# COMPARISON OF THE EM ALGORITHM AND ALTERNATIVES

THERESA SPRINGER AND KARSTEN URBAN

Abstract. The Expectation-Maximization (EM) algorithm is widely used also in industry for parameter estimation within a Maximum Likelihood (ML) framework in case of missing data. However, to the best of our knowledge, precise statements concerning the connections of the EM algorithm to other ascent methods are only available for specific probability distributions. In this paper, we analyze the connection of the EM algorithm to other ascent methods as well as the convergence rates of the EM algorithm in general and apply this to the PMHT model. We compare the EM with other known iterative schemes such as gradient and Newton-type methods. It is shown that EM reaches Newton-convergence in case of well-separated objects and a Newton-EM combination turns out to be robust and efficient even in cases of closely-spaced targets.

## 1. Introduction

Maximum Likelihood (ML) estimation is a wide-spread tool in parametric statistics. In many cases of practical interest, an analytical computation of an ML estimator is not possible so that one has to resort to numerical approximation methods. These methods need to be fast (efficient), robust and accurate, where efficiency is particularly important for large data sets, robustness is a key issue in case of perturbed data and accuracy is crucial when decisions need to be based upon numerical simulations. Hence, the choice of a numerical method for ML estimation is an important issue for a problem at hand. The intention of this paper is to analyze some widely used computational schemes in the particular case of missing data with respect to the three mentioned criteria and to establish similarities and differences that can guide the choice.

Given a probability space $(\Omega, \mathfrak{A}, \mathsf{P})$ and an (observable) random variable $Y : \Omega \to \mathbb{R}^n$, $n \in \mathbb{N}$. The associated probability density function $g_\phi^Y : \mathbb{R}^n \to \mathbb{R}_0^+ := [0, \infty)$ is assumed to be parameter-dependent with some parameter $\phi \in \Phi \subseteq \mathbb{R}^P$, $P \in \mathbb{N}$. The Likelihood function $\mathcal{L} : \Phi \to \mathbb{R}_0^+$ for an observed realization $y = Y(\omega)$ of a (fixed) event $\omega \in \Omega$ is then defined as $\mathcal{L}(\phi) := g_\phi^Y(y)$, $\phi \in \Phi$. The ML estimator is usually computed for the log-Likelihood function $L := \log \mathcal{L}$ and reads

$$\hat{\phi}^{\mathrm{ML}} = \underset{\phi\in\Phi}{\arg\sup}\; L(\phi). \tag{1.1}$$

As long as $L$ is explicitly known, the computation of $\hat{\phi}^{\mathrm{ML}}$ reduces to an optimization problem which can be solved numerically by any scheme known from numerical optimization.

*Date*: November 15, 2012.

*Key words and phrases.* EM algorithm, tracking, convergence, Gradient method, Newton's method, MSC: 47J25, 62F99, 62P30, 65C60.

We are particularly interested in the context of missing data, where the particular form of $L$ might not be known (since the full data $x = X(\omega)$ may not be observable) or at least not be efficiently computable in general. In this framework, the Expectation-Maximization (EM) algorithm is a widely used scheme. The main reason is that EM does neither require the analytic expression of the log-Likelihood function nor of the gradient, the function $L$ does not even need to be differentiable. This makes EM a good "black box" method.

However, in many applications, there is some knowledge on $L$, its gradient or smoothness properties, even though it might be cumbersome to retrieve this information at a first glance. In terms of the properties efficiency, robustness and accuracy, however, it might pay off to use this information so that numerical methods other then EM (and possibly superior) could in principle be used.

This is the motivation for a theoretical analysis of the relationship of EM to other numerical optimization schemes. Similar comparisons of the performance of the EM algorithm with unconstrained optimization methods have been performed e.g. in [20, 21, 22, 28]. We will detail the relations of our results to existing ones in literature later on in Section 3.2.

The remainder of this paper is organized as follows. In Section 2 we describe the ML framework of missing data, collect some general assumptions and introduce the EM algorithm. The connections of the EM algorithm to other ascent methods and a linearized fixpoint scheme for the Euler-Lagrange equation are in general derived in Section 3. Subsequently, in Section 4, we examine the convergence of iterative schemes and discuss the implications for the EM algorithm. This is then applied to the PMHT model analyzing a specific tracking scenario in Section 5. The theoretical results are supplemented by numerical experiments and a discussion on their relevance for practical applications, at the end of that section.

## 2. Preliminaries

In this section, we collect some basic notation and facts that will be needed throughout the paper.

### 2.1. **ML estimation for missing data.**

Let $(\Omega, \mathfrak{A}, \mathsf{P})$ again be a probability space, i.e., a sampling space. Given two random variables $X : \Omega \to \mathbb{R}^m$ and $Y : \Omega \to \mathbb{R}^n$, $m \geq n$, let $\mathbb{D} := \{X(\omega) : \omega \in \Omega\} \subset \mathbb{R}^m$, $\mathbb{O} := \{Y(\omega) : \omega \in \Omega\} \subset \mathbb{R}^n$ denote the corresponding data and observation space, respectively. It seems natural to assume that both $\mathbb{D}$ and $\mathbb{O}$ are bounded (since unbounded observations are usually truncated to a finite computational domain anyway). We assume that the density functions are given in terms of a parametric model with parameter space $\Phi \subseteq \mathbb{R}^P$. Denote the probability density functions by $f_\phi^X$ and $g_\phi^Y$, respectively.

We may think of $x \in \mathbb{D}$ as the complete data and $y \in \mathbb{O}$ as the observed or incomplete data. This is in special cases indicated by writing $\mathbb{D} = \mathbb{O} \times \mathbb{M}$, $x = (y, z)$, where $z \in \mathbb{M}$ is interpreted as missing data. This can also be expressed by a "many-to-one"-mapping $\ell : \mathbb{D} \to \mathbb{O}$, $\ell(x) := y$. More general, given the observed data $y$, the unknown complete data is only known to be in the subset of $\mathbb{D}$ given by $\{x \in \mathbb{D} : y = \ell(x)\} =: \ell^{-1}(y)$, the preimage of $\ell$. In particular ,$\ell^{-1}(y)$ is assumed to be uniformly bounded for any observation $y \in \mathbb{O}$.

The relation between observed and complete data (in terms of their probability density functions -PDF- $f_\phi^X$ and $g_\phi^Y$, respectively) can be written as

$$g_\phi^Y(y) = \int_{\ell^{-1}(y)} f_\phi^X(x)\,dx. \tag{2.1}$$

Further, let $h_\phi^{X|Y}$ denote the conditional density of the complete data $x$ given the observation $y$, i.e.,

$$h_\phi^{X|Y}(x|y) := \frac{f_\phi^X(x)}{g_\phi^Y(y)}, \qquad x \in \ell^{-1}(y), \tag{2.2}$$

for $g_\phi^Y(y) \neq 0$. Note, that by definition we have $\int_{\ell^{-1}(y)} h_\phi^{X|Y}(x|y)\,dx = 1$ and this normalization will be frequently used in the sequel. One approach to "fill in" the missing data is to use an estimate $\hat{\phi}$ for the unknown parameter $\phi$ given the observation $y$ and then to use $\hat{\phi}$ to retrieve some statistics for the unknown full data $x$. One option for $\hat{\phi}$ is a standard ML estimate, i.e.,

$$\hat{\phi}^{\mathrm{ML}} = \arg\sup_{\phi\in\Phi}\ g_\phi^Y(y). \tag{2.3}$$

Taking the logarithm on both sides of (2.2) and denoting

$$\ell_{\mathrm{full}}(x;\phi) := \log f_\phi^X(x), \quad \ell_{\mathrm{obs}}(y;\phi) := \log g_\phi^Y(y), \quad \ell_{\mathrm{miss}}(x|y;\phi) := \log h_\phi^{X|Y}(x|y),$$

equation (2.2) reads after reordering

$$L(\phi) := L_{\mathrm{md}}(\phi) := \ell_{\mathrm{obs}}(y;\phi) = \ell_{\mathrm{full}}(x;\phi) - \ell_{\mathrm{miss}}(x|y;\phi). \tag{2.4}$$

This decomposition has a useful interpretation. If we aim at maximizing $g_\phi^Y(y)$ (or its log $\ell_{\mathrm{obs}}(y;\phi)$, i.e., the log-Likelihood $L_{\mathrm{md}}$ in the missing data-case) w.r.t. $\phi$, we can hope that $\ell_{\mathrm{full}}(x;\phi)$ is relatively easy to maximize w.r.t. $\phi$ and can be used as approximation for $L(\phi)$. This is a framework particularly appropriate for the EM algorithm to be described in Section 2.3 below.

### 2.2. Some general assumptions.

Before we continue, let us collect some general assumptions. Since we are concerned with observed data, it is by no means a restriction to assume that $X : \Omega \to \mathbb{D}$ and $Y : \Omega \to \mathbb{O}$, where $\mathbb{D} \subset \mathbb{R}^m$, $\mathbb{O} \subset \mathbb{R}^n$ are bounded and closed domains (one may think of an observation window). Consequently, we have that $f_\phi^X : \mathbb{D} \to \mathbb{R}_0^+$. For later convenience, let us specify our general assumptions.

**Assumption A.** Let $f_\phi^X : \mathbb{D} \to \mathbb{R}_0^+$, $\mathbb{D} \subset \mathbb{R}^m$ compact, be a probability density function such that

(A1) $f_\phi^X(x) > 0$ for all $x \in \mathbb{D}$;
(A2) $f_\phi^X(x) \in C^1(\Phi)$ for all $x \in \mathbb{D}$;
(A3) $\frac{\partial}{\partial\phi} f_\phi^X \in L_\infty(\mathbb{D})$.

**Remark 2.1.** *Since $\mathbb{D}$ is assumed to be compact, (A1) implies the existence of a $\mathbb{R} \ni f_\phi^- > 0$ such that*

$$f_\phi^X(x) \geq f_\phi^- > 0, \qquad x \in \mathbb{D}.$$

Obviously, Assumption A is a general assumption on the model for the full data. We will also pose the following assumption on the observations.

**Assumption B.** For any given observation $y = Y(\omega)$, $\omega \in \Omega$, let $\ell^{-1}(y) \subset \mathbb{D} \subset \mathbb{R}^n$ be a set of positive measure.

**Remark 2.2.** *Assumption B implies that* $g_\phi^Y(y) > 0$ *for all* $y \in \mathbb{O}$.

Assumption B excludes such observations that cannot be associated to data in a meaningful way. Since any data processing includes a pre-selection of meaningful data, this is no restriction at all in practice. Sometimes, we need more regularity, which is reflected by the following assumption.

**Assumption C.** In addition to Assumptions A and B assume that $f_\phi^X(x) \in C^2(\Phi)$ for all $x \in \mathbb{D}$ and $\frac{\partial^2}{\partial\phi^2} f_\phi^X \in L_\infty(\mathbb{D})$.

### 2.3. **The EM algorithm.**

The EM algorithm was initially introduced in [5] and is by now a widely used black box method for ML estimation in case of missing data. The main idea to obtain an iterative procedure for approximating the ML estimate in (2.3) is to replace $L_{\mathrm{md}}(\phi) = \ell_{\mathrm{obs}}(y;\phi)$ by successive maximizations of the conditional expectation $Q(\phi, \phi^{(m)})$ of the log-Likelihood function $\ell_{\mathrm{full}}(x;\phi)$ for the complete data given the observation $y$ and the current parameter value $\phi^{(m)}$, i.e.,

$$Q(\phi,\psi) := \mathbb{E}\Big[\ell_{\mathrm{full}}(\cdot;\phi)\big|y,\psi\Big] = \int_{\ell^{-1}(y)} \big(\log\, f_\phi^X(x)\big)\, h_\psi^{X|Y}(x|y)\, dx.^{\mathrm{a}} \tag{2.5}$$

Since the observation $y$ is assumed to be given and fixed, we do not explicitly denote the dependence of $Q$ w.r.t. $y$. Then, $Q : \Phi \times \Phi \to \mathbb{R}$. With these preparations, the scheme consists of two steps:

Given an initial guess $\phi^{(0)} \in \Phi$, for $k = 0, 1, 2, \dots$ do

**(E)** Compute $Q(\phi, \phi^{(k)})$;

**(M)** Update $\phi^{(k)}$ by computing a maximizer $\phi^{(k+1)}$ of $Q(\phi, \phi^{(k)})$ w.r.t. $\phi$, i.e.,

$$\phi^{(k+1)} := \arg\max_{\phi\in\Phi} Q(\phi, \phi^{(k)}). \tag{2.6}$$

In order to shorten notation, we abbreviate for any $F : \Phi\times\Phi \to \mathbb{R}$ with $F(\cdot,\psi) \in C^1(\Phi)$, $\psi \in \Phi$,

$$D^{10}F(\tilde\phi,\psi) := \frac{\partial}{\partial\phi} F(\phi,\psi)\big|_{\phi=\tilde\phi}, \qquad \phi,\psi,\tilde\phi \in \Phi,$$

with obvious extensions to other partial derivatives. Note, that we interpret $D^{10}F \in \mathbb{R}^P$ as well as $\frac{\partial}{\partial\phi}F \in \mathbb{R}^P$ as a column vector. Moreover, we set

$$C^{10}(\Phi) := \{F \in C(\Phi\times\Phi) :\, D^{10}F \in C(\Phi\times\Phi)\},$$

again with (almost) obvious extensions to other partial derivatives. As an example, we just note that $D^{11}F(\tilde\phi,\tilde\psi) := \frac{\partial}{\partial\psi}\frac{\partial}{\partial\phi}F(\phi,\psi)|_{\phi=\tilde\phi,\psi=\tilde\psi}$ and $C^{11}(\Phi)$ accordingly.

**Lemma 2.3.** *Let Assumptions A and B hold. If* $f_\phi^X(x) \in C^k(\Phi)$, $k \in \mathbb{N}$, *for all* $x \in \mathbb{D}$, *then* $Q \in C^{kk}(\Phi)$.

[a] In the EM literature, often the notation $Q(\phi|\psi)$ is used. We avoid the vertical bar here in order to avoid possible misunderstandings with conditional expectations, densities etc.

*Proof.* First, note that

$$Q(\phi,\psi) = \frac{1}{g_\psi^Y(y)} \int_{\ell^{-1}(y)} f_\psi^X(x)\, \log\, f_\phi^X(x)\, dx.$$

By assumption, we have $g_\psi^Y(y) \neq 0$ and $f_\phi^X(x) > 0$, by (2.1) and Remark A.2 we have $g_\psi^Y(y) \in C^k(\Phi)$, so that $Q$ is a combination of $C^k$-functions w.r.t. $\phi$ and $\psi$. □

**Remark 2.4.** *For later reference, let us remark that*

$$\nabla L_{\mathrm{md}}(\phi) = \nabla_\phi \ell_{\mathrm{obs}}(y;\phi) = D^{10}Q(\phi,\phi), \tag{2.7}$$

*provided that Assumptions A and B hold.*

*Proof.* By definition, we have that

$$D^{10}Q(\phi,\psi) = \frac{\partial}{\partial\phi}Q(\phi,\psi) = \frac{\partial}{\partial\phi}\int_{\ell^{-1}(y)} (\log\, f_\phi^X(x))\, h_\psi^{X|Y}(x|y)\, dx.$$

Due to Lemma A.6, we can interchange differentiation and integration. Moreover, since $f_\phi^X(x) > 0$, $x \in \mathbb{D}$, we get by (2.1) and (2.2)

$$\begin{aligned} D^{10}Q(\phi,\phi) &= \int_{\ell^{-1}(y)} \frac{\partial}{\partial\phi} f_\phi^X(x)(f_\phi^X(x))^{-1}\, h_\phi^{X|Y}(x|y)\, dx \\ &= \int_{\ell^{-1}(y)} \Big(\frac{\partial}{\partial\phi} f_\phi^X(x)\Big)(g_\phi^Y(y))^{-1} dx = (g_\phi^Y(y))^{-1}\int_{\ell^{-1}(y)} \frac{\partial}{\partial\phi} f_\phi^X(x) dx \\ &= (g_\phi^Y(y))^{-1}\frac{\partial}{\partial\phi} g_\phi^Y(y) = \frac{\partial}{\partial\phi}\log\, g_\phi^Y(y) = \nabla_\phi \ell_{\mathrm{obs}}(y;\phi), \end{aligned}$$

where we have used Remark A.2 in the last row. This proves (2.7). □

Before we continue, let us collect one more well-known and useful fact. Defining

$$\begin{aligned} H(\phi,\psi) &:= \mathbb{E}(\log\, h_\phi^{X|Y}(\cdot)|y,\psi) = \mathbb{E}(\ell_{\mathrm{miss}}(\cdot|y;\phi)|y,\psi) \\ &= \int_{\ell^{-1}(y)} (\log\, h_\phi^{X|Y}(x|y))\, h_\psi^{X|Y}(x|y)\, dx, \end{aligned} \tag{2.8}$$

we get (recall (2.4))

$$L(\phi) = Q(\phi,\psi) - H(\phi,\psi), \quad \psi \in \Phi. \tag{2.9}$$

In fact, using the respective definitions and the integral normalization of the PDF $h_\phi^{X|Y}$, we get by (2.2)

$$\begin{aligned} Q(\phi,\psi) - H(\phi,\psi) &= \int_{\ell^{-1}(y)} (\log\, f_\phi^X(x) - \log\, h_\phi^{X|Y}(x|y))\, h_\psi^{X|Y}(x|y)\, dx \\ &= \int_{\ell^{-1}(y)} (\log\, g_\phi^Y(y))\, h_\psi^{X|Y}(x|y)\, dx = L(\phi)\int_{\ell^{-1}(y)} h_\psi^{X|Y}(x|y)\, dx \\ &= L(\phi). \end{aligned}$$

In the following, we will omit the subscript "md" and just write $L$ instead of $L_{\mathrm{md}}$.

**Lemma 2.5.** *Let Assumptions A and B hold. If $f_\phi^X(x) \in C^k(\Phi)$, $k \in \mathbb{N}$ for all $x \in \mathbb{D}$, then $L \in C^k(\Phi)$ and $H \in C^{kk}(\Phi)$.*

*Proof.* By definition, we have for given $y \in \mathbb{O}$ that $L(\phi) = \log g_\phi^Y(y)$. By Assumption B, it holds that $g_\phi^Y(y) > 0$ so that $L$ inherits its regularity w.r.t. $\phi$ from those of $f_\phi^X$. With the same reasoning and Assumption B we get $h_\phi^{X|Y}(x|y) \neq 0$ so that $H$ is a combination of $C^k$-functions. □

For later reference, let us collect some basically straightforward facts. We start by showing some relations for derivatives.

**Lemma 2.6.** *Let Assumption C hold true, then $D^{11}H(\phi,\phi) = -D^{20}H(\phi,\phi)$ for all $\phi \in \Phi$.*

*Proof.* By Corollary A.8, we get for $\phi, \psi \in \Phi$ that

$$D^{10}H(\phi,\psi) = \frac{\partial}{\partial\phi}\int_{\ell^{-1}(y)} \big(\log h_\phi^{X|Y}(x|y)\big) h_\psi^{X|Y}(x|y)\,dx \qquad (2.10)$$

$$= \int_{\ell^{-1}(y)} \frac{\frac{\partial}{\partial\phi} h_\phi^{X|Y}(x|y)}{h_\phi^{X|Y}(x|y)} h_\psi^{X|Y}(x|y)\,dx$$

and consequently by Lemma A.9 with $q = \frac{\partial}{\partial\phi} h_\phi^{X|Y}(h_\phi^{X|Y})^{-1}$

$$D^{11}H(\phi,\psi) = \int_{\ell^{-1}(y)} \frac{\frac{\partial}{\partial\phi} h_\phi^{X|Y}(x|y)}{h_\phi^{X|Y}(x|y)} \Big(\frac{\partial}{\partial\psi} h_\psi^{X|Y}(x|y)\Big)^T\,dx,$$

$$D^{20}H(\phi,\psi) = \int_{\ell^{-1}(y)} \frac{\big(\frac{\partial^2}{\partial\phi^2} h_\phi^{X|Y}(x|y)\big) h_\phi^{X|Y}(x|y) - \big(\frac{\partial}{\partial\phi} h_\phi^{X|Y}(x|y)\big)^2}{(h_\phi^{X|Y}(x|y))^2} h_\psi^{X|Y}(x|y)\,dx,$$ [b]

where the second statement is implied by Lemma A.10. This implies by Corollary A.5 that

$$\begin{aligned} D^{20}H(\phi,\phi) &= \int_{\ell^{-1}(y)} \big(\frac{\partial^2}{\partial\phi^2} h_\phi^{X|Y}(x|y)\big)\,dx - \int_{\ell^{-1}(y)} \frac{\big(\frac{\partial}{\partial\phi} h_\phi^{X|Y}(x|y)\big)^2}{h_\phi^{X|Y}(x|y)} dx \\ &= \frac{\partial^2}{\partial\phi^2}\int_{\ell^{-1}(y)} h_\phi^{X|Y}(x|y)\,dx - D^{11}H(\phi,\phi) = -D^{11}H(\phi,\phi) \end{aligned}$$

since $\int_{\ell^{-1}(y)} h_\phi^{X|Y}(x|y)\,dx = 1$. □

**Lemma 2.7.** *Let $\phi \in \Phi$ and let Assumptions A and B hold.*
(a) *$\nabla L(\phi) = D^{10}Q(\phi,\psi) - D^{10}H(\phi,\psi)$ for all $\psi \in \Phi$.*
(b) *If in addition $f_\phi^X(x) \in C^2(\Phi)$ for all $x \in \mathbb{D}$, then we have for all $\psi \in \Phi$ that $\nabla^2 L(\phi) = D^{20}Q(\phi,\psi) - D^{20}H(\phi,\psi)$.*
(c) *$D^{01}Q(\phi,\psi) = D^{01}H(\phi,\psi)$ for all $\psi \in \Phi$.*
(d) *For the EM iterates $\phi^{(k)}$, we have $\nabla L(\phi^{(k+1)}) = -D^{10}H(\phi^{(k+1)},\phi^{(k)})$.*
(e) *$D^{11}H(\phi,\psi) = D^{11}Q(\phi,\psi)$ for all $\psi \in \Phi$.*

*Proof.* (a) and (b) follow directly from (2.9) by differentiating w.r.t. $\phi$ and (c) by differentiating w.r.t. $\psi$. By definition of the M-step, we have $D^{10}Q(\phi^{(k+1)},\phi^{(k)}) = 0$ so that (d) is a consequence of (a) using $\phi = \phi^{(k+1)}$ and $\psi = \phi^{(k)}$. Finally, (e) follows from (c). □

[b] We use the notation $(\cdot)^2 := (\cdot)(\cdot)^T$.

**Corollary 2.8.** *Under the assumptions of Lemma 2.6, we have* $D^{20}H(\phi,\phi) = -D^{11}Q(\phi,\phi)$.

*Proof.* By Lemma 2.3 we have that $Q \in C^{11}(\Phi)$. Then, the claim follows directly from Lemma 2.6 and Lemma 2.7 (e). □

**Lemma 2.9.** *Let Assumptions A and B hold. Then, for* $\phi \in \Phi$

(a) $D^{10}H(\phi,\phi) = 0$.
(b) *If* $\phi^{(k)} \to \phi^*$ *and* $\phi^*$ *is a critical point of* $L$, *we have* $D^{10}Q(\phi^*,\phi^*) = 0$.

*Proof.* By definition and Corollary A.8, we obtain for $\phi, \psi \in \Phi$

$$D^{10}H(\phi,\psi) = \frac{\partial}{\partial\phi}H(\phi,\psi) = \int_{\ell^{-1}(y)} \frac{\frac{\partial}{\partial\phi}h_\phi^{X|Y}(x|y)}{h_\phi^{X|Y}(x|y)} h_\psi^{X|Y}(x|y)\,dx.$$

This implies by Lemma A.4 $D^{10}H(\phi,\phi) = \int_{\ell^{-1}(y)} \frac{\partial}{\partial\phi}h_\phi^{X|Y}(x|y)\,dx = 0$. Hence, (a) is proven. Finally, (b) follows from Remark 2.4 by passing to the limit $\phi \to \phi^*$ □

## 3. EM Algorithm and Ascent as well as Quasi-Newton methods

The precise form of the target function $Q(\phi,\tilde{\phi})$ in the optimization of the M-step (2.6) obviously depends on the specific form of the log-Likelihood function $\ell_{\text{full}}$. This, in turns, implies the precise form of the EM algorithm. The EM iteration $\phi^{(k)} \to \phi^{(k+1)}$ in (2.6) can be written as $\phi^{(k+1)} = M(\phi^{(k)})$ with $M : \Phi \to \Phi$ defined as $M(\phi) := \arg\max_{\psi\in\Phi} Q(\psi,\phi)$. It is well-known that convergence properties of the EM algorithm depend on properties of the (fixpoint) function $M$. It is obvious that statements in the most general case are difficult if not impossible. In many cases, however, the EM iteration turns out to coincide with iterative schemes that are well-investigated in numerical analysis. The aim of this section is to show some of these similarities and relationships.

We first note a result which is known from the literature [5, Theorem 4] and include a proof since our findings slightly differ from those in [5].

**Proposition 3.1.** *If Assumption C holds and* $\phi^* \in \Phi$ *is a fixpoint of* $M$, *then*

$$\nabla M(\phi^*) = (D^{20}Q(\phi^*,\phi^*))^{-1}\, D^{20}H(\phi^*,\phi^*).$$

*Proof.* We start by the Taylor expansion for $\phi_1, \phi_2 \in \Phi$, i.e.,

$$\begin{aligned} D^{10}Q(\phi_2,\phi_1) &= D^{10}Q(\phi^*,\phi^*) + D^{20}Q(\phi^*,\phi^*)(\phi_2-\phi^*) \\ &\quad + D^{11}Q(\phi^*,\phi^*)(\phi_1-\phi^*) + o(\|\phi_1-\phi^*\| + \|\phi_2-\phi^*\|). \end{aligned}$$

Let $\psi^{(\ell)} := \phi^* + h^{(\ell)}e$ for $h^{(\ell)} \in \mathbb{R}^+$, $h^{(\ell)} \searrow 0$, $\ell \to \infty$, $e \in \Phi$, $\|e\| = 1$, be an arbitrary sequence converging to $\phi^*$. Setting $\phi_1 = \psi^{(\ell)}$, $\phi_2 = M(\psi^{(\ell)})$ and recalling that $D^{10}Q(M(\psi^{(\ell)}),\psi^{(\ell)}) = 0$ by definition of the M-step yields

$$\begin{aligned} 0 &= D^{20}Q(\phi^*,\phi^*)(M(\psi^{(\ell)})-\phi^*) + D^{11}Q(\phi^*,\phi^*)(\psi^{(\ell)}-\phi^*) \\ &\qquad + o(\|\psi^{(\ell)}-\phi^*\| + \|M(\psi^{(\ell)})-\phi^*\|) \\ &= D^{20}Q(\phi^*,\phi^*)(M(\psi^{(\ell)})-M(\phi^*)) + D^{11}Q(\phi^*,\phi^*)(\psi^{(\ell)}-\phi^*) + o(\|\psi^{(\ell)}-\phi^*\|). \end{aligned}$$

Dividing by $h^{(\ell)}$, passing to the limit $\ell \to \infty$ and noting that $e \in \Phi$, $\|e\| = 1$ can be chosen arbitrarily yields

$$\begin{aligned} 0 &= D^{20}Q(\phi^*, \phi^*)\, \nabla M(\phi^*) + D^{11}Q(\phi^*, \phi^*) \\ &= D^{20}Q(\phi^*, \phi^*)\, \nabla M(\phi^*) - D^{20}H(\phi^*, \phi^*), \end{aligned}$$

where we have used Corollary 2.8 in the last step. □

**Remark 3.2.** *Let us briefly comment on the relationship of the above result to* [5, Theorem 4]*, which reads in our notation* $\nabla M(\phi^*) = D^{20}H(\phi^*, \phi^*)(D^{20}Q(\phi^*, \phi^*))^{-1}$*. Thus, both statements coincide if and only if the two matrices* $D^{20}H(\phi^*, \phi^*)$ *and* $(D^{20}Q(\phi^*, \phi^*))^{-1}$ *are normal, which -at least in general- might not be the case.*

### 3.1. **Ascent methods.**

In some cases, the EM iteration can be written as an updating procedure, i.e.,

$$\phi^{(k+1)} = \phi^{(k)} + d^{(k)}, \tag{3.1}$$

where the update $d^{(k)}$ is a projection applied to the gradient of the log-Likelihood function, i.e.,

$$d^{(k)} = \eta\; P(\phi^{(k)})\, \nabla L(\phi^{(k)}), \tag{3.2}$$

$\eta > 0$ is a fixed step size parameter and $P(\phi) \in \mathbb{R}^{n \times n}$ is some matrix uniformly bounded in the parameter $\phi \in \Phi$, i.e. $\|P(\phi)\| \leq C$ for all $\phi \in \Phi$. The specific form of the EM algorithm in this case depends on the particular choices of $P$ and $\eta$.

If the EM algorithm takes the form (3.1), the scheme is an ascent method with ascent direction $d^{(k)}$. If in addition (3.2) is true, then the EM algorithm is a Quasi-Newton (sometimes also called *gradient akin*) method where $P(\phi)$ acts as an approximation of the inverse of the Hessian $\nabla^2 L$ of the log-Likelihood function and can thus be interpreted as a preconditioner. The step size is up to the choice of the user.

### 3.2. **Preconditioning ascent methods.**

In order to study the convergence properties of (3.1, 3.2), one needs to ensure that $P(\phi)$ is s.p.d. – at least acting on the current gradient (see also Proposition 4.2 below). The following result in that regard is well-known.

**Theorem 3.3** ([22, §2])**.** *Let the Assumptions A and B hold. Assume for* (3.1) *and* (3.2) *that* $\phi^{(k)} \neq \phi^{(k+1)}$, $k \in \mathbb{N}$, *holds and that* $D^{10}Q(\phi, \phi^{(k)}) = 0$ *if and only if* $\phi = \phi^{(k+1)}$*. Then,*

$$\nabla L(\phi^{(k)})^T\, P(\phi^{(k)})\, \nabla L(\phi^{(k)}) > 0 \quad \forall \phi^{(k)} \in \Phi, \tag{3.3}$$

*i.e.,* $P(\phi^{(k)})$ *is positive definite along the direction* $\nabla L(\phi^{(k)})$*.* □

Let us briefly recall from the literature under which circumstances (3.1) and (3.2) hold true. The explicit form of the matrix $P(\phi^{(k)})$ for the EM algorithm is given in [20, 21, 28] for some specific cases. These are in [28, Theorem 1] the Gaussian Mixtures model, in [21, Appendix] and [20, §2] the Mixture of Factor Analyzers model and moreover in [20, §§2,3] the Factor Analysis model, the Hidden Markov model and the Exponential Family models. Yet, a uniform derivation of $P(\phi^{(k)})$ for more general cases is not given in literature. Furthermore, the existence of an EM preconditioner $P(\phi^{(k)})$ is postulated in these papers without statements under which circumstances this assumption is actually justified. Hence, we generalize the statements on existence, derivation and special properties of the projection matrix for all problems where the M-step consists in solving a linear system of equations. In addition, the given derivation of $P(\phi^{(k)})$ will identify a connection between the EM scheme and the Fisher Information Matrix (FIM) of the complete data which in turns establishes a relation to Scoring algorithms, see Section 3.3 below.

Let us start with the particular situation where the M-step amounts solving a linear system. This case, however, will be important also for the general understanding.

**Theorem 3.4.** *Let Assumptions A and B hold and $Q(\cdot,\psi)$ is assumed to have a unique local maximum. If*

$$D^{10}Q(\phi,\psi) = -A(\psi)\phi + b(\psi) \tag{3.4}$$

*with a regular $A(\psi) \in \mathbb{R}^{n\times n}$ and $b(\psi) \in \mathbb{R}^n$. Then, the EM algorithm can be written in the form* (3.1), (3.2) *with*

$$P(\phi^{(k)}) = \big[\eta A(\phi^{(k)})\big]^{-1} = [-\eta D^{20}Q(\phi^{(k)},\phi^{(k)})]^{-1}. \tag{3.5}$$

*Moreover, $P(\phi^{(k)})$ is s.p.d. provided that in addition $f_\phi^X(x) \in C^2(\Phi)$ for all $x \in \mathbb{D}$.*

*Proof.* By Remark 2.4 we have that

$$\nabla L(\psi) = D^{10}Q(\psi,\psi) = -A(\psi)\,\psi + b(\psi), \quad \psi \in \Phi. \tag{3.6}$$

For the EM iterates $\phi^{(k)}$, we hence get that

$$\begin{aligned}\phi^{(k+1)} - \phi^{(k)} &= \big[A(\phi^{(k)})\big]^{-1} b(\phi^{(k)}) - \phi^{(k)} = \big[A(\phi^{(k)})\big]^{-1}\big(b(\phi^{(k)}) - A(\phi^{(k)})\,\phi^{(k)}\big)\\ &= \big[A(\phi^{(k)})\big]^{-1}\nabla L(\phi^{(k)}),\end{aligned} \tag{3.7}$$

i.e., $\eta P(\phi^{(k)}) = \big[A(\phi^{(k)})\big]^{-1}$. From assumption (3.4), we immediately obtain

$$D^{20}Q(\phi,\psi) = \frac{\partial}{\partial\phi}\Big(-A(\psi)\,\phi + b(\psi)\Big) = -A(\psi), \tag{3.8}$$

so that (3.5) is proven.

If in addition $f_\phi^X(x) \in C^2(\Phi)$ for all $x \in \mathbb{D}$, we obtain $Q \in C^{20}(\Phi)$ by Lemma 2.3, and then the Hessian $D^{20}Q(\phi,\psi)$ is symmetric and so is $P(\phi^{(k)})$. Note that the Hessian in (3.8) does not depend on $\phi$ which means that $-A(\psi)$ also corresponds to the Hessian at the unique local maximum of $Q(\cdot,\psi)$, where the Hessian is symmetric and negative semidefinite. Since $A(\psi)$ is assumed to be regular, we deduce that $A(\psi)$ is s.p.d. which by (3.5) implies that $P(\phi^{(k)})$ is s.p.d. □

Of course, the assumption (3.4) is very specific and will not hold true in general. There are two separate issues. First, (3.4) may not be true at all and second, (3.4) may not be valid for all $\psi \in \Phi$. The latter one is not serious since a closer look to the proof shows that we only need (3.4) for the specific choice $\psi = \phi^{(k)}$ and arbitrary $\phi \in \Phi$, i.e.,

$$D^{10}Q(\phi,\phi^{(k)}) = -A(\phi^{(k)})\,\phi + b(\phi^{(k)}). \tag{3.9}$$

Obviously, such an assumption is delicate to check a-priorily since the iterates $\phi^{(k)}$ are unknown, which is the reason for the general formulation (3.4). However, even if (3.9) is not true, we get at least an approximation as the following statement shows.

**Proposition 3.5.** *Let Assumption C hold and $\|(D^{20}Q(\phi,\phi))^{-1}\| < \infty$ uniformly for $\phi \in \Phi$. Then, the EM iteration satisfies*

$$\phi^{(k+1)} = \phi^{(k)} - [D^{20}Q(\phi^{(k)}, \phi^{(k)})]^{-1}\, \nabla L(\phi^{(k)}) + o(\|\phi^{(k+1)} - \phi^{(k)}\|),$$

*i.e., $P(\phi^{(k)}) \approx [-\eta D^{20}Q(\phi^{(k)}, \phi^{(k)})]^{-1}$.*

*Proof.* We use Taylor's expansion (e.g. [13, XVI, §6]), i.e.

$$D^{10}Q(\phi, \psi) = D^{10}(\phi_0, \psi) + D^{20}Q(\phi_0, \psi)(\phi - \phi_0) + o(\|\phi - \phi_0\|).$$

The assertion follows using $\phi_0 = \phi^{(k)}$, $\phi = \phi^{(k+1)}$ and $\psi = \phi^{(k)}$ recalling that the M-step implies $D^{10}Q(\phi^{(k+1)}, \phi^{(k)}) = 0$ and that $D^{10}Q(\phi^{(k)}, \phi^{(k)}) = \nabla L(\phi^{(k)})$. □

This latter statement shows that the EM iteration asymptotically behaves like an EM algorithm for the specific case in Theorem 3.4.

We continue with some statements on the matrix $P$. Additionally, the remaining part of this section shall serve to give an overview on similarities and differences to existing literature. We start by a well-known short-hand notation.

**Definition 3.6.** *The* Loewner ordering *of symmetric matrices, i.e. $A \succ B$, for $A, B \in \mathbb{R}^{n\times n}$ symmetric, means that $A - B$ is positive definite.*

**Lemma 3.7.** *Let Assumptions A and B hold and $f_\phi^X(x) \in C^2(\Phi)$ for all $x \in \mathbb{D}$. If the series $\{\phi^{(k)}\}_{k\in\mathbb{N}}$ of EM iterates for which* (3.1), (3.2) *hold converges towards some $\phi^* \in \Phi$ such that $D^{10}Q(\phi^{(k+1)}, \phi^{(k)}) = 0$ and $D^{20}Q(\phi, \phi^{(k)}) \prec 0$, then*

$$P(\phi^*) = -(\eta D^{20}Q(\phi^*, \phi^*))^{-1}.$$

*Proof.* For $\phi \in \Phi$, we have $M(\phi) - \phi = \eta P(\phi)\, \nabla L(\phi)$ and due to differentiability of $M$, $P$ and $\nabla L$ (at least locally) we have

$$\nabla M(\phi) - I = \eta \nabla P(\phi)\, \nabla L(\phi) + \eta P(\phi)\, \nabla^2 L(\phi).$$

By assumption the iteration converges, i.e., $\nabla L(\phi^*) = 0$ and by passing to the limit $\phi \to \phi^*$, we obtain $\nabla M(\phi^*) = I + \eta P(\phi^*)\, \nabla^2 L(\phi^*)$, so that $P(\phi^*) = (\nabla M(\phi^*) - I)(\eta\nabla^2 L(\phi^*))^{-1}$. Using Proposition 3.1, we arrive at

$$P(\phi^*) = [(D^{20}Q(\phi^*, \phi^*))^{-1} D^{20}H(\phi^*, \phi^*) - I](\eta\nabla^2 L(\phi^*))^{-1}. \tag{3.10}$$

Now we use Lemma 2.7 and get $D^{20}H(\phi, \psi) = D^{20}Q(\phi, \psi) - \nabla^2 L(\phi)$ so that

$$\begin{aligned} P(\phi^*) &= [(D^{20}Q(\phi^*, \phi^*))^{-1}(D^{20}Q(\phi^*, \phi^*) - \nabla^2 L(\phi^*)) - I](\eta\nabla^2 L(\phi^*))^{-1} \\ &= [-(D^{20}Q(\phi^*, \phi^*))^{-1}\nabla^2 L(\phi^*)](\eta\nabla^2 L(\phi^*))^{-1} = -(\eta D^{20}Q(\phi^*, \phi^*))^{-1}. \end{aligned}$$

This proves the lemma. □

**Remark 3.8.** *Equation* (3.10) *should also be valid at least approximately by replacing the fixpoint $\phi^*$ by $\phi^{(k)}$ for sufficiently large $k$. If this is so, then the result is the same as in the affine case in Theorem 3.4,* (3.5)*. The interpretation of* (3.10) *in* [22] *reads: "When the missing information is small compared to the complete information, EM exhibits approximate Newton behavior and enjoys fast, typically superlinear convergence in the neighborhood of $\phi^*$".*
*It should be noted, however, that our derivation differs from* [22] *in two ways. First,* [5, Theorem 4] *was used there, so that the comments in Remark 3.2 apply. Second, the derivation in* [22] *is done basically along the lines described above, with $\phi^*$ replaced by $\phi^{(k)}$, however. Since $\nabla L(\phi^{(k)}) \neq 0$ and* [5, Theorem 4] *is only given for*

*$\phi = \phi^*$, the substitution can not be performed straightforwardly. In* [22] *no justification to use* [5, Theorem 4] *for $\phi = \phi^{(k)}$ and no approximation quality for their result is given. Yet, we will see within the proof of the next Proposition that given some additional assumptions on $\nabla^2 L$ and $\nabla L$ we can derive a similar approximation for $P(\phi^{(k)})$ with approximation quality $o(1)$.*

We are now going to make the approximate representation of Proposition 3.5 rigorous.

**Proposition 3.9.** *Assume Assumption C. Let $M(\phi) := \phi + \eta P(\phi)\,\nabla L(\phi)$, $P \in C^1(\Phi)$ be the fixpoint scheme of an EM algorithm. Moreover, assume the estimates $\|(\nabla^2 L(\phi^{(k)}))^{-1}\| = \mathcal{O}(1)$ and $\|\nabla L(\phi^{(k)})\| = o(1)$, $k \to \infty$. Then, as $k \to \infty$*

$$P(\phi^{(k)}) = -\big(\eta D^{20} Q(\phi^*, \phi^*)\big)^{-1} + o(1).$$

*Proof.* By assumption, we have $\nabla M(\phi) = I + \eta \nabla P(\phi)\,\nabla L(\phi) + \eta P(\phi)\,\nabla^2 L(\phi)$ for all $\phi \in \Phi$, $\|\nabla M(\phi^*) - \nabla M(\phi)\| = o(1)$ as $\phi \to \phi^*$ and in particular

$$\eta P(\phi^{(k)})\,\nabla^2 L(\phi^{(k)}) - \big(\nabla M(\phi^*) - I\big) = -\eta \nabla P(\phi^{(k)})\,\nabla L(\phi^{(k)}) + o(1) = o(1)$$

which finally leads to $P(\phi^{(k)}) = (\nabla M(\phi^*) - I)(\eta \nabla^2 L(\phi^{(k)}))^{-1} + o(1)$. Now, we use Proposition 3.1 so that the first term on the right-hand side reads

$$\begin{aligned}
(\nabla M(\phi^*) - I)(\eta \nabla^2 L(\phi^{(k)}))^{-1} \quad &= \\
&= \Big((D^{20} Q(\phi^*, \phi^*))^{-1} D^{20} H(\phi^*, \phi^*) - I\Big)(\eta \nabla^2 L(\phi^{(k)}))^{-1} \\
&= \Big((D^{20} Q(\phi^*, \phi^*)^{-1}[D^{20} Q(\phi^*, \phi^*) - \nabla^2 L(\phi^*)] - I)\Big)(\eta \nabla^2 L(\phi^{(k)}))^{-1} \\
&= -(D^{20} Q(\phi^*, \phi^*))^{-1}\,\nabla^2 L(\phi^*)\,(\eta \nabla^2 L(\phi^{(k)}))^{-1},
\end{aligned}$$

where we have used Lemma 2.7. Since $L \in C^2(\Phi)$ by assumption, we finally get $\nabla^2 L(\phi^*)\,\nabla L(\phi^{(k)})^{-1} = I + o(1)$ so that the claim is proven. □

This allows now for the following convergence statements:

**Theorem 3.10.** *Under the assumptions of Proposition 3.9, we have*

(a) $\phi^{(k+1)} - \phi^{(k)} = -(D^{20} Q(\phi^*, \phi^*))^{-1}\,\nabla L(\phi^{(k)}) + o(1)$

(b) $\phi^{(k+1)} - \phi^* = [I - (D^{20} Q(\phi^*, \phi^*))^{-1}\,\nabla^2 L(\phi^*)](\phi^{(k)} - \phi^*) + o(||\phi^{(k)} - \phi^*||)$.

*Proof.* Using the above definitions and Proposition 3.9, it holds

$$\phi^{(k+1)} - \phi^{(k)} = \eta P(\phi^{(k)})\,\nabla L(\phi^{(k)}) = -(D^{20} Q(\phi^*, \phi^*))^{-1}\,\nabla L(\phi^{(k)}) + o(1),$$

which is (a). As for (b), we have by Taylor's expansion that

$$\begin{aligned}
\phi^{(k+1)} - \phi^* \quad &= \quad \phi^{(k)} - \phi^* + \eta P(\phi^{(k)})\,\nabla L(\phi^{(k)}) \\
&= \quad \phi^{(k)} - \phi^* + \eta P(\phi^{(k)})\,\nabla^2 L(\phi^*)(\phi^* - \phi^{(k)}) + o(\|\phi^{(k)} - \phi^*\|) \\
&= \quad \big(I - (D^{20} Q(\phi^*, \phi^*))^{-1}\,\nabla^2 L(\phi^*)\big)\,(\phi^{(k)} - \phi^*) + o(\|\phi^{(k)} - \phi^*\|),
\end{aligned}$$

where we have used Proposition 3.9 in the last step. □

**Remark 3.11.** *The approximation of Theorem 3.10 (a) can even be improved yielding an approximation quality of $o(\|\phi^{(k)} - \phi^*\|)$ instead of $o(1)$. A proof therefore is given in* [9, Appendix A.1]. *The approximation is used in* [9] *to show that the EM step can approximately be viewed as a generalized gradient of the log-Likelihood $L(\phi)$ which is hence employed for conjugate gradient acceleration of the EM algorithm.*

*A corresponding approximate representation of (b) using FIMs instead of Hessians can also be found in* [17, (12)]. *Yet, without a rigorous statement or proof on the approximation quality.*

### 3.3. **Fisher Scoring Algorithm (FSA).**

The *Fisher Scoring Algorithm (FSA)* is a Newton-type method to compute a numerical approximation of an ML estimate in terms of a root of the *score* which is typically defined as $V := \nabla L$, the parametric derivative of the log-Likelihood function [14, 15]. The basic idea is to replace the Hessian $\nabla^2 L(\phi^{(k)})$ in Newton's method (i.e., the observed information) by the expected information.

We start by describing the FSA in the general case of observed data and will then detail the connections to the EM algorithm in the case of missing data. The variance of the score, also known as the *Fisher Information Matrix (FIM)*,

$$\mathcal{F}(\phi) := \mathrm{Var}[\nabla L(\phi)|\phi], \tag{3.11}$$

replaces the negative Hessian. The FSA then reads

$$\phi^{(k+1)} = \phi^{(k)} + \mathcal{F}^{-1}(\phi^{(k)})\, \nabla L(\phi^{(k)}). \tag{3.12}$$

This construction obviously offers two advantages from a numerical point of view. First, $\mathcal{F}(\phi)$ is positive semidefinite by construction and second it avoids possibly cumbersome computations of second derivatives (as required in Newton's method).

**Lemma 3.12.** *If Assumption A holds, if $g_\phi^Y(y) \in C^2(\Phi)$ for all $y \in \mathbb{O}$ and $\frac{\partial}{\partial\phi} g_\phi^Y \in L_\infty(\mathbb{O})$, then we have*

$$\mathcal{F}(\phi) = \mathrm{Var}[\nabla L(\phi)|\phi] = \mathbb{E}[(\nabla L(\phi))^2|\phi] = -\mathbb{E}[\nabla^2 L(\phi)|\phi].$$

*Proof.* First, note that (recalling that $\mathbb{O} \subset \mathbb{R}^n$ is the observed data space)

$$\begin{aligned} \mathbb{E}[\nabla L(\phi)|\phi] &= \int_{\mathbb{O}} \Big(\frac{\partial}{\partial\phi} \log\, g_\phi^Y(y)\Big) g_\phi^Y(y)\, dy = \int_{\mathbb{O}} \frac{\frac{\partial}{\partial\phi} g_\phi^Y(y)}{g_\phi^Y(y)} g_\phi^Y(y)\, dy \\ &= \int_{\mathbb{O}} \frac{\partial}{\partial\phi} g_\phi^Y(y)\, dy = \frac{\partial}{\partial\phi} \int_{\mathbb{O}} g_\phi^Y(y)\, dy = 0 \end{aligned}$$

by Remark A.2. This means that the score has vanishing expectation. Then,

$$\begin{aligned} \mathcal{F}(\phi) &= \mathbb{E}[(\nabla L(\phi))^2|\phi] - \mathbb{E}[\nabla L(\phi)|\phi]^2 = \mathbb{E}[(\nabla L(\phi))^2|\phi] \\ &= -\frac{\partial}{\partial\phi} \int_{\mathbb{O}} \frac{\partial}{\partial\phi} g_\phi^Y(y)\, dy + \mathbb{E}[(\nabla L(\phi))^2|\phi] \\ &= -\int_{\mathbb{O}} \frac{\partial^2}{\partial\phi^2} g_\phi^Y(y)\, dy + \int_{\mathbb{O}} \Big(\frac{\partial}{\partial\phi} \log\, g_\phi^Y(y)\Big)^2 g_\phi^Y(y)\, dy \\ &= -\int_{\mathbb{O}} \Big(\frac{\partial^2}{\partial\phi^2} \log\, g_\phi^Y(y)\Big)\, g_\phi^Y(y)\, dy = -\mathbb{E}[\nabla^2 L(\phi)|\phi], \end{aligned}$$

which proves the claim by Remark A.2 and Lemma A.3. □

Lemma 3.12 shows that the Jacobian in Newton's method is replaced by the expectation of the Jacobian of the score, i.e., the expectation of the Hessian of the log-Likelihood [15]. This is the reason why $\mathcal{F}$ in (3.11) is sometimes called *expected FIM* as opposed to the negative Hessian $\mathcal{I} = -\nabla^2 L$ of the log-Likelihood (i.e., standard Newton's method) which is called *observed FIM* [6]. We use the letter $\mathcal{F}$ for expected and $\mathcal{I}$ for observed information matrices.

Let us now establish the connection of the above definitions to the case of missing data and the EM algorithm. In a missing data model we have only access to the observed data $y$ and know that there exists a relation to the complete data $x$ which can, however, not be observed. We can thus fabricate a kind of mixture of the observed and expected FIMs. This is done building the FIM of the complete data Likelihood conditioned on the observed data $y$. Hence, it seems appropriate to replace $\mathcal{F}$ by

$$\mathcal{I}_{\text{full}}(\phi) := \mathbb{E}\Big[\Big(\nabla \ell_{\text{full}}(\cdot;\phi)\Big)^2 \Big| y, \phi\Big] = \int_{\ell^{-1}(y)} \Big(\frac{\partial}{\partial \phi} \log f_\phi^X(x)\Big)^2 h_\phi^{X|Y}(x|y)\, dx. \tag{3.13}$$

Note that $\mathcal{I}_{\text{full}}(\phi)$ contains both expected and observed information. The use of the conditioned expectation implies that the available observed information is used, whereas the expectation is performed for the missing information. Thus, we use the letter $\mathcal{I}$ to indicate observed information. A very similar reasoning as for proving Lemma 3.12 yields

$$\mathcal{I}_{\text{full}}(\phi) = -\mathbb{E}\Big[\frac{\partial^2}{\partial \phi^2} \log f_\phi^X(\cdot) \Big| y, \phi\Big] \tag{3.14}$$

and the variant of FSA reads

$$\phi^{(k+1)} = \phi^{(k)} + \mathcal{I}_{\text{full}}(\phi^{(k)})^{-1}\, \nabla L_{\text{md}}(\phi^{(k)}).$$

Interchanging expectation and differentiation (Lemma A.6 and Corollary A.7) in (3.14) together with (2.5) results in

$$\mathcal{I}_{\text{full}}(\phi) = -D^{20}Q(\phi,\phi). \tag{3.15}$$

In view of (3.2) and Theorem 3.4, the FSA variant hence coincides with the EM algorithm for the particular choice $P(\phi^{(k)}) = [\eta \mathcal{I}_{\text{full}}(\phi^{(k)})]^{-1}$. This motivates a closer look towards the properties of $\mathcal{I}_{\text{full}}$.

**Lemma 3.13.** *Let Assumption C hold true. Then, the observed FIM for the complete data Likelihood $\mathcal{I}_{\text{full}}$ reads*

$$\mathcal{I}_{\text{full}} = \mathcal{F}_{\text{miss}} + \mathcal{I}_{\text{obs}},$$

*where $\mathcal{F}_{\text{miss}}(\phi) := \operatorname{Var}[\nabla \ell_{\text{miss}}(\cdot|y;\phi)|y,\phi] = \mathbb{E}\big[-\nabla^2 \ell_{\text{miss}}(\cdot|y;\phi)\big|y,\phi\big]$ is the expected FIM of the missing data Likelihood (cf. Lemma 3.12).*

*Proof.* We first note that $f_\phi^X \in C^2(\Phi)$ implies that $g_\phi^Y, h_\phi^{X|Y} \in C^2(\Phi)$. This, in turns, ensures that $\ell_{\text{full}}(x;\cdot)$, $\ell_{\text{obs}}(y;\cdot)$, $\ell_{\text{miss}}(x|y;\cdot) \in C^2(\Phi)$. From (2.4), we then deduce

$$\frac{\partial^2}{\partial \phi^2} \ell_{\text{full}}(x;\phi) = \frac{\partial^2}{\partial \phi^2} \ell_{\text{obs}}(y;\phi) + \frac{\partial^2}{\partial \phi^2} \ell_{\text{miss}}(x|y;\phi).$$

Now, we form the expectation on both sides to obtain

$$\begin{aligned}
\mathcal{I}_{\text{full}}(\phi) &= -\int_{\ell^{-1}(y)} \Big(\frac{\partial^2}{\partial\phi^2} \log\, f_\phi^X(x)\Big)\, h_\phi^{X|Y}(x|y)\, dx \\
&= -\int_{\ell^{-1}(y)} \Big(\frac{\partial^2}{\partial\phi^2} \log\, g_\phi^Y(y)\Big)\, h_\phi^{X|Y}(x|y)\, dx \\
&\qquad - \int_{\ell^{-1}(y)} \Big(\frac{\partial^2}{\partial\phi^2} \log\, h_\phi^{X|Y}(x)\Big)\, h_\phi^{X|Y}(x|y)\, dx \\
&= -\frac{\partial^2}{\partial\phi^2}\Big(\log\, g_\phi^Y(y)\Big) + \mathbb{E}\Big[-\frac{\partial^2}{\partial\phi^2} \log\, h_\phi^{X|Y}(x|y)\Big| y, \phi\Big] \\
&= \mathcal{I}_{\text{obs}}(\phi) + \mathcal{F}_{\text{miss}}(\phi),
\end{aligned}$$

which proves the claim. □

It is well-known that the convergence-properties of an iteration of the form (3.1,3.2) crucially depend on the positivity properties of the matrix $P(\phi)$, $\phi \in \Phi$, i.e., in the FSA-variant of EM we have to investigate the positivity of $\mathcal{I}_{\text{full}}$.

**Remark 3.14.** *Let Assumption C hold. Then, $\mathcal{I}_{\text{full}}(\phi)$ is positive definite for some $\phi \in \Phi$ provided that (recall Definition 3.6)*

$$\mathcal{F}_{\text{miss}}(\phi) \succ -\mathcal{I}_{\text{obs}}(\phi). \tag{3.16}$$

*Note, that* (3.16) *is satisfied if $\mathcal{I}_{\text{obs}}(\phi)$ is s.p.d. since $\mathcal{F}_{\text{miss}}$ is s.p.d. as a covariance matrix in view* (3.11)*.*

**Proposition 3.15.** *Let Assumption C hold and let $\phi^* \in \Phi$ be a non-degenerate local maximal point of $L$. Then, the matrix $\mathcal{I}_{\text{full}}(\phi^*)$ is s.p.d.*

*Proof.* Under the above assumptions, the Hessian $\nabla^2 L(\phi^*)$ is negative definite. Hence, the matrix $\mathcal{I}_{\text{obs}}(\phi^*) = -\nabla^2 L(\phi^*)$ is s.p.d. so that (3.16) holds true which proves the claim. □

3.4. **A linearized fixpoint scheme for the Euler-Lagrange equation.** If the log-Likelihood function $L$ is smooth, i.e., $L \in C^1(\Phi)$, the ML estimate $\hat{\phi}^{\text{ML}}$ is a critical point of $L$, i.e.,

$$\nabla L(\hat{\phi}^{\text{ML}}) = 0.$$

In other words, $\phi^* = \hat{\phi}^{\text{ML}}$ is a root of the Euler-Lagrange equation (first-order necessary condition).

Let us consider the case in which the assumption of Theorem 3.4 holds, i.e., the affine assumption on $D^{10}Q(\phi, \psi)$ in (3.4). Then, we are looking for a root $\phi^*$ of the function

$$F(\phi) := \nabla L(\phi) = D^{10}Q(\phi, \phi) = b(\phi) - A(\phi)\phi,$$

i.e., $A(\phi^*)\phi^* = b(\phi^*)$. Since $A$ and $b$ in general depend in a nonlinear way on the argument, the latter equation is nonlinear. A simple numerical scheme is a linearized fixpoint-type method, where $\phi^{(k+1)}$ is determined as the solution of the linear system

$$A(\phi^{(k)})\phi^{(k+1)} = b(\phi^{(k)}), \quad k = 0, 1, 2, \ldots \tag{3.17}$$

given some initial guess $\phi^{(0)} \in \Phi$. If the iteration $(\phi^{(k)})_{k\in\mathbb{N}_0}$ converges towards some $\phi^*$, this $\phi^*$ is a root of $F = \nabla L$ and hence coincides with the limit of the EM algorithm. According to Theorem 3.4, the EM iteration reads (under the corresponding assumptions) $-A(\phi^{(k)})\phi^{(k+1)} + b(\phi^{(k)}) = D^{10}Q(\phi^{(k+1)}, \phi^{(k)}) = 0$, so that in fact the two schemes coincide. We may summarize these observations as follows.

**Proposition 3.16.** *Let the assumptions of Theorem 3.4 hold. Then, the EM iteration coincides with the linearized fixpoint scheme* (3.17) *for the Euler-Lagrange equation.* □

## 4. Convergence

We found several statements in the literature concerning convergence properties of the EM algorithm somehow incomplete since often only very specific cases have been investigated or the statements themselves seem to be somewhat vague. With the results of the previous section at hand, we can derive statements concerning convergence and rate of convergence of the EM algorithm. In fact, if we know that the EM algorithm coincides with a particular iterative scheme (under certain assumptions), we can use convergence properties of these schemes in order to derive corresponding results for the EM algorithm. We start by reviewing some more or less well-known general considerations concerning convergence and convergence rates of iterative schemes and will apply these results then to the EM algorithm.

### 4.1. **Convergence of iterative schemes.**

4.1.1. *Fixpoint schemes.* Let $M \in C^1(\Phi)$, $\phi^{(0)} \in \Phi$, and consider the fixpoint iteration

$$\phi^{(k+1)} := M(\phi^{(k)}), \qquad k = 0, 1, 2, \dots \tag{4.1}$$

It is well-known that the convergence properties of (4.1) crucially depend on properties of the fixpoint function $M$. If $M \in C^{p+1}(\Phi)$, $p \geq 1$, and

$$\nabla^\ell M(\phi^*) = 0, \quad 1 \leq \ell \leq p, \qquad \nabla^{p+1} M(\phi^*) \neq 0,$$

then (4.1) converges locally with order $p+1$. If $M \in C^1(\Phi)$, $\nabla M(\phi^*) \neq 0$, $\|\nabla M(\phi^*)\| < 1$, then the iteration (4.1) is locally linear convergent. In this case, the speed of convergence is also influenced by the error reduction factor $\varrho$, i.e.,

$$\|\phi^{(k+1)} - \phi^*\| \leq \varrho \|\phi^{(k)} - \phi^*\|, \qquad k \geq K_0. \tag{4.2}$$

In the above framework, the error reduction factor is hence $\varrho = \|\nabla M(\phi^*)\|$. Since this situation is particularly important for the EM algorithm, let us formulate the above observations.

**Proposition 4.1.** *Let $M \in C^1(\Phi)$ with $\nabla M(\phi^*) \neq 0$. If the spectral radius satisfies $\rho(\nabla M(\phi^*)) < 1$, then* (4.1) *is locally linear convergent with $\varrho = \rho(\nabla M(\phi^*))$ (cf.*[18]*).*

*Proof.* By Taylor's expansion, we have

$$\phi^{(k+1)} - \phi^* = M(\phi^{(k)}) - M(\phi^*) = \nabla M(\phi^*)(\phi^{(k)} - \phi^*) + o(\|\phi^{(k)} - \phi^*\|).$$

Hence, for sufficiently large $k$, $\|\phi^{(k+1)} - \phi^*\| \leq \|\nabla M(\phi^*)\| \, \|\phi^{(k)} - \phi^*\|$ for any norm. Choosing an appropriate norm, we get $\|\nabla M(\phi^*)\| = \varrho$ so that convergence is proven in this particular norm. Finally, the claim follows from the equivalence of all norms on a finite-dimensional space. □

4.1.2. *Quasi-Newton methods.* The convergence analysis of Quasi-Newton methods is well-established and can be found in several text books. One way to establish corresponding results is to rewrite Quasi-Newton schemes as fixpoint iteration and to use the statements of the previous section. Usually, full Newton exhibits (at least) locally quadratic convergence, Quasi-Newton locally linear.

4.1.3. *Preconditioned gradient ascent methods.* Let us consider a preconditioned gradient ascent scheme with *fixed* step size $\eta > 0$ and some preconditioner $B$, i.e.,

$$\phi^{(k+1)} = M_\eta(\phi^{(k)}) := \phi^{(k)} + \eta\, B(\phi^{(k)})\, \nabla L(\phi^{(k)}), \tag{4.3}$$

i.e., $M_\eta(\phi) := \phi + \eta\, B(\phi)\, \nabla L(\phi)$, $\nabla M_\eta(\phi) = I + \eta\, \nabla B(\phi)\, \nabla L(\phi) + \eta\, B(\phi)\, \nabla^2 L(\phi)$. In general, we expect locally linear convergence so that we have to ensure error reduction, i.e.,

$$\varrho_\eta := \varrho_\eta(\phi^*) < 1, \quad \text{where} \quad \varrho_\eta(\phi) := \rho(\nabla M_\eta(\phi)).$$

The following statement shows the optimal choice of the step size $\eta$.

**Proposition 4.2.** *Let $L \in C^2(\Phi)$, $\phi^* \in \Phi$ be a nondegenerate maximal point of $L$ and $\mathcal{H} := \nabla^2 L(\phi^*)$ the Hessian of $L$ at $\phi^*$. Moreover, let $B \in C^1(\Phi)$ be s.p.d. on $\Phi$. Then, the preconditioned gradient ascent method* (4.3) *is locally linear convergent with optimal step size $\eta_{\text{opt}}$ and corresponding optimal error reduction $\varrho_{\text{opt}}$ given by*

$$\eta_{\text{opt}} = \frac{2}{|r| + |R|}, \qquad \varrho_{\text{opt}} = \frac{\kappa - 1}{\kappa + 1}, \tag{4.4}$$

*where $r := \lambda_{\min}(B(\phi^*)\mathcal{H})$, $R := \lambda_{\max}(B(\phi^*)\mathcal{H})$ are minimal and maximal eigenvalues of $B(\phi^*)\mathcal{H}$, respectively, and $\kappa := \frac{|r|}{|R|}$, which corresponds to the condition number $\kappa_2(B(\phi^*)\mathcal{H})$ provided that the product $B(\phi^*)\mathcal{H}$ is normal.*

*Proof.* Since $\mathcal{H}$ is symmetric negative definite and $B(\phi)$ symmetric positive definite by assumption, the eigenvalues of the product $B(\phi^*)\mathcal{H}$ are all real and negative, so that $-\infty < r < R < 0$ (see Remark 4.3 below). Since $\nabla L(\phi^*) = 0$, we have $\nabla M_\eta(\phi^*) = I + \eta\, B(\phi^*)\, \nabla^2 L(\phi^*)$. The spectral radius of $\nabla M_\eta(\phi^*)$ is hence given by $\varrho_\eta = \max\{|1 + \eta r|, |1 + \eta R|\}$. It is readily seen that $0 < \eta < \frac{2}{|r|} =: \eta_\infty$ ensures $\varrho_\eta < 1$. It is also readily seen that

$$\varrho_\eta = (1 - \eta|R|)\chi_{(0,\eta_{\text{opt}})} + (\eta|r| - 1)\chi_{[\eta_{\text{opt}},\eta_\infty)}$$

is convex and piecewise linear so that it takes its minimum at $\eta_{\text{opt}}$ given by $1 - \eta_{\text{opt}}|R| = \eta_{\text{opt}}|r| - 1$, which shows the first formula in (4.4). Finally,

$$\varrho_{\text{opt}} = \varrho_{\eta_{\text{opt}}} = 1 - \eta_{\text{opt}}|R| = \frac{|r| - |R|}{|r| + |R|} = \frac{\kappa - 1}{\kappa + 1},$$

since $\kappa = \frac{|r|}{|R|}$. □

**Remark 4.3.** *The product of two symmetric positive definite matrices $A, B \in \mathbb{R}^{n \times n}$ has only real and positive eigenvalues.*

*Proof.* Since $A$ is s.p.d., its inverse $A^{-1}$ is also s.p.d. Hence, for the Cholesky decomposition $A^{-1} = LL^T$, the matrix $L$ is lower triangular with strictly positive diagonal entries. Thus, $A = L^{-T}L^{-1}$. Similarly, $B = RR^T$, where $R$ is also a lower triangular matrix with strictly positive diagonal entries.

Note, that $AB$ and $L^{-1}BL^{-T}$ have the same eigenvalues, which can be seen as follows: Let $v$ be an eigenvector of $AB$ with eigenvalue $\lambda$, then $L^T v$ is an eigenvector of $L^{-1}BL^{-T}$ with eigenvalue $\lambda$: $(L^{-1}BL^{-T})(L^T v) = L^{-1}Bv = L^T L^{-T} L^{-1} Bv = L^T ABv = \lambda L^T v$.

Since $B$ is s.p.d., also $L^{-1}BL^{-T}$ is s.p.d., since $(L^{-1}BL^{-T})^T = L^{-1}B^T L^{-T} = L^{-1}BL^{-T}$ and thus for all $x \neq 0$

$$x^T L^{-1}BL^{-T}x = x^T L^{-1}RR^T L^{-T}x = (R^T L^{-T}x)^T(R^T L^{-T}x) = \|R^T L^{-T}x\|_2 > 0.$$

The eigenvalues of $L^{-1}BL^{-T}$ are hence all real and positive just as those of $AB$. □

**Remark 4.4.**

(a) *The above convergence statements rely on the optimal step size choice in* (4.4) *and are only true in this case.*

(b) *One expects improvements if the step size is adapted in each iteration which corresponds to an exact line search. In this case, all convergence-statements concerning steepest descent methods apply.*

(c) *Proposition 4.2 sets the framework for numerical comparisons of different methods by computing eigenvalues of $B(\phi^*)\mathcal{H}$ for different choices of $B$ — as long as $\phi^*$ is known, of course.*

4.2. **The EM algorithm.** We conclude this section with some remarks concerning the convergence of the EM algorithm in (3.1, 3.2). In the framework of Theorem 3.3, we can interpret the EM algorithm as a Quasi-Newton method. This means by Proposition 4.1, that (3.1, 3.2) converges at least locally linear. If $P(\phi) = -\nabla^2 L(\phi)^{-1}$ (i.e. the Hessian is used), EM coincides with Newton's method being at least locally quadratical convergent. Moreover, the experiments for Gaussian mixture densities in [28] as well as our analysis and experiments for the tracking context in Sections 5.2 and 5.3 below show that the multiplication of $\nabla^2 L$ with $P$ in (3.5) significantly reduces the condition number. In the special cases considered, this makes the EM algorithm similar to a superlinear method. For details on preconditioners for gradient descent methods see [23].

## 5. The PHMT tracking model

In this section, we describe the Probabilistic Multi-Hypothesis Tracking (PMHT) and apply our previous results to the EM algorithm for the PMHT problem. PMHT is a data-association/tracking approach proposed in [25, 26]. In the literature, PMHT is often understood as a model that is always solved by means of the EM algorithm. We rather view PMHT just as a stochastic model that can be solved with any appropriate numerical scheme. We will first describe the PMHT model and then use our above results to analyze the performance of EM for PMHT.

5.1. **The PMHT model.** We start by reviewing the PMHT model. We will mostly employ the notation used in the PMHT literature, but identify later the properties with our notation used in Section 2.

Assume that we are aiming at tracking $M \in \mathbb{N}$ objects over $T$ time steps, $t = 1, \ldots, T$, given their state distribution at the initial time $t = 0$. The state of each object can be described by an $n_X$-dimensional vector, $n_X \in \mathbb{N}$. All state information is collected in a vector

$$\mathbf{X} := (X^0, \ldots, X^T), \quad X^t := (x_1^t, \ldots, x_M^t), \quad x_m^t \in \mathcal{X} \subset \mathbb{R}^{n_X},$$

for $0 \le t \le T$, $1 \le m \le M$, where $\mathcal{X} \subset \mathbb{R}^{n_X}$ is the compact data space for one object. The corresponding complete state space reads

$$\mathbb{X} := \bigtimes_{t=0}^{T} \bigtimes_{m=1}^{M} \mathcal{X} = \mathcal{X}^{(T+1)M} \subset \mathbb{R}^{N_X}, \quad N_X := (T+1) \cdot M \cdot n_X.$$

The complete state information is neither known nor observable and thus needs to be estimated.

The observed data is a measurement scan of the form

$$\mathbf{Z} := (Z^1, \dots, Z^T), \quad Z^t := (z_1^t, \dots, z_{n_t}^t), \quad z_r^t \in \mathcal{O} \subset \mathbb{R}^{n_Z},$$

for $1 \le t \le T$, $1 \le r \le n_t$. Note, that the number $n_t$ of observations (measurements) at time $t$ may differ from the number $M$ of targets. Moreover, the measurement dimension $n_Z$ may also differ from the state dimension $n_X$ and $\mathcal{O} \subset \mathbb{R}^{n_Z}$ is the compact observation window. The measurement space is then given as

$$\mathbb{Z} := \bigtimes_{t=1}^{T} \bigtimes_{r=1}^{n_t} \mathcal{O} = \mathcal{O}^{N_T} \subset \mathbb{R}^{N_Z}, \quad N_T := \sum_{t=1}^{T} n_t, \; N_Z := N_T \cdot n_Z.$$

The connection between the observed data and the state is given by the assignment which consists of two parts, namely the assignment probabilities and the assignment of given observations. The assignment probabilities read

$$\mathbf{\Pi} := (\Pi^1, \dots, \Pi^T), \quad \Pi^t := (\pi_1^t, \dots, \pi_M^t), \quad \pi_m^t \in [0,1],$$

where $\pi_m^t$ denotes the (*a priori*) probability that a measurement at time $t$ is associated to target $m$. The assignment probability space is

$$\mathbb{\Pi} := \bigtimes_{t=1}^{T} \bigtimes_{m=1}^{M} [0,1] = [0,1]^{TM}, \quad N_\Pi := T \cdot M.$$

Finally, the assignment of given observations read

$$\mathbf{K} := (K^1, \dots, K^T), \quad K^t := (k_1^t, \dots, k_{n_t}^t), \quad k_r^t \in \{1, \dots, M\},$$

where $k_r^t = m$ means that the measurement $z_r^t$ corresponds to object $m$, $1 \le m \le M$, $1 \le r \le n_t$, $1 \le t \le T$. The assignment space is then

$$\mathbb{K} := \bigtimes_{t=1}^{T} \{1, \dots, M\}^{n_t}, \qquad N_K := N_T.$$

Let us now interpret these four quantities in terms of our notation introduced in Section 2. Of course, $\mathbf{X}$ contains all information that is relevant for tracking, so that this could be interpreted as the complete data. However, since we are aiming at determining an ML estimate for $\mathbf{X}$, it makes more sense to view $\mathbf{X}$ as a parameter. Moreover, the assignment probabilities $\mathbf{\Pi}$ obviously influence the Likelihood. Thus, often the couple $(\mathbf{X}, \mathbf{\Pi})$ is interpreted as parameter. For simplicity, we will assume that a model for $\mathbf{\Pi}$ is available, so that $\mathbf{X}$ is the parameter. The observed data is obviously $\mathbf{Z}$ and the missing data is the assignment $\mathbf{K}$. If $\mathbf{Z}$ and $\mathbf{K}$ are known, we would be able to reconstruct the full data, so that $(\mathbf{Z}, \mathbf{K})$ is the full data. In summary, we have

$$\begin{aligned}
\phi &\equiv \mathbf{X} \in \Phi \equiv \mathbb{X}, \\
x &\equiv (\mathbf{Z}, \mathbf{K}) \in \mathbb{D} \equiv \mathbb{Z} \times \mathbb{K}, \quad X : \Omega \to \mathbb{D}, m = N_X, \\
y &\equiv \mathbf{Z} \in \mathbb{O} \equiv \mathbb{Z}, \qquad\qquad\quad Y : \Omega \to \mathbb{O}, n = N_Z, \\
z &\equiv \mathbf{K} \in \mathbb{M} \equiv \mathbb{K}, \\
\ell(\mathbf{Z}, \mathbf{K}) &= \mathbf{Z}, \quad \ell^{-1}(\mathbf{Z}) = \{(\mathbf{Z}, \mathbf{K}) \in \mathbb{D} : \mathbf{K} \in \mathbb{K}\}.
\end{aligned}$$

We can already remark here, that for $M > 0$ and any observation batch $\mathbf{Z}$ with $N_T > 0$ the set $\ell^{-1}(\mathbf{Z})$ has positive counting measure. Thus, Assumption B is met. The restrictions $M > 0$ and $N_T > 0$ are without relevance in practice since a multi-target tracking problem without measurements or tracks is irrelevant.

Next, we describe the PMHT model for the pdf of the random variables $Z : \Omega \ni \omega \mapsto \mathbf{Z} = \mathbf{Z}(\omega)$, $K : \Omega \ni \omega \mapsto \mathbf{K} = \mathbf{K}(\omega)$ ($(Z, K)$ corresponding to $X$ in Section 2), namely (note that we omit the superscript indices for convenience)

(5.1)

$$\begin{aligned}
&f_\phi^X(x) \equiv f_{\mathbf{X}}(\mathbf{Z}, \mathbf{K}) \\
&\quad = \left\{ \prod_{\nu=1}^{M} \varphi_\nu^0(x_\nu^0) \right\} \prod_{t=1}^{T} \left\{ \left[ \prod_{s=1}^{M} \varphi_s^t(x_s^t | x_s^{t-1}) \right] \prod_{r=1}^{n_t} \left[ \pi_m^t \, \zeta_m^t(z_r^t | x_m^t) \, |_{m=k_r^t} \right] \right\},
\end{aligned}$$

where the following quantities are given by the specific chosen movement, measure and initial model (see Section 5.3 below):

- $\varphi_\nu^0$ denotes the *a priori* pdf for the initial state of target $\nu$,
- $\varphi_s^t(x|\tilde{x})$, $1 \leq t \leq T$, denotes the transition density of target $s$ at time $t$ to move from $\tilde{x} \in \mathcal{X}$ to $x \in \mathcal{X}$,
- $\zeta_s^t(z|x)$ is the measurement pdf that $z \in \mathcal{O}$ at time $t$ is originated from object $s$ with state $x \in \mathcal{X}$.

For details on the properties of $f_{\mathbf{X}}(\mathbf{Z}, \mathbf{K})$ and its practical meaning, we refer to [24, 26]. We shortly note, however, that its domain $\mathbb{D} \equiv \mathbb{Z} \times \mathbb{K}$ is compact as the observation space $\mathcal{O}$ is compact and $\mathbb{K}$ is finite. Hence, Assumption A and C hold true for $f_{\mathbf{X}}(\mathbf{Z}, \mathbf{K})$ whenever they hold true for all individual initial state, transition and measurement densities. This is e.g. the case for a linear Gaussian tracking model, where the individual densities are assumed to be non-degenerated Gaussian pdf's which are restricted to the compact spaces $\mathcal{X} \subset \mathbb{R}^{n_X}$ and $\mathcal{O} \subset \mathbb{R}^{n_Z}$, resp., and normalized such that their integrals over $\mathcal{X}$ or $\mathcal{O}$, resp., are unity.

Next, we get

$$\begin{aligned}
g_\phi^Y(y) &\equiv g_{\mathbf{X}}(\mathbf{Z}) = \int_{\ell^{-1}(\mathbf{Z})} f_{\mathbf{X}}(\mathbf{Z}, \mathbf{K}) \, d(\mathbf{Z}, \mathbf{K}) = \sum_{\mathbf{K} \in \mathbb{K}} f_{\mathbf{X}}(\mathbf{Z}, \mathbf{K}) \\
&= \left\{ \prod_{\nu=1}^{M} \varphi_\nu^0(x_\nu^0) \right\} \prod_{t=1}^{T} \left\{ \left[ \prod_{s=1}^{M} \varphi_s^t(x_s^t | x_s^{t-1}) \right] \prod_{r=1}^{n_t} \sum_{m=1}^{M} \pi_m^t \, \zeta_m^t(z_r^t | x_m^t) \right\}.
\end{aligned}$$

Finally,

$$\begin{aligned} h_\phi^{X|Y}(x|y) \;\; &\equiv \;\; h_{\mathbf{X}}(\mathbf{Z},\mathbf{K}|\mathbf{Z}) = \frac{f_{\mathbf{X}}(\mathbf{Z},\mathbf{K})}{g_{\mathbf{X}}(\mathbf{Z})} \\ &= \;\; \prod_{t=1}^{T}\prod_{r=1}^{n_t} \frac{\pi^t_{k^t_r}\,\zeta^t_{k^t_r}(z^t_r|x^t_{k^t_r})}{\sum_{m=1}^M \pi^t_m\,\zeta^t_m(z^t_r|x^t_m)} = \prod_{t=1}^{T}\prod_{r=1}^{n_t} w^t_{k^t_r,r}, \end{aligned}$$

where the weights

$$w^t_{s,r} := \frac{\pi^t_s\,\zeta^t_s(z^t_r|x^t_s)}{\sum_{m=1}^M \pi^t_m\zeta^t_m(z^t_r|x^t_m)} \in [0,1] \tag{5.2}$$

can be interpreted as the a posteriori probabilities that a measurement $z^t_r$ is assigned to the target $s$ conditioned on all measurements and target states, i.e.,

$$w^t_{s,r} = \mathsf{P}(k^t_r(\omega) = s|\mathbf{Z},\mathbf{X}). \tag{5.3}$$

Here $k^t_r(\omega)$ denotes the corresponding component of the random variable $K : \Omega \to \mathbb{K}$. The log-Likelihood function which shall be optimized with respect to $\mathbf{X}$ is then given by

$$\begin{aligned} L(\mathbf{X}) \;\equiv\; L(\phi) &= \log g_\phi^Y(y) \\ &= \left\{\sum_{\nu=1}^{M} \log\left(\varphi^0_\nu(x^0_\nu)\right)\right\} + \sum_{t=1}^{T}\left\{\left[\sum_{s=1}^{M}\log\left(\varphi^t_s(x^t_s|x^{t-1}_s)\right)\right]\right. \\ &\qquad\qquad \left. + \sum_{r=1}^{n_t}\left[\log\left(\sum_{m=1}^{M}\pi^t_m\,\zeta^t_m(z^t_r|x^t_m)\right)\right]\right\}. \end{aligned} \tag{5.4}$$

**Remark 5.1.** *If we would consider the general case of unknown assignment probabilities* $\mathbf{\Pi}$ *and thus a parameter* $(\mathbf{X},\mathbf{\Pi})$, *we would get additional constraints, namely*

$$0 \le \pi^t_s \le 1, \qquad \sum_{s=1}^{M}\pi^t_s = 1.$$

*Thus, this would lead to a constraint optimization problem. In order to avoid such constraints, one could e.g. consider a reparameterization with the "softmax" function* $\pi^t_s = \exp(\gamma^t_s)\big(\sum_{m=1}^M \exp(\gamma^t_m)\big)^{-1}$ *with parameters* $\gamma^t_m \in \mathbb{R}$.

*As already mentioned above, we consider the case of a given* $\mathbf{\Pi}$. *One major reason for this is the fact that this setting allows for a comparison of the EM algorithm with a standard Newton's method. We will come to that point later.*

Next, we develop expressions for $Q(\mathbf{X},\tilde{\mathbf{X}})$ and $H(\mathbf{X},\tilde{\mathbf{X}})$ defined in (2.5) and (2.8), respectively, for the specific case of PMHT. First, we obtain

$$\begin{aligned} Q(\mathbf{X},\tilde{\mathbf{X}}) \;\; &= \;\; \int_{\ell^{-1}(\mathbf{Z})} \big(\log\, f_{\mathbf{X}}(\mathbf{Z},\mathbf{K})\big) h_{\tilde{\mathbf{X}}}(\mathbf{Z},\mathbf{K}|\mathbf{Z})\, d(\mathbf{Z},\mathbf{K}) \\ &= \;\; \sum_{\mathbf{K}\in\mathbb{K}} \big(\log\, f_{\mathbf{X}}(\mathbf{Z},\mathbf{K})\big)\, h_{\tilde{\mathbf{X}}}(\mathbf{Z},\mathbf{K}|\mathbf{Z}). \end{aligned}$$

As $h_{\tilde{\mathbf{X}}}(\mathbf{Z},\mathbf{K}|\mathbf{Z})$ is a probability measure for the discrete random variable $\mathbf{K}$, we have $\sum_{\mathbf{K}\in\mathbb{K}} h_{\tilde{\mathbf{X}}}(\mathbf{Z},\mathbf{K}|\mathbf{Z}) = 1$. By setting the weights in (5.2) and (5.3) for $\tilde{\mathbf{X}}$ as

$\tilde{w}^t_{s,r} := \mathsf{P}(k^t_r(\omega) = s|\mathbf{Z}, \tilde{\mathbf{X}})$, we obtain

$$\sum_{\mathbf{K}\in\mathbb{K}^t_{s,r}} h_{\tilde{\mathbf{X}}}(\mathbf{Z}, \mathbf{K}|\mathbf{Z}) = \mathsf{P}(k^t_r(\omega) = s|\mathbf{Z}, \tilde{\mathbf{X}}) = \tilde{w}^t_{s,r}, \tag{5.5}$$

where $\mathbb{K}^t_{s,r} := \{\mathbf{K} \in \mathbb{K} : k^t_r = s\}$. Using these identities we arrive at

$$Q(\mathbf{X}, \tilde{\mathbf{X}}) = \sum_{t=1}^{T} Q_{\Pi^t} \ + \sum_{s=1}^{M} Q_{X_s} \ , \tag{5.6}$$

where $Q_{\Pi^t} := \sum_{r=1}^{n_t}\sum_{s=1}^{M} \tilde{w}^t_{s,r} \log \pi^t_s$ and $Q_{X_s} := \log(\varphi^0_s(x^0_s)) + \sum_{t=1}^{T}\{\log(\varphi^t_s(x^t_s|x^{t-1}_s)) + \sum_{r=1}^{n_t} w^t_{s,r} \log(\zeta^t_s(z^t_r|x^t_s))\}$. For details on the derivation of (5.6) see [26]. As $Q_{\Pi^t}$ is independent of $\mathbf{X}$ and $Q_{X_s}$ only depends on $X_s$, the maximization problem in the M-step decouples into $M$ independent maximization problems for each target

$$\hat{X}_s = \arg\max_{X_s} Q_{X_s}, \quad \text{for } s = 1, \dots, M. \tag{5.7}$$

The ML estimate $\hat{X}_s$ can hence be obtained from the systems of equations $\frac{d}{dX_s} Q_{X_s} = 0$. For some tracking models, the optimization problem (5.7) can even be solved analytically. As an example, for a linear Gaussian tracking model the solution is given by the so-called RTS formulas [1, 19] employed with a synthetic measurement (see [26]).

Next, we obtain

$$\begin{aligned} H(\mathbf{X}, \tilde{\mathbf{X}}) &= \int_{\ell^{-1}(\mathbf{Z})} \big(\log h_{\mathbf{X}}(\mathbf{Z}, \mathbf{K})\big) h_{\tilde{\mathbf{X}}}(\mathbf{Z}, \mathbf{K}|\mathbf{Z})\, d(\mathbf{Z}, \mathbf{K}) \\ &= \sum_{\mathbf{K}\in\mathbb{K}} \big(\log h_{\mathbf{X}}(\mathbf{Z}, \mathbf{K})\big)\, h_{\tilde{\mathbf{X}}}(\mathbf{Z}, \mathbf{K}|\mathbf{Z}) \\ &= \sum_{\mathbf{K}\in\mathbb{K}} \Big(\sum_{t=1}^{T}\sum_{r=1}^{n_t} \log(w^t_{k^t_r,r})\Big) h_{\tilde{\mathbf{X}}}(\mathbf{Z}, \mathbf{K}|\mathbf{Z}). \end{aligned}$$

Similar to the derivation of $Q$ in [26], we rearrange the sum over $\mathbf{K} \in \mathbb{K}$ as follows

$$\begin{aligned} H(\mathbf{X}, \tilde{\mathbf{X}}) &= \sum_{t=1}^{T}\sum_{r=1}^{n_t}\sum_{\mathbf{K}\in\mathbb{K}} \log(w^t_{k^t_r,r})\, h_{\tilde{\mathbf{X}}}(\mathbf{Z}, \mathbf{K}|\mathbf{Z}) \\ &= \sum_{t=1}^{T}\sum_{r=1}^{n_t}\sum_{s=1}^{M} \log(w^t_{s,r}) \sum_{\mathbf{K}\in\mathbb{K}^t_{s,r}} h_{\tilde{\mathbf{X}}}(\mathbf{Z}, \mathbf{K}|\mathbf{Z}), \end{aligned}$$

so that using the identity (5.5) once more we obtain

$$H(\mathbf{X}, \tilde{\mathbf{X}}) = \sum_{s=1}^{M}\sum_{t=1}^{T}\sum_{r=1}^{n_t} \tilde{w}^t_{s,r}\ \log(w^t_{s,r}). \tag{5.8}$$

Now, we are going to determine the derivatives. Recall the above notation

**Proposition 5.2.** (a) *If* $\zeta^t_s(z|\cdot) \in C^1(\mathcal{X})$, *then* $\frac{\partial}{\partial x^0_s} H(\mathbf{X}, \tilde{\mathbf{X}}) = 0$ *and for* $t > 0$ *we have* $\frac{\partial}{\partial x^t_s} H(\boldsymbol{X}, \tilde{\mathbf{X}}) = \ \sum_{r=1}^{n_t} \Big\{ \big(-w^t_{s,r} + \ \tilde{w}^t_{s,r}\big) \cdot \frac{\partial}{\partial x^t_s} \log\big(\zeta^t_s\,(z^t_r|x^t_s)\big) \Big\}.$

(b) *If* $\zeta_s^t(z|\cdot) \in C^2(\mathcal{X})$, *then*

$$D^{20}H(\boldsymbol{X}, \tilde{\mathbf{X}}) = \begin{pmatrix} H_{1,1} & H_{1,2} & \cdots & H_{1,M} \\ H_{2,1} & H_{2,2} & \cdots & H_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ H_{M,1} & H_{M,2} & \cdots & H_{M,M} \end{pmatrix}, \tag{5.9}$$

*with* $H_{s,m} = \operatorname{diag}(H_{s,m;1}, \dots, H_{s,m;T})$, *where* $H_{s,m;0} = 0_{n_x \times n_x}$ *and* $H_{s,m;t}$, *for* $t = 1, \dots, T$, *are* $(n_x \times n_x)$ *- matrices given by*

$$\begin{aligned} H_{s,m;t} = \sum_{r=1}^{n_t} \Bigg\{ &- w_{s,r}^t \left(\delta_{s,m} - w_{m,r}^t\right) \cdot \left[\frac{\partial}{\partial x_s^t} \log\left(\zeta_s^t\left(z_r^t | x_s^t\right)\right)\right] \\ &\times \left[\frac{\partial}{\partial x_m^t} \log\left(\zeta_m^t\left(z_r^t | x_m^t\right)\right)\right]^T + \delta_{s,m} \cdot \left(\tilde{w}_{s,r}^t - w_{s,r}^t\right) \cdot \frac{\partial^2}{\partial x_s^t \, \partial x_s^t} \log\left(\zeta_s^t\left(z_r^t | x_s^t\right)\right) \Bigg\}. \end{aligned} \tag{5.10}$$

*Proof.* (a) Obviously, we have $D^{10}H(\mathbf{X}, \tilde{\mathbf{X}}) = \sum_{s=1}^M \sum_{t=1}^T \sum_{r=1}^{n_t} \tilde{w}_{s,r}^t \frac{d}{d\mathbf{X}} \log\left(w_{s,r}^t\right)$ and

$$\frac{\partial}{\partial x_{s'}^{t'}} \log\left(w_{s,r}^t\right) = \frac{\partial}{\partial x_{s'}^{t'}} \log\left(\frac{\pi_s^t \, \zeta_s^t\left(z_r^t | x_s^t\right)}{\sum_{m=1}^M \pi_m^t \, \zeta_m^t\left(z_r^t | x_m^t\right)}\right). \tag{5.11}$$

Since $w_{s,r}^t$ only depends on the values of $x_{s'}^t$ for $s' = 1, \dots, M$, all partial derivatives $\frac{\partial}{\partial x_{s'}^{t'}} \log\left(w_{s,r}^t\right)$ with $t' = 0$ or $t' \neq t$ vanish. For the partial derivatives with $t' = t$ we have to distinguish between the cases $s' = s$ and $s' \neq s$, which is done by means of the Kronecker delta $\delta_{s',s}$:

$$\begin{aligned} &\frac{\partial}{\partial x_{s'}^t} \log\left(w_{s,r}^t\right) = \\ &= \frac{1}{w_{s,r}^t} \left( \delta_{s',s} \cdot \frac{\pi_s^t \left(\frac{\partial}{\partial x_s^t} \zeta_s^t\left(z_r^t | x_s^t\right)\right) \cdot \left(\sum_{m=1}^M \pi_m^t \, \zeta_m^t\left(z_r^t | x_m^t\right)\right)}{\left(\sum_{m=1}^M \pi_m^t \, \zeta_m^t\left(z_r^t | x_m^t\right)\right)^2} \right. \\ &\qquad \left. - \frac{\pi_s^t \, \zeta_s^t\left(z_r^t | x_s^t\right) \cdot \pi_{s'}^t \left(\frac{\partial}{\partial x_{s'}^t} \zeta_{s'}^t\left(z_r^t | x_{s'}^t\right)\right)}{\left(\sum_{m=1}^M \pi_m^t \, \zeta_m^t\left(z_r^t | x_m^t\right)\right)^2} \right) \\ &= \left(\delta_{s',s} - w_{s',r}^t\right) \cdot \frac{\partial}{\partial x_{s'}^t} \log\left(\zeta_{s'}^t\left(z_r^t | x_{s'}^t\right)\right). \end{aligned}$$

This results in the following expression for $t > 0$:

$$\begin{aligned} \frac{\partial}{\partial x_{s'}^t} H(\mathbf{X}, \tilde{\mathbf{X}}) &= \sum_{s=1}^M \sum_{r=1}^{n_t} \left[\tilde{w}_{s,r}^t \left(\delta_{s's} - w_{s',r}^t\right) \cdot \frac{\partial}{\partial x_{s'}^t} \log\left(\zeta_{s'}^t\left(z_r^t | x_{s'}^t\right)\right)\right] \\ &= \sum_{r=1}^{n_t} \left\{ \frac{\partial}{\partial x_{s'}^t} \log\left(\zeta_{s'}^t\left(z_r^t | x_{s'}^t\right)\right) \left[-w_{s',r}^t \sum_{s=1}^M \tilde{w}_{s,r}^t + \tilde{w}_{s',r}^t\right] \right\} \\ &= \sum_{r=1}^{n_t} \left\{ \left(-w_{s',r}^t + \tilde{w}_{s',r}^t\right) \cdot \frac{\partial}{\partial x_{s'}^t} \log\left(\zeta_{s'}^t\left(z_r^t | x_{s'}^t\right)\right) \right\}, \end{aligned} \tag{5.12}$$

since $\sum_{s=1}^M \tilde{w}_{s,r}^t = 1$. For $t = 0$ holds $\frac{\partial}{\partial x_{s'}^0} H(\mathbf{X}, \tilde{\mathbf{X}}) = 0$.

(b) Again, all second order partial derivatives with $t' = 0$ or $t' \neq t$ vanish. Next,

$$\begin{aligned}\frac{\partial^2}{\partial x_s^t\,\partial x_{s'}^t}\,H(\mathbf{X},\tilde{\mathbf{X}}) =\ & \left(\sum_{r=1}^{n_t} \frac{\partial}{\partial x_s^t}\left\{\left(-w_{s',r}^t +\ \tilde{w}_{s',r}^t\right)\cdot\left[\frac{\partial}{\partial x_{s'}^t}\log\left(\zeta_{s'}^t\left(z_r^t|x_{s'}^t\right)\right)\right]\right\}\right)^T \\ =\ & \sum_{r=1}^{n_t}\left\{-\frac{\partial}{\partial x_s^t}\,w_{s',r}^t\cdot\left[\frac{\partial}{\partial x_{s'}^t}\log\left(\zeta_{s'}^t\left(z_r^t|x_{s'}^t\right)\right)\right]^T\right. \\ & \left. +\left(-w_{s',r}^t +\ \tilde{w}_{s',r}^t\right)\cdot\frac{\partial^2}{\partial x_s^t\,\partial x_{s'}^t}\log\left(\zeta_{s'}^t\left(z_r^t|x_{s'}^t\right)\right)\right\} \\ =\ & \sum_{r=1}^{n_t}\left\{-w_{s,r}^t\left(\delta_{s,s'} - w_{s',r}^t\right)\cdot\left[\frac{\partial}{\partial x_s^t}\log\left(\zeta_s^t\left(z_r^t|x_s^t\right)\right)\right]\right. \\ & \times\left[\frac{\partial}{\partial x_{s'}^t}\log\left(\zeta_{s'}^t\left(z_r^t|x_{s'}^t\right)\right)\right]^T + \delta_{ss'}\cdot\left(\tilde{w}_{s,r}^t - w_{s,r}^t\right) \\ & \left.\cdot\frac{\partial^2}{\partial x_s^t\,\partial x_s^t}\log\left(\zeta_s^t\left(z_r^t|x_s^t\right)\right)\right\}.\end{aligned} \tag{5.13}$$

This proves our claim. □

5.2. **Convergence analysis.** Now, we investigate the convergence of the EM algorithm for the PMHT problem in a specific tracking scenario. It will turn out that EM converges fast, if the observed targets are 'well separated', whereas the convergence of gradient ascent might be rather slow. This corresponds to observations in [28] for the parameter estimation of Gaussian mixtures.

As we already remarked before, Assumption B is always met in the PMHT context and the Assumptions A and C hold true whenever they hold true for the individual initial, transition and measurement densities of a tracking model which is e.g. the case for the linear Gaussian PMHT. Moreover, we restrict our tracking model to one where the second assumption of Proposition 3.5 is also met, i.e. $\|(D^{20}Q(\mathbf{X},\mathbf{X}))^{-1}\| < \infty$. Hence, the EM algorithm reads in the PMHT context

$$\mathbf{X}^{(k+1)} =\ \mathbf{X}^{(k)} + \eta P(\mathbf{X}^{(k)})\nabla L(\mathbf{X}^{(k)}) + o(\|\mathbf{X}^{(k+1)} - \mathbf{X}^{(k)}\|), \tag{5.14}$$

where $P(\mathbf{X}^{(k)}) = -\Big(\eta D^{20}Q(\mathbf{X}^{(k)},\mathbf{X}^{(k)})\Big)^{-1}$. Using a linear Gaussian motion and measurement model for PMHT, the assumptions of Theorem 3.4 are met and (5.14) is exact:

$$\mathbf{X}^{(k+1)} =\ \mathbf{X}^{(k)} + \eta P(\mathbf{X}^{(k)})\nabla L(\mathbf{X}^{(k)}).$$

From (3.15), we know that $P(\mathbf{X}^*) = (\eta\mathcal{I}_{\text{full}}(\mathbf{X}^*))^{-1}$ is s.p.d. and we can apply the convergence analysis of Section 4. Thus, we consider the ratio of biggest and smallest eigenvalues of $-P(\mathbf{X}^*)\,\nabla^2 L(\mathbf{X}^*)$ in order to investigate the convergence speed. Since we know by Lemma 2.7 2.7 that

$$P(\mathbf{X}^*) = -\eta^{-1}\big(D^{20}Q(\mathbf{X}^*,\mathbf{X}^*)\big)^{-1} = -\eta^{-1}\Big(\nabla^2 L(\mathbf{X}^*) + D^{20}H(\mathbf{X}^*,\mathbf{X}^*)\Big)^{-1}, \tag{5.15}$$

we will take a closer look at $D^{20}H(\mathbf{X}^*,\mathbf{X}^*)$ for PMHT, see Proposition 5.2 (b).

First of all, we note that since we have $\mathbf{X} = \tilde{\mathbf{X}} \equiv \mathbf{X}^*$, the second term of $H_{s,m;t}$ in (5.10) vanishes as $\tilde{w}_{s,r}^t = w_{s,r}^t \equiv w_{s,r}^{t*}$. Furthermore, we consider a tracking scenario with $M$ 'well separated' targets and without false alarms, i.e. measurements that are not target originated. Note, that to some extend, the notion 'well separated' is somewhat vague since it depends on the tracking model, the measurement model and an appropriate distance measure. We will report a quantitative investigation in Section 5.3 below. In the specified situation, we can assume that each obtained measurement can easily be assigned to one particular target. Thus, since the assignment weight $w_{s,r}^{t*}$ is the a posteriori probability of assigning observation $r$ of scan $t$ to target $s$, we can assume that for every index combination $(r,t)$ there exists exactly one index $s$ such that $w_{s,r}^{t*} \approx 1$ and for all other indices $m \neq s$, $w_{m,r}^{t*} \approx 0$. Consequently, the products $-w_{s,r}^{t*}\left(1-w_{s,r}^{t*}\right)$ and $-w_{s,r}^{t*} w_{m,r}^{t*}$, for $m \neq s$, are both approximately zero and the Hessian $D^{20}H(\mathbf{X}^*, \mathbf{X}^*)$ is approximately a zero matrix. This, in turn, leads to $P(\mathbf{X}^*)$ being approximately equal to the negative inverse of the Hessian of the log-Likelihood multiplied by $\eta^{-1}$ (cf. (5.15)). The product $P(\mathbf{X}^*)\,\nabla^2 L(\mathbf{X}^*)$ is thus approximately equal to $-\eta^{-1}$ times the identity matrix and

$$\frac{\lambda_{\max}\left(-P(\mathbf{X}^*)\,\nabla^2 L(\mathbf{X}^*)\right)}{\lambda_{\min}\left(-P(\mathbf{X}^*)\,\nabla^2 L(\mathbf{X}^*)\right)} \approx \frac{\lambda_{\max}\left(\eta^{-1}I\right)}{\lambda_{\min}\left(\eta^{-1}I\right)} = 1. \tag{5.16}$$

The first ratio being close to one implies superlinear convergence if the step size of each iteration is chosen in an optimal way corresponding to $\eta_{\text{opt}}$ of (4.4). Yet, since

$$\begin{aligned} \eta_{\text{opt}} &= \frac{2}{\lambda_{\min}\left(-P(\mathbf{X}^*)\,\nabla^2 L(\mathbf{X}^*)\right) + \lambda_{\max}\left(-P(\mathbf{X}^*)\,\nabla^2 L(\mathbf{X}^*)\right)} \\ &\approx \frac{2}{\lambda_{\min}\left(\eta^{-1}I\right) + \lambda_{\max}\left(\eta^{-1}I\right)} = \eta \end{aligned} \tag{5.17}$$

for the described tracking situation, the step size of the EM algorithm is close to the optimal choice. Thus, the EM algorithm resembles Newton's method for such a scenario and possesses locally quadratic convergence.

For gradient ascent on contrast, we have shown that the convergence speed depends on the condition number of $-\nabla^2 L(\mathbf{X}^*)$ which can be pretty large, resulting in slow convergence even for unambiguous situations of 'well separated' targets.

### 5.3. **Numerical Experiments.**

Finally, we want to quantify our above analysis by some numerical experiments. In order to do so, we consider the number of iterations until convergence for a tracking scenario of several objects moving parallel to each other with a fixed distance and the same velocity. The distance varies from 100 $m$ to 1 $km$ in order to quantify the notion of 'well separated targets'. The targets move with a constant speed of $v_0 = 200\, m/s$ and the time period between two consecutive scans is set to $\Delta t = 2\, s$. For the filter, we assume a linear Gaussian motion and measurement model:

$$x_s^{t+1} = F_s^t\, x_s^t + v_s^t, \qquad z_r^t = \ H_s^t\, x_s^t + w_s^t,$$

where the discrete target states consist of two-dimensional Cartesian position data, $(x,y)$, and velocity data, $(\dot{x},\dot{y})$. The elements of a target's state are ordered as $x_s^t = (x,\dot{x},y,\dot{y})^T \in \mathbb{R}^{n_x}$ with $n_x = 4$. The measurements are position-only, i.e. $z_r^t \in \mathbb{R}^{n_z}$ with $n_z = 2$. The state transition matrices, $F_s^t$, and the measurement matrices, $H_s^t$, are given as

$$F_s^t \equiv F = \begin{pmatrix} 1 & \Delta t & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \Delta t \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad H_s^t \equiv H = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}. \tag{5.18}$$

In our model, $v_s^t$ and $w_s^t$ are mutually independent zero-mean, white Gaussian process and measurement noise, respectively, with covariances $Q_s^t$ and $R_s^t$. The process noise $v_s^t$ is assumed to reflect white noise acceleration. The covariance matrices $Q_s^t$ are thus given by [2]

$$Q_s^t \equiv Q = \tilde{q} \cdot \begin{pmatrix} \frac{\Delta t^3}{3} & \frac{\Delta t^2}{2} & 0 & 0 \\ \frac{\Delta t^2}{2} & \Delta t & 0 & 0 \\ 0 & 0 & \frac{\Delta t^3}{3} & \frac{\Delta t^2}{2} \\ 0 & 0 & \frac{\Delta t^2}{2} & \Delta t \end{pmatrix}. \tag{5.19}$$

The factor $\tilde{q}$ is due to the process noise intensity and often assumed to be a tuning parameter, [4]. The measurement covariance matrices are chosen to be

$$R_s^t \equiv R = \begin{pmatrix} \sigma_x^2 & \frac{1}{2}\sigma_x\sigma_y \\ \frac{1}{2}\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}. \tag{5.20}$$

In our simulations we use $\sigma_x = 30\ m$, $\sigma_y = 100\ m$ and $\tilde{q} = 0.5\ sg^2$, where $g \approx 9.81\ m/s^2$ is the value of the gravitational acceleration. This model yields

$$\varphi_s(x_s^{t+1}|x_s^t) = \mathcal{N}(x_s^{t+1}; F_s^t x_s^t, Q_s^t), \qquad \zeta_s^t(z_r^t|x_s^t) = \mathcal{N}(z_r^t; H_s^t x_s^t, R_s^t), \tag{5.21}$$

where $\mathcal{N}(x;\mu,\Sigma)$ is the Gaussian density in variable $x$ with mean $\mu$ and covariance $\Sigma$. Furthermore, we assume a Gaussian distribution for the initial target states which yields $\varphi_s^0(x_s^0) = \mathcal{N}(x_s^0; \bar{x}_{0s}, \bar{\Sigma}_{0s})$. For a random (but realistic) initialization of $\{\bar{x}_{0s}, \bar{\Sigma}_{0s}\}$, simulated measurements of four data scans are used. The measurements of the first scan are employed for a one-point initialization [3, 16]. The correctly assigned simulated measurements of the next three scans are then used to perform Kalman updates [11] to improve track quality. The state and covariance estimates resulting from this procedure constitute the initial track estimates.

In our numerical experiments we process a measurement batch of length $T = 3$. We hence simulate measurements for another three data scans. The measurements in each scan are generated from the true target states of $M = 4$ targets using the above given measurement model together with a detection probability of $p_D = 0.9$. False alarms are not assumed to be present. The values of the prior assignment probabilities are set to $\pi_s^t \equiv \frac{1}{M}$ for all targets $s = 1, \ldots, M$.

The iterations of the algorithm are stopped if either the stopping criterion of [27] averaged over all targets is less than a given tolerance $\varepsilon_1$:

$$\frac{1}{TM} \sum_{s=1}^{M} \sum_{t=1}^{T} (x_s^{t\,(i)} - x_s^{t\,(i-1)})^T (Q_s^t)^{-1} (x_s^{t\,(i)} - x_s^{t\,(i-1)}) < \varepsilon_1, \tag{5.22}$$

or if the Euclidean norm of the gradient of the log-Likelihood at the current iterate falls below a given threshold $\varepsilon_2$:

$$\|\nabla L(\mathbf{X}^{(i)})\|_2 < \varepsilon_2. \tag{5.23}$$

The values of $\varepsilon_1$ and $\varepsilon_2$ are chosen as $10^{-5}$ and $10^{-2}$, respectively.

For gradient ascent or Newton's method, the gradient has to be computed anyway. A check of its magnitude is important to avoid numerical instabilities. In our example, the gradient can also be easily obtained as a by-product of the EM iteration. A check of the second stopping criterion hence does not produce extra cost but might recognize convergence at an early stage avoiding superfluous iterations.

For the evaluation of the number of iterations we simulate data for 100 Monte-Carlo (MC) runs for each scenario with a different target distance. The input data are then processed by (a) the standard PMHT algorithm using the EM algorithm, (b) by a hybrid Newton-EM method and (c) by a conjugate-gradient ascent algorithm. For Newton's method and the conjugate-gradient ascent algorithm, a line search is performed in order to find a step size that ensures an increase in the log-Likelihood. The EM step is known to always increase the Likelihood. For stability reasons, Newton's method is started with two iterations of the EM algorithm and switches to an EM step whenever the Hessian is not negative definite. That is why we call it Newton-EM method. A combination of Newton's method with gradient ascent was also investigated, but showed a desperate slow convergence due to frequent switches to gradient ascent at the beginning.

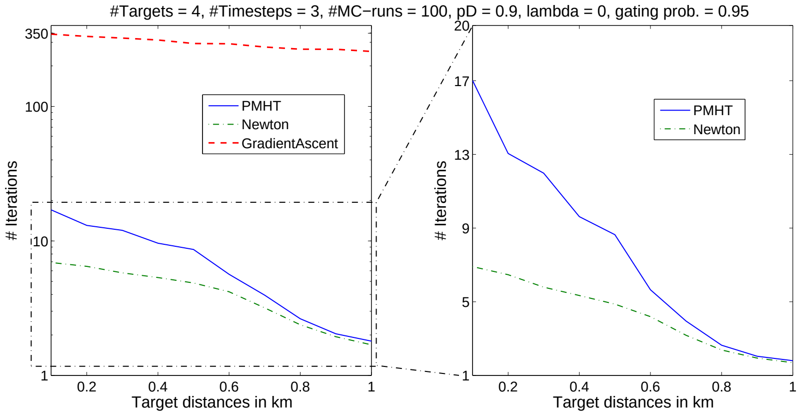

Figure 5.1. Mean iteration number as a function of target distances.

In Figure 5.1, the results of the described simulations are shown. The mean iteration number of each of the above mentioned algorithms is plotted as a function of target distances. In the left plot, it becomes evident that the number of iterations of conjugate-gradient ascent algorithm is by far larger than for the other two algorithms even for 'well separated' targets. For a better comparison of the EM algorithm with Newton's method we show an excerpt of these two methods in the right plot. For large target distances, the required iteration numbers of both methods is small and nearly the same. With decreasing distance the number of iterations increases for both methods and a larger increase is observed for the EM algorithm as compared with Newton's method.

In Figures 5.2 and 5.3, two MC runs are compared in more detail. The first one is a simulation with target distance 0.3 $km$ and the second one with 0.6 $km$. The plots in the upper row of both figures illustrate the learning curves and the

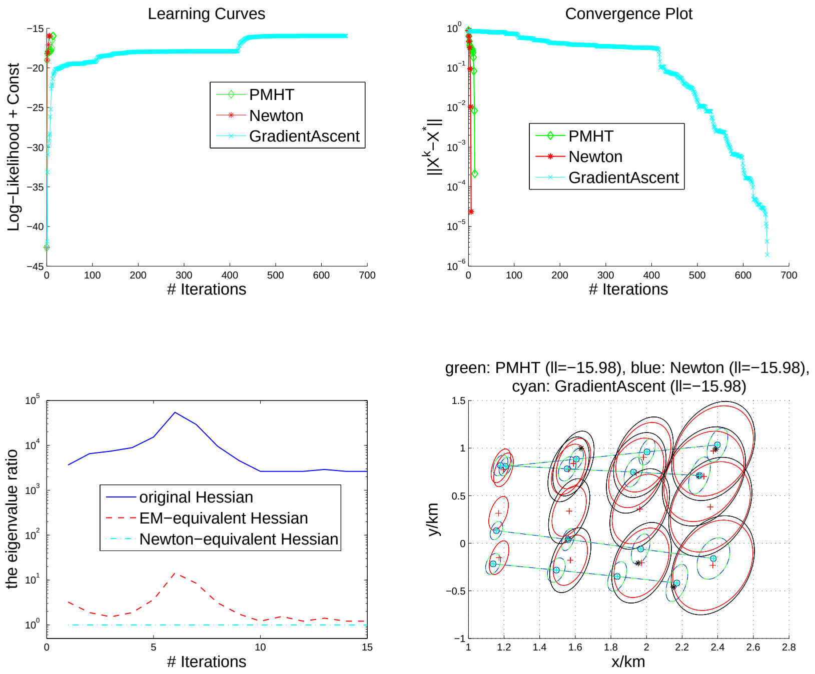

FIGURE 5.2. Single MC run with target distance 0.3 $km$.

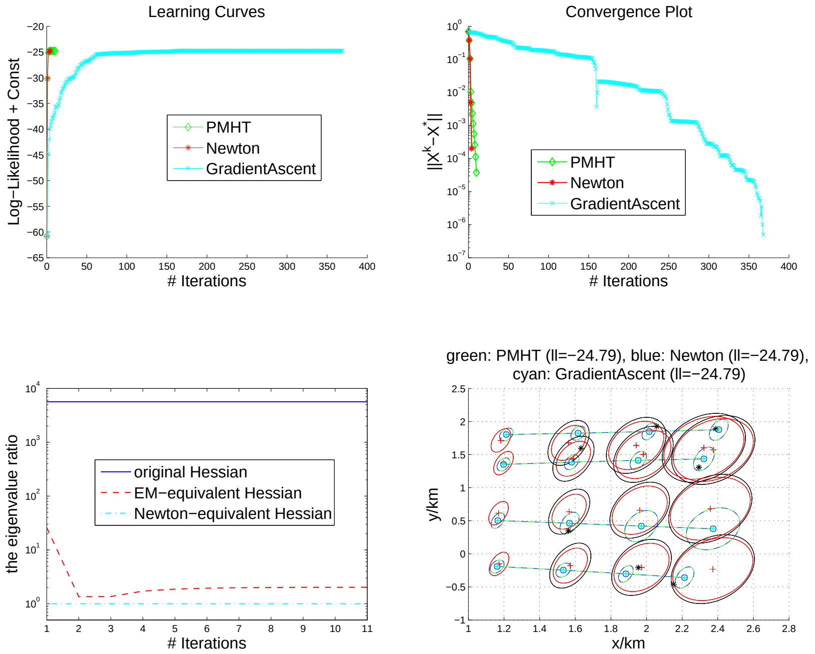

FIGURE 5.3. Single MC run with target distance 0.6 $km$.

convergence of the estimated parameter. For the convergence plot the parameter estimate of the last iteration of each algorithm is used as reference value for $\mathbf{X}^*$.

In the lower row on the left, we examine the eigenvalue ratios of $B(\mathbf{X}^{(k)})\nabla^2 L(\mathbf{X}^{(k)})$ at the iteration points of the EM algorithm. In that product, $B(\mathbf{X}^{(k)})$ corresponds to the preconditioner of an preconditioned gradient ascent method of the form (4.3). For Newton's method, the preconditioner $B$ corresponds to the inverse Hessian of $L$. The product thus corresponds to the identity matrix so that the eigenvalue ratio always equals 1. For gradient ascent there exists no preconditioner, i.e. $B(\mathbf{X}^{(k)}) \equiv I$. We thus analyze the eigenvalues of the Hessian of $L$. Finally, for the EM algorithm it was shown, that under the assumptions of Theorem 3.4 we have $B(\mathbf{X}^{(k)}) = \left[-\eta D^{20} Q(\mathbf{X}^{(k)}, \mathbf{X}^{(k)})\right]^{-1}$. In our example, the assumptions of Theorem 3.4 are met. It can be seen in Proposition 4.2 that the closer this eigenvalue ratio gets to 1 for a given algorithm, the smaller is the error reduction factor which should yield faster local convergence. Our numerical examples agree with that. The large eigenvalue ratios of the gradient ascent algorithm correspond to extremely slow convergence of that algorithm. The values of the eigenvalue ratio of the EM algorithm on the other hand come close to 1 and the algorithm is observed to converge fast.

The plots in the lower right corners of Figure 5.2 and 5.3, respectively, visualize the tracking results. The a priori target state predictions are shown as red-colored crosses together with their also red-colored 90% confidence ellipses and the black-colored innovation covariance ellipses. The black stars are the simulated measurements. The tracking results of the three analyzed algorithms are plotted in green, blue and cyan, respectively. As all algorithms converge to the same maximum, their tracking results coincide. Note that for any of the considered algorithms (as for most optimization algorithms) convergence to a global maximum is not guaranteed. Yet, in the 1000 MC runs that we performed for the generation of Figure 5.1, the maxima found by PMHT and Newton's method differed only once.

The results for the second tracking scenario depict the truth of parallel moving objects quite well whereas the tracking results of the first scenario have problems to match reality. This is however no problem of the employed algorithm but rather due to an erroneous model or insufficient data.

5.4. **Discussion.** Even for unambiguous situations of 'well separated' targets, extremely slow convergence has been observed for the conjugate-gradient method. This disqualifies gradient ascent for practical appliance.

Comparing the EM algorithm with Newton's method by only looking at the number of iterations, Newton's method converges faster. Yet, the computational costs of each iteration are higher compared to those of the EM algorithm due to the calculation of the Hessian. Thus, for 'well separated' targets the application of Newton's method does not pay off since the difference in the total number of iterations is not significant. For closely spaced targets, however, a switch to Newton's method after some EM iterations might be advisable. Yet, at the beginning and whenever the Hessian is not negative definite, one should resort to EM as the global convergence properties of EM are much better. Figure 5.4 compares the actual CPU times spent by both algorithms for the processing of the above described simulation and emphasizes the mentioned strategy. As it is quite often the case, the actual gain of one of these methods over the other depends on the specific problem at hand. Furthermore, implementational issues may also influence the

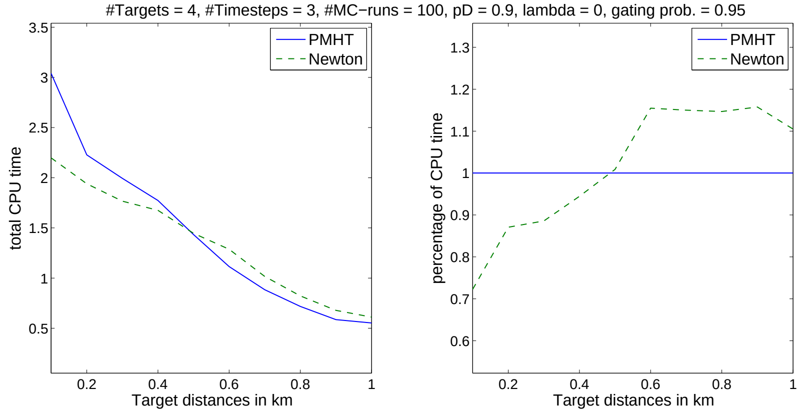

FIGURE 5.4. CPU time as a function of target distances.

choice whether to prefer the pure EM algorithm or a hybrid Newton-EM algorithm. For example, in the PMHT context, a parallelization is easily implemented for the EM algorithm since the M-step decouples into smaller independent maximization problems (cf.(5.7)). This is not the case for Newton's method.

The above derived convergence results might be relevant for automatic tracking and data association. One might argue that for 'well separated' targets without clutter, data association is usually not a problem and can be performed with much simpler modeling and algorithms than PMHT. Albeit this might be true, one first has to automatically recognize that the targets are separated well enough to employ these simpler methods. If targets come closer, one has to detect the change and switch back to a more elaborated data association algorithm. This automatic recognition of the adequate algorithm for the current situation can e.g. be performed using clustering, but requires additional routines and hence involves extra computational costs. With the above derived convergence result, we know, however, that if targets are 'well separated', the EM algorithm is very fast and does not produce much overhead. Nothing is lost by still applying it. Furthermore, we can omit the mentioned routines for analyzing the type of the current tracking situation and avoid additional computational efforts. The situation is captured quickly and there is no need to make provisions for recognizing situations where alternative algorithms are sufficient. Yet, it is important to implement a good adaptive stopping criterion to identify convergence as fast as possible and stop further processing. Otherwise nothing would be gained.

## APPENDIX A. INTERCHANGING INTEGRATION AND DIFFERENTIATION

In several places we need to exchange differentiation w.r.t. the parameter and integration over the data. In this appendix, we collect the analytical justification to do so. We start by reviewing the well-known Leibniz rule adapted to the framework and notation that we need here (for completeness we include a proof). We will use the following abbreviation

$$L_1(\mathbb{D})\times C^1(\Phi) := \{G : \mathbb{D}\times\Phi \to \mathbb{R} :\ G(\cdot,\phi) \in L_1(\mathbb{D}), \phi \in \Phi; G(x,\cdot) \in C^1(\Phi), x \in \mathbb{D}\}.$$

**Theorem A.1** (Leibniz integral rule)**.** *Let $\mathbb{D} \subset \mathbb{R}^n$ be a bounded domain, $\Phi \subset \mathbb{R}^P$ open and $F : \mathbb{D} \times \Phi \to \mathbb{R}$ with $F \in L_1(\mathbb{D}) \times C^1(\Phi)$ such that there exists a function $0 \leq \Psi \in L_1(\mathbb{D}) \cap C(\mathbb{D})$ satisfying $|D^{01}F(x,\phi)| \leq \Psi(x)$ for all $(x,\phi) \in \mathbb{D} \times \Phi$. Then, $D^{01}F(\cdot,\phi) \in L_1(\mathbb{D})$, $\int_{\mathbb{D}} F(x,\phi)\,dx \in C^1(\Phi)$ and*

$$\frac{\partial}{\partial\phi}\int_{\mathbb{D}} F(x,\phi)\ dx = \int_{\mathbb{D}} \frac{\partial}{\partial\phi} F(x,\phi)\ dx. \tag{A.1}$$

*Proof.* The proof follows the lines of [12, p. 195]. Let $\phi^{(0)} \in \Phi$ and choose $r > 0$ such that $\|\phi - \phi^{(0)}\| < r$ for all $\phi \in \Phi$. Next, choose a sequence $(h^{(k)})_{k\in\mathbb{N}} \subset \mathbb{R}$, $h^{(k)} \searrow 0$ and $|h^{(k)}| < r$, $h^{(k)} \neq 0$ for all $k \in \mathbb{N}$. Denoting by $e_\nu$, $1 \leq \nu \leq P$, the $\nu$-th unit vector in $\mathbb{R}^P$, set $\phi^{(k)} := \phi^{(0)} + h^{(k)}e_\nu$ as well as

$$\varphi^{(k)}(x) := \frac{1}{h^{(k)}}\big(F(x,\phi^{(k)}) - F(x,\phi^{(0)})\big), \qquad x \in \mathbb{D}.$$

By assumption, we obtain $\varphi^{(k)} \in L_1(\mathbb{D})$ and $\lim_{k\to\infty}\varphi^{(k)}(x) = \frac{\partial}{\partial\phi_\nu}F(x,\phi^{(0)})$, $x \in \mathbb{D}$. By $|D^{01}F(x,\phi)| \leq \Psi(x)$, we conclude that $|\varphi^{(k)}(x)| \leq \Psi(x)$ for all $x \in \mathbb{D}$ and hence $\frac{\partial}{\partial\phi_\nu}F(\cdot,\phi^{(0)}) \in L_1(\mathbb{D})$ as well as

$$\begin{aligned}\frac{\partial}{\partial\phi_\nu}\int_{\mathbb{D}} F(x,\phi)\,dx &= \lim_{k\to\infty}\frac{1}{h^{(k)}}\int_{\mathbb{D}}\big(F(x,\phi^{(k)}) - F(x,\phi^{(0)})\big)\,dx \\ &= \lim_{k\to\infty}\int_{\mathbb{D}}\varphi^{(k)}(x)\,dx = \int_{\mathbb{D}}\frac{\partial}{\partial\phi_\nu}F(x,\phi^{(0)})\,dx,\end{aligned}$$

which proves the claim. □

Slightly different versions of the Leibniz integral rule can also be found in [7, 8, 10].

**Remark A.2.** *Assume that a parametric probability density function $f_\phi^X$ satisfies Assumption A. Then,*

$$\frac{\partial}{\partial\phi}\int_{\tilde{\mathbb{D}}} f_\phi^X(x)\,dx = \int_{\tilde{\mathbb{D}}}\frac{\partial}{\partial\phi}f_\phi^X(x)\,dx$$

*for any $\tilde{\mathbb{D}} \subset \mathbb{D}$. In fact, it can easily be seen that Assumption A implies the validity of the assumptions in Theorem A.1.*

Let us now collect several circumstances allowing the above interchange of differentiation w.r.t. parameter and integration w.r.t. data.

**Lemma A.3.** *Let Assumption A hold true. If $f_\phi^X(x) \in C^2(\Phi)$ for all $x \in \mathbb{D}$ and $\frac{\partial}{\partial\phi}f_\phi^X \in L_\infty(\mathbb{D})$, then*

$$\frac{\partial}{\partial\phi}\int_{\mathbb{D}}\frac{\partial}{\partial\phi}f_\phi^X(x)\,dx = \int_D \frac{\partial^2}{\partial\phi^2}f_\phi^X(x)\,dx.$$

*Proof.* Again, it is straightforward to verify that the assumptions in Theorem A.1 hold for $F(x,\phi) = \frac{\partial}{\partial\phi}f_\phi^X(x)$. □

**Lemma A.4.** *Let Assumptions A and B hold. Then, $\int_{\ell^{-1}(y)}\frac{\partial}{\partial\phi}h_\phi^{X|Y}(x|y)\,dx = 0$.*

*Proof.* We only have to show that $\int_{\ell^{-1}(y)}\frac{\partial}{\partial\phi}h_\phi^{X|Y}(x|y)\,dx = \frac{\partial}{\partial\phi}\int_{\ell^{-1}(y)}h_\phi^{X|Y}(x|y)\,dx$, since the integral on the right-hand side is unity. First, we have by definition and assumption that $\int_{\ell^{-1}(y)}|h_\phi^{X|Y}(x|y)|\,dx = \frac{1}{g_\phi^Y(y)}\int_{\ell^{-1}(y)}|f_\phi^X(x)|\,dx < \infty$, since

$g_\phi^Y(y) > 0$ by Assumption B, $\ell^{-1}(y)$ has positive measure and $f_\phi^X \in L_1(\mathbb{D})$ by assumption. By (2.1), Theorem A.1 and (A2), we have that $g_\phi^Y(y) \in C^1(\Phi)$ for all $y \in \mathbb{O}$ and the same holds for $h_\phi^{X|Y}(x|y)$, $x \in \ell^{-1}(y)$, as well. Finally $\frac{\partial}{\partial\phi} h_\phi^{X|Y}(x) = \frac{\frac{\partial}{\partial\phi} f_\phi^X(x)}{g_\phi^Y(y)} - \frac{f_\phi^X(x)\frac{\partial}{\partial\phi} g_\phi^Y(y)}{(g_\phi^Y(y))^2} \in L_\infty(\ell^{-1}(y))$ by (A3). □

We note an immediate consequence.

**Corollary A.5.** *In addition to Assumption A and B let $f_\phi^X \in C^2(\Phi)$ for all $x \in \mathbb{D}$ and $\frac{\partial}{\partial\phi} f_\phi^X \in L_\infty(\mathbb{D})$. Then, $\frac{\partial}{\partial\phi}\int_{\ell^{-1}(y)} \frac{\partial}{\partial\phi} h_\phi^{X|Y}(x)\,dx = \int_{\ell^{-1}(y)} \frac{\partial^2}{\partial\phi^2} h_\phi^{X|Y}(x)\,dx = 0$.* □

**Lemma A.6.** *If Assumptions A and B hold, we have*

$$\frac{\partial}{\partial\phi}\int_{\ell^{-1}(y)} \Big(\log f_\phi^X(x)\Big)\, h_\psi^{X|Y}(x|y)\,dx = \int_{\ell^{-1}(y)} \frac{\frac{\partial}{\partial\phi} f_\phi^X(x)}{f_\phi^X(x)} h_\psi^{X|Y}(x|y)\,dx$$

*for all fixed $\psi \in \Phi$.*

*Proof.* We have to verify the conditions of Theorem A.1 for the specific choice $F(x,\phi) = \big(\log f_\phi^X(x)\big)\, h_\psi^{X|Y}(x|y)$, $\psi \in \Phi$ fixed. First,

$$\begin{aligned}\int_{\ell^{-1}(y)} |F(x,\phi)|\,dx &= \int_{\ell^{-1}(y)} \Big|\big(\log\, f_\phi^X(x)\big)\frac{f_\psi^X(x)}{g_\psi^Y(y)}\Big|\,dx\\ &= \frac{1}{g_\psi^Y(y)}\int_{\ell^{-1}(y)} |f_\psi^X(x)\,\log\, f_\phi^X(x)|\,dx\end{aligned}$$

since $g_\psi^Y(y) > 0$ due to Assumption B. Since $f_\psi^X \in L_\infty(\mathbb{D})$ and $\log\, f_\phi^X \in L_1(\ell^{-1}(y))$, we obtain that $F(\cdot,\phi) \in L_1(\mathbb{D})$. It is obvious that $F(x,\cdot) \in C^1(\Phi)$ for $x \in \mathbb{D}$ by (A2). Finally,

$$\frac{\partial}{\partial\phi} F(x,\phi) = \frac{\frac{\partial}{\partial\phi} f_\phi^X(x)}{f_\phi^X(x)}\,\frac{f_\psi^X(x)}{g_\psi^Y(y)},$$

the second term being in $L_\infty(\mathbb{D})$ and the first one in $L_1(\mathbb{D})$ so that $\frac{\partial}{\partial\phi} F(\cdot,\phi) \in L_1(\mathbb{D})$ for all $\phi \in \Phi$. □

**Corollary A.7.** *If Assumption C holds, we have*

$$\frac{\partial}{\partial\phi}\int_{\ell^{-1}(y)} \Big(\frac{\partial}{\partial\phi}\log f_\phi^X(x)\Big)\, h_\psi^{X|Y}(x|y)\,dx = \int_{\ell^{-1}(y)} \frac{\partial}{\partial\phi}\frac{\frac{\partial}{\partial\phi} f_\phi^X(x)}{f_\phi^X(x)} h_\psi^{X|Y}(x|y)\,dx$$

*for all fixed $\psi \in \Phi$.*

*Proof.* It suffices to remark that $f_\phi^X$ is bounded away from zero by Remark 2.1 and that Assumption C ensures sufficient regularity. □

**Corollary A.8.** *Let Assumptions A and B hold, then*

$$\frac{\partial}{\partial\phi}\int_{\ell^{-1}(y)} (\log\, h_\phi^{X|Y}(x|y))\, h_\psi^{X|Y}(x|y)\,dx = \int_{\ell^{-1}(y)} \frac{\frac{\partial}{\partial\phi} h_\phi^{X|Y}(x|y)}{h_\phi^{X|Y}(x|y)} h_\psi^{X|Y}(x|y)\,dx.$$

*Proof.* We set

$$F(x,\phi) := h_{\psi}^{X|Y}(x|y)\,\log\, h_{\phi}^{X|Y}(x|y) = h_{\psi}^{X|Y}(x|y)\,(\log\, f_{\phi}^{X}(x) - \log\, g_{\phi}^{Y}(y)).$$

The first term can be treated by Lemma A.6 and the second one is straightforward since $\log\, g_{\phi}^{Y}(y)$ is independent of the integration variable and $h_{\psi}^{X|Y}(x|y)$ is independent of $\phi$. □

**Lemma A.9.** *In addition to the Assumptions A and B let $q \in L_\infty(\mathbb{D})$, then*

$$\frac{\partial}{\partial\phi}\int_{\ell^{-1}(y)} h_{\phi}^{X|Y}(x|y)\,q(x)\,dx = \int_{\ell^{-1}(y)} \frac{\partial}{\partial\phi} h_{\phi}^{X|Y}(x|y)\,q(x)\,dx.$$

*Proof.* Consider the function $F(x,\phi) := h_{\phi}^{X|Y}(x|y)\,q(x)$. By assumption, we have that $h_{\phi}^{X|Y} \in L_1(\mathbb{D})$ so that $F \in L_1(\mathbb{D}) \times C^1(\Phi)$. Next,

$$\begin{aligned}\frac{\partial}{\partial\phi} h_{\phi}^{X|Y}(x) &= \frac{\big(\frac{\partial}{\partial\phi} f_{\phi}^{X}(x)\big) g_{\phi}^{Y}(y) - f_{\phi}^{X}(x)\,\frac{\partial}{\partial\phi} g_{\phi}^{Y}(y)}{\big(g_{\phi}^{Y}(y)\big)^2} \\ &= \frac{\frac{\partial}{\partial\phi} f_{\phi}^{X}(x)}{g_{\phi}^{Y}(y)} - f_{\phi}^{X}(x)\frac{\frac{\partial}{\partial\phi} g_{\phi}^{Y}(y)}{\big(g_{\phi}^{Y}(y)\big)^2}\end{aligned}$$

which is an $L_1(\mathbb{D})$-function by our assumptions. Thus, all requirements of Theorem A.1 hold. □

**Lemma A.10.** *Let Assumption C hold, then*

$$\frac{\partial}{\partial\phi}\int_{\ell^{-1}(y)} \frac{\frac{\partial}{\partial\phi} h_{\phi}^{X|Y}(x|y)}{h_{\phi}^{X|Y}(x|y)} h_{\psi}^{X|Y}(x|y)\,dx = \int_{\ell^{-1}(y)} \frac{\partial}{\partial\phi}\frac{\frac{\partial}{\partial\phi} h_{\phi}^{X|Y}(x|y)}{h_{\phi}^{X|Y}(x|y)} h_{\psi}^{X|Y}(x|y)\,dx.$$

*Proof.* We have to show that $F(x,\phi) := \frac{\frac{\partial}{\partial\phi} h_{\phi}^{X|Y}(x|y)}{h_{\phi}^{X|Y}(x|y)} h_{\psi}^{X|Y}(x|y)$ satisfies all requirements of Theorem A.1. To this end, note that $h_{\phi}^{X|Y}(x|y) \geq f_{\phi}^{-} \cdot (g_{\phi}^{Y}(y))^{-1} > 0$ by Remark 2.1, which implies that $F(\cdot,\phi) \in L_1(\mathbb{D})$ for all $\phi \in \Phi$. Moreover, it is then immediate that $F \in L_1(\mathbb{D}) \times C^1(\Phi)$. Next, we have

$$D^{01}F(x,\phi) = h_{\psi}^{X|Y}(x|y)\frac{h_{\phi}^{X|Y}(x|y)\frac{\partial^2}{\partial\phi^2} h_{\phi}^{X|Y}(x|y) - \big(\frac{\partial}{\partial\phi} h_{\phi}^{X|Y}(x|y)\big)^2}{\big(h_{\phi}^{X|Y}(x|y)\big)^2},$$

which is also in $L_1(\mathbb{D}) \cap L_\infty(\mathbb{D})$ by the above arguments. □

## References

[1] B. D. O. Anderson and J. B. Moore. *Optimal Filtering.* Information and system science series. Prentice Hall, Inc., Englewood Cliffs, NJ, 1979.

[2] Y. Bar-Shalom and X.-R. Li. *Estimation and Tracking: Principles, Techniques, and Software.* Artech House, Inc., 1993.

[3] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan. *Estimation with Applications to Tracking and Navigation: Theory, Algorithms and Software.* Wiley, New York ; Chichester ; Weinheim [et al.], 2001.

[4] S. Blackman and R. Popoli. *Design and Analysis of Modern Tracking Systems.* Artech House, Inc., Norwood, MA, 1999.

[5] A. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:1–38, 1977.

[6] B. Efron and D. V. Hinkley. Assessing the Accuracy of the Maximum Likelihood Estimator: Observed Versus Expected Fisher Information. *Biometrika*, 65(3):457–483, Dec 1978.

[7] O. Forster. *Differentialrechnung im $\mathbb{R}^n$, gewöhnliche Differentialgleichungen.* Vieweg + Teubner, Wiesbaden, 8th edition, 2008.

[8] Heuser, Harro. *Lehrbuch der Analysis 2.* Teubner, Stuttgart; Leipzig; Wiesbaden, 13th edition, 2004.

[9] M. Jamshidian and R. I. Jennrich. Conjugate Gradient Acceleration of the EM Algorithm. *Journal of the American Statistical Association*, 88(421):221–228, Mar 1993.

[10] W. Kaballo. *Einführung in die Analysis 2.* Spektrum, Akad. Verl., Heidelberg; Berlin; Oxford, 2nd edition, 1997.

[11] R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Transaction of the ASME-Journal of Basic Engineering*, 82:35–45, 1960.

[12] K. Königsberger. *Analysis 2.* Springer, Berlin, 2nd edition, 1993.

[13] S. Lang. *Real analysis.* Addison-Wesley Publishing Company Advanced Book Program, Reading, MA, second edition, 1983.

[14] K. Lange. *Numerical Analysis for Statisticians.* Springer-Verlag, New York, 1999.

[15] K. Lange. *Optimization.* Springer texts in statistics. Springer, New York; Berlin; Heidelberg [et al.], 2004.

[16] M. Mallick and B. La Scala. Comparison of Single-point and Two-point Difference Track Initiation Algorithms Using Position Measurements. *Acta Automatica Sinica*, 34(3):258–265, 2008.

[17] I. Meilijson. A Fast Improvement to the EM Algorithm on its Own Terms. *Journal of the Royal Statistical Society B*, 51(1):127–138, 1989.

[18] A. M. Ostrowski. *Solution of Equations and Systems of Equations.* Academic Press, New York, 2nd edition, 1966.

[19] H. E. Rauch, F. Tung, and C. T. Striebel. Maximum Likelihood Estimates of Linear Dynamic Systems. *AIAA Journal*, 3(8):1445–1450, 1965.

[20] R. Salakhutdinov and Z. Ghahramani. Relationship between gradient and EM steps in latent variable models. Technical Report, 2002. `http://www.mit.edu/~rsalakhu/papers/report.pdf`.

[21] R. Salakhutdinov, S. Roweis, and Z. Ghahramani. Expectation-Conjugate Gradient: An Alternative to EM. Technical Report, 2002. `http://www.mit.edu/~rsalakhu/papers/ecgdraft.pdf`.

[22] R. Salakhutdinov, S. Roweis, and Z. Ghahramani. Optimization with EM and Expectation-Conjugate-Gradient. In *Proceedings of the International Conference on Machine Learning*, volume 20, pages 672–679, 2003.

[23] J. R. Shewchuk. An Introduction to the Conjugate Gradient Method Without the Agonizing Pain. Technical report, Carnegie Mellon University, Pittsburgh, PA, USA, 1994.

[24] T. Springer. *Mathematical Analysis and Computational Methods for Probabilistic Multi-Hypothesis Tracking (PMHT).* PhD thesis, Ulm University, Feb. 2013.

[25] R. L. Streit and T. E. Luginbuhl. Maximum likelihood method for probabilistic multi-hypothesis tracking. In *Proceedings of SPIE International Symposium, Signal and Data Processing of Small Targets*, SPIE Proceedings Vol. 2335–24, pages 394–405, Orlando, FL, Apr 1994.

[26] R. L. Streit and T. E. Luginbuhl. Probabilistic Multi-Hypothesis Tracking. Technical Report NUWC-NPT 10,428, Naval Undersea Warfare Center, Division Newport, RI, Feb 1995.

[27] P. K. Willett, Y. Ruan, and R. L. Streit. PMHT: Problems and Some Solutions. *IEEE Transactions on Aerospace and Electronic Systems*, 38(3):738–754, 2002.

[28] L. Xu and M. I. Jordan. On Convergence Properties of the EM Algorithm for Gaussian Mixtures. *Neural Computation*, 8(1):129–151, 1996.

Theresa Springer, University of Ulm, Institute of Numerical Mathematics, Helmholtzstrasse 20, D-89069 Ulm, Germany

*E-mail address*: `theresa.springer@uni-ulm.de`

Karsten Urban, University of Ulm, Institute of Numerical Mathematics, Helmholtzstrasse 20, D-89069 Ulm, Germany

*E-mail address*: `karsten.urban@uni-ulm.de`