On variational inference and maximum likelihood estimation with the λ-exponential family

Thomas Guilmeau, Emilie Chouzenoux, Víctor Elvira

On variational inference and maximum likelihood estimation with the λ-exponential family

Thomas Guilmeau[1,a,*], Emilie Chouzenoux[1,b], and Víctor Elvira[2]

[1]Université Paris-Saclay, CentraleSupélec, INRIA, CVN, France
[a] thomas.guilmeau@inria.fr
[b] emilie.chouzenoux@centralesupelec.fr
[2]School of Mathematics, University of Edinburgh, United Kingdom
victor.elvira@ed.ac.uk

Abstract

The λ-exponential family has recently been proposed to generalize the exponential family. While the exponential family is well-understood and widely used, this is not the case yet for the λ-exponential family. However, many applications require models that are more general than the exponential family, and the λ-exponential family is often a good alternative. In this work, we propose a theoretical and algorithmic framework to solve variational inference and maximum likelihood estimation problems over the λ-exponential family. We give new sufficient optimality conditions for variational inference problems. Our conditions take the form of generalized moment-matching conditions and generalize existing similar results for the exponential family. We exhibit novel characterizations of the solutions of maximum likelihood estimation problems, that recover optimality conditions in the case of the exponential family. For the resolution of both problems, we propose novel proximal-like algorithms that exploit the geometry underlying the λ-exponential family. These new theoretical and methodological insights are tested on numerical examples, showcasing their usefulness and interest, especially on heavy-tailed target distributions.

1 Introduction

Variational inference and maximum likelihood estimation are two classes of statistical problems arising in many applications. In variational inference, one aims at approaching an intractable target distribution of interest by a distribution from a family of (usually parametric) approximating densities. This is done by minimizing a discrepancy measure, such as the Kullback-Leibler [27] or the Rényi [42] divergence, between the target distribution and its approximation over the approximating family. In maximum likelihood estimation, one gets data samples, selects a parametric model to account for the unknown data-generating distribution, and then searches for the parameter maximizing the model likelihood over the data samples.

Keywords. Variational inference, maximum likelihood estimation, Rényi divergence, λ-exponential family, generalized subdifferential, heavy-tailed distribution.

2020 Mathematics Subject Classification. Primary: 62F99, 62B11, 49K10; Secondary: 90C26.

T.G. and E.C. acknowledge support from the ERC Starting Grant MAJORIS ERC-2019-STG-850925. The work of V. E. is supported by the *Agence Nationale de la Recherche* of France under PISCES (ANR-17-CE40-0031-01), the Leverhulme Research Fellowship (RF-2021-593), and by ARL/ARO under grant W911NF-22-1-0235.

$*$ Corresponding author: Thomas Guilmeau.

These two optimization tasks are deeply related as maximum likelihood estimation is equivalent to minimizing a Kullback-Leibler divergence in the large number of samples limit [48].

In variational inference and maximum likelihood estimation, a popular choice for the approximating family is the exponential family [6]. The exponential family is a family of probability distributions indexed by a finite-dimensional parameter, with the parameter appearing in the definition of the density through a scalar product with a sufficient statistics. Many well-known families of distributions can be written as instances of the exponential family, such as the Gaussian distributions. The exponential family benefits from numerous theoretical properties, many of them coming from convex analysis [6]. For instance, the exponential family contains the distributions with a sufficient statistics, a fact known as the Pitman-Koopman-Darmois theorem [44]. This implies that the maximum likelihood estimator over the exponential family is reached when a moment-matching condition on sufficient statistics is satisfied [12]. In variational inference, minimizing the Kullback-Leibler divergence over the exponential family leads to optimality conditions which can also be written as a moment-matching condition on sufficient statistics (see [10, 13, 46]). Thus, variational inference and maximum likelihood problems over the exponential family are both solved when moment-matching conditions are satisfied.

The exponential family also benefits from many geometric properties [2, 35]. Indeed, the Kullback-Leibler divergence between two distributions from the exponential family can be seen as the Bregman divergence induced by the log-partition function of the family. Bregman divergences generalize the Euclidean distance, and can be plugged in optimization algorithms, leading for instance to the so-called Bregman proximal gradient algorithms [43]. These properties can be leveraged to design more efficient algorithms over the exponential family in many settings [5, 22, 25, 20].

Despite the advantages of using the exponential family, there exists some contexts where it is better to use other types of distributions. For instance, the exponential family cannot represent physical systems governed by large fluctuations, such as cold atoms in optical lattices [16]. In ecology, using Gaussian kernels to account for the diffusion of a population does not allow to represent species invading a territory with increasing speed, while heavier-tailed kernels can [26]. In signal processing and statistics, Student priors have been used to enforce signal sparsity [15] or for logistic regression [19], and Cauchy distributions to model noise [28]. Using Student distributions rather than Gaussian ones have also been proven beneficial to cluster heavy-tailed data in [38], while Student distributions have been used successfully in importance sampling [13, 17, 47].

Motivated by these situations, several works generalize the exponential family and extend its properties. These generalizations are often indexed by a scalar parameter, with the value zero corresponding to the exponential family. One can mention the q-exponential family studied in [3], the $\mathcal{F}^{(\alpha)}$-family and $\mathcal{F}^{(-\alpha)}$-family of [49], and the unifying λ-exponential family studied in [50]. We focus on the latter in this paper as it recovers the two former ones. The densities of distributions from the λ-exponential family are similar to those from the standard exponential family, but the scalar product between the parameter and what plays the role of sufficient statistics is replaced by a non-linear coupling. Instances of the λ-exponential family are the Student distributions (including Cauchy distributions), the Student Wishart distributions [4], the β-Gaussian distributions [32], or the Dirichlet perturbation model [50]. The geometric properties of these families have also been studied in the above papers. More precisely, and similarly to the situation for the standard exponential family, the authors of [50] established strong links between the λ-exponential family, the Rényi divergence, and a quantity that generalizes the Bregman divergence. Note that while the exponential family is studied using convex duality, the authors of [50] proposed the theory of λ-duality to study the λ-exponential family.

Generalizations of the exponential family have already been used in several tasks in statistics. Let us mention the creation of paths between distributions [33], neural attention mechanisms and regression problems with bounded noise [32], or the understanding of generative adversarial networks based on f-divergences [37, 36]. Let us also mention the work of [24] in which an optimization algorithm using a

generalization of the Bregman divergence is studied and applied for maximum likelihood estimation over the λ-exponential family.

However, the λ-exponential family has been less studied than the standard exponential family. Indeed, to our knowledge, (i) variational inference problems over generalizations of the exponential family have not been studied, (ii) maximum likelihood estimation problems are usually solved within a particular λ-exponential family (see the works of [21, 4] for instance), and (iii) no algorithm exploits explicitly the geometry of these models (see [24] for an exception).

As a summary, we propose a theoretical analysis and a novel methodological framework that allows to tackle variational inference and maximum likelihood estimation problems on the λ-exponential family. Our contributions are as follows:

- (i) We give new optimality conditions for variational inference problems on the λ-exponential family that generalize the existing moment-matching conditions for the exponential family.
- (ii) We propose novel characterizations for the solutions of maximum likelihood estimation problems. We show that these are optimal conditions in the case of the exponential family, and related (in a sense we explicit) to optimal ones in the case of the λ-exponential family.
- (iii) We introduce new algorithms generalizing moment-matching to solve the considered variational inference and maximum likelihood problems, including an expectation-maximization algorithm. Our algorithms are shown to be related to proximal algorithms in the geometry induced by the Rényi divergence.
- (iv) All the aforementioned results are obtained using a novel theoretical framework to study the exponential family and the λ-exponential family based on non-convex duality. This new framework allows us to recover known results for the exponential family and to generalize them in a simple and unified way.
- (v) We illustrate numerically the behavior of our algorithms on variational inference and maximum likelihood estimation problems involving heavy-tailed distributions, showing the benefits of our novel theoretical results.

The paper is organized as follows. We present some background in Section 2. In Section 3, we state our main assumptions, an important example, and our main technical results. In Section 4, we apply these novel results to analyze, in a systematic way, variational inference and maximum likelihood estimation problems. We also propose proximal-like algorithms to solve these problems and compare the situation between the λ-exponential family and the standard exponential family. We illustrate our findings in Section 5 through numerical experiments. Finally, we present future research developments and conclude in Section 6.

2 Preliminaries

We introduce some preliminary background on divergences [45], the λ-exponential family [50], and convex analysis [8] that we will use throughout the rest of the paper.

Notation

We introduce some notation that will hold throughout the paper. $\mathcal{H}$ is a real Hilbert space in finite dimension with scalar product $\langle \cdot, \cdot \rangle$. Given a natural number d, $\mathcal{S}_+^d$ denotes the set of positive semi-definite matrices in dimension d, $\mathcal{S}_{++}^d$ denotes the set of positive definite matrices in dimension d, and $\mathcal{S}_{--}^d$ denotes the set of matrices obtained as the opposite of matrices in $\mathcal{S}_{++}^d$. Finally, $\mathbb{R}_{++}$ is the set of positive real numbers and $\bar{\mathbb{R}}$ is the extended real line.

The set $\mathcal{X}$ with its Borel algebra is a measurable space, m is a measure on this space, and $\mathcal{P}(\mathcal{X}, m)$ is the set of probability measures on this space which admit a density with respect to m. We will often use the same notation for a distribution of $\mathcal{P}(\mathcal{X}, m)$ and its density. The letter S will be used to denote the support of a distribution. The restriction of a probability density $q \in \mathcal{P}(\mathcal{X}, m)$ to a set $\mathcal{Y}$ is denoted by $q_{|\mathcal{Y}}$. We denote the Lebesgue measure by dx. The family of Gaussian distributions in dimension d will be denoted by $\mathcal{G}^d$, and the family of Student distributions in dimension d with degree of freedom parameter $\nu > 0$ by $\mathcal{T}_\nu^d$ (the formal definition is recalled in the remaining).

Generally, we used sub-scripts to describe the dependence over a scalar parameter, an index, or an iteration count, while we used super-scripts to denote escort distributions or conjugate functions, two notions that will be defined later on in the paper.

2.1 Entropies and statistical divergences

Let us introduce statistical notions that we will leverage through the rest of the paper. The first one is the entropy of a probability distribution, which is related to the information the distribution encodes.

Definition 1. Consider $\alpha > 0$, $\alpha \neq 1$, and a probability distribution $p \in \mathcal{P}(\mathcal{X}, m)$. Then the *Rényi entropy* is defined by

$$H_\alpha(p) = \frac{1}{1-\alpha} \log \left(\int p(x)^\alpha m(dx) \right). \tag{2.1}$$

When $\alpha = 1$, we define H_1 as the standard *Shannon entropy*, that is

$$H_1(p) = -\int \log\left(p(x)\right) p(x) m(dx). \tag{2.2}$$

If the integrals above do not converge, then the corresponding entropies are equal to $+\infty$.

We now introduce the Rényi and Kullback-Leibler divergences. These divergences measure the discrepancy between two probability densities. Although they are not distances, they are non-negative, and they are null if and only if the two considered densities are equal almost everywhere.

Definition 2. Consider $\alpha > 0$, $\alpha \neq 1$, and probability distributions $p_1, p_2 \in \mathcal{P}(\mathcal{X}, m)$. Then the *Rényi divergence* between p_1 and p_2 is defined by

$$RD_\alpha(p_1, p_2) = \frac{1}{\alpha - 1} \log \left(\int p_1(x)^\alpha p_2(x)^{1-\alpha} m(dx) \right). \tag{2.3}$$

When $\alpha = 1$, we define RD_1 as the *Kullback-Leibler divergence* through

$$RD_1(p_1, p_2) = KL(p_1, p_2) = \int \log\left(\frac{p_1(x)}{p_2(x)}\right) p_1(x) m(dx). \tag{2.4}$$

If these quantities are not defined, then the divergence takes the value $+\infty$.

The notations H_1 and RD_1 in Definitions 1 and 2, respectively, are motivated by the property that when $\alpha \to 1$, the Rényi entropy identifies with the Shannon entropy and the Rényi divergence with the Kullback-Leibler divergence [45].

We conclude this section by defining a transformation that, for a given probability density, leads to its so-called escort distribution, parametrized by a scalar parameter $\alpha > 0$. When $\alpha = 1$, this transformation is simply the identity (i.e., the distribution identifies with its escort).

Definition 3. Consider $\alpha > 0$ and $p \in \mathcal{P}(\mathcal{X}, m)$. The *escort probability distribution* with exponent α of p is the probability $p^{(\alpha)} \in \mathcal{P}(\mathcal{X}, m)$ defined by

$$p^{(\alpha)}(x) = \frac{1}{\int p(x)^\alpha m(dx)} p(x)^\alpha, \tag{2.5}$$

assuming the normalization constant $\int p(x)^\alpha m(dx)$ is finite.

2.2 The exponential family and the λ-exponential family

We introduce the λ-exponential family, which is a generalization of the standard exponential family. Such family is obtained by replacing the scalar product $\langle \cdot, \cdot \rangle$, in the definition of the standard exponential family, by a non-linear coupling c_λ defined as

$$c_\lambda(u, v) = \frac{1}{\lambda} \log(1 + \lambda \langle u, v \rangle), \ \forall u, v \in \mathcal{H}. \tag{2.6}$$

Since $c_\lambda(u, v) \xrightarrow{\lambda \to 0} \langle u, v \rangle$, we denote $c_0(u, v) = \langle u, v \rangle$, $\forall u, v \in \mathcal{H}$.

We now turn to the definition of the λ-exponential family, following the formalism of [50]. This definition encompasses the definition of the standard exponential family. We set the conventions that $\log(s) = -\infty$ when $s \leq 0$ and $\exp(-\infty) = 0$. We also give examples in Figure 1 of densities from the λ-exponential family for different values of λ.

Definition 4. Consider $\lambda \in \mathbb{R}$. The *λ-exponential family* $\mathcal{Q}_\lambda$ *with sufficient statistics* T *and base measure* m is the family $\mathcal{Q}_\lambda = \{q_\vartheta \in \mathcal{P}(\mathcal{X}, m), \ \vartheta \in \operatorname{dom} \varphi_\lambda\}$, with

$$q_\vartheta(x) = \exp\left(c_\lambda(\vartheta, T(x)) - \varphi_\lambda(\vartheta)\right), \tag{2.7}$$

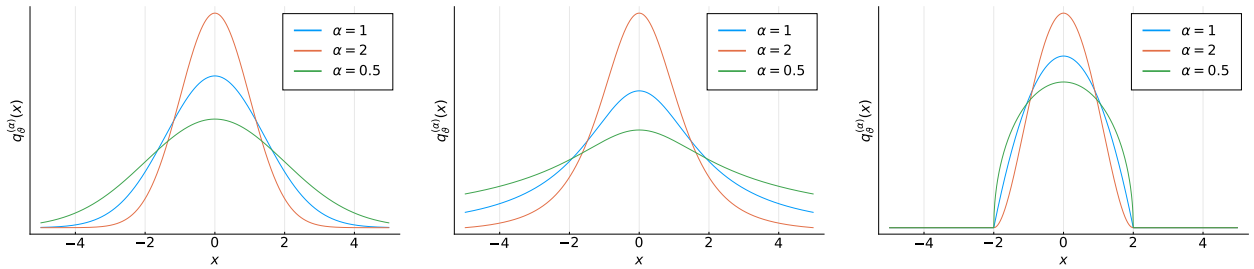
where c_λ is the non-linear coupling defined in Equation (2.6). Function φ_λ in (2.7) is the λ-log-partition function, defined for any $\vartheta \in \operatorname{dom} \varphi_\lambda$ by

$$\varphi_\lambda(\vartheta) = \log\left(\int \exp(c_\lambda(\vartheta, T(x)) m(dx)\right). \tag{2.8}$$

The support of q_ϑ is the set $S_\vartheta = \{x \in \mathcal{X}, \ 1 + \lambda \langle \vartheta, T(x) \rangle > 0\}$. When $\alpha = 1 - \lambda$ is positive, we introduce, for any $\vartheta \in \operatorname{dom} \varphi_\lambda$, the entropy function

$$\psi_\lambda(\vartheta) = -H_\alpha(q_\vartheta), \ \forall \vartheta \in \operatorname{dom} \varphi_\lambda. \tag{2.9}$$

Remark 1. When $\lambda = 0$, we have $c_0(\cdot, \cdot) = \langle \cdot, \cdot \rangle$, and we recover in (2.7) the standard notion of exponential family, that is $q_\vartheta(x) = \exp(\langle \vartheta, T(x) \rangle - \varphi_0(\vartheta))$. In this case, the family is denoted by $\mathcal{Q}$ and we drop the subscript λ.



(a) $\lambda = 0$ (b) $\lambda = -1$ (c) $\lambda = 1$

Figure 1: Plots of the densities $q_\vartheta^{(\alpha)}$, for the λ-exponential family obtained with sufficient statistics $T(x) = x^2$ and $\vartheta = 2$, for different values of $\lambda \in \{-1, 0, 1\}$ and $\alpha \in \{0.5, 1, 2\}$. When $\lambda = 0$, we recover a Gaussian distribution, while we obtain distributions with respectively heavier tails for $\lambda = -1$ and lighter tails for $\lambda > 0$. Also, values of $\alpha > 1$ lighten the tails while values $\alpha < 1$ make them heavier.

2.3 λ-duality and proximal operators

We now introduce elements of the concept of λ-duality, that will play an important role in our analysis of the considered optimization problems and the derivation of their optimality conditions.

The λ-duality, initially introduced in [49, 50], can be viewed as an extension of the usual convex duality [8] (sometimes called Fenchel-Rockafellar duality). Let us remind that the convex duality relies on a coupling between primal and dual variables through the scalar product $\langle \cdot, \cdot \rangle$. This leads in particular to the notion of convex (or Fenchel) conjugate of a function $f : \mathcal{H} \to \bar{\mathbb{R}}$, defined at $v \in \mathcal{H}$ by

$$f^*(v) = \sup_{u \in \mathcal{H}} \langle u, v \rangle - f(u). \tag{2.10}$$

Such conjugate can then used to define the subgradient of function f, by saying that v is a subgradient of f at u, denoted by $v \in \partial f(u)$, if and only if

$$f^*(v) + f(u) = \langle u, v \rangle. \tag{2.11}$$

One can then verify that $v \in \partial f(u)$ is equivalent to having that

$$f(u') \geq f(u) + \langle v, u' - u \rangle, \, \forall u' \in \operatorname{dom} f, \tag{2.12}$$

meaning that the right-hand side is a linear tangent minorant of f. The subdifferential can also be used to state optimality conditions through the Fermat rule [8].

The λ-duality is constructed by replacing the scalar product of $\mathcal{H}$, appearing for instance in (2.10), by the non-linear coupling $c_\lambda(\cdot, \cdot)$ introduced in Equation (2.6). This leads to the definition of several mathematical notions, given hereafter.

Definition 5. Consider a proper function $f : \mathcal{H} \to \bar{\mathbb{R}}$ and $\lambda \in \mathbb{R}$.

(i) We define its c_λ-*conjugate* $f^{c_\lambda} : \mathcal{H} \to \bar{\mathbb{R}}$ by

$$f^{c_\lambda}(v) = \sup_{u \in \mathcal{H}} c_\lambda(u, v) - f(u). \tag{2.13}$$

(ii) We say that $v \in \mathcal{H}$ is a c_λ-*subgradient* of f at u and belongs to the c_λ-*subdifferential* of f at u, denoted by $\partial^{c_\lambda} f(u)$ if and only if

$$f^{c_\lambda}(v) + f(u) = c_\lambda(u, v). \tag{2.14}$$

As already mentioned, the above definitions correspond to generalizations of convex analysis theory. Similar constructions were achieved for instance in [14, 29] using the so-called CAPRA couplings, in [18] to study evenly convex functions, or in [39] for general couplings in optimal transport. The standard notions of convexity have also been generalized by considering alternative notions of subgradients, such as in [9]. Let us relate this latter work to the notions introduced in Definition 5. Consider $f : \mathcal{H} \to \bar{\mathbb{R}}$, such that $v \in \partial^{c_\lambda} f(u)$. Equation (2.14) can be rewritten in the following way:

$$\begin{aligned} & f(u) + f^{c_\lambda}(v) = c_\lambda(u, v) \\ \Leftrightarrow\, & c_\lambda(u, v) - f(u) \geq c_\lambda(u', v) - f(u'), \, \forall u' \in \operatorname{dom} f \\ \Leftrightarrow\, & f(u') \geq f(u) + c_\lambda(u', v) - c_\lambda(u, v), \, \forall u' \in \operatorname{dom} f. \end{aligned}$$

This shows that $c_\lambda(\cdot, v)$ is a subgradient of f at u in the sense of the framework of abstract convexity, as outlined in [9] for instance.

Let us emphasize that Definition 5 does not focus on the same objects than the ones in the study of [49, 50]. The latter also relies on λ-duality, but the so-called λ-gradient of f is introduced before showing

the fulfillment of Equation (2.14). This requires differentiability and regularity assumptions on f. We take the opposite direction in our Definition 5, as we define the c_λ-subdifferential assuming only the properness of f. As a consequence, we lose explicit expressions for c_λ-subgradients, while the λ-gradients in [49, 50] could be computed from the gradients of f. We will show in the following that Definition 5 is sufficient to solve the considered optimization problems and that it is actually possible to exhibit c_λ-subgradients in our cases of interest, under mild hypotheses that are easy to check.

The above elements of λ-duality will be used subsequently to solve optimization problems of variational inference and maximum likelihood over a λ-exponential family providing explicit optimality conditions. We will also rely on proximal operators [8], which are an essential tool for the algorithmic resolution of the considered problems. In order to fit the geometry induced by the λ-exponential family, we will rely on the Rényi proximal operator defined below.

Definition 6. Consider $\lambda \in \mathbb{R}$ such that $\alpha = 1 - \lambda$ is positive, the family $\mathcal{Q}_\lambda$ with λ-log-partition φ_λ, and an objective function $f : \mathcal{H} \to \bar{\mathbb{R}}$. Then the *Rényi proximal operator* of f with step-size $\tau > 0$ is defined by

$$\operatorname{prox}_\tau^f(\vartheta') = \underset{\vartheta \in \operatorname{dom} \varphi_\lambda}{\arg\min}\ f(\vartheta) + \frac{1}{\tau} RD_\alpha(q_{\vartheta'}, q_\vartheta). \tag{2.15}$$

When $\lambda = 0$, i.e., the λ-exponential family recovers the standard exponential family, the Rényi divergence appearing in the definition of $\operatorname{prox}_\tau^f$ reduces to the Kullback-Leibler divergence [35]. In this case, $\operatorname{prox}_\tau^f$ can be seen as a Bregman proximal operator [7] (see also [20] for some examples of explicit Bregman proximal operators in the case of the exponential family). Note also that in [24], a proximal operator in the geometry defined by the Rényi divergence is mentioned but not studied.

In the following, we will refer to the operator (2.15) simply as proximal operator, except otherwise stated.

3 Novel results on the λ-exponential family

In this section, we present a first set of novel results about the λ-exponential family, using the notion of λ-duality introduced in Definition 5. We first state our main assumptions and recover with our framework some known results including a key reformulation of the Rényi divergence in Section 3.1. We then discuss the important example of Student distributions in Section 3.2, before presenting in Section 3.3 new technical optimality conditions that we will apply in subsequent sections to statistical problems.

3.1 Assumptions and properties of the λ-exponential family

We now introduce our main assumptions and recover known results about the λ-exponential family under mild hypotheses, including a rewriting of the Rényi divergence in a way that will be crucial to solve statistical inference problems later on.

Assumption 1. The λ-exponential family $\mathcal{Q}_\lambda$ is such that $\alpha = 1 - \lambda$ is positive and the function φ_λ in (2.8) is proper.

Assumption 1 implies in particular that $\operatorname{dom} \varphi_\lambda \neq \emptyset$ and that any $\vartheta \in \operatorname{dom} \varphi_\lambda$ is such that q_ϑ is well-defined and belongs to $\mathcal{P}(\mathcal{X}, m)$. Note also that under Assumption 1, φ_λ cannot take the value $-\infty$, meaning in particular that, for any $\vartheta \in \operatorname{dom} \varphi_\lambda$, $S_\vartheta \neq \emptyset$.

Definition 7. Consider the λ-exponential family $\mathcal{Q}_\lambda$, the scalar $\alpha = 1 - \lambda$, and a probability density $p \in \mathcal{P}(\mathcal{X}, m)$. We say that p is *q_ϑ-compatible* for $q_\vartheta \in \mathcal{Q}_\lambda$ if the restriction of p to the support of q_ϑ, denoted by S_ϑ, is such that $\int p_{|S_\vartheta}(x)^\alpha m(dx) \in (0, +\infty)$ and $\int T(x) p_{|S_\vartheta}(x)^\alpha m(dx)$ have finite components. If p is q_ϑ-compatible for any $q_\vartheta \in \mathcal{Q}_\lambda$, then we say that p is *$\mathcal{Q}_\lambda$-compatible*.

The notion of compatibility in Definition 7 is a technical condition that allows in particular to ensure the following well-posedness result.

Lemma 1. *Consider the λ-exponential family $\mathcal{Q}_\lambda$, and $q_\vartheta \in \mathcal{Q}_\lambda$. Assume that Assumption 1 is satisfied and consider $\vartheta \in \operatorname{dom}\varphi_\lambda$ and $p \in \mathcal{P}(\mathcal{X}, m)$. If p is q_ϑ-compatible, then $c_\lambda(\vartheta, p^{(\alpha)}_{|S_\vartheta}(T)) \in \mathbb{R}$.*

Proof. If $\lambda = 0$, $c_\lambda(\vartheta, p^{(\alpha)}_{|S_\vartheta}(T)) = \langle \vartheta, p_{|S_\vartheta}(T)\rangle$ and the result is straightforward. Now, consider $\lambda \neq 0$. The support of q_ϑ is the set $S_\vartheta = \{x \in \mathcal{X},\, 1 + \lambda\langle \vartheta, T(x)\rangle > 0\}$. Then we can compute

$$1 + \lambda\langle \vartheta, p^{(\alpha)}_{|S_\vartheta}(T)\rangle = \int (1 + \lambda\langle \vartheta, T(x)\rangle) p^{(\alpha)}_{|S_\vartheta}(x) m(dx). \tag{3.1}$$

We get from the compatibility assumption that $p^{(\alpha)}_{|S_\vartheta}$ is well-defined and belongs to $\mathcal{P}(\mathcal{X}, m)$. This ensures that the quantity in (3.1) is positive. Also by assumption, $p^{(\alpha)}_{|S_\vartheta}(T)$ is well-defined, ensuring that the quantity in (3.1) is also finite, hence the result. □

We now introduce an extra assumption stating that all the densities $q_\vartheta \in \mathcal{Q}_\lambda$ share the same support. In [50], this property is also assumed and called the support condition. This assumption ensures that $q^{(\alpha)}_\vartheta(T)$ is well-defined for any $\vartheta \in \operatorname{dom}\varphi_\lambda$ and $\alpha = 1 - \lambda$ as we will show.

Assumption 2. There exists a non-empty set $S_\lambda \subset \mathcal{X}$ such that

$$S_\vartheta = S_\lambda,\ \forall \vartheta \in \operatorname{dom}\varphi_\lambda, \tag{3.2}$$

with S_ϑ being the support of q_ϑ. Moreover, every $q_\vartheta \in \mathcal{Q}_\lambda$ is $\mathcal{Q}_\lambda$-compatible.

We now state a property that links the coupling c_λ, the log-partition function φ_λ, and the Rényi divergence RD_α. This technical property is used in the proof of a Rényi entropy maximization property in [50], and we will exploit it further in our subsequent developments.

Proposition 1. *Consider the λ-exponential family $\mathcal{Q}_\lambda$ under Assumption 1 with $\alpha = 1 - \lambda$. Consider a probability distribution $p \in \mathcal{P}(\mathcal{X}, m)$. For any $q_\vartheta \in \mathcal{Q}_\lambda$, $RD_\alpha(p, q_\vartheta)$ satisfies*

$$RD_\alpha(p, q_\vartheta) = \varphi_\lambda(\vartheta) - c_\lambda(\vartheta, p^{(\alpha)}_{|S_\vartheta}(T)) - H_\alpha(p_{|S_\vartheta}). \tag{3.3}$$

Further, under Assumptions 1 and 2, for every $\vartheta' \in \operatorname{dom}\varphi_\lambda$,

$$RD_\alpha(q_{\vartheta'}, q_\vartheta) = \varphi_\lambda(\vartheta) - c_\lambda(\vartheta, q^{(\alpha)}_{\vartheta'}(T)) + \psi_\lambda(\vartheta'). \tag{3.4}$$

Proof. When $\lambda = 0$, recall that q_ϑ has full support. In this case, we have

$$\begin{aligned} RD_\alpha(p, q_\vartheta) &= KL(p, q_\vartheta) \\ &= \int \log\left(\frac{p(x)}{q_\vartheta(x)}\right) p(x) m(dx), \end{aligned}$$

from which we can straightforwardly obtain the result using that $c_\lambda(\cdot,\cdot) = \langle \cdot, \cdot\rangle$ for $\lambda = 0$ and $\alpha = 1$.

For $\lambda \neq 0$, we compute the Rényi divergence RD_α (defined in Definition 2) between p and q_ϑ. Using the definitions of q_ϑ given in Definition 4, of the Rényi entropy H_α given in Definition 1, and of the coupling c_λ

from Equation (2.6), we obtain the following result.

$$\begin{aligned}
RD_\alpha(p, q_\vartheta) &= \frac{1}{\alpha-1}\log\left(\int p_{|S_\vartheta}(x)^\alpha q_\vartheta(x)^{1-\alpha} m(dx)\right) \\
&= \frac{1}{\alpha-1}\log\left(\int p_{|S_\vartheta}(x)^\alpha \exp((1-\alpha)c_\lambda(\vartheta, T(x)) - (1-\alpha)\varphi_\lambda(\vartheta))\nu(dx)\right) \\
&= \frac{1}{\alpha-1}\log\left(\int p_{|S_\vartheta}(x)^\alpha \exp(\lambda c_\lambda(\vartheta, T(x)))\nu(dx)\right) + \varphi_\lambda(\vartheta) \\
&= \frac{1}{\alpha-1}\log\left(\int p_{|S_\vartheta}(x)^\alpha (1+\lambda\langle\vartheta, T(x)\rangle)\nu(dx)\right) + \varphi_\lambda(\vartheta) \\
&= \frac{1}{\alpha-1}\log\left(\int p_{|S_\vartheta}(x)^\alpha \nu(dx)\left(1+\lambda\langle\vartheta, p^{(\alpha)}_{|S_\vartheta}(T)\rangle\right)\right) + \varphi_\lambda(\vartheta) \\
&= \frac{1}{\alpha-1}\log\left(\int p_{|S_\vartheta}(x)^\alpha \nu(dx)\right) + \frac{1}{\alpha-1}\log\left(1+\lambda\langle\vartheta, p^{(\alpha)}_{|S_\vartheta}(T)\rangle\right) + \varphi_\lambda(\vartheta) \\
&= -H_\alpha(p_{|S_\vartheta}) - c_\lambda(\vartheta, p^{(\alpha)}_{|S_\vartheta}(T)) + \varphi_\lambda(\vartheta),
\end{aligned}$$

which proves the first part of the property in Equation (3.3). The second part in Equation (3.4) follows using the assumptions and Equation (2.9). □

We establish a second property, describing the λ-duality objects associated to $\mathcal{Q}_\lambda$ in terms of moments and entropy and recovering the results of [50] in our framework.

Proposition 2. *Suppose that Assumptions 1 and 2 are satisfied. Then, for every $\vartheta \in \operatorname{dom}\varphi_\lambda$,*

$$\varphi_\lambda^{c_\lambda}(q_\vartheta^{(\alpha)}(T)) = \psi_\lambda(\vartheta), \tag{3.5}$$

$$q_\vartheta^{(\alpha)}(T) \in \partial^{c_\lambda}\varphi_\lambda(\vartheta). \tag{3.6}$$

Proof. We denote $\eta = q_\vartheta^{(\alpha)}(T)$ for sake of concision. We begin with the proof for (3.5). Using the result of Proposition 1 and the non-negativity of the Rényi divergence,

$$c_\lambda(\vartheta', \eta) - \varphi_\lambda(\vartheta') \leq \psi_\lambda(\vartheta), \tag{3.7}$$

with equality if and only if $\vartheta' = \vartheta$. This shows that $\varphi_\lambda^{c_\lambda}(\eta) = \psi_\lambda(\vartheta)$ following Equation (2.13), hence the result.

We now turn to the proof for (3.6). Consider the rewriting of the Rényi divergence from Proposition 1:

$$0 = \varphi_\lambda(\vartheta) - c_\lambda(\vartheta, q_\vartheta^{(\alpha)}(T)) + \psi_\lambda(\vartheta). \tag{3.8}$$

Then, by Equation (3.5), Equation (3.8) can be written as $\varphi_\lambda(\vartheta) + \varphi_\lambda^{c_\lambda}(\eta) = c_\lambda(\vartheta, \eta)$. This concludes the proof, using Equation (2.14). □

Remark 2. Assumption 2 and Proposition 2 ensure that, for every $\vartheta \in \operatorname{dom}\varphi_\lambda$, $q_\vartheta^{(\alpha)}(T)$ is well-defined and thus that $\partial^{c_\lambda}\varphi_\lambda(\vartheta)$ is non-empty. This can be viewed as a form of convexity result on φ_λ. Indeed, for $\lambda = 0$, which corresponds to the classical Fenchel duality theory, having a non-empty subdifferential at every point of $\operatorname{dom}\varphi$ implies that $\varphi(\vartheta) = \varphi^{**}(\vartheta)$ on $\operatorname{dom}\varphi$ [8, Proposition 16.4]. This last equality shows that φ is equal to its biconjugate and hence that it is convex.

3.2 The example of Student distributions

We now show that the Student distributions can be seen as a particular example of the λ-exponential family that satisfies the assumptions outlined in Section 3 and whose escort distributions have easily computable moments. This means that Student distributions will be an importance use-case of our coming theoretical results of Section 4, as we will illustrate on numerical experiments in Section 5. Student distributions form an important class of distributions arising in several applications from statistics and signal processing [38, 15, 19, 13, 28, 47].

Definition 8. Consider the family of *multivariate Student distributions* on $\mathbb{R}^d$ with fixed degree of freedom parameter $\nu > 0$. We denote this family by $\mathcal{T}_\nu^d$. Densities with respect to the Lebesgue measure are of the form

$$q_{\mu,\Sigma}(x) = \frac{1}{Z_\nu} \det(\Sigma)^{-\frac{1}{2}} \left(1 + \frac{1}{\nu}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)^{-\frac{\nu+d}{2}}, \forall x \in \mathbb{R}^d \tag{3.9}$$

with location parameter $\mu \in \mathbb{R}^d$, scale matrix $\Sigma \in \mathcal{S}_{++}^d$, and normalization constant $Z_\nu = \frac{\Gamma(\nu/2)\nu^{d/2}\pi^{d/2}}{\Gamma((\nu+d)/2)}$ where Γ denotes the Gamma function.

The degree of freedom parameter $\nu > 0$ controls the tail behavior of the distributions. In particular, higher values of ν lead to distributions in $\mathcal{T}_\nu^d$ with lighter (but still heavy) tails, with the limit $\nu \to +\infty$ corresponding to the family of Gaussian distributions, which is an example of the exponential family. On the contrary, distributions in $\mathcal{T}_\nu^d$ for low ν have heavier tails, an example being that $\mathcal{T}_1^d$ is the family of multivariate Cauchy distributions. In particular, distributions in $\mathcal{T}_\nu^d$ have well-defined first order moments if $\nu > 1$ and well-defined second order moments if $\nu > 2$.

The next proposition shows that the λ-exponential family, with sufficient statistics being the first and second order moments, is the family of Student distribution when $\lambda < 0$ and that it satisfies Assumption 1. We further compute the escort moments of Student distributions, which are c_λ-subgradients of φ_λ. We also compute the Rényi entropy of Student distributions, which is the c_λ-conjugate of φ_λ. Finally, we describe the distributions that are compatible with the Student distributions (following Definition 7) and show that Student distributions satisfy Assumption 2.

Proposition 3. *Consider the Student family $\mathcal{T}_\nu^d$.*

(i) The family $\mathcal{T}_\nu^d$ is an instance of the λ-exponential family (see Equation (2.7)) for $\lambda = -\frac{2}{\nu+d}$ and with sufficient statistics $T(x) = (x, xx^\top)$. Its natural parameters are $\vartheta = (\vartheta_1, \vartheta_2) \in \mathbb{R}^d \times \mathcal{S}_{--}^d$. It satisfies Assumption 1 and $\operatorname{dom}\varphi_\lambda = \{(\vartheta_1, \vartheta_2) \in \mathbb{R}^d \times \mathcal{S}_{--}^d,\ 2(\nu+d) + \vartheta_1^\top \vartheta_2^{-1} \vartheta_1 > 0\}$.

(ii) Consider $\alpha = 1 + \frac{\nu+d}{2}$. Then, for any $q_{\mu,\Sigma} \in \mathcal{T}_\nu^d$, $q_{\mu,\Sigma}^{(\alpha)}(T) = (q_{\mu,\Sigma}^{(\alpha)}(x), q_{\mu,\Sigma}^{(\alpha)}(xx^\top))^\top$ is such that

$$\begin{cases} q_{\mu,\Sigma}^{(\alpha)}(x) = \mu, \\ q_{\mu,\Sigma}^{(\alpha)}(xx^\top) = \Sigma + \mu\mu^\top. \end{cases} \tag{3.10}$$

The mapping $\vartheta \longmapsto q_\vartheta^{(\alpha)}(T)$ is a bijection from $\operatorname{dom}\varphi_\lambda$ to $\mathbb{R}^d \times \mathcal{S}_{++}^d$. Moreover, $\psi_\lambda(\vartheta) = \frac{1}{2}\log\det(\Sigma) + C$, where C is a scalar depending only on ν and d.

(iii) The $\mathcal{T}_\nu^d$-compatible distributions are the probability densities $p \in \mathcal{P}(\mathcal{X}, dx)$ such that $p^{(\alpha)}$ has finite first and second order moments. The family $\mathcal{T}_\nu^d$, seen as a λ-exponential family, satisfies Assumption 2.

Proof. The proof is deferred to Appendix A. □

Remark 3. Proposition 3 generalizes analogous results for Gaussian distributions. Indeed, Gaussian distributions in dimension d, denoted by $\mathcal{G}^d$, form an example of the exponential family with sufficient statistics $T(x) = (x, xx^\top)$, satisfying Assumptions 1 and 2. Furthermore, for any $q_{\mu,\Sigma} \in \mathcal{G}^d$ with $\mu \in \mathbb{R}^d, \Sigma \in \mathcal{S}^d_{++}$, we have

$$\begin{cases} q_{\mu,\Sigma}(x) = \mu, \\ q_{\mu,\Sigma}(xx^\top) = \Sigma + \mu\mu^\top. \end{cases} \tag{3.11}$$

Finally, the $\mathcal{G}^d$-compatible distributions are the probability densities in $\mathcal{P}(\mathcal{X}, dx)$ with finite first and second order moments. While the Gaussian case corresponds to $\lambda = 0$, we remark that the case $\lambda > 0$ corresponds to the β-Gaussians distributions discussed in [32]. However, these distributions do not satisfy the support condition of Assumption 2.

We now establish some novel properties that state how two families of Student distributions with different degree of freedom parameters relate to each other, including the computation of some escort moments and compatibility conditions. This provides a mechanism to construct an escort distribution with lighter tails than the original distribution.

Proposition 4. *Let $p \in \mathcal{T}^d_{\nu_p}$ a Student distribution with dimension d, location μ_p and shape Σ_p. Set $\nu > 0$ and consider the Student family $\mathcal{T}^d_\nu$ with associated $\alpha = 1 + \frac{2}{\nu+d}$. Then, the distribution p is $\mathcal{T}^d_\nu$-compatible if and only if $\nu_p + 2\frac{\nu_p+d}{\nu+d} > 2$, and the escort probability $p^{(\alpha)}$ is a Student distribution with $\nu^{(\alpha)}$ degrees of freedom, location $\mu^{(\alpha)}$, and shape $\Sigma^{(\alpha)}$ such that*

$$\begin{cases} \nu^{(\alpha)} &= \nu_p + 2\frac{\nu_p+d}{\nu+d}, \\ \mu^{(\alpha)} &= \mu_p, \\ \Sigma^{(\alpha)} &= \frac{\nu_p}{\nu^{(\alpha)}}\Sigma_p. \end{cases} \tag{3.12}$$

Proof. By Proposition 3, $\mathcal{T}^d_\nu$ is a λ-exponential family, with $\lambda = -\frac{2}{\nu+d}$. For such setting, $\alpha = 1 - \lambda$. The compatibility property requires $p^{(\alpha)}$ to have finite first and second order moments. Consider $x \in \mathbb{R}^d$, we compute

$$\begin{aligned} p(x)^\alpha &\propto \left(1 + \frac{1}{\nu_p}(x-\mu_p)^\top \Sigma_p^{-1}(x-\mu_p)\right)^{-\left(\frac{\nu_p+d}{2}\right)\left(1+\frac{2}{\nu+d}\right)} \\ &\propto \left(1 + \frac{1}{\nu_p}(x-\mu_p)^\top \Sigma_p^{-1}(x-\mu_p)\right)^{-\frac{1}{2}\left(\nu_p + 2\frac{\nu_p+d}{\nu+d}+d\right)} \\ &\propto \left(1 + \frac{1}{\nu^{(\alpha)}}(x-\mu_p)^\top \left(\frac{\nu_p}{\nu^{(\alpha)}}\Sigma_p\right)^{-1}(x-\mu_p)\right)^{-\frac{\nu^{(\alpha)}+d}{2}}. \end{aligned}$$

We recognize that $p^{(\alpha)}$ is a Student distribution with $\nu^{(\alpha)}$ degrees of freedom, location $\mu^{(\alpha)}$, and shape $\Sigma^{(\alpha)}$, and that $p^{(\alpha)}$ has finite first and second order moments if and only if $\nu^{(\alpha)} > 2$, showing the result. □

Proposition 4 provides a systematic way to construct, from an initial distribution with possibly infinite moments, an escort distribution for which these moments are defined. Indeed, if $p \in \mathcal{T}^d_{\nu_p}$ for some $\nu_p > 0$, we can construct $p^{(\alpha)}$ where $\alpha = 1 + \frac{2}{\nu+d}$ and $\nu > 0$. The resulting escort distribution $p^{(\alpha)}$ has $\nu^{(\alpha)} > \nu_p$ degrees of freedom, i.e., a lighter tail than the one of p itself. Indeed, we can have $\nu_p \leq 2$, meaning that p has infinite variance and $\nu^{(\alpha)} > 2$, in which case $p^{(\alpha)}$ has finite variance.

3.3 Novel technical optimality results

We present in this section two new technical results, that will later be used to study the optimality conditions of the optimization problems arising in variational inference and maximum likelihood estimation.

Proposition 5. *Consider the λ-exponential family $\mathcal{Q}_\lambda$ under Assumption 1, and $\bar{T} \in \mathcal{H}$ such that $c_\lambda(\vartheta, \bar{T}) \in \mathbb{R}$ for any $\vartheta \in \operatorname{dom} \varphi_\lambda$. Then $\vartheta_* \in \operatorname{dom} \varphi_\lambda$ minimizes $\vartheta \longmapsto \varphi_\lambda(\vartheta) - c_\lambda(\vartheta, \bar{T})$ if and only if $\bar{T} \in \partial^{c_\lambda} \varphi_\lambda(\vartheta_*)$.*

Proof. Suppose that $\vartheta_* \in \operatorname{dom} \varphi_\lambda$ minimizes $\vartheta \longmapsto \varphi_\lambda(\vartheta) - c_\lambda(\vartheta, \bar{T})$. This is equivalent to

$$c_\lambda(\vartheta_*, \bar{T}) - \varphi_\lambda(\vartheta_*) \geq c_\lambda(\vartheta, \bar{T}) - \varphi_\lambda(\vartheta), \, \forall \vartheta \in \operatorname{dom} \varphi_\lambda. \tag{3.13}$$

By definition of the c_λ-conjugate, (3.13) can be summarized as

$$c_\lambda(\vartheta_*, \bar{T}) - \varphi_\lambda(\vartheta_*) \geq \varphi_\lambda^{c_\lambda}(\bar{T}). \tag{3.14}$$

Since the opposite inequality is true by definition, the above statement is equivalent to

$$c_\lambda(\vartheta_*, \bar{T}) - \varphi_\lambda(\vartheta_*) = \varphi_\lambda^{c_\lambda}(\bar{T}). \tag{3.15}$$

That yields $\bar{T} \in \partial^{c_\lambda} \varphi_\lambda(\vartheta_*)$, which concludes the proof. □

Lemma 2. *Consider $u \in \mathcal{H}$ and the function $c_\lambda(\cdot, u) : v \longmapsto c_\lambda(v, u)$ for $\lambda \neq 0$.*

(i) If $\lambda = 0$, the function $c_\lambda(\cdot, u)$ is linear.

(ii) If $\lambda > 0$, the function $c_\lambda(\cdot, u)$ is concave.

(iii) If $\lambda < 0$, the function $c_\lambda(\cdot, u)$ is convex.

Proof. Case (i) follows from $c_0(\cdot, \cdot) = \langle \cdot, \cdot \rangle$. We now assume $\lambda \neq 0$. Consider $v_1, v_2 \in \mathcal{H}$ and $s \in [0, 1]$. Then we can compute

$$\begin{aligned} c_\lambda(sv_1 + (1-s)v_2, u) &= \frac{1}{\lambda} \log(1 + \lambda \langle sv_1 + (1-s)v_2, v \rangle) \\ &= \frac{1}{\lambda} \log(s(1 + \lambda \langle v_1, u \rangle) + (1-s)(1 + \lambda \langle v_2, u \rangle)). \end{aligned}$$

We then get the results of cases (ii) and (iii), using the convexity (resp. concavity) of $v \longmapsto \lambda^{-1} \log v$, resulting from the positive (resp. negative) sign of λ. □

Proposition 6. *Consider the λ-exponential family $\mathcal{Q}_\lambda$ under Assumption 1. Let a collection $\{\bar{T}_i\}_{i=1}^N$ of $N > 1$ elements of $\mathcal{H}$, such that for any $i \in \{1, \ldots, N\}$, $c_\lambda(\vartheta, \bar{T}_i) \in \mathbb{R}$ for any $\vartheta \in \operatorname{dom} \varphi_\lambda$, and a collection of non-negative values $\{\rho_i\}_{i=1}^N$ such that $\sum_{i=1}^N \rho_i = 1$. Suppose that there exists $\vartheta_* \in \operatorname{dom} \varphi_\lambda$ such that $\sum_{i=1}^N \rho_i \bar{T}_i \in \partial^{c_\lambda} \varphi_\lambda(\vartheta_*)$.*

(i) If $\lambda = 0$, ϑ_* minimizes $\vartheta \longmapsto \varphi_\lambda(\vartheta) - \sum_{i=1}^N \rho_i c_\lambda(\vartheta, \bar{T}_i)$.

(ii) If $\lambda < 0$, ϑ_* minimizes the function $\vartheta \longmapsto \varphi_\lambda(\vartheta) - c_\lambda\left(\vartheta, \sum_{i=1}^N \rho_i \bar{T}_i\right)$ over $\operatorname{dom} \varphi_\lambda$, itself being an upper bound of the function $\vartheta \longmapsto \varphi_\lambda(\vartheta) - \sum_{i=1}^N \rho_i c_\lambda(\vartheta, \bar{T}_i)$ over $\operatorname{dom} \varphi_\lambda$. Moreover,

$$\varphi_\lambda(\vartheta_*) - \sum_{i=1}^N \rho_i c_\lambda\left(\vartheta_*, \bar{T}_i\right) \leq -\varphi_\lambda^{c_\lambda}\left(\sum_{i=1}^N \rho_i \bar{T}_i\right). \tag{3.16}$$

(*iii*) *If* $\lambda > 0$, ϑ_* *minimizes* $\vartheta \longmapsto \varphi_\lambda(\vartheta) - c_\lambda\left(\vartheta, \sum_{i=1}^N \rho_i \bar{T}_i\right)$ *over* $\operatorname{dom}\varphi_\lambda$, *itself being an lower bound of the function* $\vartheta \longmapsto \varphi_\lambda(\vartheta) - \sum_{i=1}^N \rho_i c_\lambda(\vartheta, \bar{T}_i)$ *over* $\operatorname{dom}\varphi_\lambda$. *Moreover,*

$$-\varphi_\lambda^{c_\lambda}\left(\sum_{i=1}^N \rho_i \bar{T}_i\right) \leq \varphi_\lambda(\vartheta) - \sum_{i=1}^N \rho_i c_\lambda(\vartheta, \bar{T}_i),\ \forall \vartheta \in \operatorname{dom}\varphi_\lambda, \tag{3.17}$$

Proof. Let $\vartheta \in \operatorname{dom}\varphi_\lambda$. Let us first show that, for any λ, $c_\lambda(\vartheta, \sum_{i=1}^N \rho_i \bar{T}_i) \in \mathbb{R}$. Due to the assumption on the $\{\bar{T}_i\}_{i=1}^N$, such result trivially holds for $\lambda = 0$. When $\lambda \neq 0$, we have $1+\lambda\langle\vartheta, \bar{T}_i\rangle > 0$ for any $i \in \{1, \ldots, N\}$, hence $1 + \lambda\langle\vartheta, \sum_{i=1}^N \rho_i \bar{T}_i\rangle > 0$, showing $c_\lambda(\vartheta, \sum_{i=1}^N \rho_i \bar{T}_i) \in \mathbb{R}$.

Case (i): Let $\lambda = 0$. By Lemma 2,

$$\varphi_\lambda(\vartheta) - \sum_{i=1}^N \rho_i c_\lambda(\vartheta, \bar{T}_i) = \varphi_\lambda(\vartheta) - c_\lambda\left(\vartheta, \sum_{i=1}^N \rho_i \bar{T}_i\right),\ \forall \vartheta \in \operatorname{dom}\varphi_\lambda. \tag{3.18}$$

By Proposition 5, ϑ_* minimizes the right-hand side of (3.18), showing the result.

Case (ii): Let $\lambda < 0$. Using Lemma 2 in the case $\lambda < 0$, we get that

$$\varphi_\lambda(\vartheta) - \sum_{i=1}^N \rho_i c_\lambda(\vartheta, \bar{T}_i) \leq \varphi_\lambda(\vartheta) - c_\lambda\left(\vartheta, \sum_{i=1}^N \rho_i \bar{T}_i\right),\ \forall \vartheta \in \operatorname{dom}\varphi_\lambda. \tag{3.19}$$

Using the result of Proposition 5, we get that ϑ_* minimizes the right-hand side of (3.19). Moreover, as $\sum_{i=1}^N \rho_i \bar{T}_i \in \partial^{c_\lambda}\varphi_\lambda(\vartheta_*)$,

$$\varphi_\lambda(\vartheta_*) - c_\lambda\left(\vartheta_*, \sum_{i=1}^N \rho_i \bar{T}_i\right) = -\varphi_\lambda^{c_\lambda}\left(\sum_{i=1}^N \rho_i \bar{T}_i\right). \tag{3.20}$$

Using the inequality in Equation (3.19) and the identity in Equation (3.20) yields the upper-bound property.

Case (iii): Let $\lambda > 0$. Using Lemma 2 for $\lambda > 0$ yields

$$\varphi_\lambda(\vartheta) - c_\lambda\left(\vartheta, \sum_{i=1}^N \rho_i \bar{T}_i\right) \leq \varphi_\lambda(\vartheta) - \sum_{i=1}^N \rho_i c_\lambda(\vartheta, \bar{T}_i),\ \forall \vartheta \in \operatorname{dom}\varphi_\lambda. \tag{3.21}$$

The parameter ϑ_* minimizes the left-hand side of the above due to Proposition 5. This implies in particular that

$$\varphi_\lambda(\vartheta_*) - c_\lambda\left(\vartheta_*, \sum_{i=1}^N \rho_i \bar{T}_i\right) \leq \varphi_\lambda(\vartheta) - \sum_{i=1}^N \rho_i c_\lambda(\vartheta, \bar{T}_i),\ \forall \vartheta \in \operatorname{dom}\varphi_\lambda. \tag{3.22}$$

Using Equation (3.20), which remains true for $\lambda > 0$, we obtain the lower bound result. □

4 Statistical problems over the λ-exponential family

We now leverage all the previous notions and new technical results from Section 3 to tackle variational inference and maximum likelihood estimation problems within the λ-exponential family. We derive novel optimality conditions and algorithms to solve these problems. Finally, we compare and discuss our new findings on the λ-exponential family with known results on the standard exponential family.

4.1 Variational inference through Rényi divergence minimization

We consider in this section the problem

$$\underset{q_\vartheta \in \mathcal{Q}_\lambda}{\text{minimize}}\, RD_\alpha(\pi, q_\vartheta), \tag{P_{VI}}$$

where $\lambda + \alpha = 1$, under Assumption 1. Notice that in the case $\lambda = 0$, $\alpha = 1$, Problem (P_{VI}) corresponds to the minimization of the inclusive Kullback-Leibler divergence over the standard exponential family. We introduce the following additional assumption.

Assumption 3. The target π is in $\mathcal{P}(\mathcal{X}, m)$ and is $\mathcal{Q}_\lambda$-compatible.

4.1.1 Optimality conditions

We now derive novel optimality conditions for Problem (P_{VI}). This is done by leveraging the technical optimality conditions introduced in Section 3.3. These conditions can be seen as a moment-matching conditions on escort probabilities and are discussed in greater extent in Section 4.3. We also show that they can be used straightforwardly in the case of Student distributions.

Proposition 7. *Suppose that Assumptions 1, 2, and 3 are satisfied. If $\vartheta_* \in \operatorname{dom} \varphi_\lambda$ is such that*

$$q_{\vartheta_*}^{(\alpha)}(T) = \pi_{|S_\lambda}^{(\alpha)}(T), \tag{4.1}$$

then ϑ_* is a solution to Problem (P_{VI})

Proof. We first prove that $c_\lambda(\vartheta, \pi_{|S_\lambda}^{(\alpha)}(T)) \in \mathbb{R}$ for any $\vartheta \in \operatorname{dom} \varphi_\lambda$. Due to Assumption 2, it is sufficient to check that $c_\lambda(\vartheta, \pi_{|S_\vartheta}^{(\alpha)}(T)) \in \mathbb{R}$ for any $\vartheta \in \operatorname{dom} \varphi_\lambda$. We then get this first result from Assumption 3 and Lemma 1. Assumption 3 also implies that $H_\alpha(\pi)$ is finite.

Then, rewriting the Rényi divergence using Proposition 1 shows that solving Problem (P_{VI}) is equivalent to solving

$$\underset{\vartheta \in \operatorname{dom} \varphi_\lambda}{\text{minimize}}\, \varphi_\lambda(\vartheta) - c_\lambda(\vartheta, \pi_{|S_\lambda}^{(\alpha)}(T)). \tag{4.2}$$

We can thus apply Proposition 5 to obtain that $\vartheta \in \operatorname{dom} \varphi_\lambda$ is a solution of Problem (P_{VI}) if and only if $\pi_{|S_\lambda}^{(\alpha)}(T) \in \partial^{c_\lambda} \varphi_\lambda(\vartheta)$. Using the description of $\partial^{c_\lambda} \varphi_\lambda(\vartheta)$ from Proposition 2, we get that any ϑ_* satisfying the assumptions of this proposition is such that $\pi_{|S_\lambda}^{(\alpha)}(T) \in \partial^{c_\lambda} \varphi_\lambda(\vartheta_*)$, hence a solution to Problem (P_{VI}). □

Corollary 1. *Consider a target $\pi \in \mathcal{P}(\mathbb{R}^d, dx)$, the family of Student distributions in dimension d with ν degrees of freedom $\mathcal{T}_\nu^d$ and the family of Gaussian distributions $\mathcal{G}^d$.*

(i) If the escort probability $\pi^{(\alpha)}$ exists and has finite first and second order moments for $\alpha = 1 + \frac{2}{\nu + d}$, then the distribution $q_{\mu_*, \Sigma_*} \in \mathcal{T}_\nu^d$ such that

$$\begin{cases} \mu_* & = \pi^{(\alpha)}(x), \\ \Sigma_* & = \pi^{(\alpha)}(xx^\top) - \mu_* \mu_*^\top, \end{cases} \tag{4.3}$$

minimizes $q \longmapsto RD_\alpha(\pi, q)$ over $\mathcal{T}_\nu^d$.

(ii) If π has finite first and second order moments, then the distribution $q_{\mu_*, \Sigma_*} \in \mathcal{G}^d$ such that μ_* and Σ_* satisfy Equation (4.3) with $\alpha = 1$ minimizes $q \longmapsto KL(\pi, q)$ over $\mathcal{G}^d$.

4.1.2 An iterative variational inference algorithm

In order to resolve the optimality conditions given in Proposition 7, we propose in this section an iterative approach relying on the following novel update operator, parametrized by the target π and a stepsize $\tau > 0$.

Definition 9. Consider $\alpha > 0, \lambda = 1 - \alpha$ and the λ-exponential family $\mathcal{Q}_\lambda$ under Assumptions 1 and 2. Consider a target $\pi \in \mathcal{P}(\mathcal{X}, m)$ satisfying Assumption 3. For $\tau > 0$, we define the *operator* P_τ^π such that $\vartheta_P = P_\tau^\pi(\vartheta')$ satisfies

$$q_{\vartheta_P}^{(\alpha)}(T) = \frac{\tau}{1+\tau}\pi_{|S_\lambda}^{(\alpha)}(T) + \frac{1}{1+\tau}q_{\vartheta'}^{(\alpha)}(T),\ \forall \vartheta' \in \operatorname{dom}\varphi_\lambda. \tag{4.4}$$

The above operator shares close links with the proximal operator introduced in Definition 6, as shown below.

Proposition 8. *Consider $\alpha > 0$, $\lambda = 1 - \alpha$, the λ-exponential family $\mathcal{Q}_\lambda$, and a target distribution $\pi \in \mathcal{P}(\mathcal{X}, m)$ such that Assumptions 1, 2, and 3 are satisfied. Let $\tau > 0$. Then,*

(i) If $\lambda = 0$, then

$$\forall \vartheta' \in \operatorname{dom}\varphi_\lambda \quad P_\tau^\pi(\vartheta') = prox_\tau^{KL(\pi,q_\cdot)}(\vartheta') \tag{4.5}$$

(ii) If $\lambda < 0$ (resp. $\lambda > 0$), then $P_\tau^\pi(\vartheta')$ approximates $prox_\tau^{RD_\alpha(\pi,q_\cdot)}(\vartheta')$ in the sense that it minimizes an upper bound (resp. lower bound) of the proximal loss

$$\vartheta \longmapsto RD_\alpha(\pi, q_\vartheta) + \frac{1}{\tau}RD_\alpha(q_{\vartheta'}, q_\vartheta). \tag{4.6}$$

Proof. We begin by decomposing the objective function appearing in the computation of $\operatorname{prox}_\tau^{RD_\alpha(\pi,q_\cdot)}(\vartheta')$.

$$\begin{aligned}
&RD_\alpha(\pi, q_\vartheta) + \frac{1}{\tau}RD_\alpha(q_{\vartheta'}, q_\vartheta) \\
&= \varphi_\lambda(\vartheta) - c_\lambda(\vartheta, \pi_{|S_\lambda}^{(\alpha)}(T)) - H_\alpha(\pi_{|S_\lambda}) + \frac{1}{\tau}\left(\varphi_\lambda(\vartheta) - c_\lambda(\vartheta, q_{\vartheta'}^{(\alpha)}(T)) + \psi_\lambda(\vartheta')\right) \\
&= \left(\frac{1+\tau}{\tau}\right)\left(\varphi_\lambda(\vartheta) - \frac{\tau}{1+\tau}c_\lambda(\vartheta, \pi_{|S_\lambda}^{(\alpha)}(T)) - \frac{1}{1+\tau}c_\lambda(\vartheta, q_{\vartheta'}^{(\alpha)}(T))\right) - H_\alpha(\pi_{|S_\lambda}) + \frac{1}{\tau}\psi_\lambda(\vartheta').
\end{aligned}$$

If we conserve only the terms depending on the variable ϑ and ignore the positive multiplicative factor, we thus obtain that $\operatorname{prox}_\tau^{RD_\alpha(\pi,q_\cdot)}(\vartheta')$ is the set of solutions of the problem

$$\underset{\vartheta \in \operatorname{dom}\varphi_\lambda}{\text{minimize}}\, \varphi_\lambda(\vartheta) - \frac{\tau}{1+\tau}c_\lambda(\vartheta, \pi_{|S_\lambda}^{(\alpha)}(T)) - \frac{1}{1+\tau}c_\lambda(\vartheta, q_{\vartheta'}^{(\alpha)}(T)).$$

Now, remark that due to Assumption 2 and Lemma 1, $c_\lambda(\vartheta, q_{\vartheta'}^{(\alpha)}(T)) \in \mathbb{R}$ for any $\vartheta \in \operatorname{dom}\varphi_\lambda$. The same holds with the term $c_\lambda(\vartheta, \pi_{|S_\lambda}^{(\alpha)}(T))$ by Assumption 3. Since $\frac{\tau}{1+\tau} + \frac{1}{1+\tau} = 1$ and all the involved terms are positive, we can thus apply Propositions 6 and 2 to get the result. □

We are now ready to state our algorithm to solve Problem (P_{VI}). We then study its convergence properties.

Algorithm 1: Proposed algorithm to solve Problem (P_{VI})

Let $\vartheta_0 \in \operatorname{dom}\varphi_\lambda$, and a sequence $\{\tau_k\}_{k\in\mathbb{N}}$ of step parameters in $\mathbb{R}_{++}$.
for $k = 0, \ldots$ **do**
  Update ϑ_{k+1} using $P_{\tau_k}^\pi$, that is such that

$$q_{\vartheta_{k+1}}^{(\alpha)} = \frac{\tau_k}{1+\tau_k}\pi_{|\operatorname{supp} Q_\lambda}^{(\alpha)}(T) + \frac{1}{1+\tau_k}q_{\vartheta_k}^{(\alpha)}(T). \tag{4.7}$$

Proposition 9. *If $\vartheta_k \in \operatorname{dom} \varphi_\lambda$ for every $k \in \mathbb{N}$, then the sequence generated by Algorithm 1 is well-defined and*

$$q^{(\alpha)}_{\vartheta_k}(T) \xrightarrow[k\to+\infty]{} \pi^{(\alpha)}_{|S_\lambda}(T). \tag{4.8}$$

Proof. For any $K \in \mathbb{N} \setminus \{0\}$, we have

$$q^{(\alpha)}_{\vartheta_K}(T) = \left(\prod_{k=0}^{K-1} \frac{1}{1+\tau_k}\right) q^{(\alpha)}_{\vartheta_0}(T) + \left(1 - \prod_{k=0}^{K-1} \frac{1}{1+\tau_k}\right) \pi_{|S_\lambda}(T). \tag{4.9}$$

Since $\frac{1}{1+\tau_k} \in (0,1)$ for every $k \in \mathbb{N}$, $\prod_{k=0}^{K-1} \frac{1}{1+\tau_k} \xrightarrow[K\to+\infty]{} 0$, showing the result. □

Remark 4. When $\lambda = 0$, Algorithm 1 identifies with the Bregman proximal algorithm from [7, 43]. Further, Proposition 9 shows that Algorithm 1 produces iterates converging to the solution of Problem (P_{VI}).

Algorithm 1 involves at every iteration the computation of $\pi^{(\alpha)}_{|S_\lambda}(T)$. This quantity is in general unavailable. Actually, the updates in Algorithm 1 allow to build an alternative estimate for $\pi^{(\alpha)}_{|S_\lambda}(T)$ at every iteration, and to combine them using a step-size parameter $\tau_k \xrightarrow[k\to+\infty]{} 0$ in the spirit of stochastic approximation algorithms. This will be illustrated in Section 5.1.

4.2 Maximum likelihood estimation

We consider now the maximum likelihood problem of estimating the parameters of a distribution from the λ-exponential family $\mathcal{Q}_\lambda$ based on observed data $\{x_i\}_{i=1}^N$. This problem reads as follows.

$$\underset{\vartheta \in \operatorname{dom} \varphi_\lambda}{\text{maximize}} \sum_{i=1}^N \log q_\vartheta(x_i). \tag{P_{MLE}}$$

We need the following assumption on the data to ensure the well-posedness of Problem (P_{MLE}).

Assumption 4. For every $q_\vartheta \in \mathcal{Q}_\lambda$, $x_i \in S_\vartheta$ for every $i \in \{1, \dots, N\}$.

4.2.1 Optimality conditions and approximate solutions

We now provide novel conditions for the resolution of Problem (P_{MLE}). These conditions are optimal in the case of the standard exponential family. In the case of the λ-exponential family, the conditions are sub-optimal and we relate them explicitly to the optimal solutions of Problem (P_{MLE}).

Proposition 10. *Consider $\lambda \in \mathbb{R}$ such that $\alpha = 1 - \lambda$ is positive, and the λ-exponential family and data $\{x_i\}_{i=1}^N$ such that that Assumptions 1, 2, and 4 hold. Suppose that there exists $\vartheta_* \in \operatorname{dom} \varphi_\lambda$ such that*

$$q^{(\alpha)}_{\vartheta_*}(T) = \frac{1}{N} \sum_{i=1}^N T(x_i). \tag{4.10}$$

(i) If $\lambda = 0$, ϑ_* maximizes Problem (P_{MLE}).

(ii) If $\lambda < 0$, ϑ_* maximizes a lower bound of $\vartheta \longmapsto \sum_{i=1}^N \log q_\vartheta(x_i)$ over $\operatorname{dom} \varphi_\lambda$. Moreover,

$$\frac{1}{N} \sum_{i=1}^N \log q_{\vartheta_*}(x_i) \geq \psi_\lambda(\vartheta_*). \tag{4.11}$$

(*iii*) *If* $\lambda > 0$, ϑ_* *maximizes an upper bound of* $\vartheta \longmapsto \sum_{i=1}^N \log q_\vartheta(x_i)$ *over* $\operatorname{dom}\varphi_\lambda$. *Moreover,*

$$\frac{1}{N}\sum_{i=1}^N \log q_\vartheta(x_i) \leq \psi_\lambda(\vartheta_*),\ \forall \vartheta \in \operatorname{dom}\varphi_\lambda. \tag{4.12}$$

Proof. Remark first that solving Problem (P_{MLE}) is equivalent to solving

$$\underset{\vartheta\in\operatorname{dom}\varphi_\lambda}{\text{minimize}} -\frac{1}{N}\sum_{i=1}^N \log q_\vartheta(x_i) = \varphi_\lambda(\vartheta) - \frac{1}{N}\sum_{i=1}^N c_\lambda(\vartheta, T(x_i)).$$

Assumption 4 ensures that we can apply Proposition 6. The result comes from the results of this Proposition and the description of $\partial^{c_\lambda}\varphi_\lambda$ and $\varphi_\lambda^{c_\lambda}$ provided in Proposition 2. □

Corollary 2. *Consider Problem* (P_{MLE}) *with data points* $x_i \in \mathbb{R}^d$ *for* $i = \{1, \dots, N\}$.

(*i*) *If we consider Problem* (P_{MLE}) *over the family of Student distributions in dimension* d *with* ν *degrees of freedom* $\mathcal{T}_\nu^d$, *the distribution* $q_{\mu_*,\Sigma_*} \in \mathcal{T}_\nu^d$ *such that*

$$\begin{cases} \mu_* &= \frac{1}{N}\sum_{i=1}^N x_i, \\ \Sigma_* &= \frac{1}{N}\sum_{i=1}^N x_i x_i^\top - \mu_*\mu_*^\top, \end{cases} \tag{4.13}$$

maximizes a lower bound of Problem (P_{MLE}). *We also get that*

$$\frac{1}{N}\sum_{i=1}^N \log q_{\mu_*,\Sigma_*}(x_i) \geq \frac{1}{2}\log\det(\Sigma_*) + C, \tag{4.14}$$

where the constant C *depends only on* d *and* ν.

(*ii*) *If we consider Problem* (P_{MLE}) *over the family of Gaussian distributions* $\mathcal{G}^d$, *the distribution* $q_{\mu_*,\Sigma_*} \in \mathcal{G}^d$ *with* μ_*, Σ_* *satisfying Equation* (4.13) *maximizes Problem* (P_{MLE}).

4.2.2 An iterative algorithm for maximum likelihood estimation

We now propose a new iterative algorithm to reach the (sub-optimal) solutions to Problem (P_{MLE}), as characterized in Proposition 10. To do so, we first introduce the following operator.

Definition 10. Consider $\alpha > 0, \lambda = 1 - \alpha$ and the λ-exponential family $\mathcal{Q}_\lambda$ under Assumptions 1 and 2. Consider data points $\{x_i\}_{i=1}^N$ satisfying Assumption 4. For $\tau > 0$, we define the *operator* $P_\tau^{\{x_i\}_{i=1}^N}$ such that for any $\vartheta' \in \operatorname{dom}\varphi_\lambda$, $\vartheta_P = P_\tau^{\{x_i\}_{i=1}^N}(\vartheta')$ satisfies

$$q_{\vartheta_P}^{(\alpha)}(T) = \frac{N\tau}{1+N\tau}\sum_{i=1}^N T(x_i) + \frac{1}{1+N\tau} q_{\vartheta'}^{(\alpha)}(T). \tag{4.15}$$

This operator can be related to the proximal operator from Definition 6, as we show now.

Proposition 11. *Consider a* λ*-exponential family* $\mathcal{Q}_\lambda$ *with* $\lambda \in \mathbb{R}$ *such that* $\alpha = 1 - \lambda$ *is positive, and observed data* $\{x_i\}_{i=1}^N$ *such that Assumptions 1, 2, and 4 are satisfied. Let* $\tau > 0$. *Then,*

(*i*) *If* $\lambda = 0$, *then*

$$P_\tau^{\{x_i\}_{i=1}^N}(\vartheta') = \mathit{prox}_\tau^{-\sum_{i=1}^N \log q_\cdot(x_i)}(\vartheta'),\ \forall \vartheta' \in \operatorname{dom}\varphi_\lambda. \tag{4.16}$$

(*ii*) *If* $\lambda < 0$ *(resp.* $\lambda > 0$*), then* $P_\tau^{\{x_i\}_{i=1}^N}(\vartheta')$ *approximates* $\mathrm{prox}_\tau^{-\sum_{i=1}^N \log q_\cdot(x_i)}(\vartheta')$ *in the sense that it minimizes an upper bound (resp. lower bound) of the proximal loss*

$$\vartheta \longmapsto -\sum_{i=1}^N \log q_\vartheta(x_i) + \frac{1}{\tau} RD_\alpha(q_{\vartheta'}, q_\vartheta). \tag{4.17}$$

Proof. We first decompose the objective function appearing in $\mathrm{prox}_\tau^{-\sum_{i=1}^N \log q_\cdot(x_i)}(\vartheta')$:

$$\begin{aligned}
&-\sum_{i=1}^N \log q_\vartheta(x_i) + \frac{1}{\tau} RD_\alpha(q_{\vartheta'}, q_\vartheta) \\
&= N\varphi_\lambda(\vartheta) - \sum_{i=1}^N c_\lambda(\vartheta, T(x_i)) + \frac{1}{\tau}\left(\varphi_\lambda(\vartheta) - c_\lambda(\vartheta, q_{\vartheta'}^{(\alpha)}(T)) + \psi_\lambda(\vartheta')\right) \\
&= \left(\frac{1+N\tau}{\tau}\right)\left(\varphi_\lambda(\vartheta) - \sum_{i=1}^N \frac{\tau}{1+N\tau} c_\lambda(\vartheta, T(x_i)) - \frac{1}{1+N\tau} c_\lambda(\vartheta, q_{\vartheta'}^{(\alpha)}(T))\right) + \frac{1}{\tau}\psi_\lambda(\vartheta').
\end{aligned}$$

The above calculation shows that computing $\mathrm{prox}_\tau^{\{x_i\}}(\vartheta')$ is equivalent to solving

$$\underset{\vartheta \in \mathrm{dom}\,\varphi_\lambda}{\mathrm{minimize}}\, \varphi_\lambda(\vartheta) - \sum_{i=1}^N \frac{\tau}{1+N\tau} c_\lambda(\vartheta, T(x_i)) - \frac{1}{1+N\tau} c_\lambda(\vartheta, q_{\vartheta'}^{(\alpha)}(T)).$$

Then, one can conclude as in the proof of Proposition 8. □

We are now ready to introduce our algorithm to reach the solutions given in Proposition 10, and as such, solving (approximatly) Problem (P_{MLE}).

Algorithm 2: Proposed algorithm to solve Problem (P_{MLE})

Let $\vartheta_0 \in \mathrm{dom}\,\varphi_\lambda$, and a sequence $\{\tau_k\}_{k\in\mathbb{N}}$ of step parameters in $\mathbb{R}_{++}$.
for $k = 0, \dots$ **do**
  Update ϑ_{k+1} using $P_{\tau_k}^{\{x_i\}_{i=1}^N}$, that is such that

$$q_{\vartheta_{k+1}}^{(\alpha)}(T) = \frac{N\tau_k}{1+N\tau_k} \sum_{i=1}^N T(x_i) + \frac{1}{1+N\tau_k} q_{\vartheta_k}^{(\alpha)}(T). \tag{4.18}$$

Proposition 12. *If* $\vartheta_k \in \mathrm{dom}\,\varphi_\lambda$ *for every* $k \in \mathbb{N}$*, then the sequence generated by Algorithm 2 is well-defined and*

$$q_{\vartheta_k}^{(\alpha)}(T) \xrightarrow[k\to+\infty]{} \frac{1}{N}\sum_{i=1}^N T(x_i). \tag{4.19}$$

Proof. The proof follows the same step as the proof of Proposition 9. □

Remark 5. In the case $\lambda = 0$, Algorithm 2 is a Bregman proximal algorithm [7, 43] that converges to the solution of Problem (P_{MLE}). In the case $\lambda \neq 0$, we have the convergence to the sub-optimal solutions described in Proposition 10.

The algorithm obtained by applying the update of Equation (4.18) can for instance be used in an online context, where all the data points are not available at every iteration. This will be illustrated in Section 5.2.

4.2.3 An expectation-maximization algorithm for mixture MLE

We consider here a variant of Problem (P_{MLE}) where we aim at estimating the parameters of a mixture of $J \in \mathbb{N}$, $J > 0$, distributions from the λ-exponential family $\mathcal{Q}_\lambda$, based on observed data $\{x_i\}_{i=1}^N$. The problem is over the parameters of each component of the mixture, as well as over their weights, and reads as follows.

$$\underset{\substack{\xi_j \geq 0, \vartheta_j \in \operatorname{dom} \varphi_\lambda,\, j=1,\dots,J \\ \sum_{j=1}^J \xi_j = 1}}{\text{maximize}} \sum_{i=1}^N \log \left(\sum_{j=1}^J \xi_j q_{\vartheta_j}(x_i) \right). \tag{$P_{\text{MLE-Mixt}}$}$$

A standard algorithm to solve this type of problem is the expectation-maximization (EM) algorithm [10], that generates a sequence of weights $\{\xi_{j,k}\}_{k\in\mathbb{N}}$ and parameters $\{\vartheta_{j,k}\}_{k\in\mathbb{N}}$ for $j = 1, \dots, J$. For any $j = 1, \dots, J$ and iteration $k \in \mathbb{N}$, we denote by $\gamma_{k,j}$ the function defined by

$$\gamma_{k,j}(x) = \frac{\xi_{k,j} q_{\vartheta_{k,j}}(x)}{\sum_{j'=1}^J \xi_{k,j'} q_{\vartheta_{k,j'}}(x)}. \tag{4.20}$$

It is then possible to apply the EM algorithm in our setting, yielding updates of the form

$$\xi_{k+1,j} = \frac{1}{N} \sum_{n=1}^N \gamma_{k,j}(x_n) \tag{4.21}$$

$$\vartheta_{k+1,j} = \underset{\vartheta \in \operatorname{dom} \varphi_\lambda}{\arg\max} \sum_{i=1}^N \gamma_{k,j}(x_i) \log p_\vartheta(x_i). \tag{4.22}$$

The maximization step, often called the M-step, appearing in the update (4.22) does not always have a closed-form. We now give a result about explicit solutions of these steps, that are possibly optimal, leveraging tools from our Proposition 10.

Proposition 13. *Consider a λ-exponential family $\mathcal{Q}_\lambda$ with $\lambda \in \mathbb{R}$ such that $\alpha = 1-\lambda$ is positive, and observed data $\{x_i\}_{i=1}^N$ such that Assumptions 1, 2, and 4 are satisfied. Suppose that there exists $\vartheta_{k+1,j} \in \operatorname{dom} \varphi_\lambda$ such that*

$$q^{(\alpha)}_{\vartheta_{k+1,j}} = \sum_{i=1}^N \frac{\gamma_{k,j}(x_i)}{\sum_{i'=1}^N \gamma_{k,j}(x_{i'})} T(x_i). \tag{4.23}$$

(i) If $\lambda = 0$, then $\vartheta_{k+1,j}$ exactly solves the optimization problem in the update (4.22).

(ii) If $\lambda < 0$ (resp. $\lambda > 0$), then $\vartheta_{k+1,j}$ approximates the solution of the update (4.22) in the sense that it maximizes a lower bound (resp. upper bound) of the considered loss.

Proof. The maximization problem in the update (4.22) is the following:

$$\underset{\vartheta \in \operatorname{dom} \varphi_\lambda}{\text{maximize}} \sum_{i=1}^N \gamma_{k,j}(x_i) \log p_\vartheta(x_i), \tag{4.24}$$

which is equivalent, since the functions $\gamma_{k,j}$ take non-negative values, to

$$\underset{\vartheta \in \operatorname{dom} \varphi_\lambda}{\text{maximize}} \sum_{i=1}^N \frac{\gamma_{k,j}(x_i)}{\sum_{i'=1}^N \gamma_{k,j}(x_{i'})} \log p_\vartheta(x_i). \tag{4.25}$$

Finally, we re-write this optimization problem as

$$\underset{\vartheta \in \operatorname{dom} \varphi_\lambda}{\text{minimize}}\, \varphi_\lambda(\vartheta) - \sum_{i=1}^{N} \frac{\gamma_{k,j}(x_i)}{\sum_{i'=1}^{N} \gamma_{k,j}(x_{i'})} c_\lambda(\vartheta, T(x_i)), \tag{4.26}$$

which allows to conclude the proof as in the proof of Proposition 10. □

When $\lambda = 0$, the result of Proposition 13 implies that the M-steps, that is the updates of the form (4.22), can be solved exactly. Since $\lambda = 0$ corresponds to the exponential family, which can represent Gaussian distributions, this result recovers the EM algorithm for Gaussian mixtures presented in [10]. Otherwise, the result of Proposition 13 leads to an approximate EM algorithm, where the M-steps, are only approximately solved through an explicit expression. The resulting algorithm, is summarized in Algorithm 3.

Algorithm 3: A sub-optimal EM algorithm to solve Problem ($P_{\text{MLE-Mixt}}$)

Let $\vartheta_{0,j} \in \operatorname{dom} \varphi_\lambda$ and $\xi_{0,j} \geq 0$ for $j = 1, \dots, J$ such that $\sum_{j=1}^{J} \xi_{0,j} = 1$.

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

For every $j = 1, \dots, J$, define the function $\gamma_{k,j}$ following Equation (4.20), and update the parameters $\xi_{k+1,j}$ and $\vartheta_{k+1,j}$ such that they satisfy

$$\xi_{k+1,j} = \frac{1}{N} \sum_{i=1}^{N} \gamma_{k,j}(x_i), \tag{4.27}$$

$$q^{(\alpha)}_{\theta_{k+1,j}}(T) = \sum_{i=1}^{N} \frac{\gamma_{k,j}(x_i)}{\sum_{i'=1}^{N} \gamma_{k,j}(x_{i'})} T(x_i). \tag{4.28}$$

4.3 Discussion and comparison with the standard exponential family

Let us now discuss our results for maximum likelihood, variational inference, and iterative algorithms obtained for the λ-exponential family $\mathcal{Q}_\lambda$.

4.3.1 The particular case of the exponential family

We here discuss how our theoretical results position themselves, compared to existing results for the special case $\lambda = 0$.

We recall that for an exponential family $\mathcal{Q}$ with sufficient statistics T (which is the λ-exponential family with $\lambda = 0$), the densities of the members of the family are given by Equation (2.7) with $c_0(\cdot, \cdot) = \langle \cdot, \cdot \rangle$ and the log-partition function φ. We have proven in Proposition 1 that for any $\pi \in \mathcal{P}(\mathcal{X}, m)$ such that $H_1(\pi)$, which is the Shannon entropy of π, and $\pi(T)$ are well-defined,

$$KL(\pi, q_\vartheta) = -H_1(\pi) - \langle \vartheta, \pi(T) \rangle + \varphi(\vartheta), \, \forall \vartheta \in \operatorname{dom} \varphi. \tag{4.29}$$

In Proposition 2, we also uncovered the links between the Shannon entropy and the Fenchel conjugate of the log-partition function, and showed that the moments of a distribution from the exponential family are the subgradients of the log-partition function. These facts, although scattered in the literature, are well-known. In our Propositions 1 and 2, we generalized them to the λ-exponential family, using H_α instead of H_1, RD_α instead of KL, φ_λ instead of φ, and escort moments instead of standard moments.

In the case of Problem (P_{VI}), we can have the same type of correspondence. We have proven in Proposition 7 that Problem (P_{VI}) over $\mathcal{Q}$ with $RD_\alpha = KL$ is solved by $\vartheta_* \in \operatorname{dom} \varphi$ satisfying the moment-matching

condition $q_{\vartheta_*}(T) = \pi(T)$. This optimality condition was already known, see for instance [10, 46, 13]. In Proposition 7, we generalized this optimality condition under the form of a moment-matching condition on escort probabilities, that is $\pi^{(\alpha)}(T) = q^{(\alpha)}_{\vartheta_*}(T)$.

So far, the analysis we proposed for $\lambda \neq 0$ strictly generalizes the case $\lambda = 0$. In fact, it uses the same proofs for $\lambda = 0$ and $\lambda \neq 0$. Let us now review situations where this correspondence breaks. Due to the linearity of the scalar product and the convex subdifferential, we could obtain in Proposition 10 that for $\lambda = 0$, Problem (P_{MLE}) is minimized for $\vartheta_* \in \operatorname{dom}\varphi$ such that the moments $q_{\vartheta_*}(T)$ match the sufficient statistics of the data $\frac{1}{N}\sum_{i=1}^N T(x_i)$. In the case $\lambda \neq 0$, the similar solution obtained by plugging escort moments $q^{(\alpha)}_{\vartheta_*}(T)$ instead of standard moments $q_{\vartheta_*}(T)$ is only sub-optimal, as shown in Proposition 10. More precisely, these are only the minimizers of upper or lower bounds, depending on the sign of λ. The situation is similar when designing EM algorithms, as shown in Proposition 13 where optimality is only attained when $\lambda = 0$, in which case Algorithm 3 recovers the standard EM algorithm [10]. Such results are to be expected as no closed-form solution is known for this type of maximum likelihood estimation problems, and solving these problems, notably over Student-like distribution, is still an active field of research [21, 4]. The situation is similar for the operators defined in Definitions 9 and 10, since they can be seen as an exact proximal operator only for $\lambda = 0$, as shown in Propositions 8 and 11.

Let us now comment about the sub-optimality of the maximum likelihood estimator proposed in Proposition 10 by relating it with the solution of Problem (P_{VI}). Suppose that $x_i \sim \pi$ for any $i \in \{1, \dots, N\}$ and $\frac{1}{N}\sum_{i=1}^N T(x_i) \xrightarrow[N\to+\infty]{} \pi(T)$. This means that in the limit $N \to +\infty$ and when $\lambda = 0$, Problems (P_{MLE}) and (P_{VI}) have the same solution ϑ_* such that $q_{\vartheta_*}(T) = \pi(T)$. This relation between maximum likelihood estimation and minimization of a Kullback-Leibler divergence is well-known and applies in fact in a more general setting [48]. When $\lambda \neq 0$, the sub-optimal solution of Problem (P_{MLE}) described in Proposition 10 is such that $q^{(\alpha)}_{\vartheta_*}(T) = \pi(T)$ in the large number limit $N \to +\infty$, which is different from the solution of Problem (P_{VI}) given in Proposition 7. Notice however that the solution $\vartheta_* \in \operatorname{dom}\varphi_\lambda$ such that $q^{(\alpha)}_{\vartheta_*}(T) = \pi(T)$ is a solution to

$$\operatorname*{minimize}_{\vartheta \in \operatorname{dom}\varphi_\lambda} RD_\alpha(\pi^{(1/\alpha)}, q_\vartheta).$$

Thus, in the large number of samples regime, the sub-optimal solution of Problem (P_{MLE}) does not solve Problem (P_{VI}) but a similar variational inference problem with a deformed target.

Finally, we remark that Assumptions 1 and 2 prevent us from straightforwardly applying our results to the λ-exponential family when $\lambda > 0$. Indeed, such value of λ can lead to distributions whose support depends on the parameters (see for instance the distributions studied in [32] and in [50, Example 3.17]). Although Proposition 1 holds even for varying support, this behavior makes optimization much more challenging.

4.3.2 Comparing our works with existing results in optimization

First, remark that the proximal operators used in our algorithms can be considered as generalized Bregman proximal operators, where the scalar product of $\mathcal{H}$ is replaced by the non-linear coupling c_λ. Indeed, it is well-known that the Kullback-Leibler divergence between two members of the same exponential family can be written as a Bregman divergence [6]. In our case, we can rewrite the Rényi divergence $RD_\alpha(q_{\vartheta'}, q_\vartheta)$ under a similar form, using c_λ:

$$RD_\alpha(q_{\vartheta'}, q_\vartheta) = \varphi_\lambda(\vartheta) - \varphi_\lambda(\vartheta') - c_\lambda(\vartheta, q^{(\alpha)}_{\vartheta'}(T)) + c_\lambda(\vartheta', q^{(\alpha)}_{\vartheta'}(T)), \tag{4.30}$$

with $q^{(\alpha)}_{\vartheta'}(T) \in \partial^{c_\lambda}\varphi_\lambda(\vartheta')$, using Equation (2.14) and Proposition 2. The particular re-writing of Equation (4.30) was established in [47].

Propositions 8 and 11 show that the operators that we proposed in Definitions 9 and 10 to build our algorithms are approximating a proximal operator when $\lambda \neq 0$. We are not aware of any optimization algorithms

stated directly in a generalized convexity framework (i.e., a generalization of standard convexity theory using modified scalar product as in [14, 29, 18], or modified subgradient as in [9]). Although our operators are not exactly proximal operators (except for $\lambda = 0$), they may be a first step leading to such algorithms. Note however that our construction heavily depends on the objective function having an expression like the ones described in Propositions 5 and 6.

The authors of [24] also faced the difficulty of computing proximal operators of the form introduced in Definition 6. While we proposed ad hoc operators that are shown to be sub-optimal solutions to these optimization problems in Propositions 8 and 11, they took another route. Indeed, they studied a continuous time Riemannian gradient flow, whose metric is given by the corrected Hessian of a function that is convex in the sense of the coupling c_λ. Note that the authors consider other types of objective functions than we did. They consider convex and differentiable objectives, while we consider specifically maximum likelihood and variational inference problems, whose objectives are not necessarily convex.

In the context of variational inference, a gradient descent algorithm within the geometry induced by the Kullback-Leibler divergence is studied in [20] for the minimization of the Rényi divergence with $\alpha \in (0, 1]$ over the standard exponential family, amounting to $\lambda = 0$. A gradient descent algorithm to minimize the χ^2 divergence, which is linked to the Rényi divergence with $\alpha = 2$ over the exponential family has also been proposed in [1] for adaptive importance sampling [11]. In this work, we have only considered the setting $\lambda + \alpha = 1$, imposing a strict relation between the approximating family and the divergence.

5 Numerical experiments

We now illustrate our findings through numerical experiments. Our examples are designed as proof-of-concepts, illustrating the advantage of considering the λ-exponential family, instead of the standard exponential one, in simple situations. To do so, we consider instances of Problems (P_{VI}), (P_{MLE}), and ($P_{\text{MLE-Mixt}}$) where the approximating family $\mathcal{Q}_\lambda$ is the Student family $\mathcal{T}_\nu^d$ (see Section 3.2). We remind that this amounts to setting $\lambda = -\frac{2}{\nu+d}$ (see Proposition 3 (i)). In our comparisons, we will also consider the limiting case of Gaussian distributions, obtained by setting $\nu = +\infty$, in which case $\lambda = 0$. For pedagogical purpose, in all examples, the target distribution (in case of variational inference problem) and the distribution generating the samples (in case of maximum likelihood problem) is also a Student density, denoted $\pi \in \mathcal{T}_{\nu_\pi}^d$, and parametrized by $\nu_\pi > 0$ degrees of freedom, location parameter $\mu_\pi \in \mathbb{R}^d$ and shape matrix $\Sigma_\pi \in \mathcal{S}_{++}^d$. This controlled setting allows to access π and its escort $\pi^{(\alpha)}$, sample from them, and compute Rényi divergences, making it possible to assess quantitatively the results.

5.1 A variational inference problem with Student approximating densities

We start our experiments by an instance of Problem (P_{VI}) described as

$$\underset{q_{\mu,\Sigma} \in \mathcal{T}_\nu^d}{\text{minimize}} \, RD_\alpha(\pi, q_{\mu,\Sigma}), \tag{$P_{\text{VI-Student}}$}$$

where $\alpha = 1 + \frac{2}{\nu+d}$, in light of $\mathcal{T}_\nu^d$ being a λ-exponential family with $\lambda = -\frac{2}{\nu+d}$ (see Proposition 3 (ii)) and the optimality result of Proposition 7. The optimality conditions of Problem ($P_{\text{VI-Student}}$) are given in Equation (4.3). These conditions amount to setting μ and Σ such that the first and second order moments of $q_{\mu,\Sigma}^{(\alpha)}$ match those of the escort of the target π, that is $\pi^{(\alpha)}$.

By Proposition 4, if $\alpha = 1 + \frac{2}{\nu+d}$ for some $\nu > 0$, then $\pi^{(\alpha)}$ has first and second order moments if and only if

$$\nu_\pi + 2\frac{\nu_\pi + d}{\nu + d} > 2. \tag{5.1}$$

We consider $\pi \in \mathcal{T}_{\nu_\pi}^d$ with $\nu_\pi \in \{1, 3, 10\}$ and $d \in \{5, 20\}$. The location vector μ_π is sampled uniformly in $[-1, 1]^d$ and the shape matrix $\Sigma_\pi \in \mathcal{S}_{++}^d$ is constructed following [34] with a condition number $\kappa_\pi \in \{10, 1000\}$ (i.e., a well conditioned setting, and a poorly conditioned setting). Regarding our approximating families, we experiment various degrees of freedom $\nu \in \{1, 3, 10, +\infty\}$ such that Equation (5.1) is satisfied. The case $\nu = \infty$ corresponds to a Gaussian approximating family, which is an instance of the exponential family. In contrast, for finite ν, we are working within an instance of the λ-exponential family, $\lambda = -\frac{2}{\nu+d}$. Our experimental scenarios cover the matched case where $\nu = \nu_\pi$, as well as various mismatched cases where $\nu \neq \nu_\pi$.

Using the results from Section 4.1, we have actually two ways to solve Problem ($P_{\text{VI-Student}}$). We can either follow Corollary 1 and try to directly approximate the optimality conditions of Equation (4.3). This requires the computation of the first and second order moments of the escort of the target. Alternatively, we can implement Algorithm 1. This requires the computation of the same moments, but it allows to approximate them differently at each iteration and possibly average the errors and improve the estimators. We consider the two approaches in what follows. We also consider two distinct ways to approximate the first and second order moments of the escort of the target. In Section 5.1.1, we consider that exact samples from $\pi^{(\alpha)}$ are used to approximate Equation (4.3). This idealized setting allows to illustrate the validity of our optimality conditions with an exact sampling procedure. In Section 5.1.2, we consider a more realistic situation where only the unnormalized density of the target is available. In this situation, one needs an integration procedure to approximate the moments of the escort of the target in this setting. We choose here to use a Metropolis-adjusted Langevin algorithm (MALA) [41]. In this setting, we consider the approximation of Equation (4.3) as well as the implementation of Algorithm 1 with an adaptively scaled MALA [31, 30].

5.1.1 Using samples from the target

Problem ($P_{\text{VI-Student}}$) can be solved by approximating the optimality conditions of Corollary 1 using a standard Monte Carlo algorithm with samples from $\pi^{(\alpha)}$. This is feasible as, in this experiment, $\pi^{(\alpha)}$ is a Student distribution with parameters described by Proposition 4. This is an idealized setting since in practical scenarios of variational inference, one does not have the possibility to sample from the escort target. This leads to the following sampling algorithm.

Algorithm 4: Solving Problem ($P_{\text{VI-Student}}$) by approximating (4.3) with samples from $\pi^{(\alpha)}$

Choose an approximating family $\mathcal{T}_\nu^d$ and set $\alpha = 1 + \frac{2}{\nu+d}$. Choose $N \in \mathbb{N}$.

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

Sample $\{x_{k+1}^{(1)}, \dots, x_{k+1}^{(N)}\}$ from $\pi^{(\alpha)}$.

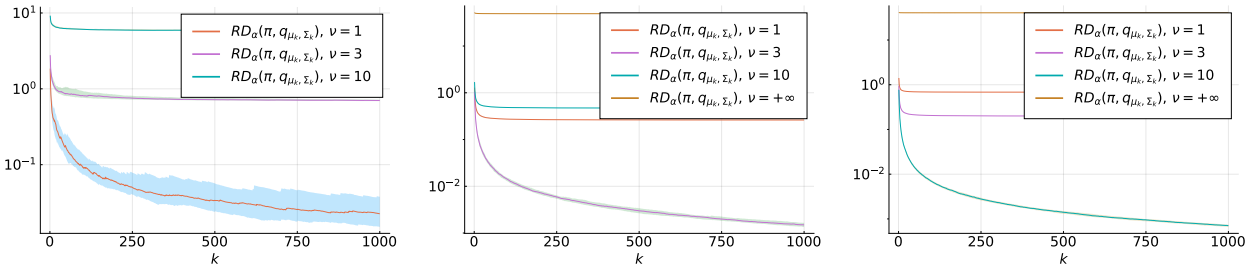
Evaluate $(\pi^{(\alpha)}(x))_{k+1}, (\pi^{(\alpha)}(xx^\top))_{k+1}$ by

$$\begin{cases} (\pi^{(\alpha)}(x))_{k+1} & = \frac{1}{kN} \sum_{l=0}^{k} \sum_{i=1}^{N} x_{l+1}^{(i)}, \\ (\pi^{(\alpha)}(xx^\top))_{k+1} & = \frac{1}{kN} \sum_{l=0}^{k} \sum_{i=1}^{N} x_{l+1}^{(i)} (x_{l+1}^{(i)})^\top. \end{cases} \tag{5.2}$$

Compute μ_{k+1}, Σ_{k+1} following

$$\begin{cases} \mu_{k+1} = (\pi^{(\alpha)}(x))_{k+1}, \\ \Sigma_{k+1} = (\pi^{(\alpha)}(xx^\top))_{k+1} - \mu_{k+1}\mu_{k+1}^\top. \end{cases} \tag{5.3}$$

We now present the results, using $N = 10d$ samples per iteration. Figure 2 shows the performance of Algorithm 4, in terms of Rényi divergence value along iterations, when setting dimension $d = 20$, and condition number $\kappa_\pi = 10$. We observe that the best values of the Rényi divergences are obtained for the matched case $\nu = \nu_\pi$, which is expected. Note also that the Gaussian approximations (i.e., $\nu = +\infty$)

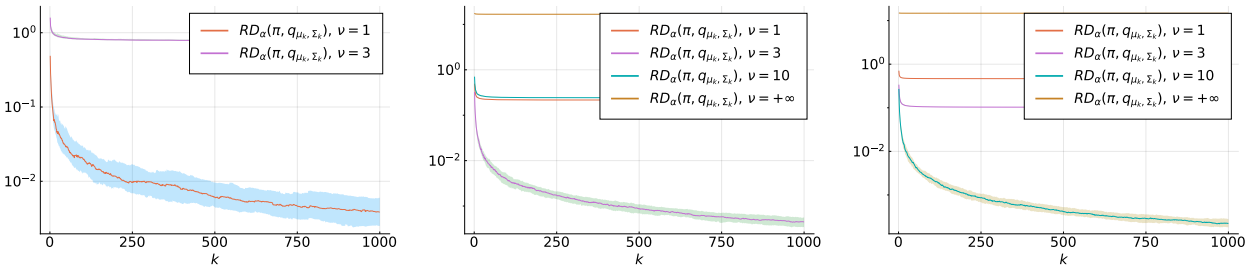


(a) $\nu_\pi = 1$, $\nu \in \{1, 3, 10\}$ (b) $\nu_\pi = 3$, $\nu \in \{1, 3, 10, +\infty\}$ (c) $\nu_\pi = 10$, $\nu \in \{1, 3, 10, +\infty\}$

Figure 2: Rényi divergence between $q_{\mu_k,\Sigma_k} \in \mathcal{T}_\nu^d$ and π in dimension $d = 20$ with $\kappa_\pi = 10$ at every iteration k. The iterates $q_{\mu_k,\Sigma_k} \in \mathcal{T}_\nu^d$ are obtained using Algorithm 4. The line is the median Rényi divergence per iteration and the shaded area is the interval between the first and third quartiles. The quartiles are obtained by running the algorithm for 100 runs of 1000 iterations.

perform very poorly. More generally, the closer ν is to ν_π, the better the performance. Remark that when $\nu_\pi = \nu = 1$, the values reached by the Rényi divergences are more spread around the median. This could be because the degree of freedom parameter of $\pi^{(\alpha)}$ in this case is the lowest, and hence, $\pi^{(\alpha)}$ has heavier tails. In Figure 2a, in the case when $\nu_\pi = 1$, some approximating families need to be excluded to comply with Equation (5.1). In particular, standard moment-matching, corresponding to $\nu = +\infty$ is not defined in this case. In constrast, as soon as $\nu_\pi > 2$, Equation (5.1) is satisfied for any $\nu > 0$, so any approximating family can be chosen, as it is done in the plots for Figures 2b and 2c.

In Figure 3, we show performance in dimension $d = 5$ and high condition number $\kappa_\pi = 1000$. Since the samples are generated directly from $\pi^{(\alpha)}$, the poor conditioning issue is mitigated. Since a low dimension has been used, more values of ν need to be excluded in order to comply with the condition in Equation (5.1) in the case $\nu_\pi = 1$.



(a) $\nu_\pi = 1$, $\nu \in \{1, 3\}$ (b) $\nu_\pi = 3$, $\nu \in \{1, 3, 10, +\infty\}$ (c) $\nu_\pi = 10$, $\nu \in \{1, 3, 10, +\infty\}$

Figure 3: Rényi divergence between $q_{\mu_k,\Sigma_k} \in \mathcal{T}_\nu^d$ and π in dimension $d = 5$ with $\kappa_\pi = 1000$ at every iteration k. The iterates $q_{\mu_k,\Sigma_k} \in \mathcal{T}_\nu^d$ are obtained using Algorithm 4. The line is the median Rényi divergence per iteration and the shaded area is the interval between the first and third quartiles. The quartiles are obtained by running the algorithm for 100 runs of 1000 iterations.

5.1.2 Using Metropolis-adjusted Langevin algorithms

We now consider a more practical resolution of Problem ($P_{\text{VI-Student}}$). We only assume that one has access to an oracle giving the unnormalized log-density $\log \tilde{\pi}$ such that for any $x \in \mathbb{R}^d$, $\log \pi(x) = \log \tilde{\pi}(x) - \log Z_\pi$ for some $Z_\pi > 0$. We also assume that one can evaluate the gradients $\nabla \log \tilde{\pi}(x)$ for any $x \in \mathbb{R}^d$. Under

these assumptions, we propose to perform the computation of $\pi^{(\alpha)}(x), \pi^{(\alpha)}(xx^\top)$ using a MALA approach, a particular Monte Carlo Markov Chain algorithm introduced in [41]. Let $x \in \mathbb{R}^d$ a starting point of the chain and suppose that we want to have samples approximately distributed following $\pi^{(\alpha)}$ for $\alpha > 0$. Then, MALA uses a proposal distribution of the form

$$y \sim \mathcal{N}\left(x + \frac{1}{2}\sigma_d^2 \alpha A \nabla \log \tilde{\pi}(x), \sigma_d^2 A\right). \tag{5.4}$$

A typical choice is $\sigma_d^2 = \frac{0.574^2}{d^{1/3}}$, following the optimal settings described in [40]. Moreover, hereabove, $A \in \mathcal{S}_{++}^d$ is the so-called scale matrix. The proposed sampled y is then accepted or not following a Metropolis-Hastings step. The scale matrix A in Equation (5.4) will be chosen either as the identity matrix leading to the standard MALA algorithm, or as to reflect the curvature of $\log \pi$ around the current point x, as it is done in [31, 30] for instance.

Standard MALA We first consider the direct approximation of the optimality conditions (4.3) by approximating the moments of $\pi^{(\alpha)}$ using samples generated with Equation (5.4) with $A = I_d$. This leads to Algorithm 5 described below.

Algorithm 5: Solving Problem ($P_{\text{VI-Student}}$) by approximating (4.3) with MALA

Choose an approximating family $\mathcal{T}_\nu^d$ and set $\alpha = 1 + \frac{2}{\nu+d}$. Choose $N \in \mathbb{N}$. Initialize x_0.

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

Sample $\{x_{k+1}^{(1)}, \dots, x_{k+1}^{(N)}\}$ samples from x_k using the MALA algorithm with proposal described in Equation (5.4) with $A = I_d$, set $x_{k+1} = x_{k+1}^{(N)}$.

Evaluate $(\pi^{(\alpha)}(x))_{k+1}, (\pi^{(\alpha)}(xx^\top))_{k+1}$ by

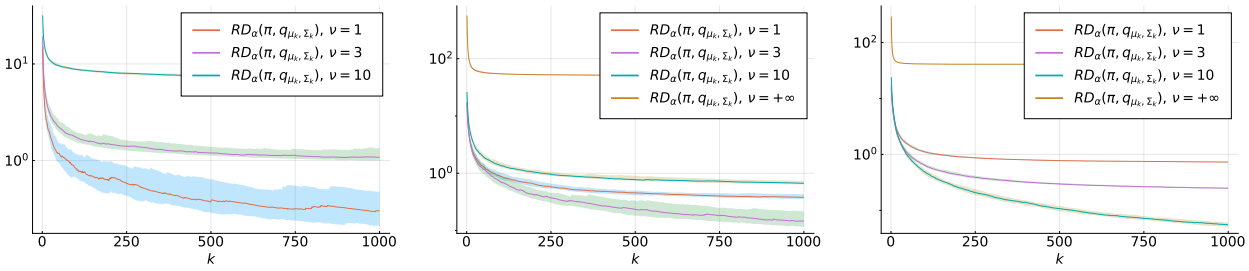
$$\begin{cases} (\pi^{(\alpha)}(x))_{k+1} &= \frac{1}{kN}\sum_{l=0}^{k}\sum_{i=1}^{N} x_{l+1}^{(i)}, \\ (\pi^{(\alpha)}(xx^\top))_{k+1} &= \frac{1}{kN}\sum_{l=0}^{k}\sum_{i=1}^{N} x_{l+1}^{(i)}(x_{l+1}^{(i)})^\top. \end{cases} \tag{5.5}$$

Compute μ_{k+1}, Σ_{k+1} following

$$\begin{cases} \mu_{k+1} = (\pi^{(\alpha)}(x))_{k+1}, \\ \Sigma_{k+1} = (\pi^{(\alpha)}(xx^\top))_{k+1} - \mu_{k+1}\mu_{k+1}^\top. \end{cases} \tag{5.6}$$

We now turn to the experiments on the parameters described previously. We set $N = 10d$ for each experiment and initialize x_0 by sampling it uniformly in $[-5, 5]^d$.

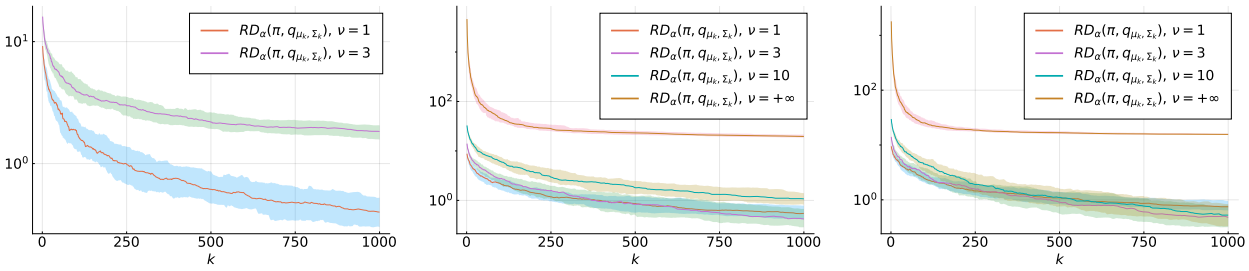
We display in Figure 4 the results obtained, for a target with low condition number $\kappa_\pi = 10$, in dimension $d = 20$. We can observe that, as in Section 5.1.1, the matched case $\nu = \nu_\pi$ yields the best results. Interestingly, the proposed MALA strategy works well even when the target is heavy-tailed. This could be surprising in light of negative results such as the ones in [23], but remark that we apply MALA on $\pi^{(\alpha)}$ and not π. Due to Equation (5.1), $\pi^{(\alpha)}$ has well-defined first and second order moments, which explains the good performance of the MALA algorithm in this case. This illustrates the interest of the optimality conditions that we prove in Proposition 7, as they allow to handle heavy-tailed targets just as if they were light-tailed.



(a) $\nu_\pi = 1$, $\nu \in \{1, 3, 10\}$ (b) $\nu_\pi = 3$, $\nu \in \{1, 3, 10, +\infty\}$ (c) $\nu_\pi = 10$, $\nu \in \{1, 3, 10, +\infty\}$

Figure 4: Rényi divergence between $q_{\mu_k,\Sigma_k} \in \mathcal{T}_\nu^d$ and π in dimension $d = 20$ with $\kappa_\pi = 10$ at every iteration k. The iterates $q_{\mu_k,\Sigma_k} \in \mathcal{T}_\nu^d$ are obtained using Algorithm 5. The line is the median Rényi divergence per iteration and the shaded area is the interval between the first and third quartiles. The quartiles are obtained by running the algorithm for 100 runs of 1000 iterations.

We now turn to a target with low dimension $d = 5$, whose scale matrix has condition number $\kappa_\pi = 1000$. This is challenging given that the proposal distribution in our MALA algorithm is isotropic. Figure 5 shows the results. Compared to the case of a low condition number in higher dimension, depicted in Figure 4, we observe that the values of the Rényi divergence are higher, sometimes by an order of magnitude. The dispersal of the values around the median is also more pronounced. This can be explained by the fact that in the standard MALA algorithm, the proposals are isotropic Gaussian distributions, and hence not well adapted to the target at hand. Note also that when ν_π grows, the negative impact of having $\nu \neq \nu_\pi$ seems to diminish, especially compared to the situation of Figure 4



(a) $\nu_\pi = 1$, $\nu \in \{1, 3\}$ (b) $\nu_\pi = 3$, $\nu \in \{1, 3, 10, +\infty\}$ (c) $\nu_\pi = 10$, $\nu \in \{1, 3, 10, +\infty\}$

Figure 5: Rényi divergence between $q_{\mu_k,\Sigma_k} \in \mathcal{T}_\nu^d$ and π in dimension $d = 5$ with $\kappa_\pi = 1000$ at every iteration k. The iterates $q_{\mu_k,\Sigma_k} \in \mathcal{T}_\nu^d$ are obtained using Algorithm 5. The line is the median Rényi divergence per iteration and the shaded area is the interval between the first and third quartiles. The quartiles are obtained by running the algorithm for 100 runs of 1000 iterations.

Scaled MALA As shown in Figure 5, the use of an isotropic proposal in MALA might not be well suited for a poorly conditioned target. We now consider the implementation of Algorithm 1 and the adaptation of the scale matrix A in the MALA sampling step (5.4). To do so, we exploit the approximation q_{μ_k,Σ_k} of $\pi^{(\alpha)}$ by setting $A = \Sigma_k$ at each iteration $k \in \mathbb{N}$. The approximating distribution at iteration $k \in \mathbb{N}$, q_{μ_k,Σ_k} is itself updated following Algorithm 1 with $\tau_k = \frac{1}{k}$ and $\pi^{(\alpha)}(T)$ being approximated by N samples from the Markov chain. Therefore, the scaling matrix is updated every N number of MALA steps and not at every iteration as in [31, 30]. The resulting procedure is detailed in Algorithm 6.

Algorithm 6: Solving Problem ($P_{\text{VI-Student}}$) with the updates (4.7) and scaled MALA.

Choose an approximating family $\mathcal{T}_\nu^d$ and set $\alpha = 1 + \frac{2}{\nu+d}$. Choose $N \in \mathbb{N}$. Initialize μ_0, Σ_0, and x_0.

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

Sample $\{x_{k+1}^{(1)}, \dots, x_{k+1}^{(N)}\}$ samples from x_k using the MALA algorithm with proposal as in Equation (5.4) with $A = \Sigma_k$, set $x_{k+1} = x_{k+1}^{(N)}$.

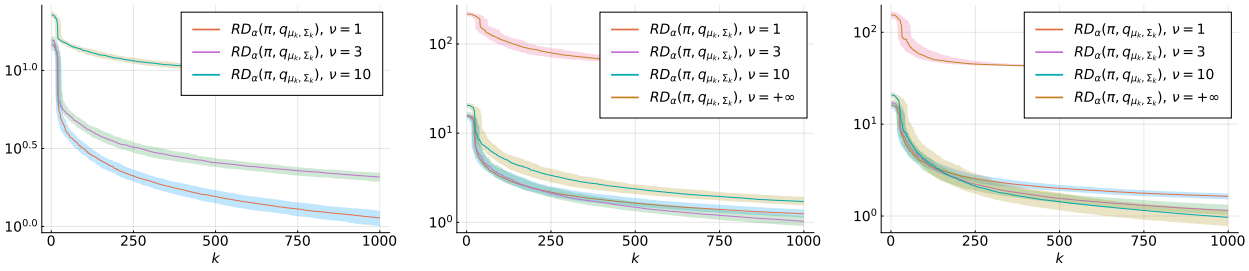
Evaluate $(\pi^{(\alpha)}(x))_{k+1}, (\pi^{(\alpha)}(xx^\top))_{k+1}$ by

$$\begin{cases} (\pi^{(\alpha)}(x))_{k+1} & = \frac{1}{N}\sum_{i=1}^N x_{k+1}^{(i)}, \\ (\pi^{(\alpha)}(xx^\top))_{k+1} & = \frac{1}{N}\sum_{i=1}^N x_{k+1}^{(i)}(x_{k+1}^{(i)})^\top. \end{cases} \tag{5.7}$$

Update μ_{k+1}, Σ_{k+1} following

$$\begin{cases} \mu_{k+1} = \frac{1}{k+1}(\pi^{(\alpha)}(x))_{k+1} + \frac{k}{k+1}\mu_k, \\ \Sigma_{k+1} = \frac{1}{k+1}(\pi^{(\alpha)}(xx^\top))_{k+1} + \frac{k}{k+1}(\Sigma_k + \mu_k\mu_k^\top) - \mu_{k+1}\mu_{k+1}^\top. \end{cases} \tag{5.8}$$

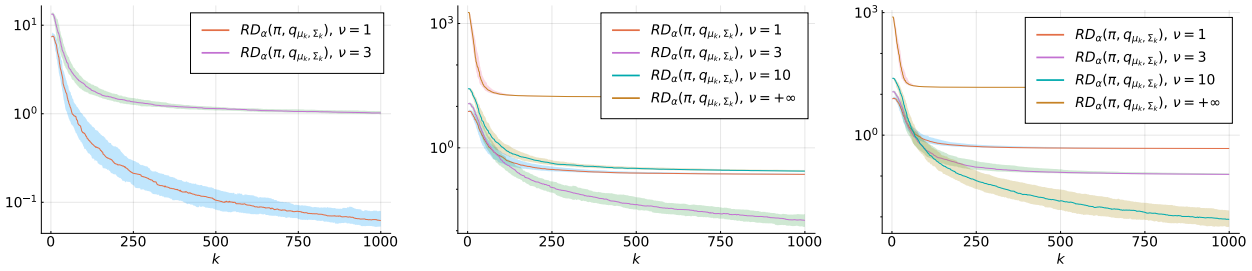
We now present our results, with $N = 10d$. For each run, we initialize the algorithm with $\Sigma_0 = I_d$, and $\mu_0 = x_0$ sampled uniformly in $[-5, 5]^d$. Figure 6 shows the performance of Algorithm 6 in dimension $d = 20$ on a well-conditioned target. As in the previous cases, we observe that the best performance are reached when the approximating family contains the target when $\nu_\pi = 1$, while performance get more similar for other choices of ν_π as soon as ν is close to ν_π.



(a) $\nu_\pi = 1$, $\nu \in \{1, 3, 10\}$ (b) $\nu_\pi = 3$, $\nu \in \{1, 3, 10, +\infty\}$ (c) $\nu_\pi = 10$, $\nu \in \{1, 3, 10, +\infty\}$

Figure 6: Rényi divergence between $q_{\mu_k,\Sigma_k} \in \mathcal{T}_\nu^d$ and π in dimension $d = 20$ with $\kappa_\pi = 10$ at every iteration k. The iterates $q_{\mu_k,\Sigma_k} \in \mathcal{T}_\nu^d$ are obtained using Algorithm 6. The line is the median Rényi divergence per iteration and the shaded area is the interval between the first and third quartiles. The quartiles are obtained by running the algorithm for 100 runs of 1000 iterations.

We now turn to a target π that has a higher condition number $\kappa_\pi = 1000$, displaying the results on Figure 7. We can observe that Algorithm 6 reaches better performance than Algorithm 5 on this poorly conditioned target. Compared to the case of Figure 6, the values reached when the approximating family contains the target are now much better than the ones obtained with the other approximating families.



(a) $\nu_\pi = 1$, $\nu \in \{1, 3\}$ (b) $\nu_\pi = 3$, $\nu \in \{1, 3, 10, +\infty\}$ (c) $\nu_\pi = 10$, $\nu \in \{1, 3, 10, +\infty\}$

Figure 7: Rényi divergence between $q_{\mu_k,\Sigma_k} \in \mathcal{T}_\nu^d$ and π in dimension $d = 5$ with $\kappa_\pi = 1000$ at every iteration k. The iterates $q_{\mu_k,\Sigma_k} \in \mathcal{T}_\nu^d$ are obtained using Algorithm 6. The line is the median Rényi divergence per iteration and the shaded area is the interval between the first and third quartiles. The quartiles are obtained by running the algorithm for 100 runs of 1000 iterations.

Synthesis of the results Table 1 summarizes the results, for the three algorithms, in terms of final value of the Rényi divergence, averaged over 100 runs, after 10^3 iterations. The results span the two scenarios, the first with a high-dimensional target with good conditioning, and the second with a poorly-conditioned target in lower dimension. The relative performance of Algorithms 5 and 6 depends on the scenario. On targets that are well-conditioned but high-dimensional, Algorithm 5 seems to behave better than Algorithm 6. On the contrary, Algorithm 6 yields here the best performance when the target is poorly conditioned, showing that our proposed scale adaptation mechanism is able to capture the geometry of the target distribution. We have also implemented the algorithms in the limit $\nu \to +\infty$, corresponding to the exponential family. In this case, the target is not properly captured by the proposals, showing the interest of our new result about the λ-exponential family. Finally, as expected, the idealized Algorithm 4 reaches the best results in most cases, confirming the validity of our optimality conditions.

		$\nu_\pi = 1$		$\nu_\pi = 3$		$\nu_\pi = 10$	
		High d	High κ_π	High d	High κ_π	High d	High κ_π
$\nu = 1$	Alg. 4	$2.25 \cdot 10^{-2}$	$3.84 \cdot 10^{-3}$	$2.61 \cdot 10^{-1}$	$2.15 \cdot 10^{-1}$	$6.81 \cdot 10^{-1}$	$4.65 \cdot 10^{-1}$
	Alg. 5	$\mathbf{3.01 \cdot 10^{-1}}$	$4.00 \cdot 10^{-1}$	$\mathbf{3.80 \cdot 10^{-1}}$	$5.45 \cdot 10^{-1}$	$\mathbf{7.30 \cdot 10^{-1}}$	$7.48 \cdot 10^{-1}$
	Alg. 6	$1.13 \cdot 10^{0}$	$\mathbf{6.22 \cdot 10^{-2}}$	$1.25 \cdot 10^{0}$	$\mathbf{2.29 \cdot 10^{-1}}$	$1.64 \cdot 10^{0}$	$\mathbf{4.72 \cdot 10^{-1}}$
$\nu = 3$	Alg. 4	$7.02 \cdot 10^{-1}$	$7.78 \cdot 10^{-1}$	$1.49 \cdot 10^{-3}$	$4.50 \cdot 10^{-4}$	$1.99 \cdot 10^{-1}$	$1.03 \cdot 10^{-1}$
	Alg. 5	$\mathbf{1.08 \cdot 10^{0}}$	$1.83 \cdot 10^{0}$	$\mathbf{1.46 \cdot 10^{-1}}$	$4.18 \cdot 10^{-1}$	$\mathbf{2.49 \cdot 10^{-1}}$	$4.76 \cdot 10^{-1}$
	Alg. 6	$2.08 \cdot 10^{0}$	$\mathbf{1.02 \cdot 10^{0}}$	$1.03 \cdot 10^{0}$	$\mathbf{1.78 \cdot 10^{-2}}$	$1.15 \cdot 10^{0}$	$\mathbf{1.10 \cdot 10^{-1}}$
$\nu = 10$	Alg. 4	$5.83 \cdot 10^{0}$	$\times$	$4.69 \cdot 10^{-1}$	$2.41 \cdot 10^{-1}$	$7.08 \cdot 10^{-4}$	$2.24 \cdot 10^{-4}$
	Alg. 5	$\mathbf{7.20 \cdot 10^{0}}$	$\times$	$\mathbf{6.68 \cdot 10^{-1}}$	$1.07 \cdot 10^{0}$	$\mathbf{5.50 \cdot 10^{-2}}$	$5.25 \cdot 10^{-1}$
	Alg. 6	$9.36 \cdot 10^{0}$	$\times$	$1.71 \cdot 10^{0}$	$\mathbf{2.74 \cdot 10^{-1}}$	$9.67 \cdot 10^{-1}$	$\mathbf{8.63 \cdot 10^{-3}}$
$\nu = +\infty$	Alg. 4	$\times$	$\times$	$4.91 \cdot 10^{1}$	$1.67 \cdot 10^{1}$	$4.06 \cdot 10^{1}$	$1.48 \cdot 10^{1}$
	Alg. 5	$\times$	$\times$	$\mathbf{5.01 \cdot 10^{1}}$	$1.96 \cdot 10^{1}$	$\mathbf{4.07 \cdot 10^{1}}$	$1.56 \cdot 10^{1}$
	Alg. 6	$\times$	$\times$	$5.74 \cdot 10^{1}$	$\mathbf{1.69 \cdot 10^{1}}$	$4.18 \cdot 10^{1}$	$\mathbf{1.48 \cdot 10^{1}}$

Table 1: Median of the Rényi divergence $RD_\alpha(\pi, q_{\mu_K,\Sigma_K})$ over 100 runs of $K = 10^3$ iteration. "High d" corresponds to $d = 20, \kappa_\pi = 10$ and "High κ_π" to $d = 5, \kappa_\pi = 10^3$. The symbol $\times$ denotes situations when Equation (5.1) is not satisfied. For each target and each approximating family $\mathcal{T}_\nu^d$, we highlighted in bold font the algorithm achieving the lowest value between Algorithm 5 and 6. The values obtained with the idealized Algorithm 4 are indicated as a reference.

We can observe on Table 1 that Algorithms 5 and 6 yield lower performance than Algorithm 4. However, implementing this last algorithm is unrealistic in practice, as it needs samples from the escort of the target. However, we can notice that, when $\nu_\pi \neq \nu$, i.e., the approximating family does not match with the target, the algorithms based on MALA are able to reach similar performance than Algorithm 4.

We see in Table 1 that Algorithm 6 outperforms Algorithm 5 on the high κ_π scenario, sometimes by one or two orders of magnitude. This gain can be explained by the fact that Algorithm 6 better handles the shape of the target. This indicates that as soon as the target may be poorly conditioned, it is best to turn to Algorithm 6 instead of Algorithm 5.

On the other hand, the situation is reversed on the high d scenario, where the performance of Algorithm 6 decreases. This indicates that on high-dimensional and well-conditioned targets, it may be beneficial to use Algorithm 5 instead of its scaled version, in Algorithm 6.

Finally, let us mention that when the algorithm matches the scenario, that is Algorithm 5 is used for high d or Algorithm 6 is used for high κ_π, it is especially important to choose $\nu = \nu_\pi$. Indeed, this is when we observe the biggest degradation if $\nu \neq \nu_\pi$.

5.2 Maximum likelihood estimation with Student distributions

We consider now maximum likelihood problems of the form (P_{MLE}) and ($P_{\text{MLE-Mixt}}$) over the λ-exponential family $\mathcal{Q}_\lambda$. We will work in the case where $\mathcal{Q}_\lambda$ is the Student family $\mathcal{T}_\nu^d$.

5.2.1 Online maximum likelihood with approximate proximal updates

We now consider a maximum likelihood estimation problem of the form (P_{MLE}). The approximating family is hereagain the Student family, $\mathcal{T}_\nu^d$. The samples processed for the maximum likelihood estimation are also distributed following a Student distribution $\pi \in \mathcal{T}_\nu^d$. Following [24], we consider an online setting, where one sample is delivered at each iteration of the algorithm. We implement Algorithm 2 in this setting and study how they approach the true maximum likelihood estimator, depending on the value of λ.

We assume that at every iteration $k \in \mathbb{N}$ one point $x_k \sim \pi$ is sampled. We implement Algorithm 2 and apply, at each iteration, the operator $P_{\tau_k}^{\{x_k\}}$, with a single data point, namely x_k, and we set $\tau_k = \frac{1}{k}$, ensuring an averaging effect. In our setting, this leads to Algorithm 7.

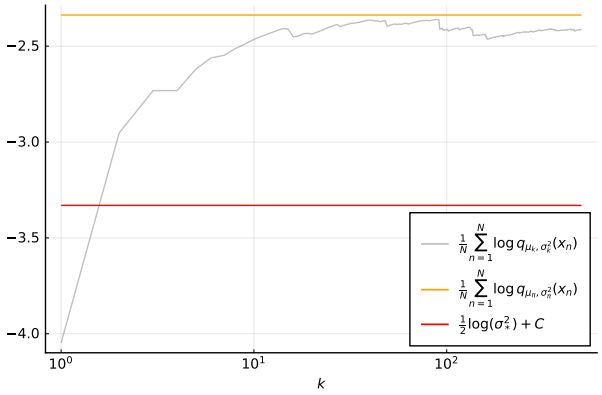
Algorithm 7: Online algorithm to solve Problem (P_{MLE}) on Student families.

Choose an approximating family $\mathcal{T}_\nu^d$ and initialize μ_0 and Σ_0.
for $k = 0, \dots$ **do**
Using the new sample x_k, update μ_{k+1}, Σ_{k+1} following

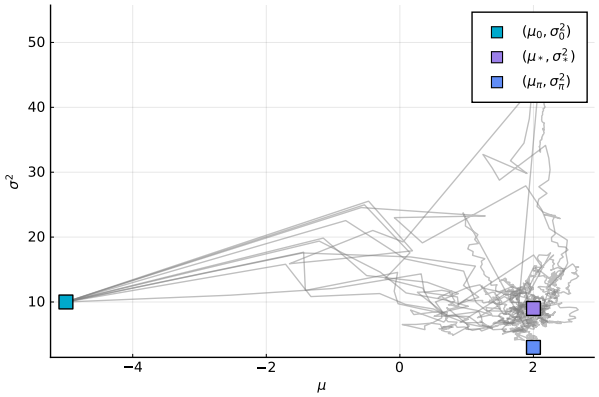
$$\begin{cases} \mu_{k+1} = \frac{1}{k+1} x_k + \frac{k}{k+1} \mu_k, \\ \Sigma_{k+1} = \frac{1}{k+1} x_k x_k^\top + \frac{k}{k+1} (\Sigma_k + \mu_k \mu_k^\top) - \mu_{k+1} \mu_{k+1}^\top. \end{cases} \tag{5.9}$$

As discussed in Section 4.3, Algorithm 7 cannot exactly recover the parameters (μ_π, Σ_π) of the distribution of the data points, even when $k \to +\infty$. From Propositions 11 and 4, the sequence $\{(\mu_k, \Sigma_k)\}_{k \in \mathbb{N}}$ converges to (μ_*, Σ_*) satisfying $\mu_* = \mu_\pi$ and $\Sigma_* = \frac{\nu}{\nu - 2} \Sigma_\pi$, provided that $\nu > 2$.

We illustrate the behavior of Algorithm 7 by showing several runs of it, in dimension $d = 1$, with $\nu \in \{3, 10\}$. This yields trajectories in the plane (μ, σ^2). In the Gaussian case, recovered when $\nu \to +\infty$, we have from Corollary 2 that $q_* = \pi$. Trying different values of ν allows to explore situations that are far from the Gaussian setting when $\nu = 3$, or closer to it when $\nu = 10$. In the latter case, we expect a lower mismatch between π and q_*.



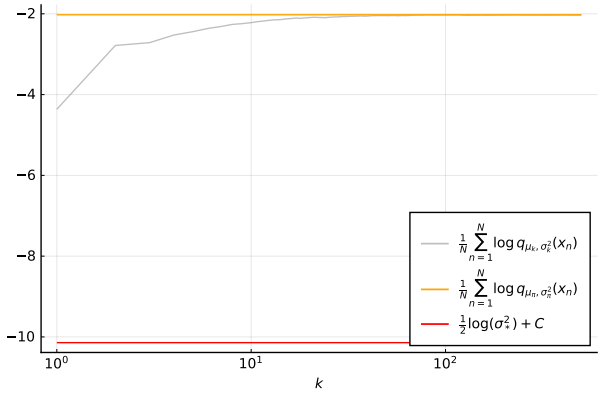
(a) Plot of the log-likelihood of one trajectory of Algorithm 7, with the log-likelihood of the true parameters in orange and the bound of Corollary 2 in red.



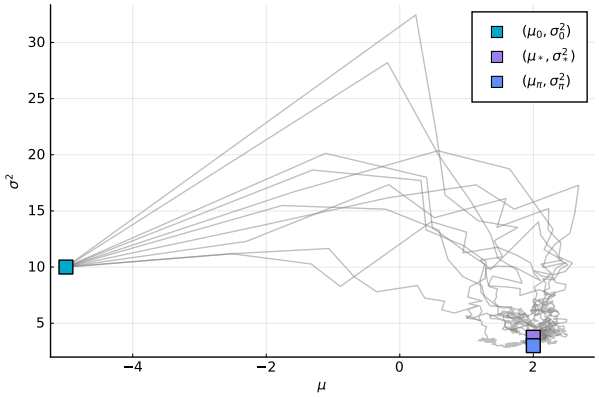
(b) Plot of 10 trajectories of Algorithm 7, with the point (μ_*, σ_*^2) to which the trajectories converge and the point (μ_π, σ_π^2) encoding the distribution of the samples.

Figure 8: Plots of trajectories of Algorithm 7, initialized at $\mu_0 = -2$ and $\sigma_0^2 = 10$, in dimension $d = 1$, with samples generated following $\pi \in \mathcal{T}_\nu^d$, $\nu = 3$, with parameters (μ_π, σ_π^2).

Figure 8 shows that the iterates $\{(\mu_k, \sigma_k^2)\}_{k\in\mathbb{N}}$ generated by Algorithm 7 converge to the point (μ_*, σ_*^2), which is different from the true parameters (μ_π, σ_π^2). We can also observe in this figure that the log-likelihood of the iterates gets very close to the one of π. The bound on the sub-optimal log-likelihood, predicted by Proposition 10 and Corollary 2, is satisfied by the iterates $\{(\mu_k, \sigma_k^2)\}_{k\in\mathbb{N}}$ after a small number of iterations.



(a) Plot of the log-likelihood of one trajectory of Algorithm 7, with the log-likelihood of the true parameters in orange and the bound of Corollary 2 in red.



(b) Plot of 10 trajectories of Algorithm 7, with the point (μ_*, σ_*^2) to which the trajectories converge and the point (μ_π, σ_π^2) encoding the distribution of the samples.

Figure 9: Plots of trajectories of Algorithm 7 initialized at $\mu_0 = -2$ and $\sigma_0^2 = 10$, in dimension $d = 1$, with samples generated following $\pi \in \mathcal{T}_\nu^d$, $\nu = 10$, with parameters (μ_π, σ_π^2).

Figure 9 considers a higher value of ν. This setting is closer to the Gaussian case, reached in the limit $\nu \to +\infty$, for which our algorithm reaches the true distribution of the samples. We thus observe that in Figure 9, the log-likelihood converge to the value of the log-likelihood of π. This is in contrast with Figure

8, in which we can observe gap. We again observe that the iterates converge to the point (μ_*, σ_*^2), which is very close this time to the true parameters (μ_π, σ_π^2). Compared to Figure 8 in the case $\nu = 3$, we see that the bound predicted by Corollary 2 is reached from the first iterates, meaning that it is not a tight bound for the log-likelihood of q_*.

According to our theoretical results, Algorithm 7 converges to a sub-optimal solution of Problem (P_{MLE}). Such solution is very easy to implement and could be used to initialize a more complex but exact maximum likelihood estimation algorithm [21, 4]. Moreover, the obtained sub-optimal solution has links with the probability distribution that generated the data, as discussed in Section 4.3 and thus remains relevant for computing exact maximum likelihood estimators.

5.2.2 Maximum likelihood estimation with mixtures using relaxed EM

We consider here a maximum likelihood estimation problem over a mixture of Student distributions, that is, Problem ($P_{\text{MLE-Mixt}}$) where $Q_\lambda = \mathcal{T}_\nu^d$. The samples are also considered to be distributed from a mixture of Student distributions from $\mathcal{T}_\nu^d$, denoted by π such that $\pi = \sum_{j=1}^J \xi_{*,j} q_{\mu_{*,j},\Sigma_{*,j}}$. We implement the relaxed EM method described in Algorithm 3 in this particular case. The resulting scheme is summarized in Algorithm 8. Algorithm 3 only requires to work with a λ-exponential family satisfying Assumptions 1 and 2, so Algorithm 8 is a particular instance for a specific choice of family. We notice that Student distributions benefit from specific properties that would also allow the design of exact EM algorithms [21], so we aim here at illustrating as a proof of concept the use of our mixture-based algorithm.

Algorithm 8: A sub-optimal EM algorithm to solve Problem ($P_{\text{MLE-Mixt}}$) on Student families.

Choose an approximating family $\mathcal{T}_\nu^d$. Initialize Let $\mu_{0,j} \in \mathbb{R}^d$, $\Sigma_{0,j} \in \mathcal{S}_{++}^d$, and $\xi_{0,j} \geq 0$ for $j = 1, \dots, J$ such that $\sum_{j=1}^J \xi_{0,j} = 1$.

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

For every $j = 1, \dots, J$, define the function $\gamma_{k,j}$ following Equation (4.20), and update the parameters $\xi_{k+1,j}$ and $\mu_{k+1,j}, \Sigma_{k+1,j}$ such that they satisfy

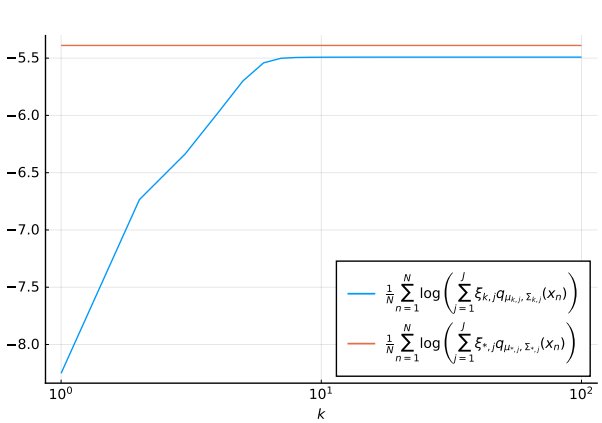
$$\xi_{j,k+1} = \frac{1}{N} \sum_{i=1}^N \gamma_{k,j}(x_i), \tag{5.10}$$

$$\begin{cases} \mu_{k+1,j} = \sum_{i=1}^N \frac{\gamma_{k,j}(x_i)}{\sum_{i'=1}^N \gamma_{k,j}(x_{i'})} x_i, \\ \Sigma_{k+1,j} = \sum_{i=1}^N \frac{\gamma_{k,j}(x_i)}{\sum_{i'=1}^N \gamma_{k,j}(x_{i'})} x_i x_i^\top - \mu_{k+1,j} \mu_{k+1,j}^\top. \end{cases} \tag{5.11}$$

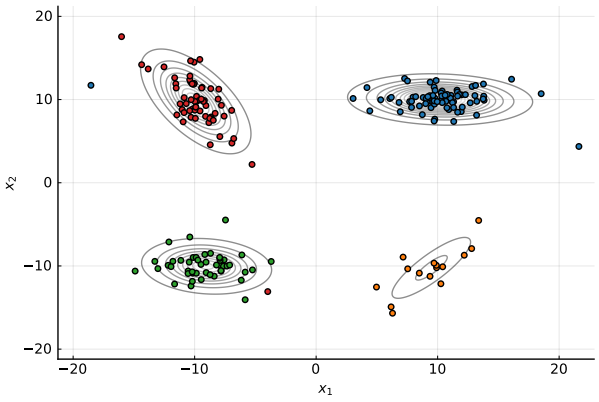
We illustrate the behavior of Algorithm 8 in dimension $d = 2$, with $\nu \in \{3, 10\}$. Note that a greater value of ν corresponds to a value of λ closer to 0, in which case the approximate M-steps are closer to being optimal (they are optimal for $\lambda = 0$). We use $N = 200$ samples, from $\pi = \sum_{j=1}^J \xi_{*,j} q_{\mu_{*,j},\Sigma_{*,j}}$ with $J = 4$. We use $\{\xi_{*,j}\}_{j=1}^J = \{0.4, 0.1, 0.2, 0.3\}$, with locations parameters $\mu_{*,1} = (10, 10)^\top$, $\mu_{*,2} = (-10, 10)^\top$, $\mu_{*,3} = -\mu_{*,1}$, and $\mu_{*,4} = -\mu_{*,3}$. The shape matrices $\Sigma_{*,j}$, $j = 1, \dots J$ are constructed in $\mathcal{S}_{++}^d$ with condition number $\kappa = 10$ following [34]. This is a controlled setting which allows to observe precisely the behavior of Algorithm 8 (i.e., an instance of Algorithm 3). Algorithm 8 is initialized with mixture weights satisfying $\xi_{j,0} = 1/J$ for $j = 1, \dots, J$, initial locations parameters $\mu_{j,0}$ sampled from a normal distribution with zero mean and covariance $10I$ and shape parameters being $\Sigma_{j,0} = 10I$ for $j = 1, \dots, J$.

Figure 10 shows the performance of Algorithm 8 when mixture components are from $\mathcal{T}_\nu^d$, with $\nu = 3$. The resulting mixture is able to identify the different components of the data-generating distribution π and to achieve a significant increase in terms of log-likelihood from initialization. In this setting, the suboptimality in solving the M-step of the EM algorithm is more pronounced, as the iterates generated by the algorithm

cannot reach the log-likelihood achieved by the data-generating distribution. This is to be expected, as the corresponding value of λ is far from $\lambda = 0$ where the M-step is optimal.



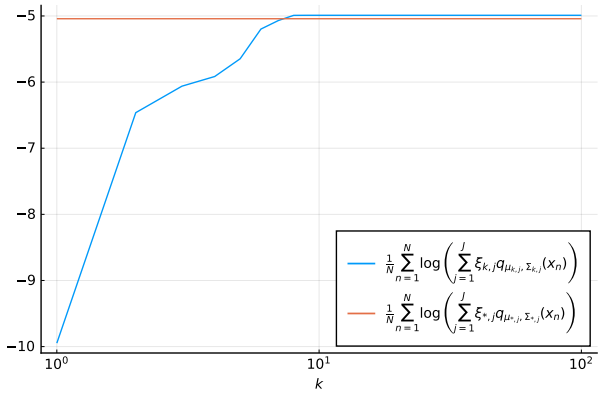
(a) Plot of the log-likelihood achieved by the iterates of Algorithm 8, with the log-likelihood of data-generating distribution in orange.



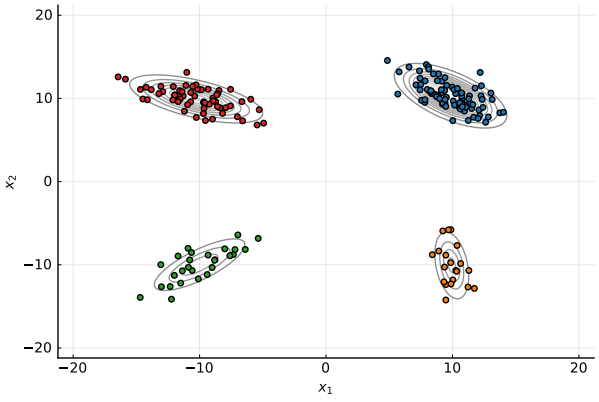
(b) Plot of the samples generated by π, the different colors denoting the component of the mixture from which they have been drawn. The level lines of the final distribution generated by Algorithm 8 are shown in grey.

Figure 10: Plot of the performance achieved by Algorithm 8 after $K = 100$ iterations in terms of log-likelihood and graphical representation in the sample space, for mixtures of distributions in $\mathcal{T}_\nu^d$, with $\nu = 3$ and $d = 2$.

Figure 11 shows the performance of Algorithm 8 when mixture components are from $\mathcal{T}_\nu^d$, with $\nu = 10$. We observe that the proposed algorithm generates iterates whose log-likelihood matches the one of the data-generating distribution. Indeed, this setting is closer to the case $\lambda = 0$ where our algorithm solves the M-step in the EM algorithm exactly, showing that the sub-optimality has no severe effect in this case.



(a) Plot of the log-likelihood achieved by the iterates of Algorithm 8, with the log-likelihood of data-generating distribution in orange.



(b) Plot of the samples generated by π, the different colors denoting the component of the mixture from which they have been drawn. The level lines of the final distribution generated by Algorithm 8 are shown in grey.

Figure 11: Plot of the performance achieved by Algorithm 8 after $K = 100$ iterations in terms of log-likelihood and graphical representation in the sample space, for mixtures of distributions in $\mathcal{T}_\nu^d$, with $\nu = 10$ and $d = 2$.

6 Conclusion

In this work, we have studied variational inference and maximum likelihood estimation problems over the λ-exponential family, and we have proposed algorithms to solve theses problems. Several known results on the standard exponential family are retrieved as special cases.

First, we have shown that variational inference problems over the λ-exponential family can be solved by satisfying a generalized moment-matching condition that extends the existing one for the standard exponential family. We have also proposed an iterative algorithm to solve this problem, which identifies with a Bregman proximal algorithm in the particular case of the exponential family. The usefulness of our optimality conditions and our algorithm is confirmed by numerical experiments on heavy-tailed targets. These experiments show that the λ-exponential family can be used to capture phenomenon that the standard exponential family fails to represent.

Second, in maximum likelihood estimation problems, we exhibited sub-optimal solutions with a novel algorithm converging to it. In the case of the exponential family, the solutions become optimal and the algorithm reads again as a Bregman proximal algorithm. In the general case, our algorithm is quick and easy to implement, as demonstrated through numerical experiments. For problems with mixtures, we also proposed a relaxed EM algorithm that recovers the standard EM algorithm in the case of the exponential family. An interesting line of research would be the combination of our algorithms, which leads to sub-optimal solutions, with exact methods.

We achieved these results by extending convex analysis notions to a more general setting, replacing the scalar product by a well-chosen non-linear coupling. By leveraging the specific structure of the problems we consider, we have been able to exhibit optimality conditions and proximal-like algorithm using such tools, which is one of the main novelties of our work. Extending our results and techniques to more general problems, including other divergences and distances over probabilities, or more general couplings, related for instance with elliptical distributions, seems to be an exciting area of research.

A Proof of Proposition 3

Proof of Proposition 3. (i) Consider a distribution in $\mathcal{T}_\nu^d$ with location parameter μ and scale matrix Σ. Then we compute for any $x \in \mathbb{R}^d$ the following.

$$\begin{aligned} q_{\mu,\Sigma}(x) &\propto \left(1 + \frac{1}{\nu}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)^{-\frac{\nu+d}{2}} \\ &\propto \left(1 + \frac{1}{\nu}\mu^\top\Sigma^{-1}\mu - \frac{2}{\nu}\mu^\top x + \frac{1}{\nu}x^\top\Sigma^{-1}x\right)^{-\frac{\nu+d}{2}} \\ &\propto \left(1 + \frac{1}{\nu}\mu^\top\Sigma^{-1}\mu\right)^{-\frac{\nu+d}{2}} \left(1 + \left(-\frac{2}{\nu+d}\right)\left(-\frac{\nu+d}{2\nu(1+\frac{1}{\nu}\mu^\top\Sigma^{-1}\mu)}\left(-2\mu^\top x + x^\top\Sigma^{-1}x\right)\right)\right)^{-\frac{\nu+d}{2}} \end{aligned}$$

and since $\mu^\top x = \langle \mu, x\rangle$ and $x^\top\Sigma^{-1}x = \langle \Sigma^{-1}, xx^\top\rangle$, we can identify that $q_{\mu,\Sigma} = q_\vartheta$.

We can identify from the above that $\mathcal{T}_\nu^d$ is an instance of the λ-exponential family with $\lambda = -\frac{2}{\nu+d}$. Its parameters are

$$\vartheta_1 = \frac{\nu+d}{\nu+\mu^\top\Sigma^{-1}\mu}\Sigma^{-1}\mu, \quad \vartheta_2 = -\frac{\nu+d}{2(\nu+\mu^\top\Sigma^{-1}\mu)}\Sigma^{-1}. \tag{A.1}$$

In order to compute φ_λ, let us inverse the mapping $\mu, \Sigma \longmapsto \vartheta_1, \vartheta_2$. First, we can easily compute that $\mu = -\frac{1}{2}\vartheta_2^{-1}\vartheta_1$. Now, we compute the intermediate quantity $\mu^\top\Sigma^{-1}\mu$. Remark that

$$\vartheta_1^\top\vartheta_2^{-1}\vartheta_1 = -\frac{2(\nu+d)}{\nu+\mu^\top\Sigma^{-1}\mu}\mu^\top\Sigma^{-1}\mu. \tag{A.2}$$

Hence we deduce that

$$\nu + \mu^\top\Sigma^{-1}\mu = \frac{2\nu(\nu+d)}{2(\nu+d) + \vartheta_1^\top\vartheta_2^{-1}\vartheta_1}. \tag{A.3}$$

From Equations (A.1) and (A.3), it comes that $\Sigma^{-1} = -\frac{4\nu}{2(\nu+d)+\vartheta_1^\top\vartheta_2^{-1}\vartheta_1}\vartheta_2$. Summarizing our results, we thus obtained

$$\mu = -\frac{1}{2}\vartheta_2^{-1}\vartheta_1, \quad \Sigma = -\frac{2(\nu+d) + \vartheta_1^\top\vartheta_2^{-1}\vartheta_1}{4\nu}\vartheta_2^{-1}. \tag{A.4}$$

Finally, we turn to the computation of φ_λ. We can identify

$$\begin{aligned} \varphi_\lambda(\vartheta) &= -\log\left(\frac{\det(\Sigma)^{-\frac{1}{2}}}{Z_\nu}\left(1 + \frac{1}{\nu}\mu\Sigma^{-1}\mu^\top\right)^{-\frac{\nu+d}{2}}\right) \\ &= \frac{1}{2}\log\det(\Sigma) + \frac{\nu+d}{2}\log\left(1 + \frac{1}{\nu}\mu\Sigma^{-1}\mu^\top\right) + \log(Z_\nu) \\ &= \frac{d}{2}\log\left(\frac{2(\nu+d) + \vartheta_1^\top\vartheta_2\vartheta_1}{4\nu}\right) + \frac{1}{2}\log\det(-\vartheta_2^{-1}) + \frac{\nu+d}{2}\log\left(\frac{2(\nu+d)}{2(\nu+d)+\vartheta_1^\top\vartheta_2\vartheta_1}\right) + \log Z_\nu \\ &= -\frac{d}{2}\log(4\nu) + \frac{1}{2}\log\det(-\vartheta_2^{-1}) + \frac{\nu+d}{2}\log(2(\nu+d)) - \frac{\nu}{2}\log(2(\nu+d) + \vartheta_1^\top\vartheta_2^{-1}\vartheta_1) + \log Z_\nu. \end{aligned}$$

This shows in particular that $\operatorname{dom}\varphi_\lambda(\vartheta) = \{\vartheta \in \mathbb{R}^d \times \mathcal{S}_{--}^d,\ 2(\nu+d) + \vartheta_1^\top\vartheta_2^{-1}\vartheta_1 > 0\}$, which is non-empty. This shows that $\mathcal{T}_\nu^d$ satisfies Assumption 1.

(*ii*) We now turn to the study of the escort probabilities. We can compute for $x \in \mathbb{R}^d$ the following:

$$\begin{aligned} q^{(\alpha)}_{\mu,\Sigma}(x) &= \frac{1}{Z^{(\alpha)}} \left(1 + \frac{1}{\nu}(x-\mu)\Sigma^{-1}(x-\mu)^\top\right)^{-\alpha\frac{\nu+d}{2}} \\ &= \frac{1}{Z^{(\alpha)}} \left(1 + \frac{1}{\nu+2}(x-\mu)\left(\frac{\nu}{\nu+2}\Sigma\right)^{-1}(x-\mu)^\top\right)^{-\frac{(\nu+2)+d}{2}}. \end{aligned}$$

We recognize that $q^{(\alpha)}_{\mu,\Sigma}$ is a Student distribution with $\nu+2 > 2$ degrees of freedom, location parameter μ and scale matrix $\frac{\nu}{\nu+2}\Sigma$. Hence, we obtain that

$$\begin{cases} q^{(\alpha)}_{\mu,\Sigma}(x) & = \mu, \\ q^{(\alpha)}_{\mu,\Sigma}((x-\mu)(x-\mu)^\top) & = \frac{\nu+2}{(\nu+2)-2}\left(\frac{\nu}{\nu+2}\Sigma\right) = \Sigma. \end{cases} \tag{A.5}$$

To show the bijection result, we show that the map $(\mu,\Sigma) \longmapsto (\vartheta_1,\vartheta_2)$ is a bijection between $\mathbb{R}^d \times \mathcal{S}^d_{++}$ and $\operatorname{dom}\varphi_\lambda$. Consider $\mu \in \mathbb{R}^d$, $\Sigma \in \mathcal{S}^d_{++}$ and ϑ_1,ϑ_2 defined as in Equation (A.1). We can first remark that $\vartheta_1 \in \mathbb{R}^d$ and that $\vartheta_2 \in \mathcal{S}^d_{--}$. Using the result of Equation (A.2), we now compute

$$\begin{aligned} 2(\nu+d) + \vartheta_1^\top \vartheta_2^{-1}\vartheta_1 &= 2(\nu+d) - \frac{2(\nu+d)}{\nu+\mu^\top\Sigma^{-1}\mu}\mu^\top\Sigma^{-1}\mu \\ &= \frac{2\nu(\nu+d)}{\nu+\mu^\top\Sigma^{-1}\mu} > 0, \end{aligned}$$

showing that $\vartheta_1,\vartheta_2 \in \operatorname{dom}\varphi_\lambda$. Consider now $\vartheta_1,\vartheta_2 \in \operatorname{dom}\varphi_\lambda$, and μ,Σ as given by Equation (A.4). By definition of $\operatorname{dom}\varphi_\lambda$, $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathcal{S}^d_{++}$, showing the result.

We now compute the Rényi entropy of $q_{\mu,\Sigma} \in \mathcal{T}^d_\nu$ for $\alpha = 1-\lambda$ with $\lambda = -\frac{2}{\nu+d}$. By using similar steps as above, we obtain

$$H_\alpha(q_{\mu,\Sigma}) = \frac{1}{1-\alpha}\log\left(\frac{1}{Z^\alpha_\nu(\det\Sigma)^{\frac{\alpha}{2}}} Z_{\nu+2}\det\left(\frac{\nu}{\nu+2}\Sigma\right)^{\frac{1}{2}}\right). \tag{A.6}$$

(*iii*) Consider $\vartheta \in \operatorname{dom}\varphi_\lambda$, and $p \in \mathcal{P}(\mathcal{X},dx)$. Consider $\mu,\Sigma \in \mathbb{R}^d \times \mathcal{S}^d_{++}$ given by Equation (A.4). We can then compute

$$1 + \lambda\langle\vartheta, p^{(\alpha)}(T)\rangle = 1 - \frac{2}{\nu+\mu^\top\Sigma^{-1}\mu}p^{(\alpha)}(x)^\top\Sigma^{-1}\mu + \frac{1}{\nu+\mu^\top\Sigma^{-1}\mu}\operatorname{tr}(\Sigma^{-1}p^{(\alpha)}(xx^\top)), \tag{A.7}$$

which is defined if $p^{(\alpha)}$ has finite first and second order moments. Introducing the quantity $V := p^{(\alpha)}((x-p^{(\alpha)}(x))(x-p^{(\alpha)}(x))^\top) = p^{(\alpha)}(xx^\top) - p^{(\alpha)}(x)p^{(\alpha)}(x)^\top \in \mathcal{S}^d_+$, we get for any $\vartheta \in \operatorname{dom}\varphi_\lambda$ that

$$\begin{aligned} 1 + \lambda\langle\vartheta, p^{(\alpha)}(T)\rangle &= 1 - \frac{2}{\nu+\mu^\top\Sigma^{-1}\mu}p^{(\alpha)}(x)^\top\Sigma^{-1}\mu + \frac{1}{\nu+\mu^\top\Sigma^{-1}\mu}\operatorname{tr}(\Sigma^{-1}V) \\ &\quad + \frac{1}{\nu+\mu^\top\Sigma^{-1}\mu}p^{(\alpha)}(x)^\top\Sigma^{-1}p^{(\alpha)}(x) \\ &= \frac{1}{\nu+\mu\Sigma^{-1}\mu}\left(\nu + (\mu-p^{(\alpha)}(x))^\top\Sigma^{-1}(\mu-p^{(\alpha)}(x))^\top + \operatorname{tr}(\Sigma^{-1}V)\right) \\ &\geq \frac{\nu}{\nu+\mu\Sigma^{-1}\mu}. \end{aligned}$$

This shows that for any $\vartheta \in \operatorname{dom}\varphi_\lambda$, and $p^{(\alpha)} \in \mathcal{P}(\mathcal{X},dx)$ with finite first and second order moments, the quantity $c_\lambda(\vartheta, p^{(\alpha)}(T))$ is in $\mathbb{R}$. With the result of (*ii*), this shows that $\mathcal{T}^d_\nu$, seen as an instance of the λ-exponential family, satisfies Assumptions 2. □

References

[1] O. Akyildiz and J. Míguez. Convergence rates for optimised adaptive importance samplers. *Statistics and Computing*, 31(12), 2021.

[2] S.-I. Amari. *Differential-Geometrical Methods in Statistics*. Springer New York, 1985.

[3] S.-I. Amari and A. Ohara. Geometry of q-exponential family of probability distributions. *Entropy*, 13(6):1170–1185, 2011.

[4] I. Ayadi, F. Bouchard, and F. Pascal. Elliptical Wishart distribution: Maximum likelihood estimator from information geometry. In *IEEE International Conference on Speech, Acoustics and Signal Processing (ICASSP)*, 2023.

[5] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6(58):1705–1749, 2005.

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

[7] H. Bauschke, J. Borwein, and P. Combettes. Bregman monotone optimization algorithms. *SIAM Journal on Control and Optimization*, 42(2):596–636, 2003.

[8] H. Bauschke and P. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2011.

[9] E. M. Bednarczuk and M. Syga. On duality for nonconvex minimization problems within the framework of abstract convexity. *Optimization*, 71(4):949–971, 2022.

[10] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[11] 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 Processing Magasine*, 34(4):60–79, 2017.

[12] L. L. Campbell. Equivalence of Gauss's principle and minimum discrimination information estimation of probabilities. *Annals of Mathematical Statistics*, 41(3):1011–1015, 1970.

[13] 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.

[14] J.-P. Chancelier and M. De Lara. Constant along primal rays conjugacies and the ℓ_0 pseudonorm. *Optimization*, 71(2):355–386, 2020.

[15] G. Chantas, N. Galatsanos, A. Likas, and M. Saunders. Variational Bayesian image restoration based on a product of t-distributions image prior. *IEEE Transactions on Image Processing*, 17(10):1795–1805, 2008.

[16] P. Douglas, S. Bergamini, and F. Renzoni. Tunable Tsallis distributions in dissipative optical lattices. *Physical Review Letters*, 96:110601, 2006.

[17] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo. Generalized multiple importance sampling. *Statistical Science*, 34(1):129–155, 2019.

[18] M. Fajardo and J. Vidal. On subdifferentials via a generalized conjugation scheme: an application to DC problems and optimality conditions. *Set-Valued and Variational Analysis*, 30:1313–1331, 2022.

[19] A. Gelman, A. Jakulin, M. G. Pittau, and Y.-S. Su. A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, 2(4):1360–1383, 2008.

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

[21] M. Hasanasab, J. Hertrich, and G. Steidl. Alternatives to the EM algorithm for estimating the parameters of the Student t-distribution. *Numerical Algorithms*, 87:77–118, 2021.

[22] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(4):1303–1347, 2013.

[23] S. F. Jarner and G. O. Roberts. Convergence of heavy-tailed Monte Carlo Markov chain algorithms. *Scandinavian Journal of Statistics*, 34(4):781–815, 2007.

[24] A. S. Kainth, T.-K. L. Wong, and F. Rudzicz. Conformal mirror descent with logarithmic divergences. *Information Geometry*, 2022.

[25] M. Khan and W. Lin. Conjugate-computation variational inference: Converting variational inference in non-conjugate models to inferences in conjugate models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 878–887, 2017.

[26] M. Kot, M. A. Lewis, and P. van Den Driessche. Dispersal data and the spread of invading organisms. *Ecology*, 77(7):2027–2042, 1996.

[27] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[28] F. Laus, F. Pierre, and G. Steidl. Nonlocal myriad filters for Cauchy noise removal. *Journal of Mathematical Imaging and Vision*, 60:1324–1354, 2018.

[29] A. Le Franc, J.-P. Chancelier, and M. De Lara. The Capra-subdifferential of the ℓ_0 pseudonorm. *Optimization*, pages 1–23, 2022.

[30] Y. Marnissi, E. Chouzenoux, A. Benazza-Benyahia, and J.-C. Pesquet. Majorize–minimize adapted Metropolis–Hastings algorithm. *IEEE Transactions on Signal Processing*, 68:2356 – 2369, 2020.

[31] J. Martin, L. C. Wilcox, C. Burstedde, and O. Ghattas. A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. *SIAM Journal of Scientific Computing*, 34(3):A1460–A1487, 2012.

[32] A. F. T. Martins, M. Treviso, A. Farinhas, P. M. Q. Aguiar, M. A. T. Figueiredo, M. Blondel, and V. Niculae. Sparse continuous distributions and Fenchel Young losses. *Journal of Machine Learning Research*, 23(257):1–74, 2022.

[33] V. Masrani, R. Brekelmans, T. Bui, F. Nielsen, A. Galstyan, G. V. Steeg, and F. Wood. q-paths: Generalizing the geometric annealing path using power means. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 161, pages 1938–1947, 2021.

[34] J. J. Moré and G. Toraldo. Algorithms for bound constrained quadratic programming problems. *Numerische Mathematik*, 55(4):377–400, 1989.

[35] F. Nielsen and R. Nock. Entropies and cross-entropies of exponential families. In *IEEE International Conference on Image Processing (ICIP)*, pages 3621–3624, 2010.

[36] R. Nock, Z. Cranko, A. K. Menon, L. Qu, and R. C. Williamson. f-GANs in an information geometric nutshell. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

[37] S. Nowozin, B. Cseke, and R. Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016.

[38] D. Peel and G. J. McLachlan. Robust mixture modelling using the t-distribution. *Statistics and Computing*, 10:339–348, 2000.

[39] S. T. Rachev and L. Rüschendorf. *Mass Transportation Problems*. Springer-Verlag, 1998.

[40] G. O. Roberts and J. S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4):351–367, 2001.

[41] G. O. Roberts and O. Stramer. Langevin diffusions and Metropolis-Hastings algorithms. *Methodology and Computing in Applied Probability*, 4(4):337–357, 2002.

[42] A. Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 547–561, 1961.

[43] M. Teboulle. A simplified view of first order methods for optimization. *Mathematical Programming*, 170(1):67–96, 2018.

[44] Y. Tikochinsky, N. Z. Tishby, and R. D. Levine. Alternative approach to maximum-entropy inference. *Physical Review A*, 30:2638–2644, 1984.

[45] T. van Erven and P. Harremoës. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[46] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

[47] S. Wang and T. Swartz. Moment matching adaptive importance sampling with skew-Student proposals. *Monte Carlo Methods and Applications*, 28(2):149–162, 2022.

[48] H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982.

[49] T.-K. L. Wong. Logarithmic divergences from optimal transport and Rényi geometry. *Information Geometry*, 1(1):39–78, 2018.

[50] T.-K. L. Wong and J. Zhang. Tsallis and Rényi deformations linked via a new λ-duality. *IEEE Transactions on Information Theory*, 68(8):5353–5373, 2022.