The f-divergence expectation iteration scheme

Kamélia Daudel, Randal Douc, François Portier, François Roueff

Submitted to the Annals of Statistics
arXiv: arXiv:0000.0000

THE *F*-DIVERGENCE EXPECTATION ITERATION SCHEME

By Kamélia Daudel*, Randal Douc†, François Portier*, and François Roueff*.

*Télécom Paris** *and Télécom SudParis*†

This paper introduces the f-EI(ϕ) algorithm, a novel iterative algorithm which operates on measures and performs f-divergence minimisation in a Bayesian framework. We prove that for a rich family of values of (f, ϕ) this algorithm leads at each step to a systematic decrease in the f-divergence and show that we achieve an optimum. In the particular case where we consider a weighted sum of Dirac measures and the α-divergence, we obtain that the calculations involved in the f-EI(ϕ) algorithm simplify to gradient-based computations. Empirical results support the claim that the f-EI(ϕ) algorithm serves as a powerful tool to assist Variational methods.

1. Introduction. Bayesian statistics for complex models often induce intractable and hard-to-compute posterior densities which need to be approximated. Variational methods such as Variational Inference (VI) [1, 2] and Expectation Propagation (EP) [3, 4] consider this objective purely as a (non-convex) optimisation problem. These approaches seek to approximate the posterior density by a variational density q_θ, characterized by a set of variational parameters $\theta \in \mathsf{T}$, where T is the parameter space. In these methods θ is optimised such that it minimizes a certain divergence D between the posterior and the variational density, typically the Kullback-Leibler (KL) divergence [5].

Modern optimisation-based approximate inference methods improved in three major directions [6, 7]. Firstly, Variational methods used to be limited to conditionally conjugate exponential family models [8]. Monte Carlo methods and Black-Box inference techniques such as [9, 10] have since been deployed, rendering Variational methods applicable to a wide range of models. These methods use gradients which are computed through automatic differentiation tools and climb the Monte Carlo approximated Variational Bound to the log-likelihood.

Secondly, classical Variational methods focused on the KL divergence as an objective function. However, in the VI case, the KL is known to often

MSC 2010 subject classifications: Primary 60K35, 60K35; secondary 60K35
Keywords and phrases: sample, LaTeX 2ε

underestimate the variance and may miss important local modes of the true posterior [11, 6]. As for the EP algorithm, which performs local minimization, it is not guaranteed to converge and does not provide an easy estimate of the marginal likelihood [12]. Modern research consequently turned to more flexible families of divergence in order to reach better accuracy, the α-divergence [13, 14] and Renyi's α-divergence [15, 16] being such examples [17, 18, 19, 20, 21].

Thirdly, scalable methods such as Stochastic Variational Inference [11] or Stochastic Expectation Propagation [22, 23] have been developed to enable large scale learning. These methods rely on stochastic optimisation techniques [24, 25] and have been applied to complex probabilistic models, e.g Latent Dirichlet Allocation [26].

Framework. In this paper, we contribute in these three directions. The divergence we choose to work with is the f-divergence [27, 28], as it is a general family of divergences that encompasses the Kullback-Leibler, the reverse Kullback-Leibler and the α-divergence. Furthermore, we offer to change the space on which the minimization occurs. While the common approach consists in minimizing over the set of densities

$$\{y \mapsto q_\theta(y) \ : \ \theta \in \mathsf{T}\}$$

we consider instead a minimization over

$$\left\{y \mapsto \int_{\mathsf{T}} \mu(\mathrm{d}\theta) q_\theta(y) \ : \ \mu \in \mathsf{M}\right\}$$

where M is a convenient subset of $\mathrm{M}_1(\mathsf{T})$, the set of probability measures on T (and in this case, we equip T with a σ-field denoted by $\mathcal{T}$). In doing so, we extend the minimizing set to a larger space since a parameter θ can be identified with its associated Dirac measure δ_θ. The paper is then organised as follows:

- In Section 2, we briefly review basic concepts around the f-divergence. We formulate explicitly the general form of the optimisation problem we consider, which includes the particular case of f-divergence posterior density approximation.
- In Section 3, we provide a new iterative scheme which we call the f-EI(ϕ) algorithm that performs an update of measures. We establish sufficient conditions on (f, ϕ) for this algorithm to lead at each step to a systematic decrease in the f-divergence and we obtain its convergence to an optimum. As the exact f-EI(ϕ) algorithm involves an integral which might be intractable, we define its approximate version and show its convergence to the exact algorithm.

• In Section 4, we apply the f-EI(ϕ) algorithm to f-divergence density approximation. In the particular case of the α-divergence and when μ is chosen as a weighted sum of Dirac measures, we obtain that the computations involved in the f-EI(ϕ) algorithm mostly rely on gradient-based calculations.

• Finally, Section 5 is devoted to numerical experiments, where we explore the impact of the hyperparameter ϕ as well as the state space dimension d and the parameter space dimension J on the convergence of the f-EI(ϕ) algorithm.

2. Formulation of the optimisation problem. Let $(\mathsf{Y}, \mathcal{Y}, \nu)$ be a measured space, where ν is a σ-finite measure on $(\mathsf{Y}, \mathcal{Y})$ and let f be a *convex function* over $(0, \infty)$ that satisfies $f(1) = 0$. We start by defining the f-divergence between two probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$.

DEFINITION 1 (f-divergence). *Let $\mathbb{P}_1$ and $\mathbb{P}_2$ be two probability measures on $(\mathsf{Y}, \mathcal{Y})$ that are absolutely continuous with respect to ν i.e. $\mathbb{P}_1 \preceq \nu$, $\mathbb{P}_2 \preceq \nu$. Let us denote by $p_1 = \frac{\mathrm{d}\mathbb{P}_1}{\mathrm{d}\nu}$ and $p_2 = \frac{\mathrm{d}\mathbb{P}_2}{\mathrm{d}\nu}$ the Radon-Nikodym derivatives of $\mathbb{P}_1$ and $\mathbb{P}_2$ with respect to ν. The f-divergence between $\mathbb{P}_1$ and $\mathbb{P}_2$ is defined as follows :*

$$D_f(\mathbb{P}_1||\mathbb{P}_2) = \int_{\mathsf{Y}} f\left(\frac{p_1(y)}{p_2(y)}\right) p_2(y)\nu(\mathrm{d}y) \,. \tag{1}$$

In (1), we adopt the conventional notation $0f(\frac{0}{0}) = 0$ and $0f(\frac{a}{0}) = \lim_{t\downarrow 0} tf(\frac{a}{t}) = a\lim_{u\to\infty} \frac{f(u)}{u}$ for all $a > 0$. We also extend the definition of f at 0 by setting $f(0) = \lim_{t\downarrow 0} f(t) \in (-\infty, \infty]$.

We now recall without proof a few results around the f-divergence and we refer to [29, 30] and [31] for more details on the f-divergence family.

PROPOSITION 2. *The f-divergence is always non-negative and for f strictly convex, it is equal to zero if and only if $\mathbb{P}_1 = \mathbb{P}_2$. Furthermore, it is jointly convex in $\mathbb{P}_1$ and $\mathbb{P}_2$ and $D_f(\mathbb{P}_1||\mathbb{P}_2) = D_{\tilde{f}}(\mathbb{P}_2||\mathbb{P}_1)$, where $\tilde{f}(u) = uf(\frac{1}{u})$ is called the conjugate function.*

Special cases include the Kullback-Leibler (KL) divergence D_{KL}, the reverse Kullback-Leibler divergence D_{rKL} and the α-divergence $D_A^{(\alpha)}$, where $\alpha \in \mathbb{R} \setminus \{0, 1\}$ (see Table 1). The Hellinger distance [32, 33] and the χ^2-divergence [21] correspond respectively to order $\alpha = 0.5$ and $\alpha = 2$ of the α-divergence. Moreover, the definition of the α-divergence can be extended to $\alpha \in \{0, 1\}$ by continuity and we recover the Kullback-Leibler when $\alpha \to 1$

TABLE 1
Special cases in the f-divergence family

$f(u)$	Corresponding divergence
$u\log(u)$	$D_{KL}(\mathbb{P}_1\|\|\mathbb{P}_2) = \int_{\mathsf{Y}} \log\left(\frac{p_1(y)}{p_2(y)}\right) p_1(y)\nu(\mathrm{d}y)$
$-\log(u)$	$D_{rKL}(\mathbb{P}_1\|\|\mathbb{P}_2) = \int_{\mathsf{Y}} -\log\left(\frac{p_1(y)}{p_2(y)}\right) p_2(y)\nu(\mathrm{d}y)$
$\frac{1}{\alpha(\alpha-1)}[u^\alpha - 1]$	$D_A^{(\alpha)}(\mathbb{P}_1\|\|\mathbb{P}_2) = \frac{1}{\alpha(\alpha-1)}\left[\int_{\mathsf{Y}} \left(\frac{p_1(y)}{p_2(y)}\right)^\alpha p_2(y)\nu(\mathrm{d}y) - 1\right]$

and the reverse Kullback-Leibler when $\alpha \to 0$.

Now consider a measurable space $(\mathsf{T}, \mathcal{T})$. Let p be a measurable positive function on $(\mathsf{Y}, \mathcal{Y})$ and $Q : (\theta, A) \mapsto \int_A q(\theta, y)\nu(\mathrm{d}y)$ be a Markov transition kernel on $\mathsf{T} \times \mathcal{Y}$ with kernel density q defined on $\mathsf{T} \times \mathsf{Y}$. Moreover, for all $y \in \mathsf{Y}$, we denote $\mu q(y) = \int_{\mathsf{T}} \mu(\mathrm{d}\theta) q(\theta, y)$ and we define, for all $\mu \in \mathrm{M}_1(\mathsf{T})$,

$$\Psi^{(f)}(\mu) = \int_{\mathsf{Y}} f\left(\frac{\mu q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y) \ . \tag{2}$$

Note that p, q and ν appear as well in $\Psi^{(f)}(\mu)$ i.e $\Psi^{(f)}(\mu) = \Psi^{(f)}(\mu; p, q, \nu)$, but we drop them for notational ease and when no ambiguity occurs. Notice also that we replaced $q_\theta(y)$ by $q(\theta, y)$ to comply with usual kernel notation.

We consider in what follows the general optimisation problem

$$\mathrm{arginf}_{\mu \in \mathsf{M}} \Psi^{(f)}(\mu) \ , \tag{3}$$

where p is a measurable positive function on $(\mathsf{Y}, \mathcal{Y})$.

This framework includes the particular case of f-divergence posterior density approximation. Indeed, let us consider the *posterior density* of the latent variables y given the data $\mathscr{D}$:

$$p(y|\mathscr{D}) = \frac{p(\mathscr{D}, y)}{p(\mathscr{D})} \ ,$$

where $p(\mathscr{D}) := \int_{\mathsf{Y}} p(\mathscr{D}, y)\nu(\mathrm{d}y)$ is *marginal likelihood* or *model evidence* (whose value is unknown). Now denoting by $\mathbb{P}$ and μQ the probability measures on $(\mathsf{Y}, \mathcal{Y})$ with corresponding associated density $p(\cdot|\mathscr{D})$ and μq with respect to ν and setting $p = p(\cdot|\mathscr{D})$ in $\Psi^{(f)}(\mu)$, the optimisation problem defined by (3) can be rewritten as

$$\mathrm{arginf}_{\mu \in \mathsf{M}} D_f(\mu Q, \mathbb{P}) \ .$$

At this stage, a first remark is that the convexity of $\Psi^{(f)}$ is straightforward from the convexity of f. Therefore, a simple yet powerful consequence of enlarging the variational family is that the optimisation problem now involves the *convex* mapping

$$\mu \mapsto \Psi^{(f)}(\mu) = \int_{\mathsf{Y}} f\left(\frac{\mu q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y) \ ,$$

whereas the initial optimisation problem was associated to the mapping $\theta \mapsto \int_{\mathsf{Y}} f\left(\frac{q_\theta(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$, which is not necessarily convex.

We now move on to Section 3, where we describe the f-EI(ϕ) algorithm and state our main theoretical results.

3. The f-Expectation Iteration algorithm f-EI(ϕ) .

3.1. *An iterative algorithm for optimising* $\Psi^{(f)}$. The following set of assumptions will be in force throughout the paper. We first gather the assumptions on q, p and ν.

(A1) The density kernel q on $\mathsf{T} \times \mathsf{Y}$, the function p on Y and the σ-finite measure ν on $(\mathsf{Y}, \mathcal{Y})$ satisfy, for all $(\theta, y) \in \mathsf{T} \times \mathsf{Y}$, $q(\theta, y) > 0$, $p(y) > 0$ and $\int_{\mathsf{Y}} p(y)\nu(\mathrm{d}y) < \infty$.

The next assumption mostly retains the assumptions on the function f used to define $\Psi^{(f)}$ in (2).

(A2) The function $f : (0, \infty) \to \mathbb{R}$ is monotonous, strictly convex and continuously differentiable, and $f(1) = 0$.

Under (A1) and (A2), we immediately obtain a lower bound on $\Psi^{(f)}$.

LEMMA 3. *Suppose that (A1) and (A2) hold. Then, for all* $\mu \in \mathrm{M}_1(\mathsf{T})$, *we have*

$$\Psi^{(f)}(\mu) \geqslant \tilde{f}\left(\int_{\mathsf{Y}} p(y)\nu(\mathrm{d}y)\right) > -\infty \ ,$$

where $\tilde{f}$ *is defined in Proposition 2.*

PROOF. Since $\tilde{f}(u) = uf(1/u)$, we have

$$\Psi^{(f)}(\mu) = \int_{\mathsf{Y}} \tilde{f}\left(\frac{p(y)}{\mu q(y)}\right) \mu q(y)\nu(\mathrm{d}y) \ .$$

Recalling that f and hence $\tilde{f}$, is convex on $\mathbb{R}_{>0}$, Jensen's inequality applied to $\tilde{f}$ yields $\Psi^{(f)}(\mu) \geqslant \tilde{f}\left(\int_{\mathsf{Y}} p(y)\nu(\mathrm{d}y)\right) > -\infty$. □

REMARK 4. *Assumption (A1) can be extended by discarding the assumption that $p(y)$ is positive for all $y \in \mathsf{Y}$. As it complicates the expression of the constant appearing in the bound without increasing dramatically the degree of generality of the results, we chose to maintain this assumption for the sake of simplicity.*

Thus, if there exists a sequence of probability measures $\{\mu_n \,:\, n \in \mathbb{N}\}$ on $(\mathsf{T}, \mathcal{T})$ such that $\Psi^{(f)}(\mu_0) < \infty$ and $\Psi^{(f)}(\mu_n)$ is non-increasing with n, Lemma 3 guarantees that this sequence converges to a limit in $\mathbb{R}$. We now focus on constructing such a sequence $\{\mu_n \,:\, n \in \mathbb{N}\}$.

For this purpose, let $\phi \in \mathbb{R}^\star := \mathbb{R} \setminus \{0\}$. The one-step transition of the *f-Expectation Iteration algorithm* f-EI(ϕ) can be formally described as an *expectation* step and an *iteration* step:

Algorithm 1: *Exact f-EI(ϕ) one-step transition*

1. Expectation step : $b_\mu(\theta) = \displaystyle\int_{\mathsf{Y}} q(\theta, y) f'\left(\frac{\mu q(y)}{p(y)}\right) \nu(\mathrm{d}y)$
2. Iteration step : $\mathcal{I}^\phi(\mu)(\mathrm{d}\theta) = \dfrac{\mu(\mathrm{d}\theta) \cdot |b_\mu(\theta)|^\phi}{\mu(|b_\mu|^\phi)}$

Given any initial measure $\mu \in \mathrm{M}_1(\mathsf{T})$ such that $\Psi^{(f)}(\mu) < \infty$, the iterative sequence of probability measures $(\mu_n)_{n\in\mathbb{N}}$ is then defined by setting

$$\begin{cases} \mu_0 = \mu, \\ \mu_{n+1} = \mathcal{I}^\phi(\mu_n)\,, \qquad n \in \mathbb{N}\,. \end{cases} \tag{4}$$

Note that under (A1) and (A2), b_μ is well-defined (since f' is of constant sign) and $|b_\mu| \in (0, \infty]$ for all $\mu \in \mathrm{M}_1(\mathsf{T})$. Given $\phi \in \mathbb{R}^\star$, the iteration $\mu \mapsto \mathcal{I}^\phi(\mu)$ is thus well-defined if moreover we have

$$0 < \mu(|b_\mu|^\phi) < \infty\,. \tag{5}$$

In the following part, we investigate some core properties of the aforementioned sequence of probability measures $(\mu_n)_{n\in\mathbb{N}}$. For all $\mu \in \mathrm{M}_1(\mathsf{T})$ such that $\Psi^{(f)}(\mu) < \infty$, we identify couples (f, ϕ) such that (i) f satisfies (A2), (ii) the sequence $(\mu_n)_{n\in\mathbb{N}}$ is well-defined and (iii) the f-EI(ϕ) algorithm diminishes $\Psi^{(f)}(\mu_n)$ at each iteration.

3.2. *Monotonicity.* To establish that the f-EI(ϕ) algorithm diminishes $\Psi^{(f)}(\mu_n)$ at each iteration, we first derive a general upper-bound for the difference $\Psi^{(f)}(\zeta) - \Psi^{(f)}(\mu)$. Here, (ζ, μ) is a couple of probability measures where ζ is dominated by μ, which we denote by $\zeta \preceq \mu$.

This first result relies on the existence of an exponent ϱ satisfying condition (A3) below, which will later on be used to specify a range of values for ϕ satisfying (5) for any $\mu \in \mathrm{M}_1(\mathsf{T})$.

(A3) We have $\varrho \in \mathbb{R}\backslash[0,1]$ and the function $f_\varrho : u \mapsto f(u^{1/\varrho})$ is non-decreasing and concave on $\mathbb{R}_{>0}$.

PROPOSITION 5. *Assume (A1), (A2) and (A3). Then, for all $\mu, \zeta \in \mathrm{M}_1(\mathsf{T})$ such that $\mu(|b_\mu|) < \infty$ and $\zeta \preceq \mu$,*

$$\Psi^{(f)}(\zeta) \leqslant \Psi^{(f)}(\mu) + |\varrho|^{-1} \left\{ \mu\left(|b_\mu| g^\varrho\right) - \mu(|b_\mu|) \right\} . \tag{6}$$

where g is the density of ζ wrt μ, i.e. $\zeta(\mathrm{d}\theta) = \mu(\mathrm{d}\theta) g(\theta)$. Moreover, equality holds in (6) if and only if $\zeta = \mu$.

PROOF. Under (A2) f' is of constant sign and under (A3), if $\varrho > 1$, f is non-decreasing and if $\varrho < 0$, f is non-increasing. This implies that for all $u > 0$ and all $\varrho \neq 0$, $\varrho^{-1} f'(u) = |\varrho|^{-1} |f'(u)|$, which will be used later in the proof.

Write by definition of f_ϱ in (A3) and ζ,

$$\begin{aligned}
\Psi^{(f)}(\zeta) &= \int_{\mathsf{Y}} f\left(\frac{\zeta q(y)}{p(y)}\right) p(y) \nu(\mathrm{d}y) \\
&= \int_{\mathsf{Y}} f_\varrho\left(\left[\frac{\zeta q(y)}{p(y)}\right]^\varrho\right) p(y) \nu(\mathrm{d}y) \\
&= \int_{\mathsf{Y}} f_\varrho\left(\left[\int_{\mathsf{T}} \mu(\mathrm{d}\theta) \frac{q(\theta, y)}{\mu q(y)} \left(\frac{g(\theta) \mu q(y)}{p(y)}\right)\right]^\varrho\right) p(y) \nu(\mathrm{d}y) \\
&\leqslant \int_{\mathsf{Y}} f_\varrho\left(\int_{\mathsf{T}} \mu(\mathrm{d}\theta) \frac{q(\theta, y)}{\mu q(y)} \left(\frac{g(\theta) \mu q(y)}{p(y)}\right)^\varrho\right) p(y) \nu(\mathrm{d}y)
\end{aligned} \tag{7}$$

where the last inequality follows from Jensen's inequality applied to the convex function $u \mapsto u^\varrho$ (since $\varrho \in \mathbb{R} \setminus [0,1]$) and the fact that f_ϱ is non-decreasing. Now set

$$\begin{aligned}
u_y &= \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \frac{q(\theta, y)}{\mu q(y)} \left(\frac{g(\theta) \mu q(y)}{p(y)}\right)^\varrho \\
v_y &= \left(\frac{\mu q(y)}{p(y)}\right)^\varrho
\end{aligned}$$

and note that

$$u_y - v_y = \left(\frac{\mu q(y)}{p(y)}\right)^{\varrho} \left(\int_{\mathsf{T}} \mu(\mathrm{d}\theta) \frac{q(\theta, y)}{\mu q(y)} g^{\varrho}(\theta) - 1\right) \tag{8}$$

Since f_ϱ is concave, $f_\varrho(u_y) \leqslant f_\varrho(v_y) + f'_\varrho(v_y)(u_y - v_y)$. Then, combining with (7), we get

$$\begin{aligned} \Psi^{(f)}(\zeta) &\leqslant \int_{\mathsf{Y}} f_\varrho(u_y) p(y) \nu(\mathrm{d}y) \\ &\leqslant \int_{\mathsf{Y}} f_\varrho(v_y) p(y) \nu(\mathrm{d}y) + \int_{\mathsf{Y}} f'_\varrho(v_y)(u_y - v_y) p(y) \nu(\mathrm{d}y) \end{aligned} \tag{9}$$

Note that the first term of the rhs can be written as

$$\int_{\mathsf{Y}} f_\varrho(v_y) p(y) \nu(\mathrm{d}y) = \int_{\mathsf{Y}} f\left(\frac{\mu q(y)}{p(y)}\right) p(y) \nu(\mathrm{d}y) = \Psi^{(f)}(\mu) \tag{10}$$

Using now $f'_\varrho(v_y) = \varrho^{-1} v_y^{1/\varrho - 1} f'(v_y^{1/\varrho})$ and (8), the second term of the rhs of (9) may be expressed as

$$\begin{aligned} &\int_{\mathsf{Y}} f'_\varrho(v_y)(u_y - v_y) p(y) \nu(\mathrm{d}y) \\ &\quad = \varrho^{-1} \int_{\mathsf{Y}} \left(\frac{\mu q(y)}{p(y)}\right)^{1-\varrho} f'\left(\frac{\mu q(y)}{p(y)}\right) \\ &\qquad\qquad \left(\frac{\mu q(y)}{p(y)}\right)^{\varrho} \left(\int_{\mathsf{T}} \mu(\mathrm{d}\theta) \frac{q(\theta, y)}{\mu q(y)} g^{\varrho}(\theta) - 1\right) p(y) \nu(\mathrm{d}y) \\ &\quad = \varrho^{-1} \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left(\int_{\mathsf{Y}} q(\theta, y) f'\left(\frac{\mu q(y)}{p(y)}\right) \nu(\mathrm{d}y)\right) g^{\varrho}(\theta) \\ &\qquad\qquad - \varrho^{-1} \int_{\mathsf{Y}} \mu q(y) f'\left(\frac{\mu q(y)}{p(y)}\right) \nu(\mathrm{d}y) \\ &\quad = |\varrho|^{-1} \left\{\mu\left(|b_\mu| g^{\varrho}\right) - \mu(|b_\mu|)\right\} . \end{aligned}$$

Combining this equality with (9) and (10) finishes the proof of the inequality.

If the equality holds in (6), then the equality in Jensen's inequality (7) shows that g is constant μ-a.e. so that $\zeta = \mu$, and the proof is completed. □

We now plan on setting $\zeta = \mathcal{I}^{\phi}(\mu)$ in Proposition 5 and obtain that one iteration of the f-EI(ϕ) algorithm yields $\Psi^{(f)} \circ \mathcal{I}^{\phi}(\mu) \leqslant \Psi^{(f)}(\mu)$. For this purpose and based on the upper bound obtained in Proposition 5, we strengthen the condition (5) as follows to take into account the exponent ϱ

$$0 < \mu(|b_\mu|^{\phi}) < \infty \quad \text{and} \quad \mu(|b_\mu| g^{\varrho}) \leqslant \mu(|b_\mu|) \quad \text{with } g = \tfrac{|b_\mu|^{\phi}}{\mu(|b_\mu|^{\phi})} \tag{11}$$

This leads to our first main theorem.

THEOREM 1. *Assume (A1), (A2) and (A3). Let $\mu \in \mathrm{M}_1(\mathsf{T})$ be such that $\mu(|b_\mu|) < \infty$ and let $\phi \in \mathbb{R}^\star$ satisfy (11). Then, the two following assertions hold.*

(i) We have $\Psi^{(f)} \circ \mathcal{I}^\phi(\mu) \leqslant \Psi^{(f)}(\mu)$.
(ii) We have $\Psi^{(f)} \circ \mathcal{I}^\phi(\mu) = \Psi^{(f)}(\mu)$ if and only if $\mu = \mathcal{I}^\phi(\mu)$.

PROOF. We apply Proposition 5 with $\zeta = \mathcal{I}^\phi(\mu)$ so that $\zeta(\mathrm{d}\theta) = \mu(\mathrm{d}\theta)g(\theta)$ with $g = |b_\mu|^\phi / \mu(|b_\mu|^\phi)$. Then,

$$\Psi^{(f)} \circ \mathcal{I}^\phi(\mu) \leqslant \Psi^{(f)}(\mu) + |\varrho|^{-1} \left\{ \mu\left(|b_\mu| g^\varrho\right) - \mu(|b_\mu|) \right\} \leqslant \Psi^{(f)}(\mu) \tag{12}$$

where the last inequality follows from condition (11).

Let us now show (ii). The *if* part is obvious. As for the *only if* part, $\Psi^{(f)} \circ \mathcal{I}^\phi(\mu) = \Psi^{(f)}(\mu)$ combined with (12) yields

$$\Psi^{(f)} \circ \mathcal{I}^\phi(\mu) = \Psi^{(f)}(\mu) + |\varrho|^{-1} \left\{ \mu\left(|b_\mu| g^\varrho\right) - \mu(|b_\mu|) \right\} \ ,$$

which is the case of equality in Proposition 5. Therefore, $\mathcal{I}^\phi(\mu) = \mu$. □

We are now able to derive our second main theorem.

THEOREM 2. *Assume that p and q are as in (A1). Let (f, ϕ) belong to any of the following cases.*

(i) Reverse Kullback-Leibler: $f(u) = -\log(u)$, and $\phi \in (0, 1]$.
(ii) α-Divergence: $f(u) = \frac{1}{\alpha(\alpha-1)}(u^\alpha - 1)$,
  (a) $\alpha \in (-\infty, -1]$ and $\phi \in (0, -1/\alpha]$;
  (b) $\alpha \in (-1, 1) \setminus \{0\}$ and $\phi \in (0, 1]$;
  (c) $\alpha \in (1, \infty)$ and $\phi \in (1/(1-\alpha), 0)$.

Then (A2) holds. Moreover, let $\mu \in \mathrm{M}_1(\mathsf{T})$ be such that $\Psi^{(f)}(\mu) < \infty$. Then the sequence $(\mu_n)_{n\in\mathbb{N}}$ defined by (4) is well-defined and the sequence $(\Psi^{(f)}(\mu_n))_{n\in\mathbb{N}}$ is non-increasing.

The proof of this theorem requires intermediate results, which are derived in Appendix A.1 alongside with the proof of Theorem 2.

REMARK 6. *In the proof of Theorem 2, we also prove along the way that either $\Psi^{(f)}(\mu_{n+1}) < \Psi^{(f)}(\mu_n)$ for all $n \geqslant 0$, or that there exists $n_0 < \infty$ such that $\mu_n = \mu_{n_0}$ for all $n \geqslant n_0$ i.e $\mu_{n_0} = \mathcal{I}^\phi(\mu_{n_0})$.*

The results we obtained at this point are summarized in Table 2.

We now study the limiting behavior of the f-EI(ϕ) algorithm for the iterative sequence of probability measure $(\mu_n)_{n\in\mathbb{N}}$.

TABLE 2
Allowed ranges for ϕ in the f-EI(ϕ) algorithm per divergence

Divergence considered		Corresponding range
Reverse KL $f(u) = -\log(u)$		$\phi \in (0, 1]$
α-divergence $f(u) = \frac{1}{\alpha(\alpha-1)}(u^\alpha - 1)$	$\alpha \in (-\infty, -1]$	$\phi \in (0, -1/\alpha]$
	$\alpha \in (-1, 1) \setminus \{0\}$	$\phi \in (0, 1]$
	$\alpha \in (1, \infty)$	$\phi \in (1/(1-\alpha), 0)$

3.3. *Limiting behavior of the Exact f-EI(ϕ) algorithm.* Let $\mu \in \mathrm{M}_1(\mathsf{T})$ and let us consider the iterative sequence of probability measures $(\mu_n)_{n\in\mathbb{N}}$ defined by (4). We seek to identify the possible weak limits of $(\mu_n)_{n\in\mathbb{N}}$. To do so, we introduce the following additive set of assumptions

(A4) (i) T is a compact metric space and $\mathcal{T}$ is the associated Borel σ-field;

(ii) for all $y \in \mathsf{Y}$, $\theta \mapsto q(\theta, y)$ is continuous;

(iii) we have $\displaystyle\int_{\mathsf{Y}} \sup_{\theta\in\mathsf{T}} \left| f\left(\frac{q(\theta,y)}{p(y)}\right)\right| p(y)\nu(\mathrm{d}y) < \infty$;

(iv) we have $\displaystyle\int_{\mathsf{Y}} \sup_{\theta\in\mathsf{T}} q(\theta,y) \times \left(\sup_{\theta'\in\mathsf{T}} \left| f'\left(\frac{q(\theta',y)}{p(y)}\right)\right|\right) \nu(\mathrm{d}y) < \infty$.

Notice that conditions (A4)-(iii) and (A4)-(iv) act as a uniform majoration of $\Psi^{(f)}(\mu)$ and $b_\mu(\theta)$ with respect to μ and θ. In addition, these two conditions are rather weak under (A4)-(i), since we consider in each a supremum taken over a compact set (and T will always be chosen as such in practice).

We define $\mathrm{M}_{1,\mu}(\mathsf{T})$ as the set of probability measures dominated by μ. We have the following theorem, which states that the possible weak limits of $(\mu_n)_{n\in\mathbb{N}}$ corresponds to the *global infimum* of $\Psi^{(f)}$.

THEOREM 3. *Assume (A1), (A2) and (A4). Then, for all $\zeta \in \mathrm{M}_1(\mathsf{T})$ any $\phi \in \mathbb{R}^\star$ satisfies (5) and $\Psi^{(f)}(\zeta) < \infty$.*

Further assume that there exists $\mu, \bar{\mu} \in \mathrm{M}_1(\mathsf{T})$ such that the (well-defined) sequence $(\mu_n)_{n\in\mathbb{N}}$ defined by (4) weakly converges to $\bar{\mu}$ as $n \to \infty$. Then the following assertions hold

(i) $\bar{\mu}$ is a fixed point of $\mathcal{I}^\phi$,
(ii) $\Psi^{(f)}(\bar{\mu}) = \inf_{\zeta\in\mathrm{M}_{1,\mu}(\mathsf{T})} \Psi^{(f)}(\zeta)$,

in any of the following cases:

(a) f non-increasing and $\phi > 0$,
(b) f non-decreasing and $\phi < 0$.

The proof of Theorem 3 can be found in Appendix A.2. Note that any couple (f, ϕ) described in Table 2 falls into one of the two categories (a) f non-increasing and $\phi > 0$ or (b) f non-decreasing and $\phi < 0$.

REMARK 7. *In the particular case of the α-divergence, for which $f(u) = \frac{1}{\alpha(\alpha-1)}[u^\alpha - 1]$, Condition (A4)-(iv) can be rewritten as*

$$\int_{\mathsf{Y}} \sup_{\theta \in \mathsf{T}} q(\theta, y) \times \sup_{\theta' \in \mathsf{T}} \left[\left(\frac{q(\theta', y)}{p(y)} \right)^{\alpha - 1} \right] \nu(\mathrm{d}y) < \infty \, ,$$

which in turns, implies that Condition (A4)-(iii) is satisfied.

3.4. *Approximate f-EI(ϕ)* . As Algorithm 1 typically involves an intractable integral in the Expectation step, we now turn to a practical version of this algorithm.

Algorithm 2: *Approximate f-EI(ϕ) one-step transition*

1. Sampling step : Draw independently $Y_1, ..., Y_K \sim \mu q$
2. Expectation step : $b_{\mu,K}(\theta) = \frac{1}{K} \sum_{k=1}^{K} \frac{q(\theta, Y_k)}{\mu q(Y_k)} f' \left(\frac{\mu q(Y_k)}{p(Y_k)} \right)$
3. Iteration step : $\mathcal{I}_K^\phi(\mu)(\mathrm{d}\theta) = \frac{\mu(\mathrm{d}\theta) \cdot |b_{\mu,K}(\theta)|^\phi}{\mu(|b_{\mu,K}|^\phi)}$

Algorithm 2 uses μq as a sampler instead of $q(\theta, \cdot)$. Indeed, as our algorithm optimises over μ, sampling with respect to μq gives preference to the interesting regions of the parameter space. Furthermore, picking a sampler that is independent from θ is less costly from a computational point of view.

In the rest of this section, we consider i.i.d random variables $Y_1, Y_2, ...$ with common density μq w.r.t ν, defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and we denote by $\mathbb{E}$ the associated expectation operator.

PROPOSITION 8. *Assume (A1) and (A2). Let $\mu \in \mathrm{M}_1(\mathsf{T})$, $\phi \in \mathbb{R}^\star$ be such that $\mu(|b_\mu|) \vee \mu(|b_\mu|^\phi) < \infty$ and*

$$\int_{\mathsf{T}} \mu(\mathrm{d}\theta) \mathbb{E} \left[\left\{ \frac{q(\theta, Y_1)}{\mu q(Y_1)} \left| f' \left(\frac{\mu q(Y_1)}{p(Y_1)} \right) \right| \right\}^\phi \right] < \infty \, . \tag{13}$$

Then,

$$\lim_{K\to\infty} \left\| \mathcal{I}_K^\phi(\mu) - \mathcal{I}^\phi(\mu) \right\|_{TV} = 0\,, \quad \mathbb{P}-\text{a.s.}$$

Proof. By the triangular inequality, for all $K \in \mathbb{N}^\star$, for all $\theta \in \mathsf{T}$,

$$\left| \frac{|b_{\mu,K}(\theta)|^\phi}{\mu(|b_{\mu,K}|^\phi)} - \frac{|b_\mu(\theta)|^\phi}{\mu(|b_\mu|^\phi)} \right| \leqslant \frac{|b_{\mu,K}(\theta)|^\phi}{\mu(|b_{\mu,K}|^\phi)} \left| 1 - \frac{\mu(|b_{\mu,K}|^\phi)}{\mu(|b_\mu|^\phi)} \right| + \frac{||b_{\mu,K}(\theta)|^\phi - |b_\mu(\theta)|^\phi|}{\mu(|b_\mu|^\phi)}$$

Thus,

$$\begin{aligned} \left\| \mathcal{I}_K^\phi(\mu) - \mathcal{I}^\phi(\mu) \right\|_{TV} &= \mu\left(\left| \frac{|b_{\mu,K}|^\phi}{\mu(|b_{\mu,K}|^\phi)} - \frac{|b_\mu|^\phi}{\mu(|b_\mu|^\phi)} \right| \right) \\ &\leqslant \left| 1 - \frac{\mu(|b_{\mu,K}|^\phi)}{\mu(|b_\mu|^\phi)} \right| + \frac{\mu(||b_{\mu,K}|^\phi - |b_\mu|^\phi|)}{\mu(|b_\mu|^\phi)} \end{aligned}$$

For the first term of the rhs, Lemma 21 yields

$$\lim_{K\to\infty} \left| 1 - \frac{\mu(|b_{\mu,K}|^\phi)}{\mu(|b_\mu|^\phi)} \right| = 0 \tag{14}$$

As for the second term of the rhs, first note that for all $K \in \mathbb{N}^\star$, for all $\theta \in \mathsf{T}$

$$0 \leqslant ||b_{\mu,K}(\theta)|^\phi - |b_\mu(\theta)|^\phi| \leqslant |b_{\mu,K}(\theta)|^\phi + |b_\mu(\theta)|^\phi\,, \tag{15}$$

and since $\mu(|b_\mu|^\phi) < \infty$ the LLN for μ-almost all $\theta \in \mathsf{T}$ yields

$$\lim_{K\to\infty} |b_{\mu,K}(\theta)|^\phi = |b_\mu(\theta)|^\phi\,. \tag{16}$$

Furthermore, since $\mu(|b_\mu|^\phi) < \infty$, Lemma 21 and (16) imply

$$\lim_{K\to\infty} \mu\left[|b_{\mu,K}|^\phi + |b_\mu|^\phi \right] = \mu\left[\lim_{K\to\infty} \left(|b_{\mu,K}|^\phi + |b_\mu|^\phi \right) \right] < \infty$$

Combining with (15) and (16), we apply Lemma 22 and obtain

$$\lim_{K\to\infty} \frac{\mu(||b_{\mu,K}|^\phi - |b_\mu|^\phi|)}{\mu(|b_\mu|^\phi)} = 0$$

which, along with (14), finishes the proof. □

Note that since Algorithm 2 is to be repeated multiple times until the convergence is reached, $\Psi^{(f)}(\mu)$ needs to be approximated as well, as it measures the convergence. In particular, the expression of $\Psi^{(f)}(\mu)$ given

in (2) legitimates the choice of μq as the sampler, as the closest available approximation of p. Let us denote by $\Psi_K^{(f)}(\mu)$ the corresponding unbiased approximation of $\Psi^{(f)}(\mu)$

$$\Psi_K^{(f)}(\mu) = \frac{1}{K}\sum_{k=1}^{K} f\left(\frac{\mu q(Y_k)}{p(Y_k)}\right)\frac{p(Y_k)}{\mu q(Y_k)} \ .$$

LEMMA 9. *Assume (A1) and that (f,ϕ) belongs to any of the cases stated in Table 2. Let $\mu \in \mathrm{M}_1(\mathsf{T})$ be such that $\Psi^{(f)}(\mu) < \infty$. Then for all $K \in \mathbb{N}^\star$,*

$$\Psi_K^{(f)} \circ \mathcal{I}_K^{\phi}(\mu) \leqslant \Psi_K^{(f)}(\mu) \ .$$

Furthermore,

$$\lim_{K\to\infty} \Psi_K^{(f)}(\mu) = \Psi^{(f)}(\mu) \ , \quad \mathbb{P}-\text{a.s.}$$

PROOF. The first point is a straightforward adaptation of the proof of Theorem 2.

As for the second point, we know from Lemma 3 that $\mathbb{E}[|f(\frac{\mu q(Y_1)}{p(Y_1)})|\frac{p(Y_1)}{\mu q(Y_1)}]$ is finite if and only if $\Psi^{(f)}(\mu)$ is finite. We can thus apply the LLN which yields the desired result. □

In the next section, we investigate how our algorithm can be applied to density approximation for the divergences identified in Theorem 2.

4. f-EI(ϕ) applied to density approximation.

4.1. *Reformulation of the optimisation problem.* Let $\tilde{p}$ be a probability density function on $(\mathsf{Y},\mathcal{Y})$ and assume that we only have access to an unnormalized version p^* of the density $\tilde{p}$, that is for all $y \in \mathsf{Y}$,

$$\tilde{p}(y) = \frac{p^*(y)}{Z} \ , \tag{17}$$

where $Z := \int_{\mathsf{Y}} p^*(y)\nu(\mathrm{d}y)$ is called the *normalizing constant* or *partition function*.

Let us denote by $\tilde{\mathbb{P}}$ the probability measure on $(\mathsf{Y},\mathcal{Y})$ with density $\tilde{p}$ with respect to ν and let us recall that for all $\mu \in \mathrm{M}_1(\mathsf{T})$, μQ corresponds to the probability measure on $(\mathsf{Y},\mathcal{Y})$ with density μq with respect to ν. We then have the following lemma, whose proof can be found in Appendix C.

LEMMA 10. *Assume (A1). Then, for both the reverse Kullback-Leibler and the α-divergence, optimising the objective $D_f(\mu Q||\tilde{\mathbb{P}})$ (with respect to μ) is equivalent to optimising the objective $\Psi^{(f)}(\mu; p)$ with $p = p^*$.*

If we now rewrite Lemma 3 for $p = p^*$ in the particular case of the reverse Kullback-Leibler and of the α-divergence, we obtain, as $\tilde{f}(u) = u\log(u)$ and $\tilde{f}(u) = \frac{1}{\alpha(\alpha-1)}[u^{1-\alpha} - u]$ respectively,

$$\int_{\mathsf{Y}} -\log\left(\frac{\mu q(y)}{p^*(y)}\right) p^*(y)\nu(\mathrm{d}y) \geqslant Z\log(Z)\ , \tag{18}$$

$$\int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)}\left(\frac{\mu q(y)}{p^*(y)}\right)^{\alpha} p^*(y)\nu(\mathrm{d}y) \geqslant \frac{1}{\alpha(\alpha-1)} Z^{1-\alpha}\ . \tag{19}$$

Here, the normalizing constant Z only appears in the r.h.s of (18) and (19). An interesting aspect is that optimising $\Psi^{(f)}(\mu; p)$ with $p = p^*$ is equivalent to optimising the bound (18) for the reverse Kullback-Leibler and the bound (19) for the α-divergence, where the optimisation does not involve the (unknown) normalizing constant Z anymore.

As it turns out, (18) is of little help to provide an explicit bound on the normalizing constant Z, however (19) can be used to provide either an upper or lower bound on Z, depending on the sign of α. To see this, let $\alpha \in \mathbb{R} \setminus \{0, 1\}$. Then, for any measurable positive function $\tilde{q}$ on $(\mathsf{Y}, \mathcal{Y})$, set

$$\xi^{(\alpha)}(\tilde{q}) := \left[\int_{\mathsf{Y}} \left(\frac{\tilde{q}(y)}{p^*(y)}\right)^{\alpha} p^*(y)\nu(\mathrm{d}y)\right]^{\frac{1}{1-\alpha}}\ . \tag{20}$$

We call α-bound the function $\tilde{q} \mapsto \xi^{(\alpha)}(\tilde{q})$. The next lemma is a straightforward consequence of (19).

LEMMA 11. *Assume (A1). Let $\mu \in \mathrm{M}_1(\mathsf{T})$. Then, for all $\alpha_+ \in (0, 1) \cup (1, +\infty)$ and all $\alpha_- < 0$, we have*

$$\xi^{(\alpha^+)}(\mu q) \leqslant Z \leqslant \xi^{(\alpha^-)}(\mu q)\ .$$

Of course, the two previous lemmas go beyond the case $\tilde{q} = \mu q$. Notably, Lemma 11 holds for any probability density function $\tilde{q}$ on $(\mathsf{Y}, \mathcal{Y})$ such that $\mathrm{supp}(p^*(y)) \subseteq \mathrm{supp}(\tilde{q}(y))$, as previously established in [17, Theorem 2]. Furthermore, $\alpha \mapsto \xi^{(\alpha)}(\tilde{q})$ is continuous on $\left\{\alpha : \xi^{(\alpha)}(\tilde{q}) < +\infty\right\}$, which is a straightforward consequence of [20, Theorem 1] using Renyi's α-divergence.

REMARK 12. *The Bayesian framework described in Section 2 corresponds to the case $p = p(\cdot|\mathscr{D})$, $\tilde{p} = p(\mathscr{D}, \cdot)$ and $Z = p(\mathscr{D})$. Note that in*

this particular configuration, Lemma 11 allows to upper and lower bound the marginal likelihood [20, Theorem 1].

In conclusion, we acquire additional information by using the α-divergence instead of the reverse Kullback-Leibler, since the α-bound can be used as a bound for the normalizing constant Z. Then, under the assumptions of Theorem 2, we can apply Algorithm 1 iteratively and benefit from the fact that (i) the α-bound acts as a measure of convergence and (ii) the monotonicity property can be observed through $(\xi^{(\alpha)}(\mu_n q))_{n\in\mathbb{N}}$, as for all $\mu \in \mathrm{M}_1(\mathsf{T})$,

$$\Psi^{(f)}(\mu; p) = \left[\xi^{(\alpha)}(\mu q)\right]^{1-\alpha} - \frac{Z}{\alpha(\alpha-1)} \quad \text{with} \quad p = p^* \,.$$

We now focus on rewriting Algorithm 2, which is the algorithm that we use in practice, in the special case of the α-divergence and with $p = p^*$.

4.2. *Approximate f-EI(ϕ) for the α-divergence.* From here on, we only consider the particular case of the α-divergence with $p = p^*$. Let (α, ϕ) be as in Table 2 and let $\mu \in \mathrm{M}_1(\mathsf{T})$. Using that $f'(u) = \frac{1}{\alpha-1}u^{\alpha-1}$, we obtain Algorithm 3.

Algorithm 3: *α-Approximate f-EI(ϕ) one-step transition*

1. Sampling step : Draw independently $Y_1, ..., Y_K \sim \mu q$
2. Expectation step : $b_{\mu,K}(\theta) = \dfrac{1}{K(1-\alpha)} \sum_{k=1}^{K} q(\theta, Y_k)\mu q(Y_k)^{\alpha-2} p^*(Y_k)^{1-\alpha}$
3. Iteration step : $\mathcal{I}_K^{\phi}(\mu)(\mathrm{d}\theta) = \dfrac{\mu(\mathrm{d}\theta)\cdot|b_{\mu,K}(\theta)|^{\phi}}{\mu(|b_{\mu,K}|^{\phi})}$

Now recall that $\xi^{(\alpha)}(\mu q)$ acts as a surrogate to $\Psi^{(f)}(\mu; p)$ which does not involve the normalizing constant Z anymore. If we are to apply Algorithm 3 repeatedly, we thus need to approximate this quantity in order to assess the convergence. Consequently, we define the Monte Carlo approximation $\xi_K^{(\alpha)}(\mu q)$ of $\xi^{(\alpha)}(\mu q)$ by

$$\xi_K^{(\alpha)}(\mu q) = \left[\frac{1}{K}\sum_{k=1}^{K}\left(\frac{p^*(Y_k)}{\mu q(Y_k)}\right)^{1-\alpha}\right]^{\frac{1}{1-\alpha}} \,.$$

where $Y_1, ..., Y_K$ are drawn independently from μq as in Algorithm 3. This estimator converges to $\xi^{(\alpha)}(\mu q)$ as $K \to \infty$ and while it is biased, the bias

can be characterized using [20, Theorem 2].

Finally, Algorithm 3 requires us to know how to sample from μq. We address the case where μ corresponds to a weighted sum of Dirac measures. This case is of particular interest to us since, as we shall see, it provides a gradient-based update formula for the mixture weights. To this end, let $J \in \mathbb{N}^\star$ and let us introduce the simplex of $\mathbb{R}^J$

$$\mathcal{S}_J = \left\{ \boldsymbol{\lambda} = (\lambda_1, ..., \lambda_J) \in \mathbb{R}^J \ : \ \forall j \in \{1, ..., J\} \, , \ \lambda_j \geqslant 0 \text{ and } \sum_{j=1}^{J} \lambda_j = 1 \right\} .$$

Let $\theta_1, ..., \theta_J \in \mathsf{T}$ be fixed. For all $\boldsymbol{\lambda} \in \mathcal{S}_J$, we define $\mu_{\boldsymbol{\lambda}} \in \mathrm{M}_1(\mathsf{T})$ by

$$\mu_{\boldsymbol{\lambda}} = \sum_{j=1}^{J} \lambda_j \delta_{\theta_j} \, .$$

Then, for all $\boldsymbol{\lambda} \in \mathcal{S}_J$, $\mu_{\boldsymbol{\lambda}} q(y) = \sum_{j=1}^{J} \lambda_j q(\theta_j, y)$ corresponds to a mixture model. Let $(\mu_n)_{n \in \mathbb{N}}$ be defined by

$$\begin{cases} \mu_0 = \mu_{\boldsymbol{\lambda}}, \\ \mu_{n+1} = \mathcal{I}_K^\phi(\mu_n) \, , \qquad n \in \mathbb{N} \, . \end{cases}$$

An immediate induction yields that for every $n \in \mathbb{N}$, μ_n can be expressed as $\mu_n = \sum_{j=1}^{J} \lambda_{j,n} \delta_{\theta_j}$ where $\boldsymbol{\lambda}_n = (\lambda_{1,n}, \ldots, \lambda_{J,n}) \in \mathcal{S}_J$ satisfies the initialisation $\boldsymbol{\lambda}_0 = \boldsymbol{\lambda}$ and the update formula: for all $n \in \mathbb{N}$ and all $j \in \{1, \ldots, J\}$,

$$\lambda_{j,n+1} = \lambda_{j,n} \frac{|b_{\mu_n,K}(\theta_j)|^\phi}{\mu_n(|b_{\mu_n,K}|^\phi)} \, , \tag{21}$$

with $Y_{1,n}, ..., Y_{K,n}$ drawn independently from $\mu_n q$ and

$$b_{\mu_n,K}(\theta_j) = \frac{1}{K(1-\alpha)} \sum_{k=1}^{K} q(\theta_j, Y_{k,n}) \mu_n q(Y_{k,n})^{\alpha-2} p^*(Y_{k,n})^{1-\alpha} \, .$$

We are thus able to derive Algorithm 4 below, where (21) is iterated until the convergence is reached and the α-bound is used at each step as a tractable measure of convergence and as a bound on the normalizing constant Z.

Algorithm 4: Mixture α-Approximate f-EI(ϕ)

Input: p^*: unnormalized version of the density $\tilde{p}$, Q: Markov transition kernel, K: number of samples, $\Theta_J = \{\theta_1, ..., \theta_J\} \subset \mathsf{T}$: parameter set.
Output: Optimised weights $\boldsymbol{\lambda}$.

Set $\boldsymbol{\lambda} = [\frac{1}{J}, ..., \frac{1}{J}]$.
while *the α-bound has not converged* **do**

Sampling step : Draw independently K samples $Y_1, ..., Y_K$ from $\mu_{\boldsymbol{\lambda}} q$.

Expectation step : Compute $\boldsymbol{A}_{\boldsymbol{\lambda}} = (a_j)_{1 \leqslant j \leqslant J}$ where

$$a_j = \frac{1}{K} \sum_{k=1}^{K} q(\theta_j, Y_k) \mu_{\boldsymbol{\lambda}} q(Y_k)^{\alpha - 2} p^*(Y_k)^{1-\alpha} \tag{22}$$

and deduce $\boldsymbol{B}_{\boldsymbol{\lambda}} = (\lambda_j a_j^{\phi})_{1 \leqslant j \leqslant J}$, $b_{\boldsymbol{\lambda}} = \sum_{j=1}^{J} \lambda_j a_j^{\phi}$ and $c_{\boldsymbol{\lambda}} = \sum_{j=1}^{J} \lambda_j a_j$.

Iteration step : Set

$$\xi_K^{(\alpha)}(\mu_{\boldsymbol{\lambda}} q) \leftarrow c_{\boldsymbol{\lambda}}^{1/(1-\alpha)}$$
$$\boldsymbol{\lambda} \leftarrow \frac{1}{b_{\boldsymbol{\lambda}}} \boldsymbol{B}_{\boldsymbol{\lambda}}$$

end

In this particular framework, most of the computing effort at each step lies within the computation of the vector $(b_{\mu_n, K}(\theta_j))_{1 \leqslant j \leqslant J}$, or equivalently the vector $\boldsymbol{A}_{\boldsymbol{\lambda}} = (a_j)_{1 \leqslant j \leqslant J}$ where the a_j are defined as in (22). Interestingly, these computations are similar to the ones required in typical gradient-based variational methods involving the α-divergence or Renyi's α-divergence [19, 20, 21]. Indeed, the objective function in these methods has a straightforward connection with the function

$$\tilde{q} \mapsto \mathcal{L}_A^{(\alpha)}(\tilde{q}) := \int_{\mathsf{Y}} \frac{1}{\alpha(\alpha - 1)} \left(\frac{\tilde{q}(y)}{p^*(y)} \right)^{\alpha} p^*(y) \nu(\mathrm{d}y) \ ,$$

and in our case, the score gradient of $\mathcal{L}_A^{(\alpha)}(\mu_{\boldsymbol{\lambda}} q)$ is directly linked to the quantities that are approximated in the Mixture α-Approximate f-EI(ϕ) algorithm, since under the proper (differentiation) assumptions

$$\nabla_{\boldsymbol{\lambda}} \mathcal{L}_A^{(\alpha)}(\mu_{\boldsymbol{\lambda}} q) = (b_{\mu_{\boldsymbol{\lambda}}}(\theta_j))_{1 \leqslant j \leqslant J} \ ,$$

where for all $j \in \{1, ..., J\}$, $b_{\mu_{\boldsymbol{\lambda}}}(\theta_j) = \frac{1}{\alpha - 1} \int_{\mathsf{Y}} q(\theta_j, y) \left(\frac{\mu_{\boldsymbol{\lambda}} q(y)}{p^*(y)} \right)^{\alpha - 1} \nu(\mathrm{d}y)$.

REMARK 13. *Our (exact) mixture weights update rule states that for all*

$n \in \mathbb{N}$, for all $j \in \{1, \ldots, J\}$,

$$\lambda_{j,n+1} = \lambda_{j,n} \frac{|b_{\mu_{\boldsymbol{\lambda}_n}}(\theta_j)|^\phi}{\mu_{\boldsymbol{\lambda}_n}(|b_{\mu_{\boldsymbol{\lambda}_n}}|^\phi)} \ ,$$

In the particular case where $(\alpha, \phi) = (0, 1)$ or $(\alpha, \phi) = (-1, 1)$, notice that we recover respectively the weights update rules from the Population Monte Carlo algorithm for mixtures applied to reverse Kullback-Leibler minimisation [34] and to Variance minimisation [35].

We now move on to numerical experiments in the next section.

5. Numerical experiments. To illustrate Algorithm 4, we first consider an example where p^* corresponds to a mixture of two one-dimensional Gaussian densities multiplied by a positive constant Z, such that

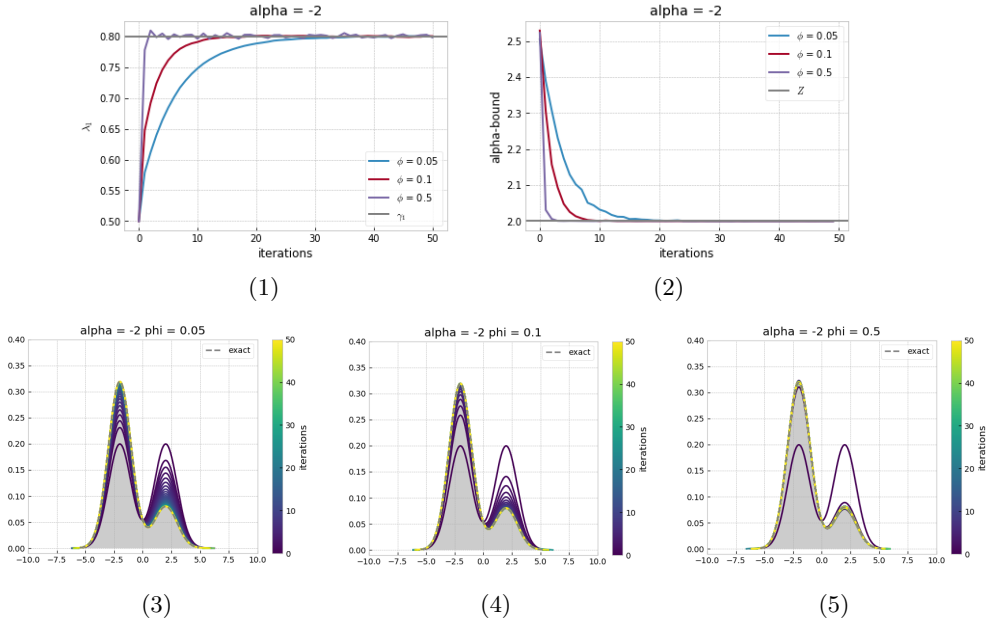
$$p^*(y) = Z \times [\gamma_1 \mathcal{N}(y; -s, 1) + \gamma_2 \mathcal{N}(y; s, 1)] \ ,$$

where $\gamma_1, \gamma_2 > 0$, $\gamma_1 + \gamma_2 = 1$ and $Z = 2$.

This simple framework allows us to visualize and characterize the behavior of the algorithm with respect to ϕ. Looking back at (21), we are prone to think that ϕ acts as a learning rate, as large values of $|\phi|$ amplify the impact of each $|b_{\mu,K}(\theta_j)|$ in the update process while $\phi = 0$ corresponds to no update at all. Let $\gamma_1 = 0.8$, $\gamma_1 = 0.2$, Q be a Gaussian kernel with variance 1, $J = 2$ with $\{\theta_1, \theta_2\} = \{-2, 2\}$ and $K = 5000$. We perform 50 iterations of the α-Approximate f-EI(ϕ) algorithm. The results for $\alpha = -2$ (with corresponding range $(0, 0.5]$ for ϕ) can be seen on Figure 1 and several more examples are available in Appendix D.

As expected, we obtain faster convergence rates as ϕ gets bigger in absolute value. However, if $|\phi|$ is too large, the algorithm may oscillate around the true value of the parameters due to the discretization, as exemplified in Figure 1-(1). Thus, the hyperparameter ϕ behaves like a learning rate. Secondly, we are able to observe the monotonicity property from Theorem 2 in Figure 1-(2), which plots the α-bound at each time step for various values of ϕ. Here, as we have picked α negative and as the conditions of support are met ($\text{supp}(p^*) = \text{supp}(\mu_{\boldsymbol{\lambda}} q)$), Z is upper-bounded by the α-bound. Furthermore, since the true value of the parameters belongs to the optimisation set, the bound is attained as the optimisation is carried out.

FIGURE 1. *Impact of the hyperparameter ϕ. Here $\alpha = -2$ with corresponding range $(0, 0.5]$ for ϕ and the grey dotted line corresponds to the exact density $\tilde{p}$.*



(1) (2)

(3) (4) (5)

Towards an adaptive algorithm. Algorithm 4 leaves $\{\theta_1, ..., \theta_J\}$ unchanged throughout the optimisation of the mixture weights (we call it an *Exploitation Step*). A natural idea is to combine this algorithm with an *Exploration step* that modifies the parameter set.

We offer to derive the new parameter set from the old one using the following simple update rule: assume that the mixture weights have been optimised using Algorithm 4. We first resample among $\{\theta_1, ..., \theta_J\}$ according to the optimised mixture weights. The obtained sample $\{\theta'_1, ..., \theta'_{J'}\}$ is then perturbed stochastically using a Gaussian transition kernel Q_r with covariance matrix rI_d $(r > 0)$ and density q_r, which gives us our new parameter set. The goal is then to iterate this Exploitation-Exploration procedure whilst using a decaying r (to ensure the convergence of the parameter set). As for the initial parameter set, it is generated randomly from an initial density q_0, where we have in mind that q_0 should allow to explore the space extensively.

The complete algorithm is summed up in Algorithm 5. Note that we kept K and J fixed for convenience, but that they could be set according to a policy $(K_i)_i$ and $(J_i)_i$ and similarly we could consider a sequence of Markov transition kernels $(Q_i)_i$.

Algorithm 5: Complete Exploitation-Exploration Algorithm

Input: p^*: unnormalized version of the density $\tilde{p}$, α: α-divergence parameter, ϕ: tuning parameter, q_0: initial sampler, Q, Q_r: Markov transition kernels, K: number of samples, J: dimension of parameter set, $(r_i)_i$: rate policy
Output: Optimised weights $\boldsymbol{\lambda}$ and parameter set Θ_J.
Draw $\theta_1, ..., \theta_J$ from q_0. Set $i = 0$.
while *the α-bound has not converged* **do**
  Exploitation step : Set $\Theta_J = \{\theta_1, ..., \theta_J\}$. Perform Mixture α-Approximate f-EI(ϕ) and obtain $\boldsymbol{\lambda}$ and $\xi_K^{(\alpha)}(\mu_{\boldsymbol{\lambda}} q)$.
  Exploration step : Draw independently J samples $\theta_1, ..., \theta_J$ from $\mu_{\boldsymbol{\lambda}} q_{r_i}$ and set $i = i + 1$.
end

We now want to assess how Algorithm 5 performs in a higher dimensional setting. Notably, we aim at observing the impact of the dimension d as well as the parameter set dimension J.

Toy example revisited. The target p^* now corresponds to a mixture of two d-dimensional Gaussian densities multiplied by a positive constant Z such that

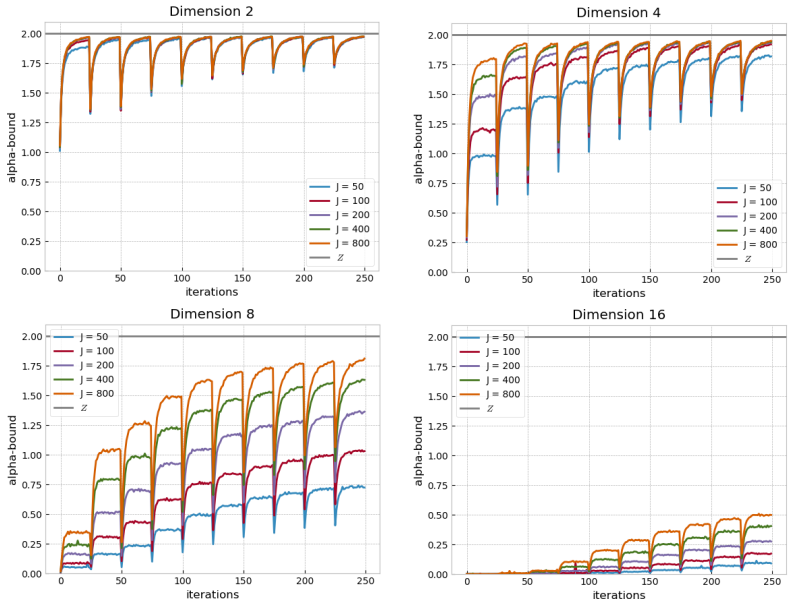
$$p^*(y) = Z \times [0.5\mathcal{N}(\boldsymbol{y}; -s\boldsymbol{u_d}, \boldsymbol{I_d}) + 0.5\mathcal{N}(\boldsymbol{y}; s\boldsymbol{u_d}, \boldsymbol{I_d})]$$

where $\boldsymbol{u_d}$ is the d-dimensional vector whose coordinates are all equal to 1, $\boldsymbol{I_d}$ is the identity matrix, $s = 2$ and $Z = 2$.

The first aspect that we need to consider is the choice of the divergence measure. The hyperparameter α allows to choose between *mass-covering* divergences which tends to cover all the modes ($\alpha \ll 0$) and *mode-seeking* divergences that are attracted to the mode with the largest probability mass ($\alpha \gg 1$), the case $\alpha \in (0, 1)$ corresponding to a mix of the two worlds. Depending on the learning task, the optimal α may differ and understanding how to select the value of α is still an area of ongoing research. In our case, since the targeted density is multimodal, we prefer having $\alpha < 1$ and we set $(\alpha, \phi) = (0.5, 1)$ in our experiments.

We set the initial sampler to be a centered normal distribution with covariance matrix $5\boldsymbol{I_d}$, Q to be a Gaussian kernel with variance $\boldsymbol{I_d}$ and $(r_i)_i$ which satisfies $r_0 = 2.5$ and $r_i = r_0/\sqrt{i+1}$ for all $i \in \mathbb{N}$. Let $K = 500$, $J \in \{50, 100, 200, 400, 800\}$, $d = \{2, 4, 8, 16\}$. We run 10 iterations of Algorithm 5, with 25 inner iterations each time the Mixture α-Approximate f-EI(ϕ) algorithm is called. We replicate the experiment 100 times and calculate the average α-bound computed over the 100 replicates. The results

FIGURE 2. *Impact of the dimension d and of the parameter set dimension J. Plotted is the α-bound computed for each pair of (J, d) over 100 replicates.*



can be seen on Figure 2.

Observe that the jumps in the α-bound correspond to an update of the parameter set. As expected, the optimisation becomes harder as the dimension grows. Yet, we are still able to climb to the normalizing constant Z up to the dimension 8 with $J = 800$, even though the Exploration step has not been optimised.

6. Conclusion and perspectives. We introduced the f-EI(ϕ) algorithm, an iterative algorithm which operates on measures. We proved that for a rich family of values of (f, ϕ) this algorithm leads at each step to a systematic decrease in the f-divergence and obtained its convergence to an optimum. In the particular case of the α-divergence with μ set as a weighted sum of Dirac measures, we obtained that the mixture weights update rule mostly relied on gradient-based calculations. Empirical results confirmed that the hyperparameter ϕ acted as a learning rate for our algorithm. They also demonstrated that the f-EI(ϕ) algorithm serves as a powerful Exploitation step, which shall be combined with an appropriate Exploration step to form a fully adaptive algorithm.

To conclude, we state several directions to extend our work on both a theoretical and a practical level.

Learning rate. We maintained ϕ constant in the f-EI(ϕ) algorithm. Exploring variants of the algorithm with different decaying learning rate policies $(\phi_n)_n$ and investigating convergence rates might result in more accurate and more stable results in practice.

Large scale learning. By noticing that the f-EI(ϕ) algorithm falls into the category of gradient-based algorithms when μ is chosen as a weighted sum of Diracs, we paved the way for large scale learning by using stochastic optimisation techniques, as deployed in [19] or more recently in [20, 21].

Exploration Step. The f-EI(ϕ) algorithm allows us to extend the parameter set and to work with a population of particles $\{\theta_1, ..., \theta_J\}$ instead of just one particle θ. In this regard, many more evolved methods could be envisioned as an Exploration step and combined with the f-EI(ϕ) algorithm.

Monte Carlo Approximation. One may want to resort to more advanced Monte Carlo methods in the estimation of b_{μ_n} at each step. For example, we did not reuse any of the past samples so far in our calculations. Since we kept the allocation policy $(K_n)_n$ constant equal to K in this paper, another interesting aspect would be to investigate how different allocation policies affect the performances of the algorithm.

References.

[1] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, Nov 1999.

[2] Matthew James. Beal. Variational algorithms for approximate bayesian inference /. *PhD thesis*, 01 2003.

[3] Manfred Opper and Ole Winther. Gaussian processes for classification: Mean-field algorithms. *Neural Comput.*, 12(11):2655–2684, November 2000.

[4] Thomas P. Minka. Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, pages 362–369, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[5] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 03 1951.

[6] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational Inference: A Review for Statisticians. *arXiv e-prints*, page arXiv:1601.00670, Jan 2016.

[7] Cheng Zhang, Judith Butepage, Hedvig Kjellstrom, and Stephan Mandt. Advances in Variational Inference. *arXiv e-prints*, page arXiv:1711.05597, Nov 2017.

[8] Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1-2):1–305, January 2008.

[9] John Paisley, David Blei, and Michael Jordan. Variational Bayesian Inference with Stochastic Search. *arXiv e-prints*, page arXiv:1206.6430, Jun 2012.

[10] Rajesh Ranganath, Sean Gerrish, and David M. Blei. Black Box Variational Inference. *arXiv e-prints*, page arXiv:1401.0118, Dec 2013.

[11] Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347, 2013.
[12] Thomas P. Minka. *A Family of Algorithms for Approximate Bayesian Inference*. PhD thesis, Cambridge, MA, USA, 2001. AAI0803033.
[13] Huaiyu Zhu and Richard Rohwer. Bayesian invariant measurements of generalization. *Neural Processing Letters*, 2:28–31, 12 1995.
[14] Huaiyu Zhu and Richard Rohwer. Information geometric measurements of generalisation. 10 1995.
[15] Alfréd Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 547–561, Berkeley, Calif., 1961. University of California Press.
[16] Tim van Erven and Peter Harremoës. Rényi Divergence and Kullback-Leibler Divergence. *ArXiv e-prints*, June 2012.
[17] Tom Minka. Divergence measures and message passing. page 17, January 2005.
[18] Tom Minka. Power ep. page 6, January 2004.
[19] José Miguel Hernández-Lobato, Yingzhen Li, Mark Rowland, Daniel Hernández-Lobato, Thang Bui, and Richard E. Turner. Black-box α-divergence Minimization. *arXiv e-prints*, page arXiv:1511.03243, Nov 2015.
[20] Yingzhen Li and Richard E. Turner. Rényi Divergence Variational Inference. *arXiv e-prints*, page arXiv:1602.02311, Feb 2016.
[21] Adji Bousso Dieng, Dustin Tran, Rajesh Ranganath, John Paisley, and David Blei. Variational inference via χ upper bound minimization. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 2732–2741. Curran Associates, Inc., 2017.
[22] Yingzhen Li, Jose Miguel Hernandez-Lobato, and Richard Turner. Stochastic expectation propagation. 06 2015.
[23] Guillaume Dehaene and Simon Barthelmé. Expectation Propagation in the large-data limit. *arXiv e-prints*, page arXiv:1503.08060, Mar 2015.
[24] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In Yves Lechevallier and Gilbert Saporta, editors, *Proceedings of COMPSTAT'2010*, pages 177–186, Heidelberg, 2010. Physica-Verlag HD.
[25] Herbert Robbins and Sutton Monro. A stochastic approximation method. *Ann. Math. Statist.*, 22(3):400–407, 09 1951.
[26] David M. Blei, Andrew Y. Ng, and Michael Jordan. Latent dirichlet allocation. volume 3, pages 601–608, 01 2001.
[27] Tetsuzo Morimoto. Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von markoffschen ketten. *Magyar Tud. Akad. Mat. Kutat Int.*, page 85–108, 08 1963.
[28] Tetsuzo Morimoto. Markov processes and the h-theorem. *Journal of The Physical Society of Japan - J PHYS SOC JPN*, 18:328–331, 03 1963.
[29] Andrzej Cichocki and Shun-ichi Amari. Families of alpha- beta- and gamma- divergences: Flexible and robust measures of similarities. *Entropy*, 12, 06 2010.
[30] María del Carmen Pardo and Igor Vajda. On asymptotic properties of information-theoretic divergences. *IEEE Transactions on Information Theory*, 49(7):1860–1867, July 2003.
[31] Igal Sason. On f-divergences: Integral representations, local behavior, and inequalities. *CoRR*, abs/1804.06334, 2018.
[32] E. Hellinger. Neue begründung der theorie quadratischer formen von unendlichvielen

veränderlichen. *Journal für die reine und angewandte Mathematik*, 136:210–271, 1909.
[33] Bruce G. Lindsay. Efficiency versus robustness: The case for minimum hellinger distance and related methods. *Ann. Statist.*, 22(2):1081–1114, 06 1994.
[34] Randal Douc, Arnaud Guillin, Jean-Michel Marin, and Christian P. Robert. Convergence of adaptive mixtures of importance sampling schemes. *arXiv e-prints*, page arXiv:0708.0711, Aug 2007.
[35] Randal Douc, Arnaud Guillin, Jean-Michel Marin, and Christian P. Robert. Minimum variance importance sampling via population monte carlo. *ESAIM: Probability and Statistics*, 11:427–447, 2007.
[36] H.L. Royden and P. Fitzpatrick. *Real Analysis (4th Edtion)*. Prentice Hall, 2010.

APPENDIX A

A.1. Proof of Theorem 2. In Proposition 5, the difference $\Psi^{(f)}(\zeta) - \Psi^{(f)}(\mu)$ is split into two terms

$$\Psi^{(f)}(\zeta) - \Psi^{(f)}(\mu) = A(\mu, \zeta) + |\varrho|^{-1} \left\{ \mu\left(|b_\mu| g^\varrho\right) - \mu(|b_\mu|) \right\} \ ,$$

where $g = \mathrm{d}\zeta/\mathrm{d}\mu$. Moreover, Proposition 5 states that $A(\mu, \zeta)$ is always non-positive.

It turns out that the second term is minimal over all positive probability densities g when it is proportional to $|b_\mu|^{1/(1-\varrho)}$, as we show in Lemma 14 below.

LEMMA 14. *For any positive probability density g w.r.t μ, we have*

$$\mu\left(|b_\mu| g^\varrho\right) \geqslant \left[\mu\left(|b_\mu|^{1/(1-\varrho)}\right)\right]^{1-\varrho} \ ,$$

with equality if and only if $g \propto |b_\mu|^{1/(1-\varrho)}$.

PROOF. The function $x \mapsto x^{1-\varrho}$ is strictly convex for $\varrho \in \mathbb{R} \setminus [0, 1]$. Thus Jensen's inequality yields, for any positive probability density g w.r.t. μ,

$$\mu\left(|b_\mu| g^\varrho\right) = \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left(\frac{|b_\mu(\theta)|^{1/(1-\varrho)}}{g(\theta)} \right)^{1-\varrho} g(\theta) \geqslant \left[\mu\left(|b_\mu|^{1/(1-\varrho)}\right)\right]^{1-\varrho} \tag{23}$$

which finishes the proof of the inequality. The next statement follows from the case of equality in Jensen's inequality: g must be proportional to $|b_\mu|^{1/(1-\varrho)}$. □

The next lemma shows that this choice leads to a non-positive second term, thus implying that $\Psi^{(f)}(\zeta) \leqslant \Psi^{(f)}(\mu)$.

LEMMA 15. *Assume (A1), (A2) and (A3). Then $\phi = 1/(1-\varrho)$ satisfies (11) for any $\mu \in \mathrm{M}_1(\mathsf{T})$ such that $\mu(|b_\mu|) < \infty$.*

PROOF. We apply (23) with $g = 1$ and get that

$$\left[\mu\left(|b_\mu|^{1/(1-\varrho)}\right)\right]^{1-\varrho} \leqslant \mu(|b_\mu|) < \infty \ . \tag{24}$$

Then (11) can be readily checked with $\phi = 1/(1-\varrho)$. Furthermore using $\mu(|b_\mu|) < \infty$ when $\phi < 0$ and (A1) combined with (A2) for $\phi > 0$, we obtain $\mu(|b_\mu|^\phi) > 0$, which concludes the proof. □

While Lemma 15 seems to advocate for $g = \mathrm{d}\zeta/\mathrm{d}\mu$ to be proportional to $|b_\mu|^{1/(1-\varrho)}$, notice that this choice of g might not be optimal to minimize $\Psi^{(f)}(\zeta) - \Psi^{(f)}(\mu)$, as $A(\mu, \zeta)$ also depends on g through ζ. In the next lemma, we thus propose another choice of the tuning parameter ϕ, which also satisfies (11) for any $\mu \in \mathrm{M}_1(\mathsf{T})$ such that $\mu(|b_\mu|) < \infty$.

LEMMA 16. *Assume* (A1), (A2) *and* (A3). *Let* $\mu \in \mathrm{M}_1(\mathsf{T})$ *such that* $\mu(|b_\mu|) < \infty$. *Assume in addition that* $|\varrho| \geqslant 1$, *then the real number* $\phi = -1/\varrho$ *satisfies* (11).

PROOF. Setting $g \propto |b_\mu|^{-1/\varrho}$, we get

$$\mu(|b_\mu| g^\varrho) = \mu(|b_\mu|^{1-\varrho/\varrho})[\mu(|b_\mu|^{-1/\varrho})]^{-\varrho} = [\mu(|b_\mu|^{-1/\varrho})]^{-\varrho} \leqslant \mu(|b_\mu|)$$

where the last inequality follows from Jensen's inequality applied to the convex function $u \mapsto u^{-\varrho}$ (since $|\varrho| \geqslant 1$). Since $\mu(|b_\mu|) < \infty$, the parameter $\phi = -1/\varrho$ satisfies (11). Furthermore using $\mu(|b_\mu|) < \infty$ when $\phi < 0$ and (A1) combined with (A2) for $\phi > 0$, we obtain $\mu(|b_\mu|^\phi) > 0$, which concludes the proof. □

Lemma 15 and Lemma 16 allow us to define a range of values for ϕ that decreases $\Psi^{(f)}(\mu_n)$ at each iteration step. Now, in order to prove Theorem 2, we need to check that the reverse Kullback-Leibler and the α-divergence are f-divergences with a function f that satisfies (A2) and (A3).

PROOF OF THEOREM 2. The proof consists in verifying that we can apply Theorem 1, that is, for the two considered functions f, we must check (A2) and find a range of constants ϱ which satisfy (A3). We then use Lemma 15 or Lemma 16 to deduce that, for the provided constants ϕ, (11) holds for all μ_n with $n \geqslant 0$.

(i) Assumption (A2) readily holds and so does (A3) for all $\varrho < 0$, with $f_\varrho(u) = -\log(u)/\varrho$. Moreover, by definition of b_{μ_n}, we get for all $n \in \mathbb{N}$,

$$\mu_n(|b_{\mu_n}|) = \int_{\mathsf{Y}} \mu_n q(y) \frac{p(y)}{\mu_n q(y)} \nu(\mathrm{d}y) = \int_{\mathsf{Y}} p(y) \nu(\mathrm{d}y) < \infty \,.$$

Combining with Lemma 15 and Lemma 16, (11) holds for all μ_n with $n \geqslant 0$ and for any $\phi \in (0, 1]$.

(ii) Again (A2) can be readily checked. Observing that for $\alpha \notin \{0, 1\}$,

$$f_\varrho(u) = \frac{1}{\alpha(\alpha - 1)} \left(u^{\alpha/\varrho} - 1 \right) \,,$$

we get that (A3) holds for

$$\begin{cases} \varrho \leqslant \alpha & \text{if } \alpha < 0, \\ \varrho < 0 & \text{if } \alpha \in (0,1), \\ \varrho \geqslant \alpha & \text{if } \alpha > 1. \end{cases}$$

Lemmas 15 and 16 provide the corresponding ranges for ϕ in Cases (a), (b) and (c). To finish the proof, we now show by induction that for all $n \in \mathbb{N}$, $\Psi^{(f)}(\mu_n)$ and $\mu_n(|b_{\mu_n}|)$ are finite, so that Lemmas 15 and 16 can indeed be applied.
Since $uf'(u) = \alpha f(u) + 1/(\alpha - 1)$, we have, for all $n \in \mathbb{N}$,

$$\begin{aligned} \mu_n(|b_{\mu_n}|) &= \int_{\mathsf{Y}} \left| \left(\frac{\mu_n q(y)}{p(y)} \right) f' \left(\frac{\mu_n q(y)}{p(y)} \right) \right| p(y)\nu(\mathrm{d}y) \\ &\leqslant |\alpha| \int_{\mathsf{Y}} \left| f \left(\frac{\mu_n q(y)}{p(y)} \right) \right| p(y)\nu(\mathrm{d}y) + \frac{1}{|\alpha - 1|} \end{aligned} \tag{25}$$

Using Lemma 3, the rhs is finite if and only if $\Psi^{(f)}(\mu_n)$ is finite. We can now check that $\Psi^{(f)}(\mu_n)$ and $\mu_n(|b_{\mu_n}|)$ are finite by induction on n:

- Start with $n = 0$. Then $\mu_0 = \mu$ and by assumption, $\Psi^{(f)}(\mu) < \infty$ and we deduce $\mu(|b_\mu|) < \infty$ by (25).
- Now, under the induction assumption and using Lemma 15 and Lemma 16, Theorem 1 can be applied and $\Psi^{(f)}(\mu_{n+1}) \leqslant \Psi^{(f)}(\mu_n) < \infty$. Using again (25) with n replaced by $n+1$, we get $\mu_{n+1}(|b_{\mu_{n+1}}|) < \infty$.

□

A.2. Proof of Theorem 3. In the following, we use the notation $\mu_n \Rightarrow \bar{\mu}$ for the weak convergence of measures in $\mathrm{M}_1(\mathsf{T})$. We first derive four useful lemmas.

LEMMA 17. *Assume (A1), (A2) and (A4). Suppose that $\mu_n \Rightarrow \bar{\mu}$. Then the following assertions hold.*

(i) For all $y \in \mathsf{Y}$, $\mu_n q(y)$ tends to $\bar{\mu} q(y)$ as $n \to \infty$.
(ii) For all $\zeta \in \mathrm{M}_1(\mathsf{T})$, the function $\theta \mapsto |b_\zeta(\theta)|$ is continuous. Furthermore for all $\theta \in \mathsf{T}$, $|b_{\mu_n}(\theta)|$ tends to $|b_{\bar{\mu}}(\theta)|$ as $n \to \infty$.
(iii) There exist $0 < m_- < m_+ < \infty$ such that, for all $\zeta \in \mathrm{M}_1(\mathsf{T})$ and $\theta \in \mathsf{T}$, $|b_\zeta(\theta)| \in [m_-, m_+]$.
(iv) For all continuous, positive and bounded function h,

$$\lim_{n\to\infty} \int_{\mathsf{T}} \mu_n(\mathrm{d}\theta) |b_{\mu_n}(\theta)|^\phi h(\theta) = \int_{\mathsf{T}} \bar{\mu}(\mathrm{d}\theta) |b_{\bar{\mu}}(\theta)|^\phi h(\theta) \, .$$

PROOF. We prove the assertions successively.

Proof of (i). For all $y \in \mathsf{Y}$, the function $\theta \mapsto q(\theta, y)$ is continuous on a compact set, hence bounded. The weak convergence $\mu_n \Rightarrow \bar{\mu}$ thus implies the pointwise convergence of $\mu_n q$ to $\bar{\mu} q$.

Proof of (ii). For all $\theta \in \mathsf{T}$ and $\zeta \in \mathrm{M}_1(\mathsf{T})$, we write

$$|b_\zeta(\theta)| = \int_{\mathsf{Y}} a_\zeta(\theta, y) \nu(\mathrm{d}y) \ ,$$

where we set for all $(\theta, y) \in \mathsf{T} \times \mathsf{Y}$, $a_\zeta(\theta, y) = q(\theta, y) \left| f' \left(\frac{\zeta q(y)}{p(y)} \right) \right|$ (the absolute value can be put inside the integral since f' is of constant sign by (A2)). The continuity of $|b_\zeta|$ follows from the Dominated Convergence Theorem, since for all $y \in \mathsf{Y}$, the function $\theta \mapsto a_\zeta(\theta, y)$ is continuous on T by (A4)-(ii) and for all $(\theta, y) \in \mathsf{T} \times \mathsf{Y}$, we have

$$|a_\zeta(\theta, y)| \leqslant \sup_{\theta' \in \mathsf{T}} q(\theta', y) \times \left(\sup_{\theta'' \in \mathsf{T}} \left| f' \left(\frac{q(\theta'', y)}{p(y)} \right) \right| \right) \ , \tag{26}$$

which is integrable w.r.t $\nu(\mathrm{d}y)$ by (A4)-(iv). The second part of (ii) is obtained similarly. Using (i) and that f is C^1 by (A2), we get that, for all $(\theta, y) \in \mathsf{T} \times \mathsf{Y}$,

$$\lim_{n \to \infty} q(\theta, y) \left| f' \left(\frac{\mu_n q(y)}{p(y)} \right) \right| = q(\theta, y) \left| f' \left(\frac{\bar{\mu} q(y)}{p(y)} \right) \right| \ ,$$

i.e $\lim_{n \to \infty} a_{\mu_n}(\theta, y) = a_{\bar{\mu}}(\theta, y)$. The bound (26) and (A4)-(iv) provide a domination criterion and we get that $|b_{\mu_n}(\theta)|$ tends to $|b_{\bar{\mu}}(\theta)|$ as $n \to \infty$, which concludes the proof of (ii).

Proof of (iii). For all $(\theta, \zeta) \in \mathsf{T} \times \mathrm{M}_1(\mathsf{T})$, we have $|b_\zeta(\theta)| \in [m_-, m_+]$ where

$$\begin{aligned} m_- &:= \int_{\mathsf{Y}} \inf_{\theta' \in \mathsf{T}} q(\theta', y) \times \left(\inf_{\theta'' \in \mathsf{T}} \left| f' \left(\frac{q(\theta'', y)}{p(y)} \right) \right| \right) \nu(\mathrm{d}y) \ , \\ m_+ &:= \int_{\mathsf{Y}} \sup_{\theta' \in \mathsf{T}} q(\theta', y) \times \left(\sup_{\theta'' \in \mathsf{T}} \left| f' \left(\frac{q(\theta'', y)}{p(y)} \right) \right| \right) \nu(\mathrm{d}y) \ . \end{aligned} \tag{27}$$

We have that m_+ is finite by (A4)-(iv). Now recall that under (A2), f' does not vanish on $(0, \infty)$. Together with (A1), we thus have that for any $y \in \mathsf{Y}$, the functions $\theta \mapsto q(\theta, y)$ and $\theta \mapsto |f'(q(\theta, y)/p(y))|$ are continuous and positive on the compact set T, from which we deduce that $m_- > 0$.

Proof of (iv). Using (ii) the function $\theta \mapsto |b_{\bar{\mu}}(\theta)|^\phi h(\theta)$ is continuous, and, since T is compact, $\mu_n \Rightarrow \bar{\mu}$ gives that

$$\lim_{n \to \infty} \int_{\mathsf{T}} \mu_n(\mathrm{d}\theta) |b_{\bar{\mu}}(\theta)|^\phi h(\theta) = \int_{\mathsf{T}} \bar{\mu}(\mathrm{d}\theta) |b_{\bar{\mu}}(\theta)|^\phi h(\theta) \ . \tag{28}$$

Next we show that

$$\lim_{n\to\infty} \int_{\mathsf{T}} \mu_n(\mathrm{d}\theta) \left| |b_{\mu_n}(\theta)|^\phi - |b_{\bar\mu}(\theta)|^\phi \right| h(\theta) = 0 \tag{29}$$

Using (iii), since $u \mapsto u^\phi$ is Lipschitz on $[m_-, m_+]$, there exists a constant C such that

$$\begin{aligned}\mu_n \left[\left| |b_{\mu_n}|^\phi - |b_{\bar\mu}|^\phi \right| h \right] &\leqslant C \sup_{\theta\in\mathsf{T}} h(\theta) \int_{\mathsf{T}} \mu_n(\mathrm{d}\theta) \, ||b_{\mu_n}(\theta)| - |b_{\bar\mu}(\theta)|| \\ &= C \sup_{\theta\in\mathsf{T}} h(\theta) \int_{\mathsf{Y}} |a_n(y)| \nu(\mathrm{d}y)\end{aligned}$$

where $a_n(y) := \mu_n q(y) \left\{ \left| f'\left(\frac{\mu_n q(y)}{p(y)}\right)\right| - \left| f'\left(\frac{\bar\mu q(y)}{p(y)}\right)\right| \right\}$. Now, for all $y \in \mathsf{Y}$,

$$|a_n(y)| \leqslant 2 \sup_{\theta\in\mathsf{T}} q(\theta, y) \times \left(\sup_{\theta'\in\mathsf{T}} \left| f'\left(\frac{q(\theta', y)}{p(y)}\right)\right| \right)$$

which is integrable w.r.t ν by (A4)-(iv). Moreover, by (i) and by continuity of f', we have $\lim_{n\to\infty} a_n(y) = 0$, and (29) follows by dominated convergence. Finally, combining (28), (29) and

$$\mu_n \left[|b_{\mu_n}|^\phi h \right] = \mu_n \left[|b_{\mu_n}|^\phi h - |b_{\bar\mu}|^\phi h \right] + \mu_n \left[|b_{\bar\mu}|^\phi h \right] \ ,$$

we obtain (iv), and the proof is concluded. □

LEMMA 18. *Assume (A1) and (A2). Let $\zeta, \zeta' \in \mathrm{M}_1(\mathsf{T})$. Then,*

$$\int_{\mathsf{T}} [\zeta - \zeta'](\mathrm{d}\theta) b_{\zeta'}(\theta) \leqslant \Psi^{(f)}(\zeta) - \Psi^{(f)}(\zeta') \ . \tag{30}$$

Let $\bar\mu, \mu \in \mathrm{M}_1(\mathsf{T})$ and assume that there exists $\mu^\star \in \mathrm{M}_{1,\mu}(\mathsf{T})$ such that $\Psi^{(f)}(\mu^\star) < \Psi^{(f)}(\bar\mu)$. Then, for f non-increasing (resp. non-decreasing), there exists $\delta > 1$ (resp. $\delta < 1$) and such that

$$\mu^\star(b_{\bar\mu} < \delta \bar\mu(b_{\bar\mu})) > 0 \ . \tag{31}$$

PROOF. By definition of $b_{\zeta'}$,

$$\begin{aligned}\int_{\mathsf{T}} (\zeta - \zeta')(\mathrm{d}\theta) b_{\zeta'}(\theta) &= \int_{\mathsf{T}} (\zeta - \zeta')(\mathrm{d}\theta) \int_{\mathsf{Y}} q(\theta, y) f'\left(\frac{\zeta' q(y)}{p(y)}\right) \nu(\mathrm{d}y) \\ &= \int_{\mathsf{Y}} \frac{\zeta q(y) - \zeta' q(y)}{p(y)} f'\left(\frac{\zeta' q(y)}{p(y)}\right) p(y) \nu(\mathrm{d}y) \ .\end{aligned}$$

Now set $u_y = \dfrac{\zeta q(y)}{p(y)}$ and $v_y = \dfrac{\zeta' q(y)}{p(y)}$. Since f is convex, $f'(v_y)(u_y - v_y) \leqslant f(u_y) - f(v_y)$ and we obtain

$$\begin{aligned}\int_{\mathsf{T}} (\zeta - \zeta')(\mathrm{d}\theta) b'_\zeta(\theta) &\leqslant \int_{\mathsf{Y}} \left[f\left(\frac{\zeta q(y)}{p(y)}\right) - f\left(\frac{\zeta' q(y)}{p(y)}\right) \right] p(y)\nu(\mathrm{d}y) \\ &= \Psi^{(f)}(\zeta) - \Psi^{(f)}(\zeta') \,,\end{aligned}$$

which proves (30).

We now prove (31) in the case where f is non-increasing. First note that for all $\delta > 1$, $(\delta - 1)\bar{\mu}(b_{\bar{\mu}}) \leqslant 0$. Let us define $A_\delta = \{b_{\bar{\mu}} < \delta \bar{\mu}(b_{\bar{\mu}})\}$ and show that $\mu^\star(A_\delta) > 0$ for some $\delta > 1$. To do so, we proceed by contradiction. Suppose that $\mu^\star(A_\delta) = 0$ for all $\delta > 1$, so that

$$\mu^\star[b_{\bar{\mu}} - \bar{\mu}(b_{\bar{\mu}})] = \mu^\star[(b_{\bar{\mu}} - \bar{\mu}(b_{\bar{\mu}}))\, \mathbf{1}_{A_\delta^c}] \geqslant (\delta - 1)\bar{\mu}(b_{\bar{\mu}}) \,.$$

Using (30), we get that, for all $\delta > 1$,

$$0 > \Psi^{(f)}(\mu^\star) - \Psi^{(f)}(\bar{\mu}) \geqslant \mu^\star[(b_{\bar{\mu}} - \bar{\mu}(b_{\bar{\mu}}))] \geqslant (\delta - 1)\bar{\mu}(b_{\bar{\mu}}) \,.$$

Letting $\delta \downarrow 1$, we obtain a contradiction, which finishes the proof. The alternative case where f is non-decreasing is obtained similarly by taking $\delta \uparrow 1$. □

LEMMA 19. *Assume (A1) and (A2). Let $\bar{\mu} \in \mathrm{M}_1(\mathsf{T})$ be a fixed point of $\mathcal{I}^\phi$, let $g_{\bar{\mu}} := |b_{\bar{\mu}}|^\phi$ and $\phi \in \mathbb{R}^\star$. Let $\mu \in \mathrm{M}_1(\mathsf{T})$ and assume that there exists $\mu^\star \in \mathrm{M}_{1,\mu}(\mathsf{T})$ such that $\Psi^{(f)}(\bar{\mu}) > \Psi^{(f)}(\mu^\star)$. Then, there exists $\delta > 1$ such that*

$$\mu^\star(g_{\bar{\mu}} > \delta \bar{\mu}(g_{\bar{\mu}})) > 0$$

in the cases (a) and (b) of Theorem 3.

PROOF. Note that (5) holds for any ϕ and ζ (in particular $\zeta = \bar{\mu}$) by Lemma 17-(iii). As $\bar{\mu}$ is a fixed point of $\mathcal{I}^\phi$, $g_{\bar{\mu}}$ is $\bar{\mu}$-almost all constant. Consequently, $\bar{\mu}(|b_{\bar{\mu}}|^\phi)^{1/\phi} = \bar{\mu}(|b_{\bar{\mu}}|)$. We separate the two cases f non-increasing and f non-decreasing:

(i) Let f be non-increasing and $\phi > 0$. Then, $|b_{\bar{\mu}}| = -b_{\bar{\mu}}$ and $u \mapsto u^{1/\phi}$ is increasing. For all $\delta > 1$, $\delta' := \delta^{1/\phi} > 1$ and

$$\begin{aligned}\mu^\star(g_{\bar{\mu}} > \delta \bar{\mu}(g_{\bar{\mu}})) &= \mu^\star(|b_{\bar{\mu}}| > \delta^{1/\phi}[\bar{\mu}(|b_{\bar{\mu}}|^\phi)]^{1/\phi}) \\ &= \mu^\star(|b_{\bar{\mu}}| > \delta' \bar{\mu}(|b_{\bar{\mu}}|)) \\ &= \mu^\star(b_{\bar{\mu}} < \delta' \bar{\mu}(b_{\bar{\mu}})) \,.\end{aligned}$$

(ii) Let f be non-decreasing and $\phi < 0$. Then, $|b_{\bar{\mu}}| = b_{\bar{\mu}}$ and $u \mapsto u^{1/\phi}$ is decreasing. For all $\delta > 1$, $\delta' := \delta^{1/\phi} < 1$ and

$$\begin{aligned}\mu^\star(g_{\bar{\mu}} > \delta\bar{\mu}(g_{\bar{\mu}})) &= \mu^\star(|b_{\bar{\mu}}| < \delta^{1/\phi}[\bar{\mu}(|b_{\bar{\mu}}|^\phi)]^{1/\phi}) \\ &= \mu^\star(|b_{\bar{\mu}}| < \delta'\bar{\mu}(|b_{\bar{\mu}}|)) \\ &= \mu^\star(b_{\bar{\mu}} < \delta'\bar{\mu}(b_{\bar{\mu}}))\ .\end{aligned}$$

We conclude by applying Lemma 18 in the two separated cases. □

LEMMA 20. *Assume (A1), (A2) and (A4). Let $\phi \in \mathbb{R}^\star$ and denote $g_\zeta := |b_\zeta|^\phi$ for any $\zeta \in \mathrm{M}_1(\mathsf{T})$. Let $\mu \in \mathrm{M}_1(\mathsf{T})$ and define the sequence $(\mu_n)_{n\in\mathbb{N}}$ according to (4). Suppose that $\mu_n \Rightarrow \bar{\mu}$ for some fixed point $\bar{\mu} \in \mathrm{M}_1(\mathsf{T})$ of $\mathcal{I}^\phi$. Further assume there exists $\mu^\star \in \mathrm{M}_{1,\mu}(\mathsf{T})$ such that $\Psi^{(f)}(\bar{\mu}) > \Psi^{(f)}(\mu^\star)$. Then, there exist $\delta > 1$ and $n \in \mathbb{N}^*$ such that*

$$\mu^\star\left(\bigcap_{m\geqslant n}\{g_{\mu_m} > \delta\mu_m(g_{\mu_m})\}\right) > 0\ ,$$

in the cases (a) and (b) of Theorem 3.

PROOF. First note that the sequence $(\mu_n)_{n\in\mathbb{N}}$ is well-defined for any $\phi \in \mathbb{R}^\star$ by Lemma 17-(iii), which implies $\mu_n(g_{\mu_n}) > 0$ for all n. We further have that

$$\begin{aligned}\lim_{n\to\infty}\mu^\star\left(\bigcap_{m\geqslant n}\{g_{\mu_m} > \delta\mu_m(g_{\mu_m})\}\right) &= \mu^\star\left(\bigcup_{n\geqslant 1}\bigcap_{m\geqslant n}\{g_{\mu_m} > \delta\mu_m(g_{\mu_m})\}\right) \\ &= \mu^\star\left(\left\{\theta\in\mathsf{T}\ :\ \liminf_{n\to\infty}\frac{g_{\mu_m}(\theta)}{\mu_n(g_{\mu_m})} > \delta\right\}\right)\ .\end{aligned}$$

Furthermore, applying (ii) and (iv) in Lemma 17, we have, for all $\theta \in \mathsf{T}$, $\lim_{n\to\infty} g_{\mu_n}(\theta) = g_{\bar{\mu}}(\theta)$ and $\lim_{n\to\infty}\mu_n(g_{\mu_m}) = \bar{\mu}(g_{\bar{\mu}})$. Hence, for all $\theta \in \mathsf{T}$,

$$\liminf_{n\to\infty}\frac{g_{\mu_m}(\theta)}{\mu_m(g_{\mu_m})} = \frac{g_{\bar{\mu}}(\theta)}{\bar{\mu}(g_{\bar{\mu}})}\ .$$

The proof is concluded by applying Lemma 19. □

PROOF OF THEOREM 3. Assume (A1), (A2) and (A4).

Lemma 17-(iii) is exactly the first result we want to obtain, that is: for all $\zeta \in \mathrm{M}_1(\mathsf{T})$, any $\phi \in \mathbb{R}^\star$ satisfies (5) for ζ. Furthermore, $|\Psi^{(f)}(\zeta)| < \infty$ by (A4)-(iii).

Assume that $(\mu_n)_{n\in\mathbb{N}}$ weakly converges to $\bar{\mu} \in \mathrm{M}_1(\mathsf{T})$. First note that Lemma 17-(iii) implies that for any $\phi \in \mathbb{R}^\star$ the sequence $(\mu_n)_{n\in\mathbb{N}}$ is well-defined and $\bar{\mu}$ satisfies (5).

We now prove Assertions (i) and (ii) successively.

Proof of (i). For all $\zeta \in \mathrm{M}_1(\mathsf{T})$ and all $y \in \mathsf{Y}$, set $a_\zeta(y) = f\left(\frac{\zeta q(y)}{p(y)}\right) p(y)$, leading to

$$\Psi^{(f)}(\zeta) = \int_{\mathsf{Y}} a_\zeta(y)\nu(\mathrm{d}y) \; . \tag{32}$$

Then, for all $y \in \mathsf{Y}$,

$$|a_\zeta(y)| \leqslant \left(\sup_{\theta\in\mathsf{T}} \left| f\left(\frac{q(\theta,y)}{p(y)}\right)\right|\right) p(y) \; , \tag{33}$$

which is integrable w.r.t $\nu(\mathrm{d}y)$ by (A4)-(iii).

Furthermore, recall that for all $y \in \mathsf{Y}$,

$$[\mathcal{I}^\phi(\mu_n)q](y) = \frac{\int_{\mathsf{T}} \mu_n(\mathrm{d}\theta)|b_{\mu_n}(\theta)|^\phi q(\theta,y)}{\int_{\mathsf{T}} \mu_n(\mathrm{d}\theta)|b_{\mu_n}(\theta)|^\phi} \; .$$

By applying twice Lemma 17-(iv) with $h(\theta) = 1$ and $h(\theta) = q(\theta, y)$, we have that for all $y \in \mathsf{Y}$,

$$\lim_{n\to\infty} [\mathcal{I}^\phi(\mu_n)q](y) = [\mathcal{I}^\phi(\bar{\mu})q](y) \; . \tag{34}$$

Now, since f is C^1 by (A2), we obtain from Lemma 17-(i) and (34) respectively that for all $y \in \mathsf{Y}$, $\lim_{n\to\infty} a_{\mu_n}(y) = a_{\bar{\mu}}(y)$ and $\lim_{n\to\infty} a_{\mathcal{I}^\phi(\mu_n)}(y) = a_{\mathcal{I}^\phi(\bar{\mu})}(y)$. Combining with (33) and (32) we can thus apply the Dominated Convergence Theorem to obtain

$$\lim_{n\to\infty} \Psi^{(f)}(\mu_n) = \Psi^{(f)}(\bar{\mu}) \tag{35}$$

and

$$\lim_{n\to\infty} \Psi^{(f)}(\mu_{n+1}) = \lim_{n\to\infty} \Psi^{(f)}(\mathcal{I}^\phi(\mu_n)) = \Psi^{(f)}(\mathcal{I}^\phi(\bar{\mu})) \; . \tag{36}$$

Finally, (35) and (36) together yield $\Psi^{(f)}(\bar{\mu}) = \Psi^{(f)} \circ \mathcal{I}^\phi(\bar{\mu})$, which in turn implies that $\bar{\mu}$ is a fixed point of $\mathcal{I}^\phi$ according to Theorem 1-(ii).

Proof of (ii). We prove (ii) by contradiction. Suppose that $\mu_n \Rightarrow \bar{\mu}$, where $\bar{\mu}$ is a fixed point of $\mathcal{I}^\phi$ that satisfies

$$\Psi^{(f)}(\bar{\mu}) > \inf_{\zeta\in\mathrm{M}_{1,\mu}(\mathsf{T})} \Psi^{(f)}(\zeta) \; .$$

Then, there exists $\mu^\star \in \mathrm{M}_{1,\mu}(\mathsf{T})$ such that $\Psi^{(f)}(\bar{\mu}) > \Psi^{(f)}(\mu^\star)$. Now for all $n \in \mathbb{N}$, set

$$B_n = \left\{ \theta \in \mathsf{T} \ : \ \bigcap_{m \geqslant n} \{ g_{\mu_m}(\theta) > \delta \mu_m(g_{\mu_m}) \} \right\} ,$$

where for all $\zeta \in \mathrm{M}_1(\mathsf{T})$, $g_\zeta := |b_\zeta|^\phi$. There exists, according to Lemma 20, for a well chosen $\delta > 1$, a sufficiently large n_0 such that $\mu^\star(B_{n_0}) > 0$.

Furthermore $\mu^\star \approx \mu$ by definition, where $\zeta \approx \mu$ if and only if for all $A \in \mathcal{T}$: $\zeta(A) > 0$ is equivalent to $\mu(A) > 0$. Since $0 < |b_\mu(\theta)|^\phi < \infty$ for μ-almost all $\theta \in \mathsf{T}$ and $\frac{\mathrm{d}\mu_1}{\mathrm{d}\mu} \propto |b_\mu|^\phi$, we also have $\mu_1 \approx \mu$. Then by induction, $\mu_n \approx \mu$ for all $n \in \mathbb{N}$. Finally, $\mu_{n_0}(B_{n_0}) > 0$. Moreover, for all $\theta \in B_{n_0}$ and all $m > n_0$, $\frac{g_{\mu_m}(\theta)}{\mu_m(g_{\mu_m})} > \delta$ and consequently

$$\mu_m(B_{n_0}) = \int_{B_{n_0}} \mu_{m-1}(\mathrm{d}\theta) \frac{g_{\mu_{m-1}}(\theta)}{\mu_{m-1}(g_{\mu_{m-1}})} \geqslant \delta \mu_{m-1}(B_{n_0}) \ .$$

By induction on m we get that, for all $m \geqslant n$, $\mu_m(B_{n_0}) \geqslant \delta^{m-n_0} \mu_{n_0}(B_{n_0})$. This contradicts the previously obtain facts that $\delta > 1$ and $\mu_{n_0}(B_{n_0}) > 0$. Therefore we get a contradiction and the proof is concluded. □

A.3. Lemma 21 : statement and proof. Recall that $Y_1, Y_2, ...$ are i.i.d random variables with common density μq w.r.t ν, defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and we denote by $\mathbb{E}$ the associated expectation operator.

LEMMA 21. *Assume (A1) and (A2). Let $\mu \in \mathrm{M}_1(\mathsf{T})$, $\phi \in \mathbb{R}^\star$ be such that $\mu(|b_\mu|) < \infty$ and*

$$\int_{\mathsf{T}} \mu(\mathrm{d}\theta) \mathbb{E} \left[\left\{ \frac{q(\theta, Y_1)}{\mu q(Y_1)} \left| f' \left(\frac{\mu q(Y_1)}{p(Y_1)} \right) \right| \right\}^\phi \right] < \infty \ . \tag{37}$$

Then,

$$\lim_{K \to \infty} \mu(|b_{\mu,K}|^\phi) = \mu(|b_\mu|^\phi) \, , \quad \mathbb{P} - \text{a.s.} \tag{38}$$

PROOF. Set $g(\theta, y) = \frac{q(\theta,y)}{\mu q(y)} |f'(\frac{\mu q(y)}{p(y)})|$ and note that $\mathbb{E}[g(\theta, Y_1)] = |b_\mu(\theta)|$ since f' is of constant sign.

(i) We start with the case $\phi \notin [0, 1]$. Our goal is to apply Lemma 22, which is a generalized version of the Dominated Convergence Theorem. To

do so, first note that $|b_\mu|^\phi$ is positive and combining with the convexity of the mapping $u \mapsto u^\phi$, we have for all $K \in \mathbb{N}^\star$ and for all $\theta \in \mathsf{T}$,

$$0 \leqslant |b_{\mu,K}(\theta)|^\phi \leqslant K^{-1} \sum_{k=1}^{K} [g(\theta, Y_k)]^\phi \, . \tag{39}$$

Since $\mu(|b_\mu|) < \infty$, the LLN for μ-almost all $\theta \in \mathsf{T}$ yields

$$\lim_{K \to \infty} b_{\mu,K}(\theta) = b_\mu(\theta) \, . \tag{40}$$

Now applying successively (a) the LLN for μ-almost all $\theta \in \mathsf{T}$ (as stated in Lemma 23), which is valid under (37), (b) Fubini's Theorem and (c) again the LLN

$$\begin{aligned} \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \lim_{K \to \infty} K^{-1} \sum_{k=1}^{K} \{g(\theta, Y_k)\}^\phi &\overset{(a)}{=} \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \mathbb{E}\left[\{g(\theta, Y_1)\}^\phi\right] \\ &\overset{(b)}{=} \mathbb{E}\left[\int_{\mathsf{T}} \mu(\mathrm{d}\theta) [g(\theta, Y_1)]^\phi\right] \overset{(c)}{=} \lim_{K \to \infty} \int_{\mathsf{T}} \mu(\mathrm{d}\theta) K^{-1} \sum_{k=1}^{K} [g(\theta, Y_k)]^\phi \end{aligned} \tag{41}$$

That is

$$\mu\left(\lim_{K \to \infty} K^{-1} \sum_{k=1}^{K} \{g(\cdot, Y_k)\}^\phi\right) = \lim_{K \to \infty} \mu\left(K^{-1} \sum_{k=1}^{K} [g(\cdot, Y_k)]^\phi\right) < \infty$$

Combining with (39) and (40), we apply Lemma 22 and obtain

$$\mu\left(|b_\mu|^\phi\right) = \mu\left(\lim_{K \to \infty} |b_{\mu,K}|^\phi\right) = \lim_{K \to \infty} \mu(|b_{\mu,K}|^\phi) \, .$$

(ii) We now turn to the case $\phi \in (0, 1]$. Let $M > 0$. Since

$$\int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left(K^{-1} \sum_{k=1}^{K} g(\theta, Y_k) \mathbf{1}_{\{g(\theta, Y_k) \leqslant M\}}\right)^\phi \leqslant \mu(|b_{\mu,K}|^\phi) \, ,$$

the LLN for μ-almost all $\theta \in \mathsf{T}$ (Lemma 23) and the Dominated Convergence Theorem yields

$$\int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left(\mathbb{E}[g(\theta, Y_1) \mathbf{1}_{\{g(\theta, Y_1) \leqslant M\}}]\right)^\phi \leqslant \liminf_{K \to \infty} \mu(|b_{\mu,K}|^\phi) \, . \tag{42}$$

Using now $(u+v)^\phi \leqslant u^\phi + v^\phi$ and then Jensen's inequality for the concave mapping $u \mapsto u^\phi$,

$$\mu(|b_{\mu,K}|^\phi) \leqslant \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left(K^{-1} \sum_{k=1}^{K} g(\theta, Y_k) \mathbf{1}_{\{g(\theta,Y_k)\leqslant M\}} \right)^\phi + \left(\int_{\mathsf{T}} \mu(\mathrm{d}\theta) K^{-1} \sum_{k=1}^{K} g(\theta, Y_k) \mathbf{1}_{\{g(\theta,Y_k)>M\}} \right)^\phi$$

By invoking the LLN for μ-almost all $\theta \in \mathsf{T}$ (Lemma 23) and the Dominated Convergence Theorem for the first term of the rhs and the LLN combined with Fubini for the second term, we get

$$\limsup_{K\to\infty} \mu(|b_{\mu,K}|^\phi) \leqslant \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left(\mathbb{E}[g(\theta, Y_1) \mathbf{1}_{\{g(\theta,Y_1)\leqslant M\}}] \right)^\phi + \left(\int_{\mathsf{T}} \mu(\mathrm{d}\theta) \mathbb{E}[g(\theta, Y_1) \mathbf{1}_{\{g(\theta,Y_1)>M\}}] \right)^\phi$$

Letting M go to infinity both in this inequality and in (42) completes the proof of (38).

□

APPENDIX B: TECHNICAL RESULTS

B.1. General Dominated Convergence Theorem. We state and prove a generalized version of the Dominated Convergence Theorem, adapted from [36, Theorem 19]. We provide here a full proof for the sake of completeness.

LEMMA 22 (General Dominated Convergence Theorem). *Let $\zeta \in \mathrm{M}_1(\mathsf{T})$. Assume there exist (a_K), (b_K), (c_K) three sequences of $(\mathcal{T}, \mathcal{B}(\mathbb{R}))$-measurable functions such that the limits $\lim_{K\to\infty} a_K(\theta)$, $\lim_{K\to\infty} b_K(\theta)$, $\lim_{K\to\infty} c_K(\theta)$ exist for ζ-almost all $\theta \in \mathsf{T}$ and*

$$\zeta|\lim_{K\to\infty} a_K| + \zeta|\lim_{K\to\infty} c_K| < \infty$$

Assume moreover that for all $K \in \mathbb{N}^\star$ and for ζ-almost all $\theta \in \mathsf{T}$

$$a_K(\theta) \leqslant b_K(\theta) \leqslant c_K(\theta)$$

and

$$\zeta(\lim_{K\to\infty} a_K) = \lim_{K\to\infty} \zeta(a_K) \tag{43}$$

$$\zeta(\lim_{K\to\infty} c_K) = \lim_{K\to\infty} \zeta(c_K) \tag{44}$$

Then,

$$\zeta(\lim_{K\to\infty} b_K) = \lim_{K\to\infty} \zeta(b_K)$$

PROOF. We apply Fatou's Lemma combined with (43) and (44) to the two non-negative, $(\mathcal{T}, \mathcal{B}(\mathbb{R}))$-measurable functions $\theta \mapsto b_K(\theta) - a_K(\theta)$ and $\theta \mapsto c_K(\theta) - b_K(\theta)$ and we obtain

$$\zeta(\liminf_{K\to\infty} b_K) \leqslant \liminf_{K\to\infty} \zeta(b_K)$$
$$\zeta(\liminf_{K\to\infty} -b_K) \leqslant \liminf_{K\to\infty} \zeta(-b_K)$$

which proves the lemma, as $\liminf_{K\to\infty} b_K(\theta) = \limsup_{K\to\infty} b_K(\theta)$ for ζ-almost all $\theta \in \mathsf{T}$. □

B.2. Integrated Law of Large Numbers. Let $Y_1, Y_2, \ldots$ be i.i.d. random variables on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let h be a non-negative real-valued $(\mathcal{T} \otimes \mathcal{F}, \mathcal{B}(\mathbb{R}_{\geqslant 0}))$-measurable function. We are interested in showing

$$\int_{\mathsf{T}} \zeta(\mathrm{d}\theta) \lim_{K\to\infty} K^{-1} \sum_{k=1}^{K} h(\theta, Y_k) = \int_{\mathsf{T}} \zeta(\mathrm{d}\theta) \mathbb{E}[h(\theta, Y_1)] \tag{45}$$

for $\zeta \in \mathrm{M}_1(\mathsf{T})$ satisfying $\int_{\mathsf{T}} \zeta(\mathrm{d}\theta)\mathbb{E}[h(\theta, Y_1)] < \infty$. While this result follows easily if we can show that

$$\mathbb{P}\left(\forall \theta \in \mathsf{T},\ \lim_{K\to\infty} K^{-1} \sum_{k=1}^{K} h(\theta, Y_k) = \mathbb{E}[h(\theta, Y_1)]\right) = 1 \tag{46}$$

unfortunately the LLN only yields

$$\mathbb{P}\left(\lim_{K\to\infty} K^{-1} \sum_{k=1}^{K} h(\theta, Y_k) = \mathbb{E}[h(\theta, Y_1)]\right) = 1$$

for ζ-almost all $\theta \in \mathsf{T}$. The following lemma allows to show (45) without resorting to the much stronger identity (46).

LEMMA 23. *Let* $\zeta \in \mathrm{M}_1(\mathsf{T})$ *and assume that* $\int_{\mathsf{T}} \zeta(\mathrm{d}\theta)\mathbb{E}[h(\theta, Y_1)] < \infty$. *Then,* $\mathbb{P}-$a.s.

$$\int_{\mathsf{T}} \zeta(\mathrm{d}\theta) \lim_{K\to\infty} K^{-1} \sum_{k=1}^{K} h(\theta, Y_k) = \int_{\mathsf{T}} \zeta(\mathrm{d}\theta)\mathbb{E}[h(\theta, Y_1)] \; .$$

PROOF. Set

$$B = \left\{ (\theta, \omega) \in \mathsf{T} \times \Omega \; : \; \lim_{K\to\infty} K^{-1} \sum_{k=1}^{K} h(\theta, Y_k(\omega)) = \mathbb{E}[h(\theta, Y_1)] \right\} \; ,$$

Let $\gamma_0 : (\theta, \omega) \mapsto \mathbf{1}_{B^c}(\theta, \omega)$ and $\gamma_1 = 1 - \gamma_0$. According to the Fubini Theorem and the LLN for $K^{-1}\sum_{k=1}^{K} h(\theta, Y_k)$ where θ is such that $\mathbb{E}[h(\theta, Y_1)] < \infty$ (which is satisfied for ζ-almost all $\theta \in \mathsf{T}$ by assumption),

$$\mathbb{E}\left[\int_{\mathsf{T}} \zeta(\mathrm{d}\theta)\gamma_0(\theta, \cdot) \right] = \int_{\mathsf{T}} \zeta(\mathrm{d}\theta)\mathbb{E}\left[\gamma_0(\theta, \cdot)\right] = 0 \; .$$

Therefore, $\int_{\mathsf{T}} \zeta(\mathrm{d}\theta)\gamma_0(\theta, \cdot)$ is $\mathbb{P}-$a.s. null that is, there exists Ω_1 such that $\mathbb{P}(\Omega_1) = 1$ and for all $\omega \in \Omega_1$, $A \mapsto \int_A \zeta(\mathrm{d}\theta)\gamma_0(\theta, \omega)$ is the null-measure on $(\mathsf{T}, \mathcal{T})$, which in turn implies that the measures ζ and $A \mapsto \int_A \zeta(\mathrm{d}\theta)\gamma_1(\theta, \omega)$ coincide. The latter property implies for all $\omega \in \Omega_1$,

$$\begin{aligned}
\int_{\mathsf{T}} \zeta(\mathrm{d}\theta)\mathbb{E}[h(\theta, Y_1)] &= \int_{\mathsf{T}} \zeta(\mathrm{d}\theta)\mathbb{E}[h(\theta, Y_1)]\gamma_1(\theta, \omega) \\
&= \int_{\mathsf{T}} \zeta(\mathrm{d}\theta) \left[\lim_{K\to\infty} K^{-1} \sum_{k=1}^{K} h(\theta, Y_k(\omega)) \right] \gamma_1(\theta, \omega) \\
&= \int_{\mathsf{T}} \zeta(\mathrm{d}\theta) \lim_{K\to\infty} K^{-1} \sum_{k=1}^{K} h(\theta, Y_k(\omega)) \; .
\end{aligned}$$

□

APPENDIX C: α-BOUND FOR Z

PROOF OF LEMMA 10. We derive the explicit link between $D_f(\mu Q||\tilde{\mathbb{P}})$ and $\Psi^{(f)}(\mu; p)$ with $p = p^*$ for each of the two divergences:

(a) Reverse Kullback-Leibler: $f(u) = -\log(u)$ and

$$\begin{aligned}
D_f(\mu Q||\tilde{\mathbb{P}}) &= \int_{\mathsf{Y}} -\log\left(\frac{\mu q(y)}{\tilde{p}(y)}\right) \tilde{p}(y)\nu(\mathrm{d}y) \\
&= \frac{1}{Z}\int_{\mathsf{Y}} -\log\left(\frac{\mu q(y)}{p^*(y)}\right) p^*(y)\nu(\mathrm{d}y) - \log Z \\
&= \frac{1}{Z}\Psi^{(f)}(\mu; p) - \log Z
\end{aligned}$$

(b) α-divergence: For all $\alpha \notin \{0,1\}$, $f(u) = \frac{1}{\alpha(\alpha-1)}[u^\alpha - 1]$ and

$$
\begin{aligned}
(47) \qquad D_f(\mu Q||\tilde{\mathbb{P}}) &= \int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)} \left[\left(\frac{\mu q(y)}{\tilde{p}(y)} \right)^\alpha - 1 \right] \tilde{p}(y)\nu(\mathrm{d}y) \\
&= Z^{\alpha-1} \int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)} \left(\frac{\mu q(y)}{p^*(y)} \right)^\alpha p^*(y)\nu(\mathrm{d}y) - \frac{1}{\alpha(\alpha-1)} \\
&= Z^{\alpha-1} \Psi^{(f)}(\mu; p) + \frac{1}{\alpha(\alpha-1)} \left[Z^\alpha - 1 \right]
\end{aligned}
$$

□

APPENDIX D: MORE ILLUSTRATIONS

FIGURE 3. *Dimension 1: Impact of the parameter ϕ. Here $\alpha = 0.5$ with corresponding range $(0, 1]$ for ϕ.*

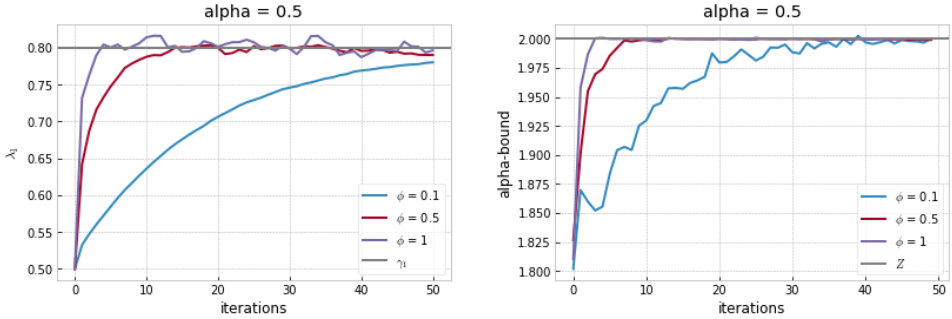
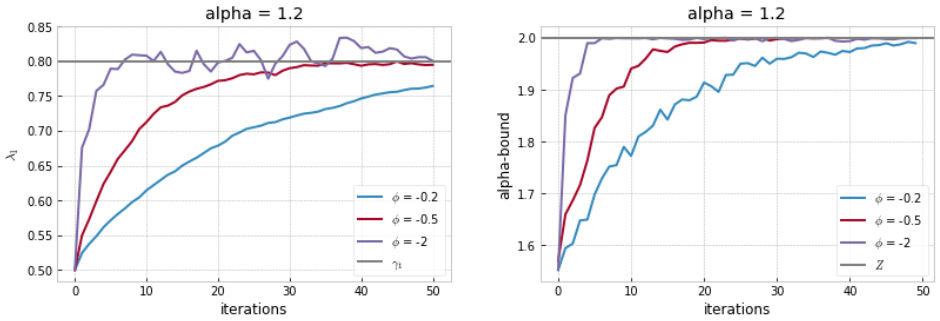


FIGURE 4. *Dimension 1: Impact of the parameter ϕ. Here $\alpha = 1.2$ with corresponding range $(-5, 0)$ for ϕ.*



LTCI, Télécom Paris
Institut Polytechnique de Paris
46, rue Barrault, 75013 Paris
E-mail: kamelia.daudel@telecom-paris.fr
francois.portier@telecom-paris.fr
francois.roueff@telecom-paris.fr

Département CITI, Télécom SudParis
Institut Polytechnique de Paris
9 rue Charles Fourier, 91000 Evry
E-mail: randal.douc@telecom-sudparis.eu