information

MDPI

Article

Reversing Jensen's Inequality for Information-Theoretic Analyses

Neri Merhav

The Viterbi Faculty of Electrical and Computer Engineering, Technion—Israel Institute of Technology, Haifa 3200003, Israel; merhav@ee.technion.ac.il; Tel.: +972-4-8294737

Abstract: In this work, we propose both an improvement and extensions of a reverse Jensen inequality due to Wunder et al. (2021). The new proposed inequalities are fairly tight and reasonably easy to use in a wide variety of situations, as demonstrated in several application examples that are relevant to information theory. Moreover, the main ideas behind the derivations turn out to be applicable to generate bounds to expectations of multivariate convex/concave functions, as well as functions that are not necessarily convex or concave.

Keywords: Jensen's inequality; reverse Jensen's inequality; converse Jensen's inequality; convex function; concave function

Citation: Merhav, N. Reversing Jensen's Inequality for Information-Theoretic Analyses. *Information* **2022**, *13*, 39. https://dx.doi.org/10.3390/info13010039

Academic Editor: Hector Zenil

Received: 15 December 2021
Accepted: 13 January 2022
Published: 13 January 2022

1. Introduction

It is very well known that Jensen's inequality is one of the most elementary mathematical tools, which is extremely useful in a variety of fields, including information theory. At the general level, it covers many other well-known inequalities, which are useful in their own right, as special cases. Examples in general applied mathematics include the Shwartz–Cauchy inequality (which in turn serves the uncertainty principle and the Cramér–Rao inequality), the Hölder inequality, the Lyapunov inequality, and the inequalities of arithmetic, geometric, and harmonic means, just to name a few. More specifically, in information theory, it is basis of the information inequality (i.e., the non-negativity of the Kullback–Leibler divergence), the data processing inequality (which, in turn, leads to the Fano inequality), and the property that conditioning reduces entropy. Moreover, it has a pivotal role in the generation of single-letter formulas in Shannon theory and in the theory of maximum entropy under moment constraints (see, for example, Chapter 12 of [1]).

As often as not, however, applied mathematicians, and in particular, information theorists (the author included), encounter the somewhat frustrating situation that Jensen's inequality works in the opposite direction than the one they would hope for along their way to obtaining their desired results. It is conceivable to speculate that it is this fact that has triggered a considerable research effort in developing a variety of versions of the so-called *reverse Jensen inequality* (RJI) (see, e.g., Refs. [2–11] for a non-exhaustive sample of articles on this topic). In most of these (and other) works, the inequalities derived are demonstrated in a variety of applications, for example, useful relationships between arithmetic and geometric means, reverse bounds on the entropy, the Kullback–Leibler divergence (and, more generally, Csiszár's f-divergence), reverse versions of the Hölder inequality, etc. In many of the abovementioned papers, the main results are given in the form of an upper bound on the difference, $\mathbb{E}\{f(X)\} - f(\mathbb{E}\{X\})$, where f is the convex function, $\mathbb{E}\{\cdot\}$ is the expectation operator, and X is the random variable. These upper bounds, however, depend mostly on global properties of the function f, such as its range and domain, but not quite on the underlying probability density function (PDF) of X (or its probability mass function, in the discrete case). For example, a desirable property of an RJI would be tightness when the PDF of X is well concentrated in the vicinity of its mean, just like the same well-known property of the ordinary Jensen inequality,

$$\mathbb{E}\{f(X)\} \geq f(\mathbb{E}\{X\}). \tag{1}$$

In this work, we build on [9] and provide several new variants of the RJI, which possess the abovementioned desired property of tightness under measure concentration. Our starting point is the same as in (the proof of) Lemma 1 of [9], but the continuation of our derivation is substantially different. As a result of this difference, the proposed approach yields considerably tighter bounds, which are reasonably convenient to analyze and calculate in many cases of interest, as we demonstrate throughout this work. We also extend the scope to functions of more than one variable, which are convex (or concave) in each variable alone, but not necessarily so jointly, in all variables. Finally, using similar ideas, we also derive upper and lower bounds on expectations of functions that are not necessarily convex or concave altogether.

To summarize, the main contributions of this work relative to [9] are as follows.

1. Improvement of the tightness of the lower bound on the expectation of a concave function.
2. Relaxing some of the assumptions on the concave function.
3. Proposing a more convenient (and perhaps more natural) passage from lower bounds to expectations of concave functions to upper bounds on expectations of convex functions.
4. Extension to bivariate (and multivariate) functions that are concave (or convex) in each variable.
5. Providing examples of usefulness in information theory (other than the mutual information estimation of [9]).

The outline of this article is as follows. In Section 2, we present the basic inequality, rooted in Lemma 1 of [9], and discuss its weaknesses. In Section 3, we present two alternative approaches to improve on the basic bound of [9] and discuss the relative strengths and weaknesses of each one. In Section 4, we provide several examples for the usefulness of our proposed bounds. In Section 5, we mostly discuss variations and modifications of our main results, and finally, in Section 6, we extend the scope to bivariate (and multivariate) functions.

2. The Basic Reverse Inequality

Our starting point, which is very similar to Lemma 1 of [9], is the following Lemma.

Lemma 1. *Let $f : \mathbb{R}^+ \to \mathbb{R}$ be a concave function with $f(x) \geq f(0)$ for every $x \geq 0$. Let X be a non-negative random variable with a finite mean, $\mathbb{E}\{X\} = \mu$. Then,*

$$\mathbb{E}\{f(X)\} \geq \sup_{a>0} \left[\frac{\mu}{a} \cdot f(a) + \left(1 - \frac{\mu}{a}\right) \cdot f(0) - \frac{f(a) - f(0)}{a} \cdot \mathbb{E}\{X \cdot \mathcal{I}[X > a]\} \right], \quad (2)$$

where $\mathcal{I}[X > a]$ denotes the indicator function of event $\{X > a\}$.

The proof is almost the same as in the first few steps in the proof of ([9], Lemma 1), but nevertheless, we present it here for completeness.

Proof of Lemma 1. For every $x \in [0, a]$,

$$\begin{aligned} f(x) &= f\left(\left[1 - \frac{x}{a}\right] \cdot 0 + \frac{x}{a} \cdot a\right) && (3) \\ &\geq \left(1 - \frac{x}{a}\right) \cdot f(0) + \frac{x}{a} \cdot f(a) && (4) \\ &= f(0) + x \cdot \frac{f(a) - f(0)}{a}, && (5) \end{aligned}$$

and so,

$$\frac{f(x) - f(0)}{x} \geq \frac{f(a) - f(0)}{a} \quad \text{whenever } 0 \leq x \leq a. \quad (6)$$

Thus,

$$\begin{aligned}
\mathbb{E}\{f(X)\} &= \mathbb{E}\left\{\frac{f(X)-f(0)}{X}\cdot X\right\}+f(0) && (7)\\
&\geq \mathbb{E}\left\{\frac{f(X)-f(0)}{X}\cdot X\cdot \mathcal{I}[X\leq a]\right\}+f(0) && (8)\\
&\geq \mathbb{E}\left\{\frac{f(a)-f(0)}{a}\cdot X\cdot \mathcal{I}[X\leq a]\right\}+f(0) && (9)\\
&= \frac{f(a)-f(0)}{a}\cdot \mathbb{E}\{X\cdot \mathcal{I}[X\leq a]\}+f(0) && (10)\\
&= \frac{f(a)-f(0)}{a}\cdot \mathbb{E}\{X\cdot (1-\mathcal{I}[X> a])\}+f(0) && (11)\\
&= \frac{f(a)-f(0)}{a}\cdot \mu+f(0)-\frac{f(a)-f(0)}{a}\cdot \mathbb{E}\{X\cdot \mathcal{I}[X> a]\} && (12)\\
&= \frac{\mu}{a}\cdot f(a)+\left(1-\frac{\mu}{a}\right)\cdot f(0)-\frac{f(a)-f(0)}{a}\cdot \mathbb{E}\{X\cdot \mathcal{I}[X> a]\}, && (13)
\end{aligned}$$

where in the first inequality, we relied on the assumption that $f(x)\geq f(0)$ for every $x\geq 0$ and in the second inequality we have used (6). Since this chain of inequalities holds true for every $a>0$, we may take the supremum of the right-hand side (r.h.s.) over all $a>0$. This completes the proof of Lemma 1. □

Lemma 1 serves as the basis of our subsequent derivations throughout this paper. The main remaining issue now is how to assess the term

$$q(a)\stackrel{\triangle}{=}\mathbb{E}\{X\cdot \mathcal{I}[X>a]\}. \tag{14}$$

In simple cases, $q(a)$ can be calculated exactly by a closed form expression, for example, when the PDF of X is uniform, or triangular, or exponential, etc. In most cases of interest, however, it is difficult, if not impossible, to derive an exact, closed-form expression for $q(a)$, and then one has to resort to upper bounds in order to further lower bound the r.h.s. of Equation (2).

In [9] (Lemma 1), it was proposed to upper bound $q(a)$ by applying the Hölder inequality and then the Markov inequality, to obtain

$$q(a)\leq \inf_{p>1}\left\{(\mathbb{E}\{X^p\})^{1/p}\cdot\left(\frac{\mu}{a}\right)^{1-1/p}\right\}. \tag{15}$$

This is a very interesting bound, but unfortunately, it suffers from two main weaknesses. The first is that in many cases, the calculation of high order moments of X may not be a trivial task, let alone the step of taking the infimum over p. More importantly, the second weakness is that, as mentioned in the Introduction, it does not yield a tight lower bound to $\mathbb{E}\{f(X)\}$ when the PDF of X concentrates strongly around μ. To see the intuition, consider the following argument. When X fluctuates in the vicinity of μ with high probability, it is clear that the optimal choice of a, in the sense of maximizing the r.h.s. of (2), is slightly larger than μ, because then $q(a)$ would be small and the main term of (2) would be close to $f(\mu)$, which, in turn, is also the Jensen upper bound to $\mathbb{E}\{f(X)\}$ and hence it is fairly tight. Now, by the Lypunov inequality, the factor $\mathbb{E}\{X^p\}]^{1/p}$ is increasing in p, and so, for every $p\geq 1$, $[\mathbb{E}\{X^p\}]^{1/p}\geq \mathbb{E}\{X\}=\mu$. At the same time, for every $a>\mu$, we have $(\mu/a)^{1-1/p}\geq \mu/a$. Consequently, if we take, for the sake of simplicity, $f(0)=0$, the lower bound of [9] (Lemma 1), cannot be larger than

$$\sup_{a>\mu}\frac{f(a)}{a}\cdot\left(\mu-\frac{\mu^2}{a}\right)=\sup_{a>\mu}f(a)\cdot\frac{\mu}{a}\left(1-\frac{\mu}{a}\right)=\sup_{0\leq\theta<1}f\left(\frac{\mu}{\theta}\right)\theta(1-\theta). \tag{16}$$

In fact, this is a very generous assessment of the bound of ([9], Lemma 1), as the inequality $[\mathbb{E}\{X^p\}]^{1/p} \geq \mathbb{E}X = \mu$ is met with equality for $p = 1$, whereas the inequality $(\mu/a)^{1-1/p} \geq \mu/a$ is met for $p \to \infty$.

For example, if $f(x) = \sqrt{x}$, [9] (Lemma 1) yields a lower bound which is smaller than

$$\sup_{0\leq\theta<1} \sqrt{\frac{\mu}{\theta}}\theta(1-\theta) = \sqrt{\mu} \cdot \sup_{0\leq\theta<1} \sqrt{\theta}(1-\theta) = \frac{2\sqrt{3}}{9} \cdot \sqrt{\mu} = 0.3849\sqrt{\mu} \tag{17}$$

even if $X = \mu$ with probability one.

These observations motivate the quest for improvements of the upper bound on $q(a)$, which is the subject of the next section.

3. Alternative Upper Bounds to $q(a)$

For cases where $q(a)$ does not lend itself to an exact closed-form expression, we propose two basic alternative approaches for upper bounding $q(a)$, which both have the property that when X concentrates around μ, $q(a)$ is small even when a is only slightly larger than μ, which in turn yields a tight bound very close to $f(\mu)$.

The choice between the two approaches depends on the problem at hand and the capability to obtain closed-form expressions for the moments involved, if they exist at all.

1. *The Chernoff approach.* The first approach is to upper bound the indicator function, $\mathcal{I}\{x > a\}$, by the exponential function $e^{s(x-a)}$ ($s \geq 0$), exactly like in the Chernoff bound. This would yield

$$q(a) \leq \inf_{s\geq 0} \mathbb{E}\{Xe^{s(X-a)}\} = \inf_{s\geq 0}\left[e^{-as}\mathbb{E}\{Xe^{sX}\}\right] = \inf_{s\geq 0}\left[e^{-as}\Phi'(s)\right] \stackrel{\triangle}{=} q_{\text{Chernoff}}(a), \tag{18}$$

where $\Phi'(s)$ is the derivative of the moment generating function (MGF), $\Phi(s) \stackrel{\triangle}{=} \mathbb{E}\{e^{sX}\}$. Thus, Equation (2) is further lower bounded as

$$\mathbb{E}\{f(X)\} \geq \sup_{a>0}\left[\frac{\mu}{a} \cdot f(a) + \left(1 - \frac{\mu}{a}\right) \cdot f(0) - \frac{f(a) - f(0)}{a} \cdot q_{\text{Chernoff}}(a)\right]. \tag{19}$$

This bound is useful when X has a finite MGF, $\Phi(s)$, at least in some range of $s > 0$, and $\Phi(s)$ is differentiable in that range. Moreover, for the bound to be useful, $q_{\text{Chernoff}}(a)$ must lend itself to a reasonably simple closed-form expression.

As a slight variation of the Chernoff approach is to upper bound the function $x \cdot \mathcal{I}[x > a]$ itself (and not just the indicator function factor) by an exponential function of the form $a \cdot e^{s(x-a)}$, where s is such that the derivative w.r.t. x at $x = a$ is not less than 1, so that it is at least tangential to the function $x \cdot \mathcal{I}[x > a]$ for $x \downarrow a$. This means $as \geq 1$, or $s \geq 1/a$. Thus,

$$q(a) \leq a \cdot \inf_{s\geq 1/a}\{e^{-as}\Phi(s)\} \stackrel{\triangle}{=} \tilde{q}_{\text{Chernoff}}(a) \tag{20}$$

which, of course, may replace $q_{\text{Chernoff}}(a)$ in Equation (19). The usefulness of this version of the bound is essentially under the same circumstances as those of $q_{\text{Chernoff}}(a)$. It has the small advantage that there is no need to differentiate $\Phi(s)$, but the range of the optimization over s is somewhat smaller.

2. *The Chebychev–Cantelli approach.* According to this approach, the function $x \cdot \mathcal{I}[x > a]$ is upper bounded by a quadratic function, in the spirit of the Chebychev–Cantelli inequality, i.e.,

$$x \cdot \mathcal{I}[x > a] \leq \frac{a(x+s)^2}{(a+s)^2}, \tag{21}$$

where the parameter $s \geq 0$ is optimized under the constraint that the derivative at $x = a$, which is $2a/(a+s)$, is at least 1 (again, to be at least tangential to the function itself at $x \downarrow a$), which is equivalent to the requirement, $s \leq a$. In this case, denoting $\sigma^2 = \mathrm{Var}\{X\}$, we obtain

$$q(a) \leq \frac{a\mathbb{E}\{(X+s)^2\}}{(a+s)^2} = \frac{a[\sigma^2 + (\mu+s)^2]}{(a+s)^2}, \tag{22}$$

which, when minimized over $s \in [0, a]$, yields

$$s^* = \min\left\{a, \frac{\sigma^2}{a-\mu} - \mu\right\}, \tag{23}$$

and then the best bound is given by

$$q(a) \leq q_{\text{Cheb-Cant}}(a) \stackrel{\triangle}{=} \begin{cases} \frac{\sigma^2+(a+\mu)^2}{4a} & a < a_{\text{c}} \\ \frac{a\sigma^2}{\sigma^2+(a-\mu)^2} & a \geq a_{\text{c}} \end{cases} \tag{24}$$

where $a_{\text{c}} \stackrel{\triangle}{=} \sqrt{\sigma^2+\mu^2}$.

The Chernoff approach yields better bounds than the Chebychev–Cantelli approach in many cases. Suppose, for example, that $X = \sum_{i=1}^n Y_i$, where $Y_1, \ldots, Y_n$ are independently and identically distributed (i.i.d.) random variables, all having mean μ_Y, variance σ_Y^2, and MGF $\Phi_Y(s)$. Then, of course, $\mu = n\mu_Y$, $\sigma^2 = n\sigma_Y^2$, and $\Phi(s) = [\Phi_Y(s)]^n$. For simplicity, suppose also that $f(0) = 0$. In this case, the Chernoff approach yields

$$\begin{aligned} \mathbb{E}\left\{f\left(\sum_{i=1}^n Y_i\right)\right\} &\geq \frac{n\mu_Y}{a} \cdot f(a) - \frac{f(a)}{a} \inf_{s\geq 0}\left\{e^{-sa}\frac{\mathrm{d}}{\mathrm{d}s}[\Phi_Y(s)]^n\right\} \\ &= \frac{n\mu_Y}{a} \cdot f(a) - \frac{nf(a)}{a} \inf_{s\geq 0}\left\{e^{-sa}[\Phi_Y(s)]^{n-1}\Phi_Y'(s)\right\} \\ &= \frac{nf(a)}{a}\left[\mu_Y - \inf_{s\geq 0}\left\{e^{-sa}[\Phi_Y(s)]^n \cdot \frac{\mathrm{d}\ln\Phi_Y(s)}{\mathrm{d}s}\right\}\right]. \end{aligned} \tag{25}$$

Now, if $Y_1, Y_2, \ldots$ obey a large deviations principle, the second term in the square brackets tends to zero exponentially for the choice $a = n(\mu_Y + \epsilon)$ with arbitrarily small $\epsilon > 0$. In this case, let $s^* > 0$ be the maximizer of $[s(\mu+\epsilon) - \ln\Phi_Y(s)]$, yielding $I(\epsilon) = s^*(\mu+\epsilon) - \ln\Phi_Y(s^*)$. Then,

$$\mathbb{E}\left\{f\left(\sum_{i=1}^n Y_i\right)\right\} \geq \frac{f[(\mu_Y+\epsilon)n]}{\mu_Y+\epsilon}\left[\mu_Y - e^{-nI(\epsilon)}\frac{\mathrm{d}\ln\Phi_Y(s)}{\mathrm{d}s}\bigg|_{s=s^*}\right]. \tag{26}$$

For large enough n, the second term in the square brackets becomes negligible, and the lower bound becomes arbitrarily close to $f[(\mu_Y+\epsilon)n] \cdot \mu_Y/(\mu_Y+\epsilon)$. On the other hand, Jensen's upper bound is $f(\mu_Y n)$. In some cases, the difference is not very large, at least for asymptotic evaluations. For example, if $f(x) = \ln(1+x)$, which is a frequently encountered concave function in information theory, $\ln[1+n(\mu_Y+\epsilon)] \geq \ln n + \ln(\mu_Y+\epsilon)$, whereas $\ln(1+n\mu_Y) \leq \ln n + \ln(\mu_Y + 1/n)$, which are very close for large n and small $\epsilon > 0$. In Section 4.4, we will get back to the example of the logarithm of the sum independent random variables.

In the Chebychev–Cantelli approach, on the other hand, we have $a_{\text{c}} = \sqrt{n^2\mu_Y^2 + n\sigma_Y^2} \sim n\mu_Y$ for large n. Thus, if we take $a = n(\mu_Y+\epsilon) > a_{\text{c}}$, we have

$$q_{\text{Cheb-Cant}}[n(\mu_Y+\epsilon)] = \frac{n\sigma_Y^2}{n\sigma_Y^2 + n^2\epsilon^2} = \frac{\sigma_Y^2}{\sigma_Y^2 + n\epsilon^2}, \tag{27}$$

which tends to zero, but only at the rate of $1/n$, as opposed to the exponential decay in the Chernoff approach. Still, for large n, the main term of the bound becomes asymptotically tight, as before.

In spite of the superiority of the Chernoff approach relative to the Chebychev–Cantelli approach, as we demonstrated, one should keep in mind that there are also situations where the random variable X does not have an MGF (i.e., when the PDF of X has a heavy tail), yet it does have a mean and a variance. In such cases, the Chebychev–Cantelli approach is applicable while the Chernoff approach is not. However, even when the MGF exists, in certain cases, the calculation of the first and the second moment is easier than the calculation of the exponential moment.

To conclude this section, we summarize its main findings in the form of a theorem.

Theorem 1. *Under the conditions of Lemma 1,*

$$\mathbb{E}\{f(X)\} \geq \sup_{a>0}\left[\frac{\mu}{a}\cdot f(a) + \left(1-\frac{\mu}{a}\right)\cdot f(0) - \frac{f(a)-f(0)}{a}\cdot q_{\min}(a)\right], \tag{28}$$

where

$$q_{\min}(a) = \min\{q_{\mathit{Chernoff}}(a), \tilde{q}_{\mathit{Chernoff}}(a), q_{\mathit{Cheb\text{-}Cant}}(a)\}. \tag{29}$$

4. Examples

In this section, we demonstrate the lower bound in several information-theoretic application examples.

4.1. Example 1—Capacity of the Gaussian Channel with Random SNR

We begin with a simple example of a zero-mean, circularly symmetric complex Gaussian channel whose signal-to-noise ratio (SNR), Z, is a random variable (e.g., due to fading), known to both the transmitter and the receiver. The capacity is given by $C = \mathbb{E}\{\ln(1+gZ)\}$, where g is a certain deterministic gain factor and the expectation is with respect to (w.r.t.) the randomness of Z. For simplicity, let us assume that Z is distributed exponentially, i.e.,

$$p_Z(z) = \theta e^{-\theta z}, \quad z \geq 0, \tag{30}$$

where the parameter $\theta > 0$ is given. In this case, $f(x) = \ln(1+gx)$, $\mu = 1/\theta$ and $q(a)$ can be easily derived in closed form, to obtain

$$q(a) = \theta \cdot \int_a^\infty z e^{-\theta z}\mathrm{d}z = \left(a + \frac{1}{\theta}\right)\cdot e^{-\theta a}. \tag{31}$$

Consequently,

$$\begin{aligned} C &\geq \sup_{a\geq 1/\theta} \tfrac{\ln(1+ga)}{a}\left[\tfrac{1}{\theta} - \left(a+\tfrac{1}{\theta}\right)\cdot e^{-a\theta}\right] \\ &= \sup_{s\geq 1}\left[\tfrac{1-(s+1)e^{-s}}{s}\right]\cdot \ln\left(1+\tfrac{gs}{\theta}\right), \end{aligned} \tag{32}$$

whereas the Jensen upper bound is $C \leq \ln(1+g/\theta)$. Figure 1 displays both the upper bound and the lower bound to C as functions of θ. As can be seen, the gap between the bounds depends strongly on θ. For large θ, which is the case where the p_Z becomes narrower and more concentrated, the gap between the bounds decays rapidly in the sense that not only does the difference shrink (which by itself is not surprising, since both bounds tend to zero), but their ratio also becomes closer to unity.

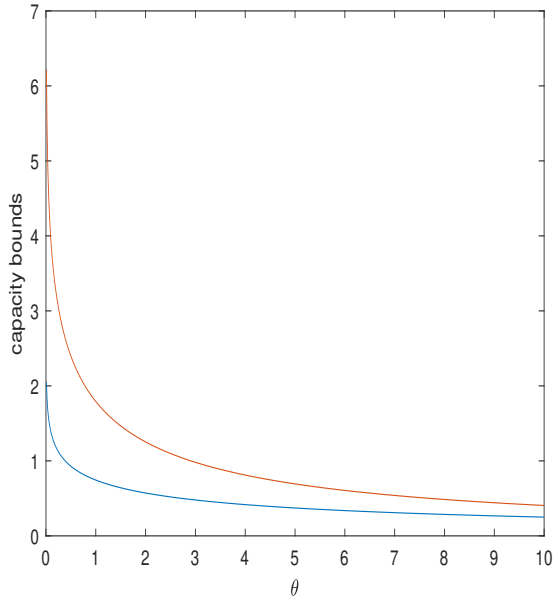

Figure 1. Capacity upper and lower bounds for $g = 5$. The blue curve is the lower bound as a function of θ and the red curve is the ordinary Jensen upper bound, given by $\ln(1 + g/\theta)$.

4.2. Example 2—Moments of the Number of Guesses in Randomized Guessing

Consider the problem of guessing the realization of a random variable, which takes on values in a finite alphabet, using a sequence of yes/no questions of the form "Is $U = u_1$?", "Is $U = u_2$?", etc., until a positive response is provided by a party that observes the actual realization of U (see, e.g., Ref. [12] whose Introduction includes an overview on this problem). Given a distribution of U, a commonly used performance metric for this problem is the ρ^{th} moment of the number of guesses until U is guessed successfully (for a given $\rho > 0$). For guessing random vectors of length n, minimizing the moments of the number of guesses by different (deterministic or randomized) guessing strategies has several applications and motivations in information theory, such as sequential decoding, guessing passwords, etc., and it is also strongly related to lossless source coding. In this vector case, the moments of the number of guesses behave asymptotically exponentially in n. Here, we refer to randomized guessing strategies, rather than deterministic strategies.

Let the random variable U take on values in a finite alphabet $\mathcal{U}$. Consider a random guessing strategy where the guesser sequentially submits a sequence of independently drawn random guesses according to a certain probability distribution, $P(\cdot)$, defined on $\mathcal{U}$. Randomized guessing strategies have the advantage that they can be used by multiple asynchronous agents, which submit their guesses concurrently (again, see [12] and references therein). Let $\boldsymbol{u} \in \mathcal{U}^n$ be any realization of $\boldsymbol{U}$, and let the guessing distribution, P, be given. The random number, X, of independent guesses until success has a geometric distribution:

$$\Pr\{X = k\} = [1 - P(\boldsymbol{u})]^{k-1} \cdot P(\boldsymbol{u}), \quad k = 1, 2, \ldots \tag{33}$$

and we are interested in a lower bound to $\mathbb{E}\{X^\rho\}$ for $\rho \in (0, 1)$, which is the case where $f(x) = x^\rho$ is concave. Obviously, the condition $f(x) = x^\rho \geq 0 = f(0)$ is met as well in this case. We next use the shorthand notation p instead of $P(\boldsymbol{u})$, and later we will think of this quantity as an exponential function of n, denoted e^{-nE}, where $E > 0$ is some constant that depends on the type of $\boldsymbol{u}$. We will also denote $q = 1 - p$. Then,

$$\begin{aligned} \mathbb{E}\{X^\rho\} &\geq \tfrac{\mu}{a} \cdot a^\rho - \tfrac{a^\rho}{a} \inf_{s \geq 0}\{e^{-sa}\Phi'(s)\} \\ &= a^{\rho-1}\left[\mu - \inf_{s \geq 0}\{e^{-sa}\Phi'(s)\}\right]. \end{aligned} \tag{34}$$

Now, X is a geometric random variable with $\mu = 1/p$, and

$$\begin{aligned} \Phi(s) &= \sum_{k=1}^{\infty} e^{ks} \Pr\{X = k\} \\ &= \frac{pe^s}{1 - qe^s}, \quad s < \ln \frac{1}{q}. \end{aligned} \tag{35}$$

which yields

$$\Phi'(s) = \frac{pe^s}{(1 - qe^s)^2}. \tag{36}$$

Choosing $s = s^* = \ln[(a-1)/aq]$, with $a = e^{n(E+\epsilon)}$ (for some arbitrarily small $\epsilon > 0$), gives

$$\begin{aligned} e^{-as}\Phi'(s) &= \frac{p}{q} \cdot \frac{a(a-1)}{(1-1/a)^a} \cdot q^a \\ &\leq \frac{e^{-nE}}{1 - e^{-nE}} \cdot \frac{e^{2nE}}{\left(1 - e^{-n(E+\epsilon)}\right)^{e^{n(E+\epsilon)}}} \cdot (1 - e^{-nE})^{e^{n(E+\epsilon)}} \\ &= \frac{e^{nE}}{(1 - e^{-nE})\left(1 - e^{-n(E+\epsilon)}\right)^{e^{n(E+\epsilon)}}} \cdot (1 - e^{-nE})^{e^{n(E+\epsilon)}}. \end{aligned} \tag{37}$$

Now, as $n \to \infty$, the denominator tends to $1/e$ and the factor $(1 - e^{-nE})^{e^{n(E+\epsilon)}}$ tends to zero double-exponentially, and so the entire expression goes to zero double-exponentially. Thus, the main term of the lower bound is

$$\mu a^{\rho - 1} = e^{nE\rho - (1-\rho)\epsilon}, \tag{38}$$

which is roughly $e^{nE\rho}$ on the exponential scale. The Jensen upper bound is $(\mathbb{E}\{X\})^\rho = 1/p^\rho = e^{nE\rho}$, which is of the same exponential order if ϵ is neglected.

4.3. Example 3—Moments of the Error in Parameter Estimation

Consider the estimation of a parameter θ from given observations. Suppose that we have an expression (or a lower bound), $\epsilon^2(\theta)$, on the mean square error (MSE) of a certain estimator, and the question is how to obtain a lower bound to other moments of the estimation error, such as $\mathbb{E}\{|\hat{\theta} - \theta|^q\}$ where $q < 2$. Using the proposed approach, we have, in this example, $X = (\hat{\theta} - \theta)^2$ and $f(x) = x^{q/2}$. Consequently,

$$\begin{aligned} \mathbb{E}\{|\hat{\theta} - \theta|^q\} &= \mathbb{E}\left\{ \left| (\hat{\theta} - \theta)^2 \right|^{q/2} \right\} \\ &\geq \frac{a^{q/2}}{a} \cdot \Big[\mathbb{E}\{(\hat{\theta} - \theta)^2\} - \mathbb{E}\{(\hat{\theta} - \theta)^2 \cdot \mathcal{I}[(\hat{\theta} - \theta)^2 \geq a]\} \Big] \\ &\geq a^{q/2-1} \cdot \Big[\epsilon^2(\theta) - \mathbb{E}\{(\hat{\theta} - \theta)^2 \cdot \mathcal{I}[(\hat{\theta} - \theta)^2 \geq a]\} \Big]. \end{aligned} \tag{39}$$

Now, if the given estimator has the additional large-deviations property to make the second term exponentially small for large n, then we can lower bound $\mathbb{E}\{|\hat{\theta} - \theta|^q\}$ by an expression whose main term is $[\epsilon^2(\theta)]^{q/2}$. For example, if θ is the mean of a Bernoulli source and $\hat{\theta}$ is the empirical relative frequency, or θ is the mean of Gaussian observations, we can calculate the bound relatively easily. We will not pursue this example any further here.

4.4. Logarithms of Sums of Independent Random Variables

Referring to the discussion near the end of Section 3, where X is given by the sum of n independent random variables, we now focus on the special case where $f(x) = \ln(1 + x)$,

which was also mentioned therein (see the discussion around Equation (26)). Here, we list several examples where this example is encountered in information theory applications.

4.4.1. Example 4—Universal Source Coding

Consider the evaluation of the expected code length associated with the universal lossless source code due to Krichevsky and Trofimov [13]. In a nutshell, this is a universal code for memoryless sources. In the binary case, at each time instant t, it sequentially assigns probabilities to the next binary symbol according to (a biased version of) the empirical distribution pertaining to the source data observed so far, $s_1, \ldots, s_t$. Specifically, consider the ideal code-length function (in nats),

$$L(s^n) = -\sum_{t=0}^{n-1} \ln Q(s_{t+1}|s_1, \ldots, s_t), \tag{40}$$

where

$$Q(s_{t+1} = s|s_1, \ldots, s_t) = \frac{N_t(s) + 1}{t + 2}, \tag{41}$$

and $N_t(s)$, $s \in \{0, 1\}$ is the number of occurrences of the symbol s in $(s_1, \ldots, s_t)$. Therefore,

$$\begin{aligned} \mathbb{E}\{L(S^n)\} &= \sum_{t=0}^{n-1} \ln(t+2) - \sum_{t=0}^{n-1} \mathbb{E}\{\ln[N_t(S_{t+1}) + 1]\} \\ &= \ln[(n+1)!] - \sum_{t=0}^{n-1} \mathbb{E}\left\{\ln\left(1 + \sum_{i=0}^{t} \mathcal{I}[S_i = S_{t+1}]\right)\right\} \\ &= \ln[(n+1)!] - p \cdot \sum_{t=0}^{n-1} \mathbb{E}\left\{\ln\left(1 + \sum_{i=0}^{t} \mathcal{I}[S_i = 1]\right)\right\} - \\ &\quad (1-p) \cdot \sum_{t=0}^{n-1} \mathbb{E}\left\{\ln\left(1 + \sum_{i=0}^{t} \mathcal{I}[S_i = 0]\right)\right\}, \end{aligned} \tag{42}$$

where $\mathcal{I}[\cdot]$ are indicator functions of the corresponding events and where p and $1 - p$ are the probabilities of "1" and "0", respectively. Now, in order to obtain an upper bound to $\mathbb{E}\{L(S^n)\}$, one can lower bound each of the terms, $\mathbb{E}\{\ln(1 + \sum_{i=0}^{t} \mathcal{I}[S_i = 1])\}$ and $\mathbb{E}\{\ln(1 + \sum_{i=0}^{t} \mathcal{I}[S_i = 0])\}$. This falls in the framework of logarithms of sums of i.i.d. random variables, where $\{U_i\}$ are binary.

4.4.2. Example 5—Ergodic Capacity of the Rayleigh SIMO Channel

Consider the single-input, multiple-output (SIMO) channel with L receiving antennas and assume that the channel transfer coefficients, $h_1, \ldots, h_L$, are independent, zero-mean, circularly symmetric complex Gaussian random variables with variances $\sigma_1^2, \ldots, \sigma_L^2$. The ergodic capacity (in nats per channel use) of the SIMO channel is given by

$$C = \mathbb{E}\left\{\ln\left(1 + A \sum_{\ell=1}^{L} |h_\ell|^2\right)\right\} = \mathbb{E}\left\{\ln\left(1 + A \sum_{\ell=1}^{L} (f_\ell^2 + g_\ell^2)\right)\right\}, \tag{43}$$

where $f_\ell = \mathrm{Re}\{h_\ell\}$, $g_\ell = \mathrm{Im}\{h_\ell\}$, and A is the SNR (see, e.g., Refs [14,15] and many references therein).

Once again, here $f(x) = \ln(1 + x)$, and now, $X = A \sum_{\ell=1}^{L} |h_\ell|^2$. Here, the transfer coefficients are assumed independent, but not identically distributed. Therefore, the means should be summed, and so should the variances, whereas the MGFs should be multiplied. Clearly, the case where they all share the same PDF is a special case.

4.4.3. Example 6—Differential Entropy of the Generalized Multivariate Cauchy Distribution

Let $(Y_1, \ldots, Y_n)$ be a multivariate Cauchy random vector, whose PDF is given by

$$p(y_1, \ldots, y_n) = \frac{C_n}{\left[1 + \sum_{i=1}^n y_i^2\right]^{(n+1)/2}}. \tag{44}$$

Using the Laplace transform relation,

$$\frac{1}{s^{(n+1)/2}} = \frac{1}{\Gamma((n+1)/2)} \int_0^\infty t^{(n-1)/2} e^{-st} \, \mathrm{d}t, \quad \mathrm{Re}(s) > 0, \tag{45}$$

f can be represented as a mixture of product measures:

$$\begin{aligned} f(y_1, \ldots, y_n) &= \frac{C_n}{\left[1 + \sum_{i=1}^n y_i^2\right]^{(n+1)/2}} \\ &= \frac{C_n}{\Gamma((n+1)/2)} \int_0^\infty t^{(n-1)/2} e^{-t} \exp\left\{-t \sum_{i=1}^n y_i^2\right\} \mathrm{d}t. \end{aligned} \tag{46}$$

Defining

$$Z(t) \stackrel{\triangle}{=} \int_{-\infty}^\infty e^{-ty^2} \, \mathrm{d}y = \sqrt{\frac{\pi}{t}}, \quad t > 0, \tag{47}$$

we get from (46),

$$\begin{aligned} 1 &= \frac{C_n}{\Gamma((n+1)/2)} \int_0^\infty t^{(n-1)/2} e^{-t} \int_{\mathbb{R}^n} \exp\left\{-t \sum_{i=1}^n y_i^2\right\} \mathrm{d}y_1 \ldots \mathrm{d}y_n \, \mathrm{d}t \\ &= \frac{C_n}{\Gamma((n+1)/2)} \int_0^\infty t^{(n-1)/2} e^{-t} \left(\int_{-\infty}^\infty e^{-ty^2} \, \mathrm{d}y\right)^n \mathrm{d}t \\ &= \frac{C_n}{\Gamma((n+1)/2)} \int_0^\infty t^{(n-1)/2} e^{-t} \left(\frac{\pi}{t}\right)^{n/2} \mathrm{d}t, \end{aligned} \tag{48}$$

and so,

$$C_n = \frac{\Gamma((n+1)/2)}{\int_0^\infty t^{(n-1)/2} e^{-t} \left(\frac{\pi}{t}\right)^{n/2} \mathrm{d}t}. \tag{49}$$

The calculation of the differential entropy is associated with the evaluation of the expectation $\mathbb{E}\left\{\ln\left[1 + \sum_{i=1}^n Y_i^2\right]\right\}$. This falls within the framework of this application example, where $X = \sum_{i=1}^n Y_i^2$. Here, Y_i^2 are not i.i.d., but they are distributed under a mixture of Gaussian i.i.d. distributions, as can be seen in (46). In particular,

$$\begin{aligned} p(y_1, \ldots, y_n) &= \frac{C_n}{\left[1 + \sum_{i=1}^n y_i^2\right]^{(n+1)/2}} \\ &= \frac{C_n}{\Gamma((n+1)/2)} \int_0^\infty t^{(n-1)/2} e^{-t} \left(\frac{\pi}{t}\right)^{n/2} \frac{\exp\left\{-t \sum_{i=1}^n y_i^2\right\}}{(\pi/t)^{n/2}} \, \mathrm{d}t \\ &= \frac{C_n \pi^{n/2}}{\Gamma((n+1)/2)} \int_0^\infty t^{-1/2} e^{-t} \frac{\exp\left\{-t \sum_{i=1}^n y_i^2\right\}}{(\pi/t)^{n/2}} \, \mathrm{d}t \end{aligned} \tag{50}$$

Let us denote

$$w(t) = \frac{C_n \pi^{n/2} e^{-t}}{\Gamma((n+1)/2)\sqrt{t}}, \quad t > 0. \tag{51}$$

Then, $(Y_1, \ldots, Y_n)$ is governed by a mixture of i.i.d., zero-mean Gaussians of variance $1/(2t)$ with weight function $w(t)$, $t > 0$. Therefore,

$$\mathbb{E}\left\{\ln\left(1+\sum_{i=1}^{n} Y_i^2\right)\right\} = \int_0^\infty \mathrm{d}t w(t) \mathbb{E}\left\{\ln\left(1+\sum_{i=1}^{n} Y_i^2\right)\middle| t\right\}, \tag{52}$$

and we can now lower bound $\mathbb{E}\left\{\ln\left(1+\sum_{i=1}^{n} Y_i^2\right)\middle| t\right\}$ for each $t > 0$ and, finally, integrate the result (choosing possibly different a and s for every t). Following the discussion near the end of Section 3, the result will be roughly

$$\int_0^\infty \mathrm{d}t w(t) \ln\left(1+\frac{n}{2t}\right), \tag{53}$$

which is in agreement with the Jensen upper bound.

5. Discussion

In this section, we discuss several additional aspects and observations regarding our main result, as well as modifications and suggestions for relaxing certain assumptions. We summarize our comments below.

1. *The maximization over a is not necessary.* Our first comment is quite trivial but, nevertheless, it is important to mention at least as a reminder. The explicit maximization over the parameter a may not be trivial to carry out in most examples, but for certain purposes, it may not be necessary. One can select an arbitrary value of a and obtain a legitimate lower bound. In some cases, however, it is not too difficult to guess what could be a good choice of this value, as we saw in some of the examples of Section 4.

2. *Softening the assumption $f(x) \geq f(0)$.* One may partially relax the assumption $f(x) \geq f(0)$ for all $x \geq 0$, and replace it with the softer assumption that there exists $\Delta \geq 0$, such that $f(x) + \Delta \cdot x \geq f(0)$ for all $x \geq 0$ (or more precisely, within the support of the PDF of X). By applying Lemma 1 and Theorem 1 to $f(x) + \Delta \cdot x$ and compensating for the term $\mathbb{E}\{\Delta \cdot X\} = \Delta\mu$, we can easily use exactly the same technique and obtain the modified lower bound,

$$\mathbb{E}\{f(X)\} \geq \sup_{a>0}\left[\frac{\mu}{a} \cdot f(a) + \left(1-\frac{\mu}{a}\right) f(0) - \left[\frac{f(a)-f(0)}{a} + \Delta\right] \cdot q(a)\right], \tag{54}$$

and so the cost of this relaxation is the extra Δ in the last term. This means that the best choice of Δ is the smallest one for which $f(x) + \Delta x \geq f(0)$ for all $x \geq 0$, namely,

$$\Delta = \Delta^* \stackrel{\triangle}{=} \sup_{x>0} \frac{f(0)-f(x)}{x}. \tag{55}$$

Nevertheless, if we can make $q(a)$ small by selecting a to be just slightly above μ, this cost can be kept small. The results presented in Section 3 then become the special case of $\Delta = 0$. Moreover, the artificially added (and subtracted) term, Δx, can also be replaced by a non-linear term, say, $g(x)$, provided that $f(x) + g(x)$ remains concave and $f(x) + g(x) \geq f(0) + g(0)$. This is useful, of course, only if $\mathbb{E}\{g(X)\}$ lends itself to an explicit calculation, for example, $g(x) = x^n$, or $g(x) = e^{ax}$.

3. *Convex functions.* So far we have dealt with RJIs for concave functions. RJIs associated with expectations of convex functions (upper bounds) can be obtained in exactly the same manner, except that the signs are flipped. Specifically, by replacing f with $-f$, we have

similar statements for convex functions: Let $f : \mathbb{R}^+ \to \mathbb{R}$ be a convex function with $f(x) \leq f(0)$ for every $x \geq 0$. Then,

$$\mathbb{E}\{f(X)\} \leq \inf_{a>0}\left[\frac{\mu}{a} \cdot f(a) + \left(1 - \frac{\mu}{a}\right) \cdot f(0) + \frac{f(0) - f(a)}{a} \cdot q(a)\right], \tag{56}$$

and, again, the assumption $f(x) \leq f(0)$ can be softened in the same manner as described in item 2 above to obtain

$$\mathbb{E}\{f(X)\} \leq \inf_{a>0}\left[\frac{\mu}{a} \cdot f(a) + \left(1 - \frac{\mu}{a}\right) \cdot f(0) + \left[\frac{f(0) - f(a)}{a} + \Delta\right] \cdot q(a)\right]. \tag{57}$$

4. *Functions that are neither convex nor concave.* Using the same line of thought as in item 2 above, one can obtain upper and lower bounds to expectations of general functions, that are not necessarily convex or concave. Indeed, let $f : [a, b] \to \mathbb{R}$ be a real, continuous function that satisfies the following condition: $y > x$ implies $f(y) \leq f(x) + \Delta(y - x)$ for all x, y. Assume also that f is bounded from below by a constant, $f_{\min}$. Then, $f(x) + \Delta \cdot (b - x) - f_{\min}$ is monotonically non-increasing and positive. Thus, for every $u \in [a, b]$,

$$f(x) + \Delta \cdot (b - x) - f_{\min} \geq [f(u) + \Delta(b - u) - f_{\min}] \cdot \mathcal{I}\{x \leq u\}, \tag{58}$$

and so,

$$\mathbb{E}\{f(X)\} + \Delta \cdot (b - \mu) - f_{\min} \geq [f(u) + \Delta(b - u) - f_{\min}] \cdot \Pr\{X \leq u\}, \tag{59}$$

or

$$\begin{aligned} \mathbb{E}\{f(X)\} \quad &\geq \quad f_{\min} - \Delta \cdot (b - \mu) + f(u) + \Delta \cdot (b - u) - f_{\min} - \\ &\qquad [f(u) + \Delta(b - u) - f_{\min}] \cdot \Pr\{X \geq u\} \qquad (60) \\ &= \quad f(u) - \Delta \cdot (u - \mu) - [f(u) + \Delta(b - u) - f_{\min}] \cdot \Pr\{X \geq u\}. \qquad (61) \end{aligned}$$

so, finally,

$$\mathbb{E}\{f(X)\} \geq \sup_{u \in [a,b]} \{f(u) - \Delta \cdot (u - \mu) - [f(u) + \Delta(b - u) - f_{\min}] \cdot \Pr\{X \geq u\}\},$$

with the understanding that we can further lower bound $\mathbb{E}\{f(X)\}$ by using any of the available upper bounds on $\Pr\{X \geq u\}$ (Markov, Chebychev, Chebychev–Cantelli, Chernoff, Hoeffding, etc.). The choice depends on considerations of tightness and the calculability of the bound, as described before.

Likewise, if $f : [a, b] \to \mathbb{R}$ is a real, continuous function that satisfies the following condition: $y > x$ implies $f(y) \geq f(x) - \Delta(y - x)$ for all x, y, and is upper bounded by $f_{\max} < \infty$, we obtain, in the same manner,

$$\mathbb{E}\{f(X)\} \leq \inf_{u \in [a,b]} \{f(u) + \Delta \cdot (u - \mu) - [f(u) - \Delta(b - u) - f_{\max}] \cdot \Pr\{X \geq u\}\}. \tag{62}$$

6. Extension to Bivariate (and Multivariate) Concave Functions

We conclude this article by outlining a possible extension of our main results to bivariate concave functions, with the understanding that the main ideas also extend to multivariate functions of more than two variables. It should also be understood that a parallel upper bound for convex functions can easily be obtained by flipping the signs as before. For simplicity, we assume parallel assumptions to those of Lemma 1 and Theorem 1. In this section, we confine attention to the Chernoff approach, but the other approaches are valid as well.

For a pair of random variables (X,Y), governed by a given joint PDF, we define the following functions:

$$\Phi(s,t) \stackrel{\triangle}{=} \mathbb{E}\{e^{sX+tY}\} \tag{63}$$

$$\Phi_s(s) \stackrel{\triangle}{=} \frac{\partial\Phi(s,0)}{\partial s} = \mathbb{E}\{Xe^{sX}\} \tag{64}$$

$$\Phi_t(t) \stackrel{\triangle}{=} \frac{\partial\Phi(0,t)}{\partial t} = \mathbb{E}\{Ye^{tY}\} \tag{65}$$

$$\Phi_{st}(s,t) \stackrel{\triangle}{=} \frac{\partial^2\Phi(s,t)}{\partial s\partial t} = \mathbb{E}\{XYe^{sX+tY}\} \tag{66}$$

$$q_1(a) = \inf_{s\geq 0}\{e^{-sa}\Phi_s(s)\} \tag{67}$$

$$q_2(b) = \inf_{t\geq 0}\{e^{-tb}\Phi_t(t)\} \tag{68}$$

$$q_3(a) = \inf_{s\geq 0}\{e^{-sa}\Phi_{st}(s,0)\} \tag{69}$$

$$q_4(b) = \inf_{t\geq 0}\{e^{-tb}\Phi_{st}(0,t)\}. \tag{70}$$

We then have the following theorem.

Theorem 2. *Let $f:\mathbb{R}_+^2\to\mathbb{R}$ have the property $f(x,y)\geq f(x,0)$ and $f(x,y)\geq f(0,y)$ for all positive x and y in their domains, and assume that f is concave in x for fixed y, and concave in y for fixed x (but not necessarily jointly concave in both variables). Then,*

$$\begin{aligned}\mathbb{E}\{f(X,Y)\} \geq \sup_{a>0}\sup_{b>0}\Bigg[&\frac{f(a,b)-f(a,0)}{ab}[\mathbb{E}\{XY\}-q_3(a)-q_4(b)]+\\ &\frac{f(a,0)}{a}[\mathbb{E}\{X\}-q_1(a)]-\frac{1}{a}\mathbb{E}\{X\cdot f(0,\mathbb{E}\{Y|X\})\}+\\ &\frac{f(0,b)-f(0,0)}{b}\cdot[\mathbb{E}\{Y\}-q_2(b)]+f(0,0)\Bigg].\end{aligned} \tag{71}$$

Likewise,

$$\begin{aligned}\mathbb{E}\{f(X,Y)\} \geq \sup_{a>0}\sup_{b>0}\Bigg[&\frac{f(a,b)-f(0,b)}{ab}[\mathbb{E}\{XY\}-q_3(a)-q_4(b)]+\\ &\frac{f(0,b)}{b}[\mathbb{E}\{Y\}-q_2(b)]-\frac{1}{b}\mathbb{E}\{Y\cdot f(\mathbb{E}\{X|Y\},0)\}+\\ &\frac{f(a,0)-f(0,0)}{a}\cdot[\mathbb{E}\{X\}-q_1(a)]+f(0,0)\Bigg].\end{aligned} \tag{72}$$

Proof of Theorem 2. We begin with the simple identity,

$$\mathbb{E}\{f(X,Y)\} = \mathbb{E}\{\mathbb{E}\{f(X,Y)|Y\}\}. \tag{73}$$

Now, applying Theorem 1 to $f(X,y)$ (for fixed y) w.r.t. the conditional PDF of X given $Y=y$, we obtain:

$$\mathbb{E}\{f(X,y)|Y=y\} \geq \frac{f(a,y)-f(0,y)}{a}\cdot\mathbb{E}\{X\cdot\mathcal{I}\{X\leq a\}|Y=y\}+f(0,y). \tag{74}$$

Thus,

$$\mathbb{E}\{f(X,Y)\} \overset{(a)}{\geq} \frac{1}{a}\mathbb{E}\{[f(a,Y)-f(0,Y)]X\cdot\mathcal{I}\{X\leq a\}\}+\mathbb{E}\{f(0,Y)\} \tag{75}$$

$$\overset{(b)}{\geq} \frac{1}{a}\mathbb{E}\{[f(a,Y)X\cdot\mathcal{I}\{X\leq a\}\}-\frac{1}{a}\mathbb{E}\{[f(0,Y)X\cdot\mathcal{I}\{X\leq a\}\}+$$
$$\frac{f(0,b)-f(0,0)}{b}\cdot(\mathbb{E}\{Y\}-\mathbb{E}\{Y\cdot\mathcal{I}\{Y\geq b\}\})+f(0,0) \tag{76}$$

$$\overset{(c)}{\geq} \frac{1}{a}\mathbb{E}\left\{\left[\frac{f(a,b)-f(a,0)}{b}\cdot Y\cdot\mathcal{I}\{Y\leq b\}+f(a,0)\right]X\cdot\mathcal{I}\{X\leq a\}\right\}-$$
$$\frac{1}{a}\mathbb{E}\{[f(0,Y)X\}+\frac{f(0,b)-f(0,0)}{b}\cdot(\mathbb{E}\{Y\}-\mathbb{E}\{Y\cdot\mathcal{I}\{Y\geq b\}\})+f(0,0) \tag{77}$$

$$\overset{(d)}{=} \frac{f(a,b)-f(a,0)}{ab}\mathbb{E}\{XY\cdot\mathcal{I}\{X\leq a\}\cdot\mathcal{I}\{Y\leq b\}\}+\frac{f(a,0)}{a}\mathbb{E}\{X\cdot\mathcal{I}\{X\leq a\}\}-$$
$$\frac{1}{a}\mathbb{E}\{\mathbb{E}\{f(0,Y)X|X\}\}+\frac{f(0,b)-f(0,0)}{b}\cdot(\mathbb{E}\{Y\}-\mathbb{E}\{Y\cdot\mathcal{I}\{Y\geq b\}\})+f(0,0)$$

$$\overset{(e)}{\geq} \frac{f(a,b)-f(a,0)}{ab}[\mathbb{E}\{XY\}-\mathbb{E}\{XY\cdot\mathcal{I}\{X\geq a\}\}-\mathbb{E}\{XY\cdot\mathcal{I}\{Y\geq b\}]+$$
$$\frac{f(a,0)}{a}[\mathbb{E}\{X\}-\mathbb{E}\{X\cdot\mathcal{I}\{X\geq a\}]-\frac{1}{a}\mathbb{E}\{X\cdot f(0,\mathbb{E}\{Y|X\})\}+$$
$$\frac{f(0,b)-f(0,0)}{b}\cdot(\mathbb{E}\{Y\}-\mathbb{E}\{Y\cdot\mathcal{I}\{Y\geq b\}\})+f(0,0)$$

$$\overset{(f)}{=} \frac{f(a,b)-f(a,0)}{ab}[\mathbb{E}\{XY\}-q_3(a)-q_4(b)]+$$
$$\frac{f(a,0)}{a}[\mathbb{E}\{X\}-q_1(a)]-\frac{1}{a}\mathbb{E}\{X\cdot f(0,\mathbb{E}\{Y|X\})\}+$$
$$\frac{f(0,b)-f(0,0)}{b}\cdot[\mathbb{E}\{Y\}-q_2(b)]+f(0,0), \tag{78}$$

where (a) is by taking the expectation over Y on both sides of (74), (b) is by applying Theorem 1 to $f(0,Y)$, (c) is by applying Theorem 1 to $f(a,Y)$ (first, w.r.t. to the conditional PDF of Y given X, and then by averaging w.r.t. X), (d) is by rearranging terms, (e) is by the inequality $\mathcal{I}\{X\leq a\}\cdot\mathcal{I}\{Y\leq b\}\geq 1-\mathcal{I}\{X\geq a\}-\mathcal{I}\{Y\geq b\}$, and (f) is by the definitions of the functions $q_1(\cdot),\ldots,q_4(\cdot)$. The second inequality in Theorem 2 is proved in the same manner by switching the roles of: (i) X and Y, (ii) a and b, (iii) $q_3(a)$ and $q_4(b)$, and (iv) $f(a,0)$ and $f(0,b)$. In other words, we first condition on X rather than on Y. This completes the proof of Theorem 2. □

The usefulness of this bound depends on the ability to calculate, or at least to upper bound the term, $\frac{1}{a}\mathbb{E}\{X\cdot f(0,\mathbb{E}\{Y|X\})\}$. If the function $g(x)=x\cdot f(0,\mathbb{E}\{Y|X=x\})$ happens to be a concave function, we can use Jensen's inequality. If it is convex, we can apply a version of reverse Jensen for convex functions. If it is neither convex nor concave, we can use the method presented in the previous section to upper bound it. When X and Y are independent, this term simplifies significantly, as it becomes $\mu_X\cdot f(0,\mu_Y)$, where μ_X and μ_Y are the means of X and Y, respectively.

Another case where the lower bound simplifies considerably is the case where $f(a,0)=f(0,b)=0$ for all a and b, as it becomes

$$\mathbb{E}\{f(X,Y)\}\geq\frac{f(a,b)}{ab}[\mathbb{E}\{XY\}-q_3(a)-q_4(b)]. \tag{79}$$

We next provide two application examples where this is the case.

Example 7—minimum between two sums of independent random variables. Let $X=\sum_{i=1}^{n}A_i$ and $Y=\sum_{j=1}^{m}B_j$, where both $\{A_i\}$ and $\{B_j\}$ are all non-negative, independent random

variables. Obviously, $\mu_X = \sum_{i=1}^{n} \mu_{A_i}$ and $\mu_Y = \sum_{j=1}^{m} \mu_{B_j}$. Further, let $f(x,y) = \min\{x,y\}$, which is concave in x for fixed y and vice versa. Then,

$$\mathbb{E}\left\{ \min\left\{ \sum_{i=1}^{n} A_i, \sum_{i=1}^{m} B_i \right\} \right\} \geq \frac{\min\{a,b\}}{ab} \left[\sum_{i=1}^{n} \sum_{j=1}^{m} \mu_{A_i}\mu_{B_j} - q_3(a) - q_4(b) \right], \tag{80}$$

which is essentially $\min\{\mu_X, \mu_Y\}$ for large n and m, provided that X and Y concentrate around their means, as discussed above. This example has a few applications, all of them are relevant in situations where there is a certain additive cost associated with a given task, there are two possible routes (or strategies), and the one with the smaller cost is chosen. For example, suppose we are compressing a realization, $u_1, \ldots, u_n$, of a random source vector, $U_1, \ldots, U_n$, and we have two side informations (available at both ends), $V_1, \ldots, V_n$ and $W_1, \ldots, W_n$, which are both conditionally independent noisy versions of $u_1, \ldots, u_n$, but for practical reasons, we use only one of them—the one for which code-length is shorter for the given realization (also adding a flag bit). In this case, $n = m$, $A_i = -\log P(u_i|V_i)$, and $B_i = -\log P(u_i|W_i)$, which are independent. As a second step, we have, of course, to take the expectation over $(U_1, \ldots, U_n)$. Other examples of costs might be prices, distances, waiting times, bit errors, etc.

Example 8—channel capacity revisited. Here, we combine Example 1 (capacity of the Gaussian channel with random SNR) and Example 5 (ergodic capacity of the SIMO channel). Consider the expression

$$C = \mathbb{E}\left\{ \log\left(1 + Z \cdot \sum_{\ell=1}^{L} |h_\ell|^2 \right) \right\}, \tag{81}$$

where, as in Example 1, $\{h_\ell\}$ are zero-mean, circularly symmetric, complex Gaussian random variables with variances $\{\sigma_\ell^2\}$, and as in Example 5, Z is an exponentially distributed random variable with parameter θ, and independent of $\{h_\ell\}$. In principle, we could have treated this problem in the framework of univariate functions, using the concavity of $f(X) = \ln(1+X)$, where the random variable X is defined as $X = Z \cdot \sum_{\ell=1}^{L} h_\ell^2$. However, the calculation of the characteristic function of X is not convenient to analyze since it is a product of two random variables. Instead, we treat it as a bivariate function, $f(X,Y) = \ln(1+XY)$, where $X = Z$ and $Y = \sum_{\ell=1}^{L} h_\ell^2$. Clearly, f is concave in each one of its arguments when the other one is kept fixed. We then have

$$C \;\geq\; \frac{\ln(1+ab)}{ab} \left[\frac{1}{\theta} \cdot \sum_{\ell=1}^{L} \sigma_\ell^2 - q_3(a) - q_4(b) \right]. \tag{82}$$

As in Example 5, instead of the Chernoff-like bound, $q_3(a)$, we have the exact expression

$$\mathbb{E}\{XY \cdot \mathcal{I}[X \geq a]\} \;=\; \mathbb{E}\{Y\} \cdot \mathbb{E}\{X \cdot \mathcal{I}[X \geq a]\} \tag{83}$$

$$\;=\; \left(\sum_{\ell=1}^{L} \sigma_\ell^2 \right) \cdot \left(a + \frac{1}{\theta} \right) \cdot e^{-\theta a}. \tag{84}$$

The other term can be readily bounded using the Chernoff approach.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Interesting discussions with the authors of [9] are acknowledged with thanks.

Conflicts of Interest: The author declares no conflict of interests.

References

1. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2006.
2. Jebara, T.; Pentland, A. On reversing Jensen's inequality. In Proceedings of the 13th International Conference on Neural Information Processing Systems (NIPS 2000), Denver, CO, USA, 1 January 2000; pp. 213–219.
3. Budimir, I.; Dragomir, S.S.; Pečarić, J. Further reverse results for Jensen's discrete inequality and applications in information theory. *J. Inequal. Pure Appl. Math.* **2001**, *2*, 5.
4. Simić, S. On a new converse of Jensen's inequality. *Math. Publ. L'Institut Math.* **2009**, *85*, 107–110. [CrossRef]
5. Dragomir, S.S. Some reverses of the Jensen inequality for functions of selfadjoint operators in Hilbert spaces. *J. Inequal. Appl.* **2010**, 496821. [CrossRef]
6. Dragomir, S.S. Some reverses of the Jensen inequality with applications. *Bull. Aust. Math. Soc.* **2013**, *87*, 177–194. [CrossRef]
7. Khan, S.; Khan, M.A.; Chu, Y.-M. New converses of Jensen inequality via Green functions with applications. *Rev. Real Acad. Cienc. Exactas Fis. Nat. Ser. A Mat.* **2020**, *114*. [CrossRef]
8. Khan, S.; Khan, M.A.; Chu, Y.-M. Converses of Jensen inequality derived from the Green functions with applications in information theory. *Math. Methods Appl. Sci.* **2020**, *43*, 2577–2587. [CrossRef]
9. Wunder, G.; Groβ, B.; Fritschek, R.; Schaefer, R.F. A reverse Jensen inequality result with application to mutual information estimation. In Proceedings of the 2021 IEEE Information Theory Workshop (ITW 2021), Kanazawa, Japan, 17–21 October 2021. Available online: https://arxiv.org/pdf/2111.06676.pdf (accessed on 12 November 2021).
10. Ali, M.A.; Budak, H.; Zhang, Z. A new extension of quantum Simpson's and quantum Newton's inequalities for quantum differentiable convex functions. *Math. Methods Appl. Sci.* **2021**, 1–19.
11. Budak H.; Ali, M.A.; Tarhanaci, M. Some new quantum Hermite-Hadamard like inequalities for co-ordinated convex functions. *J. Optim. Theory Appl.* **2020**, *186*, 899–910. [CrossRef]
12. Merhav, N.; Cohen, A. Universal randomized guessing with application to asynchronous decentralized brute-force attacks. *IEEE Trans. Inform. Theory* **2020**, *66*, 114–129. [CrossRef]
13. Krichevsky, R.E.; Trofimov, V.K. The performance of universal encoding. *IEEE Trans. Inform. Theory* **1981**, *27*, 199–207. [CrossRef]
14. Dong, A.; Zhang, H.; Wu, D.; Yuan, D. Logarithmic expectation of the sum of exponential random variables for wireless communication performance evaluation. In Proceedings of the 2015 IEEE 82nd Vehicular Technology Conference (VTC2015-Fall), Boston, MA, USA, 6–9 September 2015.
15. Tse, D.; Viswanath, P. *Fundamentals of Wireless Communication*; Cambridge University Press: Cambridge, UK, 2005.