Importance sampling-based gradient method for dimension reduction in Poisson log-normal model

Bastien Batardière, Julien Chiquet, Joon Kwon, Julien Stoehr

Importance sampling-based gradient method for dimension reduction in Poisson log-normal model

Bastien Batardière[1] **Julien Chiquet**[1], **Joon Kwon**[1], **and Julien Stoehr**[2,1]

[1] *Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, 91120, Palaiseau, France, e-mail:* bastien.batardiere@inrae.fr; julien.chiquet@inrae.fr; joon.kwon@inrae.fr

[2] *Université Paris-Dauphine, Université PSL, CNRS, CEREMADE, 75016 Paris, France, e-mail:* stoehr@ceremade.dauphine.fr

Abstract: High-dimensional count data poses significant challenges for statistical analysis, necessitating effective methods that also preserve explainability. We focus on a low rank constrained variant of the Poisson log-normal model, which relates the observed data to a latent low-dimensional multivariate Gaussian variable via a Poisson distribution. Variational inference methods have become a golden standard solution to infer such a model. While computationally efficient, they usually lack theoretical statistical properties with respect to the model. To address this issue we propose a projected stochastic gradient scheme that directly maximizes the log-likelihood. We prove the convergence of the proposed method when using importance sampling for estimating the gradient. Specifically, we obtain a rate of convergence of $O(T^{-1/2} + N^{-1})$ with T the number of iterations and N the number of Monte Carlo draws. The latter follows from a novel descent lemma for non convex L-smooth objective functions, and random biased gradient estimate. We also demonstrate numerically the efficiency of our solution compared to its variational competitor. Our method not only scales with respect to the number of observed samples but also provides access to the desirable properties of the maximum likelihood estimator.

MSC2020 subject classifications: Primary 00X00, 00X00; secondary 00X00.
Keywords and phrases: Dimension reduction, importance sampling, multivariate count data, Poisson log-normal model, projected stochastic gradient descent.

1. Introduction

Multivariate count data are prevalent in a widening range of applications such as ecology, genomics, microbiology, astronomy, and economy, just to name a few. This ubiquity has prompted the development of numerous statistical models, as unlike continuous multivariate distributions, a generic universal multivariate distribution for count data does not exist (Inouye et al., 2017). Most of the

arXiv: 2010.00000

successful proposals are latent variable models belonging to the family of generalized multivariate mixed models (GMMM). The latter offers the strength of model-based approaches, enabling the incorporation of external covariates and allowing the latent variables to be constrained in various ways to perform a specific task — regression, variable selection, and dimension reduction — while controlling the complexity of the model. These strengths contribute to the unwavering popularity of these models in the aforementioned fields of application, where both modeling and interpretability are essential prerequisites.

Significant milestones in the literature include a few generic frameworks initially developed through applications in ecology, where count and abundance tables have long been the norm. The generalized linear latent variable models (GLLVM) of Niku et al. (2019) are instances of generalized multivariate mixed models with low-dimensional latent variables, where the distribution of observed responses usually belongs to the exponential family. The Hierarchical Modeling of Species Communities (HMSC), presented in Ovaskainen et al. (2017), also falls within the class of generalized linear latent variable models with additional layers in the modeling of the latent variables. Multivariate Poisson Log-Normal models (PLN), as presented in Chiquet, Mariadassou and Robin (2021), are yet another instance of the expansive family of generalized linear latent variable models. The latter confers the advantage of developing a generic and versatile framework capable of addressing various tasks, including dimension reduction, regression, clustering, discriminant analysis.

These modeling frameworks encounter the usual inference issues inherent to latent variable models. Specifically, direct and exact likelihood maximization is difficult since it requires evaluating an integral over a space of the latent variable dimension. Hence, direct numerical integration approaches (Aitchison and Ho, 1989) are limited to small-scale problems involving solely a few variables. Approaches based on Markov chain Monte Carlo (MCMC) techniques can handle medium-size problems but are computationally expensive (Hui, 2016; Tikhonov et al., 2020). Alternatively, methods based on Laplace approximations exhibit greater computational efficiency but can potentially be inaccurate (Gómez-Rubio, 2020). An indirect approach relies on the Expectation-Maximization (EM) algorithm, a well-established method for inference in incomplete data models since Dempster, Laird and Rubin (1977). However, the M-step is practically intractable in GMMM, as it requires computing an expectation with respect to the distribution of the latent variable conditional on the observations. MCMC techniques can obviously be used in such a setting (*e.g.*, Karlis, 2005), though displaying the same shortcomings. More recently, the growing size of datasets and the porting of these methods to other fields of application where the number of variables expands drastically, such as genomics, have sparked interest in variational approaches (Hall, Ormerod and Wand, 2011; Blei, Kucukelbir and McAuliffe, 2017; Hui et al., 2017; Niku et al., 2019) as they provide a good compromise between accuracy and computational efficiency.

In the context of PLN models, which is the focus of this paper, Chiquet, Mariadassou and Robin (2021) have extensively used this variational approach in conjunction with the EM algorithm and adapted it to several contexts, in-

cluding dimension reduction, clustering, sparse covariance. The implementation provided in Chiquet et al. (2023) is efficient and can deal with problems with thousands of observations and hundreds — even a couple of thousands — of variables. The resulting estimator can be shown to converge to the maximum of the surrogate likelihood function and enjoy asymptotical normality (Westling and McCormick, 2019). However, these results pertain to the surrogate model and generally differ from the natural properties of an M-estimator associated with the likelihood (van der Vaart, 1998). In particular, while the maximum reached by the variational estimator seems at least empirically to coincide with the maximum likelihood estimator (MLE), there is no genuine estimator of the variance of the variational estimator that can be used to measure uncertainty properly: although the bootstrap method or the jackknife estimator could be used to build an estimator of the variance of the estimate, the variational solution is marred by the lack of relevant statistical guarantees. Consequently, the design of efficient algorithms that can directly maximize the likelihood and inherit the desirable properties of MLEs is still a key research issue for GMMM, particularly for PLN models. Such an algorithm allows a more direct assessment of estimator uncertainty by means of asymptotic variance estimates. It is in this spirit that Stoehr and Robin (2024) propose a variant of the Monte Carlo EM scheme that combines composite likelihood and importance sampling methods with a focus on applications in synecology. While the approach benefits from the properties of the maximum composite likelihood estimator, it necessitates splitting the data into overlapping blocks containing a small number of variables. The mimimum number of blocks required grows quadratically in the number of variables (or species). As the computational complexity of their Monte Carlo EM increases linearly with the number of blocks, the solution is primarily suited for problems involving a few dozen variables but does not scale up efficiently to larger problems from a computational perspective.

Contributions This paper introduces a projected stochastic gradient descent (SGD) scheme based on self-normalized importance sampling to obtain gradient estimates for optimizing the marginal likelihood of the observed data in the Poisson log-normal model, subject to a rank constraint on the latent space. This model, introduced by Chiquet, Mariadassou and Robin (2018a), can be seen as a probabilistic version of Principal Component Analysis (PCA) with Poisson emission law, and its standard inference solution is a variational Expectation-Maximization (VEM) algorithm. Estimating parameters according to the maximum likelihood principle with Monte Carlo simulations is a long-standing idea for an unnormalized statistical model — a class of challenging models due to their intractable partition function which is a highly multidimensional integral depending on the parameters. For instance, Monte Carlo maximum likelihood estimation (Geyer, 1994) uses importance sampling to estimate the partition function while contrastive divergence (Hinton, 2002) estimates the gradient of the log partition function via Monte Carlo methods. More recently, the noise-contrastive estimation (Gutmann and Hyvärinen, 2012) reformulates the initial problem to avoid estimating the partition function or its gradient. Here, we rely

on the fact that incomplete data models share similarities with the unnormalized models in that, under mild regularity conditions, the inference resumes to deal with an intractable integral, namely the score function for the observed likelihood (Louis, 1982). We show that the PLN-PCA model falls within the set of incomplete data models for which the score function is written as an expected value with respect to the conditional distribution of the latent given the observed and can thereby be estimated by simulation methods. The rank constraint ensures that the importance sampling estimator can handle problems with up to thousands of variables in the emission space, provided that the dimension of the latent space is controlled and limited to some tens. The projection step onto a convex compact set specifically guarantees that the objective function is L-smooth. We also show that it ensures a bounded mean squared error and bias for the gradient estimator. Such properties are common in the literature (Ghadimi and Lan, 2013; Mai and Johansson, 2021; Scaman, Malherbe and Santos, 2022).

Our major contribution is a novel convergence theorem for the gradient method presented. To establish the result within the context of a self-normalized importance sampling estimator, we first present a general descent lemma applicable under minimal assumptions — specifically, L-smoothness, and bounded bias and quadratic error for the gradient estimator. To the best of our knowledge, it is the first result on projected stochastic gradient schemes for potentially both non convex objective functions and random biased gradient estimators. Given T iterations of our method and N Monte Carlo draws, we obtain a theoretical rate of $O(T^{-1/2} + N^{-1})$ for the gradient mapping norm. This convergence rate is consistent with those demonstrated in the literature, albeit in different contexts. For a non-convex setting with an unbiased gradient estimator, Ghadimi and Lan (2013) derive an $O(T^{-1/2})$ convergence rate for the gradient's norm objective, while Mai and Johansson (2021) achieve the same rate for the gradient mapping norm but when a projection step is added. In situations where the gradient estimator is biased, with a bound b on the bias, Ajalloeian and Stich (2021) retrieve a rate of $O(T^{-1/2} + b)$, but only when no projection step is performed.

The paper also includes an efficient implementation of our algorithm using `JAX` library (Bradbury et al., 2018), and GPU computing. As a by-product, we provide a `PyTorch` (Paszke et al., 2019) version of the VEM solution, enabling the analysis of large-scale datasets with hundreds of thousands of observations and tens of thousands of variables.

Outline The paper begins with an introduction to the standard multivariate PLN model and its PCA version in Section 3. We then present our stochastic gradient scheme with convergence guarantees in Section 4. In Section 5, we propose a sequentially adapted Gaussian mixture distribution to serve as a valid importance sampling proposal distribution within our algorithm. Finally, Section 6 details a simulation study on synthetic data and an application to genomic data, where we deal with the problem of dimension reduction and visualization of a transcriptomic single-cell dataset. Technical details are postponed till Appendices A–D.

2. Notations and conventions

Let p and q be positive integers. The vector space of all $p \times q$-matrices over a ring $\mathbb{A}$ is denoted by $\mathcal{M}_{p\times q}(\mathbb{A})$. The subset of all symmetric, positive and definite $p \times p$-matrices over $\mathbb{R}$ is denoted by $\mathcal{S}^p_{++}$. We denote by $\langle \cdot, \cdot \rangle$ the scalar product on a real p-space $\mathbb{R}^p$, and $\|\cdot\|$ its associated norm. The matrix norm induced by $\|\cdot\|$ on $\mathcal{M}_{p\times p}(\mathbb{R})$ is also denoted by $\|\cdot\|$. $\mathrm{Diag}(x)$ is a diagonal matrix with diagonal equal to x for x a vector. When applied to matrices or vectors, simple functions like $\log, \exp$ or square apply element-wise.

We denote by $\mathbb{M}_1(\mathbb{R}^p)$ the set of probability measures on $\mathbb{R}^p$. Given a probability measure $\pi \in \mathbb{M}_1(\mathbb{R}^p)$, $\mathbb{M}_\pi$ is the set of probability measures that dominates π. The product measure $\prod_{i=1}^n \pi$ on $\mathbb{R}^{d\times n}$ is denoted by $\pi^{\otimes n}$. We use the same notation to refer to a measure and its associated density, meaning that if π is absolutely continuous with respect to the Lebesgue measure λ, $\pi(\mathrm{d}x) = \pi(x)\lambda(\mathrm{d}x)$. The expectation with respect to π is denoted by $\mathbb{E}_\pi$. When there is no ambiguity regarding the integration measure, we simply use the notation $\mathbb{E}$.

We denote by $\mathcal{N}(\boldsymbol{\mu}, \mathbf{S})$ a p-dimensional Gaussian variable with mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and variance $\mathbf{S} \in \mathcal{S}^p_{++}$ and $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \mathbf{S})$ its density evaluated at $\mathbf{x} \in \mathbb{R}^p$. We denote by $\mathcal{P}(\lambda)$ a Poisson variable with rate $\lambda > 0$.

The Kullback–Leibler divergence between $\pi \in \mathbb{M}_1(\mathbb{R}^p)$ and $\mu \in \mathbb{M}_\pi$ is defined by

$$\mathrm{KL}(\pi \parallel \mu) = \int \log \frac{\pi(x)}{\mu(x)} \pi(\mathrm{d}x).$$

The entropy of a random variable X distributed according to $\pi \in \mathbb{M}_1(\mathbb{R}^p)$ is defined by

$$\mathcal{H}_\pi(X) = -\int \log \pi(x)\pi(\mathrm{d}x).$$

Given $\mathcal{X} \subseteq \mathbb{R}^p$, a differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ is said to be $L-$smooth on $\mathcal{X}$ with $L \geq 0$ if its gradient is $L-$Lipschitz on $\mathcal{X}$, namely, for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathcal{X}$,

$$\|\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}')\| \leq L\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|.$$

3. Dimension reduction in multivariate Poisson log-normal models

Background: Multivariate Poisson log-normal model Consider a data matrix $\mathbf{Y} = (Y_{ij}) \in \mathcal{M}_{n\times p}(\mathbb{N})$ storing n i.i.d. observations of a p-dimensional random vector $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{ip}) \in \mathbb{N}^p$, $i = 1, \ldots, n$. The multivariate Poisson lognormal model (see Aitchison and Ho, 1989, for its original formulation) relates each of the observed vector $\mathbf{Y}_i$ to a latent (or unobserved) p-dimensional Gaussian vector $\mathbf{Z}_i$, whose covariance matrix $\boldsymbol{\Sigma}$ describes the underlying structure of dependence between the p variables. Following a formalism similar to that of GMMM, the model can also encompass a main effect due to a linear combination of d observed covariates $\mathbf{x}_i \in \mathbb{R}^d$ (including a vector of intercepts), and some possible user-specified offsets $\mathbf{o}_i = (o_{ij}) \in \mathbb{R}^p$ to take into account

the sampling efforts between the samples. The model assumes that the observations Y_{ij} are independent conditionnally on $\mathbf{Z}_i = (Z_{ij})$, and that the conditional distribution $p(Y_{ij} \mid Z_{ij})$ is a Poisson distribution, namely,

$$\begin{aligned} \{\mathbf{Z}_i\}_{1\leq i\leq n} \ i.i.d. : &\qquad \mathbf{Z}_i \sim \mathcal{N}(\mathbf{0}_q, \boldsymbol{\Sigma}); \\ \{Y_{ij}\}_{\substack{1\leq i\leq n \\ 1\leq j\leq p}} \text{ ind.} \mid \{\mathbf{Z}_i\}_{1\leq i\leq n} : &\quad Y_{ij} \mid Z_{ij} \sim \mathcal{P}\left(\exp(Z_{ij} + \mathbf{B}_j^\top \mathbf{x}_i + o_{ij})\right), \end{aligned} \tag{1}$$

where $\mathbf{B} = [\mathbf{B}_1, \ldots, \mathbf{B}_p] \in \mathcal{M}_{d\times p}(\mathbb{R})$ is a latent matrix of regression parameters. In this framework, the main goal is to estimate the vector of parameters $\boldsymbol{\theta} = (\mathbf{B}, \boldsymbol{\Sigma}) \in \mathbb{R}^D$, with $D = pm + p(p+1)/2$, from the data matrices $\mathbf{Y}$ and $\mathbf{X}$.

Poisson lognormal-model with low-rank constraint Throughout this paper, we focus on the PCA variant of model (1) as introduced in Chiquet, Mariadassou and Robin (2018a). The latter is derived by adding a rank constraint on the latent covariance matrix, such that $\text{rank}(\boldsymbol{\Sigma}) = q < p$. The constraint alleviates the number of parameters to estimate, which can become prohibitive when the number of variables p is large in (1). This key feature is particularly relevant in the perspective of importance sampling. Indeed, this allows us to deal with problems where the dimension of the observation space is potentially large, in contrast with the number of parameters in the model and the dimension of the latent space, where the particles are sampled. The PCA version can be written in a hierarchical framework by adding a layer with, for each individual, a q-dimensional standard Gaussian vector $\mathbf{W}_i$, and introducing an individual-specific linear function $f_i : \mathbb{R}^q \to \mathbb{R}^p$ defined for all $\mathbf{w} \in \mathbb{R}^q$ as

$$f_i(\mathbf{w}; \mathbf{B}, \mathbf{C}) = \mathbf{C}\mathbf{w} + \mathbf{B}^\top \mathbf{x}_i + \mathbf{o}_i,$$

where $\mathbf{C} \in \mathcal{M}_{p\times q}(\mathbb{R})$ encodes the embedding of the observations into a space of lower dimension. The model is then written as

$$\begin{aligned} \{\mathbf{W}_i\}_{1\leq i\leq n} \ i.i.d. : &\qquad \mathbf{W}_i \sim \mathcal{N}(\mathbf{0}_q, \mathbf{I}_q); \\ \{\mathbf{Z}_i\}_{1\leq i\leq n} \ i.i.d. : &\qquad \mathbf{Z}_i = \mathbf{C}\mathbf{W}_i + \mathbf{B}^\top \mathbf{x}_i + \mathbf{o}_i; \\ \{Y_{ij}\}_{\substack{1\leq i\leq n \\ 1\leq j\leq p}} \text{ ind.} \mid \{\mathbf{Z}_i\}_{1\leq i\leq n} : &\quad p\left(Y_{ij} \mid Z_{ij}\right) = \frac{\exp\left\{Y_{ij}Z_{ij} - \exp(Z_{ij})\right\}}{Y_{ij}!}. \end{aligned} \tag{2}$$

We refer to Model (2) as PLN-PCA. Adopting the PCA terminology, $\mathbf{C}$ is the $p \times q$ matrix of loadings, and $\mathbf{W}_i$ represents the vector of scores of the i-th observation in the low-dimensional latent subspace, whose dimension q corresponds to the number of components.

The vector of unknown parameters to be estimated is now $\boldsymbol{\theta} = (\mathbf{B}, \mathbf{C}) \in \mathbb{R}^d$, with $d = p(q+m)$ that is significantly smaller than D when $q \ll p$, the typical case at play in a context of dimension reduction. The complete log-likelihood of Model (2) can be written, up to additive constants with respect to model parameters, as

$$\sum_{i=1}^n \log p_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{W}_i) = \sum_{i=1}^n \left\{ \langle \mathbf{Y}_i, \mathbf{Z}_i \rangle - \sum_{j=1}^p \exp(Z_{ij}) - \frac{1}{2}\|\mathbf{W}_i\|^2 \right\}.$$

Without any further assumption on $\mathbf{C}$, remark that $\boldsymbol{\theta}$ is not identifiable since the distribution of $\mathbf{Z}_i$ is invariant when multiplying $\mathbf{C}$ by any orthogonal matrix. However, since $\mathbf{Y}_i$ depends on $\mathbf{C}$ solely through the covariance $\boldsymbol{\Sigma} = \mathbf{C}\mathbf{C}^\top$, it is enough to have the identifiability for the model parametrized by $(\mathbf{B}, \boldsymbol{\Sigma})$. We will discuss this point further when introducing the gradient ascent algorithm.

4. Biased stochastic gradient descent

4.1. Solution to maximum likelihood principle

Estimation of $\boldsymbol{\theta}$ is achieved by maximizing the likelihood of the observed data $p_{\boldsymbol{\theta}}(\mathbf{Y})$, or equivalently by solving the optimization problem

$$\boldsymbol{\theta}^\star = \arg\min_{\boldsymbol{\theta} \in \mathcal{X}} \ell(\boldsymbol{\theta}), \quad \ell(\boldsymbol{\theta}) = -\frac{1}{n} \sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(\mathbf{Y}_i), \tag{3}$$

where ℓ is referred to as the loss function. Optimizing such a function is not straightforward because the marginal $p_{\boldsymbol{\theta}}(\mathbf{Y}_i)$ requires integrating out the latent variable $\mathbf{W}_i$.

Reminder on the EM approach The Expectation-Maximization algorithm (Dempster, Laird and Rubin, 1977) circumvent this issue by using a decomposition of the log-likelihood of the observed data $\mathbf{Y}_i$ into

$$\begin{aligned}\log p_{\boldsymbol{\theta}}(\mathbf{Y}_i) &= \int_{\mathbb{R}^q} \log p_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{w}) p_{\boldsymbol{\theta}}(\mathrm{d}\mathbf{w} \mid \mathbf{Y}_i) - \int_{\mathbb{R}^q} \log p_{\boldsymbol{\theta}}(\mathbf{w} \mid \mathbf{Y}_i) p_{\boldsymbol{\theta}}(\mathrm{d}\mathbf{w} \mid \mathbf{Y}_i) \\ &= \mathbb{E}\left[\log p_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{W}_i) \mid \mathbf{Y}_i, \boldsymbol{\theta}\right] + \mathcal{H}_{p_{\boldsymbol{\theta}}(\cdot \mid \mathbf{Y}_i)}(\mathbf{W}_i),\end{aligned}$$

The algorithm proceeds by evaluating the conditional expectation of the complete log-likelihood using the current estimates $\boldsymbol{\theta}^{(t)}$ of the model parameters:

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{n} \mathbb{E}\left[\log p_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{W}_i) \;\middle|\; \mathbf{Y}_i, \boldsymbol{\theta}^{(t)}\right].$$

By iteratively maximizing this quantity, the algorithm generates a sequence that converges under suitable regularity conditions to the maximum likelihood estimator $\boldsymbol{\theta}^\star$ (Wu, 1983; Boyles, 1983). However, the conditional $p_{\boldsymbol{\theta}}(\mathbf{W}_i \mid \mathbf{Y}_i)$ is intractable for the PLN model and its PCA extension. To address this challenge, variational inference approximates $p_{\boldsymbol{\theta}}(\mathbf{W}_i \mid \mathbf{Y}_i)$ with a surrogate distribution $\phi_{\boldsymbol{\psi}}(\mathbf{V})$ from a parametric family $\mathcal{P}$. For instance, Chiquet, Mariadassou and Robin (2018b, 2019) resorted to multivariate Gaussian distributions $\mathcal{N}(\mathbf{m}_i^{\mathrm{VEM}}, \mathbf{S}_i^{\mathrm{VEM}})$ with diagonal covariance matrix. The Variational EM method alternates between updates of the variational parameter $\boldsymbol{\psi}$ and the model parameter $\boldsymbol{\theta}$, aiming to maximize a lower bound of the log-likelihood, defined as

$$\log p_{\boldsymbol{\theta}}(\mathbf{Y}_i) - \mathrm{KL}\left[\phi_{\boldsymbol{\psi}}(\mathbf{W}_i) \| p_{\boldsymbol{\theta}}(\mathbf{W}_i \mid \mathbf{Y}_i)\right] = \mathbb{E}_{\phi_{\boldsymbol{\psi}}}\left[\log p_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{V})\right] + \mathcal{H}_{\phi_{\boldsymbol{\psi}}}(\mathbf{V}).$$

At convergence, the variational solution is associated with

$$\phi_{\boldsymbol{\psi}} = \arg\min_{\phi \in \mathcal{P}} \mathrm{KL}\left[\phi \| p_{\boldsymbol{\theta}}(\cdot \mid \mathbf{Y}_i)\right]. \tag{4}$$

Stochastic gradient scheme In contrast to existing methods, we propose to address the optimization problem (3) directly with an SGD scheme. Our approach leverages that the Louis principle applies to Model (2), as outlined below (see Appendix A for a proof).

Proposition 4.1. *For all individual $i = 1, \dots, n$, the incomplete log-likelihood $\boldsymbol{\theta} \mapsto \log p_{\boldsymbol{\theta}}(\mathbf{Y}_i)$ of Model (2) is twice continuously differentiable on $\mathbb{R}^d$ and its score function can be written as*

$$s_i(\boldsymbol{\theta}) = \int_{\mathbb{R}^q} \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{w}) p_{\boldsymbol{\theta}}(\mathrm{d}\mathbf{w} \mid \mathbf{Y}_i) = \mathbb{E}\left[\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{W}_i) \mid \mathbf{Y}_i, \boldsymbol{\theta}\right]. \quad (5)$$

While the intractability of the conditional distribution renders exact calculation infeasible even for a small value of q, identity (5) is instrumental in estimating the score function with Monte Carlo methods and designing an SGD scheme. Given a learning rate $\gamma \in \mathbb{R}_+^*$ and an initial point $\boldsymbol{\theta}^{(1)} \in \mathbb{R}^d$, the SGD scheme recursively defines a sequence $\{\boldsymbol{\theta}^{(t)}\}_{t \in \mathbb{N}^*}$ through the equation

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \gamma \widehat{g}^{(t)}, \quad (6)$$

where $\widehat{g}^{(t)}$ is a possibly biased estimator of $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^{(t)})$. Here we explore the opportunity of importance sampling methods (Kahn, 1949; Kahn and Harris, 1951) to define $\widehat{g}^{(t)}$. Indeed, the lack of closed-form for $p_{\boldsymbol{\theta}}(\mathbf{W}_i \mid \mathbf{Y}_i)$ hinders Monte Carlo methods that rely on exact samples from it. However, importance sampling overcomes this difficulty by changing the integration measure. For any $\boldsymbol{\theta} \in \mathbb{R}^d$, it approximates $p_{\boldsymbol{\theta}}(\cdot \mid \mathbf{Y}_i)$ with a random probability measure based on weighted samples from a probability density function $\nu_i(\cdot\,; \boldsymbol{\theta})$, possibly depending on $\boldsymbol{\theta}$ and referred to as proposal distribution, such that $p_{\boldsymbol{\theta}}(\cdot \mid \mathbf{Y}_i)$ is absolutely continuous with respect to $\nu_i(\cdot\,; \boldsymbol{\theta})$. The importance sampling method is then based on the following identity

$$s_i(\boldsymbol{\theta}) = \int_{\mathbb{R}^q} \frac{p_{\boldsymbol{\theta}}(\mathbf{v} \mid \mathbf{Y}_i)}{\nu_i(\mathbf{v}; \boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{v}) \nu_i(\mathrm{d}\mathbf{v}; \boldsymbol{\theta}).$$

To circumvent the issue of evaluating the intractable distribution in the above, the method leverages a tractable non-normalized version of the conditional distribution, namely the joint distribution. Let us introduce the Radon-Nikodym derivative of $p_{\boldsymbol{\theta}}(\mathbf{Y}_i, \cdot)$ with respect to $\nu_i(\cdot\,; \boldsymbol{\theta})$:

$$\rho_{\boldsymbol{\theta},i}(\mathbf{v}) = \frac{p_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{v})}{\nu_i(\mathbf{v}; \boldsymbol{\theta})}, \quad \mathbf{v} \in \mathbb{R}^q.$$

The score can be written as

$$s_i(\boldsymbol{\theta}) = \int_{\mathbb{R}^q} \rho_{\boldsymbol{\theta},i}(\mathbf{v}) \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{v}) \nu_i(\mathrm{d}\mathbf{v}; \boldsymbol{\theta}) \Big/ \int_{\mathbb{R}^q} \rho_{\boldsymbol{\theta},i}(\mathbf{v}) \nu_i(\mathrm{d}\mathbf{v}; \boldsymbol{\theta}) \,.$$

Thereby, given $N \in \mathbb{N}^*$ independent samples $\mathbf{v}_{i,1}, \dots, \mathbf{v}_{i,N}$ from $\nu_i(\cdot\,; \boldsymbol{\theta})$, the self-normalized importance sampling (SNIS) estimator of $s_i(\boldsymbol{\theta})$ is

$$\widehat{s}_i^N(\boldsymbol{\theta}) = \sum_{r=1}^N \rho_{\boldsymbol{\theta},i}(\mathbf{v}_{i,r}) \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}\left(\mathbf{Y}_i, \mathbf{v}_{i,r}\right) \Big/ \sum_{s=1}^N \rho_{\boldsymbol{\theta},i}(\mathbf{v}_{i,s}) \,. \quad (7)$$

Algorithm 1: Importance Sampling based Gradient Descent (ISGD)

Input: initial point $\theta^{(1)}$, learning rate γ, number of iterations T, number of Monte Carlo draws N.
Output: the sequence $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T+1)}$
for $t = 1$ **to** T **do**
  Sample i uniformly in $\{1, \ldots, n\}$;
  Sample $(\mathbf{v}_{i,1}, \ldots, \mathbf{v}_{i,N})$ from $\nu_i^{\otimes N}\big(\cdot\,; \boldsymbol{\theta}^{(t)}\big)$;
  Update $\boldsymbol{\theta}^{(t+1)} = P_\Theta\big(\boldsymbol{\theta}^{(t)} - \gamma \widehat{g}^{(t)}\big)$ with $\widehat{g}^{(t)} = -\widehat{s}_i^N\big(\boldsymbol{\theta}^{(t)}\big)$ as in Equation (7);
end

A possible solution is then to define $\widehat{g}^{(t)}$ by averaging the estimator (7) across individuals. In what follows, we rely on a mini-batch strategy where we use a single individual at a time. Specifically, at iteration $t \in \mathbb{N}^*$, we draw an individual $i(t)$ uniformly in $\{1, \ldots, n\}$ and the gradient estimator within the update Equation (6) is

$$\widehat{g}^{(t)} = -\widehat{s}_{i(t)}^N\big(\boldsymbol{\theta}^{(t)}\big). \tag{8}$$

4.2. Convergence guarantees

The SNIS estimator (7) is strongly consistent, but exhibits bias for a fixed sample size N. As shown in this section, controlling this bias is essential to ensure the convergence of our algorithm. This requires imposing further constraints on the optimization problem. In the remainder of the paper, let $\Theta \subset \mathbb{R}^d$ be a nonempty, compact, and convex set. In place of the standard update (6), we employ a projected SGD algorithm, which is defined by

$$\boldsymbol{\theta}^{(t+1)} = P_\Theta\big(\boldsymbol{\theta}^{(t)} - \gamma \widehat{g}^{(t)}\big), \tag{9}$$

where P_Θ stands for the orthogonal projection on Θ. Algorithm 1 summarizes the overall scheme. Formally, the projection step ensures that all iterates remain bounded, which in turn guarantees the L-smoothness and the bounded gradient conditions commonly assumed in the literature to establish convergence properties.

Preliminary result for L-smooth functions In the following, we delve into the analysis of an SGD algorithm (i) for non-convex and constrained optimization with L-smooth objective, (ii) with a gradient estimator that is both biased and random. To the best of our knowledge, no established convergence theorem exists for these conditions. Mai and Johansson (2021, Lemma 3.2) achieve a similar result but for unbiased gradient estimator. Let $F : \mathbb{R}^d \to \mathbb{R}$ be defined by

$$F(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) + I_\Theta(\boldsymbol{\theta}), \quad I_\Theta(\boldsymbol{\theta}) = \begin{cases} 0 & \text{if } \boldsymbol{\theta} \in \Theta, \\ \infty & \text{if } \boldsymbol{\theta} \notin \Theta. \end{cases} \tag{10}$$

For a real $\eta > 0$, its Moreau envelope is

$$F_\eta(\boldsymbol{\theta}) = \inf_{\boldsymbol{\theta}' \in \mathbb{R}^d} \left\{ F(\boldsymbol{\theta}') + \frac{1}{2\eta} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2 \right\}.$$

The Moreau envelope is useful from an optimization perspective, as it has the same set of minimizers as F, while being differentiable, unlike F. This characteristic is instrumental in analyzing the convergence of proximal gradient methods such as the one from Equation (9). Our following result yields control over the Moreau envelope, providing we have a L-smooth loss function and a gradient estimator $\widehat{g}^{(t)}$ with bounded mean squared error and bias (see Appendix B for proof).

Lemma 4.2 (Descent lemma)**.** *Let consider the gradient scheme as defined by Equation (9). Assume that*

(i) the function ℓ is L-smooth on Θ, and denote $\Gamma = \sup_{\boldsymbol{\theta} \in \Theta} \|\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})\|$;
(ii) for $t \in \mathbb{N}^*$, one has

$$\sigma^{(t)} = \mathbb{E}\left[\left\| \widehat{g}^{(t)} - \nabla_{\boldsymbol{\theta}} \ell\left(\boldsymbol{\theta}^{(t)}\right) \right\|^2 \,\middle|\, \boldsymbol{\theta}^{(t)} \right] < \infty,$$
$$\xi^{(t)} = \left\| \mathbb{E}\left[\widehat{g}^{(t)} \,\middle|\, \boldsymbol{\theta}^{(t)} \right] - \nabla_{\boldsymbol{\theta}} \ell\left(\boldsymbol{\theta}^{(t)}\right) \right\| < \infty.$$

Then, for any real constant $\eta \in (0, 1/\max\{2\Gamma + L, 2L\}]$,

$$\begin{aligned} \mathbb{E}\left[F_\eta\left(\boldsymbol{\theta}^{(t+1)}\right) \,\middle|\, \boldsymbol{\theta}^{(t)} \right] \leq F_\eta\left(\boldsymbol{\theta}^{(t)}\right) - \frac{\gamma}{2} \left\| \nabla F_\eta\left(\boldsymbol{\theta}^{(t)}\right) \right\|^2 \\ + \frac{\gamma + \gamma^2 \Gamma}{\eta} \xi^{(t)} + \frac{\gamma^2}{2\eta} \left(\sigma^{(t)} + \Gamma^2 \right). \end{aligned} \tag{11}$$

Remark that Γ is finite as the gradient is L-Lipschitz continuous on a bounded set. Consequently the interval $(0, 1/\max\{2\Gamma + L, 2L\}]$ is nonempty.

Applicability to the PLN-PCA model We now demonstrate that the loss function for the PLN-PCA model and the importance sampling estimate (8) satisfy the assumptions from Lemma 4.2.

Proposition 4.3. *Under Model (2), for any nonempty compact subset $\Theta \subset \mathbb{R}^d$, we have that*

(i) for any individual $i = 1, \ldots, n$, there exists a real

$$0 < \zeta_i = \inf_{\theta \in \Theta} p_{\boldsymbol{\theta}}(\mathbf{Y}_i);$$

(ii) there exists a real $L \geq 0$ such that the objective function ℓ, as defined in (3), is L-smooth on Θ.

Regarding Assumption *(ii)* of Lemma 4.2, we should note that both $\sigma^{(t)}$ and $\xi^{(t)}$ are driven by, respectively, the mean squared error and the bias of the importance sampling estimate (7). Agapiou et al. (2017, Theorem 2.3) yield

sufficient conditions to have control over the bias and error. The latter led to the subsequent assumptions on the proposal distribution for our optimization problem.

Assumption 1. *For all individual $i \in \{1, \dots, n\}$, the proposal distribution ν_i is chosen such that*

$$\lambda_i = \sup_{(\boldsymbol{\theta}, \mathbf{V}) \in \Theta \times \mathbb{R}^q} \rho_{\boldsymbol{\theta}, i}(\mathbf{V}) < \infty,$$
$$\beta_i = \sup_{\boldsymbol{\theta} \in \Theta} \mathbb{E}_{\nu_i(\cdot\,; \boldsymbol{\theta})} \left[\| \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{V}) \|_1^4 \right] < \infty.$$

Following the work of Agapiou et al. (2017), we obtain finite bound when integrating the mean squared error and the bias of the importance sampling estimate (7) with respect to the random mini-batch index $i(t)$ (see Appendix C for proof).

Proposition 4.4. *Let $\{i(t)\}_{t \in \mathbb{N}^*}$ and $\{\boldsymbol{\theta}^{(t)}\}_{t \in \mathbb{N}^*}$ be the associated random sequences generated by Algorithm 1. Under Model (2), if Assumption 1 holds, then, for all $t \in \mathbb{N}$,*

$$\overline{\sigma}_{\mathrm{IS}}^{(t)} = \mathbb{E}\left[\left\| \widehat{s}_{i(t)}^N\left(\boldsymbol{\theta}^{(t)}\right) - \nabla \log p_{\boldsymbol{\theta}^{(t)}}\left(\mathbf{Y}_{i(t)}\right) \right\|^2 \,\middle|\, \boldsymbol{\theta}^{(t)} \right] \leq \frac{d}{N} M_\sigma,$$
$$\overline{\xi}_{\mathrm{IS}}^{(t)} = \mathbb{E}\left[\left\| \mathbb{E}\left[\widehat{s}_{i(t)}^N\left(\boldsymbol{\theta}^{(t)}\right) - \nabla \log p_{\boldsymbol{\theta}^{(t)}}\left(\mathbf{Y}_{i(t)}\right) \,\middle|\, \boldsymbol{\theta}^{(t)}, i(t) \right] \right\| \,\middle|\, \boldsymbol{\theta}^{(t)} \right] \leq \frac{\sqrt{d}}{N} M_\xi,$$

where M_σ and M_ξ are two finite and positive constants given by

$$M_\sigma = \frac{12}{n} \sum_{i=1}^n \frac{\lambda_i^2 \sqrt{\beta_i}}{\zeta_i^2} \left(1 + \frac{250 \lambda_i}{\zeta_i} + \frac{9 \lambda_i^2}{\zeta_i^2} \right),$$
$$M_\xi = \frac{4}{n} \sum_{i=1}^n \frac{\lambda_i^2 \beta_i^{1/4}}{\zeta_i^2} \left\{ 2 + \sqrt{3 \left(1 + \frac{250 \lambda_i}{\zeta_i} + \frac{9 \lambda_i^2}{\zeta_i^2} \right)} \right\},$$

with λ_i and β_i as in Assumption 1.

As detailled in the proof of Theorem 4.5 in Appendix C, this result implies the conditions required on $\sigma^{(t)}$ and $\xi^{(t)}$, since there is a constant A such that

$$\sigma^{(t)} \leq A \left(\overline{\sigma}_{\mathrm{IS}}^{(t)} + \overline{\xi}_{\mathrm{IS}}^{(t)} + 1 \right), \quad \text{and} \quad \xi^{(t)} \leq \overline{\xi}_{\mathrm{IS}}^{(t)}.$$

Convergence of the gradient mapping As the local minimum of the loss function may lie outside the compact set Θ, we cannot prove that the norm of $\nabla \ell(\theta)$ becomes arbitrarily small within Algorithm 1. Instead, in the context of gradient methods incorporating a projection step, the convergence rate is characterized in terms of the norm of the gradient mapping $G_\eta^{(t)}$ (Drusvyatskiy and Paquette, 2019) defined for any real $\eta > 0$ by

$$G_\eta^{(t)} = \frac{\boldsymbol{\theta}^{(t)} - P_\Theta\left(\boldsymbol{\theta}^{(t)} - \eta \nabla \ell\left(\boldsymbol{\theta}^{(t)}\right)\right)}{\eta}. \tag{12}$$

This mapping is a tailored gradient objective, specifically modified to handle the projection step, whose norm is equivalent to the gradient norm of the Moreau envelope (see Lemma B.3). Our next result shows that the norm of the gradient mapping for Algorithm 1 can be rendered arbitrarily small, provided we use a sufficiently large number of iterations T and particles N.

Theorem 4.5. *Let $\boldsymbol{\theta}^{(1)} \in \Theta$ be an initial value and $\gamma_0 \in \mathbb{R}_+^*$ a user-specified initial learning rate. Under Model (2), if Assumption 1 holds, then for any $T \in \mathbb{N}^*$ and any real constant $\eta \in (0, 1/\max\{2\Gamma, L\}]$, the sequence $\{\boldsymbol{\theta}^{(t)}\}_{1 \leq t \leq T}$ defined by Algorithm 1 with $\gamma = \gamma_0/\sqrt{T}$ satisfies*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\left\|G_\eta^{(t)}\right\|^2\right] \leq \frac{2\tau}{\gamma_0(L\eta+1)\sqrt{T}}\left(\ell\left(\boldsymbol{\theta}^{(1)}\right) - \ell(\boldsymbol{\theta}^\star) + \frac{\gamma_0^2(L\eta+1)}{2\eta}\left[\Delta^2 + \Gamma^2 + \frac{d}{N}\left\{M_\sigma + \frac{2(\Delta+\Gamma)}{\sqrt{d}}M_\xi\right\}\right]\right) + \frac{2\tau\sqrt{d}}{\eta N}M_\xi, \tag{13}$$

with L the smoothness constant of $\ell(\cdot)$, $\Gamma = \sup_{\boldsymbol{\theta}\in\Theta}\|\nabla\ell(\boldsymbol{\theta})\|$, constants M_σ and M_ξ as defined in Proposition 4.4, and

$$\tau = \frac{(2L\eta+1)^2}{L\eta+1}\left(1+\sqrt{\frac{L\eta}{L\eta+1}}\right)^2, \quad \Delta = \max_{i=1,\ldots,n}\sup_{\boldsymbol{\theta}\in\Theta}\left\|\nabla_{\boldsymbol{\theta}}\ell\left(\boldsymbol{\theta}\right) + \nabla_{\boldsymbol{\theta}}\log p_{\boldsymbol{\theta}}\left(\mathbf{Y}_i\right)\right\|.$$

Remark 1. While the upper bound $1/\max\{2\Gamma, L\}$ is unknown, Algorithm 1 does not depend on the choice of η in practice, and therefore neither Γ nor L need to be estimated.

5. Importance sampling proposal choice

In this section, we propose a specific choice of proposal distribution that satisfies Assumption 1, namely a mixture distribution. Mixture distributions are often chosen for this task (*e.g.*, Cappé et al., 2008) because of their flexibility as parametric models. In what follows, we leverage the model structure and focus on two-component Gaussian mixture distributions. Given two real constants $\alpha \in [0,1]$ and $\delta > 0$, denote the two-component Gaussian mixture with mean $\boldsymbol{\mu} \in \mathbb{R}^q$ and covariance $\mathbf{S} \in \mathcal{S}_{++}^q$ by

$$\mathcal{GM}(\cdot; \boldsymbol{\mu}, \mathbf{S}, \alpha, \delta) = (1-\alpha)\mathcal{N}(\cdot; \boldsymbol{\mu}, \mathbf{S}) + \alpha\mathcal{N}(\cdot; \boldsymbol{\mu}, \delta\mathbf{I}_q). \tag{14}$$

For each individual $i = 1, \ldots, n$, the proposal distribution $\nu_i(\cdot\,; \boldsymbol{\theta})$ is set to such a mixture, and $\boldsymbol{\mu}$ and $\mathbf{S}$ are iteratively adapted according to the current estimate $\boldsymbol{\theta}^{(t)}$ (see Algorithm 2). However, to ensure convergence, it is necessary to impose conditions on (14) (see Appendix D for proof).

Lemma 5.1. *Let $\alpha \in (0,1]$ and $\delta > 1$. If for any $i = 1, \ldots, n$, $\boldsymbol{\theta} \mapsto \boldsymbol{\mu}_i(\boldsymbol{\theta}) \in \mathbb{R}^q$ and $\boldsymbol{\theta} \mapsto \mathbf{S}_i(\boldsymbol{\theta}) \in \mathcal{S}_{++}^q$ are continuous on Θ, then the proposal distribution defined by*

$$\nu_i(\cdot\,; \boldsymbol{\theta}) = \mathcal{GM}(\cdot; \boldsymbol{\mu}_i(\boldsymbol{\theta}), \mathbf{S}_i(\boldsymbol{\theta}), \alpha, \delta) \tag{15}$$

fulfils Assumption 1.

Constraining the parameter δ to live in $(1, +\infty)$ is sufficient to guarantee that the Radon–Nikodym derivative $\rho_{\boldsymbol{\theta},i}$ is uniformly bounded with respect to $\boldsymbol{\theta} \in \Theta$. However, this does not ensure an efficient gradient estimator in terms of error or bias. The bias and the error are both related to the Kullback–Leibler divergence $\mathrm{KL}[p_{\boldsymbol{\theta}}(\cdot \mid \mathbf{Y}_i) \| \nu_i(\cdot\,; \boldsymbol{\theta})]$ (Agapiou et al., 2017; Chatterjee and Diaconis, 2018). Specifically, these results show that the estimator exhibits enhanced efficiency for a fixed computational budget, as the Kullback–Leibler divergence decreases. Adaptive importance sampling addresses the minimization of $\nu \mapsto \mathrm{KL}[p_{\boldsymbol{\theta}}(\cdot \mid \mathbf{Y}_i) \| \nu]$ over a given class of probability measure. For instance, the Population Monte Carlo proposed by Cappé et al. (2008) provides a solution when the proposal is a mixture distribution. Nevertheless, implementing such methods within an SGD scheme can be computationally intensive. Indeed, the target distribution of the adaptive scheme changes at each iteration of the gradient scheme, necessitating a full run of the adaptive method at each iteration.

Practical implementation In the following, we present a simpler heuristic that is efficient for the class of problems presented in this paper, albeit not optimal in terms of the Kullback–Leibler divergence. We consider using the mean and covariance of the conditional distribution $p_{\boldsymbol{\theta}}(\mathbf{W}_i \mid \mathbf{Y}_i)$, namely

$$\boldsymbol{\mu}_i(\boldsymbol{\theta}) = \mathbb{E}[\mathbf{W}_i \mid \mathbf{Y}_i, \boldsymbol{\theta}], \quad \mathbf{S}_i(\boldsymbol{\theta}) = \mathbb{V}[\mathbf{W}_i \mid \mathbf{Y}_i, \boldsymbol{\theta}]. \tag{16}$$

Both functions are continuous on Θ (see Appendix D). Moreover, the parameter α can be interpreted as a regularization parameter. Indeed, in the limiting case $\alpha = 0$, the proposal distribution as defined in (15) resumes to the optimal Gaussian proposal distribution, that is

$$\nu_i(\cdot\,; \boldsymbol{\theta}) = \arg\min_{\nu \in \mathcal{F}} \mathrm{KL}[p_{\boldsymbol{\theta}}(\cdot \mid \mathbf{Y}_i) \| \nu], \quad \mathcal{F} = \left\{ \mathcal{N}(\boldsymbol{\mu}, \mathbf{S})\,;\, \boldsymbol{\mu} \in \mathbb{R}^q,\, \mathbf{S} \in \mathcal{S}_{++}^q \right\}. \tag{17}$$

While the Radon–Nikodym derivative $\rho_{\boldsymbol{\theta},i}$ with respect to such a proposal may not necessarily be bounded for any $\boldsymbol{\theta} \in \Theta$, it points out the effect of α. The mixture distribution (15) balances a component that informs on the intractable conditional distribution and a regularization or defensive component that plays a similar role to that of Cappé et al. (2008). Practically speaking, we should opt for a small value of α to improve the efficiency of the importance sampling.

Both $\boldsymbol{\mu}_i(\boldsymbol{\theta})$ and $\mathbf{S}_i(\boldsymbol{\theta})$ are unknown and must be estimated. Obviously, we can use the importance sampling method, since we could simply recycle the particles simulated to estimate the gradient. However, such a solution may lead to poorly conditioned and non-positive definite matrix estimates for $\mathbf{S}_i(\boldsymbol{\theta})$. A more robust alternative can be achieved using the Hessian of the log-complete likelihood:

$$\begin{aligned} \mathbf{S}_i^H(\boldsymbol{\theta}) &= -\left[\nabla^2_{\mathbf{w}} \log p_{\boldsymbol{\theta}}\left(\mathbf{Y}_i, \mathbf{w}\right)_{|\mathbf{w}=\boldsymbol{\mu}_i(\boldsymbol{\theta})}\right]^{-1} \\ &= \left[\mathbf{I}_q + \mathbf{C}^\top \mathrm{Diag}[\exp\{f_i(\boldsymbol{\mu}_i(\boldsymbol{\theta}); \mathbf{B}, \mathbf{C})\}]\mathbf{C}\right]^{-1}. \end{aligned} \tag{18}$$

Algorithm 2: Adaptive Importance Sampling based Gradient Descent

Input: initial point $\theta^{(1)}$, learning rate γ, number of iterations T, number of Monte Carlo draws N, mixture parameter $0 < \alpha \leq 1$, parameter $\delta > 0$.
Output: the sequence $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T+1)}$
for $t = 1$ **to** T **do**
  Sample i uniformly in $\{1, \ldots, n\}$;
  Compute the estimate $\widehat{\boldsymbol{\mu}}_i$ of $\boldsymbol{\mu}_i(\boldsymbol{\theta}^{(t)})$ as defined in Equation (16) ;
  Compute the estimate $\widehat{\mathbf{S}}_i$ of $\mathbf{S}_i(\boldsymbol{\theta}^{(t)})$ as defined in Equations (16) or (18) ;
  Set $\nu_i = \mathcal{GM}(\cdot; \widehat{\boldsymbol{\mu}}_i, \widehat{\mathbf{S}}_i, \alpha, \delta)$;
  Sample $(\mathbf{v}_{i,1}, \ldots, \mathbf{v}_{i,N})$ from $\nu_i^{\otimes N}\big(\cdot\,; \boldsymbol{\theta}^{(t)}\big)$;
  Compute $\boldsymbol{\theta}^{(t+1)} = P_\Theta\big(\boldsymbol{\theta}^{(t)} - \gamma \widehat{g}^{(t)}\big)$ with $\widehat{g}^{(t)} = -\widehat{s}_i^N\big(\boldsymbol{\theta}^{(t)}\big)$ as in Equation (7);
end

This alternative stems from the second order Taylor expansion of the complete log-likelihood, and has been used in various contexts, such as posterior approximation (Tierney and Kadane, 1986) and importance sampling (Owen, 2013, Chapter 9). Interestingly, for a Gaussian distribution, the Taylor expansion exactly relates the curvature of the scalar field at its mode to the variance, namely $\mathbf{S}_i^H(\boldsymbol{\theta})$ corresponds to the variance. In contrast to a Monte Carlo estimate of the covariance of the conditional distribution, it directly follows from Equation (18) that the Hessian of the log-complete likelihood, and consequently its inverse, is definite positive. Moreover, Lemma 5.1 also applies to this choice of covariance matrix, since the function $\boldsymbol{\theta} \mapsto \mathbf{S}_i^H(\boldsymbol{\theta})$ is continuous as a composition of continuous functions.

6. Simulation study

Datasets Both synthetic data and real data analysis are based on the scMARK dataset (Diaz-Mejia, 2021). The latter is a benchmark for single-cell Ribonucleic acid (scRNA) data designed to serve as an RNA-seq equivalent of the MNIST dataset — each cell being labeled by one of the 28 possible cell types. It corresponds to $n = 19998$ samples (cells), $p = 14059$ features (gene expression), and the design matrix resumes to an intercept ($d = 1$).

Competitors We compare different variants of our algorithm corresponding to specific choices of the proposal distribution $\nu_i(\cdot; \boldsymbol{\theta})$:

- **ISGD-VEM:** we set the proposal distribution to the variational distribution ϕ_ψ, as defined in Equation (4). Although it differs from the optimal proposal distribution (17), it represents a natural choice as it serves as the optimal Gaussian surrogate for the conditional distribution in terms of the Kullback–Leibler divergence $\phi \mapsto \mathrm{KL}[\phi \| p_{\boldsymbol{\theta}}(\cdot \mid \mathbf{Y}_i)]$.
- **ISGD-VEMmix:** there is no guarantee that the Radon–Nikodym derivative with respect to the VEM proposal is bounded, and thus ensures the

convergence of Algorithm 1. To address this, we introduce a defensive component in this version, and consider

$$\nu_i(\cdot\,;\boldsymbol{\theta}) = \mathcal{GM}(\cdot\,;\mathbf{m}_i^{\text{VEM}}, \mathbf{S}_i^{\text{VEM}}, \alpha, \delta),$$

where $\mathbf{m}_i^{\text{VEM}}$ and $\mathbf{S}_i^{\text{VEM}}$ are the mean and covariance of the variational distribution ϕ_ψ.

- **AISGD-SNIS:** it corresponds to Algorithm 2 and the choice of the mixture distribution

$$\nu_i(\cdot\,;\boldsymbol{\theta}) = \mathcal{GM}(\cdot\,;\widehat{\mathbf{m}}_i(\boldsymbol{\theta}), \widehat{\mathbf{S}}_i(\boldsymbol{\theta}), \alpha, \delta),$$

where $\widehat{\mathbf{m}}_i(\boldsymbol{\theta})$ and $\widehat{\mathbf{S}}_i(\boldsymbol{\theta})$ are the SNIS estimators the mean and the covariance of the conditional distribution $p_{\boldsymbol{\theta}}(\cdot \mid \mathbf{Y}_i)$.

- **AISGD-Hessian:** it corresponds to Algorithm 2 and the choice of the mixture distribution

$$\nu_i(\cdot\,;\boldsymbol{\theta}) = \mathcal{GM}(\cdot\,;\widehat{\mathbf{m}}_i(\boldsymbol{\theta}), \mathbf{S}_i^H(\boldsymbol{\theta}), \alpha, \delta),$$

with $\mathbf{S}_i^H(\boldsymbol{\theta})$, as defined in Equation (18).

All these methods are initialized with the variational estimator $\hat{\boldsymbol{\theta}}^{\text{VEM}}$ fitted the standard VEM algorithm implemented in `pyPLNmodels`[1].

6.1. Synthetic data

Data generation We consider simulation settings such that $n = 250$ individuals, $p = 30$ variables, $d = 1$ covariate (that is, one intercept), and rank constraints $q = 3, 5, 15$. The offset term $\mathbf{o}$ is set to zero. For each value of q, we mimic the scMARK dataset by sampling $m = 1, \ldots, M = 10$ datasets with parameters $\boldsymbol{\theta}^\star_{(q,m)}$ as follows:

(i) draw n individuals from the scMARK dataset and then select the p variables with the highest variance;
(ii) fit the standard PLN Model (1) on that reduced dataset with VEM from which we get $\mathbf{B}^0_{(m)} \in \mathcal{M}_{d\times p}(\mathbb{R})$ and $\boldsymbol{\Sigma}^0_{(m)} \in \mathcal{S}^p_{++}$;
(iii) set $\mathbf{B}^\star_{(q,m)}$ to $\mathbf{B}^0_{(m)}$ and $\boldsymbol{\Sigma}^\star_{(q,m)}$ to the best q-rank approximation of $\boldsymbol{\Sigma}^0_{(m)}$ derived from SVD.
(iv) generate a count matrix $\mathbf{Y}_{(q,m)}$ according to Model (2) for each $\boldsymbol{\theta}^\star_{(q,m)} = \left(\mathbf{B}^\star_{(q,m)}, \boldsymbol{\Sigma}^\star_{(q,m)}\right)$.

Experimental design Each algorithm is initialized at the variational estimate $\hat{\boldsymbol{\theta}}^{\text{VEM}}$ and run for a total of 200 epochs with two regimes of batch sizes $\mathfrak{B} = 250$ and 1 along iterations. Batches of size 250 — which correspond to the full dataset — are used for the first 180 epochs. During these, the gradient

[1] https://github.com/PLN-team/pyPLNmodels

estimator is computed as the average of the SNIS estimators (7) over $\mathfrak{B}$ individuals simultaneously drawn from the dataset. Namely, let $\mathfrak{I}$ denotes the set of randomly drawn individuals, we use

$$\widehat{g}^{(t)} = -\frac{1}{\mathfrak{B}} \sum_{i \in \mathfrak{I}} s_i^N \left(\boldsymbol{\theta}^{(t)} \right).$$

This initial part of the run can be viewed as a burn-in period that necessitates less computational effort. We adopt this practical approach based on empirical observations that the batch size has no significant impact on the convergence of our method. Our study focuses on the last 20 epochs with a batch size of 1, which corresponds to $T = 5000$ iterations of Algorithm 1. We set the mixture hyperparameters to $\alpha = 0.001$ and $\delta = 1.1$, and use $N = 500, 1000, 5000$ samples for the SNIS estimator (7). The learning rate γ is determined via a grid search.

Quality of the importance sampling proposal distribution The proposal distributions $\nu_i(\cdot\,; \boldsymbol{\theta})$ of the four competitors are compared at initialization with the three following metrics:

- the Kullback–Leibler divergence $\mathrm{KL}[\nu_i(\cdot\,; \boldsymbol{\theta}) \| p_{\boldsymbol{\theta}}(\cdot \mid \mathbf{Y}_i)]$, which gives a discrepancy measure in terms of the variational objective function, and thus provides a comparison between the proposal distribution and the variational distribution. Given a N-sample $\mathbf{v}_{i,1}, \ldots, \mathbf{v}_{i,N}$ from $\nu_i(\cdot\,; \boldsymbol{\theta})$, its Monte Carlo estimator is

$$-\log(N) - \frac{1}{N} \sum_{r=1}^{N} \log(\omega_{\boldsymbol{\theta},i,r}), \quad \omega_{\boldsymbol{\theta},i,r} = \frac{\rho_{\boldsymbol{\theta},i}(\mathbf{v}_{i,r})}{\sum_{s=1}^{N} \rho_{\boldsymbol{\theta},i}(\mathbf{v}_{i,s})}.$$

- the Kullback–Leibler divergence $\mathrm{KL}[p_{\boldsymbol{\theta}}(\cdot \mid \mathbf{Y}_i) \| \nu_i(\cdot\,; \boldsymbol{\theta})]$, which relates to the efficiency of the importance sampling scheme in terms of bias and quadratic error. Its Monte Carlo estimator is

$$\log(N) + \sum_{r=1}^{N} \omega_{\boldsymbol{\theta},i,r} \log(\omega_{\boldsymbol{\theta},i,r}).$$

- The Effective Sample Size (ESS), which assesses how accurately the weighted samples from the importance sampling method approximates the target distribution $p_{\boldsymbol{\theta}}(\cdot \mid \mathbf{Y}_i)$: a higher effective sample size indicates a better empirical approximation of the target distribution. It is estimated by

$$\left(\sum_{r=1}^{N} \omega_{\boldsymbol{\theta},i,r}^2 \right)^{-1}.$$

The results are displayed in Figure 1.

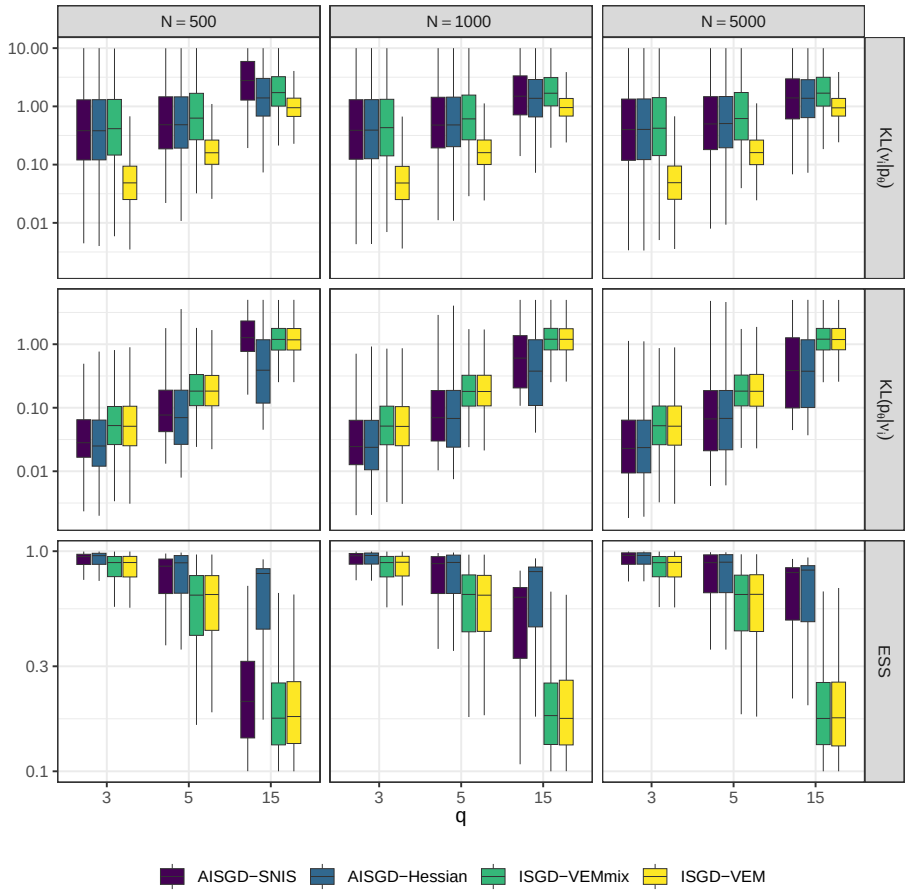

FIGURE 1. Distribution of the Kullback–Leibler divergence $\mathrm{KL}[\nu_i(\cdot\,;\boldsymbol{\theta})\|p_{\boldsymbol{\theta}}(\cdot \mid \mathbf{Y}_i)]$ *(top row), the Kullback–Leibler divergence* $\mathrm{KL}[p_{\boldsymbol{\theta}}(\cdot \mid \mathbf{Y}_i)\,\|\nu_i(\cdot\,;\boldsymbol{\theta})]$ *(middle row), and the effective sample size (bottom row) as a function of the number of Monte Carlo draws N, the rank constraint q, and the inference algorithms (AISGD-SNIS, AISGD-Hessian, ISGD-VEMmix, ISGD-VEM) at initialization (that is* $\boldsymbol{\theta}^{(0)} = \hat{\boldsymbol{\theta}}^{VEM}$*). Each boxplot is based on $M = 10$ synthetic datasets, with each metric estimated using the specified N Monte Carlo draws.*

A key distinction between the ISGD-VEM proposal and other methods is the use of a Gaussian mixture distribution. A comparison between ISGD-VEMmix and ISGD-VEM highlights the impact of adding a defensive component to the variational solution. As shown in the bottom row of Figure 1, the defensive component does not yield improvement in terms of the variational objective function. However, the ISGD-VEMmix proposal offers equivalent performances regarding the importance sampling method (middle and bottom rows of Figure 1).

Furthermore, the comparison of ISGD-VEMmix with the two AISGD variants (SNIS and Hessian) enables to evaluate the impact of replacing the diagonal

covariance from the variational distribution with either the conditional covariance estimate or the curvature estimate. Unsurprisingly, AISGD-SNIS becomes numerically unstable when we do not have enough Monte Carlo draws for a given rank constraint q. It is well known that estimating a covariance matrix is much more demanding when the dimension of the sample space increases. From the variational point of view (top row of Figure 1), AISGD-SNIS and AISGD-Hessian do not exhibit significant improvement compared to ISGD-VEMmix, with the exception of the $q = 15$ case where all methods performs similarly. This suggests that as the value of q increases, the covariance candidates are able to encompass more information about the conditional distribution than variational solutions. Finally, providing sufficient Monte Carlo draws are available, both AISGD-SNIS and AISGD-Hessian consistently demonstrate superior performance for the importance sampling scheme in comparison to variational-based proposals (middle and top row of Figure 1). Overall, AISGD-Hessian offers the most favourable practical performances.

Estimator performances We assess the performance of the resulting estimator $\boldsymbol{\theta}^{(T)}$ by comparing it to the initialization $\boldsymbol{\theta}^{(0)} = \hat{\boldsymbol{\theta}}^{\text{VEM}}$ using the likelihood ratio $p_{\boldsymbol{\theta}^{(0)}}(\mathbf{Y})/p_{\boldsymbol{\theta}^{(T)}}(\mathbf{Y})$, estimated by

$$\prod_{i=1}^{n} \left(\sum_{r=1}^{N} \omega_{\boldsymbol{\theta}^{(0)},i,r} \Big/ \sum_{r=1}^{N} \omega_{\boldsymbol{\theta}^{(T)},i,r} \right),$$

and the relative efficiency in terms of Root Mean Squared Error (RMSE), that is $(\mathbb{E}[\|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^{\star}\|^2]/\mathbb{E}[\|\boldsymbol{\theta}^{(T)} - \boldsymbol{\theta}^{\star}\|^2])^{1/2}$ estimated on the M dataset for each q. Following the previous conclusions, we only provide a comparison for AISGD-SNIS and AISGD-Hessian. The results are displayed in Figure 2. We observe an increase in the likelihood and a reduction of the RMSE for both methods.

6.2. Inference on the scMARK dataset

Due to memory limitations, we cannot address the problem with the original $p = 14059$ features. We make the inference for a dataset reduced to the $n = 50$ samples and $p = 30$ features with the largest variance.

For each rank constraint $q = 3, 5, 15$, we perform 30 independent runs of AISGD-Hessian initialized at $\boldsymbol{\theta}^{(0)} = \hat{\boldsymbol{\theta}}^{\text{VEM}}$ for 250 epochs, a batch size $\mathfrak{B} = 1, 10, 50$, and $N = 1000$ Monte Carlo draws. We do not consider the AISGD-SNIS variant since the robust version exhibits better performances, as illustrated in Section 6.1.

We monitor both the marginal log-likelihood and the gradient norm along the iterations. The results are presented in Figure 3. Although we have theoretical guarantees solely for a batch of size 1, we observe similar behavior of the method for larger batch sizes. The norm of the gradient of the objective function and the negative marginal log-likelihood decrease with T, suggesting that the VEM may not converge to the maximum likelihood estimator. Interestingly, it appears

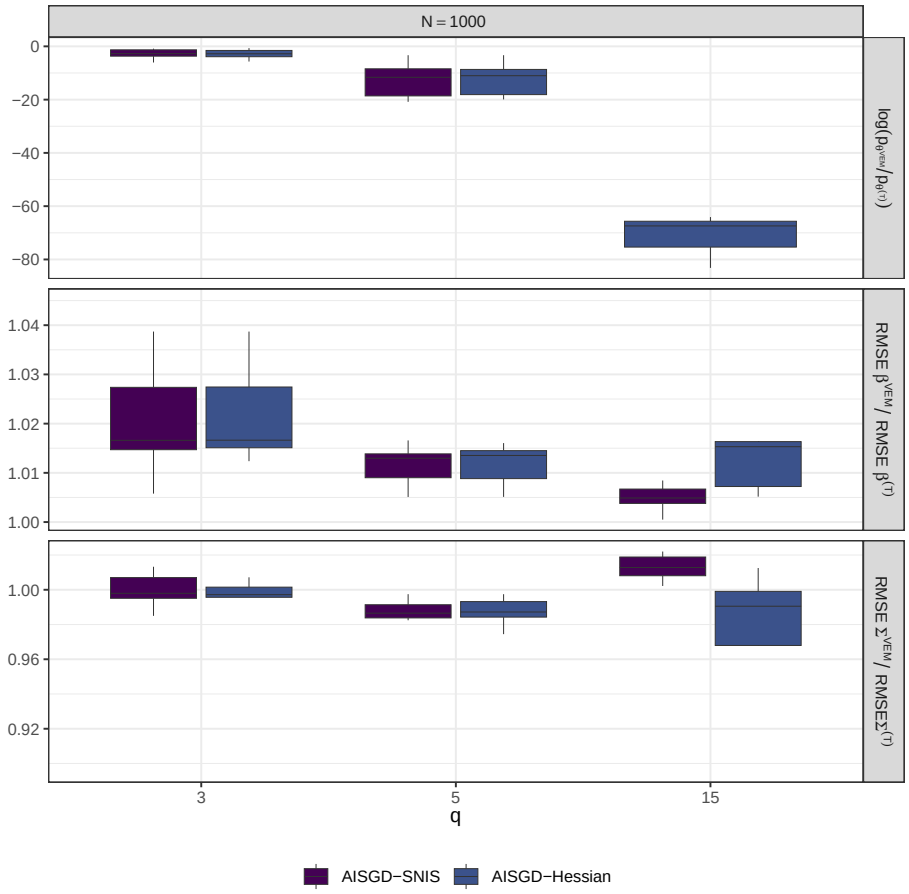

Figure 2. Distribution of the likelihood log-ratio $\log\{p_{\boldsymbol{\theta}^{\mathrm{VEM}}}(\mathbf{Y})/\, p_{\boldsymbol{\theta}^{(T)}}(\mathbf{Y})\}$ *(top row), and the Root Mean Squared Error (RMSE) relative efficiency* $(\mathbb{E}[\|\boldsymbol{\theta}^{\mathrm{VEM}} - \boldsymbol{\theta}^{\star}\|^2]/\,\mathbb{E}[\|\boldsymbol{\theta}^{(T)} - \boldsymbol{\theta}^{\star}\|^2])^{1/2}$ *specifically for* $\boldsymbol{\Sigma}$ *(middle row) and* $\mathbf{B}$ *(bottom row), as a function of the rank constraint* q*, and the inference algorithms (AISGD-SNIS, AISGD-Hessian). Each boxplot is based on* $M = 10$ *synthetic datasets, with the likelihood estimated using the specified* N *Monte Carlo draws. The log-likelihood ratio is omitted from the visualization for AISGD-SNIS when* $q = 15$ *due to its significantly high value above zero.*

that the larger the q, the more our method gains in comparison to the standard VEM approach. This can be attributed to the use of proposal distributions with full covariance structure, providing additional insights compared to the diagonal approximation used in the VEM.

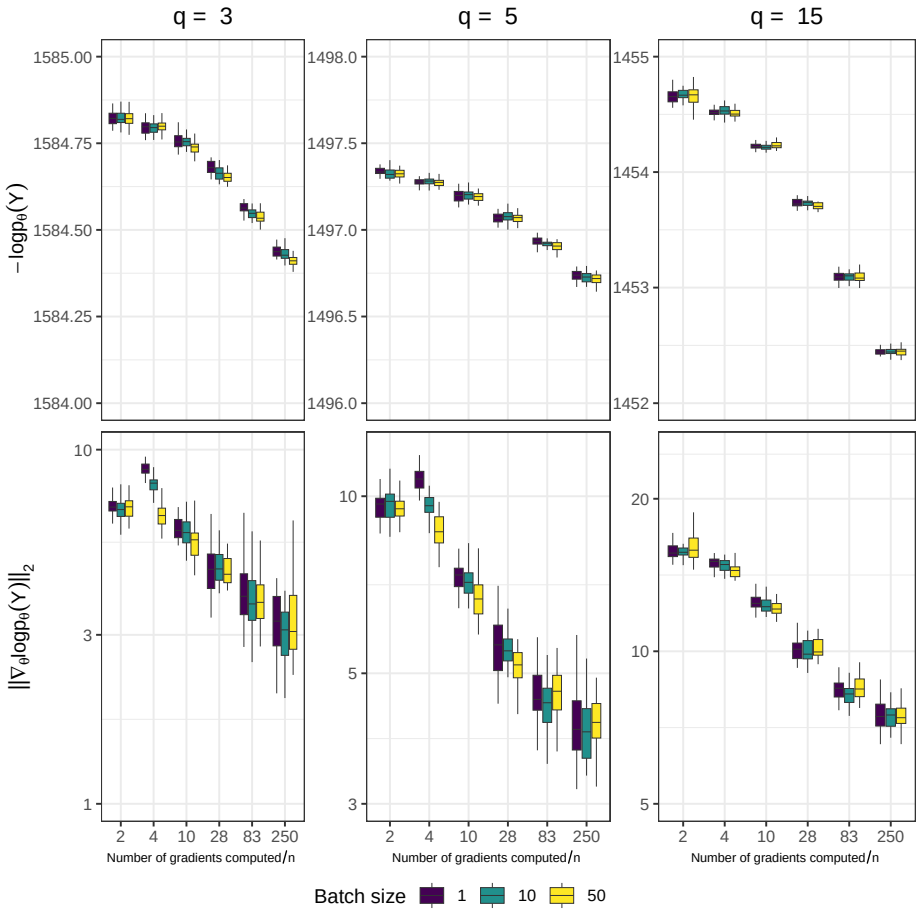

Figure 3. Numerical illustration of Theorem 4.5 on the scMARK dataset, reduced to $n = 50$ samples and $p = 30$ variables, for AISGD-Hessian initialized with the VEM estimator $\hat{\boldsymbol{\theta}}^{VEM}$. The boxplots represent the distribution of the log-likelihood (top row) and the gradient of the objective function (bottom row) over 30 independent runs as a function of the rank constraint q, the number of epochs, and the batch size.

7. Discussion

The paper explores the opportunity of a projected stochastic gradient scheme for inferring the parameters of a PLN-PCA model. Unlike competing variational approaches, our method allows to retrieve statistical guarantees on the resulting estimate. representing a significant step forward for the practitioners. This enables the construction of uncertainty measures and statistical tests. We have illustrated the benefits of our method on synthetic and real datasets using mixture distributions as proposals. Obviously we could resort to any distribution

provided it satisfies Assumption 1. A future work could be to investigate the potential of normalizing flows as proposal distribution, especially for larger rank constraints.

While we have proven the convergence for the PLN-PCA model using an importance sampling-based gradient estimator, the descent lemma 4.2 opens new vista for latent variable models. Indeed, its minimal assumptions offers broader applicability, and it can serve as the corner stone for extending Theorem 4.5 to models with L-smooth loss function and arbitrary random and biased gradient estimator.

A natural extension is to consider models where the emission distribution belongs to a natural (or canonical) exponential family. While deriving regularity of the log likelihood, and hence a score estimator, assessing the L-smoothness of the loss function is relatively straightforward for some families — typically when the natural parameter space is the entire vector space —, such as the Binomial distribution. Additional technical conditions may however be required for families with constrained natural parameter spaces, like the negative-Binomial or Gamma distributions. These conditions might involve integrability constraints on the moments of the emission distribution with respect to the latent distribution or modifications to the link function f_i to account for the parameter constraints.

Beside addressing the inference for other models, the assumption of lemma 4.2 on the gradient estimator allows to consider other, possibly more elaborated, simulation based method to estimate the gradient, such as diffusion models.

References

Agapiou, S., Papaspiliopoulos, O., Sanz-Alonso, D. and Stuart, A. M. (2017). Importance Sampling: Intrinsic Dimension and Computational Cost. *Statistical Science* **32** 405–431.

Aitchison, J. and Ho, C. H. (1989). The multivariate Poisson-log normal distribution. *Biometrika* **76** 643–653.

Ajalloeian, A. and Stich, S. U. (2021). On the Convergence of SGD with Biased Gradients.

Bertsekas, D. P. (1999). *Nonlinear Programming*. Athena Scientific.

Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association* 859–877.

Boyles, R. A. (1983). On the Convergence of the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **45** 47–50.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S. and Zhang, Q. (2018). JAX: composable transformations of Python+NumPy programs.

Bubeck, S. (2015). Convex Optimization: Algorithms and Complexity.

Cappé, O., Douc, R., Guillin, A., Marin, J. M. and Robert, C. P. (2008). Adaptive importance sampling in general mixture classes. *Statistics and Computing* **18** 447–459.

Chatterjee, S. and Diaconis, P. (2018). The sample size required in importance sampling. *The Annals of Applied Probability* **28** 1099–1135.

Chiquet, J., Mariadassou, M. and Robin, S. (2018a). Variational inference for probabilistic Poisson PCA. *Ann. Appl. Statist.* **12** 2674-2698.

Chiquet, J., Mariadassou, M. and Robin, S. (2018b). Variational inference for probabilistic Poisson PCA. *The Annals of Applied Statistics* **12** 2674 – 2698.

Chiquet, J., Mariadassou, M. and Robin, S. (2019). Variational Inference for sparse network reconstruction from count data. In *International Conference on Machine Learning*.

Chiquet, J., Mariadassou, M. and Robin, S. (2021). The Poisson-Lognormal Model as a Versatile Framework for the Joint Analysis of Species Abundances. *Frontiers in Ecology and Evolution* **9** 188.

Chiquet, J., Mariadassou, M., Robin, S., Poggiato, G. and Gindraud, F. (2023). PLNmodels: Poisson Lognormal Models PLN teams.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM algorithm. *Journal of the Royal Statistical Society: Series B* **39** 1–38.

Diaz-Mejia, J. (2021). scMARK an 'MNIST' like benchmark to evaluate and optimize models for unifying scRNA data.

Drusvyatskiy, D. and Paquette, C. (2019). Efficiency of minimizing compositions of convex functions and smooth maps. *Math. Program.* **178** 503–558.

Geyer, C. J. (1994). On the Convergence of Monte Carlo Maximum Likelihood Calculations. *Journal of the Royal Statistical Society. Series B (Methodological)* **56** 261–274.

Ghadimi, S. and Lan, G. (2013). Stochastic First- and Zeroth-order Methods for Nonconvex Stochastic Programming.

Gómez-Rubio, V. (2020). *Bayesian inference with INLA*. CRC Press.

Gutmann, M. U. and Hyvärinen, A. (2012). Noise-Contrastive Estimation of Unnormalized Statistical Models, with Applications to Natural Image Statistics. *Journal of Machine Learning Research* **13** 307–361.

Hall, P., Ormerod, J. T. and Wand, M. (2011). Theory of Gaussian variational approximation for a Poisson mixed model. *Statistica Sinica* 369–389.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural computation* **14** 1771–1800.

Hui, F. K. C. (2016). boral - Bayesian Ordination and Regression Analysis of Multivariate Abundance Data in r. *Methods in Ecology and Evolution* **7** 744–750.

Hui, F. K. C., Warton, D. I., Ormerod, J. T., Haapaniemi, V. and Taskinen, S. (2017). Variational Approximations for Generalized Linear Latent Variable Models. *Journal of Computational and Graphical Statistics* **26** 35–43.

Inouye, D. I., Yang, E., Allen, G. I. and Ravikumar, P. (2017). A review

of multivariate distributions for count data derived from the Poisson distribution. *Wiley Interdisciplinary Reviews: Computational Statistics* **9** e1398.

Kahn, H. (1949). *Stochastic (Monte Carlo) Attenuation Analysis*. Rand Corporation.

Kahn, H. and Harris, T. E. (1951). Estimation of particle transmission by random sampling. *National Bureau of Standards applied mathematics series* **12** 27–30.

Karlis, D. (2005). EM algorithm for mixed Poisson and other discrete distributions. *ASTIN Bulletin: The Journal of the IAA* **35** 3–24.

Louis, T. A. (1982). Finding the Observed Information Matrix When Using the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **44** 226-233.

Mai, V. V. and Johansson, M. (2021). Convergence of a Stochastic Gradient Method with Momentum for Non-Smooth Non-Convex Optimization.

Niku, J., Hui, F. K. C., Taskinen, S. and Warton, D. I. (2019). gllvm: Fast analysis of multivariate abundance data with generalized linear latent variable models in R. *Methods in Ecology and Evolution* **10** 2173-2182.

Ovaskainen, O., Tikhonov, G., Norberg, A., Blanchet, F. G., Duan, L., Dunson, D., Roslin, T. and Abrego, N. (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters* **20** 561–576.

Owen, A. B. (2013). *Monte Carlo theory, methods and examples.*

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L. et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32**.

Scaman, K., Malherbe, C. and Santos, L. D. (2022). Convergence Rates of Non-Convex Stochastic Gradient Descent Under a Generic Lojasiewicz Condition and Local Smoothness. In *Proceedings of the 39th International Conference on Machine Learning* **162** 19310–19327. PMLR.

Simon, M. K. (2002). *Probability distributions involving Gaussian random variables: A handbook for engineers and scientists*. Springer.

Stoehr, J. and Robin, S. S. (2024). Composite likelihood inference for the Poisson log-normal model.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the american statistical association* **81** 82–86.

Tikhonov, G., Opedal, O. H., Abrego, N., Lehikoinen, A., de Jonge, M. M. J., Oksanen, J. and Ovaskainen, O. (2020). Joint species distribution modelling with the r-package Hmsc. *Methods in Ecology and Evolution* **11** 442-447.

van der Vaart, A. (1998). *Asymptotic statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **27**. Cambridge Univ. Press, New York.

Westling, T. and McCormick, T. H. (2019). Beyond prediction: A framework for inference with variational approximations in mixture models. *Journal of Computational and Graphical Statistics* **28** 778–789.

Wu, C. F. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics* **11** 95–103.

Appendix A: Properties of the PLN-PCA model

Lemma A.1. *Under Model (2), for any compact set $\Theta \subset \mathbb{R}^d$, for all individual $i \in \{1, \dots, n\}$, there exists two constants $K_i^\Theta > 0$ and $\kappa_i^\Theta \in \mathbb{R}$, such that for any $\boldsymbol{\theta} \in \Theta$*

$$\log p_{\boldsymbol{\theta}} \left(\mathbf{Y}_i \mid \mathbf{W}_i \right) \leq K_i^\Theta \|\mathbf{W}_i\| + \kappa_i^\Theta.$$

Proof. Given $i \in \{1, \dots, n\}$, $\boldsymbol{\theta} = (\mathbf{B}, \mathbf{C}) \in \mathbb{R}^d$, due to the conditional independance, we have for $\mathbf{Z}_i = (Z_i j)$ as defined by Equation (2),

$$\log p_{\boldsymbol{\theta}}(\mathbf{Y}_i \mid \mathbf{W}_i) = \langle \mathbf{Y}_i, \mathbf{Z}_i \rangle - \sum_{j=1}^{p} \exp(Z_{ij})$$

Since exp is convex and differentiable on $\mathbb{R}$, for any $z, z_0 \in \mathbb{R}$, we have

$$\exp(z) \geq \exp(z_0) + (z - z_0) \exp(z_0).$$

It follows

$$\log p_{\boldsymbol{\theta}}(\mathbf{Y}_i \mid \mathbf{W}_i) \leq \langle \mathbf{Y}_i, \mathbf{Z}_i \rangle - \exp(z_0) \sum_{j=1}^{p} Z_{ij} - \sum_{j=1}^{p} (1 - z_0) \exp(z_0).$$

Using the definition of $\mathbf{Z}_i$ from Equation (2), we set

$$\begin{aligned} K_i(\boldsymbol{\theta}) &= \mathbf{C}^\top \{\mathbf{Y}_i - \exp(z_0) \mathbf{1}_p\}, \\ \kappa_i(\boldsymbol{\theta}) &= \langle \mathbf{Y}_i - \exp(z_0) \mathbf{1}_p, \mathbf{B}^\top \mathbf{x}_i + \mathbf{o}_i \rangle - \sum_{j=1}^{p} (1 - z_0) \exp(z_0). \end{aligned}$$

Then,

$$\log p_{\boldsymbol{\theta}}(\mathbf{Y}_i \mid \mathbf{W}_i) \leq \langle K_i(\boldsymbol{\theta}), \mathbf{W}_i \rangle + \kappa_i(\boldsymbol{\theta}),$$

and using the Cauchy–Schwarz inequality we get

$$\log p_{\boldsymbol{\theta}}(\mathbf{Y}_i \mid \mathbf{W}_i) \leq \|K_i(\boldsymbol{\theta})\| \|\mathbf{W}_i\| + \kappa_i(\boldsymbol{\theta}).$$

The functions K_i and κ_i are linear with respect to the parameter coordinates and, consequently, continuous on $\mathbb{R}^d$, and hence on Θ. Therefore, we can define the upper bounds $K_i^\Theta = \sup_{\boldsymbol{\theta} \in \Theta} \|K_i(\boldsymbol{\theta})\|$ and $\kappa_i^\Theta = \sup_{\boldsymbol{\theta} \in \Theta} \kappa_i(\boldsymbol{\theta})$, which provides the result.

□

Proposition 4.1. *For all individual $i = 1, \dots, n$, the incomplete log-likelihood $\boldsymbol{\theta} \mapsto \log p_{\boldsymbol{\theta}}(\mathbf{Y}_i)$ of Model (2) is twice continuously differentiable on $\mathbb{R}^d$ and its score function can be written as*

$$s_i(\boldsymbol{\theta}) = \int_{\mathbb{R}^q} \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{w}) p_{\boldsymbol{\theta}}(\mathrm{d}\mathbf{w} \mid \mathbf{Y}_i) = \mathbb{E}\left[\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{W}_i) \mid \mathbf{Y}_i, \boldsymbol{\theta}\right]. \quad (5)$$

Proof. The result is a direct application of the dominated convergence theorem. Indeed, for all individual $i = 1, \dots, n$, the likelihood writes as

$$p_{\boldsymbol{\theta}}(\mathbf{Y}_i) = \int_{\mathbb{R}^q} p_{\boldsymbol{\theta}}(\mathbf{Y}_i \mid \mathbf{w}) \mathcal{N}(\mathrm{d}\mathbf{w}; \mathbf{0}_q, \mathbf{I}_q).$$

The function $\boldsymbol{\theta} \mapsto p_{\boldsymbol{\theta}}(\mathbf{Y}_i \mid \mathbf{W}_i)$ is twice continuously differentiable on $\mathbb{R}^d$ as a composition of such functions. Moreover, each component Z_{ij}, $j = 1, \dots, p$, is a linear function of the components of $\boldsymbol{\theta}$, and for $r, s = 1, \dots, d$,

$$\frac{\partial}{\partial \theta_r} Z_{ij} = \sum_{k=1}^{m} x_{ik} \mathbb{1}_{\theta_r = B_{kj}} + \sum_{k=1}^{q} W_{ik} \mathbb{1}_{\theta_r = C_{jk}}, \quad \frac{\partial^2}{\partial \theta_s \partial \theta_r} Z_{ij} = 0. \quad (19)$$

Consequently, given a component θ_r, there is a unique index in $\{1, \dots, p\}$ such that the first partial derivative is non-zero. Denote by j and k such indices for the partial derivatives with respect to θ_r and θ_s, respectively. We then have

$$\frac{\partial}{\partial \theta_r} p_{\boldsymbol{\theta}}(\mathbf{Y}_i \mid \mathbf{W}_i) = p_{\boldsymbol{\theta}}(\mathbf{Y}_i \mid \mathbf{W}_i) \left\{Y_{ij} - \exp(Z_{ij})\right\} \frac{\partial}{\partial \theta_r} Z_{ij},$$

$$\frac{\partial^2}{\partial \theta_s \partial \theta_r} p_{\boldsymbol{\theta}}(\mathbf{Y}_i \mid \mathbf{W}_i) = p_{\boldsymbol{\theta}}(\mathbf{Y}_i \mid \mathbf{W}_i) \Bigg[\left\{Y_{ij} - \exp(Z_{ij})\right\} \left\{Y_{ik} - \exp(Z_{ik})\right\} \frac{\partial}{\partial \theta_r} Z_{ij} \frac{\partial}{\partial \theta_s} Z_{ik}$$

$$- \exp(Z_{ij}) \frac{\partial}{\partial \theta_r} Z_{ij} \frac{\partial}{\partial \theta_s} Z_{ij} \Bigg].$$

It follows from (19) that for any $r = 1, \dots, d$, and any $j = 1, \dots, p$

$$\left| \frac{\partial}{\partial \theta_r} Z_{ij} \right| \leq \|\mathbf{x}_i\| + \|\mathbf{W}_i\|.$$

Given a non-empty open and bounded set $\Theta \subset \mathbb{R}^d$, it follows from Lemma A.1 that it exists two constants $K_i^\Theta > 0$ and κ_i^Θ such that for any $\boldsymbol{\theta} \in \Theta$

$$p_{\boldsymbol{\theta}}(\mathbf{Y}_i \mid \mathbf{W}_i) \leq \exp\left(K_i^\Theta \|\mathbf{W}_i\| + \kappa_i^\Theta\right).$$

On the other hand, for any $\boldsymbol{\theta} \in \Theta$, any $j = 1, \dots, p$

$$|Z_{ij}| \leq A^\Theta \|\mathbf{W}_i\| + R_i^\Theta, \quad \text{with} \quad A^\Theta = \sup_{\boldsymbol{\theta} \in \Theta} \|\boldsymbol{\theta}\|, \quad R_i^\Theta = \|\mathbf{x}_i\| \sup_{\boldsymbol{\theta} \in \Theta} \|\boldsymbol{\theta}\| + \max_{j=1,\dots,p} |o_{ij}|.$$

Consequently, for any $j = 1, \dots, p$, since $\exp(Z_{ij}) \leq \exp(|Z_{ij}|)$,

$$|Y_{ij} - \exp(Z_{ij})| \leq \exp(|Z_{ij}|) \left\{\|\mathbf{Y}_i\| \exp(-|Z_{ij}|) + 1\right\} \leq \exp\left(A^\Theta \|\mathbf{W}_i\| + R_i^\Theta\right) \left(\|\mathbf{Y}_i\| + 1\right).$$

Thereby, for any $j = 1, \ldots, p$,

$$\begin{aligned}
\left|\{Y_{ij} - \exp(Z_{ij})\} \frac{\partial}{\partial \theta_r} Z_{ij}\right| &\leq (\|\mathbf{x}_i\| + \|\mathbf{W}_i\|)\,(\|\mathbf{Y}_i\| + 1) \exp\left(A^\Theta \|\mathbf{W}_i\| + R_i^\Theta\right) \\
\left|\exp(Z_{ij}) \frac{\partial}{\partial \theta_r} Z_{ij} \frac{\partial}{\partial \theta_s} Z_{ij}\right| &\leq (\|\mathbf{x}_i\| + \|\mathbf{W}_i\|)^2 \exp\left(A^\Theta \|\mathbf{W}_i\| + R_i^\Theta\right) \\
&\leq (\|\mathbf{x}_i\| + \|\mathbf{W}_i\|)^2 \exp\left(2A^\Theta \|\mathbf{W}_i\| + 2R_i^\Theta\right),
\end{aligned}$$

where the last inequality follows because A^Θ and R_i^Θ are both positive. Finally, we get that on Θ

$$\begin{aligned}
\left|\frac{\partial}{\partial \theta_r} p_{\boldsymbol{\theta}}(\mathbf{Y}_i \mid \mathbf{W}_i)\right| \exp\left(-\frac{\|\mathbf{W}_i\|^2}{2}\right) &\leq (\|\mathbf{x}_i\| + \|\mathbf{W}_i\|)\,(\|\mathbf{Y}_i\| + 1)\, Q_i^\Theta(\mathbf{W}_i, 1) \\
\left|\frac{\partial}{\partial \theta_s \partial \theta_r} p_{\boldsymbol{\theta}}(\mathbf{Y}_i \mid \mathbf{W}_i)\right| \exp\left(-\frac{\|\mathbf{W}_i\|^2}{2}\right) &\leq (\|\mathbf{x}_i\| + \|\mathbf{W}_i\|)^2 \left\{(\|\mathbf{Y}_i\| + 1)^2 + 1\right\} Q_i^\Theta(\mathbf{W}_i, 2),
\end{aligned}$$

with

$$Q_i^\Theta(\mathbf{W}_i, k) = \exp\left\{-\frac{1}{2}\|\mathbf{W}_i\|^2 + \left(K_i^\Theta + kA^\Theta\right)\|\mathbf{W}_i\| + \kappa_i^\Theta + kR_i^\Theta\right\}.$$

Crucially, each of these upper bounds is a Lebesgue integrable function on $\mathbb{R}^q$ that does not depend on $\boldsymbol{\theta} \in \Theta$. Consequently, we can conclude that the likelihood is twice continuously differentiable on Θ with the dominated convergence theorem. Since the result holds for any open and bounded set $\Theta \subset \mathbb{R}^d$, for all $\boldsymbol{\theta} \in \mathbb{R}^d$, we can apply it to an open d-ball with center $\boldsymbol{\theta}$. Therefore, the likelihood is twice continuously differentiable on $\mathbb{R}^d$. Moreover

$$\nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\mathbf{Y}_i) = \int_{\mathbb{R}^q} \nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\mathbf{Y}_i \mid \mathbf{w}) \mathcal{N}(\mathrm{d}\mathbf{w}; \mathbf{0}_q, \mathbf{I}_q)) = \int_{\mathbb{R}^q} \nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{w}) \mathrm{d}\mathbf{w}.$$

For all $\mathbf{Y}_i \in \mathbb{N}^p$, the likelihood $\boldsymbol{\theta} \mapsto p_{\boldsymbol{\theta}}(\mathbf{Y}_i)$ is positive on $\mathbb{R}^d$. Indeed, the integrand is positive everywhere since, by definition

$$\mathbf{w} \mapsto p_{\boldsymbol{\theta}}(\mathbf{Y}_i \mid \mathbf{w}) = \prod_{j=1}^{p} p(Y_{ij} \mid z_{ij}), \quad (z_{i1}, \ldots, z_{ip})^\top = f_i(\mathbf{w}; \mathbf{B}, \mathbf{C}),$$

and for any $z \in \mathbb{R}$, $y \mapsto p(y \mid z)$ is positive on $\mathbb{N}$. The continuous differentiability of the log-likelihood follows directly by composition. Finally,

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{Y}_i) = \frac{1}{p_{\boldsymbol{\theta}}(\mathbf{Y}_i)} \int_{\mathbb{R}^q} \nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{w}) \mathrm{d}\mathbf{w} &= \int_{\mathbb{R}^q} \frac{p_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{w}) \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{w})}{p_{\boldsymbol{\theta}}(\mathbf{Y}_i)} \mathrm{d}\mathbf{w} \\
&= \int_{\mathbb{R}^q} \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{w}) p_{\boldsymbol{\theta}}(\mathrm{d}\mathbf{w} \mid \mathbf{Y}_i).
\end{aligned}$$

□

Proposition 4.3. *Under Model (2), for any nonempty compact subset $\Theta \subset \mathbb{R}^d$, we have that*

(i) for any individual $i = 1, \ldots, n$, there exists a real

$$0 < \zeta_i = \inf_{\theta \in \Theta} p_{\boldsymbol{\theta}}(\mathbf{Y}_i);$$

(ii) there exists a real $L \geq 0$ such that the objective function ℓ, as defined in (3), is L-smooth on Θ.

Proof. Let $\Theta \subset \mathbb{R}^d$ be a nonempty compact set.

(i) From Proposition 4.1, the likelihood of the observed data $\boldsymbol{\theta} \mapsto p_{\boldsymbol{\theta}}(\mathbf{Y}_i)$ is continuous on the compact set Θ. Thus, it is bounded and attains its bounds on Θ. Moreover, the likelihood of the observed data is positive on $\mathbb{R}^d$, so is its lower bound on Θ.

(ii) A direct consequence of Proposition 4.1, is that $\boldsymbol{\theta} \mapsto \nabla \ell(\boldsymbol{\theta})$ is continuous differentiable on the compact set Θ. Therefore, it is Lipschitz continuous.

□

Appendix B: Intermediate optimization results and proof of Lemma 4.2

In this Section, we consider results for minimizing a function ℓ on $\mathbb{R}^d$. We denote $\Theta \subset \mathbb{R}^d$ a compact and convex set.

Lemma B.1 (Non-expansiveness, Bertsekas (1999, Prop 2.1.3))**.** *The projection P_Θ satisfies the non-expansiveness property, namely for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d$*

$$\left\| P_\Theta(\boldsymbol{\theta}) - P_\Theta(\boldsymbol{\theta}') \right\| \leq \left\| \boldsymbol{\theta} - \boldsymbol{\theta}' \right\|.$$

For a real $\eta > 0$, the proximal operator of F as defined in Equation (10) is given by

$$\operatorname{prox}_F^\eta(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}' \in \mathbb{R}^d}{\arg\min} \left\{ F(\boldsymbol{\theta}') + \frac{1}{2\eta} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2 \right\}.$$

Lemma B.2 (Drusvyatskiy and Paquette (2019, Lemma 4.3))**.** *If the function ℓ is L-smooth, for any real constant $\eta < 1/L$, the Moreau envelope F_η is differentiable on $\mathbb{R}^d$ and relates to the proximal operator through the following equation*

$$\nabla F_\eta(\boldsymbol{\theta}) = \frac{1}{\eta} \left\{ \boldsymbol{\theta} - \operatorname{prox}_F^\eta(\boldsymbol{\theta}) \right\}.$$

Lemma B.3 (Drusvyatskiy and Paquette (2019, Theorem 4.5))**.** *If the function ℓ is L-smooth, for any real constant $\eta > 0$, the gradient mapping $G_\eta^{(t)}$, as defined in Equation (12), satisfy the following inequality*

$$\left\|G_\eta^{(t)}\right\| \leq \left(1 + \frac{L\eta}{L\eta+1}\right)\left(1 + \sqrt{\frac{L\eta}{L\eta+1}}\right)\left\|\nabla F_{\frac{\eta}{L\eta+1}}\left(\boldsymbol{\theta}^{(t)}\right)\right\|.$$

A differentiable function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is said to be μ-strongly-convex, $\mu \in \mathbb{R}_+^*$, if for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d$,

$$f(\boldsymbol{\theta}') \geq f(\boldsymbol{\theta}) + \left\langle \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}), \boldsymbol{\theta}' - \boldsymbol{\theta} \right\rangle + \frac{\mu}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2.$$

Lemma B.4. *If the function ℓ is L-smooth, then, for any real constant $\eta \in (0, 1/L)$, and any $\bar{\boldsymbol{\theta}} \in \mathbb{R}^d$, the following function is $(\eta^{-1} - L)$-strongly convex*

$$f : \boldsymbol{\theta} \mapsto \ell(\boldsymbol{\theta}) + \frac{1}{2\eta}\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\|^2.$$

Proof. Let $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d$. We have

$$f(\boldsymbol{\theta}) + \left\langle \nabla f(\boldsymbol{\theta}), \boldsymbol{\theta}' - \boldsymbol{\theta} \right\rangle = \ell(\boldsymbol{\theta}) + \frac{1}{2\eta}\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\|^2 + \left\langle \nabla \ell(\boldsymbol{\theta}) + \eta^{-1}(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}), \boldsymbol{\theta}' - \boldsymbol{\theta} \right\rangle. \tag{20}$$

According to Bubeck (2015, Lemma 3.4), since ℓ is $L-$smooth, it satisfies for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d$

$$\left|\ell(\boldsymbol{\theta}') - \ell(\boldsymbol{\theta}) - \left\langle \nabla \ell(\boldsymbol{\theta}), \boldsymbol{\theta}' - \boldsymbol{\theta} \right\rangle\right| \leq \frac{L}{2}\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|^2. \tag{21}$$

It directly follows that

$$\ell(\boldsymbol{\theta}) + \langle \nabla \ell(\boldsymbol{\theta}), \boldsymbol{\theta}' - \boldsymbol{\theta} \rangle - \frac{L}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2 \leq \ell(\boldsymbol{\theta}').$$

Combining the latter inequality with Equation (20) yields

$$\begin{aligned} &f(\boldsymbol{\theta}) + \left\langle \nabla f(\boldsymbol{\theta}), \boldsymbol{\theta}' - \boldsymbol{\theta} \right\rangle + \frac{\eta^{-1} - L}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2 \\ &\quad \leq \ell(\boldsymbol{\theta}') + \frac{1}{2\eta}\left(\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\|^2 + 2\langle \boldsymbol{\theta} - \bar{\boldsymbol{\theta}}, \boldsymbol{\theta}' - \boldsymbol{\theta} \rangle + \|\boldsymbol{\theta} - \boldsymbol{\theta}\|^2\right) \end{aligned}$$

For a scalar product and its associated norm, the identity $\|a + b\|^2 = \|a\|^2 + 2\langle a, b\rangle + \|b\|^2$ gives

$$f(\boldsymbol{\theta}) + \left\langle \nabla f(\boldsymbol{\theta}), \boldsymbol{\theta}' - \boldsymbol{\theta} \right\rangle + \frac{\eta^{-1} - L}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2 \leq \ell(\boldsymbol{\theta}') + \frac{1}{2\eta}\|\boldsymbol{\theta}' - \bar{\boldsymbol{\theta}}\| = f(\boldsymbol{\theta}').$$

Since $\eta^{-1} + L > 0$, we can conlude that f is $(\eta^{-1} + L)$-strongly convex.

□

Lemma B.5. *If the function ℓ is L-smooth, then for any $\boldsymbol{\theta} \in \Theta$, and any real constant $\eta \in (0, (2\Gamma + L)^{-1}]$, with $\Gamma = \sup_{\boldsymbol{\theta}\in\Theta}\|\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta})\|$, we have*

$$\|\boldsymbol{\theta} - \operatorname{prox}_F^{\eta}(\boldsymbol{\theta})\| \leq 1.$$

Proof. Given $\boldsymbol{\theta} \in \Theta$, we show that any point at a distance more than one from $\boldsymbol{\theta}$ is not $\operatorname{prox}_F^{\eta}(\boldsymbol{\theta})$. This implies that necessarily

$$\|\boldsymbol{\theta} - \operatorname{prox}_F^{\eta}(\boldsymbol{\theta})\| \leq 1.$$

Let $\boldsymbol{\theta}' \in \mathbb{R}^d$ be a point such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\| > 1$. According to Bubeck (2015, Lemma 3.4), since ℓ is $L-$smooth, it satisfies for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d$

$$\ell(\boldsymbol{\theta}) \leq \ell(\boldsymbol{\theta}') + \langle \nabla\ell(\boldsymbol{\theta}'), \boldsymbol{\theta} - \boldsymbol{\theta}'\rangle + \frac{L}{2}\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|^2. \tag{22}$$

Let first assume that $\boldsymbol{\theta}' \in \Theta$. Using successively the Cauchy–Schwarz inequality, the positive bound Γ and the assumption $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\| > 1$, we get

$$\langle \nabla\ell(\boldsymbol{\theta}'), \boldsymbol{\theta} - \boldsymbol{\theta}'\rangle \leq \|\nabla\ell(\boldsymbol{\theta}')\|\|\boldsymbol{\theta} - \boldsymbol{\theta}'\| < \Gamma\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2. \tag{23}$$

It results from Equation (22) and Equation (23) that

$$\ell(\boldsymbol{\theta}) < \ell(\boldsymbol{\theta}') + \left(\Gamma + \frac{L}{2}\right)\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|^2.$$

For $\eta \leq (2\Gamma + L)^{-1}$, we have $(\Gamma + L/2) \leq 1/(2\eta)$. Moreover, the functions F and ℓ coincide on Θ, so that

$$F(\boldsymbol{\theta}) < F(\boldsymbol{\theta}') + \frac{1}{2\eta}\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|^2.$$

We can therefore conclude that $\boldsymbol{\theta}' \neq \operatorname{prox}_F^{\eta}(\boldsymbol{\theta})$. Conversely, if $\boldsymbol{\theta}' \notin \Theta$, by definition of F, we also have that $\boldsymbol{\theta}' \neq \operatorname{prox}_F^{\eta}(\boldsymbol{\theta})$.

□

Lemma 4.2 (Descent lemma)**.** *Let consider the gradient scheme as defined by Equation (9). Assume that*

(i) the function ℓ is L-smooth on Θ, and denote $\Gamma = \sup_{\boldsymbol{\theta}\in\Theta}\|\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta})\|$;
(ii) for $t \in \mathbb{N}^*$, one has

$$\sigma^{(t)} = \mathbb{E}\left[\left\|\widehat{g}^{(t)} - \nabla_{\boldsymbol{\theta}}\ell\left(\boldsymbol{\theta}^{(t)}\right)\right\|^2 \,\middle|\, \boldsymbol{\theta}^{(t)}\right] < \infty,$$
$$\xi^{(t)} = \left\|\mathbb{E}\left[\widehat{g}^{(t)} \,\middle|\, \boldsymbol{\theta}^{(t)}\right] - \nabla_{\boldsymbol{\theta}}\ell\left(\boldsymbol{\theta}^{(t)}\right)\right\| < \infty.$$

Then, for any real constant $\eta \in (0, 1/\max\{2\Gamma + L, 2L\}]$,

$$\mathbb{E}\left[F_\eta\left(\boldsymbol{\theta}^{(t+1)}\right) \,\middle|\, \boldsymbol{\theta}^{(t)}\right] \leq F_\eta\left(\boldsymbol{\theta}^{(t)}\right) - \frac{\gamma}{2}\left\|\nabla F_\eta\left(\boldsymbol{\theta}^{(t)}\right)\right\|^2 + \frac{\gamma + \gamma^2\Gamma}{\eta}\xi^{(t)} + \frac{\gamma^2}{2\eta}\left(\sigma^{(t)} + \Gamma^2\right). \tag{11}$$

Proof. Given $t \in \mathbb{N}^*$ and $\eta \in (0, \max\{2\Gamma + L, 2L\}^{-1}]$, we define the virtual iterates $\bar{\boldsymbol{\theta}}^{(t)}$ as

$$\bar{\boldsymbol{\theta}}^{(t)} = \operatorname{prox}_F^\eta\left(\boldsymbol{\theta}^{(t)}\right) = \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\mathbb{R}^d}\left\{F(\boldsymbol{\theta}) + \frac{1}{2\eta}\left\|\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}\right\|^2\right\}.$$

Note that $\bar{\boldsymbol{\theta}}^{(t)} \in \Theta$. Indeed, according to the definition of the Moreau envelope

$$F\left(\bar{\boldsymbol{\theta}}^{(t)}\right) = F_\eta\left(\boldsymbol{\theta}^{(t)}\right) - \frac{1}{2\eta}\left\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^{(t)}\right\|^2 \leq F\left(\boldsymbol{\theta}^{(t)}\right).$$

The upper bound is finite, since $\boldsymbol{\theta}^{(t)}$ as defined in (9) is within the compact Θ. The operator F is thus finite at $\bar{\boldsymbol{\theta}}^{(t)}$, and consequently $\bar{\boldsymbol{\theta}}^{(t)} \in \Theta$ since F takes infinite value outside of Θ.

As the virtual iterates are within Θ, the definition of the Moreau envelope yields

$$F_\eta\left(\boldsymbol{\theta}^{(t+1)}\right) \leq \ell\left(\bar{\boldsymbol{\theta}}^{(t)}\right) + \frac{1}{2\eta}\left\|\boldsymbol{\theta}^{(t+1)} - \bar{\boldsymbol{\theta}}^{(t)}\right\|^2. \tag{24}$$

Using the definition of $\boldsymbol{\theta}^{(t+1)}$ and the fact that $\bar{\boldsymbol{\theta}}^{(t)} \in \Theta$, we obtain from Lemma B.1

$$\begin{aligned}\frac{1}{2\eta}\left\|\boldsymbol{\theta}^{(t+1)} - \bar{\boldsymbol{\theta}}^{(t)}\right\|^2 &= \frac{1}{2\eta}\left\|P_\Theta\left(\boldsymbol{\theta}^{(t)} - \gamma\widehat{g}^{(t)}\right) - P_\Theta\left(\bar{\boldsymbol{\theta}}^{(t)}\right)\right\|^2 \\ &\leq \frac{1}{2\eta}\left\|\boldsymbol{\theta}^{(t)} - \gamma\widehat{g}^{(t)} - \bar{\boldsymbol{\theta}}^{(t)}\right\|^2.\end{aligned}$$

The bilinearity of the scalar product gives

$$\left\|\boldsymbol{\theta}^{(t)} - \gamma\widehat{g}^{(t)} - \bar{\boldsymbol{\theta}}^{(t)}\right\|^2 = \left\|\boldsymbol{\theta}^{(t)} - \bar{\boldsymbol{\theta}}^{(t)}\right\|^2 + 2\gamma\left\langle\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^{(t)}, \widehat{g}^{(t)}\right\rangle + \gamma^2\left\|\widehat{g}^{(t)}\right\|^2,$$

and the inequality (24) becomes

$$\begin{aligned}F_\eta\left(\boldsymbol{\theta}^{(t+1)}\right) &\leq \ell\left(\bar{\boldsymbol{\theta}}^{(t)}\right) + \frac{1}{2\eta}\left\|\boldsymbol{\theta}^{(t)} - \bar{\boldsymbol{\theta}}^{(t)}\right\|^2 + \frac{\gamma}{\eta}\left\langle\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^{(t)}, \widehat{g}^{(t)}\right\rangle + \frac{\gamma^2}{2\eta}\left\|\widehat{g}^{(t)}\right\|^2 \\ &\leq F_\eta\left(\boldsymbol{\theta}^{(t)}\right) + \frac{\gamma}{\eta}\left\langle\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^{(t)}, \widehat{g}^{(t)}\right\rangle + \frac{\gamma^2}{2\eta}\left\|\widehat{g}^{(t)}\right\|^2.\end{aligned}$$

Therefore,

$$\mathbb{E}\left[F_\eta\left(\boldsymbol{\theta}^{(t+1)}\right) \,\middle|\, \boldsymbol{\theta}^{(t)}\right] \leq F_\eta\left(\boldsymbol{\theta}^{(t)}\right) + \frac{\gamma}{\eta}\left\langle\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^{(t)}, \mathbb{E}\left[\widehat{g}^{(t)} \,\middle|\, \boldsymbol{\theta}^{(t)}\right]\right\rangle + \frac{\gamma^2}{2\eta}\mathbb{E}\left[\left\|\widehat{g}^{(t)}\right\|^2 \,\middle|\, \boldsymbol{\theta}^{(t)}\right].$$

We now study more in detail the second and the third terms of the right-hand side.

Regarding the second term, we have

$$\left\langle \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^{(t)}, \mathbb{E}\left[\widehat{g}^{(t)} \;\middle|\; \boldsymbol{\theta}^{(t)}\right] \right\rangle = \left\langle \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^{(t)}, \nabla_{\boldsymbol{\theta}} \ell\big(\boldsymbol{\theta}^{(t)}\big) \right\rangle + \left\langle \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^{(t)}, b^{(t)} \right\rangle,$$

where $b^{(t)}$ stands for the bias of the gradient estimate, namely

$$b^{(t)} = \mathbb{E}\left[\widehat{g}^{(t)} \;\middle|\; \boldsymbol{\theta}^{(t)}\right] - \nabla_{\boldsymbol{\theta}} \ell\big(\boldsymbol{\theta}^{(t)}\big).$$

By combining the Cauchy–Schwarz inequality with Lemma B.5, which applies since $\eta \leq (2\Gamma + L)^{-1}$, we get

$$\left\langle \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^{(t)}, b^{(t)} \right\rangle \leq \left\| \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^{(t)} \right\| \left\| b^{(t)} \right\| \leq \left\| b^{(t)} \right\|.$$

Since the function ℓ is L-smooth, we can use the inequality (21), namely

$$\left\langle \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^{(t)}, \nabla_{\boldsymbol{\theta}} \ell\big(\boldsymbol{\theta}^{(t)}\big) \right\rangle \leq \ell\left(\bar{\boldsymbol{\theta}}^{(t)}\right) - \ell\big(\boldsymbol{\theta}^{(t)}\big) + \frac{L}{2} \left\| \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^{(t)} \right\|^2.$$

In conclusion we obtain

$$\left\langle \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^{(t)}, \mathbb{E}\left[\widehat{g}^{(t)} \;\middle|\; \boldsymbol{\theta}^{(t)}\right] \right\rangle \leq \ell\left(\bar{\boldsymbol{\theta}}^{(t)}\right) - \ell\big(\boldsymbol{\theta}^{(t)}\big) + \frac{L}{2} \left\| \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^{(t)} \right\|^2 + \left\| b^{(t)} \right\|. \tag{25}$$

We aim at obtaining a bound not depending on the virtual iterates. As $\eta^{-1} \geq 2L > L$, Lemma B.4 applies, and the function

$$g : \boldsymbol{\theta} \mapsto \ell(\boldsymbol{\theta}) + \frac{1}{2\eta} \left\| \boldsymbol{\theta} - \boldsymbol{\theta}^{(t)} \right\|^2$$

is $(\eta^{-1} - L)$-strongly convex. By definition, this function achieves a minimum at $\bar{\boldsymbol{\theta}}^{(t)}$. Consequently its gradient is 0 at that point and the strong-convexity gives

$$\ell\big(\boldsymbol{\theta}^{(t)}\big) \geq \ell\left(\bar{\boldsymbol{\theta}}^{(t)}\right) + \frac{1}{2\eta} \left\| \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^{(t)} \right\|^2 + \frac{\eta^{-1} - L}{2} \left\| \boldsymbol{\theta}^{(t)} - \bar{\boldsymbol{\theta}}^{(t)} \right\|^2.$$

Using this last inequality, Equation (25) becomes

$$\begin{aligned}
\left\langle \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^{(t)}, \mathbb{E}\left[\widehat{g}^{(t)} \;\middle|\; \boldsymbol{\theta}^{(t)}\right] \right\rangle &\leq -\frac{1}{2\eta} \left\| \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^{(t)} \right\|^2 - \frac{\eta^{-1} - 2L}{2} \left\| \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^{(t)} \right\|^2 \\
&\quad + \left\| b^{(t)} \right\| \\
&\leq -\frac{1}{2\eta} \left\| \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^{(t)} \right\|^2 + \left\| b^{(t)} \right\|,
\end{aligned}$$

since $\eta^{-1} \geq 2L$. Finally, using Lemma B.2 leads to

$$\frac{1}{2\eta} \left\| \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^{(t)} \right\|^2 = \frac{1}{2\eta} \left\| \eta \nabla F_\eta\big(\boldsymbol{\theta}^{(t)}\big) \right\|^2 = \frac{\eta}{2} \left\| \nabla F_\eta\big(\boldsymbol{\theta}^{(t)}\big) \right\|^2.$$

So far, we have hence proven that

$$\mathbb{E}\left[F_\eta\big(\boldsymbol{\theta}^{(t+1)}\big) \,\middle|\, \boldsymbol{\theta}^{(t)}\right] \leq F_\eta\big(\boldsymbol{\theta}^{(t)}\big) - \frac{\gamma}{2}\big\|\nabla F_\eta\big(\boldsymbol{\theta}^{(t)}\big)\big\|^2 + \frac{\gamma}{\eta}\big\|b^{(t)}\big\| + \frac{\gamma^2}{2\eta}\mathbb{E}\left[\big\|\widehat{g}^{(t)}\big\|^2 \,\middle|\, \boldsymbol{\theta}^{(t)}\right].$$

We now focus on the third term. The bilinearity of the scalar product gives

$$\begin{aligned}\mathbb{E}\left[\big\|\widehat{g}^{(t)}\big\|^2 \,\middle|\, \boldsymbol{\theta}^{(t)}\right] &= \sigma^{(t)} - \big\|\nabla_{\boldsymbol{\theta}}\ell\big(\boldsymbol{\theta}^{(t)}\big)\big\|^2 + 2\left\langle \mathbb{E}\left[\widehat{g}^{(t)} \,\middle|\, \boldsymbol{\theta}^{(t)}\right], \nabla_{\boldsymbol{\theta}}\ell\big(\boldsymbol{\theta}^{(t)}\big)\right\rangle \\ &= \sigma^{(t)} + \big\|\nabla_{\boldsymbol{\theta}}\ell\big(\boldsymbol{\theta}^{(t)}\big)\big\|^2 + 2\big\langle b^{(t)}, \nabla_{\boldsymbol{\theta}}\ell\big(\boldsymbol{\theta}^{(t)}\big)\big\rangle.\end{aligned}$$

The Cauchy-Schwarz inequality along with the definition of the bound Γ yields

$$\mathbb{E}\left[\big\|\widehat{g}^{(t)}\big\|^2 \,\middle|\, \boldsymbol{\theta}^{(t)}\right] \leq \sigma^{(t)} + \Gamma^2 + 2\Gamma\big\|b^{(t)}\big\|.$$

It follows that

$$\mathbb{E}\left[F_\eta\big(\boldsymbol{\theta}^{(t+1)}\big) \,\middle|\, \boldsymbol{\theta}^{(t)}\right] \leq F_\eta\big(\boldsymbol{\theta}^{(t)}\big) - \frac{\gamma}{2}\big\|\nabla F_\eta\big(\boldsymbol{\theta}^{(t)}\big)\big\|^2 + \frac{\gamma+\gamma^2\Gamma}{\eta}\big\|b^{(t)}\big\| + \frac{\gamma^2}{2\eta}\big(\sigma^{(t)} + \Gamma^2\big).$$

□

Appendix C: Proof of Theorem 4.5

Proposition 4.4. *Let $\{i(t)\}_{t\in\mathbb{N}^*}$ and $\{\boldsymbol{\theta}^{(t)}\}_{t\in\mathbb{N}^*}$ be the associated random sequences generated by Algorithm 1. Under Model (2), if Assumption 1 holds, then, for all $t \in \mathbb{N}$,*

$$\overline{\sigma}_{\mathrm{IS}}^{(t)} = \mathbb{E}\left[\left\|\widehat{s}_{i(t)}^N\big(\boldsymbol{\theta}^{(t)}\big) - \nabla \log p_{\boldsymbol{\theta}^{(t)}}\left(\mathbf{Y}_{i(t)}\right)\right\|^2 \,\middle|\, \boldsymbol{\theta}^{(t)}\right] \leq \frac{d}{N}M_\sigma,$$

$$\overline{\xi}_{\mathrm{IS}}^{(t)} = \mathbb{E}\left[\left\|\mathbb{E}\left[\widehat{s}_{i(t)}^N\big(\boldsymbol{\theta}^{(t)}\big) - \nabla \log p_{\boldsymbol{\theta}^{(t)}}\left(\mathbf{Y}_{i(t)}\right) \,\middle|\, \boldsymbol{\theta}^{(t)}, i(t)\right]\right\| \,\middle|\, \boldsymbol{\theta}^{(t)}\right] \leq \frac{\sqrt{d}}{N}M_\xi,$$

where M_σ and M_ξ are two finite and positive constants given by

$$M_\sigma = \frac{12}{n}\sum_{i=1}^n \frac{\lambda_i^2\sqrt{\beta_i}}{\zeta_i^2}\left(1 + \frac{250\lambda_i}{\zeta_i} + \frac{9\lambda_i^2}{\zeta_i^2}\right),$$

$$M_\xi = \frac{4}{n}\sum_{i=1}^n \frac{\lambda_i^2\beta_i^{1/4}}{\zeta_i^2}\left\{2 + \sqrt{3\left(1 + \frac{250\lambda_i}{\zeta_i} + \frac{9\lambda_i^2}{\zeta_i^2}\right)}\right\},$$

with λ_i and β_i as in Assumption 1.

Proof. For any $t \in \mathbb{N}$, since the individual $i(t)$ is drawn uniformly independently of $\boldsymbol{\theta}^{(t)}$, we have

$$\begin{aligned}
\overline{\sigma}_{\mathrm{IS}}^{(t)} &= \mathbb{E}\left[\mathbb{E}\left[\left\|\widehat{s}_{i(t)}^{N} - \nabla \log p_{\boldsymbol{\theta}^{(t)}}\left(\mathbf{Y}_{i(t)}\right)\right\|^2 \,\middle|\, \boldsymbol{\theta}^{(t)}, i(t)\right]\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\left\|\widehat{s}_{i(t)}^{N} - \nabla \log p_{\boldsymbol{\theta}^{(t)}}\left(\mathbf{Y}_{i(t)}\right)\right\|^2 \,\middle|\, \boldsymbol{\theta}^{(t)}, i(t) = i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\nu_i^{\otimes N}(\cdot\,;\boldsymbol{\theta}^{(t)})}\left[\left\|\widehat{s}_{i}^{N} - \nabla \log p_{\boldsymbol{\theta}^{(t)}}\left(\mathbf{Y}_{i}\right)\right\|^2\right]
\end{aligned}$$

and, similarly

$$\begin{aligned}
\overline{\xi}_{\mathrm{IS}}^{(t)} &= \frac{1}{n}\sum_{i=1}^{n} \left\|\mathbb{E}\left[\widehat{s}_{i(t)}^{N} - \nabla \log p_{\boldsymbol{\theta}^{(t)}}\left(\mathbf{Y}_{i(t)}\right) \,\middle|\, \boldsymbol{\theta}^{(t)}, i(t) = i\right]\right\| \\
&= \frac{1}{n}\sum_{i=1}^{n} \left\|\mathbb{E}_{\nu_i^{\otimes N}(\cdot\,;\boldsymbol{\theta}^{(t)})}\left[\widehat{s}_{i}^{N} - \nabla \log p_{\boldsymbol{\theta}^{(t)}}\left(\mathbf{Y}_{i}\right)\right]\right\|
\end{aligned}$$

Consider an individual $i \in \{1, \ldots, n\}$ and a parameter $\boldsymbol{\theta} \in \Theta$. As in Agapiou et al. (2017), the r-th central moment, $r \in \mathbb{N}^*$, of a mesurable function $h : \mathbb{R}^q \to \mathbb{R}$ with respect to the proposal distribution is denoted

$$m_r[h] = \mathbb{E}_{\nu_i(\cdot\,;\boldsymbol{\theta})}\left[\left|h(\mathbf{V}) - \mathbb{E}_{\nu_i(\cdot\,;\boldsymbol{\theta})}\left[h(\mathbf{V})\right]\right|^r\right].$$

We also denote ϕ_k, $k \in \{1, \ldots, d\}$, the application that returns the k-th component of the score, namely for all $\mathbf{V} \in \mathbb{R}^q$

$$\phi_k(\mathbf{V}) = \frac{\partial}{\partial \theta_k} \log p_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{V}).$$

For a N-sample $(\mathbf{V}_1, \ldots, \mathbf{V}_N)$ from $\nu_i(\cdot\,;\boldsymbol{\theta})$, the importance sampling estimate $\widehat{s}_i^N(\boldsymbol{\theta}) = (\widehat{s}_{ik})_{1 \le k \le d}$ of the score g_i, as defined in Equation (5), can thereby be written as

$$\widehat{s}_{ik} = \frac{1}{\sum_{s=1}^{N} \rho_{\boldsymbol{\theta},i}(\mathbf{V}_s)} \sum_{r=1}^{N} \rho_{\boldsymbol{\theta},i}(\mathbf{V}_r)\phi_k(\mathbf{V}_r).$$

According to Theorem 2.3 in Agapiou et al. (2017), we can guarantee control over the bias and mean squared error of the importance sampling estimate for the k-th component of the score, $k \in \{1, \ldots, d\}$, provided that the following quantity is finite

$$\begin{aligned}
C_{\mathrm{MSE},k}(\boldsymbol{\theta}) = \frac{3}{p_{\boldsymbol{\theta}}(\mathbf{Y}_i)^2} \Bigg\{ & m_2[\phi_k \rho_{\boldsymbol{\theta},i}] \\
& + \frac{9}{p_{\boldsymbol{\theta}}(\mathbf{Y}_i)^2} \sqrt{m_4[\rho_{\boldsymbol{\theta},i}] \mathbb{E}_{\nu_i(\cdot\,;\boldsymbol{\theta})}\left[\left|\phi_k(\mathbf{V})\rho_{\boldsymbol{\theta},i}(\mathbf{V})\right|^4\right]} \\
& + \frac{125}{p_{\boldsymbol{\theta}}(\mathbf{Y}_i)} \sqrt{m_6[\rho_{\boldsymbol{\theta},i}] \mathbb{E}_{\nu_i(\cdot\,;\boldsymbol{\theta})}\left[\left|\phi_k(\mathbf{V})\right|^4\right]} \Bigg\}.
\end{aligned}$$

To prove such a statement, first note that by definition

$$0 < \lambda_i, \qquad \mathbb{E}_{\nu_i(\cdot\,;\boldsymbol{\theta})}\left[|\phi_k(\mathbf{V})|^4\right] \leq \beta_i. \tag{26}$$

The monotonicity of the expected value thus yields

$$\mathbb{E}_{\nu_i(\cdot\,;\boldsymbol{\theta})}\left[|\phi_k(\mathbf{V})\rho_{\boldsymbol{\theta},i}(\mathbf{V})|^4\right] \leq \lambda_i^4 \mathbb{E}_{\nu_i(\cdot\,;\boldsymbol{\theta})}\left[|\phi_k(\mathbf{V})|^4\right] \leq \lambda_i^4 \beta_i.$$

Furthermore, we use that for all $r \in \mathbb{N}^*$, $\mathbb{E}[|X - \mathbb{E}[X]|^r] \leq 2^r \mathbb{E}[|X|^r]$ (this is a direct consequence of the Minkowski and Jensen inequalities). Then, we have for all $r \in \mathbb{N}^*$

$$\begin{aligned} m_r[\rho_{\boldsymbol{\theta},i}] &\leq 2^r \mathbb{E}_{\nu_i(\cdot\,;\boldsymbol{\theta})}\left[\rho_{\boldsymbol{\theta},i}(\mathbf{V})^r\right] \leq 2^r \lambda_i^r, \\ m_2[\phi_k \rho_{\boldsymbol{\theta},i}] &\leq 4\mathbb{E}_{\nu_i(\cdot\,;\boldsymbol{\theta})}\left[|\phi_k(\mathbf{V})\rho_{\boldsymbol{\theta},i}(\mathbf{V})|^2\right] \leq 4\lambda_i^2 \mathbb{E}_{\nu_i(\cdot\,;\boldsymbol{\theta})}\left[|\phi_k(\mathbf{V})|^2\right]. \end{aligned} \tag{27}$$

Finally, the Jensen inequality for the square root function provides that

$$\mathbb{E}_{\nu_i(\cdot\,;\boldsymbol{\theta})}\left[|\phi_k(\mathbf{V})|^2\right] \leq \sqrt{\mathbb{E}_{\nu_i(\cdot\,;\boldsymbol{\theta})}\left[|\phi_k(\mathbf{V})|^4\right]} \leq \sqrt{\beta_i} \tag{28}$$

Combining the upper-bounds from Equations (26)–(28) leads to

$$C_{\mathrm{MSE},k}(\boldsymbol{\theta}) \leq \frac{12\lambda_i^2\sqrt{\beta_i}}{p_{\boldsymbol{\theta}}(\mathbf{Y}_i)^2}\left\{1 + \frac{9\lambda_i^2}{p_{\boldsymbol{\theta}}(\mathbf{Y}_i)^2} + \frac{250\lambda_i}{p_{\boldsymbol{\theta}}(\mathbf{Y}_i)}\right\}.$$

Since λ_i and β_i are all finite constants, and for any $\boldsymbol{\theta} \in \mathbb{R}^d$ and any $\mathbf{Y}_i \in \mathbb{N}$, $p_{\boldsymbol{\theta}}(\mathbf{Y}_i)$ is positive, it follows that $C_{\mathrm{MSE},k}(\boldsymbol{\theta})$ is finite. Consequently, Theorem 2.3 in Agapiou et al. (2017) states that

$$\begin{aligned} \mathbb{E}_{\nu_i^{\otimes N}(\cdot\,;\boldsymbol{\theta})}\left[\left\{\widehat{s}_{ik} - \frac{\partial}{\partial\theta_k}\log p_{\boldsymbol{\theta}}(\mathbf{Y}_i)\right\}^2\right] &\leq \frac{1}{N} C_{\mathrm{MSE},k}(\boldsymbol{\theta}), \\ \left|\mathbb{E}_{\nu_i^{\otimes N}(\cdot\,;\boldsymbol{\theta})}\left[\widehat{s}_{ik}\right] - \frac{\partial}{\partial\theta_k}\log p_{\boldsymbol{\theta}}(\mathbf{Y}_i)\right| &\leq \frac{2}{N p_{\boldsymbol{\theta}}(\mathbf{Y}_i)}\left\{\frac{1}{p_{\boldsymbol{\theta}}(\mathbf{Y}_i)}\sqrt{m_2[\rho_{\boldsymbol{\theta},i}] m_2[\overline{\phi_k}\rho_{\boldsymbol{\theta},i}]}\right. \\ &\qquad \left. + \sqrt{C_{\mathrm{MSE},k}(\boldsymbol{\theta})\mathbb{E}_{\nu_i(\cdot\,;\boldsymbol{\theta})}\left[\rho_{\boldsymbol{\theta},i}(\mathbf{V})^2\right]}\right\}, \end{aligned}$$

where $\overline{\phi_k} : \mathbf{V} \mapsto \phi_k(\mathbf{V}) - \mathbb{E}_{\nu_i(\cdot\,;\boldsymbol{\theta})}\left[\phi_k(\mathbf{U})\right]$. To conclude, we further need to eliminate the dependence in $\boldsymbol{\theta}$ in these upper bounds. For any $\mathbf{Y}_i \in \mathbb{N}^p$, the function $\boldsymbol{\theta} \mapsto p_{\boldsymbol{\theta}}(\mathbf{Y}_i)$ is bounded below by $\zeta_i > 0$ on Θ (Proposition 4.3). Moreover, combining the argument of Equation (27) with the positivity of the variance and Equation (28), we have

$$m_2[\overline{\phi_k}\rho_{\boldsymbol{\theta},i}] \leq 4\lambda_i^2 \mathbb{E}_{\nu_i(\cdot\,;\boldsymbol{\theta})}\left[\left|\overline{\phi_k}(\mathbf{V})\right|^2\right] \leq 4\lambda_i^2 \mathbb{E}_{\nu_i(\cdot\,;\boldsymbol{\theta})}\left[|\phi_k(\mathbf{V})|^2\right] \leq 4\lambda_i^2\sqrt{\beta_i}.$$

Overall, these last two bounds in addition with Equations (26)–(28) yield the conclusion, namely

$$\begin{aligned}\overline{\sigma}_{\mathrm{IS}}^{(t)} &= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{d}\mathbb{E}_{\nu_i^{\otimes N}(\cdot\,;\boldsymbol{\theta})}\left[\left\{\widehat{s}_{ik}-\frac{\partial}{\partial\theta_k}\log p_{\boldsymbol{\theta}}(\mathbf{Y}_i)\right\}^2\right]\\ &\leq \frac{12d}{nN}\sum_{i=1}^{n}\frac{\lambda_i^2\sqrt{\beta}_i}{\zeta_i^2}\left(1+\frac{250\lambda_i}{\zeta_i}+\frac{9\lambda_i^2}{\zeta_i^2}\right),\end{aligned}$$

and

$$\begin{aligned}\overline{\xi}_{\mathrm{IS}}^{(t)} &= \frac{1}{n}\sum_{i=1}^{n}\sqrt{\sum_{k=1}^{d}\left|\mathbb{E}_{\nu_i^{\otimes N}(\cdot\,;\boldsymbol{\theta})}\left[\widehat{s}_{ik}\right]-\frac{\partial}{\partial\theta_k}\log p_{\boldsymbol{\theta}}(\mathbf{Y}_i)\right|^2}\\ &\leq \frac{1}{n}\sum_{i=1}^{n}\frac{2\sqrt{d}}{N\zeta_i}\left\{\frac{1}{\zeta_i}4\lambda_i^2\beta_i^{1/4}+\sqrt{\frac{12\lambda_i^4\sqrt{\beta}_i}{\zeta_i^2}\left(1+\frac{250\lambda_i}{\zeta_i}+\frac{9\lambda_i^2}{\zeta_i^2}\right)}\right\}\\ &\leq \frac{4\sqrt{d}}{nN}\sum_{i=1}^{n}\frac{\lambda_i^2\beta_i^{1/4}}{\zeta_i^2}\left\{2+\sqrt{3\left(1+\frac{250\lambda_i}{\zeta_i}+\frac{9\lambda_i^2}{\zeta_i^2}\right)}\right\}.\end{aligned}$$

□

Theorem 4.5. *Let $\boldsymbol{\theta}^{(1)}\in\Theta$ be an initial value and $\gamma_0\in\mathbb{R}_+^*$ a user-specified initial learning rate. Under Model (2), if Assumption 1 holds, then for any $T\in\mathbb{N}^*$ and any real constant $\eta\in(0,1/\max\{2\Gamma,L\}]$, the sequence $\{\boldsymbol{\theta}^{(t)}\}_{1\leq t\leq T}$ defined by Algorithm 1 with $\gamma=\gamma_0/\sqrt{T}$ satisfies*

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|G_\eta^{(t)}\right\|^2\right] \leq \frac{2\tau}{\gamma_0(L\eta+1)\sqrt{T}}\Bigg(\ell\big(\boldsymbol{\theta}^{(1)}\big)-\ell(\boldsymbol{\theta}^\star)+\frac{\gamma_0^2(L\eta+1)}{2\eta}\Bigg[\Delta^2\\ +\Gamma^2+\frac{d}{N}\left\{M_\sigma+\frac{2(\Delta+\Gamma)}{\sqrt{d}}M_\xi\right\}\Bigg]\Bigg)+\frac{2\tau\sqrt{d}}{\eta N}M_\xi,\end{aligned} \tag{13}$$

with L the smoothness constant of $\ell(\cdot)$, $\Gamma=\sup_{\boldsymbol{\theta}\in\Theta}\|\nabla\ell(\boldsymbol{\theta})\|$, constants M_σ and M_ξ as defined in Proposition 4.4, and

$$\tau=\frac{(2L\eta+1)^2}{L\eta+1}\left(1+\sqrt{\frac{L\eta}{L\eta+1}}\right)^2,\quad \Delta=\max_{i=1,\ldots,n}\sup_{\boldsymbol{\theta}\in\Theta}\left\|\nabla_{\boldsymbol{\theta}}\ell\left(\boldsymbol{\theta}\right)+\nabla_{\boldsymbol{\theta}}\log p_{\boldsymbol{\theta}}\left(\mathbf{Y}_i\right)\right\|.$$

Proof. According to Proposition 4.3, the function ℓ is L-smooth. Using successively that the square function and the expectation are increasing on $\mathbb{R}_+$, for any real constant $\eta>0$, Lemma B.3 yields

$$\mathbb{E}\left[\left\|G_\eta^{(t)}\right\|^2\right]\leq\left(1+\frac{L\eta}{L\eta+1}\right)^2\left(1+\sqrt{\frac{L\eta}{L\eta+1}}\right)^2\mathbb{E}\left[\left\|\nabla F_{\frac{\eta}{L\eta+1}}\left(\boldsymbol{\theta}^{(t)}\right)\right\|^2\right]. \tag{29}$$

To elaborate on the upper bound, we aim at using Lemma 4.2. Therefore, we first need to prove that we have control over the bias $\xi^{(t)}$ and the mean squared error $\sigma^{(t)}$ of the gradient estimates, as defined in Lemma 4.2.

Let $t \in \mathbb{N}^*$. Due to the bilinearity of the scalar product, the following identity holds:

$$\begin{aligned}\sigma^{(t)} = \mathbb{E}\left[\left|\left|\widehat{g}^{(t)} + \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}^{(t)}}\left(\mathbf{Y}_{i(t)}\right)\right|\right|^2 + \left|\left|\nabla_{\boldsymbol{\theta}}\ell\left(\boldsymbol{\theta}^{(t)}\right) + \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}^{(t)}}\left(\mathbf{Y}_{i(t)}\right)\right|\right|^2 \,\middle|\, \boldsymbol{\theta}^{(t)}\right] \\ - 2\mathbb{E}\left[\left\langle \widehat{g}^{(t)} + \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}^{(t)}}\left(\mathbf{Y}_{i(t)}\right), \nabla_{\boldsymbol{\theta}}\ell\left(\boldsymbol{\theta}^{(t)}\right) + \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}^{(t)}}\left(\mathbf{Y}_{i(t)}\right)\right\rangle \,\middle|\, \boldsymbol{\theta}^{(t)}\right]\end{aligned}$$

Note that the first term of the right hand side corresponds to $\overline{\sigma}^{(t)}_{\mathrm{IS}}$. Let bound the two remaining terms. According to Proposition 4.1, for all $i = 1, \ldots, n$, the function

$$\boldsymbol{\theta} \mapsto \nabla_{\boldsymbol{\theta}}\ell\left(\boldsymbol{\theta}\right) + \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}\left(\mathbf{Y}_i\right)$$

is continuous on Θ, and thereby bounded. Denote

$$\Delta = \max_{i=1,\ldots,n} \sup_{\boldsymbol{\theta}\in\Theta} \left\|\nabla_{\boldsymbol{\theta}}\ell\left(\boldsymbol{\theta}\right) + \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}\left(\mathbf{Y}_i\right)\right\|.$$

We have

$$\mathbb{E}\left[\left|\left|\nabla_{\boldsymbol{\theta}}\ell\left(\boldsymbol{\theta}^{(t)}\right) + \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}^{(t)}}\left(\mathbf{Y}_{i(t)}\right)\right|\right|^2 \,\middle|\, \boldsymbol{\theta}^{(t)}\right] \leq \Delta^2$$

Moreover, $i(t)$ being a uniform random variable on $\{1, \ldots, n\}$ independent of the sequence $\boldsymbol{\theta}^{(t)}$,

$$\begin{aligned}&\mathbb{E}\left[\left\langle \widehat{g}^{(t)} + \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}^{(t)}}\left(\mathbf{Y}_{i(t)}\right), \nabla_{\boldsymbol{\theta}}\ell\left(\boldsymbol{\theta}^{(t)}\right) + \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}^{(t)}}\left(\mathbf{Y}_{i(t)}\right)\right\rangle \,\middle|\, \boldsymbol{\theta}^{(t)}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\left\langle -\widehat{s}^N_{i(t)}\left(\boldsymbol{\theta}^{(t)}\right) + \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}^{(t)}}\left(\mathbf{Y}_{i(t)}\right), \nabla_{\boldsymbol{\theta}}\ell\left(\boldsymbol{\theta}^{(t)}\right) + \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}^{(t)}}\left(\mathbf{Y}_{i(t)}\right)\right\rangle \,\middle|\, \boldsymbol{\theta}^{(t)}, i(t)\right]\right] \\ &= \mathbb{E}\left[\left\langle -\mathbb{E}\left[\widehat{s}^N_{i(t)}\left(\boldsymbol{\theta}^{(t)}\right) \,\middle|\, \boldsymbol{\theta}^{(t)}, i(t)\right] + \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}^{(t)}}\left(\mathbf{Y}_{i(t)}\right), \nabla_{\boldsymbol{\theta}}\ell\left(\boldsymbol{\theta}^{(t)}\right) + \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}^{(t)}}\left(\mathbf{Y}_{i(t)}\right)\right\rangle\right] \\ &= -\frac{1}{n}\sum_{i=1}^{n}\left\langle \mathbb{E}_{\nu_i^{\otimes N}(\cdot\,;\boldsymbol{\theta}^{(t)})}\left[\widehat{s}^N_i\left(\boldsymbol{\theta}^{(t)}\right)\right] - \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}^{(t)}}\left(\mathbf{Y}_i\right), \nabla_{\boldsymbol{\theta}}\ell\left(\boldsymbol{\theta}^{(t)}\right) + \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}^{(t)}}\left(\mathbf{Y}_i\right)\right\rangle.\end{aligned}$$

Providing Assumption 1 holds, the Cauchy-Schwarz inequality along with Proposition 4.4 yields

$$\begin{aligned}&-2\mathbb{E}\left[\left\langle \widehat{g}^{(t)} + \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}^{(t)}}\left(\mathbf{Y}_{i(t)}\right), \nabla_{\boldsymbol{\theta}}\ell\left(\boldsymbol{\theta}^{(t)}\right) + \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}^{(t)}}\left(\mathbf{Y}_{i(t)}\right)\right\rangle \,\middle|\, \boldsymbol{\theta}^{(t)}\right] \\ &\leq \frac{2}{n}\sum_{i=1}^{n}\left|\left|\mathbb{E}_{\nu_i^{\otimes N}(\cdot\,;\boldsymbol{\theta}^{(t)})}\left[\widehat{s}^N_i\left(\boldsymbol{\theta}^{(t)}\right)\right] - \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}^{(t)}}\left(\mathbf{Y}_i\right)\right|\right|\left|\left|\nabla_{\boldsymbol{\theta}}\ell\left(\boldsymbol{\theta}^{(t)}\right) + \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}^{(t)}}\left(\mathbf{Y}_i\right)\right|\right| \\ &\leq \frac{2\Delta\sqrt{d}}{N} M_\xi.\end{aligned}$$

In conclusion, for any $t \in \mathbb{N}^*$, we have a finite upper bound for the mean squared error $\sigma^{(t)}$, namely

$$\sigma^{(t)} \leq \frac{d}{N} M_\sigma + \Delta^2 + \frac{2\Delta\sqrt{d}}{N} M_\xi = \Delta^2 + \frac{d}{N}\left(M_\sigma + \frac{2\Delta}{\sqrt{d}} M_\xi\right).$$

Conversely, leveraging Assumption 1 once again, Proposition 4.4 shows that we have a finite upper bound for the bias $\xi^{(t)}$. Indeed

$$\begin{aligned}\xi^{(t)} &= \left\| \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[-\widehat{s}^{N}_{i(t)} \;\middle|\; \boldsymbol{\theta}^{(t)}, i(t)=i \right] + \frac{1}{n}\sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}^{(t)}}(\mathbf{Y}_i) \right\| \\ &= \left\| \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[-\widehat{s}^{N}_{i(t)} + \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}^{(t)}}(\mathbf{Y}_{i(t)}) \;\middle|\; \boldsymbol{\theta}^{(t)}, i(t)=i \right] \right\| \\ &\leq \frac{1}{n}\sum_{i=1}^{n} \left\| \mathbb{E}\left[-\widehat{s}^{N}_{i(t)} + \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}^{(t)}}(\mathbf{Y}_{i(t)}) \;\middle|\; \boldsymbol{\theta}^{(t)}, i(t)=i \right] \right\| = \overline{\xi}^{(t)}_{\mathrm{IS}} \leq \frac{\sqrt{d}}{N} M_{\xi}.\end{aligned}$$

Consequently, Lemma 4.2 applies for any real constant $\eta \in (0, 1/\max\{2\Gamma, L\}]$, as

$$\frac{\eta}{L\eta+1} \leq \frac{1}{L+\max\{2\Gamma, L\}} = \frac{1}{\max\{2\Gamma+L, 2L\}}.$$

After integrating both sides of Equation (11) with respect to $\boldsymbol{\theta}^{(t)}$ and plugging in the aforementionned upper bounds, we get

$$\begin{aligned}\frac{\gamma}{2}\mathbb{E}\left[\left\| \nabla F_{\frac{\eta}{L\eta+1}}\left(\boldsymbol{\theta}^{(t)}\right) \right\|^2 \right] \leq{}& \mathbb{E}\left[F_{\frac{\eta}{L\eta+1}}\left(\boldsymbol{\theta}^{(t)}\right) \right] - \mathbb{E}\left[F_{\frac{\eta}{L\eta+1}}\left(\boldsymbol{\theta}^{(t+1)}\right) \right] \\ &+ \frac{(\gamma+\gamma^2\Gamma)(L\eta+1)\sqrt{d}}{\eta N} M_{\xi} \\ &+ \frac{\gamma^2(L\eta+1)}{2\eta}\left\{ \Delta^2 + \frac{d}{N}\left(M_{\sigma} + \frac{2\Delta}{\sqrt{d}} M_{\xi} \right) + \Gamma^2 \right\}.\end{aligned}$$

Then, summing along the iterations $t = 1, \ldots, T$ yields

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}\left[\left\| \nabla F_{\frac{\eta}{L\eta+1}}\left(\boldsymbol{\theta}^{(t)}\right) \right\|^2 \right] \leq{}& \frac{2}{\gamma}\left\{ F_{\frac{\eta}{L\eta+1}}\left(\boldsymbol{\theta}^{(1)}\right) - \mathbb{E}\left[F_{\frac{\eta}{L\eta+1}}\left(\boldsymbol{\theta}^{(T+1)}\right) \right] \right\} \\ &+ \frac{2(L\eta+1)T\sqrt{d}}{\eta N} M_{\xi} + \frac{\gamma(L\eta+1)T}{\eta}\left[\Delta^2 \right. \\ &\left. + \frac{d}{N}\left\{ M_{\sigma} + \frac{2(\Delta+\Gamma)}{\sqrt{d}} M_{\xi} \right\} + \Gamma^2 \right].\end{aligned}$$

For $\boldsymbol{\theta}^{\star} = \arg\min_{\boldsymbol{\theta}\in\Theta} \ell(\boldsymbol{\theta})$, according to the definition of the Moreau envelope, for any $\boldsymbol{\theta} \in \Theta$

$$\ell(\boldsymbol{\theta}^{\star}) \leq F_{\eta/(L\eta+1)}(\boldsymbol{\theta}),$$

and thus,

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\left\| \nabla F_{\frac{\eta}{L\eta+1}}\left(\boldsymbol{\theta}^{(t)}\right) \right\|^2 \right] \leq{}& \frac{2}{\gamma T}\left\{ F_{\frac{\eta}{L\eta+1}}\left(\boldsymbol{\theta}^{(1)}\right) - \ell(\boldsymbol{\theta}^{\star}) \right\} + \frac{2(L\eta+1)\sqrt{d}}{\eta N} M_{\xi} \\ &+ \frac{\gamma(L\eta+1)}{\eta}\left[\Delta^2 + \frac{d}{N}\left\{ M_{\sigma} + \frac{2(\Delta+\Gamma)}{\sqrt{d}} M_{\xi} \right\} \right. \\ &\left. + \Gamma^2 \right].\end{aligned}$$

Combining the last inequality with Equation (29) and using $F_{\eta/(L\eta+1)}(\boldsymbol{\theta}^{(1)}) \leq \ell(\boldsymbol{\theta}^{(1)})$ since $\boldsymbol{\theta}^{(1)} \in \Theta$, we get for $\gamma = \gamma_0/\sqrt{T}$

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|G_\eta^{(t)}\right\|^2\right] \leq \frac{2\tau}{\gamma_0(L\eta+1)\sqrt{T}}\left\{F_{\frac{\eta}{L\eta+1}}\left(\boldsymbol{\theta}^{(1)}\right) - \ell(\boldsymbol{\theta}^\star)\right\} + \frac{2\tau\sqrt{d}}{\eta N}M_\xi + \frac{\gamma_0\tau}{\eta\sqrt{T}}\left[\Delta^2 + \frac{d}{N}\left\{M_\sigma + \frac{2(\Delta+\Gamma)}{\sqrt{d}}M_\xi\right\} + \Gamma^2\right].$$

□

Appendix D: Results on the importance sampling proposal distribution

Lemma 5.1. *Let $\alpha \in (0,1]$ and $\delta > 1$. If for any $i = 1,\ldots,n$, $\boldsymbol{\theta} \mapsto \boldsymbol{\mu}_i(\boldsymbol{\theta}) \in \mathbb{R}^q$ and $\boldsymbol{\theta} \mapsto \mathbf{S}_i(\boldsymbol{\theta}) \in \mathcal{S}_{++}^q$ are continuous on Θ, then the proposal distribution defined by*

$$\nu_i(\cdot\,;\boldsymbol{\theta}) = \mathcal{GM}(\cdot;\boldsymbol{\mu}_i(\boldsymbol{\theta}),\mathbf{S}_i(\boldsymbol{\theta}),\alpha,\delta) \tag{15}$$

fulfils Assumption 1.

Proof. Let $i \in \{1,\ldots,n\}$, $\alpha \in (0,1]$ and $\delta > 1$.

(A1.1) Each component of a mixture distribution being a non-negative function, we have for any $\mathbf{V} \in \mathbb{R}^q$ and any $\boldsymbol{\theta} \in \mathbb{R}^d$,

$$\rho_{\boldsymbol{\theta},i}(\mathbf{V}) = \frac{p_{\boldsymbol{\theta}}(\mathbf{Y}_i,\mathbf{V})}{\mathcal{GM}(\mathbf{V};\boldsymbol{\mu}_i(\boldsymbol{\theta}),\mathbf{S}_i(\boldsymbol{\theta}),\alpha,\delta)} \leq \frac{1}{\alpha}\frac{p_{\boldsymbol{\theta}}(\mathbf{Y}_i,\mathbf{V})}{\mathcal{N}(\mathbf{V};\boldsymbol{\mu}_i(\boldsymbol{\theta}),\delta\mathbf{I}_q)}.$$

On the compact set Θ, there exist real constants $K_i^\Theta > 0$ and κ_i^Θ (Lemma A.1) such that

$$\log\frac{p_{\boldsymbol{\theta}}(\mathbf{Y}_i,\mathbf{V})}{\mathcal{N}(\mathbf{V};\boldsymbol{\mu}_i(\boldsymbol{\theta}),\delta\mathbf{I}_q)} \leq K_i^\Theta\|\mathbf{V}\| + \kappa_i^\Theta - \frac{1}{2}\|\mathbf{V}\|^2 + \frac{1}{2\delta}\|\mathbf{V}-\boldsymbol{\mu}_i(\boldsymbol{\theta})\|^2 + \frac{q}{2}\log(\delta).$$

Using the Cauchy-Schwarz inequality, we get

$$\|\mathbf{V}-\boldsymbol{\mu}_i(\boldsymbol{\theta})\|^2 = \|\mathbf{V}\|^2 + \|\boldsymbol{\mu}_i(\boldsymbol{\theta})\|^2 - 2\langle\boldsymbol{\mu}_i(\boldsymbol{\theta}),\mathbf{V}\rangle \leq \|\mathbf{V}\|^2 + \|\boldsymbol{\mu}_i(\boldsymbol{\theta})\|^2 + 2\|\mathbf{V}\|\|\boldsymbol{\mu}_i(\boldsymbol{\theta})\|.$$

$\boldsymbol{\theta} \mapsto \boldsymbol{\mu}_i(\boldsymbol{\theta})$ is continuous on the compact set Θ, and hence bounded, say by $u_\star$. It follows that

$$\|\mathbf{V}-\boldsymbol{\mu}_i(\boldsymbol{\theta})\|^2 \leq \|\mathbf{V}\|^2 + u_\star^2 + 2u_\star\|\mathbf{V}\|.$$

Consequently, we can derive an upper bound independent of $\boldsymbol{\theta}$, namely

$$\frac{p_{\boldsymbol{\theta}}(\mathbf{Y}_i,\mathbf{V})}{\mathcal{N}(\mathbf{V};\boldsymbol{\mu}_i(\boldsymbol{\theta}),\delta\mathbf{I}_q)} \leq \exp\left[\left(K_i^\Theta + \frac{u_\star}{\delta}\right)\|\mathbf{V}\| - \frac{\delta-1}{2\delta}\|\mathbf{V}\|^2 + \kappa_i^\Theta + \frac{u_\star^2}{2\delta} + \frac{q}{2}\log(\delta)\right].$$

Providing $\delta > 1$, the quadratic term in $\mathbf{V}$ overweight any other terms and the supremum in $\mathbf{V}$ is finite:

$$\sup_{(\boldsymbol{\theta},\mathbf{V})\in\Theta\times\mathbb{R}^q}\rho_{\boldsymbol{\theta},i}(\mathbf{V}) \leq \frac{1}{\alpha}\sup_{(\boldsymbol{\theta},\mathbf{V})\in\Theta\times\mathbb{R}^q}\frac{p_{\boldsymbol{\theta}}(\mathbf{Y}_i,\mathbf{V})}{\mathcal{N}(\mathbf{V};\boldsymbol{\mu}_i(\boldsymbol{\theta}),\delta\mathbf{I}_q)} < \infty.$$

(A1.2) Following Equation (19), there is a unique j in $\{1,\dots,p\}$ such that

$$\left|\frac{\partial}{\partial\theta_r}\log p_{\boldsymbol{\theta}}(\mathbf{Y}_i,\mathbf{W}_i)\right| \leq (\|\mathbf{x}_i\|+\|\mathbf{W}_i\|)\left\{\|\mathbf{Y}_i\|+\exp(Z_{ij})\right\}.$$

We then have,

$$\begin{aligned}\left|\frac{\partial}{\partial\theta_r}\log p_{\boldsymbol{\theta}}(\mathbf{Y}_i,\mathbf{W}_i)\right|^4 &\leq \left(8\|\mathbf{x}_i\|^4+8\|\mathbf{W}_i\|^4\right)\left\{8\|\mathbf{Y}_i\|^4+8\exp(4Z_{ij})\right\}.\\ &\leq 64\|\mathbf{Y}_i\|^4\left(\|\mathbf{x}_i\|^4+\|\mathbf{W}_i\|^4\right)\\ &\quad+64\exp(4\mathbf{B}_j^\top\mathbf{x}_i+4o_{ij})\left\{\|\mathbf{x}_i\|^4+\|\mathbf{W}_i\|^4\right\}\exp\left(4\mathbf{C}_j^\top\mathbf{W}_i\right),\end{aligned}$$

where $\mathbf{B}_j=(B_{1j},\dots,B_{pj})^\top$ and $\mathbf{C}_j=(C_{j1},\dots,C_{jq})^\top$ stand for the j-th column and row of $\mathbf{B}$ and $\mathbf{C}$, respectively.

Let show that the upper bound admits a finite expectation with respect to a multivariate Gaussian distribution $\mathcal{N}(\mu,S)$, for any $\mu\in\mathbb{R}^q$ and $S\in\mathbb{S}_{++}^q$. To demonstrate this, we use the following identities. For any $\mathbf{w}\in\mathbb{R}^q$, a straightforward rewriting yields

$$\mathcal{N}(\mathbf{w};\mu,S)\exp\left(4\mathbf{C}_j^\top\mathbf{w}\right)=\mathcal{N}(\mathbf{w};\mu+4S\mathbf{C}_j,S)\exp\left(4\mathbf{C}_j^\top\mu+8\mathbf{C}_j^\top S\mathbf{C}_j\right). \quad (30)$$

Moreover, we have (see for instance Simon (2002))

$$\mathbb{E}_{\mathcal{N}(\mathbf{0}_q,\mathbf{I}_q)}\left[\|\mathbf{W}_i\|^4\right]\leq q(q+2).$$

Using that

$$\mathbb{E}_{\mathcal{N}(\mu,S)}\left[\|\mathbf{W}_i\|^4\right]=\mathbb{E}_{\mathcal{N}(\mathbf{0}_q,\mathbf{I}_q)}\left[\left\|\mu+S^{1/2}\mathbf{W}_i\right\|^4\right]\leq 8\|\mu\|^4+8\left\|S^{1/2}\right\|^4\mathbb{E}_{\mathcal{N}(\mathbf{0}_q,\mathbf{I}_q)}\left[\|\mathbf{W}_i\|^4\right],$$

we get

$$\mathbb{E}_{\mathcal{N}(\mu,S)}\left[\|\mathbf{W}_i\|^4\right]\leq 8\|\mu\|^4+8q(q+2)\left\|S^{1/2}\right\|^4 \quad (31)$$

Equations (30) and (31) thus lead to

$$\begin{aligned}\Phi_j(\boldsymbol{\theta},\mu,S)&=\exp\left(4\mathbf{C}_j^\top\mu+8\mathbf{C}_j^\top S\mathbf{C}_j\right)\\ &=\mathbb{E}_{\mathcal{N}(\mu,S)}\left[\exp\left(4\mathbf{C}_j^\top\mathbf{W}_i\right)\right],\\ \Psi_j(\boldsymbol{\theta},\mu,S)&=\exp\left(4\mathbf{C}_j^\top\mu+8\mathbf{C}_j^\top S\mathbf{C}_j\right)\left\{8\|\mu+4S\mathbf{C}_j\|^4+8q(q+2)\left\|S^{1/2}\right\|^4\right\}\\ &\geq\mathbb{E}_{\mathcal{N}(\mu,S)}\left[\|\mathbf{W}_i\|^4\exp\left(4\mathbf{C}_j^\top\mathbf{W}_i\right)\right].\end{aligned}$$

Therefore, for any $\mu\in\mathbb{R}^q$ and $S\in\mathbb{S}_{++}^q$

$$\begin{aligned}\mathbb{E}_{\mathcal{N}(\mu,S)}\left[\left|\frac{\partial}{\partial\theta_r}\log p_{\boldsymbol{\theta}}(\mathbf{Y}_i,\mathbf{W}_i)\right|^4\right]&\leq 64\|\mathbf{Y}_i\|^4\left\{\|\mathbf{x}_i\|^4+8\|\mu\|^4+8q(q+2)\left\|S^{1/2}\right\|^4\right\}\\ &\quad+64\exp(4\mathbf{B}_j^\top\mathbf{x}_i+4o_{ij})\|\mathbf{x}_i\|^4\Phi_j(\boldsymbol{\theta},\mu,S)\\ &\quad+64\exp(4\mathbf{B}_j^\top\mathbf{x}_i+4o_{ij})\Psi_j(\boldsymbol{\theta},\mu,S).\end{aligned}$$

If $\boldsymbol{\theta} \mapsto \boldsymbol{\mu}_i(\boldsymbol{\theta})$ and $\boldsymbol{\theta} \mapsto \mathbf{S}_i(\boldsymbol{\theta})$ are continuous on Θ, the functions

$$\begin{aligned}
&\boldsymbol{\theta} \mapsto \mathbf{S}_i(\boldsymbol{\theta})^{1/2},\\
&\boldsymbol{\theta} \mapsto (1-\alpha)\Phi_j(\boldsymbol{\theta}, \boldsymbol{\mu}_i(\boldsymbol{\theta}), \mathbf{S}_i(\boldsymbol{\theta})) + \alpha\Phi_j(\boldsymbol{\theta}, \boldsymbol{\mu}_i(\boldsymbol{\theta}), \delta\mathbf{I}_q),\\
&\boldsymbol{\theta} \mapsto (1-\alpha)\Psi_j(\boldsymbol{\theta}, \boldsymbol{\mu}_i(\boldsymbol{\theta}), \mathbf{S}_i(\boldsymbol{\theta})) + \alpha\Psi_j(\boldsymbol{\theta}, \boldsymbol{\mu}_i(\boldsymbol{\theta}), \delta\mathbf{I}_q),
\end{aligned}$$

are also continuous on the compact set Θ (by a composition argument), and therefore bounded on Θ, say by $c_\star$, $\phi_\star$ and $\psi_\star$ respectively. On the other hand,

$$\exp(4\mathbf{B}_j^\top \mathbf{x}_i + 4o_{ij}) \leq \exp(4\|\mathbf{x}_i\|\|\mathbf{B}_j\| + 4\|\mathbf{o}_i\|) \leq \exp(4\|\mathbf{x}_i\| \sup_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{\theta}\| + 4\|\mathbf{o}_i\|).$$

Consequently, with $u_\star$ the bound of $\boldsymbol{\theta} \mapsto \boldsymbol{\mu}_i(\boldsymbol{\theta})$ on Θ, for any $r \in \{1, \ldots, d\}$,

$$\begin{aligned}
\mathbb{E}_{\nu_i(\cdot\,;\boldsymbol{\theta})}\left[\left|\frac{\partial}{\partial\theta_r} \log p_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{W}_i)\right|^4\right] &\leq 64\|\mathbf{Y}_i\|^4 \left[\|\mathbf{x}_i\|^4 + 8u_\star + 8q(q+2)\left\{(1-\alpha)c_\star^4 + \alpha\delta^4\right\}\right]\\
&\quad + 64\exp(4\|\mathbf{x}_i\| \sup_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{\theta}\| + 4\|\mathbf{o}_i\|)\left(\|\mathbf{x}_i\|^4\phi_\star + \psi_\star\right),
\end{aligned}$$

which yields the conclusion. □

Lemma D.1. *When considering Model (2), for any individual $i = 1, \ldots, n$, and any $k \in \mathbb{N}^*$, the function*

$$\boldsymbol{\theta} \mapsto \int_{\mathbb{R}^q} \mathbf{w}^k p_{\boldsymbol{\theta}}(\mathrm{d}\mathbf{w} \mid \mathbf{Y}_i)$$

is continuous on $\mathbb{R}^d$.

Proof. Let $i \in \{1, \ldots, n\}$, $k \in \mathbb{N}^*$, and $\Theta \subset \mathbb{R}^d$ a non-empty bounded and open set. For any $\mathbf{w} \in \mathbb{R}^q$ and $\boldsymbol{\theta} \in \Theta$, the Bayes rule yields

$$p_{\boldsymbol{\theta}}(\mathbf{w} \mid \mathbf{Y}_i) = \frac{p_{\boldsymbol{\theta}}(\mathbf{Y}_i \mid \mathbf{w}_i)\mathcal{N}(\mathbf{w}; \mathbf{0}_q, \boldsymbol{I}_q)}{p_{\boldsymbol{\theta}}(\mathbf{Y}_i)}.$$

Given $\mathbf{w} \in \mathbb{R}^q$, $\boldsymbol{\theta} \mapsto p_{\boldsymbol{\theta}}(\mathbf{Y}_i \mid \mathbf{W})$ is continuous on Θ, and $\boldsymbol{\theta} \mapsto p_{\boldsymbol{\theta}}(\mathbf{Y}_i)$ is continuous and positive on Θ (see Proposition 4.1 and its proof). This proves the continuity of $\boldsymbol{\theta} \mapsto \mathbf{w}^k p_{\boldsymbol{\theta}}(\mathbf{w} \mid \mathbf{Y}_i)$ on Θ.

Additionally, Lemma A.1 and Proposition 4.3 state that there are real constants $K_i^\Theta > 0$, κ_i^Θ, and $\zeta_i > 0$, such that for any $\boldsymbol{\theta} \in \Theta$

$$\left|\mathbf{w}^k p_{\boldsymbol{\theta}}(\mathbf{w} \mid \mathbf{Y}_i)\right| \leq \frac{\|\mathbf{w}\|^k}{\zeta_i} \exp\left\{K_i^\Theta\|\mathbf{w}\| - \frac{1}{2}\|\mathbf{w}\|^2 + \kappa_i^\Theta - \frac{q}{2}\log(2\pi)\right\}.$$

The upper bound is Lebesgue integrable on $\mathbb{R}^q$ and independent of $\boldsymbol{\theta} \in \Theta$. Consequently, the dominated convergence theorem yields the continuity on any non-empty bounded and open set.

□