IOP Publishing Inverse Problems
Inverse Problems **36** (2020) 035013 (26pp) https://doi.org/10.1088/1361-6420/ab6d55

The ML–EM algorithm in continuum: sparse measure solutions

Camille Pouchol[1,2] **and Olivier Verdier**[1,2]

[1] Department of Mathematics, KTH Royal Institute of Technology, 100 44 Stockholm, Sweden
[2] Department of Computing, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, Bergen, Norway

E-mail: pouchol@kth.se, olivierv@kth.se and olivier.verdier@hvl.no

Received 23 September 2019, revised 18 December 2019
Accepted for publication 15 January 2020
Published 12 February 2020

Abstract

Linear inverse problems $A\mu = y$ with Poisson noise and non-negative unknown $\mu \geqslant 0$ are ubiquitous in applications, for instance in positron emission tomography (PET) in medical imaging. The associated maximum likelihood problem is routinely solved using an expectation–maximisation algorithm (ML–EM). This typically results in images which look spiky, even with early stopping. We give an explanation for this phenomenon. We first regard the image μ as a measure. We prove that if the measurements y are not in the cone $\{A\mu, \mu \geqslant 0\}$, which is typical of low injected dose, likelihood maximisers must be sparse, i.e., typically a sum of point masses. We also show a weak sparsity result for cluster points of ML–EM. On the other hand, in the low noise regime, we prove that cluster points of ML–EM are optimal measures with full support. Finally, we provide concentration bounds for the probability to be in the sparse case, and a set of numerical experiments supporting our claims.

Keywords: maximum likelihood, inverse problems, expectation–maximisation, Kullback–Leibler divergence, positron emission tomography, Richardson–Lucy

1. Introduction

In various imaging modalities, recovering the image from acquired data can be recast as solving an inverse problem of the form $A\mu = y$, where A is a linear operator, y represents noisy measurements and μ the image, with $\mu \geqslant 0$ usually a desirable property. The problem thus becomes $\min_{\mu \geqslant 0} d(y, A\mu)$ where d is some given distance or divergence.

When the model is finite-dimensional, the operator A is simply a matrix $A = (a_{ij}) \in \mathbb{R}^{m\times r}$. If we assume a Poisson noise model, i.e., $y_i \sim \mathcal{P}((A\mu)_i)$ with independent draws, the corresponding (negative log) likelihood problem is equivalent to

$$\min_{\mu \geqslant 0} \quad d(y \| A\mu), \tag{1}$$

where d is the Kullback–Leibler divergence. As it turns out, this statistical model is similar to the familiar non-negative least-squares regression corresponding to Gaussian noise, but for a different distance functional: if $y \notin \{A\mu, \mu \geqslant 0\}$, it is projected onto it in the sense of of the divergence d, whereas if it belongs to this image set, any $\mu \geqslant 0$ such that $A\mu = y$ will be optimal.

The celebrated maximum likelihood expectation–maximisation algorithm (ML–EM) precisely aims at solving (1) and was introduced by Shepp and Vardi [37, 39], in the particular context of the imaging modality called positron emission tomography (PET). It was proposed earlier in another context and is as such often called the Richardson–Lucy algorithm [20, 34].

The ML–EM algorithm is iterative and writes

$$\mu_{k+1} = \frac{\mu_k}{A^{\mathrm{T}}1} A^{\mathrm{T}} \left(\frac{y}{A\mu_k} \right), \tag{2}$$

starting from $\mu_0 > 0$, usually $\mu_0 = 1$.

This algorithm is an expectation–maximisation (EM) algorithm, and as such it has many desirable properties: it preserves non-negativity and the negative log-likelihood decreases along iterates [12]. It can also be interpreted in several other ways [3, 11, 39], see [26] for an overview and [32] for related iterative algorithms. The expectation–maximisation point of view has also led to alternative algorithms [13], but in spite of various competing approaches, ML–EM (actually, its more numerically efficient variation OSEM [16]) has remained the algorithm used in practice in many PET scanners.

Despite its success, the ML–EM algorithm (2) is known to produce undesirable spikes along the iterates, where some pixels take increasingly high values. The toy example presented in figure 1 is an example of such artefacts in the case of PET. The reconstruction of a torus of maximum 1 with 100 iterations of ML–EM indeed exhibits some pixels with values as high as about 6.

This phenomenon has long been noticed in the literature, where images are referred to as "spiky" or "speckled" [38], others talking about "the chequerboard effect" [39]. In the discrete case, the works [7–9] have provided a partial explanation for this result. The author proves that, under general conditions (which include $m < r$), the minimum of (1) is such that it has at most $m - 1$ non-zero entries whenever $y \notin \{A\mu, \mu \geqslant 0\}$.

To the best of our knowledge, a theoretical justification for the subsistence of *only a few* non-zero entries has however remained elusive.

The aim of the present paper is to better understand that phenomenon via the analysis *in a continuous setting* of the minimisation problem (1) and the corresponding ML–EM algorithm (2). The continuous setting here refers to the image not being discretised on a grid. Note however that we keep the data space discrete.

Informally, considering μ as an element in some function space, we consider forward operators A of the form

$$(A\mu)_i := \int_K a_i(x)\mu(x)\,\mathrm{d}x,$$

where K is the compact on which one aims at reconstructing the image, and a_i is some non-negative function on K. This covers a wide range of applications, including PET.

One of our motivations is to derive algorithms for Poisson inverse problems with movement, for example for PET acquisition of a moving organ [18]. In that case, movement can be

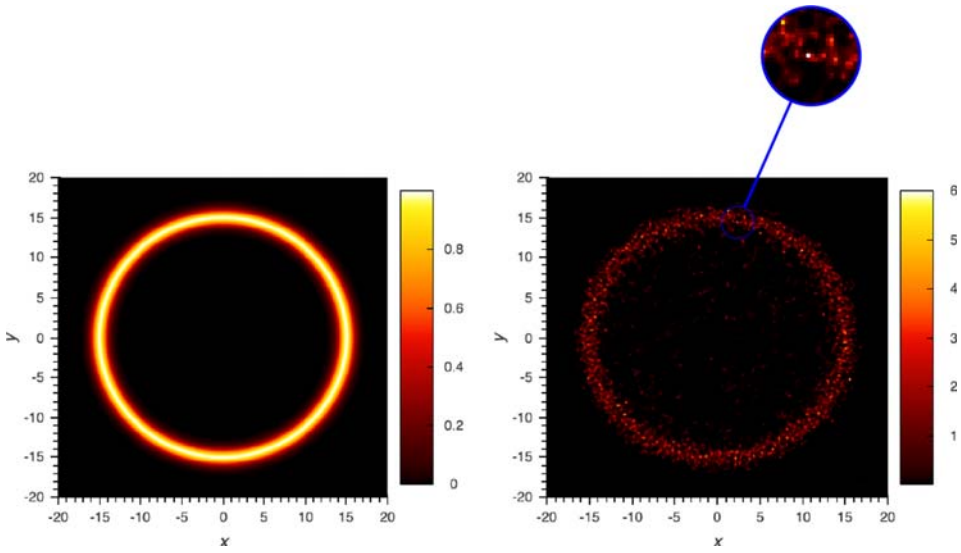

Figure 1. Phantom and reconstruction after 100 iterations of ML–EM, with a zoom on the region containing the pixel of highest value.

modelled by deformations of images which do not easily carry over to discretised images (simply because interesting deformations do not preserve the grid). It is then desirable to express the problem in a continuous setting, in order to both analyse the algorithms proposed in the literature, and to derive new ones [15, 27].

The field of inverse problems for imaging, with a continuum description of the unknown image, is abundant [2, 5]. Most often, the image is taken to be a function in some appropriate Sobolev space. To the best of our knowledge, however, there are relatively few results concerning the continuum description of the Poisson likelihood and the ML–EM algorithm for solving it.

In [23–25] and [33], both the image and data are considered in continuum, with a deterministic description of noise. These authors assume that detectors a_i lie in $L^\infty(K)$ and correspondingly assume that the image μ lies in $L^1(K)$. They study the convergence properties of the corresponding ML–EM algorithm in detail. In the first series of three papers, the compact K is restricted to $K = [0, 1]$.

Our paper differs from these works in that we do not make the two following restrictive assumptions, common to [23–25, 33]. The first restriction is to assume the existence of a non-negative solution μ to the equation $A\mu = y$, assumed to lie in $L^1(K)$. The second restriction is to assume that the functions a_i are bounded away from zero. This last assumption is unrealistic for some applications such as PET [33, remark 6.1].

The seminal paper [21] considers the optimisation problem over the set of non-negative Borel measures as we do. They obtain the corresponding likelihood function informally as the limit of the discrete one, but do not prove that it is an actual maximum likelihood problem for the PET statistical model. They then proceed to study the problem of whether minimisers can be identified with bounded functions, and not merely measures which might have a singular part. They indeed note that in some very specific cases (see also [23]), one can prove that the minimiser should be a Dirac mass. They speculate that there might be a link with the usual spiky results from ML–EM. They, however, do not provide any general conditions for sparsity.

Working in the space of non-negative measures $\mathcal{M}_+$, our main contributions are as follows:

Continuous framework: We prove that the continuous setting of measures is precisely the maximum likelihood problem with a Poisson point process model (proposition 2.1), and that the natural generalisation of the ML–EM iterates (23) indeed corresponds to the expectation–maximisation method associated to that continuous statistical model (see section 2.2).

Sparsity: We give a precise sparsity criterion (sparsity means that any optimal solution has singular support): if the data y is outside the cone $A(\mathcal{M}_+)$, then all optimal solutions are necessarily sparse (corollary 3.8); if the data y is inside the cone $A(\mathcal{M}_+)$, then there exist absolutely continuous solutions (lemma 3.12).

Properties of ML–EM iterates: We show the expected properties of the ML–EM iterates, namely monotonicity (corollary 4.2) and the fact that cluster points are fixed points of the algorithm (proposition 4.3).

Properties of ML–EM solutions: We show that in the non-sparse case, i.e., when an absolutely continuous solution exists as just mentioned, ML–EM iterates are optimal and have full support (theorem 4.10). In the sparse case, we provide a weak sparsity result for cluster points (corollary 4.4), and we give an explicit example of ML–EM converging to a sum of point masses (proposition 4.6).

Effect of noise: We derive estimates on the probability to be in the sparse case, depending on the noise level (proposition 5.1, theorem 5.2).

"Spiky" artefacts: With these results, we provide an explanation for the artefacts of figure 1: they are related to the sparsity result. By weak duality, we can indeed *certify* that optimal measures should be sums of point masses in that case, as detailed in section 6 dedicated to simulations.

Outline of the paper. The paper is organised as follows. In section 2, we introduce the functional and ML–EM in continuum in detail, with all the necessary notations, normalisations, definitions and useful properties about Kullback–Leibler divergences. Section 3 contains all results on the functional minimisers, starting from the optimality conditions to the diverging cases of the data y being inside or outside the cone $A(\mathcal{M}_+)$. Section 4 is devoted to the algorithm ML–EM itself, with the proof of its usual properties in continuum together with the implications they have on the case where the data y is in the cone $A(\mathcal{M}_+)$. In section 5, we estimate the probability that the data y ends up outside the image cone $A(\mathcal{M}_+)$. In section 6, we present simulations which confirm our theoretical predictions. Finally, in section 7 we conclude with open questions and perspectives.

2. Maximum likelihood and ML–EM in continuum

2.1. Mathematical background

2.1.1. Space of radon measures. As stated in the introduction, we model the image to reconstruct as a non-negative measure μ defined on a compact set K. Some of our results require $K \subset \mathbb{R}^p$ (typically, $p = 2$ or 3).

More precisely, we will consider the set of Radon measures, denoted $\mathcal{M}(K)$ and defined as the topological dual of the set of continuous functions over K, denoted $\mathcal{C}(K)$. The space of non-negative measures will be denoted by $\mathcal{M}_+(K)$. For brevity, we will often write $\mathcal{M}$ for $\mathcal{M}(K)$ and $\mathcal{M}_+$ for $\mathcal{M}_+(K)$ when there is no ambiguity as to the underlying compact K.

We identify a linear functional $\mu \in \mathcal{M}_+$ with its corresponding Borel measure (as per the Riesz–Markov representation theorem), using $\mu(B)$ to denote the measure of a Borel subset B of K. We will also sometimes write the dual pairing between a measure $\mu \in \mathcal{M}$ and a function $f \in \mathcal{C}(K)$ as

$$\langle \mu, f \rangle = \int_K f \mathrm{d}\mu.$$

The support of a measure $\mu \in \mathcal{M}$ is defined as the closed set

$$\mathrm{supp}(\mu) := \{\, x \in K \mid \mu(N) > 0, \quad \forall N \in N(x) \,\},$$

where $N(x)$ is the set of all open neighbourhoods of x.

Finally, recall that, by the Banach–Alaoglu theorem, bounded sets in $\mathcal{M}$ are weak-$*$ compact [35].

2.1.2. Kullback-Leibler divergence. We here recall the definition of the Kullback–Leibler (KL) divergence. Instead of giving the general definition, we make the two instances that will actually be needed in this paper explicit, for (non-normalised) non-negative vectors in $\mathbb{R}^m$, and for probability measures on K.

- *For vectors in* $\mathbb{R}^m$. For any two non-negative vectors u and v in $\mathbb{R}^m$, we define the Kullback–Leibler divergence between u and v as

$$d(u\|v) := \sum_{i=1}^{m} \left(v_i - u_i - u_i \log\left(\frac{v_i}{u_i}\right)\right),$$

  with the convention $0\log(0) = 0$ and $d(u\,\|\,v) = +\infty$ if there exists i such that $u_i = 0$ and $v_i > 0$.

- *For probability measures on K.* For any two probability measures μ and ν on K, we define the Kullback–Leibler divergence between μ and ν

$$D(\mu\|\nu) := \int_K \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\mu,$$

  if μ is absolutely continuous with respect to ν (denoted $\mu \ll \nu$) and $\log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)$ is integrable with respect to μ. Here $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ stands for the Radon–Nikodym derivative of μ with respect to ν. Otherwise, we define $D(\mu\,\|\,\nu) := +\infty$.

When a measure is absolutely continuous with respect to the Lebesgue measure on $K \subset \mathbb{R}^p$, we simply say that it is absolutely continuous. Any reference to the Lebesgue measure implicitly assumes that K stands for the closure of some bounded domain in $\mathbb{R}^p$ (i.e., a bounded, connected and open subset of $\mathbb{R}^p$).

2.2. Statistical model

We want to recover a measure $\mu \in \mathcal{M}_+$ from independent Poisson distributed measurements

$$N_i \sim \mathcal{P}\left(\int_K a_i d\mu\right), \qquad i = 1, \ldots, m, \tag{3}$$

with

$$a_i \geqslant 0, \quad a_i \in \mathcal{C}(K), \qquad i = 1, \ldots, m. \tag{4}$$

2.2.1. Positron emission tomography [28]. In PET, a radiotracer injected into the patient and, once concentrated into tissues, disintegrates by emitting a positron. This process is well known to be accurately modelled by a Poisson point process, itself defined by a non-negative measure. After a short travel distance called *positron range*, this positron interacts with an electron. The result is the emission of two photons in random opposite directions. Pairs of detectors around the body then detect simultaneous photons, and the data is given by the number of counts per pair of detectors.

In the case of PET, m is the number of detectors (i.e., pairs of single detectors). For a given point $x \in K$ and detector $i \in \{1, \dots, m\}$, $a_i(x)$ then denotes the probability that a positron emitted in x will be detected by detector i.

Finally, we will throughout the paper assume

$$\sum_{i=1}^{m} a_i > 0 \text{ on } K. \tag{5}$$

For PET, this amounts to assuming that the points in K are in the so-called field of view, namely that any emission has a non-zero probability to be detected.

2.2.2. Derivation of the statistical model (3) for PET. We proceed to give a proof that the statistical model (3) (and thus, the corresponding likelihood function) applies to PET. Here, we assume that the emission process of PET is modelled by a Poisson point process, defined by a measure $\mu \in \mathcal{M}_+$, and that each point drawn from the Poisson process has a probability $a_i(x)$ to be detected by detector i.

Proposition 2.1. *The statistical model* (3) *applies to PET.*

Proof. The proof relies on the following properties of Poisson point processes [19]:

- *law of numbers*: the number of points emitted by a Poisson process of intensity μ follows the Poisson law with parameter $\int_K m\mu = \mu(K)$.
- *thinning property*: the points that are kept with (measurable) probability $p : K \mapsto [0, 1]$ still form a Poisson point process, with intensity $p\mu$, and it is independent from that of points that are not kept. This property generalises to p_i, $1 \leqslant i \leqslant m$ with $\sum_{i=1}^{m} p_i(x) = 1$ for all $x \in K$.

By the thinning property, the families of points which lead to an emission detected in detector i, $i = 1, \dots, m$, are all independent Poisson processes with associated measure $a_i\mu$, for $i = 1, \dots, m$. Thus, the random variables N_i representing the number of points detected in detector i are independent and of law $\mathcal{P}(\int_K a_i m\mu)$, which proves the claim. □

2.2.3. Maximum likelihood problem. The likelihood corresponding to the statistical model (3) reads

$$L(N_1, \dots, N_p; \mu) = \prod_{i=1}^{m} L(N_i; \mu) = \prod_{i=1}^{m} \mathrm{e}^{-\int_K a_i \mathrm{d}\mu} \frac{(-\int_K a_i \mathrm{d}\mu)^{N_i}}{N_i!},$$

since $\mathbb{P}(N_i = n_i) = \mathrm{e}^{-\int_K a_i \mathrm{d}\mu} \frac{(-\int_K a_i \mathrm{d}\mu)^{n_i}}{n_i!}$.

Dropping the factorial terms (they do not depend on μ and will thus play no role when maximising the likelihood), we get

$$\log(L(N_1,\dots,N_p;\mu)) = -\sum_{i=1}^{m}\int_K a_i \mathrm{d}\mu + \sum_{i=1}^{m} N_i \log\left(\int_K a_i \mathrm{d}\mu\right). \tag{6}$$

The corresponding maximum likelihood problem, written for a realisation n_i of the random variable N_i, $i = 1,\dots,m$, is given by

$$\max_{\mu\in\mathcal{M}_+} \quad -\int_K \left(\sum_{i=1}^{m} a_i\right)\mathrm{d}\mu + \sum_{i=1}^{m} n_i \log\left(\int_K a_i \mathrm{d}\mu\right). \tag{7}$$

Defining the operator

$$\begin{aligned} A : \mathcal{M} &\longrightarrow \mathbb{R}^m \\ \mu &\longmapsto \big(\langle \mu, a_i\rangle\big)_{1\leqslant i\leqslant m}, \end{aligned}$$

the optimisation problem conveniently rewrites in terms of the Kullback–Leibler divergence: upon adding constants and taking the negative log-likelihood problem, it reads

$$\min_{\mu\in\mathcal{M}_+} \quad d(n\|A\mu). \tag{8}$$

2.2.4. ML–EM iterates. We now define the ML–EM algorithm, which aims at solving the optimisation problem (7). It is given by the iterates

$$\mu_{k+1} = \frac{\mu_k}{\sum_{i=1}^m a_i}\left(\sum_{i=1}^{m}\frac{n_i a_i}{\int_K a_i \mathrm{d}\mu_k}\right), \tag{9}$$

starting from an initial guess $\mu_0 \in \mathcal{M}_+$. In agreement with the Kullback–Leibler divergence, we choose the convention that divisions of the form $0/0$ are of course taken to be equal to 0.

Note that this algorithm can be shown to be an EM algorithm for the continuous problem. The proof is beyond the scope of this paper, so we decide to omit it, but we just mention that the corresponding so-called *complete data* would be given by the positions of points together with the detector that has detected each of them.

2.3. Normalisations

Due to the assumption (5), we may without loss of generality assume that

$$\sum_{i=1}^{m} a_i = 1 \tag{10}$$

on K. Otherwise we could just define $\tilde{\mu} = (\sum_{i=1}^m a_i)\mu$ and $\tilde{a}_i = a_i/(\sum_{j=1}^m a_j)$. This normalisation now implies $0 \leqslant a_i \leqslant 1$ for all $i = 1,\dots,m$.

We further normalise the measures by dividing the functional by $n := \sum_{i=1}^m n_i$, considering $\mu := \frac{\mu}{n}$ to remove the factor. We then define

$$y_i := \frac{n_i}{n}.$$

From now on, we consider the optimisation problem (minimisation of the negative log-likelihood):

$$\min_{\mu\in\mathcal{M}_+} \quad \ell(\mu), \tag{11}$$

where

$$\ell(\mu) := \int_K \mathrm{d}\mu - \sum_{i=1}^m y_i \log\left(\int_K a_i \mathrm{d}\mu\right) \tag{12}$$

$$= \langle \mu, 1\rangle - \sum_{i=1}^m y_i \log\left(\langle \mu, a_i\rangle\right), \tag{13}$$

defined to be $+\infty$ for any measure such that $\langle \mu, a_i\rangle = 0$ for some $i \in \mathrm{supp}(y)$, where

$$\mathrm{supp}(y) := \{i = 1, \dots, m \mid y_i > 0\}.$$

After normalisation, the ML–EM iterates are given by

$$\mu_{k+1} = \mu_k \left(\sum_{i=1}^m \frac{y_i a_i}{\int_K a_i \mathrm{d}\mu_k}\right) = \mu_k \left(\sum_{i=1}^m \frac{y_i a_i}{\langle \mu_k, a_i\rangle}\right). \tag{14}$$

We recall the property that ML–EM preserves the total number of counts: $\langle \mu_k, 1\rangle = 1$ for all $k \geqslant 1$, which corresponds to $\langle \mu_k, 1\rangle = n = \sum_{i=1}^m n_i$ before normalisation. We also emphasise the important property that iterations cannot increase the support of the measure, namely

$$\forall k \in \mathbb{N},\ \mathrm{supp}(\mu_{k+1}) \subset \mathrm{supp}(\mu_k).$$

ML–EM iterates are well-defined. We assume throughout that the initial measure μ_0 fulfils

$$\langle \mu_0, a_i\rangle > 0 \qquad \forall i \in \mathrm{supp}(y). \tag{15}$$

Note that usual practice is to take μ_0 to be absolutely continuous with respect to the Lebesgue measure, typically $\mu_0 = 1$, in which case (15) is satisfied.

The following simple Lemma shows that assumption (15) ensures that the iterates are well-defined.

Lemma 2.2. *The ML–EM iterates* (14) *satisfy*

$$\langle \mu_k, a_i\rangle > 0 \implies \langle \mu_{k+1}, a_i\rangle > 0 \qquad i \in \mathrm{supp}(y).$$

Proof. From the Cauchy–Schwarz inequality, $\mu_k(K)\left\langle \mu_k, a_i^2\right\rangle \geqslant \langle \mu_k, a_i\rangle^2$. Combined with the definition of ML–EM iterates, this entails for any $i \in \mathrm{supp}(y)$,

$$\langle \mu_{k+1}, a_i\rangle = \sum_{j=1}^m y_j \frac{\langle \mu_k, a_i a_j\rangle}{\langle \mu_k, a_j\rangle} \geqslant y_i \frac{\left\langle \mu_k, a_i^2\right\rangle}{\langle \mu_k, a_i\rangle} \geqslant y_i \frac{\langle \mu_k, a_i\rangle}{\mu_k(K)} > 0.$$

□

2.4. Adjoint and cone

Since $A : \mathcal{C}(K)^* \to \mathbb{R}^m$, we can define its adjoint $A^* : \mathbb{R}^m \to \mathcal{C}(K)$ (identifying $\mathbb{R}^m$ as a Euclidean space with its dual), which is given by

$$A^* w = \sum_{i=1}^m w_i a_i, \qquad w \in \mathbb{R}^m. \tag{16}$$

The set $A(\mathcal{M}_+) = \{A\mu,\ \mu \in \mathcal{M}_+\} \subset \mathbb{R}^m$ is a closed and convex cone and, as proved in [14], its dual cone $A(\mathcal{M}_+)^*$ can be characterised as being given by the set of vectors $\lambda \in \mathbb{R}^m$ such that $\sum_{i=1}^m \lambda_i a_i \geqslant 0$ on K, i.e.,

$$A(\mathcal{M}_+)^* = \{\lambda \in \mathbb{R}^m \mid A^*\lambda \geqslant 0 \quad \text{on} \quad K\}. \tag{17}$$

As a result, the interior of the dual cone $A(\mathcal{M}_+)^*$ is given by the vectors $\lambda \in \mathbb{R}^m$ such that $A^*\lambda > 0$ on K.

The normalisation condition (10) can now be rewritten

$$A^*\mathbf{1} = \mathbf{1}, \tag{18}$$

where $\mathbf{1}$ is the vector of $\mathbb{R}^m$ which all components are one: $\mathbf{1} = (1, \ldots, 1)$. Moreover, we can rewrite the ML–EM iteration (14) as

$$\mu_{k+1} = \mu_k A^* \left(\frac{y}{A\mu_k}\right),$$

which is the continuous analogue to the discrete case (2), taking into account the normalisation (18).

Minimisation over the cone. The problem $\min_{\mu\in\mathcal{M}_+} d(y\|A\mu)$, is equivalent to the following minimisation problem over the cone $A(\mathcal{M}_+)$:

$$\min_{w\in A(\mathcal{M}_+)} \quad d(y\|w). \tag{19}$$

Indeed, if $w^\star$ is optimal for the problem (19), any $\mu^\star$ such that $A\mu^\star = w^\star$ is optimal for the original problem. From the property $d(y\,\|\,w) = 0 \iff y = w$, we also infer that when $y \in A(\mathcal{M}_+)$, $\mu^\star$ is optimal if and only if $A\mu^\star = y$.

3. Properties of minimisers

In this section, we gather results concerning the functional ℓ and its minimisers, proving that they are sparse when the data y is not in the image cone $A(\mathcal{M}_+)$. First, we note that the functional ℓ defined by (12) is a convex and proper function.

3.1. Characterisation of optimality

We now derive necessary and sufficient *optimality conditions*.

We first prove that any optimum must have a fixed unit mass.

Proposition 3.1. *If $\mu^\star$ is optimal for (11), then $\langle \mu^\star, 1\rangle = \int_K \mathrm{d}\mu^\star = 1$.*

Proof. For any $\mu \in \mathcal{M}_+$, we have

$$\ell(\mu) = -\sum_{i=1}^m y_i \log\left(\frac{\langle \mu, a_i\rangle}{\langle \mu, 1\rangle}\right) + \big(\langle \mu, 1\rangle - \log\big(\langle \mu, 1\rangle\big)\big). \tag{20}$$

Observe that the second term depends only on the mass $\langle \mu, 1\rangle$, whereas the first term is scale-invariant. As a result, an optimal μ has to minimise the second term, which turns out to admit the unique minimiser $\langle \mu, 1\rangle = 1$. □

Remark 3.2. This result follows from the optimality conditions derived later in proposition 3.4, but the proof above is simple and also highlights that the maximum likelihood estimator

for μ is consistent with the maximum likelihood estimator for $\int_K \mathrm{d}\mu$, as the second term in (20) is none other than the negative log-likelihood of the total mass.

Corollary 3.3. *The infimum of ℓ is a minimum.*

Proof. From proposition 3.1, we may restrict the search of optimal solutions to $\{\,\mu \in \mathcal{M}_+ \mid \mu(K) = 1\,\}$, which by the Banach–Alaoglu theorem, is weak-$*$ compact. Since ℓ is weak-$*$ continuous, the claim follows. □

We now give the full optimality conditions. The convex function ℓ defined in (12) has the following open domain:

$$\mathrm{dom}(\ell) := \{\mu \in \mathcal{M}_+ \mid \langle \mu, a_i \rangle > 0 \ \text{ for } \ i \in \mathrm{supp}(y)\}\,.$$

Notice further that for any $\mu \in \mathrm{dom}(\ell)$, the function ℓ is Fréchet-differentiable (in the sense of the strong topology). Its gradient is given for $\mu \in \mathrm{dom}(\ell)$ is then the element in the dual $\mathcal{M}^*$ of $\mathcal{M}$ given by

$$\nabla \ell(\mu) = 1 - \sum_{i=1}^{m} \frac{y_i a_i}{\langle \mu, a_i \rangle}, \tag{21}$$

which we identify with an element of $\mathcal{C}(K)$.

For any vector $w \in \mathbb{R}^m$, we define $\lambda(w) \in \mathbb{R}^m$ by

$$\lambda_i(w) := 1 - \frac{y_i}{w_i}, \tag{22}$$

(with the convention $\lambda_i = 1$ if $y_i = 0$, that is, if $i \notin \mathrm{supp}(y)$). Using $\sum_{i=1}^m a_i = 1$, we can rewrite (21) as

$$\nabla \ell(\mu) = A^* \lambda(A\mu) = \sum_{i=1}^{m} \lambda_i(A\mu)\, a_i. \tag{23}$$

Proposition 3.4. *The measure $\mu^\star \in \mathcal{M}$ is optimal for the problem (11) if and only if the following optimality conditions hold*

$$\begin{aligned} &A^*\lambda(A\mu^\star) \geqslant 0 \quad \textit{on} \quad K, \\ &A^*\lambda(A\mu^\star) = 0 \quad \textit{on} \quad \mathrm{supp}(\mu^\star). \end{aligned} \tag{24}$$

These conditions can be equivalently written as

$$\begin{aligned} &\sum_{i=1}^{m} \frac{y_i a_i}{\langle \mu^\star, a_i \rangle} \leqslant 1 \quad \textit{on} \quad K, \\ &\sum_{i=1}^{m} \frac{y_i a_i}{\langle \mu^\star, a_i \rangle} = 1 \quad \textit{on} \quad \mathrm{supp}(\mu^\star). \end{aligned} \tag{25}$$

Recall that the *normal cone* to $\mathcal{M}_+$ at μ is defined as

$$N_{\mathcal{M}_+}(\mu) := \{f \in \mathcal{C}(K) \mid \forall \nu \in \mathcal{M}_+,\ \langle \nu - \mu, f \rangle \leqslant 0\}\,.$$

We need a characterisation of that normal cone before proceeding further.

Lemma 3.5. *The normal cone at a given $\mu \in \mathrm{dom}(\ell)$ is given by*

$$N_{\mathcal{M}_+}(\mu) = \{f \in \mathcal{C}(K) \mid f \leqslant 0 \text{ on } K, f = 0 \text{ on } \mathrm{supp}(\mu)\}.$$

Proof. Let $f \in \mathcal{C}(K)$ be in $N_{\mathcal{M}_+}(\mu)$, i.e., it satisfies $\langle \nu - \mu, f\rangle \leqslant 0$ for all ν in $\mathcal{M}_+$. First, we choose $\nu = \mu + \delta_x$ (with δ_x the Dirac mass at x), which yields $f(x) \leqslant 0$, so we must have $f \leqslant 0$ on K. Then with $\nu = 0$, we find $\langle \mu, f\rangle \geqslant 0$. Since we also have $f \leqslant 0$, $\langle \mu, f\rangle = 0$ leading to $f = 0$ on $\mathrm{supp}(\mu)$.

The reverse is also true: if $f \leqslant 0$ on K and $f = 0$ on $\mathrm{supp}(\mu)$, then $\langle \nu - \mu, f\rangle \leqslant 0$ for all ν in $\mathcal{M}_+$, which gives $f \in N_{\mathcal{M}_+}(\mu)$. □

Proof of proposition 3.4. Since f is differentiable on $\mathrm{dom}(\ell)$ and convex on the convex set $\mathcal{M}_+$, a point $\mu^\star \in \mathrm{dom}(\ell)$ is optimal if and only if

$$\nabla \ell(\mu^\star) \in -N_{\mathcal{M}_+}(\mu^\star).$$

From the characterisation of $N_{\mathcal{M}_+}(\mu)$ given below in lemma 3.5, and the fact that $\nabla \ell(\mu) = A^* \lambda(A\mu)$, the optimality condition exactly amounts to the conditions (24). □

Remark 3.6. An alternative proof of these optimality conditions can be obtained by considering instead the equivalent problem of minimising $d(y \,\|\, w)$ with w ranging over the cone $A(\mathcal{M}_+)$. The cone has a non-empty relative interior which proves that Slater's condition is fulfiled. Since the problem is convex, KKT conditions are equivalent to optimality for $\min_{w \in A(\mathcal{M}_+)} d(y \| w)$ [6].

The Lagrange dual is given by $g(\lambda) := \min d(y \,\|\, w) - \langle \lambda, w\rangle$ for $\lambda \in A(\mathcal{M}_+)^*$. A straightforward computation leads to

$$g(\lambda) = \sum_{i=1}^{m} y_i \log(1 - \lambda_i), \tag{26}$$

for $\lambda \leqslant 1$, with value $-\infty$ if there exists $i \in \mathrm{supp}(y)$ such that $\lambda_i = 1$.

The KKT conditions for a primal optimal $w^\star$ and dual optimal $\lambda^\star$ write

(a) $w^\star \in A(\mathcal{M}_+)$, $\lambda^\star \in A(\mathcal{M}_+)^*$
(b) $\langle \lambda^\star, w^\star\rangle = 0$,
(c) $\nabla_w d(y \,\|\, w^\star) - \lambda^\star = 0$ (equivalent to $\lambda^\star = \lambda(w^\star) = 1 - \frac{y}{w^\star}$)

A measure $\mu^\star$ is then optimal if and only if $A\mu^\star = w^\star$ for $w^\star$ primal optimal. Since $(\lambda^\star, A\mu^\star) = \langle \mu^\star, A^*\lambda^\star\rangle$ (by definition of A^*), the condition (ii) thus becomes

$$\langle \mu^\star, A^*\lambda^\star\rangle = \int_K A^*\lambda^\star \, \mathrm{d}\mu^\star = 0.$$

Since $\lambda^\star \in A(\mathcal{M}_+)^*$, $A^*\lambda^\star \geqslant 0$ over K. Thus, we must have $A^*\lambda^\star = 0$ on $\mathrm{supp}(\mu^\star)$ for the above integral to vanish. All in all, we exactly recover the conditions (25), with the additional interpretation that $\lambda(A\mu^\star)$ is a dual optimal variable.

With these notations concerning the dual problem now set, let us prove that the dual problem has a unique maximiser $\lambda^\star$.

Lemma 3.7. *The dual problem*

$$\max_{\lambda \in A(\mathcal{M}_+)^*} g(\lambda)$$

has a unique maximiser.

Proof. The idea is to go back the primal problem by using the identity $\lambda^\star = 1 - \frac{y}{w^\star}$ for an optimal pair $(w^\star, \lambda^\star)$. Since $w^\star$ relates to an optimal measure $\mu^\star$ by $A\mu^\star = w^\star$, we are done if we prove that $\{ A\mu^\star \mid \mu^\star \text{ optimal} \}$ is reduced to a singleton. This fact is proved in [21]-[theorem 4.1], and we here gather the main ideas for completeness. For two optimal measures μ and ν, we integrate the first KKT condition of (25) on the support of ν to uncover

$$\sum_{i=1}^m y_i \frac{(A\nu)_i}{(A\mu)_i} \leqslant 1,$$

and we may of course exchange the roles of μ and ν in this inequality.

Suppose now that a vector $c \in \mathbb{R}^m$ with $c_i = 0$ for $i \in \mathrm{supp}(y)$ satisfies both $\sum_{i=1}^m y_i c_i \leqslant 1$ and $\sum_{i=1}^m y_i (1/c_i) \leqslant 1$. From that, one obtains $\sum_{i=1}^m y_i \frac{(c_i-1)^2}{c_i} \leqslant 0$, from which we conclude that $c_i = 1$ for all i. Applying this to $c = \frac{A\mu}{A\nu}$, the result is proved. □

3.2. Case $y \notin A(\mathcal{M}_+)$

When the data y is not in the cone $A(\mathcal{M}_+)$, optimality conditions imply sparsity of any optimal measure.

Corollary 3.8. *Assume that $y \notin A(\mathcal{M}_+)$. Then any $\mu^\star$ minimiser of (11) is sparse, in the following sense*

$$\mathrm{supp}(\mu^\star) \subset \arg\ \min \left(\sum_{i=1}^m \lambda_i^\star a_i \right), \tag{27}$$

where $\lambda^\star$ is the unique maximiser for the dual problem, which satisfies $\lambda^\star \neq 0$.

Proof. Given an optimal $\mu^\star$, conditions (24) imply $\mathrm{supp}(\mu^\star) \subset \arg\min(A^*\lambda(A\mu^\star))$, with $\lambda(A\mu^\star) = 1 - \frac{y}{A\mu^\star}$, where we used $\sum_{i=1}^m a_i = 1$. The uniqueness of maximisers for the dual problem established in lemma 3.7, allows us to write $\lambda(A\mu^\star) = \lambda^\star$.

The vector $\lambda^\star$ must be non-zero: if it were not the case, then, using the definition (22), that would imply $y \in A(\mathcal{M}_+)$, which would contradict our initial assumption. □

Remark 3.9. Why does condition (27) imply sparsity? Let $\lambda^\star$ be defined as in the previous theorem, and define the function $\varphi^\star := A^*\lambda^\star = \sum_{i=1}^m \lambda_i^\star a_i$. We know from proposition 3.4 that both $\varphi^\star \geqslant 0$ and $\mathrm{supp}(\mu^\star) \subset \arg\min(\varphi^\star)$.

Assuming that the a_i's are linearly independent in $\mathcal{C}(K)$, $\varphi^\star$ cannot vanish identically since $\lambda^\star \neq 0$. Supposing further that for all i, $a_i \in \mathcal{C}^2(K)$, we have

$$\mathrm{supp}(\mu^\star) \cap \mathrm{int}(K) \subset \mathcal{S} := \{ x \in K \mid \nabla\varphi^\star(x) = 0 \}.$$

We make the final assumption that the Hessian of $\varphi^\star$ is invertible at the points $x \in \arg\min(\varphi^\star)$, which is equivalent to its positive definiteness since these are minimum points of $\varphi^\star$. This implies that $\mathcal{S}$ consists of *isolated points*. Consequently, the restriction to $\mathrm{int}(K)$ of any optimal solution $\mu^\star$ is a sum of Dirac masses.

Note that all the above regularity assumptions hold for *generic* functions a_i. One case where all of them are readily satisfied is when the functions a_i are analytic with K connected.

In fact, if we go further and assume that $\arg\min(\varphi^\star)$ is reduced to a singleton $\bar{x}$, then the set of optimal measures is itself a singleton, given by the Dirac mass at $\bar{x}$.

Remark 3.10. We can exhibit a case where only Dirac masses are optimal. Suppose that only $y_{i_0} = 1$. Then the function ℓ for measures μ such that $\mu(K) = 1$ is simply $\ell(\mu) = 1 - \log\left(\left\langle \mu, a_{i_0}\right\rangle\right)$. One can directly check that a minimiser $\mu^\star$ necessarily satisfies $\mathrm{supp}(\mu^\star) \subset \arg\max(a_{i_0})$, in agreement with condition (27). If this set is discrete, then $\mu^\star$ is a sum of Dirac masses located at these points. Note that such a data point y is outside the cone $A(\mathcal{M}_+)$ if and only if $\max(a_{i_0}) < 1$. If not, it lies on the boundary of the cone, showing that some boundary points might lead to sparse minimisers as well.

3.3. Moment matching problem and case $y \in \mathrm{int}(A(\mathcal{M}_+))$

When the data y is in the cone $A(\mathcal{M}_+)$, searching for minimisers of (11) is equivalent to solving $A\mu = y$ for $\mu \in \mathcal{M}_+$. For the applications, we are particularly interested in the existence of absolutely continuous solutions. We make use of the results of [14], which addresses this problem.

We shall use the assumption:

$$\text{the functions } a_i, i = 1, \dots, m \text{ are linearly independent in } \mathcal{C}(K). \tag{28}$$

Under (28), $A(\mathcal{M}_+)$ has non-empty interior.

We now recall a part of theorem 3 of [14] which will be sufficient of our purpose.

Theorem 3.11. ([14]). *Assume that $A(\mathcal{M}_+)$ and its dual cone $A(\mathcal{M}_+)^*$ have non-empty interior. Then for any $y \in \mathrm{int}(A(\mathcal{M}_+))$, there exists $\mu^\star$ which is absolutely continuous, with positive and continuous density, such that $A\mu^\star = y$.*

Lemma 3.12. *Under hypothesis (28), $A(\mathcal{M}_+)$ has non-empty interior, and if $y \in \mathrm{int}(A(\mathcal{M}_+))$, there exists an optimal measure $\mu^\star$ which is absolutely continuous with positive and continuous density.*

Proof. This is a direct consequence of theorem 3.11. We just need to check that $A(\mathcal{M}_+)^*$ has non-empty interior. Using the characterisation of the dual cone (17), this is straightforward since $\sum_{i=1}^m a_i = 1$. □

3.4. Case $y \in \partial A(\mathcal{M}_+)$

The previous approach settles the case where the data y is in the interior $\mathrm{int}(A(\mathcal{M}_+))$ of the image cone, which poses the natural question of its boundary $\partial A(\mathcal{M}_+)$. It routinely happens in practice that some components of the data y are zero, which means that the vector y lies at the border of the cone, $y \in \partial A(\mathcal{M}_+)$. Upon changing the compact, a further use of the results of [14] shows that if the support of the data $\mathrm{supp}(y)$ is not too small (see the precise condition (32) below), the situation is the same as for $\mathrm{int}(A(\mathcal{M}_+))$.

The idea is to remove all the zero components of the data vector y, consider only the positive ones and try to solve $\langle \mu^\star, a_i\rangle = y_i$ for $i \in \mathrm{supp}(y)$, while making sure that the measure $\mu^\star$ has a support such that $\langle \mu^\star, a_i\rangle = 0$ for $i \notin \mathrm{supp}(y)$.

We denote $\tilde{m} := \#(\mathrm{supp}(y))$, $\tilde{K} := K \backslash \cup_{i \notin \mathrm{supp}(y)} a_i^{-1}(\{0\})$, $\tilde{y} = (y_i)_{i \in \mathrm{supp}(y)}$, and finally the reduced operator,

$$\tilde{A} : \mathcal{M}(\tilde{K}) \to \mathbb{R}^{\tilde{m}} \tag{29}$$

$$\mu \longmapsto \left(\int_K a_i \mu \right)_{i \in \mathrm{supp}(y)}, \tag{30}$$

which has an associated cone $\tilde{A}(\mathcal{M}_+(\tilde{K}))$.

We will need the assumptions

$$\text{the functions } a_i,\ i \in \mathrm{supp}(y) \text{ are linearly independent in } \mathcal{C}(\tilde{K}), \tag{31}$$

and

$$\sum_{i \in \mathrm{supp}(y)} a_i > 0 \quad \text{on} \quad \tilde{K}. \tag{32}$$

Proposition 3.13. *We assume* (31) *and* (32). $\tilde{A}(\mathcal{M}_+(\tilde{K}))$ *has non-empty interior and we assume*

$$\tilde{y} \in \mathrm{int}\left(\tilde{A}(\mathcal{M}_+(\tilde{K}))\right).$$

Then there exists an absolutely continuous solution $\mu^\star$ *of* $A\mu = y$ *with positive and continuous density (on* $\tilde{K}$*).*

Proof. We make use of theorem 3.11. In order to do so, we need the dual cone of $\tilde{A}(\mathcal{M}_+(\tilde{K}))$ to have a nonempty interior, which (32) entails. Then we may build an absolutely continuous $\tilde{\mu}^\star \in \mathcal{M}_+(\tilde{K})$ with positive and continuous density, such that $\tilde{A}\tilde{\mu}^\star = \tilde{y}$. We then extend $\tilde{\mu}^\star$ to a measure on the whole of K by defining $\mu^\star$ to equal $\tilde{\mu}^\star$ on $\tilde{K}$ with support contained in $\tilde{K}$, namely $\mu^\star(B) = \tilde{\mu}^\star(B \cup \tilde{K})$ for any Borel subset B of K. Then $\mu^\star$ clearly solves $A\mu = y$ and thus minimises ℓ. □

Note that lemma 3.12 is a particular case of proposition 3.13, but we believe this presentation makes the role of $\mathrm{int}(A(\mathcal{M}_+))$ and $\partial A(\mathcal{M}_+)$ clearer.

Let us now finish this section by proving that not any point of the boundary may be associated to absolutely continuous measures. We denote S the simplex in $\mathbb{R}^m$, i.e.,

$$S := \left\{ w \in \mathbb{R}^m, w \geqslant 0 \;\middle|\; \sum_{i=1}^m w_i = 1 \right\}. \tag{33}$$

Proposition 3.14. *Assume that* $y \in \partial A(\mathcal{M}_+)$ *is an extremal point of* $A(\mathcal{M}_+) \cap S$*. Then any measure satisfying* $A\mu = y$ *is a Dirac mass.*

We omit the proof, which is straightforward and relies on the linearity of the operator A and the fact that the only extremal points among probability measures are the Dirac masses [35].

4. Properties of ML–EM

We now turn our attention to the ML–EM algorithm (14) for the minimisation of the functional ℓ (problem (11)).

4.1. Monotonicity and asymptotics

We first proceed to prove that the algorithm is monotonous, a property stemming from it being an expectation–maximisation algorithm.

We build a so-called *surrogate* function, i.e., a function Q_k such that $\ell(\mu) \leqslant Q_k(\mu)$ for all μ, with equality for $\mu = \mu_k$, where Q_k is minimised at μ_{k+1}. The precise details are in lemma 4.1.

Lemma 4.1. *For a given $k \in \mathbb{N}$, we define*

$$X_k := \{\mu \in \mathcal{M}_+ \mid \mu_{k+1} \ll \mu \ll \mu_k, \quad \langle \mu, 1\rangle = 1\}.$$

For a measure $\mu \in X_k$, and for $i = 1, \dots, m$, we define the probability distribution

$$\nu_i(\mu) := \frac{a_i \mu}{\langle \mu, a_i\rangle}.$$

as well as

$$Q_k(\mu) := \ell(\mu) + \sum_{i=1}^{m} y_i D\left(\nu_i(\mu_k) \| \nu_i(\mu)\right).$$

The following holds:

(a) $Q_k(\mu) \geq \ell(\mu), \quad \mu \in X_k$
(b) $Q_k(\mu_k) = \ell(\mu_k)$
(c) $Q_k(\mu) - Q_k(\mu_{k+1}) = D(\mu_{k+1}\|\mu), \quad \mu \in X_k$

Proof. The fact that $D(\nu_i(\mu_k)\|\nu_i(\mu))$ for $i = 1, \dots, m$ are divergences allows us to conclude about (a) and (b).

After defining

$$y_i^k := \langle \mu_k, a_i\rangle, \qquad 1 \leqslant i \leqslant m \tag{34}$$

and using the definition of ℓ in equation (12), we compute

$$Q_k(\mu) = 1 - \sum_{i=1}^{m} y_i \left\langle \nu_i(\mu_k), \log\left(y_i^k \frac{\mathrm{d}\mu}{\mathrm{d}\mu_k}\right)\right\rangle, \qquad \mu \in X_k$$

This gives

$$\begin{aligned} Q_k(\mu) - Q_k(\mu_{k+1}) &= \sum_{i=1}^{m} y_i \left\langle \nu_i(\mu_k), \log\left(y_i^k \frac{\mathrm{d}\mu_{k+1}}{\mathrm{d}\mu_k}\right) - \log\left(y_i^k \frac{\mathrm{d}\mu}{\mathrm{d}\mu_k}\right)\right\rangle \\ &= \sum_{i=1}^{m} y_i \left\langle \nu_i(\mu_k), \log\left(\frac{\mathrm{d}\mu_{k+1}}{\mathrm{d}\mu}\right)\right\rangle \\ &= \left\langle \underbrace{\sum_{i=1}^{m} y_i \nu_i(\mu_k)}_{\mu_{k+1}}, \log\left(\frac{\mathrm{d}\mu_{k+1}}{\mathrm{d}\mu}\right)\right\rangle \\ &= D\left(\mu_{k+1}\|\mu\right). \end{aligned}$$

which proves (c). □

Corollary 4.2. *For any $\mu_0 \in \mathrm{dom}(\ell)$, we have*

$$D\left(\mu_{k+1}\|\mu_k\right) \leqslant \ell(\mu_k) - \ell(\mu_{k+1}).$$

In particular,

$$\ell(\mu_{k+1}) \leqslant \ell(\mu_k)$$

Proof. First, observe that $\mu_{k+1} \in X_k$. Now, from (ii) and (i) in lemma 4.1, we obtain $Q_k(\mu_k) - Q_k(\mu_{k+1}) = \ell(\mu_k) - Q_k(\mu_{k+1}) \leqslant \ell(\mu_k) - \ell(\mu_{k+1})$. We conclude using (iii). □

Let us now prove that all cluster points of ML–EM are fixed points of the algorithm.

Proposition 4.3. *For any $\mu_0 \in \mathrm{dom}(\ell)$, any cluster point $\bar{\mu}$ of ML–EM is a fixed point of the algorithm, namely*

$$\bar{\mu} = \bar{\mu}\left(\sum_{i=1}^{m} \frac{y_i a_i}{\langle \bar{\mu}, a_i \rangle}\right).$$

Proof. We pursue an idea from [21]. Since we have $\mu_K(K) = \langle \mu_k, 1\rangle = \sum_{i=1}^m y_i = 1$ for all $k \geqslant 1$, (μ_k) is a bounded sequence in $\mathcal{M}_+$. By the Banach–Alaoglu theorem, it is thus weak-$*$ compact in $\mathcal{M}_+$, and we may extract some subsequence $\mu_{\varphi(k)}$ converging to a weak-$*$ cluster point $\bar{\mu}$ of ML–EM. Note first that ℓ is weak-$*$ continuous.

We now observe that such a cluster point must satisfy $\bar{\mu} \in \mathrm{dom}(\ell)$. Indeed, $\langle \bar{\mu}, a_i\rangle > 0$ for any $i \in \mathrm{supp}(y)$ (i.e., whenever $y_i > 0$). Otherwise, ℓ would go to infinity, a contradiction with the fact that ℓ decreases along iterates and $\ell(\mu_0) < +\infty$.

We also note that the convergence of $\ell(\mu_k)$ towards $\ell(\bar{\mu})$ is then along the whole sequence since $\{\ell(\mu_k) | k = 0, \ldots\}$ is decreasing.

Upon extracting another subsequence, we may assume that the sequence $(\mu_{\varphi(k)+1})$ is also convergent, say to $\tilde{\mu} \in \mathrm{dom}(\ell)$. Passing to the limit in the defining relation of ML–EM (as one readily checks that it is weak-$*$ continuous) along the subsequence, we find

$$\tilde{\mu} = \bar{\mu}\left(\sum_{i=1}^{m} \frac{y_i a_i}{\langle \bar{\mu}, a_i \rangle}\right),$$

and all it remains to show is that $\tilde{\mu} = \bar{\mu}$.

The inequality established in corollary 4.2 becomes

$$D\left(\mu_{\varphi(k)+1} \| \mu_{\varphi(k)}\right) \leqslant \ell(\mu_{\varphi(k)}) - \ell(\mu_{\varphi(k)+1}).$$

The right-hand side converges to 0. For the left-hand side, we use the property that the function $(\mu, \nu) \mapsto D(\mu \parallel \nu)$ is weak-$*$ lower semi-continuous [30]. This leads to $D(\tilde{\mu} \,|\, \bar{\mu}) \leqslant 0$, whence $\tilde{\mu} = \bar{\mu}$. □

Note that

$$\bar{\mu} = \bar{\mu}\left(\sum_{i=1}^{m} \frac{y_i a_i}{\langle \bar{\mu}, a_i \rangle}\right) \iff \sum_{i=1}^{m} \frac{y_i a_i}{\langle \bar{\mu}, a_i \rangle} = 1 \text{ on } \mathrm{supp}(\bar{\mu}).$$

Thus, ML–EM cluster points satisfy one of the two optimality conditions (25). Although we conjecture they actually satisfy both of them under the additional hypothesis that $\mathrm{supp}(\mu_0) = K$, we are able to prove it only when $y \in C$.

4.2. Case $y \notin A(\mathcal{M}_+)$

When $y \notin A(\mathcal{M}_+)$, we know from corollary 3.8 that optimal solutions are sparse. Note that this is also the case for boundary points which are extremal in $A(\mathcal{M}_+) \cap S$, in virtue of proposition 3.14.

We do not know whether ML–EM iterates converge to an optimal point, but we can at least state a straightforward partial sparsity result from the first optimality condition, which we call *weak sparsity*.

Corollary 4.4. *Assume that $y \notin A(\mathcal{M}_+)$ and the linear independence condition (28). Then, for $\mu_0 \in \mathrm{dom}(\ell)$, any cluster point $\bar{\mu}$ of ML–EM is such that*

$$\mathrm{supp}(\bar{\mu}) \subset \left(\sum_{i=1}^{m} \lambda_i(A\bar{\mu})\, a_i \right)^{-1} (\{0\}),$$

with $\lambda(A\bar{\mu}) \neq 0$ and the components $\lambda_i(A\bar{\mu})$'s do not have the same sign. In particular, $\mathrm{supp}(\bar{\mu}) \neq K$.

Proof. This is just a rephrasing of proposition 4.3, using the formula for $\lambda(\mu)$ given in equation (22). □

Remark 4.5. In general, the fact that the components $\lambda_i(A\bar{\mu})$'s do not have the same sign will impose that $\mathrm{supp}(\bar{\mu})$ is restricted to a lower dimensional set, of Lebesgue measure 0. Thus, one cannot expect that the cluster points of ML–EM are absolutely continuous when $y \notin A(\mathcal{M}_+)$.

We can go further in the case where $y_i = 0$ except for $y_{i_0} = 1$, for which we saw in remark 3.10 that any minimiser $\mu^\star$ of the function ℓ for normalised measures, i.e., $\ell(\mu) = 1 - \log\left(\langle \bar{\mu}, a_{i_0} \rangle\right)$, will be such that $\mathrm{supp}(\mu^\star) \subset \arg\max a_{i_0}$. The goal of this subsection is to highlight a case where one clearly identifies the limiting measure of ML–EM, and its dependence with respect to the initial measure μ_0. This suggests that in the sparse case $y \notin A(\mathcal{M}_+)$, the position of Dirac masses will in general depend on the initial condition μ_0.

We recall Laplace's method (see [40]) which holds for $f \in \mathcal{C}(K)$, $g \in \mathcal{C}^2(K)$ with a single non-degenerate interior maximum point $\bar{x}$, and reads:

$$\int_K f(x) \mathrm{e}^{g(x)t}\, \mathrm{d}x \sim (2\pi)^{p/2} \frac{f(\bar{x})}{\sqrt{|\det(H(\bar{x}))|}} \frac{\mathrm{e}^{g(\bar{x})t}}{t^{\frac{p}{2}}} \quad \text{as } t \to \infty, \tag{35}$$

where $H(\bar{x})$ is the Hessian of g at $\bar{x}$.

Proposition 4.6. *Assume that $y_i = 0$ for $i \neq i_0$, $y_{i_0} = 1$. Assume further that* $\arg\max a_{i_0} = \{\bar{x}_1, \ldots, \bar{x}_l\}$ *with $\bar{x}_j \in \mathrm{int}(K)$ for all $j = 1, \ldots, l$, that a_{i_0} is of class $\mathcal{C}^2$ and that the maximum points $\bar{x}_j$ are non-degenerate. Under these assumptions and for μ_0 absolutely continuous with continuous positive density (still denoted μ_0), the ML–EM sequence (μ_k) satisfies*

$$\mu_k \rightharpoonup \mu^\star := C \sum_{j=1}^{l} \frac{\mu_0(\bar{x}_j)}{\sqrt{|\det H_j|}} \delta_{\bar{x}_j},$$

where $C > 0$ is a normalising constant such that the limit has mass one, H_j is the Hessian of a_{i_0} at the point $\bar{x}_j$, and $\delta_{\bar{x}_j}$ is the Dirac mass centred at $\bar{x}_j$.

Proof. We first remark that the ML–EM iterates are then explicitly solved as

$$\mu_k = \frac{a_{i_0}^k \mu_0}{\int_K a_{i_0}^k(x)\mu_0(x)\,\mathrm{d}x}.$$

We denote $M := \max_{x\in K} \log(a_{i_0}(x))$ and let $f \in \mathcal{C}(K)$ be a generic function. For δ small enough such that, for all j, $\bar{x}_j$ is the unique maximum point of a_{i_0} in $B(\bar{x}_j,\eta)$, we split contributions in the integral of $a_{i_0}^k\mu_0$ against f as follows

$$\begin{aligned}\langle a_{i_0}^k\mu_0, f\rangle &= \int_K f(x)\mu_0(x)\mathrm{e}^{k\,\log(a_{i_0}(x))}\,\mathrm{d}x \\ &= \sum_{j=1}^{l}\int_{B(\bar{x}_j,\eta)} f(x)\mu_0(x)\mathrm{e}^{k\,\log(a_{i_0}(x))}\,\mathrm{d}x \\ &\quad + \int_{K\setminus\cup_{j=1}^{l}B(\bar{x}_j,\eta)} f(x)\mu_0(x)\mathrm{e}^{k\,\log(a_{i_0}(x))}\,\mathrm{d}x.\end{aligned}$$

From Laplace's method (35), each term in the first sum can be estimated as

$$\int_{B(\bar{x}_j,\eta)} f(x)\mu_0(x)\mathrm{e}^{k\,\log(a_{i_0}(x))}\,\mathrm{d}x \sim (2\pi)^{p/2}\frac{f(\bar{x}_j)\mu_0(\bar{x}_j)}{\sqrt{|\det H_j|}}\frac{\mathrm{e}^{Mk}}{k^{\frac{p}{2}}} \quad \text{as } k\to\infty,$$

whereas one can check that the second term is $o\left(\mathrm{e}^{Mk}\right)$. We end up with

$$\langle a_{i_0}^k\mu_0, f\rangle \sim (2\pi)^{p/2}\sum_{j=1}^{l}\left(\frac{f(\bar{x}_j)\mu_0(\bar{x}_j)}{\sqrt{|\det H_j|}}\right)\frac{\mathrm{e}^{Mk}}{k^{\frac{p}{2}}} \quad \text{as } k\to\infty.$$

Applying this equivalent for $f = 1$ yields an equivalent for the denominator $\int_K a_{i_0}^k(x)\mu_0(x)\,\mathrm{d}x$ in the explicit formula for μ_k, which is of the order of $\mathrm{e}^{Mk}/k^{\frac{p}{2}}$. All in all, we find

$$\langle \mu_k, f\rangle \longrightarrow C\sum_{j=1}^{l}\left(\frac{f(\bar{x}_j)\mu_0(\bar{x}_j)}{\sqrt{|\det H_j|}}\right) = \left\langle C\sum_{j=1}^{l}\frac{\mu_0(\bar{x}_j)}{\sqrt{|\det H_j|}}\delta_{\bar{x}_j}, f\right\rangle = \langle \mu^\star, f\rangle,$$

as $k\to\infty$, with $C>0$ normalising $\mu^\star$, which proves the claim. □

4.3. Case $y \in A(\mathcal{M}_+)$

When the data y is in the cone $A(\mathcal{M}_+)$, there are infinitely many measures satisfying $A\mu = y$, and when there are absolutely continuous ones, a desirable property of ML–EM is to converge to one of them rather than to a measure having a singular part. In order to address this question, we start with a proposition of independent interest, valid for any data y. It generalises a result which holds in the discrete case. It gives information on the divergence of ML–EM iterates to *any* fixed point of the ML–EM algorithm.

Proposition 4.7. *Let $\bar{\mu}$ be a fixed point of the ML–EM algorithm, and $\mu_0 \in \mathrm{dom}(\ell)$. We further assume that*

$$D(\bar{\mu}\|\mu_0) < \infty.$$

Then the ML–EM iterates are such that

$$\forall k \in \mathbb{N},\ D(\bar{\mu}\|\mu_{k+1}) \leqslant D(\bar{\mu}\|\mu_k) - \ell(\mu_k) + \ell(\bar{\mu}).$$

In particular, the KL divergence to any optimum decreases:

$$\forall k \in \mathbb{N},\ D(\bar{\mu}\|\mu_{k+1}) \leqslant D(\bar{\mu}\|\mu_k).$$

Proof. We recursively prove that $D(\bar{\mu}\|\mu_k) < \infty$. Assuming this holds at the step k, one checks that the definition of ML–EM iterates is such that if μ_k is absolutely continuous with respect to $\bar{\mu}$, then so is μ_{k+1}. Furthermore,

$$\begin{aligned}
-\log\left(\frac{\mathrm{d}\mu_{k+1}}{\mathrm{d}\bar{\mu}}\right) &= -\log\left(\frac{\mathrm{d}\mu_k}{\mathrm{d}\bar{\mu}}\right) - \log\left(\sum_{i=1}^m \frac{y_i a_i}{\langle \mu_k, a_i\rangle}\right) \\
&= -\log\left(\frac{\mathrm{d}\mu_k}{\mathrm{d}\bar{\mu}}\right) - \log\left(\sum_{i=1}^m \frac{y_i a_i}{\langle \mu_k, a_i\rangle}\right) \sum_{i=1}^m \frac{y_i a_i}{\langle \bar{\mu}, a_i\rangle} \\
&= -\log\left(\frac{\mathrm{d}\mu_k}{\mathrm{d}\bar{\mu}}\right) + \sum_{i=1}^m \frac{y_i a_i}{\langle \bar{\mu}, a_i\rangle} \log\left(\sum_{i=1}^m \frac{y_i a_i}{\langle \bar{\mu}, a_i\rangle} \Big/ \sum_{i=1}^m \frac{y_i a_i}{\langle \mu_k, a_i\rangle}\right),
\end{aligned} \tag{36}$$

where we twice took advantage of $\sum_{i=1}^m \frac{y_i a_i}{\langle \bar{\mu}, a_i\rangle} = 1$ on $\mathrm{supp}(\bar{\mu})$ by definition of a fixed point, a property we may use since we will eventually integrate against $\mathrm{d}\bar{\mu}$.

Now, we use the convexity of $(u, v) \mapsto u \log(\frac{u}{v})$ on $[0, +\infty) \times (0, +\infty)$ (following an idea of [17]) to bound the second term as follows

$$\sum_{i=1}^m \frac{y_i a_i}{\langle \bar{\mu}, a_i\rangle} \log\left(\sum_{i=1}^m \frac{y_i a_i}{\langle \bar{\mu}, a_i\rangle} \Big/ \sum_{i=1}^m \frac{y_i a_i}{\langle \mu_k, a_i\rangle}\right) \leqslant \sum_{i=1}^m y_i \frac{a_i}{\langle \bar{\mu}, a_i\rangle} \log\left(\frac{\langle \mu_k, a_i\rangle}{\langle \bar{\mu}, a_i\rangle}\right).$$

When integrated against $\mathrm{d}\bar{\mu}$, the right hand side simplifies to

$$\sum_{i=1}^m y_i \log\left(\frac{\langle \mu_k, a_i\rangle}{\langle \bar{\mu}, a_i\rangle}\right) = -\ell(\mu_k) + \ell(\bar{\mu}).$$

Wrapping up, the integration of (36) against $\mathrm{d}\bar{\mu}$ and the above inequality exactly yield the result. □

Remark 4.8. As we saw, we typically expect a fixed point $\bar{\mu}$ to be a sparse measure when $y \notin A(\mathcal{M}_+)$, which means that the assumption that μ_0 is absolutely continuous with respect to $\bar{\mu}$ will typically not be satisfied for μ_0 chosen to be constant over K. This result will instead come in handy when $y \in A(\mathcal{M}_+)$.

Corollary 4.9. *Assume that $y \in A(\mathcal{M}_+)$, and that there exists an absolutely continuous measure $\mu^\star$ with positive and continuous density (on $\tilde{K}$) such that $A\mu^\star = y$. Then, for any μ_0 absolutely continuous, with a positive and continuous density, any cluster point $\bar{\mu}$ satisfies* $\mathrm{supp}(\bar{\mu}) = \tilde{K}$.

Proof. Let $\mu^\star$ be such an absolutely continuous optimum with positive and continuous density on $\tilde{K}$. The assumption on the initial measure μ_0 ensures that it satisfies conditions (15), i.e., $\mu_0 \in \mathrm{dom}(\ell)$, and since μ_1 is then continuous and positive on $\tilde{K}$, we also have $D(\mu^\star \,\|\, \mu_1) < \infty$. We may then use proposition 4.7 to obtain

$$\forall k \in \mathbb{N}^\star,\ D(\mu^\star\|\mu_{k+1}) \leqslant D(\mu^\star\|\mu_k).$$

Let $\bar{\mu}$ be a cluster point of the iterates $\{\mu_k\}$. Note that we obviously have $\mathrm{supp}(\bar{\mu}) \subset \tilde{K}$. By weak-$*$ lower semi-continuity, we may pass to the limit in the inequality to obtain

$$D(\mu^\star \| \bar{\mu}) \leqslant \lim_{k\to+\infty} D(\mu^\star \| \mu_k).$$

In particular, $D(\mu^\star\|\bar{\mu}) < +\infty$, which by definition implies that $\mu^\star \ll \bar{\mu}$, and consequently $\bar{\mu}$ has support at least $\mathrm{supp}(\mu^\star) = \tilde{K}$. □

We are now in a position to prove the main result of this section, where we use the notations of proposition 3.13.

Theorem 4.10. *Assume that conditions* (31) *and* (32) *hold, and that* $\tilde{y} \in \mathrm{int}\left(\tilde{A}\left(\mathcal{M}_+(\tilde{K})\right)\right)$. *Then, if the initial measure* μ_0 *is absolutely continuous, with a positive and continuous density, the cluster points of the ML–EM iterates*

(*a*) *have support* $\tilde{K}$,

(*b*) *are optimal.*

In particular, the algorithm is convergent in the sense that

$$l(\mu_k) \xrightarrow[k\to+\infty]{} \inf_{\mu\in\mathcal{M}_+} l(\mu).$$

Proof. Under these assumptions, by proposition 3.13, there exists an absolutely continuous measure $\mu^\star$ with positive and continuous density on $\tilde{K}$, such that $A\mu^\star = y$. We may thus apply corollary 4.9 to conclude that any cluster point must be absolutely continuous with support equal to $\tilde{K}$, showing (a).

Letting $\bar{\mu}$ be a cluster point, it remains to show the optimality (b), namely that $A\bar{\mu} = y$. As a cluster point, it must be fixed point of the algorithm from proposition 4.3, i.e. $\bar{\mu} = \bar{\mu}\left(\sum_{i=1}^m \frac{y_i a_i}{\langle \bar{\mu}, a_i\rangle}\right)$. Since $\mathrm{supp}(\bar{\mu}) = \tilde{K}$, we obtain

$$\sum_{i=1}^m \left(1 - \frac{y_i}{\langle \bar{\mu}, a_i\rangle}\right) a_i = 0,$$

on $\tilde{K}$. By the linear independence assumption (31), this imposes $\langle \bar{\mu}, a_i\rangle = y_i$ for all $i \in \{1,\dots,m\}$, whence the optimality of $\bar{\mu}$.

As for the convergence of the algorithm, we recall that the whole sequence $(l(\mu_k))$ is decreasing and hence, converges. Extracting such that a subsequence of the ML–EM iterates converges to an optimal measure, the limit of $(l(\mu_k))$ is identified to be $\inf_{\mu\in\mathcal{M}_+} l(\mu)$. □

Note that these results also cover the case of lemma 3.12: if $y \in \mathrm{int}(A(\mathcal{M}_+))$ and the functions $\{a_i\}$ are linearly independent, cluster points of ML–EM are absolutely continuous, whenever the initial measure μ_0 is absolutely continuous with a positive and continuous density.

Remark 4.11. This result does not mean that cluster points do not have a singular part, such as Dirac masses. However, the fact that cluster points have full support is relevant in practice: cropping the image obtained by the ML–EM algorithm allows one to reduce the effect of the singular part and to uncover the continuous one (assuming that the measure does not have any continuous singular part). This is also what should be obtained by smoothing the final image, or by regularising the functional ℓ, see [29] for an analysis of such regularisation techniques.

5. Statistics

In this section, we estimate the probability that the data y stays in the image cone, i.e., $y \in A(\mathcal{M}_+)$.

Let us first go back to modelling (before any normalisation) by introducing a *dose variable* t. We assume that the real image is given by $\mu_r \in \mathcal{M}_+$ which represents the image in the relevant unit depending on the context (*Bq* for PET). The dosage t gives rise to independent random variables $N_i \sim \mathcal{P}(\gamma_i t)$ where $\gamma_i := \langle \mu_r, a_i \rangle$, whose sum $N = \sum_{i=1}^m N_i$ is $\mathcal{P}(\gamma t)$ with $\gamma := \sum_{i=1}^m \gamma_i$.

The expected frequencies are given by $y_r := (\frac{\gamma_1}{\gamma}, \dots, \frac{\gamma_m}{\gamma})$, which is an element of $A(\mathcal{M}_+) \cap S$, where we recall that S, given by (33), is the simplex in $\mathbb{R}^m$.

With our previous notations, n_i and n are thus realisations of the random variables N_i, and N, and now $y = (\frac{N_1}{N}, \dots, \frac{N_m}{N})$ is a random variable. We emphasise that it is an estimator for y_r which depends on t by using the notation $\hat{y}_t$. When conditioned on the fact that $N = n$, we will denote $\hat{y}_n := (\frac{N_1}{n}, \dots, \frac{N_m}{n})$.

By the law of large numbers, $\hat{y}_t$ tends to y_r almost surely as $t \to +\infty$, so we know that if $y_r \in \mathrm{int}(A(\mathcal{M}_+))$, a high-enough dose will ensure that $\hat{y}_t \in \mathrm{int}(A(\mathcal{M}_+))$ as well, avoiding having only sparse measures as solutions to the maximum likelihood problem.

We now wish to give quantitative bounds for $\mathbb{P}(\hat{y}_t \notin A(\mathcal{M}_+))$, one with a conditioning on the number of events n, the other without such a conditioning. The aim is to address the following questions:

- *a posteriori*, for a given number of points n, how small is the probability that $\hat{y}_n \notin A(\mathcal{M}_+)$?
- *a priori*, how large should the dosage t be for the probability that $\hat{y}_t \notin A(\mathcal{M}_+)$ to be small enough?

The celebrated Sanov's theorem [36] states that the empirical distribution has an exponentially small probability of being in a set which does not contain the real distribution, where the exponential is controlled by the Kullback–Leibler divergence from the real distribution to the set. We thus define

$$\varepsilon := \inf_{q \in S \cap A(\mathcal{M}_+)^c} \mathrm{d}(q \| y_r),$$

which is the Kullback–Leibler divergence of y_r to the boundary of the set $A(\mathcal{M}_+)$ (intersected with the simplex S).

In both cases, we shall give two different bounds which might be relevant in different regimes in the parameters (m, n) and (m, t), respectively.

Proposition 5.1. *The following concentration bounds hold:*

$$\mathbb{P}(\hat{y}_n \notin A(\mathcal{M}_+)) \leqslant \begin{cases} (n+1)^m \, \mathrm{e}^{-n\varepsilon}, \\ 2m \, \mathrm{e}^{-n\varepsilon/m}. \end{cases} \tag{37}$$

Proof. Conditioned on $N = n$, the random vector $\hat{y}_n$ follows the multinomial distribution of parameters n and y_r. The first inequality is then nothing but a direct application of Sanov's theorem [36]. The second is more recent and given in lemma 6 of [22]. □

We now proceed to the case with dose t:

Theorem 5.2. *The following concentration bounds hold:*

$$\mathbb{P}(\hat{y}_t \notin A(\mathcal{M}_+)) \leqslant \begin{cases} C(m)(1+(\gamma t)^m)\,\mathrm{e}^{-\gamma t\varepsilon}, \\ 2m\,\mathrm{e}^{-\gamma t\varepsilon/m}, \end{cases}$$

where $C(m)$ is a combinatorial constant which depends only on m and satisfies $C(m) \leqslant \left(\frac{a(m+1)}{\log(m+2)}\right)^{m+1}$, with $a = 0.792$.

Proof. We may write

$$\mathbb{P}(\hat{y}_t \notin A(\mathcal{M}_+)) = \mathbb{E}(\mathbb{P}(\hat{y}_N \notin A(\mathcal{M}_+))\,|N) \leqslant \mathbb{E}(g(N))$$

where $g(n) = (n+1)^m\,\mathrm{e}^{-n\varepsilon}$ or $2m\,\mathrm{e}^{-n\varepsilon/m}$ from the previous proposition. It is now a matter of estimating this expectation with $N \sim \mathcal{P}(\gamma t)$. In the second case,

$$\begin{aligned}\mathbb{E}(g(N)) &= 2m\,\mathrm{e}^{-\gamma t}\sum_{n=0}^{+\infty}\mathrm{e}^{-n\varepsilon/m}\frac{(\gamma t)^n}{n!} \\ &= 2m\,\mathrm{e}^{-\gamma t(1-\exp(-\varepsilon/m))} \leqslant 2m\,\mathrm{e}^{-\gamma t\varepsilon/m},\end{aligned}$$

from $1-\mathrm{e}^{-u} \geqslant u$.

In the first case, $\mathbb{E}(g(N)) = \mathrm{e}^{-\gamma t}\varphi_m(\gamma t\mathrm{e}^{-\varepsilon})$, with

$$\varphi_m(x) := \sum_{n=0}^{+\infty}(n+1)^m\,\frac{x^n}{n!},$$

and $x := \gamma t\mathrm{e}^{-\varepsilon}$, which we now estimate. We may integrate to find

$$\int_0^x \varphi_m(u)\,\mathrm{d}u = \sum_{n=1}^{+\infty} n^m\frac{x^n}{n!} =: T_m(x)\mathrm{e}^x,$$

where T_m is the so-called Touchard polynomial of order m, which has degree m.

This allows to go back to $\varphi_m(x)$ as $\varphi_m(x) = \left(T_m'(x)+T_m(x)\right)\mathrm{e}^x = \frac{T_{m+1}(x)}{x}\mathrm{e}^x$, using a well-known property of Touchard polynomials. The Touchard polynomial of order m has integer coefficients (the Stirling numbers), whose sum is given by the so-called Bell number B_m.

Using the crude bound $P(x) = \sum_{k=0}^m a_k x^k \leqslant \left(\sum_{k=0}^m a_k\right)(1+x^m)$ valid for all $x \geqslant 0$ when P is a polynomial with non-negative coefficients, we may write $\frac{T_{m+1}(x)}{x} \leqslant B_{m+1}(1+x^m)$ for $x \geqslant 0$.

Summing up, we have

$$\begin{aligned}\mathbb{E}(g(N)) &\leqslant B_{m+1}\mathrm{e}^{-\gamma t}(1+x^m)\mathrm{e}^x = B_{m+1}(1+(\gamma t\mathrm{e}^{-\varepsilon})^m)\mathrm{e}^{-\gamma t(1-\exp(-\varepsilon))} \\ &\leqslant C(m)(1+(\gamma t)^m)\,\mathrm{e}^{-\gamma t\varepsilon},\end{aligned}$$

where $C(m) := B_{m+1}$. The bound about Bell numbers such as $C(m)$, stated in the proposition, can be found in [4]. We use them here as bounds on the moments of a Poisson random variable. □

Remark 5.3. Although the bound coming from Sanov's theorem is sharper at the limit $n \to +\infty$ or $t \to +\infty$, it may be that the other one is relevant for realistic values m, n or t. If these bounds are taken as functions of ε, it can be checked that the alternative bound becomes more stringent in the regime where $\varepsilon \ll \frac{m}{n}\log(n)$.

6. Numerical simulations

We perform simulations using the Python library Operator Discretization Library [1]. The interested reader may run our numerical experiments using a Jupyter Notebook [31].

All simulations are run with a 2D PET operator A having $9.0 \times 10^{+1}$ views and $1.28 \times 10^{+2}$ tangential positions, leading to a number of (pairs of) detectors $m = 11\,520$. The image resolution is 256×256. We draw $\bar{y}_t \sim \frac{1}{t}\mathcal{P}(tA\mu_r)$ for different doses t, so that the higher t, the lower the noise level. We then normalise $(y_t)_i := (\bar{y}_t)_i / \sum_i (\bar{y}_t)_i$ for $i = 1, \dots, m$.

In figure 2, we consider five different noise levels, associated to different values of dose t. For each of these values, we are interested in seeing whether iterations lead to sparse measures or not. From our results, this is equivalent to testing if $y_t \in A(\mathcal{M}_+)$. A first crude estimate of this problem is to plot $d(y_t \| A\mu_k)$ and check whether this quantity converges to zero, which by theory implies $y_t \in A(\mathcal{M}_+)$. We also look at the 95% evolution of the percentile along the iterations. A low percentile means that the mass concentrates on the remaining five percent of the image.

6.1. Dual certificates

It is difficult to make sure that the divergence to the data y converges to zero. We thus also look for *dual certificates* λ in the dual cone $A(\mathcal{M}_+)^*$ (see section 2.4) such that the dual function g defined in (26) fulfils $g(\lambda) > 0$. Indeed, weak duality ensures that $\min d(y_t \| A\mu) \geqslant g(\lambda)$, so the existence of any dual certificate proves that $y_t \notin A(\mathcal{M}_+)$.

In order to find a good choice for the certificate λ, we compute $\lambda_k = 1 - \frac{y_t}{A\mu_k}$ along iterates, which we know should converge to the optimal dual variable $\lambda^\star$ in the case where $y_t \in A(\mathcal{M}_+)$, and we conjecture this is true in full generality. Recall that the dual optimal variable is such that $A^*\lambda^\star \geqslant 0$, i.e., $\lambda^\star \in A(\mathcal{M}_+)^*$. For a fixed number of iterations k, there is no reason that $\lambda_k \in A(\mathcal{M}_+)^*$, so we add a small appropriate constant c to λ_k and check that it provides a dual certificate, that is, we check whether $g(\lambda_k + c) > 0$.

This procedure allows us to certify that in the three noisiest cases of figure 2, the data y_t is not in the cone $A(\mathcal{M}_+)$. We thus expect sparsity in those three cases. We note that the resulting image after $4.0 \times 10^{+2}$ iterates is not completely sparse: more iterates are required for only the Dirac masses to remain. We expect the convergence to the sum of Dirac masses to be very slow.

In the two less noisy cases, we could not certify that the data y_t is not in the cone $A(\mathcal{M}_+)$, although that does not mean that the converse should be true. In fact, it could well be that even for relatively low levels of noise, sparsity is the outcome but Dirac masses only take over after a practically unrealistic number of iterates.

7. Open problems and perspectives

7.1. Convergence of iterates

As shown in theorem 4.10, cluster points of the weak-$*$ cluster points of ML–EM are optimal when $y \in A(\mathcal{M}_+)$. A problem left open is their optimality when $y \notin A(\mathcal{M}_+)$. That result would imply a strong sparsity result for ML–EM cluster points.

Another problem is the convergence of the whole sequence to a single point, which is a tall order since there are in general many optimal points. The discrete equivalent to proposition 4.7 allows to prove the full convergence of iterates in the discrete case, owing to the continuity of the discrete Kullback–Leibler divergence [39] at cluster points. In continuum, the fact that the

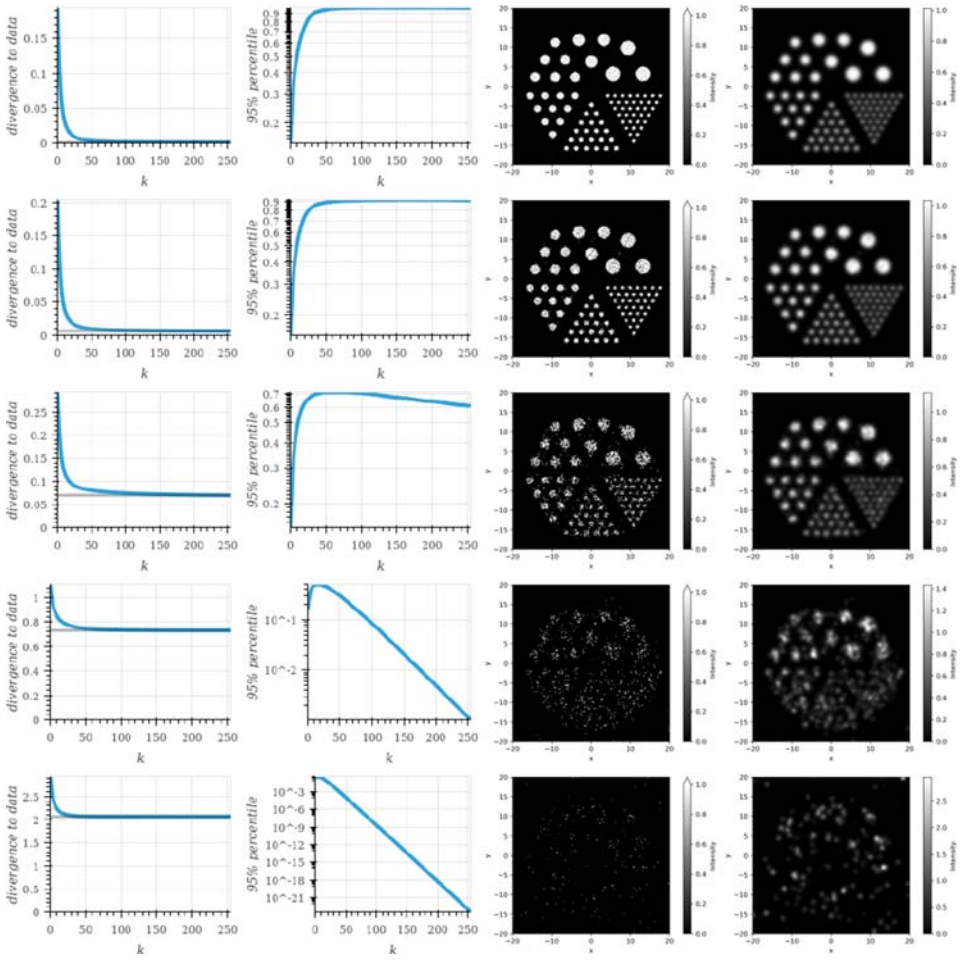

Figure 2. We show here various reconstructions for a decreasing amount of dose (the first row has $t = 10^2$ and each subsequent row has ten times less dose than the previous one). The columns depict (a) the divergence to the data $d(y_t \| A\mu^k)$ (b) the 95% percentile (logarithmic scale) (c) the reconstruction with limitations between zero and one (d) a smoothed reconstruction (three pixels wide Gaussian convolution).
It is apparent that (i) when there is too much noise, the divergence to the data does not converge to zero (ii) the percentile in the second column shows that when the noise is large enough, MLEM iterations quickly increase the sparsity.

divergence $D(\mu \| \nu)$ may be infinite even μ is absolutely continuous with respect to ν does not allow us to obtain a similar result, although we conjecture it does hold true.

7.2. Regularisation

Another interesting issue in light of our sparsity results is that of regularisation. How should one choose appropriate additional regularisation terms to alleviate the problem? Similarly and as is done in [33] for continuous data, analysing the alternative strategy of regularising by early stopping is worthy of interest, since this is usual practice in PET.

7.3. Going further in the case of PET

For PET, the functions a_i are actually close to being singular measures concentrated on a line. Studying the effect of this near-singularity on our results is of practical interest. More precisely, one could for specific geometries analyse the typical minimum set of a function of the form $A^*\lambda = \sum_{i=1}^m \lambda_i a_i$.

It would also be natural to look at the effect of binning (i.e., aggregating detectors) on the constant $\inf_{q\in S\cap A(\mathcal{M}_+)^c} d(q, y_r)$. Indeed, theorem 5.2 shows its importance when it comes to sparsity, justifying to try and make this distance as large as possible.

7.4. Sparsity results in general

We intend to investigate the generality of these sparsity results in the context of other divergences. The squared-distance is a popular one, but generalisations have recently been advocated for in the literature, such as the β-divergences [10]. Finally, we believe our results can be extended without too much difficulty to the generalised statistical model for PET where scatter and random events are taking into account, namely $y = \mathcal{P}(A\mu + s)$ with s a known vector standing for the counts of scatter and random events.

Acknowledgments

We are grateful to Sebastian Banert for fruitful discussions about optimisation in Banach spaces, as well as Axel Ringh and Johan Karlsson for bringing the moment matching problem to our attention. We acknowledge support from the Swedish Foundation of Strategic Research Grant AM13-004.

ORCID iDs

Camille Pouchol https://orcid.org/0000-0002-1152-9896
Olivier Verdier https://orcid.org/0000-0003-3699-6244

References

[1] Adler J, Kohr H and Öktem O 2017 ODL-a Python framework for rapid prototyping in inverse problems *R. Inst. Technol.*
[2] Baumeister J and Leitão A 2017 *Topics in Inverse Problems*
[3] Benvenuto F and Piana M 2014 Regularization of multiplicative iterative algorithms with nonnegative constraint *Inverse Problems* **30** 035012
[4] Berend D and Tassa T 2010 Improved bounds on Bell numbers and on moments of sums of random variables *Probab. Math. Statistics* **30** 185–205
[5] Bertero M and Boccacci P 1998 *Introduction to Inverse Problems in Imaging* (Boca Raton, FL: CRC Press)
[6] Boyd S and Vandenberghe L 2004 *Convex Optimization* (Cambridge: Cambridge University Press)
[7] Byrne C 1993 Iterative image reconstruction algorithms based on cross-entropy minimization *IEEE Trans. Image Process.* **2** 96–103
[8] Byrne C 1995 Erratum and addendum to "Iterative image-reconstruction algorithms based on cross-entropy minimization" (New York: Springer)
[9] Byrne C 1996 Iterative reconstruction algorithms based on cross-entropy minimization *Image Models (and their Speech Model Cousins)* (Berlin: Springer) pp 1–11
[10] Cavalcanti Y C, Oberlin T, Dobigeon N, Févotte C, Stute S, Ribeiro M-J and Tauber C 2019 Factor analysis of dynamic PET images: beyond Gaussian noise *IEEE Trans. Med. Imag.*
[11] Csiszár I 1984 *Information Geometry and Alternating Minimization Procedures Statistics and Decisions* vol 1 pp 205–37

[12] Dempster A P, Laird N M and Rubin D B 1977 Maximum likelihood from incomplete data via the EM algorithm *J. R. Stat. Soc. Ser. B Methodol.* **39** 1–22
[13] Fessler J A, Clinthome N H and Rogers W L 1993 On complete-data spaces for PET reconstruction algorithms *IEEE Trans. Nuc. Sci.* **40** 1055–61
[14] Georgiou T T 2005 Solution of the general moment problem via a one-parameter imbedding *IEEE Trans. Autom. Control* **50** 811–26
[15] Hinkle J, Szegedi M, Wang B, Salter B and Joshi S 2012 4D CT image reconstruction with diffeomorphic motion model *Med. Image Anal.* **16** 1307–16
[16] Hudson H M and Larkin R S 1994 Accelerated image reconstruction using ordered subsets of projection data *IEEE Trans. Med. Imag.* **13** 601–9
[17] Iusem A N 1992 A short convergence proof of the EM algorithm for a specific poisson model *Braz. J. Probab. Statistics* 57–67
[18] Jacobson M and Fessler J A 2003 Joint estimation of image and deformation parameters in motion-corrected PET 2003 IEEE Nuclear Science Symposium. Conference Record (IEEE Cat. No. 03CH37515) vol 5 (Piscataway, NJ: IEEE) pp 3290–4
[19] Last G and Penrose M 2017 *Lectures on the Poisson Process* vol 7 (Cambridge: Cambridge University Press)
[20] Lucy L B 1974 An iterative technique for the rectification of observed distributions *Astron. J.* **79** 745
[21] Mair B, Rao M and Anderson J 1996 Positron emission tomography, Borel measures and weak convergence *Inverse Problems* **12** 965
[22] Mardia J, Jiao J, Tánczos E, Nowak R D and Weissman T 2018 Concentration inequalities for the empirical distribution arXiv:1809.06522
[23] Mülthei H 1992 Iterative continuous maximum-likelihood reconstruction method *Math. Methods Appl. Sci.* **15** 275–86
[24] Mülthei H, Schorr B and Törnig W 1987 On an iterative method for a class of integral equations of the first kind *Math. Methods Appl. Sci.* **9** 137–68
[25] Mülthei H, Schorr B and Törnig W 1989 On properties of the iterative maximum likelihood reconstruction method *Math. Methods Appl. Sci.* **11** 331–42
[26] Natterer F and Wübbeling F 2001 *Mathematical Methods in Image Reconstruction* vol 5 (Philadelphia, PA: SIAM)
[27] Öktem O, Pouchol C and Verdier O 2019 Spatiotemporal PET reconstruction using ML-EM with learned diffeomorphic deformation in *International Workshop on Machine Learning for Medical Image Reconstruction* (Berlin: Springer) pp 151–62
[28] Ollinger J M and Fessler J A 1997 Positron-emission tomography *IEEE Signal Process. Mag.* **14** 43–55
[29] O'Sullivan F 1995 A study of least squares and maximum likelihood for image reconstruction in positron emission tomography *Ann. Statistics* 1267–300
[30] Posner E 1975 Random coding strategies for minimum entropy *IEEE Trans. Inf. Theory* **21** 388–91
[31] Pouchol C and Verdier O *MLEM Experiment Notebook* https://github.com/olivierverdier/mlem_notebook
[32] Qi J and Leahy R M 2006 Iterative reconstruction techniques in emission computed tomography *Phys. Med. Biol.* **51** R541
[33] Resmerita E, Engl H W and Iusem A N 2007 The expectation-maximization algorithm for ill-posed integral equations: a convergence analysis *Inverse Problems* **23** 2575
[34] Richardson W H 1972 Bayesian-based iterative method of image restoration *JoSA* **62** 55–9
[35] Rudin W 1991 Functional analysis *International Series in Pure and Applied Mathematics* 2nd edn (New York: McGraw-Hill)
[36] Sanov I N 1961 On the probability of large deviations of random variables *Sel. Transl. Math. Statistics Probab.* **1** 213–44
[37] Shepp L A and Vardi Y 1982 Maximum likelihood reconstruction for emission tomography *IEEE Trans. Med. Imaging* **1** 113–22
[38] Silverman B, Jones M, Wilson J and Nychka D 1990 A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography *J. R. Stat. Soc. Ser. B Methodol.* **52** 271–303
[39] Vardi Y, Shepp L and Kaufman L 1985 A statistical model for positron emission tomography *J. Am. Stat. Assoc.* **80** 8–20
[40] Wong R 2001 *Asymptotic Approximations of Integrals* vol 34 (Philadelphia, PA: SIAM)