

CONVERGENCE OF THE EM ALGORITHM VIA PROXIMAL TECHNIQUES

DOMINIKUS NOLL

ABSTRACT. We investigate convergence of the expectation maximization algorithm by representing it as a generalized proximal method. Convergence of iterates and not just in value is investigated under natural hypotheses such as definability of the incomplete data log-likelihood in the sense of o-minimal structure theory.

Key words: EM algorithm · definable sets · o-minimal structures · proximal method · Kullback-Leibler divergence · information geometry

MSC 2020: 65K05 · 49J52 · 62D10 · 62B11

1. INTRODUCTION

The EM algorithm assures monotone decrease of the incomplete data negative log-likelihood [30], but convergence of the iterates may fail in various ways [78]. Without coercivity iterates may escape to infinity while values converge. Even when iterates stay bounded, they may still fail to converge, cycle [76], or generate a continuum of accumulation points. Convergence analysis is further complicated when iterates tend to the boundary of the natural parameter domain, where the likelihood is typically not well-behaved.

Even when convergent, instead of approaching a global minimum of the incomplete data negative log-likelihood, EM iterates may go to local minima, saddle points, or even to local maxima [53]. This is a well-known phenomenon in non-convex optimization, where in practice it is usually acceptable to find good local extrema.

In this work we allow the maximum likelihood problem to include parameter constraints. This yields a convenient way to model curved exponential families, but constraints may also convey prior knowledge about the unknown parameter, allow to implement restricted maximum likelihood [46, p. 191], [14, 70, 63, 71], deal with truncated families [47], keep iterates away from the boundary of the natural parameter domain, or simply force boundedness, see e.g. [48, 36, 56, 41]. While practical, constraints further complicate convergence analysis of the EM algorithm.

There is a similarity between the EM algorithm and the proximal point method (PPM), which had been observed in the contributions [75, 21, 22, 23, 25]. The quadratic penalty term in PPM is replaced by a regularizer based on the Kullback-Leibler information distance. Since convergence of PPM without convexity has been investigated [4, 5, 64, 40, 72, 68], some of these techniques, so the intention, may carry over to Kullback-Leibler regularizers. Another compelling reason to investigate this link is the fact that PPM can be seen as a special instance of EM when the latter includes constraints.

Presently we take a fresh look at this line, adding as a new element definability of the incomplete data log-likelihood in the sense of o-minimal structure theory [32, 31, 33, 77], a hypothesis always met in practice. Our investigation reveals that the Kullback-Leibler distance has only a partial regularizing effect concentrated on a linear subspace, whose dimension depends on the rank of the conditional Fisher information matrix of missing data, given the observed datum. In consequence, even under definability, only convergence of the projection of iterates on this subspace can be derived. Further elements are needed to assure convergence

*Institut de Mathématiques, Université de Toulouse, France.

of the full EM sequence. The positive aspect is that this gives us clues as to why instances of EM fail to converge.

Additional insight into convergence of the full EM sequence $\theta^{(k)}$ is gained by the point of view of [59], where the EM algorithm had been interpreted as a method of alternating Bregman projections between data and model sets. EM iterates $\theta^{(k)}$ generated in the M step arise in tandem with iterates $\vartheta^{(k)}$ generated in the E step, and the algorithm alternates between these. Since the $\vartheta^{(k)}$ converge under fairly general hypotheses, also involving definability, this adds new occasions to deduce convergence of the $\theta^{(k)}$.

While the EM algorithm is in general expected to converge linearly, cases of sub-linear rates have been reported, see [53, p.34]. As part of our analysis we obtain a sublinear worst case rate $\|\theta^{(k)} - \theta^*\| = O(k^{-\rho})$ for some $0 < \rho < \infty$, which can be certified under quite natural assumptions. The case where linear rates occur is also precisely delimited.

Recent approaches to convergence of the EM algorithm are [44, 42], where the authors use mirror descent to apply results from non-linear optimization, and [19], where the Polyak-Łojasiewicz inequality allows the authors to derive complexity results in a non-parametric setting. As the Polyak-Łojasiewicz inequality is an instance of the Kurdyka-Łojasiewicz inequality, it is included in our present analysis, where it gives the case of linear convergence.

The structure of the paper is as follows. Section 2 recalls facts from optimization and definability theory. Section 3 recalls the set-up of the EM algorithm, including the case of curved exponential families. Section 4 concerns Kullback-Leibler information, its role as a regularizer, and its link with Fisher information of missing data. Parameter dimension reduction for the conditional family follows in Section 5, revealing the partial character of the Kullback-Leibler regularizer. Convergence under partial regularization is proved in Section 6, a worst case rate given in Section 6.2. Section 7 applies this to the EM algorithm, followed by cases where partial convergence can be upgraded to convergence of the full sequence $\theta^{(k)}$. Alternating Bregman projections come into play in Section 8. Extensions beyond the exponential family are discussed in Section 9. Examples are given in Section 10.

NOTATION

For a function $\Psi(x, y)$ we write $\nabla_x \Psi$ or $\nabla_1 \Psi$ for the derivative with respect to x , $\nabla_y \Psi$, $\nabla_2 \Psi$ for the derivative with respect to y . Second derivatives twice with respect to x are $\nabla_{11}^2 \Psi(x, y)$ or $\nabla_{xx}^2 \Psi(x, y)$, and similarly $\nabla_{22}^2 \Psi(x, y)$ or $\nabla_{yy}^2 \Psi(x, y)$ and $\nabla_{xy}^2 \Psi(x, y)$ for a mixed second derivative. The subdifferential is understood in the sense of [55, 69] and denoted $\partial \psi(x)$. For a function $\Psi(x, y)$ we have $\partial_1 \Psi(x, y) = \partial \Psi(\cdot, y)(x)$, and $\partial_2 \Psi(x, y) = \partial \Psi(x, \cdot)(y)$. The euclidean scalar product and norm on \mathbb{R}^n are $x \cdot y$ and $\|x\|$. Euclidean balls are $B(x, \delta)$, and the euclidean δ -neighborhood of a set K is $N(K, \delta)$.

2. PREPARATION

In this section we recall facts from optimization and definability theory. We follow the convention that iterates in general optimization algorithms are denoted x_k , while when specifying to the EM algorithm iterates, being parameters to be estimated, will be termed $\theta^{(k)}$.

2.1. Proximal point algorithm. The classical proximal point method for a proper lower semi-continuous function $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$ generates iterates x_k via

$$(1) \quad x_{k+1} \in \operatorname{argmin}_{x \in \mathbb{R}^n} f(x) + \frac{1}{2\lambda_k} \|x - x_k\|^2,$$

where $\lambda_k > 0$ and where $\|\cdot\|$ is the euclidean norm [51, 52]. For convex f it is known [67, 35] that when f has a minimum, the sequence x_k converges to $x^* \in \operatorname{argmin} f$ iff $\sum_{j=1}^k \lambda_j \rightarrow \infty$ ($k \rightarrow \infty$), and the speed is even super-linear when $\lambda_k \rightarrow \infty$, [69, 35], [49, Thm. 2.1]. In the non-convex case convergence is much harder to obtain, but partial results are known, cf. [4, 5, 64, 40, 72, 68].

2.2. Bregman proximal method. It has been proposed to replace the quadratic penalty $\frac{1}{2\lambda_k}\|x - x_k\|^2$ by a Bregman regularizer, $\lambda_k^{-1}D(x, x_k)$, which leads to the scheme

$$(2) \quad x_{k+1} \in \operatorname{argmin}_{x \in \mathbb{R}^n} f(x) + \lambda_k^{-1}D(x, x_k),$$

see [20, 10]. Here, given a function ψ of Legendre type [9],[66], the Bregman distance associated with ψ is

$$(3) \quad D(x, y) = \begin{cases} \psi(x) - \psi(y) - \nabla\psi(y) \cdot (x - y) & \text{if } y \in \operatorname{int}(\operatorname{dom} \psi) =: G \\ +\infty & \text{otherwise} \end{cases}$$

with the proximal point method corresponding to the special case $\psi(x) = \frac{1}{2}\|x\|^2$.

In the context of exponential families, the ψ arise as cumulant generating functions, or log-normalizers. Legendre-ness of ψ is called steepness [18], and $\nabla^2\psi \succ 0$ corresponds to minimality of the family. In the following we assume throughout that ψ is of class C^2 .

2.3. More general regularizers. Expanding on (1) and (2), we next envisage schemes of the form

$$(4) \quad x_{k+1} \in \operatorname{argmin}_{x \in \mathbb{R}^n} f(x) + \lambda_k^{-1}\Psi(x, x_k),$$

with even more general regularizers $\Psi(x^+, x)$, allowing $D(x^+, x)$ as special cases. Like $D(x^+, x)$, $\Psi(x^+, x)$ will only be partially defined. Taking Bregman regularizers as paragon, we propose the following set-up:

- (i) There exists an open set G with $G \times G \subset \operatorname{dom}\nabla_{11}^2\Psi = \operatorname{dom}\nabla_1\Psi \subset \operatorname{dom}\Psi \subset \bar{G} \times G$, and $\Psi, \nabla_1\Psi, \nabla_{11}^2\Psi$ are jointly continuous on their domains.
- (ii) $\Psi \geq 0$, and $\Psi(x, x) = 0$ for $x \in G$.

We call Ψ *separating* if it satisfies the stronger property

- (ii') $\Psi \geq 0$ and $\Psi(x^+, x) = 0$ iff $x^+ = x$.

We say that Ψ has a *lower norm bound* at $x \in G$ if there exist $\delta > 0$ and $m > 0$ such that

- (iii) $\Psi(y, z) \geq m\|y - z\|^2$ for all $y, z \in B(x, \delta)$.

We say that Ψ has *pointwise lower norm bounds* on a set $K \subset G$ if (iii) holds for every $x \in K$. While δ, m depend on x , a standard compactness argument (see Lemma 7) shows that we can get the same δ, m for all $x \in K$ when $K \subset G$ is compact. Finally, in view of (i) we may without loss of generality assume that $\operatorname{dom}f \subset \bar{G}$.

2.4. Partial regularizers. Suppose $\Psi_V(v^+, v)$ satisfies (i), (ii), but only for elements v^+, v of a linear subspace V . Then $\Psi(x^+, x) = \Psi_V(Px^+, Px)$, with P the orthogonal projection onto V , gives a regularizer on \mathbb{R}^n . We call such Ψ partial regularizers, because their effect is limited to V -components Px of iterates x , leaving V^\perp -components unaffected. In x -space a partial regularizer satisfies (i) and (ii), while (ii') or (iii) could at best be satisfied in V .

2.5. Kurdyka-Łojasiewicz inequality. The following definition is crucial for our approach.

Definition 1. (Kurdyka-Łojasiewicz inequality). A lower semi-continuous function $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ has the KL-property at $\bar{x} \in \operatorname{dom}(\partial f)$ if there exist $\gamma, \eta > 0$, a neighborhood U of \bar{x} , and a continuous concave function $\phi : [0, \eta] \rightarrow \mathbb{R}_+$, called *de-singularizing function*, such that

- i. $\phi(0) = 0$,
- ii. ϕ is of class C^1 on $(0, \eta)$,
- iii. $\phi'(s) > 0$ for $s \in (0, \eta)$,
- iv. For all $x \in U \cap \{x : f(\bar{x}) < f(x) < f(\bar{x}) + \eta\}$ the KL-inequality

$$(5) \quad \phi'(f(x) - f(\bar{x}))\operatorname{dist}(0, \partial f(x)) \geq \gamma$$

is satisfied, where ∂f is the subdifferential of [69].

Remark 1. We say that f satisfies the Łojasiewicz inequality when the de-singularizing function is $\phi'(s) = s^{-\theta}$ for some $\theta \in [\frac{1}{2}, 1)$, which means $\phi(s) = \frac{s^{1-\theta}}{1-\theta}$. The case $\theta = \frac{1}{2}$ is sometimes singled out under the name Polyak-Łojasiewicz inequality.

Remark 2. When K is a compact set on which f has constant value $f(\bar{x}) = f^*$ for all $\bar{x} \in K$, then (5) holds uniformly on a neighborhood U of K .

Remark 3. For convergence via the KL-property see e.g. [1, 5, 6, 7, 16, 57, 58]. It is well-known that definability in an o-minimal structure [31, 32, 33], for short *definability*, implies the KL-inequality. See [45], and for non-smooth f , [17, Thm. 11], where it is shown that ϕ may be chosen concave. We use the o-minimal structure \mathbb{R}_{an} of globally sub-analytic sets, but also the larger $\mathbb{R}_{an,exp}$, allowing exponential and logarithm, see [31, 15, 27, 74, 77, 54].

3. EM ALGORITHM

We consider a family of probability measures \mathbb{P}_θ with densities $p(x, \theta)$ with regard to a σ -finite base measure μ on the complete data space X , $d\mathbb{P}_\theta(x) = p(x, \theta)d\mu(x)$, where $\theta \in \Theta$ is the unknown parameter we want to estimate by maximum likelihood. However, it is not $x \in X$ which is observed, but a random variable $y = h(x)$ in the incomplete data space Y , where $h : X \rightarrow Y$ is measurable and typically non-invertible. The density of observed data y with regard to the marginal $\nu = \mu \circ h^{-1}$ is therefore

$$(6) \quad q(y, \theta) = \int_{h^{-1}(y)} p(x, \theta) d\mu_y(x),$$

where the family $(\mu_y)_{y \in Y}$ is a disintegration of the measure μ with regard to the marginal $\nu = \mu \circ h^{-1}$ on Y , each μ_y concentrated on $h^{-1}(y) \subset X$. Here disintegration means

$$(7) \quad \int_X f(x) d\mu(x) = \int_Y \left[\int_{h^{-1}(y)} f(x) d\mu_y(x) \right] d\nu(y)$$

for μ -integrable f . Substituting $f = \chi_{h^{-1}(B)} p(\cdot, \theta)$ leads to the relation

$$(\mathbb{P}_\theta \circ h^{-1})(B) = \mathbb{P}_\theta(h^{-1}(B)) = \int_B \left[\int_{h^{-1}(y)} p(x, \theta) d\mu_y(x) \right] d\nu(y)$$

justifying (6). This allows us to define, for every $y \in Y$, the conditional density

$$(8) \quad k(x|y, \theta) = \frac{p(x, \theta)}{q(y, \theta)}, \quad x \in h^{-1}(y),$$

with regard to the measure μ_y . For hypotheses needed to establish the existence of a disintegration see for instance [37].

Given an available sample y , maximum likelihood in incomplete data space Y is the optimization program

$$(9) \quad \hat{\theta} \in \operatorname{argmin}_{\theta \in M} -\log q(y, \theta),$$

where $M \subset \Theta$ is a set of model parameters admitted for optimization. At this stage the rationale of the EM algorithm assumes that minimization (9) is cumbersome, and that it would be preferable algorithmically to perform maximum likelihood estimation in complete data space

$$(10) \quad \tilde{\theta} \in \operatorname{argmin}_{\theta \in M} -\log p(x, \theta).$$

Since no sample x is available, (10) cannot be performed directly, and recourse is taken to the following iterative procedure. Given a current guess $\theta^{(k)} \in M$ of the unknown parameter, one computes *for every* $\theta \in M$, the conditional expectation of $\log p(x, \theta)$, given y and $\theta^{(k)}$:

$$Q(\theta, \theta^{(k)}) := \mathbb{E}_{\theta^{(k)}}(\log p(x, \theta)|y),$$

which (for fixed $y, \theta^{(k)}$) gives a function of θ . This is called the E step, based on the formula

$$\mathbb{E}_\theta(\phi(x, \theta')|y) = \int_{h^{-1}(y)} \phi(x, \theta') k(x|y, \theta) d\mu_y(x),$$

with $k(x|y, \theta)$ given by (8). Once this is obtained, one performs the M step

$$\theta^{(k+1)} \in \operatorname{argmin}_{\theta \in M} -Q(\theta, \theta^{(k)}),$$

which gives the new model parameter estimate $\theta^{(k+1)} \in M$. As is well-known, the EM algorithm reduces the negative log-likelihood $-\log q(y, \theta^{(k)})$ at each step, and this is not altered by optimizing over $\theta \in M$, nor when only a local minimum is computed. However, as mentioned before, monotone decrease in function value does *not* assure convergence of the iterates $\theta^{(k)}$, and it is convergence of the iterates we presently scrutinize.

Algorithm EM algorithm

- ▷ **Step 1 (E step).** Given current model parameter estimate $\theta^{(k)} \in M$, make the function $\theta \mapsto Q(\theta, \theta^{(k)}) = \mathbb{E}_{\theta^{(k)}}(\log p(x, \theta)|y)$ on M available for optimization.
- ▷ **Step 2 (M step).** Compute $\theta^{(k+1)} \in \operatorname{argmin}_{\theta \in M} -Q(\theta, \theta^{(k)})$. Back to step 1.
-

Remark 4. A special case of constraints are *curved families* $M = \{\theta \in \Theta : \theta = \theta(u), u \in U\}$, with $\theta(u)$ a re-parametrization of θ , but our approach allows more general sets. A typical instance of M is given in Example 3.

3.1. Properties of exponential families. The densities $p(x, \theta)$ on X form a n -dimensional exponential family with regard to the base measure μ if they are of the form

$$(11) \quad p(x, \theta) = e^{\theta \cdot T(x) - \psi(\theta)},$$

where $T(x)$ is the sufficient statistic, $\theta \in \Theta = \{\theta \in \mathbb{R}^n : p(\cdot, \theta) \in L^1(X, d\mu)\}$ the natural parameter, and $\psi(\theta)$ the log-normalizer defined on Θ , satisfying

$$(12) \quad \psi(\theta) = \log \int_X e^{\theta \cdot T(x)} d\mu(x).$$

The natural parameter space can also be written as $\Theta = \{\theta \in \mathbb{R}^n : \int_X e^{\theta \cdot T(x)} d\mu(x) < \infty\}$ and is a convex subset of \mathbb{R}^n . There is no loss in generality in assuming that Θ is of full dimension n , as otherwise a parameter reduction leading to an equivalent representation (11) with lower dimension $m < n$ can be performed:

Lemma 1. *Suppose the natural parameter space Θ is contained in an affine subspace of dimension $m < n$. Then there exists an equivalent representation of \mathbb{P}_θ as a m -dimensional exponential family $d\mathbb{P}_\theta(x) = p'(x, \theta') d\mu'(x) = e^{\theta' \cdot T'(x) - \psi'(\theta')} d\mu'(x)$, $\theta' \in \Theta' \subset \mathbb{R}^m$, where now $\dim(\Theta') = m$. If in (11) the statistic $T(x)$ is affinely independent on X , then so is $T'(x)$.*

Proof: Without loss of generality write the parameter as $\theta = (\theta_1, \theta_2)$ with $\theta_2 = A\theta_1 + a$ for a matrix A of size $(n - m) \times m$ of rank $n - m$. Then $d\mathbb{P}_\theta(x) = p(x, \theta) d\mu(x) = e^{\theta_1 \cdot T_1(x) + \theta_2 \cdot T_2(x) - \psi(\theta)} d\mu(x) = e^{\theta_1 \cdot (T_1(x) + A^T T_2(x)) - \psi(\theta_1, A\theta_1 + a)} e^{a \cdot T_2(x)} d\mu(x) = e^{\theta_1 \cdot T'(x) - \psi'(\theta_1)} d\mu'(x)$, with $T'(x) = T_1(x) + A^T T_2(x)$, $\psi'(\theta_1) = \psi(\theta_1, A\theta_1 + a)$, and $d\mu'(x) = e^{a \cdot T_2(x)} d\mu(x)$, and where the parameter space $\Theta' = \{\theta_1 : (\theta_1, A\theta_1 + a) \in \Theta\}$ is now of full dimension m . Since m is smallest possible, there could no longer be any affine dependence among the $\theta_1 \in \Theta'$. Note also that $\mu \ll \mu'$ and $\mu' \ll \mu$ gives equivalence of the two representations.

To conclude, suppose $a \cdot T(x)$ constant a.e. implies $a = 0$. Then if $a' \cdot T'(x) = c$ for almost all $x \in X$, we have $(a', A^T a') \cdot (T_1(x), T_2(x)) = c$, hence $(a', A^T a') = (0, 0)$, which gives $a' = 0$, so that T' is also affinely independent. \square

Assuming therefore that Θ has already full dimension n in (11), we denote the interior of Θ by G . The function ψ is also known as the cumulant generating function, because it satisfies

$$\mathbb{E}_\theta[T(x)] = \nabla\psi(\theta), \quad \mathbb{V}_\theta[T(x)] = \nabla^2\psi(\theta),$$

and similar relations for higher moments; cf. [18].

Definition 2. A family (11) is called *minimal* if the functions $T_i(\cdot)$ are affinely independent, i.e., if there exists no $a \neq 0$ such that $a \cdot T(x) = \text{const}$ for μ -a.a. $x \in X$.

Lemma 2. Under minimality we have $\nabla^2\psi(\theta) \succ 0$ for every $\theta \in G = \text{int}(\Theta)$. Moreover, the mapping $\theta \mapsto p(\cdot, \theta)$ is injective on G .

Proof: 1) From $\nabla^2\psi(\theta)a = 0$ follows $0 = a \cdot \nabla^2\psi(\theta)a = a \cdot \mathbb{V}_\theta[T(x)]a = \mathbb{V}_\theta[a \cdot T(x)] = \mathbb{E}_\theta|a \cdot T(x) - \mathbb{E}_\theta(a \cdot T(x))|^2$, hence $a \cdot T(x) = \mathbb{E}_\theta[a \cdot T(x)] = a \cdot \nabla\psi(\theta) = \text{const}$ μ -a.e., forcing $a = 0$.

2) Let $p(\cdot, \theta) = p(\cdot, \theta')$ μ -a.e., then $\theta \cdot T(x) - \psi(\theta) = \theta' \cdot T(x) - \psi(\theta')$ a.e., hence $(\theta - \theta') \cdot T(x) + \psi(\theta') - \psi(\theta) = 0$ a.e., so that the vector $a = \theta - \theta'$ renders $a \cdot T(x) = \psi(\theta) - \psi(\theta')$ constant a.e., forcing $\theta = \theta'$. \square

An exponential family is called *steep* if the log-normalizer ψ is of Legendre type [69, 9, 18]. The family is called *regular* if Θ is an open set, i.e., $\Theta = G$. In that case the family is automatically steep, but the steep class is larger [18].

3.2. EM algorithm for the exponential family. The EM algorithm is sometimes characterized as going back and forth between *completing the data* in the E step, and *maximum likelihood in complete data space* in the M step. This is not true in general, cf. [34], but holds when $p(x, \theta)$ belongs to an exponential family (11). Namely, in that case, $\log p(x, \theta) = \theta \cdot T(x) - \psi(\theta)$, hence

$$\mathbb{E}_{\theta^{(k)}}[\log p(x, \theta)|y] = \theta \cdot \mathbb{E}_{\theta^{(k)}}[T(x)|y] - \psi(\theta),$$

and the first term on the right *selects* a complete data statistic $t^{k+1} = T(x_{k+1})$, a fact which one expresses by saying that the E step consists in *completing the data*. The M step is unchanged, but due to the substitution of $T(x_{k+1})$, leads to $Q(\theta, \theta^{(k)}) = \log p(x_{k+1}, \theta) = \theta \cdot T(x_{k+1}) - \psi(\theta)$, and may therefore rightfully be referred to as *maximum likelihood in complete data space*. Altogether, for exponential families the EM algorithm has the form

Algorithm EM algorithm for the exponential family

- ▷ **Step 1 (E step).** Given current model parameter estimate $\theta^{(k)} \in M$, complete the data by computing $T(x_{k+1}) = \mathbb{E}_{\theta^{(k)}}[T(x)|y]$.
- ▷ **Step 2 (M step).** Compute $\theta^{(k+1)} \in \arg\min_{\theta \in M} \psi(\theta) - \theta \cdot T(x_{k+1})$. Back to step 1.
-

Remark 5. When the sufficient statistic is affine, $T(x) = Ax + b$, the E step may even be based on computing $x_{k+1} = \mathbb{E}_{\theta^{(k)}}(x|y)$, as then $\mathbb{E}_{\theta^{(k)}}[T(x)|y] = \mathbb{E}_{\theta^{(k)}}(Ax + b|y) = A\mathbb{E}_{\theta^{(k)}}(x|y) + b$, making the expression *completing the data* even more suggestive.

We conclude this section by remarking that when complete data are from an exponential family $p(x, \theta)$ on X as in (11), then the conditional densities $k(x|y, \theta)$ also constitute, for given $y \in Y$, an exponential family on the sample space $h^{-1}(y)$ with regard to the base measure μ_y arising from the disintegration of μ . This can be seen from

$$(13) \quad k(x|y, \theta) = \frac{p(x, \theta)}{q(y, \theta)} = \frac{e^{\theta \cdot T(x) - \psi(\theta)}}{\int_{h^{-1}(y)} e^{\theta \cdot T(x') - \psi(\theta)} d\mu_y(x')} =: e^{\theta \cdot T(x) - \psi_y(\theta)},$$

obtained on putting

$$(14) \quad \psi_y(\theta) = \log \int_{h^{-1}(y)} e^{\theta \cdot T(x)} d\mu_y(x),$$

which parallels (12), and which we can write as $d\mathbb{P}_\theta^{x|y} = k(x|y, \theta) d\mu_y(x)$. For an exponential family the measures $k(\cdot|y, \theta) d\mu_y$ are mutually equivalent, i.e., $k(\cdot|y, \theta) d\mu_y \ll k(\cdot|y, \theta') d\mu_y$ for all $\theta, \theta' \in \Theta$. In a general setting, this may for practical considerations be added as a hypothesis (cf. [21, Sect III B]).

Remark 6. Even when complete data $p(x, \theta)$ are from an exponential family, *this need not be true for incomplete data $q(y, \theta)$* . The exponential structure may be lost just because data are missing. But there are also cases where specific families $q(y, \theta)$ *not* of exponential type are deliberately arranged as *missing data from an exponential family*. This terminology goes back to Sundberg [73], who gives a variety of examples $q(y, \theta)$, including finite mixtures of exponential families, censored data, convolutions, folded distributions, the negative binomial distribution, and much else. Presently we extend this to include parameter constraints $\theta \in M$, so that complete data $p(x, \theta)$ from which $q(y, \theta)$ are derived may e.g. be curved.

Concerning the well-posedness of the EM sequence, we have to bear in mind that M has to be a closed set, because the objective f has to be lower semi-continuous. This may be in conflict with the fact that Θ is in general not closed. We may therefore only assume $M \subset \bar{\Theta}$, so that $M \cap \partial\Theta$ may be non-empty, and points in this set cause trouble.

4. KULLBACK-LEIBLER INFORMATION MEASURE

The Kullback-Leibler information distance on X is defined as

$$\mathcal{K}(q||p) = \mathbb{E}_q \left(\log \frac{q}{p} \right) = \int_X q(x) \log \frac{q(x)}{p(x)} d\mu(x),$$

and in the parameter-dependent case we use the notation

$$K(\theta||\theta^+) = \mathcal{K}(p(\cdot, \theta)||p(\cdot, \theta^+)).$$

When restricted to $h^{-1}(y)$ the KL-distance takes the form

$$(15) \quad K_y(\theta||\theta^+) = \mathbb{E}_\theta \left(\log \frac{k(\cdot|y, \theta)}{k(\cdot|y, \theta^+)} \middle| y \right) = \int_{h^{-1}(y)} k(x|y, \theta) \log \frac{k(x|y, \theta)}{k(x|y, \theta^+)} d\mu_y(x),$$

where μ_y arises from the disintegration of μ , and where the value is finite due to the hypothesis $k(\cdot|y, \theta) d\mu_y \ll k(\cdot|y, \theta') d\mu_y$ for all $\theta, \theta' \in \Theta$. The following is now a crucial observation.

Proposition 1. (See [21, Prop. 1]). *The EM algorithm is a realization of the general scheme (4) with $f(\theta) = -\log q(y, \theta) + i_M(\theta)$, $\lambda_k = 1$, and $\Psi(\theta, \theta^{(k)}) = K_y(\theta^{(k)}||\theta)$ given by (15).*

Proof: Re-arranging (8) and integrating, we have

$$\log q(y, \theta) = \mathbb{E}_{\theta^{(k)}}(\log p(x, \theta)|y) - \mathbb{E}_{\theta^{(k)}}(\log k(x|y, \theta)|y)$$

for arbitrary k . Hence the M step in the EM algorithm is

$$\theta^{(k+1)} \in \underset{\theta \in \mathbb{R}^n}{\operatorname{argmin}} -\log q(y, \theta) + i_M(\theta) - \mathbb{E}_{\theta^{(k)}}(\log k(x|y, \theta)|y).$$

Adding the constant term $\mathbb{E}_{\theta^{(k)}}(\log k(x|y, \theta^{(k)})|y)$ to the objective does not change the optimization program, hence we have

$$(16) \quad \theta^{(k+1)} \in \underset{\theta \in \mathbb{R}^n}{\operatorname{argmin}} -\log q(y, \theta) + i_M(\theta) - \mathbb{E}_{\theta^{(k)}} \left(\log \frac{k(x|y, \theta)}{k(x|y, \theta^{(k)})} \middle| y \right),$$

and the last term equals $K_y(\theta^{(k)}||\theta)$. □

Remark 7. 1) This explains why the EM algorithm decreases the negative log-likelihood $-\log q(y, \theta)$, as this is a general property of the scheme (4), see also Theorem 1.

2) Kullback-Leibler divergence is separating in function space, i.e., $\mathcal{K}(q||p) = 0$ implies $q = p$ a.e. Hence $K_y(\theta||\theta^+) = 0$ implies $k(\cdot|y, \theta) = k(\cdot|y, \theta^+)$ μ_y -a.e. However, the latter does not always give $\theta = \theta^+$, because the family $k(\cdot|y, \theta)$ is not necessarily minimal on $h^{-1}(y)$.

4.1. Interpretation for the exponential family. We now specify the Kullback-Leibler divergence to the case of an exponential family.

Proposition 2. *The Kullback-Leibler divergence of two distributions $p(x, \theta)$ and $p(x, \theta')$ belonging to the same exponential family is $K(\theta||\theta^+) = D(\theta^+, \theta)$, where D is the Bregman divergence induced by the log-normalizer ψ .*

Proof: From (11), and since $\int_X p(x, \theta) d\mu(x) = 1$, we have

$$\psi(\theta) = \log \int_X e^{\theta \cdot T(x)} d\mu(x).$$

Differentiation with respect to θ gives

$$\nabla \psi(\theta) = \int_X T(x) e^{\theta \cdot T(x)} d\mu(x) / \int_X e^{\theta \cdot T(x)} d\mu(x).$$

Now $e^{\psi(\theta)} = \int_X e^{\theta \cdot T(x)} d\mu(x)$, hence $\nabla \psi(\theta) = \int_X T(x) e^{\theta \cdot T(x) - \psi(\theta)} d\mu(x) = \int_X T(x) p(x, \theta) d\mu(x) = E_\theta[T(x)]$, the expectation of the random variable $T(x)$ with respect to the distribution $d\mathbb{P}_\theta = p(\cdot, \theta) d\mu$ (see also [53, (1.57)]). Then

$$\begin{aligned} K(\theta||\theta^+) &= \int_X p(x, \theta) \log \frac{p(x, \theta)}{p(x, \theta^+)} d\mu(x) \\ &= \int_X p(x, \theta) [\psi(\theta^+) - \psi(\theta) + (\theta - \theta^+) \cdot T(x)] d\mu(x) \\ (17) \quad &= \int_X p(x, \theta) [D(\theta^+, \theta) + (\theta^+ - \theta) \cdot \nabla \psi(\theta) + (\theta - \theta^+) \cdot T(x)] d\mu(x) \\ &= D(\theta^+, \theta) + \int_X p(x, \theta) [(\theta^+ - \theta) \cdot (\nabla \psi(\theta) - T(x))] d\mu(x) \\ &= D(\theta^+, \theta) + (\theta^+ - \theta) \cdot (\nabla \psi(\theta) - E_\theta[T(x)]) \\ &= D(\theta^+, \theta). \end{aligned}$$

This proves the claim. \square

Bregman distances or divergences are usually considered for functions ψ of Legendre type [9, 69]. As already mentioned, for log-normalizers this is called steepness [18, Def. 3.2]. Most exponential families in practice are regular, i.e., Θ is open, in which case steepness follows automatically.

Lemma 3. *Suppose the exponential family $p(x, \theta)$ is minimal. Then the Bregman divergence induced by the log-normalizer ψ is separating, i.e., $D(\theta^+, \theta) = 0$ implies $\theta^+ = \theta$.*

Proof: From $D(\theta^+, \theta) = 0$ we get $\psi(\theta^+) - \psi(\theta) - \nabla \psi(\theta) \cdot (\theta^+ - \theta) = 0$. Taylor expansion at θ gives $\psi(\theta^+) = \psi(\theta) + \nabla \psi(\theta) \cdot (\theta^+ - \theta) + \frac{1}{2}(\theta^+ - \theta) \cdot \nabla^2 \psi(\bar{\theta})(\theta^+ - \theta)$ for some $\bar{\theta}$ on the open segment joining θ^+ and θ , and depending on θ, θ^+ . Hence $(\theta^+ - \theta) \cdot \nabla^2 \psi(\bar{\theta})(\theta^+ - \theta) = 0$. But minimality gives $\nabla^2 \psi(\bar{\theta}) \succ 0$ by Lemma 2, hence $\theta^+ = \theta$. \square

Lemma 4. *Suppose the exponential family $p(x, \theta)$ is minimal. Then $\nabla \psi$ is injective on G . For $\eta = \nabla \psi(\theta) \in G^* = \text{int}(\text{dom } \psi^*)$ we have $\nabla \psi^*(\eta) = \theta$. In particular, if the family is steep, then $\nabla \psi^*$ is the inverse of $\nabla \psi$, with $G = \text{int}(\text{dom } \psi)$ mapped 1-1 onto $G^* = \text{int}(\text{dom } \psi^*)$.*

Proof: Let $\nabla\psi(\theta) = \nabla\psi(\theta^+)$. For a test vector h , Taylor expansion of $\theta \mapsto h \cdot \nabla\psi(\theta)$ at θ^+ gives $h \cdot \nabla\psi(\theta) = h \cdot \nabla\psi(\theta^+) + h \cdot \nabla^2\psi(\bar{\theta})(\theta - \theta^+)$ for some $\bar{\theta} = \bar{\theta}(\theta, \theta^+, h)$ on the open segment joining θ and θ^+ and depending on θ, θ^+, h . Matching this with the first equation above shows $h \cdot \nabla^2\psi(\bar{\theta})(\theta - \theta^+) = 0$. Taking as test vector $h = \theta - \theta^+$ gives $(\theta - \theta^+) \cdot \nabla^2\psi(\bar{\theta})(\theta - \theta^+) = 0$ for $\bar{\theta} = \bar{\theta}(\theta, \theta^+, \theta - \theta^+)$, and since $\nabla^2\psi(\bar{\theta}(\theta, \theta^+, \theta - \theta^+)) \succ 0$, we have $\theta = \theta^+$.

For the second part, recall that $\partial\psi, \partial\psi^*$ are inverses of each other in the sense $\eta \in \partial\psi(\theta)$ iff $\theta \in \partial\psi^*(\eta)$; cf. [69, Cor. 23.5.1]. By strict convexity of ψ its conjugate ψ^* is differentiable on $G^* = \text{int dom } \psi^*$. Hence if $\eta = \nabla\psi(\theta) \in G^*$, then $\nabla\psi^*(\eta) = \theta$. Since we may have $\nabla\psi(G) \not\subset G^*$, all we know about $\eta = \nabla\psi(\theta) \in \partial G^*$ is $\theta \in \partial\psi^*(\eta)$. When $p(x, \theta)$ is steep, then $\partial\psi(\theta) = \emptyset$ for $\theta \in \partial\Theta$, and then $\nabla\psi$ maps G 1-1 into G^* with inverse $(\nabla\psi)^{-1} = \nabla\psi^*$. \square

Remark 8. In general one has $G = \text{int}(\text{dom } \psi) \subset \text{dom}(\nabla\psi) \subset \text{dom } \psi = \Theta$, and $G^* = \text{int}(\text{dom } \psi^*) \subset \text{dom}(\nabla\psi^*) \subset \text{dom } \psi^*$; cf. [66, Thm. 23.4].

4.2. Fisher information of missing data. The regularizer in (16) has a statistical interpretation. It is well-known that $K_y(\theta||\theta) = 0$, and also $K_y \geq 0$, hence for fixed θ , the global minimum 0 of $\theta^+ \mapsto K_y(\theta||\theta^+)$ is attained in particular at $\theta^+ = \theta$. Then clearly $\nabla_2 K_y(\theta||\theta) = 0$ and $\nabla_{22}^2 K_y(\theta||\theta) \succeq 0$ from the necessary optimality conditions. We investigate whether we may expect the stronger sufficient optimality condition $\nabla_{22}^2 K_y(\theta||\theta) \succ 0$. Going back to the definition, we have

$$K_y(\theta||\theta^+) = \int_{h^{-1}(y)} k(x|y, \theta) \log \frac{k(x|y, \theta)}{k(x|y, \theta^+)} d\mu_y(x) = \mathbb{E}_\theta \left[\log \frac{k(\cdot|y, \theta)}{k(\cdot|y, \theta^+)} \middle| y \right].$$

Differentiating twice with respect to θ^+ (cf. [8, Thm. 5.8, sect. 7.1]) gives

$$\nabla_{22}^2 K_y(\theta||\theta^+) = \mathbb{E}_\theta [-\nabla_{\theta+\theta^+}^2 \log k(\cdot|y, \theta^+)|y],$$

hence we obtain

$$\nabla_{22}^2 K_y(\theta||\theta^+) \big|_{\theta^+=\theta} = \mathbb{E}_\theta [-\nabla_{\theta\theta}^2 \log k(\cdot|y, \theta)|y] =: \mathcal{I}_m(\theta, y),$$

which is recognized as the conditional expected Fisher information matrix of missing data, given y ; cf. [53, 3.52]. According to the missing information principle [61], $\mathcal{I}_m(\theta, y)$ gives the expected loss of information between complete and incomplete data. Differentiating the identity $\mathbb{E}_\theta [\nabla_\theta \log k(\cdot|y, \theta)|y] = 0$ with respect to θ (see again [8, Thm. 5.8, sect. 7.1]) gives the alternative formula

$$(18) \quad \mathcal{I}_m(\theta, y) = \mathbb{E}_\theta [\nabla_\theta k(\cdot|y, \theta) \nabla_\theta k(\cdot|y, \theta)^T |y] \succeq 0.$$

A bit more can be said in the case of an exponential family.

Lemma 5. *Suppose $k(x|y, \theta)$ is an exponential family. Then $\nabla_{22}^2 K_y(\theta||\theta) = \nabla^2 \psi_y(\theta)$. In addition, if the family is minimal with regard to the sample space $h^{-1}(y)$, then $\nabla^2 \psi_y(\theta) \succ 0$.*

Proof: From (17) we get $K_y(\theta||\theta^+) = D_y(\theta^+, \theta)$ for the Bregman distance D_y induced by ψ_y , and then $\nabla_{\theta+\theta^+}^2 K_y(\theta||\theta^+) = \nabla^2 \psi_y(\theta^+)$. The last part follows with Lemma 2. \square

5. DIMENSION REDUCTION FOR THE CONDITIONAL FAMILY

We consider constrained maximum likelihood with incomplete data from a n -dimensional exponential family, i.e., $\text{im}(T) \subset \mathbb{R}^n$, where $\dim(\Theta) = n$. In view of Lemma 1, we also assume that the complete data family $p(x, \theta)$ is minimal. However, this does not mean that the conditional n -dimensional exponential family $k(x|y, \theta)$ is also minimal on $h^{-1}(y)$. In fact, the missing data case (Example 1) shows that we should rather expect the opposite. This calls for a dimension reduction argument.

Proposition 3. *The n -dimensional conditional exponential family $d\mathbb{P}_\theta^{x|y} = k(\cdot|y, \theta)d\mu_y$ has an equivalent minimal representation as m -dimensional exponential family $d\mathbb{P}_{\theta'}^{x|y} = \bar{k}(x|y, \theta')d\mu_y$, where $\theta' = P\theta$ for an orthogonal projection P on a m -dimensional subspace V of \mathbb{R}^n . The conditional distributions satisfy $\mathbb{P}_\theta^{x|y} = \mathbb{P}_{\theta'}^{x|y}$, with sufficient statistic $T' = P \circ T$ now minimal. Under this reduction the Kullback-Leibler divergences are related as $K_y(\theta||\theta^+) = \bar{K}_y(\theta'||\theta'^+)$.*

Proof: We have $T : X \rightarrow \mathbb{R}^n$, and $T(h^{-1}(y)) \subset t_0 + V$ for a linear subspace V of \mathbb{R}^n of minimal dimension m , where we may assume without loss of generality that $t_0 \in V^\perp$. Let P be the orthogonal projection onto V . Write $T(x) = t_0 + v(x)$ for $v(x) \in V$. We have

$$\begin{aligned} \psi_y(\theta) &= \log \int_{h^{-1}(y)} e^{\theta \cdot t_0} e^{\theta \cdot v(x)} d\mu_y(x) \\ (19) \quad &= \theta \cdot t_0 + \log \int_{h^{-1}(y)} e^{P\theta \cdot v(x)} d\mu_y(x) =: \theta \cdot t_0 + \bar{\psi}_y(P\theta), \end{aligned}$$

where we use the fact that $\theta \cdot v = P\theta \cdot v$ for $v \in V$, and where in consequence the rightmost term $\bar{\psi}_y(P\theta)$ depends only on $P\theta$. Now $Pt_0 = 0$ implies $PT(x) = v(x)$, hence

$$\begin{aligned} \log k(x|y, \theta) &= \theta \cdot T(x) - \psi_y(\theta) \\ (20) \quad &= \theta \cdot t_0 + \theta \cdot v(x) - \psi_y(\theta) \\ &= \theta \cdot t_0 + \theta \cdot v(x) - \bar{\psi}_y(P\theta) - \theta \cdot t_0 \quad (\text{using (19)}) \\ &= P\theta \cdot v(x) - \bar{\psi}_y(P\theta) =: \log \bar{k}(x|y, P\theta), \end{aligned}$$

using again $\theta \cdot v = P\theta \cdot v$ for $v \in V$. This gives the representation

$$(21) \quad d\mathbb{P}_\theta^{x|y} = k(\cdot|y, \theta)d\mu_y = \bar{k}(\cdot|y, \theta')d\mu_y = d\mathbb{P}_{\theta'}^{x|y},$$

where the new sufficient statistic is $v = P \circ T$, and the new parameter is $\theta' = P\theta$. Note that $v = P \circ T$ is affinely independent on $h^{-1}(y)$, because $T(h^{-1}(y)) - t_0 \subset V$ has full dimension m in V by the choice of V , and we have $T(h^{-1}(y)) - t_0 = P[T(h^{-1}(y)) - t_0] = v(h^{-1}(y))$, so that the latter has also full dimension in V .

It remains to compare the Kullback-Leibler divergences generated by both representations.

$$\begin{aligned} \bar{K}_y(\theta'||\theta'^+) &= \int_{h^{-1}(y)} \log \frac{\bar{k}(x|y, \theta')}{\bar{k}(x|y, \theta'^+)} \bar{k}(x|y, \theta') d\mu_y(x) \\ &= \int_{h^{-1}(y)} \log \frac{\bar{k}(x|y, \theta')}{\bar{k}(x|y, \theta'^+)} k(x|y, \theta) d\mu_y(x) \quad (\text{using (21)}) \\ &= \int_{h^{-1}(y)} [(P\theta - P\theta^+) \cdot v(x) - \bar{\psi}_y(P\theta) + \bar{\psi}_y(P\theta^+)] k(x|y, \theta) d\mu_y(x) \\ &= \int_{h^{-1}(y)} [(\theta - \theta^+) \cdot v(x) - \psi_y(\theta) + \theta \cdot t_0 + \psi_y(\theta^+) - \theta^+ \cdot t_0] k(x|y, \theta) d\mu_y(x) \\ &= \int_{h^{-1}(y)} [(\theta - \theta^+) \cdot (v(x) + t_0) - \psi_y(\theta) + \psi_y(\theta^+)] k(x|y, \theta) d\mu_y(x) \\ &= \int_{h^{-1}(y)} [(\theta - \theta^+) \cdot T(x) - \psi_y(\theta) + \psi_y(\theta^+)] k(x|y, \theta) d\mu_y(x) \\ &= \int_{h^{-1}(y)} \log \frac{k(x|y, \theta)}{k(x|y, \theta^+)} k(x|y, \theta) d\mu_y(x) = K_y(\theta||\theta^+). \end{aligned}$$

Finally, for notational beauty we write v as T' . □

The relationship between $\nabla_{\theta\theta}^2 \psi_y(\theta)$ and $\nabla_{\theta'\theta'}^2 \bar{\psi}_y(\theta')$ is as follows.

Corollary 1. *There exists a $n \times n$ orthogonal matrix Q with*

$$(22) \quad Q \nabla_{\theta\theta}^2 \psi_y(\theta) Q^T = \left[\begin{array}{c|c} \nabla_{\theta'\theta'}^2 \bar{\psi}_y(\theta') & 0 \\ \hline 0 & 0 \end{array} \right], \quad Q\theta = \begin{bmatrix} \theta' \\ \theta'' \end{bmatrix}, \quad \nabla_{\theta'\theta'}^2 \bar{\psi}_y(\theta') \succ 0,$$

and in particular $m = \dim(\theta')$ is the rank of $\nabla_{\theta\theta}^2 \psi_y(\theta)$. The KL-divergence $K_y(\theta||\theta^+)$ depends only on the coordinate θ' and is separating on the subspace $V = \text{im}(P)$, where P is the orthogonal projection $\theta' = P\theta$. Moreover $\nabla_{22}^2 \bar{K}_y(\theta'|\theta') \succ 0$.

Proof: The projections $P : \mathbb{R}^n \rightarrow V$ and $I - P : \mathbb{R}^n \rightarrow V^\perp$ used in the proof of Proposition 3 are represented by an orthogonal $n \times n$ matrix Q giving the change of variables in (22), where $\theta' \in V$, $\theta'' \in V^\perp$. Then k and the log-normalizer $\bar{\psi}_y$ depend only on θ' .

Since the family $\bar{k}(x|y, \theta')$ is minimal, we have indeed $\nabla_{\theta'\theta'}^2 \bar{\psi}_y(\theta') \succ 0$ by Lemma 2. Since in a minimal family $\theta' \mapsto \bar{k}(\cdot|y, \theta')$ is 1-1, $\bar{K}_y(\theta'|\theta'^+) = 0$ gives in the first place $\bar{k}(\cdot|y, \theta') = \bar{k}(\cdot|y, \theta'^+)$ μ_y -a.e., and then $\theta' = \theta'^+$. The last claim follows from Lemma 5. \square

Remark 9. The rank m depends on y . Since $\dim(\theta') = m$, we call θ' the *accurate* parameter, unique up to an orthogonal change of coordinates. Its meaning is that the family $k(x|y, \theta)$ is overparametrized by the $n - m$ *spare* parameters θ'' , and has a statistic with $n - m$ too many components $T_j(x)$, while θ' maintains only the accurate number m of parameters needed.

Remark 10. While the topological dimension of $T(h^{-1}(y))$ is typically smaller than n , it is possible that $m = n$ for the affine dimension of $T(h^{-1}(y))$. For instance, $T(h^{-1}(y))$ might be a space curve in 3d-space, which has topological dimension 1, but affine dimension 3. This has the ironic consequence that our convergence result for such curved fibers $h^{-1}(y)$ is *a priori* better than for the more likely case where fibers are affine subspaces.

What we have found is that $\Psi(\theta^+, \theta) = K_y(\theta||\theta^+)$ in (16) is a partial regularizer on the subspace $V = \text{im}(P)$ in the sense of Section 2.4. This calls now for our central convergence result under partial regularization, which we give in the next section.

6. CONVERGENCE WITH INTERIORITY

We prove partial convergence of the generalized proximal method (4) under the assumption that the sequence of iterates is bounded and together with its set of accumulation points stays in the interior G of the domain of Ψ . Since the results are of general nature, we switch to the notation familiar in optimization.

We consider an extension where (4) is solved approximatively in the sense that

$$(23) \quad e_k = g_k + \lambda_{k-1}^{-1} \nabla_1 \Psi(x_k, x_{k-1}) \text{ and } f(x_k) + \lambda_{k-1}^{-1} \Psi(x_k, x_{k-1}) \leq f(x_{k-1}),$$

with $g_k \in \partial f(x_k)$ and a subgradient error e_k satisfying one of the following conditions:

- a. $\sum_k \lambda_{k-1} \|e_k\| < \infty$;
- b. $\lambda_{k-1} \|e_k\| \leq M' \|\nabla_1 \Psi(x_k, x_{k-1})\|$ for some fixed big $M' > 0$;
- c. $\|e_k\| \leq M'' \|g_k\|$ for some big $M'' > 0$.

Theorem 1. (Partial convergence). *Suppose f satisfies the KL-inequality on G . Consider a sequence x_k generated by the approximate proximal method, which is bounded and together with its accumulation points stays in G . Let Ψ_V have pointwise lower norm bounds and be separating on a subspace V with projection P . Assume $\lambda_k/\lambda_{k-1} \leq r < \infty$ and $\lambda_k \leq R < \infty$. Then the sequence Px_k converges. When $\sum_k \lambda_k = \infty$, then at least one accumulation point of the x_k is critical, and if $\lambda_k \geq \eta > 0$, then all accumulation points are critical.*

Proof: 1) From (4) in the case where $e_{k+1} = 0$, respectively from (23), we have

$$(24) \quad f(x_{k+1}) + \lambda_k^{-1} \Psi(x_{k+1}, x_k) \leq f(x_k) + \lambda_k^{-1} \Psi(x_k, x_k) = f(x_k),$$

using $\Psi(x, x) = 0$. For $\Psi(x_{k+1}, x_k) > 0$ values are strictly decreasing. The case $f(x_{k+1}) = f(x_k)$ only occurs when $\Psi(x_{k+1}, x_k) = 0$, which by (23) implies $e_{k+1} \in \partial f(x_{k+1})$. Here the algorithm stops when $e_{k+1} = 0 \in \partial f(x_{k+1})$, but continues for $e_{k+1} \neq 0$. We may therefore concentrate on the case where the algorithm does not terminate finitely.

2) By hypothesis the sequences x_k and $f(x_k)$ are bounded, accumulation points of the x_k are in G , and by monotone convergence we have $f(x_k) \rightarrow f^* \in \mathbb{R}$. From this and (24) we immediately get $\lambda_k^{-1}\Psi(x_{k+1}, x_k) \rightarrow 0$, using $\Psi \geq 0$. Since $\lambda_k^{-1} \geq R^{-1} > 0$, we deduce $\Psi(x_{k+1}, x_k) \rightarrow 0$. We argue that this implies $P(x_{k+1} - x_k) \rightarrow 0$. Indeed, assume that there is an infinite subsequence $k \in N \subset \mathbb{N}$ with $\|P(x_{k+1} - x_k)\| \geq \epsilon > 0$. Using boundedness of x_k , extract a sub-subsequence $k' \in N' \subset N$ such that $x_{k'} \rightarrow x$, $x_{k'+1} \rightarrow x'$. Then $\Psi(x_{k'+1}, x_{k'}) \rightarrow \Psi(x', x) = 0$, and since Ψ is separating on V , we have $Px' = Px$, forcing $P(x_{k'+1} - x_{k'}) \rightarrow 0$, a contradiction.

3) We argue that this implies $\Psi(x_{k+1}, x_k) \geq m\|P(x_{k+1} - x_k)\|^2$ from some counter onwards. Indeed, let K be the compact set of accumulation points of the x_k . Applying Lemma 7 to the regularizer Ψ_V in the space V , we find that there exist $m, \delta > 0$ such that $m\|P(y - z)\|^2 \leq \Psi(y, z)$ for all $y, z \in N(K, \delta)$ with $\|P(y - z)\| < \delta$. But clearly $x_{k+1}, x_k \in N(K, \delta)$ from some counter k onward, and as $P(x_{k+1} - x_k) \rightarrow 0$, we also have $\|P(x_{k+1} - x_k)\| < \delta$ from some counter onwards. This proves the claim.

4) The sequence x_k being bounded, its set of accumulation points K is compact, and f has constant value f^* on K . Hence the KL-inequality (5) holds on a neighborhood U of K . Since there are only finitely many x_k outside U , on re-labeling the sequence, we may assume that (5) holds for the entire sequence.

By concavity of the de-singularizing function ϕ in (5) we have

$$\begin{aligned}
 \phi(f(x_k) - f^*) - \phi(f(x_{k+1}) - f^*) &\geq \phi'(f(x_k) - f^*)(f(x_k) - f^* - (f(x_{k+1}) - f^*)) \\
 &= \phi'(f(x_k) - f^*)(f(x_k) - f(x_{k+1})) \\
 &\geq \phi'(f(x_k) - f^*)\lambda_k^{-1}\Psi(x_{k+1}, x_k) \\
 &\geq \phi'(f(x_k) - f^*)\lambda_k^{-1}m\|Px_{k+1} - Px_k\|^2.
 \end{aligned}
 \tag{25}$$

Here the third line uses (24), while the last line uses the lower norm bound on V , which as mentioned above is based on Lemma 7, applied to Ψ_V .

5) By the Kurdyka-Łojasiewicz inequality (5) we have

$$\phi'(f(x_k) - f^*)\|g_k\| \geq \gamma,$$

using $g_k \in \partial f(x_k)$. Applying the partial upper norm bound and approximate optimality at stage $k - 1$, we get

$$\begin{aligned}
 \phi'(f(x_k) - f^*)^{-1} &\leq \gamma^{-1}\|g_k\| = \gamma^{-1}(\|e_k\| + \lambda_{k-1}^{-1}\|\nabla_1\Psi(x_k, x_{k-1})\|) \\
 &\leq \gamma^{-1}(\|e_k\| + M\lambda_{k-1}^{-1}\|Px_k - Px_{k-1}\|),
 \end{aligned}
 \tag{26}$$

where the upper norm bound on V occurs in the last estimate. This bound uses Lemma 6 in the next section applied to Ψ_V . Combining this with (25) gives

$$\phi(f(x_k) - f^*) - \phi(f(x_{k+1}) - f^*) \geq \frac{\gamma\lambda_k^{-1}m\|Px_{k+1} - Px_k\|^2}{\|e_k\| + M\lambda_{k-1}^{-1}\|Px_k - Px_{k-1}\|}.$$

Since $a^2 \leq bc$ for $a, b, c \geq 0$ implies $a \leq \frac{1}{2}b + \frac{1}{2}c$, we get

$$\begin{aligned}
 \|Px_{k+1} - Px_k\| &\leq \frac{1}{2}(\|Px_k - Px_{k-1}\| + M^{-1}\lambda_{k-1}\|e_k\|) \\
 &\quad + \frac{1}{2}\gamma^{-1}\lambda_k/\lambda_{k-1}M/m(\phi(f(x_k) - f^*) - \phi(f(x_{k+1}) - f^*))
 \end{aligned}
 \tag{27}$$

and setting $C = \frac{Mr}{\gamma m}$ while using $\lambda_k/\lambda_{k-1} \leq r$, we get

$$\|Px_k - Px_{k+1}\| \leq \frac{1}{2}\|Px_k - Px_{k-1}\| + \frac{\lambda_{k-1}}{2M}\|e_k\| + \frac{C}{2}[\phi(f(x_k) - f^*) - \phi(f(x_{k+1}) - f^*)].
 \tag{28}$$

Summing this from $k = 1$ to $k = n$ gives

$$\sum_{k=1}^n \|Px_{k+1} - Px_k\| \leq \frac{1}{2} \sum_{k=1}^n \|Px_k - Px_{k-1}\| + \frac{1}{2M} \sum_{k=1}^n \lambda_{k-1} \|e_k\| + \frac{C}{2} [\phi(f(x_1) - f^*) - \phi(f(x_{n+1}) - f^*)].$$

Hence

$$\begin{aligned} \sum_{k=1}^n \|Px_{k+1} - Px_k\| &\leq \|Px_1 - Px_0\| + C [\phi(f(x_1) - f^*) - \phi(f(x_{n+1}) - f^*)] \\ &\quad - \|Px_{n+1} - Px_n\| + \frac{1}{2M} \sum_{k=1}^n \lambda_{k-1} \|e_k\| \\ &\leq \|x_1 - x_0\| + C\phi(f(x_1) - f^*) + \frac{1}{2M} \sum_{k=1}^{\infty} \lambda_{k-1} \|e_k\|. \end{aligned}$$

Under condition a. the series on the right converges, hence $\sum_k \|Px_{k+1} - Px_k\| < \infty$, so that Px_k is a Cauchy sequence, which converges to some $Px^* \in V$, where x^* is an accumulation point of the x_k . In fact, Px^* is the same for all accumulation points x^* of the x_k .

6) It remains to show that at least one accumulation point x^* is critical under condition a. Since $x_k \in G$, we have

$$e_{k+1} \in \partial f(x_{k+1}) + \lambda_k^{-1} \nabla_1 \Psi(x_{k+1}, x_k),$$

so we can write $\lambda_k e_{k+1} = \lambda_k g_{k+1} + v_{k+1}$ for $g_{k+1} \in \partial f(x_{k+1})$ and $v_{k+1} = \nabla_1 \Psi(x_{k+1}, x_k)$. Now $\|\nabla_1 \Psi(x_{k+1}, x_k)\| \leq M \|Px_{k+1} - Px_k\|$ and $\sum_k \|Px_{k+1} - Px_k\| < \infty$, hence $v \in \ell_1$, and since also $\lambda \cdot e \in \ell_1$ by hypothesis a., we must have $\lambda \cdot g \in \ell_1$. By assumption, $\lambda \notin \ell_1$, and this means g cannot be bounded away from 0. In other words, $g_{k'} \rightarrow 0$ for at least a subsequence k' , and then $0 \in \partial f(x^*)$ from $x_{k'} \rightarrow x^*$, $g_{k'} \in \partial f(x_{k'})$, ∂f being upper semi-continuous.

7) If the stronger $\lambda_k \geq \eta > 0$ holds, then every accumulation point x^* is critical, because in that case we must have $g \in \ell_1 \subset c_0$ for the entire sequence.

8) It remains to discuss conditions b. and c. Under b. we can directly get rid of the term $\|e_k\|$ in the estimate (26), and the same goes for condition c. The remainder of the proof is then simplified as we can work as if $e_k = 0$. \square

Remark 11. 1) For $P = I$, $\Psi(x^+, x) = \frac{1}{2} \|x^+ - x\|^2$, $\eta \leq \lambda_k \leq R$ and $e_k = 0$ this was proved by Attouch and Bolte [5]. See also [16, Thm. 24].

2) When $P = I$ then $\sum_k \lambda_{k-1} \|e_k\| < \infty$ suffices for the limit point x^* to be critical.

3) Constraints $x \in M$ are included directly by letting $f = f_0 + i_M$. Inf-compactness of $f = f_0 + i_M$ on G assures boundedness of the sequence x_k and is trivially satisfied if $M \subset G$ is closed bounded.

4) We can dispense with the hypothesis of separatingness of Ψ_V if we assume $Px_k - Px_{k-1} \rightarrow 0$.

5) If we assume that upper and lower norm bounds still hold with the same m, M as iterates approach the boundary ∂G , then convergence holds also at the boundary. However, in the context of EM this is not a realistic assumption.

6.1. Lemmas for convergence. Recall that the general regularizer $\Psi(x^+, x)$ is of class C^2 in the first variable on $G \times G$, where $G \subset \text{dom} \Psi(\cdot, y)$ for every $y \in G$, and $G \subset \text{dom} \Psi(x, \cdot)$ for every $x \in G$. Moreover, $\nabla_1 \Psi(x, y)$ and $\nabla_{11}^2 \Psi(x, y)$ are jointly continuous. We have $\Psi(\cdot, x) \geq 0$ and $\Psi(x, x) = 0$, hence $\nabla_1 \Psi(x, x) = 0$ and $\nabla_{11}^2 \Psi(x, x) \succeq 0$ from the necessary optimality conditions.

Lemma 6. (Upper norm bound). *Let $K \subset G$ be compact convex. Then there exist $M > 0$ and $\delta > 0$ such that $\|\nabla_1 \Psi(x, y)\| \leq M \|x - y\|$ for all $x, y \in N(K, \delta)$.*

Proof: Using compactness of $K \times K$ and continuity of $(x, y) \mapsto \nabla_{11}^2 \Psi(x, y)$ on $G \times G$, choose $M > 0$ such that $\lambda_{\max}(\nabla_{11}^2 \Psi(x, y)) \leq M/2$ for all $x, y \in K$. Then find a neighborhood

$N(K, \delta)$ of K such that $\lambda_{\max}(\nabla_{11}^2 \Psi(x, y)) \leq M$ for all $x, y \in N(K, \delta)$. Fix $\|h\| = 1$, then Taylor expansion of $x \mapsto h \cdot \nabla_1 \Psi(x, y)$ at y gives

$$h \cdot \nabla_1 \Psi(x, y) = h \cdot \nabla_1 \Psi(y, y) + h \cdot \nabla_{11}^2 \Psi(\bar{y}, y)(x - y)$$

for some $\bar{y} \in (x, y)$, the open segment, where \bar{y} depends on x, y and h . Using $\nabla_1 \Psi(y, y) = 0$ and $\lambda_{\max}(\nabla_{11}^2 \Psi(\bar{y}, y)) \leq M$, we obtain $h \cdot \nabla_1 \Psi(x, y) \leq \|h\| M \|x - y\| = M \|x - y\|$, which gives the claimed estimate, since $\|h\| = 1$ is arbitrary. \square

We call this the upper norm bound. When $\nabla_{11}^2 \Psi$ is strictly positive, we have the following

Lemma 7. (Lower norm bound). *Let K be a compact convex subset of G . Suppose $\nabla_{11}^2 \Psi(x, x) \succ 0$ on K . Then there exist $\delta > 0$ and $m > 0$ such that $m\|x - y\|^2 \leq \Psi(x, y)$ for all x, y with $y \in N(K, \delta)$ and $\|x - y\| < \delta$.*

Proof: 1) Since K is compact and $\nabla_{11}^2 \Psi(x, x) \succ 0$ on G , there exists $m > 0$ such that $\nabla_{11}^2 \Psi(x, x) \succeq m > 0$ for all $x \in K$. From that we obtain a neighborhood $U = N(K, \delta)$ of K such that $\nabla_{11}^2 \Psi(y, z) \succeq m/2$ for all y, z with $z \in N(K, \delta)$ and $\|y - z\| < \delta$.

Indeed, for $x \in K$ choose $\epsilon_x > 0$ such that $\nabla_{11}^2 \Psi(y, z) \succeq m/2$ for all $y, z \in B(x, \epsilon_x)$. This is possible due to continuity of $(y, z) \mapsto \nabla_{11}^2 \Psi(y, z)$. Now let $\Delta_K = \{(x, x) : x \in K\}$ be the diagonal, then $\Delta_K \subset \bigcup_{x \in K} B(x, \epsilon_x/2) \times B(x, \epsilon_x/2)$, hence by compactness of Δ_K there are finitely many $x_i \in K$ such that $\Delta_K \subset B(x_1, \epsilon_1/2) \times B(x_1, \epsilon_1/2) \cup \dots \cup B(x_n, \epsilon_n/2) \times B(x_n, \epsilon_n/2)$, where $\epsilon_i = \epsilon_{x_i}$.

Now let $\delta := \min_{i=1, \dots, n} \epsilon_i/4$. Suppose $z \in N(K, \delta)$ and $\|y - z\| < \delta$. Find $x \in K$ with $\|z - x\| < \delta$, then $\|y - x\| < 2\delta$. For some x_i we have $(x, x) \in B(x_i, \epsilon_i/2) \times B(x_i, \epsilon_i/2)$, therefore $\|y - x_i\| < 2\delta + \epsilon_i/2 < \epsilon_i$ and $\|y - x_i\| < \delta + \epsilon_i/2 < \epsilon_i$. Hence $\nabla_{11}^2 \Psi(y, z) \succeq m/2$ by the definition of $B(x_i, \epsilon_{x_i})$.

2) Now second order Taylor expansion of $\Psi(\cdot, y)$ at y gives

$$\Psi(x, y) = \Psi(y, y) + \nabla_1 \Psi(y, y) \cdot (x - y) + \frac{1}{2}(x - y) \cdot \nabla_{11}^2 \Psi(\bar{y}, y)(x - y)$$

for some $\bar{y} \in (x, y)$, the open segment. Therefore, if $\|x - y\| < \delta$, and $y \in N(K, \delta)$, then also $\|\bar{y} - y\| < \delta$, hence by part 1), $\lambda_{\min}(\nabla_{11}^2 \Psi(\bar{y}, y)) \geq m/2$, and then using $\Psi(y, y) = 0$ and $\nabla_1 \Psi(y, y) = 0$, we get $\Psi(x, y) \geq (m/4)\|x - y\|^2$. \square

With the proof of Lemma 7 we can also get the following

Lemma 8. *Let $K \subset G$ be compact and let $\Psi(x^+, x)$ have a lower norm bound at every $x \in K$. Then there exist $\delta > 0$ and m such that $m\|y - z\|^2 \leq \Psi(y, z)$ holds for all $y, z \in N(K, \delta)$ with $\|y - z\| < \delta$.*

6.2. Rate of convergence.

Corollary 2. *Consider the case $\lambda_k \geq \eta > 0$ and $e_k = 0$ in Theorem 1. Further suppose that $\phi(s) = s^{1-\theta}/(1-\theta)$ for $\theta \in [\frac{1}{2}, 1)$. If $\theta \in (\frac{1}{2}, 1)$, then the speed of convergence is $\|Px_k - Px^*\| = O(k^{-\frac{1-\theta}{2\theta-1}})$. For $\theta = \frac{1}{2}$ the speed is R -linear.*

Proof: In the Łojasiewicz case equation (28) specializes to

$$\|Px_k - Px_{k+1}\| \leq \frac{1}{2}\|Px_k - Px_{k-1}\| + \frac{C}{2} [(f(x_k) - f^*)^{1-\theta} - (f(x_{k+1}) - f^*)^{1-\theta}].$$

Summing this from $k = N$ to $k = M$ gives

$$\begin{aligned} -\frac{1}{2}\|Px_{N-1} - Px_N\| + \frac{1}{2} \sum_{k=N}^{M-1} \|Px_k - Px_{k+1}\| + \|Px_M - Px_{M+1}\| \\ \leq \frac{C}{2} [(f(x_N) - f^*)^{1-\theta} - (f(x_{M+1}) - f^*)^{1-\theta}]. \end{aligned}$$

Passing to the limit $M \rightarrow \infty$ gives

$$-\frac{1}{2}\|Px_{N-1} - Px_N\| + \frac{1}{2}\sum_{k=N}^{\infty}\|Px_k - Px_{k+1}\| \leq \frac{C}{2}(f(x_N) - f^*)^{1-\theta}.$$

Putting $S_N = \sum_{k=N}^{\infty}\|Px_k - Px_{k+1}\|$, this becomes

$$-\frac{1}{2}(S_{N-1} - S_N) + \frac{1}{2}S_N \leq \frac{C}{2}(f(x_N) - f^*)^{1-\theta}.$$

Now from (26) we have $\phi'(f(x_N) - f^*)^{-1} \leq \gamma^{-1}M\lambda_{k-1}^{-1}\|Px_N - Px_{N-1}\| = \gamma^{-1}M\lambda_{k-1}^{-1}(S_{N-1} - S_N) \leq \gamma^{-1}M\eta^{-1}(S_{N-1} - S_N)$. Since $\phi'(s) = s^{-\theta}$, this implies

$$\begin{aligned} \phi(f(x_N) - f^*) &= (1 - \theta)^{-1}(f(x_N) - f^*)^{1-\theta} \\ &= (1 - \theta)^{-1}[\phi'(f(x_N) - f^*)^{-1}]^{\frac{1-\theta}{\theta}} \\ &\leq (1 - \theta)^{-1}(M\gamma^{-1}\eta^{-1})^{\frac{1-\theta}{\theta}}(S_{N-1} - S_N)^{\frac{1-\theta}{\theta}}. \end{aligned}$$

So altogether we get

$$(29) \quad \frac{1}{2}S_N \leq C'(S_{N-1} - S_N)^{\frac{1-\theta}{\theta}} + \frac{1}{2}(S_{N-1} - S_N)$$

for $C' = (1 - \theta)^{-1}(M\gamma^{-1}\eta^{-1})^{\frac{1-\theta}{\theta}}$. Now for $\theta > \frac{1}{2}$ we have $\frac{1-\theta}{\theta} < 1$, so the first term on the right of (29) dominates the second term, and we get

$$S_N^{\frac{\theta}{1-\theta}} \leq C''(S_{N-1} - S_N)$$

for N large enough and yet another constant C'' . Following [60, Cor. 4(24)ff], this leads to an estimate $S_N \leq C'''N^{-\frac{1-\theta}{2\theta-1}}$.

It remains to discuss the case $\theta = \frac{1}{2}$. Here (29) gives

$$\frac{1}{2}S_N \leq C'(S_{N-1} - S_N) + \frac{1}{2}(S_{N-1} - S_N),$$

hence

$$S_N \leq \frac{1 + SC'}{2 + 2C'}S_{N-1}$$

which gives Q-linear convergence of the S_N , hence R-linear convergence of the Px_k . \square

Remark 12. When $\phi(s) = s^{1/2}$, which is the best possible case, the Łojasiewicz inequality (5) specializes to the Polyak-Łojasiewicz inequality. In the unconstrained case, for a critical point \bar{x} , (5) is then $(f(x) - f(\bar{x}))^{-1/2}\|\nabla f(x)\| \geq \gamma$, and that means f is locally bounded below by a quadratic, and in particular, has a strict local minimum at \bar{x} . Indeed, assuming $f(\bar{x}) = 0$, $\bar{x} = 0$ and letting $y(t) = f(th)$ for fixed $\|h\| = 1$, (5) gives $y' \geq \gamma\sqrt{y}$, hence $dy/\sqrt{y} \geq \gamma dt$, hence $\sqrt{y} \geq \frac{\gamma}{2}t$, i.e. $y \geq \frac{\gamma^2}{4}t^2$, using $y(0) = 0$. This is of course too good to be true, so we expect the Polyak-Łojasiewicz inequality to be satisfied in exceptional cases only.

In the general case of the Kurdyka-condition we can still say something:

Corollary 3. *Consider the case $\lambda_k \geq \eta > 0$ and $e_k = 0$ in Theorem 1. Then the speed of partial convergence is $\|Px_k - Px^*\| \leq C\phi(f(x_k) - f^*)$ with ϕ the de-singularizing function.*

Proof: Since $\Psi(x, x) = 0$ and $\nabla_1\Psi(x, x) = 0$, Taylor expansion of $\Psi(\cdot, x)$ at x gives $\Psi(u, x) = \Psi(x, x) + \nabla_1\Psi(x, x) \cdot (u - x) + \frac{1}{2}(u - x) \cdot \nabla_{11}^2\Psi(\bar{x}, x)(u - x) = \frac{1}{2}(u - x) \cdot \nabla_{11}^2\Psi(\bar{x}, x)(u - x)$ for \bar{x} on the segment (x, u) . Since for a compact set $K \subset G$ we find $c > 0$ such that $\nabla_{11}^2\Psi(x', x) \preceq c^2I$ for all $x', x \in K$, choosing as K the convex hull of the set of iterates and its accumulation points, we have the estimate $\Psi(x^+, x)^{1/2} \leq c\|x^+ - x\|$ for the sequence generated by (4).

Since the solution x^+ of (4) from the current x is exact, letting $f(x^+) = r^+$, $f(x) = r$, we have $\Psi(x^+, x) \leq \Psi(u, x)$ for all $u \in \{f \leq r^+\}$. Hence $\Psi(x^+, x)^{1/2} \leq \min_{u \in \{f \leq r^+\}} \Psi(u, x)^{1/2} \leq c \min_{u \in \{f \leq r^+\}} \|u - x\| = c \operatorname{dist}(\{f \leq r^+\}, x) \leq c \max_{v \in \{f \leq r\}} \operatorname{dist}(\{f \leq r^+\}, v) \leq c \operatorname{haus}(\{f \leq r^+\}, \{f \leq r\})$, where we used that x is one of the $v \in \{f \leq r\}$.

According to [16, Thm. 20(vi), Cor. 4] the KL-property is equivalent to Lipschitz continuity of the sublevel operator. More precisely, there exists $k > 0$ such that $\operatorname{haus}(\{f \leq r^+\}, \{f \leq r\}) \leq k|\phi(r^+) - \phi(r)|$, where ϕ is the de-singularizing function of (5). Substituting this gives $\Psi(x^+, x)^{1/2} \leq c \cdot \operatorname{haus}(\{f \leq r^+\}, \{f \leq r\}) \leq ck|\phi(r^+) - \phi(r)|$. Using the lower norm bound, we deduce $\|Px_{k+1} - Px_k\| \leq m^{-1/2}ck[\phi(f(x_k) - f^*) - \phi(f(x_{k+1}) - f^*)]$, and summing both sides from $k = n$ to $k = r$ gives $\sum_{k=n}^r \|Px_{k+1} - Px_k\| \leq m^{-1/2}ck[\phi(f(x_n) - f^*) - \phi(f(x_{r+1}) - f^*)]$. Letting $r \rightarrow \infty$ gives the claimed rate $\|Px_n - Px^*\| \leq m^{-1/2}ck\phi(f(x_n) - f^*)$. \square

For the Kurdyka inequality we may also argue as follows. From (28) we obtain the estimate

$$-\frac{1}{2}(S_{n-1} - S_n) + \frac{1}{2}S_n \leq \frac{c}{2}\phi(f(x_n) - f^*).$$

Fix $\alpha \in (\frac{1}{2}, 1)$, and divide integers in two classes $\mathcal{N}_1 = \{n : S_n \leq \alpha S_{n-1}\}$ and $\mathcal{N}_2 = \{n : S_n > \alpha S_{n-1}\}$. Now for $n \in \mathcal{N}_2$ we have

$$(1 - \frac{1}{2\alpha})S_n \leq S_n - \frac{1}{2}S_{n-1} \leq \frac{c}{2}\phi(f(x_n) - f^*).$$

On the other hand, for $n \in \mathcal{N}_1$ we have $S_n \leq \alpha S_{n-1}$, so here the error shrinks with linear rate. Altogether we get a sequence $n_1 < m_1 < n_2 < m_2 < \dots$ such that

$$\begin{aligned} S_n &\leq c\phi(f(x_n) - f^*) \text{ for } n = n_k, \dots, m_k - 1 \\ S_n &\leq \alpha^{n-m_k} S_{m_k}, \quad \text{for } n = m_k, \dots, n_{k+1} - 1, \end{aligned}$$

with $c = \frac{c}{2}(1 - \frac{1}{2\alpha})^{-1}$, which is a slight refinement of Corollary 3, as it leaves the option of the entire sequence S_n converging linearly with rate α even for ϕ less desingularizing than $s^{1/2}$.

7. CONVERGENCE OF THE EM ALGORITHM

Now we apply this to the EM algorithm. In the first place, we assume interiority.

Theorem 2. (Convergence for constrained exponential family). *Let $q(y, \theta)$ be incomplete data from a minimal n -dimensional exponential family $p(x, \theta)$. Suppose $-\log q(y, \cdot)$ and M are definable. Let $\theta^{(k)} \in M$ be a bounded sequence generated by the constrained EM algorithm which together with its accumulation points stays in G . Then the sequence $\theta'^{(k)}$ of accurate parameters converges, $\theta'^{(k)} \rightarrow \theta^*$. Every accumulation point θ^* of the sequence $\theta^{(k)}$ solves the constrained incomplete data MLE problem, has the same projection $P\theta^* = \theta^*$, and the conditional distributions $\mathbb{P}_{\theta^{(k)}}^{x|y}$ converge weakly to $\mathbb{P}_{\theta^*}^{x|y}$, the limit being the same for all θ^* . Moreover, the sequence $\bar{T}(x_k)$ is also convergent.*

Proof: Since $-\log q(y, \cdot)$ and M are definable, so is the objective $f = -\log q(y, \cdot) + i_M$ in (16). Hence f has the KL-property (5). By hypothesis the sequence together with its accumulation points stays in G , as required for our main convergence result.

By Proposition 3 we have $\nabla_{\theta'\theta'}^2 \bar{\psi}_y(\theta') \succ 0$ for the accurate parameter θ' , hence the partial regularizer $\Psi(\theta^+, \theta) = \bar{K}_y(\theta' || \theta^+)$ satisfies $\nabla_{22}^2 \bar{K}_y(\theta' || \theta') \succ 0$. Therefore it has a lower norm bound on the compact set of accumulation points of the $\theta'^{(k)}$ contained in the subspace V of dimension m . Also, since the family $\bar{k}(\cdot | y, \theta')$ is minimal, the partial regularizer is separating on the subspace V . Therefore we can apply the main convergence theorem (with $\lambda_k = 1$), and this gives convergence of the $\theta'^{(k)}$.

Since $\mathbb{P}_{\theta^{(k)}}^{x|y} = \mathbb{P}_{\theta'^{(k)}}^{x|y}$ by Proposition 3, and since the right hand sequence converges weakly, so does the left hand sequence. It also follows from a classical result of F. Riesz that $\bar{k}(\cdot | y, \theta'^{(k)})$ converges to $\bar{k}(\cdot | y, \theta^*)$ in $L^1(h^{-1}(y), d\mu_y)$. But then due to (20) the sequence $k(\cdot | y, \theta^{(k)})$ also converges in L^1 , regardless of whether the $\theta''^{(k)}$ converge.

As a consequence of the main convergence theorem, every accumulation point θ^* of the full sequence $\theta^{(k)}$ is critical, and $P\theta^* = \theta^*$ for every such θ^* . Convergence of $\bar{T}(x_{k+1})$ follows because $\bar{T}(x_{k+1})$ is the result of computing $\mathbb{E}_{\theta^{(k)}}[\bar{T}(x)|y]$, which depends continuously on $\theta^{(k)}$. \square

Remark 13. Partial convergence is also guaranteed under the more general choices of λ_k in Theorem 1, but that requires computing $-\log q(y, \theta^+)$ and $\bar{K}_y(\theta' || \theta^+)$ separately. In practice one prefers to use the Q-function.

Corollary 4. *Under the hypotheses of Theorem 2, suppose $T(h^{-1}(y))$ is not contained in a proper affine subspace of \mathbb{R}^n . Then the entire sequence $\theta^{(k)}$ converges.*

Proof: In this situation no dimension reduction takes place and we have $\theta^{(k)} = \theta'^{(k)}$. Hence $\theta^{(k)}$ converges. \square

7.1. Trouble at the boundary. From (6) we have $\text{dom } q(y, \cdot) = \Theta$, while (8) gives $\Theta \subset \text{dom } \Psi(\cdot, \theta')$ for every $\theta' \in G$. But this does not exclude the possibility that iterates lie on the boundaries of $\text{dom}(-\log q(y, \cdot))$ and $\text{dom } \Psi(\cdot, \theta')$ simultaneously, causing problems (see Example 1). We take a closer look.

Definition 3. We call the incomplete data problem *regular* if $f(\theta) = -\log q(y, \theta) + i_M(\theta) = \infty$ for all $\theta \in \partial\Theta$, and we call it *steep* if $\partial f(\theta) = \partial(-\log q(y, \cdot) + i_M)(\theta) = \emptyset$ for all $\theta \in \partial\Theta$.

Remark 14. 1) We have $\partial f(\theta) = \partial(-\log q(y, \cdot) + i_M)(\theta) = \partial(-\log q(y, \cdot))(\theta) + \mathcal{N}_M(\theta)$ for $\theta \in G$ by [69, 8.8. c] or [55], but the sum rule fails at the boundary $\partial\Theta$, unless additional regularity hypotheses are made. This is why the definition uses $\partial(-\log q(y, \cdot) + i_M)$.

2) Regular implies steep. If $M \subset G = \text{int}(\Theta)$, then the problem is automatically regular.

3) When $q(y, \cdot)$ is from an exponential family and $M = \bar{\Theta}$, then steepness in the sense of the definition is equivalent to steepness of the log-normalizer ψ_q of q .

4) Suppose the complete data family is steep in the sense that whenever $\theta \in \partial\Theta$, then $\partial_\theta(p(x, \cdot) + i_M)(\theta) = \emptyset$ for μ_y -a.a. $x \in h^{-1}(y)$. Then steepness of the incomplete data problem follows from the inclusion $\partial_\theta(q(y, \cdot) + i_M)(\theta) \subset \int_{h^{-1}(y)} \partial_\theta(p(x, \cdot) + i_M)(\theta) d\mu_y(x)$ (see [26, Thm. 2.7.2]). This happens when $p(x, \theta)$ is of exponential type, as then $\partial_\theta p(x, \theta) = p(x, \theta)(T(x) + \partial\psi(\theta))$, so that $\partial\psi(\theta) = \emptyset$ for $\theta \in \partial\Theta \cap M$ forces $\partial_\theta p(x, \theta) = \emptyset$ for all x . This justifies the definition.

Let $\theta^{(k)}$ be the sequence generated by the constrained EM algorithm, then every accumulation point θ^* must have finite value $f(\theta^*) < \infty$, because $\theta^{(k)} \in \{f \leq f(\theta^{(1)})\}$, and by lower semi-continuity of f this set is closed. Therefore in the regular case no θ^* can be on the boundary $\partial\Theta$. Hence the interiority hypothesis in Theorem 2 is automatically satisfied for bounded $\theta^{(k)}$. Boundedness is assured e.g. under inf-compactness of f . In other words, if the incomplete data problem is regular, there ain't any trouble at the boundary $\partial\Theta$.

Now suppose f is steep, and let $e_k = 0$. By optimality $-\lambda_{k-1}^{-1} \nabla_1 \Psi(\theta^{(k)}, \theta^{(k-1)}) \in \partial f(\theta^{(k)})$, hence $\theta^{(k)}$ cannot be on the boundary $\partial\Theta$, and the method is well-defined. Suppose $\theta^{(k)} \rightarrow \theta^* \in \partial\Theta$ for a subsequence $k \in N \subset \mathbb{N}$. Then by steepness θ^* cannot be a solution of the constrained incomplete data MLE program, because that would give $0 \in \partial f(\theta^*) = \partial(-\log q(y, \cdot) + i_M)(\theta^*) = \emptyset$. Unfortunately, steepness alone does *not* prevent iterates from approaching $\partial\Theta$, but at least we know in that case that these iterates go astray.

Remark 15. When $M \subset G = \text{int}(\Theta)$ is bounded, then the sequence $\theta^{(k)}$ together with its accumulation points stays in G and all trouble at the boundary $\partial\Theta$ is avoided. Conversely, suppose we wish to make a statement about a sequences $\theta^{(k)}$ respecting interiority, i.e., bounded and contained in G together with its accumulation points. Then we can replace the constraint set $M \subset \bar{\Theta}$ by a closed subset $M' \subset M \cap G$ so that the sequence may be

considered as generated under the constraints $\theta^{(k)} \in M'$. This set M' can be chosen bounded, definable if M is definable, and convex if M is convex. See [59, Sect. 2.2].

7.2. Consequences for the M step. Having established partial convergence $\theta'^{(k)} \rightarrow \theta'^*$ under interiority, we now investigate under what conditions we may upgrade this to convergence of the full parameter sequence $\theta^{(k)}$. Observe that in the new coordinates $Q\theta = (\theta', \theta'')$ the M-step (16) has the equivalent form:

$$(\theta'^{(k+1)}, \theta''^{(k+1)}) \in \underset{\theta', \theta''}{\operatorname{argmin}} -\log q(y, \theta', \theta'') + i_M(\theta', \theta'') + \lambda_k^{-1} \Psi(\theta', \theta'^{(k)}),$$

with

$$\Psi(\theta', \theta'^{(k)}) = -\mathbb{E}_{\theta'^{(k)}} \left[\log \frac{\bar{k}(x|y, \theta')}{\bar{k}(x|y, \theta'^{(k)})} \middle| y \right] = \bar{K}_y(\theta'^{(k)} || \theta'),$$

a partial regularizer independent of the variable θ'' . Identifying for simplicity θ with (θ', θ'') , define $M(\theta') = \{\theta'' : (\theta', \theta'') \in M\}$, then we can split the M step optimization program as follows:

$$(P_k) \quad \theta''^{(k+1)} \in \underset{\theta'' \in M(\theta'^{(k+1)})}{\operatorname{argmin}} -\log q(y, \theta'^{(k+1)}, \theta''),$$

which is just a sequence of parametrized optimization programs in θ'' , with $\theta'^{(k+1)}$ the parameter, and no longer any regularization affecting θ'' . The limiting program is clearly:

$$(P_\infty) \quad \theta''^* \in \underset{\theta'' \in M(\theta'^*)}{\operatorname{argmin}} -\log q(y, \theta'^*, \theta''),$$

and since $\theta'^{(k)} \rightarrow \theta'^*$, every accumulation point (θ'^*, θ''^*) of the EM sequence $\theta^{(k)}$ gives a solution θ''^* of (P_∞) with the same incomplete data MLE value $q^* = q(y, \theta^*)$. This has the following immediate consequence:

Proposition 4. *If the limiting program (P_∞) has a unique critical point $\theta''^* \in M(\theta'^*)$ among those with critical value q^* , then the EM sequence $\theta^{(k)}$ converges.*

This is a weaker hypothesis than requesting as e.g. in [78], that the full incomplete data MLE program has a unique solution θ^* with the correct MLE value.

Remark 16. Note, however, that in each program (P_k) we are free to choose *any* of the local solutions in case there are several. If there exists $\epsilon > 0$ such that in every (P_k) one can choose two local solutions a distance ϵ apart, then failure of convergence of the spare sequence $\theta''^{(k)}$ is inevitable. This happens for instance in the counterexample in [76, Sect. 4].

A second consequence is based on the following.

Proposition 5. *Suppose the complete data exponential family $p(x, \theta)$ is minimal. Then $-\nabla_{\theta''\theta''}^2 \log q(y, \theta', \theta'') \succ 0$ for fixed θ' .*

Proof: Using standard notation, one defines $I(\theta, y) = -\nabla_{\theta\theta}^2 \log q(y, \theta)$, which makes $\mathcal{I}(\theta) = \mathbb{E}_\theta[I(\theta, y)]$ the expected Fisher information of incomplete data. For complete data one defines $I_c(\theta, x) = -\nabla_{\theta\theta}^2 \log p(x, \theta)$, then $\mathcal{I}_c(\theta) = \mathbb{E}_\theta[I_c(x, \theta)]$ is the expected Fisher information of complete data. In the same vein, one also lets $\mathcal{I}_c(\theta, y) = \mathbb{E}_\theta(I_c(\theta, x)|y) = \nabla^2 \psi(\theta)$, the conditional expected Fisher information of complete data given y .

Now from (8) we get $\log p(x, \theta) = \log q(y, \theta) + \log k(x|y, \theta)$, hence differentiating twice gives

$$I_c(\theta, x) = I(\theta, y) - \nabla_{\theta\theta}^2 \log k(x|y, \theta).$$

Taking conditional expectations over x given y , we obtain

$$(30) \quad \mathcal{I}_c(\theta, y) = I(\theta, y) + \mathcal{I}_m(\theta, y),$$

where $\mathcal{I}_m(\theta, y) = \mathbb{E}_\theta(-\nabla_{\theta\theta}^2 \log k(x|y, \theta)|y)$ is the expected Fisher information of missing data conditioned on y . The latter, however, was previously identified as the second derivative of the regularizer.

Now for $k(x|y, \theta)$ an exponential family, and adopting the change of coordinates in (22), we know that

$$\mathcal{I}_m(\theta, y) = \left[\begin{array}{c|c} \nabla_{\theta'\theta'}^2 \bar{\psi}_y(\theta') & 0 \\ \hline 0 & 0 \end{array} \right], \quad \nabla_{\theta'\theta'}^2 \bar{\psi}_y(\theta') \succ 0,$$

while

$$I(\theta, y) = -\nabla_{\theta\theta}^2 \log q(y, \theta', \theta'') = \left[\begin{array}{c|c} -\nabla_{\theta'\theta'}^2 \log q(y, \theta', \theta'') & -\nabla_{\theta'\theta''}^2 \log q(y, \theta', \theta'') \\ \hline * & -\nabla_{\theta''\theta''}^2 \log q(y, \theta', \theta'') \end{array} \right].$$

Since $p(x, \theta)$ is minimal by hypothesis, we have $I_c(\theta, x) = \nabla^2 \psi(\theta) \succ 0$. Therefore also $\mathcal{I}_c(\theta, y) \succ 0$. In consequence, the matrices $\mathcal{I}_m(\theta, y)$ and $I(\theta, y)$ on the right of (30) must add up to a matrix of full rank. Due to the structure of $\mathcal{I}_m(\theta, y)$ this forces $-\nabla_{\theta''\theta''}^2 \log q(y, \theta) \succ 0$ for the lower diagonal block in $I(\theta, y)$. \square

Applying this with θ^* shows that the objective function $-\log q(y, \theta^*, \cdot)$ of (P_∞) is strictly convex. We therefore have the following consequence:

Theorem 3. *Under the assumptions of Theorem 2, suppose $M(\theta^*)$ is convex. Then the EM sequence $\theta^{(k)}$ converges.*

Proof: A strictly convex function has a unique minimum on a convex domain. \square

Note that $M(\theta^*)$ is clearly convex if M is convex, but convexity of $M(\theta^*)$ is a weaker hypothesis. In particular we have the following

Corollary 5. *Suppose the constraint set is $M = \{\theta(u) : u \in U\} = \{(\theta'(u), \theta''(u)) : u \in U\}$, with $U \subset \mathbb{R}^m$, $\theta'(\cdot)$ of class C^1 and $\theta''(\cdot)$ continuous, both definable. Let $\theta'(u^*) = \theta^*$, and suppose the rank of the Jacobian $\frac{d\theta'}{du}(u^*)$ is m . Then the EM sequence $\theta^{(k)}$ converges. When $\theta''(\cdot)$ is locally Lipschitz, then the speed of convergence of $\theta^{(k)}$ is the same as that of $\theta'^{(k)}$.*

Proof: Under the rank hypothesis the mapping $u \mapsto \theta'(u)$ has locally a left inverse, i.e., we have a C^1 mapping $\theta' \mapsto u(\theta')$ defined in a neighborhood of θ^* such that $u(\theta^*) = u^*$ and $u(\theta'(u)) = u$. Then $\theta'' = \theta''(u) = \theta''(u(\theta'))$, so that θ'' is a function of θ' . Then $M(\theta^*)$ is singleton, hence convex. An even more direct argument is that $\theta''^{(k)} = \theta''(u(\theta'^{(k)}))$ converges by continuity of $\theta''(\cdot)$ and $u(\cdot)$. The statement concerning speed follows because when $\theta''(\cdot)$ is locally Lipschitz, then so is $\theta''(\cdot) \circ u(\cdot)$. \square

Remark 17. 1) Assuming that each (P_k) has a unique solution is not sufficient for convergence, as the $\theta''^{(k)}$ obtained may still have several accumulation points.

2) Under the somewhat artificial assumption that the set of accumulation points of the sequence $\theta''^{(k)}$ is discrete, one obtains convergence as soon as $\theta''^{(k)} - \theta''^{(k-1)} \rightarrow 0$.

3) When dependence of the solution set $\text{argmin}_{\theta''} -\log q(y, \theta', \cdot) + i_{M(\theta')}$ on the parameter θ' is upper Lipschitz (cf. [43]) on the compact set $\{\theta'^{(k)} : k \in \mathbb{N}\} \cup \{\theta^*\}$, then convergence follows, because this forces $\|\theta''^{(k)} - \theta''^{(k-1)}\| \leq L\|\theta'^{(k)} - \theta'^{(k-1)}\|$ for some $L > 0$, and since $\sum_k \|\theta'^{(k)} - \theta'^{(k-1)}\| < \infty$, the sequence $\theta''^{(k)}$ is also Cauchy. For NLP constraints M , sufficient conditions are discussed in [65, 43], are typically local, and require mild regularity hypotheses. Here these have to be satisfied at all accumulation points θ''^* of the spare sequence $\theta''^{(k)}$.

Theorem 4. (Regularized EM for exponential family). *Under the hypotheses of Theorem 2, suppose the M step is regularized as $\min_{\theta \in M} -Q(\theta, \theta^{(k)}) + \lambda_k^{-1} \|\theta'' - \theta''^{(k)}\|^2$. Then the sequence $\theta^{(k)}$ converges to a critical point θ^* which is a MLE for the incomplete data problem. The value of the incomplete data negative log-likelihood is still monotonically decreasing.*

Proof: In view of Proposition 1 we have modified the M step such that the regularizer is now $\Psi(\theta, \theta^{(k)}) = \bar{K}_y(\theta'^{(k)} || \theta') + \|\theta'' - \theta''^{(k)}\|^2$, which is no longer partial but full. We apply the

main convergence theorem with $P = I$, which gives convergence of the $\theta^{(k)}$. Since (4) gives always decrease of the objective, the last statement follows. \square

7.3. Definable objectives. We inquire whether, or when, objectives $f(\theta) = -\log q(y, \theta) + i_M(\theta)$ in (16) are definable in an o-minimal structure, because this is how we assure the KL-inequality (5). Starting with $-\log q(y, \theta)$, we first run through those cases where $q(y, \theta)$ is by itself from an exponential family, say $q(y, \theta) = \exp\{\langle \theta, T(y) \rangle - \psi_q(\theta) + h(y)\}$. Here definability hinges on definability of the corresponding log-normalizer ψ_q .

Inspecting lists of exponential families, one finds that along with algebraic expressions, log-normalizers $\psi(\theta)$ sometimes include terms of the form $\log \theta_i$ for certain components θ_i of θ . Those are definable in $\mathbb{R}_{an,exp}$. Moreover, if these θ_i can a priori be bounded and bounded away from 0, one gets definability in \mathbb{R}_{an} . Inverse gamma distribution and χ^2 -distribution call for terms of the form $\log \Gamma(\theta_i)$ for certain components of θ , which require definability of the Gamma function. This has recently been addressed e.g. in [62], and for our purpose it is again sufficient to bound these θ_i away from 0.

The second case is when $q(y, \theta)$ are incomplete data from an exponential family, as termed in [73], but do not by themselves stem from an exponential family. Here due to $-\log q(y, \theta) = -\log p(x_k, \theta) + \lambda_k^{-1} \bar{K}_y(\theta^{(k)} || \theta')$ definability of $\log q(y, \cdot)$ may be derived from definability of $\log p(x_k, \cdot)$ in tandem with definability of $\bar{K}_y(\theta^{(k)} || \cdot)$. The first is assured when ψ is definable, as discussed above. For the second we use Proposition 2, which shows that definability of the Bregman distance induced by ψ_y is required, and this follows from definability of ψ_y .

Definability of M is even less complicated, as M typically gathers equality and inequality constraints of the form $M = \{\theta \in \mathbb{R}^n : f_i(\theta) = 0, i \in I, g_j(\theta) \leq 0, j \in J\}$ for finite sets I, J and definable functions f_i, g_j , typically sub-analytic or even algebraic. Note that M may even have the benefit to restrict components θ_i to a bounded interval, which allows to replace $\mathbb{R}_{an,exp}$ by the more convenient structure \mathbb{R}_{an} , where (5) turns into the Łojasiewicz inequality.

Remark 18. When f is definable in \mathbb{R}_{an} , Corollary 2 gives a convergence rate $\|\theta^{(k)} - \theta^*\| = O(k^{-\rho})$. If in addition $\theta'' = \theta''(\theta')$ is locally Lipschitz, we get the same rate for the full parameter sequence. This holds in Corollary 4 and Theorem 4, but also in the case in Corollary 5. In Theorem 3 it also holds due to $\nabla_{\theta''\theta''}^2 - \log q(y, \theta', \cdot) \succeq \epsilon > 0$ for θ' in the compact set $\{\theta^*\} \cup \{\theta^{(k)} : k \in \mathbb{N}\}$, provided M is given by sufficiently smooth definable equality and inequality constraints, where the MFCQ is satisfied, see [43].

Linear speed is obtained in the case $\phi(s) = s^{1/2}$, which corresponds to the Polyak-Łojasiewicz inequality. Unfortunately this is a rather strong hypothesis (see also Section 8 for that aspect).

7.4. Reasons for failure. We can now list the following reasons why the EM algorithm for incomplete data from a constrained exponential family may fail to converge to critical points:

- (1) Iterates may be unbounded or tend to the boundary of Θ . Once those are ruled out:
- (2) Convergence may still fail because $-\log q(y, \cdot) + i_M$ does not have the KL-property. But even when f does have the KL-property:
- (3) It may still happen that only the accurate parameter $\theta^{(k)}$ converges, while the spare sequence $\theta''^{(k)}$ fails to converge. This may happen if $M(\theta^*)$ is not convex.
- (4) But even when $M(\theta^*)$ is convex, including the unconstrained case, convergence of the $\theta''^{(k)}$ may still fail because $-\log q(y, \theta^*, \cdot)$ is not strictly convex. This may be the case because $p(x, \theta)$ is not minimal. The latter may be avoided when setting up the problem.

In curved families $M = \{\theta(u) : u \in U\}$, chances of convergence are paradoxically even better, as some of the degrees of freedom are removed. As we had seen, convergence of the entire sequence is forced when there is a continuous dependence $\theta'' = \theta''(\theta')$. Even when this is too optimistic, as the portion of missing data is likely to be smaller than the portion of observed ones, one still gets $\theta = (\theta', \theta'', \theta''')$, where θ' is the accurate parameter, $\theta'' = \theta''(\theta')$

a portion of the spare parameter actually dependent on θ' , and therefore forced to converge, with θ''' fewer remaining spare coordinates which require additional conditions to converge.

Remark 19. Dimension reduction due to affine constraints on Θ (Lemma 1) is not critical for the question of convergence of the EM parameter sequence $\theta^{(k)}$. This is different for dimension reduction due to non-minimality of the sufficient statistic (Proposition 3).

8. ALTERNATING BREGMAN PROJECTIONS

This section finds more instances where convergence of the $\theta^{(k)}$ can be guaranteed, by matching the EM algorithm with the *em*-algorithm of [2]. We consider the case of missing data from a constrained exponential family assumed minimal, where $T(x) = (T_1(x), T_2(x)) = (y, z)$ with y observed and z hidden, and we partition $\theta = (\theta_y, \theta_z)$ accordingly. Minimality of the complete data family assures that $\eta = \nabla\psi(\theta) = \mathbb{E}_\theta[T(x)]$ is a diffeomorphism from G to G^* , with inverse $(\nabla\psi)^{-1} = \nabla\psi^*$, and we may therefore work with the expectation parameter η . Partitioning $\eta = (\eta_y, \eta_z) = (\nabla_{\theta_y}\psi(\theta), \nabla_{\theta_z}\psi(\theta))$ in the same way, we define the data set as

$$D = \{\vartheta \in \Theta : \vartheta = \nabla\psi^*(\eta), \eta_y = y\},$$

where we note data parameters as $\vartheta \in D$, keeping $\theta \in M$ for model parameters. We have

Lemma 9. (Amari [2]). *Let $\theta \in \Theta$. Then the right Bregman projection $\vartheta = \vec{P}_D(\theta)$ on the data set D is unique, satisfies $\vartheta_z = \theta_z$, and $\mathbb{E}_\theta[z|y] = \mathbb{E}_{\vartheta}[z|y]$.*

In information geometry $\vartheta_e = \vec{P}_D(\theta)$ is called the *e*-step from $\theta \in M$. It turns out that the E step from $\theta \in M$ can also be represented as a point $\vartheta_E \in D$ in the data set, and it generates the next M step as a left Bregman projection $\theta^+ \in \vec{P}_M(\vartheta_E)$.

Lemma 10. (Amari [2]). *Let $\eta_E = \nabla\psi(\vartheta_E)$, $\eta_e = \nabla\psi(\vartheta_e)$ be the expectation parameters of E and *e*-step from $\theta \in M$. Then $\eta_E = (y, \mathbb{E}_{\vartheta_e}[z|y]) \in \nabla\psi(D)$ and $\eta_e = (y, \mathbb{E}_{\vartheta_e}[z]) \in \nabla\psi(D)$.*

The question when E step and *e*-step coincide is answered by the following:

Lemma 11. (Amari [2]). *E step and e-step from θ coincide iff $\mathbb{E}_{\vartheta_e}[z|y] = \mathbb{E}_{\vartheta_e}[z]$ for $\vartheta_e = \vec{P}_D(\theta)$. If this is true all along, then EM and em-algorithm generate the same iterates.*

When Amari's condition $\mathbb{E}(Z|Y = y) = \mathbb{E}(Z)$ is satisfied, we say that Z is unpredictable based on knowledge of Y . This is a property settled between the stronger independence (of Z, Y) and the weaker uncorrelatedness ($\text{cov}(Z, Y) = 0$).

Let Amari's unpredictability condition be satisfied. Then the E step is the right Bregman projection of the iterate $\theta^{(k)} \in M$ onto the data set, $\vartheta^{(k+1)} = \vec{P}_D(\theta^{(k)})$, while the M step is the left Bregman projection of the E step iterate $\vartheta^{(k)}$ onto the model set M , that is, $\theta^{(k)} \in \vec{P}_M(\vartheta^{(k)})$, the Bregman distance being the one induced by the log-normalizer ψ of the complete data family $p(x, \theta)$. As in [59], we visualize this by a building block diagram

$$(31) \quad \vartheta \xrightarrow[m]{l} \theta \xrightarrow[e]{r} \vartheta^+ \xrightarrow[m]{l} \theta^+ \quad \vartheta^{(k)} \xrightarrow[m]{l} \theta^{(k)} \xrightarrow[e]{r} \vartheta^{(k+1)} \xrightarrow[m]{l} \theta^{(k+1)}$$

Theorem 5. (Convergence via information geometry). *Let Amari's condition be satisfied. Then the constrained EM algorithm for missing data from an exponential family generates sequences $\theta^{(k)} \in M$, $\vartheta^{(k)} \in D$ of alternating Bregman projections between D and M , where $\theta^{(k)} \in M$ is generated by the M step $\theta^{(k)} \in \vec{P}_M(\vartheta^{(k)})$, $\vartheta^{(k)} = \vec{P}_D(\theta^{(k-1)})$ the E step. Assume $p(x, \theta)$ is minimal, ψ, M are definable, and suppose $M \subset G$ is bounded. Then:*

- (a) *The sequence $\vartheta^{(k)}$ converges to some $\vartheta^* \in D$.*
- (b) *Every accumulation point θ^* of the sequence $\theta^{(k)}$ solves the constrained incomplete data MLE problem, and satisfies $\vartheta^* = \vec{P}_D(\theta^*)$, $\theta^* \in \vec{P}_M(\vartheta^*)$. There exists z^* such that $\log p(y, z^*, \vartheta^*) = \mathbb{E}_{\theta^*}(\log p(y, z, \theta)|y)$ is the same for every θ^* .*

- (c) Suppose $D \cap M \neq \emptyset$. Then $D \cap M$ has a neighborhood U such that any EM sequence which enters U converges to a point $\theta^* \in D \cap M$, i.e., $\theta^{(k)} \rightarrow \theta^*$ and $\vartheta^{(k)} \rightarrow \theta^* = \vartheta^*$.
- (d) The sequence $\theta^{(k)}$ of accurate parameters converges to some θ^* , the conditional distributions converge weakly, and every θ^* gives rise to the same $P\theta^* = \theta'^*$, $\vartheta^* = \vec{P}_D(\theta^*)$.
- (e) When $M(\theta^*)$ is convex, then the sequence $\theta^{(k)}$ converges to some $\theta^* \in M$ with $\vec{P}_D(\theta^*) = \vartheta^*$ and $\vec{P}_M(\vartheta^*) = \theta^*$.
- (f) Suppose the EM instance is unconstrained. Then every sequence $\theta^{(k)}$ with interiority converges.

Proof: Case (a). We adopt the notation (1.4) from [59], where $a_k \xrightarrow{l} b_k \xrightarrow{r} a_{k+1}$ means left and right projections. Matching with (31) gives $A = D$ and $B = M$, $\vartheta^k = a_k$ and $\theta^{(k)} = b_k$, making the results of [59] accessible. Then [59, Thm. 8.1] gives convergence of the $a_k = \vartheta^{(k)}$. Here the lr -angle condition [59, Def. 8.1] follows from definability of L, M and ψ . The lr -three-point inequality follows via [59, Prop. 6.4] from convexity of $\nabla\psi(A) = \nabla\psi(D)$, which holds because $\nabla\psi(D)$ is the intersection of the affine space $L = \{\eta : \eta_y = y\}$ with $\text{im}(\nabla\psi)$. That proves (a).

Case (b). This is a general fact using that under Amari's condition the EM algorithm coincides with alternating Bregman projections between D and M as given above.

Case (c). This uses [59, Cor. 7.4], which gives an even stronger statement using prox-regularity (see also Proposition 6). Case (d) is Theorem 2. Case (e) is Theorem 3.

Case (f). The last part is when the constraint is $\bar{\Theta}$ and the sequence $\theta^{(k)}$ is bounded and together with its accumulation points stays in G . Then we may find a closed bounded convex definable set $M \subset G$ such that $\theta^{(k)}$ alternates between D and M (see [59, Sect. 2.2]). Then both projections are unique, \vec{P}_M because M is convex, and \vec{P}_D because $\nabla\psi(D) = L \cap \text{im}(\nabla\psi)$ as the intersection of an affine subspace with $\text{im}(\nabla\psi)$ is also convex. Convexity of M and $\nabla\psi(D)$ also guarantees that the rl - and lr -three-point inequalities are satisfied (see [59, Prop. 6.4]). Definability of ψ implies definability of Θ and $\bar{\Theta}$, hence of M , but also definability of $\nabla\psi$, hence of $\text{im}(\nabla\psi)$, and since L as an affine subspace is algebraic, we get definability of $\nabla\psi(D)$. Definability of D now follows because D is the image of the definable set $\nabla\psi(D)$ under the definable diffeomorphism $\nabla\psi^*$, see [27]. In consequence, both rl - and lr -angle conditions are satisfied (see [59, Prop. 5.3] for rl and using duality [59, Sect. 5.2] for lr).

Now convergence of the sequence $b_k = \theta^{(k)}$ follows from [59, Thm. 7.1], while convergence of the sequence $a_k = \vartheta^{(k)}$ follows from [59, Thm. 8.1]. In general the sequences converge to a gap (ϑ^*, θ^*) , that is, $\vartheta^* = \vec{P}_D(\theta^*)$, $\theta^* = \vec{P}_M(\vartheta^*)$, possibly with $\vartheta^* \neq \theta^*$. \square

Remark 20. 1) In information geometry [2, 3], \vec{P}_M is called the m -projection, \vec{P}_D the e -projection. m -geodesics, or perpendiculars to M at a left-projected point, are curved in θ -coordinates, while e -geodesics, or perpendiculars to D at a right-projected point, are straight in θ -coordinates, (see [59, 13]). A set M is m -flat if the left-projection onto M is unique, while a set D is e -flat if the right-projection onto D is unique. Using [13], and assuming ψ is 1-coercive, m -flat is equivalent to M convex, while e -flat is equivalent to $\nabla\psi(D)$ convex.

2) Alternatively, Legendreness and strict convexity of $D(x, \cdot)$ also assure uniqueness of \vec{P}_C for C convex (see [11]), so here convex sets are also e -flat. However, this is less useful in the present setting, where it is $\nabla\psi(D)$ which is convex, not D .

3) In the information geometry literature statement (f) has been made repeatedly, but without the hypothesis of definability of ψ . We are aware of a couple of published incorrect proofs. Our own proof requires the KL-condition, and one would of course like to know whether this can be avoided. Note that we can treat the case $D \cap M = \emptyset$.

4) Case (f) can also be derived from Theorem 3.

5) Case (f) may have local minima or convergence to saddle points, which confirms that even this simplest instance of EM is not from the realm of convexity despite $-\log q(y, \cdot)$ and $-\log p(x^+, \cdot)$ being convex. A simple case with two points in $D \cap M$ and a non-zero gap between D and M is given in Example 5.

6) The case $\phi(s) = s^{1/2}$ in the KL-condition gives linear convergence. This occurs for instance in the neighborhood of a point $\theta^\# \in D \cap M$ where D, M intersect transversally. (For the definition of transversal intersection see [59]).

Remark 21. Instead of e - and m -flatness it is preferable to request uniqueness of the projections only for points close enough. That leads to the concept of prox-regularity, or positive reach, which is used in [59]. The significant difference is that positive reach is invariant under duality, i.e., D has positive reach iff $\nabla\psi(D)$ has, and the same for M , while auto-duality fails for e - and m -flatness. A sample result is:

Proposition 6. *Consider the missing data case under the hypotheses of Theorem 2. Suppose the set M has positive left Bregman reach $r^* > 0$ at the accumulation points θ^* of the sequence $\theta^{(k)}$. Let $\vartheta^* = \vec{P}_D(\theta^*)$ and suppose $\min_{\theta \in M} D(\theta, \vartheta^*) < \frac{1}{2}r^{*2}$. Then the sequence $\theta^{(k)}$ converges.*

Proof: We know that the data set sequence $\vartheta^{(k)}$ converges to ϑ^* and $\theta^{(k)} \in \tilde{P}_M(\vartheta^{(k)})$. But $\tilde{P}_M(\vartheta^*) = \theta^\circ$ is singleton due to the hypothesis on left Bregman reach of M . That clearly implies $\theta^{(k+1)} \rightarrow \theta^\circ$. \square

9. MORE GENERAL FAMILIES

Several of the arguments used for exponential families can be extended to a more general setting. We consider the case where $d\mathbb{P}_\theta = p(T(\cdot), \theta)d\mu$, and for the conditional family, $d\mathbb{P}_\theta^{x|y} = k(T(\cdot)|y, \theta)d\mu_y$ for a sufficient statistic $T(x)$. Suppose there is an orthogonal change of coordinates $Q\theta = (\theta', \theta'')$ and a possibly non-linear reduction to a minimal sufficient statistic $T'(x)$ such that the conditional family has the equivalent representation

$$d\mathbb{P}_\theta^{x|y} = d\mathbb{P}_{\theta'}^{x|y} = k'(T'(x)|y, \theta')d\mu_y(x),$$

depending only on θ' . Let us consider the following property extending affine independence of the T_j in the case of exponential families (Definition 2):

(32) In an affine-minimal representation $k'(T'(x)|y, \theta')$ of $\mathbb{P}_\theta^{x|y}$ there exists no $v \neq 0$ with $v \cdot \nabla_{\theta'} \log k'(T'(x)|y, \theta') = 0$ for μ_y -almost all $x \in h^{-1}(y)$.

Proposition 7. *Suppose $\mathbb{P}^{x|y}$ has the above property, and let $k'(T'(x)|y, \theta')d\mu_y(x)$ be an affine-minimal representation of $\mathbb{P}_\theta^{x|y}$. Then $\nabla_{22}^2 \bar{K}_y(\theta' || \theta') \succ 0$.*

Proof: Let $k(\cdot|y, \theta)d\mu_y$ be affine-minimal for the ease of notation. Since $\nabla_{22}^2 K_y(\theta || \theta) = \mathbb{E}_\theta[\nabla_\theta k(\cdot|y, \theta) \nabla_\theta k(\cdot|y, \theta)^T | y] \succeq 0$ by (18), $v \cdot \nabla_{22}^2 K_y(\theta || \theta) v = 0$ implies $\mathbb{E}_\theta[|v \cdot \nabla_\theta k(\cdot|y, \theta)|^2 | y] = 0$, hence $v \cdot \nabla_\theta \log k(x|y, \theta) = 0$ μ_y -a.s., which by minimality implies $v = 0$. Therefore $\nabla_{22}^2 K_y(\theta || \theta) \succ 0$. \square

This means we can get a situation as previously found for the exponential family. There exists an orthogonal $n \times n$ -matrix Q such that

$$Q \nabla_{22}^2 K_y(\theta || \theta) Q^T = \left[\begin{array}{c|c} \nabla_{22}^2 \bar{K}_y(\theta' || \theta') & 0 \\ \hline 0 & 0 \end{array} \right], \quad Q\theta = \begin{bmatrix} \theta' \\ \theta'' \end{bmatrix}, \quad \nabla_{22}^2 \bar{K}_y(\theta' || \theta') \succ 0,$$

where θ' are the accurate parameters remaining after removing the affine dependence in θ . In consequence, we can again prove convergence of the accurate parameter sequence $\theta'^{(k)}$ using the partial convergence theorem.

Theorem 6. *Suppose $-\log q(y, \cdot) + i_M$ satisfies the KL-inequality and $M \subset G$ is bounded. Suppose the conditional family has the minimality property above. Let $\theta^{(k)}$ be generated by the constrained EM algorithm. Then the sequence $\theta^{(k)}$ of accurate parameters converges.*

The last step is to consider the split program (P_∞) . We have the following result, which uses the proof of Proposition 5.

Proposition 8. *Under the hypotheses of Theorem 6. Suppose the conditional expected Fisher information matrix $\mathcal{I}_c(\theta, y)$ of complete data given y is positive definite on G . Then we have $-\nabla_{\theta''\theta''}^2 \log q(y, \theta', \theta'') \succ 0$ for fixed θ' .*

This allows again to upgrade convergence of $\theta^{(k)}$ to convergence of the full parameter sequence, e.g. for $M(\theta')$ convex, when the complete data family has a full rank expected conditional Fisher information of complete data, given y , i.e., $\mathcal{I}_c(\theta, y) \succ 0$.

10. EXAMPLES

In this section we discuss several limiting examples.

Example 1. We consider the missing data case $x = (y, z)$, where, $h : (y, z) \mapsto y$ with y observed and z hidden, and with the special statistic $T(x) = x$. Let $d\mu = m(y, z)d\mu_Y \otimes d\mu_Z$, where $\mu_Y \otimes \mu_Z$ is a product measure on $X = Y \times Z$. We have $p(x, \theta) = e^{\theta_y \cdot y + \theta_z \cdot z - \psi(\theta_y, \theta_z)}$. In the notation of the disintegration, $\nu = \mu_Y$ and $d\mu_y = m(y, z)d(\delta_{\{y\}} \otimes \mu_Z)$ for $y \in Y$, because

$$\begin{aligned} \int_{Y \times Z} f(y, z) d\mu(y, z) &= \int_Y \left[\int_Z f(y, z) m(y, z) d\mu_Z(z) \right] d\mu_Y(y) \\ &= \int_Y \left[\int_{\{y\} \times Z} f(y, z) m(y, z) d(\delta_{\{y\}} \otimes \mu_Z)(y, z) \right] d\mu_Y(y) \\ &= \int_Y \left[\int_{h^{-1}(y)} f(y, z) d\mu_y(y, z) \right] d\nu(y). \end{aligned}$$

That gives

$$\begin{aligned} \psi_y(\theta) &= \log \int_{h^{-1}(y)} e^{\theta_y \cdot y} e^{\theta_z \cdot z} d\mu_y(y, z) \\ &= \log \int_{\{y\} \times Z} e^{\theta_y \cdot y} e^{\theta_z \cdot z} m(y, z) d(\delta_{\{y\}} \otimes \mu_Z)(y, z) \\ &= \theta_y \cdot y + \log \int_Z e^{\theta_z \cdot z} m(y, z) d\mu_Z(z). \end{aligned}$$

Here $T(h^{-1}(y)) = \{y\} \times Z$, so that we need dimension reduction in Proposition 3. We obtain

$$\begin{aligned} (33) \quad \log k(y, z | y, \theta_y, \theta_z) &= \theta_y \cdot y + \theta_z \cdot z - \theta_y \cdot y - \log \int_Z e^{\theta_z \cdot z} m(y, z) d\mu_Z(z) \\ &=: \theta_z \cdot z - \chi_y(\theta_z) =: \log \bar{k}(z | y, \theta_z) \end{aligned}$$

on defining $\chi_y(\theta_z) = \log \int_Z e^{\theta_z \cdot z} m(y, z) d\mu_Z(z)$. Therefore the conditional family is

$$k(y, z | y, \theta_y, \theta_z) d\mu_y(y, z) = \bar{k}(z | y, \theta_z) d\mu_Z(z)$$

and the accurate parameter is θ_z , at least when z is minimal for the conditional family, the spare parameter being θ_y . That means even under definability we can only expect convergence of the θ_z -part of θ . For convergence of the θ_y -part we must rely on the split program (P_∞) .

Example 2. (Continued...) We compare the domains of objective and regularizer in (16). From $q(y, \theta_y, \theta_z) = \int_Z e^{\theta_y \cdot y} e^{\theta_z \cdot z} e^{-\psi(\theta_y, \theta_z)} m(y, z) d\mu_Z(z) = e^{\theta_y \cdot y - \psi(\theta_y, \theta_z)} \int_Z e^{\theta_z \cdot z} m(y, z) d\mu_Z(z)$ follows $\text{dom } q(y, \cdot) = \text{dom } p(x, \cdot) = \text{dom } (\psi) = \Theta = \{(\theta_y, \theta_z) : \int_Z e^{\theta_y \cdot y} e^{\theta_z \cdot z} m(y, z) d\mu_y \otimes \mu_Z < \infty\}$,

while $\text{dom } \psi_y = Y \times \{\theta_z : \int_Z e^{\theta_z \cdot z} m(y, z) d\mu_Z(z) < \infty\} = \text{dom } \bar{K}_y(\theta^{(k)} || \cdot)$ is larger. Nonetheless, Θ and $\text{dom } \psi_y$ have common boundary points.

Example 3. Taken from [2]. Consider an independent normal sample x_1, \dots, x_N with statistic $T(x) = (T_1(x), T_2(x)) = (\sum_{i=1}^N x_i/N, \sum_{i=1}^N x_i^2/N)$, and suppose $y = T_1(x)$ is observed, while $z = T_2(x)$ is hidden. The corresponding exponential family $p(x, \mu, \sigma^2)$ is written as $p(x, \theta)$ with $\theta = (\theta_1, \theta_2) = (N\mu/\sigma^2, -N/2\sigma^2)$, hence $\mu = -\theta_1/2\theta_2$, $\sigma^2 = -N/2\theta_2$, $\Theta = G = \mathbb{R} \times (-\infty, 0)$.

Specializing to $N = 2$ for simplicity, we have $\psi(\theta_1, \theta_2) = -\theta_1^2/4\theta_2 - \log(-\theta_2)$ defined on G , with Legendre transform $\psi^*(\eta_1, \eta_2) = -1 - \log(\eta_2 - \eta_1^2)$ defined on $G^* = \{\eta : \eta_2 > \eta_1^2\}$, gradient, inverse gradient and expectation parameter being

$$\nabla \psi(\theta) = \begin{bmatrix} -\frac{\theta_1}{2\theta_2} \\ \frac{\theta_1^2}{4\theta_2^2} - \frac{1}{\theta_2} \end{bmatrix} =: \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} \quad \theta_1 = -\frac{2\eta_1}{\eta_1^2 - \eta_2}, \theta_2 = \frac{1}{\eta_1^2 - \eta_2}, \quad \nabla \psi^*(\eta) = \begin{bmatrix} -\frac{2\eta_1}{\eta_1^2 - \eta_2} \\ \frac{\eta_1^2}{\eta_1^2 - \eta_2} \end{bmatrix}$$

and in the original coordinates $\eta_1 = \mu$, $\eta_2 = \mu^2 + \sigma^2$. We have $h(x_1, x_2) = y = \frac{x_1 + x_2}{2}$. The family is two-dimensional, but $T(h^{-1}(y)) = (y, \frac{1}{2}(x_1^2 + (2y - x_1)^2))$ is included in the one-dimensional affine space $T_1 = y$ of (T_1, T_2) . Hence we need parameter reduction. The conditional family is

$$\begin{aligned} k(x_1, x_2 | y, \theta) &= C(\sigma) \exp\{-(x_1 - \mu)^2 + (x_2 - \mu)^2 / 2\sigma^2\} / \exp\{-(y - \mu)^2 / \sigma^2\} \\ &= \exp\{-(x_1 - y)^2 / \sigma^2\} / 2\sqrt{\pi}\sigma =: \bar{k}(x_1 | y, \theta_2) \\ &= \exp\left\{-(x_1^2 - 2yx_1) / \sigma^2 - [y^2 / \sigma^2 - \frac{1}{2} \log \sigma^{-2}]\right\} / 2\sqrt{\pi} \\ &= \exp\left\{\theta_2 \cdot (x_1^2 - 2yx_1) - [-y^2\theta_2 - \frac{1}{2} \log(-\theta_2)]\right\} / 2\sqrt{\pi} \end{aligned}$$

the reduced family depending only on $\theta' = \theta_2$, which is the accurate parameter. The incomplete data family is obtained as follows: Since $\mathbb{E}_\theta[(x_1 + x_2)/2] = \mu$ and $\mathbb{V}_\theta[(x_1 + x_2)/2] = \sigma^2/2$, we have $q(y, \theta) \sim N(\mu, \sigma^2/2)$.

Now we consider the M step. We discuss two cases, constrained and unconstrained. In the first scenario a constraint is introduced in the form of a curved exponential family, namely $\mu^2 = \sigma^2$, so that the model family is $N(\mu, \mu^2)$. In natural parameters this is $M = \{\theta : \theta_1^2 = -4\theta_2\}$. Here Theorem 2 assures convergence of the accurate parameter sequence $\theta_2^{(k)}$. But the constraint gives θ_1^2 as a function of θ_2 , so $\theta_1^{(k)}$ converges, too.

In the unconstrained case the parameter $\theta_2^{(k)}$ converges, hence so does $\sigma^{(k)2}$, while the parameter θ_1 is free. We have to consider the M step program (P_k) , which is $\min_{\theta_1} -\log q(y, \theta_1, \theta_2^{(k)})$. Since $-\log q(y, \theta_1, \theta_2^{(k)}) = (y - \mu)^2 / \sigma^{(k)2} + \log C(\sigma^{(k)})$, the solution is always $\mu = y = \mu^{(k)}$, which implies $\theta_1^{(k)} = 2y / \sigma^{(k)2} = -2y\theta_2^{(k)}$, which again converges. This is interesting, as in the first place the available information does not seem sufficient to estimate μ and σ^2 .

Example 4. (Continued...) As observed in [2], in this example a difference between the EM algorithm and the *em*-algorithm occurs. In our present terminology this is due to the fact that $\mathbb{E}_\theta[(x_1^2 + x_2^2)/2] = \mu^2 + \sigma^2 = y^2 + \sigma^2$ and $\mathbb{E}_\theta[(x_1^2 + x_2^2)/2 | y] = y^2 + \sigma^2/2$ are different.

This discrepancy can be easily remedied by letting $T_2(x) = \sum_{i=1}^N (x_i - \bar{x})^2 / (N - 1)$, as then T_1, T_2 are independent, so that Amari's condition is satisfied.

Example 5. Consider the Kullback-Leibler distance in \mathbb{R}_+^2 given as $K(a || b) = \sum_{i=1}^2 a_i \log \frac{a_i}{b_i} - a_i + b_i$, which is a special Bregman distance induced by $\psi(x) = \sum_{i=1}^2 x_i \log x_i - x_i$. Choose two points $p, q \in (-\infty, 0)^2$ and compute $a = \exp(p)$, $b = \exp(q)$. Let $p_t = tp + (1 - t)q$, $t \in [0, 1]$ be the points on the segment $[p, q]$, then $a_t = \exp(p_t) = \exp(tp) \exp((1 - t)q)$ forms a curve in $(0, 1)^2$, which in general is not straight. Let $A = \{a_t : t \in [0, 1]\}$. Then $\nabla \psi(A) = \log(A) = [p, q]$ the segment, hence A is *e*-flat. In other words, Kullback-Leibler right projections on A are unique. Now let $B = [a, b]$ be the segment joining a, b . Then B is convex, hence Kullback-Leibler left projections on B are unique, and B is *m*-flat. We

have $A \cap B = \{a, b\}$ in the case where A is curved. This means Bregman projections \vec{P}_A , \vec{P}_B are unique, and $\vec{P}_A \circ \vec{P}_B$ has two points of attraction. However, starting with a point a_t somewhere in between, one can see iterates either go left, or go right toward a or b . Except a point pair a', b' which satisfies $a' = \vec{P}_A(b')$, $b' = \vec{P}_B(a')$, building a gap or fixed point pair. If started at a' , the alternating projection method will remain at the fixed point pair.

Example 6. (Failure of convergence for curved family). We consider independent gaussian random variables X_1, X_2, X_3 with $\mathbb{E}(X_i) = \mu_i$ and $\mathbb{V}(X_i) = 1$. We want to estimate $\mu = (\mu_1, \mu_2, \mu_3)$ based on an observation y of $Y = X_3$, where it is assumed that $\mu_3 = f(\mu_1, \mu_2)$ with a known f . Concerning the statistic that means $x = (x_1, x_2, x_3) = (z_1, z_2, y)$ with y observed and z hidden. This gives the model set $M = \{\mu : \mu_3 = f(\mu_1, \mu_2)\}$, hence the family is curved. Suppose $y = 0$. Then the E step $x^+ = \mathbb{E}(X|Y = y, \mu)$ yields $x_1^+ = \mu_1, x_2^+ = \mu_2, x_3^+ = 0$. The M step is $\mu^+ \in \operatorname{argmin}_{\mu \in M} -\log p(x^+, \mu)$, where

$$p(x^+, \mu) = C \exp\{-\frac{1}{2}(x_1^+ - \mu_1)^2 - \frac{1}{2}(x_2^+ - \mu_2)^2 - \frac{1}{2}(x_3^+ - \mu_3)^2\}.$$

Hence the E step is orthogonal projection of $\mu \in \mathbb{R}^3$ onto the data set $D = \{x_3 = 0\}$. The M step is orthogonal projection of $x^+ = (x_1^+, x_2^+, 0)$ onto the model set $M = \operatorname{graph}(f)$. The result of this projection is $\mu^+ \in M$, and then the procedure is repeated. We therefore have a case, where EM and *em*-algorithm coincide.

Altogether the method is now the alternating projection method between the sets $A = \{(x_1, x_2, x_3) : x_3 = 0\}$ and $B = \{(x_1, x_2, x_3) : x_3 = f(x_1, x_2)\}$, and this can readily be extended to $x \in \mathbb{R}^n$, where $A = \{(x, 0) : x \in \mathbb{R}^n\}$ and $B = \{(x, f(x)) : x \in \mathbb{R}^n\}$ the graph of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$. Assume $f(x) \geq 0$. Then

$$(34) \quad P_A(x_k, f(x_k)) = (x_k, 0), \quad (x_k, f(x_k)) \in P_B(x_{k-1}, 0) \text{ iff } x_{k-1} = x_k + f(x_k)\nabla f(x_k).$$

Infinitesimally, this method follows steepest ascent backwards.

Going back to $n = 2$, we let $B = M$ be the graph of the mexican hat function [1] on $x_1^2 + x_2^2 \leq 1$. Similar to the argument given for steepest descent with infinitesimal steps in [1], AP with infinitesimal steps will also follow the valley of the hat downward, endlessly circling around and approaching the boundary curve $x_1^2 + x_2^2 = 1$, where $f = 0$. What is amiss for convergence is the KL-property of M , which is not definable. For a picture see [1].

Example 7. (PPM and EM). Take again the situation $A = \{(x, 0) : x \in \mathbb{R}^n\}$, $B = \{(x, f(x)) : x \in \mathbb{R}^n\}$, where $f \geq 0$. Consider the proximal point step

$$(35) \quad x_k \in \operatorname{argmin} \frac{1}{2}f(x)^2 + \frac{1}{2}\|x - x_{k-1}\|^2.$$

The necessary optimality condition is $0 = f(x_k)\nabla f(x_k) + x_k - x_{k-1}$, which is the alternating projection step (34) above. This means, (35) must be the scheme (4), respectively, its realization in Proposition 1, for our example above. Choosing f such that AP fails to converge, we also produce an example, where the non-convex proximal point method (with fixed $\lambda_k = 1$) fails to converge, now with objective $\frac{1}{2}f(x)^2$. This construction makes every instance of PPM with an objective bounded below a special instance of EM.

REFERENCES

- [1] Absil, P.A., Mahony, R., Andrews, B.: Convergence of the iterates of descent methods for analytic cost functions. *SIAM J. Optim.*, 16(2), 531-547 (2005).
- [2] Sh.-I. Amari. Information geometry of the EM- and em-algorithms for neural networks. *Neural Networks*, 8(9):1995,1379-1408.
- [3] Sh.-I. Amari. *Information Geometry and its Applications*. Springer Applied Math. Sci. 194, 2016.
- [4] F.J. Aragón Artacho, A.L. Donchev, M.H. Geoffrey. Convergence of the proximal point method for metrically regular mappings. *ESAIM: Proceedings* 17:2007,1-8.
- [5] H. Attouch, J. Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Math. Programming*, 116(1-2, Ser. B), 5-16 (2009).

- [6] H. Attouch, J. Bolte, P. Redont, A. Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):2010, 438–457.
- [7] H. Attouch, J. Bolte, B.F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. *Math. Programming* 137:2013,91-129.
- [8] O.E. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*, Wiley 1978.
- [9] H.H. Bauschke, J.M. Borwein. Legendre functions and the method of random Bregman projections. *J. Convex Anal.* 4(1):1997,27-67.
- [10] H.H. Bauschke, M.N. Dao, S.B. Lindstrom. Regularizing with Bregman-Moreau envelopes. *SIAM J. Optim.* 28(4):2018,3208-3228.
- [11] H.H. Bauschke, D. Noll. The method of forward projections. *J. Nonlin. Convex Anal.* 3:2002,191-205.
- [12] H.H. Bauschke, D. Noll, A. Celler, J.M. Borwein. An EM algorithm for dynamic SPECT. *IEEE Transactions on Medical Imaging*, 18(3):1999,252-261.
- [13] H.H. Bauschke, X. Wang, J. Ye, X. Yuan. Bregman distances and Chebyshev sets. *J. Approx. Theory* 159:2009,3-25.
- [14] P.M. Bentler, J. Liang, M.-L. Tang, K.-H. Yuan. Constrained maximum likelihood estimation for two-level mean and covariance structure models. *Educ. Psychol. Meas.*, 71(2):2011, 325-345.
- [15] E. Bierstone, P. Milman. Semianalytic and subanalytic sets. *IHES Publ. Math.*, 67:1988, 5-42.
- [16] J. Bolte, A. Daniilidis, O. Ley, L. Mazet. Characterizations of Łojasiewicz inequalities: subgradient flows, talweg, convexity. *Trans. Amer. Math. Soc.* 362(6):2009,3319-3363.
- [17] J. Bolte, A. Daniilidis, A.S. Lewis, M. Shiota. Clarke subgradients of stratifiable functions. *SIAM J. Optim.* 18:2007,556-572.
- [18] L. D. Brown. Fundamentals of statistical exponential families with applications in statistical decision theory. *IMS Lect. Notes Monog., Ser.* 9:1986, , 1986.
- [19] R. Caprio, A.M. Johansen. Fast convergence of the Expectation Maximization algorithm under a logarithmic Sobolev inequality. *Biometrika* 103:2024,1-18.
- [20] G. Chen, M. Teboulle. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIOPT* 3(3):1993,538–543.
- [21] S. Chrétien, A.O. Hero. Kullback proximal algorithms for maximum likelihood estimation. *IEEE Trans. Inform. Theory* 46(5):2000,1800–1810.
- [22] S. Chrétien, A.O. Hero. On EM algorithms and their proximal generalizations. *ESAIM: PS* 12:2008,308–326.
- [23] S. Chrétien, A.O. Hero. Kullback proximal algorithms for maximum likelihood estimation. [arXiv:1201.5907v1 \[stat.CO\]](https://arxiv.org/abs/1201.5907v1) 27 Jan 2012
- [24] S. Chrétien, A. O. Hero. Generalized proximal point algorithms, *Technical Report, The University of Michigan*, 1998.
- [25] S. Chrétien, A.O. Hero. Acceleration of the EM algorithm via proximal point iterations. *Proceedings of the 1998 IEEE International Symposium on Information Theory*, MIT, Cambridge,444.
- [26] F. Clarke. *Optimization and Nonsmooth Analysis*. John Wiley and Sons, New York (1983)
- [27] M. Coste. An introduction to o-minimal geometry. *RAAD Notes, Institut de Recherche de Rennes*, 1999.
- [28] I. Csiszár. I-divergence geometry of probability distributions and minimization problems, *Ann. Probability*, 3:1975,146-158.
- [29] I. Csiszár, G. Tusnády. Information geometry and alternating minimization procedures. *Statistics and Decisions*, 1(1):1984,205-237.
- [30] A.P. Dempster, N.M. Laird, D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Stat. Soc. B* 39:1977,1-38.
- [31] L. van den Dries. *Tame Topology and O-minimal Structures*. Camb. Univ. Press, 1998.
- [32] L. van den Dries, C. Miller. Geometric categories and o-minimal structures. *Duke Math. J.* 85:1996,487-540.
- [33] L. van den Dries, C. Miller. On the real exponential field with restricted analytic functions, *Isr. J. Math.* 85, No. 1-3, 19-56 (1994).
- [34] B. Flury, A. Zoppé. Exercises in EM. *American Statistician*, 54:2000,207–209.
- [35] O. Guler. On the convergence of the proximal point algorithm for convex minimization. *SICON* 29:1991,403–419.
- [36] R.J. Hathaway. A constrained em algorithm for univariate normal mixtures. *J. Stat. Comp. Sim.* 23:1986,211-230.
- [37] J. Hoffmann-Jørgensen. Existence of conditional probabilities. *Math. Scand.* 28:1971,257-264.
- [38] A.N. Iusem. A short convergence proof of the EM algorithm for a specific Poisson model. *Brazilian Journal of Probability and Statistics* 6(1): 1992, 57-67.

- [39] A. Kagan, Z. Landsman. Relation between the covariance and Fisher information matrices. *Statistics & Probability Letters* 42:1999,7-13.
- [40] A. Kaplan, R. Tichatschke. Proximal point method and nonconvex optimization. *J. Global Opt.* 13:1998,389-406.
- [41] D.K. Kim, J.M.G. Taylor. The restricted EM algorithm for maximum likelihood estimation under linear restrictions on the parameters. *J. Amer. Stat. Assoc.* 90:1995,708-716.
- [42] M.E. Khan, P. Baqué, F. Fleuret, P. Fua. Kullback-Leibler proximal variational inference. *Preprint 2025*.
- [43] D. Klatte. Upper Lipschitz behavior of solutions to perturbed $C^{1,1}$ programs. *Math. Progr. B* 88:2000,285-311.
- [44] F. Kunstner, R. Kumar, M. Schmidt. Homeomorphic-invariance of EM: Non-asymptotic convergence in KL divergence for exponential families with mirror descent. *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. arXiv:2011.01170*
- [45] K. Kurdyka. On gradients of functions definable in o-minimal structures. *Ann. Inst. Fourier*, 48:1998,769-783.
- [46] E. L. Lehmann, G. Casella. *Theory of point estimation*. Springer Texts in Statistics, 2nd ed., 1998.
- [47] M. Levine, D. Richards, J. Su. Non-steepness and maximum likelihood estimation properties of the truncated multivariate normal distributions. *arXiv:2303.10287v3 [math.ST]*
- [48] Ch. Lui, D.B. Rubin. The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika* 81(4):1994,633-648.
- [49] F.J. Luque. Asymptotic convergence analysis of the proximal point algorithm. *SIAM J. Contr. Optim.* 22,1984,277-293.
- [50] J. Maeght, S.P. Boyd, D. Noll. Dynamic emission tomography - regularization and inversion. *Can. Math. Soc. Conf. Proc.* 27, 2000, 211-234.
- [51] B. Martinet. Regularisation d'inéquations variationnelles par approximations successives. *Rev. Française d'Autom. Inform. Rech. Opér.* 4:1970,154-159.
- [52] B. Martinet. Determination approchée d'un point fixe d'une application pseudocontractante. *C. R. Acad. Sci. Paris*, 274:1972,163-165.
- [53] G.J. McLachlan, T. Krishnan. *The EM algorithm and extensions*. Wiley Series in Prob. and Stat. 2nd ed. 2008.
- [54] C. Miller. Exponentiation is hard to avoid. *Proc. Amer. Math. Soc* 122(1):1994,257-259.
- [55] B.S. Mordukhovich. *Variational Analysis and Generalized Differentiation I: Basic Theory, II: Applications*. Springer, New York, 2006.
- [56] D. Nettleton. Convergence properties of the EM algorithm in constrained parameter spaces. *Can. J. Stat.* 27(3):1999,639-648.
- [57] D. Noll. Convergence of nonsmooth descent methods using the Kurdyka-Łojasiewicz inequality. *J. Optim. Theory Appl.* 160(2):2014,553-572.
- [58] D. Noll. Alternating projections with applications to Gerchberg-Saxton error reduction. *Set-Valued and Variational Analysis*, 29:2021,771-802.
- [59] D. Noll. Alternating Bregman projections and convergence of the EM algorithm. *J. Pure Appl. Funct. Anal.* 10(4):2025, 979-1021.
- [60] D. Noll, A. Rondepierre. On local convergence of the method of alternating projections. *Foundations of Computational Mathematics*, vol. 16, no. 2, 2016, pp. 425-455.
- [61] T. Orchard, M.A. Woodbury. A missing information principle: Theory and applications. *Proc. 6th Berkeley Symp. Math. Stat. Prob.* vol. 1, Berkeley, CA: University of California Press, pp. 697-715.
- [62] A. Padgett, P. Speissegger. Definability of complex functions in o-minimal structures. *arXiv:2506.15119.v1 [math.LO] 18 Jun 2025*.
- [63] R.F. Phillips. A constrained maximum-likelihood approach to estimating switching regressions. *Journal of Econometrics*, 48(1-2):1991,241-262
- [64] T. Pennanen. Local convergence of the proximal point algorithm and multiplier methods without monotonicity. *Math. Oper. Res.* 27(1):2002,170-191.
- [65] S.M. Robinson. Generalized equations and their solutions. Part II: Applications to nonlinear programming. *Math. Program. Study* 19:1982, 200-221
- [66] R.T. Rockafellar. *Convex Analysis*. Princeton University Press, NJ, 1970.
- [67] R.T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM J. Control and Optimization* 14:1976,877-898.
- [68] R.T. Rockafellar. Advances in convergence and scope of the proximal point algorithm. *Journal of Nonlinear and Convex Analysis*, 22(11):2021, 2347-2374.
- [69] R.T. Rockafellar, R.J.B. Wets. *Variational Analysis*. Grundlehren der mathematischen Wissenschaften, Springer 317:2009.

- [70] R. Schoenberg. Constrained maximum likelihood. *Computational Economics* 10(3):1997,251-266.
- [71] Ning-Zhong Shi, Shu-Rong Zheng, and Jianhua Guoa. The restricted EM algorithm under inequality restrictions on the parameters. *Journal of Multivariate Analysis* 92 (2005) 53-76.
- [72] J.E. Spingarn. Submonotone mappings and the proximal point algorithm. *Num. Funct. Anal. and Optim.* 4(2):1982,123-150.
- [73] R. Sundberg. Maximum likelihood theory for incomplete data from an exponential family. *Scand. J. Stat.* 1:1974,49-58.
- [74] M. Shiota. *Geometry of Subanalytic and Semialgebraic Sets*. Birkhäuser Verlag 1997.
- [75] P. Tseng. An analysis of the EM algorithm and entropy-like proximal point methods. *Math. Oper. Res.* 29(1):2004,27-44.
- [76] F. Vaida. Parameter convergence for EM and MM algorithms. *Statistica Sinica* 15:2005,831-840.
- [77] A. Wilkie. Model completeness results for expansions of the ordered field of real numbers by restricted Pfaffian functions and the exponential function. *J. Amer. Math. Soc.* 9:1996,1051-1094.
- [78] C.F.J. Wu. On the convergence properties of the EM algorithm. *Annals of Statistics* 11:1983,95-103.

INSTITUT DE MATHÉMATIQUES, UNIVERSITÉ DE TOULOUSE, FRANCE
Email address: dominikus.noll@math.univ-toulouse.fr