Ann Inst Stat Math (2012) 64:1139–1160
DOI 10.1007/s10463-012-0350-4

Converting information into probability measures with the Kullback–Leibler divergence

Pier Giovanni Bissiri · Stephen G. Walker

Received: 25 June 2010 / Revised: 15 June 2011 / Published online: 13 March 2012

Abstract This paper uses a decision theoretic approach for updating a probability measure representing beliefs about an unknown parameter. A cumulative loss function is considered, which is the sum of two terms: one depends on the prior belief and the other one on further information obtained about the parameter. Such information is thus converted to a probability measure and the key to this process is shown to be the Kullback–Leibler divergence. The Bayesian approach can be derived as a natural special case. Some illustrations are presented.

Keywords Bayesian inference · Posterior distribution · Loss function · Kullback–Leibler divergence · g-divergence

1 Introduction and preliminaries

Unstructured information, denoted by I, is used by a Bayesian to construct a prior distribution about a parameter of interest, say θ_0, and stochastic observations can then be used to update the prior to the posterior, see e.g. Bernardo and Smith (1994). However, if future unstructured or non-stochastic information is subsequently available, there is no formal procedure for updating the belief probability measure. This paper works on the idea of updating a belief probability measure using unstructured information, of the type a Bayesian would use to construct a prior distribution, where the only

P. G. Bissiri (✉)
Dipartimento di Statistica, Università degli Studi di Milano-Bicocca, Edificio U7,
via Bicocca degli Arcimboldi 8, 20126 Milan, Italy
e-mail: pier.bissiri@unimib.it

S. G. Walker
SMSAS, University of Kent, Canterbury, Kent CT2 7NZ, UK
e-mail: S.G.Walker@kent.ac.uk

requirement is to be able to connect the information and parameter of interest via a loss function. If information I is represented by pieces from a finite set $I = (I_0, I_1, \ldots, I_n)$ then the idea here is that I_j and I_k are, for $j \neq k$, disjoint, in the sense that nothing about I_k could be inferred from knowledge of I_j, and vice versa.

Hence, we are interested in taking an initial or prior distribution $\pi_0(\theta)$, given information I, to a posterior distribution. We point out that such information need not be stochastic: "word of mouth" information can also be used and such information is often available from experts in a Bayesian approach. Therefore, here "information" is just knowledge that is connected in some way, to be explained exactly how later on, to θ_0. Such knowledge can be of different kinds. For instance, the information may consist of learning that θ_0 is a value close to zero. More examples will be considered later.

For us, as we have indicated, information pieces about θ_0 are represented by $(I_0, I_1, \ldots, I_n)$, which for any $0 \leq m \leq n$ is equivalent to $(\pi_m, I_{m+1}, \ldots, I_n)$ where π_m is the probability measure constructed from $(I_0, \ldots, I_m)$. We translate the information to the probability measure via the use of cumulative loss functions; specifically, and to be defined further later on, we use a loss function $L(\nu, I^{(n)}, \pi_0)$, where $I^{(n)}$ stands for $(I_1, \ldots, I_n)$, ν is the action and hence π_n minimizes $L(\nu, I^{(n)}, \pi_0)$, i.e. $L(\pi_n, I^{(n)}, \pi_0) = \min_\nu L(\nu, I^{(n)}, \pi_0)$, for every $n \geq 1$.

To approach statistical problems in a decision theoretical framework is a well-established practice, see for instance, Berger (1980). Therefore, our plan is to use decision theory to select a probability measure which represents information about a particular object. The decision space is the space of probability measures which are absolutely continuous with respect to π_0.

The loss $L(\nu, I^{(n)}, \pi_0)$ will be required to satisfy a certain type of coherence. By this coherence we mean that no matter which m we use, the problem based on loss functions is the same; that is the minimizer of $L(\nu, I^{(m+1,\,n)}, \pi_m)$, where $I^{(m+1,\,n)}$ stands for $(I_{m+1}, \ldots, I_n)$, is the same for all $0 \leq m < n$.

Let the unknown quantity θ_0 belong to some space Θ. Moreover, assume that I_0 belongs to some set $\mathcal{I}_0$ and I_i belongs to some set $\mathcal{I}$ for every $i \geq 1$. Let $H(\nu, I)$ denote the loss when the information I in $\mathcal{I} \cup \mathcal{I}_0$ is summarized by a probability measure ν on Θ. The prior π_0 is defined as the probability measure on Θ that minimizes the loss $H(\nu, I_0)$ corresponding to the initial piece of information I_0, i.e. $H(\pi_0, I_0) = \min_\nu H(\nu, I_0)$.

Then consider the following cumulative loss function:

$$L(\nu,\ I^{(n)},\ \pi_0)\ :=\ \sum_{i=1}^{n} H(\nu,\ I_i) + l(\nu, \pi_0), \tag{1}$$

where ν is a probability measure on Θ. Here, the loss l is chosen to be the g-divergence D_g defined by

$$D_g(Q_1, Q_2) = \int g\left(\frac{\mathrm{d}Q_1}{\mathrm{d}Q_2}\right) \mathrm{d}Q_2, \tag{2}$$

for any couple (Q_1, Q_2) of probability measures such that $Q_1 \ll Q_2$, where g is a convex function from $(0, \infty)$ into $\mathbb{R}$ such that $g(1) = 0$. The class (2) of probability divergences has been introduced and studied independently by Ali and Silvey (1966) and Csiszár (1967). We use a g-divergence as the loss function, based on certain characterizations of such divergences, including decomposability and information monoticity. See, for example, Amari (2009). Such characteristics are obviously important for a loss function.

Choosing $l(\nu, \pi_0)$ to be $D_g(\nu, \pi_0)$, then ν is required to be absolutely continuous with respect to π_0. Indeed, the updated probability π_n should be zero on every event whose prior π_0 probability is zero.

The loss H is most suitably taken in integral form, i.e. the average loss

$$H(\nu,\ I) = \int_{\Theta} h(\theta, I)\, \nu(\mathrm{d}\theta), \tag{3}$$

for every I in $\mathcal{I}$ and every $\nu \ll \pi_0$ for which such an integral exists. Since θ_0 is the object of interest, it is more realistic to be able to construct $h(\theta, I)$, i.e. a loss relating information to θ_0. Then, obviously, the appropriate $H(\nu, I)$ becomes the average loss. In this way, the loss (1) becomes

$$L(\nu) := L(\nu,\ I^{(n)},\ \pi_0)\ :=\ \int_{\Theta} h_n(\theta,\ I^{(n)})\, \nu(\mathrm{d}\theta)\ +\ D_g(\nu, \pi_0), \tag{4}$$

where

$$h_n(\theta,\ I^{(n)})\ =\ \sum_{i=1}^{n} h(\theta,\ I_i). \tag{5}$$

The loss (4) is defined on the class of probability measures ν on Θ that are absolutely continuous with respect to π_0 and such that the integral $\int_{\Theta} h_n(\theta,\ I^{(n)})\, \nu(\mathrm{d}\theta)$ exists. Denote such a class by $\mathscr{P}(\Theta, \pi_0, I^{(n)})$.

The loss defined by (5) is cumulative and moreover is symmetric with respect to $(I_1, \ldots, I_n)$. Therefore, the order in which the single pieces of information are given is not relevant.

A difference to the Bayesian approach here is that we are not demanding that the pieces of information are specifically an independent sample from a probability density function indexed by the parameter θ, say $f(x, \theta)$. Or even that θ is a parameter of a probability density function. Nevertheless, if it is, and information are samples from such a density, then we can recover the Bayesian approach by taking D_g to be the Kullack–Leibler divergence and $h(\theta, I) = h(\theta, X) = -\log f(X, \theta)$, which is the self-information loss function and the most commonly used and "honest" loss function in such cases.

When combining (4) and (5), our loss becomes:

$$L(\nu) := L(\nu, I^{(n)}, \pi_0) = \sum_{i=1}^{n} \int h(\theta, I_i)\, \nu(\mathrm{d}\theta) + D_g(\nu, \pi_0). \tag{6}$$

A loss of the form given by (6) has been previously considered by Walker (2006) for general parametric models and by Bissiri and Walker (2010) for Bernoulli observations. These authors took $g(x) = x \log(x)$ for $x > 0$, so that D_g turns out to be the Kullback–Leibler divergence; so $D_g(\nu, \pi_0) = \int \nu(\mathrm{d}\theta) \log\{\nu(\mathrm{d}\theta)/\pi_0(\mathrm{d}\theta)\}$.

The next proposition provides the solution for the general case, where D_g is the Kullback–Leibler divergence.

Proposition 1 *Let D_g be the Kullback–Leibler divergence. There exists a probability measure π_n in $\mathscr{P}(\Theta, \pi_0, I^{(n)})$ that minimizes the loss L in (4) if and only if*

$$\int_\Theta e^{-h_n(\theta;\, I^{(n)})}\, \pi_0(\mathrm{d}\theta) \;<\; \infty. \tag{7}$$

If such a π_n exists, then it is unique and a version of $\pi_n(\mathrm{d}\theta)/\pi_0(\mathrm{d}\theta)$ is equal to

$$\exp\{-h_n(\theta;\; I^{(n)})\} \Big/ \int_\Theta \exp\{-h_n(\tau;\; I^{(n)})\}\, \pi_0(\mathrm{d}\tau). \tag{8}$$

In a statistical problem, the place of $I_1, I_2, \ldots$ is taken by the observations $X_1, X_2, \ldots$, which are stochastic, i.e. random variables on a probability space $(\Omega, \mathscr{F}, P)$, where the unknown probability measure P is assumed to belong to a family of distributions $\{P_\theta : \theta \in \Theta\}$ and θ_0 plays the role of the true value of the parameter. The support $\Theta_0 \subset \Theta$ of the prior π_0 is sometimes chosen to be an infinite-dimensional space so that the model turns out to be non-parametric.

Generally, the observations $X_1, X_2, \ldots$ are taken identically distributed and the law of each observation is absolutely continuous with respect to some measure μ. Denote by $f(x, \theta)$ the density of X_1 with respect to μ under the probability measure P_θ, for every θ in Θ_0. If

$$h(\theta, x) = -\log f(x, \theta) \tag{9}$$

for every θ in Θ, then the minimizing probability distribution (8) is the classical posterior obtained under the assumption that the observations $(X_i)_{i \geq 1}$ are conditionally i.i.d. given θ. This is tantamount to taking

$$h_n(\theta;\; x_1, \ldots, x_n) \;=\; -\sum_{i=1}^{n} \log f(x_i, \theta), \tag{10}$$

for each $\theta \in \Theta$. It is known that if a maximum likelihood estimator exists for the model Θ_0, then it has to minimize (10). Moreover, in a Bayesian setting, the (unconditional) distribution of the sequence of observations turns out to be exchangeable. Exchangeability expresses the idea that the order in which the observations are sampled does not provide any information about θ_0. In the present paper, this idea is generalized to the case where the observations are replaced by non-stochastic pieces of informations $I^{(n)}$ by taking a function h_n that is symmetric with respect to $I_1, \ldots, I_n$ as in (5). If the observed information is stochastic, then the most natural choice for h is (9). In

fact, this choice yields the only smooth, proper and local utility function, see Bernardo (1979), Good (1952), and also Aczel and Pfanzagl (1966).

The layout of the paper is as follows: Sect. 2 contains a statement of the key preliminary results and Sect. 3 the main result which determines the necessity of the Kullback–Leibler divergence from the class of g-divergences when a coherence property is required. Section 4 provides two examples in which we believe it is useful to adopt the approach presented in this paper, Sect. 5 concludes with a brief discussion and the proofs are contained in the Appendix.

2 The minimization problem

To begin with, notice that if $h_n(\theta\,;\, I^{(n)})$ is a constant function of θ, π_0-a.s., then the prior minimizes L. In fact, in such case, $L(\nu;\ I^{(n)};\ \pi_0) = D_g(\nu, \pi_0)$. The following proposition deals with the opposite implication.

Proposition 2 *Let π_0 be a probability measure on Θ and let g be differentiable at one. If π_0 belongs to $\mathscr{P}(\Theta, \pi_0, I^{(n)})$ and*

$$L(\pi_0,\ I^{(n)},\ \pi_0) = \min_{\nu \in \mathscr{P}(\Theta,\pi_0),\, I^{(n)}} L(\nu,\ I^{(n)},\ \pi_0), \tag{11}$$

then $h_n(\theta\,;\ I^{(n)})$ is a constant map of θ, π_0 - a.s.

Before proceeding, recall that being convex, g admits left and right derivatives and denote by g'_- and g'_+ the left and the right derivative of g, respectively.

Theorem 1 *If there is a probability measure $\pi_n \in \mathscr{P}(\Theta, \pi_0, I^{(n)})$ such that*

$$\operatorname*{ess\,inf}_{\theta\in\Theta} h_n(\theta; I^{(n)}) + g'_+\left(\frac{\mathrm{d}\pi_n}{\mathrm{d}\pi_0}(\theta)\right) \geq \operatorname*{ess\,sup}_{\theta\in\Theta} h_n(\theta; I^{(n)}) + g'_-\left(\frac{\mathrm{d}\pi_n}{\mathrm{d}\pi_0}(\theta)\right), \tag{12}$$

then π_n satisfies

$$L(\pi_n) = \min_{\nu \in \mathscr{P}(\Theta,\pi_0,I^{(n)})} L(\nu). \tag{13}$$

At this stage, recall that a subderivative ϕ of g is a function defined on $(0, \infty)$ such that $g'_-(x) \leq \phi(x) \leq g'_+(x)$, for every positive x. The inequality (12) is tantamount to the existence of a subderivative ϕ such that $h_n + \phi(\mathrm{d}\pi_n/\mathrm{d}\pi_0)$ is a constant map, π_0-a.s.

Before proceeding, some additional notation and a few remarks are useful. If g is differentiable, denote by g' its derivative and denote $\lim_{x\uparrow\infty} g'(x) = g'(\infty)$, recalling that this limit exists since g' is non-decreasing by convexity of g. If g is strictly convex and E denotes the image of g', then $g' : (0, \infty) \to E$ is strictly increasing and therefore there exists its inverse, which will be denoted by G, defined on E.

The following proposition gives a necessary and sufficient condition for the solution of the minimization problem in the case g is differentiable.

Proposition 3 *Let g be differentiable. A probability measure π_n in $\mathscr{P}(\Theta, \pi_0, I^{(n)})$ minimizes L if and only if $h_n(\theta, I^{(n)}) + g'(\mathrm{d}\pi_n(\theta)/\mathrm{d}\pi_0(\theta))$ is constant, π_0-a.s.*

Corollary 1 *Let g be differentiable and strictly convex and denote by $G : E \to (0, \infty)$ the inverse function of g' and by E the image of g'.*

Hence, there exists a probability measure $\pi_n \in \mathscr{P}(\Theta, \pi_0, I^{(n)})$ minimizing L if and only if there exists a real constant c such that $c - h_n(\theta, I^{(n)})$ belongs to E for every θ in Θ and $\int_\Theta G(c - h_n(\theta, I^{(n)}))\,\mathrm{d}\pi_0$ exists and is equal to one.

Moreover, if π_n exists, then it is unique and satisfies:

$$\pi_n(A) = \int_A G(c - h_n(\theta, I^{(n)}))\ \pi_0(\mathrm{d}\theta), \tag{14}$$

for every measurable subset A of Θ.

For different relevant g-divergences, the function g is strictly convex and Corollary 1 can be applied. The Kullback–Leibler divergence was already considered by Proposition 1. Here are a few other examples:

(i) If D_g is the χ^2-distance ($g(x) = (x-1)^2$), then there exists a probability measure π_n in $\mathscr{P}(\Theta, \pi_0, I^{(n)})$ that minimizes L if and only if $h_n(\theta,\ I^{(n)}) < 2 + \int_\Theta h_n(\theta,\ I^{(n)})$ for every θ in Θ, the integral $\int_\Theta |h_n(\theta,\ I^{(n)})|\,\pi_0(\mathrm{d}\theta)$ is finite and $\mathrm{d}\pi_n(\theta)/\mathrm{d}\pi_0(\theta) = 1 + \frac{1}{2}\int_\Theta h_n(\tau,\ I^{(n)})\,\pi_0(\mathrm{d}\tau) - \frac{1}{2}h_n(\theta,\ I^{(n)})$.

(ii) If D_g is the square of the Hellinger metric ($g(x) = (\sqrt{x}-1)^2$) and a minimizing probability measure π_n exists for L, then $\mathrm{d}\pi_n(\theta)/\mathrm{d}\pi_0(\theta) = (h_n(\theta,\ I^{(n)}) + c)^{-2}$ for some real constant c.

(iii) If D_g is the power divergence ($g(x) = \{x^\alpha - \alpha(x-1) - 1\}/\{\alpha(\alpha-1)\}$ where $\alpha \notin \{0, 1\}$) and a minimizing probability measure π_n exists for L, then $\mathrm{d}\pi_n(\theta)/\mathrm{d}\pi_0(\theta) = \{1 + (\alpha-1)(c - h_n(\theta,\ I^{(n)}))\}^{1/(\alpha-1)}$ for some real constant c. Note that $\alpha = 1/2$ gives (ii) and $\alpha = 2$ gives (i).

The following proposition provides sufficient conditions for the existence and uniqueness of a probability measure π_n in $\mathscr{P}(\Theta, \pi_0, I^{(n)})$ that minimizes L.

In what follows, given a map $\psi : \Theta \to \mathbb{R}$, denote by ess sup ψ and ess inf ψ the essential supremum and the essential infimum of ψ with respect to π_0, respectively. Given two real numbers x and y, let $x \wedge y$ be their minimum.

Proposition 4 *Assume that g is a strictly convex and differentiable function such that $g(1) = 0$ and that*

$$\begin{aligned} &\operatorname*{ess\,sup}_{\theta\in\Theta} h_n(\theta,\ I^{(n)}) - \operatorname*{ess\,inf}_{\theta\in\Theta} h_n(\theta,\ I^{(n)}) \\ &\quad\le (g'(1) - g'_+(0)) \wedge (g'(\infty) - g'(1)). \end{aligned} \tag{15}$$

Moreover, assume that $h_n(\cdot, I^{(n)})$ is essentially bounded (with respect to π_0), i.e.

$$\operatorname*{ess\,inf}_{\theta\in\Theta} h_n(\theta,\ I^{(n)}) > -\infty, \quad \operatorname*{ess\,sup}_{\theta\in\Theta} h_n(\theta,\ I^{(n)}) < \infty.$$

Hence, there exists a unique probability measure $\pi_n \in \mathscr{P}(\Theta, \pi_0, I^{(n)})$ *such that* (13) *holds true. By Corollary 1,* π_n *satisfies* (14)*, for some real constant c.*

3 The coherence property

Given the prior π_0 and the sequence of pieces of information $(I_n)_{n\geq 1}$, denote by $\mathscr{H}(\pi_0)$ the class of measurable functions $h : \Theta \times \mathcal{I} \to \mathbb{R}$ such that $h(\cdot, I) : \Theta \to \mathbb{R}$ is measurable for every $I \in \mathcal{I}$ and that a probability measure minimizing the loss $L(\nu) := L(\nu, I^{(n)}, \pi_0)$ defined by (6) exists for every $n \geq 1$. In other words, $\mathscr{H}(\pi_0)$ is the class of all available functions that the loss h can match. Moreover, denote by $\mathscr{P}(h, I^{(n)})$ the class of probability measures π on Θ such that there is a probability measure minimizing $L(\cdot, I^{(n)}, \pi)$.

At this stage, it is possible to define a rule to update the probability measure on Θ representing our believes on the base of the observed information. Formally, given the information $I^{(n)} = (I_1, \ldots, I_n)$ and $n \geq 1$, there is an operator $U_n(\cdot, I^{(n)}) := U_n(\cdot, I^{(n)}, h)$ from $\mathscr{P}(h, I^{(n)})$ into the space of probability measures on Θ such that:

$$L(U_n(\pi, I^{(n)}), I^{(n)}, \pi) = \min_{\nu \in \mathscr{P}(\Theta, \pi)} L(\nu, I^{(n)}, \pi), \tag{16}$$

for every probability measure π in $\mathscr{P}(h, I^{(n)})$. So, the operator $U_n(\pi, I^{(n)})$ updates any probability measure π in $\mathscr{P}(h, I^{(n)})$ on the basis of the information $I^{(n)}$. Clearly, the sequence of operators $(U_n(\cdot, I^{(n)}))_{n\geq 1}$ has to satisfy a natural coherence property. Indeed, given $n, m \geq 1$ and a probability measure π on Θ, updating π on the basis of $I^{(n+m)} = (I_1, \ldots, I_{n+m})$ and updating $U_n(\pi, I^{(n)})$ on the basis of $I^{(n+1, n+m)} = (I_{n+1}, \ldots, I_{n+m})$, should yield the same probability measure. One can easily verify that such coherence property is satisfied if D_g is the Kullback–Leibler divergence and therefore $U_n(\pi_0, I^{(n)})$ is given by (8). The following theorem shows that such coherence property is satisfied only if D_g is the Kullback–Leibler divergence.

Theorem 2 *Let* $g : (0, \infty) \to \mathbb{R}$ *be a convex function such that* $g(1) = 0$ *and define the loss L by* (6)*. Moreover, assume that for every two integer* $m, n \geq 1$*, the following coherence condition*

$$U_{n+m}(\pi_0, I^{(n+m)}) = U_m(U_n(\pi_0, I^{(n)}), I^{(n+1, n+m)}) \tag{17}$$

holds true for every sequence $(I_n)_{n\geq 1} \in \mathcal{I}^\infty$*, every prior* π_0 *on* Θ *and every map* $h : \Theta \times \mathcal{I} \to \mathbb{R}$ *in* $\mathscr{H}(\pi_0)$*.*

Then D_g *is the Kullback–Leibler divergence, i.e.* $D_g(\nu, \mu) = k \int_\Theta \ln(\mathrm{d}\nu/\mathrm{d}\mu)\mathrm{d}\nu$*, for some positive constant k. Therefore, a map* $h : \Theta \times \mathcal{I} \to \mathbb{R}$ *belongs to* $\mathscr{H}(\pi_0)$ *if and only if*

$$\int_\Theta e^{-k\, h(\theta; I)} \pi_0(\mathrm{d}\theta) < \infty, \tag{18}$$

for every I in $\mathcal{I}$, and

$$U_n(\pi_0,\ I^{(n)})(A) = \frac{\int_A e^{-k\,h_n(\theta,\,I^{(n)})}\pi_0(\mathrm{d}\theta)}{\int_\Theta e^{-k\,h_n(\tau,\,I^{(n)})}\,\pi_0(\mathrm{d}\tau)}, \tag{19}$$

holds true for every measurable subset A of Θ.

It is important that (17) holds for every possible loss function h, i.e. for every h in the class $\mathscr{H}(\pi_0)$. In fact, we need to take into consideration any possible information I which the loss $h(\theta, I)$ could be based on.

A relevant special case for the updating operator U_n defined by (19) is given by the posterior distribution in a Bayesian dominated model. Indeed, if $I^{(n)}$ is a sample of n i.i.d. random variables $X_1, \ldots, X_n$ with common density f_{θ_0} and π_0 is the prior for the unknown parameter θ_0, then (19) is the posterior distribution, provided that h is the self-information loss function, i.e. the opposite of the log-likelihood function.

4 Illustrations

In this section, we provide two examples in which we believe it is useful to adopt the approach presented in this paper. The first one deals with stochastic information, whereas the second one deals with non-stochastic information. Both are based on the results in Sect. 3 and hence both use the Kullback–Leibler divergence.

Example 1 In this example, we consider stochastic information about the parameter of interest using loss functions that are used in robust estimation such as M-estimators and are not based on any density function. Let $h(\theta; X)$ be the loss in adopting θ as the true value of the parameter in the presence of information X. Assume that the observed information is stochastic and consists of n i.i.d. observations $X_1, \ldots, X_n$. Under these circumstances, minimization of the loss

$$\sum_{i=1}^{n} h(\theta, X_i) \tag{20}$$

should yield a reasonable frequentist estimator of the parameter. Estimators minimizing a loss of that form are called M-estimators and their usage has been motivated by robustness reasons. One can take, for instance, $h(\theta, X_i) = (\theta - X_i)^2$ if θ_0 is the mean of the probability distribution from which the X_i's are coming. If it is the median, then one can take $h(\theta, X_i) = |X_i - \theta|$. For more examples of M-estimators and more details about them, see Huber and Ronchetti (2009) and references cited therein.

From a Bayesian point of view, the loss (20) can be used to update a prior distribution π_0 of θ on the basis of the information given by $(X_1, \ldots, X_n)$ according to our approach, i.e. by minizing the loss

$$\tilde{L}(\nu) = \sum_{i=1}^{n} \int_\Theta h(\theta,\ X_i)\,\nu(\mathrm{d}\theta)\ +\ D_g(\nu, \pi_0). \tag{21}$$

The distribution π_n minimizing (21) is given by

$$\mathrm{d}\pi_n/\mathrm{d}\pi_0(\theta) = \exp\left(-\sum_{i=1}^{n} h(\theta, X_i)\right) \Big/ \int_\Theta \exp\left(-\sum_{i=1}^{n} h(\tau, X_i)\right) \pi_0(\mathrm{d}\tau).$$

In this way, one can obtain a Bayesian estimator that is the counterpart of the M-estimator and such an estimator is given by the mean of the probability distribution π_n, i.e.

$$\int_\Theta \theta \exp\left\{-\sum_{i=1}^{n} h(\theta,\ X_i)\right\} \pi_0(\mathrm{d}\theta) \Big/ \int_\Theta \exp\left\{-\sum_{i=1}^{n} h(\theta,\ X_i)\right\} \pi_0(\mathrm{d}\theta).$$

Example 2 Here we present another example whereby information is non-stochastic about a parameter of interest. Assume that θ is the mean of the probability distribution with density $f(x, \theta)$ and that θ_0 is known to be close to zero due to the information I_1 given by an expert. Hence, it is natural to assess $h(\theta, I_1) = w\,\theta^2$, where w is a positive weight, and the probability distribution minimizing the loss (4) with prior π_0 is given by

$$\pi_1(A) = \int_A e^{-w\,\theta^2}\, \pi_0(\mathrm{d}\theta) \Big/ \int_\Theta e^{-w\,\theta^2}\, \pi_0(\mathrm{d}\theta),$$

for every measurable subset A of Θ. A Bayes estimator would be the mean of π_1, i.e. $\int_\Theta \theta\, e^{-w\,\theta^2}\, \pi_0(\mathrm{d}\theta) / \int_\Theta e^{-w\,\theta^2}\, \pi_0(\mathrm{d}\theta)$. This is an interesting example. We have a framework for which it is possible to update word of mouth information which is common to the Bayesian approach when employing so-called expert opinions. More interestingly, we have a coherent framework whereby it is of no matter exactly when this word of mouth information is received. It could be post data, i.e. after a stochastic sample from a density $f(x; \theta)$ has been observed. If such information is provided to a Bayesian post data then there is currently no framework in which this information can be processed.

For more on Bayesian constructions from expert opinion, see Johnson et al. (2010) and references therein.

5 Discussion

A general framework has been presented for the translation of information into probability measure for a parameter θ of interest, using principles of standard decision theory. For general piece of information I it is only required to establish a loss function connecting θ and I. This encompasses, using a natural loss function, the Bayesian learning approach in the case when θ is a parameter of a density function and the information are samples from the density.

The key for coherence of this translation is the Kullback–Leibler divergence. By coherence, it is meant that the n stage solution, π_n, must serve as the prior for the $n+1$ piece of information.

Appendix

Here we provide the proofs of Propositions 1, 2, Theorem 1 and Propositions 3, 4.

Proof (of Proposition 1) If D_g is the Kullback–Leibler divergence, then

$$L(\nu;\, I^{(n)};\, \pi_0) = D_g(\nu, \pi_n) - \log\left(\int_\Theta e^{-h_n(\theta\,;\, I^{(n)})}\, \pi_0(\mathrm{d}\theta)\right),$$

where π_n is the probability measure such that a version of $\mathrm{d}\pi_n/\mathrm{d}\pi_0$ is (8). Clearly, if (7) holds true, then π_n is the unique probability measure (absolutely continuous with respect to π_0) that minimizes L.

If the integral in (7) diverges, then consider a sequence $(A_m)_{m\geq 1}$ of measurable subsets of Θ such that $A_m \uparrow \Theta$ and the integral

$$k_m := \int_{A_m} e^{-h_n(\theta\,;\, I^{(n)})}\, \pi_0(\mathrm{d}\theta)$$

is finite. One can take, for instance, $A_m = \{\theta \in \Theta : h_n(\theta\,;\, I^{(n)}) > c_m\}$, where $c_m \downarrow -\infty$. Moreover, define the sequence $(\mu_m)_{m\geq 1}$ of probability measures on Θ such that $\mathrm{d}\mu_m(\theta)/\mathrm{d}\pi_0(\theta) = \mathbb{I}_{A_m}\, e^{-h_n(\theta\,;\, I^{(n)})}/k_m$, π_0-a.s. By the monotone convergence theorem, $k_m \to \infty$ as m diverges. This entails that $L(\mu_m) = -\log k_m$ diverges to $-\infty$ and therefore L does not attain a minimum. □

To proceed with the other proofs, let us introduce some further notation. Define

$$\begin{aligned}\mathscr{F} &:= \mathscr{F}(\Theta, \pi_0)\\ &:= \left\{ f \in L^1(\Theta, \pi_0) : \int_\Theta f\, \mathrm{d}\pi_0 = 1,\ h_n(\cdot\,,\, I^{(n)})\, f \,+\, g \circ f \in L^1(\Theta, \pi_0)\right\}.\end{aligned}$$

To minimize $L : \mathscr{P}(\Theta, \pi_0, I^{(n)}) \to \mathbb{R}$ is equivalent to minimizing the loss $\bar{L} : \mathscr{F} \to \mathbb{R}$ defined as follows:

$$\bar{L}(f) := \bar{L}(f,\, I^{(n)},\, \pi_0) = \int_\Theta h_n(\theta,\, I^{(n)})\, f(\theta) + g \circ f(\theta)\, \pi_0(\mathrm{d}\theta).$$

Proof (of Proposition 2) The proof will be done by contradiction. Assume that (11) holds true, i.e. $f_0 \equiv 1$ belongs to $\mathscr{F}$ and satisfies:

$$\min_{f\in\mathscr{F}(\Theta,\pi_0)} \bar{L}(f,\, I^{(n)},\, \pi_0) = \bar{L}(f_0,\, I^{(n)},\, \pi_0) = \int_\Theta h_n(\theta,\, I^{(n)})\, \pi_0(\mathrm{d}\theta). \qquad (22)$$

Moreover, assume that $h_n(\theta;\ I^{(n)})$ is not constant on a set with positive π_0 - probability, i.e. there exists $c \in \mathbb{R}$ such that $\pi_0\{\theta \in \Theta :\ h_n(\theta;\ I^{(n)}) > c\}$ and $\pi_0\{\theta \in \Theta :\ h_n(\theta;\ I^{(n)}) \leq c\}$ are both positive. For every $0 < \varepsilon < 1$, define

$$f_\varepsilon(\theta) = 1 + k(\varepsilon)\,\mathbb{I}_{A^c} - \varepsilon\,\mathbb{I}_A,$$

where $A := \{\theta \in \Theta :\ h_n(\theta;\ I^{(n)}) > c\}$, A^c is the complementary of A and $k(\varepsilon) := \varepsilon\,\pi_0(A)/\pi_0(A^c)$.

Notice that f_ε belongs to $\mathscr{F}$ for every $\varepsilon \in (0,\ 1)$. In fact, its integral is one and

$$\begin{aligned}\int_\Theta \left|h_n(\theta,\ I^{(n)})\, f_\varepsilon(\theta) + g \circ f_\varepsilon(\theta)\right| \pi_0(\mathrm{d}\theta)\\ \leq (1 + k(\varepsilon) + \varepsilon) \int_\Theta \left|h_n(\theta,\ I^{(n)})\right| \pi_0(\mathrm{d}\theta) + |g(1 + k(\varepsilon))| + |g(1 - \varepsilon)|,\end{aligned}$$

which is finite. In fact, $f_0 \equiv 1$ is assumed to belong to $\mathscr{F}$ and therefore $\int_\Theta \left|h_n(\theta,\ I^{(n)})\right| \pi_0(\mathrm{d}\theta)$ is finite. Notice that

$$\begin{aligned}\bar{L}(f_\varepsilon) - \bar{L}(f_0) &= \int_\Theta h_n(\theta;\ I^{(n)})(f_\varepsilon(\theta) - 1) + g(f_\varepsilon(\theta))\,\pi_0(\mathrm{d}\theta)\\ &= k(\varepsilon) \int_{A^c} h_n(\theta,\ I^{(n)})\,\pi_0(\mathrm{d}\theta) - \varepsilon \int_A h_n(\theta,\ I^{(n)})\,\pi_0(\mathrm{d}\theta)\\ &\quad + \pi_0(A^c)\, g(1 + k(\varepsilon)) + \pi_0(A)\, g(1 - \varepsilon). \qquad (23)\end{aligned}$$

Being $\pi_0(A)$ positive, there is $\delta > 0$ such that the set $B := \{\theta \in \Theta :\ h_n(\theta;\ I^{(n)}) > c + \delta\}$ has positive π_0 - probability. Therefore,

$$\begin{aligned}&\int_A h_n(\theta,\ I^{(n)})\,\pi_0(\mathrm{d}\theta)\\ &= \int_B h_n(\theta,\ I^{(n)})\,\pi_0(\mathrm{d}\theta) + \int_{A \setminus B} h_n(\theta,\ I^{(n)})\,\pi_0(\mathrm{d}\theta)\\ &> (c + \delta)\,\pi_0(B) + c\,\pi_0(A \setminus B)\\ &= \delta\,\pi_0(B) + c\,\pi_0(A). \qquad (24)\end{aligned}$$

Combination of (23) with (24) entails that

$$\bar{L}(f_\varepsilon) - \bar{L}(f_0) < -\varepsilon\,\delta\,\pi_0(B) + \pi_0(A^c)\, g(1 + k(\varepsilon)) + \pi_0(A)\, g(1 - \varepsilon). \quad (25)$$

Being g convex and $g(1) = 0$, $g(1 + \eta) = g(1 + \eta) - g(1) \leq g'_-(1 + \eta)\,\eta$ and $g(1 - \eta) = g(1 - \eta) - g(1) \leq -g'_+(1 - \eta)\,\eta$ for every $0 < \eta < 1$ (see for instance, Zălinescu 2002, page 49). Hence, (25) entails:

$$\bar{L}(f_\varepsilon) - \bar{L}(f_0) < -\delta\,\pi_0(B)\,\varepsilon + (g'_-(1 + k(\varepsilon)) - g'_+(1 - \varepsilon))\,\pi_0(A)\,\varepsilon.$$

Being g differentiable at one, there is $\bar{\varepsilon} > 0$ such that $g'_-(1 + k(\varepsilon)) - g'_+(1 - \varepsilon) < \delta\, \pi_0(B)/\pi_0(A)$ for every $0 \leq \varepsilon \leq \bar{\varepsilon}$. Therefore, $\bar{L}(f_\varepsilon) < \bar{L}(f_0)$ for some $\varepsilon > 0$, which contradicts (22) and the proof is complete. □

Proof (of Theorem 1) Define $\varphi_\theta(x) := h_n(\theta,\, I^{(n)})\, x \,+\, g(x)$, for every $x > 0$ and every $\theta \in \Theta$. For every $\theta \in \Theta$, $\varphi_\theta : (0, \infty) \to \mathbb{R}$ is a convex function and therefore, denoting by $\frac{\partial_+}{\partial x}\varphi_\theta(x)$ and by $\frac{\partial_-}{\partial x}\varphi_\theta(x)$ the right and left derivatives of φ_θ respectively,

$$\varphi_\theta(x) - \varphi_\theta(x_0) \geq \frac{\partial_+}{\partial x}\varphi_\theta(x)\Big|_{x=x_0}(x - x_0)$$
$$\varphi_\theta(x) - \varphi_\theta(x_0) \geq \frac{\partial_-}{\partial x}\varphi_\theta(x)\Big|_{x=x_0}(x - x_0)$$

holds true for every (x, x_0) in $(0, \infty) \times (0, \infty)$. Hence, letting $f_0 = \mathrm{d}\pi_n/\mathrm{d}\pi_0$,

$$\varphi_\theta \circ f(\theta) - \varphi_\theta \circ f_0(\theta) \geq \frac{\partial_+}{\partial x}\varphi_\theta(x)\Big|_{x=f_0(\theta)}(f(\theta) - f_0(\theta))\, \mathbb{I}_{\{f > f_0\}}(\theta) + \frac{\partial_-}{\partial x}\varphi_\theta(x)\Big|_{x=f_0(\theta)}(f(\theta) - f_0(\theta))\, \mathbb{I}_{\{f < f_0\}}(\theta), \quad (26)$$

holds true for every $f \in \mathscr{F}$. By hypothesis, there exists a real constant c such that

$$\frac{\partial_+}{\partial x}\varphi_\theta(x)\Big|_{x=f_0(\theta)} = h_n(\theta;\, I^{(n)}) \,+\, g'_+ \circ f_0(\theta) \,\geq\, c$$
$$\frac{\partial_-}{\partial x}\varphi_\theta(x)\Big|_{x=f_0(\theta)} = h_n(\theta;\, I^{(n)}) \,+\, g'_- \circ f_0(\theta) \,\leq\, c,$$

for every $\theta \in \Theta$. Therefore, (26) yields:

$$\varphi_\theta \circ f(\theta) - \varphi_\theta \circ f_0(\theta) \;\geq\; c\,(f(\theta) - f_0(\theta)), \quad (27)$$

for every $\theta \in \Theta$. Integrating both sides of (27) with respect to π_0, one obtains that $\bar{L}(f) \geq \bar{L}(f_0)$ for every $f \in \mathscr{F}$. □

To prove Proposition 3, it is useful to state and prove the following lemma, which gives a necessary condition for the solution of this minimization problem. Notice that the map $\bar{L}$ is convex and therefore, any local minimum of L is a global minimum.

Lemma 1 *If f_0, $f_1 \in \mathscr{F}$ and $\bar{L}(f_0) = \min_{f \in \mathscr{F}} \bar{L}(f)$, then*

$$\int_{\{f_1 > f_0\}} \xi_+\,(f_1 - f_0)\,\mathrm{d}\pi_0 \,+\, \int_{\{f_1 < f_0\}} \xi_-\,(f_1 - f_0)\,\mathrm{d}\pi_0 \geq 0, \quad (28)$$

where $\xi_+(\theta) := h_n(\theta; I^{(n)}) + g'_+ \circ f_0(\theta)$, $\xi_-(\theta) := h_n(\theta; I^{(n)}) + g'_- \circ f_0(\theta)$, for every $\theta \in \Theta$.

Proof Define $\varphi_\theta(x) := h_n(\theta; I^{(n)})\, x + g(x)$, for every $x > 0$ and every $\theta \in \Theta$. Notice that $\mathscr{F}$ is a convex set. In fact, if f_2 and f_3 belongs to $\mathscr{F}$ and $0 < \varepsilon < 1$, then $\varepsilon f_2 + (1-\varepsilon) f_3$ is a density and by convexity of φ_θ and Jensen's inequality,

$$\int_\Theta \left|\varphi_\theta(\varepsilon f_2(\theta) + (1-\varepsilon) f_3(\theta))\right| \pi_0(\mathrm{d}\theta) \leq \varepsilon \int_\Theta \left|\varphi_\theta(f_2(\theta))\right| \pi_0(\mathrm{d}\theta) + (1-\varepsilon) \int_\Theta \left|\varphi_\theta(f_3(\theta))\right| \pi_0(\mathrm{d}\theta) < \infty.$$

Therefore the function $f_\varepsilon := (1-\varepsilon) f_0 + \varepsilon f_1$ belongs to $\mathscr{F}$, for every $0 \leq \varepsilon \leq 1$. Denote $\psi_m(\theta) := \{\varphi_\theta(f_{1/m}(\theta)) - \varphi_\theta(f_0(\theta))\}/\{1/m\}$, for every integer $m \geq 1$. Recall that a convex function η of real variable has non-decreasing incremental ratios and admits one sided derivatives $\eta'_+(x) = \frac{\partial_+}{\partial x}\eta(x)$ and $\eta'_-(x) = \frac{\partial_-}{\partial x}\eta(x)$. Hence, by convexity of φ_θ, $\psi_m(\theta) \downarrow \frac{\partial_+}{\partial \varepsilon}\varphi_\theta(f_\varepsilon(\theta))\Big|_{\varepsilon=0}$ as m diverges, for every fixed $\theta \in \Theta$. Moreover, $\int_\Theta \varphi_\theta \circ f_\varepsilon(\theta)\, \pi_0(\mathrm{d}\theta) < \infty$ (being $f_\varepsilon \in \mathscr{F}$) for every $0 \leq \varepsilon \leq 1$ and therefore ψ_m is in $L^1(\Theta, \pi_0)$ for each $m \geq 1$. Hence, by the monotone convergence theorem, it is possible to differentiate $\bar{L}(f_\varepsilon)$ under the integral sign, obtaining:

$$\begin{aligned}\frac{\partial_+}{\partial \varepsilon}\bar{L}(f_\varepsilon)\Big|_{\varepsilon=0} &= \int_\Theta \frac{\partial_+}{\partial \varepsilon}\varphi_\theta(f_\varepsilon(\theta))\Big|_{\varepsilon=0} \pi_0(\mathrm{d}\theta) \\ &= \int_{\{f_1 > f_0\}} \left\{h_n(\theta; I^{(n)}) + g'_+ \circ f_0(\theta)\right\}(f_1 - f_0)\, \pi_0(\mathrm{d}\theta) \\ &\quad + \int_{\{f_1 < f_0\}} \left\{h_n(\theta; I^{(n)}) + g'_- \circ f_0(\theta)\right\}(f_1 - f_0)\, \pi_0(\mathrm{d}\theta)\end{aligned}$$

which must be nonnegative since the map $\varepsilon \to \bar{L}(f_\varepsilon)$ defined on $(0, 1)$ attains its minimum at $\varepsilon = 0$ and (28) is proved. □

Proof (of Proposition 3) It needs to be proved that a density $f_0 \in \mathscr{F}$ satisfies

$$\bar{L}(f_0) = \min_{f \in \mathscr{F}} \bar{L}(f) \tag{29}$$

if and only if

$$h_n(\theta; I^{(n)}) + g' \circ f_0(\theta) \tag{30}$$

is constant, π_0- a.s.

The "if" part trivially follows from Theorem 1. In fact, if g is differentiable, $g' = g'_+ = g'_-$ and therefore the fact that (30) is constant with $f_0 = \mathrm{d}\pi_n/\mathrm{d}\pi_0$ implies (12).

Now, let us prove the "only if" part. Let

$$D_n := \{\theta \in \Theta :\ 1/n \leq f_0 \leq n,\ -n \leq h_n(\theta, I^{(n)})\, f(\theta) \leq n\},$$

for a fixed positive integer n, which is big enough so that $\pi_0(D_n) > 0$. Moreover, fix a measurable subset A of Θ such that

$$\int_{A^c \cap D_n} f_0 \, \mathrm{d}\pi_0 \geq \int_{A \cap D_n} f_0 \, \mathrm{d}\pi_0, \tag{31}$$

and

$$\int_{A^c \cap D_n} f_0 \, \mathrm{d}\pi_0 > 0. \tag{32}$$

Moreover, set:

$$f_1 := \mathbb{I}_{D_n^c} f_0 + (1+\varepsilon)\, \mathbb{I}_{A \cap D_n} f_0 + c\, \mathbb{I}_{A^c \cap D_n} f_0, \tag{33}$$

where $0 < \varepsilon < 1$ and $c := 1 - \varepsilon \int_{A \cap D_n} f_0 \, \mathrm{d}\pi_0 / \int_{A^c \cap D_n} f_0 \, \mathrm{d}\pi_0$. Notice that f_1 is nonnegative by (31) and its integral with respect to π_0 is one. Moreover, it belongs to $\mathscr{F}$. In fact,

$$\begin{aligned}
&\int_\Theta \left| h_n(\theta\,;\, I^{(n)})\, f_1(\theta) + g(f_1(\theta)) \right| \pi_0(\mathrm{d}\theta) \\
&\leq \int_{D_n^c} \left| h_n(\theta\,;\, I^{(n)})\, f_0(\theta) + g(f_0(\theta)) \right| \pi_0(\mathrm{d}\theta) \\
&\quad + 2 \int_{D_n} \left| h_n(\theta\,;\, I^{(n)})\, f_0(\theta) \right| \pi_0(\mathrm{d}\theta) \\
&\quad + \int_{A \cap D_n} |g((1+\varepsilon)\, f_0(\theta))| \, \pi_0(\mathrm{d}\theta) \\
&\quad + \int_{A^c \cap D_n} |g(c\, f_0(\theta))| \, \pi_0(\mathrm{d}\theta)
\end{aligned} \tag{34}$$

is finite. This follows from the fact that each integrand of each addendum of the right term in (34) is bounded. For the first addendum this is true since $f_0 \in \mathscr{F}$, for the second one it follows from the way D_n is defined, for the third and the fourth one it follows from that fact that g, being convex, is also continuous and therefore bounded on every compact interval.

At this stage, one can consider the function $f_2 := 2f_0 - f_1$, which is nonnegative since $f_1 \leq 2f_0$. One can easily verify that f_2 is a density with respect to π_0 and also that it belongs to $\mathscr{F}$, similarly as we did for f_1. Moreover, $f_1 - f_0 = f_0 - f_2$. Therefore, defining the function ξ such that $\xi(\theta) := h_n(\theta;\, I^{(n)}) + g' \circ f_0(\theta)$, for every $\theta \in \Theta$, one can write that

$$\int_\Theta \xi\, (f_1 - f_0) = \int_\Theta \xi\, (f_0 - f_2). \tag{35}$$

Since g is differentiable, the map ξ coincides with the maps ξ_+ and ξ_- defined in the statement of Lemma 1. By Lemma 1, the two integrals $\int_\Theta \xi\, (f_1 - f_0)$ and $\int_\Theta \xi\, (f_2 - f_0)$

are both nonnegative since f_1 and f_2 belong to $\mathscr{F}$. Hence, by (35), they must equal to zero. In particular, $\int_\Theta \xi\,(f_1 - f_0) = 0$, which by (33) yields:

$$\int_{A\cap D_n} \xi\, f_0\, \mathrm{d}\pi_0 \int_{A^c\cap D_n} f_0\, \mathrm{d}\pi_0 \;=\; \int_{A^c\cap D_n} \xi\, f_0\, \mathrm{d}\pi_0 \int_{A\cap D_n} f_0\, \mathrm{d}\pi_0. \tag{36}$$

Adding $\int_{A\cap D_n} \xi\, f_0\, \mathrm{d}\pi_0 \int_{A\cap D_n} f_0\, \mathrm{d}\pi_0$ to each side of (36), one can see that (36) is equivalent to:

$$\int_{A\cap D_n} \xi\, f_0\, \mathrm{d}\pi_0 \int_{D_n} f_0\, \mathrm{d}\pi_0 \;=\; \int_{A\cap D_n} f_0\, \mathrm{d}\pi_0 \int_{D_n} \xi\, f_0\, \mathrm{d}\pi_0. \tag{37}$$

Notice that (36) is trivially true if (32) is not satisfied. Moreover, it is true when A is replaced by A^c and viceversa. Clearly, for each measurable subset A of Θ, either (31) holds true or the equivalent inequality obtained replacing A with A^c and viceversa holds true. Therefore, (37) must be satisfied by every measurable subset A of Θ. In other words, for every measurable subset A of Θ, the integral $\int_A \mathbb{I}_{D_n} f_0 \left(\xi \int_{D_n} f_0\, \mathrm{d}\pi_0 - \int_{D_n} \xi\, f_0\, \mathrm{d}\pi_0\right) \mathrm{d}\pi_0$ is zero and therefore the integrand is constantly zero, π_0-a.s. This implies that

$$\mathbb{I}_{D_n}\,\xi \;=\; \left(\int_{D_n} \xi\, f_0\, \mathrm{d}\pi_0 \Big/ \int_{D_n} f_0\, \mathrm{d}\pi_0\right) \mathbb{I}_{D_n},$$

π_0-a.s., i.e. the function ξ is constant on D_n, π_0-a.s. Since this is true eventually for every n and $D_n \uparrow \{f_0 > 0\}$ as n diverges, ξ is constant on $\{f_0 > 0\}$, π_0-a.s. Being g convex and differentiable, g' is continuous (in particular at zero) and therefore ξ is constant on Θ, π_0-a.s. □

Proof (of Proposition 4) By Corollary 1, if g is differentiable and strictly convex, then (14) is equivalent to (13). Therefore, the existence of a probability measure π_n satisfying (14) needs to be proved. To this aim, denote

$$\begin{aligned} c_1 &= g'_+(0) \;+\; \operatorname*{ess\,sup}_{\theta\in\Theta}\; h_n(\theta,\, I^{(n)}) \\ c_2 &= g'(\infty) \;+\; \operatorname*{ess\,inf}_{\theta\in\Theta}\; h_n(\theta,\, I^{(n)}) \end{aligned} \tag{38}$$

and define

$$f_c(\theta) \;:=\; G(c - h_n(\theta, I^{(n)})), \tag{39}$$

for every $c \in (c_1,\, c_2)$. Notice that f_c is properly defined. In fact, by (38), for every $c \in (c_1,\, c_2)$ and every $\theta \in \Theta$, $c - h_n(\theta;\, I^{(n)})$ belongs to $(g'_+(0),\, g'(\infty))(\pi_0$-a.s.), which is the domain of G.

It needs to be proved that for some $c^* \in (c_1,\, c_2)$, f_{c^*} is a density on Θ with respect to π_0. Since the image of G is the domain $(0, \infty)$ of g', f_c is nonnegative for every

$c \in (c_1, c_2)$. Moreover, it is measurable. In fact, $h_n(\cdot, I^{(n)})$ is measurable being the integral (4) defined, and G is the inverse of a measurable function and therefore measurable. Therefore, its integral $I_c := \int_\Theta f_c \, d\pi_0$ is properly defined.

At this stage, it needs to be proved that there exists $c^* \in (c_1, c_2)$ such that $I_{c^*} = 1$. To this aim, for every fixed $\bar{c} \in (c_1, c_2)$, take $0 < \varepsilon < c_2 - \bar{c}$. Hence, being G strictly increasing, if $c \in (\bar{c} - \varepsilon, \bar{c} + \varepsilon)$, then $\sup_\Theta f_c < G(c_2 - \operatorname{ess\,inf}_\Theta h(\cdot; I^{(n)})) = \infty$. Therefore, one can apply the dominated convergence theorem to write:

$$\lim_{c \to \bar{c}} I_c = \int_\Theta \lim_{c \to \bar{c}} f_c \, d\pi_0. \tag{40}$$

Moreover, for every fixed $\theta \in \Theta$, $f_c(\theta)$ is a continuous function of $c \in (c_1, c_2)$. In fact, G is continuous, since its inverse g' is a continuous function defined on an interval, i.e. $(0, \infty)$. Hence, by (40), I_c is a continuous function of c on the interval (c_1, c_2).

Notice that being G increasing, (15) yields that

$$\begin{aligned} f_{c_1} &\le G(c_1 - \operatorname*{ess\,inf}_{\theta \in \Theta} h_n(\theta, I^{(n)})) \\ &\le G(g'_+(0) + \operatorname*{ess\,sup}_{\theta \in \Theta} h_n(\theta, I^{(n)}) - \operatorname*{ess\,inf}_{\theta \in \Theta} h_n(\theta, I^{(n)})) \le 1, \end{aligned}$$

and

$$\begin{aligned} f_{c_2} &\ge G(c_2 - \operatorname*{ess\,sup}_{\theta \in \Theta} h_n(\theta, I^{(n)})) \\ &\ge G(g'(\infty) + \operatorname*{ess\,inf}_{\theta \in \Theta} h_n(\theta, I^{(n)}) - \operatorname*{ess\,sup}_{\theta \in \Theta} h_n(\theta, I^{(n)})) \ge 1. \end{aligned}$$

Therefore, integrating with respect to π_0, one can see that $I_{c_1} \le 1 \le I_{c_2}$. Being I_c a continuous function defined on an interval, which is (c_1, c_2), one can apply the intermediate value theorem to prove that there exists $c^* \in (c_1, c_2)$ such that $I_{c^*} = 1$. Hence, one can properly define a probability measure π_n such that $d\pi_n / d\pi_0 = f_{c^*}$, π_0-a.s., where

$$f_{c^*}(\theta) := G(c^* - h_n(\theta, I^{(n)})), \tag{41}$$

for some real constant c^*.

Since g is a strictly convex map, so is $\bar{L}$ and therefore $\bar{L}$ admits at most one minimum point. Therefore, f_0 must be the only one. In other words, (41) implies (29) and the proof is complete. □

6 Proof of Theorem 2

Assume that Θ contains at least two distinct points, say θ_0 and θ_1. Otherwise, π_0 is degenerate and is equal to $U_n(\pi_0; I)$ for every $I \in \mathcal{I}^n$ and every $n \ge 1$, and therefore (19) is trivially satisfied.

To prove Theorem 2, a very specific choice for π_0 will be considered. Let

$$\pi_0 = p_0\delta_{\theta_0} + (1 - p_0)\delta_{\theta_1}, \tag{42}$$

where $0 < p_0 < 1$. Hence,

$$\nu = p_1\delta_{\theta_0} + (1 - p_1)\delta_{\theta_1}, \tag{43}$$

where $0 \leq p_1 \leq 1$ so that $\nu \ll \pi_0$.

Hence, the probability measure ν is identified by a real number p_1 and the functional (4) can be replaced by a function of real variable of the following form:

$$\begin{aligned} l(p) \;=\; l(p, p_0, I) :=\; & h_n(\theta_0,\, I)\, p \;+\; g\left(\frac{p}{p_0}\right) p_0 \\ & +\; h_n(\theta_1,\, I)\,(1 - p) \;+\; g\left(\frac{1 - p}{1 - p_0}\right)(1 - p_0) \end{aligned} \tag{44}$$

where $I \in \mathcal{I}^n$. Under (42), the minimization problem in (16) is tantamount to minimizing the function $l(p)$ with the constraint: $0 \leq p \leq 1$.

In this simple situation, the operator U_n satisfying (16) can be replaced by a map $\bar{U}_n : [0, 1] \times \mathcal{I}^n \to [0, 1]$ such that $l(\bar{U}_n(p_0, I),\ p_0,\ I) = \min_{p\in[0,1]} l(p, p_0, I)$, for every $I \in \mathcal{I}^n$, every $n \geq 1$ and every $0 < p_0 < 1$. Being $\nu \ll \pi_0$,

$$\bar{U}_n(0, I) = 0, \quad \bar{U}_n(1, I) = 1, \tag{45}$$

for every $I \in \mathcal{I}^n$ and every $n \geq 1$. In the rest of the proof, it will be assumed that $0 < p_0 < 1$ so that $\bar{U}(p_0\,,\, I)$ is the minimum point of $l(\cdot\,,\ p_0,\ I)$, for every $I \in \mathcal{I}^n$.

Assumption (17) entails that

$$\bar{U}_n(\bar{U}_m(p_0, I), I') = \bar{U}_{n+m}(p_0, I, I') \tag{46}$$

holds true for every $0 \leq p_0 \leq 1$, $I \in \mathcal{I}^n$, $I' \in \mathcal{I}^m$, $n, m \geq 1$. Hence,

$$\bar{U}_n(\bar{U}_m(p_0, I'), I) = \bar{U}_m(\bar{U}_n(p_0, I), I') \tag{47}$$

for every $0 \leq p_0 \leq 1$, $I \in \mathcal{I}^n$, $I' \in \mathcal{I}^m$, $n, m \geq 1$.

Recalling that composition with an affine function and weighted summation preserve convexity, one notices that by (44), l is convex and

$$l'_+(p) = h_n(\theta_0,\, I) \;+\; g'_+\left(\frac{p}{p_0}\right) \;-\; h_n(\theta_1,\, I) \;-\; g'_-\left(\frac{1 - p}{1 - p_0}\right), \tag{48a}$$

$$l'_-(p) = h_n(\theta_0,\, I) \;+\; g'_-\left(\frac{p}{p_0}\right) \;-\; h_n(\theta_1,\, I) \;-\; g'_+\left(\frac{1 - p}{1 - p_0}\right), \tag{48b}$$

for every $0 < p < 1$. An analogous equality is satisfied by the derivatives $l'(p)$ and $g'(p)$ for every p at which g is differentiable.

Before proceeding, we need to prove the following lemma.

Lemma 2 *If the hypotheses of Theorem 2 are satisfied, then* $g'_+(0)$ *must be* $-\infty$ *and moreover* $0 < \bar{U}_n(p_0, I) < 1$ *for every* $p_0 \in (0, 1)$*, every* $I \in \mathcal{I}^n$ *and every* $n \geq 1$.

In other words, if the updating mechanism U_n is consistent and the prior π_0 satisfies (42), then $U_n(\pi_0, I^{(n)})$ cannot be degenerated and $g'_+(0)$ is not finite, as for $g(x) = x \log x$, which yields the Kullback–Leibler divergence. Recall also it is important that (17) holds for every possible loss function h, i.e. for every h in the class $\mathscr{H}(\pi_0)$. In fact, we need to take into consideration any possible information I which the loss $h(\theta, I)$ could be based on.

Proof The fact that $g'_+(0) = -\infty$ will be proved by contradiction. To this aim, assume that $g'_+(0)$ is finite and denote

$$d = \begin{cases} g'_-(1/(1-p_0)) & \text{if } h_n(\theta_0;\ I) - h_n(\theta_1;\ I) > 0 \\ g'_-(1/p_0) & \text{if } h_n(\theta_0;\ I) - h_n(\theta_1;\ I) < 0. \end{cases}$$

Let us consider the case in which $|h_n(\theta_0,\ I) - h_n(\theta_1,\ I)| > d - g'_+(0)$. Recalling (48a), notice that if $h_n(\theta_0,\ I) - h_n(\theta_1,\ I) > d - g'_+(0)$, then $l'_+(0) > 0$. Since l'_+ is a non decreasing function by convexity of l, this entails that l'_+ is positive on $(0, 1)$. In this case, zero is the (unique) minimum point for l. By (48b), if $h_n(\theta_1,\ I) - h_n(\theta_0,\ I) > d - g'_+(0)$, then $l'_-(1) < 0$. In this other case, l turns out to be decreasing on $(0, 1)$ and therefore one is its (unique) minimum point.

At this stage, consider a function h in $\mathscr{H}(\pi_0)$ and two pieces of information I and I' in $\mathcal{I}^n$ (for some $n \geq 1$) such that

$$h_n(\theta_0,\ I) - h_n(\theta_1,\ I) > d - g'_+(0) \tag{49}$$

and that

$$h_n(\theta_1,\ I') - h_n(\theta_0,\ I') > d - g'_+(0), \tag{50}$$

where h_n satisfies (5) and $g'_+(0)$ is finite by hypothesis.

By (49) $\bar{U}_n(p_0, I) = 0$ and, recalling (45), (47) becomes

$$\bar{U}_n(\bar{U}_n(p_0, I'), I) = 0. \tag{51}$$

By (50), $\bar{U}_n(p_0, I') = 1$ and therefore the left hand side of (51) is one, which yields a contradiction. Hence, $g'_+(0) = -\infty$. By (48), this entails that $l'_+(0) = -\infty$ and $l'_-(1) = \infty$. Therefore, l cannot attain its minimum at zero or one, i.e. $p_1 := \bar{U}_n(p_0, I)$ must belong to $(0, 1)$. □

The following lemma provides necessary and sufficient conditions for a minimum point of $l(\cdot\,, p_0, I)$ with $0 < p_0 < 1$ under the hypotheses of Theorem 2.

Lemma 3 *Assume that the hypotheses of Theorem 2 are satisfied and let* $0 < p_0 < 1$. *Hence,* $p_1 = U_n(p_0, I)$ *if and only if*

$$g'_-\left(\frac{p_1}{p_0}\right) - g'_+\left(\frac{1-p_1}{1-p_0}\right) \leq h_n(\theta_1, I) - h_n(\theta_0, I) \tag{52a}$$

$$h_n(\theta_1, I) - h_n(\theta_0, I) \leq g'_+\left(\tfrac{p_1}{p_0}\right) - g'_-\left(\tfrac{1-p_1}{1-p_0}\right). \tag{52b}$$

Proof If $0 < p_0 < 1$ and p_1 is the minimum point of $l(\cdot, p_0, I)$, then by Lemma 2, p_1 must belong to $(0, 1)$. Hence, $l'_-(p_1)$ and $l'_+(p_1)$ exist and

$$l'_-(p_1) \leq 0 \leq l'_+(p_1). \tag{53}$$

Since $l(\cdot\,, p_0, I)$ is a convex function and its minimum point must be an interior point of $(0, 1)$, (53) is a sufficient and necessary condition for p_1 to be the minimum point of l. By (48), the inequality (53) is tantamount to (52). □

At this stage, denote by $\mathscr{H}$ the class of all maps $h : \Theta \times \mathcal{I} \to \mathbb{R}$ such that $h(\cdot\,, I_1)$ is measurable for every $I_1 \in \mathcal{I}$.

Lemma 4 *If* $0 < p_0, p_1, p_2 < 1$, n *is a positive integer and* $I, I' \in \mathcal{I}^n$, $I \neq I'$ *then there exists* $h \in \mathscr{H}$, *such that* p_1 *is the minimum point of* $l(\cdot; p_0; I)$, *and* p_2 *is the minimum point of* $l(\cdot; p_1; I')$.

Proof At this stage, let $\psi : (0, 1) \to \mathbb{R}$ be the following function:

$$\psi(p, p_0) = p_0\, g\left(\frac{p}{p_0}\right) + (1 - p_0)\, g\left(\frac{1-p}{1-p_0}\right).$$

Hence, (52) can be re-written in the following form:

$$\frac{\partial_-}{\partial p_1}\psi(p_1, p_0) \leq h_n(\theta_1; I) - h_n(\theta_0; I) \leq \frac{\partial_+}{\partial p_1}\psi(p_1, p_0). \tag{54}$$

By Lemma 3, p_2 is the minimum point of $l(\cdot, p_1, I')$ if and only if

$$\frac{\partial_-}{\partial p_2}\psi(p_2, p_1) \leq h_n(\theta_1; I') - h_n(\theta_0; I') \leq \frac{\partial_+}{\partial p_2}\psi(p_2, p_1). \tag{55}$$

Take a function h satisfying (54), and (55) under (5), and the thesis follows. □

Lemma 5 *If the assumptions of Theorem 2 are satisfied, then* g *is a differentiable function and its derivative* g' *is continuous. Moreover,*

$$g'(xy) = g'(x) + g'(y) - g'(1) \tag{56}$$

holds true for every $x, y > 0$.

Proof Fix $p_0 \in (0,1), n > 0, I, I' \in \mathcal{I}^n$ and let $p_1 = \bar{U}_n(p_0, I)$ and $p_2 = \bar{U}_n(p_1, I')$. Since (52) is satisfied by every $0 < p_0 < 1$ with $p_1 = \bar{U}_n(p_0; I)$ and $I \in \mathcal{I}^n$, one can write mutatis mutandis:

$$g'_-\left(\frac{p_2}{p_1}\right) - g'_+\left(\frac{1-p_2}{1-p_1}\right) \leq h_n(\theta_1;\, I') - h_n(\theta_0;\, I') \tag{57a}$$

$$h_n(\theta_1;\, I') - h_n(\theta_0;\, I') \leq g'_+\left(\tfrac{p_2}{p_1}\right) - g'_-\left(\tfrac{1-p_2}{1-p_1}\right). \tag{57b}$$

By (47), $p_2 = \bar{U}_{2n}(p_0, I, I')$. Moreover, by (5) $h_{2n}(\cdot\,, I, I') = h_n(\cdot\,, I) + h_n(\cdot\,, I')$, and therefore by Lemma 3,

$$\begin{aligned} &g'_-\left(\frac{p_2}{p_0}\right) - g'_+\left(\frac{1-p_2}{1-p_0}\right) \\ &\quad \leq h_n(\theta_1,\, I) + h_n(\theta_1,\, I') - h_n(\theta_0,\, I) - h_n(\theta_0,\, I'), \end{aligned} \tag{58a}$$

and

$$\begin{aligned} &h_n(\theta_1,\, I) + h_n(\theta_1,\, I') - h_n(\theta_0,\, I) - h_n(\theta_0,\, I') \\ &\quad \leq g'_+\left(\frac{p_2}{p_0}\right) - g'_-\left(\frac{1-p_2}{1-p_0}\right). \end{aligned} \tag{58b}$$

Summing up term by term (52b) and (57b) and considering (58a), one obtains:

$$\begin{aligned} &g'_-\left(\frac{p_2}{p_0}\right) - g'_+\left(\frac{1-p_2}{1-p_0}\right) \\ &\leq g'_+\left(\frac{p_1}{p_0}\right) - g'_-\left(\frac{1-p_1}{1-p_0}\right) + g'_+\left(\frac{p_2}{p_1}\right) - g'_-\left(\frac{1-p_2}{1-p_1}\right). \end{aligned} \tag{59a}$$

In an analogous way, (52a),(57a) and (58b) yield:

$$\begin{aligned} &g'_+\left(\frac{p_2}{p_0}\right) - g'_-\left(\frac{1-p_2}{1-p_0}\right) \\ &\geq g'_-\left(\frac{p_1}{p_0}\right) - g'_+\left(\frac{1-p_1}{1-p_0}\right) + g'_-\left(\frac{p_2}{p_1}\right) - g'_+\left(\frac{1-p_2}{1-p_1}\right). \end{aligned} \tag{59b}$$

In virtue of Lemma 4, (59) needs to hold for every $p_0, p_1, p_2 \in (0,1)$. By substituting $t = p_0, x = p_1/p_0, y = p_2/p_1$, one obtains that

$$\begin{aligned} &g'_-(xy) - g'_+\left(\frac{1-txy}{1-t}\right) \\ &\quad \leq g'_+(x) - g'_-\left(\frac{1-tx}{1-t}\right) + g'_+(y) - g'_-\left(\frac{1-txy}{1-tx}\right), \end{aligned} \tag{60a}$$

and

$$g'_+(xy) - g'_-\left(\frac{1-txy}{1-t}\right) \geq g'_-(x) - g'_+\left(\frac{1-tx}{1-t}\right) + g'_-(y) - g'_+\left(\frac{1-txy}{1-tx}\right). \tag{60b}$$

must hold for every $0 < t < 1$ and every $x, y > 0$ such that $x < 1/t$ and $y < 1/(tx)$. At this stage, recall that being g convex,

$$\lim_{u \downarrow u_0} g'_+(u) = \lim_{u \downarrow u_0} g'_-(u) = g'_+(u_0) \tag{61a}$$

$$\lim_{u \uparrow u_0} g'_+(u) = \lim_{u \uparrow u_0} g'_-(u) = g'_-(u_0), \tag{61b}$$

for every $u_0 > 0$. See, for instance, Zălinescu (2002).

Hence, taking $x = 1 + \varepsilon$, $y = 1 - \varepsilon$ and $0 < t < 1$ so that as $\varepsilon \downarrow 0$, $x \downarrow 1$, $y \uparrow 1$, $xy \uparrow 1$, (60b) entails that $g'_+(1) \leq g'_-(1)$. By convexity of g, $g'_- \leq g'_+$ and therefore $g'_+(1) = g'_-(1)$. Hence, g is differentiable at one and therefore letting $t \downarrow 0$, (60a) entails that for every $x, y > 0$

$$g'_-(xy) \leq g'_+(x) + g'_+(y) - g'(1). \tag{62}$$

Hence, taking $x = x_0 - \varepsilon$, $y = (x_0 + \varepsilon)/(x_0 - \varepsilon)$, for some fixed $x_0 > 0$ and letting $\varepsilon \downarrow 0$ (so that $x \uparrow x_0$, $y \downarrow 1$, $xy \downarrow x_0$), (62) yields that $g'_-(x_0) - g'_+(x_0) \geq 0$. By convexity of g, $g'_- \leq g'_+$ and therefore $g'_+(x_0) = g'_-(x_0)$ for every $x > 0$. Hence, the derivative g' of g exists and $g' = g'_+ = g'_-$. In virtue of (61), the derivative g' is continuous.

Being $g'_- = g'_+$, letting $t \downarrow 0$, (60) entails that (56) holds true for every $x, y > 0$. □

At this stage, it is possible to prove Theorem 2.

Proof (of Theorem 2) It can be proved by induction from (56) that

$$g'\left(y^n\right) = n\left(g'(y) - g'(1)\right) + g'(1), \tag{63}$$

holds true for every $y > 0$ and every integer $n > 0$. In fact, notice that the case $n = 1$ is trivial and that applying (56) taking $x = y^n$ and then (63), one can write:

$$\begin{aligned} g'(y^{n+1}) &= g'(y^n) + g'(y) - g'(1) \\ &= (n+1)\left(g'(y) - g'(1)\right) + g'(1). \end{aligned} \tag{64}$$

Notice that (63) implies that

$$g'(y^{1/m}) = \tfrac{1}{m}\left(g'(y) - g'(1)\right) + g'(1), \tag{65}$$

for every integer $m > 0$ and every $y > 0$. Applying (65) and (63), one obtains that

$$\begin{aligned} g'(y^{n/m}) &= \tfrac{1}{m}\,(g'(y^n) - g'(1)) + g'(1) \\ &= \tfrac{n}{m}\,(g'(y) - g'(1)) + g'(1) \end{aligned} \tag{66}$$

holds true for every pair of positive integers (n, m) and every $y > 0$. At this stage, fix $y = 2$. By continuity of g', this entails that

$$g'(2^z) = z\,(g'(2) - g'(1)) + g'(1)$$

holds true for every $z > 0$. By substituting $x = 2^z$, this implies that

$$g'(x) = k\,\ln(x) + g'(1), \tag{67}$$

where $k = (g'(2) - g'(1))/\ln(2)$. Being g convex, g' is not decreasing and therefore $k \geq 0$. If $k = 0$, then g' is constant, which is impossible since $g'_+(0) = -\infty$ by Lemma 2. Therefore, k must be positive. Being $g(1) = 0$ by assumption, (67) implies that

$$g(x) = k\,x\ln(x) + (g'(1) - k)(x - 1). \tag{68}$$

Hence, $D_g(Q_1, Q_2) = k \int \ln(\mathrm{d}Q_1/\mathrm{d}Q_2)\,\mathrm{d}Q_1$ holds true for some $k > 0$ and for every couple of measures (Q_1, Q_2) such that $Q_1 \ll Q_2$. Therefore, D_g turns out to be the Kullback–Leibler divergence, which yields (18) and (19) by Proposition 1. □

Acknowledgments This work was partially supported by ESF and Regione Lombardia (by the grant "Dote Ricercatori"). We thank two referees and an AE for valuable comments on earlier versions of the paper.

References

Aczel, J., Pfanzagl, J. (1966). Remarks on the measurement of subjective probability and information. *Metrika, 11*, 91–105.

Ali, S. M., Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B, 28*(1), 131–142.

Amari, S.-I. (2009). α-divergence is unique, belonging to both f-divergence and Bregman divergence classes. *IEEE Transactions on Information Theory, 55*(11), 4925–4931.

Berger, J. O. (1980). *Statistical decision theory and Bayesian analysis*. New York: Springer.

Bernardo, J. M. (1979). Expected information as expected utility. *The Annals of Statistics, 7*(3), 686–690.

Bernardo, J., Smith, A. F. M. (1994). *Bayesian theory*. Chichester: Wiley.

Bissiri, P. G., Walker, S. G. (2010). On Bayesian learning from Bernoulli observations. *Journal of Statistical Planning and Inference, 140*(11), 3520–3530.

Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica, 2*, 229–318.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society: Series B, 14*, 107–114.

Huber, P. J., Ronchetti, E. M. (2009). *Robust statistics*. New Jersey: Wiley.

Johnson, S., Tomlinson, G., Hawker, G., Granton, J., Feldman, B. (2010). Methods to elicit beliefs for bayesian priors. a systematic review. *Journal of Clinical Epidemiology, 63*(4), 355–369.

Walker, S. G. (2006). Bayesian inference via a minimization rule. *Sankhyā, 68*, 542–553.

Zălinescu, C. (2002). *Convex analysis in general vector spaces*. Singapore: World Scientific.