



**HAL**  
open science

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

Thomas Guilmeau, Emilie Chouzenoux, Víctor Elvira

► **To cite this version:**

Thomas Guilmeau, Emilie Chouzenoux, Víctor Elvira. A divergence-based condition to ensure quantile improvement in black-box global optimization. 2024. hal-04616771

**HAL Id: hal-04616771**

**<https://inria.hal.science/hal-04616771>**

Preprint submitted on 19 Jun 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.





Distributed under a Creative Commons Attribution 4.0 International License

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


Thomas Guilmeau<sup>1,a</sup>, Emilie Chouzenoux<sup>1,b</sup>, and Víctor Elvira<sup>2</sup>

<sup>1</sup>Université Paris-Saclay, CentraleSupélec, INRIA, CVN, France

<sup>a</sup> [thomas.guilmeau@inria.fr](mailto:thomas.guilmeau@inria.fr) 

<sup>b</sup> [emilie.chouzenoux@centralesupelec.fr](mailto:emilie.chouzenoux@centralesupelec.fr) 

<sup>2</sup>School of Mathematics, University of Edinburgh, United Kingdom

[victor.elvira@ed.ac.uk](mailto:victor.elvira@ed.ac.uk) 

## Abstract

Black-box global optimization aims at minimizing 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 theoretical study of these methods remains difficult. In this work, we introduce a new framework, based on divergence-decrease conditions, to study and design black-box global optimization algorithms. Our approach allows to establish and quantify the improvement of proposals at each iteration, in terms of expected value or quantile of the objective. We show that the information-geometric optimization approach fits within our framework, yielding a new approach for its analysis. We also establish proposal improvement results for two novel algorithms, one related with the cross-entropy approach with mixture models, and another one using heavy-tailed sampling proposal distributions.

**Keywords.** Black-box optimization, Variational inference, Mixture models, Heavy-tailed distributions, Kullback-Leibler divergence.

## 1 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 [29]. 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 [17], typically Gaussians, to generate samples close to the minimizers of the objective. Among these methods, one can mention the class of estimation-of-distribution algorithms [24], the cross-entropy algorithm [23], or the CMA-ES algorithm [18, 16].

A useful perspective to gain theoretical insights is to understand these algorithms as optimization schemes aiming at minimizing an expectation-based reformulation of the original problem over a set of proposals. The resulting reformulated problem can consist in minimizing the expected objective value [14, 4, 25], or the expected transformed objective value, for some well-chosen transformation [34, 33, 9]. These transformations can also be rank-based [34, 33]. Rank-based transforms only require a ranking of solutions and thus preserve

invariance properties with respect to monotonic transformation of the objective function. Algorithms with such invariance properties behave identically on two problems with the same ranking of solutions, allowing to generalize insights from one problem to another [21]. Invariance properties usually yield better-performing algorithms [19, 8]. Rank-based transformations rely on reformulations that depend on the retained proposal. In the infinite-population limit, a quantile-based reformulation of the objective function is obtained [27].

In order to solve the reformulated optimization problem over a set of proposals, many algorithms resort to natural gradient updates. These updates have been proposed in [6], and consist in a gradient descent step preconditioned by the Fisher information matrix of the proposal. Natural gradients have been used in the context of estimation-of-distribution algorithms [25, 9], evolution strategies [4, 33], or discrete permutation problems [12]. Natural gradients yield invariance properties with respect to parametrization of the proposals [33, 27]. They are straightforward to be computed when the proposal lies in the exponential family [22]. The natural gradient descent has been used jointly with a (rank-based) quantile-based reformulation of the objective in [27, 5], leading to the so-called information-geometric optimization (IGO) framework. IGO recovers many existing algorithms, such as the cross-entropy (CE) algorithm [23], also based on quantiles, as well as various estimation-of-distribution algorithms [24].

The theoretical study of the aforementioned natural gradient methods mostly follows two main approaches. The first one is to establish asymptotic convergence of the proposals to a limit proposal well-suited to solve the original problem. This is the approach of [3, 8] showing the convergence of Gaussian proposals used in IGO. Note that these results are established in an infinite sample size regime, and for infinitesimally small step sizes, amounting to continuous time. Their applicability in practical implementations, characterized by non-zero step sizes and stochastic errors, is up to our knowledge still an open problem. The second approach consists in proving an improvement on the reformulated optimization problem at every iteration. Such results are useful in practice as they hold without having to wait for an eventual asymptotic regime to be reached. This is done in [4], in the case of expected objective value minimization over Gaussian proposals, and in [27, 5] for the IGO reformulation of the original problem, although assuming infinitesimally small step-sizes, or proposals within an exponential family. In the latter studies, the improvement implies a quantile-based improvement on the original problem. This translates into proposals having a larger fraction of their mass where the objective function is low, as the algorithm iterations progress. We are not aware of any study connecting the two approaches, which can be explained by the very different mathematical tools and paradigms used in both research lines (e.g., continuous versus discrete time).

We follow in this work the second approach, with the aim to establish new improvement results for black-box optimization algorithms. Ideally, one would want proven improvement results that are valid at every iteration, for realistic step sizes, and for a wide variety of proposal models. Typical motivating examples, widely used in black-box global optimization, are heavy-tailed [31], and mixture [26, 1, 20, 2], proposals. However, such proposals would require infinitesimally small step sizes to benefit from the available improvement results from [27, 5], since they do not form an exponential family. Further, even for proposals within the exponential family, the results from [27, 5] do not quantify the magnitude of the improvements. Therefore, there is still a need for novel wide-ranging criteria to ensure that black-box optimization algorithms improve, and if so, how much, either in terms of reformulated problems or quantiles, that we address in this work.

Our contributions are the following:

- We introduce novel divergence-based conditions, measuring, through a Kullback-Leibler (KL) or a Rényi divergence, the discrepancy between a given target distribution and successive proposals. We show that any generic algorithm satisfying those conditions improves in terms of the expectation-based reformulation of the objective at every iteration, and we quantify the improvement. Namely, if the divergence is decreasing between two consecutive proposals, then the decrease in divergence translates into an improvement both in the expectation-based reformulated objective and quantile. It is worthy to emphasize that our results do not depend on the way the next proposal is designed, making our

conditions a versatile tool to study evolution strategy algorithms.

- We show that the IGO framework fits within the introduced divergence-decrease conditions, illustrating the wide scope of our results. In the case of the IGO reformulation of the objective, we quantify the quantile improvement that comes when the introduced divergence-based conditions hold.
- 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 [23, Example 3.2], and we show that it fits within our framework. We also propose a new algorithm for Student proposals (having heavier tails than Gaussians and including Cauchy), and we show that it satisfies our divergence-based conditions.

Let us position our contributions with respect to existing literature. Our results allow to derive new proofs for the quantile improvement in the IGO framework of [27, 5]. As we discuss later in detail, our results are stated for the IGO quantile-based reformulation of the objective. They also hold for other expectation-based reformulations of the objective such as the ones in [34, 33, 9]. We furthermore quantify the improvement in terms of expectation-based reformulation and quantile, yielding more precise results than in [27, 5]. Contrary to existing works, our results can be applied on proposal models beyond the exponential family, as we show in several examples, and hold without the stringent assumption of infinitesimal step sizes.

The paper is organized as follows. We give preliminary notions about black-box global optimization problems, and algorithms to solve them, in Section 2. We then state our main results in Section 3, including our novel conditions for improvement, and new results obtained under these conditions. Finally, we discuss perspectives in Section 4.

## 2 Preliminary notions

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

### 2.1 Problem setting and reformulation

We consider throughout the paper the generic black-box minimization problem

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

where  $f : \mathbb{X} \rightarrow \mathbb{R}$  may be 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 mixed between continuous and discrete variables. We assume the existence of a measure  $m$  on  $\mathbb{X}$  for some  $\sigma$ -algebra over  $\mathbb{X}$ . For instance, if  $\mathbb{X} = \mathbb{R}^d$ , one can consider the Lebesgue measure, while if  $\mathbb{X} = \mathbb{N}^d$ , one can take the counting measure.

We focus in our work on algorithms that solve Problem (1) through a sampling-based approach. The aim is to construct a parametric probability distribution  $p_\theta$  over  $\mathbb{X}$  such that  $p_\theta$  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  $p_\theta \in \{p_\theta, \theta \in \Theta\}$ . In the following, we make the standard assumption that the considered proposals  $p_\theta$  have a density with respect to  $m$ , also denoted by  $p_\theta$ . One way to transform Problem (1) into a problem over  $\Theta$  is to consider the minimization of  $\theta \mapsto \mathbb{E}_{X \sim p_\theta}[f(X)]$ . This reformulation, which has been studied in [4, 25] for instance, makes however the resulting algorithm sensitive to transformation of  $f$ .

In this work, we focus on an alternative reformulation of Problem (1), which has been proposed in the context of IGO [27], and has the advantage of ensuring more invariance properties. Let the  $p_\theta$ - $f$ -quantiles at  $x \in \mathbb{X}$  be defined by

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

For a given  $x \in \mathbb{X}$ ,  $q_\theta^<$  and  $q_\theta^\leq$  measure the mass that  $p_\theta$  puts on points that achieve (strictly) better value of  $f$  than  $x$ . Select next a weighting non-increasing function  $w : [0, 1] \rightarrow \mathbb{R}_+$ . The authors of [27] then introduced the preference function  $W_\theta^f : \mathbb{X} \rightarrow \mathbb{R}$  which is defined for any  $x \in \mathbb{X}$  as

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

The function  $W_\theta^f$  is a quantile-based rewriting of  $f$  that is invariant under increasing transformation of the objective  $f$ , as  $W_\theta^f \equiv W_\theta^{\phi \circ f}$  for any increasing function  $\phi$  and  $\theta \in \Theta$ . Also,  $W_\theta^f$  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,  $W_\theta^f(x) \geq W_\theta^f(x')$ . Under such definitions, given a proposal with parameter  $\theta' \in \Theta$ , the authors of [27] considered the search for a good proposal  $p_\theta$  to solve (1) as the maximization of the function  $J(\cdot|\theta') : \Theta \rightarrow \mathbb{R}$  defined, for any  $\theta \in \Theta$ , by

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

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

Measuring the quality of a proposal  $p_\theta$  to solve Problem (1) using quantiles has also been proposed in the framework of the cross-entropy (CE) method [23]. Given a proposal  $p_\theta$  and a scalar  $q \in (0, 1)$ , the CE method relies on  $q$ -quantiles of  $f(X)$  where  $X \sim p_\theta$ , that is, any value  $u \in \mathbb{R}$  such that

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

Let us denote, as in [5],  $Q_\theta^q(f)$  as the largest of such values,

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

If  $x$  is sampled from  $p_\theta$ , then  $f(x)$  is below  $Q_\theta^q(f)$  with a probability greater than  $q$ . Therefore, good proposals to solve Problem (1) should be such that  $Q_\theta^q(f)$  is as low as possible. Remark that  $Q_\theta^q(f)$  only depends on the current proposal, contrary to  $J$ .

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 [5] where the authors showed that an increase in term of  $\theta \mapsto J(\theta|\theta')$  relates to a decrease in terms of  $\theta \mapsto Q_\theta^q(f)$ , as stated in the lemma hereafter.

**Lemma 1** (Lemma 8 in [5]). *Consider the weighting function  $w(u) = \delta_{u \leq q}(u)$  with  $q \in (0, 1)$ . If  $(\theta, \theta') \in \Theta^2$  satisfies the increase condition*

$$J(\theta|\theta') > J(\theta'|\theta') = Z_w, \quad (6)$$

*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 algorithms, for either discrete or continuous optimization. Indeed, showing that consecutive proposals achieve the increase condition (6) allows to apply Lemma 1 yielding a proposal with more mass where the objective function is low.

## 2.2 The information-geometric optimization algorithm

We now recall here the information-geometric (IGO) framework from [27, 5]. 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 [27] 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 [23].

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 a practical implementation than continuous flows. Two types of updates have been proposed in [5, 27] to satisfy the increase condition (6), leading to two distinct IGO-like algorithms, that we will recall here.

The first algorithm in [27, 5] is based on natural gradient ascent updates. Consider an iteration  $k \in \mathbb{N}$ , with  $\theta_k$  parameterizing the current proposal, and the objective function being  $J(\cdot|\theta_k)$ . The natural gradient of  $J(\cdot|\theta_k)$  at  $\theta$  is the quantity  $\tilde{\nabla}_\theta J(\theta|\theta_k) = I(\theta)^{-1} \nabla_\theta J(\theta|\theta_k)$ , where  $I(\theta) = -\mathbb{E}_{X \sim p_\theta} [\nabla_\theta^2 \ln p_\theta(X)]$  is the Fisher information matrix of  $p_\theta$ . Given the above gradient expression, iterating a gradient ascent scheme over  $k \in \mathbb{N}$  leads to Algorithm 1. We remark that natural gradient updates have been used in other contexts than IGO, see for instance [33].

---

### Algorithm 1 IGO algorithm (natural gradient update)

---

Initialize  $\theta_0$  and choose the step size  $\tau > 0$ .

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

Update  $\theta_{k+1}$  such that

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

**end for**

---

The second algorithm proposed in [27, 5] estimates the proposal parameters by performing a weighted maximum likelihood update at every iteration. We provide its description in Algorithm 2. The CE algorithm of [23] is recovered as a special case when the step size is  $\tau = 1$  and the weighting function is  $w(u) = \delta_{u \leq q}(u)$  [27].

---

### Algorithm 2 IGO algorithm (IGO-ML update)

---

Initialize  $\theta_0$  and choose the step size  $\tau > 0$ .

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

Update  $\theta_{k+1}$  such that

$$\theta_{k+1} = \arg \max_{\theta \in \Theta} \left( (1 - \tau) \int \ln(p_\theta(x)) p_{\theta_k}(x) m(dx) + \tau \int W_{\theta_k}^f(x) \ln p_\theta(x) p_{\theta_k}(x) m(dx) \right). \quad (8)$$

**end for**

---

The theoretical properties of Algorithms 1 and 2 have been studied in [27, 5]. 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  $\tau \in (0, 1]$  [5, Theorem 6]. This result gives in turn improvement 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 [27, Proposition 7].

Moreover, Algorithms 1 and 2 have been shown to coincide when  $\{p_\theta, \theta \in \Theta\}$  forms an exponential family [7], ensuring that Algorithm 1 satisfies (6) for  $\tau \in (0, 1]$  in this case (see [5, Corollary 7]).

Algorithms 1 and 2 coincide with many existing black-box global optimization algorithms on discrete or continuous domains [27, Section 5], allowing to get improvement guarantees for these algorithms as well. However, the aforementioned study requires, as a preliminary step, to show that the considered algorithms fit within the IGO framework, which is not always possible nor straightforward. In the next section, we present our contribution, that aims at giving novel broader conditions under which the increase condition (6) is satisfied. This allows, in particular, to exhibit new improvement guarantees beyond the IGO framework, the latter being retrieved as a special case.

### 3 A general divergence-based condition for quantile improvement

We present our main results in this section. We start in Section 3.1 with the introduction of novel, divergence-based, conditions. We show that they imply the increase condition (6). We go further by quantifying the improvements in terms of reformulated objective and quantile. We then show in Section 3.2 that IGO algorithms satisfy our conditions, allowing us to provide a new and refined perspective on these methods. We finally exploit our divergence-based conditions to show new improvement guarantees for algorithms using proposals that do not form an exponential family. Namely, in Section 3.3, we study a mixture-based algorithm and discuss his tight links with the mixture-based CE algorithm of [23, Example 3.2]. In Section 3.4, we study an algorithm with heavy-tailed proposals, namely Student distributions with arbitrary degree of freedom parameter.

#### 3.1 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. We also go further and quantify the improvement on the reformulated objective and quantile that come from our divergence-based conditions.

##### 3.1.1 Proposed divergence-based condition

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 the Kullback-Leibler or Rényi divergence. Let us start by specifying the target probability distribution we are going to consider. For  $\theta \in \Theta$ , we introduce  $\pi_\theta^f$ , the probability density with respect to  $m$  defined for any  $x \in \mathbb{X}$  by

$$\pi_\theta^f(x) = \frac{1}{Z_w} W_\theta^f(x) p_\theta(x). \quad (9)$$

When  $w(u) = \delta_{u \leq q}$  for some  $q \in (0, 1)$ ,  $\pi_\theta^f$  is a truncated version of  $p_\theta$  with support being the points  $x \in \mathbb{X}$  such that  $q_\theta^f(x) < q$ , 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  $\pi_{\theta_k}^f$  and either the current proposal, or the next one. This discrepancy is measured using the KL or a Rényi 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

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

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

If for some  $x \in \mathbb{X}$ ,  $p_1(x) = 0$ , then we use  $\ln(p_1(x))p_1(x) = 0$  (see [28, Definition 7.1] for more details on these singular cases). Note that we have  $D_\alpha(p_1, p_2) \xrightarrow{\alpha \rightarrow 1} KL(p_1, p_2)$  [32, Theorem 5], so that the KL divergence can be viewed as a limit case of the Rényi divergence.

We are now ready to state our first result, obtained when using the KL divergence to measure the discrepancy between probability densities. We show hereafter that, when a generic algorithm constructs its next proposal so that it gets closer, by a certain amount, to the target  $\pi_{\theta_k}^f$  defined in (9), then it improves upon the reformulated objective defined in (3), by an amount that we quantify.

**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, for some  $\Delta_k \in \mathbb{R}$ ,*

$$KL(\pi_{\theta_k}^f, p_{\theta_{k+1}}) + \Delta_k \leq KL(\pi_{\theta_k}^f, p_{\theta_k}). \quad (10)$$

*Then,  $J(\theta_{k+1}|\theta_k) \geq \exp(\Delta_k)J(\theta_k|\theta_k) = \exp(\Delta_k)Z_w$ . In particular,  $(\theta_{k+1}, \theta_k)$  satisfy the increase condition (6), i.e.,  $J(\theta_{k+1}|\theta_k) > Z_w$  with  $J$  defined in (3), if  $\Delta_k > 0$ .*

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

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

and remark that it is equivalent to having

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

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

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

which shows, by definition of  $\pi_{\theta_k}^f$ , and using that  $J(\theta_k|\theta_k) = Z_w$ , that  $J(\theta_{k+1}|\theta_k) \geq \exp(\Delta_k)Z_w$ .  $\square$

We now state a second similar result, that arises when one now measures the discrepancy between the target density and the proposals using a Rényi 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, for some  $\Delta_k \in \mathbb{R}$ ,*

$$D_\alpha(\pi_{\theta_k}^f, p_{\theta_{k+1}}) + \Delta_k \leq D_\alpha(\pi_{\theta_k}^f, p_{\theta_k}), \quad (11)$$

*for some  $\alpha \in (0, 1)$ .  $J(\theta_{k+1}|\theta_k) \geq \exp(\Delta_k)J(\theta_k|\theta_k) = \exp(\Delta_k)Z_w$ . In particular,  $(\theta_{k+1}, \theta_k)$  satisfy the increase condition (6) if  $\Delta_k > 0$ .*



*Proof.* By definition of  $\pi_{\theta_k}^f$  and since by assumption,  $W_{\theta_k}^f(x)^\alpha = W_{\theta_k}^f(x)$  for any  $x \in \mathbb{X}$ , Equation (11) is equivalent to

$$\begin{aligned} & \frac{1}{\alpha - 1} \ln \left( \int \left( \frac{p_{\theta_{k+1}}(x)}{p_{\theta_k}(x)} \right)^{1-\alpha} W_{\theta_k}^f(x) p_{\theta_k}(x) m(dx) \right) + \Delta_k \\ & \leq \frac{1}{\alpha - 1} \ln \left( \int W_{\theta_k}^f(x) p_{\theta_k}(x) m(dx) \right) = \frac{1}{\alpha - 1} \ln Z_w. \end{aligned}$$

We deduce from there that

$$\int \left( \frac{p_{\theta_{k+1}}(x)}{p_{\theta_k}(x)} \right)^{1-\alpha} \pi_{\theta_k}^f(x) m(dx) \geq \exp((1 - \alpha)\Delta_k).$$

Finally, since  $u \mapsto u^{1-\alpha}$  is concave due to the assumption on  $\alpha$ , we apply Jensen's inequality to obtain that

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

showing by definition of  $\pi_{\theta_k}^f$ , that  $\int W_{\theta_k}^f(x) p_{\theta_{k+1}}(x) m(dx) \geq \exp(\Delta_k) Z_w$ , and hence establishing the result.  $\square$

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 [27, 5], the increase condition (6) was achieved for specific algorithms only.

*Remark 1.* Results equivalent to Propositions 1 and 2 can be stated for other expectation-based reformulations of Problem (1) (under very mild hypotheses). For instance, consider the minimization over  $\Theta$  of  $J : \theta \mapsto \mathbb{E}_{X \sim p_\theta}[\phi(f(X))]$  for some transform  $\phi : \mathbb{R} \rightarrow \mathbb{R}_+$  (see for instance [34]). Consider as well the tilted densities  $\pi_\theta^f$  defined for  $\theta \in \Theta$  by  $\pi_\theta^f(x) \propto \phi(f(x)) p_\theta(x)$  for any  $x \in \mathbb{X}$ . Being able to define the probability density  $\pi_\theta^f$  in this way is the only requirement on  $\phi$ . Then, decrease conditions of the form (10), or (11) if  $\phi$  takes values in  $\{0, 1\}$ , imply that

$$\mathbb{E}_{X \sim p_{\theta_{k+1}}}[\phi(f(X))] \geq \exp(\Delta_k) \mathbb{E}_{X \sim p_{\theta_k}}[\phi(f(X))],$$

translating the improvement in terms of divergence to an improvement in the reformulation of Problem (1).

### 3.1.2 Quantile improvement quantification

We now present an additional result that quantifies how an improvement in term of the reformulation  $J(\theta|\theta_k)$ , defined in (3), results in an improvement in terms of  $Q_\theta^q(f)$ , defined in (5). As we explained in the previous Section, in the particular case when  $w(u) = \delta_{u \leq q}(u)$ , there is a link between  $J(\theta|\theta_k)$  and  $Q_\theta^q(f)$ . Our result can thus be seen as a quantitative and more precise version of Lemma 1. Although the previous results hold in the continuous and discrete settings, this result hereafter requires assumptions on  $f$  and  $\mathbb{X}$  that will hold for most continuous optimization problems but will usually not hold for discrete problems.

**Proposition 3.** *Assume that  $w(u) = \delta_{u \leq q}(u)$  for some  $q \in (0, 1)$ . Let  $k \in \mathbb{N}$  and  $(\theta_k, \theta_{k+1}) \in \Theta^2$  such that  $J(\theta_{k+1}|\theta_k) \geq \exp(\Delta_k) Z_w$  with  $\exp(\Delta_k) \mathbb{P}_{X \sim p_{\theta_{k+1}}}[f(X) \leq Q_{\theta_{k+1}}^q(f)] \in [0, 1]$ . Suppose that  $F_{\theta_{k+1}} : u \mapsto \mathbb{P}_{X \sim p_{\theta_{k+1}}}[f(X) \leq u]$  is continuous and strictly monotonic, and hence bijective from the range of  $f$  (which is thus an interval of  $\mathbb{R}$ ) to  $[0, 1]$ . Then, we have*

$$Q_{\theta_k}^q(f) \geq F_{\theta_{k+1}}^{-1} \left( \exp(\Delta_k) F_{\theta_{k+1}}(Q_{\theta_{k+1}}^q(f)) \right). \quad (12)$$

If  $F_{\theta_{k+1}}$  is bijective and  $\mathcal{C}^1$  in a neighbourhood of  $Q_{\theta_{k+1}}^q(f)$  with  $F'_{\theta_{k+1}}(Q_{\theta_{k+1}}^q(f)) \neq 0$ , then

$$Q_{\theta_k}^q(f) \geq Q_{\theta_{k+1}}^q(f) + q\Delta_k \left(F_{\theta_{k+1}}^{-1}\right)'(q) + o(\Delta_k^2) \quad (13)$$

for  $\Delta_k$  small enough, with  $\left(F_{\theta_{k+1}}^{-1}\right)'(q) > 0$ .

*Proof.* We first show that  $F_{\theta_{k+1}}(Q_{\theta_k}^q(f)) \geq q\Delta_k$ , by adapting parts of the proof of [5, Lemma 8]. First, following the arguments laid in the proof of [5, Lemma 8], we have that  $W_{\theta_k}^f(x) = 0$  for any  $x \in \mathbb{X}$  such that  $f(x) > Q_{\theta_k}^q(f)$ . This implies that  $J(\theta_{k+1}|\theta_k) \leq \mathbb{P}_{X \sim p_{\theta_{k+1}}}[f(X) \leq Q_{\theta_k}^q(f)] = F_{\theta_{k+1}}(Q_{\theta_k}^q(f))$ .

From our assumptions, we thus have that  $\exp(\Delta_k)q \leq F_{\theta_{k+1}}(Q_{\theta_k}^q(f))$  since our choice of  $w$  implies  $Z_w = q$ . On the other hand, our assumption on  $F_{\theta_{k+1}}$  implies that  $Q_{\theta_{k+1}}^q(f)$  is the only value satisfying  $q = F_{\theta_{k+1}}(Q_{\theta_{k+1}}^q(f))$ . These two facts imply that  $F_{\theta_{k+1}}(Q_{\theta_k}^q(f)) \geq \exp(\Delta_k)F_{\theta_{k+1}}(Q_{\theta_{k+1}}^q(f))$ , establishing Equation (12) with the bijectivity of  $F_{\theta_{k+1}}$ .

In order to prove Equation (13), we use Equation (12) and develop  $F_{\theta_{k+1}}^{-1} \circ \exp$  around  $\ln F_{\theta_{k+1}}(Q_{\theta_{k+1}}^q(f))$ . We get the well-posedness and sign of  $\left(F_{\theta_{k+1}}^{-1}\right)'$  from the inverse function theorem.  $\square$

Proposition 3 shows a complex interplay between the proposals and the objective  $f$  through the cumulative density function  $F_{\theta_{k+1}}$ . This Proposition may be used to assess how a certain family of proposals is adapted to the problem at hand. Indeed, one could aim for a family of proposals such that  $\left(F_{\theta_{k+1}}^{-1}\right)'(q)$  is as high as possible to ensure the largest quantile improvement possible.

In the particular case when  $w(u) = \delta_{u \leq q}(u)$ , we can then extend the result of Propositions 1 and 2 with Lemma 1 and Proposition 3 to obtain the following quantile improvement results from divergence-decrease conditions.

**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*

$$D_\alpha(p_{\theta_{k+1}}, \pi_{\theta_k}^f) + \Delta_k \leq D_\alpha(p_{\theta_k}, \pi_{\theta_k}^f) \quad (14)$$

for some  $\Delta_k \in \mathbb{R}$  and  $\alpha \in (0, 1]$ , with  $\alpha = 1$  corresponding to the inequality  $KL(\pi_{\theta_k}^f, p_{\theta_{k+1}}) + \Delta_k \leq KL(\pi_{\theta_k}^f, p_{\theta_k})$ .

- (i) Suppose that  $\Delta_k > 0$ . If  $\alpha = 1$ , or if  $\alpha \in (0, 1)$  and  $W_{\theta_k}^f$  takes values in  $\{0, 1\}$ , then  $Q_{\theta_{k+1}}^q(f) \leq Q_{\theta_k}^q(f)$ .
- (ii) If  $\Delta_k > 0$  and  $\mathbb{P}_{X \sim p_\theta}[f(X) = u] = 0$  for any  $\theta \in \Theta$ ,  $u \in \mathbb{R}$ , then  $Q_{\theta_{k+1}}^q(f) < Q_{\theta_k}^q(f)$ .
- (iii) Suppose that  $F_{\theta_{k+1}} : u \mapsto \mathbb{P}_{X \sim p_\theta}[f(X) \leq u]$  is continuous and strictly monotonic. If  $\alpha = 1$ , or if  $\alpha \in (0, 1)$  and  $W_{\theta_k}^f$  takes values in  $\{0, 1\}$ , then Equation (12) holds.
- (iv) Suppose that  $F_{\theta_{k+1}} : u \mapsto \mathbb{P}_{X \sim p_\theta}[f(X) \leq u]$  is bijective and  $\mathcal{C}^1$  around  $Q_{\theta_{k+1}}^q(f)$ , while we have  $F'_{\theta_{k+1}}(Q_{\theta_{k+1}}^q(f)) \neq 0$ . If  $\alpha = 1$ , or if  $\alpha \in (0, 1)$  and  $W_{\theta_k}^f$  takes values in  $\{0, 1\}$ , then Equation (13) holds for  $\Delta_k$  small enough.

*Proof.* Point (i) follows from Propositions 1 and 2 with the first part of Lemma 1. Point (ii) follows by remarking that, under our assumptions, for any  $\theta \in \Theta$ ,  $x \in \mathbb{X}$ ,  $q_\theta^<(x) = q_\theta^>(x)$ , ensuring that  $W_{\theta_k}^f(x)$  takes values in  $\{0, 1\}$ . This also implies that  $\mathbb{P}_{X \sim p_{\theta_{k+1}}}[f(X) = Q_{\theta_k}^q(f)] = 0$ . The result comes by applying the second part of Lemma 1. Finally, points (iii) and (iv) are proven using the results from Propositions 1 and 2 together with the results of Proposition 3.  $\square$

### 3.1.3 Monitoring target construction

We now show that the KL and Rényi divergences can also be used to control the discrepancy between the proposal  $p_\theta$  and the resulting target  $\pi_\theta^f$ . The resulting bound only depends on the choice of the weighting function  $w$ .

**Proposition 4.** *Consider  $\theta \in \Theta$  and the probability densities  $p_\theta$  and  $\pi_\theta^f$ . We have the following results, writing  $D_1(\pi_\theta^f, p_\theta)$  for  $KL(\pi_\theta^f, p_\theta)$ .*

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

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

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

$$D_\alpha(\pi_\theta^f, p_\theta) = \frac{1}{\alpha - 1} \ln \left( \int \frac{W_\theta^f(x)^\alpha}{Z_w^\alpha} p_\theta(x) m(dx) \right). \quad (15)$$

(i) We have that  $W_\theta^f(x)^{1-\alpha} = W_\theta^f(x)$  for any  $x \in \mathbb{X}$ , thus showing with Equation (15) that  $D_\alpha(p_\theta, \pi_\theta^f) = -\ln Z_w$  for any  $\alpha \in (0, 1) \cup (1, +\infty)$  which implies the result, using [32, Theorem 3] for the case of the KL divergence.

(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,  $W_\theta^f(x)^\alpha \geq W_\theta^f(x)$  for any  $x \in \mathbb{X}$  when  $\alpha \in (0, 1)$ . We thus get from Equation (15) when  $\alpha \in (0, 1)$  that  $D_\alpha(\pi_\theta^f, p_\theta) \leq -\ln Z_w$ . Taking the limit  $\alpha \rightarrow 1$ ,  $\alpha < 1$ , we finally obtain that  $KL(\pi_\theta^f, p_\theta) \leq -\ln Z_w$ .  $\square$

*Remark 2.* Consider, following Remark 1, that we use a target density of the form  $\pi_\theta^f(x) \propto \phi(f(x))p_\theta(x)$ , using  $\phi \circ f$  instead of  $W_\theta^f$ . It would be possible to derive results like Proposition 4. However, the normalization constant of  $\pi_\theta^f$  is  $\int \phi(f(x))p_\theta(x)m(dx)$  which depends on  $\theta$ , while using  $W_\theta^f$  ensures that the normalization constant of  $\pi_\theta^f$  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 4, we have furthermore 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 also 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.

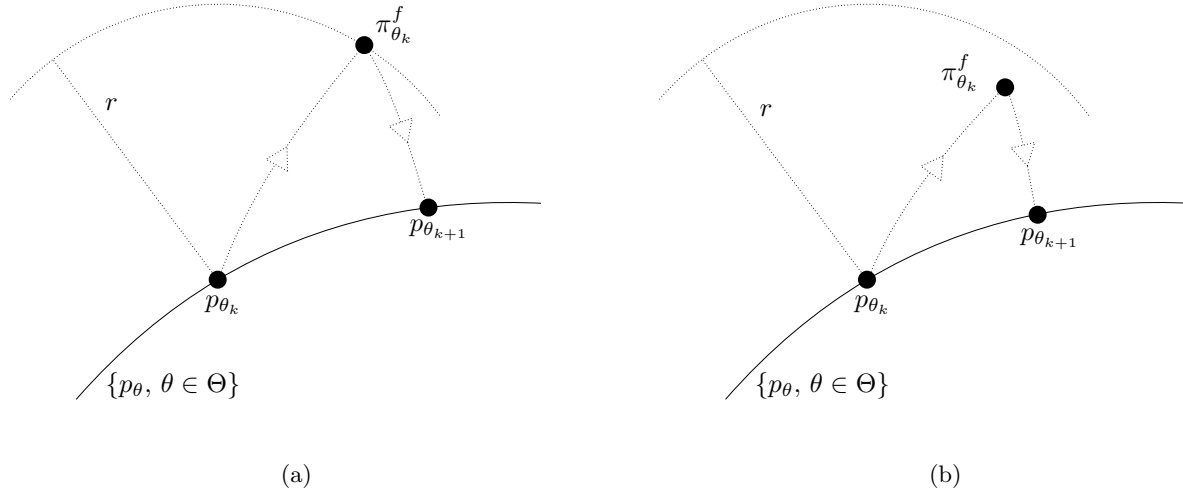


Figure 1: A schematic view of one step of an algorithm covered by our divergence-based framework. Starting from a proposal  $p_{\theta_k}$ , one first constructs the target  $\pi_{\theta_k}^f$  following (9), i.e., a process that benefits from the results of Proposition 4, with  $r = -\ln Z_w$ . Then, one adapts  $p_{\theta_{k+1}}$  such that a divergence-decrease (14), for some  $\alpha \in (0, 1]$ , is achieved. (a) Case  $\alpha \in (0, 1]$  and  $W_{\theta_k}^f$  taking values in  $\{0, 1\}$ , corresponding to Propositions 1 or 2 and Proposition 4 (i). (b) Case  $\alpha \in (0, 1]$  and  $w$  taking values in  $[0, 1]$ , corresponding to Propositions 1 or 2 and Proposition 4 (ii).

### 3.1.4 Finite sample regime

Our divergence-decrease conditions, presented in Propositions 1 and 2, are useful to analyze algorithms working in the finite-sample regime. Indeed, our results hold independently of the construction strategy adopted to define the next proposal. Actually, they even hold when the next proposal degrades the performance upon the current one, which corresponds to the case  $\Delta_k < 0$  in Equations (10) and (11). Such situation can typically arise in the finite sample regime, as we discuss hereafter.

In the case of finite sample size, the construction of the next proposal becomes inexact (i.e., noisy) as gradients need to be approximated. This makes the analysis of the algorithm more difficult. If the noise is controlled, in such a way that its effect is ‘absorbed’ by the parameter  $\Delta_k$  in the divergence-decrease conditions of Equations (10) and (11), then all of our results still apply (for instance in expectation or with high probability, depending on the control one has on the noise). Namely, it remains possible to establish improvement, or control the degradation if  $\Delta_k < 0$ , in terms of expectation-based reformulation and then in terms of quantile in such noisy contexts.

## 3.2 Analyzing the IGO algorithms with our framework

### 3.2.1 Main result

We now show that both IGO algorithms, namely Algorithms 1 and 2, proposed in [27], fall within our divergence-decrease framework, showing the applicability of our construction. We quantify the improvement achieved at each iteration in terms of divergence, which can then be used to quantify the quantile improvement using Proposition 3.

**Proposition 5.** 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 assumptions detailed in the proof), that  $KL(\pi_{\theta_k}^f, p_{\theta_{k+1}}) + \Delta_k \leq KL(\pi_{\theta_k}^f, p_{\theta_k})$  with  $\Delta_k = \frac{1-\tau Z_w}{\tau Z_w} KL(p_{\theta_k}, p_{\theta_{k+1}})$ .
- (iii) If Algorithm 2 is used, then  $KL(\pi_{\theta_k}^f, p_{\theta_{k+1}}) + \Delta_k \leq KL(\pi_{\theta_k}^f, p_{\theta_k})$  with  $\Delta_k = \frac{1-\tau}{\tau Z_w} KL(p_{\theta_k}, p_{\theta_{k+1}})$ .

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

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

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

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

(ii) Suppose that  $\{p_\theta, \theta \in \Theta\}$  forms a minimal exponential family [7] with sufficient statistics  $\Gamma$  and log-partition function  $A$  with  $\Theta = \text{dom } A$ . Then,  $A$  is differentiable on  $\text{int dom } A$  with  $\nabla A(\theta) = \mathbb{E}_{X \sim p_\theta}[\Gamma(X)]$  [7, Theorem 8.1]. We also suppose that  $(\theta_{k+1}, \theta_k) \in (\text{int } \Theta)^2$  and that for any  $\theta \in \text{dom } A$ ,  $KL(\pi_{\theta_k}^f, p_\theta) < +\infty$ .

From [5, 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 having

$$\mathbb{E}_{X \sim \pi_{\theta_k}^f} [\Gamma(X)] - \nabla A(\theta_{k+1}) + \frac{1 - \tau Z_w}{\tau Z_w} (\nabla A(\theta_k) - \nabla A(\theta_{k+1})) = 0. \quad (16)$$

We now interpret Equation (16) as an optimality condition, showing that

$$\theta_{k+1} = \arg \min_{\theta \in \Theta} \left( KL(\pi_{\theta_k}^f, p_\theta) + \frac{1 - \tau Z_w}{\tau Z_w} KL(p_{\theta_k}, p_\theta) \right).$$

This implies that  $\theta_{k+1}$  is such that

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

We now show that Equation (16) is the optimality condition of the problem solved by  $\theta_{k+1}$ . Consider any  $\theta \in \Theta$  and a probability density with respect to  $m$ , denoted by  $p$ . We have

$$KL(p, p_\theta) = \mathbb{E}_{X \sim p} [\ln p(X)] - \langle \mathbb{E}_{X \sim p} [\Gamma(X)], \theta \rangle + A(\theta),$$

and thus obtain that  $\nabla_\theta KL(p, p_\theta) = \nabla A(\theta) - \mathbb{E}_{X \sim p} [\Gamma(X)]$ , which we then use to show the desired result.

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

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

We thus obtain by definition of  $\theta_{k+1}$  that

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

□

### 3.2.2 Discussion

Proposition 5 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 5 (i) that the increase (6), that is  $J(\theta_{k+1}|\theta_k) > Z_w$ , is guaranteed for infinitesimal step sizes [27, Proposition 7]. Similarly, we recover in Proposition 5 (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 [5, Theorem 6].

Proposition 5 (ii) establishes a similar result for Algorithm 1 when the proposals form an exponential family. In the proof of [5, Corollary 7], this result was achieved by remarking that in this case, Algorithms 1 and 2 coincide. We do a direct proof, showing that Algorithm 1 is equivalent to a proximal update with a KL divergence objective. Note that [5, Corollary 7] ensured improvement for step sizes in  $(0, 1]$  while our results allow for possibly larger step sizes (if  $w(u) = \delta_{u \leq q}(u)$  with  $q \in (0, 1)$ ,  $1/Z_w = 1/q > 1$ ). This is because the authors of [5] defined  $W_\theta^f$  such that  $Z_w = 1$ , while we chose here a different convention.

*Remark 3.* Proposition 5 (ii) actually holds in the setting of Remark 1. More explicitly, we get the same result, with same  $\Delta_k$ , when optimizing  $\theta \mapsto \mathbb{E}_{X \sim p_\theta}[\phi(f(X))]$  over an exponential family using natural gradient descent with  $\pi_\theta^f$  defined as in Remark 1. This allows to control the improvement of  $\theta \mapsto \mathbb{E}_{X \sim p_\theta}[\phi(f(X))]$  over iterations thanks to the result outlined in Remark 1.

### 3.3 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. As already explained, 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 [11], a type of expectation-maximization (EM) algorithm proposed in the context of computational statistics, and with the mixture-based CE method proposed in [23, Example 3.2]. Similarly to IGO, our proposed algorithm can be applied to discrete and continuous optimization problems. By exploiting the paradigm we introduced in Section 3.1, 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 Corollary 1 to establish quantile improvement.

### 3.3.1 Proposed algorithm

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  $p_\theta = \sum_{j=1}^J \lambda^{(j)} p_{\vartheta^{(j)}}$  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  $\theta = (\{\lambda^{(j)}\}_{j=1}^J, \{\vartheta^{(j)}\}_{j=1}^J)$ .

---

#### 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  $\rho_k^{(j)} : \mathbb{X} \rightarrow \mathbb{R}$  defined for any  $x \in \mathbb{X}$  by

$$\rho_k^{(j)}(x) = \frac{\lambda_k^{(j)} p_{\vartheta_k^{(j)}}(x)}{\sum_{i=1}^J \lambda_k^{(i)} p_{\vartheta_k^{(i)}}(x)}. \quad (17)$$

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

$$\lambda_{k+1}^{(j)} = \mathbb{E}_{X \sim \pi_{\theta_k}^f} [\rho_k^{(j)}(X)], \quad (18)$$

$$\vartheta_{k+1}^{(j)} = \arg \max_{\vartheta \in \Theta} \mathbb{E}_{X \sim \pi_{\theta_k}^f} [\ln p_\vartheta(X) \rho_k^{(j)}(X)]. \quad (19)$$

**end for**

---

Algorithm 3 shares links with the EM point of view adopted in [9]. In this work, several estimation-of-distribution algorithms [24] were 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 [34] and the discussion in [4, Section 5.3]). Algorithm 3 recovers the M-PMC algorithm of [11], which is also an EM-like algorithm, but with rank-based weights (see [27, Equation (14)] or Equation (22)) instead of importance weights. Note that, contrary to [9] 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.

### 3.3.2 Main result

We now show in Proposition 6 that Algorithm 3 achieves a decrease in terms of KL divergence at every iteration. Our proof techniques are reminiscent from the recent work [13] on variational inference. The result of Proposition 6 can then be used to apply Proposition 1 and Corollary 1 and get insights on the optimization performance of Algorithm 3.

**Proposition 6.** *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  $k \in \mathbb{N}$ . Suppose that the problem in (19) is uniquely maximized at every iteration. Then, at every iteration  $k \in \mathbb{N}$ , Algorithm 3 achieves the decrease*

$$KL(\pi_{\theta_k}^f, p_{\theta_{k+1}}) + \Delta_k \leq KL(\pi_{\theta_k}^f, p_{\theta_k}), \quad (20)$$

with  $\Delta_k > 0$ , unless  $\lambda_{k+1}^{(j)} = \lambda_k^{(j)}$  and  $\vartheta_{k+1}^{(j)} = \vartheta_k^{(j)}$  for every  $j = 1, \dots, J$ , in which case  $\Delta_k = 0$ .

*Proof.* We adapt the ideas of the proof of [13, Theorem 2]. We compute the quantity

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

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,

$$\begin{aligned}
\int -\sum_{j=1}^J \rho_k^{(j)}(x) \ln \left( \frac{\lambda_{k+1}^{(j)} p_{\vartheta_{k+1}^{(j)}}(x)}{\lambda_k^{(j)} p_{\vartheta_k^{(j)}}(x)} \right) \pi_{\theta_k}^f(x) m(dx) &= -\sum_{j=1}^J \ln \left( \frac{\lambda_{k+1}^{(j)}}{\lambda_k^{(j)}} \right) \int \rho_k^{(j)}(x) \pi_{\theta_k}^f(x) m(dx) \\
&\quad + \sum_{j=1}^J \int \rho_k^{(j)}(x) \ln \left( \frac{p_{\vartheta_k^{(j)}}(x)}{p_{\vartheta_{k+1}^{(j)}}(x)} \right) \pi_{\theta_k}^f(x) m(dx). \quad (21)
\end{aligned}$$

Due to the definition of  $\lambda_{k+1}$ , given in Equation (18), 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 (21) 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 (21) is a sum of  $J$  terms, each being non-positive from the definition of  $\vartheta_{k+1}^{(j)}$  given in Equation (19). 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 (19) is uniquely maximized. We have thus shown the decrease (20), with equality holding if and only if  $\lambda_{k+1} = \lambda_k$  and  $\theta_{k+1} = \theta_k$ .  $\square$

### 3.3.3 Discussion

Since our Proposition 6 can be used to apply Corollary 1, it can be viewed, to our knowledge, as the first result to establish quantile improvement for black-box global optimization algorithms that are explicitly mixture-based. Indeed, mixtures were not explicitly considered in [27, 5], and they often do not admit closed-form solutions for the maximization problem (8) in Algorithm 2. The strategy adopted in the literature was usually to perform EM-like updates, as it was done in [23, Example 3.2] for instance, which can now be handled with our divergence-decrease condition. Many variational inference or adaptive importance sampling methods explicitly consider mixtures, see for instance [11, 10, 13], 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 [26, 1, 2], whose convergence has only been verified empirically, the proposed Algorithm 3 has a fixed number of mixture components.

We now discuss the links between Algorithm 3 and the CE algorithm of [23, Example 3.2]. To make these links more explicit, we discuss the finite sample size implementation of Algorithm 3. This requires the



computation of integrals with respect to  $\pi_{\theta_k}^f$ , as discussed in [27]. To do so,  $N$  points  $x_{k,n}$ ,  $n = 1, \dots, N$ , are first sampled 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  $p_{\vartheta_k^{(j)}}$ . Each sample  $x_{k,n}$  receives a rank-based weight  $\widehat{\omega}_{k,n}$  defined as in [27, Equation (14)] by

$$\widehat{\omega}_{k,n} = \frac{1}{N} w \left( \frac{\text{rank}(x_{k,n}) + 1/2}{N} \right), \quad (22)$$

where  $\text{rank}(x_{k,n})$  is the number of samples in  $\{x_{k,n}\}_{n=1}^N$  with value of  $f$  strictly less than  $f(x_{k,n})$ . Then, one can show using [27, Proposition 27] and Slutsky's Lemma that for any integrand  $h$  such that  $\mathbb{E}_{X \sim p_{\theta_k}} [h(X)^2] < +\infty$ , and conditioned on  $\theta_k$ , that

$$\frac{1}{\sum_{n=1}^N \widehat{\omega}_{k,n}} \sum_{n=1}^N \widehat{\omega}_{k,n} h(x_{k,n}) \xrightarrow[N \rightarrow +\infty]{a.s.} \mathbb{E}_{X \sim \pi_{\theta_k}^f} [h(X)]. \quad (23)$$

In light of the above formula, Algorithm 3 can be approximated at iteration  $k \in \mathbb{N}$  by setting, for every  $j = 1, \dots, J$ ,

$$\begin{aligned} \lambda_{k+1}^{(j)} &= \frac{1}{\sum_{n=1}^N \widehat{\omega}_{k,n}} \sum_{n=1}^N \widehat{\omega}_{k,n} \rho_k^{(j)}(x_{k,n}), \\ \vartheta_{k+1}^{(j)} &= \arg \max_{\vartheta \in \Theta} \frac{1}{\sum_{n=1}^N \widehat{\omega}_{k,n}} \sum_{n=1}^N \widehat{\omega}_{k,n} \ln p_{\vartheta}(x_{k,n}) \rho_k^{(j)}(x_{k,n}). \end{aligned}$$

In order to compare with the mixture-based CE algorithm of [23, Example 3.2], let us introduce  $\xi_{k,n}^{(j)}$  which is a latent variable being equal to 1 if  $x_{k,n}$  has been sampled from the component  $j$  of the mixture and zero otherwise. Then, the CE algorithm of [23, Example 3.2] has the following update at iteration  $k \in \mathbb{N}$  and for every  $j = 1, \dots, J$ .

$$\begin{aligned} \lambda_{k+1}^{(j)} &= \frac{1}{\sum_{n=1}^N \widehat{\omega}_{k,n}} \sum_{n=1}^N \widehat{\omega}_{k,n} \xi_{k,n}^{(j)}, \\ \vartheta_{k+1}^{(j)} &= \arg \max_{\vartheta \in \Theta} \frac{1}{\sum_{n=1}^N \widehat{\omega}_{k,n}} \sum_{n=1}^N \widehat{\omega}_{k,n} \ln p_{\vartheta}(x_{k,n}) \xi_{k,n}^{(j)}. \end{aligned}$$

We thus observe that the approximated version of Algorithm 3 and the mixture-based CE algorithm are very similar, except that  $\xi_{k,n}^{(j)}$  is used instead of  $\rho_k^{(j)}(x_{k,n})$  in the latter. 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  $\xi_{k,n}^{(j)}$  amounts to a Rao-Blackwellized version (i.e., a random variable is replaced by its conditional expectation [11]) of the CE algorithm from [23, Example 3.2]. The procedure used in the approximated version of Algorithm 3 does not entail additional evaluations of the objective  $f$ , while providing better numerical stability [11], as all the components of the mixtures appear in every update. This shows how our divergence-based conditions can be used to better understand and design algorithms for mixture-based proposals and also connects our approach to methods in adaptive importance sampling such as [11].

### 3.4 A new result for heavy-tailed proposals with our framework

We finally 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. Note that, contrary to our previous analysis which also held for discrete problems, in this section, we now assume that  $\mathbb{X} = \mathbb{R}^d$  and take  $m$  to be the Lebesgue measure.

### 3.4.1 Proposed algorithm

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

$$\mathcal{T}(x; \mu, \Sigma, \nu) \propto \left(1 + \frac{1}{\nu}(x - \mu)^\top \Sigma^{-1}(x - \mu)\right)^{-\frac{\nu+d}{2}} \quad (24)$$

with normalization constant being equal to  $\frac{\Gamma(\nu/2)}{\Gamma((\nu+d)/2)}(\nu^d \pi^d \det(\Sigma))^{1/2}$ ,  $\Gamma$  denoting the Gamma function and  $\det$  the determinant. When  $\nu = 1$ , the Cauchy distributions are recovered, while Gaussian distributions are recovered in the limit  $\nu \rightarrow +\infty$ . Alternatively, the density in (24) can be written as the continuous mixture

$$\mathcal{T}(x; \mu, \Sigma, \nu) = \int_0^{+\infty} \mathcal{N}\left(x; \mu, \frac{1}{z}\Sigma\right) \mathcal{G}\left(z; \frac{\nu}{2}, \frac{\nu}{2}\right) dz, \quad (25)$$

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 Gaussian distribution with mean  $\mu$  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 (25) 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  $\gamma_k^{(\nu)} : \mathbb{X} \rightarrow \mathbb{R}$  defined for any  $x \in \mathbb{X}$  by

$$\gamma_k^{(\nu)}(x) = \frac{\nu + d}{\nu + (x - \mu_k)^\top \Sigma^{-1}(x - \mu_k)}. \quad (26)$$

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

$$\mu_{k+1} = \frac{\mathbb{E}_{X \sim \pi_{\theta_k}^f} [\gamma_k^{(\nu)}(X)X]}{\mathbb{E}_{X \sim \pi_{\theta_k}^f} [\gamma_k^{(\nu)}(X)]}, \quad (27)$$

$$\Sigma_{k+1} = \frac{\mathbb{E}_{X \sim \pi_{\theta_k}^f} [\gamma_k^{(\nu)}(X)XX^\top]}{\mathbb{E}_{X \sim \pi_{\theta_k}^f} [\gamma_k^{(\nu)}(X)]} - \mu_{k+1}\mu_{k+1}^\top. \quad (28)$$

**end for**

---

When the degree of freedom parameter  $\nu$  goes to infinity, we have that  $\mathcal{T}(x; \mu, \Sigma, \nu) \rightarrow \mathcal{N}(x; \mu, \Sigma)$  for any  $x \in \mathbb{R}^d$ , meaning that Student distributions recover the Gaussian distributions. Moreover, we have at

any iteration  $k \in \mathbb{N}$  that  $\gamma_k^{(\nu)}(x) \rightarrow 1$  when  $\nu \rightarrow +\infty$ . In this case, the updates (27) and (28) 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  $f$ .

### 3.4.2 Main result

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 use Corollary 1 to get quantile improvement when  $w(u) = \delta_{u \leq q}(u)$  is used.

**Proposition 7.** *Consider a sequence  $\{\theta_k\}_{k \in \mathbb{N}}$  generated by Algorithm 4. At every iteration  $k \in \mathbb{N}$ , we have the decrease*

$$KL(\pi_{\theta_k}^f, p_{\theta_{k+1}}) + \Delta_k \leq KL(\pi_{\theta_k}^f, p_{\theta_k}), \quad (29)$$

with  $\Delta_k > 0$ , unless  $\theta_{k+1} = \theta_k$  in which case  $\Delta_k = 0$ .

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

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

Hence, we have that

$$\int \ln p_\theta(x) \pi_{\theta_k}^f(x) dx \geq \iint \ln \left( \frac{p_\theta(x, z)}{p(z|x)} \right) p(z|x) dz \pi_{\theta_k}^f(x) dx, \quad (30)$$

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 (30), we obtain

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

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

$$KL(\pi_{\theta_k}^f, p_{\theta_{k+1}}) - KL(\pi_{\theta_k}^f, p_{\theta_k}) \leq - \iint \ln p_{\theta_{k+1}}(x|z) p_{\theta_k}(z|x) dz \pi_{\theta_k}^f(x) dx + \iint \ln p_{\theta_k}(x|z) p_{\theta_k}(z|x) dz \pi_{\theta_k}^f(x) dx.$$

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

$$\theta_{k+1} = \arg \max_{\theta \in \Theta} \iint \ln p_\theta(x|z) p_{\theta_k}(z|x) dz \pi_{\theta_k}^f(x) dx, \quad (31)$$

establishes the decrease in Equation (29) 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 (31).

For any  $\theta \in \Theta$ , we have that  $p_\theta(x|z) = \mathcal{N}(x; \mu, \frac{1}{2}\Sigma)$ . Hence, we can compute that

$$\iint \ln p_\theta(x|z) p_{\theta_k}(z|x) dz \pi_{\theta_k}^f(x) dx = - \ln \det(\Sigma) - \frac{1}{2} \iint z p_{\theta_k}(z|x) dz (x - \mu)^\top \Sigma^{-1} (x - \mu) \pi_{\theta_k}^f(x) dx.$$

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  $(\frac{\nu+d}{2}, \frac{1}{2}(\nu + (x - \mu_k)^\top \Sigma_k^{-1} (x - \mu_k)))$ . We then remark that  $\gamma_k^{(\nu)}$  as defined in (26) satisfies for any  $x \in \mathbb{R}^d$

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

and we thus get an explicit expression for the objective in (31) of the form

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

from which the result follows.  $\square$

### 3.4.3 Discussion

The result of our Proposition 7 allows to give improvement guarantees for heavy-tailed distributions, that do not form an exponential family. In particular, this applies for any Student family, including Cauchy distributions when  $\nu = 1$ , and to Gaussian distributions in the limit  $\nu \rightarrow +\infty$ . It has been shown in [30]

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 it is done for instance in [13, 15].

Algorithm 4 requires the computation of expectations with respect to  $\pi_{\theta_k}^f$  at every iteration  $k \in \mathbb{N}$ . In practice, these expectations can be approximated with samples from the current proposal  $p_{\theta_k}$  that are then weighted according to their rank, as we discussed in Section 3.3. Such approximations are consistent when the number of samples goes to infinity, as discussed in Section 3.3 and [27]. In the context of computational statistics, algorithms similar to Algorithm 4 have been implemented for experiments in [11].

## 4 Conclusion and perspectives

We have proposed in this work divergence-based conditions that imply the quantile improvement results achieved by the IGO framework, and can also be used to show improvements in terms of other expectation-based reformulations of the original problem. Therefore, our results can be seen as an alternative way to IGO, to prove that an algorithm achieves a quantile improvement result. The introduced divergence-based conditions are more general, in the sense that IGO algorithms satisfy them, and our results further allow to predict the magnitude of the quantile improvement from the decrease in divergence. Our divergence-based conditions also allow to cover more general families of proposals than exponential families, including mixtures or heavy-tailed distributions.

In our proofs, we leveraged existing results from statistics and machine learning, related to divergence minimization in the context of variational inference. This connection between the two fields opens new perspectives for the design and study of black-box global optimization algorithms. Future works could exploit this connection to use more complex proposal distribution and adaptation strategies.

## References

- [1] A. Ahrari, K. Deb, and M. Preuss. Multimodal optimization by covariance matrix self-adaptation evolution strategy with repelling subpopulations. *Evolutionary Computation*, 25(3):439–471, 2017.
- [2] 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*, 26(3):527–541, 2022.
- [3] 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)*, pages 42–51, 2012.
- [4] Y. Akimoto, Y. Nagata, I. Ono, and S. Kobayashi. Theoretical foundation for CMA-ES from information geometry perspective. *Algorithmica*, 64(4):698–716, 2012.
- [5] Y. Akimoto and Y. Ollivier. Objective improvement in information-geometric optimization. In *Proceedings of the Conference on Foundations of Genetic Algorithms (FOGA)*, pages 1–10, 2013.
- [6] S. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.
- [7] O. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. John Wiley & Sons, Ltd, 2014.

- [8] H.-G. Beyer. Convergence analysis of evolutionary algorithms that are based on the paradigm of information geometry. *Evolutionary Computation*, 22(4):679–709, 2014.
- [9] 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)*, pages 189–190, 2022.
- [10] M. F. Bugallo, V. Elvira, L. Martino, D. Luengo, J. Míguez, and P. M. Djuric. Adaptive importance sampling: The past, the present, and the future. *IEEE Signal Process. Mag.*, 34(4):60–79, 2017.
- [11] O. Cappé, R. Douc, A. Guillin, J. M. Marin, and C. P. Robert. Adaptive importance sampling in general mixture classes. *Statistics and Computing*, 18:447–459, 2008.
- [12] J. Ceberio and V. Santucci. Model-based gradient search for permutation problems. *ACM Transactions on Evolutionary Learning and Optimization*, 3:1–35, 2023.
- [13] K. Daudel, R. Douc, and F. Roueff. Monotonic alpha-divergence minimisation for variational inference. *Journal of Machine Learning Research*, 24(62):1–76, 2023.
- [14] 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)*, pages 393–400, 2010.
- [15] T. Guilmeau, N. Branchini, E. Chouzenoux, and V. Elvira. Adaptive importance sampling for heavy-tailed distributions via  $\alpha$ -divergence minimization. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 3871–3879, 2024.
- [16] N. Hansen. The CMA evolution strategy: A tutorial. <https://arxiv.org/abs/1604.00772>, 2023.
- [17] N. Hansen, D. V. Arnold, and A. Auger. *Evolution strategies*. Springer, 2015.
- [18] N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001.
- [19] 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*, 11(8):5755–5769, 2011.
- [20] X. He, Z. Zheng, and Y. Zhou. MMES: Mixture model-based evolution strategy for large-scale optimization. *IEEE Transactions on Evolutionary Computation*, 25(2):320–333, 2021.
- [21] L. Hernando, A. Mendiburu, and J. A. Lozano. Characterising the rankings produced by combinatorial optimisation problems and finding their intersections. In *Proceedings of the Conference on Genetic and Evolutionary Computation (GECCO)*, pages 266–273, 2019.
- [22] M. E. Khan and D. Nielsen. Fast yet simple natural-gradient descent for variational inference in complex models. In *International Symposium on Information Theory and Its Application (ISITA)*, pages 31–35, 2018.
- [23] D. Kroese, S. Porotsky, and R. Rubinstein. The cross-entropy method for continuous multi-extremal optimization. *Methodology and Computing in Applied Probability*, 8:383–407, 2006.
- [24] P. Larrañaga. *A Review on Estimation of Distribution Algorithms*. Springer, 2002.

- [25] 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)*, pages 150–162, 2015.
- [26] 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)*, pages 713–720, 2017.
- [27] Y. Ollivier, L. Arnold, A. Auger, and N. Hansen. Information-geometric optimization algorithms: a unifying picture via invariance principles. *Journal of Machine Learning Research*, 18(18):1–65, 2017.
- [28] Y. Polyanskiy and Y. Wu. Information theory: From coding to learning. <https://people.lids.mit.edu/yp/homepage/papers.html>, 2023.
- [29] L. M. Rios and N. V. Sahinidis. Derivative-free optimization: a review of algorithms and comparison of software implementations. *Journal of Global Optimization*, 56:1247–1293, 2013.
- [30] 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)*, pages 3409–3416, 2016.
- [31] 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)*, pages 845–852, 2011.
- [32] T. van Erven and P. Harremoës. Rényi divergence and kullback-leibler divergence. *IEEE Transactions of Information Theory*, 60(7):3797–3820, 2014.
- [33] D. Wierstra, T. Schaul, T. Glasmachers, Y. Sun, J. Peters, and J. Schmidhuber. Natural evolution strategies. *Journal of Machine Learning Research*, 15:949–980, 2014.
- [34] D. Wierstra, T. Schaul, J. Peters, and J. Schmidhuber. Fitness expectation-maximization. In *Proceedings of the Conference on Parallel Problem Solving from Nature (PPSN)*, pages 337–346, 2008.