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

Thomas Lartigue, Stanley Durrleman, Stéphanie Allassonnière

Statistics and Computing manuscript No.
(will be inserted by the editor)

Deterministic Approximate EM algorithm

Application to the Riemann approximation EM and the tempered EM

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

the date of receipt and acceptance should be inserted later

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 For instance, the Expectation (E) step has been replaced by Monte Carlo (MC), Markov Chain Monte Carlo or tempered approximations... 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 cases with intractable E steps, we introduce a deterministic alternative to the MC-EM, using Riemann sums. This method is easy to implement and does not require the tuning of hyper-parameters. Then, we consider the tempered approximation, borrowed from the Simulated Annealing optimisation technique and meant to improve the EM solution. We prove that the tempered EM verifies the convergence guarantees for a wide range of temperature profiles. We showcase empirically how it is able to escape adversarial initialisations. Finally, we combine the Riemann and tempered approximations to accomplish both their purposes.

Keywords Expectation Maximisation algorithm · exponential family · approximation · Riemann sum · tempering · annealing

Mathematics Subject Classification (2010) 62F99 · 49N30

Thomas Lartigue
thomas.lartigue@inria.fr
Present e-mail: thomas.lartigue@dzne.de

Stanley Durrleman
stanley.durrleman@inria.fr

Stéphanie Allassonnière
stephanie.allassonniere@parisdescartes.fr

[1] Aramis project-team, Inria Paris, France

[2] CMAP, CNRS, École polytechnique, I.P. Paris, France

[3] Centre de Recherche des Cordeliers, Université de Paris, INSERM, Sorbonne Université, France

1 Introduction

The Expectation Maximisation (EM) algorithm was introduced by Dempster, Laird and Rubin (DLR, 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 results 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, 2019). Random approximation of the E step have also been used in the case where the data is too voluminous to allow a full E step computation. See for instance 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) and the Stochastic Path-Integrated Differential EstimatoR EM (SPIDER-EM) (Fort 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), our without such strong-concavity assumption by Dwivedi et al. (2018).

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.

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 (2019) 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 estimate 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 propose a new, unified class of EM algorithms with deterministic approximations of the E step. We prove that members of this class benefit from the state of the art theoretical convergence guarantees of Wu (1983); Lange (1995); Delyon et al. (1999), under mild regularity conditions on the approximation. Then, we provide examples of approximations that fall under this framework and have practical applications. 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. Second, we prove that the deter-

ministic 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, 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 3, the "tempered EM" in Section 4, and their association, "tempered Riemann approximation EM", in Section 5.

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 Deterministic Approximate EM algorithm and its convergence for the curved exponential family

2.1 Context and motivation

In this section, we propose a new class of deterministic EM algorithms with approximated E step. This class of algorithms is general and includes both methods that estimate intractable E steps as well as methods that strive to improve the algorithm's solution. 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 (2019), we work with probability density functions belonging to the curved exponential family. The specific properties of which are given in the hypothesis $M1$ of Theorem 1.

The general framework of the EM is the following: a random variable x follows a parametric probability distribution with probability density function (pdf) depending on a parameter $\theta \in \Theta \subset \mathbb{R}^l$. We observe independent and identically distributed (iid) realisations of the distribution: $(x^{(1)}, ..., x^{(N)})$ and wish to maximise with respect to θ the resulting likelihood, which is noted $g(\theta)$. In the notations and the discourse, we mostly ignore x as a variable since the observations $(x^{(1)}, ..., x^{(N)})$ are supposed fixed throughout the reasoning. In particular, the sample size $N < +\infty$ is finite and fixed throughout this paper. We are not considering the asymptotic case where $N \longrightarrow +\infty$. We assume there exists a latent variable z informing the behaviour of the observed variable x such that $g(\theta)$ is the integral of the complete likelihood $h(z;\theta)$: $g(\theta) = \int_z h(z;\theta)\mu(dz)$, with μ the reference measure. The conditional density function of z is then $p_\theta(z) := h(z;\theta)/g(\theta)$. 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 $\theta \mapsto \mathbb{E}_z\left[ln\, h(z;\theta)\right]$ 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 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 $\theta \mapsto \mathbb{E}_{z\sim p_{\theta_n}(z)}\left[ln\, h(z;\theta)\right]$;

Which can be summarised as:

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

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):

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

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 idea behind this technique was originally introduced by . The steps of the resulting algorithm, called "Stable Approximate EM", can be written as:

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

This re-initialisation of the EM sequence may seem like a hurdle, however, this truncation is mostly 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.

2.2 Theorem

In the following, we will state the convergence Theorem of Eq. (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 Eq. (3). Let us assume two sets of hypotheses:*

- ***The $M1-3$ conditions of Fort and Moulines (2003).***
  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 $L : \mathcal{S}\times\Theta \to \mathbb{R}$ and $h : \mathcal{X}\times\Theta \to \mathbb{R}_+ \setminus \{0\}$:*

$$L(s;\theta) := \psi(\theta) + \langle s, \phi(\theta)\rangle\,,\ h(z;\theta) := exp(L(S(z);\theta))\,.$$

  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. *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:*

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

Or

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

Then,

(i) (a) $\lim\limits_{n\,\infty} j_n < \infty$ and $\sup\limits_{n\in\mathbb{N}} \|\theta_n\| < \infty$, with probability 1;
(b) $g(\theta_n)$ converges towards a connected component of $g(\mathcal{L})$.
(ii) If, additionally, $g\left(\mathcal{L} \cap Cl\left(\{\theta_n\}_{n\in\mathbb{N}}\right)\right)$ has an empty interior, then:

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

With $\mathcal{L} := \{\theta\in\Theta | \nabla g(\theta) = 0\}$ and $\mathcal{L}_{g^*} := \{\theta\in\mathcal{L} | g(\theta) = g^*\}$.

Remark 1 – $M2$ (a) is modified with regards to Fort and Moulines (2003), we remove the hypothesis that S has to be a continuous function of z that is not needed when the approximation is not stochastic. We call M2 (a*) this new sub-hypothesis.
– The condition $\int_z S_u^2(z)dz < \infty$ of the condition (4) can seem hard to verify since S is not integrated against a probability function. However, when z is a finite variable, as is the case for finite mixtures, this integral becomes a finite sum.
– 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 their have a measure of zero. In particular, this excludes Quasi-Monte Carlo approximations. However, one look at the proof of Theorem 1 (in supplementary materials) or at the following sketch of proof reveals that verifying the convergence $\sup_{\theta\in K} \left\|\tilde{S}_n(\theta) - \bar{S}(\theta)\right\| \underset{n\infty}{\longrightarrow} 0$ for any compact set K is actually a sufficient condition to benefit from the results of Theorem 1. This condition can be verified by finite approximations.

2.3 Sketch of proof

The detailed proof of this results can be found in supplementary materials, we propose here a 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. In our case, 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. First, let us state these two propositions:

Proposition 1 ("Proposition 9") *Let $\Theta \subseteq \mathbb{R}^l, K$ compact $\subset \Theta, \mathcal{L} \subseteq \Theta$ such that $\mathcal{L} \cap K$ compact. Let us assume*

- *WC^0 Lyapunov function with regards to $(T, \mathcal{L})$.*
- *$\exists u_n \in K^{\mathbb{N}}$ such that $|W(u_{n+1}) - W \circ T(u_n)| \underset{n\infty}{\longrightarrow} 0$*

Then

- *$\{W(u_n)\}_{n\in\mathbb{N}}$ converges towards a connected component of $W(\mathcal{L} \cap K)$*
- *If $W(\mathcal{L}\cap K)$ has an empty interior, then $\{W(u_n)\}_n$ converges towards w^* and $\{u_n\}_n$ converges towards the set $\mathcal{L}_{w^*} \cap K$*

Where $\mathcal{L}_{w^*} := \{\theta \in \mathcal{L} | W(\theta) = w^*\}$.

Proposition 2 ("Proposition 11") *Let $\Theta \subseteq \mathbb{R}^l, T$ and $\{F_n\}_n$ point to point maps on Θ. Let $\{\theta_n\}_n$ be the sequence defined by the Stable Approximate EM with likelihood g and approximate maps sequence $\{F_n\}_n$. Let $\mathcal{L} \subset \Theta$. We assume*

- *the $A1-2$ conditions of Proposition 10 of Fort and Moulines (2003).*
  - ***(A1)*** *There exists W, a C^0 Lyapunov function with regards to $(T, \mathcal{L})$ such that for all $M > 0$, $\{\theta \in \Theta, W(\theta) > M\}$ is compact, and:*

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

  - ***(A2)*** *$W(\mathcal{L})$ is compact OR **(A2')** $W(\mathcal{L}\cap K)$ is finite for all compact $K \subseteq \Theta$.*
- *$\forall u \in K_0, \quad \lim_{n\infty} |W \circ F_n - W \circ T|(u) = 0$*
- *$\forall$ compact $K \subseteq \Theta$:*

$$\lim_{n\infty} |W \circ F_n(u_n) - W \circ T(u_n)| \mathbb{1}_{u_n\in K} = 0$$

Then:
With probability 1, $\limsup_{n\infty} j_n < \infty$ and $\{u_n\}_n$ is a compact sequence.

For the proofs of these two results, see Fort and Moulines (2003). The proof of Theorem 1 is structured as follows: verifying the conditions of Proposition 2, applying Proposition 2, verifying the conditions of Proposition 1 and finally applying Proposition 1.

Verifying the conditions of Proposition 2. We first make explicit which object of our model plays which part in the Proposition. Let g be the likelihood function of a model of the curved exponential family.

- The set of its critical points is called $\mathcal{L}$:

$$\mathcal{L} := \{\theta \in \Theta | \nabla g(\theta) = 0\} .$$

- We call T the point to point map describing the transition between θ_n and θ_{n+1} in the exact EM algorithm, that is to say $T := \hat{\theta} \circ \bar{S}$.
- The general properties of the EM tell us that its stationary points are the critical points of g: $\mathcal{L} = \{\theta \in \Theta | T(\theta) = \theta\}$. Additionally, we have that g is a C^0 Lyapunov function associated to $(T, \mathcal{L})$, hence it is fit to play the part of W from Proposition 2.
- Let $\{\theta_n\}_n$ be the sequence defined by the Stable Approximate EM, and $\{F_n\}_n$ the corresponding sequence of point to point maps.

With this setup, the assumptions $M1-3$ of Theorem 1 directly imply that A1 and A2 or A2' are verified.

We need to prove that the last two conditions for Proposition 2 are verified:

$$\forall \theta \in K_0, \quad \lim_{n\infty} |g \circ F_n - g \circ T|(\theta) = 0 , \tag{6}$$

$$\forall \text{compact } K \subseteq \Theta, \ \lim_{n\infty} |g \circ F_n(\theta_n) - g \circ T(\theta_n)| \mathbb{1}_{\theta_n\in K} = 0. \tag{7}$$

At this stage, the steps to prove these two conditions differ from Fort and Moulines (2003). We denote $\tilde{S}_n(\theta_n)$ the approximated E step in the Stable Approximate

EM (so that $F_n = \hat{\theta} \circ \tilde{S}_n$). By using uniform continuity properties on compacts, we first obtain that

$$\forall \text{ compact } K,\ \sup_{\theta \in K} \left\| \tilde{S}_n(\theta) - \bar{S}(\theta) \right\| \underset{n \infty}{\longrightarrow} 0\,, \tag{8}$$

is a sufficient condition to obtain both (6) and (7), and conclude that we can apply Proposition 2. Writing $\tilde{S}_n$ and $\bar{S}$ as integrals in z makes it clear that the two hypothesis (4) and (5) of Theorem 1 are both sufficient to have (8). Which concludes this section of the Proof.

Applying Proposition 2. Since we verify all the condition of Proposition 2, we can apply its conclusion. With probability 1:
$\underset{n \infty}{limsup}\, j_n < \infty$ and $\{\theta_n\}_n$ is a compact sequence.
Which is specifically the result $(i)(a)$ of Theorem 1.

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

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

The last condition that remains to be shown to apply Proposition 1 is that:

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

In our previous efforts to prove (6) and (7), we have more or less already proven that 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 j_n experiences an increment of 1.

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

We have proven with Proposition 2 that there is only a finite number of such increments and that $Cl(\{\theta_k\}_k)$ is a compact. Since θ_n is always in $Cl(\{\theta_k\}_k)$ by definition, we can apply to $K := Cl(\{\theta_k\}_k)$ the result:

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

that we proved in order to verify Proposition 2, and get the needed condition:

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

Applying Proposition 1 Since we verify all we need to apply the conclusions of Proposition 1:

- $\{g(\theta_n)\}_{n \in \mathbb{N}}$ 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 $\mathcal{L}_{g^*} \cap Cl(\{\theta_n\}_n)$. Where $\mathcal{L}_{g^*} := \{\theta \in \mathcal{L} | g(\theta) = g^*\}$

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

3 Riemann approximation EM

3.1 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.
When the conditional probability $p_\theta(z)$ is a continuous function, and even if $h(z;\theta)$ 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 (2019). When the dimension is reasonably small, a deterministic approximation through Riemann sums can also be performed. Unlike the stochastic methods, which often require to define and tune a Markov Chain, the Riemann approximation involves almost no parameter. The user only needs to choose the position of the Riemann intervals, a choice which is very guided by the well known theories of integration (Lagrange, Legendre...).
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 results, 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.

3.2 Theorem and proof

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

Theorem 2 *Under conditions $M1-3$ of Theorem 1, and when z is bounded, the (Stable) Approximate EM with $\tilde{p}_{n,\theta}(z) := \frac{h(\lfloor nz \rfloor/n;\theta)}{\int_{z'} h(\lfloor nz' \rfloor/n;\theta)dz'}$, which we call "Riemann approximation EM", verifies the remaining conditions of applicability of Theorem 1 as long as $z \mapsto S(z)$ is continuous.*

Proof This is the detailed proof of Theorem 2.
The conditions $M1-3$ on the model are already assumed to be verified. In order to apply Theorem 1, we need to verify either Eq. (4) or (5). Here, with $z \mapsto S(z)$ continuous, we prove Eq. (4): For any compact $K \subseteq \Theta$,

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

Since z is bounded (and assumed to be in $[0,1]$ for simplicity) and S is continuous, the first part of the condition is easily verified: $\int_{z=0}^1 S_u^2(z)dz < \infty$. Only the second part remains to be proven.
First we 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:

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

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

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

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

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

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

In other words, $\{\tilde{h}_n\}_n$ converges uniformly towards h. Let ϵ be given, we assume that $n \ge N$, then $\forall(z,\theta) \in [0,1]\times K$:

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

Hence, $\forall n \ge N$:

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

Then, $\sup\limits_{\theta\in K} \int_{z=0}^1 \left(\tilde{p}_{\theta,n}(z) - p_\theta(z)\right)^2 dz \underset{n\infty}{\longrightarrow} 0$, by definition. This was the last hypothesis needed to apply Theorem 1. Which concludes the proof.

3.3 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:

$$\begin{aligned}
\frac{1}{\hat{\alpha}} &= -\frac{1}{N}\sum_{i=1}^{N} \Psi^{(i)} \circ ln(z)\,,\\
\hat{\lambda} &= \frac{\sum_{i=1}^{N} \Psi^{(i)} \circ (x^{(i)}z)}{\sum_{i=1}^{N} \Psi^{(i)} \circ z^2}\,,\\
\hat{\sigma}^2 &= \frac{1}{N}\sum_{i=1}^{N} \Psi^{(i)} \circ \left(x^{(i)} - \hat{\lambda}z\right)^2\,.
\end{aligned} \tag{9}$$

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 Fig. 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 Fig. 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, Fig. 2 suggests that the performances will not depend too much on the profile.

3.4 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.
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:

$$\begin{aligned}
x_1 &= \lambda_1 z_1 + z_2 + \sigma_1\epsilon_1\\
x_2 &= z_1 + \lambda_2 z_2 + \sigma_2\epsilon_2\,,
\end{aligned}$$

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:

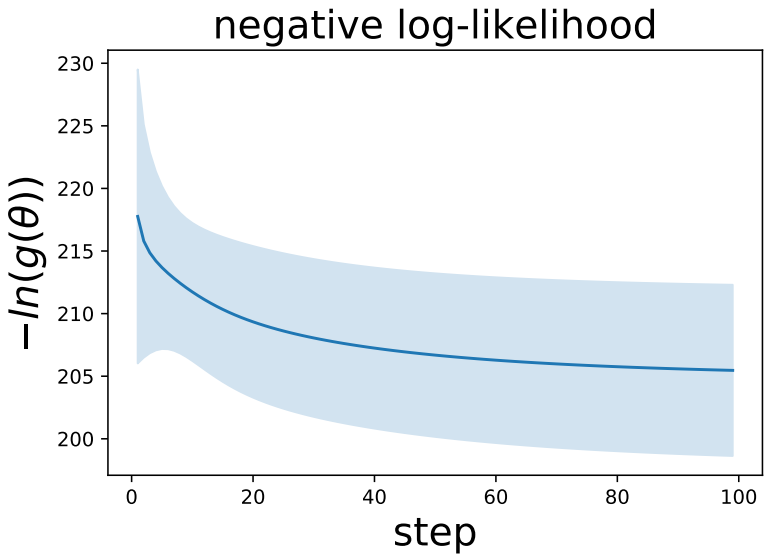
$$\tilde{p}^{(i)}_{\theta,n}\left(\frac{k_1}{n}, \frac{k_2}{n}\right) = \frac{h^{(i)}(\frac{k_1}{n}, \frac{k_2}{n};\theta)}{\frac{1}{n^2}\sum_{l_1,l_2=0}^{n-1} h^{(i)}(\frac{l_1}{n}, \frac{l_2}{n};\theta)}\,.$$

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:

$$\begin{aligned}
\frac{1}{\hat{\alpha}_1} &= -\frac{1}{N}\sum_{i=1}^{N} \Psi^{(i)} \circ ln(z_1)\,,\\
\hat{\lambda}_1 &= \frac{\sum_i \Psi^{(i)} \circ (x_1^{(i)}z_1 - z_2z_1)}{\sum_i \Psi^{(i)} \circ z_1^2}\,,\\
\hat{\sigma}_1 &= \frac{1}{N}\sum_{i=1}^{N} \Psi^{(i)} \circ \left(x_1^{(i)} - \hat{\lambda}_1 z_1 - z_2\right)^2\,,
\end{aligned}$$



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)

Fig. 1 (Right). 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. (Left). Average and standard deviation of the relative parameter reconstruction errors at the end of the Riemann EM.

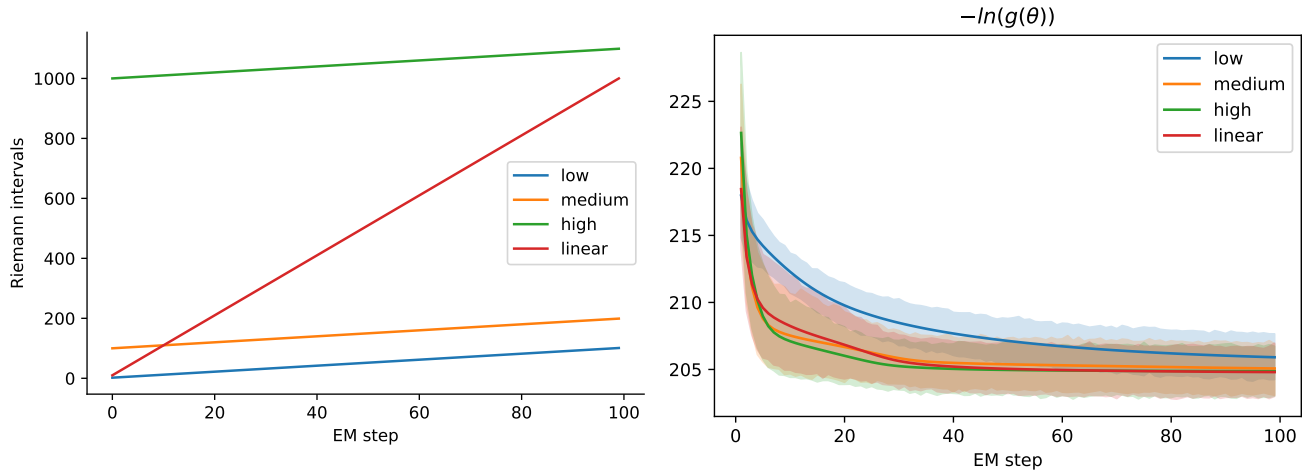


Fig. 2 (Right). Visual representation of the number of Riemann intervals over the EM steps for each profile φ_i. The total number of Riemann intervals computed over 100 EM iterations are: 5 150 for "low", 14 950 for "medium", 50 500 for "linear" and 104 950 for "high". (Left). For each profile, average evolution of the negative log-likelihood, with standard deviation, over 50 simulations. The results are fairly similar, in particular between "medium", "high" and "linear".

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 Fig. 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 Fig. 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.

4 Tempered EM

4.1 Context and motivation

In this section, we consider another particular case of Deterministic Approximate EM: the Tempered EM (or "tmp-EM"). We first motivate this algorithm. Then, we prove that under mild conditions, it verifies the hypothesis of Theorem 1, hence has the state of the art EM convergence guarantees. In particular, we prove that the choice of the temperature profile is almost completely free.

When optimising a non-convex function, following the gradients leads to one of the local extrema closest to the initialisation. If the method was allowed to explore more the profile of the function to be optimised, it would encounter points with better values and areas with stronger gradients missed because of its early commitment to one of the nearest potential wells.

A very well known way to encourage such an exploratory behaviour is the tempering, also called annealing. In its simplest form, the function to optimised g is elevated to the power $\frac{1}{T_n}$: $g \mapsto g^{\frac{1}{T_n}}$. Where T_n is a temperature tending towards 1 as the number n of steps of the procedure increases. This manipulation equalises the value of the function in the different points of the space, renders the gradients less strong, and makes the potential wells less attractive the higher the temperature T_n is. As a result, the optimisation procedure is not incited to limit itself to its starting region. Additionally, the general shape of the function g, in particular the hierarchy of values, is still preserved, meaning that the early course of the algorithm is still made on a function replicating the reality. As T_n gets closer to 1, the optimised function becomes identical to g and the potential wells become attractive again. By this point, the assumption is that the algorithm will be in a better place than it was at the initialisation.

These concepts are put in application in many state of the art procedures. The most iconic maybe 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.

The idea of applying a tempering to a classical EM was introduced in the Deterministic Annealed EM of Ueda and Nakano (1998) with a specific decreasing temperature scheme. Another specific, non-monotonous, temperature scheme was later proposed by Naim and Gildea (2012). In both cases, theoretical convergence guarantees are lacking. In Allassonnière and Chevallier (2019), tempering is applied to the SAEM, and convergence guarantees for the resulting algorithm are provided with any temperature scheme $T_n \longrightarrow 1$.

Here, we introduce the tmp-EM as a specific case of the Approximate EM of Section 2. We use the approximated distribution: $\tilde{p}_{n,\theta}(z) := p_\theta^{\frac{1}{T_n}}(z) / \int_{z'} p_\theta^{\frac{1}{T_n}}(z')dz' = h(z;\theta)^{\frac{1}{T_n}} / \int_{z'} h(z';\theta)^{\frac{1}{T_n}} dz'$ (renormalised to sum to 1). Unlike Ueda and Nakano (1998) and Naim and Gildea (2012), we do not specify any temperature scheme T_n, and prove in the following Theorem 3 that, under very mild conditions on the model, any sequence $\{T_n\}_n \in (\mathbb{R}^*)^{\mathbb{N}}$, $T_n \underset{n\infty}{\longrightarrow} 1$ guarantees the state of the art convergence.

Remark 2 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 $\mathbb{E}_{z\sim\tilde{p}_{n,\theta}(z)} \longrightarrow \mathbb{E}_{z\sim p_\theta(z)}$.

4.2 Theorem

We now give 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.

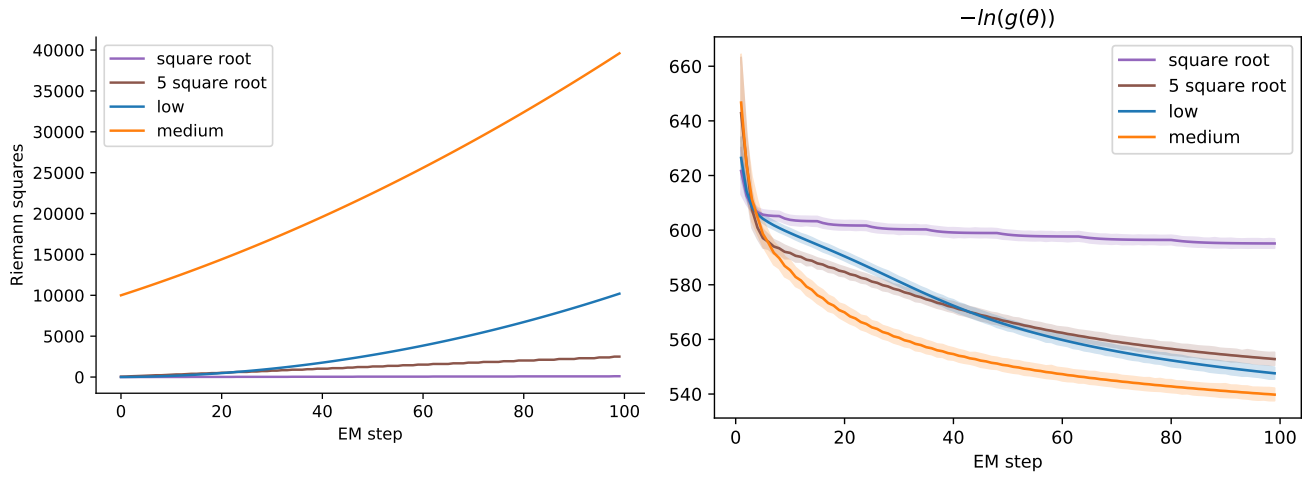


Fig. 3 (Right). Visual representation of the number of Riemann intervals over the EM steps for each profile φ_i. In higher dimension, the computational complexity of the profiles are very different. More precisely, the total number of Riemann squares computed over 100 EM iterations are: 4 534 for "square root", 125 662 for "5 square root", 348 550 for "low" and 2 318 350 for "medium". (Left). 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.

Theorem 3 *Let T_n be a sequence of non-zero real numbers. Under conditions $M1$-3 of Theorem 1, the (Stable) Approximate EM with $\tilde{p}_{n,\theta}(z) := \frac{p_\theta^{\frac{1}{T_n}}(z)}{\int_{z'} p_\theta^{\frac{1}{T_n}}(z')dz'}$, which we call "Tempered EM", verifies all other remaining conditions of applicability of Theorem 1 as long as $T_n \underset{n\,\infty}{\longrightarrow} 1$ and for any compact $K \in \Theta$, $\exists \epsilon \in]0,1[$, $\forall \alpha \in \overline{\mathcal{B}}(1,\epsilon)$:*

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

Where $\overline{\mathcal{B}}(1,\epsilon)$ is the closed ball centered in 1 and with radius ϵ in $\mathbb{R}$, and the index u of $S_u(z)$ indicates each of the real component of the $S(z) \in \mathcal{S} \subset \mathbb{R}^q$. The conditions on the integrability of $p_\theta^\alpha(z)$ and $S_u^2(z)p_\theta^\alpha(z)$ brought by the tempering are very mild. Indeed, in Section 4.4, we will show classical examples that easily verify the much stronger conditions: for any compact $K \in \Theta, \forall \alpha \in \mathbb{R}_+^*$,

$$\begin{array}{l} \underset{\theta\in K}{sup} \int_z p_\theta^\alpha(z)dz < \infty \quad , \\ \forall u \in [\![1,q]\!], \quad \underset{\theta\in K}{sup} \int_z S_u^2(z) p_\theta^\alpha(z)dz < \infty \quad . \end{array}$$

4.3 Sketch of proof

The detailed proof of Theorem 3 can be found in supplementary materials, we propose here a abbreviated version.

In order to apply Theorem 1, we need to verify five conditions. The three inevitable are $M1$, $M2$ and $M3$. The last two can either be that, $\forall$ compact $K \in \Theta$:

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

Or

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

The hypothesis of Theorem 3 already include $M1$, $M2$, $M3$ and:

$$\forall \text{ compact } K \in \Theta,\ \forall u, \quad \underset{\theta\in K}{sup} \int_z S_u^2(z) p_\theta(z)dz < \infty \,.$$

As a result, to apply Theorem 1, it is sufficient to verify that, with the tempering approximation, we have:

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

The proof of Theorem 3 revolves around proving this result.
With a Taylor development in $\left(\frac{1}{T_n} - 1\right)$, which converges toward 0 when $n \to \infty$, we control the difference

$(\tilde{p}_{\theta,n}(z) - p_\theta(z))^2$:

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

The terms $A(\theta,T_n)$, $B(\theta,T_n)$ and $a(z,\theta,T_n)$ come from the Taylor development. With the previous inequality, we control the integral of interest:

$$\begin{aligned}\int_z \frac{\left(\frac{p_\theta(z)^{\frac{1}{T_n}}}{\int_{z'} p_\theta(z')^{\frac{1}{T_n}}} - p_\theta(z)\right)^2}{p_\theta(z)} dz \\ \leq 2\left(\frac{1}{T_n} - 1\right)^2 A(\theta,T_n) \int_z p_\theta(z) e^{2a(z,\theta,T_n)} ln^2\, p_\theta(z) dz \\ + 2\left(\frac{1}{T_n} - 1\right)^2 B(\theta,T_n)\,.\end{aligned} \tag{10}$$

$A(\theta,T_n)$ and $B(\theta,T_n)$ both have upper bounds involving $\int_z p_\theta(z)^{\frac{1}{T_n}} ln\, p_\theta(z)$. In a similar fashion, the term $\int_z p_\theta(z) e^{2a(z,\theta,T_n)} ln^2\, p_\theta(z)$ is bounded by terms involving $\int_z p_\theta(z)^{\frac{2}{T_n}-1} ln^2\, p_\theta(z) dz$.

Thanks to the hypothesis of the Theorem, we prove that for any $\alpha \in \overline{\mathcal{B}}(1,\epsilon)$ and $\theta \in K$ the two 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 $T_n \underset{n\mapsto\infty}{\longrightarrow} 1$, then when n is large enough, $\frac{1}{T_n} \in \overline{\mathcal{B}}(1,\epsilon)$ and $\frac{2}{T_n} - 1 \in \overline{\mathcal{B}}(1,\epsilon)$ meaning that the previous result applies to the three terms $A(\theta,T_n)$, $B(\theta,T_n)$ and $\int_z p_\theta(z) e^{2a(z,\theta,T_n)} ln^2\, p_\theta(z) dz$: they are upper bounded by constants C_1, C_2 and C_3 respectively, all independent of θ and T_n.

The inequality (10) 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:

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

4.4 Examples of models that verify the conditions

In this section we illustrate that the conditions of Theorem 3 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.
In order to apply Theorem 3, we need to verify the conditions

- $M1$, $M2$ and $M3$
- for any compact $K \in \Theta, \exists \epsilon \in]0,1[, \forall \alpha \in \overline{\mathcal{B}}(1,\epsilon)$,
  $\underset{\theta\in K}{sup} \int_z p_\theta^\alpha(z) dz < \infty$,
  $\forall u, \quad \underset{\theta\in K}{sup} \int_z S_u^2(z) p_\theta^\alpha(z) dz < \infty$.

As previously stated, in both examples, we will actually verify the much stronger conditions: for any compact $K \in \Theta, \forall \alpha \in \mathbb{R}_+^*$:

$$\begin{cases} \underset{\theta\in K}{sup} \int_z p_\theta^\alpha(z) dz < \infty\,, \\ \forall u, \ \underset{\theta\in K}{sup} \int_z S_u^2(z) p_\theta^\alpha(z) dz < \infty\,. \end{cases}$$

4.4.1 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. (2018, 2020), 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 hypothesis of the general Theorem 1 as well as the tempering hypothesis of Theorem 3. 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 3 applies, and the sequence is guaranteed to converge towards a critical point of the likelihood function.
In the following, we give more details on the GMM likelihoods and the theorem hypotheses they verify. First,

note that the GMM belongs to the curved exponential family with the complete likelihood:

$$h(z;\theta)=\prod_{i=1}^{N}\sum_{k=1}^{K} exp\Bigg(\frac{1}{2}\Big(-(x^{(i)}-\mu_k)^T\Theta_k(x^{(i)}-\mu_k) + ln\left(|\Theta_k|\right)+2ln\left(\pi_k\right) - pln\left(2\pi\right)\Big)\Bigg)\mathbb{1}_{z^{(i)}=k}\,, \tag{11}$$

and the observed likelihood:

$$g(\theta)=\prod_{i=1}^{N}\sum_{k=1}^{K} exp\Bigg(\frac{1}{2}\Big(-(x^{(i)}-\mu_k)^T\Theta_k(x^{(i)}-\mu_k) + ln\left(|\Theta_k|\right)+2ln\left(\pi_k\right) - pln\left(2\pi\right)\Big)\Bigg)\,. \tag{12}$$

This is an exponential model with parameter:

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

The verification of conditions $M1-3$ for the GMM (exept $M2$ (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 hypothesis 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.

4.4.2 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 $M1-3$.
The complete likelihood of the model, not accounting for irrelevant constants, is:

$$h(z;\theta) = e^{\theta\sum_k Y_k}.exp\left(-e^\theta\sum_k e^{z_k}\right)\,. \tag{13}$$

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

$$\begin{aligned} g(\theta) &= \int_{z\in\mathbb{R}^d} h(z;\theta)dz \\ &= e^{\theta\sum_k Y_k}\int_{z\in\mathbb{R}^d} exp\left(-e^\theta\sum_k e^{z_k}\right)dz \\ &= e^{\theta\sum_k Y_k}\prod_{k=1}^d\int_{z_k\in\mathbb{R}} exp\left(-e^\theta e^{z_k}\right)dz_k \\ &= e^{\theta\sum_k Y_k}\left(\int_{u\in\mathbb{R}_+}\frac{exp\left(-u\right)}{u}du\right)^d \\ &= e^{\theta\sum_k Y_k}E_1(0)^d\,, \end{aligned} \tag{14}$$

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 Eq. (14):

$$\begin{aligned}\int_{z\in\mathbb{R}^d} h(z;\theta)^\alpha &= e^{\alpha\theta\sum_k Y_k}\prod_{k=1}^d\int_{z_k\in\mathbb{R}} exp\left(-\alpha e^\theta e^{z_k}\right)dz_k \\ &= e^{\alpha\theta\sum_k Y_k}E_1(0)^d\,.\end{aligned}$$

Hence:

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

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, $S(z) = \sum_k e^{z_k} \in \mathbb{R}$. We have:

$$S^2(z)p_\theta^\alpha(z) = \left(\sum_k e^{z_k}\right)^2 \frac{exp\left(-\alpha e^\theta \sum_k e^{z_k}\right)}{E_1(0)^{\alpha d}} . \quad (15)$$

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 $\prod_k e^{z_k} = \prod_k u_k$.

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

Which is proportional to:

$$\int_{u_1=0}^{+\infty} u_1 \int_{u_2=u_1}^{+\infty} u_2 ... \int_{u_d=u_{d-1}}^{+\infty} u_d^3\, e^{-\alpha e^\theta u_d}\, du_d ... du_2 du_1 .$$

Where we removed the finite constant $\frac{1}{E_1(0)^{\alpha d}}$ 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 Eq. (15), we have that

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

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 d 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 $d+2$. We chose a faster path.

4.5 Experiments with Mixtures of Gaussian

4.5.1 Context and experimental protocol

In this section, we will assess the capacity of tmp-EM to escape from deceptive local maxima, on a very well know toy example: likelihood maximisation within the Gaussian Mixture Model. We confront the algorithm 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.
For the sake of comparison, the experimental design is similar to the one in Allassonnière and Chevallier (2019) 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. Fig. 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 two 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 Fig. 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 Fig. 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 Fig. 4, 1000 datasets with 500 observations each are simulated, and the two EMs are ran with both the *barycenter* and the *2v1* initialisation. Regarding the temperature profile of tmp-EM, the only constraint is that $T_n \longrightarrow 1$. We use an oscillating profile inspired from Allassonnière and Chevallier (2019): $T_n = th(\frac{n}{2r}) + (T_0 - b\frac{2\sqrt{2}}{3\pi})\, a^{n/r} + b\, sinc(\frac{3\pi}{4} + \frac{n}{r})$. 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 "lockin-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. Although, we observe that in the case of *2v1*, the oscillations are not critical, and a simple decreasing exponential profile: $T_n = 1 + (T_0 - 1)\, exp(-r.n)$, with $T_0 = 100$ and $r = 1.5$, works as well. 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.

4.5.2 Quantitative analysis

In this section, we quantify the performances of EM and tmp-EM over all the simulations.

Fig. 5 and 6 depict the results of 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 is 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.

On these examples, we note that tmp-EM is more correct than 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 EM and tmp-EM is highlighted in bold. The recovery of the true class averages μ_k is spotlighted as it is the essential success metric for this experiment.

First we focus on the effect of the different initialisations and placement of (μ_1, μ_2) on the performance of the classical EM. In the first parameter family of Table 1, μ_1 and μ_2 are still far from one another. The relative error on these two positions is around 0.50 when the initialisation is a the neutral position at the barycenter of the dataset, and 1.50 when the initialisation is made by placing two centroids in the right cluster (*2v1*), which is a much more adversarial initialisation. In the second parameter family, μ_1 and μ_2 are getting closer. The relative error with the *barycenter* initialisation has doubled to reach 1.00, and, with the adversarial *2v1*, it has increased to 1.70. Finally, in the third parameter family, where μ_1 and μ_2 are so close that their distributions are hard to distinguish with the naked eye, the relative error with the *barycenter* initialisation has gained another 0.50 points to reach over 1.50, which was the initial error level with the *2v1* initialisation when μ_1 and μ_2 were well separated (parameter family 1). In this very ambiguous setting however, the relative error with *2v1* initialisation has gone up to around 1.80-1.90. As expected, we see that the performances are always hindered in average by the 2v1 initialisation, and that they also worsen when the relative positions of μ_1 and μ_2 become more ambiguous, regardless of the initialisation. The *barycenter* initialisation however is the one that suffers the most from the increasing ambiguity, gaining 0.5 points of relative error at every transition, whereas *2v1* gain "only" around 0.2 points.

We compare these results and their progression with the ones of tmp-EM in Table 1. In the first parameter family - the least ambiguous situation - the relative errors on μ_1 and μ_2 are around 0.05 with the *barycenter* initialisation and 0.30 with *2v1*. In other words, with the tempered E step, we divide by 10 and 5 respectively the relative errors with the *barycenter* and *2v1* initialisation. In the next position of μ_1 and μ_2, in the second parameter family, the relative error with the

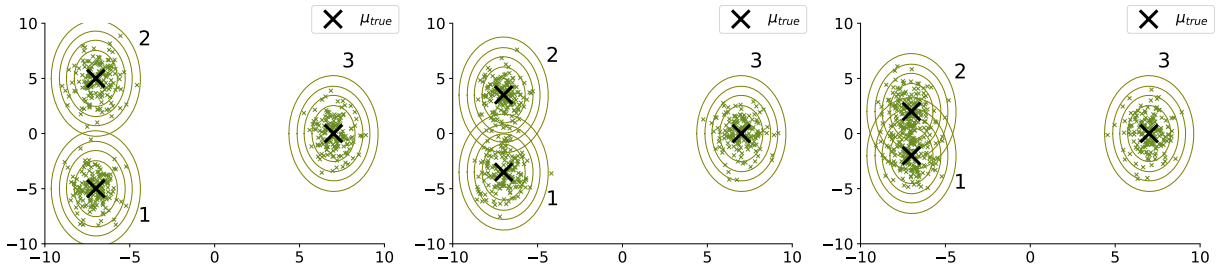


Fig. 4 500 sample points from a Mixture of Gaussians with 3 classes. The true centroid of each Gaussian are depicted by black crosses, and their true covariance matrices are represented by the confidence ellipses of level 0.8, 0.99 and 0.999 around the centre. There are three different versions of the true parameters. From left to right: the true μ_k of the two left clusters (mu_1 and mu_2) are getting closer while everything else stays identical.

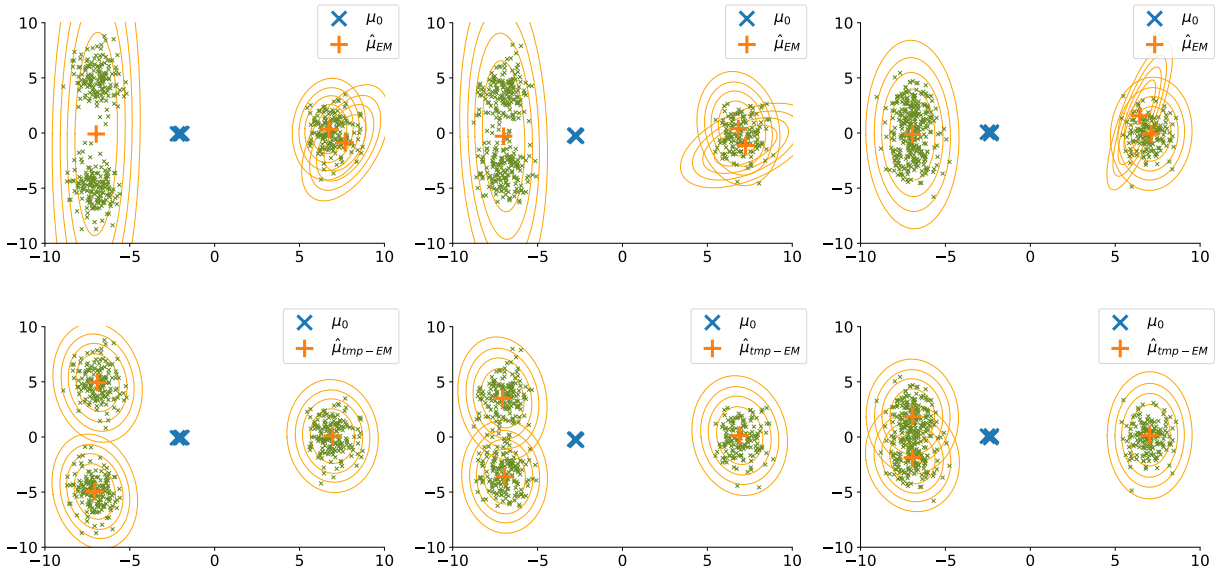


Fig. 5 Typical final positioning of the centroids by EM (first row) and tmp-EM (second row) **when the initialisation is made at the barycenter of all data points** (blue crosses). The three columns represent the three gradually more ambiguous parameter sets. Each figure represents the positions of the 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.

barycenter initialisation is now around 0.10, staying 10 times smaller than without tempering. With *2v1*, the relative error stayed fairly stable, reaching now 0.35 in average, and remaining approximately 5 times smaller than without tempering. We underline that up until this point (parameter families 1 and 2), the standard deviation of these errors was 3 times smaller with tempering in the case of the *barycenter* initialisation, and around 2 times smaller in the case of the *2v1* initialisation. In the final configuration, parameter family 3, the relative errors with tempering are 0.30 with the *barycenter* initialisation (5 times smaller than without tempering) and 0.40 with the *2v1* initialisation (more than 4.5 times smaller than without tempering). Moreover, the standards deviations are at least 1.8 times smaller with tempering. We note that, in similar fashion to EM, the errors on μ_1 and μ_2 with the *barycenter* initialisation reached, in the most ambiguous configuration, the level of error seen with the *2v1* initialisation in the least ambiguous situation: 0.30. Which, as stated, remains 5 times smaller than the corresponding level of error without tempering: 1.50.

In the end, the progression of errors when μ_1 and μ_2 get closer is alike between EM and tmp-EM: the *barycenter* initialisation is the most affected, the *2v1* initialisation error being higher but fairly stable. However the level of error is much smaller with tmp-EM, being 5 to 10 times smaller in the case of the *barycenter* initialisation, and

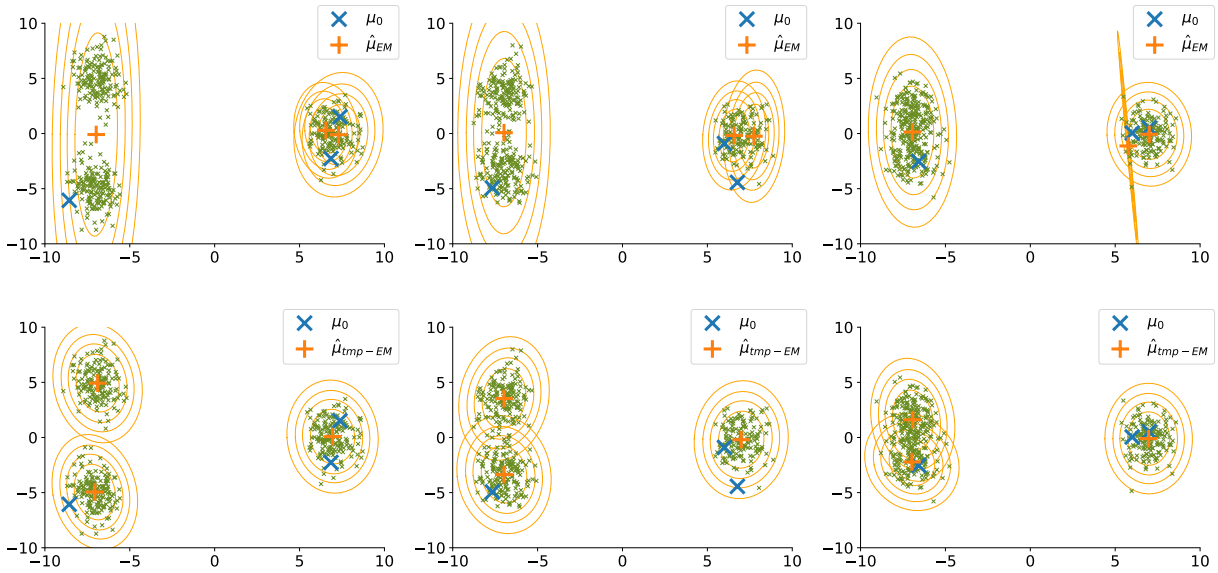


Fig. 6 Typical final positioning of the centroids by EM (first row) and tmp-EM (second row) **when the initialisation is made by selecting two points in the isolated cluster and one in the lower ambiguous cluster** (blue crosses). The three columns represent the three gradually more ambiguous parameter 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.

4.5 to 5 times smaller for the *2v1* initialisation. Similarly, the standard deviation around those average levels is 1.8 to 2 times smaller with tmp-EM.

These quantitative results on the reconstruction error of μ_1 and μ_2 confirm exactly what was observed on the illustrative examples: with tempering, the EM procedure is much more likely to discern the true position of the three clusters regardless of the initialisation, and able to reach a very low error rate even with the most adversarial initialisations. To bolster this last point, we underline that even in the worst case scenario, *2v1* initialisation and very close μ_1 and μ_2, tmp-EM still outperforms EM in the best scenario, *barycenter* initialisation and well separated clusters, with an error rate of 0.40 versus 0.50.

5 Tempered Riemann approximation EM

5.1 Context, Theorem and proof

The Riemann approximation of Section 3 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...) that are commonly used in those cases. The tempering approximation of Section 4 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. 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 $\{T_n\}_n \in (\mathbb{R}_+^*)^{\mathbb{N}}$, $T_n \underset{n\infty}{\longrightarrow} 1$.

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 2.

Theorem 4 *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 remain-*

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 and tmp-EM over 1000 simulated dataset with two different initialisations. The three different parameter families, described in Fig. 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.

		EM		tmp-EM	
Parameter family	cl.	barycenter	2v1	barycenter	2v1
1	1	0.52 (1.01)	1.52 (1.24)	**0.04 (0.26)**	**0.29 (0.64)**
	2	0.55 (1.05)	1.53 (1.25)	**0.05 (0.31)**	**0.30 (0.64)**
	3	**0.01 (0.06)**	**0.01 (0.03)**	0.03 (0.17)	0.03 (0.19)
2	1	1.00 (1.42)	1.69 (1.51)	**0.09 (0.47)**	**0.37 (0.86)**
	2	1.03 (1.44)	1.71 (1.52)	**0.12 (0.57)**	**0.32 (0.79)**
	3	0.01 **(0.05)**	**0.02 (0.03)**	$\mathbf{5.10^{-3}}$ (0.05)	0.04 (0.22)
3	1	1.56 (1.75)	1.79 (1.77)	**0.31 (0.97)**	**0.39 (0.98)**
	2	1.51 (1.74)	1.88 (1.76)	**0.30 (0.93)**	**0.39 (0.97)**
	3	0.02 (0.04)	**0.02 (0.04)**	**0.01 (0.04)**	0.07 (0.30)

ing conditions of applicability of Theorem 1 as long as $z \mapsto S(z)$ *is continuous and* $\{T_n\}_n \in (\mathbb{R}_+^*)^{\mathbb{N}}$, $T_n \underset{n\infty}{\longrightarrow} 1$.

Proof This proof of Theorem 4 is very similar to the proof of Theorem 2 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:

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

In the proof of Theorem 2, 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 $h(z;\theta)$ hence 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 $T_n > 0$, $T_n \underset{n\infty}{\longrightarrow} 1$, 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 2, that:

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

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:

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

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

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

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 2.

5.2 Application to a Gaussian model with the Beta prior

We illustrate the method with the model of Section 3.3:

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

We apply the tempered Riemann approximation. As in Section 3.3, 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 Eq. (9), 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. Fig. 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 $T_0 = 150, r = 3, a = 0.02, b = 40$.

6 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 step, 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.

For cases where one wants to improve the solution of the EM, we proved that the tempered EM, introduced under a different form in Ueda and Nakano (1998), is a specific case of the Deterministic Approximate EM. Moreover, we showed that the commonly used models benefit from the convergence property as long as the temperature profile converges towards 1. This justifies the use of many more temperature profiles than the ones tried in Ueda and Nakano (1998) and Naim and Gildea (2012). We ran an in-depth empirical comparison between tmp-EM and the regular EM. In particular, we showed how tmp-EM was able to escape from adversarial initial positions, a task that sometimes required complex non-monotonous temperature schemes, which are covered by our Theorem.

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.

Acknowledgments

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).

References

Aarts E, Korst J (1988) Simulated annealing and Boltzmann machines. New York, NY; John Wiley and Sons Inc.

Allassonnière S, Chevallier J (2019) A New Class of EM Algorithms. Escaping Local Minima and Handling Intractable Sampling, URL https://hal.archives-ouvertes.fr/hal-02044722, working paper or preprint

Allassonnière S, Kuhn E, Trouvé A, et al. (2010) Construction of bayesian deformable models via a

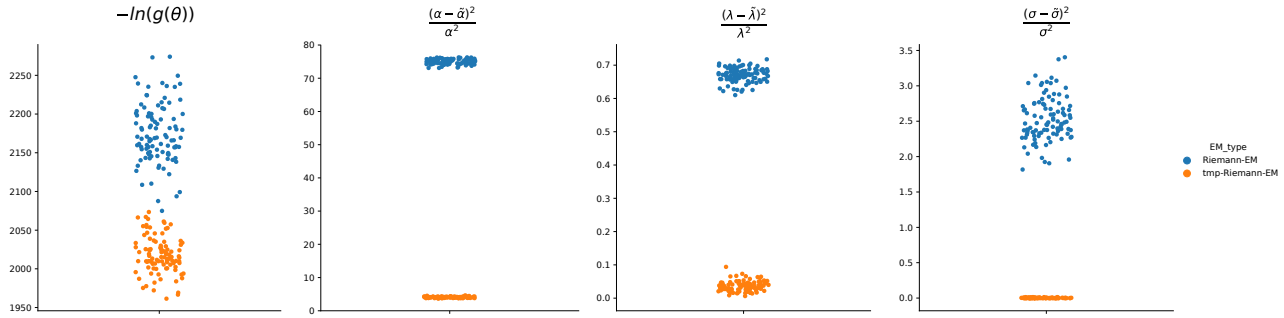


Fig. 7 Results over many simulations of the Riemann EM and tmp-Riemann EM on the Beta-Gaussian model. The tempered Riemann EM reaches relative errors on the real parameters that are orders of magnitude below the Riemann EM with no temperature. The likelihood reached is also lower with the tempering.

stochastic approximation algorithm: a convergence study. Bernoulli 16(3):641–678

Balakrishnan S, Wainwright MJ, Yu B, et al. (2017) Statistical guarantees for the em algorithm: From population to sample-based analysis. The Annals of Statistics 45(1):77–120

Booth JG, Hobert JP (1999) Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61(1):265–285

Boyles RA (1983) On the convergence of the em algorithm. Journal of the Royal Statistical Society: Series B (Methodological) 45(1):47–50

Cappé O, Moulines E (2009) On-line expectation–maximization algorithm for latent data models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71(3):593–613

Chen HF, Guo L, Gao AJ (1987) Convergence and robustness of the robbins-monro algorithm truncated at randomly varying bounds. Stochastic Processes and their Applications 27:217–231

Chen J, Zhu J, Teh YW, Zhang T (2018) Stochastic expectation maximization with variance reduction. Advances in Neural Information Processing Systems 31:7967–7977

Delyon B, Lavielle M, Moulines E, et al. (1999) Convergence of a stochastic approximation version of the em algorithm. The Annals of Statistics 27(1):94–128

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society: Series B (Methodological) 39(1):1–22

Dwivedi R, Ho N, Khamaru K, Jordan MI, Wainwright MJ, Yu B (2018) Singularity, misspecification, and the convergence rate of em. arXiv preprint arXiv:181000828

Dwivedi R, Ho N, Khamaru K, Wainwright M, Jordan M, Yu B (2020) Sharp analysis of expectation-maximization for weakly identifiable models. In: International Conference on Artificial Intelligence and Statistics, PMLR, pp 1866–1876

Fort G, Moulines E (2003) Convergence of the monte carlo expectation maximization for curved exponential families. The Annals of Statistics 31(4):1220–1259

Fort G, Moulines E, Wai HT (2020) A stochastic path integral differential estimator expectation maximization algorithm. Advances in Neural Information Processing Systems 33

Geyer CJ, Thompson EA (1995) Annealing markov chain monte carlo with applications to ancestral inference. Journal of the American Statistical Association 90(431):909–920

Ho N, Nguyen X, et al. (2016) Convergence rates of parameter estimation for some weakly identifiable finite mixtures. The Annals of Statistics 44(6):2726–2755

Hukushima K, Nemoto K (1996) Exchange monte carlo method and application to spin glass simulations. Journal of the Physical Society of Japan 65(6):1604–1608

Jank W (2005) Quasi-monte carlo sampling to improve the efficiency of monte carlo em. Computational statistics & data analysis 48(4):685–701

Karimi B, Wai HT, Moulines E, Lavielle M (2019) On the global convergence of (fast) incremental expectation maximization methods. In: Advances in Neural Information Processing Systems, pp 2837–2847

Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. science 220(4598):671–680

Kuhn E, Lavielle M (2005) Maximum likelihood estimation in nonlinear mixed effects models. Computational Statistics & Data Analysis 49(4):1020–1038

Lange K (1995) A gradient algorithm locally equivalent to the em algorithm. Journal of the Royal Statistical Society: Series B (Methodological) 57(2):425–437

Levine RA, Casella G (2001) Implementations of the monte carlo em algorithm. Journal of Computational and Graphical Statistics 10(3):422–439

Levine RA, Fan J (2004) An automated (markov chain) monte carlo em algorithm. Journal of Statistical Computation and Simulation 74(5):349–360

Naim I, Gildea D (2012) Convergence of the em algorithm for gaussian mixtures with unbalanced mixing coefficients. arXiv preprint arXiv:12066427

Neal RM, Hinton GE (1998) A view of the em algorithm that justifies incremental, sparse, and other variants. In: Learning in graphical models, Springer, pp 355–368

Ng SK, McLachlan GJ (2003) On the choice of the number of blocks with the incremental em algorithm for the fitting of normal mixtures. Statistics and Computing 13(1):45–55

Pan JX, Thompson R (1998) Quasi-monte carlo em algorithm for mles in generalized linear mixed models. In: COMPSTAT, Springer, pp 419–424

Swendsen RH, Wang JS (1986) Replica monte carlo simulation of spin-glasses. Physical review letters 57(21):2607

Titterington D, Smith A, Makov U (1985) Statistical Analysis of Finite Mixture Distributions. Wiley, New York

Ueda N, Nakano R (1998) Deterministic annealing em algorithm. Neural networks 11(2):271–282

Van Laarhoven PJ, Aarts EH (1987) Simulated annealing. In: Simulated annealing: Theory and applications, Springer, pp 7–15

Wei GC, Tanner MA (1990) A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. Journal of the American statistical Association 85(411):699–704

Winkelbauer A (2012) Moments and absolute moments of the normal distribution. arXiv preprint arXiv:12094340

Wu CJ (1983) On the convergence properties of the em algorithm. The Annals of statistics 11(1):95–103