Integration by parts and representation of information functionals

Ivan Nourdin[*], G. Peccati[†] and Y. Swan[‡]

December 19, 2013

Abstract

We introduce a new formalism for computing expectations of functionals of arbitrary random vectors, by using generalised integration by parts formulae. In doing so we extend recent representation formulae for the score function introduced in [18] and also provide a new proof of a central identity first discovered in [8]. We derive a representation for the standardized Fisher information of sums of i.i.d. random vectors which use our identities to provide rates of convergence in information theoretic central limit theorems (both in Fisher information distance and in relative entropy).

1 Introduction

Let X be a random vector in $\mathbb{R}^d$, with differentiable density f. The *score function* $\rho_X(x) = \nabla \log f(x)$ is a well-understood and oft-used quantity whose behaviour has long been known to provide useful summaries of the law of X. In particular the covariance matrix of the random vector $\rho_X(X)$ (the so-called *score of* X) is the *Fisher information matrix of* X; this matrix is much used by statisticians and probabilists alike. A much less studied object is the *Stein matrix* of X, defined in (2), which can in some sense be seen as a counterpart to the score where rather than taking log-derivatives one considers a special form of integration. This matrix (whose properties when $d = 1$ are closely related to the so-called *zero-bias transform*, see [7]) has only recently started to attract the attention of the community. For instance, this matrix was exploited in [17] to study central convergence on Wiener Chaos. See also [1] for a detailed study. And as we shall demonstrate, the Stein matrix of a random vector X is a no less fundamental quantity than its score.

Rather than defining these two quantities explicitly in terms of the density f, we choose to characterise them by their behaviour through integration by parts formulae

[*] Université de Lorraine; IN is partially supported by the (french) ANR grant 'Malliavin, Stein and Stochastic Equations with Irregular Coefficients' [ANR- 10-BLAN-0121].

[†] Université du Luxembourg

[‡] Université du Luxembourg

tailored for integration with respect to the law of X (see definitions (1) and (2)). These formulae are inspired by results usually exploited within the context of the so-called *Stein's method* (see [16, 5, 6]). Though in some sense elementary, our approach allows us to obtain a variety of representation formulae for the Fisher information of an arbitrary random vector X and, as we shall demonstrate, provides powerful information theoretic bounds for Gaussian approximation problems. The results in this paper are closely related to the work [18] and tiptoes around the results from [9, 10] as well as [3, 4]. The idea of exploiting representation formulae for Fisher information was pioneered in [8] and, as we shall see, our results are closely to theirs.

The outline of the paper is as follows. All formulae and definitions are provided in Section 2. In Section 3 we show how elementary manipulations of these integration by parts formulae allow to generalize the central identity from [18] (see forthcoming Lemma 3.1) and to prove a version of the celebrated MMSE formula from [8]. In Section 5 we exploit our identities to provide a general *"Stein bound"* on the standardized Fisher information of sums of iid random vectors.

2 Score function and Stein factor

Fix an integer $d \geq 1$. Let X, Y be centered random d-vectors (all elements in $\mathbb{R}^d$ are taken as $d \times 1$ column vectors) which we throughout assume to admit a density (with respect to the Lebesgue measure) with support $S \subset \mathbb{R}^d$.

Definition 2.1. *The* score *of X is the random vector $\rho_X(X)$ which satisfies*

$$E\left[\rho_X(X)\varphi(X)\right] = -E\left[\nabla\varphi(X)\right] \tag{1}$$

(with ∇ the usual gradient in $\mathbb{R}^d$) for all test functions $\varphi \in C_c^\infty(\mathbb{R}^d)$. Any random $d \times d$ matrix $\tau_X(X)$ which satisfies

$$E\left[\tau_X(X)\nabla\varphi(X)\right] = E\left[X\varphi(X)\right] \tag{2}$$

for all test functions $\varphi \in C_c^\infty(\mathbb{R}^d)$ is called a Stein matrix *for X.*

Remark 2.2. The existence of a score function and of a Stein matrix for a random vector X are substantial (and quite restrictive) assumptions. Let X have density f with support $S \subset \mathbb{R}^d$. Then assumptions on the behavior of the density at the edges of the support are implicit : indeed it is necessary that $f|_{\delta(S)} = 0$ ($\delta(S)$ is the border of S) in order to have (1) and (2). In the sequel we suppose that, when required, all such assumptions are satisfied.

If X has covariance matrix C, then a direct application of the definition of the Stein matrix yields

$$E\left[\tau_X(X)\right] = C;$$

similarly we get

$$E\left[\rho_X(X)\right] = 0 \text{ and } E\left[\rho_X(X)X^T\right] = -Id,$$

where A^T denotes the transpose of A and Id is the $d \times d$ identity matrix.

For a Gaussian random vector Z with covariance matrix C one uses the well-known Stein identity (see, e.g., [9] for proof and a history)

$$E\left[Z\varphi(Z)\right] = CE\left[\nabla\varphi(Z)\right] \tag{3}$$

to prove that $\rho_Z(Z) = -C^{-1}Z$ and $\tau_Z(Z) = C$ are the Gaussian score and Stein matrix, respectively. Identity (3) characterizes the Gaussian distribution in the sense that a random vector X with support $\mathbb{R}^d$ satisfies (3) if and only if X is itself Gaussian with covariance C. More generally, the following result holds (see also, e.g., [1, 9, 21]).

Proposition 2.3. *Let X have density f and let $\rho_X(x) = \nabla \log f(x)$. Then $\rho_X(X)$ is the score of X and is unique.*

Proof. Proving that $\rho_X(X)$ as defined in the statement is a score for X follows from an easy integration by parts argument (whose validity rests on the border condition on f). To see uniqueness let $\rho(X)$ be another score of X and note that, then, we have

$$E\left[(\rho(X) - \rho_X(X))\varphi(X)\right] = 0$$

for all test functions φ. Consequently $\rho(X) = \rho_X(X)$ almost surely. □

In the case $d = 1$, under standard assumptions of regularity of the density f, the existence of the Stein's matrix τ (which is indeed a one-dimensional mapping sometimes called *Stein's factor*) follows from standard integration by parts arguments, from which one deduces that

$$\tau(x) = f(x)^{-1}\int_x^\infty f(y)dy.$$

In higher dimensions, the existence of a Stein matrix for X also follows easily from an integration by parts argument, once one can find a matrix valued function $x \mapsto A(x)$ whose components a_{ij} with $1 \leq i, j \leq d$ satisfy

$$\sum_{j=1}^d \frac{\partial}{\partial x_j}\left(a_{ij}(x)f(x)\right) = -x_i \tag{4}$$

for all $i = 1, \ldots, d$. As demonstrated in the huge body of literature revolving around Malliavin calculus (see [18] as well as the monograph [16]), a Stein's matrix alway exist for random vectors that are given by a smooth transformation of a given Gaussian field. Contrarily to the score, however, there is no reason for which the Stein matrix, at least according to our definition, should be unique. See the appendix for a deeper discussion on this issue.

The following useful properties follow immediately from the definitions.

Proposition 2.4. *Let $\rho_X(X)$ be the score function of X. Then $\rho_{aX}(aX) = \frac{1}{a}\rho_X(X)$ and $\rho_{AX+b}(AX+b) = (A^{-1})^T\rho_X(X)$ for all $a > 0$, for all invertible matrices $A \in \mathcal{M}(d)$ and all vectors $b \in \mathbb{R}^d$ (A^T is the transpose of A).*

Similarly, if $\tau_X(X)$ is a Stein matrix of X, then $a^2\tau_X(X)$ is a Stein matrix for aX and $A\tau_X(X)A^T$ is a Stein matrix of $AX+b$.

Proof. Invariance under translation is immediate. To verify that the effect of scaling is as announced, we check that the right-hand side of the equalities satisfy the corresponding definitions. Take a test function φ. Using $\nabla(\varphi(AX)) = A^T\nabla\varphi(AX)$, we can write

$$E\left[(A^{-1})^T\rho_X(X)\varphi(AX)\right] = -(A^{-1})^T E\left[A^T\nabla\varphi(AX)\right] = -E\left[\nabla\varphi(AX)\right].$$

Similarly, we see that

$$E\left[A\tau_X(X)A^T\nabla\varphi(AX)\right] = AE\left[\tau_X(X)\nabla\left(\varphi(AX)\right)\right] = E\left[AX\nabla\varphi(AX)\right].$$

In both cases the conclusion follows by definition. □

Given a random vector X (with density f) we define its (differential) entropy and its Fisher information as

$$H(X) = -E\left[\log f(X)\right] \text{ and } I(X) = E\left[\rho_X(X)\rho_X(X)^T\right], \tag{5}$$

respectively. Using Proposition 2.4 we obtain the following result.

Proposition 2.5. *For a some positive constant we have*

$$I(aX) = \frac{1}{a^2}I(X) \text{ and } H(aX) = \log a + H(X). \tag{6}$$

More generally, for all invertible matrices $A \in \mathcal{M}(d)$ we get

$$I(AX+b) = (A^{-1})^T I(X)A^{-1} \text{ and } H(AX+b) = \log|\mathrm{det}A| + H(X) \tag{7}$$

with $b \in \mathbb{R}^d$ and $\mathrm{det}A$ *the determinant of A.*

If Z is Gaussian with covariance C then one can use the explicit expression of its density to compute

$$H(Z) = \frac{1}{2}\log((2\pi e)^d\mathrm{det}C) \text{ and } I(Z) = C^{-1}. \tag{8}$$

Definition 2.6. *Let X be a d-random vector with density f and covariance B, and let ϕ be the density of $Z \sim \mathcal{N}_d(0, C)$. The* relative entropy *of X (with respect to Z) is*

$$D(X \,\|\, Z) = E\left[\log(f(X)/\phi(X))\right] \tag{9}$$

and its relative Fisher information matrix *is*

$$\mathcal{J}(X) = E\left[(\rho_X(X) + B^{-1}X)(\rho_X(X) + B^{-1}X)^T\right]. \tag{10}$$

The standardized Fisher information distance *is*

$$J_{st}(X) = \mathrm{tr}\left(B\mathcal{J}(X)\right), \tag{11}$$

with 'tr' *the usual trace operator.*

Entropy and Fisher information are related to one another via the so-called *de Bruijn's identity* (see [11, Lemma 2.2] for the original statement, as well as [18] for the forthcoming version).

Lemma 2.7 (Multivariate de Bruijn's identity)**.** *Let X be a random d-vector with covariance C (invertible); let Z be Gaussian with covariance B and define $X_t = \sqrt{t}X + \sqrt{1-t}Z$, $t \in [0,1]$. Let Γ_t be the covariance matrix of X_t. Then*

$$\begin{aligned} D(X\|Z) = & \int_0^1 \frac{1}{2t} \mathrm{tr}\left(C\Gamma_t^{-1} J_{st}(X_t)\right) dt \\ & + \frac{1}{2}\left(\mathrm{tr}\left(C^{-1}B\right) - d\right) + \int_0^1 \frac{1}{2t} tr\left(C\Gamma_t^{-1} - I_d\right) dt. \end{aligned} \tag{12}$$

If $C = B$ then

$$D(X\,\|Z) = \int_0^1 \frac{1}{2t}\mathrm{tr}(J_{st}(X_t))dt. \tag{13}$$

There are a number of fundamental deep inequalities that are known on the behavior of information and entropy over convolutions. For instance, using an elementary representation formula for the score of a sum of random variables, one can prove (see, e.g., [9, Lemma 1.21]) that information (and therefore standardized information) decreases along convolutions.

Lemma 2.8. *If X and Y are independent real-valued random variables then*

$$I(\sqrt{t}X + \sqrt{1-t}Y) \le tI(X) + (1-t)I(Y) \tag{14}$$

with equality if and only if X and Y are both Gaussian.

In particular, from (14) and if X is real-valued with unit variance and if Z is standard Gaussian, we have that for all $0 \le t \le 1$,

$$J_{st}(X_t) \le tJ_{st}(X) + (1-t)J_{st}(Z) = tJ_{st}(X)$$

so that

$$D(X \,\|\, Z) \le \frac{1}{2}\mathrm{tr}(J_{st}(X)). \tag{15}$$

Hence bounds on the standardized Fisher information translate directly into bounds on the relative entropy hereby providing, via Pinsker's inequality

$$\sqrt{2}d_{TV}(X, Z) \le D(X \,\|\, Z), \tag{16}$$

bounds on the *total variation distance* between the law of X and the law of Z. Applying de Bruijn's identity one can then integrate the above relation to deduce

$$H(\sqrt{t}X + \sqrt{1-t}Y) \ge tH(X) + (1-t)H(Y) \tag{17}$$

for all $0 \le t \le 1$. This is equivalent to the so-called *entropy power inequality* first put forth by Shannon in [19] (see [20] or [23] for a proof).

Lemma 2.9 (Entropy power inequality)**.** *If X and Y are independent real-valued random variables with density having entropy $H(X)$ then*

$$exp\,(2H(X+Y)) \geq exp(2H(X)) + exp(2H(Y)). \tag{18}$$

Fisher information decreases with convolution, while entropy increases accordingly. The amenability of the Gaussian density for computations allows one to easily prove that relative entropy is positive definite and satisfies

$$0 \leq D(X \,\|\, Z) = H(Z) - H(X) + \frac{tr(C^{-1}B) - d}{2}; \tag{19}$$

also direct computations yield

$$0 \leq J_{st}(X) = \mathrm{tr}(BI(X) - Id). \tag{20}$$

Hence it makes sense to quantify the discrepancy between the law of X and the Gaussian in terms of its relative entropy or of its Fisher information. Adding to this the fact that the Gaussian has (in dimension 1) maximum entropy and minimal Fisher information among all random variables with given variance, many authors have sought to provide intrinsic interpretations of the CLT via maximum entropy or minimal Fisher information arguments. These observations spawned a series of papers wherein the authors use either the Fisher information distance or the relative entropy to prove so-called "information theoretic central limit theorems"; see [9] for references and details.

In the present paper we take a new angle on these matters via a novel representation formula which we detail in the forthcoming section.

3 Representation formulae

The following lemma is a generalization of the central formula from [18] and is the cornerstone of the present paper. The device contained in the proof (namely a probabilistic integration by parts formula) will be used throughout the subsequent arguments.

Lemma 3.1. *Let X and Y be stochastically independent centered random vectors in $\mathbb{R}^d$. Suppose that X (resp., Y) allows a score $\rho_X(X)$ (resp., $\rho_Y(Y)$) as well as a Stein matrix $\tau_X(X)$ (resp., $\tau_Y(Y)$). Then, for all $0 < t < 1$, we have*

$$\begin{aligned}&\rho_{W_t}(W_t) + \Gamma_t^{-1}W_t \\ &\quad = E\left[\frac{t}{\sqrt{1-t}}(Id - \Gamma_t^{-1}\tau_X(X))\rho_Y(Y) + \frac{1-t}{\sqrt{t}}(Id - \Gamma_t^{-1}\tau_Y(Y))\rho_X(X) \,\Big|\, W_t\right]\end{aligned} \tag{21}$$

with $W_t = \sqrt{t}X + \sqrt{1-t}Y$ and Γ_t the covariance matrix of W_t.

Proof. Let $\varphi \in C_c^1$ be a test function. Applying first (1) (with respect to Y) then (2) (with respect to X) we get

$$\begin{aligned}
&E\left[E\left[(Id - \Gamma_t^{-1}\tau_X(X))\rho_Y(Y)\,|\,W_t\right]\varphi(W_t)\right] \\
&= E\left[(Id - \Gamma_t^{-1}\tau_X(X))\rho_Y(Y)\varphi(W_t)\right] \\
&= -\sqrt{1-t}E\left[(Id - \Gamma_t^{-1}\tau_X(X))\nabla\varphi(W_t)\right] \\
&= -\sqrt{1-t}\left(E\left[\nabla\varphi(W_t)\right] - \Gamma_t^{-1}\frac{1}{\sqrt{t}}E\left[X\varphi(W_t)\right]\right).
\end{aligned}$$

Likewise

$$\begin{aligned}
&E\left[E\left[(Id - \Gamma_t^{-1}\tau_Y(Y))\rho_X(X)\,|\,W_t\right]\varphi(W_t)\right] \\
&= -\sqrt{t}\left(E\left[\nabla\varphi(W_t)\right] - \Gamma_t^{-1}\frac{1}{\sqrt{1-t}}E\left[Y\varphi(W_t)\right]\right).
\end{aligned}$$

Hence

$$\begin{aligned}
&E\left[E\left[\frac{t}{\sqrt{1-t}}(Id - \Gamma_t^{-1}\tau_X(X))\rho_Y(Y) + \frac{1-t}{\sqrt{t}}(Id - \Gamma_t^{-1}\tau_Y(Y))\rho_X(X)\,\Big|\,W_t\right]\varphi(W_t)\right] \\
&= -E\left[\nabla\varphi(W_t)\right] + E\left[\Gamma_t^{-1}W_t\varphi(W_t)\right] = E\left[(\rho_{W_t}(W_t) + \Gamma_t^{-1}W_t)\varphi(W_t)\right],
\end{aligned}$$

and the conclusion (21) follows. □

In the sequel we simply write ρ_t instead of ρ_{W_t}. In [18] we use a version of (21) specialised to the case where X has covariance C and $Y = Z$ is a Gaussian random vector also with covariance C. Then $\Gamma_t = C$ and, setting here and throughout $X_t = \sqrt{t}X + \sqrt{1-t}Z$, we get

$$\rho_t(X_t) + C^{-1}X_t = -\frac{t}{\sqrt{1-t}}E\left[\left(Id - C^{-1}\tau_X(X)\right)C^{-1}Z\,|\,X_t\right] \tag{22}$$

for all $0 < t < 1$ (recall that $\rho_Z(Z) = -C^{-1}Z$ and $\tau_Z(Z) = C$). Taking squares we obtain the following representations for the (standardized) Fisher information of an arbitrary random vector with density.

Theorem 3.2. *For all* $0 < t < 1$

$$I(X_t) = \frac{t^2}{1-t}E\left[E\left[\left(Id - C^{-1}\tau_X(X)\right)C^{-1}Z\,|\,X_t\right]^2\right] + C^{-1} \tag{23}$$

and

$$J_{st}(X_t) = \frac{t^2}{1-t}\mathrm{tr}\left(CE\left[E\left[\left(Id - C^{-1}\tau_X(X)\right)C^{-1}Z\,|\,X_t\right]^2\right]\right) \tag{24}$$

for all $0 < t < 1$.

Proof. Equation (24) follows from (23), (22) and (20) without further ado. We prove (23). Take $C = Id$ for simplicity and apply (22) to get

$$\begin{aligned} I(X_t) &= \frac{t^2}{1-t} E\left[E\left[(Id - \tau_X(X))\, Z \,|\, X_t \right]^2 \right] + E\left[X_t X_t^T \right] \\ &- \frac{t}{\sqrt{1-t}} E\left[E\left[(Id - \tau_X(X))\, Z \,|\, X_t \right] X_t^T \right] \\ &- \frac{t}{\sqrt{1-t}} E\left[X_t E\left[(Id - \tau_X(X))\, Z \,|\, X_t \right]^T \right]. \end{aligned}$$

It is immediate that

$$E\left[E\left[(Id - \tau_X(X))\, Z \,|\, X_t \right] X_t^T \right] = E\left[X_t E\left[(Id - \tau_X(X))\, Z \,|\, X_t \right]^T \right] = 0.$$

The conclusion follows. □

4 Connection with a formula of Guo, Shamai and Verdú

It was brought to our attention (by Oliver Johnson, personal communications) that representation (22) resembled, at least in principle, an identity for Fisher information discovered in [8]. An explicit connection between the two approaches is easily obtained, as follows. We start with a different proof of their identity.

Lemma 4.1 (Guo, Shamai and Verdú [8])**.** *Let X be a centered random vector with covariance C and let Z be Gaussian with the same covariance as X. Then, for all $0 < t < 1$, the random vector $X_t = \sqrt{t}X + \sqrt{1-t}Z$ has a score*

$$\rho_t(X_t) = -\frac{1}{1-t} C^{-1} \left(X_t - \sqrt{t} E\left[X \,|\, X_t \right] \right) \tag{25}$$

and its Fisher information is

$$I(X_t) = \frac{1}{1-t} C^{-1} - \frac{t}{(1-t)^2} C^{-1} E\left[\left(X - E\left[X \,|\, X_t \right] \right) \left(X - E\left[X \,|\, X_t \right] \right)^T \right] C^{-1}. \tag{26}$$

Proof. Clearly X_t has a differentiable density with support $\mathbb{R}^d$. Let φ be a test function. Then (once again recall $\rho_Z(Z) = -C^{-1}Z$)

$$\begin{aligned} E\left[C^{-1} \left(X_t - \sqrt{t} E\left[X \,|\, X_t \right] \right) \varphi(X_t) \right] &= E\left[C^{-1} \left(X_t - \sqrt{t} X \right) \varphi(X_t) \right] \\ &= \sqrt{1-t} E\left[C^{-1} Z \varphi(X_t) \right] \\ &= (1-t) E\left[\nabla \varphi(X_t) \right]. \end{aligned}$$

The first identity ensues. Next, to prove (26), we simply use (25) and compute

$$\begin{aligned} I(X_t) &= \frac{1}{(1-t)^2} C^{-1} E\left[\left(X_t - \sqrt{t} E\left[X \,|\, X_t \right] \right) \left(X_t - \sqrt{t} E\left[X \,|\, X_t \right] \right)^T \right] C^{-1} \\ &= \frac{1}{(1-t)^2} C^{-1} \left\{ E\left[X_t X_t^T \right] + t E\left[E\left[X \,|\, X_t \right] E\left[X \,|\, X_t \right]^T \right] \right. \\ &\left. - \sqrt{t} E\left[X_t E\left[X \,|\, X_t \right]^T \right] - \sqrt{t} E\left[E\left[X \,|\, X_t \right] X_t^T \right] \right\} C^{-1}. \end{aligned}$$

By independence of X and Z we can write

$$E\left[X_t E\left[X \mid X_t\right]^T\right] = E\left[X_t X^T\right] = \sqrt{t}C = E\left[E\left[X \mid X_t\right] X_t^T\right]$$

so that

$$\begin{aligned} I(X_t) &= \frac{1}{(1-t)^2} C^{-1}\left\{C - 2tC + tE\left[E\left[X \mid X_t\right] E\left[X \mid X_t\right]^T\right]\right\} C^{-1} \\ &= \frac{1}{(1-t)^2} C^{-1}\left\{C(1-t) - t\left(C - E\left[E\left[X \mid X_t\right] E\left[X \mid X_t\right]^T\right]\right)\right\} C^{-1}. \end{aligned}$$

Finally, we have

$$\begin{aligned} &C - E\left[E\left[X \mid X_t\right] E\left[X \mid X_t\right]^T\right] = E\left[XX^T\right] - E\left[E\left[X \mid X_t\right] E\left[X \mid X_t\right]^T\right] \\ &= E\left[\left(X - E\left[X \mid X_t\right]\right)\left(X - E\left[X \mid X_t\right]\right)^T\right], \end{aligned}$$

whence the claim. □

As in [8] we define the MMSE as the minimal mean square error when estimating X with observation X_t, that is,

$$\mathrm{MMSE}(X,t) = E\left[\left(X - E\left[X \mid X_t\right]\right)\left(X - E\left[X \mid X_t\right]\right)^T\right]. \tag{27}$$

Combining the previous relationships we obtain the connection between our identities and theirs (results are obvious and stated without proof).

Proposition 4.2. *Let all notations be as above, and let $\tau_X(X)$ be a Stein matrix for X. Let A_t be as above. Then*

$$Id - \frac{1}{1-t}\mathrm{MMSE}(X,t)C^{-1} = tCE\left[E\left[\left(Id - C^{-1}\tau_X(X)\right) C^{-1}Z \mid X_t\right]^2\right] \tag{28}$$

so that

$$J_{st}(X_t) = \frac{t}{1-t}\mathrm{tr}\left(Id - \frac{1}{(1-t)}\mathrm{MMSE}(X,t)C^{-1}\right). \tag{29}$$

5 Information bounds for sums of random vectors

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent random vectors having a density on $\mathbb{R}^d$. We first extend (21) to an arbitrary number of summands. Proof is immediate by following the same route and is left to the reader.

Lemma 5.1. *For all $t = (t_1, \ldots, t_n) \in [0,1]^d$ such that $\sum_{i=1}^n t_i = 1$ we define $W_t = \sum_{i=1}^n \sqrt{t_i}X_i$ and denote Γ_t the corresponding covariance matrix. Then*

$$\rho_t(W_t) + \Gamma_t^{-1}W_t = \sum_{i=1}^n \frac{t_i}{\sqrt{t_{i+1}}} E\left[\left(Id - \Gamma_t^{-1}\tau_i(X_i)\right)\rho_{i+1}(X_{i+1}) \mid W_t\right] \tag{30}$$

where we identify $X_{n+1} = X_1$ and $t_{n+1} = t_1$, and where we set $\tau_i = \tau_{X_i}$, $\rho_i = \rho_{X_i}$ and $\rho_t = \rho_{W_t}$ for simplicity.

In the sequel we suppose (for simplicity) that the X_i are isotropic, that is, we suppose that their covariance matrix is the identity. We first prove that Fisher information distance is bounded by a measure of the discrepancy between the Stein matrix and the identity.

Theorem 5.2. *Let $W_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i$ where the X_i have Stein matrix $\tau_i(X_i)$ and score function $\rho_i(X_i)$. Define $W_n^{(t)} = \sqrt{t}W_n + \sqrt{1-t}Z$. Then*

$$J_{st}(W_n^{(t)}) \leq \frac{1}{n^2}\frac{t^2}{1-t}\sum_{i=1}^n \mathrm{tr}\left(E\left[(Id-\tau_i(X_i))\,(Id-\tau_i(X_i))^T\right]\right) \tag{31}$$

for all $0 \leq t \leq 1$. If the X_i are i.i.d. copies of X then

$$J_{st}(W_n^{(1/2)}) \leq \frac{1}{2n}\mathrm{tr}\left(E\left[(Id-\tau_X(X))\,(Id-\tau_X(X))^T\right]\right). \tag{32}$$

Proof. First note that Stein matrices behave similarly as score functions over convolutions, in the sense that

$$\tau_{W_n}(W_n) = \frac{1}{n}\sum_{i=1}^n E\left[\tau_i(X_i)\,|\,W_n\right], \tag{33}$$

(see [21, 18] for a proof). Hence, by Jensen's inequality in (24),

$$\begin{aligned}
J_{st}(W_n^{(t)}) &\leq \frac{t^2}{1-t}\mathrm{tr}\left(E\left[(Id-\tau_{W_n}(W_n))(Id-\tau_{W_n}(W_n))^T\right]\right)\\
&= \frac{t^2}{1-t}\mathrm{tr}\left(E\left[\left(\frac{1}{n}\sum_{i=1}^n(Id-E\left[\tau_i(X_i)\,|\,W_n\right])\right)^2\right]\right)\\
&\leq \frac{1}{n^2}\frac{t^2}{1-t}\mathrm{tr}\left(E\left[\left(\sum_{i=1}^n(Id-\tau_i(X_i))\right)^2\right]\right)
\end{aligned}$$

Independence of the X_i as well as the fact that $E\left[Id-\tau_i(X_i)\right]=0$ allow to deduce the first claim. The second claim is then immediate. □

By Cramer's theorem [12], convergence of W_n to the Gaussian is equivalent to convergence of $W_n^{(1/2)}$, and (32) provides rates of convergence (of order $1/n$) of the Fisher information under the assumption that X has a well-defined Stein matrix $\tau_X(X)$. Using Pinsker's inequality as well as (15), we then obtain rates of convergence in total variation which have the correct order.

6 Stein representations for Fisher information

In all the above identities, one translates the problem of controlling the variance of a random vector into that of control the squared conditional expectation of another random vector. Our next lemma (communicated to us by Guillaume Poly) exploits the duality representation of the norm on a Hilbert space to allow to control moments of conditional expectations very efficiently.

Lemma 6.1. *Let X and Y be square-integrable random variables. Assume moreover that $E[X] = 0$. Then*

$$E\left[(E[X \mid Y])^2\right] = \sup_{\varphi \in \mathcal{H}(Y)} (E[X\varphi(Y)])^2, \tag{34}$$

where the supremum is taken over the collection $\mathcal{H}(Y)$ of test functions φ such that $E[\varphi(Y)] = 0$ and $E\left[\varphi(Y)^2\right] \leq 1$.

Proof. First, by Cauchy-Schwarz,

$$\begin{aligned}\sup_{\varphi \in \mathcal{H}(Y)} (E[X\varphi(Y)])^2 &= \sup_{\varphi \in \mathcal{H}(Y)} (E[E[X|Y]\varphi(Y)])^2 \\ &\leq \sup_{\varphi \in \mathcal{H}(Y)} E\left[E[X|Y]^2\right] E\left[\varphi(Y)^2\right] \leq E\left[E[X|Y]^2\right].\end{aligned}$$

To prove the reverse inequality define $\varphi(y) = E[X|Y = y] / \sqrt{E[E[X|Y]^2]}$. Clearly $E[\varphi(Y)] = 0$ and $E[\varphi(Y)^2] \leq 1$ so that $\varphi \in \mathcal{H}(Y)$ and

$$\begin{aligned}\sup_{\varphi \in \mathcal{H}(Y)} (E[X\varphi(Y)])^2 &\geq \left(E\left[X\frac{E[X|Y]}{\sqrt{E[E[X|Y]^2]}}\right]\right)^2 \\ &= \frac{(E[XE[X|Y]])^2}{E[E[X|Y]^2]} = E\left[E[X|Y]^2\right].\end{aligned}$$

Equality ensues. □

We immediately deduce an original proof (not relying on Stein's method!) of a recently discovered fact (see e.g. [13, 14]) that the Fisher information distance is dominated by expressions which appear naturally within the context of Stein's method.

Theorem 6.2 (Stein representation for Fisher information)**.** *Let $W_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$ where the X_i have Stein matrix $\tau_i(X_i)$ and score function $\rho_i(X_i)$. Then*

$$\mathcal{J}(W_n) = \sup_{\varphi \in \mathcal{H}(W_n)} (E[\nabla\varphi(W_n) - W_n\varphi(W_n)])^2. \tag{35}$$

Proof. We combine (30) (in the special case $t_1 = t_2 = \ldots = t_n = 1/n$) and (34) to deduce (we abuse of notation by writing A^2 instead of AA^T)

$$\begin{aligned}\mathcal{J}(W_n) &= E\left[(\rho_n(W_n) + W_n)^2\right] \\ &= E\left[\left(\frac{1}{\sqrt{n}}E\left[\left(\sum_{i=1}^n (Id - \tau_i(X_i))\rho_{i+1}(X_{i+1})\right) \mid W_n\right]\right)^2\right] \\ &= \frac{1}{n} \sup_{\varphi \in \mathcal{H}(W_n)} \left(\sum_{i=1}^n E[(Id - \tau_i(X_i))\rho_{i+1}(X_{i+1})\varphi(W_n)]\right)^2 \\ &= \frac{1}{n} \sup_{\varphi \in \mathcal{H}(W_n)} E\left[\sqrt{n}\nabla\varphi(W_n) - \sum_{i=1}^n X_i\varphi(W_n)\right]^2,\end{aligned}$$

and the conclusion follows. □

7 Acknowledgements

We are grateful to Oliver Johnson for providing the connexion with [23] and to Guillaume Poly for sharing his Lemma 6.1 with us.

A Appendix: Non-uniqueness of the Stein matrix

Take two random matrices τ and $\tilde{\tau}$ satisfying (2) and define $\Psi(X) = \tau_X(X) - \tilde{\tau}_X(X) = (\Psi_1(X), \Psi_2(X), \ldots, \Psi_d(X))^T$. Then, for all $1 \leq i \leq d$, the vector-valued function Ψ_i is integrable (with respect to X) with expectation 0 and satisfies

$$E\left[\Psi_i(X)\nabla\varphi(X)\right] = E\left[\frac{\mathrm{div}(\Psi_i(X)f(X))}{f(X)}\varphi(X)\right] = 0$$

for all test functions φ (with div the divergence operator), with f the density of X. In other words $\tau_X(X)$ and $\tilde{\tau}_X(X)$ are Stein matrices for X if and only if $\Psi(x) = \tau_X(x) - \tilde{\tau}_X(x)$ satisfies $E\left[\Psi(X)\right] = 0$ and

$$\mathrm{div}(\Psi_i(x)f(x)) = 0 \tag{36}$$

almost surely for all $1 \leq i \leq d$. Of course there are many ways to construct a matrix function Ψ such that $\mathrm{div}(\Psi_i(x)f(x)) = F(x)$ for any given function F, see for instance the interesting discussion at end of the proof of [2, Theorem 4].

Example A.1. *We provide an explicit construction in case $d = 2$ and $(X, Y) \sim f$ some regular 2-dimensional distribution. Let $\Psi_j = (\Psi_{j1}, \Psi_{j2})$ for $j = 1, 2$. Then* (36) *is satisfied as soon as $\partial_x\left(\Psi_{j1}(x, y)f(x, y)\right) = -\partial_y\left(\Psi_{j2}(x, y)f(x, y)\right)$ for all $(x, y) \in \mathbb{R}^2$. Integrating, we see that for any choice of sufficiently regular Ψ_{j2} such that $E\left[\Psi_{j2}(X, Y)\right] = 0$ it suffices to take*

$$\Psi_{j1}(x, y)f(x, y) = -\int_{-\infty}^{x} \partial_y\Psi_{j2}(u, y)f(u, y)du.$$

to ensure that (36) *is satisfied. Moreover, the condition $E\left[\Psi_j(X, Y)\right] = 0$ is then automatically satisfied. Hence, letting $W = (X, Y)$, if $\tau_W(W)$ is a Stein matrix for W then so is*

$$\tau_W(W) + \begin{pmatrix} \Psi_{11}(X, Y) & \Psi_{12}(X, Y) \\ \Psi_{21}(X, Y) & \Psi_{22}(X, Y) \end{pmatrix}$$

for Ψ_{ij} as constructed above. Thus the Stein matrix is not unique.

Example A.2. *Starting from the condition $E\left[\Psi_i(X)\nabla\varphi(X)\right] = 0$, the furthest we can go with respect to the identification of the Stein matrix of X is by requiring that $E\left[\tau\right] = E\left[\tilde{\tau}\right]$ and*

$$E\left[\tau_{ij}(X) \,|\, X_j\right] = E\left[\tilde{\tau}_{ij}(X) \,|\, X_j\right];$$

this conclusion is reached by considering all test functions $\varphi(X) = \varphi(X_i)$ for $1 \leq i \leq d$.

Example A.3. *If we further require in the definition that the Stein matrix be symmetric then we can take Ψ as in Example A.1 with the further requirement that $\Psi_{21}(X,Y) = \Psi_{12}(X,Y)$; the Stein matrix remains non-unique. Choosing $\Psi_{22}(x,y) = \Psi_{11}(x,y)$ then the function Ψ_2 not only need have mean 0 with respect to the law f but also (using (4)) it needs to satisfy the wave equation*

$$\partial_x^2 \left(\Psi_2(x,y)f(x,y)\right) = \partial_y^2 \left(\Psi_2(x,y)f(x,y)\right) \tag{37}$$

for all (x,y). Still it is not unique.

References

[1] H. Airault, P. Malliavin, and F. Viens. Stokes formula on the Wiener space and n-dimensional Nourdin-Peccati analysis. *Journal of Functional Analysis*, 258(5):1763–1783, 2010.

[2] S. Artstein, K. Ball, F. Barthe, and A. Naor. Solution of Shannon's problem on the monotonicity of entropy. *Journal of the American Mathematical Society*, 17(4):975–982, 2004.

[3] K. Ball, F. Barthe, and A. Naor. Entropy jumps in the presence of a spectral gap. *Duke Math. J.*, 119(1):41–63, 2003.

[4] K. Ball and V. Nguyen. Entropy jumps for random vectors with log-concave density and spectral gap. Preprint, arxiv:1206.5098v3, 2012.

[5] A. D. Barbour and L. H. Y. Chen. *An introduction to Stein's method*, volume 4 of *Lect. Notes Ser. Inst. Math. Sci. Natl. Univ. Singap.* Singapore University Press, Singapore, 2005.

[6] L. H. Y. Chen, L. Goldstein, and Q.-M. Shao. *Normal approximation by Stein's method.* Probability and its Applications (New York). Springer, Heidelberg, 2011.

[7] L. Goldstein and G. Reinert. Stein's method and the zero bias transformation with application to simple random sampling. *Ann. Appl. Probab.*, 7(4):935–952, 1997.

[8] D. Guo, S. Shamai, and S. Verdú. Mutual information and minimum mean-square error in gaussian channels. *Information Theory, IEEE Transactions on*, 51(4):1261–1282, 2005.

[9] O. Johnson. *Information theory and the central limit theorem.* Imperial College Press, London, 2004.

[10] O. Johnson and A. Barron. Fisher information inequalities and the central limit theorem. *Probab. Theory Related Fields*, 129(3):391–409, 2004.

[11] O. Johnson and Y. Suhov. Entropy and random vectors. *J. Statist. Phys.*, 104(1-2):145–192, 2001.

[12] A. Kagan. A multivariate analog of the Cramer theorem on components of the Gaussian distributions. In *Stability problems for stochastic models*, pages 68–77. Springer, 1989.

[13] C. Ley and Y. Swan. Local Pinsker inequalities via Stein's discrete density approach. *IEEE Trans. Info. Theory*, 59(9):5584–4491, 2013.

[14] C. Ley and Y. Swan. Stein's density approach and information inequalities. *Electron. Comm. Probab.*, 18(7):1–14, 2013.

[15] I. Nourdin and G. Peccati. Stein's method on Wiener chaos. *Probab. Theory Related Fields*, 145(1-2):75–118, 2009.

[16] I. Nourdin and G. Peccati. *Normal approximations with Malliavin calculus : from Stein's method to universality.* Cambridge Tracts in Mathematics. Cambridge University Press, 2012.

[17] I. Nourdin, G. Peccati and A. Réveillac. Multivariate normal approximation using Stein's method and Malliavin calculus . *Ann. I.H.P. Proba. Stat.*, 46(1):45–58, 2010.

[18] I. Nourdin, G. Peccati, and Y. Swan. Entropy and the fourth moment phenomenon. *Journal of Functional Analysis*, to appear, 2013.

[19] C. E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

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

[21] C. Stein. *Approximate computation of expectations.* Institute of Mathematical Statistics Lecture Notes—Monograph Series, 7. Institute of Mathematical Statistics, Hayward, CA, 1986.

[22] A. M. Tulino and S. Verdú. Monotonic decrease of the non-gaussianness of the sum of independent random variables: A simple proof. *Information Theory, IEEE Transactions on*, 52(9):4295–4297, 2006.

[23] S. Verdú and D. Guo. A simple proof of the entropy-power inequality. *IEEE Transactions on Information Theory*, 52(5):2165–2166, 2006.