Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The windowEM algorithm

Carsten Wiuf

wiuf@math.ku.dk

University of Copenhagen

Malthe Sebro Rasmussen

University of Copenhagen

Research Article

Keywords: EM algorithm, Exponential family, Convergence, Contraction

Posted Date: July 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10163958/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

The `windowEM` algorithm

Carsten Wiuf[1*] and Malthe Sebro Rasmussen[2]

[1*]Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, Copenhagen, 2100, Denmark.
[2]Globe Institute, University of Copenhagen, Øster Voldgade 5-7, Copenhagen, 2200, Denmark.

*Corresponding author(s). E-mail(s): wiuf@math.ku.dk;
Contributing authors: malthe.rasmussen@sund.ku.dk;

Abstract

We propose a stochastic version of the EM algorithm that partitions the data into $\boldsymbol{B}$ blocks and places the blocks on a circle. The standard EM algorithm is used to update the block estimates sequentially such that the estimate of the current block is used as seed for the estimate of the next block. The estimates might further be smoothed by averaging the estimates in a rolling window of blocks. In this way, a population of $\boldsymbol{B}$ estimates is produced. A single estimate might be produced taking the average of all $\boldsymbol{B}$ estimates. Contraction theory and random iteration theory are used to infer convergence properties assuming $\boldsymbol{B} < \boldsymbol{\infty}$ or $\boldsymbol{B} = \boldsymbol{\infty}$. Empirical examples show that this approach might prevent over-fitting.

Keywords: EM algorithm, Exponential family, Convergence, Contraction

MSC Classification: 62-02 , 62F12

Contents

1 Introduction **2**

2 The `windowEM` algorithm **3**
2.1 The setting 4
2.2 The algorithm 5
2.3 Special cases 5
2.4 Extensions 6

3 Exponential families **7**
3.1 General exponential families 7
3.2 Mixture of distributions from an exponential family 8
3.3 Partitioning the sufficient statistics 9

4 Examples **10**
4.1 An illustrative example 10
4.2 Univariate Gaussian mixtures 13

5 Convergence **16**
5.1 Fixed points for $B < \infty$ 17
5.2 Lipschitz condition 18
5.3 Convergence for $B = \infty$ 20
5.4 Theorem 10 and convergence 21

A Mixture of Gaussian distributions **23**

B Concavity **23**

1 Introduction

This paper is concerned with a stochastic version of the expectation-maximization (EM) algorithm [6], based on averaging parameter estimates within a rolling window of smaller block estimates. This algorithm, which we refer to as `windowEM`, was originally presented in [16] and adapted for use in [12]. The motivation is to do inference on large-scale genetic data with latent variables, such as inference on site frequency spectra [16] or probabilistic PCA [19]. In these settings, where the dataset potentially comprises billions of data points, the standard EM algorithm becomes impractical since it makes use of all data in each iteration.

In response to the increasing data issue, various stochastic versions of the EM algorithm have been proposed, in which the data is split into smaller blocks (sometimes called "mini-batches") [4, 14]. With large amounts of data, such an approach may drastically speed up convergence by allowing multiple parameter updates per pass through the data. With enough data, it may even be possible to reach convergence in a single (or very few) passes, such that data can be kept out of RAM.

We find, however, that when the parameter space is complex, and particularly when some parameter estimates rely on latent variables that are scarcely represented in the input data, a stochastic approach based on individual blocks result in roughed or noisy parameter estimates (e.g., in the case of the site frequency spectra, the probability of a particular allele frequency is expected to vary monotonically with the allele frequency), even when introducing decaying weights [4].

Similar to other stochastic EM algorithms, `windowEM` splits data randomly into blocks. However, in the `windowEM` algorithm, these blocks are placed on a circle. Inspired by Polyak-Ruppert averaging [15, 17], estimates are then averaged over a rolling window of blocks on this circle. The effect of this windowing is to produce a sequence of *dampened estimates*, counteracting the aforementioned noise, while still

allowing frequent updates and fast convergence. For block k, a preliminary estimate is obtained by performing one iteration of the standard EM algorithm with data from block k, using the estimate of the previous block k as input value. That is, the preliminary estimate is $\psi_k = \arg\max_\psi Q_k(\psi|\theta_{k-1})$, where Q_k is the maximisation function of the standard EM algorithm applied to the data of block k. A final estimate is obtained by averaging all w estimates in the window, $\theta_k = (\psi_k + \psi_{k-1} + \ldots + \psi_{k-w+1})/w$.

While previous work has shown the utility of this scheme in practice, no work has been done on its theoretical foundation. Thus, in the present paper, we investigate the convergence properties of the `windowEM` algorithm. We distinguish two cases, depending on the number of blocks B:

- B is finite, or
- B is infinite.

In the first case, the whole data is repeatedly used to update parameter estimates in a circular manner. In effect, this yields a population of B block estimates, rather than a single estimate. If the update function is a contraction, then we have convergence of estimates and a fixed point exists for one block if and only if a fixed point exists for all blocks. A single estimate might be obtained by averaging over all B estimates. The distribution of block estimates depends on the window size, and a smaller variance in block estimates is observed empirically with larger window sizes.

In the second case, the number of blocks is considered so large that we never see the same block twice. For this case, we give conditions under which convergence of estimates occur to a fixed point (for large block indices), based on random function iteration theory for dependent data [5, 7]. Moreover, a single combined estimate might be obtained by taking the (ergodic) average of block estimates, and we give conditions that guarantee convergence of this average in the limit as $B \to \infty$. In both scenarios, B is finite, as well as B is infinite, we assume that the update function is a contraction, though in the latter case, this is only required to hold in expectation with respect to the input data.

Having a population of estimates rather than a single 'best' estimate might give a better idea about the true parameter. The window size determines how much variation in the estimates is allowed. In concrete applications in genomics [12, 16], we have shown that when estimating the site frequency spectrum (SFS) of the rarer allele in large samples, averaging across a window smooths out the estimate of the SFS, contrary to maximum likelihood estimation that leaves 'bump' in the SFS because of rare events that artificially raises an estimated frequency, and we see a similar tendency for mixtures of normals below. We speculate, therefore, that `windowEM` might also help prevent over-fitting.

The results are illustrated with familiar examples.

2 The `windowEM` algorithm

In this section we present the `windowEM` algorithm, first introduced in [16]. The primary questions are in what sense the `windowEM` algorithm might converge and under what conditions it might do so. We give partial answers to both in Section 5 below.

2.1 The setting

Let $x_1, \ldots, x_B$ be a partition of the available data into B blocks, $B \geq 1$. The blocks might not be of the same size, nor be independent, nor have the same distribution, even if the blocks are of the same size. Hence, we assume $x_k \in \mathcal{X}_k \subseteq \mathbb{R}^{m_k}$ and that $(\mathcal{X}_1 \times \ldots \times \mathcal{X}_B, \mathcal{B}_1 \otimes \ldots \otimes \mathcal{B}_B, P)$ is a measure space, where P is the joint distribution of the data, and $\mathcal{B}_k$ the Borel algebra on $\mathcal{X}_k$. Contrary to this, we do assume the model parameters are comparable and vary in the same *convex* domain $\Theta \subseteq \mathbb{R}^m$ for all blocks. We allow $B = \infty$. The blocks might be arranged in random order to reduce correlation between neighbouring blocks.

For $t \geq 1$, define $b(t) = (t-1) \bmod B + 1$, such that $b(Bi+k) = k$ for $k = 1, \ldots, B$, $i \geq 0$. The function $b(\cdot)$ maps the tth iteration of the algorithm to a corresponding block. If $B = \infty$, then $b(t) = t$. The sequence of blocks are paved with windows, consisting of w consecutive blocks, $1 \leq w \leq B$. The kth window consists of the k blocks numbered

$$b(k-w+1), \ldots, b(k-1), k, \quad k = 1, \ldots, B,$$

noting that $b(k) = k$ always, whereas for example, $b(k-1) = k-1$ or $b(k-1) = B$, depending on whether $k > 1$ or $k = 1$, respectively. There are B distinct windows.

Denote by

$$\ell_k(\theta) = \log(p(x_k|\theta)), \quad k = 1, \ldots, B, \quad \theta \in \Theta \subseteq \mathbb{R}^m,$$

the log-likelihood function of the data in the kth block, where Θ is a convex parameter space, and $p(x_k|\theta)$ is the likelihood function. The log-likelihood might be given as

$$\ell_k(\theta) = \log(p(x_k, z_k|\theta)) - \log(p(z_k|x_k, \theta)), \tag{1}$$

where $z_k \in \mathcal{Z}_k \subseteq \mathbb{R}^{m'_k}$, $k = 1, \ldots, B$, are unobserved (latent) data [6]. Equation (1) is well-known from the EM algorithm [6]. Hence, by taking the conditional expectation of $\ell_k(\theta)$, given x_k, with respect to $\theta' \in \Theta$,

$$\begin{aligned}\ell_k(\theta) &= \sum_{z_k} p(z_k|x_k, \theta') \log(p(x_k, z_k|\theta)) - \sum_{z_k} p(z_k|x_k, \theta') \log(p(z_k|x_k, \theta)) \\ &= Q_k(\theta|\theta') - H_k(\theta|\theta'),\end{aligned}$$

where $Q_k(\theta|\theta')$ and $H_k(\theta|\theta')$ are the two sums in the equation. Computation and maximisation of $Q_k(\cdot|\theta')$ are the E- and the M-step of the standard EM algorithm, restricted to the kth block [6]. If data in different blocks are independent of each other, then the full log-likelihood is obtained as the sum over k of ℓ_k.

Above we assume discrete data for convenience and that the involved sums and functions are well-defined. However, the theory also holds more generally under standard regularity conditions [6].

2.2 The algorithm

The division of the data into blocks suggests an optimisation procedure, here called `windowEM`, which is related to the Polyak-Ruppert averaging scheme though different [15, 17]. To state the algorithm, we define the functions $f_k \colon \Theta \to \Theta$, $k = 1, \ldots, B$, by

$$f_k(\theta) = \operatorname{argmax}_{\theta' \in \Theta} Q_k(\theta'|\theta),$$

assuming there exists a *unique* $\theta' \in \Theta$ such that $f_k(\theta) = \theta'$. The steps of `windowEM` are the following:

A. Initialise $(\theta_1, \ldots, \theta_w) \in \Theta^w$. Define $\psi_{t+1} = f_{b(t+1)}(\theta_t)$, $t = 1, \ldots, w$,
B. For $t \geq w+1$, recursively compute θ_t as the average of the past w consecutive values of ψ_i,

$$\theta_t = \frac{1}{w} \sum_{i=t-w+1}^{t} \psi_i, \quad \text{and let} \quad \psi_{t+1} = f_{b(t+1)}(\theta_t).$$

The estimate θ_t for block $b(t)$ is computed from the estimates $\theta_{t-1}, \ldots, \theta_{t-w}$ from the neighbouring blocks $b(t-1), \ldots, b(t-w)$. After B iterations, the algorithm has returned to the first block again, unless $B = \infty$, and there is only one infinite round of iterations. We refer to k successive iterations as a *k-iteration* and a B-iteration as an *epoch*. When $B = \infty$, there is therefore only one epoch.

Since Θ is convex (by assumption), then $\psi_{t+1}, \theta_t \in \Theta$ for all $t \geq 1$. This convexity requirement is naturally fulfilled for the canonical and the mean value parameterisations of a regular exponential family [8]. In particular, the mean value parameterisation seems a natural choice as it often leads to an update function that is a weighted sum of sufficient statistics, see Section 3. Averaging across a window as in B) thus preserves this property. It is, however, noteworthy that averaging across a window is not invariant under parameter transformations as the operations + and f_k do not commute in general. Let $\phi = g(\theta)$ be an alternative parameterisation, then

$$\theta_{t+1} = \frac{1}{w} \sum_{i=t-w+1}^{t} f_{b(i+1)}(\theta_i) \quad \Rightarrow \quad \phi_{t+1} = g\Big(\frac{1}{w} \sum_{i=t-w+1}^{t} f_{b(i+1)}(g^{-1}(\phi_i))\Big).$$

If $w = 1$, then averaging commutes with g, $\phi_{t+1} = g(f_{b(t+1)}(g^{-1}(\phi_t)))$, whereas in general the average commutes with g only if g is linear, $g(\theta) = A\theta$,

$$g\Big(\frac{1}{w} \sum_{i=t-w+1}^{t} f_{b(i+1)}(g^{-1}(\phi_i))\Big) = \frac{1}{w} \sum_{i=t-w+1}^{t} A f_{b(i+1)}(A^{-1}\phi_i).$$

2.3 Special cases

If $w = B = 1$, then we retrieve the standard EM algorithm and the likelihood increases after each epoch (one iteration in this case). However, in contrast to the standard EM algorithm, `windowEM` is not designed to increase the likelihood in each successive

iteration or epoch. If $w = 1$ and $B \geq 1$, then $\theta_{t+1} = \psi_{t+1}$ and

$$\ell_{b(t+1)}(\theta_{t+1}) \geq \ell_{b(t+1)}(\theta_t), \quad t \geq 1,$$

as for the standard EM algorithm, but there are no guarantees the log-likelihood value one epoch later is larger than the current log-likelihood value.

If a fixed point q_k, $k = 1, \ldots, B$, exists for each block, implying q_k remains the same after iterating one epoch, then

$$q_{b(k+1)} = \frac{1}{w} \sum_{i=k-w+1}^{k} f_{b(i+1)}(q_{b(i)}), \quad k = 1, \ldots, B.$$

If $w = B$, then the right hand side does not depend on k and consequently, $q_k = q_\ell$ for all $k, \ell = 1, \ldots, B$, and there is a common q^* solving the equations. This estimate might or might not equal the maximum likelihood estimate, even in cases where the maximum likelihood estimate is unique. However, for exponential families, the update function f_k considered as a function of the input data x_k and the parameter θ, is in many cases linear in x_k, see Section 3. If this is so, then the equation reads (with q the fixed point)

$$q = \frac{1}{B} \sum_{k=1}^{B} f_k(q) = f(q),$$

where $f \colon \Theta \to \Theta$ is the update function with data $\frac{1}{w} \sum_{k=1}^{B} x_k$. Since the standard EM algorithm is guaranteed to find a local maximum [6], then so will `windowEM` in this particular case.

On the other hand, if $w = 1$, then

$$\begin{aligned} q_1 &= f_1(q_B), \quad q_{k+1} = f_{k+1}(q_k), \quad k = 1, \ldots, B-1, \\ q_B &= f_B \circ f_{B-1} \circ \ldots \circ f_2 \circ f_1(q_B), \end{aligned}$$

and similar for q_k, $k \neq B$. Thus, one might solve for q_B and then find q_k, $k \neq B$, afterwards. In particular, if the composition $f_B \circ f_{B-1} \circ \ldots \circ f_2 \circ f_1$ is a contraction, then there is a unique q_B. If $f_1, \ldots, f_B$ are increasing and concave, then so is the composition, see the Appendix, Section B. For a 1-dimensional parameter this implies there is at most two fixed points, only one of which is attractive; unless the composition is the identity.

2.4 Extensions

One might replace the single updating step, $(*)$ $\psi_{t+1} = f_{b(t+1)}(\theta_t)$, by multiple updating steps, say N, before averaging over block estimates:

$$\xi_0 = \theta_t, \quad \xi_{n+1} = f_{b(t+1)}(\xi_n), \quad n = 0, \ldots, N-1, \quad \psi_{t+1} = \xi_N.$$

With large N, ψ_{t+1} becomes a local maximum point for block $b(t+1)$. If there is a unique local maximum, then this maximum point does not depend on the other blocks and is the maximum likelihood estimate of block $b(t+1)$.

Another alternative to the single updating step is to replace EM with generalized EM [6], where $(*)$ is replaced by choosing $\widehat{\psi}_{t+1}$, such that $Q_{b(t+1)}(\widehat{\psi}_{t+1}|\theta_t) \geq Q_{b(t+1)}(\theta_t|\theta_t)$, or replacing $Q_k(\cdot|\cdot)$ with another function, as in the case of the MM (majorize–minimize) algorithm [20].

Finally, each term in the sum in B has equal weight. As an alternative, one might weigh the w terms differentially. For example, the weight of $\widehat{\psi}_t$ might be put higher than that of $\widehat{\psi}_{t-1}$ and so forth, or if the data structure differs from block to block, for example, if $x_k \sim \text{Bi}(n_k, \theta)$, then one might weigh ψ_t by n_k whenever $b(t) = k$.

In [12], an approximation to PCA is considered where data also is divided into blocks. The block size is increased after each round of iterations and something similar might be considered for `windowEM`.

3 Exponential families

We consider the special case of an exponential family with densities,

$$\frac{dP_\theta}{dP_0}(x, z) = \exp(\theta^\top T(x, z) - a(\theta)), \quad (x, z) \in \mathcal{X} \times \mathcal{Z}, \quad \theta \in \Theta \subseteq \mathbb{R}^n,$$

where x is observed, z is latent data, and Θ is the (convex) canonical parameter space, containing $0 \in \Theta$ (which might always be arranged). The family is assumed to be *regular*, that is, the canonical parameter space is an open set.

Let $\tau(\theta) = \mathbb{E}_\theta(T(X, Z))$ be the mean value parameter as a function of θ. It might be obtained as the derivative of $a(\theta)$: $\tau(\theta) = \nabla_\theta a(\theta)$, which is an invertible function if the family is *minimal* [8]. Hence, θ and τ are in one-to-one correspondence. We will formulate the EM algorithm using the mean value parameterisation as this turns out to be a natural choice, but apply both θ, τ and other parameterisations in equations for notational convenience. In the following, we assume the model is minimal unless otherwise said. However, in applications this might not be so. For details, we refer to textbooks on the subject [8, 18].

3.1 General exponential families

For a general exponential family, the update step for the `windowEM` in the mean value parameterisation becomes:

$$\tau_{t+1} = \frac{1}{w} \sum_{k=t-w+1}^{t} \mathbb{E}_{\theta_k}(T(x_{b(k+1)}, Z_{b(k+1)})|x_{b(t+1)}),$$

relying on the update step of the standard EM algorithm for exponential families [18].

In genomics the data often consist of millions of markers arranged in a linear order, hence each block of data is itself a sequence of data points. The observed data might

be measurements of genotypes and the latent variables, the true genotypes, or the number of dominant genotypes versus the allele counts. If so, then

$$\begin{aligned}\mathbb{E}_{\theta_k}(T(x_{b(k+1)}, Z_{b(k+1)})|x_{b(k+1)}) &= \frac{1}{n_{b(k+1)}} \sum_{i=1}^{n_{b(k+1)}} \mathbb{E}_{\theta_k}(T(x_{b(k+1),i}, Z_{b(k+1),i})|x_{b(k+1),i}) \\ &= \frac{1}{n_{b(k+1)}} \sum_{i=1}^{n_{b(k+1)}} \mathbb{E}_{\theta_k}(T(x_{b(k+1),i}, Z_{b(k+1),i})|x_{b(k+1)}),\end{aligned}$$

where $n_{b(k)}$ is the number of data points in block $b(k)$. The latter equality assumes independence between sites to change the conditioning to be independent of i. Thus, the mean value parameterisation appears as a natural choice since the update τ_{t+1} becomes a weighted average of terms that are similar to each other.

3.2 Mixture of distributions from an exponential family

A standard example is a mixture of normal distributions [13, 21]. Here, we consider a mixture of distributions from a general exponential family with the previous section in mind.

To be precise, we assume the data $x_k = (x_{k1}, \ldots, x_{kn_k})$, $k = 1, \ldots, B$, of block k is distributed as n_k independent draws from a mixture of M distributions from the same regular exponential family. Let $(\pi, \phi) \in \Delta_M \times \Phi^M$ be the parameter vector, where Δ_M denotes the M-simplex, $\phi = (\phi_1, \ldots, \phi_M)$ are the canonical parameters of the mixture components, and $\pi = (\pi_1, \ldots, \pi_M)$, $\sum_{m=1}^M \pi_m = 1$ are the mixture proportions. Then, the density function for an observation x_{ki}, respectively, the density function of a complete observation (x_{ki}, z_{ki}), where z_{ki} is an unobserved class indicator, takes the form,

$$\begin{aligned}p(x_{ki}|\pi, \phi) &= \sum_{m=1}^M \pi_m p(x_{ki}|\phi_m) = \sum_{m=1}^M \pi_m \exp(\phi_m^\top T(x_{ki}) - a(\phi_m)), \\ p(x_{ki}, z_{ki}|\pi, \phi) &= \pi_{z_{ki}} \exp(\phi_{z_{ki}}^\top T(x_{ki}) - a(\phi_{z_{ki}})) \\ &= \exp\Big(\sum_{m=1}^M (\log(\pi_m) - a(\phi_m)) 1_{\{m\}}(z_{ki}) + \phi_m^\top 1_{\{m\}}(z_{ki}) T(x_{ki})\Big).\end{aligned}$$

Thus, the model of the complete data is another exponential family. As the indicators sum to one, this representation is not minimal. A minimal representation can be obtained by substituting $1 - \sum_{m=1}^{M-1} 1_{\{m\}}(z_{ki})$ for $1_{\{M\}}(z_{ki})$, and $1 - \sum_{m=1}^{M-1} \pi_m$ for π_M in the first part of the sum.

The (non-minimal) sufficient statistic is

$$\widehat{T}(X, Z) = \Big(1_{\{1\}}(Z), \ldots, 1_{\{M\}}(Z), 1_{\{1\}}(Z)T(X), \ldots, 1_{\{M\}}(Z)T(X)\Big),$$

with expectation (mean value parameter) $\psi = (\pi, \tau_1, \ldots, \tau_M)$, $\pi_m = \mathbb{E}_{\pi,\phi}(1_{\{m\}}(Z))$ and $\tau_m = \mathbb{E}_{\pi,\phi}(1_{\{m\}}(Z)T(X))$. From the standard EM algorithm, the update functions of the `windowEM` in the mean value parameterisation become

$$\pi_m^{(t+1)} = \frac{1}{w} \sum_{k=t-w+1}^{t} \frac{1}{n_{b(k+1)}} \sum_{i=1}^{n_{b(k+1)}} q_{k+1,i}(m), \quad \tau_m^{(t+1)} = \frac{1}{w} \sum_{k=t-w+1}^{t} \frac{\sum_{i=1}^{n_{b(k+1)}} q_{k+1,i}(m) T(x_{k+1,i})}{\sum_{i=1}^{n_{b(k+1)}} q_{k+1,i}(m)},$$

where

$$q_{k+1,i}(m) = p(z_{b(k+1),i} = m | x_{b(k+1),i}, \psi^{(k)}), \quad \psi^{(k)} = (\pi^{(k)}, \tau^{(k)}).$$

Averaging over the estimates in a window remains a sum of observed statistics, weighted by estimated class probabilities. Thus, the update functions are linear in the observed statistics with coefficients that depend on the observations.

Note that a (non-minimal) sufficient statistics might alternatively be given as

$$\begin{aligned}\widetilde{T}(X, Z) &= A\widehat{T}(X, Z) \\ &= \Big(1_{\{1\}}(Z), \ldots, 1_{\{M-1\}}(Z), 1, 1_{\{1\}}(Z)T(X), \ldots, 1_{\{M-1\}}(Z)T(X), T(X)\Big),\end{aligned}$$

where

$$A = \begin{pmatrix} I_{M-1} & 0 & 0 & 0 \\ e_{M-1} & 1 & 0 & 0 \\ 0 & 0 & I_{M-1} & 0 \\ 0 & 0 & e_{M-1} & 1 \end{pmatrix},$$

I_{M-1} is the identity matrix of dimension $M-1$ and e_{M-1} is a row vector of $M-1$ ones. The example might thus be seen as an example where one component of the sufficient statistics is observed, namely $(1, T(X))$.

3.3 Partitioning the sufficient statistics

Let $T = (U, V)$ be a partition of T of dimension d_1 and d_2, respectively, and assume $V = v$ is observed. Consider the corresponding partitions of $\theta = (\theta_1, \theta_2)$ and $\tau = (\tau_1, \tau_2)$.

Lemma 1 *The update function of the standard EM algorithm in the mean value parameterisation is a linear function of v,*

$$\tau_1' = \mathbb{E}_{\theta_1}(U|v) = A(\tau)v, \qquad \tau_2' = v, \tag{2}$$

where $\theta_1 = \theta_1(\tau)$, and τ is the current parameter estimate. After one iteration of the standard EM algorithm, τ_2 is fixed at v.

Proof It follows from [18] that the collection of conditional probability measures of $T(X, Z)$ given $V = v$ forms a regular exponential family with canonical parameter $\theta_1 \in \Theta_1$ (the restriction of Θ to the first d_1 coordinates) and densities,

$$\frac{dP_{\theta_1}}{dP_0}(u|v) = \exp(\theta_1^\top u - a_v(\theta_1)),$$

where $a_v(\theta_1)$ is a normalisation constant.

Let $\tau' = (\tau_1', \tau_2')$ and $\theta' = (\theta_1', \theta_2')$ be corresponding parameter values. In the canonical parameterisation, we find

$$\begin{aligned} Q(\theta'|\theta) &= \int (\theta'^{\top} T - a(\theta')) \exp(\theta_1^{\top} U - a_v(\theta_1) dP_0(\cdot|x) \\ &= \theta_1'^{\top} \mathbb{E}_{\theta_1}(U|v) + \theta_2'^{\top} v - a(\theta'). \end{aligned}$$

Differentiating with respect to (θ_1', θ_2') yields

$$\nabla_{\theta_1'} Q(\theta'|\theta) = \mathbb{E}_{\theta_1}(U|v)^{\top} - \tau_1'^{\top}, \qquad \nabla_{\theta_2'} Q(\theta'|\theta) = v^{\top} - \tau_2'^{\top}.$$

Equating the derivative to zero yields $\tau_2' = v$ and $\tau_1' = \mathbb{E}_{\theta_1}(U(Z)|v)$. Existence and uniqueness of such τ' follow from regularity [18]. This gives the first equality for τ_1' in (5).

For the second equality for τ_1' in (5), we consider the projection in L^2, equipped with the measure $P_{\theta_1}(\cdot|v)$, of each coordinate function of $U(Z)$ onto the linear span of the coordinate functions of V. Let the coordinate-wise projection be $\widehat{V}$. Then, $\mathrm{Cov}_{\theta_1}(U - \widehat{V}, \widehat{V}|v) = 0$. Since we are dealing with an exponential family, zero covariance implies independence, hence $U - \widehat{V}$ and $\widehat{V}$ are independent with respect to $P_{\theta_1}(\cdot|v)$ [1]. Finally, the conditional expectation yields

$$\mathbb{E}_{\theta_1}(U|v) = \mathbb{E}_{\theta_1}(U - \widehat{V}|v) + \mathbb{E}_{\theta_1}(\widehat{V}|v) = 0 + \widehat{V} = \widehat{V},$$

by independence and V-measurability of $\widehat{V}$. It remains to say that the linear projection is a linear map of V. See [11] for an elaborated proof for the Gaussian exponential family. □

In the context of `windowEM`, this implies that the mean value parameter is updated according to the rule

$$\tau_1^{(t+1)} = \frac{1}{w} \sum_{k=t-w+1}^{t} A(\tau^{(k)}) v_{b(k+1)}, \quad \tau_2^{(t+1)} = \frac{1}{w} \sum_{k=t-w+1}^{t} v_{b(k+1)},$$

with data v_k, $k = 1, \ldots, B$. Thus, $\tau^{(t)}$ is a weighted sum of the observations. Moreover, $\tau_2^{(t+1)}$ is unchanged after one epoch and is the average of the observations in the corresponding window, and only $\tau_1^{(t+1)}$ needs updating.

4 Examples

4.1 An illustrative example

This example is a classical application of the EM algorithm for which the maximum likelihood estimate can be found by simple calculation. However, it foresees the theoretical development in Section 5 about convergence of estimates.

Consider a gene with three genotypes AA, Aa and aa, where the A allele is recessive, such that there are two phenotypes, AA and aa-Aa (non-AA). We observe the number x of AA genotypes out of n. Assume Hardy-Weinberg proportions and that the probability for the A allele is $\theta \in [0, 1]$. The likelihood is binomial with probability $p = \theta^2$,

$$\binom{n}{x} \theta^{2x} (1 - \theta^2)^{n-x}, \quad x = 0, \ldots, n. \tag{3}$$

Let z be the latent count of the number of aa genotypes, such that the likelihood with data (x, z) is

$$\binom{n}{x\ z}\theta^{2x}(1-\theta)^{2z}(2\theta(1-\theta))^{n-x-z} = \binom{n}{x\ z}2^{n-x-z}\theta^{n+x-z}(1-\theta)^{n-x+z}. \tag{4}$$

The family is not regular (see Section 3) as the full parameter space of (4) is 2-dimensional.

Furthermore, the conditional distribution and expectation of z given x is

$$\binom{n-x}{z}2^{n-x-z}\frac{\theta^{n-x-z}(1-\theta)^z}{(1+\theta)^{n-x}}, \qquad \mathbb{E}_\theta(z|x) = (n-x)\,\frac{1-\theta}{1+\theta},$$

respectively. Taking expectation of the log of (4) for θ' with respect to the conditional distribution of z given x and θ (and removing terms not depending on θ), yields

$$Q(\theta'|\theta) = (n+x)\log(\theta') + (n-x)\log(1-\theta') - (n-x)\frac{1-\theta}{1+\theta}\log\left(\frac{\theta'}{1-\theta'}\right).$$

Finally, finding the optimal argument for $Q(\theta'|\theta)$ yields

$$\theta' = \frac{1}{2n}\left(n+x-(n-x)\frac{1-\theta}{1+\theta}\right) = \frac{u+\theta}{1+\theta}, \quad \text{with} \quad u = \frac{x}{n}. \tag{5}$$

The maximum likelihood estimate is $\widehat{\theta} = \sqrt{u}$, which can be found directly from (3). It is the only fixed point of the standard EM algorithm. It can readily be seen that $\theta < \theta'$ if $\theta < \sqrt{u}$, and $\theta > \theta'$ if $\theta > \sqrt{u}$ under the standard EM algorithm. Hence, convergence to the maximum likelihood estimate is guaranteed as this is the only local extremum.

Consider the `windowEM` algorithm with data $u_k = x_k/n_k$, $k = 1, \ldots, B$. From (5),

$$\psi_{t+1} = f_{b(t+1)}(\theta_t) = \frac{u_{b(t+1)} + \theta_t}{1+\theta_t},$$

hence

$$\theta_{t+1} = \frac{1}{w}\sum_{i=1}^{w}\frac{u_{b(t-i+2)} + \theta_{t-i+1}}{1+\theta_{t-i+1}} = 1 - \frac{1}{w}\sum_{i=1}^{w}\frac{1-u_{b(t-i+2)}}{1+\theta_{t-i+1}}. \tag{6}$$

Note that the update function, considered as a function of data and parameter, is linear in the data.

The map $f_{b(t+1)}$ is non-expansive,

$$\begin{aligned}\left\|\frac{\theta_t + u_{b(t+1)}}{1+\theta_t} - \frac{\theta'_t + u_{b(t+1)}}{1+\theta'_t}\right\| &= \frac{1-u_{b(t+1)}}{(1+\theta_t)(1+\theta'_t)}\|\theta_t - \theta'_t\| \\ &\leq (1-u_{b(t+1)})\|\theta_t - \theta'_t\| = A_{b(t+1)}\|\theta_t - \theta'_t\|,\end{aligned}$$

and a contraction if $u_{b(t+1)} > 0$. Here, $A_{b(t+1)} = 1 - u_{b(t+1)} \leq 1$ (non-expansive) and < 1 (contraction) if and only if $u_{b(t+1)} > 0$. It follows that the theory developed in Theorem 8 and 9 below applies for finite B, and that there is a unique fixed point for each block provided there is at least one non-zero observation x_k, $k = 1, \ldots, B$. If all observations are zero, the algorithm also converges to zero, which might be seen directly from (6). These conclusions do not require independence between blocks or identity in distribution.

Denote the fixed point of block k by q_k, $k = 1, \ldots, B < \infty$. Then,

$$q_{b(k+1)} = 1 - \frac{1}{w} \sum_{i=1}^{w} \frac{1 - u_{b(k-i+2)}}{1 + q_{b(k-i+1)}}. \tag{7}$$

In few cases, the fixed points can be found explicitly:

Lemma 2 *Assume $n_k = n$, $k = 1, \ldots, B$. If $w = B$, then all q_ks are the same and equal the maximum likelihood estimate of θ,*

$$\widehat{\theta} = \sqrt{\frac{1}{B} \sum_{k=1}^{B} u_k}.$$

Proof The right hand side of (7) is independent of k, hence $q_k = q$. The last part follows from (7) by putting $q_k = q$ and solving for q. Alternatively, by reference to Section 2.3 and the linearity of the update function in the input data. □

Lemma 3 *If $w = 1$, then*

$$\frac{1}{B} \sum_{k=1}^{B} q_k q_{b(k+1)} = \widehat{\theta}^2.$$

If in addition $B = 2$, then q_1 and q_2 are

$$q_1 = \frac{1}{4}(u_1 - u_2) + \frac{1}{4}\sqrt{(u_1 - u_2)^2 + 8(u_1 + u_2)},$$
$$q_2 = \frac{1}{4}(u_2 - u_1) + \frac{1}{4}\sqrt{(u_1 - u_2)^2 + 8(u_1 + u_2)}.$$

Consequently,

$$q_0 := \frac{1}{2}(q_1 + q_2) \geq \widehat{\theta}, \qquad \theta\left(1 + \frac{1}{4n}(1 - \theta^2)\right) \geq \mathbb{E}(q_0) \geq \mathbb{E}(\widehat{\theta}),$$

where the latter requires $n := n_1 = n_2$.

Proof The first part follows from (7), taking $w = 1$ and summing over the B blocks. The statement for $B = 2$ also follows from (7). The last part follows from Jensen's inequality, $u_i = x_i/n$, $x_i \sim \mathrm{Bi}(n, \theta^2)$, and $\widehat{\theta} = \sqrt{\frac{1}{2}(u_1 + u_2)}$. □

Let $B = \infty$. Assume $x_k \sim \mathrm{Bi}(n, \theta^2)$, $k \geq 1$, is an ergodic process. This is in particular the case if the observations are independent. We have

$$\mathbb{E}(\log(A_k)) = \mathbb{E}(\log(1 - x_k)) < \infty,$$

in fact, the expection is $-\infty$ Hence, the assumptions of Theorem 10 below are fulfilled. In particular, by ergodicity, this implies that θ_t stabilises in distribution for large t, say, to the law of T_∞, independently of the initial conditions, and that the average,

$$\theta(B) = \frac{1}{B}\sum_{k=1}^{B} \theta_k \quad \xrightarrow{\mathcal{D}} \quad T_\infty,$$

converges in distribution as $B \to \infty$. Assuming independent input data (rather than just ergodicity), then further

$$\mathbb{E}(\theta_{t+1}) = 1 - \frac{1}{w}\sum_{i=1}^{w}(1-\theta^2)\mathbb{E}\left(\frac{1}{1+\theta_{t-i+1}}\right) \leq 1 - \frac{1-\theta^2}{w}\sum_{i=1}^{w}\frac{1}{1+\mathbb{E}(\theta_{t-i+1})},$$
$$\mathbb{E}(\theta(\infty)) \leq 1 - (1-\theta^2)\frac{1}{1+E(\theta(\infty))}, \quad \text{or} \quad \mathbb{E}(\theta(\infty)) \leq \theta,$$

where $\mathbb{E}(\theta(\infty)) = \lim_{B\to\infty}\mathbb{E}(\theta(B))$.

Figure 1 shows the result of applying `windowEM` to a small data set consisting of $B = 50$ independent blocks with $n_k = 10$ and a true $\theta = 0.6$. Even with relatively small B, if w is not too large, then convergence of estimates is obtained after one epoch, perhaps two, of iteration. In this case, the standard EM algorithm requires several rounds (approximately 4-5) to converge.

Figure 2 shows the effect of varying the initial estimates and only aiming for one epoch, so $B = \infty$. For both window sizes convergence appears rapidly.

4.2 Univariate Gaussian mixtures

As a special case of a mixture family, consider the univariate Gaussian distribution,

$$p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right), \quad x \in \mathbb{R},$$

which is a two-parameter exponential family distribution with canonical statistics, canonical parameter and mean value parameterisation, respectively,

$$\begin{aligned} T(x) &= (x^2, x) \\ \phi = (\phi_1, \phi_2) &= \Big(-\frac{1}{2\sigma^2}, \frac{\mu}{2\sigma^2}\Big) \\ \tau(\phi) = (\tau_1(\phi), \tau_2(\phi)) &= \Big(\frac{\phi^2}{\phi^2} - \frac{1}{2\phi_1}, -\frac{\phi_1}{\phi_2}\Big). \end{aligned}$$

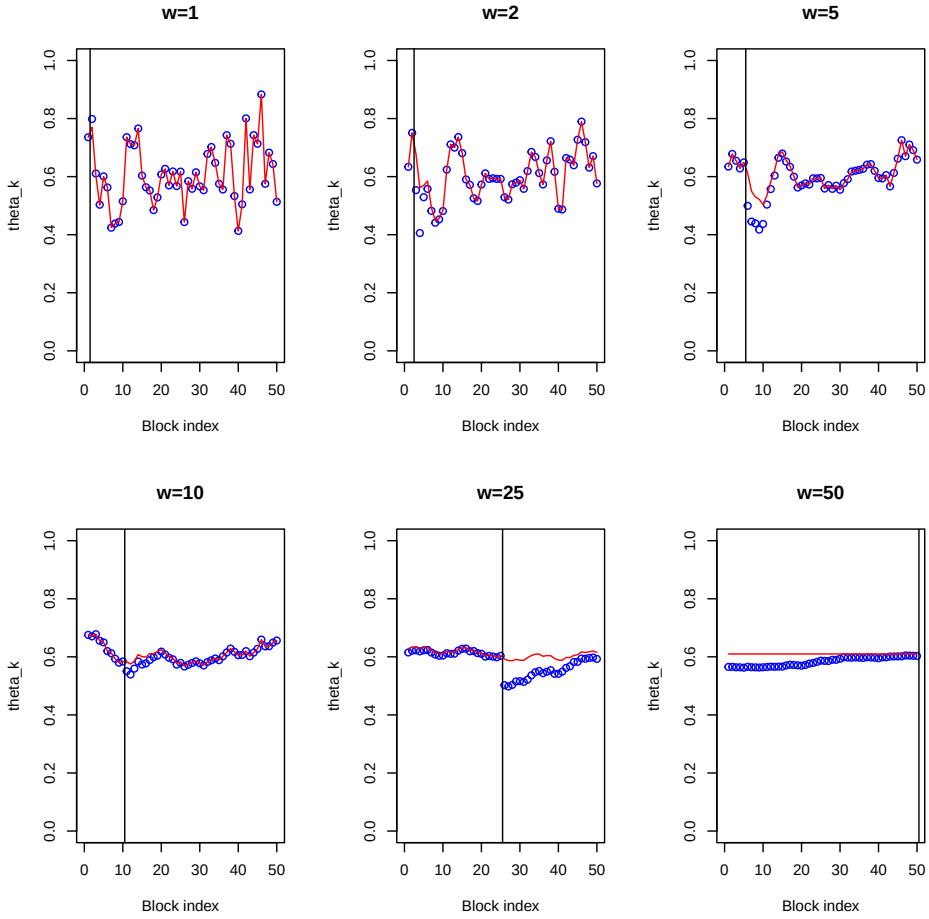


Fig. 1 $B < \infty$. The blue dots show the estimates $\theta_{B+1}, \ldots, \theta_{B+w}, \theta_{w+1}, \ldots, \theta_B$, that is, the estimates after one epoch of iteration. The initial estimates for $\theta_1, \ldots, \theta_w$ are randomly chosen between 0 and 1 (not shown). The red line shows estimates after 100 rounds (converged estimates). Here, $B = 50$ and $w = 1, 2, 5, 10, 25, 50$, and the same input data is used for all plots. The variation in the estimates is higher for small w than large w, while the convergence rate is slower for large w than small w. Compare the red line and the blue dots, which in the case of $w = 25, 50$ shows it takes roughly 25 blocks before convergence is achieved.

Properties of mixtures of univariate Gaussian distributions and the EM algorithm are well-studied [13, 21].

Assume x_{ik}, $i = 1, \ldots, B$, $k = 1, \ldots, n$, are observations from a univariate Gaussian mixture model, divided into B blocks. We find

$$\tau_{1j}^{(t+1)} = \frac{1}{w} \sum_{i=1}^{w} \frac{\sum_{\ell=1}^{n} x_{b(t-i+2)\ell}^2 e_{\ell j}^{(t-i+1)}}{\sum_{\ell=1}^{n} e_{\ell j}^{(t-i+1)}}$$

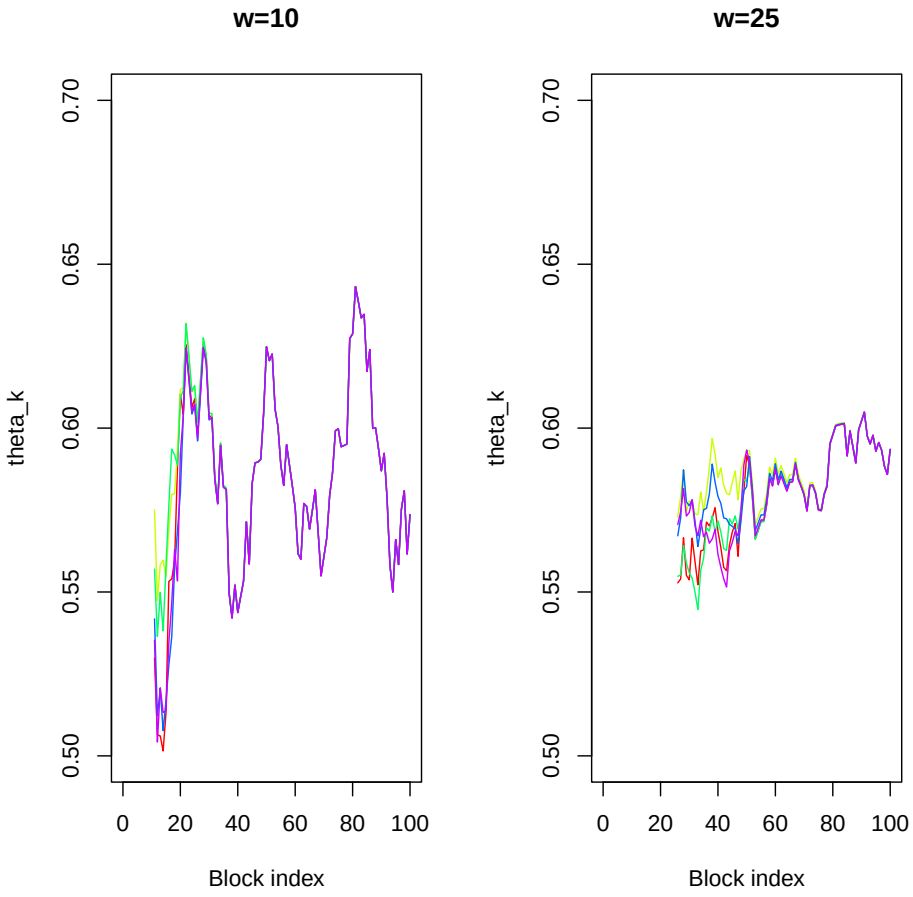


Fig. 2 $B = \infty$. Multiple initiations (here, 5) on the same input data with initial estimates for $\theta_1, \ldots, \theta_w$ randomly chosen between 0 and 1 (not shown). After relatively few iterations the estimates become independent of the random initial estimates.

$$\tau_{2j}^{(t+1)} = \frac{1}{w} \sum_{i=1}^{w} \frac{\sum_{\ell=1}^{n} x_{b(t-i+2)\ell} e_{\ell j}^{(t-i+1)}}{\sum_{\ell=1}^{n} e_{\ell j}^{(t-i+1)}},$$

$$\pi_j^{(t+1)} = \frac{1}{wn} \sum_{i=1}^{w} \sum_{\ell=1}^{n} e_{\ell j}^{(t-i+1)},$$

where $e_{\ell j}^{(t)} = E_{\tau^{(t)}}(I_j(z_\ell)|x)$.

In this case Θ might be taken to be compact subset of $\mathbb{R}^{3k}$, delimitated by the maximal and minimal observations. Hence by the Brouwer fixed point theorem there is a fixed point.

As an illustration, we consider fitting an equal-proportion mixture of ten Gaussian distributions with unit variance and mean values equal to $0, 1, \ldots, 9$. In total, there are $10 + 10 + 9 = 29$ free parameters. We randomly sample $n = 10^5$ data points from the population, and divide these into into $B = 10$ blocks at random, after which we run `windowEM` with a window size of $w = 5$ blocks in both cases. For `windowEM`, one epoch is B iterations and corresponds to processing all data exactly once. After 500 rounds of iteration of standard EM and 500 epochs of `windowEM`, we retrieve the estimates shown in Figure 3. We observe in particular that the estimate resulting from averaging the block estimates achieves a log-likelihood close to the complete data MLE significantly faster than standard EM. We note, however, that the standard EM estimate eventually achieves a log-likelihood that is higher than then `windowEM` estimate, but it also exceeds the complete data MLE. We interpret this finding to suggest that the averaging over a population of estimates induces a measure of regularisation, similar to what was observed for the site frequency spectrum in previous work [16].

For comparison, we show in Figure 4 a smaller dataset ($n = 10^4$) with the same hyperparameter settings, which has been allowed to run for 5000 iterations (resp. epochs). As before, the log-likelihood of the `windowEM` solution stabilises closer to the complete-data MLE faster than the standard EM estimate. The final `windowEM` estimate looks somewhat curious compared to the standard EM solution, but we note again that standard EM overshoots the complete-data log-likelihood, suggesting that the apparent visual benefit is arguably not statistically meaningful.

Finally, the same setup is repeated with a larger dataset $n = 10^6$ (see Figure A1 in Appendix A), where `windowEM` once again converges significantly faster and recovers a more meaningful solution.

To compare hyperparameter settings, we repeated the same experiment at $n = 10^5$, while either keeping $B = 20$ fixed and varying the window size, or keeping the window size fixed at $B/2$ and varying B. The resulting log-likelihood curves are shown in Figure 5. As expected, we observe that as the window size increases towards the B, the convergence behaviour startes to resemble standard EM. Similarly, as the number of blocks decreases (with fixed window size as a proportion), convergence likewise tends to the standard EM behaviour.

5 Convergence

We distinguish between $B < \infty$ and $B = \infty$. In the first case, by iterating over several epochs, the same data is repeatedly being used, whereas in the second case, only one epoch, one iteration is done for each block.

We equip the space Θ^w with a norm

$$\|(\zeta_1, \ldots, \zeta_w)\|_{\max} = \max_{i=1,\ldots,w} \|\zeta_i\|, \quad (\zeta_1, \ldots, \zeta_w) \in \Theta^w,$$

where the norm $\|\cdot\|$ on Θ can be any norm.

Lemma 4 *The space Θ^w equipped with $\|\cdot\|_{\max}$ is a normed space.*

Proof Let $\zeta = (\zeta_1, \ldots, \zeta_w)$ and $\xi = (\xi_1, \ldots, \xi_w)$. If $\|\zeta\|_{\max} = \max_{i=1,\ldots,w} \|\zeta_i\| = 0$, then $\|\zeta_i\| = 0$ for each $i = 1, \ldots, w$, hence $\zeta = 0$. For the triangular inequality,

$$\|\zeta + \xi\|_{\max} = \max_{i=1,\ldots,w} \|\zeta_i + \xi_i\| \leq \max_{i=1,\ldots,w} \|\zeta_i\| + \max_{i=1,\ldots,w} \|\xi_i\| = \|\zeta\|_{\max} + \|\xi\|_{\max}.$$

Finally, for $a \in \mathbb{R}$, $\|a\zeta\|_{\max} = \max_{1,\ldots,w} \|a\zeta_i\| = |a| \, \|\zeta\|_{\max}$. □

5.1 Fixed points for $B < \infty$

Consider the kth window, $k \in \{1, \ldots, B\}$, with block indices $b(k-w+1), \ldots, b(k-1), b(k)$, noting that $b(k) = k$ always. Define the *one-block update* based on the kth window by

$$\sigma_k \colon \Theta^w \to \Theta^w, \quad \sigma_k(\zeta_1, \ldots, \zeta_w) = (\zeta_2, \ldots, \zeta_{w+1}) \in \Theta^w,$$

where

$$\zeta_{w+1} = \frac{1}{w} \sum_{i=1}^{w} f_{b(k-w+i)}(\zeta_i).$$

Similarly, define the *i-block update*, $i \geq 0$, based on the kth window by

$$S_{k,0}(\zeta) = \zeta, \quad S_{k,i}(\zeta) := \sigma_{b(k+i-1)} \circ \ldots \circ \sigma_{b(k+1)} \circ \sigma_k(\xi), \quad i \geq 1,$$

where $\zeta = (\zeta_1, \ldots, \zeta_w) \in \Theta^w$. It holds that

$$S_{k,i+j} = S_{b(k+j),i}(S_{k,j}), \quad i, j \geq 0, \quad 1 \leq k \leq B. \tag{8}$$

Of particular interest is $S_{k,B}$ (update all blocks) and $S_{k,w}$ (update all blocks in one window), starting from block k.

A point $\zeta \in \Theta^w$ is a *fixed point* for the kth window if ζ is a fixed point for the B-step update, that is, if

$$S_{k,B}(\zeta) = \zeta.$$

Lemma 5 *Assume $\zeta \in \Theta^w$ is a fixed point for the kth window, with $k \in \{1, \ldots, B\}$. Then, $S_{k,iB}(\zeta) = \zeta$ for $i \geq 0$. Furthermore, let $\xi = S_{k,j}(\zeta)$ for some $j \geq 0$, then $S_{b(k+j),iB}(\xi) = \xi$ for all $i \geq 0$. As a consequence, all windows have a fixed point if and only if one window has a fixed point.*

Proof For both statements we use (8). For the first, $S_{k,0}(\zeta) = \zeta$ and $S_{k,B}(\zeta) = \zeta$. Then, by induction, $S_{k,(i+1)B}(\zeta) = S_{b(k+B),iB}(S_{k,B}(\zeta)) = S_{k,iB}(S_{k,B}(\zeta)) = S_{k,iB}(\zeta) = \zeta$. The first part is clear. For the second, $S_{b(k+j),iB}(\xi) = S_{b(k+j),iB}(S_{k,j}(\zeta)) = S_{k,iB+j}(\zeta) = S_{k,j}(S_{k,iB}(\zeta)) = S_{k,j}(\zeta) = \xi$. □

Lemma 6 *Let $k \in \{1, \ldots, B\}$, and assume $\zeta \in \Theta^w$ and $S_{k,iB}(\zeta) \to \zeta^* \in \Theta^w$ as $i \to \infty$. Then, $\zeta^* \in \Theta^w$ is a fixed point for the kth window, provided $S_{k,B}$ is continuous in $\zeta^* \in \Theta^w$.*

Moreover, also $S_{\ell,iB}(S_{k,d}(\zeta)) \to S_{k,d}(\zeta^*) \in \Theta$ *with* $d = \ell - k$ *for* $B \geq \ell > k$, *and* $d = B + \ell - k$ *for* $k > \ell \geq 1$, *provided* $S_{k,d}$ *is continuous in* $\zeta \in \Theta^w$. *As a consequence, a converging sequence exists for none or all windows, provided the relevant continuity assumptions hold.*

Proof We make use (8). To prove ζ^* is a fixed point, consider $S_{k,(i+1)B}(\zeta) = S_{k,B}(S_{k,iB}(\zeta)) \to S_{k,B}(\zeta^*)$ for $i \to \infty$, by continuity. Also $S_{k,(i+1)B}(\zeta) \to \zeta^*$ for $i \to \infty$, proving ζ^* is a fixed point. By continuity, then $S_{\ell,iB}(S_{k,d}(\zeta)) = S_{k,d}(S_{k,iB}(\zeta)) \to S_{k,d}(\zeta^*)$. □

If Θ is compact, then the Brouwer fixed point theorem implies the existence of a at least one fixed point, but there might be more fixed points [3]. In general, there are no simple ways to find the fixed point postulated by Brouwer fixed point theorem, in particular, the iterated sequence is not guaranteed to converge. The iterated sequence contains however a converging subsequence due to compactness of Θ.

Lemma 7 *Assume there is a compact convex set* $\Theta_0 \subseteq \Theta$, *such that* $f_k(\Theta) \subseteq \Theta_0$ *for all* $k \in \{1, \ldots, B\}$. *Then, there exists at least one fixed point of* $S_{k,B}$, $k \in \{1, \ldots, B\}$, *in* Θ_0^w *and none in* $(\Theta \setminus \Theta_0)^w$.

Proof The function $S_{k,B}$, $k \in \{1, \ldots, B\}$, is a multi-composition of the functions f_ℓ and and $h_\ell \colon \Theta^w \to \Theta_0$, $\ell \in \{1, \ldots, B\}$, where

$$h_\ell(\zeta_1, \ldots, \zeta_w) = \frac{1}{w} \sum_{i=1}^{w} f_{b(\ell - w + i)}(\zeta_i).$$

If $f_\ell(\Theta) \subseteq \Theta_0$ and Θ_0 is convex, then $h_\ell(\Theta^w) \subseteq \Theta_0$, and the functions restricted to the domain Θ_0 (Θ_0^w) are well-defined. Brouwer fixed point theorem guarantees the existence of at least one fixed point within the domain as it is compact. There cannot be any in $(\Theta \setminus \Theta_0)^w$ as Θ is mapped to Θ_0. □

5.2 Lipschitz condition

A function $F \colon D \to D$, $D \subseteq \mathbb{R}^n$, is *Lipschitz* if there is $\alpha \geq 0$ such that $\|F(x) - F(y)\| \leq \alpha \|x - y\|$ for all $x, y \in D$. For convenience, we assume $\alpha > 0$, since it is always possible to arrange that. Furthermore, the function is *non-expansive* if it is Lipschitz with constant $\alpha \leq 1$, *contractive* (or a *contraction*) if $\alpha < 1$, and *iteratively contractive* if $\|F^n(x) - F^n(y)\| \leq \delta \alpha^n$, $x, y \in D$, $n \geq 1$, for some $\delta > 0$ and $0 < \alpha < 1$. A contraction is iteratively contractive. If D is a complete metric space, then any iteratively contractive function has a unique fixed point, that is, there is a unique $x \in D$ such that $F(x) = x$.

In this section, we study the particular situation, where the functions f_k, $k = 1, \ldots, B$, are all Lipschitz. Let $\theta_{t,w} = (\theta_t, \ldots, \theta_{t-w+1}) \in \Theta^w$ and $\theta'_{t,w} = (\theta'_t, \ldots, \theta'_{t-w+1}) \in \Theta^w$, $t \geq w$.

Theorem 8 *Assume f_k is Lipschitz on Θ with constant $A_k > 0$, such that*

$$\|f_k(\theta) - f_k(\theta')\| \le A_k\|\theta - \theta'\|, \quad \theta, \theta' \in \Theta, \quad k \in \{1, \dots, B\}.$$

Then, for $t \ge 1$,

$$\|\theta_t - \theta'_t\| \le C_t \, \|\theta_{w,w} - \theta'_{w,w}\|_{\max},$$

and for $t \ge w$,

$$\|\theta_{t,w} - \theta'_{t,w}\|_{\max} \le M_t \, \|\theta_{w,w} - \theta'_{w,w}\|_{\max},$$

where $C_1 = \ldots = C_w = 1$, $M_w = 1$, and for $t \ge w$,

$$C_{t+1} = \frac{1}{w} \sum_{k=t-w+1}^{t} A_{b(k+1)} C_k, \quad M_{t+1} = \max\{C_{t+1}, \dots, C_{t-w+2}\}.$$

Proof The two inequalities are clearly fulfilled with $C_1 = \ldots = C_w = 1$ and $M_w = 1$. Assume the first inequality holds for all t', such that $t \ge t' \ge 1$ for some $t \ge w$. Then,

$$\begin{aligned}
\|\theta_{t+1} - \theta'_{t+1}\| &= \left\| \frac{1}{w} \sum_{k=t-w+2}^{t+1} f_{b(k)}(\theta_{k-1}) - f_{b(k)}(\theta'_{k-1}) \right\| \\
&\le \frac{1}{w} \sum_{k=t-w+2}^{t+1} A_{b(k)} \|\theta_{k-1} - \theta'_{k-1}\| \\
&\le \frac{1}{w} \sum_{k=t-w+1}^{t} A_{b(k+1)} C_k \|\theta_{w,w} - \theta'_{w,w}\|_{\max} \\
&= C_{t+1} \|\theta_{w,w} - \theta'_{w,w}\|_{\max}.
\end{aligned}$$

Therefore, for $t \ge w$,

$$\begin{aligned}
\|\theta_{t+1,w} - \theta'_{t+1,w}\|_{\max} &\le \max_{i=0,\dots,w-1} \|\theta_{t+1-i} - \theta'_{t+1-i}\| \\
&\le \max\{C_{t+1}, \dots, C_{t-w+2}\} \|\theta_{w,w} - \theta'_{w,w}\|_{\max} \\
&= M_{t+1} \|\theta_{w,w} - \theta'_{w,w}\|_{\max},
\end{aligned}$$

and the proof is completed. □

The individual constants A_k, $k = 1, \dots, B$, might depend on the data, x_k, but not on the data x_ℓ, $\ell \ne k$, from other blocks. Hence, if x_k, $k \ge 1$, forms a sequence of independent random variables, then one might choose A_k, $k \ge 1$, to be a sequence of independent random variables as well. Contrary to this, M_t depends on $A_t, \dots, A_{t-2w+1}$, so potentially also on $x_t, \dots, x_{t-2w+1}$, that is, $2w$ values.

Theorem 9 *Assume as in Theorem 8. If there exists $i \ge 1$, such that $M_{w+iB} < 1$, then*

$$M_{k+i'B} \le \delta_k \alpha^{i'B}, \quad i' \ge 0,$$

for some $0 < \alpha < 1$, $\delta_k > 0$, and $k = 1, \dots, B$. Consequently, $S_{k,B}$ is iteratively contractive, and $\theta_{k+i'B} \to \theta^*_k$ for $i' \to \infty$ and some $\theta^*_k \in \Theta$.

Proof Since $C_1 = \ldots = C_w = 1$, $M_w = 1$ and $M_{w+iB} < 1$, then by definition of M_k, we have $M_{w+miB} \leq (M_{w+iB})^m \to 0$ as $m \to \infty$. Let $S = \max(1, \sup_{k=1,\ldots,B} D_k)$, where

$$D_k = \frac{1}{w} \sum_{i=0}^{w-1} A_{b(k-i)}, \quad k = 1, \ldots, B,$$

is the average of the w preceding A_is. If $M_{w+miB} = \max\{C_{w+miB}, \ldots, C_{1+miB}\} \leq \epsilon$, then $C_{w+1+miB} \leq \epsilon S$ and by induction $M_{w+j+miB} \leq \epsilon S^j \leq \epsilon S^{iB}$, $j = 0, \ldots, iB-1$. Hence, for $m \geq 0$, with $\epsilon = (M_{w+iB})^m$,

$$\begin{aligned} M_{w+j+miB} \leq (M_{w+iB})^m S^{iB} &= \frac{S^{iB}}{(M_{w+iB})^{j/(iB)}} ((M_{w+iB})^{1/(iB)})^{j+miB} \\ &\leq \frac{S^{iB}}{M_{w+iB}} ((M_{w+iB})^{1/(iB)})^{j+miB}. \end{aligned}$$

Any $k + i'B$ with $i' \in \mathbb{N}_0$, $k = w, w+1, \ldots, B+w-1$, can be written as $j + miB$ by choosing $j \in \{0, 1, \ldots, iB-1\}$ and $m \in \mathbb{N}_0$ appropriately. Hence, this leads to

$$M_{k+i'B} \leq \delta_k \alpha^{i'B}, \;\; i' \geq 0, \quad \delta_k = \frac{S^{iB} \alpha^{k-w}}{M_{w+iB}}, \quad \alpha = (M_{w+iB})^{1/(iB)},$$

for $k = w, w+1, \ldots, B+w-1$, with $\delta_k > 0$ and $0 < \alpha < 1$. By changing the constant δ_k appropriately, one might choose k to run from $k = 1$ to $k = B$. Lemma 5 gives the convergence part and the proof is completed. □

5.3 Convergence for $B = \infty$

In many contexts, the amount of data is huge and we might treat B as being infinite. If so, for practical and theoretical reasons, one might consider whether convergence occurs in some sense after a single round of iteration, or at least to a reasonable degree.

The case $B < \infty$ is quite different from the case $B = \infty$. In the former case, the algorithm repeatedly returns to the same block and convergence of the algorithm depends on the finite data at hand. In contrast, in the case $B = \infty$, each block is visited only once and the probabilistic nature of the (infinite) data becomes important, as we expect to see the data x_k in frequency proportional to its probability.

We will lean on theory for random function iteration, see [7] for an extensive review and [5] for the setting we apply. This setting extends the standard set-up with iid random functions to the non-iid case that fits the non-independence introduced by averaging across windows and using the estimate of the previous block as set-off for the estimate of the current block. Specifically, we will apply a theorem from [5], quoting [9, 10] for its origin, and formulated here according to our needs. Consider

$$\theta_{k+tw,w} = (\theta_{k+tw}, \ldots, \theta_{k+tw-1}, \theta_{k+(t-1)w+1}), \quad t \geq 0,$$

for fixed $k \in \{1, \ldots, w\}$ as a random process depending on the data (and random initial parameter values). Note that $\theta_{k+tw,w}$ is a w-tuple and one time step has the size of one window. In our context, the iterated function is $S_{k,w}$ that depends on the data:

$$(\theta_{k+tw,w}, x_{k+(t+1)w,2w}) \mapsto S_{k,w}(\theta_{k+tw,w}) = \theta_{k+(t+1)w,w},$$

where $x_{k+(t+1)w,2w} = (x_{k+(t+1)w,w}, x_{k+tw,w})$. In the terminology of [5], the first w-tuple of $x_{k+(t+1)w,2w}$ is called the *covariate* process (as function of t) and the second, the *noise* process (as function of t). All coordinates of $x_{k+(t+1)w,2w}$, except its last coordinate, is used to compute $\theta_{k+(t+1)w,w}$ from $\theta_{k+tw,w}$.

In random function iteration theory, one is typically interested in the *forward iteration* and the *backward iteration*, respectively, that is,

$$\begin{aligned} S_{k,tw}(\zeta) &= S_{k+(t-1)w,w} \circ \ldots \circ S_{k+2w,w} \circ S_{k+w,w} \circ S_{k,w}(\zeta), \quad \zeta \in \Theta^w, \quad \text{and} \\ T_{k,tw}(\zeta) &= S_{k,w} \circ S_{k+w,w} \circ S_{k+2w,w} \circ \ldots \circ S_{k+(t-1)w,w}(\zeta), \quad \zeta \in \Theta^w, \end{aligned}$$

respectively. The equality of the forward iteration follows from (8), whereas the equality of the backward iteration is a definition. The forward iteration is the t-block update for the window ending at block k, and the backward iteration is similar, however, here we consider the data laid out differently from right to left. See Table 1 for an illustration.

By laying out the data in the backward iteration way, we obtain from Theorem 10 below (under the conditions of the theorem) that

$$\theta_{k+tw,w} = T_{k,tw}(\theta_{k,w}) \to T_{k,\infty} \quad P\text{-a.s.} \quad \text{for} \quad t \to \infty, \quad T_{k,\infty} \sim T_\infty,$$

for $k = 1, \ldots, w$, where $\theta_{k,w} \in \Theta^w$ are arbitrary initial values. In other words, for large enough t, the estimate $\theta_{k+tw,w}$ stabilises to a value $T_{k,\infty} \in \Theta^w$, which is independent of the initial value. Furthermore, $T_{k,\infty}$ has a distribution that is independent of k. In turn, the sequence θ_k, $k = 1, 2, \ldots$, obeys the law of large numbers,

$$\theta(B) = \frac{1}{B}\sum_{k=1}^{B} \theta_k = \frac{1}{B}\sum_{k=1}^{w} \sum_{i \in A_k(B)} \theta_{k+iw} \to \mathbb{E}(T_\infty), \qquad A_k(B) = \{i \colon k + iw \le B\},$$

as $B \to \infty$. Thus, one might apply $\theta(B)$ as an estimate of the parameter $\theta \in \Theta$. Alternatively, one might estimate the distribution of T_∞ from the empirical distribution of θ_t, and use this as an estimate of θ. The distribution of T_∞ depends on w.

5.4 Theorem 10 and convergence

Assume $S_{k,w}$ is Lipschitz with constant $L_{k,w}$, and that

$$\mathbb{E}(\log(L_{k,w})) < \infty, \quad \mathbb{E}(\log(\|\zeta, S_{k,w}(\zeta)\|_{\max}) < \infty \tag{9}$$

(though both might be $-\infty$) for some $\zeta \in \Theta^w$, where the expectation is over all data realisations. The following theorem is an extract from [5, Theorem 1], see also [9]. In the following ergodic is short for mean ergodic. The Lipschitz constant $L_{k,w}$ might be obtained from those of f_ℓ, $\ell = k, \ldots, k+w-1$, by multiplication $L_{k,w} = A_k \cdots A_{k+w-1}$,

Forward iteration

Block	1	2	3	...	$k-1$	k	$k+1$
Data before adding block $k+1$	x_1	x_2	x_3	...	x_{k-1}	x_k	
Data after adding block $k+1$	x_1	x_2	x_3	...	x_{k-1}	x_k	x_{k+1}

Backward iteration

Block	1	2	3	...	$k-1$	k	$k+1$
Data before adding block $k+1$	x_k	x_{k-1}	x_{k-2}	...	x_2	x_1	
Data after adding block $k+1$	x_{k+1}	x_k	x_{k-1}	...	x_3	x_2	x_1

Table 1 In the perspective of the forward iteration (here illustrated with $w = 1$), we consider blocks and data added to the right, until there are B blocks (and then we let $B \to \infty$). In the perspective of the backward iteration, blocks are still added to the right, but the data is shifted once to the right at each addition and the new data point is added at block 1. We consider convergence of θ_k as $k \to \infty$. For the backward iteration, we have convergence P-a.s. (under regularity conditions) as θ_k and θ_{k+1} differ in data points 'many blocks away' for large k. In contrast, for the forward iteration, the best we can achieve is convergence in distribution.

Theorem 10 *Assume the process x_{k+tw}, $t \geq 0$, is stationary and ergodic, and that (9) is fulfilled. Assume $\mathbb{E}(\log(L_{k,w})) < 0$, that is, $S_{k,w}$ is on average a contraction. Then, for $k = 1, \ldots, w$,*

- *$T_{k,\infty} = \lim_{t\to\infty} T_{k,tw}(\zeta)$ exists P-a.s. and does not depend on $\zeta \in \Theta^w$,*
- *There is a random function T_∞, such that $T_{k,\infty} \sim T_\infty$ and $E(\|T_\infty\|_{\max}) < \infty$,*
- *$S_{k,tw}(\zeta)$ converges in distribution to T_∞ as $t \to \infty$, and $S_{k,w}(T_\infty) \sim T_\infty$,*
- *The process $S_{\ell,tw}(T_\infty)$, $t \geq 0$, is stationary and ergodic.*

Proof If the data $x_{k+tw,w}$, $t \geq 0$, forms a stationary process, then the random functions $S_{k+tw,w}$, $t \geq 0$, have identical distribution, independent of t, and form a stationary process as well. The Lipschitz constant $L_{k,tw}$ of the forward iteration $S_{k,tw}$, $t \geq 1$, is bounded by $L_{k,w}L_{k+w,w}\cdots L_{k+(t-1)w,w}$, hence

$$\begin{aligned}\frac{1}{t}\mathbb{E}(\log(L_{k,tw})) &\leq \frac{1}{t}\mathbb{E}(\log(L_{k,w}L_{k+w,w}\cdots L_{k+(t-1)w,w}))\\ &= \frac{1}{t}\sum_{i=0}^{t-1}\mathbb{E}(\log(L_{k+iw,w})) = \mathbb{E}(\log(L_{k,w})) < 0.\end{aligned}$$

The claim of the theorem is now part of [5, Theorem 1]. □

The condition $(*)$ $\mathbb{E}(A_k) \leq 1-\epsilon$ for all k implies $\mathbb{E}(L_{k,w}) \leq 1-\epsilon$ and by Jensen's inequality that $\mathbb{E}(\log(L_{k,w})) \leq \log(1-\epsilon) < 0$. Hence, the Lipschitz condition of Theorem 10 is fulfilled. The condition $(*)$ itself has strong implications for convergence.

Lemma 11 *Assume as in Lemma 8 and that $\mathbb{E}(A_k) \leq 1-\epsilon$ for all $k \geq 1$ and some $\epsilon > 0$ (but not necessarily $A_k \leq 1$). If A_k, $k \geq 1$, are independent, then for $i \geq 0$ and $1 \leq n \leq w$,*

$$\mathbb{E}(C_{wi+n}) \leq (1-\epsilon)^i, \quad 0 < wi + n \leq w(i+1),$$

and for $\theta_{w,w}, \theta'_{w,w} \in \Theta^w$, we have

$$\mathbb{E}(\|\theta_{t,w} - \theta'_{t,w}\|_{\max}) \leq w(1-\epsilon)^{\lfloor (t-1)/w \rfloor}\, \|\theta_{w,w} - \theta'_{w,w}\|_{\max}.$$

The initial parameter vectors, $\theta_{w,w}$ and $\theta'_{w,w}$, are considered fixed.

Proof According to Lemma 8, we have $C_1 = \ldots = C_w = 1$, and for $t \geq w$,

$$C_{t+1} = \frac{1}{w} \sum_{k=t-w+1}^{t} A_{b(k+1)} C_k.$$

For the induction basis with $i = 0$ and $n = 1, \ldots, w$, $\mathbb{E}(C_n) = 1 \leq (1-\epsilon)^0$. Assume the induction hypothesis holds for all i', n' such that $wi + n \geq wi' + n' > 0$. If $n < w$, then

$$\mathbb{E}(C_{wi+n+1}) = \frac{1}{w} \sum_{k=w(i-1)+n+1}^{wi+n} \mathbb{E}(A_{k+1})\mathbb{E}(C_k) \leq (1-\epsilon)(1-\epsilon)^{i-1} = (1-\epsilon)^i,$$

where independence of A_{k+1} and C_k is used, and that $n + 1 \leq w$. If $n = w$, then

$$\mathbb{E}(C_{wi+n+1}) = \mathbb{E}(C_{w(i+1)+1}) = \frac{1}{w} \sum_{k=wi+1}^{w(i+1)} \mathbb{E}(A_{k+1})\mathbb{E}(C_k) \leq (1-\epsilon)(1-\epsilon)^i = (1-\epsilon)^{i+1}.$$

The last statement is a consequence of Lemma 8 and the first result. □

Acknowledgements. This work was supported by a research grant (VIL78716) from Villum Fonden. We thank Anders Albrechtsen for many fruitful discussions and ideas, and his willingness to share them. We thank Ulises Bercovich Szulmajster for sharing some early ideas.

Competing interests

The authors declare no competing interests.

Contributions

The authors jointly developed the methodology and co-wrote the manuscript.

Appendix A Mixture of Gaussian distributions

See Figure A1.

Appendix B Concavity

A function $F = (F_1, \ldots, F_k) \colon C \to \mathbb{R}^k$ defined on a convex set $C \subseteq \mathbb{R}^n$ is said to be *concave* if each coordinate function F_i, $i = 1, \ldots, k$, is concave. As a consequence, we have

$$F(tx + (1-t)y) \geq tF(x) + (1-t)F(y), \quad x, y \in C, \quad t \in (0,1),$$

where the inequality is taken component-wise.

The function $F \colon C \to \mathbb{R}^k$ is said to be *non-decreasing* if each coordinate function F_i, $i = 1, \ldots, k$, is non-decreasing in each coordinate x_i, $i = 1, \ldots, n$, $x = (x_1, \ldots, x_n)$. Equivalently,

$$\text{if} \quad x \geq y, \quad \text{then} \quad F_i(x) \geq F_i(y), \quad i = 1, \ldots, k, \quad \text{and hence,} \quad F(x) \geq F(y).$$

Lemma 12 *Let $G\colon C \to C'$ and $F\colon C' \to \mathbb{R}^k$ be two functions, where $C \subseteq \mathbb{R}^n$ and $C' \subseteq \mathbb{R}^m$ are convex sets. Consider $H = F \circ G\colon C \to \mathbb{R}^n$. Assume F is concave and non-decreasing, and G is concave. Then, H is concave.*

Proof In principle, it follows from [2, p86]. However, the reference requires the extended function of F to be non-decreasing, which we do not (and are not willing to) assume. Since, in our set-up, C, C' are convex sets, $G(C) \subseteq C'$ and the domain of F is C', we do not encounter the problems of the general set-up.

The proof is adapted from [2, p86]. Assume $x, y \in C$ and $0 \leq a \leq 1$. Since C is convex, we conclude that $ax + (1-a)y \in C$, hence from concavity of G, we have

$$G(ax + (1-a)y) \geq aG(x) + (1-a)G(y), \tag{B1}$$

where the inequality is component-wise. Since $G(x), G(y) \in C'$ by assumption, we conclude that $aG(x) + (1-a)G(y) \in C'$ by convexity of C'. Also, $G(ax + (1-a)y) \in C'$ (in the general set-up, non-decreasingness of the extended function is used to conclude the same).

Using the above conclusions, the fact that F is non-decreasing, and (B1), we get

$$\begin{aligned} F \circ G(ax + (1-a)x_2) &= F(G(ax_1 + (1-a)y)) \\ &\geq F(aG(x) + (1-a)G(y)) \\ &\geq aF(G(x)) + (1-a)F(G(y)) \\ &= aF \circ G(x) + (1-a)F \circ G(y) \end{aligned}$$

where concavity of F is used to derive the last inequality. □

References

[1] O. Barndorff-Nielsen. *Information and Exponential Families.* John Wiley & Sons, 1978.

[2] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2009.

[3] L. E. J. Brouwer. Über abbildungen von mannigfaltigkeiten. *Mathematische Annalen*, 71:97–115, 1911.

[4] O. Cappé and E. Moulines. Online Expectation-Maximization algorithm for latent data models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(3):593–613, 2009.

[5] Z. M. Debaly and L. Truquet. Iterations of dependent random maps and exogeneity in nonlinear dynamics. *Econometric Theory*, 37:1135–1172, 2021.

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

[7] P. Diaconis and D. Freedman. Iterated random functions. *SIAM Review*, 41(1):45–76, 1999.

[8] B. Efron. *Exponential Families in Theory and Practice*. Cambridge University Press, 2022.

[9] J. H. Elton. A multiplicative ergodic theorem for Lipschitz maps. *Stochastic Processes and their Applications*, 34:39–47, 1990.

[10] M. Iosifescu. Iterated function systems. a critical survey. *Math. Rep., Bucur.*, 11(61):181–229, 2009.

[11] J.-F. Le Gall. *Measure Theory, Probability, and Stochastic Pprocesses*. Springer, 2022.

[12] Z. Li, J. Meisner, and A. Albrechtsen. Fast and accurate out-of-core pca framework for large scale biobank data. *Genome Research*, 33(9):1599–1608, 2023.

[13] J. Ma, L. Xu, and M. I. Jordan. Asymptotic convergence rate of the EM algorithm for Gaussian mixtures. *Neural Computation*, 12:2881–2907, 2000.

[14] R. M. Neil and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse and other variants. In *Jordan, M.I. (eds) Learning in Graphical Models. NATO ASI Series, vol 89*. Springer, Dordrecht, 1988.

[15] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

[16] M. S. Rasmussen, G. Garcia-Erill, T. S. Korneliussen, C. Wiuf, and A. Albrechtsen. Estimation of site frequency spectra from low-coverage sequencing data using stochastic em reduces overfitting, runtime, and memory usage. *Genetics*, 222(4), 2022.

[17] D. Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. *Technical report, Cornell University, Operation Research and Industrial Engineering*, 1988.

[18] R. Sundberg. An iterative method for solution of the likelihood equations for incomplete data from exponential families. *Communications in Statistics*, 5(1):55–64, 1976.

[19] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *J. R. Stat. Soc. B*, 61:611–622, 1999.

[20] T. T. Wu and K. Lange. The MM alternative to EM. *Statistical Science*, 4:494–505, 2010.

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

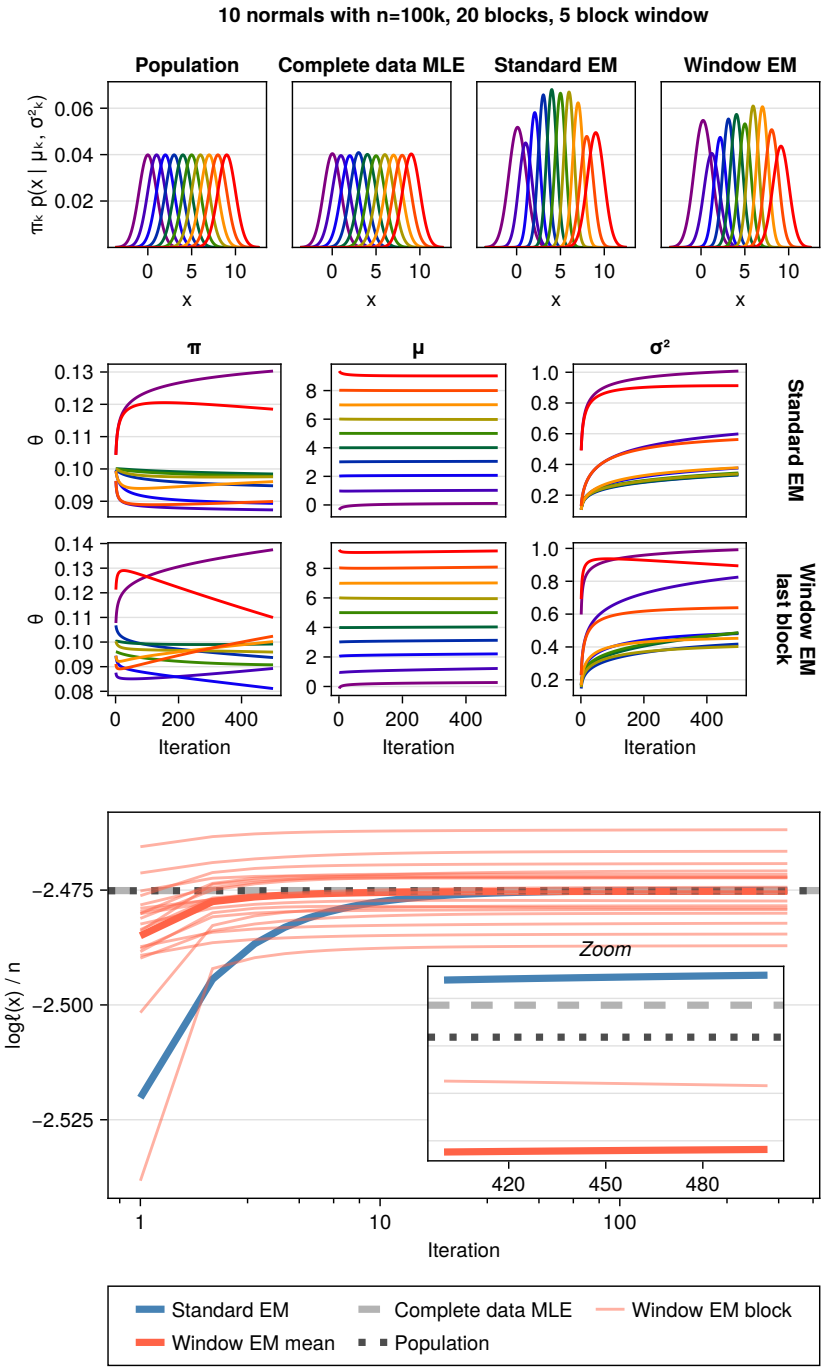


Fig. 3 Comparison of fitting a mixture of 10 unit variance Gaussian distributions with standard EM and `windowEM` from 100000 randomly sampled data points. Both algorithms are run for 500 full passes over the data (corresponding to iterations for standard EM, or epochs for `windowEM`). The final estimates are shown together with the population and complete data MLE at the top. In the middle panel, the evolution of individual parameter estimates are shown, while the bottom shows the log-likelihood over iterations (resp. epochs). The individual block estimates are shown, together with their average taken at the end of each epoch.

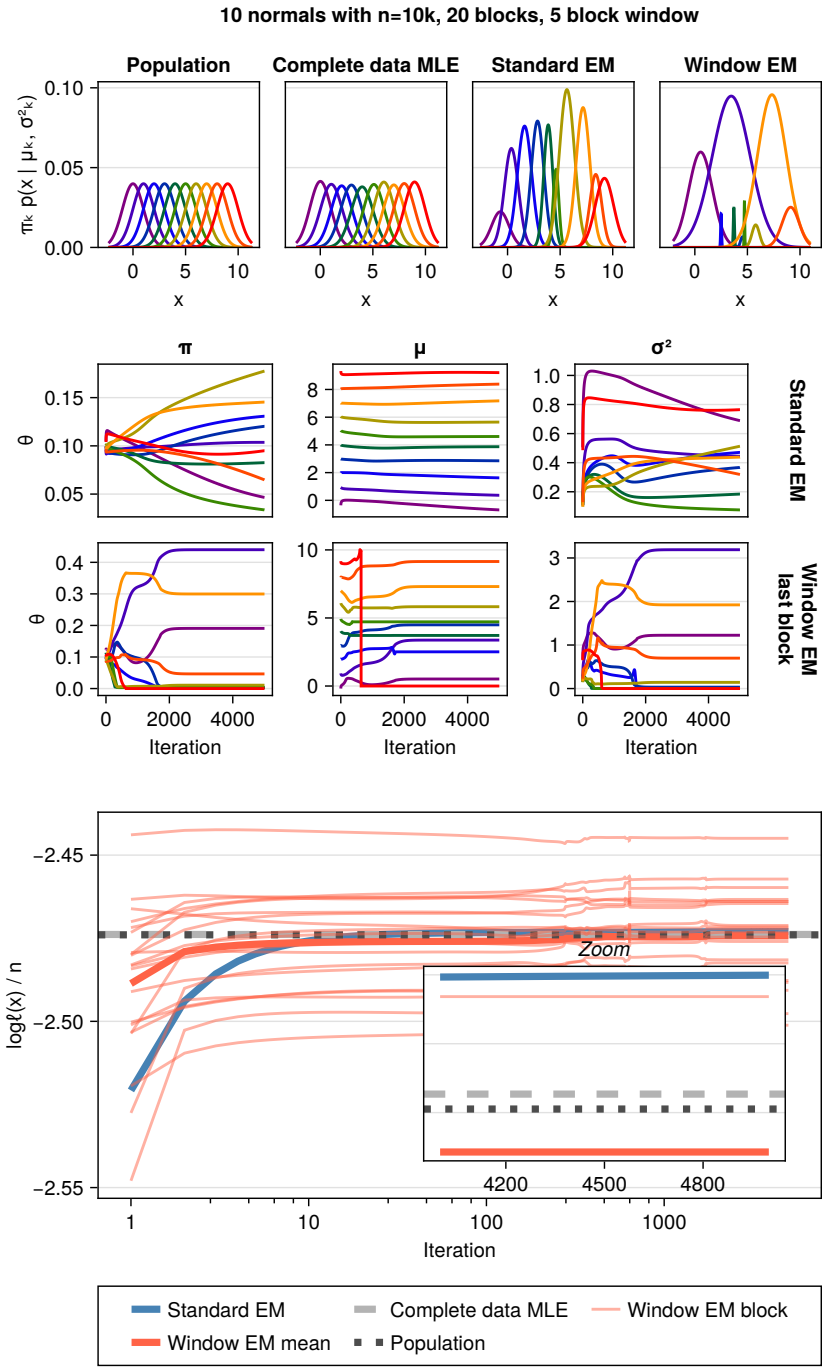
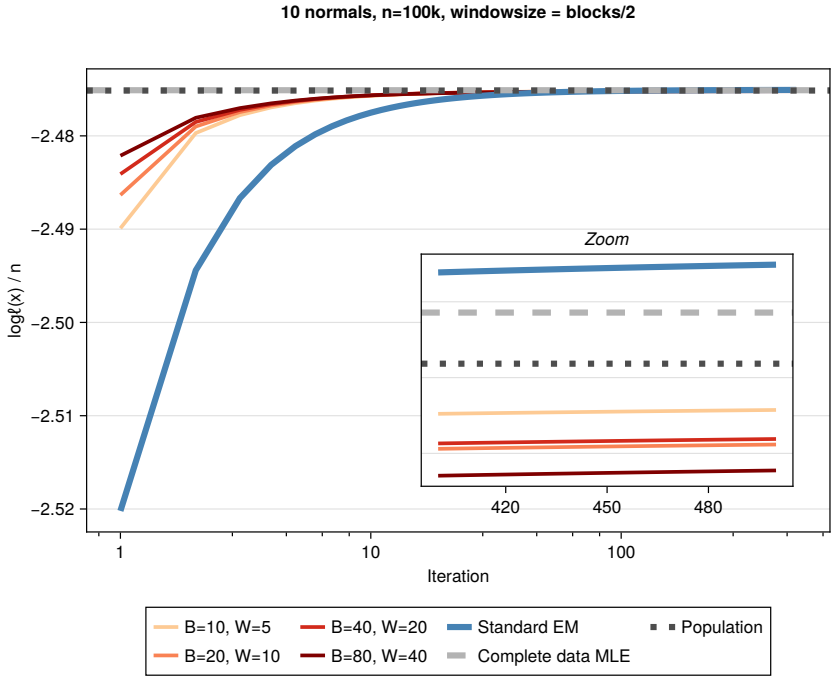


Fig. 4 The same set-up as Figure 3, but fitting $n = 10^4$ data points.



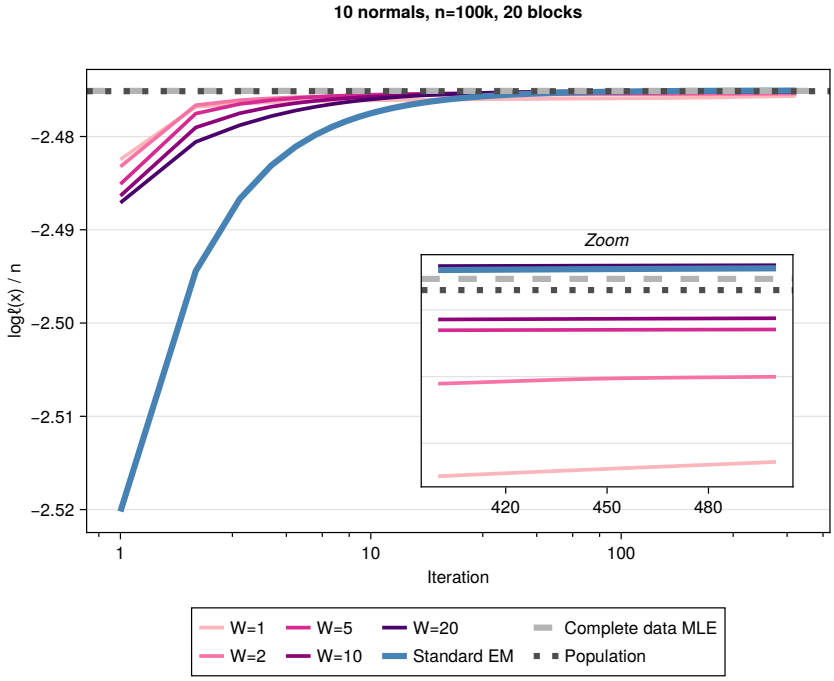


Fig. 5 Effects of varying window size (for fixed number of blocks) or number of blocks (with window size a fixed proportion of blocks) when fitting the same scenario as in Figure 3.

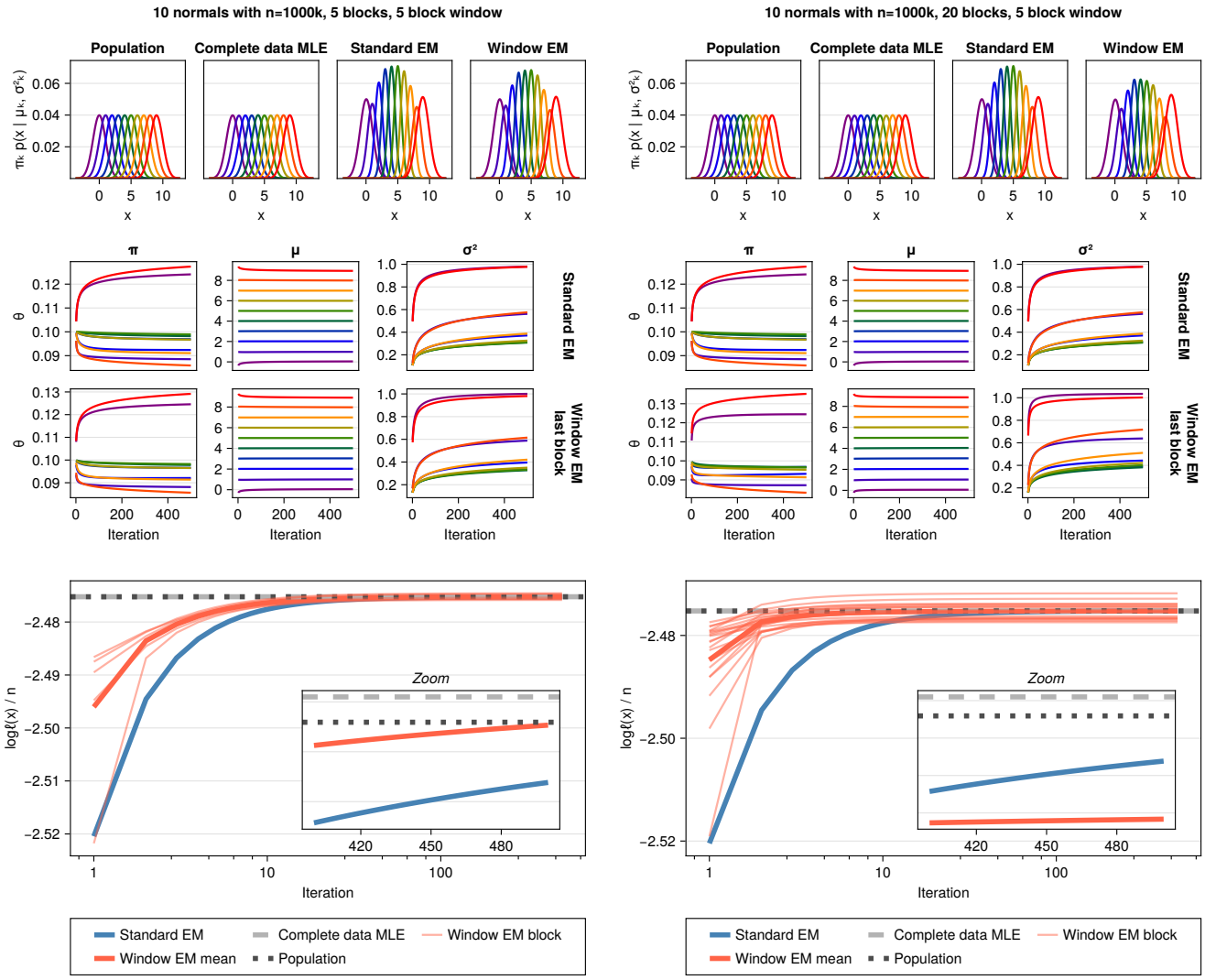


Fig. A1 The same set-up as Figure 3 and 4, but fitting $n = 10^6$ data points.