Deterministic Approximate EM Algorithm; Application to the Riemann Approximation EM and the Tempered EM

Thomas Lartigue, Stanley Durrleman, Stéphanie Allassonnière

Deterministic Approximate EM Algorithm; Application to the Riemann Approximation EM and the Tempered EM

Thomas Lartigue[1, 2], Stanley Durrleman[1], and Stéphanie Allassonnière[3]

[1] *Aramis Project-Team, Inria, 75012 Paris, France*
[2] *CMAP, CNRS, École Polytechnique, Institut Polytechnique de Paris, 91120 Palaiseau, France*
[3] *HeKA, Centre de recherche des Cordeliers, Université de Paris, INRIA, INSERM, Sorbonne Université, 75012 Paris, France*

Abstract

The Expectation Maximisation (EM) algorithm is widely used to optimise non-convex likelihood functions with latent variables. Many authors modified its simple design to fit more specific situations. For instance, the Expectation (E) step has been replaced by Monte Carlo (MC), Markov Chain Monte Carlo or tempered approximations, etc. Most of the well-studied approximations belong to the stochastic class. By comparison, the literature is lacking when it comes to deterministic approximations. In this paper, we introduce a theoretical framework, with state-of-the-art convergence guarantees, for any deterministic approximation of the E step. We analyse theoretically and empirically several approximations that fit into this framework. First, for intractable E-steps, we introduce a deterministic version of MC-EM using Riemann sums. A straightforward method, not requiring any hyper-parameter fine-tuning, useful when the low dimensionality does not warrant a MC-EM. Then, we consider the tempered approximation, borrowed from the Simulated Annealing literature and used to escape local extrema. We prove that the tempered EM verifies the convergence guarantees for a wider range of temperature profiles than previously considered. We showcase empirically how new non-trivial profiles can more successfully escape adversarial initialisations. Finally, we combine the Riemann and tempered approximations into a method that accomplishes both their purposes.

Keywords— Expectation Maximisation; exponential family; approximate EM; Riemann sums; tempering; annealing

1 Introduction

The Expectation Maximisation (EM) algorithm was introduced by Dempster, Laird and Rubin (DLR, Dempster *et al.* (1977)) to maximise likelihood functions $g(\theta)$ defined from inherent latent variables z and that were non-convex and had intricate gradients and Hessians. The EM algorithm estimates θ in an iterative fashion, starting from a certain initial value θ_0 and where the new estimate θ_{n+1} at step $n+1$ is function the estimate θ_n from the previous step n. In addition to presenting the method, DLR provide convergence guarantees on the sequence of estimated parameters $\{\theta_n\}_n$, namely that it converges towards a critical point of the likelihood function as the step n of the procedure grows. Their flawed proof for this result was later corrected by Wu (1983), and more convergence guarantees were studied by Boyles (1983). Since some likelihood functions are too complex to apply DLR's raw version of the EM, authors of later works have proposed alternative versions, usually with new convergence guarantees. On the one hand, when the maximisation step (M step) is problematic, other optimisation methods such as coordinate descent Wu (1983) or gradient

descent Lange (1995) have been proposed. On the other hand, several works introduce new versions of the algorithm where the expectation step (E step), which can also be intractable, is approximated. Most of them rely on Monte Carlo (MC) methods and Stochastic Approximations (SA) to estimate this expectation. Notable examples include the Stochastic Approximation EM (SAEM, Delyon *et al.* (1999)), the Monte Carlo EM (MC-EM, Wei and Tanner (1990)), the Markov Chain Monte Carlo EM (MCMC-EM, Fort and Moulines (2003)), the MCMC-SAEM Kuhn and Lavielle (2005); Allassonnière *et al.* (2010) and the Approximate SAEM Allassonnière and Chevallier (2021). Random approximation of the E step have also been used in the case where the data are too voluminous to allow a full E step computation. A non-exhaustive list includes the Incremental EM Neal and Hinton (1998); Ng and McLachlan (2003), the Online EM Cappé and Moulines (2009) and, more recently, the stochastic EM with variance reduction (sEM-vr) Chen *et al.* (2018), the fast Incremental EM (FIEM) Karimi *et al.* (2019), the Stochastic Path-Integrated Differential EstimatoR EM (SPIDER-EM) Fort *et al.* (2020), and the mini-batch MCMC-SAEM Kuhn *et al.* (2020). Most of these variants come with their own theoretical convergence guarantees for the models of the exponential family. Recent works have also provided theoretical analysis of the EM algorithm outside of the exponential family, with locally strongly-concave log-likelihood function around the global maxima by Balakrishnan *et al.* (2017), or without such strong-concavity assumption by Dwivedi *et al.* (2020a).

The stochastically approximated EM algorithms constitute an extensive catalogue of methods. Indeed, there are many possible variants of MCMC samplers that can be considered, as well as the additional parameters, such as the burn-in period length and the gain decrease sequence, that have to be set. All these choices have an impact on the convergence of these "EM-like" algorithms and making the appropriate ones for each problem can be overwhelming, see Booth and Hobert (1999); Levine and Casella (2001); Levine and Fan (2004), among others, for discussions on tuning the MC-EM alone. On several cases, one might desire to dispose of a simpler method, possibly non-stochastic, and non-parametric to run an EM-like algorithm for models with no closed forms. However the literature is lacking in that regards. The Quasi-Monte Carlo EM, introduced by Pan and Thompson (1998), is a deterministic version of Monte Carlo EM, however theoretical guarantees are not provided. In that vein, Jank (2005) introduces the randomised Quasi-Monte Carlo EM, which is not deterministic, and does not have theoretical guarantees either. Another example of deterministic approximation can be found in the Variational EM Attias (1999); Bishop (2006); Tzikas *et al.* (2008), where the conditional distribution is approximated within a certain distribution family, and optimised over before each E step. Note that with such a design, if the distance between the true conditional and the chosen family is non-zero, then the approximation can never converge towards the true conditional.

In addition to intractable E steps, EM procedures also commonly face a second issue: their convergence, when guaranteed, can be towards any maximum. This theoretical remark has crucial numerical implications. Indeed, most of the time, convergence is reached towards a sub-optimal local maximum, usually very dependent on the initialisation. To tackle this issue and improve the solutions of the algorithm, other types of, usually deterministic, approximations of the E step have been proposed. One notable example is the tempering (or "annealing") of the conditional probability function used in the E step. Instead of replacing an intractable problem by a tractable one, the tempering approximation is used to find better local maxima of the likelihood profile during the optimisation process, in the spirit of the Simulated Annealing of Kirkpatrick et al. (1983) and the Parallel Tempering (or Annealing MCMC) of Swendsen and Wang (1986); Geyer and Thompson (1995). The deterministic annealing EM was introduced by Ueda and Nakano (1998) with a decreasing temperature profile; another temperature profile was proposed by Naim and Gildea (2012). Contrary to most of the studies on stochastic approximations, these two works do not provide theoretical convergence guarantees for the proposed tempered methods, which, as a consequence, does not provide insight on the choice of the temperature scheme. Moreover, the tempered methods do not allow the use of the EM in case of an intractable E step. In their tempered SAEM algorithm, Allassonnière and Chevallier (2021) combine the stochastic and tempering approximations, which allows the SAEM to run, even with an intractable E step, while benefiting from the improved optimisation properties brought by the tempering. In addition, theoretical convergence guarantees are provided. However, this method is once again stochastic and parametric.

Overall, most of the literature on approximated E steps focuses on stochastic approximations that esti-

mate intractable conditional probability functions. The few purely deterministic approximations proposed, such as the tempered/annealed EM, are used for other purposes, improving the optimisation procedure, and lack convergence guarantees.

In this paper, we prove that, with mild model assumptions and regularity conditions on the approximation, any Deterministic Approximate EM benefits from the state of the art theoretical convergence guarantees of Wu (1983); Lange (1995); Delyon *et al.* (1999). This theorem covers several already existing methods, such as the tempered EM, and paves the way for the exploration of new ideas. To illustrate this, we study a small selection of methods with practical applications that verify our theorem's hypotheses. First, for E steps without closed form, we propose to use Riemann sums to estimate the intractable normalising factor. This "Riemann approximation EM" is a deterministic, less parametric, alternative to the MC-EM and its variants. It is suited to the small dimension cases, where the analytic estimation of the intractable integral is manageable, and MC estimation unnecessary. Second, we prove that the deterministic annealed EM (or "tempered EM") of Ueda and Nakano (1998) is a member of our general deterministic class as well. We prove that the convergence guarantees are achieved with almost no condition of the temperature scheme, justifying the use of a wider range of temperature profile than those proposed by Ueda and Nakano (1998) and Naim and Gildea (2012). Finally, since the Riemann and tempered approximations are two separate methods that fulfil very different practical purposes, we also propose to associate the two approximations in the "tempered Riemann approximation EM" when both their benefits are desired.

In Section 2.1, we introduce our general class of deterministic approximated versions of the EM algorithm and prove their convergence guarantees, for models of the exponential family. We discuss the "Riemann approximation EM" in Section 2.2, the "tempered EM" in Section 2.3, and their association, "tempered Riemann approximation EM", in Section 2.4. In Section 3, we study empirically each of these three methods. Section 4 discusses and concludes this work.

We demonstrate empirically that the Riemann EM converges properly on a model with and an intractable E step, and that adding the tempering to the Riemann approximation allows in addition to get away from the initialisation and recover the true parameters. On a tractable Gaussian Mixture Model, we compare the behaviours and performances of the tempered EM and the regular EM. In particular, we illustrate that the tempered EM is able to escape adversarial initialisations, and consistently reaches better values of the likelihood than the unmodified EM, in addition to better estimating the model parameters.

2 Materials and Methods

2.1 Deterministic Approximate EM Algorithm and Its Convergence for the Exponential Family

2.1.1 Context and Motivation

In this section, we propose a new class of deterministic EM algorithms with approximated E step. This class of algorithms is very general. In particular, it includes methods with deterministic approximations of intractable normalisation constant. It also includes optimisation-oriented approximations, such as the annealing approximation used to escape sub-optimal local extrema of the objective. In addition, combinations of these two types of approximations are also covered. We prove that members of this class benefit from the same convergence guarantees that can be found in the state of the art references (Wu, 1983; Lange, 1995; Delyon *et al.*, 1999) for the classical EM algorithm, and under similar model assumptions. The only condition on the approximated distribution being that it converges towards the real conditional probability distribution with a l_2 regularity. Like the authors of Delyon *et al.* (1999); Fort and Moulines (2003); Allassonnière and Chevallier (2021), we work with probability density functions belonging to the exponential family. The specific properties of which are providen in the hypothesis $M1$ of Theorem 1.

EM algorithms are most often considered in the context of the *missing data problem*, see Delyon *et al.* (1999). The formulation of the problem is the following: we observe a random variable x, described by a parametric probability density function (pdf) noted $g(\theta)$ with parameter $\theta \in \Theta \subset \mathbb{R}^l$, $l \in \mathbb{N}^*$. We assume

that there exists a hidden variable z informing the behaviour of the observed variable x such that $g(\theta)$ can be expressed as the integral of the complete likelihood $h(z;\theta)$: $g(\theta) = \int_z h(z;\theta)\mu(dz)$, with μ the reference measure. We denote $p_\theta(z) := h(z;\theta)/g(\theta)$ the conditional density function of z. For a given sample x, the goal is to maximise the likelihood $g(\theta)$ with respect to θ. In all these notations, and throughout the rest of the paper, we omit x as a variable since it is considered fixed once and for all, and everything is done conditionally to x. In particular, in many applications, the observed data x is made of $N \in \mathbb{N}^*$ samples: $x := (x^{(1)}, x^{(N)})$. In such cases, the sample size $N < +\infty$ is supposed finite and fixed once and for all. We are never in the asymptotic regime $N \longrightarrow +\infty$.

The foundation of the EM algorithm is that while $\ln g(\theta)$ is hard to maximise in θ, the functions $\theta \mapsto \ln h(z;\theta)$ and even $\theta \mapsto \mathbb{E}_z[\ln h(z;\theta)]$ are easier to work with because of the information added by the latent variable z (or its distribution). In practice however, the actual value of z is unknown and its distribution $p_\theta(z)$ dependent on θ. Hence, the EM was introduced by DLR Dempster *et al.* (1977) as the two-stages procedure starting from an initial point θ_0 and iterated over the number of steps n:

(E) With the current parameter θ_n, calculate the conditional probability $p_{\theta_n}(z)$;

(M) To get θ_{n+1}, maximise in $\theta \in \Theta$ the function $\theta \mapsto \mathbb{E}_{z\sim p_{\theta_n}(z)}[\ln h(z;\theta)]$;

Which can be summarised as:

$$\theta_{n+1} := T(\theta_n) := \underset{\theta\in\Theta}{\operatorname{argmax}}\, \mathbb{E}_{z\sim p_{\theta_n}(z)}[\ln h(z;\theta)]\ , \tag{1}$$

where we call T the point to point map in Θ corresponding to one EM step. We will not redo the basic theory of the exact EM here, but this procedure noticeably increase $g(\theta_n)$ at each new step n. However, as discussed in the introduction, in many cases, one may prefer to or need to use an approximation of $p_{\theta_n}(z)$ instead of the exact analytical value.

In the following, we consider a deterministic approximation of $p_\theta(z)$ noted $\tilde{p}_{\theta,n}(z)$ which depends on the current step n and on which we make no assumption at the moment. The resulting steps, defining the "Approximate EM", can be written under the same form as (1):

$$\theta_{n+1} := F_n(\theta_n) := \underset{\theta\in\Theta}{\operatorname{argmax}}\, \mathbb{E}_{z\sim \tilde{p}_{\theta_n,n}(z)}[\ln h(z;\theta)]\ , \tag{2}$$

where $\{F_n\}_{n\in\mathbb{N}}$ is the sequence of point to point maps in Θ associated with the sequence of approximations $\{\tilde{p}_{\theta,n}(z)\}_{\theta\in\Theta;\, n\in\mathbb{N}}$. In order to ensure the desired convergence guarantees, we add a slight modification to this EM sequence: *re-initialisation of the EM sequence onto increasing compact sets*. This modification was introduced by Chen *et al.* (1987) and adapted by Fort and Moulines (2003). Assume that you dispose of an increasing sequence of compacts $\{K_n\}_{n\in\mathbb{N}}$ such that $\cup_{n\in\mathbb{N}} K_n = \Theta$ and $\theta_0 \in K_0$. Define $j_0 := 0$. Then, the transition $\theta_{n+1} = F_n(\theta_n)$ is accepted only if $F_n(\theta_n)$ belongs to the current compact K_{j_n}, otherwise the sequence is reinitialised at θ_0. The steps of the resulting algorithm, called *Stable Approximate EM*, can be written as:

$$\begin{cases} if\, F_n(\theta_n) \in K_{j_n}, \text{ then } \theta_{n+1} = F_n(\theta_n), \text{ and } j_{n+1} := j_n \\ if\, F_n(\theta_n) \notin K_{j_n}, \text{ then } \theta_{n+1} = \theta_0, \text{ and } j_{n+1} := j_n + 1\,. \end{cases} \tag{3}$$

This re-initialisation of the EM sequence may seem like a hurdle, however, this truncation is only a theoretical requirement. In practice, the first compact K_0 is taken so large that it covers the most probable areas of Θ and the algorithms (2) and (3) are identical as long as the sequence $\{\theta_n\}_n$ does not diverges towards the border of Θ.

2.1.2 Theorem

In the following, we will state the convergence Theorem of Equation (3) and provide a brief description of the main steps of the proof.

Theorem 1 (Convergence of the Stable Approximate EM)**.** *Let $\{\theta_n\}_{n\in\mathbb{N}}$ be a sequence of the Stable Approximate EM defined in Equation (3). Let us assume two sets of hypotheses:*

- ***Conditions M1–3 on the model.***

  M1. *$\Theta \subseteq \mathbb{R}^l$, $\mathcal{X} \subseteq \mathbb{R}^d$ and μ is a σ-finite positive Borel measure on $\mathcal{X}$. Let $\psi : \Theta \to \mathbb{R}$, $\phi : \Theta \to \mathbb{R}^q$ and $S : \mathcal{X} \to \mathcal{S} \subseteq \mathbb{R}^q$. Define $L : \mathcal{S} \times \Theta \to \mathbb{R}$, $h : \mathcal{X} \times \Theta \to \mathbb{R}_+^*$ and $g : \Theta \to \mathbb{R}_+^*$ as:*

$$\begin{aligned} L(s;\theta) &:= \psi(\theta) + \langle s, \phi(\theta)\rangle \,, \\ h(z;\theta) &:= \exp(L(S(z);\theta)) \,, \\ g(\theta) &:= \int_{z\in\mathcal{X}} h(z;\theta)\mu(dz) \,. \end{aligned}$$

  M2. *Assume that*

  (a) ψ and ϕ are continuous on Θ;*

  (b) for all $\theta \in \Theta$, $\bar{S}(\theta) := \int_z S(z)p_\theta(z)\mu(dz)$ is finite and continuous on Θ;

  (c) there exists a continuous function $\hat{\theta} : \mathcal{S} \to \Theta$ such that for all $s \in \mathcal{S}$, $L(s;\hat{\theta}(s)) = \sup_{\theta\in\Theta} L(s;\theta)$;

  (d) g is positive, finite and continuous on Θ and the level set $\{\theta \in \Theta, g(\theta) \geq M\}$ is compact for any $M > 0$.

  M3. *Define $\mathcal{L} := \left\{\theta \in \Theta | \hat{\theta} \circ \bar{S}(\theta) = \theta\right\}$ and, for any $g^* \in \mathbb{R}_+^*$, $\mathcal{L}_{g^*} := \{\theta \in \mathcal{L} | g(\theta) = g^*\}$. Assume either that:*

  (a) The set $g(\mathcal{L})$ is compact or

  (a′) for all compact sets $K \subseteq \Theta$, $g(K \cap \mathcal{L})$ is finite.

- ***The conditions on the approximation.*** *Assume that $\tilde{p}_{\theta,n}(z)$ is deterministic. Let $S(z) = \{S_u(z)\}_{u=1}^q$. For all indices u, for any compact set $K \subseteq \Theta$, one of the two following configurations holds:*

$$\begin{cases} \int_z S_u^2(z)dz < \infty \,, \\ \sup_{\theta\in K} \int_z \left(\tilde{p}_{\theta,n}(z) - p_\theta(z)\right)^2 dz \underset{n\to\infty}{\longrightarrow} 0 \,. \end{cases} \tag{4}$$

  Or

$$\begin{cases} \sup_{\theta\in K} \int_z S_u^2(z)p_\theta(z)dz < \infty \,, \\ \sup_{\theta\in K} \int_z \left(\frac{\tilde{p}_{\theta,n}(z)}{p_\theta(z)} - 1\right)^2 p_\theta(z)dz \underset{n\to\infty}{\longrightarrow} 0 \,. \end{cases} \tag{5}$$

Then,

(i) (a) $\lim_{n\to\infty} j_n < \infty$ and $\sup_{n\in\mathbb{N}} \|\theta_n\| < \infty$, with probability 1;

(b) $g(\theta_n)$ converges towards a connected component of $g(\mathcal{L})$.

(ii) Let $Cl : \mathcal{P}(\Theta) \to \Theta$ be the set closure function and $d : \Theta \times \mathcal{P}(\Theta) \to \mathbb{R}_+$ be any point-to-set distance function within Θ. If $g\left(\mathcal{L} \cap Cl\left(\{\theta_n\}_{n\in\mathbb{N}}\right)\right)$ has an empty interior, then $\exists g^* \in \mathbb{R}_+^*$ such that:

$$\begin{aligned} g(\theta_n) &\underset{n\to\infty}{\longrightarrow} g^* \,, \\ d(\theta_n, \mathcal{L}_{g^*}) &\underset{n\to\infty}{\longrightarrow} 0 \,. \end{aligned}$$

Remark 1.

- *When g is differentiable on Θ, its stationary points are the stable points of the EM, i.e. $\mathcal{L} = \{\theta \in \Theta | \nabla g(\theta) = 0\}$.*
- *The M1–3 conditions in Theorem 1 are almost identical to the similarly named M1–3 conditions in Fort and Moulines (2003). The only difference is in $M2$ (a^*), where we remove the hypothesis that S has to be a continuous function of z, since it is not needed when the approximation is not stochastic.*
- *The property $\int_z S_u^2(z)dz < \infty$ of the condition (4) can seem hard to verify since S is not integrated here against a probability density function. However, when z is a finite variable, as is the case for finite mixtures, this integral becomes a finite sum. Hence, condition (4) is better adapted to finite mixtures, while condition (5) is better adapted to continuous hidden variables.*
- *The two sufficient conditions (4) and (5) involve a certain form of l_2 convergence of $\tilde{p}_{\theta,n}$ towards p_θ. If the latent variable z is continuous, this excludes countable (and finite) approximations such as sums of Dirac functions, since they have a measure of zero. In particular, this excludes Quasi-Monte Carlo approximations. However, one look at the proof of Theorem 1 (in Appendix A.1) or at the following sketch of proof reveals it is actually sufficient to verify $\sup_{\theta \in K} \left\| \tilde{S}_n(\theta) - \bar{S}(\theta) \right\| \underset{n \to \infty}{\longrightarrow} 0$ for any compact set K. Where $\tilde{S}_n(\theta) := \int_z S(z)\tilde{p}_{\theta,n}(z)\mu(dz)$ denotes the approximated E step in the Stable Approximate EM. This condition can be verified by finite approximations.*

2.1.3 Sketch of Proof

The detailed proof of this result can be found in Appendix A.1, we propose here an abbreviated version where we highlight the key steps.

Two intermediary propositions, introduced and proven in Fort and Moulines (2003), are instrumental in the proof of Theorem 1. These two propositions are called Propositions 9 and 11 by Fort and Moulines, and used to prove their Theorem 3. Within our framework, with the new condition $M2(a^*)$ and the absence of Monte Carlo sum, the reasoning for verifying the conditions of applicability of the two proposition is quite different from Fort and Moulines (2003) and will be highlighted below. Let us state these two propositions using the notations of this paper:

Proposition 2 ("Proposition 9")**.** *Consider a parameter space $\Theta \subseteq \mathbb{R}^l$, a point to point map T on Θ, a compact $K \subset \Theta$ and a subset $\mathcal{L} \subseteq \Theta$ such that $\mathcal{L} \cap K$ is compact. Let g be a C^0, Lyapunov function relative to $(T, \mathcal{L})$. Assume that there exist a $K-$valued sequence $\{\theta_n\}_{n \in \mathbb{N}^*} \in K^{\mathbb{N}}$ such that:*

$$\lim_{n \to \infty} |g(\theta_{n+1}) - g \circ T(\theta_n)| = 0 \,.$$

Then

- *$\{g(\theta_n)\}_{n \in \mathbb{N}}$ converges towards a connected component of $g(\mathcal{L} \cap K)$*
- *If $g(\mathcal{L} \cap K)$ has an empty interior, then $\exists g^* \in \mathbb{R}_+^*$ such that $\{g(\theta_n)\}_n$ converges towards g^*. Moreover, $\{\theta_n\}_n$ converges towards the set $\mathcal{L}_{g^*} \cap K$. Where $\mathcal{L}_{g^*} := \{\theta \in \mathcal{L} | g(\theta) = g^*\}$.*

Proposition 3 ("Proposition 11")**.** *Consider a parameter space $\Theta \subseteq \mathbb{R}^l$, a subset $\mathcal{L} \subseteq \Theta$, a point to point map T on Θ and a sequence of point to point maps $\{F_n\}_{n \in \mathbb{N}^*}$ also on Θ. Let $\{\theta_n\}_{n \in \mathbb{N}} \in \Theta^{\mathbb{N}}$ be a sequence defined from $\{F_n\}_n$ by the Stable Approximate EM equation (3) for some increasing sequence $\{K_n\}_{n \in \mathbb{N}}$ of compacts of Θ. Let $\{j_n\}_{n \in \mathbb{N}}$ be the corresponding sequence of indices, also defined in (3), such that $\forall n$, $\theta_n \in K_{j_n}$. We assume:*

(a) the $C1-2$ conditions of Proposition 10 of Fort and Moulines (2003).

- ***(C1)** There exists g, a C^0 Lyapunov function relative to $(T, \mathcal{L})$ such that for all $M > 0$, $\{\theta \in \Theta | g(\theta) > M\}$ is compact, and:*

$$\Theta = \cup_{n \in \mathbb{N}} \left\{\theta \in \Theta | g(\theta) > n^{-1}\right\} \,.$$

– ***(C2)*** *$g(\mathcal{L})$ is compact OR* ***(C2')*** *$g(\mathcal{L} \cap K)$ is finite for all compact $K \subseteq \Theta$.*

(b) $\forall \theta \in K_0, \quad \lim_{n\to\infty} |g \circ F_n - g \circ T|(\theta) = 0$

(c) $\forall$ *compact* $K \subseteq \Theta$*:*

$$\lim_{n\to\infty} |g \circ F_n(\theta_n) - g \circ T(\theta_n)| \mathbb{1}_{\theta_n \in K} = 0\,.$$

Then:

With probability 1, $\limsup_{n\to\infty} j_n < \infty$ *and* $\{\theta_n\}_n$ *is a compact sequence.*

Remark 2. *In Fort and Moulines (2003), condition $C1$ of Proposition 3 is mistakenly written as:*

$$\Theta = \cup_{n\in\mathbb{N}} \{\theta \in \Theta | W(\theta) > n\}\,.$$

This is a typo that we have corrected here.

The proof of Theorem 1 is structured as follows: verifying the conditions of Proposition 3, applying Proposition 3, verifying the conditions of Proposition 2 and finally applying Proposition 2.

Verifying the conditions of Proposition 3. Let g be the observed likelihood function defined in hypothesis $M1$ of Theorem 1, T the exact EM point to point map defined in (1), $\mathcal{L} := \{\theta \in \Theta | T(\theta) = \theta\}$ the set of its stable points, $\{F_n\}_n$ a sequence of approximated point to point map as defined in (2). With some increasing sequence $\{K_n\}_{n\in\mathbb{N}}$ of compacts of Θ, let $\{\theta_n, j_n\}_{n\in\mathbb{N}}$ be defined from the Stable Approximate EM equation (3).

By design of the regular EM, the following two properties are true. First, g is a C^0, Lyapunov function relative to $(T, \mathcal{L})$. Second, the map T can be written $T := \hat{\theta} \circ \bar{S}$, with the $\hat{\theta}$ and $\bar{S}$ defined in condition $M2$ of Theorem 1. These properties, in conjunction with hypotheses $M1 - 3$ of Theorem 1, directly imply that condition (a) of Proposition 3 is verified.

The steps to verify conditions (b) and (c) differ from those in the proof of Fort and Moulines (2003). Let $\tilde{S}_n(\theta_n) = \int_z S(z)\tilde{p}_{\theta,n}(z)\mu(dz)$ be the approximated E step in the Stable Approximate EM, such that $F_n = \hat{\theta} \circ \tilde{S}_n$. By using uniform continuity properties on compacts, we prove that the following condition is sufficient to verify both (a) and (b):

$$\forall \text{ compact } K, \ \sup_{\theta\in K} \left\| \tilde{S}_n(\theta) - \bar{S}(\theta) \right\| \underset{n\to\infty}{\longrightarrow} 0\,. \tag{6}$$

Then, upon replacing $\tilde{S}_n$ and $\bar{S}$ by their integral forms, it becomes clear that each of the hypotheses (4) or (5) of Theorem 1 is sufficient to verify (6). In the end, all conditions are verified to apply Proposition 3.

Applying Proposition 3. The application of Proposition 3 to the Stable Approximate EM tells us that with probability 1:

$\limsup_{n\to\infty} j_n < \infty$ and $\{\theta_n\}_{n\in\mathbb{N}}$ is a compact sequence.

Which is specifically the result $(i)(a)$ of Theorem 1.

Verifying the conditions of Proposition 2. We already have that the likelihood g is a C^0, Lyapunov function relative to $(T, \mathcal{L})$. Thanks to Proposition 3, we have that $K := Cl\left(\{\theta_n\}_n\right)$ is a compact. Then, $\mathcal{L} \cap K$ is also compact thanks to hypothesis $M3$. Moreover, by definition, the EM sequence verifies: $\{\theta_n\}_n \in K^{\mathbb{N}}$. The last condition that remains to be shown to apply Proposition 2 is that:

$$\lim_{n\to\infty} |g(\theta_{n+1}) - g \circ T(\theta_n)| = 0\,.$$

If we apply (c) of Proposition 3 with $K = Cl\left(\{\theta_n\}_n\right)$, we get an almost identical result, but with θ_{n+1} replaced by $F_n(\theta_n)$. However, $F_n(\theta_n) \neq \theta_{n+1}$ only when $j_{n+1} = j_n + 1$:

$$|g(\theta_{n+1}) - g \circ T(\theta_n)| = |g(\theta_0) - g \circ T(\theta_n)| \mathbb{1}_{j_{n+1}=j_n+1} + |g \circ F_n(\theta_n) - g \circ T(\theta_n)| \mathbb{1}_{j_{n+1}=j_n}\,.$$

We have proven with Proposition 3 that there is only a finite number of such increments. Hence, when n is large enough, $F_n(\theta_n) = \theta_{n+1}$ always, and we have the desired result.

Applying Proposition 2 Since we verify all we need to apply the conclusions of Proposition 2:

- $\{g(\theta_n)\}_{n\in\mathbb{N}}$ converges towards a connected component of $g(\mathcal{L} \cap Cl(\{\theta_n\}_n)) \subset g(\mathcal{L})$.
- If $g(\mathcal{L} \cap Cl(\{\theta_n\}_n))$ has an empty interior, then the sequence $\{g(\theta_n)\}_{n\in\mathbb{N}}$ converges towards a $g^* \in \mathbb{R}$ and $\{\theta_n\}_n$ converges towards the set $\mathcal{L}_{g^*} \cap Cl(\{\theta_n\}_n) \subseteq \mathcal{L}_{g^*}$.

Which are both respectively exactly $(i)(b)$ and (ii) of Theorem 1 and conclude the proof of the Theorem.

2.2 Riemann Approximation EM

2.2.1 Context and Motivation

In this section, we introduce one specific case of Approximate EM useful in practice: approximating the conditional probability density function $p_\theta(z)$ at the E step by a Riemann sum, in the scenario where the latent variable z is continuous and bounded. We call this procedure the "Riemann approximation EM". After motivating this approach, we prove that it is an instance of the Approximate EM algorithm and verifies the hypotheses of Theorem 1, therefore benefits from the convergence guarantees.

Consider the case where z is a continuous variable and its conditional probability $p_\theta(z)$ is a continuous function. Even when $h(z;\theta)$ can be computed point by point, a closed form may not exist for the renormalisation term $g(\theta) = \int_z h(z;\theta)dz$. In that case, this integral is usually approximated stochastically with a Monte Carlo estimation, see for instance Delyon *et al.* (1999); Fort and Moulines (2003); Allassonnière and Chevallier (2021). When the dimension is reasonably small, a deterministic approximation through Riemann sums can also be performed. For the user this can be a welcome simplification, since MCMC methods have a high hyper-parameter complexity (burn-in duration, gain decrease sequence, Gibbs sampler, etc.), whereas the Riemann approximation involves only the position of the Riemann intervals. The choice of which is very guided by the well known theories of integration (Lagrange, Legendre, etc.), and demonstrated in our experiments to have little impact.

We introduce the Riemann approximation as a member of the Approximate EM class. Since z is supposed bounded in this section, without loss of generality, we will assume that z is a real variable and $z \in [0,1]$. We recall that $p_\theta(z) = h(z;\theta)/g(\theta) = h(z;\theta)/\int_z h(z;\theta)dz$. Instead of using the exact joint likelihood $h(z;\theta)$, we define a sequence of step functions $\left\{\tilde{h}_n\right\}_{n\in\mathbb{N}^*}$ as: $\tilde{h}_n(z;\theta) := h(\lfloor\varphi(n)z\rfloor/\varphi(n);\theta)$. Where φ is an increasing function from $\mathbb{N}^* \to \mathbb{N}^*$ such that $\varphi(n) \underset{n\to\infty}{\longrightarrow} \infty$. For the sake of simplicity, we will take $\varphi = Id$, hence $\tilde{h}_n(z;\theta) = h(\lfloor nz\rfloor/n;\theta)$. The following result, however, can be applied to any increasing function φ with $\varphi(n) \underset{n\to\infty}{\longrightarrow} \infty$.

With these steps functions, the renormalising factor $\tilde{g}_n(\theta) := \int_z \tilde{h}_n(z;\theta)dz$ is now a finite sum. That is: $\tilde{g}_n(\theta) = \frac{1}{n}\sum_{k=0}^{n-1} h(\lfloor kz\rfloor/n;\theta)$. The approximate conditional probability $\tilde{p}_n(\theta)$ is then naturally defined as: $\tilde{p}_n(\theta) := \tilde{h}_n(z;\theta)/\tilde{g}_n(\theta)$. Thanks to the replacement of the integral by the finite sum, this deterministic approximation is much easier to compute than the real conditional probability.

2.2.2 Theorem and Proof

We state and prove the following Theorem for the convergence of the EM with a Riemann approximation.

Theorem 4. *Assume that the hidden variable z is continuous and bounded. Consider the approximation*

$$\tilde{p}_{n,\theta}(z) := \frac{h(\lfloor nz\rfloor/n;\theta)}{\int_{z'} h(\lfloor nz'\rfloor/n;\theta)dz'}\,.$$

We call "Riemann approximation EM" the associated Stable Approximate EM. Under conditions $M1-3$ of Theorem 1, if $z \mapsto S(z)$ is continuous, then the conclusions of Theorem 1 hold for the Riemann approximation EM.

Proof. Under the current assumptions, it is sufficient to verify condition (4) in order to apply Theorem 1. Without loss of generality, we will assume that the bounded variable z is contained in $[0,1]$. Since S is continuous, the first part of the condition is easily verified: $\int_{z=0}^{1} S_u^2(z)dz < \infty$.

For the second part of the condition, we consider a compact $K \subseteq \Theta$. First, note that $h(z;\theta) = \exp(\psi(\theta) + \langle S(z), \phi(\theta)\rangle)$ is continuous in (z,θ), hence uniformly continuous on the compact set $[0,1]\times K$. Additionally, we have:

$$0 < m := \min_{(z,\theta)\in[0,1]\times K} h(z;\theta) \leq h(z;\theta)\,,$$
$$h(z;\theta) \leq \max_{(z,\theta)\in[0,1]\times K} h(z;\theta) =: M < \infty\,.$$

where m and M are constants independent of z and θ. This also means that $m \leq g(\theta) = \int_{z=0}^{1} h(z;\theta) \leq M$. Moreover, since $\tilde{h}_n(z;\theta) = h\left(\lfloor nz\rfloor/n;\theta\right)$, then we also have $\forall z \in [0,1]\,, \theta \in K, n \in \mathbb{N},\ m \leq \tilde{h}_n(z;\theta) \leq M$ and $m \leq \tilde{g}_n(\theta) = \int_{z=0}^{1} \tilde{h}_n(z;\theta) \leq M$.

As h is uniformly continuous, $\forall \epsilon > 0, \exists \delta > 0, \forall (z,z') \in [0,1]^2\,, (\theta,\theta') \in K^2$:

$$|z - z'| \leq \delta \text{ and } \|\theta - \theta'\| \leq \delta \Rightarrow |h(z;\theta) - h(z';\theta')| \leq \epsilon\,.$$

By definition, $\lfloor nz\rfloor/n - z \leq 1/n$. Hence $\exists N \in \mathbb{N}, \forall n \geq N, \lfloor nz\rfloor/n - z \leq \delta$. As a consequence:

$$\forall \epsilon > 0,\ \exists N \in \mathbb{N},\ \forall n \geq N,\ \forall (z,\theta) \in [0,1]\times K, \quad \left|h(z;\theta) - \tilde{h}_n(z;\theta)\right| \leq \epsilon\,.$$

In other words, $\{\tilde{h}_n\}_n$ converges uniformly towards h. Let us fix ϵ and N now. Then, $\forall n \geq N,\ \forall (z,\theta) \in [0,1]\times K$:

$$\begin{aligned}
\tilde{p}_{\theta,n}(z) - p_\theta(z) &= \frac{\tilde{h}_n(z;\theta)}{\int_z \tilde{h}_n(z;\theta)dz} - \frac{h(z;\theta)}{\int_z h(z;\theta)dz} \\
&= \frac{\tilde{h}_n(z;\theta) - h(z;\theta)}{\int_z \tilde{h}_n(z;\theta)dz} + h(z;\theta)\frac{\int_z \left(h(z;\theta) - \tilde{h}_n(z;\theta)\right)dz}{\int_z h(z;\theta)dz \int_z \tilde{h}_n(z;\theta)dz} \\
&\leq \frac{\epsilon}{m} + M\frac{\epsilon}{m^2} \\
&= \epsilon\frac{m+M}{m^2}\,.
\end{aligned}$$

Hence, for this ϵ, $\forall n \geq N$:

$$\sup_{\theta\in K} \int_{z=0}^{1} \left(\tilde{p}_{\theta,n}(z) - p_\theta(z)\right)^2 dz \leq \epsilon^2 \left(\frac{m+M}{m^2}\right)^2\,,$$

which, by definition, means that

$$\sup_{\theta\in K} \int_{z=0}^{1} \left(\tilde{p}_{\theta,n}(z) - p_\theta(z)\right)^2 dz \underset{n\to\infty}{\longrightarrow} 0\,.$$

With this, condition (4) is fully verified, and the conclusions of Theorem 1 are applicable, which concludes the proof. □

2.3 Tempered EM

2.3.1 Context and Motivation

In this section, we consider another particular case of Deterministic Approximate EM: the Tempered EM (or "tmp-EM"), first introduced in Ueda and Nakano (1998). We first prove that under mild conditions, tmp-EM verifies the hypotheses of Theorem 1, hence benefits from the state of the art EM convergence guarantees. In particular, we prove that the choice of the temperature profile is almost completely free. This justifies the use of a wider array of temperature profiles than the ones specified in Ueda and Nakano (1998); Naim and Gildea (2012). Then, we demonstrate the interest of tmp-EM with non-monotonous profiles, on Mixture estimation problems with hard to escape initialisations.

Tempering or *annealing* is a common technique in the world of non-convex optimisation. With a non-convex objective function g, naively following the gradients usually leads to undesirable local extrema close to the initial point. A common remedy is to elevate g to the power $\frac{1}{T_n}$, with $\{T_n\}_{n\in\mathbb{N}}$ a sequence of temperatures tending towards 1 as the number n of steps of the procedure increases. When $T_n > 1$, the shape of the new objective function $g^{\frac{1}{T_n}}$ is flattened, making the gradients less strong, and the potential wells less attractive, but without changing the hierarchy of the values. This allows the optimisation procedure to explore more before converging. This concept is put in application in many state of the art procedures. The most iconic probably being the Simulated Annealing, introduced and developed in Kirkpatrick et al. (1983); Van Laarhoven and Aarts (1987); Aarts and Korst (1988), where in particular $T_n \longrightarrow 0$ instead of 1. It is one of the few optimisation technique proven to find global optimum of non-convex functions. The Parallel Tempering (or Annealing MCMC) developed in Swendsen and Wang (1986); Geyer and Thompson (1995); Hukushima and Nemoto (1996) also makes use of these ideas to improve the MCMC simulation of a target probability distribution.

Since non-convex objectives are a common occurrence in the EM literature, Ueda and Nakano (1998) introduced the *Deterministic Annealed EM*, where, in the E step, the conditional probability is replaced by a tempered approximated distribution: $\tilde{p}_{n,\theta}(z) \propto p_\theta^{\frac{1}{T_n}}(z) \propto h(z;\theta)^{\frac{1}{T_n}}$ (renormalised such that $\int_z \tilde{p}_{n,\theta}(z)dz = 1$). For the temperature profile they consider specific sequences $\{T_n\}_{n\in\mathbb{N}}$ decreasingly converging to 1. Another specific, non-monotonous, temperature scheme was later proposed by Naim and Gildea (2012). In both cases, theoretical convergence guarantees are lacking. Later, in Allassonnière and Chevallier (2021), tempering has been applied to the SAEM, with convergence guarantees provided for any temperature scheme $T_n \longrightarrow 1$.

With Theorem 5, we show that the Deterministic Anealing EM, or tmp-EM, is a specific case of the Deterministic Approximate EM of Section 2.1, hence can benefit from the same convergence properties. In particular, we show that any temperature scheme $\{T_n\}_n \in (\mathbb{R}^*)^{\mathbb{N}}$, $T_n \underset{n\to\infty}{\longrightarrow} 1$ guarantees the state of the art convergence.

Remark 3. *Elevating $p_\theta(z)$ to the power $\frac{1}{T_n}$, as is done here and in Ueda and Nakano (1998); Naim and Gildea (2012), is not equivalent to elevating to the power $\frac{1}{T_n}$ the objective function $g(\theta)$, which would be expected for a typical annealed or tempered optimisation procedure. It is not equivalent either to elevating to the power $\frac{1}{T_n}$ the proxy function $\mathbb{E}_{z\sim p_{\theta_n}(z)}[h(z;\theta)]$ that is optimised in the M step. Instead, the weights $p_{\theta_n}(z)$ (or equivalently, the terms $h(z;\theta_n)$) used in the calculation of $\mathbb{E}_{z\sim p_{\theta_n}(z)}[h(z;\theta)]$ are the tempered terms. This still results in the desired behaviour and is only a more "structured" tempering. Indeed, with this tempering, it is the estimated distribution of the latent variable z that are made less unequivocal, with weaker modes, at each step. This forces the procedure to spend more time considering different configurations for those variables, which renders as a result the optimised function $\mathbb{E}_{z\sim p_{\theta_n}(z)}[h(z;\theta)]$ more ambiguous regarding which θ is the best, just as intended. Then, when n large, the algorithm is allowed to converge, as $T_n \longrightarrow 1$ and $\mathbb{E}_{z\sim\tilde{p}_{n,\theta}(z)} \longrightarrow \mathbb{E}_{z\sim p_\theta(z)}$.*

2.3.2 Theorem

We now provide the convergence Theorem for the Approximate EM with the tempering approximation. In particular, this result highlights that there are almost no constraints on the temperature profile to achieve convergence.

Theorem 5. *Let T_n be a sequence of non-zero real numbers. Consider the approximation introduced in Ueda and Nakano (1998):*

$$\tilde{p}_{n,\theta}(z) := \frac{p_\theta^{\frac{1}{T_n}}(z)}{\int_{z'} p_\theta^{\frac{1}{T_n}}(z')dz'} .$$

We call "Tempered EM" the associated Stable Approximate EM. Define $\overline{\mathcal{B}}(1,\epsilon)$ be the closed ball centred in 1 and with radius $\epsilon \in \mathbb{R}_+$. Under conditions $M1-3$ of Theorem 1, if $T_n \underset{n\to\infty}{\longrightarrow} 1$ and for any compact $K \subseteq \Theta$, $\exists \epsilon \in]0,1[$, $\forall \alpha \in \overline{\mathcal{B}}(1,\epsilon)$:

(T1) $\sup_{\theta\in K} \int_z p_\theta^\alpha(z)dz < \infty$,

(T2) $\forall u \in [\![1,q]\!]$, $\sup_{\theta\in K} \int_z S_u^2(z) p_\theta^\alpha(z)dz < \infty$,

then the conclusions of Theorem 1 hold for the Tempered EM.

Remark 4. *The added condition that $(T1)$ and $(T2)$ must hold for all α in a ball $\overline{\mathcal{B}}(1,\epsilon)$ is very mild. Indeed, in Section 2.3.4, we show classical examples that easily verify the much stronger condition that $(T1)$ and $(T2)$ hold for all $\alpha \in \mathbb{R}_+^*$.*

2.3.3 Sketch of Proof

The detailed proof of Theorem 5 can be found in Appendix A.2, we propose here an abbreviated version where we highlight the key steps.
Under the current assumptions, it is sufficient to verify the second part of condition (5) in order to apply Theorem 1. To that end, we must control the integral

$$\int_z \left(\frac{\tilde{p}_{\theta,n}(z)}{p_\theta(z)} - 1\right)^2 p_\theta(z)dz ,$$

for all θ in a compact $K \subseteq \Theta$. First, with a Taylor development in the term $\left(\frac{1}{T_n} - 1\right)$, which converges toward 0 when $n \to \infty$, we control the difference $(\tilde{p}_{\theta,n}(z) - p_\theta(z))^2$:

$$\left(\frac{p_\theta(z)^{\frac{1}{T_n}}}{\int_{z'} p_\theta(z')^{\frac{1}{T_n}}} - p_\theta(z)\right)^2 \leq 2\left(\frac{1}{T_n} - 1\right)^2 p_\theta(z)^2 \left(\left(\ln p_\theta(z)\, e^{a(z,\theta,T_n)}\right)^2 A(\theta,T_n) + B(\theta,T_n)\right) .$$

where the terms $A(\theta,T_n)$, $B(\theta,T_n)$ and $a(z,\theta,T_n)$ come from the Taylor development. Then, we can control the integral of interest:

$$\begin{aligned}\int_z \frac{\left(\frac{p_\theta(z)^{\frac{1}{T_n}}}{\int_{z'} p_\theta(z')^{\frac{1}{T_n}}} - p_\theta(z)\right)^2}{p_\theta(z)} dz \leq & 2\left(\frac{1}{T_n} - 1\right)^2 A(\theta,T_n) \int_z p_\theta(z) e^{2a(z,\theta,T_n)} \ln^2 p_\theta(z)dz \\ & + 2\left(\frac{1}{T_n} - 1\right)^2 B(\theta,T_n) .\end{aligned} \tag{7}$$

From the properties of the Taylor development, we prove that $A(\theta, T_n)$ and $B(\theta, T_n)$ both have upper bounds involving only $\int_z p_\theta(z) \ln p_\theta(z)$ and $\int_z p_\theta(z)^{\frac{1}{T_n}} \ln p_\theta(z)$. In a similar fashion, the term $\int_z p_\theta(z) e^{2a(z,\theta,T_n)} \ln^2 p_\theta(z)$ has an upper bound involving only $\int_z p_\theta(z) \ln^2 p_\theta(z) dz$ and $\int_z p_\theta(z)^{\frac{2}{T_n}-1} \ln^2 p_\theta(z) dz$.

Using the hypotheses of the Theorem, we prove that for any $\alpha \in \overline{\mathcal{B}}(1, \epsilon)$ and $\theta \in K$ the terms $\int_z p_\theta(z)^\alpha \ln p_\theta(z)$ and $\int_z p_\theta(z)^\alpha \ln^2 p_\theta(z)$ are both upper bounded by a constant C independent of θ and α.

Since $T_n \underset{n\to\infty}{\longrightarrow} 1$, then when n is large enough, $\frac{1}{T_n} \in \overline{\mathcal{B}}(1, \epsilon)$ and $\frac{2}{T_n} - 1 \in \overline{\mathcal{B}}(1, \epsilon)$. Hence, the previous result applies to the upper bounds of $A(\theta, T_n)$, $B(\theta, T_n)$ and $\int_z p_\theta(z) e^{2a(z,\theta,T_n)} \ln^2 p_\theta(z) dz$. As a result, these three terms are respectively upper bounded by C_1, C_2 and C_3, three constants independent of θ and T_n.

The inequality (7) then becomes:

$$\int_z \frac{1}{p_\theta(z)} \left(\frac{p_\theta(z)^{\frac{1}{T_n}}}{\int_{z'} p_\theta(z')^{\frac{1}{T_n}}} - p_\theta(z) \right)^2 dz \leq 2 \left(\frac{1}{T_n} - 1 \right)^2 C_1 C_2 + 2 \left(\frac{1}{T_n} - 1 \right)^2 C_3 \,.$$

By taking the supremum in $\theta \in K$ and the limit when $n \longrightarrow \infty$, we get the desired result:

$$\sup_{\theta \in K} \int_z \frac{1}{p_\theta(z)} \left(\frac{p_\theta(z)^{\frac{1}{T_n}}}{\int_{z'} p_\theta(z')^{\frac{1}{T_n}}} - p_\theta(z) \right)^2 dz \underset{n\to\infty}{\longrightarrow} 0 \,.$$

With condition (5) verified, the conclusions of Theorem 1 are applicable, which concludes the proof.

2.3.4 Examples of Models That Verify the Conditions

In this section, we illustrate that the conditions of Theorem 5 are easily met by common models. We take two examples, first the Gaussian Mixture Model (GMM) where the latent variables belong to a finite space, then the Poisson count with random effect, where the latent variables live in a continuous space. As mentioned, these examples are shown to not only verify all hypotheses of Theorem 5, but also to verify $(T1)$ and $(T2)$ for any $\alpha \in \mathbb{R}_+^*$.

Gaussian Mixture Model Despite being one of the most common models the EM is applied to, the GMM have many known irregularities and pathological behaviours, see Titterington *et al.* (1985); Ho *et al.* (2016). Although some recent works, such as Dwivedi *et al.* (2020a,b), tackled the theoretical analysis of EM for GMM, none of the convergence results for the traditional EM and its variants proposed by Wu (1983); Lange (1995); Delyon *et al.* (1999); Fort and Moulines (2003) apply to the GMM. The hypothesis that the GMM fail to verify is the condition that the level lines have to be compact ($M2$ (d) in this paper). All is not lost however for the GMM, indeed, the model verifies all the other hypotheses of the general Theorem 1 as well as the tempering hypotheses of Theorem 5. Moreover, in this paper as in the others, the only function of the unverified hypothesis $M2$ (d) is to ensure in the proof that the EM sequence stays within a compact. The latter condition is the actual relevant property to guarantee convergence at this stage of the proof. This means that, in practice, if an tempered EM sequence applied to a GMM is observed to remain within a compact, then all the conditions for convergence are met, Theorem 5 applies, and the sequence is guaranteed to converge towards a critical point of the likelihood function.

In the following, we provide more details on the GMM likelihoods and the theorem hypotheses they verify. First, note that the GMM belongs to the exponential family with the complete likelihood:

$$h(z;\theta) = \prod_{i=1}^{N} \sum_{k=1}^{K} exp\Bigg(\frac{1}{2} \Big(- (x^{(i)} - \mu_k)^T \Theta_k (x^{(i)} - \mu_k) + \ln(|\Theta_k|) + 2\ln(\pi_k) - p\ln(2\pi) \Big) \Bigg) \mathbb{1}_{z^{(i)}=k} \,, \tag{8}$$

and the observed likelihood:

$$g(\theta) = \prod_{i=1}^{N} \sum_{k=1}^{K} exp\left(\frac{1}{2}\Big(- (x^{(i)} - \mu_k)^T \Theta_k (x^{(i)} - \mu_k) + \ln(|\Theta_k|) + 2\ln(\pi_k) - p\ln(2\pi)\Big)\right). \tag{9}$$

This is an exponential model with parameter:

$$\theta := \left(\{\pi_k\}_{k=1}^K, \{\mu_k\}_{k=1}^K, \{\Theta_k\}_{k=1}^K\right) \in \left\{ \{\pi_k\}_k \in [0,1]^K \,\middle|\, \sum_k \pi_k = 1 \right\} \otimes \mathbb{R}^{p\times K} \otimes S_p^{++K}.$$

where S_p^{++} is the cone of symmetric positive definite matrices of size p. The verification of conditions *M1–3* for the GMM (except $M2$ (d) of course) is a classical exercise since these are the conditions our Theorem shares with any other EM convergence result on the exponential family. We focus here on the hypotheses specific to our Deterministic Approximate EM.

Condition on $\int_z p_\theta^\alpha(z)dz$. Let $\alpha \in \mathbb{R}_+^*$, in the finite mixture case, the integrals on z are finite sums:

$$\int_z p_\theta^\alpha(z)dz = \sum_k p_\theta^\alpha(z=k)\,.$$

Which is continuous in θ since $\theta \mapsto p_\theta(z=k) = h(z=k;\theta)/g(\theta)$ is continuous. Hence

$$\forall \alpha \in \mathbb{R}_+^*, \quad \sup_{\theta\in K} \int_z p_\theta^\alpha(z)dz < \infty\,.$$

Condition on $\int_z S_u^2(z) p_\theta^\alpha(z)dz$. The previous continuity argument is still valid.

Poisson Count with Random Effect This model is discussed in Fort and Moulines (2003), the authors prove, among other things, that this model verifies the conditions $M1-3$. Here is a brief description of the model: the hidden variables $\left\{z^{(i)}\right\}_{i=1}^N$ are distributed according to an autoregressive process of order 1: $z^{(i)} := az^{(i-1)} + \sigma\epsilon_i$, with $|a| < 1$, $\sigma > 0$ and the $\left\{\epsilon^{(i)}\right\}_{i=1}^N$ are iid standard gaussian. Conditionally to $\left\{z^{(i)}\right\}_{i=1}^N$, the observed variables $x^{(i)}$ are independent Poisson variables with parameter $\exp(\theta + z^{(i)})$. The complete likelihood of the model, not accounting for irrelevant constants, is:

$$h(z;\theta) = e^{\theta\sum_{i=1}^N x^{(i)}}.exp\left(-e^\theta \sum_{i=1}^N e^{z^{(i)}}\right). \tag{10}$$

$g(\theta) = \int_z h(z;\theta)dz$ can be computed analytically up to a constant:

$$\begin{aligned} g(\theta) &= \int_{z\in\mathbb{R}^N} h(z;\theta)dz \\ &= e^{\theta\sum_{i=1}^N x^{(i)}} \int_{z\in\mathbb{R}^N} exp\left(-e^\theta \sum_{i=1}^N e^{z^{(i)}}\right)dz \\ &= e^{\theta\sum_{i=1}^N x^{(i)}} \prod_{i=1}^N \int_{z^{(i)}\in\mathbb{R}} exp\left(-e^\theta e^{z^{(i)}}\right)dz^{(i)} \\ &= e^{\theta\sum_{i=1}^N x^{(i)}} \left(\int_{u\in\mathbb{R}_+} \frac{exp\,(-u)}{u}du\right)^N \\ &= e^{\theta\sum_{i=1}^N x^{(i)}} E_1(0)^N\,, \end{aligned} \tag{11}$$

where $E_1(0)$ is a finite, non zero, constant, called "exponential integral", in particular independent of α and θ.

Condition on $\int_z p_\theta^\alpha(z)dz$. Let K be a compact in Θ.
We have $p_\theta(z) = \frac{h(z;\theta)}{g(\theta)}$. Let us compute $\int_z h(z;\theta)^\alpha$ for any positive α. The calculations work as in Equation (11):

$$\begin{aligned}\int_{z\in\mathbb{R}^N} h(z;\theta)^\alpha &= e^{\alpha\theta\sum_{i=1}^N x^{(i)}} \prod_{i=1}^N \int_{z^{(i)}\in\mathbb{R}} exp\left(-\alpha e^\theta e^{z^{(i)}}\right) dz^{(i)} \\ &= e^{\alpha\theta\sum_{i=1}^N x^{(i)}} E_1(0)^N \,.\end{aligned}$$

Hence:

$$\int_z p_\theta^\alpha(z)dz = E_1(0)^{(1-\alpha)N} \,.$$

Since $E_1(0)$ is finite, non zero, and independent of θ, we easily have:

$$\forall\alpha\in\mathbb{R}_+^*, \quad \sup_{\theta\in K}\int_z p_\theta^\alpha(z)dz < \infty \,.$$

θ does not even have to be restricted to a compact.

Condition on $\int_z S_u^2(z)p_\theta^\alpha(z)dz$. Let K be a compact in Θ and α a positive real number.
In this Poisson count model, $S(z) = \sum_{i=1}^N e^{z^{(i)}} \in \mathbb{R}$. We have:

$$S^2(z)p_\theta^\alpha(z) = \left(\sum_{i=1}^N e^{z^{(i)}}\right)^2 \frac{exp\left(-\alpha e^\theta \sum_{i=1}^N e^{z^{(i)}}\right)}{E_1(0)^{\alpha N}} \,. \tag{12}$$

First, let us prove that the integral is finite for any θ. We introduce the variables $u_k := \sum_{l=1}^k e^{z_l}$. The Jacobi matrix is triangular and its determinant is $\prod_k e^{z^{(i)}} = \prod_k u_k$.

$$\int_z S^2(z)p_\theta^\alpha(z)dz = \frac{\int_{z\in\mathbb{R}^N}\left(\sum_k e^{z^{(i)}}\right)^2 exp\left(-\alpha e^\theta \sum_k e^{z^{(i)}}\right)dz}{E_1(0)^{\alpha N}} \,.$$

Which is proportional to:

$$\int_{u_1=0}^{+\infty} u_1 \int_{u_2=u_1}^{+\infty} u_2 ... \int_{u_N=u_{N-1}}^{+\infty} u_N^3\, e^{-\alpha e^\theta u_N}\, du_N ... du_2 du_1 \,.$$

where we removed the finite constant $\frac{1}{E_1(0)^{\alpha N}}$ for clarity. This integral is finite for any θ because the exponential is the dominant term around $+\infty$. Let us now prove that $\theta \mapsto \int_z S^2(z)p_\theta^\alpha(z)dz$ is continuous. From Equation (12), we have that

- $z \mapsto S^2(z)p_\theta^\alpha(z)$ is measurable on $\mathbb{R}^N$.
- $\theta \mapsto S^2(z)p_\theta^\alpha(z)$ is continuous on K (and on $\Theta = \mathbb{R}$).
- With $\theta_M := \min_{\theta\in K}\theta$, then $\forall\theta\in K,\ 0 \le S^2(z)p_\theta^\alpha(z) \le S^2(z)p_{\theta_M}^\alpha(z)$

Since we have proven that $S^2(z)p_{\theta_M}^\alpha(z) < \infty$, then we can apply the intervertion Theorem and state that $\theta \mapsto \int_z S^2(z)p_\theta^\alpha(z)dz$ is continuous.
It directly follows that:

$$\forall\alpha\in\mathbb{R}_+^*, \quad \sup_{\theta\in K}\int_z S^2(z)p_\theta^\alpha(z)dz < \infty \,.$$

Note that after the change of variable, the integral could be computed explicitly, but involves N successive integration of polynomial $\times$ exponential function products of the form $P(x)e^{-\alpha e^{\theta} x}$. This would get tedious, especially since after each successful integration, the product with the next integration variable u_{k-1} increases by one the degree of the polynomial, i.e. starting from 3, the degree ends up being $N+2$. We chose a faster path.

2.4 Tempered Riemann Approximation EM

Context, Theorem and Proof

The Riemann approximation of Section 2.2 makes the EM computations possible in hard cases, when the conditional distribution has no analytical form for instance. It is an alternative to the many stochastic approximation methods (SAEM, MCMC-SAEM, etc.) that are commonly used in those cases. The tempering approximation of Section 2.3 is used to escape the initialisation by allowing the procedure to explore more the likelihood profile before committing to convergence. We showed that both these approximation are particular cases of the wider class of Deterministic Approximate EM, introduced in Section 2.1. However, since they fulfil different purposes, it is natural to use them in coordination and not as alternatives of one another. In this section, we introduce another instance of the Approximate EM: a combination of the tempered and Riemann sum approximations. This "tempered Riemann approximation EM" (tmp-Riemann approximation) can compute EM steps when there is no closed form thanks to the Riemann sums as well as escape the initialisation thanks to the tempering. For a bounded latent variable $z \in [0,1]$, we define the approximation as: $\tilde{p}_{n,\theta}(z) := h(\lfloor nz \rfloor/n;\theta)^{\frac{1}{T_n}} / \int_{z'} h(\lfloor nz' \rfloor/n;\theta)^{\frac{1}{T_n}} dz'$, for a sequence $\{T_n\}_n \in (\mathbb{R}_+^*)^{\mathbb{N}}$, $T_n \underset{n\to\infty}{\longrightarrow} 1$.

In the following Theorem, we prove that the tempered Riemann approximation EM verifies the applicability conditions of Theorem 1 with no additional hypothesis from the regular Riemann approximation EM covered by Theorem 4.

Theorem 6. *Under conditions $M1-3$ of Theorem 1, and when z is bounded, the (Stable) Approximate EM with $\tilde{p}_{n,\theta}(z) := \frac{h(\lfloor nz \rfloor/n;\theta)^{\frac{1}{T_n}}}{\int_{z'} h(\lfloor nz' \rfloor/n;\theta)^{\frac{1}{T_n}} dz'}$, which we call "tempered Riemann approximation EM", verifies the remaining conditions of applicability of Theorem 1 as long as $z \mapsto S(z)$ is continuous and $\{T_n\}_n \in (\mathbb{R}_+^*)^{\mathbb{N}}$, $T_n \underset{n\to\infty}{\longrightarrow} 1$.*

Proof. This proof of Theorem 6 is very similar to the proof of Theorem 4 for the regular Riemann approximation EM. The first common element is that for the tempered Riemann approximation EM, the only remaining applicability condition of the general Theorem 1 to prove is also:

$$\forall \text{compact } K \subseteq \Theta,\ \sup_{\theta\in K} \int_z \left(\tilde{p}_{\theta,n}(z) - p_\theta(z)\right)^2 dz \underset{n\to\infty}{\longrightarrow} 0\,.$$

In the proof of Theorem 4, we proved that having the uniform convergence of the approximated complete likelihood $\{\tilde{h}_n\}_n$ towards the real h - with both $\tilde{h}_n(z;\theta)$ and $h(z;\theta)$ uniformly bounded - was sufficient to fulfil this condition. Hence, we prove In this section, that these sufficient properties still hold, even with the tempered Riemann approximation, where $\tilde{h}_n(z;\theta) := h(\lfloor nz \rfloor/n;\theta)^{\frac{1}{T_n}}$.

We recall that $h(z;\theta)$ is uniformly continuous on the compact set $[0,1]\times K$, and verifies:

$$0 < m \leq h(z;\theta) \leq M < \infty\,.$$

where m and M are constants independent of z and θ.

Since $T_n > 0$, $T_n \underset{n\to\infty}{\longrightarrow} 1$, then the sequence $\{1/T_n\}_n$ is bounded. Since $\tilde{h}_n(z;\theta) = h\left(\lfloor nz \rfloor/n;\theta\right)^{\frac{1}{T_n}}$, with $0 < m \leq h\left(\lfloor nz \rfloor/n;\theta\right) \leq M < \infty$ for any z, θ and n, then we also have:

$$0 < m' \leq \tilde{h}_n(z;\theta) \leq M' < \infty\,,$$

with m' and M' constants independent of z, θ and n.

We have seen in the proof of Theorem 4, that:

$$\forall \epsilon > 0,\ \exists N \in \mathbb{N},\ \forall n \geq N,\ \forall (z,\theta) \in [0,1] \times K, \quad |h(z;\theta) - h\left(\lfloor nz \rfloor / n; \theta\right)| \leq \epsilon\,.$$

To complete the proof, we control in a similar way the difference $h\left(\lfloor nz \rfloor / n; \theta\right) - h\left(\lfloor nz \rfloor / n; \theta\right)^{\frac{1}{T_n}}$. The function $(h, T) \in [m, M] \times [T_{min}, T_{max}] \mapsto h^{\frac{1}{T}} \in \mathbb{R}$ is continuous on a compact, hence uniformly continuous in (h, T). As a consequence: $\forall \epsilon > 0, \exists \delta > 0, \forall (h, h') \in [m, M]^2, (T, T') \in [T_{min}, T_{max}]^2$,

$$|h - h'| \leq \delta \text{ and } |T - T'| \leq \delta \implies \left| h^{\frac{1}{T}} - (h')^{\frac{1}{T'}} \right| \leq \epsilon\,.$$

Hence, with $N \in \mathbb{N}$ such that $\forall n \geq N, |T_n - 1| \leq \delta$, we have:

$$\begin{aligned} &\forall n \geq N, \forall (z,\theta) \in [0,1] \times K, \\ &\qquad \left| h\left(\lfloor nz \rfloor / n; \theta\right) - h\left(\lfloor nz \rfloor / n; \theta\right)^{\frac{1}{T_n}} \right| \leq \epsilon\,. \end{aligned}$$

In the end, $\forall \epsilon > 0, \exists N \in \mathbb{N}, \forall n \geq N, \forall (z,\theta) \in [0,1] \times K$:

$$\begin{aligned} \left| h(z;\theta) - \tilde{h}_n\left(z;\theta\right) \right| &= \left| h(z;\theta) - h\left(\lfloor nz \rfloor / n; \theta\right)^{\frac{1}{T_n}} \right| \\ &\leq |h(z;\theta) - h\left(\lfloor nz \rfloor / n; \theta\right)| + \left| h\left(\lfloor nz \rfloor / n; \theta\right) - h\left(\lfloor nz \rfloor / n; \theta\right)^{\frac{1}{T_n}} \right| \\ &\leq 2\epsilon\,. \end{aligned}$$

In other words, we have the uniform convergence of $\{\tilde{h}_n\}$ towards h. From there, we conclude following the same steps as in the proof of Theorem 4. □

3 Results

In this section, we describe experiments that explore each of the three studied methods: Riemann EM, tempered EM and tempered Riemann.

3.1 Riemann Approximation EM: Two Applications

3.1.1 Application to a Gaussian Model with the Beta Prior

We demonstrate the interest of the method on a example with a continuous bounded random variable following a Beta distribution $z \sim Beta(\alpha, 1)$, and an observed random variable following $x \sim \mathcal{N}(\lambda z, \sigma^2)$. In other words, with $\epsilon \sim \mathcal{N}(0,1)$ independent of z:

$$x = \lambda z + \sigma\epsilon\,.$$

This results in a likelihood belonging to the exponential family:

$$h(z;\theta) = \frac{\alpha z^{\alpha-1}}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x-\lambda z)^2}{2\sigma^2}\right)\,.$$

Since z is bounded, and everything is continuous in the parameter $(\alpha, \lambda, \sigma^2)$, this model easily verifies each of the conditions *M1–3*. The E step with this model involves the integral $\int_z z^\alpha exp\left(-\frac{(x-\lambda z)^2}{2\sigma^2}\right) dz$, a fractional moment of the Gaussian distribution. Theoretical formulas exists for these moments, see Winkelbauer (2012), however they involve Kummer's confluent hypergeometric functions, which are infinite series. Instead, we use the Riemann approximation to run the EM algorithm with this model: $\tilde{h}_n(z;\theta) := h(\lfloor \varphi(n) z \rfloor / \varphi(n); \theta)$.

As done previously, we take, without loss of generality, $\varphi(n) := n$ for the sake of simplicity. The E step only involves the n different values taken by the step function probabilities $h(\lfloor nz \rfloor / n; \theta)$:

$$\tilde{p}^{(i)}_{\theta,n}\left(\frac{k}{n}\right) = \frac{h^{(i)}(\frac{k}{n};\theta)}{\frac{1}{n}\sum_{l=0}^{n-1} h^{(i)}(\frac{l}{n};\theta)}\,.$$

where the exponent (i) indicates the index of the observation $x^{(i)}$. To express the corresponding M step in a digest way, let us define the operator $\Psi^{(i)} : \mathbb{R}^{[0,1]} \longrightarrow \mathbb{R}$ such that, for any $f : [0,1] \longrightarrow \mathbb{R}$:

$$\Psi^{(i)} \circ f = \sum_{k=0}^{n-1} \tilde{p}^{(i)}_{\theta,n}\left(\frac{k}{n}\right) \int_{z=k/n}^{(k+1)/n} f(z)dz\,.$$

Then, the M step can be expressed as:

$$\begin{aligned}
\frac{1}{\hat{\alpha}} &= -\frac{1}{N}\sum_{i=1}^{N} \Psi^{(i)} \circ \ln(z)\,,\\
\hat{\lambda} &= \frac{\sum_{i=1}^{N} \Psi^{(i)} \circ (x^{(i)} z)}{\sum_{i=1}^{N} \Psi^{(i)} \circ z^2}\,,\\
\hat{\sigma}^2 &= \frac{1}{N}\sum_{i=1}^{N} \Psi^{(i)} \circ \left(x^{(i)} - \hat{\lambda} z\right)^2\,.
\end{aligned} \tag{13}$$

where we took the liberty of replacing f by $f(z)$ in these equations for the sake of simplicity. Here N is the total number of observations: $x := (x^{(1)}, ..., x^{(N)})$ iid.

We test this algorithm on synthetic data. With real values $\alpha = 2, \lambda = 5, \sigma = 1.5$, we generate a dataset with $N = 100$ observations and run the Riemann EM with random initialisation. This simulation is ran 2000 times. We observe that the Riemann EM is indeed able to increase the likelihood, despite the EM being originally intractable. On Figure 1, we display the average trajectory, with standard deviation, of the negative log-likelihood $-\ln(g(\theta))$ during the Riemann EM procedure. The profile is indeed decreasing. The standard deviation around the average value is fairly high, since each run involves a different dataset and a different random initialisation, hence different value of the likelihood, but the decreasing trend is the same for all of the runs. We also display the average relative square errors on the parameters at the end of the algorithm. They are all small, with reasonably small standard deviation, which indicates that the algorithm consistently recovers correctly the parameters.

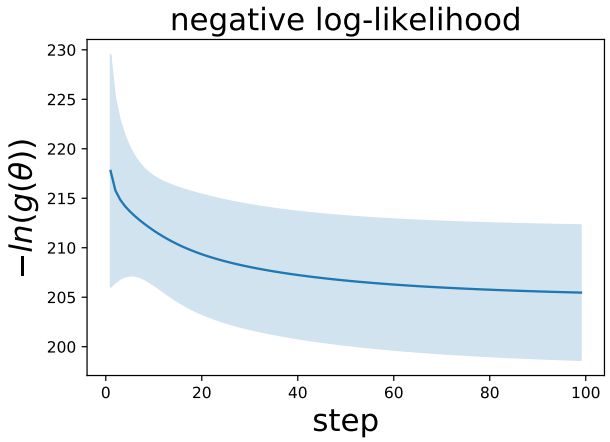
To evaluate the impact of the number of Riemann intervals $\varphi(n)$, we run a second experiment where we compare four different profiles over 50 simulations:

$$\begin{aligned}
\textbf{(low)}\quad & \varphi_1(n) := n + 1\\
\textbf{(medium)}\quad & \varphi_2(n) := n + 100\\
\textbf{(high)}\quad & \varphi_3(n) := n + 1000\\
\textbf{(linear)}\quad & \varphi_4(n) := 10 \times n + 1\,.
\end{aligned}$$

The results are displayed on Figure 2. We can see that, despite the very different profiles, the optimisations are very similar. The "low" profile performs slightly worst, which indicates that a high number of Riemann intervals is most desirable in practice. As long as this number is high enough, Figure 2 suggests that the performances will not depend too much on the profile.

3.1.2 Application in Two Dimensions

The difficulty faced by Riemann methods in general is their geometric complexity when the dimension increases. In this section, we propose a similar experiment in two dimensions to show that the method is still functional and practical in that case.

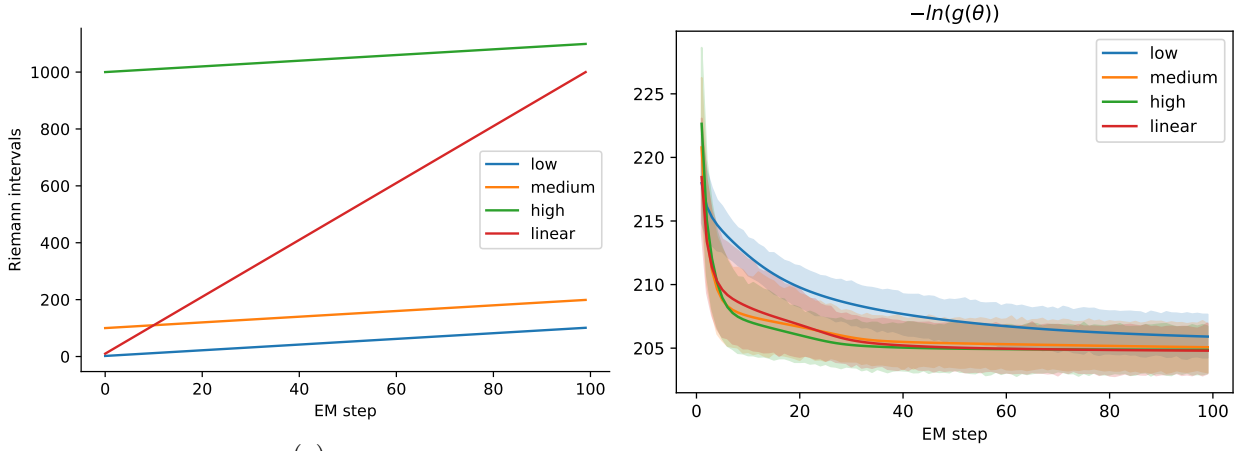


(**a**)

Metric	Average (std)
$\frac{(\alpha-\hat{\alpha})^2}{\alpha^2}$	0.129 (0.078)
$\frac{(\lambda-\hat{\lambda})^2}{\lambda^2}$	0.048 (0.058)
$\frac{(\sigma-\hat{\sigma})^2}{\sigma^2}$	0.108 (0.119)

(**b**)

Figure 1: (**a**). Average values, with standard deviation, over 2000 simulations of the negative log-likelihood along the steps of the Riemann EM. The Riemann EM increases the likelihood. (**b**). Average and standard deviation of the relative parameter reconstruction errors at the end of the Riemann EM.



(**a**) (**b**)

Figure 2: (**a**). Visual representation of the number of Riemann intervals over the EM steps for each profile φ_i. The total number of Riemann intervals computed over 100 EM iterations are: 5150 for "low", 14,950 for "medium", 50,500 for "linear" and 104,950 for "high". (**b**). For each profile, average evolution of the negative log-likelihood, with standard deviation, over 50 simulations. The results are fairly similar, in particular between "medium", "high" and "linear".

For this 2D-model, we consider two latent independent Beta random variables $z_1 \sim Beta(\alpha_1, 1)$ and $z_2 \sim Beta(\alpha_2, 1)$, and two observed variables defined as:

$$x_1 = \lambda_1 z_1 + z_2 + \sigma_1 \epsilon_1$$
$$x_2 = z_1 + \lambda_2 z_2 + \sigma_2 \epsilon_2 \,,$$

with $\epsilon_1 \sim \mathcal{N}(0,1)$, $\epsilon_2 \sim \mathcal{N}(0,1)$, and $(z_1, z_2, \epsilon_1, \epsilon_2)$ independent. The 2-dimension version of the Riemann E step with n intervals on each dimension is:

$$\tilde{p}^{(i)}_{\theta,n}\left(\frac{k_1}{n}, \frac{k_2}{n}\right) = \frac{h^{(i)}(\frac{k_1}{n}, \frac{k_2}{n}; \theta)}{\frac{1}{n^2}\sum_{l_1,l_2=0}^{n-1} h^{(i)}(\frac{l_1}{n}, \frac{l_2}{n}; \theta)} \,.$$

As before, we define an operator $\Psi^{(i)} : \mathbb{R}^{[0,1]^2} \longrightarrow \mathbb{R}$ such that, for any $f : [0,1]^2 \longrightarrow \mathbb{R}$:

$$\Psi^{(i)} \circ f = \sum_{k_1,k_2=0}^{n-1} \tilde{p}^{(i)}_{\theta,n}\left(\frac{k_1}{n}, \frac{k_2}{n}\right) \int_{z_1,z_2=k/n}^{(k+1)/n} f(z_1, z_2) dz \,.$$

Then, the M step can be expressed as:

$$\frac{1}{\hat{\alpha}_1} = -\frac{1}{N}\sum_{i=1}^{N} \Psi^{(i)} \circ \ln(z_1) \,,$$
$$\hat{\lambda}_1 = \frac{\sum_i \Psi^{(i)} \circ (x_1^{(i)} z_1 - z_2 z_1)}{\sum_i \Psi^{(i)} \circ z_1^2} \,,$$
$$\hat{\sigma}_1 = \frac{1}{N}\sum_{i=1}^{N} \Psi^{(i)} \circ \left(x_1^{(i)} - \hat{\lambda}_1 z_1 - z_2\right)^2 \,,$$

with symmetric formulas for $\hat{\alpha}_2, \hat{\lambda}_2$ and $\hat{\sigma}_2$.

For the simulations, we take $(\alpha_1, \alpha_2) = (1, 3)$, $(\lambda_1, \lambda_2) = (10, -10)$ and $(\sigma_1, \sigma_2) = (2, 3)$. From the previous experiment, we keep only the two least costly profiles: "low" $\varphi_1(n) := n + 1$ and "medium" $\varphi_2(n) := n + 100$. We also consider two new, sub-linear, profiles "square root" $\varphi_5(n) := \lfloor\sqrt{n}\rfloor + 1$ and "5 square root" $\varphi_6(n) := 5 \times \lfloor\sqrt{n}\rfloor$. $\varphi_5(n)$ and $\varphi_6(n)$ are designed to have linear complexity even in 2-dimensions.

The results of the EM algorithm runs are displayed on Figure 3. On the left, we follow the number of Riemann squares mapping the 2D space. The difference in computational complexity between profiles is accentuated by the higher dimension. In particular, "medium" performs 6.7 times more computations than "low" and 18.4 times more than "5 square root". However, on the right of Figure 3, we observe that these three profiles perform similar optimisations. This observation justifies cutting computation costs by using lower resolution profiles to compensate the higher dimension.

3.2 Tempered EM: Application to Mixtures of Gaussian

3.2.1 Context and Experimental Protocol

In this section, we will assess the capacity of tmp-EM to escape from deceptive local maxima. We compare the classical EM to tmp-EM with both a monotonous and an oscillating temperature profile on a very well know toy example: likelihood maximisation within the Gaussian Mixture Model. We confront the algorithms to situations where the true classes have increasingly more ambiguous positions, combined with initialisations designed to be hard to escape from. Although the EM is an optimisation procedure, and the log-likelihood reached is a critical metric, in this example, we put more emphasis on the correct positioning of the cluster centroids, that is to say on the recovery of the μ_k. The other usual metrics are also in favour of tmp-EM, and can be found in supplementary materials.

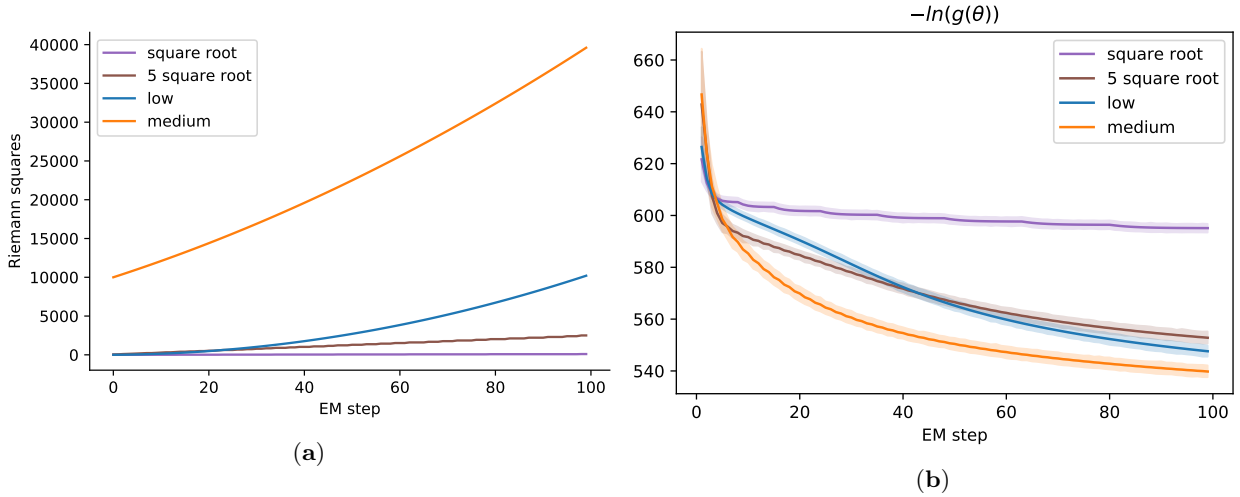


Figure 3: (**a**). Visual representation of the number of Riemann intervals over the EM steps for each profile φ_i. In higher dimension, the computational complexity of the profiles are very different. More precisely, the total number of Riemann squares computed over 100 EM iterations are: 4534 for "square root", 125,662 for "5 square root", 348,550 for "low" and 2,318,350 for "medium". (**b**). For each profile, average evolution of the negative log-likelihood, with standard deviation, over 50 simulations. The "square root" profile performs poorly. However, the other three are comparable despite their different computational complexities.

For the sake of comparison, the experimental design is similar to the one in Allassonnière and Chevallier (2021) on the tmp-SAEM. It is as follows: we have three clusters of similar shape and same weight. One is isolated and easily identifiable. The other two are next to one another, in a more ambiguous configuration. Figure 4 represents the three, gradually more ambiguous configurations. Each configuration is called a "parameter family".

We use two different initialisation types to reveal the behaviours of the three EMs. The first—which we call "*barycenter*"—puts all three initial centroids at the centre of mass of all the observed data points. However, none of the EM procedures would move from this initial state if the three GMM centroids were at the exact same position, hence we actually apply a tiny perturbation to make them all slightly distinct. The blue crosses on Figure 5 represent a typical *barycenter* initialisation. With this initialisation method, we assess whether the EM procedures are able to correctly estimate the positions of the three clusters, despite the ambiguity, when starting from a fairly neutral position, providing neither direction nor misdirection. On the other hand, the second initialisation type - which we call "*2v1*" - is voluntarily misguiding the algorithm by positioning two centroids on the isolated right cluster and only one centroid on the side of the two ambiguous left clusters. The blue crosses on Figure 6 represent a typical *2v1* initialisation. This initialisation is intended to assess whether the methods are able to escape the potential well in which they start and make theirs centroids traverse the empty space between the left and right clusters to reach their rightful position. For each of the three parameter families represented on Figure 4, 1000 datasets with 500 observations each are simulated, and the three EMs are ran with both the *barycenter* and the *2v1* initialisation.

For tmp-EM, we try two profiles. First, a simple *decreasing* exponential profile as seen in Ueda and Nakano (1998): $T_n = 1 + (T_0 - 1)\exp(-r.n)$. Through a grid search, the values $T_0 = 5$, $r = 2$ for the *barycenter* initialisation and $T_0 = 100\,r = 1.5$ for the *2v1* initialisation are picked for this profile. Since Theorem 5 only requires $T_n \longrightarrow 1$, we also propose an *oscillating* profile inspired from Allassonnière and Chevallier (2021). The exact formula of these oscillations is: $T_n = th(\frac{n}{2r}) + (T_0 - b\frac{2\sqrt{2}}{3\pi})\,a^{n/r} + b\,sinc(\frac{3\pi}{4} + \frac{n}{r})$. Where $0 < T_0$, $0 < r$, $0 < b$ and $0 < a < 1$. The oscillations in this profile are meant to achieve a two-regimes behaviour. When the temperature reaches low values, the convergence speed is momentarily increased which has the effect of "locking-in" some of the most obviously good decisions of the algorithm.

Then, the temperature is re-increased to continue the exploration on the other, more ambiguous, parameters. Those two regimes are alternated in succession with gradually smaller oscillations, resulting in a multi-scale procedure that "locks-in" gradually harder decisions. For some hyper-parameter combinations, the sequence T_n can have a (usually small) finite number of negative values. Since only the asymptotic behaviour of T_n is the step n matters for convergence, then the theory allows a finite number of negative values. However, in practice, at least for the sake of interpretation, one may prefer to use only positive values for T_n. In which case, one can either restrain themselves to parameter combinations that result in no negatives values for T_n, or enforce positivity by taking $T_n \leftarrow max(T_n, \epsilon)$ with a certain $\epsilon > 0$.

For our experiments, we select the hyper-parameter values with a grid-search. The "normalised" *sinc* function is used $sinc(x) = sin(\pi x)/(\pi x)$ and the chosen tempering parameters are $T_0 = 5$, $r = 2$, $a = 0.6$, $b = 20$ for the experiments with the *barycenter* initialisation, and $T_0 = 100$, $r = 1.5$, $a = 0.02$, $b = 20$ for the *2v1* initialisation. We have two different sets of tempering hyper-parameters values, one for each of the two very different initialisation types. However, these values then remain the same for the three different parameter families and for every data generation within them. This underlines that the method is not excessively sensitive to the tempering parameters, and that the prior search for good hyper-parameter values is a worthwhile time investment. Likewise, a simple experiment with 6 clusters, in supplementary materials, demonstrates that the same hyper-parameters can be kept over different initialisation (and different data generations as well) when they were made in a non-adversarial way, by drawing random initial centroids uniformly among the data points.

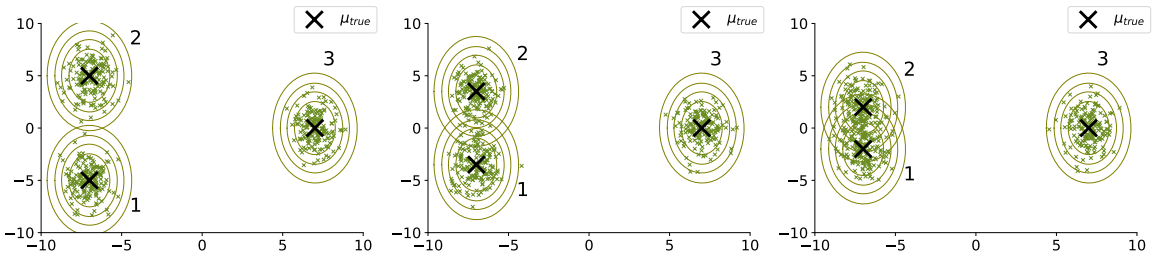


Figure 4: 500 sample points from a Mixture of Gaussians with 3 classes. The true centroid of each Gaussian are depicted by black crosses, and their true covariance matrices are represented by the confidence ellipses of level 0.8, 0.99 and 0.999 around the centre. Each sub-figure corresponds to one of the three different versions of the true parameters. From (**a**) to (**c**): the true μ_k of the two left clusters (μ_1 and μ_2) are getting closer while everything else stays identical.

3.2.2 Experimental Results Analysis

In this section, we analyse the results of EM, *decreasing* tmp-EM and *oscillating* tmp-EM over all the simulations.

First, an illustration: Figures 5 and 6 depict the final states of EM and *oscillating* tmp-EM on one typical simulation for each of the three ambiguity level (the three parameter families) starting from the *barycenter* and *2v1* initialisation respectively. The simulated data are represented by the green crosses. The initial centroids are in blue. The orange cross represents the estimated centroids positions $\hat{\mu}_k$, and the orange confidence ellipses are visual representations of the estimated covariance matrices $\hat{\Sigma}_k$. In supplementary materials, we show step by step the path taken by the estimated parameters of tmp-EM before convergence, providing much more detail on the method's behaviours. These illustrative examples show *oscillating* tmp-EM better succeeding at clustering recovery than the classical EM. The results over all simulations are aggregated in Table 1, and confirm this observation.

Table 1 presents the average and the standard deviation of the relative l_2 error on μ_k of the EMs. For each category, the better result over the three EM is highlighted in bold. The recovery of the true class

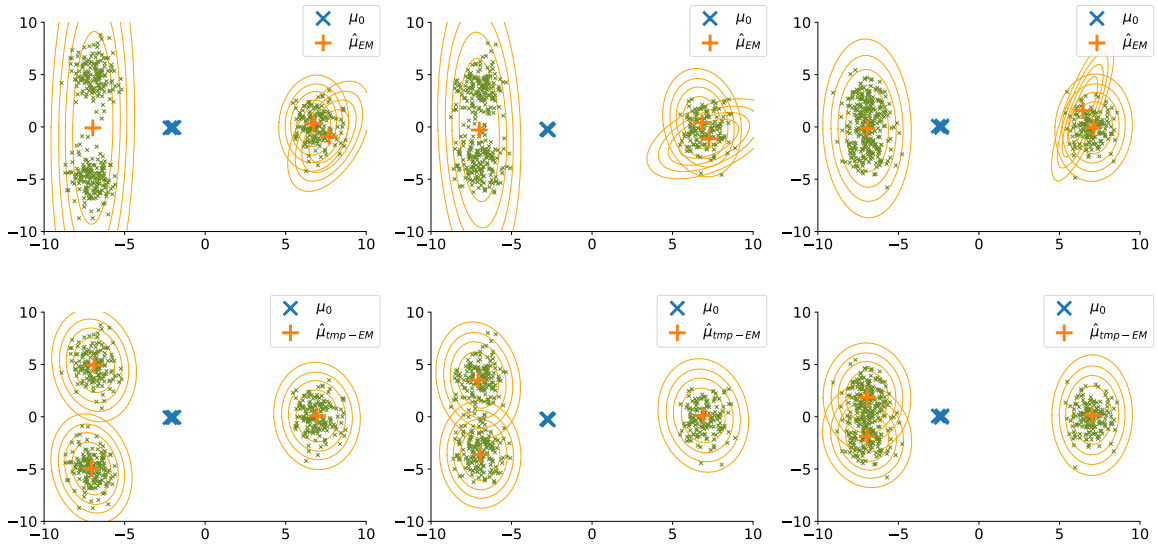


Figure 5: Typical final positioning of the centroids by EM (first row) and tmp-EM with *oscillating* profile (second row) **when the initialisation is made at the barycenter of all data points** (blue crosses). The three columns represent the three gradually more ambiguous parameter sets. Each figure represents the positions of the estimated centroids after convergence of the EM algorithms (orange cross), with their estimated covariance matrices (orange confidence ellipses). In each simulation, 500 sample points were drawn from the real GMM (small green crosses). In those example, tmp-EM managed to correctly identify the position of the three real centroids.

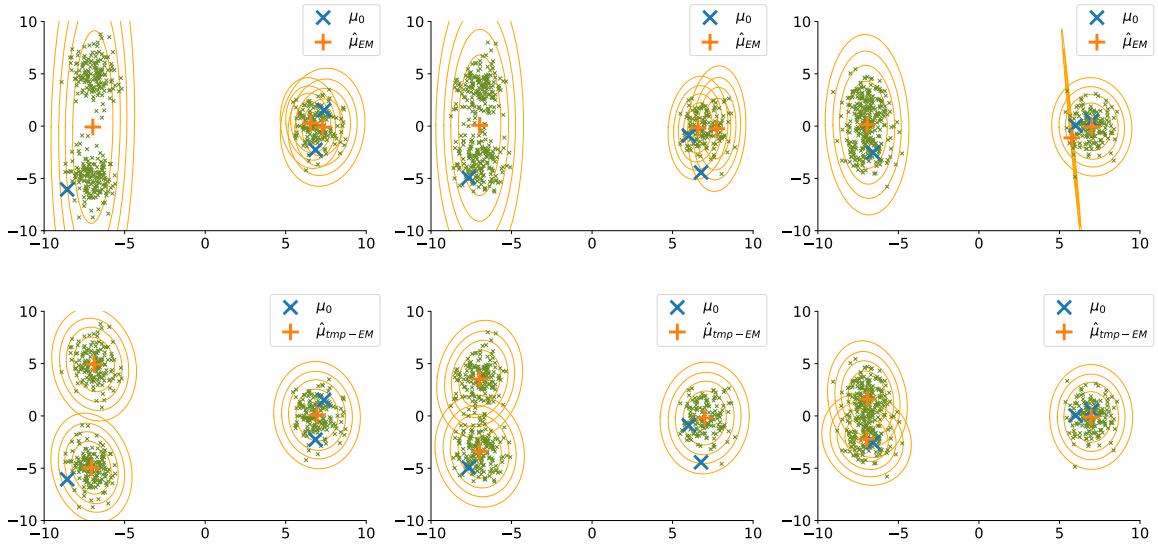


Figure 6: Typical final positioning of the centroids by EM (first row) and tmp-EM with *oscillating* profile (second row) **when the initialisation is made by selecting two points in the isolated cluster and one in the lower ambiguous cluster** (blue crosses). The three columns represent the three gradually more ambiguous parameter sets. Each figure represents the positions of the estimated centroids after convergence of the EM algorithms (orange cross), with their estimated covariance matrices (orange confidence ellipses). In each simulation, 500 sample points were drawn from the real GMM (small green crosses). In those examples, although EM kept two centroids on the isolated cluster, tmp-EM managed to correctly identify the position of the three real centroids.

averages μ_k is spotlighted as it is the essential success metric for this experiment. More specifically, class 1 and 2, the two leftmost classes, are the hardest to correctly recover and the ones whose estimation is the differentiating criterion between the algorithms. Indeed, as can be seen in Table 1, μ_3 is always well estimated by all methods. Hence, in the following, we discuss the error on μ_1 and μ_2.

With the *barycenter* initialisation, classical EM and *decreasing* tmp-EM have similar average relative error levels. With classical EM actually being slightly better. However, *oscillating* tmp-EM is much better than both of them, with error levels smaller by a factor of 10 on parameter families 1 and 2, and by a factor of 5 on parameter family 3. The standard deviation of *oscillating* tmp-EM is also lower, by a factor of roughly 3 on parameter families 1 and 2, and by a factor of 2 on parameter family 3. With the *2v1* initialisation, all error levels are higher. This time, *decreasing* tmp-EM is better in average than classical EM by a factor of 1.7 to 1.4, depending on the parameter family. In turn, *oscillating* tmp-EM is better than *decreasing* tmp-EM by a factor of 3.1 to 3.4 depending on the parameter family. Its standard deviation is also lower by a factor of about 2.

Overall, *oscillating* tmp-EM dominates the simulation. Its error rates on the recovery of μ_1 and μ_2 are always the best, and they remain at low levels even with the most adversarial initialisations. To bolster this last point, we underline that the highest relative error reached by *oscillating* tmp-EM over all the various scenarios (0.39 on parameter family 3 with *2v1* initialisation) is still lower than the lowest relative error of both classical EM (0.52 on parameter family 1 with *barycenter* initialisation) and *decreasing* tmp-EM (0.60 on parameter family 1 with *barycenter* initialisation).

Table 1: Average and standard deviation of the relative error on μ_k, $\frac{\|\hat{\mu}_k - \mu_k\|^2}{\|\mu_k\|^2}$, made by EM, tmp-EM with *decreasing* temperature and tmp-EM with *oscillating* temperature over 1000 simulated dataset with two different initialisations. The three different parameter families, described in Figure 4, correspond to increasingly ambiguous positions of classes 1 and 2. For both initialisations type, the identification of these two clusters is drastically improved by the tempering. Best results highlighted in **bold**.

		EM		**tmp-EM (*Decreasing T*)**		**tmp-EM (*Decreasing oscillating T*)**	
Parameter Family	cl.	barycenter	2v1	barycenter	2v1	barycenter	2v1
1	1	0.52 (1.01)	1.52 (1.24)	0.60 (1.08)	0.87 (1.20)	**0.04 (0.26)**	**0.29 (0.64)**
	2	0.55 (1.05)	1.53 (1.25)	0.64 (1.10)	0.96 (1.25)	**0.05 (0.31)**	**0.30 (0.64)**
	3	**0.01 (0.06)**	0.01 (0.03)	0.01 (0.10)	**0.01 (0.02)**	0.03 (0.17)	0.03 (0.19)
2	1	1.00 (1.42)	1.69 (1.51)	0.96 (1.41)	1.10 (1.46)	**0.09 (0.47)**	**0.37 (0.86)**
	2	1.03 (1.44)	1.71 (1.52)	1.08 (1.46)	1.11 (1.46)	**0.12 (0.57)**	**0.32 (0.79)**
	3	0.01 (0.05)	**0.02 (0.03)**	0.01 **(0.03)**	0.01 (0.04)	**0.01** (0.05)	0.04 (0.22)
3	1	1.56 (1.75)	1.79 (1.77)	1.63 (1.76)	1.38 (1.71)	**0.31 (0.97)**	**0.39 (0.98)**
	2	1.51 (1.74)	1.88 (1.76)	1.52 (1.74)	1.30 (1.68)	**0.30 (0.93)**	**0.39 (0.97)**
	3	0.02 (0.04)	**0.02 (0.04)**	0.01 **(0.03)**	0.02 (0.06)	**0.01** (0.04)	0.07 (0.30)

3.3 Tempered Riemann Approximation EM: Application to a Gaussian Model with Beta Prior

We illustrate the method with the model of Section 3.1.1:

$$h(z;\theta) = \frac{\alpha z^{\alpha-1}}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(y-\lambda z)^2}{2\sigma^2}\right).$$

We apply the tempered Riemann approximation. As in Section 3.1.1, the resulting conditional probability density is a step function defined by the n different values it takes on $[0,1]$. For the observation $x^{(i)}$,

$\forall k \in [\![0, n-1]\!]$:

$$\tilde{p}^{(i)}_{\theta,n}\left(\frac{k}{n}\right) = \frac{h^{(i)}\left(\frac{k}{n};\theta\right)^{\frac{1}{T_n}}}{\frac{1}{n}\sum_{l=0}^{n-1} h^{(i)}\left(\frac{l}{n};\theta\right)^{\frac{1}{T_n}}}\,.$$

The M step, seen in Equation (13), is unchanged. We compare the tempered Riemann EM to the simple Riemann EM on a case where the parameters are ambiguous. With real parameters $\alpha = 0.1, \lambda = 10, \sigma = 0.8$, for each of the 100 simulations, the algorithms are initialised at $\alpha_0 = 10, \lambda_0 = 1, \sigma_0 = 7$. The initialisation is somewhat adversarial, since the mean and variance of the marginal distribution of y are approximately the same with the real of the initialisation parameter, even though the distribution is different. Figure 7 shows that the tempered Riemann EM better escapes the initialisation than the regular Riemann EM, and reaches errors on the parameters orders of magnitude below. The tempering parameters are here $T_0 = 150, r = 3, a = 0.02, b = 40$.

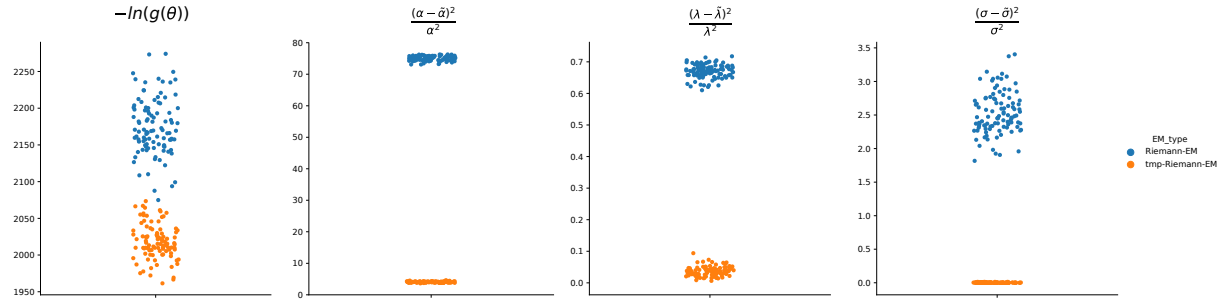


Figure 7: Results over many simulations of the Riemann EM and tmp-Riemann EM on the Beta-Gaussian model. The tempered Riemann EM reaches relative errors on the real parameters that are orders of magnitude below the Riemann EM with no temperature. The likelihood reached is also lower with the tempering.

4 Discussion and Conclusions

We proposed the Deterministic Approximate EM class to bring together the many possible deterministic approximations of the E step. We proved a unified Theorem, with mild conditions on the approximation, which ensures the convergence of the algorithms in this class. Then, we showcased members of this class that solve the usual practical issues of the EM algorithm. For intractable E steps in low dimension, we introduced the Riemann approximation EM, a less parametric and deterministic alternative to the extensive family of MC-EM. We showed on an empirical intractable example how the Riemann approximation EM was able to increase the likelihood and recover every parameter in a satisfactory manner with its simplest design, and no hyper parameter optimisation.

Second, we studied the tempered EM, introduced by Ueda and Nakano (1998) to escape the attractive sub-optimal local extrema of non-convex objectives. We proved that tmp-EM is a specific case of the Deterministic Approximate EM, benefiting from the convergence property as long as the temperature profile converges towards 1. This mild condition justifies the use of many more temperature profiles than the ones tried in Ueda and Nakano (1998); Naim and Gildea (2012). To illustrate the interest of complex, non-monotonous, temperature profiles, we demonstrated on experiments with adversarial initial positions the superiority of an *oscillating* profile over a simple *decreasing* one.

Finally, we added the Riemann approximation in order to apply the tempering in intractable cases. We were then able to show that the tmp-Riemann approximation massively improved the performances of the Riemann approximation, when the initialisation is ambiguous.

Future works will improve both methods. The Riemann approximation will be generalised to be applicable even when the latent variable is not bounded, and an intelligent slicing of the integration space

will improve the computational performances in high dimension. Regarding the tempered EM, since the theory allows the usage of any temperature profile, the natural next step is to look for efficient profiles with few hyper-parameters for fast tuning. Afterwards, implementing an adaptive tuning of the temperature parameters during the procedure will remove the necessity for preliminary grid search altogether.

Acknowledgements The research leading to these results has received funding from the European Research Council (ERC) under grant agreement No 678304, European Union's Horizon 2020 research and innovation program under grant agreement No 666992 (EuroPOND) and No 826421 (TVB-Cloud), and the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and reference ANR-10-IAIHU-06 (IHU-A-ICM).

A Proofs of the Two Main Theorems

In this Appendix, we provide in full details the proofs that were sketched in the main body of the paper. Section A.1 details the proof of convergence for the Deterministic Approximate EM algorithm, our central result, Theorem 1 of the paper. Section A.2 details the proof of convergence for tmp-EM, Theorem 5 of the paper.

A.1 Proof of the General Theorem

In this Section, we prove Theorem 1, which guarantees the convergence of the Deterministic Approximate EM algorithm.

The proof is based on the application of Propositions 2 and 3, taken from Fort and Moulines (2003).

We need to prove that, under the conditions of Theorem 1, we verify the conditions of Proposition Proposition 2 and Proposition 3. Then we will have the results announced in Theorem 1.

A.1.1 Verifying the Conditions of Proposition 3

g is the likelihood function of a model of the curved exponential family. Let T be the point to point map describing the transition between θ_n and θ_{n+1} in the exact EM algorithm. $\mathcal{L}$ the set of stationary points by T: $\mathcal{L} := \{\theta \in \Theta | T(\theta) = \theta\}$. (Note that if g is differentiable, the general properties of the EM tell us that its critical points of g are the stationary points: $\mathcal{L} = \{\theta \in \Theta | \nabla g(\theta) = 0\}$). Additionally, g is a C^0 Lyapunov function associated to $(T, \mathcal{L})$. Let $\{\theta_n\}_n$ be the sequence defined by the stable approximate EM with $\{F_n\}_{n\in\mathbb{N}}$ our sequence of point to point maps.

We verify that under this framework—and with the assumptions of Theorem 1—we check the conditions of Proposition 3.

As in Fort and Moulines (2003), *M1–3* directly implies *C1–2*.

Let us show that we have the last two conditions for Proposition 3:

$$\forall \theta \in K_0, \quad \lim_{n\to\infty} |g \circ F_n - g \circ T|(\theta) = 0\,, \tag{14}$$

and

$$\forall \text{ compact } K \subseteq \Theta, \quad \lim_{n\to\infty} |g \circ F_n(\theta_n) - g \circ T(\theta_n)| \mathbb{1}_{\theta_n \in K} = 0\,. \tag{15}$$

We focus on (15), since (14) is easier to verify and will come from the same reasoning. The first steps are similar to Fort and Moulines (2003). We underline the most consequent deviations from the proof of Fort and Moulines (2003) when they occur.

Equivalent Formulation of the Convergence We write Equation (15) under an equivalent form. Let $\tilde{S}_n(\theta) := \left\{\int_z S_u(z)\tilde{p}_{\theta,n}(z)dz\right\}_{u=1}^q$ and $\bar{S}(\theta) := \left\{\int_z S_u(z)p_\theta(z)dz\right\}_{u=1}^q$. Then $F_n(\theta_n) = \hat{\theta}(\tilde{S}_n(\theta_n))$ and $T(\theta_n) = \hat{\theta}(\bar{S}(\theta_n))$. Hence $|g \circ F_n(u_n) - g \circ T(u_n)| = |g \circ \hat{\theta}(\tilde{S}_n(\theta_n)) - g \circ \hat{\theta}(\bar{S}(\theta_n))|$. To show Equation (15):

$$|g \circ \hat{\theta}(\tilde{S}_n(\theta_n)) - g \circ \hat{\theta}(\bar{S}(\theta_n))|\mathbb{1}_{\theta_n \in K} \underset{n\to\infty}{\to} 0\,,$$

it is sufficient and necessary to have:

$$\forall \epsilon > 0,\, \exists N \in \mathbb{N},\, \forall n \geq N,\, |g \circ \hat{\theta}(\tilde{S}_n(\theta_n)) - g \circ \hat{\theta}(\bar{S}(\theta_n))|\mathbb{1}_{\theta_n \in K} \leq \epsilon\,.$$

An other equivalent formulation is that there are a finite number of integers n such that $|g \circ \hat{\theta}(\tilde{S}_n(\theta_n)) - g \circ \hat{\theta}(\bar{S}(\theta_n))|\mathbb{1}_{\theta_n \in K} > \epsilon$, in other words:

$$\forall \epsilon > 0,\, \sum_{n=1}^{\infty} \mathbb{1}_{|g\circ\hat{\theta}(\tilde{S}_n(\theta_n)) - g\circ\hat{\theta}(\bar{S}(\theta_n))|\mathbb{1}_{\theta_n\in K} > \epsilon} < \infty\,.$$

Use the Uniform Continuity We aim to relate the proximity between the images $g \circ \hat{\theta}$ of to the proximity between the antecedents of $g \circ \hat{\theta}$. The function $g \circ \hat{\theta} : \mathbb{R}^q \to \mathbb{R}$ is continuous, but not necessarily uniformly continuous on $\mathbb{R}^q$. As a consequence, we will need to restrict ourselves to a compact to get uniform continuity properties. We already have a providen compact K. $\tilde{S} : \Theta \to \mathbb{R}^l$ is continuous, hence $S(K)$ is a compact as well. Let δ be a strictly positive real number. Let $\bar{S}(K,\delta) := \left\{ s \in \mathbb{R}^q \middle| \underset{t\in K}{inf} ||\bar{S}(t) - s|| \leq \delta \right\}$. Where we use any norm $||.||$ on $\mathbb{R}^q$ since they are all equivalent. $\bar{S}(K,\delta)$ is a compact set as well. As a consequence $g \circ \theta$ is uniformly continuous on $\bar{S}(K,\delta)$, which means that:

$$\forall \epsilon > 0,\, \exists \eta(\epsilon,\delta) > 0,\, \forall x, y \in \bar{S}(K,\delta),\, \|x - y\| \leq \eta(\epsilon,\delta) \implies |g \circ \hat{\theta}(x) - g \circ \hat{\theta}(y)| \leq \epsilon\,. \tag{16}$$

Let us show that, with $\alpha := min(\delta, \eta(\epsilon,\delta))$, $\forall n$,

$$|g \circ \hat{\theta}(\tilde{S}_n(\theta_n)) - g \circ \hat{\theta}(\bar{S}(\theta_n))|\mathbb{1}_{\theta_n \in K} > \epsilon \implies \left\|\tilde{S}_n(\theta_n) - \bar{S}(\theta_n)\right\| \mathbb{1}_{\theta_n \in K} > \alpha\,. \tag{17}$$

To that end, we show that:

$$\left\|\tilde{S}_n(\theta_n) - \bar{S}(\theta_n)\right\| \mathbb{1}_{\theta_n \in K} \leq \alpha \implies |g \circ \hat{\theta}(\tilde{S}_n(\theta_n)) - g \circ \hat{\theta}(\bar{S}(\theta_n))|\mathbb{1}_{\theta_n \in K} \leq \epsilon\,.$$

Let us assume that $\left\|\tilde{S}_n(\theta_n) - \bar{S}(\theta_n)\right\| \mathbb{1}_{\theta_n \in K} \leq \alpha$.
If $\theta_n \notin K$, then $|g \circ \hat{\theta}(\tilde{S}_n(\theta_n)) - g \circ \hat{\theta}(\bar{S}(\theta_n))|\mathbb{1}_{\theta_n \in K} = 0 \leq \epsilon$.
If, in contrary, $\theta_n \in K$, then $\bar{S}(\theta_n) \in \bar{S}(K) \subset \bar{S}(K,\delta)$.
Since $\left\|\tilde{S}_n(\theta_n) - \bar{S}(\theta_n)\right\| = \left\|\tilde{S}_n(\theta_n) - \bar{S}(\theta_n)\right\| \mathbb{1}_{\theta_n \in K} \leq \alpha \leq \delta$, then $\tilde{S}_n(\theta_n) \in \bar{S}(K,\delta)$.
Since $(\bar{S}(\theta_n), \tilde{S}_n(\theta_n)) \in \bar{S}(K,\delta)^2$ and $\left\|\tilde{S}_n(\theta_n) - \bar{S}(\theta_n)\right\| \leq \alpha \leq \eta(\epsilon,\delta)$, then we get from Equation (16)

$$|g \circ \hat{\theta}(\tilde{S}_n(\theta_n)) - g \circ \hat{\theta}(\bar{S}(\theta_n))|\mathbb{1}_{\theta_n \in K} \leq \epsilon\,.$$

In both cases, we get that:

$$\left\|\tilde{S}_n(\theta_n) - \bar{S}(\theta_n)\right\| \mathbb{1}_{\theta_n \in K} \leq \alpha \implies |g \circ \hat{\theta}(\tilde{S}_n(\theta_n)) - g \circ \hat{\theta}(\bar{S}(\theta_n))|\mathbb{1}_{\theta_n \in K} \leq \epsilon\,,$$

which proves Equation (17).

Sufficient Condition for Convergence We use Equation (17) to find a sufficient condition for (15). This part differs from Fort and Moulines (2003) as our approximation is not defined as a random sum. Equation (17) is equivalent to

$$\mathbb{1}_{|g\circ\hat{\theta}(\tilde{S}_n(\theta_n))-g\circ\hat{\theta}(\bar{S}(\theta_n))|\mathbb{1}_{\theta_n\in K}>\epsilon} \leq \mathbb{1}_{\left\|\tilde{S}_n(\theta_n)-\bar{S}(\theta_n)\right\|\mathbb{1}_{\theta_n\in K}>\alpha}\,.$$

From that, we get

$$\forall\epsilon>0,\ \exists\alpha>0\ \sum_{n=1}^{\infty}\mathbb{1}_{|g\circ\hat{\theta}(\tilde{S}_n(\theta_n))-g\circ\hat{\theta}(\bar{S}(\theta_n))|\mathbb{1}_{\theta_n\in K}>\epsilon} \leq \sum_{n=1}^{\infty}\mathbb{1}_{\left\|\tilde{S}_n(\theta_n)-\bar{S}(\theta_n)\right\|\mathbb{1}_{\theta_n\in K}>\alpha}\,.$$

As a consequence, if

$$\forall\alpha>0,\ \sum_{n=1}^{\infty}\mathbb{1}_{\left\|\tilde{S}_n(\theta_n)-\bar{S}(\theta_n)\right\|\mathbb{1}_{\theta_n\in K}>\alpha}<\infty$$

Then

$$\forall\epsilon>0,\ \sum_{n=1}^{\infty}\mathbb{1}_{|g\circ\hat{\theta}(\tilde{S}_n(\theta_n))-g\circ\hat{\theta}(\bar{S}(\theta_n))|\mathbb{1}_{\theta_n\in K}>\epsilon}<\infty$$

In other, equivalent, words:

$$\begin{aligned}\text{If}\quad & \left\|\tilde{S}_n(\theta_n)-\bar{S}(\theta_n)\right\|\mathbb{1}_{\theta_n\in K}\underset{n\to\infty}{\longrightarrow}0\\ \text{Then}\quad & |g\circ\hat{\theta}(\tilde{S}_n(\theta_n))-g\circ\hat{\theta}(\bar{S}(\theta_n))|\mathbb{1}_{\theta_n\in K}\underset{n\to\infty}{\longrightarrow}0\,.\end{aligned} \tag{18}$$

Hence, having for all compact sets $K\subset\Theta$, $\left\|\tilde{S}_n(\theta_n)-\bar{S}(\theta_n)\right\|\mathbb{1}_{\theta_n\in K}\underset{n\to\infty}{\longrightarrow}0$ is sufficient to have the desired condition (15). Similarly, we find that $\forall\theta\in K_0$:

$$\begin{aligned}&\left\|\tilde{S}_n(\theta)-\bar{S}(\theta)\right\|\underset{n\to\infty}{\longrightarrow}0\\ \Longrightarrow\ &|g\circ\hat{\theta}(\tilde{S}_n(\theta))-g\circ\hat{\theta}(\bar{S}(\theta))|\underset{n\to\infty}{\longrightarrow}0\,,\end{aligned} \tag{19}$$

which provides us a sufficient condition for (14).

Further Simplifications of the Desired Result with Successive Sufficient Conditions We find another, simpler, sufficient condition for (15) from Equation (18). This part is unique to our proof and absent from Fort and Moulines (2003). It is here that we relate the formal conditions of Proposition 3 to the specific assumptions of our Theorem 1.

We first remove the dependency on the terms $\{\theta_n\}_n$ of the EM sequence:

$$\left\|\tilde{S}_n(\theta_n)-\bar{S}(\theta_n)\right\|\mathbb{1}_{\theta_n\in K}\leq\sup_{\theta\in K}\left\|\tilde{S}_n(\theta)-\bar{S}(\theta)\right\|. \tag{20}$$

From Equations (18)–(20) we get that:

$$\forall\text{ compact }K\subset\Theta,\quad\sup_{\theta\in K}\left\|\tilde{S}_n(\theta)-\bar{S}(\theta)\right\|\underset{n\to\infty}{\longrightarrow}0\,,$$

is a sufficient condition to have both Equations (14) and (15).

To show that the hypotheses of Theorem 1 imply this sufficient condition, we express it in integral form. Let $S=\{S_u\}_{u=1,\dots,q}$. We recall that $\tilde{S}_n(\theta)=\left\{\int_z S_u(z)\tilde{p}_{\theta,n}(z)dz\right\}_{u=1}^{q}$ and $\bar{S}(\theta)=\left\{\int_z S_u(z)p_\theta(z)dz\right\}_{u=1}^{q}$. Hence:

$$\tilde{S}_n(\theta)-\bar{S}(\theta)=\left\{\int_z S_u(z)\left(\tilde{p}_{\theta,n}(z)-p_\theta(z)\right)dz\right\}_{u=1}^{q}.$$

These q terms can be upper bounded by two different terms depending on the existence of the involved quantities:

$$\int_z S_u(z)\left(\tilde{p}_{\theta,n}(z)-p_\theta(z)\right)dz \leq \left(\int_z S_u(z)^2 dz\right)^{\frac{1}{2}}\left(\int_z \left(\tilde{p}_{\theta,n}(z)-p_\theta(z)\right)^2 dz\right)^{\frac{1}{2}},$$

and

$$\int_z S_u(z)\left(\tilde{p}_{\theta,n}(z)-p_\theta(z)\right)dz \leq \left(\int_z S_u(z)^2 p_\theta(z) dz\right)^{\frac{1}{2}}\left(\int_z \left(\frac{\tilde{p}_{\theta,n}(z)}{p_\theta(z)}-1\right)^2 p_\theta(z) dz\right)^{\frac{1}{2}}.$$

As a consequence, if $\int_z S_u(z)^2 dz$ exists, then it is sufficient to show have:

$$\sup_{\theta\in K}\int_z \left(\tilde{p}_{\theta,n}(z)-p_\theta(z)\right)^2 dz \underset{n\to\infty}{\longrightarrow} 0\,,$$

and if $\int_z S_u(z)^2 p_\theta(z) dz$ exists, then it is sufficient to show have:

$$\sup_{\theta\in K}\int_z \left(\frac{\tilde{p}_{\theta,n}(z)}{p_\theta(z)}-1\right)^2 p_\theta(z) dz \underset{n\to\infty}{\longrightarrow} 0\,.$$

Among the assumptions of Theorem 1 is one that states that for all compacts $K \subseteq \Theta$, one of those scenarios has to be true. Hence our sufficient condition is met.

Conclusions With the hypotheses of Theorem 1, we have

$$\forall \text{ compact } K\subseteq\Theta,\quad \sup_{\theta\in K}\left\|\tilde{S}_n(\theta)-\bar{S}(\theta)\right\| \underset{n\to\infty}{\longrightarrow} 0\,,$$

which is a sufficient condition to verify both Equation (14) and (15). With these two conditions, we can apply Proposition 3.

A.1.2 Applying Proposition 3

Since we verify all the conditions of Proposition 3, we can apply its conclusions:

$$\text{With probability 1, } \limsup_{n\to\infty} p_n < \infty \text{ and } \{\theta_n\}_n \text{ compact sequence},$$

which is specifically the result $(i)(a)$ of Theorem 1.

A.1.3 Verifying the Conditions of Proposition 2

With Proposition 2, we prove the remaining points of Theorem 1: $(i)(b)$ and (ii).

For the application of Proposition 2:

- $Cl\left(\{\theta_n\}_n\right)$ (set closure) plays the part of the compact K
- $\{\theta\in\Theta | T(\theta)=\theta\}$ plays the part of the set $\mathcal{L}$
- $\mathcal{L}\cap K$ is also compact thanks to hypothesis $M3$
- The likelihood g is the C^0 Lyapunov function with regards to $(T,\mathcal{L})$
- $\{\theta_n\}_n$ is the K valued sequence (since K is $Cl\left(\{\theta_n\}_n\right)$).

The last condition that remains to be shown to apply Proposition 2 is that:

$$\lim_{n\to\infty} |g(\theta_{n+1}) - g\circ T(\theta_n)| = 0\,.$$

We have more or less already proven that, in the previous section of the Proof, with $F_n(\theta_n)$ in place of θ_{n+1}. The only indices where $F_n(\theta_n) \neq \theta_{n+1}$ are when the value of the sequence p_n experiences an increment of 1. We have proven with Proposition 3 that there is only a finite number of such increments.

$$|g(\theta_{n+1}) - g\circ T(\theta_n)| = |g(\theta_0) - g\circ T(\theta_n)|\mathbb{1}_{p_{n+1}=p_n+1} \;+\; |g\circ F_n(\theta_n) - g\circ T(\theta_n)|\mathbb{1}_{p_{n+1}=p_n}\,.$$

Since there is only a finite number of increments of the value of p_n, then $\exists N\in\mathbb{N}$, $\forall n\geq N$, $\mathbb{1}_{p_{n+1}=p_n+1} = 0$ and $\mathbb{1}_{p_{n+1}=p_n} = 1$. In other words:

$$\begin{aligned}
&\exists N\in\mathbb{N},\,\forall n\geq N,\;\; |g(\theta_{n+1}) - g\circ T(\theta_n)| = |g\circ F_n(\theta_n) - g\circ T(\theta_n)|\\
&\exists N\in\mathbb{N},\,\forall n\geq N,\;\; |g(\theta_{n+1}) - g\circ T(\theta_n)| = |g\circ F_n(\theta_n) - g\circ T(\theta_n)|\mathbb{1}_{\theta_n\in Cl(\{\theta_k\}_k)}\,.
\end{aligned}$$

Since θ_n is always in $Cl(\{\theta_k\}_k)$ by definition. Additionally Proposition 3 tells us that $Cl(\{\theta_k\}_k)$ is a compact. Moreover, in order to use Proposition 3 in the first place, we had proven that:

$$\forall \text{ compact } K\subseteq\Theta,\quad \lim_{n\to\infty} |g\circ F_n(\theta_n) - g\circ T(\theta_n)|\mathbb{1}_{\theta_n\in K} = 0.$$

We can apply this directly with $K = Cl(\{\theta_k\}_k)$ to conclude the desired result:

$$\lim_{n\to\infty} |g(\theta_{n+1}) - g\circ T(\theta_n)| = 0$$

Hence we verify all the conditions to apply Proposition 2.

A.1.4 Applying Proposition 2

Since we verify all we need, we have the conclusions of Proposition 2:

- $\{g(\theta_n)\}_{n\in\mathbb{N}}$ converges towards a connected component of $g(\mathcal{L}\cap Cl(\{\theta_n\}_n))\subset g(\mathcal{L})$
- If $g(\mathcal{L}\cap Cl(\{\theta_n\}_n))$ has an empty interior, then $\{g(\theta_n)\}_{n\in\mathbb{N}}$ converges towards a $g^*\in\mathbb{R}$ and $\{\theta_n\}_n$ converges towards $\mathcal{L}_{g^*}\cap Cl(\{\theta_n\}_n)$. Where $\mathcal{L}_{g^*} := \{\theta\in\mathcal{L}|g(\theta) = g^*\}$

Both points are respectively the statements $(i)(b)$ and (ii) of Theorem 1.
Which concludes the proof of the Theorem.

A.2 Proof of the Tempering Theorem

In this Section, we prove Theorem 5 of the main paper, the convergence of the tempered EM algorithm. For that, we need to show that we verify each of the hypotheses of the more general Theorem 1.

We already assumed the conditions M1, M2 and M3 in the hypotheses of Theorem 5. To apply Theorem 1, we need to show that when $\tilde{p}_{\theta,n}(z) := \frac{p_\theta^{\frac{1}{T_n}}(z)}{\int_{z'} p_\theta^{\frac{1}{T_n}}(z')dz'}$, then $\forall$ compact $K\subseteq\Theta$, one of the two following configurations holds:

$$\int_z S(z)^2dz < \infty \text{ and } \sup_{\theta\in K}\int_z \left(\tilde{p}_{\theta,n}(z) - p_\theta(z)\right)^2 dz \underset{n\to\infty}{\longrightarrow} 0\,,$$

or

$$\sup_{\theta\in K}\int_z S(z)^2 p_\theta(z)dz < \infty \text{ and } \sup_{\theta\in K}\int_z \left(\frac{\tilde{p}_{\theta,n}(z)}{p_\theta(z)} - 1\right)^2 p_\theta(z)dz \underset{n\to\infty}{\longrightarrow} 0\,.$$

Since we have assumed:

$$\forall \text{ compact } K \in \Theta,\ \forall \alpha \in \overline{\mathcal{B}}(1,\epsilon), \forall u, \quad \sup_{\theta \in K} \int_z S_u^2(z) p_\theta^\alpha(z) dz < \infty\,,$$

then we already verify the first half of the second configuration for all the compacts K. Hence it is sufficient to prove that:

$$\forall \text{ compact } K \in \Theta,\ \sup_{\theta \in K} \int_z \left(\frac{\tilde{p}_{\theta,n}(z)}{p_\theta(z)} - 1 \right)^2 p_\theta(z) dz \underset{n \to \infty}{\longrightarrow} 0\,, \tag{21}$$

to have the desired result. The rest of the proof is dedicated to this goal.

A.2.1 Taylor Development

We use the Taylor's formula of the first order with the mean-value form of the reminder. For a derivable function f:

$$f(x) = f(0) + f'(a)x, \quad a \in [0, x]\,, \tag{22}$$

where the interval $[0, x]$ has a flexible meaning since x could be negative.

We apply it to:

$$f(x) = e^x, \quad f'(x) = e^x, \quad f(x) = 1 + xe^a,\ a \in [0, x]\ ,$$

and:

$$f(x) = \frac{1}{1+x}, \quad f'(x) = -\frac{1}{(1+x)^2}, \quad f(x) = 1 - \frac{x}{(1+a)^2},\ a \in [0, x]\ .$$

To make the upcoming calculation more readable, we momentarily replace $p_\theta(z)$ by simply p and T_n by T.

$$\begin{aligned} p^{\frac{1}{T}} &= p\left(p^{\frac{1}{T}-1}\right) \\ &= pe^{(\frac{1}{T}-1)\ln p} \\ &= p + \left(\frac{1}{T} - 1\right) p\, \ln p\, e^a, \quad a \in \left[0, \left(\frac{1}{T} - 1\right) \ln p\right]\ , \end{aligned}$$

where $a = a(z, \theta, T_n)$ since it depends on the value of $p_\theta(z)$ and T_n. Provided that the following quantities are defined, we have:

$$\int_z p^{\frac{1}{T}} = 1 + \left(\frac{1}{T} - 1\right) \int_z p\, \ln p\, e^a\,,$$

Hence:

$$\frac{1}{\int_z p^{\frac{1}{T}}} = 1 - \left(\frac{1}{T} - 1\right) \frac{\int_z p\, \ln p\, e^a}{(1+b)^2}, \quad b \in \left[0, \left(\frac{1}{T} - 1\right) \int_z p\, \ln p\, e^a\right]\ ,$$

where $b = b(\theta, T_n)$ since it depends on the value of T_n the integral over z of a function of z and θ.

In the end, we have:

$$\frac{p^{\frac{1}{T}}}{\int_z p^{\frac{1}{T}}} = p + \left(\frac{1}{T} - 1\right) p\, \ln p\, e^a \left(1 - \left(\frac{1}{T} - 1\right) \frac{\int_z p\, \ln p\, e^a}{(1+b)^2}\right) - \left(\frac{1}{T} - 1\right) p \frac{\int_z p\, \ln p\, e^a}{(1+b)^2}\,. \tag{23}$$

Since for any real numbers $(x+y)^2 \leq 2(x^2 + y^2)$, then:

$$\begin{aligned} \left(\frac{p^{\frac{1}{T}}}{\int_z p^{\frac{1}{T}}} - p\right)^2 &\leq 2\left(\frac{1}{T} - 1\right)^2 p^2 \left((\ln p\, e^a)^2 \left(1 - \left(\frac{1}{T} - 1\right) \frac{\int_z p\, \ln p\, e^a}{(1+b)^2}\right)^2 + \left(\frac{\int_z p\, \ln p\, e^a}{(1+b)^2}\right)^2 \right) \\ &= 2\left(\frac{1}{T} - 1\right)^2 p^2 \left((\ln p\, e^a)^2 A + B \right)\ . \end{aligned}$$

where $A = A(\theta, T_n)$ and $B = B(\theta, T_n)$. So far the only condition that has to be verified for all the involved quantities to be defined is that $\int_z p \ln p \, e^a$ exists. With this Taylor development on hand, we state, prove and apply two lemmas which allow us to get (21) and conclude the proof of the theorem.

A.2.2 Two Intermediary Lemmas

The two following lemmas provides every result we need to finish the proof.

Lemma 7. *With*

$$p_\theta(z) = exp\left(\psi(\theta) + \langle S(z), \phi(\theta)\rangle\right),$$

then

$$\int_z p_\theta^\alpha(z) \, \ln^2 p_\theta(z) dz \leq 2\psi(\theta)^2 \int_z p_\theta^\alpha(z) dz + 2\left\|\phi(\theta)\right\|^2 . \sum_u \int_z S_u^2(z) p_\theta^\alpha(z).$$

and

$$\int_z p_\theta^\alpha(z) \, \left|\ln p_\theta(z)\right| dz \leq \left|\psi(\theta)\right| \int_z p_\theta^\alpha(z) dz + \left\|\phi(\theta)\right\| . \left(\sum_u \int_z S_u^2(z) p_\theta^\alpha(z) \int_z p_\theta^\alpha(z)\right)^{\frac{1}{2}}.$$

Proof. For the first inequality, using the fact that $(a+b)^2 \leq 2(a^2+b^2)$, we have:

$$\int_z p_\theta^\alpha(z) \, \ln^2 p_\theta(z) dz \leq 2\psi(\theta)^2 \int_z p_\theta^\alpha(z) dz + 2\int_z p_\theta^\alpha(z) \left\langle S(z), \phi(\theta)\right\rangle^2,$$

We use Cauchy-Schwartz:

$$\left\langle S(z), \phi(\theta)\right\rangle^2 \leq \left\|\phi\right\|^2 \left\|S(z)\right\|^2 = \left\|\phi\right\|^2 \sum_u S_u(z)^2,$$

to get the desired result:

$$\int_z p_\theta^\alpha(z) \, \ln^2 p_\theta(z) dz \leq 2\psi(\theta)^2 \int_z p_\theta^\alpha(z) dz + 2\left\|\phi(\theta)\right\|^2 . \sum_u \int_z S_u^2(z) p_\theta^\alpha(z).$$

For the second inequality, we start with Cauchy-Schwartz on $\left\langle \int_z S(z) p_\theta^\alpha(z), \phi(\theta)\right\rangle$:

$$\int_z p_\theta^\alpha(z) \, \left|\ln p_\theta(z)\right| dz \leq \left|\psi(\theta)\right| \int_z p_\theta^\alpha(z) dz + \left\|\phi(\theta)\right\| . \left\|\int_z S(z) p_\theta^\alpha(z)\right\|.$$

Moreover, since:

$$\int_z S_u(z) p_\theta^\alpha(z) dz \leq \left(\int_z S_u^2(z) p_\theta^\alpha(z) dz\right)^{\frac{1}{2}} \left(\int_z p_\theta^\alpha(z) dz\right)^{\frac{1}{2}},$$

then

$$\int_z p_\theta^\alpha(z) \, \left|\ln p_\theta(z)\right| dz \leq \left|\psi(\theta)\right| \int_z p_\theta^\alpha(z) dz + \left\|\phi(\theta)\right\| . \left(\sum_u \int_z S_u^2(z) p_\theta^\alpha(z) \int_z p_\theta^\alpha(z)\right)^{\frac{1}{2}}.$$

□

Lemma 8. *With K compact and $\epsilon \in \mathbb{R}_+^*$,*

$$p_\theta(z) = exp\left(\psi(\theta) + \langle S(z), \phi(\theta)\rangle\right),$$

and

$$\tilde{p}_{\theta,n}(z) := \frac{p_\theta^{\frac{1}{T_n}}(z)}{\int_{z'} p_\theta^{\frac{1}{T_n}}(z') dz'},$$

if

(i) $T_n \in \mathbb{R}_+^* \underset{n\to\infty}{\longrightarrow} 1$,

(ii) $\sup_{\theta\in K} \psi(\theta) < \infty$,

(iii) $\sup_{\theta\in K} \|\phi(\theta)\| < \infty$,

(iv) $\forall \alpha \in \overline{\mathcal{B}}(1,\epsilon), \quad \sup_{\theta\in K} \int_z p_\theta^\alpha(z)dz < \infty$,

(v) $\forall \alpha \in \overline{\mathcal{B}}(1,\epsilon),\ \forall u, \quad \sup_{\theta\in K} \int_z S_u^2(z) p_\theta^\alpha(z)dz < \infty$.

then

$$\sup_{\theta\in K} \int_z \left(\frac{\tilde{p}_{\theta,n}(z)}{p_\theta(z)} - 1\right)^2 p_\theta(z)dz \underset{n\to\infty}{\longrightarrow} 0\,.$$

Proof. Provided that the following integrals exist, we have, thanks to the Taylor development:

$$\begin{aligned}\int_z \frac{1}{p}\left(\frac{p^{\frac{1}{T}}}{\int_z p^{\frac{1}{T}}} - p\right)^2 &\le 2\int_z \left(\frac{1}{T}-1\right)^2 p\left((\ln p\, e^a)^2 A + B\right)\\ &= 2\left(\frac{1}{T}-1\right)^2 A\int_z pe^{2a}\ln^2 p + 2\left(\frac{1}{T}-1\right)^2 B\,.\end{aligned} \tag{24}$$

In this proof, we find finite upper bounds independent of θ and T_n for $A(\theta,T_n)$, $B(\theta,T_n)$ and $\int_z pe^{2a}\ln^2 p$, then - since $\left(\frac{1}{T_n}-1\right) \longrightarrow 0$ - we have the desired result.

We start by studying $A(\theta,T_n) = \left(1 - \left(\frac{1}{T}-1\right)\frac{\int_z p\,\ln p\, e^a}{(1+b)^2}\right)^2$. The first term of interest here is $\int_z p\,\ln p\, e^a$. We have:

$$\begin{aligned} a &\in \left[0, \left(\frac{1}{T}-1\right)\ln p\right]\,,\\ e^a &\in \left[1, p^{\frac{1}{T}-1}\right]\,,\\ p\,\ln p\, e^a &\in \left[p\,\ln p, p^{\frac{1}{T}}\ln p\right]\,.\end{aligned}$$

where we recall that the interval is to be taken in a flexible sense, since we do not now a priory which bound is the largest and which is the smallest. What we have without doubt though is:

$$|p\,\ln p\, e^a| \le max\left(|p\,\ln p|, \left|p^{\frac{1}{T}}\ln p\right|\right)\,.$$

We find an upper bound on both those term. Let $\alpha \in \overline{\mathcal{B}}(1,\epsilon)$, the second result of Lemma 7 provides us:

$$\int_z p_\theta^\alpha(z)\ |\ln p_\theta(z)|\, dz \le |\psi(\theta)|\int_z p_\theta^\alpha(z)dz + \|\phi(\theta)\|\,.\left(\sum_u \int_z S_u^2(z)p_\theta^\alpha(z)\int_z p_\theta^\alpha(z)\right)^{\frac{1}{2}}\,.$$

Thanks to the hypotheses (ii), (iii), (iv) and (v), we have:

$$\begin{aligned}\int_z p_\theta^\alpha(z)\ |\ln p_\theta(z)|\, dz &\le \sup_{\theta\in K}\ |\psi(\theta)|\,.\sup_{\theta\in K}\ \int_z p_\theta^\alpha(z)dz\\ &\quad + \sup_{\theta\in K}\ \|\phi(\theta)\|\,.\sum_u\left(\sup_{\theta\in K}\ \int_z S_u^2(z)p_\theta^\alpha(z)\right)^{\frac{1}{2}}.\left(\sup_{\theta\in K}\ \int_z p_\theta^\alpha(z)\right)^{\frac{1}{2}}\\ &=: C(\alpha)\\ &< \infty\,.\end{aligned}$$

The upper bound $C(\alpha)$ in the previous inequality is independent of θ and z but still dependant of the exponent α. However, since $\overline{\mathcal{B}}(1,\epsilon)$ is closed ball, hypotheses (iv) and (v) can be rephrased as:

$$(iv) \sup_{\alpha\in\overline{\mathcal{B}}(1,\epsilon)} \sup_{\theta\in K} \int_z p_\theta^\alpha(z)dz < \infty\,,$$
$$(v)\ \forall u,\ \sup_{\alpha\in\overline{\mathcal{B}}(1,\epsilon)} \sup_{\theta\in K} \int_z S_u^2(z)p_\theta^\alpha(z)dz < \infty\,.$$

Hence we can actually take the supremum over α in the right term of the inequation as well. We have: $\int_z p_\theta^\alpha(z)\ |\ln p_\theta(z)|\,dz$

$$\begin{aligned}
&\leq \sup_{\theta\in K} |\psi(\theta)|\,.\sup_{\alpha\in\overline{\mathcal{B}}(1,\epsilon)} \sup_{\theta\in K} \int_z p_\theta^\alpha(z)dz \\
&\quad + \sup_{\theta\in K} \|\phi(\theta)\|\,.\sum_u \left(\sup_{\alpha\in\overline{\mathcal{B}}(1,\epsilon)} \sup_{\theta\in K} \int_z S_u^2(z)p_\theta^\alpha(z)\right)^{\frac{1}{2}}.\left(\sup_{\alpha\in\overline{\mathcal{B}}(1,\epsilon)} \sup_{\theta\in K} \int_z p_\theta^\alpha(z)\right)^{\frac{1}{2}} \\
&=: C' \\
&< \infty\,.
\end{aligned}$$

This new upper bound C' is independent of α.

Since $T_n \mapsto 1$, then $\exists N\in\mathbb{N}, \forall n\geq N, \frac{1}{T_n}\in\overline{\mathcal{B}}(1,\epsilon)$. Hence for $n\geq N$, we can apply the previous inequation to either $\alpha=1$ or $\alpha=\frac{1}{T_n}$, which provides us that $\int_z p_\theta(z)\ |\ln p_\theta(z)|$, $\int_z p_\theta^{\frac{1}{T_n}}(z)\ |\ln p_\theta(z)|$ and their supremum in θ are all finite, all of them upper bounded by C'.

In the end, when $n\geq N$, we have the control $\sup_{\theta\in K} \left|\int_z p\,\ln p\,e^a\right| < C'$.

The next term to control is $\frac{1}{(1+b)^2}$.

Since $b\in\left[0,\left(\frac{1}{T}-1\right)\int_z p\,\ln p\,e^a\right]$, then $|b|\leq\left(\frac{1}{T}-1\right)\sup_{\theta\in K}\int_z p\,\ln p\,e^a$. We already established that for all $n\geq N$, $\sup_{\theta\in K}\left|\int_z p\,\ln p\,e^a\right|\leq C'<\infty$, hence $\sup_{\theta\in K}|b(\theta,T_n)| \underset{T_n\longrightarrow 1}{\longrightarrow} 0$. In particular, $\exists N'\in\mathbb{N}, \forall n\geq N', \forall\theta\in K$ we have $|b(\theta,T_n)|\leq\frac{1}{2}$. In that case:

$$\begin{aligned}
(1+b)^2 &> (1-|b|)^2 \geq \frac{1}{4} \\
\frac{1}{(1+b)^2} &< \frac{1}{(1-|b|)^2} \leq 4\,.
\end{aligned}$$

In the end, when $n\geq max(N,N')$, for any $\theta\in K$:

$$\begin{aligned}
A(\theta,T_n) &\leq 2 + 2\left(\frac{1}{T_n}-1\right)^2\left(\frac{\int_z p\,\ln p\,e^a}{(1+b)^2}\right)^2 \\
&\leq 2 + 32\left(\frac{1}{T_n}-1\right)^2\left(\sup_{\theta\in K}\int_z p\,\ln p\,e^a\right)^2 \\
&\leq 2 + 32\left(\frac{1}{T_n}-1\right)^2 C'^2 \\
&\leq 2 + 32\epsilon^2 C'^2 \\
&=: C_1\,.
\end{aligned}$$

This upper bound does not depend en θ anymore and the part in T_n simply converges towards 0 when $T_n \longrightarrow 1$.

Treating the term $B(\theta, T_n) = \left(\frac{\int_z p \ln p\, e^a}{(1+b)^2}\right)^2 \leq 16 \left(\sup_{\theta \in K} \int_z p \ln p\, e^a\right)^2 \leq 16C'^2 =: C_2$ is immediate after having dealt with $A(\theta, T_n)$.

We now treat the term $\int_z p\, e^{2a} \ln^2 p$ in the exact same fashion as we did $A(\theta, T_n)$:

$$\begin{aligned} & p \ln p\, e^a \in \left[p \ln p, p^{\frac{1}{T}} \ln p\right] \\ \implies & p (\ln p\, e^a)^2 \in \left[p \ln^2 p, p^{\frac{2}{T}-1} \ln^2 p\right] \\ \implies & p (\ln p\, e^a)^2 \leq max(p \ln^2 p, p^{\frac{2}{T}-1} \ln^2 p)\,. \end{aligned}$$

We control those two terms as previously. First we apply Lemma 7 (its first result this time) with $\alpha \in \overline{\mathcal{B}}(1, \epsilon)$.

$$\int_z p_\theta^\alpha(z) \ln^2 p_\theta(z) dz \leq 2\psi(\theta)^2 \int_z p_\theta^\alpha(z) dz + 2 \left\|\phi(\theta)\right\|^2 . \sum_u \int_z S_u^2(z) p_\theta^\alpha(z)\,.$$

Thanks to the hypotheses (ii), (iii), (iv) and (v), we can once again take the supremum of the bound over $\theta \in K$, then over $\alpha \in \overline{\mathcal{B}}(1, \epsilon)$ and conserve finite quantities:

$$\begin{aligned} \int_z p_\theta^\alpha(z) \ln^2 p_\theta(z) dz \leq & 2 \sup_{\theta \in K} \psi(\theta)^2 . \sup_{\alpha \in \overline{\mathcal{B}}(1,\epsilon)} \sup_{\theta \in K} \int_z p_\theta^\alpha(z) dz \\ & + 2 \sup_{\theta \in K} \left\|\phi(\theta)\right\|^2 . \sum_u \sup_{\alpha \in \overline{\mathcal{B}}(1,\epsilon)} \sup_{\theta \in K} \int_z S_u^2(z) p_\theta^\alpha(z) \\ =: & C_3 \\ < & \infty\,. \end{aligned}$$

The previous result is true for $\alpha = 1$, and since once again $\exists N'', \forall n \geq N'', \frac{2}{T_n} - 1 \in \overline{\mathcal{B}}(1, \epsilon) \cap \mathbb{R}_+^*$, it is also true for $\alpha = \frac{2}{T_n} - 1$ when n is large enough. C_3 is independent of z, θ and T_n.

In the end $\forall n \geq N''$, $\int_z p\, e^{2a} \ln^2 p \leq C_3 < \infty$.

We replace the three terms $A(\theta, T_n)$, $B(\theta, T_n)$ and $\int_z p\, e^{2a} \ln^2 p$ by their upper bounds in the inequality (24). When $n \geq max(N, N', N'')$:

$$\int_z \frac{1}{p} \left(\frac{p^{\frac{1}{T}}}{\int_z p^{\frac{1}{T}}} - p\right)^2 \leq 2 \left(\frac{1}{T_n} - 1\right)^2 C_1 C_3 + 2 \left(\frac{1}{T_n} - 1\right)^2 C_2\,.$$

Which converges towards 0 when $T_n \longrightarrow 1$, i.e. when $n \longrightarrow \infty$. This concludes the proof of the lemma. □

A.2.3 Verifying the Conditions of Lemma 8

Now that the lemmas are proven, all that remains is to apply Lemma 8.

(i) We have $T_n \in \mathbb{R}_+^* \underset{n \to \infty}{\longrightarrow} 1$ by hypothesis.

(ii)* and *(iii) $\sup_{\theta \in K} \psi(\theta) < \infty$ and $\sup_{\theta \in K} \left\|\phi(\theta)\right\| < \infty$ are implied by the fact that $\psi(\theta) = \psi'(\theta) - \ln g(\theta)$ and $\phi(\theta)$ are continuous

(iv)* and *(v) are also hypotheses of the theorem.

Hence we can apply Lemma 8. This means that:

$$\sup_{\theta \in K} \int_z \left(\frac{\tilde{p}_{\theta,n}(z)}{p_\theta(z)} - 1 \right)^2 p_\theta(z) dz \underset{n \to \infty}{\longrightarrow} 0.$$

with this last condition verified, we can apply Theorem 1, which concludes the proof.

References

Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. (Methodological)* **1977**, *39*, 1–22.

Wu, C.J. On the convergence properties of the EM algorithm. *Ann. Stat.* **1983**, *11*, 95–103.

Boyles, R.A. On the convergence of the EM algorithm. *J. R. Stat. Soc. Ser. (Methodological)* **1983**, *45*, 47–50.

Lange, K. A gradient algorithm locally equivalent to the EM algorithm. *J. R. Stat. Soc. Ser. (Methodological)* **1995**, *57*, 425–437.

Delyon, B.; Lavielle, M.; Moulines, E. Convergence of a stochastic approximation version of the EM algorithm. *Ann. Stat.* **1999**, *27*, 94–128.

Wei, G.C.; Tanner, M.A. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Am. Stat. Assoc.* **1990**, *85*, 699–704.

Fort, G.; Moulines, E. Convergence of the Monte Carlo expectation maximization for curved exponential families. *Ann. Stat.* **2003**, *31*, 1220–1259.

Kuhn, E.; Lavielle, M. Maximum likelihood estimation in nonlinear mixed effects models. *Comput. Stat. Data Anal.* **2005**, *49*, 1020–1038.

Allassonnière, S.; Kuhn, E.; Trouvé, A. Construction of Bayesian deformable models via a stochastic approximation algorithm: a convergence study. *Bernoulli* **2010**, *16*, 641–678.

Allassonnière, S.; Chevallier, J. A New Class of EM Algorithms. Escaping Local Maxima and Handling Intractable Sampling. *Comput. Stat. Data Anal.* **2021**, *159*, 107159

Neal, R.M.; Hinton, G.E. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*; Springer: Dordrecht, South Holland, Netherlands; 1998; pp. 355–368.

Ng, S.K.; McLachlan, G.J. On the choice of the number of blocks with the incremental EM algorithm for the fitting of normal mixtures. *Stat. Comput.* **2003**, *13*, 45–55.

Cappé, O.; Moulines, E. On-line expectation–maximization algorithm for latent data models. *J. R. Stat. Soc. Ser. (Statistical Methodol.* **2009**, *71*, 593–613.

Chen, J.; Zhu, J.; Teh, Y.W.; Zhang, T. Stochastic expectation maximization with variance reduction. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 7967–7977.

Karimi, B.; Wai, H.T.; Moulines, E.; Lavielle, M. On the global convergence of (fast) incremental expectation maximization methods. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2019; pp. 2837–2847.

Fort, G.; Moulines, E.; Wai, H.T. A Stochastic Path Integral Differential EstimatoR Expectation Maximization Algorithm. *Adv. Neural Inf. Process. Syst.* **2020** *34*, 16972–16982

Kuhn, E.; Matias, C.; Rebafka, T. Properties of the stochastic approximation EM algorithm with mini-batch sampling. *Stat. Comput.* **2020**, *30*, 1725–1739.

Balakrishnan, S.; Wainwright, M.J.; Yu, B. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Ann. Stat.* **2017**, *45*, 77–120.

Dwivedi, R.; Ho, N.; Khamaru, K.; Wainwright, M.J.; Jordan, M.I.; Yu, B. Singularity, misspecification and the convergence rate of EM. *Ann. Stat.* **2020**, *48*, 3161–3182.

Booth, J.G.; Hobert, J.P. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J. R. Stat. Soc. Ser. (Statistical Methodol.)* **1999**, *61*, 265–285.

Levine, R.A.; Casella, G. Implementations of the Monte Carlo EM algorithm. *J. Comput. Graph. Stat.* **2001**, *10*, 422–439.

Levine, R.A.; Fan, J. An automated (Markov chain) Monte Carlo em algorithm. *J. Stat. Comput. Simul.* **2004**, *74*, 349–360.

Pan, J.X.; Thompson, R. Quasi-Monte Carlo EM algorithm for MLEs in generalized linear mixed models. In *COMPSTAT*; Physica-Verlag: Heidelberg, BW, Germany; 1998; pp. 419–424.

Jank, W. Quasi-Monte Carlo sampling to improve the efficiency of Monte Carlo EM. *Comput. Stat. Data Anal.* **2005**, *48*, 685–701.

Attias, H. Inferring Parameters and Structure of Latent Variable Models by Variational Bayes. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1999; pp. 21–30.

Bishop, C. *Pattern Recognition and Machine Learning*; Springer: New York, NY, USA, 2006.

Tzikas, D.G.; Likas, A.C.; Galatsanos, N.P. The variational approximation for Bayesian inference. *IEEE Signal Process. Mag.* **2008**, *25*, 131–146.

Kirkpatrick, S.; Gelatt, C.D.; Vecchi, M.P. Optimization by simulated annealing. *Science* **1983**, *220*, 671–680.

Swendsen, R.H.; Wang, J.S. Replica Monte Carlo simulation of spin-glasses. *Phys. Rev. Lett.* **1986**, *57*, 2607.

Geyer, C.J.; Thompson, E.A. Annealing Markov chain Monte Carlo with applications to ancestral inference. *J. Am. Stat. Assoc.* **1995**, *90*, 909–920.

Ueda, N.; Nakano, R. Deterministic annealing EM algorithm. *Neural Netw.* **1998**, *11*, 271–282.

Naim, I.; Gildea, D. Convergence of the EM algorithm for Gaussian mixtures with unbalanced mixing coefficients. In *Proceedings of the 29th International Conference on Machine Learning*; Omnipress: Madison, WI, USA; 2012; pp. 1655–1662.

Chen, H.F.; Guo, L.; Gao, A.J. Convergence and robustness of the Robbins-Monro algorithm truncated at randomly varying bounds. *Stoch. Process. Their Appl.* **1987**, *27*, 217–231.

Van Laarhoven, P.J.; Aarts, E.H. Simulated annealing. In *Simulated Annealing: Theory and Applications*; Springer: Dordrecht, South Holland, Netherlands; 1987; pp. 7–15.

Aarts, E.; Korst, J. *Simulated Annealing and Boltzmann Machines*; John Wiley and Sons Inc.: New York, NY, USA, 1988.

Hukushima, K.; Nemoto, K. Exchange Monte Carlo method and application to spin glass simulations. *J. Phys. Soc. Jpn.* **1996**, *65*, 1604–1608.

Titterington, D.; Smith, A.; Makov, U. *Statistical Analysis of Finite Mixture Distributions*; Wiley: New York, NY, USA, 1985.

Ho, N.; Nguyen, X. Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *Ann. Stat.* **2016**, *44*, 2726–2755.

Dwivedi, R.; Ho, N.; Khamaru, K.; Wainwright, M.; Jordan, M.; Yu, B. Sharp Analysis of Expectation-Maximization for Weakly Identifiable Models. In *International Conference on Artificial Intelligence and Statistics*; PMLR, 2020; pp. 1866–1876.

Winkelbauer, A. Moments and absolute moments of the normal distribution. *arXiv Prepr.* **2012**, arXiv:1209.4340.