PREDICTION OF THE EM-ALGORITHM SPEED OF CONVERGENCE WITH CRAMER-RAO BOUNDS

Cédric Herzet, Luc Vandendorpe

Communications Laboratory, Université catholique de Louvain,
Pl. du Levant 2, B1348 Louvain-la-Neuve, Belgium

ABSTRACT

This paper aims at characterizing the (mean) speed of convergence of the EM algorithm. We derive, under some simplifying assumptions, a relation between the EM algorithm mean convergence rate (MCR) and Cramer-Rao bounds (CRBs) associated to the so-called *incomplete* and *complete* data sets defined within the EM algorithm framework. We illustrate our derivations in the case of carrier-phase estimation based on the EM algorithm. As far as our simulation setups are concerned, we show that the (mean) EM-algorithm behavior may be well predicted by means of the proposed CRB-based expression.

Index Terms— Convergence of numerical methods, Maximum likelihood estimation

1. INTRODUCTION

The Expectation-Maximization (EM) algorithm is an iterative methodology for solving maximum-likelihood (ML) problems. Since its first statement by Dempster, Laird and Rubin [1], much literature has been devoted to the study of its behavior and convergence properties, see e.g. [2] and references therein.

The speed of convergence is often considered as the main drawback of the EM algorithm because very low in some cases. Dempster, Laird and Rubin [1] showed that the convergence of the EM algorithm is generally linear, with a rate of convergence obtained from the information matrices associated to the *missing* and *complete* data sets. More recently, some authors [3, 4] have given further insights into the EM-algorithm convergence. In particular, in [3, 4] the authors emphasize that the EM algorithm may locally achieve quasi-Newton behavior in some specific situations. Although these contributions gives an exact mathematical formulation of the local convergence of the EM algorithm, they do not necessarily enable to *easily* predict its behavior as a function of the parameters of the problem at hand.

In this contribution, we address this problem. More specifically, we derive an expression relating, under some simplifying assumptions, the EM-algorithm mean convergence rate (MCR) to the Cramer-Rao bounds [5] associated to the *incomplete* and *complete* data estimation problems [1]. Taking benefit from the numerous contributions dedicated to the study of the CRBs, we will emphasize that the proposed MCR expression enables to easily predict the EM algorithm speed of convergence as a function of the system parameters.

2. ML ESTIMATION AND EM ALGORITHM

Let $\mathbf{r}$ denote a vector of observations depending on an unknown *scalar* parameter b. The maximum-likelihood (ML) estimate of b is defined as the solution of the following maximization problem

$$\hat{b}_{ML} = \arg\max_{\tilde{b}} \ p(\mathbf{r}|\tilde{b}), \tag{1}$$

where $\tilde{b}$ is a trial value of b. The EM algorithm, first defined by Dempster, Laird and Rubin in [1], is a powerful iterative methodology to deal with ML problem. Formally, the EM algorithm is based on the following two steps:

$$\text{E-step: } \mathcal{Q}(\tilde{b}, \hat{b}^{(n)}) = \int_{\mathcal{Z}} p(\mathbf{z}|\mathbf{r}, \hat{b}^{(n)}) \, \log p(\mathbf{z}|\tilde{b}) \, d\mathbf{z}, \tag{2}$$

$$\text{M-step: } \hat{b}^{(n+1)} = \arg\max_{\tilde{b}} \ \mathcal{Q}(\tilde{b}, \hat{b}^{(n)}), \tag{3}$$

where $\hat{b}^{(n)}$ is the estimate computed by the EM algorithm at iteration n and $\mathbf{r} = f(\mathbf{z})$, where $f(\mathbf{z})$ is a *many-to-one* mapping. Vectors $\mathbf{r}$ and $\mathbf{z}$ are often referred to as the *incomplete* and the *complete* data set, respectively.

Talking about iterative processing, the question naturally arises of the speed of convergence of the EM algorithm. Dempster, Laird and Rubin showed in their seminal paper [1] that the convergence of the EM algorithm is usually linear i.e. we have in a neighborhood of $\hat{b}_{ML}$ that

$$e^{(n+1)} = C(\mathbf{r}) \, e^{(n)}, \tag{4}$$

where $e^{(n)} = \| \hat{b}^{(n)} - \hat{b}_{ML} \|$ and $C(\mathbf{r})$ is the *rate of convergence* of the EM algorithm. The authors showed moreover that the rate of convergence is related to the amount of *miss-*

ing information [1] in the considered problem i.e.

$$C(\mathbf{r}) = I_c^{-1}(\mathbf{r})\, I_m(\mathbf{r}) \tag{5}$$

where $I_c(\mathbf{r})$ and $I_m(\mathbf{r})$ are respectively the information matrices associated to the complete and the missing data i.e.,

$$I_c(\mathbf{r}) \triangleq -\left(\int_{\mathcal{Z}} p(\mathbf{z}|\mathbf{r},\tilde{b})\ \frac{\partial^2}{\partial \tilde{b}^2}\ \log p(\mathbf{z}|\tilde{b})\, d\mathbf{z}\right)_{\big|\tilde{b}=\hat{b}_{ML}} \tag{6}$$

$$I_m(\mathbf{r}) \triangleq -\left(\int_{\mathcal{Z}} p(\mathbf{z}|\mathbf{r},\tilde{b})\ \frac{\partial^2}{\partial \tilde{b}^2}\ \log p(\mathbf{z}|\mathbf{r},\tilde{b})\, d\mathbf{z}\right)_{\big|\tilde{b}=\hat{b}_{ML}}. \tag{7}$$

3. A CRB-BASED EXPRESSION OF THE EM MEAN CONVERGENCE RATE

In this section, we derive an approximated expression of the mean convergence rate (MCR) of the EM algorithm. We define the MCR, say M_C, as the coefficient relating the mean distances $E_{\mathbf{r}|b}[e^{(n+1)}]$ at two successive iterations i.e.,

$$E_{\mathbf{r}|b}[e^{(n+1)}] = M_C\ E_{\mathbf{r}|b}[e^{(n)}], \tag{8}$$

where $E_{\mathbf{r}|b}[\cdot]$ denotes the expectation with respect to $p(\mathbf{r}|b)$. In the sequel, considering simplifying assumptions, the MCR will be shown to be related to the CRBs associated to the *incomplete* and *complete* data set.

Our derivations are based on the following two assumptions:

1. We assume that $\forall\, \epsilon > 0$, we have

$$\Pr\{\|I_m(\mathbf{r}) - K_1\| < \epsilon\} \simeq 1, \tag{9}$$
$$\Pr\{\|I_c(\mathbf{r}) - K_2\| < \epsilon\} \simeq 1, \tag{10}$$

   where K_1 and K_2 are two constants. In words, assumption (9) (resp. (10)) means that the probability of observing a vector $\mathbf{r}$ such that $I_m(\mathbf{r})$ (resp. $I_c(\mathbf{r})$) is ϵ-close to some value K_1 (resp. K_2) is almost equal to 1.

2. We assume that the ML estimate $\hat{b}_{ML}$ is close to the actual parameter value b, i.e., $\hat{b}_{ML} \simeq b$.

At first sight, assumption 1 and 2 may appear quite restrictive. However, they are reasonable in many practical scenarios as a direct consequence of the law of large numbers (assumption 1) and the ML asymptotic efficiency (assumption 2) [5], respectively. For example, in digital communication problems the size of observation vector $\mathbf{r}$ is typically large (roughly between 100 and 10000) and the observations are only locally correlated. In such a situation, assumption 1 is often reasonable as a consequence of the law of large numbers. Moreover, the system reliability requires a good precision on the estimated parameters, and therefore assumption 2 is usually also satisfied.

Based on assumptions 1 and 2, we now derive an expression relating the EM-algorithm MRC to the CRBs associated to the complete and the incomplete data sets. Starting from (4) and taking the expectation of both sides with respect to $p(\mathbf{r}|b)$, we have

$$\int_{\mathcal{R}} p(\mathbf{r}|b)\ e^{(n+1)}\ d\mathbf{r} = \int_{\mathcal{R}} p(\mathbf{r}|b)\ C(\mathbf{r})\ e^{(n)}\ d\mathbf{r} \tag{11}$$

$$\simeq \frac{\int_{\mathcal{R}} p(\mathbf{r}|b)\ I_m(\mathbf{r})\ d\mathbf{r}}{\int_{\mathcal{R}} p(\mathbf{r}|b)\ I_c(\mathbf{r})\ d\mathbf{r}} \int_{\mathcal{R}} p(\mathbf{r}|b)\ e^{(n)}\ d\mathbf{r}, \tag{12}$$

since by assumption 1, matrices $I_m(\mathbf{r})$ and $I_c(\mathbf{r})$ are equal to some constants with probability (almost) one. From (12), it immediately follows that

$$M_C \simeq \frac{\int_{\mathcal{R}} p(\mathbf{r}|b)\ I_m(\mathbf{r})\ d\mathbf{r}}{\int_{\mathcal{R}} p(\mathbf{r}|b)\ I_c(\mathbf{r})\ d\mathbf{r}}. \tag{13}$$

Based on assumption 2, we will now show that (13) may also be expressed as

$$M_C \simeq 1 - \frac{\mathrm{CRB}_{\mathbf{z}}(b)}{\mathrm{CRB}_{\mathbf{r}}(b)}, \tag{14}$$

where $\mathrm{CRB}_{\mathbf{r}}(b)$ and $\mathrm{CRB}_{\mathbf{z}}(b)$ are the CRBs associated to the incomplete data set $\mathbf{r}$ and the complete data set $\mathbf{z}$, respectively. In order to show (14), we will show that

$$E_{\mathbf{r}|b}\big[I_c(\mathbf{r})\big] \simeq \mathrm{CRB}_{\mathbf{z}}^{-1}, \tag{15}$$

$$E_{\mathbf{r}|b}\big[I_m(\mathbf{r})\big] \simeq \mathrm{CRB}_{\mathbf{z}}^{-1}(b) - \mathrm{CRB}_{\mathbf{r}}^{-1}(b). \tag{16}$$

Let us first show (15). Using the definition of the complete-data information matrix (6) and taking the expectation with respect to $p(\mathbf{r}|b)$, we have

$$E_{\mathbf{r}|b}\big[I_c(\mathbf{r})\big] = -\int_{\mathcal{R}} p(\mathbf{r}|b) \int_{\mathcal{Z}} p(\mathbf{z}|\mathbf{r},\tilde{b}) \times \frac{\partial^2}{\partial \tilde{b}^2}\ \log p(\mathbf{z}|\tilde{b})\, d\mathbf{z}\, d\mathbf{r}_{|\tilde{b}=\hat{b}_{ML}} \tag{17}$$

Using assumption 2, i.e., $\hat{b}_{ML} \simeq b$, and the Bayes rule we have

$$E_{\mathbf{r}|b}\big[I_c(\mathbf{r})\big] \simeq -\int_{\mathcal{R}}\int_{\mathcal{Z}} p(\mathbf{z},\mathbf{r}|\tilde{b})\ \frac{\partial^2}{\partial \tilde{b}^2}\ \log p(\mathbf{z}|\tilde{b})\, d\mathbf{z}\, d\mathbf{r}_{|\tilde{b}=b}$$
$$= -\int_{\mathcal{Z}} p(\mathbf{z}|\tilde{b})\ \frac{\partial^2}{\partial \tilde{b}^2}\ \log p(\mathbf{z}|\tilde{b})\, d\mathbf{z}_{|\tilde{b}=b}. \tag{18}$$

[1] The missing information may actually be seen as the difference between the amount of information contained in the (so-called) *complete* data set and the *incomplete* data set.

Now, taking the definition of the CRB into account, i.e.

$$\mathrm{CRB}_{\mathrm{z}}(b) = -\left(E_{\mathbf{z}|b} \left[\frac{\partial^2}{\partial\, \tilde{b}^2} \log p(\mathbf{z}|\tilde{b}) \right]_{\big| \tilde{b}=b} \right)^{-1}, \tag{19}$$

we end up with (15).

Let us now consider (16). First notice that $I_m(\mathbf{r})$ and $I_c(\mathbf{r})$ may be related [2] as

$$-\left(\frac{\partial}{\partial \tilde{b}^2} \log p(\mathbf{r}|\tilde{b}) \right)_{\big| \tilde{b}=\hat{b}_{ML}} = I_c(\mathbf{r}) - I_m(\mathbf{r}). \tag{20}$$

Based on (20), we may write

$$E_{\mathbf{r}|b}\left[I_m(\mathbf{r}) \right] = E_{\mathbf{r}|b} \left[I_c(\mathbf{r}) + \left(\frac{\partial}{\partial \tilde{b}^2} \log p(\mathbf{r}|\tilde{b}) \right)_{\big| \tilde{b}=\hat{b}_{ML}} \right]. \tag{21}$$

Using assumption 2, we finally have

$$E_{\mathbf{r}|b^\star}\left[I_m(\mathbf{r}) \right] = \mathrm{CRB}_{\mathbf{z}}^{-1}(b) + E_{\mathbf{r}|b} \left[\left(\frac{\partial}{\partial \tilde{b}^2} \log p(\mathbf{r}|\tilde{b}) \right)_{\big| \tilde{b}=b} \right], \tag{22}$$

$$= \mathrm{CRB}_{\mathbf{z}}^{-1}(b) - \mathrm{CRB}_{\mathbf{r}}^{-1}(b), \tag{23}$$

where (22) follows from (15), and (23) follows from the definition of the CRB (19). This shows (16).

As far as our building assumptions are valid, (12) and (14) establish a relationship between the rate of improvement of $E_{\mathbf{r}|b}[\|b^{(n)} - \hat{b}_{ML}\|]$ and the CRBs associated to the complete and incomplete data sets. In particular, we see from (14) that the (mean) rate at which the EM algorithm converges to the ML estimate decreases a function of the ratio $\mathrm{CRB}_{\mathbf{z}}/\mathrm{CRB}_{\mathbf{r}}$. This ratio is actually a measure of the improvement of the estimation quality which can be achieved by observing the complete-data set instead of the incomplete-data set. Since the behavior of the CRBs associated to many estimation problems have already been extensively studied in the literature, (14) provides an easy way to predict what will be the EM-algorithm behavior by simply looking at the CRB one. This approach will be illustrated in section 4.

4. A PRACTICAL EXAMPLE: EM-BASED ITERATIVE CARRIER PHASE SYNCHRONIZATION

In this section, we illustrate our derivations in the practical case of iterative carrier-phase synchronization of a digital receiver. The model of the received observations is as follows:

$$\mathbf{r} = \mathbf{a}\, e^{j\theta} + \mathbf{v}, \tag{24}$$

where $\mathbf{a}$ is a vector of data symbols, θ is the carrier-phase offset and $\mathbf{v}$ is a vector of zero-mean white Gaussian noise with complex variance σ_v^2. The EM algorithm is applied to the problem of computing the ML estimate of the carrier phase offset. Following the approach proposed in [6], the complete data set is defined as $\mathbf{z} \triangleq [\mathbf{r}^T, \mathbf{a}^T]^T$.

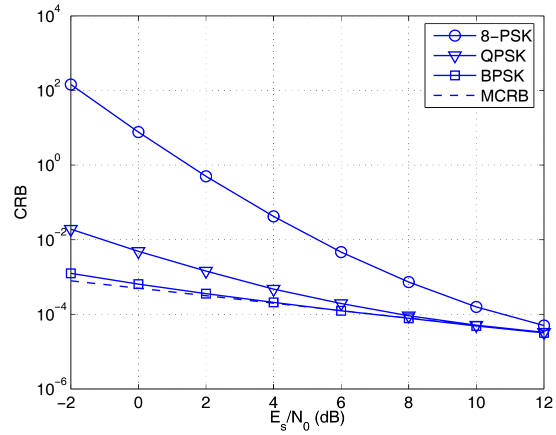

Fig. 1. Cramer-Rao bounds versus E_s/N_0-ratio for uncoded transmission with different constellation sizes.

Note that the behavior of the CRBs associated to carrier-phase estimation have been extensively studied in the literature, see e.g. [7] and references therein. Taking benefit from this knowledge, we will show that the EM algorithm convergence may be well-predicted via (14). As illustrative examples, we will consider the EM-algorithm sensitivity to the symbol-constellation size and the SNR. In each scenario, the EM-algorithm performance computed via Monte-Carlo simulations will be compared to the one predicted by means of (14).

Let us first investigate the EM-algorithm behavior when the size of the symbol constellation alphabet varies. We consider the following setup. The transmitted frames consist of 1000 uncoded PSK symbols. The size of the constellation alphabet is set to either 2 (BPSK), 4 (QPSK) or 8 (8-PSK). We use a Gray mapping.

The CRBs associated to this setup are represented versus the E_s/N_0-ratio in Fig. 1. From (14), we have that the EM convergence should be all the slower as the gap between the incomplete-data and the complete-data CRBs increases. For example, in the considered setup, the incomplete-data BSPK CRB is always closer to the complete-data CRB than the incomplete-data 8-PSK CRB. Hence, according to (14), the EM algorithm should converge slower when applied to a 8-PSK transmission than when applied an a BPSK transmission.

Fig. 2 illustrates the relevance of the proposed approach: we compare the EM-algorithm performance as predicted by (14) with actual performance computed via Monte-Carlo simulations. More particularly, we have represented the mean distance between the EM-algorithm and the ML estimate, i.e.

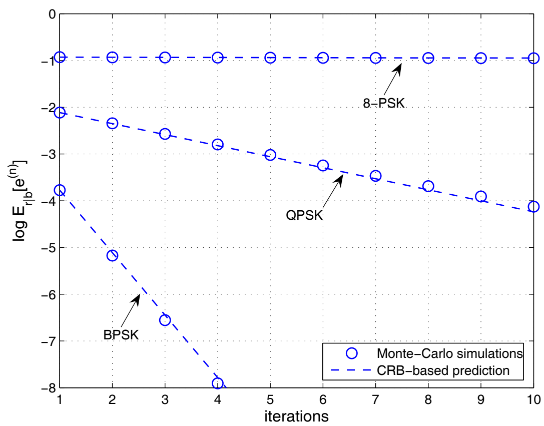

Fig. 2. Mean distance between the ML estimate and the EM-algorithm estimate at a given iteration for different constellation sizes. Monte-Carlo simulations are compared with the performance predicted by (14).

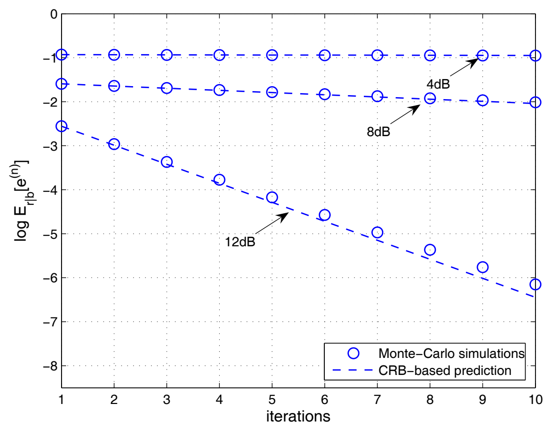

Fig. 3. Mean distance between the ML estimate and the EM-algorithm estimate for different operating SNRs. Monte-Carlo simulations are compared with the performance predicted by (14).

$E_{\mathbf{r}|b}[||\hat{b}^{(n)} - \hat{b}_{ML}||]$, versus the number of EM iterations. The E_s/N_0-ratio has been set to 4dB. The dashed curves corresponds to the prediction computed from (14) and the circles to the actual performance computed via Monte-Carlo simulations. As far as our simulation setup is concerned, we see the proposed CRB-based expression (14) enables to accurately predict the EM algorithm behavior.

We now illustrate the sensitivity of the EM algorithm to the system operating SNR. We keep the same setup as in the previous point. The CRBs plotted in Fig. 1 are therefore still valid for computing the MCR via (14). We see from Fig. 1, that an increase of the SNR reduces the gap between the incomplete-data CRB and the complete-data CRB. From our previous considerations, it seems therefore that an increase of the system operating SNR is beneficial for the EM-algorithm speed of convergence.

This is illustrated in Fig. 3 where we compare the performance predicted via (14) with the one computed by Monte-Carlo simulations for different operating SNRs. The CRB-based predictions are plotted with dashed curves and the simulated points with circles. The constellation alphabet is a Gray-mapped 8-PSK and we have considered E_s/N_0 equal to 4, 8 and 12dB, respectively. We see from this figure that the behavior predicted by (14) is in good accordance with the results computed by Monte-Carlo simulations.

5. CONCLUSIONS

In this contribution, we focus on the mean convergence rate (MCR) of the EM algorithm. In particular, based on some building assumptions, we derived an expression relating the EM-algorithm MCR to the Cramer-Rao bounds (CRBs) associated to the incomplete and the complete data set, respectively. This expression enables an easy intuition of the EM-algorithm behavior: the further is the incomplete-data CRB from the complete-data CRB, the slower the EM-algorithm speed of convergence. We illustrate our derivation by simulation results in the case of EM-based iterative carrier-phase synchronization. In particular, we showed that the performance predicted by our approach is in good accordance with Monte-Carlo simulation results.

6. REFERENCES

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum-likelihood from incomplete data via the em algorithm. *J. Roy. Stat. Soc.*, 39(1):pp. 1–38, January 1977.

[2] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics, USA, 1997.

[3] L. Xu and M.I. Jordan. ”On convergence properties of the EM algorithm for Gaussian mixtures”. *Neural Computation*, 8(1):pp. 129–151, 1996.

[4] R. Salakhutdinov, S. Roweis, and Z. Ghahramani. On the convergence of bound optimization algorithms. url: citeseer.ist.psu.edu/584732.html.

[5] S.M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall, New Jersey, USA, 1993.

[6] N. Noels, C. Herzet, A. Dejonghe, V. Lottici, H. Steendam, M. Moeneclaey, M Luise and L. Vandendorpe. Turbo-synchronization: an EM algorithm approach. In *Proc. IEEE ICC*, Anchorage, May 2003.

[7] N. Noels, H. Steendam and M. Moeneclaey. ”The Cramer-Rao Bound for Phase Estimation from Coded Linearly Modulated Signals”. *IEEE Commun. Lett.*, 7(5):pp. 207–209, May 2003.