Mathematical Programming (2025) 214:441–490
https://doi.org/10.1007/s10107-024-02189-7

FULL LENGTH PAPER

Series A

Maximum entropy on the mean and the Cramér rate function in statistical estimation and inverse problems: properties, models, and algorithms

Yakov Vaisbourd[1] · Rustum Choksi[1] · Ariel Goodwin[1] · Tim Hoheisel[1] · Carola-Bibiane Schönlieb[2]

Received: 6 June 2023 / Accepted: 21 December 2024 / Published online: 4 January 2025

Abstract

We explore a method of statistical estimation called *Maximum Entropy on the Mean* (MEM) which is based on an information-driven criterion that quantifies the compliance of a given point with a reference prior probability measure. At the core of this approach lies the *MEM function* which is a partial minimization of the Kullback–Leibler divergence over a linear constraint. In many cases, it is known that this function admits a simpler representation (known as the *Cramér rate function*). Via the connection to exponential families of probability distributions, we study general conditions under which this representation holds. We then address how the associated *MEM estimator* gives rise to a wide class of MEM-based regularized linear models for solving inverse problems. Finally, we propose an algorithmic framework to solve these problems efficiently based on the Bregman proximal gradient method, alongside proximal operators for commonly used reference distributions. The article is complemented by a software package for experimentation and exploration of the MEM approach in applications.

Keywords Maximum Entropy on the Mean · Statistical Estimation · Cramér Rate Function · Kullback–Leibler Divergence · Prior Distribution · Regularization · Linear Inverse Problems · Bregman Proximal Gradient · Convex Duality · Large Deviations

Mathematics Subject Classification 49M27 · 29M29 · 60F10 · 62B10 · 62H12 · 90C25 · 90C46

✉ Tim Hoheisel
tim.hoheisel@mcgill.ca

[1] Department of Mathematics and Statistics, McGill University, Montreal, Canada

[2] Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK

1 Introduction

Many models for modern applications in various disciplines are based on some form of *statistical estimation*, for example, the very common *maximum likelihood (ML)* principle. In this study, we consider an alternative approach known as the *maximum entropy on the mean* (MEM). At its core lies the MEM function κ_P induced by some *reference distribution* P and defined as

$$\kappa_P(y) := \inf \left\{ D_{\mathrm{KL}}(Q \,||\, P) : \mathbb{E}_Q = y, \, Q \in \mathcal{P}(\Omega) \right\},$$

where $\mathcal{P}(\Omega)$ stands for the set of probability measures on $\Omega \subseteq \mathbb{R}^d$, $\mathbb{E}_Q$ is the expected value of $Q \in \mathcal{P}(\Omega)$ and $D_{\mathrm{KL}}(Q \,||\, P)$ stands for the Kullback-Leibler (KL) divergence of Q with respect to P [49] (see Sect. 2 for precise definitions). Thus, the MEM modeling paradigm stems from the principle of minimum discrimination information [48] which generalizes the well-known principle of maximum entropy [47]. In the context of information theory [32], the argmin of $\kappa_P(y)$ is often referred to as the *information projection* of P onto the set $\{Q \in \mathcal{P}(\Omega) \,:\, \mathbb{E}_Q = y\}$, the *closest* member of the set to P.

Many forms and interpretations of MEM have been studied (see, for example, [34, 40–42, 44, 50, 51]) and found applications in various disciplines, including earth sciences [39, 53, 54, 57, 65], and medical imaging [1, 23, 27, 43, 46]. A version of the MEM method was recently explored for blind deblurring of images possessing some form of fixed symbology (for example, in barcodes) [58, 59]. There one exploited the ability of of the MEM framework to facilitate the incorporation of nonlinear constraints via the introduction of a prior distribution.

Despite its many interesting properties in both theory and applications, the MEM methodology has yet to find its place as a mainstream tool for statistical estimation, particularly as it pertains to solving inverse problems. One factor that might have contributed to this centers on the practical issue that there are no dedicated optimization algorithms designed to tackle models based on the MEM methodology. Indeed, the MEM function is defined by means of an infinite-dimensional optimization problem. Previous attempts to solve models involving the MEM function relied on its finite-dimensional dual problem. To the best of the authors' knowledge, there are no dedicated optimization algorithms designed to tackle models based on the MEM methodology. Therefore, any researcher or practitioner wishing to employ the MEM framework must first overcome a notable barrier of deriving an appropriate optimization algorithm for its solution. In this work, our goal is to fill in this gap, providing an access gate to the MEM methodology.

Our approach is based on the fundamental work by Brown [22, Chapter 6] and complements [50] by first proving the equivalence of the MEM function to the *Cramér's rate* function, mostly known from its role in *large deviation theory*. Cramér's rate function is defined by means of a finite-dimensional optimization problem as it is simply the convex conjugate of the log-normalizer (aka the cumulant generating function) of the reference distribution P. In many cases (i.e., choices of P) it admits a closed-form expression while in others it can still be evaluated efficiently. The connection between these seemingly different functions is well established in the large deviations [35], statistics [22], and information theory [50] literature. Nonetheless, various assumptions imposed in the aforementioned works limit the scope of existing results. Employing the framework of exponential families of probability distributions [22], we establish the equivalence between the two functions under very mild and natural conditions, allowing us to cover many distributions of practical interest. Thus, models involving MEM functions can be explicitly stated using the corresponding Cramér functions.

Central to our study is *the MEM estimator* which is shown to be well-defined under very mild conditions. We further recall an insightful connection between the MEM and ML estimators as presented in [22] for the case of a reference distribution from an exponential family. As with the ML counterpart, the MEM estimator has vast applications, and hence we restrict the remainder of the paper to a wide class of regularized linear models for solving inverse problems. Each model in this class involves two MEM functions, one in the role of a fidelity term and another as a regularizer (comparable to the *maximum a posteriori (MAP) estimation* framework which extends ML). Let us provide an example: given a measurement matrix $A \in \mathbb{R}^{m\times d}$, an observation vector $\hat{y} \in \mathbb{R}^m$ and an additional vector $p \in [0, 1]^d$ representing some prior knowledge, the following optimization problem

$$\min\left\{\underbrace{\frac{1}{2}\|Ax-\hat{y}\|_2^2}_{Fidelity}+\underbrace{\sum_{i=1}^{d}\left[x_i\log\left(\frac{x_i}{p_i}\right)+(1-x_i)\log\left(\frac{1-x_i}{1-p_i}\right)\right]}_{Regularization} : x\in[0,1]^d\right\},$$

fits the MEM framework with normal (Gaussian) and Bernoulli reference distributions of the fidelity and regularization terms, respectively. Other choices of reference distributions will lead to additional models that admit a similar additive composite structure. Moreover, the closed-form expressions of the two functions in our example follow from the definition of Cramér's rate function. In models of these forms, concrete expressions and structures with distinct geometry can be exploited to customize appropriate operators that pave the way to the utilization of contemporary optimization strategies. Here we highlight the class of *Bregman proximal gradient* (BPG) methods as an especially suitable choice for this family of models. Nevertheless, other methods are also viable alternatives; for example, adaptive and scaled, accelerated variants and dual decomposition methods which are defined by means of the same operators developed here.

Our overall aim is to provide a self-contained, mathematically sound toolbox for working with the MEM methodology for a wide variety of models. To achieve this goal, we include a rigorous presentation of the method and a comprehensive review of the current literature. Additionally, we make several novel contributions:

(i) We present rigorous proof (under mild and natural assumptions) for the equivalence between the MEM function and the Cramér rate function. Notably, our proof encompasses the scenario where the variable lies on the boundary of the domain, which holds particular significance when dealing with discrete reference distributions.

(ii) We compile an extensive list of Cramér rate functions, which includes distributions that seem to have previously been overlooked in the literature (including multivariate Normal-inverse Gaussian, Laplace, and Logistic distributions).

(iii) We expand upon a modeling paradigm initially introduced in [22] for exponential families, extending it to the general setting. This paradigm involves the MEM estimator, which exhibits an intriguing relationship with the ML estimator when the reference distribution belongs to an exponential family. We demonstrate how the MEM estimator enables the creation of a diverse range of models for addressing linear inverse problems, leveraging a custom-designed regularizer derived from MEM.

(iv) All of the above paves the way to the most important outcome of this work. Building upon the concrete expressions of the Cramér rate functions and their distinct geometry, we introduce novel operators tailored to the suggested models. These operators are vital for the application of contemporary optimization algorithms to solve the problems emerge from the MEM framework. We emphasize the utilization of the Bregman proximal gradient method, which proves to be particularly well-suited for these families of models. Remarkably, despite the wide applicability of the MEM model and the vast literature, to the best of our knowledge, this is the first attempt to design tailored optimization algorithms to tackle the resulting problems. Finally, we provide an extensive software package to complement our findings that includes some numerical illustrations for classical image processing applications.

We believe that this sets the basis for, and hopefully triggers, further experimentation and exploration of the MEM approach in contemporary applications.

The paper is organized as follows. In Sect. 2, we recall some concepts and preliminary results from convex analysis and probability theory which will be used in this work. In Sect. 3, we study the MEM and Cramér rate functions and establish the equivalence between the two under very mild and natural conditions. This allows us to use the accessible definition of the Cramér function and derive tractable expressions for a wide class of possible reference distributions which closes this section (see Table 1). Section 4 is devoted to the MEM models considered in this work, and in Sect. 5, we present the algorithms for solving such models. We end with a few concrete examples of problems and corresponding algorithms crafted from the operators derived in this work. An appendix provides the details of a variety of Cramér rate function computations.

2 Preliminaries

2.1 Convex analysis

We recall here some definitions and results from convex analysis. Further details and proofs can be found in various textbooks such as [11, 13, 60].

The *affine hull* of a set $S \subseteq \mathbb{R}^d$ is the smallest affine subspace containing S. For any point $y \in S$, we have the following relation

$$\operatorname{aff} S = y + \operatorname{span}(S - y), \tag{2.1}$$

where $\operatorname{span} S$ stands for the linear hull of S. The dimension of $\operatorname{aff} S$ is defined as $\dim(\operatorname{aff} S) := \dim(\operatorname{span}(S - y))$. The interior, closure, and boundary of a set are denoted as $\operatorname{int} S$, $\operatorname{cl} S$ and $\operatorname{bd} S$, respectively.

The (Fenchel) conjugate of $\psi : \mathbb{R}^d \to [-\infty, \infty]$ is defined as

$$\psi^*(y) := \sup\{\langle y, x\rangle - \psi(x) : x \in \mathbb{R}^d\}.$$

The function ψ is proper if $\psi(x) > -\infty$ for all $x \in \mathbb{R}^d$ and $\operatorname{dom}\psi := \{x \in \mathbb{R}^d : \psi(x) < \infty\} \neq \emptyset$. In addition, ψ is closed, if its epigraph $\{(x, \alpha) \in \mathbb{R}^d \times \mathbb{R} : \psi(x) \leq \alpha\}$ is a closed set.

If ψ is proper and convex then ψ^* is closed, proper, and convex. For a proper function $\psi : \mathbb{R}^d \to (-\infty, +\infty]$, the *Fenchel-Young inequality* states that $\psi(x) + \psi^*(y) \geq \langle y, x\rangle$. If ψ is proper, closed and convex then we obtain that [13, Theorem 4.20]

$$\psi(x) + \psi^*(y) = \langle y, x\rangle \iff y \in \partial\psi(x) \iff x \in \partial\psi^*(y), \tag{2.2}$$

where $\partial\psi(x) := \{g \in \mathbb{R}^d : \psi(y) \geq \psi(x) + \langle g, y - x\rangle \ (y \in \mathbb{R}^d)\}$ is the *subdifferential* of ψ at $x \in \mathbb{R}^d$.

The *indicator function* of a set $S \subseteq \mathbb{R}^d$ is denoted by δ_S and defined as $\delta_S(x) = 0$ if $x \in S$ and $\delta_S(x) = +\infty$ otherwise. Its convex conjugate is known as the *support function* $\sigma_S(y) := \delta_S^*(y) = \sup\{\langle y, x\rangle : x \in S\}$.

Definition 2.1 *(Essential smoothness and Legendre type)* Let $\psi : \mathbb{R}^d \to (-\infty, +\infty]$ be proper and convex. Then, ψ is called *essentially smooth* if it satisfies the following conditions:

1. $\operatorname{int}(\operatorname{dom}\psi) \neq \emptyset$;
2. ψ is differentiable on $\operatorname{int}(\operatorname{dom}\psi)$;
3. $\|\nabla\psi(x^k)\| \to \infty$ for any sequence $\{x^k \in \operatorname{int}(\operatorname{dom}\psi)\}_{k\in\mathbb{N}} \to \bar{x} \in \operatorname{bd}(\operatorname{dom}\psi)$.

The last condition listed above is called *steepness*. An essentially smooth function ψ is said to be of *Legendre type* if it is strictly convex on $\operatorname{int}(\operatorname{dom}\psi)$.

For $\psi : \mathbb{R}^d \to (-\infty, +\infty]$ closed and of Legendre type, the following hold [60, Theorem 26.5]:

1. ψ^* is of Legendre type.
2. $\nabla\psi : \operatorname{int}(\operatorname{dom}\psi) \to \operatorname{int}(\operatorname{dom}\psi^*)$ is a bijection with $(\nabla\psi)^{-1} = \nabla\psi^*$.

The *Bregman distance* induced by a function ψ of Legendre type is defined as [21]

$$D_\psi(y, x) = \psi(y) - \psi(x) - \langle\nabla\psi(x), y - x\rangle \qquad (x \in \operatorname{int}(\operatorname{dom}\psi), y \in \operatorname{dom}\psi).$$

For any $(x, y) \in \operatorname{int}(\operatorname{dom}\psi) \times \operatorname{dom}\psi$, the Bregman distance is nonnegative $D_\psi(y, x) \geq 0$, and equality holds if and only if $x = y$ due to strict convexity of ψ [21]. However, in general, D_ψ is not symmetric, unless $\psi = (1/2)\|\cdot\|^2$ [9, Lemma 3.16]. The Bregman distance induced by a function ψ of Legendre type satisfies the following additional properties [10, Theorem 3.7]: For any $x, y \in \operatorname{int}(\operatorname{dom}\psi)$ it holds that

$$D_\psi(y, x) = D_{\psi^*}(\nabla\psi(x), \nabla\psi(y)). \tag{2.3}$$

The Bregman distance is strictly convex with respect to its first argument. Moreover, for two functions ψ_1 and ψ_2 differentiable at $x \in \operatorname{int}(\operatorname{dom}\psi_1) \cap \operatorname{int}(\operatorname{dom}\psi_2)$

$$D_{\alpha\psi_1+\beta\psi_2}(y, x) = \alpha D_{\psi_1}(y, x) + \beta D_{\psi_2}(y, x) \quad (y \in \operatorname{dom}\psi_1 \cap \operatorname{dom}\psi_2,\ \alpha, \beta \in \mathbb{R}). \tag{2.4}$$

2.2 Probability theory and exponential families

We recall some concepts from probability theory with an emphasis on exponential families. For further detail, see e.g. [5, 22].

For $\Omega \subseteq \mathbb{R}^d$, let $\mathcal{M}(\Omega)$ denote the σ-finite Borel measures on $\mathbb{R}^d$ whose *support*[1] is contained in Ω. We denote by $\Omega_\rho^{cc} := \operatorname{cl}(\operatorname{conv}\Omega_\rho)$ the closure of the convex hull of the support Ω_ρ, which is known as the *convex support* of ρ.

Let $\mathcal{P}(\Omega)$ denote the probability measures in $\mathcal{M}(\Omega)$. For $P \in \mathcal{P}(\Omega)$ and $\nu \in \mathcal{M}(\Omega)$ with $P \ll \nu$ (i.e., P is absolutely continuous with respect to ν), the Radon-Nikodym derivative[2] is

$$f_P := \frac{dP}{d\nu}.$$

Throughout we restrict ourselves to exactly two scenarios: 1.) $\Omega = \mathbb{R}^d$ and ν is the Lebesgue measure (hence $\nu \in \mathcal{M}(\Omega)$). 2.) Ω is countable (thus Borel) and $\nu \in \mathcal{M}(\Omega)$ is such that $\Omega_\nu = \Omega$.

In both cases, we will refer to f_P as the (ν-)density of P, and we will refer to $P \in \mathcal{P}(\Omega)$ interchangeably as a distribution or a measure.

The *expected value* (if it exists) and *moment generating function* of $P \in \mathcal{P}(\Omega)$ are given by

$$\mathbb{E}_P := \int_\Omega y dP(y) \in \mathbb{R}^d \qquad \text{and} \qquad M_P[\theta] := \int_\Omega \exp(\langle\cdot, \theta\rangle) dP,$$

[1] The *support* Ω_ρ of a Borel measure ρ is the smallest closed (hence Borel) set $A \subset \mathbb{R}^d$ such that $\rho(\mathbb{R}^d \setminus A) = 0$.

[2] A measure ρ is said to be *absolutely continuous* with respect to a measure ν, denoted $\rho \ll \nu$, if for any Borel subset $A \subseteq \mathbb{R}^d$, $\nu(A) = 0$ implies $\rho(A) = 0$. In this case, there exists a unique function $h = \frac{d\rho}{d\nu}$, called the *Radon-Nikodym derivative*, such that for any Borel subset $A \subseteq \mathbb{R}^d$, $\rho(A) = \int_A h d\nu$.

respectively. Given $P \in \mathcal{P}(\Omega)$ let

$$\Theta_P := \left\{ \theta \in \mathbb{R}^d : \int_\Omega \exp(\langle \cdot, \theta \rangle) dP < \infty \right\},$$

and consider the function $\psi_P : \mathbb{R}^d \to (-\infty, +\infty]$ given by

$$\psi_P(\theta) := \begin{cases} \log \int_\Omega \exp\left(\langle \cdot, \theta \rangle\right) dP, & \theta \in \Theta_P, \\ +\infty, & \theta \notin \Theta_P. \end{cases} \tag{2.5}$$

Then $\mathcal{F}_P := \left\{ f_{P_\theta}(y) := \exp\left(\langle y, \theta \rangle - \psi_P(\theta)\right) : \theta \in \Theta_P \right\}$, is a *standard exponential family* generated by P. Note that, the probability measure P_θ satisfying $dP_\theta = f_{P_\theta} dP$ is, by construction, a probability measure such that P_θ and P are mutually absolutely continuous, hence $\Omega_{P_\theta} = \Omega_P$ for all $\theta \in \Theta_P$ [5, Section 8.1]. The function ψ_P is called the *log-normalizer* (also known as the *log-partition* or *log-Laplace transform* of P). The vector $\theta \in \mathbb{R}^d$ is known as the *natural parameter* and the set $\Theta_P = \operatorname{dom} \psi_P$ is called the *natural parameter space*,[3]

The following results summarize some well-known properties of the log-normalizer ψ_P.

Proposition 2.1 (Convexity, [22, Theorem 1.13]) *Let $\mathcal{F}_P$ be an exponential family generated by $P \in \mathcal{M}(\Omega)$. Then, the natural parameter space Θ_P is a convex set, and the log-normalizer function $\psi_P : \mathbb{R}^d \to (-\infty, +\infty]$ is closed, proper, and convex.*

Proposition 2.2 (Differentiability, [22, Theorem 2.2, Corollary 2.3]) *Let $\mathcal{F}_P$ be an exponential family generated by $P \in \mathcal{M}(\Omega)$ and let $\theta \in \operatorname{int} \Theta_P$. Then, the log-normalizer $\psi_P : \mathbb{R}^d \to (-\infty, +\infty]$ is infinitely differentiable at θ and it holds that $\nabla \psi_P(\theta) = \mathbb{E}_{P_\theta}$.*

The dimension of a convex set $S \subseteq \mathbb{R}^d$, denoted by $\dim S$, is equal to the affine dimension of $\operatorname{aff} S$. We assume that the exponential family generated by $P \in \mathcal{M}(\Omega)$ is *minimal*, i.e., $\dim \Theta_P = \dim \Omega_P^{cc} = d$ or, equivalently, $\operatorname{int} \Theta_P \neq \emptyset$ and $\operatorname{int} \Omega_P^{cc} \neq \emptyset$. This is not restrictive as a non-minimal exponential family can be always reduced to a minimal form [22, Theorem 1.9]. The following result strengthens Proposition 2.1 for minimal exponential families.

Proposition 2.3 (Strict convexity, [22, Theorem 1.13]) *Let $\mathcal{F}_P$ be a minimal exponential family generated by $P \in \mathcal{M}(\Omega)$. Then, the log-normalizer function $\psi_P : \mathbb{R}^d \to (-\infty, +\infty]$ is strictly convex over Θ_P.*

If the log-normalizer ψ_P is essentially smooth (or 'steep' in the exponential family terminology, see, e.g., [5, Theorem 5.27] and [22, Definition 3.2]), we say that the exponential family $\mathcal{F}_P$ is *steep*. This condition is automatically satisfied when Θ_P is

[3] It is possible to define the exponential family $\mathcal{F}_P$ over a subset of the natural parameter space [22, Definition 1.1] but this is not needed for our study.

open [5, Theorem 8.2]. While most exponential families encountered in practice have this property, there are relevant cases when this assumption is too restrictive (e.g., [22, Example 3.4]). Thus, in order to cover all examples provided in this work, we will assume that the exponential family is steep. Summarizing the above discussion and recalling Definition 2.1 we have the following corollary.

Corollary 2.1 *Let $\mathcal{F}_P$ be a minimal and steep exponential family generated by $P \in \mathcal{M}(\Omega)$. Then, the log-normalizer function ψ_P is of Legendre type.*

From the last corollary we can see that $\nabla\psi_P$ forms a bijection between $\operatorname{int}(\operatorname{dom}\psi_P) = \operatorname{int}\Theta_P$ and $\operatorname{int}(\operatorname{dom}\psi_P^*)$. This relation provides a dual representation of the log-normalizer ψ_P and, consequently, the distribution in question. The so-called *mean value parametrization* is obtained by applying a change of variables where the natural parameter θ is replaced by $\mu \in \mathbb{R}^d$ such that $\mu = \mathbb{E}_{P_\theta} = \nabla\psi_P(\theta)$, i.e., $\theta = \nabla\psi_P^*(\mu)$.

The *Kullback–Leibler (KL) divergence* (also known as the relative entropy) of a probability measure $Q \in \mathcal{P}(\Omega)$ with respect to $P \in \mathcal{P}(\Omega)$ is given by (see [49])

$$D_{\mathrm{KL}}(Q \,||\, P) := \begin{cases} \displaystyle\int_\Omega \log\left(\frac{dQ}{dP}\right) dQ, & Q \ll P, \\ +\infty, & \text{otherwise.} \end{cases}$$

It holds that $D_{\mathrm{KL}}(Q \,||\, P) \geq 0$ with equality if and only if $Q = P$ [49, Lemma 3.1]. Thus, the Kullback-Leibler information quantifies the dissimilarity between two probability measures. We note that, in general, $D_{\mathrm{KL}}(Q \,||\, P)$ is not symmetric. Furthermore, $D_{\mathrm{KL}}(Q \,||\, P)$ is jointly convex in $(Q|P)$. We record a special case for which the KL divergence is of particular interest.

Remark 2.1 *(Kullback–Leibler divergence for exponential family)* Let $\mathcal{F}_P$ be an exponential family generated by $P \in \mathcal{M}(\Omega)$. Let $\theta_1 \in \Theta_P$ and $\theta_2 \in \operatorname{int}\Theta_P$, thus for $i = 1, 2$ we have that $f_{P_{\theta_i}} \in \mathcal{F}_P$. In this case, the KL divergence between the two measures $P_{\theta_i} \in \mathcal{P}(\Omega)$ such that $dP_{\theta_i} := f_{P_{\theta_i}} dP$ $(i = 1, 2)$ satisfies $D_{\mathrm{KL}}(P_{\theta_2} \,||\, P_{\theta_1}) = D_{\psi_P}(\theta_1, \theta_2)$ [22, Proposition 6.3]. $\Diamond$

3 Maximum entropy on the mean and Cramér's rate function

For $y \in \mathbb{R}^d$, the density

$$f_P(y) := \frac{dP}{d\nu}(y) \tag{3.1}$$

provides an indication of the likelihood of y under the distribution $P \in \mathcal{P}(\Omega)$. The method of *Maximum Entropy on the Mean* (MEM) suggests an alternative, information-driven function $\kappa_P : \mathbb{R}^d \to (-\infty, +\infty]$ given by

$$\kappa_P(y) := \inf\left\{D_{\mathrm{KL}}(Q \,||\, P) : \mathbb{E}_Q = y,\, Q \in \mathcal{P}(\Omega)\right\}. \tag{3.2}$$

Here, κ_P measures how y complies with the distribution P, by seeking a distribution Q with expected value y that minimizes $D_{\mathrm{KL}}(\cdot \mid\mid P)$. The distance, in terms of the KL divergence (the information gain) between the resulting and the original distributions, quantifies the compliance of y with P. We will refer to κ_P as the *MEM function* and to P as the *reference distribution*. Since $D_{\mathrm{KL}}(Q \mid\mid P) \geq 0$ and $D_{\mathrm{KL}}(Q \mid\mid P) = 0$ if and only if $Q = P$, we find that the MEM function satisfies $\kappa_P(y) \geq 0$ for any $y \in \mathbb{R}^d$ and $\kappa_P(y) = 0$ if and only if $y = \mathbb{E}_P$.

In most cases of interest, the MEM function admits an alternative representation that sheds light on many of its additional properties (cf. Theorem 3.2). More precisely, under suitable conditions (cf. Theorem 3.1), the MEM function coincides with the *Cramér rate function* [33], to which we turn now. For a given reference distribution $P \in \mathcal{P}(\Omega)$, recall the log-normalizer previously defined for a general measure in (2.5):

$$\psi_P(\theta) := \log M_P[\theta] = \log \int_\Omega \exp\left(\langle \cdot, \theta \rangle\right) dP.$$

In the context of probability measures P, ψ_P is often known as the *cumulant generating function*. The *Cramér rate function* ψ_P^* associated with P is the conjugate of ψ_P, that is,

$$\psi_P^*(y) = \sup\{\langle y, \theta \rangle - \psi_P(\theta) : \theta \in \mathbb{R}^d\}.$$

Our central assumption (which is not too restrictive in view of our discussion above) on the prior P and its exponential family $\mathcal{F}_P$ is provided below. The additional condition $0 \in \operatorname{int} \Theta_P$ ensures the existence of $\mathbb{E}_P$.

Assumption A The reference distribution $P \in \mathcal{P}(\Omega)$ generates a minimal and steep exponential family $\mathcal{F}_P$ such that $0 \in \operatorname{int} \Theta_P$.

The equivalence between the two seemingly different functions[4] ψ_P^* and κ_P was previously established under various assumptions: the authors of [35, Theorem 5.2] (see also [38]) impose the (restrictive) assumption that ψ_P is finite. On the other hand, the results in [22, Theorem 6.17] and [50, Proposition 1] (see also [15] and a closely related result in [67, Theorem 3.4]) do not address the challenging case when y resides on the boundary of the domain. This scenario turns out to be important if (and only if) the reference distribution is defined over a countable set. Here, we provide complete proof that overcomes these assumptions previously imposed. Our approach emphasizes the role played by the convex support of the reference distribution and leads to natural and easy-to-verify conditions.

The main result of this section is summarized in the following theorem.

Theorem 3.1 (Equivalence between Cramér's rate function and the MEM function) *Let $P \in \mathcal{P}(\Omega)$ satisfy Assumption A, and assume that one of the following two conditions holds:*

[4] ψ_P^* appears in *Cramér's Theorem* central in large deviations theory [38]. A more general form of κ_P appears in *Sanov's Theorem*.

(i) *Ω_P is uncountable.*
(ii) *Ω_P is countable and* $\operatorname{conv}\Omega_P$ *is closed (as is the case when Ω_P is finite).*

Then, $\kappa_P = \psi_P^*$. In particular, κ_P is closed, proper, and convex.

To prove Theorem 3.1, we will first need to examine the domains $\operatorname{dom}\kappa_P$ and $\operatorname{dom}\psi_P^*$. For Cramér's rate function ψ_P^*, a characterization of the domain is summarized in the following proposition.

Proposition 3.1 (Domain of the Cramér rate function ψ_P^* [5, Theorems 9.1, 9.4 and 9.5]) *Let $P \in \mathcal{P}(\Omega)$ be a reference distribution satisfying Assumption A. Then,* $\operatorname{int}\Omega_P^{cc} \subseteq \operatorname{dom}\psi_P^* \subseteq \Omega_P^{cc}$. *Moreover, the following hold:*

(a) *If Ω_P is finite, then* $\operatorname{dom}\psi_P^* = \Omega_P^{cc}$.
(b) *If Ω_P is countable, then* $\operatorname{dom}\psi_P^* \supseteq \operatorname{conv}\Omega_P$.
(c) *If Ω_P is uncountable, then* $\operatorname{dom}\psi_P^* = \operatorname{int}\Omega_P^{cc}$.

In order to establish a similar characterization for the domain of the MEM function, we will need to make precise the relation between Ω_P and the expected value $\mathbb{E}_P$ for a given probability measure $P \in \mathcal{P}(\Omega)$. To this end, we first recall some additional definitions and results (see, for example, [60, Section 6]). Consider two subsets $S, \hat{S} \subseteq \mathbb{R}^d$ and assume further that $S \subseteq \hat{S}$. Then $\operatorname{cl} S \subseteq \operatorname{cl}\hat{S}$, $\operatorname{int} S \subseteq \operatorname{int}\hat{S}$ and $\operatorname{conv} S \subseteq \operatorname{conv}\hat{S}$.

Denote the closed Euclidean unit ball in $\mathbb{R}^d$ by $\mathcal{B}_d$. The *relative interior* [60, Section 6] of a convex set $S \subseteq \mathbb{R}^d$ is defined as

$$\operatorname{ri} S := \left\{ x \in \mathbb{R}^d : \exists \tau > 0 \text{ such that } (x + \tau\mathcal{B}_d) \cap \operatorname{aff} S \subseteq S \right\}.$$

E.g., for the *unit simplex* $\Delta_d := \{y \in \mathbb{R}_+^d : \langle e, y\rangle = 1\}$ we have $\operatorname{ri}\Delta_d := \{y \in \mathbb{R}_{++}^d : \langle e, y\rangle = 1\}$. Some facts which will be used in the sequel are summarized in the following lemma. Further details and proofs can be found in [60, Section 6, Theorem 13.1].

Lemma 3.1 (On the relative interior) *Let $S \subseteq \mathbb{R}^d$ be nonempty and convex. Then:*

(a) *It holds that* $\operatorname{ri}(\operatorname{cl} S) = \operatorname{ri} S$ *and* $\operatorname{ri} S \subseteq S \subseteq \operatorname{cl} S$.
(b) *If* $\dim S = d$ *then* $\operatorname{ri} S = \operatorname{int} S$ *and, in particular,* $\operatorname{int} S \neq \emptyset$.
(c) *It holds that $x \in \operatorname{ri} S$ if and only if $\sigma_{S-x}(v) \geq 0$ where the last inequality is strict for every $v \in \mathbb{R}^d$ such that $-\sigma_S(-v) \neq \sigma_S(v)$.*

Lemma 3.2 (Domain of expected value) *Let $P \in \mathcal{P}(\Omega)$ and assume that $\mathbb{E}_P$ exists. Then* $\mathbb{E}_P \in \operatorname{ri}\Omega_P^{cc} = \operatorname{ri}(\operatorname{conv}\Omega_P)$.

Proof By definition of σ_{Ω_P}, for any $v \in \mathbb{R}^d$, it holds that $-\sigma_{\Omega_P}(-v) \leq \langle v, y\rangle \leq \sigma_{\Omega_P}(v)$. As $P \in \mathcal{P}(\Omega)$, this implies, for all $v \in \mathbb{R}^d$, that

$$\langle v, \mathbb{E}_P\rangle = \int_{\Omega_P} \langle v, y\rangle dP(y) \leq \sigma_{\Omega_P}(v) \int_{\Omega_P} dP(y) = \sigma_{\Omega_P}(v). \tag{3.3}$$

If there exists some subset $A \subseteq \Omega_P$ such that $P(\{y \in A : \langle v, y\rangle < \sigma_{\Omega_P}(v)\}) > 0$, then the inequality in (3.3) is strict. We will show that, for any $v \in \mathbb{R}^d$ such that $-\sigma_{\Omega_P}(-v) \neq \sigma_{\Omega_P}(v)$, such a subset exists; the desired result then follows from Lemma 3.1 (c) and the equivalence $\sigma_{\Omega_P^{cc}}(v) = \sigma_{\Omega_P}(v)$ [64, Theorem 8.24]. Indeed, let $v \in \mathbb{R}^d$ such that $-\sigma_{\Omega_P}(-v) \neq \sigma_{\Omega_P}(v)$, i.e. $-\sigma_{\Omega_P}(-v) < \sigma_{\Omega_P}(v)$. Pick $\tau \in (-\sigma_{\Omega_P}(-v), \sigma_{\Omega_P}(v))$ and consider $A = \{y \in \Omega_P : \langle v, y\rangle \leq \tau\}$. As $\tau < \sigma_{\Omega_P}(v)$, we have $A \subset \{y \in \Omega_P : \langle v, y\rangle < \sigma_{\Omega_P}(v)\}$, and

$$\begin{aligned} P(A) &= P(\{y \in \Omega_P : \langle -v, y\rangle \geq -\tau\}) \\ &= P(\{y \in \Omega_P : \sigma_{\Omega_P}(-v) \geq \langle -v, y\rangle \geq -\tau\}) > 0, \end{aligned}$$

where the strict inequality follows from the definition of $\sigma_{\Omega_P}(-v)$ and $\sigma_{\Omega_P}(-v) > -\tau$. Hence, A satisfies the desired conditions, which establishes the result. □

We are now in a position to present and prove a characterization for the domain of the MEM function, analogous to Proposition 3.1. We will use the following notation

$$\mathcal{Q}_P(y) := \{Q \in \mathcal{P}(\Omega) : \mathbb{E}_Q = y,\ Q \ll P\}.$$

Observe that $y \in \operatorname{dom} \kappa_P$ if and only if $\mathcal{Q}_P(y) \neq \emptyset$.

Lemma 3.3 (Domain of the MEM function κ_P) *Let $P \in \mathcal{P}(\Omega)$ be a reference distribution satisfying Assumption A. Then:*

(a) If Ω_P is countable, then $\operatorname{dom} \kappa_P = \operatorname{conv} \Omega_P$. Hence, if Ω_P is finite, then $\operatorname{dom} \kappa_P = \Omega_P^{cc}$.
(b) If Ω_P is uncountable, then $\operatorname{dom} \kappa_P = \operatorname{int} \Omega_P^{cc}$.

Proof (a) Let $y \in \operatorname{dom} \kappa_P$, hence there exists $Q \in \mathcal{Q}_P(y)$. As $Q \ll P$, we obtain $\Omega_Q \subseteq \Omega_P$, thus $\operatorname{conv} \Omega_Q \subseteq \operatorname{conv} \Omega_P$. Hence, by Lemma 3.1 (a) and Lemma 3.2, we know that $y = \mathbb{E}_Q \in \operatorname{ri} \Omega_Q^{cc} \subseteq \operatorname{conv} \Omega_Q \subseteq \operatorname{conv} \Omega_P$. Thus, $\operatorname{dom} \kappa_P \subseteq \operatorname{conv} \Omega_P$. For the converse inclusion, let $y \in \operatorname{conv} \Omega_P$. By Carathéodory's theorem [24], there exist $n \leq d + 1$ points $p_1, \ldots, p_n$ in Ω_P such that $y = \sum_{i=1}^n \lambda_i p_i$ for some $\lambda \in \Delta_n$. Consider a distribution $Q \in \mathcal{P}(\Omega)$ satisfying $Q(\{p_i\}) = \lambda_i$ for all $i = 1, \ldots, n$. Then, $Q \in \mathcal{Q}_P(y)$ by construction. Thus, $y \in \operatorname{dom} \kappa_P$, and we can conclude that $\operatorname{conv} \Omega_P \subseteq \operatorname{dom} \kappa_P$.

(b) First, let $y \in \operatorname{dom} \kappa_P$, then there exists $Q \in \mathcal{Q}_P(y)$. Since $Q \ll P$ which satisfies Assumption A, it holds that $\dim \Omega_Q^{cc} = \Omega_P^{cc} = d$. Otherwise, the probability measure Q ($Q(\Omega_Q) = 1$) is concentrated on a lower dimensional affine subspace in contradiction to the absolute continuity of Q with respect to P. Hence, using Lemma 3.2 and Lemma 3.1 (b), we obtain that $y = \mathbb{E}_Q \in \operatorname{ri} \Omega_Q^{cc} = \operatorname{int} \Omega_Q^{cc} \subseteq \operatorname{int} \Omega_P^{cc}$. For the converse inclusion, by Proposition 3.1, $y \in \operatorname{int} \Omega_P^{cc} = \operatorname{dom} \psi_P^* = \operatorname{int} (\operatorname{dom} \psi_P^*) = \operatorname{dom} \nabla \psi_P^*$, and we conclude that $y = \mathbb{E}_{P_\theta}$ for $\theta = \nabla \psi_P^*(y)$. Since $P_\theta \ll P$ for P_θ from the exponential family generated by P, we find that $P_\theta \in \mathcal{Q}_P(y)$ and therefore $y \in \operatorname{dom} \kappa_P$. □

Combining Lemma 3.3 with Proposition 3.1 yields the following corollary.

Corollary 3.1 *Let $P \in \mathcal{P}(\Omega)$ be a reference distribution satisfying Assumption A. Then,*

(a) If Ω_P is countable and $\operatorname{conv}\Omega_P$ is closed (i.e., $\operatorname{conv}\Omega_P = \Omega_P^{cc}$), then $\operatorname{dom}\kappa_P = \operatorname{dom}\psi_P^* = \Omega_P^{cc}$. In particular, $\operatorname{dom}\kappa_P = \operatorname{dom}\psi_P^* = \Omega_P^{cc}$ if Ω_P is finite.
(b) If Ω_P is uncountable, then $\operatorname{dom}\kappa_P = \operatorname{dom}\psi_P^* = \operatorname{int}\Omega_P^{cc}$.

The following lemma will be crucial for proving the equivalence between the MEM function κ_P and Cramér's rate function ψ_P^*. The proof of the lower bound follows similar arguments as in [22, Theorem 6.17] and [50, Proposition 1] and we include it here for completeness.

Lemma 3.4 *Let $P \in \mathcal{P}(\Omega)$ be a reference distribution satisfying Assumption A. Then:*

$$\psi_P^*(y) \le \kappa_P(y) \le \psi_P^*(y) + D_{KL}(Q \,||\, P_\theta) - D_{\psi_P^*}(y, \nabla\psi_P(\theta)),$$

for any $y \in \operatorname{dom}\kappa_P$, $Q \in \mathcal{Q}_P(y)$ and $\theta \in \operatorname{int}\Theta_P$.

Proof For any $\theta \in \operatorname{int}\Theta_P$ and $Q \in \mathcal{Q}_P(y)$ we obtain that $Q \ll P_\theta$ due to the mutual absolute continuity between P_θ and P. Hence,

$$\begin{aligned} D_{\mathrm{KL}}(Q \,||\, P) &= \int_\Omega \log\left(\frac{dQ}{dP}\right) dQ = \int_\Omega \log\left(\frac{dQ}{dP_\theta}\right) dQ + \int_\Omega \log\left(\frac{dP_\theta}{dP}\right) dQ \\ &= D_{\mathrm{KL}}(Q \,||\, P_\theta) + \int_\Omega [\langle z, \theta\rangle - \psi_P(\theta)] dQ(z) = D_{\mathrm{KL}}(Q \,||\, P_\theta) + \langle y, \theta\rangle - \psi_P(\theta), \end{aligned} \tag{3.4}$$

where the last identity uses $y = \mathbb{E}_Q$. Since (3.4) holds for all $\theta \in \operatorname{int}\Theta_P$ and $D_{\mathrm{KL}}(Q \,||\, P_\theta) \ge 0$,

$$D_{\mathrm{KL}}(Q \,||\, P) \ge \sup\{\langle y, \theta\rangle - \psi_P(\theta) : \theta \in \operatorname{int}\Theta_P\} = \psi_P^*(y), \tag{3.5}$$

due to the closedness of ψ_P, see Proposition 2.1. The lower bound for κ_P follows immediately from its definition and the above inequality.

As for the upper bound: by (3.4) and (2.2), for any $Q \in \mathcal{Q}_P(y)$ and $\theta \in \operatorname{int}\Theta_P$, we have

$$\begin{aligned} D_{\mathrm{KL}}(Q \,||\, P) &= D_{\mathrm{KL}}(Q \,||\, P_\theta) + \langle y, \theta\rangle - \psi_P(\theta) \\ &= D_{\mathrm{KL}}(Q \,||\, P_\theta) + \langle y - \nabla\psi_P(\theta), \theta\rangle + \langle \nabla\psi_P(\theta), \theta\rangle - \psi_P(\theta) \\ &= D_{\mathrm{KL}}(Q \,||\, P_\theta) - \left[\psi_P^*(y) - \psi_P^*(\nabla\psi_P(\theta)) - \langle y - \nabla\psi_P(\theta), \theta\rangle\right] + \psi_P^*(y) \\ &= D_{\mathrm{KL}}(Q \,||\, P_\theta) - D_{\psi_P^*}(y, \nabla\psi_P(\theta)) + \psi_P^*(y). \end{aligned}$$

Then the result follows due to the fact that $\kappa_P(y) \le D_{\mathrm{KL}}(Q \,||\, P)$ for all $Q \in \mathcal{Q}_P(y)$. □

We are now in the position to prove the main result of this section.

Proof of Theorem 3.1 First, let $y \in \operatorname{int} \Omega_P^{cc}$. By Assumption A, $\nabla\psi_P$ is a bijection between $\operatorname{int}(\operatorname{dom}\psi_P) = \operatorname{int}\Theta_P$ and $\operatorname{int}(\operatorname{dom}\psi_P^*) = \operatorname{int}\Omega_P^{cc}$, where the latter uses Proposition 3.1. Thus, there exists $\theta \in \operatorname{int}\Theta_P$ such that $y = \nabla\psi_P(\theta) = \mathbb{E}_{P_\theta}$. Applying Lemma 3.4 with $Q = P_\theta$ yields

$$\kappa_P(y) = \psi_P^*(y) \qquad (y \in \operatorname{int}\Omega_P^{cc}). \tag{3.6}$$

Due to Corollary 3.1, this establishes the result when Ω_P is uncountable. To complete the proof, we only need to address the case when $y \in \operatorname{bd}\Omega_P^{cc}$ under assumption (ii). By Corollary 3.1, in this case $\operatorname{dom}\kappa_P = \operatorname{dom}\psi_P^* = \Omega_P^{cc}$ and $\mathcal{Q}_P(y) \neq \emptyset$ for $y \in \operatorname{bd}\Omega_P^{cc}$. Consider any $Q \in \mathcal{Q}_P(y)$, then, by definition of κ_P, we have that

$$\kappa_P(y) \le D_{\mathrm{KL}}(Q \,||\, P) < +\infty. \tag{3.7}$$

Choose any $\hat{y} \in \operatorname{int}\Omega_P^{cc}$ and set $\hat{\theta} = \nabla\psi_P^*(\hat{y})$ (i.e., $\hat{y} = \nabla\psi(\hat{\theta})$). For any $\lambda \in [0, 1)$ consider $Q_\lambda = \lambda Q + (1-\lambda)P_{\hat{\theta}}$. Then, by linearity of $Q \mapsto \mathbb{E}_Q$ [58, Lemma 2], we obtain

$$y_\lambda := \mathbb{E}_{Q_\lambda} = \lambda\mathbb{E}_Q + (1-\lambda)\mathbb{E}_{P_{\hat{\theta}}} = \lambda y + (1-\lambda)\hat{y}.$$

By convexity of Ω_P^{cc} and the line segment principle [12, Lemma 6.28] we conclude that $y_\lambda \in \operatorname{int}\Omega_P^{cc}$. Set $\theta_\lambda := \nabla\psi_P^*(y_\lambda)$ and observe that, by Lemma 3.4 and the nonnegativity of the Bregman distance, it holds that

$$\psi_P^*(y) \le \kappa_P(y) \le \psi_P^*(y) + D_{\mathrm{KL}}(Q \,||\, Q_\lambda). \tag{3.8}$$

In addition, due to (3.7) and the fact that $Q \ll P \ll P_{\hat{\theta}}$, we conclude that $D_{\mathrm{KL}}(Q \,||\, P_{\hat{\theta}}) < \infty$. Thus, by (3.8) and convexity of $D_{\mathrm{KL}}(Q \,||\, \cdot)$, we obtain

$$D_{\mathrm{KL}}(Q \,||\, Q_\lambda) \le \lambda D_{\mathrm{KL}}(Q \,||\, Q) + (1-\lambda)D_{\mathrm{KL}}(Q \,||\, P_{\hat{\theta}}) \to 0 \quad \text{as} \quad \lambda \to 1.$$

□

We refer to a solution of the optimization problem (3.2) as the *MEM distribution* and denote it as Q_{MEM}. By similar arguments to the ones used in order to establish the lower bound in Lemma 3.4, one can show that, when $y \in \operatorname{int}(\operatorname{dom}\kappa_P) = \operatorname{int}(\operatorname{conv}\Omega_P)$, the MEM distribution is a particular member of the exponential family generated by the reference distribution P. More precisely, it holds that $Q_{MEM} = P_\theta$ where $\theta = \nabla\psi_P^*(y)$ and consequently

$$f_{Q_{MEM}}(x) = \frac{dP_\theta}{dP}(x) = \exp\left(\langle x, \theta\rangle - \log\int_\Omega \exp(\langle\cdot, \theta\rangle)dP\right) = \frac{\exp(\langle x, \theta\rangle)}{\int_\Omega \exp(\langle\cdot, \theta\rangle)dP}.$$

This, again, highlights the intimate connection between the MEM function and exponential families. The case $y \in \operatorname{bd}(\operatorname{dom}\kappa_P)$ is more subtle and will be the topic of future research.

In what follows, we assume that the reference distribution of the MEM function satisfies the conditions stated in Theorem 3.1, that is:

Assumption B The distribution $P \in \mathcal{P}(\Omega)$ satisfies one of the following conditions:

(i) Ω_P is uncountable.
(ii) Ω_P is countable and $\operatorname{conv} \Omega_P$ is closed (as is the case when Ω_P is finite).

Under Assumptions A and B, the MEM function and the Cramér rate function coincide. As an immediate consequence, we obtain that the MEM function κ_P is of Legendre type. More importantly, we will see that the alternative representation by means of Cramér's rate function is more tractable compared to the original definition given in (3.2).

Theorem 3.2 (Properties of the MEM function) *Let $P \in \mathcal{P}(\Omega)$ satisfy Assumptions A and B. Then the following hold:*

(a) $\kappa_P(y) \geq 0$ and equality holds if and only if $y = \mathbb{E}_P$.
(b) κ_P is of Legendre type.
(c) κ_P is coercive in the sense that $\lim_{\|y\|\to\infty} \kappa_P(y) = +\infty$ [11, Definition 11.10]. *In particular, $\kappa_P(y)$ is level bounded.*
(d) If M_P is finite (which holds, in particular, when Ω_P is bounded), then κ_P is supercoercive in the sense that $\lim_{\|y\|\to\infty} \kappa_P(y)/\|y\| = +\infty$ [11, Definition 11.10].

Proof Part (a) is evident from the definition of κ_P as given in (3.2) and [22, Proposition 6.2]. Part (b) follows directly from the equivalence to the Cramér rate function ψ_P^* and Corollary 2.1. To see (c), observe that (a) implies that κ_P admits a unique minimizer $\mathbb{E}_P$ which combined with the fact that κ_P is closed, proper and convex (since κ_P is of Legendre type due to (b)) establishes the result by [3, Proposition 3.1.3]. Lastly, if the moment generating function is finite, then so is ψ_P, and the supercoercivity of $\kappa_P = \psi_P^*$ follows from [64, Theorem 11.8(d)].[5] If Ω_P is bounded then $\operatorname{dom} \kappa_P$ is bounded due to Lemma 3.3. In this case, $\kappa_P = \psi_P^*$ is trivially supercoercive and the claim that ψ_P is finite follows from [64, Theorem 11.8(d)]. □

The results presented in the remainder of this work are established under Assumptions A and B which, in particular, ensure the equivalence between the MEM and Cramér rate functions. For this reason, we take this opportunity to standardize our nomenclature: between the two options (κ_P or ψ_P^*) we will opt for the one that corresponds to the Cramér rate function ψ_P^*. This choice is motivated by our intent to emphasize the more computationally appealing definition and the connection to the log-normalizer function ψ_P. Nevertheless, in the definition of some new concepts defined by means of Cramér's rate function, we will adopt the MEM terminology in order to emphasize the motivation in the context of estimation.

If the reference distribution belongs to an exponential family generated by some measure $P \in \mathcal{M}(\Omega)$, i.e., if for some $\hat{\theta} \in \Theta_P$ we consider a new exponential family

[5] The definition of supercoercive convex functions we use here follows [11, Definition 11.10] In [64] the authors refer to such functions as coercive (see [64, Definition 3.25]).

generated by the probability measure $P_{\hat{\theta}}$,[6] then the corresponding moment-generating function takes the form

$$M_{P_{\hat{\theta}}}[\theta] = \exp\left(\psi_P(\hat{\theta}+\theta) - \psi_P(\hat{\theta})\right). \tag{3.9}$$

In this case, the Cramér rate functions that corresponds to $P_{\hat{\theta}}$ and P share a useful relation summarized in the following lemma.

Lemma 3.5 *Let $\mathcal{F}_P$ be a minimal and steep exponential family generated by $P \in \mathcal{M}(\Omega)$ and assume further that, for any $\theta \in \operatorname{int}\Theta_P$, Assumption B holds for $P_\theta \in \mathcal{P}(\Omega)$. Then, for any $\hat{\theta} \in \operatorname{int}\Theta_P$ and $y \in \operatorname{dom}\psi_P^*$, we have $\psi^*_{P_{\hat{\theta}}}(y) = D_{\psi_P^*}(y, \hat{y})$ where $\hat{y} := \nabla\psi_P(\hat{\theta}) \in \operatorname{int}\Omega_P^{cc}$.*

Proof For $y \in \operatorname{dom}\psi_P^*$, we have

$$\begin{aligned}\psi^*_{P_{\hat{\theta}}}(y) &\overset{(3)}{=} \sup\left\{\langle y,\theta\rangle - \log\left(M_{P_{\hat{\theta}}}[\theta]\right) : \theta \in \mathbb{R}^d\right\} \\ &\overset{(3.9)}{=} \sup\left\{\langle y,\theta\rangle - [\psi_P(\hat{\theta}+\theta) - \psi_P(\hat{\theta})] : \theta \in \mathbb{R}^d\right\} \\ &= \psi_P^*(y) + \psi_P(\hat{\theta}) - \langle y, \hat{\theta}\rangle.\end{aligned}$$

The result follows from the definition of the Bregman distance, (2.2) and $\hat{\theta} \in \operatorname{int}(\operatorname{dom}\psi_P)$. □

We list in Table 1 below a number of examples of Cramér rate functions that correspond to most of the popular distributions (i.e. choices of the reference distribution $P \in \mathcal{P}(\Omega)$). Some of the functions admit a closed form expression while others are given implicitly.[7] The derivations and further details are included in Appendix A. Observe that all cases considered below satisfy Assumptions A and B which guarantees the equivalence established in Theorem 3.1: indeed, with some exceptions, all the distributions in Table 1 are minimal with a natural parameter space Θ_P open which implies steepness. These exceptions are: the multinomial distribution which is minimal under an appropriate reformulation and the multivariate normal-inverse Gaussian which is steep (see Appendix A). Here, we provide the Cramér rate function of the multinomial distribution in minimal form. Thus, Assumption A holds for all the distributions given in Table 1. This comprehensive list complements and extends some previously established formulas [50, 67].

Many computations are facilitated in the presence of separability as described in the following remark.

Remark 3.1 *(Separability of ψ_P^*)* In most examples, the reference distribution $P \in \mathcal{P}(\Omega)$ admits a separable structure of the form $P(y) = P_1(y_1)P_2(y_2)\cdots P_d(y_d)$ where

[6] Recall from the definition of $\mathcal{F}_P$ that $P_{\hat{\theta}}$ is the probability measure with $\frac{dP_{\hat{\theta}}}{dP}(y) = \exp(\langle y,\hat{\theta}\rangle - \psi_P(\hat{\theta}))$.

[7] One can evaluate Cramér's rate function value at a point of interest by solving a nonlinear system.

$P_i \in \mathcal{P}(\Omega_i)$, $\Omega_i \subset \mathbb{R}$, i.e., each component corresponds to an i.i.d. random variable. In this case, since $\mathbb{M}_P[\theta] = \prod_{i=1}^d \mathbb{M}_{P_i}[\theta_i]$ [61, Section 4.4], we have

$$\psi_P^*(y) = \sup\left\{\langle y, \theta\rangle - \log\left(\mathbb{M}_P[\theta]\right) : \theta \in \mathbb{R}^d\right\} = \sum_{i=1}^d \sup\left\{y_i\theta_i - \log\left(\mathbb{M}_{P_i}[\theta_i]\right) : \theta_i \in \mathbb{R}\right\}.$$

Hence, in most of our examples below we will consider only the case $d = 1$. $\Diamond$

In Table 1 we employ the convention that $0\log(0) = 0$ and define

$$\Delta_{(d)} := \left\{y \in \mathbb{R}_+^d : \sum_{i=1}^d y_i \leq 1\right\} \quad \text{and} \quad I(p) := \{y \in \mathbb{R}^d : y_i = 0\ (p_i = 0)\} \quad (p \in \mathbb{R}^d).$$

Remark 3.2 *(On Table 1)* We provide some additional comments on Table 1 here.

(a) (Special cases)

- As special cases of the Gamma distribution we obtain Chi-squared with parameter k ($\alpha = k/2, \beta = 1/2$), Erlang ($\alpha$ positive integer), and exponential ($\alpha = 1$) distributions.
- As special cases of the multinomial distribution, we obtain binomial ($d = 1$, $n > 1$), Bernoulli ($d = 1, n = 1$), and categorical ($d > 1, n = 1$) distributions.
- As special cases of the negative multinomial distribution we obtain the negative binomial ($d = 1$) and (shifted) geometric ($d = 1$, $y_0 = 1$) distributions.

(b) (Statistical interpretation) For many reference distributions, ψ_P^* recovers well-known functions from information theory and related areas. Here, the MEM provides an information-driven, statistical interpretation for these functions. Examples include the squared Mahalanobis distance (multivariate normal), pseudo-Huber loss (multivariate normal-inverse Gaussian), Itakura-Saito distance (Gamma), Burg entropy (exponential), Fermi-Dirac entropy (Bernoulli), and the generalized cross entropy (Poisson). $\Diamond$

4 The MEM estimator and models for inverse problems

In this section, we show how the MEM function can be used in various modeling paradigms. We start by presenting the MEM estimator and exploring some of its properties. We then discuss its (primal and dual) analogy to the maximum likelihood (ML) estimator. Finally, we will illustrate its efficacy by considering a class of linear models involving a regularization term.

4.1 The maximum entropy on the mean estimator

The maximum entropy on the mean (MEM) function gives rise to an information-driven criterion for measuring the compliance of given data with a prior distribution.

Table 1 Cramér rate functions for popular distributions

Reference distribution (P)	Cramér rate function ($\psi_P^*(y)$)	dom ψ_P^*
Multivariate normal $(\mu \in \mathbb{R}^d, \Sigma \in \mathbb{S}^d : \Sigma \succ 0)$	$\frac{1}{2}(y-\mu)^T \Sigma^{-1}(y-\mu)$	$\mathbb{R}^d$
Multivar. normal-inverse Gaussian $\big(\mu, \beta \in \mathbb{R}^d,\ \alpha, \delta \in \mathbb{R}, \Sigma \in \mathbb{R}^{d\times d}$: $\delta > 0,\ \Sigma \succ 0, \alpha \geq \sqrt{\beta^T \Sigma \beta}\big)$ $\gamma := \sqrt{\alpha^2 - \beta^T \Sigma \beta}$	$\alpha\sqrt{\delta^2 + (y-\mu)^T \Sigma^{-1}(y-\mu)} - \beta^T(y-\mu) - \delta\gamma$	$\mathbb{R}^d$
Gamma $(\alpha, \beta \in \mathbb{R}_{++})$	$\beta y - \alpha + \alpha \log\left(\frac{\alpha}{\beta y}\right)$	$\mathbb{R}_{++}$
Laplace $(\mu \in \mathbb{R}, b \in \mathbb{R}_{++})$	$\begin{cases} 0, & y = \mu, \\ \sqrt{1+\rho(y)^2} - 1 + \log\left(\frac{\sqrt{1+\rho(y)^2}-1}{\rho(y)^2/2}\right), & y \neq \mu, \end{cases}$ $(\rho(y) := (y-\mu)/b)$	$\mathbb{R}$
Poisson $(\lambda \in \mathbb{R}_{++})$	$y \log(y/\lambda) - y + \lambda$	$\mathbb{R}_+$
Multinomial $(n \in \mathbb{N}, p \in \Delta_{(d)}$: $\sum_{i=1}^d p_i < 1)$	$\sum_{i=1}^d y_i \log\left(\frac{y_i}{np_i}\right) + \left(n - \sum_{i=1}^d y_i\right) \log\left(\frac{n - \sum_{i=1}^d y_i}{n(1-\sum_{i=1}^d p_i)}\right)$	$n\Delta_{(d)} \cap I(p)$
Negative multinomial $(p \in [0,1)^d$, $y_0 \in \mathbb{R}_{++},\ p_0 := 1 - \sum_{i=1}^d p_i > 0)$	$\sum_{i=0}^d y_i \log\left(\frac{y_i}{p_i \bar{y}}\right)\quad (\bar{y} := \sum_{i=0}^d y_i)$	$\mathbb{R}_+^d \cap I(p)$
Discrete uniform $\big(a, b \in \mathbb{Z} : a \leq b$, $\mu := (a+b)/2, n := b-a+1\big)$	$\begin{cases} 0, & y = \mu, \\ (y-\mu)\theta - \log\left(\frac{e^{(b-\mu+1)\theta} - e^{(a-\mu)\theta}}{n(e^\theta - 1)}\right), & y \neq \mu, \end{cases}$ where $\theta \in \mathbb{R} :\ y + \frac{e^\theta}{e^\theta - 1} = \frac{(b+1)e^{(b+1)\theta} - ae^{a\theta}}{e^{(b+1)\theta} - e^{a\theta}}$	$[a, b]$
Continuous uniform $\big(a, b \in \mathbb{R} : a < b, \mu := (a+b)/2\big)$	$\begin{cases} 0, & y = \mu, \\ (y-\mu)\theta - \log\left(\frac{e^{(b-\mu)\theta} - e^{(a-\mu)\theta}}{(b-a)\theta}\right), & y \neq \mu, \end{cases}$ where $\theta \in \mathbb{R} :\ y + \frac{1}{\theta} = \frac{be^{b\theta} - ae^{a\theta}}{e^{b\theta} - e^{a\theta}}$	(a, b)
Logistic $(\mu \in \mathbb{R}, s \in \mathbb{R}_{++})$	$\begin{cases} 0, & y = \mu, \\ (y-\mu)\theta - \log(B(1-s\theta, 1+s\theta)), & y \neq \mu, \end{cases}$ where $\theta \in \mathbb{R}_+ :\ y - \mu = \frac{1}{\theta} + \frac{\pi s}{\tan(-\pi s\theta)}$	$\mathbb{R}$

Based on this function, we can define the MEM estimator as given in Definition 4.1 below. First, we introduce some additional terminology and notation that will be used in the sequel. Let $\Omega \subseteq \mathbb{R}^d$ and let $F_\Lambda = \{P_\lambda : \lambda \in \Lambda \subseteq \mathbb{R}^d\} \subset \mathcal{P}(\Omega)$ be a parameterized family of distributions indexed by $\lambda \in \Lambda$ such that $\mathbb{E}_{P_{\lambda_1}} = \mathbb{E}_{P_{\lambda_2}}$ if and only if $\lambda_1 = \lambda_2$. We call F_Λ as the *reference family* and say that it satisfies Assumptions A and B if they hold for each $P_\lambda \in F_\Lambda$. When F_Λ is an exponential family (in this case Λ is the natural parameter space Θ_P for some $P \in \mathcal{M}(\Omega)$) the MEM estimator was studied in [22, Chapter 6]. We stress that, in our presentation, F_Λ need *not* be an exponential family.

Definition 4.1 *(MEM estimator)* Let $F_\Lambda \subset \mathcal{P}(\Omega)$ be a reference family satisfying Assumptions A and B and assume that $\mathbb{E}_{P_{\lambda_1}} = \mathbb{E}_{P_{\lambda_2}}$ if and only if $\lambda_1 = \lambda_2$. For an observation $\hat{y} \in \mathbb{R}^d$, let $P_{\hat{\lambda}} \in F_\Lambda$ be such that $\hat{y} = \mathbb{E}_{P_{\hat{\lambda}}}$, and let $S^* \subseteq \mathbb{R}^d$ be (nonempty) closed. The MEM estimator is defined as

$$y_{MEM}(\hat{y}, F_\Lambda, S^*) := \operatorname{argmin}\{\psi^*_{P_{\hat{\lambda}}}(y) : y \in S^*\}.$$

In order to simplify notation, in what follows, we will write $y_{MEM} := y_{MEM}(\hat{y}, F_\Lambda, S^*)$ when the dependence on the triple $(\hat{y}, F_\Lambda, S^*)$ is clear from the context.

Remark 4.1 *(The observation vector and its domain)* In Definition 4.1, the condition that $P_{\hat{\lambda}} \in F_\Lambda$ is chosen such that $\hat{y} = \mathbb{E}_{P_{\hat{\lambda}}}$ implies that the reference distribution is indexed by the observation vector $\hat{y}$. This condition combined with Assumption A entails that $\hat{y} \in \operatorname{int} \Omega^{cc}_{P_{\hat{\lambda}}}$ must hold due to Lemma 3.2. ♢

In order to establish the well-definedness of the MEM estimator, we will use the following extension of [22, Lemma 5.4].

Lemma 4.1 *Let $\phi : \mathbb{R}^d \to (-\infty, +\infty]$ be closed and Legendre-type, let $\varphi : \mathbb{R}^d \to (-\infty, +\infty]$ be proper, closed and convex such that* $\operatorname{int}(\operatorname{dom}\phi) \cap \operatorname{dom}\varphi \neq \emptyset$. *Assume that one of the functions is coercive while the other is bounded from below. Then there exists a unique solution $y^* \in \mathbb{R}^d$ to* $\min\{\phi(y) + \varphi(y) : y \in \mathbb{R}^d\}$, *which also satisfies $y^* \in$* $\operatorname{int}(\operatorname{dom}\phi) \cap \operatorname{dom}\varphi$.

Proof The existence and uniqueness of the solution follow from [11, Corollary 11.15]. It remains to show that $y^* \in \operatorname{int}(\operatorname{dom}\phi) \cap \operatorname{dom}\varphi$. Evidently, $y^* \in \operatorname{dom}\phi \cap \operatorname{dom}\varphi$ thus it is sufficient to show that $y^* \in \operatorname{int}(\operatorname{dom}\phi)$. Using [11, Theorem 16.2] and [11, Corollary 16.38] we have $0 \in \partial\phi(y^*) + \partial\varphi(y^*)$, in particular $\partial\phi(y^*) \neq \emptyset$. Since ϕ is of Legendre type we conclude that $y^* \in \operatorname{int}(\operatorname{dom}\phi)$ [60, Theorem 26.1]. □

Theorem 4.1 (Well-definedness of the MEM estimator) *Let $F_\Lambda \subset \mathcal{P}(\Omega)$ be a reference family satisfying Assumptions A and B. For $\hat{y} \in \mathbb{R}^d$, let $P_{\hat{\lambda}} \in F_\Lambda$ such that $\hat{y} = E_{P_{\hat{\lambda}}}$, and let $S^* \subseteq \mathbb{R}^d$ be closed with $S^* \cap$* $\operatorname{dom}\psi^*_{P_{\hat{\lambda}}} \neq \emptyset$. *Then, the MEM estimator y_{MEM} exists. If, in addition, S^* is convex and* $\operatorname{int}(\operatorname{dom}\psi^*_{P_{\hat{\lambda}}}) \cap S^* \neq \emptyset$, *$y_{MEM}$ is unique and in* $\operatorname{int}(\operatorname{dom}\psi^*_{P_{\hat{\lambda}}}) \cap S^*$.

Proof Recall that, by Theorem 3.2, $\psi^*_{P_{\hat{\lambda}}}$ is coercive and of Legendre type (proper, closed, steep and strictly convex on the interior of its domain). Observe that $S^* \subset \mathbb{R}^d$ is closed and $S^* \cap \operatorname{dom} \psi^*_{P_{\hat{\lambda}}} \neq \emptyset$. Thus, the function $\psi^*_{P_{\hat{\lambda}}} + \delta_{S^*}$ is proper, closed and coercive. Hence, the existence of the MEM estimator follows from [3, Remark 3.4.1, Theorem 3.4.1]. The case when S^* is convex and $\operatorname{int}(\operatorname{dom} \psi^*_{P_{\hat{\lambda}}}) \cap S^* \neq \emptyset$ follows from Lemma 4.1 with $\phi = \psi^*_{P_{\hat{\lambda}}}$ and $\varphi = \delta_S$ due to the coercivity of $\psi^*_{P_{\hat{\lambda}}}$ and the fact that δ_S is bounded from below. □

4.1.1 Analogy between MEM and ML (for exponential families)

Maximum likelihood (ML) is arguably the most popular principle for statistical estimation. Here, the estimated parameters are chosen as the most likely to produce a given sample of observed data while satisfying model assumptions. More precisely, for some $\Omega \subseteq \mathbb{R}^d$, the model is defined by means of a nonempty, closed set $S \subseteq \mathbb{R}^d$ of admissible parameters and a parameterized family of distributions $F_\Lambda = \{P_\lambda : \lambda \in \Lambda \subseteq \mathbb{R}^m\} \subset \mathcal{P}(\Omega)$ with densities f_{P_λ}. Given a sample of observed data $\hat{y} \in \mathbb{R}^d$, the ML estimator $\lambda_{ML}(\hat{y}, F_\Lambda, S)$ is defined as

$$\lambda_{ML}(\hat{y}, F_\Lambda, S) := \operatorname{argmax}\{\log f_{P_\lambda}(\hat{y}) : \lambda \in S \cap \Lambda\}.$$

In order to simplify notation, we will write $\lambda_{ML} := \lambda_{ML}(\hat{y}, F_\Lambda, S)$ when the dependence on the triple $(\hat{y}, F_\Lambda, S)$ is clear from the context.

An intriguing connection between the ML and MEM estimator comes to light when Λ is the natural parameter space Θ_P of an exponential family induced by $P \in \mathcal{M}(\Omega)$. The MEM estimator can then be retrieved by solving one of two alternative optimization problems each of which has a closely related problem that yields the ML estimator. One problem is driven by information-theoretic arguments, while the other emphasizes a connection motivated by convex duality. These connections were previously observed in [22, Chapter 6] and are summarized in the following theorem. For consistency, we denote the ML estimator as θ_{ML}.

Theorem 4.2 (MEM and ML estimator analogy) *Let $\mathcal{F}_P$ be a minimal and steep exponential family generated by $P \in \mathcal{M}(\Omega)$ and assume that, for any $\theta \in \operatorname{int}\Theta_P$, Assumption B holds with respect to $P_\theta \in \mathcal{P}(\Omega)$. Let $S, S^* \subseteq \mathbb{R}^d$ such that $S \cap \operatorname{dom} \psi_P \neq \emptyset$ and $S^* \cap \operatorname{dom} \psi^*_P \neq \emptyset$. Finally, let $\hat{y} \in \operatorname{int} \Omega^{cc}_P$ and set $\hat{\theta} := \nabla\psi^*_P(\hat{y})$. Then the following hold:*

(a) (Primal analogy) If $S^* \cap \operatorname{int}(\operatorname{dom} \psi^*_P) \neq \emptyset$ and $\nabla\psi^*_P(S^* \cap \operatorname{int}(\operatorname{dom} \psi^*_P)) = S \cap \operatorname{int}(\operatorname{dom} \psi_P)$, then $y_{MEM} = \nabla\psi_P(\theta_{MEM})$ where

$$\theta_{MEM} \in \operatorname{argmin}\{D_{KL}(P_\theta \,||\, P_{\hat{\theta}}) : \theta \in S\} \quad \textit{and} \quad \theta_{ML} \in \operatorname{argmin}\{D_{KL}(P_{\hat{\theta}} \,||\, P_\theta) : \theta \in S\}. \tag{4.1}$$

(b) (Dual analogy): We have

$$y_{MEM} \in \operatorname{argmin}\{D_{\psi_P^*}(y, \hat{y}) : y \in S^*\} \quad \textit{and} \quad \theta_{ML} \in \operatorname{argmin}\{D_{\psi_P}(\theta, \hat{\theta}) : \theta \in S\}. \tag{4.2}$$

Proof Since $\mathcal{F}_P$ is assumed to be minimal and steep, it is easy to verify (recall (3.9)) that P_θ satisfies Assumption A for any $\theta \in \operatorname{int} \Theta_P$. As we assume $S \cap \operatorname{dom} \psi_P \neq \emptyset$ and $S^* \cap \operatorname{dom} \psi_P^* \neq \emptyset$, the MEM and ML estimator exist due to Theorem 4.1 and [22, Theorem 5.7], respectively. We now prove (b). Since $\mathcal{F}_P$ is an exponential family, we have $\log f_{P_\theta}(\hat{y}) = \langle \hat{y}, \theta \rangle - \psi_P(\theta)$ and the ML estimator is a solution to

$$\begin{aligned}\max\{\log f_{P_\theta}(\hat{y}) : \theta \in S\} &= \max\{\langle \hat{y}, \theta \rangle - \psi_P(\theta) : \theta \in S\} \\ &= -\min\{D_{\psi_P}(\theta, \nabla \psi_P^*(\hat{y})) : \theta \in S\} - \psi_P(\nabla \psi_P^*(\hat{y})) \\ &\quad + \langle \hat{y}, \nabla \psi_P^*(\hat{y}) \rangle.\end{aligned}$$

Omitting terms independent of the minimization and using that $\hat{\theta} = \nabla \psi_P^*(\hat{y})$, the formulation for the ML estimator follows. To obtain the formulation for the MEM estimator, observe that, due to Lemma 3.5, we have

$$\min\{\psi_{P_{\hat{\theta}}}^*(y) : y \in S^*\} = \min\{D_{\psi_P^*}(y, \nabla \psi_P(\hat{\theta})) : y \in S^*\}.$$

Thus, the result follows by recalling that $\hat{y} = \nabla \psi_P(\hat{\theta})$.

We now turn to prove (a). Since $S^* \cap \operatorname{int}(\operatorname{dom} \psi_P^*) \neq \emptyset$ we obtain by Theorem 4.1 that $y_{MEM} \in S^* \cap \operatorname{int}(\operatorname{dom} \psi_P^*)$. This fact combined with the assumption $\nabla \psi_P^*(S^* \cap \operatorname{int}(\operatorname{dom} \psi_P^*)) = S \cap \operatorname{int}(\operatorname{dom} \psi_P)$ implies that $\nabla \psi_P^*(y_{MEM}) \in S \cap \operatorname{int}(\operatorname{dom} \psi_P)$. Thus, (a) follows from (b) due to the Bregman distance dual representation property (2.3) and Remark 2.1. □

The primal and dual analogy between the MEM and ML estimator for exponential families clarifies that the two are symmetric principles.

4.2 Examples - linear models

To illustrate the versatility of the MEM estimation framework, we will consider the broad class of linear models which are among the most popular paradigms in statistical estimation with applications in numerous fields such as image processing, bio-informatics, machine learning etc.

We assume that the set S^* of admissible mean value parameters is the image of a convex set $X \subseteq \mathbb{R}^d$ under a linear mapping defined by a measurement matrix $A \in \mathbb{R}^{m \times d}$. In many practical scenarios, this matrix satisfies some application-related properties, which in combination with the set X restricts the image space to a subset of $\mathbb{R}^m$. We will denote by $\mathcal{C}$ the set of all matrices that satisfy such a condition for the application in question. The second component in the model is $F_\Lambda = \{P_\lambda : \lambda \in \Lambda \subseteq \mathbb{R}^m\} \subset \mathcal{P}(\Omega)$, a reference family indexed by $\lambda \in \Lambda$ such that $\mathbb{E}_{P_{\lambda_1}} = \mathbb{E}_{P_{\lambda_2}}$ if and only if $\lambda_1 = \lambda_2$. The reference distribution is specified from this family by means of the observation vector $\hat{y}$. From Remark 4.1 it follows that such a family of distributions must satisfy $\hat{y} \in \operatorname{int} \Omega^{cc}_{P_{\hat{\lambda}}}$ for $\hat{\lambda}$ such that $\mathbb{E}_{P_{\hat{\lambda}}} = \hat{y}$. In some cases, this condition imposes additional assumptions that must be satisfied by the measurement vector. We will denote the set of measurement vectors that satisfy such an assumption with respect to the family of distributions under consideration by $D := \{y \in \mathbb{R}^m : \mathbb{E}_{P_\lambda} = y \ (\lambda \in \Lambda)\}$. To summarize, an MEM estimator of the linear model outlined above is obtained by solving

$$\min \left\{ \psi^*_{P_{\hat{\lambda}}}(Ax) : x \in X \right\} \qquad (\hat{\lambda} \in \Lambda : \mathbb{E}_{P_{\hat{\lambda}}} = \hat{y}), \tag{4.3}$$

under the following set of assumptions:

Assumption C *(MEM estimation for linear models)*

1. The reference family F_Λ satisfies Assumptions A and B.
2. The set $X \subseteq \mathbb{R}^d$ is nonempty and convex.
3. $A \in \mathcal{C}$ and for any $x \in X$ it holds that $Ax \in \operatorname{dom} \psi^*_P$.
4. The observation vector satisfies $\hat{y} \in D$.

In Table 2, we present some examples of MEM linear models that correspond to particular choices of a reference family. In all cases, we assume that the reference family admits a separable structure as outlined in Remark 3.1. The vectors a_i $(i = 1, \ldots, m)$ stand for the ith row of the matrix A. We set

$$\mathcal{C}_0 := \{A \in \mathbb{R}^{m \times d}_+ : \text{A has no zero rows or columns}\}.$$

Remark 4.2 Additional models are readily available by choosing any of the reference distributions presented in Table 1. Alternatively, one may consider a family of linear models where the natural parameters are the ones restricted to the image of a convex set under a linear mapping. This class of models is commonly referred to as *generalized linear models* with a *canonical link function* [55]. ♢

The MEM linear model with reference family that corresponds to the normal distribution coincides with its ML counterpart, resulting in the celebrated least-squares model [18]. This phenomenon is unique for the normal distribution and is a direct consequence of the fact that the squared Euclidean norm is the only self-conjugate function [60, Section 12].

Linear inverse models under the Poisson noise assumption have been successfully applied in various disciplines including fluorescence microscopy, optical/infrared astronomy, and medical applications such as positron emission tomography (PET) (see, for example, [16, 66]). The MEM linear model with Poisson reference distribution outlined in Table 2 was previously suggested in [8, Subsection 5.3] as an example

Table 2 Linear models under the MEM estimation framework for various reference families

Reference family	Objective function ($\psi^*_{P_{\hat{\lambda}}} \circ A$)	$\mathcal{C}$	X	D
Normal	$\frac{1}{2}\|Ax - \hat{y}\|_2^2$	$\mathbb{R}^{m\times d}$	$\mathbb{R}^d$	$\mathbb{R}^m$
Poisson	$\sum_{i=1}^{m}\left[\langle a_i, x\rangle \log\left(\langle a_i, x\rangle / \hat{y}_i\right) - \langle a_i, x\rangle + \hat{y}_i\right]$	$\mathcal{C}_0$	$\mathbb{R}^d_+$	$\mathbb{R}^m_{++}$
Gamma ($\beta = 1$)	$\sum_{i=1}^{m}\left[\langle a_i, x\rangle - \hat{y}_i \log\left(\langle a_i, x\rangle\right) - \left(\hat{y}_i - \hat{y}_i \log\left(\hat{y}_i\right)\right)\right]$	$\mathcal{C}_0$	$\mathbb{R}^d_{++}$	$\mathbb{R}^m_+$

for the algorithmic setting considered in that work (see further details in Sect. 5 where we expand on the framework considered in [8]).

If, for example, $X = \mathbb{R}^d$ and $\operatorname{rge} A = \mathbb{R}^m$ with $m < d$, then $x \in \mathbb{R}^d$ such that $y_{ML} = y_{MEM} = Ax = \hat{y}$. This outcome is not a result of a deep statistical characteristic but a simple consequence of the model's ill-posedness, a situation when the desired solution is not uniquely characterized by the model. Situations like this are among the reasons which motivate the use of *regularizers* which allow for the incorporation of some additional (prior) knowledge of the solution. This approach gives rise to the following extended version of model (4.3)

$$\min \left\{ \psi^*_{P_{\hat{\lambda}}}(Ax) + \varphi(x) : x \in X \right\} \qquad (\hat{\lambda} \in \Lambda : \mathbb{E}_{P_{\hat{\lambda}}} = \hat{y}), \tag{4.4}$$

where, in our setting, $\varphi : \mathbb{R}^d \to (-\infty, +\infty]$ stands for a proper, closed, and convex function. In (4.4), the optimization formulation is designed to find a solution (model estimator) that balances between two criteria represented by the *fidelity* term $\psi^*_{P_{\hat{\lambda}}} \circ A$ and the *regularization* term φ. While the fidelity term penalizes the violation between the model and observations, the regularization term incorporates prior information (belief) on the solution, and in many cases, when the problem with the fidelity term alone is ill-posed, it also serves as a regularizer. In the context of MEM, the Cramér rate function can be used to penalize violations of the solution vector $x \in \mathbb{R}^d$ with respect to some prior reference measure $R \in \mathcal{P}(\Omega)$ that satisfies Assumptions A and B. In other words, we can set $\varphi(x) = \psi^*_R(x)$.

In many applications, the desired reference distribution of the regularizer will admit a separable structure (à la Remark 3.1). While this is advantageous from an algorithmic perspective (cf. Remark 5.1), other alternatives are viable. Non-separable priors can be considered in order to promote desirable correlations between the entries of the solution to problem (4.4). E.g., by considering the multinomial, negative multinomial, multivariate normal inverse Gaussian or multivariate normal (with non-diagonal correlation matrix in the latter) reference distributions intrinsically give rise to non-separable modeling. But there are other options that involve separable reference distributions with a composite structure such as

$$\varphi(x) = \psi^*_R(Lx) \qquad \text{or} \qquad \varphi(x) = \sum_{i=1}^{d} \psi^*_R(L_i x), \tag{4.5}$$

where $L \in \mathbb{R}^{r \times d}$, $L_i \in \mathbb{R}^{r \times d}$. For example, new variants of the well-known (discrete) *total variation* (TV) regularizer [62] can be considered by replacing the norm appearing in the original definition with a Cramér rate function while keeping the first-order finite difference matrix (further details are given at the end of Sect. 5). Different reference distributions might be used to promote desirable, application-specific, properties of the solution. Nevertheless, for all choices of reference distribution, the resulting function will admit some desirable properties, including convexity, differentiability, and coerciveness as established in Theorem 3.2. As we will see in the following section, these properties allow us to consider a unified algorithmic approach for tackling problem (4.4).

5 Algorithms

5.1 The Bregman proximal gradient method

The optimization formulations of statistical estimation problems as presented in the previous section are solved by optimization algorithms. Customized methods, such as the ones we consider here, allow us to leverage the structure of a given problem, thus resulting in a significant efficiency improvement compared to general-purpose solvers. The structure of problems which are of interest to us is given by the *additive composite model*

$$\min\{F(x) := f(x) + g(x) : x \in \mathbb{R}^d\}, \tag{5.1}$$

where $f, g : \mathbb{R}^d \to (-\infty, +\infty]$ are proper, closed, and convex.

We will assume that both the fidelity and regularization terms, represented by f and g, respectively, are continuously differentiable on the interior of their domain. This assumption holds for all the modeling paradigms discussed in the previous section. In particular, model (4.4) is recovered with $f = \psi_P^* \circ A$ and $g = \psi_R^*$. Our focus on this type of problem is for convenience only as our goal is merely to illustrate how modern first-order methods can be used for computing MEM estimators, much like their popular ML counterparts. We point out that we are not limited to this setting. Other models can be considered as well, e.g., by blending a fidelity term originating from an MEM modeling paradigm with a traditional regularizer or vice versa. In this case, similar algorithms are applicable under suitable adjustments.

The method we consider is the *Bregman proximal gradient* (BPG) method. This first-order iterative algorithm admits a comparably mild per-iteration complexity and as such, it is particularly suitable for contemporary large-scale applications.

Before we present the BPG method, we need to define its fundamental components [8, 19].

Smooth adaptable kernel: Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be proper, closed and continuously differentiable on $\operatorname{int}(\operatorname{dom} f)$. Then $h : \mathbb{R}^d \to (-\infty, +\infty]$ of Legendre type is a *smooth adaptable kernel* with respect to f if $\operatorname{dom} h \subseteq \operatorname{dom} f$ and there exists $L > 0$ such that $Lh - f$ is convex.

Bregman proximal operator: Let $g : \mathbb{R}^d \to (-\infty, +\infty]$ be closed and proper and $h : \mathbb{R}^d \to (-\infty, +\infty]$ of Legendre type. Then the *Bregman proximal operator* is defined as

$$\operatorname{prox}_g^h(\bar{x}) := \operatorname{argmin}\left\{g(x) + D_h(x, \bar{x}) : x \in \mathbb{R}^d\right\} \qquad (\bar{x} \in \operatorname{int}(\operatorname{dom} h)). \tag{5.2}$$

The Bregman distance induced by a smooth adaptable kernel h is not necessarily symmetric. A symmetry measure $\alpha(h)$ was introduced in [8, Subsection 2.3] and it plays an important role for assuring the convergence of the BPG method.

The BPG method is applicable under the following assumption.

Assumption D Consider problem (5.1) and assume that there exists a function of Legendre type $h : \mathbb{R}^d \to (-\infty, +\infty]$ such that:

1. h is a smooth adaptable kernel with respect to f.
2. h induces a computationally efficient Bregman proximal operator with respect to g.

The BPG method reads:

(BPG Method) Pick $t \in (0, (1 + \alpha(h))/(2\,L)]$ and $x^0 \in \operatorname{int}(\operatorname{dom} h)$. For $k = 0, 1, 2, \ldots$ compute

$$x^{k+1} = \operatorname{prox}^h_{tg}\left(\nabla h^*\left(\nabla h(x^k) - t\nabla f(x^k)\right)\right).$$

For $h = (1/2)\|\cdot\|_2^2$ and f convex, $Lh - f$ is convex if and only if ∇f is L-Lipschitz. In this case, the Bregman proximal operator reduces to the classical proximal operator and the BPG method is the well-known proximal gradient algorithm [13].

For completeness, in the following proposition we recall the objective sublinear convergence rate for the BPG method. Under suitable assumptions, the convergence improves to linear [7].

Proposition 5.1 (BPG Convergence Rate, [8, Corollary 1]) *Let* $\{x^k\}_{k\in\mathbb{N}}$ *be a sequence generated by the BPG method with step size* $t = (1 + \alpha(h))/(2L)$. *Assume further that* $\operatorname{dom} h = \operatorname{cl}\operatorname{dom} h$ *and that there exists a solution* x^* *to problem* (5.1). *Then,*

$$F(x^k) - F(x^*) \leq \frac{2\,LD_h(x^*, x^0)}{(1 + \alpha(h))k}.$$

The convergence of the sequence $\{x^k\}_{k\in\mathbb{N}}$ generated by the BPG method was established under some additional technical assumptions, for further details see [8, Theorem 2].

It is important to notice that many other methods can be also considered in some scenarios. For simplicity's sake, we confine ourselves to the basic BPG scheme, but the operators to be presented in the following subesection can be readily applied to many classes of more sophisticated algorithms. We point out blow some of the prominent alternatives.

Accelerated First Order Methods In the seminal work [56] and its extension in [14], the authors introduced accelerated variants to the gradient descent and proximal gradient (BPG in the Euclidean setting) methods, respectively. These methods employ a clever modification of the basic scheme that results in a superior theoretical and practical performance without a sacrifice in the per-iteration complexity.

Similar modifications are considered also for the general non-Euclidean case, for example, in [4] (see also its extension to the additive composite model in [45]). Superior theoretical guarantees cannot be established for the general non-Euclidean case as has been shown in [36, 37]. Nevertheless, in practice, the proposed algorithms tend to improve the performance of the basic schemes.

Primal-Dual Splitting Methods When the problem at hand is more complicated like in the case when the regularization term is given by a composition of a MEM function with a linear mapping as in (4.5) the resulting proximal operator might not be efficiently computed. In such cases, we can consider primal-dual splitting algorithms that exploit the structure of the problem in a way that give rise to efficient updates at each iteration. We cannot review here the vast literature on the topic, instead, we refer the reader to some classic references [20, 29] and recent reviews [30, 63].

Proximal operators are fundamental for primal-dual splitting methods and the operators developed in this work can be employed by such methods to solve models involving MEM functions (see an example at the end of this section for an image deblurring problem solved with a primal-dual method proposed in [25]).

Primal-dual splitting methods that rely on general Bregman proximal operators can be also found in the literature, for example, see [26, 28].

5.2 The BPG method for MEM linear models

In order to customize the BPG method to a particular instance of problem (5.1), a smooth adaptable kernel and corresponding Bregman proximal operator must be specified. To illustrate this idea for MEM estimation, we focus on the linear models discussed in the previous section. In particular, we consider the model (4.4) where $\varphi = \psi_R^*$. We assume that Assumption C holds and that the prior reference measure $R \in \mathcal{P}(\Omega)$ satisfies Assumptions A and B. Furthermore, we assume that $\operatorname{dom} \psi_R \subseteq X$ which allows us to disregard the constraint $x \in X$. The latter assumption holds in many practical situations and we assume it here for simplicity. Otherwise, one can simply apply the BPG method with $g = \psi_R^* + \delta_X$ (under the appropriate adjustments to the proximal operator). In Table 3 below, we summarize the smooth adaptable kernels suitable for the models described in the previous section, see Table 2. In all cases, the smooth adaptable function admits a separable structure of the form $h(x) = \sum_{j=1}^{d} h_j(x_j)$ where $h_j : \mathbb{R} \to (-\infty, +\infty]$ $(j = 1, \ldots, d)$ is a (univariate) function of Legendre type. As we will see in what follows, this property is very desirable as it gives rise to a computationally efficient implementation of the Bregman proximal operator. For completeness, we include the the constants and explicit formulas for the operators involved in the BPG method.

The kernel and related constant that corresponds to the normal reference family is a well-known consequence due to the Lipschitz gradient continuity, a special case of the

Table 3 Smooth adaptable kernels and related operators and constants that correspond to the objective function ($f = \psi^*_{P_{\hat{\theta}}} \circ A$) of the linear models listed in Table 2

Reference family	Kernel (h_j)	Constant (L)	$[\nabla h(x)]_j$	$[\nabla h^*(z)]_j$	$\alpha(h)$
Normal	$(1/2)x_j^2$	$\|A\|_2 := \sqrt{\lambda_{\max}(A^T A)}$	x_j	z_j	1
Poisson	$x_j \log(x_j)$	$\|A\|_1 := \max_{j=1,2,\ldots,d} \sum_{i=1}^{m} \|A_{i,j}\|$	$\log(x_j) + 1$	$\exp(z_j - 1)$	0
Gamma ($\beta = 1$)	$-\log(x_j)$	$\|\hat{y}\|_1 := \sum_{i=1}^{m} \|\hat{y}_i\|$	$-1/x_j$	$-1/z_j$	0

smooth adaptability property considered here.[8] The kernel and related constant that corresponds to the Poisson reference family is due to [8, Lemma 8]. The kernel and related constant that corresponds to the Gamma distribution follows from [8, Lemma 7].

We now discuss the special form of the Bregman proximal operator in the setting of the linear model (4.4) with $\varphi = \psi_R^*$. According to (5.2), for any $t > 0$, the Bregman proximal operator is defined by the smooth adaptable kernel h and the regularizer $g = \psi_R^*$ as follows:

$$\operatorname{prox}^h_{t\psi_R^*}(\bar{x}) = \operatorname{argmin}\left\{t\psi_R^*(u) + D_h(u, \bar{x}) : u \in \mathbb{R}^d\right\}. \tag{5.3}$$

The following theorem records that, in our setting, the above operator is well-defined.

Theorem 5.1 (Well-definedness of the Bregman proximal operator) *Let $h : \mathbb{R}^d \to (-\infty, +\infty]$ be of Legendre type and let $R \in \mathcal{P}(\Omega)$ be a reference distribution satisfying the conditions in Assumptions A and B. Assume further that* $\operatorname{int}(\operatorname{dom} h) \cap \operatorname{dom} \psi_R^* \neq \emptyset$*. Then, for any $t > 0$ and $\bar{x} \in$* $\operatorname{int}(\operatorname{dom} h)$*, the Bregman proximal operator defined in (5.3) produces a unique point in* $\operatorname{int}(\operatorname{dom} h) \cap \operatorname{dom} \psi_R^*$.

Proof Since $\bar{x} \in \operatorname{int}(\operatorname{dom} h)$, the function $D_h(\cdot, \bar{x})$ is proper. In addition, since h is of Legendre type, so is $D_h(\cdot, \bar{x})$. Finally, $D_h(\cdot, \bar{x})$ is bounded below (by zero) by the convexity of h. The result follows from Lemma 4.1 with $\phi = D_h$ and $\varphi = t\psi_R^*$ due to the aforementioned properties of D_h and the coercivity of $t\psi_R^*$ (Theorem 3.2 and $t > 0$). □

We now show that this operator is also computationally tractable. For many reference distributions, this fact stems from the following separability property.

Remark 5.1 *(Separability of the Bregman proximal operator)* In all cases under consideration, the smooth adaptable kernel $h : \mathbb{R}^d \to (-\infty, +\infty]$ admits a separable structure $h(x) = \sum_{j=1}^d h_j(x_j)$. Therefore, by (2.4), the induced Bregman distance satisfies: $D_h(x, y) = \sum_{i=1}^d D_{h_i}(x_i, y_i)$. If, in addition, the Cramér rate function admits a separable structure $\psi_R^* = \sum_{i=1}^d \psi_{R_i}^*$ (cf. Remark 3.1), then the optimization problem defining the Bregman proximal operator is separable and can be evaluated for each component of $\bar{x}$. ♢

Given a particular instance of problem (5.1), with fidelity term $f = \psi_{P_{\hat{\lambda}}}^* \circ A$ and regularizer $g = \psi_R^*$, one can derive a formula for the corresponding Bregman proximal operator. These formulas are summarized in Tables 4, 5 and 6 for each of the combinations of linear models (by using a compatible kernel generating distance from Table 3) and regularizers from Table 1. Some formulas are given in a closed form, others must be evaluated numerically through a solution of a nonlinear system.[9] Due

[8] More precisely, the equivalence holds for convex functions such as the ones considered here. For the nonconvex case see an extension of the smooth adaptability condition presented in [19].

[9] The solution of the nonlinear system can be efficiently approximated by various methods. In our implementation, building upon the fact that the systems involve monotonic functions (since they stem from the optimality conditions of a convex problem), we used a variant of safeguarded Newton-Raphson method.

to Remark 5.1, for most of the regularizer reference distributions (excluding only the multivariate normal, multinomial and negative multinomial) the resulting subproblem is separable. Thus, for the sake of simplicity and without loss of generality, we assume that $d = 1$, i.e., the resulting formulas correspond to one entry of the vector produced by the operator. The general case follows by applying the operator components-wise on all the elements of a vector $\bar{x} \in \mathbb{R}^d$. An implementation of the operators along with selected algorithms, applications, and detailed derivations of the operators can be found under:

https://github.com/yakov-vaisbourd/memmpy.

Table 4 lists the formulas of Bregman proximal operators for the normal linear family. In this case, the operator reduces to the classical proximal operator [52].

Recall that the Cramér rate function induced by a uniform (discrete/continuous) or logistic reference distribution does not admit a closed form. To compute their proximal operator we appeal to the corresponding dual of the subproblem in (5.3). This is done via Moreau decomposition (see, e.g., [13, Theorem 6.45]) which applies when the Bregman proximal operator (5.3) reduces to the classical proximal operator (i.e., when $h = (1/2)\|\cdot\|_2^2$). For the general case, we will employ a result summarized in Lemma 5.1 and Corollary 5.1 below. Some notation is needed: for a function $g : \mathbb{R}^d \to (-\infty, +\infty]$ proper, closed and convex and of $h : \mathbb{R}^d \to (-\infty, +\infty]$ of Legendre type we set

$$\operatorname{iconv}_g^h(\bar{x}) := \operatorname{argmin}\left\{g(x) + h(\bar{x} - x) : x \in \mathbb{R}^d\right\}. \tag{5.4}$$

This is the (possibly empty) solution of the optimization problem defining the *infimal convolution* $(g \square h)(\bar{x}) := \inf\left\{g(x) + h(\bar{x} - x) : x \in \mathbb{R}^d\right\}$.

Lemma 5.1 *Let $g : \mathbb{R}^d \to (-\infty, +\infty]$ be proper, closed, and convex, and let $h : \mathbb{R}^d \to (-\infty, +\infty]$ be of Legendre type. Let $\bar{x} \in \operatorname{int}(\operatorname{dom} h)$ and assume that there exists a unique point $x^+ := \operatorname{prox}_g^h(\bar{x})$ satisfying $x^+ \in \operatorname{int}(\operatorname{dom} h) \cap \operatorname{dom} g$. Then, $y^+ := \operatorname{iconv}_{g^*}^{h^*}(\nabla h(\bar{x}))$ exists and it holds that $\nabla h(x^+) + y^+ = \nabla h(\bar{x})$.*

Proof By the optimality condition of the optimization problem in the definition of the Bregman proximal operator (5.2) we obtain that

$$\nabla h(\bar{x}) - \nabla h(x^+) \in \partial g(x^+).$$

Since g is assumed to be proper, closed and convex, (2.2) yields

$$x^+ \in \partial g^*\left(\nabla h(\bar{x}) - \nabla h(x^+)\right). \tag{5.5}$$

Setting $\tilde{y} := \nabla h(\bar{x}) - \nabla h(x^+)$ and observing that $x^+ = \nabla h^*(\nabla h(\bar{x}) - \tilde{y})$ we can rewrite (5.5) as

$$\nabla h^*(\nabla h(\bar{x}) - \tilde{y}) \in \partial g^*(\tilde{y}).$$

Table 4 Bregman proximal operators—normal linear model ($h = \frac{1}{2}\|\cdot\|^2$)

Reference distribution (R)	Proximal operator ($x^+ = \mathrm{prox}_{t\psi_R^*}(\bar{x})$)
Multivariate normal ($\mu \in \mathbb{R}^d, \Sigma \in \mathbb{S}^d : \Sigma \succ 0$)	$x^+ = (tI + \Sigma)^{-1}(\Sigma\bar{x} + t\mu)$
Multivariate normal-inverse Gaussian$\big(\mu, \beta \in \mathbb{R}^d,\ \alpha, \delta \in \mathbb{R},$ $\Sigma \in \mathbb{R}^{d\times d} : \delta > 0,\ \Sigma \succ 0,$ $\alpha^2 \geq \beta^T\Sigma\beta, \gamma := \sqrt{\alpha^2 - \beta^T\Sigma\beta}\big)$	$x^+ = \left(I + \rho\Sigma^{-1}\right)^{-1}\left(t\beta + \bar{x} + \rho\Sigma^{-1}\mu\right)$, where $\rho \in \mathbb{R}_+ :$ $(\rho\delta)^2 + \left\| \left(\rho^{-1}I + \Sigma^{-1}\right)^{-1}(t\beta + \bar{x} - \mu) \right\|_{\Sigma^{-1}}^2 = (\alpha t)^2$
Gamma ($\alpha, \beta \in \mathbb{R}_{++}$)	$x^+ = \left(\bar{x} - t\beta + \sqrt{(\bar{x} - t\beta)^2 + 4t\alpha}\right)/2$
Laplace ($\mu \in \mathbb{R},\ b \in \mathbb{R}_{++}$)	$x^+ = \begin{cases} \mu, & \bar{x} = \mu, \\ \mu + b\rho, & \bar{x} \neq \mu, \end{cases}$ where $\rho \in \mathbb{R} :\quad \alpha_1\rho^3 + \alpha_2\rho^2 + \alpha_3\rho + \alpha_4 = 0,$ with $\alpha_1 = (b/t)^2 b^2,\ \alpha_2 = 2(b/t)^2 b(\mu - \bar{x}),$ $\alpha_3 = (b/t)^2(\mu - \bar{x})^2 - 2(b/t)b - 1,\ \alpha_4 = -2(b/t)(\mu - \bar{x})$
Poisson[11] ($\lambda \in \mathbb{R}_{++}$)	$x^+ = tW\left(\frac{\lambda e^{\bar{x}/t}}{t}\right)$
Multinomial ($n \in \mathbb{N}, p \in \Delta_{(d)} :$ $\sum_{i=1}^d p_i < 1$)	$x^+ \in \mathbb{R}_+^d \cap I(p) :\quad (x_i^+ - \bar{x}_i)/t + \log\left(\frac{x_i^+(1-\sum_{j=1}^d p_j)}{p_i(n-\sum_{j=1}^d x_j^+)}\right) = 0$
Negative multinomial ($p \in [0, 1)^d$, $x_0 \in \mathbb{R}_{++},\ p_0 := 1 - \sum_{i=1}^d p_i > 0$)	$x^+ \in \mathbb{R}_+^d \cap I(p) :\ (x_i^+ - \bar{x}_i)/t + \log\left(\frac{x_i^+}{p_i(x_0+\sum_{j=1}^d x_j^+)}\right) = 0,$
Discrete uniform ($a, b \in \mathbb{R} : a < b$)	$x^+ = \bar{x} - t\theta^+$ where $\theta^+ = 0$ if $\bar{x} = (a + b)/2$, otherwise: $\theta^+ \in \mathbb{R} \setminus \{0\}$: $t(\theta^+ - \bar{x}/t) + \frac{(b+1)e^{(b+1)\theta^+} - ae^{a\theta^+}}{e^{(b+1)\theta^+} - e^{a\theta^+}} = \frac{e^{\theta^+}}{e^{\theta^+} - 1}$
Continuous uniform ($a, b \in \mathbb{R} : a \leq b$)	$x^+ = \bar{x} - t\theta^+$ where $\theta^+ = 0$ if $\bar{x} = (a + b)/2$, otherwise: $\theta^+ \in \mathbb{R} \setminus \{0\}$: $t(\theta^+ - \bar{x}/t) + \frac{be^{b\theta^+} - ae^{a\theta^+}}{e^{b\theta^+} - e^{a\theta^+}} = \frac{1}{\theta^+}$
Logistic ($\mu \in \mathbb{R},\ s \in \mathbb{R}_{++}$) :	$x^+ = \bar{x} - t\theta^+$ where $\theta^+ = 0$ if $\bar{x} = \mu$, otherwise: $\theta^+ \in \mathbb{R} \setminus \{0\}$: $t\theta^+ + \frac{1}{\theta^+} + \frac{\pi s}{\tan(-\pi s\theta^+)} = \bar{x} - \mu$

[11] We denote by $W : \mathbb{R} \to \mathbb{R}$ the Lambert W function (see, for example, [31])

It is now easy to verify that the above is nothing else but the optimality condition for $\bar{y}$, thus, $\tilde{y} = y^+$ and we can conclude that $\nabla h(x^+) + y^+ = \nabla h(\bar{x})$, establishing the desired result. □

The following corollary adapts the above lemma to the setting considered in our study. Furthermore, we complement this result with a simple observation which is particularly useful for Bregman proximal operator computations.

Corollary 5.1 *Let $h : \mathbb{R}^d \to (-\infty, +\infty]$ be of Legendre type and let $R \in \mathcal{P}(\Omega)$ satisfy Assumptions A and B. Assume further that* $\mathrm{int}\,(\mathrm{dom}\, h) \cap \mathrm{dom}\, \psi_R^* \neq \emptyset$*. For*

$t > 0$ *and* $\bar{x} \in \operatorname{int}(\operatorname{dom} h)$*, let* $x^+ := \operatorname{prox}^h_{t\psi_R^*}(\bar{x})$ *and* $\theta^+ := \operatorname{iconv}^{h^*}_{t\psi_R(\cdot/t)}(\bar{x})$*. Then,* $\nabla h(x^+) + \theta^+ = \nabla h(\bar{x})$*. In particular,* $\theta^+ = 0$ *(and* $x^+ = \bar{x}$*) if and only if* $\bar{x} = \mathbb{E}_R$*.*

Proof By Theorem 3.2 we have that ψ_R^* is proper, closed and convex and thus $\psi_R^{**} = \psi_R$ due to [13, Theorem 4.8]. By Theorem 5.1 we know that x^+ is well-defined. The proof of the first part then follows directly from Lemma 5.1 (with $g = t\psi_R^*$ and $y^+ = \theta^+$) and [13, Theorem 4.14(a)]. To see that $\theta^+ = 0$ if and only if $\bar{x} = \mathbb{E}_R$, observe that the objective function in the subproblem defining the Bregman proximal operator (5.3) is greater equal than zero, and equality holds if and only if $\bar{x} = \mathbb{E}_R$ with $x^+ = \bar{x}$. Thus, the statement holds true in view of the first part of the current corollary. □

Tables 5 and 6 list the formulas of Bregman proximal operators for the Poisson and Gamma ($\beta = 1$) linear families, respectively. Observe that by Theorem 5.1 the Bregman proximal operator is well defined if $\operatorname{int}(\operatorname{dom} h) \cap \operatorname{dom} \psi_R^* \neq \emptyset$. Since $\operatorname{int}(\operatorname{dom} h) = \mathbb{R}^d_{++}$ this implies that for the multinomial and negative multinomial distributions we must assume that $p_i > 0$ for all $i = 1, 2, \ldots, d$. Furthermore, for the sake of simplicity, we include the normal and normal inverse-Gaussian distributions. The multivariate variants can be found in the software documentation along with further explanations.

5.3 Examples

We close our study with particular models and algorithms.

Barcode Image Deblurring. Restoration of a blurred and noisy image represented by a vector $\hat{y} \in \mathbb{R}^d$ can be cast as the following optimization problem:

$$\min\left\{\frac{1}{2}\|Ax - \hat{y}\|_2^2 + \tau\varphi_R^*(x) : x \in \mathbb{R}^d\right\}. \tag{5.6}$$

$A \in \mathbb{R}^{d\times d}$ is the blurring operator and $\tau > 0$ is a regularization parameter. The noise is assumed to be Gaussian which explains the least-squares fidelity term. As discussed in Sect. 4.2 (see the discussion after Remark 4.2), the least-squares model can be justified from the viewpoint of both the ML and, as we know from our study, the MEM framework. If the original image is a 2D barcode, a natural choice for the reference measure $R \in \mathcal{P}(\Omega)$ inducing φ_R^* is a separable Bernoulli distribution with $p = 1/2$ due to the binary nature of each pixel and no preference at each pixel to take either value.[10] Additional information (symbology) can be easily incorporated by an appropriate adjustment of the parameter for each known pixel (see [59]). Using the

[10] As mentioned in Remark 3.2, Bernoulli is a special case of the multinomial distribution. This, one dimensional, distribution is used to form a d-dimensional i.i.d as described in Remark 3.1.

Table 5 Bregman proximal operators—poisson linear m odel ($h_j(x) = x_j \log x_j$)

Reference distribution (R)	Bregman proximal operator ($x^+ = \mathrm{prox}^h_{t\psi^*_R}(\bar{x})$)
Normal ($\mu, \sigma \in \mathbb{R} : \sigma > 0$)	$x^+ = \frac{\sigma}{t} W\left(\frac{t}{\sigma}\bar{x}e^{\frac{t\mu}{\sigma}}\right)$
Normal-inverse Gaussian $\big(\mu, \alpha, \beta, \delta \in \mathbb{R} : \delta > 0,$ $\alpha \geq \lvert\beta\rvert,\ \gamma := \sqrt{\alpha^2 - \beta^2}\big)$	$x^+ \in \mathbb{R}_{++}:$ $(t\alpha/\sigma)(x^+ - \mu) = \big(t\beta - \log(x^+/\bar{x})\big)\sqrt{\delta^2 + (x^+ - \mu)^2/\sigma}$
Gamma ($\alpha, \beta \in \mathbb{R}_{++}$)	$x^+ = \frac{\alpha t}{W\left(\frac{\alpha t \exp(t\beta)}{\bar{x}}\right)}$
Laplace ($\mu \in \mathbb{R},\ b \in \mathbb{R}_{++}$)	$x^+ = \begin{cases} \mu, & \bar{x} = \mu, \\ \mu + b\rho, & \bar{x} \neq \mu, \end{cases}$ where $\rho \in \mathbb{R}: \quad \rho + \frac{2b}{t}\log\left(\frac{\mu+b\rho}{\bar{x}}\right) = \frac{b^2\rho}{t^2}\log^2\left(\frac{\mu+b\rho}{\bar{x}}\right)$
Poisson ($\lambda \in \mathbb{R}_{++}$)	$x^+ = \bar{x}^{1-\tau}\lambda^\tau \qquad (\tau := \frac{t}{t+1})$
Multinomial ($n \in \mathbb{N},\ p \in \mathrm{int}\,\Delta_{(d)}$)	$x_i^+ = \gamma_i\,(n - \rho)^\tau \quad \left(\tau := \frac{t}{t+1},\ \gamma_i := \left[\frac{p_i \bar{x_i}^{1/t}}{1 - \sum_{j=1}^d p_j}\right]^\tau\right)$ where $\rho \in \mathbb{R}:\ \rho = (n - \rho)^{\frac{t}{t+1}}\left(\sum_{i=1}^d \gamma_i\right)$
Negative multinomial ($p \in (0, 1)^d$, $x_0 \in \mathbb{R}_{++},\ p_0 := 1 - \sum_{i=1}^d p_i > 0$)	$x^+ \in \mathbb{R}^d_+ \cap I(p):\ \log\left(\frac{x_i^+}{\bar{x}_i}\right) + t\log\left(\frac{x_i^+}{p_i\,(x_0 + \sum_{j=1}^d x_j^+)}\right) = 0,$
Discrete uniform ($a, b \in \mathbb{R} : a < b$)	$x^+ = \bar{x}e^{-t\theta^+}$ where $\theta^+ = 0$ if $\bar{x} = (a+b)/2$, otherwise: $\theta^+ \in \mathbb{R} \setminus \{0\}:$ $\frac{(b+1)\exp((b+1)\theta^+) - a\exp(a\theta^+)}{\exp((b+1)\theta^+) - \exp(a\theta^+)} = \frac{\exp(\theta^+)}{\exp(\theta^+) - 1} + \exp(\bar{x} - t\theta^+ - 1)$
Continuous uniform ($a, b \in \mathbb{R} : a \leq b$)	$x^+ = \bar{x}e^{-t\theta^+}$ where $\theta^+ = 0$ if $\bar{x} = (a+b)/2$, otherwise: $\theta^+ \in \mathbb{R} \setminus \{0\}:$ $\frac{b\exp(b\theta^+) - a\exp(a\theta^+)}{\exp(b\theta^+) - \exp(a\theta^+)} = \frac{1}{\theta^+} + \exp(\bar{x} - t\theta^+ - 1)$
Logistic ($\mu \in \mathbb{R},\ s \in \mathbb{R}_{++}$)	$x^+ = \bar{x}e^{-t\theta^+}$ where $\theta^+ = 0$ if $\bar{x} = \mu$, otherwise: $\theta^+ \in \mathbb{R} \setminus \{0\}:$ $\frac{1}{\theta^+} + \frac{\pi s}{\tan(-\pi s\theta^+)} + \mu = \exp\left(\bar{x} - t\theta^+ - 1\right)$

appropriate proximal operator from Table 4, the BPG method for solving the model takes the form

$$x_i^{k+1} \in \mathbb{R}: \quad x_i^{k+1} + t\tau\log\left(\frac{x_i^{k+1}}{1 - x_i^{k+1}}\right) = x_i^k - t[A^T(Ax^k - \hat{y})]_i, \quad (i = 1, 2, \ldots, d).$$

As mentioned above, our focus on the Bregman proximal gradient method is only for illustration purposes. Favorable accelerated algorithms that employ the proximal operators derived in this work are readily available and should be used in practice. The acceleration scheme applicable here is known as the Fast Iterative Shrinkage Thresholding Algorithm (FISTA) [14].

Natural Image Deblurring. For natural image deblurring there is no obvious structure such as the binary one for barcodes. However, it is customary to assume that the image is piecewise smooth. A popular model that promotes piecewise constant restoration is the Rudin, Osher, and Fatemi (ROF) model [62] based on the total variation (TV) regularizer $\sum_{i=1}^d g(L_i x)$. Here, $L_i \in \mathbb{R}^{2\times d}$ extracts the difference between the pixel i

Table 6 Bregman proximal operators—Gamma ($\beta = 1$) Linear Model ($h_j(x) = -\log(x_j)$)

Reference distribution (R)	Bregman proximal operator ($x^+ = \operatorname{prox}^h_{t\psi^*_R}(\bar{x})$)
Normal ($\mu, \sigma \in \mathbb{R} :\ \sigma > 0$)	$x^+ = \left((t/\sigma)\mu - 1/\bar{x} + \sqrt{((t/\sigma)\mu - 1/\bar{x})^2 + 4(t/\sigma)}\right)/(2t/\sigma)$
Normal-inverse Gaussian $\big(\mu, \alpha, \beta, \delta \in \mathbb{R} :\ \delta > 0,$ $\alpha \geq \lvert\beta\rvert,\ \gamma := \sqrt{\alpha^2 - \beta^2}\big)$	$x^+ \in \mathbb{R}_{++}$: $t\alpha(x^+ - \mu)x^+ = \left((t\beta - 1/\bar{x})x^+ + 1\right)\sqrt{\delta^2 + (x^+ - \mu)^2}$
Multivariate normal-inverse Gaussian $\big(\mu, \beta \in \mathbb{R}^d,\ \alpha, \delta \in \mathbb{R},$ $\Sigma = \sigma I, \sigma > 0 : \delta > 0,\ \Sigma \succ 0,$ $\alpha^2 \geq \beta^T\Sigma\beta, \gamma := \sqrt{\alpha^2 - \beta^T\Sigma\beta}\big)$	$x_i^+ = (w_i + \rho\mu_i + \sqrt{(w_i + \rho\mu_i)^2 + 4\rho})/(2\rho),$ with $w_i = t\beta_i - 1/\bar{x}_i$ and $\rho \in \mathbb{R}_+$: $(\rho\delta)^2 + \frac{1}{4\sigma}\sum_{i=1}^d \left(w_i + \sqrt{(w_i + \mu_i\rho)^2 + 4\rho}\right)^2 = (\alpha t/\sigma)^2$
Gamma ($\alpha, \beta \in \mathbb{R}_{++}$)	$x^+ = \bar{x}(t\alpha + 1)/(\bar{x}t\beta + 1)$
Laplace ($\mu \in \mathbb{R},\ b \in \mathbb{R}_{++}$)	$x^+ = \begin{cases} \mu, & \bar{x} = \mu, \\ \mu + b\rho, & \bar{x} \neq \mu, \end{cases}$ where $\rho \in \mathbb{R} :\quad \alpha_1\rho^3 + \alpha_2\rho^2 + \alpha_3\rho + \alpha_4 = 0,$ with $\alpha_1 = b^2((b/\bar{x})^2 - t^2),\ \alpha_2 = 2b(\mu((b/\bar{x})^2 - t^2) - b^2(t+1)/\bar{x}),$ $\alpha_3 = b^2((1 - \mu/\bar{x})^2 + 2t(1 - 2\mu/\bar{x})) - t^2\mu^2,\ \alpha_4 = 2tb\mu(1 - \mu/\bar{x})$
Poisson ($\lambda \in \mathbb{R}_{++}$)	$x^+ \in \mathbb{R}_+ :\ t\log\left(\frac{x^+}{\lambda}\right) = \frac{1}{x^+} - \frac{1}{\bar{x}}$
Multinomial ($n \in \mathbb{N}$, $p \in \operatorname{ri}\Delta_{(d)}$)	$x^+ \in \operatorname{ri} n\Delta_{(d)} :\ t\log\left(\frac{x_i^+(1 - \sum_{j=1}^d p_j)}{p_i(n - \sum_{j=1}^d x_j^+)}\right) = \frac{1}{x_i^+} - \frac{1}{\bar{x}_i}$
Negative multinomial ($p \in (0,1)^d$, $x_0 \in \mathbb{R}_{++},\ p_0 := 1 - \sum_{i=1}^d p_i > 0$)	$x^+ \in \mathbb{R}^d_{++} :\ t\log\left(\frac{x_i^+}{p_i(x_0 + \sum_{i=j}^d x_j^+)}\right) = \frac{1}{x_i^+} - \frac{1}{\bar{x}_i},$
Discrete uniform ($a, b \in \mathbb{R} : a < b$)	$x^+ = \bar{x}/(\bar{x}t\theta^+ + 1)$ where $\theta^+ = 0$ if $\bar{x} = (a+b)/2$, otherwise: $\theta^+ \in \mathbb{R} \setminus \{0\}$: $\frac{(b+1)\exp((b+1)\theta) - a\exp(a\theta)}{(\exp((b+1)\theta) - \exp(a\theta)} = \frac{\exp(\theta)}{\exp(\theta) - 1} + \frac{\bar{x}}{t\bar{x}\theta^+ + 1}$
Continuous uniform ($a, b \in \mathbb{R} : a \leq b$)	$x^+ = \bar{x}/(\bar{x}t\theta^+ + 1)$ where $\theta^+ = 0$ if $\bar{x} = (a+b)/2$, otherwise: $\theta^+ \in \mathbb{R} \setminus \{0\}$: $\frac{b\exp(b\theta^+) - a\exp(a\theta^+)}{\exp(b\theta^+) - \exp(a\theta^+)} = \frac{1}{\theta^+} + \frac{\bar{x}}{t\bar{x}\theta^+ + 1}$
Logistic ($\mu \in \mathbb{R},\ s \in \mathbb{R}_{++}$)	$x^+ = \bar{x}/(\bar{x}t\theta^+ + 1)$ where $\theta^+ = 0$ if $\bar{x} = \mu$, otherwise: $\theta^+ \in \mathbb{R} \setminus \{0\}$: $\frac{1}{\theta^+} + \frac{\pi s}{\tan(-\pi s\theta^+)} + \mu = \frac{\bar{x}}{\bar{x}t\theta^+ + 1}$

and two adjacent pixels while g stands for either the l_1 (isotropic TV) or l_2 (anisotropic TV) norm. Variants that admit the same structure with other choices of g are also considered in the literature: in [25, Subsection 6.2.3], a model with the Huber norm for g was shown to promote restoration prone to artificial flat areas. Alternatively, one may consider the pseudo-Huber norm that corresponds to an MEM regularizer induced by the multivariate normal inverse-Gaussian reference distribution with parameters $\mu = \beta = 0$, $\alpha = 1$, and $\Sigma = I$. The resulting model is similar to (5.6) where the regularization term is substituted by $\sum_{i=1}^d \psi_R^*(L_i x)$. This model can be tackled by a primal-dual decomposition method that employs the appropriate proximal operator from Table 4. For example, using the separability of the proximal operator [13, Theorem 6.6] and the extended Moreau decomposition [13, Theorem 6.45], the update formula of the Chambolle-Pock algorithm [25, Algorithm 1] reads

$$\begin{aligned}
y_i^{k+1} &= \tfrac{\rho_i}{1+\rho_i}(y^k + sL_i z^k) \qquad\qquad (i = 1, 2, \ldots, d),\\
&\text{with } \rho_i \in \mathbb{R}_+ : \rho_i^2 (s\delta)^2 + \left(\tfrac{\rho_i}{1+\rho_i}\right)^2 \|y_i^k + sL_i z^k\|_2^2 = 1,\\
x^{k+1} &= (I + \tau A^T A)^{-1} \left(x^k - \tau(L^T y^{k+1} - A^T \hat{y})\right),\\
z^{k+1} &= 2x^{k+1} - x^k,
\end{aligned}$$

where $L^T = [L_1^T, \ldots, L_d^T] \in \mathbb{R}^{d\times 2d}$, $y^k \in \mathbb{R}^{2d} : (y^k)^T = [(y_1^k)^T, \ldots, (y_d^k)^T]$ with $y_i^k \in \mathbb{R}^2$ for all $i = 1, 2, \ldots, d$) and s, τ are some positive step-sizes satisfying $s\tau \|L\|_2^2 < 1$.

We point out that an efficient implementation of the above algorithm that takes into account the sparse and structured nature of the matrices L and A, respectively, will result in a per-iteration complexity of the order $O(d \log d)$. The same statement is true with regard to the BPG method in the previous and following examples.

Poisson Linear Inverse Problem. Poisson linear inverse problems play a prominent role in various physical and medical imaging applications. The linear model proposed in [8, Subsection 5.3] is simply the MEM linear model with Poisson reference distribution. The authors of [8] suggest l_1-regularization to deploy their BPG method. Alternatively, one may consider the MEM function induced by the Laplace distribution with parameters $\mu = 0$ and $b = 1$. This setting leads to the following update formula of the BPG method. For $i = 1, 2, \ldots, d$:

$$\bar{x}_i^{k+1} = \exp\left(\log(x_i^k) - t\sum_{j=1}^m a_{ji}\log(\langle a_j, x^k\rangle/\hat{y}_j)\right),$$
$$x_i^{k+1} \in \mathbb{R} : t^2 x_i^{k+1} + 2t\log\left(\frac{x_i^{k+1}}{\bar{x}_i^{k+1}}\right) = x_i^{k+1}\left[\log\left(\frac{x_i^{k+1}}{\bar{x}_i^{k+1}}\right)\right]^2.$$

Matrix Games. Consider the formulation of matrix games with mixed strategies[11]

$$\min_{x\in\Delta_{(n)}} \max_{y\in\Delta_{(m)}} \{\langle x, Ay\rangle + \langle a_r, x\rangle + \langle a_c, y\rangle\}, \tag{5.7}$$

where $a_r \in \mathbb{R}^m$, $a_c \in \mathbb{R}^n$ and $A \in \mathbb{R}^{m\times n}$ such that $\|A\|_1 = 1$.[12]

The vectors x and y are known as the mixed strategies of the row and column players, respectively. These vectors represent a measure over the pure strategies that can be interpreted as a categorical distribution prior.

[11] The matrix game formulation presented in (5.7) is somewhat non-standard as the feasible set is given by $\Delta_{(n)} \times \Delta_{(m)}$ (cf. definition after Remark 3.1) instead of $\Delta_n \times \Delta_m$. Clearly, the classical formulation can be recovered by an introduction of slack variables to (5.7).

[12] The assumption that the l_1-norm of the matrix A equals to one is not essential; we make it solely to simplify our presentation.

Alternatively, one can consider a model that uses the MEM functions $\psi^*_{P_r}$ and $\psi^*_{P_c}$ as priors for the mixed strategy vectors of the row and column players, respectively. The measures P_r and P_c that induce the MEM functions are given by a categorical distribution[13] with parameter vectors $(1/(m+1), \dots, 1/(m+1))^T \in \mathbb{R}^m$ and $(1/(n+1), \dots, 1/(n+1))^T \in \mathbb{R}^n$ for the row and column players, respectively. This is to emphasize that no additional prior knowledge on the strategies is available. Nevertheless, additional prior knowledge can be easily incorporated by updating the parameter vectors if such a knowledge is available. The resulting model takes the following form:

$$\min_{x\in\Delta_{(n)}} \max_{y\in\Delta_{(m)}} \left\{\langle x, Ay\rangle + \langle a_r, x\rangle + \psi^*_{P_r}(x) + \langle a_c, y\rangle - \psi^*_{P_c}(y)\right\}.$$

Unlike the classical formulation of matrix games, the proposed MEM-regularized model guarantees a unique solution due to the strict convexity of the MEM function.

The above can be solved by the primal-dual algorithms proposed in [26]. These algorithms are based on a general update rule that for a given $(x, y) \in \Delta_{(m)} \times \Delta_{(n)}$ generates

$$\begin{aligned} x^+ &= \operatorname{prox}^{h_r}_{t_r\psi^*_{P_r}}(\bar{x}), \text{ with } \bar{x} = \nabla h_r^*(\nabla h_r(x) - (Ay + a_r)),\\ \text{and } y^+ &= \operatorname{prox}^{h_c}_{t_c\psi^*_{P_r}}(\bar{y}), \text{ with } \bar{y} = \nabla h_c^*(\nabla h_c(y) + (A^T x + a_c)), \end{aligned} \tag{5.8}$$

where $t_r, t_c \in \mathbb{R}_{++}$ are some given step-size parameters and h_r, h_c are the kernel functions that induce the Bregman proximal operator. Since (up to the MEM function terms) the objective function is linear, the choice of kernels h_r and h_c can be arbitrary. Choosing the kernel $h(x) = x\log(x)$ and $t_r = t_c = 1$ (see [26, Remark 1] and the fact that our choice of h is 1-strongly convex with respect to the l_1 norm) we can use the equation from Table 5 to get a closed form expression to the update rule (5.8).[14] For each $i = 1, 2, \dots, m$ and $j = 1, 2, \dots, n$

$$\begin{aligned} \bar{x}_i &= \exp(\log(x_i) - (Ay + a_r)_i),\\ \bar{y}_j &= \exp(\log(y_j) + (A^T x + a_c)_j),\\ x_i^+ &= \sqrt{(1 + 0.5(\gamma_r - \sqrt{\gamma_r^2 + 4\gamma_r}))\bar{x}_i}, \text{ where } \gamma_r = \left(\textstyle\sum_{i=1}^m \sqrt{\bar{x}_i}\right)^2,\\ y_j^+ &= \sqrt{(1 + 0.5(\gamma_c - \sqrt{\gamma_c^2 + 4\gamma_c}))\bar{y}_i}, \text{ where } \gamma_c = \left(\textstyle\sum_{i=1}^n \sqrt{\bar{y}_i}\right)^2. \end{aligned}$$

An implementation of some of the above algorithms along with an illustration for applications in image processing are available with the MEM Python package accompanying this work.

[13] Recall from Remark 3.2 that the categorical distribution is a special case of the multinomial one.

[14] The choice $h(x) = x\log(x)$ corresponds to the Poisson MEM linear model (recall Table 3). Nevertheless, here this choice is arbitrary and unrelated to the Poisson MEM linear model. The role of MEM in this example is confined to the priors.

6 Conclusion

Many of the inherent properties of the maximum entropy on the mean (MEM) function, such as differentiability and strong convexity, motivated its use for modeling and solving inverse problems in different fields of science and engineering. Nevertheless, this method hasn't gained widespread acceptance as a mainstream statistical tool. One of the possible contributing factors to this outcome could have been the variational definition of the MEM function which adds considerable complexity to the optimization process of MEM models. This work aims to overcome this barrier by introducing an optimization toolbox specifically tailored for MEM functions. To achieve this, a novel proof of the well-known equivalence between the MEM and Cramér's rate functions that hold under very mild and natural assumptions is provided. This clears the way for an alternative representation that gives rise to explicit (or tractable) formulas for the MEM function induced by many popular distributions. We highlight Bregman proximal gradient (BPG) as an exceptionally well-suited method for addressing the class of MEM linear models. We provide a library of Bregman proximal operators that cover over a dozen statistical linear MEM models. These operators are not limited to BPG and can be used within many other modern optimization methods. A software package with an implementation of most of the proposed operators complements this work. It is our hope that this research will serve to push forward the utilization and further study of the MEM method.

A Cramér rate functions

We present here the computations of all Cramér rate functions provided during our study. To this end, recall that the Cramér rate function ψ_P^* is the conjugate

$$\psi_P^*(y) := \sup\{\langle y, \theta\rangle - \psi_P(\theta) : \theta \in \mathbb{R}^d\}$$

of the cumulant-generating function

$$\psi_P(\theta) := \log M_P[\theta],$$

where $M_P[\theta]$ is the moment-generating function of the reference distribution $P \in \mathcal{P}(\Omega)$ (which one can simply look up at various places in the literature for the distributions considered here).

Multivariate normal

For a normal distribution with mean μ and covariance $\Sigma \succ 0$, its moment generating function is $M_P[\theta] = \exp(\langle \mu, \theta\rangle + \frac{1}{2}\langle \theta, \Sigma\theta\rangle)$. Therefore, we find

$$\begin{aligned}\psi_P^*(y) &= \sup\left\{\langle y, \theta\rangle - \log\left(\exp(\langle \mu, \theta\rangle + \tfrac{1}{2}\langle \theta, \Sigma\theta\rangle)\right) : \theta \in \mathbb{R}^d\right\}\\ &= \sup\left\{\langle y, \theta\rangle - \langle \mu, \theta\rangle - \tfrac{1}{2}\langle \theta, \Sigma\theta\rangle : \theta \in \mathbb{R}^d\right\}.\end{aligned}$$

The maximumizer of the above quadratic optimization problem is $\theta^* = \Sigma^{-1}(y-\mu)$, hence

$$\psi_P^*(y) = \frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu).$$

Multivariate normal-inverse Gaussian

The Multivariate Normal-inverse Gaussian distribution is defined by means of location ($\mu \in \mathbb{R}^d$), tail heaviness ($\alpha \in \mathbb{R}$), asymmetry ($\beta \in \mathbb{R}^d$), and scale ($\delta \in \mathbb{R},\ \Sigma \in \mathbb{R}^{d\times d}$) parameters satisfying $\alpha \geq \sqrt{\langle \beta, \Sigma\beta\rangle}$, $\delta > 0$ and $\Sigma \succ 0$ [6]. In addition, let $\gamma := \sqrt{\alpha^2 - \langle \beta, \Sigma\beta\rangle}$. Its moment-generating function is

$$M_P[\theta] = \exp\left(\langle \mu, \theta\rangle + \delta(\gamma - \sqrt{\alpha^2 - \langle \beta+\theta, \Sigma(\beta+\theta)\rangle})\right) \quad (\theta \in B_\alpha),$$

for the ellipsoid $B_\alpha = \{\theta \in \mathbb{R}^d : \sqrt{\langle \beta+\theta, \Sigma(\beta+\theta)\rangle} \leq \alpha\}$. Observe that in this case $\psi_P(\theta) = \log(M_P[\theta])$ is indeed steep and minimal.

Now, in order to compute the Cramér rate function, we find that

$$\psi_P^*(y) = \sup\left\{\langle y-\mu, \theta\rangle - \delta(\gamma - \sqrt{\alpha^2 - \langle \beta+\theta, \Sigma(\beta+\theta)\rangle}) : \theta \in B_\alpha\right\} \tag{A.1}$$

We consider two cases: if $y = \mu$, then it is evident that the optimal solution of the problem above is given by $\theta = -\beta$ and thus $\psi_P^*(\mu) = \delta(\alpha - \gamma)$. Consider the case $y \neq \mu$. Disregarding the feasibility constraints (which will be justified in the sequel), the first-order optimality condition is given by

$$y - \mu = \frac{\delta\Sigma(\beta+\theta)}{\sqrt{\alpha^2 - \langle \beta+\theta, \Sigma(\beta+\theta)\rangle}}.$$

From the above, we can derive

$$\begin{aligned}\langle \beta+\theta, \Sigma(\beta+\theta)\rangle = \frac{\alpha^2\langle y-\mu, \Sigma^{-1}(y-\mu)\rangle}{\delta^2 + \langle y-\mu, \Sigma^{-1}(y-\mu)\rangle} \quad \text{and} \quad \theta\\ = -\beta + \frac{\alpha\Sigma^{-1}(y-\mu)}{\sqrt{\delta^2 + \langle y-\mu, \Sigma^{-1}(y-\mu)\rangle}}.\end{aligned}$$

It is straightforward to verify that $\theta \in \operatorname{int} B_\alpha$, which retroactively justifies our choice to disregard the constraint before. Now, we can write the Cramér rate function as

$$\begin{aligned}
&\psi_P^*(y)\\
&= \langle y-\mu, -\beta + \frac{\alpha\Sigma^{-1}(y-\mu)}{\sqrt{\delta^2 + \langle y-\mu, \Sigma^{-1}(y-\mu)\rangle}}\rangle\\
&\quad -\delta\left(\gamma - \sqrt{\alpha^2 - \tfrac{\alpha^2\langle y-\mu, \Sigma^{-1}(y-\mu)\rangle}{\delta^2+\langle y-\mu, \Sigma^{-1}(y-\mu)\rangle}}\right)\\
&= \alpha\sqrt{\delta^2 + (y-\mu)^T\Sigma^{-1}(y-\mu)} - \langle \beta, y-\mu\rangle - \delta\gamma.
\end{aligned}$$

Gamma

The Gamma distribution is parametrized by $\alpha, \beta > 0$ and its moment generating function is given by

$$M_P[\theta] = \left[1 - \frac{\theta}{\beta}\right]^{-\alpha} \quad (\theta < \beta).$$

Hence, its Cramér rate function reads

$$\begin{aligned}
\psi_P^*(y) &= \sup\left\{y\theta - \log\left(\left[1 - \tfrac{\theta}{\beta}\right]^{-\alpha}\right) : \theta < \beta\right\}\\
&= \sup\left\{y\theta + \alpha\log\left(1 - \tfrac{\theta}{\beta}\right) : \theta < \beta\right\}.
\end{aligned}$$

If $y \leq 0$, then $\psi_P^*(y) = +\infty$ (with $\theta \to -\infty$). If $y > 0$ then the first-order optimality conditions imply

$$y - \frac{\alpha}{\beta}\left(1 - \frac{\theta}{\beta}\right)^{-1} = 0 \quad \Rightarrow \quad \theta = \beta - \frac{\alpha}{y}.$$

Thus,

$$\psi_P^*(y) = \beta y - \alpha + \alpha\log\left(\frac{\alpha}{\beta y}\right), \quad y \in \mathbb{R}_{++}.$$

Laplace

The Laplace distribution is parameterized by its mean $\mu \in \mathbb{R}$ and scale $b > 0$. Its MGF reads

$$M_P[\theta] = \frac{\exp(\mu\theta)}{1 - b^2\theta^2} \quad (|\theta| < 1/b)$$

Hence, its Cramér rate function reads

$$\psi_P^*(y) = \sup\left\{(y-\mu)\theta + \log\left(1 - b^2\theta^2\right) : |\theta| < 1/b\right\}.$$

It is easy to see that $\log\left(1 - b^2\theta^2\right) \leq 0$ for any θ such that $|\theta| < 1/b$ and that $\log\left(1 - b^2\theta^2\right) \to -\infty$ when $|\theta| \to 1/b$. Thus, we can conclude that $\psi_P^*(\mu) = 0$ and for any $y \neq \mu$ the maximum of the above problem is attained at some point in the open interval $(0, 1/b)$ for $y > \mu$ or in $(-1/b, 0)$ for $y < \mu$. The first-order optimality conditions boil down to the quadratic equation

$$\theta^2 + \left(\frac{2}{y-\mu}\right)\theta - \frac{1}{b^2} = 0$$

Evaluating the roots of the resulting quadratic equation we conclude that the optimal solution is

$$\theta = \frac{1}{y-\mu}\left(\sqrt{1 + \left(\frac{y-\mu}{b}\right)^2} - 1\right) = \frac{1}{b\rho}\left(\sqrt{1+\rho^2} - 1\right),$$

where we set $\rho := \frac{y-\mu}{b}$. Evidently, $|\theta| < 1/b$ holds for the solution we just derived. Thus

$$\begin{aligned}
\psi_P^*(y) &= (y-\mu)\theta + \log\left(1 - (b\theta)^2\right) \\
&= \rho(b\theta) + \log\left(1 - (b\theta)^2\right) \\
&= \sqrt{1+\rho^2} - 1 + \log\left(1 - \tfrac{1}{\rho^2}(\sqrt{1+\rho^2} - 1)^2\right) \\
&= \sqrt{1+\rho^2} - 1 + \log\left(1 - \tfrac{1}{\rho^2}(1 + \rho^2 + 1 - 2\sqrt{1+\rho^2})\right) \\
&= \sqrt{1+\rho^2} - 1 + \log\left(\tfrac{2}{\rho^2}(\sqrt{1+\rho^2} - 1)\right),
\end{aligned}$$

and we can conclude that

$$\psi_P^*(y) = \begin{cases} 0, & y = \mu, \\ \sqrt{1 + \left(\frac{y-\mu}{b}\right)^2} - 1 + \log\left(2\left(\frac{y-\mu}{b}\right)^{-2}\left[\sqrt{1 + \left(\frac{y-\mu}{b}\right)^2} - 1\right]\right), & y \neq \mu. \end{cases}$$

Poisson

The Poisson distribution is parameterized by its rate $\lambda > 0$. Its MGF reads

$$M_P[\theta] = \exp(\lambda(\exp(t) - 1)$$

Consequently, its Cramér rate function is given by

$$\psi_P^*(y) = \sup\{y\theta - \lambda(\exp(\theta) - 1) : \theta \in \mathbb{R}\}.$$

If $y < 0$ then it is evident from the above that $\psi_P^*(y) = +\infty$ (indeed, take $\theta \to -\infty$). Similarly, we can see that $\psi_P^*(0) = \lambda$. Otherwise, due to the first-order optimality conditions

$$y = \lambda\exp(\theta) \quad \Rightarrow \quad \theta = \log(y/\lambda),$$

we obtain that $\psi_P^*(y) = y\log(y/\lambda) - y + \lambda$.

Multinomial

We will use the following notation. The ith canonical unit vector is denoted by e_i and the vector of all ones is denoted by e. The unit simplex is given by $\Delta_d := \{y \in \mathbb{R}_+^d : \langle e, y\rangle = 1\}$.

For $n \in \mathbb{N}$ and $p \in \Delta_{d+1}$ we can write

$$\begin{aligned}\psi_P^*(\theta) &= \sup\left\{l(y,\theta) := \langle y,\theta\rangle - \log\left(M_P[\theta]\right) : \theta \in \mathbb{R}^{d+1}\right\}\\ &= \sup\left\{\langle y,\theta\rangle - n\log\left(\sum_{i=1}^{d+1} p_i\exp(\theta_i)\right) : \theta \in \mathbb{R}^{d+1}\right\}.\end{aligned}$$

Let $I(p) := \left\{y \in \mathbb{R}^{d+1} : y_i = 0 \ (p_i = 0,\ i = 1,2,\ldots,d+1)\right\}$. We can see that $\operatorname{dom}\psi_P^* = n\Delta_d \cap I(p)$. Indeed, if there exists $k \in \{1,2,\ldots,d+1\}$ such that $y_k < 0$ then by setting $\theta = -\alpha e_k$ we obtain that

$$l(y,\theta) = \alpha|y_k| - n\log\left(p_k\exp(-\alpha) + \sum_{i\neq k} p_i\right).$$

If, $y \in \mathbb{R}^{d+1}$ but $\langle e, y\rangle \neq n$ then by choosing $\theta = \alpha\sigma e$ where $\sigma = \operatorname{sign}(\langle e, y\rangle - n)$ we obtain that

$$l(y,\theta) = \alpha\sigma\langle e, y\rangle - n\log\left(\exp(\alpha\sigma)\langle e, p\rangle\right) = \alpha|\langle e, y\rangle - n|.$$

If there exists $k \in \{i \in \{1,2,\ldots,d+1\} : p_i = 0\}$ such that $y_k > 0$ then by setting $\theta = \alpha e_k$ we obtain

$$l(y,\theta) = \alpha y_k - n\log\left(\sum_{i\neq k} p_i\right).$$

In all cases, by taking $\alpha \to \infty$ it is evident that the problem is unbounded.

We now address the case when $y \in \operatorname{dom} \psi_P^* = n\Delta_{d+1} \cap I(p)$. From the first-order optimality condition, we can deduce that for any $j = 1, \dots, d+1$ such that $p_j > 0$

$$y_j = \frac{np_j \exp(\theta_j)}{\sum_{i=1}^{d+1} p_i \exp(\theta_i)} \quad \Rightarrow \quad \theta_j = \log\left(\frac{y_j}{np_j}\right),$$

for all $j = 1, 2, \dots, d+1$. Thus, under the convention that $0/0 = 1$, we can conclude that for $y \in n\Delta_{d+1} \cap I(p)$

$$\psi_P^*(y) = \sum_{i=1}^{d+1} y_i \log\left(\frac{y_i}{np_i}\right).$$

Cramér's rate function that corresponds to the multinomial distribution after reduction to a minimal form can be obtained from the above by eliminating one component of the vectors $y \in \mathbb{R}^{d+1}$ and $p \in \mathbb{R}^{d+1}$. Assuming, without the loss of generality, that $p_{d+1} > 0$ we can plug in the above

$$y_{d+1} = n - \sum_{i=1}^{d} y_i, \qquad \text{and} \qquad p_{d+1} = 1 - \sum_{i=1}^{d} p_i,$$

in order to obtain the Cramér rate function $\psi_P^* : \mathbb{R}^d \to (-\infty, +\infty]$. Hence, for $y \in \mathbb{R}^d$ and $p \in \Delta_{(d)} := \{z \in \mathbb{R}_+^d : \langle e, z\rangle \leq 1\}$ such that $\langle e, p\rangle < 1$

$$\psi_P^*(y) = \sum_{i=1}^{d} y_i \log\left(\frac{y_i}{np_i}\right) + (n - \langle e, y\rangle) \log\left(\frac{n - \langle e, y\rangle}{n(1 - \langle e, p\rangle)}\right),$$

where, in this case, $\operatorname{dom} \psi_P^* = I(p) \cap \Delta_{(d)}$.

Negative multinomial

Observing that $\Theta_P := \{\theta \in \mathbb{R}^d : \sum_{i=1}^{d} p_i \exp(\theta_i) < 1\}$ and using the definition of Cramér's rate function we can write

$$\begin{aligned}
\psi_P^*(\theta) &= \sup\left\{l(y, \theta) := \langle y, \theta\rangle - \log\left(M_P[\theta]\right) : \theta \in \mathbb{R}^d\right\} \\
&= \sup\left\{\langle y, \theta\rangle - \log\left(\left[\frac{p_0}{1 - \sum_{i=1}^{d} p_i \exp(\theta_i)}\right]^{y_0}\right) : \theta \in \Theta_P\right\} \\
&= \sup\left\{\langle y, \theta\rangle + y_0 \log\left(1 - \sum_{i=1}^{d} p_i \exp(\theta_i)\right) : \theta \in \Theta_P\right\} - y_0 \log(p_0).
\end{aligned}$$

Let $I(p) := \left\{y \in \mathbb{R}^d : y_i = 0\ (p_i = 0,\ i = 1, 2, \ldots, d)\right\}$. We can see that $\operatorname{dom} \psi_P^* = \mathbb{R}_+^d \cap I(p)$. Indeed, if there exists $k \in \{1, \ldots, d\}$ such that $y_k < 0$ then by setting $\theta = -\alpha e_k$ (recall that e_k stands for the kth canonical unit vector) we obtain that

$$l(y, \theta) + y_0 \log(p_0) = \alpha |y_k| + y_0 \log \left(1 - p_k \exp(-\alpha) - \sum_{i \neq k} p_i\right).$$

If there exists $k \in \{i \in \{1, 2, \ldots, d\} : p_i = 0\}$ such that $y_k > 0$ then by setting $\theta = \alpha e_k$ we obtain that

$$l(y, \theta) + y_0 \log(p_0) = \alpha y_k + y_0 \log \left(1 - \sum_{i \neq k} p_i\right).$$

In both cases, by taking $\alpha \to \infty$ it is evident that the problem is unbounded.

We now address the case when $y \in \operatorname{dom} \psi_P^* = \mathbb{R}_+^d \cap I(p)$. From the first-order optimality condition, we can deduce that

$$y_j = \frac{y_0 p_j \exp(\theta_j)}{1 - \sum_{i=1}^d p_i \exp(\theta_i)} \quad \Rightarrow \quad \frac{y_j}{y_0}\left(1 - \sum_{i=1}^d p_i \exp(\theta_i)\right) = p_j \exp(\theta_j), \tag{A.2}$$

for all $j = 1, 2, \ldots, d$. Denoting $\sigma := \sum_{i=1}^d p_i \exp(\theta_i)$, $\bar{y} := \sum_{i=0}^d y_i$ and summing (A.2) for $j = 1, 2, \ldots, d$ yields

$$(\bar{y} - y_0)\left(\frac{1 - \sigma}{y_0}\right) = \sigma \qquad \Rightarrow \qquad \sigma = \frac{\bar{y} - y_0}{\bar{y}}.$$

The above, combined with (A.2) we obtain that for any $j = 1, 2, \ldots, d$ such that $p_j \neq 0$

$$\theta_j = \log\left(\frac{y_j}{p_j \bar{y}}\right).$$

Thus, we can conclude that for $y \in \mathbb{R}_+^d \cap I(p)$

$$\psi_P^*(y) = \sum_{i=1}^d y_i \log\left(\frac{y_i}{p_i \bar{y}}\right) + y_0 \log\left(\frac{y_0}{\bar{y}}\right) - y_0 \log(p_0) = \sum_{i=0}^d y_i \log\left(\frac{y_i}{p_i \bar{y}}\right).$$

It is important to note that in the above $y \in \mathbb{R}^d$ is the function variable while $y_0 \in \mathbb{R}$ is a fixed parameter.

Discrete uniform

The discrete uniform distribution is parameterized by $a, b \in \mathbb{Z}$ with $a \leq b$. We set $\mu := (a+b)/2$ and $n := b-a+1$. Its MGF reads

$$M_P[\theta] = \begin{cases} \frac{\exp((b+1)\theta)-\exp(a\theta)}{n(\exp(\theta)-1)}, & \theta \neq 0, \\ 1, & \theta = 0. \end{cases}$$

If $b = a$ then it is straightforward to verify that $\psi_P^* = \delta_{\{a\}}$ (degenerate distribution). We now turn to consider the case $b > a$. Since $M_P[\theta]$ is continuous at zero, we have

$$\begin{aligned} \psi_P^*(y) &= \sup\left\{ y\theta - \log\left(\frac{\exp((b+1)\theta)-\exp(a\theta)}{n(\exp(\theta)-1)}\right) : \theta \in \mathbb{R} \right\} \\ &= \sup\left\{ (y-b)\theta - \log\left(\frac{\exp(\theta)-\exp(-(b-a)\theta)}{n(\exp(\theta)-1)}\right) : \theta \in \mathbb{R} \right\} \\ &= \sup\left\{ (y-a)\theta - \log\left(\frac{\exp((b-a+1)\theta)-1}{n(\exp(\theta)-1)}\right) : \theta \in \mathbb{R} \right\} \\ &= \sup\left\{ (y-\mu)\theta - \log\left(\frac{\exp((b-\mu+1)\theta)-\exp((a-\mu)\theta)}{n(\exp(\theta)-1)}\right) : \theta \in \mathbb{R} \right\}. \end{aligned} \tag{A.3}$$

If $y > b$ then from the second formulation above we can conclude that $\psi_P^*(y) = +\infty$ by taking $\theta \to +\infty$. Similarly, if, $y < a$, then from the third formulation above we can conclude that $\psi_P^*(y) = +\infty$ by taking $\theta \to -\infty$. If $y = \mu$ then the last formulation of (A.3) can be written as

$$\sup\left\{ -\log\left(\frac{\exp(\gamma\theta)-\exp(-\gamma\theta)}{2\gamma(\exp(\theta/2)-\exp(-\theta/2))}\right) : \theta \in \mathbb{R} \right\} = -\log\left(\inf\{\phi(\theta) : \theta \in \mathbb{R}\}\right),$$

where $\gamma := (b-a+1)/2 > 1/2$ and

$$\phi(\theta) := \begin{cases} \frac{\exp(\gamma\theta)-\exp(-\gamma\theta)}{2\gamma(\exp(\theta/2)-\exp(-\theta/2))}, & \theta \neq 0, \\ 1, & \theta = 0. \end{cases}$$

By using L'Hôpital's rule and some straightforward arguments, it is easy to verify that

$$\lim_{|\theta|\to+\infty} \phi(\theta) = +\infty, \quad \lim_{|\theta|\to 0} \phi(\theta) = 1 \quad \text{and} \quad \phi(\theta) = \phi(-\theta).$$

Thus, ϕ is continuous at zero (which justifies its definition), coercive and symmetric. Since the log-normalizer function $\psi_P(\theta) = \log(M_P[\theta])$ is strictly convex, we conclude that if a solution exists it must be unique. The coercivity of ϕ implies that a solution exists, and due to the symmetry of ϕ we can conclude that it must be zero. To summarize, in this case, $\psi_P^*(\mu) = 0$ (with $\theta = 0$). If $y \neq \mu$ such that $a \leq y \leq b$ then the optimal solution to (A.3) is nonzero and by the first-order optimality conditions it must satisfy

$$y - \frac{(b+1)\exp((b+1)\theta) - a\exp(a\theta)}{\exp((b+1)\theta) - \exp(a\theta)} + \frac{\exp(\theta)}{\exp(\theta)-1} = 0. \tag{A.4}$$

Therefore, using (A.3) we can summarize that for $y \in [a,b] = \operatorname{dom} \psi_P^*$:

$$\psi_P^*(y) = \begin{cases} 0, & y = \mu, \\ (y-\mu)\theta - \log\left(\frac{\exp((b-\mu+1)\theta)-\exp((a-\mu)\theta)}{n(\exp(\theta)-1)}\right), & y \neq \mu, \end{cases}$$

where θ is the root of (A.4).

Continuous uniform

By definition

$$\psi_P^*(y) = \sup\left\{y\theta - \log\left(M_P[\theta]\right) : \theta \in \mathbb{R}\right\},$$

where for $a < b$ we have that

$$M_P[\theta] = \begin{cases} \frac{\exp(b\theta)-\exp(a\theta)}{(b-a)\theta}, & \theta \neq 0, \\ 1, & \theta = 0. \end{cases}$$

Since $M_P[\theta]$ is continuous at zero, then, without loss of generality, we obtain

$$\begin{aligned} \psi_P^*(y) &= \sup\left\{y\theta - \log\left(\frac{\exp(b\theta)-\exp(a\theta)}{(b-a)\theta}\right) : \theta \in \mathbb{R}\right\} \\ &= \sup\left\{(y-b)\theta - \log\left(\frac{1-\exp(-(b-a)\theta)}{(b-a)\theta}\right) : \theta \in \mathbb{R}\right\} \\ &= \sup\left\{(y-a)\theta - \log\left(\frac{\exp((b-a)\theta)-1}{(b-a)\theta}\right) : \theta \in \mathbb{R}\right\} \\ &= \sup\left\{(y-\mu)\theta - \log\left(\frac{\exp((b-\mu)\theta)-\exp((a-\mu)\theta)}{(b-a)\theta}\right) : \theta \in \mathbb{R}\right\}. \end{aligned} \tag{A.5}$$

where $\mu = (a+b)/2$. If $y \geq b$ then from the second formulation above we can conclude that $\psi_P^*(y) = \infty$ by taking $\theta \to \infty$. Similarly, if, $y \leq a$, then from the third formulation above we can conclude that $\psi_P^*(y) = \infty$ by taking $\theta \to -\infty$. If $y = \mu$ then the last formulation of (A.5) can be written as

$$\sup\left\{-\log\left(\frac{\exp(\gamma\theta)-\exp(-\gamma\theta)}{2\gamma\theta}\right) : \theta \in \mathbb{R}\right\} = -\log\left(\inf\left\{\phi(\theta) : \theta \in \mathbb{R}\right\}\right),$$

where $\gamma := (b-a)/2 > 0$ and

$$\phi(\theta) := \begin{cases} \frac{\exp(\gamma\theta)-\exp(-\gamma\theta)}{2\gamma\theta}, & \theta \neq 0, \\ 1, & \theta = 0. \end{cases}$$

By using L'Hôpital's rule and some straightforward arguments, it is easy to verify that

$$\lim_{|\theta|\to+\infty} \phi(\theta) = +\infty, \quad \lim_{|\theta|\to 0} \phi(\theta) = 1 \quad \text{and} \quad \phi(\theta) = \phi(-\theta).$$

Thus, ϕ is continuous at zero (which justifies its definition), coercive and symmetric. Since the log-normalizer function $\psi_P(\theta) = \log(M_P[\theta])$ is strictly convex we can conclude that if a solution exists it must be unique. The coercivity of ϕ implies that a solution exists, and due to the symmetry of ϕ we can conclude that it must be zero. To summarize, in this case, $\psi_P^*(\mu) = 0$ (with $\theta = 0$). If $y \neq \mu$ such that $a < y < b$ then the optimal solution to (A.5) is nonzero and by the first-order optimality conditions it must satisfy

$$y - \frac{b\exp(b\theta) - a\exp(a\theta)}{\exp(b\theta) - \exp(a\theta)} + \frac{1}{\theta} = 0. \tag{A.6}$$

Therefore, using (A.5) we can summarize that for $y \in (a, b) = \operatorname{dom} \psi_P^*$:

$$\psi_P^*(y) = \begin{cases} 0, & y = \mu, \\ (y-\mu)\theta - \log\left(\frac{\exp((b-\mu)\theta)-\exp((a-\mu)\theta)}{(b-a)\theta}\right), & y \neq \mu, \end{cases}$$

where θ is the root of (A.6).

Logistic

The moment generating function for Logistic distribution with location and scaling parameters μ and $s > 0$, respectively, is given by

$$M_P[\theta] = \exp(\mu y)B(1 - s\theta, 1 + s\theta), \qquad s\theta \in (-1, 1),$$

where $B(\cdot, \cdot)$ stands for the *Beta function*

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt.$$

The beta function and the closely related *gamma function*

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}\exp(-t)dt, \qquad \alpha > 0,$$

share the following well-known relation

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}. \tag{A.7}$$

The gamma function is an extension of the factorial as for a positive integer α it holds that $\Gamma(\alpha) = (\alpha - 1)!$. In the following, we will use the well-known function equations

$$B(\alpha+1, \beta) = B(\alpha, \beta)\frac{\alpha}{\alpha+\beta}, \tag{A.8}$$

and

$$B(\alpha, 1-\alpha) = \Gamma(1-\alpha)\Gamma(\alpha) = \frac{\pi}{\sin(\pi\alpha)}, \qquad \alpha \notin \mathbb{Z}. \tag{A.9}$$

The latter is known as Euler's reflection formula or Euler's function equation. Further details and proofs for both (A.8) and (A.9) can be found, for example, in [2].

Since $s\theta \in (-1, 1)$, the above relations imply that for any $\theta \neq 0$

$$\phi_s(\theta) := B(1-s\theta, 1+s\theta) \overset{(A.8)}{=} B(-s\theta, 1+s\theta)\frac{-s\theta}{-s\theta+1+s\theta} \overset{(A.9)}{=} \frac{-\pi s\theta}{\sin(-\pi s\theta)}.$$

For $\theta = 0$ we can verify by (A.7) that

$$\phi_s(\theta) = B_s(1-s\theta, 1+s\theta) = 1.$$

Thus, we can summarize

$$\phi_s(\theta) = B(1-s\theta, 1+s\theta) = \begin{cases} 1, & s\theta = 0, \\ \frac{-\pi s\theta}{\sin(-\pi s\theta)}, & s\theta \in (-1, 1)\backslash\{0\}. \end{cases} \tag{A.10}$$

Using L'Hôpital's rule we can verify that ϕ_s is continuous at $\theta = 0$. Since $-\pi s\theta \geq \sin(-\pi s\theta)$ for all $s\theta \in (-1, 1)$ we can conclude that $\phi_s(\theta) \geq 1$ for all $s\theta \in (-1, 1)$ and equality ($\phi_s(\theta) = 1$) holds if and only if $s\theta = 1$. Taking $|s\theta| \to 1$ it is evident that $\phi_s(\theta) \to \infty$. In addition, for any $\theta \neq 0$ the derivative of ϕ is given by

$$\phi_s'(\theta) = -\pi s\left[\frac{\sin(-\pi s\theta) + \pi s\theta\cos(-\pi s\theta)}{\sin^2(-\pi s\theta)}\right],$$

and consequently

$$\frac{\phi_s'(\theta)}{\phi_s(\theta)} = \frac{\sin(-\pi s\theta) + \pi s\theta\cos(-\pi s\theta)}{\theta\sin(-\pi s\theta)}. \tag{A.11}$$

We are now ready to evaluate Cramér's rate function that corresponds to the logistic distribution.

$$\begin{aligned}\psi_P^*(y) &= \sup\left\{y\theta - \log\left(M_P[\theta]\right) : \theta \in \mathbb{R}\right\} \\ &= \sup\left\{(y-\mu)\theta - \log\left(\phi_s(\theta)\right) : \theta \in \mathbb{R}\right\}.\end{aligned} \tag{A.12}$$

If $y = \mu$ then the discussion that follows equation (A.10) implies that $\sup\{-\log(\phi_s(\theta)) : \theta \in \mathbb{R}\} \leq 0$ where the upper bound is attained for $\theta = 0$ (since $\phi_s(\theta) \geq 1$ and $\phi_s(0) = 1$). Thus, we can conclude that $\psi_P^*(\mu) = 0$. If $y \neq \mu$ then the optimal solution to (A.12) satisfies $\theta \neq 0$. Since, in addition, for $|s\theta| \to 1$ we have that $\phi_s(\theta) \to \infty$, and consequently, $-\log(\phi_s(\theta)) \to -\infty$, an optimal solution to (A.12) for the case $y \neq \mu$ must satisfy the first-order optimality conditions

$$0 = y - \mu - \frac{\phi_s'(\theta)}{\phi_s(\theta)} = y - \mu - \frac{1}{\theta} - \frac{\pi s}{\tan\left(-\pi s\theta\right)}, \tag{A.13}$$

where the above follows from (A.11). To summarize,

$$\psi_P^*(y) = \begin{cases} 0, & y = \mu, \\ (y-\mu)\theta - \log\left(B(1-s\theta, 1+s\theta)\right), & y \neq \mu, \end{cases}$$

where $\theta \in \mathbb{R}$ is the nonzero root of (A.13).

References

1. Amblard, Cécile., Lapalme, E., Lina, J.-M.: Biomagnetic source detection by maximum entropy and graphical models. IEEE T. Biomed. Eng. **51**(3), 427–442 (2004)
2. Artin, Emil: The Gamma Function. Courier Dover Publications, New York (1964)
3. Auslender, Alfred, Teboulle, Marc: Asymptotic Cones and Functions in Optimization and Variational Inequalities. Springer Science & Business Media, Berlin (2006)
4. Auslender, Alfred, Teboulle, Marc: Interior gradient and proximal methods for convex and conic optimization. SIAM J. Optim. **16**(3), 697–725 (2006)
5. Barndorff-Nielsen, Ole: Information and Expontial Families: in Statistical Theory. Wiley, Hoboken (2014)
6. Barndorff-Nielsen, Ole E.: Normal inverse Gaussian distributions and stochastic volatility modeling. Scand. J. Stat. **24**(1), 1–13 (1997)
7. Bauschke, Heinz H., Bolte, Jérôme., Chen, Jiawei, Teboulle, Marc, Wang, Xianfu: On linear convergence of non-Euclidean gradient methods without strong convexity and Lipschitz gradient continuity. J. Optim. Theory Appl. **182**(3), 1068–1087 (2019)
8. Bauschke, Heinz H., Bolte, Jérôme., Teboulle, Marc: A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. Math. Oper. Res. **42**(2), 330–348 (2017)
9. Bauschke, H. H., Borwein, J. M.: Joint and separate convexity of the Bregman distance. In Studies in Computational Mathematics, vol. 8, pp. 23–36. Elsevier, (2001)
10. Bauschke, Heinz H., Borwein, Jonathan M., et al.: Legendre functions and the method of random Bregman projections. J. Convex Anal. **4**(1), 27–67 (1997)
11. Bauschke, Heinz H., Combettes, Patrick L., et al.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces, vol. 408. Springer, Berlin (2011)

12. Beck, Amir: Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with MATLAB. SIAM, Philadelphia (2014)
13. Beck, Amir: First-Order Methods in Optimization. SIAM, Philadelphia (2017)
14. Beck, Amir, Teboulle, Marc: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. J. Imaging Sci. **2**(1), 183–202 (2009)
15. Ben-Tal, A., Charnes, A.: A dual optimization framework for some problems of information theory and statistics. Technical report, Texas Univ. at Austin Center for Cybernetic Studies, (1979)
16. Ben-Tal, Aharon, Teboulle, Marc, Charnes, Abraham: The role of duality in optimization problems involving entropy functions with applications to information theory. J. Optim. Theory Appl. **58**(2), 209–223 (1988)
17. Bertsekas, Dimitri P.: Incremental proximal methods for large scale convex optimization. Math. Program. **129**(2), 163–195 (2011)
18. Björck, Åke.: Numerical Methods for Least Squares Problems. SIAM, Philadelphia (1996)
19. Bolte, Jérôme., Sabach, Shoham, Teboulle, Marc, Vaisbourd, Yakov: First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. J. Optim. **28**(3), 2131–2151 (2018)
20. Boyd, Stephen, Parikh, Neal, Chu, Eric, Peleato, Borja, Eckstein, Jonathan, et al.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Regist. Mach. Learn. **3**(1), 1–122 (2011)
21. Bregman, Lev M.: The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. U.S.S.R. Comp. Math. & Math. Phys. **7**(3), 200–217 (1967)
22. Brown, Lawrence D.: Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory. Institute of Mathematical Statistics, Ann Arbor (1986)
23. Cai, Zhengchen, Machado, Alexis, Chowdhury, Rasheda Arman, Spilkin, Amanda, Vincent, Thomas, Aydin, Ümit., Pellegrino, Giovanni, Lina, Jean-Marc., Grova, Christophe: Diffuse optical reconstructions of functional near infrared spectroscopy data using maximum entropy on the mean. Sci. Rep. **12**(1), 1–18 (2022)
24. Carathéodory, Constantin: Über den variabilitätsbereich der Fourier'schen konstanten von positiven harmonischen funktionen. Rend. Circ. Mat. Palermo **32**(1), 193–217 (1911)
25. Chambolle, Antonin, Pock, Thomas: A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis. **40**(1), 120–145 (2011)
26. Chambolle, Antonin, Pock, Thomas: On the ergodic convergence rates of a first-order primal-dual algorithm. Math. Program. **159**(1), 253–287 (2016)
27. Chowdhury, Rasheda Arman, Lina, Jean Marc, Kobayashi, Eliane, Grova, Christophe: MEG source localization of spatially extended generators of epileptic activity: comparing entropic and hierarchical Bayesian approaches. PloS One **8**(2), e55969 (2013)
28. Cohen, Eyal, Sabach, Shoham, Teboulle, Marc: Non-Euclidean proximal methods for convex-concave saddle-point problems. J. Appl. Num. Optim. **3**(1), 43–60 (2021)
29. Combettes , Patrick L., Pesquet, Jean-Christophe: Proximal splitting methods in signal processing. In: Fixed-Point Algorithms for Inverse Problems in Science and Engineering, pp. 185–212 (2011)
30. Condat, Laurent, Kitahara, Daichi, Contreras, Andrés, Hirabayashi, Akira: Proximal splitting algorithms for convex optimization: A tour of recent advances, with new twists. SIAM Rev. **65**(2), 375–435 (2023)
31. Corless, Robert M., Gonnet, Gaston H., Hare, David EG., Jeffrey, David J., Knuth, Donald E.: On the LambertW function. Adv. Comput. Math. **5**(1), 329–359 (1996)
32. Cover, Thomas M.: Elements of Information Theory. Wiley, Hoboken (1999)
33. Cramér, Harald: Sur un nouveau théoreme-limite de la théorie des probabilités. Actual. Sci. Ind. **736**, 5–23 (1938)
34. Dacunha-Castelle, Didier, Gamboa, Fabrice: Maximum d'entropie et problème des moments. Annales de l'IHP Probabilités et Statistiques **26**, 567–596 (1990)
35. Donsker, Monroe D., Varadhan, SR Srinivasa.: Asymptotic evaluation of certain Markov process expectations for large time-III. Commun. Pure Appl. Math. **29**(4), 389–461 (1976)
36. Dragomir, R.-A.: Bregman gradient methods for relatively-smooth optimization. PhD thesis, UT1 Capitole, (2021)
37. Dragomir, Radu-Alexandru., Taylor, Adrien B., d'Aspremont, Alexandre, Bolte, Jérôme.: Optimal complexity and certification of Bregman first-order methods. Math. Program. **19**, 41–83 (2022)

38. Ellis, Richard S.: Entropy, Large Deviations, and Statistical Mechanics, vol. 1431. Taylor & Francis, New York (2006)
39. Fermin, A.K., Loubes, J.E.A.N.-M.I.C.H.E.L., Ludena, C.A.R.E.N.N.E.: Bayesian methods for a particular inverse problem seismic tomography. Int. J. Tomogr. Stat. **4**(W06), 1–19 (2006)
40. Gamboa, F.: Méthode du maximum d'entropie sur la moyenne et applications. PhD thesis, Paris 11, (1989)
41. Gamboa, Fabrice, Gassiat, Elisabeth: Bayesian methods and maximum entropy for ill-posed inverse problems. Ann. Stat. **25**(1), 328–350 (1997)
42. Gamboa, Fabrice, Guéneau, Christine, Klein, Thierry, Lawrence, Eva: Maximum entropy on the mean approach to solve generalized inverse problems with an application in computational thermodynamics. RAIRO Oper. Res. **55**(2), 355–393 (2021)
43. Grova, Christophe, Jean Daunizeau, J.-M., Lina, Christian G., Bénar, Habib Benali, Gotman, Jean: Evaluation of EEG localization methods using realistic simulations of interictal spikes. Neuroimage **29**(3), 734–753 (2006)
44. Gzyl, H.: Maximum entropy in the mean: a useful tool for constrained linear problems. In: AIP Conference Proceedings, vol. 659, pp. 361–385. American Institute of Physics, (2003)
45. Hanzely, Filip, Richtarik, Peter, Xiao, Lin: Accelerated Bregman proximal gradient methods for relatively smooth convex optimization. Comput. Optim. Appl. **79**, 405–440 (2021)
46. Heers, Marcel, Chowdhury, Rasheda A., Hedrich, Tanguy, Dubeau, François, Hall, Jeffery A., Lina, Jean-Marc., Grova, Christophe, Kobayashi, Eliane: Localization accuracy of distributed inverse solutions for electric and magnetic source imaging of interictal epileptic discharges in patients with focal epilepsy. Brain Topogr. **29**(1), 162–181 (2016)
47. Jaynes, Edwin T.: Information theory and statistical mechanics. Phys. Rev. **106**(4), 620 (1957)
48. Kullback, Solomon: Information Theory and Statistics. Courier Corporation, Chelmsford (1997)
49. Kullback, Solomon, Leibler, Richard A.: On information and sufficiency. Ann. Math. Stat. **22**(1), 79–86 (1951)
50. Le Besnerais, Guy, Bercher, J.-F., Demoment, Guy: A new look at entropy for solving linear inverse problems. IEEE Trans. Inf. Theory **45**(5), 1565–1578 (1999)
51. Maréchal, Pierre, Lannes, André: Unification of some deterministic and probabilistic methods for the solution of linear inverse problems via the principle of maximum entropy on the mean. Inverse Probl. **13**(1), 135 (1997)
52. Moreau, Jean-Jacques.: Proximité et dualité dans un espace Hilbertien. Bulletin de la Société mathématique de France **93**, 273–299 (1965)
53. Navaza, Jorge: On the maximum-entropy estimate of the electron density function. Acta Crystallogr. A **41**(3), 232–244 (1985)
54. Navaza, Jorge: The use of non-local constraints in maximum-entropy electron density reconstruction. Acta Crystallogr. A **42**(4), 212–223 (1986)
55. Nelder, John Ashworth, Wedderburn, Robert WM.: Generalized linear models. J. R. Stat. Soc. Ser. A-G. **135**(3), 370–384 (1972)
56. Nesterov, Yurii: A method of solving a convex programming problem with convergence rate o (1/k** 2). Dokl. Akad. Nauk. SSSR **269**(3), 543 (1983)
57. Rietsch, E., et al.: The maximum entropy approach to inverse problems-spectral analysis of short data records and density structure of the Earth. J. Geophys. **42**(1), 489–506 (1977)
58. Rioux, Gabriel, Choksi, Rustum, Hoheisel, Tim, Maréchal, Pierre, Scarvelis, Christopher: The maximum entropy on the mean method for image deblurring. Inverse Probl. **37**(1), 015011 (2021)
59. Rioux, Gabriel, Scarvelis, Christopher, Choksi, Rustum, Hoheisel, Tim, Marechal, Pierre: Blind deblurring of barcodes via Kullback–Leibler divergence. IEEE Trans. Pattern Anal. Mach. Intel. **43**(1), 77–88 (2021)
60. Rockafellar, R Tyrrell: Convex Analysis. Princeton University Press, Princeton (1970)
61. Rohatgi, Vijay K., AK Md Saleh, Ehsanes: An Introduction to Probability and Statistics. Wiley, Hoboken (2015)
62. Rudin, Leonid I., Osher, Stanley, Fatemi, Emad: Nonlinear total variation based noise removal algorithms. Phys. D **60**(1–4), 259–268 (1992)
63. Sabach, Shoham, Teboulle, Marc: Lagrangian methods for composite optimization. Handb. Num. Anal. **20**, 401–436 (2019)
64. Tyrrell Rockafellar, R., Wets, Roger J-B.: Variational Analysis, vol. 317. Springer Science & Business Media, Berlin (2009)

65. Urban, B.: Retrieval of atmospheric thermodynamical parameters using satellite measurements with a maximum entropy method. Inverse Probl. **12**(5), 779 (1996)
66. Vardi, Yehuda, Shepp, Larry A., Kaufman, Linda: A statistical model for positron emission tomography. J. Am. Stat. Assoc. **80**(389), 8–20 (1985)
67. Wainwright, Martin J., Jordan, Michael Irwin: Graphical Models, Exponential Families, and Variational Inference. Now Publishers Inc, Norwell (2008)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.