Aequat. Math. 90 (2016), 817–848

0001-9054/16/040817-32
published online January 13, 2016
DOI 10.1007/s00010-015-0396-5

Aequationes Mathematicae

CrossMark

Basic inequalities for weighted entropies

Yuri Suhov, Izabella Stuhl, Salimeh Yasaei Sekeh
and Mark Kelbert

Abstract. The concept of weighted entropy takes into account values of different outcomes, i.e., makes entropy context-dependent, through the weight function. In this paper, we establish a number of simple inequalities for weighted entropies (general as well as specific), mirroring similar bounds on standard (Shannon) entropies and related quantities. The required assumptions are written in terms of various expectations of weight functions. Examples are weighted Ky Fan and weighted Hadamard inequalities involving determinants of positive-definite matrices, and weighted Cramér-Rao inequalities involving the weighted Fisher information matrix.

Mathematics Subject Classification. 60A10, 60B05, 60C05.

Keywords. Weighted entropy, weighted conditional entropy, weighted relative entropy, weighted mutual entropy, weighted Gibbs inequality, convexity, concavity, weighted Hadamard inequality, weighed Fisher information, weighed Cramér-Rao inequalities.

1. The weighted Gibbs inequality and its consequences

The definition and initial results on weighted entropy were introduced in [1,11]. The purpose was to introduce disparity between outcomes of the same probability: in the case of a standard entropy such outcomes contribute the same amount of information/uncertainty, which is appropriate in context-free situations. However, imagine two equally rare medical conditions, occurring with probability $p << 1$, one of which carries a major health risk while the other is just a peculiarity. Formally, they provide the same amount of information $-\log p$ but the value of this information can be very different. The weight, or a weight function, was supposed to fulfill this task, at least to a certain extent. The initial results have been further extended and deepened in [7,14,15,23–25,31], and, more recently, in [2,6,22,26,27]. Certain applications emerged, see [8,13], along with a number of theoretical suggestions.

The purpose of this note is to extend a number of inequalities, established previously for a standard (Shannon) entropy, to the case of the weighted entropy. We particularly mention Ky Fan- and Hadamard-type inequalities from [3,9,20] which are related to (standard) Gaussian entropies. Extended inequalities for weighted entropies already found applications and further developments in [28–30]. Another kind of bounds, weighted Cramér–Rao inequalities, may be useful in statistics.

An additional motivation for studying weighted entropy (WE) can be provided in the following questions. (I) What is the rate at which the WE is produced by a sample of a random process (and what could be an analog of the Shannon–McMillan–Breiman theorem)? (II) What would be an analog of Shannon's Second Coding theorem when an incorrect channel output causes a penalty but does not make the transmission session invalid? Properties of the WE established in the current paper could be helpful in this line of research.

One of the naturally emerging questions is about the form/structure of the weight function (WF). In this paper we focus on some simple inequalities (as suggested by the title). Our results hold for fairly general WFs, subject to some mild conditions (in the form of inequalities). A systematic verification of these conditions may require a separate work.

Let $(\Omega, \mathfrak{B}, \mathbb{P})$ be a standard probability space (see, e.g., [12]). We consider random variables (RVs) as (measurable) functions $\Omega \to \mathcal{X}$, with values in a measurable space $(\mathcal{X}, \mathfrak{M})$ equipped with a countably additive reference measure ν. Probability mass functions (PMFs) or probability density functions (PDFs) are denoted by the letter f with various indices and defined relative to ν. The difference between PMFs (discrete parts of probability measures) and PDFs (continuous parts) is insignificant for most of the presentation; this will be reflected in a common acronym PM/DF. In a few cases we will address directly the probabilities $\mathbb{P}(X = i)$ (when $\mathcal{X}$ is a finite or countable set, assuming that $\nu(i) = 1\ \forall\ i \in \mathcal{X}$). On the other hand, some important facts will remain true without the assumption that $\displaystyle\int_{\mathcal{X}} f(x)\nu(\mathrm{d}x) = 1$. When we deal with a collection of RVs X_i, the space of values $\mathcal{X}_i$ and the reference measure ν_i may vary with i. Some RVs X_i may be random $1 \times n$ vectors, viz., $\mathbf{X}_1^n = (X_1, \ldots, X_n)$, with random components $X_i : \Omega \to \mathcal{X}_i$, $1 \le i \le n$.

Definition 1.1. Given a function $x \in \mathcal{X} \mapsto \varphi(x) \ge 0$, and an RV $X :\ \Omega \to \mathcal{X}$, with a PM/DF f, the **weighted entropy** (WE) of X (or f) with weight function (WF) φ and reference measure ν is defined by

$$h^{\mathrm{w}}_{\varphi}(X) = h^{\mathrm{w}}_{\varphi}(f) = -\mathbb{E}[\varphi(X) \log f(X)] = -\int_{\mathcal{X}} \varphi(x) f(x) \log f(x) \nu(\mathrm{d}x) \tag{1.1}$$

whenever the integral $\int_{\mathcal{X}} \varphi(x) f(x) \Big(1 \vee |\log f(x)|\Big) \nu(\mathrm{d}x) < \infty$. (A standard agreement $0 = 0 \cdot \log 0 = 0 \cdot \log \infty$ is adopted throughout the paper.) If

$f(x) \leq 1 \ \forall \ x \in \mathcal{X}$, $h^{\rm w}_\varphi(f)$ is non-negative. (This is the case when $\nu(\mathcal{X}) \leq 1$.) The dependence of $h^{\rm w}_\varphi(X) = h^{\rm w}_\varphi(f)$ on ν is omitted.

Given two functions, $x \in \mathcal{X} \mapsto f(x) \geq 0$ and $x \in \mathcal{X} \mapsto g(x) \geq 0$, the **relative** WE of g relative to f with WF φ is defined by

$$D^{\rm w}_\varphi(f\|g) = \int_{\mathcal{X}} \varphi(x) f(x) \log \frac{f(x)}{g(x)} \nu({\rm d}x). \tag{1.2}$$

Alternatively, the quantity $D^{\rm w}_\varphi(f\|g)$ can be termed a weighted Kullback–Leibler divergence (of g from f) with WF φ. If f is a PM/DF, one can use an alternative form of writing:

$$D^{\rm w}_\varphi(f\|g) = \mathbb{E}\Big[\varphi(X) \log \frac{f(X)}{g(X)}\Big].$$

In what follows, all WFs are assumed non-negative and positive on a set of positive f-measure.

Remark 1.2. Passing to standard entropies, an obvious formula reads as

$$h^{\rm w}_\varphi(f) = h(\varphi f) + D(\varphi f\|f) = -D(\varphi f\|\varphi), \tag{1.3}$$

provided that one can guarantee that the integrals involved converge. However, in general neither φf nor φ are PM/DFs, which can be a nuisance. Besides, the interpretation of φ as a weight function in $h^{\rm w}_\varphi(f)$ makes the inequalities more transparent.

Theorem 1.3. (The weighted Gibbs inequality; cf. [4], Lemma 1, [3], Theorem 2.6.3, [5] Lemma 1, [20], Theorem 1.2.3 (c)) *Given non-negative functions f, g, assume the bound*

$$\int\limits_{\mathcal{X}} \varphi(x)\big[f(x) - g(x)\big]\nu({\rm d}x) \geq 0. \tag{1.4}$$

Then

$$D^{\rm w}_\varphi(f\|g) \geq 0. \tag{1.5}$$

Moreover, equality in (1.5) *holds iff the ratio* $\dfrac{g}{f}$ *equals* 1 *modulo function* φ. *In other words,* $\left[\dfrac{g(x)}{f(x)} - 1\right]\varphi(x) = 0$ *for f-almost all $x \in \mathcal{X}$.*

Proof. Following a standard calculation (see, e.g., [3], Theorem 2.6.3 or [20], Theorem 1.2.3 (c)) and using (1.2), we write

$$\begin{aligned} -D^{\mathrm{w}}_\varphi(f\|g) &= \int_{\mathcal{X}} \varphi(x)f(x)\mathbf{1}(f(x)>0)\log\frac{g(x)}{f(x)}\nu(\mathrm{d}x) \\ &\leq \int_{\mathcal{X}} \varphi(x)f(x)\mathbf{1}(f(x)>0)\left[\frac{g(x)}{f(x)}-1\right]\nu(\mathrm{d}x) \\ &= \int_{\mathcal{X}} \varphi(x)\mathbf{1}(f(x)>0)\big[g(x)-f(x)\big]\nu(\mathrm{d}x) \\ &\leq \int_{\mathcal{X}} \varphi(x)\big[g(x)-f(x)\big]\nu(\mathrm{d}x)\leq 0. \end{aligned} \tag{1.6}$$

The equality in (1.6) occurs iff $\varphi(g/f-1)$ vanishes f-a.s. □

Theorem 1.4. (Bounding the WE via a uniform distribution) *Suppose an RV X takes at most m values, i.e., $\mathcal{X}=\{1,\dots,m\}$, and set $p_i=\mathbb{P}(X=i)$, $1\leq i\leq m$. Suppose that for given $0<\beta\leq 1$*

$$\sum_{i=1}^m \varphi(i)(p_i-\beta)\geq 0. \tag{1.7}$$

Then $h^{\mathrm{w}}_\varphi(X)=-\sum_{i=1}^m\varphi(i)p_i\log p_i$ obeys

$$h^{\mathrm{w}}_\varphi(X)\leq -\log\beta\sum_{i=1}^m\varphi(i)p_i, \quad \text{or} \quad -\mathbb{E}[\varphi(X)\log p_X]\leq -(\log\beta)\mathbb{E}[\varphi(X)], \tag{1.8}$$

with equality iff for all $i=1,\dots,m$, $\varphi(i)(p_i-\beta)=0$.

In the case of a general space $\mathcal{X}$, assume that for a constant $\beta>0$ we have

$$\int_{\mathcal{X}}\varphi(x)\,[f(x)-\beta]\,\nu(\mathrm{d}x)\geq 0. \tag{1.9}$$

Then

$$h^{\mathrm{w}}_\varphi(X)\leq -\log\beta\int_{\mathcal{X}}\varphi(x)f(x)\nu(\mathrm{d}x); \tag{1.10}$$

with equality iff $\varphi(x)\,[f(x)-\beta]=0$ for f-almost all $x\in\mathcal{X}$.

Proof. The proof follows directly from Theorem 1.3, with $g(x)=\beta$, $x\in\mathcal{X}$. □

Definition 1.5. Let (X_1,X_2) be a pair of RVs $X_i:\ \Omega\to\mathcal{X}_i$, with a joint PM/DF $f(x_1,x_2)$, $x_i\in\mathcal{X}_i$, $i=1,2$, relative to measure $\nu_1(\mathrm{d}x_1)\times\nu_2(\mathrm{d}x_2)$, and marginal PM/DFs

$$\begin{aligned} f_1(x_1) &= \int_{\mathcal{X}_2} f(x_1,x_2)\nu_2(\mathrm{d}x_2), \quad x_1\in\mathcal{X}_1, \\ f_2(x_2) &= \int_{\mathcal{X}_1} f(x_1,x_2)\nu_1(\mathrm{d}x_1), \quad x_2\in\mathcal{X}_2. \end{aligned}$$

Let $(x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2 \mapsto \varphi(x_1, x_2)$ be a given WF. We use Eq. (1.1) to define the **joint** WE of X_1, X_2 with WF φ (under an assumption of absolute convergence of the integrals involved):

$$\begin{aligned} h^{\rm w}_\varphi(X_1, X_2) &= -\mathbb{E}[\varphi(X_1, X_2) \log f(X_1, X_2)] \\ &= -\int_{\mathcal{X}_1\times\mathcal{X}_2} \varphi(x_1, x_2) f(x_1, x_2) \log\, f(x_1, x_2)\nu_1({\rm d}x_1)\nu_2({\rm d}x_2). \end{aligned} \tag{1.11}$$

Next, the **conditional** WE of X_1 given X_2 with WF φ is defined by

$$\begin{aligned} h^{\rm w}_\varphi(X_1|X_2) &= -\mathbb{E}\Big[\varphi(X_1, X_2) \log\, \frac{f(X_1, X_2)}{f_2(X_2)}\Big] = h^{\rm w}_\varphi(X_1, X_2) - h^{\rm w}_{\psi_2}(X_2) \\ &= -\int_{\mathcal{X}_1\times\mathcal{X}_2} \varphi(x_1, x_2) f(x_1, x_2) \log\, \frac{f(x_1, x_2)}{f_2(x_2)}\nu_1({\rm d}x_1)\nu_2({\rm d}x_2), \end{aligned} \tag{1.12}$$

here and below

$$\psi_2(X_2) = \int_{\mathcal{X}_1} \varphi(x_1, x_2)\frac{f(x_1, x_2)}{f_2(x_2)}\nu_1({\rm d}x_1).$$

Further, the **mutual** WE between X_1 and X_2 is given by

$$\begin{aligned} i^{\rm w}_\varphi(X_1 : X_2) &= D^{\rm w}_\varphi(f\|f_1 \otimes f_2) = \mathbb{E}\left[\varphi(X_1, X_2) \log\, \frac{f(X_1, X_2)}{f_1(X_1)f_2(X_2)}\right] \\ &= \int_{\mathcal{X}_1\times\mathcal{X}_2} \varphi(x_1, x_2) f(x_1, x_2) \log \frac{f(x_1, x_2)}{f_1(x_1)f_2(x_2)}\nu_1({\rm d}x_1)\nu_2({\rm d}x_2). \end{aligned} \tag{1.13}$$

We will use the notation $\mathbf{X}_i^k = (X_i, \dots, X_k)$ and $\mathbf{x}_i^k = (x_i, \dots, x_k)$, $1 \le i < k \le n$, for collections of RVs and their sample values (particularly for pairs and triples of RVs) allowing us to shorten equations throughout the paper. In addition, we employ Cartesian products $\mathcal{X}_i^k = \mathcal{X}_i \times \dots \times \mathcal{X}_k$ and product-measures $\nu_i^k({\rm d}\mathbf{x}_i^k) = \nu_i({\rm d}x_i) \times \dots \times \nu_k({\rm d}x_k)$. Given a random $1 \times n$ vector $\mathbf{X}_1^n$ with a PM/DF f, we denote by f_i, f_{ij} and f_{ijk} the PM/DFs for component X_i, pair $\mathbf{X}_{ij} = (X_i, X_j)$ and triple $\mathbf{X}_{ijk} = (X_i, X_j, X_k)$, respectively. The arguments of f_i, f_{ij} and f_{ijk} are written as $x_i \in \mathcal{X}_i$, $\mathbf{x}_{ij} = (x_i, x_j) \in \mathcal{X}_{ij} = \mathcal{X}_i \times \mathcal{X}_j$ and $\mathbf{x}_{ijk} = (x_i, x_j, x_k) \in \mathcal{X}_{ijk} = \mathcal{X}_i \times \mathcal{X}_j \times \mathcal{X}_k$. Next, symbols $f_{i|j}$, $f_{ij|k}$ and $f_{i|jk}$ are used for conditional PM/DFs:

$$f_{i|j}(x_i|x_j) = \frac{f_{ij}(\mathbf{x}_{ij})}{f_j(x_j)}, \ \ f_{ij|k}(\mathbf{x}_{ij}|x_k) = \frac{f_{ijk}(\mathbf{x}_{ijk})}{f_k(x_k)}, \ \ f_{i|jk}(x_i|\mathbf{x}_{jk}) = \frac{f_{ijk}(\mathbf{x}_{ijk})}{f_{jk}(\mathbf{x}_{jk})}.$$

For a pair of RVs $\mathbf{X}_1^2$, set

$$\psi_1(x_1) = \int_{\mathcal{X}_2} \varphi(x_1, x_2) f_{2|1}(x_2|x_1)\nu_2({\rm d}x_2), \ x_1 \in \mathcal{X}_1; \tag{1.14}$$

the quantity $\psi_2(x_2)$, $x_2 \in \mathcal{X}_2$, is defined in a similar (symmetric) fashion. See above.

Next, given a triple of RVs $\mathbf{X}_1^3$, with a joint PM/DF $f(\mathbf{x}_1^3)$, set:

$$\begin{aligned}\psi_3^{12}(x_3) &= \int_{\mathcal{X}_1^2} \varphi(\mathbf{x}_1^3) f_{12|3}(\mathbf{x}_1^2|x_3)\nu_1^2(\mathrm{d}\mathbf{x}_1^2) = \mathbb{E}\Big[\varphi(\mathbf{X}_1^3)|X_3 = x_3\Big], \quad x_3 \in \mathcal{X}_3,\\ \psi_{12}(\mathbf{x}_1^2) &= \int_{\mathcal{X}_3} \varphi(\mathbf{x}_1^3) f_{3|12}(x_3|\mathbf{x}_1^2)\nu_3(\mathrm{d}x_3) = \mathbb{E}\Big[\varphi(\mathbf{X}_1^3)|\mathbf{X}_1^2 = \mathbf{x}_1^2\Big], \quad \mathbf{x}_1^2 \in \mathcal{X}_1^2,\end{aligned} \tag{1.15}$$

and define functions ψ_k^{ij} and ψ_{ij} for distinct labels $1 \le i, j, k \le 3$, in a similar manner.

Lemma 1.6. (Bounds on conditional WE, I) *Let $\mathbf{X}_1^2$ be a pair of RVs with a joint PM/DF $f(\mathbf{x}_1^2)$. Suppose that a WF $\mathbf{x}_1^2 \in \mathcal{X}_1^2 \mapsto \varphi(\mathbf{x}_1^2)$ obeys*

$$\mathbb{E}\Big[\varphi(\mathbf{X}_1^2)\big[f_{1|2}(X_1|X_2) - 1\big]\Big] = \int_{\mathcal{X}_1^2} \varphi(\mathbf{x}_1^2) f(\mathbf{x}_1^2)\Big[f_{1|2}(x_1|x_2) - 1\Big]\nu_1^2(\mathrm{d}\mathbf{x}_1^2) \le 0. \tag{1.16}$$

Then

$$h_\varphi^{\mathrm{w}}(\mathbf{X}_1^2) \ge h_{\psi_2}^{\mathrm{w}}(X_2), \quad \textit{or, equivalently,} \quad h_\varphi^{\mathrm{w}}(X_1|X_2) \ge 0, \tag{1.17}$$

with equality iff $\varphi(\mathbf{x}_1^2)\big[f_{1|2}(x_1|x_2) - 1\big] = 0$ *for f-almost all $\mathbf{x}_1^2 \in \mathcal{X}_1^2$.*

Proof. The statement is derived similarly to that of Theorem 1.3:

$$\begin{aligned}&\int_{\mathcal{X}_1^2} \varphi(\mathbf{x}_1^2) f(\mathbf{x}_1^2) \log\, f_{1|2}(x_1|x_2)\nu_1^2(\mathrm{d}\mathbf{x}_1^2)\\ &\qquad \le \int_{\mathcal{X}_1^2} \varphi(\mathbf{x}_1^2) f(\mathbf{x}_1^2)\Big[f_{1|2}(x_1|x_2) - 1\Big]\nu_1^2(\mathrm{d}\mathbf{x}_1^2).\end{aligned}$$

The argument is concluded as in (1.6). The cases of equalities also follow. □

Remark 1.7. In particular, suppose that X_1 takes finitely or countably many values and ν_1 is a counting measure with $\nu_1(i) = 1$, $i \in \mathcal{X}_1$. Then the value $f_{1|2}(x_1|x_2)$ yields the conditional probability $\mathbb{P}(X_1 = x_1|x_2)$, which is ≤ 1 for f_2-almost all $x_2 \in \mathcal{X}_2$. Then $h_\varphi^{\mathrm{w}}(X_1|X_2) \ge 0$, and the bound is strict unless, modulo φ, RV X_1 is a function of X_2. That is, there exists a map $\upsilon : \mathcal{X}_2 \to \mathcal{X}_1$ such that $\big[x_1 - \upsilon(x_2)\big]\varphi(\mathbf{x}_1^2) = 0$ for f-almost every $\mathbf{x}_1^2 \in \mathcal{X}_1^2$.

For future use, we can consider a triple of RVs, $\mathbf{X}_1^3$, and a pair, $\mathbf{X}_2^3$, and assume that

$$\mathbb{E}\Big[\varphi(\mathbf{X}_1^3)\big[f_{1|23}(X_1|\mathbf{X}_2^3) - 1\big]\Big] = \int_{\mathcal{X}_1^3} \varphi(\mathbf{x}_1^3) f(\mathbf{x}_1^3)\Big[f_{1|23}(x_1|\mathbf{x}_2^3) - 1\Big]\nu_1^3(\mathrm{d}\mathbf{x}_1^3) \le 0. \tag{1.18}$$

Then

$$h_\varphi^{\mathrm{w}}(\mathbf{X}_1^3) \ge h_{\psi_{23}}^{\mathrm{w}}(\mathbf{X}_2^3), \quad \text{or, equivalently,} \quad h_\varphi^{\mathrm{w}}(X_1|\mathbf{X}_2^3) \ge 0, \tag{1.19}$$

with equality iff $\varphi(\mathbf{x}_1^3)\big[f_{1|23}(x_1|\mathbf{x}_2^3) - 1\big] = 0$ for f-almost all $\mathbf{x}_1^3 \in \mathcal{X}_1^3$.

Theorem 1.8. (Sub-additivity of the WE) *Let* $\mathbf{X}_1^2 = (X_1, X_2)$ *be a pair of RVs with a joint PM/DF* $f(\mathbf{x}_1^2)$ *and marginals* $f_1(x_1)$, $f_2(x_2)$, *where* $\mathbf{x}_1^2 \in \mathcal{X}_1^2$. *Suppose that a WF* $\mathbf{x}_1^2 \in \mathcal{X}_1^2 \mapsto \varphi(\mathbf{x}_1^2)$ *obeys*

$$\mathbb{E}\varphi(\mathbf{X}_1^2) - \mathbb{E}\varphi(\mathbf{X}_{12}^{\otimes}) = \int_{\mathcal{X}_1^2} \varphi(\mathbf{x}_1^2)\Big[f(\mathbf{x}_1^2) - f_1(x_1)f_2(x_2)\Big]\nu_1^2(\mathrm{d}\mathbf{x}_1^2) \geq 0. \quad (1.20)$$

Here $\mathbf{X}_{12}^{\otimes}$ *stands for the pair of independent RVs having the same marginal distributions as* X_1, X_2. (*The joint PDF for* $\mathbf{X}_{12}^{\otimes}$ *is the product* $f_1(x_1)f_2(x_2)$.) *Then*

$$h_\varphi^{\mathrm{w}}(\mathbf{X}_1^2) \leq h_{\psi_1}^{\mathrm{w}}(X_1) + h_{\psi_2}^{\mathrm{w}}(X_2), \quad \textit{or, equivalently,} \quad h_\varphi^{\mathrm{w}}(X_1|X_2) \leq h_{\psi_1}^{\mathrm{w}}(X_1),$$
$$\textit{or, equivalently,} \quad i_\varphi^{\mathrm{w}}(X_1 : X_2) \geq 0. \quad (1.21)$$

The equalities hold iff X_1, X_2 *are independent modulo* φ, *i.e.,*

$$\varphi(\mathbf{x}_1^2)\left[1 - \frac{f_1(x_1)f_2(x_2)}{f(\mathbf{x}_1^2)}\right] = 0$$

for f*-almost all* $\mathbf{x}_1^2 \in \mathcal{X}_1^2$.

Proof. The subsequent argument works for the proof of Theorem 1.10 as well. Set $(f_1 \otimes f_2)(x_1, x_2) = f_1(x_1)f_2(x_2)$. According to (1.2), (1.11)–(1.13) and owing to Theorem 1.3 and Lemma 1.6,

$$\begin{aligned} 0 \geq -D_\varphi^{\mathrm{w}}(f\|f_1 \otimes f_2) &= \int_{\mathcal{X}_1^2} \varphi(\mathbf{x}_1^2)f(\mathbf{x}_1^2)\log\frac{f_1(x_1)f_2(x_2)}{f(\mathbf{x}_1^2)}\nu_1^2(\mathrm{d}\mathbf{x}_1^2) \\ &= h_\varphi^{\mathrm{w}}(X_1, X_2) - h_{\psi_1}^{\mathrm{w}}(X_1) - h_{\psi_2}^{\mathrm{w}}(X_2) \\ &= h_\varphi^{\mathrm{w}}(X_1|X_2) - h_{\psi_1}^{\mathrm{w}}(X_1) = -i_\varphi^{\mathrm{w}}(X_1 : X_2). \quad (1.22) \end{aligned}$$

This yields the inequalities in (1.21). The cases of equality are also identified from Theorem 1.3. □

Note that if in (1.20) we use the function $\psi_{12}(\mathbf{x}_1^2)$ emerging from triple $\mathbf{X}_1^3$, the assumption becomes

$$\begin{aligned} &\mathbb{E}\varphi(\mathbf{X}_1^3) - \mathbb{E}\varphi(\mathbf{X}_{12}^{\otimes} \to X_3) \\ &\quad = \int_{\mathcal{X}_1^3} \varphi(\mathbf{x}_1^3)\Big[f_{12}(\mathbf{x}_1^2) - f_1(x_1)f_2(x_2)\Big]f_{3|12}(x_3|\mathbf{x}_1^2)\nu_1^3(\mathrm{d}\mathbf{x}_1^3) \geq 0 \end{aligned} \quad (1.23)$$

and the conclusion

$$h_{\psi_{12}}^{\mathrm{w}}(X_1|X_2) \leq h_{\psi_1^{23}}^{\mathrm{w}}(X_1). \quad (1.24)$$

Here $\mathbf{X}_{12}^{\otimes} \to X_3$ denotes the triple of RVs where X_1 and X_2 have been made independent, keeping intact their marginal distributions, and X_3 has the same conditional PM/DF $f_{3|12}$ as within the original triple $\mathbf{X}_1^3$.

Lemma 1.9. (Bounds on conditional WE, II) *Let $\mathbf{X}_1^3$ be a triple of RVs, with a joint PM/DF $f(\mathbf{x}_1^3)$. Given a WF $\mathbf{x}_1^3 \mapsto \varphi(\mathbf{x}_1^3)$, assume that*

$$\mathbb{E}\Big[\varphi(\mathbf{X}_1^3)\big[f_{1|23}(X_1|\mathbf{X}_2^3)-1\big]\Big] = \int_{\mathcal{X}_1^3}\varphi(\mathbf{x}_1^3)f(\mathbf{x}_1^3)\Big[f_{1|23}(x_1|\mathbf{x}_2^3)-1\Big]\nu_1^3(\mathrm{d}\mathbf{x}_1^3) \le 0. \tag{1.25}$$

Then

$$h^{\mathrm{w}}_{\psi_{23}}(X_2|X_3) \le h^{\mathrm{w}}_{\varphi}(\mathbf{X}_1^2|X_3); \tag{1.26}$$

with equality iff $\varphi(\mathbf{x}_1^3)\big[f_{1|23}(x_1|\mathbf{x}_2^3)-1\big]=0$ *for f-almost all $\mathbf{x}_1^3 \in \mathcal{X}_1^3$.*

As in Remark 1.7, assume X_1 takes finitely or countably many values and $\nu_1(i)=1$, $i \in \mathcal{X}_1$. Then the value $f_{1|23}(x_1|\mathbf{x}_2^3)$ yields the conditional probability $\mathbb{P}(X_1=x_1|\mathbf{x}_2^3)$, for f_{23}-almost all $\mathbf{x}_2^3 \in \mathcal{X}_2^3$. Then $h^{\mathrm{w}}_{\varphi}(\mathbf{X}_1^2|X_3) \ge h^{\mathrm{w}}_{\psi_{23}}(X_2|X_3)$, with equality iff modulo φ, RV X_1 is a function of $\mathbf{X}_2^3$.

Proof. Observe that $h^{\mathrm{w}}_{\varphi}(\mathbf{X}_1^2|X_3) = h^{\mathrm{w}}_{\varphi}(\mathbf{X}_1^3) - h^{\mathrm{w}}_{\psi_3^{12}}(X_3)$ and $h^{\mathrm{w}}_{\psi_{23}}(X_2|X_3) = h^{\mathrm{w}}_{\psi_{23}}(\mathbf{X}_2^3) - h^{\mathrm{w}}_{\psi_3^{12}}(X_3)$, so that we need to prove that $h^{\mathrm{w}}_{\varphi}(\mathbf{X}_1^3) \ge h^{\mathrm{w}}_{\psi_{23}}(\mathbf{X}_2^3)$. The proof follows that of Lemma 1.6, with obvious modifications. $\square$

Of course, if we swap labels 1 and 3 in (1.25), assuming that

$$\mathbb{E}\varphi(\mathbf{X}_1^3)\Big[f_{3|12}(X_3|\mathbf{X}_1^2)-1\Big] = \int_{\mathcal{X}_1^3}\varphi(\mathbf{x}_1^3)f(\mathbf{x}_1^3)\Big[f_{3|12}(x_3|\mathbf{x}_1^2)-1\Big]\nu_1^3(\mathrm{d}\mathbf{x}_1^3) \le 0 \tag{1.27}$$

we get

$$h^{\mathrm{w}}_{\psi_{12}}(X_1|X_2) \le h^{\mathrm{w}}_{\varphi}(\mathbf{X}_{13}|X_2),$$

with equality iff $\varphi(\mathbf{x}_1^3)\big[f_{3|12}(x_3|\mathbf{x}_1^2)-1\big]=0$ for f-almost all $\mathbf{x}_1^3 \in \mathcal{X}_1^3$.

Theorem 1.10. (Sub-additivity of the conditional WE) *Let $\mathbf{X}_1^3$ be a triple of RVs, with a joint PM/DF f. Given a WF $\mathbf{x}_1^3 \mapsto \varphi(\mathbf{x}_1^3)$, assume the following bound*

$$\begin{aligned}\mathbb{E}\varphi(\mathbf{X}_1^3) &- \mathbb{E}\varphi(X_2 \to \mathbf{X}_{13}^{\otimes})\\ &= \int_{\mathcal{X}_1^3}\varphi(\mathbf{x}_1^3)\left[f(\mathbf{x}_1^3) - f_2(x_2)\prod_{i=1,3} f_{i|2}(x_i|x_2)\right]\nu_1^3(\mathrm{d}\mathbf{x}_1^3) \ge 0.\end{aligned} \tag{1.28}$$

Here $X_2 \to \mathbf{X}_{13}^{\otimes}$ stands for the triple of RVs where X_2 keeps its distribution as within the triple $\mathbf{X}_1^3$ whereas X_1 and X_3 have been made conditionally independent given X_2, with the same marginal conditional PDFs $f_{1|2}$ and $f_{3|2}$ as in $\mathbf{X}_1^3$. Then

$$h^{\mathrm{w}}_{\varphi}(\mathbf{X}_{13}|X_2) \le h^{\mathrm{w}}_{\psi_{12}}(X_1|X_2) + h^{\mathrm{w}}_{\psi_{32}}(X_3|X_2), \tag{1.29}$$

with equality iff, modulo φ, RVs X_1 and X_3 are conditionally independent given X_2. That is: $\varphi(\mathbf{x}_1^3)\Big[f(\mathbf{x}_1^3) - f_2(x_2)f_{1|2}(x_1|x_2)f_{3|2}(x_3|x_2)\Big] = 0$ for f-almost all $\mathbf{x}_1^3 \in \mathcal{X}_1^3$.

Proof. The proof is based on the equation (1.30):

$$\begin{aligned} & h^{\rm w}_{\varphi}(\mathbf{X}_{13}|X_2) - h^{\rm w}_{\psi_{12}}(X_1|X_2) - h^{\rm w}_{\psi_{32}}(X_3|X_2) \\ & \quad = \int_{\mathcal{X}_1^3} \varphi(\mathbf{x}_1^3) f(\mathbf{x}_1^3) \log \frac{f_{1|2}(x_1|x_2) f_{3|2}(x_3|x_2)}{f_{13|2}(\mathbf{x}_{13}|x_2)} \\ & \quad = \int_{\mathcal{X}_1^3} \varphi(\mathbf{x}_1^3) f(\mathbf{x}_1^3) \log \frac{f_2(x_2) f_{1|2}(x_1|x_2) f_{3|2}(x_3|x_2)}{f(\mathbf{x}_1^3)}. \end{aligned} \tag{1.30}$$

After that we apply the same argument as in (1.22). □

Lemma 1.11. (Bounds on conditional WE, III) *For a triple of RVs* $\mathbf{X}_1^3$ *with a joint PM/DF* $f(\mathbf{x}_1^3)$ *and a WF* $\mathbf{x}_1^3 \mapsto \varphi(\mathbf{x}_1^3)$, *assume the bound as in* (1.28). *Then*

$$\begin{aligned} h^{\rm w}_{\varphi}(X_1|\mathbf{X}_2^3) \le h^{\rm w}_{\psi_{12}}(X_1|X_2); \quad & \textit{with equality iff } X_1 \textit{ and } X_3 \\ & \textit{are conditionally independent given } X_2 \textit{ modulo } \varphi. \end{aligned} \tag{1.31}$$

Proof. Write (1.31) as

$$h_{\psi_{12}}(\mathbf{X}_1^2) - h_{\psi_2^{13}}(X_2) \ge h^{\rm w}_{\varphi}(\mathbf{X}_1^3) - h^{\rm w}_{\psi_{23}}(\mathbf{X}_2^3)$$

and then pass to an equivalent form $h^{\rm w}_{\varphi}(\mathbf{X}_{13}|X_2) \le h^{\rm w}_{\psi_{12}}(X_1|X_2) + h^{\rm w}_{\psi_{32}}(X_3|X_2)$ which is exactly (1.29). □

Summarizing, we have an array of inequalities (1.32) for $h^{\rm w}_{\varphi}(X_1|\mathbf{X}_2^3)$ and its upper bounds, each requiring its own assumption (and with its own case for equality):

$$\begin{array}{lll} \text{by Lemma 1.6:} & 0 & \le h^{\rm w}_{\varphi}(X_1|\mathbf{X}_2^3), \text{ assuming (1.18)} \\ & & \quad \text{(a modified form of (1.16))}, \\ \text{by Lemma 1.11:} & h^{\rm w}_{\varphi}(X_1|\mathbf{X}_2^3) & \le h^{\rm w}_{\psi_{12}}(X_1|X_2), \text{ assuming (1.28)}, \\ \text{by Theorem 1.8:} & h^{\rm w}_{\psi_{12}}(X_1|X_2) & \le h_{\psi_1^{23}}(X_1), \text{ assuming (1.23)}, \\ \text{by Lemma 1.9:} & h^{\rm w}_{\psi_{12}}(X_1|X_2) & \le h^{\rm w}_{\varphi}(\mathbf{X}_{13}|X_2), \text{ assuming (1.27)}, \\ \text{by Theorem 1.10:} & h^{\rm w}_{\varphi}(\mathbf{X}_{13}|X_2) & \le h^{\rm w}_{\psi_{12}}(X_1|X_2) + h^{\rm w}_{\psi_{32}}(X_3|X_2), \text{ assuming (1.28)}. \end{array} \tag{1.32}$$

It is worth noting that the assumptions listed in Eq. (1.32) express an impact on the total expected weight when we perform various manipulations with RVs forming a pair or a triple under consideration.

Theorem 1.12. (Strong sub-additivity of the WE) *Given a triple of RVs* $\mathbf{X}_1^3$, *assume that bound* (1.28) *is fulfilled. Then*

$$h^{\rm w}_{\varphi}(\mathbf{X}_1^3) + h^{\rm w}_{\psi_2^{13}}(X_2) \le h^{\rm w}_{\psi_{12}}(\mathbf{X}_1^2) + h^{\rm w}_{\psi_{23}}(\mathbf{X}_2^3). \tag{1.33}$$

The equality in (1.33) *holds iff, modulo* φ, X_1 *and* X_3 *are conditionally independent given* X_2.

Proof. Write the inequality in Eq. (1.33) in an equivalent form:

$$h^{\rm w}_{\varphi}(\mathbf{X}_1^3) - h^{\rm w}_{\psi_2^{13}}(X_2) \le h^{\rm w}_{\psi_{12}}(\mathbf{X}_1^2) - h^{\rm w}_{\psi_2^{13}}(X_2) + h^{\rm w}_{\psi_{23}}(\mathbf{X}_2^3) - h^{\rm w}_{\psi_2^{13}}(X_2). \tag{1.34}$$

The LHS in (1.34) equals $h^{\mathrm{w}}_{\varphi}(\mathbf{X}_{13}|X_2)$ while the RHS yields $h^{\mathrm{w}}_{\psi_{12}}(X_1|X_2) + h^{\mathrm{w}}_{\psi_{32}}(X_3|X_2)$. The inequality then follows from Theorem 1.10. $\square$

2. Convexity, concavity, data-processing and Fano inequalities

Theorem 2.1. (Concavity of the WE; cf. [3], Theorem 2.7.3) *The functional $f \mapsto h^{\mathrm{w}}_{\varphi}(f)$ is concave in argument f. Namely, for given PM/DFs $f_1(x)$, $f_2(x)$, a non-negative function $x \in \mathcal{X} \mapsto \varphi(x)$ and $\lambda_1, \lambda_2 \in [0,1]$ with $\lambda_1 + \lambda_2 = 1$,*

$$h^{\mathrm{w}}_{\varphi}(\lambda_1 f_1 + \lambda_2 f_2) \geq \lambda_1 h^{\mathrm{w}}_{\varphi}(f_1) + \lambda_2 h^{\mathrm{w}}_{\varphi}(f_2). \tag{2.1}$$

The inequality in (2.1) is strict unless one of the values λ_1, λ_2 vanishes (and the other equals 1) or when f_1 and f_2 coincide modulo φ, that is, $\varphi(x)\big[f_1(x) - f_2(x)\big] = 0$ for $(\lambda_1 f_1 + \lambda_2 f_2)$-almost all $x \in \mathcal{X}$.

Proof. Let $X_1, X_2 : \ \Omega \to \mathcal{X}$ be RVs with PM/DF f_1 and f_2, respectively. Consider a binary RV Θ with

$$\Theta = \begin{cases} 1, & \text{with probability } \lambda_1, \\ 2, & \text{with probability } \lambda_2. \end{cases} \tag{2.2}$$

Setting $Z = X_\theta$ yields an RV Z with values from $\mathcal{X}$ and with PM/DF $f = \lambda_1 f_1 + \lambda_2 f_2$. Thus,

$$h^{\mathrm{w}}_{\varphi}(Z) = h^{\mathrm{w}}_{\varphi}(\lambda_1 f_1 + \lambda_2 f_2).$$

On the other hand, take the conditional WE $h^{\mathrm{w}}_{\widetilde{\varphi}}(Z|\Theta)$ with the WF $\widetilde{\varphi}(z,\theta) = \varphi(z)$ depending on the first argument $z \in \mathcal{X}$ and not on the value $\theta = 1, 2$ of RV Θ. Then the WF $\psi_1(z) = \mathbb{E}\Big[\widetilde{\varphi}(Z,\Theta)|Z = z\Big]$ coincides with $\varphi(z)$. It means that condition (1.20) holds true for the pair of RVs Z, Θ. According to Theorem 1.8 [cf. Eq. (1.21)], $h^{\mathrm{w}}_{\widetilde{\varphi}}(Z|\Theta) \leq h^{\mathrm{w}}_{\varphi}(Z)$, with equality iff Z and Θ are independent modulo φ. The latter holds when the product $\lambda_1\lambda_2 = 0$ or when $f_1 = f_2$ modulo φ. Now,

$$h^{\mathrm{w}}_{\varphi}(Z|\Theta) = -\sum_{\theta=1}^{2} \lambda_\theta \int_{\mathcal{X}} \varphi(z) f_\theta(z) \log\, f_\theta(z) \nu(\mathrm{d}z) = \lambda_1 h^{\mathrm{w}}_{\varphi}(f_1) + \lambda_2 h^{\mathrm{w}}_{\varphi}(f_2).$$

This completes the proof. $\square$

Theorem 2.2. (a) (Convexity of relative WE; cf. [3], Theorem 2.7.2) *Consider two pairs of non-negative functions, (f_1, g_1) and (f_2, g_2), on $\mathcal{X}$. Given a WF $x \in \mathcal{X} \mapsto \varphi(x)$ and $\lambda_1\lambda_2 \in (0,1)$ with $\lambda_1 + \lambda_2 = 1$, the following property is satisfied:*

$$\lambda_1 D^{\mathrm{w}}_{\varphi}(f_1\|g_1) + \lambda_2 D^{\mathrm{w}}_{\varphi}(f_2\|g_2) \geq D^{\mathrm{w}}_{\varphi}(\lambda_1 f_1 + \lambda_2 f_2\|\lambda_1 g_1 + \lambda_2 g_2), \tag{2.3}$$

with equality iff $\lambda_1\lambda_2 = 0$ or $f_1 = f_2$ and $g_1 = g_2$ modulo φ.

(b) (Data-processing inequality for relative WE; cf. [3], Theorem 2.8.1) *Let* (f, g) *be a pair of non-negative functions and* φ *a WF on* $\mathcal{X}$. *Let* $\mathbf{\Pi} = (\Pi(x,y), x, y \in \mathcal{X})$ *be a stochastic kernel. (That is,* $\forall\ x, y \in \mathcal{X}$, $\Pi(x,y) \geq 0$ *and* $\displaystyle\int_{\mathcal{X}} \Pi(x,y)\nu(\mathrm{d}y) = 1$*; in other words,* $\Pi(x,y)$ *is a transition function of a Markov chain). Set* $\Psi(u) = \displaystyle\int_{\mathcal{X}} \varphi(x)\Pi(u,x)\nu(\mathrm{d}x)$. *Then*

$$D^{\mathrm{w}}_{\Psi}(f||g) \geq D^{\mathrm{w}}_{\varphi}(f\mathbf{\Pi} \,\|\, g\mathbf{\Pi}) \tag{2.4}$$

where $\big(f\mathbf{\Pi}\big)(x) = \displaystyle\int_{\mathcal{X}} f(u)\Pi(u,x)\nu(\mathrm{d}u)$ *and* $\big(g\mathbf{\Pi}\big)(x) = \displaystyle\int_{\mathcal{X}} g(u)\Pi(u,x)\nu(\mathrm{d}u)$. *The equality occurs iff* $f\mathbf{\Pi} = f$ *and* $g\mathbf{\Pi} = g$.

Proof. (a) The log-sum inequality yields

$$\begin{aligned} &\lambda_1\varphi(x)f_1(x)\log\frac{\lambda_1\varphi(x)f_1(x)}{\lambda_1 g_1(x)} + \lambda_2\varphi(x)f_2(x)\log\frac{\lambda_2\varphi(x)f_2(x)}{\lambda_2 g_2(x)} \\ &\quad\geq (\lambda_1\varphi(x)f_1(x) + \lambda_2\varphi(x)f_2(x))\log\frac{\lambda_1\varphi(x)f_1(x) + \lambda_2\varphi(x)f_2(x)}{\lambda_1 g_1(x) + \lambda_2 g_2(x)}, \quad x \in \mathcal{X}. \end{aligned} \tag{2.5}$$

Integrating in $\nu(\mathrm{d}x)$ yields the asserted inequality (2.3). The cases of equality emerge from the log-sum equality cases.

(b) Again, a straightforward application of the log-sum inequality gives the result. $\square$

Theorem 2.3. *Let* $\mathbf{X}_1^3$ *be a triple of RVs with joint PM/DF* $f(\mathbf{x}_1^3)$. *Let* $\mathbf{x}_1^3 \in \mathcal{X}_1^3 \mapsto \varphi(\mathbf{x}_1^3)$ *be a WF such that* X_1 *and* X_3 *are conditionally independent given* X_2 *modulo* φ. (*This property can be referred to as a Markov property modulo* φ.)

(a) (Data-processing inequality for conditional WE) *Assume inequality* (2.6) (*which is* (1.28) *with* X_1 *and* X_2 *swapped*)*:*

$$\int_{\mathcal{X}_1^3} \varphi(\mathbf{x}_1^3)\left[f(\mathbf{x}_1^3) - f_1(x_1)\prod_{i=2,3} f_{i|1}(x_i|x_1)\right]\nu_1^3(\mathrm{d}\mathbf{x}_1^3) \geq 0. \tag{2.6}$$

Then the conditional WEs satisfy property (2.7)*:*

$$h^{\mathrm{w}}_{\psi_{32}}(X_3|X_2) \leq h^{\mathrm{w}}_{\psi_{31}}(X_3|X_1), \tag{2.7}$$

with equality iff X_2 *and* X_3 *are independent modulo* φ. *Furthermore, assume in addition that bound* (2.8) *holds true*

$$\int_{\mathcal{X}_1^3} \varphi(\mathbf{x}_1^3)f(\mathbf{x}_1^3)\Big[f_{2|13}(x_2|\mathbf{x}_{13}) - 1\Big]\nu_1^3(\mathrm{d}\mathbf{x}_1^3) \leq 0 \tag{2.8}$$

(*which becomes* (1.25) *after a cyclic substitution* $X_1 \to X_2 \to X_3 \to X_1$) *and suppose* $h^{\rm w}_{\psi_{32}}(X_3|X_2) = h^{\rm w}_{\psi_{21}}(X_2|X_1)$ (*a stationarity-type property*). *Then*

$$h^{\rm w}_{\psi_{31}}(X_3|X_1) \leq 2h^{\rm w}_{\psi_{32}}(X_3|X_2). \tag{2.9}$$

(b) (Data-processing inequality for mutual WE; cf. [3], Theorem 2.8.1.) *Assume inequality* (2.10)*:*

$$\int_{\mathcal{X}_1^3} \varphi(\mathbf{x}_1^3)\Big[f(\mathbf{x}_1^3) - f_3(x_3)\prod_{i=1,2} f_{i|3}(x_i|x_3)\Big]\nu_1^3(\mathrm{d}\mathbf{x}_1^3) \geq 0 \tag{2.10}$$

(*similar to* (1.28), *with* X_3 *and* X_2 *swapped*). *Then*

$$i^{\rm w}_{\psi_{13}}(X_1 : X_3) \leq i^{\rm w}_{\psi_{12}}(X_1 : X_2). \tag{2.11}$$

Here, equality in (2.11) *holds iff, modulo* φ, *RVs* X_1 *and* X_2 *are conditionally independent given* X_3.

Proof. (a) Following the argument in Lemma 1.11, we observe that

$$h^{\rm w}_{\varphi}(X_3|\mathbf{X}_1^2) \leq h^{\rm w}_{\psi_{31}}(X_3|X_1).$$

On the other hand, owing to conditional independence,

$$h^{\rm w}_{\varphi}(X_3|\mathbf{X}_1^2) = h^{\rm w}_{\psi_{32}}(X_3|X_2). \tag{2.12}$$

This yields the inequality in (2.7); for equality we need that, modulo φ, RVs X_2 and X_3 are conditionally independent given X_1. Together with the conditional independence of X_1 and X_3 given X_2, it implies that for $i = 1, 2$, the conditional PM/DF $f_{3|i}$ does not depend on i.

Next, using Lemma 1.9, we can write

$$h^{\rm w}_{\psi_{31}}(X_3|X_1) \leq h^{\rm w}_{\varphi}(\mathbf{X}_2^3|X_1) := h^{\rm w}_{\varphi}(X_3|\mathbf{X}_1^2) + h^{\rm w}_{\psi_{21}}(X_2|X_1). \tag{2.13}$$

Applying (2.12) yields the following assertion:

$$h^{\rm w}_{\psi_{31}}(X_3|X_1) \leq h^{\rm w}_{\psi_{32}}(X_3|X_2) + h^{\rm w}_{\psi_{21}}(X_2|X_1). \tag{2.14}$$

Now, the assumption that $h^{\rm w}_{\psi_{32}}(X_3|X_2) = h^{\rm w}_{\psi_{21}}(X_2|X_1)$ implies (2.9). The cases of equality follow from Lemmas 1.11 and 1.9.

(b) As before, we use Lemma 1.11 and Eq. (2.12) (implied by conditional independence):

$$h^{\rm w}_{\psi_{12}}(X_1|X_2) = h^{\rm w}_{\varphi}(X_1|\mathbf{X}_2^3) \leq h^{\rm w}_{\psi_{13}}(X_1|X_3).$$

Consequently,

$$\begin{aligned} i^{\rm w}_{\psi_{12}}(X_1 : X_2) &= h^{\rm w}_{\psi_1^{23}}(X_1) - h^{\rm w}_{\psi_{12}}(X_1|X_2) \geq h^{\rm w}_{\psi_1^{23}}(X_1) - h^{\rm w}_{\psi_{13}}(X_1|X_3) \\ &= i^{\rm w}_{\psi_{13}}(X_1 : X_3), \end{aligned}$$

with the case of equality also determined from Lemma 1.9. □

Theorem 2.4. (Cf. [3], Theorem 2.7.4.) Let $\mathbf{X}_1^2$ be a pair of RVs with joint PM/DF $f(\mathbf{x}_1^2) = f_1(x_1)f_{2|1}(x_2|x_1) = f_2(x_2)f_{1|2}(x_1|x_2)$.

- (I) The mutual WE $i^{\rm w}_\varphi(X_1:X_2)$ is convex in $f_{2|1}(x_2|x_1)$ for fixed $f_1(X)$.
- (II) Suppose that the WF $\varphi(x_1,x_2)$ depends only on x_2: $\varphi(x_1,x_2)=\varphi(x_2)$. Then $i^{\rm w}_\varphi(X_1:X_2)$ is a concave function in $f_1(X)$ for fixed $f_{2|1}(x_2|x_1)$.

Proof. (I) For a fixed f_1, take two conditional PM/DFs, $f^{(1)}_{2|1}(x_2|x_1)$ and $f^{(2)}_{2|1}(x_2|x_1)$, and set

$$\widetilde{f}_{2|1}(x_2|x_1)=\lambda_1 f^{(1)}_{2|1}(x_2|x_1)+\lambda_2 f^{(2)}_{2|1}(x_2|x_1)$$

and

$$\widetilde{f}(\mathbf{x}_1^2)=f_1(x_1)\widetilde{f}_{2|1}(x_2|x_1)=\lambda_1 f^{(1)}(\mathbf{x}_1^2)+\lambda_2 f^{(2)}(\mathbf{x}_1^2)$$

where $f^{(j)}(\mathbf{x}_1^2)=f_1(x_1)f^{(j)}_{2|1}(x_2|x_1)$, $j=1,2$. Also, set:

$$\widetilde{f}_2(x_2)=\int_{\mathcal{X}_1}\widetilde{f}(\mathbf{x}_1^2)\nu_1^2(\mathrm{d}\mathbf{x}_1^2)\quad\text{and}\quad f^{(j)}_2(x_2)=\int_{\mathcal{X}_1}f^{(j)}(\mathbf{x}_1^2)\nu_1^2(\mathrm{d}\mathbf{x}_1^2)$$

and

$$\widetilde{g}(\mathbf{x}_1^2)=f_1(x_1)\widetilde{f}_2(x_2),\quad\text{and}\quad g^{(j)}(\mathbf{x}_1^2)=f_1(x_1)f^{(j)}_2(x_2),\ \ j=1,2.$$

Next, the mutual WE $i^{\rm w}_\varphi(X_1:X_2)$ for joint PM/DFs $\widetilde{f}(\mathbf{x}_1^2)$ and $f^{(j)}(\mathbf{x}_1^2)$ is given, respectively, by relative WEs

$$D^{\rm w}_\varphi(\widetilde{f}\|\widetilde{g})\quad\text{and}\quad D^{\rm w}_\varphi(f^{(j)}\|g^{(j)}),\ j=1,2.$$

Now assertion (I) follows from Theorem 2.2 (a).

(II) Under the condition of the theorem, the reduced WF does not depend on the choice of PM/DF f_1

$$\psi_2(x_2)=\int_{\mathcal{X}_1}\varphi(x_1,x_2)f_{1|2}(x_1|x_2)\nu_1(dx_1)=\varphi(x_2).$$

Next, write

$$\begin{aligned}i^{\rm w}_\varphi(X_1:X_2)&=h^{\rm w}_{\psi_2}(X_2)-h^{\rm w}_\varphi(X_2|X_1)\\&=h^{\rm w}_\varphi(X_2)-\textstyle\int_{\mathcal{X}_1^2}\varphi(x_2)f_1(x_1)f_{2|1}(x_2|x_1)\log f_{2|1}(x_2|x_1)\nu_1^2(\mathrm{d}\mathbf{x}_1^2)\\&=h^{\rm w}_\varphi(X_2)-\textstyle\int_{\mathcal{X}_1}f_1(x_1)h^{\rm w}_\varphi(X_2|X_1=x_1)\nu_1(\mathrm{d}x_1)\end{aligned}$$

where

$$h^{\rm w}_\varphi(X_2|X_1=x_1)=\int_{\mathcal{X}_2}\varphi(x_2)f_{2|1}(x_2|x_1)\log f_{2|1}(x_2|x_1)\nu_2(\mathrm{d}x_2).$$

Owing to Theorem 2.1, for fixed WF $x_2\mapsto\varphi(x_2)$ and conditional PM/DF $f_{2|1}(x_2|x_1)$, the WE $h^{\rm w}_\varphi$ is concave in f_1. The negative term is linear in f_1. This completes the proof of statement (II). $\square$

Theorem 2.5. (The weighted Fano inequality; cf. [3], Theorem 2.10.1, [20], Theorem 1.2.8)

Suppose an RV X takes a value $x^*\in\mathcal{X}$ with probability $p^*=\mathbb{P}(X=x^*)<1$ (i.e., $p^*=f(x^*)\nu(\{x^*\})$). Given a WF $x\in\mathcal{X}\mapsto\varphi(x)$, assume that

$$\int_{\mathcal{X}\setminus\{x^*\}}\varphi(x)\left[f(x)-\frac{1-p^*}{\nu(\mathcal{X}\setminus\{x^*\})}\right]\nu(\mathrm{d}x)\geq 0. \tag{2.15}$$

Then

$$h^{\mathrm{w}}_{\varphi}(X)\leq -\varphi(x^*)p^*\log\,p^*+\varphi_*\log\left(\frac{\nu(\mathcal{X}\setminus\{x^*\})}{1-p^*}\right). \tag{2.16}$$

Here $\varphi_*=\int_{\mathcal{X}}\varphi(x)\nu(\mathrm{d}x)-\varphi(x^*)p^*$.

The equality in (2.16) is achieved iff $\varphi(x)\left[f(x)-\dfrac{1-p^*}{\nu(\mathcal{X}\setminus\{x^*\})}\right]=0$, for f-almost all $x\in\mathcal{X}\setminus\{x^*\}$, i.e., iff RV X is (conditionally) uniform on $\mathcal{X}\setminus\{x^*\}$ modulo φ.

Proof. We write

$$\begin{aligned}h^{\mathrm{w}}_{\varphi}(X)&=-\varphi(x^*)p^*\log\,p^*-\int_{\mathcal{X}\setminus\{x^*\}}\varphi(x)f(x)\log\,f(x)\nu(\mathrm{d}x)\\&=-\varphi(x^*)p^*\log\,p^*-\log\,(1-p^*)\int_{\mathcal{X}\setminus\{x^*\}}\varphi(x)f(x)\nu(\mathrm{d}x)\\&\quad-(1-p^*)\int_{\mathcal{X}\setminus\{x^*\}}\varphi(x)\frac{f(x)}{1-p^*}\log\,\frac{f(x)}{1-p^*}\nu(\mathrm{d}x).\end{aligned} \tag{2.17}$$

Theorem 1.4, with $\beta=\dfrac{1}{\nu(\mathcal{X}\setminus\{x^*\})}$, yields that the last line in Eq. (2.17) is bounded above by $\varphi_*\log\,\nu(\mathcal{X}\setminus\{x^*\})$. This leads to (2.16). □

Theorem 2.6. (The weighted generalized Fano inequality; cf. [20], Theorem 1.2.11) *Let $X_i:\Omega\to\mathcal{X}_i$, be a pair of RVs, $i=1,2$. Suppose that X_2 takes exactly m values $1,\dots,m$ (that is, $\mathcal{X}_2=\{1,\dots,m\}$) while X_1 takes values $1,\dots,m$ and possibly other values (that is, $\mathcal{X}_1\supseteq\{1,\dots,m\}$), and set: $\varepsilon_j=\mathbb{P}(X_1\neq j|X_2=j)$. Let a WF $(x_1,x_2)\in\mathcal{X}_1^2\mapsto\varphi(x_1,x_2)$ be given such that for all $j=1,\dots,m$,*

$$\int_{\mathcal{X}_1\setminus\{j\}}\varphi(x_1,j)\left[f_{1|2}(x_1|j)-\frac{\varepsilon_j}{\nu(\mathcal{X}_1\setminus\{j\})}\right]\nu_1(\mathrm{d}x_1)\geq 0. \tag{2.18}$$

Then

$$\begin{aligned}&h^{\mathrm{w}}_{\varphi}(X_1|X_2)\leq\\&\sum_{1\leq j\leq m}\mathbb{P}(X_2=j)\left[-\varphi^*_j(0)(1-\varepsilon_j)\log\,(1-\varepsilon_j)+\varphi^*_j(1)\log\left(\frac{\nu_1(\mathcal{X}_1\setminus\{j\})}{\varepsilon_j}\right)\right].\end{aligned} \tag{2.19}$$

Here RV X_j^* *takes two values, say* 0 *and* 1*, with* $\mathbb{P}(X^* = 0) = 1 - \varepsilon_j = 1 - \mathbb{P}(X^* = 1)$*, and the WF* φ^* *has*

$$\varphi_j^*(0) = \varphi(j, j) \quad \textit{and} \quad \varphi_j^*(1) = \int_{\mathcal{X}\setminus\{j\}} \varphi(x_1, j) f(x, j) \nu_1(\mathrm{d}x_1). \tag{2.20}$$

Proof. By the definition of the conditional WE, the weighted Fano inequality, Theorem 1.4 and with definitions (2.20) at hand, we obtain that

$$h_\varphi^{\mathrm{w}}(X_1|X_2) \le \sum_j \mathbb{P}(X_2 = j) \Bigg[-\varphi(j, j)(1 - \varepsilon_j) \log(1 - \varepsilon_j) + \int_{\mathcal{X}\setminus\{j\}} \varphi(x_1, j) f(x, j) \nu_1(\mathrm{d}x_1) \log \frac{\nu_1(\mathcal{X}_1 \setminus \{j\})}{\varepsilon_j} \Bigg].$$

This yields inequality (2.19). □

3. Maximum WE properties

In this section we establish some extremality properties for the WE; cf. [4], Chap. 12.

Theorem 3.1. *Suppose* $X^* : \Omega \to \mathcal{X}$ *is an RV with a PM/DF* f^* *and* $x \in \mathcal{X} \to \varphi(x)$ *is a given WF.*

- (I) *Then* f^* (*or* X^*) *is the unique maximizer, modulo* φ*, of the WE* $h_\varphi^{\mathrm{w}}(f)$ *under the constraints*

$$\int_{\mathcal{X}} \varphi(x) \big[f(x) - f^*(x)\big] \nu(\mathrm{d}x) \ge 0 \quad \textit{and} \tag{3.1}$$

$$\int_{\mathcal{X}} \varphi(x) \big[f(x) - f^*(x)\big] \log f^*(x) \nu(\mathrm{d}x) \ge 0. \tag{3.2}$$

- (II) *On the other hand, consider a constraint*

$$\int_{\mathcal{X}} \varphi(x) f(x) \beta(x) \mathrm{d}\nu(x) = c \tag{3.3}$$

where $x \in \mathcal{X} \mapsto \beta(x)$ *is a given function and* c *a given constant neither of which is assumed non-negative. Suppose that* $f^*(x) = \dfrac{1}{Z} \exp[-b\beta(x)]$ *is a* (*Gibbsian-type*) *PM/DF such that*

$$\int_{\mathcal{X}} \varphi(x) f^*(x) \mathrm{d}\nu(x) = 1 \quad \textit{and} \quad \int_{\mathcal{X}} \varphi(x) f^*(x) \beta(x) \mathrm{d}\nu(x) = c.$$

Here b *is a constant* (*an analog of inverse temperature*) *and* $Z = \int_{\mathcal{X}} \exp[-b\beta(x)] \mathrm{d}\nu(x) \in (0, \infty)$ *is the normalizing denominator* (*an analog of a partition function*)*. Introduce the second constraint:*

$$(\log Z)\int_{\mathcal{X}}\varphi(x)[f^*(x)-f(x)]\mathrm{d}\nu(x)\geq 0. \tag{3.4}$$

Then, under (3.3) *and* (3.4)*, the WE* $h^{\rm w}_\varphi(f)$ *is maximized at* $f=f^*$*. As above, it is a unique maximizer, modulo* φ*.*

Proof. (I) Using definition (1.2) and Theorem 1.3, we obtain

$$0\geq -D^{\rm w}_\varphi(f\|f^*)=h^{\rm w}_\varphi(f)+\int\limits_{\mathcal{X}}\varphi(x)f(x)\log f^*(x)\nu(\mathrm{d}x). \tag{3.5}$$

Under our constraint (3.1) it yields

$$h^{\rm w}_\varphi(f)\leq -\int\limits_{\mathcal{X}}\varphi(x)f^*(x)\log f^*(x)\nu(\mathrm{d}x)=h^{\rm w}_\varphi(f^*). \tag{3.6}$$

The uniqueness of the maximizer follows from the uniqueness case for equality in the weighted Gibbs inequality.

(II) Again use (3.5):

$$\begin{aligned}h^{\rm w}_\varphi(f)&\leq -\int_{\mathcal{X}}\varphi(x)f(x)\Big[-\log\, Z-b\beta(x)\Big]\mathrm{d}\nu(x)\\&=(\log\, Z)\int_{\mathcal{X}}\varphi(x)f(x)\mathrm{d}\nu(x)+b\int_{\mathcal{X}}\varphi(x)f(x)\beta(x)\mathrm{d}\nu(x)\\&\leq(\log\, Z)\int_{\mathcal{X}}\varphi(x)f^*(x)\mathrm{d}\nu(x)+b\int_{\mathcal{X}}\varphi(x)f^*(x)\beta(x)\mathrm{d}\nu(x)=h^{\rm w}_\varphi(f^*).\end{aligned}$$

□

Note that when $Z\geq 1$, the factor $\log\, Z$ can be omitted from (3.4); otherwise $\log\, Z$ can be replaced by -1.

Example 3.2. Consider a random vector $\mathbf{X}=\mathbf{X}_1^d:\Omega\to\mathbb{R}^d$ with PDF f (relative to the d-dimensional Lebesgue measure), mean vector $\mathbf{0}$ and covariance matrix $\mathbf{C}=(C_{ij})$ with $C_{ij}=\mathbb{E}\big[X_iX_j\big]$, $1\leq i,j\leq d$. Let $f^{\rm No}_{\mathbf{C}}$ be the normal PDF with the same $\boldsymbol{\mu}$ and $\mathbf{C}$. Let $\mathbf{x}=\mathbf{x}_1^d\in\mathbb{R}^d\mapsto\varphi(\mathbf{x})$ be a given WF which is positive on an open domain in $\mathbb{R}^d$. Introduce $d\times d$ matrices $\mathbf{\Phi}=(\Phi_{ij})$, $\mathbf{\Phi}^{\rm No}_{\mathbf{C}}=(\Phi^{\rm No}_{ij})$ and $\mathbf{x}^T\mathbf{x}$, where $(\mathbf{x}^T\mathbf{x})_{ij}=x_ix_j$ and

$$\begin{aligned}\Phi&=\int_{\mathbb{R}^d}\varphi(\mathbf{x})f(\mathbf{x})\mathbf{x}^T\mathbf{x}\,\mathrm{d}\mathbf{x}\\ \Phi^{\rm No}_{\mathbf{C}}&=\int_{\mathbb{R}^d}\varphi(\mathbf{x})f^{\rm No}_{\mathbf{C}}(\mathbf{x})\mathbf{x}^T\mathbf{x}\,\mathrm{d}\mathbf{x}.\end{aligned}$$

Suppose that

$$\begin{aligned}&\int_{\mathbb{R}^d}\varphi(\mathbf{x})\Big[f(\mathbf{x})-f^{\rm No}_{\mathbf{C}}(\mathbf{x})\Big]\mathrm{d}\mathbf{x}\geq 0\quad\text{and}\\&\log\big[(2\pi)^d(\det\mathbf{C})\big]\int_{\mathbb{R}^d}\varphi(\mathbf{x})\Big[f(\mathbf{x})-f^{\rm No}_{\mathbf{C}}(\mathbf{x})\Big]\mathrm{d}\mathbf{x}+\mathrm{tr}\Big[\mathbf{C}^{-1}\Big(\mathbf{\Phi}-\mathbf{\Phi}^{\rm No}_{\mathbf{C}}\Big)\Big]\leq 0.\end{aligned} \tag{3.7}$$

Then

$$h^{\rm w}_\varphi(f) \le h^{\rm w}_\varphi(f^{\rm No}_{\mathbf{C}}) = \frac{1}{2}\log\left[(2\pi)^d(\det\mathbf{C})\right]\int\limits_{\mathbb{R}^d}\varphi(\mathbf{x})f^{\rm No}_{\mathbf{C}}(\mathbf{x}){\rm d}\mathbf{x} + \frac{\log e}{2}{\rm tr}\,\mathbf{C}^{-1}\mathbf{\Phi}^{\rm No}_{\mathbf{C}}, \tag{3.8}$$

with equality iff $f = f^{\rm No}_{\mathbf{C}}$ modulo φ.

Proof. Using the same idea as before, write

$$0 \ge -D^{\rm w}_\varphi(f\|f^{\rm No}_{\mathbf{C}}) = h^{\rm w}_\varphi(f) - \frac{1}{2}\log\Big[(2\pi)^d(\det\mathbf{C})\Big]\int\limits_{\mathbb{R}^d}\varphi(\mathbf{x})f(\mathbf{x}){\rm d}\mathbf{x} - \frac{\log e}{2}{\rm tr}\,\mathbf{C}^{-1}\mathbf{\Phi}. \tag{3.9}$$

Equivalently,

$$h^{\rm w}_\varphi(f) \le \frac{1}{2}\log\left[(2\pi)^d(\det\mathbf{C})\right]\int\limits_{\mathbb{R}^d}\varphi(\mathbf{x})f(\mathbf{x}){\rm d}\mathbf{x} + \frac{\log e}{2}{\rm tr}\,\mathbf{C}^{-1}\mathbf{\Phi},$$

which leads directly to the result. □

To further illustrate the above methodology, we provide some more examples, omitting the proofs.

Example 3.3. Let $f^{\rm Exp}$ denote an exponential PDF on $\mathbb{R}_+ = (0,\infty)$ (relative to the Lebesgue measure ${\rm d}x$) with mean λ^{-1}. Suppose a PDF f on $\mathbb{R}_+$ satisfies the constraints

$$\begin{aligned}&\int\limits_{\mathbb{R}_+}\varphi(x)\big[f(x) - f^{\rm Exp}(x)\big]{\rm d}x \ge 0 \quad \text{and}\\ &\big(\log\lambda\big)\int\limits_{\mathbb{R}_+}\varphi(x)\big[f(x) - f^{\rm Exp}(x)\big]{\rm d}x - \lambda\int\limits_{\mathbb{R}_+}x\,\varphi(x)\big[f(x) - f^{\rm Exp}(x)\big]{\rm d}x \ge 0,\end{aligned} \tag{3.10}$$

where $x \in \mathbb{R}_+ \mapsto \varphi(x)$ is a given WF positive on an open interval. Then

$$h^{\rm w}_\varphi(f) \le h^{\rm w}_\varphi(f^{\rm Exp}) = -\big(\lambda\log\lambda\big)\int_{\mathbb{R}_+}\varphi(x)e^{-\lambda x}{\rm d}x + \lambda^2\int_{\mathbb{R}_+}x\varphi(x)e^{-\lambda x}{\rm d}x,$$

and $f^{\rm Exp}$ is a unique maximizer modulo φ.

Example 3.4. Take $\mathcal{X} = \mathbb{Z}_+ = \{0,1,\ldots\}$ and let ν be the counting measure: $\nu(i) = 1\ \forall\, i \in \mathbb{Z}_+$. Then, for a RV X with PMF $f(i)$ we have $f(i) = \mathbb{P}(X = i)$. Fix a WF $i \in \mathbb{Z}_+ \mapsto \varphi(i)$.

(a) Let $f^{\rm Ge}$ be a geometric PMF: $f^{\rm Ge}(x)=(1-p)^xp$, $x\in\mathbb{Z}_+$. Then for any PMF $f(x)$, $i\in\mathbb{Z}_+$, satisfying the constraints

$$\begin{array}{l}\displaystyle\sum_{i\in\mathbb{Z}_+}\varphi(i)\big[f(i)-f^{\rm Ge}(i)\big]\geq 0\quad\text{and}\\ \displaystyle\log\,p\sum_{i\in\mathbb{Z}_+}\varphi(i)\big[f(i)-f^{\rm Ge}(i)\big]+\log(1-p)\sum_{i\in\mathbb{Z}_+}i\varphi(i)\big[f(i)-f^{\rm Ge}(i)\big]\geq 0\end{array}\tag{3.11}$$

we have $h^{\rm w}_\varphi(f)\leq h^{\rm w}_\varphi(f^{\rm Ge})$, with equality iff $f=f^{\rm Ge}$ modulo φ.

(b) Let $f^{\rm Po}$ be a Poissonian PMF: $f^{\rm Po}(k)=\dfrac{e^{-\lambda}\lambda^k}{k!}$, $k\in\mathbb{Z}_+$. Then for any PMF $f(k)$, $k\in\mathbb{Z}_+$, satisfying the constraints

$$\begin{array}{l}\displaystyle\sum_{k\in\mathbb{Z}_+}\varphi(k)\big[f(k)-f^{\rm Po}(k)\big]\geq 0\quad\text{and}\\ \displaystyle\log\,\lambda\sum_{k\in\mathbb{Z}_+}k\varphi(k)\big[f(k)-f^{\rm Po}(k)\big]\\ \displaystyle\qquad-\lambda\sum_{k\in\mathbb{Z}_+}\varphi(k)\big[f(k)-f^{\rm Po}(k)\big]-\sum_{k\in\mathbb{Z}_+}(\log\,k!)\varphi(k)\big[f(k)-f^{\rm Po}(k)\big]\geq 0\end{array}\tag{3.12}$$

we have $h^{\rm w}_\varphi(f)\leq h^{\rm w}_\varphi(f^{\rm Po})$, with equality iff $f=f^{\rm Po}$ modulo φ.

Theorem 3.5 below offers an extension of the Ky Fan inequality so that log det $\mathbf{C}$ is a concave function of a positive definite $d\times d$ matrix $\mathbf{C}$. Cf. [16–18, 21]. We follow the method proposed by Cover-Dembo-Thomas. As before, $f^{\rm No}_{\mathbf{C}}$ denotes the normal PDF with zero mean and covariance matrix $\mathbf{C}$.

Theorem 3.5. (The weighted Ky Fan inequality; cf. [3], Theorem 17.9.1, [4], Theorem 1, [5], Theorem 8, [20], Worked Example 1.5.9) *Assume that* $\mathbf{x}_1^d\in\mathbb{R}^d\mapsto\varphi(\mathbf{x}_1^d)\geq 0$ *is a given WF positive on an open domain. Suppose that, for* $\lambda_1,\lambda_2\in[0,1]$ *with* $\lambda_1+\lambda_2=1$ *and positive-definite* $\mathbf{C}_1$, $\mathbf{C}_2$, *with* $\mathbf{C}=\lambda_1\mathbf{C}_1+\lambda_2\mathbf{C}_2$,

$$\int_{\mathbb{R}^d}\varphi(\mathbf{x})\Big[\lambda_1 f^{\rm No}_{\mathbf{C}_1}(\mathbf{x})+\lambda_2 f^{\rm No}_{\mathbf{C}_2}(\mathbf{x})-f^{\rm No}_{\mathbf{C}}(\mathbf{x})\Big]{\rm d}\mathbf{x}\geq 0,\quad\textit{and}\tag{3.13}$$

$$\begin{array}{l}\displaystyle\log\big[(2\pi)^d(\det\mathbf{C})\big]\int_{\mathbb{R}^d}\varphi(\mathbf{x})\Big[\lambda_1 f^{\rm No}_{\mathbf{C}_1}(\mathbf{x})+\lambda_2 f^{\rm No}_{\mathbf{C}_2}(\mathbf{x})-f^{\rm No}_{\mathbf{C}}(\mathbf{x})\Big]{\rm d}\mathbf{x}\\ \displaystyle\qquad+\frac{\log e}{2}{\rm tr}\Big[\mathbf{C}^{-1}\mathbf{\Psi}\Big]\leq 0,\end{array}\tag{3.14}$$

$$\textit{where}\quad\mathbf{\Psi}=\int_{\mathbb{R}^d}\varphi(\mathbf{x})\Big[\lambda_1 f^{\rm No}_{\mathbf{C}_1}(\mathbf{x})+\lambda_2 f^{\rm No}_{\mathbf{C}_2}(\mathbf{x})-f^{\rm No}_{\mathbf{C}}(\mathbf{x})\Big](\mathbf{x}-\boldsymbol{\mu})^{\rm T}(\mathbf{x}-\boldsymbol{\mu}){\rm d}\mathbf{x}.\tag{3.15}$$

Then, with $\sigma_\varphi(\mathbf{C})=h^{\rm w}_\varphi(f^{\rm No}_{\mathbf{C}})$, $\sigma_\varphi(\mathbf{C}_1)=h^{\rm w}_\varphi(f^{\rm No}_{\mathbf{C}_1})$ *and* $\sigma_\varphi(\mathbf{C}_2)=h^{\rm w}_\varphi(f^{\rm No}_{\mathbf{C}_2})$

$$\sigma_\varphi(\mathbf{C}) - \lambda_1\sigma_\varphi(\mathbf{C}_1) - \lambda_2\sigma_\varphi(\mathbf{C}_2) \geq 0; \tag{3.16}$$

with equality iff $\lambda_1\lambda_2 = 0$ *or* $\mathbf{C}_1 = \mathbf{C}_2$.

Proof. Take values $\lambda_1, \lambda_2 \in [0, 1]$, such that $\lambda_1 + \lambda_2 = 1$. Let $\mathbf{C}_1$ and $\mathbf{C}_2$ be two positive definite $d \times d$ matrices. Let $\mathbf{X}_1$ and $\mathbf{X}_2$ be two multivariate normal vectors, with PDFs $f_k \sim \mathrm{N}(0, \mathbf{C}_k)$, $k = 1, 2$. Set $\mathbf{Z} = \mathbf{X}_\Theta$, where the RV Θ, takes two values, $\theta = 1$ and $\theta = 2$ with probability λ_1 and λ_2 respectively, and is independent of $\mathbf{X}_1$ and $\mathbf{X}_2$. Then vector $\mathbf{Z}$ has covariance $\mathbf{C} = \lambda_1\mathbf{C}_1 + \lambda_2\mathbf{C}_2$. Also set:

$$\alpha(\mathbf{C}) = \int\limits_{\mathbb{R}^d} \varphi(\mathbf{x}) f^{\mathrm{No}}_{\mathbf{C}}(\mathbf{x})\mathrm{d}\mathbf{x}. \tag{3.17}$$

Let $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d \mapsto \varphi(\mathbf{x})$ be a given WF and set $\widetilde{\varphi}(\mathbf{x}, \theta) = \varphi(\mathbf{x})$. Following the same arguments as in the proof of Theorem 2.1, $h^{\mathrm{w}}_{\widetilde{\varphi}}(\mathbf{Z}|\Theta) \leq h^{\mathrm{w}}_\varphi(\mathbf{Z})$. It is plain that

$$\begin{aligned} h^{\mathrm{w}}_{\widetilde{\varphi}}(\mathbf{Z}|\Theta) &= \lambda_1 h^{\mathrm{w}}_\varphi(X_1) + \lambda_2 h^{\mathrm{w}}_\varphi(X_2) \\ &= \sum_{k=1,2} \lambda_k \left\{ \frac{1}{2} \log\left[(2\pi)^d(\det \mathbf{C}_k)\right] \int\limits_{\mathbb{R}^d} \varphi(\mathbf{x}) f^{\mathrm{No}}_{\mathbf{C}_k}(\mathbf{x})\mathrm{d}\mathbf{x} \right. \\ &\qquad \left. + \tfrac{\log\, e}{2} \mathrm{tr}\, \mathbf{C}_k^{-1} \mathbf{\Phi}^{(k)} \right\} \end{aligned}$$

where

$$\mathbf{\Phi}^{(k)} = \int\limits_{\mathbb{R}^d} \mathbf{x}^{\mathrm{T}}\mathbf{x}\, \varphi(\mathbf{x}) f^{\mathrm{No}}_{\mathbf{C}_k}(\mathbf{x})\mathrm{d}\mathbf{x}, \quad k = 1, 2,$$

and $(\mathbf{x}^{\mathrm{T}}\mathbf{x})_{ij} = x_i x_j$. According to Example 3.2, we have

$$h^{\mathrm{w}}_\varphi(\mathbf{Z}) \leq \frac{1}{2}\Big\{ \log\left[(2\pi)^d(\det \mathbf{C})\right] \Big\}\alpha(\mathbf{C}) + \frac{\log\, e}{2}\mathrm{tr}\, \mathbf{C}^{-1}\mathbf{\Phi}, \tag{3.18}$$

where

$$\mathbf{\Phi} = \int\limits_{\mathbb{R}^d} \mathbf{x}^{\mathrm{T}}\mathbf{x}\, \varphi(\mathbf{x}) f^{\mathrm{No}}_{\mathbf{C}}(\mathbf{x})\mathrm{d}\mathbf{x}. \tag{3.19}$$

The inequality (3.16) then follows. The cases of equality are covered by Theorem 2.1. $\square$

The following lemma is an immediate extension of Lemma 1.6.

Lemma 3.6. *Let* $\mathbf{X}_1^n = (X_1, \ldots, X_n)$ *be a random vector, with components* $X_i : \Omega \to \mathcal{X}_i$, $1 \leq i \leq n$, *and the joint PM/DF* f. *Extending the notation used in Sect* 1, *set:*

$$\mathbf{x}_1^n = (x_1, \ldots, x_n) \in \mathcal{X}_1^n := \mathop{\times}_{1\leq i\leq n} \mathcal{X}_i \quad \text{and} \quad \nu_1^n(\mathrm{d}\mathbf{x}_1^n) = \prod_{1\leq i\leq n} \mathrm{d}\nu_i(\mathrm{d}x_i),$$

and more generally,

$$\mathbf{x}_k^l = (x_k, \ldots, x_l) \in \mathcal{X}_k^l := \underset{k\le i\le l}{\times} \mathcal{X}_i \text{ and } \nu_k^l(\mathrm{d}\mathbf{x}_k^l) = \prod_{k\le i\le l} \mathrm{d}\nu_i(\mathrm{d}x_i), \ \ 1 \le k \le l.$$

Next, introduce

$$f_i(x_i) = \int_{\mathcal{X}_1^{i-1}\times\mathcal{X}_{i+1}^n} f(\mathbf{x}_1^{i-1}, x_i, \mathbf{x}_{i+1}^n)\nu_1^{i-1}(\mathrm{d}\mathbf{x}_1^{i-1})\nu_{i+1}^n(\mathrm{d}\mathbf{x}_{i+1}^n)$$
$$\textit{(the marginal PM/DF for RV } X_i\textit{)},$$

and

$$f_{\,|i}\,(\mathbf{x}_1^n|x_i) = \frac{f(\mathbf{x}_1^n)}{f_i(x_i)} \qquad \textit{(the conditional PM/DF given that } X_i = x_i\textit{)}.$$

Given a WF $\mathbf{x}_1^n \in \mathcal{X}_1^n \mapsto \varphi(\mathbf{x}_1^n)$*, suppose that*

$$\int_{\mathcal{X}_1^n} \varphi(\mathbf{x}_1^n)\left[f(\mathbf{x}_1^n) - \prod_{i=1}^n f_i(x_i)\right]\nu_1^n(\mathrm{d}\mathbf{x}_1^n) \ge 0. \tag{3.20}$$

Then

$$h_\varphi^{\mathrm{w}}(\mathbf{X}_1^n) \le \sum_{i=1}^n h_{\psi_i}^{\mathrm{w}}(\mathbf{X}_i), \tag{3.21}$$

where

$$\psi_i(x_i) = \int_{\mathcal{X}_1^{i-1}\times\mathcal{X}_{i+1}^n} \varphi(\mathbf{x}_1^n) f_{\,|i}\,(\mathbf{x}_1^n|x_i)\nu_1^{i-1}(\mathrm{d}\mathbf{x}_1^{i-1})\nu_{i+1}^n(\mathrm{d}\mathbf{x}_{i+1}^n).$$

Here, equality in (3.21) *holds iff, modulo* φ*, components* $X_1, \ldots, X_n$ *are independent.*

Theorem 3.7. (The weighted Hadamard inequality; cf. [3], Theorem 17.9.2, [4], Theorem 3, [5], Theorem 26, [20], Worked Example 1.5.10). *Let* $\mathbf{C} = (C_{ij})$ *be a positive definite* $d \times d$ *matrix and* $f_{\mathbf{C}}^{\mathrm{No}}$ *the normal PDF with zero mean and covariance matrix* $\mathbf{C}$*. Given a WF function* $\mathbf{x}_1^d = (x_1, \ldots, x_d) \in \mathbb{R}^d \mapsto \varphi(\mathbf{x}_1^d)$*, positive on an open domain in* $\mathbb{R}^d$*, introduce the quantity* $\alpha = \alpha(\mathbf{C})$ *by* (3.17) *and matrix* $\mathbf{\Phi} = (\Phi_{ij})$ *by* (3.19)*. Let* f_i^{No} *stand for the* $\mathrm{N}(0, C_{ii})$*-PDF (the marginal PDF of the i-th component). Then under the condition*

$$\int_{\mathbb{R}^d} \varphi(\mathbf{x}_1^d)\Big[f_{\mathbf{C}}^{\mathrm{No}}(\mathbf{x}_1^d) - \prod_{i=1}^d f_i^{\mathrm{No}}(x_i)\Big]\mathrm{d}\mathbf{x}_1^d \ge 0, \tag{3.22}$$

we have:

$$\alpha\log\prod_i(2\pi C_{ii}) + (\log\, e)\sum_i C_{ii}^{-1}\Phi_{ii} - \alpha\log\,\Big[(2\pi)^d(\det\mathbf{C})\Big] - (\log\, e)\mathrm{tr}\,\mathbf{C}^{-1}\mathbf{\Phi} \ge 0, \tag{3.23}$$

with equality iff $\mathbf{C}$ *is diagonal.*

Proof. If $X_1, \ldots, X_d \sim \mathrm{N}(0, \mathbf{C})$, then in Lemma 3.6, by following (3.21) we can write

$$\begin{aligned}\frac{1}{2}\log\left[(2\pi)^d(\det\mathbf{C})\right]\int\limits_{\mathbb{R}^d}\varphi(\mathbf{x}_1^d)f(\mathbf{x}_1^d)\mathrm{d}\mathbf{x}_1^d+\frac{\log e}{2}\mathrm{tr}\,\mathbf{C}^{-1}\mathbf{\Phi}\\ \leq\frac{1}{2}\sum_{i=1}^d\left[\log\left(2\pi C_{ii}\right)\int\limits_{\mathbb{R}}\psi_i(x)f_i^{\mathrm{No}}(x)\mathrm{d}x+(\log e)\,C_{ii}^{-1}\Psi_{ii}\right].\end{aligned}\tag{3.24}$$

Here

$$\psi_i(x_i)=\int_{\mathbb{R}^{d-1}}\varphi(\mathbf{x}_1^d)f_{|i}^{\mathrm{No}}(\mathbf{x}_1^d|x_i)\prod_{j:j\neq i}\mathrm{d}x_j,\quad\Psi_{ii}=\int\limits_{\mathbb{R}^d}x_i^2\psi_i(x_i)f_i^{\mathrm{No}}(x_i)\mathrm{d}x_i=\Phi_{ii}$$

and

$$f_{|i}^{\mathrm{No}}(\mathbf{x}_1^d|x_i)=\frac{f_{\mathbf{C}}^{\mathrm{No}}(\mathbf{x}_1^d)}{f_i^{\mathrm{No}}(x_i)}\ \text{ (the conditional PDF).}$$

With

$$\alpha=\int\limits_{\mathbb{R}}\psi_i(x_i)f_i^{\mathrm{No}}(x_i)\mathrm{d}x_i=\int\limits_{\mathbb{R}^d}\varphi(\mathbf{x}_1^d)f_{\mathbf{C}}^{\mathrm{No}}(\mathbf{x}_1^d)\mathrm{d}\mathbf{x}_1^d,$$

the bound (3.23) follows. □

Remark 3.8. As above, maximizing the left-hand side in (3.23) would give a bound between $\det\mathbf{C}$ and the product $\prod_{i=1}^d C_{ii}$.

4. Weighted Fisher information and related inequalities

In this section we introduce a weighted version of the Fisher information matrix and establish some straightforward facts. The bulk of these properties is derived by following Ref. [32].

Definition 4.1. Let $\mathbf{X}=(X_1,\ldots,X_n)$ be a random $1\times n$ vector with probability density function (PDF) $f_{\underline{\theta}}(\mathbf{x})=f_{\mathbf{X}}(\mathbf{x};\underline{\theta})$, $\mathbf{x}=(x_1,\ldots,x_n)\in\mathbb{R}^n$, where $\underline{\theta}=(\theta_1,\ldots,\theta_m)\in\mathbb{R}^m$ is a parameter vector. Suppose that dependence $\underline{\theta}\mapsto f_{\underline{\theta}}$ is C^1. The $m\times m$ weighted Fisher information matrix $\mathrm{J}_\varphi^{\mathrm{w}}(\mathbf{X};\underline{\theta})$, with a given WF $\mathbf{x}\in\mathbb{R}^n\mapsto\varphi(\mathbf{x})\geq 0$, is defined by

$$\begin{aligned}\mathrm{J}_\varphi^{\mathrm{w}}(\mathbf{X};\underline{\theta})&=\mathbb{E}\Big[\varphi(\mathbf{X})\mathbf{S}(\mathbf{X},\underline{\theta})^{\mathrm{T}}\mathbf{S}(\mathbf{X},\underline{\theta})\Big]\\&=\int\frac{\varphi(\mathbf{x})}{f_{\underline{\theta}}(\mathbf{x})}\left(\frac{\partial f_{\underline{\theta}}(\mathbf{x})}{\partial\underline{\theta}}\right)^{\mathrm{T}}\frac{\partial f_{\underline{\theta}}(\mathbf{x})}{\partial\underline{\theta}}\mathbf{1}\left(f_{\underline{\theta}}(\mathbf{x})>0\right)\mathrm{d}\mathbf{x},\end{aligned}\tag{4.1}$$

assuming the integrals are absolutely convergent. Here and below, $\dfrac{\partial}{\partial \underline{\theta}}$ stands for the $1 \times m$ gradient in $\underline{\theta}$ and $\mathbf{S}(\mathbf{X}, \underline{\theta}) = \mathbf{1}(f_{\underline{\theta}}(\mathbf{x}) > 0) \dfrac{\partial}{\partial \underline{\theta}} \log f_{\underline{\theta}}(\mathbf{x})$ denotes the *score vector.*

When $\varphi(\mathbf{x}) \equiv 1$, $\mathtt{J}^{\rm w}_{\varphi}(\mathbf{X}; \underline{\theta}) = \mathtt{J}(\mathbf{X}; \underline{\theta})$, the standard Fisher information matrix, cf. [4,5,20].

Definition 4.2. Let $(\mathbf{X}, \mathbf{Y})$ be a pair of RVs with a joint PDF $f_{\underline{\theta}}(\mathbf{x}, \mathbf{y}) = f_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}; \underline{\theta})$ and conditional PDF $f_{\underline{\theta}}(\mathbf{y}|\mathbf{x}) = f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}; \underline{\theta}) := \dfrac{f_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}; \underline{\theta})}{f_{\mathbf{X}}(\mathbf{x}; \underline{\theta})}$. Given a joint WF $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^n \mapsto \varphi(\mathbf{x}, \mathbf{y}) \geq 0$, we set:

$$\begin{aligned}\mathtt{J}^{\rm w}_{\varphi}(\mathbf{X}, \mathbf{Y}; \underline{\theta}) &= \mathbb{E}\left[\varphi(\mathbf{X}, \mathbf{Y}) \left(\frac{\partial \log\, f_{\underline{\theta}}(\mathbf{X}, \mathbf{Y})}{\partial \underline{\theta}}\right)^{\rm T} \frac{\partial \log\, f_{\underline{\theta}}(\mathbf{X}, \mathbf{Y})}{\partial \underline{\theta}} \mathbf{1}\left(f_{\underline{\theta}}(\mathbf{X}, \mathbf{Y}) > 0\right)\right] \\ &= \int \frac{\varphi(\mathbf{x}, \mathbf{y})}{f_{\underline{\theta}}(\mathbf{x}, \mathbf{y})} \left(\frac{\partial f_{\underline{\theta}}(\mathbf{x}, \mathbf{y})}{\partial \underline{\theta}}\right)^{\rm T} \frac{\partial f_{\underline{\theta}}(\mathbf{x}, \mathbf{y})}{\partial \underline{\theta}} \mathbf{1}\left(f_{\underline{\theta}}(\mathbf{x}, \mathbf{y}) > 0\right) {\rm d}\mathbf{x}{\rm d}\mathbf{y} \end{aligned} \tag{4.2}$$

and

$$\begin{aligned}\mathtt{J}^{\rm w}_{\varphi}(\mathbf{Y}|\mathbf{X}; \underline{\theta}) &= \mathbb{E}\left[\varphi(\mathbf{X}, \mathbf{Y}) \left(\frac{\partial \log f_{\underline{\theta}}(\mathbf{Y}|\mathbf{X})}{\partial \underline{\theta}}\right)^{\rm T} \frac{\partial \log f_{\underline{\theta}}(\mathbf{Y}|\mathbf{X})}{\partial \underline{\theta}} \mathbf{1}\left(f_{\underline{\theta}}(\mathbf{Y}|\mathbf{X}) > 0\right)\right] \\ &= \int \frac{\varphi(\mathbf{x}, \mathbf{y}) f_{\underline{\theta}}(\mathbf{x}, \mathbf{y})}{f_{\underline{\theta}}(\mathbf{y}|\mathbf{x})^2} \left(\frac{\partial f_{\underline{\theta}}(\mathbf{y}|\mathbf{x})}{\partial \underline{\theta}}\right)^{\rm T} \frac{\partial f_{\underline{\theta}}(\mathbf{y}|\mathbf{x})}{\partial \underline{\theta}} \mathbf{1}\left(f_{\underline{\theta}}(\mathbf{y}|\mathbf{x}) > 0\right) {\rm d}\mathbf{x}{\rm d}\mathbf{y}. \end{aligned} \tag{4.3}$$

Next, consider an $m \times m$ matrix $\mathtt{S}_{\underline{\theta}} = \mathtt{S}_{\underline{\theta}}(f_{\mathbf{X},\mathbf{Y}})$ and a $1 \times m$ vector $\mathbf{B}_{\underline{\theta}} = \mathbf{B}_{\underline{\theta}}(\mathbf{x}, f_{\mathbf{Y}|\mathbf{X}})$:

$$\mathbf{B}_{\underline{\theta}} = \mathbb{E}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}\left[\varphi(\mathbf{x}, \mathbf{Y}) \frac{\partial \log f_{\underline{\theta}}(\mathbf{Y}|\mathbf{x})}{\partial \underline{\theta}}\right] = \int \frac{\varphi(\mathbf{x}, \mathbf{y})}{f_{\underline{\theta}}(\mathbf{y}|\mathbf{x})} \frac{\partial f_{\underline{\theta}}(\mathbf{y}|\mathbf{x})}{\partial \underline{\theta}} \mathbf{1}\left(f_{\underline{\theta}}(\mathbf{y}|\mathbf{x}) > 0\right) {\rm d}\mathbf{y}, \tag{4.4}$$

$$\mathtt{S}_{\underline{\theta}} = \mathbb{E}\left\{\left[\left(\frac{\partial \log f_{\underline{\theta}}(\mathbf{X})}{\partial \underline{\theta}}\right)^{\rm T} \mathbf{B}_{\underline{\theta}}(\mathbf{X}) + \mathbf{B}_{\underline{\theta}}(\mathbf{X})^{\rm T} \frac{\partial \log f_{\underline{\theta}}(\mathbf{X})}{\partial \underline{\theta}}\right] \mathbf{1}\left(f_{\underline{\theta}}(\mathbf{X}) > 0\right)\right\}. \tag{4.5}$$

When $\varphi(\mathbf{x}, \mathbf{y})$ depends only on $\mathbf{x}$ and under standard regularity assumptions, vector $\mathbf{B}_{\underline{\theta}}$ vanishes (and so does matrix $\mathtt{S}_{\underline{\theta}}$):

$$\mathbf{B}_{\underline{\theta}} = \varphi(\mathbf{x}) \int \frac{\partial f_{\underline{\theta}}(\mathbf{y}|\mathbf{x})}{\partial \underline{\theta}} {\rm d}\mathbf{y} = \frac{\partial}{\partial \underline{\theta}} \int f_{\underline{\theta}}(\mathbf{y}|\mathbf{x}) {\rm d}\mathbf{y} = 0.$$

For the sake of brevity, in formulas that follow we routinely omit indicators of positivity of PDFs involved: their presence can be easily derived from the local context.

Lemma 4.3. (The chain rule: cf. [32], Lemma 1) *Given a pair* $(\mathbf{X},\mathbf{Y})$ *of random vectors and a joint WF* $(\mathbf{x},\mathbf{y})\mapsto\varphi(\mathbf{x},\mathbf{y})$, *set:*

$$\psi(\mathbf{x})=\psi_{\mathbf{X}}(\mathbf{x})=\int\varphi(\mathbf{x},\mathbf{y})f_{\underline{\theta}}(\mathbf{y}|\mathbf{x})\,\mathrm{d}\mathbf{y}=\mathbb{E}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}\varphi(\mathbf{x},\mathbf{Y}). \tag{4.6}$$

Then

$$\mathtt{J}^{\mathrm{w}}_{\varphi}(\mathbf{X},\mathbf{Y};\underline{\theta})=\mathtt{J}^{\mathrm{w}}_{\psi}(\mathbf{X};\underline{\theta})+\mathtt{J}^{\mathrm{w}}_{\varphi}(\mathbf{Y}|\mathbf{X};\underline{\theta})+\mathtt{S}_{\underline{\theta}}. \tag{4.7}$$

Proof. For simplicity, assume that $\underline{\theta}$ is scalar: $\underline{\theta}=\theta$; a generalization to a vector case is straightforward. Therefore, we have

$$\mathtt{J}^{\mathrm{w}}_{\varphi}(\mathbf{X},\mathbf{Y};\theta)=\mathbb{E}\left[\varphi(\mathbf{X},\mathbf{Y})\left(\frac{\partial\log f_{\theta}(\mathbf{X},\mathbf{Y})}{\partial\theta}\right)^2\right]. \tag{4.8}$$

Furthermore, we know

$$\log f_{\theta}(\mathbf{x},\mathbf{y})=\log f_{\theta}(\mathbf{x})+\log f_{\theta}(\mathbf{y}|\mathbf{x}).$$

Using (4.8) yields:

$$\begin{aligned}\mathtt{J}^{\mathrm{w}}_{\varphi}(\mathbf{X},\mathbf{Y};\theta)=\mathbb{E}&\left[\varphi(\mathbf{X},\mathbf{Y})\left(\frac{\partial\log f_{\theta}(\mathbf{X})}{\partial\theta}\right)^2\right]\\&+\mathbb{E}\left[\varphi(\mathbf{X},\mathbf{Y})\left(\frac{\partial\log f_{\theta}(\mathbf{Y}|\mathbf{X})}{\partial\theta}\right)^2\right]\\&+2\,\mathbb{E}\left[\varphi(\mathbf{X},\mathbf{Y})\left(\frac{\partial\log f_{\theta}(\mathbf{X})}{\partial\theta}\right)\left(\frac{\partial\log f_{\theta}(\mathbf{Y}|\mathbf{X})}{\partial\theta}\right)\right].\end{aligned} \tag{4.9}$$

We can also write

$$\begin{aligned}&\mathbb{E}\left[\varphi(\mathbf{X},\mathbf{Y})\left(\frac{\partial\log f_{\theta}(\mathbf{X})}{\partial\theta}\right)\left(\frac{\partial\log f_{\theta}(\mathbf{Y}|\mathbf{X})}{\partial\theta}\right)\right]\\&\quad=\mathbb{E}\left\{\frac{\partial\log f_{\theta}(X)}{\partial\theta}\mathbb{E}\left[\varphi(\mathbf{X},\mathbf{Y})\left(\frac{\partial\log f_{\theta}(\mathbf{Y}|\mathbf{X})}{\partial\theta}\right)\Big|\mathbf{X}\right]\right\}.\end{aligned} \tag{4.10}$$

This cancels the last term in (4.7) when applying inner expectation in the RHS of (4.7). □

Throughout the paper, an inequality $\mathtt{A}\leq\mathtt{B}$ between matrices $\mathtt{A}$ and $\mathtt{B}$ means that $\mathtt{B}-\mathtt{A}$ is a positive-definite matrix.

Lemma 4.4. (Data-refinement inequality: cf. [32], Lemma 2) *For a given joint WF* $(\mathbf{x},\mathbf{y})\mapsto\varphi(\mathbf{x},\mathbf{y})$,

$$\mathtt{J}^{\mathrm{w}}_{\varphi}(\mathbf{X},\mathbf{Y};\underline{\theta})\geq\mathtt{J}^{\mathrm{w}}_{\psi}(\mathbf{X};\underline{\theta})+\mathtt{S}_{\underline{\theta}}, \tag{4.11}$$

with equality if $\mathbf{X}$ *is a sufficient statistic for* $\underline{\theta}$. *Here WF* $\psi=\psi_{\mathbf{X}}$ *is defined as in* (4.6).

Proof. Bound (4.11) follows from Lemma 4.3 using the non-negativity of matrix

$$\mathtt{J}^{\mathrm{w}}_{\varphi}(\mathbf{Y}|\mathbf{X}=\mathbf{x};\underline{\theta}) = \int f_{\underline{\theta}}(\mathbf{y}|\mathbf{x})\varphi(\mathbf{x},\mathbf{y})\left(\frac{\partial \log f_{\underline{\theta}}(\mathbf{y}|\mathbf{x})}{\partial\underline{\theta}}\right)^{\mathrm{T}}\frac{\partial \log f_{\underline{\theta}}(\mathbf{y}|\mathbf{x})}{\partial\underline{\theta}}\mathrm{d}\mathbf{y}.$$

Equality holds when $\mathtt{J}^{\mathrm{w}}_{\varphi}(\mathbf{Y}|\mathbf{X}=\mathbf{x};\underline{\theta})=0$, which leads to the statement. □

Lemma 4.5. (Data-processing inequality: cf. [32], Lemma 3) *For a given joint WF* $(\mathbf{x},\mathbf{y})\mapsto\varphi(\mathbf{x},\mathbf{y})$ *and a function* $\mathbf{x}\mapsto g(\mathbf{x})$*, set*

$$\varrho_g(\mathbf{x})=\varphi(\mathbf{x},g(\mathbf{x})) \quad \textit{and} \quad \rho_g(\mathbf{x})=\varphi(\mathbf{x},g(\mathbf{x}))f_{\underline{\theta}}(\mathbf{x}|g(\mathbf{x})). \tag{4.12}$$

Then we have

$$\mathtt{J}^{\mathrm{w}}_{\varrho_g}(\mathbf{X};\underline{\theta}) \geq \mathtt{J}^{\mathrm{w}}_{\rho_g}(g(\mathbf{X});\underline{\theta}). \tag{4.13}$$

The equality holds iff function $g(\mathbf{X})$ *is invertible.*

Proof. We make use of Lemma 4.4. Let $\mathbf{Y}=g(\mathbf{X})$, then $\mathtt{S}_{\underline{\theta}}=\mathtt{0}$. This yields

$$\mathtt{J}^{\mathrm{w}}_{\rho_g}(g(\mathbf{X});\underline{\theta}) \leq \mathtt{J}^{\mathrm{w}}_{\varphi}(\mathbf{X},g(\mathbf{X});\underline{\theta}). \tag{4.14}$$

Note that the equality holds true if $\mathtt{J}^{\mathrm{w}}_{\varphi}(\mathbf{X}|g(\mathbf{X});\underline{\theta})=0$, that is $g(\mathbf{X})$ is a sufficient statistic for $\underline{\theta}$. Now use the chain rule, Lemma 4.3, where $\mathtt{J}^{\mathrm{w}}_{\varphi}(g(\mathbf{X})|\mathbf{X};\underline{\theta})$ $=0$. Hence,

$$\mathtt{J}^{\mathrm{w}}_{\varphi}(\mathbf{X},g(\mathbf{X});\underline{\theta}) = \mathtt{J}^{\mathrm{w}}_{\varrho_g}(\mathbf{X};\underline{\theta}). \tag{4.15}$$

The assertions (4.14) and (4.15) lead directly to the result. □

Lemma 4.6. (Parameter transformation: cf. [32], Lemma 4) *Suppose we have a family of PDFs* $f_{\underline{\eta}}(\mathbf{x})$ *parameterized by a* $1\times m'$ *vector* $\underline{\eta}=(\eta_1,\ldots,\eta_{m'})\in\mathbb{R}^{m'}$*. Suppose that vector* $\underline{\eta}$ *is a function of* $\underline{\theta}\in\mathbb{R}^m$*. Then*

$$\mathtt{J}^{\mathrm{w}}_{\varphi}(\mathbf{X};\underline{\theta}) = \left(\frac{\partial\underline{\eta}}{\partial\underline{\theta}}\right)^{\mathrm{T}}\mathtt{J}^{\mathrm{w}}_{\varphi}\Big(\mathbf{X};\underline{\eta}(\underline{\theta})\Big)\left(\frac{\partial\underline{\eta}}{\partial\underline{\theta}}\right), \tag{4.16}$$

with an $m'\times m$ *matrix* $\dfrac{\partial\underline{\eta}}{\partial\underline{\theta}}=\left(\dfrac{\partial\eta_i}{\partial\theta_j}\right)$, $1\leq i\leq m'$, $1\leq j\leq m$.

In the linear case where $\underline{\eta}(\underline{\theta})=\underline{\theta}\mathtt{Q}$ *for some* $m\times m'$ *matrix* $\mathtt{Q}$*, we obtain:*

$$\mathtt{J}^{\mathrm{w}}_{\varphi}(\mathbf{X};\underline{\theta}) = \mathtt{Q}\mathtt{J}^{\mathrm{w}}_{\varphi}\big(\mathbf{X};\underline{\eta}(\underline{\theta})\big)\mathtt{Q}^{\mathrm{T}}. \tag{4.17}$$

Proof. Formula (4.16) becomes straightforward by substituting the expression

$$\frac{\partial\log f_{\underline{\eta}(\underline{\theta})}(x)}{\partial\underline{\theta}} = \left(\frac{\partial\underline{\eta}(\underline{\theta})}{\partial\underline{\theta}}\right)\left(\frac{\partial\log f_{\underline{\eta}}(x)}{\partial\underline{\eta}}\right)^{\mathrm{T}}. \tag{4.18}$$

□

Concluding this section, we consider a linear model where the parameter is related to an additive shift. Suppose a random vector $\mathbf{X}$ in $\mathbb{R}^n$ has a joint PDF $f_{\mathbf{X}}$ and $\mathbf{x} \in \mathbb{R}^n \mapsto \varphi(\mathbf{x})$ is a given WF. Set:

$$\mathtt{L}^{\mathrm{w}}_{\varphi}(\mathbf{X}) := \int \frac{\varphi(\mathbf{x})}{f(\mathbf{x})}\Big(\nabla f(\mathbf{x})\Big)^{\mathrm{T}} \nabla f(\mathbf{x}) \mathrm{d}\mathbf{x}. \tag{4.19}$$

Here and below, we use the symbol ∇ for the spatial gradient $1 \times n$ vectors as opposed to parameter gradients $\dfrac{\partial}{\partial \underline{\theta}}$ and $\dfrac{\partial}{\partial \underline{\eta}}$.

Furthermore, set

$$\mathbf{X} = \underline{\theta}\mathtt{Q} + \mathbf{Y}\mathtt{P}. \tag{4.20}$$

Here $\mathtt{Q}$ and $\mathtt{P}$ are two matrices, of sizes $m \times n$ and $k \times n$ respectively, with $m \leq k \leq n$. Next, $\mathbf{X} \in \mathbb{R}^n$ and $\mathbf{Y} \in \mathbb{R}^k$. Let $\mathbf{x} \in \mathbb{R}^n \mapsto \varphi(\mathbf{x}) \geq 0$ be a given WF and set

$$\psi(\mathbf{y}) = \psi_{\mathtt{P}}(\mathbf{y}) = \int_{\mathbb{R}^{n-k}} \varphi(\mathbf{x})\mathbf{1}(\mathbf{x}\mathtt{P}^{\mathrm{T}} = \mathbf{y}) f_{\mathbf{X}|\mathbf{X}\mathtt{P}^{\mathrm{T}}}(\mathbf{x}|\mathbf{y})\mathrm{d}\mathbf{x}^{\mathtt{C}}_{\mathtt{P}}, \quad \mathbf{y} \in \mathbb{R}^n\mathtt{P}^{\mathrm{T}},$$

where $\mathbf{x}^{\mathtt{C}}_{\mathtt{P}}$ stands for the complementary variable in $\mathbf{x}$, given that $\mathbf{x}\mathtt{P}^{\mathrm{T}} = \mathbf{y}$. In Lemma 4.7 we present relationships between $\mathtt{J}^{\mathrm{w}}_{\varphi}(\mathbf{X};\underline{\theta})$, $\mathtt{J}^{\mathrm{w}}_{\psi}(\mathbf{Y};\underline{\theta})$, $\mathtt{L}^{\mathrm{w}}_{\varphi}(\mathbf{X})$ and $\mathtt{L}^{\mathrm{w}}_{\psi}(\mathbf{X}\mathtt{P}^{\mathrm{T}})$ for the above model. (The proofs are straightforward and omitted.)

Lemma 4.7. (Cf. [32], Lemmas 5 and 6) *Assume the model* (4.20)*. Then*

$$\mathtt{J}^{\mathrm{w}}_{\varphi}(\mathbf{X};\underline{\theta}) = \mathtt{Q}\mathtt{L}^{\mathrm{w}}_{\varphi}(\mathbf{X})\mathtt{Q}^{\mathrm{T}}, \quad \mathtt{J}^{\mathrm{w}}_{\psi}(\mathbf{Y};\underline{\theta}) = \mathtt{Q}\mathtt{P}^{\mathrm{T}}\mathtt{L}^{\mathrm{w}}_{\psi}(\mathbf{X}\mathtt{P}^{\mathrm{T}})\mathtt{P}\mathtt{Q}^{\mathrm{T}}, \quad \textit{and}$$
$$\mathtt{J}^{\mathrm{w}}_{\varphi}(\mathbf{X};\underline{\theta}) \geq \mathtt{J}^{\mathrm{w}}_{\psi}(\mathbf{Y};\underline{\theta}). \tag{4.21}$$

Corollary 4.8. (Cf. [32], Corollary 1) *Let* $\mathtt{P}$ *be an* $m \times m$ *matrix. Let* $\mathbf{X}$ *be a random vector in* $\mathbb{R}^m$ *and WFs* φ *and* $\psi = \psi_{\mathtt{P}}$ *be as above. Then*

- (i) $\mathtt{J}^{\mathrm{w}}_{\varphi}(\mathbf{X};\underline{\theta}) \geq \mathtt{P}^{\mathrm{T}}\mathtt{J}^{\mathrm{w}}_{\psi}(\mathbf{X}\mathtt{P}^{\mathrm{T}};\underline{\theta})\mathtt{P}$.
- (ii) *For* $\mathtt{P}$ *with orthonormal rows (i.e., with* $\mathtt{P}\mathtt{P}^{\mathrm{T}}$ *equal to* I_m*, the unit* $m \times m$ *matrix),*

$$\mathtt{J}^{\mathrm{w}}_{\psi}(\mathbf{X}\mathtt{P}^{\mathrm{T}};\underline{\theta}) \leq \mathtt{P}^{\mathrm{T}}\mathtt{J}^{\mathrm{w}}_{\varphi}(\mathbf{X};\underline{\theta})\mathtt{P}. \tag{4.22}$$

- (iii) *For* $\mathtt{P}$ *with a full row rank* m*, and* $\mathbf{X} \in \mathbb{R}^m$ *with nonsingular* $\mathtt{J}^{\mathrm{w}}_{\varphi}$*,*

$$\mathtt{J}^{\mathrm{w}}_{\psi}(\mathbf{X}\mathtt{P}^{\mathrm{T}}) \leq \left(\mathtt{P}^{\mathrm{T}}\mathtt{J}^{\mathrm{w}}_{\psi}(\mathbf{X};\underline{\theta})^{-1}\mathtt{P}\right)^{-1}. \tag{4.23}$$

5. Weighted Cramér-Rao and Kullback inequalities

We start with multivariate weighted Cramér-Rao inequalities (WCRIs). As usually, consider a family of PDFs $f_{\underline{\theta}}(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^n$, dependent on a parameter $\underline{\theta} \in \mathbb{R}^m$ and let $\mathbf{X} = \mathbf{X}_{\underline{\theta}}$ denote the random vector with PDF $f_{\underline{\theta}}$. Let a statistic $\mathbf{x} \mapsto \mathbf{T}(\mathbf{x}) = (T_1(\mathbf{x}), ..., T_s(\mathbf{x}))$ and a WF $\mathbf{x} \mapsto \varphi(\mathbf{x}) \geq 0$ be given. With $\mathbb{E}_{\underline{\theta}}$ standing for the expectation relative to $f_{\underline{\theta}}$, set:

$$\alpha(\underline{\theta}) = \mathbb{E}_{\underline{\theta}}\,\varphi(\mathbf{X}), \quad \underline{\eta}(\underline{\theta}) = \mathbb{E}_{\underline{\theta}}\left[\varphi(\mathbf{X})\mathbf{T}(\mathbf{X})\right]. \tag{5.1}$$

We also suppose that the operations of taking expectation and the gradient are interchangeable:

$$\mathbb{E}_{\underline{\theta}}\Big[\varphi(\mathbf{X})\mathbf{S}(\mathbf{X},\underline{\theta})\Big] = \frac{\partial\alpha(\underline{\theta})}{\partial\underline{\theta}}, \quad \mathbb{E}_{\underline{\theta}}\Big[\varphi(\mathbf{X})\mathbf{T}(\mathbf{X})^{\mathrm{T}}\mathbf{S}(\mathbf{X},\underline{\theta})\Big] = \frac{\partial\underline{\eta}(\underline{\theta})}{\partial\underline{\theta}}, \tag{5.2}$$

assuming C^1-dependence in $\underline{\theta} \mapsto \alpha(\underline{\theta})$ and $\underline{\theta} \mapsto \underline{\eta}(\underline{\theta})$ and absolute convergence of the integrals involved. Let $\mathtt{C}^{\mathrm{w}}_{\varphi}(\underline{\theta})$ denote the weighted covariance matrix for $\mathbf{X}$:

$$\mathtt{C}^{\mathrm{w}}_{\varphi}(\underline{\theta}) = \mathbb{E}_{\underline{\theta}}\Big\{\varphi(\mathbf{X})\Big[\mathbf{T}(\mathbf{X}) - \underline{\eta}(\mathbf{X})\Big]^{\mathrm{T}}\Big[\mathbf{T}(\mathbf{X}) - \underline{\eta}(\mathbf{X})\Big]\Big\} \tag{5.3}$$

and $\mathtt{J}^{\mathrm{w}}_{\varphi}(\mathbf{X};\underline{\theta}) = \mathbb{E}\left[\varphi(\mathbf{X})\mathbf{S}(\mathbf{X},\underline{\theta})^{\mathrm{T}}\mathbf{S}(\mathbf{X},\underline{\theta})\right]$ be the weighted Fisher information matrix under the WF φ; cf. Eq. (4.1).

Theorem 5.1. (A weighted Cramér-Rao inequality, version I; [4], Theorem 11.10.1, [5], Theorem 20.) *Assuming $\mathtt{J}^{\mathrm{w}}_{\varphi}(\mathbf{X};\underline{\theta})$ is invertible and under condition* (5.2), *vectors $\underline{\eta}(\underline{\theta})$, $\dfrac{\partial\alpha(\underline{\theta})}{\partial\underline{\theta}}$ and matrices $\mathtt{C}^{\mathrm{w}}_{\varphi}(\underline{\theta})$, $\mathtt{J}^{\mathrm{w}}_{\varphi}(\mathbf{X};\underline{\theta})$, $\dfrac{\partial\underline{\eta}(\underline{\theta})}{\partial\underline{\theta}}$ obey*

$$\mathtt{C}^{\mathrm{w}}_{\varphi}(\underline{\theta}) \geq \left[\frac{\partial\underline{\eta}(\underline{\theta})}{\partial\underline{\theta}} - \big(\underline{\eta}(\underline{\theta})\big)^{\mathrm{T}}\frac{\partial\alpha(\underline{\theta})}{\partial\underline{\theta}}\right]\,\Big[\mathtt{J}^{\mathrm{w}}_{\varphi}(\mathbf{X};\underline{\theta})\Big]^{-1}\left[\frac{\partial\underline{\eta}(\underline{\theta})}{\partial\underline{\theta}} - \big(\underline{\eta}(\underline{\theta})\big)^{\mathrm{T}}\frac{\partial\alpha(\underline{\theta})}{\partial\underline{\theta}}\right]^{\mathrm{T}}. \tag{5.4}$$

Proof. We start with a simplified version where $s = 1$ and $\mathbf{T}(\mathbf{X}) = T(\mathbf{X})$ and $\underline{\eta}(\underline{\theta}) = \eta(\underline{\theta})$ are scalars, keeping general $n, m \geq 1$. By using (5.2), write:

$$\begin{aligned}&\mathbb{E}_{\underline{\theta}}\left\{\varphi(\mathbf{X})\left[T(\mathbf{X}) - \eta(\underline{\theta})\right]\mathbf{S}(\mathbf{X};\underline{\theta})\right\} \\ &\quad = \mathbb{E}_{\underline{\theta}}\left[\varphi(\mathbf{X})T(\mathbf{X})\mathbf{S}(\mathbf{X};\underline{\theta})\right] - \eta(\underline{\theta})\mathbb{E}_{\underline{\theta}}\left[\varphi(\mathbf{X})\mathbf{S}(\mathbf{X};\underline{\theta})\right] = \frac{\partial\eta(\underline{\theta})}{\partial\underline{\theta}} - \eta(\underline{\theta})\frac{\partial\alpha(\underline{\theta})}{\partial\underline{\theta}}.\end{aligned} \tag{5.5}$$

Then for any $1 \times m$ vector $\underline{\mu} \in \mathbb{R}^m$,

$$\begin{aligned}0 &\leq \mathbb{E}_{\underline{\theta}}\left\{\varphi(\mathbf{X})\Big[T(\mathbf{X}) - \eta(\underline{\theta}) - \mathbf{S}(\mathbf{X},\underline{\theta})\underline{\mu}^{\mathrm{T}}\Big]^2\right\} \\ &= \mathbb{E}_{\underline{\theta}}\left\{\varphi(\mathbf{X})\Big[T(\mathbf{X}) - \eta(\underline{\theta})\Big]^2\right\} + \underline{\mu}\mathtt{J}^{\mathrm{w}}_{\varphi}(\mathbf{X};\underline{\theta})\underline{\mu}^{\mathrm{T}} \\ &\quad - 2\underline{\mu}\left(\frac{\partial\eta(\underline{\theta})}{\partial\underline{\theta}} - \eta(\underline{\theta})\frac{\partial\alpha(\underline{\theta})}{\partial\underline{\theta}}\right)^{\mathrm{T}}.\end{aligned} \tag{5.6}$$

Taking $\underline{\mu} = \left(\dfrac{\partial\eta(\underline{\theta})}{\partial\underline{\theta}} - \eta(\underline{\theta})\dfrac{\partial\alpha(\underline{\theta})}{\partial\underline{\theta}}\right)\left[\mathtt{J}^{\mathrm{w}}_{\varphi}(\mathbf{X};\underline{\theta})\right]^{-1}$ [which is the minimiser for the RHS in (5.6)], we obtain

$$\begin{aligned}\mathrm{Var}^{\mathrm{w}}_{\varphi}[T(\mathbf{X})] &:= \mathbb{E}_{\underline{\theta}}\left\{\varphi(\mathbf{X})\Big(T(\mathbf{X})-\eta(\underline{\theta})\Big)^2\right\}\\ &\geq \left(\frac{\partial\eta(\underline{\theta})}{\partial\underline{\theta}}-\eta(\underline{\theta})\frac{\partial\alpha(\underline{\theta})}{\partial\underline{\theta}}\right)\left[\mathtt{J}^{\mathrm{w}}_{\varphi}(\mathbf{X};\underline{\theta})\right]^{-1}\left(\frac{\partial\eta(\underline{\theta})}{\partial\underline{\theta}}-\eta(\underline{\theta})\frac{\partial\alpha(\underline{\theta})}{\partial\underline{\theta}}\right)^{\mathrm{T}}.\end{aligned}\tag{5.7}$$

Turning to the general case $s\geq 1$, set: $T(\mathbf{X})=\mathbf{T}(\mathbf{X})\underline{\lambda}^{\mathrm{T}}$ where the $1\times s$ vector $\underline{\lambda}\in\mathbb{R}^s$. Then (5.7) yields that for all $\underline{\lambda}$,

$$\begin{aligned}\underline{\lambda}\mathtt{C}^{\mathrm{w}}_{\varphi}(\underline{\theta})\underline{\lambda}^{\mathrm{T}} = \mathrm{Var}^{\mathrm{w}}_{\varphi}[\mathbf{T}(\mathbf{X})\underline{\lambda}^{\mathrm{T}}] &:= \mathbb{E}_{\underline{\theta}}\left\{\varphi(\mathbf{X})\Big[\mathbf{T}(\mathbf{X})\underline{\lambda}^{\mathrm{T}}-\underline{\eta}(\underline{\theta})\underline{\lambda}^{\mathrm{T}}\Big]^2\right\}\\ &\geq \underline{\lambda}\left(\frac{\partial\underline{\eta}(\underline{\theta})}{\partial\underline{\theta}}-\big(\underline{\eta}(\underline{\theta})\big)^{\mathrm{T}}\frac{\partial\alpha(\underline{\theta})}{\partial\underline{\theta}}\right)\Big[\mathtt{J}^{\mathrm{w}}_{\varphi}(\mathbf{X};\underline{\theta})\Big]^{-1}\\ &\quad\times\left(\frac{\partial\underline{\eta}(\underline{\theta})}{\partial\underline{\theta}}-\big(\underline{\eta}(\underline{\theta})\big)^{\mathrm{T}}\frac{\partial\alpha(\underline{\theta})}{\partial\underline{\theta}}\right)^{\mathrm{T}}\underline{\lambda}^{\mathrm{T}},\end{aligned}$$

implying (5.4). □

Definition 5.2. The **calibrated** relative WE $K^{\mathrm{w}}_{\varphi}(f||g)$ of f and g with WF φ is defined by

$$K^{\mathrm{w}}_{\varphi}(f||g)=\int\frac{\varphi(\mathbf{x})f(\mathbf{x})}{\alpha(f)}\log\frac{f(\mathbf{x})\alpha(g)}{g(\mathbf{x})\alpha(f)}\mathrm{d}\mathbf{x}=\frac{D^{\mathrm{w}}_{\varphi}(f\|g)}{\alpha(f)}+\log\frac{\alpha(g)}{\alpha(f)}=D(\widetilde{f}\|\widetilde{g}).\tag{5.8}$$

Here $\widetilde{f}$ and $\widetilde{g}$ are PDFs produced from φf and φg after normalizing by $\alpha(f)$ and $\alpha(g)$:

$$\begin{aligned}&\alpha(f)=\int\varphi(\mathbf{x})f(\mathbf{x})\mathrm{d}\mathbf{x},\ \alpha(g)=\int\varphi(\mathbf{x})g(\mathbf{x})\mathrm{d}\mathbf{x},\ \widetilde{f}(\mathbf{x})=\frac{\varphi(\mathbf{x})f(\mathbf{x})}{\alpha(f)},\\ &\widetilde{g}(\mathbf{x})=\frac{\varphi(\mathbf{x})g(\mathbf{x})}{\alpha(g)},\end{aligned}\tag{5.9}$$

and $D(\,\cdot\,\|\,\cdot\,)$ is the standard Kullback–Leibler divergence.

Theorem 5.3. (Weighted Kullback inequalities, cf. [10]) *For given φ and f, g as above, the following bounds hold true. First, for the $1\times n$ vector $\boldsymbol{\zeta}$,*

$$K^{\mathrm{w}}_{\varphi}(f\|g)\geq\sup\left[\frac{\mathbf{e}_{\varphi}(f)\boldsymbol{\zeta}^{\mathrm{T}}}{\alpha(f)}+\log\alpha(g)-\log M_g(\boldsymbol{\zeta}):\ \boldsymbol{\zeta}\in\mathbb{R}^n\right],\tag{5.10}$$

where

$$\mathbf{e}_{\varphi}(f)=\int\varphi(\mathbf{x})f(\mathbf{x})\mathbf{x}\,\mathrm{d}\mathbf{x},\quad M_g(\boldsymbol{\zeta})=\int\varphi(\mathbf{x})g(\mathbf{x})\left[\exp(\mathbf{x}\boldsymbol{\zeta}^{\mathrm{T}})\right]\mathrm{d}\mathbf{x}.\tag{5.11}$$

Second,

$$D^{\mathrm{w}}_{\varphi}(f\|g)\geq\sup\left[\mathbf{e}_{\varphi}(f)\boldsymbol{\zeta}^{\mathrm{T}}:\ \boldsymbol{\zeta}\in\mathbb{M}\right],\tag{5.12}$$

where

$$\mathbb{M}=\left\{\boldsymbol{\zeta}:\ \int\varphi(\mathbf{x})\Big(f(\mathbf{x})-g(\mathbf{x})[\exp(\mathbf{x}\boldsymbol{\zeta}^{\rm T})]\Big)\mathrm{d}\mathbf{x}\geq 0\right\}. \tag{5.13}$$

Proof. First, given $\boldsymbol{\zeta}\in\mathbb{R}^n$, set $\widetilde{G}_{\boldsymbol{\zeta}}(\mathbf{x})=\dfrac{\varphi(\mathbf{x})g(\mathbf{x})\left[\exp(\mathbf{x}\boldsymbol{\zeta}^{\rm T})\right]}{M_g(\boldsymbol{\zeta})}$. Following (5.11) and (5.8), obtain:

$$K^{\rm w}_\varphi(f\|g)=D(\widetilde{f}\|\widetilde{G}_{\boldsymbol{\zeta}})+\int\widetilde{f}(\mathbf{x})\log\frac{\widetilde{G}_{\boldsymbol{\zeta}}(\mathbf{x})}{\widetilde{g}(\mathbf{x})}\mathrm{d}\mathbf{x}\geq\int\widetilde{f}(\mathbf{x})\log\frac{\alpha(g)[\exp(\mathbf{x}\boldsymbol{\zeta}^{\rm T})]}{M_g(\boldsymbol{\zeta})}\mathrm{d}\mathbf{x}; \tag{5.14}$$

the bound holds as $D(\widetilde{f}\|\widetilde{G}_{\boldsymbol{\zeta}})\geq 0$ by the Gibbs inequality for the standard Kullback–Leibler divergence. By taking the supremum, we arrive at (5.10).

Second, write: $G_{\boldsymbol{\zeta}}(\mathbf{x})=g(\mathbf{x})[\exp(\mathbf{x}\boldsymbol{\zeta}^{\rm T})]$ and

$$D^{\rm w}_\varphi(f\|g)=D^{\rm w}_\varphi(f\|G_{\boldsymbol{\zeta}})+\mathbf{e}_\varphi(f)\boldsymbol{\zeta}^{\rm T}. \tag{5.15}$$

For $\boldsymbol{\zeta}\in\mathbb{M}$, the bound $D^{\rm w}_\varphi(f\|G_{\boldsymbol{\zeta}})\geq 0$ holds true [the weighted Gibbs inequality (1.3)]. This yields (5.12). □

An application of the weighted Kullback's inequality is given in the next theorem where we obtain another version of the weighted Cramér–Rao inequality.

Theorem 5.4. (A weighted Cramér-Rao inequality, version II; [4], Theorem 11.10.1, [5], Theorem 20) *Suppose we have a family of* $1\times n$ *random vectors* $\mathbf{X}$, *with PDFs* $f_{\underline{\theta}}(\mathbf{x})$, $\mathbf{x}\in\mathbb{R}^n$, *indexed by* $\underline{\theta}\in\mathbb{R}^m$. *Suppose that* $\dfrac{f_{\underline{\theta}}(\mathbf{x})\alpha(\underline{\theta}+\underline{\varepsilon})}{f_{\underline{\theta}+\underline{\varepsilon}}(\mathbf{x})\alpha(\underline{\theta})}\to 1$ *as* $\underline{\varepsilon}\to 0$ *uniformly in* $\mathbf{x}$. *Let* $\mathbf{x}\mapsto\varphi(\mathbf{x})$ *be a given WF. Denoting, as before, the expectation relative to* $f_{\underline{\theta}}$ *by* $\mathbb{E}_{\underline{\theta}}$, *set* $\alpha(\underline{\theta})=\mathbb{E}_{\underline{\theta}}[\varphi(\mathbf{X})]$, $\mathbf{e}(\underline{\theta})=\mathbb{E}_{\underline{\theta}}[\varphi(\mathbf{X})\mathbf{X}]$ *and*

$$\widetilde{\mathtt{C}}^{\rm w}_\varphi(\underline{\theta})=\frac{1}{\alpha(\underline{\theta})}\mathbb{E}_{\underline{\theta}}\big[\varphi(\mathbf{X})\mathbf{X}^{\rm T}\mathbf{X}\big]-\mathbf{e}(\underline{\theta})^{\rm T}\mathbf{e}(\underline{\theta}). \tag{5.16}$$

Under the assumptions needed to define matrix $\mathtt{J}^{\rm w}_\varphi(\mathbf{X};\underline{\theta})$,

$$\mathtt{J}^{\rm w}_\varphi(\mathbf{X};\underline{\theta})\geq\frac{\partial\mathbf{e}(\underline{\theta})}{\partial\underline{\theta}}^{\rm T}\Big[\widetilde{\mathtt{C}}^{\rm w}_\varphi(\underline{\theta})\Big]^{-1}\frac{\partial\mathbf{e}(\underline{\theta})}{\partial\underline{\theta}}+\alpha(\underline{\theta})^{-1}\frac{\partial\alpha(\underline{\theta})}{\partial\underline{\theta}}^{\rm T}\frac{\partial\alpha(\underline{\theta})}{\partial\underline{\theta}}. \tag{5.17}$$

Proof. By definition (5.8), for $\underline{\varepsilon}\in\mathbb{R}^m$,

$$K^{\rm w}_\varphi(f_{\underline{\theta}+\underline{\varepsilon}}||f_{\underline{\theta}})=-\int\varphi(\mathbf{x})\frac{f_{\underline{\theta}+\underline{\varepsilon}}(\mathbf{x})}{\alpha(\underline{\theta}+\underline{\varepsilon})}\log\frac{f_{\underline{\theta}}(\mathbf{x})\alpha(\underline{\theta}+\underline{\varepsilon})}{f_{\underline{\theta}+\underline{\varepsilon}}(\mathbf{x})\alpha(\underline{\theta})}\mathrm{d}\mathbf{x}. \tag{5.18}$$

Next, set $M(\underline{\theta},\boldsymbol{\zeta})=\mathbb{E}_{\underline{\theta}}\left\{\varphi(\mathbf{X})[\exp(\mathbf{X}\boldsymbol{\zeta}^{\rm T})]\right\}$ and

$$\Psi(\underline{\theta},\underline{\varepsilon})=\sup\Big[\mathbf{e}(\underline{\theta}+\underline{\varepsilon})\boldsymbol{\zeta}^{\rm T}+\log\alpha(\underline{\theta})-\log M(\underline{\theta},\boldsymbol{\zeta}):\ \boldsymbol{\zeta}\in\mathbb{R}^n\Big]. \tag{5.19}$$

Then, owing to Theorem 5.3, we obtain:

$$K^{\mathrm{w}}_{\varphi}(f_{\underline{\theta}+\underline{\varepsilon}}||f_{\underline{\theta}}) \geq \Psi(\underline{\theta}, \underline{\varepsilon}). \tag{5.20}$$

The LHS of (5.20) is

$$\begin{aligned} &-\int \varphi(\mathbf{x}) f_{\underline{\theta}+\underline{\varepsilon}}(\mathbf{x}) \log \frac{f_{\underline{\theta}}(\mathbf{x})\alpha(\underline{\theta}+\underline{\varepsilon})}{f_{\underline{\theta}+\underline{\varepsilon}}(\mathbf{x})\alpha(\underline{\theta})} \mathrm{d}\mathbf{x} = \int \varphi(\mathbf{x}) f_{\underline{\theta}+\underline{\varepsilon}}(\mathbf{x}) \Bigg\{ \left[1 - \frac{f_{\underline{\theta}}(\mathbf{x})\alpha(\underline{\theta}+\underline{\varepsilon})}{f_{\underline{\theta}+\underline{\varepsilon}}(\mathbf{x})\alpha(\underline{\theta})}\right] \\ &+ \frac{1}{2}\left[1 - \frac{f_{\underline{\theta}}(\mathbf{x})\alpha(\underline{\theta}+\underline{\varepsilon})}{f_{\underline{\theta}+\underline{\varepsilon}}(\mathbf{x})\alpha(\underline{\theta})}\right]^2 + O\left(\left[1 - \frac{f_{\underline{\theta}}(\mathbf{x})\alpha(\underline{\theta}+\underline{\varepsilon})}{f_{\underline{\theta}+\underline{\varepsilon}}(\mathbf{x})\alpha(\underline{\theta})}\right]^3\right)\Bigg\} \mathrm{d}\mathbf{x}. \end{aligned} \tag{5.21}$$

Here we have used the Taylor expansion of $\log(1+z)$. The first-order term disappears:

$$\int \varphi(\mathbf{x}) f_{\underline{\theta}+\underline{\varepsilon}}(\mathbf{x}) \left[1 - \frac{f_{\underline{\theta}}(\mathbf{x})\alpha(\underline{\theta}+\underline{\varepsilon})}{f_{\underline{\theta}+\underline{\varepsilon}}(\mathbf{x})\alpha(\underline{\theta})}\right] \mathrm{d}\mathbf{x} = \alpha(\underline{\theta}+\underline{\varepsilon}) - \alpha(\underline{\theta}+\underline{\varepsilon}) = 0. \tag{5.22}$$

Next, for small $\underline{\varepsilon}$,

$$\begin{aligned} &\int \varphi(\mathbf{x}) f_{\underline{\theta}+\underline{\varepsilon}}(\mathbf{x}) \left[1 - \frac{f_{\underline{\theta}}(\mathbf{x})\alpha(\underline{\theta}+\underline{\varepsilon})}{f_{\underline{\theta}+\underline{\varepsilon}}(\mathbf{x})\alpha(\underline{\theta})}\right]^2 \mathrm{d}\mathbf{x} \\ &\quad = \underline{\varepsilon}\left[\mathrm{J}^{\mathrm{w}}_{\varphi}(\mathbf{X};\underline{\theta}) - \frac{1}{\alpha(\underline{\theta})}\frac{\partial\alpha(\underline{\theta})}{\partial\underline{\theta}}\left(\frac{\partial\alpha(\underline{\theta})}{\partial\underline{\theta}}\right)^{\mathrm{T}}\right]\underline{\varepsilon}^{\mathrm{T}} + o(\|\underline{\varepsilon}\|^2). \end{aligned} \tag{5.23}$$

Finally, the remainder

$$\lim_{\underline{\varepsilon}\to 0} \frac{1}{\|\underline{\varepsilon}\|^2} \int \varphi(\mathbf{x}) f_{\underline{\theta}+\underline{\varepsilon}}(\mathbf{x}) O\left(\left[1 - \frac{f_{\underline{\theta}}(\mathbf{x})\alpha(\underline{\theta}+\underline{\varepsilon})}{f_{\underline{\theta}+\underline{\varepsilon}}(\mathbf{x})\alpha(\underline{\theta})}\right]^3\right) \mathrm{d}\mathbf{x} = o(\|\underline{\varepsilon}\|^2). \tag{5.24}$$

For the RHS in (5.20), we take the point $\boldsymbol{\tau}$ where the gradient has the form $\nabla_{\boldsymbol{\zeta}}\left[\mathbf{e}(\underline{\theta}+\underline{\varepsilon})\boldsymbol{\zeta}^{\mathrm{T}} + \log\alpha(\underline{\theta}) - \log M(\underline{\theta},\boldsymbol{\zeta})\right]\Big|_{\boldsymbol{\zeta}=\boldsymbol{\tau}} = 0$, i.e.,

$$\mathbf{e}(\underline{\theta}+\underline{\varepsilon}) = \nabla_{\boldsymbol{\zeta}} \log M(\underline{\theta},\boldsymbol{\zeta})\Big|_{\boldsymbol{\zeta}=\boldsymbol{\tau}} = \frac{1}{M(\underline{\theta},\boldsymbol{\tau})}\nabla_{\boldsymbol{\zeta}} M(\underline{\theta},\boldsymbol{\tau})\Big|_{\boldsymbol{\zeta}=\boldsymbol{\tau}}.$$

Consider the limit

$$\lim_{\underline{\varepsilon}\to 0} \frac{1}{\|\underline{\varepsilon}\|^2} \sup_{\mathbf{t}\in\mathbb{R}^n} \left\{\mathbf{t}^{\mathrm{T}}\boldsymbol{\mu}_{\varphi}(\underline{\theta}+\underline{\varepsilon}) - \overline{\Psi}(\mathbf{t})\right\}. \tag{5.25}$$

Here $\overline{\Psi}(\mathbf{t}) = \log\alpha(\underline{\theta}) + \log\int f_{\underline{\theta}}(\mathbf{x})[\exp(\mathbf{x}\mathbf{t}^{\mathrm{T}})]\mathrm{d}\mathbf{x}$ denotes the weighted cumulant-generating function for PDF $\widetilde{f}_{\underline{\theta}}$. The supremum is attained at a value of $\mathbf{t} = \boldsymbol{\tau} = \boldsymbol{\tau}(\underline{\varepsilon})$ where the first derivative of the weighted cumulant-generating function equals $\nabla_{\mathbf{t}}\overline{\Psi}(\mathbf{t}=\boldsymbol{\tau}) = \boldsymbol{\mu}_{\varphi}(\underline{\theta}+\underline{\varepsilon})$. Here $\boldsymbol{\mu}_{\varphi}(\underline{\theta}) = \mathbb{E}_{\underline{\theta}}[\mathbf{X}\varphi(\mathbf{X})]/\mathbb{E}_{\underline{\theta}}\varphi(\mathbf{X})$. We also have $\nabla_{\mathbf{t}}\overline{\Psi}(0) = \boldsymbol{\mu}_{\varphi}(\underline{\theta})$, and therefore the Hessian

$$\nabla_{\mathbf{t}}\overline{\Psi}(0) = \frac{\partial}{\partial\underline{\theta}}\boldsymbol{\mu}_{\varphi}(\underline{\theta}) \lim_{\underline{\varepsilon}\to 0}\frac{\partial\underline{\varepsilon}}{\partial\boldsymbol{\tau}}. \tag{5.26}$$

It can also be seen that

$$\nabla\nabla\overline{\Psi}(0) = \frac{\mathbb{E}_{\underline{\theta}}\big[\mathbf{X}^{\mathrm{T}}\mathbf{X}\varphi(\mathbf{X})\big]}{\mathbb{E}_{\underline{\theta}}[\varphi(\mathbf{X})]} - \boldsymbol{\mu}_{\varphi}(\underline{\theta})^{\mathrm{T}}\boldsymbol{\mu}_{\varphi}(\underline{\theta}) := \overline{\mathbf{V}}_{\varphi}(\mathbf{X};\underline{\theta}). \tag{5.27}$$

In addition, by using the Taylor formula at an intermediate point between $\underline{\theta}$ and $\underline{\theta}+\underline{\varepsilon}$,

$$\lim_{\underline{\varepsilon}\to 0}\frac{1}{\|\underline{\varepsilon}\|^2}\Big\{\boldsymbol{\tau}^{\mathrm{T}}\boldsymbol{\mu}_{\varphi}(\underline{\theta}+\underline{\varepsilon}) - \overline{\Psi}(\boldsymbol{\tau})\Big\} = \left(\frac{\partial}{\partial\underline{\theta}}\boldsymbol{\mu}_{\varphi}(\underline{\theta})\right)\frac{1}{2}[\nabla\nabla\overline{\Psi}(0)]^{-1}\left(\frac{\partial}{\partial\underline{\theta}}\boldsymbol{\mu}_{\varphi}(\underline{\theta})\right)^{\mathrm{T}}. \tag{5.28}$$

Now let us go back to the RHS of (5.25) which becomes:

$$\begin{aligned}\lim_{\underline{\varepsilon}\to 0}&\frac{1}{\|\underline{\varepsilon}\|^2}\left[\boldsymbol{\tau}^{\mathrm{T}}\boldsymbol{\mu}_{\varphi}(\underline{\theta}+\underline{\varepsilon}) - \Psi(\boldsymbol{\tau}) + \log\alpha(\underline{\theta})\right]\\ &= \frac{1}{2}\left(\frac{\partial}{\partial\underline{\theta}}\boldsymbol{\mu}_{\varphi}(\underline{\theta})\right)[\overline{\mathbf{V}}_{\varphi}(\mathbf{X};\underline{\theta})]^{-1}\left(\frac{\partial}{\partial\underline{\theta}}\boldsymbol{\mu}_{\varphi}(\underline{\theta})\right)^{\mathrm{T}}.\end{aligned} \tag{5.29}$$

Now (5.29) gives the required result (5.17). $\square$

Remark 5.5. When $\varphi(\mathbf{x})\equiv 1$ then $\alpha(\underline{\theta})=1$, $\mathbf{e}(\underline{\theta})=\mathbb{E}_{\underline{\theta}}\mathbf{X}$, $\mathtt{C}^{\mathrm{w}}_{\varphi}(\underline{\theta})=\widetilde{\mathtt{C}}^{\mathrm{w}}_{\varphi}(\underline{\theta})$, and the two inequalities (5.4) and (5.17) coincide.

In general, these inequalities are competing; the question which inequality is stronger is not discussed in this paper. We also note that both inequalities (5.4) and (5.17) lack a covariant property: multiplying WF φ by a constant has a different impact on the left- and right-hand sides.

Acknowledgements

YS thanks the Office of the Rector, University of Sao Paulo (USP) for the financial support during the academic year 2013-4. YS thanks Math Department, Penn State University, USA for the hospitality and support during the academic years 2014-6. IS is supported by FAPESP Grant—process number 11/51845-5, and expresses her gratitude to IMS, University of São Paulo, Brazil, and to Math Department, University of Denver, USA for the warm hospitality. SYS thanks the CAPES PNPD-UFSCAR Foundation for the financial support in the year 2014. SYS thanks the Federal University of Sao Carlos, Department of Statistics, for hospitality during the year 2014. MK thanks the Higher School of Economics for the support in the framework of the Global Competitiveness Program.

References

[1] Belis, M., Guiasu, S.: A Quantitative and qualitative measure of information in cybernetic systems. IEEE Trans. Inf. Theory **14**, 593–594 (1968)

[2] Clim, A.: Weighted entropy with application. In: Analele Universităţii Bucureşti, Matematică, Anul LVII, pp. 223–231 (2008)
[3] Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley, New York (2006)
[4] Cover, T.M., Thomas, J.A.: Determinant inequalities via information theory. SIAM J. Matrix Anal. Appl. **9**, 384–392 (1988)
[5] Dembo, A., Cover, T.M., Thomas, J.A.: Information theoretic inequalities. IEEE Trans. Inf. Theory **37**, 1501–1518 (1991)
[6] Di Crescenzo, A., Longobardi, M.: Entropy based measure of uncertainty in past lifetime distributions. J. Appl. Prob. **39**(3), 434–440 (2002)
[7] Dial, G., Taneja, I.J.: On weighted entropy of type (α,β) and its generalizations. Appl. Math. **26**, 418–425 (1981)
[8] Frizelle, G., Suhov, Y.M.: An entropic measurement of queueing behaviour in a class of manufacturing operations. Proc. R. Soc. A **457**, 1579–1601 (2001)
[9] Frizelle, G., Suhov, Y.M.: The measurement of complexity in production and other commercial systems. Proc. R. Soc. A **464**, 2649–2668 (2008)
[10] Fuchs, A., Letta, G.: L'inégalité de Kullback. Application à la théorie de l'estimation. Séminaire de probabilités 4, pp. 108–131. Strasbourg (1970)
[11] Guiasu, S.: Weighted entropy. Rep. Math. Phys. **2**, 165–179 (1971)
[12] Ito, K.: Introduction to Probability Theory. Cambridge University Press, Cambridge (1984)
[13] Johnson, D.H., Glantz, R.M.: When does interval coding occur? Neurocopmuting **59:60**, 13–18 (2004)
[14] Kannappan, P.L., Sahoo, P.K.: On the general solution of a functional equation connected to sum form information measures on open domain. Math. Sci. **9**, 545–550 (1986)
[15] Kapur, J.N.: Measures of Information and Their Applications, Chapter 17. Wiley, New Delhi (1994)
[16] Fan, K.: On a theorem of Weyl concerning eigenvalues of linear transformations. I. Proc. Natl. Acad. USA **35**, 652–655 (1949)
[17] Fan, K.: On a theorem of Weyl concerning eigenvalues of linear transformations. II. Proc. Natl. Acad. USA **36**, 31–35 (1950)
[18] Fan, K.: Maximum properies and inequalities for the eigenvalues of completely continuous operators. Proc. Natl. Acad. USA **37**, 760–766 (1951)
[19] Kelbert, M., Suhov, Y.: Continuity of mutual entropy in the limiting signal-to-noise ratio regimes. In: Stochastic Analysis, pp. 281–299. Springer, Berlin (2010)
[20] Kelbert, M., Suhov, Y.: Information Theory and Coding by Example. Cambridge University Press, Cambridge (2013)
[21] Moslehian, M.: Ky Fan inequalities. arXiv:1108.1467 (2011)
[22] Muandet, K., Marukatat, S., Nattee, C.: Query selection via weighted entropy in graph-based semi-supervised classification. In: Advances in Machine Learning. Lecture Notes in Computer Science, vol. 5828, pp. 278–292 (2009)
[23] Parkash, O., Taneja, H.C.: Characterization of the quantitative-qualitative measure of inaccuracy for discrete generalized probability distributions. Commun. Stat. Theory Methods **15**, 3763–3771 (1986)
[24] Sharma, B.D., Mitter, J., Mohan, M.: On measure of 'useful' information. Inf. Control **39**, 323–336 (1978)
[25] Singh, R.P., Bhardwaj, J.D.: On parametric weighted information improvement. Inf. Sci. **59**, 149–163 (1992)
[26] Sreevally, A., Varma, S.K.: Generating measure of cross entropy by using measure of weighted entropy. Soochow J. Math. **30**(2), 237–243 (2004)
[27] Srivastava, A.: Some new bounds of weighted entropy measures. Cybern. Inf. Technol. **11**(3), 60–65 (2011)
[28] Suhov, Y., Yasaei Sekeh, S., Kelbert, M.: Entropy-power inequality for weighted entropy. arXiv:1502.02188

[29] Suhov, Y., Stuhl, I., Yasaei Sekeh, S.: Weighted Gaussian entropy and determinant inequalities. arXiv:1505.01753

[30] Suhov, Y., Stuhl, I., Kelbert, M.: Weight functions and log-optimal investment portfolios. arXiv:1505.01437

[31] Tuteja, R.K., Chaudhary, Sh., Jain, P.: Weighted entropy of orders α and type β information energy. Soochow J. Math. **19**(2), 129–138 (1993)

[32] Zamir, R.: A proof of the Fisher information inequality via a data processing argument. IEEE Trans. Inf. Theory **44**(3), 1246–1250 (1998)

Yuri Suhov
DPMMS, University of Cambridge
Cambridge
UK

Yuri Suhov
Mathematics Department, Penn State University
Philadelphia PA
USA

Yuri Suhov
IPIT RAS
Moscow
Russian Federation

Izabella Stuhl
IMS, University of São Paulo
São Paulo, SP
Brazil

Izabella Stuhl
Mathematics Department, University of Denver
Denver CO
USA
e-mail: izabella.stuhl@du.edu

Izabella Stuhl
University of Debrecen
Debrecen
Hungary

Salimeh Yasaei Sekeh
Statistics Department, Federal University of São Carlos
São Carlos SP
Brazil

Mark Kelbert
Mathematics Department, University of Swansea
Swansea
UK

Mark Kelbert
Moscow Higher School of Economics
Moscow
Russian Federation

Received: October 28, 2015
Revised: December 7, 2015