A divergence-based condition to ensure quantile improvement in black-box global optimization

Thomas Guilmeau, Emilie Chouzenoux, *Senior Member, IEEE*, and Víctor Elvira, *Senior Member, IEEE*

***Abstract*—Black-box global optimization aims at seeking for the minimizers of an objective function whose analytical form is not known. To do so, many state-of-the-art methods rely on sampling-based strategies, where sampling distributions are built in an iterative fashion, so that their mass concentrate where the objective function is low. Despite empirical success, the convergence of these methods remains difficult to show theoretically. In this work, we introduce a new framework, based on divergence-decrease conditions, to study and design black-box global optimization algorithms. We show that the information-geometric optimization approach fits within our framework, which yields a new proof for its convergence analysis. We also establish a quantile improvement result for two novel algorithms, one related with the cross-entropy approach with mixture models, and another using heavy-tailed sampling distributions.**

***Index Terms*—Black-box optimization ; Variational inference ; Mixture models ; Heavy-tailed distributions ; Kullback-Leibler divergence.**

I. Introduction

Finding the minimizer of a possibly non-convex objective function that is only accessible through a black-box oracle is a challenging, yet important task, which has motivated many works [1]. Given the presence of eventual local minima and the difficulty of evaluating or even approximating gradients, many methods resort to sampling procedures. These rely on evolution strategies to construct proposal distributions [2], typically Gaussian ones, to generate samples close to the minimizers of the objective. Among these methods, one can mention the class of estimation of distribution algorithms [3], the cross-entropy algorithm [4], or the CMA-ES algorithm [5], [6].

A useful perspective to gain theoretical insights is to understand these algorithms as optimization schemes aiming at finding the proposal with the lowest expected objective function value [7]–[9], or with the lowest expected transformed objective value, for some well-chosen transformation [10], [11]. However, the resulting optimization dynamics usually does not exactly map the actual algorithms used in practice. Algorithms such as CMA-ES exhibit invariance properties with respect to transformations of the initial problem, such as translation, rotation, and even monotonic transformation of the objective function. Apart from allowing to generalize any insight one has on such algorithms to a broader class of objective functions, algorithms with invariance properties also seem to perform better [12], [13].

T.G. and E.C. are with Université Paris-Saclay, CentraleSupélec, INRIA, CVN, France. V.E. is with the School of Mathematics, University of Edinburgh, UK.

E.C. and T.G. acknowledge support from the European Research Council Starting Grant MAJORIS ERC-2019-STG850925. The work of V. E. is supported by ARL/ARO under grant W911NF-22-1-0235.

The authors of [14] transformed the original black-box optimization problem into an optimization problem over the proposals while preserving good invariance properties. This is done by considering the optimization, over proposals, of the expectation of a quantile-based rewriting of the objective function. They furthermore introduced the so-called information-geometric optimization (IGO) framework to minimize the resulting objective, which was further studied in [15]. The IGO framework recovers many existing algorithms, such as the cross-entropy (CE) algorithm [4], also based on quantiles, or various estimation of distribution algorithms [3]. Within this framework, the authors of [14], [15] showed that the IGO algorithm yields an improvement on the original optimization problem resolution, in a quantile-based sense. For quadratic convex problems and Gaussian proposals, further convergence results have also been established in [13], [16]. IGO is based on natural gradient updates shown to improve the proposals. Such an improvement can be shown assuming infinitesimal step sizes, or for larger step-sizes when the proposals form an exponential family. However, in practice, one would prefer to avoid both taking too small sizes and to be limited to the exponential family, for instance by exploring heavy-tailed [17] or mixture distributions [18]–[21] for the proposals.

Our first contribution is the introduction of novel conditions, not tied to any particular choice of algorithm, to establish quantile-based improvement results in the original optimization problem in the same sense as in [14], [15]. Our conditions rely on measuring, through a Kullback-Leibler (KL) or an α divergence, the discrepancy between a given target distribution and successive proposals. If the divergence is decreasing between two consecutive proposals, then quantile-based improvement is shown. It is worthy to emphasize that our result does not depend on the way the next proposal is designed, meaning that the divergence-decrease conditions can be checked on any algorithm.

Our divergence-decrease conditions are close in essence to methods from the field of variational inference in statistics [22], [23]. This allows us to leverage tools from this domain to get novel convergence results for global black-box optimization algorithms, leading us to our second contribution. First, we show that the IGO algorithm satisfies the introduced divergence-decrease condition, which allows us to derive new proofs for results in [14], [15]. Second, we go beyond the scope of the aforementioned IGO works, by considering mixture and heavy-tailed proposals. We propose a novel mixture-based algorithm, reminiscent from the mixture-based CE algorithm of [4, Example 3.2], and we show that it fits within our framework, and, as such, that it ensures quantile improvement at every iteration. We also propose a

novel algorithm for Student distribution, which have heavier tails than Gaussians and include Cauchy distributions, and use our framework to ensure quantile improvement at every iteration.

The paper in organized as follows. We give preliminary notions about black-box global optimization problems and algorithms to solve them in Section II. We then give our main results in Section III, including our novel conditions for improvement, and new results obtained under these conditions. Finally, we discuss perspectives in Section IV.

II. Preliminary notions

Let us start with some preliminary notions on sampling-based black-box global optimization algorithms. These algorithms sample points from a proposal distribution updated iteratively so it concentrates around the solutions of the sought optimization problem. We first present how the initial optimization problem can be relaxed into an optimization problem on proposal distributions. Then, we discuss algorithms to solve this resulting problem, making a particular focus on the IGO framework.

A. Problem setting and problem rewriting

We consider throughout the paper the general form of black-box minimization problem

$$\underset{x\in\mathbb{X}}{\text{minimize}}\, f(x), \tag{1}$$

where $f : \mathbb{X} \to \mathbb{R}$ is possibly non-convex and can only be accessed through a black-box that, for a given $x \in \mathbb{X}$, returns the value $f(x)$. The search space $\mathbb{X}$ can be continuous, discrete, or involving mixed continuous and discrete variables.

We focus in our work on algorithms that solve Problem (1) through a sampling-based approach. The aim is to construct parametric probability distributions p_θ over $\mathbb{X}$ such that p_θ is concentrated around the minimizers of f over $\mathbb{X}$. In this context, one does not search anymore for an optimal point $x \in \mathbb{X}$, but instead for an optimal parameter $\theta \in \Theta$, or alternatively, for an optimal probability distribution $p_\theta \in \{p_\theta, \theta \in \Theta\}$. One way to transform Problem (1) into a problem over Θ is to consider the minimization of $\theta \longmapsto \mathbb{E}_{X\sim p_\theta}[f(X)]$. This construction, which has been studied in [8], [9] for instance, makes however the resulting algorithm sensitive to transformation of f.

Another way to reformulate Problem (1) has been proposed in the context of IGO, in [14], and has the advantage of ensuring more invariance properties. Let the p_θ-f-quantiles at $x \in \mathbb{X}$ be defined by

$$q_\theta^{\leq}(x) = \mathbb{P}_{X\sim p_\theta}[f(X) \leq f(x)],$$
$$q_\theta^{<}(x) = \mathbb{P}_{X\sim p_\theta}[f(X) < f(x)],$$

and select a weighting non-increasing function $w : [0, 1] \to \mathbb{R}$. The authors of [14] then introduced the preference function $W_\theta^f : \mathbb{X} \to \mathbb{R}$ which is defined for any $x \in \mathbb{X}$ as

$$W_\theta^f(x) = \begin{cases} w(q_\theta^{\leq}(x)) \text{ if } q_\theta^{\leq}(x) = q_\theta^{<}(x), \\ \frac{1}{q_\theta^{\leq}(x) - q_\theta^{<}(x)} \int_{q_\theta^{<}(x)}^{q_\theta^{\leq}(x)} w(u)du \text{ otherwise.} \end{cases} \tag{2}$$

The function W_θ^f is a quantile-based rewriting of f that is invariant under increasing transformation of the objective f and reflects the behavior of f. Indeed, consider $(x, x') \in \mathbb{X}^2$ such that $f(x) \leq f(x')$, $q_\theta^{\leq}(x) = q_\theta^{<}(x)$, and $q_\theta^{\leq}(x') = q_\theta^{<}(x')$. Then, one has $W_\theta^f(x) \geq W_\theta^f(x')$. Under these definitions, given a proposal with parameter $\theta' \in \Theta$, the authors of [14] consider the search for a good proposal p_θ to solve (1) as the maximization of the function $J(\cdot|\theta') : \Theta \to \mathbb{R}$ defined for any $\theta \in \Theta$ by

$$J(\theta|\theta') = \mathbb{E}_{X\sim p_\theta}\left[W_{\theta'}^f(X)\right]. \tag{3}$$

Note that we have that $J(\theta|\theta) = Z_w$ for any $\theta \in \Theta$, with the notation $Z_w = \int_0^1 w(u)du$.

Using quantiles to link Problem (1) and the choice of a good proposal p_θ has also been proposed in the framework of the cross-entropy (CE) method [4]. Given a proposal p_θ and a scalar $q \in (0, 1)$, such method relies on q-quantiles of $f(X)$ where $X \sim p_\theta$, that is, any value $u \in \mathbb{R}$ such that

$$\mathbb{P}_{X\sim p_\theta}[f(X) \leq u] \geq q \text{ and } \mathbb{P}_{X\sim p_\theta}[f(X) \geq u] \geq 1 - q. \tag{4}$$

Let us denote, as in [15], $Q_\theta^q(f)$ as the largest of those values, that is

$$Q_\theta^q(f) = \sup\{u \in \mathbb{R} \text{ such that (4) is satisfied}\}. \tag{5}$$

The link between Problem (1) and $\theta \longmapsto Q_\theta^q(f)$ becomes straightforward. For a given value $q \in (0, 1)$, the best proposal p_θ is such that $Q_\theta^q(f)$ is as low as possible. Indeed, if one samples a point x from p_θ, then $f(x)$ is below $Q_\theta^q(f)$ with a probability at least equal to q.

It is actually possible to relate the behavior of the quantities $J(\theta|\theta')$ from the IGO framework, and $Q_\theta^q(f)$ from CE methods, for a particular case of weighting scheme w. This is done in [15] where it is shown that an increase in term of $\theta \longmapsto J(\theta|\theta')$ relates to a decrease in terms of $\theta \longmapsto Q_\theta^q(f)$.

Lemma 1 (Lemma 8 in [15])**.** *Consider the weighting function $w(u) = \delta_{u\leq q}(u)$ with $q \in (0, 1)$. If $(\theta, \theta') \in \Theta^2$ satisfy the increase condition*

$$J(\theta|\theta') > J(\theta'|\theta') = Z_w, \tag{6}$$

then we have $Q_\theta^q(f) \leq Q_{\theta'}^q(f)$. If further, $\mathbb{P}_{X\sim p_\theta}[f(X) = Q_{\theta'}^q(f)] = 0$, then $Q_\theta^q(f) < Q_{\theta'}^q(f)$.

The above result gives insights into the design and study of theoretically sounded black-box global optimization algorithm. Indeed, in order to study a given algorithm, a good course of action is to show that given a current proposal $p_{\theta'}$, the algorithm constructs a proposal p_θ such that the increase condition (6) is satisfied, which then allows to apply Lemma 1.

B. The information-geometric optimization algorithm

We now recall here the information-geometric (IGO) framework from [14], [15]. The latter is an iterative proposal construction algorithm, explicitly designed to achieve the increase condition (6) at every iteration. The IGO framework has been shown in [14] to recover many existing algorithms to solve Problem (1), both in discrete or continuous domains.

Among the algorithms recovered by the IGO framework, let us mention the CE algorithm of [4].

The quantity $J(\theta|\theta')$ defined in (3) is generally an intractable integral that needs in practice to be approximated with sampling. Throughout this paper, we focus on idealized algorithms that are deterministic, corresponding to the limit of an infinite number of samples. In such idealized setting, we only consider discrete-time updates since they are closer to implementation than continuous flows. Two types of updates have been proposed in [14], [15] to satisfy the increase condition (6), leading to two distinct IGO-like algorithms, that we will recall here.

The first algorithm in [14], [15] is based on natural gradient ascent updates. Let consider an iteration $k \in \mathbb{N}$, with θ_k being the parameter of the current proposal, and the objective function being $J(\cdot|\theta_k)$. The natural gradient of $J(\cdot|\theta_k)$ at θ is the quantity $\widetilde{\nabla}_\theta J(\theta|\theta_k) = I(\theta)^{-1}\nabla_\theta J(\theta|\theta_k)$, where $I(\theta) = -\mathbb{E}_{X\sim p_\theta}[\nabla^2_\theta \ln p_\theta(X)]$ is the Fisher information matrix of p_θ. Given the above gradient expression, iterating a gradient ascent scheme over $k \in \mathbb{N}$ leads to Algorithm 1.

Algorithm 1 IGO algorithm (natural gradient update)

Initialize θ_0 and choose the step-size $\tau > 0$.
for $k = 0, \dots$ **do**
Update θ_{k+1} such that

$$\theta_{k+1} = \theta_k + \tau \widetilde{\nabla}_\theta J(\theta|\theta_k)_{|\theta=\theta_k}. \tag{7}$$

end for

The second algorithm proposed in [14], [15] updates the proposal parameters by performing a weighted maximum likelihood update at every iteration. We provide its description in Algorithm 2. The CE algorithm of [4] is recovered as a special case when the step-size is $\tau = 1$ and the weighting function is $w(u) = \delta_{u\leq q}(u)$ [14].

Algorithm 2 IGO algorithm (IGO-ML update)

Initialize θ_0 and choose the step-size $\tau > 0$.
for $k = 0, \dots$ **do**
Update θ_{k+1} such that

$$\theta_{k+1} = \arg\max_{\theta\in\Theta} \Big((1-\tau) \int \ln(p_\theta(x)) p_{\theta_k}(x) m(dx) + \tau \int W^f_{\theta_k}(x) \ln p_\theta(x) p_{\theta_k}(x) m(dx) \Big). \tag{8}$$

end for

The theoretical properties of Algorithms 1 and 2 have been studied in [14], [15]. Algorithm 2 achieves the increase condition (6), that is $J(\theta_{k+1}|\theta_k) > Z_w$ at every iteration $k \in \mathbb{N}$, for step-sizes $\tau \in (0, 1]$ [15, Theorem 6]. This result gives in turn theoretical guarantees for the CE algorithm thanks to Lemma 1. Algorithm 1 has been shown to satisfy the increase condition (6) for sufficiently small step-sizes [14, Proposition 7]. Moreover, Algorithms 1 and 2 have been shown to coincide when $\{p_\theta,\, \theta \in \Theta\}$ forms an exponential family [24], ensuring that Algorithm 1 satisfies (6) for $\tau \in (0, 1]$ in this case (see [15, Corollary 7]).

Algorithms 1 and 2 coincide with many existing black-box global optimization algorithms [14, Section 5], allowing to get theoretical guarantees for these algorithms as well. However, this study requires to show that the considered algorithms fits within the IGO framework, which is not always possible nor straightforward. In the next section, we give novel broader conditions under which the increase condition (6) is satisfied. This allows us to exhibit theoretical guarantees beyond the IGO framework, the latter being retrieved as a special case.

III. A GENERAL DIVERGENCE-BASED CONDITION FOR QUANTILE IMPROVEMENT

We present our main results in this section. We start in Section III-A with the introduction of novel, divergence-based, conditions. We show that they imply the increase condition (6). We then show in Section III-B that IGO algorithms satisfy our conditions, allowing us to provide a new perspective on these methods. Finally, we exploit our divergence-based conditions to obtain new theoretical guarantees for algorithms using proposals that do not belong to an exponential family. In Section III-C, we study a mixture-based algorithm linked with the mixture-based CE algorithm of [4, Example 3.2]. In Section III-D, we study an algorithm with heavy-tailed proposals, namely Student distributions with arbitrary degree of freedom parameter.

A. Quantile improvement with divergence-decreasing steps

The goal of this section is to show that the increase condition (6), i.e., the theoretical guarantee achieved in the IGO framework, can be expressed as a consequence of divergence-based conditions, that we detail below. Combining these conditions with Lemma 1 then yields a quantile improvement result.

Our divergence-based conditions can be interpreted as the search for a proposal closer to a specific target distribution than the previous proposal, in the sense of a Kullback-Leibler, or an α divergence. Let us start by specifying the target probability distribution we are going to consider. For any $\theta \in \Theta$, we introduce π^f_θ, the probability density with respect to m defined for any $x \in \mathbb{X}$ by

$$\pi^f_\theta(x) = \frac{1}{Z_w} W^f_\theta(x) p_\theta(x). \tag{9}$$

When $w(u) = \delta_{u\leq q}(u)$ for some $q \in (0, 1)$, π^f_θ is a truncated version of p_θ with support being the points $x \in \mathbb{X}$ such that $q^<_\theta(x) < q$, meaning that areas of $\mathbb{X}$ where the values reached by f are too high are given zero mass. Let a given iteration $k \in \mathbb{N}$. We aim at measuring the discrepancy between the target $\pi^f_{\theta_k}$ and either the current proposal, or the next one. This discrepancy is measured using the KL or an α divergence, with $\alpha \in (0, 1) \cup (1, +\infty)$. These are defined, respectively, for any probability densities with respect to m p_1 and p_2 by

$$KL(p_1, p_2) = \int \ln\left(\frac{p_1(x)}{p_2(x)}\right) p_1(x) m(dx)$$

$$D_\alpha(p_1, p_2) = \frac{1}{\alpha(\alpha-1)} \left(\int p_1(x)^\alpha p_2(x)^{1-\alpha} m(dx) - 1 \right).$$

If for some $x \in \mathbb{X}$, $p_1(x) = 0$, then we use $\ln(p_1(x))p_1(x) = 0$ (see [25, Definition 7.1] for more details on these singular cases). Note that we have $D_\alpha(p_1, p_2) \xrightarrow[\alpha \to 0]{} KL(p_2, p_1)$ [26], so that KL can be viewed as a limit case of the α-divergence.

Remark 1. *All the following results could be stated with any tilted density of the form $\pi_\theta^f(x) \propto V_\theta^f(x)p_\theta(x)$ with $V_\theta^f : \mathbb{X} \to \mathbb{R}$ instead of W_θ^f. Then, the normalization constant of π_θ^f is $\int V_\theta^f(x)p_\theta(x)m(dx)$ and the results are stated with $G(\cdot|\theta') : \theta \longmapsto \mathbb{E}_{X\sim p_\theta}[V_{\theta'}^f(X)]$ instead of $J(\cdot|\theta')$. For instance, a possible choice is $V_\theta^f(x) = \phi(f(x))$ for some transform $\phi : \mathbb{R} \to \mathbb{R}$ (see for instance [10, Section 3.1]). We stick with W_θ^f defined in* (2) *and $J(\cdot|\theta)$ defined in* (3) *as this choice is connected to the quantiles $Q_\theta^q(f)$ as shown by Lemma 1 and has been shown to yield faster convergence than other choices [13].*

We are now ready to state our first result, obtained when using the KL divergence to measure the discrepancy between probability densities.

Proposition 1. *Let $k \in \mathbb{N}$ and $\theta_k \in \Theta$. Suppose that $\pi_{\theta_k}^f$ is given by Equation* (9)*, and $p_{\theta_{k+1}}$ satisfies*

$$KL(\pi_{\theta_k}^f, p_{\theta_{k+1}}) < KL(\pi_{\theta_k}^f, p_{\theta_k}). \tag{10}$$

Then, (θ_{k+1}, θ_k) satisfy the increase condition (6)*, i.e., $J(\theta_{k+1}|\theta_k) > Z_w$ with J defined in* (3)*. The non-strict version of* (10) *implies the non-strict version of* (6)*.*

Proof. By construction of $\pi_{\theta_k}^f$, we can rewrite condition (10) as

$$\int \ln\left(\frac{W_{\theta_k}^f(x)p_{\theta_k}(x)}{Z_w p_{\theta_{k+1}(x)}}\right)\pi_{\theta_k}^f(x)m(dx) < \int \ln\left(\frac{W_{\theta_k}^f(x)}{Z_w}\right)\pi_{\theta_k}^f(x)m(dx),$$

and remark that it is equivalent to having

$$-\int \ln\left(\frac{p_{\theta_{k+1}(x)}}{p_{\theta_k}(x)}\right)\pi_{\theta_k}^f(x)m(dx) < 0.$$

We then get from Jensen's inequality that

$$-\ln\left(\int \frac{p_{\theta_{k+1}}(x)}{p_{\theta_k}(x)}\pi_{\theta_k}^f(x)m(dx)\right) \leq -\int \ln\left(\frac{p_{\theta_{k+1}(x)}}{p_{\theta_k}(x)}\right)\pi_{\theta_k}^f(x)m(dx),$$

implying that

$$\int \frac{p_{\theta_{k+1}(x)}}{p_{\theta_k}(x)}\pi_{\theta_k}^f(x)m(dx) > 1,$$

which shows, by definition of $\pi_{\theta_k}^f$, that $J(\theta_{k+1}|\theta_k) > Z_w$. The case where condition (10) is written with a non-strict inequality follows with the same steps. □

We now state a second result, that arises when one measures the discrepancy between the target density and the proposals using an α-divergence.

Proposition 2. *Let $k \in \mathbb{N}$, $\theta_k \in \Theta$ and suppose that $W_{\theta_k}^f$ takes values in $\{0, 1\}$. Suppose that $\pi_{\theta_k}^f$ is given by Equation* (9) *and that $p_{\theta_{k+1}}$ satisfies*

$$D_\alpha(p_{\theta_{k+1}}, \pi_{\theta_k}^f) < D_\alpha(p_{\theta_k}, \pi_{\theta_k}^f), \tag{11}$$

for some $\alpha \in (0, 1)$. Then, (θ_{k+1}, θ_k) satisfy the increase condition (6)*, i.e., $J(\theta_{k+1}|\theta_k) > Z_w$. If the inequality in* (11) *is not strict, then we get* (6) *with a non-strict inequality.*

Proof. Since $\alpha < 1$, given the assumption on the preference function $W_{\theta_k}^f$, the condition (11) is equivalent to

$$\int p_{\theta_{k+1}}(x)^\alpha W_{\theta_k}^f(x)p_{\theta_k}(x)^{1-\alpha}m(dx) > \int W_{\theta_k}^f(x)p_{\theta_k}(x)m(dx),$$

from which we can deduce that $\int \left(\frac{p_{\theta_{k+1}}(x)}{p_{\theta_k}(x)}\right)^\alpha \pi_{\theta_k}^f(x)m(dx) > 1$. Remark now that $u \longmapsto u^\alpha$ is concave since $\alpha < 1$, so Jensen's inequality allows to upper-bound the left-hand side in the above and obtain that $\left(\frac{1}{Z_w}\int W_{\theta_k}^f(x)p_{\theta_{k+1}}(x)m(dx)\right)^\alpha > 1$, implying the result. If the inequality in (11) is non-strict, we follow the same steps and obtain that the inequality in (6) is also non-strict. □

Propositions 1 and 2 establish divergence-decrease conditions under which the increase condition (6) is satisfied. Note that the construction mechanism of $p_{\theta_{k+1}}$ does not intervene in our result, while in in [14], [15], the increase condition (6) was achieved for specific algorithms only.

As we already explained in the previous Sections, in the particular case when $w(u) = \delta_{u \leq q}(u)$, there is a link between $J(\theta|\theta_k)$, defined in (3), and $Q_\theta^q(f)$, defined in (5). In this case, we can extend the result of Propositions 1 and 2 with Lemma 1 and obtain the following result.

Corollary 1. *Assume that $w(u) = \delta_{u \leq q}(u)$ for some $q \in (0, 1)$ and that, at a given iteration $k \in \mathbb{N}$, $p_{\theta_{k+1}}$ is constructed such that*

$$D_\alpha(p_{\theta_{k+1}}, \pi_{\theta_k}^f) < D_\alpha(p_{\theta_k}, \pi_{\theta_k}^f) \tag{12}$$

is satisfied for some $\alpha \in [0, 1)$ (the case $\alpha = 0$ corresponding to the inequality $KL(\pi_{\theta_k}^f, p_{\theta_{k+1}}) < KL(\pi_{\theta_k}^f, p_{\theta_k})$). Then, we have the following quantile improvement results.

- (i) *If $\alpha = 0$, then $Q_{\theta_{k+1}}^q(f) \leq Q_{\theta_k}^q(f)$.*
- (ii) *If $\alpha \in (0, 1)$ and $W_{\theta_k}^f$ takes values in $\{0, 1\}$, then $Q_{\theta_{k+1}}^q(f) \leq Q_{\theta_k}^q(f)$.*
- (iii) *If $\mathbb{P}_{X\sim p_\theta}[f(X) = u] = 0$ for any $\theta \in \Theta$, $u \in \mathbb{R}$, then $Q_{\theta_{k+1}}^q(f) < Q_{\theta_k}^q(f)$.*

Proof. Points (i) and (ii) follow from Propositions 1 and 2 together with the first part of Lemma 1. Point (iii) follows by remarking that, under our assumptions, for any $\theta \in \Theta$, $x \in \mathbb{X}$, $q_\theta^<(x) = q_\theta^\leq(x)$, ensuring that $W_{\theta_k}^f(x)$ takes values in $\{0, 1\}$. This also implies that $\mathbb{P}_{X\sim p_{\theta_{k+1}}}[f(X) = Q_{\theta_k}^q(f)] = 0$, allowing to apply the second part of Lemma 1, and conclude the proof. □

We now show that the KL and α divergences can also be used to control the discrepancy between the proposal p_θ and

the resulting target π_θ^f. The resulting bound only depends on the choice of the weighting function w.

Proposition 3. *Consider $\theta \in \Theta$ and the probability densities p_θ and π_θ^f. We have the following results.*

(i) If W_θ^f takes values in $\{0,1\}$, then $D_\alpha(p_\theta, \pi_\theta^f) = \frac{1}{\alpha(\alpha-1)}(Z_w^\alpha - 1)$ for any $\alpha \in (0,1)$ and $KL(\pi_\theta^f, p_\theta) = -\ln Z_w$.

(ii) If w takes values in $[0,1]$, then $D_\alpha(p_\theta, \pi_\theta^f) \leq \frac{1}{\alpha(\alpha-1)}(Z_w^\alpha - 1)$ for any $\alpha \in (0,1)$ and $KL(\pi_\theta^f, p_\theta) \leq -\ln Z_w$.

Proof. Consider any $\alpha \in (0,1) \cup (1, +\infty)$, we have

$$D_\alpha(p_\theta, \pi_\theta^f) = \frac{1}{\alpha(\alpha-1)} \left(\frac{\int W_\theta^f(x)^{1-\alpha} p_\theta(x) m(dx)}{Z_w^{1-\alpha}} - 1 \right). \quad (13)$$

(i) We have that $W_\theta^f(x)^{1-\alpha} = W_\theta^f(x)$ for any $x \in \mathbb{X}$, thus showing with Equation (13) that $D_\alpha(p_\theta, \pi_\theta^f) = \frac{1}{\alpha(\alpha-1)}(Z_w^\alpha - 1)$ for any $\alpha \in (0,1) \cup (1, +\infty)$ which implies the result. In the limit $\alpha \to 0$, we get that $D_\alpha(p_\theta, \pi_\theta^f) \to KL(\pi_\theta^f, p_\theta)$ following [26] and that $\frac{1}{\alpha(\alpha-1)}(Z_w^\alpha - 1) \to -\ln Z_w$, establishing the second part of the result.

(ii) Since w takes values in $[0,1]$, we also have $W_\theta^f(x) \in [0,1]$ for any $x \in \mathbb{X}$. Therefore, $W_\theta^f(x)^{1-\alpha} \geq W_\theta^f(x)$ for any $x \in \mathbb{X}$ when $\alpha \in (0,1)$. We thus get from Equation (13) when $\alpha \in (0,1)$ that $D_\alpha(p_\theta, \pi_\theta^f) \leq \frac{1}{\alpha(\alpha-1)}(Z_w^\alpha - 1)$. Taking the limit $\alpha \to 0$, we finally obtain that $KL(\pi_\theta^f, p_\theta) \leq -\ln Z_w$. □

Remark 2. *Consider, following Remark 1, that we use a target density of the form $\pi_\theta^f(x) \propto V_\theta^f(x) p_\theta(x)$, choosing $V_\theta^f : \mathbb{X} \to \mathbb{R}$ instead of W_θ^f. Then, the normalization constant of π_θ^f is $\int V_\theta^f(x) p_\theta(x) m(dx)$ which may depend on the parameter θ. On the contrary, using W_θ^f ensures that the normalization constant of π_θ^f is equal to Z_w for any $\theta \in \Theta$.*

With Propositions 1 and 2, we have shown that if the divergence between the target and the next proposal is lower than the divergence between the target and the current proposal, Equation (6) is satisfied. In the particular case of an indicator weighting function, Corollary 1 shows that this leads to a quantile improvement. With Proposition 3, we have further shown that the divergence between the target and the current proposal, from which the target is constructed, can be controlled by a quantity that depends only on the weighting function w. This means that for any algorithm satisfying a divergence-decrease conditions at every step, divergences can be used to understand both steps of the algorithms, namely the construction of the target, and the construction of the next proposal. We illustrate this fact in Figure 1.

B. Analyzing the IGO algorithms with our framework

We now show that both IGO algorithms, namely Algorithms 1 and 2, proposed in [14], fall within our divergence-decrease framework, showing the applicability of our construction. Using our framework, we also show that the theoretical guarantees for Algorithm 1 in the case of proposals from an exponential family can be ensured for larger step-sizes than in [14]. This result is obtained with a different proof mechanism than in [15], by leveraging a variational inference result from [27] and thus shedding new lights on Algorithm 1.

Proposition 4. *Consider a sequence $\{\theta_k\}_{k \in \mathbb{N}}$ constructed either from Algorithm 1 or Algorithm 2. Then, at every iteration $k \in \mathbb{N}$, we have the following.*

(i) If Algorithm 1 is used, the proposal $p_{\theta_{k+1}}$ satisfies $KL(\pi_{\theta_k}^f, p_{\theta_{k+1}}) \leq KL(\pi_{\theta_k}^f, p_{\theta_k})$ for step-sizes $\tau > 0$ small enough, with equality if and only if $\theta_{k+1} = \theta_k$.

(ii) If Algorithm 1 is used and $\{p_\theta, \theta \in \Theta\}$ is an exponential family, we have (under some regularity assumptions detailed in the proof), that $KL(\pi_{\theta_k}^f, p_{\theta_{k+1}}) \leq KL(\pi_{\theta_k}^f, p_{\theta_k})$ for any step-size $\tau \in (0, 1/Z_w]$ with equality if and only if $\theta_{k+1} = \theta_k$.

(iii) If Algorithm 2 is used and the problem in (8) is uniquely maximized, then $KL(\pi_{\theta_k}^f, p_{\theta_{k+1}}) \leq KL(\pi_{\theta_k}, p_{\theta_k})$ for any step-size $\tau \in (0,1]$, with equality if and only $\theta_{k+1} = \theta_k$.

Proof. Let $k \in \mathbb{N}$.

(i) The update (7) in Algorithm 1 can be written as $\theta_{k+1} = \theta_k + \tau \int W_{\theta_k}^f(x) \widetilde{\nabla}_\theta \left(p_\theta(x)\right)_{|\theta=\theta_k} m(dx)$. Then, using $\widetilde{\nabla}_\theta \left(\ln p_\theta(x)\right)_{|\theta=\theta_k} = (1/p_{\theta_k}(x)) \widetilde{\nabla}_\theta \left(p_\theta(x)\right)_{|\theta=\theta_k}$, we obtain that (7) is equivalent to having

$$\theta_{k+1} = \theta_k + \tau \int W_{\theta_k}^f(x) p_{\theta_k}(x) \widetilde{\nabla}_\theta \left(\ln p_\theta(x)\right)_{|\theta=\theta_k} m(dx).$$

We can then notice that this is equivalent to performing $\theta_{k+1} = \theta_k - \tau Z_w \widetilde{\nabla}_\theta \left(KL(\pi_{\theta_k}^f, p_\theta) \right)_{|\theta=\theta_k}$, from which we deduce the result.

(ii) Consider that $\{q_\theta, \theta \in \Theta\}$ forms an exponential family with sufficient statistics Γ and log-partition function A with $\Theta = \operatorname{dom} A$. Assume that $\{q_\theta, \theta \in \Theta\}$ is minimal and steep (see [24]), that there exists $\theta \in \operatorname{int} \operatorname{dom} A$ such that $\mathbb{E}_{X \sim \pi_{\theta_k}^f}[\Gamma(X)] = \mathbb{E}_{X \sim p_\theta}[\Gamma(X)]$, and that for any $\theta \in \operatorname{dom} A$, $KL(\pi_{\theta_k}^f, p_\theta) < +\infty$. From [15, Equation (15)], the IGO update over an exponential family at iteration k reads

$$\eta_{k+1} = \eta_k + \tau \int \left(W_{\theta_k}^f(x)(\Gamma(x) - \eta_k) \right) p_{\theta_k}(x) dx,$$

where $\eta_k = \nabla A(\theta_k)$ and $\eta_{k+1} = \nabla A(\theta_{k+1})$, both well-defined under our assumptions. This is equivalent to $\eta_{k+1} = \eta_k + \tau Z_w \left(\pi_{\theta_k}^f(\Gamma) - \eta_k \right)$, which can be recognized to be equivalent using [27, Proposition 6] to

$$\eta_{k+1} = \eta_k - \tau Z_w \nabla_\theta KL(\pi_{\theta_k}^f, p_\theta)_{|\theta=\theta_k}.$$

This characterization ensures the result using [27, Proposition 14 (i) and (ii)], for the step sizes τ satisfying $\tau Z_w \in (0,1]$.

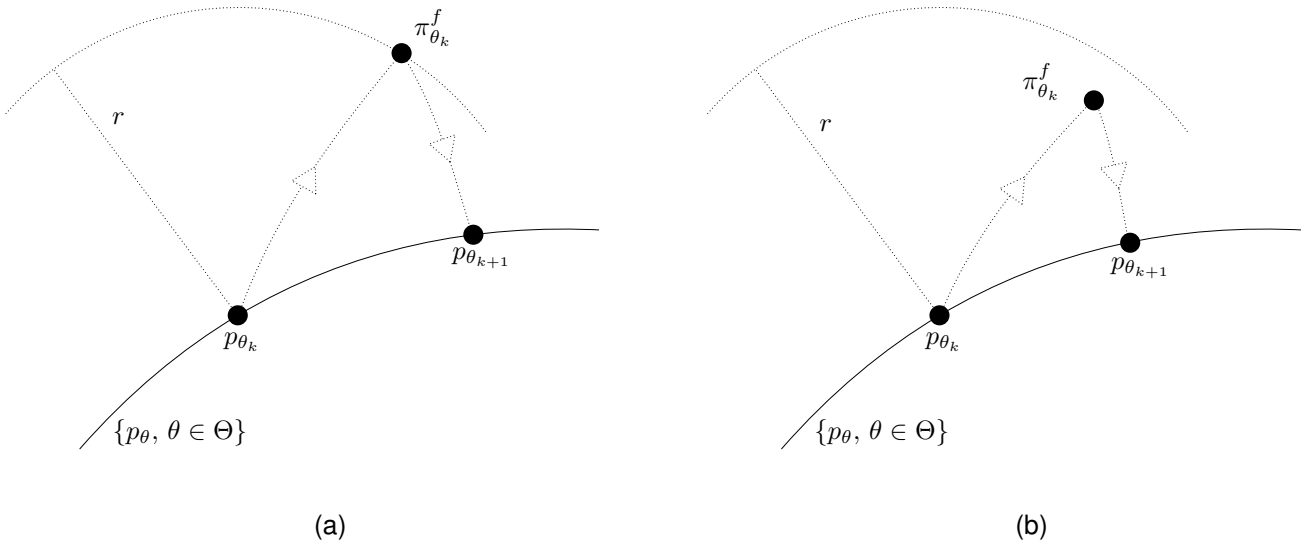

Fig. 1. A schematic view of one step of an algorithm covered by our divergence-based framework. Starting from a proposal p_{θ_k}, one first constructs the target $\pi^f_{\theta_k}$ following (9), i.e., a process that benefits from the results of Proposition 3. Then, one adapts $p_{\theta_{k+1}}$ such that a divergence-decrease (12), for some $\alpha \in [0,1)$, is achieved. (a) Case $\alpha \in (0,1)$ and $W^f_{\theta_k}$ taking values in $\{0,1\}$, corresponding to Proposition 2 and Proposition 3 (i) with $r = \frac{1}{\alpha(\alpha-1)}(Z_w^\alpha - 1)$. (b) Case $\alpha = 0$ and w taking values in $[0,1]$, corresponding to Proposition 1 and Proposition 3 (ii) with $r = -\ln Z_w$.

(iii) The IGO-ML update (8) can be rewritten as

$$\begin{aligned}\theta_{k+1} &= \arg\min_{\theta\in\Theta} \Bigg((1-\tau)\int \ln\left(\frac{1}{p_\theta(x)}\right) p_{\theta_k}(x)m(dx) \\ &\qquad + \tau \int \ln\left(\frac{1}{p_\theta(x)}\right) W^f_{\theta_k}(x) p_{\theta_k}(x)m(dx)\Bigg) \\ &= \arg\min_{\theta\in\Theta} \left((1-\tau)KL(p_{\theta_k}, p_\theta) + \tau Z_w KL(\pi_{\theta_k}, p_\theta)\right) \\ &= \arg\min_{\theta\in\Theta} \left(KL(\pi^f_{\theta_k}, p_\theta) + \frac{1-\tau}{\tau Z_w} KL(p_{\theta_k}, p_\theta)\right).\end{aligned}$$

We thus obtain by definition of θ_{k+1} that

$$KL(\pi^f_{\theta_k}, p_{\theta_{k+1}}) + \frac{1-\tau}{\tau Z_w} KL(p_{\theta_k}, p_{\theta_{k+1}}) \leq KL(\pi^f_{\theta_k}, p_{\theta_k}),$$

from which we obtain the result, using that $KL(p_{\theta_k}, p_{\theta_{k+1}}) = 0$ if and only if $\theta_{k+1} = \theta_k$. □

Proposition 4 shows that the IGO algorithms can be studied using our divergence-decrease condition with the KL divergence. For Algorithm 1, i.e., the IGO algorithm using natural gradients, we recover in Proposition 4 (i) that the increase (6), that is $J(\theta_{k+1}|\theta_k) > Z_w$, is guaranteed for infinitesimal step-sizes [14, Proposition 7]. Similarly, we recover in Proposition 4 (iii) a similar result for Algorithm 2, i.e., the maximum-likelihood IGO algorithm, for any step-size in $(0,1]$, which was established in [15, Theorem 6].

Proposition 4 (ii) establishes a new result for Algorithm 1 when the proposal form an exponential family. Indeed, the theoretical guarantees of Algorithm 1 over an exponential family were established in [15, Corollary 7] for step-sizes in $(0,1]$ by remarking that in this case, Algorithms 1 and 2 coincide. We do things differently in Proposition 4 (ii) by showing that Algorithm 1 is equivalent to a mirror descent update [28] on a KL divergence objective. These algorithms have been studied in our recent work [27], and we leverage these results to obtain a KL divergence decrease for Algorithm 1 for possibly larger step-sizes (if $w(u) = \delta_{u\leq q}(u)$ with $q \in (0,1)$, $1/Z_w = 1/q > 1$). Even if the weighting function can be normalized such that $Z_w = 1$, this still brings a new point of view on Algorithm 1, allowing to establish guarantees independently of Algorithm 2.

C. A new result on mixture-based methods with our framework

We now show how our proposed framework can be applied, for the study of black-box global optimization algorithms with mixture proposals. Such situation is challenging to analyze by sticking to the IGO framework, as mixture proposals do not form an exponential family nor yield a closed-form solution for the IGO-ML update (8) in Algorithm 2. We focus on a particular algorithm, linked both with the M-PMC algorithm of [29], a type of expectation-maximization (EM) algorithm proposed in the context of computational statistics, and with the mixture-based CE method proposed in [4, Example 3.2]. By exploiting the paradigm we introduced in Section III-A, we show that each iteration of the considered algorithm achieves a divergence-decrease, thus implying that the increase condition (6) is fulfilled. We can then apply Lemma 1 to establish a quantile improvement result.

Let us first introduce the algorithm we are going to consider, summarized in Algorithm 3. In this algorithm, the weight as well as the parameters of each component of the mixture are adapted at every iteration. We consider in the following mixture models with $J \in \mathbb{N}$ components $p_\theta = \sum_{j=1} \lambda^{(j)} p_{\vartheta^{(j)}}$ such that, for every $j \in \mathbb{N}$, $\lambda^{(j)} \in [0,1]$ and $p_{\vartheta^{(j)}} \in \{p_\vartheta,\, \vartheta \in \Theta\}$, with $\sum_{j=1}^J \lambda^{(j)} = 1$. We thus have the global parameters $\theta = (\{\lambda^{(j)}\}_{j=1}^J, \{\theta^{(j)}\}_{j=1}^J)$.

Algorithm 3 Mixture-based ML algorithm

Initialize the parameters $\vartheta_0^{(j)}$ and the mixture weights $\lambda_0^{(j)}$ for $j = 1, \dots, J$, and form the global parameter $\theta_0 = (\{\lambda_0^{(j)}\}_{j=1}^J, \{\vartheta_0^{(j)}\}_{j=1}^J)$.
for $k = 0, \dots$ **do**
For each $j = 1, \dots, J$, define the function $\rho_k^{(j)} : \mathbb{X} \to \mathbb{R}$ defined for any $x \in \mathbb{X}$ by

$$\rho_k^{(j)}(x) = \frac{\lambda_k^{(j)} p_{\vartheta_k^{(j)}}(x)}{\sum_{i=1}^J \lambda_k^{(i)} p_{\vartheta_k^{(i)}}(x)}. \tag{14}$$

Update $\theta_{k+1} = (\{\lambda_{k+1}^{(j)}\}_{j=1}^J, \{\vartheta_{k+1}^{(j)}\}_{j=1}^J)$ such that for every $j = 1, \dots, J$,

$$\lambda_{k+1}^{(j)} = \mathbb{E}_{X \sim \pi_{\theta_k}^f}[\rho_k^{(j)}(X)], \tag{15}$$

$$\vartheta_{k+1}^{(j)} = \arg\max_{\vartheta \in \Theta} \mathbb{E}_{X \sim \pi_{\theta_k}^f}[\ln p_\vartheta(X) \rho_k^{(j)}(X)]. \tag{16}$$

end for

Algorithm 3 shares links with the EM point of view adopted in [11]. In this work, several estimation of distribution algorithms [3] are shown to be EM algorithms with maximum likelihood steps that are reweighted using the objective to be minimized f (see also the fitness EM algorithm of [10] and the discussion in [8, Section 5.3]). Algorithm 3 recovers the M-PMC algorithm of [29], which is also an EM-like algorithm, with rank-based weights (see [14, Equation (14)] or Equation (19)) instead of importance weights. Note that contrary to [11] which does not explicitly consider mixture models, we do so here. Let us also remark that Algorithm 2 can also be linked to an EM, using a similar analysis.

We now show in Proposition 5 that Algorithm 3 achieves a decrease in terms of KL divergence at every iteration. Our proof techniques are reminiscent from the work [30] on variational inference. The result of Proposition 5 can then be used to apply Proposition 1 and Corollary 1 and get insights on the optimization performance of Algorithm 3.

Proposition 5. *Consider a sequence $\{\theta_k\}_{k \in \mathbb{N}}$ generated by Algorithm 3 with $\theta_k = (\{\lambda_k^{(j)}\}_{j=1}^J, \{\vartheta_k^{(j)}\}_{j=1}^J)$ for every $k \in \mathbb{N}$. At every iteration $k \in \mathbb{N}$, Algorithm 3 achieves the decrease*

$$KL\left(\pi_{\theta_k}^f, p_{\theta_{k+1}}\right) \leq KL\left(\pi_{\theta_k}^f, p_{\theta_k}\right), \tag{17}$$

with equality if and only if $\lambda_{k+1}^{(j)} = \lambda_k^{(j)}$ and $\vartheta_{k+1}^{(j)} = \vartheta_k^{(j)}$ for every $j = 1, \dots, J$, provided that the problem in (16) *is uniquely maximized.*

Proof. We adapt the ideas of the proof of [30, Theorem 2]. We compute the quantity

$$\begin{aligned}
&KL(\pi_{\theta_k}^f, p_{\theta_{k+1}}) - KL(\pi_{\theta_k}^f, p_{\theta_k}) \\
&\quad = \int -\ln\left(\frac{\sum_{j=1}^J \lambda_{k+1}^{(j)} p_{\vartheta_{k+1}^{(j)}}}{\sum_{i=1}^J \lambda_k^{(i)} p_{\vartheta_k^{(i)}}}\right) \pi_{\theta_k}^f(x) m(dx) \\
&\quad = \int -\ln\left(\sum_{j=1}^J \rho_k^{(j)}(x) \frac{\lambda_{k+1}^{(j)} p_{\vartheta_{k+1}^{(j)}}(x)}{\lambda_k^{(j)} p_{\vartheta_k^{(j)}}(x)}\right) \pi_{\theta_k}^f(x) m(dx) \\
&\quad \leq \int -\sum_{j=1}^J \rho_k^{(j)}(x) \ln\left(\frac{\lambda_{k+1}^{(j)} p_{\vartheta_{k+1}^{(j)}}(x)}{\lambda_k^{(j)} p_{\vartheta_k^{(j)}}(x)}\right) \pi_{\theta_k}^f(x) m(dx) \\
&\quad = \int -\sum_{j=1}^J \rho_k^{(j)}(x) \ln\left(\frac{\lambda_{k+1}^{(j)} p_{\vartheta_{k+1}^{(j)}}(x)}{\lambda_k^{(j)} p_{\vartheta_k^{(j)}}(x)}\right) \pi_{\theta_k}^f(x) m(dx)
\end{aligned}$$

using Jensen's inequality and that the $\rho_k^{(j)}(x)$ sum to one for any $x \in \mathbb{X}$. We can then decompose the above quantity into two terms, namely,

$$\begin{aligned}
&\int -\sum_{j=1}^J \rho_k^{(j)}(x) \ln\left(\frac{\lambda_{k+1}^{(j)} p_{\vartheta_{k+1}^{(j)}}(x)}{\lambda_k^{(j)} p_{\vartheta_k^{(j)}}(x)}\right) \pi_{\theta_k}^f(x) m(dx) \\
&\quad = -\sum_{j=1}^J \ln\left(\frac{\lambda_{k+1}^{(j)}}{\lambda_k^{(j)}}\right) \int \rho_k^{(j)}(x) \pi_{\theta_k}^f(x) m(dx) \\
&\quad + \sum_{j=1}^J \int \rho_k^{(j)}(x) \ln\left(\frac{p_{\vartheta_k}^{(j)}(x)}{p_{\vartheta_{k+1}}^{(j)}(x)}\right) \pi_{\theta_k}^f(x) m(dx).
\end{aligned} \tag{18}$$

Due to the definition of λ_{k+1}, given in Equation (15), that is $\lambda_{k+1}^{(j)} = \int \rho_k^{(j)}(x) \pi_{\theta_k}^f(x) m(dx)$, the first term in the right-hand side of Equation (18) is equal to $-\sum_{j=1}^J \ln\left(\frac{\lambda_{k+1}^{(j)}}{\lambda_k^{(j)}}\right) \lambda_{k+1}^{(j)}$ which is non-positive from Jensen's inequality, being null if and only if $\lambda_k = \lambda_{k+1}$. The second term in the right-hand side of (18) is a sum of J terms, each being non-negative from the definition of $\vartheta_{k+1}^{(j)}$ given in Equation (16). Each term is zero if and only if $\vartheta_{k+1}^{(j)} = \vartheta_k^{(j)}$, due to our assumption that each maximization problem of the form (16) is uniquely maximized. We have thus shown the decrease (17) and that there is equality if and only if $\lambda_{k+1} = \lambda_k$ and $\theta_{k+1} = \theta_k$. □

The result of Proposition 5 is to our knowledge the first to establish convergence guarantees of this kind for black-box global optimization algorithms that are explicitly mixture-based. Indeed, mixtures were not explicitly considered in [14], [15], and they often do not admit closed-form solutions for the maximization problem (8) in Algorithm 2. The solution is usually to perform EM-like updates, as it is done in [4, Example 3.2] for instance, which can be handled with our divergence-decrease condition. Many variational inference methods explicitly consider mixtures, see for instance [29]–[31], showing the potential for further links between black-box global optimization with mixture models and variational inference. Let us also remark that compared to more complex mixture-based algorithms, such as the ones proposed in [18],

[19], [21], whose convergence has only been verified empirically, Algorithm 3 has a fixed number of mixture components.

In order to describe how Algorithm 3 is related to the CE algorithm of [4, Example 3.2], we first discuss the Monte Carlo approximation of Algorithm 3 and explicit its link with the CE algorithm given in [4, Example 3.2]. In order to implement an iteration $k \in \mathbb{N}$ of Algorithm 3, one can approximate the integrals with respect to π_{θ_k} using samples and rank-based weights, as shown in [14]. To do so, we first sample N points $x_{k,n}$, $n = 1, \dots, N$ from the mixture distribution $p_{\theta_k} = \sum_{j=1}^{J} \lambda_k^{(j)} q_{\vartheta_k^{(j)}}$. This is done by drawing a component j with probability $\lambda_k^{(j)}$ via multinomial sampling, and then drawing from the corresponding component. Each sample $x_{k,n}$ receives a rank-based weight $\widehat{\omega}_{k,n}$ defined as in [14, Equation (14)] by

$$\widehat{\omega}_{k,n} = \frac{1}{N} w \left(\frac{\mathrm{rank}(x_{k,n}) + 1/2}{N} \right), \tag{19}$$

where $\mathrm{rank}(x_{k,n})$ is the number of samples in $\{x_{k,n}\}_{n=1}^N$ with value of f strictly less than $f(x_{k,n})$. Then, if h is such that $\mathbb{E}_{X \sim p_{\theta_k}}[h(X)^2] < +\infty$, we have from [14, Proposition 27] that, conditioned on θ_k,

$$\sum_{n=1}^{N} \widehat{\omega}_{k,n} h(x_{k,n}) \xrightarrow[N \to +\infty]{a.s.} \mathbb{E}_{X \sim p_{\theta_k}}[W_{\theta_k}(X) h(X)].$$

In particular, $\sum_{n=1}^{N} \widehat{\omega}_{k,n} \xrightarrow[N \to +\infty]{a.s.} Z_w$, meaning that the weights $\{\widehat{\omega}_{k,n}\}_{1 \leq n \leq N}$ can be self-normalized.

Now, consider the Monte Carlo approximation of Algorithm 3, but with $\rho_k^{(j)}(x_{k,n})$ being replaced by $\xi_{k,n}^{(j)}$. Consider a random variable $X \sim p_{\theta_k}$, with the latent variable $\xi_{k,n}^{(j)}$ which is equal to 1 if $x_{k,n}$ has been sampled from the component j of the mixture and zero otherwise.. Then, $\mathbb{P}[\xi_{k,n}^{(j)} = 1] = \lambda_k^{(j)}$ and $\mathbb{E}_{X \sim p_{\theta_k}}[h(X) | \xi_{k,n}^{(j)} = 1] = \mathbb{E}_{X \sim p_{\vartheta_k^{(j)}}}[h(X)]$. We can thus check that in the large number of samples limit and conditioned on θ_k,

$$\sum_{n=1}^{N} \widehat{\omega}_{k,n} \xi_{k,n}^{(j)} h(x_{k,n}) \xrightarrow[N \to +\infty]{a.s.} \lambda_k^{(j)} \mathbb{E}_{X \sim p_{\vartheta_k^{(j)}}}[W_{\theta_k}^f(X) h(X)].$$

We thus obtain the algorithm outlined in [29, Equation (12)] with rank-based weights instead of importance weights. If further, $w(u) = \delta_{u \leq q}(u)$ and the components of the mixture are Gaussians, we obtain the algorithm of [4, Example 3.2]. Since $\rho_k^{(j)}(x_{k,n}) = \mathbb{E}[\xi_{k,n}^{(j)} | x_{k,n}]$, using $\rho_k^{(j)}(x_{k,n})$ instead of $\xi_{k,n}$ amounts to a Rao-Blackwellized (meaning that a random variable is replaced by its conditional expectation [29]) version of the algorithm from [4, Example 3.2]. This procedure does not entail additional evaluations of the objective f, while providing better numerical stability [29], as all the components of the mixtures appear in every update.

D. A new result for heavy-tailed proposals with our framework

We now apply our theoretical tools to study a black-box global optimization algorithm with proposals being Student distributions with a fixed degree of freedom parameter $\nu > 0$. Specifically, we propose an algorithm to update the location and scale parameters of the proposals at every iteration, and show that it satisfies our divergence-decrease conditions.

We consider Student distributions in dimension d with $\nu > 0$ degrees of freedom indexed by their location parameters $\mu \in \mathbb{R}^d$ and scale parameters $\Sigma \in \mathcal{S}_{++}^d$, the set of positive definite matrices in dimensions d. The density with respect to the Lebesgue measure of the Student distribution $\mathcal{T}(\cdot; \mu, \Sigma, \nu)$ is defined for all $x \in \mathbb{R}^d$ by

$$\mathcal{T}(x; \mu, \Sigma, \nu) \propto \left(1 + \frac{1}{\nu} (x - \mu)^\top \Sigma^{-1} (x - \mu) \right)^{-\frac{\nu + d}{2}} \tag{20}$$

with normalization constant being equal to $\frac{\Gamma(\nu/2)}{\Gamma((\nu+d)/2)} (\nu^d \pi^d \det(\Sigma))^{1/2}$, Γ denoting the Gamma function and $\det$ the determinant. When $\nu = 1$, the Cauchy distributions are recovered, while normal distributions are recovered in the limit $\nu \to +\infty$. Alternatively, the density in (20) can be written as the continuous mixture

$$\mathcal{T}(x; \mu, \Sigma, \nu) = \int_0^{+\infty} \mathcal{N}\left(x; \mu, \frac{1}{z} \Sigma \right) \mathcal{G}\left(z; \frac{\nu}{2}, \frac{\nu}{2} \right) dz, \tag{21}$$

where the latent variable Z is distributed following the Gamma distribution with parameters $(\frac{\nu}{2}, \frac{\nu}{2})$ and probability density $\mathcal{G}(z; \frac{\nu}{2}, \frac{\nu}{2})$ for any $z \in (0, +\infty)$. Conditionally on Z, X follows a normal distribution with mean μ and covariance $\frac{1}{Z}\Sigma$, and density $\mathcal{N}(x; \mu, \frac{1}{z}\Sigma)$ for any $x \in \mathbb{R}^d$. We will use the point of view from (21) in the following. We fix $\nu > 0$, and consider parameters $\theta = (\mu, \Sigma)$ with associated densities $p_\theta = \mathcal{T}(\cdot; \mu, \Sigma, \nu)$. In this context, we propose the heavy-tailed black-box global optimization algorithm, summarized in Algorithm 4.

Algorithm 4 Heavy-tail ML algorithm

Initialize the parameters $\theta_0 = (\mu_0, \Sigma_0)$ and choose the degree of freedom parameter $\nu > 0$.

for $k = 0, \dots$ **do**

Define the function $\gamma_k^{(\nu)} : \mathbb{X} \to \mathbb{R}$ defined for any $x \in \mathbb{X}$ by

$$\gamma_k^{(\nu)}(x) = \frac{\nu + d}{\nu + (x - \mu_k)^\top \Sigma^{-1} (x - \mu_k)}. \tag{22}$$

Update $\theta_{k+1} = (\mu_{k+1}, \Sigma_{k+1})$ such that

$$\mu_{k+1} = \frac{\mathbb{E}_{X \sim \pi_{\theta_k}^f}[\gamma_k^{(\nu)}(X) X]}{\mathbb{E}_{X \sim \pi_{\theta_k}^f}[\gamma_k^{(\nu)}(X)]}, \tag{23}$$

$$\Sigma_{k+1} = \frac{\mathbb{E}_{X \sim \pi_{\theta_k}^f}[\gamma_k^{(\nu)}(X) X X^\top]}{\mathbb{E}_{X \sim \pi_{\theta_k}^f}[\gamma_k^{(\nu)}(X)]} - \mu_{k+1} \mu_{k+1}^\top. \tag{24}$$

end for

When the degree of freedom parameter ν goes to infinity, we have that $\mathcal{T}(x; \mu, \Sigma, \nu) \to \mathcal{N}(x; \mu, \Sigma)$ for any $x \in \mathbb{R}^d$, meaning that Student distributions recover the normal distributions. Moreover, we have at any iteration $k \in \mathbb{N}$ that $\gamma_k^{(\nu)}(x) \to 1$ when $\nu \to +\infty$. In this case, the updates (23) and (24) in Algorithm 4 recover the updates of Algorithm 2 with step size $\tau = 1$ when Gaussian distributions are

used. Moreover, evaluating the function $\gamma_k^{(\nu)}$ does not imply a heavy computational burden, as it does not involve additional computations of the objective function f.

We now show that Algorithm 4 achieves our divergence-decrease condition. This means that the improvement (6) is satisfied at every iteration, and thus that one can apply Lemma 1 to get quantile improvement when $w(u) = \delta_{u \leq q}(u)$ is used.

Proposition 6. *Consider a sequence $\{\theta_k\}_{k\in\mathbb{N}}$ generated by Algorithm 4. At every iteration $k \in \mathbb{N}$, we have the decrease*

$$KL(\pi^f_{\theta_k}, p_{\theta_{k+1}}) \leq KL(\pi^f_{\theta_k}, p_{\theta_k}), \tag{25}$$

with equality if and only if $\theta_{k+1} = \theta_k$.

Proof. Consider any $\theta \in \Theta$ and any distribution p over the optimization variables $x \in \mathbb{R}^d$ and the latent variables $z \in (0, +\infty)$. We then have

$$\begin{aligned}
&\int \ln p_\theta(x) \pi^f_{\theta_k}(x) dx \\
&\quad = \iint \ln p_\theta(x) p(z|x) dz\, \pi^f_{\theta_k}(x) dx \\
&\quad = \iint \ln \left(\frac{p_\theta(x,z)}{p_\theta(z|x)} \right) p(z|x) dz\, \pi^f_{\theta_k}(x) dx \\
&\quad = \iint \ln \left(\frac{p_\theta(x,z)}{p(z|x)} \right) p(z|x) dz\, \pi^f_{\theta_k}(x) dx \\
&\qquad - \iint \ln \left(\frac{p_\theta(z|x)}{p(z|x)} \right) p(z|x) dz\, \pi^f_{\theta_k}(x) dx \\
&\quad = \iint \ln \left(\frac{p_\theta(x,z)}{p(z|x)} \right) p(z|x) dz\, \pi^f_{\theta_k}(x) dx \\
&\qquad + \int KL(p(\cdot|x), p_\theta(\cdot|x)) \pi^f_{\theta_k}(x) dx.
\end{aligned}$$

Hence, we have that

$$\begin{aligned}
&\int \ln p_\theta(x) \pi^f_{\theta_k}(x) dx \\
&\qquad \geq \iint \ln \left(\frac{p_\theta(x,z)}{p(z|x)} \right) p(z|x) dz\, \pi^f_{\theta_k}(x) dx,
\end{aligned} \tag{26}$$

with equality if and only if $p_\theta(z|x) = p(z|x)$ for any $z \in (0, +\infty)$ and $x \in \mathbb{R}^d$.

We now compute the gap in Kullback-Leibler divergence and using Equation (26), we obtain

$$\begin{aligned}
&KL(\pi^f_{\theta_k}, p_{\theta_{k+1}}) - KL(\pi^f_{\theta_k}, p_{\theta_k}) \\
&\quad = -\int \ln p_{\theta_{k+1}}(x) \pi^f_{\theta_k}(x) dx + \int \ln p_{\theta_k}(x) \pi^f_{\theta_k}(x) dx \\
&\quad \leq -\iint \ln \left(\frac{p_{\theta_{k+1}}(x,z)}{p_{\theta_k}(z|x)} \right) p_{\theta_k}(z|x) dz\, \pi^f_{\theta_k}(x) dx \\
&\qquad + \iint \ln \left(\frac{p_{\theta_k}(x,z)}{p_{\theta_k}(z|x)} \right) p_{\theta_k}(z|x) dz\, \pi^f_{\theta_k}(x) dx \\
&\quad = -\iint \ln p_{\theta_{k+1}}(x,z) p_{\theta_k}(z|x) dz\, \pi^f_{\theta_k}(x) dx \\
&\qquad + \iint \ln p_{\theta_k}(x,z) p_{\theta_k}(z|x) dz\, \pi^f_{\theta_k}(x) dx.
\end{aligned}$$

Since the degree of freedom parameter is kept constant, we have that $p_\theta(x,z) = p_\theta(x|z)p(z)$, with $p_\theta(x|z) = \mathcal{N}(x; \mu, \frac{1}{z}\Sigma)$ and $p(z) = \Gamma(z; \frac{\nu}{2}, \frac{\nu}{2})$ that does not depend on θ. In particular, we can write

$$\begin{aligned}
&KL(\pi^f_{\theta_k}, p_{\theta_{k+1}}) - KL(\pi^f_{\theta_k}, p_{\theta_k}) \\
&\qquad \leq -\iint \ln p_{\theta_{k+1}}(x|z) p_{\theta_k}(z|x) dz\, \pi^f_{\theta_k}(x) dx \\
&\qquad\qquad + \iint \ln p_{\theta_k}(x|z) p_{\theta_k}(z|x) dz\, \pi^f_{\theta_k}(x) dx.
\end{aligned}$$

Therefore, showing that $\theta_{k+1} = (\mu_{k+1}, \Sigma_{k+1})$ is such that

$$\theta_{k+1} = \arg\max_{\theta\in\Theta} \iint \ln p_\theta(x|z) p_{\theta_k}(z|x) dz\, \pi^f_{\theta_k}(x) dx, \tag{27}$$

establishes the decrease in Equation (25) with equality if and only if $\theta_{k+1} = \theta_k$. We now show that $\theta_{k+1} = (\mu_{k+1}, \Sigma_{k+1})$ as constructed in Algorithm 4 satisfies (27).

For any $\theta \in \Theta$, we have that $p_\theta(x|z) = \mathcal{N}(x; \mu, \frac{1}{z}\Sigma)$. Hence, we can compute that

$$\begin{aligned}
&\iint \ln p_\theta(x|z) p_{\theta_k}(z|x) dz\, \pi^f_{\theta_k}(x) dx = -\ln\det(\Sigma) \\
&\quad - \frac{1}{2} \iint z p_{\theta_k}(z|x) dz\, (x-\mu)^\top \Sigma^{-1} (x-\mu) \pi^f_{\theta_k}(x) dx.
\end{aligned}$$

For any $x \in \mathbb{R}^d$, one can check that $p_{\theta_k}(\cdot|x)$ is the density of a Gamma distribution with parameters $\left(\frac{\nu+d}{2}, \frac{1}{2}(\nu + (x-\mu_k)\Sigma_k^{-1}(x-\mu_k))\right)$. We then remark that $\gamma_k^{(\nu)}$ as defined in (22) satisfies for any $x \in \mathbb{R}^d$

$$\gamma_k^{(\nu)}(x) = \int z p_{\theta_k}(z|x) dz$$

and thus get an explicit expression for the objective in (27) of the form

$$\begin{aligned}
&\int \ln p_\theta(x|z) p_{\theta_k}(z|x) dz\, \pi^f_{\theta_k}(x) m(dx) = -\ln\det(\Sigma) \\
&\qquad - \frac{1}{2} \mathbb{E}_{X\sim\pi^f_{\theta_k}} [\gamma_k^{(\nu)}(X)(X-\mu)^\top \Sigma^{-1}(X-\mu)],
\end{aligned}$$

from which the result follows. □

The result of our Proposition 6 allows to give theoretical guarantees for heavy-tailed distributions, that do not form an exponential family. In particular, they apply for any Student family, including Cauchy distributions when $\nu = 1$, and to Gaussian distributions in the limit $\nu \to +\infty$. It has been shown in [32] that Cauchy proposals perform better than Gaussian proposals in low dimension, while the reverse is true when the dimension of the problem grows. Our algorithm allows to interpolate these two regimes, possibly opening the way to a tail-adaptive algorithm able to select good values of the degree of freedom parameter as well.

IV. Discussion and perspectives

We have proposed in this work divergence-based conditions that imply the quantile improvement results achieved by the IGO framework. Therefore, to prove that an algorithm achieves a quantile improvement result, one can show that this algorithm fits in the IGO framework, or that it satisfies our divergence-based conditions. We have shown that the IGO algorithms satisfy our divergence-based condition, meaning

that we have provided broader conditions to establish quantile improvement results.

We used our divergence-based conditions to study the IGO framework, showing that it satisfies our divergence-based conditions, but also using our conditions to get a new proof of the quantile improvement for the IGO algorithm based on natural gradients for larger step-sizes. We also applied our divergence-based conditions to study two other algorithms, namely a mixture-based algorithm and one with heavy-tailed proposals, obtaining a new quantile improvement result for both these cases which were difficult to study within the IGO framework. The first algorithm is related to the mixture-based version of the CE algorithm, but is supposed to be more stable across iterations, without increasing the required number of objective evaluations. The second algorithm uses Student proposals with arbitrary degree of freedom parameter. This includes Cauchy distributions, but also other Student distributions that may yield better performance depending on the problem at hand and its dimension.

In order to show that existing black-box global optimization algorithms satisfy our divergence-decrease conditions, we leveraged existing results from statistics and machine learning. Indeed, our conditions require to show that the divergence between the proposal and the proposed target decreases. Many statistics and machine learning algorithms have actually been proposed to minimize divergences, especially in the field of variational inference. This connection between the two fields opens new perspectives for the design and study of black-box global optimization algorithms. Moreover, recent variational inference algorithms handle complex proposal distributions and sophisticated adaptation strategies which could in turn lead to improved optimization performance and better understanding of global optimization methods.

REFERENCES

[1] L. M. Rios and N. V. Sahinidis, "Derivative-free optimization: a review of algorithms and comparison of software implementations," *Journal of Global Optimization*, vol. 56, pp. 1247–1293, 2013.

[2] N. Hansen, D. V. Arnold, and A. Auger, *Evolution strategies*. Springer, 2015.

[3] P. Larrañaga, *A Review on Estimation of Distribution Algorithms*. Springer, 2002.

[4] D. Kroese, S. Porotsky, and R. Rubinstein, "The cross-entropy method for continuous multi-extremal optimization," *Methodology and Computing in Applied Probability*, vol. 8, pp. 383–407, 2006.

[5] N. Hansen and A. Ostermeier, "Completely derandomized self-adaptation in evolution strategies," *Evolutionary Computation*, vol. 9, no. 2, pp. 159–195, 2001.

[6] N. Hansen, "The CMA evolution strategy: A tutorial," 2023, https://arxiv.org/abs/1604.00772.

[7] T. Glasmachers, T. Schaul, S. Yi, D. Wierstra, and J. Schmidhuber, "Exponential natural evolution strategies," in *Proceedings of the Conference on Genetic and Evolutionary Computation (GECCO)*, 2010, pp. 393–400.

[8] Y. Akimoto, Y. Nagata, I. Ono, and S. Kobayashi, "Theoretical foundation for CMA-ES from information geometry perspective," *Algorithmica*, vol. 64, no. 4, pp. 698–716, 2012.

[9] L. Malagò and G. Pistone, "Information geometry of the Gaussian distribution in view of stochastic optimization," in *Proceedings of the Conference on Foundations of Genetic Algorithms (FOGA)*, 2015, pp. 150–162.

[10] D. Wierstra, T. Schaul, J. Peters, and J. Schmidhuber, "Fitness expectation-maximization," in *Proceedings of the Conference on Parallel Problem Solving from Nature (PPSN)*, 2008, pp. 337–346.

[11] D. Brookes, A. Busia, C. Fannjiang, K. Murphy, and J. Listgarten, "A view of estimation of distribution algorithms through the lens of expectation-maximization," in *Proceedings of the Genetic and Evolutionary Computation Conference Companion (GECCO)*, 2022, pp. 189–190.

[12] N. Hansen, R. Ros, N. Mauny, M. Schoenauer, and A. Auger, "Impacts of invariance in search: When CMA-ES and PSO face ill-conditioned and non-separable problems," *Applied Soft Computing*, vol. 11, no. 8, pp. 5755–5769, 2011.

[13] H.-G. Beyer, "Convergence analysis of evolutionary algorithms that are based on the paradigm of information geometry," *Evolutionary Computation*, vol. 22, no. 4, pp. 679–709, 2014.

[14] Y. Ollivier, L. Arnold, A. Auger, and N. Hansen, "Information-geometric optimization algorithms: a unifying picture via invariance principles," *Journal of Machine Learning Research*, vol. 18, no. 18, pp. 1–65, 2017.

[15] Y. Akimoto and Y. Ollivier, "Objective improvement in information-geometric optimization," in *Proceedings of the Conference on Foundations of Genetic Algorithms (FOGA)*, 2013, pp. 1–10.

[16] Y. Akimoto, A. Auger, and N. Hansen, "Convergence of the continuous time trajectories of isotropic evolution strategies on monotonic C^2-composite functions," in *Proceedings of the Conference on Parallel Problem Solving from Nature (PPSN)*, 2012, pp. 42–51.

[17] T. Schaul, T. Glasmachers, and J. Schmidhuber, "High dimensions and heavy tails for natural evolution strategies," in *Proceedings of the Conference on Genetic and Evolutionary Computation (GECCO)*, 2011, pp. 845–852.

[18] S. C. Maree, T. Alderliesten, D. Thierens, and P. A. N. Bosman, "Niching an estimation-of-distribution algorithm by hierarchical Gaussian mixture learning," in *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO)*, 2017, pp. 713–720.

[19] A. Ahrari, K. Deb, and M. Preuss, "Multimodal optimization by covariance matrix self-adaptation evolution strategy with repelling subpopulations," *Evolutionary Computation*, vol. 25, no. 3, pp. 439–471, 2017.

[20] X. He, Z. Zheng, and Y. Zhou, "MMES: Mixture model-based evolution strategy for large-scale optimization," *IEEE Transactions on Evolutionary Computation*, vol. 25, no. 2, pp. 320–333, 2021.

[21] A. Ahrari, S. Elsayed, R. Sarker, D. Essam, and C. A. C. Coelo, "Static and dynamic multimodal optimization by improved covariance matrix self-adaptation evolution strategy with repelling subpopulations," *IEEE Transactions on Evolutionary Computation*, vol. 26, no. 3, pp. 527–541, 2022.

[22] D. Blei, A. Kucukelbir, and J. McAuliffe, "Variational inference: A review for the statistician," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[23] C. Zhang, J. Bütepage, H. Kjellström, and S. Mandt, "Advances in variational inference," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 2008–2026, 2019.

[24] O. Barndorff-Nielsen, *Information and Exponential Families in Statistical Theory*. John Wiley & Sons, Ltd, 2014.

[25] Y. Polyanskiy and Y. Wu, "Information theory: From coding to learning," 2023, https://people.lids.mit.edu/yp/homepage/papers.html.

[26] A. Cichocki and S.-I. Amari, "Families of alpha- beta- and gamma-divergences: Flexible and robust measures of similarities," *Entropy*, vol. 12, no. 6, pp. 1532–1568, 2010.

[27] T. Guilmeau, E. Chouzenoux, and V. Elvira, "Regularized Rényi divergence minimization through Bregman proximal gradient algorithms," 2022, https://arxiv.org/abs/2211.04776.

[28] A. Beck and M. Teboulle, "Mirror descent and nonlinear projected subgradient optimization," *Operations Research Letters*, vol. 31, no. 3, pp. 167–175, 2003.

[29] O. Cappé, R. Douc, A. Guillin, J. M. Marin, and C. P. Robert, "Adaptive importance sampling in general mixture classes," *Statistics and Computing*, vol. 18, pp. 447–459, 2008.

[30] K. Daudel, R. Douc, and F. Roueff, "Monotonic alpha-divergence minimisation for variational inference," *Journal of Machine Learning Research*, vol. 24, no. 62, pp. 1–76, 2023.

[31] W. Lin, M. E. Khan, and M. Schmidt, "Fast and simple natural-gradient variational inference with mixture of exponential-family approximations," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2019, pp. 3992–4002.

[32] M. L. Sanyang, R. J. Durant, and A. Kabán, "How effective is Cauchy-EDA in high dimensions?" in *Proceedings of the IEEE Congress on Evolutionary Computation (CEC)*, 2016, pp. 3409–3416.

BIOGRAPHY SECTION

Thomas Guilmeau received an engineering degree from ENSTA Paris, Palaiseau, France, in 2020, and an MSc in applied mathematics from Université Paris-Saclay, Gif-sur-Yvette, France, in 2020. He is now working towards a doctoral degree at INRIA Saclay in project team OPIS, in Centre pour la Vision Numérique, CentraleSupélec. His research interests are in optimization, Monte Carlo methods, and their interactions.

Emilie Chouzenoux (IEEE Senior Member) received the engineering degree from Ecole Centrale, Nantes, France, in 2007, and the Ph.D. degree in signal processing from the Institut de Recherche en Communications et Cybernétique (IRCCyN, UMR CNRS 6597), Nantes, in 2010. Between 2011 and 2019, she was a Maître de conferences at the University of Paris-Est Marne-la-Vallée, Champs-sur-Marne, France (LIGM, UMR CNRS 8049). Since September 2019, she has been a Researcher at Inria Saclay, within the project team OPIS, in Centre pour la Vision Numérique, CentraleSupélec. She has been promoted as a Research Director in October 2023. Since May 2023, she is the group leader of OPIS Inria Saclay entity. She is an Associated Editor of SIAM Journal on Imaging Sciences, and of SIAM Journal on Data Sciences. Since January 2020, she has been the PI of the ERC Starting Grant MAJORIS. Her research interests are in large scale optimization algorithms for inverse problems and machine learning problems of image processing.

Víctor Elvira (IEEE Senior Member) received his Ph.D. degree in statistical signal processing in 2011 from the University of Cantabria (Spain). He is currently a Professor with a Personal Chair of Statistics and Data Science at the School of Mathematics at the University of Edinburgh (UK). From 2016 to 2019, he was an Associate Professor at the engineering school IMT Lille Douai (France). From 2013 to 2016, he was an Assistant Professor at University Carlos III of Madrid (Spain). He has also been a visiting researcher at several institutions such as Stony Brook University of New York (USA) and Paris-Dauphine University (France). Prof. Elvira's has co-authored more than 120 journal and refereed conference papers. His research interests are mostly in the fields of computational statistics, statistical signal processing, and machine learning, in particular in Bayesian inference and Monte Carlo methods with different applications including sensor networks, wireless communications, target tracking, ecology, and biomedicine. He is a Fulbright Fellow, Marie Curie Fellow, Leverhulme Fellow, and Alan Turing Fellow.