J. Appl. Numer. Optim. 7 (2025), No. 3, pp. 399-420
Available online at http://jano.biemdas.com
https://doi.org/10.23952/jano.7.2025.3.07

VARIABLE BREGMAN MAJORIZATION-MINIMIZATION ALGORITHM AND ITS APPLICATION TO DIRICHLET MAXIMUM LIKELIHOOD ESTIMATION

SÉGOLÈNE MARTIN[1,*], JEAN-CHRISTOPHE PESQUET[2], GABRIELE STEIDL[1], ISMAIL BEN AYED[3]

[1]*TU Berlin, Germany*
[2]*Université Paris-Saclay, Inria, CentraleSupélec, CVN, France*
[3]*ÉTS Montréal, Canada*

Abstract. We propose a novel Bregman descent algorithm for minimizing a convex function that is expressed as the sum of a differentiable part (defined over an open set) and a possibly nonsmooth term. The approach, referred to as the Variable Bregman Majorization-Minimization (VBMM) algorithm, extends the Bregman Proximal Gradient method by allowing the Bregman function used in the divergence to adaptively vary at each iteration, provided it satisfies a majorizing condition on the objective function. This adaptive framework enables the algorithm to approximate the objective more precisely at each iteration, thereby allowing for accelerated convergence compared to the traditional Bregman Proximal Gradient descent. We establish the convergence of the VBMM algorithm to a minimizer under mild assumptions on the family of metrics used. Furthermore, we introduce a novel application of both the Bregman Proximal Gradient method and the VBMM algorithm to the estimation of the multidimensional parameters of a Dirichlet distribution through the maximization of its log-likelihood. Numerical experiments confirm that the VBMM algorithm outperforms existing approaches in terms of convergence speed.

Keywords. Bregman majorization-minimization algorithm; Bregman proximal gradient method; Dirichlet maximum likelihood estimation.

2020 Mathematics Subject Classification. 90C26, 49J52.

1. INTRODUCTION

The objective of this article is to solve the classical convex minimization problem:

$$\underset{\boldsymbol{x}\in\mathbb{R}^d}{\text{minimize}}\; F(\boldsymbol{x}), \tag{1.1}$$

where $F = f + g$. Here, the function $f : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ is convex and differentiable on an open set $\mathscr{D} \subseteq \mathbb{R}^d$, while $g : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ is proper, convex, and lower semi-continuous, with $\operatorname{dom} g = \{\boldsymbol{x} \in \mathbb{R}^d \mid g(\boldsymbol{x}) < +\infty\} \subseteq \mathscr{D}$.

When $\mathscr{D} = \mathbb{R}^d$ and the gradient of f is L-Lipschitz continuous, problem (1.1) can be efficiently solved by the *Proximal Gradient* (PG) method [1, 2], also known as Forward Backward

*Corresponding author.
E-mail address: segolene.tiffany.martin@gmail.com (S. Martin).
Received 14 January 2025; Accepted 29 August 2025; Published online 24 November 2025.

Splitting. Given an initial point $\boldsymbol{x}^{(0)} \in \mathbb{R}^d$, the PG algorithm generates a sequence $(\boldsymbol{x}^{(\ell)})_{\ell\in\mathbb{N}}$ by solving

$$(\forall \ell \in \mathbb{N}) \quad \boldsymbol{x}^{(\ell+1)} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^d} \left\{ g(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}^{(\ell)}), \boldsymbol{x} - \boldsymbol{x}^{(\ell)} \rangle + \frac{1}{2\gamma_\ell} \|\boldsymbol{x} - \boldsymbol{x}^{(\ell)}\|^2 \right\}, \tag{1.2}$$

where $\gamma_\ell < 2/L$.

However, in many practical applications the gradient of f is not Lipschitz continuous, or the domain $\mathscr{D}$ is a strict subset of $\mathbb{R}^d$ (see also recent conditions in [3]). This is for instance the case of minimization problems involving a Kullback-Leibler divergence such as Poisson restoration [4] or estimation of the parameters of a distribution through a Maximum Likelihood approach in statistics [5]. For such cases, PG is not directly applicable.

Bregman proximal methods address this limitation. First introduced in [6] and followed by [7, 8, 9], this family of approaches replace the Euclidean metric in proximal schemes with a Bregman distance D_h with respect to a convex and continuously differentiable function $h : \mathbb{R}^d \to (-\infty, +\infty]$ on its domain. We refer to [10] for an overview of Bregman descent methods. In particular, in the *Bregman Proximal Gradient* (BPG) algorithm [11, 12] (a.k.a. Mirror Descent algorithm when the non-differentiable term g is set to zero [13, 14]), the ℓ-th iterates reads

$$\boldsymbol{x}^{(\ell+1)} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^d} \left\{ g(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}^{(\ell)}), \boldsymbol{x} - \boldsymbol{x}^{(\ell)} \rangle + \frac{1}{L} D_h(\boldsymbol{x}, \boldsymbol{x}^{(\ell)}) \right\}, \tag{1.3}$$

for a fixed constant $L > 0$. This formulation generalizes the PG method in (1.2), which can be recovered by setting $h = \|\cdot\|^2/2$.

The BPG method has proven valuable in a range of applications. In imaging, BPG has shown effectiveness in addressing Poisson noise in deblurring problems [15, 16, 17] and has recently been adapted within a Plug-and-Play framework for enhanced performance [18]. In parametric statistics, BPG is useful for estimating parameters of distributions, in particular distributions within the exponential family, for which it is closely linked to the Expectation-Maximization algorithm [19, 20, 21]. Moreover, constrained sampling through mirror-Langevin dynamics has also incorporated BPG techniques, illustrating its versatility [22].

The convergence of BPG has been studied in [11, 23] under the Lipschitz-like condition:

$$Lh - f \text{ is convex on } \operatorname{int}(\operatorname{dom} h),$$

where $\operatorname{int}(S)$ denotes the interior of a subset S of $\mathbb{R}^d$. This condition was shown in [11, Lemma 1] to be equivalent to the following *majorization* property, for every $\boldsymbol{y} \in \operatorname{int}(\operatorname{dom} h)$,

$$(\forall \boldsymbol{x} \in \operatorname{int}(\operatorname{dom} h)) \quad f(\boldsymbol{x}) \leq f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle + L D_h(\boldsymbol{x}, \boldsymbol{y}).$$

This condition is also sometimes called an L-relative smoothness property of f with respect to function h. Under this property, BPG can be interpreted as a *Majorization-Minimization* (MM) method with the following update rule:

$$\boldsymbol{x}^{(\ell+1)} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^d} q_{\boldsymbol{x}^{(\ell)}}(\boldsymbol{x}),$$

where the surrogate function q is defined by

$$(\forall \boldsymbol{x} \in \operatorname{int}(\operatorname{dom} h)) \quad q_{\boldsymbol{y}}(\boldsymbol{x}) = g(\boldsymbol{x}) + f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle + L D_h(\boldsymbol{x}, \boldsymbol{y}).$$

In this paper, we introduce an extended version of the BPG framework that allows the Bregman metric D_h in (1.3) to vary at each iteration, adapting to the current iterate. We refer to this adaptive approach as the Variable Bregman Majorization-Minimization (VBMM) algorithm. By adapting the Bregman divergence to the local geometry at each iteration, we aim to achieve more accurate approximation to the objective F and to accelerate convergence.

The Bregman proximal gradient approach with variable Bregman metric has been relatively unexplored. It has been studied in the context of monotone operator theory in [12] and [24], where decaying conditions on the Bregman divergences are imposed. Our approach differs by not requiring such conditions. Additionally, [25] analyzed the convergence of a generalized majorization-minimization framework, of which VBMM is a special case. Their convergence theory, however, depends on a "growth condition" formulated as

$$|F(\boldsymbol{x}) - q_{\boldsymbol{y}}(\boldsymbol{x})| \leq \omega(\|\boldsymbol{x}-\boldsymbol{y}\|),$$

where $\omega : [0,+\infty) \longrightarrow [0,+\infty)$ is a differentiable function such that $\omega(0) = 0$, $w' > 0$, and $\lim_{t\to+\infty} \omega'(t) = \lim_{t\to+\infty} \omega(t)/\omega'(t) = 0$. In practice, however, it may be challenging to identify a suitable function ω that fulfills these conditions, limiting the practical use of this approach for some applications. Finally, the recent work [26] proposed a variable Bregman proximal gradient approach for a Poisson deconvolution problem and showed that using a variable metric can accelerate convergence. However, this work did not provide any convergence analysis.

The paper is organized as follows. In Section 2, we introduce the Variable Bregman Majorization-Minimization algorithm. In Section 3, we provide a convergence analysis that relies on mild assumptions on the Bregman metric, enhancing the practical utility of the method. Finally, Section 4 is dedicated to an original application of the VBMM algorithm to the Maximum Likelihood estimation of the parameters of a Dirichlet distribution. In particular, we prove that minimization problem admits a solution, derive closed form iterates for the VBMM algorithm, and provide numerical experiments on synthetic data to illustrate the good performance of the algorithm.

Throughout the article, $\partial\phi$ denotes the Moreau subidfferential of a convex function $\phi : \mathbb{R}^d \longrightarrow \mathbb{R}\cup\{+\infty\}$.

2. Problem and Algorithm

Let $F : \mathbb{R}^d \longrightarrow \mathbb{R}\cup\{+\infty\}$ be defined as

$$F = f + g,$$

where $f : \mathbb{R}^d \longrightarrow \mathbb{R}\cup\{+\infty\}$ and $g : \mathbb{R}^d \longrightarrow \mathbb{R}\cup\{+\infty\}$ are functions satisfying the following assumption:

Assumption 2.1.

(i) f is convex and differentiable on an open set $\mathscr{D} \subseteq \mathbb{R}^d$,
(ii) g is proper, convex, and lower semi-continuous with $\operatorname{dom} g \subseteq \mathscr{D}$,
(iii) F is lower bounded.

Our objective is to solve the classical convex minimization problem

$$\underset{\boldsymbol{x}\in\mathbb{R}^d}{\text{minimize}}\ F(\boldsymbol{x}).$$

In particular, if g is the indicator function of a nonempty closed convex set $\mathscr{C} \subset \mathscr{D}$, the following constrained optimization problem is obtained:

$$\underset{\boldsymbol{x}\in\mathscr{C}}{\text{minimize}}\ f(\boldsymbol{x}).$$

A central mathematical tool in our methodology is the Bregman divergence, originally introduced in [27].

Definition 2.1 (Bregman Divergence)**.** Consider a function $h : \mathbb{R}^d \to (-\infty, +\infty]$ that is differentiable on $\operatorname{int}(\operatorname{dom} h)$. The Bregman divergence associated with h is defined, for every pair $(\boldsymbol{x}, \boldsymbol{y})$ in $\operatorname{dom} h \times \operatorname{int}(\operatorname{dom} h)$, as

$$D_h(\boldsymbol{x}, \boldsymbol{y}) := h(\boldsymbol{x}) - h(\boldsymbol{y}) - \langle \nabla h(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle.$$

Essentially, $D_h(\boldsymbol{x}, \boldsymbol{y})$ quantifies the difference between the value of the function h at a point $\boldsymbol{x} \in \operatorname{dom} h$ and its linear approximation around $\boldsymbol{x}$, calculated at $\boldsymbol{y} \in \operatorname{int}(\operatorname{dom} h)$. If h is convex on $\operatorname{dom} h$, then $D_h(\boldsymbol{x}, \boldsymbol{y}) \geq 0$ and $D_h(\cdot, \boldsymbol{y})$ is convex. If h is strictly convex on $\operatorname{int}(\operatorname{dom} h)$ and $\boldsymbol{x} \in \operatorname{int}(\operatorname{dom} h)$, then $D_h(\boldsymbol{x}, \boldsymbol{y}) = 0$ if and only if $\boldsymbol{x} = \boldsymbol{y}$. Furthermore, $D_h(\cdot, \boldsymbol{y})$ is strictly convex. Note that

$$\nabla_{\boldsymbol{x}} D_h(\boldsymbol{x}, \boldsymbol{y}) = \nabla h(\boldsymbol{x}) - \nabla h(\boldsymbol{y}). \tag{2.1}$$

Definition 2.2 (Bregman Majorant Function)**.** For a given $\boldsymbol{y} \in \mathscr{D}$, let $h_{\boldsymbol{y}} \colon \mathbb{R}^d \longrightarrow \mathbb{R} \cup \{+\infty\}$ be a lower semi-continuous, strictly convex function, which is differentiable on $\operatorname{int}(\operatorname{dom} h_{\boldsymbol{y}}) \supseteq \mathscr{D}$. Let $q_{\boldsymbol{y}} \colon \mathscr{D} \longrightarrow \mathbb{R}$ be defined as

$$q_{\boldsymbol{y}}(\boldsymbol{x}) := f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle + D_{h_{\boldsymbol{y}}}(\boldsymbol{x}, \boldsymbol{y}).$$

We call $q_{\boldsymbol{y}}$ a Bregman tangent majorant of f at $\boldsymbol{y}$ associated with $h_{\boldsymbol{y}}$ if, for every $x \in \mathscr{D}$,

$$f(\boldsymbol{x}) \leq q_{\boldsymbol{y}}(\boldsymbol{x}) \tag{2.2}$$

or, in other words,

$$D_f(\boldsymbol{x}, \boldsymbol{y}) \leq D_{h_{\boldsymbol{y}}}(\boldsymbol{x}, \boldsymbol{y}).$$

The definition expands upon concepts previously introduced as those in [11, 23]. A distinctive aspect of our approach is the flexibility, in the *Majorization-Minimization* (MM) strategy, to allow the Bregman metric to vary according to the reference point $\boldsymbol{y}$.

Let $(h_{\boldsymbol{y}})_{\boldsymbol{y}\in\mathscr{D}}$ be a family of function such that, for every $\boldsymbol{y} \in \mathscr{D}$, $h_{\boldsymbol{y}} \colon \mathbb{R}^d \longrightarrow \mathbb{R} \cup \{+\infty\}$. We make the following assumption.

Assumption 2.2. For f and g as in Assumption 2.1, the family of functions $(h_{\boldsymbol{y}})_{\boldsymbol{y}\in\mathscr{D}}$ fulfills the conditions: for every $\boldsymbol{y} \in \mathscr{D}$,

(i) $h_{\boldsymbol{y}}$ is a lower semi-continuous function which is strictly convex and differentiable on $\operatorname{int}(\operatorname{dom} h_{\boldsymbol{y}})$, and $\mathscr{D} \subseteq \operatorname{int}(\operatorname{dom} h_{\boldsymbol{y}})$.

(ii) The function

$$\boldsymbol{x} \mapsto g(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle + D_{h_{\boldsymbol{y}}}(\boldsymbol{x}, \boldsymbol{y}) \tag{2.3}$$

is coercive.

(iii) The majorizing property (2.2) is satisfied.

Remark 2.1. A sufficient condition for Assumption 2.2(i)-(ii) to be satisfied is that $h_{\boldsymbol{y}}$, $\boldsymbol{y} \in \mathscr{D}$, are lower-semicontinuous Legendre functions with $\mathscr{D} \subseteq \operatorname{int}(\operatorname{dom} h_{\boldsymbol{y}})$ and the functions

$$\varphi_{\boldsymbol{y}} : \boldsymbol{x} \mapsto g(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle, \quad \boldsymbol{y} \in \mathscr{D}$$

are lower bounded. This can be seen as follows: recall that $h_{\boldsymbol{y}}$ is a Legendre function if it is essentially strictly convex and essentially smooth. The first property implies that $h_{\boldsymbol{y}}$ is strictly convex on $\operatorname{int}(\operatorname{dom} h_{\boldsymbol{y}}) \subseteq \operatorname{dom} \partial h_{\boldsymbol{y}}$. As shown in [28, Thm. 3.7], $D_{h_{\boldsymbol{y}}}(\cdot, \boldsymbol{y})$ is coercive because $h_{\boldsymbol{y}}$ is essentially strictly convex. Since $\varphi_{\boldsymbol{y}}$ is lower-bounded, this implies that (2.3) is coercive.

Example 2.1. A useful example of metrics $(h_{\boldsymbol{y}})_{\boldsymbol{y} \in \mathscr{D}}$ leading to separable Bregman majorants, and thus easy to minimize, can be defined as follows. For every $\boldsymbol{y} \in \mathscr{D}$, let

$$(\forall \boldsymbol{x} = (x_i)_{1 \leq i \leq d} \in \mathbb{R}^d) \quad h_{\boldsymbol{y}}(\boldsymbol{x}) \colon = \sum_{i=1}^{d} a_i(\boldsymbol{y}) \nu_i(x_i) + \sum_{i=1}^{d} b_i(\boldsymbol{y}) \frac{x_i^2}{2}, \tag{2.4}$$

where, for every $i \in \{1, \ldots, d\}$, the functions $a_i, b_i \colon \mathbb{R}^d \longrightarrow [0, +\infty)$ are continuous, and $\nu_i \colon \mathbb{R} \longrightarrow \mathbb{R} \cup \{+\infty\}$ is convex and differentiable on the interior of its domain with $\mathscr{D} \subseteq \operatorname{int}\left(\bigtimes_{j=1}^{d} \operatorname{dom} \nu_j\right)$. Note that the convexity of the differentiable function ν_i implies that $\nabla \nu_i$ is continuous, see [29, Corollary 17.42]. Functions of the form (2.4) are in particular useful for representing quadratic majorants with a diagonal curvature matrix when setting $a_i = 0$ for all $i \in \{1, \ldots, d\}$. This form is also well-suited for the maximum likelihood Dirichlet problem that will be discussed in the application section.

The Variable Bregman Majorization-Minimization (VBMM) algorithm aims at minimizing the sum of the non-smooth function g and a Bregman majorant of f calculated from the previous iterate $\boldsymbol{x}^{(\ell)}$, i.e., finding for $\boldsymbol{y} = \boldsymbol{x}^{(\ell)}$ the minimizer of

$$\begin{aligned} \Phi_{\boldsymbol{y}}(\boldsymbol{x}) &\coloneqq g(\boldsymbol{x}) + q_{\boldsymbol{y}}(\boldsymbol{x}) \\ &= g(\boldsymbol{x}) + f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle + h_{\boldsymbol{y}}(\boldsymbol{x}) - h_{\boldsymbol{y}}(\boldsymbol{y}) - \langle \nabla h_{\boldsymbol{y}}(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \\ &= g(\boldsymbol{x}) + h_{\boldsymbol{y}}(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{y}) - \nabla h_{\boldsymbol{y}}(\boldsymbol{y}), \boldsymbol{x} \rangle - \langle \nabla f(\boldsymbol{y}) - \nabla h_{\boldsymbol{y}}(\boldsymbol{y}), \boldsymbol{y} \rangle + f(\boldsymbol{y}). \end{aligned}$$

Neglecting the constants yields Algorithm 1.

Algorithm 1: VBMM Algorithm

Input: Initial point $\boldsymbol{x}^{(0)} \in \mathscr{D}$
for $\ell = 0, 1, \ldots$ **do**
  $\boldsymbol{x}^{(\ell+1)} = \arg\min_{\boldsymbol{x} \in \mathbb{R}^d} \; g(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}^{(\ell)}), \boldsymbol{x} \rangle + D_{h_{\boldsymbol{x}^{(\ell)}}}(\boldsymbol{x}, \boldsymbol{x}^{(\ell)})$

Proposition 2.1. *Under Assumptions 2.1 and 2.2, Algorithm 1 is well-defined and generates a sequence* $(\boldsymbol{x}^{(\ell)})_{\ell \geq 1}$ *in* $\operatorname{dom} g$.

Proof. For $\boldsymbol{x}^{(\ell)} \in \mathscr{D}$, the function

$$\boldsymbol{x} \mapsto \langle \nabla f(\boldsymbol{x}^{(\ell)}), \boldsymbol{x} - \boldsymbol{x}^{(\ell)} \rangle + D_{h_{\boldsymbol{x}^{(\ell)}}}(\boldsymbol{x}, \boldsymbol{x}^{(\ell)}) + g(\boldsymbol{x})$$

is well-defined since $\mathscr{D} \subseteq \operatorname{int}(\operatorname{dom} h_{\boldsymbol{x}^{(\ell)}})$. This function is proper since $\varnothing \neq \operatorname{dom} g \subseteq \mathscr{D}$. In addition, the lower-semicontinuity and coercivity properties in Assumption 2.2 implies the existence of a minimizer $\boldsymbol{x}^{(\ell+1)}$ of this function, which necessarily belongs to $\operatorname{dom} g$. This minimizer is

unique since $h_{\boldsymbol{x}^{(\ell)}}$ and thus $D_{h_{\boldsymbol{x}^{(\ell)}}}(\cdot,\boldsymbol{x}^{(\ell)})$ are strictly convex on $\mathscr{D}$. The result is then shown by induction. □

Remark 2.2. This strategy of adapting the metric to the current iterate in non-Bregman MM schemes has already been widely studied both theoretically and numerically for the case of quadratic majorants [30, 31, 32]. In particular, our Variable Bregman Majorization-Minimization algorithm can be seen as a generalization of the Variable Metric Forward-Backward (VMFB) algorithm [33, 34, 35] which considers quadratic majorants of the form

$$(\forall \boldsymbol{x} \in \mathbb{R}^d) \quad q(\boldsymbol{x};\boldsymbol{y}) = f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{x}-\boldsymbol{y}\rangle + \frac{1}{2}\langle \boldsymbol{x}-\boldsymbol{y}, \boldsymbol{A}_{\boldsymbol{y}}(\boldsymbol{x}-\boldsymbol{y})\rangle,$$

with $\boldsymbol{A}_{\boldsymbol{y}}$ a positive definite matrix that depends on the point $\boldsymbol{y}$.

Remark 2.3. For every $\ell \in \mathbb{N}$, $\boldsymbol{x}^{(\ell+1)}$ is actually the $D_{h_{\boldsymbol{x}^{(\ell)}}}$-proximity operator of function g at $\nabla h_{\boldsymbol{x}^{(\ell)}}(\boldsymbol{x}^{(\ell)}) - \nabla f(\boldsymbol{x}^{(\ell)})$. Algorithm 1 is thus a variable metric forward-backward splitting method based on a Bregman divergence [12, 24]. The main difference with the analysis conducted in [24], in the general context of monotone operator theory, is that we do not assume that the Bregman divergences $(D_{h_{\boldsymbol{x}^{(\ell)}}})_{\ell\in\mathbb{N}}$ satisfy a decaying condition like in [24, Algorithme 2.4b] (see also [12, Definition 2.2]), but we rely on a Majorization-Minimization property. When $g = 0$, the algorithm simplifies to

$$(\forall \ell \in \mathbb{N}) \quad \begin{cases} \widetilde{\boldsymbol{z}}^{(\ell)} = \nabla h_{\boldsymbol{x}^{(\ell)}}(\boldsymbol{x}^{(\ell)}) \\ \boldsymbol{z}^{(\ell+1)} = \widetilde{\boldsymbol{z}}^{(\ell)} - \nabla f(\boldsymbol{x}^{(\ell)}) \\ \boldsymbol{x}^{(\ell+1)} = (\nabla h_{\boldsymbol{x}^{(\ell)}})^{-1}(\boldsymbol{z}^{(\ell+1)}). \end{cases}$$

This shows that the mirror map is modified along the gradient steps based on the current iterate.

3. Convergence Analysis

Our convergence result is stated in Proposition 3.1. It shows that, if the sequence $(\boldsymbol{x}^{(\ell)})_{\ell\in\mathbb{N}}$ generated by Algorithm 1 lies in a compact subset of $\mathscr{D}$, then subsequential convergence towards a minimizer of F is achieved under suitable assumptions on the Bregman functions. The existence of such compact set can be addressed independently. For instance, a sufficient condition is that the objective function is coercive on $\mathscr{D}$. This can be ensured by a proper choice of function g.

We start proving that, without further assumptions, Algorithm 1 produces a non increasing sequence of cost values. First, we recall the *Three Points Identity*, which can be found, e.g., in [8, Lemma 3.1].

Lemma 3.1 (Three Points Identity)**.** *Let $h\colon \mathbb{R}^d \longrightarrow \mathbb{R}\cup\{+\infty\}$ be strictly convex and differentiable on* $\operatorname{int}(\operatorname{dom} h)$*. Then, for any three points $\boldsymbol{x},\boldsymbol{y}$ in* $\operatorname{int}(\operatorname{dom} h)$ *and $\boldsymbol{z} \in \operatorname{dom} h$, it holds*

$$D_h(\boldsymbol{z},\boldsymbol{x}) - D_h(\boldsymbol{z},\boldsymbol{y}) - D_h(\boldsymbol{y},\boldsymbol{x}) = \langle \nabla h(\boldsymbol{x}) - \nabla h(\boldsymbol{y}), \boldsymbol{y}-\boldsymbol{z}\rangle.$$

Lemma 3.2. *Suppose that Assumptions 2.1 and 2.2 are satisfied. Let $(\boldsymbol{x}^{(\ell)})_{\ell\in\mathbb{N}}$ be the sequence generated by VBMM Algorithm 1. Then, for every $\ell \in \mathbb{N}$ and $\boldsymbol{v} \in \mathscr{D}$, we have*

$$F(\boldsymbol{x}^{(\ell+1)}) - F(\boldsymbol{v}) \leq D_{h_{\boldsymbol{x}^{(\ell)}}}(\boldsymbol{v},\boldsymbol{x}^{(\ell)}) - D_{h_{\boldsymbol{x}^{(\ell)}}}(\boldsymbol{v},\boldsymbol{x}^{(\ell+1)}) - D_f(\boldsymbol{v},\boldsymbol{x}^{(\ell)}).$$

In addition, sequence $(F(\boldsymbol{x}^{(\ell)}))_{\ell\in\mathbb{N}}$ is non-increasing.

Proof. Let $\boldsymbol{v} \in \mathscr{D}$ and $\ell \in \mathbb{N}$. By definition of $\boldsymbol{x}^{(\ell+1)}$ and (2.1), there exists $\boldsymbol{w}^{(\ell+1)} \in \partial g(\boldsymbol{x}^{(\ell+1)})$ such that

$$\nabla f(\boldsymbol{x}^{(\ell)}) = \nabla h_{\boldsymbol{x}^{(\ell)}}(\boldsymbol{x}^{(\ell)}) - \nabla h_{\boldsymbol{x}^{(\ell)}}(\boldsymbol{x}^{(\ell+1)}) - \boldsymbol{w}^{(\ell+1)}.$$

Using the fact that $\boldsymbol{w}^{(\ell+1)}$ is a subgradient of g at $\boldsymbol{x}^{(\ell+1)}$, we obtain

$$\begin{aligned}\langle \nabla f(\boldsymbol{x}^{(\ell)}), \boldsymbol{x}^{(\ell+1)} - \boldsymbol{v}\rangle &= \langle \nabla h_{\boldsymbol{x}^{(\ell)}}(\boldsymbol{x}^{(\ell)}) - \nabla h_{\boldsymbol{x}^{(\ell)}}(\boldsymbol{x}^{(\ell+1)}), \boldsymbol{x}^{(\ell+1)} - \boldsymbol{v}\rangle - \langle \boldsymbol{w}^{(\ell+1)}, \boldsymbol{x}^{(\ell+1)} - \boldsymbol{v}\rangle \\ &\leq \langle \nabla h_{\boldsymbol{x}^{(\ell)}}(\boldsymbol{x}^{(\ell)}) - \nabla h_{\boldsymbol{x}^{(\ell)}}(\boldsymbol{x}^{(\ell+1)}), \boldsymbol{x}^{(\ell+1)} - \boldsymbol{v}\rangle + g(\boldsymbol{v}) - g(\boldsymbol{x}^{(\ell+1)}).\end{aligned}$$

Furthermore, it follows from Lemma 3.1 with $\boldsymbol{x} = \boldsymbol{x}^{(\ell)}$, $\boldsymbol{y} = \boldsymbol{x}^{(\ell+1)}$, and $\boldsymbol{z} = \boldsymbol{v}$ that

$$\begin{aligned}\langle \nabla f(\boldsymbol{x}^{(\ell)}), \boldsymbol{x}^{(\ell+1)} - \boldsymbol{v}\rangle \leq{}& D_{h_{\boldsymbol{x}^{(\ell)}}}(\boldsymbol{v}, \boldsymbol{x}^{(\ell)}) - D_{h_{\boldsymbol{x}^{(\ell)}}}(\boldsymbol{v}, \boldsymbol{x}^{(\ell+1)}) - D_{h_{\boldsymbol{x}^{(\ell)}}}(\boldsymbol{x}^{(\ell+1)}, \boldsymbol{x}^{(\ell)}) \\ &+ g(\boldsymbol{v}) - g(\boldsymbol{x}^{(\ell+1)}).\end{aligned} \tag{3.1}$$

On the other hand, the majorization property at iteration ℓ reads as

$$f(\boldsymbol{x}^{(\ell+1)}) \leq f(\boldsymbol{x}^{(\ell)}) + \langle \nabla f(\boldsymbol{x}^{(\ell)}), \boldsymbol{x}^{(\ell+1)} - \boldsymbol{x}^{(\ell)}\rangle + D_{h_{\boldsymbol{x}^{(\ell)}}}(\boldsymbol{x}^{(\ell+1)}, \boldsymbol{x}^{(\ell)}),$$

and subtracting $f(\boldsymbol{v}) + \langle \nabla f(\boldsymbol{x}^{(\ell)}), \boldsymbol{x}^{(\ell+1)} - \boldsymbol{v}\rangle$ on both sides, yields

$$\begin{aligned}f(\boldsymbol{x}^{(\ell+1)}) - f(\boldsymbol{v}) - \langle \nabla f(\boldsymbol{x}^{(\ell)}), \boldsymbol{x}^{(\ell+1)} - \boldsymbol{v}\rangle &\leq f(\boldsymbol{x}^{(\ell)}) - f(\boldsymbol{v}) + \langle \nabla f(\boldsymbol{x}^{(\ell)}), \boldsymbol{v} - \boldsymbol{x}^{(\ell)}\rangle \\ &\quad + D_{h_{\boldsymbol{x}^{(\ell)}}}(\boldsymbol{x}^{(\ell+1)}, \boldsymbol{x}^{(\ell)}) \\ &= -D_f(\boldsymbol{v}, \boldsymbol{x}^{(\ell)}) + D_{h_{\boldsymbol{x}^{(\ell)}}}(\boldsymbol{x}^{(\ell+1)}, \boldsymbol{x}^{(\ell)}).\end{aligned} \tag{3.2}$$

Finally, summing up (3.1) and (3.2),

$$f(\boldsymbol{x}^{(\ell+1)}) - f(\boldsymbol{v}) + g(\boldsymbol{x}^{(\ell+1)}) - g(\boldsymbol{v}) \leq D_{h_{\boldsymbol{x}^{(\ell)}}}(\boldsymbol{v}, \boldsymbol{x}^{(\ell)}) - D_{h_{\boldsymbol{x}^{(\ell)}}}(\boldsymbol{v}, \boldsymbol{x}^{(\ell+1)}) - D_f(\boldsymbol{v}, \boldsymbol{x}^{(\ell)}).$$

By setting $\boldsymbol{v} = \boldsymbol{x}^{(\ell)}$, we deduce that $(F(\boldsymbol{x}^{(\ell)}))_{\ell \in \mathbb{N}}$ is non-increasing. □

We shall now prove that, under the condition that the Bregman divergences are uniformly lower-bounded by an increasing function of the norm, the difference between two consecutive iterates tends to zero.

Assumption 3.1. For any nonempty bounded set $\mathscr{C} \subseteq \mathscr{D}$, there exists an increasing function $\rho : \mathbb{R}_+ \to \mathbb{R}_+$ vanishing only at 0, such that

$$(\forall (\boldsymbol{x}, \boldsymbol{y}) \in \mathscr{C}^2) \quad D_{h_{\boldsymbol{x}}}(\boldsymbol{x}, \boldsymbol{y}) \geq \rho(\|\boldsymbol{x} - \boldsymbol{y}\|). \tag{3.3}$$

Note that similar conditions are classical in the case when $h_{\boldsymbol{x}}$ is independent of $\boldsymbol{x}$ [36, Property (*) in Sec. 4.2]. Assumption 3.1 is satisfied for Bregman functions of the form given in (2.4), particularly when, for all $\boldsymbol{y} \in \mathscr{C}$, there exists a constant $c \geq 0$ such that, for every $i \in \{1, \ldots, d\}$ $b_i(\boldsymbol{y}) \geq c$. In this scenario, (3.3) is satisfied with $\rho = c\,(\cdot)^2/2$.

Remark 3.1. Assumption 3.1 can be seen as a generalization to the Bregman framework of the assumption made on the metrics in the VMFB algorithm [35]. Indeed, VMFB was shown to converge to a minimizer under the assumption that there exits $\nu \in (0, +\infty)$ such that for all $\boldsymbol{y} \in \mathbb{R}^d$, $\nu \boldsymbol{I}_d \preceq \boldsymbol{A}_{\boldsymbol{y}}$, where $\boldsymbol{A}_{\boldsymbol{y}}$ is the curvature matrix associated with the quadratic majorant at $\boldsymbol{y}$.

Lemma 3.3. *Let $(\boldsymbol{x}^{(\ell)})_{\ell \in \mathbb{N}}$ be the sequence generated by the VBMM algorithm 1. Assume that $(\boldsymbol{x}^{(\ell)})_{\ell \in \mathbb{N}}$ is bounded. Suppose that Assumptions 2.1, 2.2, and 3.1 are satisfied. Then $\boldsymbol{x}^{(\ell+1)} - \boldsymbol{x}^{(\ell)} \to 0$ as $\ell \to +\infty$.*

Proof. Applying Lemma 3.2 with $\boldsymbol{v} = \boldsymbol{x}^{(\ell)}$ yields

$$F(\boldsymbol{x}^{(\ell+1)}) - F(\boldsymbol{x}^{(\ell)}) \leq -D_{h_{\boldsymbol{x}^{(\ell)}}}(\boldsymbol{x}^{(\ell)}, \boldsymbol{x}^{(\ell+1)}).$$

According to Assumption 3.1, there exists a function $\rho : \mathbb{R}_+ \to \mathbb{R}_+$, increasing and vanishing only at 0, such that, for every $\ell \in \mathbb{N}$,

$$D_{h_{\boldsymbol{x}^{(\ell)}}}(\boldsymbol{x}^{(\ell)}, \boldsymbol{x}^{(\ell+1)}) \geq \rho(\|\boldsymbol{x}^{(\ell)} - \boldsymbol{x}^{(\ell+1)}\|).$$

Hence, summing from $\ell = 0$ to $\ell = L$,

$$\sum_{\ell=0}^{L} \rho(\|\boldsymbol{x}^{(\ell)} - \boldsymbol{x}^{(\ell+1)}\|) \leq F(\boldsymbol{x}^{(0)}) - F(\boldsymbol{x}^{(L+1)}) \leq F(\boldsymbol{x}^{(0)}) - \underline{F},$$

where $\underline{F}$ is a lower bound on F. This shows that $\sum_{\ell=0}^{+\infty} \rho(\|\boldsymbol{x}^{(\ell)} - \boldsymbol{x}^{(\ell+1)}\|) < +\infty$, which implies that $\rho(\|\boldsymbol{x}^{(\ell)} - \boldsymbol{x}^{(\ell+1)}\|) \to 0$, that is $\boldsymbol{x}^{(\ell)} - \boldsymbol{x}^{(\ell+1)} \to 0$. □

Assumption 3.2. For any bounded sequences $(\boldsymbol{x}^{(\ell)})_{\ell\in\mathbb{N}}$, $(\boldsymbol{y}^{(\ell)})_{\ell\in\mathbb{N}}$ in $\mathscr{D}$ such that $\boldsymbol{x}^{(\ell)} - \boldsymbol{y}^{(\ell)} \to 0$, it holds

$$\nabla h_{\boldsymbol{y}^{(\ell)}}(\boldsymbol{y}^{(\ell)}) - \nabla h_{\boldsymbol{y}^{(\ell)}}(\boldsymbol{x}^{(\ell)}) \to 0.$$

It is easy to check that Assumption 3.2 is satisfied by Bregman functions of the form (2.4).

Proposition 3.1. *Let $(\boldsymbol{x}^{(\ell)})_{\ell\in\mathbb{N}}$ be the sequence generated by the VBMM Algorithm 1. Assume that there exists a compact subset $\mathscr{C} \subset \mathscr{D}$, such that $\{\boldsymbol{x}^{(\ell)}\}_{\ell\in\mathbb{N}} \subseteq \mathscr{C}$. Further, suppose that Assumptions 2.1-3.2 are satisfied. Then the following holds true:*

(i) the set of cluster points of $(\boldsymbol{x}^{(\ell)})_{\ell\in\mathbb{N}}$ is a nonempty compact and connected subset of Argmin F*;*
(ii) $(F(\boldsymbol{x}^{(\ell)}))_{\ell\in\mathbb{N}}$ converges to the minimum of F;
(iii) if F admits a unique minimizer, then $(\boldsymbol{x}^{(\ell)})_{\ell\in\mathbb{N}}$ converges to this minimizer.

Proof. (i) We show that the cluster points are minimizers of F. Then the fact that the set of cluster points of $(\boldsymbol{x}^{(\ell)})_{\ell\in\mathbb{N}}$ is a nonempty compact and connected subset of Argmin F follows from the boundness of $(\boldsymbol{x}^{(\ell)})_{\ell\in\mathbb{N}}$ and Lemma 3.3, by invoking Ostrowski's theorem [29, Lemma 2.53].

Let $\bar{\boldsymbol{x}} \in \mathscr{C} \subset \mathscr{D}$ be a cluster point of $(\boldsymbol{x}^{(\ell)})_{\ell\in\mathbb{N}}$. There exists a subsequence $(\boldsymbol{x}^{(\ell_j)})_{j\in\mathbb{N}}$ converging to $\bar{\boldsymbol{x}}$. Let us show that $\bar{\boldsymbol{x}}$ is a critical point of F, i.e., $0 \in \partial F(\bar{\boldsymbol{x}})$, or equivalently, $-\nabla f(\bar{x}) \in \partial g(\bar{x})$. According to the definition of the subdifferential, it remains to show that

$$(\forall \boldsymbol{y} \in \mathbb{R}^d) \quad g(\boldsymbol{y}) \geq g(\bar{\boldsymbol{x}}) + \langle -\nabla f(\bar{\boldsymbol{x}}), \boldsymbol{y} - \bar{\boldsymbol{x}} \rangle.$$

By definition of the sequence generated by the algorithm, for all $j \in \mathbb{N}^*$, there exists $\boldsymbol{w}^{(\ell_j)} \in \partial g(\boldsymbol{x}^{(\ell_j)})$ such that

$$\boldsymbol{w}^{(\ell_j)} = \nabla h_{\boldsymbol{x}^{(\ell_j-1)}}(\boldsymbol{x}^{(\ell_j-1)}) - \nabla h_{\boldsymbol{x}^{(\ell_j-1)}}(\boldsymbol{x}^{(\ell_j)}) - \nabla f(\boldsymbol{x}^{(\ell_j-1)}). \tag{3.4}$$

Using again the definition of the subdifferential, we get

$$\boldsymbol{w}^{(\ell_j)} \in \partial g(\boldsymbol{x}^{(\ell_j)}) \quad \Leftrightarrow \quad (\forall \boldsymbol{y} \in \mathbb{R}^d) \;\; g(\boldsymbol{y}) \geq g(\boldsymbol{x}^{(\ell_j)}) + \langle \boldsymbol{w}^{(\ell_j)}, \boldsymbol{y} - \boldsymbol{x}^{(\ell_j)} \rangle.$$

By (3.4), we obtain

$$\langle \boldsymbol{w}^{(\ell_j)}, \boldsymbol{y}-\boldsymbol{x}^{(\ell_j)}\rangle = \langle \nabla h_{\boldsymbol{x}^{(\ell_j-1)}}(\boldsymbol{x}^{(\ell_j-1)}) - \nabla h_{\boldsymbol{x}^{(\ell_j-1)}}(\boldsymbol{x}^{(\ell_j)}), \boldsymbol{y}-\boldsymbol{x}^{(\ell_j)}\rangle - \langle \nabla f(\boldsymbol{x}^{(\ell_j-1)}), \boldsymbol{y}-\boldsymbol{x}^{(\ell_j)}\rangle. \tag{3.5}$$

Since we know from Lemma 3.3 that $\boldsymbol{x}^{(\ell_j-1)} - \boldsymbol{x}^{(\ell_j)} \to 0$, as $j \to +\infty$, it follows from Assumption 3.2 that the first term on the right-hand side of (3.5) tends to zero as $j \to +\infty$. Additionally, since f is convex and differentiable on $\mathscr{D}$, its gradient is continuous on $\mathscr{D}$ [29, Corollary 17.42]. Consequently, since $\boldsymbol{x}^{(\ell_j-1)} \to \bar{\boldsymbol{x}}$ as $j \to +\infty$,

$$\langle \nabla f(\boldsymbol{x}^{(\ell_j-1)}), \boldsymbol{y}-\boldsymbol{x}^{(\ell_j)}\rangle \underset{j\to+\infty}{\longrightarrow} \langle \nabla f(\bar{\boldsymbol{x}}), \boldsymbol{y}-\bar{\boldsymbol{x}}\rangle.$$

Concluding the proof, we employ the lower semi-continuity of g to deduce:

$$\begin{aligned}(\forall \boldsymbol{y}\in\mathbb{R}^d)\quad g(\boldsymbol{y}) &\geq \liminf_{j\to+\infty}\left\{g(\boldsymbol{x}^{(\ell_j)})\right\} + \lim_{j\to+\infty}\left\{\langle \boldsymbol{w}^{(\ell_j)}, \boldsymbol{y}-\boldsymbol{x}^{(\ell_j)}\rangle\right\}\\ &\geq g(\bar{\boldsymbol{x}}) + \langle -\nabla f(\bar{\boldsymbol{x}}), \boldsymbol{y}-\bar{\boldsymbol{x}}\rangle.\end{aligned}$$

Thus, $\bar{\boldsymbol{x}}$ is a critical point of F, and by convexity of F, it is also a minimizer.

(ii) For all $j \in \mathbb{N}^*$, since $\boldsymbol{w}^{(\ell_j)} + \nabla f(\boldsymbol{x}^{(\ell_j)}) \in \partial F(\boldsymbol{x}^{(\ell_j)})$, we obtain

$$F(\bar{\boldsymbol{x}}) \geq F(\boldsymbol{x}^{(\ell_j)}) + \langle \boldsymbol{w}^{(\ell_j)} + \nabla f(\boldsymbol{x}^{(\ell_j)}), \bar{\boldsymbol{x}} - \boldsymbol{x}^{(\ell_j)}\rangle,$$

that is

$$0 \leq F(\boldsymbol{x}^{(\ell_j)}) - F(\bar{\boldsymbol{x}}) \leq \langle \boldsymbol{w}^{(\ell_j)}, \boldsymbol{x}^{(\ell_j)} - \bar{\boldsymbol{x}}\rangle + \langle \nabla f(\boldsymbol{x}^{(\ell_j)}), \boldsymbol{x}^{(\ell_j)} - \bar{\boldsymbol{x}}\rangle.$$

From (3.4) and the continuity of ∇f on $\mathscr{D}$, it follows for $j \to +\infty$ that

$$\begin{aligned}\langle \boldsymbol{w}^{(\ell_j)}, \boldsymbol{x}^{(\ell_j)} - \bar{\boldsymbol{x}}\rangle &\to 0\\ \langle \nabla f(\boldsymbol{x}^{(\ell_j)}), \boldsymbol{x}^{(\ell_j)} - \bar{\boldsymbol{x}}\rangle &\to 0,\end{aligned}$$

which shows that $\lim_{j\to+\infty} F(\boldsymbol{x}^{(\ell_j)}) = F(\bar{\boldsymbol{x}})$. Since we have proved in Lemma 3.2 that $(F(\boldsymbol{x}^{(\ell)}))_{\ell\in\mathbb{N}}$ is non increasing, and it is also lower bounded, it converges, and its limit is $F(\bar{\boldsymbol{x}})$.

(iii) When F has a single minimizer, $(\boldsymbol{x}^{(\ell)})_{\ell\in\mathbb{N}}$ is a bounded sequence with a unique cluster point, and it therefore converges to the unique minimizer.

□

Remark 3.2. Assume that

$$\sigma = \inf\left\{\frac{D_{h_{\boldsymbol{y}}}(\boldsymbol{y},\boldsymbol{x})}{D_{h_{\boldsymbol{y}}}(\boldsymbol{x},\boldsymbol{y})} \mid (\boldsymbol{x},\boldsymbol{y}) \in \mathscr{D}^2, \boldsymbol{y}\neq\boldsymbol{x}\right\} > 0.$$

The parameter σ acts as a symmetry measure [11] for the family of Bregman divergences of functions $(h_{\boldsymbol{y}})_{\boldsymbol{y}\in\mathscr{D}}$. For example, if $h_{\boldsymbol{y}}\colon \boldsymbol{x} \mapsto \frac{1}{2}\langle \boldsymbol{x}, \boldsymbol{Q}_{\boldsymbol{y}}\boldsymbol{x}\rangle$ where $\boldsymbol{Q}_{\boldsymbol{y}} \in \mathbb{R}^{d\times d}$ is a symmetric positive definite matrix, then $\sigma = 1$. Now introduce a parameter $\lambda \in [1, 1+\sigma-\varepsilon]$ with $\varepsilon \in (0,\sigma)$ in VBMM algorithm:

$$(\forall \ell\in\mathbb{N})\quad \boldsymbol{x}^{(\ell+1)} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^d} g(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}^{(\ell)}), \boldsymbol{x}\rangle + \frac{1}{\lambda} D_{h_{\boldsymbol{x}^{(\ell)}}}(\boldsymbol{x}, \boldsymbol{x}^{(\ell)}). \tag{3.6}$$

Choosing a value of λ greater than 1 lessens the weight on the proximity term and therefore permits larger iterate moves. Adapting the proof of Lemma 3.2 and applying in $\boldsymbol{v} = \boldsymbol{x}^{(\ell)}$ yields

$$(\forall \ell \in \mathbb{N}) \quad F(\boldsymbol{x}^{(\ell+1)}) - F(\boldsymbol{x}^{(\ell)}) \leq -\frac{1}{\lambda} D_{h_{\boldsymbol{x}^{(\ell)}}}(\boldsymbol{x}^{(\ell)}, \boldsymbol{x}^{(\ell+1)}) + \left(1 - \frac{1}{\lambda}\right) D_{h_{\boldsymbol{x}^{(\ell)}}}(\boldsymbol{x}^{(\ell+1)}, \boldsymbol{x}^{(\ell)}),$$

from which we deduce

$$(\forall \ell \in \mathbb{N}) \quad \lambda(F(\boldsymbol{x}^{(\ell+1)}) - F(\boldsymbol{x}^{(\ell)})) \leq -\varepsilon D_{h_{\boldsymbol{x}^{(\ell)}}}(\boldsymbol{x}^{(\ell+1)}, \boldsymbol{x}^{(\ell)})$$

and the convergence result in Proposition 3.1 remains valid for Algorithm (3.6).

4. Application to the Estimation of the Parameter of a Dirichlet Distribution

In this section, we present a novel application of the VBMM algorithm for estimating the parameters of a Dirichlet distribution. The Dirichlet distribution, defined on the unit simplex in $\mathbb{R}^d$, is widely used in various domains. It has applications in modeling text data, such as word appearance in documents [37], hyperspectral unmixing [38], customer segmentation based on spending patterns [39] and, more recently, in image classification using text-vision models like CLIP [40] and image restoration [41]. Additionally, it has been employed for generating DNA sequences [42].

We begin by recalling the definition of the Dirichlet distribution and its associated log-likelihood function. We then establish that the negative log-likelihood is coercive, which guarantees the existence of a minimizer. Following this, we demonstrate how the VBMM algorithm converges to a minimizer of the negative log-likelihood function. By leveraging its variable metric feature, VBMM achieves faster convergence compared to traditional methods.

Finally, through numerical experiments, we illustrate that VBMM outperforms existing methods for solving the Dirichlet Maximum Likelihood Estimation problem, particularly in terms of convergence speed.

4.1. Maximum likelihood estimation for a Dirichlet distribution.

Let $M \in \mathbb{N}^*$ denote the number of samples and let $(\boldsymbol{z}_m)_{1\leq m\leq M} \in (\Delta_d)^M$, where Δ_d denotes the open unit simplex of $\mathbb{R}^d$. Assume the vectors $(\boldsymbol{z}_m)_{1\leq m\leq M}$ are samples drawn from a same Dirichlet distribution with parameter $\boldsymbol{\alpha} = (\alpha_i)_{1\leq i\leq d} \in (0,+\infty)^d$, corresponding to the probability density function

$$(\forall \boldsymbol{z} = (z_i)_{1\leq i\leq d} \in [0,+\infty)^d) \quad \mathrm{p}(\boldsymbol{z};\boldsymbol{\alpha}) := \frac{1}{\mathscr{B}(\boldsymbol{\alpha})} \prod_{i=1}^{d} z_i^{\alpha_i - 1}\, \mathbb{1}_{\boldsymbol{z}\in\Delta_d},$$

where

$$\mathscr{B}(\boldsymbol{\alpha}) := \prod_{i=1}^{d} \Gamma(\alpha_i) \Big/ \Gamma\left(\sum_{i=1}^{d} \alpha_i\right)$$

and Γ denotes the Gamma function.

To determine the parameter of this distribution, we intend to minimize the negative log-likelihood function, defined as

$$(\forall \boldsymbol{\alpha} = (\alpha_i)_{1\le i\le d} \in \mathbb{R}^d) \quad f(\boldsymbol{\alpha}) := \begin{cases} MG(\boldsymbol{\alpha}) - \displaystyle\sum_{m=1}^{M}\sum_{i=1}^{d}(\alpha_i - 1)\ln z_{m,i} & \text{if } \boldsymbol{\alpha} \in \mathscr{D} := (0,+\infty)^d, \\ +\infty & \text{otherwise,} \end{cases} \tag{4.1}$$

where

$$(\forall \boldsymbol{\alpha} = (\alpha_i)_{1\le i\le d} \in \mathscr{D}) \quad G(\boldsymbol{\alpha}) := \sum_{i=1}^{d} \ln\Gamma(\alpha_i) - \ln\Gamma\Big(\sum_{i=1}^{d}\alpha_i\Big).$$

4.2. **Existence of a unique minimizer.**

Proposition 4.1. *Assume that $d \ge 2$ and $M \ge 2$. If at least two samples in the set $\{\boldsymbol{z}_m\}_{1\le m\le M}$ are not identical, then function f is lower-semicontinuous, strictly convex, and coercive. Thus, it admits a unique minimizer.*

Proof.

Lower-semicontinuity. Function f is continuous on its domain $\mathscr{D}$. Let us study what happens at the boundary of this domain. Let $(\boldsymbol{\alpha}^{(\ell)})_{\ell\in\mathbb{N}}$ be a sequence of $(0,+\infty)^d$ converging to $\boldsymbol{\alpha}^*$, where $\boldsymbol{\alpha}^*$ belongs to the boundary of $(0,+\infty)^d$. If $\{i \in \{1,\dots,d\} \mid \alpha_i^* = 0\} \neq \{1,\dots,d\}$, then it is clear that $f(\boldsymbol{\alpha}^{(\ell)}) \underset{\ell\to+\infty}{\longrightarrow} +\infty$. Now let $\boldsymbol{\alpha}^* = \boldsymbol{0}$. Then, since $\Gamma(\alpha) \sim \frac{1}{\alpha}$ as $\alpha \to 0+$, we get

$$\frac{\Gamma(\alpha_1)\dots\Gamma(\alpha_d)}{\Gamma(\sum_{i=1}^d \alpha_i)} \sim \frac{1}{\alpha_1\dots\alpha_d}\sum_{i=1}^{d}\alpha_i = \sum_{i=1}^{d}\frac{1}{\prod_{j\neq i}\alpha_j}$$

and, consequently,

$$G(\boldsymbol{\alpha}) \sim \ln\left(\sum_{i=1}^{d}\frac{1}{\prod_{j\neq i}\alpha_j^{(\ell)}}\right) \underset{\ell\to+\infty}{\longrightarrow} +\infty,$$

so that $\lim_{\ell\to+\infty} f(\boldsymbol{\alpha}^{(\ell)}) = +\infty = f(\boldsymbol{\alpha}^*)$. This shows that f is lower-semicontinuous.

Coercivity. Let us show that

$$\lim_{\substack{\|\boldsymbol{\alpha}\|\to+\infty \\ \boldsymbol{\alpha}\in\mathscr{D}}} f(\boldsymbol{\alpha}) = +\infty. \tag{4.2}$$

According to a variant of Stirling's formula [43, page 257, 6.1.38], for all $\alpha > 0$, there exists $\theta \in (0,1)$ such that

$$\begin{aligned}\Gamma(\alpha) = \frac{\Gamma(\alpha+1)}{\alpha} &= \alpha^{-1}\sqrt{2\pi}\alpha^{\alpha+1/2}\exp\Big(-\alpha+\frac{\theta}{12\alpha}\Big) \\ &= \sqrt{2\pi}\alpha^{\alpha-1/2}\exp\Big(-\alpha+\frac{\theta}{12\alpha}\Big).\end{aligned} \tag{4.3}$$

For all vectors $\boldsymbol{\alpha} = (\alpha_i)_{1\le i\le d} \in (0,+\infty)^d$, we define $\bar{\alpha} := \sum_{i=1}^d \alpha_i$. Additionally, we denote by $(\theta_i)_{1\le i\le d} \in (0,1)^d$ the set of constants corresponding to each α_i as per equation (4.3). Similarly,

$\bar{\theta} \in (0,1)$ is associated with $\bar{\alpha}$. We can write, for all $\boldsymbol{\alpha} \in (0,+\infty)^d$,

$$\begin{aligned} G(\boldsymbol{\alpha}) &= \sum_{i=1}^{d} \left((\alpha_i - 1/2)\ln\alpha_i - \alpha_i + \frac{\theta_i}{12\alpha_i} + \frac{1}{2}\ln(2\pi) \right) \\ &\quad - (\bar{\alpha} - 1/2)\ln\bar{\alpha} + \bar{\alpha} - \frac{\bar{\theta}}{12\bar{\alpha}} - \frac{1}{2}\ln(2\pi), \\ &= A(\boldsymbol{\alpha}) + B(\boldsymbol{\alpha}), \end{aligned}$$

where

$$\begin{aligned} A(\boldsymbol{\alpha}) &\coloneqq \sum_{i=1}^{d} (\alpha_i - \frac{1}{2})\ln\alpha_i - (\bar{\alpha} - 1/2)\ln\bar{\alpha}, \\ B(\boldsymbol{\alpha}) &\coloneqq \frac{1}{12}\left[\sum_{i=1}^{d} \frac{\theta_i}{\alpha_i} - \frac{\bar{\theta}}{\bar{\alpha}} \right] + \frac{d-1}{2}\ln(2\pi). \end{aligned}$$

Setting $\eta_i \coloneqq \frac{\alpha_i}{\bar{\alpha}} \in (0,1]$, we have

$$\begin{aligned} A(\boldsymbol{\alpha}) &= \sum_{i=1}^{d} (\eta_i\bar{\alpha} - 1/2)\ln(\eta_i\bar{\alpha}) - (\bar{\alpha} - 1/2)\ln\bar{\alpha}, \\ &= \sum_{i=1}^{d} \eta_i\bar{\alpha}\ln\eta_i + \sum_{i=1}^{d} \eta_i\bar{\alpha}\ln\bar{\alpha} - \frac{1}{2}\sum_{i=1}^{d}\ln\eta_i - \frac{d}{2}\ln\bar{\alpha} - (\bar{\alpha} - 1/2)\ln\bar{\alpha}, \\ &= \bar{\alpha}\sum_{i=1}^{d} \eta_i\ln\eta_i - \frac{1}{2}\sum_{i=1}^{d}\ln\eta_i + \frac{1-d}{2}\ln\bar{\alpha}. \end{aligned}$$

Since, for all $i \in \{1,\dots,d\}$, $\eta_i \leq 1$, the following lower bound is obtained:

$$A(\boldsymbol{\alpha}) \geq \bar{\alpha}\sum_{i=1}^{d} \eta_i\ln\eta_i + \frac{1-d}{2}\ln\bar{\alpha}.$$

Additionally, since for all $i \in \{1,\dots,d\}$, $\theta_i > 0$ and $\bar{\theta} < 1$, we obtain

$$B(\boldsymbol{\alpha}) > -\frac{1}{12\bar{\alpha}} + \frac{d-1}{2}\ln(2\pi).$$

Therefore, we can lower-bound G as follows

$$G(\boldsymbol{\alpha}) > \bar{\alpha}\sum_{i=1}^{d} \eta_i\ln\eta_i + \frac{1-d}{2}\ln\bar{\alpha} - \frac{1}{12\bar{\alpha}} + \frac{d-1}{2}\ln(2\pi).$$

Let us now go back to the objective function (4.1). The following holds:

$$\begin{aligned} f(\boldsymbol{\alpha}) &= MG(\boldsymbol{\alpha}) - \sum_{i=1}^{d} (\alpha_i - 1)\ln\left(\prod_{m=1}^{M} z_{m,i} \right) \\ &= M\left(G(\boldsymbol{\alpha}) - \sum_{i=1}^{d} (\eta_i\bar{\alpha} - 1)\ln\tilde{z}_i \right), \end{aligned}$$

where

$$(\forall i \in \{1,\dots,d\}) \quad \tilde{z}_i \coloneqq \left(\prod_{m=1}^{M} z_{m,i} \right)^{1/M}.$$

Then,

$$f(\boldsymbol{\alpha}) > M\left(-\sum_{i=1}^{d}(\eta_i\bar{\alpha}-1)\ln\tilde{z}_i+\bar{\alpha}\sum_{i=1}^{d}\eta_i\ln\eta_i+\frac{1-d}{2}\ln\bar{\alpha}-\frac{1}{12\bar{\alpha}}+\frac{d-1}{2}\ln(2\pi)\right),$$
$$= M\left(\bar{\alpha}\sum_{i=1}^{d}\eta_i\ln\left(\frac{\eta_i}{\tilde{z}_i}\right)+\sum_{i=1}^{d}\ln\tilde{z}_i+\frac{1-d}{2}\ln\bar{\alpha}-\frac{1}{12\bar{\alpha}}+\frac{d-1}{2}\ln(2\pi)\right).$$

Let us recall the definition of the Kullback-Leibler (KL), divergence:

$$(\forall \boldsymbol{u}=(u_i)_{1\le i\le d}\in[0,+\infty)^d)(\forall \boldsymbol{v}=(v_i)_{1\le i\le d}\in(0,+\infty)^d)$$
$$D_{\mathrm{KL}}(\boldsymbol{u},\boldsymbol{v})=\sum_{i=1}^{d}u_i\ln\left(\frac{u_i}{v_i}\right)-\sum_{i=1}^{d}u_i+\sum_{i=1}^{d}v_i.$$

Since $\sum_{i=1}^{d}\eta_i=1$, the following holds, for $\boldsymbol{\eta}=(\eta_i)_{1\le i\le d}$ and $\tilde{\boldsymbol{z}}=(\tilde{z}_i)_{1\le i\le d}$,

$$D_{\mathrm{KL}}(\boldsymbol{\eta},\tilde{\boldsymbol{z}})=\sum_{i=1}^{d}\eta_i\ln\left(\frac{\eta_i}{\tilde{z}_i}\right)-1+\sum_{i=1}^{d}\tilde{z}_i,$$

which implies that

$$f(\boldsymbol{\alpha}) > M\left(\bar{\alpha}\left(D_{\mathrm{KL}}(\boldsymbol{\eta},\tilde{\boldsymbol{z}})+1-\sum_{i=1}^{d}\tilde{z}_i\right)+\sum_{i=1}^{d}\ln\tilde{z}_i+\frac{1-d}{2}\ln\bar{\alpha}-\frac{1}{12\bar{\alpha}}+\frac{d-1}{2}\ln(2\pi)\right).$$

Using the nonnegativity of the KL divergence, we deduce that

$$f(\boldsymbol{\alpha}) > M\left(\bar{\alpha}\left(1-\sum_{i=1}^{d}\tilde{z}_i\right)+\sum_{i=1}^{d}\ln\tilde{z}_i+\frac{1-d}{2}\ln\bar{\alpha}-\frac{1}{12\bar{\alpha}}+\frac{d-1}{2}\ln(2\pi)\right). \tag{4.4}$$

If we show that $1-\sum_{i=1}^{d}\tilde{z}_i>0$, it follows that the term on the right-hand side of equation (4.4) tends to infinity as $\|\boldsymbol{\alpha}\|$ goes to infinity. According to the inequality of arithmetic and geometric means, for all $i\in\{1,\dots,d\}$,

$$\tilde{z}_i=\left(\prod_{m=1}^{M}z_{m,i}\right)^{1/M}\le\frac{1}{M}\sum_{m=1}^{M}z_{m,i},$$

with equality if and only if for all $i\in\{1,\dots,d\}$, $z_{m,i}=z_{0,i}$. Since the samples $(\boldsymbol{z}_m)_{1\le m\le M}$ belong to the unit simplex of $\mathbb{R}^d$,

$$\sum_{i=1}^{d}\tilde{z}_i\le\frac{1}{M}\sum_{i=1}^{d}\sum_{m=1}^{M}z_{m,i}\le 1. \tag{4.5}$$

The inequality (4.5) is strict if there exists $i\in\{1,\dots,d\}$ and $(m,\ell)\in\{1,\dots,M\}$ with $m\neq\ell$ such that $z_{m,i}\neq z_{\ell,i}$. In other words, the inequality is strict if at least two samples are not identical, which has been assumed. Finally, we have proved that the coercivity condition (4.2) holds.

Strict convexity. The Dirichlet distribution is part of the Exponential family. In addition, it is associated to the sufficient statistics

$$(\forall z=(z_i)_{1\le i\le m}\in\Delta_d)\quad \boldsymbol{T}(\boldsymbol{z})=(-\ln z_i)_{1\le i\le M}.$$

Since there is no linear dependence between the components of $\boldsymbol{T}(\boldsymbol{z})$, the Dirichlet distribution defines a minimal exponential family. In such a case, the negative log-likelihood is known to

be strictly convex [44, Thm. 1.13], [45]. As an alternative proof, it was shown in [46] that the Hessian of the negative log-likelihood is positive definite. □

4.3. **Existence of Bregman majorants.** Building upon the following lemma, we construct a Bregman tangent majorant of f.

Lemma 4.1 ([47])**.** *Let φ be a twice-continuously differentiable function on $[0,+\infty)$. Assume that φ'' is decreasing on $[0,+\infty)$. Let $t \in [0,+\infty)$ and let*

$$c(t) = \begin{cases} \varphi''(0) & \text{if } t = 0, \\ 2\dfrac{\varphi(0) - \varphi(t) + \varphi'(t)t}{t^2} & \text{otherwise.} \end{cases} \tag{4.6}$$

Then, for every $x \in [0,+\infty)$, it holds

$$\varphi(x) \leq \varphi(t) + \varphi'(t)(x-t) + \frac{1}{2}c(t)(x-t)^2.$$

Proposition 4.2. *Let $\boldsymbol{\beta} \in (0,+\infty)^d$. The negative log-likelihood in (4.1) admits a Bregman tangent majorant at $\boldsymbol{\beta}$ associated with the Bregman function*

$$(\forall \boldsymbol{\alpha} = (\alpha_i)_{1\leq i\leq d} \in (0,+\infty)^d) \quad h_{\boldsymbol{\beta}}(\boldsymbol{\alpha}) = h^{\mathrm{cst}}(\boldsymbol{\alpha}) + h^{\mathrm{var}}_{\boldsymbol{\beta}}(\boldsymbol{\alpha}), \tag{4.7}$$

where h^{cst} and $h^{\mathrm{var}}_{\boldsymbol{\beta}}$ are respectively the constant and variable components defined as

$$\begin{cases} h^{\mathrm{cst}}(\boldsymbol{\alpha}) := -M\displaystyle\sum_{i=1}^{d} \ln \alpha_i \\ h^{\mathrm{var}}_{\boldsymbol{\beta}}(\boldsymbol{\alpha}) := M\displaystyle\sum_{i=1}^{d} c(\beta_i)\frac{\alpha_i^2}{2}, \end{cases}$$

where $c\colon [0,+\infty) \longrightarrow [0,+\infty)$ is defined by Lemma 4.1 applied to the function $\varphi = \ln\Gamma(\cdot+1)$.

Proof. Recalling that $\ln\Gamma(\cdot+1) = \ln\Gamma + \ln$, we decompose f into

$$f(\boldsymbol{\alpha}) = f_1(\boldsymbol{\alpha}) + f_2(\boldsymbol{\alpha}) + f_3(\boldsymbol{\alpha}),$$

where

$$f_1(\boldsymbol{\alpha}) = -\sum_{m=1}^{M}\left(\sum_{i=1}^{d}(\alpha_i - 1)\ln z_{m,i} + \ln\Gamma\left(\sum_{i=1}^{d}\alpha_i\right)\right),$$

$$f_2(\boldsymbol{\alpha}) = M\sum_{i=1}^{d}\ln\Gamma(\alpha_i+1),$$

and

$$f_3(\boldsymbol{\alpha}) = -M\sum_{i=1}^{d}\ln\alpha_i.$$

We now derive upper tangent bounds at $\boldsymbol{\beta}$ for each of these functions. Recalling that the Gamma function is log-convex, f_1 is the sum of a linear function and a concave function, and the following tangent inequality holds

$$f_1(\boldsymbol{\alpha}) \leq f_1(\boldsymbol{\beta}) + \nabla f_1(\boldsymbol{\beta})^\top(\boldsymbol{\alpha} - \boldsymbol{\beta}). \tag{4.8}$$

In addition, the function $\ln\Gamma(\cdot+1)$ is twice-continuously differentiable on $(0,+\infty)^d$ and its second derivative, which is a shifted version of the Trigamma function, admits the following expansion

$$(\forall t \in \mathbb{R}) \quad (\ln\Gamma(\cdot+1))''(t) = \sum_{\ell=0}^{+\infty} \frac{1}{(t+1+\ell)^2},$$

provided that $-t \notin \mathbb{N}$. This shows that $(\ln\Gamma(\cdot+1))''$ is decreasing on $[0,+\infty)$. Therefore, Lemma 4.1 applies to $\ln\Gamma(\cdot+1)$, yielding the upper-bound

$$f_2(\boldsymbol{\alpha}) \leq f_2(\boldsymbol{\beta}) + \nabla f_2(\boldsymbol{\beta})^\top(\boldsymbol{\alpha}-\boldsymbol{\beta}) + \frac{M}{2}\sum_{i=1}^{d} c(\beta_i)(\alpha_i-\beta_i)^2,$$

which can also be written as

$$f_2(\boldsymbol{\alpha}) \leq f_2(\boldsymbol{\beta}) + \nabla f_2(\boldsymbol{\beta})^\top(\boldsymbol{\alpha}-\boldsymbol{\beta}) + D_{h^{\mathrm{var}}_{\boldsymbol{\beta}}}(\boldsymbol{\alpha},\boldsymbol{\beta}). \tag{4.9}$$

Finally, by definition of the Bregman divergence and since f_3 coincides with h^{cst}, we obtain

$$f_3(\boldsymbol{\alpha}) = f_3(\boldsymbol{\beta}) + \nabla f_3(\boldsymbol{\beta})^\top(\boldsymbol{\alpha}-\boldsymbol{\beta}) + D_{h^{\mathrm{cst}}}(\boldsymbol{\alpha},\boldsymbol{\beta}). \tag{4.10}$$

Thus, since $D_{h^{\mathrm{cst}}+h^{\mathrm{var}}_{\boldsymbol{\beta}}} = D_{h^{\mathrm{cst}}} + D_{h^{\mathrm{var}}_{\boldsymbol{\beta}}}$, we deduce from (4.8), (4.9), and (4.10) that

$$f(\boldsymbol{\alpha}) \leq f(\boldsymbol{\beta}) + \nabla f(\boldsymbol{\beta})^\top(\boldsymbol{\alpha}-\boldsymbol{\beta}) + D_{h_{\boldsymbol{\beta}}}(\boldsymbol{\alpha},\boldsymbol{\beta}).$$

□

Remark 4.1. The majorants established in Proposition 4.2 can be written in the form (2.4), with for all $i \in \{1,\ldots,d\}$, for all $\boldsymbol{\beta} \in (0,+\infty)^d$, $a_i(\boldsymbol{\beta}) = M$ and $b_i(\boldsymbol{\beta}) = Mc(\beta_i)$, and for all $t \in (0,+\infty)$, $v_i(t) = -\ln t$. Additionally, the curvature function c defined in (4.6) is represented in Figure 1. It satisfies $c(0) = \pi^2/6$ and

$$(\forall t \in [0,+\infty)) \quad c(t) > 0 \text{ and } \lim_{t\to+\infty} c(t) = 0.$$

The family of Bregman functions defined in equation (4.7) satisfies the assumptions of Proposition 3.1 for suitable choices of function g. For all $\boldsymbol{\beta} \in (0,+\infty)^d$, the function $h_{\boldsymbol{\beta}}$ is strictly convex and differentiable on its domain which is $\mathscr{D} = (0,+\infty)^d$. Moreover, the majorization property is verified as per Proposition 4.2, thereby satisfying Assumption 2.2. Regarding Assumptions 3.1 and 3.2, they are clearly met as indicated in the remarks made after stating these assumptions.

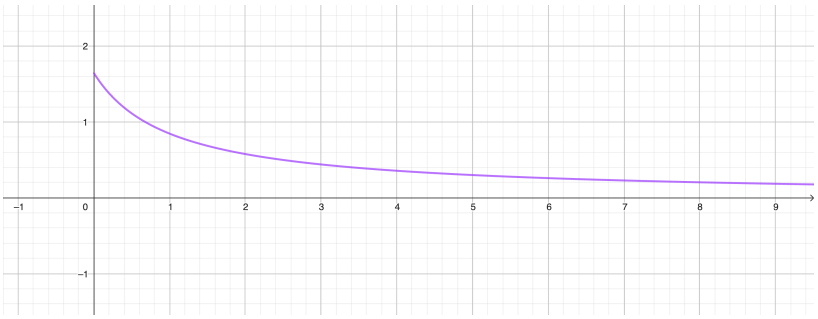

FIGURE 1. Curvature function (4.6) when $\varphi = \ln\Gamma(\cdot+1)$.

4.4. **Algorithm.** For a given value of $\boldsymbol{\beta} = (\beta_i)_{1\leq i\leq d} \in (0,+\infty)^d$, we can find a closed-form expression of the minimizer $\widehat{\boldsymbol{\alpha}}$ of the Bregman majorant given by Proposition 4.2. Since the majorant has a separable form, each component $\widehat{\alpha}_i$, $i \in \{1,\dots,d\}$, of $\widehat{\boldsymbol{\alpha}}$ is the minimizer of the function defined on $(0,+\infty)$ as

$$\alpha_i \mapsto \left((\ln\Gamma)'(\beta_i) - (\ln\Gamma)'\Big(\sum_{j=1}^{d}\beta_j\Big) - \frac{1}{M}\sum_{m=1}^{M}\ln z_{m,i} \right)(\alpha_i-\beta_i) + \frac{c(\beta_i)}{2}(\alpha_i-\beta_i)^2 - \ln\alpha_i + \frac{\alpha_i-\beta_i}{\beta_i},$$

where, as already mentioned, the curvature $c(\beta_i)$ is positive. We deduce that $\widehat{\alpha}_i$ is the unique positive root of the second order polynomial equation

$$c(\beta_i)\alpha_i^2 + \delta_i(\boldsymbol{\beta})\alpha_i = 1,$$

where $\delta_i(\boldsymbol{\beta})$ is given by

$$\delta_i(\boldsymbol{\beta}) := (\ln\Gamma)'(\beta_i+1) - (\ln\Gamma)'\Big(\sum_{j=1}^{d}\beta_j\Big) - c(\beta_i)\beta_i - \frac{1}{M}\sum_{m=1}^{M}\ln z_{m,i}.$$

Therefore, we obtain the following closed-form expression for the minimizer:

$$\widehat{\alpha}_i = \frac{-\delta_i(\boldsymbol{\beta}) + \sqrt{\big(\delta_i(\boldsymbol{\beta})\big)^2 + 4c(\beta_i)}}{2c(\beta_i)}.$$

The final Variable Bregman MM updates are described in Algorithm 2. Note that it only requires the evaluation of the log-Gamma function and of its derivative, the Digamma function.

Algorithm 2: VBMM for Dirichlet parameter estimation

Initialize $\boldsymbol{\alpha}^{(0)} \in (0,+\infty)^d$.
for $\ell = 0,1,\dots,$ **do**
  for $i \in \{1,\dots,d\}$ **do**

$$c_i^{(\ell)} = 2\left(-\ln\Gamma(\alpha_i^{(\ell)}+1) + \alpha_j^{(\ell)}(\ln\Gamma)'(\alpha_i^{(\ell)}+1)\right)/(\alpha_i^{(\ell)})^2$$

$$\delta_i^{(\ell)} = (\ln\Gamma)'(\alpha_i^{(\ell)}+1) - (\ln\Gamma)'\left(\sum_{j=1}^{d}\alpha_j^{(\ell)}\right) - c_i^{(\ell)}\alpha_i^{(\ell)} - \frac{1}{M}\sum_{m=1}^{M}\ln z_{m,i}.$$

$$\alpha_i^{(\ell+1)} = \left(-\delta_i^{(\ell)} + \sqrt{(\delta_i^{(\ell)})^2 + 4c_i^{(\ell)}}\right)\Big/2c_i^{(\ell)}.$$

4.5. **Numerical experiments.** All numerical experiments were run in Python 3.8 on an Apple M1 CPU.

4.5.1. *Unconstrained case.* We first consider the case when f is defined by(4.1) and the penalty function g is zero. The cost function thus reduces to $F = f$. We apply the VBMM algorithm (see Algorithm 2) to minimize F.

We compare the computational efficiency of VBMM with two state-of-the-art algorithms used to solve the Dirichlet maximum-likelihood problem. The first one is the algorithm described in [46, 48, 49], which is a Newton algorithm adapted to deal with the constraint $\boldsymbol{\alpha} \in$

$(0,+\infty)^d$ and requires computing the second derivative of the log-Gamma function at each iteration. The second one, proposed in [49], constructs an upper bound on the negative log-likelihood, which is tight at the current iterate. Unlike ours, this upper bound has no closed-form minimizer. However, the sub-optimization problem can be rewritten as a fixed-point problem, which can be solved with a Newton method separately on each component. Finally, we highlight the advantages of using Bregman functions that adapt dynamically to the iterates by comparing the Variable Bregman Majorization-Minimization (VBMM) algorithm with the fixed-metric Bregman Majorization-Minimization (BMM) approach. In BMM, the variable component $\boldsymbol{\alpha} \mapsto h_{\boldsymbol{\beta}}^{\text{var}}(\boldsymbol{\alpha})$ is replaced with a function independent of $\boldsymbol{\beta}$, namely

$$\boldsymbol{\alpha} \mapsto M \sup_{t\in[0,+\infty)} \{c(t)\} \sum_{i=1}^{d} \frac{\alpha_i^2}{2} = \frac{M\pi^2}{12} \sum_{i=1}^{d} \alpha_i^2.$$

Experimental setup. We conduct a series of experiments where we set the dimension $d$of the data to 1000 and the number of samples M to 500 across all experiments. We sample from a Dirichlet distribution with parameter $\boldsymbol{\alpha}_{\text{true}} \in \mathbb{R}^d$ for different values of $\boldsymbol{\alpha}_{\text{true}}$, allowing us to scan a wide range of means and variances for the data distribution. We first define the following three vectors:

- $\boldsymbol{m}_1 = \mathbf{1}_N$;
- $\boldsymbol{m}_2 = (m_{2,i})_{1\le i\le d}$ where, for all $i \in \{1,\dots,d\}$, $m_{2,i} = 10$ if $i = 1$, $m_{2,i} = 1$ otherwise;
- $\boldsymbol{m}_3 = (m_{3,i})_{1\le i\le d}$ where, for all $i \in \{1,\dots,d\}$, $m_{3,i} = i$.

For $j \in \{1,2,3\}$, we define the normalized vector $\bar{\boldsymbol{m}}_j = \boldsymbol{m}_j / \sum_{i=1}^{d} m_{j,i} \in \Delta_d$. We also introduce three scaling factors $s_1 = 100$, $s_2 = 10$, and $s_3 = 1$. Given $\bar{\boldsymbol{m}} \in \{\bar{\boldsymbol{m}}_1, \bar{\boldsymbol{m}}_2, \bar{\boldsymbol{m}}_3\}$ and $s \in \{s_1, s_2, s_3\}$, we set the true parameter $\boldsymbol{\alpha}_{\text{true}}$ as

$$\boldsymbol{\alpha}_{\text{true}} = s\bar{\boldsymbol{m}}.$$

In this case, the mean of the Dirichlet distribution is $\bar{\boldsymbol{m}}$, and the variance is $\frac{\bar{\boldsymbol{m}}(1-\bar{\boldsymbol{m}})}{s+1}$. Therefore, a large value of s results in a small variance of the samples.

Once $\boldsymbol{\alpha}_{\text{true}}$ is set, we draw our data samples from the Dirichlet distribution with the chosen true parameter. Subsequently, we aim to estimate the Dirichlet parameter solution to the Maximum Likelihood problem. For each method, the initial estimate $\alpha^{(0)}$ is uniformly set to a vector with value $10\,\mathbf{1}_N$. Each experimental setup was averaged over 1000 random experiments.

Results. To assess the convergence speed of all three methods on the experimental setup previously described, we evaluate the relative squared error (RSE) with respect to the optimal parameter $\boldsymbol{\alpha}_{\text{opt}}$ at each iteration ℓ against time in seconds. The RSE at iteration ℓ is given by $\frac{\|\boldsymbol{\alpha}^{(\ell)} - \boldsymbol{\alpha}_{\text{opt}}\|^2}{\|\boldsymbol{\alpha}_{\text{opt}}\|^2}$ where $\boldsymbol{\alpha}^{(\ell)}$ is the estimated parameter at iteration ℓ. Note that, as the number of samples is finite, the maximizer of the likelihood $\boldsymbol{\alpha}_{\text{opt}}$ differs from $\boldsymbol{\alpha}_{\text{true}}$ in general.

In Figure 2, we present convergence plots displaying the distance to the optimum versus time for different values of $\boldsymbol{\alpha}_{\text{true}}$. We observe that, in every configuration, VBMM converges faster than the three other methods.

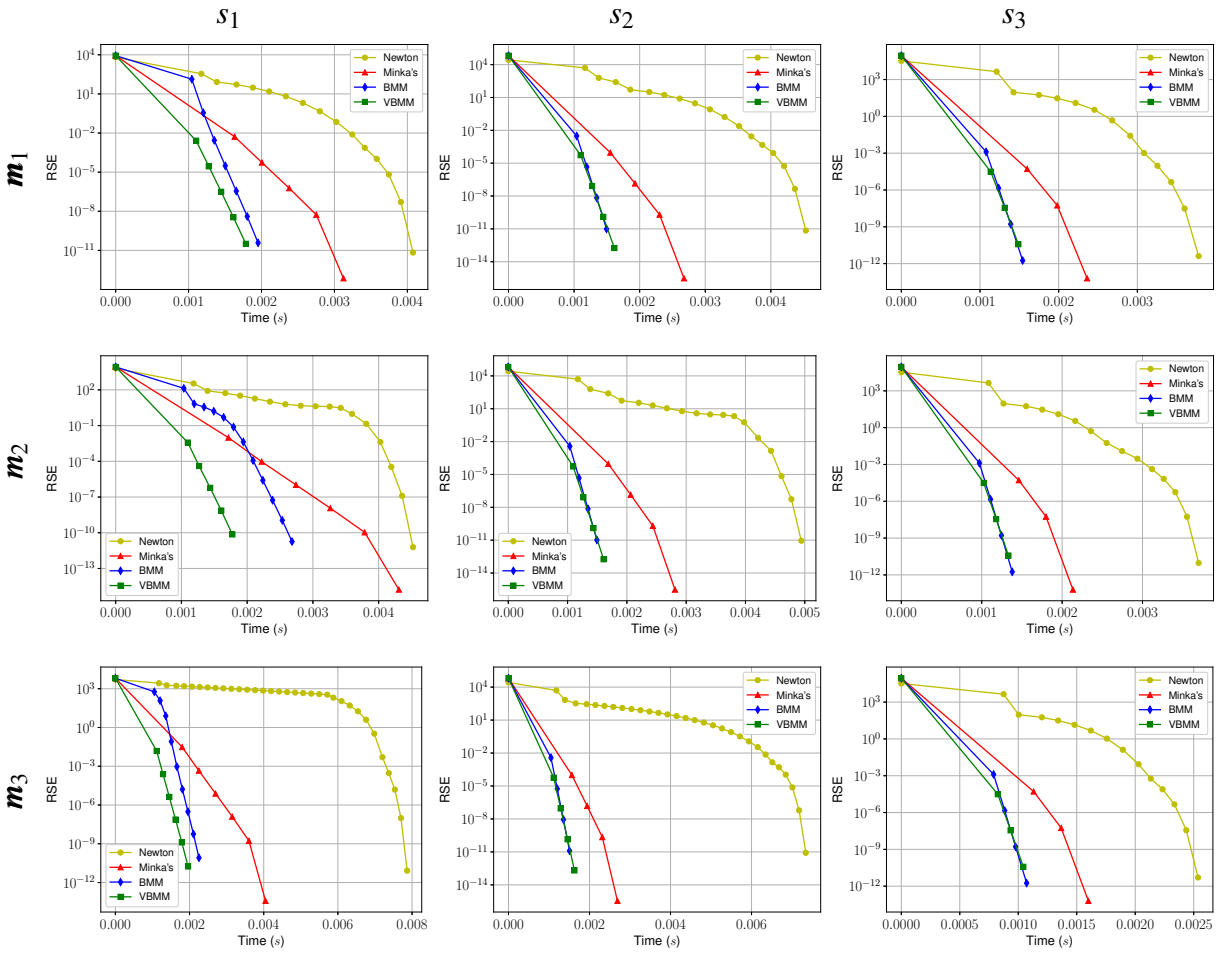

FIGURE 2. Distance to optimum versus time for different values of $\boldsymbol{\alpha}_{\text{true}}$, with $M = 500$ samples and sample size $d = 1000$. Rows from top to bottom: $\boldsymbol{\alpha}_{\text{true}}$ defined with respectively $\boldsymbol{m}_1$, $\boldsymbol{m}_2$, and $\boldsymbol{m}_3$. Columns from left to right: $\boldsymbol{\alpha}_{\text{true}}$ defined with respectively s_1, s_2, and s_3.

4.5.2. *Use of a separable constraint.* Our approach also allows for the inclusion of a non-zero function g, which can be useful for enforcing constraints on the estimated Dirichlet parameters. Specifically, we consider a function having the following separable form:

$$(\forall \boldsymbol{x} = (x_i)_{1\leq i\leq d} \in \mathbb{R}^d) \quad g(\boldsymbol{x}) := \sum_{i=1}^{d} \iota_{[r_i^-, r_i^+]}(x_i),$$

where for each $i \in \{1,\dots,d\}$, $(r_i^-, r_i^+) \in (0,+\infty)^2$.

An example of such a constraint arises frequently in the Latent Dirichlet Allocation (LDA) problem, where setting $r_i^- = \varepsilon$ with $\varepsilon > 0$ a constant arbitrarily close to zero, and $r_i^+ = 1$ is common. In practice, this constraint reflects the sparsity of word counts in documents.

In this case, minimizing the Bregman majorant simply reduces to minimizing a convex one-variable function on $[r_i^-, r_i^+]$, which is equivalent to projecting the unconstrained update onto $[r_i^-, r_i^+]$. The complete algorithm is detailed in Algorithm 3. In our experiments, we set $r_i^- = 10^{-10}$ and $r_i^+ = 1$, for all $i \in \{1,\dots,d\}$.

In Figure 3, we plot both the RSE and the negative log-likelihood evaluated at the iterates of VBMM as a function of time. As theoretically expected, the loss is monotonically decreasing (linearly) and converges fast. Note that neither of the two alternative methods we compared against previously are designed to handle constrained problems, highlighting an additional benefit from our approach.

Algorithm 3: VBMM for Dirichlet parameter estimation with a box constraint

Initialize $\boldsymbol{\alpha}^{(0)} \in (0,+\infty)^d$.
for $\ell = 0, 1, \ldots,$ **do**
  for $i \in \{1, \ldots, d\}$ **do**

$$c_i^{(\ell)} = 2\left(\ln\Gamma(1) - \ln\Gamma(\alpha_i^{(\ell)}+1) + (\ln\Gamma)'(\alpha_i^{(\ell)}+1)\alpha_i^{(\ell)}\right)/(\alpha_i^{(\ell)})^2$$

$$\delta_i^{(\ell)} = (\ln\Gamma)'(\alpha_i^{(\ell)}+1) - (\ln\Gamma)'\left(\sum_{j=1}^{d}\alpha_j^{(\ell)}\right) - c_i^{(\ell)}\alpha_i^{(\ell)} - \frac{1}{M}\sum_{m=1}^{M}\ln z_{m,i}$$

$$\alpha_i^{(\ell+1)} = \mathrm{proj}_{[r_i^-, r_i^+]}\left(\left(-\delta_i^{(\ell)} + \sqrt{(\delta_i^{(\ell)})^2 + 4c_i^{(\ell)}}\right)\Big/ 2c_i^{(\ell)}\right).$$

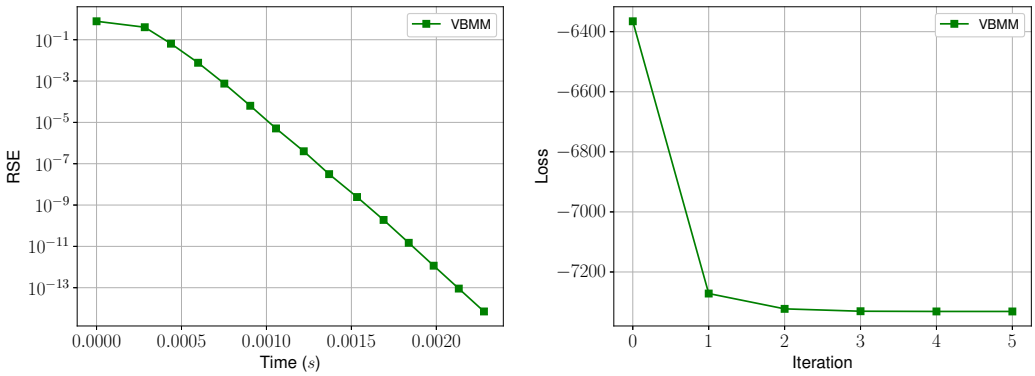

FIGURE 3. (Left) RSE versus lapsed time. (Right) Function f minus the optimal value f^* versus elapsed time. The loss is averaged over 1000 experiments where $\boldsymbol{\alpha}_{\text{true}}$ is uniformly sampled in $[0,2]^{1000}$. The number of samples for each experiment is $M = 500$.

5. CONCLUSION

We introduced the Variable Bregman Majorization-Minimization (VBMM) algorithm as a versatile extension of the Bregman Proximal Gradient method. By allowing the Bregman divergence to adapt dynamically at each iteration, VBMM increases flexibility in optimization and achieves faster convergence. We established the subsequential convergence of the algorithm under mild assumptions on the family of Bregman metrics. A novel application to the estimation of Dirichlet distribution parameters demonstrated the practical efficiency of VBMM, outperforming existing methods in terms of convergence speed.

Future research could focus on strengthening our convergence results. For instance, leveraging the Kurdyka-Łojasiewicz property, as in [23], might lead to stronger convergence results. Moreover, relaxing the convexity requirement on the objective function f could broaden the applicability of VBMM to a wider class of optimization problems.

Acknowledgments
The authors thank Michael Adipoetra for his help on the numerical experiments of this paper. Ségolène Martin's work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – The Berlin Mathematics Research Center MATH+ (EXC-2046/1, project ID: 390685689).

REFERENCES

[1] M. Fukushima and H. Mine, A generalized proximal point algorithm for certain non-convex minimization problems, International Journal of Systems Science, 12 (1981), 989-1000.
[2] P. L. Combettes and V. R. Wajs, Signal recovery by proximal forward-backward splitting, Multiscale Modeling & Simulation, 4 (2005), 1168-1200.
[3] P. Pérez-Aros and E. Vilches, An enhanced Baillon–Haddad theorem for convex functions defined on convex sets, Applied Mathematics & Optimization, 83 (2021), 2241-2252.
[4] M. Bertero, P. Boccacci, G. Desiderà, and G. Vicidomini, Image deblurring with Poisson data: from cells to galaxies, Inverse Problems, 25 (2009), 123006.
[5] J. M. Joyce, Kullback-Leibler Divergence, Springer, Berlin, Heidelberg, 2011.
[6] Y. Censor and S. A. Zenios, Proximal minimization algorithm with D-functions, Journal of Optimization Theory and Applications, 73 (1992), 451-464.
[7] M. Teboulle, Entropic proximal mappings with applications to nonlinear programming, Mathematics of Operations Research, 17 (1992), 670-690.
[8] G. Chen and M. Teboulle, Convergence analysis of a proximal-like minimization algorithm using Bregman functions, SIAM Journal on Optimization, 3 (1992), 538-543.
[9] J. Eckstein, Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming, Mathematics of Operations Research, 18 (1993), 202-226.
[10] M. Benning and E. S. Riis, Bregman methods for large-scale optimisation with applications in imaging, andbook of Mathematical Models and Algorithms in Computer Vision and Imaging: Mathematical Imaging and Vision, pp. 1–42, 2021.
[11] H. H. Bauschke, J. Bolte, and M. Teboulle, A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications, Mathematics of Operations Research, 42 (2017), 330-348.
[12] Q. Van Nguyen, Forward-backward splitting with Bregman distances, Vietnam Journal of Mathematics, 45 (2017), 519-539.
[13] A. Ben-Tal, T. Margalit, and A. Nemirovski, The ordered subsets mirror descent optimization method with applications to tomography, SIAM Journal on Optimization, 12 (2001), 79-108.
[14] A. Beck and M. Teboulle, Mirror descent and nonlinear projected subgradient methods for convex optimization, Operations Research Letters, 31 (2003), 167-175.
[15] S. Setzer, G. Steidl, and T. Teuber, Deblurring Poissonian images by split Bregman techniques, Journal of Visual Communication and Image Representation, 21 (2010), 193-199.
[16] P. Tan, F. Pierre, and M. Nikolova, Inertial alternating generalized forward–backward splitting for image colorization, Journal of Mathematical Imaging and Vision, 61 (2019), 672-690.
[17] M. Kahl, S. Petra, C. Schnörr, G. Steidl, and M. Zisler, On the remarkable efficiency of SMART, In: International Conference on Scale Space and Variational Methods in Computer Vision (SSVM), pp. 418-430, Springer, 2023.

[18] S. Hurault, U. Kamilov, A. Leclaire, and N. Papadakis, Convergent Bregman plug-and-play image restoration for Poisson inverse problems, In: Advances in Neural Information Processing Systems (NeurIPS), vol. 36, 2024.
[19] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, Robust stochastic approximation approach to stochastic programming, SIAM Journal on Optimization, 19 (2009), 1574-1609.
[20] R. L. Priol, F. Kunstner, D. Scieur, and S. Lacoste-Julien, Convergence rates for the MAP of an exponential family and stochastic mirror descent–an open problem, arXiv preprint arXiv:2111.06826, 2021.
[21] F. Kunstner, R. Kumar, and M. Schmidt, Homeomorphic-invariance of EM: Non-asymptotic convergence in KL divergence for exponential families via mirror descent, In: International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 3295-3303, PMLR, 2021.
[22] K. Ahn and S. Chewi, Efficient constrained sampling via the mirror-Langevin algorithm, Advances in Neural Information Processing Systems (NeurIPS), 34 (2021), 28405–28418.
[23] M. Teboulle, A simplified view of first order methods for optimization, Mathematical Programming, 170 (2018), 67-96.
[24] M. N. Bùi and P. L. Combettes, Bregman forward-backward operator splitting, Set-Valued and Variational Analysis, 29 (2021), 583–603.
[25] P. Ochs, J. Fadili, and T. Brox, Non-smooth non-convex Bregman minimization: Unification and new algorithms, Journal of Optimization Theory and Applications, 181 (2019), 244-278.
[26] C. Rossignol, F. Sureau, E. Chouzenoux, C. Comtat, and J.-C. Pesquet, A Bregman majorization-minimization framework for PET image reconstruction, In: 2022 IEEE International Conference on Image Processing (ICIP), pp. 1736–1740, IEEE, 2022.
[27] L. M. Bregman, The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming, USSR Computational Mathematics and Mathematical Physics, 7 (1967), 200-217.
[28] H. H. Bauschke and J. M. Borwein, Legendre functions and the method of random Bregman projections, Journal of Convex Analysis, 4 (1997), 27-67.
[29] H. H. Bauschke and P. L. Combettes, Convex Analysis and Monotone Operator Theory in Hilbert Spaces, 2nd ed., Springer, 2017.
[30] D. Geman and G. Reynolds,Constrained restoration and the recovery of discontinuities, IEEE Transactions on Pattern Analysis & Machine Intelligence, 14 (1992), 367-383.
[31] É. Chouzenoux, A. Jezierska, J.-C. Pesquet, and H. Talbot, A majorize-minimize subspace approach for $\ell_2 - \ell_0$ image regularization, SIAM Journal on Imaging Sciences, 6 (2013), 563-591.
[32] C. M. Verdun, O. Melnyk, F. Krahmer, and P. Jung, Fast, blind, and accurate: Tuning-free sparse regression with global linear convergence, In: The Thirty Seventh Annual Conference on Learning Theory, pp. 3823–3872, PMLR, 2024.
[33] J. D. Lee, Y. Sun, and M. Saunders, Proximal newton-type methods for convex optimization, In: Advances in Neural Information Processing Systems (NeurIPS), vol. 25, 2012.
[34] P. L. Combettes and B. C. Vũ, Variable metric forward–backward splitting with applications to monotone inclusions in duality, Optimization, 63 (2014), 1289-1318.
[35] E. Chouzenoux, J.-C. Pesquet, and A. Repetti, Variable metric forward–backward algorithm for minimizing the sum of a differentiable function and a convex function, Journal of Optimization Theory and Applications, 162 (2014), 107-132.
[36] D. Butnariu, A. N. Iusem, and C. Zalinescu, On uniform convexity, total convexity and convergence of the proximal point and outer Bregman projection algorithm in Banach spaces, Journal of Convex Analysis, 10 (2003), 35-62.
[37] D. M. Blei, A. Y. Ng, and M. I. Jordan, Latent Dirichlet allocation, Journal of Machine Learning Research, 3 (2003), 993-1022.
[38] J. M. P. Nascimento and J. M. Bioucas-Dias, Hyperspectral unmixing based on mixtures of Dirichlet components, IEEE Transactions on Geoscience and Remote Sensing, 50 (2011), 863-878.
[39] S. Pal and C. Heumann, Clustering compositional data using Dirichlet mixture model, PLOS One, 17 (2022), e0268438.

[40] S. Martin, Y. Huang, F. Shakeri, J.-C. Pesquet, and I. Ben Ayed, Transductive zero-shot and few-shot CLIP, In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 28816–28826, 2024.
[41] S. Martin, A. Gagneux, P. Hagemann, and G. Steidl, PnP-Flow: Plug-and-play image restoration with flow matching, arXiv preprint arXiv:2410.02423, 2024.
[42] H. Stark, B. Jing, C. Wang, G. Corso, B. Berger, R. Barzilay, and T. Jaakkola, Dirichlet flow matching with applications to dna sequence design, arXiv preprint arXiv:2402.05841, 2024.
[43] M. Abramowitz and I. A. Stegun, Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, vol. 55. US Government Printing Office, 1948.
[44] L. D. Brown, Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory. Institute of Mathematical Statistics, 1986.
[45] O. Barndorff-Nielsen, Information and Exponential Families: in Statistical Theory, John Wiley & Sons, 2014.
[46] G. Ronning, Maximum likelihood estimation of Dirichlet distributions, Journal of Statistical Computation and Simulation, 32 (1989), 215-221.
[47] H. Erdogan and J. A. Fessler, Monotonic algorithms for transmission tomography, In: 2002 5th IEEE EMBS International Summer School on Biomedical Imaging, pp. 14, IEEE, 2002.
[48] A. Narayanan, Algorithm AS 266: maximum likelihood estimation of the parameters of the Dirichlet distribution, Applied Statistics, pp. 365–374, 1991.
[49] T. Minka, Estimating a Dirichlet distribution, Tech. Rep., MIT, 2000.