Upper bounds for the relative entropy of the squared norm of a Gaussian projection onto a convex cone

Ivan Nourdin*
Université du Luxembourg, L-1359 Luxembourg; ivan.nourdin@uni.lu

Abstract

We provide upper bounds for the relative entropy of the squared norm of the Gaussian projection onto a convex cone, that are specially relevant in the regime where the statistical dimension of the cone is large. Our findings complement some results proved in [3].

Keywords: Fisher information; total variation distance; relative entropy; Gaussian projection.

1 Introduction and main results

Projections on closed convex cones play a pivotal tool in a constrained statistical inference setup, see, e.g., the reference [8]. Every closed convex cone $C \subset \mathbb{R}^d$ can be associated with a random variable $V_C$, with support on \{0, \ldots, d\} whose distribution $\mathcal{L}(V_C)$ coincides with the intrinsic volumes of $C$, see Section 2.2 for more details. Its mean $\delta_C = \mathbb{E}[V_C]$ (called the statistical dimension of $C$) measures in some sense the ‘effective’ dimension of $C$, and is a direct generalisation of the dimension of a linear space. It is also closely related to the squared Gaussian width.

Within the past decade, the concentration of measure theory has been extraordinarily successful in proving sharp phase transitions that were observed empirically in a number of diverse situations. The beautiful reference [1] (see also the nice contribution [2] of Candès to the Proceedings of the 2014 ICM at Seoul) contains deep and important examples where such a transition for convex programs can be explained by the Gaussian concentration of the intrinsic volumes of a convex cone around its statistical dimension.

The elegant ‘Master Steiner formula’ from McCoy and Tropp [5] connects $V_C$ to the random variable $\|\Pi_C(g)\|^2$ where, here and throughout the text, $g$ denotes a standard Gaussian vector on $\mathbb{R}^d$, $\Pi_C$ is the metric projection onto $C$, and $\|\cdot\|$ stands for the Euclidean norm. Shifting from $V_C$ to $\|\Pi_C(g)\|^2$ allows one (among others) to obtain asymptotic properties for $V_C$ by first proving them for $\|\Pi_C(g)\|^2$. For instance, under an additional technical assumption on the asymptotic behavior of the variance of $V_C$ it is shown in [3, Theorem 3.3(3)] that the random variable $W_C := (V_C - \delta_C)/\sqrt{\text{Var}(V_C)}$ satisfies a central limit theorem (CLT) when $\delta_C \to \infty$ if and only if the same holds for $F_C := (\|\Pi_C(g)\|^2 - \delta_C)/\sqrt{\text{Var}(V_C)} + 2\delta_C$. In [3, Proposition 3.1], it is also shown that $d_{TV}(F_C, N) \leq 8/\sqrt{\delta_C}$ for all closed convex cone $C$, where $N \sim N(0, 1)$ is a standard Gaussian and $d_{TV}$ stands for the total variation distance; in particular, $F_C \overset{\text{Law}}{\to} N$ as $\delta_C \to \infty$. Putting these two results together yields that $W_C \overset{\text{Law}}{\to} N$ as $\delta_C \to \infty$, thus showing that, in the high-dimensional limit, most conic intrinsic volumes encountered in applications can be approximated by a suitable Gaussian distribution.

Our goal in the present paper is to complement some of the results proved in [3], by analyzing stronger notions of convergences for $F_C$. We refer to this latter reference for a complete discussion about the relevance and the interests of the results we are presenting below. More specifically, do we have convergence of $F_C$ to the standard Gaussian in relative entropy? in relative Fisher information? (See Section 2.1 for a quick summary about these strong notions of convergence.) Our Theorem 1.1 provides an answer to these two questions and corresponds to the main result of this short note.

---

* $C$ is a convex cone if $ax + by \in C$ whenever $x$ and $y$ are in $C$ and $a$ and $b$ are positive reals.
Theorem 1.1 Let \( \{d_n : n \geq 1\} \) be a sequence of non-negative integers and let \( \{C_n \subset \mathbb{R}^{d_n} : n \geq 1\} \) be a sequence of non-empty closed convex cones such that \( \delta_n := \delta_{C_n} \to \infty \). Furthermore, assume that \( P(V_{C_n} \geq 5) = 1 \) for every \( n \) (large enough). Consider a standard Gaussian random vector \( g_n \) on \( \mathbb{R}^{d_n} \), and set \( \tau_n^2 = \text{Var}(V_{C_n}) \), \( \sigma_n^2 = \text{Var}(\|\Pi_{C_n}(g_n)\|)^2 = \tau_n^2 + 2\delta_n \), and \( F_n = \frac{1}{\delta_n}(\|\Pi_{C_n}(g_n)\|^2 - \delta_n) \). Then, the following holds as \( n \to \infty \).

1. The relative entropy of \( F_n \) satisfies \( H(F_n) = O\left(\frac{\log \delta_n}{\delta_n}\right) \); thus, \( H(F_n) \to 0 \) as \( n \to \infty \).

2. The Fisher information of \( F_n \) satisfies \( J(F_n) = 1 + O\left(\max\{\sqrt{\log \delta_n}, \frac{\tau_n^2}{\delta_n}\}\right) \). In particular, if \( \tau_n^2 = o(\delta_n) \) then \( F_n \) converges to \( N(0,1) \) in Fisher information.

Let us give some remarks and comments about Theorem 1.1.

- Section 2.1 contains definitions and explanations for convergence in relative entropy, resp. Fisher information.
- In the statement of Theorem 1.1 and as far as the relative entropy is concerned, it is actually possible to replace \( P(V_{C_n} \geq 5) = 1 \) by the weakest assumption \( P(V_{C_n} \geq 1) = 1 \). The proof of this claim is more technical than the one presented here, and can be obtained as a consequence of Theorem 1.7 and Lemma 4.4 in [7]. (Details are not provided here.) Nevertheless, we stress that one cannot avoid assuming that \( P(V_{C_n} \geq 1) = 1 \); indeed, if \( P(V_C = 0) > 0 \) then \( \|\Pi_C(g)\|^2 \) has an atom at 0 and, as a consequence, its relative entropy is infinite.
- When \( d > s \geq 6 \), the descent cone \( C \subset \mathbb{R}^d \) of the \( \ell_1 \)-norm at a given \( s \)-sparse vector \( x \) satisfies \( P(V_C \geq 5) = 1 \). This important cone, which is at the core of the compressed sensing theory, is used for the recovery of sparse vectors via \( \ell_1 \)-norm minimization. See, e.g., [1, 2, 3] for further details.
- We conjecture that \( J(F_n) \to 1 \) if and only if \( \tau_n^2 = o(\delta_n) \). However, we have to stress that this latter condition is not satisfied for the most common examples of convex cones that appear in practice, see indeed Table 1 as well as Sections 4.2 and 4.3 of [3].

The rest of the paper is organized as follows. Section 2 contains a description of the framework in which our study takes place and gathers several preliminary results. Finally, proof of Theorem 1.1 is given in Section 3.

2 Preliminaries

2.1 Fisher information, relative entropy, Stein discrepancy and the HSI inequality

Let \( F \) be a real-valued random variable with, say, mean zero and unit variance. Assume further that \( F \) has a density, noted \( p_F : \mathbb{R} \to \mathbb{R} \). The relative entropy \( H(F) \) of \( F \) with respect to \( N \sim N(0,1) \) is defined as \( H(F) = \int_{\mathbb{R}} p_F(x) \log \frac{p_F(x)}{p_N(x)} \, dx \), where \( p_N(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \) is the density of the standard Gaussian distribution on \( \mathbb{R} \). One can prove that \( H(F) \geq 0 \). Our interest in \( H(F) \) comes from its link with the total variation distance \( d_{TV} \), as provided by the celebrated Pinsker inequality, according to which

\[
d_{TV}(F,N) := \sup_{A \in \mathcal{B}(\mathbb{R})} |P(F \in A) - P(N \in A)| \leq \sqrt{\frac{1}{2} H(F)}. \tag{1}\end{equation}

Inequality (1) actually shows that bounds on the relative entropy translate directly into bounds on the total variation distance. Thus, it makes perfectly sense to quantify the discrepancy between the law of \( F \) and that of the standard Gaussian \( N \) in terms of its relative entropy. One can go even further by considering the Fisher information of \( F \). Let us recall its definition. Let \( s_F \) denote the score function associated with \( F \), that is, the function uniquely determined by the following integration by parts:

\[
E[\phi'(F)] = -E[s_F(F)\phi(F)] \quad \text{for all test function } \phi : \mathbb{R} \to \mathbb{R}. \tag{2}\end{equation}
When it makes sense, it is easy to compute that \( s_F = p_F' / p_F \) (with the convention that \( s_F(x) = 0 \) when \( x \) is outside the support of \( p_F \)). When \( s_F(F) \) is square-integrable, the Fisher information of \( F \) is set to be

\[
J(F) = E[s_F(F)^2] = \int_R \frac{p_F'(x)^2}{p_F(x)} \, dx;
\]

otherwise, we set \( J(F) = +\infty \). In the former case, it is a straightforward exercise to check that \( J(F) - 1 = E[(s_F(F) + F)^2] \). In particular, \( J(F) \geq 1 = J(N) \) with equality if and only if \( F \) is standard Gaussian. The quantity \( J(F) - 1 \) is called the relative Fisher information of \( F \); it is related to the relative entropy through the celebrated log-Sobolev inequality:

\[
H(F) \leq \frac{1}{2} (J(F) - 1). \tag{4}
\]

By comparing (4) with (1), we see that the gap between \( J(F) \) and \( 1 = J(N) \) is an even stronger measure of how close the law of \( F \) is to the standard Gaussian \( N \). The interest of (4) is that, due to the representation (3) (first identity), it is often an easier task to compute (or to estimate) \( J(F) - 1 \) rather than \( H(F) \). Inequality (4) becomes then particularly useful. But sometimes one may have \( H(F) \approx 0 \) without \( J(F) \approx 1 \). Inequality (4) becomes then irrelevant and one needs another strategy to bound the relative entropy. The HSI inequality of [6, Theorem 2.2], which asserts that

\[
H(F) \leq \frac{S^2(F)}{2} \log \left( 1 + \frac{J(F) - 1}{S^2(F)} \right), \tag{5}
\]

may happen to be the right tool in such situations. In (5), \( S^2(F) \) stands for the so-called Stein discrepancy, which is defined by \( S^2(F) = E[(1 - \tau_F(F))^2] \), where the Stein kernel \( \tau_F \) of \( F \) is the uniquely determined function satisfying \( E[F \phi(F)] = E[\tau_F(F) \phi'(F)] \) for all test function \( \phi : \mathbb{R} \to \mathbb{R} \), compare with (2).

**Remarks 2.1** Since \( (x, y) \mapsto x \log(1 + y/x) \) is increasing in \( x \) (for \( y \) fixed) and in \( y \) (for \( x \) fixed), in (5) one can freely replace \( J(F) - 1 \) (resp. \( S^2(F) \)) by any upper bound.

### 2.2 Intrinsic volumes of a convex cone

One can express the size of an angular expansion of a closed convex cone \( C \subset \mathbb{R}^d \) by means of the ‘angular’ Steiner formula, according to which \( P \{ d^2(\theta, C) \leq \lambda \} = \sum_{j=0}^d \beta_j d(\lambda) v_j(C) \). Here, \( d(\cdot, C) \) stands for the Euclidean distance to \( C \), \( \theta \) is a random variable uniformly distributed on the unit sphere in \( \mathbb{R}^d \), the coefficients \( \beta_j d(\lambda) \) do not depend on \( C \), and the conic intrinsic volumes \( v_0, \ldots, v_d \) are determined by \( C \) only, and can be shown to be nonnegative and sum to one. As a consequence, we may associate to the conic intrinsic volumes of \( C \) an integer-valued random variable \( V_C \), whose probability distribution is given by

\[
P(V_C = j) = v_j(C), \quad \text{for } j = 0, \ldots, d. \tag{6}
\]

Let \( g \) be a standard Gaussian random vector on \( \mathbb{R}^d \). As shown in [5, Corollary 3.2], the squared norm \( \| \Pi_C(g) \|^2 \) behaves like a chi-squared random variable with a random number \( V_C \) of degrees of freedom: in symbols,

\[
\| \Pi_C(g) \|^2 \overset{\text{law}}{=} \sum_{k=1}^{V_C} X_k^2 \tag{7}
\]

where \( X_1, X_2, \ldots \sim N(0,1) \) are independent and independent from \( V_C \). Reference [5] (inequality (4.14) therein) also provides a concentration bound for \( V_C \), which shall play a crucial role in our proof of Theorem 1.1 for any \( \lambda > 0 \) and with \( \delta_c = E[V_C] \),

\[
P(\mid V_C - \delta_C \mid \geq \lambda) \leq 2 \exp \left\{ -\frac{\lambda^2}{4(\delta_c + \lambda/3)} \right\}. \tag{8}
\]
3 Proof of Theorem 1.1

We start with a preliminary lemma.

**Lemma 3.1** Let $C \subset \mathbb{R}^d$ be a non-trivial closed convex set such that $P(V_C \geq 5) = 1$, and let $g$ be a standard Gaussian random vector on $\mathbb{R}^d$. Then, the Fisher information of $\|\Pi_C(g)\|^2$ satisfies $J(\|\Pi_C(g)\|^2) \leq \frac{1}{2} E \left[ \frac{1}{\|V_C - x\|} \right]$. 

**Proof.** We use the representation provided by the first identity of (7). Let $\phi$ be a test function. We can write, considering $Z_i = \sum_{k=1}^i X_k^2$ and using the independence between $(Z_i)_{i \geq 1}$ and $V_C$,

$$-E \left[ \phi' \left( \sum_{k=1}^{V_C} X_k^2 \right) \right] = -\sum_{l=5}^d E[\phi'(Z_l)] P(V_C = l) = \sum_{l=5}^\infty E[\phi(Z_l) s_l(Z_l)] P(V_C = l).$$

Here, $s_l$ is the score function associated with $Z_l \sim \chi^2(l)$, which is given by $s_l(x) = (\frac{1}{2} - 1) \frac{x}{l} - \frac{l}{2}, x > 0$ (see, e.g., [4, Example 1.5, page 22]). As a result,

$$-E \left[ \phi' \left( \sum_{k=1}^{V_C} X_k^2 \right) \right] = E \left[ \phi \left( \sum_{k=1}^{V_C} X_k^2 \right) \left( \frac{V_C}{2} - 1 \right) \sum_{l=5}^\infty \frac{V_C}{2} \sum_{k=1}^{V_C} X_k^2 - \frac{1}{2} \right]$$

from which we deduce, going back to the representation (3) of the Fisher information and using that $J(Z_l) = \frac{1}{2(l-4)}$ (for this latter fact, see again [4, Example 1.5, page 22]),

$$J(\|\Pi_C(g)\|^2) = E \left[ \frac{V_C}{2} - 1 \right] \sum_{l=5}^\infty \frac{V_C}{2} \sum_{k=1}^{V_C} X_k^2 - \frac{1}{2} \right]$$

$$\leq E \left[ \left( \frac{V_C}{2} - 1 \right) \sum_{l=5}^\infty \frac{V_C}{2} \sum_{k=1}^{V_C} X_k^2 - \frac{1}{2} \right]$$

$$= \sum_{l=5}^d J(Z_l) P(V_C = l) = \frac{1}{2} \sum_{l=5}^d P(V_C = l) = \frac{1}{2} E \left[ \frac{1}{V_C - 4} \right].$$

We are now in a position to prove Theorem 1.1.

**Proof of Theorem 1.1** For the sake of clarity, it is divided into several steps.

**Step 1.** Using the very definition of $J$, it is immediate to check that $J(aF + b) = J(F) + \frac{b}{\sigma^2}$ for any $a \neq 0$ and $b \in \mathbb{R}$. In our case, we obtain that $J(F_n) = \sigma^2 n J(\|\Pi_C(g_n)\|^2)$. From Lemma 3.1 and since we do the assumption that $P(V_{C_n} \geq 5) = 1$ for all $n$ large enough, we deduce that $J(F_n) \leq \sigma^2 \frac{1}{2} E \left[ \frac{1}{V_{C_n} - 4} \right]$ for all $n$ large enough.

**Step 2.** Let us now bound $E \left[ \frac{1}{V_{C_n} - 4} \right]$. The idea is that, since $V_{C_n}$ concentrates around its mean according to (8), it is legitimate to expect that the quantity $E \left[ \frac{1}{V_{C_n} - 4} \right]$ is somehow close to $\frac{1}{E[V_{C_n}]} \approx \frac{1}{\delta_n}$. To make this idea rigorous let us write, for $n$ large enough,

$$E \left[ \frac{1}{V_{C_n} - 4} \right] = \sum_{k=1}^{\infty} \frac{1}{k} P(V_{C_n} = k + 4) \leq \sum_{k=1}^{\infty} P(V_{C_n} = k + 4) \sum_{l=k}^{\infty} \frac{1}{l^2}$$

$$\leq \sum_{l=0}^{\infty} \frac{1}{l^2} P(5 \leq V_{C_n} \leq l + 4) \leq \sum_{l=1}^{\lfloor \delta_n \rfloor - 4} \frac{1}{l^2} P(V_{C_n} \leq l + 4) + \sum_{l=\lfloor \delta_n \rfloor - 4}^{\infty} \frac{1}{l^2}$$

$$\leq \sum_{l=1}^{\lfloor \delta_n \rfloor - 4} \frac{1}{l^2} P(V_{C_n} \leq l + 4) + \frac{1}{\delta_n - 6}.$$
To evaluate the sum in the previous right-hand side, let us introduce a factor $\rho \in (0, 1)$; we can then write, using (8) among others,

$$
\sum_{l=1}^{\lfloor \delta_n \rfloor - 5} \frac{1}{l^2} P(V_{C_n} \leq l + 4) = \sum_{l=1}^{\lfloor \delta_n \rfloor - 5} \frac{1}{(\lfloor \delta_n \rfloor - 4 - l)^2} P(V_{C_n} \leq \lfloor \delta_n \rfloor - 4)
$$

$$
\leq 2 \sum_{l=1}^{\lfloor \delta_n \rfloor - 5} \frac{1}{(\lfloor \delta_n \rfloor - 4 - l)^2} \leq 2 \exp \left( - \frac{l^2}{4(\delta_n + \frac{l}{3})} \right)
$$

$$
\leq \frac{2\rho}{(1-\rho)^2(\lfloor \delta_n \rfloor - 5)} + 2 \exp \left( - \frac{3\rho^2(\lfloor \delta_n \rfloor - 5)^2}{16\delta_n} \right).
$$

**Step 3.** By combining the results of Steps 1 and 2 together with the fact that

$$
\sigma_n^2 \leq 4\delta_n, \quad (9)
$$

one obtains

$$
J(F_n) \leq \frac{\sigma_n^2}{2} E \left[ \frac{1}{V_{C_n} - 4} \right] \leq \frac{\rho}{\delta_n - 6} \left( \frac{\rho}{(1-\rho)^2 + \frac{1}{2}} \right) + \sigma_n^2 \exp \left( - \frac{3\rho^2(\delta_n - 6)^2}{16\delta_n} \right)
$$

$$
\leq \frac{1 + \rho^2}{2(1-\rho)^2} \times \frac{\sigma_n^2}{\delta_n - 6} + 4\delta_n \exp \left( - \frac{3\rho^2(\delta_n - 6)^2}{16\delta_n} \right).
$$

Let us choose $\rho = \sqrt{\frac{32\log 5}{(\delta_n - 6)^2}} \quad (\rho < 1 \text{ for } n \text{ large enough})$ so that $\exp \left( - \frac{3\rho^2(\delta_n - 6)^2}{16\delta_n} \right) = \frac{1}{57} \; \text{(some elementary simplifications (using in particular that } \sigma_n^2 = 2\delta_n + \tau_n^2 \text{) lead to the first point of Theorem 1.1}}$.

Let us now prove the second point of Theorem 1.1. From (9), one deduces that $\tau_n^2 \leq 2\delta_n$. Combining this fact with the first point of Theorem 1.1 leads to $J(F_n) = O(1)$ as $n \to \infty$. On the other hand, recall from [3] Proof of Theorem 3.1 (\(\mu = 0\) therein) that $S^2(F_n) = O(\frac{1}{n})$. Finally, plugging these two bounds, on $J(F_n)$ and $S^2(F_n)$ respectively, into [8] allows to conclude that the desired conclusion takes place, see also Remark 2.1. □

**References**


