Nonparametric likelihood maximization for missing values models

Jean-Marie Aubry* Yuichi Hirose†

September 17, 2010

Abstract

We study the maximization problem for a generalized log-likelihood functional defined on a space of probability measures, issued from a model where a random variable is observed both directly and indirectly through auxiliary variables with know conditional densities. We show existence of the maximum, give a condition for its uniqueness and absolute continuity with respect to the direct observation, and derive its Fréchet (or Hadamard) differential with respect to other parameters of the model and with respect to the observation itself.

1 Introduction

1.1 Nonparametric likelihood

Our framework in this paper is the following. Let S be an integer ≥ 1 and let $\mathscr{X}, \mathscr{Y}_s$ $(1 \leq s \leq S)$ be Polish (i.e. metrisable, complete, separable) spaces. Let $\mathscr{M}_{\mathscr{X}}, \mathscr{M}_{\mathscr{Y}_s}$ denote the sets of all Borel completed probability measures on $\mathscr{X}, \mathscr{Y}_s$ respectively. An $\mathscr{X}$-valued random variable X has (i.i.d.) realisations observed n_0 times; on the other hand, each Y_s has conditional density $f_s(y|x)$ with respect to a reference probability measure ζ_s and is observed n_s times. The measure ζ_s does not play a particular role here, except for the fact that for all $x \in \mathscr{X}$, $\int_{\mathscr{Y}_s} f_s(y|x) d\zeta_s(y) = 1$. Let $n := \sum_{s=0}^{S} n_s$ denote the total number of observations and let $w_s := n_s/n$ be their relative frequencies; the primary goal is to estimate the law of X.

As was pointed out by Gill [3], in a nonparametric setting there is typically no dominating measure, so one cannot simply "maximize a density". This difficulty can be avoided by looking for the maximizer directly in the larger space

*Laboratoire d'Analyse et de Mathématiques Appliquées, UMR CNRS 8050, Université de Paris Est à Créteil, France.

†School of Mathematics, Statistics and Operations Research, Victoria University of Wellington, New Zealand.

of probability measures. More specifically, we are interested in estimating the true law $\nu_0 \in \mathscr{M}_{\mathscr{X}}$ of X by maximising the likelihood of $\nu \in \mathscr{M}_{\mathscr{X}}$, defined as

$$L(\nu) := \prod_{i=1}^{n_0} \nu(\{X_i\}) \prod_{s=1}^{S} \prod_{i=1}^{n_s} f_{Y_s}(Y_{s,i}; \nu) \tag{1}$$

where

$$f_{Y_s}(y; \nu) := \int_{\mathscr{X}} f_s(y|x) d\nu(x) \tag{2}$$

is a probability density with respect to ζ_s. Note that for (1) to make sense, the density f_{Y_s} should a priori be a function defined *pointwise* and not just ζ_s-almost everywhere.

Taking the logarithm of (1) and rescaling, we obtain the log-likelihood

$$\mathfrak{L}(\nu) := w_0 \int_{\mathscr{X}} \log\left(\frac{d\tilde{\nu}}{d\mu_X}(x)\right) d\mu_X(x) + \sum_{s=1}^{S} w_s \int_{\mathscr{Y}_s} \log(f_{Y_s}(y; \nu)) d\mu_{Y_s}(y) \tag{3}$$

where the following notation has been used: $w_s := \frac{n_s}{n}$,

$$\mu_X := \frac{1}{n_0} \sum_{i=1}^{n_0} \delta_{X_i} \quad \text{and} \quad \mu_{Y_s} := \frac{1}{n_s} \sum_{i=1}^{n_s} \delta_{Y_{s,i}} \tag{4}$$

are the empirical measures observed from X and Y_s (respectively called direct and indirect observation measures). In (3) we wrote $\nu =: \tilde{\nu} + \hat{\nu}$ for the Lebesgue decomposition of ν in its absolutely continuous and orthogonal parts with respect to μ_X and $\frac{d\tilde{\nu}}{d\mu_X}$ denotes the Radon-Nicodym derivative.

1.2 Semiparametric setting

Log-likelihoods of type (3) are common in the litterature, notably in a semiparametric setting for case-control or outcome-dependent sampling (Prencice and Pyke [7], Breslow and Holubkov [1], Scott and Wild [8, 9], Lawless *et al.* [5], Zhou *et al.* [14], Weaver and Zhou [12] to name a few). In these situation, the densities f_s also depend on a parameter of interest $\theta \in \Theta$ (thus, so does f_{Y_s}). The full likelihood can then be written in the form

$$L(\nu, \theta) := \prod_{i=1}^{n_0} \nu(\{X_i\}) f_s(Y_i|X_i; \theta) \prod_{s=1}^{S} \prod_{i=1}^{n_s} f_{Y_s}(Y_{s,i}; \nu, \theta)$$

In the so-called *profile likelihood* method, one fixes first θ to find ν (depending on θ) maximizing $L(\nu, \theta)$; then $L(\nu_\theta, \theta)$ is maximized in θ. Since the terms $f_s(Y_i|X_i; \theta)$ do not depend on ν, the first step in this method amounts to maximizing (1) or, equivalently, (3).

The crucial point we would like to make is that maximising (1) or (3) over $\nu \in \mathscr{M}_{\mathscr{X}}$ is in general not the same as maximising

$$\mathscr{L}(\nu_1, \ldots, \nu_{n_0}) := \frac{w_0}{n_0} \sum_{i=1}^{n_0} \log(\nu_i) + \sum_{s=1}^{S} \frac{w_s}{n_s} \sum_{i=1}^{n_s} \log \left(\frac{1}{n_0} \sum_{j=1}^{n_0} \nu_j f(Y_{s,i}|X_j) \right) \quad (5)$$

over the set $\left\{(\nu_1, \ldots, \nu_{n_0}) \in (\mathbb{R}^+)^{n_0} : \sum_{i=0}^{n_0} \nu_i = 1\right\}$. In fact, maximising (5) amounts to maximising (3) over the subset of measures ν that are absolutely continuous with respect to μ_X (notation $\nu \ll \mu_X$). But consider the following very simple example: $\mathscr{X} = \mathscr{Y} = \{0,1\}$, $f(y|x) = \mathbf{1}_{x=y}$, $\mu_X = \delta_0$, $\mu_Y = \delta_1$ and $\frac{n_0}{n} = \frac{n_1}{n} = \frac{1}{2}$: it is not difficult to see that here $\mathfrak{L}(\nu)$ is maximised for $\nu = \frac{1}{2}(\delta_0 + \delta_1) \not\ll \mu_X$.

Nevertheless, (5) seems to be universally accepted in the literature as the quantity to maximise. To quote Lawless *et al.* [5]: "As is standard with nonparametric maximum likelihood, we maximise $[\mathfrak{L}(\nu)]$ over all discrete distributions whose support contains the X-values observed in the data". It is certainly a much easier problem to solve than having to find a maximum over a whole set of measures; but (3) makes more sense from a mathematical point of view, and with less constraint it uses the available data more efficiently.

1.3 Outline of the results

Our first result in this paper is to give a sufficient and (almost) necessary condition guaranteeing that, for any μ_X, μ_{Y_s}, the maximum of (3) exists, is unique, and is reached for $\nu \ll \mu_X$ which therefore has a density $g := \frac{d\nu}{d\mu_X}$. Note that in this paper we study only the maximization problem; the random generation of μ_X and μ_{Y_s} (including the selection of the realizations of X that are directly observed), thus the *estimating* properties of ν, are not considered.

In fact, in § 2 we shall prove this result in a slightly more general framework for (3). First, the numbers w_0, w_s can be any positive real numbers summing to 1 and second, the measures μ_X, μ_{Y_s} can be any Borel completed probability measures on $\mathscr{X}$ and $\mathscr{Y}_s$ respectively. Besides the technical necessity (to define differentiability, see below), this extension has intrinsic interest: general μ_X, μ_{Y_s} can be interpreted as "information measures" containing all that is known about the random variables X and Y_s, with respective weights w_0 and w_s. Empirical measures (4) are the basic example; to model noisy measurements μ_X and μ_{Y_s} could be empirical measures convolved with Gaussian distributions; more generally, any prior information about X or Y_s can be cast in this form.

This extension of the range of μ_X, μ_{Y_s} is costless and allows us to study in § 3 the regularity of the optimal density g as a function of these measures. We will show that g must satisfy a fixed-point equation that may be used for numerical computations, and that it depends smoothly (C^1) on μ_X, μ_{Y_s}. If, as in

a semiparametric setting, the functions f_s also depend on a parameter $\theta \in \Theta$, then the optimal g also depends smoothly on θ.

2 Absolute continuity of the likelihood maximiser

We now consider $\mathfrak{L}(\nu)$ defined by (3) with $\sum_{s=0}^{S} w_s = 1$ and μ_X, μ_{Y_s} arbitrary probability measures on $\mathscr{X}, \mathscr{Y}_s$ $(1 \leq s \leq S)$.

2.1 Existence of the maximiser

We first show that the problem

$$\max_{\nu \in \mathscr{M}_{\mathscr{X}}} \mathfrak{L}(\nu) \tag{$\mathscr{P}$}$$

always has a solution if f_s satisfies basic measurability and continuity hypotheses. Remark that the first term of (3) is the same as $-w_0 D(\mu_X \| \nu)$, where

$$D(\mu\|\nu) := \begin{cases} \int_{\mathscr{X}} \log\left(\frac{d\mu}{d\nu}(x)\right) d\mu(x) & \text{if } \mu \ll \nu \\ +\infty & \text{else} \end{cases}$$

is the Kullback-Leibler relative entropy (or information divergence). If $C_B(\mathscr{X})$ denotes the space of continuous bounded functions on $\mathscr{X}$, with uniform convergence topology, then $\mathscr{M}_{\mathscr{X}}$ is naturally a convex subset of the dual $C_B(\mathscr{X})'$ endowed with the weak* topology (a.k.a. convergence in law). By a theorem of Posner [6], the map $(\mu, \nu) \mapsto D(\mu\|\nu)$ is jointly lower-semicontinuous in μ and ν on $\mathscr{M}_{\mathscr{X}}^2$ for this topology, so the first term of (3) is upper-semicontinuous.

For the subsequent terms of (3) we will need the following result. Let μ_Y be a Borel completed probability measure on a Polish space $\mathscr{Y}$ and for $f \in L^1(\mu_Y)$, let $f \diamond \mu_Y$ denote the measure having density f with respect to μ_Y (i.e. $d(f \diamond \mu_Y)(y) := f(y) d\mu_Y(y)$).

Proposition 2.1. *Suppose that there exists $C < \infty$ such that for every $x_0 \in \mathscr{X}$,*

(i). the function $y \mapsto f(y|x_0)$ is μ_Y-measurable.

(ii). for μ_Y-a.e. y, the function: $x \mapsto f(y|x)$ is continuous at x_0;

(iii). there exists a neighbourhood $N(x_0)$ and $h \in L^1(\mu_Y)$, $\|h\|_{L^1} \leq C$, such that for every $x \in N(x_0)$, for μ_Y-a.e. y, $\left|f(y|x)\right| \leq h(y)$.

Then the linear operator: $\nu \mapsto f_Y(\cdot, \nu) \diamond \mu_Y$ is weak continuous: $\mathscr{M}_{\mathscr{X}} \to C_B(\mathscr{Y})'$.*

Proof. Fix a $\psi \in C_B(\mathscr{Y})$. Thanks to Lebesgue's dominated convergence theorem, the map $x \mapsto \int_{\mathscr{Y}} f(y|x)\psi(y)d\mu_Y(y)$ is continuous and bounded (by $C\|\psi\|_{L^\infty}$) on $\mathscr{X}$, thus integrable against any $\nu \in \mathscr{M}_{\mathscr{X}}$. Using Fubini's theorem, we obtain that the function $y \mapsto \int_{\mathscr{X}} f(y|x)\psi(y)d\nu(x)$ is μ_Y-integrable; in particular, taking $\psi = 1$, $f_Y(\cdot;\nu) \in L^1(\mu_Y)$ and $f_Y(\cdot;\nu) \diamond \mu_Y \in C_B(\mathscr{Y})'$. Suppose now that $\nu_n \xrightarrow{*} \nu$ in $\mathscr{M}_{\mathscr{X}}$. Using Fubini's theorem again,

$$\begin{aligned}
\int_{\mathscr{Y}} f_Y(y;\nu_n)\psi(y)d\mu_Y(y) &= \int_{\mathscr{Y}}\int_{\mathscr{X}} f(y|x)d\nu_n(x)\psi(y)d\mu_Y(y) \\
&= \int_{\mathscr{X}}\int_{\mathscr{Y}} f(y|x)\psi(y)d\mu_Y(y)d\nu_n(x) \\
&\to \int_{\mathscr{X}}\int_{\mathscr{Y}} f(y|x)\psi(y)d\mu_Y(y)d\nu(x) \\
&\to \int_{\mathscr{Y}} f_Y(y;\nu)\psi(y)d\mu_Y(y)
\end{aligned}$$

As $\mathscr{X}$ is separable, the weak* topology on $\mathscr{M}_{\mathscr{X}}$ is metrisable (for example by the Lévy-Prohorov metric) and sequential continuity implies continuity. □

However, we may want (3) to be defined for *any* probability measures μ_X, μ_{Y_s}. To guarantee this, we need $y \mapsto f_{Y_s}(y,\nu)$ to be *universally measurable*, that is, measurable for any Borel completed measure μ_Y; the following proposition is a straightforward consequence of Fubini's theorem.

Proposition 2.2. *Suppose that $(x,y) \mapsto f(y|x)$ is universally measurable and bounded. Then for any $\nu \in \mathscr{M}_{\mathscr{X}}$, the function $f_Y(\cdot,\nu)$ is universally measurable and bounded.*

Putting it all together, we obtain:

Theorem 1. *Suppose that for every $1 \le s \le S$,*

$$(x,y) \mapsto f_s(y|x) \text{ is universally measurable, continuous in } x \text{ and bounded.} \quad (6)$$

Then for any probability measures μ_X, μ_{Y_s}, the map $\nu \mapsto \mathfrak{L}(\nu)$ is well defined, (weak)-upper-semicontinuous and reaches its maximum on $\mathscr{M}_{\mathscr{X}}$.*

Proof. By Proposition 2.2, $\mathfrak{L}(\nu)$ is well defined for any $\nu \in \mathscr{M}_{\mathscr{X}}$ and we can apply Proposition 2.1 to any probability measure μ_{Y_s} on $\mathscr{Y}_s$. Since

$$\int_{\mathscr{Y}_s} \log(f_{Y_s}(y;\nu))d\mu_{Y_s}(y) = -D(\mu_{Y_s}\|f_{Y_s}(\cdot;\nu)\diamond\mu_{Y_s})$$

by Posner's theorem each map: $\nu \mapsto \int_{\mathscr{Y}_s} \log(f_{Y_s}(y;\nu))d\mu_{Y_s}(y)$ is weak* upper-semicontinuous. It follows that $\mathfrak{L}$ is also upper-semicontinuous and reaches its

maximum on $\bigcup_{0\leq\lambda\leq 1}\lambda\mathscr{M}_{\mathscr{X}}$, because this set is weak*-closed in the unit ball of $C_B(\mathscr{X})'$ which is, by the Banach-Alaoglu theorem, compact in the weak* topology. Obviously $\mathfrak{L}(\lambda\nu)\leq\mathfrak{L}(\nu)$ so the maximum is in fact reached on $\mathscr{M}_{\mathscr{X}}$. □

Naturally, the universal measurability condition may be replaced by continuity, stronger but easier to verify.

At "true value" It is interesting to check what happens when the information measures μ_X and μ_{Y_s} are the true probability distributions from the model. In that case $\mu_X=\nu_0$ and μ_{Y_s} has marginal density $f_{Y_s}(\cdot;\nu_0)$ with respect to ζ_s. Then the log-likelihood

$$\mathfrak{L}(\nu)=w_0\int_{\mathscr{X}}\log\left(\frac{d\tilde{\nu}}{d\nu_0}(x)\right)d\nu_0(x)+\sum_{s=1}^{S}w_s\int_{\mathscr{Y}_s}\log(f_{Y_s}(y;\nu))f_{Y_s}(y;\nu_0)d\zeta_s(y)$$

is indeed, by Gibbs' inequality, uniquely maximized for $\nu=\nu_0$.

From now on we assume that (6) is satisfied.

2.2 Condition for absolute continuity

As we noticed before, the solution to $(\mathscr{P})$ in the case of missing data is not necessarily absolutely continuous with respect to μ_X. However, in nonparametric or semiparametric statistics it is often assumed that the log-likelihood is maximised for a density: we now give a condition for this to hold for any μ_X and $\mu_{Y_s}, 1\leq s\leq S$. As before, we are given conditional densities $(x,y)\mapsto f_s(y|x)$ and define $f_{Y_s}(y;\nu):=\int_{\mathscr{X}}f_s(y|x)d\nu(x)$.

Theorem 2. *For each $s\in\{1,\dots,S\}$ and $y\in\mathscr{Y}_s$ let*

$$\alpha_s(y):=\inf_{x\in\mathscr{X}}f_s(y|x),\quad \beta_s(y):=\sup_{x\in\mathscr{X}}f_s(y|x)\quad \textit{and}\quad \gamma_s:=\sup_{y\in\mathscr{Y}_s}\frac{\beta_s(y)}{\alpha_s(y)}\tag{7}$$

Under the hypothesis

$$\sum_{s=1}^{S}w_s\gamma_s<1\tag{8}$$

then for any probability measures μ_X,μ_{Y_s}, the log-likelihood $\mathfrak{L}(\nu)$ reaches its maximum at $\nu\ll\mu_X$. Conversely, if $\sum_{s=1}^{S}w_s\gamma_s>1$, there exist measures μ_X,μ_{Y_s} such that $\mathfrak{L}(\nu)$ reaches its maximum at $\nu\not\ll\mu_X$.

Proof. Suppose that ν realises the maximum of $\mathfrak{L}(\nu)$ and that $\hat{\nu}(\mathscr{X})>0$. For a (small) $h>0$ consider the probability measure

$$\nu_h:=\left(1+h\frac{\hat{\nu}(\mathscr{X})}{\tilde{\nu}(\mathscr{X})}\right)\tilde{\nu}+(1-h)\hat{\nu}$$

We have

$$\begin{aligned}
f_{Y_s}(y;\nu_h) &= \int_{\mathscr{X}} f_s(y|x)d\nu_h(x) \\
&= \left(1+h\frac{\hat{\nu}(\mathscr{X})}{\tilde{\nu}(\mathscr{X})}\right)\int_{\mathscr{X}} f_s(y|x)d\tilde{\nu}(x) + (1-h)\int_{\mathscr{X}} f_s(y|x)d\hat{\nu}(x) \\
&= \int_{\mathscr{X}} f_s(y|x)d\nu(x) \\
&\qquad\qquad + h\left(\frac{\hat{\nu}(\mathscr{X})}{\tilde{\nu}(\mathscr{X})}\int_{\mathscr{X}} f_s(y|x)d\tilde{\nu}(x) - \int_{\mathscr{X}} f_s(y|x)d\hat{\nu}(x)\right) \\
&\geq f_{Y_s}(y;\nu) + h\big(\alpha_s(y)-\beta_s(y)\big)\hat{\nu}(\mathscr{X})
\end{aligned}$$

and since $f_{Y_s}(y;\nu) \geq \alpha_s(y)$,

$$f_{Y_s}(y;\nu_h) \geq f_{Y_s}(y;\nu)\left(1+h\left(1-\frac{\beta_s(y)}{\alpha_s(y)}\right)\hat{\nu}(\mathscr{X})\right)$$

so that

$$\mathfrak{L}(\nu_h) \geq \mathfrak{L}(\nu) + w_0\log\left(1+h\frac{\hat{\nu}(\mathscr{X})}{\tilde{\nu}(\mathscr{X})}\right) + \sum_{s=1}^{S} w_s \log\big(1+h\big(1-\gamma_s\big)\hat{\nu}(\mathscr{X})\big)$$

If (8) is satisfied, then for h small enough we get $\mathfrak{L}(\nu_h) > \mathfrak{L}(\nu)$, a contradiction with the hypothesis that ν is the maximum of $\mathfrak{L}$.

For the proof of optimality let us assume, without loss of generality, that $S = 1$. We will show that there exists measures μ_X and μ_Y such that $\mathfrak{L}(\nu)$ reaches its maximum for $\nu \not\ll \mu_X$. The inequality in the hypothesis being strict, there exist $y_+ \in \mathscr{Y}$, $x_-, x_+ \in \mathscr{X}$ and $\alpha < \beta$ such that $\alpha = f(x_-, y_+)$, $\beta = f(x_+, y_+)$ and $w_1\frac{\beta}{\alpha} > 1$. Let $\mu_X := \delta_{x_-}$ and $\mu_Y := \delta_{y_+}$. The only probability measure that is absolutely continuous with respect to μ_X is $\mu_X = \delta_{x_-}$ itself and $f_Y(y;\mu_X) = f(y|x_-)$, so we have

$$\mathfrak{L}(\mu_X) = w_1\log(\alpha)$$

Compare this with the log-likelihood obtained for $\nu_h := h\delta_{x_+} + (1-h)\delta_{x_-}$, $0 < h < 1$: then $f_Y(y;\nu_h) = hf(y|x_+) + (1-h)f(y|x_-)$ and

$$\begin{aligned}
\mathfrak{L}(\nu_h) &= w_0\log(1-h) + w_1\log\big(h\beta + (1-h)\alpha\big) \\
&= -hw_0 + hw_1\left(\frac{\beta}{\alpha}-1\right) + w_1\log(\alpha) + o(h) \\
&> w_1\log(\alpha) = \mathfrak{L}(\mu_X)
\end{aligned}$$

if h is small enough. But $\nu_h \not\ll \mu_X$ if $h > 0$.

□

An interpretation of Hypothesis (8) may be easier in the equivalent form $w_0 > \sum_{s=1}^{S} w_s(\gamma_s - 1)$: the weight assigned to the direct observations must be large enough to ensure that the optimal ν stays absolutely continuous with respect to the direct observation measure.

From now on we suppose that (8) is satisfied.

2.3 Uniqueness and fixed point equation

Corollary 2.3. *The maximum of $\mathfrak{L}(\nu)$ is unique.*

Proof. The functional $g \mapsto \int_{\mathscr{X}} \log(g(x))d\mu_X(x)$ is strictly concave and proper on the set of μ_X-probability densities and the other terms are concave $> -\infty$, so the maximum of

$$\mathscr{L}(g) := w_0 \int_{\mathscr{X}} \log\big(g(x)\big)d\mu_X(x) + \sum_{s=1}^{S} w_s \int_{\mathscr{Y}_s} \log(f_{Y_s}(y; g \diamond \mu_X))d\mu_{Y_s}(y) \quad (9)$$

(if it exists) is unique. So is the maximum of $\mathfrak{L}(\nu)$, for we know that it is reached at $\nu =: g_1 \diamond \mu_X$. Moreover, this g_1 realises the maximum of (9) on $L^1(\mu_X)$. □

Corollary 2.4. *The fixed point equation*

$$g = \frac{w_0}{1 - \sum_{s=1}^{S} w_s \int_{\mathscr{Y}_s} \frac{f_s(y|\cdot)}{f_{Y_s}(y; g \diamond \mu_X)} d\mu_{Y_s}(y)} \quad (10)$$

has a solution $g_0 \in C_B(\mathscr{X})$, which coincides on the support of μ_X with the density maximising (9).

Proof. By Corollary 2.3, the maximum of $\mathscr{L}$ is reached for a density $g_1 \in L^1(\mu_X)$ and since $\mathscr{L}$ is differentiable at that point (obviously at the maximum, $g_1(x) > 0$ on the support of μ_X), the Lagrange multipliers theorem asserts that there exists $\lambda \in \mathbb{R}$ such that, for μ_X-almost all x,

$$\nabla_g \mathscr{L}(g_1)(x) = \frac{w_0}{g_1(x)} + \sum_{s=1}^{S} w_s \int_{\mathscr{Y}_s} \frac{f_s(y|x)}{f_{Y_s}(y; g_1 \diamond \mu_X)} d\mu_{Y_s}(y) = \lambda$$

and the fact that $\lambda = 1$ follows from the constraint $\int_{\mathscr{X}} g_1(x)d\mu_X(x) = 1$.

So far $g_1(x)$ is defined only μ_X-almost everywhere: for any $x \in \mathscr{X}$ let

$$g_0(x) := \frac{w_0}{1 - \sum_{s=1}^{S} w_s \int_{\mathscr{Y}_s} \frac{f_s(y|x)}{f_{Y_s}(y; g_1 \diamond \mu_X)} d\mu_{Y_s}(y)}$$

With Hypotheses (6) and (8), this function is continuous in x and bounded by $\frac{w_0}{1 - \sum_{s=1}^{S} w_s \gamma_s} < \infty$. Clearly g_0 and g_1 coincide μ_X-almost everywhere so g_0 also satisfies (10). □

Assuming the fixed point is attractive (see Lemmata 3.2 and 3.3 for sufficient conditions), (10) can be used iteratively to compute g_0 numerically.

3 Implicit equation for the density

To lighten the notation we assume (without loss of generality) that $S = 1$, drop the subscript s in (2) and (7) and, as long as μ_X is fixed, write $f_Y(y; g)$ instead of $f_Y(y; g \diamond \mu_X)$. By Corollary 2.4, the optimal density extended to a continuous function g_0 satisfies $A(g_0) = 0$, where

$$A(g) := g - \frac{w_0}{1 - w_1 \int_{\mathscr{Y}} \frac{f(y|\cdot)}{f_Y(y;g)} d\mu_Y(y)} \tag{11}$$

In this way g_0 is implicitly defined as a function of μ_X, μ_Y and optionally a parameter θ if $f(y|x) = f(y|x;\theta)$. We now prove C^1 regularity of this implicit function and compute its differential with respect to these variables (unless otherwise specified, differentiability will always be in the sense of Fréchet).

3.1 Invertibility of $\nabla_g A$

Initially the map A is defined on the set of continuous bounded functions that are also μ_X-probability densities

$$\Pi(\mu_X) := \left\{ g \in C_B(\mathscr{X}) : g \geq 0 \int_{\mathscr{X}} g(x) d\mu_X(x) = 1 \right\}$$

However, this set is not open (for the topology of uniform convergence) in $C_B(\mathscr{X})$ so it is convenient, for the purpose of differentiability, to enlarge it a little.

Lemma 3.1. *There exists a $C_B(\mathscr{X})$-open set $\Omega(\mu_X) \supset \Pi(\mu_X)$, such that A is differentiable on $\Omega(\mu_X)$ and its differential is the operator $\nabla_g A(g)$ defined by*

$$\nabla_g A(g)h := h + \int_{\mathscr{X}} \kappa(g)(\cdot, z) h(z) d\mu_X(z) \tag{12}$$

where

$$\kappa(g)(x,z) := \frac{w_0 w_1 \int_{\mathscr{Y}} \frac{f(y|x) f(y|z)}{f_Y(y;g)^2} d\mu_Y(y)}{\left(1 - w_1 \int_{\mathscr{Y}} \frac{f(y|x)}{f_Y(y;g)} d\mu_Y(y)\right)^2}$$

Proof. When $S = 1$, (8) reduces to $w_1 \gamma < 1$ and there exists $\epsilon > 0$ small enough so that $\frac{\gamma}{1 - 2\gamma\epsilon} \leq \frac{\gamma + 1/w_1}{2}$. Let

$$\Omega(\mu_X) := \Pi(\mu_X) + \left\{ h \in C_B(\mathscr{X}) : \|h\|_{L^\infty} < \epsilon \right\}$$

Consider the map $\phi : C_B(\mathscr{X}) \to C_B(\mathscr{X})$ defined by

$$\phi(g)(x) := \int_{\mathscr{Y}} \frac{f(y|x)}{f_Y(y;g)} d\mu_Y(y)$$

If $g \in \Pi(\mu_X)$, then for all $y \in \mathscr{Y}$,

$$\alpha(y) \le f_Y(y;g) \le \beta(y)$$

whereas for a general $h \in C_B(\mathscr{X})$ only $\left|f_Y(y;h)\right| \le \beta(y)\|h\|_{L^\infty}$ holds. Thus

$$\alpha(y) - \beta(y)\|h\|_{L^\infty} \le f_Y(y;g+h) \le \beta(y) + \beta(y)\|h\|_{L^\infty} \tag{13}$$

and if furthermore $\|h\|_{L^\infty} < \epsilon$,

$$\frac{\alpha(y)}{\beta(y)(1+\epsilon)} \le \frac{f(y|x)}{f_Y(y;g+h)} \le \frac{\beta(y)}{\alpha(y)-\beta(y)\epsilon}$$

So finally if $g \in \Omega(\mu_X)$

$$\frac{1}{\gamma(1+\epsilon)} \le \phi(g) \le \frac{\gamma}{1-\gamma\epsilon} \tag{14}$$

This map is actually differentiable at any $g \in \Omega(\mu_X)$ because as soon as $\|h\|_{L^\infty} < \epsilon$,

$$\begin{aligned} 0 \le \phi(g+h)(x) - \phi(g)(x) + \int_{\mathscr{Y}} \frac{f(y|x)f_Y(y;h)}{f_Y(y;g)^2} d\mu_Y(y) \\ = \int_{\mathscr{Y}} \frac{f(y|x)f_Y(y;h)^2}{f_Y(y;g)^2 f_Y(y;g+h)} d\mu_Y(y) \le \left(\frac{\gamma}{1-2\gamma\epsilon}\right)^3 \|h\|_{L^\infty}^2 \end{aligned}$$

and the continuous linear map $\nabla_g \phi(g) : h \mapsto -\int_{\mathscr{Y}} \frac{f(y|\cdot)f_Y(y;h)}{f_Y(y;g)^2} d\mu_Y(y)$ is its differential. If $g \in \Omega(\mu_X)$ then $\left\|\phi(g)\right\|_{L^\infty} \le \frac{\gamma+1/w_1}{2} < \frac{1}{w_1}$, so the map $\psi : \beta \mapsto \frac{w_0}{1-w_1\beta}$ is defined and differentiable at $\beta = \phi(g)$; its differential is $\nabla_\beta \psi(\beta) : h \mapsto \frac{w_0 w_1 h}{(1-w_1\beta)^2}$. The announced result follows by applying the chain rule to $\psi \circ \phi$. □

Lemma 3.2. *Suppose that $\mathscr{X}$ is compact or that $\frac{w_0 w_1 \gamma^2}{(1-w_1\gamma)^2} < 1$. Then for all $g \in \Pi(\mu_X)$, the operator $\nabla_g A(g)$ is an isomorphism: $C_B(\mathscr{X}) \to C_B(\mathscr{X})$.*

Proof. First assume that $\mathscr{X}$ is compact. Hypothesis (6) and the bound on $f_Y(y;g)$ imply that $\kappa(g)$ is uniformly continuous on the compact set $\mathscr{X}^2$. Then by the Arzelà-Ascoli theorem the integral operator

$$T(g) : h \mapsto \int_{\mathscr{X}} h(z)\kappa(g)(x,z) d\mu_X(z)$$

is compact: $C_B(\mathscr{X}) \to C_B(\mathscr{X})$. Thus $\nabla_g A(g) = \mathrm{Id} + T(g)$ is Fredholm of index zero [13, 5.C].

On the other hand $T(g)$ is composed of a positive multiplier and the integral operator $S(g)$ with kernel $\lambda_g(x,z) := \int_{\mathscr{Y}} \frac{f(y|x)f(y|z)}{f_Y(y;g)^2} d\mu_Y(y)$. In fact $S(g)$

is the second derivative of the convex map $g \mapsto -\int_{\mathscr{Y}} \log(f_Y(y;g))d\mu_Y(y)$ so by Kachurovskii's theorem $S(g)$ is positive (on the Banach algebra $C_B(\mathscr{X})$) and so is $T(g)$. Therefore $\nabla_g A(g) = \mathrm{Id} + T(g)$ is injective and by Fredholm's alternative it is also surjective; in addition $(\mathrm{Id}+T(g))^{-1}$ is continuous by the closed graph theorem.

If $\mathscr{X}$ fails to be compact, then even though $\kappa(g)$ may be uniformly continuous on $\mathscr{X}^2$, the Arzelà-Ascoli theorem asserts only the compactness of $T(g) : C_B(\mathscr{X}) \to \widetilde{C_B}(\mathscr{X})$, where $\widetilde{C_B}(\mathscr{X})$ is the space of bounded continuous functions on $\mathscr{X}$ endowed with the weaker topology of uniform convergence on compact subsets of $\mathscr{X}$; this space is no longer a Banach space and Fredholm's alternative doesn't apply anymore. The second hypothesis, $\frac{w_0 w_1 \gamma^2}{(1-w_1\gamma)^2} < 1$, can be used as a backup in that case: it implies that $\left\|\kappa(g)\right\|_{L^\infty} < 1$ so $T(g)$ has norm < 1 and $\mathrm{Id}+T(g)$ is invertible. □

Remark: the same proofs shows that A is also differentiable on a $L^1(\mu_X)$-open enlargement of $\Pi(\mu_X)$ and, when $\mathscr{X}$ is compact, that its differential $\nabla_g A(g)$ is a bijection: $L^1(\mu_X) \to L^1(\mu_X)$.

At "true value" If, as it is often the case in statistics, we are specifically interested in differentiability when μ_X and μ_Y are the true probability distributions for the model, then invertibility of $\nabla_g A$ is easier (this aspect is studied in more detail, with applications, in our other article [4]). We already saw that $g = 1$ is the maximizing density in that case.

Lemma 3.3. *Suppose that μ_X, μ_Y are such that μ_Y has density $f_Y(\cdot;1)$ with respect to ζ and that $w_1 < w_0$. Then $\nabla_g A(1)$ is an isomorphism: $C_B(\mathscr{X}) \to C_B(\mathscr{X})$.*

Proof. At true value and $g = 1$ we have for any $h \in C_B(\mathscr{X})$

$$\begin{aligned}
|T(1)h(x)| &= \left| \frac{w_0 w_1 \int_{\mathscr{X}} \int_{\mathscr{Y}} \frac{f(y|x)f(y|z)}{f_Y(y;1)^2} f_Y(y;1)d\zeta(y)h(z)d\mu_X(z)}{\left(1 - w_1 \int_{\mathscr{Y}} \frac{f(y|x)}{f_Y(y;1)} f_Y(y;1)d\zeta(y)\right)^2} \right| \\
&\leq \frac{w_0 w_1 \int_{\mathscr{Y}} f(y|x) \frac{\int_{\mathscr{X}} f(y|z)|h(z)|d\mu_X(z)}{f_Y(y;1)} d\zeta(y)}{\left(1 - w_1 \int_{\mathscr{Y}} f(y|x)d\zeta(y)\right)^2} \\
&\leq \frac{w_1}{w_0} \|h\|_{L^\infty}
\end{aligned}$$

so $T(1)$ has norm $\leq \frac{w_1}{w_0} < 1$ and $\mathrm{Id}+T(1)$ is invertible. □

3.2 Differentiability with respect to θ

We now suppose that the conditional density f also depends on some parameter θ in an open subset Θ of $\mathbb{R}^d$. At this point we should recall the dependency on θ that was hidden in (11), namely

$$A(g;\theta) := g - \frac{w_0}{1 - w_1 \int_{\mathscr{Y}} \frac{f(y|\cdot;\theta)}{f_Y(y;g,\theta)} d\mu_Y(y)}$$

with $f_Y(y;g,\theta) := \int_{\mathscr{X}} f(y|x;\theta) g(x) d\mu_X(x)$.

Lemma 3.4. *Suppose that the map $\mathfrak{f} : \Theta \to L^\infty(\mathscr{X} \times \mathscr{Y})$ defined by*

$$\mathfrak{f}(\theta) : (x,y) \mapsto \frac{f(y|x;\theta)}{\beta(y)}$$

is C^1 in θ. Then f_Y is differentiable in θ and

$$\nabla_\theta f_Y(y;g,\theta) = \int_{\mathscr{X}} \nabla_\theta f(y|x;\theta) g(x) d\mu_X(x) \tag{15}$$

Moreover, A is also differentiable in θ and its differential $\nabla_\theta A : \Omega(\mu_X) \times \Theta \to \mathscr{L}(\mathbb{R}^d, C_B(\mathscr{X}))$ is given by

$$\nabla_\theta A(g,\theta) := w_0 w_1 \frac{\int_{\mathscr{Y}} \frac{f(y|\cdot;\theta)\nabla_\theta f_Y(y;g,\theta)}{\left(f_Y(y;g,\theta)\right)^2} - \frac{\nabla_\theta f(y|\cdot;\theta)}{f_Y(y;g,\theta)} d\mu_Y(y)}{\left(1 - w_1 \int_{\mathscr{Y}} \frac{f(y|\cdot;\theta)}{f_Y(y;g,\theta)} d\mu_Y(y)\right)^2} \tag{16}$$

Proof. Since $\nabla_\theta \mathfrak{f}$ is continuous it is locally bounded, more precisely for each $\bar\theta \in \Theta$ there exists $U \supset \bar\theta$ and a constant C such that for all $\theta \in U$ and $(x,y) \in \mathscr{X} \times \mathscr{Y}$, $\left\|\nabla_\theta f(y|x;\theta)\right\| \le C\beta(y)$. By differentiating (cf. Lemma A.1) under the integral defining f_Y we obtain (15) and $\left\|\nabla_\theta f_Y(y;g,\theta)\right\| \le C\beta(y)$.

Let $Q : \Omega(\mu_X) \times \Theta \to L^\infty(\mathscr{X} \times \mathscr{Y})$ be defined by $Q(g,\theta) : (x,y) \mapsto \frac{f(y|x;\theta)}{f_Y(y;g,\theta)}$. Clearly Q is C^1 in θ and thanks to (7)

$$\nabla_\theta Q(g,\theta)(x,y) = \frac{\nabla_\theta f(y|x;\theta)}{f_Y(y;g,\theta)} - \frac{f(y|x;\theta)\nabla_\theta f_Y(y;g,\theta)}{\left(f_Y(y;g,\theta)\right)^2}$$

is bounded (for $\theta \in U$ and for all x,y,g) by $C(\gamma + \gamma^2)$. Thus $\phi : (g,\theta) \mapsto \int_{\mathscr{Y}} Q(g,\theta)(\cdot,y) d\mu_Y(y)$ can also be differentiated under the integral and

$$\nabla_\theta \phi(g,\theta) := \int_{\mathscr{Y}} \frac{\nabla_\theta f(y|\cdot;\theta)}{f_Y(y;g,\theta)} - \frac{f(y|\cdot;\theta)\nabla_\theta f_Y(y;g,\theta)}{\left(f_Y(y;g,\theta)\right)^2} d\mu_Y(y)$$

Finally (16) is obtained by chain rule on $\psi \circ \phi$, as in the proof of Lemma 3.1. □

Lemma 3.5. *Suppose as in Lemma 3.4 that* $\mathfrak{f}$ *is* C^1 *in* θ*. Then the map:* $(g,\theta) \mapsto \nabla_\theta A(g,\theta)$ *is continuous on* $\Omega(\mu_X) \times \Theta$*.*

Proof. For $\nabla_\theta A$: first notice that

$$\sup_{y\in\mathscr{Y}} \frac{\left|f_Y(y;g_1,\theta_1) - f_Y(y;g_2,\theta_2)\right|}{\beta(y)} \leq \left\|g_1\right\|_{L^\infty} \left\|\mathfrak{f}(\theta_1) - \mathfrak{f}(\theta_2)\right\|_{L^\infty} + \left\|\mathfrak{f}(\theta_2)\right\|_{L^\infty} \left\|g_1 - g_2\right\|_{L^\infty}$$

so the map: $(g,\theta) \mapsto \frac{f_Y(\cdot;g,\theta)}{\beta(\cdot)}$ is continuous (and bounded from below by the function $\alpha - \epsilon\beta$, cf. (13)). It follows that Q (from the previous proof) is also continuous and so is ϕ, which we recall is bounded from below by $\frac{1}{\gamma(1+\epsilon)}$ (cf. (14)). By a similar reasoning one proves continuity in (g,θ) for the numerator of (16), hence the announced result. □

Lemma 3.6. *The map* $(g,\theta) \mapsto \nabla_g A(g,\theta)$ *is continuous on* $\Omega(\mu_X) \times \Theta$*.*

Proof. We focus on the operator-valued map T defined by

$$T(g,\theta)h := \int_{\mathscr{X}} h(z)\kappa(g,\theta)(\cdot,z)d\mu_X(z)$$

where κ is the $C_B(\mathscr{X}^2)$-valued function given by

$$\kappa(g,\theta) : (x,z) \mapsto \frac{w_0 w_1 \int_{\mathscr{Y}} \frac{f(y|x;\theta)f(y|z;\theta)}{f_Y(y;g,\theta)^2} d\mu_Y(y)}{\left(1 - w_1 \int_{\mathscr{Y}} \frac{f(y|x;\theta)}{f_Y(y;g,\theta)} d\mu_Y(y)\right)^2}$$

The numerator of κ is continuous by a similar reasoning as above and its denominator is exactly the same as that of (16), continuous and bounded from below, so κ itself is continuous. Hence the result for T and $\nabla_g A$. □

Theorem 3. *Suppose that the hypotheses of Lemmata 3.4 and (3.2 or 3.3) hold. Then the optimal density* $g_0 \in C_B(\mathscr{X})$ *solution of (10) is* C^1 *as a function of* θ*. Moreover,*

$$\nabla_\theta\, g_0(\theta) = -\left(\nabla_g A(g_0(\theta),\theta)\right)^{-1} \nabla_\theta A(g_0(\theta),\theta) \tag{17}$$

where $\nabla_g A$ *and* $\nabla_\theta A$ *are given by (12) and (16).*

Proof. Let $\bar{\theta}$ be fixed and let $\bar{g_0}$ be the (unique) corresponding solution to (10). Combining Lemmata 3.1, 3.4, 3.5 and 3.6 shows that A a C^1 function in a neighbourhood of $(\bar{g_0},\bar{\theta})$ and from Lemma (3.2 or 3.3) its differential (in the variable g) $\nabla_g A(\bar{g_0},\bar{\theta})$ is an isomorphism. By the implicit function theorem, there exists a neighbourhood of $\bar{\theta}$ and a $C_B(\mathscr{X})$-valued C^1 function: $\theta \mapsto g_0(\theta)$ defined on this neighbourhood, satisfying (10). Formula (17) is also a consequence of the implicit function theorem. □

3.3 Differentiability with respect to μ_X, μ_Y

To emphasise the dependency on $\mu := (\mu_X, \mu_Y)$ in (11) we write

$$A(g;\mu) := g - \frac{w_0}{1 - w_1 \int_{\mathscr{Y}} \frac{f(y|\cdot)}{f_Y(y; g \diamond \mu_X)} d\mu_Y(y)} \tag{18}$$

For the purpose of Fréchet differentiation of g with respect to the information measure μ, the space $C_B(\mathscr{X})'$ is now endowed with the *total variation* norm

$$\left\|\mu_X\right\|_{\text{tv}} := \sup_{\|f\|_{L^\infty} \leq 1} \left| \int_{\mathscr{X}} f(x) d\mu_X(x) \right|$$

which makes it a Banach space (this is the *strong* dual of $C_B(\mathscr{X})$). The same for $C_B(\mathscr{Y})'$ and on the product space, $\left\|\mu\right\|_{\text{tv}} := \left\|\mu_X\right\|_{\text{tv}} + \left\|\mu_Y\right\|_{\text{tv}}$. Obviously $\mathscr{M}_{\mathscr{X}} \times \mathscr{M}_{\mathscr{Y}}$ is not open in that space but as in Lemma 3.1 it is easy to find $\epsilon > 0$ such that for any $\nu \in B_{\text{tv}}(\mu, \epsilon) := \left\{\nu : \left\|\nu - \mu\right\|_{\text{tv}} < \epsilon\right\}$, for any $g \in \Omega(\mu_X)$, the denominator of (18) is bounded from below by some positive constant.

Lemma 3.7. *As a function of μ, A is differentiable and its differential is the linear operator acting on $\lambda := (\lambda_X, \lambda_Y)$ as*

$$\nabla_\mu A(g,\mu)\lambda = w_0 w_1 \frac{\int_{\mathscr{Y}} \frac{f_Y(y; g \diamond \tilde{\lambda}_X) f(y|\cdot)}{f_Y(y; g \diamond \mu_X)^2} d\mu_Y(y) - \int_{\mathscr{Y}} \frac{f(y|\cdot)}{f_Y(y; g \diamond \mu_X)} d\lambda_Y(y)}{\left(1 - w_1 \int_{\mathscr{Y}} \frac{f(y|\cdot)}{f_Y(y; g \diamond \mu_X)} d\mu_Y(y)\right)^2} \tag{19}$$

Proof. Let $\phi(g,\mu) := \int_{\mathscr{Y}} \frac{f(y|\cdot)}{f_Y(y; g \diamond \mu_X)} d\mu_Y(y)$. In μ_Y this map is linear, while in μ_X

$$\nabla_{\mu_X} \frac{f(y|\cdot)}{f_Y(y; g \diamond \mu_X)} : \lambda_X \mapsto -\frac{f_Y(y; g \diamond \tilde{\lambda}_X) f(y|\cdot)}{f_Y(y; g \diamond \mu_X)^2}$$

is a bounded linear operator ($\tilde{\lambda}_X$ denoting the μ_X-absolutely continuous part of λ_X). Differentiating under the integral (lemma A.1), we obtain a Fréchet differential

$$\nabla_\mu \phi(g,\mu) : \lambda \mapsto \int_{\mathscr{Y}} \frac{f(y|\cdot)}{f_Y(y; g \diamond \mu_X)} d\lambda_Y(y) - \int_{\mathscr{Y}} \frac{f_Y(y; g \diamond \tilde{\lambda}_X) f(y|\cdot)}{f_Y(y; g \diamond \mu_X)^2} d\mu_Y(y)$$

and we conclude as in the proof of Lemma 3.1 by composition with $\psi : \beta \mapsto \frac{w_0}{1 - w_1 \beta}$. □

Lemma 3.8. *The map: $(g,\mu) \mapsto \nabla_\mu A(g,\mu)$ is continuous on $\Omega(\mu_X) \times \mathscr{M}(\mathscr{X}) \times \mathscr{M}(\mathscr{Y})$.*

Proof. The map $\nu \mapsto f_Y(\cdot;\nu)$ being linear, $(g,\mu) \mapsto f_Y(\cdot; g \diamond \mu_X)$ is continuous and bounded from below by $C\alpha$ for some $C > 0$. Thus $\nabla_\mu \phi$ (see previous proof) is also continuous and the conclusion follows by composition with ψ. □

Lemma 3.9. *The map:* $(g,\mu) \mapsto \nabla_g A(g,\mu)$ *is continuous on* $\Omega(\mu_X) \times \mathscr{M}(\mathscr{X}) \times \mathscr{M}(\mathscr{Y})$.

Proof. As for Lemma 3.6 it suffices to show that the kernel κ is continuous, now as a $C_B(\mathscr{X}^2)$-valued function of (g,μ):

$$\kappa(g,\mu) : (x,z) \mapsto \frac{w_0 w_1 \int_{\mathscr{Y}} \frac{f(y|x)f(y|z)}{f_Y(y;g\diamond\mu_X)^2} d\mu_Y(y)}{\left(1 - w_1 \int_{\mathscr{Y}} \frac{f(y|x)}{f_Y(y;g\diamond\mu_X)} d\mu_Y(y)\right)^2}$$

Continuity is now obvious for the numerator and the denominator is the same as for $\nabla_\mu A$. □

Theorem 4. *Suppose that the hypotheses of Lemma (3.2 or 3.3) hold. Then the optimal density* $g_0 \in C_B(\mathscr{X})$ *solution of (10) is* C^1 *as a function of* μ *and*

$$\nabla_\mu g(\mu) = -\left(\nabla_g A(g(\mu),\mu)\right)^{-1} \nabla_\mu A(g(\mu),\mu) \tag{20}$$

where $\nabla_g A$ *and* $\nabla_\mu A$ *are given by (12) and (19).*

Proof. Let $\bar{\mu}$ be fixed and let $\bar{g_0}$ be the (unique) corresponding solution to (10). Combining Lemmata 3.1, 3.7 3.8 and 3.9 shows that A a C^1 function in a neighbourhood of $(\bar{g_0},\bar{\mu})$ and from Lemma (3.2 or 3.3) its differential (in the variable g) $\nabla_g A(\bar{g_0},\bar{\mu})$ is an isomorphism. By the implicit function theorem there exists a neighbourhood of $\bar{\mu}$ and a $C_B(\mathscr{X})$-valued C^1 function: $\mu \mapsto g_0(\mu)$ defined on this neighbourhood, satisfying (10). Formula (20) is also a consequence of the implicit function theorem. □

3.4 Joint differentiability

For simplicity, we treated differentiability in θ and μ separately. But joint differentiability is straightforward, if tedious, along the same lines.

Lemma 3.10. *The maps:* $(g,\theta,\mu) \mapsto \nabla_g A(g,\theta,\mu)$ *(resp.* $\nabla_\theta A$ *and* $\nabla_\mu A$*) are continuous (in operator norm) on* $\Omega(\mu_X) \times \Theta \times \mathscr{M}(\mathscr{X}) \times \mathscr{M}(\mathscr{Y})$.

Theorem 5. *The optimal density g solution of* $(\mathscr{P})$ *is* C^1 *as a function of* θ *and* μ*; its differential is the direct sum of (17) and (20).*

The proofs are left to the reader.

3.5 Hadamard differentiability in weaker topologies

For practical purposes, e.g. in probability or statistics, the total variation topology on the set of probability measures is much too strong (empirical measures cannot converge to nonatomic probabilities, for instance). On the other hand, most useful topologies are too weak to allow Fréchet differentiability. A good compromise is provided by the notion of Hadamard differentiability [3, 10, 11].

Definition. *Let E, F be topological vector spaces. A map $\phi : E \to F$ is* [sequentially] Hadamard differentiable *at $x \in E$ if there exists a bounded linear operator H such that*

$$\frac{1}{h}\big(\phi(x+hu) - \phi(x)\big) - Hu \to_{h\to 0} 0$$

uniformly in u on the [sequentially] compact subsets of E.

As a consequence of Lemma A.2, our results of strong Fréchet differentiability (Theorems 4 and 5) translate into weak* sequential Hadamard differentiability. Weak* compact sets in $C_B(\mathscr{X})'$ are not necessarily sequentially compact but if $C_B(\mathscr{X})'$ is endowed with a metrisable topology τ stronger than weak* (such as the uniform convergence or the Skorokhod topology for cumulative distribution functions), then its compact sets are sequentially compact (a fortiori in the weak* topology) and we get the τ-Hadamard differentiability equivalent of Theorems 4 and 5.

A Technical lemmata

Notation: for a Banach space E, we write E'_s (resp. E'_w) its topological dual endowed with the strong (resp. weak*) topology. Recall that E'_s is a Banach space normed by

$$\|A\|_{E'_s} := \sup_{\|u\|_E \le 1} |Au|$$

Lemma A.1. *Let $(\mathscr{X}, \mathscr{F}, \mu)$ be a measured space and let $U \subset E$ be open. Let $f : \mathscr{X} \times U \to \mathbb{R}$ be such that for μ-almost every $x \in \mathscr{X}$:*

(i). for all $u \in U$, $f(x,u)$ has a Gâteaux derivative $\nabla_u f(x,u)$;

(ii). the map $u \mapsto \nabla_u f(x,u)$ is continuous: $U \to E'_s$;

(iii). for all $u \in U$, $\left\|\nabla_u f(x,u)\right\|_{E'_s} \le h(x)$ for some $h \in L^1(\mu)$.

Then the map $u \mapsto F(u) := \int_{\mathscr{X}} f(x,u)d\mu(x)$ is Fréchet C^1 with differential

$$\nabla_u F(u) := \int_{\mathscr{X}} \nabla_u f(x,u) d\mu(x) \tag{21}$$

Note that in (21) we are integrating a Banach space (E'_s) valued function, so the integral is to be understood in the sense of Bochner.

Proof of Lemma A.1. Let $v \in E$: for $\epsilon > 0$ small enough, $u + \epsilon v \in U$, so by the mean value theorem $\left|f(x,u+\epsilon v) - f(x,u)\right| \le \epsilon \|v\|_E h(x)$. By Lebesgue's

dominated convergence theorem it follows that

$$\begin{aligned}\lim_{\epsilon\to 0}\frac{F(u+\epsilon v)-F(u)}{\epsilon} &= \int_{\mathscr{X}}\lim_{\epsilon\to 0}\frac{f(x,u+\epsilon v)-f(x,u)}{\epsilon}d\mu(x)\\ &= \int_{\mathscr{X}}\nabla_u f(x,u)v\ d\mu(x)\\ &= \nabla_u F(u)v\end{aligned}$$

where $\nabla_u F(u)$, defined by (21), is a bounded linear operator with norm

$$\left\|\nabla_u F(u)\right\|_{E_s'} \le \int_{\mathscr{X}} h(x)d\mu(x)$$

Moreover, by the dominated convergence theorem applied to the Bochner integral (21), $\nabla_u F(u)$ is continuous on U. By a classical result (see for instance Flett [2, 4.1.7]) this implies that F is actually Fréchet C^1 on U. □

Lemma A.2. *Any sequentially compact subset of E_w' is bounded in E_s'.*

Proof. If (μ_n) is a sequence converging in E_w', then for any $f \in E$ the sequence $\left(\int_{\mathscr{X}} f(x)d\mu_n(x)\right)$ is bounded, so by the Banach-Steinhaus theorem (μ_n) is bounded in E_s. If U is unbounded in E_s' it contains a sequence whose norm $\to\infty$: this sequence E_s' does not converge in E_w', nor does any of its subsequences and this proves that U cannot be sequentially compact in E_w'. □

References

[1] Breslow, N. E., and Holubkov, R. Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *J. Roy. Statist. Soc. Ser. B 59*, 2 (1997), 447–461.

[2] Flett, T. M. *Differential Analysis*. Cambridge Univ. Press, 1980.

[3] Gill, R. D. Non- and semi-parametric maximum likelihood estimators and the von Mises method (part 1). *Scand. J. Statist. 16*, 2 (1989), 97–128. With a discussion by J. A. Wellner and J. Præstgaard and a reply by the author.

[4] Hirose, Y., and Aubry, J.-M. On differentiability of implicitly defined function in semi-parametric profile likelihood estimation. In preparation, 2010.

[5] Lawless, J. F., Kalbfleisch, J. D., and Wild, C. J. Semiparametric methods for response-selective and missing data problems in regression. *J. R. Stat. Soc. Ser. B Stat. Methodol. 61*, 2 (1999), 413–438.

[6] POSNER, E. C. Random coding strategies for minimum entropy. *IEEE Trans. Inform. Theory 21*, 4 (1975), 388–391.

[7] PRENTICE, R. L., AND PYKE, R. Logistic disease incidence models and case-control studies. *Biometrika 66*, 3 (1979), 403–411.

[8] SCOTT, A. J., AND WILD, C. J. Fitting regression models to case-control data by maximum likelihood. *Biometrika 84*, 1 (1997), 57–71.

[9] SCOTT, A. J., AND WILD, C. J. Maximum likelihood for generalised case-control studies. *J. Statist. Plann. Inference 96*, 1 (2001), 3–27. Statistical design of medical experiments, II.

[10] SHAPIRO, A. On concepts of directional differentiability. *J. Optim. Theory Appl. 66*, 3 (1990), 477–487.

[11] VAN DER VAART, A. W. Efficiency and Hadamard differentiability. *Scandinavian Journal of Statistics 18*, 1 (1991), 63–75.

[12] WEAVER, M. A., AND ZHOU, H. An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling. *J. Amer. Statist. Assoc. 100*, 470 (2005), 459–469.

[13] ZEIDLER, E. *Applied functional analysis*, vol. 109 of *Applied Mathematical Sciences*. Springer, New York, 1995.

[14] ZHOU, H., WEAVER, M. A., QIN, J., LONGNECKER, M. P., AND WANG, M. C. A semiparametric empirical likelihood method for data from an outcome-dependent sampling scheme with a continuous outcome. *Biometrics 58*, 2 (2002), 413–421.

Corresponding author's e-mail address: `jmaubry@math.cnrs.fr`