# An Online Minorization-Maximization Algorithm: Theory and Applications

**TrungKhang Tran**[⋆1], **TrungTin Nguyen**[⋆2], **Tung Doan**[1]

**Binh T. Nguyen**[1], **Hien Duy Nguyen**[3,4], **Florence Forbes**[5], **Gersende Fort**[6]

[1]University of Science - VNUHCM, Ho Chi Minh City, Vietnam;
[2]School of Mathematical Sciences, Queensland University of Technology, Brisbane, Australia;
[3]Department of Mathematics and Physical Science, La Trobe University, Melbourne Australia;
[4]Institute of Mathematics for Industry, Kyushu University, Fukuoka, Japan;
[5]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000, Grenoble, France;
[6]Institut de Mathématiques de Toulouse, CNRS, Toulouse, France.

## Abstract

Processing high-volume and streaming data in modern statistical and machine learning applications often renders batch-mode algorithms impractical due to their requirement to access the entire dataset at each iteration. This limitation has spurred significant interest in the development of efficient online estimation algorithms. The online Expectation–Maximization (EM) algorithm extends the widely-used EM algorithm to accommodate such scenarios by employing a stochastic approximation framework. In this work, we study an online variant of the Minorization–Maximization (MM) algorithm, which generalizes the online EM algorithm as a special case while offering more advantageous features by relaxing restrictive assumptions, such as convexity and latent variable stochastic representations. We provide new robust theoretical guarantees, including consistency and asymptotic normality. Notably, the proposed algorithm achieves convergence to a stationary point–where the gradient of the objective function vanishes–at an optimal rate, matching that of the maximum likelihood estimator. To showcase the practicality and effectiveness of our method, we apply it to the softmax-gated Gaussian mixture of experts regression problem, where the online EM algorithm fails in this context. This MoE model has recently emerged as a pivotal element in cutting-edge deep neural networks for heterogeneous data analysis, underpinning a broad spectrum of applications in machine learning and statistics.

**Keywords:** Mixture of experts, minorization-maximization, expectation-maximization, parameter estimation, online algorithms, stochastic approximation.

## Contents



⋆Equal contribution.

# 1 Introduction

## 1.1 Mixture of experts models

Mixture of Experts (MoE) models, originally introduced by [39, 44], are extensively utilized to capture intricate non-linear relationships between input and output variables in heterogeneous datasets. These models effectively address the dual challenges of regression and clustering by decomposing the predictive framework into a combination of gating models and expert models, both of which depend on the input variables. As a notable example of conditional computation [2, 15], MoE models allocate distinct experts to specific regions within the input space. This architecture allows MoE models to substantially enhance model capacity while maintaining nearly constant training and inference costs by leveraging only a subset of parameters for each example.

Moreover, MoE models have gained prominence due to their universal approximation capabilities and nearly optimal estimation rates in various contexts. These include mixture models [31, 95, 85, 36, 37, 84, 83, 16], mixture of regression models [25, 38], and more generally, mixture of experts models [42, 81, 86, 78, 75, 80, 71, 70, 72]. For a comprehensive overview of MoE models and their applications, we direct readers to reviews such as [103, 63, 74, 82].

## 1.2 Minorization-Maximization algorithms

The Expectation-Maximization (EM) [24, 65, 64] algorithm has become a fundamental tool in computational statistics, boasting a wide range of statistical and signal processing applications. Its significance was quickly recognized and embraced by the international statistical community. In contrast, the body of literature on the more general Minorization-Maximization (MM) algorithm [87, 20, 21, 53, 101, 73] has been traditionally been smaller. A key strength of these algorithms is their ability to leverage computationally efficient surrogate functions, which effectively replace challenging optimization objectives, thereby simplifying the computational iterations. While the MM principlecan be recognized in as early as the work of [87] in the context of line search methods, its first statistical application appeared in [20] and [21], in the context of multidimensional scaling. The limited recognition of these pioneering papers arguably delayed the broader adoption and development of MM algorithms within computational statistics. They are well-established optimization frameworks that play a pivotal role in the development of estimation methodologies for a wide range of data analysis models. While these frameworks are commonly applied to finite mixture models, their application to the more general MoE models, such as softmax-gated Gaussian MoE models [39, 44], remains limited.

With the increasing volume of data and the streaming nature of data acquisition, significant advancements have been made in the development of online and mini-batch algorithms. These approaches enable model estimation without requiring simultaneous access to the entire dataset. Online and mini-batch extensions of EM algorithms can be derived within the classical Stochastic Approximation framework (see, e.g., [6, 49]). Numerical evaluations have demonstrated the effectiveness of these methods in tackling estimation problems within mixture models and machine learning, with notable examples presented in [9, 8, 30, 77, 48, 46, 19, 45, 88, 89]. In contrast, online and mini-batch adaptations of MM algorithms have primarily been developed using techniques from online learning and convex optimization (see, *e.g.,* [17, 34, 41, 51, 59, 96, 99]), with illustrative examples outlined in [62, 67, 76, 18, 60, 61, 57, 55, 56, 29]. These methods have similarly proven valuable for addressing the challenges of optimizing complex models in streaming data environments.

### 1.3 Contributions and practical implications

In this paper, we study and analyse a stochastic approximation-based formulation of an online MM algorithm introduced in [76] and developed within the framework proposed by [9]. This approach adheres more closely to the fundamental principles of the original batch-mode EM algorithm, ensuring consistency with its theoretical and methodological foundations. A key advantage of our approach is that we do not impose convexity assumptions, instead replacing them with oracle assumptions concerning the surrogates as in Assumption 1. In contrast to the online EM algorithm outlined in [9, 8], upon which this work is based, the online MM algorithm [76] broadens the scope by accommodating surrogate functions that do not require latent variable stochastic representations. This feature is particularly advantageous when developing estimation algorithms for MoE models (see, *e.g.,* [79]). Additionally, the proposed approach offers theoretical guarantees, including consistency and asymptotic normality. Notably, the algorithm produces a consistent and efficient estimator in the sense that it outputs a sequence that converges to a stationary point–where the gradient of the objective function vanishes–at an optimal rate comparable to that of the maximum likelihood estimator, see more details in Section 3.

To demonstrate the practicality and effectiveness of our method, we apply it to the softmax-gated Gaussian MoE regression problem in Sections 4 and 5, a context in which the online EM algorithm, as proposed in *e.g.,* [9, 8, 77], fails to be applicable. The proposed stochastic algorithms not only encompass but also extend various existing online EM algorithmic frameworks for fitting softmax-gated Gaussian MoE models. Moreover, they can be adapted to develop feasible and practical algorithms for more complex MoE models, such as tensor-variate mixtures of experts [40]. This is particularly significant as the MoE model has recently become a cornerstone in state-of-the-art deep neural networks [100, 27] for heterogeneous data analysis, supporting a wide range of applications, including speech recognition [102], natural language processing [26, 91, 94], computer vision [54], medicine [58, 7], remote sensing in planetary science [28, 47, 22], bioinformatics [3, 69], and the physical sciences [50].

**Notation.** Throughout the paper, the set $\{1, 2, \ldots, N\}$ is abbreviated as $[N]$ for any positive integer $N \in \mathbb{N}$, and $\mathbf{1}_N$ denotes the $N$-dimensional vector of ones. In accordance with standard conventions, $\boldsymbol{A}^\top$ represents the transpose of a given vector or matrix $\boldsymbol{A}$, and all vectors are expressed as column vectors. The Loewner order between two matrices is denoted by $\boldsymbol{A} \succ \boldsymbol{B}$, implying that the matrix $\boldsymbol{A} - \boldsymbol{B}$ is positive semi-definite. Additionally, the symbol $\boldsymbol{A} \otimes \boldsymbol{B}$ refers to the Kronecker product (also known as the matrix direct product) of the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. Let $\langle \cdot, \cdot \rangle$ denote the Euclidean scalar product. Given any differentiable function $f = (f_1, \ldots, f_M)^\top$ from $\mathbb{R}^T$ to $\mathbb{R}^M$, we denote by $\nabla_{\boldsymbol{\theta}} f = (\partial f / \partial \boldsymbol{\theta}_1, \ldots, \partial f / \partial \boldsymbol{\theta}_T)^\top$ the gradient of $f$ when $M = 1$ and denote by $\nabla_{\boldsymbol{\theta}} f^\top$ the $T \times M$ matrix whose columns are the gradients, that is, $\nabla_{\boldsymbol{\theta}} f^\top = (\nabla_{\boldsymbol{\theta}} f_1, \ldots, \nabla_{\boldsymbol{\theta}} f_M)$. Accordingly, the symbol $\nabla_{\boldsymbol{\theta}} f^\top$ is to be understood as either a vector or a matrix, depending on whether the function $f$ is scalar or vector-valued. In the latter case, the usual Jacobian matrix is the transpose of $\nabla_{\boldsymbol{\theta}} f^\top$. When $f$ is twice differentiable, we use the notation $\nabla^2_{\boldsymbol{\theta}} f$ as the Hessian matrix which is a $T \times T$ matrix with components given by $[\nabla^2_{\boldsymbol{\theta}} f]_{i,j} = \partial^2 f / \partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j, i, j \in [T]$. For any $s \in \mathbb{R}^D$ and $\mathbb{S} \subset \mathbb{R}^D$, let $d(s, \mathbb{S}) = \inf_{r \in \mathbb{S}} |s - r|$.

**Paper organization.** The remainder of this paper is organized as follows. In Section 2, we provide an overview of the online MM algorithm, setting the foundation for subsequent discussions. In Section 3, we establish the consistency and convergence rates of the online MM algorithm, which are further substantiated through empirical evaluations. Applications to softmax-gated Gaussian MoE models are presented in Section 4, followed by an experimental study in Section 5 to validate the theoretical findings. Finally, all proofs of theoretical results are deferred to the supplementary material in Appendices A to C.

## 2 The online MM algorithm

### 2.1 Primitive optimization problem

We begin by examining the following primitive optimization problem in a statistical framework:

$$\arg\max_{\boldsymbol{\theta} \in \mathbb{T}} \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_\pi} \left[ f\left(\boldsymbol{\theta}; \mathbf{x}\right) \right] \equiv \arg\max_{\boldsymbol{\theta} \in \mathbb{T}} \mathbb{E}_{\pi} \left[ f\left(\boldsymbol{\theta}; \mathbf{x}\right) \right]. \tag{1}$$

In this context, the set $\mathbb{T}$ is defined as a measurable open subset of $\mathbb{R}^T$, and $\mathbb{X} \subset \mathbb{R}^D$ is a topological space endowed with its corresponding Borel sigma-algebra. The function $f : \mathbb{T} \times \mathbb{X} \to \mathbb{R}$ is measurable and represents a parametric family indexed by the parameter $\boldsymbol{\theta} \in \mathbb{T}$. Additionally, $\mathbf{x}$ is an $\mathbb{X}$-valued random variable on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with its probability density function denoted by $\pi$. The symbols $\mathbb{P}_\pi$ and $\mathbb{E}_\pi$ are used to represent the probability measure and the expected value, respectively, under the distribution of $\mathbf{x}$. In this paper, we focus on the scenario where the expectation $\mathbb{E}_\pi\left[f\left(\boldsymbol{\theta}; \mathbf{x}\right)\right]$ cannot be expressed in closed form, and the corresponding optimization problem is addressed using an MM-based algorithm.

### 2.2 Exponential family minorizer surrogate assumptions

According to the terminology introduced in [53], the function $g : (\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{\tau}) \mapsto g\left(\boldsymbol{\theta}, \boldsymbol{x}; \boldsymbol{\tau}\right)$, defined over the domain $\mathbb{T} \times \mathbb{X} \times \mathbb{T}$, is referred to as a *minorizer of* $f$ if, for any $\boldsymbol{\tau} \in \mathbb{T}$ and any $(\boldsymbol{\theta}, \boldsymbol{x}) \in \mathbb{T} \times \mathbb{X}$, the following conditions are satisfied:

$$f(\boldsymbol{\theta}; \boldsymbol{x}) - f(\boldsymbol{\tau}; \boldsymbol{x}) \geq g(\boldsymbol{\theta}, \boldsymbol{x}; \boldsymbol{\tau}) - g(\boldsymbol{\tau}, \boldsymbol{x}; \boldsymbol{\tau}) \text{ and } f(\boldsymbol{\tau}; \boldsymbol{x}) = g(\boldsymbol{\tau}, \boldsymbol{x}; \boldsymbol{\tau}). \tag{2}$$

We study the scenario in which the minorizer function $g$ satisfies the following conditions:

**Assumption 1.**

*A1 The minorizer surrogate $g$ belongs to an exponential family:*

$$g\left(\boldsymbol{\theta}, \boldsymbol{x}; \boldsymbol{\tau}\right) := -\psi\left(\boldsymbol{\theta}\right) + \left\langle \bar{S}(\boldsymbol{\tau}; \boldsymbol{x}), \phi(\boldsymbol{\theta}) \right\rangle, \tag{3}$$

*where $\psi : \mathbb{T} \to \mathbb{R}$, $\phi : \mathbb{T} \to \mathbb{R}^D$ and $\bar{S} : \mathbb{T} \times \mathbb{X} \to \mathbb{R}^D$ are measurable functions. In addition, $\phi$ and $\psi$ are continuously differentiable on $\mathbb{T}$.*

*A2 There exists a measurable, open, and convex set $\mathbb{S} \subseteq \mathbb{R}^D$ such that for any $s \in \mathbb{S}$, $\gamma \in [0, 1)$, and any $(\boldsymbol{\tau}, \boldsymbol{x}) \in \mathbb{T} \times \mathbb{X}$, the following condition holds:*

$$s + \gamma\left\{\bar{S}(\boldsymbol{\tau}; \boldsymbol{x}) - s\right\} \in \mathbb{S}.$$

*A3 The expectation $\mathbb{E}_\pi[\bar{S}(\boldsymbol{\theta}; \mathbf{x})]$ exists, lies within $\mathbb{S}$, and is finite for any $\boldsymbol{\theta} \in \mathbb{T}$, although it may not have a closed-form expression. Additionally, an online sequence of independent oracles $\{\mathbf{x}_n, n \geq 0\}$, distributed identically to $\mathbf{x}$, is available.*

*A4 For any $s \in \mathbb{S}$, there exists a unique root to the function $\boldsymbol{\theta} \mapsto -\nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta})^\top s$. This root corresponds to the unique global maximum of the function $\boldsymbol{\theta} \mapsto h(s; \boldsymbol{\theta}) := -\psi(\boldsymbol{\theta}) + \langle s, \phi(\boldsymbol{\theta})\rangle$ over $\mathbb{T}$. The root is denoted by $\bar{\boldsymbol{\theta}}(s)$, meaning:*

$$\bar{\boldsymbol{\theta}}(s) := \arg\max_{\boldsymbol{\theta} \in \mathbb{T}}[-\psi(\boldsymbol{\theta}) + \langle s, \phi(\boldsymbol{\theta})\rangle], \quad -\nabla_{\boldsymbol{\theta}}\psi(\bar{\boldsymbol{\theta}}(s)) + \nabla_{\boldsymbol{\theta}}\phi(\bar{\boldsymbol{\theta}}(s))^\top s = 0. \tag{4}$$

**Remark 1.** *Note that, viewed as a function of $\boldsymbol{\theta}$, $g(\cdot, \boldsymbol{x}; \boldsymbol{\tau})$ can be expressed as the sum of two components: $-\psi$ and a linear combination of the elements of $\phi = (\phi_1, \ldots, \phi_D)$. Assumption 1.A1 ensures that the minorizer surrogate resides within the functional space generated by these $(d + 1)$ basis functions. From Equation (2) and Assumption 1.A1–A3, the following inequality holds:*

$$\mathbb{E}_\pi\left[f(\boldsymbol{\theta}; \mathbf{x})\right] - \mathbb{E}_\pi\left[f(\boldsymbol{\tau}; \mathbf{x})\right] \geq \psi(\boldsymbol{\tau}) - \psi(\boldsymbol{\theta}) + \left\langle \mathbb{E}_\pi\left[\bar{S}(\boldsymbol{\tau}; \mathbf{x})\right], \phi(\boldsymbol{\theta}) - \phi(\boldsymbol{\tau}) \right\rangle,$$

*thereby establishing $g$ as a valid minorizer function for the objective function $\boldsymbol{\theta} \mapsto \mathbb{E}_\pi\left[f(\boldsymbol{\theta}; \mathbf{x})\right]$.*

Given that the expectation in Equation (1) may lack a closed form, but infinitely large datasets are available (as stated in Assumption 1.A3), we consider the `Online MM` algorithm introduced in [76] and described in Algorithm 1. This algorithm defines the sequence $\{s_n, n \geq 0\}$, where the update mechanism, detailed in Equation (5), is a stochastic approximation iteration that generates an $\mathbb{S}$-valued sequence (see Assumption 1.A2). The approach involves constructing a sequence of minorizer functions by defining their parameter $s_n$ within the functional space spanned by $-\psi, \phi_1, \ldots, \phi_D$.

## 3 Convergence properties of the online MM algorithm

In this section, we investigate the convergence properties of the proposed Algorithm 1, demonstrating its convergence to the set of stationary points of the primitive optimization function in Equation (1).

**Algorithm 1** Online MM algorithm

**Require:** An initial value $s_0 \in \mathbb{S}$ and positive step sizes $\{\gamma_{n+1}, n \geq 1\}$ in $(0, 1)$.
**Ensure:** A $\mathbb{T}$-valued sequence $\{\boldsymbol{\theta}_n, n \geq 0\}$ and a $\mathbb{S}$-valued sequence $\{s_n, n \geq 0\}$.
1: **for** $n = 0, 1, \cdots,$ **do**
2: Compute

$$s_{n+1} = s_n + \gamma_{n+1} \left\{ \bar{S} \left( \bar{\boldsymbol{\theta}}(s_n); \boldsymbol{x}_{n+1} \right) - s_n \right\}, \tag{5}$$

$$\boldsymbol{\theta}_{n+1} = \bar{\boldsymbol{\theta}}(s_{n+1}) := \arg\max_{\boldsymbol{\theta} \in \mathbb{T}} g\left(\boldsymbol{\theta}, \boldsymbol{x}_{n+1}; s_{n+1}\right). \tag{6}$$

3: **end for**

Proposition 1 was initially proposed and proved as Lemma 1 in [76]. For the sake of completeness and to maintain consistency with the notations used in our paper, we have also included the proof of Proposition 1 in Appendix A.1. Additionally, while the authors of [76] conjectured Proposition 3 without providing formal statements or proofs, we present these results formally, supported by rigorous mathematical proofs in Appendix A.3. Beyond establishing consistency, we also derive the convergence rate of the proposed Algorithm 1 under mild regularity conditions, demonstrating $\gamma_n^{-1/2}$-consistency and asymptotic normality. The proof primarily relies on the weak convergence result in Theorem 1 of [90], which was similarly employed in [9] to establish the convergence rate of the online EM algorithm.

### 3.1 Limiting points and stationary points of the original optimization problem

If Algorithm 1 converges, any limiting point $s^0$ satisfies $\mathbb{E}_\pi \left[ \bar{S}(\bar{\boldsymbol{\theta}}(s^0); \mathbf{x}) \right] = s^0$. Thus, Algorithm 1 is formulated to approximate the intractable expectation at $\bar{\boldsymbol{\theta}}(s^0)$, where $s^0$ adheres to this fixed-point condition. Furthermore, Proposition 1, which is proved in Appendix A.1 elucidates the connection between the limiting points of Equation (5) and the optimization problem defined in Equation (1). Notably, it demonstrates that any limiting value $s^0$ corresponds to a stationary point $\boldsymbol{\theta}^0 := \bar{\boldsymbol{\theta}}(s^0)$ of the objective function $\mathbb{E}_\pi \left[ f\left(\boldsymbol{\theta}; \mathbf{x}\right) \right]$, indicating that $\boldsymbol{\theta}^0$ is a root of its derivative. The proof leverages the methodology presented in [9]. We commence by defining the following notations:

$$\eta(s) := \mathbb{E}_\pi \left[ \bar{S} \left( \bar{\boldsymbol{\theta}} \left( s \right); \mathbf{x} \right) \right] - s, \qquad \mathbb{F} := \{ s \in \mathbb{S} : \eta(s) = 0 \}. \tag{7}$$

**Proposition 1.** *Suppose that the function $\boldsymbol{\theta} \mapsto \mathbb{E}_\pi \left[ f(\boldsymbol{\theta}; \mathbf{x}) \right]$ is continuously differentiable on $\mathbb{T}$, and let $\mathbb{L}$ represent the set of stationary points of this function, defined as $\mathbb{L} = \{ \boldsymbol{\theta} \in \mathbb{T} : \nabla_{\boldsymbol{\theta}} \mathbb{E}_\pi \left[ f(\boldsymbol{\theta}; \mathbf{x}) \right] = 0 \}$. If $s^0 \in \mathbb{F}$, then $\bar{\boldsymbol{\theta}}(s^0) \in \mathbb{L}$. Conversely, if $\boldsymbol{\theta}^0 \in \mathbb{L}$, it follows that $s^0 := \mathbb{E}_\pi \left[ \bar{S} \left( \boldsymbol{\theta}^0; \mathbf{x} \right) \right] \in \mathbb{F}$.*

**Additional regularity assumptions for Lyapunov function.** Leveraging the results from [23] on the asymptotic convergence of stochastic approximation algorithms, along with the additional regularity condition specified in Assumption 2 for $\psi$, $\phi$, and $\bar{\boldsymbol{\theta}}$, Proposition 2 (whose proof is detailed in Appendix A.2) establishes that the algorithm described in Equation (5) admits a continuously differentiable Lyapunov function $V$, defined on $\mathbb{S}$ as

$$s \mapsto -\mathbb{E}_\pi \left[ f(\bar{\boldsymbol{\theta}}(s); \mathbf{x}) \right] =: V(s),$$

which satisfies $\langle \nabla_{\boldsymbol{\theta}} V(s), \eta(s) \rangle \leq 0$, with strict inequality holding outside the set $\mathbb{F}$ (refer to [9, Prop. 2]).

Furthermore, the additional regularity assumptions detailed below must also be satisfied, in addition to those set out in Assumption 1.

**Assumption 2.**

*A5 The parameter space $\mathbb{T}$ is defined as a convex and open subset of $\mathbb{R}^D$, with the functions $\psi$ and $\phi$ in Equation (3) assumed to be twice continuously differentiable throughout $\mathbb{T}$.*

*A6 The global maximum function $s \mapsto \bar{\boldsymbol{\theta}}(s)$, defined in Equation (4), is continuously differentiable on $\mathbb{T}$.*

*A7 For all compact subsets $\mathbb{K} \subset \mathbb{S}$ and some $p > 2$, it holds that*

$$\sup_{s \in \mathbb{K}} (\mathbb{E}_\pi \left[ |\bar{S} \left( \bar{\boldsymbol{\theta}}(s); \mathbf{x} \right)|^p \right]) < \infty.$$

**Proposition 2.** *Under Assumptions 1 and 2, it holds that:*

*(a) The function $V(s)$ is continuously differentiable on $\mathbb{S}$.*

*(b) For any $s \in \mathbb{S}$, we have $\langle \nabla_s V(s), \eta(s) \rangle \leq 0$ and the equality occurs if and only if $s$ lies in the set $\mathbb{F}$ defined in Equation (7), that is, $\{s \in \mathbb{S} : \langle \nabla_s V(s), \eta(s) \rangle = 0\} = \mathbb{F}$.*

*(c) For all compact subsets $\mathbb{K} \subset \mathbb{S} \setminus \mathbb{F}$, it follows that:*

$$\sup_{s \in \mathbb{K}} \langle \nabla_s V(s), \eta(s) \rangle < 0.$$

Building upon Proposition 1, additional conditions on the distribution of $\mathbf{x}$ and the stability of the sequence $\{s_n, n \geq 0\}$ are outlined in [23, Thm. 2 and Lem. 1]. When these conditions are coupled with the standard requirements for step sizes, specifically $\sum_n \gamma_n = +\infty$ and $\sum_n \gamma_n^2 < \infty$, they ensure the almost-sure convergence of the sequence $\{s_n, n \geq 0\}$ to the set $\mathbb{F}$. Moreover, the sequence $\{\bar{\boldsymbol{\theta}}(s_n), n \geq 0\}$ is also proven to converge almost surely to the set $\mathbb{L}$, which corresponds to the stationary points of the objective function $\boldsymbol{\theta} \mapsto \mathbb{E}_\pi [f(\boldsymbol{\theta}; \mathbf{x})]$.

### 3.2 Consistency of the online MM algorithm

We now prove the almost-sure convergence of the sequence $\{s_n, n \geq 0\}$ in Proposition 3, which is proved in Appendix A.3. To this end, we denote by $\mathbb{L} = \{\boldsymbol{\theta} \in \mathbb{T} : \nabla_{\boldsymbol{\theta}} \mathbb{E}_\pi [f(\boldsymbol{\theta}; \mathbf{x})] = 0\}$ the stationary points of the objective function and require the following assumptions on the distribution of $\mathbf{x}$, on the stability of the sequence $\{s_n, n \geq 0\}$, and the usual conditions on the step sizes:

**Assumption 3.**

*A8 The set $V(\mathbb{F})$ is nowhere dense.*

*A9 It holds that $\limsup |s_n| < \infty$ and $\liminf \{d(s_n, \mathbb{S}^c)\} > 0$ with probability 1.*

*A10 $\sum_n \gamma_n = +\infty$ and $\sum_n \gamma_n^2 < \infty$ for every $0 < \gamma_n < 1$.*

**Proposition 3** (Consistency)**.** *Under Assumptions 1 to 3, with probability 1,*

$$\lim_{n \to \infty} \{d(s_n, \mathbb{F})\} = 0 \text{ and } \lim_{n \to \infty} \{d(\boldsymbol{\theta}_n, \mathbb{L})\} = 0.$$

**Remark 2.** *The condition Assumption 3.A9 in Proposition 3 imposes a stability assumption, which is not straightforward. In general, the stability of the algorithm can be guaranteed by truncating the algorithm updates, either within a fixed set, as outlined in [49, Chapter 2], or within a progressively expanding sequence of sets, as demonstrated in [1]. To maintain conciseness in the exposition, we refrain from explicitly performing these constructions and leave them as directions for future work. The Assumption 3.A10 stated in Proposition 3 represents a conventional requirement for stochastic approximation methods employing diminishing step sizes, as highlighted in [49]. This condition is satisfied, for instance, by selecting the step size $\gamma_n = \gamma_0 n^{-\alpha}$, with $\alpha$ belonging to the interval $(\frac{1}{2}, 1]$. The supplementary conditions that $\gamma_n$ remain less than 1 and that $s_0$ be selected within $\mathbb{S}$ are intended solely to guarantee that the entire sequence $\{s_n, n \geq 0\}$ remains contained within $\mathbb{S}$ (see Assumption 1.A2).*

### 3.3 Convergence rate of the online MM algorithm

We aim to establish $\gamma_n^{-1/2}$-consistency and asymptotic normality in Theorem 1, with the proof provided in Appendix A.4. To establish the rate of convergence of the online MM algorithm, we first reframe it as a stochastic approximation procedure applied to $\boldsymbol{\theta}_n$, as detailed in Theorem 1 (a). To achieve this, we define $\boldsymbol{\theta}^0$ as a (possibly local) maximum of the objective function $\boldsymbol{\theta} \mapsto \mathbb{E}_\pi [f(\boldsymbol{\theta}; \mathbf{x})]$, and introduce the following matrix:

$$\boldsymbol{I}_\pi(\boldsymbol{\theta}^0) := -\mathbb{E}_\pi \left[\nabla^2_{\boldsymbol{\theta}^0} f(\mathbf{x}; \boldsymbol{\theta}^0)\right], \ \boldsymbol{H}(\boldsymbol{\theta}^0) := [\boldsymbol{I}_\pi(\boldsymbol{\theta}^0)]^{-1} \left\{\nabla^2_{\boldsymbol{\theta}} \mathbb{E}_\pi \left[f(\mathbf{x}; \boldsymbol{\theta}^0)\right]\right\},$$
$$\boldsymbol{\Gamma}(\boldsymbol{\theta}^0) := [\boldsymbol{I}_\pi(\boldsymbol{\theta}^0)]^{-1} \mathbb{E}_\pi \left[\nabla_{\boldsymbol{\theta}} f(\mathbf{x}; \boldsymbol{\theta}^0) \nabla_{\boldsymbol{\theta}} f(\mathbf{x}; \boldsymbol{\theta}^0)^\top\right] [\boldsymbol{I}_\pi(\boldsymbol{\theta}^0)]^{-1}.$$

**Theorem 1** (Convergence rate)**.** *Under Assumptions 1 to 3, the following holds:*

*(a) The online MM sequence $\{\boldsymbol{\theta}_n\}_{n\geq 0}$, defined by Equation (5) in Algorithm 1, satisfies the recursion:*

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n + \gamma_{n+1} \left[\boldsymbol{I}_\pi(\boldsymbol{\theta}_n)\right]^{-1} \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_n; \boldsymbol{x}_{n+1}) + \gamma_{n+1}\rho_{n+1},$$

*where $\lim_{n\to\infty} \rho_n = 0$ almost surely. Here, $\boldsymbol{I}_\pi(\boldsymbol{\theta}_n)$ denotes the Fisher information matrix, $\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_n; \boldsymbol{x}_{n+1})$ represents the gradient of the objective function evaluated at $\boldsymbol{\theta}_n$ with respect to the new observation $\boldsymbol{x}_{n+1}$, and $\rho_n$ captures a vanishing error term.*

*(b) The matrix $\boldsymbol{H}(\boldsymbol{\theta}^0)$ is a stable matrix, and its eigenvalues are such that their real parts are upper bounded by $-\lambda(\boldsymbol{\theta}^0)$, where $\lambda(\boldsymbol{\theta}^0) > 0$.*

*(c) Let $\gamma_n = \gamma_0 n^{-\alpha}$, where $\gamma_0$ can be freely chosen from the interval $(0, 1)$ when $\alpha \in (\frac{1}{2}, 1)$, but must satisfy $\gamma_0 > \dfrac{1}{2\lambda(\boldsymbol{\theta}^0)}$ when $\alpha = 1$. Then, on the event $\Omega(\boldsymbol{\theta}^0) = \left\{\lim_{n\to\infty} \boldsymbol{\theta}_n = \boldsymbol{\theta}^0\right\}$, the sequence $\gamma_n^{-1/2}(\boldsymbol{\theta}_n - \boldsymbol{\theta}^0)$ converges in distribution to a zero-mean Gaussian distribution with covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta}^0)$. The matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta}^0)$ is the solution of the Lyapunov equation:*

$$\left[\boldsymbol{H}(\boldsymbol{\theta}^0) + \zeta \boldsymbol{I}\right] \boldsymbol{\Sigma}(\boldsymbol{\theta}^0) + \boldsymbol{\Sigma}(\boldsymbol{\theta}^0) \left[\boldsymbol{H}^\top(\boldsymbol{\theta}^0) + \zeta \boldsymbol{I}\right] = -\boldsymbol{\Gamma}(\boldsymbol{\theta}^0), \tag{8}$$

*where $\zeta = 0$ if $\alpha \in \left(\frac{1}{2}, 1\right)$, and $\zeta = \dfrac{\gamma_0^{-1}}{2}$ if $\alpha = 1$.*

**Remark 3** (Polyak–Ruppert averaging)**.** *When the step size is chosen as $\gamma_n = \gamma_0 n^{-\alpha}$ with $\alpha = 1$, the algorithm achieves the optimal convergence rate of $n^{-1/2}$. However, this optimal rate comes with the significant drawback of requiring a specific constraint on the scale $\gamma_0$, a condition that is typically impractical to verify in real-world scenarios. On the other hand, selecting $\alpha \in (\frac{1}{2}, 1)$ leads to a slower convergence rate but avoids the need for stringent constraints on the scale $\gamma_0$ of the step size. The only requirement in this case is that $\gamma_0$ must remain smaller than 1, making it a more flexible and practical choice in most applications. To address this challenge, we propose employing the Polyak–Ruppert averaging technique [68, 93, 92, 98] as an effective post-processing strategy, see more details in Section 5.1.3.*

## 4 Application to softmax-gated Gaussian MoE models

To demonstrate the effectiveness of the proposed method, we consider a regression model that, as outlined in Section 2.1, represents a scenario where the expectation $\mathbb{E}_\pi\left[f\left(\boldsymbol{\theta}; \mathbf{x}\right)\right]$ in Equation (1) cannot be explicitly computed in closed form. The associated optimization problem is therefore tackled using an algorithm based on the MM framework in Algorithm 1 of Section 2.2.

We refer to the outputs $\mathbf{y} \in \mathcal{Y} \subset \mathbb{R}^Q,\ Q \in \mathbb{N}$, as the target or response variables and the inputs $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^P,\ P \in \mathbb{N}$, as the explanatory or predictor variables. Consider the dataset $\left(\mathbf{x}_{[N]}, \mathbf{y}_{[N]}\right) = \left(\left(\mathbf{x}_1, \mathbf{y}_1\right), \ldots, \left(\mathbf{x}_N, \mathbf{y}_N\right)\right)$ represents $N$ pairs of real-valued random samples of the variables $(\mathbf{x}, \mathbf{y})$. The corresponding observed values are denoted by $(\boldsymbol{x}, \boldsymbol{y})$.

### 4.1 MoE models for heterogeneous data

**Softmax-gated Gaussian MoE model.** We assume that for each $n \in [N]$, the covariates $\mathbf{x}_n$ are independent and not necessarily identically distributed, and the response variables $\mathbf{y}_n$ are independent given $\mathbf{x}_n$. Moreover, $\mathbf{y}_n$ follows a distribution characterized by the true but unknown probability density function $\pi(\cdot \mid \mathbf{x} = \boldsymbol{x})$. Motivated by universal approximation theorems for MoE models, the function $\pi$ can be estimated using a softmax-gated Gaussian MoE model, defined as:

$$s_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}) = \sum_{k=1}^{K} \mathrm{g}_k(\boldsymbol{w}(\boldsymbol{x}))\mathcal{N}(\boldsymbol{y}; \boldsymbol{v}_k(\boldsymbol{x}), \boldsymbol{\Sigma}_k) =: f\left(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y}\right). \tag{9}$$

The softmax gating network and expert network are defined respectively as follows:

$$\mathrm{g}_k(\boldsymbol{w}(\boldsymbol{x})) = \frac{\exp\left(w_k(\boldsymbol{x})\right)}{\sum_{l=1}^{K} \exp\left(w_l(\boldsymbol{x})\right)},\ \mathcal{N}(\boldsymbol{y}, \boldsymbol{v}_k(\boldsymbol{x}), \boldsymbol{\Sigma}_k) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{v}_k(\boldsymbol{x}))^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{y} - \boldsymbol{v}_k(\boldsymbol{x}))\right)}{\sqrt{(2\pi)^Q |\boldsymbol{\Sigma}_k|}},$$

where $\boldsymbol{w}(\boldsymbol{x}) = (w_1(\boldsymbol{x}), \ldots, w_K(\boldsymbol{x}))$ represents the weight functions and the Gaussian expert, $\mathcal{N}(\boldsymbol{y}, \boldsymbol{v}_k(\boldsymbol{x}), \boldsymbol{\Sigma}_k)$, is parameterized by the mean function $\boldsymbol{v}_k(\boldsymbol{x})$ and the covariance matrix $\boldsymbol{\Sigma}_k$. Moreover, both the weights $w_k$ and the means $\boldsymbol{v}_k$ are modeled as polynomials of the input variables $\boldsymbol{x}$ as follows: given any with $\boldsymbol{\omega}_{kd} = (\boldsymbol{\omega}_{kdp})_{p\in[P]} \in \mathbb{R}^P$,

$$w_k(\boldsymbol{x}) = \sum_{d=0}^{D_W} \left( \sum_{p=1}^{P} \boldsymbol{\omega}_{kdp} x_p^d \right), \ \boldsymbol{v}_k(\boldsymbol{x}) = \sum_{d=0}^{D_V} \boldsymbol{\Upsilon}_{kd} \boldsymbol{x}^d, \text{ with } \boldsymbol{\Upsilon}_{kd} \in \mathbb{R}^{Q\times P}.$$

Then, let $\boldsymbol{\omega} = (\boldsymbol{\omega}_k)_{k\in[K]}$, $\boldsymbol{\omega}_k = (\boldsymbol{\omega}_{kdp})_{p\in[P],d\in\{0,\ldots,D_W\}}$ and $\boldsymbol{\Upsilon}_k = (\boldsymbol{\Upsilon}_{kd})_{d\in\{0,1,\ldots,D_V\}}$ be the tuples of unknown coefficients with the maximum degrees $D_W$ and $D_V$ of polynomials for the weight and mean functions, respectively. Finally, the unknown parameters of the model are denoted as follows: $\boldsymbol{\theta} = (\boldsymbol{\omega}_k, \boldsymbol{\Upsilon}_k, \boldsymbol{\Sigma}_k)_{k\in[K]}$ .

**Identifiability of softmax-gated Gaussian MoE models.** Based on [43, 35], we parameterize the gating parameters via the constraints, without loss of generality, $\{\boldsymbol{\omega}_{Kdp}\}_{p\in[P],d\in\{0,\ldots,D_W\}} = \mathbf{0}$ such that

$$\mathrm{g}_K(\boldsymbol{x}, \boldsymbol{\omega}) \equiv \mathrm{g}_K(\boldsymbol{w}(\boldsymbol{x}; \boldsymbol{\omega})) = 1 - \sum_{k=1}^{K-1} \mathrm{g}_k(\boldsymbol{x}, \boldsymbol{\omega}) \quad \text{with}$$

$$\mathrm{g}_k(\boldsymbol{x}, \boldsymbol{\omega}) \equiv \mathrm{g}_k(\boldsymbol{w}(\boldsymbol{x}; \boldsymbol{\omega})) = \frac{\exp(w_k(\boldsymbol{x}; \boldsymbol{\omega}_k))}{1 + \sum_{l=1}^{K-1} \exp(w_l(\boldsymbol{x}; \boldsymbol{\omega}_l))}, \ \forall k \in [K-1].$$

**Motivation for polynomial regression.** To address the heterogeneous regression problem, certain authors have employed softmax-gated Gaussian Mixture of Experts (MoE) models under specific simplifying assumptions. Notably, [13, 14] explored MoE models for multiple regression with univariate output variables, where both the weights and means are modeled as linear functions of the input variables. However, this linear assumption restricts the capacity of MoE models. Indeed, in the context of convolutional neural networks, [15] empirically demonstrated that while mixtures of linear experts outperform single-expert models, they fall significantly short when compared to mixtures of non-linear experts. Motivated by these findings, we aim to enhance the flexibility of MoE models by incorporating nonlinearities. Specifically, we define the weights and mean experts as polynomials of the input variables. Regarding the convergence properties of such polynomial-based MoE models, we refer to [66], which provides insights into the optimal convergence rates of MoE models where each expert employs a polynomial regression framework.

### 4.2 Online MM algorithm for softmax-gated Gaussian MoE models

We denote by $\boldsymbol{\theta}^0$ a (possibly local) maximum of the objective function:

$$\boldsymbol{\theta} \mapsto \mathbb{E}_{\mathbf{y}|\mathbf{x}\sim\mathbb{P}_\pi}\left[\log[f(\boldsymbol{\theta}; \mathbf{x}, \mathbf{y})]\right].$$

Since the true conditional density $\pi$ is typically unknown in practice, we aim to explore the maximum log-likelihood estimator (MLE) for the softmax-gated Gaussian MoE models defined in Equation (9), using the dataset $\left(\mathbf{x}_{[N]}, \boldsymbol{y}_{[N]}\right)$. The MLE is formally expressed as:

$$\widehat{\boldsymbol{\theta}}_N = \operatorname*{arg\,max}_{\boldsymbol{\theta}\in\mathbb{T}} \left\{ \frac{1}{N} \log[f(\boldsymbol{\theta}; \boldsymbol{x}_{[N]}, \boldsymbol{y}_{[N]})] \right\}, \text{ where}$$

$$\log[f(\boldsymbol{\theta}; \boldsymbol{x}_{[N]}, \boldsymbol{y}_{[N]})] = \sum_{n=1}^{N} \log\left[ \sum_{k=1}^{K} \mathrm{g}_k(\boldsymbol{w}(\boldsymbol{x}_n)) \mathcal{N}(\boldsymbol{y}_n; \boldsymbol{v}_k(\boldsymbol{x}_n), \boldsymbol{\Sigma}_k) \right] =: L(\boldsymbol{\theta}). \tag{10}$$

For simplicity, this study on the application to softmax-gated Gaussian MoE models is restricted to scalar variables for both $\boldsymbol{x}$ and $\boldsymbol{y}$, *i.e.*, $\boldsymbol{\theta} = \left(\boldsymbol{\omega}_k, \boldsymbol{\upsilon}_k, \sigma_k^2\right)_{k\in[K]}$. Extending the analysis to the multivariate case is left for future work. To develop the online MM algorithm as in Algorithm 1 for softmax-gated Gaussian MoE models in Algorithm 2, we begin by introducing the notations:

- The term $s_{4,n,[k-1,k]}$ refers to the segment of the vector $s_{4,n}$ spanning from position $(k-1)(D_V+1)+1$ to $k(D_V+1)$. Similarly, $s_{5,n,[k-1,k]}$ denotes the segment of the vector $s_{5,n}$ ranging from position $(k-1)(D_V+1)^2+1$ to $k(D_V+1)^2$.

- The term $s_{3,n,k}$ represents the $k$-th component of the vector $s_{3,n}$, while $s_{6,n,k}$ denotes the $k$-th component of the vector $s_{6,n}$.
- The term $\mathrm{mat}(s_{2,n+1})$ refers to the process of reshaping the vector $s_{2,n+1}$ into a matrix with dimensions $K(D_W+1)\times K(D_W+1)$. Similarly, $\mathrm{mat}(s_{5,n+1,k})$ denotes the reshaping of the vector $s_{5,n+1,k}$ into a square matrix of dimensions $(D_V+1)\times(D_V+1)$.
- $\tau_{kn}^{(t)} = \dfrac{\mathsf{g}_k(\boldsymbol{w}(\boldsymbol{x}_n))^{(t)}\mathcal{N}(\boldsymbol{y}_n,\boldsymbol{v}_k(\boldsymbol{x}_n),\sigma_k)^{(t)}}{\sum_{l=1}^{K}\mathsf{g}_l(\boldsymbol{w}(\boldsymbol{x}_n))^{(t)}\mathcal{N}(\boldsymbol{y}_n;\boldsymbol{v}_l(\boldsymbol{x}_n),\sigma_l)^{(t)}}$.
- $\boldsymbol{\xi}_n = [\tau_{1n}^{(t)}\boldsymbol{x}_n^0, \tau_{1n}^{(t)}\boldsymbol{x}_n^1,\ldots,\tau_{1n}^{(t)}\boldsymbol{x}_n^{D_W},\ldots,\tau_{(K-1)n}^{(t)}\boldsymbol{x}_n^0,\ldots,\tau_{(K-1)n}^{(t)}\boldsymbol{x}_n^{D_W}]^\top$.
- $\hat{\boldsymbol{x}}_n = \left[\boldsymbol{x}_n^0,\ldots,\boldsymbol{x}_n^{D_W}\right]^\top$, $\boldsymbol{B}_n = \left(1.5\boldsymbol{I}_{K-1} - \dfrac{\mathbf{1}\mathbf{1}^\top}{K-1}\right)/2 \otimes \hat{\boldsymbol{x}}_n\hat{\boldsymbol{x}}_n^\top$.
- $\boldsymbol{\tau}_n^{(t)} = [\tau_{1n}^{(t)},\tau_{2n}^{(t)},\ldots,\tau_{Kn}^{(t)}]^\top$, $\boldsymbol{r}_n = \left[1,\boldsymbol{x}_n,\boldsymbol{x}_n^2,\ldots,\boldsymbol{x}_n^{D_V}\right]^\top$.

Here, we denote $\boldsymbol{\theta}^{(t)} = \left[\boldsymbol{\omega}^{(t)},\boldsymbol{\upsilon}^{2(t)},\sigma^{(t)}\right]$ as the parameter vector at the $t$-th iteration, and $\tau_{kn}^{(t)}$ represents the posterior probability that the data point $(\boldsymbol{x}_n,\boldsymbol{y}_n)$ belongs to the $k$-th expert.

**Algorithm 2** Online MM algorithm for softmax-gated Gaussian MoE models

**Require:** An initial value $s_0\in\mathbb{S}$ and positive step sizes $\{\gamma_{n+1}, n\geq 1\}$ in $(0,1)$.
**Ensure:** A $\mathbb{T}$-valued sequence $\{\boldsymbol{\theta}_n, n\geq 0\}$ and a $\mathbb{S}$-valued sequence $\{s_n, n\geq 0\}$.

1: **for** $n=0,1,\ldots,$ **do**
2: Compute

$$
\begin{aligned}
s_{1,n+1} &= s_{1,n} + \gamma_{n+1}\left(-\boldsymbol{\xi}_{n+1} + \nabla f(\boldsymbol{\omega}^{(n)}) - \boldsymbol{B}_{n+1}\boldsymbol{\omega}^{(n)} - s_{1,n}\right),\\
s_{2,n+1} &= s_{2,n} + \gamma_{n+1}\left[\frac{1}{2}\,\mathrm{vec}\left(\boldsymbol{B}_{n+1}\right) - s_{2,n}\right],\ \mathrm{mat}(s_{2,0})\succ 0,\\
s_{3,n+1} &= s_{3,n} + \gamma_{n+1}\left(y_{n+1}^2\boldsymbol{\tau}_{n+1}^{(n)} - s_{3,n}\right),\ s_{3,0} = \mathbf{1}_K,\\
s_{4,n+1} &= s_{4,n} + \gamma_{n+1}\left(\boldsymbol{\tau}_{n+1}^{(n)}\otimes(-2y_{n+1}\boldsymbol{r}_{n+1}) - s_{4,n}\right),\ s_{4,0} = \mathbf{1}_K\otimes\boldsymbol{z},\ \boldsymbol{z}\in\mathbb{R}^{D_V+1},\\
s_{5,n+1} &= s_{5,n} + \gamma_{n+1}\left(\boldsymbol{\tau}_{n+1}^{(n)}\otimes\left(\mathrm{vec}\left(\boldsymbol{r}_{n+1}\boldsymbol{r}_{n+1}^\top\right)\right) - s_{5,n}\right),\ s_{5,0} = \mathbf{1}_K\otimes\mathrm{vec}(\boldsymbol{z}\boldsymbol{z}^\top + \boldsymbol{I}_{D_V+1}),\\
s_{6,n+1} &= s_{6,n} + \gamma_{n+1}\left(\boldsymbol{\tau}_{n+1}^{(n)} - s_{6,n}\right),\ s_{6,0} > 0,\\
\boldsymbol{\theta}_{n+1} &= \left(\boldsymbol{\omega}_{k,n+1},\boldsymbol{\upsilon}_{k,n+1},\sigma_{k,n+1}^2\right)_{k\in[K]}\ \text{where}\\
\boldsymbol{\omega}_{n+1} &= -\left(\mathrm{mat}(s_{2,n+1}) + \mathrm{mat}(s_{2,n+1})^\top\right)^{-1}s_{1,n+1},\\
\boldsymbol{\upsilon}_{k,n+1} &= -\left(\mathrm{mat}\left(s_{5,n+1,[k-1,k]}\right) + \mathrm{mat}\left(s_{5,n+1,[k-1,k]}\right)^\top\right)^{-1}s_{4,n+1,[k-1,k]},\\
\sigma_{k,n+1}^2 &= \frac{s_{3,n+1,k} + s_{4,n+1,[k-1,k]}^\top\boldsymbol{\upsilon}_k + s_{5,n+1,[k-1,k]}^\top\mathrm{vec}(\boldsymbol{\upsilon}_k\boldsymbol{\upsilon}_k^\top)}{s_{6,n+1,k}}.
\end{aligned}
$$

3: **end for**

We direct the reader to Appendix A.6 for a detailed derivation of the update equations presented in Algorithm 2, which are supported by the theoretical guarantees established in Theorem 2.

**Theorem 2** (Stability, consistency, and convergence rate of Algorithm 2)**.** *If the output sequence $\{\boldsymbol{\theta}_n, n\geq 0\}$ of Algorithm 2 and the true (but unknown) conditional probability density function $\pi(\cdot\mid\mathbf{x}=\boldsymbol{x})$ satisfy Assumptions 1 to 3, all the theoretical guarantees established in Propositions 1 to 3 and Theorem 1 are met.*

The proof of Theorem 2 is deferred to Appendix A.5.

# 5 Experimental study

## 5.1 Experiments on simulation data sets

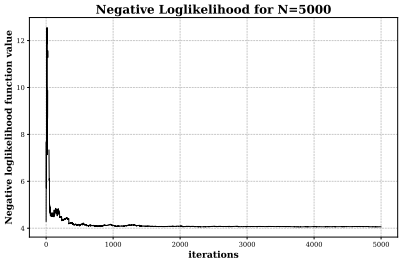
In this section, we perform experiments on a well-specified dataset, *i.e.,* $\pi(\cdot \mid \mathbf{x} = \boldsymbol{x}) = f\left(\boldsymbol{\theta}^0; \boldsymbol{x}, \boldsymbol{y}\right)$ comprising $N = 5000$ data points (*i.e.,* , $N = 5000$) with the true number of component expert set to $K_0 = 5$. Furthermore, we select $D_W = D_V = 2$, and specify the true parameters as follows:

$$\boldsymbol{\omega}^0 = \begin{bmatrix} 0.2 & 0.5 & 0.1 \\ 0.8 & 1.2 & 0.1 \\ 0.3 & 1.5 & 0.1 \\ 0.5 & 0.5 & 0.1 \\ 0 & 0 & 0 \end{bmatrix}, \quad \boldsymbol{\upsilon}^0 = \begin{bmatrix} 1 & 5 & 0.1 \\ 300 & 7 & 0.2 \\ 100 & 6 & 0.3 \\ 200 & 9 & 0.4 \\ 400 & 3 & 0.5 \end{bmatrix}, \quad \boldsymbol{\sigma}^0 = \begin{bmatrix} 5.2 \\ 6.8 \\ 5.8 \\ 6.2 \\ 5 \end{bmatrix}.$$

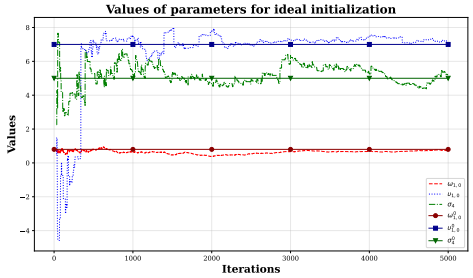
We investigated two approaches for parameter initialization. The first approach incorporates the addition of a noise term to each true parameter, simulating small random perturbations. The second approach leverages K-means clustering for initialization, which does not require prior knowledge of the true parameters and provides a data-driven starting point.

### 5.1.1 Well-specified with ideal initialization

In the first method, the experimental setup assumes a well-specified model with parameters initialized close to their true values, ensuring good initialization for the algorithm. More precisely, noise variables are independently drawn from a Gaussian distribution with a mean of 0 and a variance of 1, then scaled by a factor of 0.005. The initial values for each series, $s_{i,0}$, are computed as the average of the corresponding $S(\boldsymbol{x})$ values over the first 85 observations of $\boldsymbol{x}$, parameterized by the initialized parameters. Figure 1a illustrates the progression of the negative log-likelihood function across iterations. The function demonstrates rapid convergence, nearing its optimal value within approximately 1500 iterations out of a total of 5000. This highlights the efficiency and effectiveness of the proposed method. Similarly, Figure 1b presents the evolution of selected parameter values during the algorithm's execution. The results indicate that all parameters converge closely to their true values, further validating the robustness of the approach.




(a) Convergence of negative log-likelihood.

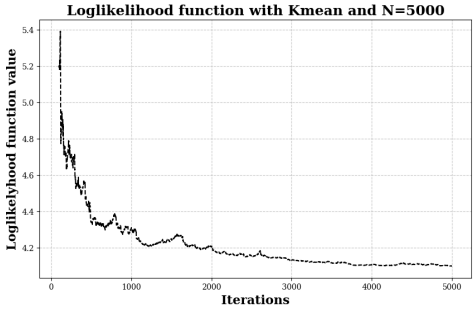



(b) Convergence of parameter estimation.

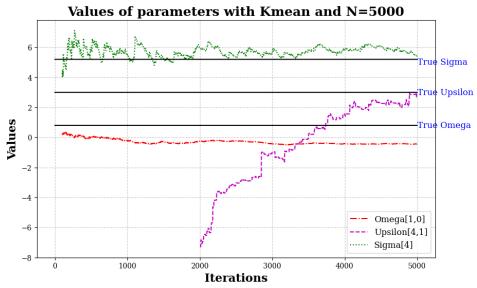
Figure 1: Ideal initialization: (a) The progression of the negative log-likelihood function is depicted from iteration 100 to 5000. (b) The evolution of the parameter values for $\omega_{1,0}$, $\upsilon_{1,1}$, and $\sigma_4$ is shown over 5000 iterations.

### 5.1.2 Well-specified with K-means initialization

To demonstrate the robustness of our algorithm, we employ a second initialization method based on K-means clustering. Specifically, K-means is performed using randomly selected centroids, repeated up to ten times, and the result with the lowest distortion is chosen as the initialization for our algorithm. The initial values for each series, $s_{i,0}$, are determined similarly to the method outlined previously. The results are illustrated in Figure 2a and Figure 2b, showing the loss function trajectory and the progression of parameter estimates during the algorithm's execution. Since the initialization provided by K-means is farther from the optimal values compared to the earlier method, the algorithm requires more iterations to achieve convergence. Nevertheless, all parameters eventually converge to their true values, validating the algorithm's robustness.




(a) Convergence of negative log-likelihood.




(b) Convergence of parameter estimation.

Figure 2: K-mean initialization: (a) The progression of the negative log-likelihood function is depicted from iteration 100 to 5000. (b) The evolution of the parameter values for $\omega_{1,0}$, $\upsilon_{1,1}$, and $\sigma_4$ is shown over 5000 iterations.

### 5.1.3 Polyak–Ruppert averaging

In the context of online optimization, Polyak averaging [93] is a widely adopted technique to mitigate the variability of results and avoid the influence of initial highly volatile estimates. Starting from a designated iteration $n_0$, the averaged sequence is initialized as $\boldsymbol{\theta}^P_{n_0} = 0$. For $n \geq n_0$, the updates are given by:

$$\boldsymbol{\theta}^P_{n+1} = \boldsymbol{\theta}^P_n + \alpha_{n-n_0+1}(\boldsymbol{\theta}_n - \boldsymbol{\theta}^P_n), \text{ where } \alpha_n \text{ is usually chosen to be } 1/n.$$

In the case of our first initialization method using noise terms, we applied Polyak averaging from $n_0 = 100$, with $\alpha_n = 1/n$. The corresponding loss function values are presented in Figure 3, alongside those obtained without Polyak averaging. This approach effectively reduces the influence of early-stage variability and stabilizes the loss function trajectory.

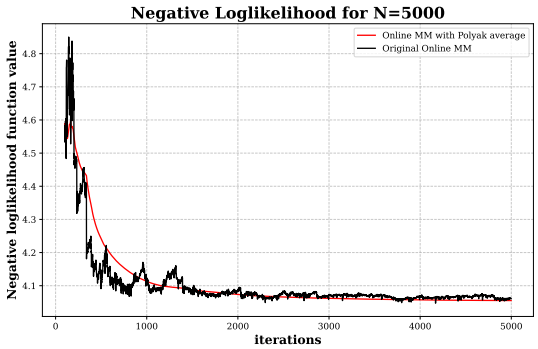



Figure 3: Negative log-likelihood function with Polyak average.

### 5.1.4 Train-Test evaluation with ideal initialization

In this section, we apply the proposed model to the tasks of training and testing (*i.e.,* the model parameters are optimized using a designated training dataset, while the performance is evaluated on a separate testing dataset to measure the accuracy). This setup simulates real-world scenarios where historical streaming data is utilized to train the model, enabling it to predict subsequent values effectively.

In terms of the dataset, we follow the initialization described by Chamroukhi [12, 10] in Table 1, which is as follows:

$$\boldsymbol{\omega}^0 = \begin{bmatrix} 0 & 10 \\ 0 & 0 \end{bmatrix}, \quad \boldsymbol{\upsilon}^0 = \begin{bmatrix} 0 & 1 \\ 0 & -1 \end{bmatrix}, \quad \boldsymbol{\sigma}^0 = \begin{bmatrix} 0.1 \\ 0.1 \end{bmatrix}.$$

Additionally, the training dataset consists of 2,000 data points, evenly distributed across two clusters, with 1,000 points per cluster. Similarly, the test dataset comprises 400 data points, with 200 points allocated to each cluster. Illustrations of this setup are provided in Figure 4.

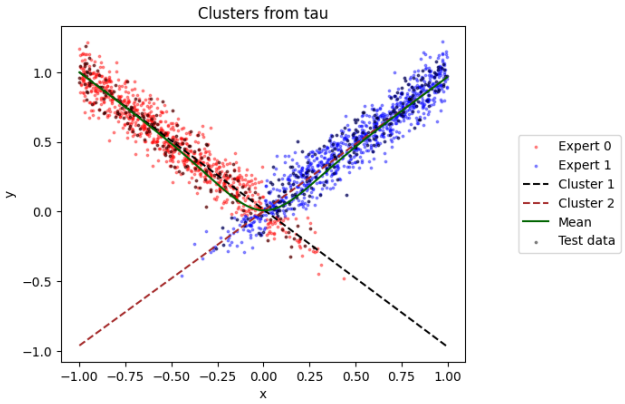
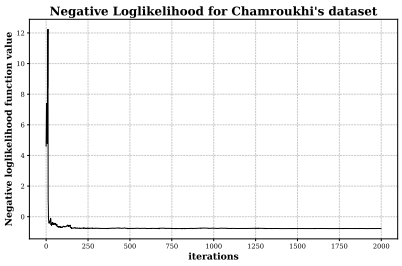


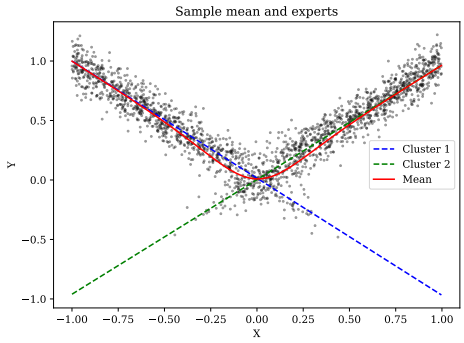

Figure 4: Typical realization of softmax-gated Gaussian MoE models with $K_0 = 2$.

The results of our experiments are presented in Figure 5. Specifically, Figure 5a illustrates the negative log-likelihood function of the training dataset during the execution of the algorithm, while Figure 5b depicts the model's predictions for the test dataset. Additionally, we include the prediction mean line for each cluster and the prediction overall mean line across both clusters for easier visualization. Furthermore, the accuracy of our model are provided in Table 1.

Overall, the results demonstrate the effectiveness of our model. It converges to near its optimal point in fewer than 250 iterations, and the resulting mean line shown in Figure 5b as well as the numbers in Table 1 exhibits excellent accuracy.




(a) True negative log-likelihood




(b) Test set results

Figure 5: (a) Negative log-likelihood values during training. (b) Predictions on the test set with cluster-specific and overall mean lines.

### 5.2 Application to a real-world dataset

In this experiment, we analyze a real-world dataset related to climate change, focusing on global surface temperature trends. The NASA GISS Surface Temperature (GISTEMP) analysis provides monthly measures of global surface temperature changes dating back to 1880, when a sufficient global distribution of meteorological stations became available. Although the GISS analysis is updated monthly, the dataset utilized here [33] is updated annually, sourced from the Carbon Dioxide Information Analysis Center (CDIAC)[1]. This center has served as the primary repository for climate-change data and analysis under the US Department of Energy since 1982. The dataset includes

[1] https://data.ess-dive.lbl.gov/view/doi:10.3334/CDIAC/CLI.001

$N = 136$ yearly measurements of global annual temperature anomalies (in degrees Celsius), derived from land-based meteorological stations, spanning the period from 1882 to 2012. These data have been previously analyzed by researchers [32, 97, 11, 12, 10, 79] to explore long-term temperature trends and their implications for climate change.

We partitioned the dataset into two subsets: the first 100 data points were utilized for training, while the remaining 36 data points were reserved for testing. The training loss function is presented in Figure 7a, and the resulting clustering of the test set is depicted in Figure 7b. Furthermore, we evaluated the accuracy of our model, and the results are provided in Table 2. Despite the temperature anomaly dataset being relatively small, containing only 136 data points, our model exhibits remarkable adaptability. It converges to a solution near its optimal value after just over 40 iterations, achieving high accuracy despite the limited size of the dataset. These findings highlight the robustness of our proposed method.

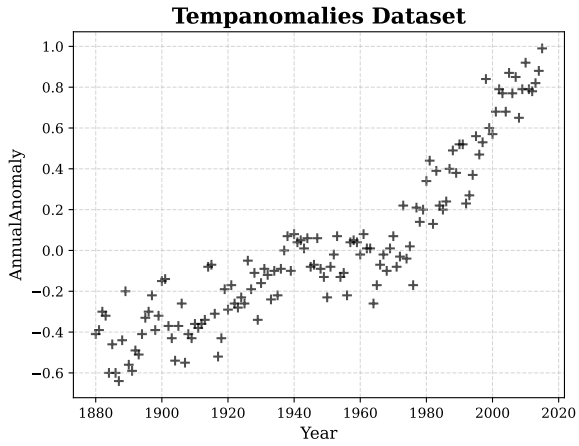
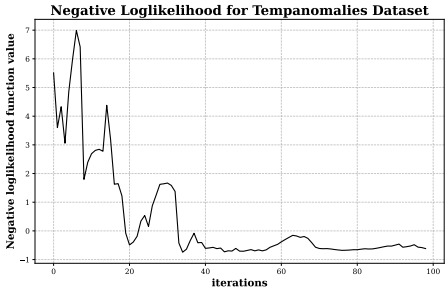
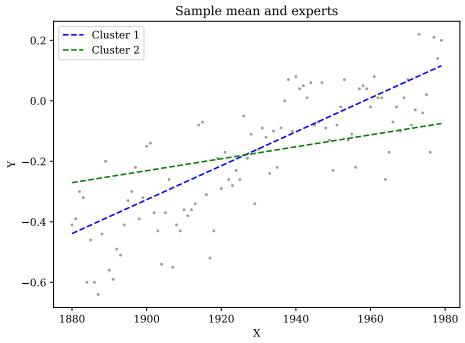



Figure 6: Unprocessed temperature anomalies dataset.




(a) Negative log-likelihood function of the training process.




(b) The clustering of the test set.

Figure 7: (a) The values of the true negative log-likelihood function of the training set. (b) The clustering values of the test set.

| Metric | Value |
|---|---|
| MSE | 0.015 |
| MAPE | 0.398 |
| RAE | 0.341 |
| RSE | 0.140 |

Table 1: The accuracy of the online MM algorithm applied to Chamroukhi's dataset.

.

| Metric | Value |
|---|---|
| MSE | 0.228 |
| MAPE | 0.731 |
| RAE | 2.130 |
| RSE | 4.107 |

Table 2: Accuracy of the online MM for the temperature anomalies dataset.

.


## Acknowledgments

TrungTin Nguyen, Hien Duy Nguyen, Florence Forbes, and Gersende Fort acknowledge funding from the Australian Research Council grant DP230100905, and from Inria Project WOMBAT.


# Supplementary Materials for "An Online Minorization-Maximization Algorithm : Theory and Applications"

In this supplementary material, we present the proofs of the main results in Appendix A and the technical proofs and results in Appendix B and Appendix C, respectively.

## A Proofs of main results

### A.1 Proof of Proposition 1

A4 implies that

$$-\nabla_{\boldsymbol{\theta}}\psi(\bar{\boldsymbol{\theta}}(s)) + \nabla_{\boldsymbol{\theta}}\phi(\bar{\boldsymbol{\theta}}(s))^{\top} s = 0, \qquad s \in \mathbb{S}. \tag{11}$$

By using Equation (2) and A1, and apply the expectation w.r.t. $\mathbb{P}_\pi$ (under Assumption 1.A3), for any $\boldsymbol{\theta}, \boldsymbol{\tau} \in \mathbb{T}$, we obtain

$$\mathbb{E}_\pi\left[f(\boldsymbol{\theta};\mathbf{x})\right] - \mathbb{E}_\pi\left[f(\boldsymbol{\tau};\mathbf{x})\right] \geq \mathbb{E}_\pi\left[g(\boldsymbol{\theta},\mathbf{x};\boldsymbol{\tau}) - g(\boldsymbol{\tau},\mathbf{x};\boldsymbol{\tau})\right].$$

This inequality establishes a minorizer function for $\boldsymbol{\theta} \mapsto \mathbb{E}_\pi\left[f(\boldsymbol{\theta};\mathbf{x})\right]$, ensuring that the difference is nonnegative and achieves its minimum value (zero) when $\boldsymbol{\theta} = \boldsymbol{\tau}$.

In accordance with the definition of $g$ as the minorizer of $f$ in Equation (2), namely, $f(\boldsymbol{\tau};\mathbf{x}) = g(\boldsymbol{\tau},\mathbf{x};\boldsymbol{\tau})$, and Assumption 1.A1, we have the following conclusion:

$$\begin{aligned}\nabla_{\boldsymbol{\theta}}\mathbb{E}_\pi\left[f(\boldsymbol{\tau};\mathbf{x})\right] + \nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\tau}) - \nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\tau})^{\top}\mathbb{E}_\pi\left[\bar{S}(\boldsymbol{\tau};\mathbf{x})\right] &= \nabla_{\boldsymbol{\theta}}\mathbb{E}_\pi\left[f(\boldsymbol{\tau};\mathbf{x})\right] - \nabla_{\boldsymbol{\theta}}\mathbb{E}_\pi\left[g(\boldsymbol{\theta},\mathbf{x};\boldsymbol{\tau})\right]|_{\boldsymbol{\theta}=\boldsymbol{\tau}} \\ &= 0. \end{aligned} \tag{12}$$

Let $s^0 \in \mathbb{F}$ and apply Equation (12) with $\boldsymbol{\tau} \leftarrow \bar{\boldsymbol{\theta}}(s^0)$. It then follows that

$$\nabla_{\boldsymbol{\theta}}\mathbb{E}_\pi\left[f(\cdot;\mathbf{x})\right]|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}(s^0)} + \nabla_{\boldsymbol{\theta}}\psi(\bar{\boldsymbol{\theta}}(s^0)) - \nabla_{\boldsymbol{\theta}}\phi(\bar{\boldsymbol{\theta}}(s^0))^{\top} s^0 = 0,$$

which implies $\bar{\boldsymbol{\theta}}(s^0) \in \mathbb{L}$ by Equation (11).

Conversely, if $\boldsymbol{\theta}^0 \in \mathbb{L}$, then by Equation (12), we have

$$\nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta}^0) - \nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta}^0)^{\top}\mathbb{E}_\pi\left[\bar{S}(\boldsymbol{\theta}^0;\mathbf{x})\right] = 0,$$

which, by A3 and A4, implies that $\boldsymbol{\theta}^0 = \bar{\boldsymbol{\theta}}\left(\mathbb{E}_\pi\left[\bar{S}(\boldsymbol{\theta}^0;\mathbf{x})\right]\right) = \bar{\boldsymbol{\theta}}(s^0)$. By definition of $s^0$, this yields $s^0 = \mathbb{E}_\pi\left[\bar{S}\left(\bar{\boldsymbol{\theta}}(s^0);\mathbf{x}\right)\right]$; *i.e.,* $s^0 \in \mathbb{F}$.

### A.2 Proof of Proposition 2

(a) Using Assumption 2, parts A5 and A6, it holds that $V(s)$ is continuously differentiable on $\mathbb{S}$.

(b) In accordance with the definition of $g$ as the minorizer of $f$ in Equation (2), namely $f(\boldsymbol{\tau};\mathbf{x}) = g(\boldsymbol{\theta},\mathbf{x};\boldsymbol{\tau})$, where we choose $\boldsymbol{\tau} = \bar{\boldsymbol{\theta}}(s)$, and Assumption 1.A1, we obtain that

$$\nabla_{\boldsymbol{\theta}}\mathbb{E}_\pi\left[f(\bar{\boldsymbol{\theta}}(s);\mathbf{x})\right] = -\nabla_{\boldsymbol{\theta}}\psi(\bar{\boldsymbol{\theta}}(s)) + \nabla_{\boldsymbol{\theta}}\phi(\bar{\boldsymbol{\theta}}(s))^\top\mathbb{E}_\pi\left[\bar{S}(\bar{\boldsymbol{\theta}}(s);\mathbf{x})\right]. \tag{13}$$

Then, via Equation (13), we apply the chain rule of differentiation to get

$$\begin{aligned}
-\nabla_s V(s) &= \nabla_s\mathbb{E}_\pi\left[f(\bar{\boldsymbol{\theta}}(s);\mathbf{x})\right] = \nabla_s\bar{\boldsymbol{\theta}}(s)^\top\nabla_{\boldsymbol{\theta}}\mathbb{E}_\pi\left[f(\bar{\boldsymbol{\theta}}(s);\mathbf{x})\right] \\
&= \nabla_s\bar{\boldsymbol{\theta}}(s)^\top\{-\nabla_{\boldsymbol{\theta}}\psi(\bar{\boldsymbol{\theta}}(s)) + \nabla_{\boldsymbol{\theta}}\phi(\bar{\boldsymbol{\theta}}(s))^\top\mathbb{E}_\pi\left[\bar{S}(\bar{\boldsymbol{\theta}}(s);\mathbf{x})\right]\} \\
&= \nabla_s\bar{\boldsymbol{\theta}}(s)^\top\{-\nabla_{\boldsymbol{\theta}}\phi(\bar{\boldsymbol{\theta}}(s))^\top s + \nabla_{\boldsymbol{\theta}}\phi(\bar{\boldsymbol{\theta}}(s))^\top\mathbb{E}_\pi\left[\bar{S}(\bar{\boldsymbol{\theta}}(s);\mathbf{x})\right]\} \text{ (using Equation (4))} \\
&= \nabla_s\bar{\boldsymbol{\theta}}(s)^\top\{\nabla_{\boldsymbol{\theta}}\phi(\bar{\boldsymbol{\theta}}(s))^\top\}\eta(s) \text{ (using definition of } \eta(s) \text{ in Equation (7)).}
\end{aligned} \tag{14}$$

Recall that

$$\bar{\boldsymbol{\theta}}(s) := \arg\max_{\boldsymbol{\theta}\in\mathbb{T}}[-\psi(\boldsymbol{\theta}) + \langle s, \phi(\boldsymbol{\theta})\rangle], \quad \nabla_{\boldsymbol{\theta}}h(s;\boldsymbol{\theta})|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}(s)} = -\nabla_{\boldsymbol{\theta}}\psi(\bar{\boldsymbol{\theta}}(s)) + \nabla_{\boldsymbol{\theta}}\phi(\bar{\boldsymbol{\theta}}(s))^\top s = 0$$

$$h(s;\boldsymbol{\theta}) := -\psi(\boldsymbol{\theta}) + \langle s, \phi(\boldsymbol{\theta})\rangle = -\psi(\boldsymbol{\theta}) + \phi(\boldsymbol{\theta})^\top s. \tag{15}$$

This implies that

$$\begin{aligned}
\nabla_s[-\nabla_{\boldsymbol{\theta}}\psi(\bar{\boldsymbol{\theta}}(s))] &= -\nabla_s[\nabla_{\boldsymbol{\theta}}\phi(\bar{\boldsymbol{\theta}}(s))^\top s], \\
\nabla_s[\nabla_{\boldsymbol{\theta}}h(s;\boldsymbol{\theta})]^\top &= \nabla_s[\nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta})^\top s] = (\nabla_s s)[\nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta})^\top]^\top = [\nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta})^\top]^\top.
\end{aligned} \tag{16}$$

The last equality comes from the fact that $(\nabla_s s^\top)$ is an identity $d \times d$ matrix. We have

$$\begin{aligned}
\nabla_s[\nabla_{\boldsymbol{\theta}}h(s;\boldsymbol{\theta})|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}(s)}]^\top &= \nabla_s\{\nabla_{\boldsymbol{\theta}}h(s;\boldsymbol{\theta})|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}(s)}\} + \nabla_s\bar{\boldsymbol{\theta}}(s)^\top\nabla_{\boldsymbol{\theta}}[\nabla_{\boldsymbol{\theta}}h(s;\boldsymbol{\theta})|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}(s)}] \\
&\quad \text{(since } \nabla_{\boldsymbol{\theta}}h(s;\boldsymbol{\theta})|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}(s)} = 0\text{)} \\
&= [\nabla_{\boldsymbol{\theta}}\phi(\bar{\boldsymbol{\theta}}(s))^\top]^\top + \nabla_s\bar{\boldsymbol{\theta}}(s)^\top\nabla^2_{\boldsymbol{\theta}}[h(s;\boldsymbol{\theta})|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}(s)}] \\
&\quad \text{(using Equation (16))} \\
&= 0.
\end{aligned}$$

The last equality comes from the fact that the function $\nabla_{\boldsymbol{\theta}}h(s;\boldsymbol{\theta})|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}(s)}$ is identically equal to 0, hence $\nabla_s[\nabla_{\boldsymbol{\theta}}h(s;\boldsymbol{\theta})|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}(s)}]^\top = 0$ with $\nabla_s[\nabla_{\boldsymbol{\theta}}h(s;\boldsymbol{\theta})|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}(s)}]^\top$ is a $d \times p$ matrix. Hence,

$$\nabla_s\bar{\boldsymbol{\theta}}(s)^\top = -[\nabla_{\boldsymbol{\theta}}\phi(\bar{\boldsymbol{\theta}}(s))^\top]^\top\{\nabla^2_{\boldsymbol{\theta}}[h(s;\boldsymbol{\theta})|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}(s)}]\}^{-1}. \tag{17}$$

Assumption 1.A4 implies that $\nabla^2_{\boldsymbol{\theta}}h(s;\boldsymbol{\theta})|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}(s)}$ and its inverse are both negative definite matrices and hence $v^\top\{\nabla^2_{\boldsymbol{\theta}}[h(s;\boldsymbol{\theta})|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}(s)}]\}^{-1}v \leq 0$ for any vector $v$ with equality if and only if $v = 0$. Hence, for any $s \in \mathbb{S}$,

$$\begin{aligned}
\langle\nabla_s V(s), \eta(s)\rangle &= -\langle\nabla_s\bar{\boldsymbol{\theta}}(s)^\top[\nabla_{\boldsymbol{\theta}}\phi(\bar{\boldsymbol{\theta}}(s))^\top]\eta(s), \eta(s)\rangle \\
&= \eta(s)^\top[\nabla_{\boldsymbol{\theta}}\phi(\bar{\boldsymbol{\theta}}(s))^\top]^\top\{\nabla^2_{\boldsymbol{\theta}}[h(s;\boldsymbol{\theta})|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}(s)}]\}^{-1}[\nabla_{\boldsymbol{\theta}}\phi(\bar{\boldsymbol{\theta}}(s))^\top]\eta(s) \\
&= [\nabla_{\boldsymbol{\theta}}\phi(\bar{\boldsymbol{\theta}}(s))^\top\eta(s)]^\top\{\nabla^2_{\boldsymbol{\theta}}[h(s;\boldsymbol{\theta})|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}(s)}]\}^{-1}[\nabla_{\boldsymbol{\theta}}\phi(\bar{\boldsymbol{\theta}}(s))^\top\eta(s)] \leq 0.
\end{aligned} \tag{18}$$

The equality in Equation (18) happens if and only if there exists $s^0$ such that $\nabla_{\boldsymbol{\theta}}\phi(\bar{\boldsymbol{\theta}}(s^0))^\top h(s^0) = 0$. This is equivalent to

$$\nabla_{\boldsymbol{\theta}}\phi(\bar{\boldsymbol{\theta}}(s^0))^\top\mathbb{E}_\pi\left[\bar{S}\left(\bar{\boldsymbol{\theta}}\left(s^0\right);\mathbf{x}\right)\right] = \nabla_{\boldsymbol{\theta}}\phi(\bar{\boldsymbol{\theta}}(s^0))^\top s^0. \tag{19}$$

Using Assumption 1.A4, for $s^0 \in \mathbb{S}$, there exists a unique root to the function $\boldsymbol{\theta} \mapsto -\nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta})^\top s^0$, which is the unique global maximum on $\mathbb{T}$ of the function $\boldsymbol{\theta} \mapsto h(s^0;\boldsymbol{\theta}) := -\psi(\boldsymbol{\theta}) + \left\langle s^0, \phi(\boldsymbol{\theta})\right\rangle$. This root is denoted by $\bar{\boldsymbol{\theta}}(s^0)$, that is,

$$\bar{\boldsymbol{\theta}}(s^0) := \arg\max_{\boldsymbol{\theta}\in\mathbb{T}}[-\psi(\boldsymbol{\theta}) + \left\langle s^0, \phi(\boldsymbol{\theta})\right\rangle], \quad -\nabla_{\boldsymbol{\theta}}\psi(\bar{\boldsymbol{\theta}}(s^0)) + \nabla_{\boldsymbol{\theta}}\phi(\bar{\boldsymbol{\theta}}(s^0))^\top s^0 = 0. \tag{20}$$

By combining the equations referenced in both Equation (19) and Equation (20), it can be demonstrated that

$$\nabla_{\boldsymbol{\theta}}\phi(\bar{\boldsymbol{\theta}}(s^0))^\top\mathbb{E}_\pi\left[\bar{S}\left(\bar{\boldsymbol{\theta}}\left(s^0\right);\mathbf{x}\right)\right] = \nabla_{\boldsymbol{\theta}}\psi(\bar{\boldsymbol{\theta}}(s^0)).$$

In accordance with the definition of $g$ as the minorizer of $f$ in Equation (2), namely $f(\boldsymbol{\tau};\mathbf{x}) = g(\boldsymbol{\theta},\mathbf{x};\boldsymbol{\tau})$, where we choose $\boldsymbol{\tau} = \bar{\boldsymbol{\theta}}(s^0)$, and Assumption 1.A1, we obtain that

$$\nabla_{\boldsymbol{\theta}}\mathbb{E}_\pi\left[f(\bar{\boldsymbol{\theta}}(s^0);\mathbf{x})\right] = -\nabla_{\boldsymbol{\theta}}\psi(\bar{\boldsymbol{\theta}}(s^0)) + \nabla_{\boldsymbol{\theta}}\phi(\bar{\boldsymbol{\theta}}(s^0))^\top\mathbb{E}_\pi\left[\bar{S}(\bar{\boldsymbol{\theta}}(s^0);\mathbf{x})\right] = 0.$$

This implies that $\bar{\boldsymbol{\theta}}(s^0)$ belongs to the stationary points of $\mathbb{E}_\pi[f(\boldsymbol{\theta};\mathbf{x})]$, that is, $\bar{\boldsymbol{\theta}}(s^0) \in \mathbb{L}$. Hence, $s^0 = \mathbb{E}_\pi\left[\bar{S}\left(\boldsymbol{\theta}^0;\mathbf{x}\right)\right] \in \mathbb{F}$ via Proposition 1.

(c) For any $s \in \mathbb{S}$, the functions $s \mapsto \langle\nabla_s V(s), \eta(s)\rangle$ is continuous via using Assumption 2.A5 and Proposition 2 (a). Hence, the extreme value theorem implies that there exists $s_{\max} \in \mathbb{K}$ such that $\sup_{s\in\mathbb{K}}\langle\nabla_s V(s),\eta(s)\rangle = \langle\nabla_s V(s_{\max}), h(s_{\max})\rangle < 0$. The last inequality is due to the fact that $s_{\max} \notin \mathbb{F}$ since $s_{\max} \in \mathbb{K} \subset \mathbb{S}\setminus\mathbb{F}$ and Proposition 2 (b). Hence, for all compact subsets $\mathbb{K} \subset \mathbb{S}\setminus\mathbb{F}$, it follows that:

$$\sup_{s\in\mathbb{K}}\langle\nabla_s V(s),\eta(s)\rangle < 0.$$

### A.3 Proof of Proposition 3

Denote:

$$\eta(s) := \mathbb{E}_\pi\left[\bar{S}\left(\bar{\boldsymbol{\theta}}\left(s\right);\mathbf{x}\right)\right] - s, \qquad \xi_{n+1} = \bar{S}\left(\bar{\boldsymbol{\theta}}\left(s_n\right);\boldsymbol{x}_{n+1}\right) - \mathbb{E}_\pi\left[\bar{S}\left(\bar{\boldsymbol{\theta}}\left(s_n\right);\mathbf{x}\right)\right].$$

Thus, our algorithm Equation (5) can be rewritten as:

$$s_{n+1} = s_n + \gamma_{n+1}\left\{\eta(s_n) + \xi_{n+1}\right\}. \tag{21}$$

There exists a compact $\mathbb{K} \subset \mathbb{S}$ and $n$ for every $\varepsilon > 0$ under the given assumptions, such that $\mathbb{P}\left(\bigcap_{k\geq n}\{s_n \in \mathbb{K}\}\right) \geq 1-\varepsilon$. Therefore, for any $\kappa > 0$,

$$\begin{aligned}\mathbb{P}\left(\sup_{k\geq n}\left|\sum_{i=n}^{k}\gamma_i\xi_i\right| \geq \kappa\right) &\leq \mathbb{P}\left(\sup_{k\geq n}\left|\sum_{i=n}^{k}\gamma_i\xi_i\mathbf{1}_{\mathrm{s_i}\in\mathbb{K}}\right| \geq \kappa, \bigcap_{i\geq n}\{s_i\in\mathbb{K}\}\right) + \mathbb{P}\left(\bigcup_{i\geq n}\{s_i\notin\mathbb{K}\}\right)\\ &\leq \varepsilon + \mathbb{P}\left(\sup_{k\geq n}\left|\sum_{i=n}^{k}\gamma_i\xi_i\mathbf{1}_{\mathrm{s_i}\in\mathbb{K}}\right| \geq \kappa\right).\end{aligned}$$

Note that $M_{n,k} = \sum_{i=n}^{k}\gamma_i\xi_i\mathbf{1}_{\mathrm{s_i}\in\mathbb{K}}$ is an $L_2$-martingale. Its angle bracket is calculated as:

$$\begin{aligned}\sum_{i=n}^{k}\mathbb{E}_\pi\left[(M_{n,i}-M_{n,i-1})^2\right] &= \sum_{i=n}^{k}\mathbb{E}_\pi\left[(\gamma_i\xi_i)^2\mathbf{1}_{\mathrm{s_i}\in\mathbb{K}}\right] = \sum_{i=n}^{k}\gamma_i^2\mathbb{E}_\pi\left[\xi_i^2\right]\mathbf{1}_{\mathrm{s_i}\in\mathbb{K}}\\ &= \sum_{i=n}^{k}\gamma_i^2\mathrm{Var}\left[\bar{S}\left(\bar{\boldsymbol{\theta}}\left(s_i\right);\mathbf{x}\right)\right]\mathbf{1}_{\mathrm{s_i}\in\mathbb{K}} < \sum_{i=n}^{k}\gamma_i^2\mathbb{E}_\pi\left[\left|\bar{S}\left(\bar{\boldsymbol{\theta}}\left(s_i\right);\mathbf{x}\right)\right|^2\right]\mathbf{1}_{\mathrm{s_i}\in\mathbb{K}}\\ &< \sup_{s\in\mathbb{K}}\left(\mathbb{E}_\pi\left[\left|\bar{S}\left(\bar{\boldsymbol{\theta}}\left(s\right);\mathbf{x}\right)\right|^2\right]\sum_{i=n}^{k}\gamma_i^2\right) < \infty.\end{aligned}$$

By means of the Chebyshev inequality in connection with the Doob martingale inequality, we conclude that

$$\mathbb{P}\left(\sup_{k\geq n}|M_{n,k}| \geq \kappa\right) \leq 2\kappa^{-2}\sup_{s\in\mathbb{K}}\mathbb{E}_\pi\left[\left|\bar{S}\left(\bar{\boldsymbol{\theta}}\left(s\right);\mathbf{x}\right)\right|^2\right]\sum_{i=n}^{\infty}\gamma_i^2.$$

Next we derive the following Lemma 1, which is proved in Appendix B.1, to show that: $\lim_{n\to\infty}\sum_{i=n}^{\infty}\gamma_i^2 \to 0$.

**Lemma 1.** *Let* $(a)_{k=0}^{\infty}$ *be a sequence of positive real number such that:* $\sum_{i=0}^{\infty} a_i < c < \infty$*, then we have* $\lim_{n\to\infty} \sum_{i=n}^{\infty} a_i \to 0$.

Furthermore, notice that due to Assumption 2.A7, we have:

$$\sup_{s\in\mathbb{K}} \mathbb{E}_\pi \left[ \left| \bar{S}\left(\bar{\boldsymbol{\theta}}\left(s\right); \mathbf{x}\right) \right|^2 \right] < \sup_{s\in\mathbb{K}} \mathbb{E}_\pi \left[ \left| \bar{S}\left(\bar{\boldsymbol{\theta}}\left(s\right); \mathbf{x}\right) \right|^p \right] + 1 < \infty. \tag{22}$$

Thus, by applying Lemma 1 and Equation (22), we can show that:

$$\limsup_n \sup_{k\geq n} |M_{n,k}| = 0 \quad \text{with probability 1.}$$

The proof follows from Theorem 2.3 in [1], which establishes that the sequence of sufficient statistics, as defined by Equation (21), satisfies

$$\limsup_n d(s_n, \mathbb{F}) = 0 \quad \text{almost surely.}$$

For the second half of the problem, from condition [A6], we get $\bar{\boldsymbol{\theta}}\left(s_n\right)$ is continuous. Furthermore, since $\lim_{n\to\infty} s_n = s^* \in \mathbb{F}$ with probability 1, we get

$$\lim_{n\to\infty} \boldsymbol{\theta}_n = \lim_{n\to\infty} \bar{\boldsymbol{\theta}}\left(s_n\right) \to \bar{\boldsymbol{\theta}}\left(s^*\right) \text{ with probability 1.}$$

On the other hand, from Proposition 1 we get $\bar{\boldsymbol{\theta}}\left(s^*\right) \in \mathbb{L}$ which conclude:

$$\lim_{n\to\infty} \left\{ d(\boldsymbol{\theta}_n, \mathbb{L}) \right\} = 0.$$

### A.4 Proof of Theorem 1

#### A.4.1 Proof of Theorem 1 (a)

This is shown by an expansion of the Taylor series with an integral remainder:

$$\begin{aligned} \boldsymbol{\theta}_{n+1} &= \bar{\boldsymbol{\theta}}\left[ s_n + \gamma_{n+1} \left\{ \bar{S}(\boldsymbol{\theta}_n; \boldsymbol{x}_{n+1}) - s_n \right\} \right] \\ &= \boldsymbol{\theta}_n + \gamma_{n+1} \left( \nabla_s \bar{\boldsymbol{\theta}}^\top(s_n) \right)^\top \left\{ \bar{S}(\boldsymbol{\theta}_n; \boldsymbol{x}_{n+1}) - s_n \right\} + \gamma_{n+1} \boldsymbol{r}_{n+1}. \end{aligned} \tag{23}$$

Here the remainder $\boldsymbol{r}_{n+1}$ is given by:

$$\boldsymbol{r}_{n+1}^\top := \left\{ \bar{S}(\boldsymbol{\theta}_n; \boldsymbol{x}_{n+1}) - s_n \right\}^\top \int_0^1 \left( \nabla_s \bar{\boldsymbol{\theta}}^\top \left[ s_n + \gamma_{n+1} t \{ \bar{S}(\boldsymbol{\theta}_n; \boldsymbol{x}_{n+1}) - s_n \} \right] - \nabla_s \bar{\boldsymbol{\theta}}^\top(s_n) \right) dt.$$

We will first demonstrate that $\lim_{n\to\infty} \boldsymbol{r}_{n+1} = 0$ almost surely, using Lemma 2, which is proved in Appendix B.2.

**Lemma 2.** *Let* $p \geq 1$*. Assume that, for any compact subset* $\mathcal{K} \subset \mathbb{S}$*,*

$$\sup_{s\in\mathcal{K}} \left( \mathbb{E}_\pi [ | \bar{S}\left(\bar{\boldsymbol{\theta}}\left(s\right); \mathbf{x}\right) |^p ] \right) < \infty$$

*for some* $p > 0$*, and that* $\mathbb{P}_\pi$ *almost surely* $\lim(s_n)$ *exists and belongs to* $\mathbb{S}$*. Then, the sequence* $\{ \bar{S}\left(\bar{\boldsymbol{\theta}}\left(s_n\right); \boldsymbol{x}_{n+1}\right) \}_{n\geq 0}$ *is bounded in probability,* i.e.,

$$\lim_{M\to\infty} \limsup_{n\to\infty} \mathbb{P}\big( | \bar{S}\left(\bar{\boldsymbol{\theta}}\left(s_n\right); \boldsymbol{x}_{n+1}\right) | \geq M \big) = 0.$$

Lemma 2 demonstrates that the sequence $\{\bar{S}(\boldsymbol{\theta}_n; \boldsymbol{x}_{n+1})\}_{n\geq 0}$ is bounded in probability, which is denoted as $\bar{S}(\boldsymbol{\theta}_n; \boldsymbol{x}_{n+1}) = \mathbf{O}_{\mathbb{P}}(1)$. Under the assumption stated in Proposition 3, we have $s_n = \mathbf{O}_{\mathbb{P}}(1)$, which leads to the conclusion that $\bar{S}(\boldsymbol{\theta}_n; \boldsymbol{x}_{n+1}) - s_n = \mathbf{O}_{\mathbb{P}}(1)$. Let $\epsilon > 0$ be arbitrary, and choose a compact set $\mathcal{K}$ and a constant $M$ large enough such that

$$\limsup_{n\to\infty} \mathbb{P}(s_n \notin \mathcal{K}) + \limsup_{n\to\infty} \mathbb{P}(|\bar{S}(\boldsymbol{\theta}_n; \boldsymbol{x}_{n+1}) - s_n| \geq M) \leq \epsilon. \tag{24}$$

Since $\nabla_s \bar{\boldsymbol{\theta}}(\cdot)$ is continuous, it is also uniformly continuous over any compact subset. Specifically, there exists a constant $\delta_0$ such that the set $\mathcal{K}_{\delta_0} = \{s \in \mathbb{S} : d(s, \mathcal{K}) \leq \delta_0\} \subset \mathbb{S}$ satisfies

$$\sup_{|h| \leq \delta_0, s \in \mathcal{K}} |\nabla_s \bar{\boldsymbol{\theta}}(s+h) - \nabla_s \bar{\boldsymbol{\theta}}(s)| \leq \epsilon. \tag{25}$$

Since $\lim_{n\to\infty} \gamma_n = 0$ and $\bar{S}(\boldsymbol{\theta}_n; \boldsymbol{x}_{n+1}) - s_n$ is bounded in probability, the term $\gamma_{n+1}\{\bar{S}(\boldsymbol{\theta}_n; \boldsymbol{x}_{n+1}) - s_n\}$ converges to 0 in probability, which is expressed as $\gamma_{n+1}\{\bar{S}(\boldsymbol{\theta}_n; \boldsymbol{x}_{n+1}) - s_n\} = \mathbf{o}_{\mathbb{P}}(1)$. For any $\delta_0 > 0$ which satisfies the inequality Equation (25), it follows that

$$\lim_{n\to\infty} \mathbb{P}\left(\gamma_{n+1}|\bar{S}(\boldsymbol{\theta}_n; \boldsymbol{x}_{n+1}) - s_n| \geq \delta_0\right) = 0.$$

Hence, we obtain $\boldsymbol{r}_n = \mathbf{O}_{\mathbb{P}}(1)$ due to the fact that

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{P}(|\boldsymbol{r}_{n+1}| \geq M\epsilon) \leq & \limsup_{n\to\infty} \mathbb{P}\left(\gamma_{n+1}|\bar{S}(\boldsymbol{\theta}_n; \boldsymbol{x}_{n+1}) - s_n| \geq \delta_0\right) + \limsup_{n\to\infty} \mathbb{P}(s_n \notin \mathcal{K}) \\ & + \limsup_{n\to\infty} \mathbb{P}\left(|\bar{S}(\boldsymbol{\theta}_n; \boldsymbol{x}_{n+1}) - s_n| \geq M\right) \leq \epsilon.\end{aligned}$$

Let us continue with the first term in Equation (23). Note that Equation (17) gives us

$$\nabla_s \bar{\boldsymbol{\theta}}(s)^\top = -[\nabla_{\boldsymbol{\theta}} \phi(\bar{\boldsymbol{\theta}}(s))^\top]^\top \{\nabla^2_{\boldsymbol{\theta}}[h(s; \boldsymbol{\theta})|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}(s)}]\}^{-1}. \tag{26}$$

Here we use he definition of $h(s, \boldsymbol{\theta})$ in Equation (15). According to Equation (2) we have:

$$\nabla_{\boldsymbol{\theta}} f(\bar{\boldsymbol{\theta}}(s); \mathbf{x}) + \nabla_{\boldsymbol{\theta}} \psi(\bar{\boldsymbol{\theta}}(s)) - \nabla_{\boldsymbol{\theta}} \phi(\bar{\boldsymbol{\theta}}(s))^\top \bar{S}(\bar{\boldsymbol{\theta}}(s); \mathbf{x}) = 0.$$

Combining this with Equation (11) , we can show that

$$\nabla_{\boldsymbol{\theta}} \phi(\bar{\boldsymbol{\theta}}(s))^\top \left[\bar{S}(\bar{\boldsymbol{\theta}}(s); \mathbf{x}) - s\right] = \nabla_{\boldsymbol{\theta}} f(\bar{\boldsymbol{\theta}}(s); \mathbf{x}). \tag{27}$$

Hence, by coupling Equation (26) and Equation (27), the first-order term in Equation (23) can be expressed as

$$\begin{aligned}(\nabla_s \bar{\boldsymbol{\theta}}^\top(s_n))^\top [\bar{S}(\boldsymbol{\theta}_n; \boldsymbol{x}_{n+1}) - s_n] = & \left[-\{\nabla^2_{\boldsymbol{\theta}}[h(s; \boldsymbol{\theta})|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}(s)}]\}^{-1} - [\boldsymbol{I}_\pi(\boldsymbol{\theta}_n)]^{-1}\right] \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_n; \boldsymbol{x}_{n+1}) \\ & + [\boldsymbol{I}_\pi(\boldsymbol{\theta}_n)]^{-1} \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_n; \boldsymbol{x}_{n+1}).\end{aligned} \tag{28}$$

By definition, the term $\boldsymbol{I}_\pi$ can be reformulated as

$$\boldsymbol{I}_\pi(\boldsymbol{\theta}) = -\nabla^2_{\boldsymbol{\theta}} h(s; \boldsymbol{\theta})\big|_{s=\mathbb{E}_\pi[\bar{s}(\boldsymbol{\theta}; \mathbf{x})]}.$$

It is important to note that $\boldsymbol{\theta}_n$ converges almost surely to $\boldsymbol{\theta}^0$ and that $\boldsymbol{I}_\pi(\boldsymbol{\theta}^0)$ is assumed to be positive definite. Consequently, to justify neglecting the term in square brackets in Equation (28), it is necessary to demonstrate that

$$\nabla^2_{\boldsymbol{\theta}} h(s_n; \boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_n} - \nabla^2_{\boldsymbol{\theta}} h(s; \boldsymbol{\theta})\big|_{(s,\boldsymbol{\theta})=(\mathbb{E}[\bar{S}(\boldsymbol{\theta}_n; \boldsymbol{x}_{n+1}|\mathcal{F}_n)], \boldsymbol{\theta}_n)} = \mathbf{o}_{\mathbb{P}}(1). \tag{29}$$

Given the continuity of the function $(s, \boldsymbol{\theta}) \mapsto \nabla^2_{\boldsymbol{\theta}} h(s; \boldsymbol{\theta})$, there exists a $\delta_1 > 0$ such that $\mathcal{K}_{\delta_1} = \{s \in \mathbb{S} : d(s, \mathcal{K}) \leq \delta_1\} \subset \mathbb{S}$, and

$$\sup_{|h| \leq \delta_1, s \in \mathcal{K}} \left|\nabla^2_{\boldsymbol{\theta}} h\{s+h; \bar{\boldsymbol{\theta}}(s)\} - \nabla^2_{\boldsymbol{\theta}} h\{s; \bar{\boldsymbol{\theta}}(s)\}\right| \leq \epsilon,$$

where the set $\mathcal{K}$ is as specified in Equation (24). Under the given assumption, it follows that $\lim_{n\to\infty} d(s_n, \mathbb{L}) = 0$ almost surely, which leads to the conclusion that

$$\lim_{n\to\infty} \{\eta(s_n)\} = \lim_{n\to\infty} \left(\mathbb{E}\left[\bar{S}\left(\boldsymbol{\theta}_n; \boldsymbol{x}_{n+1}|\mathcal{F}_n\right)\right] - s_n\right) = 0, \quad \text{almost surely.}$$

By combining these two results, we obtain

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{P}\Big(\left|\nabla^2_{\boldsymbol{\theta}} h(s_n; \boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_n} - \nabla^2_{\boldsymbol{\theta}} h(s; \boldsymbol{\theta})\big|_{(s,\boldsymbol{\theta})=(\mathbb{E}[\bar{S}(\boldsymbol{\theta}_n; \boldsymbol{x}_{n+1}|\mathcal{F}_n)], \boldsymbol{\theta}_n)}\right| \geq \epsilon\Big) \\ \leq \limsup_{n\to\infty} \mathbb{P}(s_n \notin \mathcal{K}) + \limsup_{n\to\infty} \mathbb{P}(|\eta(s_n)| \geq \delta_0) \leq \epsilon,\end{aligned}$$

thereby establishing Equation (29).

### A.4.2 Proof of Theorem 1 (b)

Since $\boldsymbol{\theta}^0$ is a (possibly) local maximum of $\mathbb{E}_\pi\left[f(\boldsymbol{\theta};\mathbf{x})\right]$, we have:

$$\nabla^2_{\boldsymbol{\theta}}\mathbb{E}_\pi\left[f(\boldsymbol{\theta}^*;\mathbf{x})\right] \prec 0.$$

Furthermore, since $\nabla^2_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^0, \boldsymbol{x})$ is symmetric, we have $\boldsymbol{I}_\pi(\boldsymbol{\theta}^0)$ is also symmetric. Hence, for any non-zero vector $\boldsymbol{h}$ we have:

$$\boldsymbol{h}\boldsymbol{I}_\pi(\boldsymbol{\theta}^0)^{-1/2}\nabla^2_{\boldsymbol{\theta}}\mathbb{E}_\pi\left[f(\boldsymbol{\theta};\mathbf{x})\right]\boldsymbol{I}_\pi(\boldsymbol{\theta}^0)^{-1/2}\boldsymbol{h}^\top = (\boldsymbol{h}\boldsymbol{I}_\pi(\boldsymbol{\theta}^0)^{-1/2})\nabla^2_{\boldsymbol{\theta}}\mathbb{E}_\pi\left[f(\boldsymbol{\theta};\mathbf{x})\right](\boldsymbol{h}\boldsymbol{I}_\pi(\boldsymbol{\theta}^0)^{-1/2})^\top < 0.$$

This show us that $\left[\boldsymbol{I}_\pi(\boldsymbol{\theta}^0)\right]^{-1/2}\nabla^2\mathbb{E}_\pi\left[f(\boldsymbol{\theta};\mathbf{x})\right]\left[\boldsymbol{I}_\pi(\boldsymbol{\theta}^0)\right]^{-1/2} \prec 0$. This implies that all eigenvalues of $\left[\boldsymbol{I}_\pi(\boldsymbol{\theta}^0)\right]^{-1/2}\nabla^2\mathbb{E}_\pi\left[f(\boldsymbol{\theta};\mathbf{x})\right]\left[\boldsymbol{I}_\pi(\boldsymbol{\theta}^0)\right]^{-1/2}$ are real and strictly negative.

On the other hand, since

$$\boldsymbol{I}_\pi(\boldsymbol{\theta}^0)^{-1}\nabla^2\mathbb{E}_\pi\left[f(\boldsymbol{\theta};\mathbf{x})\right] = \left[\boldsymbol{I}_\pi(\boldsymbol{\theta}^0)\right]^{1/2}\left\{\left[\boldsymbol{I}_\pi(\boldsymbol{\theta}^0)\right]^{-1/2}\nabla^2\mathbb{E}_\pi\left[f(\boldsymbol{\theta};\mathbf{x})\right]\left[\boldsymbol{I}_\pi(\boldsymbol{\theta}^0)\right]^{-1/2}\right\}\left[\boldsymbol{I}_\pi(\boldsymbol{\theta}^0)\right]^{-1/2},$$

it follows that $\boldsymbol{I}_\pi(\boldsymbol{\theta}^0)^{-1}\nabla^2\mathbb{E}_\pi\left[f(\boldsymbol{\theta};\mathbf{x})\right]$ and $\left[\boldsymbol{I}_\pi(\boldsymbol{\theta}^0)\right]^{-1/2}\nabla^2\mathbb{E}_\pi\left[f(\boldsymbol{\theta};\mathbf{x})\right]\left[\boldsymbol{I}_\pi.(\boldsymbol{\theta}^0)\right]^{-1/2}$ are similar matrices. As a result, they share the same eigenvalues, including multiplicities. Combining this fact with the results established above leads to the desired conclusion.

### A.4.3 Proof of Theorem 1 (c)

We use the definition of the recursive MM sequence that was given by Theorem 1 (a). The mean field that is associated with this sequence is given by

$$\eta_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \mathbb{E}_\pi\left\{\left[\boldsymbol{I}_\pi(\boldsymbol{\theta})\right]^{-1}\nabla_{\boldsymbol{\theta}} f(\mathbf{x};\boldsymbol{\theta})\right\} = \left[\boldsymbol{I}_\pi(\boldsymbol{\theta})\right]^{-1}\left\{\nabla_{\boldsymbol{\theta}}\mathbb{E}_\pi\left[f(\mathbf{x};\boldsymbol{\theta})\right]\right\}.$$

Thus, the Jacobian of this vector field at $\boldsymbol{\theta}^0$ is equal to $\boldsymbol{H}(\boldsymbol{\theta}^0)$.

By directly applying Theorem 1 and its corresponding remarks in [90], the proof of the second part of Theorem 1 follows.

## A.5 Proof of Theorem 2

**Exponential family minorizer surrogate construction.** Using key inequalities essential for constructing surrogate functions, as outlined in Appendix C, and following the procedure detailed in Section 2.2, we derive the exponential family minorizer surrogate function for Equation (10). This is formally presented in Proposition 4 and its proof is provided in Appendix A.5.1.

**Proposition 4** (Exponential family minorizer surrogate construction)**.** *Given the current estimate values after $t$ iterations of the algorithm as follows:*

$$\tau_{kn}^{(t)} = \frac{\mathrm{g}_k(\boldsymbol{w}(\boldsymbol{x}_n))^{(t)}\mathcal{N}(\boldsymbol{y}_n, \boldsymbol{v}_k(\boldsymbol{x}_n), \boldsymbol{\Sigma}_k)^{(t)}}{\sum_{l=1}^{K}\mathrm{g}_l(\boldsymbol{w}(\boldsymbol{x}_n))^{(t)}\mathcal{N}(\boldsymbol{y}_n; \boldsymbol{v}_l(\boldsymbol{x}_n), \boldsymbol{\Sigma}_l)^{(t)}},$$

*these serve as the starting point for computing updates in the $(t+1)$-th iteration, utilizing the constructed surrogate function to refine parameter estimates and improve convergence. Hence, it holds that*

$$\begin{aligned}
-L(\boldsymbol{\theta}) &\leq \sum_{n=1}^{N}\sum_{k=1}^{K}\tau_{kn}^{(t)}log(\tau_{kn}^{(t)}) - \sum_{n=1}^{N}\sum_{k=1}^{K}\tau_{kn}^{(t)}\left[log(\mathrm{g}_k(\boldsymbol{w}(\boldsymbol{x}_n))) + log(\mathcal{N}(\boldsymbol{y}_n; \boldsymbol{v}_k(\boldsymbol{x}_n), \boldsymbol{\Sigma}_k))\right] \\
&\leq \sum_{n=1}^{N}\sum_{k=1}^{K}\tau_{kn}^{(t)}log(\tau_{kn}^{(t)}) + \sum_{n=1}^{N}f_n(\boldsymbol{\omega}^{(t)}) - \sum_{n=1}^{N}\boldsymbol{\omega}^{(t)\top}\nabla f_n(\boldsymbol{\omega}^{(t)}) - \sum_{n=1}^{N}\sum_{k=1}^{K}\tau_{kn}^{(t)}w_k(\boldsymbol{x}_n) \\
&+ \sum_{n=1}^{N}\left\{\boldsymbol{\omega}^\top\nabla f_n(\boldsymbol{\omega}^{(t)}) + \frac{1}{2}(\boldsymbol{\omega}-\boldsymbol{\omega}^{(t)\top})\left(\left(1.5\boldsymbol{I}_{K-1} - \frac{\mathbf{1}\mathbf{1}^\top}{K-1}\right)/2 \otimes \hat{\boldsymbol{x}}_n\hat{\boldsymbol{x}}_n^\top\right)(\boldsymbol{\omega}-\boldsymbol{\omega}^{(t)})\right\} \\
&- \sum_{n=1}^{N}\sum_{k=1}^{K}\tau_{kn}^{(t)}\left[log(\phi(\boldsymbol{y}_n, \boldsymbol{v}_k(\boldsymbol{x}_n), \boldsymbol{\Sigma}_k))\right] := -\log[g(\boldsymbol{\theta}, \boldsymbol{x}_{[N]}, \boldsymbol{y}_{[N]}; \boldsymbol{\theta}^{(t)})].
\end{aligned} \tag{30}$$

### A.5.1 Proof of surrogate function construction in Proposition 4

Now we go into the main part of finding the surrogate function of Equation (10) via Equation (30). By applying Lemma 7 for the function $f = -log$ and:

$$\begin{aligned}
\boldsymbol{c} &= [1, 1, \ldots, 1]^\top, \\
\boldsymbol{x} &= [\mathrm{g}_k(\boldsymbol{w}(\boldsymbol{x}_i))\mathcal{N}(\boldsymbol{y}_i; \boldsymbol{v}_k(\boldsymbol{x}_i), \boldsymbol{\Sigma}_k)]_{k\in[K]}^\top, \\
\boldsymbol{y} &= [\mathrm{g}_k(\boldsymbol{w}(\boldsymbol{x}_i))^{(t)}\mathcal{N}(\boldsymbol{y}_i; \boldsymbol{v}_k(\boldsymbol{x}_i), \boldsymbol{\Sigma}_k)^{(t)}]_{k\in[K]}^\top.
\end{aligned}$$

We obtain:

$$-\log\left[\sum_{k=1}^{K}\mathrm{g}_k(\boldsymbol{w}(\boldsymbol{x}_i))\mathcal{N}(\boldsymbol{y}_i; \boldsymbol{v}_k(\boldsymbol{x}_i), \boldsymbol{\Sigma}_k)\right] \leq -\sum_{k=1}^{K}\tau_{kn}^{(t)} log\left[\frac{1}{\tau_{kn}^{(t)}}\mathrm{g}_k(\boldsymbol{w}(\boldsymbol{x}_i))\mathcal{N}(\boldsymbol{y}_i; \boldsymbol{v}_k(\boldsymbol{x}_i), \boldsymbol{\Sigma}_k)\right]$$

where $\tau_{kn}^{(t)} = \dfrac{\mathrm{g}_k(\boldsymbol{w}(\boldsymbol{x}_n))^{(t)}\mathcal{N}(\boldsymbol{y}_n, \boldsymbol{v}_k(\boldsymbol{x}_n), \boldsymbol{\Sigma}_k)^{(t)}}{\sum_{l=1}^{K}\mathrm{g}_l(\boldsymbol{w}(\boldsymbol{x}_n))^{(t)}\mathcal{N}(\boldsymbol{y}_n; \boldsymbol{v}_l(\boldsymbol{x}_n), \boldsymbol{\Sigma}_l)^{(t)}}$.

Hence:

$$\begin{aligned}
-L(\boldsymbol{\theta}) &\leq -\sum_{n=1}^{N}\sum_{k=1}^{K}\tau_{kn}^{(t)} log\left[\frac{1}{\tau_{kn}^{(t)}}\mathrm{g}_k(\boldsymbol{w}(\boldsymbol{x}_n))\mathcal{N}(\boldsymbol{y}_n; \boldsymbol{v}_k(\boldsymbol{x}_n), \boldsymbol{\Sigma}_k)\right] \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K}\tau_{kn}^{(t)} log(\tau_{kn}^{(t)}) - \sum_{n=1}^{N}\sum_{k=1}^{K}\tau_{kn}^{(t)}\left[log(\mathrm{g}_k(\boldsymbol{w}(\boldsymbol{x}_n))) + log(\mathcal{N}(\boldsymbol{y}_n; \boldsymbol{v}_k(\boldsymbol{x}_n), \boldsymbol{\Sigma}_k))\right].
\end{aligned}$$

### A.5.2 Exponential family minorizer surrogate for the online MM algorithm

Denote:

$$f(\boldsymbol{\omega}) = \log\left(\sum_{k=1}^{K}\exp\left(w_k(\boldsymbol{x})\right)\right) \text{ where } w_r(\boldsymbol{x}) = \sum_{r=0}^{D_W}\boldsymbol{x}^r\omega_{kr}.$$

By applying the Taylor series we have:

$$f(\boldsymbol{\omega}) - f(\boldsymbol{\omega}^{(t)}) = (\boldsymbol{\omega} - \boldsymbol{\omega}^{(t)})^\top\nabla f(\boldsymbol{\omega}^{(t)}) + \frac{1}{2}(\boldsymbol{\omega} - \boldsymbol{\omega}^{(t)})^\top\nabla^2 f(\boldsymbol{\omega}^{(t)} - \alpha(\boldsymbol{\omega} - \boldsymbol{\omega}^{(t)}))(\boldsymbol{\omega} - \boldsymbol{\omega}^{(t)}). \tag{31}$$

### A.5.3 Calculate the gradient of $f$

For some $k \in [K]$ and $r \in 0, 1, \ldots, D_W$, the gradient can be calculate as follow:

$$\nabla_{\omega_{kr}} f = \frac{\boldsymbol{x}^r\exp\left(\sum_{r=0}^{D_W}\boldsymbol{x}^r\omega_{kr}\right)}{\sum_{l=1}^{K}\exp\left(\sum_{r=0}^{D_W}\boldsymbol{x}^r\omega_{lr}\right)} = \boldsymbol{x}^r\mathrm{g}_k(\boldsymbol{w}(\boldsymbol{x})).$$

Hence, the gradient of $f$ is given as:

$$\nabla f = \left[\boldsymbol{x}^0 g_1(\boldsymbol{w}(\boldsymbol{x})), \ldots, \boldsymbol{x}^0\mathrm{g}_K(\boldsymbol{w}(\boldsymbol{x})), \ldots, \boldsymbol{x}^{D_W} g_1(\boldsymbol{w}(\boldsymbol{x})), \ldots, \boldsymbol{x}^{D_W}\mathrm{g}_K(\boldsymbol{w}(\boldsymbol{x}))\right]^\top = \hat{\boldsymbol{x}} \otimes \hat{\boldsymbol{g}}.$$

where $\hat{\boldsymbol{x}} = \left[\boldsymbol{x}^0, \ldots, \boldsymbol{x}^{D_W}\right]^\top$, $\hat{\boldsymbol{g}} = [g_1(\boldsymbol{w}(\boldsymbol{x})), \ldots, \mathrm{g}_K(\boldsymbol{w}(\boldsymbol{x}))]^\top$ and $\otimes$ is the Kronecker product.

### A.5.4 Calculate for the Hessian the gradient of $f$

For some $k_1, k_2 \in [K]$ and $r_1, r_2 \in \{0, 1, \ldots, D_W\}$, the gradient can be calculate as follow:

$$\nabla_{\omega_{k_1 r_1}\omega_{k_2 r_2}} f = \frac{-\boldsymbol{x}^{r_1}\exp(w_{k_1}(\boldsymbol{x}))\boldsymbol{x}^{r_2}\exp(w_{k_2}(\boldsymbol{x}))}{\left[\sum_{l=1}^{K}\exp(w_l(\boldsymbol{x}))\right]^2} = -\boldsymbol{x}^{r_1+r_2} g_{k_1}(\boldsymbol{w}(\boldsymbol{x}))g_{k_2}(\boldsymbol{w}(\boldsymbol{x}))$$

for $k_1 \neq k_2$ and

$$\begin{aligned}\nabla_{\omega_{k_1 r_1}\omega_{k_1 r_2}} f &= \frac{\boldsymbol{x}^{r_1+r_2}\exp(w_{k_1}(\boldsymbol{x}))\left(\sum_{l=1}^{K}\exp(w_l(\boldsymbol{x})) - \exp(w_{k_1}(\boldsymbol{x}))\right)}{\left[\sum_{l=1}^{K}\exp(w_l(\boldsymbol{x}))\right]^2} \\ &= \boldsymbol{x}^{r_1+r_2} g_{k_1}(\boldsymbol{w}(\boldsymbol{x}))(1 - g_{k_1}(\boldsymbol{w}(\boldsymbol{x}))).\end{aligned}$$

Hence, the Hessian is given as:

$$\nabla^2 f = \left(\boldsymbol{\Lambda} - \hat{\boldsymbol{g}}\hat{\boldsymbol{g}}^\top\right) \otimes \hat{\boldsymbol{x}}\hat{\boldsymbol{x}}^\top \text{ where } \boldsymbol{\Lambda} = \operatorname{diag}(\hat{\boldsymbol{g}}).$$

**Lemma 3** (See *e.g.,* [4])**.** *If $\boldsymbol{A} \leq \boldsymbol{B}$ then for symmetric, nonnegative definite $\boldsymbol{C}$:*

$$\boldsymbol{A} \otimes \boldsymbol{C} \leq \boldsymbol{B} \otimes \boldsymbol{C}.$$

**Proposition 5.** *Let $p_i$ be be real values in $(0, 1)$ for $i \in [N+1]$ and $\sum_{i=1}^{N+1} p_i = 1$. Denote $\hat{\boldsymbol{p}} = [p_1, p_2, \ldots, p_N]^\top$. We have:*

$$\boldsymbol{\Lambda}_{\boldsymbol{p}} - \hat{\boldsymbol{p}}\hat{\boldsymbol{p}}^\top \leq (1.5\boldsymbol{I}_N - \boldsymbol{1}\boldsymbol{1}^\top/N)/2.$$

The proofs of Lemma 3 are provided in Appendix B.3, while the proofs of Proposition 5 are detailed in Appendix B.4.

Applying Lemma 3 and Proposition 5 to Equation (31) we receive the majorization:

$$\begin{aligned}f(\boldsymbol{\omega}) \leq\ & f(\boldsymbol{\omega}^{(t)}) + (\boldsymbol{\omega} - \boldsymbol{\omega}^{(t)})^\top \nabla f(\boldsymbol{\omega}^{(t)}) \\ & + \frac{1}{2}(\boldsymbol{\omega} - \boldsymbol{\omega}^{(t)})^\top \left(\left(1.5\boldsymbol{I}_{K-1} - \frac{\boldsymbol{1}\boldsymbol{1}^\top}{K-1}\right)/2 \otimes \hat{\boldsymbol{x}}\hat{\boldsymbol{x}}^\top\right)(\boldsymbol{\omega} - \boldsymbol{\omega}^{(t)}).\end{aligned}$$

Hence, our overall majorization function take the form:

$$\begin{aligned}& -L(\boldsymbol{\theta}) \\ & \leq C_{kn}^{(t)} - \sum_{n=1}^{N}\sum_{k=1}^{K}\tau_{kn}^{(t)} w_k(\boldsymbol{x}_n) \\ & + \sum_{n=1}^{N}\left\{f_n(\boldsymbol{\omega}^{(t)}) + (\boldsymbol{\omega} - \boldsymbol{\omega}^{(t)})^\top \nabla f_n(\boldsymbol{\omega}^{(t)}) + \frac{1}{2}(\boldsymbol{\omega} - \boldsymbol{\omega}^{(t)})^\top \left(\left(1.5\boldsymbol{I}_{K-1} - \frac{\boldsymbol{1}\boldsymbol{1}^\top}{K-1}\right)/2 \otimes \hat{\boldsymbol{x}}_n\hat{\boldsymbol{x}}_n^\top\right)(\boldsymbol{\omega} - \boldsymbol{\omega}^{(t)})\right\} \\ & - \sum_{n=1}^{N}\sum_{k=1}^{K}\tau_{kn}^{(t)}\left[\log(\phi(\boldsymbol{y}_n, \boldsymbol{v}_k(\boldsymbol{x}_n), \sigma_k^2))\right],\end{aligned}$$

where $f_n(\boldsymbol{\omega}) = \log\left(\sum_{k=1}^{K}\exp\left(w_k(\boldsymbol{x}_n)\right)\right)$ and $\boldsymbol{\omega} = [\omega_{10}, \ldots, \omega_{1D_w}, \ldots, \omega_{(K-1)0}, \ldots, \omega_{(K-1)D_W}]^\top$,

$$\begin{aligned}& = \sum_{n=1}^{N}\sum_{k=1}^{K}\tau_{kn}^{(t)}\log(\tau_{kn}^{(t)}) + \sum_{n=1}^{N} f_n(\boldsymbol{\omega}^{(t)}) - \sum_{n=1}^{N}\boldsymbol{\omega}^{(t)\top}\nabla f_n(\boldsymbol{\omega}^{(t)}) - \sum_{n=1}^{N}\sum_{k=1}^{K}\tau_{kn}^{(t)} w_k(\boldsymbol{x}_n) \\ & + \sum_{n=1}^{N}\left\{\boldsymbol{\omega}^\top \nabla f_n(\boldsymbol{\omega}^{(t)}) + \frac{1}{2}(\boldsymbol{\omega} - \boldsymbol{\omega}^{(t)\top})\left(\left(1.5\boldsymbol{I}_{K-1} - \frac{\boldsymbol{1}\boldsymbol{1}^\top}{K-1}\right)/2 \otimes \hat{\boldsymbol{x}}_n\hat{\boldsymbol{x}}_n^\top\right)(\boldsymbol{\omega} - \boldsymbol{\omega}^{(t)})\right\} \\ & - \sum_{n=1}^{N}\sum_{k=1}^{K}\tau_{kn}^{(t)}\left[\log(\phi(\boldsymbol{y}_n, \boldsymbol{v}_k(\boldsymbol{x}_n), \sigma_k^2))\right] := -\log[g(\boldsymbol{\theta}^{(t)}, \boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta})].\end{aligned}$$

### A.6 Derivation of the Algorithm 2

#### A.6.1 Construction of the online MM algorithm for $O_{gate}$

$$O_{gate}(\boldsymbol{\omega}, \boldsymbol{\theta}^{(t)}) = -\sum_{k=1}^{K} \tau_k^{(t)} w_k(\boldsymbol{x}) + \left\{ \boldsymbol{\omega}^\top \nabla f(\boldsymbol{\omega}^{(t)}) + \frac{1}{2}(\boldsymbol{\omega} - \boldsymbol{\omega}^{(t)})^\top \left( \left( 1.5\boldsymbol{I}_{K-1} - \frac{\mathbf{1}\mathbf{1}^\top}{K-1} \right) / 2 \otimes \hat{\boldsymbol{x}}\hat{\boldsymbol{x}}^\top \right) (\boldsymbol{\omega} - \boldsymbol{\omega}^{(t)}) \right\}.$$

Let us denote vector:

$$\boldsymbol{\xi} = [\tau_1^{(t)}\boldsymbol{x}^0, \tau_1^{(t)}\boldsymbol{x}^1, \ldots, \tau_1^{(t)}\boldsymbol{x}^{D_W}, \ldots, \tau_{K-1}^{(t)}\boldsymbol{x}^0, \ldots., \tau_{K-1}^{(t)}\boldsymbol{x}^{D_W}]^\top.$$

Hence:

$$O_{gate}(\boldsymbol{\omega}, \boldsymbol{\theta}^{(t)}) = \left( -\boldsymbol{\xi}^\top + \nabla f(\boldsymbol{\omega}^{(t)})^\top - \boldsymbol{\omega}^{(t)T}\boldsymbol{B} \right) \boldsymbol{\omega} + \frac{1}{2}\boldsymbol{\omega}^\top \boldsymbol{B}\boldsymbol{\omega} + C,$$

$\boldsymbol{B} = \left( 1.5\boldsymbol{I}_{K-1} - \frac{\mathbf{1}\mathbf{1}^\top}{K-1} \right) / 2 \otimes \hat{\boldsymbol{x}}\hat{\boldsymbol{x}}^\top$ and $C$ represents the constants that are independent of $\boldsymbol{\omega}$.

Here the online MM algorithms can be written as:

$$\phi(\boldsymbol{\omega}) := \begin{bmatrix} \boldsymbol{\omega} \\ \text{vec}(\boldsymbol{\omega}\boldsymbol{\omega}^\top) \end{bmatrix}, \ \bar{S}(\boldsymbol{\tau}; \boldsymbol{x}) = \begin{bmatrix} -\boldsymbol{\xi} + \nabla f(\boldsymbol{\tau}) - \boldsymbol{B}\boldsymbol{\tau} \\ \frac{1}{2}\,\text{vec}(\boldsymbol{B}) \end{bmatrix}, \ \boldsymbol{\theta}(\boldsymbol{\omega}) = C.$$

#### A.6.2 Construction of the online MM algorithm for $O_{expert}$

Recall that the expert function to be minimized is given by:

$$O_{expert}(\boldsymbol{\upsilon}, \boldsymbol{\sigma}^2, \boldsymbol{\theta}^{(t)}) = \sum_{k=1}^{K} \left[ \frac{1}{\sigma_k^2} \tau_k^{(t)} \left( \boldsymbol{y} - \boldsymbol{\upsilon}_k^\top \boldsymbol{r} \right)^2 + \tau_k^{(t)} \log \sigma_k^2 \right] \text{ where } \boldsymbol{r} = \left[ 1, \boldsymbol{x}, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^{D_V} \right]^\top.$$

Here the online MM algorithms can be written as:

$$\phi(\boldsymbol{\upsilon}, \boldsymbol{\sigma}) := \begin{bmatrix} \boldsymbol{\kappa} \\ \boldsymbol{\zeta} \\ \boldsymbol{\Delta} \\ \boldsymbol{\Sigma} \end{bmatrix}, \ \bar{S}(\boldsymbol{\tau}; \boldsymbol{x}) = \begin{bmatrix} \boldsymbol{y}^2\boldsymbol{\tau}^{(t)} \\ \boldsymbol{\tau}^{(t)} \otimes (-2\boldsymbol{y}\boldsymbol{r}) \\ \boldsymbol{\tau}^{(t)} \otimes \left( \text{vec}\left( \boldsymbol{r}\boldsymbol{r}^\top \right) \right) \\ \boldsymbol{\tau}^{(t)} \end{bmatrix}, \text{ where}$$

$$\boldsymbol{\Sigma} = [\log \sigma_1^2, \log \sigma_2^2, \ldots, \log \sigma_K^2]^\top, \ \boldsymbol{\kappa} = \left[ \frac{1}{\sigma_1^2}, \frac{1}{\sigma_2^2}, \ldots, \frac{1}{\sigma_K^2} \right]^\top,$$

$$\boldsymbol{\zeta} = \text{vec}\left( \left[ \frac{\boldsymbol{\upsilon_1}}{\sigma_1^2}, \frac{\boldsymbol{\upsilon}_2}{\sigma_2^2}, \ldots, \frac{\boldsymbol{\upsilon}_K}{\sigma_K^2} \right] \right), \ \boldsymbol{\tau}^{(t)} = [\tau_1^{(t)}, \tau_2^{(t)}, \ldots, \tau_K^{(t)}]^\top,$$

$$\boldsymbol{\Delta} = \text{vec}\left( \frac{\text{vec}(\boldsymbol{\upsilon}_1\boldsymbol{\upsilon}_1^\top)}{\sigma_1^2}, \frac{\text{vec}(\boldsymbol{\upsilon}_2\boldsymbol{\upsilon}_2^\top)}{\sigma_2^2}, \ldots, \frac{\text{vec}(\boldsymbol{\upsilon}_K\boldsymbol{\upsilon}_K^\top)}{\sigma_K^2} \right).$$

Results from Appendices A.6.1 and A.6.2 imply the desired update equations of Equation (5) presented in Algorithm 2.

#### A.6.3 Parameters update

**Lemma 4.** *Consider the series $u_n$ defined as follows:*

$$u_0 = \boldsymbol{A} \in \mathbb{R}^{n \times n}, \ u_{n+1} = u_n + \Lambda(\boldsymbol{B} - u_n), \ \forall n \geq 1, \ \textit{here } \boldsymbol{B} \in \mathbb{R}^{n \times n}.$$

*Then, $u_n \succ 0$ for all $n$ if the following conditions are satisfied: $u_0 \succ 0$, $\boldsymbol{B} \succeq 0$ and $1 > \Lambda > 0$.*

The proofs for Lemma 4 are shown in Appendix B.5.

**Gating Update:**

At the $t+1$ iteration, our gating parameters are calculated as followed:

$$\boldsymbol{\omega}^{(t+1)} = \arg\min_{\boldsymbol{\omega}} \left[ \boldsymbol{\omega}^\top s_{1,(t+1)} + \text{vec}(\boldsymbol{\omega}\boldsymbol{\omega}^\top)^\top s_{2,n+1} \right].$$

The first and second derivatives of $r_1(\boldsymbol{\omega}) = \boldsymbol{\omega}^\top s_{1,(t+1)} + \text{vec}(\boldsymbol{\omega}\boldsymbol{\omega}^\top)^\top s_{2,n+1}$ are:

$$\nabla r_1(\boldsymbol{\omega}) = s_{1,t+1} + (\text{mat}(s_{2,n+1}) + \text{mat}(s_{2,n+1})^\top)\boldsymbol{\omega}, \tag{32}$$

$$\nabla^2 r_1(\boldsymbol{\omega}) = \text{mat}(s_{2,n+1}) + \text{mat}(s_{2,n+1})^\top. \tag{33}$$

Since

$$s_{2,n+1} = s_{2,n} + \gamma_{n+1}\left[\frac{1}{2}\,\text{vec}\left(\left(1.5\boldsymbol{I}_{K-1} - \frac{\mathbf{1}\mathbf{1}^\top}{K-1}\right)/2 \otimes \hat{\boldsymbol{x}}\hat{\boldsymbol{x}}^\top\right) - s_{2,n}\right],$$

hence:

$$\text{mat}(s_{2,n+1}) = \text{mat}(s_{2,n}) + \gamma_{n+1}\left[\frac{1}{2}\left(1.5\boldsymbol{I}_{K-1} - \frac{\mathbf{1}\mathbf{1}^\top}{K-1}\right)/2 \otimes \hat{\boldsymbol{x}}\hat{\boldsymbol{x}}^\top - \text{mat}(s_{2,n})\right].$$

Now by applying Lemma 4 (knowing $\text{mat}(s_{2,0}) \succ 0$), we have $\text{mat}(s_{2,n})$ is positive definite for all $n$. Combine this with Equation (33), we have $\boldsymbol{\omega}^{(t+1)}$ is the solution of Equation (32).

Hence,

$$\boldsymbol{\omega}^{(t+1)} = -\left(\text{mat}(s_{2,n+1}) + \text{mat}(s_{2,n+1})^\top\right)^{-1} s_{1,t+1},$$

where $\text{mat}(s_{2,n+1})$ represents the reshaping of $s_{2,n+1}$ into a $K(D_W+1) \times K(D_W+1)$ square matrix.

**Upsilon Update:**
At the $t+1$ iteration, our gating parameters are calculated as followed:

$$\boldsymbol{\upsilon}_k^{(t+1)} = \arg\min_{\boldsymbol{\upsilon}_k}\left[s_{4,t+1,k}^\top \frac{\boldsymbol{\upsilon}_k}{\sigma_k^2} + s_{5,n+1,k}^\top \frac{\text{vec}(\boldsymbol{\upsilon}_k\boldsymbol{\upsilon}_k^\top)}{\sigma_k^2}\right].$$

The first and second derivatives of

$$r_2(\boldsymbol{\upsilon}_k) = s_{4,t+1,k}^\top \frac{\boldsymbol{\upsilon}_k}{\sigma_k^2} + s_{5,n+1,k}^\top \frac{\text{vec}(\boldsymbol{\upsilon}_k\boldsymbol{\upsilon}_k^\top)}{\sigma_k^2}$$

are:

$$\nabla r_2(\boldsymbol{\upsilon}_k) = \frac{s_{4,t+1,k}}{\sigma_k^2} + \frac{\left(\text{mat}\left(s_{5,n+1,k}\right) + \text{mat}\left(s_{5,n+1,k}\right)^\top\right)}{\sigma_k^2}\boldsymbol{\upsilon}_k, \tag{34}$$

$$\nabla^2 r_2(\boldsymbol{\upsilon}_k) = \frac{\left(\text{mat}\left(s_{5,n+1,k}\right) + \text{mat}\left(s_{5,n+1,k}\right)^\top\right)}{\sigma_k^2}. \tag{35}$$

Since

$$s_{5,n+1,[k-1,k]} = s_{5,n,[k-1,k]} + \gamma_{n+1}\left(\text{vec}\left(\boldsymbol{r}\boldsymbol{r}^\top\right) - s_{5,n,[k-1,k]}\right),$$

hence:

$$\text{mat}(s_{5,n+1,[k-1,k]}) = \text{mat}(s_{5,n,[k-1,k]}) + \gamma_{n+1}\left(\boldsymbol{r}\boldsymbol{r}^\top - \text{mat}(s_{5,n,[k-1,k]})\right).$$

Now by applying Lemma 4, knowing that:

$$\text{mat}(s_{5,0,k}) = \boldsymbol{z}\boldsymbol{z}^\top + \boldsymbol{I}_{D_V+1} \succ 0,$$

we have $\text{mat}(s_{5,n,[k-1,k]})$ is positive definite for all $n$ and $k \in \{1,\ldots,K\}$. Combine this with Equation (35), we have the solution of Equation (34) is the global minimum of $r_2(\boldsymbol{\upsilon}_k)$.

Hence:

$$\boldsymbol{v}_k^{t+1} = -\left(\text{mat}\left(s_{5,n+1,k}\right) + \text{mat}\left(s_{5,n+1,k}\right)^\top\right)^{-1} s_{4,t+1,k},$$

where $\text{mat}(s_{5,n+1,k})$ is $s_{5,n+1,k}$ reshape into $(D_V+1)$ x $(D_V+1)$ square matrix.

**Sigma Update:**
At the $t+1$ iteration, our sigma parameters are calculated as followed:

$$\sigma_k^{(t+1)} = \arg\min_{\sigma}\left[s_{3,t+1,[k]}\frac{1}{\sigma_k^2} + s_{4,t+1,k}^\top\frac{\boldsymbol{v}_k}{\sigma_k^2} + s_{5,n+1,k}^\top\frac{\text{vec}(\boldsymbol{v}_k\boldsymbol{v}_k^\top)}{\sigma_k^2} + s_{6,t+1,[k]}\log(\sigma_k^2)\right].$$

The first and second derivatives of:

$$r_3(\sigma_k^2) = s_{3,t+1,[k]}\frac{1}{\sigma_k^2} + s_{4,t+1,k}^\top\frac{\boldsymbol{v}_k}{\sigma_k^2} + s_{5,n+1,k}^\top\frac{\text{vec}(\boldsymbol{v}_k\boldsymbol{v}_k^\top)}{\sigma_k^2} + s_{6,t+1,[k]}\log(\sigma_k^2)$$

are:

$$\nabla r_3(\sigma_k^2) = -\frac{s_{3,t+1,[k]} + s_{4,t+1,k}^\top\boldsymbol{v}_k + s_{5,n+1,k}^\top\text{vec}(\boldsymbol{v}_k\boldsymbol{v}_k^\top)}{\sigma_k^4} + \frac{s_{6,t+1,[k]}}{\sigma_k^2}, \tag{36}$$

$$\nabla^2 r_3(\sigma_k^2) = 2\frac{s_{3,t+1,[k]} + s_{4,t+1,k}^\top\boldsymbol{v}_k + s_{5,n+1,k}^\top\text{vec}(\boldsymbol{v}_k\boldsymbol{v}_k^\top)}{\sigma_k^6} - \frac{s_{6,t+1,[k]}}{\sigma_k^4}.$$

We shall prove that the solution to Equation (36) is a global minimum of:

$$s_{3,t+1,[k]}\frac{1}{\sigma_k^2} + s_{4,t+1,k}^\top\frac{\boldsymbol{v}_k}{\sigma_k^2} + s_{5,n+1,k}^\top\frac{\text{vec}(\boldsymbol{v}_k\boldsymbol{v}_k^\top)}{\sigma_k^2} + s_{6,t+1,[k]}\log(\sigma_k^2).$$

Denote $\bar{\sigma}_{t,k}^2 = \dfrac{s_{3,t+1,[k]} + s_{4,t+1,k}^\top\boldsymbol{v}_k + s_{5,n+1,k}^\top\text{vec}(\boldsymbol{v}_k\boldsymbol{v}_k^\top)}{s_{6,t+1,[k]}}$ be such solution, we have:

$$\begin{aligned}
\bar{\sigma}_{t,k}^6\nabla^2 r_3(\bar{\sigma}_{t,k}^2) &= s_{3,t+1,[k]} + s_{4,t+1,k}^\top\boldsymbol{v}_k + s_{5,n+1,k}^\top\text{vec}(\boldsymbol{v}_k\boldsymbol{v}_k^\top) \\
&= s_{3,t,[k]} + \gamma_{t+1}\left(\boldsymbol{y}^2\boldsymbol{\tau}_k^{(t)} - s_{3,t,[k]}\right) + \left[s_{4,t,k} + \gamma_{t+1}\left((-2\boldsymbol{y}\tau_k^{(t)}\boldsymbol{r}) - s_{4,t,k}\right)\right]^\top\boldsymbol{v}_k \\
&\quad + \left[s_{5,t,k} + \gamma_{t+1}\left(\tau_k^{(t)}\left(\text{vec}\left(\boldsymbol{r}\boldsymbol{r}^\top\right)\right) - s_{5,t,k}\right)\right]^\top\text{vec}(\boldsymbol{v}_k\boldsymbol{v}_k^\top) \\
&= (1-\gamma_{t+1})\left[s_{3,t,[k]} + s_{4,t,k}^\top\boldsymbol{v}_k + s_{5,t,k}^\top\text{vec}(\boldsymbol{v}_k\boldsymbol{v}_k^\top)\right] + \gamma_{t+1}\tau_k^{(t)}\left(\boldsymbol{y} - \boldsymbol{r}^\top\boldsymbol{v}_k\right)^2.
\end{aligned}$$

Now notice that for all value of $\boldsymbol{v}_k$ we have:

$$s_{3,0,[k]} + s_{4,0,k}^\top\boldsymbol{v}_k + s_{5,0,k}^\top\text{vec}(\boldsymbol{v}_k\boldsymbol{v}_k^\top) = \left(\frac{1}{2} + \boldsymbol{z}^\top\boldsymbol{v}_k\right)^2 + \boldsymbol{v}_k^\top\boldsymbol{v}_k + \frac{3}{4} > 0.$$

Hence using induction we can prove that $\bar{\sigma}_{t,k}^6\nabla^2 r_3(\bar{\sigma}_{t,k}^2) > 0$ for all $t, k, \boldsymbol{v}_k$. Which directly leading to $\bar{\sigma}_{t,k}$ being a local minimum. Now since $\bar{\sigma}_{t,k}^2$ is the only solution of $\nabla r_3(\sigma_k^2)$, to prove that it is the global minimum, we only need to show that $\lim_{\sigma_k^2\to 0^+} r_3(\sigma_k^2) \geq r_3(\bar{\sigma}_{t,k}^2)$ and $\lim_{\sigma_k^2\to+\infty} r_3(\sigma_k^2) \geq r_3(\bar{\sigma}_{t,k}^2)$.
Denote $U = s_{3,t+1,[k]} + s_{4,t+1,k}^\top\boldsymbol{v}_k + s_{5,n+1,k}^\top\text{vec}(\boldsymbol{v}_k\boldsymbol{v}_k^\top) > 0$, and $V = s_{6,t+1,[k]} > 0$, we have:

$$\lim_{\sigma_k^2\to+\infty} r_3(\sigma_k^2) = \lim_{\sigma_k^2\to+\infty}\frac{U}{\sigma_k^2} + \lim_{\sigma_k^2\to+\infty} V\log(\sigma_k^2) = 0 + \infty = +\infty \geq r_3(\bar{\sigma}_{t,k}^2),$$

$$\lim_{\sigma_k^2\to 0^+} r_3(\sigma_k^2) = \lim_{\sigma_k^2\to+\infty} r_3(\frac{1}{\sigma_k^2}) = \lim_{\sigma_k^2\to+\infty}(U\sigma_k^2 - V\log(\sigma_k^2)) = +\infty \geq r_3(\bar{\sigma}_{t,k}^2).$$

Hence this has proven that $\bar{\sigma}_{t,k}^2$ is the global minimum, thus:

$$\sigma_k^{2^{(t+1)}} = \bar{\sigma}_{t,k}^2 = \frac{s_{3,t+1,[k]} + s_{4,t+1,k}^\top\boldsymbol{v}_k + s_{5,n+1,k}^\top\text{vec}(\boldsymbol{v}_k\boldsymbol{v}_k^\top)}{s_{6,t+1,[k]}}.$$

**Notes on the updates of $\upsilon$ and $\sigma$**

**Lemma 5.** *Consider the function*

$$f(x, \boldsymbol{y}) = \frac{g(\boldsymbol{y})}{x} + c\log(x).$$

*Where $\boldsymbol{y}$ is a vector, $x$ is a postivie real number, $c$ is a positive constant and $g(\boldsymbol{y}) : \mathbb{R}^n \to \mathbb{R}^+$ be a function that has exactly 1 global minimum.*

*Denote $\boldsymbol{y}_0 = \arg\min_{\boldsymbol{y}} g(\boldsymbol{y})$ be the global minimum of $g(\boldsymbol{y})$ and $x_0 = \arg\min_x \left(\frac{g(\boldsymbol{y}_0)}{x} + \log(x)\right)$.*
*We have $f(x_0, \boldsymbol{y}_0)$ is the global minimum of $f$ over all $x,\ \boldsymbol{y}$.*

*Proof of Lemma 5.* For all $\boldsymbol{y}$, we have:

$$\nabla_x f(x, \boldsymbol{y}) = \frac{1}{x^2}(-g(\boldsymbol{y}) + cx),\ \nabla_x^2 f(x, \boldsymbol{y}) = \frac{1}{x^3}(2g(\boldsymbol{y}) - cx).$$

Hence $x = \frac{g(\boldsymbol{y})}{c}$ is the local minimum of $f(x, \boldsymbol{y})$ for all $\boldsymbol{y}$. But since $\nabla_x f(x, \boldsymbol{y})$ has exactly 1 solution for x and both $\lim_{x\to+\infty} f(x, \boldsymbol{y})$, $\lim_{x\to+0} f(x, \boldsymbol{y})$ is $+\infty$. We can conclude that $x = \frac{g(\boldsymbol{y})}{c}$ is the global minimum of $f(x, \boldsymbol{y})$ for all given $\boldsymbol{y}$ .
Assume that $(x_1, \boldsymbol{y}_1)$ is the minimum of $f(x, \boldsymbol{y})$ ,we have

$$f(x_1, \boldsymbol{y}_1) \geq f\left(\frac{g(\boldsymbol{y}_1)}{c}, \boldsymbol{y}_1\right) = c\log\left(\frac{g(\boldsymbol{y}_1)}{c}\right) \geq c\log\left(\frac{g(\boldsymbol{y}_0)}{c}\right) = f\left(\frac{g(\boldsymbol{y}_0)}{c}, \boldsymbol{y}_0\right).$$

Which lead to $g(\boldsymbol{y}_1) = g(\boldsymbol{y}_0)$. But since $\boldsymbol{y}_0$ is the only global minimum of $g(\boldsymbol{y})$, we must have $\boldsymbol{y}_1 = \boldsymbol{y}_0$. Thus, leading to $x_1$ also is equal to $x_0$. □

**Remark 4.** *In our problem, $\boldsymbol{y} = \boldsymbol{\upsilon}_k, x = \sigma_k^2$ and $g(\boldsymbol{y}) = \sigma_k^2 r_2(\boldsymbol{\upsilon}_k)$. It has been proven in the Upsilon update section that $\sigma_k^2 r_2(\boldsymbol{\upsilon}_k)$ has exactly 1 global minimum (since it is strictly convex). Hence, Lemma 5 shows that our way of updating Upsilon and then Sigma is the same as updating the expert function all at once.*

# B Technical proofs

## B.1 Proof of Lemma 1

Let us assume that there exist $\varepsilon > 0$ such that:

$$\sum_{i=n}^{\infty} a_i > \varepsilon \quad \forall n.$$

For all positive integer $s$, denote:

$$c_{u_i} = a_{u_{i-1}+1} + a_{u_{i-1}+2} + \cdots + a_{u_i} \quad \forall i \in [s] \text{ where } a_{u_0} = a_0. \tag{37}$$

Now we have:

$$a_0 + \sum_{i=1}^{s} c_{u_i} < c \quad \forall u_1, u_2, \ldots, u_s. \tag{38}$$

However, due to Equation (37) there exist $\bar{u}_i$ such that $c_{\bar{u}_i} > \varepsilon$ for all $i \in [s]$. Thus combining this with Equation (38), we have

$$a_0 < c - s\varepsilon \quad \forall s \in \mathbb{Z}_+.$$

This contradicts the positivity of $a_0$, hence completes the proof.

### B.2 Proof of Lemma 2

Let $\mathcal{K}$ be a compact subset of $\mathbb{S}$ that contain the limiting point of $s_n$. We may decompose $\mathbb{P}(|\bar{S}\left(\bar{\boldsymbol{\theta}}\left(s_n\right); \boldsymbol{x}_{n+1}\right)| \geq M)$ as follows:

$$\begin{aligned}\mathbb{P}\big(|\bar{S}\left(\bar{\boldsymbol{\theta}}\left(s_n\right); \boldsymbol{x}_{n+1}\right)| \geq M\big) &\leq \mathbb{P}(s_n \notin \mathcal{K}) + \mathbb{P}\big(|\bar{S}\left(\bar{\boldsymbol{\theta}}\left(s_n\right); \boldsymbol{x}_{n+1}\right)| \geq M, s_n \in \mathcal{K}\big) \\ &\leq \mathbb{P}(s_n \notin \mathcal{K}) + M^{-p} \sup_{s \in \mathcal{K}} \big(\mathbb{E}_\pi[|\bar{S}\left(\bar{\boldsymbol{\theta}}\left(s_n\right); \boldsymbol{x}_{n+1}\right)|^p]\big)\mathbb{P}(s_n \in \mathcal{K}),\end{aligned}$$

which implies that

$$\limsup_{n\to\infty} \mathbb{P}\big(|\bar{S}\left(\bar{\boldsymbol{\theta}}\left(s_n\right); \boldsymbol{x}_{n+1}\right)| \geq M\big) \leq \mathbb{P}\big(\lim_{n\to\infty} s_n \notin \mathcal{K}\big) + M^{-p} \sup_{s \in \mathcal{K}} \big(\mathbb{E}_\pi[|\bar{S}\left(\bar{\boldsymbol{\theta}}\left(s_n\right); \boldsymbol{x}_{n+1}\right)|^p]\big).$$

Since $\mathbb{P}(\lim_{n\to\infty} s_n \notin \mathcal{K}) = 0$ as $s_n$ converges to a value in $\mathcal{K}$, we have

$$\limsup_{n\to\infty} \mathbb{P}\big(|\bar{S}\left(\bar{\boldsymbol{\theta}}\left(s_n\right); \boldsymbol{x}_{n+1}\right)| \geq M\big) \leq M^{-p} \sup_{s \in \mathcal{K}} \big(\mathbb{E}_\pi[|\bar{S}\left(\bar{\boldsymbol{\theta}}\left(s_n\right); \boldsymbol{x}_{n+1}\right)|^p]\big).$$

Taking the limit as $M \to \infty$, we obtain

$$\lim_{M\to\infty} \limsup_{n\to\infty} \mathbb{P}\big(|\bar{S}\left(\bar{\boldsymbol{\theta}}\left(s_n\right); \boldsymbol{x}_{n+1}\right)| \geq M\big) = 0,$$

which completes the proof.

### B.3 Proof of Lemma 3

Using the spectral decomposition we can write:

$$\boldsymbol{C} = \boldsymbol{M}_1^\top \boldsymbol{D}_1 \boldsymbol{M}_1, \ \boldsymbol{B} - \boldsymbol{A} = \boldsymbol{M}_2^\top \boldsymbol{D}_2 \boldsymbol{M}_2.$$

Hence:

$$\boldsymbol{C} \otimes (\boldsymbol{B} - \boldsymbol{A}) = (\boldsymbol{M}_1^\top \boldsymbol{D}_1 \boldsymbol{M}_1) \otimes (\boldsymbol{M}_2^\top \boldsymbol{D}_2 \boldsymbol{M}_2) = (\boldsymbol{M}_1 \otimes \boldsymbol{M}_2)^\top (\boldsymbol{D}_1 \otimes \boldsymbol{D}_2)(\boldsymbol{M}_1 \otimes \boldsymbol{M}_2),$$

which lead to $\boldsymbol{C} \otimes (\boldsymbol{B} - \boldsymbol{A})$ are nonnegative definite, thus proven the lemma.

### B.4 Proof of Proposition 5

Denote $\varepsilon = \dfrac{1}{p_1 + p_2 + \cdots + p_N}$. Hence $\varepsilon \hat{p}$ is a stochastic vector which according to Lemma 6:

$$\varepsilon \boldsymbol{\Lambda}_{\hat{\boldsymbol{p}}} - \varepsilon^2 \hat{\boldsymbol{p}} \hat{\boldsymbol{p}}^\top \leq (\boldsymbol{I}_N - \mathbf{1}\mathbf{1}^\top/N)/2.$$

Hence:

$$\begin{aligned}\boldsymbol{\Lambda}_{\hat{\boldsymbol{p}}} - \hat{\boldsymbol{p}} \hat{\boldsymbol{p}}^\top &\leq (\boldsymbol{I}_N - \mathbf{1}\mathbf{1}^\top/N)/2\varepsilon^2 + \left(1 - \frac{1}{\varepsilon}\right) \boldsymbol{\Lambda}_{\hat{\boldsymbol{p}}} \leq (\boldsymbol{I}_N - \mathbf{1}\mathbf{1}^\top/N)/2 + p_{N+1} \max(p_i) \boldsymbol{I}_N \\ &\leq (\boldsymbol{I}_N - \mathbf{1}\mathbf{1}^\top/N)/2 + \frac{1}{4} \boldsymbol{I}_N \leq (1.5 \boldsymbol{I}_N - \mathbf{1}\mathbf{1}^\top/N)/2.\end{aligned}$$

### B.5 Proof of Lemma 4

Let say that all the conditions are met, we shall prove that $u_n \succ 0$ for all $n$ using induction.

Since $u_1 \succ 0$, we assume that $u_k \succ 0$ for all $k \leq n$. Our target is to prove that $u_{n+1} \succ 0$.

This is true since the term $u_n(1 - \Lambda) \succ 0$ and the term $\Lambda \boldsymbol{B} \succeq 0$ , hence their sum which is $u_{n+1}$ is also positive definite.

## C Technical results

Surrogate function construction plays a pivotal role in optimization techniques such as MM algorithms. By leveraging mathematical inequalities, it is possible to approximate or bound complex objective functions with simpler, more tractable alternatives. Below are several fundamental inequalities commonly employed in the development of surrogate functions:

**Lemma 6** (Theorem 5.3 in [5]). *Let $\boldsymbol{p}_N$ be a stochastic vector of dimension $N$. We have:*

$$\boldsymbol{\Lambda}_{\boldsymbol{p}} - \boldsymbol{p}_N \boldsymbol{p}_N^\top \leq (\boldsymbol{I}_N - \mathbf{1}_N \mathbf{1}_N^\top / N)/2, \text{ where } \mathbf{1}_N \text{ is the } N \text{ dimensional vector } 1.$$

**Lemma 7** (See, *e.g.,* page 191 in [52]). *Given three vectors $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{c} \in \mathbb{R}^D$ such that all components of $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{c}$ are positive and a convex function $f$. The inequality below holds:*

$$f(\boldsymbol{c}^\top \boldsymbol{x}) \leq \sum_{i=1}^{D} \frac{c_i y_i}{\boldsymbol{c}^\top \boldsymbol{y}} f\left(\frac{\boldsymbol{c}^\top \boldsymbol{y}}{y_i} x_i\right), \text{ where the equality holds when } x = y.$$

**Lemma 8** (See, *e.g.,* page 191 in [52]). *Given three vectors $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{c} \in \mathbb{R}^D$ and a convex function $f$. The inequality below holds:*

$$f(\boldsymbol{c}^\top \boldsymbol{x}) \leq \sum_{i=1}^{D} \alpha_i f\left(\frac{c_i}{\alpha_i}(x_i - y_i)\right) + \boldsymbol{c}^\top \boldsymbol{y}), \text{ where the equality holds when } x = y.$$

**Lemma 9** (See, *e.g.,* page 206 in [52]). *Given 2 positive reals $x, x_n$. The inequality below holds:*

$$\log(x) \leq \frac{x}{x_n} + \log(x_n) - 1, \text{ where the equality holds when } x = x_n.$$

## References


[1] C. Andrieu, E. Moulines, and P. Priouret. Stability of Stochastic Approximation under Verifiable Conditions. *SIAM Journal on Control and Optimization*, 44(1):283–312, 2005. 7, 18

[2] Y. Bengio. Deep Learning of Representations: Looking Forward. In *Statistical Language and Speech Processing*, pages 1–37, Berlin, Heidelberg, 2013. 3

[3] M. Blein-Nicolas, E. Devijver, M. Gallopin, and E. Perthame. Nonlinear network-based quantitative trait prediction from biological data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, page qlae012, Mar. 2024. 4

[4] D. Böhning. Multinomial logistic regression algorithm. *Annals of the institute of Statistical Mathematics*, 44(1):197–200, 1992. 22

[5] D. Böhning and B. G. Lindsay. Monotonicity of quadratic-approximation algorithms. *Annals of the Institute of Statistical Mathematics*, 40(4):641–663, 1988. 28

[6] V. S. Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 9. Springer, 2008. 3

[7] F. Boux, F. Forbes, J. Arbel, B. Lemasson, and E. L. Barbier. Bayesian inverse regression for vascular magnetic resonance fingerprinting. *IEEE Transactions on Medical Imaging*, 40(7):1827–1837, 2021. 4

[8] O. Cappé. Online EM Algorithm for Hidden Markov Models. *Journal of Computational and Graphical Statistics*, 20(3):728–749, Jan. 2011. 3, 4

[9] O. Cappé and E. Moulines. On-Line Expectation–Maximization Algorithm for latent Data Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(3):593–613, Feb. 2009. 3, 4, 6

[10] F. Chamroukhi. Robust mixture of experts modeling using the t distribution. *Neural Networks*, 79:20–36, 2016. 12, 14

[11] F. Chamroukhi. Skew-normal Mixture of Experts. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 3000–3007, 2016. 14

[12] F. Chamroukhi. Skew t mixture of experts. *Neurocomputing*, 266:390–408, 2017. 12, 14

[13] F. Chamroukhi and B. T. Huynh. Regularized maximum-likelihood estimation of mixture-of-experts for regression and clustering. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2018. 9

[14] F. Chamroukhi and B. T. Huynh. Regularized maximum likelihood estimation and feature selection in mixtures-of-experts models. *Journal de la Société Française de Statistique*, 160(1):57–85, 2019. 9

[15] Z. Chen, Y. Deng, Y. Wu, Q. Gu, and Y. Li. Towards Understanding the Mixture-of-Experts Layer in Deep Learning. In *Advances in Neural Information Processing Systems*, 2022. 3, 9

[16] M. C. Chong, H. D. Nguyen, and T. Nguyen. Risk Bounds for Mixture Density Estimation on Compact Domains via the h-Lifted Kullback–Leibler Divergence. *Transactions on Machine Learning Research*, 2024. 3

[17] E. Chouzenoux and J.-B. Fest. Sabrina: A stochastic subspace majorization-minimization algorithm. *Journal of Optimization Theory and Applications*, 195:919–952, 2022. 3

[18] E. Chouzenoux and J.-C. Pesquet. A Stochastic Majorize-Minimize Subspace Algorithm for Online Penalized Least Squares Estimation. *IEEE Transactions on Signal Processing*, 65(18):4770–4783, 2017. 3

[19] S. L. Corff and G. Fort. Online Expectation Maximization based algorithms for inference in Hidden Markov Models. *Electronic Journal of Statistics*, 7(none):763 – 792, 2013. 3

[20] J. De Leeuw. Application of convex analysis to multidimensional scaling. *Recent developments in statistics*, pages 133–145, 1977. 3

[21] J. De Leeuw and W. J. Heiser. Convergence of correction matrix algorithms for multidimensional scaling. *Geometric representations of relational data*, 36:735–752, 1977. 3

[22] A. Deleforge, F. Forbes, S. Ba, and R. Horaud. Hyper-Spectral Image Analysis With Partially Latent Regression and Spatial Markov Dependencies. *IEEE Journal of Selected Topics in Signal Processing*, 9(6):1037–1048, 2015. 4

[23] B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the EM algorithm. *Annals of statistics*, pages 94–128, 1999. 6, 7

[24] 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, mar 1977. 3

[25] D. Do, L. Do, and X. Nguyen. Strong identifiability and parameter learning in regression with heterogeneous response. *arXiv preprint arXiv:2212.04091*, 2022. 3

[26] T. Do, L. Khiem, Q. Pham, T. Nguyen, T.-N. Doan, B. Nguyen, C. Liu, S. Ramasamy, X. Li, and S. Hoi. HyperRouter: Towards Efficient Training and Inference of Sparse Mixture of Experts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5754–5765, Singapore, Dec. 2023. 4

[27] W. Fedus, B. Zoph, and N. Shazeer. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. 4

[28] F. Forbes, H. D. Nguyen, T. Nguyen, and J. Arbel. Summary statistics and discrepancy measures for approximate Bayesian computation via surrogate posteriors. *Statistics and Computing*, 32(5):85, Oct. 2022. 4

[29] G. Fort, F. Forbes, and H. D. Nguyen. Sequential Sample Average Majorization–Minimization. *hal-04607609*, June 2024. 3

[30] G. Fort, P. Gach, and E. Moulines. Fast incremental expectation maximization for finite-sum optimization: nonasymptotic convergence. *Statistics and Computing*, 31(4):48, June 2021. 3

[31] C. R. Genovese and L. Wasserman. Rates of convergence for the Gaussian mixture sieve. *The Annals of Statistics*, 28(4):1105–1127, aug 2000. 3

[32] J. Hansen, R. Ruedy, J. Glascoe, and M. Sato. GISS analysis of surface temperature change. *Journal of Geophysical Research: Atmospheres*, 104(D24):30997–31022, 1999. 14

[33] J. Hansen, R. Ruedy, M. Sato, and K. Lo. Global surface temperature change. *Reviews of Geophysics*, 48:RG4004, 2010. 13

[34] E. Hazan. *Introduction to online convex optimization*, volume 2. Now Publishers, Inc., 2016. 3

[35] C. Hennig. Identifiablity of Models for Clusterwise Linear Regression. *Journal of Classification*, 17(2):273–296, 2000. 9

[36] N. Ho and X. Nguyen. Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *The Annals of Statistics*, 44(6):2726 – 2755, 2016. 3

[37] N. Ho and X. Nguyen. On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electronic Journal of Statistics*, 10(1):271–307, 2016. 3

[38] N. Ho, C.-Y. Yang, and M. I. Jordan. Convergence Rates for Gaussian Mixtures of Experts. *Journal of Machine Learning Research*, 2022. 3

[39] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991. 3

[40] N. Jaquier, R. Haschke, and S. Calinon. Tensor-variate mixture of experts for proportional myographic control of a robotic hand. *Robotics and Autonomous Systems*, 142:103812, 2021. 4

[41] H. Jiang and J. Xu. The stochastic proximal distance algorithm. *Statistics and Computing*, 34:210, 2024. 3

[42] W. Jiang and M. A. Tanner. Hierarchical mixtures-of-experts for exponential family regression models: approximation and maximum likelihood estimation. *Annals of Statistics*, pages 987–1011, 1999. 3

[43] W. Jiang and M. A. Tanner. On the identifiability of mixtures-of-experts. *Neural Networks*, 12(9):1253–1258, 1999. 9

[44] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural computation*, 6(2):181–214, 1994. 3

[45] B. Karimi, B. Miasojedow, E. Moulines, and H.-T. Wai. Non-asymptotic Analysis of Biased Stochastic Approximation Scheme. In A. Beygelzimer and D. Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1944–1974. PMLR, June 2019. 3

[46] B. Karimi, H.-T. Wai, E. Moulines, and M. Lavielle. On the Global Convergence of (Fast) Incremental Expectation Maximization Methods. In *Advances in Neural Information Processing Systems*, volume 32, 2019. 3

[47] B. Kugler, F. Forbes, and S. Douté. Fast Bayesian inversion for high dimensional inverse problems. *Statistics and Computing*, 32(2):31, 2022. 4

[48] E. Kuhn, C. Matias, and T. Rebafka. Properties of the stochastic approximation EM algorithm with mini-batch sampling. *Statistics and Computing*, 30(6):1725–1739, Nov. 2020. 3

[49] H. Kushner and G. G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*, volume 35. Springer Science & Business Media, 2003. 3, 7

[50] M. Kuusela, T. Vatanen, E. Malmi, T. Raiko, T. Aaltonen, and Y. Nagai. Semi-supervised anomaly detection – towards model-independent searches of new physics. *Journal of Physics: Conference Series*, 368(1):012032, June 2012. 4

[51] G. Lan. *First-order and stochastic optimization methods for machine learning*, volume 1. Springer, 2020. 3

[52] K. Lange. *Numerical analysis for statisticians*, volume 1. Springer, 2010. 28

[53] K. Lange. *MM Optimization Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2016. 3, 5

[54] S. Lathuilière, R. Juge, P. Mesejo, R. Muñoz-Salinas, and R. Horaud. Deep mixture of linear inverse regressions applied to head-pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4817–4825, 2017. 4

[55] H. A. Le Thi and V. T. Ho. Online Learning Based on Online DCA and Application to Online Classification. *Neural Computation*, 32(4):759–793, Apr. 2020. 3

[56] H. A. Le Thi and V. T. Ho. DCA for online prediction with expert advice. *Neural Computing and Applications*, 33(15):9521–9544, Aug. 2021. 3

[57] H. A. Le Thi, H. P. H. Luu, and T. P. Dinh. Online Stochastic DCA With Applications to Principal Component Analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 35(5):7035–7047, 2024. 3

[58] Q. Li, R. Shi, and F. Liang. Drug sensitivity prediction with high-dimensional mixture regression. *PLOS ONE*, 14(2):1–18, Feb. 2019. 4

[59] D. Lupu and I. Necoara. Convergence analysis of stochastic higher-order majorization–minimization algorithms. *Optimization Methods and Software*, 39:384–413, 2024. 3

[60] H. Lyu. Stochastic regularized majorization-minimization with weakly convex and multi-convex surrogates. *Journal of Machine Learning Research*, 25:1–83, 2024. 3

[61] H. Lyu, C. Strohmeier, and D. Needell. Online Nonnegative CP-dictionary Learning for Markovian Data. *Journal of Machine Learning Research*, 23(148):1–50, 2022. 3

[62] J. Mairal. Incremental Majorization-Minimization Optimization with Application to Large-Scale Machine Learning. *SIAM Journal on Optimization*, 25(2):829–855, 2015. 3

[63] S. Masoudnia and R. Ebrahimpour. Mixture of experts: a literature survey. *Artificial Intelligence Review*, 42(2):275–293, 2014. 3

[64] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, 1997. 3

[65] I. Meilijson. A Fast Improvement to the Em Algorithm on its Own Terms. *Journal of the Royal Statistical Society: Series B (Methodological)*, 51(1):127–138, Sept. 1989. 3

[66] E. F. Mendes and W. Jiang. On convergence rates of mixtures of polynomial experts. *Neural computation*, 24(11):3025–3051, 2012. 9

[67] A. Mensch, J. Mairal, B. Thirion, and G. Varoquaux. Stochastic Subsampling for Factorizing Huge Matrices. *IEEE Transactions on Signal Processing*, 66(1):113–128, 2018. 3

[68] A. Mokkadem and M. Pelletier. Convergence rate and averaging of nonlinear two-time-scale stochastic approximation algorithms. *The Annals of Applied Probability*, 16(3):1671 – 1702, 2006. 8

[69] D. N. Nguyen and Z. Li. Joint learning of Gaussian graphical models in heterogeneous dependencies of high-dimensional transcriptomic data. In *The 16th Asian Conference on Machine Learning (Conference Track)*, 2024. 4

[70] H. Nguyen, P. Akbarian, T. Nguyen, and N. Ho. A General Theory for Softmax Gating Multinomial Logistic Mixture of Experts. In *Proceedings of The 41st International Conference on Machine Learning*, 2024. 3

[71] H. Nguyen, T. Nguyen, and N. Ho. Demystifying Softmax Gating in Gaussian Mixture of Experts. In *Advances in Neural Information Processing Systems*, Dec. 2023. 3

[72] H. Nguyen, T. Nguyen, K. Nguyen, and N. Ho. Towards Convergence Rates for Parameter Estimation in Gaussian-gated Mixture of Experts. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238, pages 2683–2691, May 2024. 3

[73] H. D. Nguyen. An introduction to Majorization-Minimization algorithms for machine learning and statistical estimation. *WIREs Data Mining and Knowledge Discovery*, 7(2):e1198, 2017. 3

[74] H. D. Nguyen and F. Chamroukhi. Practical and theoretical aspects of mixture-of-experts modeling: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1246, 2018. 3

[75] H. D. Nguyen, F. Chamroukhi, and F. Forbes. Approximation results regarding the multiple-output Gaussian gated mixture of linear experts model. *Neurocomputing*, 366:208–214, 2019. 3

[76] H. D. Nguyen, F. Forbes, G. Fort, and O. Cappé. An Online Minorization-Maximization Algorithm. In *Classification and Data Science in the Digital Age - 17th Conference of the International Federation of Classification Societies, IFCS 2022, Proceedings*, pages 263–271. Springer, 2023. 3, 4, 5, 6

[77] H. D. Nguyen, F. Forbes, and G. J. McLachlan. Mini-batch learning of exponential family finite mixture models. *Statistics and Computing*, 30(4):731–748, July 2020. 3, 4

[78] H. D. Nguyen, L. R. Lloyd-Jones, and G. J. McLachlan. A universal approximation theorem for mixture-of-experts models. *Neural computation*, 28(12):2585–2593, 2016. 3

[79] H. D. Nguyen and G. J. McLachlan. Laplace mixture of linear experts. *Computational Statistics & Data Analysis*, 93:177–191, 2016. 4, 14

[80] H. D. Nguyen, T. Nguyen, F. Chamroukhi, and G. J. McLachlan. Approximations of conditional probability density functions in Lebesgue spaces via mixture of experts models. *Journal of Statistical Distributions and Applications*, 8(1):13, Aug. 2021. 3

[81] H. D. Nguyen, T. Nguyen, and F. Forbes. Bayesian Likelihood Free Inference using Mixtures of Experts. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2024. 3

[82] T. Nguyen. *Model Selection and Approximation in High-dimensional Mixtures of Experts Models: from Theory to Practice*. PhD Thesis, Normandie Université, Dec. 2021. 3

[83] T. Nguyen, F. Chamroukhi, H. D. Nguyen, and G. J. McLachlan. Approximation of probability density functions via location-scale finite mixtures in Lebesgue spaces. *Communications in Statistics - Theory and Methods*, pages 1–12, May 2022. 3

[84] T. Nguyen, H. D. Nguyen, F. Chamroukhi, and G. J. McLachlan. Approximation by finite mixtures of continuous density functions that vanish at infinity. *Cogent Mathematics & Statistics*, 7(1):1750861, 2020. 3

[85] X. Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, 41(1):370–400, 2013. 3

[86] A. Norets. Approximation of conditional densities by smooth mixtures of regressions. *The Annals of Statistics*, 38(3):1733 – 1766, 2010. 3

[87] J. Ortega and W. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*, volume 30. SIAM, 1970. 3

[88] G. Oudoumanessah, T. Coudert, C. Lartizien, M. Dojat, T. Christen, and F. Forbes. Scalable magnetic resonance fingerprinting: Incremental inference of high dimensional elliptical mixtures from large data volumes, 2024. 3

[89] G. Oudoumanessah, T. Coudert, L. Meyer, A. Delphin, M. Dojat, C. Lartizien, and F. Forbes. Cluster globally, reduce locally: Scalable efficient dictionary compression for magnetic resonance fingerprinting. In *IEEE International Symposium on Biological Imaging (ISBI)*, 2025. 3

[90] M. Pelletier. Weak convergence rates for stochastic approximation with application to multiple targets and simulated annealing. *Annals of Applied Probability*, pages 10–44, 1998. 6, 20

[91] Q. Pham, G. Do, H. Nguyen, T. Nguyen, C. Liu, M. Sartipi, B. T. Nguyen, S. Ramasamy, X. Li, S. Hoi, and others. CompeteSMoE–Effective Training of Sparse Mixture of Experts via Competition. *arXiv preprint arXiv:2402.02526*, 2024. 4

[92] B. T. Polyak. A new method of stochastic approximation type. *Avtomatika i telemekhanika*, pages 98–107, 1990. 8

[93] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992. 8, 12

[94] J. Puigcerver, C. R. Ruiz, B. Mustafa, and N. Houlsby. From Sparse to Soft Mixtures of Experts. In *The Twelfth International Conference on Learning Representations*, 2024. 4

[95] A. Rakhlin, D. Panchenko, and S. Mukherjee. Risk bounds for mixture density estimation. *ESAIM: PS*, 9:220–229, 2005. 3

[96] M. Razaviyayn, M. Sanjabi, and Z.-Q. Luo. A Stochastic Successive Minimization Method for Nonsmooth Nonconvex Optimization with Applications to Transceiver Design in Wireless Communication Networks. *Mathematical Programming*, 157(2):515–545, June 2016. 3

[97] R. Ruedy, W. Lawrence, and T. Peterson. A closer look at United States and global surface temperature change. *Journal of Geophysical Research: Atmospheres*, 106(D20):23947–23963, 2001. 14

[98] D. Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988. 8

[99] S. Shalev-Shwartz. Online Learning and Online Convex Optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012. 3

[100] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *International Conference on Learning Representations*, 2017. 4

[101] Y. Sun, P. Babu, and D. P. Palomar. Majorization-Minimization Algorithms in Signal Processing, Communications, and Machine Learning. *IEEE Transactions on Signal Processing*, 65(3):794–816, 2017. 3

[102] Z. You, S. Feng, D. Su, and D. Yu. Speechmoe2: Mixture-of-Experts Model with Improved Routing. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 7217–7221. IEEE, 2022. 4

[103] S. E. Yuksel, J. N. Wilson, and P. D. Gader. Twenty Years of Mixture of Experts. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8):1177–1193, 2012. 3