HAL open science

On the entropy and information of Gaussian mixtures

Alexandros Eskenazis, Lampros Gavalakis

▶ **To cite this version:**

Alexandros Eskenazis, Lampros Gavalakis. On the entropy and information of Gaussian mixtures. 2023. hal-04272904

HAL Id: hal-04272904

https://hal.science/hal-04272904

Preprint submitted on 6 Nov 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ON THE ENTROPY AND INFORMATION OF GAUSSIAN MIXTURES

ALEXANDROS ESKENAZIS AND LAMPROS GAVALAKIS

Abstract. We establish several convexity properties for the entropy and Fisher information of mixtures of centered Gaussian distributions. First, we prove that if X_1, X_2 are independent scalar Gaussian mixtures, then the entropy of $\sqrt{t}X_1 + \sqrt{1-t}X_2$ is concave in $t \in [0,1]$, thus confirming a conjecture of Ball, Nayar and Tkocz (2016) for this class of random variables. In fact, we prove a generalisation of this assertion which also strengthens a result of Eskenazis, Nayar and Tkocz (2018). For the Fisher information, we extend a convexity result of Bobkov (2022) by showing that the Fisher information matrix is operator convex as a matrix-valued function acting on densities of mixtures in $\mathbb{R}^d$. As an application, we establish rates for the convergence of the Fisher information matrix of the sum of weighted i.i.d. Gaussian mixtures in the operator norm along the central limit theorem under mild moment assumptions.

2020 Mathematics Subject Classification. Primary: 94A17; Secondary: 60E15, 26B25.
Key words. Entropy, Fisher information, Gaussian mixture, Central Limit Theorem, rates of convergence.

1. Introduction

1.1. **Entropy.** Let X be a continuous random vector in $\mathbb{R}^d$ with density $f : \mathbb{R}^d \to \mathbb{R}_+$. The (differential) entropy of X is the quantity

$$h(X) \stackrel{\text{def}}{=} -\int_{\mathbb{R}^d} f(x)\log f(x)\,\mathrm{d}x = \mathbb{E}\Big[-\log f(X)\Big], \tag{1}$$

where log always denotes the natural logarithm. The celebrated entropy power inequality of Shannon and Stam [30, 31] (see also [25]) implies that for every independent continuous random vectors X_1, X_2 in $\mathbb{R}^d$, we have

$$\forall\ t \in [0,1], \qquad h\Big(\sqrt{t}X_1 + \sqrt{1-t}X_2\Big) \geq t h(X_1) + (1-t)h(X_2). \tag{2}$$

In general, the entropy power inequality cannot be reversed (see, e.g., the construction of [11, Proposition 4]). However, reverse entropy power inequalities have been considered under different assumptions on the random vectors, such as log-concavity [9, 15, 3, 26].

It follows directly from (2) that if X_1, X_2 are i.i.d. random vectors, then the entropy function $t \mapsto h(\sqrt{t}X_1 + \sqrt{1-t}X_2)$ is minimised at $t = 0$ and $t = 1$. In the spirit of reversing the entropy power inequality, Ball, Nayar and Tkocz [3] raised the question of maximising this function. In particular, they gave an example of a random variable X_1 for which the maximum is not attained at $t = \frac{1}{2}$ but conjectured that for i.i.d. log-concave random variables this function must be concave in $t \in [0,1]$, in which case it is in particular maximised at $t = \frac{1}{2}$. It is worth noting that the conjectured concavity would also be a *strengthening* of the entropy power inequality for i.i.d. random variables, as (2) amounts to the concavity condition for the points $0, t, 1$. So far, no special case of the conjecture of [3] seems to be known.

In this work, we consider (centered) Gaussian mixtures, i.e. random variables of the form

$$X = YZ, \tag{3}$$

where Y is an almost surely positive random variable and Z is a standard Gaussian random variable, independent of Y. The resulting random variable can be seen as a centered Gaussian with random variance Y^2 and has density of the form

$$\forall\ x \in \mathbb{R}, \qquad f_X(x) = \mathbb{E}\Big[\frac{1}{\sqrt{2\pi Y^2}} e^{-\frac{x^2}{2Y^2}}\Big]. \tag{4}$$

In particular, as observed in [18], (4) combined with Bernstein's theorem readily implies that a symmetric random variable X is a Gaussian mixture if and only if $x \mapsto f_X(\sqrt{x})$ is completely monotonic on $(0,\infty)$. Therefore, distributions with density proportional to $e^{-|x|^p}$, symmetric p-stable random variables, where $p \in (0,2]$, and the Cauchy distribution are Gaussian mixtures. Let us mention that Costa [16] also considered symmetric stable laws to prove a strengthened version of the entropy power inequality that fails in general.

Our first result proves the concavity of entropy conjectured in [3] for Gaussian mixtures.

Theorem 1. *Let X_1, X_2 be independent Gaussian mixtures. Then the function*

$$t \longmapsto h\Big(\sqrt{t}X_1 + \sqrt{1-t}X_2\Big) \tag{5}$$

is concave on the interval $[0,1]$.

Theorem 1 will be a straightforward consequence of a more general result for the Rényi entropy of a weighted sum of n Gaussian mixtures. Let $\triangle^{n-1}$ be the standard simplex in $\mathbb{R}^n$,

$$\triangle^{n-1} \stackrel{\text{def}}{=} \Big\{(\pi_1,\dots,\pi_n) \in [0,1]^n : \ \pi_1 + \cdots + \pi_n = 1\Big\}. \tag{6}$$

The Rényi entropy of order $\alpha \neq 1$ of a random vector X with density f is given by

$$h_\alpha(X) \stackrel{\text{def}}{=} \frac{1}{1-\alpha} \log\Big(\int_{\mathbb{R}^d} f^\alpha(x)\,\mathrm{d}x\Big), \tag{7}$$

and $h_1(X)$ is simply the Shannon entropy $h(X)$. We will prove the following general concavity.

Theorem 2. *Let $X_1,\dots,X_n$ be independent Gaussian mixtures. Then, the function*

$$\triangle^{n-1} \ni (a_1^2,\dots,a_n^2) \longmapsto h_\alpha\Big(\sum_{i=1}^n a_i X_i\Big) \tag{8}$$

is concave on $\triangle^{n-1}$ for every $\alpha \geq 1$.

When $n = 2$ and $\alpha = 1$, Theorem 2 reduces exactly to Theorem 1.

In [18, Theorem 8], it was shown that if $X_1,\dots,X_n$ are i.i.d., then the function (8) is *Schur concave*, namely that if $(a_1,\dots,a_n)$ and $(b_1,\dots,b_n)$ are two unit vectors in $\mathbb{R}^n$, then

$$(a_1^2,\dots,a_n^2) \preceq_{\mathrm{m}} (b_1^2,\dots,b_n^2) \quad \Longrightarrow \quad h_\alpha\Big(\sum_{i=1}^n a_i X_i\Big) \geq h_\alpha\Big(\sum_{i=1}^n b_i X_i\Big), \tag{9}$$

for any $\alpha \geq 1$, where $\preceq_{\mathrm{m}}$ is the majorisation ordering of vectors (see [18]). As the unit vector with all coordinates equal to $\frac{1}{n}$ is majorised by any other vector in $\triangle^{n-1}$, (9) implies that the function (8) achieves its maximum on the main diagonal for Gaussian mixtures.

As any permutationally invariant concave function is Schur concave (see [28, p. 97]), (9) follows from Theorem 2. On the other hand, the function $x_1 \cdots x_n$ is permutationally invariant and Schur concave on $\mathbb{R}_+^n$ (see [28, p. 115]) but it is evidently not concave on the hyperplane $x_1 + \cdots + x_n = 1$ when $n \geq 3$. Therefore, Theorem 2 is a strict refinement of [18, Theorem 8].

We note in passing that, while the conclusion of Theorem 1 has been conjectured in [3] to hold for every i.i.d. log-concave random variables X_1, X_2, the conclusion of Theorem 2 cannot hold for this class of variables. In [27], Madiman, Nayar and Tkocz constructed a symmetric log-concave random variable X for which the Schur concavity (9) does not hold for i.i.d. copies of X and thus, as a consequence of [28, p. 97], the concavity of Theorem 2 must also fail.

1.2. **Fisher Information.** Let X be a continuous random vector in $\mathbb{R}^d$ with smooth density $f : \mathbb{R}^d \to \mathbb{R}_+$. The Fisher information of X is the quantity

$$I(X) \stackrel{\text{def}}{=} \int_{\mathbb{R}^d} \frac{|\nabla f(x)|^2}{f(x)}\,\mathrm{d}x = \mathbb{E}\Big[\big|\rho(X)\big|^2\Big], \tag{10}$$

where $\rho(x) \stackrel{\text{def}}{=} \frac{\nabla f(x)}{f(x)}$ is the *score function* of X. Fisher information and entropy are connected by the classical de Bruijn identity (see, e.g., [31]), due to which most results for Fisher information are formally stronger than their entropic counterparts. In particular, the inequality

$$\forall\ t \in [0,1], \qquad \frac{1}{I(\sqrt{t}X_1 + \sqrt{1-t}X_2)} \geq \frac{t}{I(X_1)} + \frac{1-t}{I(X_2)} \tag{11}$$

of Blachman and Stam [31, 8], which holds for all independent random vectors X_1, X_2 in $\mathbb{R}^d$, implies the entropy power inequality (2). In the spirit of the question of Ball, Nayar and Tkocz [3] and of the result of [18], we raise the following problem.

Question 3. *Let $X_1, \ldots, X_n$ be i.i.d. Gaussian mixtures. For which unit vectors $(a_1, \ldots, a_n)$ in $\mathbb{R}^n$ is the Fisher information of $\sum_{i=1}^n a_i X_i$ minimised?*

While Question 3 still remains elusive, we shall now explain how to obtain some useful bounds for the Fisher information of mixtures. In order to state our results in the greatest possible generality, we consider random *vectors* which are mixtures of centered multivariate Gaussians. Recall that the Fisher information matrix of a random vector X in $\mathbb{R}^d$ is given by

$$\mathcal{I}(X)_{ij} \stackrel{\text{def}}{=} \int_{\mathbb{R}^d} \frac{\partial_i f(x) \partial_j f(x)}{f(x)}\, \mathrm{d}x, \tag{12}$$

where $f : \mathbb{R}^d \to \mathbb{R}_+$ is the smooth density of X, so that $I(X) = \mathrm{tr}\mathcal{I}(X)$.

Let $\mathcal{F}_d \subset L_1(\mathbb{R}^d)$ be the space of smooth probability densities on $\mathbb{R}^d$. By abuse of notation, we will also write $I(f)$ and $\mathcal{I}(f)$ to denote the Fisher information and Fisher information matrix respectively of a random vector with smooth density f on $\mathbb{R}^d$. In his recent treatise on estimates for the Fisher information, Bobkov made crucial use of the *convexity* of the Fisher information functional $I(X)$ as a function of the density of the random variable X, see [10, Proposition 15.2]. For our purposes we shall need the following matricial extension of this.

Proposition 4. *Fix $d \in \mathbb{N}$. If π is a Borel probability measure on $\mathcal{F}_d$, then*

$$\mathcal{I}\Big(\int_{\mathcal{F}_d} g\, \mathrm{d}\pi(g)\Big) \preceq \int_{\mathcal{F}_d} \mathcal{I}(g)\, \mathrm{d}\pi(g), \tag{13}$$

provided that $\int_{\mathcal{F}_d} \|\mathcal{I}(g)\|_{\mathrm{op}}\, \mathrm{d}\pi(g) < \infty$. Here $\preceq$ denotes the positive semi-definite ordering of matrices.

We propose the following definition of Gaussian mixtures in arbitrary dimension.

Definition 5. *A random vector X in $\mathbb{R}^d$ is a (centered) Gaussian mixture if X has the same distribution as $\mathbf{Y}Z$, where $\mathbf{Y}$ is a random symmetric $d \times d$ matrix which is almost surely positive definite and Z is a standard Gaussian random vector in $\mathbb{R}^d$, independent of $\mathbf{Y}$.*

As in the scalar case, a Gaussian mixture X in $\mathbb{R}^d$ has density of the form

$$\forall\ x \in \mathbb{R}^d, \qquad f_X(x) = \mathbb{E}\Big[\frac{1}{\det(\sqrt{2\pi}\mathbf{Y})} e^{-|\mathbf{Y}^{-1}x|^2/2}\Big]. \tag{14}$$

Employing Proposition 4 for Gaussian mixtures we deduce the following bound.

Corollary 6. *Fix $d \in \mathbb{N}$ and let X be a random vector in $\mathbb{R}^d$ admitting a Gaussian mixture representation $\mathbf{Y}Z$. Then, we have*

$$\mathcal{I}(X) \preceq \mathbb{E}\Big[(\mathbf{Y}\mathbf{Y}^T)^{-1}\Big]. \tag{15}$$

This upper bound should be contrasted with the general lower bound

$$\mathcal{I}(X) \succeq \mathrm{Cov}(X)^{-1} = \Big(\mathbb{E}\mathbf{Y}\mathbf{Y}^T\Big)^{-1}, \tag{16}$$

where the first inequality is the multivariate Crámer–Rao bound [7, Theorem 3.4.4].

1.2.1. *Quantitative CLT for the Fisher information matrix of Gaussian mixtures.* Equality in the Cramér–Rao bound (16) is attained if and only if X is Gaussian. The deficit in the scalar version of this inequality is the relative Fisher information $I(X\|Z)$ between X and Z and may be interpreted as a strong measure of distance of X from Gaussianity. In particular, in view of Gross' logarithmic Sobolev inequality [21] and Pinsker's inequality [29, 17, 23], closeness in relative Fisher information implies closeness in relative entropy and a fortiori in total variation distance. Therefore, a very natural question is under which conditions and with what rate the relative Fisher information of a weighted sum tends to zero along the central limit theorem, thus offering a strenthening of the entropic central limit theorem [4]. As an application of Corollary 6, we obtain a bound for a matrix analogue of the relative Fisher information of Gaussian mixtures. Here and throughout, $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm of a square matrix.

Theorem 7. *Fix $d \in \mathbb{N}$, $\delta \in (0,1]$ and let $X_1,\dots,X_n$ be i.i.d. random vectors in $\mathbb{R}^d$, each admitting a Gaussian mixture representation $\mathbf{Y}Z$ as above. Assume also that*

$$\mathbb{E}\left\|\mathbf{Y}\mathbf{Y}^T\right\|_{\mathrm{op}}^{1+\delta} < \infty \quad \textit{and} \quad \mathbb{E}\left\|\left(\mathbf{Y}\mathbf{Y}^T\right)^{-1}\right\|_{\mathrm{op}}^{1+\delta} < \infty. \tag{17}$$

Then, for every unit vector $a = (a_1,\dots,a_n)$ in $\mathbb{R}^n$ the weighted sum $S_n = \sum_{i=1}^n a_i X_i$ satisfies

$$\left\|\mathrm{Cov}(S_n)^{\frac{1}{2}}\mathcal{J}(S_n)\mathrm{Cov}(S_n)^{\frac{1}{2}} - \mathrm{I}_d\right\|_{\mathrm{op}} \le C(\mathbf{Y})\log^{\delta}(d+1)\|a\|_{2+2\delta}^{\frac{2\delta}{1+\delta}}, \tag{18}$$

where $C(\mathbf{Y})$ is a constant that depends only on the moments of $\|\mathbf{Y}\mathbf{Y}^T\|_{\mathrm{op}}$.

There is a vast literature on quantitative versions of the central limit theorem. The first to obtain efficient bounds for the relative Fisher information of weighted sums were Artstein, Ball, Barthe and Naor [1] (see also the work [22] of Johnson and Barron) who obtained a $O(\|a\|_4^4)$ upper bound on $I(S_n\|X)$, where $S_n = \sum_{i=1}^n a_i X_i$ for $X_1,\dots,X_n$ i.i.d. random variables satisfying a Poincaré inequality. In particular, this bound reduces to the sharp rate $O(\frac{1}{n})$ on the main diagonal. Following a series of works on the relative entropy of weighted sums [12, 13], Bobkov, Chistyakov and Götze investigated in [14] upper bounds for the relative Fisher information along the main diagonal under finite moment assumptions. More specifically, their main result asserts that if $\mathbb{E}|X_1|^s < \infty$ for some $s \in (2,4)$, then

$$I\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i \middle\| Z\right) = O\left(\frac{1}{n^{\frac{s-2}{2}+o(1)}}\right), \tag{19}$$

where the $n^{o(1)}$ term is a power of $\log n$, provided that the Fisher information of the sum is finite for some n. The exponent $\frac{s-2}{2}$ is sharp in this estimate. Moreover, it is also shown in [14] that if $\mathbb{E}X_1^4 < \infty$, then the relative Fisher information decays with the optimal $O(\frac{1}{n})$ rate of convergence. This is a far-reaching extension of the results of [1, 22] on the main diagonal as the Poincaré inequality assumption in particular implies finiteness of all moments.

The scalar version of Theorem 7 (corresponding to $d = 1$) is in various ways weaker than the results of [14]. Firstly, it applies only within the class of Gaussian mixtures and it requires the finiteness of a *negative* moment of the random variable besides a positive one. Moreover, even if these assumptions are satisfied, the bound (18) yields the rate $O(\frac{1}{n^{c_\delta}})$ with $c_\delta = \frac{\delta^2}{(1+\delta)^2}$ along the main diagonal if X has a finite $2+2\delta$ moment. This is weaker than the sharp $O(\frac{1}{n^{\delta+o(1)}})$ which follows from [14]. On the other hand, Theorem 7 applies to general coefficients beyond the main diagonal and, in contrast to [1, 22], does not require the finiteness of all positive moments. More importantly though, (18) is multi-dimensional bound with a *subpolynomial* dependence on the dimension d. To the best of our knowledge, this is the first such bound for the relative Fisher information *matrix* of a weighted sum and it would be very interesting to extend it to more general classes of random vectors and to obtain sharper rates.

The logarithmic dependence on the dimension in Theorem 7 is a consequence of a classical result of Tomczak-Jaegermann [32] on the uniform smoothness of Schatten classes. While

Theorem 7 is stated in terms of the operator norm, the proof yields an upper bound for any operator monotone matrix norm (see Remark 13) in terms of its Rademacher type constants.

Acknowledgements. We are grateful to Léonard Cadilhac for helpful discussions.

2. Concavity of entropy

This section is devoted to the proof of Theorem 2. We shall make use of the standard variational formula for entropy which asserts that if X is a continuous random variable, then

$$h(X) = \min\Big\{\mathbb{E}[-\log g(X)] : \ g : \mathbb{R} \to \mathbb{R}_+ \text{ is a density function}\Big\}. \tag{20}$$

Proof of Theorem 2. We start with the Shannon entropy, which corresponds to $\alpha = 1$. Fix two unit vectors $(a_1, \dots, a_n)$ and $(b_1, \dots, b_n)$ in $\mathbb{R}^n$. For $t \in [0,1]$, consider

$$X_t \stackrel{\text{def}}{=} \sum_{i=1}^n \sqrt{ta_i^2 + (1-t)b_i^2}\, X_i \qquad \text{and} \qquad f(t) \stackrel{\text{def}}{=} h(X_t), \tag{21}$$

and denote by $g_t : \mathbb{R} \to \mathbb{R}_+$ the density of X_t. The statement of the theorem is equivalent to the concavity of the function f on the interval $[0,1]$.

Let $\lambda, t_1, t_2 \in [0,1]$ and set $t = \lambda t_1 + (1-\lambda)t_2$. By the variational formula for entropy, we have

$$\begin{aligned} \lambda f(t_1) + (1-\lambda) f(t_2) &= \lambda \mathbb{E}[-\log g_{t_1}(X_{t_1})] + (1-\lambda)\mathbb{E}[-\log g_{t_2}(X_{t_2})] \\ &\le \lambda \mathbb{E}[-\log g_t(X_{t_1})] + (1-\lambda)\mathbb{E}[-\log g_t(X_{t_2})] \end{aligned} \tag{22}$$

Moreover, since X_i has the same distribution as the independent product $Y_i Z_i$, the stability of Gaussian measure implies the equality in distribution

$$X_t \stackrel{(\mathrm{d})}{=} \sqrt{\sum_{i=1}^n \Big(ta_i^2 + (1-t)b_i^2\Big) Y_i^2}\; Z. \tag{23}$$

Therefore, X_t is itself a Gaussian mixture. By the characterisation of [18, Theorem 2], this is equivalent to the complete monotonicity of the function $g_t(\sqrt{\cdot})$. Thus, by Bernstein's theorem, $g_t(\sqrt{\cdot})$ is the Laplace transform of a non-negative Borel measure on $(0,\infty)$ and therefore the function $\varphi_t \stackrel{\text{def}}{=} -\log g_t(\sqrt{\cdot})$ is concave on $(0,\infty)$. Hence, by (22) and (23), we have

$$\begin{aligned} &\lambda f(t_1) + (1-\lambda) f(t_2) \\ &\quad \le \lambda \mathbb{E}\Big[\varphi_t\Big(\sum_{i=1}^n \Big(t_1 a_i^2 + (1-t_1)b_i^2\Big) Y_i^2 Z^2\Big)\Big] + (1-\lambda)\mathbb{E}\Big[\varphi_t\Big(\sum_{i=1}^n \Big(t_2 a_i^2 + (1-t_2)b_i^2\Big) Y_i^2 Z^2\Big)\Big] \\ &\quad \le \mathbb{E}\Big[\varphi_t\Big(\sum_{i=1}^n \Big(\lambda\Big(t_1 a_i^2 + (1-t_1)b_i^2\Big) + (1-\lambda)\Big(t_2 a_i^2 + (1-t_2)b_i^2\Big)\Big) Y_i^2 Z^2\Big)\Big] \\ &\quad = \mathbb{E}\Big[\varphi_t\Big(\sum_{i=1}^n \Big(ta_i^2 + (1-t)b_i^2\Big) Y_i^2 Z^2\Big)\Big] = \mathbb{E}\Big[-\log g_t\Big(\sum_{i=1}^n \sqrt{ta_i^2 + (1-t)b_i^2}\, X_i\Big)\Big] = f(t). \end{aligned} \tag{24}$$

This completes the proof of the concavity of Shannon entropy.

Next, let $\alpha > 1$ and consider again $t = \lambda t_1 + (1-\lambda)t_2$. Denoting by $\psi_t = g_t^{\alpha-1}(\sqrt{\cdot})$ and applying the same reasoning, we get

$$\begin{aligned} \int_{\mathbb{R}} g_t^\alpha(x)\,\mathrm{d}x &= \int_{\mathbb{R}} g_t(x) g_t^{\alpha-1}(x)\,\mathrm{d}x = \mathbb{E} g_t^{\alpha-1}(X_t) \\ &= \mathbb{E}\Big[\psi_t\Big(\sum_{i=1}^n \Big(\lambda\Big(t_1 a_i^2 + (1-t_1)b_i^2\Big) + (1-\lambda)\Big(t_2 a_i^2 + (1-t_2)b_i^2\Big)\Big) Y_i^2 Z^2\Big)\Big]. \end{aligned} \tag{25}$$

Now $\psi_t = e^{-(\alpha-1)\varphi_t}$ is log-convex and thus

$$\begin{aligned}\int_{\mathbb{R}} g_t^\alpha(x)\,\mathrm{d}x &\le \mathbb{E}\Big[\psi_t^\lambda\Big(\sum_{i=1}^n \big(t_1 a_i^2 + (1-t_1)b_i^2\big)Y_i^2 Z^2\Big)\psi_t^{1-\lambda}\Big(\sum_{i=1}^n \big(t_2 a_i^2 + (1-t_2^2)b_i^2\big)Y_i^2 Z^2\Big)\Big] \\ &\le \mathbb{E}\Big[g_t^{\alpha-1}(X_{t_1})\Big]^\lambda \mathbb{E}\Big[g_t^{\alpha-1}(X_{t_2})\Big]^{1-\lambda}\end{aligned} \tag{26}$$

by Hölder's inequality and (23). By two more applications of Hölder's inequality, we get

$$\int_{\mathbb{R}} g_{t_1}(x) g_t(x)^{\alpha-1}\,\mathrm{d}x \le \Big(\int_{\mathbb{R}} g_{t_1}^\alpha(x)\,\mathrm{d}x\Big)^{\frac{1}{\alpha}}\Big(\int_{\mathbb{R}} g_t^\alpha(x)\,\mathrm{d}x\Big)^{\frac{\alpha-1}{\alpha}} \tag{27}$$

and

$$\int_{\mathbb{R}} g_{t_2}(x) g_t(x)^{\alpha-1}\,\mathrm{d}x \le \Big(\int_{\mathbb{R}} g_{t_2}^\alpha(x)\,\mathrm{d}x\Big)^{\frac{1}{\alpha}}\Big(\int_{\mathbb{R}} g_t^\alpha(x)\,\mathrm{d}x\Big)^{\frac{\alpha-1}{\alpha}}. \tag{28}$$

Combining (26), (27) and (28) we thus obtain

$$\Big(\int_{\mathbb{R}} g_t^\alpha(x)\,\mathrm{d}x\Big)^{\frac{1}{\alpha}} \le \Big(\int_{\mathbb{R}} g_{t_1}^\alpha(x)\,\mathrm{d}x\Big)^{\frac{\lambda}{\alpha}}\Big(\int_{\mathbb{R}} g_{t_2}^\alpha(x)\,\mathrm{d}x\Big)^{\frac{1-\lambda}{\alpha}} \tag{29}$$

which is exactly the claimed concavity of Rényi entropy. □

Remark 8. *One may wonder whether Theorem 2 can be extended to Gaussian mixtures on $\mathbb{R}^d$ in the sense of Definition 5. A repetition of the above argument in this setting would require the validity of the inequality*

$$\forall\ \lambda \in (0,1), \qquad g\Big(\sqrt{\lambda A + (1-\lambda)B}z\Big) \le g\Big(\sqrt{A}z\Big)^\lambda g\Big(\sqrt{B}z\Big)^{1-\lambda} \tag{30}$$

where $g : \mathbb{R}^d \to \mathbb{R}_+$ is the density of a Gaussian mixture, A and B are positive semidefinite $d \times d$ matrices and z is a vector in $\mathbb{R}^d$. The validity of (30) *for a Gaussian density with arbitrary covariance is equivalent to the operator concavity of the matrix function*

$$f(X) \stackrel{\text{def}}{=} \sqrt{X}Y\sqrt{X} \tag{31}$$

for an arbitrary positive semidefinite matrix Y. The following counterexample to this statement was communicated to us by Léonard Cadilhac. As the function f takes values in the cone of positive semidefinite matrices, operator concavity is equivalent to operator monotonicity (see the proof of [6, Theorem V.2.5]*). Take two non-negative matrices A, Y such that $Y \preceq A$ but $Y^2 \not\preceq A^2$. Then, the corresponding function $f(X) = \sqrt{X}Y\sqrt{X}$ satisfies $f(Y) = Y^2$ and $f(A) = \sqrt{A}Y\sqrt{A} \preceq A^2$ since $Y \preceq A$. Therefore, $f(Y) \not\preceq f(A)$ and thus f is not operator monotone or concave.*

3. Convexity of Fisher information

3.1. **Warm-up: the Fisher information of independent products.** Before showing the general argument which leads to Proposition 4, we present a short proof for the case of mixtures of dilates of a *fixed* distribution which corresponds exactly to the Fisher information of a product of independent random variables. As this is a special case of Bobkov's [14, Proposition 15.2], we shall disregard rigorous integrability assumptions for the sake of simplicity of exposition.

Theorem 9. *Let W be a random variable with zero mean and smooth-enough density and let Y be an independent positive random variable. Then,*

$$\frac{1}{\mathbb{E}Y^2 \mathrm{Var}(W)} \le I(YW) \le \mathbb{E}\Big[\frac{I(W)}{Y^2}\Big]. \tag{32}$$

Proof. The first inequality is the Cramér-Rao lower bound. Suppose that W has density e^{-V} with V nice enough. Then, X has density

$$f(x) \stackrel{\text{def}}{=} \mathbb{E}\Big[\frac{1}{Y}e^{-V\left(\frac{x}{Y}\right)}\Big] \tag{33}$$

and thus, differentiating under the expectation and using Cauchy–Schwarz, we get

$$f'(x)^2 = \mathbb{E}\Big[\frac{V'(\frac{x}{Y})}{Y^2}e^{-V\left(\frac{x}{Y}\right)}\Big]^2 \le \mathbb{E}\Big[\frac{1}{Y}e^{-V\left(\frac{x}{Y}\right)}\Big]\mathbb{E}\Big[\frac{V'(\frac{x}{Y})^2}{Y^3}e^{-V\left(\frac{x}{Y}\right)}\Big] = f(x)\,\mathbb{E}\Big[\frac{V'(\frac{x}{Y})^2}{Y^3}e^{-V\left(\frac{x}{Y}\right)}\Big]. \tag{34}$$

Thus,

$$\begin{aligned} I(X) = \int_{\mathbb{R}} \frac{f'(x)^2}{f(x)}\,\mathrm{d}x &\le \int_{\mathbb{R}} \mathbb{E}\Big[\frac{V'(\frac{x}{Y})^2}{Y^3}e^{-V\left(\frac{x}{Y}\right)}\Big]\,\mathrm{d}x = \mathbb{E}\Big[\frac{1}{Y^2}\int_{\mathbb{R}}\frac{V'(\frac{x}{Y})^2}{Y}e^{-V\left(\frac{x}{Y}\right)}\,\mathrm{d}x\Big] \\ &= \mathbb{E}\Big[\frac{1}{Y^2}\Big]\mathbb{E}\Big[V'(W)^2\Big] = \mathbb{E}\Big[\frac{I(W)}{Y^2}\Big]. \end{aligned}$$

□

3.2. **Proof of Proposition 4.** We start by proving the two-point convexity of $\mathcal{I}$.

Proposition 10. *The Fisher information matrix is operator convex on $\mathcal{F}_d$, that is, for $f_1, f_2 \in \mathcal{F}_d$,*

$$\forall\ \theta \in [0,1], \qquad \mathcal{I}\Big(\theta f_1 + (1-\theta) f_2\Big) \preceq \theta \mathcal{I}(f_1) + (1-\theta)\mathcal{I}(f_2). \tag{35}$$

Proof. First we claim that the function $R : \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}^{d\times d}$, given by $R(x,\lambda) = \frac{xx^T}{\lambda}$ is jointly operator convex. To prove this, we need to show that for every $\theta \in (0,1)$, $x, y \in \mathbb{R}^d$ and $\lambda, \mu > 0$,

$$R(\theta x + (1-\theta)y, \theta\lambda + (1-\theta)\mu) \preceq \theta R(x,\lambda) + (1-\theta)R(y,\mu). \tag{36}$$

After rearranging, this can be rewritten as

$$\theta(1-\theta)\Big(\lambda^2 xx^T + \mu^2 yy^T - \lambda\mu xy^T - \lambda\mu yx^T\Big) \succeq 0, \tag{37}$$

which is true since it is equivalent to $(\lambda x - \mu y)(\lambda x - \mu y)^T \succeq 0$.

Since the Fisher information matrix can be written as

$$\mathcal{I}(f) = \int_{\mathbb{R}^d} R\Big(\nabla f(x), f(x)\Big)\,\mathrm{d}x, \tag{38}$$

the conclusion follows by the convexity of R and the linearity of ∇ and $\int$. □

In order to derive the general Jensen inequality of Proposition 4 from Proposition 10, we will use a somewhat involved compactness argument that was invoked in [14]. We point out that these intricacies arise since the space $\mathcal{F}_d$ of smooth densities in $\mathbb{R}^d$ is infinite-dimensional. As our argument shares similarities with Bobkov's, we shall only point out the necessary modifications which need to be implemented. We stard by proving the following technical lemma.

Lemma 11. *Let $X, \{X_k\}_{k\ge1}$ be random vectors in $\mathbb{R}^d$ such that $X_k \Rightarrow X$ weakly.*

(i) If $\sup_k \big\|\mathcal{I}(X_k)\big\|_{\mathrm{op}} < \infty$, then for every $x \in \mathbb{S}^{d-1}$,

$$\langle \mathcal{I}(X)x, x\rangle \le \liminf_{k\to\infty} \langle \mathcal{I}(X_k)x, x\rangle. \tag{39}$$

(ii) Moreover, we always have

$$\big\|\mathcal{I}(X)\big\|_{\mathrm{op}} \le \liminf_{k\to\infty} \big\|\mathcal{I}(X_k)\big\|_{\mathrm{op}}. \tag{40}$$

Proof. We start with (39). It clearly suffices to show that any subsequence of $\{X_k\}$ has a further subsequence for which the conclusion holds. If $\big\|\mathcal{I}(X_k)\big\|_{\mathrm{op}} \le I < \infty$ for all $k \ge 1$, then

$$I(X_k) = \mathrm{tr}(\mathcal{I}(X_k)) \le d\big\|\mathcal{I}(X_k)\big\|_{\mathrm{op}} \le dI < \infty. \tag{41}$$

Write f_k and f for the densities of X_k and X respectively. Choose and fix any subsequence of $\{f_k\}$. By the proof of [10, Proposition 14.2], using the boundedness of Fisher informations, there is a further subsequence, say f_{k_j}, for which $f_{k_j} \to f$ and $\nabla f_{k_j} \to \nabla f$ a.e. as $j \to \infty$. Therefore

$$\lim_{j\to\infty}\Big\langle \frac{\nabla f_{k_j}(u)\nabla f_{k_j}(u)^T}{f_{k_j}(u)}x, x\Big\rangle \mathbb{I}_{\{f_{k_j}(u)>0\}} = \Big\langle \frac{\nabla f(u)\nabla f(u)^T}{f(u)}x, x\Big\rangle \mathbb{I}_{\{f(u)>0\}} \tag{42}$$

for almost every u. Integration with respect to u, linearity and Fatou's lemma yield (39).

To prove (40), fix a subsequence X_{k_j} for which the lim inf in (40) is attained. Then the subsequence satisfies $\sup_j \|\mathcal{I}(X_{k_j})\|_{\mathrm{op}} < \infty$ and thus by (39) for every $x \in \mathbb{S}^{d-1}$ we have

$$\langle \mathcal{I}(X)x, x\rangle \le \liminf_{j\to\infty} \langle \mathcal{I}(X_{k_j})x, x\rangle \le \liminf_{j\to\infty} \left\|\mathcal{I}(X_{k_j})\right\|_{\mathrm{op}} = \liminf_{k\to\infty} \left\|\mathcal{I}(X_k)\right\|_{\mathrm{op}}. \tag{43}$$

Taking a supremum over $x \in \mathbb{S}^{d-1}$ concludes the proof as $\mathcal{I}(X)$ is positive semi-definite. □

Equipped with the lower semi-continuity of $\mathcal{I}$, we proceed to the main part of the proof.

Proof of Proposition 4. Inequality (35) may be extended to arbitrary finite mixtures by induction, that is if $p_1, \dots, p_N \ge 0$ satisfy $\sum_{i=1}^n p_i = 1$, then

$$\mathcal{I}\Big(\sum_{i=1}^N p_i f_i\Big) \preceq \sum_{i=1}^N p_i \mathcal{I}(f_i). \tag{44}$$

We need to extend (44) to arbitrary mixtures. We write $\mathcal{F}_d(I) = \{f \in \mathcal{F}_d : \left\|\mathcal{I}(f)\right\|_{\mathrm{op}} \le I\}$ and $\mathcal{F}_d(\infty) = \cup_I \mathcal{F}_d(I)$. By the assumption $\int_{\mathcal{F}_d} \|\mathcal{I}(g)\|_{\mathrm{op}}\, \mathrm{d}\pi(g) < \infty$, we deduce that the measure π is supported on $\mathcal{F}_d(\infty)$. We shall prove that

$$\forall x \in \mathbb{S}^{d-1}, \qquad \left\langle \mathcal{I}\Big(\int_{\mathcal{F}_d} g\, \mathrm{d}\pi(g)\Big)x, x\right\rangle \le \int_{\mathcal{F}_d} \langle \mathcal{I}(g)x, x\rangle\, \mathrm{d}\pi(g). \tag{45}$$

Fix $x \in \mathbb{S}^{d-1}$ and $I \in \mathbb{N}$. By the operator convexity of the Fisher information matrix (Proposition 10), the functional

$$f \to \langle \mathcal{I}(f)x, x\rangle \tag{46}$$

is convex and by Lemma 11 lower semi-continuous on $\mathcal{F}_d(I)$. Again by operator convexity, the set $\mathcal{F}_d(I)$ is convex and by Lemma 11 it is closed. Now we may repeat exactly the same proof as in [10, Proposition 15.1, Steps 1-2], but working with the functional $\langle \mathcal{I}(f)x, x\rangle$ instead of the Fisher information $I(f)$, to obtain (45) if the measure π is supported on $\mathcal{F}_d(I)$.

To derive inequality (45) in general, fix I_0 large enough such that $\pi(\mathcal{F}_d(I_0)) > \frac{1}{2}$ and for $I \ge I_0$ write the inequality (45) for the restriction of π to $\mathcal{F}_d(I)$, namely

$$\left\langle \mathcal{I}\Big(\frac{1}{\pi(\mathcal{F}_d(I))}\int_{\mathcal{F}_d(I)} g\, \mathrm{d}\pi(g)\Big)x, x\right\rangle \le \frac{1}{\pi(\mathcal{F}_d(I))}\int_{\mathcal{F}_d(I)} \langle \mathcal{I}(g)x, x\rangle\, \mathrm{d}\pi(g). \tag{47}$$

Denoting by f_I the density on the left-hand side of the inequality, we have that f_I converges weakly to the density $\int_{\mathcal{F}_d} g\, \mathrm{d}\pi(g)$ as $I \to \infty$ and moreover (47) yields

$$\forall\ I \ge I_0, \qquad \left\|\mathcal{I}(f_I)\right\|_{\mathrm{op}} \le \frac{1}{\pi(\mathcal{F}_d(I))}\int_{\mathcal{F}_d(I)} \left\|\mathcal{I}(g)\right\|_{\mathrm{op}}\, \mathrm{d}\pi(g) \le 2\int_{\mathcal{F}_d} \left\|\mathcal{I}(g)\right\|_{\mathrm{op}}\, \mathrm{d}\pi(g) < \infty. \tag{48}$$

Therefore, the assumptions of (39) are satisfied for $\{f_I\}_{I\ge I_0}$ and thus

$$\begin{aligned}\left\langle \mathcal{I}\Big(\int_{\mathcal{F}_d} g\, \mathrm{d}\pi(g)\Big)x, x\right\rangle &\le \liminf_{I\to\infty}\left\langle \mathcal{I}\Big(\frac{1}{\pi(\mathcal{F}_d(I))}\int_{\mathcal{F}_d(I)} g\, \mathrm{d}\pi(g)\Big)x, x\right\rangle \\ &\overset{(47)}{\le} \liminf_{I\to\infty} \frac{1}{\pi(\mathcal{F}_d(I))}\int_{\mathcal{F}_d(I)} \langle \mathcal{I}(g)x, x\rangle\, \mathrm{d}\pi(g) = \int_{\mathcal{F}_d} \langle \mathcal{I}(g)x, x\rangle\, \mathrm{d}\pi(g),\end{aligned} \tag{49}$$

and this concludes the proof. □

Proof of Corollary 6. In view of (14) and Proposition 4, we have

$$\mathcal{I}(\mathbf{Y}Z) = \mathcal{I}\left(\mathbb{E}_{\mathbf{Y}}\Big[\frac{1}{\det(\sqrt{2\pi}\mathbf{Y})} e^{-|\mathbf{Y}^{-1}\cdot|^2/2}\Big]\right) \preceq \mathbb{E}_{\mathbf{Y}}\Big[\mathcal{I}(\mathbf{Y}Z)\Big] = \mathbb{E}\Big[(\mathbf{Y}\mathbf{Y}^T)^{-1}\Big], \tag{50}$$

since the Fisher information matrix of a Gaussian vector with covariance matrix Σ is Σ^{-1}. □

4. CLT for the Fisher information matrix

Before delving into the proof of Theorem 7, we shall discuss some geometric preliminaries. Recall that a normed space $(V, \|\cdot\|_V)$ has Rademacher type $p \in [1,2]$ with constant $T \in (0,\infty)$ if for every $n \in \mathbb{N}$ and every $v_1, \ldots, v_n \in V$, we have

$$\frac{1}{2^n} \sum_{\varepsilon \in \{-1,1\}^n} \Big\| \sum_{i=1}^n \varepsilon_i v_i \Big\|_V^p \leq T^p \sum_{i=1}^n \|v_i\|_V^p. \tag{51}$$

The least constant T for which this inequality holds will be denoted by $\mathsf{T}_p(V)$. A standard symmetrisation argument (see, for instance, [24, Proposition 9.11]) shows that for any $n \in \mathbb{N}$ and any V-valued random vectors $V_1, \ldots, V_n$ with $\mathbb{E}[V_i] = 0$ we have

$$\mathbb{E}\Big\| \sum_{i=1}^n V_i \Big\|_V^p \leq \big(2\mathsf{T}_p(V)\big)^p \sum_{i=1}^n \mathbb{E}\|V_i\|_V^p. \tag{52}$$

We denote by $\mathsf{M}_d(\mathbb{R})$ the vector space of all $d \times d$ matrices with real entries. We shall consider the p-Schatten trace class S_p^d of $d \times d$ matrices. This is the normed space $\mathsf{S}_p^d = (\mathsf{M}_d(\mathbb{R}), \|\cdot\|_{\mathsf{S}_p})$, where for a $d \times d$ real matrix A, we denote

$$\|A\|_{\mathsf{S}_p} \stackrel{\text{def}}{=} \Big(\sum_{i=1}^d \sigma_i(A)^p \Big)^{1/p} \tag{53}$$

and by $\sigma_1(A) \geq \cdots \geq \sigma_d(A)$ the singular values of A. Evidently, $\|\cdot\|_{\mathrm{op}} = \|\cdot\|_{\mathsf{S}_\infty}$. A classical result of Tomczak-Jaegermann [32] (see also [2] for the exact values of the constants) asserts that if $p \in [1,2]$, then S_p^d has Rademacher type p constant $T_p\big(\mathsf{S}_p^d\big) = 1$ and if $p \geq 2$, then S_p^d has Rademacher type 2 constant $\mathsf{T}_2\big(\mathsf{S}_p^d\big) \leq \sqrt{p-1}$. We shall use the following consequence of this.

Lemma 12. *Fix $n, d \in \mathbb{N}$ and let $W_1, \ldots, W_n$ be i.i.d. random $d \times d$ matrices with $\mathbb{E}[W_i] = 0$. For any $\delta \in (0,1]$ and any vector $b = (b_1, \ldots, b_n) \in \mathbb{R}^n$, we have*

$$p \in [2,\infty) \qquad \Longrightarrow \qquad \mathbb{E}\Big\| \sum_{i=1}^n b_i W_i \Big\|_{\mathsf{S}_p}^{1+\delta} \leq 2^{1+\delta}(p-1)^\delta \mathbb{E}\Big[\|W_1\|_{\mathsf{S}_p}^{1+\delta}\Big] \|b\|_{1+\delta}^{1+\delta} \tag{54}$$

and

$$p \in [1+\delta, 2] \qquad \Longrightarrow \qquad \mathbb{E}\Big\| \sum_{i=1}^n b_i W_i \Big\|_{\mathsf{S}_p}^{1+\delta} \leq 2^{1+\delta} \mathbb{E}\Big[\|W_1\|_{\mathsf{S}_p}^{1+\delta}\Big] \|b\|_{1+\delta}^{1+\delta}. \tag{55}$$

Moreover,

$$\mathbb{E}\Big\| \sum_{i=1}^n b_i W_i \Big\|_{\mathrm{op}}^{1+\delta} \leq (2e)^{1+\delta} \log^\delta(d+1) \mathbb{E}\Big[\|W_1\|_{\mathrm{op}}^{1+\delta}\Big] \|b\|_{1+\delta}^{1+\delta} \tag{56}$$

Proof. We first prove (54). In view of inequality (52), it suffices to prove that the Rademacher type $(1+\delta)$-constant of S_p^d satisfies $\mathsf{T}_{1+\delta}(\mathsf{S}_p^d) \leq (p-1)^{\frac{\delta}{1+\delta}}$. Given a normed space $(X, \|\cdot\|_X)$ and $n \in \mathbb{N}$, consider the linear operator $T_n : \ell_p^n(X) \to L_p(\{-1,1\}^n; X)$ given by

$$\forall\ x = (x_1, \ldots, x_n) \in \ell_p^n(X), \qquad [T_n x](\varepsilon) = \sum_{i=1}^n \varepsilon_i x_i, \tag{57}$$

where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n) \in \{-1,1\}^n$. Then, it follows from (51) that

$$T_p(X) = \sup_{n \in \mathbb{N}} \big\| T_n \big\|_{\ell_p^n(X) \to L_p(\{-1,1\}^n; X)}. \tag{58}$$

In fact, if X is finite-dimensional (like S_p^d) then it was shown in [20, Lemma 6.1] that the supremum is attained for some $n \leq \dim(X)(\dim(X)+1)/2$. Either way, by complex interpolation

of vector-valued L_p spaces (see [5, Section 5.6]), we thus deduce that

$$\mathsf{T}_{1+\delta}\big(\mathsf{S}_p^d\big) \le \mathsf{T}_1\big(\mathsf{S}_p^d\big)^{\theta}\mathsf{T}_2\big(\mathsf{S}_p^d\big)^{1-\theta}, \tag{59}$$

where $\frac{\theta}{1}+\frac{1-\theta}{2}=\frac{1}{1+\delta}$. The conclusion of (54) follows by plugging-in the value of θ and the result of [32, 2]. The proof of inequality (55) is similar, interpolating between 1 and p.

Finally, to deduce (56), note that for any $A \in \mathsf{M}_d(\mathbb{R})$,

$$\|A\|_{\mathrm{op}} \le \|A\|_{\mathsf{S}_p} \le d^{1/p}\|A\|_{\mathrm{op}} \tag{60}$$

and thus plugging $p=\log(d+1)+1$ in (54) we derive the desired inequality. □

Equipped with these inequalities, we can now proceed to the main part of the proof.

Proof of Theorem 7. Since $\mathbb{E}S_n=0$ and $\mathrm{Cov}(S_n)=\mathbb{E}\mathbf{Y}\mathbf{Y}^T$, we have

$$\Big\|\mathrm{Cov}(S_n)^{\frac12}\mathcal{I}(S_n)\mathrm{Cov}(S_n)^{\frac12}-\mathrm{I}_d\Big\|_{\mathrm{op}} \le \Big\|\mathbb{E}\mathbf{Y}\mathbf{Y}^T\Big\|_{\mathrm{op}}\Big\|\mathcal{I}(S_n)-\Big(\mathbb{E}\mathbf{Y}\mathbf{Y}^T\Big)^{-1}\Big\|_{\mathrm{op}}, \tag{61}$$

using that for any PSD matrices A,B, $\big\|AB\big\|_{\mathrm{op}} \le \big\|A\big\|_{\mathrm{op}}\big\|B\big\|_{\mathrm{op}}$ and $\big\|A^{\frac12}\big\|_{\mathrm{op}}=\big\|A\big\|_{\mathrm{op}}^{\frac12}$. Now, S_n is a Gaussian mixture itself and it satisfies

$$S_n=\sum_{i=1}^n a_i\mathbf{Y}_iZ_i \overset{(\mathrm{d})}{=} \Big(\sum_{i=1}^n a_i^2\mathbf{Y}_i\mathbf{Y}_i^T\Big)^{1/2}Z, \tag{62}$$

Corollary 6 yields the estimate

$$\mathcal{I}(S_n) \preceq \mathbb{E}\Big(\sum_{i=1}^n a_i^2\mathbf{Y}_i\mathbf{Y}_i^T\Big)^{-1}. \tag{63}$$

Moreover by the multivariate Cramér-Rao lower bound [7, Theorem 3.4.4], we have

$$\mathcal{I}(S_n) \succeq \Big(\mathbb{E}\mathbf{Y}\mathbf{Y}^T\Big)^{-1} \tag{64}$$

and thus the matrix in the right-hand side of (61) is positive semi-definite. Therefore, since $\|\cdot\|_{\mathrm{op}}$ is increasing with respect to the matrix ordering on positive matrices, (63) and (64) yield

$$\Big\|\mathcal{I}(S_n)-\Big(\mathbb{E}\mathbf{Y}\mathbf{Y}^T\Big)^{-1}\Big\|_{\mathrm{op}} \le \Big\|\mathbb{E}\Big(\sum_{i=1}^n a_i^2\mathbf{Y}_i\mathbf{Y}_i^T\Big)^{-1}-\Big(\mathbb{E}\mathbf{Y}\mathbf{Y}^T\Big)^{-1}\Big\|_{\mathrm{op}}. \tag{65}$$

For $i=1,\dots,n$ consider the i.i.d. random matrices $W_i \overset{\mathrm{def}}{=} \mathbf{Y}_i\mathbf{Y}_i^T-\mathbb{E}\mathbf{Y}\mathbf{Y}^T$ and denote the event $E_\varepsilon \overset{\mathrm{def}}{=} \Big\{\Big\|\sum_{i=1}^n a_i^2W_i\Big\|_{\mathrm{op}} \le \varepsilon\Big\}$. To bound the probability of the complement of E_ε, notice that

$$\begin{aligned}\mathbb{P}\{E_\varepsilon^{\mathrm{c}}\} = \mathbb{P}\Big\{\Big\|\sum_{i=1}^n a_i^2W_i\Big\|_{\mathrm{op}}^{1+\delta} > \varepsilon^{1+\delta}\Big\} &\le \frac{1}{\varepsilon^{1+\delta}}\mathbb{E}\Big\|\sum_{i=1}^n a_i^2W_i\Big\|_{\mathrm{op}}^{1+\delta} \\ &\overset{(56)}{\le} \Big(\frac{2e}{\varepsilon}\Big)^{1+\delta}\log^{\delta}(d+1)\mathbb{E}\Big[\|W_1\|_{\mathrm{op}}^{1+\delta}\Big]\,\big\|a\big\|_{2+2\delta}^{2+2\delta}.\end{aligned} \tag{66}$$

Moreover, since $\mathbb{E}\|W_1\|_{\mathrm{op}}^{1+\delta} \le 2^{1+\delta}\mathbb{E}\|\mathbf{Y}\mathbf{Y}^T\|_{\mathrm{op}}^{1+\delta}$,we get the bound

$$\mathbb{P}\{E_\varepsilon^{\mathrm{c}}\} \le \Big(\frac{4e}{\varepsilon}\Big)^{1+\delta}\log^{\delta}(d+1)\mathbb{E}\Big[\|\mathbf{Y}\mathbf{Y}^{\mathbf{T}}\|_{\mathrm{op}}^{1+\delta}\Big]\,\big\|a\big\|_{2+2\delta}^{2+2\delta}. \tag{67}$$

Next, we write

$$\mathbb{E}\Big(\sum_{i=1}^n a_i^2\mathbf{Y}_i\mathbf{Y}_i^T\Big)^{-1} = \mathbb{E}\Big[\Big(\sum_{i=1}^n a_i^2\mathbf{Y}_i\mathbf{Y}_i^T\Big)^{-1}\mathbb{I}_{E_\varepsilon}\Big]+\mathbb{E}\Big[\Big(\sum_{i=1}^n a_i^2\mathbf{Y}_i\mathbf{Y}_i^T\Big)^{-1}\mathbb{I}_{E_\varepsilon^{\mathrm{c}}}\Big] \tag{68}$$

and use the triangle inequality to get

$$\begin{aligned}\Big\|\mathbb{E}\Big(\sum_{i=1}^n a_i^2\mathbf{Y}_i\mathbf{Y}_i^T\Big)^{-1}-\big(\mathbb{E}\mathbf{Y}\mathbf{Y}^T\big)^{-1}\Big\|_{\mathrm{op}} &\le \Big\|\mathbb{E}\Big[\Big(\Big(\sum_{i=1}^n a_i^2\mathbf{Y}_i\mathbf{Y}_i^T\Big)^{-1}-\big(\mathbb{E}\mathbf{Y}\mathbf{Y}^T\big)^{-1}\Big)\mathbb{I}_{E_\varepsilon}\Big]\Big\|_{\mathrm{op}}\\ &\quad+\mathbb{P}\{E_\varepsilon^{\mathrm{c}}\}\,\big\|\big(\mathbb{E}\mathbf{Y}\mathbf{Y}^T\big)^{-1}\big\|_{\mathrm{op}}+\Big\|\mathbb{E}\Big[\Big(\sum_{i=1}^n a_i^2\mathbf{Y}_i\mathbf{Y}_i^T\Big)^{-1}\mathbb{I}_{E_\varepsilon^{\mathrm{c}}}\Big]\Big\|_{\mathrm{op}}.\end{aligned}\tag{69}$$

To control the first term in (69), we use Jensen's inequality for $\|\cdot\|_{\mathrm{op}}$ to get

$$\begin{aligned}\Big\|\mathbb{E}\Big[\Big(\Big(\sum_{i=1}^n a_i^2\mathbf{Y}_i\mathbf{Y}_i^T\Big)^{-1}-\big(\mathbb{E}\mathbf{Y}\mathbf{Y}^T\big)^{-1}\Big)\mathbb{I}_{E_\varepsilon}\Big]\Big\|_{\mathrm{op}} &\le \mathbb{E}\Big[\Big\|\Big(\sum_{i=1}^n a_i^2\mathbf{Y}_i\mathbf{Y}_i^T\Big)^{-1}-\big(\mathbb{E}\mathbf{Y}\mathbf{Y}^T\big)^{-1}\Big\|_{\mathrm{op}}\mathbb{I}_{E_\varepsilon}\Big]\\ &\le \big\|\big(\mathbb{E}\mathbf{Y}\mathbf{Y}^T\big)^{-1}\big\|_{\mathrm{op}}\mathbb{E}\Big[\Big\|\Big(\sum_{i=1}^n a_i^2\mathbf{Y}_i\mathbf{Y}_i^T\Big)^{-1}\Big\|_{\mathrm{op}}\Big\|\sum_{i=1}^n a_i^2\mathbf{Y}_i\mathbf{Y}_i^T-\mathbb{E}\mathbf{Y}\mathbf{Y}^T\Big\|_{\mathrm{op}}\mathbb{I}_{E_\varepsilon}\Big]\\ &= \big\|\big(\mathbb{E}\mathbf{Y}\mathbf{Y}^T\big)^{-1}\big\|_{\mathrm{op}}\mathbb{E}\Big[\Big\|\Big(\sum_{i=1}^n a_i^2\mathbf{Y}_i\mathbf{Y}_i^T\Big)^{-1}\Big\|_{\mathrm{op}}\Big\|\sum_{i=1}^n a_i^2W_i\Big\|_{\mathrm{op}}\mathbb{I}_{E_\varepsilon}\Big],\end{aligned}\tag{70}$$

where the second line follows from the inequality $\|X^{-1}-Y^{-1}\|_{\mathrm{op}}\le\|X^{-1}\|_{\mathrm{op}}\|Y^{-1}\|_{\mathrm{op}}\|X-Y\|_{\mathrm{op}}$ for positive matrices X,Y. Now, by the definition of the event E_ε the last factor is at most ε and thus we derive the bound

$$\Big\|\mathbb{E}\Big[\Big(\Big(\sum_{i=1}^n a_i^2\mathbf{Y}_i\mathbf{Y}_i^T\Big)^{-1}-\big(\mathbb{E}\mathbf{Y}\mathbf{Y}^T\big)^{-1}\Big)\mathbb{I}_{E_\varepsilon}\Big]\Big\|_{\mathrm{op}}\le\big\|\big(\mathbb{E}\mathbf{Y}\mathbf{Y}^T\big)^{-1}\big\|_{\mathrm{op}}\mathbb{E}\Big[\Big\|\Big(\sum_{i=1}^n a_i^2\mathbf{Y}_i\mathbf{Y}_i^T\Big)^{-1}\Big\|_{\mathrm{op}}\Big]\varepsilon.\tag{71}$$

Finally, the function $A\mapsto A^{-1}$ is operator convex on positive matrices, thus

$$\Big(\sum_{i=1}^n a_i^2\mathbf{Y}_i\mathbf{Y}_i^T\Big)^{-1}\le\sum_{i=1}^n a_i^2\big(\mathbf{Y}_i\mathbf{Y}_i^T\big)^{-1}\qquad\text{and}\qquad\big(\mathbb{E}\mathbf{Y}\mathbf{Y}^T\big)^{-1}\le\mathbb{E}\big(\mathbf{Y}\mathbf{Y}^T\big)^{-1}.\tag{72}$$

Applying the operator norm on both sides, plugging this in (71) and using the triangle inequality after taking the expectation, we conclude that

$$\Big\|\mathbb{E}\Big[\Big(\Big(\sum_{i=1}^n a_i^2\mathbf{Y}_i\mathbf{Y}_i^T\Big)^{-1}-\big(\mathbb{E}\mathbf{Y}\mathbf{Y}^T\big)^{-1}\Big)\mathbb{I}_{E_\varepsilon}\Big]\Big\|_{\mathrm{op}}\le\Big(\mathbb{E}\big\|\big(\mathbf{Y}\mathbf{Y}^T\big)^{-1}\big\|_{\mathrm{op}}\Big)^2\varepsilon.\tag{73}$$

In view of (67) and (72), the second term in (69) is bounded by

$$\mathbb{P}\{E_\varepsilon^{\mathrm{c}}\}\,\big\|\big(\mathbb{E}\mathbf{Y}\mathbf{Y}^T\big)^{-1}\big\|_{\mathrm{op}}\le\Big(\frac{4e}{\varepsilon}\Big)^{1+\delta}\log^\delta(d+1)\mathbb{E}\big\|\mathbf{Y}\mathbf{Y}^{\mathbf{T}}\big\|_{\mathrm{op}}^{1+\delta}\,\mathbb{E}\big\|\big(\mathbf{Y}\mathbf{Y}^T\big)^{-1}\big\|_{\mathrm{op}}\,\|a\|_{2+2\delta}^{2+2\delta}.\tag{74}$$

To bound the third term in (69), we use Jensen's inequality and (72) to get

$$\begin{aligned}\Big\|\mathbb{E}\Big[\Big(\sum_{i=1}^n a_i^2\mathbf{Y}_i\mathbf{Y}_i^T\Big)^{-1}\mathbb{I}_{E_\varepsilon^{\mathrm{c}}}\Big]\Big\|_{\mathrm{op}} &\le \mathbb{E}\Big[\Big\|\Big(\sum_{i=1}^n a_i^2\mathbf{Y}_i\mathbf{Y}_i^T\Big)^{-1}\Big\|_{\mathrm{op}}\mathbb{I}_{E_\varepsilon^{\mathrm{c}}}\Big]\\ &\overset{(72)}{\le}\mathbb{E}\Big[\Big\|\sum_{i=1}^n a_i^2\big(\mathbf{Y}_i\mathbf{Y}_i^T\big)^{-1}\Big\|_{\mathrm{op}}\mathbb{I}_{E_\varepsilon^{\mathrm{c}}}\Big]\le\mathbb{E}\Big[\Big(\sum_{i=1}^n a_i^2\big\|\big(\mathbf{Y}_i\mathbf{Y}_i^T\big)^{-1}\big\|_{\mathrm{op}}\Big)\mathbb{I}_{E_\varepsilon^{\mathrm{c}}}\Big]\end{aligned}\tag{75}$$

where the last estimate follows from the triangle inequality. Now, by Hölder's inequality,

$$\begin{aligned}\mathbb{E}\Big[\Big(\sum_{i=1}^n a_i^2\big\|\big(\mathbf{Y}_i\mathbf{Y}_i^T\big)^{-1}\big\|_{\mathrm{op}}\Big)\mathbb{I}_{E_\varepsilon^{\mathrm{c}}}\Big] &\le\mathbb{E}\Big[\Big(\sum_{i=1}^n a_i^2\big\|\big(\mathbf{Y}_i\mathbf{Y}_i^T\big)^{-1}\big\|_{\mathrm{op}}\Big)^{1+\delta}\Big]^{\frac{1}{1+\delta}}\mathbb{P}\{E_\varepsilon^{\mathrm{c}}\}^{\frac{\delta}{1+\delta}}\\ &\le\Big(\mathbb{E}\big\|\big(\mathbf{Y}\mathbf{Y}^T\big)^{-1}\big\|_{\mathrm{op}}^{1+\delta}\Big)^{\frac{1}{1+\delta}}\mathbb{P}\{E_\varepsilon^{\mathrm{c}}\}^{\frac{\delta}{1+\delta}},\end{aligned}\tag{76}$$

where the last line follows from the triangle inequality in $L_{1+\delta}$. Combining this with (75) and (67) we thus conclude that

$$\left\|\mathbb{E}\left[\left(\sum_{i=1}^n a_i^2 \mathbf{Y}_i\mathbf{Y}_i^T\right)^{-1}\mathbb{I}_{E_\varepsilon^c}\right]\right\|_{\mathrm{op}} \le \left(\frac{4e}{\varepsilon}\right)^\delta \log^{\frac{\delta^2}{1+\delta}}(d+1)\left(\mathbb{E}\left\|\mathbf{Y}\mathbf{Y}^{\mathbf{T}}\right\|_{\mathrm{op}}^{1+\delta}\right)^{\frac{\delta}{1+\delta}}\left(\mathbb{E}\left\|\left(\mathbf{Y}\mathbf{Y}^T\right)^{-1}\right\|_{\mathrm{op}}^{1+\delta}\right)^{\frac{1}{1+\delta}}\left\|a\right\|_{2+2\delta}^{2\delta}.$$

Plugging this bound along with (73) and (74) in (69), we get that for every $\varepsilon > 0$,

$$\left\|\mathbb{E}\left(\sum_{i=1}^n a_i^2 \mathbf{Y}_i\mathbf{Y}_i^T\right)^{-1} - \left(\mathbb{E}\mathbf{Y}\mathbf{Y}^T\right)^{-1}\right\|_{\mathrm{op}} \lesssim_{\mathbf{Y}} \varepsilon + \frac{\log^\delta(d+1)\|a\|_{2+2\delta}^{2+2\delta}}{\varepsilon^{1+\delta}} + \frac{\log^{\frac{\delta^2}{1+\delta}}\|a\|_{2+2\delta}^{2\delta}}{\varepsilon^\delta} \tag{77}$$

where the implicit constant depends only on the moments of $\|\mathbf{Y}\mathbf{Y}^T\|_{\mathrm{op}}$. Finally, the (almost) optimal choice $\varepsilon = \|a\|_{2+2\delta}^{\frac{2\delta}{1+\delta}}$ yields the desired bound. □

Remark 13. *We insisted on stating Theorem 7 as a bound for the operator norm of the (normalised) Fisher information matrix of S_n but this is not necessary. An inspection of the proof reveals that given any norm $\|\cdot\|$ on $\mathsf{M}_d(\mathbb{R})$ which is operator monotone, i.e.*

$$0 \preceq A \preceq B \qquad \Longrightarrow \qquad \|A\| \le \|B\| \tag{78}$$

and satisfies the ideal property

$$\forall\ A, B \in \mathsf{M}_d(\mathbb{R}), \qquad \|AB\| \le \|A\|_{\mathrm{op}}\|B\|, \tag{79}$$

we can derive a bound of the form

$$\left\|\mathrm{Cov}(S_n)^{\frac{1}{2}}\mathcal{I}(S_n)\mathrm{Cov}(S_n)^{\frac{1}{2}} - \mathrm{I}_d\right\| \le C\left(\mathbf{Y}, \|\cdot\|\right)\|a\|_{2+2\delta}^{\frac{2\delta}{1+\delta}} \tag{80}$$

for random matrices $\mathbf{Y}$ satisfying (17). *The implicit constant depends on moments of $\|\mathbf{Y}\mathbf{Y}^T\|$ and $\|\mathbf{Y}\mathbf{Y}^T\|_{\mathrm{op}}$ and on the Rademacher type $(1+\delta)$-constant of $\|\cdot\|$. These conditions are, in particular, satisfied for all S_p^d norms and the corresponding type constant is subpolynomial in d for $p \ge 1+\delta$.*

Remark 14. *As was already mentioned in the introduction, bounding the relative Fisher information of a random vector automatically implies bounds for the relative entropy in view of the Gaussian logarithmic Sobolev inequality* [21]. *However, bounds for the Fisher information* matrix *allow one to get better bounds for the relative entropy using more sophisticated functional inequalities which capture the whole spectrum of $\mathcal{I}(X)$. We refer to* [19] *for more on this kind of inequalities.*

Finally, we present some examples of Gaussian mixtures related to conditions (17).

Examples. 1. Fix $p \in (0,2)$ and consider the random variable X_p with density $c_p e^{-|x|^p}$, where $x \in \mathbb{R}$. It was shown in [18, Lemma 23] that X can be expressed as

$$X_p \overset{(\mathrm{d})}{=} (2V_{\frac{p}{2}})^{-\frac{1}{2}}Z, \tag{81}$$

where $V_{\frac{p}{2}}$ has density proportional to $t^{-\frac{1}{2}}g_{\frac{p}{2}}(t)$ and g_a is the density of the standard positive a-stable law. The moments of $Y_p = V_{\frac{p}{2}}^{-1/2}$ then satisfy

$$\forall \alpha \in \mathbb{R}, \qquad \mathbb{E}Y_p^\alpha = \mathbb{E}V_{\frac{p}{2}}^{-\alpha/2} = \kappa_p\int_0^\infty t^{-\frac{\alpha+1}{2}}g_{\frac{p}{2}}(t)\,\mathrm{d}t, \tag{82}$$

for some $\kappa_p > 0$. Since positive $\frac{p}{2}$-stable random variables have finite β-moments for all powers $\beta \in \left(-\infty, \frac{p}{2}\right)$, the assumptions (17) are satisfied when

$$\min\{2\delta+2, -2\delta-2\} > -p-1 \tag{83}$$

or, equivalently, $\delta < \frac{p-1}{2}$. Therefore, Theorem 7 applies for these variables when $p \in (1,2)$.

2. It is well-known (see, for instance [18, Lemma 23]) that for $p \in (0,2)$, the standard symmetric p-stable random variable X_p can be written as

$$X_p \overset{(d)}{=} (2G_{\frac{p}{2}})^{\frac{1}{2}} Z, \tag{84}$$

where $G_{p/2}$ is a standard positive $\frac{p}{2}$ stable random variable. In this setting, the factor $G_{\frac{p}{2}}^{\frac{1}{2}}$ does not have a finite $2+2\delta$ moment for any value of p, so Theorem 7 does not apply.

References

[1] Shiri Artstein, Keith M. Ball, Franck Barthe, and Assaf Naor. On the rate of convergence in the entropic central limit theorem. *Probability theory and related fields*, 129(3):381–390, 2004.

[2] Keith Ball, Eric A. Carlen, and Elliott H. Lieb. Sharp uniform convexity and smoothness inequalities for trace norms. *Invent. Math.*, 115(3):463–482, 1994.

[3] Keith Ball, Piotr Nayar, and Tomasz Tkocz. A reverse entropy power inequality for log-concave random vectors. *Studia Mathematica*, 235:17–30, 2016.

[4] Andrew R. Barron. Entropy and the central limit theorem. *The Annals of Probability*, pages 336–342, 1986.

[5] Jöran Bergh and Jörgen Löfström. *Interpolation spaces. An introduction*. Springer-Verlag, Berlin-New York, 1976. Grundlehren der Mathematischen Wissenschaften, No. 223.

[6] Rajendra Bhatia. *Matrix analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1997.

[7] Peter J. Bickel and Kjell A. Doksum. *Mathematical statistics—basic ideas and selected topics. Vol. 1*. Texts in Statistical Science Series. CRC Press, Boca Raton, FL, second edition, 2015.

[8] Nelson M. Blachman. The convolution inequality for entropy powers. *IEEE Trans. Inform. Theory*, IT-11:267–271, 1965.

[9] Sergey Bobkov and Mokshay Madiman. Reverse Brunn–Minkowski and reverse entropy power inequalities for convex measures. *Journal of Functional Analysis*, 262(7):3309–3339, 2012.

[10] Sergey G. Bobkov. Upper bounds for Fisher information. *Electronic Journal of Probability*, 27:1–44, 2022.

[11] Sergey G. Bobkov and Gennadiy P. Chistyakov. Entropy power inequality for the Rényi entropy. *IEEE Trans. Inform. Theory*, 61(2):708–714, February 2015.

[12] Sergey G. Bobkov, Gennadiy P. Chistyakov, and Friedrich Götze. Rate of convergence and Edgeworth-type expansion in the entropic central limit theorem. *The Annals of Probability*, pages 2479–2512, 2013.

[13] Sergey G. Bobkov, Gennadiy P. Chistyakov, and Friedrich Götze. Berry–Esseen bounds in the entropic central limit theorem. *Probability Theory and Related Fields*, 159(3-4):435–478, 2014.

[14] Sergey G. Bobkov, Gennadiy P. Chistyakov, and Friedrich Götze. Fisher information and the central limit theorem. *Probability theory and related fields*, 159(1-2):1–59, 2014.

[15] Sergey G. Bobkov and Mokshay M. Madiman. On the problem of reversibility of the entropy power inequality. In *Limit Theorems in Probability, Statistics and Number Theory: In Honor of Friedrich Götze*, pages 61–74. Springer, 2013.

[16] Max Costa. A new entropy power inequality. *IEEE Transactions on Information Theory*, 31(6):751–760, 1985.

[17] Imre Csiszár. Information-type measures of difference of probability distributions and indirect observation. *studia scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.

[18] Alexandros Eskenazis, Piotr Nayar, and Tomasz Tkocz. Gaussian mixtures: entropy and geometric inequalities. *The Annals of Probability*, 46(5):2908–2945, 2018.

[19] Alexandros Eskenazis and Yair Shenfeld. Intrinsic dimensional functional inequalities on model spaces. Preprint available at `https://arxiv.org/abs/2303.00784`, 2023.

[20] Tadeusz Figiel, Joram Lindenstrauss, and Vitali D. Milman. The dimension of almost spherical sections of convex bodies. *Acta Math.*, 139(1-2):53–94, 1977.

[21] Leonard Gross. Logarithmic Sobolev inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975.

[22] Oliver Johnson and Andrew Barron. Fisher information inequalities and the central limit theorem. *Probability Theory and Related Fields*, 129:391–409, 2004.

[23] Solomon Kullback. A lower bound for discrimination information in terms of variation (corresp.). *IEEE transactions on Information Theory*, 13(1):126–127, 1967.

[24] Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1991. Isoperimetry and processes.

[25] Elliott H. Lieb. Proof of an entropy conjecture of Wehrl. *Comm. Math. Phys.*, 62(1):35–41, 1978.

[26] Mokshay Madiman, James Melbourne, and Peng Xu. Forward and reverse entropy power inequalities in convex geometry. In *Convexity and concentration*, pages 427–485. Springer, 2017.

[27] Mokshay Madiman, Piotr Nayar, and Tomasz Tkocz. Two remarks on generalized entropy power inequalities. In *Geometric aspects of functional analysis. Vol. II*, volume 2266 of *Lecture Notes in Math.*, pages 169–185. Springer, Cham, [2020] ©2020.
[28] Albert W. Marshall, Ingram Olkin, and Barry C. Arnold. *Inequalities: theory of majorization and its applications*. Springer Series in Statistics. Springer, New York, second edition, 2011.
[29] Mark S. Pinsker. *Information and information stability of random variables and processes*. Holden-Day, Inc., San Francisco, Calif.-London-Amsterdam,, 1964.
[30] Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.
[31] Aart J. Stam. Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Information and Control*, 2:101–112, 1959.
[32] Nicole Tomczak-Jaegermann. The moduli of smoothness and convexity and the Rademacher averages of trace classes $S_p (1 \leq p < \infty)$. *Studia Math.*, 50:163–182, 1974.

(A. E.) CNRS, Institut de Mathématiques de Jussieu, Sorbonne Université, France and Trinity College, University of Cambridge, UK.
Email address: alexandros.eskenazis@imj-prg.fr, ae466@cam.ac.uk

(L. G.) Laboratoire d'Analyse et de Mathématiques Appliquées, Université Gustave Eiffel, France.
Email address: lampros.gavalakis@univ-eiffel.fr