathrm{E}\left(\frac{d}{dt}I_w(\mu_t^{X|W})\right) &\leq \mathrm{E}\left(\sqrt{\left(\frac{d^2}{dt^2}I_w(\mu_t^{X|W})\right)}\sqrt{\left(I_w(\mu_t^{X|W})\right)}\right)\\ &\overset{Cau-Sch}{\leq} \sqrt{\mathrm{E}\left(\frac{d^2}{dt^2}I_w(\mu_t^{X|W})\right)\mathrm{E}\left(I_w(\mu_t^{X|W})\right)} = \sqrt{\left(\frac{d^2}{dt^2}I(\mu_t^{X|W})\right)\left(I(\mu_t^{X|W})\right)}.\end{aligned}$$

Here monotone convergence theorem can be used to justify the exchange of differentiation and expectation.

We state this in the following Corollary.

Corollary 2.3. *The conditional Fisher information of X_t with respect to W is log-convex in t, i.e.*

$$\ln I(\mu_t^{X|W}) = \ln \int_{\mathbb{R}^2} \left(\frac{\partial}{\partial x} \ln f_t(x|w)\right)^2 f_t(x|w) f_W(w)dxdw$$

is convex in t.

From Lemma 1.7 we see that $\frac{d}{dt}I(\mu_t^{X|W}) \leq 0$. This allows us to show the following Corollary.

Corollary 2.4 (with Yunrui Guan)**.** *For any $t_0, \tau > 0$ we have*

$$\big(h(X_{t_0+3\tau}|W) - h(X_{t_0+2\tau}|W)\big)\big(h(X_{t_0+\tau}|W) - h(X_{t_0}|W)\big) \geq \big(h(X_{t_0+2\tau}|W) - h(X_{t_0+\tau}|W)\big)^2.$$

Proof. Observe that

$$\begin{aligned}\left(\int_0^\tau I(\mu^{X|W}_{t_0+2\tau+x})dx\right)\left(\int_0^\tau I(\mu^{X|W}_{t_0+x})dx\right) &\overset{(a)}{\geq} \int_0^\tau I(\mu^{X|W}_{t_0+2\tau+x})I(\mu^{X|W}_{t_0+x})dx\\ &\overset{(b)}{\geq} \int_0^\tau \left(I(\mu^{X|W}_{t_0+\tau+x})\right)^2 dx\\ &\overset{Cau-Sch}{\geq} \left(\int_0^\tau I(\mu^{X|W}_{t_0+\tau+x})dx\right)^2.\end{aligned}$$

Here (a) follows from the FKG inequality since $\frac{d}{dt}I(\mu_t^{X|W}) \leq 0$ from Lemma 1.7. Inequality (b) follows from the log-convexity of $I(\mu_t^{X|W})$, and the last inequality is an immediate consequence of Cauchy-Schwarz

inequality. The statement in the Corollary follows immediately using DeBruijn's identity which implies that $I(\mu_t^{X|W}) = 2\frac{d}{dt}h(X_t|W)$. □

Remark 2.5. *One clear question that is definitely worth addressing is to determine whether the log-convexity of Fisher information along the heat flow also holds for random vectors. In particular we ask, whether*

$$\left(\frac{d^3h(\mathbf{X}+\sqrt{t}\mathbf{Z})}{dt^3}\right)\left(\frac{dh(\mathbf{X}+\sqrt{t}\mathbf{Z})}{dt}\right) \geq \left(\frac{d^2h(\mathbf{X}+\sqrt{t}\mathbf{Z})}{dt^2}\right)^2$$

where $\mathbf{X}$ *and* $\mathbf{Z}(\sim \mathcal{N}(0, I_d))$ *are independent random vectors taking values in* $\mathbb{R}^d$*. The proofs do not extend naturally to the random vectors instance. However Corollary 2.3 implies that it does extend to random vectors to independent components in a trivial manner. While the techniques applied in the scalar case do have natural extensions to the vector case, preliminary investigations by the authors indicate that these extensions seem insufficient to establish the log-convexity for vector valued random variables.*

Another way to extend the log-convexity to random vectors is to show (by some tensorization/single-letterization argument) that

$$\big(h(\mathbf{X}_{t_0+3\tau}) - h(\mathbf{X}_{t_0+2\tau})\big)\big(h(\mathbf{X}_{t_0+\tau}) - h(\mathbf{X}_{t_0})\big) \geq \big(h(\mathbf{X}_{t_0+2\tau}) - h(\mathbf{X}_{t_0+\tau})\big)^2,$$

holds for all $t_0, \tau > 0$*, and it is this approach which motivated Corollary 2.4.*

3. Discrete Markov Semigroups

To get a feeling of the veracity of Conjecture 1.1 or the weaker form of complete monotonicity, we consider families of Markov operators $W_{n,t}$ in finite dimensional probability spaces with dimension n that share similar the semi-group structures as Gaussian random variable, i.e. $W_{n,t_2}(W_{n,t_1}(p_{X_n})) = W_{n,t_1+t_2}(p_{X_n})$. To carry over the symmetry structure of the continuous world, we restrict ourselves to circulant (to capture $X_t = X + \sqrt{t}Z$), symmetric (to capture $Z \overset{d}{=} -Z$) Markov operators. Further one also imposes that as $t \to \infty$, W_t would tend to a completely noisy channel, which for an input-size n would correspond to $W_\infty = \frac{1}{n}\mathbb{1}_n$ where $\mathbb{1}_n$ is the $n \times n$ all-ones matrix. We identify a particular class of channels that satisfies the above assumptions.

Lemma 3.1. *Let* $n \in \mathbb{N}^+$ *denote the channel input size. Let* $W_{n,t}$ *be a collection of circulant symmetric stochastic matrices (parameterized by* t*) whose row and column entries are indexed from* 0 *to* $n-1$ *with the first row characterized by*

$$w_{0,0} = \frac{1+(n-1)e^{-t}}{n}, w_{0,i} = \frac{1-e^{-t}}{n} \forall i = 1, ..., n-1.$$

Then $W_{n,t}$ *satisfies the semi-group property, i.e.* $W_{n,t_1+t_2} = W_{n,t_1}W_{n,t_2}$ *and* $W_{n,\infty} = \frac{1}{n}\mathbf{1}_n$*, where* $\mathbf{1}_n$ *denotes the* $n \times n$ *matrix whose entries are all unity.*

Proof. Let $n \in \mathbb{N}^+$ and $W_{n,t}$ first row entries as defined in assumption. Since $W_{n,t}$ is a circulant matrix, we have

$$w_{i,j} = \begin{cases} \frac{1+(n-1)e^{-t}}{n}, & i = j \\ \frac{1-e^{-t}}{n}, & i \neq j \end{cases}$$

where $i, j = 0, ..., n-1$. It is immediate to verify have $W_{n,t_1+t_2} = W_{n,t_1}W_{n,t_2}$. Further note that $w_{i,j} \to \frac{1}{n}$ as $t \to \infty$. It follows that $W_{n,\infty} = \frac{1}{n}\mathbf{1}_n$ □

With this characterization, we will see it is convenient to parameterize probability distribution p_{X_n} to be the form of $p_{X_n}(i) = \frac{1+x_{n,i}}{n}$ where $\sum_{i=0}^{n-1} x_{n,i} = 0$ and $x_{n,i} \in [-1, n-1]$. By matrix multiplication, i.e. $p_{X_{n,t}} = W_{n,t}p_{X_n}$, we have $p_{X_{n,t}}(i) = \frac{1+x_{n,i}e^{-t}}{n}$.

Similar to before, let $p_{X_{n,t}} = W_{n,t}p_{X_n}$ denote a flow of probability distributions induced by the semi-group $W_{n,t}$ on $(n-1)$-dimensional probability simplex. Here $W_{n,t}$ denotes the specific Markov family described in Lemma 3.1. Let $I_{n,X_n}(t) := \frac{d}{dt}H(p_{X_{n,t}})$, where $H(p_X) = -\sum_x p_X(x)\ln p_X(x)$ denotes the Shannon-entropy of p_X.

The main goal of this section is check whether $I_{n,X_n}(t)$ is completely monotone for any $n \in \mathbb{N}^+$ and probability distribution p_{X_n}.

3.1. **Statement of Results.** Our first result is an affirmative result of the complete monotnicity for the binary alphabet.

Proposition 3.2 (with Daoyuan Max Chen, Chin Wa Ken Lau)**.** *$I_{2,X_2}(t)$ is completely monotone for any probability distribution p_{X_2}.*

Our second result is that that the signs of the first 9 derivatives alternate, but this alternating sign breaks down at the 10th derivative for large n and certain initial distributions.

Theorem 3.3 (with Daoyuan Max Chen, Chin Wa Ken Lau)**.** *The following holds:*

(i) For any $n \in \mathbb{N}^+$ and probability distribution p_{X_n}, we have

$$(-1)^k I^{(k)}_{n,X_n}(t) \geq 0, \forall t > 0$$

when $k = 0, ..., 9$.

(ii) There exists n and p_{X_n} such that $(-1)^{10} I^{(10)}_{n,X_n}(t) < 0$ for some $t > 0$.

Remark 3.4. *As our theorem shows, while the first 9 derivatives alternate in sign for any $n \in \mathbb{N}^+$, the function $I_{n,X_n}(t)$ is not completely monotone in general.*

3.2. **Proofs of Results.**

3.2.1. *Proof of Proposition 3.2.*

Proof. When $n = 2$, by Lemma 3.1, the circulant symmetric matrix is

$$W_{2,t} = \begin{bmatrix} \frac{1+e^{-t}}{2} & \frac{1-e^{-t}}{2} \\ \frac{1-e^{-t}}{2} & \frac{1+e^{-t}}{2} \end{bmatrix}.$$

Suppose $p_{X_2} = [\frac{1+x}{2}, \frac{1-x}{2}]^T$ where $x \in [-1, 1]$. Since $p_{X_{2,t}} = W_{2,t} p_{X_2}$, then $p_{X_{2,t}} = [\frac{1+xe^{-t}}{2}, \frac{1-xe^{-t}}{2}]^T$. We have

$$\begin{aligned} H(p_{X_{2,t}}) &= -\frac{1-xe^{-t}}{2} \ln \frac{1-xe^{-t}}{2} - \frac{1+xe^{-t}}{2} \ln \frac{1+xe^{-t}}{2} \\ &\overset{(a)}{=} \ln 2 - \left(\frac{1-xe^{-t}}{2} \sum_{k\geq 1} \frac{-x^k e^{-kt}}{k} + \frac{1+xe^{-t}}{2} \sum_{k\geq 1} \frac{(-1)^{k+1} x^k e^{-kt}}{k}\right) \\ &= \ln 2 - \sum_{k\geq 1} \frac{x^{2k} e^{-2kt}}{2k(2k-1)} \end{aligned}$$

where (a) uses the Taylor expansion of $\ln(1+xe^{-t}), \ln(1-xe^{-t})$, which is valid as $xe^{-t} \in (-1, 1)$ when $t > 0$. Then we have

$$I_{2,X_2}(t) = \sum_{k\geq 1} \frac{x^{2k} e^{-2kt}}{2k-1}.$$

Note that $I_{2,X_2}(t)$ is a summation of completely monotone functions, i.e. $\frac{x^{2k}}{2k-1} e^{-2kt}$. Hence $I_{2,X_2}(t)$ is completely monotone. □

3.2.2. *Proof of Theorem 3.3.* We will use the following lemma to establish Theorem 3.3.

Lemma 3.5. *Let $\phi_x(t) = xe^{-t} \ln(1+xe^{-t})$ where $x \in [-1, \infty)$. For any non-negative integer k, the following holds:*

(i) $(-1)^k \phi^{(k)}_x(t)$ is non-negative on $t > 0$ for any $x \in [-1, \infty)$ if and only if $(-1)^k I^{(k)}_{n,X_n}(t)$ is non-negative on $t > 0$ for any $n \in \mathbb{N}^+$ and probability distribution p_{X_n}.

(ii) $\phi_x(t)$ is completely monotone for any $x \in [-1, \infty)$ if and only if $I_{n,X_n}(t)$ is completely monotone for any $n \in \mathbb{N}^+$ and probability distribution p_{X_n}.

Proof. We will prove part (i) of Lemma 3.5. Note that part (ii) follows as an immediate corollary. Note that for any $n \in \mathbb{N}^+$, we have $P(p_{X_{n,t}} = i) = \frac{1+x_{n,i}e^{-t}}{n}$ for any $i = 0, ..., n-1$ where $\sum_{i=0}^{n-1} x_{n,i} = 0$ and $x_{n,i} \in [-1, n-1]$. Then we have

$$H(p_{X_{n,t}}) = \ln n - \frac{1}{n}\sum_{i=0}^{n-1}(1 + x_{n,i}e^{-t})\ln(1 + x_{n,i}e^{-t}).$$

It follows that

$$I_{n,X_n}(t) = \frac{1}{n}\sum_{i=0}^{n-1} x_{n,i}e^{-t}\ln(1 + x_{n,i}e^{-t}) = \frac{1}{n}\sum_{i=0}^{n-1}\phi_{x_{n,i}}(t).$$

Let $k \in \mathbb{N}$. Suppose $(-1)^k\phi_x^{(k)}(t)$ is non-negative on $t > 0$ for any $x \in [-1, \infty)$. Then for any $n \in \mathbb{N}^+$ and probability distribution p_{X_n}, we have $(-1)^k\phi_{x_{n,i}}^{(k)}(t)$ is non-negative on $t > 0$ for any $i = 0, ..., n-1$. Note that

$$(-1)^k I_{n,X_n}^{(k)}(t) = \frac{1}{n}\sum_{i=0}^{n-1}(-1)^k\phi_{x_{n,i}}^{(k)}(t).$$

Hence $(-1)^k I_{n,X_n}^{(k)}(t)$ is non-negative on $t > 0$.

Conversely, suppose there exists $x \in [-1, \infty)$, and $t_0 > 0$ such that $(-1)^k\phi_x^{(k)}(t_0) < 0$. Take a probability distribution p_{X_n} with $x_{n,0} = x$ and $x_{n,j} = -\frac{x}{n-1}$ for $j = 1, ..., n-1$. This is a valid distribution for any n as long as $n \geq |x| + 1$. It follows that

$$\begin{aligned}
(-1)^k n I_{n,X_n}^{(k)}(t_0) &= (-1)^k \sum_{i=0}^{n-1}\phi_{x_{n,i}}^{(k)}(t_0) \\
&= (-1)^k\phi_x^{(k)}(t_0) + (-1)^k\sum_{j=1}^{n-1}\phi_{x_{n,j}}^{(k)}(t_0) \\
&\overset{(a)}{=} (-1)^k\phi_x^{(k)}(t_0) + \sum_{j=1}^{n-1}(z-1)\ln(z) + \sum_{j=1}^{n-1}\frac{(z-1)^2}{z^k}Q_{k-1}(z)
\end{aligned}$$

where (a) uses the reparameterization $z = 1 + x_j e^{-t_0} = 1 + (-\frac{x}{n-1})e^{-t_0}$ and the expansion of $\phi_x^{(k)}(t)$ in Section 3.2.4, (in particular see Equation (11)). Note that $Q_k(z)$ is a polynomial in degree-k polynomial in z. Consider $\sum_{j=1}^{n-1}(z-1)\ln(z)$ and $\sum_{j=1}^{n-1}\frac{(z-1)^2}{z^k}Q_{k-1}(z)$ separately. Note that

$$\sum_{j=1}^{n-1}(z-1)\ln(z) = -xe^{-t_0}\ln(z) \to 0$$

$$\sum_{j=1}^{n-1}\frac{(z-1)^2}{z^k}Q_{k-1}(z) = \frac{1}{n-1}x^2e^{-2t_0}\frac{Q_{k-1}(z)}{z^k} \to 0$$

when $n \to \infty$. It follows that $(-1)^k n I_{n,X_n}^{(k)}(t_0) \to (-1)^k\phi_x^{(k)}(t_0)$ when $n \to \infty$. Then there exists $n \in \mathbb{N}^+$ such that $(-1)^k I_{n,X_n}^{(k)}(t_0) < 0$. Hence $(-1)^k I_{n,X_n}^{(k)}(t)$ is not non-negative on $t > 0$ for any $n \in \mathbb{N}^+$ and probability distribution p_{X_n}. □

3.2.3. *Proof of Theorem 3.3.*

Proof. In Section 3.2.4, we perform that the computation of the derivatives $\phi_x(t)$ and the k-th derivative can be expressed (see Equation (11)) as

$$\phi_x^{(k)}(t) = (-1)^k(z-1)\ln(z) + (-1)^k\frac{(z-1)^2}{z^k}Q_{k-1}(z).$$

Here $z = 1 + xe^{-t}$ and $Q_{k-1}(z)$ is a polynomial of degree $(k-1)$. By Lemma 3.5(i), in order to check whether $(-1)^k I_{n,X_n}^{(k)}(t)$ is non-negative on $t > 0$ for any $n \in \mathbb{N}^+$ and probability distribution p_{X_n}, it is equivalent to check whether $(-1)^k\phi_x^{(k)}(t)$ is non-negative on $t > 0$ for any $x \in [-1, \infty)$. Note that $z > 0$ when $x \in [-1, \infty)$ and $t > 0$. It follows that $(z-1)\ln z \geq 0$. Then $Q_{k-1}(z) \geq 0$ for any $z > 0$ is a sufficient (not necessary)

condition for $(-1)^k\phi_x^{(k)}(t) \geq 0$ for any $t > 0$ and $x \in [-1, \infty)$. In Lemma 3.7 of the Appendix, we show that $Q_0(z), Q_1(z), ..., Q_8(z)$ is non-negative when $z > 0$. Therefore Theorem 3.3(i) is established using Lemma 3.5(i).

We will see that $Q_9(z)$ (see (12)) is negative for some $z > 0$. In particular, take $z = 1.15$, we have $Q_9(z) \in (-88, -87)$, then

$$\begin{aligned}
-(z-1)\ln(z) + \frac{(z-1)^2}{z^{10}}Q_9(z) &= 0.15 \ln\frac{20}{3} + \frac{0.0225}{1.15^{10}}Q_9(1.15) \\
&\leq 0.15 \times 1.9 + \frac{0.0225}{4.05}(-87) \\
&= -\frac{119}{600} \\
&< 0
\end{aligned}$$

It follows that $(-1)^k\phi_x^{(10)}(t) < 0$ for some $x \in [-1, \infty)$, $t > 0$. Therefore Theorem 3.3(ii) is established. □

3.2.4. *Derivatives of $\phi_x(t)$.* Define $f(t) := x\log(1 + xe^{-t})$ and $y = xe^{-t}$. Note that $\phi_x(t) = f(t)e^{-t}$. By chain rule and induction, we obtain that

$$\phi_x^{(n)}(t) + (-1)^{n+1}f(t)e^{-t} = \sum_{k=1}^{n}(-1)^{n-k}\binom{n}{k}f^{(k)}(t)e^{-t}. \tag{9}$$

Lemma 3.6. *The kth derivative of f, for $k \geq 2$, satisfies*

$$f^{(k)}(t)e^{-t} = \frac{(-1)^k y^2 P_{k-2}(y)}{(1+y)^k}$$

where $P_{k-2}(y)$ is a polynomial of degree $(k-2)$.

Proof. The proof proceeds by induction. Note that, explicit calculation yields, the first few derivatives to satisfy.

$$\begin{aligned}
f^{(1)}(t)e^{-t} &= \frac{-x^2e^{-2t}}{1+xe^{-t}} = \frac{-y^2}{1+y} \\
f^{(2)}(t)e^{-t} &= \frac{x^2e^{-2t}}{(1+xe^{-t})^2} = \frac{y^2}{(1+y)^2} \\
f^{(3)}(t)e^{-t} &= \frac{-x^2e^{-2t}(1-xe^{-t})}{(1+xe^{-t})^3} = \frac{-y^2(1-y)}{(1+y)^3} \\
f^{(4)}(t)e^{-t} &= \frac{x^2e^{-2t}(1-4xe^{-t}+x^2e^{-2t})}{(1+xe^{-t})^4} = \frac{y^2(1-4y+y^2)}{(1+y)^4}.
\end{aligned}$$

Thus, let the lemma hold until the ℓth derivative, i.e.

$$f^{(\ell)}(t)e^{-t} = \frac{(-1)^\ell y^2 P_{\ell-2}(y)}{(1+y)^\ell},$$

where $P_{\ell-2}(y)$ is a polynomial of degree $\ell - 2$. Then by chain rule we have

$$\begin{aligned}
f^{(\ell+1)}(t)e^{-t} &= \frac{d}{dt}(f^{(\ell)}(t)e^{-t}) + f^{(\ell)}(t)e^{-t} \\
&= \frac{d}{dy}\frac{(-1)^\ell y^2 P_{\ell-2}(y)}{(1+y)^\ell}\frac{dy}{dt} + \frac{(-1)^\ell y^2 P_{\ell-2}(y)}{(1+y)^\ell} \\
&= (-1)^{\ell+1}y\left(\frac{[2yP_{\ell-2}(y) + y^2P'_{\ell-2}(y)](1+y)^\ell - \ell y^2 P_{\ell-2}(y)(1+y)^{\ell-1}}{(1+y)^{2\ell}}\right) + \frac{(-1)^\ell y^2 P_{\ell-2}(y)}{(1+y)^\ell} \\
&= \frac{(-1)^{\ell+1}y^2[(1-(\ell-1)y)P_{\ell-2}(y) + y(1+y)P'_{\ell-2}(y)]}{(1+y)^{\ell+1}} \\
&= \frac{(-1)^{\ell+1}y^2 P_{\ell-1}(y)}{(1+y)^{\ell+1}},
\end{aligned}$$

where

$$P_{\ell-1}(y) = (1-(\ell-1)y)P_{\ell-2}(y) + y(1+y)P'_{\ell-2}(y). \tag{10}$$

It is immediate that $P_{\ell-1}(y)$ is a polynomial of degree $\ell-1$. □

For notational consistency, let us denote $P_{-1}(y) = P_0(y) = 1$. Using (9) and Lemma 3.6 we see that

$$\phi_x^{(n)}(t) + (-1)^{n+1} f(t) e^{-t} = \frac{(-1)^n y^2}{(1+y)^n} \sum_{k=1}^{n} \binom{n}{k} (1+y)^{n-k} P_{k-2}(y).$$

Define a degree $n-1$ polynomial $\hat{Q}_{n-1}(y)$ according to

$$\hat{Q}_{n-1}(y) = \sum_{k=1}^{n} \binom{n}{k} (1+y)^{n-k} P_{k-2}(y).$$

Let $z = 1+y$ and $Q_{n-1}(z) = \hat{Q}_{n-1}(z-1)$. Then, we have

$$\phi_x^{(n)}(t) = (-1)^n (z-1)\ln z + (-1)^n \frac{(z-1)^2 Q_{n-1}(z)}{z^n}. \tag{11}$$

Lemma 3.7. *The polynomials $Q_k(z)$ are non-negative when $z > 0$, when $0 \le k \le 8$.*

Proof. We re-express the corresponding polynomials as a sum of non-negative terms below. By explicit calculation, we obtain

$$\begin{aligned}
Q_0(z) &= 1 \\
Q_1(z) &= 1 + 2z \\
Q_2(z) &= 2 + 2z + 3z^2 \\
Q_3(z) &= 6 + 2z + 3z^2 + 4z^3 \\
Q_4(z) &= 24 - 6z + 4z^2 + 4z^3 + 5z^4 \\
&= 15 + (3-z)^2 + 3z^2 + 4z^3 + 5z^4 \\
Q_5(z) &= 120 - 96z + 24z^2 + 4z^3 + 5z^4 + 6z^5 \\
&= 24 + 24(2-z)^2 + 4z^3 + 5z^4 + 6z^5 \\
Q_6(z) &= 720 - 960z + 384z^2 - 36z^3 + 6z^4 + 6z^5 + 7z^6 \\
&= 320(z - \frac{3}{2})^2 + 10z^2 + 6z^2(z-3)^2 + 6z^5 + 7z^6 \\
Q_7(z) &= 5040 - 9360z + 5760z^2 - 1296z^3 + 90z^4 + 6z^5 + 7z^6 + 8z^7 \\
&= 8z^7 + 7(z^3 + \frac{3}{7}z^2 - \frac{3}{4}z - 2)^2 + \frac{1389}{14}(z^2 - \frac{17689}{2778}z + \frac{13}{2})^2 + (\frac{4673}{229}z - \frac{229}{8})^2 + \frac{319634849689}{8158141488}z^2 + \frac{361}{448} \\
Q_8(z) &= 40320 - 95760z + 81360z^2 - 29520z^3 + 4248z^4 - 162z^5 + 8z^6 + 8z^7 + 9z^8 \\
&= z^8 + 8(z^4 + \frac{1}{2}z^3 - z^2 - z - \frac{2}{5})^2 + 22(z^3 - \frac{69}{22}z^2 - z + 4)^2 + \frac{449019}{110}(z^2 - \frac{1641574}{449019}z + \frac{147}{50})^2 \\
&\quad + (\frac{98101}{17109}z - \frac{17109}{250})^2 + \frac{13807378294210133547}{40160966452670250}z^2 + \frac{188763}{1375000}
\end{aligned}$$

Thus $Q_n(z)$ is non-negative for $n = 0, ..., 8$ when $z > 0$. □

Remark 3.8. *Note that similarly one can obtain*

$$Q_9(z) = 362880 - 1048320z + 1144080z^2 - 583920z^3 + 139320z^4 - 13392z^5 + 348z^6 + 8z^7 + 9z^8 + 10z^9. \tag{12}$$

4. Bibliographic Notes

The results on the log-convexity of Fisher Information in the scalar case is a joint work with Prof. Michel Ledoux and the author's doctoral student Yannan Wang. This result was published in Yannan Wang's thesis. Corollaries 2.3 and 2.4 were obtained as a result of the author's discussions with a current undergraduate student, Yunrui Guan. The section on discrete alphabets is the result of a collaboration with a then undergraduate advisee, Daoyuan Chen, and the author's doctoral student, Ken Lau. These results were presented in the undergraduate thesis of Daoyuan Chen.

References

[1] Fan Cheng and Yanlin Geng. Higher order derivatives in costa's entropy power inequality. *IEEE Transactions on Information Theory*, 61(11):5892–5905, 2015.

[2] T Cover and J Thomas. *Elements of Information Theory*. Wiley Interscience, 1991.

[3] A.M Fink. Kolmogorov-landau inequalities for monotone functions. *Journal of Mathematical Analysis and Applications*, 90(1):251–258, 1982.

[4] M. Ledoux. Heat flow derivatives and minimum mean-square error in gaussian noise. *IEEE Transactions on Information Theory*, 62(6):3401–3409, 2016.

[5] H. P. McKean. Speed of approach to equilibrium for kac's caricature of a maxwellian gas. *Archive for Rational Mechanics and Analysis*, 21(5):343–367, jan 1966.

[6] A.J. Stam. Some inequalities satisfied by the quantities of information of fisher and shannon. *Information and Control*, 2(2):101 – 112, 1959.

[7] C. Villani. A short proof of the "concavity of entropy power". *IEEE Transactions on Information Theory*, 46(4):1695–1696, 2000.

[8] A. Wyner and J. Ziv. A theorem on the entropy of certain binary sequences and applications: Part I. *IEEE Trans. Inform. Theory*, IT-19(6):769–772, Nov 1973.

[9] Xiaobing Zhang, Venkat Anantharam, and Yanlin Geng. Gaussian optimality for derivatives of differential entropy using linear matrix inequalities. *Entropy (Basel, Switzerland)*, 20(3), March 2018.