MEM pixel correlated solutions for generalized moment and interpolation problems

Imre Csiszár, Fabrice Gamboa, Elisabeth Gassiat

Abstract

In generalized moment problems (signed) measures are searched to fit given observations, or continuous functions are searched to fit given constraints. Known convex methods for solving such problems, and their stochastic interpretations via Maximum Entropy on the Mean (MEM) and in a Bayesian sense are reviewed, with some improvements on previous results. Then the MEM and Bayesian approaches are extended to default models with a dependence structure, yielding new families of solutions. One family involves a transfer kernel, and allows using prior information such as modality, convexity, or Sobolev norms. Another family of solutions with possibly non-convex criteria, is arrived at using default models with exchangeable random variables. The main technical tools are convex analysis and large deviations theory.

Keywords

Maximum entropy on the mean, interpolation, correlated pixels, prior information, convex distances, large deviations.

I. Introduction

In many physical problems, a real valued function $g(x)$ defined on a given set U ought to be reconstructed based on the information that for a vector-valued function $\varphi(x) = (\varphi_1(x), \ldots, \varphi_k(x))$, the integral of $g(x)\varphi(x)$ belongs to a given closed convex set $\mathcal{K} \subset \mathbb{R}^k$, i.e.,

$$\int_U g\varphi dP_U \in \mathcal{K} \tag{I.1}$$

where P_U is a given reference measure. In the sequel, P_U will be assumed to be a probability measure (p.m.).

Typically, the integral in (I.1) represents k "ideal" measurement results, and if the actual measurements are completely reliable, with results $(c_1, \ldots, c_k)$, then $\mathcal{K}$ consists of a single point $c = (c_1, \ldots, c_k)$. Otherwise, $\mathcal{K}$ is a confidence set reflecting that the measurements were subject to errors. Often, some prior knowledge about g is also available, such as upper and lower bounds, an energy bound, smoothness, or shape information such as unimodality or convexity. The problem of reconstructing a probability density subject to moment constraints is included in this frame by setting

I. Csiszár is with the A. Rényi Institute of Mathematics, Hungarian Academy of Sciences H-1364 Budapest, P.O.B. 127. e-mail: csiszar@math-inst.hu, partially supported by the Hungarian Academy Research Grant 96/2-678 2,1, F. Gamboa is Ecole Polytechnique, C.M.A.P., F-91128 Palaiseau Cedex and Université de Picardie Jules Verne Faculté de Mathématiques et d'Informatique. e-mail: Fabrice.Gamboa@math.u-psud.fr, E. Gassiat is with Laboratoire de Statistiques, Université Paris Sud F-91405 Orsay. email: Elisabeth.Gassiat@math.u-psud.fr

$\varphi_1(x) = 1$, letting each $c \in \mathcal{K}$ have first component $c_1 = 1$, and requiring non-negativity as prior information about g.

Typically, g is not determined uniquely by the available information, and one is lead to the mathematically not well-defined problem of selecting a "good" function g from the set of possible candidates. One possible theoretical approach is to postulate certain properties that a "logically consistent" method of selection ought to have, and to determine whether those postulates characterize a unique method. Shore and Johnson [38] showed that for reconstructing a probability density g subject to constraints as in (I.1), with $\mathcal{K} = \{c\}$, the only method satisfying some natural postulates is to select that g satisfying the constraints whose Shannon entropy is maximum, or whose I-divergence (also called cross entropy or Kullback-Leibler distance) from a default model g_0 is minimum. If $g_0 = \text{const}$ then minimizing I-divergence from g_0 is the same as maximizing entropy, and even if $g_0 \neq \text{const}$, this method will be referred to as maximum entropy (ME) in the sequel. Of course, ME has been a popular method prior to [38], primarily due to the works of Kullback and Jaynes cf. [28], [24]. It is often used also for reconstructing positive valued functions not necessarily probability densities, and the axiomatic approach of [38] was shown to cover that case by Skilling [39]. For reconstructing arbitrary real-valued functions, the time-honored method is least squares, i.e., to select that g satisfying the constraints whose L_2 norm, or L_2-distance from a default model g_0, is minimum.

A systematic axiomatic study of "logically consistent" reconstruction methods has been carried out by Csiszár [12] (restricting attention to functions defined on a finite set). It confirmed that ME is uniquely singled out by at least two different sets of intuitively appealing axioms, when the function to be reconstructed is positive valued, and so is least squares in the absense of the latter condition. Further, somewhat weaker axioms were shown to permit alternative methods of minimizing some other distance from the default model than I-divergence or L_2-distance. Specifically, two families of distances generated by convex functions f were distinguished: f-divergences (Csiszár [8], [9], Ali and Silvey [2]), and Bregman distances (Bregman [5], cf also Jones and Byrne [25] as "projective distances"). An extension of the ME method, called maximum entropy on the mean (MEM) has been developed by Dacunha-Castelle, Gamboa, and Gassiat [15], [21], [23], motivated by the work of Navaza [35], see also [29] and for applications [41], [34]. MEM sheds new light on the reconstruction problem; in particular, it permits a stochastic interpretation of the method of f-divergence minimization, for a large class of convex functions f. We note that Bregman distance minimization has also been arrived at via MEM (in the discrete case, see [29], [14]) but this will not be treated in this paper.

We will consider MEM-based approaches to the immediate extensions of the moment problem (I.1) stated as Problems (i) and (ii) below. These problems will be notationally distinguished by denoting the underlying set and its generic element by U and x, respectively by V and t.

In the sequel, the term measure will always mean a finite real-valued mesure (signed measure) unless stated otherwise. The set of all measures on a set V (equipped with a σ-algebra, not indicated in the notation) will be denoted by $\mathcal{M}(V)$. An $\mathbb{R}^k$-valued measure is a k-tuple of measures.

Problem (i): Find a real-valued function g with domain U when knowing that

$$\int_U g(x)d\Phi(x) \in \mathcal{K} \tag{I.2}$$

for a given $\mathbb{R}^k$-valued measure Φ on U, and a closed convex set $\mathcal{K} \subset \mathbb{R}^k$.

The integral in (I.2) means $(\int g(x)d\Phi_i(x))_{i=1}^k$ if $\Phi = (\Phi_1, \ldots, \Phi_k)$.

Problem (ii): Find a measure $G \in \mathcal{M}(V)$ when knowing that

$$\int_V \varphi(t)dG(t) \in \mathcal{K} \tag{I.3}$$

for a given $\mathbb{R}^k$-valued function φ on V, and a closed convex set $\mathcal{K} \subset \mathbb{R}^k$.

Here g resp. G may be subject to additional requirements representing prior information.

Notice that (I.2) reduces to (I.1) if $\Phi = \varphi P_U$; this notation means that each component Φ_i of Φ is absolutely continuous with respect to P_U with Radon-Nikodym derivative φ_i. (I.3) also reduces to (I.1) if $V = U$ and G is required to be absolutely continuous with respect to P_V. Problem (i) with $\Phi = (\delta_{x_1}, \ldots \delta_{x_k})$, where δ_x denotes the point mass at x, reduces to the *interpolation problem* of finding g satisfying $(g(x_1), \ldots, g(x_k)) \in \mathcal{K}$, whereas $\Phi = (\delta_{x_1}, \ldots, \delta_{x_\ell}, \varphi_{\ell+1} P_U, \ldots, \varphi_k P_U)$ yields an *interpolation problem coupled with a moment problem.* Of course, interpolation problems are meaningful only under additional requirements on g such as smoothness.

The key idea of MEM is that the function g or measure G to be reconstructed is visualized as the expectation of a random function or measure with respect to an unknown p.m. Q on the space of functions or measures. This Q is constrained by the requirement that the expectation of the random function or measure should satisfy the given constraints on g or G. Thus the problem is reduced to that of reconstructing Q, and this is done via ME, by minimizing I-divergence from a default model p.m. Q_0. The expectation of the random function or measure with respect to the so obtained Q is declared to be the MEM solution of the original problem. Typically, this idea does not work directly, since there is no natural default model p.m. Q_0 on the space of functions or measures, unless the underlying set is finite or countable. In particular, only in the latter case can one take Q_0 such that the "random amplitudes" (the values of the random function or measure at distinct points) be i.i.d. random variables under Q_0. For this reason, the MEM idea has been developed in [15], [21], [23] via discrete approximations.

MEM solutions often permit an alternate interpretation that requires an approximation process even if the underlying set is finite. Namely, an MEM solution to

Problem (ii) is the limit in probability of a sequence of random measures ν_n under the probability assignment obtained from the default model by conditioning on the event that $\int \varphi d\nu_n \in \mathcal{K}$.

In Section II we will review known results and include one new result about minimizing an f-divergence-like convex functional, and discuss how, for a large class of such functionals, this method arises via MEM and admits a stochastic interpretation.

In Section III we propose a new family of MEM-based solutions obtained, in effect, with amplitudes correlated in a specific way under the default model (rather then i.i.d.). The idea is that if G satisfies (I.3) with

$$\varphi(t) = \int_U K(x,t)d\Phi(x) \tag{I.4}$$

then

$$g(x) = \int_V K(x,t)dG(t) \tag{I.5}$$

satisfies (I.2). Previous works using this idea in the context of ME include [39] and [6]. We will explain how, by suitable choice of the kernel K, this allows to use prior knowledge, such as smoothness, unimodality or convexity. We also give criteria for the existence of functions g satisfying the moment constraints (I.2) or (I.1) as well as specific prior information.

In Section IV we treat MEM with amplitudes correlated differently from Section III, viz. exchangeable ones. The rigorous results we obtain via advanced large deviations techniques are in accord with conclusions of a previous heuristic treatment of the exchangeable case in the context of ME [11].

For the reader's orientation, we send forward consequences of Theorem II.4 (i.i.d. amplitudes) and Theorem IV.1 (exchangeable amplitudes) for the discrete case $V = \{1,\ldots,m\}$, when functions and measures reduce to m-dimensional vectors and $\mathbb{R}^k$-valued measures reduce to $k \times m$ matrices.

Thus, to find a vector $\boldsymbol{g} = (g_1,\ldots,g_m)^T$ subject to the constraint that, for a given matrix $\boldsymbol{\Phi}$ of rank k, $\boldsymbol{\Phi g}$ belongs to a given (closed, convex) $\mathcal{K} \subset \mathbb{R}^k$, consider a sequence of random vectors of form

$$\nu_n = \left(\frac{1}{n} \sum_{j=1}^{\ell(i,n)} Y_{ij} \right)_{i=1}^{m} \tag{I.6}$$

where the Y_{ij} are (real-valued) random variables and the numbers $\ell(i,n)$ satisfy $\ell(1,n) + \ldots + \ell(m,n) = n$ and

$$\lim_{n\to\infty} \frac{1}{n}\ell(i,n) = p_i > 0, \qquad i = 1,\ldots,m. \tag{I.7}$$

Then, subject to some technical conditions, the following holds: If the Y_{ij}'s are i.i.d. with distribution F, the random vectors (I.6) conditioned on the event $\boldsymbol{\Phi}\nu_n \in \mathcal{K}$

converge in probability to the unique minimizer of the divergence

$$D_\gamma(\boldsymbol{g},\boldsymbol{p}) = \sum_{i=1}^{m} p_i \gamma\left(\frac{g_i}{p_i}\right) \tag{I.8}$$

subject to $\boldsymbol{\Phi g} \in \mathcal{K}$, where $\gamma = \gamma_F$ is a strictly convex function determined by F. If the Y_{ij}'s are only exchangeable, i.e., their joint distributions are mixtures of Cartesian powers of distributions F_ϑ, $\vartheta \in \Theta$, the role of $D_\gamma(\boldsymbol{g},\boldsymbol{p})$ in (I.8) is played by

$$\mathcal{I}(\boldsymbol{g}) = \min_{\vartheta} D_{\gamma_{F_\vartheta}}(\boldsymbol{g},\boldsymbol{p}). \tag{I.9}$$

However, the (perhaps non-convex) function $\mathcal{I}(\boldsymbol{g})$ need not have a unique minimizer subject to $\boldsymbol{\Phi g} \in \mathcal{K}$. Then the assertion replacing convergence in probability to the minimizer is that the conditional distribution of ν_n on the condition $\boldsymbol{\Phi}\nu_n \in \mathcal{K}$ accumulates on the set M of minimizers, i.e., the sequence of these conditional distributions is relatively compact and its accumulation points are p.m.'s concentrated on M.

II. Classical convex methods

In this Section, we review the method of solving problems as stated in Section I by minimizing certain convex functionals, and discuss the stochastic interpretation of this method via MEM.

For the concepts and simple facts about convex functions used below see e.g. [37].

Throughout this paper, γ will denote a closed convex function $\gamma\colon \mathbb{R} \to [0,+\infty]$ such that $\mathrm{dom}\gamma = \{y\colon \gamma(y) < +\infty\}$ is an interval with endpoints $a_\gamma < b_\gamma$. The closedness of γ means that if a_γ or b_γ is finite then $\gamma(y)$ tends to $\gamma(a_\gamma)$ or $\gamma(b_\gamma)$ when y approaches a_γ from the right or b_γ from the left, respectively.

The convex conjugate (or Legendre-Fenchel transform) of γ will be denoted by ψ, i.e.,

$$\psi(\tau) = \sup_{y\in\mathbb{R}} (y\tau - \gamma(y)). \tag{II.1}$$

Then ψ is a closed convex function, the endpoints of $\mathrm{dom}\psi = \{\tau\colon \psi(\tau) < +\infty\}$ are

$$a_\psi = \lim_{y\to-\infty} \frac{\gamma(y)}{y}, \qquad b_\psi = \lim_{y\to\infty} \frac{\gamma(y)}{y}, \tag{II.2}$$

and the convex conjugate of ψ equals γ.

A. *Deterministic formulation*

Let $\gamma\colon \mathbb{R} \to [0,+\infty]$ be a closed convex function such that

- (i) γ is differentiable and strictly convex in (a_γ, b_γ), the interior of $\mathrm{dom}\gamma$
- (ii) the minimum of $\gamma(y)$ is 0, attained at some $y_0 \in (a_\gamma, b_\gamma)$.

The strict convexity of γ in (a_γ, b_γ) is equivalent to the condition that the convex conjugate ψ of γ is *essentially smooth*, i.e., differentiable in (a_ψ, b_ψ) with

$$\begin{aligned} \lim_{\tau \to a_\psi + 0} \psi'(\tau) &= -\infty \quad \text{if} \quad a_\psi > -\infty, \\ \lim_{\tau \to b_\psi - 0} \psi'(\tau) &= +\infty \quad \text{if} \quad b_\psi < +\infty. \end{aligned} \tag{II.3}$$

Conversely, γ is essentially smooth iff ψ is strictly convex in (a_ψ, b_ψ). Note that $a_\psi < 0 < b_\psi$.

One approach to Problem (i) or (ii) is to adopt as a solution the minimizer g^* of

$$I_\gamma(g) = \int \gamma(g) dP_U \tag{II.4}$$

subject to (I.2), or the minimizer G^* of

$$D_\gamma(G, P_V) = \int \gamma(f_G) dP_V + b_\psi \sigma_G^+(V) - a_\psi \sigma_G^-(V) \tag{II.5}$$

subject to (I.3), providing a (unique) minimizer exists. Here

$$G = f_G P_V + \sigma_G, \qquad \sigma_G = \sigma_G^+ - \sigma_G^- \tag{II.6}$$

is the Lebesgue decomposition of G, and the Jordan decomposition of the singular part σ_G, respectively. Note that $D_\gamma(\cdot, P_V)$ defined by (II.5) is the unique convex and lower semicontinuous extension to $\mathcal{M}(V)$ of the functional $I_\gamma\left(\frac{dG}{dP_V}\right)$ defined for $G \ll P_V$, at least if V is a Polish space; lower semicontinuity is meant for the weak topology of measures.

The functional (II.4) is called γ-entropy in [3], [40]. It is, however, $-I_\gamma(g)$ that (for $V = [0,1]$, P_V=Lebesgue measure) gives Shannon differential entropy or Burg entropy if

$$\gamma(y) = y \log y - y + 1 \tag{II.7}$$

or

$$\gamma(y) = -\log y + y - 1, \tag{II.8}$$

the favored choices when a positive solution g is sought. Here we understand $\log y = -\infty$ if $y \leq 0$. In absence of a positivity requirement, the familiar choice is

$$\gamma(y) = y^2. \tag{II.9}$$

The functional (II.5) was first introduced in [8] as "f-divergence" (where f was the present γ), see also [9], [2], in the special case when G is a p.m. and $a_\gamma = 0$, $b_\gamma = \infty$, generalizing the concept of I-divergence that arises with the choice (II.7). For a systematic treatment of these divergences see [33]. In this paper, we will use the term γ-divergence for the functional (II.5) in the present generality.

Since

$$D_\gamma(G, P_V) \geq 0, \quad \text{equality iff} \quad G = y_0 P_V \tag{II.10}$$

(recall that y_0 denotes the unique point where $\gamma(y)$ attains its minimum 0), $D_\gamma(G, P_V)$ is a measure of distance of G from $y_0 P_V$, and in particular from P_V if $y_0 = 1$, as in the cases (II.7), (II.8). The γ-divergence (II.5) plays a key role in [4], for measures on a compact set V. As shown there, in that case

$$D_\gamma(G, P_V) = \sup_{f \in C(V)} [\int f dG - \int \psi(f) dP_V] \tag{II.11}$$

where $C(V)$ is the set of continuous functions on V. In other words, $D_\gamma(\cdot, P_V)$ is the convex conjugate of the convex functional $I_\psi(f) = \int \psi(f) dP_V$ on $C(V)$.

In this subsection, we present results about the existence and uniqueness of the minimizer of the γ-entropy (II.4) subject to (I.1) and of the γ-divergence (II.5) subject to (I.3), as well as about the algebraic form of the minimizer. Note that these minimization problems (with $U = V$) are identical if $a_\psi = -\infty$, $b_\psi = \infty$, e.g., for γ in (II.7) or (II.9), since then $D_\gamma(G, P_V) = \infty$ whenever $G \not\ll P_V$. Otherwise, as for γ in (II.8), it may happen even in "regular" cases that the first minimum is not attained while the second one is attained for some $G^* \not\ll P_V$, see [4].

The following Theorem was proved in [13], Theorem 1, for the case $a_\gamma = 0$, $b_\gamma = +\infty$; the same proof works in general. The Bregman (γ) distance of functions g and g^* on U such that $\gamma(g)$ and $\gamma(g^*)$ are finite P_U – a.s., is defined as

$$B_\gamma(g, g^*) = \int [\gamma(g) - \gamma(g^*) - \gamma'(g^*)(g - g^*)] dP_U. \tag{II.12}$$

Notice that the bracketed expression is strictly positive when $g \neq g^*$.

Theorem II.1 ([13]) : Given any convex set E of measurable functions on U, such that $I_\gamma(E) = \inf_{g \in E} I_\gamma(g)$ is finite, there exists g^* not necessarily in E such that

$$I_\gamma(g) \geq I_\gamma(E) + B_\gamma(g, g^*) \quad \text{if} \quad g \in E,\ I_\gamma(g) < +\infty. \tag{II.13}$$

This g^* is unique P_U-a.s. and any sequence of functions $g_n \in E$ with $I_\gamma(g_n) \to I_\gamma(E)$ converges to g^* in P_U measure.

Clearly, the minimum of $I_\gamma(g)$ subject to $g \in E$ is attained iff $g^* \in E$. If the given γ is such that $\alpha_\psi = -\infty$, $b_\psi = +\infty$, cf. (II.2), then convergence in measure of functions g_n with $\sup_n I_\gamma(g_n) < +\infty$ implies convergence in $L_1(P_U)$ norm. Thus in this case, the $L_1(P_U)$-closedness of E is a sufficient condition for the attainment of $\min_{g \in E} I_\gamma(g)$. In particular, a sufficient condition for the attainment of $\min I_\gamma(g)$ subject to (I.1) is the boundedness of φ. A less restrictive sufficient condition (still under the assumption

$a_\psi = -\infty$, $b_\psi = +\infty$) has been given in [13] via Orlicz space techniques; cf. also [30].

The following Theorem characterizes $I_\gamma(E) = \inf_{g \in E} I_\gamma(g)$ when E is the set of those (measurable) functions on U that satisfy (I.1). In this case, the function g^* of Theorem II.1 is also characterized, and a general sufficient condition for $g^* \in E$, i.e. the attainment of $\min_{g \in E} I_\gamma(g)$ is given. This Theorem is new in this generality, though the characterization of $I_\gamma(E)$ via a finite dimensional maximization, the dual problem of minimizing $I_\gamma(g)$ in the sense of convex programming, cf. [3], [4], has been known for less general situations, together with the fact that under some additonal conditions, the "dual optimal solution" provides the minimizer g^* of $I_\gamma(g)$. The observation that it provides g^* of Theorem II.1 even if $g^* \notin E$ (when $\min I_\gamma(g)$ is not attained) appears new, except for the case of I-divergence, see [10].

Given an $\mathbb{R}^k$-valued function $\varphi = (\varphi_1, \ldots, \varphi_k)$ on U where $\varphi_i \in L_1(P_U)$, $i = 1, \ldots, k$, respectively an $\mathbb{R}^k$-valued measure Φ, and a closed convex set $\mathcal{K} \subset \mathbb{R}^k$, write

$$E_{\varphi,\mathcal{K}} = \left\{ g : \int g \varphi dP_U \in \mathcal{K} \right\}, \quad E_{\Phi,\mathcal{K}} = \left\{ g : \int g d\Phi \in \mathcal{K} \right\}. \tag{II.14}$$

Theorem II.2: Supposing the "constraint qualification"

$$a_\gamma < g(x) < b_\gamma \qquad P_U\text{-a.s. for some } g \in E_{\varphi,\mathcal{K}} \tag{II.15}$$

holds, we have

$$I_\gamma(E_{\varphi,\mathcal{K}}) = \max_{v \in \mathbb{R}^k} \left[\inf_{c \in \mathcal{K} \cap L} \langle v, c \rangle - l(v) \right], \tag{II.16}$$

where L is the subspace of $\mathbb{R}^k$ consisting of the vectors representable as $\int g \varphi dP_U$, and

$$l(v) = \int \psi(\langle v, \varphi \rangle) dP_U. \tag{II.17}$$

For v^* attaining the maximum in (II.16),

$$g^* = \psi'(\langle v^*, \varphi \rangle) \tag{II.18}$$

satisfies (II.13) with $E = E_{\varphi,\mathcal{K}}$, and if this v^* is an interior point of

$$\operatorname{dom} l = \{v \colon \int \psi(\langle v, \varphi \rangle) dP_U < +\infty\}$$

then g^* defined by (II.18) belongs to $E_{\varphi,\mathcal{K}}$. In the latter case g^* attains $\min I_\gamma(g)$ subject to $g \in E_{\varphi,\mathcal{K}}$.

Theorem II.2 will be proved in Appendix A, using ideas of [3] and [4] with some modifications. We will give a self-contained proof not depending on deeper results from convex analysis than the Fenchel duality theorem.

Let now $\Phi = \varphi P_U + \Sigma$ where Σ is an $\mathbb{R}^k$-valued measure concentrated on a set U_0 with $P_U(U_0) = 0$. Denote by N the subspace of $\mathbb{R}^k$ consisting of the vectors representable as $\int g d\Sigma$. Then

$$I_\gamma(E_{\Phi,\mathcal{K}}) = I_\gamma(E_{\varphi,\mathcal{K}+N})$$

since g can be arbitrarily changed on the set U_0 without changing $I_\gamma(g)$, making $\int g d\Sigma$ be equal to any element of N. Thus

Corollary II.3: Theorem II.2 holds replacing everywhere $E_{\varphi,\mathcal{K}}$ by $E_{\Phi,\mathcal{K}}$, and $\mathcal{K}$ by $\mathcal{K} + N$ on the right hand side of eq. (II.16). The last assertions hold, however, not necessarily for g^* but for some g^{**} that may differ from g^* on U_0.

Corollary II.3 makes more precise the intuitive idea that straightforward γ-entropy minimization is not a good method for "solving" Problem (i) with $\Phi \not\ll P_V$, e.g., for solving interpolation problems coupled with moment problems. Methods treated in Section III are more suitable for that purpose.

The following Theorem of Borwein and Lewis [4] concerns problem (ii). It gives a sufficient condition for the existence of a $G^* \in \mathcal{M}(V)$ that minimizes $D_\gamma(G, P_V)$ subject to the constraint (I.3) and -under that condition- characterizes G^*. Write

$$S_{\varphi,\mathcal{K}} = \left\{ G : \int \varphi dG \in \mathcal{K} \right\}. \tag{II.19}$$

Assume that

$(H1)$ V is a compact metric space, P_V is a p.m. having full support, $\varphi : V \to \mathbb{R}^k$ is continuous, its components φ_i are linearly independent.

Here, *full support* means that every open subset of V has positive P_V measure. Notice that Hypothesis $(H1)$ implies that each vector in $\mathbb{R}^k$ is representable as $\int g\varphi dP_V$. In the following Theorem, reference to equations (II.16)-(II.18) is understood with U replaced by V, and with $L = \mathbb{R}^k$.

Theorem II.4 ([4]) : Under Hypothesis $(H1)$, supposing the "constraint qualification"

$$\frac{dG}{dP_V} \in (a_\gamma, b_\gamma) \qquad P_V\text{-a.s. for some } G \ll P_V \text{ in } S_{\varphi,\mathcal{K}} \tag{II.20}$$

holds, the minimum of $D_\gamma(G, P_V)$ subject to $G \in S_{\varphi,\mathcal{K}}$ is attained, and equals (II.16). Moreover, each minimizer G^* is of form $G^* = g^* P_V + \sigma_{G^*}$ with g^* given by (II.18), and $\sigma_{G^*} \in \mathcal{M}(V)$ singular with respect to P_V.

Comparison of Theorems II.2 and II.4 (with $U = V$) shows that under the assumptions of Theorem II.4, the infimum of $I_\gamma(g)$ subject to $\int g\varphi dP_V \in \mathcal{K}$ equals the minimum of $D_\gamma(G, P_V)$ subject to $\int \varphi dG \in \mathcal{K}$, and the density of the absolutely continuous part of G^* attaining that minimum equals the function g^* appearing in Theorem II.2. Actually, Theorem II.4 could be obtained as a corollary of Theorem II.2, via the argument in [13], Theorem 4.

Remarks:

i) If $a_\psi = -\infty$, $b_\psi = +\infty$ then σ_{G^*} in Theorem II.4 necessarily vanishes by (II.5). More generally, the same holds if the maximum in (II.16) is attained for v^* satisfying $a_\psi < \langle v^*, \varphi(t)\rangle < b_\psi$ for all $t \in V$. Indeed, then (using Hypothesis (H1)) v^* is in the interior of $\{v : \int \psi(\langle v, \varphi\rangle)dP_V < +\infty\}$, and hence $g^* P_V \in S_{\varphi,\mathcal{K}}$ by Theorem II.2. In these cases, G^* is unique.

ii) If $(H1)$ holds and $\mathrm{Int}\mathcal{K} \neq \emptyset$ where $\mathrm{Int}\mathcal{K}$ denotes the interior of $\mathcal{K}$, the constraint qualification (II.20) is equivalent to

$$\int \varphi dG \in \mathrm{Int}\mathcal{K} \quad \text{for some} \quad G \in \mathcal{M}(V) \quad \text{with} \quad D_\gamma(G, P_V) < +\infty. \tag{II.21}$$

This will be shown in Appendix A.

iii) Strictly speaking, Theorem II.4 was established in [4] for the case $\mathcal{K} = \{c\}$. It was also shown in [4], under Hypothesis $(H1)$, that

$$\tilde{f}(c) = \inf\{D_\gamma(G, P_V) : \int \varphi dG = c\} = \sup_{v \in \mathbb{R}^k} [\langle v, c\rangle - l(v)] \tag{II.22}$$

holds for all $c \in \mathbb{R}^k$, not only for those with $\mathcal{K} = \{c\}$ satisfying (II.20), where $l(v)$ is defined by (II.17), with U replaced by V. Thus, $\tilde{f}(c)$ is the convex conjugate of $l(v)$ which -by the proof of Theorem II.2 in Appendix A- is the convex conjugate of $f(c) = \inf\{I_\gamma(g) : \int g\varphi dP_V = c\}$. Hence, (II.22) means that $\tilde{f}(c)$ equals the closure of the convex function $f(c)$, cf. [37] p. 104. Of course, the latter can be easily checked directly.

B. Stochastic interpretation

In [15], [21] and [23], following the work of Navaza [35], a new method called maximum entropy on the mean (MEM) has been developed. It has lead to the same solution of Problem (ii) as given by Theorem II.4, that has been proven simultaneously and independently [23], for convex functions γ_F as below. MEM admits a stochastic interpretation of this solution.

We briefly recall the main ideas of MEM; for details, cf. [15], [21], [23]. Here, Hypothesis $(H1)$ will be assumed. Convergence and other topological concepts for measures on a Polish space are always understood in the weak topology. Note that V and $\mathcal{M}(V)$ are Polish spaces.

First, the given p.m. P_V on V is represented as the limit of a sequence of discrete p.m.'s

$$P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{t_i} \tag{II.23}$$

where $(t_i)_{i\in\mathbb{N}}$ is a suitable (deterministic) sequence of points of V.

With each t_i a real valued random variable Y_i is associated (random amplitude at pixel t_i), and for each n a random measure $\nu_n \ll P_n$ is defined on V by

$$\nu_n = \frac{1}{n}\sum_{i=1}^{n} Y_i \delta_{t_i}. \tag{II.24}$$

The joint distribution of $Y_1, \ldots, Y_n$ is assigned using the ME method. Let the default model be that $Y_1, \ldots, Y_n$ are i.i.d. with common distribution F, a non-degenerate p.m. on $\mathbb{R}$ whose choice represents prior knowledge. Then the joint distribution Q_n^{MEM} assigned to $Y_1, \ldots, Y_n$ will be that p.m. Q on $\mathbb{R}^n$ whose I-divergence (see example 1) after Theorem II.5) from the default model joint distribution $Q_0 = F^{\otimes n}$ is minimum subject to the constraint that the measure $G_n = E_Q(\nu_n)$ satisfies (I.3).

In other words, let

$$Q_n^{MEM} = \arg\min_{Q \in \mathcal{R}_n} D(Q \| F^{\otimes n}) \tag{II.25}$$

where $\mathcal{R}_n$ is the following set of p.m.'s on $\mathbb{R}^n$:

$$\mathcal{R}_n = \left\{ Q\colon E_Q\left(\frac{1}{n}\sum_{i=1}^{n} \varphi(t_i) Y_i\right) \in \mathcal{K}\right\}. \tag{II.26}$$

The n'th approximation G_n^{MEM} of the MEM solution to Problem (ii) is defined as the expectation of ν_n in (II.24) when the joint distribution of $Y_1, \ldots, Y_n$ is Q_n^{MEM}, i.e.,

$$G_n^{MEM} = E_{Q_n^{MEM}}(\nu_n) = \frac{1}{n}\sum_{i=1}^{n} E_{Q_n^{MEM}}(Y_i)\delta_{t_i}. \tag{II.27}$$

It should be noted that the Q_n^{MEM} do not form, in general, a consistent family of joint distributions defining a process. Nevertheless, the discrete measures $G_n^{MEM} \ll P_n$ often converge to a unique limit G^*, or else the set of possible accumulation points of the sequence $\{G_n^{MEM}\}$ can be described, see Theorem II.5 below. In addition to G_n^{MEM} we will also consider the Bayesian estimate

$$G_n^{Bay} = E_{F^{\otimes n}}\left[\ \nu_n \ \middle|\ \int_V \varphi d\nu_n \in \mathcal{K}\right] \tag{II.28}$$

defined if the event $\int \varphi d\nu_n \in \mathcal{K}$, i.e. $\frac{1}{n}\sum_{i=1}^{n} \varphi(t_i) Y_i \in \mathcal{K}$, has positive probability under $F^{\otimes n}$. If F is a continuous distribution, the latter requires Int $\mathcal{K} \neq \emptyset$.

Let $a_F < b_F$ denote the endpoints of the smallest interval with F-measure 1. We need the following technical assumption:

$(H2)$ The logarithm of the moment generating function of F, viz.

$$\psi_F(\tau) = \log \int \exp(\tau y) dF(y), \tag{II.29}$$

is finite in an interval with endpoints $a_\psi < 0 < b_\psi$, and ψ_F satisfies (II.3).

$(H2)$ always holds if a_F and b_F are finite, since then $a_\psi = -\infty$, $b_\psi = +\infty$. The function ψ_F is a closed convex function, strictly convex and differentiable in the interval (a_ψ, b_ψ), and by Hypothesis $(H2)$ it is essentially smooth. Note that

$$\psi_F'(\tau) = \int y \exp[\tau y - \psi_F(\tau)] dF(y) \tag{II.30}$$

equals the expectation of the "tilted distribution" $\exp(\tau y - \psi_F(\tau))F$. Provided that the given (closed, convex) set $\mathcal{K} \subset \mathbb{R}^k$ contains some point of form $\frac{1}{n}\sum_{i=1}^n \varphi(t_i) y_i$, $y_i \in (a_F, b_F)$, $i = 1, \ldots, n$, by standard theory of the ME method the (unique) minimizer Q_n^{MEM} of $D(Q \| F^{\otimes n})$ in (II.25) exists, and it belongs to the exponential family through $F^{\otimes n}$ spanned by the k-dimensional statistic $\frac{1}{n}\sum_{i=1}^n \varphi(t_i) y_i$. Thus for some $v_n \in \mathbb{R}^k$,

$$Q_n^{MEM} = \exp\left(\sum_{i=1}^n [\tau_i y_i - \psi_F(\tau_i)]\right) F^{\otimes n}, \quad \tau_i = \langle v_n, \varphi(t_i)\rangle, \tag{II.31}$$

i.e., the Y_i's are independent under Q_n^{MEM}, each with a tilted distribution as above. Using (II.30), it follows from (II.27), (II.31) that

$$G_n^{MEM} = \frac{1}{n}\sum_{i=1}^n \psi_F'(\langle v_n, \varphi(t_i)\rangle)\delta_{t_i}. \tag{II.32}$$

The parameter $v_n \in \mathbb{R}^k$ can be determined as the maximizer of the function

$$H_n(v) = \inf_{c \in \mathcal{K}} \langle v, c\rangle - \frac{1}{n}\sum_{i=1}^n \psi_F(\langle v, \varphi(t_i)\rangle). \tag{II.33}$$

Let γ_F be the convex conjugate of ψ_F, cf. (II.1). Then the endpoints a_{γ_F} and b_{γ_F} of $\mathrm{dom}\gamma_F$ are equal to a_F and b_F. Further, γ_F is essentially smooth (since ψ_F is strictly convex in the interior of $\mathrm{dom}\psi_F$) and is strictly convex in the interior (a_F, b_F) of $\mathrm{dom}\gamma_F$ (since ψ_F is essentially smooth). Thus γ_F satisfies assumption (i) in Section II A. Further, $\gamma_F(y)$ attains its minimum 0 at $y_0 = \int y dF(y)$, hence γ_F satisfies assumption (ii). The following result has been proven in [23], for the case $a_F = 0$, $b_F = +\infty$, under the technical assumptions that $\varphi_1(x) = 1$ and $\mathcal{K}$ is compact; for the minor modifications in the proof needed in the present generality, see Appendix B.

We will use the following terminology: A sequence of elements z_n of a Polish space E (below, $E = \mathcal{M}(V)$) *accumulates in a set* $F \subset E$ if (z_n) is relatively compact and its accumulation points are in F. A sequence of p.m.'s μ_n on E *accumulates on a set* $F \subset E$ if (μ_n) is relatively compact and its accumulation points are p.m.'s on F. In the special case $F = \{z_0\}$, these concepts reduce to $z_n \to z_0$, respectively to convergence in probability to z_0 of E-valued random variables with distributions μ_n.

Theorem II.5: Assume $(H1)$, $(H2)$, and the constraint qualification that for some $G \ll P_V$ with $\int \varphi dG \in \mathcal{K}$,

$$\frac{dG}{dP_V} \in (a_F, b_F) \qquad P_V \text{ - a.s.} \tag{II.34}$$

Then the sequence (G_n^{MEM}) accumulates on the set of minimizers of $D_{\gamma_F}(G, P_V)$ subject to $\int \varphi dG \in \mathcal{K}$. In case Int$\mathcal{K} \neq \emptyset$ the same holds also for (G_n^{Bay}), moreover, the conditional distribution of ν_n in (II.24) on the condition $\int \varphi d\nu_n \in \mathcal{K}$ (under the i.i.d. default model) accumulates on the set of minimizers as above.

The result of Theorem II.5 is most interesting when $D_{\gamma_F}(G, P_V)$ has a unique minimizer G^* subject to $\int \varphi dG \in \mathcal{K}$. Then G_n^{MEM} and G_n^{Bay} converge to G^*, and ν_n converges to G^* in conditional probability on the condition $\int \varphi d\nu_n \in \mathcal{K}$. For sufficient conditions of uniqueness of G^* see Remark i) after Theorem II.4. For the representation of a minimizer (unique or not) see Theorem II.4.

Let us return to the problem of selecting a good functional for the reconstruction method. The first simple idea is that the choice of domγ or dom ψ allows to take into account the following possible prior informations:

- **Positivity.** If dom $\gamma \subset [0, +\infty[$, then the solution of problem (ii) lies in the set of positive measures on V.
- **Absolute continuity.** If $a_\psi = -\infty$ and $b_\psi = +\infty$, the minimization leads to a solution with no singular part σ_{G^*}.
- **Bounds on the density.** If dom $\gamma \subset [0, L]$, $L > 0$, the minimization leads to measures lying in

$$\mathcal{B}_L(P_V) := \left\{ G \ : \ G \ll P_V,\ 0 \leq \frac{dG}{dP_V} \leq L \right\}. \tag{II.35}$$

Theorem II.5 gives a point of view to choose between different functionals for solving Problem (ii). In particular, it gives a Bayesian criterion for choosing between γ's having the same domain. It may be of interest to recall examples of functionals that may be minimized this way and to which default model F for the amplitudes they are related. Below we understand that $G = f_G P_V + \sigma_G$ is the Lebesgue decomposition of a measure $G \in \mathcal{M}(V)$ with respect to P_V.

1) <u>I-divergence:</u> Let F be the Poisson distribution with mean ϑ. Then, $(a_F, b_F) = (0, +\infty)$, $\psi_F(\tau) = \log \sum_{y=0}^{\infty} e^{-\vartheta} \frac{\vartheta^y}{y!} e^{\tau y} = \vartheta(e^\tau - 1)$, and $\gamma_F(y) = y \log y - (1 + \log \vartheta) y + \vartheta$ for $y \in [0, +\infty) = \text{dom}\gamma_F$, yielding

$$D_{\gamma_F}(G, P_V) = \begin{cases} \displaystyle\int f_G \log f_G dP_V - (1 + \log \vartheta) G(V) + \vartheta \text{ if } f_G \geq 0,\ \sigma_G = 0 \\ +\infty \quad \text{otherwise} \end{cases}$$

With $\vartheta = 1$, γ_F will be the function (II.7) and D_{γ_F} becomes the I-divergence $D(G \| P_V)$.

2) Burg Entropy: Let F be the exponential distribution with mean ϑ. Then, $(a_F, b_F) = (0, +\infty)$, $\psi_F(\tau) = \log \int\limits_0^\infty \frac{1}{\vartheta} e^{-\frac{y}{\vartheta}} e^{\tau y} dy = -\log(1-\vartheta\tau)$, $\tau < \frac{1}{\vartheta}$, and $\gamma_F(y) = \log\frac{\vartheta}{y} + \frac{y}{\vartheta} - 1$ for $y \in (0, +\infty) = \mathrm{dom}\gamma_F$, yielding

$$D_{\gamma_F}(G, P_V) = \begin{cases} -\int \log f_G dP_V + \dfrac{G(V)}{\vartheta} + \log\vartheta - 1 \text{ if } f_G \geq 0, \sigma_G \geq 0 \\ +\infty \quad \text{otherwise} \end{cases}$$

With $\vartheta = 1$, D_{γ_F} will be the Burg entropy [7] of G, or equivalently, the reversed I-divergence $D(P_V \| G)$.

3) Fermi-Dirac Entropy: Let F be the two points distribution $(1-\vartheta)\delta_0 + \vartheta\delta_1$, $0 < \vartheta < 1$. Then, $(a_F, b_F) = (0, 1)$, $\psi_F(\tau) = \log(1 - \vartheta + \vartheta e^\tau)$, and $\gamma_F(y) = y\log\frac{y}{\vartheta} + (1-y)\log\frac{1-y}{1-\vartheta}$ for $y \in [0, 1] = \mathrm{dom}\gamma_F$, yielding

$$D_{\gamma_F}(G, P_V) = \begin{cases} \int \left[f_G \log\dfrac{f_G}{\vartheta} + (1-f_G)\log\dfrac{1-f_G}{1-\vartheta} \right] dP_V \text{ if } 0 \leq f_G \leq 1, \sigma_G = 0 \\ +\infty \quad \text{otherwise} \end{cases}$$

With $\vartheta = \frac{1}{2}$, D_{γ_F} will be the Fermi-Dirac entropy, see [26], p.164. This comes into account when the shape information is $0 \leq f_G \leq 1$.

4) Energy: Let F be the Gaussian distribution with mean ϑ and variance 1. Then, $(a_F, b_F) = (-\infty, +\infty)$, $\psi_F(\tau) = \log \int \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\vartheta)^2}{2}} e^{\tau y} = \frac{\tau^2}{2} + \vartheta\tau$, and $\gamma_F(y) = \frac{1}{2}(y-\vartheta)^2$ for $y \in \mathbb{R}$ yielding

$$D_{\gamma_F}(G, P_V) = \begin{cases} \dfrac{1}{2}\int (f_G - \vartheta)^2 dP_V \text{ if } \sigma_G = 0 \\ +\infty \quad \text{otherwise} \end{cases}$$

With $\vartheta = 0$, D_{γ_F} will be $\frac{1}{2}$ times the L_2 norm squared, or energy, of f_G. This is used when $G \ll P_V$ is the only prior information.

5) Hellinger distance: Let F be the Poissonized convolution of the exponential distribution with mean 1, whose moment generating function is $(1-\tau)^{-1}$, $\tau < 1$. Then $(a_F, b_F) = (0, +\infty)$, $\psi_F(\tau) = \log \sum\limits_{k=0}^{\infty} \frac{1}{ek!}(1-\tau)^{-k} = \frac{\tau}{1-\tau}$, $\tau < 1$, and $\gamma_F(y) = y - 2\sqrt{y} + 1$ for $y \in [0, +\infty) = \mathrm{dom}\gamma_F$. Thus

$$D_{\gamma_F}(G, P_V) = \begin{cases} \int (\sqrt{f}_G - 1)^2 dP_V + \sigma_G(V) \text{ if } f_G \geq 0, \sigma_G \geq 0 \\ +\infty \quad \text{otherwise} \end{cases}$$

is the squared Hellinger distance of G from P_V.

6) <u>α-divergences:</u> The α-divergence of G from P_V is the γ-divergence with the convex function

$$\gamma_\alpha(y) = \frac{y^\alpha - \alpha(y-1) - 1}{\alpha - 1} \qquad (\alpha > 0, \alpha \neq 1).$$

For $0 < \alpha < 1$, this can be obtained as in the previous example $\left(\alpha = \frac{1}{2}\right)$, taking for F the Poissonized convolution of the Γ-distribution with density $\frac{1}{\Gamma(\beta)}\beta^\beta x^{\beta-1}e^{-\beta x}$, $x > 0$, and moment generating function $\left(\frac{\beta}{\beta-\tau}\right)^\beta$, $\tau < \beta$, where $\beta = \frac{\alpha}{1-\alpha}$. Indeed, then $\psi_F(\tau) = \left(\frac{\beta}{\beta-\tau}\right)^\beta - 1$, $\tau < \beta$, and $\gamma_F(y) = \beta y - (1+\beta)y^{\frac{\beta}{1+\beta}} + 1 = \gamma_\alpha(y)$. We note, however, that α-divergences with $\alpha > 1$ do not admit an MEM interpretation, except for the case $\alpha = 2$ covered under 4.

C. Existence criteria

Theorems II.2, II.4 and II.5, with suitable choice of γ or ψ, propose a way to find solutions to Problem (ii) in the Introduction that have some desired a priori properties (positivity, absolute continuity, bounds on the density). These Theorems, however, do not directly adress the question whether such solutions given a set of data exist at all. We recall here results in [15], [21], [23] about existence, in the case $\mathcal{K} = \{c\}$. These rely upon two main ideas. One is to give a specific γ such that $G \in \mathcal{M}(V)$ with the required prior property and satisfying $\int \varphi dG = c$ exists iff $\tilde{f}(c) = \inf\{D_{\gamma_F}(G, P_V) : \int \varphi dG = c\}$ meets a specified upper bound (perhaps depending on c). The other idea is to use eq. (II.22) to turn that bound on $\tilde{f}(c)$ into an effectively computable criterion. The following criteria were obtained in this way. The first one appears in [15] and [22], the second one in [23] and [31]. The third one is proved in Appendix C.

- **Criterion (E1)**. Problem (ii) has a solution in $\mathcal{B}_L(P_V)$ of (II.35) if and only if

$$h_L^\varphi(c) = \sup_v \left(\langle v, c\rangle - \int_V \log(1 + \exp[L\langle v, \varphi\rangle])dP_V\right) \leq 0. \tag{II.36}$$

- **Criterion (E2)**. If $\varphi_1 = 1$, there exists a positive measure solving problem (ii) if and only if

$$h_+^{1,\varphi}(c) = \sup_v \left(\langle v, c\rangle + \int_V \log(1 - \exp[\langle v, \varphi\rangle])dP_V\right) \leq 0. \tag{II.37}$$

- **Criterion (E3)**. If $\varphi_1 = 1$ is not assumed but $\langle \beta, \varphi\rangle \geq m$ holds for a vector $\beta \in \mathbb{R}^k$ and a positive constant m, there exists a positive measure solving problem (ii) if and only if

$$h_+^\varphi(c) = \sup_v \left(\langle v, c\rangle + \int_V \log(1 - \exp[\langle v, \varphi\rangle - 1])dP_V\right) \leq \frac{\langle \beta, c\rangle}{m}. \tag{II.38}$$

Moreover, the functions in the above criteria take the value $+\infty$ at points c for which no solution with the required prior property exists.

III. MEM WITH CORRELATED AMPLITUDES: LINEAR TRANSFER.

We will now consider problem (i). The novelty of solving this problem, compared to problem (ii), is that it allows interpolation and moment problems, with perhaps non continuous moment functions. As said in the introduction, the idea is to transfer linearly problem (ii) to problem (i), see Section III-A below. The stochastic MEM construction will show that it relies, in fact, on a particular correlation structure for the default model on the amplitudes of the random measure involved in the MEM construction, and that it admits a Bayesian interpretation, see Sections III-B. Then, this new method will allow to take into account new kinds of prior information, see Section III-C; in a particular case, a computational criterion to decide the existence of a solution fitting the data and meeting the prior information will be given, see Section III-D.

A. Transfer principle

Let $K(\cdot,\cdot)$ be a measurable bounded real valued function on $U \times V$. Fubini Theorem gives that for any measure G on V and $\mathbb{R}^k$-valued measure Φ on U

$$\int_U \left(\int_V K(x,t)dG(t)\right) d\Phi(x) = \int_V \left(\int_U K(x,t)d\Phi(x)\right) dG(t).$$

In other words, if $G \in \mathcal{M}(V)$ solves problem (ii) with

$$\varphi(t) := \int_U K(x,t)d\Phi(x) \tag{III.1}$$

then

$$g(x) := \int_V K(x,t)dG(t)$$

solves problem (i). In particular, we can take G^* minimizing $D_\gamma(G, P_V)$ subject to $G \in S_{\varphi,\mathcal{K}}$, with φ in (III.1), as in Theorem II.4, and then adopt $g^*(x) := \int_V K(x,t)dG^*(t)$ as the solution of problem (i). This procedure leads to a unique result whenever G^* is unique, for which $a_\psi = -\infty$, $b_\psi = +\infty$ is a sufficient condition.

B. Stochastic interpretation

As problem (i) is linked with a problem of type (ii) via a linear transfer, we may obtain solutions using the same linear transfer applied to the MEM procedure. Write (with the notation (II.24)):

$$p_n(x) := \int_V K(x,t)\nu_n(dt) = \frac{1}{n}\sum_{j=1}^{n} Y_j \cdot K(x,t_j), \tag{III.2}$$

interpreted as a random amplitude at point x. This amplitude is a linear transform of i.i.d. Y_i. As said before, in some physical problems, it is more intuitive to introduce some dependence structure between the amplitudes of neighbouring pixels. Let us detail the linear transfer of MEM. First, we assign as the joint distribution of $Y_1, \ldots, Y_n$ in (III.2) that p.m. Q on $\mathbb{R}^n$ whose I-divergence from $F^{\otimes n}$ is minimum subject to the constraint that the expectation of p_n under this assignement satisfies (1.2), that is,

$$\int E_Q(p_n(x))d\Phi(x) \in \mathcal{K}. \tag{III.3}$$

Since by (III.1), (III.2)

$$\int p_n(x)d\Phi(x) = \int \varphi(t)\nu_n(dt) = \frac{1}{n}\sum_{i=1}^{n} \varphi(t_i)Y_i, \tag{III.4}$$

the latter constraint is equivalent to $Q \in \mathcal{R}_n$, cf. (II.26). Thus in this way exactly Q_n^{MEM} in (II.25) will be assigned. Next, we take as the n'th approximation of our solution to problem (i) via linear transfer by K the expectation under Q_n^{MEM} of p_n, i.e., with G_n^{MEM} defined in (II.27),

$$g_{n,K}^{MEM}(x) := E_{Q_n^{MEM}}(p_n(x)) = \int_V K(x,t)dG_n^{MEM}(t). \tag{III.5}$$

Let us also consider the Bayesian estimate of $p_n(x)$:

$$g_{n,K}^{Bay}(x) = E_{F^{\otimes n}}[p_n(x) \mid \int p_n d\Phi \in \mathcal{K}].$$

Using (III.2), (III.4) and (II.28), this is related to G_n^{Bay} for problem (ii) as

$$g_{n,K}^{Bay}(x) = \int K(x,t)dG_n^{Bay}(t). \tag{III.6}$$

The following is an immediate consequence of Theorem II.5.

Theorem III.1: Suppose the assumptions of Theorem II.5 hold for φ defined by (III.1) and that for each $x \in U$, $K(x,\cdot)$ is a P_V-a.e. continuous function of t. Then, if in Theorem II.5 there is a unique minimizer G^* (in particular if $a_\psi = -\infty$, $b_\psi = +\infty$), $g_{n,K}^{MEM}(x)$ converges to

$$g_K^*(x) = \int K(x,t)G^*(dt) \in E_{\Phi,\mathcal{K}}. \tag{III.7}$$

If in addition Int$\mathcal{K} \neq \emptyset$ then the same holds for $g_{n,K}^{Bay}(x)$.

C. How to incorporate new prior information

The use of a *transfer* kernel K leads to solutions of problem (i) of particular forms. It then allows to consider new prior informations. We assume here that $V = U = [-1, 1]$ and that P_U and P_V are both the uniform probability.

- **Smoothness.** If a smooth solution is required, with initial value 0 at point -1, the kernel
$$K_{der,p}(x,t) = \frac{(x-t)^{p-1}}{(p-1)!} \mathbb{1}_{[-1,x]}(t)$$
leads to a solution with p derivatives as long as the solution G^* involved in Theorem II.4, with $\varphi(t)$ defined by (III.1), is absolutely continuous (see Remark i) after the Theorem). This has been considered by Kemperman in [27].
- **Unimodality.** A non-negative function g on $\mathbb{R}$ is unimodal at 0 if $g(x)$ is non-decreasing for $x < 0$ and non-increasing for $x > 0$. If a unimodal solution to problem (i) is sought, the kernel
$$K_{mod}(x,t) = \begin{cases} \dfrac{1}{|t|} \text{ if } 0 \le x < t \text{ or } t < x \le 0 \\ 0 \quad \text{otherwise} \end{cases}$$
leads to a solution as desired, as long as G^* is positive. We recover here results published by Brockett et al. in [6].
- **Interpolation and moments.** The $\mathbb{R}^{k_1+k_2}$-valued measure $\Phi = (\delta_{s_1}, \ldots, \delta_{s_{k_1}}, \phi_1 P_U, \ldots, \phi_{k_2} P_U)$ leads to an interpolation problem coupled with a generalized moment problem. It may be solved using a continuous kernel K; then $\varphi(t)$ in (III.1) will be
$\varphi(t) = (K(s_1,t), \ldots, K(s_{k_1},t), \int_U K(x,t)\phi_1(x)dP_U(x), \ldots, \int_U K(x,t)\phi_{k_2}(x)dP_U(x))$.

Now, for particular kernels K, the prior information taken into account by the choice of γ may be conserved:

1) If G^* is a positive measure and K is non negative then g^* is a non negative function.

2) If $G^* \in \mathcal{B}_L(P_V)$ and K is a Markovian kernel, that is, $K(x,\cdot)P_V$ is for all $x \in U$ a p.m. on V, then $0 \le g^* \le L$ (see equation (II.35) for the definition of $\mathcal{B}_L(P_V)$).

Finally, all those properties may be coupled with the minimization to be performed. Let us give some examples where we use simultaneously different possibilities of the method to treat particular sets of prior informations. The following list of examples is not exclusive.

- **Minimizing Sobolev norms.** If a solution g to problem (i) possessing p derivatives and minimizing $\int (g^{(p)})^2(x)dx$ is sought, use kernel $K_{der,p}$ and the energy functional associated with the Gaussian distribution F.
- **Convexity.** If a convex solution to problem (i) is sought, use the kernel $K_{der,2}$ together with a γ-divergence leading to an absolutely continuous positive solution (for instance I-divergence), so that the solution will have a non negative second derivative.

- **Interpolation and moments with prior bounds.** If a positive solution with upper bound L is sought to an interpolation problem coupled with a moment problem, use a continuous Markovian kernel and a γ-divergence leading to solutions in $\mathcal{B}_L(P_V)$ (for instance a rescaled Fermi-Dirac entropy).

D. Existence criteria

While the existence of solutions to problem (ii) that fit the data and are positive, or have specific bounds on the density, may be decided using the criteria in Section II-C, it is not obvious whether such criteria may be used to decide if problem (i) admits a solution that fits the data and satisfies some prior information such as convexity, unimodality or smoothness. The latter may be possible when the particular prior information has an equivalent formulation via linear transfer. This is the case for the following problem, with a minor change of the setup in Section III-A. The notation is the same as in Section III-C.

Suppose a p.m. H on $U = [-1, 1]$ is sought that solves the problem

$$\int \Phi(x) dH(x) = c \tag{III.8}$$

subject to the prior information that H is unimodal at 0. The latter means that $H = gP_U + c\delta_0$ where the function g is unimodal at 0 and $c \geq 0$; note that H may have an atom at 0.

A well-known characterization of unimodality (see [19], p. 155) implies that $H = gP_U + c\delta_0$ is a p.m. unimodal at 0 iff

$$g(x) = \int 2K_{mod}(x,t) dG(t), \qquad c = G(\{0\}),$$

for some p.m. G on V. It follows that (III.8) has a desired unimodal solution (given $\Phi = (\Phi_1, \ldots, \Phi_k)$ and $c \in \mathbb{R}^k$) iff a positive solution G of

$$\int \varphi(t) dG(t) = (1, c)$$

exists, where $\varphi = (\varphi_0, \varphi_1, \ldots \varphi_k)$ is defined by

$$\varphi_i(t) = \begin{cases} \int 2K_{mod}(x,t)\Phi_i(x) dP_U(x) = \dfrac{1}{t}\displaystyle\int_0^t \Phi_i(x)dx & \text{if } t \neq 0 \\ \Phi_i(0) & \text{if } t = 0 \end{cases}$$

for $i = 1, \ldots, k$, and $\varphi_0(t) = 1$. Note that this φ is continuous if the functions Φ_i are integrable on U and continuous at $x = 0$. Hence, applying Criterion (E2) in Section II. C, we obtain

Theorem III.2: Let $U = [-1, 1]$. There exists a p.m. H on U which is unimodal at 0 and fits the moments

$$\int \Phi(x) dH(x) = c,$$

where $\Phi = (\Phi_1, \ldots, \Phi_k)$ is integrable on U and continuous at $x = 0$, if and only if

$$h_+^{1,\varphi}((1,c)) \leq 0$$

for $\varphi = (\varphi_0, \ldots, \varphi_k)$ defined above (see (II.37) for the definition of $h_+^{1,\varphi}$).

IV. MEM WITH EXCHANGEABLE AMPLITUDES

In this Section we return to the Problem (ii) stated in Section I where a measure $G \in \mathcal{M}(V)$ is searched that fits the moment constraints (I.3). We will proceed as in Section II-B, considering first discrete random measures

$$\nu_n = \frac{1}{n}\sum_{i=1}^{n} Y_i \delta_{t_i} \quad \text{where} \quad \frac{1}{n}\sum_{i=1}^{n} \delta_{t_i} \to P_V, \tag{IV.1}$$

but now the random amplitudes $Y_1, Y_2, \ldots$ are assumed only exchangeable, rather than i.i.d., under the default model. Thus, according to De Finetti's Theorem (cf. [16] or [1]), the joint distribution F_n of $(Y_1, \ldots, Y_n)$ under the default model is:

$$F_n = \int_\Theta (F_\vartheta)^{\otimes n} d\mu(\vartheta) \tag{IV.2}$$

where $(F_\vartheta)_{\vartheta \in \Theta}$ is a family of p.m.'s on $\mathbb{R}$, and μ is a p.m. on the parameter set Θ. We define the p.m. Q_n^{MEM} on $\mathbb{R}^n$ similarly to (II.25), replacing $F^{\otimes n}$ by F_n, i.e.,

$$Q_n^{MEM} = \text{argmin}_{Q \in \mathcal{R}_n} D(Q \| F_n), \tag{IV.3}$$

with $\mathcal{R}_n$ defined by (II.26), and consider the expectation G_n^{MEM} of ν_n under Q_n^{MEM}, i.e.,

$$G_n^{MEM} = E_{Q_n^{MEM}}(\nu_n) = \frac{1}{n}\sum_{i=1}^{n} E_{Q_n^{MEM}}(Y_i)\delta_{t_i}. \tag{IV.4}$$

We also consider the Bayesian estimate of ν_n, viz.

$$G_n^{Bay} = E_{F_n}\left[\ \nu_n \ \Big| \int_V \varphi d\nu_n \in \mathcal{K}\right]. \tag{IV.5}$$

For convenience, the logarithm of the moment generating function of F_ϑ will be denoted by ψ_ϑ (see (II.29)), and its convex conjugate by γ_ϑ. We will use the notation

$$\mathcal{I}_\vartheta(G) = D_{\gamma_\vartheta}(G, P_V), \qquad \mathcal{I}(G) = \min_{\vartheta \in \Theta} \mathcal{I}_\vartheta(G). \tag{IV.6}$$

If the amplitudes Y_i were i.i.d. with distribution F_ϑ, under the hypotheses of Theorem II.5 the conditional distribution of ν_n on the condition $\int \varphi d\nu_n \in \mathcal{K}$ would accumulate on the set of minimizers of $\mathcal{I}_\vartheta(G)$ subject to $\int \varphi dG \in \mathcal{K}$. We will prove a similar result in the case of exchangeable amplitudes, with $\mathcal{I}_\vartheta(G)$ replaced by $\mathcal{I}(G)$; we will also establish (less direct) extensions of the assertions of Theorem II.5 about G_n^{MEM} and G_n^{Bay}.

The extension of Theorem II.5 to the present setting is easy if Θ is a finite set, but otherwise some technical assumptions appear necessary. Making no attempt at maximum generality, we will adopt the following ones:

$(H3)$ Θ is a compact metric space, μ has full support, $a_{F_\vartheta} = a$, $b_{F_\vartheta} = b$ for each $\vartheta \in \Theta$, and $\psi_\vartheta(\tau)$ is finite valued and jointly continuous on $\mathrm{I\!R} \times \Theta$.

Remark: If a and b are finite, the last assumption requires but the continuous dependence of F_ϑ on ϑ.

Under Hypotheses $(H1)$ and $(H3)$, $\mathcal{I}_\vartheta(G) = D_{\gamma_\vartheta}(G, P_V)$ is a jointly lower semicontinuous function of (G, ϑ), on account of (II.11). In particular, the minimum in (IV.6) is certainly attained. It follows, too, that $\mathcal{I}(G)$ is lower semicontinuous; note, however, that $\mathcal{I}(G)$ need not be convex.

Theorem IV.1: Assume $(H1)$, $(H3)$ and the constraint qualification that for some $G \ll P_V$ with $\int \varphi dG \in \mathcal{K}$

$$\frac{dG}{dP_V} \in (a, b) \qquad P_V \text{ -a.s.} \tag{IV.7}$$

Assume moreover that $\mathrm{Int}\mathcal{K} \neq \emptyset$. Then the conditional distribution of ν_n in (IV.1) on the condition $\int \varphi d\nu_n \in \mathcal{K}$ (under the exchangeable default model) accumulates on the compact set of minimizers of $\mathcal{I}(G)$ subject to $\int \varphi dG \in \mathcal{K}$, and G_n^{Bay} accumulates in the convex hull of that set.

The proof relies on the fact that if the joint distribution of the amplitudes $Y_1, \dots, Y_n$ is given by (IV.2), the random measures ν_n satisfy a large deviations principle with rate function $\mathcal{I}$. It will be given in Appendix D.

The convergence of G_n^{MEM} will be treated, for simplicity, when $\mathcal{K}$ consists of a single point $c \in \mathrm{I\!R}^k$. It turns out that for the limiting behaviour of G_n^{MEM} the relevant functional is the affine regularization of $\mathcal{I}(G)$, to be denoted by $\tilde{\mathcal{I}}(G)$. It is defined ([17], p. 135) as the maximum convex and lower semicontinuous functional of G bounded above by $\mathcal{I}(G)$. The existence of the latter is clear since the supremum of convex and lower semicontinuous functionals also has these properties.

Theorem IV.2: Assume $(H1)$, $(H3)$, and the constraint qualification (IV.7), where $\mathcal{K} = \{c\}$. Then the sequence G_n^{MEM} defined by (IV.4) accumulates in the set of minimizers of $\tilde{\mathcal{I}}(G)$ subject to $\int \varphi dG = c$.

The proof again relies on a L.D.P. for ν_n, this time when the random amplitudes $Y_1, \dots, Y_n$ have joint distribution Q_n^{MEM}; it will be given in Appendix D.

Remark: Theorems IV.1 and IV.2 are most interesting when $\mathcal{I}(G)$ respectively $\tilde{\mathcal{I}}(G)$ has a unique minimizer subject to $\int \varphi dG \in \mathcal{K}$; this is, however, less expected than uniqueness in Theorem II.5. Note that since $\mathrm{dom}\psi_\vartheta = \mathbb{R}$ by Hypothesis (H3), the minimizer of $\mathcal{I}_\vartheta(G)$ subject to $\int \varphi dG \in \mathcal{K}$ is unique for each fixed $\vartheta \in \Theta$ and is of form

$$G^*_\vartheta = g^*_\vartheta P_V \quad \text{with} \quad g^*_\vartheta = \psi'_\vartheta(\langle v_\vartheta, \varphi\rangle), \quad v_\vartheta \in \mathbb{R}^k, \tag{IV.8}$$

see Theorem II.4 and Remark i) after it. It follows that each minimizer of $\mathcal{I}(G)$ subject to $\int \varphi dG \in \mathcal{K}$ is of form (IV.8), for some $\vartheta \in \Theta$, and consequently each accumulation point of G_n^{Bay} in Theorem IV.1 is a mixture of measures of that form. The same holds also for the accumulation points of G_n^{MEM} in Theorem IV.2 because, as seen from the proof, they are mixtures of measures in a set $\mathcal{N}$ such that each $G \in \mathcal{N}$ is a minimizer of $\mathcal{I}(G)$ given the value of $\int \varphi dG$, see (D.21), (D.26).

In Examples 1)-4) at the end of Section II-B, $D_{\gamma_F}(\cdot, P_V)$ has been determined for various choices of F, depending on a parameter equal to the expectation of F. Continuing those examples, we now determine the corresponding functional $\mathcal{I}(G)$ of (IV.6), for some natural choices of the set Θ of permitted ϑ values. Note, however, that case 2) (Burg entropy) is not covered by Theorems IV.1 and IV.2 since Hypothesis (H3) does not hold in that case.

Let F_ϑ be the distribution F with expectation ϑ in either of Examples 1-4 in Section II-B, and let γ_ϑ be the corresponding function γ_F. Then, using the explicit formulas in Examples 1-4, elementary algebra gives that $\mathcal{I}_\vartheta(G) = D_{\gamma_\vartheta}(G, P_V)$ satisfies

$$\mathcal{I}_\vartheta(G) = \mathcal{I}_{G(V)}(G) + \gamma_\vartheta(G(V)). \tag{IV.9}$$

Note that $\gamma_\vartheta(y)$ as a function of ϑ is decreasing for $\vartheta < y$ and increasing for $\vartheta > y$; of course, it equals 0 for $\vartheta = y$. Hence, if $\Theta = [\vartheta_1, \vartheta_2]$ is a closed subinterval of the set of ϑ values admissible in the corresponding example, it follows from (IV.9) that in all four cases

$$\mathcal{I}(G) = \min_{\vartheta\in\Theta} \mathcal{I}_\vartheta(G) = \begin{cases} \mathcal{I}_{G(V)}(G) & \text{if } \vartheta_1 \le G(V) \le \vartheta_2 \\ \mathcal{I}_{\vartheta_1}(G) & \text{if } G(V) \le \vartheta_1 \\ \mathcal{I}_{\vartheta_2}(G) & \text{if } G(V) \ge \vartheta_2. \end{cases} \tag{IV.10}$$

The functional (IV.10) is convex in all four cases, hence $\tilde{\mathcal{I}} = \mathcal{I}$. On the other hand, letting Θ be the two-point set $\{\vartheta_1, \vartheta_2\}$ leads to the non-convex functional

$$\min\left(\mathcal{I}_{\vartheta_1}(G), \mathcal{I}_{\vartheta_2}(G)\right) = \begin{cases} \mathcal{I}_{\vartheta_1}(G) & \text{if } G(V) \le \overline{\vartheta} \\ \mathcal{I}_{\vartheta_2}(G) & \text{if } G(V) \ge \overline{\vartheta}, \end{cases}$$

where $\overline{\vartheta}$ is determined by $\gamma_{\vartheta_1}(\overline{\vartheta}) = \gamma_{\vartheta_2}(\overline{\vartheta})$. It can be shown, using eq. (IV.9), that the affine regularization of the latter is equal to $\mathcal{I}(G)$ given by (IV.10), again in all four cases. Note finally the obvious implication of (IV.9) that in case $\varphi_1 = 1$, the minimizer of $\mathcal{I}_\vartheta(G)$ subject to $\int \varphi dG = c$ is the same for every ϑ, and hence it is also the (unique) minimizer of $\mathcal{I}(G)$ or $\tilde{\mathcal{I}}(G)$ subject to $\int \varphi dG = c$.

Appendices

A. Proof of Theorem II.2

Lemma A.1: Given $\varphi_1, \ldots, \varphi_k$ in $L_1(P_U)$, suppose that g is any (measurable) function on U such that

$$a < g(x) < b, \qquad g\varphi_i \in L_1(P_U), \qquad i = 1, \ldots, k. \tag{A.1}$$

Then there exist $a' > a$, $b' < b$ and $\tilde{g}$ such that

$$a' \leq \tilde{g}(x) \leq b', \qquad \int \tilde{g}\varphi_i dP_U = \int g\varphi_i dP_U, \qquad i = 1, \ldots, k. \tag{A.2}$$

Proof: (cf. [3], proof of Theorem 2.9).
Let L denote the subspace of $\mathbb{R}^k$ consisting of the vectors representable as $\int h\varphi dP_U$, with some $h: U \to \mathbb{R}$. Fix strictly decreasing respectively increasing sequences $a_n \to a$, $b_n \to b$, and write

$$T_n = \{x : a_n \leq g(x) \leq b_n\}. \tag{A.3}$$

We first claim that for n sufficiently large, the set of vectors $v \in \mathbb{R}^k$ representable as $\int \varphi h dP_U$ with bounded $h: U \to \mathbb{R}$ that vanishes outside T_n, is equal to L. Indeed, were this set a proper subspace of L, there would exist $v_n \in L$ with $\|v_n\| = 1$ such that

$$\langle v_n, \int \varphi h dP_U \rangle = \int \langle v_n, \varphi \rangle h dP_U = 0$$

for every h as above, thus $\langle v_n, \varphi \rangle = 0$ a.s. on T_n. Picking a convergent subsequence $v_{n_\ell} \to v$, this would imply $\langle v, \varphi \rangle = 0$ a.s. on U, hence $v \perp L$ by the definition of L, a contradiction since $v \in L$, $\|v\| = 1$.

Fix n_0 for which the above claim holds. Then for any $\delta > 0$, the affine hull of the convex set

$$C_\delta = \left\{ \int \varphi h dP_U \text{ with } h: |h(x)| \leq \delta \text{ on } T_{n_0},\ h(x) = 0 \text{ on } T_n^c \right\}. \tag{A.4}$$

equals L, and the origin belongs to the interior of C_δ in the relative topology of L.

Write

$$\begin{aligned} g_n(x) \;&=\; a_n \text{ if } g(x) < a_n \\ &\; g(x) \text{ if } a_n \leq g(x) \leq b_n \\ &\; b_n \text{ if } g(x) > b_n. \end{aligned}$$

Then $\int(g_n - g)\varphi dP_U \in L$ approaches 0 as $n \to +\infty$, hence $\int(g_n - g)\varphi dP_U \in C_\delta$ for $n \geq n_1(\delta)$. This means that

$$\int (g_n + h)\varphi dP_U = \int g\varphi dP_U$$

for some h as in (A.4), and $\tilde{g} = g_n + h$ satisfies the assertion with $a' = \min(a_n, a_{n_0} - \delta)$, $b' = \max(b_n, b_{n_0} + \delta)$. ■

Proof of Theorem II.2: Write

$$f(c) = \inf\left\{ I_\gamma(g)\colon \int g\varphi dP_U = c\right\}. \tag{A.5}$$

Clearly, this is a convex function, with $\mathrm{dom} f = \{c\colon f(c) < +\infty\} \subset L$, and

$$I_\gamma(E_{\varphi,\mathcal{K}}) = \inf_{g\in E_{\varphi,\mathcal{K}}} I_\gamma(g) = \inf_{c\in\mathcal{K}} f(c) = \inf_{c\in\mathbb{R}^k} [f(c) + h(c)] \tag{A.6}$$

where

$$\begin{aligned} h(c) &= 0 \text{ if } c \in \mathcal{K}\cap L \\ &\quad +\infty \text{ otherwise.} \end{aligned} \tag{A.7}$$

As in ([3], Theorem 4.2), we will apply the Fenchel duality theorem ([37], p. 327). By this theorem, for convex functions f and h such that the relative interiors of $\mathrm{dom} f$ and $\mathrm{dom} h$ intersect,

$$\inf_{c\in\mathbb{R}^k}[f(c) + h(c)] = \max_{v\in\mathbb{R}^k}[-f^*(v) - h^*(-v)] \tag{A.8}$$

where * denotes convex conjugation, i.e. $f^*(v) = \sup_{c\in\mathbb{R}^k}[\langle c, v\rangle - f(c)]$.

We first check that f and h defined by (A.5) and (A.7) meet the condition on relative interiors. To this, since $\mathrm{dom} f \subset L$, it suffices to show the existence of $c_0 \in \mathrm{dom} f \cap \mathcal{K}$ such that c_0 is in the interior of $\mathrm{dom} f$ in the relative topology of L.

By Lemma A.1, the Hypothesis (II.15) implies the existence of $\tilde{g} \in E_{\varphi,\mathcal{K}}$ satisfying $a_\gamma < a' \leq \tilde{g}(x) \leq b' < b_\gamma$. Then clearly $I_\gamma(\tilde{g}) < +\infty$ hence $c_0 = \int \tilde{g}\varphi dP_U \in \mathcal{K}$ belongs to $\mathrm{dom} f$. Further, since the origin is in the interior of C_δ in the relative topology of L, cf. (A.4), some neighborhood of c_0 in that topology also belongs to $\mathrm{dom} f$, because $\tilde{g} + h$ with $|h| < \delta$ is still bounded away from a_γ and b_γ if δ is sufficiently small.

Next, the convex conjugate of f in (A.5) is calculated as

$$\begin{aligned} f^*(v) &= \sup_c \left(\langle v, c\rangle - \inf\left\{ \int \gamma(g) dP_U \text{ with } g\colon \int g\varphi dP_U = c\right\}\right) \\ &= \sup_c \sup \left\{ \langle v, c\rangle - \int \gamma(g) dP_U \text{ with } g\colon \int g\varphi dP_U = c\right\} \\ &= \sup_g \left(\langle v, \int g\varphi dP_U\rangle - \int \gamma(g) dP_U\right) = \sup_g \left(\int (\langle v, \varphi\rangle g - \gamma(g)\right) dP_U \\ &= \int \psi(\langle v, \varphi\rangle) dP_U. \end{aligned}$$

For a rigorous justification of the last step ("obvious" from (II.1)) cf. [36].

Since clearly $h^*(-v) = \sup_{c\in\mathcal{K}\cap L}\langle -v, c\rangle = -\inf_{c\in\mathcal{K}\cap L}\langle v, c\rangle$, (A.6) and (A.8) prove the assertion (II.16) of Theorem II.2.

For v^* attaining the maximum in (II.16), write $\eta(x) = \langle v^*, \varphi(x)\rangle$. Then clearly $\eta \in \mathrm{dom}\psi$ a.s. Further, even if the interval $\mathrm{dom}\psi$ contains its endpoints a_ψ, b_ψ (where $a_\psi < 0 < b_\psi$ by (II.2)), v^* could not be a maximizer if $\eta = \langle v^*, \varphi\rangle$ were equal to a_ψ or b_ψ on a set of positive measure, on account of (II.3). Thus $a_\psi < \eta(x) < b_\psi$ a.s. must hold, and therefore $\psi'(\eta(x))$ is well defined.

The assertion that $g^* = \psi'(\eta)$ satisfies (II.13) for $E = E_{\varphi,\mathcal{K}}$ means, on account of (II.12), (II.16), that

$$\int[\gamma(\psi'(\eta)) + \gamma'(\psi'(\eta))(g - \psi'(\eta))]dP_U \geq \inf_{c\in\mathcal{K}\cap L}\langle v^*, c\rangle - \int \psi(\eta)dP_U, \tag{A.9}$$

whenever $g \in E_{\varphi,\mathcal{K}}$, $I_\gamma(g) < +\infty$.

Since γ is the convex conjugate of ψ,

$$\gamma(\psi'(\eta)) = \sup_\tau(\tau\psi'(\eta) - \psi(\tau)) = \eta\psi'(\eta) - \psi(\eta).$$

Further,

$$\int g\eta dP_U = \langle v^*, \int g\varphi dP_U\rangle \geq \inf_{c\in\mathcal{K}\cap L}\langle v^*, c\rangle \quad \text{if} \quad g \in E_{\varphi,\mathcal{K}}.$$

Hence (A.9) will be proved if we show that

$$\int[\eta - \gamma'(\psi'(\eta))][\psi'(\eta) - g]dP_U \geq 0 \text{ if } I_\gamma(g) < +\infty. \tag{A.10}$$

Now, it is (known and) easy to see that

$$\gamma'(\psi'(\tau)) = \tau \quad \text{if} \quad \lim_{y\to a_\gamma+0}\gamma'(y) < \tau < \lim_{y\to b_\gamma-0}\gamma'(y).$$

In particular, if γ is essentially smooth then $\gamma'(\psi'(\tau)) = \tau$ for all $\tau \in (a_\psi, b_\psi)$, hence in that case (A.10) holds with the equality.

Otherwise, if $a_\gamma > -\infty$ and $\gamma'(a_\gamma) = \lim_{y\to a_\gamma+0}\gamma'(y) > -\infty$, then

$$\psi(\tau) = \tau a_\gamma - \gamma(a_\gamma), \qquad \psi'(\tau) = a_\gamma \quad \text{if} \quad \tau \leq \gamma'(a_\gamma).$$

Thus in case $\eta \leq \gamma'(a_\gamma)$ we have $\eta - \gamma'(\psi'(\eta)) \leq 0$, and $\psi'(\eta) - g = a_\gamma - g \leq 0$ whenever $I_\gamma(g) < +\infty$. One sees similarly that in case $\eta \geq \gamma'(b_\gamma)$ we have $\eta - \gamma'(\psi'(\eta)) \geq 0$ and $\psi'(\eta) - g = b_\gamma - g \geq 0$, thus (A.10) always holds even if γ is not essentially smooth.

Finally, to prove the last assertion of Theorem II.2, notice that since v^* defined by (II.18) is the maximizer in (A.8) or equivalently the minimizer of the convex function

$$\int \psi(\langle v, \varphi\rangle)dP_U - \inf_{c\in\mathcal{K}\cap L}\langle v, c\rangle = f^*(v) + h^*(-v),$$

0 must be a subgradient at v^* of the latter. Since ψ is an essentially smooth convex function, using the assumption that v^* is an interior point of

$$\mathrm{dom} f^* = \left\{ v\colon \int \psi(\langle v, \varphi\rangle) dP_U < +\infty \right\}$$

it follows that f^* is differentiable at v^* with gradient

$$\left(\int \psi'(\langle v^*, \varphi\rangle)\varphi_i dP_U \right)_{i=1}^{k}. \tag{A.11}$$

Further, every subgradient of h^* is an element of $\mathcal{K}$ ([37], Corollary 23.5.3, p. 219). Hence, using the theorem of ([37], p. 223) about the subdifferential of the sum of convex functions, it follows that the vector (A.11) must belong to $\mathcal{K}$, i.e., $g^* \in E_{\varphi,\mathcal{K}}$. ■

Proof of the claim in Remark (ii) after Theorem II.4: Suppose (II.20). Then, G may be chosen such that $g = \frac{dG}{dP_V}$ satisfies $a_\gamma < a' \le g(x) \le b' < b_\gamma$, by Lemma A.1. As the subspace L in the proof of Lemma A.1 now equals the whole $\mathbb{R}^k$, $\int \varphi dG \in \mathcal{K}$ and $\mathrm{Int}\mathcal{K} \neq \emptyset$ imply that for any $\delta > 0$ there exists h with $|h| < \delta$ such that $\tilde{G} = (g+h)P_V$ satisfies $\int \varphi d\tilde{G} \in \mathrm{Int}\mathcal{K}$. If $\delta < \min(a' - a_\gamma, b_\gamma - b')$ then clearly $D_\gamma(\tilde{G}, P_V) < +\infty$, proving that (II.20) implies (II.21). The converse implication follows since it is easy to check, using Hypothesis $(H1)$, that to any $G \in \mathcal{M}(V)$ with $D_\gamma(G, P_V) < +\infty$ and any $\epsilon > 0$ there exists $\tilde{G} \ll P_V$ with $a_\gamma < \frac{d\tilde{G}}{dP_V} < b_\gamma$ such that

$$\| \int \varphi d\tilde{G} - \int \varphi dG \| < \epsilon, \quad |D_\gamma(\tilde{G}, P_V) - D_\gamma(G, P_V)| < \epsilon.$$

B. Proof of Theorem II.5

Consider first the assertion about G_n^{MEM}. Checking carefully the proof given in [23] (proof of Theorem 2.1) it may be seen to remain valid under the present weaker hypotheses provided that the relative compactness of the sequence (G_n^{MEM}), i.e., the uniform boundedness of both the positive and negative parts of G_n^{MEM} can be shown. We will assume that $a_F = -\infty$ and $b_F = +\infty$, the case where a_F and b_F are both finite is obvious and that of a semiinfinite interval is left to the reader. For any measure μ supported by $(t_i)_{i=1\ldots n}$ set

$$\Gamma_{F,n}(\mu) := \frac{1}{n} \sum_{i=1}^{n} \gamma_F\left(\mu(\{t_i\})\right). \tag{B.1}$$

Note that $\Gamma_{F,n}(\mu)$ may be decomposed as

$$\Gamma_{F,n}(\mu) = \Gamma_{F,n}(\mu^+) + \Gamma_{F,n}(-\mu^-) - \gamma_F(0), \tag{B.2}$$

where $\mu = \mu^+ - \mu^-$ is the Jordan decomposition of μ. From obvious considerations on the convex function γ_F, whose minimum 0 is attained at $y_0 = \int y dF(y)$, for $a_\gamma <$

$y_1 < y_0 < y_2 < b_\gamma$ we have

$$\Gamma_{F,n}(-\mu^-) - \gamma_F(y_1) + y_1\gamma_F'(y_1) \geq -\gamma_F'(y_1)\int_V d\mu^- \tag{B.3}$$

$$\Gamma_{F,n}(\mu^+) - \gamma_F(y_2) + y_2\gamma_F'(y_2) \geq \gamma_F'(y_2)\int_V d\mu^+, \tag{B.4}$$

where $\gamma_F'(y_1) < 0 < \gamma_F'(y_2)$.
The key point is now the constraint qualification (II.34). This implies, using a perturbation argument, that there exists a continuous function g with $gP_V \in S_{\varphi,\mathcal{K}}$ such that

$$\int_V \gamma_F(g)dP_V = M < +\infty.$$

Arguing as in Proposition 3.1 of [23], it follows that for large enough n, $0 \leq \Gamma_{F,n}(G_n^{MEM}) \leq M+\epsilon$. Using (B.2) and inequalities (B.3), (B.4) to $\mu = G_n^{MEM}$, we may conclude that

$$\int_V dG_n^{MEM,-} \leq \frac{-(M+\epsilon-\gamma_F(y_1)+y_1\gamma_F'(y_1))}{\gamma_F'(y_1)}$$

$$\int_V dG_n^{MEM,+} \leq \frac{M+\epsilon-\gamma_F(y_2)+y_2\gamma_F'(y_2)}{\gamma_F'(y_2)}.$$

Hence, the sequence (G_n^{MEM}) is relatively compact.

Consider next the assertion about G_n^{Bay}. As in [23], it is a consequence of a large deviations principle (L.D.P.) for the sequence (ν_n) of $\mathcal{M}(V)$-valued random variables, see Section D below about the L.D.P.. We need the following simple fact: if the p.m. F satisfies Hypothesis $(H2)$ then for each τ with $a_\psi < -\tau < \tau < b_\psi$, positive integer n, and $C > (n\tau)^{-1}$,

$$\int_{\frac{1}{n}\sum_{i=1}^n |y_i|>C} \left(\frac{1}{n}\sum_{i=1}^n |y_i|\right) dF^{\otimes n} \leq Cn\exp\left\{-n[C\tau - \max(\psi_F(\tau),\psi_F(-\tau)) - \log 2]\right\}. \tag{B.5}$$

Indeed, since $xe^{-\tau x}$ is a decreasing function of $x > \tau^{-1}$

$$\begin{aligned}\int_{\sum_{i=1}^n |y_i|>nC} \left(\sum_{i=1}^n |y_i|\right) dF^{\otimes n} &\leq Cne^{-Cn\tau}\int_{\sum_{i=1}^n |y_i|>nC} e^{\tau\sum_{i=1}^n |y_i|}dF^{\otimes n} \\ &\leq Cne^{-Cn\tau}\left[\int e^{n\tau|y|}dF\right]^n \leq Cne^{-Cn\tau}\left[2\exp\{\max(\psi_F(\tau),\psi_F(-\tau)\}\right]^n.\end{aligned}$$

Now, let R_n denote the distribution of ν_n under the default model $F^{\otimes n}$. The proof in [23] that the sequence (R_n) of p.m.'s on $\mathcal{M}(V)$, or equivalently the random sequence (ν_n), satisfies a L.D.P. with good rate function $L(\cdot) = D_{\gamma_F}(\cdot, P_V)$, remains valid since exponential tightness required for that proof holds by (B.5).

We will apply Lemma D.1 in Appendix D with $E = \mathcal{M}(\mathcal{V})$, to ν_n and $S_{\varphi,\mathcal{K}}$ in the role of Z_n and F. Condition (D.2) there becomes

$$\min_{G \in S_{\varphi,\mathcal{K}}} D_{\gamma_F}(G, P_V) = \inf_{G \in \mathrm{Int}\, \mathrm{S}_{\varphi,\mathcal{K}}} D_{\gamma_F}(G, P_V) < \infty. \tag{B.6}$$

To check that (B.6) holds, recall that since Int$\mathcal{K} \neq \emptyset$, the constraint qualification (II.34) in Theorem II.5 is equivalent to (II.21), with $\gamma = \gamma_F$. Then (B.6) follows by considering $G_\alpha = (1-\alpha)G^* + \alpha G_1$ where G^* attains the minimum on the left hand side, G_1 satisfies (II.21), and $\alpha \downarrow 0$.

Lemma D.1 immediately gives the assertion of Theorem II.5 about the conditional distribution of ν_n on the condition $\int \varphi d\nu_n \in \mathcal{K}$. To complete the proof, it suffices to show that if for some subsequence $\{n_i\}$ the above conditional distributions converge to a p.m. R^*, $G_{n_i}^{Bay}$ converges to the R^*-mixture G^* of measures in $\mathcal{M}(V)$. Indeed, by the previous result, R^* is concentrated on the compact set of minimizers of $D_{\gamma_F}(G, P_V)$ subject to $G \in S_{\varphi,\mathcal{K}}$, and therefore, using convexity, G^* belongs to that set of minimizers. Note that the mixture G^* is well defined since the total variation of the measures in the support of R^* is uniformly bounded, by compactness.

Now, take $s(G) = \int f dG$ in Lemma D.1, where $f \in C(V)$ is arbitrary. Condition (D.4) is satisfied with this choice, on account of (B.5). It follows that $\int f(t) dG_n^{Bay} = E[\int f(t) d\nu_{n_i}(t) \mid \nu_{n_i} \in S_{\varphi,\mathcal{K}}]$ converges to $\int (\int f(t) dG(t)) dR^*(G) = \int f(t) dG^*(t)$, by (D.5). Since this holds for every $f \in C(V)$, we have $G_{n_i}^{Bay} \to G^*$ as claimed. ■

C. Proof of the existence criterion (E3)

Take

$$F = \frac{e-1}{e} \sum_{n=0}^{\infty} e^{-n} \delta_n.$$

Easy computation gives

$$\psi_F(\tau) = \log(\frac{e-1}{e}) - \log[1 - \exp(\tau - 1)]$$

and

$$\gamma_F(y) = -\log(\frac{e-1}{e}) + y + y\log y - (y+1)\log(1+y).$$

If there exists a positive measure G satisfying $\int \varphi dG = c$, equation (II.22) gives

$$\sup_v (\langle v, c\rangle - l(v)) \leq D_{\gamma_F}(G, P_V), \tag{C.1}$$

where

$$\ell(v) = \int \psi(\langle v, \varphi\rangle) dP_V = \log\frac{e-1}{e} - \int \log[1 - \exp(\langle v, \varphi\rangle - 1)] dP_V. \tag{C.2}$$

As $\gamma_F(y) \leq -\log(\frac{e-1}{e}) + y$ if $y \geq 0$, here

$$D_{\gamma_F}(G, P_V) \leq -\log\left(\frac{e-1}{e}\right) + G(V). \tag{C.3}$$

But $1 \leq \frac{\langle \beta, \varphi \rangle}{m}$ implies $G(V) \leq \frac{\langle \beta, c \rangle}{m}$, so that (C.1), (C.2), (C.3) give

$$\sup_v \Big(\langle v, c \rangle + \int \log \left[1 - \exp(\langle v, \varphi \rangle - 1) \right] \Big) \leq \frac{\langle \beta, c \rangle}{m}.$$

If, on the other hand, no positive measure G satisfies $\int \varphi dG = c$, then (II.22) leads to $\sup_v(\langle v, c \rangle + \int \log[1 - \exp(\langle v, \varphi \rangle - 1)]) = +\infty$. ■

D. Large deviations and proof of Theorems IV.1, IV.2

D.1. Large deviations

We briefly recall here some basic concepts of large deviations theory, restricting attention to p.m.'s on a Polish space and requiring a good rate function. For more on this topic, the reader is referred to Dembo and Zeitouni [17]. Given a Polish space E endowed with its Borel σ-field $\mathcal{B}(E)$, a sequence (R_n) of p.m.'s on $(E, \mathcal{B}(E))$ is said to satisfy a large deviations principle (L.D.P.) with good rate function L if

(i) L is a non-negative function with compact level sets, i.e., $\{z \in E : L(z) \leq b\}$ is compact for each $b \geq 0$;

(ii) for every open set $B \subset E$

$$\liminf_{n \to +\infty} \frac{1}{n} \log R_n(B) \geq - \inf_{z \in B} L(z),$$

(iii) for every closed set $B \subset E$

$$\limsup_{n \to +\infty} \frac{1}{n} \log R_n(B) \leq - \inf_{z \in B} L(z).$$

A sequence of E-valued random variables Z_n is said to satisfy a L.D.P. if their distributions R_n satisfy a L.D.P. Several criteria available in the literature to establish a L.D.P. involve the concept of exponential tightness. A sequence (R_n) of p.m's on E is called exponentially tight if to every $\alpha > 0$ there exists a compact set $K_\alpha \subset E$ such that

$$\limsup_{n \to \infty} \frac{1}{n} \log R_n(K_\alpha^c) < -\alpha. \tag{D.1}$$

The following Lemma connects large deviations with conditional expectations, in a form convenient for the proof of our Bayesian results. This Lemma appears in a slightly different form in [20], and expresses a classical phenomenon, cf. "Gibbs conditioning principle" in [17]. For more abstract considerations on the link between conditioning and large deviations cf. [32].

Lemma D.1: Let (Z_n) be a sequence of E-valued random variables with distributions R_n, satisfying a L.D.P. with good rate function L. Let $F \subset E$ be a closed set such that

$$\min_{z \in F} L(z) = \inf_{z \in \mathrm{Int} F} L(z) = L_0 < \infty. \tag{D.2}$$

Then the sequence of conditional distributions $R_n(\cdot \mid F)$ of Z_n on the condition $Z_n \in F$ accumulates on the compact set

$$F_0 = \{z : z \in F,\ L(z) = L_0\}. \tag{D.3}$$

Further, given any continuous function $s: E \to R$ such that for some $C > 0$

$$\limsup_{n\to\infty} \frac{1}{n} \log \int_{F\cap\{z:|s(z)|>C\}} |s(z)| dR_n(z) < -L_0, \tag{D.4}$$

$R_{n_i}(\cdot \mid F) \to R^*$ implies that

$$E(s(Z_{n_i}) \mid Z_{n_i} \in F) \to \int_{F_0} s(z) dR^*(z). \tag{D.5}$$

Proof: The compactness of F_0 follows from the L.D.P. property (i). Assumption (D.2) and properties (ii), (iii) imply that

$$\lim_{n\to\infty} \frac{1}{n} \log R_n(F) = \lim_{n\to\infty} \frac{1}{n} \log R_n(\mathrm{Int}\, F) = -L_0. \tag{D.6}$$

The sequence of p.m.'s $R_n(\cdot \mid F)$, where

$$R_n(B \mid F) = \frac{R_n(B \cap F)}{R_n(F)}, \tag{D.7}$$

is relatively compact since it is tight, even exponentially. The latter follows, using (D.6), (D.7), because the assumed L.D.P. implies that $\{R_n\}$ is exponentially tight, see [17], p. 106.

To check that each accumulation point of $\{R_n(\cdot \mid F)\}$ is a p.m. concentrated on F_0, set $B_k = \{z : d(z, F_0) \geq \frac{1}{k}\}$ where d denotes the distance in E. Then $\min_{z\in B_k\cap F} L(z) > L_0$ by (D.2), thus (D.7) gives by (D.6) and the L.D.P. property (iii) that $R_n(B_k \mid F) \to 0$ as $n \to \infty$. Hence $R_{n_i}(\cdot \mid F) \to R^*$ implies for the closed sets $C_k = \{z : d(z, F_0) \leq \frac{1}{k}\} \supset B_k^c$ that

$$R^*(C_k) \geq \liminf_{n\to\infty} R_{n_i}(C_k \mid F) \geq \liminf_{i\to\infty} R_{n_i}(B_k^c \mid F) = 1.$$

This proves that

$$R^*(F_0) = \lim_{k\to\infty} R^*(C_k) = 1. \tag{D.8}$$

Finally, (D.5) follows since $E(s(Z_{n_i}) \mid Z_{n_i} \in F)$ equals, by definition, the integral of $s(z)$ with respect to $R_{n_i}(\cdot \mid F)$, and (D.4), (D.6), (D.7) imply that the continuous function $s(z)$ is uniformly integrable with respect to the p.m.'s $R_{n_i}(\cdot \mid F)$. ∎

D.2. Exchangeable amplitudes, proof of Theorem IV.1

In this subsection and the next one, we establish L.D.P.'s for the sequence of E-valued random variables (ν_n) defined by (IV.1), where $E = \mathcal{M}(V)$ (with V compact metric), under different assumptions on the joint distributions of the random amplitudes Y_i. In the sequel, the generic element of $\mathcal{M}(V)$ will be denoted by ν rather than G.

Notice that $E = \mathcal{M}(V)$, with the topology of weak convergence, is a topological vector space with dual $E^* = C(V)$. Note also that a closed set $F \subset \mathcal{M}(V)$ is compact iff $\sup_{\nu \in F} \|\nu\| < \infty$, where $\|\nu\|$ denotes the total variation of ν.

Theorem D.2: If the random amplitudes $Y_1, \ldots, Y_n$ have distribution F_n as in (IV.2), subject to Hypothesis $(H3)$, the sequence (ν_n) of (IV.1) satisfies a L.D.P. with good rate function $\mathcal{I}$ given by (IV.6).

Proof: We shall apply a result of Dinwoodie and Zabell ([18], Theorem 2.3), stated in our terminology as follows.

Theorem D.3: ([18]) Let $(P^n_\vartheta)_{\vartheta \in \Theta}$ be a family of sequences of p.m.'s on a Polish space E, with Θ compact metric, such that whenever $\vartheta_n \to \vartheta$, the sequence $(P^n_{\vartheta_n})$ satisfies a L.D.P. with good rate function $L(z, \vartheta)$. Suppose that $L(z, \vartheta)$ is a jointly lower semicontinuous function of (z, ϑ). Then the sequence of mixtures $P_n = \int_\Theta P^n_\vartheta d\mu(\vartheta)$, with arbitrary p.m. μ of full support, satisfies a L.D.P. with good rate function $L(z) = \min_{\vartheta \in \Theta} L(z, \vartheta)$.

To deduce Theorem D.2 from Theorem D.3, it suffices to check that if $\vartheta_n \to \vartheta$ and Q_n denotes the distribution of ν_n when $Y_1, \ldots, Y_n$ have joint distribution $F^{\otimes n}_{\vartheta_n}$, the sequence (Q_n) satisfies a L.D.P. with good rate function $\mathcal{I}_\vartheta(\cdot)$. Indeed, the lower semicontinuity hypothesis required in Theorem D.3 is satisfied in our case, as noted in Section IV.

The desired L.D.P. for (Q_n) can be obtained using the abstract Gärtner-Ellis-Baldi Theorem (see [17], p. 139), similarly to the proof of the L.D.P. for ν_n in the case of i.i.d. amplitudes in [23]. Let Λ_n be the logarithmic moment generating function of Q_n. Thus, Λ_n is defined on $E^* = C(V)$, and for each $f \in C(V)$

$$\begin{aligned} \Lambda_n(nf) &= \log E_{Q_n}\left(\exp n \int f d\nu_n\right) \\ &= \log E_{F^{\otimes n}_{\vartheta_n}}\left(\exp \sum_{i=1}^n f(t_i) Y_i\right) = \sum_{i=1}^n \psi_{\vartheta_n}(f(t_i)). \end{aligned} \tag{D.9}$$

Since $\psi_\vartheta(\tau)$ is continuous on $\mathbb{R} \times \Theta$ by $(H3)$, (D.9) implies on account of $\vartheta_n \to \vartheta$ and (IV.1) that $\frac{1}{n}\Lambda_n(nf)$ converges to $\Lambda(f) = \int \psi_\vartheta(f(t)) dP_V(t)$. The convex conjugate Λ^* of this Λ equals $D_{\gamma_\vartheta}(\cdot, P_V) = \mathcal{I}_\vartheta(\cdot)$ by (II.11). Hence the Gärtner-Ellis-Baldi Theorem gives the desired L.D.P. for (Q_n) if one checks that the sequence Q_n is exponentially tight, and that for every open set $G \subset \mathcal{M}(V)$

$$\inf_{\nu \in G \cap F} \mathcal{I}_\vartheta(\nu) = \inf_{\nu \in G} \mathcal{I}_\vartheta(\nu), \tag{D.10}$$

where F is the set of those $\overline{\nu} \in \mathcal{M}(V)$ for which an $f \in C(V)$ exists such that

$$\int f d\overline{\nu} - \mathcal{I}_\vartheta(\overline{\nu}) > \int f d\nu - \mathcal{I}_\vartheta(\nu), \quad \text{when } \nu \neq \overline{\nu}. \tag{D.11}$$

Since $\|\nu_n\| = n^{-1} \sum_{i=1}^n |Y_i|$, the required exponential tightness immmediately follows from equation (B.5) and the continuity of $\psi_\vartheta(\tau)$ as a function of ϑ. To prove (D.10), it suffices to show that $\overline{\nu} \in F$ holds whenever $\overline{\nu} = \overline{g} P_V$, $\overline{g} \in C(V)$, $a < \overline{g} < b$, where a, b are as in Hypothesis $(H3)$. For such $\overline{\nu}$, (D.11) is satisfied with $f(t) = \gamma'(\overline{g}(t))$, $\gamma = \gamma_\vartheta$; this is trivial for $\nu \not\ll P_V$, while if $\nu = g P_V$ then (D.11) with the above f is equivalent to $B_\gamma(g, \overline{g}) > 0$, cf. (II.12) (with U replaced by V). ■

Proof of Theorem IV.1: Having available Theorem D.2, the claim follows as the similar one in the case of i.i.d. amplitudes, cf. Appendix B, applying Lemma D.1 with the same setting as there. The only changes are that the rate function is now the perhaps non-convex $\mathcal{I}(\nu) = \min_{\vartheta \in \Theta} \mathcal{I}_\vartheta(\nu)$, thus mixtures of measures belonging to the set of minimizers of $\mathcal{I}(\nu)$ subject to $\nu \in S_{\varphi, \mathcal{K}}$ need not belong to that set, and that to check hypothesis (D.4) of Lemma D.1, the following obvious consequence of (B.5) and (IV.2) is used:

$$\int_{\frac{1}{n}\sum_{i=1}^n |y_i| > C} \left(\frac{1}{n} \sum_{i=1}^n |y_i| \right) dF_n \leq$$
$$C n \exp\left\{ -n \left[C\tau - \max(\max_{\vartheta \in \Theta} \psi_\vartheta(\tau), \max_{\vartheta \in \Theta} \psi_\vartheta(-\tau)) - \log 2 \right] \right\}. \tag{D.12}$$

■

D.3. Proof of Theorem IV.2

By standard theory of the ME method, Q_n^{MEM} belongs to the exponential family through F_n spanned by the k-dimensional statistic $\frac{1}{n}\sum_{i=1}^n \varphi(t_i) y_i$, and to the analogy of equation (II.31)

$$Q_n^{MEM} = \frac{\exp(\sum_{i=1}^n \langle v_n, \varphi(t_i) \rangle y_i)}{E_{F_n}[\exp(\sum_{i=1}^n \langle v_n, \varphi(t_i) \rangle Y_i)]} \cdot F_n, \tag{D.13}$$

where v_n maximizers the analogue of (II.33) given by

$$\mathcal{H}_n(v) = \langle v, c \rangle - \ell_n(v), \tag{D.14}$$

$$\ell_n(v) = \frac{1}{n} \log[E_{F_n}(\exp(\sum_{i=1}^n \langle v, \varphi(t_i) \rangle Y_i))]. \tag{D.15}$$

We first show that for each $v \in \mathrm{I\!R}^k$

$$\lim_{n\to\infty} \ell_n(v) = l(v) = \max_{\vartheta\in\Theta} \int \psi_\vartheta(\langle v, \varphi(t)\rangle) dP_V(t). \tag{D.16}$$

Indeed, by Theorem D.2, (ν_n) satisfies a L.D.P. under F_n with good rate function $\mathcal{I}(\nu)$. Hence

$$\ell_n(v) = \frac{1}{n}\log\left[E_{F_n}(\exp(n\int\langle v, \varphi(t)\rangle d\nu_n(t)))\right]$$

converges, by Varadhan's Theorem ([17], p. 120) to

$$\begin{aligned}\sup_{\nu\in\mathcal{M}(V)}\left[\int\langle v,\varphi(t)\rangle d\nu(t) - \mathcal{I}(\nu)\right] &= \sup_{\vartheta\in\Theta}\sup_{\nu\in\mathcal{M}(V)}\left[\int\langle v,\varphi(t)\rangle d\nu(t) - \mathcal{I}_\vartheta(\nu)\right] \\ &= \sup_{\vartheta\in\Theta}\int \psi_\vartheta(\langle v,\varphi(t)\rangle)dP_V(t) = l(v),\end{aligned} \tag{D.17}$$

where the first equality holds by (IV.6) and the second one follows from (II.11) by convex duality. The fact that in (D.16) we may write max (rather than sup) follows by hypothesis $(H3)$ that insures the continuity of

$$l(v,\vartheta) = \int \psi_\vartheta(\langle v,\varphi(t)\rangle)dP_V(t) \tag{D.18}$$

as a function of $(v,\vartheta) \in \mathrm{I\!R}^k \times \Theta$. Next we show that

$$\mathcal{H}(v) = \langle v, c\rangle - l(v) \tag{D.19}$$

has a unique maximizer $\tilde{v}$, and that

$$v_n \to \tilde{v},\ \ell_n(v_n) \to l(\tilde{v}) \quad \text{as} \quad n\to\infty. \tag{D.20}$$

Indeed, $\max_v[\langle v,c\rangle - l(v,\vartheta)]$ is attained, by Theorem II.2. Thus the strictly concave functions $\langle v,c\rangle - l(v,\vartheta)$ go to $-\infty$ as $\|v\|\to\infty$, hence so does their minimum $\mathcal{H}(v)$, which is also strictly concave. This implies that $\mathcal{H}(v)$ has a unique maximizer $\tilde{v}$. Notice that the convergence in (D.16) must be uniform on each compact subset of $\mathrm{I\!R}^k$ ([37], p.50); this implies that the maximizer v_n of $\mathcal{H}_n(v)$ in (D.14) converges to the maximizer $\tilde{v}$ of $\mathcal{H}(v)$ in (D.19), and (D.20) holds.
The key to the proof of Theorem IV.2 is the following

Theorem D.4: Under the hypotheses of Theorem IV.2, if the random amplitudes $Y_1,\ldots,Y_n$ have joint distribution Q_n^{MEM} given by (D.13), the sequence (ν_n) of (IV.1) satisfies a L.D.P. with good rate function

$$\mathcal{J}(\nu) = \mathcal{I}(\nu) - \int\langle\tilde{v},\varphi(t)\rangle d\nu(t) + l(\tilde{v}), \qquad \tilde{v} = \lim_{n\to\infty} v_n. \tag{D.21}$$

Proof: Let R_n denote the distribution of ν_n under Q_n^{MEM}. Note first that, using (D.13), similar arguments as used for (B.5) and (D.12) give for any $C>0$, $\tau>0$ and sufficiently large n

$$\int_{\frac{1}{n}\sum_{i=1}^{n}|y_i|>C}\left(\frac{1}{n}\sum_{i=1}^{n}|y_i|\right)dQ_n^{MEM} \\ \leq Cn\exp\left\{-n\left[C\tau-k_n(\tau)+\ell_n(v_n)\right]\right\}, \tag{D.22}$$

where

$$k_n(\tau)= \\ =\max\left[\int\max_{\vartheta\in\Theta}\psi_\vartheta(\tau+\langle v_n,\varphi(t)\rangle)dP_V(t),\int\max_{\vartheta\in\Theta}\psi_\vartheta(-\tau+\langle v_n,\varphi(t)\rangle)dP_V(t)\right]-\log 2.$$

As $\|\nu_n\|=\frac{1}{n}\sum_{i=1}^{n}|Y_i|$, it follows from (D.23) and (D.20) that (R_n) is exponentially tight. Since (R_n) is exponentially tight and $\mathcal{J}(\nu)$ defined by (D.21) is lower semicontinuous (as $\mathcal{I}(\nu)$ is), to prove Theorem D.4 it suffices to check the lower bound (ii) in Section D-A for open sets $B\subset\mathcal{M}(V)$, and the upper bound (iii) for compact sets $B\subset\mathcal{M}(V)$ (cf. [17], p.8).
Now, by (D.13) and (D.15), we have for any $B\subset\mathcal{M}(V)$

$$\begin{aligned}\frac{1}{n}\log R_n(B) &= \frac{1}{n}\log\int_{\frac{1}{n}\sum_{i=1}^{n}y_i\delta_{t_i}\in B}(\exp\sum_{i=1}^{n}\langle v_n,\varphi(t_i)\rangle y_i)dF_n(y_1,\ldots,y_n)-\ell_n(v_n)\\ &= \frac{1}{n}\log E_{F_n}[\exp(n\int\langle v_n,\varphi(t)\rangle d\nu_n(t))1_B(\nu_n)]-\ell_n(v_n)\end{aligned} \tag{D.23}$$

where $1_B(\nu)=1$ if $\nu\in B$ and 0 if $\nu\notin B$. Since $v_n\to\tilde{v}$ implies that $\langle v_n,\varphi(t)\rangle\to\langle\tilde{v},\varphi(t)\rangle$ uniformly in t, it follows from (D.23) and (D.20) that

$$\frac{1}{n}\log R_n(B)=\frac{1}{n}\log E_{F_n}[\exp(n\int\langle\tilde{v},\varphi(t)\rangle d\nu_n(t))1_B(\nu_n)]-l(\tilde{v})+\epsilon_n(B) \tag{D.24}$$

where $\epsilon_n(B)\to 0$ as $n\to\infty$ provided that

$$\sup_{\nu\in B}\|\nu\|<\infty. \tag{D.25}$$

Since (ν_n) satisfies a L.D.P. under F_n with good rate function $\mathcal{I}$, by Theorem D.2, a variant of Varadhan's Theorem ([17], p. 123, Exercise 4.3.11) may be invoked to obtain that the liminf or limsup of the $\frac{1}{n}\log E_{F_n}$ term in (D.24) is bounded below or above by

$$\sup_{\nu\in B}\left\{\int\langle\tilde{v},\varphi(t)\rangle d\nu(t)-\mathcal{I}(\nu)\right\},$$

according as B is open or closed, respectively. Hence (D.24) implies (ii) and (iii) in Section D-A for those open, respectively closed sets B that satisfy (D.25). But the validity of (ii) for open sets satisfying (D.25) implies the same for each open B, while the closed sets B satisfying (D.25) are exactly the compact sets. Thus the proof of

Theorem D.4 is complete.

Returning to the proof of Theorem IV.2, we shall use Lemma D.1 as in the proofs of Theorems II.5 and IV.1, with $E = \mathcal{M}(V)$, $Z_n = \nu_n$ as there, but now setting $F = E = \mathcal{M}(V)$. Then the conditional distribution $R_n(\cdot \mid F)$ in the Lemma is R_n itself, and F_0 in (D.3) is the set

$$\mathcal{N} = \{\nu \ : \ \mathcal{J}(\nu) = 0\}. \tag{D.26}$$

The first conclusion is that the sequence (R_n) accumulates on the compact set $\mathcal{N}$. Next, for arbitrary $f \in C(V)$, $s(\nu) = \int f d\nu$ is a continuous function on $E = \mathcal{M}(V)$ that meets condition (D.4), on account of (D.23). Hence by (D.5), $R_{n_i} \to R^*$ implies that

$$\int f(t) dG_{n_i}^{MEM}(t) = E(\int f(t) d\nu_{n_i}(t))$$

converges to

$$\int_{\mathcal{N}} (\int f d\nu) dR^*(\nu) = \int f dG^*$$

where G^* is the R^*-mixture of measures $\nu \in \mathcal{N}$. This proves that $R_{n_i} \to R^*$ implies $G_{n_i}^{MEM} \to G^*$, and to complete the proof of Theorem IV.2 it remains to show that $\nu = G^*$ minimizes $\tilde{\mathcal{I}}(\nu)$ subject to $\int \varphi d\nu = c$.

Since $\int \varphi dG_n^{MEM} = c$ by definition, $\int \varphi dG^* = c$ clearly holds since $G_{n_i}^{MEM} \to G^*$. Finally, let $\tilde{\mathcal{J}}$ denote the affine regularization of the functional $\mathcal{J}$ defined by (D.21). Then

$$\tilde{\mathcal{J}}(\nu) = \tilde{\mathcal{I}}(\nu) - \int \langle \tilde{v}, \varphi(t) \rangle d\nu(t) + l(\tilde{v})$$

and hence

$$\tilde{\mathcal{I}}(\nu) = \tilde{\mathcal{J}}(\nu) + \langle \tilde{v}, c \rangle - l(\tilde{v}) \ \text{ if } \ \int \varphi d\nu = c. \tag{D.27}$$

As $\tilde{\mathcal{J}}$ attains its minimum 0 on a closed convex set that contains $\mathcal{N}$ of (D.26), $\tilde{\mathcal{J}}$ vanishes on the convex closure of $\mathcal{N}$; in particular $\tilde{\mathcal{J}}(G^*) = 0$. Hence (D.27) gives that G^* is a minimizer of $\tilde{\mathcal{I}}(\nu)$ subject to $\int \varphi d\nu = c$, as claimed. ∎

References

[1] Aldous D J, *Exchangeability and related topics*, Ecole d'Eté de Probabilités de St Flour XIII 1983. Lecture notes in math., 1985.

[2] Ali S M, Silvey S D, "A general class of coefficients of divergence of one distribution from another," *J. Roy. Stat. Soc.,* Ser. B, vol. 28, pp. 131–142, 1966.

[3] Borwein J M, Lewis A S, "Duality relationships for entropy-like minimization problems," *SIAM Journal of Control and Optimization*, vol. 29, pp. 325–338, 1991.

[4] Borwein J M, Lewis A S, "Partially-finite programming in L_1 and the existence of maximum entropy estimates," *SIAM Journal of Optimization*, vol. 3, pp. 248–267, May 1993.

[5] Bregman L M, "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming," *U.S.S.R. Comput. Math. Phys.*, vol. 7, pp. 200–217, 1967.

[6] Brockett P L, Charnes A, Paick K H, "Information-theoretic approach to unimodal density estimation," *IEEE Transactions on Information Theory*, vol. 41, pp. 825–829, 1995.

[7] Burg J P, *Ph.D Thesis*, Standford University, 1975.

[8] Csiszár I, "Eine informationtheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markovschen Ketten," *Publ. Math. Inst. Hungar. Acad. Sci.*, vol. 8, pp. 85–108, 1963.

[9] Csiszár I, "Information-type measures of difference of probability distributions and indirect observations," *Studia Sci. Math. Hungar.*, vol. 2, pp. 299–318, 1967.

[10] Csiszár I, "Sanov property, generalized I-projections and a conditional limit theorem," *The Annals of Probability*, vol. 12, pp. 768–793, 1984.

[11] Csiszár I, "An extended maximum entropy principle and a Bayesian justification," in *Bayesian Statistics 2* (Bernardo J M, ed.) pp. 83–98, Elsevier, 1985.

[12] Csiszár I, "Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems," *The Annals of Statistics*, vol. 19, pp. 2032–2066, 1991.

[13] Csiszár I, "Generalized projection for non negative functions," *Acta Math. Hungar.*, vol. 68, pp. 161–185, 1995.

[14] Csiszár I, "Maxent, mathematics, and information theory," in *Maximum Entropy and Bayesian Methods* (Hanson K M, Silver R N, eds.) pp. 35–50, Kluwer, 1996.

[15] Dacunha-Castelle D, Gamboa F, "Maximum d'entropie et problème des moments," *Annales de l'I.H.P. Proba.*, vol. 26, pp. 567–596, 1990.

[16] De Finetti B, "La prévision: ses lois logiques, ses sources subjectives," *Annales de l'I.H.P. Série B*, vol. 7, pp. 1–68, 1937.

[17] Dembo A, Zeitouni O, *Large deviations techniques and applications*, Jones and Bartlett, Boston, 1993.

[18] Dinwoodie I H, Zabell S L, "Large deviations for exchangeable random vectors," *The Annals of Probability*, pp. 1147–1166, 1992.

[19] Feller W, *An introduction to probability theory and its applications*, John Wiley & Sons, 1968.

[20] Gamboa F, "New Bayesian methods for ill posed problems," To appear in Statistics and Decisions.

[21] Gamboa F, Gassiat E, "Maximum d'entropie et problème des moments cas multidimensionnel," *Probability and Mathematical Statistics*, vol. 12, pp. 67–83, 1991.

[22] Gamboa F, Gassiat E, "The maximum entropy method on the mean: Applications to linear programming and superresolution," *Mathematical Programming, Series A*, vol. 66, pp. 103–122, 1994.

[23] Gamboa F, Gassiat E, "Bayesian methods for ill posed problems," *The Annals of Statistics*, vol. 25, pp. 328–350, 1997.

[24] Jaynes E T, *Papers on Probability, Statistics and Statistical Physics*, (Rosenkrantz R D, ed.), Reidel, Dordrecht, 1983.

[25] Jones L, Byrne C, "General entropy criteria for inverse problems with application to data compression, pattern classification and cluster analysis," *IEEE Transactions on Information Theory*, vol. 36, pp. 23–30, 1990.

[26] Kapur J N, "Maximum entropy models in science and engineering," *John Wiley*, New-York, 1989.

[27] Kemperman J H, "Moment problems with convexity conditions," in *Optimizing methods in statistics*, pp. 115–178. Academic Press, 1971.

[28] Kullback S, *Information Theory and Statistics*, John Wiley and Sons, 1959.

[29] Le Besnerais G, Bercher J F, Demoment G, "A new look at entropy for solving linear inverse problems," *IEEE Transactions on Information Theory*, vol. (to appear), 1999.

[30] Léonard C, "Minimizer of energy functionals," Preprint, Orsay, 1996.

[31] Lewis A S, "Consistency of moment systems," *Canadian J. of Math.*, vol. 47, pp. 995–1006, 1995.

[32] Lewis J, Pfister C, Sullivan W, "Entropy, concentration of probability, and conditional limit theorems," *Markov Processes and Related Fields*, vol. 1., pp. 319–386, 1995.

[33] Liese F, Vajda I, *Convex Statistical Distances*, Teubner, 1987.

[34] Maréchal P, Lannes A, "Unification of some deterministic and probabilistic methods for the solution of linear inverse problems via the principle of the maximum entropy on the mean," *Inverse Problems*, vol. 13, pp. 135–151, 1997.

[35] Navaza J, "On the maximum entropy estimate of electron density function," *Acta Crist*, pp. 232–244, 1985.
[36] Rockafellar R T, "Integrals which are convex functionals," *Pacific J. Math.*, vol. 24, pp. 525–539, 1968.
[37] Rockafellar R T, *Convex analysis*, Princeton University Press, 1970.
[38] Shore J, Johnson R, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Transactions on Information Theory*, vol. 26, pp. 26–37, 1980.
[39] Skilling J, "Fundamental of MaxEnt in data analysis," in *Maximum entropy in action*, (Buck B, Macaulay V A, eds.) pp. 19–40, Oxford Science Publications, Oxford, 1991.
[40] Teboulle M, Vajda I, "Convergence of best ϕ-entropy estimates," *IEEE Transactions on Information Theory*, vol. 39, pp. 297–301, 1993.
[41] Urban B, "Retrieval of atmospheric thermodynamical parameters using satelite measurements with maximum entropy method," *Inverse Problems*, vol. 12, pp. 779–796, 1996.

Imre Csiszár (M'87-SM'91-F'92) was born in Miscolc, Hungary, on February 7, 1938. He received the diploma in mathematics from the L. Eötvös University, Budapest, Hungary, in 1961, and the Doctor of Mathematical Science degree from the Hungarian Academy of Sciences, Budapest, in 1977.

Since 1961, he has been with the Mathematical Institute of the Hungarian Academy of Sciences, and has been Head of the Information Theory Group there since 1968. He is also Professor of Mathematics at the L. Eötvös University. He has been Visiting Professor at various universities including The University of Bielefeld, Germany, the University of Maryland at College Park, Stanford University, Stanford, CA, University of Virginia, Charlottesville, and the University of Tokyo, Japan. His research interests are centered on Shannon theory and on application of information theory in mathematics, primarily in probability and statistics. He is coauthor of the book *Information Theory: Coding Theorems for Discrete Memoryless Systems* (New York: Academic Press, 1981).

Dr. Csiszár is President of the Hungarian Mathematical Society. In 1990 he was elected Corresponding Member and in 1995 Regular Member of the Hungarian Academy of Sciences. He was the recipient of the 1988 Paper Award Of the IEEE Information Theory Society, the Academy Award for Interdisciplinary Research of the Hungarian Academy of Sciences in 1989, and the Shannon Award of the IEEE Information Theory Society in 1996.

Fabrice Gamboa received the Ph. D degree in mathematics in Orsay University (France) in 1989. He is actually with Amiens University (France) and Ecole Polytechnique. His research interests lie in the area of large deviations applied to statistics and signal processing.

Elisabeth Gassiat received both the *Diplôme de l'École Polytechnique* (Palaiseau France) and the Ph. D degree in mathematics in Orsay University (France), in 1984 and 1988 respectively. She is actually with Paris Sud University (France) where she conducts a group on signal processing and statistics. Her research interests lie in the areas of statistics, applied probability and signal processing.